Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Departamento de Computação e Sistemas

Uma Revisão Narrativa sobre Modelos de IA Generativa com aplicação para o Problema de Visual Paragraph Generation

Tatiane Vitória de Oliveira

TRABALHO DE CONCLUSÃO DE CURSO

ORIENTAÇÃO:
Alexandre Magno de Sousa

Agosto, 2025
João Monlevade–MG

Tatiane Vitória de Oliveira

Uma Revisão Narrativa sobre Modelos de IA Generativa com aplicação para o Problema de Visual Paragraph Generation

Orientador: Alexandre Magno de Sousa

Monografia apresentada ao curso de Sistemas de Informação do Instituto de Ciências Exatas e Aplicadas, da Universidade Federal de Ouro Preto, como requisito parcial para aprovação na Disciplina "Trabalho de Conclusão de Curso II".

Universidade Federal de Ouro Preto

João Monlevade

Agosto de 2025

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

O484r Oliveira, Tatiane Vitoria de.
Uma revisão narrativa sobre modelos de IA generativa com aplicação para o problema de visual paragraph generation. [manuscrito] / Tatiane Vitoria de Oliveira. - 2025.
73 f.: il.: color., gráf., tab..

Orientador: Prof. Dr. Alexandre Magno de Sousa.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Aplicadas. Graduação em Sistemas de Informação .

1. Aprendizado do computador. 2. Inteligência artificial. 3. Processamento de imagens. 4. Processamento de linguagem natural (Computação). 5. Redes neurais (Computação). 6. Visão por computador. I. Sousa, Alexandre Magno de. II. Universidade Federal de Ouro Preto. III. Título.

CDU 004.8

Bibliotecário(a) Responsável: Flavia Reis - CRB6-2431

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E APLICADAS
DEPARTAMENTO DE COMPUTACAO E SISTEMAS

FOLHA DE APROVAÇÃO

Tatiane Vitória de Oliveira

Uma Revisão Narrativa sobre Modelos de IA Generativa
com aplicação para o Problema de Visual Paragraph Generation

Monografia apresentada ao Curso de Sistemas de Informação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação

Aprovada em 08 de setembro de 2025

Membros da banca

Prof. Dr. Alexandre Magno de Sousa - Orientador (Universidade Federal de Ouro Preto)
Prof. Dr. Eduardo da Silva Ribeiro (Universidade Federal de Ouro Preto)
Prof. Dr. Elton Máximo Cardoso (Universidade Federal de Ouro Preto)

Alexandre Magno de Sousa, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 22/09/2025

Documento assinado eletronicamente por **Alexandre Magno de Sousa**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 25/09/2025, às 08:46, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0984175** e o código CRC **01C05367**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3808-0819 - www.ufop.br

Dedico este trabalho a todos que, com carinho, apoio e palavras de incentivo, estiveram ao meu lado e contribuíram para a realização desta importante etapa da minha vida.

Agradecimentos

Agradeço profundamente à minha mãe, por todo amor, paciência e apoio incondicional em cada etapa da minha vida. À minha irmã, pela companhia, incentivo e por compartilhar comigo cada conquista e desafio. E à minha tia Zezé, que sempre acreditou em mim e me apoiou em todos os momentos, sendo uma presença constante e inspiradora. Vocês são parte fundamental desta jornada e desta vitória.

"Happiness can be found, even in the darkest of times, if one only remembers to turn on the light."

— Albus Dumbledore,
in: Harry Potter and the Prisoner of Azkaban.

Resumo

Este trabalho investiga a aplicação de modelos de Inteligência Artificial Generativa no problema de Visual Paragraph Generation, com ênfase na arquitetura Transformer, que se consolidou como a abordagem mais utilizada em razão de sua capacidade de lidar com dependências de longo alcance, escalabilidade e paralelização. O objetivo central foi analisar os principais modelos descritos na literatura e demonstrar, por meio de experimentos, a aplicação prática de um modelo pré-treinado baseado em Transformer para este contexto. A metodologia foi desenvolvida em cinco etapas complementares. Primeiramente, realizou-se uma revisão da literatura sobre os fundamentos de Inteligência Artificial Generativa, abordando sua evolução, aplicações e implicações. Em seguida, foi estudada a arquitetura Transformer e apresentado como funciona o mecanismo de atenção e alguns modelos de referência. Na terceira etapa, analisaram-se os trabalhos relacionados ao problema de Visual Paragraph Generation, identificando avanços, pontos fortes e limitações. A quarta etapa consistiu na seleção do Vision Transformer como modelo pré-treinado para experimentação, justificando sua escolha pela relevância e viabilidade de aplicação. Por fim, conduziram-se experimentos para exemplificar o uso do modelo selecionado no problema, com análise quantitativa dos resultados. Os experimentos demonstraram desempenho estável em termos de acurácia, mas apresentaram limitações relevantes em métricas semânticas, tais como Bilingual Evaluation Understudy, revelando a dificuldade do modelo em produzir descrições ricas e diversificadas. Além disso, restrições técnicas, como a alta demanda de GPU e memória, limitaram a exploração de ajustes mais sofisticados e treinos de maior escala. Conclui-se que a arquitetura Transformer representa a abordagem mais promissora para Visual Paragraph Generation, justificando seu amplo uso em aplicações de Inteligência Artificial Generativa. Contudo, também se evidenciam desafios práticos que precisam ser superados para ampliar a qualidade e a aplicabilidade dessas soluções. O estudo contribui de forma exploratória ao consolidar a base teórica, discutir trabalhos relacionados e demonstrar um exemplo prático, abrindo caminhos para pesquisas futuras.

Palavras-chaves: Inteligência Artificial. Inteligência Artificial Generativa. Processamento de Linguagem Natural. Visual Paragraph Generation. Transformers. Vision Transformers.

Abstract

This work investigated the application of Generative Artificial Intelligence models to the problem of Visual Paragraph Generation, focusing on the Transformer architecture, which has become the most widely used approach due to its ability to handle wide-ranging dependencies, as well as its scalability and parallelization. The main objective was to analyze the main models described in the literature and to demonstrate through experiments the practical application of a pre-trained Transformer-based model in this context. The methodology was developed in five complementary steps. First, a literature review was conducted on the foundations of Generative Artificial Intelligence, covering its development, applications and implications. Secondly, the Transformer architecture was examined, highlighting the operation of the attention mechanism and presenting reference models. The third phase analyzed related work on the problem of visual paragraph generation and identified advances, strengths and limitations. In the fourth phase, the Vision Transformer was selected as a pre-trained model for experimentation, which was justified by its relevance and applicability. Finally, experiments were conducted to illustrate the application of the chosen model to the problem, with a quantitative analysis of the results. The experiments showed stable performance in terms of accuracy, but relevant limitations on semantic metrics such as the Bilingual Evaluation Understudy, highlighting the difficulty of the model in producing rich and diversified descriptions. In addition, technical limitations, such as high GPU and memory requirements, restricted the exploration of more sophisticated adaptations and extensive training. We conclude that the Transformer architecture is the most promising approach for the Visual Paragraph Generation, justifying its wide use in Generative Artificial Intelligence applications. However, practical challenges still need to be overcome to improve the quality and applicability of such solutions. This study makes an exploratory contribution by consolidating the theoretical foundations, discussing related work and demonstrating a practical example, thus opening avenues for future research.

Key-words: Artificial Intelligence. Generative Artificial Intelligence. Visual Paragraph Generation. Natural Language Processing. Transformers. Vision Transformer.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

IA Inteligência Artificial

DL *Deep Learning*

ML *Machine Learning*

PLN Processamento de Linguagem Natural

CNN *Convolutional Neural Networks*

VAE *Variational Autoencoder*

GAN *Generative Adversarial Networks*

RNN *Recurrent Neural Network*

BERT *Bidirectional Encoder Representations from Transformers*

LSTM *Long Short-Term Memory*

ViT *Vision Transformer*

BLEU *Bilingual Evaluation Understudy*

MLP *Multi-layer Perceptron*

LLM *Large Language Model*

RSICD *Remote Sensing Image Captioning Dataset*

AIGC *Artificial Intelligence-Generated Content*

IC Intervalo de Confiança

Sumário

1 Introdução

A Inteligência Artificial (IA) Generativa tem despertado crescente interesse como ferramenta promissora para a criação de conteúdo inovador. Trata-se de uma área dentro do *Deep Learning* (DL) dedicada à geração de imagens, textos, músicas e vídeos (FOSTER, 2022). Essa área utiliza algoritmos e modelos previamente treinados com técnicas de *Machine Learning* (ML) em conjuntos de dados específicos do contexto da aplicação (ALTO, 2023).

Segundo Foster nos últimos anos, avanços notáveis têm sido feitos na aplicação do aprendizado de máquina em tarefas generativas, como a geração de imagens faciais (FOSTER, 2022). Embora historicamente seja mais fácil usar modelos de aprendizado de máquina para resolver problemas específicos, empresas estão começando a explorar o potencial da IA Generativa para atender a necessidades específicas de negócios, tais como a criação de conteúdo original. Isso não apenas amplia as possibilidades de inovação, mas também sugere um futuro em que a IA generativa desempenhará um papel significativo em várias indústrias, incluindo jogos, cinematografia e marketing.

Nesse contexto, a geração de parágrafos descritivos a partir de imagens, do inglês, Visual Paragraph Generation, torna-se uma área de estudo crucial e desafiadora, pois consiste em transformar imagens em texto ou texto em imagens de forma coerente e informativa (ZHENG; WANG; WANG, 2022). Tradicionalmente, métodos baseados em técnicas de processamento de imagens e processamento de linguagem natural têm sido utilizados para abordar esse problema. No entanto, essas abordagens muitas vezes enfrentam limitações na captura e expressão de informações complexas contidas nas imagens, o que resulta em descrições genéricas ou imprecisas (KRAUSE et al., 2017).

O problema de Visual Paragraph Generation consiste na necessidade de desenvolver sistemas capazes de analisar imagens e gerar parágrafos descritivos que capturem com precisão os elementos visuais e conceituais presentes nas mesmas (XU et al., 2020). Isso requer não apenas a identificação de objetos e cenas, mas também a habilidade de contextualizá-los em linguagem natural de forma coerente e significativa (KRAUSE et al., 2017).

A importância de resolver esse problema é evidente em diversas áreas, incluindo acessibilidade, busca de imagens e geração automática de conteúdo (KRAUSE et al., 2017). Uma solução eficaz poderia melhorar significativamente a capacidade de sistemas de IA em compreender e descrever o conteúdo visual proporcionando benefícios tangíveis para os usuários finais (XU et al., 2020). Além disso, a integração de modelos de curiosidade computacional na geração de parágrafos pode aumentar a diversidade e o interesse das

descrições textuais produzidas, o que pode resultar em experiências mais envolventes e informativas (LUO et al., 2019).

1.1 Motivação e Justificativa

A geração de descrições detalhadas e contextualizadas de imagens, conhecida como Visual Paragraph Generation, tem se tornado um tema de grande relevância na área de visão computacional e Processamento de Linguagem Natural (PLN). Esse tipo de abordagem é essencial para aplicações que demandam uma interpretação mais profunda do conteúdo visual, como sistemas de acessibilidade, recuperação de informações multimodais, análise de vídeos e imagens em larga escala e até mesmo suporte a processos de tomada de decisão.

Diversos estudos têm contribuído significativamente para esse campo ao buscar formas de superar as limitações dos métodos tradicionais de Visual Paragraph Generation, que normalmente geram descrições curtas e pouco informativas. Por exemplo, o trabalho de Che *et al.* mostra a importância de explorar as relações entre os objetos de uma imagem, em vez de analisá-los de forma isolada. Esse avanço abre espaço para descrições mais ricas e coerentes, aproximando os modelos da forma como humanos interpretam cenas visuais (CHE et al., 2020). Por sua vez, o estudo de Xie *et al.* propõe uma abordagem para extrair informações significativas de imagens de forma a combinar modelos pré-treinados de visão computacional e modelos de linguagem (XIE et al., 2022). Os autores utilizam pistas visuais geradas a partir da representação semântica da imagem e o método utilizado gera descrições abrangentes do conteúdo visual por meio de um modelo de linguagem. Por fim, a qualidade das descrições é avaliada e a eficácia dessa abordagem estruturada é comprovada por meio dos resultados alcançados pelos autores.

Um modelo de legenda de parágrafos para vídeos não editados foi proposto por Song *et al.* (SONG; CHEN; JIN, 2021). Esse modelo usa memórias de vídeo dinâmicas para descrever eventos de forma coerente e diversificada e emprega uma estratégia de treinamento para melhorar a variedade linguística dos parágrafos. Resultados experimentais mostram que o modelo supera abordagens atuais em métricas de precisão e diversidade sem necessidade de anotações de limites de evento. Ainda, o trabalho de Yant *et al.* aborda o problema de gerar automaticamente um parágrafo descritivo a partir de uma imagem, esse problema é considerado mais complexo pois os parágrafos são, geralmente, mais longos, informativos e linguisticamente mais complicados (YANG; YANG; HSU, 2021). A principal questão abordada é como gerar frases consistentes e que não sejam contraditórias em um parágrafo. Para resolver isso,os autores propõem um método que incorpora a coerência espacial dos objetos em um modelo de geração de linguagem. O estudo mostra que essa abordagem produz características eficazes dos objetos para a geração de parágrafos de

imagem e que também supera os métodos existentes. O trabalho destaca a importância de integrar informações visuais e linguísticas para descrever conteúdo visual de forma mais detalhada e precisa.

Por fim, a pesquisa de Luo *et al.* (LUO et al., 2019) propõe um framework de aprendizado por reforço dirigido orientado à curiosidade para melhorar a diversidade e a precisão da descrição gerada para Visual Paragraph Generation. O modelo apresenta soluções mais precisas as quais são raramente identificadas no contexto da imagem a ser descrita ao invés de soluções genéricas. Experimentos foram realizados no dataset Stanford Image-paragraph demonstram a eficácia e eficiência do framework proposto melhorando o desempenho em 38.4% se comparado com métodos do estado da arte.

Esses estudos demonstram não apenas a evolução das técnicas de geração de parágrafos descritivos, mas também a relevância dessa área para o desenvolvimento de sistemas que compreendem e comunicam conteúdo visual de maneira cada vez mais humana, detalhada e útil. Essa trajetória de avanços reforça a motivação para pesquisas que busquem soluções inovadoras e que ampliem as possibilidades de aplicação dessa tecnologia em diversos contextos.

1.2 Definição do Problema

A geração de parágrafos visuais é um campo interdisciplinar e emergente entre a inteligência artificial, a visão computacional e o PLN. É uma abordagem de descrição de imagens que produz um parágrafo longo, narrativo e informativo, contendo uma história coerente e unificada com detalhes ricos que descrevem o conteúdo semântico de uma imagem (ALMOHSEN, 2023). O processo de geração de parágrafos se propõe a ser diferente de outras abordagens básicas que apenas concatenam múltiplas frases curtas geradas por abordagens tradicionais de legendagem de imagens. Isso porque essas abordagens não abordam as complexidades dos parágrafos, como a coerência entre frases consecutivas, uma estrutura global consistente ou a diversidade nas descrições de parágrafos (CHATTERJEE; SCHWING, 2018).

A geração automática de legendas para imagens ou Visual Paragraph Generation é um problema desafiador e exige não apenas a identificação precisa dos objetos em uma imagem, mas também a expressão dessas relações em linguagem natural. É essencial que os modelos de geração de legendas sejam capazes de compreender a cena representada na imagem e traduzi-la em texto significativo (KRAUSE et al., 2017). No entanto, isso enfrenta problemas complexos tais como a necessidade de representar corretamente os objetos e suas relações na imagem, bem como a capacidade de expressar essas relações de forma coerente em linguagem humana (XU et al., 2015). Dessa forma, o desafio é encontrar um equilíbrio entre a precisão na descrição da imagem e a fluência na linguagem.

É essencial usar abordagens com mecanismos de atenção, que ajudem o modelo a focar nas partes mais relevantes da imagem ao gerar a legenda. Isso ajuda a melhorar a qualidade e a interpretabilidade das legendas geradas e permite uma compreensão mais profunda do processo de geração de legendas automáticas para imagens (XU et al., 2015).

1.3 Objetivos Geral e Específicos

Este projeto pretende apresentar uma revisão narrativa sobre IA Generativa na aplicação do problema de Visual Paragraph Generation. Diante disso, o objetivo geral deste trabalho é investigar os modelos de IA Generativa existentes na literatura para solução do problema de Visual Paragraph Generation. Para alcançar o objetivo geral, os seguintes objetivos específicos são definidos:

- Realizar uma revisão da literatura sobre os principais modelos de IA Generativa para Visual Paragraph Generation;
- Identificar um modelo baseado em Transformer que seja mais utilizado mediante sua viabilidade de aplicação em termos de custo benefício para o problema de Visual Paragraph Generation;
- Realizar experimentos com o modelo selecionado para exemplificar o uso de aplicação e demonstrar suas possíveis limitações para solução.

1.4 Metodologia

A metodologia deste trabalho foi estruturada de forma a transformar os objetivos específicos em etapas práticas que, em conjunto, possibilitam atingir o objetivo geral da pesquisa. A ideia central é que cada etapa contribua de maneira incremental, formando um percurso coerente desde a fundamentação teórica até a análise dos resultados obtidos. Assim, busca-se não apenas compreender os conceitos e modelos de IA Generativa, mas também aplicá-los em um cenário prático voltado ao problema de Visual Paragraph Generation.

As etapas metodológicas podem ser descritas da seguinte forma:

- Revisão da literatura: inicialmente, será realizada uma investigação bibliográfica acerca dos principais fundamentos de IA Generativa, com o intuito de consolidar o embasamento teórico e situar o trabalho no estado da arte da área;
- Análise de modelos existentes: em seguida, serão explorados os principais modelos de IA Generativa baseados na arquitetura Transformers, identificando suas características, aplicações e limitações;

- Comparação entre modelos: a partir dessa análise, será conduzida uma avaliação comparativa entre os modelos estudados, de modo a evidenciar pontos positivos e negativos que orientem a escolha de uma abordagem mais adequada;
- Seleção de um modelo pré-treinado: identificar um modelo pré-treinado que se mostre viável para a resolução do problema de Visual Paragraph Generation, considerando critérios de aplicabilidade e eficiência;
- Exploração do dataset Flickr30k por meio de um exemplo de aplicação: posteriormente, será utilizado o dataset Flickr30k (PLUMMER et al., 2015) como base para a geração de legendas visuais. O objetivo é possibilitar o uso do modelo pré-treinado de forma a viabilizar uma solução de baixo custo computacional;
- Análise de resultados: por fim, os resultados obtidos serão analisados de forma crítica, contemplando tanto o desempenho alcançado quanto as limitações observadas. Essa etapa busca fornecer subsídios para discussões futuras e para possíveis melhorias no desenvolvimento de soluções similares.

