DANIEL LOUZADA FERNANDES

ABORDAGENS COMPUTACIONAIS BASEADAS EM MODELOS DE APRENDIZADO PROFUNDO E VOLTADAS AO DEFICIENTE VISUAL PARA GERAÇÃO E AVALIAÇÃO AUTOMÁTICA DE DESCRIÇÕES TEXTUAIS DE CENAS DE WEBINÁRIOS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Doctor Scientiae*.

Orientador: Fabio Ribeiro Cerqueira

Coorientador: Michel Melo da Silva

VIÇOSA - MINAS GERAIS
2024

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
F363a
2024

Fernandes, Daniel Louzada, 1993-
Abordagens computacionais baseadas em modelos de aprendizado profundo e voltadas ao deficiente visual para geração e avaliação automática de descrições textuais de cenas de webinários / Daniel Louzada Fernandes. – Viçosa, MG, 2024.
1 tese eletrônica (176 f.): il. (algumas color.).

Orientador: Fábio Ribeiro Cerqueira.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Informática, 2024.
Referências bibliográficas: f. 160-176.
DOI: https://doi.org/10.47328/ufvbbt.2024.588
Modo de acesso: World Wide Web.

1. Inteligência artificial. 2. Visão computacional. 3. Processamento de imagens. 4. Processamento de linguagem natural (Computação). 5. Dispositivos de autoajuda para pessoas com deficiência. 6. Pessoas com deficiência visual. I. Cerqueira, Fábio Ribeiro, 1971-. II. Universidade Federal de Viçosa. Departamento de Informática. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDD 22. ed. 006.37

Bibliotecário(a) responsável: Bruna Silva CRB-6/2552

DANIEL LOUZADA FERNANDES

ABORDAGENS COMPUTACIONAIS BASEADAS EM MODELOS DE APRENDIZADO PROFUNDO E VOLTADAS AO DEFICIENTE VISUAL PARA GERAÇÃO E AVALIAÇÃO AUTOMÁTICA DE DESCRIÇÕES TEXTUAIS DE CENAS DE WEBINÁRIOS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Doctor Scientiae*.

APROVADA: 17 de junho de 2024.

Assentimento:

Documento assinado digitalmente
gov.br
DANIEL LOUZADA FERNANDES
Data: 16/09/2024 13:13:15-0300
Verifique em https://validar.iti.gov.br

Daniel Louzada Fernandes
Autor

Documento assinado digitalmente
gov.br
FABIO RIBEIRO CERQUEIRA
Data: 16/09/2024 13:16:07-0300
Verifique em https://validar.iti.gov.br

Fabio Ribeiro Cerqueira
Orientador

Dedico este trabalho aos meus queridos avós, fontes inesgotáveis de amor e sabedoria. Ao avô, cuja presença sempre acompanhada de um delicioso café me ensinou o valor das pequenas pausas e do aconchego. À avó, cujo sorriso iluminado e encorajador sempre me inspirou a seguir em frente, mesmo nos momentos mais escuros. À outra avó, cujo cafuné carinhoso e abraço caloroso são lembranças constantes de um amor incondicional que me aquece o coração. Este trabalho é para vocês, que com simplicidade e ternura, transformaram cada dia da minha vida em uma jornada especial. Com todo o meu amor e gratidão.

AGRADECIMENTOS

Agradeço primeiramente a Deus por sempre ter estado comigo, me dando forças e coragem para superar os obstáculos do dia a dia. Por nunca ter me abandonado, tanto nos momentos tristes e difíceis quanto nos felizes da vida, iluminando sempre o meu caminho com Sua luz. Obrigado, Senhor, pela minha vida, saúde, família, amigos e por ter me permitido chegar ao término desta longa jornada.

Mesmo sendo difícil listar todos os envolvidos e correndo o risco de não mencionar alguém em específico, não poderia deixar de citar algumas pessoas que foram cruciais para a conclusão deste doutorado.

Meus agradecimentos profissionais vão inicialmente para o meu orientador, professor Fabio R. Cerqueira, por ter, mais uma vez, me aceitado como seu aluno. Foram cinco anos e três meses de trabalho remoto, enfrentando a pandemia de COVID-19 e três mudanças no tema da tese. Agradeço pelo empenho na orientação a distância, pelo tempo dedicado, pelo esforço em guiar esta pesquisa, mesmo fora de sua área principal, pela preocupação em transformar o estudante em um profissional, pelos necessários puxões de orelha, pelos diversos *podcasts* em intermináveis áudios no *WhatsApp*, e pelas incontáveis correções dos meus textos em PDF. Muito obrigado, Fabio, por toda a amizade ao longo desses nove anos em que trabalhamos juntos, desde o mestrado e as pesquisas com profissionais da saúde até a conclusão deste doutorado. Suas considerações foram fundamentais para a formação do doutor que hoje se apresenta.

Estendo meus agradecimentos ao professor Michel M. da Silva, que, desde sua recém-chegada ao departamento em 2020, aceitou de bom grado nos ajudar a conduzir este projeto de pesquisa. Sua parceria e seus ensinamentos foram fundamentais para o desenvolvimento desta tese. Agradeço por me introduzir ao mundo da Visão Computacional e por confiar na minha capacidade de concluir este trabalho. Expresso também minha gratidão ao professor Marcos Henrique F. Ribeiro, por toda a ajuda prestada, sempre a uma porta de distância, especialmente nos momentos em que meus orientadores não puderam estar presentes.

Aos demais professores e técnicos do Programa de Pós-Graduação em Ciência da Computação (PPGCC) e do Departamento de Informática (DPI) da Universidade Federal de Viçosa (UFV), que contribuíram de alguma forma para a minha formação e sempre proporcionaram a nós, estudantes, um excelente ambiente de ensino, pesquisa e extensão, meus sinceros agradecimentos. Obrigado por todo o apoio e acolhimento ao longo desses anos, tanto como pós-graduando quanto como professor substituto. As interações na cozinha ou nos corredores, sempre acompanhadas de um copo de café, foram momentos descontraídos e enriquecedores.

Agradeço também aos servidores e funcionários da UFV em geral, pela dedicação e empenho em fazer com que esta universidade pública funcione tão bem para nós, sendo reconhecida pelo seu selo de qualidade. O fato de a universidade possuir o título de "A Mais Bonita do Brasil" certamente contribuiu para tornar nossas idas e vindas diárias ao DPI, sob sol, chuva, frio ou à luz do luar, mais agradáveis.

Não esquecendo os membros do laboratório MaVILab (*Machine Vision and Intelligence Laboratory*), especialmente meus coorientados de TCC e de Iniciação Científica, que foram responsáveis por diferentes partes deste grande projeto que resultou na minha tese: Ricson Vilaça, Luísa Ferreira, Allan Lopes, Júlia Vieira, Júlia Lopes, Sophia Jorge e Erick Figueiredo, meus sinceros agradecimentos.

O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001. Quero agradecer também às agências de pesquisa e financiamento: Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e ao *Google Cloud Platform*, por todo suporte financeiro concedido em diferentes partes desta tese. A qualidade do trabalho não seria a mesma sem o auxílio de vocês.

Meus agradecimentos pessoais vão inicialmente para meus pais – Jocimar e Ana Lucia – por serem minha principal fonte de inspiração e amor, e que, em breve, também concluirão seus cursos de Pós-Graduação *Stricto Sensu* em Educação. Estendo meus agradecimentos aos meus irmãos – Gabriella e Lucas – e sobrinhos – Gabriel, Miguel, José Emmanuel e Sammuel – pela compreensão, paciência, atenção, carinho e apoio incondicional durante esta trajetória, especialmente nos momentos em que não pude estar presente. Agradeço imensamente a eles por sempre estarem comigo, buscando, mesmo de longe, meu bem-estar, saúde e educação. Agradeço também aos meus familiares por sempre me motivarem a seguir meus sonhos e por torcerem pelas minhas conquistas.

Em especial, quero agradecer aos meus amigos de laboratório e do programa de Pós-Graduação – Vinícius Paiva, Mariana Schaefer, Yago Santos, Isabela Gomes, Rubens Moraes, Lucas Mucida e Gabriel Franco – que estiveram comigo durante vários momentos ao longo desses últimos cinco anos. Desde o acesso ao departamento nas primeiras até as últimas horas do dia, passando pelos lanches e conversas nos laboratórios, café na lanchonete da BBT, corridas na reta da UFV, desespero nas aulas de PAA, preparações de seminários e qualificações, compartilhamento de materiais e artigos, videoaulas de indianos, jogatinas de *board games*, até os desânimos por experimentos mal-sucedidos, lágrimas, abraços, risadas e conquistas de resultados e publicações. Meus sinceros agradecimentos por todo o apoio. Este doutorado não seria o que foi sem a amizade de vocês!

Aos meus antigos amigos da época de mestrado, que hoje se encontram espalhados pelo mundo, tanto no mercado quanto na academia. Nos agradecimentos da minha dissertação, eu desejei que continuássemos amigos além das pilastras da UFV, e graças a Deus, minha prece foi atendida. Assim, agradeço a Alba, Aly, Charles, Fabio, Jonatas (*in memoriam*), Liliane, Victor, Renan, Bárbara, Guidson, Joana, Maurício, Roberta e Vinício pela longa e duradoura amizade.

Agradeço também a uma série de pessoas que passaram pela minha vida nesses anos e cujas amizades foram fundamentais: Amanda Moitinho, Angélica Silva, Sócrates Júnior, Gabriela Cruz, Matheus Ferreira, Romário Will, meus afilhados Bárbara Gonçalves e Lucas Neiva, Arthur Amaral e demais membros do SIGEOnPA, meu trio apocalíptico concurseiro (Isadora Ribeiro e Angelina Melo), meus companheiros da república ERMO (Rafael Rocha, Gabriel Coura, Rafael Iasczczaki e João Victor Fonseca), diversos amigos do curso de Ciência da Computação, membros da peteca da AAACE, pesquisadoras do LAMECC e VigSUS, meu grupo interdisciplinar de café, meus alunos, minha personal trainer Laís Silveira e minha psicóloga Kátia Cruz. E, por fim, aos meus amigos de Cachoeiro de Itapemirim, em especial Mariana Guimarães, Ludymilla Paineiras e Mayara Batista. Sou grato pelo companheirismo de vocês!

A charmosa cidade de Viçosa (também conhecida por Viciosa ou Viçosa City), que tem sido meu lar há nove anos, agradeço por todo o acolhimento. Por fim, a todos aqueles que contribuíram de alguma forma, meu sincero agradecimento!

"Every day we make it, we'll make it the best we can." (Jack Daniel Distillery)

RESUMO

FERNANDES, Daniel Louzada, D.Sc., Universidade Federal de Viçosa, junho de 2024. **Abordagens computacionais baseadas em modelos de aprendizado profundo e voltadas ao deficiente visual para geração e avaliação automática de descrições textuais de cenas de webinários**. Orientador: Fabio Ribeiro Cerqueira. Coorientado: Michel Melo da Silva.

Estudos recentes preveem que pelo menos 2,2 bilhões de pessoas no mundo sofrem de cegueira ou alguma deficiência visual (como a baixa visão) e que esse número continuará a crescer. Essas pessoas precisarão de algum tipo de cuidado apropriado e o uso de Tecnologias Assistivas é uma forma valiosa para que elas possam mitigar seus obstáculos diários. Nesse contexto, com o rápido avanço da Inteligência Artificial e dos sistemas portáteis embarcados, tem-se testemunhado um aumento no desenvolvimento e oferecimento de vários serviços e tecnologias que proporcionam comodidade e suporte para esse público. Apesar desses avanços, muitas dessas tecnologias têm fatores restritivos, como funcionalidades limitadas ou preços elevados. Além disso, estão disponíveis apenas para uma pequena parcela da população necessitada. Com a pandemia de COVID-19, a vida cotidiana e o local de trabalho tornaram-se mais dependentes das tecnologias, como o consumo intensivo de conteúdos online e o aumento significativo no uso de ferramentas de videoconferência. Embora um mundo altamente conectado permita o trabalho remoto como substituto para o deslocamento e o trabalho de escritório – assim como webinários/videoconferências como sucessores de conferências presenciais, entrevistas, reuniões ou até mesmo aulas – isso também levanta novas barreiras de acessibilidade para as pessoas com deficiência visual. Como a informação visual é complementar à própria mensagem, a baixa ou nenhuma visão impede que essas pessoas capturem informações visuais, o que pode dificultar a compreensão do contexto geral do conteúdo compartilhado em uma apresentação remota. Com isso, aumentou-se a necessidade de prover mais acesso a informações contidas em webinários, em especial, sobre contexto. Para suprir essa necessidade, iniciativas vêm sendo realizadas no sentido de incentivar os usuários da Internet a produzirem descrições textuais de imagens on-line. No entanto, esse é um processo manual e lento que depende da disposição das pessoas com visão em ajudar. Como consequência, muitas imagens carecem de descrições ou apresentam explicações de baixa qualidade. A maioria dos métodos existentes na literatura sobre descrição automática de imagens baseados em Inteligência Artificial, quando utilizados como Tecnologias Assistivas, negligencia as necessidades de indivíduos cegos ou com

baixa visão. Esses métodos tendem a comprimir todos os elementos visuais em legendas breves, criar frases desconexas para cada região da imagem ou fornecer descrições extensas, não se concentrando no fornecimento das informações pertinentes para esse grupo específico. Isso ocorre também devido à escassez de conjuntos de dados específicos para atender necessidades de deficientes visuais; logo, esses métodos são treinados em conjuntos para domínios de dados gerais, considerando o uso por pessoas com visão. Para lidar com essas limitações, nesta tese, propõe-se um conjunto de metodologias por meio da integração de técnicas de Visão Computacional e Processamento de Linguagem Natural que possibilitam a implementação e avaliação de uma abordagem para construir descrições de imagens baseada em normas e diretrizes de acessibilidade direcionadas a pessoas com deficiência visual, focando em cenas de webinários. Como parte do processo, o trabalho também desenvolve um conjunto de dados direcionado para este público e propõe uma métrica de avaliação de adequabilidade de descrição textual, levando em conta os aspectos importantes para pessoas cegas ou de baixa visão. Os experimentos demonstraram estatisticamente que a abordagem proposta produziu descrições alinhadas com o conteúdo das imagens, com características linguísticas escritas por humanos e com as diretrizes de acessibilidade para deficientes visuais, apresentando melhor desempenho nesses aspectos quando comparada a métodos anteriores de descrição de imagens.

Palavras-chave: Inteligência artificial. Visão computacional. Processamento de linguagem natural. Descrição de imagens. Tecnologias assistivas. Deficiente visual.

ABSTRACT

FERNANDES, Daniel Louzada, D.Sc., Universidade Federal de Viçosa, June, 2019. **Computational approaches based on deep learning models and aimed at the visually impaired for automatic generation and evaluation of textual descriptions of webinar scenes**. Advisor: Fabio Ribeiro Cerqueira. Co-adviser: Michel Melo da Silva.

Recent studies predict that at least 2.2 billion people worldwide suffer from blindness or some form of visual impairment (such as low vision) and that this number will continue to grow. These people will require some form of appropriate care and the use of Assistive Technologies is a valuable way for them to mitigate their daily obstacles. In this context, with the rapid advancement of Artificial Intelligence and portable embedded systems, there has been an increase in the development and provision of various services and technologies that provide convenience and support to this audience. Despite these advancements, many of these technologies have restrictive factors such as limited functionality or high prices. Additionally, they are available to only a small portion of the population in need. With the COVID-19 pandemic, daily life and workplace have become more dependent on technologies, such as the intensive consumption of online content and the significant increase in the use of videoconferencing tools. While a highly connected world allows for remote work as a substitute for commuting and office work – as well as webinars/videoconferences as successors to in-person conferences, interviews, meetings, or even classes – it also raises new accessibility barriers for visually impaired people. Since visual information complements the message itself, low or no vision prevents these individuals from capturing visual information, which can make it difficult for them to understand the overall context of the content shared in a remote presentation. Consequently, there has been an increased need to provide more access to the information contained in webinars, particularly regarding context. To meet this need, initiatives have been carried out to motivate Internet users to produce textual descriptions of online images. However, this is a manual and slow process that depends on the willingness of sighted people to help. As a result, many images lack descriptions or have low-quality explanations. Most existing methods in the literature for automatic image description based on Artificial Intelligence, when used as Assistive Technologies, neglect the needs of blind or low-vision individuals. These methods tend to compress all visual elements into brief captions, create disjointed sentences for each region of the image, or provide extensive descriptions without focusing on providing the relevant information for this specific group. This

is also due to the scarcity of datasets specifically designed to meet the needs of visually impaired individuals; thus, these methods are trained on datasets for general domains, considering use by sighted people. To address these limitations, this thesis proposes a set of methodologies through the integration of Computer Vision and Natural Language Processing techniques that enable the implementation and evaluation of an image description approach based on accessibility standards and guidelines aimed at visually impaired people, focusing on webinar scenes. As part of the process, this work also develops a dataset targeted at this audience and proposes a metric for evaluating the adequacy of textual descriptions, taking into account the important aspects for blind or low-vision individuals. The experiments statistically demonstrated that the proposed approach produced descriptions aligned with the content of the images, with human-written linguistic characteristics, and with accessibility guidelines for visually impaired individuals, presenting better performance in these aspects when compared to previous image description methods.

Keywords: Artificial intelligence. Computer vision. Natural language processing. Image description. Assistive technologies. Visually impaired.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
amod	*Adjectival Modifier*
API	*Application Programming Interface*
BART	*Bidirectional and Auto-Regressive Transformers*
BERT	*Bidirectional Encoder Representations from Transformers*
BIC	Bolsistas de Iniciação Científica
BLEU	*BiLingual Evaluation Understudy*
BiLSTM	*Bidirectional Long Short-Term Memory*
BLIP	*Bootstrapping Lang.-Image Pre-train. for Unified Vision-Lang. Underst. and Gen.*
bbox	*Bounding Box*
CAVP	*Context-Aware Visual Policy*
CBV	Cegueira ou Baixa Visão
CIDEr	*Consensus-based Image Description Evaluation*
CLIP	*Contrastive Language-Image Pre-Training*
CNN	*Convolutional Neural Network*
CoLA	*Corpus of Linguistic Acceptability*
CO_2	Dióxido de Carbono
CPU	*Central Processing Unit*
C4	*Colossal Clean Crawled Corpus*
DAN	*Deep Averaging Network*
DEX	*Deep EXpectation*
DistilRoBERTa	*Distilled version of RoBERTa*
DOI	*Digital Object Identifier*
FAPEMIG	Fundação de Amparo à Pesquisa do Estado de Minas Gerais
FCLN	*Fully Convolutional Localization Network*
FER	*Facial Expression Recognition*
FID	*Fréchet Inception Distance*
GAN	*Generative Adversarial Network*
GB	*Gigabyte*
GHz	*Gigahertz*
gLSTM	*Guiding Long Short-Term Memory*
GPU	*Graphics Processing Unit*

GRIT	*Grid- and Region-based Image captioning Transformer*
GRU	*Gated Recurrent Unit*
GT	*Ground-Truth*
HRNN	*Hierarchical Recurrent Neural Network*
IA	Inteligência Artificial
IDF	*Inverse Document Frequency*
InViDe	*Inclusive Visual Description*
LAION	*Large-scale Artificial Intelligence Open Network*
LLaVA	*Large Language and Vision Assistant*
LLM	*Large Language Model*
LSTM	*Long Short-Term Memory*
LTS	*Long-Term Support*
METEOR	*Metric for Evaluation of Translation with Explicit ORdering*
MLP	*Multi-Layer Perceptron*
MS COCO	*Microsoft Common Objects in Context*
MTCNN	*Multi-task Cascaded Convolutional Network*
NLTK	*Natural Language Toolkit*
NSFW	*Not Safe For Work*
nsubj	*Nominal Subject*
npadvmod	*Noun Phrase as Adverbial Modifier*
NUBIA	*NeUral Based Interchangeability Assessor for Text Generation*
O	Notação *Big* O
OE	Objetivo Específico
OFA	*One For All*
OMS	Organização Mundial da Saúde
PAZ	*Perception for Autonomous Systems*
PEGASUS	*Pre-training with Extracted Gap-sentences for Abstractive Summarization*
PLM	*Pre-trained Language Model*
PLN	Processamento de Linguagem Natural
pobj	*Object of a Preposition*
POS	*Part-of-Speech*
PROPN	*Proper Noun*
RAM	*Random Access Memory*
ReLU	*Rectified Linear Unit*
ResNet	*Residual Network*
RNN	*Recurrent Neural Network*

RoBERTa	*Robustly Optimized BERT Pretraining Approach*
ROI	*Region of Interest*
RV	*Realistic Vision*
SD1.4	*Stable Diffusion 1.4*
SD2.0	*Stable Diffusion 2.0*
SFW	*Safe For Work*
SIBGRAPI	*Conference on Graphics, Patterns and Images*
SIP	*Stanford Image-Paragraph*
SGD	*Stochastic Gradient Descent*
TF-IDF	*Term Frequency-Inverse Document Frequency*
TOR	*Teachable Object Recognizer*
TPU	*Tensor Processing Unit*
T5	*Text-to-Text-Transfer-Transformer*
UFV	Universidade Federal de Viçosa
VAE	*Variational Auto-Encoder*
VG	*Visual Genome*
VGGNet	*Visual Geometry Group Network*
VIIDA	*Visually Impaired Image Description Approach*
VISAPP	*International Conference on Computer Vision Theory and Applications*
VISIGRAPP	*Inter. Joint Conf. on Comp. Vision, Imaging and Comp. Graphics Theory and App.*
ViT	*Vision Transformer*
VQA	*Visual Question Answering*
VLP	*Vision-Language Pre-training*
VRAM	*Video Random Access Memory*
YOLO	*You Only Look Once*

SUMÁRIO

Capítulo 1

Introdução

Estatísticas sobre deficiência visual (por exemplo, cegueira, baixa visão, fotofobia, diplopia, distorção visual, etc.) revelaram que pelo menos 2,2 bilhões de pessoas no mundo sofrem de alguma deficiência visual, devido principalmente a erros de refração, catarata, retinopatia diabética, glaucoma e degeneração macular relacionada à idade (Flaxman et al. (2017); Steinmetz et al. (2021)). Estudos recentes preveem que esse número continuará a crescer em todo o mundo, e essas pessoas precisarão de algum tipo de cuidado apropriado (Noorjahan and Punitha (2020); Patel and Parmar (2022)). Sendo a visão o mais dominante dos nossos sentidos, ela desempenha um papel fundamental na vida de todos (Steinmetz et al. (2021); Patel and Parmar (2022)). Nesta tese, o termo "deficiência visual" irá se referir tanto a cegueira quanto a baixa visão, independentemente do grau.

O uso de Tecnologias Assistivas, como o aplicativo *Seeing AI*[1], é uma forma valiosa para que pessoas com cegueira ou baixa visão (CBV) possam mitigar seus obstáculos diários e melhorar suas habilidades cognitivas (Senjam et al. (2020)). Embora haja uma ampla gama de Tecnologias Assistivas disponíveis que promovem uma melhoria significativa na vida dessas pessoas, essas tecnologias têm funcionalidade limitada, preços altos, e estão acessíveis apenas para uma pequena parcela da população necessitada (Guo et al. (2022); Patel and Parmar (2022); Dokania and Chattaraj (2024)).

Recentemente, com o rápido avanço da Inteligência Artificial (IA) e dos sistemas portáteis embarcados, observou-se um aumento no desenvolvimento e na oferta de diversos produtos, serviços e tecnologias voltados para indivíduos com CBV. Exemplos incluem textos alternativos automáticos para fotos em redes sociais, sistemas que reconhecem objetos e fornecem retorno audível ou descrito em braille, aplicativos para auxiliar na navegação em espaços internos, entre outros (Wu et al. (2017); Alhichri et al. (2019); Guo et al. (2022); Dokania and Chattaraj (2024)).

A literatura também apresenta estudos recentes nas áreas de IA e Tecnologia Assistiva, que se concentram no desenvolvimento de sensores e métodos computacionais por meio de Visão Computacional, Processamento Digital de

[1]Disponível em https://www.seeingai.com/

Imagens e Processamento de Linguagem Natural (PLN). Esses avanços proporcionam maior comodidade e suporte para esse público específico (Gurari et al. (2020); Noorjahan and Punitha (2020); Granquist et al. (2021); Shah et al. (2021); dos Santos et al. (2022); Guo et al. (2022); Dokania and Chattaraj (2024)).

Apesar desses avanços tecnológicos, indivíduos com deficiência visual ainda enfrentam desafios significativos, como a navegação em seus ambientes, a realização de compras, o reconhecimento de pessoas e objetos ao seu redor, e a compreensão de conteúdos em noticiários de TV, vídeos on-line, plataformas de mídia social ou webinários (Stangl et al. (2020)). Os webinários, que são seminários ou conferências realizados pela Internet, geralmente no domínio educacional, envolvem um ou mais palestrantes ou apresentadores que se expressam enquanto outras pessoas assistem.

1.1 Contextualização

A introdução de novos hábitos durante a pandemia de COVID-19 integrou amplamente a vida das pessoas com conteúdo on-line, ferramentas de videoconferência e interações remotas (Wanga et al. (2020); Doush et al. (2023)). Por exemplo, a prática conhecida como *home office* (exercer o trabalho a partir de casa) foi amplamente utilizada como substituta para o deslocamento dos profissionais até seus locais de trabalho usuais. Para interações como conferências, entrevistas, reuniões e aulas, dentre outros, a prática substituta foi o amplo uso de webinários.

Independentemente do aumento significativo no uso de ferramentas de videoconferência, elas ainda apresentam expressivas limitações em relação à acessibilidade. Isto é, há uma falta de ferramentas acessíveis que permitam que pessoas com deficiência visual compreendam o contexto desses recursos on-line, garantindo acesso igualitário (Gurari et al. (2020); Simons et al. (2020)). Uma parte significativa desse conteúdo fornecido é exclusivamente visual, o que consiste em uma barreira de acesso para pessoas com sérias limitações de visão.

Na literatura atual, apenas um pequeno número de estudos explora questões de acessibilidade relacionadas ao uso de ferramentas de videoconferência para pessoas com deficiência visual (Acosta-Vargas et al. (2021); Leporini et al. (2021); Doush et al. (2023)). Neste mundo altamente conectado, as pessoas com CBV lutam para se socializar e entender o contexto do conteúdo compartilhado.

A informação visual é complementar à mensagem sendo transferida. Por exemplo, um aviso solicitando distanciamento social vindo de um membro da linha de frente da saúde, em um hospital lotado, é mais sensível aos ouvintes do que se o mesmo texto de aviso viesse de um empresário falando a partir de seu escritório em um arranha-céu. A informação textual permaneceria a mesma. O que diferenciaria os teores das mensagens seria a localização, ligada a contexto, neste exemplo. Assim,

pessoas com baixa ou nenhuma visão são privadas exatamente desse contexto e sua compreensão pode ficar limitada ou até mesmo comprometida.

O contexto visual desempenha um papel fundamental na comunicação, pois adiciona camadas de significado que vão além do conteúdo textual. Ele pode transmitir emoções, intenções, e até mesmo a urgência ou seriedade de uma mensagem, influenciando como essa mensagem é percebida e entendida pelo receptor. No exemplo dado, a localização de quem transmite a mensagem carrega informações contextuais que alteram a forma como o aviso sobre distanciamento social é recebido. No primeiro caso, a urgência e a seriedade da mensagem são amplificadas pela gravidade do ambiente hospitalar durante uma pandemia, o que sensibiliza os ouvintes para a importância de seguir as orientações. Já no segundo caso, a mesma mensagem, vinda de um contexto menos impactante, pode não ter o mesmo efeito emocional ou cognitivo, mesmo que as palavras sejam idênticas. Assim, a ausência desse contexto visual prejudica a capacidade de pessoas com deficiência visual reagirem de forma adequada às situações comunicadas, devido à falta de detalhes nas informações transmitidas pela expressão facial, pela postura e pelo cenário em geral.

1.2 Problema e sua Importância

Para indivíduos com CBV, ter acesso a ferramentas que elucidem o conteúdo de imagens digitais é algo extremamente relevante, pois tais ferramentas têm o potencial de auxiliar até mesmo no caso de vídeos, uma vez que decodificar imagens também ajuda a descrever o conteúdo de vídeos. Isto porque os mesmos podem ser analisados a partir de seus quadros individuais, ou seja, de imagens isoladas.

Dada uma imagem, uma possível abordagem para decifrá-la seria o fornecimento de uma descrição textual sobre seu conteúdo, de modo que esse texto seja dado como entrada para uma ferramenta de narração. Nos últimos anos, houve algumas iniciativas que incentivaram as pessoas a produzir descrições textuais de imagens on-line (por exemplo, o projeto brasileiro #PraCegoVer ou #ParaTodosVerem (dos Santos et al. (2022))). No entanto, esse é um processo manual e lento, que depende da disposição das pessoas com visão em ajudar. Como consequência, muitas imagens carecem de descrições ou apresentam explicações de baixa qualidade.

Em relação às abordagens automáticas baseadas em IA, as tecnologias atuais de descrição de imagens muitas vezes falham em considerar as preferências, necessidades e diretrizes de acessibilidade da população com baixa ou nenhuma visão (Stangl et al. (2020)). Assim, o desafio está em arquitetar descrições eficazes que transmitam o contexto da imagem junto com as informações significativas necessárias para o usuário com deficiência visual.

As técnicas de legendagem de imagens podem ser aplicadas como Tecnologia Assistiva para fornecer descrições rápidas e automáticas do conteúdo das imagens (Gurari et al. (2020)), ou dos quadros de um vídeo. Apesar dos avanços recentes e resultados notáveis, essas técnicas são limitadas quando aplicadas para extrair o contexto das imagens, pois comprimem todos os elementos visuais em uma única frase. Isso resulta em descrições que negligenciam os detalhes cognitivos e visuais necessários para pessoas com deficiência visual, produzindo uma legenda genérica, simples e de alto nível (Ng et al. (2021); Dognin et al. (2022)), como apresentado na Figura 1.1. Portanto, essas técnicas ainda apresentam lacunas como Tecnologias Assistivas.

Figura 1.1: Exemplos de legendas produzidas por uma técnica de legendagem de imagem. É possível observar a simplicidade de detalhes nas legendas. Fonte: Adaptado de Karpathy and Fei-Fei (2015).

As técnicas de legendagem densa (Johnson et al. (2016)) e parágrafo de imagem (Krause et al. (2017)) surgiram como abordagens promissoras e eficazes para gerar descrições detalhadas de todo o conteúdo visual de uma imagem, visando lidar com as limitações da legendagem de imagem tradicional (Chatterjee and Schwing (2018)). Essas técnicas detectam informações relevantes das imagens, destacando objetos e regiões de interesse, cujas características são extraídas e subsequentemente agrupadas em uma representação que expressa ricamente a semântica da imagem (Krause et al. (2017)). Para alcançar seus objetivos, essas técnicas capturam todos os elementos salientes de uma cena e os combinam para descrever a imagem por meio de múltiplas sentenças.

Embora ambas as técnicas mencionadas tenham proporcionado um importante avanço no problema de geração de descrições de imagens, sendo capazes de gerar frases conectadas e informativas (Allen and Smith (2020); Ng et al. (2021)), elas também falham na tarefa de extrair o contexto da imagem para indivíduos com deficiência visual. A técnica de legendagem densa não sintetiza múltiplas informações em sentenças estruturadas e coerentes (Krause et al. (2017)). Além disso, mesmo para pessoas com visão, entender o resultado produzido por essa técnica é uma tarefa desafiadora sem o auxílio de uma interface visual que proporcione um

Figura 1.2: Exemplos de saídas de uma técnica de legendagem densa. Cada frase de uma cor representa uma legenda densa do seu respectivo quadro. Na imagem à direita, nota-se a dificuldade em se conseguir compreender o contexto da imagem devido à sobrecarga de informação. Fonte: Adaptado de Johnson et al. (2016).

melhor entendimento das descrições retornadas (Figura 1.2).

Por outro lado, a técnica de parágrafo de imagem, com seus parágrafos longos e densos (Figura 1.3), frequentemente não mantêm a atenção do espectador com CBV, uma vez que conta uma história a partir da imagem negligenciando detalhes contextuais e relações espaciais entre os objetos detectados, o que leva a um desempenho insatisfatório em termos de ordem das palavras e preposições (Delloul and Larabi (2022)). A falta de ênfase nessas relações espaciais e no contexto resulta em descrições que podem ser confusas ou pouco úteis para uma pessoa com deficiência visual, que depende dessas informações para formar uma imagem mental precisa e completa da cena. Além disso, a ordem das palavras e o uso correto das preposições são fundamentais para a compreensão do texto, especialmente em descrições que precisam ser precisas e diretas. Se esses aspectos não são bem

Two children are sitting at a table in a restaurant. The children are one little girl and one little boy. The little girl is eating a pink frosted donut with white icing lines on top of it. The girl has blonde hair and is wearing a green jacket with a black long sleeve shirt underneath. The little boy is wearing a black zip up jacket and is holding his finger to his lip but is not eating. A metal napkin dispenser is in between them at the table. The wall next to them is white brick. Two adults are on the other side of the short white brick wall. The room has white circular lights on the ceiling and a large

Figura 1.3: Exemplo de um parágrafo de imagem. Fonte: Adaptado de Krause et al. (2017).

executados, a descrição pode se tornar de difícil compreensão por parte da pessoa com CBV.

À vista disso, as técnicas de legendagem densa e de parágrafo de imagem também são limitadas e não adequadas para que pessoas com deficiência visual compreendam plenamente uma cena de videoconferência. Isso ocorre porque a maioria das técnicas de IA são específicas para domínios de dados de pessoas com visão. Este fato também pode ser explicado pela escassez de conjuntos de dados direcionados para o contexto de aplicações para deficientes visuais.

1.3 Hipótese

É possível desenvolver uma abordagem computacional que produza automaticamente descrições textuais inclusivas do contexto de imagens de webinários para pessoas com deficiência visual, garantindo a presença de informações relevantes e necessárias para esse público, por meio da aplicação de modelos de Aprendizado Profundo.

1.4 Questões de Pesquisa

As seguintes questões, relacionadas ao problema apresentado, guiam a discussão por esta tese:

- Os modelos existentes de descrição de imagens são capazes de descrever cenas de webinários para pessoas com deficiência visual de maneira inclusiva? Quais

erros estão presentes em suas saídas? Os conjuntos de dados de treinamento e as métricas de avaliação são adequados para esse problema?

- Como desenvolver um método capaz de gerar descrições precisas e coerentes a partir de imagens, capturando não apenas objetos, mas também as relações contextuais entre eles?
- Quais técnicas de Aprendizado Profundo e arquiteturas de redes neurais são mais adequadas para combinar informações visuais e textuais de maneira eficaz em um método de descrição de imagens?
- Quais são as informações relevantes e necessárias em imagens para deficientes visuais, e como garantir que o método produza descrições imparciais e inclusivas?
- Como garantir que o método seja robusto o suficiente para lidar com uma ampla variedade de cenários visuais?
- De que forma o pré-treinamento em grandes conjuntos de dados (visuais e textuais) pode impactar a qualidade das descrições geradas?
- Quais métricas de avaliação são mais apropriadas para medir tanto a qualidade quanto a precisão semântica das descrições geradas?
- É possível avaliar a adequação de descrições geradas a partir de imagens?
- Como minimizar o viés inerente nos conjuntos de dados visuais e textuais frequentemente utilizados, e como incorporar a diversidade na geração de descrições?
- Quais são os desafios éticos envolvidos na geração automática de descrições de imagens, e como eles podem ser abordados?

1.5 Objetivos

A motivação para este trabalho foi contribuir para o campo da Tecnologia Assistiva, com o propósito de mitigar as barreiras de acessibilidade e proporcionar melhorias na vida de pessoas com deficiência visual. Diante dessa motivação, da importância do problema discutido e das questões de pesquisa elaboradas, o objetivo principal desta tese foi desenvolver abordagens computacionais baseadas em modelos de Aprendizado Profundo que descrevam automaticamente cenas de webinários para pessoas com CBV com precisão. Isso permitirá que essas pessoas entendam, principalmente, as características do apresentador e de seu entorno, obtendo uma

melhor compreensão da mensagem do webinário e de seu contexto, sem depender de um usuário com visão.

Além do objetivo principal, esta tese propôs como objetivos específicos:

- OE-1: Desenvolver e avaliar uma abordagem computacional inicial para geração de descrições de imagens de webinários para produzir textos mais interpretáveis e focadas nas informações relevantes para pessoas com deficiência visual;
- OE-2: Investigar o impacto da adequação dos conjuntos de dados e métricas de avaliação existentes na tarefa de geração de descrições textuais para pessoas com CBV, com base nos resultados obtidos pela abordagem inicial;
- OE-3: Criar uma abordagem automática para coleta de imagens de cenas de webinários a partir de vídeos;
- OE-4: Desenvolver um protocolo de anotação para descrever manualmente o contexto visual das imagens de cenas de webinários coletadas, isto é, as características do apresentador em destaque e do seu entorno, seguindo as normas e práticas apropriadas recomendadas para o público cego e com baixa visão;
- OE-5: Construir um conjunto de dados composto das imagens de cenas de webinários coletadas e anotadas manualmente;
- OE-6: Desenvolver e validar uma abordagem computacional aprimorada que, utilizando modelos de Aprendizado Profundo, seja capaz de extrair características visuais relevantes das imagens de webinários e gerar descrições textuais coerentes e estruturadas baseadas nessas características extraídas;
- OE-7: Verificar a possibilidade de geração de conteúdos prejudiciais e identificar questões éticas e sociais associadas às abordagens desenvolvidas;
- OE-8: Construir uma métrica de avaliação automática, alinhada às diretrizes e boas práticas de acessibilidade recomendadas, para verificar o nível de adequação das descrições textuais de imagens produzidas para indivíduos com deficiência visual.

1.6 Aspectos Éticos

Esta tese não envolveu a participação de seres humanos. Desta forma, não foi necessário submetê-la à revisão pelo Comitê de Ética em Pesquisa da Universidade Federal de Viçosa.

1.7 Organização do Trabalho

Além da introdução já apresentada, esta tese foi estruturada em mais seis capítulos, que serão descritos a seguir:

- O Capítulo 2 e o Capítulo 3 fornecem uma explicação sobre o referencial teórico e os trabalhos relacionados envolvidos neste estudo;
- No Capítulo 4, são apresentados os detalhes da primeira abordagem computacional proposta e os resultados encontrados durante a investigação inicial para a resolução do problema de descrição de imagens para pessoas com deficiência visual (OE-1 e 2). Este capítulo refere-se ao trabalho (Fernandes. et al. (2022)).
- O Capítulo 5 fornece detalhes sobre a criação e anotação do conjunto de dados de domínio específico (descrição de cenas de webinários) proposto para o problema a ser resolvido nesta tese (OE-3 a 5). Este capítulo refere-se ao trabalho (Ferreira et al. (2023)).
- No Capítulo 6, são explicados, principalmente, a visão geral e os detalhes da abordagem computacional aprimorada para geração de descrição de imagens, juntamente com a métrica de adequação proposta (OE-6 a 8). Este capítulo refere-se ao trabalho (Fernandes et al. (2024)).
- Finalmente, no Capítulo 7, são apresentadas as conclusões da tese, juntamente com as suas contribuições e futuras orientações para a pesquisa.

1.8 Publicações

Os experimentos do Capítulo 6 foram incorporados em um artigo que foi submetido e que se encontra em processo final de revisão no periódico *Disability and Rehabilitation: Assistive Technology*. A resposta à primeira decisão foi favorável à publicação:

- Fernandes, Daniel L.; Ribeiro, Marcos H. F.; Silva, Michel M.; Cerqueira, Fabio R. **VIIDA and InViDe: Computational approaches for generating and evaluating inclusive image paragraphs for the visually impaired**. Disability and Rehabilitation: Assistive Technology, [2024?]. No prelo.

Os resultados do Capítulo 4 e do Capítulo 5 desta tese foram publicados nas conferências VISAPP em 2022 e no SIBGRAPI em 2023:

- Fernandes, D.; Ribeiro, M.; Cerqueira, F.; Silva, M. **Describing image focused in cognitive and visual details for visually impaired people: An approach to generating inclusive paragraphs**. In Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 5: VISAPP. Setúbal: SciTePress, 2022. p. 526-534. DOI: <https://doi.org/10.5220/0010845700003124>.

- Ferreira, L.; Fernandes, D.; Cerqueira, F.; Ribeiro, M.; Silva, M. **Presenter-Centric Image Collection and Annotation: Enhancing Accessibility for the Visually Impaired**. In Proceedings of the 36th Conference on Graphics, Patterns and Images (SIBGRAPI). Rio Grande: IEEE, 2023. p. 199-204. DOI: <https://doi.org/10.1109/SIBGRAPI59091.2023.10347135>.

Capítulo 2

Referencial Teórico

Este capítulo tem como objetivo apresentar os principais conceitos e características relacionados ao tema desta tese. A seguir, é realizada uma revisão da literatura, descrevendo as principais arquiteturas de redes neurais que serviram de base para o desenvolvimento dos modelos de Aprendizado Profundo estudados, utilizados e comparados neste trabalho.

2.1 Redes Neurais Convolucionais

As Redes Neurais Convolucionais (do inglês, *Convolutional Neural Networks* - CNNs) são técnicas de Aprendizado Profundo aplicadas principalmente ao processamento de informações visuais, como classificação de imagens, detecção de objetos, extração de características, entre outros (Ponti et al. (2017)). As CNNs têm o objetivo de reduzir o tamanho das imagens de entrada para um formato dimensional mais simples de ser processado, sem que haja perda das características fundamentais das imagens.

Basicamente, as CNNs são estruturadas por meio da alternância entre as suas camadas, destacando-se as camadas de convolução, *pooling*, *flatten* e totalmente conectadas. No entanto, com exceção das camadas de convolução, as demais camadas não são obrigatórias nessas estruturas. A Figura 2.1 apresenta um possível exemplo de CNN. Neste exemplo, ao receber uma imagem como entrada, é realizada a extração das suas características por meio de diferentes tipos de camadas (como, convolução e *pooling*) e de funções (por exemplo, ReLU) para a posterior compactação destas características em uma representação vetorial (*flatten*). Em seguida, esse vetor de características é fornecido como entrada de camadas totalmente conectadas que irão realizar ao fim e, por meio da distribuição de probabilidades gerada pela função *softmax*, uma tarefa de classificação.

As camadas convolucionais são responsáveis por executar a operação de convolução, que funciona via deslizamento dos filtros convolucionais, também conhecidos como conjuntos de parâmetros ou matrizes de pesos, por toda a imagem de entrada durante o treinamento. Cada filtro destina-se a detectar algum tipo de característica presente na imagem (Guo et al. (2016)). Matematicamente, a operação

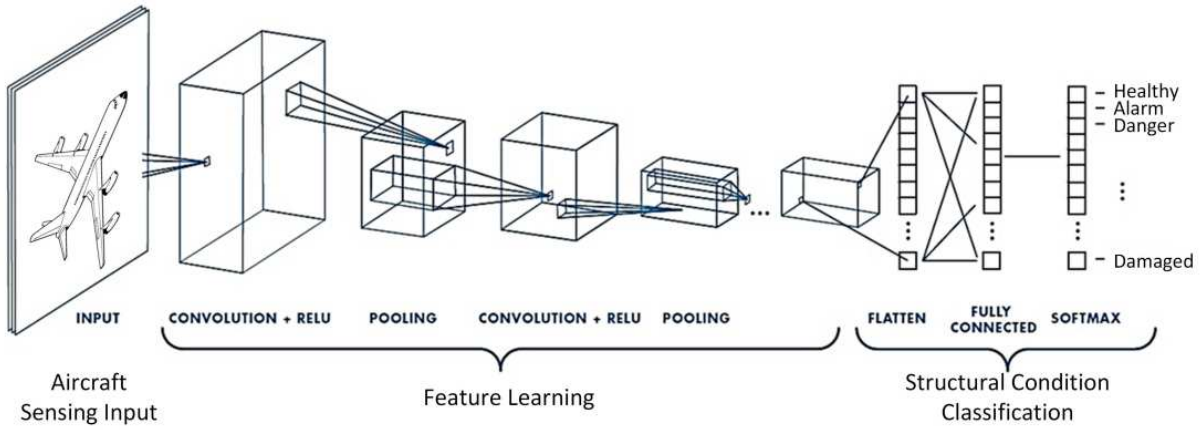


Figura 2.1: Exemplo de arquitetura de uma CNN para tarefa de monitoramento da integridade estrutural de aeronaves. Fonte: Tabian et al. (2019).

de convolução é representada pela soma do produto das matrizes de pesos pelas regiões da imagem. Inicialmente, os filtros da primeira camada de convolução atuam diretamente nos *pixels* das imagens de entrada, extraindo padrões mais simples, como linhas, brilho ou bordas, e propagando-os à frente na rede. Os filtros das próximas camadas irão atuar nas saídas das camadas anteriores, denominadas mapas de ativação, e tenderão a extrair padrões mais complexos, como formas geométricas, com base nos padrões antecedentes detectados.

Os pesos são inicializados aleatoriamente e atualizados de acordo com o algoritmo de otimização durante o processo de retropropagação da rede, podendo ser a cada nova entrada, a cada novo *batch* ou a cada época. Após a camada de convolução, é aplicada uma função de ativação responsável por trazer a não-linearidade ao modelo, permitindo assim que a rede tenha capacidade representativa (Goodfellow et al. (2016)). Dentre as funções de ativação existentes, a ReLU é amplamente utilizada nas CNNs. A saída de cada camada de convolução seguida pela função de ativação é usada como entrada para a próxima camada.

A camada de *pooling* pode suceder a camada de convolução e é responsável por reduzir a dimensionalidade dos mapas de ativação gerados, mantendo somente as informações mais relevantes detectadas e, consequentemente, diminuindo a quantidade de parâmetros treinados pela rede e as chances de *overfitting* (Wu and Gu (2015)). A ideia por trás dessa camada é que *pixels* vizinhos em uma imagem são altamente correlacionados, portanto, muitas vezes, torna-se possível identificar informações relevantes em uma determinada região da imagem com apenas o valor máximo, médio ou mínimo daquela região (Goodfellow et al. (2016)). Assim, a camada de *pooling*, que não possui parâmetros a serem aprendidos, aplica uma técnica que permite comprimir os valores de cada área dos mapas de ativação por meio do deslizamento de um filtro de tamanho f considerando o valor s do parâmetro *stride*, sendo este a quantidade de movimentos (saltos ou passos)

realizados pelos filtros por iteração no mapa de ativação de entrada da camada de *pooling*. Se $s > f$, o filtro não será deslizado por todo o mapa de ativação, pois ultrapassará o limite das suas dimensões. Logo, a parte do mapa não analisada será desconsiderada.

Diferente dos filtros que são aprendidos durante o processo de treinamento, as suas dimensões (altura, largura e profundidade), o tamanho do *stride* e o *padding* são hiperparâmetros, logo, precisam ser definidos antes de iniciar o processo de treinamento das CNNs. O *padding* geralmente é utilizado quando se deseja evitar que o tamanho da imagem, isto é, a quantidade de informação, seja reduzido rapidamente devido à utilização de várias camadas de convolução (Goodfellow et al. (2016)). Dessa forma, o *padding* pode acrescentar linhas e colunas preenchidas pelo valor zero ao redor da matriz da imagem antes da operação de convolução, de forma a manter a dimensionalidade.

Atualmente, existem diversas arquiteturas de CNNs disponíveis na literatura que têm sido fundamentais para a construção de modelos para as mais diversas tarefas de Visão Computacional. As arquiteturas mais comuns são a LeNet, AlexNet, GoogLeNet, Inception, VGGNet e ResNet. Dentre elas, as arquiteturas das famílias VGGNet (Simonyan and Zisserman (2015)) e ResNet (He et al. (2016)) foram as mais aplicadas nos modelos de legendagem de imagem dos últimos anos. As suas principais características são o aumento da profundidade por meio da adição de mais camadas de convoluções para a extração de um maior número de características e o uso de filtros menores para a redução do custo computacional (Hossain et al. (2019)).

É importante mencionar que, embora as arquiteturas tenham evoluído bastante nos últimos tempos, a ausência da fácil disponibilidade de aceleradores de *hardware*, como Unidades de Processamento Gráfico (do inglês, *Graphics Processing Units* - GPUs), que suportem o treinamento de CNNs profundas em conjuntos de dados muito grandes, tornou-se um empecilho. Diante dessa limitação, é possível o reaproveitamento dos conjuntos de pesos de redes pré-treinadas que alcançaram um alto desempenho, isto é, pesos de redes que foram previamente treinadas para um problema relacionado à tarefa a ser resolvida, por meio de técnicas de transferência de aprendizagem, *fine-tuning*, treinamento continuado, etc. (Goodfellow et al. (2016)), para mitigar essa situação.

2.2 Redes Neurais Recorrentes

Diferente das CNNs, que são incapazes de memorizar saídas anteriores, transmitindo apenas a informação da camada de entrada em direção à saída, as Redes Neurais Recorrentes (do inglês, *Recurrent Neural Networks* - RNNs)

possibilitam, por meio do uso da relação de recorrência entre o estado oculto atual (ou variável latente - entrada presente) e o estado oculto anterior (o que aprendeu com as entradas que recebeu anteriormente - memória), a tomada de decisão simulando um comportamento semelhante ao funcionamento do cérebro humano (Goodfellow et al. (2016)). Basicamente, as RNNs são chamadas de recorrentes porque executam a mesma tarefa para todos os elementos de uma sequência, com a saída sendo dependente dos cálculos anteriores realizados.

Resumidamente, as RNNs podem ser representadas por meio de ciclos que permitem que as informações sejam armazenadas e passadas de uma etapa da rede para a próxima ao realimentar os neurônios de uma camada com as saídas anteriores (Figura 2.2 - esquerda). Ao desenrolar os ciclos, forma-se um grafo direcionado que representa as camadas individuais da rede neural ao longo da modelagem de uma sequência temporal (Figura 2.2 - direita). Graças à sua capacidade de memorização recente, essa técnica tornou-se ideal para tarefas de processamento de dados onde as saídas anteriores devem ser consideradas para realizar a próxima predição, como, por exemplo, em previsão de bolsas de valores, tradução de linguagem, geração de música, etc. (Vinyals et al. (2015); Hossain et al. (2019)).

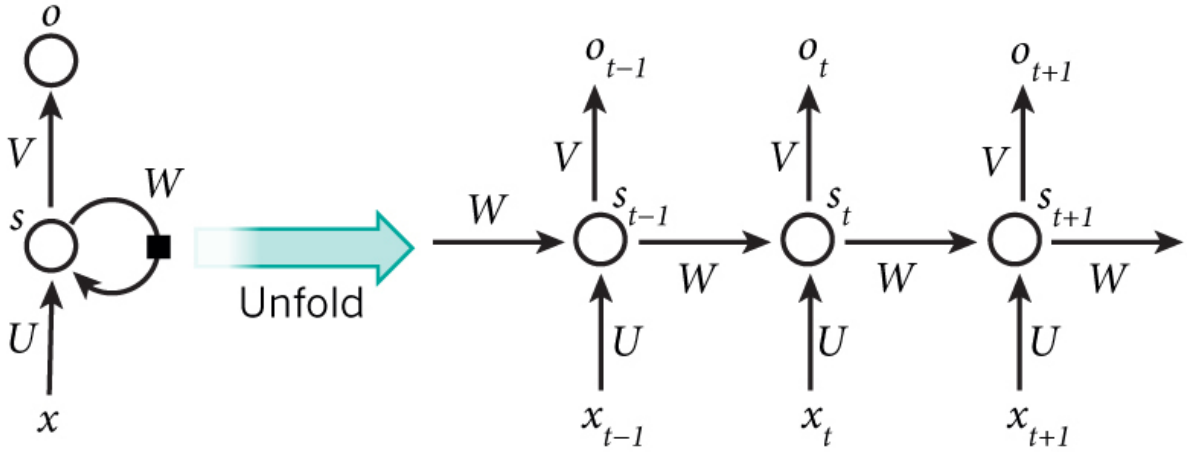


Figura 2.2: Representação compacta e desenrolada da estrutura de uma RNN. A representação desenrolada pode ser considerada como várias cópias da mesma rede, cada uma passando uma mensagem a uma sucessora. Fonte: LeCun et al. (2015).

Outra característica distintiva das RNNs é que elas compartilham o mesmo conjunto de pesos em cada camada da rede. Esses pesos são ajustados por meio do processo de retropropagação através do tempo e de algoritmos de otimização, como o gradiente descendente, SGD, Adam, entre outros, para facilitar o aprendizado. Por outro lado, é importante mencionar que as RNNs enfrentam dificuldades em aprender a usar as informações anteriores durante o processo de treinamento, especialmente quando lidam com sequências de palavras longas. Quando os valores de entrada, representados por um vetor, compõem uma sequência longa de palavras,

o estado oculto atual será multiplicado várias vezes pelo conjunto de pesos, o que pode resultar em dois problemas de dependências de longo prazo: a explosão e o desaparecimento dos gradientes (Pascanu et al. (2013); Goodfellow et al. (2016)).

Quando o gradiente, isto é, o valor usado para atualizar os pesos das redes neurais durante o treinamento, cresce muito devido à multiplicação por valores de pesos elevados, o processo de otimização torna-se instável e, consequentemente, a rede não aprende suficientemente com os dados de treinamento, deixando de fora informações importantes vistas no início da sequência de entrada. Contudo, ao limitar o valor máximo do gradiente, esse fenômeno pode ser controlado na prática (Goodfellow et al. (2016)).

Se o gradiente é reduzido a um valor muito pequeno, devido à multiplicação por valores de pesos pequenos, a atualização dos pesos será muito pequena e tenderá a continuar a se tornar menor ao longo do tempo, até que os pesos se tornem insignificantes. Logo, o resultado desse processo será uma aprendizagem lenta e focada nas informações mais próximas ao final da sequência, em vez do início. Esse problema geralmente acontece quando são utilizadas funções de ativação específicas, como a ReLU ou a Tangente Hiperbólica, para processar sequências de dados muito longas, resultando em uma memória de curto prazo (Goodfellow et al. (2016)).

Neste contexto, em uma situação em que um modelo de linguagem tenta prever a próxima palavra com base nas palavras anteriores, caso a lacuna entre as informações relevantes e o ponto onde elas são necessárias se torne muito grande, as RNNs tornam-se incapazes de aprender mais contexto e de conectar as informações. Uma forma de amenizar esse problema de dependência de longo prazo é por meio do uso de estruturas chamadas de portões.

Os portões são uma forma opcional de permitir a passagem de informações, possibilitando que a própria rede aprenda a decidir o que deve ser lembrado e esquecido. Dessa forma, o valor do peso do início das sequências pode ser aumentado, fazendo com que os gradientes correspondentes também sejam maiores (Goodfellow et al. (2016)). Portanto, as RNNs dos tipos *Long Short-Term Memory* (LSTM) e *Gated Recurrent Unit* (GRU) foram concebidas para mitigar o problema da memória de curto prazo, principalmente o desaparecimento do gradiente. No presente referencial, será dada ênfase somente à rede LSTM.

As LSTMs foram amplamente utilizadas junto com as CNNs para descrever em linguagem natural o conteúdo de imagens e vídeos, graças ao seu bom desempenho em tarefas de tradução e geração de sentenças (Vinyals et al. (2015); Hossain et al. (2019)). A arquitetura de uma LSTM, conforme a Figura 2.3, é composta, além da célula de memória, por três estruturas internas, isto é, portões (um de entrada, um de saída e um de esquecimento) que controlam o fluxo de informações por meio da filtragem seletiva de qualquer informação irrelevante em uma sequência (Goodfellow

et al. (2016)). Assim, os portões transmitem somente informações importantes ao longo de uma sequência para prever a saída da rede. Os portões são compostos por uma função de ativação sigmoide e uma operação de multiplicação para obter valores entre zero e um e, assim, fazer as operações de esquecimento (multiplicação do valor da sigmoide por zero) e de lembrança (multiplicação por um).

