Analysis of Machine Translators on Sentences Generated by Portuguese *Image Captioning Models*

Natan Moura[1], João Medrado Gondim [1], Daniela Barreiro Claro[1], Babacar Mane[1],

[1]FORMAS Research Center on Data and Natural Language –
Institute of Computing – Federal University of Bahia (UFBA) – Salvador - Bahia - Brazil
{natan.moura, joao.gondim, dclaro, babacarm}@ufba.br

Abstract

Recent works in the fields of computer vision and natural language processing have enabled the recognition and identification of objects in images, generating automatic descriptions. Despite these advancements, the main research in this field is primarily related to the English language, requiring some adaptation when dealing with other languages, such as Portuguese. One of these methods is the translate-train approach, which involves translating the training dataset into the desired language. However, there are various translators with different levels of effectiveness available. The primary objective of this work is to evaluate the behavior of image captioning models when trained on datasets translated into Portuguese by different automatic translators, both quantitatively (cost, training time, metrics on the test set) and qualitatively (comparative evaluation form, error analysis). The results indicate that it is possible to obtain valid automatic descriptions in Portuguese from image captioning models trained on translated datasets, and that more robust translators produce more meaningful descriptions.

1 Introduction

The human capacity to observe the external world and to communicate such observations to others including the ability to articulate what is perceived is an intrinsic and routine aspect of human cognition. However, replicating these behaviors in computational systems remains a substantive challenge, as it presupposes sophisticated integration of methods from both Natural Language Processing (NLP) and Computer Vision (El-Komy et al., 2022).

NLP encompasses a broad set of theoretical frameworks and computational techniques designed to analyze, interpret, and represent naturally produced linguistic expressions. These methods aim, either implicitly or explicitly, to approximate the cognitive processes underlying human language understanding across diverse tasks and applications (Liddy, 2001). Computer Vision, conversely, is primarily concerned with endowing machines with the capacity to perceive, interpret, and reason about visual stimuli in a manner analogous to human visual cognition (El-Komy et al., 2022).

Both domains necessarily involve evaluative and interpretative mechanisms: NLP requires systems to infer meaning, structure, and context from text, while Computer Vision demands the recognition of three-dimensional form, spatial configuration, and the visual appearance of objects in images (Szeliski, 2010). Consequently, the intersection of these fields reflects a broader endeavor to computationally model complex human perceptual and communicative competencies.

Recent demands for the joint interpretation of both images and text have led to sustained interaction between these two research areas, fostering the development of methods capable of addressing such multimodal needs (Bernardi et al., 2016). Among the various tasks that emerge from this intersection is image captioning (IC), which enables the automatic generation of textual descriptions for images, including those present in videos or online posts, thereby providing increased accessibility for individuals with visual impairments (dos Santos et al., 2023).

Despite the significant progress achieved in IC, much of the existing research remains predominantly focused on the English language as stated by (Gondim et al., 2025). Only a limited number of studies address Portuguese, and most available techniques and models are originally developed for English. One strategy to mitigate this limitation is the application of adaptation techniques to existing methods, such as translate-train, which consists of translating the training dataset and subsequently training a model specifically tailored to the target language; zero-shot, which leverages the abundance of datasets and models in a source lan-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 360–368

guage to make predictions in a target language; and translate-infer, in which inferences are generated in the source language and later translated into the target language (Artetxe et al., 2020; Rosa et al., 2021).

Given the requirement for a multilingual base model, the zero-shot approach was excluded, and the translate-infer strategy was also disregarded due to its ongoing computational cost, since translation occurs directly during inference. By contrast, translate-train entails a finite cost (Rosa et al., 2021) and has demonstrated promising results in the field of image captioning (Vishnu Kumar and Lalithamani, 2022), and was therefore adopted in the present study.

Alternative architectural strategies can also be considered in multilingual image captioning systems. For example, captions may be generated in the source language and translated only at inference time (translate-infer), multilingual captioning models may be trained directly on mixed-language datasets, or multiple translated versions of the training corpus may be combined to mitigate translator-specific artifacts. The translate-train approach was selected in this study due to its simpler implementation and its finite computational cost compared with alternatives that require translation during inference.