1.5 Resultados e Contribuições

Na revisão da literatura, foi possível consolidar os fundamentos da Inteligência Artificial Generativa e identificar como diferentes arquiteturas têm sido aplicadas ao problema de Visual Paragraph Generation. Esse estudo evidenciou o papel central da arquitetura Transformer, que se consolidou como abordagem dominante devido à sua escalabilidade, flexibilidade e capacidade de lidar com dependências de longo alcance. Além disso, a análise crítica dos trabalhos relacionados permitiu identificar tanto avanços significativos na área quanto limitações persistentes, como a dificuldade em gerar descrições semanticamente ricas e contextualmente adequadas.

Durante a fase experimental, foi utilizado um modelo com codificador visual, o *Vision Transformer* (ViT), e Transformer puro como decodificador para a tarefa de geração de legendas, a partir de um código-base já existente. Os experimentos serviram como etapa de teste e validação metodológica, com o objetivo de compreender o comportamento do modelo e estabelecer um ponto de partida para aprimoramentos futuros.

Os resultados obtidos na fase de finetunning por meio de validação cruzada identificou o melhor modelo o qual apresentou o melhor equilíbrio entre as métricas, com função de perda média de 1,34, acurácia de 74,89% e pontuação *Bilingual Evaluation Understudy* (BLEU) média de 0,0337. De forma geral, a acurácia manteve-se estável entre as configurações (73% a 75%), mostrando que o modelo conseguiu capturar padrões básicos entre imagens e descrições. No entanto, as pontuações BLEU, que variaram entre 0,0184 e

0,0337, evidenciam limitações significativas na qualidade semântica e na proximidade com as legendas de referência.

A principal contribuição deste trabalho não reside na obtenção de métricas desempenho ótimas para o exemplo de aplicação, mas na análise crítica que combina revisão narrativa e experimentação prática. A partir dela, foi possível: (*i*) destacar a relevância e predominância da arquitetura Transformer no campo da IA Generativa; (*ii*) discutir pontos fortes e limitações dos trabalhos existentes na literatura; (*iii*) exemplificar, por meio de um estudo prático, como um modelo pré-treinado pode ser aplicado ao problema de Visual Paragraph Generation e quais barreiras técnicas ainda precisam ser superadas. Dessa forma, este estudo contribui tanto para a sistematização teórica da área quanto para a indicação de caminhos de aprimoramento em pesquisas futuras.

1.6 Estrutura da monografia

Este trabalho está organizado conforme descrito a seguir. O próximo capítulo apresenta a fundamentação teórica tais como os fundamentos da inteligência artificial generativa e a arquitetura de redes neurais Transformers. O capítulo 3 mostra os trabalhos relacionados que fornecem uma perspectiva geral da literatura existente, analisando alguns estudos que se relacionam diretamente com o tema. O Capítulo 4 apresenta a metodologia de desenvolvimento e os resultados alcançados. Para isso, descreve o planejamento e a execução do trabalho, em que consta a definição do *dataset*, o modelo de IA generativa e as métricas de avaliação, além da apresentação e análise dos resultados dos experimentos. Por fim, o Capítulo 5 apresenta as conclusões e trabalhos futuros, sintetiza os dados obtidos durante a pesquisa, discute as implicações e aponta direções para futuras investigações.

2 Revisão da Literatura

Este capítulo tem como objetivo apresentar a fundamentação teórica que serve de base para o estudo da Inteligência Artificial Generativa, com foco especial na arquitetura Transformer. Abordaremos desde os conceitos fundamentais da IA Generativa, sua evolução e aplicações, até os processos de funcionamento dos Transformers, destacando o mecanismo de atenção. A compreensão desses tópicos é importante para o entendimento das tecnologias que fomentam os avanços recentes em diversas áreas, como processamento de linguagem natural e geração de conteúdo.

A estrutura deste capítulo está organizada conforme descrito a seguir. A Seção 2.1 trata dos fundamentos sobre IA Generativa. Nessa seção, é apresentada a definição do que é IA Generativa, apresentando sua evolução histórica e destacando suas principais aplicações. Também são apresentadas as empresas líderes nessa área, as tecnologias subjacentes e as importantes considerações éticas, além de discutir as vantagens e desvantagens dos modelos generativos. Por sua vez, na Seção 2.2 descreve a arquitetura Transformer fazendo uma introdução sobre o tema e ainda descreve o funcionamento de sua arquitetura. Também apresenta os conceitos do mecanismo de atenção, elemento central para o desempenho desses modelos e, por fim, descreve dos principais modelos baseados em Transformers. Finalmente, a Seção 2.3 sumariza as considerações finais sobre este capítulo.

2.1 Fundamentos de IA Generativa

Os modelos de IA generativa representam uma classe fascinante de algoritmos de inteligência artificial capazes de criar conteúdo novo e original. Ao contrário dos modelos convencionais, que se focam em reconhecer padrões e fazer previsões com base em dados existentes, os modelos dessa classe podem produzir dados totalmente novos, similares aos exemplos nos quais foram treinados (RANI; SINGH; KHANNA, 2023). A IA generativa pode produzir não só textos, mas também formas multimídia, como imagens, vídeos, áudios e códigos. A Figura 1 mostra um exemplo de geração de imagens.

Essa tecnologia faz uso de modelos avançados de aprendizagem automática, particularmente redes neurais de aprendizagem profunda, para analisar e aprender a partir de grandes quantidades de dados, identificar padrões e, em seguida, utilizar esses padrões para gerar conteúdos novos e semelhantes (CAO et al., 2023).

A base para a IA generativa é a rede neural profunda, uma arquitetura neural com várias “camadas” de neurônios, incluindo algumas que ficam ocultas entre as camadas de entrada e saída. Neurônios, os blocos fundamentais de uma rede neural, são funções

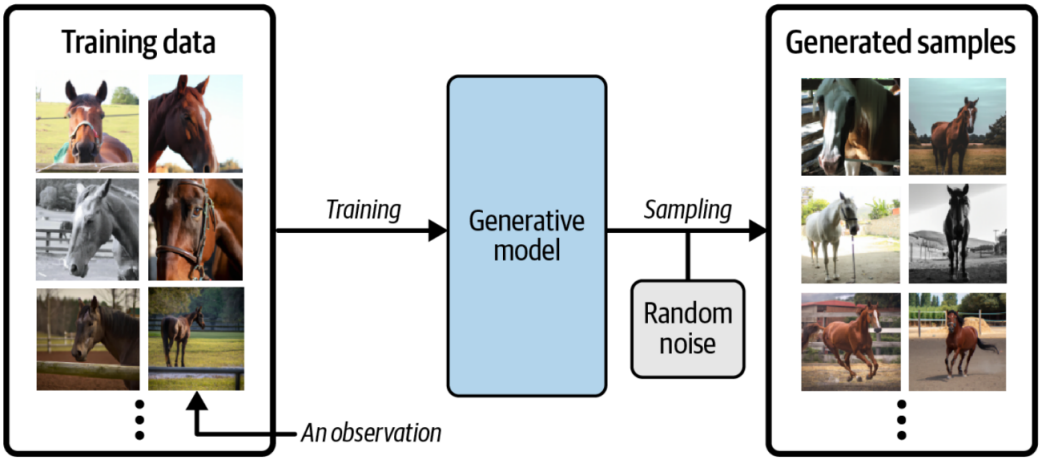


Figura 1 – Modelo de IA Generativa treinado para gerar fotos realistas de cavalos. Fonte: (FOSTER, 2023).

matemáticas que imitam, de maneira simplificada, o comportamento dos neurônios biológicos. Embora as redes neurais tenham sido introduzidas já na década de 1960, só a partir de cerca de 2009, quando uma rede neural recorrente, um tipo específico de rede neural profunda, venceu várias competições de reconhecimento de caligrafia pela primeira vez, é que elas conseguiram superar outros métodos de aprendizado de máquina mais simples (STRATIS, 2023).

Essa tecnologia encontra aplicações em diversos campos como arte, entretenimento, design e pesquisa científica. Ela possibilita a criação de imagens realistas, síntese de novas composições musicais, desenvolvimento de personagens realistas em jogos de vídeo, e auxilia na descoberta de medicamentos ao projetar moléculas novas. A capacidade de gerar conteúdo de forma autônoma abre novas possibilidades para a criatividade humana e expande as capacidades das máquinas (RAMDURAI; ADHITHYA, 2023).

A inteligência artificial generativa é crucial por sua capacidade de gerar conteúdo criativo, além de facilitar a augmentação de dados sintéticos em áreas onde a coleta de dados reais é difícil. Ela também é utilizada para simulação e modelagem de sistemas complexos, planejamento estratégico, personalização de recomendações e assistência criativa. Além disso, apoia descobertas científicas ao sugerir hipóteses e explorar novos territórios de pesquisa, preenchendo lacunas em dados incompletos para melhorar a tomada de decisões (RAMDURAI; ADHITHYA, 2023).

2.1.1 Evolução Histórica

Apesar de parecer que a IA Generativa surgiu do nada nos últimos um ou dois anos, a pesquisa nesse campo tem sido contínua por décadas (STRATIS, 2023). Os modelos generativos iniciais, como Modelos Ocultos de Markov e Modelos de Mistura Gaussiana,

evoluíram significativamente com o aprendizado profundo. Em processamento de linguagem natural, a geração de sentenças melhorou com redes neurais recorrentes e transformadores, permitindo a modelagem de dependências longas (CAO et al., 2023). As Redes Neurais Celulares/Não Lineares foram propostas em 1988 por Leon O. Chua e Lee Yang como uma arquitetura de computador fácil de implementar e programar para processamento de imagens e sinais (ZARáNDY et al., 2015).

Em 1991, Sepp Hochreiter introduziu a *Long Short-Term Memory* (LSTM) que é um tipo de rede neural recorrente que pode aprender relacionamentos de longo prazo em dados sequenciais. Já em 2014 o *Variational Autoencoder* (VAE) foi introduzido por Diederik Kingma e Max Welling. É um tipo de modelo que pode aprender representações de dados e gerar novos dados com base nessas representações aprendidas. Além do VAE, a *Generative Adversarial Networks* (GAN) também foi introduzida em 2014 pelo Ian Goodfellow e seus colegas. É um tipo de modelo generativo que compreende duas redes neurais: um gerador e um discriminador. O gerador visa gerar dados realistas, enquanto o discriminador visa diferenciar entre dados reais e falsos. Por sua vez, no ano de 2015 mais um modelo apareceu, o modelo de difusão foi proposto pelo Yann LeCun e sua equipe. É um modelo generativo que aprende a reverter um processo que gradativamente transforma dados em ruído. Então, o modelo Transformer foi apresentado em 2017 pelo Ashish Vaswani e sua equipe. Esse modelo é um projeto de rede neural que aproveita mecanismos de atenção para aprender a partir de informações sequenciais, como linguagem ou fala (CAO et al., 2023).

2.1.2 Principais Aplicações

Um modelo de geração de texto é um modelo de aprendizado de máquina que emprega redes neurais, especialmente a arquitetura de Transformers, para criar textos contextualmente relevantes com base em padrões linguísticos aprendidos com grandes conjuntos de dados linguísticos. Esses modelos são treinados com grandes volumes de dados textuais para modelar e compreender conceitos complexos em qualquer idioma, como gramática, vocabulário, frases e estilos.

Eles têm o potencial de aumentar a produtividade humana nos processos de negócios atuais, automatizando a criação de conteúdo em várias indústrias, incluindo relatórios, resumos, e-mails, entre outros. Além disso, possibilitam um nível superior de personalização na comunicação entre empresas e clientes. Esses modelos também são capazes de resumir artigos extensos e livros, tornando as informações mais compreensíveis e acessíveis em menos tempo. A área em que esses modelos têm impactado de maneira mais significativa é a tradução automática, aproveitando grandes corpora multilíngues para executar uma variedade de tarefas em idiomas além do inglês.

Geralmente utilizados para ampliar as capacidades humanas, esses modelos funcio-

nam através de uma abordagem de colaboração homem-máquina, ajudando na geração de ideias e na redução do espaço de soluções disponíveis, ao invés de serem meros geradores de soluções definitivas. Com o tempo, a diversidade de casos de uso continua a crescer, permitindo que os humanos se concentrem em tarefas de maior valor ou realizem suas atividades cotidianas com maior eficiência (PANDEY et al., 2024).

Para geração de texto incluem GPT-4, LaMDA, LLaMA e BLOOM (RANI; SINGH; KHANNA, 2023). As técnicas de texto para imagem evoluíram significativamente, permitindo a geração de imagens a partir de descrições textuais. Anteriormente, legendar imagens, uma abordagem de imagem para texto, era mais prevalente. No entanto, com o surgimento de aplicações notáveis como DALL-E e Midjourney, juntamente com outros modelos existentes, a síntese de texto para imagem ganhou destaque. Diversas arquiteturas, como GANs, difusão, VAEs e Transformers, são empregadas para facilitar a geração de texto para imagem. O conteúdo de uma imagem pode ser descrito usando técnicas *Artificial Intelligence-Generated Content* (AIGC), que aproveitam o poder da visão computacional e do processamento de linguagem natural. Um modelo notável neste campo é o Show and Tell: Neural Image Caption, desenvolvido pela equipe de pesquisa do Google, que serve como base para aplicações de legenda de imagem.

O processo arquitetural do Neural Image Caption envolve o uso de *Convolutional Neural Networks* (CNN) e LSTM, um tipo de *Recurrent Neural Network* (RNN). A CNN atua como um codificador, extraindo características visuais da imagem. Essas características capturam as informações salientes necessárias para gerar legendas significativas. As características extraídas são então passadas para a LSTM, que atua como decodificador e gera uma sequência de palavras ou uma descrição com base nas características visuais obtidas da CNN. No campo do processamento e síntese de imagens, técnicas de imagem para imagem oferecem uma ampla variedade de tarefas e modelos que nos permitem manipular e transformar imagens de diversas maneiras, incluindo a síntese de imagens altamente realistas. Uma técnica poderosa para gerar imagens sintéticas altamente realistas e diversas é o StyleGAN. Sua arquitetura inclui uma rede geradora que produz imagens sintéticas com base em um mapeamento aprendido de um espaço latente para o espaço da imagem. Um aspecto intrigante do StyleGAN é a introdução da normalização de instância adaptativa, que permite controle sobre diferentes aspectos do estilo e aparência da imagem, resultando na geração de imagens sintéticas altamente realistas e visualmente diversas (BANDI; ADAPA; KUCHI, 2023).

Os modelos de texto para imagem alcançaram progressos notáveis na geração de conteúdo visual com base em descrições textuais. Construindo sobre esse avanço, os pesquisadores agora voltaram sua atenção para a geração de texto para vídeo. Nos primeiros trabalhos neste campo, uma técnica notável que surgiu foi o uso de redes generativas adversariais temporais condicionadas a legendas (TGANs-C). TGANs-C focou na geração

de sequências de vídeo a partir de descrições textuais. Expansões subsequentes levaram ao desenvolvimento de diversos modelos baseados em difusão para geração de texto para vídeo. Uma técnica significativa é *"Make-A-Video"* da Meta AI. Esta abordagem utiliza dados pareados texto-imagem para capturar a aparência visual e descrições do mundo, enquanto emprega filmagens de vídeo não supervisionadas para entender dinâmicas de movimento. No modelo *"Make-A-Video"*, a entrada de texto passa por um decodificador para criar embeddings de imagem. Essas imagens geradas são então interpoladas para influenciar os quadros por segundo no vídeo resultante, utilizando camadas espaciotemporais para produzir saída de vídeo de alta resolução. Outro modelo notável é o IMAGEN VIDEO, introduzido pelo Google, que vai além da geração de vídeos simplesmente guiados por texto, permitindo a criação de animações de texto com diversos estilos artísticos. No IMAGEN VIDEO, entradas textuais são codificadas em embeddings textuais usando o codificador de texto T5. O modelo de difusão de vídeo é então utilizado para gerar um vídeo de 16 quadros, que é refinado ainda mais usando técnicas de super-resolução espacial e super-resolução temporal. O Tune-a-video, outro modelo baseado em difusão, utiliza um único par texto-vídeo para treinar o gerador T2V, uma técnica conhecida como ajuste de vídeo em uma única passagem. Este modelo possui grande potencial para várias aplicações, incluindo edição de objetos (por exemplo, substituição de uma zebra por um cavalo na entrada de vídeo fornecida), troca de fundo e transferência de estilo. PHENAKI é um modelo de geração de vídeo que utiliza uma arquitetura de Transformer bidirecional. Ao usar descrições textuais como entrada, ele tem a capacidade de gerar sequências de vídeo, destacando-se na geração de vídeos que correspondem a diferentes prompts textuais variáveis no tempo, permitindo saídas dinâmicas e diversas (BANDI; ADAPA; KUCHI, 2023).

Recentemente, a inteligência artificial generativa tem atraído uma atenção sem precedentes tanto na academia quanto na indústria. Vale ressaltar que o sucesso do modelo de difusão em visão computacional inspirou inúmeros trabalhos para geração de fala. A fala permite aos humanos expressar seus pensamentos e se comunicar de maneira precisa e eficiente. Portanto, a síntese de fala é um componente indispensável em sistemas de IA modernos. Especificamente, as tarefas de texto para fala e aprimoramento de fala são duas das principais tarefas ativas, que geram fala a partir de um texto fornecido e melhoram a qualidade de uma fala existente, respectivamente. O desenvolvimento da tarefa de texto para fala pode ser dividido em três estágios: trabalhos iniciais (por exemplo, síntese por formantes e síntese concatenativa), métodos baseados em síntese paramétrica estatística e estágio baseado em redes neurais. Mais recentemente, o modelo de difusão tem atraído grande atenção em múltiplos campos (por exemplo, visão computacional) e também tem sido aplicado na tarefa de texto para fala (ZHANG et al., 2023).

