A Comparative Evaluation of Transformer-Based Vision Encoder-Decoder Models for Brazilian Portuguese Image Captioning

Gabriel Bromonschenkel, Hilário Oliveira, Thiago M. Paixão
Programa de Pós-graduação em Computação Aplicada (PPComp)
Instituto Federal do Espírito Santo (IFES), Serra, Brazil
Email: gabriel.mota.b.lima@gmail.com, {hilario.oliveira, thiago.paixao}@ifes.edu.br

***Abstract*—Image captioning refers to the process of creating a natural language description for one or more images. This task has several practical applications, from aiding in medical diagnoses through image descriptions to promoting social inclusion by providing visual context to people with impairments. Despite recent progress, especially in English, low-resource languages like Brazilian Portuguese face a shortage of datasets, models, and studies. This work seeks to contribute to this context by fine-tuning and investigating the performance of vision language models based on the Transformer architecture in Brazilian Portuguese. We leverage pre-trained vision model checkpoints (ViT, Swin, and DeiT) and neural language models (BERTimbau, DistilBERTimbau, and GPorTuguese-2). Several experiments were carried out to compare the efficiency of different model combinations using the #PraCegoVer-63K, a native Portuguese dataset, and a translated version of the Flickr30K dataset. The experimental results demonstrated that configurations using the Swin, DistilBERTimbau, and GPorTuguese-2 models generally achieved the best outcomes. Furthermore, the #PraCegoVer-63K dataset presents a series of challenges, such as descriptions made up of multiple sentences and the presence of proper names of places and people, which significantly decrease the performance of the investigated models.**

I. Introduction

Humans possess the capacity to envision a visual scenario, comprehend and recognize the occurrence of existing elements, and grasp the connections between those elements. This capability relates to a computational task known as Image Captioning (IC), which lies at the intersection of Natural Language Processing (NLP) and Computer Vision (CV). IC focuses on generating natural language captions describing one or more images [1]. IC systems have several real-world applications, including generating captions for social media posts, automatic image indexing, supporting social inclusion for individuals with visual impairments, and guiding medical diagnostics through medical image descriptions [1], [2].

The IC research field has seen considerable improvements following the evolution of Deep Learning (DL) algorithms such as Long Short-Term Memory (LSTM) Networks, Convolutional Neural Networks (CNNs), Graph Neural Networks (GNNs), and Transformers [1]. From 2015 to 2022, there has been an increase in the adoption of attention-based approaches, particularly Transformer-based architectures [2]. Generally, these architectures follow a standard design consisting of a visual encoder and a natural language model decoder. The encoder learns a contextual encoded representation of the image content, while the decoder translates this representation into a natural language description [1], [2].

Developing IC models in low-resource languages (e.g., Portuguese) poses significant challenges compared to widely used languages (e.g., English). Most IC datasets are comprised of descriptions written in English, and previous research results point out that translating the generated captions to the desired language is less efficient and precise than training the image captioning language using a dataset on the desired dialect [3].

The Brazilian Portuguese is an example of an idiom poorly explored in the IC domain [3]. The lack of studies, models, and datasets in Brazilian Portuguese for IC significantly limits the development of applications involving translation tasks of images to natural language. These limitations hamper the training and evaluation of models specifically designed for the linguistic and cultural idiosyncrasies of Portuguese [3]–[5].

To tackle these issues, there is a need for advanced architectures that overcome the linguistic challenges of Portuguese, integrate state-of-the-art pre-trained language and vision models, and explore appropriate datasets to learn the linguistic nuances.

This work aims to contribute to the advancement of the Portuguese IC field through a comprehensive comparative analysis of different vision and language models combined in a fully Transformed-based visual encoder-decoder architecture, as illustrated in Figure 1. The pre-trained vision models of Vision Transformers (ViT) [6], Shifted Windows Transformer (Swin Transformer) [7], and Data-efficient Image Transformer (DeiT) [8] were investigated as encoders. We also evaluated the pre-trained natural language models, such as BERTimbau [9] and GPorTuguese-2 [10], as decoders.

