Revista ARACÊ
ISSN: 2358-2472

APRENDIZADO PROFUNDO BASEADO EM CNN-LSTM PARA LEGENDAGEM AUTOMÁTICA DE IMAGENS

https://doi.org/10.56238/arev6n3-145

Data de submissão: 13/10/2024 **Data de publicação:** 13/11/2024

Maria Vitória Sousa Ribeiro
Graduada em Gestão da Tecnologia da Informação (IFBAIANO)
Instituto Federal Baiano (IFBAIANO)
E-mail: vicksousa503@gmail.com

Tiago do Carmo Nogueira
Doutor em Engenharia Elétrica e de Computação (UFG)
Instituto Federal Baiano (IFBAIANO)
E-mail: tiago.nogueira@ifbaiano.edu.br

Gelson da Cruz Junior
Doutor em Engenharia Elétrica (Unicamp)
Universidade Federal de Goiás (UFG)
E-mail: gcruzjr@ufg.br

Cássio Dener Noronha Vinhal
Doutor em Engenharia Elétrica (Unicamp)
Universidade Federal de Goiás (UFG)
E-mail: vinhal@ufg.br

Matheus Rudolfo Diedrich Ullmann
Doutorando em Engenharia Elétrica e de Computação (UFG)
Instituto Federal da Bahia (IFBA)
E-mail: matheusullmann@ifba.edu.br

Deller James Ferreira
Doutora em Educação (UNB)
Universidade Federal de Goiás (UFG)
E-mail: deller@inf.ufg.br

Caio Henrique Rodrigues Carvalho
Mestre em Engenharia Elétrica (UFPI)
Instituto Federal Baiano (IFBAIANO)
E-mail: caio.carvalho@ifbaiano.edu.br

Danyele de Oliveira Santana
Mestranda em Ciência da Computação (UEFS)
Instituto Federal Baiano (IFBAIANO)
E-mail: danyele.santana@ifbaiano.edu.br

Revista ARACÊ
ISSN: 2358-2472

RESUMO
A evolução da Visão Computacional e Aprendizado de Máquina permite que técnicas de descrição de imagens em linguagem natural sejam mais eficientes e precisas, por meio de redes neurais profundas. Este estudo utilizou uma estrutura codificador-decodificador para identificação e legendagem de objetos, através de uma imagem de entrada. O modelo proposto utilizou as arquiteturas VGG16 e Inception-V3 como codificadores e LSTM como decodificador. Para a realização dos experimentos, foi utilizado o conjunto de dados Flickr8k, possuindo 8.000 imagens. O modelo foi avaliado pelas métricas Bleu, Meteor, CIDEr e Rouge. Alcançando 58,40% de precisão conforme a métrica Bleu, garantindo assim descrições compreensíveis para o ser humano.

Palavras-chave: Aprendizado de Máquina. Aprendizado Profundo. Redes Neurais Convolucionais. Memória Longa de Curto Prazo. Legendagem de Imagens.

Revista ARACÊ
ISSN: 2358-2472

1 INTRODUÇÃO

A legendagem automática informa em textos descritivos ou em única palavra, a classificação do objeto analisado na imagem, tendo como principal objetivo a análise de imagens para descrição do seu conteúdo (RINALDI; RUSSO; TOMMASINO, 2023; NOGUEIRA, 2020). Com o avanço da Inteligência Artificial (IA), modelos de Aprendizado Profundo (do inglês *deep learning* - DL) têm contribuído para que técnicas de legendagem automática, possam ser mais precisas e tenham maior desempenho (SINGH; GUPTA, 2023).

O DL é uma subárea da IA, sendo uma das técnicas de Aprendizado de Máquina (do inglês *machine learning* - ML). O ML possui a capacidade de processar uma variedade de dados, encontrando padrões e podendo ser programado para aprender modelos, além de utilizar as redes neurais profundas para a análise (SEHGAL; MANDAN, 2022).

Recentemente, estudos no campo de redes neurais profundas estão sendo aplicados para a análise de imagens, Visão Computacional (VC), Processamento de Linguagem Natural (PLN) e legendagem (RINALDI; RUSSO; TOMMASINO, 2023; AL-MALLA; JAFAR; GHNEIM, 2022; JAIN; DOSHI; DWIVEDI, 2023). Segundo Sehgal e Mandan (2022), a Visão Computacional é um campo da Ciência da Computação que estuda a capacidade das máquinas em extraírem elementos de uma imagem ou vídeo. Através da VC é possível realizar o reconhecimento de objetos em imagens, reconhecimento facial e análise de movimento, treinando a máquina para compreender o que está sendo visto de forma semelhante aos seres humanos.

A quantidade de imagens disponíveis na Internet faz com que exija a necessidade de identificá-las e descrevê-las, como destacado por Al-Malla, Jafar e Ghneim (2022). Apesar da habilidade dos seres humanos de identificar objetos, as máquinas possuem dificuldade nessa classificação (NOGUEIRA, 2020). Sendo assim, as máquinas podem ser treinadas por diferentes tipos de conjuntos de dados para adquirir essa habilidade de identificar os objetos e relacioná-los na imagem (VERMA et al., 2024).

Estudos recentes aplicaram a estrutura de codificador-decodificador para identificar a relação dos objetos na imagem. Essa estrutura utiliza arquiteturas de redes neurais profundas para codificar a imagem e a descrevê-la, sendo utilizados principalmente Redes Neurais Convolucionais (do inglês *convolutional neural networks* - CNN's) e Redes Neurais Recorrentes (do inglês *recurrent neural networks* - RNN's) para uma melhor qualidade das sentenças (NOGUEIRA et al., 2023). A CNN é o codificador, sendo utilizada para analisar as imagens, classificá-las e reconhecer os seus padrões. Enquanto, a RNN é o decodificador, sendo aplicada na legendagem das imagens (AL-MALLA; JAFAR; GHNEIM, 2022; NOGUEIRA, 2020).

Revista ARACÊ
ISSN: 2358-2472

No entanto, autores que utilizaram a RNN tradicional localizaram limitações em suas aplicações, sendo elas: tempo de treinamento longo, explosão ou desaparecimento de gradientes (NOGUEIRA et al., 2020; NOGUEIRA et al., 2023). Sendo assim, pesquisadores obtiveram melhores resultados, aplicando modelos de Unidade Recorrente Fechadas (GRU) ou Memória Curta de Longo-Prazo (do inglês *Long Short-Term Memory - LSTM*) (SINGH; SINGH; ANANDHAN, 2022; BHALEKAR; BEDEKAR, 2022).

A legendagem automática de objetos em imagens é uma técnica que possui diversas aplicações, podendo ser utilizada no auxílio de pessoas com deficiência visual, recomendações em mídias sociais, indexação de imagens, pesquisa visual, sistemas de segurança, entre outras aplicações de processamento de linguagem natural l (VERMA et al., 2024). Portanto, este estudo tem como principal objetivo o desenvolvimento de um modelo que realize a legendagem automática de imagens por meio do aprendizado profundo.

O modelo proposto é baseado na arquitetura codificador-decodificador para a legendagem dos objetos em imagens. Assim, para a extração das características das imagens, será utilizado o CNN como codificador. E para gerar a descrição dos objetos, será utilizado o LSTM como decodificador. O modelo proposto fornece resultados de última geração para legendagem de imagens em conjuntos de dados públicos, como Flickr8k.