2.1.3 Principais Empresas de Tecnologias de IA Generativa

O ChatGPT é um chatbot baseado em inteligência artificial e aprendizado de máquina, lançado pela OpenAI, um laboratório de pesquisa fundado em 2015. A OpenAI realizou grandes avanços no ramo de inteligência artificial e lançou muitos produtos de aprendizado de máquina, como o ChatGPT e o DALL-E. O ChatGPT utiliza a arquitetura GPT-3, que pode gerar texto natural semelhante ao humano. Antes do GPT-3, a OpenAI desenvolveu os modelos GPT-1 e GPT-2. O GPT-1 foi lançado em junho de 2018 e o GPT-2 em fevereiro de 2019. O GPT-1 era capaz de gerar respostas gramaticalmente corretas e coerentes, mas essas respostas frequentemente eram repetitivas e careciam de diversidade. O GPT-2, por sua vez, era mais diverso, mas seu modelo completo inicialmente não foi lançado pela OpenAI devido ao seu potencial de uso indevido. Posteriormente, a OpenAI liberou o GPT-2 com algumas restrições. O ChatGPT pode ser aplicado em diversos campos. Ele pode responder perguntas, ajudar em trabalhos criativos como escrever ensaios e artigos, e gerar e depurar código. Na educação, resolve dúvidas de matemática e ciências rapidamente e pode traduzir textos entre várias línguas. Além disso, pode criar novos chatbots, resolver problemas de geofísica e realizar análises em ciências sociais, filosofia e finanças (SINGH; KUMAR; MEHRA, 2023).

Gemini, um sistema conversacional multimodal pioneiro criado pela Google, marca uma mudança significativa na tecnologia de IA ao superar os modelos tradicionais de *Large Language Model* (LLM) baseados em texto, como o GPT-3, e até mesmo seu equivalente multimodal, o ChatGPT-4. A arquitetura do Gemini foi projetada para incorporar o processamento de diversos tipos de dados, como texto, imagens, áudio e vídeo, um feito facilitado por seu codificador multimodal exclusivo, rede de atenção cruzada entre modalidades e decodificador multimodal. O núcleo arquitetônico do Gemini é sua estrutura de codificadores duplos, com codificadores separados para dados visuais e textuais, permitindo uma contextualização multimodal sofisticada. Gemini se destaca pela amplitude de modalidades, desempenho em benchmarks de multimodalidade, escalabilidade com versões Ultra, Pro e Nano, habilidades avançadas em geração de código e foco na explicabilidade. No entanto, seu desempenho em tarefas complexas que exigem integração de conhecimento de senso comum entre modalidades ainda precisa ser totalmente avaliado (MCINTOSH et al., 2023).

GitHub e OpenAI lançaram recentemente o Copilot, que utiliza o poder do processamento de linguagem natural, análise estática, síntese de código e inteligência artificial. Dada uma descrição em linguagem natural da funcionalidade desejada, o Copilot pode gerar o código correspondente em várias linguagens de programação. As sugestões do Copilot em Java têm a maior taxa de correção (57%), enquanto JavaScript tem a menor (27%). No geral, as sugestões do Copilot têm baixa complexidade, sem diferenças notáveis entre as linguagens de programação. Também identificamos algumas possíveis limitações

do Copilot, como a geração de código que pode ser simplificado e código que depende de métodos auxiliares indefinidos (WERMELINGER, 2023).

A empresa Tabnine criou uma assistente de codificação chamada "Tabnine". Essa ferramenta usa algoritmos de aprendizado de máquina para fornecer conclusões, previsões e sugestões de código com reconhecimento de contexto em várias linguagens de programação e ambientes de codificação. Esses recursos ajudam os desenvolvedores a melhorar a eficiência da codificação, reduzir erros de tempo de execução e também ajudar a reduzir o pressionamento de teclas (LLERENA-IZQUIERDO et al., 2024).

2.1.4 Modelos de Redes Neurais Profundas para IA Generativa

Foi a introdução do VAE e das GANs em 2014 que realmente impulsionou a era moderna da IA generativa. Até então, as redes neurais profundas eram usadas principalmente para tarefas de classificação, mas com essas arquiteturas, elas começaram a ser utilizadas para inteligência artificial generativa. Os VAEs são utilizados para detecção de anomalias, geração e aumento de dados entre outras aplicações, e são capazes de gerar dados de texto e áudio, além de imagens.

As redes generativas adversariais representam a técnica principal de IA Generativa utilizada atualmente. Uma GAN consiste em um par de redes neurais: uma, chamada de geradora, sintetiza o conteúdo (como por exemplo, uma imagem de um rosto humano), enquanto a outra, conhecida como discriminadora, avalia a veracidade desse conteúdo gerado, isto é, determina se o rosto é natural ou falso. Essas redes repetem esse ciclo de gerar e discriminar até que a geradora produza conteúdo que a discriminadora não consiga distinguir entre real e sintético (JOVANOVIC; CAMPBELL, 2022).

A grande inovação que levou ao desempenho dos atuais modelos de IA generativa de ponta foi a introdução do Transformer. Em uma arquitetura de Transformer geral, pode-se usar um número igual de codificadores e decodificadores. Esse modelo revolucionou a IA generativa, formando a base dos atuais LLMs disponíveis hoje, como a família GPT, PaLM e LLaMA, entre outros, juntamente com a família de modelos geradores de imagens DALL-E (BENGESI et al., 2024).

Modelos autorregressivos são modelos estatísticos ou arquiteturas de redes neurais que geram dados sequenciais um elemento de cada vez, condicionados aos elementos anteriores. Eles modelam a probabilidade condicional de cada elemento na sequência com base nos elementos que o precedem. Geralmente usados em tarefas como previsão de séries temporais, modelagem de linguagem e geração de música, os modelos autorregressivos são treinados usando máxima verossimilhança (RANI; SINGH; KHANNA, 2023).

A rede neural convolucional, é uma arquitetura essencial para análise de dados visuais, amplamente utilizada em detecção de objetos e reconhecimento facial. Na classificação

de imagens, as CNNs processam imagens de entrada para identificar e categorizar diferentes elementos visuais com precisão. Sua estrutura inclui camadas convolucionais para extração de características, camadas de *pooling* para redução de dimensões e camadas totalmente conectadas para combinar informações na saída final. Diversas funções de ativação nas camadas melhoram a compreensão de relações complexas nos dados (HOSSAIN; ZAMAN; ISLAM, 2023).

O modelo de difusão é um modelo generativo probabilístico caracterizado por um processo de duas etapas. Primeiramente, o processo de difusão direta introduz ruído gaussiano nos dados de treinamento. Em seguida, o processo de difusão reversa, conhecido como redução de ruído, ou, do inglês, *denoising*, reverte gradualmente o passo de difusão passo a passo para gerar novos dados amostrais. Esses modelos superam efetivamente os desafios encontrados no alinhamento das distribuições posteriores dentro dos VAEs, mitigam a instabilidade inerente nos objetivos adversariais das GANs, oferecendo um objetivo de treinamento mais estável, e abordam os encargos computacionais associados aos métodos de Cadeia de Markov. O modelo de difusão basicamente engloba três formulações principais: *Denoising Diffusion Probabilistic Models*, *Stochastic Differential Equations* e *Score-based Generative Models* (BENGESI et al., 2024).

2.1.5 Ética na IA Generativa

Os modelos de IA Generativa podem gerar componentes em grande escala para diversos contextos, desde a educação até a tomada de decisões médicas. No entanto, antes de implementá-los em produção, os criadores dos modelos devem definir claramente seus objetivos; identificar os beneficiários; e validar os cenários de uso com os usuários-alvo para evitar comportamentos não éticos não intencionais. Isso exige a identificação e a participação ativa de todos os stakeholders envolvidos, cientistas, engenheiros de inteligência artificial, especialistas na área, autoridades regulatórias e usuários-alvo (JOVANOVIC; CAMPBELL, 2022).

A coleta e o uso de dados de back-end como entrada para inteligência artificial generativa devem estar em conformidade com a legislação vigente. Exemplos disso incluem o Regulamento Geral de Proteção de Dados na Europa, a Lei de Privacidade do Consumidor da Califórnia nos Estados Unidos e a Lei de Proteção de Dados de 2018 no Reino Unido. A saída de ferramentas de IA generativa como Bard ou ChatGPT pode conter dados pessoais, infringindo a privacidade e podendo configurar violação do consentimento informado. Essas violações de privacidade podem incluir a exposição do status financeiro de clientes, a divulgação de detalhes de uma marca ou empresa ainda não lançada, a coleta e o compartilhamento não autorizados de informações privadas e o rastreamento de movimentos online através do aplicativo sem o consentimento do usuário (OOI et al., 2025).

Os LLMs normalmente dependem fortemente dos dados de treinamento e, quando esses dados contêm vieses ou anomalias, isso pode resultar em resultados injustos. Por exemplo, se os dados de treinamento forem tendenciosos contra certas pessoas ou culturas, o modelo pode produzir saídas injustas ou discriminatórias. Portanto, é crucial garantir que os dados de treinamento sejam diversos e bem equilibrados. Modelos de linguagem como o ChatGPT podem ser usados para gerar notícias falsas, discurso de ódio e outros conteúdos prejudiciais, o que pode levar a distúrbios sociais, danos à reputação e até mesmo danos físicos. Além disso, os mecanismos e processos internos desses modelos não são suficientemente abertos e transparentes para os usuários. É importante garantir que os processos de tomada de decisão desses modelos sejam claros e compreensíveis. Como o ChatGPT gera respostas sem intervenção humana direta, pode ser difícil responsabilizar alguém pelas respostas geradas, complicando o tratamento de questões éticas ou vieses. Esses modelos também envolvem a coleta e o processamento de dados pessoais, levantando preocupações sobre privacidade e segurança. Medidas adequadas devem ser implementadas para proteger os dados pessoais contra acesso não autorizado (RAHMAN; WATANOBE, 2023).

Dado o potencial de uso indevido e os riscos associados, torna-se evidente a necessidade de implementar regulamentações que abordem os parâmetros éticos dos modelos de IA generativa. Isso inclui estabelecer mecanismos robustos de governança, assegurar a transparência nos processos de tomada de decisão e garantir que os dados de treinamento sejam diversificados e livres de preconceitos. Além disso, a colaboração entre desenvolvedores de inteligência artificial, autoridades regulatórias e a sociedade em geral é fundamental para mitigar riscos e garantir que a IA generativa seja utilizada de maneira responsável e benéfica para todos.

A Tabela 1 resume as principais vantagens e desvantagens de cada modelo de inteligência artificial generativa.

Tabela 1 – Vantagens e desvantagens dos modelos baseados em Transformers.

Modelo	**Foco Principal**	**Vantagens**	**Desvantagens**
GANs	Geração de dados e realismo	Produz saídas de alta qualidade e diversidade	Estabilização desafiadora e colapso de modo
VAEs	Geração de dados e síntese controlável	Capacidade de capturar incerteza de dados e fornecer representação de dados eficiente	Reconstruções borradas e desafios na avaliação da qualidade das amostras geradas
Autoregressive Models	Geração de texto e dados sequenciais	Fornece a probabilidade exata dos dados	Um processo de geração mais lento, com natureza sequencial, pode limitar o paralelismo
Transformer	Captura global de contexto e relações	Alta escalabilidade, capacidade de modelagem e processamento paralelo de dados	Alto Custo Computacional e Complexidade de Implementação
CNN	Extração eficiente de padrões visuais locais	Detectam características automaticamente e exigem pouco pré-processamentom	Alto custo computacional e baixa performance em tempo real
Diffusion Models	Geração progressiva de dados realistas	Geração de Imagens de Alta Qualidade	Lentidão no Processo de Geração, Complexidade de Implementação

2.2 Arquitetura Transformers

Foi introduzido pela primeira vez no trabalho de Vaswani et al. (2017) em 2017 e foi rapidamente popularizado como a arquitetura líder para a maioria dos aplicativos de dados de texto.

Os Transformers foram inicialmente introduzidos no contexto do processamento de linguagem natural, onde uma "linguagem natural" é uma língua como o inglês ou o mandarim. Eles superaram amplamente as abordagens anteriores baseadas em redes neurais recorrentes. Posteriormente, descobriu-se que essas arquiteturas também alcançam excelentes resultados em muitos outros domínios. Por exemplo, os modelos voltados para visão frequentemente superam as redes convolucionais em tarefas de processamento de imagens, enquanto os modelos multimodais, que combinam múltiplos tipos de dados, como texto, imagens, áudio e vídeo, estão entre os modelos de aprendizado profundo mais poderosos (BISHOP; BISHOP, 2023).

Considerado um modelo de aprendizado profundo de grande renome, o Transformer tem sido vastamente adotado em diversas áreas, incluindo PLN, *Computer Vision* e proces-

samento de fala. Esse modelo foi originalmente apresentado como um modelo de sequência para sequência para tradução automática. Trabalhos posteriores mostram que modelos pré-treinados baseados em Transformer podem atingir desempenhos de última geração em várias tarefas. Esse modelo se tornou a arquitetura favorita em PLN, especialmente para modelos pré-treinados, como consequência. Além de aplicações associadas à linguagem, o Transformer também foi adotado em processamento de áudio e também em outras disciplinas, como ciências biológicas e química (LIN et al., 2021).

Os Transformers são um dos avanços mais importantes no aprendizado profundo. Eles são baseados em um conceito de processamento chamado atenção, que possibilita que uma rede atribua pesos diferentes a diferentes entradas, com coeficientes de ponderação que, por sua vez, dependem dos valores de entrada, capturando assim vieses indutivos poderosos associados a dados sequenciais e outras formas de dados.

A denominação *transformers* é atribuída a esses modelos por sua capacidade de transformar um um conjunto de vetores em um determinado espaço de representação em um conjunto correspondente de vetores, preservando a mesma dimensionalidade, mas em um novo espaço. Essa transformação tem o objetivo de que o novo espaço tenha uma representação interna mais rica, melhor adaptada para a solução de tarefas subsequentes. Os Transformers possuem uma ampla aplicabilidade porque suas entradas podem assumir a forma de conjuntos de vetores não estruturados, sequências ordenadas ou representações mais gerais (BISHOP; BISHOP, 2023).

2.2.1 Funcionamento e Arquitetura

O mecanismo de atenção é um conceito impressionante em redes neurais, especialmente em tarefas como processamento de linguagem natural. É como dar um holofote ao modelo, permitindo que ele foque em algumas partes da sequência de entrada enquanto ignora outras, de maneira semelhante a como nós, humanos, prestamos atenção a palavras ou frases específicas ao entender uma sentença.

No contexto dos Transformers, os mecanismos de atenção servem para ponderar a importância de diferentes *tokens* de entrada ao produzir uma saída. Isso não é apenas uma reprodução da atenção humana, mas uma melhoria, permitindo que as máquinas superem o desempenho humano em certas tarefas. Considere os seguintes pontos que destacam a importância dos mecanismos de atenção (VASWANI et al., 2017):

(a) eles fornecem um meio de focar nas partes mais pertinentes dos dados;

(b) conseguem capturar dependências de longo alcance que modelos anteriores, como RNNs, tinham dificuldades para lidar;

(c) facilitam paralelização, levando a melhorias consideráveis na eficiência computacional.

A graciosidade dos mecanismos de atenção reside em sua simplicidade e capacidade. Ao permitir que os modelos considerem todo o contexto de uma entrada, eles abriram novas possibilidades em aprendizagem de máquina, levando a avanços na área de processamento de linguagem natural.

O mecanismo de atenção opera sobre três entradas:

- *Query* (Q): essa matriz representa a palavra na qual é o foco da aplicação;
- *Key* (K): essa matriz representa as palavras no documento com as quais se compara a consulta;
- *Value* (V): essa matriz contém as representações de cada palavra no documento de entrada.

Esses são os componentes fundamentais do mecanismo de *self-attention*. Eles permitem que o modelo analise uma palavra (*query*), a compare com todas as outras palavras (*keys)* e, por fim, decida quanta atenção dar a cada palavra (*values*) (BISHOP; BISHOP, 2023).

O mecanismo de *self-attention* é uma inovação fundamental dentro do modelo Transformer, permitindo que ele identifique relações complexas nos dados. Ele possibilita que o modelo pese a importância de diferentes partes da entrada de forma independente da posição na sequência. Isso é especialmente útil em cenários onde o contexto é essencial para a compreensão do significado, como distinguir se *"it"* se refere ao "lobo" ou ao "coelho" em uma determinada sentença. A elegância do *self-attention* está em sua capacidade de modelar relacionamentos independentemente da distância entre os elementos na sequência. Essa característica representa uma mudança significativa em relação às abordagens anteriores de modelagem de sequência, que frequentemente enfrentavam dificuldades para lidar com dependências de longo alcance.

Os seguintes passos resumem o funcionamento do mecanismo de *self-attention*:

1. calcula os vetores de consulta (*query*), chave (*key*) e valor (*value*) para cada elemento de entrada;
2. realiza o produto escalar entre as consultas e as chaves para obter as pontuações;
3. aplica a função SoftMax para normalizar as pontuações;
4. multiplica as pontuações normalizadas pelos vetores de valor para obter os vetores de atenção.

Ao separar e processar esses vetores, o mecanismo de *self-attention* permite uma compreensão mais profunda da entrada, sendo fundamental para as tarefas complexas com as quais os Transformers são frequentemente encarregados (VASWANI et al., 2017).

Em um nível muito alto, um modelo codificador-decodificador pode ser pensado como dois blocos, o codificador e o decodificador, conectados por um vetor que chamaremos de "vetor de contexto". O codificador processa cada *token* na sequência de entrada. Ele tenta comprimir todas as informações sobre a sequência de entrada em um vetor de comprimento fixo, ou seja, o "vetor de contexto". Depois de passar por todos os *tokens*, o codificador passa esse vetor para o decodificador. Por sua vez, o vetor é construído de tal forma que se espera que ele encapsule todo o significado da sequência de entrada e ajude o decodificador a fazer previsões precisas. Mais tarde será visto que esses são os estados internos finais do nosso bloco codificador. Por fim, o decodificador lê o vetor de contexto e tenta prever a sequência de destino *token* por *token*. De forma mais clara, o codificador e o decodificador são os componentes centrais de um Transformer. O codificador funciona como uma equipe que processa e organiza as informações de entrada, criando uma representação que destaca as relações importantes entre elas. Já o decodificador usa essa representação para gerar a saída passo a passo, como uma equipe que elabora o plano final de um evento com base nos dados e nas decisões iniciais (BISHOP; BISHOP, 2023). A arquitetura é ilustrada na Figura 2.

