Towards Image Captioning for the Portuguese Language: Evaluation on a Translated Dataset

João Gondim[a], Daniela Barreiro Claro[b] and Marlo Souza[c]
FORMAS Research Group, Institute of Computing, Federal University of Bahia,
Av. Milton Santos, s/n, PAF2., Campus de Ondina, Salvador, Bahia, Brazil

Keywords: Image Captioning, Natural Language Processing, Machine Translation.

Abstract: Automatic describing an image comprehends the representation from the scene elements to generate a concise natural language description. Few resources, particularly annotated datasets for the Portuguese language, discourage the development of new methods in languages other than English. Thus, we propose a new image captioning method for the Portuguese language. We provide an analysis empowered by an encoder-decoder model with an attention mechanism when employing a multimodal dataset translated into Portuguese. Our findings suggest that: 1) the original and translated datasets are pretty similar considering the measure achievements; 2) the translation approach includes some dirty sentence formulations that disturb our model for the Portuguese language.

1 INTRODUCTION

Image captioning is the task of, given an image, analyzing its visual contents and generating a textual description to it (Bernardi et al., 2017). It is a challenging task as the model not only has to predict the objects present on the scene but also express their relationships in natural language (Xu et al., 2016).

Automatic captioning can bring advances in information systems, as they can make predictions based on the scene description and help the visually impaired with more accessible interfaces (Bernardi et al., 2017). Nevertheless, despite its importance and potential incorporation into systems, this problem has not received much attention for languages other than English. Although there are many Natural Language Processing methods for the English language, few resources can be applied to languages other than English. Furthermore, it is increasingly recognized in the literature that the focus on the English language may introduce some biases into the research area (Bender, 2009; Bender, 2019). As far as we know, no previous work has explored image captioning for the Portuguese language.

Common datasets for training image captioning models, such as Flickr8k (Hodosh et al., 2013), Flickr30k (Young et al., 2014) and COCO (Common Objects in Context) (Lin et al., 2015) are all annotated using English captions. To our knowledge, there are no corpora for image captioning for the Portuguese language[1]. This work analyzes image captioning for the Portuguese language based on automated machine translation. We constructed a multimodal corpus for image captioning in Portuguese from an English corpus, and we evaluate its quality by training an image captioning neural architecture based on Encoder-Decoder adapted from (Xu et al., 2016) for the Portuguese language.

The quality of our neural architecture was evaluated two-fold: (a) an automatic evaluation of the predicted captions based on standard metrics in the literature, namely the Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) and the Metric for Evaluation of Translation with Explicit ORdering (METEOR); and (b) a qualitative human evaluation of our model. The contributions of this work are (1) the multimodal dataset for image captioning for the Portuguese language created through machine translation, (2) a model capable of captioning images in the Portuguese language, (3) an analysis of the outcomes

[a] https://orcid.org/0000-0001-7225-1165
[b] https://orcid.org/0000-0001-8586-1042
[c] https://orcid.org/0000-0002-5373-7271

[1] While we are aware of the existence of the PraCegoVer (dos Santos et al., 2021) multimodal corpus construction project for the Portuguese language, the corpus is currently not publicly available.

Gondim, J., Claro, D. and Souza, M.
Towards Image Captioning for the Portuguese Language: Evaluation on a Translated Dataset.
DOI: 10.5220/0011080000003179
In *Proceedings of the 24th International Conference on Enterprise Information Systems (ICEIS 2022) - Volume 1*, pages 384-393
ISBN: 978-989-758-569-2; ISSN: 2184-4992

of our model and which types of errors such method is susceptible and (4) a comparison of obtained metrics between English and Portuguese captions generated by the same model trained with, respectively, the original and the translated dataset.

The following sections are organized as follows: Section 2 presents the related literature on image captioning based on neural networks; Section 3 presents the proposed architecture of our model, adapted from (Cho et al., 2014), (Xu et al., 2016) and (Tan and Le, 2020); Section 4 describes our experimental methodology, discussing the dataset creation and empirical setup. We present the results of our evaluations in Section 5 and our discussions in Section 6, analyzing possible threats to the validity of our results. Finally, we present our final considerations discuss possible future work in Section 7.

2 RELATED WORK

Early image caption systems relied on rule-based techniques to fill template captions with detected data obtained from object detectors and scene recognition systems (Yao et al., 2010; Socher and Fei-Fei, 2010).

Recent work on image captioning are neural-based being inspired by techniques from automatic machine translation encoder-decoder models, posing the problem of generating captioning as “translating” images into text.