Our main contributions are summarized as follows:

1) To the best of our knowledge, this is the first work to conduct a comprehensive experimental investigation on a fully Transformer-based visual encoder-decoder architecture for Brazilian Portuguese. By fine-tuning and comparing the performance of different options to encoder and decoder models, we provide insights into the most effective combinations for IC in Portuguese.
2) Our extensive evaluation was conducted on two datasets: the native Portuguese #PraCegoVer-63K and our trans-

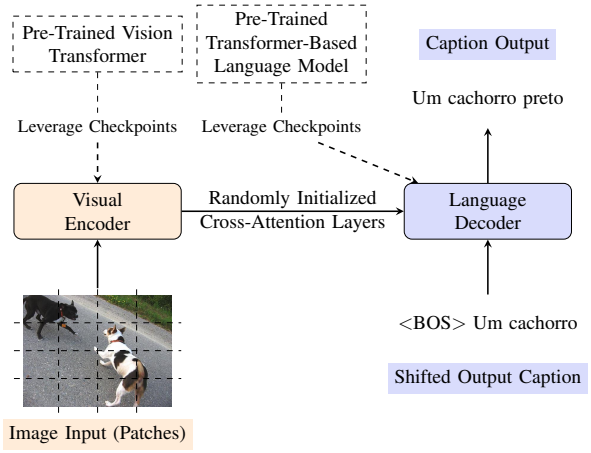


Fig. 1. Transformer-based visual encoder-decoder architecture.

lated version of the traditional Flickr30K. The performance of the models was evaluated using different automatic evaluation measures.

3) Our source code, Portuguese translated version of Flickr30K, and the models that achieved the highest performance are available at: github.com/laicsiifes/ved-transformer-caption-ptbr.

II. Related Work

The initial works on exploring deep learning architectures for IC are dated from 2015, and they use a CNN-based encoder and a LSTM-based decoder. In 2017, the Transformer-based models brought up a novel way of building language models. Since then, several architectures based on neural networks have been proposed and evaluated for the IC task [1], [2].

Traditionally, IC approaches are comprised of two main modules: the image encoder block and the natural language decoder block. During the image block processing, the input image is encoded into the model structure in a process called visual encoding. Then, the output of the encoder is passed to the natural language decoder as embeddings that represent the initial image. In the decoder block processing, a language model is responsible for decoding the image embeddings into a natural language caption that describes the image [1], [2].

For end-to-end Transformer-based architectures in IC, the adoption of a self-attention image encoder is prevalent. Typically, this kind of encoder makes use of an adaptation of Transformers for vision called ViT, in which the image patches[1] – or regions – are embedded to be processed by a Transformer-based block [11]. Novel approaches of Transformers for vision are being applied to IC, including the Swin Transformer method, which incorporates window self-attention and shifted window self-attention for image feature extraction [12].

In the context of the Brazilian Portuguese language, the earliest architecture explored in a native dataset for IC was an Attention-on-Attention Network (AoANet) [13] as reported by Santos et al. [3]. The AoANet uses a self-attention layer for image encoding and a LSTM-based layer for language decoding. It underwent training on a large multimodal dataset of paired images and reference captions proposed by the authors, which is derived from Instagram API[2] content tagged with #PraCegoVer. Gondim et al. [4] conducted a study on the fusion of a CNN and a Gated Recurrent Unit (GRU) network integrated with an attention mechanism for IC trained on a Portuguese translated version of Flickr8K dataset. Later, the approach of Alencar et al. [5] emerged as the first alternative for Brazilian Portuguese IC using a **G**rid–and **R**egion-based **I**mage Captioning **T**ransformer (GRIT), a Transformer-only neural architecture, trained on a Portuguese translated version of MSCOCO Captions dataset.

According to the Table I, our work addresses four points for Brazilian Portuguese IC not previously addressed [3]–[5]: **(1)** it uses several end-to-end Transformer-based architectures that **(2)** leverage checkpoints pre-trained in Brazilian Portuguese datasets, and the models built in this work are fine-tuned in **(3)** a native Brazilian Portuguese dataset and **(4)** a translated dataset.

TABLE I
Our work compared to other Brazilian Portuguese IC works.

Author(s)	Fully Transformer-Based	Leverage Pre-Trained Checkpoints	Compare Several Models	Brazilian Portuguese Dataset	Translated Dataset
[3]				✓	
[4]					✓
[5]	✓				✓
Ours	✓	✓	✓	✓	✓

III. Materials and Methods

This section presents the two datasets used in this work, as well as the encoder-decoder architecture adopted. The pre-trained visual and language models investigated are briefly described. Lastly, we describe the experimental design and evaluation metrics applied.