Building upon the translate-train technique as a foundational approach, the present study aims to compare both free (LibreTranslate Groq) and paid (Google Cloud Translation) machine translation systems that support translation into Portuguese. Accordingly, the primary objective of this work is to assess how a model trained on a translated version of the Flickr30k dataset into Portuguese is affected in terms of the captions it generates. The analysis involves identifying monetary cost, translation time, caption quality, and performance metrics on the test set, with the goal of determining the strengths and limitations of each translation method and the key factors to consider when employing them within a translate-train workflow.

Based on the criteria established in this evaluation, it was observed that translating a dataset from English to Portuguese using different translation systems results in distinct corpora, each yielding models capable of producing valid inferences. These findings suggest that translate-train is a viable technique for the image captioning task and that the effort invested in dataset translation directly influences the quality of the resulting inferences. To the best of our knowledge, this is the first study to conduct a translation-focused evaluation within the context of image captioning.

This article is organized as follows. Section 2 presents the related work; Section 3 details the methodology; Section 4 describes the experiments performed; Section 5 reports the results; Section 6 provides a discussion; and Section 7 concludes the paper.

2 Related Works

This section introduces related work in the field of image captioning, categorized according to dataset translation, the development of IC models for Portuguese, and the comparison of multilingual IC methods.

Regarding the translation of datasets for training image captioning models, the authors of (Vishnu Kumar and Lalithamani, 2022) employ, in one of their methods, a dataset translated from English into Tamil using the Google Translate API, and highlight the promising potential of generating models from translated datasets. They note that, despite obtaining relatively low metric scores, the model produced through this method was still capable of capturing meaningful details in images.

With respect to the development of Portuguese image captioning models using translated datasets, the authors of (Gondim et al., 2025) propose the construction of a Portuguese-translated version of the Flickr8k dataset using LibreTranslate. They apply a custom architecture combining an attention model, a transformer, and a classifier to generate a Portuguese IC model. For evaluation, they employ the METEOR (Lavie and Agarwal, 2007) and BLEU (Papineni et al., 2002) metrics, human assessment, and a comparison of results with a model trained on the original dataset.

Concerning the comparative effectiveness of transfer learning methods, the studies presented in (Rosa et al., 2021) and (Artetxe et al., 2020) primarily focused on the task of question answering analyze the performance (in terms of metrics) of each method relative to the others, as well as the associated costs. Their findings indicate that, for the translate-train approach, differences arise depending on whether translations are produced by commercial or open-source systems (Rosa et al., 2021). They further observe that translations introduce unintentional artifacts that warrant more careful examination (Artetxe et al., 2020).

In contrast to these approaches, the present work proposes to evaluate different machine translation systems in the construction of Portuguese datasets derived from Flicker30k, as well as the quality of the sentences generated by models trained on these respective datasets. Similar to (Vishnu Kumar and Lalithamani, 2022), translations are performed from English into another language—Portuguese, in this case. Unlike (Gondim et al., 2022), no modifications are made to the underlying image captioning system; instead, adaptations are restricted to the data input, with a broader variety of translation tools being examined. This study is specifically designed for the image captioning task, which exhibits characteristics distinct from those of question answering (Rosa et al., 2021)(Artetxe et al., 2020).

3 The process of generating sentences in Portuguese through translation

The process of generating Portuguese sentences through translation, corresponding to the image descriptions, comprises several stages illustrated in Figure 1. It begins with the dataset to be translated in this study, the Flicker30k dataset. Subsequently, the workflow of the translation algorithm is described for the three translation systems considered (LibreTranslate, Groq, and Google Cloud Translation). The next stage details the model selected for the experiments (AoANet (Huang et al., 2019)), followed by the presentation of the models trained on each of the translated datasets.

3.1 Dataset - *Flickr30k* (Young et al., 2014)

The dataset used in this study is Flickr30k (Young et al., 2014), a collection comprising 31,783 images depicting everyday activities (sourced from the Flickr platform) and 158,915 captions (five per image) obtained through collaborative annotation. The key characteristics of this dataset include its relatively large size, diversity, and its widespread use as a baseline dataset particularly in its smaller version, employed in (Gatt and Krahmer, 2018). Moreover, it is fully compatible with the AoANet architecture adopted in this work.