Authors from (Vinyals et al., 2015), to our knowledge, are the first inspiration from machine translation *encoder-decoder* models to image captioning. In this work, the authors employ an encoder-decoder model to maximize the likelihood of generating a caption given an input image, similar to neural machine translation models that aim to maximize the likelihood of generating a translation given an input sentence. This is achieved by replacing the encoder neural network, a Recurrent Neural Network (RNN), with a Convolutional Neural Network (CNN), which is responsible for generating feature vectors describing semantic characteristics of the image. The generation of the image description is, thus, performed by a decoder network, composed by an RNN that generates words by receiving, at every time step, the image vector, the previous hidden state, and words generated before.

Such work (Vinyals et al., 2015) has inspired other neural architecture models for image captioning, combining computer vision and natural language processing. Authors in (Xu et al., 2016) expanded the CNN-RNN model with the attention mechanism introduced in (Bahdanau et al., 2016). This mechanism helps the model learn better interpretations from images and adds the ability to visualize what the model ”sees”. In (Lu et al., 2017), a novel adaptive model with a *visual sentinel* is proposed to help the captioning model decide when to ”look” at the image to generate the next word.

More recently, authors in (Huang et al., 2019) extend the use of *attention* by applying it on both the decoder and the encoder of their image captioning model, extending the conventional attention mechanism to determine the relevance between its results and queries. Authors in (Li et al., 2020) (the state of the art model for MS COCO Captions[2]) uses pretraining and objects tags in images as anchor to facilitate the learning of alignments.

While, to our knowledge, no work on Image Captioning has been conducted for the Portuguese language, recently, the work of (dos Santos et al., 2021) proposes the creation of multimodal corpus, which may be used for image captioning. The authors exploit voluntary image captioning performed by users of Instagram [3] social network using the tag *#pracegover*. The tag is commonly used for increasing the accessibility of the image-based network by providing image descriptions for the screen readers. While this corpus is of great value to studies on multimodal processing and image captioning for the Portuguese language, it is not yet available to the community. Thus, we could not evaluate it.

Therefore, to the best of our knowledge, this is the first work to attempt image captioning for the Portuguese language employing an attention method and a translated dataset.

3 MODEL

This section describes the architecture we employ in this work. We adapted the model from (Xu et al., 2016) which, to our knowledge, is the first model to use the attention mechanism on the task of image captioning. This is important since we want to evaluate how image captioning early models perform on a language with a different complexity other than English. We chose this model to train with Portuguese captions to observe how a simpler architecture behaves when trained with a much more different language.

The architecture takes advantage of the sequence-to-sequence training approach from (Cho et al., 2014) replacing the first RNN as an encoder with a CNN model for image classification. The CNN creates feature vectors to a second RNN as a decoder that out-

[2] https://paperswithcode.com/sota/image-captioning-on-coco-captions

[3] http://www.instagram.com

puts each word of the final caption. However, differently from (Vinyals et al., 2015), by employing a mechanism of *attention* (Bahdanau et al., 2016), our decoder can select different aspects from the image at each time while generating the caption, as seen in Figure 1.

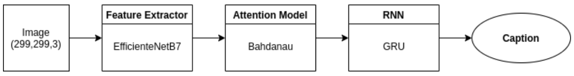

Figure 1: Pipeline of caption generation using CNN and RNN with attention. Source: created by authors.

While (Xu et al., 2016) employ the CNNs from GoogleNet (Szegedy et al., 2014) and VGG16 (Simonyan and Zisserman, 2015) in their models, in our architecture, we make use of EfficientNet (Tan and Le, 2020) due to its outstanding performance for image classification[4]. EfficientNet is a family of neural network models (going from B0 to B7) that takes advantage of scaling on all model dimensions to achieve better accuracy at image classification tasks. In our approach, we employed the EfficientNetB7 because of its higher accuracy.

In the next sections, we describe some experiments to validate our approach for the Portuguese language.

4 EXPERIMENTAL SETUP

In this section, we describe our experimental setup. We provide the methodology employed for generating the data, splitting the dataset, and training the network. We also describe our dataset and the hyperparameters used in our empirical evaluation of the proposed architecture.

4.1 Methodology

The first aspect of our empirical validation of our architecture is the construction of a dataset to train and evaluate the model. Since, as discussed before, there is no publicly available image captioning dataset for the Portuguese language, we created one by automatically translating the image descriptions in an image captioning corpus for the English language, namely the Flickr8k corpus (Hodosh et al., 2013).