A. Datasets

a) #PraCegoVer-63K: This is the first large multimodal dataset for image captioning in Portuguese with freely annotated images [3]. It was built by gathering images from Instagram with the hashtag of the social project PraCegoVer. This social project aimed to include people with visual impairments, besides having an educational proposal. For such, users were encouraged to post images marked with the hashtag #PraCegoVer and write a brief description of their image content. There are two versions of the dataset, called #PraCegoVer-173K and #PraCegoVer-63K, which differ in terms of the number of images. In this work, we used the #PraCegoVer-63K subset, comprising 62,935 examples. A challenging feature of this dataset is the large variability in description length, with a mean of 37.9 and a standard deviation of 27.0 words per caption.

[1]Partitions with a fixed size.

[2]https://www.instagram.com/

b) Flickr30K: This dataset comprises 31,014 images paired with five descriptive captions provided by human annotators for each image [14]. The original English captions were translated into Portuguese using the Google Translator API[3], following an approach similar to related work in Portuguese image captioning [4].

Table II presents descriptive statistics of the Flickr30K and #PraCegoVer-63K datasets. The statistics on caption length (average and standard deviation) were generated using the SpaCy toolkit[4], with punctuation symbols removed. These statistics show that the reference captions in #PraCegoVer-63K are, in general, three times more extensive than those present in Flickr30K. In this work, Flickr30K and #PraCegoVer-63K follow the split reported by Karpathy et al. [15] and Santos et al. [3], respectively.

TABLE II
DESCRIPTIVE STATISTICS OF THE EVALUATION DATASETS.

	#PraCegoVer-63K		**Flickr30K**	
Split	Samples	Avg. Caption Length (Words)	Samples	Avg. Caption Length (Words)
Train	37,881	39.2 ± 27.6	29,000	12.1 ± 5.1
Validation	12,442	34.6 ± 26.0	1,014	12.3 ± 5.3
Test	12,612	37.4 ± 25.9	1,000	12.2 ± 5.4
Total	62,935	37.9 ± 27.0	31,014	12.1 ± 5.2

B. Transformer-based Vision Encoder-Decoder Model

For our experiments, we adopted a fully Transformer-based Vision Encoder-Decoder architecture, using pre-trained vision models as encoders and pre-trained language models as decoders. This kind of approach has been successfully applied in previous works involving image-to-text tasks, as demonstrated in [16].

The following three visual models were evaluated as the encoder.

- **ViT**: The Vision Transformer (ViT) was proposed to solve image classification tasks. ViT can learn from images by embedding them into patches and feeding a Transformer Encoder with these embeddings [6].
- **Swin**: The **S**hifted **Win**dow (Swin) Transformer was introduced as a hierarchical structure of the Transformer model, incorporating shifted windows to perform self-attention computations efficiently within distinct local windows [7].
- **DeiT**: The Data-efficient Image Transformer (DeiT) is a model designed for image classification tasks with a focus on efficient training and inference [8].

The encoders were adopted in their base architecture version, and they were pre-trained on the ImageNet1k dataset for the object recognition task. They accept images of size 224×224 as input.

[3]https://translate.google.com
[4]https://spacy.io/

For the natural language decoder, we evaluated two variations (a base and a distilled architecture version) of the Bidirectional Encoder Representations for Transformers (BERT) [17] and a small model version based on the Generative Pre-trained Transformer version (GPT).

- **BERTimbau**: BERTimbau is a BERT-based language model designed for Brazilian Portuguese pre-trained on brWaC [9].
- **DistilBERTimbau**: DistilBERTimbau is a small and fast model trained by distilling BERTimbau base architecture [18].
- **GPorTuguese-2**: GPorTuguese-2 is a causal language model for Portuguese based on the Generative Pre-training Transformer version 2 (GPT-2) small architecture. The model was pre-trained on Portuguese Wikipedia articles [10].