The preprocessing stage consisted of translating Flickr30k from English into Portuguese using the previously mentioned translation systems. The translation was carried out automatically through Python-based algorithms, which are illustrated through the corresponding flowchart.

3.2 Flowchart of the Translation Algorithm

The translation flowchart is based on an algorithm that operates independently of the translation service selected. The output structure corresponds to a JSON file containing a layout similar to the original Flickr30k dataset, but with all captions already translated. The translation pipeline can be described through the process illustrated in Figure 2.

This flowchart also defines a control file that records which sentences have been translated, ensuring a checkpoint mechanism in case the translation process is interrupted.

3.3 AoANet — Model Architecture for Image Captioning

The baseline architecture used for generating models responsible for producing image captions was the Attention on Attention (AoANet) (Huang et al., 2019). This architecture enhances existing attention mechanisms by performing the attention operation a second time, in order to measure the relevance of the attention context produced in the initial pass. The choice of AoANet was motivated by its performance metrics on the MS COCO (Lin et al., 2014) test set and by the availability of the authors' implementation.

3.4 Automatics Translators

The three translators were selected based on distinct criteria to enable a more comprehensive analysis of the generated sentences. LibreTranslate[1], which is built upon the OpenNMT model (Klein et al., 2017), supports English–Portuguese translation, offers a self-hosted API, and has previously been employed in image captioning research (Gondim et al., 2025). Groq[2] was included because it provides translation capabilities through various Large Language Models (LLMs). In this study, the model selected for Groq was llama3-8b-8192, chosen for the quality of its outputs and its high token throughput (30,000 tokens per minute). Finally, the Google Cloud Translation API (Google, 2008), a commercial translator provided by Google, performs text translation via an API and is priced according to the number of characters sent. It is one of the most established automatic translation tools available. The Google model used corresponds to a Neural Machine Translation (NMT) system (Wu

[1]www.libretranslate.com
[2]www.groq.com

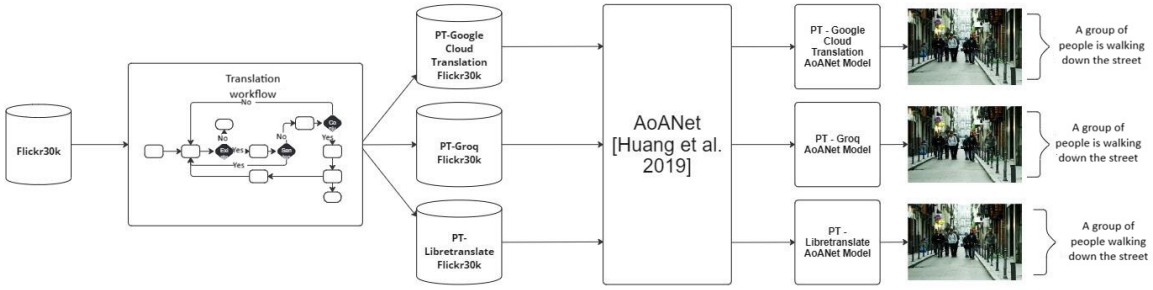

Figure 1: Architecture

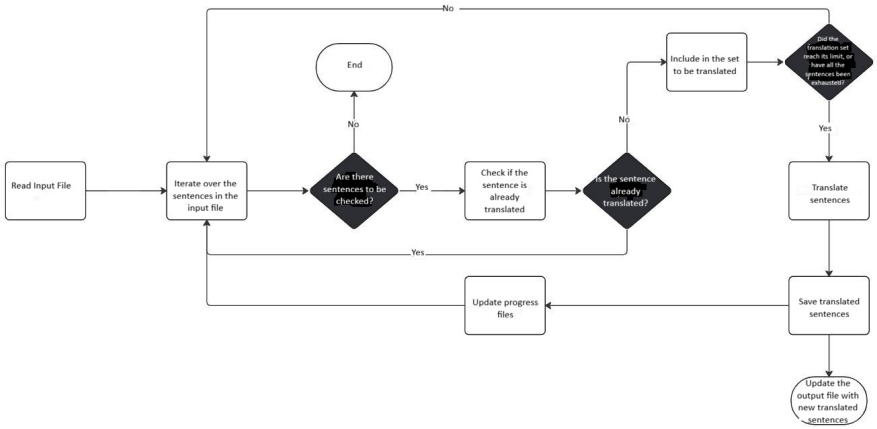

Figure 2: Flow diagram of the translation algorithm

et al., 2016).

