


CNN-LSTM-BASED DEEP LEARNING FOR AUTOMATIC IMAGE CAPTIONING

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

Submitted on: 13/10/2024

Publication date: 13/11/2024

**Maria Vitória Sousa Ribeiro¹, Tiago do Carmo Nogueira², Gelson da Cruz Junior³,
Cássio Dener Noronha Vinhal⁴, Matheus Rudolfo Diedrich Ullmann⁵, Deller James
Ferreira⁶, Caio Henrique Rodrigues Carvalho⁷ and Danyele de Oliveira Santana⁸**

ABSTRACT

The evolution of Computer Vision and Machine Learning allows natural language image description techniques to be more efficient and accurate, through deep neural networks. This study used an encoder-decoder structure for object identification and captioning, through an input image. The proposed model used the VGG16 and Inception-V3 architectures as encoders and LSTM as decoder. To carry out the experiments, the Flickr8k dataset was used, with 8,000 images. The model was evaluated by the Bleu, Meteor, CIDEr and Rouge metrics. Achieving 58.40% accuracy according to the Bleu metric, thus ensuring human-understandable descriptions.

Keywords: Machine Learning. Deep Learning. Convolutional Neural Networks. Long Short-Term Memory. Image Subtitling.

¹ Graduated in Information Technology Management (IFBAIANO)
Federal Institute of Bahia (IFBAIANO)
E-mail: vicksousa503@gmail.com

² Dr. in Electrical and Computer Engineering (UFG)
Federal Institute of Bahia (IFBAIANO)
E-mail: tiago.nogueira@ifbaiano.edu.br

³ Dr. in Electrical Engineering (Unicamp)
Federal University of Goiás (UFG)
E-mail: gcruzjr@ufg.br

⁴ Dr. in Electrical Engineering (Unicamp)
Federal University of Goiás (UFG)
E-mail: vinhal@ufg.br

⁵ Doctorate student in Electrical and Computer Engineering (UFG)
Instituto Federal da Bahia (IFBA)
Email: matheusullmann@ifba.edu.br

⁶ Dr. in Education (UNB)
Federal University of Goiás (UFG)
E-mail: deller@inf.ufg.br

⁷ Master in Electrical Engineering (UFPI)
Federal Institute of Bahia (IFBAIANO)
E-mail: caio.carvalho@ifbaiano.edu.br

⁸ Master's student in Computer Science (UEFS)
Federal Institute of Bahia (IFBAIANO)
E-mail: danyele.santana@ifbaiano.edu.br

INTRODUCTION

Automatic subtitling informs in descriptive texts or in a single word, the classification of the object analyzed in the image, with the main objective of analyzing images to describe their content (RINALDI; RUSSIAN; TOMMASINO, 2023; NOGUEIRA, 2020). With the advancement of Artificial Intelligence (AI), Deep Learning (DL) models have contributed to automatic subtitling techniques being more accurate and performing better (SINGH; GUPTA, 2023).

DL is a sub-area of AI, being one of the Machine Learning (ML) techniques. ML has the ability to process a variety of data, finding patterns and being able to be programmed to learn models, in addition to using deep neural networks for analysis (SEHGAL; MANDAN, 2022).

Recently, studies in the field of deep neural networks are being applied to image analysis, Computer Vision (VC), Natural Language Processing (NLP) and subtitling (RINALDI; RUSSIAN; TOMMASINO, 2023; AL-MALLA; JAFAR; GHNEIM, 2022; JAIN; DOSHI; DWIVEDI, 2023). According to Sehgal and Mandan (2022), Computer Vision is a field of Computer Science that studies the ability of machines to extract elements from an image or video. Through VC, it is possible to perform object recognition in images, facial recognition, and motion analysis, training the machine to understand what is being seen in a similar way to humans.

The amount of images available on the Internet makes it require the need to identify and describe them, as highlighted by Al-Malla, Jafar, and Ghneim (2022). Despite the ability of humans to identify objects, machines have difficulty in this classification (NOGUEIRA, 2020). Therefore, machines can be trained by different types of datasets to acquire this ability to identify objects and relate them in the image (VERMA et al., 2024).

Recent studies have applied the encoder-decoder structure to identify the relationship of objects in the image. This structure uses deep neural network architectures to encode the image and describe it, using mainly Convolutional Neural Networks (CNN's) and *Recurrent Neural Networks* (RNN's) for better sentence quality (NOGUEIRA et al., 2023). CNN is the encoder, being used to analyze the images, classify them and recognize their patterns. Meanwhile, RNN is the decoder, being applied in the captioning of images (AL-MALLA; JAFAR; GHNEIM, 2022; NOGUEIRA, 2020).

However, authors who used traditional RNN found limitations in its applications, namely: long training time, explosion or disappearance of gradients (NOGUEIRA et al.,

2020; NOGUEIRA et al., 2023). Thus, researchers obtained better results by applying Closed Recurrent Unit (GRU) or Long *Short-Term Memory* (LSTM) (SINGH) models; SINGH; ANANDHAN, 2022; BHALEKAR; BEDEKAR, 2022).

Automatic captioning of objects in images is a technique that has several applications, and can be used to help people with visual impairments, recommendations on social media, image indexing, visual search, security systems, among other natural language processing applications (VERMA et al., 2024). Therefore, this study has as its main objective the development of a model that performs automatic image captioning through deep learning.

The proposed model is based on the encoder-decoder architecture for the captioning of objects in images. Thus, for the extraction of the characteristics of the images, CNN will be used as an encoder. And to generate the description of the objects, the LSTM will be used as a decoder. The proposed model provides state-of-the-art results for image captioning in public datasets such as Flickr8k.