C. Evaluation Metrics

To assess the performance of the investigated models, we used five automatic evaluation measures commonly adopted in the literature [1]. These metrics compute the similarity between automatically generated descriptions and one or more reference descriptions.

a) CIDEr-D: The Consensus-based Image Description Evaluation (CIDEr) is calculated by measuring the cosine similarity between n-grams weighted based on the Term Frequency-Inverse Document Frequency (TF-IDF) [19].

b) BLEU-4: The Bilingual Evaluation Understudy (BLEU) is a metric for evaluating machine translation defined as the precision of the *n*-grams in a geometric average with a brevity penalty [20]. We adopted the BLEU-4 version that considers *n*-grams up to length 4.

c) ROUGE-L: The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is a collection of metrics commonly adopted to assess automatic text summarization approaches [21]. One of the most adopted ROUGE measures is the ROUGE Longest Common Subsequence (LCS), referred to as ROUGE-L. This metric computes the LCS between a candidate and one or more reference texts. Here, we adopted the f1-score computed using the ROUGE-L.

d) METEOR: The Metric for Evaluation of Translation with Explicit Ordering (METEOR) is a measure commonly used for evaluating machine translation. METEOR is a weighted harmonic mean of unigram precision and recall [22].

e) BERTScore: This metric computes the resemblance of candidate sentence tokens with reference sentence tokens using contextual embeddings extracted from the BERT model [23]. We computed the f1-score metric using the BERTScore and the BERTimbau model to generate the contextual embeddings.

Except for CIDEr-D, all other measures have normalized outcome values ranging from 0 to 1. For better textual conciseness, the measures CIDEr-D, BLEU-4, ROUGE-L, METEOR, and BERTScore will be abbreviated as C, B@4, RL, M, and BS, respectively.

D. Experimental Setup

Nine configurations of our Vision Encoder-Decoder model were evaluated, considering the three visual encoder options and the three language decoder models. All pre-trained models have public checkpoints available on the Hugging Face platform[5].

The models were fine-tuned on a computer equipped with a NVIDIA RTX 4090 GPU with 24GB of memory for 20 epochs. The learning rate was set to 5.0×10^{-5}, and due to memory constraints, the batch size was either 16 or 8, depending on the model size. The Beam Search algorithm was used to generate the captions, with the number of beams set to five. Based on an exploratory analysis of caption lengths in both datasets, the maximum caption length was set to 25 tokens for the Flickr30K and 70 tokens for the #PraCegoVer-63K.

A monitoring strategy was used during the fine-tuning epochs to avoid overfitting. This strategy involved applying the resultant model at the end of each epoch to the validation set and computing the ROUGE-L score. We use this measure during the fine-tuning because it is lightweight to compute and has been successfully adopted as an objective function in automatic text summarization models [24]. Throughout the epochs, only the model with the highest ROUGE-L value was saved as the best model. This best model was then used to generate captions in the test set.

IV. Experimental Results

This section presents the quantitative and qualitative results of the experiments. The experiments aimed to evaluate the following: ***(i)*** Which encoder and decoder configuration, among the assessed model options, achieves the best performance based on the adopted evaluation metrics and ***(ii)*** To better understand the models' performance, a qualitative inspection was conducted on some examples (images and reference captions) and the captions generated by the models in both datasets.

A. Encoder-Decoder Models Assessment

Table III presents the results of the experiments using the Flickr30K dataset. Configurations using $\text{Swin}_{\text{BASE}}$ as an encoder achieved the best performance in all evaluation metrics. The best performance was obtained by the model employing $\text{Swin}_{\text{BASE}}$ (encoder) and $\text{DistilBERT}_{\text{BASE}}$ (decoder). An important point to highlight is that the $\text{DistilBERT}_{\text{BASE}}$ model has around 66 million parameters, which represents a reduction of approximately 40% compared to the $\text{BERT}_{\text{BASE}}$ model's 110 million parameters.

The performance metrics for the #PraCegoVer-63K dataset were significantly lower than those achieved with the Flickr30K dataset, as shown in Table IV. The $\text{GPT-2}_{\text{SMALL}}$ decoder achieved the best results, and the $\text{Swin}_{\text{BASE}}$ model as an encoder showed better performance in almost all evaluation metrics, except for the METEOR (M) score, where the ViT_{BASE} model as an encoder obtained the top result.

[5]https://huggingface.co/

TABLE III
Evaluation results (%) for the Flickr30K Portuguese dataset. The three best results in each metric are highlighted in bold, and the highest performance is indicated with a †.