Specific characteristics of the translation methods. Certain characteristics were defined for each translation method: both LibreTranslate and Cloud Translation follow a similar procedure in which sentences are submitted to the translator's API and the corresponding translated outputs are retrieved. Groq (Groq, 2016), in contrast, operates as an LLM and therefore requires the use of a prompt.

4 Experiments

Before conducting the experiments proposed in this study, the AoANet model (Huang et al., 2019) was reproduced using the Flickr30k dataset (Young et al., 2014) in English in order to obtain a baseline under identical hyperparameter settings. The features used were the same pre-extracted bottom-up features, which are more lightweight and require less storage than those extracted with ResNet (He et al., 2016). Training was performed for 15 epochs on one of the Google Colab Pro environments, equipped with 64 GB of RAM and L4 GPUs. Access to Google Colab Pro was made available for the experiment.

The experiments conducted with the three models generated in Portuguese followed exactly the same hyperparameters used for the English baseline, differing only in the translated datasets produced by each translation method. For financial reasons, the models were not trained until full convergence, as the computational cost required to do so was prohibitively high. Two categories of experiments were designed in this study: a quantitative evaluation, which includes translation cost, translation time, and standard captioning metrics; and a qualitative evaluation, based on human assessment of caption quality through an evaluation form, as illustrated in Figure 3.

Figure 3: Illustration of a question included in the evaluation form

4.1 Experiment 1 – Quantitative Analysis

The quantitative analysis employed standard evaluation metrics commonly used in the validation of image captioning models and also supported by the AoANet implementation: BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), and CIDEr (Vedantam et al., 2014). These metrics were calculated on the Flickr30k test set. Additional quantitative factors recorded included the monetary cost of translation and the total time required to translate the dataset.

4.2 Experiment 2 – Qualitative Analysis

The qualitative analysis is conducted through a survey designed to evaluate the semantic accuracy of the captions generated by the Portuguese image captioning models. Each question presented an image from the test set together with multiple captions automatically produced by models trained on datasets translated using different machine translation systems.

A total of 30 images were randomly selected using a Python algorithm to be included in the survey. This procedure prevented the manual selection of only favorable examples and ensured a more representative sample of the model outputs. In addition, the order of the answer choices was automatically randomized to avoid positional bias during the selection process.

Annotators are asked to identify which sentence best described the presented image. To capture cases in which the generated captions failed to represent the visual content, an additional response option was included in each question:

- None of the sentences partially or fully describe the image;
- Evaluators were instructed to rely on their own judgment regarding the semantic correspondence between the captions and the image when selecting their answers.

The collected responses are aggregated using a scoring scheme in which the caption most frequently selected by the evaluators for each image received one point for the translation system responsible for generating that description. In cases where both Google and Groq produced identical captions that are jointly selected as the most appropriate description, both systems received a point simultaneously.

Two scoring schemes are considered: one including all evaluated images, even when the majority of evaluators indicated that none of the captions are adequate, and another restricted to images for which the majority of evaluators identified at least one caption as partially or fully compliant with the visual content.

The survey are implemented using Google Forms, responses were collected anonymously, and the questionnaire are distributed through messaging applications.

5 Results

5.1 Quantitative Analysis

The metrics selected for this experiment were METEOR, BLEU (only BLEU-1[3]), and CIDEr. The results for the Portuguese models and the English baseline model are presented in Table 1.