For the translation of the captions, we employed the LibreTranslate[5] machine translation system. LibreTranslate employs OpenNMT (Klein et al., 2017) for neural machine translation. We discarded the use of proprietary software due to the cost or the restriction of requisitions when using online tools such as Google Translate. The use of open-source software also facilitates the replicability of the experiments.

We trained our architecture within the new Portuguese corpus and conducted two evaluations: *automatic evaluation* and *human evaluation*.

In the first step, we employ an *automatic evaluation* of the system, according to different hyperparameters of the model. For this evaluation, we employ BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007) scores between the generated caption and a set of possible captions for an image in the corpus, as commonly done in the language generation literature. We used the scores obtained from training with both the translated and the original dataset to compare metrics of generated captions (English and Portuguese) for the same images. We analyze this comparison to evaluate how two equal models behave when the descriptions' languages differ and if the Portuguese model would produce similar metrics compared with the already tested English one.

In the second step, we perform a *qualitative evaluation* of the generated captions and the analysis of errors made by the model. In this analysis, we selected 100 images and the captions generated by the best model in the automatic evaluation, which were analyzed by 34 human annotators. Each annotator was asked to check: (1) if caption contain errors in the description of the subject on the image (gender, quantity, age, or action done by subject), (2) if the caption contain errors regarding color descriptions of an object, (3) if caption contained errors in the description of the scenario where the action occurred, (4) if the caption contained errors in the descriptions of the objects on the scene, (5) if the generated caption was poorly structured (verb not agreeing with the subject, word repetition), (6) if the generated caption described the image correctly, or (7) if the generated caption was wrong or improperly described the image. With the annotation results, we performed qualitative analysis of the errors in our model.

4.2 Datasets

The data used to generate our corpus was obtained from the Flickr8k (Hodosh et al., 2013) dataset, which was created for the task of image description and retrieval. It is composed of 8091 images that capture a wide range of common activities, each of them containing five descriptions. These original captions are independently produced, and they have 8488 unique

[4] https://paperswithcode.com/sota/image-classification-on-imagenet

[5] https://github.com/LibreTranslate/LibreTranslate

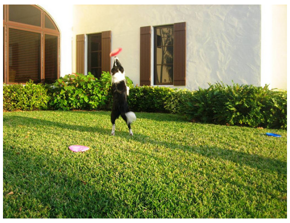

Figure 2: Example of image from Flickr8k. Source (Hodosh et al., 2013).

words among their captions, and the average size of each caption is 10.82 words with a standard deviation of 3.77. The Flickr8k dataset has a default split of training, dev, and testing images and captions. We followed these default splits to select the images for training (7091 images) and testing (1000 images). Figure 2 shows an example of image with the following captions annotated:

- A black and white dog is catching a Frisbee in the yard.
- A black and white dog is trying to catch a Frisbee in the air.
- A dog jumps to catch a red Frisbee in the yard.
- Dog is jumping up on a very green lawn to catch a Frisbee.
- The black and white dog tries to catch a red Frisbee on green grass.

We translated each caption in the *Flickr8k* dataset with *LibreTranslate* and fit them into the same format for the model training. After translating, the above example gets the following Portuguese captions:

- Um cão preto e branco está pegando um Frisbee no pátio.
- Um cão preto e branco está tentando pegar um Frisbee no ar.
- Um cão salta para pegar um Frisbee vermelho no pátio.
- O cão está pulando em um gramado muito verde para pegar um Frisbee.
- O cão preto e branco tenta pegar um Frisbee vermelho na grama verde.

After translating each caption, we obtained a corpus of 9780 unique words, an average caption length of 11.16, and 4.00 as a standard deviation.

Since we employed a dataset with translated sentences, we analyzed some errors inherited from the translation approach. For example the sentence “A black dog is jumping over a log along a beach.” was translated to “Um cão preto está saltando sobre um *log* ao longo de uma praia.” and the word “log” was not translated; another example is “A dirt bike racer jumps over a slope.” that was translated into “Um motociclista de sujeira salta sobre uma inclinação.”.

Although the automatic translation from English into Portuguese may introduce some errors in the corpus, this is a first attempt to build a Portuguese annotated dataset for image captioning. Undoubtedly, in our pipeline of image captioning, we inherited these errors into our model, and thus, they can induce errors in the generation of our captions.

4.3 Experimental Framework

We implemented our architecture in Section 3 with Python and the Tensorflow library. As hyperparameters, we set a fixed learning rate of 0.001, with a batch size of 128 and 50 epochs. We evaluated 300 and 600 dimensions for the Embedding layer of our RNN decoder, depending on the word embedding pre-trained weights. Each image was resized to have 299 pixels of width and height.