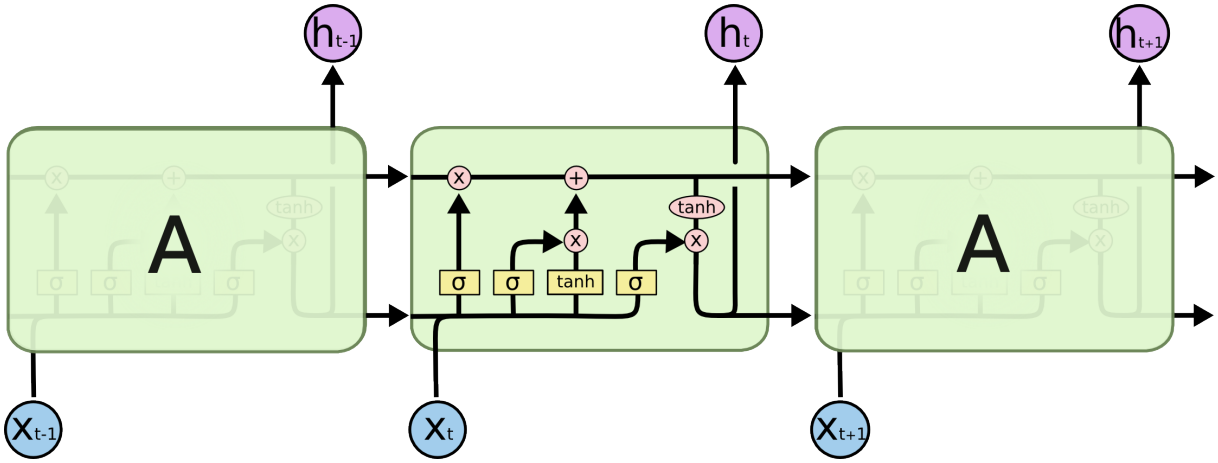


Figura 2.3: Estrutura de uma célula LSTM. Fonte: Olah (2015).

Mesmo que as LSTMs sejam capazes de manter a memória de longo prazo até determinado ponto, essas redes ainda enfrentam alguns desafios para a geração de sentenças longas e contextualmente bem formuladas (Jia et al. (2015); Hossain et al. (2019)). Neste cenário, uma vez que as informações da imagem são incluídas apenas no início do processo de uma LSTM e as próximas palavras são geradas com base no intervalo de tempo entre o estado oculto atual e o anterior, essa rede também pode enfrentar o problema do desaparecimento dos gradientes, ou seja, a sua capacidade de recordar os objetos que já foram descritos no início da sentença tende a ficar cada vez mais fraca ao longo do tempo.

Nesse contexto, Jia et al. (2015) propuseram uma extensão da LSTM denominada LSTM guiada (do inglês, *Guiding Long Short-Term Memory*, ou gLSTM). Nesta arquitetura, os autores adicionaram as informações semânticas extraídas da imagem como entrada extra de cada portão e do estado de célula da LSTM, além de considerarem diferentes estratégias de normalização de comprimento da pesquisa de feixe para controlar o tamanho das legendas. Dessa forma, foi possível orientar o modelo para gerar saídas mais fortemente aproximadas ao conteúdo da imagem, evitando frases não relacionadas e um viés para frases muito curtas.

É importante mencionar que as LSTMs unidirecionais são relativamente rasas em profundidade, portanto, não conseguem gerar sentenças contextualmente bem formuladas. Em modelos de linguagem unidirecionais, a palavra seguinte será prevista somente conforme os contextos textuais anteriores (Hossain et al. (2019)). Nessa conjuntura, Wang et al. (2016) desenvolveram um modelo ponta a ponta

baseado em LSTMs bidirecionais (do inglês, *Bidirectional Long Short-Term Memorys* - BiLSTMs) profundas, isto é, com duas LSTMs separadas, capaz de gerar sentenças contextual e semanticamente ricas para imagens (Figura2.4). A vantagem da bidirecionalidade desse modelo é conseguir adicionar mais informações de contexto, levando em consideração tanto palavras anteriores quanto posteriores, e aprender interações de linguagem a longo prazo.

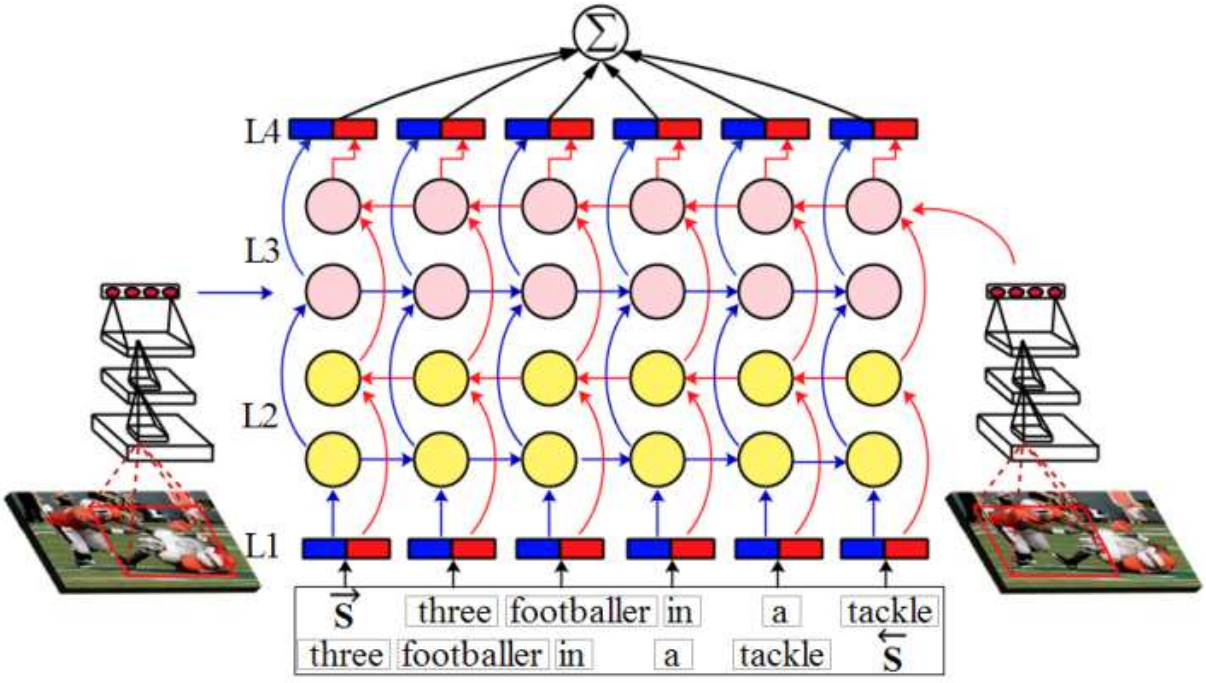


Figura 2.4: Visão geral do modelo de LSTM bidirecional. Os autores fornecem como entrada as sentenças tanto na ordem para frente (setas azuis) quanto para trás (setas vermelhas), o que permite que modelo ponta a ponta proposto resuma as informações de contexto do lado esquerdo e direito para gerar frases palavra por palavra ao longo do tempo. Fonte: Wang et al. (2016).

Além das LSTMs, como alternativa para modelagem de longas sequências, existem as Redes Neurais Recorrentes Hierárquicas (do inglês, *Hierarchical Recurrent Neural Networks* - HRNNs). As HRNNs são modelos de RNNs empilhados projetados com o objetivo de modelar, em diferentes escalas de tempo, estruturas hierárquicas distintas sobre dados sequenciais (Krause et al. (2017)).

As HRNNs são constituídas por várias camadas empilhadas, e cada camada possui uma ou mais RNNs. Assim, uma longa sequência de entrada é processada hierarquicamente até que a saída final seja gerada (Du et al. (2015); Zhao et al. (2017)). Motivadas pelo desejo de melhorar a capacidade de exploração da dependência temporal, as HRNNs, que foram inspiradas pela operação de convolução 1D (unidimensional), promovem a redução tanto da perda de informações na modelagem de sequências longas quanto da quantidade de

operações de computação em comparação com as RNNs tradicionais, devido às operações estarem diretamente associadas ao somatório do comprimento das camadas ao invés do comprimento da sequência de entrada (Krause et al. (2017); Zhao et al. (2017)).

2.3 Transformers

As RNNs foram firmemente estabelecidas como abordagens de última geração para problemas de modelagem e tradução automática de sequências. Desde então, diversos esforços têm sido feitos para expandir os limites dos modelos de linguagem recorrentes (Jia et al. (2015); Wang et al. (2016)). Todavia, a natureza inerentemente sequencial desses modelos impede a paralelização dos exemplos de treinamento, o que se torna crítico para sequências mais longas devido às restrições de memória e tempo computacional (Vaswani et al. (2017)). Além disso, mesmo que os mecanismos de atenção tenham se tornado parte integrante de várias tarefas em modelos de linguagem para lidar com dependências de longo prazo, em quase todos os casos, esses mecanismos são usados em conjunto com um modelo recorrente (Parikh et al. (2016)). Logo, a restrição fundamental da computação sequencial prevalece.

Inspirados nos sistemas biológicos dos seres humanos, que tendem a se concentrar nas partes distintas ao processar grandes quantidades de informações, os mecanismos de atenção têm sido utilizados em diversas aplicações no campo de Aprendizado Profundo (Niu et al. (2021)). Proposto inicialmente por Bahdanau et al. (2015), o mecanismo de atenção é uma extensão de melhoria da arquitetura codificadora-decodificadora utilizada na tarefa de tradução automática de sequências.

O mecanismo funciona como uma interface que conecta esses dois blocos da arquitetura, fornecendo ao decodificador, por meio da soma ponderada dos estados ocultos, um vetor de contexto com as informações e a importância relativa de todos os estados ocultos do codificador. Assim, sempre que o modelo com mecanismo de atenção for gerar a próxima palavra, ele buscará um conjunto de posições nos estados ocultos do codificador onde as informações mais relevantes estejam disponíveis e concentrar-se-á seletivamente apenas nessas informações para a realização da tarefa. Posteriormente, esse mecanismo de atenção e suas variantes foram utilizados em outras aplicações, incluindo legendagem de imagem, graças ao bom desempenho alcançado (Xu et al. (2015)).

Neste contexto, Vaswani et al. (2017) propuseram os Transformers, uma nova arquitetura de rede neural que evita o uso da recorrência e, em vez disso, passa a depender inteiramente de um mecanismo de atenção para modelar as dependências globais e as relações de longo alcance entre os elementos de uma sequência, sendo

este um grande avanço para o estado da arte no uso do mecanismo de atenção. Logo, por usar somente operações de soma ponderada e ativações, os Transformers possibilitaram a paralelização e, consequentemente, uma redução significativa do tempo de treinamento em comparação com as RNNs.

Os Transformers seguem a arquitetura codificadora-decodificadora apresentada na Figura 2.5. O codificador (lado esquerdo) é composto por um aglomerado de componentes empilhados, sendo eles um codificador posicional (*positional encoding*) seguido por N blocos idênticos ($N = 6$), contendo duas subcamadas cada: um mecanismo de auto-atenção (*self-attention*) com múltiplas cabeças e uma rede *feed-forward* totalmente conectada. Em torno de cada subcamada, há uma conexão residual que realiza a soma da saída de cada subcamada com sua entrada, seguida por uma camada de normalização. O conjunto de vetores codificados resultante é fornecido como uma das entradas do decodificador.

Semelhante ao codificador, o decodificador (lado direito) também compreende um codificador posicional e N blocos idênticos. Contudo, esta estrutura diverge do codificador nos seguintes pontos: possui como entrada tanto as saídas anteriores quanto a saída do codificador; cada bloco idêntico é constituído por três subcamadas ao invés de somente duas; e, por fim, a saída dos blocos é fornecida como entrada para uma camada linear seguida de uma função *softmax* (Vaswani et al. (2017)).

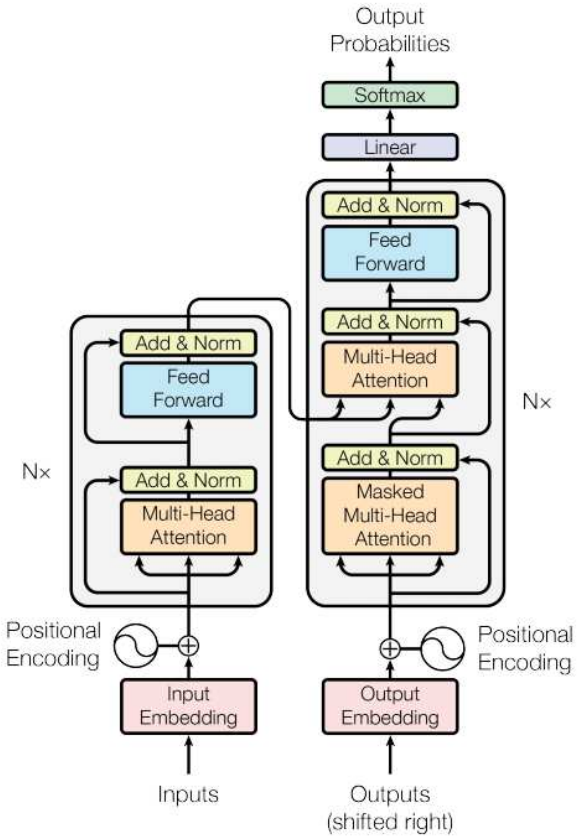


Figura 2.5: Arquitetura geral dos Transformers. Fonte: Vaswani et al. (2017).

Em suma, o codificador do Transformers começa recebendo uma sequência de entrada e gerando, para cada palavra da sequência, vetores de representação ou incorporação (*embeddings*). Codificadores posicionais são adicionados aos vetores para mapear a posição relativa de cada palavra na sequência codificada. Esses codificadores posicionais são vetores que podem ser aprendidos ou predefinidos, por exemplo, por funções de seno ou cosseno, e fornecem contexto de acordo com a posição da palavra em uma frase, uma vez que cada palavra em frases diferentes pode ter significados distintos (Khan et al. (2022)).

Em seguida, vem a essência principal do modelo Transformers, a camada de auto-atenção. Essa camada tem como finalidade permitir que as entradas interajam entre si para estimar a relevância de cada palavra em relação às outras palavras da mesma frase, por meio do cálculo do produto escalar dos vetores de entrada representados através dos elementos Chave, Consulta e Valor. Logo, para cada palavra, as pontuações de atenção retornadas são agregadas em um vetor de atenção, que captura a relação contextual entre as palavras daquela frase. Esta etapa é então repetida várias vezes em paralelo para todas as palavras, gerando sucessivamente novos vetores de atenção por palavra, com o intuito duplo de evitar que sejam produzidos os mesmos resultados e de fornecer um maior poder de discriminação à camada de auto-atenção, atendendo a diferentes segmentos das palavras. Por fim, para calcular o vetor de atenção final de cada palavra, é realizada uma média ponderada dos vários vetores, dividindo-os por um número fixo de cabeças (h) (Vaswani et al. (2017); Khan et al. (2022)).

Após o mecanismo de auto-atenção, uma rede *feed-forward* totalmente conectada simples, composta por duas transformações lineares com uma função de ativação ReLU, é aplicada a todos os vetores de atenção. Seu objetivo é transformar esses vetores, um de cada vez, em um formato que seja aceitável para o decodificador (Vaswani et al. (2017)). Graças à independência entre os vetores de atenção, a paralelização nessa etapa torna-se possível. Como resultado, é obtido um conjunto de vetores codificados para cada palavra simultaneamente.

O decodificador do Transformers, por ser um modelo auto-regressivo, opera nas previsões geradas anteriormente para produzir a próxima palavra da sequência, uma palavra por vez, da esquerda para a direita até o fim da sequência (Vaswani et al. (2017)). Dessa forma, o decodificador recebe tanto a sequência de saídas anteriores (inicialmente) quanto as representações finais produzidas pelo codificador (posteriormente, por meio de uma ramificação na arquitetura). A sequência de saídas anteriores é obtida por meio do deslocamento dessa sentença em uma posição à direita e anexando o *token* de início de sentença no início. Tal alteração tem como objetivo impedir que o modelo simplesmente aprenda a copiar a entrada do decodificador para a saída (Khan et al. (2022)).

De forma similar ao codificador, o decodificador recebe uma entrada que é convertida em vetores de representação seguidos pelos codificadores posicionais. Ao serem fornecidos aos N blocos, os vetores passam primeiramente pela camada de atenção de múltiplas cabeças mascaradas, onde os vetores de atenção são gerados para cada palavra nas frases para representar o quanto cada palavra está relacionada às demais palavras na mesma frase. A diferença desse módulo para a auto-atenção usada no codificador é que ele oculta a próxima palavra da sequência para que o próprio modelo consiga prever as palavras subsequentes com base nos resultados anteriores (Vaswani et al. (2017); Khan et al. (2022)). Logo, para que o aprendizado realmente aconteça, torna-se necessário mascarar a próxima palavra da sequência de entrada. Essa oclusão é realizada pela matriz de Valor, que transformará em zero as palavras posteriores para que a rede de atenção não possa usá-las (Vaswani et al. (2017)).

Em seguida, os vetores de atenção resultantes da camada anterior e os vetores fornecidos pelo codificador são mapeados e enviados como entrada para outra camada de atenção com múltiplas cabeças, seguidos por uma *feed-forward*. De forma similar ao ocorrido no codificador, nesse momento, é realizado o mapeamento entre as palavras e a descoberta da relação entre elas, a compactação dessas informações em novos vetores de atenção e a transformação desses vetores em representações que sejam aceitas por uma camada linear fora do bloco decodificador.

A camada linear é outra *feed-forward* que realiza a expansão das dimensões para a quantidade de palavras do vocabulário utilizado após o processo de geração. Por fim, a saída da camada linear é informada a uma função *softmax*, que transforma a entrada em uma distribuição de probabilidade, e a palavra resultante é retornada com base na maior probabilidade gerada (Vaswani et al. (2017); Khan et al. (2022)).

2.3.1 Vision Transformer

Com o crescente sucesso dos Transformers em PLN, Dosovitskiy et al. (2021) propuseram um modelo baseado na arquitetura Transformer adaptado para Visão Computacional, nomeado como *Vision Transformer* (ViT). Esse modelo é conhecido por aplicar diretamente a arquitetura Transformer em sequências de *patches* de imagem para realizar tarefas visuais, como a classificação. O ViT alcançou desempenho de ponta em vários *benchmarks* de reconhecimento de imagem. Além da tarefa de classificação, o ViT tem sido utilizado para abordar uma variedade de outras tarefas, incluindo detecção de objetos, segmentação semântica e compreensão de vídeo (Han et al. (2023)).

Conforme apresentado anteriormente, o Transformer recebe como entrada uma sequência de palavras (1D), que se transforma em uma representação. Em vez de

processar a imagem (2D) inteira de uma vez, Dosovitskiy et al. (2021) remodelaram a imagem de entrada, dividindo-a em uma sequência de pequenos blocos ou fragmentos chamados *patches*, onde cada *patch* é tratado como uma "palavra" em uma tarefa de PLN. Esses *patches* são então achatados e transformados em vetores de dimensão D fixa, usando uma projeção linear treinável. Como resultado dessa projeção, obtêm-se as representações dos *patches*. Em seguida, codificadores posicionais aprendidos são acrescentados a essas representações para reter informações sobre a posição espacial dos *patches* na imagem original.

A sequência de representações de *patches*, somada aos codificadores posicionais, é então alimentada em um codificador Transformer (Figura 2.6). Esse codificador aplica camadas de auto-atenção e redes *feed-forward* totalmente conectadas (neste caso, uma MLP) para processar as informações. O uso de auto-atenção permite que o modelo foque em diferentes partes da imagem ao mesmo tempo, capturando relações complexas entre diferentes *patches*. Além disso, camadas de normalização e conexões residuais são aplicadas, respectivamente, antes e depois de cada subcamada, sendo essa ordem uma diferença em relação ao codificador padrão do Transformer. Ressalta-se também que, conforme a Figura 2.6, o ViT não utiliza o bloco decodificador do Transformer, apenas o codificador. Por fim, após passar pelas camadas do codificador, a saída é processada por uma camada final de MLP para a tarefa específica, como a classificação de imagem.

Em comparação com as CNNs, o ViT oferece maior flexibilidade e desempenho como vantagens. Os codificadores posicionais, no momento da inicialização, não carregam nenhuma informação sobre as posições 2D dos *patches*, e todas as relações espaciais entre os *patches* precisam ser aprendidas do zero (Dosovitskiy et al. (2021)).

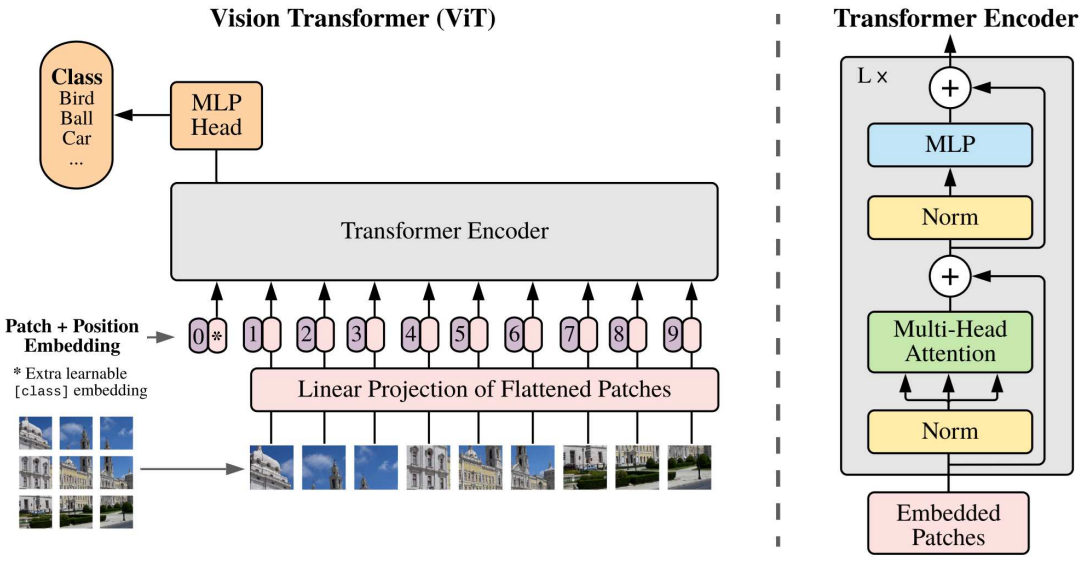


Figura 2.6: Arquitetura geral do ViT. Fonte: Dosovitskiy et al. (2021).

Por isso, o ViT apresenta um menor viés indutivo específico relacionado à estrutura da imagem, ao contrário das CNNs, que tendem a focar inicialmente em características locais da imagem, como linhas, brilho ou bordas. Da mesma forma, quando treinado em grandes volumes de dados, como JFT-300M (Sun et al. (2017)) e ImageNet-21k (Ridnik et al. (2021)), o ViT pode superar os modelos tradicionais de CNNs em tarefas de Visão Computacional (Han et al. (2023)).

Apesar das vantagens supracitadas, o ViT apresenta limitações em relação a sua aplicação. Por exemplo, o ViT não generaliza bem quando é treinado com quantidades insuficientes de dados, normalmente recomenda-se na a quantidade mínima de 14 milhões de imagens para iniciar um treinamento do zero (Han et al. (2023)). Ademais, a arquitetura Transformer necessita de mais recursos computacionais em comparação com as CNNs, principalmente quando se trata de processar grandes quantidades de dados (Dosovitskiy et al. (2021)). Em razão desses desafios, na maioria dos casos, o ViT é pré-treinado em grandes conjuntos de dados e, em seguida, ajustado para tarefas posteriores com menos dados.

2.4 Sumarização de Texto Abstrativa

A sumarização de texto abstrativa (do inglês, *Abstractive Text Summarization*) é a tarefa de reescrever um texto original de forma mais concisa, preservando sua essência semântica. Em contraste com a sumarização de texto extrativa (do inglês, *Extractive Text Summarization*), que simplesmente copia sentenças-chave do documento original, a sumarização abstrativa tem o potencial de produzir resumos que se assemelham mais ao estilo humano e são mais eficazes na compressão de informações (NithyaKalyani and Jothilakshmi (2019); Wang et al. (2019)).

Trabalhos sobre sumarização abstrativa, com foco em modelos neurais, têm sido aplicados em diversos domínios, especialmente em notícias (Chen and Zhuge (2018); Raffel et al. (2020); Zhu et al. (2021)). Por exemplo, Chen and Zhuge (2018) propuseram um modelo de sumarização abstrativa de texto-imagem para gerar resumos informativos de imagens. Nesse trabalho, os autores empregaram um modelo de codificador-decodificador hierárquico com atenção para resumir um documento de texto e suas imagens simultaneamente, alinhando frases e imagens nos resumos. Eles utilizaram uma HRNN bidirecional para codificar o texto e as legendas, e uma estrutura composta por uma CNN seguida de uma RNN para codificar o conjunto de imagens durante a etapa de codificação. Durante a decodificação, combinaram tanto a codificação do texto quanto a codificação das imagens como estado inicial, empregando a HRNN de atenção multimodal proposta por eles para gerar o resumo do texto. Na etapa de inferência, propuseram o uso da estratégia de busca de feixe multimodal, pontuando os feixes através das

sobreposições de bigramas das frases geradas.

Existem evidências em estudos que afirmam que os modelos Transformers pré-treinados com conjuntos de dados massivos alcançaram desempenhos satisfatórios em muitas tarefas de PLN, incluindo sumarização de texto abstrativa (Goodwin et al. (2020); Raffel et al. (2020); Zhu et al. (2021)). Estes estudos exploraram a capacidade de sumarização de Modelos de Linguagem Pré-treinados (do inglês, *Pre-trained Language Models* - PLMs), tais como BART (Lewis et al. (2020)), T5 (Raffel et al. (2020)) e PEGASUS (Zhang et al. (2020a)), tanto por meio do uso manipulado do *lead bias* para ensinar o modelo a discriminar e extrair informações importantes em geral, e não somente com as primeiras partes de um documento (Zhu et al. (2021)), quanto em configurações de aprendizagem *zero-shot* e *few-shot* para avaliar a qualidade de sumarização dos modelos treinados com poucos exemplos rotulados (Goodwin et al. (2020)).

Entre os modelos mencionados, o *Text-to-Text-Transfer-Transformer* (T5) é um modelo codificador-decodificador baseado na arquitetura Transformer e treinado no *Colossal Clean Crawled Corpus* (C4) em uma variedade de tarefas de aprendizagem supervisionada e não supervisionada, para as quais cada tarefa é convertida em um formato texto para texto (Raffel et al. (2020); Wolf et al. (2020)). Esse modelo utiliza uma arquitetura semelhante à do Transformer original (Vaswani et al. (2017)), com algumas adaptações, como o uso de codificação posicional relativa em vez da codificação baseada em seno e o posicionamento do viés da camada de normalização fora da conexão residual (Raffel et al. (2020)).

Para determinar a tarefa que o T5 deve realizar, um prefixo específico da tarefa é adicionado antes de alimentá-lo com o texto de entrada. Após a definição da tarefa, o T5 fornece ao decodificador o texto de entrada codificado por meio de suas camadas de atenção. Por fim, o decodificador gera auto-regressivamente uma sequência de saída usando a heurística de pesquisa de feixe. Essa heurística é responsável por gerar a frase-alvo palavra por palavra, da esquerda para a direita, enquanto expande uma largura fixa de feixes (w) em cada etapa de tempo, mantendo as w palavras mais prováveis, através da sequência de probabilidades condicionais de um modelo de geração de linguagem auto-regressivo treinado, neste caso, o T5 (Freitag and Al-Onaizan (2017)).

Capítulo 3

Trabalhos Relacionados

Neste capítulo, é fornecida uma visão geral dos principais trabalhos relacionados a esta tese, organizados por sua proximidade, o que possibilita uma melhor contextualização do tema a ser abordado.

3.1 Legendagem de Imagem

Legendagem de imagem (do inglês, *Image Captioning*) é a tarefa que descreve automaticamente o conteúdo observado de uma imagem em linguagem natural. Esta tarefa tem gradualmente atraído o interesse dos pesquisadores na comunidade científica, tornando-se um campo interdisciplinar importante que conecta as áreas de Visão Computacional e PLN (Hossain et al. (2019); Wang et al. (2020)).

Na última década, vários *frameworks* promissores foram propostos por pesquisadores para a tarefa de legendagem de imagens com base nas arquiteturas codificadora-decodificadora profundas substituindo até então os modelos probabilísticos (Huang et al. (2019); Ng et al. (2021); Sharif et al. (2021)). Nessas arquiteturas, conforme a Figura 3.1, uma imagem de entrada tem suas características globais extraídas por uma CNN, que as codifica em uma representação vetorial de comprimento fixo. Essa representação é, então, fornecida como entrada para uma pilha de LSTMs, que decodificam esse vetor em uma sequência de palavras (Vinyals et al. (2015)).

Embora o modelo proposto por Vinyals et al. (2015) tenha demonstrado robustez em termos de resultados qualitativos e quantitativos, sua principal limitação é a incapacidade de criar legendas detalhando os objetos presentes na imagem. Posteriormente, mecanismos de atenção e suas variações foram implementados na arquitetura codificadora-decodificadora para mitigar essa limitação.

Esses mecanismos têm desempenhado um papel crucial na incorporação do contexto visual em modelos de legendagem de imagens, uma vez que se concentram seletivamente em partes específicas da imagem para melhorar a geração de sequências de texto descritivo (Xu et al. (2015)). Nesse sentido, a atenção adaptativa e os sentinelas visuais foram introduzidos para determinar quando ativar a atenção

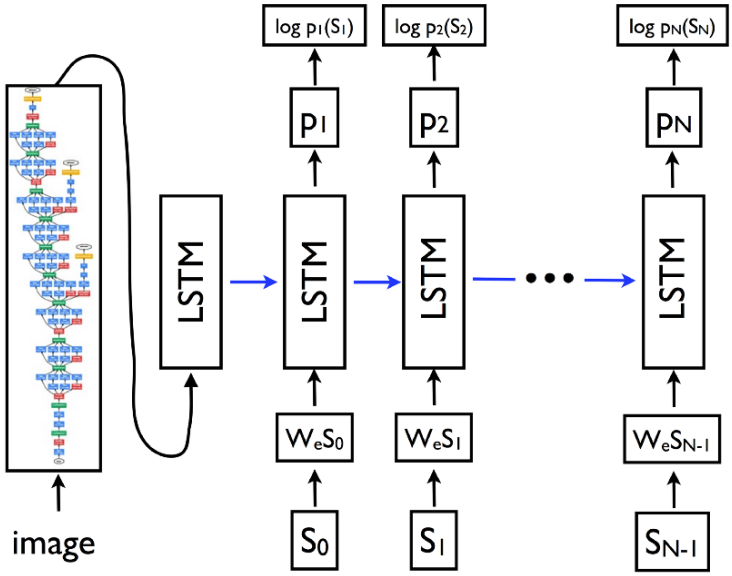


Figura 3.1: Modelo codificador-decodificador para geração de legendas de imagens. Fonte: Vinyals et al. (2015).

visual, melhorando ainda mais o desempenho do modelo (Lu et al. (2017)). Além disso, Huang et al. (2019) abordaram a questão da atenção irrelevante tanto para o codificador quanto para o decodificador por meio da extensão do mecanismo de atenção convencional com a adição de outra camada de atenção.

Recentemente, um novo *framework* para gerar histórias coerentes a partir de uma sequência de imagens de entrada foi proposto por Malakan. et al. (2021). Os autores primeiro modularam os vetores de contexto capturando a relação temporal entre a sequência de imagens de entrada em duas direções, usando BiLSTMs. Em seguida, projetaram as características da imagem e os vetores de contexto em um espaço latente conjunto, que foi submetido a um mecanismo de atenção que aprende as relações espaço-temporais entre as imagens. A saída do mecanismo de atenção é então modulada para capturar a interação entre as entradas e seu contexto e, finalmente, fornecida ao modelo de linguagem baseado na arquitetura *Mogrifier*-LSTM para gerar a sequência de sentenças que compõem a história.

Avanços recentes incorporaram arquiteturas baseadas em Transformers, que atualmente representam as técnicas do estado-da-arte para a legendagem de imagens (Rotstein et al. (2024)). O modelo Transformer, conforme explicado no Capítulo 2, consiste em um codificador e um decodificador. O codificador é composto por uma pilha de camadas de auto-atenção e *feed-forward*, que processam a sequência de entrada e geram uma representação matricial dessa entrada. O decodificador, por sua vez, utiliza mecanismos de auto-atenção para focar em palavras dentro da representação codificada e atenção cruzada para considerar a saída da última camada do codificador ao gerar iterativamente a sequência de saída

(Cornia et al. (2020)).

Nesse contexto, Herdade et al. (2019) propuseram uma arquitetura baseada em Transformers com um bloco de codificação, que incorpora informações sobre as relações espaciais entre objetos de entrada detectados através de atenção geométrica. He et al. (2020) também propuseram uma mudança na arquitetura interna do Transformer original para adaptá-lo à estrutura da imagem, explorando as relações espaciais hierárquicas entre as regiões da imagem.

Ao contrário desses dois estudos mencionados que utilizaram uma CNN e um Transformer como *framework* para legendagem de imagens, Liu et al. (2021) também abordaram o problema da descrição de imagens usando Transformers, substituindo o codificador baseado em CNN por um codificador Vision Transformer (ViT) (Dosovitskiy et al. (2021); Han et al. (2023)). Os Transformers utilizam mecanismos de atenção que consideram todas as partes da imagem simultaneamente, o que lhes permite capturar relações de longo alcance entre pixels. Essa capacidade proporciona uma modelagem mais eficaz dos contextos globais na imagem. Consequentemente, a substituição das CNNs por Transformers no estudo de Liu et al. (2021) demonstrou a eficácia do método proposto, superando os resultados obtidos com arquiteturas profundas baseadas em CNNs seguidas por um Transformer.

A integração do Pré-treinamento de Visão-Linguagem (do inglês, *Vision-Language Pre-training* - VLP) melhorou significativamente a compreensão do conteúdo visual e, consequentemente, o desempenho de inúmeras tarefas de linguagem visual (Huang et al. (2024); Rotstein et al. (2024)), incluindo legendagem de imagens e Resposta a Perguntas Visuais (do inglês, *Visual Question Answering* - VQA). O VLP (por exemplo, o modelo CLIP (Radford et al. (2021))) aproveita conjuntos de dados contendo pares de imagem-texto coletados da web em larga escala para pré-treinar modelos de visão e linguagem dentro de um espaço de incorporação compartilhado (Li et al. (2022); Xie et al. (2022)). Subsequentemente, esses modelos podem ser ajustados para tarefas específicas. Essa abordagem, que enfatiza a fusão de tarefas de geração e compreensão de visão e linguagem, levou ao desenvolvimento de modelos multimodais como BLIP (Li et al. (2022)) e OFA (Wang et al. (2022)).

Diferentes abordagens foram propostas para melhorar a capacidade discriminativa das legendas geradas e atenuar as limitações apresentadas por métodos previamente propostos (Li et al. (2019); Cornia et al. (2020); Sharif et al. (2021)). Recentemente, Gondim et al. (2022), dos Santos et al. (2022) e de Alencar et al. (2024) expandiram as técnicas de legendagem de imagens através do estudo, desenvolvimento e aplicação de modelos voltados para o idioma português brasileiro.

Apesar dos esforços, principalmente para enriquecer a representação visual e gerar descrições mais naturais, a maioria dos modelos ainda produz legendas genéricas e semelhantes, consistindo em uma única sentença (Ng et al. (2021);

Delloul and Larabi (2022); Cao et al. (2024)), tornando-as muitas vezes inadequadas para deficientes visuais (Dognin et al. (2022)). Dadas as limitações da abordagem clássica, novas tarefas derivadas da legendagem de imagens foram propostas, como a legendagem densa e parágrafos de imagens.

Acredita-se que uma única sentença pode ser insuficiente para descrever uma imagem para pessoas com deficiência visual, pois pode não conter todas as informações relevantes presentes na cena. Portanto, as abordagens propostas nesta tese diferem das técnicas tradicionais de legendagem de imagens, pois geram descrições mais detalhadas e informativas, atendendo às necessidades específicas desse público.

É importante mencionar que não foram utilizadas técnicas de legendagem de vídeo (do inglês, *Video Captioning*) nesta tese, todos os métodos estudados e aplicados foram limitados à imagem. O principal motivo dessa escolha é o objetivo deste trabalho, que foca em descrever o conteúdo visual presente na cena de webinários. A legendagem de vídeo, por sua vez, se concentra em gerar uma sentença para descrever o movimento ou a ação, ou seja, o evento do vídeo, a partir do seu entendimento (Chen et al. (2019); Perez-Martin et al. (2022)). Logo, sua aplicação não se enquadra no escopo desta tese.

3.2 Legendagem Densa

Como as descrições baseadas em regiões tendem a ser mais detalhadas do que uma única descrição global, a tarefa de legendagem densa (do inglês, *Dense Captioning*) visa generalizar as tarefas de Detecção de Objetos e Legendagem de Imagens em uma tarefa conjunta, localizando e descrevendo simultaneamente as regiões salientes de uma imagem (Johnson et al. (2016)). Para lidar com essa tarefa, Johnson et al. (2016) propuseram uma arquitetura de Rede de Localização Totalmente Convolucional (do inglês, *Fully Convolutional Localization Network* - FCLN) que suporta treinamento de ponta a ponta e possui desempenho eficiente em tempo de teste.

A arquitetura, conforme a Figura 3.2, é composta por uma CNN para processar a imagem de entrada e uma camada de localização densa para prever um conjunto de regiões de interesse (do inglês, *Regions of Interest* - ROIs) na imagem. A camada de localização densa utiliza um mecanismo de atenção diferencial espacial e interpolação bilinear em vez do mecanismo de *pooling*. Essa modificação auxilia o método na retropropagação pela rede e na seleção suave das regiões ativas de uma imagem. Por fim, as descrições para cada região são criadas por uma LSTM com linguagem natural previamente processada por uma rede totalmente conectada. Essa arquitetura resultou no modelo DenseCap (Johnson et al. (2016)).

Abordagens adicionais na área da legendagem densa têm explorado estratégias,

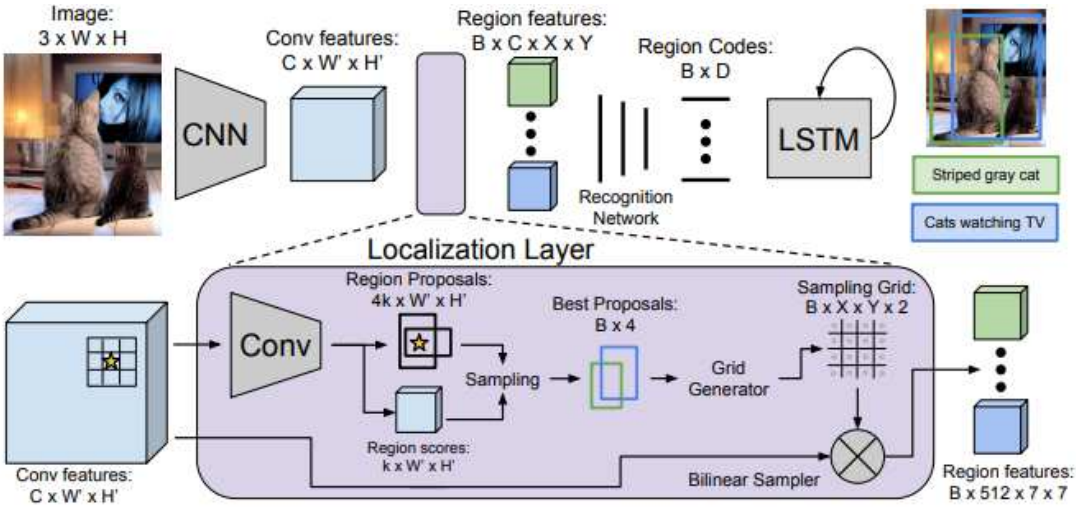


Figura 3.2: Visão geral da arquitetura da FCLN. Fonte: Johnson et al. (2016).

como a inferência conjunta e a fusão de contexto (Yang et al. (2017)), e as informações visuais sobre a região-alvo, bem como as dicas contextuais em múltiplas escalas (Yin et al. (2019)), para aprimorar a qualidade descritiva das legendas densas geradas. Além disso, a extensão da tarefa de legendagem densa para eventos de vídeo foi possível incorporando o contexto temporal, conforme demonstrado no estudo de Krishna et al. (2017a).

Comumente os métodos para legendagem densa utilizam um detector baseado em CNN para extrair características de região e fornecer informações em nível de objeto essenciais para descrever o conteúdo da imagem. No entanto, esses métodos enfrentam desafios, incluindo a falta de informações contextuais, o risco de detecção imprecisa e altos custos computacionais (Nguyen et al. (2022)). Em resposta, Nguyen et al. (2022) propuseram o *Grid- and Region-based Image captioning Transformer* (GRIT), uma arquitetura baseada em Transformers que visa resolver essas questões.

Essa recente arquitetura extrai tanto características baseadas em grade quanto características baseadas em região das imagens de entrada para gerar legendas densas com informações contextuais mais ricas. Isso é alcançado substituindo o detector de objetos baseado em CNN por outro baseado em Transformers, resultando em um desempenho computacional aprimorado.

Os métodos de legendagem densa geram múltiplas legendas para diferentes regiões de uma imagem, enfatizando a capacidade discriminatória dessas legendas. Em contraste, as abordagens propostas nesta tese buscam criar legendas que integrem informações específicas e essenciais da imagem, de acordo com as diretrizes e estudos levantados (ABNT; Lewis; Stangl et al.), atendendo de forma mais precisa às necessidades de usuários com deficiência visual.

3.3 Parágrafos de Imagem

Parágrafos de imagem (do inglês, *Image Paragraphs*), também conhecido como legendagem de parágrafos, é uma tarefa de descrição de imagem que combina os pontos fortes da legendagem de imagens e da legendagem densa, possibilitando a geração de parágrafos longos, estruturados e coerentes que descrevem detalhadamente as imagens (Krause et al. (2017)). Essa abordagem foi motivada pela falta de detalhes descritos na única frase de alto nível elaborada pelas técnicas de legendagem de imagens e pela ausência de coesão ao descrever a imagem inteira por meio da grande quantidade de legendas curtas e independentes produzidas pelas técnicas de legendagem densa.

O estudo pioneiro sobre esta tarefa propôs, conforme a Figura 3.3, um modelo que decompõe uma imagem em várias regiões através de um detector de ROIs, seguido pela agregação desse conteúdo extraído em uma representação semanticamente rica por meio do *pooling* de região (Krause et al. (2017)). Esse vetor é fornecido como entrada para uma HRNN composta por dois módulos (sentença RNN e palavra RNN, sendo cada RNN uma LSTM) para decompor o parágrafo da imagem em suas frases correspondentes.

Basicamente, a sentença RNN recebe o vetor de características, determina quantas frases serão geradas no parágrafo resultante e produz um vetor de tópico de entrada para cada frase. Dado esse vetor, a palavra RNN irá gerar as palavras constituintes de uma única frase. Esse processo é repetido até que todas as palavras de todas as frases sejam geradas e estruturadas em um parágrafo.

Essa abordagem levou ao desenvolvimento do modelo conhecido como Regions-Hierarchical. Posteriormente, este modelo foi ampliado com um mecanismo de atenção e uma Rede Adversária Generativa (do inglês, *Generative Adversarial Network* - GAN) que possibilitaram a coerência entre frases sucessivas, concentrando-se nas regiões salientes dinâmicas durante a geração de parágrafos (Liang et al. (2017)).

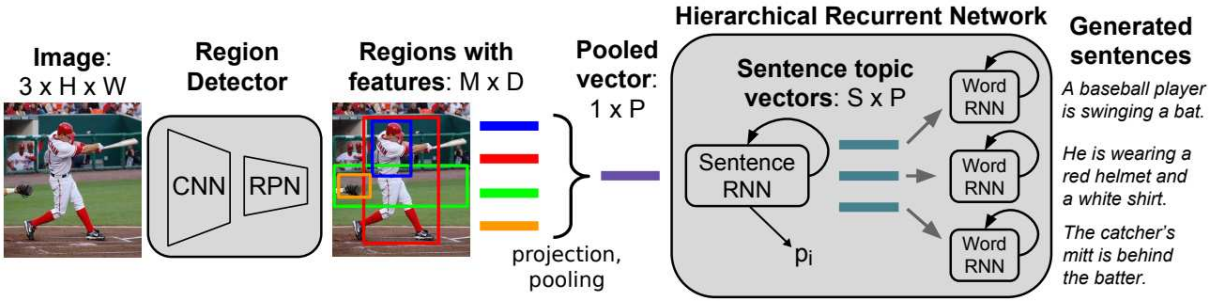


Figura 3.3: Visão geral do modelo gerador de parágrafos de imagens. Fonte: Krause et al. (2017).

Um estudo adicional relacionado à geração de parágrafos de imagem utilizou vetores de coerência e um Codificador Automático Variacional (do inglês, *Variational Auto-Encoder* - VAE) para assegurar a coerência dos temas entre as frases e aumentar a diversidade durante a geração de parágrafos (Chatterjee and Schwing (2018)). Além disso, a Rede de Política Visual Sensível ao Contexto (do inglês, *Context-Aware Visual Policy* - CAVP), proposta por Zha et al. (2022), também possibilitou a geração de uma descrição refinada da imagem em formato de parágrafo. Esse modelo não se concentra apenas na política de linguagem, mas também explora o contexto visual para atender ao raciocínio composicional ao longo do tempo. Dessa forma, ao invés de fixar apenas uma única característica visual em cada etapa, como ocorre na atenção visual convencional, a CAVP permite que as características visuais anteriores sirvam como contexto visual para a ação atual em um processo de previsão de sequência.

Nos últimos anos, estudos têm introduzido abordagens importantes: um Codificador-Decodificador de Grafo de Cena Hierárquico (do inglês, *Hierarchical Scene Graph Encoder-Decoder*) (Yang et al. (2020)), um modelo inovador projetado para capturar relações espaciais e semânticas entre objetos (Liu et al. (2022)), e a utilização de pistas visuais para conectar grandes modelos tanto de visão quanto de linguagem pré-treinados (Xie et al. (2022)). Esses métodos visam extrair informações valiosas das imagens incorporando conhecimento semântico e espacial rico em parágrafos, resultando em descrições mais coerentes, distintas e alinhadas com a imagem.

No contexto de aplicação para indivíduos com deficiência visual, Delloul and Larabi (2022) mencionam a importância de incorporar às técnicas de parágrafo de imagem a capacidade de gerar descrições egocêntricas de cena de objetos e arredores. Isso é para ajudar os indivíduos cegos ou com baixa visão a entender seu ambiente do seu ponto de vista, já que as técnicas clássicas de parágrafo de imagem tendem a descrever cenas globalmente, focando principalmente em objetos salientes.

As abordagens propostas nesta tese também divergem da tarefa parágrafo de imagem clássica. Em vez de descrever livremente toda a imagem, seguiu-se padrões e recomendações de acessibilidade (ABNT (2016); Lewis (2018); Stangl et al. (2020)). Portanto, concentrou-se em descrever as características relevantes do apresentador do webinário e seu entorno para proporcionar uma melhor compreensão às pessoas com deficiência visual. Além disso, a abordagem computacional final fornece descrições de cena, seus objetos e suas relações espaciais, abordando as lacunas identificadas por Delloul and Larabi (2022) em seu trabalho.

3.4 Grandes Modelos de Linguagem

Grandes Modelos de Linguagem (do inglês, *Large Language Models* - LLMs) é o termo usado tanto pela academia quanto pela indústria para se referir aos PLMs de tamanhos significativos (Zhao et al. (2023)). Esses modelos são frequentemente acessados através de uma interface de *prompt*, contendo dezenas ou centenas de bilhões de parâmetros e treinados com conteúdo diversificado proveniente da web (Thirunavukarasu et al. (2023)).

Atualmente, os LLMs são baseados principalmente na arquitetura Transformer, o que lhes permite capturar relações associativas entre palavras no conjunto de dados de treinamento. Através desse processo, os LLMs aprendem o uso contextual das palavras e podem aproveitar esses padrões aprendidos para realizar várias tarefas de PLN (Rotstein et al. (2024)). No entanto, os LLMs aumentam significativamente o tamanho do modelo, a quantidade de dados e os requisitos totais de computação (Zhao et al. (2023)).

Os LLMs, como LLaMA (Touvron et al. (2023)) e Vicuna (Chiang et al. (2023)), estão atraindo cada vez mais a atenção da comunidade de pesquisa devido às suas habilidades emergentes e ao bom desempenho em tarefas complexas de *zero-shot* e *few-shot*. Essas tarefas incluem tradução, sumarização de texto, resposta a perguntas, além de geração e compreensão de linguagem natural (Thirunavukarasu et al. (2023); Zhao et al. (2023); Rotstein et al. (2024)). Isso sugere que esses modelos conseguem compreender e gerar texto com pouca ou nenhuma adaptação adicional. Como resultado, as capacidades demonstradas pelos LLMs superam o desempenho de PLMs menores, permitindo seu uso tanto como assistentes de propósito geral quanto como ferramentas de geração de dados para pesquisadores (Liu et al. (2023); Rotstein et al. (2024)).

Para permitir que os LLMs executem diversas tarefas do mundo real, os pesquisadores têm explorado métodos de ajuste por instruções, visando aprimorar a capacidade do modelo de seguir instruções arbitrárias (Dai et al. (2023); Liu et al. (2023)). Recentemente, os LLMs ajustados por instruções também foram aplicados a tarefas de linguagem visual, onde são treinados em pares de texto e imagem (Dai et al. (2023)). Isso permitiu que esses modelos incorporassem o conteúdo visual das imagens de entrada em descrições textuais, resultando na geração de textos mais longos e detalhados (Huang et al. (2024)).

Aproveitando um vasto conhecimento visual e linguístico externo, essa abordagem lançou a base para o desenvolvimento de LLMs multimodais como LLaVA (Liu et al. (2023)) e InstructBLIP (Dai et al. (2023)). Esses modelos facilitaram uma ampla gama de tarefas de linguagem visual, incluindo a legendagem de imagens e descrições visuais detalhadas, conforme demonstrado por Rotstein et al. (2024) em seu trabalho

utilizando o modelo FuseCap.

Apesar do progresso e do impacto, os LLMs são propensos ao fenômeno de alucinação de IA, resultando na geração de longos textos que incluem detalhes e afirmações falsas ou enganosas (Cao et al. (2024); Huang et al. (2024)). *Prompts* textuais incompletos ou mal estruturados e a ausência de dados específicos do domínio são exemplos de fatores que desencadeiam alucinações.

Embora os LLMs utilizem vastos dados de treinamento, há uma necessidade de refinamento contínuo para aumentar a confiabilidade dos assistentes baseados em LLMs projetados especificamente para cenários práticos (Salvagno et al. (2023); Khurana et al. (2024)), como para indivíduos com deficiência visual. Assim, ao contrário de modelos baseados em LLMs, que podem gerar conteúdo com base na interpretação do *prompt* e alucinar, a abordagem final proposta nesta tese utiliza um conjunto de perguntas predefinidas que são enviadas a um modelo de VQA e retornam especificamente a informação de interesse, reduzindo as chances de produzir descrições que não são visualmente compatíveis com as imagens.

3.5 Conjuntos de Dados Personalizados para Aplicações do Mundo Real

Além dos conjuntos de dados de uso geral e massivos propostos por grandes corporações ou consórcios de institutos de pesquisa de alto nível, conjuntos de dados específicos ainda são necessários para abordar aplicações reais locais ou regionais. Nesse contexto, visando aumentar a precisão dos modelos aplicados a carros brasileiros, Kuhn and Moreira (2021) propuseram o BRCars, um conjunto de dados composto por aproximadamente 2.000.000 imagens para auxiliar na tarefa de classificação no domínio dos carros brasileiros de um website de vendas de carros variando em mais de mil modelos.

Nascimento et al. (2022) propuseram um conjunto de dados com 1.000 imagens de mudas de soja com anotações sobre o genótipo e a condição do solo para auxiliar o desenvolvimento de modelos de IA para a agricultura, o que não era possível usando apenas os conjuntos de dados massivos de uso geral. Diferente dos exemplos anteriores voltados para a tarefa de classificação, dos Santos et al. (2022) apresentaram o #PraCegoVer, um conjunto de dados multimodal com imagens e descrições em português brasileiro composto por 533.523 postagens coletadas de 14.000 perfis diferentes no Instagram, destinado à tarefa de legendagem de imagem em língua portuguesa.

À medida que é testemunhado um crescimento exponencial tanto no número quanto na qualidade das aplicações de IA de uso geral, um progresso muito mais

lento é observado para tarefas específicas. Além disso, há poucos conjuntos de dados disponíveis publicamente que visam problemas de minorias, como pessoas portadoras de deficiência visual (Gurari et al. (2018); Tang et al. (2023)).

No contexto do reconhecimento de objetos, a tarefa de *Teachable Object Recognizer* (TOR) tende a melhorar a vida das pessoas com deficiência visual, aprendendo novas classes de objetos a partir de apenas algumas amostras. O conjunto de dados ORBIT com 4.733 vídeos de 588 objetos foi proposto para habilitar o TOR em cenários do mundo real, especificamente para pessoas com deficiência visual (Massiceti et al. (2021)). Recentemente, Tang et al. (2023) introduziram um conjunto de dados com 7.915 imagens de 15 tipos de obstáculos comuns no caminho de pessoas com baixa visão, visando modelos focados em possibilitar a travessia em ambientes externos.

Gurari et al. (2020) propuseram um conjunto de dados para tarefas de legendagem de imagens focando nos interesses das pessoas com deficiência visual. O conjunto de dados consiste em 39.000 fotos capturadas por pessoas cegas e descritas textualmente com cinco legendas. O mesmo grupo propôs um conjunto de dados com 5.537 fotos para impulsionar o desenvolvimento de algoritmos que abordem o problema de divulgações não intencionais de informações privadas (Gurari et al. (2019)).

A tarefa de VQA tem grande potencial para auxiliar indivíduos com CBV em suas atividades diárias. Ao combinar tarefas de visão e linguagem, o VQA possibilita a previsão de respostas a perguntas baseadas em imagens (Kim et al. (2021); Li et al. (2022)). Assim, os métodos derivados dessa tarefa tornam-se ideais para reduzir barreiras visuais e permitir que indivíduos com deficiência visual acessem informações de imagens capturadas, sejam elas oriundas da web ou do mundo real (Le et al. (2021)). No entanto, muitos métodos de VQA enfrentam dificuldades ao lidar com imagens de baixa qualidade. Com o objetivo de incentivar o surgimento de soluções para esse problema, foi proposto o conjunto de dados VizWiz-VQA com imagens capturadas por pessoas cegas em cenários reais, acompanhadas de uma pergunta relacionada e 10 respostas anotadas para cada imagem (Gurari et al. (2018)).

Shah et al. (2021) criaram um conjunto de dados contendo 2.500 imagens, fotografadas manualmente e coletadas da Internet, para treinar um modelo capaz de identificar cinco classes de objetos perigosos, com o objetivo de desenvolver um sistema de alerta para pessoas com deficiência visual. De maneira semelhante, nesta tese, uma das abordagens propostas envolveu a criação de um conjunto de dados específico, composto por imagens e descrições de cenas de webinários coletadas da Internet, voltado para pessoas com CBV. Esse conjunto de dados foi utilizado para o desenvolvimento e validação da abordagem computacional final proposta.

3.6 Métricas de Avaliação

Em legendagem de imagem, como em qualquer tarefa de inteligência computacional, as métricas de avaliação são cruciais para determinar o desempenho de um método ou modelo. Essas métricas devem pontuar a qualidade das legendas ou textos gerados, comparando a saída do sistema com os dados gerados por humanos (González-Chávez et al. (2024)). Atualmente, existe um conjunto de métricas de desempenho. A seguir, serão descritas brevemente as métricas utilizadas neste trabalho.