The main objects of this research are: to identify technologies and models in the literature for image captioning; analyze how AI and Deep Neural Networks have contributed to subtitling; and propose a Deep Learning model with greater accuracy and efficiency for automatic image captioning.

The summary of our research contributions is as follows: to propose VGG16 and Inception23 as encoders and LSTM as decoder for automatic captioning of objects in images; as well as reporting experimental results using the most commonly used metrics such as METEOR, BLEU, CIDEr, and ROUGE in the Flickr8k dataset, contributing to the state-of-the-art approaches. In addition, present the results of the proposed model, which were validated in randomly selected live images.

This article is organized as follows: section 2 presents the related works on CNN and RNN in image object detection and automatic captioning; section 3 addresses the procedures and methodological instruments used; section 4 addresses the experiments performed; Section 5 presents the results; section 6 discusses the discussions and limitations of the proposed model; Section 7 addresses the conclusion of the study and future work.

RELATED JOBS

Automatic subtitling is a challenging and constantly evolving field in Computer Vision, and deep learning architecture is used for image analysis and generation of natural language subtitles (NOGUEIRA, 2020). Machines still have the challenge of identifying the object, describing it and analyzing its relationship with the environment. If necessary, the present study and related works in Deep Learning.

According to Al-Malla, Jafar, and Ghneim (2022), the encoder-decoder architecture is the most used in the deep learning technique for automatic image captioning. The encoder-decoder structure can be divided into two categories, namely: bottom-up and top-down approaches (NOGUEIRA et al., 2020; NOGUEIRA et al., 2023).

In order to increase the quality of the legends, object resources and convolutional resources of a model are used. CNN acts as an encoder from the input data, extracting the patterns from the images and the tags from the objects. It is subdivided into layers, namely: input layer, convolution layer, *pooling* layer, fully connected layer, and exit layer (FARABY et al., 2020; SINGH; VIJ, 2022). Meanwhile, RNN acts as a decoder, subtitling these images (NOGUEIRA, 2020).

Sehgal and Mandan (2022) proposed a model for captioning photographs using neural networks, in addition to the LSTM model, for greater accuracy in automatic captions. LSTM is a type of RNN architecture with the ability to learn from sequential data. In their results, a better percentage of accuracy was found in the subtitles generated using this model. LSTM analyzes the characteristics of the image and generates a text or description that matches it. The image is transcribed by decoding the input vectors a (NOGUEIRA, 2020; GOEL et al., 2023).

In this context, Thangavel et al. (2023) explored in their research the use of the encoding layer, employing the architecture of the Recurrent Neural Network Mask (Faster R-CNN), while the decoding layer adopted the use of the attention mechanism with LSTM, achieving a performance of 25% higher than the methods analyzed in the research. Although LSTM has a higher performance in the accuracy of image captions, to implement it requires a longer training time (NOGUEIRA, 2020).

Rohitharun, Reddy, and Sujana (2022) proposed in their study a one-to-one model based on ResNet-101 as an encoder and LSTM as a decoder. Through it, the resources of the images are extracted in the format of vectors. ResNet-101, because it has fewer

parameters, allows for deeper training and makes it more efficient in terms of using computational resources.

Meanwhile, Nogueira et al. (2020) to solve the problems of traditional RNN in the disappearance of gradients, used the R-GRU model in their study for better power and result. From the Flickr30k database, they obtained 69.40% accuracy in the Bleu metric. This approach uses CNN to analyze the image in small parts and extract the most important visual features from the input image. The multimodal GRU is a type of RNN that receives data from CNN and other types of data, with the objective of generating the sentence by reference.

CNN-based models require a large amount of data for analysis and training, avoiding problems such as lack of accuracy in captions (AYESHA et al., 2021; WAHEED et al., 2023). The performance of image generation varies according to the multiple CNN architectures used (KATIYAR; BORGOHAIN, 2021).

In this sense, Verma et al. (2024) conducted a survey of pre-trained CNN architectures in their study to discuss and compare the best model. The first model analyzed was the VGG16, which achieved more than 90% accuracy. The second model examined was the Inception V3, which demonstrated superior accuracy to the previously analyzed network. Finally, the last model evaluated was ResNet-50, which is considered the most recent and effective for deep neural networks compared to the other models. When comparing the architectures surveyed, the authors proposed in their study the VGG16 Hybrid Places 1365 model as an encoder, being used to provide specific object and scene results. The proposed model achieved 66.66% accuracy in the Bleu evaluation metric and 50.60% in the Meteor.

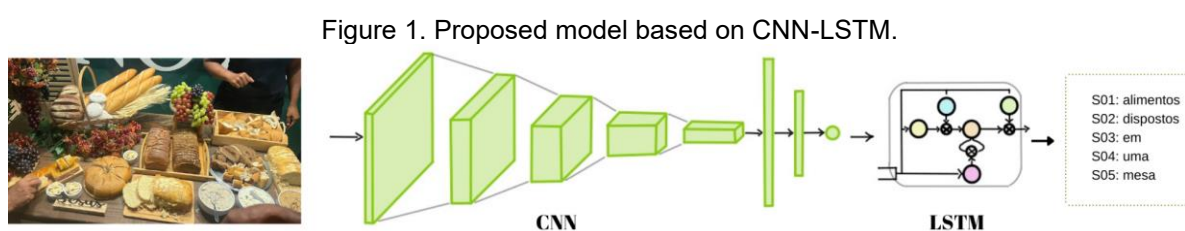
Seshadri, Srikanth and Belov (2020) developed a web system to receive an image as input and generate the captioning that corresponds to it, obtaining 13% and 14% accuracy according to the Bleu and Meteor metrics. A CNN encoder model and a bidirectional LSTM decoder were used. Encoder-decoder approaches can be applied in two ways, namely: injection architecture and merge architecture. In injection architecture, the image is encoded in a fixed set of numbers. After that, the combination is made with each word in the text description, the decoder uses this information to predict what the next word in the description will be. While, the merge architecture extracts the encoding of the image and the encoding of the description. The decoder uses both encodings to generate the description sequence.