Different from the authors in (Xu et al., 2016), we adopted a pre-trained word embedding model to test different representations of the Portuguese words on our training corpus among with starting training from random weights. We employed GloVe (Pennington et al., 2014) with different dimensions (300 and 600). GloVe is a vectorized word representation suited for capturing semantic and syntactic regularities using vector arithmetic.

5 RESULTS

In this section, we analyze the results from our image captioning model for the Portuguese language. First we present the results of our automatic evaluation process among with the comparison of metrics with a model trained on the original Flickr8k captions (English language), we analyze positive and negative outliers from this comparison and than we present the agreement scores obtained from the human evaluations on the errors enumerated in Section 4.1.

5.1 System Evaluation

The model’s scores can be seen on Table 1. The differences in the BLEU scores between the models trained were not significantly high. Although the model with embedding layer initiated with random weights presented the best results, we decided to go further within our human evaluations with the model trained with GloVe 300 (the second in BLEU score metrics). We state that a known generalized representation of the

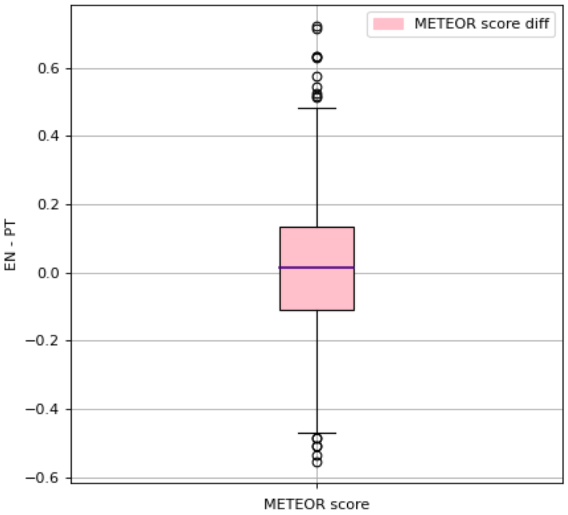

Figure 3: Difference of METEOR scores for each caption on the test set.

words can decrease the risk of overfitting a model with captions in the training set.

Table 1: Metrics from the model trained with Portuguese captions.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
300d Embedding	40.31	29.67	23.06	19.82	26.93
600d Embedding	41.92	30.45	23.96	20.46	28.11
GloVe 300	40.84	29.61	23.34	19.77	27.94
GloVe 600	39.92	29.36	23.15	19.79	27.34

We trained another model with the original (English captions) dataset to compare the metrics obtained when an English and a Portuguese caption are generated for the same image. This model was trained with the same hyperparameters as the one chosen before, with a 300 dimensions GloVe embedding appropriate for English language, this embedding is available online[6]. The results are shown in Table 2.

Table 2: Metrics from the model trained with English captions.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
GloVe 300	43.38	31.39	24.68	21.00	29.56

We calculate the difference between BLEU and METEOR scores for each image in the test set. We show these differences in Figure 3 and Figure 4.

Figure 3 shows a boxplot of Eq. 1 for each of the 1000 images on the test set. The mean difference is 0.016, and the boxplot shows a slight asymmetry toward positive values, indicating that English metrics are a little higher.

$$difference = ENMETEOR - PTMETEOR \quad (1)$$

Figure 4 shows the same results but with BLEU scores. The mean differences from BLEU 1, 2, 3, and 4 scores are, respectively, 0.025, 0.017, 0.013, and 0.012, the same slight asymmetry from the METEOR score. The different approaches to calculating BLEU scores (with 1, 2, 3, or 4 n-grams counting) show a decreasing inter quantile range, indicating that the difference of scores gets more concentrated as the number n-grams counted increases. Similar metrics might lead to this, or BLEU-4 metrics are generally smaller.

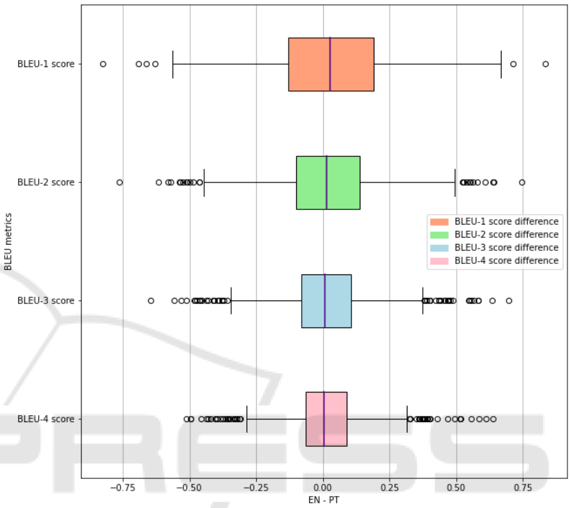

Figure 4: Difference of BLEU scores for each caption on test set.