3.6.1 BLEU

BiLingual Evaluation Understudy (BLEU) é a métrica pioneira de avaliação automática de tradução de textos realizada por máquinas (Papineni et al. (2002)) e amplamente utilizada em várias tarefas de geração de textos, como a legendagem de imagens (Hossain et al. (2019); Wang et al. (2020)). Fundamentada na proximidade entre o texto traduzido automaticamente, isto é, o candidato gerado, e a tradução de referência informada, a BLEU avalia os textos traduzidos com base na precisão das sequências de n palavras (n-gramas). Essa precisão é informada por meio da divisão da quantidade de n-gramas ($n = 1, 2, 3, 4$) correlacionadas entre a tradução candidata e a tradução de referência pela quantidade total de palavras da tradução candidata (de Melo et al. (2015)).

Pelo fato de que a tradução candidata precisa ser similar à de referência, tanto na escolha quanto na ordem e na quantidade de palavras, a BLEU faz uso da penalização de brevidade no nível do corpus para penalizar traduções curtas geradas e evitar pontuações superestimadas. Por fim, a pontuação BLEU é calculada por meio da média geométrica das precisões dos n-gramas, multiplicando-a pela penalização de brevidade (Papineni et al. (2002); de Melo et al. (2015)).

Destaca-se ainda que, embora a métrica BLEU seja amplamente utilizada, ela é criticada por retornar pontuações altas apenas para sentenças curtas, especialmente nas variantes BLEU-1 e BLEU-2, e por não considerar as variações sintáticas das palavras.

3.6.2 METEOR

Metric for Evaluation of Translation with Explicit ORdering (METEOR), assim como a BLEU, é uma métrica usada para avaliar a qualidade do resultado da tarefa de tradução automática. Posteriormente, essas duas métricas foram adotadas para avaliação de legendas de imagens por serem apresentadas como medidas de

similaridade que comparam uma frase candidata a um conjunto de frases de referência com base na correspondência de *n*-gramas (Anderson et al. (2016)). Contudo, diferente da BLEU, a METEOR foi projetada com base na média harmônica entre a precisão e o *recall*, com mais ênfase no *recall* (Banerjee and Lavie (2005)).

No cálculo da sua pontuação, a METEOR gera um conjunto de mapeamentos alinhando ordenadamente os unigramas entre os textos candidatos e de referência com base nas correspondências exatas dos *tokens*, seguido pelos *synsets* da WordNet (Miller (1995)) e, finalmente, pelos *tokens* derivados por meio do lematizador de Porter (Banerjee and Lavie (2005)). Graças a esse critério, essa métrica tornou-se altamente correlacionada com o julgamento humano em nível da frase ou do segmento, levando em consideração sinônimos e radicais de uma frase, isto é, fatores da gramaticalidade. Além disso, a METEOR também utiliza uma função de penalidade de fragmentação de alinhamento que analisa a ordem incorreta das palavras de acordo com a ocorrência de mapeamentos adjacentes entre os textos comparados (Anderson et al. (2016); Wang and Gaizauskas (2016)).

3.6.3 CIDEr

Ao contrário das métricas BLEU e METEOR, que se destinam à tarefa de tradução automática, a *Consensus-based Image Description Evaluation* (CIDEr) foi projetada especificamente para tarefas de legendagem de imagens (Vedantam et al. (2015)). Essa métrica avalia a consistência da descrição textual de uma imagem baseada no consenso humano por meio da sobreposição ponderada da medida *Term Frequency-Inverse Document Frequency* (TF-IDF) (Aizawa (2003)) para cada *n*-grama.

O TF-IDF é uma medida estatística que quantifica tanto a frequência quanto a relevância de uma determinada palavra em um texto. Essa medida é frequentemente usada como um fator de ponderação na recuperação de informações e mineração de texto (Christian et al. (2016)).

Para calcular a pontuação, cada frase é representada por um conjunto de *n*-gramas, seguido pela verificação das coocorrências dos *n*-gramas no conjunto de referência e no texto candidato. Com base no TF-IDF, os *n*-gramas que são comuns em todas as descrições da imagem são menos ponderados. Finalmente, a similaridade do cosseno entre os *n*-gramas do texto candidato gerado e o conjunto de referências é calculada (Kilickaya et al. (2017)).

Ao medir a similaridade entre as descrições (candidata e de referência), a CIDEr assimila intrinsecamente as ideias de gramaticalidade, saliência, importância, precisão e *recall*. Isso a tornou mais fortemente correlacionada com o julgamento humano em comparação com as métricas BLEU e METEOR (Vedantam et al. (2015)).

3.6.4 BERTScore

A BERTScore é uma métrica de avaliação automática não supervisionada proposta para a geração de texto baseada em incorporações (*embeddings*) contextuais de modelos BERT pré-treinados (Zhang et al. (2020b)). Essa métrica calcula a correspondência de duas sentenças textuais (no formato de representações vetoriais) como a soma das similaridades dos cossenos, opcionalmente ponderada com pontuações de IDF, entre os *embeddings* de seus tokens (isto é, palavras segmentadas processadas) para criar uma matriz de pontuação. Para isso, a métrica utiliza a correspondência gananciosa para maximizar a pontuação de similaridade de correspondência, ou seja, entre palavras nas sentenças candidatas e de referência (Kane et al. (2020)). Ao final é retornado o valor de precisão, *recall* e pontuação F1 da BERTScore.

A BERTScore, por meio do uso de incorporações de tokens contextualizadas, consegue superar duas limitações comuns em métricas baseadas em n-gramas que utilizam a correspondência exata de textos (BLEU) ou heurísticas de correspondência (METEOR). Essas limitações são a incapacidade de lidar de maneira robusta com paráfrases e a dificuldade em capturar dependências distantes. As métricas BLEU e METEOR penalizam textos candidatos que são semanticamente semelhantes, mas diferem na forma superficial ou que apresentam mudanças de ordenação semanticamente críticas. Em contraste, os *embeddings* contextualizados são treinados para detectar paráfrases e capturar efetivamente dependências e ordenações distantes (Lee et al. (2020); Zhang et al. (2020b)).

Os autores validaram a BERTScore em tarefas de tradução automática de texto e legendagem de imagens, demonstrando, em seus extensos experimentos, que a métrica se correlaciona altamente com avaliações humanas, superando até mesmo a SPICE, uma métrica baseada em grafos de cena (Anderson et al. (2016)).

3.6.5 CLIPScore

A CLIPScore é uma métrica recente que se destaca por ser a única métrica, entre as utilizadas neste trabalho, que dispensa o uso de referência. Isso significa que essa métrica não requer textos de referência escritos por humanos para medir a compatibilidade semântica entre o texto candidato gerado por máquinas e a imagem associada.

Baseada no CLIP (Radford et al. (2021)), um modelo *cross-modal* treinado por meio de aprendizagem contrastiva em 400 milhões de pares imagem-legenda da Internet, a CLIPScore demonstrou desempenho superior em comparação com várias métricas de *benchmark* existentes, incluindo CIDEr, BERTScore, NUBIA (Kane et al. (2020)) e até mesmo ViLBERTScore (Lee et al. (2020)), que também incorpora recursos de *grounding*

imagem-texto, além das referências (Hessel et al. (2021)).

O CLIP utiliza dois codificadores, um para a imagem e outro para a descrição, que aprendem a transformar ambos em vetores semelhantes, enquanto simultaneamente tenta distinguir uma imagem das outras descrições. O objetivo era criar um modelo *zero-shot* capaz de classificar imagens em uma variedade de conjuntos de dados (González-Chávez et al. (2024)).

A CLIPScore avalia a semelhança entre uma imagem e uma legenda de maneira não supervisionada, calculando a similaridade do cosseno com *embeddings* derivados do CLIP. Sua característica distintiva é a capacidade de avaliar não apenas com base na referência à imagem, mas também em um contexto livre de referência. Quando as referências estão disponíveis, ela mede a similaridade da legenda prevista com elas e calcula a média harmônica para obter a RefCLIPScore.

Hessel et al. (2021) demonstraram que essa métrica se correlaciona mais fortemente com o julgamento humano do que outras métricas mencionadas. No entanto, recentemente, González-Chávez et al. (2024) apontaram que, embora a CLIPScore se destaque na identificação das palavras que descrevem os elementos ou objetos presentes na imagem, ela dá pouca atenção à avaliação da estrutura gramatical da frase candidata.

Conforme supracitado, as métricas de avaliação existentes geralmente focam em aspectos técnicos como precisão, cobertura e similaridade entre textos gerados e textos e/ou imagens de referência de forma genérica, sem considerar a qualidade textual e a adequação à necessidade de clareza para um público com deficiência visual. Nesta tese, foi proposta uma métrica de avaliação com enfoque especial na acessibilidade e na adequação das descrições para pessoas com deficiência visual, garantindo, por meio da análise de vários critérios, que as descrições sejam informativas e úteis para esse público específico.

Capítulo 4

Descrevendo imagens focadas em detalhes cognitivos e visuais para pessoas com deficiência visual: Uma abordagem para gerar parágrafos inclusivos

Alinhado com o primeiro Objetivo Específico (OE-1) desta tese, apresentado na Introdução (Capítulo 1), e com os trabalhos relacionados mencionados, este capítulo descreve a abordagem computacional proposta inicialmente para a geração de descrições textuais de contextos de imagens de webinários para pessoas com deficiência visual. A abordagem foca nos detalhes cognitivos e visuais necessários para esse público específico, seguindo as diretrizes e recomendações de acessibilidade extraídos de ABNT (2016) e Lewis (2018), que são frequentemente negligenciados na literatura existente.

Além da metodologia, este capítulo também detalha a avaliação experimental realizada, os resultados obtidos, as discussões que validaram a abordagem proposta e as limitações identificadas. Ao final, com o propósito de atingir o OE-2, foi investigado e relatado o impacto da adequação dos conjuntos de dados e das métricas de avaliação utilizados.

4.1 Materiais e Métodos

A abordagem proposta recebe como entrada uma imagem de webinário, caracterizada por um palestrante em primeiro plano que se expressa por meio de uma ferramenta de videoconferência, e, após uma sequência de passos, gera como saída um parágrafo que descreve o contexto da imagem para pessoas com baixa ou nenhuma visão. Conforme a Figura 4.1, a estrutura da abordagem consiste em duas fases principais: Analisador de Imagem e Geração de Contexto.

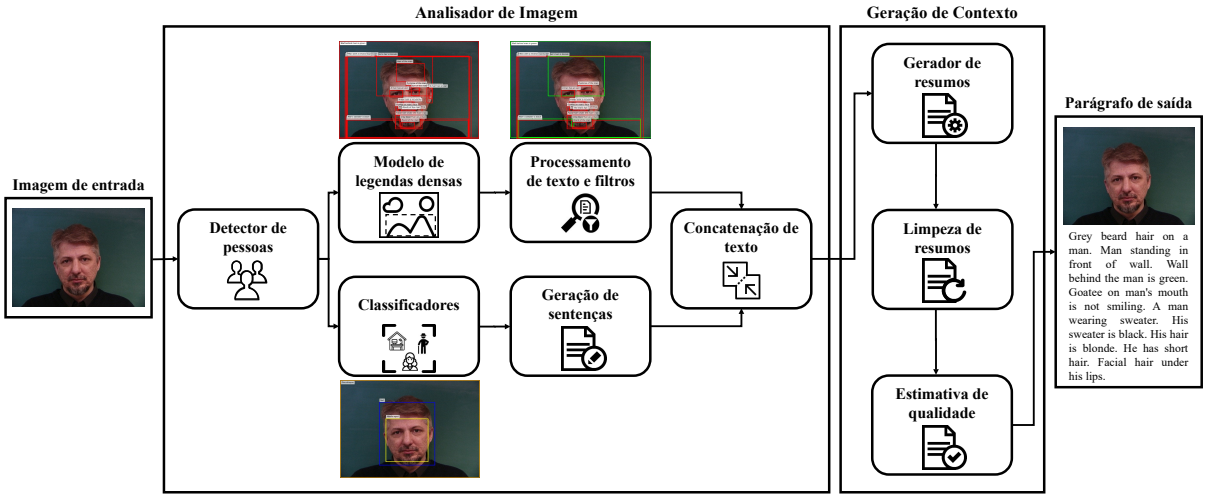


Figura 4.1: Visão geral da abordagem proposta. Fonte: Material próprio.

Na primeira fase, as descrições textuais (legendas densas) da imagem de entrada são geradas. Em seguida, essas descrições são filtradas e informações de alto nível da imagem (idade, emoção e cena) são extraídas. Ao final, todas essas informações até então obtidas são concatenadas em um texto.

Na segunda fase, um modelo de linguagem é alimentado com a saída textual da etapa anterior e cria-se um resumo coerente e conectado com base nos dados de entrada recebidos. Esse resumo passa por um processo de limpeza e filtragem para selecionar o texto de melhor qualidade gerado, que será usado como o contexto da imagem para auxiliar a pessoa com CBV.

As subseções a seguir descrevem as etapas de ambas as fases, cujos detalhes de implementação, como limiares, modelos e seus hiperparâmetros, podem ser conferidos posteriormente na Seção 4.2.2.

4.1.1 Analisador de Imagem

O objetivo desta fase de cinco etapas consiste na extração de informações relacionadas ao conteúdo visual relevante para pessoas com deficiência visual da imagem de entrada e, em seguida, realizar a codificação dessas informações em um formato de texto para continuidade na próxima fase.

Detector de pessoas

Como primeiro passo, aplica-se um detector de objetos, estendendo-se para pessoas, e o número de pessoas P na imagem é computado. Se $P = 0$, o processo é interrompido, uma vez que a abordagem é restrita apenas a imagens com pessoas.

Modelo de legendas densas

Como segunda etapa, caso $P > 0$, um modelo de legendagem densa é utilizado para produzir uma legenda em formato de sentença curta e independente para cada objeto ou ROI da imagem. Cada legenda possui um valor de confiança sobre a sentença gerada, suas propriedades descritivas (relacionamentos e características) e uma caixa delimitadora (do inglês, *bounding box* - bbox) que representa a localização espacial da região por meio de coordenadas. Ao final dessa etapa, um conjunto de legendas é retornado e encaminhado para a próxima etapa.

Processamento de texto e filtros

O modelo de legendagem densa produz um grande número de legendas por imagem, o que pode resultar em descrições semelhantes, fora do contexto de interesse e redundantes. Deste modo, esta etapa tem como finalidade a filtragem desses tipos de descrições, selecionando as legendas mais propícias para o público-alvo, conforme especificado pela a Associação Brasileira de Normas Técnicas (ABNT (2016)) e a *Perkins School for the Blind* (Lewis (2018)).

Inicialmente, todas as legendas são convertidas para minúsculas, pontuações e espaços em branco duplicados são removidos e um método de tokenização é aplicado nas sentenças para realizar a segmentação das palavras, isto é, a quebra da sequência de caracteres (Bird et al. (2009)). As palavras segmentadas são identificadas como *tokens*, que são as unidades básicas de texto processadas em PLN.

Em seguida, para reduzir a quantidade de legendas densas semelhantes, calcula-se a similaridade semântica entre elas. Para isso, uma representação vetorial de alta dimensionalidade é criada para cada legenda pré-processada usando a incorporação de palavras (*word embedding*). Em seguida, computa-se a similaridade do cosseno (Ravichandran et al. (2005); Faruqui et al. (2016)) sobre todas as representações das legendas densas. Essa métrica mede o cosseno do ângulo entre dois vetores n-dimensionais projetados em um espaço multidimensional (Han et al. (2012)).

Dado que as legendas densas são representadas por esses vetores de alta dimensionalidade, é possível mensurar o quão semelhantes essas descrições são quantificando o cosseno do ângulo entre seus vetores. A similaridade do cosseno captura a direção (o ângulo) entre as legendas no espaço, de modo que, quanto menor for o ângulo, maior será o valor do cosseno e, consequentemente, maior a similaridade entre os vetores comparados (Ravichandran et al. (2005); Han et al. (2012); Faruqui et al. (2016)).

As legendas que apresentam uma similaridade do cosseno maior ou igual ao limite definido, denominado T_{text_sim}, são descartadas. Desta forma, com exceção da

primeira legenda presente no conjunto de descrições, apenas são mantidas as próximas legendas que, ao serem comparadas com as anteriores já permanecidas no conjunto, retornem um valor de similaridade abaixo do limiar definido. Esse processo se repete até que todas as legendas densas do conjunto inicial de descrições sejam comparadas semanticamente, conforme demonstrado no Algoritmo 1.

Algoritmo 1: Redução da quantidade de legendas densas semelhantes

```
1:  function similarCaptions(tokenized_subset, T_text_sim)
2:      captions ← [ ]
3:      not_similar ← [ ]
4:      for each dense_caption in tokenized_subset do
5:          if not_similar = ∅ then
6:              captions_out_of_context_check(dense_caption, not_similar)
7:          else
8:              is_not_similar ← true
9:              for each caption_not_similar in not_similar do
10:                 cosine_s ← semantic_s(dense_caption, caption_not_similar)
11:                 if cosine_s ≥ T_text_sim then
12:                     is_not_similar ← false
13:                     break
14:                 end if
15:             end for
16:             if is_not_similar = true then
17:                 captions_out_of_context_check(dense_caption, not_similar)
18:             end if
19:         end if
20:     end for
21:     return captions
```

A próxima tarefa, conforme detalhada na Figura 4.2, é destinada à remoção das legendas densas que estão fora do contexto de interesse, ou seja, as legendas produzidas pelo modelo de legendagem densa para a imagem de entrada que não estão associadas ao palestrante e seu entorno. Para esse fim, os atributos de anotação linguística de substantivos, pronomes e sujeitos nominais são verificados para cada *token* das sentenças sequencialmente. Se nenhum *token* relacionado a um desses atributos for encontrado na legenda analisada, ela é descartada.

Em seguida, uma nova verificação é realizada nas legendas que foram mantidas após a aplicação desta segunda filtragem para garantir que as restrições baseadas nas características do palestrante e seu entorno sejam atendidas. Usando os conjuntos de sinônimos cognitivos da WordNet (Miller (1995)), é verificado se os *tokens* das legendas estão associados ao conceito semântico de pessoa (ser humano), isto é, se os *tokens* são hierarquicamente dependentes do ancestral comum de pessoa. Caso essa premissa seja verdadeira, o *token* analisado é considerado um hipônimo de

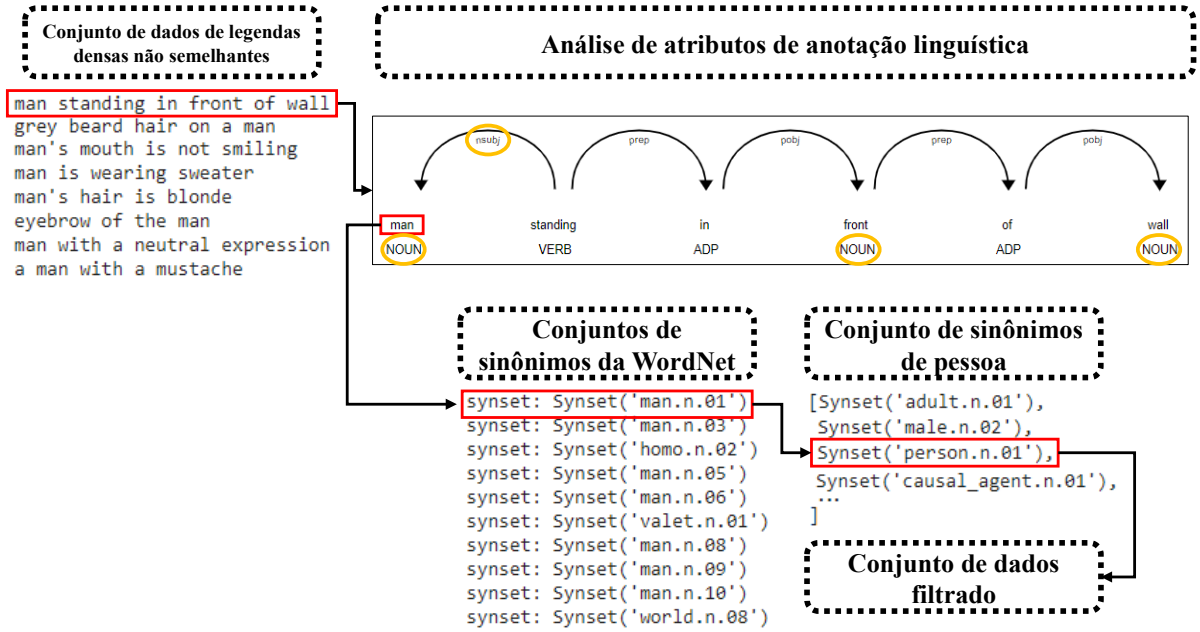


Figura 4.2: Fluxograma do processo de filtragem das legendas fora do contexto. Fonte: Material próprio.

ser humano (Krishna et al. (2017b)) e a sua legenda não é excluída do conjunto de descrições.

A WordNet é uma base de dados lexical da língua inglesa que agrupa palavras em conjuntos de sinônimos denominados como *synsets*. Cada *synset* representa um conceito específico e está interconectado a outros *synsets* através de relações semânticas, como hiperonímia (relações de "é um tipo de") e hiponímia (relações de "subtipo") (Miller (1995)). Na WordNet, conceitos são organizados em uma estrutura hierárquica onde conceitos mais específicos (hipônimos) derivam de conceitos mais gerais (hiperônimos). Nesse contexto, "ancestral comum" refere-se ao conceito mais geral de onde os conceitos mais específicos derivam. Por exemplo, um *synset* para o conceito de pessoa pode incluir palavras como ser humano, indivíduo, e pessoa. Além disso, conceitos como médico e professor são hipônimos do conceito mais geral pessoa.

De acordo com a Associação Brasileira de Normas Técnicas (ABNT (2016)) e a *Perkins School for the Blind* (Lewis (2018)), é importante evitar o uso de sentenças curtas, em virtude de serem mais propensas a possuírem detalhes óbvios e menos informativas, como, por exemplo, "*man's eyebrow*", "*a man has two eyes*" e "*"the man's lips*". Portanto, calcula-se o comprimento de todas as legendas densas que permaneceram no conjunto de descrições após as filtragens anteriores. Posteriormente, obtém-se o comprimento mediano desse conjunto de legendas, removendo-se todas as sentenças que não atingem esse comprimento. Optou-se pela mediana em vez da média, uma vez que essa medida não é influenciada por valores

extremos, eliminando assim aproximadamente 50% das legendas mais curtas.

Ao analisar diferentes partes da imagem, o modelo de legendagem densa pode produzir legendas que descrevam diferentes sujeitos (por exemplo, "*man*" e "*woman*") com base nos padrões encontrados mesmo não havendo mais do que uma pessoa na imagem. Dessa forma, se $P = 1$, haverá a padronização do substantivo do sujeito com base no substantivo mais frequente no conjunto de legendas que permaneceram após todo o processo de filtragem descrito.

A padronização do substantivo é realizada visando a obtenção de melhores resultados no gerador de resumos (Seção 4.1.2), uma vez que será oferecido como entrada legendas que descrevam unicamente a mesma pessoa, possibilitando como resultados textos mais precisos sobre a imagem de entrada. Em caso de empate, é retornado o primeiro sujeito descrito encontrado, uma vez que o conjunto de legendas gerado pelo modelo de legendagem densa é ordenado de acordo com o nível de confiança, do mais alto para o mais baixo. Portanto, o primeiro sujeito descrito tem a maior confiança associada.

Ao final desta etapa, é obtido um conjunto reduzido de descrições textuais semanticamente diversas, relacionadas ao conceito de pessoa (e seu entorno) e que tendem a ser longas o suficiente para não possuírem informações óbvias ou redundantes. As etapas juntamente com a ordem de todo o processo de filtragem podem ser visualizadas na Figura 4.3.

Figura 4.3: Fluxograma de todo processo de filtragem das legendas densas. Fonte: Material próprio.

Classificadores

Com o intuito de complementar os resultados produzidos pelo modelo de legendagem densa, nesta etapa, são aplicados modelos de Aprendizado de Máquina treinados especificamente para tarefas focada na pessoa presente na imagem, como detecção de idade, reconhecimento de emoção e classificação de cena, todos de forma independente. Esses modelos acrescentam informações de alto nível sobre características relevantes para a criação de contexto de imagem para pessoas com CBV (ABNT (2016); Lewis (2018)), uma vez que não há garantias de que o uso exclusivo de um modelo de legendagem densa seja suficiente para capturar todas essas informações a partir da imagem de entrada.

Devido ao desafio de prever a idade correta de uma pessoa, o valor retornado pelo modelo detector de idade é agrupado por meio da classificação adaptada de faixa etária proposta pela Organização Mundial da Saúde (OMS) (Ahmad et al. (2001); OMS (2014)). Sendo assim, dado o valor inteiro inferido pelo modelo para a idade (APPARENT_AGE), será informado se a pessoa da imagem pertence à faixa etária criança ("*child*") se APPARENT_AGE $<$ 10, jovem ("*young*") se APPARENT_AGE $<$ 20, adulta ("*adult*") se APPARENT_AGE $<$ 45, meia-idade ("*middle-aged*") se APPARENT_AGE $<$ 65 ou idosa ("*elderly*") caso APPARENT_AGE $\geq$ 65.

Em relação aos modelos de reconhecimento de emoção e classificação de cena, suas saídas são utilizadas apenas quando o valor de confiança de cada modelo é superior ou igual a um limite definido, denominado $T_{model\_confidence}$. Além disso, para evitar inconsistências nas saídas, os modelos de detecção de idade e reconhecimento de emoção são aplicados apenas quando uma única pessoa é detectada na imagem, ou seja, se $P = 1$.

Geração de sentenças e concatenação de texto

A partir da saída produzida anteriormente pelos classificadores aplicados, sentenças coerentes são criadas para incluir as informações a respeito da faixa etária, emoção e da cena no conjunto de legendas densas filtradas para a imagem de entrada. As sentenças geradas nessa etapa seguem uma estrutura padrão: "*there is a/an <AGE> <SUBJECT>*", "*there is a/an <SUBJECT> who is <EMOTION>*", e "*"there is a/an <SUBJECT> in the <SCENE>*", onde *AGE*, *EMOTION* e *SCENE* são, respectivamente, as saídas dos modelos classificadores de idade, emoção e cena, e *SUBJECT* é o substantivo mais frequente relacionado à pessoa.

No caso de $P > 1$, as sentenças com as informações sobre as faixas etárias e emoções das pessoas não são criadas, e a sentença sobre a cena passa a ser "*in the <SCENE>*". Exemplos de sentenças geradas são: "*there is a middle-aged woman*", "*there is a child who is sad*", e "*"there is a boy in the office*". Por fim, na última etapa dessa primeira fase, as sentenças complementares produzidas são concatenadas, de forma simples, com o conjunto de legendas densas previamente filtradas, formando um único texto.

4.1.2 Geração de contexto

Nesta fase de três etapas, é gerado o contexto da imagem de entrada por meio do resumo das informações textuais das características de interesse extraídas durante a fase de Analisador de Imagem.

Gerador de resumos

Para criar uma sentença semelhante à humana, ou seja, uma estrutura textual coerente e semanticamente conectada, um PLM é aplicado para produzir um resumo (sumarização) abstrativo do conjunto de legendas retornadas na fase anterior. Nessa etapa, o PLM é executado cinco vezes sobre o texto com os dados de entrada para gerar diferentes resumos a respeito das informações de interesse da imagem.

São produzidas cinco opções de resumos na tentativa de obter melhores resultados, isto é, resumos de maior qualidade, ao invés de apenas um. Para tal, é alternado o valor do parâmetro de largura da heurística de pesquisa de feixe do PLM. No contexto geral, o número cinco foi motivado pela quantidade de legendas independentes por imagens presentes nos principais conjuntos de dados para a tarefa de legendagem de imagens, como o MS COCO (Lin et al. (2014)) e o Flickr30 (Plummer et al. (2015)).

Limpeza de resumos

Após a geração dos resumos, um passo importante é garantir que cada um dos resumos produzidos tenha uma alta similaridade semântica com o texto de entrada. Para isso, assegura-se a similaridade entre a saída do PLM e sua entrada por meio da filtragem das sentenças contidas nos resumos.

De forma similar apresentada na Seção 4.1.1, utiliza-se a representação de palavras e a similaridade do cosseno para calcular a similaridade semântica entre os textos. Se o valor da similaridade entre cada sentença do resumo gerado e as sentenças do texto de entrada for menor que um limite definido, denominado α, a sentença do resumo é removida. Esse processo se repete até que todas as sentenças que compõem o resumo sejam analisadas.

Também são removidas repetições e sentenças duplicadas dos resumos produzidos. Se em pares de sentenças dentro do resumo, o valor da similaridade do cosseno ultrapassar um outro limite estabelecido, chamado β, a sentença subsequente à inicial ou àquela que já permanece no resumo é removida para diminuir a redundância do texto gerado. Esse processo também é repetido enquanto houver sentenças a serem verificadas no resumo.

Como resultado dessa etapa, após a aplicação dos dois filtros baseados em limiares, apresentados no Algoritmo 2, são mantidos somente os resumos cujas sentenças são relacionadas com a entrada e com baixa probabilidade de redundância entre si.

Algoritmo 2: Limpeza das frases presentes nos resumos

```
1: function summaryCleaning(summary_output, references_input, α, β)
2:     list_remove_sentences ← [ ]
3:     list_alpha ← [ ]
4:     list_beta ← [ ]
5:     for each sentence in summary_output do
6:         sentence_tokenized ← text_preprocessing(sentence)
7:         for each input in references_input do
8:             input_tokenized ← text_preprocessing(input)
9:             cosine_s ← semantic_s(sentence_tokenized, input_tokenized)
10:            if cosine_s ≥ α then
11:                break
12:            else
13:                list_remove_sentences ← sentence
14:            end if
15:        end for
16:    end for
17:    for each sentence in summary_output do
18:        if sentence not in list_remove_sentences then
19:            list_alpha ← sentence
20:        end if
21:    end for
22:    similar_summary_input ← concatenate(list_alpha)
23:    for each next_sentence in similar_summary_input do
24:        if list_beta = ∅ then
25:            list_beta ← next_sentence
26:        else
27:            ns_tokenized ← text_preprocessing(next_sentence)
28:            count ← 0
29:            number_sentences_beta ← length(list_beta)
30:            for each previous_sentence in list_beta do
31:                count ← count + 1
32:                ps_tokenized ← text_preprocessing(previous_sentence)
33:                cosine_s ← semantic_s(ns_tokenized, ps_tokenized)
34:                if cosine_s < β then
35:                    if count ≥ number_sentences_beta then
36:                        list_beta ← next_sentence
37:                        break
38:                    else
39:                        continue
40:                    end if
41:                else
42:                    break
43:                end if
44:            end for
45:        end if
46:    end for
47:    not_redundance_summary ← concatenate(list_beta)
48:    return not_redundance_summary
```

Estimativa de qualidade

De acordo com Krause et al. (2017), os parágrafos mais informativos geralmente apresentam características linguísticas, como múltiplas sentenças conectadas por conjunções coordenativas (por exemplo, "*and*", "*or*", "*but*"), uso de determinantes (por exemplo, "*a*", "*an*", "*the*"), e fenômenos linguísticos complexos como correferência[1], além de uma maior frequência de verbos e pronomes. Com base nisso, foi desenvolvida uma estratégia para estimar a qualidade dos resumos gerados sem a necessidade de acessar as referências *Ground-Truth*, utilizando apenas as características linguísticas mencionadas.

Para garantir que o resumo retornado seja o mais informativo possível, e, consequentemente, de melhor qualidade textual, contabiliza-se a frequência total das características mencionadas (conjunções coordenativas, determinantes, verbos e pronomes) para cada um dos cinco resumos gerados e filtrados, com base na verificação dos atributos de anotação linguística dos *tokens* de cada sentença, conforme descrito na Seção 4.1.1. Em caso de empate, é retornado o último resumo empatado, pois este terá a maior largura da pesquisa de feixe e, possivelmente, a melhor performance.

O aumento na largura da pesquisa heurística permite explorar um conjunto mais amplo de possíveis sequências durante a geração do texto, o que pode levar a escolhas mais robustas e otimizadas em termos de probabilidade cumulativa das sequências. Isso reduz a chance de o algoritmo ficar preso em ótimos locais e melhora a qualidade geral da sequência gerada (Wiseman and Rush (2016)). Assim, a estratégia proposta, conforme descrito no Algoritmo 3, resulta na seleção do resumo (parágrafo) da imagem que apresenta a maior quantidade de características linguísticas de interesse.

4.2 Experimentos

Nesta seção, são fornecidos os detalhes a respeito da avaliação experimental realizada na abordagem proposta neste capítulo.

4.2.1 Conjunto de dados

O domínio deste trabalho são imagens de webinário. No entanto, até o desenvolvimento desta abordagem, não existiam conjuntos de dados específicos para esse fim. O mais próximo encontrado foi o conjunto de dados *Stanford*

[1]Maneira de identificar as diversas formas que uma entidade nomeada pode assumir em determinado texto, como substantivo e pronome.

Algoritmo 3: Avaliação dos resumos conforme as características linguísticas informativas

```
1:  function QUALITYESTIMATION(list_summaries)
2:      number_characteristic ← [ ]
3:      for each summary in list_summaries do
4:          summary_tokenized ← tokenization(summary)
5:          count ← 0
6:          for each token in summary_tokenized do
7:              is_interest ← linguistic_attribute_check(token)
8:              if is_interest = true then
9:                  count ← count + 1
10:             end if
11:         end for
12:         number_characteristic ← count
13:     end for
14:     id_highest ← −1
15:     for each i and value in number_characteristic do
16:         if value = max(number_characteristic) then
17:             id_highest ← i
18:         end if
19:     end for
20:     final_summary ← list_summaries[id_highest]
21:     return final_summary
```

Image-Paragraph[2] (SIP), mas várias adaptações precisaram ser realizadas, conforme descritas posteriormente. O SIP é um subconjunto do *Visual Genome* (VG), amplamente aplicado em tarefas de geração de parágrafos para imagens, sendo composto por 19.561 imagens, cada uma com um único parágrafo escrito por humanos (Krause et al. (2017)).

O conjunto de dados do VG, por sua vez, contém aproximadamente 108.000 imagens com suas respectivas legendas densas, também anotadas por humanos, que descrevem os objetos detectados, seus atributos e relacionamentos (Krishna et al. (2017b)). O VG foi utilizado nesta abordagem para possibilitar o acesso às legendas densas das imagens presentes no SIP, uma vez que todas as imagens estão contidas no VG, embora para propósitos distintos.

Como o SIP contém uma vasta variedade de conteúdos, por exemplo, imagens de pessoas, animais, esportes, etc., as imagens desse conjunto de dados precisaram ser selecionadas. Portanto, foram escolhidas apenas as imagens próximas ao domínio de interesse deste projeto, isto é, imagens de webinário. Nesse caso, essas imagens seriam aquelas exclusivamente com pessoas em primeiro plano ou que ocupassem uma grande área da imagem. Esse enquadramento possibilita, principalmente,

[2]Disponível publicamente em https://cs.stanford.edu/people/ranjaykrishna/im2p/index.html

evidenciar expressões, semblantes, gestos, fisionomias e emoções, além das vestimentas e um pouco do ambiente em que o sujeito em destaque se encontra.

Para realizar a filtragem mencionada, inicialmente, foram analisadas as legendas densas anotadas por humanos do VG para cada imagem do SIP. Assim, apenas as imagens com, no mínimo, uma legenda densa diretamente relacionada a uma pessoa foram mantidas. Para isso, foram utilizados novamente os *synsets* da WordNet. Se algum *token* da legenda densa anotada presente na imagem analisada pertencesse ao mesmo campo semântico do significado mais abrangente de "pessoa", a imagem não era removida do conjunto de dados.

Em seguida, foi realizada uma verificação dupla para garantir a presença de pessoas nas imagens. Para isso, utilizou-se o YOLOv3 (Redmon and Farhadi (2018)) para detectar a presença de pessoas. Imagens nas quais as pessoas foram detectadas com uma confiança inferior a 60% foram descartadas. Além disso, foram removidas todas as imagens em que as pessoas ocupavam apenas uma pequena área ao fundo (por exemplo, uma imagem de rua com uma pessoa distante no fundo). Foram mantidas apenas as imagens que possuíssem pelo menos uma bbox associada ao conceito de pessoa com uma área maior ou igual a 50% da área total da imagem. Esses critérios foram estabelecidos para garantir que as características de interesse relacionadas às pessoas fossem suficientemente representadas nas imagens selecionadas.

Ao término de todo processo de filtragem, o conjunto de dados SIP adaptado passou a conter cerca de 2.147 imagens exclusivamente de pessoas em primeiro plano. O processo de filtragem do conjunto de dados mencionado é apresentado por meio do fluxograma da Figura 4.4.

Em relação às anotações, uma vez que os parágrafos anotados por humanos, também conhecidos como *Ground-Truth* (GT), do SIP contam uma história geral

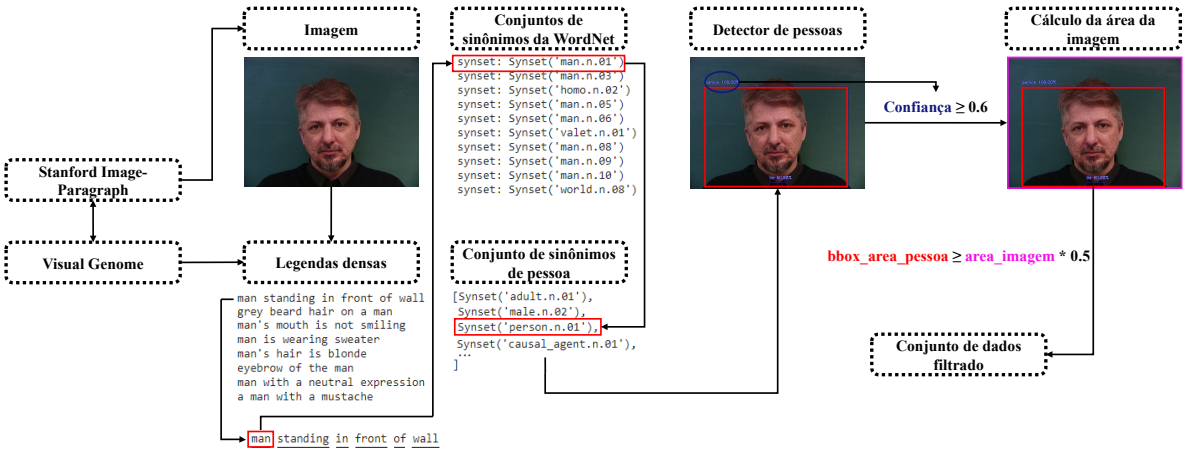


Figura 4.4: Fluxograma do processo de filtragem das imagens do conjunto de dados SIP. Fonte: Material próprio.

sobre a imagem, eles contêm sentenças que estão além do contexto de interesse desta tese. Para fazer uma comparação justa posteriormente, as sentenças que não possuíam palavras relacionadas a pessoas, verificadas por meio dos atributos linguísticos e do uso dos *synsets* da WordNet, foram removidas de seus parágrafos anotados. Como resultado, as anotações do SIP também foram moldadas para o domínio do problema tratado nesta tese.

As diferenças nas anotações entre o SIP original e o SIP adaptado para as 2.147 imagens utilizadas podem ser visualizadas na Tabela 4.1. As principais observações foram a respeito do maior comprimento médio dos parágrafos, quantidade média de palavras e tamanho do vocabulário contido no SIP sem o ajuste de suas anotações, correspondendo cerca de 28-30% de diferença média entre esses elementos.

Elem. linguísticos	**SIP Original**	**SIP Adaptado**
Caracteres	312 ± 125	225 ± 110
Palavras	63 ± 25	45 ± 22
Frases	5 ± 2	4 ± 2
Substantivos	24,6%	25,4%
Verbos	8,8%	10,0%
Adjetivos	9,8%	9,8%
Pronomes	5,5%	5,9%
Conj. Coordenativas	2,9%	3,1%
Tam. Vocabulário	5.116	3.623

Tabela 4.1: Comparação entre as estatísticas dos elementos linguísticos dos parágrafos gerados por humanos do SIP antes (original) e depois (adaptado) da filtragem das anotações GT.

4.2.2 Detalhes da implementação

Em relação às etapas mencionadas na Seção 4.1.1, para realizar a detecção de pessoas, utilizou-se a YOLOv3[3] (Redmon and Farhadi (2018)) com uma confiança mínima de $0,6$. Como modelo de legendagem densa, foi utilizado o DenseCap[4] (Johnson et al. (2016)). Para reduzir a quantidade de legendas densas semelhantes produzidas, adotou-se o limite de similaridade T_{text_sim} de $0,95$.

A segmentação das palavras foi realizada por meio do método Tokenizer do spaCy com as configurações padrões e regras para o idioma inglês. Para os cálculos de similaridade semântica e verificação dos atributos de anotação linguística entre as sentenças, usou-se o modelo `en_core_web_lg` da biblioteca spaCy (Honnibal et al. (2020)) com a adição do `universal_sentence_encoder` da Google que utiliza por

[3]Disponível publicamente em https://github.com/OlafenwaMoses/ImageAI
[4]Disponível publicamente em https://github.com/jcjohnson/densecap

padrão uma rede de média profunda (*Deep Averaging Network* - DAN) para fornecer incorporações de sentenças com uma dimensão de 512 (Cer et al. (2018)).

Em relação aos detalhes dos classificadores de Aprendizado de Máquina utilizados, empregou-se o Deep EXpectation[5] (DEX) (Rothe et al. (2015)) para prever a idade aparente das pessoas em imagens estáticas do rosto. Para o reconhecimento de emoção, utilizou-se o modelo Emotion-detection[6] (Goodfellow et al. (2013)), treinado no conjunto de dados FER-2013. Esse modelo classifica as emoções em uma das seguintes sete categorias: "*angry*", "*disgusted*", "*fearful*", "*happy*", "*neutral*", "*sad*" e "*surprised*". Para a classificação de cena, utilizou-se o PlacesCNN[7] (Zhou et al. (2017)), treinado no subconjunto *Places365-Standard* para prever 365 tipos de categorias de cena, sejam internas ou externas. Para ambos os modelos de reconhecimento de emoção e classificação de cena, foi estabelecido um limite de confiança $T_{model\_confidence}$ de $0,6$.

Na Seção 4.1.2, utilizou-se a tarefa de sumarização abstrativa com o modelo T5 pré-treinado (Raffel et al. (2020)) para gerar parágrafos. A implementação do Huggingface do T5 (Wolf et al. (2020)) foi empregada com o modelo ajustado `T5-base` (*checkpoint* `mrm8488/t5-base-finetuned-summarize-news`), treinado no conjunto de dados *News Summary* para a tarefa de sumarização abstrativa. As larguras da heurística de pesquisa de feixe variaram entre $2 - 6$. Para garantir que as sentenças nos resumos gerados fossem relevantes e não repetitivas, os valores de $0,7$ e $0,5$ foram definidos para os limites α e β, respectivamente. Além desses ajustes, nenhum outro parâmetro foi modificado, mantendo-se os valores padrão propostos pelos autores.

Para a implementação da abordagem proposta, foi utilizada a linguagem de programação Python (versão 3.8.5) em um ambiente de desenvolvimento Jupyter Notebook. O *hardware* empregado foi um desktop padrão, equipado com 8 GB de RAM, processador Intel Core i5-44605S com 2,90 GHz. O sistema operacional utilizado foi o Linux Ubuntu versão 20.04 LTS, versão de 64 bits.

4.2.3 Baselines

A abordagem proposta, descrita na Seção 4.1, utiliza um *pipeline* de técnicas e modelos de Aprendizado Profundo e resultou no método denominado Ours, com uma variante chamada Ours-GT (*Ours Ground-Truth*). Essa variante utiliza, como entrada, as legendas densas anotadas por humanos presentes no conjunto de dados do VG, em vez das legendas densas geradas automaticamente pelo DenseCap.

[5]Disponível publicamente em https://sefiks.com/2020/09/07/age-and-gender-prediction-with-deep-learning-in-opencv/

[6]Disponível publicamente em https://github.com/atulapra/Emotion-detection

[7]Disponível publicamente em https://github.com/CSAILVision/places365

Portanto, ela serve como um limite superior para o método Ours, alterando apenas a entrada, isto é, as legendas densas utilizadas.

O Ours e sua variante foram comparados com outros dois métodos concorrentes, ambos baseados no trabalho de Krause et al. (2017): Concat e Concat-Filter. O método Concat cria um parágrafo concatenando de maneira simples todas as saídas previstas pelo DenseCap. Já o método Concat-Filter, concatena apenas as saídas do DenseCap que possuem pelo menos uma palavra associada ao conceito de pessoa, conforme mencionado na Seção 4.1.1.

4.2.4 Métricas de avaliação

O desempenho dos concorrentes foi medido por meio das seguintes métricas automáticas: BLEU-{1, 2, 3, 4} (Papineni et al. (2002)), METEOR (Banerjee and Lavie (2005)) e CIDEr (Vedantam et al. (2015)). Todas essas métricas são comumente utilizadas em trabalhos anteriores de geração de parágrafos (Krause et al. (2017); Chatterjee and Schwing (2018); Zha et al. (2022)), sendo adotadas para manter o padrão da literatura.

As pontuações foram calculadas usando o código de implementação do coco-caption[8] (Lin et al. (2014)). Valores mais altos retornados são melhores para todas as métricas.

4.3 Resultados e Discussão

4.3.1 Resultados quantitativos

A seguir, são apresentados os resultados da avaliação experimental realizada para validar o método proposto. Os resultados presentes na Tabela 4.2 mostram que o método Concat apresentou o pior desempenho, principalmente na métrica CIDEr, que mede a qualidade do texto em relação ao ideal para seres humanos. Acredita-se que isso ocorra devido ao excesso de legendas produzidas de forma independente, demonstrando a incapacidade desse método em produzir sentenças coerentes (Krause et al. (2017)).

Comparado ao Concat, o Ours obteve melhor desempenho na maioria das métricas, com exceção da METEOR. Essa métrica, que usa correspondências exatas, radicais, sinônimos e paráfrases entre n-gramas para alinhar sentenças (Banerjee and Lavie (2005); Anderson et al. (2016)), favorece parágrafos candidatos que contêm mais palavras, como os produzidos pelo Concat, aumentando a probabilidade de

[8]Disponível publicamente em https://github.com/tylin/coco-caption

Método	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Concat	6,4	3,9	2,3	1,3	11,4	0,0
Concat-Filter	**29,8**	**17,5**	**9,9**	**5,5**	**16,4**	**9,6**
Ours	16,9	9,6	5,3	3,0	10,9	9,1
Ours-GT	*22,3*	*12,1*	*6,6*	*3,7*	*12,5*	*15,3*

Tabela 4.2: Comparação entre métodos concorrentes geradores de parágrafos usando as métricas automáticas para medir a qualidade entre os textos candidatos e os de referência. Os resultados abaixo da linha tracejada mostram a qualidade atingida pela variante Ours-GT nas métricas automáticas. Todos os valores são relatados como porcentagem (%). Os melhores valores estão em negrito.

gerar mapeamentos entre os parágrafos comparados e, consequentemente, pontuações mais altas.

As descrições geradas pelo método Concat-Filter produziram pontuações mais altas em comparação com a abordagem proposta. No entanto, é importante notar que ambos os métodos alcançaram pontuações próximas na métrica CIDEr. Diferentemente das métricas BLEU e METEOR, que são destinadas à tarefa de tradução automática, a CIDEr foi projetada especificamente para avaliar métodos de legendagem de imagens com base no consenso humano (Vedantam et al. (2015)). Além disso, a pontuação CIDEr do Concat-Filter é menor do que o valor obtido pelo Ours-GT, demonstrando o potencial do método proposto como um limite superior.

Ao analisar as estatísticas de linguagem apresentadas na Tabela 4.3, notou-se que o Concat-Filter produziu parágrafos com comprimento médio (considerando o número de caracteres) e quantidade média de palavras mais próximos dos parágrafos anotados por humanos. No entanto, seus parágrafos têm três vezes mais sentenças que os parágrafos humanos.

Com base nas informações da Tabela 4.3, pode-se inferir que as sentenças do Concat-Filter são curtas, enquanto os métodos Ours e Ours-GT produzem parágrafos mais próximos dos humanos em relação à quantidade média de sentenças por

Método	Caracteres	Palavras	Sentenças
Concat	2.108 ± 311	428 ± 58	89 ± 12
Concat-Filter	294 ± 116	62 ± 23	12 ± 5
Ours	109 ± 42	22 ± 8	3 ± 1
Ours-GT	139 ± 54	26 ± 10	3 ± 1
Humanos	225 ± 110	45 ± 22	4 ± 2

Tabela 4.3: Estatísticas de linguagem sobre a média e o desvio padrão no número de caracteres, palavras e frases nos parágrafos gerados por cada método concorrente. Os valores foram arredondados.

parágrafo. Além disso, esses são os únicos métodos que produziram parágrafos com a quantidade média de caracteres dentro da quantidade sugerida por Lewis (2018), sendo até 280 caracteres (comprimento de um *tweet*). A ampla gama de informações contidas na saída gerada pelo Concat-Filter explica as maiores pontuações alcançadas em BLEU-{1,2,3,4} e METEOR, uma vez que essas métricas estão diretamente relacionadas à grande quantidade de palavras presentes nos parágrafos produzidos.

É importante observar que, para todas as características de linguagem analisadas (quantidade média de caracteres, palavras e sentenças) na Tabela 4.3, as saídas produzidas pelos métodos Ours e Ours-GT apresentaram uma menor quantidade em relação aos parágrafos humanos. Esses resultados ajudam a explicar as pontuações retornadas pela métrica BLEU para ambos os métodos, uma vez que a BLEU aplica a penalização de brevidade para textos candidatos (os parágrafos gerados) mais curtos em relação aos textos de referência (os parágrafos humanos, isto é, as GTs). Assim, como todos os parágrafos produzidos pelos métodos da abordagem proposta são menores do que os parágrafos anotados pelos humanos, eles são penalizados pela métrica, retornando pontuações inferiores em comparação ao Concat-Filter.

Para demonstrar que as melhores pontuações alcançadas pelo Concat-Filter foram devido ao maior número de sentenças e palavras, experimentou-se remover aleatoriamente uma porcentagem das sentenças dos parágrafos gerados pelo Concat-Filter. O número médio de sentenças ficou mais próximo dos parágrafos descritos por humanos quando 75% das sentenças foram retiradas. Ao executar as métricas automáticas novamente, percebeu-se que os valores do Concat-Filter diminuíram substancialmente, conforme demonstrado na Tabela 4.4.

Considerando apenas 25% das sentenças do Concat-Filter, o Ours ultrapassou os resultados em todas as métricas analisadas. O Concat-Filter só começou a obter um

Método	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Concat-Filter-25%	7,4	4,0	2,1	1,1	8,2	4,7
Ours	**16,9**	**9,6**	**5,3**	**3,0**	**10,9**	**9,1**
Concat-Filter-30%	11,3	6,1	3,2	1,7	9,2	6,2
Concat-Filter-35%	16,0	8,7	4,6	2,4	10,4	8,1
Concat-Filter-40%	19,4	10,6	5,6	3,0	11,0	9,0
Concat-Filter-45%	22,6	12,4	6,6	3,6	11,7	9,9
Concat-Filter-50%	25,9	14,4	7,8	4,2	12,5	11,8

Tabela 4.4: Comparação entre a saída do método Ours com a saída filtrada do Concat-Filter para corresponder ao número de sentenças dos parágrafos escritos por humanos. Os resultados abaixo da linha tracejada mostram as pontuações das métricas quando a filtragem de frases é reduzida. Todos os valores são relatados como porcentagem (%). Os melhores valores estão em negrito.

desempenho comparável com o Ours quando foram preservados ao menos 40% do total das sentenças dos seus parágrafos (uma média de aproximadamente cinco sentenças por parágrafo). Esta é uma quantidade maior em comparação com a média de sentenças presentes nos parágrafos descritos por humanos e nos métodos da abordagem proposta, de acordo com a Tabela 4.3.

Ao analisar a Tabela 4.5, percebe-se que tanto o Concat quanto o Ours produziram parágrafos com proporções de substantivos mais próximas dos parágrafos escritos por humanos. Embora o Concat tenha um vocabulário mais amplo, contendo mais adjetivos e conjunções coordenativas em comparação ao Concat-Filter e ao Ours, sabe-se que essas quantidades estão diretamente relacionadas com o comprimento dos parágrafos (baseados na quantidade de caracteres) gerados por esse método, conforme a Tabela 4.3.

Em geral, o Ours foi o método que produziu parágrafos com quantidades de elementos linguísticos mais próximas dos parágrafos anotados por humanos na maioria das marcações de classes gramaticais (do inglês, *Part-of-Speech (POS) tagging*) analisadas. Além disso, o Ours foi o único método capaz de gerar uma quantidade superior de verbos e pronomes. Segundo Krause et al. (2017), dada a natureza dos parágrafos, é de se esperar que eles possuam mais verbos e pronomes, uma vez que esse tipo de descrição textual vai além da presença de objetos salientes, incluindo também informações sobre suas propriedades e relacionamentos.

Ainda conforme a Tabela 4.5, os valores abaixo da linha tracejada destacam novamente o potencial do método proposto como uma aplicação de limite superior. Isso ocorre porque, ao utilizar diretamente as legendas densas anotadas por humanos, amplia-se consideravelmente o vocabulário empregado, e a quantidade de elementos linguísticos se aproxima dos parágrafos humanos em todas as marcações POS, com exceção da classe de substantivos.

Método	**S**	**V**	**A**	**P**	**CC**	**VC**
Concat	**30,3**	3,3	14,3	0,0	**1,9**	**876**
Concat-Filter	32,6	8,6	8,4	0,1	0,5	418
Ours	**30,3**	**10,6**	**9,9**	**1,6**	1,5	642
Ours-GT	31,9	9,6	9,7	2,8	2,0	3.411
Humanos	*25,4*	*10,0*	*9,8*	*5,9*	*3,1*	*3.623*

Tabela 4.5: Estatísticas dos elementos linguísticos dos parágrafos gerados para a classe gramatical de substantivos (S), verbos (V), adjetivos (A), pronomes (P), conjunções coordenativas (CC) e tamanho do vocabulário (VC). Todos os valores são relatados como porcentagem (%), exceto para VC. Os valores em negrito são os mais próximos dos parágrafos humanos.

Os resultados apresentados evidenciam que os parágrafos gerados pela

abordagem proposta são mais adequados para pessoas com CBV, uma vez que suas características linguísticas são mais semelhantes às descritas por humanos e, principalmente, pelo comprimento mais curto. Conforme será visto na seção a seguir, o comprimento do texto é fundamental para uma melhor compreensão do contexto da imagem (Lewis (2018)).

4.3.2 Resultados qualitativos

A Figura 4.5 exibe exemplos dos parágrafos gerados pelos métodos competidores, com exceção do Concat, devido ao extenso tamanho de sua saída. O Concat-Filter retornou descrições de texto com estrutura simples, contendo numerosos e óbvios detalhes em geral. Essas observações corroboram os dados apresentados na Tabela 4.5, que demonstram uma baixa quantidade de pronomes e conjunções coordenativas presentes nos parágrafos produzidos.

Mesmo usando um PLM que não foi treinado especificamente para a tarefa de geração de parágrafos de imagem, tanto o Ours quanto o Ours-GT produziram os melhores parágrafos em relação à quantidade de elementos linguísticos, como pode ser visto na Tabela 4.5, sendo estes altamente relevantes em termos de interpretabilidade humana. Destaca-se, como uma propriedade interessante dos métodos Ours e Ours-GT, o uso de vírgulas e de conjunções coordenativas que possibilitam a suavização das transições entre as sentenças dos parágrafos.