Os principais objetos desta pesquisa, são: identificar tecnologias e modelos na literatura para legendagem de imagens; analisar de que forma a IA e Redes Neurais Profundas têm contribuído para a legendagem; e propor um modelo em Aprendizado Profundo com maior precisão e eficiência para legendagem automática de imagens.

O resumo de nossas contribuições de pesquisa é o seguinte: propor VGG16 e Inception23 como codificadores e LSTM como decodificador para legendagem automática de objetos em imagens; além de relatar resultados experimentais usando as métricas mais utilizadas, como METEOR, BLEU, CIDEr e ROUGE no conjunto de dados Flickr8k, contribuindo para as abordagens de última geração. Além disso, apresentar os resultados do modelo proposto, os quais foram validados em imagens ao vivo selecionadas aleatoriamente.

Este artigo está organizado da seguinte forma: a seção 2 apresenta os trabalhos relacionados sobre CNN e RNN na detecção de objetos em imagens e legendagem automática; a seção 3 aborda os procedimentos e instrumentos metodológicos utilizados; a seção 4 aborda os experimentos realizados; a seção 5 apresenta os resultados; a seção 6 discorre as discussões e limitações do modelo proposto; a seção 7 aborda a conclusão do estudo e os trabalhos futuros.

Revista ARACÊ
ISSN: 2358-2472

2 TRABALHOS RELACIONADOS

A legendagem automática é um campo desafiador e com constante evolução na Visão Computacional, sendo utilizado arquitetura de aprendizado profundo para a análise de imagens e geração de legendas em linguagem natural (NOGUEIRA, 2020). As máquinas ainda possuem o desafio de identificar o objeto, descrevê-lo e analisar sua relação com o ambiente. Sendo necessário, o presente estudo e trabalhos relacionados em Aprendizado Profundo.

Segundo Al-Malla, Jafar e Ghneim (2022), a arquitetura codificador-decodificador é a mais utilizada na técnica de aprendizagem profunda para legendagem automática de imagens. A estrutura codificador-decodificador pode ser dividida em duas categorias, sendo elas: abordagens de baixo para cima e de cima para baixo o (NOGUEIRA et al., 2020; NOGUEIRA et al., 2023).

Com objetivo de aumentar a qualidade das legendas, é utilizado recursos de objetos e recursos convolucionais de um modelo. A CNN atua como codificador a partir dos dados de entrada, realizando a extração dos padrões das imagens e das tags dos objetos. Ela é subdividida em camadas, sendo elas: camada de entrada, camada de convolução, camadas de *pooling*, camada totalmente conectada e camada de saída (FARABY et al., 2020; SINGH; VIJ, 2022). Enquanto, a RNN atua como decodificador, realizando a legendagem dessas imagens (NOGUEIRA, 2020).

Sehgal e Mandan (2022) propuseram um modelo de legendagem de fotografias utilizando redes neurais, além do modelo de LSTM, para uma maior precisão nas legendas automáticas. O LSTM é um tipo de arquitetura de RNN com a capacidade de aprender com dados sequenciais. Em seus resultados, foi constatado uma melhor porcentagem de precisão nas legendas geradas utilizando esse modelo. O LSTM analisa as características da imagem e gera um texto ou descrição que a corresponde. É realizada a transcrição da imagem decodificando os vetores de entrada a (NOGUEIRA, 2020; GOEL et al., 2023).

Nesse contexto, Thangavel et al. (2023) explorou em sua pesquisa a utilização da camada de codificação, empregando a arquitetura da Máscara de Redes Neurais Recorrentes (Faster R-CNN), enquanto a camada de decodificação adotou o uso do mecanismo de atenção com LSTM, alcançando um desempenho de 25% superior aos métodos analisados na pesquisa. Apesar do LSTM ter um maior desempenho na precisão das legendas das imagens, para implementá-lo é necessário um tempo de treinamento mais longo (NOGUEIRA, 2020).

Rohitharun, Reddy e Sujana (2022) propuseram em seu estudo um modelo one-to-one baseado na ResNet-101 como codificador e na LSTM como decodificador. Através dela, os recursos das imagens são extraídos no formato de vetores. A ResNet-101, por possuir menos parâmetros, permite um treinamento mais profundo e a torna mais eficiente em termos de uso de recursos computacionais.

Revista ARACÊ
ISSN: 2358-2472

Enquanto isso, Nogueira et al. (2020) para solucionar os problemas de RNN tradicionais no desaparecimento de gradientes, utilizaram em seu estudo o modelo R-GRU para uma melhor potência e resultado. A partir do banco de dados Flickr30k, obtiveram 69,40% de precisão na métrica Bleu. Esta abordagem utiliza a CNN para analisar a imagem em pequenas partes e extrair as características visuais mais importantes da imagem de entrada. A GRU multimodal é um tipo de RNN que recebe os dados da CNN e outros tipos de dados, tendo como objetivo gerar a sentença por referência.

Modelos baseados em CNN's demandam uma grande quantidade de dados para análise e treinamento, evitando problemas como a falta de precisão nas legendas (AYESHA et al., 2021; WAHEED et al., 2023). O desempenho da geração de imagens sofre uma variação de acordo com as múltiplas arquiteturas de CNN's utilizadas (KATIYAR; BORGOHAIN, 2021).

Nesse sentido, Verma et al. (2024) realizaram em seu estudo um levantamento de arquiteturas de CNN pré-treinadas para discussão e comparação do melhor modelo. O primeiro modelo analisado foi o VGG16, que alcançou mais de 90% de precisão. O segundo modelo examinado foi o Inception V3, que demonstrou uma precisão superior à rede analisada anteriormente. Por fim, o último modelo avaliado foi o ResNet-50, considerado o mais recente e eficaz para redes neurais profundas em comparação com os outros modelos. Ao realizar a comparação entre as arquiteturas levantadas, os autores propuseram em seu estudo o modelo VGG16 Hybrid Places 1365 como codificador, sendo utilizado para fornecer resultados específicos de objeto e cena. O modelo proposto alcançou 66,66% de precisão na métrica de avaliação Bleu e 50,60% na Meteor.

Seshadri, Srikanth e Belov (2020) desenvolveram um sistema Web para receber uma imagem como entrada e gerar a legendagem que a corresponde, obtendo 13% e 14% de precisão conforme as métricas Bleu e Meteor. Foi utilizado um modelo codificador CNN e um decodificador LSTM bidirecional. As abordagens de codificador-decodificador podem ser aplicadas de duas maneiras, sendo elas: arquitetura de injeção e arquitetura de mesclagem. Na arquitetura de injeção, a imagem é codificada em um conjunto fixo de números. Após isso, é realizada a combinação com cada palavra da descrição do texto, o decodificador utiliza essa informação para prever qual será a próxima palavra na descrição. Enquanto, a arquitetura de mesclagem são extraídas a codificação da imagem e a codificação da descrição. O decodificador utiliza de ambas codificações para gerar a sequência da descrição.