Sasibhooshan, Kumaraswamy, and Sasidharan (2023) used the convolutional neural network based on discrete wavelet decomposition (WCNN) in their model. The use of WCNN makes it possible to extract the elements from the image, as well as the approximate location of the objects in it, along with frequency information. This model allows the identification of the most important parts of an image and demonstrates greater accuracy in highly complex images, reaching 38.20% in the Bleu-4 metric.

Several works related to literature develop their models following the encoder-decoder structure. Most of these models are aimed at describing images using architectures such as CNN, RNN, LSTM, GRU, among others. However, some of these models when developed presented problems, such as the inaccuracy of subtitles or loss of information. Although recent studies mostly use the CNN-RNN architecture for image captioning, some authors have added other methods and auxiliary techniques for better performance (CHEN et al., 2020; HOSSEN et al., 2024; FERREIRA et al., 2022; SCOPARO; SERAPIÃO, 2019; PADATE et al., 2023). In the next section, an encoder-decoder model will be proposed to mitigate these problems.

METHODOLOGICAL PROPOSAL

This work proposes a model based on the encoder-decoder architecture, using CNN-RNN for the captioning of objects in images. Thus, VGG16 and Inception-V3 were used as encoders to extract the characteristics of the images. And to generate the description of the objects, the LSTM was used as a decoder, as shown in Figure 1.



The choice of CNN is justified by the fact that it is a network with the ability to learn and extract complex patterns from the images, in addition to the possibility of being pre-trained with datasets to reduce training time and obtain greater accuracy in the analysis of the images. In addition, CNN has the ability to analyze images of varying sizes and differentiate between training images and input images.

The use of VGG16 is explained by the fact that it is a convolutional model that is simple to implement and effective in computer vision tasks, such as image detection and description (YESHASVI; SUBETHA, 2022). Inception-V3 was used because it has a low computational cost compared to the other models, in addition to the availability of pre-trained implementations and proof of its performance (NOGUEIRA et al., 2020).

On the other hand, the choice of LSTM is justified by having a long-term memory, in addition to the ability to keep information and greater storage (PA; NWE et al., 2020). In addition, it has better accuracy results in the subtitles generated in the studies developed previously (ROHITHARUN; REDDY; SUJANA, 2022; YESHASVI; SUBETHA, 2022; AOTE et al., 2022).

The architecture of the proposed model can be divided mainly into three modules. The first module is the extraction of resources using CNN, the second is the proposed architectures of VGG16 and Inception-V3 (used as an encoder) and the third is the generation of subtitles using LSTM.

This section describes the methodology adopted in the present study. In the following subsections (3.1, 3.2 and 3.3), the description of the proposed model will be addressed.

RESOURCE EXTRACTION

CNNs have the ability to recognize and identify objects in an image, through the processing of an input image and classification according to predefined models (NOGUEIRA, 2020) CNN is subdivided into three main layers, which are present in all CNNs, namely: convolution layers, *pooling* layers, and fully connected deep layers (FC) (VERMA et al., 2024).

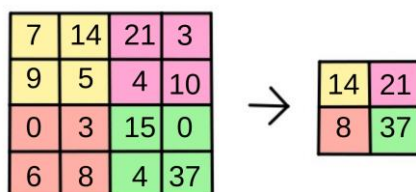
The convolution layer is the first layer to extract the features of the image, with the main objective of detecting image features such as edges, patterns, and objects. A filter (*kernel*) is applied to small parts of it, at width, height, and depth levels, generating characteristic maps for each *pixel* (NOGUEIRA, 2020). These maps reflect the image at different levels of abstraction, the more convolutional filters applied, the higher the processing and memory cost.

Meanwhile, the *Pooling* layer aims to reduce the size of the input image, decreasing the computational cost of a neural network when applied after the convolution layer. In a CNN, the second layer performed is *pooling*, and three techniques can be applied to

compress the characteristic maps, namely: average *pooling*, maximum *pooling* and *L2 pooling*.

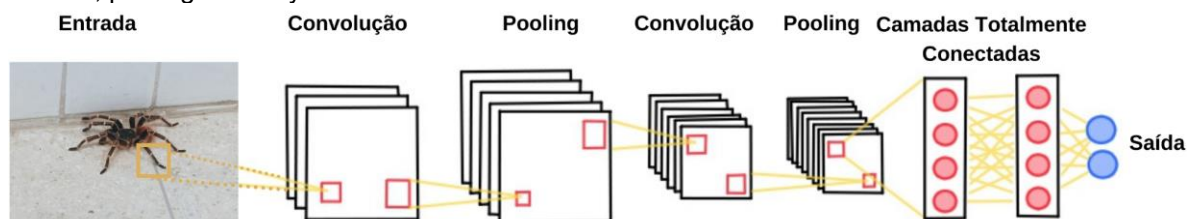
Maximum *pooling* is the most commonly used way to perform *pooling*. In this technique, size 22 filters are applied to 44 images, which allows only 25% of the activations to be used, as shown in figure 2.xx

Figure 2. Example of how max-pooling works with a 2x2 filter in a 4x4 image.