Nem todas as saídas produzidas pelo método Ours são precisas, como pode ser visto na terceira coluna da Figura 4.5. O principal causador é o erro propagado pelos modelos pré-treinados usados na abordagem proposta, incluindo o DenseCap. Apesar de alguns casos falhos, os resultados ainda são inteligíveis. Graças à natureza flexível do Ours, os erros propagados podem ser minimizados substituindo os modelos utilizados no *pipeline* de técnicas por outros semelhantes com maior precisão. Um exemplo é o Ours-GT, no qual o DenseCap foi substituído pelas anotações descritas por humanos e os resultados retornados foram superiores em todas as métricas e análises qualitativas, demonstrando, assim, o potencial do método proposto.

Pode-se mencionar as seguintes limitações encontradas no estudo proposto: a propagação do erro a partir dos modelos utilizados; a possibilidade de amplificação do erro propagado na etapa de padronização do sujeito mais frequente; a mesclagem entre as legendas densas de pessoas diferentes quando a imagem possui mais de uma pessoa ($P > 1$); e o desaparecimento das informações de alto nível extraídas dos classificadores externos durante o processo de geração de resumos pelo PLM. Esta última limitação pode ser observada na Figura 4.5, onde nem a faixa etária, nem a emoção ou a cena classificada estiveram presentes nos parágrafos de imagem, com

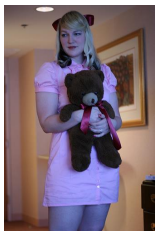
	Ground-Truth	Concat-Filter	Ours	Ours-GT
	A man is sitting in a room in front of a long desks. He is wearing a short sleeve shirt. There is a black phone sitting beside the man. The man is sitting on a black chair. A white door can be seen next to a wall behind the man.	Man wearing a black shirt. Man sitting at table. Man with short hair. A man sitting on a desk. A computer monitor. Black shirt on man. White wall behind man. Man holding a book. The hair of a man. The man is holding a phone. Man has a beard. The arm of a man. The man is wearing a hat.	Man wearing a black shirt. A man sitting at table. White wall behind the man. The man is holding a phone.	A man in black shirt sitting at a desk with a wrinkled brow, looking at a camera. The man's right hand is on the computer mouse. A white door behind the man is visible.
	A woman is dressed in a short light pink dress and holding a brown teddy bear with a dark pink bow. The woman is looking off to the left she has blond hair and a plaid bow in her hair. Her makeup is done. She is in a room.	Girl holding teddy bear. A woman is in a pink dress. The woman has brown hair. Wall behind the woman. Girl with long hair. The girl is wearing a pink shirt. The woman is wearing a skirt. Pink shirt on a woman. The woman is wearing glasses. White wall behind the man. A woman wearing a black shirt. The arm of a girl. A blue and white box. A white shirt on a man. A woman wearing a hat. The head of a woman. A wall behind the man. The man is wearing a brown shirt.	Woman holding teddy bear. A white wall behind her. A woman wearing black hat, a white shirt and a brown shirt.	Blond woman holding a stuffed teddy bear in her left hand. She is wearing a pink dress and has multicolored hair. The buttons are on the woman's dress.
	A man is sitting down. He has short hair and wear eye glasses. He is holding a remote control in his hand while posing for his picture. You can see he is sitting in a living room inside a home.	A man wearing a black shirt. Man with black hair. A man holding a phone. A man sitting on a couch. A black and white wii remote. The man has glasses. The hand of a man. The man has a beard. Man with a beard. The man has short hair. A picture of a woman. The man is wearing a tie. The ear of a man. The hair of a woman. White wall behind the woman.	Man wearing a black shirt. A man sitting on a couch. The man has glasses. He has short hair. He is wearing a white shirt. He is wearing a tie. White wall behind the man.	The man is dressed in a black sweater and is sitting by a black chair. He is wearing glasses with a frame made of plastic. The scene shows flowers beneath the lamp behind the man.
	A middle aged man with gray hair is looking straight ahead. He is dressed in a black suit gold tie gray colored shirt.	Man with glasses. Man has short hair. The man has a beard. The collar of a man. Man wearing a black suit. The nose of a man. The man has glasses. White shirt on man. The man is wearing glasses. The man is smiling. A man in a white shirt. The hair of a man.	A man in a discotheque has short hair. He has a beard. A man wearing a white shirt.	Grey haired man with dark eyes in a discotheque. A pink purple and green colorful screen behind a man's head. Old man wearing dress shirt and tie. Brown and gray hair on his head. Man in front of a darkened background.

Figura 4.5: Exemplos de parágrafos gerados pelos métodos comparados. A primeira coluna refere-se ao GT do SIP Adaptado, enquanto as outras colunas são os resultados de cada método, exceto para Concat. Fonte: Material próprio.

exceção do indicativo de faixa etária ("*old man*") e da cena ("*discotheque*") apenas no último exemplo.

A Figura 4.6 demonstra dois exemplos de casos falhos. Na primeira linha, uma mulher está presente na imagem e a saída do Ours a descreve como homem. Esse erro pode ser explicado porque o substantivo homem ("*man*") está mais presente na saída do DenseCap, conforme visto na segunda coluna por meio das palavras destacadas em vermelho. Esse erro não foi cometido pelo Ours-GT, uma vez que essa abordagem não é influenciada pela saída do DenseCap. No entanto, a alteração semântica dos sujeitos é considerada menos grave do que mencionar duas pessoas diferentes em uma imagem que apresenta apenas uma, como indicado na segunda coluna para essa mesma imagem.

O segundo exemplo da Figura 4.6 é uma imagem com duas pessoas em que todos os métodos produziram parágrafos com descrições equivocadas, incluindo o

	Ground-Truth	Concat-Filter	Ours	Ours-GT
	A mature woman with short red hair that is graying is speaking on the telephone. She has a surprised look on her face as she listens to the caller. She is sitting in a chair in a brightly lit office. She's wearing glasses and a yellow sweatshirt. Around her neck is a lanyard that has many colorful and exotic looking pins attached to it.	A woman wearing a tie. A woman with brown hair. White shirt on man. Man holding a cell phone. A white shirt on a man. The man is wearing a black watch. The man is wearing glasses. Man has glasses. The hair of a man. A man with a beard. The woman is holding a white umbrella. The arm of a man. White wall behind the woman. Hair on the womans head. The man has a beard. The ear of a man. The man is wearing a necklace.	A man with brown hair. A man holding a cell phone. A white shirt on a man. The man is wearing glasses and a black watch. He is wearing a necklace.	A black cell phone in the woman's hand. Blonde and gray hair on the woman's head. Pins attached to her lanyard. Glasses covering the woman's eyes. A clock on the wall behind the woman.
	A red headed man in thin wire rimmed glasses is sitting in front of a window overlooking a river with a woman. The man is wearing a brown raglan sleeved top and the woman is wearing a white camisole under a light blue sheer top. The woman is wearing her hair in a ponytail in the back of her head. The woman is holding a frisbee reading cheese shop wow. There is a white and red pole on the right of the man.	A woman with a dark hair. Man with short hair. A red and white cup. Man has a beard. Man wearing a blue shirt. The man has glasses. The woman is smiling. The woman is wearing a necklace. The man is wearing a tie. The collar of a man. The hair of a man. A woman with a beard. Red writing on the back of a man. The ear of a man. A man wearing a hat. A blue strap on the womans shoulder. A man in the water. A woman in a white shirt.	Man in a white shirt, wearing a black shirt and wearing a necklace. He has short hair. Red writing on the back of a man. A blue strap on his shoulder. A man in a dark hair.	Picture shows a middle-aged woman and a woman smiling together. Picture shows a woman combing with a ponytail. Picture shows a woman holding a red lid.

Figura 4.6: Exemplos de parágrafos gerados com diferentes casos falhos pelos métodos comparados. As sentenças e palavras destacadas em vermelho representam esses casos. A primeira coluna refere-se ao GT do SIP Adaptado, enquanto as outras colunas são os resultados de cada método, exceto para Concat. Fonte: Material próprio.

Ours-GT. Uma possível razão para isso é que o modelo gerador de resumos enfrenta dificuldades para associar várias legendas densas que se referem a diferentes pessoas de uma imagem em um único texto, uma vez que o modelo não foi treinado para essa finalidade, mas sim para resumir longos textos. Nesse caso específico, notou-se que algumas características de uma pessoa foram descritas como se fossem de outra. Esta é uma limitação da capacidade de transferência de conhecimento linguístico do modelo pré-treinado utilizado. Apesar disso, exemplos com mais de uma pessoa na imagem puderam ser descritos de forma coerente pelo Ours e Ours-GT, conforme mostrado na Figura 4.7.

Menciona-se que a falta de um conjunto de dados anotado adequado para o problema de descrição de imagens para deficientes visuais impacta diretamente a avaliação experimental realizada. Ao usar o conjunto de dados SIP, mesmo filtrando as legendas densas para manter apenas as sentenças que mencionassem pessoas, a maioria das GTs são genéricas, ou seja, não dão total ênfase nos detalhes das pessoas e em suas características. Por outro lado, a abordagem proposta inclui detalhes relativos às pessoas, como faixa etária e emoção, gerando parágrafos com informações específicas. Logo, mesmo fornecendo descrições precisas, o método proposto pode obter uma pontuação baixa nas métricas automáticas. Por exemplo, o método pode produzir um parágrafo que mencione a idade da pessoa em foco na

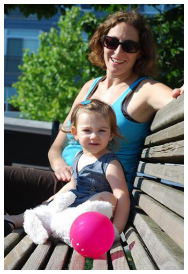
	Ground-Truth	Concat-Filter	Ours	Ours-GT
	A man and a woman are looking at each other. The man and women are facing each other. The woman has her arms folded. The woman has blonde hair. She is wearing a blue suit coat. She is wearing a white button down shirt under the jacket. The man is wearing a gray suit.	Man looking at a window. Man with short brown hair. A woman wearing a black shirt. The woman is smiling. A tie on a man. A man with dark hair. A man with a hat. Black jacket on man. The hair of a man. The woman has a brown hair. The man is wearing a gray shirt. White shirt on man. The ear of a woman. The woman is wearing a necklace. Black tie on man. White wall behind the man.	A man with short brown hair. A woman wearing a black shirt. A man looking at a window. White wall behind the man.	Woman has white collar. Black buttons on her jacket. Window behind blonde woman. Man is wearing a suit. Man's hair is dark brown.
	A woman is smiling and sitting on a bench with her daughter. The lady has some nice shades on brown hair a blue tank top with a brown shirt on under it and some brown pants. Her daughter is wearing a jean dress and has a blue and orange barrette in her hair	Woman with brown hair. Woman sitting on a bench. The girl is wearing a blue shirt. Woman wearing sunglasses. A girl with blonde hair. White teddy bear with a pink nose. Blue shirt on woman. The baby has a blue eyes. A man wearing a hat. A man holding a bat. Woman wearing a blue necklace. The woman is holding a camera. The woman is wearing sunglasses. The arm of a man. A gray shirt on a man. The man is wearing a watch. The woman has long hair.	Woman sitting on a bench. The girl is wearing a blue shirt. Woman wearing sunglasses. White teddy bear with pink eyes. A man is wearing a watch. Notably, the baby has blue eyes.	Little girl has a stuffed animal. She is wearing a blue jumper. The woman is wearing dark glasses. Orange and blue clip in the girl's hair. A toddler sitting next to the woman. A little girl dressed in black sandal.
	Two people are posing for a picture. On the right is a man. On the left is a woman.	A woman with a dark hair. Man with a black shirt. Man with glasses on face. The man has short hair. A man wearing a red shirt. The man has glasses. Woman wearing a necklace. The hand of a man. The man is smiling. A man wearing a hat. The man has a beard. A black strap on a woman. White wall behind the man.	A man with glasses on face. A woman wearing a necklace. A man wearing a red shirt. A woman with dark hair. White wall behind the man.	A man and a woman holding mostly empty glasses of wine. A gold chain and a silver pendant on the woman's neck.

Figura 4.7: Exemplos de parágrafos gerados de imagens com mais de uma pessoa. A primeira coluna refere-se ao GT do SIP Adaptado, enquanto as outras colunas são os resultados de cada método, exceto para Concat. Fonte: Material próprio.

imagem que não faz parte da GT. Neste exemplo, foi adicionada uma informação relevante e complementar que afetou negativamente as métricas.

Na Figura 4.7, foi possível notar alguns desses casos mencionados. As saídas geradas pelo Ours e Ours-GT no segundo exemplo fazem referência aos óculos de sol da mulher, sendo um objeto em destaque em seu rosto, mas não há menção a esse objeto na GT, assim como no brinquedo da criança. Já na última imagem, percebe-se que o parágrafo anotado por humanos não menciona nenhuma característica das pessoas, diferente dos parágrafos do Ours e Ours-GT que focam nessas características.

A ausência dessas menções acima podem ser por causa da não inclusão das mesmas pelos anotadores humanos ou devido ao processo de filtragem realizado. Ainda sobre o segundo exemplo, observou-se também o uso da palavra "*nice*" na sentença "*the lady has some nice shades...*" da GT. De acordo com as diretrizes apresentadas pela ABNT (2016), deve-se evitar o uso de descrições subjetivas que fazem uso de adjetivos ou expressões que remetem a sentimentos, isto é, palavras que expressam o ponto de vista de quem as escreveu, uma vez que a interpretação da imagem deve ficar a cargo do usuário com CBV.

4.4 Desfechos Alcançados

Este capítulo apresentou os primeiros esforços em direção ao objetivo desta tese, focando na geração de descrições contextuais para imagens de webinários destinadas a pessoas com deficiência visual. A abordagem proposta se diferencia das abordagens existentes na literatura, que muitas vezes falham como Tecnologias Assistivas, ao sintetizar todos os elementos visuais de uma imagem em uma única sentença, resultando em descrições superficiais e genéricas. Alternativamente, algumas técnicas criam várias sentenças desconectadas para descrever diferentes regiões da imagem, resultando, no melhor cenário, em uma descrição longa e ainda assim genérica. Essas abordagens não atendem adequadamente às necessidades cognitivas e visuais dos deficientes visuais, prejudicando a compreensão do contexto das imagens de webinários e limitando a promoção da inclusão social.

A estratégia da abordagem proposta combinou legendagem densa com um conjunto de filtros para ajustar as legendas ao domínio de interesse e um PLM para a tarefa de sumarização abstrativa. Isso resultou na geração de parágrafos descritivos que enfatizam detalhes cognitivos e visuais para pessoas com CBV. Os parágrafos gerados foram avaliados utilizando diversas métricas automáticas sobre o conjunto de dados SIP adaptado para o problema.

A avaliação experimental demonstrou que é possível desenvolver um método de inclusão social para melhorar a vida de pessoas com deficiência visual ao combinar os benefícios de métodos existentes para análise de imagens – como modelos de legendagem densa, detectores de pessoas e classificadores de idade, emoções e cenas – com modelos de linguagem pré-treinados. Essa combinação permite explorar de forma eficaz o conhecimento adquirido por esses métodos para gerar descrições mais detalhadas e úteis.

Os resultados alcançados demonstraram que foi possível produzir automaticamente descrições textuais sobre contextos de imagens de webinários sem a intervenção humana, com maior nível de interpretabilidade e quantidades de elementos linguísticos similares aos dos parágrafos anotados por humanos. Embora tenha-se listado uma série de limitações, os achados levantados evidenciam o potencial e a importância da continuação do desenvolvimento do estudo proposto como Tecnologia Assistiva.

Capítulo 5

Coleta e anotação de imagens centradas no apresentador: Aprimorando a acessibilidade para deficientes visuais

Guiando os avanços em IA, um número significativo de conjuntos de dados de diversos domínios tem sido gerado. A combinação de conjuntos de dados e modelos de Aprendizado de Máquina revolucionou muitos campos, proporcionando sucesso na realização de tarefas do mundo real e disponibilizando ferramentas presentes em nossas vidas diárias. Exemplos notáveis incluem melhorias no campo da agricultura de precisão (Nascimento et al. (2022)), detecção de objetos em tempo real (Massiceti et al. (2021); Tang et al. (2023)) e estimativa do risco de contaminação por COVID-19 com base no distanciamento social e detecção de máscaras faciais (Passamani et al. (2022)).

A revolução da IA também visou a Tecnologia Assistiva, por exemplo, por meio da detecção de objetos perigosos (Shah et al. (2021)), legendagem de imagens (Gurari et al. (2019, 2020)) e sistemas de resposta a perguntas visuais para pessoas com deficiência visual (Gurari et al. (2018)). Apesar dos resultados impressionantes, esses métodos estão limitados a tarefas generalistas e domínios de dados de pessoas com visão devido a escassez de conjuntos de dados direcionado para o contexto de aplicações para deficientes visuais.

No contexto de legendagem de imagens, embora alguns esforços tenham sido propostos para automatizar o processo de descrição e eliminar a necessidade de pessoas sem deficiência visual descreverem imagens (Gurari et al. (2018, 2020)), como o projeto #PraCegoVer, a principal barreira para essa tarefa é o domínio dos dados, uma vez que estes não seguem a mesma distribuição de conjuntos de dados generalistas em massa (Gurari et al. (2018, 2019); Shah et al. (2021)). Além disso, as descrições de imagens para esse público não seguem as mesmas regras aplicadas ao conjunto de dados criado para pessoas sem deficiência visual (ABNT (2016); Lewis (2018)). Portanto, a necessidade de conjuntos de dados específicos para modelos supervisionados é fundamental.

Os achados apresentados no capítulo anterior confirmam a necessidade de um conjunto de dados específico para a tarefa de descrição de imagens para pessoas cegas ou com baixa visão, especialmente em cenas de webinários. A ausência de um conjunto de dados adequado impactou diretamente a avaliação experimental realizada. Embora o conjunto de dados SIP tenha sido adaptado e utilizado, ele não é ideal para a aplicação investigada nesta tese, principalmente por conter imagens com mais de uma pessoa e parágrafos anotados que não atendem aos requisitos da tarefa.

Diante dessa lacuna identificada, para atingir o Objetivo Específico 3 (OE-3), este capítulo apresenta uma abordagem para coletar imagens de conteúdo compartilhado por ferramentas de videoconferência com um único apresentador em destaque. Além disso, também é descrito, como parte do OE-4, um protocolo de anotação para descrever manualmente o contexto visual das imagens para esse público específico. O objetivo de ambas as propostas, alinhado ao OE-5, é construir um conjunto de dados de imagens que apoie tarefas de Visão Computacional relacionadas a webinários, voltadas especificamente para pessoas com CBV. Subsequentemente, as análises e experimentos realizados sobre o conjunto de dados construído são detalhados, seguidos pelos resultados, discussões e conclusões alcançadas.

5.1 Conjunto de Dados

Nesta seção, inicialmente é descrito o processo sistemático empregado para extrair e armazenar metadados e seus respectivos quadros (ou *frames*) dos vídeos de entrada. Em seguida, é detalhado o protocolo de anotação proposto para descrever os quadros dos vídeos, isto é, as imagens fixas coletadas. Ao final, é apresentada a criação do conjunto de dados de domínio específico deste trabalho por meio da junção do processo de coleta e anotação.

5.1.1 Coleta de dados

Para cada vídeo de entrada, analisa-se cada Δ-*ésimo* quadro do vídeo e uma base de dados vazia é criada para armazenar os metadados dos quadros, que incluem o identificador do quadro, o descritor facial (ou do rosto) e o valor de confiança da detecção facial. O fator Δ é usado para evitar o processamento de quadros com conteúdo muito semelhante, o que leva a uma alta probabilidade de ser o mesmo apresentador ou palestrante.

Conforme apresentado na Figura 5.1, o primeiro passo é aplicar um modelo de detecção facial para determinar se há um rosto no quadro (Figura 5.1a), seguido por uma verificação se apenas um único rosto foi identificado (Figura 5.1b), uma vez que

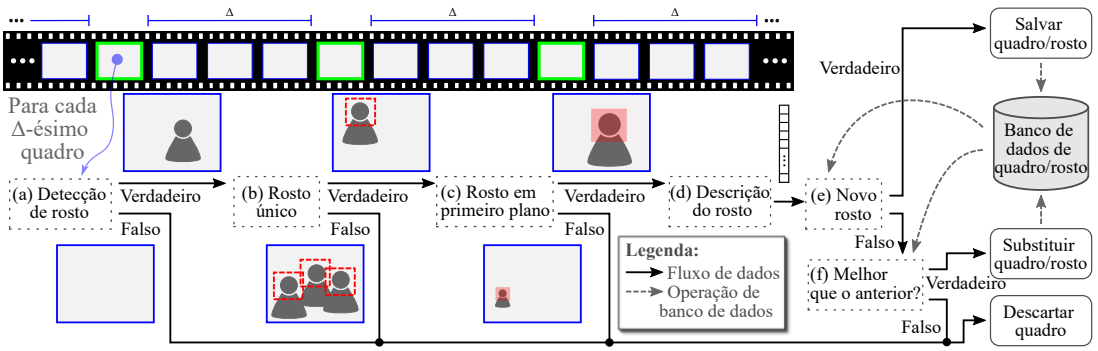


Figura 5.1: Visão geral do processo de coleta de dados para cada vídeo de entrada. Para cada quadro, se analisa a presença de um rosto (a), se o rosto é único (b) e de um apresentador em primeiro plano (c). Se alguma verificação retornar falso, o quadro é descartado. Caso contrário, é descrito o rosto (d) para verificar se é a sua primeira aparição (e). Em caso afirmativo, armazena-se o descritor do rosto. Em caso negativo, a versão em cache é substituída pela versão atual se a atual for melhor em nível de confiança. Fonte: Adaptado de Ferreira et al. (2023).

o método proposto para coleta visa quadros com apenas um indivíduo, já que o objetivo é criar um conjunto de dados de webinários com apenas um único apresentador. Em seguida, é verificado se o único rosto detectado está em primeiro plano, comparando se sua área é maior que $\gamma\%$ da área total do quadro (Figura 5.1c). Um retorno afirmativo indica que o rosto é de um indivíduo adequado para ser descrito e que o quadro vai ser extraído. Se alguma das três verificações retornar falso, o quadro é descartado e a análise do próximo é iniciada.

Após a extração do quadro, o rosto identificado é descrito (codificado) automaticamente por um modelo de detecção e codificação facial (Figura 5.1d) e usa-se o descritor criado para verificar se esta é sua primeira ocorrência (Figura 5.1e). Se este for o caso, os metadados do quadro são gravados no banco de dados correspondente. Isso evita salvar múltiplos quadros do mesmo indivíduo dentro do mesmo vídeo, uma vez que mais de uma anotação do mesmo apresentador no mesmo vídeo seria redundante.

Esse processo de verificação é realizado por meio da comparação da distância do cosseno entre o descritor do rosto atual e cada descritor já armazenado no banco de dados. Se os descritores faciais comparados alcançarem uma distância do cosseno menor que um limiar definido, denominado α, significa que o indivíduo atual pode já ter sido detectado. Assim, para garantir maior precisão, realiza-se uma dupla verificação por meio da distância euclidiana entre a potencial aparição anterior e o quadro atual. Se o valor da distância euclidiana for menor que um outro limite estabelecido, chamado β, isso indica que o indivíduo atual já foi identificado no vídeo em processamento e armazenado no banco de dados.

Para os casos em que o rosto já tenha sido detectado no vídeo, os valores da

confiança na detecção facial do quadro em análise e do quadro previamente gravado são comparados. Ao final, é mantido apenas o quadro com a maior confiança, independente do seu valor (Figura 5.1f).

A escolha pela adoção da dupla verificação com a distância euclidiana foi motivada pela menor precisão em salvar múltiplos quadros do mesmo apresentador ao utilizar apenas a distância do cosseno. Além disso, a ordem da escolha das distâncias foi determinada por meio de testes empíricos, que demonstraram melhores resultados no uso da distância do cosseno seguida pela euclidiana, ao invés do contrário.

Ressalta-se também que foi preferido a criação e o uso de uma abordagem de coleta de quadros em vez da técnica de sumarização dinâmica de vídeo (do inglês, *Video Skimming*) para selecionar os quadros dos vídeos de entrada, uma vez que o objetivo é obter o quadro com o rosto de melhor qualidade visual. A sumarização dinâmica de vídeo é utilizada para criar uma versão temporalmente resumida de um vídeo longo, identificando componentes significativos de seu conteúdo. Seu objetivo é fornecer uma representação geral rápida do vídeo, permitindo uma melhor compreensão a partir do seu resumo (K. et al. (2019)). Portanto, sua aplicação não se enquadra na tarefa de coleta abordada.

5.1.2 Protocolo de anotação

Foi proposto um protocolo para descrever o contexto visual de imagens, especificamente para pessoas com CBV, com base na análise das diretrizes sugeridas pela Associação Brasileira de Normas Técnicas (ABNT (2016)) e das boas práticas recomendadas pela *Perkins School for the Blind* (Lewis (2018)). Essas entidades, embora de países distintos, fornecem diretrizes em comum que visam à padronização da descrição de dados audiovisuais para garantir a qualidade da acessibilidade e atender às necessidades das pessoas com deficiência visual, assegurando equidade de direitos.

As diretrizes sugerem que as descrições textuais incluam características relevantes da aparência dos indivíduos, como gênero, faixa etária, cor da pele, cabelo e olhos, expressões faciais, além de outras informações visuais notáveis. Além disso, as descrições devem mencionar roupas, acessórios, objetos e a cena (ABNT (2016)). Mesmo que pessoas cegas nunca tenham visto cores, muitas vezes possuem conhecimento das associações de cores, o que pode ajudar na compreensão da imagem (Lewis (2018)).

Uma boa prática para construir uma descrição é garantir que o texto resultante seja preciso, simples, coerente, fluido, no tempo presente e dentro da faixa de 125 a 280 caracteres (ABNT (2016); Lewis (2018)). Todas essas informações são necessárias e, preferencialmente, devem ser fornecidas para garantir a qualidade da descrição e

facilitar a compreensão da imagem.

Uma vez estabelecidas as instruções, iniciou-se a tarefa de anotação em duas etapas conduzida por três anotadores humanos, que foram apresentados as diretrizes de descrição de imagem para deficientes visuais. Na primeira etapa, os anotadores utilizaram um *software* de anotação personalizado (desenvolvido pelo autor desta tese) para descrever cada imagem em formato de parágrafo, aderindo às normas e melhores práticas delineadas no conjunto de instruções do protocolo.

Na segunda etapa, foi realizado um ajuste cruzado das descrições, na qual um anotador revisa uma imagem juntamente com sua descrição previamente anotada por um dos outros dois anotadores. Assim, o anotador atual foi então encarregado de ajustar a descrição de acordo com o protocolo. O processo de ajuste cruzado foi conduzido de forma cega e aleatória, envolvendo possíveis adições, edições ou remoções à descrição.

Como resultado, cada imagem anotada resultou em três descrições: a descrição original e duas descrições ajustadas. As descrições foram inicialmente escritas em português brasileiro e, periodicamente, passavam por um processo de supervisão e curadoria dos dados de forma manual pelo autor desta tese para garantir que as anotações estivessem conforme o protocolo estabelecido. Posteriormente, as descrições foram traduzidas para o inglês usando uma ferramenta de tradução automática. Ressalta-se que os anotadores não tiveram acesso as descrições traduzidas.

Neste trabalho, contou-se com a colaboração de três estudantes do segundo ano do Ensino Médio da rede pública municipal e estadual da cidade de Viçosa (Minas Gerais) para a tarefa de anotação manual. Os estudantes foram selecionados através do processo seletivo de Bolsistas de Iniciação Científica do Programa BIC-Júnior/FAPEMIG 2022-2023, em parceria com a Universidade Federal de Viçosa (UFV). Eles foram encaminhados pela Comissão Interna de Seleção e Acompanhamento para atuar e ser orientados por professores e pesquisadores da UFV em um projeto científico. O BIC-Júnior tem como objetivo ampliar a formação desses estudantes, permitindo que mantenham contato com um projeto científico e despertando o interesse pela pesquisa científica.

Os estudantes, antes de iniciar a tarefa de anotação, passaram por um treinamento no qual lhes foram apresentados o objetivo e a importância deste trabalho de anotação de descrição de imagem para deficientes visuais, conforme as diretrizes estipuladas. Além da capacitação, durante as primeiras semanas, os estudantes foram acompanhados de perto pelos envolvidos nesta tese para garantir que entendessem as instruções necessárias e se ambientassem com a tarefa e o *software* de anotação.

5.1.3 Construção do conjunto de dados

Com o objetivo de criar um conjunto de dados composto por imagens diversas e representativas contendo apenas um único apresentador em primeiro plano, foi selecionado manualmente um conjunto de listas de reprodução de vídeos do YouTube publicamente disponíveis que atendiam a esse critério. Foram incluídos vídeos de webinários, notícias, entrevistas, aulas on-line, tutoriais, *podcasts*, e conteúdos similares, onde um apresentador se expressa enquanto outras pessoas assistem. Para cada vídeo dessas listas de reprodução, aplicou-se o processo de coleta de dados descrito na Seção 5.1.1, resultando em um banco de dados e um conjunto de quadros. Após processar todos os vídeos, os quadros extraídos foram organizados em três conjuntos de dados, conforme detalhado a seguir.

É importante destacar que, neste trabalho, não se restringiu apenas a vídeos de webinários de domínio educacional, como a apresentação de um seminário ou trabalho via ferramenta de videoconferência (Google Meet, Microsoft Teams e Zoom), para alcançar a maior quantidade possível de imagens que se encaixassem no contexto desta tese.

Conjunto de dados completo

Esta versão do conjunto de dados representa a união dos quadros selecionados de todos os vídeos em todas as listas de reprodução. Ressalta-se que, para cada vídeo, apenas uma imagem foi selecionada para cada pessoa.

A presença de um número significativo de vídeos com programas apresentados por um único anfitrião ou *host* (por exemplo, noticiários) resultou em várias imagens da mesma pessoa. Embora essas imagens provenham de vídeos diferentes, elas apresentam apenas pequenas variações, principalmente em roupas e acessórios.

O conjunto de dados completo contém 10.939 imagens, cada uma com dimensões de 1280×720 *pixels*, coletadas de 4.173 vídeos.

Conjunto de dados de pessoa única

Com o objetivo de aumentar a diversidade dos dados, foi proposta uma versão do conjunto de dados na qual cada pessoa aparece apenas uma única vez. Para isso, criou-se um banco de dados global de quadros por meio da concatenação do banco de dados de quadros do vídeo que está sendo processado com os bancos de dados de quadros dos vídeos processados anteriormente, na ordem armazenada no conjunto de dados completo.

Em seguida, inseriu-se apenas o quadro atualmente selecionado se a pessoa não foi encontrada no banco de dados global. Para essa verificação, empregou-se o processo

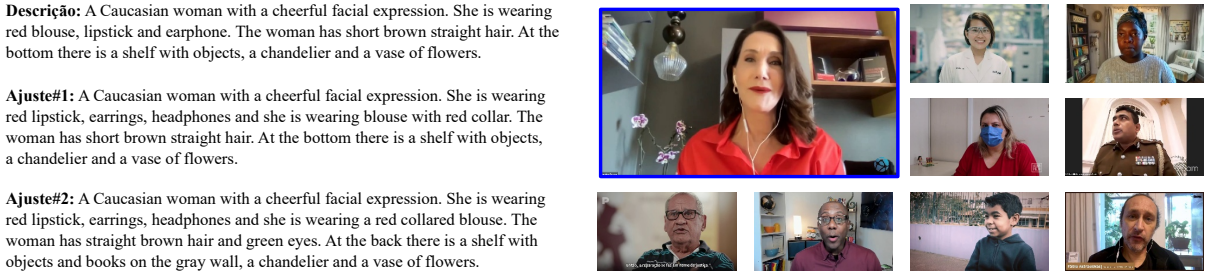


Figura 5.2: Amostras de conjuntos de dados mostrando a diversidade e representatividade dos dados, e o conjunto de rótulos da imagem destacados em azul. Fonte: Adaptado de Ferreira et al. (2023).

descrito na Seção 5.1.1 a respeito da coleta de dados.

Esta versão, ao evitar a inserção de quadros de pessoas já armazenadas no banco de dados global, resultou em um total de 5.689 imagens, processadas a partir dos mesmos 4.173 vídeos do conjunto de dados completo. Isso representa pouco mais de 50% das imagens presentes no conjunto original.

Conjunto de dados de pessoa única anotado

Para gerar esta versão do conjunto de dados, foram selecionadas aleatoriamente imagens do conjunto de dados de pessoa única e aplicou-se o processo descrito na Seção 5.1.2 para rotular as imagens. As imagens selecionadas foram distribuídas igualmente entre os três anotadores para a tarefa de anotação. Na tarefa de ajuste cruzado, cada anotador ajustou, de forma cega e aleatória, todas as descrições dos outros anotadores sem a ciência se era o primeiro ou segundo ajuste a ser realizado.

Esse processo levou cerca de cinco meses durante o ano de 2023. A versão do conjunto de dados disponibilizada[1] contém 684 imagens com três descrições, 190 imagens com duas descrições e 93 imagens com uma única descrição, totalizando 967 imagens anotadas. A Figura 5.2 apresenta exemplos dessas amostras, destacando a representatividade e diversidade dos dados.

5.2 Análise de Dados

Avaliou-se a qualidade, diversidade e representatividade dos conjuntos de dados criados usando métodos analíticos e modelos de Aprendizado de Máquina com o objetivo de mitigar vieses.

[1]Disponível publicamente em https://github.com/MaVILab-UFV/presenter-centric-dataset-SIBGRAPI-2023

5.2.1 Qualidade da imagem

Analisou-se as características de baixo nível das imagens, como brilho, contraste e nitidez, e características de alto nível, como o número de objetos. Para medir a nitidez da imagem, foi aplicado o método Tenengrad (Huang and Jing (2007)), que calcula a variância do gradiente resultante da versão em escala de cinza da imagem. Em relação à análise de contraste, foi calculado a razão entre o nível de cinza do *pixel* mais brilhante e o mais escuro. A medida de brilho retorna o valor médio entre todos os *pixels* da versão em cinza da imagem. Como uma característica de alto nível da imagem, contou-se o número de objetos detectados pela YOLOv3 (Redmon and Farhadi (2018)).

5.2.2 Diversidade e representatividade

Para avaliar a diversidade e representatividade dos dados, foram analisadas as sete características mencionadas abaixo. Para cada característica, empregou-se modelos de classificação pré-treinados e um modelo de VQA.

Gênero

Foram utilizados os modelos pré-treinados DeepFace (Serengil and Ozpinar (2020)), FairFace (Karkkainen and Joo (2021)), e DEX (Rothe et al. (2015)), que classificam o apresentador da imagem como sendo do gênero feminino ("*female*") ou masculino ("*male*"). Também foi aplicado o modelo de VQA conhecido como BLIP (Li et al. (2022)), com a pergunta "*What is the person's gender?*".

Faixa etária

De forma semelhante a realizada na Seção 4.1.1, as possíveis idades previstas também foram classificadas em cinco categorias de faixa etária recomendadas pela OMS (Ahmad et al. (2001)): ≤ 9 para criança ("*child*"), $10-19$ para jovem ("*young*"), $20-44$ para adulta ("*adult*"), $45-64$ para meia-idade ("*middle-aged*"), e ≥ 65 para idosa ("*elderly*"). Os mesmos modelos usados para a classificação de gênero foram aplicados para estimar a idade. Para o BLIP, foi fornecida a pergunta "*How old is the person?*".

Acessórios

Considerando a variedade de opções de acessórios, focou-se em nove classes específicas com base na frequência nas imagens de domínio (por exemplo, webinários) utilizadas e na união das saídas dos modelos aplicados: óculos/óculos

de sol ("*glasses/sunglasses*"), fones de ouvido ("*headphones/headsets*"), joias ("*jewelry*") (incluindo colares ("*necklaces*"), brincos ("*earrings*"), anéis ("*rings*"), etc.), chapéu ("*hat*"), lenço/cachecol ("*scarf*"), máscara ("*mask*"), microfone ("*microphone*"), turbante ("*turban*"), e uma categoria de outros ("*others*") para casos em que nenhuma das outras classes fossem contempladas. Utilizou-se o BLIP com a pergunta binária: "*Is this person wearing <ACCESSORY>?*", sendo *ACCESSORY* uma das classes mencionadas acima.

Além disso, o modelo pré-treinado DenseCap (Johnson et al. (2016)) foi utilizado para gerar legendas densas que descrevam cada objeto ou ROIs da imagem. Apenas as legendas com pontuações de confiança positivas foram mantidas. Para cada palavra presente nas legendas densas, verificou-se se a palavra está associada ao conceito de alguma das oito classes de acessórios (excluindo a classe "outros") por meio de sinônimos, hipônimos ou correspondências exatas dessas classes usando os *synsets* da WordNet (Miller (1995)), conforme previamente utilizado na filtragem proposta na Seção 4.1.1.

Cor da pele

Foram incluídas seis classes com base na interseção das saídas dos modelos aplicados: indiano ("*indian*"), oriente médio ("*middle eastern*"), negro ("*black*"), latino ("*latino*"), branco ("*white*"), e asiático ("*asian*"). Empregou-se os modelos DeepFace e FairFace para classificação. Para este último, considerou-se a saída sudeste asiático ("*southeast asian*)" ou leste asiático ("*east asian*") como asiático.

O BLIP novamente foi usado com a pergunta binária *"Is this person <ETHNICITY>?*", na ordem das classes acima mencionada, aceitando a primeira resposta positiva. Além disso, foi realizada uma pergunta aberta ao BLIP, "*What is the person's ethnicity?*". O motivo da pergunta é duplo: ter uma resposta nos casos em que as respostas para todas as seis perguntas binárias foram negativas e desempatar cenários em que as respostas às perguntas relacionadas as classes oriente médio e indiano foram positivas, uma vez que foi percebido que o BLIP retornava, frequentemente, positivo para ambas as classes. Logo, foi uma forma encontrada para reduzir o viés do modelo.

Emoção facial

Selecionou-se sete classes a partir da interseção das saídas dos modelos: bravo ("*angry*"), medo ("*fearful*"), neutro ("*neutral*"), triste ("*sad*"), enojado ("*disgusted*"), feliz ("*happy*"), e surpreso ("*surprised*"). Empregou-se além do DeepFace, o Emotion-detection (Goodfellow et al. (2013)), FER (Arriaga et al. (2019)), e PAZ (Arriaga et al. (2020)). Utilizou-se o BLIP com a pergunta "*Between angry, fear, neutral,*

sad, disgust, happy or surprise, what is the person's emotion?".

Cena

A classificação de cenas (ambiente ao redor do apresentador) abrangeu nove classes, selecionadas com base em sua relevância e frequência nas imagens utilizadas: sala de estar ("*living room*"), escritório ("*office*"), biblioteca ("*library*"), cozinha ("*kitchen*"), estúdio ("*studio*") – incluindo estúdio de música ("*music studio*"), estúdio de arte ("*art studio*"), estúdio de reunião ("*meeting studio*"), entre outros –, quarto ("*bedroom*"), áreas externas ("*outside*"), e hospital ("*hospital*"). A classe "outros" ("*others*") foi utilizada para imagens que não se encaixam em nenhuma dessas cenas específicas.

Para essa tarefa, foi utilizado o modelo pré-treinado PlacesCNN (Zhou et al. (2017)) com 365 classes de cenas e um modelo PlacesCNN que foi ajustado especificamente para as 19 classes de cenas mais relacionadas a vídeos de webinários. O BLIP foi usado com a pergunta "*What room is the person in?*".

Roupas

Como a maioria das imagens mostra apenas a parte superior do corpo das pessoas, os modelos de classificação de roupas podem apresentar confusões entre as classes; por exemplo, distinguir entre uma camisa e um vestido torna-se desafiador. Portanto, as classes de vestuário foram agrupadas em três classes principais: informal ("*informal*"), formal ("*formal*") e uniforme ("*uniform*").

Utilizou-se o DenseCap em conjunto com a WordNet para gerar legendas densas e identificar palavras associadas ao conceito de vestuário. Também aplicou-se o BLIP com a pergunta "*What is the person wearing?*" e a resposta foi categorizada manualmente entre as três classes definidas.

As respostas como terno ("*suit*"), gravata ("*tie*"), colete ("*vest*") e blazer ("*blazer*") foram designadas a classe formal, enquanto respostas incluindo uniforme, jaleco ("*lab coat*") e uniforme de enfermeira ("*nurse uniform*") foram classificadas como uniforme. Qualquer outra resposta foi associada a classe informal.

5.2.3 Detalhes da Implementação

No processo de criação do conjunto de dados, utilizou-se o MTCNN (Zhang et al. (2016)) para detecção facial e o pacote Python Face Recognition[2] baseado em um modelo pré-treinado de uma ResNet-29 como o descritor (codificador) facial. Para traduzir as descrições anotadas, foi usada a ferramenta do Deep-Translator[3] com o GoogleTranslator como tradutor integrado.

[2]Disponível publicamente em https://github.com/ageitgey/face_recognition
[3]Disponível publicamente em https://github.com/nidhaloff/deep-translator

Em relação às constantes, $\Delta = 120$, $\gamma = 5$, $\alpha = 0,11$, e $\beta = 0,6$ foram definidos empiricamente. Além disso, o classificador Haar Cascade Face usado na implementação original do DEX, Emotion-detection e PAZ foi substituído pelo modelo MTCNN para garantir a equidade.

Para a implementação da abordagem proposta, foi utilizada a linguagem de programação Python (versão 3.8.5). O *hardware* empregado foi um desktop padrão, equipado com 16 GB de RAM, processador Intel i7-7700 a 4,20 GHz, e uma GPU NVIDIA GeForce GTX 1070 com 8 GB de VRAM. O sistema operacional utilizado foi o Linux Ubuntu versão 22.04 LTS, versão de 64 bits.

5.3 Resultados e Discussão

A partir dos resultados relacionados à análise de qualidade de imagem apresentados na Figura 5.3, foi possível observar que, em relação ao brilho, apenas algumas imagens estavam excessivamente brilhantes ou escuras. Em termos de contraste, a maioria das imagens possui alto contraste, indicando uma ampla gama de tons. Na análise de nitidez, a maioria das imagens apresentou um bom nível de detalhes. A maioria das imagens retrata cenas que não estão densamente povoadas com objetos. Mais importante ainda, todas as três versões do conjunto de dados exibiram distribuições semelhantes nesses aspectos, sugerindo que as versões menores são amostras representativas das maiores.

Foi incluída a análise de diversidade e representatividade dos conjuntos de dados criados na Tabela 5.1, onde a coluna "cobertura" indica a porcentagem de dados com saída válida para cada método. A média amostral e o desvio padrão são calculados nas três versões dos conjuntos de dados, com o objetivo de avaliar a representatividade das versões menores em relação ao conjunto completo. Portanto, visou-se a um valor menor de desvio padrão.

Em relação à análise de gênero (Tabela 5.1a), todos os métodos exibiram consistentemente um padrão de média e desvio padrão, exceto o DeepFace, que apresentou um viés significativo em favor da classe "masculino". Ao analisar as

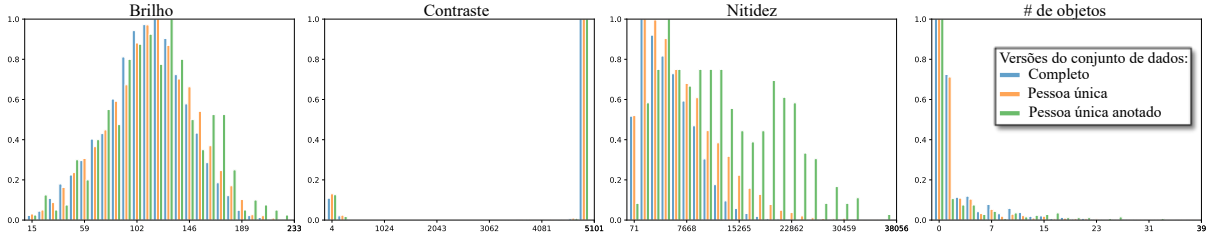


Figura 5.3: Histogramas normalizados para a análise de qualidade de imagem, representando características de baixo e alto nível das imagens. Fonte: Adaptado de Ferreira et al. (2023).

Tabela 5.1: Análise de diversidade e representatividade dos dados que compõem o conjunto de dados de acordo com os modelos de Aprendizado de Máquina usados.

Modelo	Feminino	Masculino	Cobertura
BLIP	$40,2 \pm 6,9$	$59,8 \pm 6,9$	$100,0 \pm 0,0$
DeepFace	$22,4 \pm 2,4$	$77,6 \pm 2,4$	$100,0 \pm 0,0$
DEX	$46,0 \pm 2,8$	$54,0 \pm 2,8$	$100,0 \pm 0,0$
FairFace	$46,0 \pm 5,5$	$54,0 \pm 5,5$	$100,0 \pm 0,0$

(a) Análise de gênero.

Modelo	Informal	Formal	Uniforme	Cobertura
BLIP	$83,9 \pm 1,5$	$15,7 \pm 1,3$	$0,2 \pm 0,2$	$100,0 \pm 0,0$
Densecap	$63,6 \pm 2,5$	$13,5 \pm 1,5$	$0,0 \pm 0,0$	$75,5 \pm 1,3$

(b) Análise do tipo de roupa.

Modelo	Criança	Jovem	Adulta	Meia-idade	Idosa	Cobertura
BLIP	$0,2 \pm 0,2$	$1,6 \pm 0,3$	$67,6 \pm 1,3$	$26,3 \pm 0,6$	$4,3 \pm 0,4$	$100,0 \pm 0,0$
DeepFace	$0,0 \pm 0,0$	$0,0 \pm 0,0$	$95,3 \pm 0,5$	$4,7 \pm 0,5$	$0,0 \pm 0,0$	$100,0 \pm 0,0$
DEX	$0,0 \pm 0,0$	$0,4 \pm 0,1$	$72,1 \pm 1,1$	$26,9 \pm 0,9$	$0,6 \pm 0,3$	$100,0 \pm 0,0$
FairFace	$0,7 \pm 0,0$	$5,8 \pm 0,9$	$90,6 \pm 1,2$	$2,4 \pm 0,5$	$0,4 \pm 0,2$	$100,0 \pm 0,0$

(c) Análise relacionada às faixas etárias definidas pela Organização Mundial da Saúde.

Modelo	Bravo	Medo	Neutro	Triste	Enojado	Feliz	Surpreso	Cobertura
BLIP	$2,8 \pm 1,4$	$0,0 \pm 0,0$	$0,0 \pm 0,0$	$67,0 \pm 6,0$	$0,0 \pm 0,0$	$30,1 \pm 7,3$	$0,0 \pm 0,0$	$100,0 \pm 0,0$
DeepFace	$9,7 \pm 0,2$	$14,6 \pm 0,4$	$34,3 \pm 1,0$	$23,3 \pm 2,5$	$0,9 \pm 0,3$	$14,4 \pm 3,8$	$2,8 \pm 0,2$	$100,0 \pm 0,0$
Emotion	$11,5 \pm 0,3$	$20,4 \pm 0,3$	$10,0 \pm 0,2$	$35,7 \pm 0,9$	$0,1 \pm 0,1$	$19,5 \pm 1,2$	$2,8 \pm 0,3$	$100,0 \pm 0,0$
FER	$16,0 \pm 1,9$	$6,5 \pm 0,6$	$42,0 \pm 2,2$	$18,0 \pm 0,5$	$0,0 \pm 0,0$	$9,3 \pm 1,8$	$8,2 \pm 0,5$	$100,0 \pm 0,0$
PAZ	$8,9 \pm 0,8$	$5,0 \pm 1,2$	$31,4 \pm 3,1$	$27,2 \pm 0,6$	$0,0 \pm 0,0$	$22,7 \pm 1,7$	$4,7 \pm 0,2$	$100,0 \pm 0,0$

(d) Análise da emoção facial do apresentador (inferida inteiramente com base em informações visuais)

Modelo	Asiático	Branco	Oriente médio	Indiano	Latino	Negro	Cobertura
BLIP	$1,3 \pm 0,1$	$33,5 \pm 1,6$	$12,9 \pm 1,6$	$5,0 \pm 1,5$	$30,2 \pm 4,2$	$16,8 \pm 8,2$	$99,7 \pm 0,1$
DeepFace	$11,2 \pm 1,1$	$50,2 \pm 2,1$	$11,5 \pm 2,7$	$1,6 \pm 0,5$	$14,8 \pm 3,3$	$10,8 \pm 3,6$	$100,0 \pm 0,0$
FairFace	$57,4 \pm 2,9$	$21,3 \pm 0,3$	$7,7 \pm 1,3$	$0,9 \pm 0,2$	$5,5 \pm 1,1$	$7,2 \pm 2,6$	$100,0 \pm 0,0$

(e) Análise da cor da pele do apresentador (inferida inteiramente com base em informações visuais).

Modelo	Óculos	F. de ouvido	Joias	Chapéu	Lenço	Máscara	Microfone	Turbante	Outros	Cobertura
BLIP	$26,2 \pm 1,1$	$17,5 \pm 3,4$	$23,2 \pm 3,2$	$2,0 \pm 1,1$	$2,1 \pm 0,5$	$0,5 \pm 0,4$	$7,7 \pm 0,8$	$1,0 \pm 0,2$	$19,8 \pm 2,0$	$100,0 \pm 0,0$
DenseCap	$50,1 \pm 2,9$	$1,2 \pm 2,1$	$4,2 \pm 1,5$	$6,2 \pm 2,5$	$0,5 \pm 0,1$	$0,0 \pm 0,0$	$0,0 \pm 0,0$	$0,0 \pm 0,0$	$37,9 \pm 1,7$	$100,0 \pm 0,0$

(f) Análise dos acessórios que estão sendo usados pelo apresentador.

Modelo	Sala de estar	Escritório	Biblioteca	Cozinha	Estúdio	Quarto	Hospital	Externo	Outros	Cobertura
BLIP	$42,3 \pm 3,3$	$29,0 \pm 0,7$	$8,5 \pm 4,0$	$3,3 \pm 0,8$	$0,1 \pm 0,1$	$2,3 \pm 0,3$	$0,1 \pm 0,0$	$2,2 \pm 1,5$	$12,3 \pm 4,4$	$100,0 \pm 0,0$
PlacesCNN	$0,1 \pm 0,1$	$9,0 \pm 2,0$	$0,0 \pm 0,0$	$0,1 \pm 0,1$	$2,5 \pm 0,5$	$0,0 \pm 0,0$	$0,3 \pm 0,1$	$0,0 \pm 0,0$	$88,1 \pm 1,7$	$100,0 \pm 0,0$
PlacesCNN-Ajustado	$0,2 \pm 0,1$	$1,5 \pm 0,4$	$3,3 \pm 1,1$	$3,8 \pm 1,3$	$5,7 \pm 1,9$	$73,0 \pm 1,4$	$0,1 \pm 0,1$	$0,5 \pm 0,4$	$12,0 \pm 1,0$	$100,0 \pm 0,0$

(g) Análise do ambiente ao redor do apresentador.

roupas (Tabela 5.1b), foi notado que todos os modelos reconheceram mais roupas da classe "informal" do que "formal", e apenas o BLIP reconheceu a classe "uniforme".

Sobre a idade (Tabela 5.1c), observou-se que o DeepFace teve uma forte tendência a identificar a classe "adulta", apesar da presença das faixas etárias "criança", "jovem" e "idosa" nas imagens. O FairFace é o único modelo com uma estratégia de treinamento para mitigar viés (Karkkainen and Joo (2021)), e seus resultados demonstram o compromisso do conjunto de dados proposto com a diversidade e representatividade dos dados.

A análise dos resultados apresentados na Tabela 5.1d revela que os modelos DeepFace, Emotion, FER, e PAZ se destacam na capacidade de reconhecer uma maior variedade de emoções faciais, ao contrário do BLIP, que se mostrou limitado, identificando principalmente as emoções "triste" e "feliz". Essas duas emoções, junto com "bravo", foram as únicas emoções classificadas por todos os cinco modelos comparados.

Conforme apresentado na Tabela 5.1e, a cor da pele não é um consenso entre os modelos. A maioria das saídas retornadas como "asiático" pelo FairFace são classificações incorretas devido a indivíduos com os olhos fechados. O DeepFace previu a classe "branco" para metade das imagens, enquanto as previsões restantes foram distribuídas uniformemente entre as outras classes, exceto "indiano". Diferente dos outros modelos, os resultados do BLIP foram mais uniformemente distribuídos, exceto para "asiático".

A respeito dos acessórios (Tabela 5.1f), o DenseCap apresenta um viés conhecido, identificando com maior frequência as classes como "óculos", "chapéu", "joias" e "lenço/cachecol", enquanto falha em reconhecer outras classes. Em contrapartida, o BLIP obteve uma distribuição mais equilibrada entre as classes, devido ao seu conjunto de treinamento mais representativo.

Observando a Tabela 5.1g, percebe-se que o PlacesCNN original exibiu um número excessivo de saídas para a classe "outros", devido à sua ampla gama de possibilidades de cenas. Embora o PlacesCNN ajustado tenha mostrado uma maior variação em comparação ao PlacesCNN original, ele também apresentou um forte viés, desta vez para a classe "quarto". Em contrapartida, o BLIP alcançou resultados mais diversificados, com uma maior frequência para as classes "sala de estar" e "escritório".

Por fim, destaca-se que apenas DenseCap e BLIP não alcançaram 100% de cobertura em uma tarefa cada, demonstrando a qualidade das imagens selecionadas. Os valores baixos de desvio padrão indicam que ambas as versões menores do conjunto de dados são representações confiáveis do conjunto de dados completo.

5.3.1 Limitações

Definir um limite de distância do cosseno e euclidiana para a construção do conjunto de dados provou ser desafiador. Enquanto um limite baixo inclui múltiplas imagens da mesma pessoa, um valor alto não cobre todos os indivíduos em um vídeo. Em raros casos, foram encontradas múltiplas imagens da mesma pessoa para o mesmo vídeo.

Embora as descrições de imagens anotadas por humanos geralmente sejam compostas por frases e estruturas gramaticais simples, o processo de tradução

automática do português para o inglês pode introduzir erros. Essas imperfeições ocorrem principalmente devido às particularidades de cada língua, acarretando desafios como a perda de contexto semântico, ambiguidade linguística, problemas de sintaxe e gramática, além de diferenças culturais. Portanto, uma análise do impacto da tradução automática e uma revisão humana tornam-se necessárias para garantir maior precisão e qualidade nas descrições traduzidas.

5.4 Desfechos Alcançados

Este capítulo propôs uma abordagem automática de coleta de dados, mais especificamente, quadros ou (*frames*) de listas de reproduções de vídeos do YouTube para criar um conjunto de dados de imagens com apresentadores em primeiro plano, como, webinários, *talk shows* e notícias, para apoiar a execução de tarefas de Visão Computacional. Com o objetivo de possibilitar acessibilidade para pessoas com CBV por meio de descrições imagens, também foi proposto um protocolo de anotação manual baseado em diretrizes e boas práticas recomendadas para esse público.

A viabilidade das soluções propostas foi demonstrada por meio da criação um conjunto de dados disponível em três versões: o conjunto de dados completo, uma versão com aparições únicas de pessoas e outra versão com imagens rotuladas das aparições únicas por meio da aplicação do protocolo de anotação proposto. Através de uma avaliação experimental extensiva, demonstrou-se a qualidade, diversidade e representatividade dos dados que compõem o conjunto de dados.

Capítulo 6

VIIDA e InViDe: Abordagens computacionais para geração e avaliação de parágrafos de imagens inclusivos para deficientes visuais

Foi introduzido no Capítulo 4 uma abordagem pioneira, inicialmente nomeado como Ours, para a geração automática de descrições textuais de imagens de webinários para pessoas com CBV, a fim de mitigar as barreiras de acessibilidade enfrentadas por esses indivíduos no uso de ferramentas de videoconferência. Nessa abordagem, foram combinados as vantagens das técnicas de legendagem densa com um conjunto de filtros para ajustar as descrições focadas no apresentador, métodos de Aprendizado de Máquina específicos e um PLM de sumarização abstrativa para possibilitar a produção de descrições adequadas, em formato de parágrafo, sobre as características relevantes do palestrante (por exemplo, características fisionômicas e expressões faciais) e seu entorno em uma cena de webinário para o grupo-alvo desta tese.

