An Introductory Study about Image Captioning in Portuguese using Deep Learning

1st Ravi B.D. Figueiredo, 2nd Bruno H. L. dos Anjos, 3rd Richardson B.S. Andrade
Center of Informatics (CIN)
Federal University of Pernambuco (UFPE)
Recife, Brazil
{rbdf, bhlaf, rbsa2}@cin.ufpe.br

***Abstract*—Image caption is the field of study that generates sentences describing the content of a given image. Many solutions has been proposed recently, the majority in English. However, Portuguese image captioning works has been lacking, with at the moment of this article, none papers has been published. Our purpose is to create an image caption dataset in Portuguese language using the MSCOCO as base. Also, test some deep neural networks to generate sentences in Portuguese. The results were validated using BLEU metric.**

***Index Terms*—image captioning, deep learning, Portuguese language**

I. Introduction

The image caption is the process of creating an automatic description of the content of a natural image. This process is a combination between natural language process (NLP) and computer vision [1]. The task of image captioning can be applied in different fields like bio-medicine, market, education, and digital library [2]. The objective of the research in this area is to find the most efficient pipeline for an image, as an input, and transform its content into a sequence of words, creating connections between visual elements and textual, also maintaining the spelling norm.

Fusional language, like Portuguese, may be more complex when using end-to-end deep learning methods in comparison with more simple languages, like English [3]. Most of the research in the field is made in English, with some works applied in other languages. This can be explained by the low quantity of datasets in other languages [4]. A viable approach is building a resource by automatically translating the annotations from an existing dataset: this is much less expensive than manually annotating images, but of course, it leads to a loss of human-like quality in the language model [5].

This article presents some results in Image caption in Portuguese language using three architectures based on deep learning.

II. Related Works

Masotti et al. [6] presents an approach to create a dataset with 600 thousands images and caption in Italian. The model uses InceptionV3 [7] to extract the features of images and generates an embedding features, which is used as input to a LSTM network. The best result presented for BLEU-4 is 0.238.

Bartosiewicz et al. [3] creates an image caption dataset in Polish using two datasets as base. First is a translation of Flickr8k using the Azure translator API [8], the second dataset is the AIDe dataset [9]. InceptionV3 is used to extract features from images which feed a LSTM network to create the captions. The best result found was for BLEU of 0.33 for Flickr8k. For AIDe dataset, the results was nearly zero in all metrics. Even though, the results in this work are promising for data generated through automatic translation.

MISHRA et al. [10] propose the first known dataset by authors in Hindu. They have used the MSCOCO dataset translated using Google Translator, and after the translation was manually verified by human annotators and corrected accordingly to build the Hindi image captioning corpus. They develop a model using RESNET101 as the encoder and GRU as the decoder, also, have experimented with many different attention mechanisms, such as Visual attention, Bahdanau attention, and Luong attention.

As far as we know, this is the first attempt to create a image caption generator to Portuguese language.

III. PROPOSED METHODOLOGY

This work is divided into two parts. First, for the lack of dataset in Portuguese, the COCO dataset [11] is translated using the Google Translation API the is integrated into the *deep_translator* package [1]. The translation is made sentence by sentence to keep the context of individuals' captions. The COCO dataset is formatted to competitions, i.e., the image test does not have captions, then the validation set is used like the test set, and 10% of the train set is separated for the validation task. In the second part is described the three architectures used to generate image captions in Portuguese, VGG16 + LSTM, EfficientNet + Transformer Encoder-Decoder e CLIP + GPT2, and shows results with BLEU metrics. O resumo da metodologia aplicada pode ser vista na Figura 1.