Finally, the fully connected layer transforms a data matrix into a single vector. As demonstrated in Figure 3, CNN starts with the convolution and *pooling layers*, splitting the images into features and analyzing them separately. Then, these features are applied to the *FC layer*, where the vector classification occurs (NOGUEIRA, 2020).

Figure 3. Complete example of a CNN for the classification of objects in images, using the basic layers of convolution, pooling and fully connected.



To reduce CNN's training time when trained from scratch, the VGG16 and Inception-V3 models were used in this study.

ARCHITECTURE OF THE PROPOSED MODEL

As mentioned in the previous subsection, the proposed model is based on the VGG16 and Inception-V3 architectures to extract the features of the images in the training and validation processes, both of which are convolutional neural networks.

According to Table 1, Inception-V3 traditionally has the input size 2992993 (299299 referring to input and 3 referring to filters) and produces an output *xxxpool* of 882048 (88 referring to input and 2048 referring to filters), this allows the reduction of the number of parameters and makes the model more efficient.xxx

Table 1. Inception-V3 Architecture.

| Layer | Entry | Filters |
|-------------|----------|---------|
| Conv | 299 299x | 3 |
| Conv | 149 149x | 32 |
| Conv padded | 147 147x | 32 |
| Pool | 147 147x | 64 |
| Conv | 73 73x | 64 |
| Conv | 71 71x | 80 |
| Conv | 35 35x | 192 |
| 3 Inception | 17 17x | 288 |
| 5 Inception | 8 8x | 768 |
| 2 Inception | 8 8x | 1280 |
| Pool | 8 8x | 2048 |
| Linear | 1 1x | 2048 |
| Softmax | 1 1x | 1000 |
| - | - | - |

Meanwhile, the VGG16 has 16 deep layers, allowing the analysis of more complex images (PA; NWE et al., 2020). In addition, it has a 2242243 input (224224 referring to the input and 3 referring to the filters). According to Table 2, VGG16 contains five convolution layersxxx (*conv*), five pool layers (*maxpool*), and three fully connected layers (*fc*). The last layer described by *Softmax* is responsible for the probability classification of the object in the image.

Table 2. VGG16 architecture.

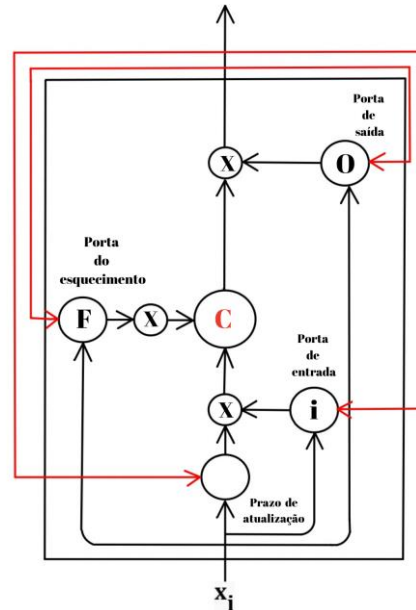
| Layer | Entrance | Filters |
|---------|------------|---------|
| Input | 224 224x | 3 |
| 2 conv | 224 224x | 64 |
| Maxpool | 112 112x | 128 |
| 2 conv | 112 112x | 128 |
| Maxpool | 56 56x | 256 |
| 3 conv | 56 56x | 256 |
| Maxpool | 28 28x | 512 |
| 3 conv | 28 28x | 512 |
| Maxpool | 14 14x | 512 |
| 3 conv | 14 14x | 512 |
| Maxpool | 7 7x | 512 |
| 2 fc | 1 1x | 4096 |
| 1 fc | 1 1x | 1000 |
| Softmax | Classifier | - |

SUBTITLE GENERATION

In this study, LSTM, which is a type of RNN, was used for better accuracy of descriptions and to ensure long-term sentence memory. Through memory cells, LSTM has the ability to delete or archive a certain piece of information (ROHITHARUN; REDDY; SUJANA, 2022), in addition to reducing the problem of gradient explosion in traditional RNN (KHANT et al., 2021).

Figure 4 shows three gates of the LSTM cell architecture, namely: *Input Gate*, *Output Gate*, and *Forget Gate*, controlled by memory cell C (located in the center of the image). These ports are responsible for performing a certain operation, having all the relevant information stored in their memory cell. The forget port determines whether the current cell value should be discarded or kept in memory, the input port defines whether the new cell value should be read or discarded, and the output port defines whether it is necessary to generate a new value for the cell (SURESH; JARAPALA; SUDEEP, 2022).

Figure 4. Basic architecture of the LSTM cell.



Step i receives input from different sources: the past hidden state (h_{i-1}); the current entry (X_i); and the previous state of the memory cell (C_{i-1}). In the time step t , the updated port values for the given inputs X and u , h_{i-1} and C_{i-1} , are:

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

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

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

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

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

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

where X_i represents the inputs of the input port, X_F of the oblivion port, X_O of the output port, and X_C of the memory cell, BV represents vectors, and WM represents weight metrics. In addition to ϕ is the hyperbolic tangent, which can be calculated:

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

In addition, the sigmoid activation function is ρ , and is calculated by:

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

EXPERIMENTS

This section presents the data sets used, as well as the configurations of the experiments, the parameters used, and the evaluation metrics applied for analysis of the results.