Encoder	Decoder	C	B@4	RL	M	BS
	$\text{BERT}_{\text{BASE}}$	49.53	19.20	36.00	39.80	69.58
$\text{DeiT}_{\text{BASE}}$	$\text{DistilBERT}_{\text{BASE}}$	50.58	19.24	35.77	39.93	69.50
	$\text{GPT-2}_{\text{SMALL}}$	50.61	19.83	36.30	40.52	69.66
	$\text{BERT}_{\text{BASE}}$	**62.42**	**22.78**	**38.71**	**43.47**	**71.19**
$\text{Swin}_{\text{BASE}}$	$\text{DistilBERT}_{\text{BASE}}$	**66.73**†	**24.65**†	**39.98**†	**44.71**†	**72.30**†
	$\text{GPT-2}_{\text{SMALL}}$	**64.71**	**23.15**	**39.39**	**44.36**	**71.70**
	$\text{BERT}_{\text{BASE}}$	57.32	22.12	37.50	41.72	70.63
ViT_{BASE}	$\text{DistilBERT}_{\text{BASE}}$	59.32	21.19	37.74	42.70	71.15
	$\text{GPT-2}_{\text{SMALL}}$	59.02	21.39	37.68	42.64	71.03

TABLE IV
Evaluation results (%) for the #PraCegoVer-63K dataset. The three best results in each metric are highlighted in bold, and the highest performance is indicated with a †.

Encoder	Decoder	C	B@4	RL	M	BS
	$\text{BERT}_{\text{BASE}}$	0.99	0.00	4.02	3.49	36.20
$\text{DeiT}_{\text{BASE}}$	$\text{DistilBERT}_{\text{BASE}}$	1.59	0.11	9.22	7.74	45.36
	$\text{GPT-2}_{\text{SMALL}}$	**5.95**	**1.00**	**12.44**	**13.87**	**49.11**
	$\text{BERT}_{\text{BASE}}$	1.29	0.00	4.53	3.90	30.84
$\text{Swin}_{\text{BASE}}$	$\text{DistilBERT}_{\text{BASE}}$	0.31	0.01	7.91	5.76	40.95
	$\text{GPT-2}_{\text{SMALL}}$	**9.45**†	**1.60**†	**13.43**†	**15.58**	**49.85**†
	$\text{BERT}_{\text{BASE}}$	0.83	0.00	3.03	2.61	27.69
ViT_{BASE}	$\text{DistilBERT}_{\text{BASE}}$	1.70	0.12	9.01	7.89	45.71
	$\text{GPT-2}_{\text{SMALL}}$	**8.27**	**1.49**	**13.23**	**15.74**†	**49.57**

The very low performance achieved by the models on the PraCegoVer dataset is consistent with the results obtained by their original authors [3], and these poor results can be attributed to the complexity of the dataset. For instance, the reference captions are very long and highly variable compared to commonly used image captioning datasets like Flickr30K. Additionally, Bencke et al. [25] identify linguistic errors in the image descriptions, which can degrade model performance. Other significant issues, such as the context of some captions, also contributed to the low outcomes, as discussed in the following section.

B. Qualitative Evaluation

This section presents a qualitative evaluation of some images and generated captions from both datasets. Four images (two from each dataset) were selected to assist the discussions. These images illustrate scenarios where the best models, Swin-DistilBERT (Flickr30K) and Swin-GPT-2 (#PraCegoVer-63K), generated captions with high and low values in the evaluation metrics.

On Flickr30K, a more nuanced analysis focusing on the generated captions reveals a complex challenge the models faced in dealing with the gender of the people in some pictures, for instance, "Uma garota (a girl)", as illustrated in Figure 2a. Moreover, the models generally produced very concise

descriptions in comparison with the reference, often neglecting specific details, such as the color of the person's shirt. Regardless of this issue, the model consistently maintained the reasoning about the main event in the picture, e.g., "jogando futebol" (playing soccer).

Still concerning Flickr30K, when analyzing other images, as depicted in Figure 2b, the model generated the caption precisely as one of the reference captions. The model has not mistaken crucial details, e.g., collective, gender, entities, or actions. Overall, the models produced high-quality outcomes in this kind of image on Flickr30K.