It is relevant to point out that the Portuguese dataset contains errors introduced by the automatic translation employed, which might influence the performance of the model. This is an important point after observing slight asymmetries toward positive differences (indicating higher metrics with the English dataset).

5.2 Human Comparison

Figure 3 presented some outliers on the captions' metrics differences. We selected these outliers in Portuguese and English to compare the mistakes that occur and analyze if there are common aspects among outliers. For this task, we chose 3 images whose differences in the METEOR metrics were either 1) a positive outlier (better metric for the English caption) or 2) a negative outlier (better metric for the Portuguese caption).

First, we present images where the METEOR score of the Portuguese sentences was high. In Figure 5, the main subject of the image - namely the golden/brown dog - is not described in the English

[6] https://github.com/stanfordnlp/GloVe

generated caption. As for Figure 6, there is a reference to the main character in the image, but with the wrong action being performed. It is important to note the use of common Portuguese words that, when employed together, can add ambiguity to the sentence, with a vulgar sense that probably would not be used by a native speaker of such language. Such words probably would not happen if we were not using automatic translation to generate the dataset. Figure 7 presents a different mistake where the men shown playing rugby are depicted as dogs; this is quite curious since we used the same image classifier to extract features from the images.

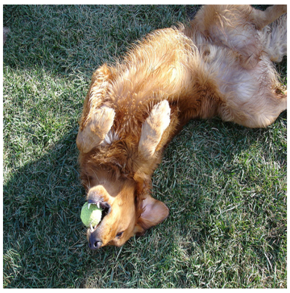

Figure 5: Generated caption PT: "um cão marrom está pulando em uma bola de tênis em sua boca". Generated caption EN: "a brown and white ball". Source (Hodosh et al., 2013).

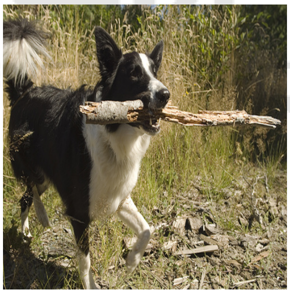

Figure 6: Generated caption PT: "um cachorro preto e branco com um pau grande". Generated caption EN: "a dog jumping from the grass". Source (Hodosh et al., 2013).

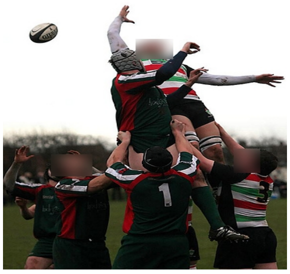

Figure 7: Generated caption PT: "um jogador de futebol entram na linha de rugby". Generated caption EN: "dogs stand in a game of rugby". Source (Hodosh et al., 2013).

The following images show examples where the EN caption had better METEOR scores than the PT one. Figure 8 shows that the Portuguese caption fails to adequately describe the scene - without mentioning the sea - and the number of characters. In Figure 9, the Portuguese caption fails to describe the fishing action from the young man. Looking at Figure 10, the object used by the person is misdescribed in the Portuguese sentence. A bad translation can cause this in the training dataset as both the words "bicicleta" and "motocicleta" can be translated to the informal word "bike" in English.

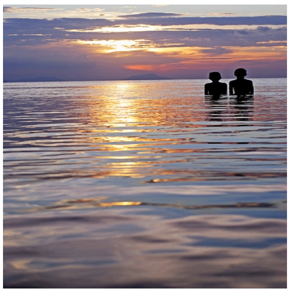

Figure 8: Generated caption EN: "two people in a boat is walking in water". Generated caption PT: "uma pessoa com uma membro da cidade no horizonte". Source (Hodosh et al., 2013).

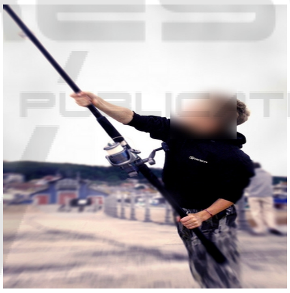

Figure 9: Generated caption EN: "a young man holding a fishing pole". Generated caption PT: "um jovem leva uma pai pontapé". Source (Hodosh et al., 2013).