Sasibhooshan, Kumaraswamy e Sasidharan (2023 utilizaram em seu modelo a rede neural convolucional baseada na decomposição Wavelet discreta (WCNN). O uso da WCNN possibilita a extração dos elementos da imagem, bem como a localização aproximada dos objetos nela, juntamente com informações de frequência. Esse modelo permite a identificação das partes mais importantes de

Revista ARACÊ
ISSN: 2358-2472

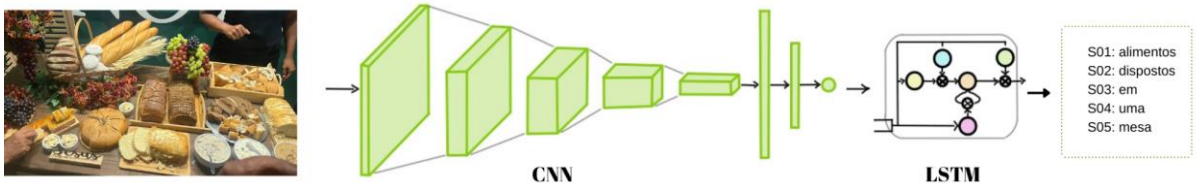
uma imagem e demonstra maior precisão em imagens altamente complexas, atingindo 38,20% na métrica Bleu-4.

Diversos trabalhos relacionados à literatura desenvolvem seus modelos seguindo a estrutura codificador-decodificador. A maioria desses modelos é voltada para a descrição de imagens utilizando arquiteturas como CNN, RNN, LSTM, GRU, entre outras. Entretanto, alguns desses modelos quando desenvolvidos apresentaram problemas, como a imprecisão das legendas ou perda de informações. Apesar de estudos recentes utilizarem em sua grande maioria a arquitetura de CNN-RNN para legendagem de imagens, alguns autores agregaram outros métodos e técnicas auxiliares para um melhor desempenho (CHEN et al., 2020; HOSSEN et al., 2024; FERREIRA et al., 2022; SCOPARO; SERAPIÃO, 2019; PADATE et al., 2023). Na próxima seção, será proposto um modelo codificador-decodificador para amenizar esses problemas.

3 PROPOSTA METODOLÓGICA

Este trabalho propõe-se um modelo baseado na arquitetura codificador-decodificador, utilizando CNN-RNN para a legendagem dos objetos em imagens. Assim, para a extração das características das imagens, foi utilizado o VGG16 e Inception-V3 como codificador. E para gerar a descrição dos objetos, foi utilizado o LSTM como decodificador, conforme mostrado na Figura 1.

Figura 1. Modelo proposto baseado em CNN-LSTM.



A escolha da CNN justifica-se por ser uma rede com capacidade de aprender e extrair padrões complexos das imagens, além da possibilidade de ser pré-treinada com conjuntos de dados para reduzir o tempo de treinamento e obter maior precisão na análise das imagens. Além disso, a CNN tem a habilidade de analisar imagens com tamanhos variados e diferenciar as imagens de treinamento das imagens de entrada.

O uso da VGG16 explica-se por ser um modelo convolucional simples de ser implementado e eficaz nas tarefas de visão computacional, como detecção e descrição de imagens (YESHASVI; SUBETHA, 2022). Já o Inception-V3 foi utilizado por ter baixo custo computacional em comparação dos outros modelos, além da disponibilidade de implementações pré-treinadas e comprovação do seu desempenho (NOGUEIRA et al., 2020).

Revista ARACÊ
ISSN: 2358-2472

Enquanto, a escolha do LSTM justifica-se por ter uma memória de longo prazo, além da capacidade de manter informações e maior armazenamento (PA; NWE et al., 2020). Além do mais, possui melhores resultados de precisão nas legendas geradas nos estudos desenvolvidos anteriormente (ROHITHARUN; REDDY; SUJANA, 2022; YESHASVI; SUBETHA, 2022; AOTE et al., 2022).

A arquitetura do modelo proposto pode ser dividida principalmente em três módulos. O primeiro módulo é a extração de recursos utilizando CNN, o segundo são as arquiteturas propostas de VGG16 e Inception-V3 (utilizadas como codificador) e o terceiro é a geração de legendas utilizando o LSTM.

Esta seção descreve a metodologia adotada no presente estudo. Nas subseções seguintes (3.1, 3.2 e 3.3), serão abordados a descrição do modelo proposto.

3.1 EXTRAÇÃO DE RECURSOS

As CNN's possuem a capacidade de reconhecer e identificar objetos em uma imagem, através do processamento de uma imagem de entrada e classificação conforme os modelos predefinidos (NOGUEIRA, 2020) A CNN é subdividida em três principais camadas, que estão presentes em todas as CNN's, sendo elas: as camadas de convoluções, camadas de *pooling* e as camadas profundas totalmente conectadas (do inglês *fully connected* - FC) (VERMA et al., 2024).

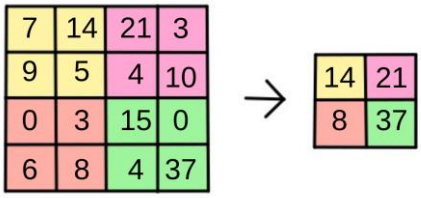
A camada de convolução é a primeira camada a extrair os recursos da imagem, tendo como principal objetivo a detecção de características da imagem, como bordas, padrões e objetos. É aplicado um filtro (*kernel*) em pequenas partes dela, em níveis de largura, altura e profundidade, gerando mapas de características para cada *pixel* (NOGUEIRA, 2020). Esses mapas refletem a imagem em diferentes níveis de abstração, quanto mais filtros convolucionais aplicados, maior é o custo de processamento e memória.

Enquanto isso, a camada de *Pooling* tem como objetivo a redução do tamanho da imagem de entrada, diminuindo o custo computacional de uma rede neural quando aplicada após a camada de convolução. Em uma CNN, a segunda camada executada é a de *pooling*, podendo ser aplicado três técnicas para compactar os mapas de características, sendo elas: *pooling* médio, *pooling* máximo e *pooling* L2.

O *pooling* máximo é a forma mais utilizada para executar o *pooling*. Nesta técnica, é aplicado filtros de tamanho $2x2$ em imagens $4x4$, isso permite que seja utilizado somente 25% das ativações, conforme demonstrado na figura 2.

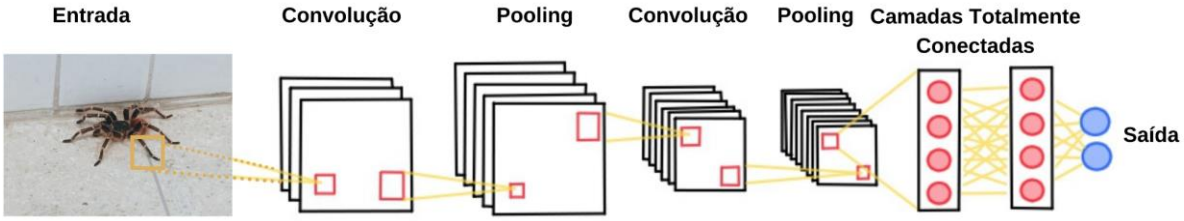
Revista ARACÊ
ISSN: 2358-2472

Figura 2. Exemplo do funcionamento do max-pooling com filtro 2x2 em uma imagem 4x4.



Enfim, a camada totalmente conectada transforma uma matriz de dados em um único vetor. Conforme demonstrado na Figura 3, a CNN inicia com as camadas de convolução e *pooling*, dividindo as imagens em recursos e analisando-as separadamente. Em seguida, esses recursos são aplicados na camada *FC*, onde ocorre a classificação do vetor (NOGUEIRA, 2020).