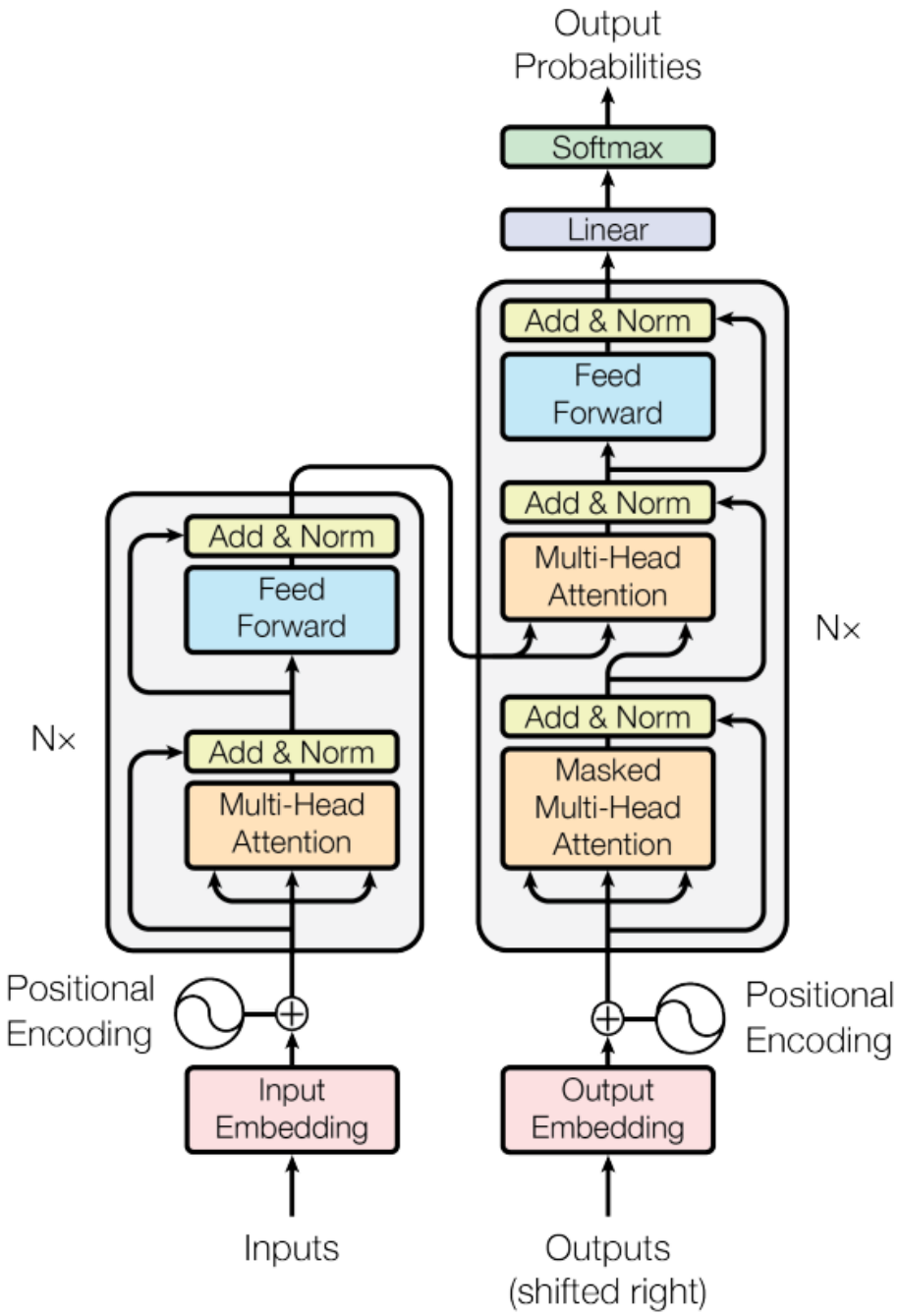


Figura 2 – Arquitetura Transformer.

Fonte: (VASWANI et al., 2017).

2.2.2 Principais Modelos Baseados em Transformers

Bidirectional Encoder Representations from Transformers (BERT) é um modelo revolucionário de processamento de linguagem natural introduzido pelo Google AI em 2018. Ele é baseado na arquitetura Transformer e foi projetado para pré-treinar representações bidirecionais profundas por condicionamento conjunto em contexto esquerdo e direito em todas as camadas. Ao contrário dos modelos tradicionais unidirecionais ou superficiais, o BERT lê texto em ambas as direções, tornando-o altamente eficaz para tarefas que

exigem compreensão contextual profunda (DEVLIN et al., 2019). A vantagem desse modelo está em sua facilidade de uso, exigindo apenas a adição de uma única camada de saída à arquitetura neural existente para obter modelos de texto que superam a imprecisão de todos os existentes em diversos problemas de processamento de linguagem natural (KOROTEEV, 2021).

O GPT se tornou famoso após o lançamento do ChatGPT pela OpenAI, uma empresa de pesquisa focada no desenvolvimento de tecnologias de IA. O GPT é um modelo de aprendizado profundo pré-treinado em grandes corpora de dados textuais e pode ser ajustado para tarefas específicas, como geração de linguagem, análise de sentimento, modelagem de linguagem, tradução automática e classificação de texto. A arquitetura Transformer usada no GPT representa um avanço significativo em relação a abordagens anteriores de PLN, como RNN e CNN. Ela utiliza um mecanismo de autoatenção, permitindo que o modelo considere o contexto de toda a sentença ao gerar a próxima palavra, o que aprimora sua capacidade de compreender e gerar linguagem (YENDURI et al., 2024).

Os sucessores do GPT-1 são o GPT-2 e o GPT-3. O GPT-2 e o GPT-3 são extensões do GPT-1, com um número maior de camadas do decodificador Transformer. Consequentemente, eles possuem mais parâmetros treináveis e podem ser treinados com um volume maior de dados para melhorar a modelagem de linguagem e a inferência. Por exemplo, o GPT-2 possui 1,5 bilhão de parâmetros. O GPT-2 e, especialmente, o GPT-3 foram treinados com um volume muito maior de dados da Internet, abrangendo diversos temas gerais e acadêmicos, para gerar textos em qualquer assunto e estilo de interesse. Como exemplo, o GPT-2 foi treinado em 8 milhões de páginas da web, contendo um total de 40 GB de dados textuais da internet (GHOJOGH; GHODSI, 2020).

2.3 Considerações Finais

O presente capítulo apresentou os fundamentos teóricos indispensáveis para a compreensão dos conceitos e tecnologias que sustentam este trabalho. Inicialmente, foram abordados os principais aspectos da Inteligência Artificial Generativa, contemplando sua definição, evolução histórica, aplicações mais relevantes, empresas protagonistas, tecnologias associadas, implicações éticas e as vantagens e limitações de seus modelos. Essa discussão foi essencial para contextualizar o uso de técnicas generativas no enfrentamento do problema proposto.

Na sequência, foi detalhada a arquitetura Transformer, destacando seu funcionamento interno, seus principais componentes e, sobretudo, o mecanismo de atenção, elemento central para a eficiência e flexibilidade desses modelos. Também foram apresentados modelos de referência que se baseiam nessa arquitetura e que hoje constituem alicerces

fundamentais para aplicações de ponta em diferentes áreas.

A partir dessa análise, consolidou-se o entendimento de que, embora diversas arquiteturas tenham sido aplicadas ao problema de Visual Paragraph Generation, como CNNs e RNNs, o Transformer se sobressai como a abordagem mais difundida e robusta. Sua capacidade de lidar com dependências de longo alcance, associada à escalabilidade e ao processamento paralelo, justificam sua ampla adoção em modelos de destaque como BERT, GPT e ViT. Assim, confirma-se o foco dado a essa arquitetura no presente trabalho, visto que ela representa o estado da arte e concentra as perspectivas mais promissoras para o avanço da área.

3 Trabalhos Relacionados

Este capítulo apresenta os principais trabalhos relacionados ao problema de Visual Paragraph Generation, com ênfase em abordagens que utilizam modelos de Inteligência Artificial Generativa e, em especial, arquiteturas baseadas em Transformers. São discutidas as propostas identificadas na literatura, ressaltando seus objetivos, metodologias e resultados, de modo a oferecer uma visão geral do estado da arte na área.

Primeiramente, as Seções 3.1 a 3.5 apresentam os trabalhos que exploram modelos baseados em Transformers aplicados à geração de descrições textuais, destacando suas metodologias e resultados mais relevantes. Em seguida, a Seção 3.6 realiza uma análise comparativa das propostas estudadas mostrando os pontos positivos e negativos, além de apresentar uma síntese crítica que conecta as discussões realizadas neste capítulo às etapas práticas da metodologia adotada neste trabalho.

3.1 Geração de Legendas para o Idioma Tailandês

O trabalho de Jaknamon e Marukatat (2022), propõe uma solução para a geração automática de legendas de imagens na língua tailandesa. O objetivo geral é desenvolver um sistema que utilize a arquitetura Transformer, tanto para a extração de características visuais quanto para a geração de texto, apresentando melhores resultados que abordagens anteriores que se baseavam em CNNs e RNNs para essa tarefa. O estudo visa especificamente aprimorar um campo ainda pouco explorado, a descrição de imagens de alimentos e viagens em tailandês. Foram utilizados três conjuntos de dados distintos para os experimentos: o dataset "Travel" com 61.000 imagens de locais de viagem na Tailândia e dois tipos de legendas (apenas o nome do local e uma descrição completa), o *dataset* "Food" com 30.000 imagens de alimentos e dois tipos de legendas (apenas o nome do alimento e uma descrição completa), e o Flickr30k, que consiste em cerca de 30.000 imagens com cinco legendas originais em inglês, traduzidas para o tailandês utilizando o Google Translate.

A metodologia usa uma arquitetura de *encoder-decoder* baseada em Transformer. O *encoder* utiliza modelos Vision Transformer pré-treinados, como SwinT, ViT e DeiT, para extrair características da imagem, tratando-a como uma sequência de *"patches"*. O *decoder*, por sua vez, aplica modelos de texto Transformer pré-treinados, como RoBERTa, BERT, WangchanBERTa, XLM-ROBERTa, GPT2 e BART, para gerar as legendas em tailandês. Para o treinamento em dataset com menor relação entre imagem e legenda completa, foi adotada uma estratégia de duas etapas: a primeira etapa treina o *encoder* em dataset com apenas o nome, e a segunda etapa utiliza o *encoder*, mantendo os pesos aprendidos, para treinar com legendas completas. A métrica de avaliação utilizada para

medir a correção das legendas geradas foi o BLEU, com valores variando de 0 (mínimo) a 1 (máximo), onde valores mais altos indicam melhor desempenho.

Os resultados demonstram que o ThaiTC, especialmente com a combinação SwinT-Large como *encoder* e RoBERTa como *decoder*, superou significativamente a abordagem tradicional de CNN+Attention nos dataset Food e Flickr30k. No *dataset* Flickr30k, o SwinT+RoBERTa obteve um score BLEU de 5.70, em contraste com 2.96 do CNN+Attention. No *dataset* Food, o SwinT+RoBERTa alcançou 4.83, enquanto o CNN+Attention registrou 4.46. No entanto, no *dataset Travel*, o desempenho do SwinT+RoBERTa (BLEU de 2.91) foi ligeiramente inferior ao do CNN+Attention (BLEU de 3.07). Os autores também observaram que a estratégia de treinamento em duas etapas proporcionou resultados superiores em comparação com o treinamento de legenda completa em uma única etapa.

Entre os principais pontos fortes do trabalho, destaca-se a proposta inovadora de um sistema *end-to-end* de geração de legendas de imagens em tailandês baseado puramente em arquiteturas Transformer, eliminando a dependência de modelos CNN convencionais para compreensão de imagem. A utilização de modelos Vision Transformer e Text Transformer pré-treinados permite o uso de embeddings de palavras em tailandês e treinamento em dataset tailandeses específicos. A metodologia de treinamento em duas etapas é uma vantagem, especialmente para dataset onde a relação entre imagem e legenda não é tão estreita, melhorando a adaptação do modelo aos dados. Como ponto fraco e limitação, o modelo SwinT+RoBERTa, embora com melhor desempenho em grande parte dos datasets, exige significativamente mais RAM durante a fase de treinamento em comparação com a abordagem CNN+Attention, devido ao seu tamanho consideravelmente maior. Além disso, o desempenho no *dataset* Travel não foi tão robusto quanto nos outros datasets, indicando que as legendas de viagem e comida podem ter menos características nas imagens.

3.2 Geração de Legendas para Imagens de Sensoriamento Remoto

Por sua vez, Wang et al. (2022) abordam a necessidade de descrever com precisão imagens de sensoriamento remoto de alta resolução espacial, que exigem a compreensão de atributos internos de objetos e relações externas entre eles. O objetivo geral é propor uma arquitetura "*pure* Transformer" (CapFormer) que capture a representação global da imagem, superando a limitação de algoritmos existentes que focam apenas em características locais e carecem da capacidade de sumarizar cenas complexas. O *dataset* utilizado para os experimentos é o *Remote Sensing Image Captioning Dataset* (RSICD), que contém 10.921 imagens alta resolução espacial de várias resoluções, coletadas do Google Earth e Baidu Map, entre outros. Todas as imagens foram redimensionadas para 224×224 pixels, e o *dataset* abrange 30 cenas diferentes, com cada imagem anotada com cinco sentenças

características. A divisão dos dados seguiu a configuração padrão: 80% para treinamento, 10% para validação e 10% para teste.

A metodologia adotada pelo CapFormer é estruturada em uma arquitetura *encoder-decoder* baseada integralmente em Transformers. O *encoder* utiliza um Vision Transformer escalável, capaz de capturar características globais da imagem por meio de camadas de *multi-head self-attention*. Nesse processo, a imagem bidimensional é convertida em uma sequência unidimensional de segmentos achatados, aos quais são adicionados embeddings posicionais para preservar a informação espacial. O *decoder*, por sua vez, é uma versão modificada do Transformer, projetada para converter progressivamente as representações visuais em descrições textuais coerentes. Ele considera explicitamente as palavras geradas anteriormente e interage com as representações da imagem por meio de camadas de *cross-attention*.

Para avaliar o desempenho do CapFormer em comparação a outros métodos, foram utilizadas métricas consolidadas na área, como BLEU (nas variantes BLEU-1 a BLEU-4), ROUGE-L e CIDEr. Os resultados obtidos demonstraram que o CapFormer superou os métodos de descrição de imagens de sensoriamento remoto de última geração. Especificamente, a configuração do CapFormer com Swin-Tiny apresentou as maiores precisões em seis métricas, incluindo BLEU-4 (32,62), ROUGE-L (49,78) e CIDEr (91,17). Isso evidencia que a arquitetura Transformer promove melhorias significativas tanto no *encoder* quanto no *decoder* no contexto de descrição de imagens de sensoriamento remoto.

Entre os principais diferenciais do CapFormer, destaca-se a habilidade de capturar informações globais da imagem por meio do Vision Transformer, um aspecto essencial para a síntese de cenas complexas em imagens de alta resolução espacial. Esse recurso permite superar limitações comuns em abordagens baseadas em CNNs, que tendem a priorizar apenas características locais. Outra vantagem relevante está na forma como o *decoder* modela explicitamente as relações entre palavras previamente geradas, além de interagir de maneira eficiente com as representações visuais por meio de camadas de *cross-attention*. Esses mecanismos contribuem para mitigar problemas recorrentes em modelos baseados em LSTMs, como o esquecimento de informações contextuais e a difusão de gradientes.

Além disso, o CapFormer apresenta boa generalização quando utilizado com diferentes variantes de *encoders* baseados em Transformer, ampliando sua aplicabilidade. O estudo, entretanto, não aborda limitações específicas do modelo, concentrando-se principalmente nas vantagens proporcionadas pela arquitetura adotada.

3.3 Geração de Legendas para Imagens com Suporte Multilíngue

A proposta de Lam et al. (2023) aborda o desafio de criar sentenças descritivas automaticamente para imagens, um processo que exige a combinação de compreensão

visual e processamento de linguagem natural. O objetivo geral do estudo é investigar o desempenho de modelos baseados em Transformer para legendagem de imagens, propondo um modelo específico que emprega um Vision Transformer com mecanismo de re-atenção como *encoder* e um modelo baseado em T5 como *decoder*. O foco é alcançar legendas de alta qualidade, especialmente para a língua vietnamita.

Os experimentos foram realizados utilizando o *dataset* Flickr8k, tanto em sua versão original em inglês quanto em uma versão traduzida para o vietnamita. Esse conjunto de dados contém 8.000 imagens, com a divisão seguindo a proporção padrão: 80% para treinamento, 10% para validação e 10% para teste.

A abordagem proposta, denominada DeepViT+ViT5, adota uma arquitetura do tipo *encoder-decoder*. O *encoder* é composto por um Vision Transformer, o DeepViT, que incorpora um mecanismo de re-atenção, introduzido para contornar uma limitação do *self-attention* tradicional, especialmente em camadas mais profundas do ViT, onde os mapas de características tendem a se tornar excessivamente semelhantes. O *decoder* é baseado no ViT5, um modelo de linguagem pré-treinado fundamentado na arquitetura Transformer e derivado do T5. O processo inicia-se com o redimensionamento das imagens, seguido da segmentação em *patches* não sobrepostos, que são então convertidos em vetores e enviados ao *encoder*.

A geração das legendas ocorre de forma progressiva, camada por camada, com o *decoder* utilizando tanto as saídas anteriores quanto as representações extraídas pelo *encoder* para produzir cada palavra da sentença. A métrica de avaliação principal utilizada é o BLEU-4.