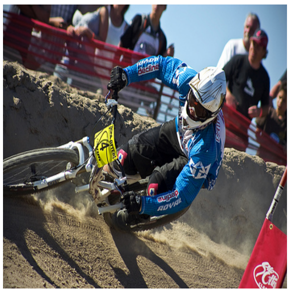

Figure 10: Generated caption EN: "a person on a bike". Generated caption PT: "um motociclista está realizando um truque em sua bicicleta". Source (Hodosh et al., 2013).

5.3 Human Evaluation

We performed a qualitative evaluation of 100 randomly selected dataset images with 34 human annotators. We separated common errors found into five clusters: 1) *Wrongly described the subject on the image* (gender, quantity, age, or action done by subject) (Error 1); 2) *Wrong color of an object* (Error 2); 3) *Wrong scenario* (beach, lagoon, mountain) (Error 3); 4) *Wrong objects in the scene* (Error 4) and 5) *Generated sentence poorly structured* (verb not agreeing with the subject, word repetition) (Error 5).

We presented these errors within an online form added with two other options to indicate if the caption is correct or not: 1) Generated sentence describes the image correctly and 2) Generated sentence is wrong and does not reflect the presented image. After asking annotators to signalize which errors they could detect on given captions, we obtained 34 complete answers and calculated the agreement between observations using Krippendorff's alpha inter-rater reliability. The overall agreement score was 0.4052, and the agreement for each type of error is described in Table 3.

Table 3: Krippendorff's agreement scores for each error.

Errors	Score
Wrongly described the subject on the image (gender, quantity, age or action done by subject) - (1)	21.01%
Wrong color of an object - (2)	42.96%
Wrong scenario (beach, lagoon, mountain) - (3)	23.31%
Wrong objects in the scene - (4)	75.58%
Generated sentence poorly structured (verb not agreeing with subject, word repetition) - (5)	27.81%
Generated sentence describes the image correctly - (6)	36.74
Generated sentence is wrong and does not reflect the presented image - (7)	19.52%
Overall Agreement	40.52%

It is noticeable that a higher agreement rate occurs when there are object-related errors.

6 DISCUSSIONS

In this section, we discuss the errors we mentioned in 5.3. Figure 11 shows the number of times each error appeared on the evaluations. Although the error with the highest frequency was *Error 1* followed by *Error 5*, both errors have low agreement among annotators. This may indicate a divergence concerning the wrong subject description or a poorly structured sentence generation. The error with fewer appearances (*Error 4*) also has the highest agreement value, suggesting that object errors do not show up much in captions or are well described.

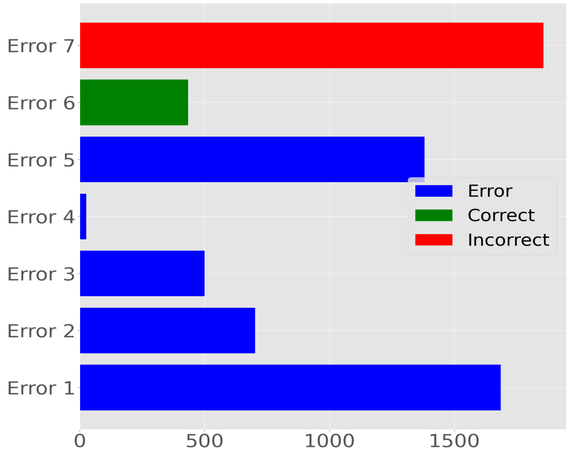

Figure 11: Counts of each annotations.

In the following subsections, we present examples input into the model that had the higher count of votes on each of the five errors and one case of error (7) (a wrong sentence and does not reflect the image).

6.1 Wrongly Described the Subject on the Image

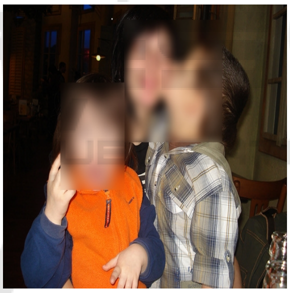

Figure 12: Generated caption: "duas crianças com dois caras" — meaning "two children with two guys".

The generated description for Figure 12 was "duas crianças com dois caras", in this caption, the model describes four subjects, instead of three, and the description of one of the subjects is erroneous - a woman described as a man.

In Figure 13 we shown the second image with most votes on error 1, and it can be seen how the representation of person had some lacks as the generated sentence is "três pessoas em frente da parede de uma cidade perto do edifício" describing less people than the reality.