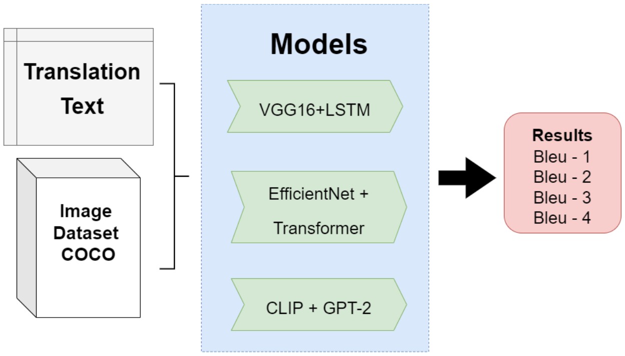

Fig. 1: Overview of workflow

For the experimental evaluation it was used python 3.7, GPU RTX 3060. For the first architecture, tensorflow 2.8 was used, and for the second, pytroch 1.11.

A. VGG16 + LSTM

This deep architecture is composed by encoder and decoder [12]. Commonly, the encoder utilizes Convolution Neural Network (CNN) to extract features and the decoder uses some variation of Recurrent Neural Networks (RNN) for sequence word generation [13]. The classic image captioning system encodes the image using pre-trained CNN and produces the hidden state. The hidden state is decoding by

[1] https://github.com/nidhaloff/deep-translator

Architecture	Blue-1	Bleu-2	Bleu-3	Bleu-4
VGG16 + LSTM	0.118	0.153	0.097	0.047
EfficientNet + Transformer	0.610	0.411	0.244	0.127
CLIP + GPT2	**0.664**	**0.492**	**0.329**	**0.209**

TABLE I: Performance of the models.

	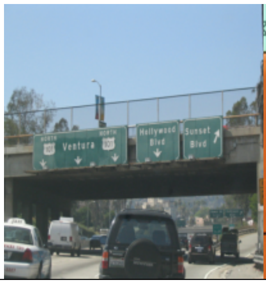	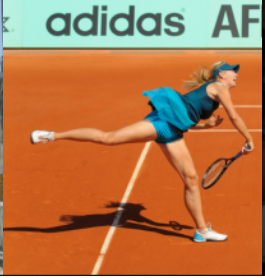	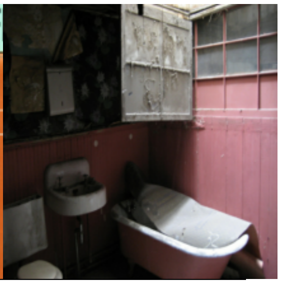
Ground Truth English	A street scene with focus on the street signs on an overpass.	a woman is striking a ball with a racket	A white bath tub sitting under a window.
Ground Truth	Uma cena de rua com foco nas placas de rua em um viaduto.	uma mulher esta batendo uma bola com uma raquete	Uma banheira branca embaixo de uma janela.
VGG16 + LSTM	Um Caminhão De Bombeiros Está Estacionado Em Um Campo.	Um Homem De Terno E Gravata.	Um Gato Está Sentado Em Uma Cadeira De Madeira.
EfficientNet + Transformer	Um carro esta dirigindo por uma rua movimentada.	Uma mulher de vestido azul e branco jogando tenis	Uma banheira branca ao lado de uma janela uma.
CLIP + GPT2	Uma estrada cheia de tráfego ao lado de um sinal de autoestrada.	Uma mulher batendo uma bola de tênis com uma raquete de tênis.	Um banheiro sujo com pia, vaso sanitário e banheira.

TABLE II: Results of 3 images in the our test set.

RNN and the decoder generates word to caption. In this work, we use pre-trained VGG-16 on the ImageNet dataset, as encoder, to extract visual features from the images (embedding) and Long Short-Term Memory (LSTM), as decoder, to receive as input the dense feature vector of image from encoder output and to translate the words in sequence to represent separately each element of image. In general, encoder-decoder model is very effective, but expensive to train in terms of time and hardware resources [6].

B. *EfficientNet + Transformer*

In this architecture is used EfficientNet [14] to extract the features from images, and transformers [15] to encode the features and decode the text. The objective of EfficientNet is to increase the scalability of the net based on the input, i.e., the number of layers and filters depends on the size of the image to better extract features of the image. We use an EfficientNet with pre-trained weights from ImageNet with weights freezer. After the extraction, the features are saved in an embedding layer with a size 1024. The transformers use the multi-head attention layers to encode, and other multi-head attention to decode, similar to Vaswani et al [15]. The overview of the architecture is presented in 2.