Os resultados alcançados pelo modelo DeepViT+ViT5 são notáveis, com um score BLEU-4 de 37.98 no *dataset* Flickr8k vietnamita, sendo este o melhor resultado obtido até o momento para legendagem de imagens em vietnamita. Um dos principais pontos fortes do trabalho está na proposta inovadora de utilizar um Vision Transformer com mecanismo de re-atenção no *encoder*, solução que enfrenta diretamente o desafio da baixa diversidade dos mapas de características em camadas profundas, um problema recorrente em ViTs convencionais. Outro ponto de destaque é a adoção de um *decoder* baseado no modelo T5, o que permite explorar o potencial de modelos de linguagem pré-treinados na geração de legendas mais precisas e naturais. A aplicação do modelo à tarefa de legendagem de imagens em vietnamita também representa uma contribuição relevante, especialmente por se tratar de um idioma com menor presença na literatura da área em comparação ao inglês. Como principal limitação, o artigo aponta as restrições de hardware disponíveis para os experimentos, especificamente, uma GPU compatível com CUDA e 12 GB de RAM. Essa limitação impediu a utilização de conjuntos de dados mais robustos, como o MS COCO, amplamente utilizado em outros estudos, restringindo as avaliações ao *dataset* Flickr8k e, consequentemente, a uma escala experimental mais modesta.

3.4 Geração de Legendas com Mecanismo de Atenção Adaptativa para a Arquietura Transformers

O quarto trabalho de Zhang et al. (2019) foca em aprimorar a geração de legendas de imagens dentro do framework *encoder-decoder*, que tem mostrado progresso significativo. O objetivo geral é melhorar a performance do modelo de legendagem ao empregar um *decoder* Transformer mais poderoso e ao combinar mecanismos de atenção espacial e atenção adaptativa, permitindo que o *decoder* determine onde e quando usar informações de regiões da imagem para gerar legendas mais precisas e acelerar o processo de treinamento. O *dataset* utilizado para os experimentos foi o Flickr30k. O conjunto de dados foi dividido em 29.000 amostras para treinamento e 1.000 para validação e teste, respectivamente. A metodologia proposta, chamada Adaptive-Trans, baseia-se em um framework *encoder-decoder*. O *encoder* utiliza uma rede neural convolucional, especificamente a ResNet-101, para extrair características visuais da imagem. O *decoder*, que tradicionalmente usava LSTM, foi substituído por uma arquitetura Transformer. A inovação principal reside na incorporação de atenção espacial e atenção adaptativa ao Transformer. A atenção espacial permite que o modelo saiba qual região da imagem é mais importante ao gerar uma palavra, enquanto a atenção adaptativa permite que o modelo decida quando usar a característica da imagem, ou seja, se deve depender da informação visual ou da própria sequência de palavras já geradas. As métricas de avaliação utilizadas para medir o desempenho incluem BLEU (BLEU-1, BLEU-2, BLEU-3, BLEU-4), METEOR e CIDEr. Os resultados alcançados no *dataset* Flickr30k demonstraram que o modelo Adaptive-Trans superou a maioria dos métodos comparados. Comparado ao modelo adaptativo original baseado em LSTM (Adaptive-LSTM), o Adaptive-Trans obteve um desempenho ligeiramente superior na maioria das métricas (B-1: 0.670 vs 0.667; B-2: 0.496 vs 0.494; B-3: 0.355 vs 0.354; B-4: 0.252 vs 0.251; METEOR: 0.204 vs 0.204), com exceção do CIDEr, onde obteve 0.530 contra 0.531 do Adaptive-LSTM. Isso é atribuído à melhoria do *decoder* com o Transformer, que é considerado mais poderoso que o LSTM. Além disso, o uso do Transformer no *decoder* permite um treinamento mais eficiente devido à sua capacidade de paralelização.

Os principais pontos fortes deste trabalho incluem a substituição do *decoder* LSTM por um Transformer, o que acelera o processo de treinamento e melhora a eficiência do modelo devido à capacidade de paralelização. A combinação da atenção espacial e adaptativa no Transformer é uma grande vantagem, permitindo que o modelo determine de forma inteligente quando e onde focar na informação visual, resultando em legendas mais precisas. O trabalho não detalha pontos fracos ou desvantagens significativas da abordagem, focando nas melhorias de desempenho e eficiência.

3.5 Geração de Legendas em Grandes Bases de Imagens

Os autores Eluri et al. (2024) tratam do avanço significativo na geração de legendas de imagens através da integração de modelos de *Deep Learning*. O objetivo geral é desenvolver um método que utilize técnicas de DL, como InceptionResNetV2 para extração de características e arquiteturas baseadas em Transformer para processamento de linguagem natural, a fim de gerar legendas descritivas para imagens, superando as limitações das abordagens tradicionais de RNN em termos de gradientes evanescentes e paralelização. O foco é capturar relações semânticas intrincadas entre o conteúdo visual e as descrições textuais, resultando em legendas mais precisas e contextualmente relevantes. O *dataset* utilizado para os experimentos é o *COCO Captions Dataset*, uma versão do COCO 2014 que inclui imagens, *bounding boxes*, rótulos e legendas, e que aborda preocupações de qualidade de dados, como imagens sem legendas correspondentes.

A metodologia proposta integra InceptionResNetV2 para extração de características da imagem e um *Detection Transformer* para a parte de compreensão visual e detecção de objetos. Essa combinação pretende alavancar os pontos fortes de ambas as arquiteturas, alcançando desempenho de ponta em tarefas de detecção de objetos. O modelo é treinado de forma conjunta para detectar imagens e gerar legendas para imagens com até 50% de oclusão. As métricas de avaliação mencionadas para a tarefa de detecção em imagens ocluídas são precisão (0.982), recall (0.931), F1 Score (0.942) e sensibilidade (0.892). O trabalho destaca a sinergia entre os modelos de DL e os modelos de linguagem natural para capturar relações semânticas complexas. Os resultados alcançados demonstram que, através do treinamento conjunto, o modelo é capaz de detectar imagens e gerar legendas em imagens com até 50% de oclusão, atingindo uma precisão de 0.982, recall de 0.931, F1 score de 0.942 e sensibilidade de 0.892.

Entre os principais pontos fortes da abordagem, destaca-se a superação das limitações associadas às RNNs, como o problema dos gradientes evanescentes e a dificuldade de paralelização, o que resulta em maior eficiência e escalabilidade do modelo. A combinação do InceptionResNetV2 com o Detection Transformer também se mostra vantajosa, ao unir uma extração robusta de características visuais com capacidades avançadas de detecção de objetos. Essa integração permite ao sistema capturar relações semânticas complexas entre a imagem e o texto, contribuindo para a geração de legendas mais precisas e contextualmente relevantes. O trabalho não apresenta de forma explícita limitações da própria abordagem, mas menciona dificuldades enfrentadas por trabalhos anteriores, como o risco de *overfitting* em função da memorização de imagens semelhantes no conjunto de dados, além da necessidade de aprimorar o alinhamento entre as legendas geradas e as referências humanas, especialmente no que se refere à métrica CIDEr.

3.6 Considerações Finais

A Tabela 2 apresenta um resumo dos trabalhos relacionados. Cada coluna da tabela foi organizada de forma a destacar informações específicas:

- **Autor:** identifica os responsáveis pelo estudo e serve de referência para situar os trabalhos na literatura;
- **Objetivo:** descreve a principal motivação de cada pesquisa, isto é, o problema que se buscou resolver;
- **Dataset:** indica os conjuntos de dados utilizados, o que é fundamental para avaliar a complexidade do treinamento, a diversidade das amostras e a generalização dos resultados;
- **Modelo:** apresenta a arquitetura adotada (*encoders* e *decoders*), revelando as tendências metodológicas da área.
- **Pontos positivos:** sintetiza os principais avanços, contribuições e inovações de cada proposta;
- **Pontos negativos:** destaca as limitações encontradas, sejam técnicas (como uso de hardware de alta capacidade) ou metodológicas (como dependência de CNNs ou dataset reduzidos).

A análise comparativa desses trabalhos permite identificar padrões importantes. A principal observação é a clara transição e preferência pelo uso de Transformers em detrimento das RNNs no papel de *decoders* e, cada vez mais, também como *encoders* visuais. Essa mudança é justificada pela capacidade dos Transformers de lidar com dependências de longo alcance de forma mais eficiente, bem como pela sua capacidade de paralelização, o que acelera o treinamento e melhora a escalabilidade.

Além disso, nota-se que cada trabalho buscou resolver desafios distintos:

- **ThaiTC:** apresenta uma solução voltada para o idioma tailandês, com um treinamento em duas fases que busca adaptar melhor o encoder e o decoder;
- **CapFormer:** foca em imagens de sensoriamento remoto, mostrando como Transformers podem lidar com cenas complexas e de alta resolução;
- **DeepViT+ViT5:** explora a geração de legendas multilíngues (inglês e vietnamita), ampliando o alcance das aplicações;
- **Adaptive Transformer:** propõe mecanismos de atenção adaptativa para melhorar a seleção de informações visuais e linguísticas;

- **DETR:** apontam para os desafios de escalabilidade e necessidade de hardware de alto desempenho, reforçando as limitações práticas para pesquisadores com recursos limitados.

Dessa forma, os trabalhos analisados não apenas representam o estado da arte da área, mas também evidenciam como diferentes variações do Transformer vêm sendo aplicadas para superar limitações de modelos anteriores. A proposta deste trabalho se insere diretamente nessa linha, pois busca explorar os Transformers especificamente no contexto do Visual Paragraph Generation. Ao reunir contribuições de diferentes áreas (multilíngue, imagens complexas, mecanismos de atenção), a pesquisa aqui desenvolvida tem a oportunidade de não apenas aplicar essa tecnologia avançada, mas também inovar e contribuir com novas perspectivas e resultados para o avanço do conhecimento na área.

Tabela 2 – Comparação entre os trabalhos relacionados.

Autor	**Objetivo**	***Dataset***	**Modelo**	**Pontos Positivos**	**Pontos Negativos**
Jaknamon e Marukatat (2022)	Desenvolver um sistema *end-to-end* de geração de legendas de imagens Transformer-based para a língua tailandesa	Travel, Food e Flickr30k	Vision Transformers como *encoder*, Text Transformers como *decoder* (ThaiTC)	Pioneiro para o tailandês	Necessidade de hardware de alto custo (Tesla V100 com 32GB RAM)
Wang et al. (2022)	Propor uma arquitetura "*pure* Transformer" (CapFormer) para descrever imagens de sensoriamento remoto de alta resolução, focando na representação global	RSICD	*encoder* "*scalable vision transformer*", *decoder* Transformer modificado	Captura características globais, modelagem explícita de relações entre palavras	Necessidade de grande capacidade de processamento para lidar com imagens de alta resolução
Lam et al. (2023)	Investigar Transformers para geração de legendas de imagens, propondo DeepViT+ViT5 com re-atenção no *encoder* para alta qualidade em vietnamita	Flickr8k	DeepViT+ViT5	Inovação com re-atenção	Uso de hardware limitado (GPU Colab com 12GB), dificultando o treinamento em datasets maiores
Zhang et al. (2019)	Aprimorar a geração de legendas de imagens com um *decoder* Transformer mais poderoso e atenção espacial/adaptativa para determinar onde e quando usar informação da imagem	Flickr30k	ResNet-101 como *encoder*, Transformer com atenção espacial e adaptativa como *decoder*	Maior eficiência e paralelização	Dependência de CNN como encoder, mantendo limitações de extração de características visuais
Eluri et al. (2024)	Gerar legendas descritivas combinando InceptionResNetV2 e Detection Transformer, superando RNNs e capturando relações semânticas complexas	COCO Captions Dataset	Integra InceptionResNetV2 e Detection Transformer	Supera limitações de RNNs, integração robusta para relações semânticas complexas e imagens ocluídas	Necessidade de hardware de alta capacidade para treinar com grandes volumes de dados

4 Exemplo de Aplicação do Modelo Vision Transformer

Este capítulo descreve o desenvolvimento dos experimentos realizados neste trabalho, apresentando de forma detalhada os recursos, ferramentas e etapas envolvidas no desenvolvimento. São abordados os dados e modelos empregados, assim como as métricas de avaliação que guiaram a análise do desempenho do modelo de Visual Paragraph Generation.

A estrutura deste capítulo está organizada conforme descrito a seguir. A Seção 4.1 aborda a definição e escolha do modelo, fundamentado na arquitetura Transformer, descrevendo sua estrutura e funcionamento no contexto de Visual Paragraph Generation. Na Seção 4.2, é apresentada a definição e escolha do *dataset* utilizado como base para os experimentos, o Flickr30k, detalhando o motivo de sua escolha e suas principais características. Em seguida, a Seção 4.3 detalha as métricas de avaliação adotadas, com ênfase na métrica BLEU, utilizada para medir o desempenho do modelo. A Seção 4.4 apresenta a realização dos experimentos realizados com o modelo ViT, incluindo os resultados finais das métricas e exemplos de legendas geradas pelo modelo. Também são discutidas as dificuldades encontradas durante a fase experimental, como a grande demanda de recursos computacionais necessária para a execução do código. Finalmente, a Seção 4.5 sumariza as considerações finais do capítulo, destacando os principais conceitos abordados, os resultados obtidos e as contribuições da metodologia aplicada.

4.1 Definição e Escolha do Modelo

Transformers caracterizam uma nova arquitetura de redes neurais que se destacou inicialmente no campo do processamento de linguagem natural, obtendo avanços relevantes em diversas tarefas. Seu principal diferencial está no uso do mecanismo de autoatenção, capaz de identificar e modelar relacionamentos complexos dentro dos dados, mesmo a longas distâncias. Essa característica os tornou especialmente eficazes na extração de características internas dos dados, contribuindo para seu uso crescente em aplicações de inteligência artificial (HAN et al., 2023).

Habitualmente, predominavam as redes neurais convolucionais nas tarefas de visão computacional, cuja estrutura em grade favorece a detecção de padrões locais em imagens. No entanto, as CNNs apresentam limitações quando se trata de compreender relações globais entre regiões distantes da imagem. Para superar essas restrições, os Transformers, originalmente projetados para lidar com sequências textuais, começaram a ser adaptados

para processar imagens. Como as imagens são compostas por uma grade bidimensional de pixels, foi necessário transformá-las em uma sequência que pudesse ser interpretada por um Transformer. Isso foi feito por meio da divisão da imagem em pequenos blocos (*patches*), que são tratados como se fossem palavras em uma frase. Essa abordagem aproveita a capacidade dos Transformers de capturar conexões globais entre os diferentes segmentos da imagem (VASWANI et al., 2017)

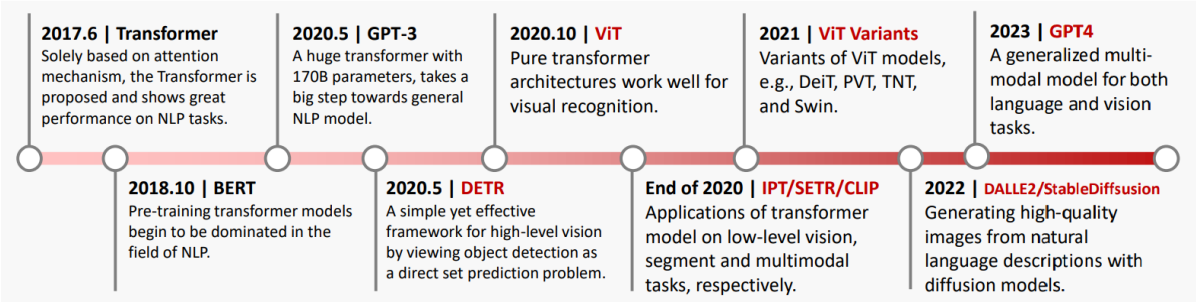


Figura 3 – Linha do tempo de desenvolvimento do Vision Transformer. Os modelos estão destacados em vermelho.

Fonte: (HAN et al., 2023).

A Figura 3 ilustra a evolução do Vision Transformer, um dos modelos mais representativos dessa adaptação. O ViT aplica diretamente um Transformer puro sobre a sequência de patches da imagem, utilizando apenas a parte codificadora da arquitetura original (com exceção de ajustes pontuais, como a normalização em camadas). O resultado da codificação é então encaminhado para uma camada de rede neural do tipo *Multi-layer Perceptron* (MLP), responsável por classificar a imagem como um todo. Em geral, o ViT é treinado inicialmente com grandes volumes de dados e, posteriormente, ajustado (*fine-tuned*) para tarefas específicas com conjuntos menores (HAN et al., 2023). A Figura 4 mostra a arquitetura do Vision Transformer.

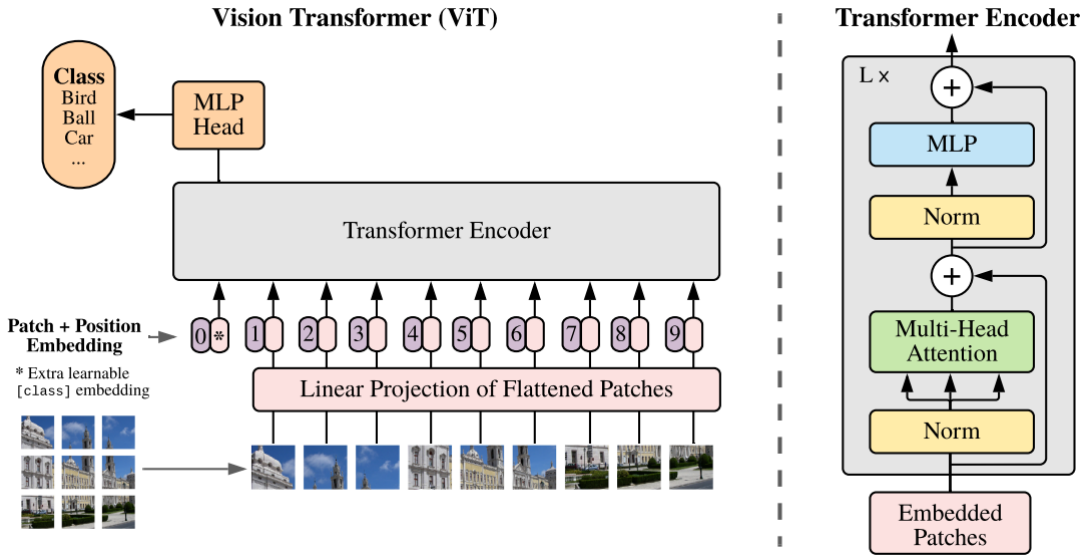


Figura 4 – Visão geral do modelo ViT.