DATASETS

During the literature analysis, several data sets were used for training, validation, and testing of the generated subtitles (ROHITHARUN; REDDY; SUJANA, 2022). As shown in Table 3, the authors used the Flickr8k datasets (KHANT et al., 2021; INDUMATHI et al., 2023), Flickr30k (NOGUEIRA et al., 2023) and MS-COCO (KESKIN et al., 2021) for the experiment, as they have a variety of images and more than one descriptive caption.

Table 2. Set of data present in the literature analysis.

| Authors | Database |
|-----------------------------------|---------------------|
| 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 and Ghneim (2022) | Flickr30k e MS COCO |

For this study of automatic image captioning, the Flickr8k public database was used. For each image in the database, five different captions were associated, generated by human beings in the English language, as shown in Figure 5. According to table 4, the Flickr8k dataset has 8,000 images, 6,000 of which are for training, 1,000 for validation, and 1,000 for testing.

Figure 5. Sample image with the corresponding captions.



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 .

Table 4. Sample numbers in the Flickr8K dataset.

| Dataset | Flickr8K |
|------------|----------|
| Total | 8.000 |
| Training | 6.000 |
| Test | 1.000 |
| Validation | 1.000 |

EXPERIMENT SETTINGS

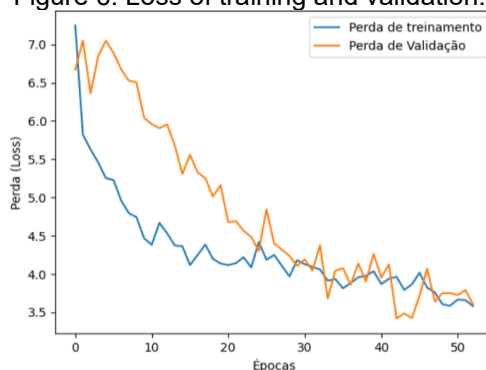
The first stage of the experiments is the pre-processing of the images and descriptions. It is necessary to reduce the training time of a model without compromising the effectiveness of the generated subtitles (ANSARI; SRIVASTAVA, 2024). For the pre-processing of the descriptions, the spaces and punctuation between words are removed, in addition to all letters are converted to lowercase. The bank captions are separated by commas and stored in a single list, in addition to the start and end tokens being added for each caption, thus allowing for better training. In the end, a total of 40,455 subtitles were counted, with 8,768 being the vocabulary size and 34 the maximum length of the subtitle.

The input images have a resolution of 256500 to 500500. In its pre-processing, the images were resized to 224224 pixels when testing the VGG16 model and to 299299 pixels when testing the Inception-V3 model. *xxxx*

The training data is used for the model to learn the patterns and adjust the parameters, while the validation data is used to verify the performance of the model while still in training. The test data is presented after the creation of the model, to validate the effectiveness of the algorithm. This division is important for the algorithm to learn and predict captioning on new images.

The *Early Stopping* technique was used to monitor performance at each time in order to avoid *overfitting* - the model's difficulty in predicting when tested on new data (KAVITHA et al., 2022). In the proposed model, 70 epochs and a "*patience*" value of 10 epochs were defined, i.e., if performance did not improve or the loss of validation increased for 10 consecutive epochs, training would be interrupted. As shown in Figure 6, the algorithm's training was interrupted when there was *overfitting*, ensuring that the model performed automatic captioning on test images in a way that was understandable to humans.

Figure 6. Loss of training and validation.



To reduce the explosion of gradients and possible losses, the Adam optimizer was used. In addition, Adam requires little computational cost, being efficient on larger data sets and shorter training period. In addition, the reLU activation function was used to improve the constancy of gradients during training.

Therefore, we set the encoder to run 70 epochs and *batch_size* 512, training the full set of images available for training and validation in the database. During the model validation process, the parameters were used, as shown in Table 5.

Table 5. Parameters used during the experiments.

| Parameters | Value |
|---------------|-------------------|
| Times | 70 |
| Batch_size | 512 |
| Optimizer | Adam |
| Learning Rate | $1 \cdot 10^{-4}$ |
| Dropout | 0,1 |

The experiments were carried out in the Google Collab development environment. To develop the proposed model, we used some libraries such as Tensorflow, nltk, Keras, os, pickle, numpy, among others.

EVALUATION METRICS

To measure model performance, it is essential to apply evaluation metrics to ensure the quality of the generated captions. This is done by comparing the reference subtitles with their content, the grammatical correctness of the subtitling, among other parameters (VERMA et al., 2024; AL-MALLA; JAFAR; GHNEIM, 2022; NOGUEIRA et al., 2020; NOGUEIRA et al., 2023). For this reason, we evaluated the model of this study using the METEOR, BLEU, CIDEr, and ROUGE evaluation metrics.

The METEOR method is an evaluation metric used to measure the quality of the generated subtitles. This metric aligns the automatic captions and the reference phrases, allowing the analysis of the degree of similarity between them (NOGUEIRA et al., 2020). In addition, METEOR builds on the concept of unigram matching by comparing which words are present individually in the model-generated captions and the words in the reference caption. The Meteor metric can be calculated as follows:

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

where it is defined by unigram accuracy and retrieval. The conditional sentence is calculated by: PR

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

where C is defined by the count of matching unigrams, while M_u indicating the minimum number of sentences required for it to match the unigrams in the reference translations.