Embora a abordagem proposta anteriormente tenha contribuído para o avanço na geração de descrições de imagens para deficientes visuais, especialmente em relação ao conteúdo de imagens de webinários, ela apresentou uma série de limitações. Entre essas limitações, destacam-se a amplificação de erros propagados pelos modelos usados em série durante o procedimento, o desaparecimento de informações importantes durante o processo de geração de texto (também devido à aplicação de diferentes modelos) e os cenários experimentais adotados (como o conjunto de dados e as métricas de avaliação utilizadas). Essas limitações podem comprometer a qualidade dos parágrafos produzidos, prejudicando a compreensão do conteúdo das imagens por pessoas com CBV.

Diante das limitações apresentadas e em alinhamento com o sexto Objetivo Específico (OE-6), neste capítulo, estendeu-se a abordagem anterior aprimorando o *pipeline* computacional para a geração de descrições textuais do conteúdo de cenas de

webinários. Esse aprimoramento visou abordar as restrições e lacunas encontradas, melhorando assim a qualidade e a adequação dos parágrafos produzidos. Os problemas de propagação de erros e desaparecimento de informações são inerentes aos métodos de Aprendizado de Máquina e PLMs, ocorrendo especialmente quando aplicados a dados ou tarefas não vistos e aprendidos durante a fase de treinamento, resultando em imprecisões nas descrições geradas. Portanto, o *pipeline* foi reformulado, com foco na aquisição de características de interesse presentes nas imagens utilizando um modelo de Aprendizado Profundo de VQA. Adicionalmente, foi proposto um algoritmo determinístico que utiliza um grafo de cena para analisar e extrair informações das sentenças em linguagem natural. Isso substitui o uso de um PLM de sumarização, garantindo que as informações relevantes necessárias estejam presentes nas descrições das imagens para pessoas com deficiência visual.

Essas duas substituições permitiram minimizar o acúmulo e a propagação de erros, uma vez que, ao evitar o uso de múltiplos métodos de Aprendizado de Máquina e impedir que os erros desses métodos fossem utilizados como entrada para o PLM de sumarização, foi possível reduzir a propagação de erros. Dado que ainda não existem dados suficientes para treinar um modelo específico voltado para a tarefa de descrição de imagens para deficientes visuais, a abordagem proposta continua a utilizar modelos e técnicas existentes para alcançar os resultados desejados.

Além dessas alterações mencionadas, todos os experimentos realizados neste capítulo fizeram uso do conjunto de dados de webinários apresentado no Capítulo 5. Dessa forma, a limitação destacada no Capítulo 4, a respeito do uso de dados inadequados nos experimentos, foi abordada.

Neste capítulo, além da nova abordagem, também são apresentadas as comparações experimentais com diversos métodos concorrentes disponíveis, alguns dos quais são baseados em modelos de linguagem multimodais de última geração, por meio de várias métricas de avaliação. Os resultados quantitativos e qualitativos são discutidos e validam a extensão elaborada. Além disso, ligado ao OE-7, foi realizada uma análise do conteúdo "Não Seguro para o Trabalho" nos resultados produzidos, mostrando que a ocorrência de tais dados é insignificante, e considerações éticas sobre a abordagem são levantadas.

Por fim, para atingir o oitavo e último Objetivo Específico (OE-8) desta tese, também foi proposta uma nova métrica multicritério ajustada para avaliar a adequação da descrição textual de imagens para o público com deficiência visual, baseada em diretrizes de acessibilidade. O desenvolvimento da métrica de avaliação cobriu a última lacuna evidenciada no final do Capítulo 4.

6.1 Método VIIDA

A partir da abordagem estendida, foi desenvolvido um novo método, denominado VIIDA (do inglês, ***V**isually **I**mpaired **I**mage **D**escription **A**pproach*)[1]. Conforme apresentado na Figura 6.1, o VIIDA consiste em duas etapas principais, sendo elas: Extração de Informações Visuais e Produção Estruturada de Parágrafos. Dada uma imagem como entrada, a primeira etapa visa identificar e descrever ROIs em linguagem natural, enquanto a segunda etapa produz um parágrafo estruturado e coerente que representa o conteúdo visual semântico da imagem, com ênfase nos detalhes para pessoas com deficiência visual.

As etapas juntamente com seus módulos são apresentados a seguir e seus detalhes de implementação, como limiares, modelos e seus hiperparâmetros, podem ser conferidos posteriormente na Seção 6.3.2.

6.1.1 Extração de informações visuais

A primeira etapa, Extração de Informações Visuais, é dividida em duas abordagens: Na abordagem principal de geração de sentenças, primeiro, é fornecido uma imagem com um único apresentador em primeiro plano (por exemplo, de webinários, notícias, entrevistas, etc.) para um modelo VQA. Além da imagem, o modelo VQA requer perguntas para prever respostas (Antol et al. (2015)). Portanto, o modelo utiliza um conjunto de perguntas predefinidas manualmente elaboradas pelo autor desta tese. As perguntas foram elaboradas conforme a análise das diretrizes de acessibilidade recomendadas (ABNT (2016); Lewis (2018); Stangl et al. (2020)).

A partir das respostas do VQA, são extraídas informações semânticas (por exemplo, gênero, idade, vestimenta da pessoa, objetos e ambiente) presentes na imagem de entrada redimensionada, representando suas características relevantes. Foram aplicados procedimentos próprios para formular uma descrição em linguagem natural atribuindo uma ou mais saídas obtidas do modelo VQA. Ao final, um conjunto de sentenças que descrevem o conteúdo visual semântico da imagem foi construído.

Na abordagem complementar, a imagem é alimentada em um modelo de legendagem densa que gera legendas densas descrevendo ROIs adicionais na imagem. Para cada legenda gerada, realiza-se um processo de filtragem que consiste na análise semântica e gramatical e na confirmação do conteúdo da legenda gerada na imagem. Finalmente, as legendas densas restantes são adicionadas ao conjunto de sentenças da abordagem principal.

[1]Disponível publicamente em https://github.com/daniellf/VIIDA-and-InViDe

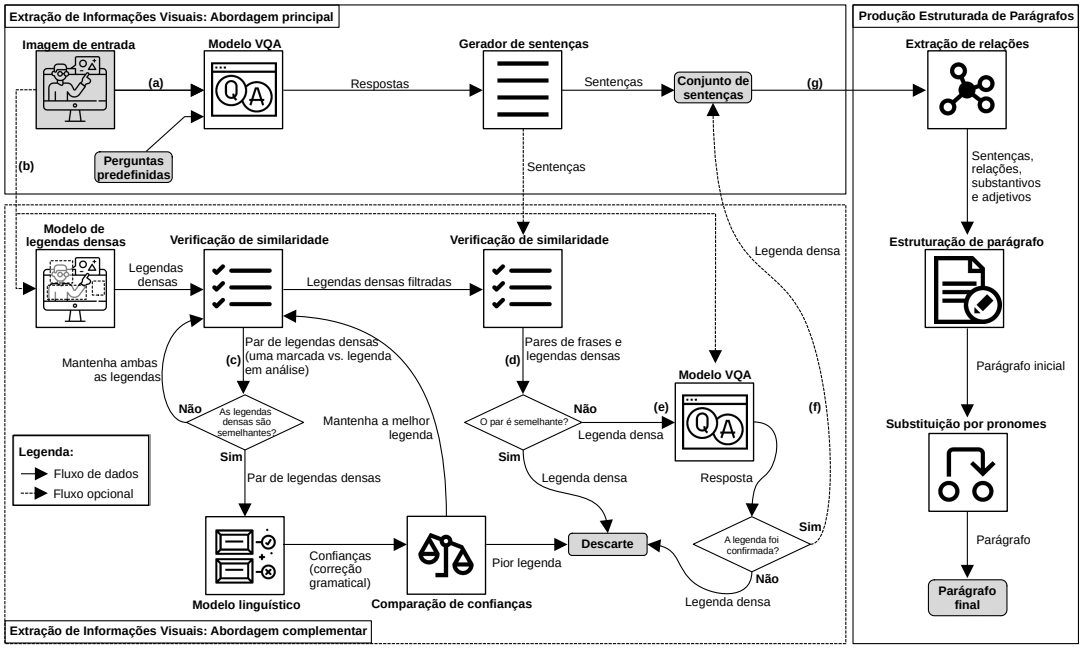


Figura 6.1: Visão geral do método gerador de parágrafos aprimorado. Dada uma imagem de webinário como entrada, um conjunto de perguntas predefinidas é usado para extrair informações relevantes da cena por meio da aplicação de um modelo de Resposta a Pergunta Visual (Visual Question Answering - VQA) para gerar sentenças textuais (a). Como opção, a imagem pode passar por outra abordagem, aumentando o conjunto de sentenças geradas anteriormente com descrições detalhadas (legendas densas) produzidas por um modelo de legendagem densa (b). Cada legenda densa criada passa por filtros de similaridade e adequação, mantendo apenas aquelas diferentes com a melhor correção gramatical, de acordo com um modelo linguístico específico (c). As legendas densas filtradas são então comparadas com as sentenças da abordagem principal novamente por meio de similaridade semântica (d), com as legendas densas dissimilares passando por uma confirmação adicional pelo modelo VQA (e). As legendas densas confirmadas são agregadas ao conjunto de sentenças da abordagem principal (f). Posteriormente, as relações gramaticais são extraídas das sentenças usando um grafo de cena, e após um processo de estruturação e ajuste, o parágrafo final é gerado como saída (g). Fonte: Material próprio.

Abordagem principal

Nesta seção, são fornecidas as explicações de cada módulo que compõe a abordagem principal da fase de Extração de Informações Visuais.

Perguntas e respostas. Na abordagem descrita no Capítulo 4, foram usados vários modelos de Aprendizado de Máquina para tarefas distintas (por exemplo, detecção de pessoas, legendagem densa, classificadores de idade, sentimento e cena) juntamente com seus respectivos limiares de confiança para criar uma descrição a respeito do contexto da imagem fornecida que mencionasse as características relevantes sobre o apresentador e seus arredores. No VIIDA, todos os modelos de

aprendizado foram substituídos por apenas um modelo VQA (Figura 6.1a). Essa substituição possibilitou a redução dos erros de previsão e a sua propagação. Em vez de trabalhar com vários modelos independentes que propagavam saídas por diferentes etapas do *pipeline* anterior, agora todas essas tarefas são concentradas em uma única etapa usando um modelo com maior precisão do que os modelos usados anteriormente.

Desta vez, juntamente com a imagem a ser descrita, é fornecido um conjunto de perguntas de interesse, como "*is the person a man or woman?*", "*how old is that person?*", "*is the person bald?*", entre outras, como entrada para o modelo VQA. Ao processar cada pergunta, esse modelo analisa a imagem e retorna uma resposta binária ("*yes/no*") ou aberta (por exemplo, "*woman*", "40", "*black*", "*short*", "*shirt*", etc.). Foi através dessas perguntas diretas que obteve-se as informações relevantes para a elaboração do contexto da imagem.

Em relação ao conjunto de perguntas de interesse, um total de 23 perguntas foram fornecidas ao modelo VQA, e algumas dessas perguntas dependiam da resposta da pergunta anterior, como visto na Tabela 6.1. Por exemplo, primeiro pergunta-se "*what accessory is the person wearing?*". Quando "*necklace*" é retornado como acessório, usou-se essa resposta na estrutura da próxima pergunta "*what color is the necklace?*". Das 23 perguntas, oito são processadas apenas conforme a verificação da resposta (como "*yes*") de outras perguntas, e três utilizam em sua estrutura as respostas obtidas anteriormente. Essa estratégia de utilizar um modelo VQA é aplicada para direcionar melhor as perguntas, ajudando a obter detalhes sobre um objeto específico ou a aparência da pessoa.

Gerador de sentenças. Em seguida, sentenças coerentes em linguagem natural são criadas a partir das saídas produzidas pelo modelo VQA para incluir as informações sobre gênero, cor da pele, faixa etária, emoção, olhos/cabelo/calvície/barba, vestuário, acessórios/maquiagem, objetos e cena extraídas das imagens. Sentenças coerentes referem-se a frases estruturadas, que são logicamente conectadas e que fluem de maneira natural dentro de um texto, por meio do uso adequado de pronomes, conectivos, e um encadeamento lógico das ideias, o que facilita a leitura e compreensão do texto como um todo.

As sentenças geradas seguem uma estrutura padrão, por exemplo: "*<DET> <SKIN_COLOR> <AGE> <NOUN> with <EMOTION> expression.*", "*The <NOUN> has <ADJECTIVE> hair.*", "*The <NOUN> is wearing <ADJECTIVE> <TOP_CLOTHES>.*", "*The <NOUN> is <PREP> <SCENE>, with <DET> <OBJECT> in the background.*", onde SKIN_COLOR, AGE, NOUN, EMOTION, ADJECTIVE, TOP_CLOTHES, SCENE e OBJECT são as saídas do modelo VQA, e DET e PREP são determinantes ("*a/an*") e preposições ("*at/in*"). Exemplos de

Tabela 6.1: Lista de perguntas de interesse que foram fornecidas ao modelo de VQA. As colunas "Opcional" e "Dependência" indicam, respectivamente, quais perguntas dependem da resposta de outra pergunta e qual é o número da pergunta de dependência. As palavras entre <> representam a saída gerada pelas perguntas anteriores ou dependem delas.

Nº	Pergunta	Opcional	Dependência
01	"*Is the person a man or a woman?*"	X	X
02	"*How old is the person?*"	X	X
03	"*What is the person's skin color?*"	X	X
04	"*Are the person's eyes open?*"	X	X
05	"*What color are the person's eyes?*"	✓	04
06	"*Is the person bald?*"	X	X
07	"*Is the person's hair short or long?*"	✓	06
08	"*What color is the person's hair?*"	✓	06
09	"*Does the person have a mustache?*"	X	X
10	"*What color is the mustache?*"	✓	09
11	"*Does the person have a beard?*"	X	X
12	"*What color is the beard?*"	✓	11
13	"*How is the beard?*"	✓	11
14	"*Is the person wearing makeup?*"	X	X
15	"*Is the person wearing any accessories?*"	X	X
16	"*What accessory is the person wearing?*"	✓	15
17	"*What color is the <ACCESSORY>?*"	✓	16
18	"*What is the person's facial expression?*"	X	X
19	"*What top clothing is the person wearing?*"	X	X
20	"*Does the <TOP CLOTHING> have more than one color?*"	X	19
21	"*What <COLOR/S> <IS/ARE> the <TOP CLOTHING>?*"	X	19, 20
22	"*What environment is the person in?*"	X	X
23	"*What objects are behind the person?*"	X	X

sentenças geradas: "*A brown woman with surprised expression.*", "*The woman has black long hair.*", "*The woman is wearing a white blouse.*" e "*The woman is in the office, with a plant in the background.*". Essas sentenças formam um conjunto e são concatenadas de maneira simples em um único texto.

Para tornar as sentenças criadas mais compreensíveis, com uma estrutura linguística adequada, são usados modelos de PLN e métodos de filtragem para, por exemplo, gerar corretamente palavras no plural e verificar o melhor determinante e preposição a serem usados com base nas saídas do modelo VQA. Em relação à escolha dos determinantes "*a*" e "*an*", é importante mencionar que um modelo foi usado para analisar a pronúncia fonética das palavras em inglês a fim de garantir que o artigo "*a*" seja usado antes de uma palavra com som de consoante e "*an*" antes de uma vogal.

Ainda no contexto do processamento, devido ao desafio de prever a idade correta de uma pessoa, os valores retornados pelo modelo VQA foram agregados pela faixa etária proposta pela OMS (Ahmad et al. (2001)) assim como feito anteriormente na Seção 4.1.1, isto é, em "*child*", "*young*", "*adult*", "*midlife*", e "*elderly*". Além disso,

embora as perguntas fornecidas ao modelo VQA sejam padronizadas, dependendo da resposta retornada, uma determinada sentença pode ou não ser criada. Por exemplo, só haverá uma sentença que descreva o cabelo da pessoa se ela não for careca. A mesma lógica é aplicada a barba, maquiagem, cor dos olhos, entre outros. Dessa forma, é garantido uma maior precisão na descrição das sentenças e evita-se a presença de vieses retornados pelo modelo VQA.

Ao final, as sentenças geradas são concatenadas de forma simples, só com a adição de espaço e vírgula entre elas, em um único texto. Posteriormente, este único texto serve como entrada para a etapa de produção de parágrafos ou pode ser agregado com a saída da abordagem complementar descrita a seguir.

Abordagem complementar

Nesta seção, os parágrafos seguintes descrevem detalhadamente cada um dos módulos da abordagem complementar da fase de Extração de Informações Visuais.

Modelo de legendas densas. Como complemento opcional à abordagem principal, foi empregado um modelo de legendagem densa (Figura 6.1b) para produzir descrições adicionais que podem ser incorporadas ao conjunto de sentenças entregues na abordagem principal do VIIDA. Cada legenda densa produzida é curta, independente e possui um valor de confiança juntamente com uma bbox para cada ROI.

Como as sentenças criadas usando as saídas do modelo VQA são pré-estruturadas, a inclusão de legendas densas geradas por um modelo externo, isto é, que não seja o modelo VQA, possibilita um maior enriquecimento semântico no conjunto de sentenças que é fornecido para a etapa de Produção de Parágrafo Estruturado. Essas legendas adicionais podem oferecer informações e detalhes não abordados ou identificados pelo modelo VQA, que é guiado pelas 23 perguntas predefinidas (Tabela 6.1).

Se, por um lado, essas perguntas trazem foco ao apresentador, diminuindo drasticamente a probabilidade de não descrever suas características importantes, por outro lado, o uso de legendas densas pode enriquecer o parágrafo final, acrescentando informações além do escopo limitado das 23 perguntas predefinidas. Assim, integrar esse complemento ao VIIDA aprimora tanto a sua capacidade de generalização quanto a produção de descrições mais precisas e de maior qualidade, facilitando sua utilidade prática.

Filtros e processamento de texto. Como descrito na Seção 4.1.1, o modelo de legendagem densa produz um grande número de legendas por imagem, o que

poderia resultar em descrições semelhantes ou fora de contexto. Portanto, essas legendas precisam ser filtradas.

Inicialmente, para reduzir o grande número de legendas densas, remove-se aquelas com um valor de confiança abaixo de um limiar pré-definido, denominado ϵ. Regiões detectadas pelo modelo de legendagem densa com alto valor de confiança têm mais probabilidade de corresponder a regiões reais de interesse (Johnson et al. (2016)). Portanto, ao remover as legendas com um valor de confiança baixo, aumenta-se a probabilidade de usar um conjunto de legendas densas de maior qualidade.

Deve-se mencionar que em cenários onde há apenas algumas legendas com um valor de confiança maior ou igual ao limite ϵ, foi permitido a inserção de legendas densas adicionais com um valor de confiança abaixo de ϵ até alcançar um total de pelo menos cinco legendas. Isso ocorre porque legendas densas com valores de confiança ligeiramente abaixo de ϵ não são necessariamente incorretas, então essa medida foi adotada para permitir uma possível margem de erro, se necessário.

O modelo de legendagem densa pode gerar legendas semelhantes devido a vários fatores, como limitações no contexto visual repetitivo (regiões diferentes com características semelhantes ou objetos parecidos), arquitetura do modelo (função de perda e capacidade do modelo em capturar variações), dados de treinamento (falta de diversidade nos dados e anotações de baixa qualidade), configurações de parâmetros de inferência e redundância nas informações da imagem (regiões diferentes, mas a informação visual é ambígua e não distinta suficientemente para gerar legendas diferentes) (Johnson et al. (2016)). Assim, para resolver o problema de legendas densas semelhantes, primeiro, converte-se todas as palavras presentes nas legendas em minúsculas, remove-se as pontuações e espaços em branco duplicados, e aplica-se um método de tokenização às palavras.

Após o uso das técnicas de pré-processamento de texto, uma representação vetorial de alta dimensão é criada para cada legenda usando a incorporação contextual de sentenças e, por fim, é calculado a correspondência entre todas as representações vetoriais das legendas por meio da similaridade de cosseno (Figura 6.1c). Uma ferramenta de PLN foi empregada tanto para o pré-processamento de texto e criação das incorporações quanto para cálculo da similaridade de cosseno.

As legendas densas com similaridade de cosseno menor que o limiar T_{text_sim} são mantidas (ou seja, são suficientemente distintas). Além disso, os pares de legendas densas com similaridade de cosseno maior ou igual a T_{text_sim} são fornecidos como entrada para um modelo de classificação de texto. Este modelo verifica qual das duas legendas semelhantes é mais clara e gramaticalmente correta, isto é, qual segue a norma padrão da língua inglesa. O modelo classifica a sentença correta como 1 e a sentença incorreta como 0. Dessa forma, a melhor legenda é mantida e a outra é

descartada. Em caso de empate, onde ambas as legendas são classificadas como 1 ou 0, os valores de confiança são usados como critério de desempate.

Após a realização do processo de filtragem, uma nova comparação semântica das legendas densas com as sentenças geradas pela abordagem principal é iniciada (Figura 6.1d). Para evitar a repetição de informações, mais uma vez, as sentenças são pré-processadas e uma representação vetorial de alta dimensionalidade é criada para calcular a similaridade de cosseno entre todas as descrições. Desta vez, ao serem comparadas com as sentenças produzidas pela abordagem principal, as legendas densas consideradas semelhantes, ou seja, com similaridade de cosseno maior ou igual a T_{text_sim}, são diretamente descartadas sem mais verificações.

As legendas densas com baixa similaridade, consideradas distintas semanticamente das sentenças da abordagem principal, são fornecidas ao modelo VQA para uma confirmação do conteúdo descrito (Figura 6.1e). Se o modelo retornar "*yes*" com um valor de confiança maior ou igual ao limiar T_{text_yes}, isso significa que o conteúdo presente na legenda densa foi confirmado e, portanto, ela é mantida; caso contrário, a legenda é descartada. Este processo de verificação visual possibilita remover legendas com ruídos, vieses, informações duvidosas e enganosas.

No final, as legendas densas confirmadas são adicionadas ao conjunto de sentenças da abordagem principal (Figura 6.1f). A Figura 6.2 apresenta um exemplo de execução da abordagem complementar descrita passo a passo.

6.1.2 Produção estruturada de parágrafos

Na etapa de Geração de Contexto da abordagem proposta no Capítulo 4 (Seção 4.1.2), a partir das legendas densas extraídas da imagem de entrada, recorreu-se a um PLM para a tarefa de sumarização abstrativa de texto, seguido pelas etapas de limpeza e estimativa de qualidade com o intuito de gerar um parágrafo descritivo para pessoas com CBV. Já para o método VIIDA, é proposto um algoritmo determinístico que utiliza um grafo de cena para produzir, por meio da extração de informação e identificação de relacionamentos semânticos no conjunto de sentenças originado na fase anterior de Extração de Informações Visuais, um parágrafo coerente e descritivo em linguagem natural que represente os elementos essenciais (entidades e relacionamentos) extraídos sobre o conteúdo da imagem.

Extração de relações. Com a ideia de gerar parágrafos mais concisos, coesos e eficientes em termos de comunicação, proporcionando uma leitura mais focada que enfatize as características relevantes da imagem, o primeiro passo do algoritmo é extrair e identificar tanto as relações entre substantivos (entidades) quanto seus respectivos adjetivos nas sentenças textuais. Essa extração é realizada para permitir o

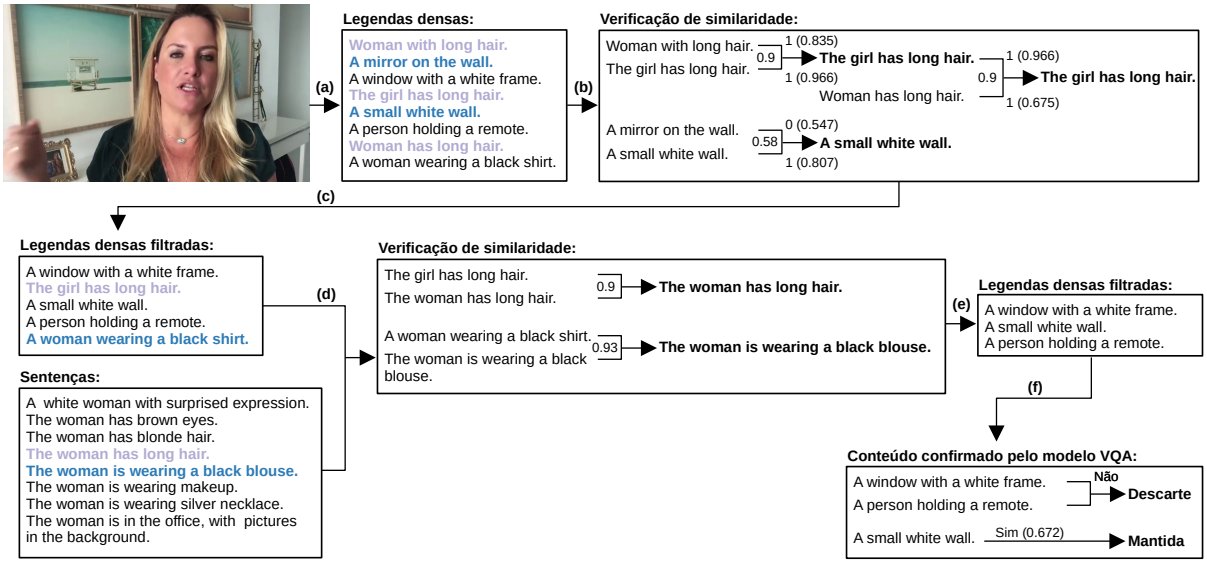


Figura 6.2: Exemplo de execução passo a passo da abordagem complementar da etapa de Extração de Informações Visuais. Dado uma imagem de webinário como entrada para um modelo de legendagem densa, o modelo irá produzir um grande conjunto de legendas. Inicialmente, essas legendas densas são filtradas com base em seu valor de confiança (a). Em seguida, essas legendas filtradas passam por um processo de verificação de similaridade semântica, e as legendas consideradas similares de acordo com um limiar definido são analisadas por um modelo de linguagem (b). As legendas densas consideradas gramaticalmente corretas ou com o maior valor de confiança são mantidas no conjunto de legendas filtradas (c). Posteriormente, as legendas filtradas são comparadas semanticamente com as sentenças produzidas pela abordagem principal da etapa de Extração de Informações Visuais (d). As legendas densas filtradas consideradas similares conforme um segundo limiar definido são descartadas, e as que mantiveram-se no conjunto (e) são fornecidas como entrada para um modelo VQA para a confirmação do conteúdo descrito na legenda densa gerada da imagem (f). Se o modelo retornar "*no*", a legenda é descartada; caso contrário, a confiança da resposta é analisada. As sentenças em roxo e azul demonstram as legendas densas consideradas semanticamente similares. Fonte: Material próprio.

agrupamento de informações no módulo seguinte. Para isso, emprega-se em conjunto ferramentas de PLN e de grafo de cena.

O objetivo é extrair relações semânticas entre duas ou mais entidades na mesma sentença por meio de uma análise linguística detalhada e modelagem semântica explícita das relações textuais. Esse processo de extração de informação utiliza um conjunto pré-existente de relações semânticas e regras bem definidas para retornar as características linguísticas essenciais. Isso é fundamental para o entendimento e a estruturação da informação proveniente de dados não-estruturados (Caseli and Nunes (2023)).

Um grafo de cena é uma estrutura de dados eficaz para fornecer descrições detalhadas em linguagem natural (Schuster et al. (2015); Li et al. (2017)). Para cada sentença processada, um grafo de cena é gerado por meio de análise de dependência,

uma técnica de análise linguística usada em PLN para descobrir relações gramaticais entre palavras de uma sentença. Cada nó do grafo representa uma entidade com modificadores (atributos), como determinantes ou adjetivos. As arestas do grafo denotam os relacionamentos entre as entidades (um sujeito e um objeto, também conhecido como "cauda") por meio de um verbo ou uma adposição (Wu et al. (2019)).

O algoritmo analisa se a quantidade de relações retornada pelo grafo de cena é maior do que zero. Se houver alguma relação, é verificado se a entidade da sentença é a "cabeça" da relação gramatical. Caso essa premissa seja verdadeira, as relações gramaticais são extraídas. A "cabeça" (ou "*head*") é o núcleo sintático que determina as propriedades gramaticais da sentença.

Nesta tese, foi considerado como "cabeça" qualquer substantivo com uma dependência igual a raiz (ou "*root*"), sujeito nominal (`nsubj`), ou frase nominal como modificador adverbial (`npadvmod`) (De Marneffe and Manning (2008)). As relações extraídas são armazenadas com a seguinte estrutura: <HEAD>_<RELATION>, por exemplo, "*man_has*", "*woman_wearing*", etc. Além disso, o algoritmo também identifica e extrai adjetivos e suas respectivas entidades.

Ao final deste módulo, as relações, os substantivos e os adjetivos são agregados, juntamente com as suas quantidades e associações, no o conjunto de sentenças da fase de Extração de Informações Visuais. A identificação dessas informações é necessária para que o algoritmo prossiga nos módulos seguintes.

Estruturação de parágrafo. Com a extração de relações gramaticais e das características linguísticas, o próximo passo do algoritmo é a estruturação dos parágrafos. Neste módulo, não foi mais utilizado um PLM para produzir um novo texto com base na entrada. Em vez disso, uma estratégia baseada no agrupamento das informações gramaticais foi utilizada para estruturar um parágrafo de modo a manter um sentido de unidade, focando apenas em um único ponto a ser discutido.

Agrupar as informações relacionadas à pessoa em uma única sentença no parágrafo, em vez de espalhá-las, evita repetições desnecessárias. Isso torna o parágrafo mais sucinto e natural de ler, pois as informações sobre a pessoa são apresentadas de forma mais compacta e organizada, facilitando para o usuário com deficiência visual entender a descrição do apresentador do webinário e seu ambiente. Esta estratégia de agrupar as informações sobre a pessoa para destacar as características relevantes (por exemplo, faixa etária e emoção) de uma só vez é uma melhoria em relação à abordagem proposta no Capítulo 4, que produzia sentenças espalhadas e sofria com o desaparecimento de informações.

Para estruturar o parágrafo, o algoritmo primeiro verifica, com base nas relações gramaticais e características linguísticas extraídas das sentenças textuais, se há uma relação na sentença atual a ser processada. Se não houver relação (por exemplo, "*The*

man's eyes are open.", "*A brown and white book.*", "*A white informational sign*", etc.), a sentença é adicionada ao corpo do parágrafo da forma que se encontra. Caso contrário, é verificado, entre as demais sentenças do conjunto (por exemplo, "*A midlife white man with surprised expression.*" e "*The man is wearing a blue shirt.*"), se existem mais relações iguais àquela identificada ("*man_has*"). Se não houver, a sentença atual (por exemplo, "*The man has brown eyes.*") também é adicionada ao parágrafo.

Em casos em que houver duas ou mais relações iguais (por exemplo, a relação "*man_has*" aparece nas seguintes sentenças: "*The man has brown eyes.*", "*The man has black hair.*", "*The man has short hair.*", e "*The man has black mustache.*"), o algoritmo realiza o processo de agrupamento das informações nas sentenças. Desta forma, em vez de repetir a descrição da relação ("*has*") em sentenças separadas, o algoritmo descreve a relação apenas uma vez, seguida de adjetivos, se houver, e de uma entidade que não seja "cabeça", isto é, um objeto ou "cauda" (por exemplo, "*eyes*", "*hair*" ou "*mustache*") mas que esteja associada à "cabeça" da relação gramatical tanto na sentença atual quanto nas subsequentes. Esse processo continua até que todas as "caudas" e suas características da mesma relação com um substantivo "cabeça" (por exemplo, "*man*") sejam agrupadas em uma única sentença. Por exemplo: "*The man has brown eyes, black short hair, and black mustache.*".

O algoritmo avança dessa maneira, processando cada sentença em relação às outras e estruturando o parágrafo a partir de cada uma delas, até que todas sejam analisadas e inseridas no texto referente ao parágrafo inicial.

Substituição por pronomes. Ainda visando produzir parágrafos mais coerentes, coesos e fáceis de entender, o algoritmo, após a estruturação do parágrafo, substitui os sujeitos por seus pronomes no texto produzido. O uso frequente de substantivos pode tornar o texto monótono e repetitivo. A introdução de pronomes, por meio da correferência, adiciona variedade linguística, tornando o texto mais fluido e melhorando sua coesão, especialmente ao se referir a objetos ou pessoas mencionadas anteriormente (Krause et al. (2017)).

Assim, como último passo, o algoritmo percorre todo o parágrafo inicial gerado anteriormente, analisando as sentenças *token* por *token* em busca de um substantivo. Se encontrar um substantivo com dependência igual a sujeito nominal (`nsubj`), ele é adequadamente substituído por seu pronome a partir da segunda ocorrência do substantivo, como "*he*" ou "*she*" (por exemplo, "***The man** has brown eyes and brown short hair.*" → "***He** has brown eyes and brown short hair.*"). Esse caso geralmente ocorre quando o substantivo está no início da sentença.

Para os casos em que o substantivo tem a dependência de objeto de uma preposição (`pobj`), ele é substituído por "*him*" ou "*her*" (por exemplo, "*Brown chair*

behind ***man***." → "*Brown chair behind* ***him***."). Por fim, o parágrafo final que descreve a imagem do webinário fornecida como entrada para o VIIDA é produzido. Um exemplo explicativo da etapa de Produção de Parágrafo Estruturado é mostrado na Figura 6.3.

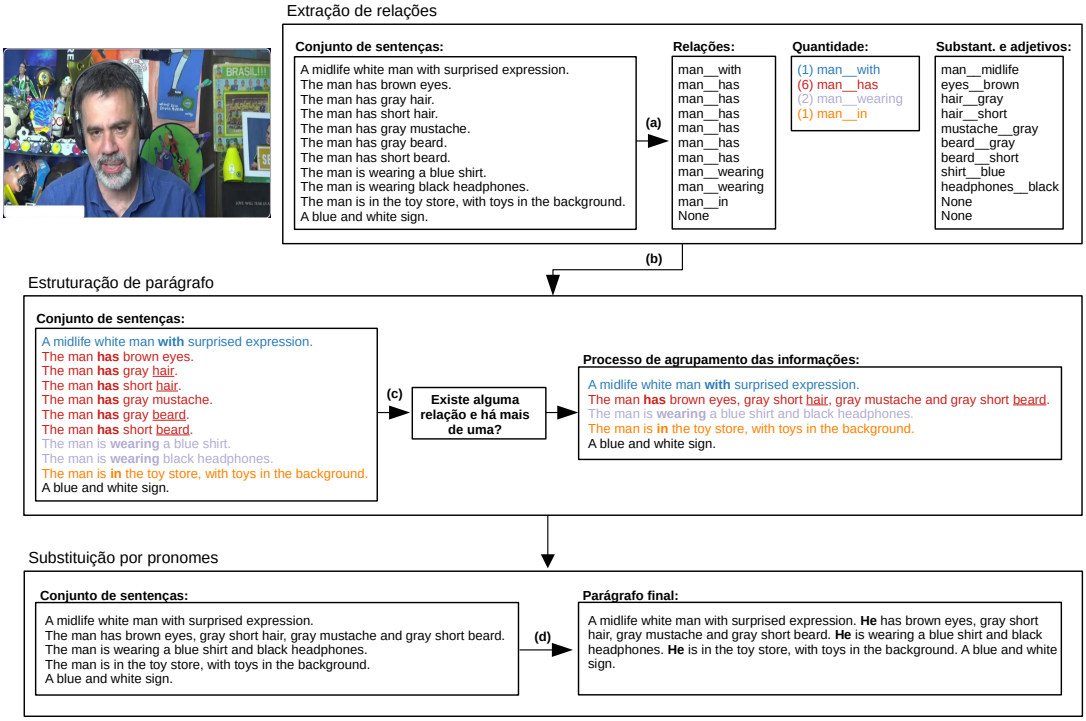


Figura 6.3: Exemplo de execução da etapa de Produção de Parágrafo Estruturado passo a passo. Para cada sentença do conjunto fornecido ao término da etapa de Extração de Informações Visuais, realiza-se uma extração de informação por meio da análise detalhada de dependência linguística utilizando ferramentas de PLN e grafos de cena (a). Ao final, são extraídas as relações gramaticais entre os substantivos e suas características. Em seguida, inicia-se o processo de estruturação do parágrafo (b), no qual é verificado se as sentenças analisadas possuem uma relação gramatical comum e se esta aparece mais de uma vez no conjunto de sentenças (c). Caso a relação não exista ou seja única no conjunto, a sentença é adicionada ao corpo do parágrafo da forma como está (por exemplo, as sentenças nas cores azul, laranja e preta na caixa "Processo de agrupamento das informações"). Caso contrário, o processo de agrupamento de informações é acionado, estruturando em uma única sentença todas as sentenças do conjunto com a mesma relação gramatical, respeitando a ordem de adjetivos e substantivos (por exemplo, as sentenças em roxo e vermelho, respectivamente, na caixa "Processo de agrupamento das informações"). Esse processo se repete até que todas as relações tenham sido analisadas. Por fim, os substantivos são substituídos por seus respectivos pronomes a partir da segunda ocorrência. (d). As cores das sentenças correspondem à relação gramatical extraída, as palavras em negrito representam os relacionamentos, e as palavras sublinhadas são os substantivos que não são a "cabeça" da sentença e que possuem mais de um adjetivo associado. Fonte: Material próprio.

6.1.3 Análise de complexidade

Analisando-se as principais fases do método VIIDA, em termos de esforço computacional, o mesmo percorre perguntas (de quantidade fixa) fornecidas ao modelo VQA, elabora sentenças com base nas respostas retornadas, estrutura um parágrafo conforme essas sentenças e realiza substituições por pronomes no texto. Além disso, o VIIDA pode utilizar legendas densas geradas para enriquecer o parágrafo produzido. Embora o método proposto apresente várias etapas, estas possuem custos computacionais reduzidos, já que o número de sentenças e palavras processadas é baixo, ou seja, não há uma questão crítica concernente ao crescimento da entrada. De toda forma, vários módulos têm complexidade linear (por exemplo, laços simples para analisar uma sequência de sentenças) e poucos apresentam complexidade quadrática; por exemplo, quando é verificada a similaridade semântica entre legendas densas, dois laços aninhados são utilizados para verificar similaridade par a par entre as legendas, resultando em uma complexidade $O(l^2)$, onde l é o número de legendas densas. Destaca-se mais uma vez o tamanho reduzido das entradas. No exemplo dado, o número de legendas gerado para uma imagem na prática é pouco significativo em termos de análise de complexidade.

Como o VIIDA é composto por diversos módulos que interagem com ferramentas e modelos de Visão Computacional e PLN, a complexidade final do método aqui proposto, no pior caso, será determinada pela função de maior ordem de crescimento entre as complexidades dos módulos proprietários que compõem o VIIDA (neste caso, dominado por uma função quadrática) e as complexidades associadas aos modelos e ferramentas externos utilizados. Em outras palavras, a complexidade do método VIIDA como um todo dependerá não apenas da implementação interna de seus módulos, mas também da eficiência dos modelos e ferramentas de terceiros, cujas complexidades podem variar de acordo com suas arquiteturas e características e são, seguramente, dominantes com relação aos módulos implementados neste trabalho. Por exemplo, o modelo VQA baseado em ViT pode ser computacionalmente caro, especialmente ao processar imagens de alta resolução, já que calcular a auto-atenção (quadrática) sobre os *patches* da imagem pode gerar altos custos de memória e processamento. Portanto, a complexidade final do VIIDA será determinada pela função que apresentar o comportamento assintótico mais significativo dentre aquelas relacionadas aos modelos e ferramentas externos empregados.

É importante destacar que as documentações oficiais dos modelos e ferramentas externos incorporados no VIIDA não abordam suas complexidades. Portanto, seria necessário depurar o código desses modelos para prover uma análise de complexidade mais detalhada, o que foge ao escopo desta tese.

6.2 Métrica InViDe

A construção de métricas de avaliação automática que alinham a avaliação humana de forma mais próxima com a qualidade dos textos produzidos por máquinas a partir de imagens sempre representou um grande desafio para os campos de PLN e Visão Computacional. Embora este seja um problema importante para a comunidade científica, o desenvolvimento de tais métricas para contextos específicos, como descrições de imagens para pessoas com CBV, enfrenta uma completa falta de iniciativas de pesquisa (Gurari et al. (2020); Stangl et al. (2020)).

Confrontados com o gargalo crítico tanto da inadequação dos métodos existentes quanto da ausência de métricas de avaliação específicas que priorizam as informações necessárias para as pessoas com deficiência visual em descrições de imagens, neste capítulo também foi introduzido um novo *framework* para avaliar automaticamente a adequação da descrição gerada para esse público-alvo: a métrica InViDe[2] (do inglês, ***In****clusive* ***Vi****sual* ***De****scription*).

Dessa forma, a InViDe foi desenvolvida com o objetivo de avaliar a adequação de descrições textuais geradas a partir de imagens, enfrentando os desafios relacionados à inadequação e à ausência de métodos específicos. A métrica proposta tem como alvo pessoas com CBV e foi fundamentada em diretrizes e recomendações apropriadas (ABNT (2016); Lewis (2018); Stangl et al. (2020)). Seu procedimento concentra-se em realizar uma análise multicritério com base em cinco características: características linguísticas, cobertura dos elementos-chave, descrição do conteúdo sobre o apresentador, descrição do conteúdo sobre o ambiente do apresentador e comprimento do texto produzido.

Para avaliar a adequação de um texto candidato (por exemplo, um parágrafo), é necessário um texto de referência (como uma anotação humana). Em seguida, são calculadas pontuações para cada uma das cinco características da análise multicritério mencionadas anteriormente, dentro do intervalo $[0,1]$. A pontuação final da InViDe é determinada pela média dessas cinco pontuações, sendo que, quanto mais próxima de 1, melhor. Para calcular a pontuação da InViDe no nível do corpus (um conjunto de parágrafos resultante da análise de um conjunto de imagens), calcula-se a média das pontuações de todos os textos candidatos.

Em virtude das características de interpretabilidade e transparência da InViDe, é possível identificar os pontos em que os modelos ou métodos de geração de texto são inconsistentes em termos de adequação para pessoas com deficiência visual. Além disso, por ser uma métrica flexível, ela pode ser personalizada para diferentes estudos. Por exemplo, pode-se calcular a pontuação de maneira diferente (usando pesos variados para a média das pontuações individuais), o que permite abordar

[2]Disponível publicamente em https://github.com/daniellf/VIIDA-and-InViDe

fraquezas e fazer melhorias de acordo com o contexto específico de seu uso. Destaca-se também que a versão atual da métrica não requer o uso de uma GPU.

6.2.1 Análise multicritério

Na InViDe, foi proposta uma análise multicritério com o objetivo de analisar características de interesse que aprimoram a compreensão do contexto visual em imagens para pessoas com CBV. Essa análise é baseada principalmente em diretrizes da Associação Brasileira de Normas Técnicas (ABNT (2016)), boas práticas recomendadas pela *Perkins School for the Blind* (Lewis (2018)), e achados relatados no estudo de Stangl et al. (2020).

Embora esses documentos sejam de países diferentes, a relação comum entre eles está na promoção da acessibilidade e inclusão para pessoas com CBV. Eles oferecem orientações baseadas na experiência do usuário sobre como criar descrições de imagens de forma eficaz, permitindo uma compreensão plena do conteúdo visual para este público específico. As características de interesse levantadas nos documentos e suas análises são apresentadas a seguir.

Características linguísticas

Conforme mencionado na Seção 4.1.2, os parágrafos mais informativos geralmente apresentam características linguísticas que tornam o texto mais fácil de ler e interpretar, como múltiplas sentenças conectadas por conjunções coordenativas, uso de correferência e uma frequência razoável de verbos e pronomes (Krause et al. (2017)). Além disso, os parágrafos frequentemente contêm informações detalhadas sobre as propriedades e relações de objetos salientes devido às suas descrições textuais mais longas em comparação com legendas em nível de sentença (Krause et al. (2017)).

Portanto, dado um texto candidato tokenizado (C), a análise inicia-se calculando a frequência de todas as características linguísticas dentro do texto. Essas características dizem respeito à marcação POS associada a cada elemento da sentença. As seis *tags* (classes gramaticais) de interesse a serem coletadas foram definidas como: substantivos, verbos, adjetivos, advérbios, pronomes e conjunções coordenativas.

A proporção da frequência de cada uma das seis *tags* de interesse mencionadas anteriormente é calculada em relação à soma das frequências de todos os *tokens* identificados no texto. Posteriormente, as proporções das *tags* são somadas, resultando em um valor normalizado (C_{ling}) em $[0, 1]$. Esse processo é formulado da seguinte forma:

$$C_{ling}(C) = \sum_{i=1}^{nt} \frac{f_i}{\|T\|}, \tag{6.1}$$

onde nt é o número de *tags* de interesse (nesse caso, seis), f_i é a frequência total de cada uma das seis *tags* de interesse e $\|T\|$ é a soma da quantidade de todas as *tags* no texto candidato (por exemplo, "substantivo": 15, "determinante": 12, "adjetivo": 10, logo, $\|T\| = 37$).

Parágrafos escritos por humanos foram utilizados como referência para avaliar a adequação do texto candidato. Para realizar essa avaliação, o texto de referência (R) passa pelo mesmo processo mencionado acima, resultando em outro valor normalizado (R_{ling}). Então, uma primeiro pontuação considerando esse critério de características linguísticas (CL) é obtida por:

$$CL = 1 - |R_{ling} - C_{ling}|. \tag{6.2}$$

Valores mais próximos de 1 indicam alta similaridade, isto é, uma concordância, entre ambos os textos (candidato e referência) em relação à proporção de *tags* utilizadas.

Cobertura de elementos-chave

Este critério visa verificar se a descrição textual da imagem inclui uma quantidade mínima de conteúdo relevante. Nesta tese, fundamentada pelas diretrizes da ABNT (2016), foram definidas seis categorias diferentes de elementos-chave como conteúdos relevantes da descrição, a saber: corpo, sentimento, acessório, vestuário, local e arredores. As quatro primeiras categorias estão associadas a elementos relacionados à pessoa, enquanto as duas categorias restantes referem-se à cena geral e aos elementos ao redor do apresentador.

Foi atribuído manualmente a cada categoria um ou mais conjuntos de sinônimos cognitivos (*synsets*) da WordNet (Miller (1995)) para cobrir uma ampla gama de elementos, como partes do corpo, expressões faciais/sentimentos, joias, cosméticos, instrumentos ópticos, roupas, cômodos, locais de negócios, entre outros.

Após o pré-processamento, a remoção de *stop words*[3] e a tokenização do texto candidato, o procedimento verifica se cada *token* identificado como substantivo ou nome próprio (`PROPN`) está associado aos conceitos de uma das seis categorias. Isso é feito analisando se o *token* pertence ao conjunto de sinônimos ou hipônimos das categorias (Krishna et al. (2017a)). Se a associação for confirmada, a categoria identificada é contabilizada. É importante observar que, se a mesma categoria for

[3]São um conjunto de palavras comumente usadas em um idioma que não carregam muitas informações úteis e são frequentemente consideradas irrelevantes para análise em PLN.

identificada em outros *tokens*, ela é contada apenas uma vez. O objetivo é determinar a cobertura, ou seja, identificar quais das seis categorias relevantes estão presentes no texto. Além disso, como alguns elementos podem ter vários *synsets*, a análise foi restrita aos primeiros cinco *synsets*, se disponíveis.

O valor retornado ao final desta análise também é normalizado para o intervalo $[0, 1]$, dividindo a contagem final por seis (o número de categorias). Diferente do critério anterior e dos subsequentes, este é livre de referência. Espera-se que uma descrição de imagem adequada para deficientes visuais inclua pelo menos uma palavra associada a cada categoria, o que significa que o valor retornado por este critério deve ser 1.

Descrição do conteúdo sobre o apresentador e seu entorno

A quantidade de conteúdo e o nível de detalhes desejados em uma descrição podem variar dependendo do contexto da imagem digital. Para imagens em contextos semelhantes a webinários, como em websites de notícias, redes sociais e websites de relacionamento, onde as pessoas são o foco central, os usuários com deficiência visual desejam ter o máximo de conteúdo disponível, com foco nas características identificáveis da pessoa e em detalhes importantes sobre seu entorno (Stangl et al. (2020)).

Nesse sentido, conforme apresentado na Seção 5.1.2, as diretrizes orientam a criação de descrições que incluam informações sobre as características das pessoas, seus arredores e os objetos na imagem (ABNT (2016)). Em relação à caracterização das pessoas, recomenda-se mencionar sua aparência física e visual, como gênero, faixa etária, cor da pele, cabelo e olhos, expressões faciais e outras características notáveis, bem como as roupas e acessórios que estão usando.

Para avaliar se a descrição contém a quantidade de conteúdo e detalhes adequados, este critério da InViDe foi dividido em duas partes. A primeira parte concentra-se no conteúdo associado à pessoa, pois geralmente possui mais características de interesse, sendo o foco principal da imagem. Assim, foram consideradas as mesmas quatro categorias usadas no critério de cobertura de elementos-chave: corpo, sentimento, acessório e vestuário. A segunda parte avalia o conteúdo relacionado à cena e aos objetos ao redor, abrangendo as categorias de local e arredores.

Inicialmente, para a análise de cada categoria, o texto candidato passa por um pré-processamento (incluindo o processo de lematização), durante o qual as *stop words* são desconsideradas. Após a tokenização, os *synsets* da WordNet são novamente empregados para associar os *tokens* de substantivos aos grupos de categorias sobre a pessoa ou a cena. Posteriormente, os *tokens* de adjetivos ou aqueles com dependência igual a modificador adjetival (`amod`) ou composto (*compound*) associados aos *tokens* de

substantivos também são coletados.

Com os substantivos e suas respectivas características identificadas, é calculada a razão entre a contagem dos *tokens* de interesse coletados (ou seja, aqueles associados à pessoa ou à cena) e o número total de *tokens* no texto candidato. Como resultado, é retornado um valor entre $[0, 1]$, indicando a proporção do conteúdo de interesse em relação a todos os elementos pré-processados do texto. Esse processo ocorre separadamente para ambos os grupos de categorias, resultando em dois valores gerados pela InViDe, a saber, C_{pes} e C_{cen}.

Similarmente ao critério de características linguísticas, um parágrafo escrito por humanos é usado como referência para avaliar a correspondência do texto candidato. Dessa forma, o texto de referência também passa pelo processo de identificação do conteúdo referente ao mesmo grupo de categorias, resultando em dois valores adicionais, R_{pes} e R_{cen}. Consequentemente, o valor final de cada um desses dois grupos (DC_{pes} e DC_{cen}) a respeito da descrição dos conteúdos de interesse é obtido por:

$$DC_{gc} = 1 - |R_{gc} - C_{gc}|, \tag{6.3}$$

onde gc representa o grupo de categorias analisado (*pes* e *cen*). Valores mais próximos de 1 indicam uma maior similaridade entre os textos (candidato e referência) em termos da proporção de conteúdo sobre a pessoa ou a cena.

Comprimento do texto

De acordo com a *Perkins School for the Blind*, recomenda-se que uma descrição de imagem para pessoas CVB tenha aproximadamente 280 caracteres (Lewis (2018)). No entanto, não foram encontradas recomendações específicas sobre o comprimento mínimo para essas descrições. Portanto, estabeleceu-se o valor de 125 caracteres como o limite inferior, alinhado com o comprimento sugerido para textos alternativos (do inglês, "*alt text*"), que fornecem uma compreensão breve da imagem (Lewis (2018)). Considerando que descrições de imagens geralmente oferecem informações mais detalhadas do que textos alternativos, os limites mínimo (L_{Min}) e máximo (L_{Max}) para descrições destinadas a pessoas com deficiência visual foram definidos como 125 e 280 caracteres, respectivamente.

Assim, como mostrado na Figura 6.4, ao processar um texto candidato, a recompensa (RP) de 1 é aplicada se o comprimento do texto, representado pelo número de caracteres (λ), estiver dentro da faixa definida por L_{Min} e L_{Max}. Caso contrário, a recompensa é reduzida para valores de comprimento fora dessa faixa. Essa redução foi parametrizada usando valores arbitrários α e β.

A redução ou decaimento (D) para comprimentos menores que L_{Min} é linear até

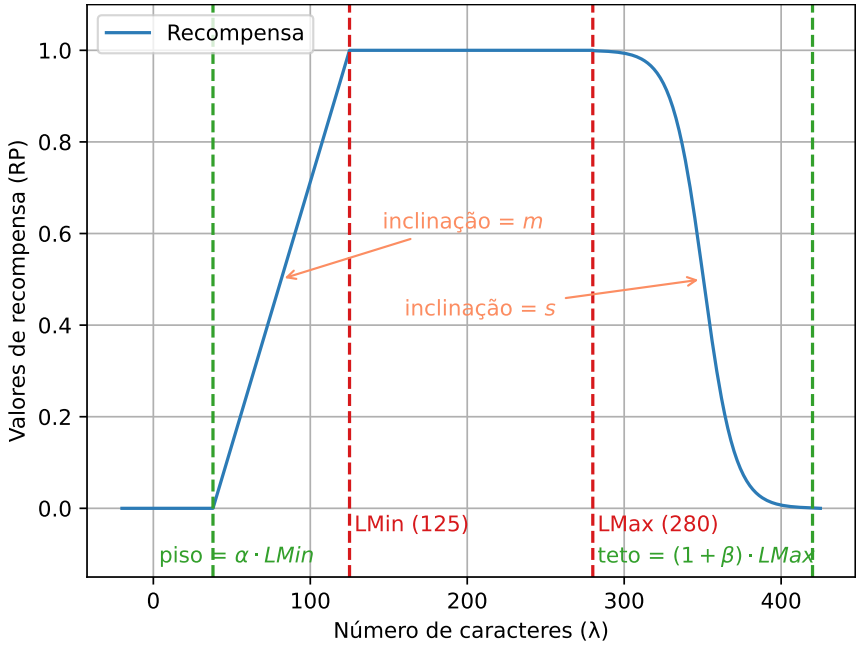


Figura 6.4: Processo de recompensa aplicado em relação ao comprimento do texto candidato. Fonte: Material próprio.

alcançar um comprimento de $\lambda = 38$ (*piso*), ponto em que a recompensa atinge seu valor mínimo de zero. Por outro lado, a redução para comprimentos maiores que L_{Max} segue uma função inversa da sigmoide até $\lambda = 420$ (*teto*), onde a recompensa também atinge seu valor mínimo.

Ambas as reduções (linear e sigmoidal) são calculadas usando os valores de inclinação m e s, respectivamente. O processo pode ser formulado da seguinte forma:

$$piso = \alpha \cdot L_{Min} \tag{6.4}$$

$$m = \frac{-1}{L_{Min} - piso} \tag{6.5}$$

$$D_{Linear} = m(\lambda - L_{Min}) \tag{6.6}$$

$$RP_{Linear} = 1 - D_{Linear} \tag{6.7}$$

$$teto = L_{Max} \cdot (\beta + 1) \tag{6.8}$$

$$\lambda' = \lambda - (\frac{teto - L_{Max}}{2}) - L_{Max} \tag{6.9}$$

$$D_{InvSig} = \frac{\tanh(s \cdot \lambda' - \frac{s}{2})}{2} + 0.5 \tag{6.10}$$

$$RP_{InvSig} = 1 - D_{InvSig} \tag{6.11}$$

Em relação à escolha de duas funções distintas para o decaimento, a abordagem linear foi selecionada para textos curtos devido à baixa probabilidade de fornecerem detalhes visuais suficientes para indivíduos com CBV. Dessa forma, a redução da recompensa ocorre linearmente — e de forma mais acentuada nas proximidades dos limites, em comparação com a função inversa da sigmoide — à medida que o número de caracteres se afasta do comprimento mínimo de 125 em direção a zero.

Para o decaimento sigmoidal aplicado aos textos com comprimento acima de 280 caracteres, a taxa de redução varia dependendo da inclinação da curva. Nas proximidades do limite superior, a penalidade é baixa, aumentando gradualmente em direção a valores mais altos. A decisão de aplicar uma penalidade mais suave a textos mais longos baseia-se na crença de que, embora esses textos possam exceder o comprimento recomendado, é mais provável que contenham detalhes importantes sobre os conteúdos das imagens em comparação com textos mais curtos.