Figura 3. Exemplo completo de uma CNN para a classificação de objetos em imagens, utilizando as camadas básicas de convolução, pooling e totalmente conectadas.



Para reduzir o tempo de treinamento da CNN quando treinada do zero, foram utilizados os modelos VGG16 e Inception-V3 neste estudo.

3.2 ARQUITETURA DO MODELO PROPOSTO

Conforme mencionado na subseção anterior, o modelo proposto é baseado nas arquiteturas VGG16 e Inception-V3 para extrair os recursos das imagens nos processos de treinamento e validação, ambas são redes neurais convolucionais.

De acordo com a Tabela 1, o Inception-V3 tradicionalmente tem o tamanho de entrada $299x299x3$ ($299x299$ referente a entrada e 3 referente os filtros) e produz um *pool* de saída de $8x8x2048$ ($8x8$ referente a entrada e 2048 referente os filtros), isso permite a redução da quantidade de parâmetros e torna o modelo mais eficiente.

Tabela 1. Arquitetura Inception-V3.

Camada	Entrada	Filtros
Conv	299 x 299	3
Conv	149 x 149	32

Revista ARACÊ
ISSN: 2358-2472

Conv padded	147 x 147	32
Pool	147 x 147	64
Conv	73 x 73	64
Conv	71 x 71	80
Conv	35 x 35	192
3 Inception	17 x 17	288
5 Inception	8 x 8	768
2 Inception	8 x 8	1280
Pool	8 x 8	2048
Linear	1 x 1	2048
Softmax	1 x 1	1000
-	-	-

Enquanto, o VGG16 possui 16 camadas profundas, permitindo a análise de imagens mais complexas (PA; NWE et al., 2020). Além disso, possui uma entrada de $224x224x3$ ($224x224$ referente a entrada e 3 referente os filtros). De acordo com a Tabela 2, a VGG16 contém cinco camadas de convolução *(conv)*, cinco camadas de *pool* (*maxpoo*l) e três camadas totalmente conectadas (*fc*). A última camada descrita por *Softmax* é responsável pela classificação por probabilidade do objeto na imagem.

Tabela 2. Arquitetura VGG16.

Camada	Entrada	Filtros
Input	224 x 224	3
2 conv	224 x 224	64
Maxpool	112 x 112	128
2 conv	112 x 112	128
Maxpool	56 x 56	256
3 conv	56 x 56	256
Maxpool	28 x 28	512
3 conv	28 x 28	512
Maxpool	14 x 14	512

Revista ARACÊ
ISSN: 2358-2472

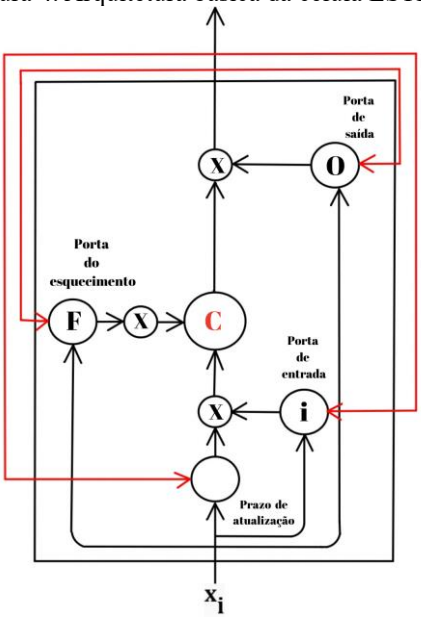
3 conv	14 *x* 14	512
Maxpool	7 *x* 7	512
2 fc	1 *x* 1	4096
1 fc	1 *x* 1	1000
Softmax	Classifier	-

3.3 GERAÇÃO DE LEGENDAS

Neste estudo foi utilizado a LSTM, que é um tipo de RNN, para uma melhor precisão das descrições e garantia de uma memória de sentenças a longo prazo. Através das células de memória, o LSTM tem a capacidade de excluir ou arquivar uma determinada informação (ROHITHARUN; REDDY; SUJANA, 2022), além de reduzir o problema da explosão de gradientes do RNN tradicional (KHANT et al., 2021).

Na Figura 4, são demonstradas três portas da arquitetura da célula LSTM, sendo elas: porta de entrada (*Input Gate*), porta de saída (*Output Gate*) e porta do esquecimento (*Forget Gate*), tendo como controle a célula de memória C (localizada no centro da imagem). Essas portas são responsáveis por realizar uma determinada operação, tendo todas as informações relevantes armazenadas em sua célula de memória. A porta de esquecimento determina se o valor atual da célula deve ser descartado ou mantido na memória, a porta de entrada define se o novo valor da célula deve ser lido ou descartado e a porta de saída define se é necessário gerar um novo valor para a célula (SURESH; JARAPALA; SUDEEP, 2022).

Figura 4. Arquitetura básica da célula LSTM.



Revista ARACÊ
ISSN: 2358-2472

A etapa i, recebe a entrada de diferentes fontes: o estado oculto passado (hi−1); a entrada atual (Xi); e o estado anterior da célula de memória (Ci−1). No passo de tempo t, os valores de porta atualizados para as entradas fornecidas X e u, hi−1 e Ci−1, são:

$$I_i = \sigma(WM_{X_I} X_i + WM_{HI} h_{i-1} + BV_I) \quad (1)$$

$$F_i = \sigma(WM_{X_F} X_i + WM_{HF} h_{i-1} + BV_F) \quad (2)$$

$$O_i = \sigma(WM_{X_O} X_i + WM_{HO} h_{i-1} + BV_O) \quad (3)$$

$$G_i = \phi(WM_{X_C} X_i + WM_{HC} h_{i-1} + BV_C) \quad (4)$$

$$C_i = F_i * C_{i-1} + I_i * G_i \quad (5)$$

$$h_i = O_i * \phi(C_i) \quad (6)$$

onde Xi representa as entradas da porta de entrada, XF da porta de esquecimento, XO da porta de saída e XC da célula de memória, BV representa vetores e WM representa métricas de peso. Além de ϕ é a tangente hiperbólica, podendo ser calculada:

$$\phi(X) = \frac{expo(X) - expo(-X)}{expo(X) + expo(-X)} \quad (7)$$

Além disso, a função de ativação sigmoide é ρ, sendo calculada por:

$$\rho(X) = \frac{1}{1 + expo(-X)} \quad (8)$$

4 EXPERIMENTOS

Esta seção apresenta os conjuntos de dados utilizados, bem como as configurações dos experimentos, os parâmetros utilizados e as métricas de avaliação aplicadas para análise dos resultados.

Revista ARACÊ
ISSN: 2358-2472

4.1 CONJUNTOS DE DADOS

Durante a análise da literatura, diversos conjuntos de dados foram utilizados para o treinamento, validação e teste das legendas geradas s (ROHITHARUN; REDDY; SUJANA, 2022). Conforme Tabela 3, os autores utilizaram os conjuntos de dados Flickr8k (KHANT et al., 2021; INDUMATHI et al., 2023), Flickr30k (NOGUEIRA et al., 2023) e MS-COCO (KESKIN et al., 2021) para o experimento, por possuírem uma variedade de imagens e mais de uma legenda descritiva.

Tabela 2. Conjunto de dados presentes na análise da literatura.