6.2 Wrong Color of an Object

In Figure 14 the description was "um jovem que usa uma camisa verde" with a clear wrong color detection on the person's shirt.

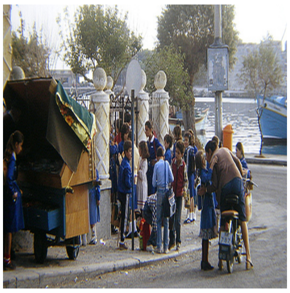

Figure 13: Generated caption: "três pessoas em frente da parede de uma cidade perto do edifício" — meaning "three people in front of the wall of a city near the buiding".

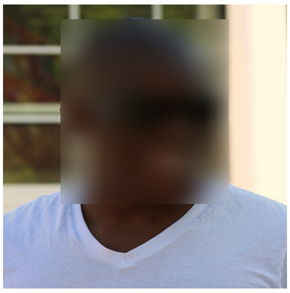

Figure 14: Generated caption: "um jovem que usa uma camisa verde" — meaning "a young that wears a green shirt".

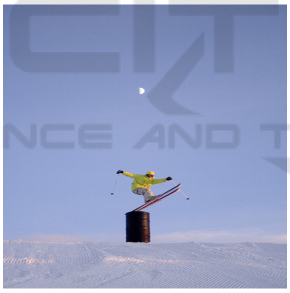

Figure 15: Generated caption: "pessoa em uma camisa vermelha a fazer um handstand em um bluff com um céu dando o pôr do sol na montanha" — meaning "person on a red shirt doing a handstand on a bluff with a sky doing a sunset on the mountain".

As for Figure 15, once again a shirt color is the mistake detected by the annotators. Here we notice the same word "camisa" before a color mistake happens.

6.3 Wrong Scenario

A wrong scenario description was the mistake appointed on Figure 16 as the caption is "um cão marrom está na grama"

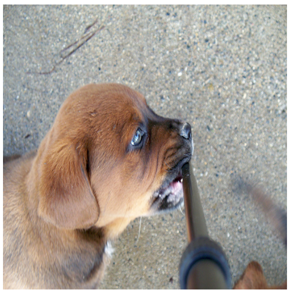

Figure 16: Generated caption: "um cão marrom está na grama" — meaning "a brown dog is on tha grass".

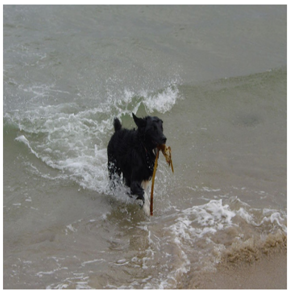

Figure 17: Generated caption: "dois pequenos cães jogam em um lago" — meaning "two small dogs play on a lake".

Figure 17 shows how a lack of information about the whole scene can lead to a mistake in the caption as the model's output was "dois pequenos cães jogam em um lago", the water and waves from the beach are not enough for correctly depicting where the dog is playing.

6.4 Wrong Objects in the Scene

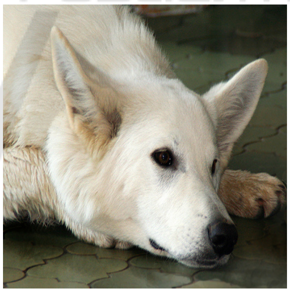

Figure 18: Generated caption: "um cão branco grande branco com pássaro está de plástico branco" — meaning "a white big dog with bird is on white plastic".

The caption of Figure 18 added what the annotators considered to be a wrong object in "um cão branco grande branco com pássaro está de plástico branco", since there are no birds on the image. The second and third images with more votes on this error are, respectively, 14 and 20.

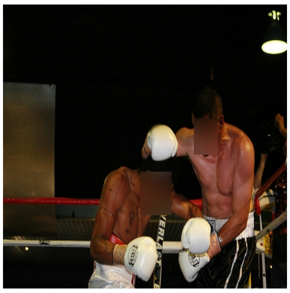

Figure 19: Generated caption: “um homem com um com um homem com um homem em roupa por um homem com um homem está usando um homem em roupa por homens em um tem um tem um pau” — meaning “a man with a with a man with a man on cloth by a man with a man is using a man on cloth by men on a has a has a stick”.

6.5 Generated Sentence Poorly Structured

A caption with Error 5 is shown on Figure 19, “um homem com um com um homem com um homem em roupa por um homem com um homem está usando um homem em roupa por homens em um tem um tem um pau”, where most of the words are repeated over the generated sentence.