Figures 2c and 2d show samples from #PraCegoVer-63K, showcasing low and high performance, respectively, for Swin-GPT-2 (the best encoder-decoder configuration, according to the evaluation metrics). Analyzing the generated captions and seeking to understand the low performance obtained by the evaluated models, we observed several challenging scenarios for the image captioning task. The first challenge was related to the length of the reference captions (37.4 ± 25.9) on the test set. In contrast, the captions generated by Swin-GPT-2, which produced the longest descriptions, had an average length of 27.04 with a standard deviation of 13.24. Therefore, the captions generated by the models are much smaller than the reference ones.

The second challenge was the text content in the evaluated images, as in the example in Figure 2d. In some cases, the models were able to recognize the text content and also other elements from the image and use this information to compose the caption. However, the high complexity of the text content observed in many samples hindered proper recognition, thereby preventing the use of this information to generate the descriptions.

The third and most significant challenge arose from captions that include information not self-contained within the image, thereby requiring external or world knowledge. For instance, the caption in Figure 2c mentions "Terceira Ponte" and a toll booth. For the model or even a person to generate this description, they would need to associate the image with the specific name of a bridge located in Vitória–ES and recognize that it has a toll booth. Similarly, we observed descriptions that mention the proper names of individuals and specific places.

To quantify the percentage of name references for individuals and places, we applied a Named Entity Recognition (NER) pipeline available in the spaCy tool to the reference captions. We computed the total number of mentions of Person (PER) and Location (LOC) entities. In total, 35,514 mentions of people were identified in 22,415 captions, representing 35.62% of the descriptions mentioning at least one person. For location entities, 49,011 mentions were identified in 26,995 captions (42.89%).

Therefore, smaller models without external knowledge or auxiliary resources, like Optical Character Recognition (OCR), are not viable options for image captioning in #PraCegoVer-63K.

V. Conclusion

This work conducted a comprehensive evaluation of a fully Transformer-based Visual encoder-decoder architecture for Portuguese image captioning. Several experiments were performed to assess different encoder and decoder combinations using public pre-trained visual and language models. Two datasets, Flickr30K (translated) and #PraCegoVer-63K (native), were used to generate captions, which were evaluated using traditional automatic evaluation metrics. The experimental results demonstrated that the Swin-DistilBERT and Swin-GPT-2 models achieved the best performance on Flickr30K and #PraCegoVer-63K, respectively.

The performance of the models in the Flickr30K dataset demonstrates that they were generally able to generate concise and descriptive image captions. However, there is still significant room for improvement. On the other hand, the performance of the models in #PraCegoVer-63K was very low but comparable to those obtained by the original authors of the dataset. #PraCegoVer-63K proved to be a challenging dataset for image caption models, requiring external knowledge to generate some captions accurately.

Future work will expand the study with new Portuguese datasets, such as FM30K [26], and investigate the use of multimodal Large Language Models (LLMs), such as Phi-3 Vision [27] and PaliGemma [28]. Additionally, we aim to evaluate the application of simplification and anonymization strategies to remove mentions of names of people and places in #PraCegoVer captions.

Acknowledgment

The authors would like to thank FAPES/UnAC (Nº FAPES 1228/2022 P 2022-CD0RQ, Nº SIAFEM 2022-CD0RQ) for the financial support given through the UniversidadES system.

References

[1] H. Sharma and D. Padha, "A comprehensive survey on image captioning: from handcrafted to deep learning-based techniques, a taxonomy and open research issues," *Artificial Intelligence Review*, pp. 1–43, 2023.

[2] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From show to tell: A survey on deep learning-based image captioning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 539–559, 2022.

[3] 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. [Online]. Available: https://www.mdpi.com/2306-5729/7/2/13

[4] J. Gondim, D. B. Claro, and M. Souza, "Towards image captioning for the portuguese language: Evaluation on a translated dataset." in *ICEIS (1)*, 2022, pp. 384–393.

[5] R. S. de Alencar, W. A. C. Castañeda, and M. Amadeus, "Image captioning for brazilian portuguese using grit model," *arXiv preprint arXiv:2402.05106*, 2024.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

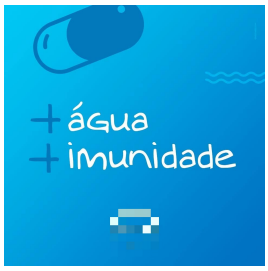
[7] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[8] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

(a) **References:** “Um jovem vestindo jeans e camiseta está sentado na grama, com uma bola no ar.”;
“Um homem de camisa vermelha está sentado na grama e uma bola voa em sua direção.”;
“Um homem de camisa rosa está sentado na grama e uma bola está no ar.”;
“Um homem está deitado na grama de um lindo parque.”;
“Um homem sentado na grama enquanto uma bola passa voando.”.
Generated: “Uma garota sentada no chão jogando futebol.”.