Por fim, optou-se por construir um procedimento totalmente parametrizado principalmente pela flexibilidade que oferece. Essa abordagem permite o reuso e a adaptação da métrica em futuros experimentos, possibilitando a avaliação do impacto de diferentes configurações de parâmetros no comportamento da InViDe. Além disso, a parametrização torna a métrica adequada para distintos contextos, permitindo a utilização de valores de parâmetros ajustados às necessidades específicas, como diferentes limiares, parâmetros ou funções de decaimento.

6.3 Experimentos

Nesta seção, são apresentados os experimentos realizados para avaliar o processo de geração de parágrafos para pessoas com CBV. Isso inclui o detalhamento do conjunto de dados utilizado, as configurações dos parâmetros, os concorrentes e as métricas escolhidas para a comparação quantitativa.

6.3.1 Conjunto de dados

Quando se trata de estudos relacionados à tarefa de parágrafo de imagem, o conjunto de dados SIP mencionado na Seção 4.2.1 é uma referência importante (Krause et al. (2017)). Este conjunto de dados compreende 19.551 imagens, cada uma acompanhada por um único parágrafo anotado por seres humanos, cobrindo uma

ampla gama de conteúdos, incluindo animais, pessoas, esportes, paisagens e mais. No entanto, como o SIP não foi especificamente projetado para contextos de webinários ou adaptado para atender a públicos com deficiência visual, ele tornou-se inadequado para o prosseguimento deste trabalho.

No Capítulo 4, essa questão foi suavizada adaptando esse conjunto de dados. Especificamente, focou-se na seleção de imagens relevantes para o domínio de interesse desta tese, como aquelas que apresentam pessoas proeminentemente no primeiro plano, ou seja, imagens semelhantes às de apresentadores de seminários. Além disso, os parágrafos anotados por humanos foram modificados por meio da aplicação de filtros, mantendo apenas as sentenças dos parágrafos associadas a indivíduos (focando nos apresentadores). Como resultado, o SIP adaptado consistiu em 2.147 imagens e parágrafos adaptados.

Embora tenham sido feitos esforços para adaptar o SIP para melhor se adequar ao contexto desta tese, ele ainda apresentou limitações relacionadas ao domínio dos dados e à adequação dos parágrafos voltados para deficientes visuais (Gurari et al. (2018)). Em virtude disso, foi empregado o conjunto de dados mais recente proposto no Capítulo 5, que apresenta imagens de apresentadores em primeiro plano, tipicamente encontradas em webinários, *talk shows*, notícias, etc.

O conjunto de dados de webinários oferece três versões, sendo selecionada a versão contendo imagens rotuladas. Essa versão é composta por 684 imagens com dimensões de 1280×720 *pixels*, cada uma com três descrições anotadas por seres humanos: a descrição original e duas descrições ajustadas. As descrições ajustadas são possíveis alterações feitas por um anotador humano após o processo de revisão da imagem juntamente com a sua descrição anteriormente elaborada por um dos outros dois anotadores. No entanto, os experimentos detalhados neste capítulo foram conduzidos usando apenas a última descrição ajustada para cada imagem.

Este conjunto de dados foi desenvolvido seguindo o protocolo de anotação baseado em instruções específicas (por exemplo, normas e boas práticas recomendadas) proposto no Capítulo 5. Essas instruções foram projetadas para apoiar tarefas de Visão Computacional, particularmente aquelas destinadas a descrever o contexto visual dessas imagens para pessoas com deficiência visual.

6.3.2 Detalhes de implementação

Na abordagem principal da etapa de Extração de Informações Visuais do VIIDA, assim como na Seção 5.2.2, foi usado o BLIP (Li et al. (2022)) ajustado como modelo VQA (*checkpoint* `model_base_vqa_capfilt_large.pth` e pesos `ViT-Base`). As imagens de entrada são redimensionadas para uma resolução de 480×480 *pixels* pelo BLIP. A seleção dos determinantes "*a*" ou "*an*" com base na pronúncia fonética foi realizada

por meio dos módulos Python Big Phoney[4] e Inflect[5].

Na abordagem complementar, para a tarefa de legendagem densa, foram empregados os modelos DenseCap (Johnson et al. (2016)) e GRIT (Nguyen et al. (2022)). Para reduzir o grande número de legendas densas, definiu-se ϵ como $0,0$ e $0,5$ para o DenseCap e GRIT, respectivamente. A escolha de $\epsilon = 0,0$ foi fundamentada no fato de que os valores de confiança do DenseCap, na amostragem utilizada, variaram aproximadamente entre $-7,6$ e $9,6$. Considerando esse intervalo, o valor $0,0$ foi escolhido como um meio termo. No caso do GRIT, já que seus valores de confiança variam entre 0 e 1 e que $0,5$ é o valor padrão adotado pelos autores no modelo, decidiu-se manter $\epsilon = 0,5$.

Para reduzir o número de legendas densas semelhantes geradas por ambos os modelos, foi adotado $T_{text\_sim} = 0,55$. Um modelo RoBERTa (Liu et al. (2019)) (implementação `roberta-base-CoLA` da Huggingface) ajustado no Corpus de Aceitabilidade Linguística (do inglês, *Corpus of Linguistic Acceptability* - CoLA) foi utilizado para a tarefa de classificação de texto com base na gramática da sentença. Finalmente, para confirmar o conteúdo descrito na sentença de acordo com a imagem usando o modelo VQA, adotou-se $T_{text\_yes} = 0,6$.

Na etapa de Produção de Estrutura de Parágrafo, foi utilizada a ferramenta SceneGraphParser (Wu et al. (2019)) para analisar as sentenças e extrair as relações gramaticais por meio de grafos de cena. Os valores de α, β e s da métrica InViDe foram definidos como $0,3$, $0,5$ e $0,05$, respectivamente. Utilizou-se a SpaCy (Honnibal et al. (2020)) (*pipeline* `en_core_web_lg`) como ferramenta de PLN e a NLTK (Bird et al. (2009)) para a WordNet. Para incorporação contextual de sentenças, integrou-se o `universal_sentence_encoder` (Cer et al. (2018)) a SpaCy como um componente de *pipeline*. Por fim, ressalta-se que os valores de todas as variáveis e limiares descritos acima foram definidos empiricamente.

Todas as execuções do método VIIDA e da métrica InViDe foram realizadas em um ambiente de desenvolvimento Jupyter Notebook, utilizando a linguagem de programação Python (versão 3.9.13). O *hardware* empregado foi um laptop padrão, equipado com 16 GB de RAM, processador AMD Ryzen 7 5800H com Radeon Graphics a 3,20 GHz, e uma GPU NVIDIA GeForce RTX 3060 com 6 GB de VRAM. O sistema operacional utilizado foi o Windows 11 Home Single Language, versão de 64 bits.

[4]Disponível publicamente em https://pypi.org/project/big-phoney/
[5]Disponível publicamente em https://pypi.org/project/inflect/

6.3.3 Baselines

O VIIDA, método proposto neste capítulo, foi comparado com o método previamente proposto no Capítulo 4, anteriormente denominado Ours e agora renomeado como *baseline*, além de outros sete competidores. Entre esses métodos, foi incluído o BLIP (Li et al. (2022)) (*checkpoint* `model_base_capfilt_large` e pesos `ViT-Base`) para a tarefa clássica de legendagem de imagem; Concat, que cria um parágrafo concatenando todas as saídas do modelo de legendagem densa; e Concat-Filter, que concatena apenas as legendas densas associadas a uma pessoa, conforme descrito na Seção 4.2.3. Ressalta-se que o Concat e Concat-Filter também foram apresentados no Capítulo 4.

Em relação aos demais métodos concorrentes, menciona-se o Regions-Hierarchical[6] (Krause et al. (2017)), que foi pioneiro na tarefa de parágrafo de imagem; os recentes LLaVA (Liu et al. (2023)) (implementação `llava-v1.5-13b-3GB`); InstructBLIP (Dai et al. (2023)) (implementação `instructblip-vicuna-7b` da Huggingface); e FuseCap (Rotstein et al. (2024)).

As variantes do VIIDA são referidas como VIIDA-Dense e VIIDA-GRIT. Essas variantes usam as legendas densas geradas pelos modelos DenseCap e GRIT, respectivamente. Os métodos Concat e Concat-Filter também têm variantes para ambos os modelos de legendagem densa (Concat-Dense, Concat-GRIT, Concat-Filter-Dense e Concat-Filter-GRIT).

Tanto LLaVA quanto InstructBLIP são modelos que incluem a tarefa de VQA e usam *prompts* textuais como entrada. Assim, dois *prompts* diferentes foram usados para gerar a descrição da imagem: um padrão, recomendado pelos respectivos trabalhos ("*provide a detailed description of the given image*"), e outro com foco na pessoa e no ambiente ("*write a detailed description about the person and the environment*") para uma comparação justa com o VIIDA que visa focar na pessoa e no ambiente da imagem.

É importante mencionar que havia sido testado um terceiro *prompt* ("*describe the image to a visually impaired person*") para o LLaVA e o InstructBLIP. No entanto, os modelos começaram a gerar descrições mencionando que a pessoa da imagem possuía deficiência visual. Devido a esses resultados, os experimentos com esse *prompt* foram suspensos. Além disso, não foram utilizados *prompts* negativos, isto é, instruções de texto fornecidas para que os modelos evitem usar determinadas características na geração do conteúdo solicitado, mantendo assim, as configurações padrões.

No final, 15 competidores, incluindo todas as variantes, foram considerados nos experimentos.

[6]Foi utilizado a implementação em PyTorch baseada no artigo original disponível publicamente em https://github.com/soloist97/region-hierarchical-pytorch

6.3.4 Métricas de avaliação

O desempenho dos competidores em relação ao texto produzido foi avaliado usando nove métricas: BLEU-{1, 2, 3, 4} (Papineni et al. (2002)), METEOR (Banerjee and Lavie (2005)), CIDEr (Vedantam et al. (2015)), BERTScore (pontuação F1) (Zhang et al. (2020b)), CLIPScore e RefCLIPScore (Hessel et al. (2021)).

Embora BLEU, METEOR e CIDEr, também utilizadas nos experimentos do Capítulo 4.2.4, sejam comumente empregadas na tarefa de parágrafo de imagem, a dependência dessas métricas na sobreposição entre os *n*-gramas de uma legenda gerada automaticamente e uma escrita por seres humanos para o cálculo de similaridade tem levado a preocupações sobre suas limitações e desempenhos sub-ótimos (Zhang et al. (2020b); Hessel et al. (2021); González-Chávez et al. (2024)). Consequentemente, houve um aumento no desenvolvimento de métricas orientadas a dados, algumas das quais utilizam diretamente as características da imagem (Wada et al. (2024)). Seguindo essa tendência, foram incluídas as métricas BERTScore, CLIPScore e sua derivação RefCLIPScore para avaliar os parágrafos gerados.

A métrica BLEU calcula a proximidade entre o texto candidato e o de referência ao avaliar a precisão da correspondência de *n*-gramas (Papineni et al. (2002)). Por outro lado, a METEOR utiliza a média harmônica entre precisão e *recall*, com maior ênfase no *recall*, e avalia a correspondência de unigramas com base na forma exata, forma do radical e sinônimos (Banerjee and Lavie (2005)). A métrica CIDEr avalia a consistência do texto gerado, considerando o consenso humano, por meio da sobreposição ponderada de TF-IDF para cada *n*-grama (Vedantam et al. (2015)). Esta métrica é considerada um bom indicador de desempenho do modelo, especialmente quando comparada com a BLEU e a METEOR (González-Chávez et al. (2024)).

A BERTScore calcula a similaridade entre duas sentenças somando as similaridades do cosseno entre os *embeddings* contextualizados pré-treinados do BERT para seus *tokens* (Zhang et al. (2020b)). A CLIPScore é uma métrica livre de referência que complementa as métricas existentes ao extrair características tanto da imagem quanto do texto candidato, calculando a similaridade do cosseno entre os *embeddings* visuais e textuais gerados pelo CLIP (Radford et al. (2021)). A RefCLIPScore é uma versão estendida da CLIPScore, que incorpora referências, caso estejam disponíveis, e é calculada como a média harmônica da CLIPScore com a similaridade do cosseno da referência máxima (Hessel et al. (2021)).

As métricas mencionadas acima são voltadas para avaliações automáticas da qualidade de textos gerados, seja por similaridade texto a texto ou imagem a texto (González-Chávez et al. (2024)). Para avaliar a adequação do texto gerado para pessoas com deficiência visual, foi utilizada a métrica proposta InViDe, a qual é a única desenvolvida especificamente para esse propósito até o momento.

6.3.5 Geração de texto para imagem

Foram empregados três modelos generativos baseados em difusão (Croitoru et al. (2023)) para produzir 684 imagens sintéticas com base nos textos de saída de cada um dos 15 competidores: Stable Diffusion 1.4 (SD1.4)[7] (Rombach et al. (2022)) (implementação `stable-diffusion-v-1-4-original` da Huggingface), Stable Diffusion 2.0 (SD2.0)[8] (implementação `stable-diffusion-2-base` da Huggingface), e Realistic Vision (RV)[9] (implementação `Realistic_Vision_V5.1` da Civitai). Além disso, também foram geradas imagens sintéticas com base nos parágrafos escritos por humanos presentes no conjunto de dados de webinários.

Os modelos foram executados utilizando os parâmetros padrão da ferramenta de interface de usuário Web AUTOMATIC1111[10]. Para avaliar a similaridade entre as imagens do conjunto de dados de webinários e as imagens sintéticas, utilizou-se a medida de similaridade do cosseno baseada em embeddings CLIP (Radford et al. (2021)) com pesos `ViT-B/32` e a pontuação FID[11] (Heusel et al. (2017)), ambas com parâmetros padrão.

A similaridade do cosseno avalia a similaridade entre imagens com base no cosseno do ângulo entre seus vetores de características, refletindo a semelhança na direção desses vetores. Quanto mais próxima de 1 for a similaridade do cosseno, maior é a semelhança entre os vetores de características, indicando que as imagens representadas são altamente semelhantes.

A pontuação FID é uma métrica que calcula a distância entre as distribuições multivariadas dos vetores de características extraídos pela InceptionV3 de imagens sintéticas e das imagens originais do domínio alvo (Jayasumana et al. (2024)). Essa métrica resume o grau de semelhança entre dois conjuntos de dados em termos de estatísticas sobre características das imagens. Pontuações mais baixas indicam que os dois conjuntos de imagens são mais semelhantes ou apresentam estatísticas mais próximas. A FID é utilizada para avaliar a qualidade de imagens geradas por GANs, e pontuações mais baixas têm sido associadas a imagens de maior qualidade (Heusel et al. (2017)). Para aplicar essa métrica, as imagens do conjunto de dados de webinário foram redimensionadas para a resolução de 512×512 *pixels*.

A geração dessas imagens sintéticas pode ser considerada uma forma de simular a representação mental que pessoas com deficiência visual poderiam formar ao ouvir

[7] Disponível publicamente em https://huggingface.co/CompVis/stable-diffusion-v-1-4-original

[8] Disponível publicamente em https://huggingface.co/stabilityai/stable-diffusion-2-base

[9] Disponível publicamente em https://civitai.com/models/4201?modelVersionId=130072

[10] Disponível publicamente em https://github.com/AUTOMATIC1111/stable-diffusion-webui

[11] Implementação PyTorch disponível publicamente em https://github.com/mseitzer/pytorch-fid

o parágrafo gerado pelos métodos. Portanto, essa parte dos experimentos visa identificar qual método de geração de parágrafo produz textos que resultam em imagens sintéticas que são o mais semelhantes possível à imagem original a partir da qual os textos foram criados.

6.3.6 Análise de conteúdo NSFW

Para identificar conteúdo "Não Seguro para o Trabalho" (do inglês, "*Not Safe For Work*" - NSFW), particularmente de natureza sexual, nas saídas textuais geradas pelo VIIDA, suas variantes e demais competidores, empregou-se um modelo DistilRoBERTa (implementação `distilroberta-nsfw-prompt-stable-diffusion` da Huggingface)[12], ajustado no subconjunto Civitai-2m-prompts para a tarefa de classificação de texto.

Além disso, para identificar conteúdo visual NSFW nas imagens geradas pelos modelos generativos, utilizou-se um modelo ViT ajustado (implementação `nsfw_image_detection` da Huggingface)[13] para a tarefa de classificação de imagens. Complementarmente, foi empregada uma API especializada para identificar imagens NSFW de natureza sexual[14].

6.4 Resultados e Discussão

6.4.1 Análise quantitativa

Avaliação do conteúdo do parágrafo

A Tabela 6.2 apresenta os resultados, expressos através da mediana[15], dos experimentos realizados por meio da aplicação das métricas de avaliação para cada método comparado. Como previsto, tanto os métodos de legendagem de imagem quanto os de legendagem densa exibiram as piores performances gerais. Embora o BLIP seja considerado estado da arte, esse método gera descrições de conteúdo de imagem simplistas em sentenças únicas, frequentemente ignorando detalhes visuais cruciais (Johnson et al. (2016); Rotstein et al. (2024)).

Os métodos de concatenação (Concat-{Dense, GRIT}) produziram sentenças excessivamente desconexas, indicando sua incapacidade de criar descrições coerentes

[12]Disponível publicamente em https://huggingface.co/AdamCodd/distilroberta-nsfw-prompt-stable-diffusion

[13]Disponível publicamente em https://huggingface.co/Falconsai/nsfw_image_detection

[14]Disponível publicamente em https://api4.ai/apis/nsfw

[15]A mediana foi utilizada para resumir as pontuações em vez da média, visto que o teste de normalidade de Shapiro-Wilk retornou um p-valor $< 0,05$, indicando que os conjuntos de pontuações seguem uma distribuição não normal.

Tabela 6.2: Comparação entre métodos usando as métricas clássicas e recentes para medir a similaridade entre os parágrafos candidatos e os de referência. Os valores são a mediana multiplicada por 100 (para mostrar menos casas decimais), com os melhores destacados em negrito. O símbolo - indica que a GPU utilizada ficou sem memória ao processar os dados do método para a métrica, enquanto os símbolos {*, ”} indicam que não houve diferença significativa entre o VIIDA e suas variantes em comparação consigo mesmas e com os outros métodos usando o teste de comparação em pares de Dunn com o p-valor $< 0,05$. B, MT, CDr, BTS, CS e RCS referem-se a BLEU, METEOR, CIDEr, BERTScore, CLIPScore e RefCLIPScore, respectivamente.

Método / Métricas	**B-1**	**B-2**	**B-3**	**B-4**	**MT**	**CDr**	**BTS**	**CS**	**RCS**
BLIP	0,08	0,03	0,00	0,00	5,15	0,00	64,46	58,48	56,64
Concat-Dense	6,38	3,92	2,41	1,45	10,71	0,00	-	58,80	59,35
Concat-Filter-Dense	32,41	19,72	12,13	7,42	18,52	0,12	58,20	56,61	58,76
Concat-GRIT	11,32	6,60	3,79	1,96	15,62	0,00	-	61,30	62,67
Concat-Filter-GRIT	33,33	19,72	12,04	6,66	17,14	0,94	60,45	60,56	61,37
Baseline	17,34	9,43	3,40	0,00	12,84	0,00	68,56	54,22	56,40
Regions-Hierarchical	37,50	21,58	12,08	0,00	16,66	1,27	63,76	49,76	52,24
LLaVA	33,33	19,44	10,60	0,00	17,75	0,04	66,22	62,20	65,77
LLaVA-Directed	30,54	17,54	9,77	0,00	17,96	0,00	66,13	61,52	65,60
InstructBLIP	26,17	14,82	7,38	0,00	17,12	0,00	66,75	64,50	65,92
InstructBLIP-Directed	27,72	16,31	8,80	0,00	17,51	0,00	65,55	63,70	65,53
FuseCap	36,37	21,15	11,08	0,00	17,22	1,46”	73,31	63,70	63,53
VIIDA	46,76	30,40	19,86	12,11*	22,02	2,38”	**82,41**	65,84	69,21
VIIDA-Dense	53,80*	34,07*	21,77*	**12,94***	23,49*	13,77*	79,51*	**67,38***	**70,45***
VIIDA-GRIT	**54,36***	**34,23***	**21,81***	12,91*	**23,67***	**17,70***	78,72*	66,94*	70,31*

(Krause et al. (2017)). As descrições geradas pelos métodos Concat-Filter-{Dense, GRIT} apresentaram pontuações mais altas em comparação com o *baseline* e com os métodos baseados em LLMs (LLaVA e InstructBLIP), particularmente para métricas clássicas (BLEU, METEOR e CIDEr). No entanto, isso pode ser atribuído ao maior número de palavras associadas a uma pessoa produzidas por esses métodos, conforme discutido anteriormente na Seção 4.3. Essa situação muda ao considerar as métricas baseadas em dados (BERTScore, CLIPScore e RefCLIPScore).

Ao analisar as pontuações dos modelos concorrentes que não foram incluídos nos experimentos do Capítulo 4, percebeu-se que, para as métricas clássicas, o Regions-Hierarchical e o FuseCap se destacaram em comparação com ambas as formas (normal e direcionada) de LLaVA e InstructBLIP. Quanto às outras métricas baseadas em similaridade de dados, houve uma melhora para o LLaVA e o InstructBLIP, enquanto o método Regions-Hierarchical apresentou uma piora. Acredita-se que essa discrepância nas pontuações decorra da utilização de conjuntos de dados maiores e mais recentes durante o treinamento de LLaVA, InstructBLIP e FuseCap.

Os resultados na Tabela 6.2 indicam que o VIIDA e suas variantes, VIIDA-Dense e, principalmente, VIIDA-GRIT, superaram os outros métodos em todas as métricas.

Isso sugere que o método proposto poderia gerar descrições de imagem mais próximas das anotações humanas destinadas a pessoas com deficiência visual, presentes no conjunto de dados de webinários, quando comparado com os métodos existentes. É perceptível que VIIDA-GRIT, Concat-GRIT e Concat-Filter-GRIT alcançaram pontuações mais altas do que as versões desses métodos com o modelo DenseCap na maioria das métricas. Essa melhoria pode ser atribuída à substituição do detector baseado em CNN, usado nos métodos anteriores, por um detector baseado em Transformer no modelo GRIT, o que resultou em melhorias de desempenho, especialmente na redução da falta de informações contextuais e do risco de detecção imprecisa durante a descrição das regiões da imagem (Nguyen et al. (2022)).

Considerando que o VIIDA é especificamente projetado para facilitar a geração de descrição automática de imagens no contexto de webinários para pessoas com CBV, e que a maioria das métricas usadas em sua avaliação depende de textos de referência para comparação, deve-se estar ciente de possíveis vieses nos textos de referência do conjunto de dados de webinários. Para demonstrar a robustez do VIIDA e suas variações na produção de descrições que se correlacionam de perto com o conteúdo da imagem, independentemente das comparações com as anotações do conjunto de dados, explorou-se o uso da métrica CLIPScore.

Como pode ser visto na Tabela 6.2 na coluna da CLIPScore, o VIIDA e suas variações alcançaram as pontuações mais altas. Notavelmente, o VIIDA-Dense exibiu uma diferença de $2,88$ pontos percentuais em comparação com a maior pontuação alcançada pelos outros concorrentes, especificamente pelo InstructBLIP, sendo essa diferença estatisticamente significativa (como pode ser visto na Tabela 6.2 e na Figura 6.5). Consequentemente, mesmo usando um conjunto de dados específico para o problema em questão (imagens de webinários anotadas por humanos direcionadas a pessoas com CBV), foi demonstrado com sucesso a eficácia do método aqui proposto em alcançar uma forte compatibilidade semântica entre imagem e texto.

Como os resultados dos testes de Shapiro-Wilk não indicaram uma distribuição normal, foi aplicado o teste não paramétrico de Kruskall-Wallis para avaliar a significância estatística das diferenças entre a mediana das pontuações. Para todas as métricas analisadas, o teste de Kruskal-Wallis produziu resultados estatisticamente significativos (p-valor $< 0,05$). Em seguida, o teste *post-hoc* (ou de comparações múltiplas) de Dunn com o procedimento de Benjamini-Hochberg foi usado para determinar exatamente quais competidores são estatisticamente diferentes (Chen et al. (2017); Jafari and Ansari-Pour (2019)).

A Figura 6.5, através da visualização gráfica de uma matriz de comparações pareadas, ilustra quais métodos são estatisticamente diferentes entre si em relação às

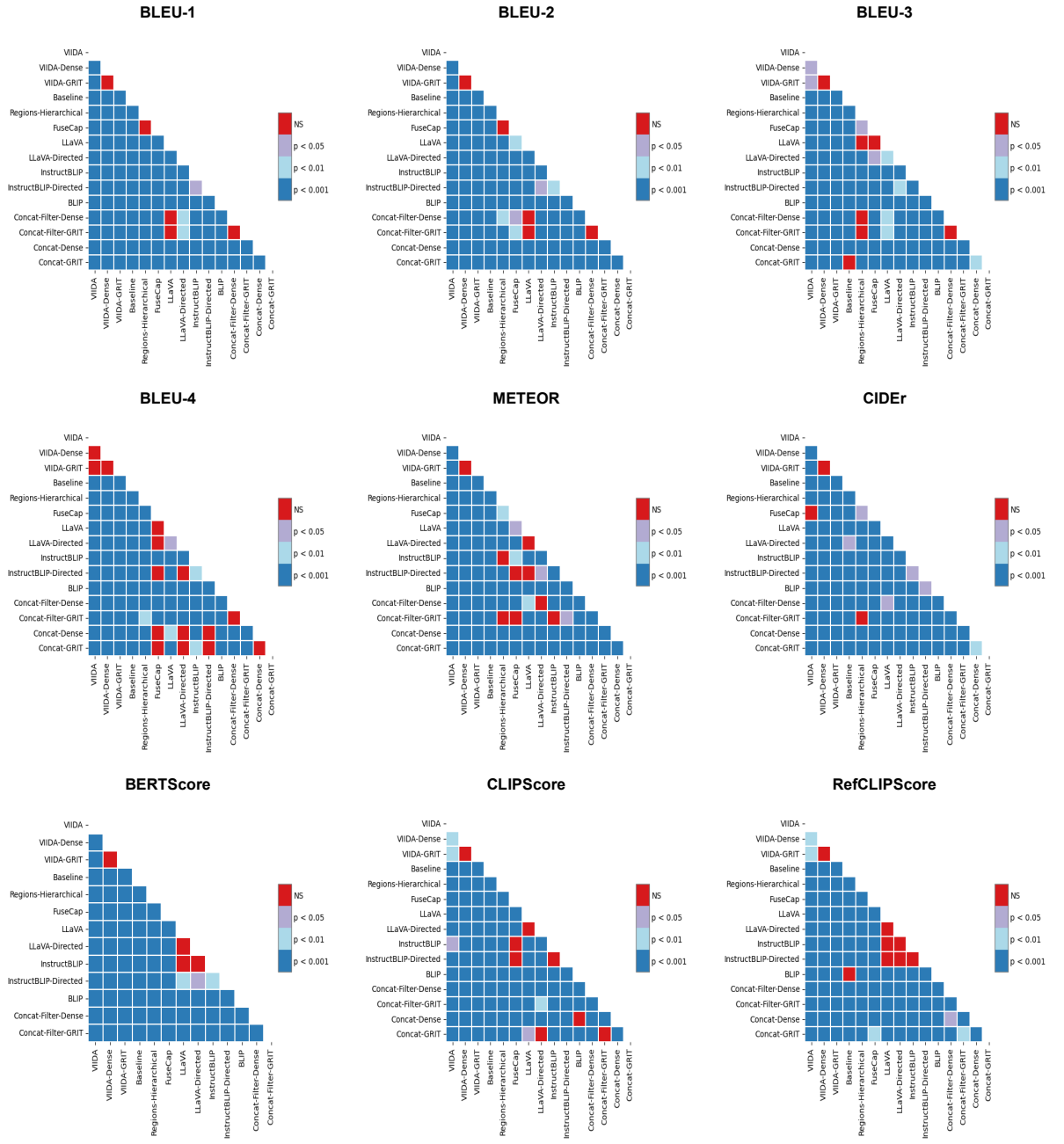


Figura 6.5: Resultados do teste *post-hoc* de Dunn com procedimento de Benjamini-Hochberg para avaliar a significância estatística entre os valores medianos dos métodos para cada métrica de avaliação de texto utilizada. NS e p referem-se a "Não Significativo" e "*p*-valor". Fonte: Material próprio.

pontuações de cada métrica de avaliação, após a aplicação do teste *post-hoc* de Dunn com o procedimento de Benjamini-Hochberg. Cada célula da matriz representa o resultado de uma comparação entre dois métodos, indicando se a diferença é significativa (p-valor $< 0,05$) ou não, sendo esta última destacada em vermelho.

A Tabela 6.3 apresenta as estatísticas de linguagem que incluem o número médio e desvio padrão de caracteres, palavras e sentenças nos parágrafos gerados por cada

Tabela 6.3: Estatísticas de linguagem sobre a média e o desvio padrão referentes ao número de caracteres, palavras e sentenças nos parágrafos gerados por cada método. Os valores foram arredondados.

Método	**Caracteres**	**Palavras**	**Sentenças**
BLIP	27 ± 9	6 ± 2	1 ± 0
Concat-Dense	2.370 ± 223	489 ± 43	104 ± 8
Concat-Filter-Dense	325 ± 76	69 ± 15	14 ± 3
Concat-GRIT	1.307 ± 460	278 ± 99	59 ± 23
Concat-Filter-GRIT	289 ± 93	59 ± 18	12 ± 4
Baseline	103 ± 33	22 ± 7	3 ± 1
Regions-Hierarchical	283 ± 67	64 ± 14	5 ± 0
LLaVA	402 ± 84	72 ± 14	4 ± 1
LLaVA-Directed	453 ± 85	81 ± 15	4 ± 0
InstructBLIP	431 ± 185	82 ± 35	4 ± 2
InstructBLIP-Directed	518 ± 107	95 ± 20	5 ± 1
FuseCap	171 ± 12	35 ± 3	1 ± 0
VIIDA	185 ± 19	35 ± 3	4 ± 0
VIIDA-Dense	216 ± 31	41 ± 6	5 ± 1
VIIDA-GRIT	225 ± 38	43 ± 7	6 ± 1
Humanos	*256 ± 38*	*49 ± 7*	*4 ± 0*

método. A linha "Humanos" diz respeito às anotações humanas no conjunto de dados de webinários e serve como referência para avaliar o desempenho dos métodos automáticos.

Nesse contexto, ao analisar a Tabela 6.3, fica evidente que a discrepância nas quantidades retornadas por BLIP e Concat-{Dense, GRIT} em comparação com outros métodos está alinhada com as descobertas apresentadas na Tabela 6.2. Os métodos Concat-Filter-{Dense, GRIT} geram parágrafos que estão próximos em características dos parágrafos escritos por humanos. No entanto, seus parágrafos exibem um desvio padrão maior e contêm quase três vezes o número de sentenças, um comportamento relatado anteriormente na Seção 4.3.

É possível observar que o *baseline*, FuseCap e VIIDA tendem a gerar parágrafos com menos caracteres, palavras e sentenças em comparação com outros métodos. A redução nas quantidades produzidas pelo *baseline* é devido ao uso do PLM para a tarefa de sumarização, enquanto para VIIDA, é devido à ausência de um modelo de legendagem densa para adicionar novas sentenças. O menor número de sentenças geradas pelo FuseCap pode ser atribuído ao seu objetivo de enriquecer as legendas com informações visuais adicionais (Rotstein et al. (2024)). Similarmente ao BLIP, o FuseCap gera apenas uma única sentença, mas com mais detalhes, o que aumenta seu comprimento, como é evidente no maior número de caracteres e palavras.

Os métodos baseados em LLMs produzem parágrafos com um número de sentenças semelhante aos escritos por humanos, mas maiores em termos de caracteres e palavras, ficando atrás apenas dos métodos Concat-{Dense, GRIT}. Isso

pode ser devido ao treinamento dos LLMs para gerar respostas de longo formato.

O número de características aumenta nas versões desses métodos direcionadas em descrever a pessoa e o ambiente na imagem. Ambos os métodos (LLaVA e InstructBLIP) passaram por ajuste de instrução, um processo destinado a melhorar o desempenho de generalização desses modelos para tarefas que não encontraram durante seus respectivos treinamentos (Dai et al. (2023); Liu et al. (2023)). Portanto, espera-se que eles sejam capazes de abordar diretamente a intenção do usuário via *prompt*, ajustando a resposta de saída com mais detalhes visuais. Em particular, o InstructBLIP usa uma variedade mais ampla de dados de instrução de linguagem visual em comparação com o LLaVA (26 conjuntos contra apenas um), o que explica seus parágrafos mais longos (Dai et al. (2023)).

Os métodos VIIDA, VIIDA-{Dense, GRIT}, Regions-Hierarchical e FuseCap geralmente apresentam parágrafos de comprimento médio mais próximo dos escritos por humanos, correlacionando com pontuações mais altas nas métricas clássicas como BLEU, METEOR e CIDEr. Os VIIDA-{Dense, GRIT} mostram uma menor variabilidade, indicando sua similaridade com os parágrafos humanos de acordo com os gráficos de intervalo de confiança de 95% da Figura 6.6, principalmente no número de caracteres e palavras. O VIIDA demonstra uma precisão em todas as características, principalmente no número de sentenças.

Tabela 6.4: Estatísticas das características linguísticas dos parágrafos para Substantivos (S), Verbos (V), Adjetivos (AJ), Advérbios (AV), Pronomes (P), Conjunções Coordenativas (CC) e Tamanho do Vocabulário (VC). Todos os valores foram normalizados pelo total de parágrafos analisados no conjunto de dados de webinários (684). Os valores em negrito representam a maior aproximação aos valores dos parágrafos anotados por humanos.

Método / Marcações POS	**S**	**V**	**AJ**	**AV**	**P**	**CC**	**VC**
BLIP	2,24	0,41	0,60	0,00	0,05	0,29	0,16
Concat-Dense	176,40	14,30	90,24	**0,03**	0,22	13,32	**0,67**
Concat-Filter-Dense	26,02	6,33	7,41	0,00	0,15	0,23	0,30
Concat-GRIT	110,73	8,22	27,23	0,34	0,30	2,04	1,75
Concat-Filter-GRIT	21,17	**4,92**	6,10	0,19	0,08	0,18	**0,67**
Baseline	7,64	2,05	2,43	0,02	0,84	0,34	0,31
Regions-Hierarchical	15,29	5,70	8,45	0,29	3,03	**3,14**	0,88
LLaVA	18,98	10,83	4,48	1,59	**3,77**	4,06	1,42
LLaVA-Directed	21,07	11,51	5,49	1,85	4,23	5,00	1,54
InstructBLIP	21,71	9,51	6,24	1,79	4,63	3,13	1,40
InstructBLIP-Directed	25,95	11,50	7,02	2,02	5,51	3,94	1,62
FuseCap	9,85	3,05	6,66	0,35	0,96	2,43	0,42
VIIDA	9,56	2,00	7,29	0,00	3,06	1,64	0,27
VIIDA-Dense	12,00	2,09	8,21	0,00	3,13	1,86	0,44
VIIDA-GRIT	**12,59**	2,19	**8,54**	**0,03**	3,13	1,80	**0,67**
Humanos	*13,77*	*4,02*	*9,49*	*0,06*	*3,50*	*3,49*	*0,57*

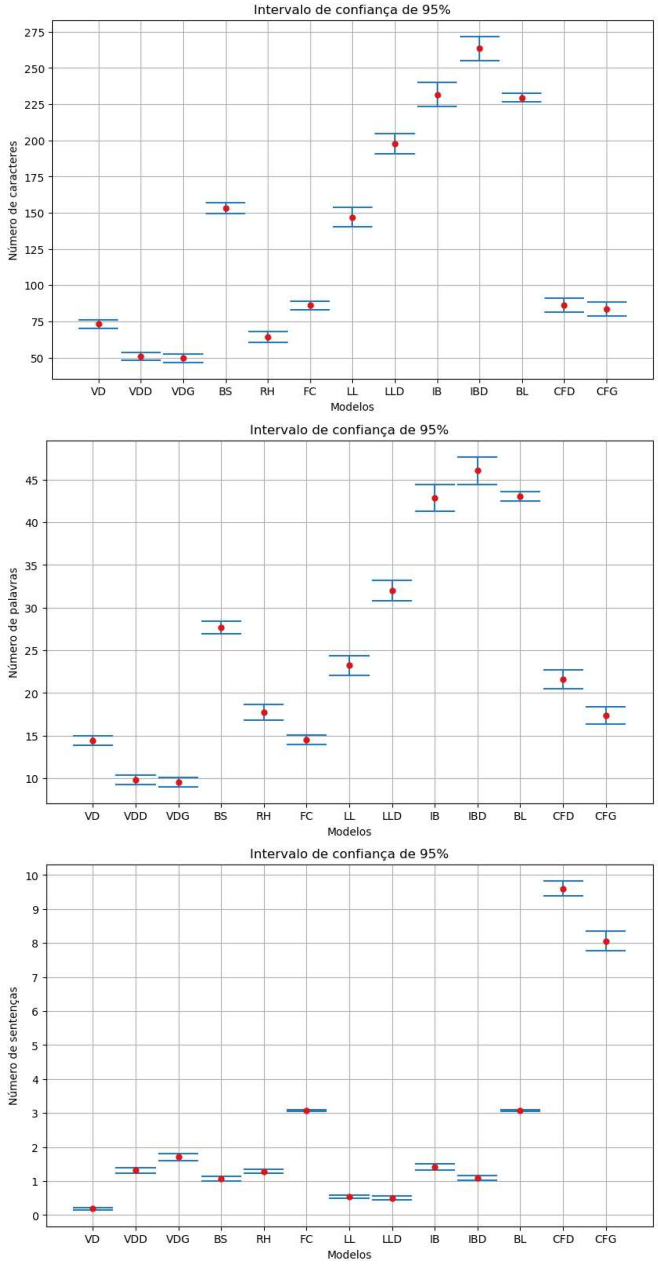


Figura 6.6: Gráficos de intervalos de confiança de 95% para análise das estatísticas de linguagem relacionadas ao número de caracteres, palavras e sentenças nos parágrafos gerados pelos seguintes métodos: VIIDA (VD), VIIDA-Dense (VDD), VIIDA-GRIT (VDG), *baseline* (BS), Regions-Hierarchical (RH), FuseCap (FC), LLaVA (LL), LLaVA-Directed (LLD), InstructBLIP (IB), InstructBLIP-Directed (IBD), BLIP (BL), Concat-Filter-Dense (CFD) e Concat-Filter-GRIT (CFG). Para melhor visualização da escala, os resultados de Concat-{Dense, GRIT} foram omitidos. O intervalo de confiança serve como medida da diferença entre parágrafos humanos e aqueles gerados pelos métodos. Valores mais próximos de zero indicam maior similaridade com parágrafos escritos por humanos. Fonte: Material próprio.

A Tabela 6.4 mostra que o VIIDA-GRIT alcançou resultados mais próximos dos parágrafos escritos por humanos para metade das marcações POS analisadas e para o tamanho do vocabulário. Para as outras *tags*, o VIIDA-GRIT não retornou os

valores mais próximos, mas ainda permanece perto às anotações humanas. Além disso, o VIIDA, VIIDA-Dense, Regions-Hierarchical e FuseCap também se destacam na geração de características linguisticamente semelhantes.

É evidente que os valores em negrito de Concat-Dense e Concat-Filter-GRIT estão diretamente relacionados ao comprimento dos parágrafos, como visto na Tabela 6.3. Os métodos baseados em LLMs exibem uma maior diversidade em relação à maioria das *tags*, pois são treinados em conjuntos de dados de visão e linguagem em grande escala (Dai et al. (2023); Liu et al. (2023)). No entanto, observou-se uma menor tendência no uso de adjetivos por esses métodos.

Avaliação de adequação para deficientes visuais

Nesta seção, avaliou-se a adequação das descrições de imagens geradas para pessoas com CBV, utilizando a métrica InViDe proposta. A Tabela 6.5 exibe os valores medianos[16] para cada critério analisado, juntamente com o valor final da métrica.

A análise dos resultados finais da InViDe revelou que o VIIDA e suas variantes alcançaram as pontuações mais altas e estes métodos não exibiram diferenças estatísticas entre si, mas sim quando comparados aos demais concorrentes. Isso sugere a superior capacidade de adequação do VIIDA e de suas variantes para a tarefa de produzir textos que transmitam o contexto visual das imagens para indivíduos com CBV, alinhando-se com padrões, diretrizes, melhores práticas e pesquisas referenciadas no campo (ABNT (2016); Lewis (2018); Stangl et al. (2020)).

Ao examinar os critérios individualmente, foi observado que as características linguísticas e a descrição do conteúdo do entorno exibiram a menor variabilidade entre os métodos. Notavelmente, a descrição do conteúdo relacionado à pessoa forneceu achados valiosos sobre as capacidades descritivas dos métodos avaliados. Além do VIIDA e de suas variações, apenas o FuseCap alcançou pontuações superiores a 85% neste critério. Em cenários como webinários, onde os indivíduos são o foco central do contexto apresentado, esperava-se pontuações altas neste critério devido à sua cobertura abrangente de atributos como características corporais, emoções, acessórios e roupas. Essa variabilidade observada sugere a ausência desses conteúdos, características (adjetivos) e detalhes nos conjuntos de dados de treinamento utilizados pelos métodos que alcançaram pontuações baixas nesse critério. Além disso, é importante lembrar que os critérios de características linguísticas, descrição da pessoa e descrição do ambiente dependem do texto de referência. Consequentemente, foi deduzido que o critério relacionado ao conteúdo sobre à pessoa exibe a maior discrepância em relação aos padrões recomendados.

[16] A mediana foi utilizada para resumir as pontuações em vez da média, visto que o teste de normalidade Shapiro-Wilk retornou um p-valor $< 0,05$, indicando que os conjuntos de pontuações seguem uma distribuição não normal.

Tabela 6.5: Comparação dos métodos usando a métrica InViDe para avaliar a adequação dos textos para indivíduos com deficiência visual por meio de análise multicritério no nível do corpus. Os valores medianos são mostrados multiplicados por 100, com os melhores destacados em negrito. Além da mediana, a última coluna "Pontuação InViDe" inclui também o valor do desvio padrão. Os símbolos {*, ”} indicam que não houve diferença significativa entre o VIIDA e suas variantes em comparação com eles mesmos e com os outros métodos usando o teste de comparação em pares de Dunn com o p-valor $< 0,05$. CL, CE, DP, DE e CT referem-se a Características Linguísticas, Cobertura de Elementos-chave, Descrição da Pessoa (apresentador), Descrição do Entorno e Comprimento do Texto, respectivamente.

Método / Critérios	**CL**	**CEC**	**DP**	**DE**	**CT**	**Pontuação InViDe**
BLIP	89,06	0,00	36,00	85,18	0,00	42,01 ± 0,02
Concat-Dense	87,50	83,33	49,76	66,44	0,00	57,20 ± 0,05
Concat-Filter-Dense	86,72	66,67	79,16	92,78	93,40	80,21 ± 0,08
Concat-GRIT	81,52	83,33	69,13	82,68	0,00	61,64 ± 0,09
Concat-Filter-GRIT	84,12	50,00	80,01	90,25*”	99,72	78,45 ± 0,07
Baseline	89,87	50,00	75,59	91,15*”	73,56	75,86 ± 0,10
Regions-Hierarchical	93,92	50,00	80,72	90,47*”	99,87	81,74 ± 0,08
LLaVA	92,03	66,67	53,08	92,22	1,21	62,68 ± 0,10
LLaVA-Directed	91,96	66,67	51,34	91,75	0,00	61,33 ± 0,07
InstructBLIP	90,03	66,67	56,59	86,66*	0,00	61,41 ± 0,07
InstructBLIP-Directed	90,66	66,67	53,75	86,27	0,00	60,91 ± 0,05
FuseCap	**96,48***	50,00	88,88*	92,03*	**100,00***	86,22 ± 0,05
VIIDA	96,46*	83,33	87,28*”	**94,44**	**100,00***	92,98* ± 0,03
VIIDA-Dense	95,16”	**100,00***	**90,47”**	91,45*	**100,00***	**93,45*** ± 0,03
VIIDA-GRIT	94,93”	**100,00***	89,19*”	90,58*”	**100,00***	92,64* ± 0,03

Quanto aos dois critérios de referência livre, ou seja, cobertura de elementos-chave e comprimento do texto, ao analisar a Tabela 6.5, foi assumido que ambos critérios exercem as influências mais significativas no cálculo da pontuação final da métrica InViDe. Apenas os métodos Concat-{Dense, GRIT}, juntamente com o VIIDA e suas variantes, alcançaram uma alta taxa ($\geq$ 80%) de cobertura do conteúdo mínimo relevante nas descrições. Isso pode ser atribuído ao número substancial de legendas densas geradas por esses métodos. Consequentemente, os concorrentes deixam de incluir alguns conteúdos pertinentes para indivíduos com CBV.

O critério de comprimento do parágrafo aparece como o mais penalizador para os métodos avaliados. Os parágrafos que estão fora da faixa de 125 a 280 caracteres sofrem penalidades, com a penalidade máxima aplicada se os parágrafos apresentarem uma quantidade de caracteres abaixo do valor mínimo ou acima do valor máximo (as variáveis "*piso*" e "*teto*", respectivamente na Figura 6.4). Conforme mencionado anteriormente na Seção 4.3, o comprimento do parágrafo é uma característica importante para melhorar a compreensão do contexto da imagem, destacando a importância de aderir ao comprimento apropriado ao produzir

parágrafos adequados para pessoas com deficiência visual.

Uma melhoria significativa nos resultados é observada ao comparar o VIIDA e suas variantes com o *baseline*. No entanto, é importante destacar que, embora o *baseline* tenha alcançado valores mais baixos nas métricas de qualidade do texto, ao analisar sua adequação para pessoas com deficiência visual, ele demonstra superioridade em relação aos modelos atuais baseados em LLMs.

Como os resultados dos testes de Shapiro-Wilk não indicaram uma distribuição normal, aplicou-se o teste de Kruskal-Wallis (p-valor $< 0,05$) e o teste de *post-hoc* de Dunn com o procedimento de Benjamini-Hochberg para avaliar a significância estatística das diferenças entre as medianas de cada critério analisado e a pontuação final da métrica InViDe. Os resultados das comparações múltiplas realizadas podem ser vistos na Figura 6.7.

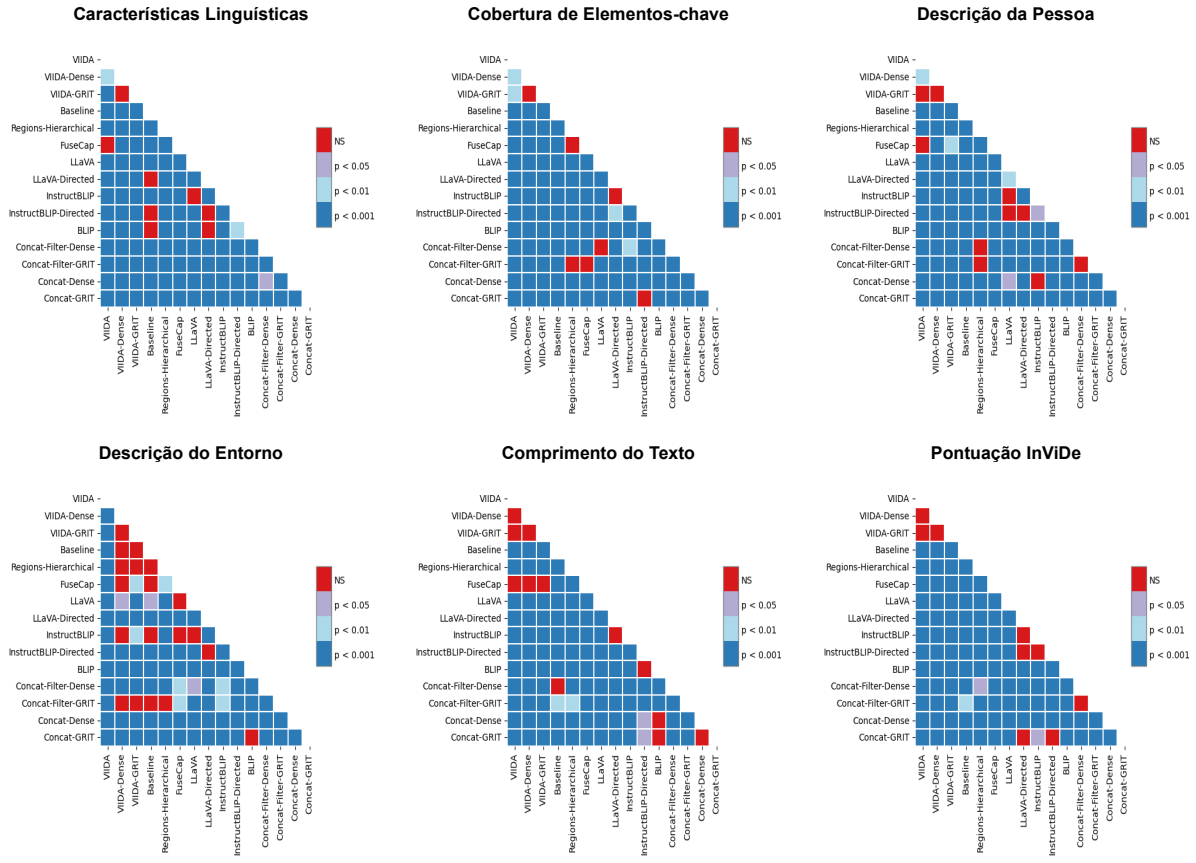


Figura 6.7: Resultados do teste *post-hoc* de Dunn com procedimento de Benjamini-Hochberg para avaliar a significância estatística das diferenças entre os valores medianos dos métodos para cada critério analisado da métrica InViDe, juntamente com o valor final da métrica. NS e p referem-se a "Não Significativo" e "p-valor". Fonte: Material próprio.

Avaliação imagem a imagem

Na tarefa de legendagem de imagens e suas variações, uma imagem de entrada é fornecida a um modelo de Aprendizado Profundo, que produz uma saída textual descrevendo o conteúdo da imagem. Nesse contexto, foi realizado um experimento adicional fazendo o procedimento inverso, ou seja, dado um parágrafo, a tarefa é gerar uma imagem de acordo com o conteúdo do texto. Aqui, usou-se os parágrafos produzidos artificialmente a partir das imagens existentes no conjunto de dados de webinários.

O objetivo do experimento foi comparar as imagens geradas a partir desses parágrafos com as imagens originais que deram origem a esses mesmos parágrafos. Para esse fim, foram usadas as saídas dos 15 métodos avaliados juntamente com os parágrafos anotados do conjunto de dados do webinários como entrada para os três modelos generativos baseados em difusão (Croitoru et al. (2023)), ou seja, geradores de texto para imagem, para produzir imagens sintéticas.

Cada um dos três modelos generativos (SD1.4, SD2.0 e RV) produziu 10.944 imagens sintéticas, cada uma com dimensões de 512×512 *pixels*, resultando em um total de 32.832 imagens. As imagens originais foram comparadas com as imagens sintéticas para verificar se o conteúdo descrito nos parágrafos fornecia as informações necessárias para criar imagens semelhantes às originais. A Tabela 6.6 resume os resultados obtidos das avaliações realizadas.

Em termos de similaridade do cosseno, embora seguidos de perto por InstructBLIP, InstructBLIP-Directed e FuseCap, o VIIDA e suas variantes consistentemente alcançaram os maiores valores em todos os modelos generativos. Portanto, o método proposto demonstrou um desempenho satisfatório na geração de descrições que resultam em imagens mais semelhantes às originais, sugerindo a eficácia do VIIDA na captura e representação precisa do conteúdo visual.

De acordo com os resultados da Tabela 6.6, o VIIDA e VIIDA-{Dense,GRIT} exibiram as menores pontuações FID em todos os modelos generativos, indicando que as imagens sintéticas geradas através dos parágrafos produzidos pelo método proposto neste capítulo estão mais próximas das imagens originais em termos de distribuição estatística. Além disso, o FuseCap também demonstra desempenho competitivo em termos de pontuações FID. Esses resultados corroboram com as descobertas apresentadas pela similaridade do cosseno. No entanto, a FID mostrou uma discrepância maior entre os resultados do VIIDA e de suas variantes em comparação com os demais concorrentes.

Em geral, os resultados destacam a eficácia das saídas do VIIDA e suas variantes na geração de imagens sintéticas semanticamente similares às imagens originais, conforme indicado pelos altos valores da similaridade do cosseno e baixas

Tabela 6.6: A comparação das imagens sintéticas geradas pelos modelos Stable Diffusion 1.4 (SD1.4), Stable Diffusion 2.0 (SD2.0) e Realistic Vision (RV) com as imagens do conjunto de dados de webinários é realizada usando como medidas a similaridade do cosseno e a FID. Os valores medianos de similaridade do cosseno são mostrados multiplicados por 100, enquanto os valores da FID são relatados como valores absolutos. Os melhores resultados são destacados em negrito. Os símbolos {*, ″} indicam que não houve diferença significativa entre o VIIDA e suas variantes em comparação com eles mesmos e com os outros métodos usando o teste de comparação em pares de Dunn com o p-valor $< 0,05$, aplicado apenas à similaridade do cosseno.

	Similaridade do cosseno (↑)			**Pontuação FID (↓)**		
Método / Modelo generativo	**SD1.4**	**SD2.0**	**RV**	**SD1.4**	**SD2.0**	**RV**
BLIP	51,92	52,88	49,24	139,62	144,21	151,61
Concat-Dense	49,14	47,83	48,53	153,68	162,53	144,75
Concat-Filter-Dense	50,29	49,92	49,71	151,07	153,42	164,64
Concat-GRIT	50,53	50,00	51,31	136,19	139,55	126,76
Concat-Filter-GRIT	50,70	50,46	50,65	149,38	136,70	138,45
Baseline	48,57	50,73	46,70	147,10	146,61	157,67
Regions-Hierarchical	48,82	49,41	46,28	158,50	155,34	161,63
LLaVA	55,51	56,49	54,61	131,43	132,91	138,89
LLaVA-Directed	55,85	55,85	54,71	134,74	138,10	143,00
InstructBLIP	57,17″	58,32″	55,83*	126,11	132,53	136,36
InstructBLIP-Directed	56,61	57,64	55,81*	135,64	141,52	146,02
FuseCap	56,29	57,91″	56,42*	116,73	116,30	121,84
VIIDA	**58,00***″	58,88*″	56,46*	**110,24**	114,97	111,46
VIIDA-Dense	57,95*″	59,76*	56,88*	110,61	**111,24**	110,83
VIIDA-GRIT	57,64″	**59,96***	**57,10***	111,81	113,32	**110,51**
Humanos	*58,20**	*59,64**	*58,86*	*107,36*	*105,72*	*110,23*

pontuações FID. Essa constatação reforça a coerência dos parágrafos em termos do conteúdo da imagem indicado pelos valores da CLIPScore relatados na Tabela 6.2.

É importante mencionar que a capacidade de geração dos modelos generativos não está sendo analisada nesta tese. Além disso, assim como em experimentos anteriores, os resultados da similaridade do cosseno foram validados estatisticamente por meio da aplicação do teste de Shapiro-Wilk (p-valor $< 0,05$), do teste de Kruskal-Wallis (p-valor $< 0,05$) e do teste *post-hoc* de Dunn com o procedimento de Benjamini-Hochberg, conforme apresentado na Figura 6.8. Nenhum teste de hipótese estatística foi aplicado aos valores da FID porque a métrica retorna um único valor de comparação entre os conjuntos de dados em vez de valores individuais para cada amostra.