Autores	Banco de Dados
Khant et al. (2021)	Flickr8k
Indumathi et al. (2023)	Flickr8k
Keskin et al. (2021)	MS COCO
Nogueira et al. (2023)	Flickr30k e MS COCO
Al-Malla, Jafar e Ghneim (2022)	Flickr30k e MS COCO

Para este estudo de legendagem automática de imagens, foi utilizado o banco de dados público do Flickr8k. Para cada imagem do banco foram associadas cinco legendas diferentes, geradas por seres humanos no idioma inglês, conforme demonstrado na Figura 5. De acordo com a tabela 4, o conjunto de dados do Flickr8k possui 8.000 imagens, sendo 6.000 imagens para treinamento, 1.000 para validação e 1.000 para teste.

Figura 5. Imagem de amostra com as legendas correspondentes.

1. A man in a hat is displaying pictures next to a skier in a blue hat .
2. A man skis past another man displaying paintings in the snow .
3. A person wearing skis looking at framed pictures set up in the snow .
4. A skier looks at framed pictures in the snow next to trees .
5. Man on skis looking at artwork for sale in the snow .

Tabela 4. Números de amostras no conjunto de dados Flickr8K.

Conjunto de dados	Flickr8K
Total	8.000
Treinamento	6.000

Revista ARACÊ
ISSN: 2358-2472

Teste	1.000
Validação	1.000

4.2 CONFIGURAÇÕES DOS EXPERIMENTOS

A primeira etapa dos experimentos, é o pré-processamento das imagens e das descrições. Sendo necessário para diminuir o tempo de treinamento de um modelo sem comprometer a eficácia das legendas geradas (ANSARI; SRIVASTAVA, 2024). Para o pré-processamento das descrições, são removidos os espaços e a pontuação entre as palavras, além disso todas as letras são convertidas para minúsculas. As legendas do banco são separadas por vírgulas e armazenadas em uma única lista, além de ser adicionado *tokens* de início e fim para cada legenda, permitindo assim um melhor treinamento. Ao final, foram contabilizadas um total de 40.455 legendas, sendo 8.768 o tamanho do vocabulário e 34 o comprimento máximo da legenda.

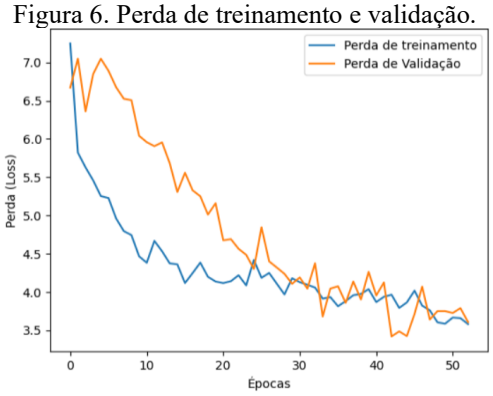
As imagens de entrada possuem a resolução de $256x500$ a $500x500$. Em seu pré-processamento, as imagens foram redimensionadas para $224x224$ pixels quando testado o modelo do VGG16 e para $299x299$ pixels quando testado o modelo Inception-V3.

Os dados de treinamento são utilizados para o modelo aprender os padrões e ajustar os parâmetros, enquanto os dados de validação servem para verificar o desempenho do modelo ainda no treinamento. Já os dados de teste são apresentados após a criação do modelo, para validação da eficácia do algoritmo. Essa divisão é importante para que o algoritmo aprenda e preveja a legendagem em imagens novas.

A técnica de parada antecipada (do inglês *Early Stopping*)foi utilizada para monitorar o desempenho a cada época, a fim de evitar o *overfitting* - dificuldade do modelo na predição quando testado em dados novos (KAVITHA et al., 2022). No modelo proposto, foram definidas 70 épocas e um valor de *"patience"* de 10 épocas, ou seja, se o desempenho não melhorasse ou a perda de validação aumentasse por 10 épocas consecutivas, o treinamento seria interrompido. Conforme demonstrado na Figura 6, o treinamento do algoritmo foi interrompido quando houve o *overfitting*, assegurando que o modelo realizasse a legendagem automática em imagens de teste de maneira compreensível aos seres humanos.

Revista ARACÊ
ISSN: 2358-2472

Figura 6. Perda de treinamento e validação.



Para diminuir a explosão de gradientes e possíveis perdas, foi utilizado o otimizador Adam. Além disso, o Adam requer pouco custo computacional, sendo eficiente em maiores conjuntos de dados e menor período de treinamento. Além do mais, foi utilizada a função de ativação reLU, para melhorar a constância dos gradientes durante o treinamento.

Portanto, definimos o codificador para executar 70 épocas e *batch_size* de 512, treinando o conjunto completo de imagens disponíveis para treinamento e validação no banco de dados. Durante o processo de validação do modelo foram utilizados os parâmetros, conforme a Tabela 5.

Tabela 5. Parâmetros utilizados durante os experimentos.

Parâmetros	Valor
Épocas	70
Batch_size	512
Otimizador	Adam
Taxa de Aprendizagem	1 x 10
Dropout	0,1

Os experimentos foram realizados no ambiente de desenvolvimento do Google Collab. Para desenvolver o modelo proposto, usamos algumas bibliotecas como Tensorflow, nltk, Keras, os, pickle, numpy, entre outros.

4.3 MÉTRICAS DE AVALIAÇÃO

Para medir o desempenho do modelo, é essencial aplicar métricas de avaliação para garantir a qualidade das legendas geradas. Isso é feito através da comparação entre as legendas de referência e o seu conteúdo, a correção gramatical da legendagem, entre outros parâmetros (VERMA et al., 2024; AL-MALLA; JAFAR; GHNEIM, 2022; NOGUEIRA et al., 2020; NOGUEIRA et al., 2023). Por esse

Revista ARACÊ
ISSN: 2358-2472

motivo, avaliamos o modelo deste estudo usando as métricas de avaliação METEOR, BLEU, CIDEr e ROUGE.

O método METEOR é uma métrica de avaliação utilizada para medir a qualidade das legendas geradas. Essa métrica realiza o alinhamento entre as legendas automáticas e as frases de referência, permitindo a análise do grau de similaridade entre elas (NOGUEIRA et al., 2020). Além disso, o METEOR baseia-se no conceito de correspondência de unigramas, comparando quais palavras estão presentes individualmente nas legendas geradas pelo modelo e as palavras da legenda de referência. A métrica Meteor pode ser calculada da seguinte forma:

$$\mathbf{METEOR} = \left(\frac{10PR}{R+9P}\right)(1-P_m), \tag{9}$$

onde P é definido pela precisão dos unigramas e R pela recuperação. A pena condicional é calculado por:

$$\mathbf{P_m} = 0.5\left(\frac{C}{M_u}\right), \tag{10}$$

onde Mu é definido pela contagem de unigramas correspondentes, enquanto C indica o número mínimo de sentenças requeridas para que haja correspondência com os unigramas nas traduções de referência.