M_u C

METEOR is designed to be more closely aligned with human-generated captions, focusing on recall and accuracy of descriptions. While, the BLEU evaluation metric is

implemented by analyzing the quality of the captions generated by the reference captions of humans (VERMA et al., 2024). This evaluation is calculated by n -gram precision between the generated sentences and the reference sentences. It has the limitation of not analyzing the meaning of the sentences and their structure, analyzing only the grams. There are several types of calculations for this metric, the main calculation being:

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

where B_p represents the value of the brevity of the sentence generated by the model, that is, how concise the caption is. A is the geometric mean of the modified precision of n -grams. And C is the maximum length of the n -grams of the candidate phrase (NOGUEIRA et al., 2023). Thus, brevity is calculated by:

$$B_p = \begin{cases} 1 & \text{if } C > r \\ e^{(1-r/C)} & \text{if } C \leq r \end{cases} \quad (12)$$

where C is the length of the model-generated legend and r is the length of the reference dataset. The BLEU score is rated as excellent when it is higher than 0.5. While, the legend below 0.15 indicates that it needs improvement.

The CIDEr metric has as its main characteristic salience and grammatical analysis (AL-MALLA; JAFAR; GHNEIM, 2022). It consists of three processes, namely: collection of different human-generated captions, consensus measurement to identify similar words in captions, consensus automation, and two datasets that describe the images (NOGUEIRA et al., 2023). The calculation for this valuation metric is:

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

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

Finally, the ROUGE metric evaluates the quality of the caption by comparing the number of words in the description generated by the model to the human description used as a reference. The Rouge metric is calculated as follows:

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

where m represents the length of n -gram; and m and n are the maximum number of n -grams that occur in a summary of our model and a set of reference summaries S_H .

To verify the effectiveness of the model, we use the MS-COCO evaluation API that allows the classification of the generated subtitles. The API uses BLEU, METEOR, CIDEr and ROUGE evaluation metrics, ensuring accurate performance analysis of the results.

RESULTS

The main objective of this study is to achieve understandable results in captioning after training by correlating the elements of the input image with the generated caption. This section presents the scores of the evaluation metrics mentioned in subsection 4.3, with the purpose of analyzing the results and comparing these scores with other state-of-the-art works.

In the Flickr8k dataset, the proposed model produced a higher BLEU score in the Inception-V3 model, reaching 0.584 (or 58.40%). Table 6 and Table 7 show the scores obtained in the proposed model in the Inception-V3 and VGG16 architectures, respectively. As the average presented increases, so does the number of correct predictions made by the proposed model.

Table 6. Performance of the proposed model with the Inception-V3 architecture.

| Metric | Punctuation |
|--------|-------------|
| Bleu-1 | 0,584 |
| Meteor | 0,176 |
| Cider | 0,338 |
| Red-1 | 0,383 |
| Red-L | 0,370 |

Table 7. Performance of the proposed model with the VGG16 architecture.

| Metric | Punctuation |
|--------|-------------|
| Blue-1 | 0,560 |
| Meteor | 0,138 |
| Cider | 0,200 |
| Red-1 | 0,357 |
| Red-L | 0,348 |

As shown in Tables 6 and 7, we performed the test with two different models. Thus, the InceptionV3-LSTM model obtained the following scores: Bleu-1 (0.584), Meteor (0.176), Cider (0.338) and Rouge (0.383). Meanwhile, the VGG-16 model scored Bleu-1 (0.560), Meteor (0.138), Cider (0.200) and Rouge (0.357). Therefore, the model proposed with the greatest accuracy in automatic subtitling is InceptionV3 with LSTM.