Fonte: (HAN et al., 2023).

Entretanto, o Vision Transformer possui menos viés indutivo específico para imagens em comparação com as CNNs. Nas redes convolucionais, propriedades como localidade, estrutura bidimensional de vizinhança e equivariância por translação estão embutidas nas camadas do modelo. Já no ViT, essas características são bastante limitadas: apenas as camadas MLP mantêm localidade e uma certa resposta proporcional, enquanto as camadas de autoatenção atuam de maneira global. A estrutura espacial em duas dimensões é utilizada apenas no início, quando a imagem é dividida em *patches*, e, mais adiante, para ajustar os *position embeddings* durante o *fine-tuning* em imagens com resoluções diferentes. Fora esses momentos, os embeddings de posição inicial não carregam informações sobre a localização espacial dos *patches*, o que significa que todas as relações espaciais precisam ser aprendidas do zero pelo modelo (DOSOVITSKIY et al., 2020).

Inicialmente, foi realizada uma busca por estudos que aplicassem Transformers à tarefa de Visual Paragraph Generation e disponibilizassem o respectivo código-fonte para reprodução e adaptação. Nesse levantamento, foi identificado o trabalho de (NAIK, 2024), cujo código foi publicado no Kaggle. O modelo proposto utiliza o Vision Transformer como *encoder* e um Transformer puro como *decoder*, servindo de base para a implementação deste estudo.

4.2 Definição e Escolha do Dataset

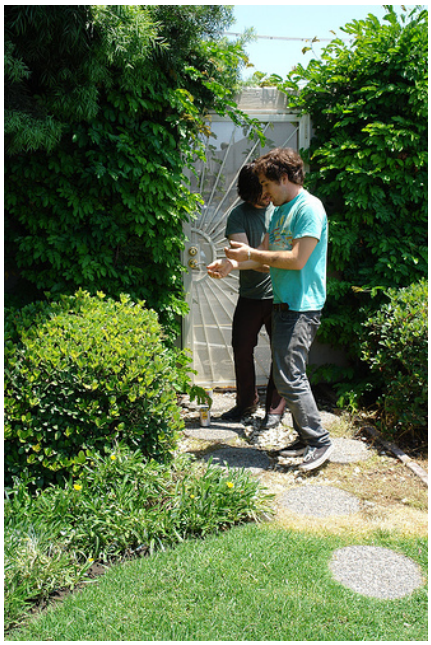
O estudo utilizou o *dataset* Flickr30k (PLUMMER et al., 2015), uma coleção abrangente composta por 31.783 imagens em formato JPG, totalizando aproximadamente 8,3 GB de dados. As imagens apresentam dimensões e tamanhos de arquivo variados,

image_name	comment_number	comment
1000092795.jpg	0	Two young guys with shaggy hair look at their hands while hanging out in the yard .
1000092795.jpg	1	Two young , White males are outside near many bushes .
1000092795.jpg	2	Two men in green shirts are standing in a yard .
1000092795.jpg	3	A man in a blue shirt standing in a garden .
1000092795.jpg	4	Two friends enjoy time spent together .

Figura 5 – Cinco primeiras legendas de uma imagem, extraídas do arquivo CSV do *dataset* Flickr30k.

refletindo a diversidade e complexidade das cenas representadas. Cada imagem retrata cenas do cotidiano, incluindo pessoas em diversas atividades, eventos e animais, e possui cinco legendas descritivas, cada uma escrita por uma pessoa diferente, permitindo múltiplas interpretações da mesma cena, as Figuras 6 a 8 apresentam algumas imagens do dataset junto com suas legendas correspondentes. As legendas estão organizadas em um arquivo CSV com três colunas: o nome da imagem, o número do comentário e o comentário em si, como apresentado na Figura 5.

O Flickr30k foi escolhido para este estudo por ser o *dataset* mais utilizado na Seção 3, oferecendo maior cobertura e diversidade em comparação com o Flickr8k, que possui apenas 8 mil imagens. Apesar de estruturalmente semelhante, com imagens e múltiplas legendas por imagem, o Flickr8k contém um número menor de exemplos, tornando o Flickr30k mais adequado para treinar e avaliar modelos de forma robusta, reduzindo o risco de *overfitting* e aumentando a generalização das legendas geradas.



(a) Imagem do *dataset*.

Two young guys with shaggy hair look at their hands while hanging out in the yard .

Two young , White males are outside near many bushes .

Two men in green shirts are standing in a yard .

A man in a blue shirt standing in a garden .

Two friends enjoy time spent together .

(b) Legendas da imagem.

Figura 6 – Exemplo de uma imagem do *dataset* e suas legendas correspondentes.

A child in a pink dress is climbing up a set of stairs in an entry way .

A little girl in a pink dress going into a wooden cabin .

A little girl climbing the stairs to her playhouse .

A little girl climbing into a wooden playhouse

A girl going into a wooden building .

(a) Imagem do *dataset*. (b) Legendas da imagem.

Figura 7 – Exemplo de uma imagem do *dataset* e suas legendas correspondentes.

Two men , one in a gray shirt , one in a black shirt , standing near a stove .

Two guy cooking and joking around with the camera .

Two men in a kitchen cooking food on a stove .

Two men are at the stove preparing food .

Two men are cooking a meal .

(a) Imagem do *dataset*. (b) Legendas da imagem.

Figura 8 – Exemplo de uma imagem do *dataset* e suas legendas correspondentes.

4.3 Definição das Métricas de Avaliação

Interpretar a acurácia e a perda do treinamento e da validação é crucial para avaliar o desempenho de um modelo de aprendizado de máquina e identificar possíveis problemas, como subajuste e superajuste.

A acurácia, neste trabalho, foi utilizada para medir a proporção de palavras previstas corretamente em relação às palavras da legenda de referência, e não no sentido clássico de classificação de rótulos. Assim, a acurácia corresponde à razão entre o número de *tokens* (palavras) corretamente previstos e o número total de *tokens* da sequência de referência (ROY et al., 2025). Essa métrica pode ser definida como:

$$\text{Acurácia} = \frac{\text{Número de previsões corretas}}{\text{Número total de previsões}}, \tag{4.1}$$

como exemplo de aplicação, tem-se:

- **Legenda de referência:** “Um cachorro corre no parque”;
- **Legenda prevista pelo modelo:** “Um cão corre no parque”.

Diante disso, comparando palavra por palavra:

- “Um” → correto;
- “cachorro” → incorreto (modelo previu “cão”);
- “corre” → correto;
- “no” → correto;
- “parque” → correto.

a acurácia desse exemplo seria calculada como:

$$\text{Acurácia} = \frac{4}{5} = 0{,}8\ (80\%)$$

A função de perda fornece uma medida de quão bem as previsões do modelo correspondem aos valores reais. Ela mede a diferença entre a distribuição de probabilidade prevista e a distribuição real dos valores verdadeiros. O objetivo durante o treinamento é minimizar essa perda, resultando em previsões mais precisas (FOSTER, 2023). A função de perda é definida conforme descrito a seguir:

$$\mathcal{L} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, x)$$

em que:

- T é o comprimento da sequência de referência;
- y_t é a palavra correta na posição t;
- $P(y_t \mid y_{<t}, x)$ é a probabilidade atribuída pelo modelo à palavra correta y_t, dado o contexto anterior $y_{<t}$ e a imagem x.

A métrica BLEU foi utilizada para complementar o trabalho, ela é usada para comparar um texto gerado, como uma tradução, com uma ou mais versões de referência. Embora tenha sido originalmente criado para avaliação de traduções automáticas, seu uso se estende a outras tarefas de processamento de linguagem natural, como geração de paráfrases e resumos. Entre suas vantagens, destacam-se a rapidez no cálculo, o baixo

custo computacional, a independência de idioma e, principalmente, a forte correlação com avaliações humanas, o que o torna um recurso valioso para pesquisadores e desenvolvedores.

O funcionamento do BLEU está mais voltado para a precisão, ou seja, quão próximo o resultado gerado está do texto de referência, aplicando ainda uma penalidade para saídas excessivamente curtas, de modo a evitar que respostas incompletas recebam pontuação alta(CHATOUI; ATA, 2021).

Segundo (PAPINENI et al., 2002), a pontuação BLEU é calculada pela seguinte fórmula:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \cdot \ln(p_n)\right)$$

em que:

- p_n é a precisão dos n-gramas, obtida como a razão entre o número de n-gramas que aparecem tanto na tradução gerada pela máquina quanto nas traduções de referência e o número total de n-gramas na tradução gerada pela máquina;
- BP é a Penalidade de Brevidade (*Brevity Penalty*), que ajusta a pontuação quando a tradução é mais curta que a referência. É calculada por:

  $$BP = \min\left(1, \frac{\text{comprimento\_de\_referência}}{\text{comprimento\_traduzido}}\right)$$

  em que:

  - comprimento_de_referência é o número total de palavras nas traduções de referência;
  - comprimento_traduzido é o número total de palavras na tradução gerada pela máquina.

A pontuação vai de 0 a 1 (ou de 0% a 100%, quando expressa em porcentagem). Um valor de 1 indica correspondência total entre a saída e a referência, enquanto 0 significa ausência completa de similaridade. Por ser automatizado, rápido e frequentemente alinhado ao julgamento humano, o BLEU consolidou-se como uma das métricas mais utilizadas em benchmarks e pesquisas na área de PLN.

Como exemplo ilustrativo tem-se:

- **Legenda de referência:** “Um cachorro corre no parque”;
- **Legenda prevista:** “Um cão corre no parque”.

Pare este exemplo, o cálculo dos *n-gramas* é dado conforme descrito a seguir:

- **Unigramas (1-palavra):**
  Referência: [Um, cachorro, corre, no, parque]
  Prevista: [Um, cão, corre, no, parque]
  Corretos: 4 de 5 $\Rightarrow$ $p_1 = \frac{4}{5} = 0{,}80$
- **Bigramas (2-palavras):**
  Referência: [Um cachorro, cachorro corre, corre no, no parque]
  Prevista: [Um cão, cão corre, corre no, no parque]
  Corretos: 2 de 4 $\Rightarrow$ $p_2 = \frac{2}{4} = 0{,}50$

diante disso, considerando $N = 2$, pesos iguais $w_1 = w_2 = 0{,}5$ e sem penalidade por comprimento ($BP = 1$):

$$BLEU = 1 \cdot \exp\left(0{,}5 \cdot \log(0{,}8) + 0{,}5 \cdot \log(0{,}5)\right)$$

$$BLEU \approx \exp\left(-0{,}111 - 0{,}347\right) = \exp(-0{,}458) \approx 0{,}632$$

Portanto, o valor de $BLEU$ neste exemplo é aproximadamente 0,63.

4.4 Realização dos Experimentos e Resultados

Antes da execução dos experimentos, o código original disponibilizado por (NAIK, 2023) no Kaggle passou por ajustes pontuais para realização de finetuning do modelo pré-treinado. Foram removidas imagens que apresentavam ao menos uma legenda com inconsistências, além de ter sido implementada uma correção usando *top-k sampling* para mitigar a repetição de palavras nas legendas geradas.

Os experimentos foram conduzidos no ambiente Kaggle, utilizando um notebook com duas GPUs Nvidia Tesla T4, cada uma com 15 GB de memória dedicada. O ambiente de execução também disponibiliza 4 núcleos de CPU Intel(R) Xeon(R) @ 2.20GHz e aproximadamente 30 GB de memória RAM para processamento geral, além de armazenamento em disco de até 57,6 GB e um limite de 12 horas por sessão e 30h por semana.

Foram testadas quatro combinações de hiperparâmetros, variando o número de épocas e o tamanho do *batch*, conforme descrito a seguir:

- Configuração 1: 5 épocas, *batch size* 96;
- Configuração 2: 10 épocas, *batch size* 96;
- Configuração 3: 5 épocas, *batch size* 192;

- Configuração 4: 10 épocas, *batch size* 192.

Os experimentos foram conduzidos com o objetivo de determinar a melhor configuração entre as quatro testadas. Cada configuração de hiperparâmetros foi executada três vezes, resultando em um total de 12 execuções, de modo a reduzir o impacto de variações aleatórias e possibilitar uma avaliação mais robusta.

O conjunto de dados foi inicialmente dividido em 80% para a fase experimental e 20% para o teste final. Dentro da parcela destinada aos experimentos (80% do total), realizou-se uma nova partição em 80% para treinamento, 10% para validação e 10% para teste interno, exemplificado na Figura 9.

A Tabela 3 apresenta os resultados obtidos ao longo de 5 épocas para cada métrica, considerando um *batch size* de 96, referente a um dos conjuntos de hiperparâmetros analisados. A *loss function* indica o erro do modelo, a acurácia mede a proporção de previsões corretas e o BLEU avalia a similaridade entre a legenda gerada e as referências. Observa-se que a *loss* tende a diminuir nas primeiras épocas, indicando aprendizado do modelo, com pequenas variações nas épocas posteriores. A acurácia aumenta rapidamente nas primeiras épocas e se mantém estável a partir da terceira, sugerindo que o modelo já alcançou uma boa convergência. Os valores de BLEU apresentam pequenas oscilações ao longo das épocas, refletindo a dificuldade da tarefa de geração de legendas. Os resultados completos para todos os conjuntos de hiperparâmetros estão disponíveis nas Tabelas 5, 6, 7 e 8, localizadas no Apêndice A.

Tabela 3 – Resultados da Configuração 1.

Epoch	Replicação 1			Replicação 2			Replicação 3		
	Loss	Acurácia	BLEU	*Loss*	Acurácia	BLEU	*Loss*	Acurácia	BLEU
1	1.5578	0.7249	0.0233	1.6748	0.7108	0.0212	1.5642	0.7258	0.0237
2	1.4427	0.7358	0.0291	1.4763	0.7327	0.0326	1.4553	0.7350	0.0358
3	1.4039	0.7409	0.0237	1.4335	0.7374	0.0357	1.4369	0.7374	0.0280
4	1.4328	0.7386	0.0131	1.4430	0.7356	0.0322	1.4343	0.7369	0.0294
5	1.4139	0.7398	0.0181	1.4500	0.7352	0.0301	1.4503	0.7362	0.0308

Após a realização dos experimentos, foram calculados a média, o desvio padrão e o Intervalo de Confiança (IC) de 95% para cada métrica, utilizando os resultados obtidos no conjunto de teste interno. Esses cálculos foram realizados em Python, utilizando a biblioteca `scipy.stats`, seguindo a documentação (SciPy Developers, 2025).

Os resultados em mais detalhes estão presentes nas Tabelas 9, 10, 11 e 12 do Apêndice A, bem como na Figura 10.

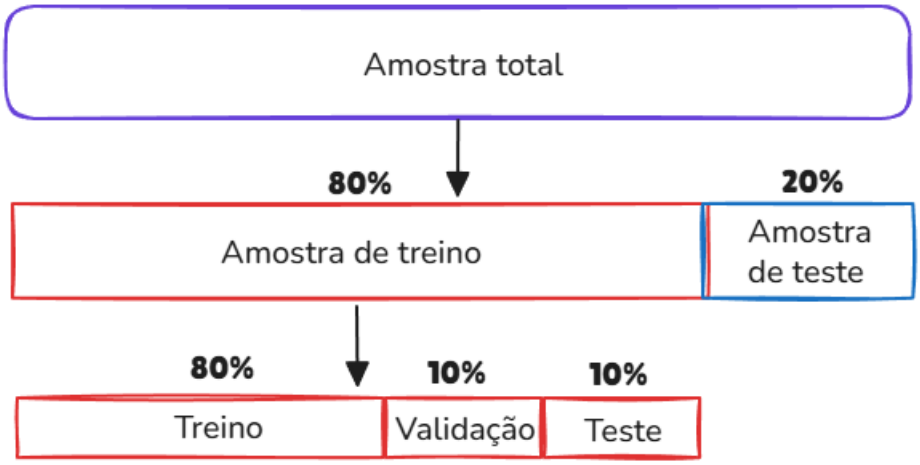


Figura 9 – Divisão do *dataset* nos experimentos.

Tabela 4 – Cálculos da Configuração 1.

Replicação	**Loss**	**Acurácia**	**BLEU**
1	1.3985	0.7416	0.0239
2	1.4495	0.7354	0.0297
3	1.4503	0.7362	0.0308
Média	1.4327	0.7377	0.0281
Desvio Padrão	0.0297	0.0034	0.0037
IC 95%	(1.3590, 1.5065)	(0.7293, 0.7461)	(0.0190, 0.0373)

A Tabela 4, referente à Configuração 1, apresenta a *loss function*, a acurácia e o BLEU obtidos em cada replicação, juntamente com a média, o desvio padrão e o intervalo de confiança de 95%. A *loss* média foi de 1.4327, com IC 95% entre 1.3590 e 1.5065, indicando estabilidade do modelo entre replicações. A acurácia média alcançou 0.7377 (IC 95%: 0.7293–0.7461), mostrando consistência nas previsões corretas, enquanto o BLEU médio foi de 0.0281 (IC 95%: 0.0190–0.0373), refletindo a similaridade entre as legendas geradas e as de referência.

Comparando todas as configurações, a Configuração 3 apresentou o melhor desempenho, caracterizando-se pela menor *loss* média e pelo menor intervalo de confiança na acurácia, como mostra a Figura 10, o que indica não apenas erro reduzido, mas também maior confiabilidade nas previsões. Além disso, a pontuação BLEU manteve valores estáveis entre replicações, evidenciando robustez na geração de legendas. Esses resultados permitem avaliar o desempenho médio e a segurança estatística de cada métrica, destacando a Configuração 3 como a mais eficaz e confiável para a tarefa proposta.