A METEOR foi projetada para estar mais alinhada com as legendas geradas pelo ser humano, focando na recordação e na precisão das descrições. Enquanto, a métrica de avaliação BLEU é implementada analisando a qualidade das legendas geradas pelas legendas de referência dos seres humanos (VERMA et al., 2024). Essa avaliação é calculada através da precisão de n-gramas entre as frases geradas e as frases de referência. Possui a limitação de não analisar o significado das frases e sua estrutura, analisando somente as n-gramas. Existem diversos tipos de cálculos para essa métrica, o principal cálculo sendo:

$$\mathbf{BLEU} = B_p \cdot exp\left(\sum_{n=1}^{N} w_n log p_n\right) \tag{11}$$

Revista ARACÊ
ISSN: 2358-2472

onde o Bp representa o valor da brevidade da frase gerada pelo modelo, ou seja, do quão concisa é a legenda. A pn é a média geométrica da precisão modificada de n-gramas. E o n é o comprimento máximo dos n-gramas da frase candidata (NOGUEIRA et al., 2023). Assim, a brevidade é calculada por:

$$\mathbf{B_p} = \begin{cases} 1 & if c > r \\ e^{(1-r/c)} & if c \leq r \end{cases} \qquad (12)$$

onde o c é o comprimento da legenda gerada pelo modelo e r é o comprimento do conjunto de dados de referência. A pontuação do BLEU é classificada como excelente, quando superior a 0,5. Enquanto, a legenda inferior a 0,15 indica que precisa de melhorias.

A métrica CIDEr tem como principal característica a saliência e a análise gramatical (AL-MALLA; JAFAR; GHNEIM, 2022). Ela consiste em três processos, sendo eles: coleta de diferentes legendas geradas por humanos, medição de consenso para identificar palavras similares nas legendas, automação do consenso e dois conjuntos de dados que descrevem as imagens s (NOGUEIRA et al., 2023). O cálculo para essa métrica de avaliação é:

$$\mathbf{CIDEr_n(c_i, S_i)} = \frac{1}{m}\sum_j \frac{g^n(c_i \cdot g^n(s_{ij}))}{\|g^n(c_i\| \, \|g^n(s_{ij})\|} \qquad (13)$$

$$\mathbf{CIDEr(c_i, S_i)} = \sum_{n=1}^{N} w_n CIDEr_n(c_i, S_i)) \qquad (14)$$

Por fim, a métrica ROUGE avalia a qualidade da legenda ao comparar a quantidade de palavras na descrição gerada pelo modelo com a descrição humana usada como referência. A métrica Rouge é calculada da seguinte forma:

$$\mathbf{ROUGE} = \frac{\sum_{S \in S_H} \sum_{g_n \in S} C_m(g_n)}{\sum_{S \in S_H} \sum_{g_n \in S} C(g_n)}, \qquad (15)$$

Revista ARACÊ
ISSN: 2358-2472

onde n representa o comprimento de n-grama; e Cm e gn são o número máximo de n-gramas que ocorrem em um resumo do nosso modelo e um conjunto de resumos de referência

Para verificação da eficácia do modelo, utilizamos a API de avaliação do MS-COCO que permite a classificação das legendas geradas. A API utiliza as métricas de avaliação BLEU, METEOR, CIDEr e ROUGE, assegurando uma análise do desempenho precisa dos resultados.

5 RESULTADOS

O principal objetivo deste estudo é alcançar resultados compreensíveis na legendagem após o treinamento, correlacionando os elementos da imagem de entrada com a legenda gerada. Esta seção apresenta as pontuações das métricas de avaliação mencionadas na subseção 4.3, com o propósito de analisar os resultados e comparar essas pontuações com outros trabalhos do estado da arte.

No conjunto de dados Flickr8k, o modelo proposto produziu uma maior pontuação BLEU no modelo Inception-V3, atingindo 0,584 (ou 58,40%). A Tabela 6 e Tabela 7 mostram as pontuações obtidas no modelo proposto nas arquiteturas Inception-V3 e VGG16, respectivamente. À medida que a média apresentada aumenta, também cresce o número de previsões corretas feitas pelo modelo proposto.

Tabela 6. Desempenho do modelo proposto com a arquitetura Inception-V3.

Métricas	Pontuação
Bleu-1	0,584
Meteor	0,176
Cider	0,338
Rouge-1	0,383
Rouge-L	0,370

Tabela 7. Desempenho do modelo proposto com a arquitetura VGG16.

Métricas	Pontuação
Bleu-1	0,560
Meteor	0,138
Cider	0,200
Rouge-1	0,357
Rouge-L	0,348

Revista ARACÊ
ISSN: 2358-2472

Conforme mostrado nas Tabelas 6 e 7 realizamos o teste com dois modelos diferentes. Assim, o modelo InceptionV3-LSTM obteve as seguintes pontuações: Bleu-1 (0,584), Meteor (0,176), Cider (0,338) e Rouge (0,383). Enquanto, o modelo VGG-16 pontuou Bleu-1 (0,560), Meteor (0,138), Cider (0,200) e Rouge (0,357). Portanto, o modelo proposto com maior precisão na legendagem automática é o InceptionV3 com a LSTM.

5.1 COMPARAÇÃO COM O ESTADO DA ARTE

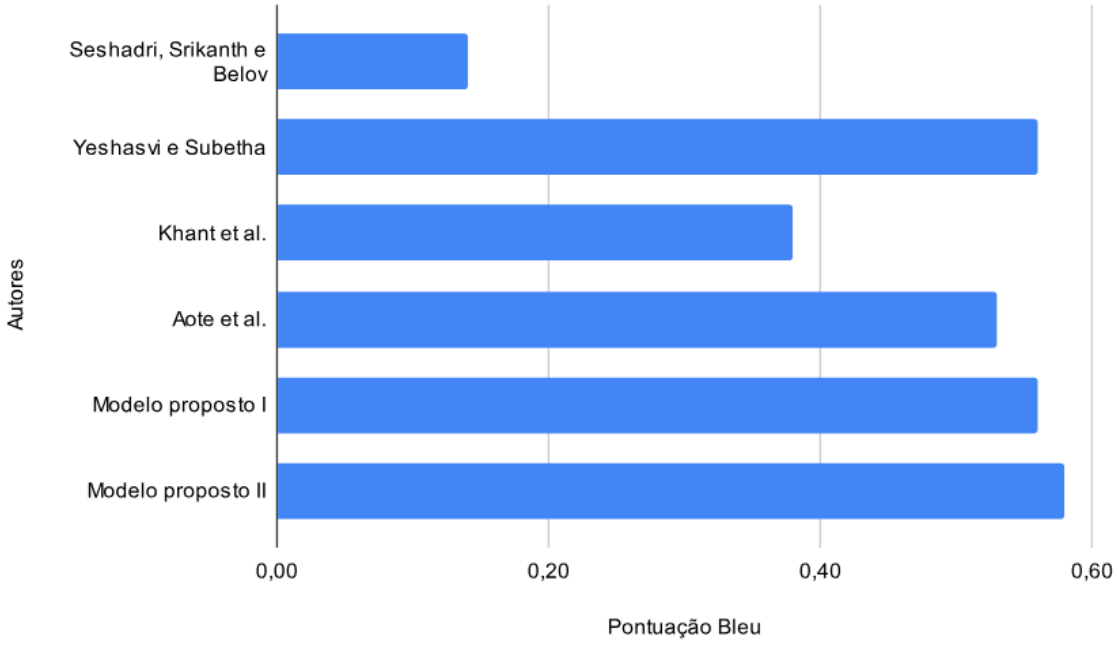
Para comprovar a eficácia do modelo proposto, comparamos seus resultados com os modelos do estado da arte, conforme ilustrado na Tabela 8 e na Figura 7. As tabelas evidenciam que o modelo proposto contribuiu para as abordagens mais recentes no conjunto de dados do Flickr8k. Foi observado que alguns estudos desenvolvidos anteriormente, não aplicaram as métricas de avaliação mais conhecidas, impossibilitando a verificação da total eficácia do modelo. O uso de técnicas avançadas de otimização e métodos para diminuir o *overfitting*, permitiu que o modelo desenvolvido tivesse pontuações satisfatórias. Sendo assim, a Tabela 8 compara os resultados dos modelos mais recentes na base do Flickr8k e a Figura 7 retrata a pontuação do Bleu em todos os modelos avaliados, o (-) indica métrica não retratada.