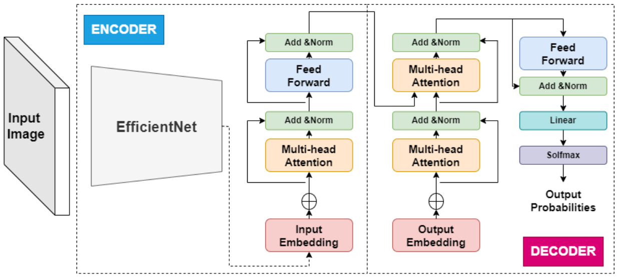

Fig. 2: Overview of the EfficientNet + Transformer

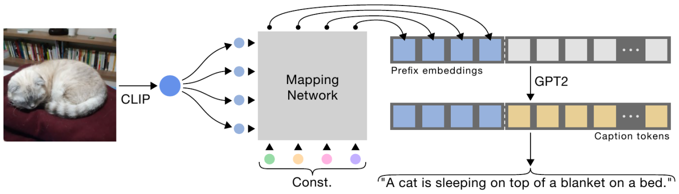

Fig. 3: Overview of the transformer-based architecture CLIP+GPT-2

C. *CLIP + GPT-2*

The overview of the architecture is presented in This architecture uses a CLIP (Contrastive Language-Image Pre-Training) encoder. This method produces a prefix for each caption by applying a mapping network over the CLIP embedding. These are fed to a language model, which is fine-tuned along with the mapping network training [16]. The mapping network, denoted F, maps the CLIP embedding to k embedding vectors:

$$p_1^i, p_2^i, \ldots, p_k^i = F(CLIP(x^i)). \quad (1)$$

Each vector p_j^i has the same dimension as a word embedding. The k embedding vectors are concatenated with the caption c^i:

$$Z^i = p_1^i, p_2^i, \ldots, p_k^i, c_1^i, c_2^i, \ldots, c_l^i. \quad (2)$$

Also, it use GPT-2 [17] as our language model, which has been demonstrated to generate rich and diverse texts. The language model is feed with the prefix-caption concatenation $\{Z^i\}_{i=1}^N$. The goal is predicting the caption tokens conditioned on the prefix in an autoregressive fashion [16]. For this, we train the mapping component F using the simple cross-entropy loss:

$$\mathcal{L}_X = -\sum_{i=1}^{N}\sum_{j=1}^{\ell} \log p_\theta\left(c_j^i \mid p_1^i, \ldots, p_k^i, c_1^i, \ldots, c_{j-1}^i\right). \quad (3)$$

For inference, the model generate the caption from the visual prefix, and predict the net tokens one by one, guided by the model output.

D. Dataset

We have created a Portuguese version of the MSCOCO dataset, which is one of the most used ones in image captioning. Its annotations are based on Common Objects in Context (COCO) dataset. The COCO dataset don't contain labels in the test dataset, so we split the train set in two, test set and validation set. 10% of the original test set is used in validation. The original validation set of the COCO is used as test set. The COCO dataset contain 530.874 captions for training, 58.986 for validation, and 5000 captions for testing. Each image have at least 5 caption describing its contents.

E. Evaluation Metric

We use bleu (Bilingual Evaluation Understudy Score) as an evaluation metric. Bleu analyzes each sentence against a set of reference sentences composed by humans themselves. An average score is computed to evaluate the overall quality of the generated text. The performance of BLEU metric depends on the generated text size and the number of references text. We use the BLEU-1, BLEU-2, BLEU-3 and BLEU-4 [18], BLEU-1 being 1-gram similarity, BLEU-2 is 2-gram, and so on. Mostly 4 grams range is the mostly used BLEU-4.

IV. Results and Discussions

In this section we report the results of generated captions using a qualitative and quantitative analysis.