No contexto de descrições de imagens para usuários com deficiência visual, esse tipo de experimento poderia ser pensado como uma aproximação das imagens que indivíduos com CBV – ou até mesmo pessoas com visão sem acesso à imagem original – poderiam criar em suas mentes ao ouvir descrições textuais de imagens de webinários. Supondo que isso seja uma boa aproximação, o VIIDA e suas variantes

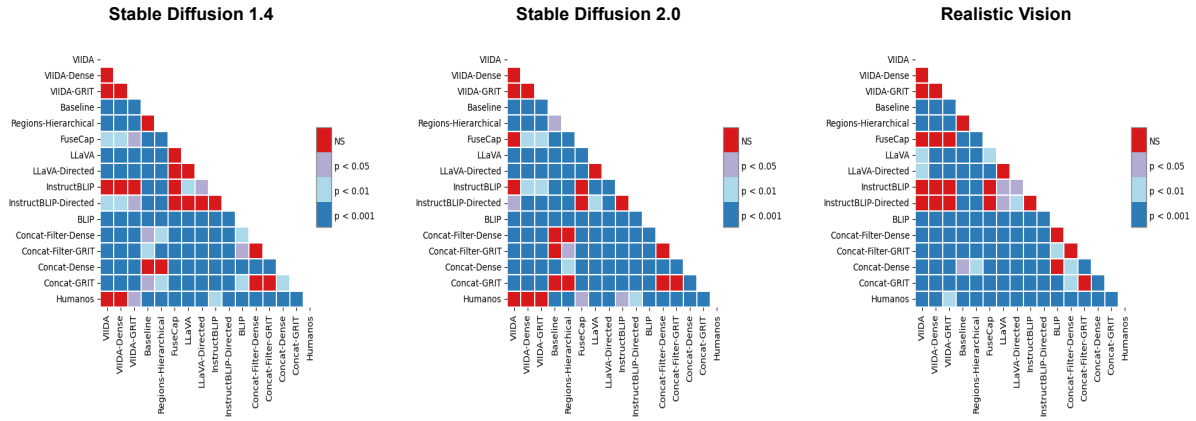


Figura 6.8: O resultado dos testes *post-hoc* de Dunn com procedimento de Benjamini-Hochberg para avaliar a significância estatística das diferenças entre os valores medianos de similaridade de cosseno das imagens sintéticas geradas por cada modelo generativo. NS e p referem-se a "Não Significativo" e "*p*-valor". Fonte: Material próprio.

foram as abordagens que poderiam produzir os melhores parágrafos em termos de quão precisamente uma pessoa imaginaria uma cena de webinário, por exemplo, com base apenas em uma descrição textual.

6.4.2 Análise qualitativa

Avaliação dos parágrafos

A Figura 6.9 apresenta exemplos de imagens de entrada e seus parágrafos correspondentes gerados pelos métodos comparados nesta tese. O BLIP retornou descrições com uma estrutura simples, carentes de detalhes e especificidade. Os métodos Concat e Concat-Filter para ambos os modelos de legendagem densa (DenseCap e GRIT) produziram descrições com estruturas simples, contendo numerosos e óbvios detalhes no geral, particularmente o Concat-Dense e o Concat-Filter-Dense, que também geraram textos significativamente longos (conforme apresentado na Tabela 6.3). Esses resultados foram omitidos da figura para maior clareza na visualização.

Embora o *baseline* tenha utilizado um modelo de linguagem não específico para a tarefa de parágrafo de imagem, esse método gerou descrições superiores em termos de elementos linguísticos em comparação com o BLIP. No entanto, notou-se algumas falhas em seus resultados devido a erros propagados pelos modelos usados no *pipeline*, incluindo o DenseCap.

Em relação ao Regions-Hierarchical, mesmo que tenha produzido parágrafos mais longos do que o BLIP e o *baseline*, foi observado um maior número de erros nas

BLIP:

A man wearing headphones in front of a book shelf.

Baseline:

Red shirt on a man, wearing a necklace.

Regions-Hierarchical:

A woman is standing on the left side of the photo. the bookshelf has a white star on it 's chin. The man has a watch on his wrist. The shelf is a pale brown color. The shelf has a lot of books in it.

LLaVA:

The image features a man wearing glasses and a red shirt, sitting in front of a bookshelf. He is wearing headphones, likely listening to music or an audio program. The bookshelf is filled with numerous books of various sizes, indicating a diverse collection of reading materials. The man appears to be enjoying his time, possibly engaging in a leisurely activity or studying.

LLaVA-Directed:

The image features a man wearing a red shirt, sitting in front of a bookshelf. He is wearing glasses and has earbuds in his ears, indicating that he might be listening to music or an audio program. The bookshelf is filled with various books, showcasing a diverse collection of reading materials. The man appears to be engaged in a conversation or focused on something, possibly related to the books or the content he is listening to.

InstructBLIP:

The image features a man wearing glasses and a red shirt sitting in front of a bookshelf filled with various books. He is holding a microphone and speaking into it, possibly for a podcast or video recording. There are several books visible on the shelves behind him, ranging from the top to the bottom of the frame. In addition to the books, there are two cups placed on the shelves, one on the left side and another on the right side. The man's head is positioned slightly to the left of the center of the image, and his mouth is open as he speaks into the microphone. Overall, the scene depicts a man engaging in a podcast or video recording in front of a well-stocked bookshelf.

InstructBLIP-Directed:

In the image, a bald man wearing glasses is sitting in front of a bookshelf filled with various books. He is holding a microphone and speaking into it, possibly for a podcast or video recording. There are several books visible on the shelves, ranging from the top to the bottom of the frame. The man's head is positioned towards the left side of the image, while the bookshelf takes up most of the right side of the frame. In addition to the books, there are two cups placed on the shelves, one near the top and the other near the bottom.

FuseCap:

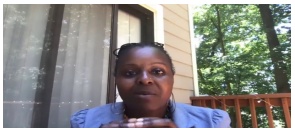
A picture of a bald man in a red shirt stands in front of a bookshelf displaying a variety of books, including a red book, a blue book, and a white book he.

VIIDA, VIIDA-Dense, and VIIDA-GRIT:

A midlife white man with surprised expression. He has blue eyes. He is bald. He is wearing a red shirt and white headphones. He is in the library, with books in the background. A wooden shelf with books. Yellow and yellow flowers. A pair of glasses. A colorful coffee mug. A stack of books.

Humanos:

Photograph of a white man with a serious face. The man is bald and he has brown eyes. He is wearing a red turtleneck shirt and he is wearing glasses and white headphones. In the back are shelves of books.



BLIP:

A woman in a blue shirt.

Baseline:

Trees behind the man. The mouth of a man. A man is wearing a necklace.

Regions-Hierarchical:

A young man is sitting in the window sill. The man is wearing a blue shirt. The man is standing in front of the window. The woman has brown hair. A little girl is sitting on the chair. The window is reflecting the window.

LLaVA:

The image features a woman with a nose piercing, wearing a white shirt and a necklace, sitting in front of a window. She appears to be engaged in a conversation or possibly giving a presentation. The woman is wearing earrings, and her hair is styled in a bun. The window behind her is open, allowing natural light to enter the room.

LLaVA-Directed:

The image features a woman with a nose piercing, wearing a blue shirt, and she is standing in front of a window. She appears to be engaged in a conversation or possibly giving a presentation. The woman is wearing large hoop earrings, which adds a stylish touch to her outfit. The environment around her seems to be a comfortable and casual setting, possibly at home or in a relaxed office space.

InstructBLIP:

The image depicts a woman sitting outside on a porch, looking directly at the camera. She is wearing a blue shirt and appears to be holding something in her hand, possibly a cell phone or a piece of food. There are several trees visible in the background, providing a natural setting for the woman's outdoor activity. In addition to the woman, there are two chairs placed on the porch, one on the left side and the other on the right side of the image. A railing can be seen running along the edge of the porch, providing support for the woman as she sits and enjoys her surroundings.

InstructBLIP-Directed:

In the image, a woman is sitting outside on a porch, looking directly at the camera. She is wearing a blue shirt and has a pair of earrings on her ears. The porch is surrounded by trees, which provide a natural backdrop for the scene. There are several chairs placed around the porch, giving the impression of a cozy outdoor living space. In addition to the woman, there are two other people visible in the image, one on the left side and another on the right side of the porch. They appear to be engrossed in their own activities, possibly enjoying the outdoor surroundings as well.

FuseCap:

A picture of a woman with black hair and a blue shirt smiles while looking out of a window, with a tall green tree in the background her brown eyes, large nose, and open mouth.

VIIDA, VIIDA-Dense, and VIIDA-GRIT:

A black woman with serious expression. She has brown eyes and black short hair. She is wearing a blue blouse, makeup and silver earrings. She is in the outside, with trees in the background. A wooden fence. Green leaves on tree.

Humanos:

A middle-aged woman with black curly hair. She has dark skin, black eyes, and a neutral facial expression. She is wearing a necklace, blue blouse, nose ring and hoop earrings. The woman is on a porch with several trees in the background and a glass door in the house.

BLIP:

A woman with glasses on.

Baseline:

White wall behind a man with a beard. A man wearing glasses. A man wearing a black hat. Hair is in the mans hand.

Regions-Hierarchical:

A young man is smiling and looking at a camera. He has a black tie around his neck. The man has a black nose and eyes.

LLaVA:

The image features a woman with dreadlocks, wearing glasses and a white hat. She is standing in a room, possibly a living room, and appears to be looking at the camera. The woman is wearing a white shirt and a colorful necklace, adding a vibrant touch to her outfit. The room has a cozy atmosphere, with a chair visible in the background.

LLaVA-Directed:

The image features a woman with long, curly hair and glasses, wearing a white hat. She is standing in a room, possibly a living area or a bedroom, with a door visible in the background. The woman appears to be looking at the camera, possibly posing for a picture. The room seems to be well-lit, with a light source above her head, which could be a ceiling fan or a light fixture.

InstructBLIP:

A woman with dreadlocks and glasses in a living room.

InstructBLIP-Directed:

In the image, a woman with long dreadlocks and glasses is standing in a living room. She is wearing a white shirt and a necklace, and there is a ceiling fan above her head. The living room appears to be well-furnished, with a couch and a chair visible in the background. The woman's facial expression is serious and focused, as if she is delivering a message or addressing an important topic. The overall atmosphere of the scene is calm and relaxed, with the ceiling fan adding a touch of serenity to the setting.

FuseCap:

A picture of a woman with long hair and black glasses stands in front of a white wall and ceiling, wearing a blue and white shirt and a red earring she looks directly at the camera.

VIIDA, VIIDA-Dense, and VIIDA-GRIT:

An elderly white woman with serious expression. She has black long hair. She is wearing a white shirt and white headband. She is in the living room, with a door in the background. A white door. A light on the wall.

Humanos:

Photograph of a middle-aged black woman with a blank facial expression. She has dreadlocks in her long hair and black eyes. The woman is wearing white turban, glasses, earrings, colorful necklace and she is wearing a blue blouse. She is sitting in a room with a ceiling fan and doors in the background.

BLIP:

A man wearing headphones.

Baseline:

A man in black jacket with beard. A man has glasses. A white wall behind a man. A man is holding a camera.

Regions-Hierarchical:

A man is wearing a black t shirt and black tie. He has a white beard and a white beard. He is wearing a white shirt and black shorts. The man has a dark colored beard and glasses. The man has a dark colored shirt on and a dark colored tie.

LLaVA:

The image features a man wearing headphones and a pink shirt. He is sitting in front of a wall, possibly in a room or a studio. The man appears to be focused on something, possibly listening to music or participating in a video call. The headphones are placed on his head, and he seems to be engaged in the activity or conversation.

LLaVA-Directed:

The image features a man wearing a pink shirt and glasses, sitting in front of a wall. He is wearing headphones, which suggests that he is either listening to music or participating in a video call. The man appears to be focused and engaged in his activity. In the background, there are two potted plants, one on the left side and the other on the right side of the man. These plants add a touch of greenery to the scene and create a more relaxed atmosphere.

InstructBLIP:

The image features a man wearing headphones and glasses, sitting in front of a purple background. He is smiling and looking directly at the camera. In addition to his headphones and glasses, he is also wearing a pink shirt. There are two potted plants visible in the background, one on the left and another on the right side of the screen.

InstructBLIP-Directed:

In the image, there is a man wearing headphones and glasses, sitting in front of a computer screen. He appears to be engaged in a video call with another person, who is also wearing headphones. The scene takes place in a room with a potted plant visible in the background. The man's headphones are positioned on his left ear, while the other person's headphones can be seen on their right ear. The presence of both individuals wearing headphones suggests that they are actively participating in the video call.

FuseCap:

A picture of a man with a black and brown beard, black glasses, and a brown and black mustache wearing black headphones stands in front of a white wall he wears a pink shirt and.

VIIDA, VIIDA-Dense, and VIIDA-GRIT:

A brown man with smile expression and glasses. He has brown eyes, black short hair, black mustache and black short beard. He is wearing a pink shirt and black headphones. He is in the office, with plants in the background. The shirt is black. A pair of glasses. A branch with leaves on it. A long green leaf.

Humanos:

A man with black skin and a blank facial expression. He has short black hair, black eyes and a full black beard. The man is wearing a pink shirt with a black blouse over it and he is wearing glasses and black headphones. In the background there are plants and a white wall.

Figura 6.9: Exemplos de parágrafos gerados pelos métodos comparados, exceto por Concat-{Dense, GRIT} e Concat-Filter-{Dense, GRIT}. As sentenças em roxo e azul na figura mostram o complemento feito por VIIDA-Dense e VIIDA-GRIT, respectivamente, em relação ao VIIDA. Fonte: Material próprio.

descrições geradas, especialmente em relação ao gênero do apresentador nas imagens.

Acredita-se que isso se deve principalmente ao conjunto de dados SIP utilizado, que compreende um número relativamente baixo de imagens (14.575) em seu conjunto de treinamento e cobre uma ampla variedade de conteúdos além de apenas pessoas e cenários (Krause et al. (2017)), o que faz com que o modelo não aprenda a distinguir bem esses casos.

Os métodos LLaVA e InstructBLIP, juntamente com suas variações, forneceram extensas descrições de imagem com um vocabulário mais amplo. Apesar de geralmente transmitirem o contexto apresentado, esses métodos fizeram muitas suposições sobre o conteúdo da mesma através do uso de palavras como "provavelmente", "possivelmente" e "parece". Acredita-se que essas suposições deveriam ser implícitas dado o tipo de imagem (por exemplo, de webinários, notícias, entrevistas, etc.).

O comportamento acima é conhecido como "alucinação de IA", um termo que se refere a resultados incorretos ou enganosos gerados pelos modelos (Liu et al. (2023)). Isso pode ocorrer por uma variedade de razões, como dados de treinamento insuficientes, ajuste excessivo do modelo, *prompt* incompleto, supervisão inadequada, ausência de referências visuais diretas e interrupção do conhecimento. A ocorrência dessas alucinações pode levar a descrições que não são visualmente congruentes com as características reais dos objetos representados (Huang et al. (2024)). Isso enfatiza a necessidade do refinamento contínuo dos modelos para aumentar a confiabilidade dos sistemas de IA em cenários práticos (Salvagno et al. (2023)).

Algumas características importantes como expressões faciais, faixa etária e cor da pele dos apresentadores foram frequentemente omitidas pelo LLaVA e InstructBLIP. Isso corrobora com os dados apresentados na Tabela 6.4, que mostra um baixo número de adjetivos e um alto número de outras características linguísticas. Além disso, esses modelos exibiram comportamentos como focar em detalhes não centrais para a imagem e expressar opiniões pessoais ou pontos de vista, o que não é aconselhável em descrições de imagem para indivíduos com deficiência visual (ABNT (2016); Lewis (2018)).

O InstructBLIP aparentou descrições mais detalhadas e o mesmo utilizou mais elementos linguísticos (por exemplo, vírgulas e conjunções coordenativas) para suavizar as transições entre as sentenças, provavelmente devido ao seu treinamento com um conjunto diversificado de dados de instrução. Sua versão direcionada foi a única entre os quatro métodos mencionados que descreveu mais características, como "*a bald man*" e "*the woman's facial expression is serious and focused*". Como o InstructBLIP passou por ajustes em suas instruções, fornecendo um *prompt* mais detalhado, esse método exibiu uma maior capacidade de generalização para dados e tarefas não vistas (Dai et al. (2023)). No entanto, é importante mencionar que implementar o InstructBLIP requisitou recursos computacionais e tempo de

execução elevados para realizar as inferências[17].

O FuseCap gerou sentenças mais próximas em comprimento em relação ao VIIDA e às escritas por humanos, além de transmitir o contexto da imagem em detalhes. No entanto, foi observado que os textos nas colunas 1 e 4 não estão corretamente finalizados, e há uma menor frequência de conjunções coordenativas e vírgulas nas sentenças. Esse padrão se repete em muitas descrições produzidas por este método. O problema com as sentenças que exibem esse comportamento é a falta de clareza e a má organização das informações. Essas sentenças são longas, contêm muitos detalhes e, por isso, são difíceis de entender. Além disso, o FuseCap também exibiu uma alta frequência dos termos "*open mouth*" e "*large nose*" em seus resultados. Portanto, o modelo se concentrou em conteúdos sem grande relevância em vez de, por exemplo, mencionar a faixa etária, cor da pele e emoção do apresentador.

Embora não sejam modelos específicos de ponta a ponta para a tarefa de geração de parágrafos de imagem ou texto, o VIIDA e suas variantes VIIDA-{Dense, GRIT} produziram textos mais semelhantes aos escritos por humanos em termos de elementos linguísticos e qualidade, como demonstrado pelas estatísticas na Tabela 6.4 e pelas pontuações das métricas de avaliação na Tabela 6.2. Além disso, destaca-se que esses textos produzidos possuem tamanhos semelhantes aos escritos por humanos e são capazes de descrever o contexto apresentado na imagem do webinário com detalhes e coesão.

As saídas do VIIDA mostrados nas Figuras 6.9 e 6.10 contêm partes destacadas em roxo e azul para indicar os complementos do DenseCap e GRIT, respectivamente. Como pode ser visto, uma melhor generalização é alcançada usando esses modelos de legendagem densa em comparação com o VIIDA sozinho (isto é, sem a abordagem complementar da etapa de Extração de Informações Visuais).

Além de incluir vírgulas, pronomes e conjunções coordenativas, os parágrafos do VIIDA e de suas variantes foram os únicos produzidos entre os métodos avaliados a mencionar uma ampla variedade de adjetivos, como faixa etária ("*midlife*" e "*elderly*"), cor da pele ("*white*", "*black*" e "*brown*") e expressão facial ("*surprised*", "*serious*" e "*smile*") dos apresentadores. O foco principal está sempre no apresentador; no entanto, elementos da cena, como o ambiente e alguns objetos, também são mencionados. Assim, o método proposto e suas variantes conseguiram cobrir o maior número possível de elementos importantes da cena sem aumentar significativamente o comprimento do parágrafo. Essas percepções corroboram os resultados da análise multicritério realizada pela métrica InViDe, disponível na Tabela 6.5, que evidenciam que os parágrafos gerados pelo VIIDA e suas variantes são os mais adequados para pessoas com deficiência visual, de acordo com padrões, orientações e boas práticas

[17]Foram necessários no mínimo 34 GB de memória RAM para execução do modelo e cada inferência em CPU durou cerca de dois a três minutos, conforme apresentado na Tabela 6.10.

recomendadas (ABNT (2016); Lewis (2018); Stangl et al. (2020)).

Quanto às menções de idade, cor da pele e emoção, reconhece-se que essas são as características mais desafiadoras de medir nesse tipo de imagem. Essa dificuldade decorre do fato de que as imagens capturam cenas onde os indivíduos podem estar envolvidos em ações como movimento ou fala. Além disso, fatores como qualidade da Internet, resolução da imagem e condições do ambiente também podem interferir na precisão das medições. No entanto, destaca-se a importância de mencionar esses atributos para uma melhor compreensão do contexto por parte dos usuários com deficiência visual.

A inclusão dessas características ajuda a mitigar a injustiça algorítmica e os vieses presentes em ferramentas de IA desse tipo, causados por desequilíbrios estatísticos na representação de grupos diversos, com base em gênero, idade, cor da pele e localização geográfica nos conjuntos de dados usados para treinar os modelos (Garcia et al. (2023); Naik and Nushi (2023); Pagano et al. (2023); Dominguez-Catena et al. (2024)). Portanto, destaca-se a importância dessa contribuição, visto que o VIIDA e suas variantes foram os únicos (como demonstrado na Figura 6.9) a produzir parágrafos contendo esse tipo sensível de conteúdo.

É importante destacar que, embora o VIIDA e suas variantes tenham sido desenvolvidos para descrever cenas de webinários, eles não se limitam apenas à descrição de ambientes internos. Esses métodos também funcionam bem em cenas de ambientes externos, conforme demonstrado na coluna 2 da Figura 6.9, onde o ambiente externo ("*outside*") é mencionado. Outros exemplos de ambientes externos encontrados nos resultados gerados incluem "*urban*" e "*park*".

Como mencionado anteriormente, os indivíduos presentes nas imagens podem estar envolvidos em movimentos. No entanto, como entre as 23 perguntas fornecidas ao modelo VQA não foi elaborada nenhuma referente à ação presente na imagem, essa informação não será considerada na formação final do parágrafo, a menos que o movimento seja capturado pelo modelo de legendagem densa e permaneça até as últimas etapas do método descritivo. Por fim, vale ressaltar que a ação do apresentador na cena do webinário não é considerada uma característica de interesse relevante, já que não desempenha um papel importante para o entendimento do contexto visual.

Nem todas as saídas do VIIDA e suas variantes são completamente precisas. Conforme apresentado na Figura 6.9, as colunas 1 e 4 destacam algumas falhas: olhos azuis relatados para um homem com olhos castanhos; a presença de óculos mencionada duas vezes; e sentenças malformadas resultantes da concatenação das saídas dos modelos de legendagem densa. As principais razões para essas falhas incluem erros propagados pelos modelos usados no *pipeline* e os limiares adotados. Apesar dessas falhas, os resultados ainda são inteligíveis. Graças à natureza flexível

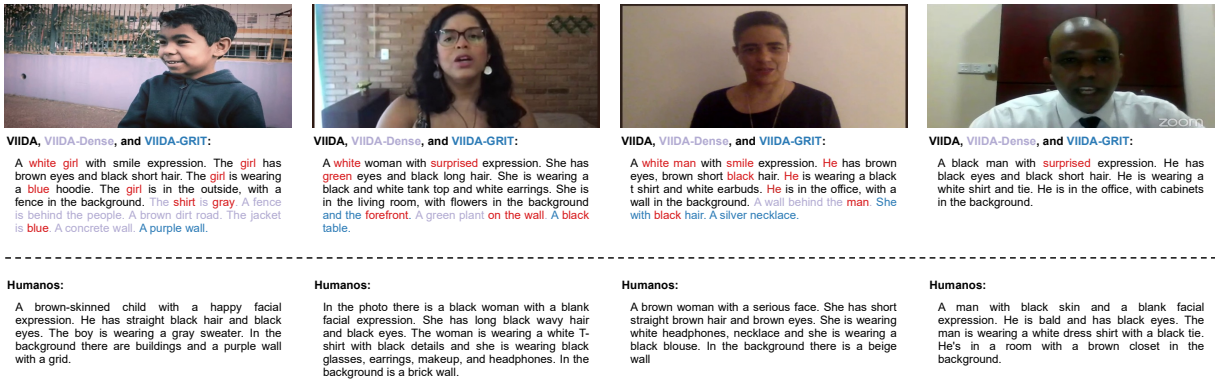


Figura 6.10: Exemplos de parágrafos gerados pelo VIIDA e suas variantes com diferentes casos de falha. Palavras em vermelho são casos falhos. As sentenças em roxo e azul na figura mostram o complemento feito por VIIDA-Dense e VIIDA-GRIT, respectivamente, em relação ao VIIDA. Fonte: Material próprio.

do método proposto e de suas variantes, é possível minimizar esses erros no futuro substituindo os modelos e ferramentas atuais por abordagens com maior precisão.

A Figura 6.10 exibe alguns dos casos de falhas gerados pelo VIIDA e suas variantes. Alguns erros são evidentes ao examinar diretamente as imagens, enquanto outros se tornam aparentes ao compará-los com os parágrafos escritos por humanos. Na primeira imagem, um menino é descrito erroneamente como uma menina (*girl*) devido à saída incorreta do modelo VQA. Ao analisar as legendas densas concatenadas geradas pelo DenseCap, observa-se que o modelo utilizou a palavra "*people*" em uma das sentenças. Um comportamento semelhante ocorreu na terceira imagem, onde o modelo VQA atribuiu o substantivo "*man*" à pessoa na imagem em vez de "*woman*", resultando em um uso incorreto dos pronomes no restante do texto. Além disso, foi observado que o GRIT usou o pronome "*she*" em uma de suas sentenças, indicando uma discrepância entre os modelos na identificação de gênero.

Na segunda imagem, é possível observar um exemplo de falha vinculada ao algoritmo proposto na etapa de produção de parágrafos (Seção 6.1.2). Neste caso, a palavra "*forefront*" na sentença "*she is in the living room, with flowers in the background and the forefront*" foi usada incorretamente, tornando a sentença incompleta ou mal estruturada. Esse erro ocorreu devido às relações gramaticais extraídas ("*woman_in*") pela ferramenta de grafo de cena. Ao avançar para a etapa do algoritmo que utiliza a estratégia baseada no agrupamento de informações gramaticais para estruturar o parágrafo, a relação extraída já estava presente. Assim, o substantivo "*forefront*" associado à "*woman*" foi incorporado à sentença com a relação gramatical "*woman_in*" já processada, resultando em uma sentença desconexa.

Na quarta imagem, há um exemplo em que o VIIDA não produziu parágrafos distintos com as saídas do DenseCap e do GRIT, resultando no mesmo parágrafo para os três métodos (VIIDA, VIIDA-Dense e VIIDA-GRIT). Além disso, foi observada uma

frequência maior das expressões "*surprised*" e "*serious*" nos parágrafos. Esse fenômeno pode ser atribuído ao desafio de inferir emoções em imagens que retratam pessoas em conversas.

Avaliação das imagens sintéticas

Ao analisar os resultados na Figura 6.11, observa-se que as imagens na coluna da esquerda são as que mais diferem das imagens originais localizadas na parte inferior da figura (conjunto de dados de webinários). Essas observações são consistentes com os valores de similaridade do cosseno e as pontuações FID, conforme mostrado na Tabela 6.6.

O BLIP, com suas legendas simples, genéricas e sem detalhes, permitiu que os modelos generativos produzissem imagens com características aprendidas a partir dos dados de treinamento. Por outro lado, os métodos Concat-{Dense, GRIT}, ao gerarem textos com inúmeras sentenças, transmitiram uma quantidade significativa de informações para a geração de imagens. Como resultado, os modelos generativos enfrentaram dificuldades para criar imagens que se assemelhassem de perto às originais.

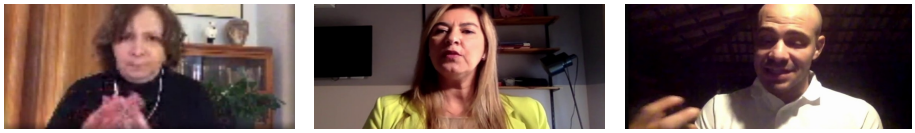
Os métodos Concat-Filter-{Dense, GRIT} filtram legendas densas que não estão associadas a pessoas, direcionando melhor o contexto dos textos e, consequentemente, apresentando resultados mais precisos em comparação com suas versões não filtradas (Concat-{Dense, GRIT}). Os resultados inesperados do *baseline* podem ser atribuídos aos erros propagados e ao modelo de sumarização utilizado, enquanto os resultados do Regions-Hierarchical podem ser relacionados ao conjunto de dados e à arquitetura empregada pela Rede Neural Recorrente Hierárquica (Krause et al. (2017)).

A discrepância semântica entre as imagens sintéticas e as originais começa a diminuir após a geração das imagens das descrições textuais do LLaVA. Essa aproximação sugere que as imagens sintéticas se tornam mais semelhantes às originais em termos de interpretação semântica. Isso indica uma melhoria na qualidade das descrições retornadas pelos métodos avaliados, o que influenciou positivamente a capacidade dos modelos generativos de criar imagens sintéticas que se alinham de forma mais precisa com as imagens originais, tanto em termos de conteúdo quanto de significado.

As imagens geradas pelas descrições dos modelos baseados em LLMs ainda apresentam discrepâncias devido ao fenômeno das alucinações. No entanto, entre esses modelos, o InstructBLIP se destacou por apresentar os resultados semânticos que possibilitaram a criação de imagens sintéticas mais próximas das imagens originais. As saídas do FuseCap, por outro lado, geraram imagens que se concentraram predominantemente em pessoas e apresentaram cenários mais

Figura 6.11: Exemplos de imagens sintéticas geradas pelos modelos generativos com base nas saídas produzidas pelos competidores, bem como nas descrições escritas por humanos (conjunto de dados de webinários). As saídas produzidas pelos competidores, utilizadas para gerar as imagens sintéticas, podem ser visualizadas na Figura 6.12. Ressalta-se que as imagens originais, localizadas na parte inferior da figura, aparecem visualmente mais estreitas devido ao redimensionamento aplicado. Fonte: Material próprio.

simples. Portanto, as imagens resultantes de suas saídas foram as que mais se assemelharam a uma imagem de webinário quando comparadas com as imagens dos



BLIP:

A woman in a black shirt.

Baseline:

White wall behind the man. A man with brown hair. A man wearing a shirt. He is looking at the camera.

Regions-Hierarchical:

A man is sitting in a room with a white shirt. He is looking down at a man in a black sweater. The window has a cartoon character. The window has a light brown wood door. The window is dark brown. The house is brown and has a brown roof.

LLaVA:

The image features a woman wearing a black shirt and a necklace. She is sitting in a chair and appears to be clapping her hands. The woman is surrounded by numerous books, which are scattered around her in various positions. Some books are located near her, while others are placed further away. The scene suggests that the woman might be in a library or a personal space with a strong interest in reading.

LLaVA-Directed:

The image features a woman with curly hair, wearing a black shirt and a necklace. She is standing in a room with a bookshelf filled with books. The woman appears to be clapping her hands, possibly in a celebratory or enthusiastic manner. The room also contains a potted plant, adding a touch of greenery to the space.

InstructBLIP:

The image depicts a woman sitting in front of a camera, wearing a black shirt. She is holding a piece of meat in her right hand and appears to be about to take a bite out of it. In the background, there is a potted plant located on the left side of the room. A book can be seen on the right side of the room, close to the woman's head. Additionally, there are two clocks visible in the scene, one on the left side and another on the right side of the room.

InstructBLIP-Directed:

In the image, a woman wearing a black shirt is sitting in front of a computer. She is holding a piece of meat in her right hand, which appears to be a pork chop. The woman's left hand is resting on her lap, and she is looking directly at the camera. Behind her, there is a bookshelf with several books visible. A potted plant can also be seen in the background, adding a touch of nature to the scene. The overall setting suggests that the woman might be preparing or cooking the pork chop, possibly for a meal or snack.

FuseCap:

A picture of a woman with long brown hair stands in front of a white wall, wearing a black shirt and a silver necklace a green plant sits on the left side of the image, while.

VIIDA, VIIDA-Dense, and VIIDA-GRIT:

A midlife white woman with serious expression. She has brown eyes and brown short hair. She is wearing a black sweater, makeup and silver necklace. She is in the living room, with pictures in the background. A green plant by the window. A vase on a shelf.

Humanos:

Photograph of a middle-aged woman with fair skin and a serious face. She has brown, straight, short hair and brown eyes. She is wearing a necklace and she is wearing a black blouse. The woman is in a room where there is a bookcase with books, a painting hanging on the white wall, a brown curtain and a vase with flowers.

BLIP:

A woman in a yellow jacket.

Baseline:

A picture of a woman with long brown hair. Woman wearing a white shirt

Regions-Hierarchical:

A woman is looking at her cell phone. She is wearing a white dress. She is wearing a white dress shirt and a black tie. She is holding a white wii remote in her hand. She is wearing a black and white shirt and black pants. The wall behind the woman is white .

LLaVA:

The image features a woman with long, blonde hair sitting in a room. She is wearing a yellow jacket and appears to be engaged in a conversation or presentation. The room has a bookshelf in the background, which is filled with various books of different sizes. The woman is the main focus of the scene, and her presence and attire create a professional and confident atmosphere.

LLaVA-Directed:

The image features a woman with long, blonde hair, wearing a yellow jacket. She appears to be sitting in a room, possibly in front of a bookshelf. The woman is looking at the camera, possibly posing for a picture. The room has a few books scattered around, with some placed on the bookshelf and others on the floor. The woman's outfit and the presence of books suggest that she might be in a professional or academic setting.

InstructBLIP:

The image features a woman with long blonde hair wearing a yellow jacket. She is standing in front of a bookshelf and appears to be speaking into a microphone. There are several books on the bookshelf, with some placed closer to the woman and others further away. In addition to the books, there are two televisions visible in the background, one on the left and another on the right side of the room. The woman's facial expression suggests that she may be delivering a speech or making an announcement.

InstructBLIP-Directed:

In the image, a woman wearing a yellow jacket is standing in front of a bookshelf. She is looking directly at the camera, making eye contact with the viewer. The bookshelf is positioned to the left of the woman, and there are several books visible on the shelves. A television is also present in the scene, placed on the right side of the room. The environment appears to be a well-organized and cozy living space.

FuseCap:

A picture of a woman with long hair and a yellow shirt stands in front of a white wall, with a wood shelf in the background her face, with a large nose and brown eyes,.

VIIDA, VIIDA-Dense, and VIIDA-GRIT:

A white woman with confused expression. She has brown eyes and blonde long hair. She is wearing a yellow blouse and makeup. She is in the living room, with a bookshelf in the background. Black television on wall. Flat screen television on wall.

Humanos:

A Caucasian female with a serious facial expression. She has long straight blonde hair and brown eyes. The woman is wearing a beige shirt with a green blazer on top and she is wearing makeup and earrings. She is in a room with an air conditioner, a television and shelves with objects on the back wall.

BLIP:

A man in a white shirt.

Baseline:

Man with glasses on face. A red wall behind the man. The man is wearing a black shirt. His hair is black.

Regions-Hierarchical:

A man is sitting on the table with his cell phone. He is wearing a black shirt with white buttons. The face of the man has a black face. The man has light brown hair and is looking at the camera. There is a white and brown bear in the background. The man is wearing a light brown shirt and white pants.

LLaVA:

The image features a bald man with a beard, wearing a white shirt, standing in a dark room. He appears to be looking at the camera, possibly posing for a picture. The man is the main subject of the image, and his facial expression suggests that he is engaged in a conversation or making a statement. The room's dark atmosphere adds a sense of mystery to the scene.

LLaVA-Directed:

The image features a bald man wearing a white shirt, standing in a dark room. He appears to be engaged in a conversation or giving a speech, as he is looking at the camera. The man is the main focus of the scene, and his bald head is a noticeable feature. The room seems to be a dimly lit space, which adds to the seriousness of the man's demeanor.

InstructBLIP:

In the image, a bald man is wearing a white shirt and smiling. He is standing in front of a wooden structure, possibly a bar or a restaurant. The man's face is close to the camera, making it appear as if he is speaking directly to the viewer. There are several other people visible in the scene, including one person sitting on the left side of the screen and another person sitting on the right side of the screen. Additionally, there are two more people standing behind the bald man, one on the left and one on the right side of the frame.

InstructBLIP-Directed:

In the image, a bald man is wearing a white shirt and smiling. He is standing in front of a wooden structure, which appears to be a barn or a shed. The man's face is close to the camera, making it possible to see his facial expression clearly. There are several other people visible in the background, but they are not as prominent as the bald man in the foreground. The environment appears to be outdoors, possibly in a rural or countryside setting.

FuseCap:

A picture of a bald man with a large nose and open mouth wearing a white shirt and collar, with a white ear and hand visible.

VIIDA, VIIDA-Dense, and VIIDA-GRIT:

A white man with smile expression. He is bald. He is wearing a white shirt. He is in the bar, with shelves in the background. A wall behind the man. White letters on the wall. He is in the foreground.

Humanos:

Photograph of a bald white man. The man is serious. He is wearing a white collared shirt and his eyes are brown. The man is making hand gestures. The background is night.

Figura 6.12: As saídas produzidas pelos competidores e aquelas escritas por humanos (conjunto de dados de webinários) foram fornecidas como entrada para os modelos generativos para criar as imagens sintéticas apresentadas na Figura 6.11, exceto para os métodos Concat-{Dense, GRIT} e Concat-Filter-{Dense, GRIT}. As sentenças em roxo e azul na figura mostram o complemento feito por VIIDA-Dense e VIIDA-GRIT, respectivamente, em relação ao VIIDA. Fonte: Material próprio.

modelos anteriores, conforme mostrado na Figura 6.11.

Em geral, o VIIDA e suas variantes produziram textos que facilitaram a geração de imagens com características mais similares às das imagens originais, especialmente o VIIDA-GRIT. Acredita-se que a estrutura direta e coesa de seus parágrafos, juntamente com as características-chave mencionadas, ao serem fornecidas como *prompts* de entrada para os modelos generativos, ajudaram na criação de imagens semanticamente mais semelhantes às originais. Outros exemplos podem ser vistos na Figura 6.13.

A constatação acima está alinhada com os resultados da métrica CLIPScore, apresentados na Tabela 6.2, que mede o alinhamento textual diretamente com as imagens por meio de incorporações contextuais. Além disso, dentro desse contexto, destaca-se a capacidade do VIIDA e suas variantes de gerar descrições a partir de imagens com baixa qualidade, como exibido pelos exemplos, especialmente o primeiro. Isso se deve ao uso do modelo VQA, que emprega perguntas direcionadas a ROIs. Consequentemente, essa abordagem demonstrou maior eficácia em comparação com os métodos de extração de características utilizados por outros modelos ao analisar a imagem inteira.

As imagens na última linha da segunda coluna foram criadas usando as descrições escritas por humanos do conjunto de dados de webinários apresentado no Capítulo 5. Como era de se esperar, as imagens sintéticas resultantes exibem

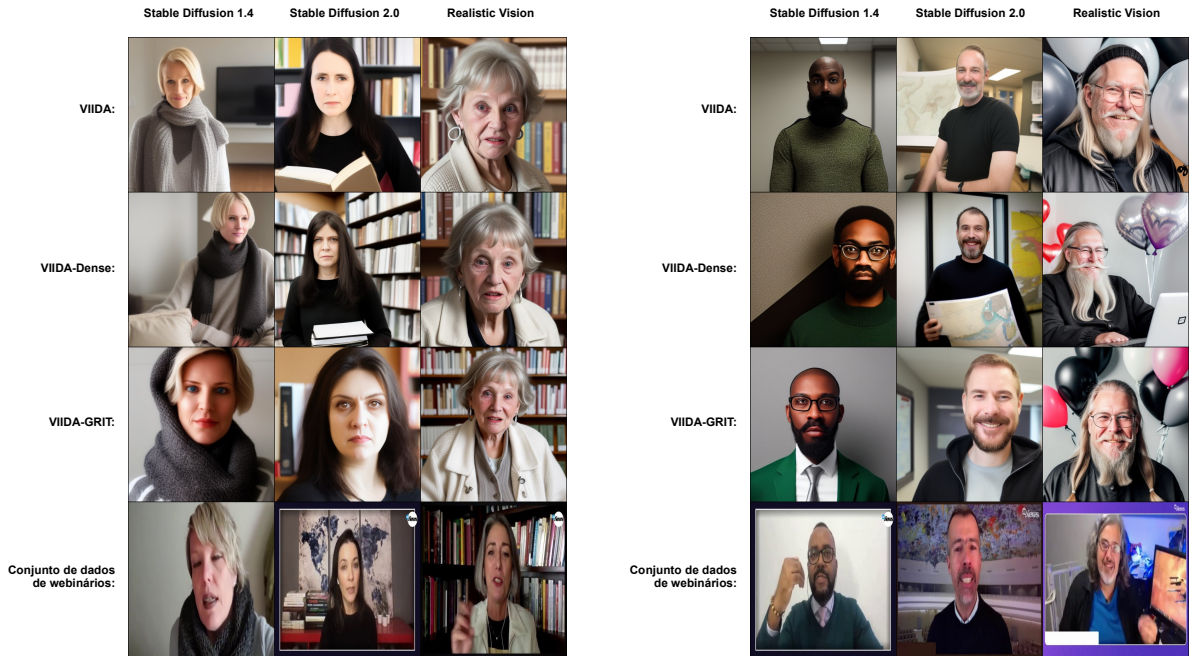


Figura 6.13: Outros exemplos de imagens sintéticas geradas com base nas saídas produzidas pelo VIIDA e suas variantes. Ressalta-se que as imagens originais, localizadas na parte inferior da figura, aparecem visualmente mais estreitas devido ao redimensionamento aplicado. Fonte: Material próprio.

características semelhantes às das originais, indicando uma alta semelhança semântica. No entanto, em termos de qualidade visual, os exemplos de imagens geradas pelo VIIDA-GRIT são superiores.

Dado que os parágrafos escritos por humanos têm comprimentos maiores e estruturas diferentes, como pode ser visto na Figura 6.12, deduz-se que o tamanho do *prompt* e o posicionamento das palavras dentro dele estão diretamente relacionados à atenção e capacidade generativa dos modelos. Isso implica na habilidade desses modelos de fazer previsões, discernindo padrões nos dados, o que ajuda a explicar a variância nos resultados. Ressalta-se também que esses parágrafos passaram pelo processo de tradução do português para o inglês e, mesmo com o processo de curadoria, essa tradução pode impactar a estrutura dos parágrafos.

Os resultados qualitativos deste experimento corroboram as conclusões discutidas na Seção 6.4.1, mostrando que as imagens sintéticas geradas a partir de parágrafos produzidos pelo VIIDA e VIIDA-{Dense, GRIT} se assemelham de perto às imagens originais, tanto em termos de similaridade semântica quanto na distribuição estatística de características. Portanto, este experimento adicional complementa a validação estatística realizada anteriormente.

Em última análise, esses resultados destacam que, no contexto de descrever imagens para indivíduos com CBV, o método proposto nesta tese pode ser de grande ajuda. Ele facilita a compreensão do contexto de aplicações de videoconferência, como webinários. O VIIDA proporciona uma percepção abrangente do apresentador e dos elementos relevantes na cena por meio de descrições adaptadas, aumentando a inclusão e reduzindo as limitações de acessibilidade em tais aplicações. Consequentemente, o VIIDA pode ser aplicado para melhorar a experiência de aprendizado de indivíduos com deficiência visual.

Avaliação do conteúdo NSFW

A popularidade e o progresso da IA, juntamente com o amplo acesso e facilidade de uso dessas ferramentas para geração de conteúdo, têm levantado preocupações sobre suas implicações éticas e sociais, bem como seu uso responsável (Bansal et al. (2022); Figueiredo et al. (2023)). Diante desse cenário, foi conduzida uma análise do conteúdo gerado neste capítulo.

Inicialmente, verificou-se automaticamente todas as saídas textuais geradas pelos 15 métodos mencionados na Seção 6.3 por meio de um modelo ajustado (DistilRoBERTa) para a classificação de conteúdo NSFW. O DistilRoBERTa foi projetado para identificar conteúdos NSFW em *prompts* textuais e foi treinado em um conjunto de dados composto por aproximadamente dois milhões de *prompts*, igualmente divididos entre as categorias NSFW e "Seguro para o Trabalho" (do

inglês, "*Safe For Work*" - SFW).

Como resultado, apenas os métodos VIIDA-Dense, FuseCap e LLaVA-Directed não tiveram nenhum parágrafo classificado como NSFW. O BLIP, juntamente com Regions-Hierarchical e o *baseline*, apresentaram as maiores quantidades de conteúdos classificados como NSFW, com 11, 12 e 18 parágrafos, respectivamente. Todos os outros métodos tiveram, no máximo, quatro parágrafos classificados como NSFW.

Em relação ao VIIDA e VIIDA-GRIT, cada um produziu apenas uma saída classificada como NSFW. Ao analisar individualmente esses parágrafos, percebeu-se que as classificações se deviam às palavras regata (*tank top*) e biquíni (*bikini*), geradas pelo BLIP (modelo VQA) e GRIT (modelo de legendagem densa), presentes nas sentenças que compõem os parágrafos. Quando essas palavras foram removidas, os parágrafos foram classificados como SFW. Isso destaca um viés implícito no modelo de classificação de conteúdo NSFW, associando palavras relacionadas a roupas femininas com conteúdo sexual.

De acordo com a documentação do DistilRoBERTa, esse modelo avalia a probabilidade de um *prompt* ser NSFW com base em ocorrências estatísticas. Portanto, a precisão do DistilRoBERTa tende a aumentar com o comprimento do *prompt*, e *prompts* mais curtos podem estar sujeitos a avaliações menos precisas. Contudo, apesar de algum conteúdo do VIIDA e de sua variante ter sido classificado como NSFW, observou-se que a quantidade é extremamente baixa em comparação com o número total de parágrafos produzidos por cada método.

Embora o DistilRoBERTa tenha demonstrado uma precisão satisfatória, também foi realizada uma análise para verificar se as imagens sintéticas geradas com base nas saídas dos métodos avaliados continham conteúdo explícito ou inadequado. Devido ao grande número de imagens produzidas pelos três modelos generativos baseados em difusão, a análise foi limitada aos resultados gerados pelo VIIDA e suas variantes.

Ao aplicar um modelo de classificação de conteúdo NSFW às imagens sintéticas – um ViT ajustado, conforme mencionado na Seção 6.3 – constatou-se que 25 de 6.156 imagens analisadas dos três métodos examinados foram classificadas como inadequadas, com uma probabilidade média de 72,58%. Esses resultados foram posteriormente revisados usando uma API de reconhecimento de conteúdo sexual em imagens, que atribuiu uma probabilidade média de 73,22% para conteúdo NSFW em 14 das 25 imagens inicialmente classificadas.

Entre os três modelos generativos, o RV produziu 13 imagens, das quais 10 foram duplamente classificadas como NSFW. O SD1.4 seguiu com 4 imagens entre as 9 inicialmente classificadas. O SD2.0 foi o único modelo que não teve nenhuma das 3 imagens duplamente classificadas como NSFW pela API. Isso se deve à menor ocorrência de conteúdo pornográfico explícito no conjunto de dados de treinamento usado para esta versão do Stable Diffusion, em comparação com a versão 1.4.

Figura 6.14: Exemplos de imagens sintéticas duplamente classificadas como NSFW, juntamente com seus respectivos parágrafos. As imagens no canto superior esquerdo, superior direito, inferior esquerdo e inferior direito foram geradas a partir das saídas produzidas pelos métodos e modelos: VIIDA e SD1.4, VIIDA-GRIT e RV, VIIDA-GRIT e RV, e VIIDA-GRIT e SD1.4, respectivamente. Fonte: Material próprio.

Os dois parágrafos classificados como NSFW, gerados por VIIDA e VIIDA-GRIT na análise anterior, também resultaram em imagens classificadas como NSFW (como, por exemplo, a imagem no canto inferior esquerdo da Figura 6.14). Notavelmente, das 25 imagens classificadas pelo ViT ajustado como NSFW, 24 continham mulheres, enquanto apenas uma imagem apresentou um homem. Esse achado ressalta novamente o viés na associação entre mulheres e conteúdo sexual presente nos conjuntos de dados (por exemplo, LAION-2B e LAION-Aesthetics) utilizados para treinar os modelos generativos e os classificadores de conteúdo NSFW.

A Figura 6.14 apresenta alguns exemplos de imagens geradas a partir das saídas produzidas pelo VIIDA e suas variantes que foram duplamente classificadas como

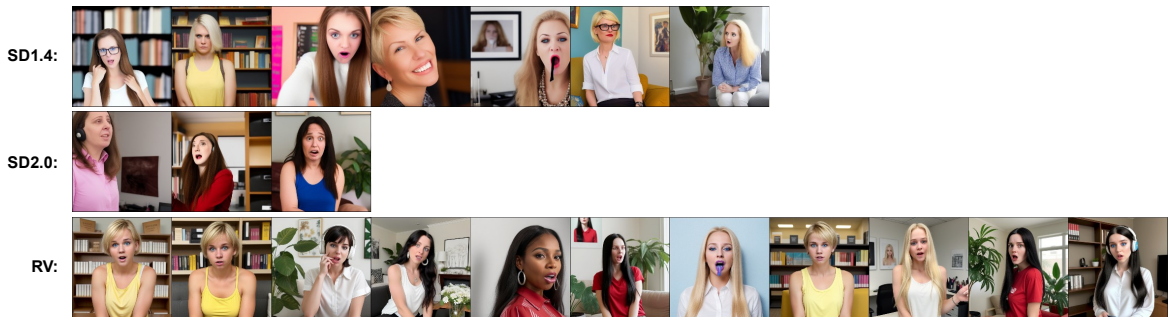


Figura 6.15: Outras imagens sintéticas classificadas como NSFW geradas pelos modelos Stable Diffusion 1.4 (SD1.4), Stable Diffusion 2.0 (SD2.0) e Realistic Vision (RV). Fonte: Material próprio.

NSFW juntamente com seus parágrafos. As demais imagens sintéticas inicialmente classificadas como NSFW podem ser vistas na Figura 6.15.

Considerações éticas

Os modelos baseados em LLMs geralmente são treinados em conjuntos de dados em grande escala e nem sempre passam por um processo rigoroso de filtragem e curadoria. Consequentemente, esses modelos podem herdar alguns dos problemas associados aos LLMs (por exemplo, LLaMA e Vicuna) e aos codificadores de visão (por exemplo, modelo CLIP), como a replicação de vieses inerentes e conteúdo explícito presente nos dados usados durante o treinamento nos resultados gerados (Liu et al. (2023); Rotstein et al. (2024)).

Conforme apresentado na seção anterior, os resultados obtidos com o VIIDA e suas variantes não produziram nenhum conteúdo verdadeiramente classificado como NSFW. Os casos identificados como NSFW foram atribuídos a vieses nos modelos utilizados. Embora esse tipo de análise tenha sido conduzido, é importante salientar que, como o VIIDA gera sentenças com base em respostas a perguntas enviadas a um modelo VQA (conforme apresentado na Seção 6.1.1), não foram formuladas perguntas que pudessem resultar em conteúdo NSFW.

Para confirmar que a abordagem proposta é segura, foram realizados experimentos adicionais com a pergunta sobre roupas, utilizando a seguinte questão: "*what top clothing is the person wearing?*". Ao fornecer imagens de pessoas sem roupas na parte superior do corpo ou pornografia explícita, "*none*" foi retornado. Explorou-se também a adição de outras perguntas, como "*what is the person wearing?*", "*what is the person doing?*", "*is the image not safe for work?*", "*is the person shirtless?*", "*is the person wearing any clothes?*", "*is the person naked?*", e "*is there pornography in the image?*". Para a primeira pergunta, a resposta incluiu algum tipo de roupa ou acessório, ou a palavra "*nothing*". Para a segunda pergunta, ações como "*posing*" ou "*taking photo*" foram retornadas. As perguntas restantes simplesmente retornaram como respostas "*yes*" ou "*no*".

Com base nos experimentos realizados, conclui-se que, sem o uso de perguntas específicas e sentenças pré-estruturadas para assuntos NSFW, o VIIDA não produzirá conteúdo potencialmente prejudicial. No entanto, não se pode garantir que os modelos de legendagem densa adicionados na abordagem complementar não gerem esse tipo de conteúdo, especialmente o GRIT (Nguyen et al. (2022)), que utilizou outros conjuntos de dados de treinamento além do Visual Genome (Krishna et al. (2017b)), que é anotado por humanos. Para evitar esses casos, métodos ou filtros específicos podem ser incorporados aos *pipeline* aprimorado para filtrar conteúdo NSFW, por exemplo, ao analisar a figura de entrada ou o parágrafo de

saída.

Apesar dos pequenos riscos envolvidos, acredita-se que as vantagens de nosso método tanto para a comunidade de deficientes visuais quanto para a comunidade científica superam qualquer dano potencial. O VIIDA não apenas permite uma investigação adicional e avanços neste campo de pesquisa, mas também incentiva o desenvolvimento de novas aplicações e Tecnologias Assistivas que utilizam tarefas de visão e linguagem.

6.5 Estudos de Ablação

Para investigar a influência e a contribuição de cada módulo do método proposto, ou seja, do VIIDA e de suas variantes, foram realizados estudos de ablação. A Tabela 6.7 apresenta as variações dos métodos comparados e os respectivos módulos isolados. Como o VIIDA-{Dense, GRIT} compartilham a mesma metodologia, os experimentos foram conduzidos apenas com o VIIDA-GRIT, uma vez que este apresentou melhores resultados na maioria dos casos.

Conforme a visão geral apresentada na Figura 6.1, analisou-se a influência dos seguintes módulos do VIIDA: Gerador de Sentenças (Linha 1 da Tabela 6.7), Extração de Relações e Estruturação de Parágrafos (Linha 2), e Substituição por Pronomes (Linha 3). Para a variante VIIDA-GRIT, foram investigados ambos os módulos de Verificação de Similaridade da abordagem complementar da etapa de Extração de Informações Visuais, diferenciando-os pelo modelo linguístico (Linha 5) e pelo VQA (Linha 6). As Linhas 4 e 7 representam, respectivamente, o VIIDA e o VIIDA-GRIT, que geram parágrafos considerando todos os módulos de suas abordagens. A sequência de caracteres "-s/" após o nome dos métodos refere-se à abreviação de "sem". Por exemplo, VIIDA-s/Sentenças significa o VIIDA sem o módulo responsável pela geração de sentenças. Os demais nomes dos métodos

Tabela 6.7: Módulos isolados no VIIDA e VIIDA-GRIT durante os estudos de ablação. No cabeçalho, as colunas "Sentenças", "Similaridade 1", "Similaridade 2", "Relações-Parágrafo" e "Pronomes" referem-se, respectivamente, à remoção dos módulos "Gerador de Sentenças", "Verificação de Similaridade" (o primeiro, com modelo linguístico), "Verificação de Similaridade" (o segundo, com modelo VQA), "Extração de Relações e Estruturação de Parágrafo" e "Substituição por Pronomes".

Método / Módulos	Sentenças	Similaridade 1	Similaridade 2	Relações-Parágrafo	Pronomes
VIIDA-s/Sentenças				✓	✓
VIIDA-s/Parágrafo	✓				✓
VIIDA-s/Pronomes	✓			✓	
VIIDA	✓			✓	✓
VIIDA-GRIT-s/Simil.Linguística	✓		✓	✓	✓
VIIDA-GRIT-s/Simil.VQA	✓	✓		✓	✓
VIIDA-GRIT	✓	✓	✓	✓	✓

Tabela 6.8: Comparação entre métodos durante os estudos de ablação usando as métricas clássicas e recentes para medir a similaridade entre os parágrafos candidatos e os de referência. Os valores são a mediana multiplicada por 100. B, MT, CDr, BTS, CS e RCS referem-se a BLEU, METEOR, CIDEr, BERTScore, CLIPScore e RefCLIPScore, respectivamente.

Método	**B-1**	**B-2**	**B-3**	**B-4**	**MT**	**CDr**	**BTS**	**CS**	**RCS**
VIIDA-s/Sentenças	9,82	0,00	0,00	0,00	8,88	0,00	47,97	60,01	54,15
VIIDA-s/Parágrafo	48,82	34,01	22,88	13,71	22,74	19,01	78,75	68,07	69,87
VIIDA-s/Pronomes	47,21	30,91	21,28	15,11	22,65	6,36	82,16	64,31	68,41
VIIDA	*46,76*	*30,40*	*19,86*	*12,11*	*22,02*	*2,38*	*82,41*	*65,84*	*69,21*
VIIDA-GRIT-s/Simil.Linguística	56,30	35,23	22,53	13,13	24,40	23,37	77,54	67,19	70,56
VIIDA-GRIT-s/Simil.VQA	54,38	34,31	21,46	12,48	24,88	18,96	73,66	66,80	69,53
VIIDA-GRIT	*54,36*	*34,23*	*21,81*	*12,91*	*23,67*	*17,70*	*78,72*	*66,94*	*70,31*

seguem essa mesma lógica.