Tabela 8. Desempenho do modelo proposto comparado com modelos de última geração no conjunto de dados Flickr8k.

Autor	BL-1	ME	CI	RO-1	RO-L
Seshadri, Srikanth e Belov (2020)	0,14	0,14	-	0,20	-
Yeshasvi e Subetha (2022)	0,56	-	-	-	-
Khant et al. (2021)	0,38	-	-	-	-
Aote et al. (2022)	0,53	-	-	-	-
Método proposto (VGG16)	0,560	0,142	0,200	0,357	0,348
Método proposto (InceptionV3)	0,584	0,176	0,338	0,383	0,370

Revista ARACÊ
ISSN: 2358-2472

Figura 7. Comparação da métrica Bleu para diferentes modelos.



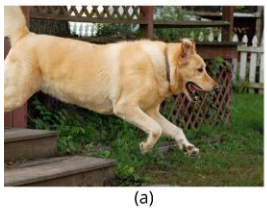
6 DISCUSSÕES E LIMITAÇÕES DO MODELO PROPOSTO

Conforme demonstrado na seção 5, o modelo proposto forneceu resultados satisfatórios nas métricas de avaliação quando utilizado a estrutura InceptionV3-LSTM no Flickr8K. Além disso, foram testadas imagens aleatórias para verificar a legendagem das imagens e sua eficácia, conforme as Figuras 8 e 9.

Foram selecionados 6 (seis) imagens com as suas legendas de referência, legenda prevista e a pontuação Bleu correspondente. Além disso, na Figura 9(f) foi verificado uma falha no modelo, na legenda de referência diz “dois cães correm em área de terra perto da floresta”, enquanto a legenda prevista foi “dois cachorros correm pela neve”. Durante a validação do modelo, houve dificuldades em descrever imagens que possuíam muitos objetos, porém o modelo se adapta a essas imagens, como a Figura 8(c) que gerou a legenda "homem de camisa preta e cachorro preto e branco estão brincando com bola na grama".

Revista ARACÊ
ISSN: 2358-2472

Figura 8. Imagens de amostra com suas legendas de referência e legenda gerada usando o modelo proposto.



(a)

----------**Legendas Atuais**----------
startseq blond dog runs down flight of stairs to the backyard endseq
startseq dog jumps off the stairs endseq
startseq tan dog runs down wooden staircase to the green grass endseq
startseq yellow dog is jumping across grassy yard in front of wooden deck endseq
startseq yellow dog jumping off of porch endseq
----------**Legenda Prevista**----------
startseq brown dog is running through the grass endseq
----------**BLEU Scores**----------
BLEU-1: 0.75

(b)

----------**Legendas Atuais**----------
startseq kite surfer is doing flip over the waves endseq
startseq man jumps over wave on his surfboard endseq
startseq person on parasail jumps off wave endseq
startseq silver craft rides the waves endseq
startseq windsurfer angles over wave endseq
----------**Legenda Prevista**----------
startseq person is jumping over the air endseq
----------**BLEU Scores**----------
BLEU-1: 0.75

(c)

----------**Legendas Atuais**----------
startseq black dog leaps for ball held by man endseq
startseq man is playing with black and white dog endseq
startseq man wearing glasses and his black and white dog wearing black collar are playing with tennis ball endseq
startseq man holding ball while dog jumps up for it endseq
startseq man with dog who is jumping to catch tennis ball endseq
----------**Legenda Prevista**----------
startseq man in black shirt and black and white dog are playing with ball in the grass endseq
----------**BLEU Scores**----------
BLEU-1: 0.68

Figura 9. Imagens de amostra com suas legendas de referência e legenda gerada usando o modelo proposto.

(d)

----------**Legendas Atuais**----------
startseq man wearing red helmet jumps up while riding skateboard endseq
startseq young man wearing red jacket performs jump on red skateboard endseq
startseq the helmeted boy is doing stunt on skateboard endseq
startseq the young man is skateboarding at skate park endseq
startseq young man is performing trick on skateboard in park endseq
----------**Legenda Prevista**----------
startseq boy in red shirt is jumping off ramp endseq
----------**BLEU Score**----------
BLEU-1: 0.60

(e)

----------**Legendas Atuais**----------
startseq few younger boys play around fountain endseq
startseq group of people gather around large fountain endseq
startseq three boys play around fountain in an office building courtyard endseq
startseq three kids are playing at fountain in front of building endseq
startseq three kids playing around fountain endseq
----------**Legenda Prevista**----------
startseq two women in the water endseq
----------BLEU Scores----------
BLEU-1: 0.42

(f)

----------**Legendas Atuais**----------
startseq black and tan dog is running with white and gray dog along dirt endseq
startseq the two dogs are running into the woods endseq
startseq two dogs run down dirt path in the forest endseq
startseq two dogs running away from the camera in the woods endseq
startseq two dogs run on dirt area near forest endseq
----------**Legenda Prevista**----------
startseq two dogs run through the snow endseq
----------**BLEU Scores**----------
BLEU-1: 0.58

Revista ARACÊ
ISSN: 2358-2472

Foram identificadas algumas limitações nos experimentos desta pesquisa, como a necessidade de um longo tempo de treinamento e um maior custo computacional para períodos de testes mais prolongados. Apesar dessas limitações, o modelo proposto foi eficiente na legendagem automática de imagens, podendo ser melhorado no uso de base de dados maiores e recursos computacionais mais robustos, permitindo um maior período de treinamento.

7 CONCLUSÕES E TRABALHOS FUTUROS

O processo da legendagem automática de imagens é de extrema importância no campo da Visão Computacional e Aprendizado de Máquina, visto que as máquinas possuem dificuldades em identificar objetos e relacioná-los corretamente. Por esse motivo, foi proposto o modelo com estrutura codificador-decodificador, o codificador baseado em CNN e o decodificador baseado em RNN. Este estudo obteve resultados satisfatórios nas métricas de avaliação mais conhecidas, sendo Bleu, Meteor, Cider e Rouge, no conjunto de dados do Flickr8k. Além disso, o modelo foi testado em imagens aleatórias ao vivo para verificar a capacidade das descrições geradas.

Nesta pesquisa, foram utilizadas as arquiteturas VGG16 e InceptionV3 para a extração de recursos das imagens, e o LSTM para gerar as descrições. Após o treinamento, o modelo que obteve o melhor resultado foi o InceptionV3-LSTM, alcançando 58,40% de precisão na métrica BLEU. A comparação do modelo proposto com os do estado da arte mostrou uma pontuação satisfatória, garantindo descrições compreensíveis para o ser humano, atingindo assim o objetivo desta pesquisa.