In this work we use metrics BLEU $1, 2, 3$ and 4 for quantitatie analysis, the results is presented in I. The CLIP + GPT2 architectures presents the best result in all cases, even showing a 64.56% of improvement. The VGG16+LSTM show the worst results in all cases. This can be explained by that the transformers model shows a superior performance in comparison with LSTM.

In table II is presented some results for VGG16 + LSTM, CNN+Transformers and CLIP+GPT-2 architecture. For the CNN+Transformers model, the generated caption describe very well the image, but it can't create a well formed phrase in Portuguese language, e.g., it generate captions where sentence ends with article or prepositions. Nevertheless, it maintain cohesion between textual elements.

The CLIP+GPT2 model presents a good quality for syntax as seen in II. The caption generated don't show flaws in formulation of captions in Portuguese languages, meaning that it can learn about punctuation even not present,e.g., as can be seen in the white bath tub generated caption.

Both models generates caption that are within the scope of the image, but describe the elements in different ways, e.g., for the white tub image, both models identify the main elements,but the CLIP+GPT-2 describes more elements than the reference caption, resulting in a low BLEU score.

The VGG16 + LSTM presents the worst results for generated captions. The captions shows a good syntax and grammatical agreement, but is possible to see that the caption don't describe well the objects on the image. And so, the results is the result is far below comparing with others architectures

V. Conclusion

In this work we generate image captions in Portuguese language over MSCOCO dataset using different models. We create a new dataset translates from MSCOCO dataset with a good quality in a gramatical sense using Google Translate. The results shows that the CLIP+GPT2 have the best results in every scenario with an improvment up to 64% in comparison with the others architectures. The VGG16+LSTM shows the worst results in every cases.

For future works, we pretend to create more datasets in Portuguese like translate the Flickr30k. Have been created many image caption architectures and we pretend to test more of the to measure their performance in Portuguese language.

References

[1] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in *NeurIPS*, 2019.

[2] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys (CSUR)*, vol. 51, pp. 1 – 36, 2019.

[3] M. Bartosiewicz, I. Krupińska, M. Bany, A. Konieczna, M. Ostrowski, M. Zalewski, and M. Iwanowski, "Generating image captions in polish – experimental study," in *2021 14th International Conference on Human System Interaction (HSI)*, pp. 1–6, 2021.

[4] G. O. dos Santos, E. L. Colombini, and S. Avila, "pracegover: A large dataset for image captioning in portuguese," *Data*, vol. 7, no. 2, 2022.

[5] D. C. Caterina Masotti and R. Basili, "Deep learning for automatic image captioning in poor training conditions," in *Emerging Topics at the Fouth Italian Conference on Computational Linguistics*, 2018.

[6] C. Masotti, D. Croce, and R. Basili, "Deep learning for automatic image captioning in poor training conditions," *IJCoL. Italian Journal of Computational Linguistics*, vol. 4, no. 4-1, pp. 43–55, 2018.

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

[8] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[9] A. Wróblewska, "Polish corpus of annotated descriptions of images," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, (Miyazaki, Japan), European Language Resources Association (ELRA), May 2018.

[10] S. K. Mishra, R. Dhir, S. Saha, P. Bhattacharyya, and A. K. Singh, "Image captioning in hindi language using transformer networks," *Computers Electrical Engineering*, vol. 92, p. 107114, 2021.

[11] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014.

[12] Y. Goldberg, "A primer on neural network models for natural language processing," *Journal of Artificial Intelligence Research*, vol. 57, pp. 345–420, 2016.

[13] H. Wang, Y. Zhang, and X. Yu, "An overview of image caption generation methods," *Computational intelligence and neuroscience*, vol. 2020, 2020.

[14] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *CoRR*, vol. abs/1905.11946, 2019.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[16] R. Mokady, A. Hertz, and A. H. Bermano, "Clipcap: Clip prefix for image captioning," 2021.

[17] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2018.

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: A method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL ’02, (USA), p. 311–318, Association for Computational Linguistics, 2002.