Para o treinamento final, a amostra de treino foi dividida apenas em dois subconjuntos: 80% para treinamento e 20% para validação. A etapa de teste foi realizada

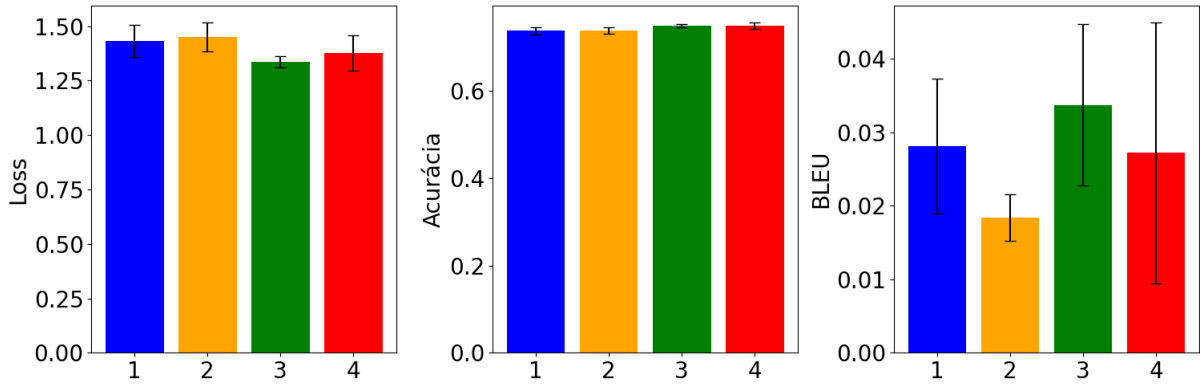


Figura 10 – Intervalo de confiança de 95% para cada métrica: (a) Configuração 1 em azul (Config.1): número de época iqual a 5 e *batch size* de 96; (b) Configuração 2 em amarelo (Config.2): número de épocas igual 10 e *batch size* de 96; (c) Configuração 3 em verde (Config.3): número de épocas de 5 e *batch size* de 192; (d) Configuração 4 em vermelho (Config.4): número de épocas de 10 e *batch size* de 192.

utilizando exclusivamente a amostra de teste previamente separada no início dos experimentos, garantindo que nenhuma das imagens desse conjunto tivesse sido utilizada em qualquer fase de treinamento. A divisão dos dados está ilustrada na Figura 11.

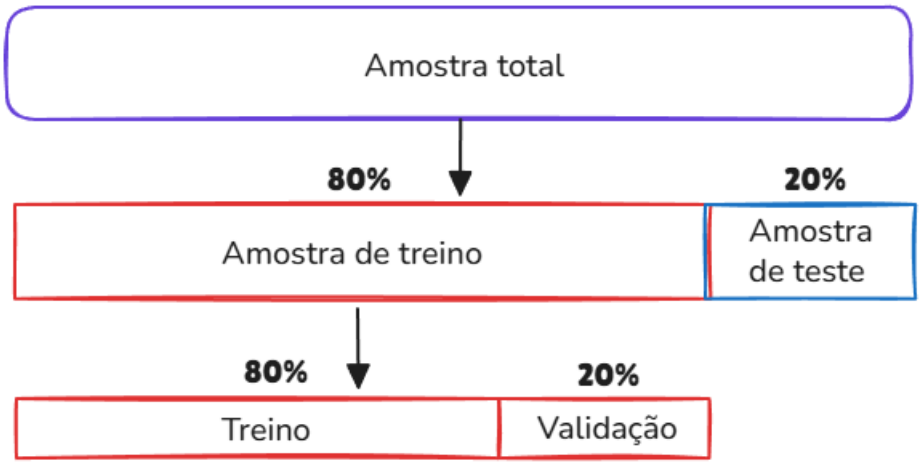


Figura 11 – Divisão do *dataset* na avaliação final.

A Figura 12 apresenta os resultados obtidos na avaliação final realizada sobre a amostra de teste, com *loss function* de 1,3469, acurácia de 0,7468 e pontuação BLEU de 0,0236. Esses valores estão próximos aos observados durante o treinamento e a validação, o que indica que o modelo conseguiu generalizar adequadamente, sem sinais de *overfitting* nem de *underfitting*. O *overfitting* ocorre quando o modelo "decora" os dados de treinamento, apresentando ótimo desempenho nessa fase, mas baixo desempenho em novos dados. Já o *underfitting* acontece quando o modelo não aprende os padrões relevantes dos dados, obtendo resultados ruins tanto no treino quanto no teste (LÓPEZ; LÓPEZ;

CROSSA, 2022). Assim, a consistência dos resultados entre as fases reforça a estabilidade do aprendizado. Apesar da acurácia satisfatória e da boa estabilidade do *loss*, a pontuação BLEU relativamente baixa revela espaço para melhorias na qualidade das legendas, que podem ser alcançadas com ajustes de hiperparâmetros, aumento do tempo de treinamento ou aplicação de outras técnicas de *fine-tuning*.

Loss	Acurácia	BLEU
1.3469	0.7468	0.0236

Figura 12 – Resultado do teste final.

Na Figura 13, é apresentado um exemplo comparando a legenda de referência do *dataset* com a legenda gerada pelo modelo. A legenda de referência descreve com precisão a cena: *"a man and a woman standing on a sidewalk in between buildings"*, indicando a presença de duas pessoas paradas em uma calçada entre os prédios.

A legenda criada pelo modelo, *"a man is walking down a canal of steps"*, demonstra uma interpretação parcial da cena. O modelo conseguiu identificar a presença de um homem e associar a imagem a um ambiente urbano com canal, mas falhou em dois aspectos importantes:

1. Não reconheceu a presença da mulher na cena;
2. Descreveu uma ação inexistente (*"walking down a canal of steps"*), possivelmente confundindo os degraus próximos ao canal.

Esse exemplo evidencia uma limitação recorrente observada nas legendas geradas: dificuldade em identificar múltiplos elementos humanos simultaneamente e tendência a descrever ações genéricas ou incorretas. Apesar disso, nota-se que o modelo conseguiu captar parte do contexto visual, relacionando o cenário ao canal, o que demonstra um aprendizado parcial do arranjo espacial.

Já na Figura 14, observa-se uma cena esportiva em que um praticante de *snowboard* realiza um salto. A legenda de referência descreve corretamente a ação e fornece detalhes precisos: *"a snowboarder in gray pants is doing a jump"*.

A legenda gerada pelo modelo, *"snowboarder in pink is performing a jump on a snow"*, demonstra que o modelo conseguiu capturar o contexto geral da imagem, identificando corretamente o esporte e a ação de salto. No entanto, há uma falha relevante na descrição: a cor da roupa foi identificada incorretamente como rosa, quando, na verdade, o atleta veste calças cinzas e jaqueta camuflada.

Esse exemplo também ilustra uma limitação recorrente do modelo, semelhante ao exemplo da Figura 13: dificuldade em descrever com precisão detalhes específicos, como cores ou padrões de roupas. Apesar disso, a estrutura geral da legenda está coerente e demonstra que o modelo aprendeu a associar corretamente a postura e o ambiente ao contexto esportivo.

A Figura 15 reforça esse padrão de desempenho do modelo. A legenda de referência do *dataset* descreve com precisão a cena: *"the large brown dog is running on the beach by the ocean"*, destacando o porte do animal, sua cor, a ação de correr e o cenário da praia próximo ao mar. Já a legenda gerada pelo modelo, *"a small brown and white dog stands at the sand"*, apresenta acertos e falhas importantes. Por um lado, o modelo identificou corretamente que se trata de um cachorro em um ambiente de areia. No entanto, houve erros consideráveis: descreveu o animal como pequeno, mencionou uma cor inexistente (branco) e, além disso, trocou a ação de correr por "estar parado" (*stands*).

Assim como nos exemplos anteriores, nota-se que o modelo capta bem o contexto geral da imagem, mas tem dificuldade em descrever características específicas ou ações com maior precisão. Esse comportamento consistente ao longo dos exemplos analisados indica que, embora o modelo tenha aprendido padrões visuais relevantes, sua capacidade de generalização ainda é limitada quando se trata de detalhes mais refinados, sugerindo a necessidade de mais dados de treinamento ou de estratégias de *fine-tuning* mais direcionadas.

As principais dificuldades enfrentadas neste trabalho foram duas. A primeira foi encontrar estudos que utilizassem modelos Transformers como referência, com códigos disponíveis que pudessem ser treinados para a realização dos experimentos e que não exigissem recursos computacionais elevados. A segunda dificuldade esteve relacionada à demanda computacional. O tempo médio de execução variou conforme a configuração: 2h30min para a configuração 1, 4h35min para a configuração 2, 2h23min para a configuração 3 e 4h20min para a configuração 4. Ao utilizar apenas os recursos básicos do Kaggle, sem GPU, o código não conseguia ser concluído dentro do limite máximo de 12 horas por sessão. Além disso, o limite de uso da GPU é de 30 horas por semana, o que restringia significativamente a exploração de diferentes parâmetros e a realização de múltiplas replicações, dificultando a diversidade dos experimentos.

TRUE: [START] a man and a woman standing on a sidewalk in between buildings [END]

PRED: [START] a man is walking down a canal of steps [END]

(a) Imagem do *dataset*. (b) *TRUE*: Legenda do *dataset*, PRED: legenda gerada pelo modelo.

Figura 13 – Exemplo de legenda gerada no modelo final.

TRUE: [START] a snowboarder in gray pants is doing a jump [END]

PRED: [START] snowboarder in pink is performing a jump on a snow [END]

(a) Imagem do *dataset*. (b) *TRUE*: Legenda do *dataset*, PRED: legenda gerada pelo modelo.

Figura 14 – Exemplo de legenda gerada no modelo final.

TRUE: [START] the large brown dog is running on the beach by the ocean [END]

PRED: [START] a small brown and white dog stands at the sand [END]

(a) Imagem do *dataset*. (b) *TRUE*: Legenda do *dataset*, PRED: legenda gerada pelo modelo.

Figura 15 – Exemplo de legenda gerada no modelo final.

As Figuras 16 a 19, localizadas no Apêndice B, apresentam exemplos de imagens acompanhadas de uma legenda original do *dataset* e da legenda correspondente gerada pelo modelo.

4.5 Considerações Finais

Os experimentos conduzidos tiveram caráter ilustrativo, com o objetivo de demonstrar, na prática, a aplicação de um modelo pré-treinado baseado em Transformer

para o problema de Visual Paragraph Generation. Nesse sentido, os resultados devem ser interpretados não como uma solução definitiva, mas como uma prova de conceito que evidencia as potencialidades e limitações da abordagem.

Inicialmente, foi selecionado o *dataset* Flickr30k, composto por cerca de 30 mil imagens, cada uma acompanhada de cinco descrições textuais. A escolha desse conjunto se justifica pela ampla utilização em tarefas de Visual Paragraph Generation, bem como pela sua diversidade semântica e contextual, permitindo avaliar a capacidade do modelo em gerar descrições mais complexas e detalhadas.

Na etapa seguinte, adotou-se a arquitetura Transformer, combinando um codificador visual baseado em ViT com um decodificador textual. Essa configuração foi aplicada à tarefa de geração de parágrafos visuais, e o desempenho do modelo foi avaliado por meio da métrica BLEU. Foram testadas quatro configurações distintas de treinamento, variando o tamanho do *batch* e o número de épocas.

Os resultados obtidos indicaram que a Configuração 3 (5 épocas, *batch size* de 192) apresentou o melhor equilíbrio entre as métricas, alcançando acurácia de 74,89%, *loss* média de 1,3364 e pontuação BLEU média de 0,0337. Em contraste, a Configuração 2 (10 épocas, *batch size* de 96) apresentou a menor pontuação BLEU (0,0184), apesar de manter acurácia similar (73,86%). Isso sugere que aumentar apenas o número de épocas não resulta necessariamente em melhorias qualitativas nas legendas geradas.

As figuras produzidas pelo modelo (Figuras 13 a 15) evidenciam essa limitação: embora a estrutura sintática das legendas esteja correta, nota-se pobreza semântica e pouca variação vocabular. Por exemplo, em imagens com contextos distintos, as legendas se restringiram à descrição superficial de elementos, sem capturar as relações entre os elementos ou a dinâmica da cena.

Além disso, o ambiente de execução representou um fator restritivo importante: a alta demanda de processamento em GPU, associada ao consumo elevado de memória e ao tempo de treinamento reduzido, comprometeu a possibilidade de ajustes mais sofisticados, como o *fine-tuning* com mais épocas e lotes maiores. Esses aspectos técnicos, aliados à complexidade inerente da tarefa de Visual Paragraph Generation, ajudam a explicar as deficiências observadas nos resultados.

Portanto, este capítulo não apenas apresentou os resultados obtidos, mas também discutiu suas implicações e limitações, oferecendo uma base teórica e empírica para avanços futuros na geração de parágrafos visuais por meio de IA generativa.

5 Conclusões e Trabalhos Futuros

O objetivo geral deste trabalho é investigar os modelos de IA Generativa existentes na literatura para a solução do problema de Visual Paragraph Generation. Esse objetivo foi alcançado por meio da consolidação de uma base teórica, da análise crítica de trabalhos relacionados, da escolha de um modelo pré-treinado baseado em Transformer e da execução de experimentos que permitiram avaliar seu desempenho no contexto estudado.

O primeiro objetivo específico consistiu em realizar uma revisão da literatura sobre os principais modelos de IA Generativa aplicados ao problema de Visual Paragraph Generation. Esse objetivo foi atingido na primeira etapa da metodologia, com a apresentação dos fundamentos de IA Generativa, sua evolução histórica, aplicações mais relevantes e as arquiteturas de maior destaque. Essa análise foi fundamental para compreender a importância do tema e situar o problema de Visual Paragraph Generation dentro do campo da inteligência artificial contemporânea.

O segundo objetivo específico buscou identificar um modelo baseado em Transformer que fosse viável em termos de custo-benefício para aplicação no problema. Para alcançar esse objetivo, duas etapas da metodologia foram relevantes. Primeiramente, no estudo da arquitetura Transformer, foi possível detalhar seu funcionamento interno e destacar o papel do mecanismo de atenção como elemento central de sua eficiência. Em seguida, na análise dos trabalhos relacionados, foi realizada uma discussão crítica sobre as propostas da literatura, evidenciando pontos positivos, como a capacidade dos Transformers em lidar com dependências de longo alcance, e limitações, como a necessidade de grandes volumes de dados e recursos computacionais. Esse conjunto de análises justificou a escolha do Vision Transformer como modelo experimental deste trabalho, devido à sua ampla utilização em visão computacional e ao equilíbrio entre desempenho e custo de aplicação.

O terceiro objetivo específico propôs a realização de experimentos com o modelo selecionado, de modo a exemplificar seu uso e demonstrar suas possíveis limitações. Essa etapa foi desenvolvida com a aplicação do ViT como codificador visual integrado a um decodificador Transformer para a geração de descrições textuais. Os experimentos foram conduzidos com divisão do conjunto de dados em treino, validação e teste, e avaliados por métricas como acurácia, perda e BLEU. Os resultados demonstraram que a Configuração 3, com 5 épocas e *batch size* de 192, apresentou o desempenho mais equilibrado entre as métricas avaliadas, alcançando acurácia de 74,89%, *loss* média de 1,3364 e pontuação BLEU média de 0,0337. Por outro lado, a Configuração 2, que utilizou dez épocas e *batch size* de 96, obteve a menor pontuação BLEU (0,0184), mesmo mantendo uma acurácia semelhante, de 73,86%. Esses resultados indicam que o aumento do número de épocas, sem

ajustes adicionais na arquitetura ou nos parâmetros de treinamento, não garante melhorias qualitativas nas legendas geradas. Os resultados obtidos na avaliação final sobre a amostra de teste apresentaram *loss function* de 1,3469, acurácia de 0,7468 e pontuação BLEU de 0,0236. Esses valores são próximos aos observados durante as fases de treinamento e validação, indicando que o modelo foi capaz de generalizar de maneira adequada, sem sinais de *overfitting* ou *underfitting*. Dessa forma, a consistência entre os resultados das diferentes fases reforça a estabilidade do aprendizado. Então, os resultados mostraram estabilidade nos valores de acurácia, mas também revelaram deficiências semânticas, como a baixa diversidade lexical das descrições geradas e a dificuldade em capturar relações mais complexas entre os elementos da imagem. Além disso, as limitações computacionais, como a alta demanda de GPU, o consumo elevado de memória e o tempo limitado para execução no Kaggle, impactaram diretamente na possibilidade de explorar *fine-tuning* em maior escala ou testar configurações mais sofisticadas.

De modo geral, este trabalho permitiu concluir que a arquitetura Transformer se mostra como a mais promissora para o problema de Visual Paragraph Generation, justificando seu amplo uso em diferentes aplicações de IA Generativa. Contudo, também se verificou que desafios permanecem, principalmente no que se refere à qualidade semântica das descrições e às restrições de hardware, fatores que limitaram a obtenção de resultados mais robustos. Assim, este estudo contribui de forma exploratória e crítica, consolidando a relevância do Transformer na área e abrindo espaço para novas investigações.

Como perspectiva para trabalhos futuros, vislumbra-se a aplicação do Visual Paragraph Generation em cenários de impacto social, como a descrição detalhada de imagens voltada para pessoas com deficiência visual. Essa aplicação permitiria oferecer uma experiência mais rica e contextualizada, superando as limitações de descrições curtas ou fragmentadas atualmente disponíveis em sistemas de acessibilidade.

Além disso, uma possível linha de pesquisa envolve a incorporação de mecanismos de curiosidade computacional ao processo de geração textual. Essa abordagem poderia ser integrada ao mecanismo de atenção dos modelos baseados em Transformers, de modo a estimular o modelo a explorar representações visuais mais diversificadas e menos previsíveis, resultando em parágrafos com maior riqueza de detalhes, fluidez narrativa e relevância contextual. Dessa forma, alia-se uma aplicação prática de utilidade social com avanços metodológicos capazes de ampliar as capacidades de modelos de IA Generativa no domínio de Visual Paragraph Generation.

Referências

ALMOHSEN, K. Visual paragraph generation: Review. In: *2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD)*. [S.l.: s.n.], 2023. p. 1–6. Citado na página 17.

ALTO, V. *Modern Generative AI with ChatGPT and OpenAI Models*. [S.l.]: Editora, 2023. Citado na página 15.

BANDI, A.; ADAPA, P. V. S. R.; KUCHI, Y. E. V. P. K. The power of generative ai: A review of requirements, models, input–output formats, evaluation metrics, and challenges. *Future Internet*, MDPI, v. 15, n. 8, p. 260, 2023. Citado 2 vezes nas páginas 24 e 25.