Como proposta para trabalhos futuros, pretende-se aprimorar o método apresentado implementando um maior banco de dados, como o Flickr30k e o MS COCO. Além disso, poderão ser aplicadas técnicas auxiliares para melhorar a pontuação das métricas de avaliação, a fim de aumentar a precisão e a qualidade das legendas geradas. Espera-se que essas melhorias possam tornar o modelo mais eficaz, contribuindo para o avanço da legendagem automática de imagens.

Revista ARACÊ
ISSN: 2358-2472

REFERÊNCIAS

AL-MALLA, M. A.; JAFAR, A.; GHNEIM, N. Image captioning model using attention and object features to mimic human image understanding. Journal of Big Data, Springer, v. 9, n. 1, p. 20, 2022.

ANSARI, K.; SRIVASTAVA, P. An efficient automated image caption generation by the encoder decoder model. Multimedia Tools and Applications, Springer, p. 1–26, 2024. 11 AOTE, S. S. et al. Image caption generation using deep learning technique. Journal of Algebraic Statistics, v. 13, n. 3, p. 2260–2267, 2022.

AYESHA, H. et al. Automatic medical image interpretation: State of the art and future directions. Pattern Recognition, Elsevier, v. 114, p. 107856, 2021.

BHALEKAR, M.; BEDEKAR, M. D-cnn: a new model for generating image captions with text extraction using deep learning for visually challenged individuals. Engineering, Technology & Applied Science Research, v. 12, n. 2, p. 8366–8373, 2022.

CHEN, C. et al. Figure captioning with relation maps for reasoning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. [S.l.: s.n.], 2020. p. 1537–1545.

FARABY, H. A. et al. Image to bengali caption generation using deep cnn and bidirectional gated recurrent unit. In: IEEE. 2020 23rd international conference on computer and information technology (ICCIT). [S.l.], 2020. p. 1–6.

FERREIRA, L. A. et al. Caption: Caption analysis with proposed terms, image of objects, and natural language processing. SN Computer Science, Springer, v. 3, n. 5, p. 390, 2022. 5GOEL, N. et al. An analysis of image captioning models using deep learning. In: IEEE. 2023 International Conference on Disruptive Technologies (ICDT). [S.l.], 2023. p. 131–136. HOSSEN,M.B.etal. Gva: guided visual attention approach for automatic image caption generation. Multimedia Systems, Springer, v. 30, n. 1, p. 50, 2024.

INDUMATHI, N. et al. Apply deep learning-based cnn and lstm for visual image caption generator. In: IEEE. 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). [S.l.], 2023. p. 1586–1591.

JAIN, B.; DOSHI, K.; DWIVEDI, P. Hybrid cnn-rnn model for accurate image captioning with age and gender detection. In: IEEE. 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS). [S.l.], 2023. p. 568–573.

KATIYAR, S.; BORGOHAIN, S. K. Comparative evaluation of cnn architectures for image caption generation. arXiv preprint arXiv:2102.11506, 2021.

KAVITHA, M. et al. Performance evaluation of deep e-cnn with integrated spatial spectral features in hyperspectral image classification. Measurement, v. 191, p. 110760, 2022. ISSN 0263-2241.

KESKIN, R. et al. Multi-gru based automated image captioning for smartphones. In: IEEE. 2021 29th Signal Processing and Communications Applications Conference (SIU). [S.l.], 2021. p. 1–4.

Revista ARACÊ
ISSN: 2358-2472

KHANT, P. et al. Image caption generator using cnn-lstm. International Research Journal of Engineering and Technology, v. 8, n. 07, p. 4100–4105, 2021.

NOGUEIRA, T. d. C. Modelo baseado em redes neurais profundas com unidades recorrentes bloqueadas para legendagem de imagens por referˆ encias. 2020.

NOGUEIRA, T. do C. et al. Reference-based model using multimodal gated recurrent units for image captioning. Multimedia Tools and Applications, Springer, v. 79, p. 30615–30635, 2020.

NOGUEIRA, T. do C. et al. A reference-based model using deep learning for image captioning. Multimedia Systems, Springer, v. 29, n. 3, p. 1665–1681, 2023.

PA, W. P.; NWE, T. L. et al. Automatic myanmar image captioning using cnn and lstm-based language model. In: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL). [S.l.: s.n.], 2020. p. 139–143.

PADATE, R. et al. Combining semi-supervised model and optimized lstm for image caption generation based on pseudo labels. Multimedia Tools and Applications, Springer, p. 1–21, 2023.

RINALDI, A. M.; RUSSO, C.; TOMMASINO, C. Automatic image captioning combining natural language processing and deep neural networks. Results in Engineering, Elsevier, v. 18, p. 101107, 2023.

ROHITHARUN, S.; REDDY, L. U. K.; SUJANA, S. Image captioning using cnn and rnn. In: IEEE. 2022 2nd Asian Conference on Innovation in Technology (ASIANCON). [S.l.], 2022. p. 1–8.

SASIBHOOSHAN, R.; KUMARASWAMY, S.; SASIDHARAN, S. Image caption generation using visual attention prediction and contextual spatial relation extraction. Journal of Big Data, Springer, v. 10, n. 1, p. 18, 2023. 5 SCOPARO, M.; SERAPI˜ AO, A. Deep learning para geração automática de legenda de imagem. In: SBC. Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional. [S.l.], 2019. p. 551–562.

SEHGAL, L.; MANDAN, S. Automated image capturing using cnn and rnn. International Journal of Research in Engineering, Science and Management, v. 5, n. 1, p. 13–17, 2022.

SESHADRI, M.; SRIKANTH, M.; BELOV, M. Image to language understanding: captioning approach. arXiv preprint arXiv:2002.09536, 2020.

SINGH, A.; GUPTA, S. Image based action recognition and captioning using deep learning. In: Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing. [S.l.: s.n.], 2023. p. 252–261.

SINGH, A.; VIJ, D. Cnn-lstm based social media post caption generator. In: IEEE. 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM). [S.l.], 2022. v. 2, p. 205–209.

Revista ARACÊ
ISSN: 2358-2472

SINGH, V.; SINGH, A. S.; ANANDHAN, K. Image captioning using machine/deep learning. In: IEEE. 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N). [S.l.], 2022. p. 849–854.

SURESH, K. R.; JARAPALA, A.; SUDEEP, P. Image captioning encoder–decoder models using cnn-rnn architectures: A comparative study. Circuits, Systems, and Signal Processing, Springer, v. 41, n. 10, p. 5719–5742, 2022.

THANGAVEL, K. et al. A novel method for image captioning using multimodal feature fusion employing mask rnn and lstm models. Soft Computing, Springer, v. 27, n. 19, p. 14205–14218, 2023.

VERMA,A. et al. Automatic image caption generation using deep learning. Multimedia Tools and Applications, Springer, v. 83, n. 2, p. 5309–5325, 2024.

WAHEED, S. R. et al. Cnn deep learning-based image to vector depiction. Multimedia Tools and Applications, Springer, v. 82, n. 13, p. 20283–20302, 2023.

YESHASVI, M.; SUBETHA, T. Image caption generator using machine learning and deep neural networks. In: Advances in Intelligent Computing and Communication: Proceedings of ICAC 2021. [S.l.]: Springer, 2022. p. 137–144.