A Tabela 6.8 apresenta os resultados dos estudos de ablação, expressos pela mediana[18]. Ao analisar os resultados do VIIDA-s/Sentença, percebe-se claramente que a remoção do módulo responsável pela geração de sentenças com base nas saídas do modelo VQA resultou em uma queda significativa nas pontuações de todas as métricas de avaliação. Isso sugere que a estruturação das sentenças é crucial para o desempenho do VIIDA e suas variantes.

Observando os resultados do VIIDA-s/Parágrafo e do VIIDA-s/Pronomes, percebe-se que, para a maioria das métricas avaliadas, as pontuações aumentaram em vez de diminuir com o isolamento dos módulos, principalmente na métrica CIDEr. O aumento nas pontuações, especialmente nas métricas clássicas baseadas em *n*-gramas que avaliam a correspondência entre textos, pode ser atribuído à maior quantidade de palavras e, consequentemente, de caracteres presentes nos parágrafos gerados pelos métodos investigados, que mais se aproximam dos parágrafos anotados por humanos, como ilustrado na Tabela 6.9. Isso resulta em um maior número de correspondências entre as palavras dos textos candidatos e dos textos de referência.

Além disso, como a métrica BLEU aplica uma penalização de brevidade, textos candidatos mais curtos do que os textos de referência são penalizados. Portanto, é esperado que o VIIDA receba pontuações menores em relação ao VIIDA-s/Parágrafo e VIIDA-s/Pronomes nessa métrica específica por apresentar textos mais curtos em comparação aos textos anotados por humanos. No entanto, essa diferença nas pontuações é menos acentuada ao analisar as métricas baseadas em dados. Por exemplo, como a BERTScore utiliza *embeddings* contextuais e pondera a importância

[18]A mediana foi utilizada para resumir as pontuações em vez da média, visto que o teste de normalidade de Shapiro-Wilk retornou um p-valor $< 0,05$, indicando que os conjuntos de pontuações seguem uma distribuição não normal.

Tabela 6.9: Estatísticas de linguagem sobre a média e o desvio padrão referentes ao número de caracteres, palavras, sentenças, além do vocabulário, nos parágrafos gerados durante os estudos de ablação. Os valores foram arredondados.

Método	**Caracteres**	**Palavras**	**Sentenças**	**Vocabulário**
VIIDA-s/Sentenças	148 ± 8	24 ± 1	23 ± 0	174
VIIDA-s/Parágrafo	218 ± 32	41 ± 6	7 ± 0	188
VIIDA-s/Pronomes	200 ± 19	37 ± 3	4 ± 0	186
VIIDA	185 ± 19	35 ± 3	4 ± 0	188
VIIDA-GRIT-s/Simil.Linguística	236 ± 37	45 ± 7	6 ± 1	461
VIIDA-GRIT-s/Simil.VQA	294 ± 60	55 ± 11	7 ± 2	718
VIIDA-GRIT	225 ± 38	43 ± 7	6 ± 1	461
Humanos	*256 ± 38*	*49 ± 7*	*4 ± 0*	*388*

das palavras por meio da medida IDF, essa métrica atribui menos peso a palavras comuns ou repetições. Assim, textos que contêm maior quantidade desses elementos tendem a receber pontuações menores.

O VIIDA-s/Parágrafo, sem os módulos de Extração de Relações e Estruturação de Parágrafos, não realiza a extração de relações semânticas entre entidades nem o agrupamento adequado de informações gramaticais. Portanto, embora apresente pontuações mais altas nas métricas, seus parágrafos não são bem estruturados, contendo repetições desnecessárias e pouca fluidez entre as sentenças, conforme mostrado na Figura 6.16. O VIIDA-s/Pronomes, por outro lado, apenas não substitui os sujeitos por pronomes. Assim, mesmo que este método apresente menor fluidez e coesão em comparação ao VIIDA, ele ainda gera parágrafos mais estruturados e de melhor qualidade em relação ao VIIDA-s/Parágrafo.

A remoção dos módulos de Extração de Relações, Estruturação de Parágrafos e Substituição por Pronomes impactou negativamente os resultados qualitativos, embora não de forma tão severa quanto a remoção do módulo Gerador de Sentenças. No entanto, apesar de a coesão textual, representada pelo uso adequado de pronomes, ser importante, uma melhor estruturação dos parágrafos contribui de maneira mais significativa para a fluidez e coerência na geração textual.

Na Tabela 6.8, também é possível visualizar a comparação dos resultados do VIIDA-GRIT com suas versões de módulos isolados. De forma semelhante ao observado nas versões do VIIDA, o VIIDA-GRIT-s/Simil.Linguística e o VIIDA-GRIT-s/Simil.VQA apresentaram uma melhoria nas pontuações da maioria das métricas em comparação com a versão completa do VIIDA-GRIT, embora de maneira mais sutil.

Conforme apresentado na Tabela 6.7, tanto o VIIDA-GRIT-s/Simil.Linguística quanto o VIIDA-GRIT-s/Simil.VQA utilizam os módulos de Extração de Relações, Estruturação de Parágrafo e Substituição por Pronomes, o que garante que seus

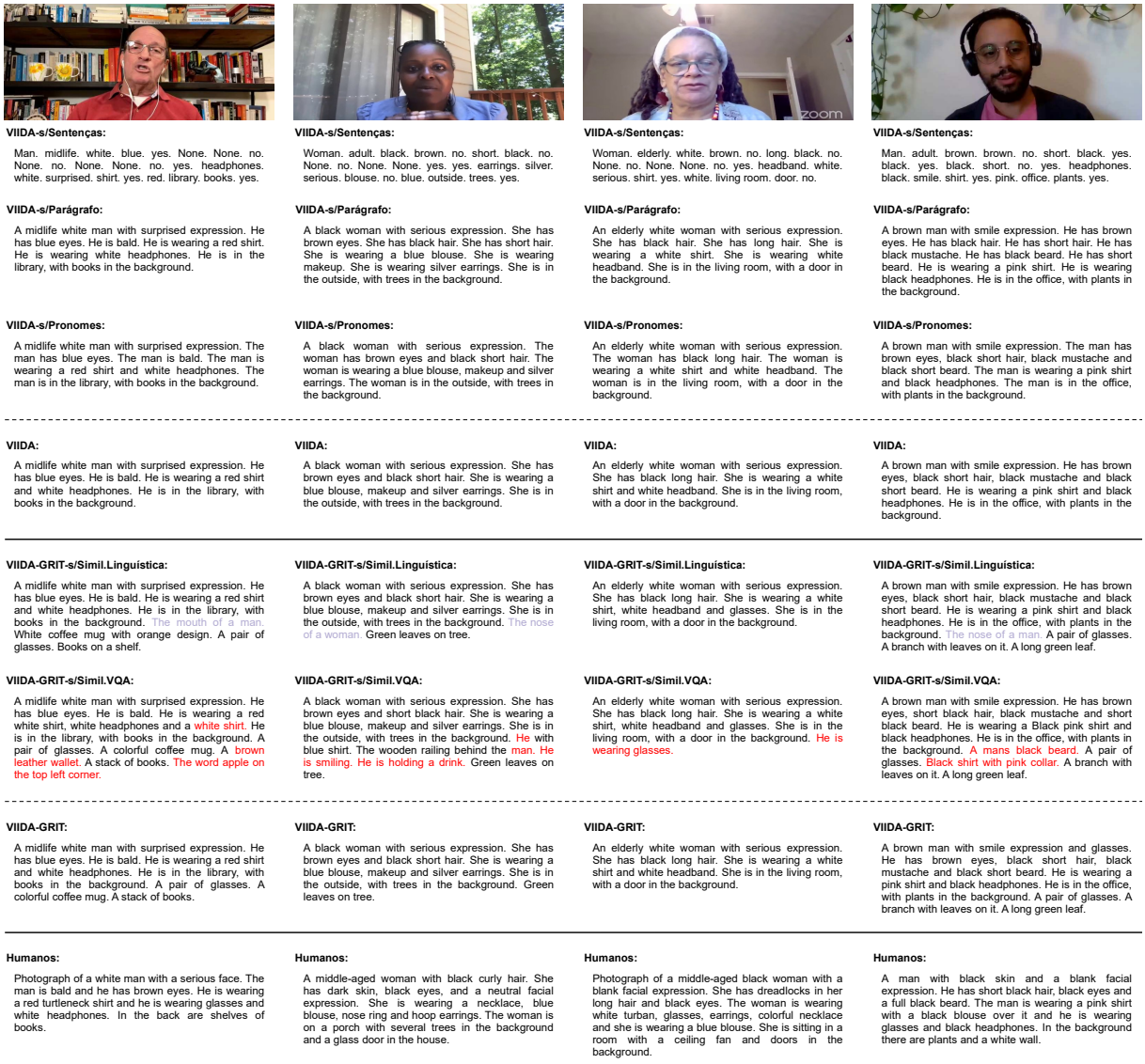


Figura 6.16: Exemplos de parágrafos gerados pelos métodos comparados durante os estudos de ablação. As sentenças em roxo e vermelho na figura mostram, respectivamente, os casos redundantes e incorretos ou repetidos. Fonte: Material próprio.

parágrafos sejam bem estruturados. A remoção dos módulos de similaridade influencia diretamente a quantidade e a qualidade dos conteúdos extras adicionados pelos modelos de legendagem densa aos parágrafos gerados, uma vez que esses módulos são responsáveis por aplicar filtros que descartam as legendas densas redundantes, incorretas ou fora de contexto. Os efeitos da remoção dos módulos de similaridade ficam evidentes ao analisar os parágrafos da Figura 6.16.

Os resultados na Tabela 6.9 destacam que o VIIDA-GRIT-s/Simil.VQA gerou parágrafos com uma quantidade maior de caracteres, palavras e sentenças, em comparação tanto à versão completa do VIIDA-GRIT quanto ao VIIDA-GRIT-s/Simil.Linguística. Além disso, o vocabulário utilizado no

VIIDA-GRIT-s/Simil.VQA é bem mais amplo, com 718 palavras distintas, maior que o de qualquer outro método, inclusive os textos de referência anotados por humanos.

Esses resultados indicam que, ao isolar o segundo módulo de Verificação de Similaridade, que inicialmente removia legendas densas similares às sentenças produzidas pela abordagem principal e confirmava visualmente as informações com o modelo VQA, mais conteúdo é gerado, o que pode incluir repetição de informações e sentenças adicionais, explicando o aumento nos números apresentados. A maior quantidade de palavras e o maior vocabulário sugerem que o conteúdo, por meio das legendas densas, não é filtrado adequadamente, permitindo uma diversidade de palavras maior, embora potencialmente redundante ou incorreta (por exemplo, o parágrafo gerado pelo VIIDA-GRIT-s/Simil.VQA para a segunda imagem da Figura 6.16). A variabilidade mais alta nos resultados, refletida pelo desvio padrão mais elevado, sugere também que os parágrafos gerados podem ser menos previsíveis e uniformes em termos de qualidade.

As ablações realizadas demonstraram que cada módulo contribui de maneira distinta para o desempenho geral do método proposto e de suas variantes nesta tese, sendo que alguns módulos exercem maior impacto do que outros. No caso do VIIDA, a geração de sentenças e a estruturação de parágrafos se mostraram mais cruciais para o desempenho, enquanto no VIIDA-GRIT a combinação dos módulos de filtragem por similaridade apresenta um efeito mais equilibrado, com o módulo de Verificação de Similaridade por meio do modelo linguístico se mostrando menos essencial.

Esse tipo de análise é fundamental para compreender a contribuição específica de cada componente e orientar melhorias futuras nos métodos. Os resultados indicam que a integração das técnicas através dos módulos oferece benefícios gerais, e que um equilíbrio (ou *trade-off*) entre resultados quantitativos e qualitativos deve ser buscado para alcançar resultados mais confiáveis e precisos.

6.6 Vantagens e Limitações

6.6.1 Vantagens

O VIIDA e suas variantes oferecem diversas vantagens, como interpretabilidade, flexibilidade, fácil ajuste, personalização e proteção contra a geração de conteúdo inseguro. Uma de suas principais forças é a capacidade de produzir parágrafos precisos e adequados para o público com CBV, apresentando um vocabulário diversificado e compatibilidade com imagens de baixa qualidade.

Ao contrário dos métodos concorrentes, o VIIDA e suas variantes não requerem treinamento, são compatíveis tanto com CPU quanto com GPU e apresentam

Tabela 6.10: Análise de custo dos recursos computacionais. Além do VIIDA e suas variantes, foram selecionados apenas o FuseCap, LLaVA e InstructBLIP com suas versões direcionadas, devido aos melhores resultados qualitativos obtidos em comparação com outros métodos. O símbolo - indica que a GPU utilizada ficou sem memória ao processar o método, enquanto o símbolo * indica que o método foi executado apenas em CPU. A coluna "Processamento" apresenta o tempo médio (em segundos) de inferência necessário para que o método avaliado processe a imagem de entrada e gere o texto de saída, com desvio padrão. As colunas relativas às memórias RAM e VRAM são expressas em GB.

Método	**CPU**	**GPU**	**Processamento**	**RAM**	**VRAM**
LLaVA	X	✓	$10{,}90 \pm 2{,}23$	5,9	9,7
LLaVA-Directed	X	✓	$13{,}30 \pm 2{,}17$	5,9	9,7
InstructBLIP*	✓	✓	$232{,}09 \pm 26{,}61$	34,2	-
InstructBLIP-Directed*	✓	✓	$210{,}77 \pm 35{,}43$	34,2	-
FuseCap	✓	✓	$0{,}95 \pm 0{,}21$	2,9	1,2
VIIDA	✓	✓	$4{,}11 \pm 0{,}34$	2,5	1,8
VIIDA-Dense	✓	✓	$11{,}89 \pm 3{,}32$	2,5	1,8
VIIDA-GRIT	✓	✓	$11{,}09 \pm 5{,}12$	2,5	1,8

eficiência em termos de consumo de memória e tempo de processamento, conforme apresentado na Tabela 6.10. Ressalta-se que os requisitos referentes à memória e ao processamento não se aplicam ao DenseCap e ao GRIT, pois esses modelos são executados de forma independente do método proposto, e somente suas saídas são utilizadas.

No VIIDA, são gastos em média cerca de $4 \pm 0,32$ e $0,1 \pm 0,02$ segundos de processamento, respectivamente, nas etapas de Extração de Informações Visuais e Produção Estruturada de Parágrafos. Isso demonstra que o método determinístico para estruturação de parágrafos, mesmo utilizando uma ferramenta de grafo de cena, é extremamente eficiente em termos de tempo.

Para o VIIDA-Dense e VIIDA-GRIT, o tempo de processamento adicional deve-se às verificações realizadas sobre as legendas densas geradas. Dessa forma, a abordagem complementar da etapa de Extração de Informações Visuais desses métodos passa a demandar, em média, aproximadamente 7 segundos, enquanto a abordagem principal mantém 4 segundos. A etapa de Produção Estruturada de Parágrafos passa para $0,23 \pm 0,11$ segundos devido à maior quantidade de sentenças adicionadas ao parágrafo final. No entanto, mesmo com o aumento de tempo em relação ao VIIDA, ambas as variantes ainda apresentam tempos de processamento baixos.

Treinar modelos de Aprendizado Profundo do zero exige uma quantidade significativa de recursos computacionais devido à complexidade do modelo e à quantidade de dados. Esse processo envolve a realização de múltiplas operações em grande escala, utilizando GPUs ou Unidades de Processamento Tensorial (*Tensor*

Processing Unit - TPUs), que consomem muita energia elétrica. A geração de energia, dependendo da fonte, pode resultar na emissão de grandes quantidades de dióxido de carbono (CO_2) e outros gases de efeito estufa (Castaño et al. (2023)). Portanto, ao utilizar modelos pré-treinados, como o VIIDA e suas variantes, as emissões associadas ao treinamento inicial são diluídas entre todos os usuários que aproveitam o modelo, tornando o uso mais eficiente em termos de impacto ambiental. Em outras palavras, usá-los evita a necessidade de repetir todo o treinamento, que pode levar dias ou semanas, economizando assim uma quantidade substancial de energia e recursos computacionais, o que contribui para a redução da emissão de carbono (Savci and Das (2023)).

Além do VIIDA, a métrica InViDe também apresenta vantagens significativas. Ela é interpretável, parametrizada, não requer treinamento, opera exclusivamente em CPU e representa uma abordagem pioneira em seu campo.

6.6.2 Limitações

Embora tanto o VIIDA quanto a InViDe tenham demonstrado eficácia na produção e avaliação de parágrafos de imagens de webinários para pessoas com CBV, existem várias limitações a serem mencionadas. O VIIDA e suas variantes são limitados a imagens que contêm apenas uma pessoa, tornando-os menos precisos quando a imagem inclui duas ou mais pessoas, pois foram projetados para descrever cenas de webinário com uma única pessoa. Além disso, esses métodos são restritos pelas perguntas fornecidas ao modelo VQA, resultando em sentenças pré-estruturadas baseadas exclusivamente nas saídas desse modelo.

Apesar de substituir o PLM e os modelos de Aprendizado de Máquina pré-treinados usados no *baseline* do Capítulo 4, o VIIDA e suas variantes ainda podem acumular e propagar erros de previsão, levando à geração de parágrafos incorretos. Isso ocorre porque os métodos continuam a depender de outros modelos e ferramentas (por exemplo, BLIP, RoBERTa, DenseCap, GRIT, SpaCy, SceneGraphParser, WordNet, entre outros) que também têm limitações, expondo-os às deficiências desses componentes. Além disso, ressalta-se que o processo de agrupamento das informações gramaticais utilizado para estruturar os parágrafos no algoritmo determinístico proposto só funciona em sentenças que compartilham as mesmas relações gramaticais; caso contrário, as sentenças analisadas são simplesmente inseridas no texto.

A métrica InViDe apresenta várias limitações que devem ser observadas. Primeiramente, a métrica não avalia a qualidade do texto ou sua relevância em relação ao conteúdo da imagem, limitando-se a comparações com um texto de referência. Além disso, a InViDe depende dos *synsets* do WordNet e da ferramenta

SpaCy, o que a torna sujeita às fragilidades desses componentes. Outro ponto importante é que a InViDe é baseada exclusivamente em sentenças em inglês e não é compatível com outros idiomas.

6.7 Desfechos Alcançados

Neste capítulo, foi apresentado o VIIDA, um *pipeline* automatizado desenvolvido para gerar descrições textuais de cenas de webinários adaptadas para indivíduos com CBV. O VIIDA representa uma melhoria significativa em relação ao método anterior (*baseline*) descrito no Capítulo 4, alcançando avanços notáveis em precisão.

O VIIDA e suas variantes, em suas estruturas, aproveitam os benefícios de um modelo multimodal de VQA, complementado por um conjunto de filtros projetados para adaptar as descrições às necessidades específicas das pessoas com deficiência visual. Além disso, foi desenvolvido um algoritmo determinístico que utiliza um grafo de cena para estruturar as descrições em parágrafos coerentes e descritivos em linguagem natural.

Também foi introduzida a InViDe, uma métrica personalizada de avaliação automática projetada para medir a adequação das descrições textuais geradas por modelos computacionais, em termos de acessibilidade, para pessoas com CBV.

Realizou-se um estudo experimental abrangente sobre o VIIDA e a InViDe usando o conjunto de imagens com apresentadores em primeiro plano proposto no Capítulo 5 (webinários). Por meio de experimentos e análises, foi demonstrada a eficácia dos métodos propostos tanto na geração quanto na avaliação de descrições de imagens para indivíduos com deficiência visual, superando métodos considerados estado da arte.

Por fim, exemplos qualitativos mostraram as diversas características e vantagens do VIIDA e suas variantes. Destaca-se, entre essas vantagens, a capacidade de incluir atributos representativos nas descrições, o que é uma característica importante para mitigar os vieses algorítmicos inerentes às ferramentas de IA, e a não geração de conteúdos NSFW.

Capítulo 7

Conclusão

Nesta tese, foi abordada a tarefa desafiadora e não trivial de criar automaticamente descrições de imagens adaptadas, mais especificamente, de cenas de webinários para pessoas com CBV. Para isso, foram aplicadas conjuntamente técnicas e modelos de Aprendizado Profundo baseados em Visão Computacional e PLN, com o objetivo de desenvolver abordagens computacionais que consigam prover acesso, de maneira textual, ao conteúdo das imagens digitais oriundas de ferramentas de videoconferência (como, Zoom, Google Meet, YouTube, etc.) para esse público específico de forma inclusiva, mitigando barreiras de acessibilidade.

Para alcançar o objetivo principal desta tese, foram realizados três grandes estudos que exploraram as questões de acessibilidade relacionadas ao uso de ferramentas de videoconferência para pessoas com deficiência visual. Esses estudos englobaram os oito objetivos específicos definidos na Introdução (Capítulo 1) e possibilitaram o desenvolvimento e validação da abordagem final apresentada no Capítulo 6, o VIIDA. O primeiro estudo, atrelado ao OE-1 e ao OE-2, concentrou-se no desenvolvimento de uma abordagem computacional inicial (denominada *baseline*) e na investigação da área pesquisada.

O *baseline* foi estruturado em duas fases principais: a extração de características visuais das imagens em formato de descrição textual e a posterior geração de um parágrafo em linguagem natural a partir da sumarização dessas descrições. Para esse fim, foram aplicados modelos pré-treinados de Aprendizado de Máquina para tarefas específicas, em combinação com conjuntos de filtros de PLN, para ajustar as descrições ao domínio de interesse. Com os resultados alcançados, foi possível identificar tanto o potencial quanto as limitações do estudo desenvolvido, assim como os achados e as lacunas que serviram como base para o prosseguimento dos estudos posteriores que compõem esta tese, como, por exemplo, a necessidade de um conjunto de dados e de uma métrica de avaliação específicos.

Dada a crucial necessidade de um conjunto de dados específico para o problema abordado, identificada no primeiro estudo, o segundo estudo desta tese focou na construção desse conjunto para apoiar tarefas de Visão Computacional e, consequentemente, o desenvolvimento de sistemas de Tecnologia Assistiva. Para

isso, foi proposta uma abordagem para a coleta automática de dados (neste caso, imagens ou *frames*) a partir de *playlists* de webinários, notícias, entrevistas e similares do YouTube (OE-3), além de um protocolo para a anotação (descrição) desses dados especificamente para o público com CBV, com base em normas e diretrizes de acessibilidade (OE-4).

Os experimentos e análises realizados demonstraram a viabilidade tanto da abordagem de coleta quanto do protocolo de anotação, resultando na criação de um conjunto de dados (em três versões) e na avaliação de sua qualidade, diversidade e representatividade por meio de métodos analíticos e modelos de Aprendizado de Máquina (OE-5). O conjunto de dados resultante deste estudo foi a principal contribuição para a área de Visão Computacional desta tese, uma vez que ele pode ser utilizado em diversas tarefas além da geração de descrições de imagens, como reconhecimento de objetos, segmentação de imagens, detecção de atributos visuais, classificação de imagens, classificação de cenas, entre outros. Além disso, o conjunto criado auxilia na redução da escassez de conjuntos de dados direcionados para o contexto de aplicações para deficientes visuais.

Com a realização dos estudos anteriores e seus respectivos resultados, foram identificadas possíveis melhorias que permitiram o refinamento do *baseline*. Assim, o terceiro e último estudo desta tese focou, principalmente, na remodelagem do *pipeline* automatizado para alcançar melhores descrições textuais (parágrafos). O VIIDA, como foi chamado o método refinado, também foi composto de duas fases para a extração de características visuais das imagens e para a produção de um parágrafo final por meio de modelos de Aprendizado Profundo (OE-6). Contudo, como estratégia, exploraram-se principalmente as vantagens de aplicar um modelo VQA para uma melhor aquisição de informações relevantes, bem como de utilizar uma ferramenta de grafo de cena e o agrupamento das informações extraídas pelo grafo para produzir parágrafos mais descritivos, coerentes e estruturados. Contrastando com outras técnicas da literatura, que realizam a descrição de imagens de maneira genérica, o VIIDA é específico para descrever as características pertinentes das imagens relacionadas ao apresentador e ao ambiente (como, webinários), garantindo que deficientes visuais possam ter uma boa compreensão da cena.

Os experimentos demonstraram que o VIIDA e suas variantes superam os métodos clássicos e atuais de descrições de imagens, em termos de compatibilidade semântica entre imagem e texto, geração de características linguisticamente semelhantes às dos seres humanos, e alinhamento com as diretrizes de acessibilidade. Em relação ao *baseline*, os experimentos também evidenciaram as melhorias proporcionadas pela modelagem refinada do *pipeline*, tanto na maior precisão na aquisição das características visuais quanto na melhor produção e

estruturação do parágrafo.

Realizou-se também uma análise de geração de imagens com base nas descrições produzidas pelos métodos comparados e uma análise de conteúdo NSFW. Na primeira análise, demonstrou-se que o VIIDA e suas variantes forneceram textos que, ao serem enviados para modelos generativos, resultaram em imagens sintéticas semanticamente mais próximas às imagens originais. Na segunda análise, atrelada ao OE-7, verificou-se que o VIIDA e suas variantes não produzem conteúdo potencialmente prejudicial e suas considerações éticas foram levantas, assim como sugestões para se garantir possíveis gerações.

No último estudo, ligado ao OE-8, também foi introduzida a InViDe, uma métrica pioneira personalizada, fundamentada em diretrizes, boas práticas recomendadas e estudos na área de acessibilidade, para avaliar automaticamente a adequação das descrições textuais geradas artificialmente a partir de imagens por modelos e métodos computacionais para deficientes visuais. Diferentemente das métricas de avaliação comumente utilizadas, a InViDe não analisa o alinhamento semântico entre texto e imagem, mas verifica a adequação das descrições, ou seja, se os textos produzidos estão de acordo com os requisitos de acessibilidade, proporcionando uma avaliação mais precisa e relevante a respeito das informações necessárias para esse público específico. A métrica InViDe representa um avanço metodológico para a validação de tecnologias assistivas e foi a principal contribuição deste último estudo para a área de PLN.

Diante dos estudos desenvolvidos ao longo desta tese e dos resultados alcançados, conclui-se que o VIIDA, ao combinar técnicas avançadas de Visão Computacional e PLN, introduz um método específico para gerar descrições de imagens otimizadas para pessoas com deficiência visual. Diferente das abordagens tradicionais de geração e avaliação de descrições de imagens, o VIIDA e a InViDe concentram-se em aspectos críticos para a acessibilidade, visando aprimorar a precisão, relevância e adequação dos textos produzidos. As integrações realizadas evidenciam uma aplicação inovadora de diferentes áreas da IA para solucionar problemas práticos e de importância social.

Ao desenvolver um sistema que torna o conteúdo visual mais acessível, o VIIDA contribui diretamente para o campo da Tecnologia Assistiva por meio da inclusão digital, alinhando-se com as tendências e necessidades contemporâneas da sociedade e abrindo portas para novas investigações. Além disso, ao utilizar modelos pré-treinados e técnicas que economizam recursos computacionais, o VIIDA também participa da discussão sobre sustentabilidade na computação, um tema cada vez mais relevante em tempos de preocupação com o impacto ambiental de grandes modelos de IA. Portanto, esses aspectos fazem do produto final desta tese uma contribuição valiosa e multifacetada para a Ciência da Computação, com impacto

significativo tanto no campo teórico quanto no prático.

7.1 Trabalhos Futuros

O campo de estudo ao qual esta tese se dedica é relativamente novo e apresenta diversas oportunidades para expandir suas fronteiras. A seguir, são apresentadas algumas possibilidades de melhorias para a tarefa de geração automática de descrições de imagens para pessoas com deficiência visual.

Como trabalho futuro, pretende-se aprimorar a correção de frases malformadas resultantes da concatenação das saídas dos modelos de legendagem densa nas versões complementares do VIIDA, aproveitando o poder dos LLMs. Essa melhoria visa aumentar a coesão e a clareza das descrições geradas, tornando-as mais compreensíveis e úteis para pessoas com CBV.

Além disso, considerando o desempenho alcançado e o potencial para ajustes adicionais, planeja-se utilizar o VIIDA e suas variações como rotuladores automáticos de conjuntos de dados. Por exemplo, é possível utilizar o VIIDA-GRIT para anotar a versão completa, com mais de 10.000 imagens, do conjunto de dados de webinários proposto no Capítulo 5, produzindo descrições precisas e detalhadas de cada imagem. Esse processo de rotulação automatizada permitirá a criação de um conjunto de dados ainda mais robusto, diversificado e com menor custo, tanto monetário quanto humano.

Com esse conjunto de dados rotulado, torna-se possível treinar ou ajustar finamente (*fine-tuning*) um modelo ponta a ponta especificamente projetado para descrever imagens para deficientes visuais. Esse modelo será otimizado para capturar variações e detalhes importantes, garantindo que as descrições geradas não apenas atendam às normas e diretrizes de acessibilidade, mas também proporcionem uma experiência rica e informativa para usuários com baixa visão ou cegueira.

Contribuições adicionais podem ser alcançadas utilizando o conjunto de dados de pessoa única anotado, também proposto no Capítulo 5. Para cada imagem, há pelo menos uma anotação em português sobre a cena de webinário observada. Assim, um estudo pode ser realizado para estender as abordagens computacionais desenvolvidas nesta tese à geração de descrições de imagens em português.

A realização de uma avaliação dos resultados do VIIDA e suas variantes envolvendo seres humanos, com o objetivo de enriquecer seus resultados, além de uma investigação mais aprofundada e quantificação dos vieses encontrados nos modelos de classificação de conteúdos NSFW, tanto textuais quanto visuais, também se enquadram como possíveis trabalhos futuros.

Uma vez que o contexto adicional proporciona uma compreensão mais profunda da cena, indo além do que é observado, planeja-se investigar novas técnicas de fusão

de contexto. Para isso, serão consideradas a integração de informações multimodais, a legendagem em nível de região e o uso de máscaras de segmentação de objetos para enriquecer ainda mais a qualidade e a precisão das descrições geradas.

É possível adaptar o VIIDA ao contexto de vídeos, mas para isso, o método precisaria ser ajustado, considerando algumas modificações e expansões em sua metodologia. Por exemplo, para que o VIIDA possa receber como entrada um vídeo em vez de uma imagem, seria necessária a integração de um mecanismo para a captação e processamento de quadros, semelhante ao descrito no Capítulo 5, para a posterior descrição do quadro selecionado. Nesse caso, a modificação seria referente apenas à mudança da entrada do método. No entanto, se o objetivo do trabalho futuro for estender o método para lidar com eventos de vídeo, torna-se essencial incorporar a modelagem temporal e a extração de recursos espaciais-temporais dos quadros, garantindo que as descrições geradas sejam representativas do conteúdo completo do vídeo. Além disso, ao considerar ações e eventos contínuos, a criação de novas estruturas de sentenças e ajustes nos modelos de geração de linguagem natural seriam necessários para capturar adequadamente a progressão dos eventos ao longo do tempo. Nesse segundo caso, destaca-se que haveria uma mudança na tarefa de legendagem de imagem para legendagem de vídeo, o que implicaria também em alterações nos conjuntos de dados, métricas de avaliação e recursos computacionais envolvidos.

Por fim, é fundamental continuar a desenvolver e refinar métricas de avaliação que considerem tanto a qualidade semântica quanto a acessibilidade das descrições geradas, assegurando que estas atendam plenamente às expectativas e necessidades dos deficientes visuais. Em relação a métrica InViDe, uma sugestão seria o adiciono da análise de legibilidade como critério.

Em resumo, o foco futuro deste trabalho está em refinar as capacidades do VIIDA, expandir a anotação de conjuntos de dados e desenvolver modelos mais especializados para oferecer descrições de imagens de alta qualidade para a comunidade de deficientes visuais. Espera-se que esta tese inspire futuras pesquisas no desenvolvimento de modelos de IA que sirvam efetivamente como Tecnologias Assistivas, garantindo a acessibilidade de imagens para pessoas com cegueira ou baixa visão.

Referências Bibliográficas

ABNT (2016). ABNT NBR 16452: Acessibilidade na comunicação — audiodescrição. *Associação Brasileira de Normas Técnicas*.

Acosta-Vargas, P., Guaña-Moya, J., Acosta-Vargas, G., Villegas-Ch, W., and Salvador-Ullauri, L. (2021). Method for assessing accessibility in videoconference systems. In *Intelligent Human Systems Integration 2021: Proceedings of the 4th International Conference on Intelligent Human Systems Integration (IHSI 2021)*, pages 669–675. Springer International Publishing.

Ahmad, O. B., Boschi-Pinto, C., Lopez, A. D., Murray, C. J., Lozano, R., and Inoue, M. (2001). Age standardization of rates: A new WHO standard. *World Health Organization*, 9(10):1–14.

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65.

Alhichri, H., Bazi, Y., Alajlan, N., and Bin Jdira, B. (2019). Helping the visually impaired see via image multi-labeling based on squeezenet CNN. *Applied Sciences*, 9(21):4656.

Allen, P. M. and Smith, L. (2020). Sars-cov-2 self-isolation: recommendations for people with a vision impairment. *Eye*, 34(7):1183–1184.

Anderson, P., Fernando, B., Johnson, M., and Gould, S. (2016). Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016*, pages 382–398. Springer.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). VQA: Visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2425–2433. IEEE Computer Society.

Arriaga, O., Valdenegro-Toro, M., Muthuraja, M., Devaramani, S., and Kirchner, F. (2020). Perception for autonomous systems (PAZ). *arXiv preprint*.

Arriaga, O., Valdenegro-Toro, M., and Plöger, P. (2019). Real-time convolutional neural networks for emotion and gender classification. In *2019 Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, pages 221–226.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3th International Conference on Learning Representations (ICLR)*, pages 1–15.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72. Association for Computational Linguistics.

Bansal, H., Yin, D., Monajatipoor, M., and Chang, K. (2022). How well can text-to-image generative models understand ethical natural language interventions? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Cao, S., An, G., Cen, Y., Yang, Z., and Lin, W. (2024). CAST: Cross-modal retrieval and visual conditioning for image captioning. *Pattern Recognition*, pages 1–26.

Caseli, H. d. M. and Nunes, M. d. G. V. (2023). *Processamento de linguagem natural: conceitos, técnicas e aplicações em português*. Universidade de São Paulo.

Castaño, J., Martínez-Fernández, S., Franch, X., and Bogner, J. (2023). Exploring the carbon footprint of hugging face's ml models: A repository mining study. In *2023 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–12.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 169–174. Association for Computational Linguistics.

Chatterjee, M. and Schwing, A. G. (2018). Diverse and coherent paragraph generation from images. In *Computer Vision – ECCV 2018*, pages 747–763. Springer International Publishing.

Chen, J. and Zhuge, H. (2018). Abstractive text-image summarization using multi-modal attentional hierarchical rnn. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4046–4056.

Chen, S., Yao, T., and Jiang, Y.-G. (2019). Deep learning for video captioning: A review. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, volume 1, page 2.

Chen, S.-Y., Feng, Z., and Yi, X. (2017). A general introduction to adjustment for multiple comparisons. *Journal of Thoracic Disease*, 9(6):1725–1729.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Christian, H., Agus, M. P., and Suhartono, D. (2016). Single document automatic text summarization using term frequency-inverse document frequency (tf-idf). *ComTech: Computer, Mathematics and Engineering Applications*, 7(4):285–294.

Cornia, M., Stefanini, M., Baraldi, L., and Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10578–10587. IEEE Computer Society.

Croitoru, F.-A., Hondru, V., Ionescu, R. T., and Shah, M. (2023). Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(9):10850–10869.

Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. (2023). InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 49250–49267. Curran Associates, Inc.

de Alencar, R. S., Castañeda, W. A. C., and Amadeus, M. (2024). Image captioning for brazilian portuguese using GRIT model. *arXiv preprint*.

De Marneffe, M.-C. and Manning, C. D. (2008). Stanford typed dependencies manual. Technical report, Technical report, Stanford University.

de Melo, F. R., de Oliveira Matos, H. C., and Dias, E. R. B. (2015). Aplicação da métrica bleu para avaliação comparativa dos tradutores automáticos bing tradutor e google tradutor. *Revista e-scrita: Revista do Curso de Letras da UNIABEU*, 5(3):33–45.

Delloul, K. and Larabi, S. (2022). Image captioning state-of-the-art: Is it enough for the guidance of visually impaired in an environment? In *Advances in Computing Systems and Applications*, pages 385–394. Springer International Publishing.

Dognin, P., Melnyk, I., Mroueh, Y., Padhi, I., Rigotti, M., Ross, J., Schiff, Y., Young, R. A., and Belgodere, B. (2022). Image captioning as an assistive technology: Lessons learned from VizWiz 2020 challenge. *Journal of Artificial Intelligence Research*, 73:437–459.

Dokania, H. and Chattaraj, N. (2024). An assistive interface protocol for communication between visually and hearing-speech impaired persons in internet platform. *Disability and Rehabilitation: Assistive Technology*, 19(1):233–246.

Dominguez-Catena, I., Paternain, D., and Galar, M. (2024). Metrics for dataset demographic bias: A case study on facial expression recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 1–18.

dos Santos, G. O., Colombini, E. L., and Avila, S. (2022). #PraCegoVer: A large dataset for image captioning in portuguese. *Data*, 7(2).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations (ICLR)*, pages 1–21.

Doush, I. A., Al-Jarrah, A., Alajarmeh, N., and Alnfiai, M. (2023). Learning features and accessibility limitations of video conferencing applications: Are people with visual impairment left behind. *Universal Access in the Information Society*, 22(4):1353–1368.

Du, Y., Wang, W., and Wang, L. (2015). Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118.

Faruqui, M., Tsvetkov, Y., Rastogi, P., and Dyer, C. (2016). Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35, Berlin, Germany. Association for Computational Linguistics.

Fernandes., D. L., Ribeiro., M. H. F., Cerqueira., F. R., and Silva., M. M. (2022). Describing image focused in cognitive and visual details for visually impaired people: An approach to generating inclusive paragraphs. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 5: VISAPP*, pages 526–534. INSTICC, SciTePress.

Fernandes, D. L., Ribeiro, M. H. F., Silva, M. M., and Cerqueira, F. R. (2024). VIIDA and InViDe: Computational approaches for generating and evaluating inclusive image paragraphs for the visually impaired. *Disability and Rehabilitation: Assistive Technology*, 19.

Ferreira, L., Fernandes, D., Cerqueira, F., Ribeiro, M., and Silva, M. (2023). Presenter-centric image collection and annotation: Enhancing accessibility for the visually impaired. In *2023 36th Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 199–204. IEEE Computer Society, SBC.

Figueiredo, E. L., Fernandes, D. L., and Reis, J. C. S. (2023). Text-to-image generation tools: A survey and nsfw content analysis. In *Anais Estendidos do XXIX Simpósio Brasileiro de Sistemas Multimídia e Web*, pages 59–62. SBC.

Flaxman, S. R., Bourne, R. R., Resnikoff, S., Ackland, P., Braithwaite, T., Cicinelli, M. V., Das, A., Jonas, J. B., Keeffe, J., and Kempen, J. H. (2017). Global causes of blindness and distance vision impairment 1990–2020: A systematic review and meta-analysis. *The Lancet Global Health*, 5(12):1221–1234.

Freitag, M. and Al-Onaizan, Y. (2017). Beam search strategies for neural machine translation. In *Proc. of the First Workshop on Neural Machine Translation*, pages 56–60, Vancouver. Association for Comput. Linguistics.

Garcia, N., Hirota, Y., Wu, Y., and Nakashima, Y. (2023). Uncurated image-text datasets: Shedding light on demographic bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6957–6966. IEEE Computer Society.

Gondim, J. M., Claro, D. B., and Souza, M. (2022). Towards image captioning for the portuguese language: Evaluation on a translated dataset. In *ICEIS (1)*, pages 384–393.

González-Chávez, O., Ruiz, G., Moctezuma, D., and Ramirez-delReal, T. (2024). Are metrics measuring what they should? An evaluation of image captioning task metrics. *Signal Processing: Image Communication*, 120:117071.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*. MIT press Cambridge.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., et al. (2013). Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pages 117–124. Springer.

Goodwin, T., Savery, M., and Demner-Fushman, D. (2020). Flight of the PEGASUS? comparing transformers on few-shot and zero-shot multi-document abstractive summarization. In *Proceedings of the 28th International Conference on Computational Linguistics (ICCL)*, pages 5640–5646.

Granquist, C., Sun, S. Y., Montezuma, S. R., Tran, T. M., Gage, R., and Legge, G. E. (2021). Evaluation and comparison of artificial intelligence vision aids: Orcam myeye 1 and seeing ai. *Journal of Visual Impairment and Blindness*, 115(4):277–285.

Guo, Y., Chen, Y., Xie, Y., Ban, X., and Obaidat, M. S. (2022). An offline assistance tool for visually impaired people based on image captioning. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 969–976. IEEE Computer Society.

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., and Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, 187:27–48.

Gurari, D., Li, Q., Lin, C., Zhao, Y., Guo, A., Stangl, A., and Bigham, J. P. (2019). Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 939–948.

Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., and Bigham, J. P. (2018). VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3608–3617. IEEE Computer Society.

Gurari, D., Zhao, Y., Zhang, M., and Bhattacharya, N. (2020). Captioning images taken by people who are blind. In *Computer Vision – ECCV 2020*, pages 417–434. Springer International Publishing.

Han, J., Kamber, M., Pei, J., et al. (2012). Getting to know your data. In *Data mining*, pages 39–82. Elsevier Amsterdam, Netherlands.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., and Tao, D. (2023). A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(1):87–110.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

He, S., Liao, W., Tavakoli, H. R., Yang, M., Rosenhahn, B., and Pugeault, N. (2020). Image captioning through image transformer. In *Proceedings of the Asian Conference on Computer Vision*.

Herdade, S., Kappeler, A., Boakye, K., and Soares, J. (2019). Image captioning: Transforming objects into words. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 1–11. Curran Associates Inc.

Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., and Choi, Y. (2021). CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7514–7528. Association for Computational Linguistics.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, volume 30, page 6629–6640. Curran Associates Inc.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength natural language processing in python. Zenodo.

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., and Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys*, 51(6):1–36.

Huang, L., Wang, W., Chen, J., and Wei, X.-Y. (2019). Attention on attention for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4634–4643.

Huang, W. and Jing, Z. (2007). Evaluation of focus measures in multi-focus image fusion. *Pattern recognition letters*, 28(4):493–500.

Huang, Y., Wu, Z., Gao, C., Peng, J., and Yang, X. (2024). Exploring the distinctiveness and fidelity of the descriptions generated by large vision-language models. *arXiv preprint*, pages 1–11.

Jafari, M. and Ansari-Pour, N. (2019). Why, when and how to adjust your p-values? *Cell Journal (Yakhteh)*, 20(4):604–607.

Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., and Kumar, S. (2024). Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9307–9315.

Jia, X., Gavves, E., Fernando, B., and Tuytelaars, T. (2015). Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2407–2415.

Johnson, J., Karpathy, A., and Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4565–4574. IEEE Computer Society.

K., V. V., Sen, D., and Raman, B. (2019). Video skimming: Taxonomy and comprehensive survey. *ACM Computing Surveys*, 52(5).

Kane, H., Kocyigit, M. Y., Abdalla, A., Ajanoh, P., and Coulibali, M. (2020). NUBIA: NeUral based interchangeability assessor for text generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 28–37, Online (Dublin, Ireland). Association for Computational Linguistics.

Karkkainen, K. and Joo, J. (2021). Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1548–1558.

Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys*, 54(10s).

Khurana, A., Subramonyam, H., and Chilana, P. K. (2024). Why and when llm-based assistants can go wrong: Investigating the effectiveness of prompt-based interactions for software help-seeking. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, page 288–303. Association for Computing Machinery.

Kilickaya, M., Erdem, A., Ikizler-Cinbis, N., and Erdem, E. (2017). Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 199–209, Valencia, Spain. Association for Computational Linguistics.

Kim, W., Son, B., and Kim, I. (2021). ViLT: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 5583–5594.

Krause, J., Johnson, J., Krishna, R., and Fei-Fei, L. (2017). A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3337–3345. IEEE Computer Society.

Krishna, R., Hata, K., Ren, F., Fei-Fei, L., and Niebles, J. C. (2017a). Dense-captioning events in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 706–715. IEEE Computer Society.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. (2017b). Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73.

Kuhn, D. M. and Moreira, V. P. (2021). BRCars: a dataset for fine-grained classification of car images. In *2021 34th Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 231–238.

Le, T., Nguyen, H. T., and Nguyen, M. L. (2021). Multi visual and textual embedding on visual question answering for blind people. *Neurocomputing*, 465:451–464.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Lee, H., Yoon, S., Dernoncourt, F., Kim, D. S., Bui, T., and Jung, K. (2020). Vilbertscore: Evaluating image caption using vision-and-language bert. In *Proceedings of the first workshop on evaluation and comparison of NLP systems*, pages 34–39.

Leporini, B., Buzzi, M., and Hersh, M. (2021). Distance meetings during the COVID-19 pandemic: Are video conferencing tools accessible for blind people? In *Proceedings of the 18th International Web for All Conference*, pages 1–10. Association for Computing Machinery.

Lewis, M. et al. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the ACL*, pages 7871–7880. Association for Comput. Linguistics.

Lewis, V. (2018). How to write alt text and image descriptions for the visually impaired. *Perkins School for the Blind*.

Li, G., Zhu, L., Liu, P., and Yang, Y. (2019). Entangled transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8928–8937.

Li, J., Li, D., Xiong, C., and Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162, pages 12888–12900. PMLR.

Li, Y., Ouyang, W., Zhou, B., Wang, K., and Wang, X. (2017). Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1270–1279. IEEE Computer Society.

Liang, X., Hu, Z., Zhang, H., Gan, C., and Xing, E. P. (2017). Recurrent topic-transition gan for visual paragraph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3382–3391. IEEE Computer Society.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023). Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 34892–34916. Curran Associates, Inc.

Liu, W., Chen, S., Guo, L., Zhu, X., and Liu, J. (2021). CPTR: Full transformer network for image captioning. *arXiv preprint*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint*, pages 1–13.

Liu, Y., Shi, Y., Feng, F., Li, R., Ma, Z., and Wang, X. (2022). Improving image paragraph captioning with dual relations. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE Computer Society.

Lu, J., Xiong, C., Parikh, D., and Socher, R. (2017). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 375–383. IEEE Computer Society.

Malakan., Z. M., Aafaq., N., Hassan., G. M., and Mian., A. (2021). Contextualise, attend, modulate and tell: Visual storytelling. In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 5: VISAPP*, pages 196–205. INSTICC, SciTePress.

Massiceti, D., Zintgraf, L., Bronskill, J., Theodorou, L., Harris, M. T., Cutrell, E., Morrison, C., Hofmann, K., and Stumpf, S. (2021). ORBIT: A real-world few-shot dataset for teachable object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10798–10808.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Naik, R. and Nushi, B. (2023). Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, page 786–808. Association for Computing Machinery.

Nascimento, B., Ribeiro, M., Silva, L., Capobiango, N., and Silva, M. (2022). A soybean seedlings dataset for soil condition and genotype classification. In *2022 35th Conference on Graphics, Patterns and Images (SIBGRAPI)*, volume 1, pages 85–90.

Ng, E. G., Pang, B., Sharma, P., and Soricut, R. (2021). Understanding guided image captioning performance across domains. In *Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL)*, pages 183–193. Association for Computational Linguistics.

Nguyen, V.-Q., Suganuma, M., and Okatani, T. (2022). GRIT: Faster and better image captioning transformer using dual visual features. In *Computer Vision – ECCV 2022*, pages 167–184. Springer Nature Switzerland.

NithyaKalyani, A. and Jothilakshmi, S. (2019). Speech summarization for tamil language. In *Intelligent Speech Signal Processing*, pages 113–138. Elsevier.

Niu, Z., Zhong, G., and Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62.

Noorjahan, M. and Punitha, A. (2020). An electronic travel guide for visually impaired–vehicle board recognition system through computer vision techniques. *Disability and Rehabilitation: Assistive Technology*, 15(2):238–241.

Olah, C. (2015). Understanding lstm networks. *Colah's blog*.

OMS (2014). Adolescence: A period needing special attention. *Health for the World's Adolescents: A second chance in the second decade*.

Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimarães, G. A. S., Cruz, G. O. R., Araujo, M. M., Santos, L. L., Cruz, M. A. S., Oliveira, E. L. S., Winkler, I., and Nascimento, E. G. S. (2023). Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, 7(1):1–31.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Parikh, A., Täckström, O., Das, D., and Uszkoreit, J. (2016). A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 1310–1318. PMLR.

Passamani, C. A. G., Neves, V. N., Júnior, L. J. L., Oliveira-Santos, T., Badue, C., and De Souza, A. F. (2022). A method to estimate covid-19 contamination risk based on social distancing and face mask detection using convolutional neural networks. In *2022 35th Conference on Graphics, Patterns and Images (SIBGRAPI)*, volume 1, pages 282–287.

Patel, K. and Parmar, B. (2022). Assistive device using computer vision and image processing for visually impaired; review and current status. *Disability and Rehabilitation: Assistive Technology*, 17(3):290–297.

Perez-Martin, J., Bustos, B., Guimaraes, S. J. F., Sipiran, I., Pérez, J., and Said, G. C. (2022). A comprehensive review of the video-to-text problem. *Artificial Intelligence Review*, pages 1–75.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2641–2649.

Ponti, M. A., Ribeiro, L. S. F., Nazare, T. S., Bui, T., and Collomosse, J. (2017). Everything you wanted to know about deep learning for computer vision but were afraid to ask. In *2017 30th Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T)*, pages 17–41. IEEE.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ravichandran, D., Pantel, P., and Hovy, E. (2005). Randomized algorithms and nlp: Using locality sensitive hash functions for high speed noun clustering. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 622–629.

Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint*.

Ridnik, T., Ben-Baruch, E., Noy, A., and Zelnik-Manor, L. (2021). Imagenet-21k pretraining for the masses. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1–12.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695. IEEE Computer Society.

Rothe, R., Timofte, R., and Van Gool, L. (2015). Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 10–15.

Rotstein, N., Bensaïd, D., Brody, S., Ganz, R., and Kimmel, R. (2024). FuseCap: Leveraging large language models for enriched fused image captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5689–5700. IEEE Computer Society.

Salvagno, M., Taccone, F. S., and Gerli, A. G. (2023). Artificial intelligence hallucinations. *Critical Care*, 27(1):180.

Savci, P. and Das, B. (2023). Comparison of pre-trained language models in terms of carbon emissions, time and accuracy in multi-label text classification using automl. *Heliyon*, 9(5):1–13.

Schuster, S., Krishna, R., Chang, A., Fei-Fei, L., and Manning, C. D. (2015). Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the 4th Workshop on Vision and Language*, pages 70–80. Association for Computational Linguistics.

Senjam, S. S., Foster, A., Bascaran, C., Vashist, P., and Gupta, V. (2020). Assistive technology for students with visual disability in schools for the blind in Delhi. *Disability and Rehabilitation: Assistive Technology*, 15(6):663–669.

Serengil, S. I. and Ozpinar, A. (2020). Lightface: A hybrid deep face recognition framework. In *ASYU*, pages 23–27.

Shah, H., Shah, R., Shah, S., and Sharma, P. (2021). Dangerous object detection for visually impaired people using computer vision. In *2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)*, pages 1–6.

Sharif, N., Bennamoun, M., Liu, W., and Shah, S. A. A. (2021). Subicap: Towards subword-informed image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3540–3549.

Simons, R. N., Gurari, D., and Fleischmann, K. R. (2020). "I hope this is helpful": Understanding crowdworkers' challenges and motivations for an image description task. In *Proceedings of the ACM on Human-Computer Interaction*, pages 1–26, New York, NY, USA. Association for Computing Machinery.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *3th International Conference on Learning Representations (ICLR)*, pages 1–14.

Stangl, A., Morris, M. R., and Gurari, D. (2020). Person, shoes, tree. Is the person naked? What people with vision impairments want in image descriptions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13. Association for Computing Machinery.

Steinmetz, J. D., Bourne, R. R., Briant, P. S., Flaxman, S. R., Taylor, H. R., Jonas, J. B., Abdoli, A. A., Abrha, W. A., Abualhasan, A., Abu-Gharbieh, E. G., et al. (2021). Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to vision 2020: the right to sight: an analysis for the global burden of disease study. *The Lancet Global Health*, 9(2):e144–e160.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 843–852.

Tabian, I., Fu, H., and Sharif Khodaei, Z. (2019). A convolutional neural network for impact detection and characterization of complex composite structures. *Sensors*, 19(22):4933.

Tang, W., Liu, D.-e., Zhao, X., Chen, Z., and Zhao, C. (2023). A dataset for the recognition of obstacles on blind sidewalk. *UAIS*, 22(1):69–82.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. (2023). Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008.

Vedantam, R., Zitnick, C., and Parikh, D. (2015). CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575. IEEE Computer Society.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164. IEEE Computer Society.

Wada, Y., Kaneda, K., Saito, D., and Sugiura, K. (2024). Polos: Multimodal metric learning from human feedback for image captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–14.

Wang, C., Yang, H., Bartz, C., and Meinel, C. (2016). Image captioning with deep bidirectional lstms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 988–997.

Wang, H. et al. (2019). Swell-and-shrink: Decomposing image captioning by transformation and summarization. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5226–5232.

Wang, H., Zhang, Y., and Yu, X. (2020). An overview of image caption generation methods. *Computational intelligence and neuroscience*, 2020.

Wang, J. and Gaizauskas, R. (2016). Cross-validating image description datasets and evaluation metrics. In *Proceedings of the 10th Language Resources and Evaluation Conference*, pages 3059–3066. European Language Resources Association.

Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. (2022). OFA: Unifying architectures, tasks, and modalities through a simple

sequence-to-sequence learning framework. In *Proceedings of the 39nd International Conference on Machine Learning (ICML)*, volume 162, pages 23318–23340. PMLR.

Wanga, H., Joseph, T., and Chuma, M. B. (2020). Social distancing: Role of smartphone during coronavirus (COVID-19) pandemic era. *International Journal of Computer Science and Mobile Computing*, 9(5):181–188.

Wiseman, S. and Rush, A. M. (2016). Sequence-to-sequence learning as beam-search optimization. In Su, J., Duh, K., and Carreras, X., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1296–1306. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 38–45, Online. Association for Computational Linguistics.

Wu, H. and Gu, X. (2015). Towards dropout training for convolutional neural networks. *Neural Networks*, 71:1–10.

Wu, H., Mao, J., Zhang, Y., Jiang, Y., Li, L., Sun, W., and Ma, W.-Y. (2019). Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6609–6618. IEEE Computer Society.

Wu, S., Wieland, J., Farivar, O., and Schiller, J. (2017). Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, page 1180–1192. Association for Computing Machinery.

Xie, Y., Zhou, L., Dai, X., Yuan, L., Bach, N., Liu, C., and Zeng, M. (2022). Visual clues: Bridging vision and language foundations for image paragraph captioning. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:17287–17300.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37, pages 2048–2057. PMLR.

Yang, L., Tang, K., Yang, J., and Li, L.-J. (2017). Dense captioning with joint inference and visual context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2193–2202. IEEE Computer Society.

Yang, X., Gao, C., Zhang, H., and Cai, J. (2020). Hierarchical scene graph encoder-decoder for image paragraph captioning. In *Proceedings of the 28th ACM International Conference on Multimedia*, page 4181–4189. Association for Computing Machinery.

Yin, G., Sheng, L., Liu, B., Yu, N., Wang, X., and Shao, J. (2019). Context and attribute grounded dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6241–6250. IEEE Computer Society.

Zha, Z.-J., Liu, D., Zhang, H., Zhang, Y., and Wu, F. (2022). Context-aware visual policy network for fine-grained image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, 44(2):710–722.

Zhang, J. et al. (2020a). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 11328–11339. PMLR.

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020b). BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations (ICLR)*, pages 1–43. OpenReview.net.

Zhao, B., Li, X., and Lu, X. (2017). Hierarchical recurrent neural network for video summarization. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 863–871.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint*, pages 1–124.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Zhu, C., Yang, Z., Gmyr, R., Zeng, M., and Huang, X. (2021). Leveraging lead bias for zero-shot abstractive news summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1462–1471. Association for Computing Machinery.