BENGESI, S. et al. Advancements in generative ai: A comprehensive review of gans, gpt, autoencoders, diffusion model, and transformers. *IEEE Access*, v. 12, p. 69812–69837, 2024. Citado 2 vezes nas páginas 27 e 28.

BISHOP, C. M.; BISHOP, H. *Deep learning: Foundations and concepts*. [S.l.]: Springer Nature, 2023. Citado 4 vezes nas páginas 30, 31, 32 e 33.

CAO, Y. et al. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*, 2023. Citado 2 vezes nas páginas 21 e 23.

CHATOUI, H.; ATA, O. Automated evaluation of the virtual assistant in bleu and rouge scores. In: *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*. [S.l.: s.n.], 2021. p. 1–6. Citado na página 52.

CHATTERJEE, M.; SCHWING, A. G. Diverse and coherent paragraph generation from images. In: *Proceedings of the European conference on computer vision (ECCV)*. [S.l.: s.n.], 2018. p. 729–744. Citado na página 17.

CHE, W. et al. Visual relationship embedding network for image paragraph generation. *IEEE Transactions on Multimedia*, v. 22, n. 9, p. 2307–2320, 2020. Citado na página 16.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. [S.l.: s.n.], 2019. p. 4171–4186. Citado na página 35.

DOSOVITSKIY, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. Citado na página 48.

ELURI, Y. et al. Image captioning using visual attention and detection transformer model. In: IEEE. *2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*. [S.l.], 2024. p. 1–4. Citado 2 vezes nas páginas 42 e 45.

FOSTER, D. *Generative deep learning*. [S.l.]: "O'Reilly Media, Inc.", 2022. Citado na página 15.

FOSTER, D. *Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play*. 2nd. ed. Sebastopol, CA: O'Reilly Media, Inc., 2023. ISBN 9781098134181. Disponível em: <https://www.oreilly.com/library/view/generative-deep-learning/9781098134174/>. Citado 2 vezes nas páginas 22 e 51.

GHOJOGH, B.; GHODSI, A. *Attention Mechanism, Transformers, BERT, and GPT: Tutorial and Survey*. 2020. Preprint, available at <https://doi.org/10.31219/osf.io/m6gcn>. Citado na página 35.

HAN, K. et al. A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Institute of Electrical and Electronics Engineers (IEEE), v. 45, n. 1, p. 87–110, jan. 2023. ISSN 1939-3539. Disponível em: <http://dx.doi.org/10.1109/TPAMI.2022.3152247>. Citado 3 vezes nas páginas 46, 47 e 48.

HOSSAIN, M. Z.; ZAMAN, F. U.; ISLAM, M. R. Advancing ai-generated image detection: Enhanced accuracy through cnn and vision transformer models with explainable ai insights. In: *2023 26th International Conference on Computer and Information Technology (ICCIT)*. [S.l.: s.n.], 2023. p. 1–6. Citado na página 28.

JAKNAMON, T.; MARUKATAT, S. Thaitc: thai transformer-based image captioning. In: IEEE. *2022 17th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*. [S.l.], 2022. p. 1–4. Citado 2 vezes nas páginas 37 e 45.

JOVANOVIC, M.; CAMPBELL, M. Generative artificial intelligence: Trends and prospects. *Computer*, IEEE Computer Society, Los Alamitos, CA, USA, v. 55, n. 10, p. 107–112, oct 2022. ISSN 1558-0814. Citado 2 vezes nas páginas 27 e 28.

KOROTEEV, M. V. BERT: A review of applications in natural language processing and understanding. *CoRR*, abs/2103.11943, 2021. Disponível em: <https://arxiv.org/abs/2103.11943>. Citado na página 35.

KRAUSE, J. et al. A hierarchical approach for generating descriptive image paragraphs. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2017. p. 317–325. Citado 2 vezes nas páginas 15 e 17.

LAM, K. N. et al. Deep vision transformer and t5-based for image captioning. In: IEEE. *2023 RIVF International Conference on Computing and Communication Technologies (RIVF)*. [S.l.], 2023. p. 306–311. Citado 2 vezes nas páginas 39 e 45.

LIN, T. et al. *A Survey of Transformers*. 2021. Disponível em: <https://arxiv.org/abs/2106.04554>. Citado na página 31.

LLERENA-IZQUIERDO, J. et al. Towards an experience in ai-driven development for programming applied to multimedia using tabnine. In: *2024 IEEE URUCON*. [S.l.: s.n.], 2024. p. 1–5. Citado na página 27.

LÓPEZ, O. A. M.; LÓPEZ, A. M.; CROSSA, J. Overfitting, model tuning, and evaluation of prediction performance. In: *Multivariate statistical machine learning methods for genomic prediction*. [S.l.]: Springer, 2022. p. 109–139. Citado na página 57.

LUO, Y. et al. Curiosity-driven reinforcement learning for diverse visual paragraph generation. In: *Proceedings of the 27th ACM International Conference on Multimedia*. [S.l.: s.n.], 2019. p. 2341–2350. Citado 2 vezes nas páginas 16 e 17.

MCINTOSH, T. R. et al. *From Google Gemini to OpenAI Q* (Q-Star): A Survey of Reshaping the Generative Artificial Intelligence (AI) Research Landscape*. 2023. Disponível em: <https://arxiv.org/abs/2312.10868>. Citado na página 26.

NAIK, K. *Image Captioning | Vision Transformer*. 2023. Kaggle Notebook. Acessado em: 23 mai. 2025. Disponível em: <https://www.kaggle.com/code/kedarnathnaik/image-captioning-vision-transformer>. Citado na página 53.

NAIK, K. *Image Captioning using Vision Transformer (ViT) | Transfer learning*. 2024. <https://medium.com/@kednaik/image-captioning-using-vision-transformer-vit-transfer-learning-4d59730b4e92>. Citado na página 48.

OOI, K.-B. et al. The potential of generative artificial intelligence across disciplines: Perspectives and future directions. *Journal of Computer Information Systems*, Taylor & Francis, v. 65, n. 1, p. 76–107, 2025. Citado na página 28.

PANDEY, R. et al. Generative ai-based text generation methods using pre-trained gpt-2 model. *arXiv preprint arXiv:2404.01786*, 2024. Citado na página 24.

PAPINENI, K. et al. Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. [S.l.: s.n.], 2002. p. 311–318. Citado na página 52.

PLUMMER, B. A. et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2015. p. 2641–2649. Citado 2 vezes nas páginas 19 e 48.

RAHMAN, M. M.; WATANOBE, Y. Chatgpt for education and research: Opportunities, threats, and strategies. *Applied Sciences*, MDPI, v. 13, n. 9, p. 5783, 2023. Citado na página 29.

RAMDURAI, B.; ADHITHYA, P. The impact, advancements and applications of generative ai. *International Journal of Computer Science and Engineering*, v. 10, n. 6, p. 1–8, 2023. Citado na página 22.

RANI, G.; SINGH, J.; KHANNA, A. Comparative analysis of generative ai models. In: *2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)*. [S.l.: s.n.], 2023. p. 760–765. Citado 3 vezes nas páginas 21, 24 e 27.

ROY, A. K. et al. Resemotenet: Bridging accuracy and loss reduction in facial emotion recognition. *IEEE Signal Processing Letters*, v. 32, p. 491–495, 2025. Citado na página 50.

SciPy Developers. *Statistical Functions (scipy.stats) — SciPy Reference Manual*. [S.l.], 2025. Acesso em: 29 jul. 2025. Disponível em: <https://docs.scipy.org/doc/scipy/reference/stats.html>. Citado na página 54.

SINGH, S.; KUMAR, S.; MEHRA, P. Chat gpt & google bard ai: A review. In: *2023 International Conference on IoT, Communication and Automation Technology (ICICAT)*. [S.l.: s.n.], 2023. p. 1–6. Citado na página 26.

SONG, Y.; CHEN, S.; JIN, Q. Towards diverse paragraph captioning for untrimmed videos. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2021. p. 11245–11254. Citado na página 16.

STRATIS, K. *What is Generative AI?* USA: O'Reilly, 2023. ISBN 978-1-098-17087-5. Citado na página 22.

VASWANI, A. et al. Attention is all you need. *Advances in neural information processing systems*, Long Beach, CA, v. 30, n. 1, p. 5999–6009, 2017. Citado 5 vezes nas páginas 30, 31, 33, 34 e 47.

WANG, J. et al. Capformer: Pure transformer for remote sensing image caption. In: IEEE. *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*. [S.l.], 2022. p. 7996–7999. Citado 2 vezes nas páginas 38 e 45.

WERMELINGER, M. Using github copilot to solve simple programming problems. In: . New York, NY, USA: Association for Computing Machinery, 2023. (SIGCSE 2023), p. 172–178. ISBN 9781450394314. Disponível em: <https://doi.org/10.1145/3545945.3569830>. Citado na página 27.

XIE, Y. et al. Visual clues: Bridging vision and language foundations for image paragraph captioning. *Advances in Neural Information Processing Systems*, v. 35, p. 17287–17300, 2022. Citado na página 16.

XU, C. et al. Interactive key-value memory-augmented attention for image paragraph captioning. In: *Proceedings of the 28th international conference on computational linguistics*. [S.l.: s.n.], 2020. p. 3132–3142. Citado na página 15.

XU, K. et al. Show, attend and tell: Neural image caption generation with visual attention. In: PMLR. *International conference on machine learning*. [S.l.], 2015. p. 2048–2057. Citado 2 vezes nas páginas 17 e 18.

YANG, L.-C.; YANG, C.-Y.; HSU, J. Y.-j. Object relation attention for image paragraph captioning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2021. v. 35, n. 4, p. 3136–3144. Citado na página 16.

YENDURI, G. et al. Gpt (generative pre-trained transformer)— a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE Access*, v. 12, p. 54608–54649, 2024. Citado na página 35.

ZARáNDY, et al. Overview of cnn research: 25 years history and the current trends. In: *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*. [S.l.: s.n.], 2015. p. 401–404. Citado na página 23.

ZHANG, C. et al. A survey on audio diffusion models: Text to speech synthesis and enhancement in generative ai. *arXiv preprint arXiv:2303.13336*, 2023. Citado na página 25.

ZHANG, W. et al. Image caption generation with adaptive transformer. In: IEEE. *2019 34rd youth academic annual conference of Chinese association of automation (YAC)*. [S.l.], 2019. p. 521–526. Citado 2 vezes nas páginas 41 e 45.

ZHENG, Q.; WANG, C.; WANG, D. Bypass network for semantics driven image paragraph captioning. *arXiv preprint arXiv:2206.10059*, 2022. Citado na página 15.

Apêndices

APÊNDICE A – Resultados de cada configuração

Tabela 5 – Resultados da Configuração 1.

Epoch	Replicação 1			Replicação 2			Replicação 3		
	Loss	Acurácia	Bleu	Loss	Acurácia	Bleu	Loss	Acurácia	Bleu
1	1.5578	0.7249	0.0233	1.6748	0.7108	0.0212	1.5642	0.7258	0.0237
2	1.4427	0.7358	0.0291	1.4763	0.7327	0.0326	1.4553	0.7350	0.0358
3	1.4039	0.7409	0.0237	1.4335	0.7374	0.0357	1.4369	0.7374	0.0280
4	1.4328	0.7386	0.0131	1.4430	0.7356	0.0322	1.4343	0.7369	0.0294
5	1.4139	0.7398	0.0181	1.4500	0.7352	0.0301	1.4503	0.7362	0.0308

Tabela 6 – Resultados da Configuração 2.

Epoch	Replicação 1			Replicação 2			Replicação 3		
	Loss	Acurácia	Bleu	Loss	Acurácia	Bleu	Loss	Acurácia	Bleu
1	1.6378	0.7150	0.0292	1.7474	0.7054	0.0249	1.6090	0.7210	0.0099
2	1.5184	0.7247	0.0352	1.5902	0.7188	0.0242	1.4971	0.7304	0.0273
3	1.4670	0.7306	0.0413	1.4709	0.7336	0.0228	1.4740	0.7329	0.0315
4	1.4879	0.7297	0.0263	1.4542	0.7352	0.0269	1.4539	0.7350	0.0259
5	1.4422	0.7352	0.0315	1.4322	0.7381	0.0310	1.4601	0.7343	0.0293
6	1.4357	0.7370	0.0300	1.5345	0.7270	0.0291	1.4353	0.7385	0.0251
7	1.4247	0.7376	0.0338	1.4903	0.7327	0.0183	1.4286	0.7402	0.0339
8	1.4544	0.7353	0.0367	1.4428	0.7388	0.0281	1.4520	0.7382	0.0252
9	1.4560	0.7347	0.0320	1.4632	0.7368	0.0295	1.4428	0.7393	0.0265
10	1.4610	0.7358	0.0384	1.4500	0.7386	0.0240	1.4272	0.7423	0.0235

Tabela 7 – Resultados da Configuração 3.

Epoch	Replicação 1			Replicação 2			Replicação 3		
	Loss	Acurácia	Bleu	Loss	Acurácia	Bleu	Loss	Acurácia	Bleu
1	1.7021	0.7101	0.0248	1.7058	0.7167	0.0168	1.6000	0.7255	0.0217
2	1.4475	0.7367	0.0201	1.4570	0.7357	0.0241	1.4146	0.7394	0.0197
3	1.3907	0.7435	0.0338	1.3884	0.7429	0.0100	1.3491	0.7474	0.0274
4	1.3597	0.7462	0.0525	1.3632	0.7448	0.0078	1.3315	0.7496	0.0257
5	1.3464	0.7495	0.0491	1.3580	0.7474	0.0286	1.3252	0.7512	0.0233

Tabela 8 – Resultados da Configuração 4.

Epoch	Replicação 1			Replicação 2			Replicação 3		
	Loss	Acurácia	Bleu	Loss	Acurácia	Bleu	Loss	Acurácia	Bleu
1	1.8167	0.7023	0.0050	1.6293	0.7216	0.0156	1.7076	0.7125	0.0203
2	1.4918	0.7326	0.0123	1.4318	0.7377	0.0294	1.4511	0.7361	0.0355
3	1.3991	0.7403	0.0222	1.3857	0.7417	0.0166	1.4114	0.7392	0.0280
4	1.3639	0.7459	0.0199	1.3550	0.7462	0.0269	1.3781	0.7432	0.0353
5	1.3503	0.7477	0.0206	1.3627	0.7470	0.0341	1.3721	0.7451	0.0328
6	1.3425	0.7497	0.0225	1.3476	0.7493	0.0215	1.3780	0.7457	0.0282
7	1.3472	0.7509	0.0377	1.3590	0.7485	0.0246	1.3922	0.7444	0.0300
8	1.3533	0.7508	0.0304	1.3746	0.7482	0.0252	1.3806	0.7469	0.0332
9	1.3628	0.7511	0.0224	1.3756	0.7490	0.0287	1.3772	0.7479	0.0354
10	1.3737	0.7510	0.0248	1.3802	0.7489	0.0296	1.3869	0.7486	0.0414

Tabela 9 – Cálculos da Configuração 1.

Replicação	**Loss**	**Acurácia**	**BLEU**
1	1.3985	0.7416	0.0239
2	1.4495	0.7354	0.0297
3	1.4503	0.7362	0.0308
Média	1.4327	0.7377	0.0281
Desvio Padrão	0.0297	0.0034	0.0037
IC 95%	(1.3590, 1.5065)	(0.7293, 0.7461)	(0.0190, 0.0373)

Tabela 10 – Cálculos da Configuração 2.

Replicação	**Loss**	**Acurácia**	**BLEU**
1	1.4654	0.7368	0.0192
2	1.4700	0.7370	0.0190
3	1.4213	0.7420	0.0169
Média	1.4522	0.7386	0.0184
Desvio Padrão	0.0269	0.0030	0.0013
IC 95%	(1.3855, 1.5190)	(0.7312, 0.7459)	(0.0152, 0.0215)

Tabela 11 – Cálculos da Configuração 3.

Replicação	**Loss**	**Acurácia**	**BLEU**
1	1.3246	0.7508	0.0363
2	1.3452	0.7474	0.0286
3	1.3394	0.7486	0.0362
Média	1.3364	0.7489	0.0337
Desvio Padrão	0.0106	0.0017	0.0044
IC 95%	(1.3100, 1.3627)	(0.7446, 0.7532)	(0.0227, 0.0446)

Tabela 12 – Cálculos da Configuração 4.

Replicação	**Loss**	**Acurácia**	**BLEU**
1	1.3672	0.7503	0.0277
2	1.3493	0.7520	0.0342
3	1.4125	0.7459	0.0199
Média	1.3763	0.7494	0.0272
Desvio Padrão	0.0326	0.0031	0.0071
IC 95%	(1.2954, 1.4572)	(0.7416, 0.7572)	(0.0095, 0.0450)

APÊNDICE B – Exemplos de legendas geradas pelo modelo

TRUE: [START] several older people with red aprons work at a book fair [END]

PRED: [START] a group of people are working together in a room while some their other [END]

(a) Imagem do *dataset*. (b) *TRUE*: Legenda do *dataset*, PRED: legenda gerada pelo modelo.

Figura 16 – Exemplo de legenda gerada no modelo final.

TRUE: [START] a couple kiss on the street at night in a busy spanish speaking city [END]

PRED: [START] a city street is busy focal on night [END]

(a) Imagem do *dataset*. (b) *TRUE*: Legenda do *dataset*, PRED: legenda gerada pelo modelo.

Figura 17 – Exemplo de legenda gerada no modelo final.

TRUE: [START] two chefs standing in front of a large boiling pot offer a spoonful of their dish [END]

PRED: [START] an people are cooking foods in front of a large fire bar in an outdoor environment area [END]

(a) Imagem do *dataset*. (b) *TRUE*: Legenda do *dataset*, PRED: legenda gerada pelo modelo.

Figura 18 – Exemplo de legenda gerada no modelo final.

(a) Imagem do *dataset*.

TRUE: [START] a woman standing in front of trees and smiling [END]

PRED: [START] the asian lady in a gray shirt on a high mountain [END]

(b) *TRUE*: Legenda do *dataset*, PRED: legenda gerada pelo modelo.

Figura 19 – Exemplo de legenda gerada no modelo final.