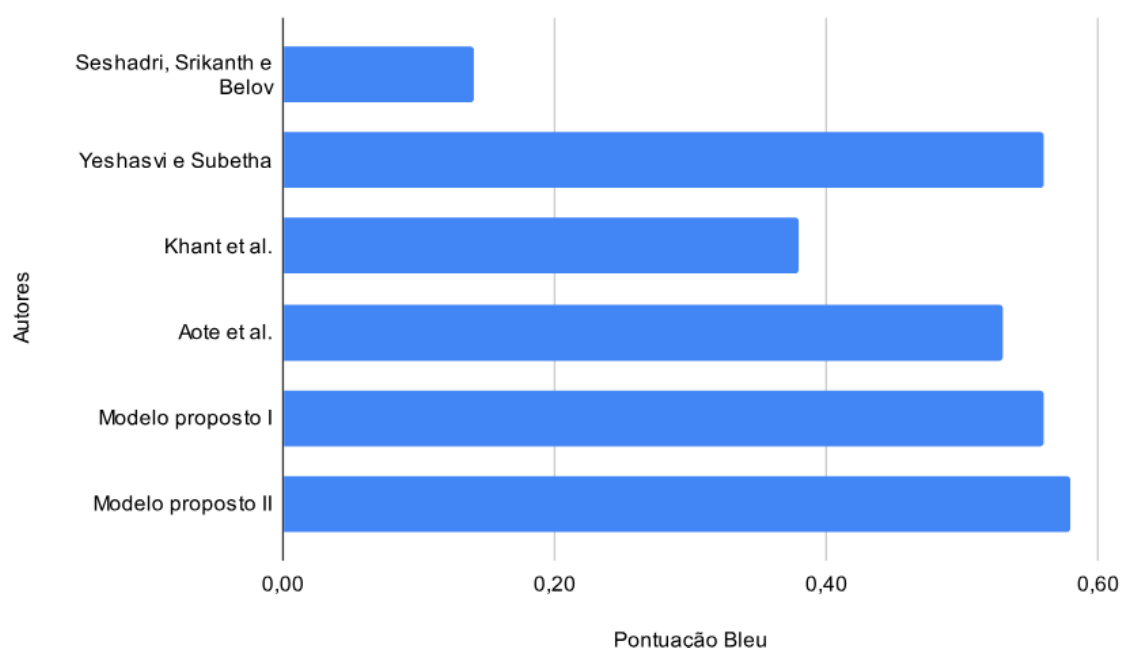
COMPARISON WITH THE STATE OF THE ART

To prove the effectiveness of the proposed model, we compared its results with the state-of-the-art models, as illustrated in Table 8 and Figure 7. The tables show that the proposed model contributed to the most recent approaches in the Flickr8k dataset. It was observed that some studies previously developed did not apply the most well-known evaluation metrics, making it impossible to verify the total effectiveness of the model. The use of advanced optimization techniques and methods to decrease *overfitting* allowed the developed model to have satisfactory scores. Thus, Table 8 compares the results of the most recent models in the Flickr8k database and Figure 7 depicts the Bleu score in all the models evaluated, the (-) indicates a metric not depicted.

Table 8. Performance of the proposed model compared to state-of-the-art models in the Flickr8k dataset.

| Author | BL-1 | ME | THERE | RO-1 | RO-L |
|-----------------------------------|-------|-------|-------|-------|-------|
| Seshadri, Srikanth E Belov (2020) | 0,14 | 0,14 | - | 0,20 | - |
| Yashasvi A Subetha (2022) | 0,56 | - | - | - | - |
| Khant et al. (2021) | 0,38 | - | - | - | - |
| Aote et al. (2022) | 0,53 | - | - | - | - |
| Proposed method (VGG16) | 0,560 | 0,142 | 0,200 | 0,357 | 0,348 |
| Proposed Method (InceptionV3) | 0,584 | 0,176 | 0,338 | 0,383 | 0,370 |

Figure 7. Comparison of the Bleu metric for different models.



DISCUSSIONS AND LIMITATIONS OF THE PROPOSED MODEL

As demonstrated in section 5, the proposed model provided satisfactory results in the evaluation metrics when using the InceptionV3-LSTM framework on Flickr8K. In addition, random images were tested to verify the captioning of the images and their effectiveness, as shown in Figures 8 and 9.

6 (six) images were selected with their reference captions, expected caption and the corresponding Bleu score. In addition, in Figure 9(f) a flaw in the model was verified, in the reference legend it says "two dogs run in an area of land near the forest", while the

predicted legend was "two dogs run through the snow". During the validation of the model, there were difficulties in describing images that had many objects, but the model adapts to these images, such as Figure 8(c) that generated the caption "man in black shirt and black and white dog are playing with a ball in the grass".

Figure 8. Sample images with their reference captions and caption generated using the proposed template.



(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

Figure 9. Sample images with their reference captions and caption generated using the proposed template.



(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

Some limitations were identified in the experiments of this research, such as the need for a long training time and a higher computational cost for longer test periods. Despite these limitations, the proposed model was efficient in the automatic captioning of images, and can be improved in the use of larger databases and more robust computational resources, allowing a longer training period.

CONCLUSIONS AND FUTURE WORK

The process of automatic image captioning is extremely important in the field of Computer Vision and Machine Learning, since machines have difficulties in identifying objects and relating them correctly. For this reason, the model with an encoder-decoder structure, the CNN-based encoder and the RNN-based decoder were proposed. This study obtained satisfactory results in the most well-known evaluation metrics, being Bleu, Meteor, Cider and Rouge, in the Flickr8k dataset. In addition, the model was tested on random live images to verify the capability of the generated descriptions.

In this research, the VGG16 and InceptionV3 architectures were used to extract features from the images, and LSTM to generate the descriptions. After training, the model that obtained the best result was InceptionV3-LSTM, achieving 58.40% accuracy in the

BLEU metric. The comparison of the proposed model with those of the state of the art showed a satisfactory score, ensuring comprehensible descriptions for the human being, thus achieving the objective of this research.

As a proposal for future works, it is intended to improve the method presented by implementing a larger database, such as Flickr30k and MS COCO. In addition, ancillary techniques may be applied to improve the scoring of evaluation metrics in order to increase the accuracy and quality of the generated captions. It is hoped that these improvements can make the model more effective, contributing to the advancement of automatic image captioning.

REFERENCES

1. Al-Malla, M. A., Jafar, A., & Ghneim, N. (2022). Image captioning model using attention and object features to mimic human image understanding. *Journal of Big Data*, 9(1), 20. <https://doi.org/10.1186/s40537-022-00688-1>
2. Ansari, K., & Srivastava, P. (2024). An efficient automated image caption generation by the encoder-decoder model. *Multimedia Tools and Applications*, 1–26. <https://doi.org/10.1007/s11042-024-15023-1>
3. Aote, S. S., et al. (2022). Image caption generation using deep learning technique. *Journal of Algebraic Statistics*, 13(3), 2260–2267.
4. Ayesha, H., et al. (2021). Automatic medical image interpretation: State of the art and future directions. *Pattern Recognition*, 114, 107856. <https://doi.org/10.1016/j.patcog.2021.107856>
5. Bhalekar, M., & Bedekar, M. (2022). D-CNN: A new model for generating image captions with text extraction using deep learning for visually challenged individuals. *Engineering, Technology & Applied Science Research*, 12(2), 8366–8373. <https://doi.org/10.48084/etasr.4720>
6. Chen, C., et al. (2020). Figure captioning with relation maps for reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1537–1545).
7. Faraby, H. A., et al. (2020). Image to Bengali caption generation using deep CNN and bidirectional gated recurrent unit. In *IEEE 23rd International Conference on Computer and Information Technology (ICCIT)* (pp. 1–6). <https://doi.org/10.1109/ICCIT49330.2020.9316821>
8. Ferreira, L. A., et al. (2022). Caption: Caption analysis with proposed terms, image of objects, and natural language processing. *SN Computer Science*, 3(5), 390. <https://doi.org/10.1007/s42979-022-01660-z>
9. Goel, N., et al. (2023). An analysis of image captioning models using deep learning. In *IEEE 2023 International Conference on Disruptive Technologies (ICDT)* (pp. 131–136). <https://doi.org/10.1109/ICDT59023.2023.00034>
10. Hossen, M. B., et al. (2024). GVA: Guided visual attention approach for automatic image caption generation. *Multimedia Systems*, 30(1), 50. <https://doi.org/10.1007/s00542-023-07317-4>
11. Indumathi, N., et al. (2023). 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)* (pp. 1586–1591).