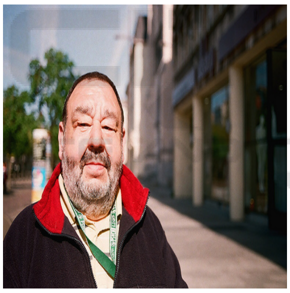

Figure 20: Generated caption: “um homem com um casaco de compras e casaco preto e casaco de peles e casaco de compras e casaco preto e casaco de couro e casaco preto e casaco de compras e casaco preto e casaco de compras está” — meaning “a man with a shopping coat and black coat and a fur coat and shopping coat and black coat and leather coat and black coat and shopping coat and black coat and shopping coat is”.

For Figure 20 we have the second image in votes for Error 5 with the following caption “um homem com um casaco de compras e casaco preto e casaco de peles e casaco de compras e casaco preto e casaco de couro e casaco preto e casaco de compras e casaco preto e casaco de compras está”, here the RNN decoder seems to be giving too much attention on the coat worn by the man as the word “casaco” is constantly repeated.

6.6 Generated Sentence Is Wrong and Does Not Reflect the Presented Image

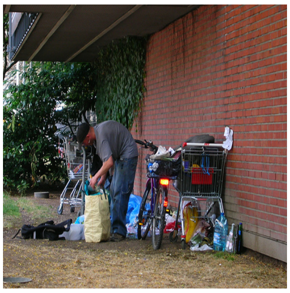

Figure 21: Generated caption: “um homem vestido com uma mulher está tirando uma com um homem em frente a uma rua” — meaning “a man dressed with a woman is taking off a with a man in front of the a stree”.

Figure 21 had a wrong generated sentence with the following caption: “um homem vestido com uma mulher está tirando uma com um homem em frente a uma rua” in which other subjects, a wrong verb and a reference to street appears at the same time.

It is worth noticing that some mistakes, subsections 6.1, 6.2, 6.3 and 6.4, lead to understanding that a wrong image description was sent to our RNN decoder: a missing person added, wrong color, erroneous ground description, or a bird that is not on the image being referenced. These might indicate that better image representations are still needed for a better “translation” of the image into a sentence.

7 CONCLUSIONS AND FUTURE WORK

We presented a qualitative evaluation of the types of errors made by a Portuguese image captioning system based on a neural translation model. For that, we performed a human evaluation of the generated sentences. Though the agreement scores obtained in this qualitative process were low for some of the errors analyzed, we were able to see how poor image representations might negatively influence when giving an automated description of an image. We compared scores of differently trained models (both with the same images, but one with Portuguese an the other with English captions) and obtained results indicating that the mean difference between these scores is close to zero: same images having similar scores with either Portuguese or English captions. For future works, we intend to explore how to enhance the way an encoded image representation is presented to the decoder responsible for generating a sentence and keep improving our dataset with better translations. We hope that this work will encourage future works on image captioning for the Portuguese language.

ACKNOWLEDGEMENTS

Authors would like to thank FAPESB TIC PROJECT Number TIC0002/2015, CAPES Finantial Code 001 and CNPQ for financial support.

REFERENCES

Bahdanau, D., Cho, K., and Bengio, Y. (2016). Neural machine translation by jointly learning to align and translate.

Bender, E. (2019). English isn't generic for language, despite what nlp papers might lead you to believe. In *Symposium and Data Science and Statistics*. [Online; accessed 15-may-2020].

Bender, E. M. (2009). Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., and Plank, B. (2017). Automatic description generation from images: A survey of models, datasets, and evaluation measures.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation.

dos Santos, G. O., Colombini, E. L., and Avila, S. (2021). #pracegover: A large dataset for image captioning in portuguese.

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

Huang, L., Wang, W., Chen, J., and Wei, X.-Y. (2019). Attention on attention for image captioning.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (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.

Lavie, A. and Agarwal, A. (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.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., and Gao, J. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2015). Microsoft coco: Common objects in context.

Lu, J., Xiong, C., Parikh, D., and Socher, R. (2017). Knowing when to look: Adaptive attention via a visual sentinel for image captioning.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition.

Socher, R. and Fei-Fei, L. (2010). Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 966–973.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2014). Going deeper with convolutions.

Tan, M. and Le, Q. V. (2020). Efficientnet: Rethinking model scaling for convolutional neural networks.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. (2016). Show, attend and tell: Neural image caption generation with visual attention.

Yao, B. Z., Yang, X., Lin, L., Lee, M. W., and Zhu, S.-C. (2010). I2t: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.