Regarding translation costs for the dataset, neither LibreTranslate nor Groq incurs a direct finan-

[3]Results for BLEU-2, BLEU-3, and BLEU-4 presented negative exponentiation and were therefore considered irrelevant

Table 1: Test set metrics and translation time

Model	BLEU	METEOR	CIDEr	Time
English	0.383	0.09	0.056	Not applicable
Google Cloud	0.007	0.03	0	5h 40m
Groq	0.009	0.03	0	24h
LibreTranslate	0.036	0.09	0	8h 30m

cial cost; however, both require computational resources and the corresponding energy consumption for running the translation algorithms. In the case of Groq, because it is a third-party hosted service, there are no usage charges, although request limits apply. LibreTranslate, by contrast, must be executed locally and imposes no usage limits. Google Cloud Translation, on the other hand, has a cost associated with its NMT-based model. It is free provided that a Google Cloud subscription is active for up to 500,000 characters per month, with an additional cost of 20 USD per million characters beyond this limit. Translating Flickr30k required approximately 9.7 million characters, resulting in an approximate cost of 180 USD.

Translation time varies considerably, as shown in Table 1. The 24 hours runtime required for Groq was distributed over three days due to daily request limits.

5.2 Qualitative Analysis

The survey received a total of 88 anonymous responses. With respect to the conformity of the automated descriptions for the 30 images, 56.7% (17) of the images had at least one sentence that, according to the evaluators, described them, while 43.3% (13) did not. In the cases where the majority selected non-conformity, there was no unanimity in choosing the option "None of the sentences partially or fully describes the image". Thus, at least one evaluator considered one of the captions to be partially or fully aligned with the image (Figure 4

Regarding the scoring of the translation tools, Google Cloud Translation emerged as the most frequently selected option, regardless of caption conformity, with a score of 16, followed by Groq with 9 and LibreTranslate with 7 as depicted in Figure 5. The total exceeds 30 because, in two of the three cases where Google and Groq produced identical translations, the corresponding caption was the most frequently chosen.

Considering only the cases in which the majority of evaluators selected at least one sentence that

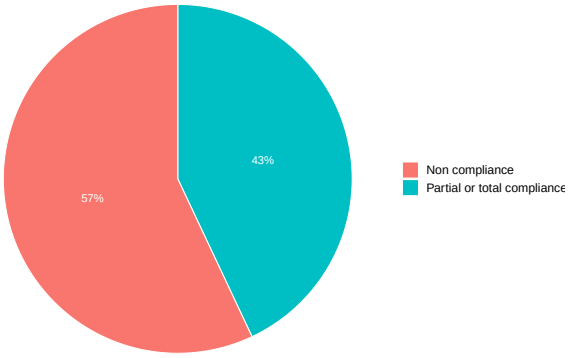

Figure 4: Distribution of caption-to-image conformity

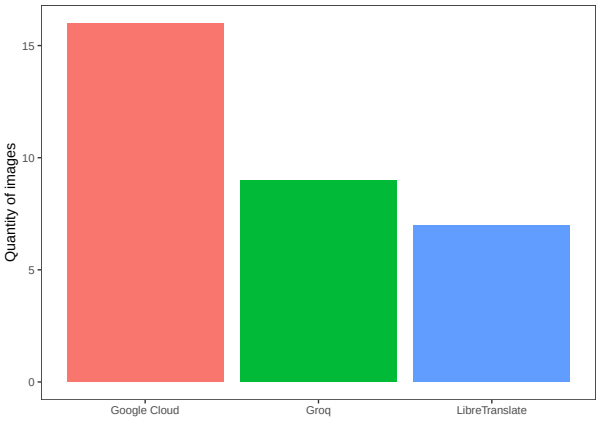

Figure 5: Translator scores (independent of caption conformity)

partially or fully described the image, the ordering remains the same, but the scores change to 10, 6, and 3, respectively. Again, the total exceeds 17 due to the previously mentioned overlap, as depicted in Figure 6.