12. Jain, B., Doshi, K., & Dwivedi, P. (2023). 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) (pp. 568–573).
13. Katyar, S., & Borgohain, S. K. (2021). Comparative evaluation of CNN architectures for image caption generation. arXiv Preprint arXiv:2102.11506.
14. Kavitha, M., et al. (2022). Performance evaluation of deep E-CNN with integrated spatial spectral features in hyperspectral image classification. *Measurement*, 191, 110760. <https://doi.org/10.1016/j.measurement.2022.110760>
15. Keskin, R., et al. (2021). Multi-GRU based automated image captioning for smartphones. In IEEE 2021 29th Signal Processing and Communications Applications Conference (SIU) (pp. 1–4).
16. Khant, P., et al. (2021). Image caption generator using CNN-LSTM. *International Research Journal of Engineering and Technology*, 8(7), 4100–4105.
17. Nogueira, T. d. C. (2020). Modelo baseado em redes neurais profundas com unidades recorrentes bloqueadas para legendagem de imagens por referências.
18. Nogueira, T. do C., et al. (2020). Reference-based model using multimodal gated recurrent units for image captioning. *Multimedia Tools and Applications*, 79, 30615–30635. <https://doi.org/10.1007/s11042-020-10037-4>
19. Nogueira, T. do C., et al. (2023). A reference-based model using deep learning for image captioning. *Multimedia Systems*, 29(3), 1665–1681. <https://doi.org/10.1007/s00542-022-00896-4>
20. Pa, W. P., Nwe, T. L., et al. (2020). 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) (pp. 139–143).
21. Padate, R., et al. (2023). Combining semi-supervised model and optimized LSTM for image caption generation based on pseudo labels. *Multimedia Tools and Applications*, 1–21. <https://doi.org/10.1007/s11042-023-17872-5>
22. Rinaldi, A. M., Russo, C., & Tommasino, C. (2023). Automatic image captioning combining natural language processing and deep neural networks. *Results in Engineering*, 18, 101107. <https://doi.org/10.1016/j.rineng.2023.101107>
23. Rohitharun, S., Reddy, L. U. K., & Sujana, S. (2022). Image captioning using CNN and RNN. In IEEE 2022 2nd Asian Conference on Innovation in Technology (ASIANCON) (pp. 1–8).
24. Sasibhooshan, R., Kumaraswamy, S., & Sasidharan, S. (2023). Image caption generation using visual attention prediction and contextual spatial relation extraction. *Journal of Big Data*, 10(1), 18. <https://doi.org/10.1186/s40537-023-00615-1>

25. Scoparo, M., & Serapião, A. (2019). Deep learning para geração automática de legenda de imagem. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional* (pp. 551–562). Sociedade Brasileira de Computação.
26. Sehgal, L., & Mandan, S. (2022). Automated image capturing using CNN and RNN. *International Journal of Research in Engineering, Science and Management*, 5(1), 13–17.
27. Seshadri, M., Srikanth, M., & Belov, M. (2020). Image to language understanding: Captioning approach. *arXiv Preprint arXiv:2002.09536*.
28. Singh, A., & Gupta, S. (2023). Image-based action recognition and captioning using deep learning. In *Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing* (pp. 252–261).
29. Singh, A., & Vij, D. (2022). CNN-LSTM based social media post caption generator. In *IEEE 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM)* (v. 2, pp. 205–209).
30. Singh, V., Singh, A. S., & Anandhan, K. (2022). Image captioning using machine/deep learning. In *IEEE 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)* (pp. 849–854).
31. Suresh, K. R., Jarapala, A., & Sudeep, P. (2022). Image captioning encoder–decoder models using CNN-RNN architectures: A comparative study. *Circuits, Systems, and Signal Processing*, 41(10), 5719–5742. <https://doi.org/10.1007/s00034-022-01913-1>
32. Thangavel, K., et al. (2023). A novel method for image captioning using multimodal feature fusion employing mask RNN and LSTM models. *Soft Computing*, 27(19), 14205–14218. <https://doi.org/10.1007/s00542-023-07302-w>
33. Verma, A., et al. (2024). Automatic image caption generation using deep learning. *Multimedia Tools and Applications*, 83(2), 5309–5325. <https://doi.org/10.1007/s11042-023-17984-0>
34. Waheed, S. R., et al. (2023). CNN deep learning-based image to vector depiction. *Multimedia Tools and Applications*, 82(13), 20283–20302. <https://doi.org/10.1007/s11042-023-17459-5>
35. Yeshasvi, M., & Subetha, T. (2022). Image caption generator using machine learning and deep neural networks. In *Advances in Intelligent Computing and Communication: Proceedings of ICAC 2021* (pp. 137–144). Springer. https://doi.org/10.1007/978-981-16-9422-7_14