(b) **References:** “Um grupo de homens reunidos em torno de uma mesa se preparando para jogar cartas.”;
“Grupo de homens sentados em volta de uma mesa conversando.”;
“Vários homens idosos estão agrupados em torno de uma mesa.”;
“Um grupo de homens está sentado ao redor de uma mesa.”;
“Idosos reunidos em torno de uma mesa.”.
Generated: “Um grupo de homens está sentado ao redor de uma mesa.”.

(c) **Reference:** “Fotografia aérea sobre o pedágio da Terceira Ponte. A foto contém alguns prédios, um pedaço da Terceira Ponte e o fluxo de carros.”.
Generated: “Foto aérea da orla da Lagoa dos Ipês.”.



(d) **Reference:** “Em um fundo azul, está a frase “Mais água, mais imunidade” com um símbolo de uma pílula de vitaminas.”.
Generated: “Em um fundo azul, está a frase “ Mais água, mais imunidade ” com um símbolo de uma pílula de vitaminas.”.

Fig. 2. Example of images where Swin-DistilBERT (in Flickr30K) generates captions with (a) low and (b) high score values in the evaluation metrics. (c) and (d) similarly for Swin-GPT-2 (in #PraCegoVer-63K).

[9] F. Souza, R. Nogueira, and R. Lotufo, “Bertimbau: pretrained bert models for brazilian portuguese,” in *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*. Springer, 2020, pp. 403–417.

[10] P. Guillou, “Gportuguese-2 (portuguese gpt-2 small): a language model for portuguese text generation (and more nlp tasks...),” 2020.

[11] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” *arXiv preprint arXiv:2301.12597*, 2023.

[12] J. C. Hu, R. Cavicchioli, and A. Capotondi, “Expansionnet v2: Block static expansion in fast end to end training for image captioning,” *arXiv preprint arXiv:2208.06551*, 2022.

[13] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, “Attention on attention for image captioning,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4634–4643.

[14] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[15] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[16] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei, “Trocr: Transformer-based optical character recognition with pre-trained models,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 13 094–13 102.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[18] R. Silva Barbon and A. T. Akabane, “Towards transfer learning techniques—bert, distilbert, bertimbau, and distilbertimbau for automatic text classification from different languages: a case study,” *Sensors*, vol. 22, no. 21, p. 8184, 2022.

[19] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, “Cider: Consensus-based image description evaluation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[20] 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 of the Association for Computational Linguistics*, 2002, pp. 311–318.

[21] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Text summarization branches out*, 2004, pp. 74–81.

[22] S. Banerjee and A. Lavie, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[23] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with bert,” *arXiv preprint arXiv:1904.09675*, 2019.

[24] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, “Pegasus: Pre-training with extracted gap-sentences for abstractive summarization,” in *International conference on machine learning*. PMLR, 2020, pp. 11 328–11 339.

[25] L. Bencke, F. V. Pereira, M. K. Santos, and V. Moreira, “Inferbr: A natural language inference dataset in portuguese,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 9050–9060.

[26] M. Viridiano, A. Lorenzi, T. T. Torrent, E. E. Matos, A. S. Pagano, N. S. Sigiliano, M. Gamonal, H. de Andrade Abreu, L. V. Dutra, M. Samagaio *et al.*, “Framed multi30k: A frame-based multimodal-multilingual dataset,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 7438–7449.

[27] M. Abdin, S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, H. Behl *et al.*, “Phi-3 technical report: A highly capable language model locally on your phone,” *arXiv preprint arXiv:2404.14219*, 2024.

[28] X. Chen, X. Wang, L. Beyer, A. Kolesnikov, J. Wu, P. Voigtlaender, B. Mustafa, S. Goodman, I. Alabdulmohsin, P. Padlewski *et al.*, “Pali-3 vision language models: Smaller, faster, stronger,” *arXiv preprint arXiv:2310.09199*, 2023.