6 Discussion

6.1 Quantitative and Qualitative Aspects

The first quantitative aspect to be examined concerns the financial constraints, which prevented the models from being trained until loss stabilization due to the high training cost. Nevertheless, the

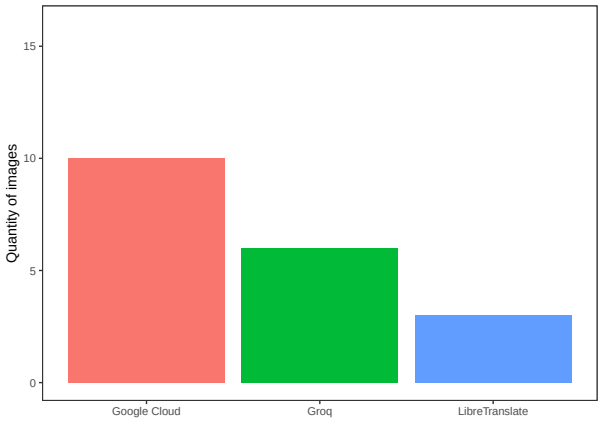

Figure 6: Translator scores (dependent on caption conformity)

generated sentences enabled the analyses described in this work. In this regard, the Portuguese models diverge from the English baseline in terms of BLEU and CIDEr metrics, although LibreTranslate achieved a comparable performance for METEOR. Future work aims to evaluate the tokenizer, as this study adopted the Stanford PTB Tokenizer, the default in AoANet.

When comparing the quantitative results, the metric values alone are insufficient to determine the best translator, requiring the assessment of additional factors. In this context, Google Cloud incurs the highest monetary translation cost, approximately 180 USD, whereas the other tools do not involve any financial cost. Regarding translation time, Google Cloud presents the shortest duration, approximately 5 hours and 40 minutes, followed by LibreTranslate with 8 hours and 30 minutes, and Groq with 24 hours—the longest processing time, which must be split across at least three days due to platform limitations.

Considering the qualitative aspects, the models are able to produce satisfactory captions for 56% of the test sample and generate at least one caption deemed partially or fully compliant in 100% of the cases. Google Cloud Translation emerges as the most effective automatic translation option, outperforming the others by 7 points when all images are considered, and by 4 points in cases where the majority of evaluators agreed on the conformity between captions and images.

6.2 Limitations

The Portuguese models are trained using the same hyperparameters adopted for the English baseline in order to ensure comparability across experiments. However, these hyperparameters are originally designed for English datasets and may not be optimal for Portuguese. Languages exhibit distinct linguistic properties, including morphology, verbal inflection, average sentence length, and grammatical marking of gender and number, which may influence tokenization, sequence length, and the distribution of words. Consequently, alternative hyperparameter configurations specifically tuned for Portuguese can potentially lead to improved performance. Due to computational constraints and the cost associated with training the models, hyperparameter optimization for Portuguese are not performed in this study.

Because this experiment relied on human evaluation, and the evaluators were instructed to adopt their own assessment criteria, evaluation bias is inherently present and may have influenced cases in which the majority indicated non-conformity while a small number of evaluators chose to mark a caption as acceptable. Furthermore, since the survey was distributed exclusively among evaluators residing in Brazil and fluent in Brazilian Portuguese, the linguistic proximity of the captions to the variety of Portuguese spoken by these evaluators may have biased decisions in favor of the LLM or Google Cloud, both of which produced captions that more closely resemble Brazilian Portuguese than European Portuguese, which is the variety reflected in LibreTranslate.

An example of this latter point can be observed in the evaluation of image identifier 222, in which most evaluators selected the caption that was more aligned with Brazilian Portuguese, even though both captions conveyed the same meaning. The corresponding responses can be seen in Figure 7.

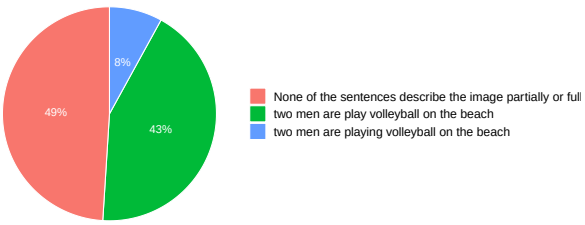

Figure 7: Result for the image labeled 222

6.3 Ethical considerations

Additional relevant factors concern the ethical aspects inherent in this experiment. The use of the Flickr30k dataset is, by itself, legitimate due to its stated policies permitting research use. Nevertheless, the creators of Flickr30k warn that the images used to construct the dataset remain subject to the terms and conditions of the Flickr platform. As such, the dataset authors do not hold the rights to the images, which could potentially raise concerns regarding image rights and copyright. Concerning the choice of Flickr30k, despite other IC datasets have emerged as a native Portuguese dataset, i.e. *PraCegoVer* (Santos et al., 2021), it still has some limitations for IC task: only one reference to each image and both mean and variance of the reference sentence length are significantly greater than those of IC datasets. Thus, we used Flickr30k as either was the most employed for IC tasks.

7 Conclusion and Future Work

This study evaluates different automatic translators for Portuguese in the task of image captioning. Both quantitative and qualitative experiments were conducted. Monetary cost and computational power may influence decisions regarding which translation systems to adopt. As part of this work, the authors will make available on the GitHub platform the Flickr30k dataset translated by the three translation systems used, as well as the translation algorithms and the AoANet implementation adapted for execution on Flickr30k. A Google Colab notebook containing all the necessary instructions for training and validation will also be provided.

Future work includes developing more robust models using the same datasets, re-evaluating the captions and metrics under these improved conditions, incorporating a morphosyntactic analysis of the sentences generated by each model, and identifying or defining a more representative metric for the automatic evaluation of image captioning models in Portuguese.

8 Acknowledgments

This work was supported by FAPESB (TIC002/2015 and CCE022/2023), CAPES, CNPQ, and INCT-TILDIAR/CNPq (408490/2024-1).

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442.

Gabriel Oliveira dos Santos, Diego Alysson Braga Moreira, Alef Iury Ferreira, Jhessica Silva, Luiz Pereira, Pedro Bueno, Thiago Sousa, Helena Maia, Nádia Da Silva, Esther Colombini, Helio Pedrini, and Sandra Avila. 2023. CAPIVARA: Cost-efficient approach for improving multilingual CLIP performance on low-resource languages. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 184–207, Singapore. Association for Computational Linguistics.

Amir El-Komy, Osama R Shahin, Rasha M Abd El-Aziz, and Ahmed I Taloba. 2022. Integration of computer vision and natural language processing in multimedia robotics application. *Inf. Sci*, 7(6):1–12.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

João Gondim, Daniela Claro, and Marlo Souza. 2022. Towards image captioning for the portuguese language: Evaluation on a translated dataset. In *Proceedings of the 24th International Conference on Enterprise Information Systems - Volume 2: ICEIS*, pages 384–393. INSTICC, SciTePress.

João Gondim, Daniela Barreiro Claro, and Marlo Souza. 2025. A bilingual analysis of multi-head attention mechanism for image captioning based on morphosyntactic information. *Journal of the Brazilian Computer Society*, 31(1):1063–1076.

Google. 2008. Cloud translation. Accessed on August 02, 2024.

Groq. 2016. Groq. Accessed on August 02, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4633–4642.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, page 228–231, USA. Association for Computational Linguistics.

Elizabeth D. Liddy. 2001. Natural language processing. In *Encyclopedia of Library and Information Science*, 2nd edition. Marcel Decker, Inc., NY.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Guilherme Moraes Rosa, Luiz Henrique Bonifacio, Leandro Rodrigues de Souza, Roberto Lotufo, and Rodrigo Nogueira. 2021. A cost-benefit analysis of cross-lingual transfer methods. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4231–4242, Online. Association for Computational Linguistics.

Gabriel Oliveira dos Santos, Esther Luna Colombini, and Sandra Avila. 2021. Pracegover: A large dataset for image captioning in portuguese. In *Proceedings of the 34th Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE.

Richard Szeliski. 2010. *Computer Vision: Algorithms and Applications*, 1st edition. Springer-Verlag, Berlin, Heidelberg.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726.

V H Vishnu Kumar and N Lalithamani. 2022. English to tamil multi-modal image captioning translation. In *2022 IEEE World Conference on Applied Intelligence and Computing (AIC)*, pages 332–338.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, and 12 others. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.