XXVIII Congresso {virtual} de Iniciação Científica da Unicamp 2020

#PraCegoVer: Automatic Image Audiodescription

Gabriel O. dos Santos*, Esther L. Colombini, Sandra Avila
Institute of Computing (IC), University of Campinas (Unicamp), Brazil

Abstract
The Internet has been becoming increasingly accessible, reaching a wide and diverse audience. However, visually impaired people still face many problems to use it because a significant part of the published content is exclusively visual. Generating image descriptions automatically is still an open challenge; this problem is known as *Image Captioning*. It involves identifying the objects present in the image and describing the semantic relationship among them verbally. The main available datasets contain only English captions, and datasets with captions described in other languages are scarce. Thus, most of the proposed models are capable of generating solely English descriptions. The initiative #PraCegoVer has been producing many Portuguese image descriptions, and we have been leveraging them to construct a dataset with Portuguese Captions. We have developed a tool to collect public posts from Instagram automatically, and we use it to retrieve images and their associated texts. Also, we proposed an algorithm to remove duplicate posts, in which both visual and textual information is leveraged to group posts considered duplicates, *i.e* posts with similar images or texts. This scientific report introduces the #PraCegoVer dataset, detailing the collection, preparation, preprocessing, and data analysis.

Key words
PraCegoVer, Image Captioning, Attention Mechanisms, Deep Neural Networks

1 Introduction

The Internet is becoming increasingly accessible, reaching a wide variety of audiences. However, little progress has been made on including people with disabilities. The scenario is even worse for visually impaired people since a significant part of the Internet's content is exclusively visual, for instance, photos and advertising images. Although there are screen readers well evolved, they are still mostly dependant on some notations added on the source code of websites, which in turn, in general, are not that descriptive or are left empty.

In light of this situation, in 2012 arose *#PraCegoVer* [1] as a social movement, idealized by Patrícia Braille, that stands for the inclusion of people with visual impairments besides it has an educational propose. The initiative aims to call attention to the accessibility question. It stimulates users to post images tagged with *#PraCegoVer* and add a short description of their content. This project has inspired many local laws that establish all posts made by public agencies on social media must refer to *#PraCegoVer* and contain a short description of the image.

The task of automatic image description using natural sentences helps to include people with visual impairments onto the Internet. This task is known as *image captioning*. It is still a big challenge that requires understanding the semantic relation of the objects, as well as their attributes and actions. Thus, in addition to visual interpretation methods, linguistic models are also needed to verbalize the semantic relations.

Many works have addressed this problem by generating English descriptions with a few words because the most famous dataset, MS COCO [2], has images labeled with English descriptions containing about 10 words on average. Thus, inspired by the *#PraCegoVer* project, we created a multi-modal dataset with images and descriptions in Portuguese by leveraging the content posted by those who have joined *#PraCegoVer* initiative. Besides being one of the first datasets for Image Captioning in Portuguese, our dataset has a varied number of words in the captions. It has more than 40 words on average, which is a challenge to state-of-the-art models that tend to repeat the same words repeatedly to increase the length, as we have demonstrated experimentally. Finally, we intend to contribute to the blind Portuguese speaker community. We hope this dataset encourages more works addressing the automatic generation of descriptions in Portuguese.

2 #PraCegoVer Dataset

#PraCegoVer dataset leverages the data posted on social media by people that have joined the initiative #PraCegoVer. Thus, we collect data of posts tagging #pracegover from public profiles on Instagram, then we clean and analyze them, and we train models to generate automatic image descriptions.

Fig. 1 illustrates the pipeline of the construction of our dataset, and it also highlights the percentage of posts lost in each step concerning the total amount of collected posts. First, we collect the data from Instagram (Sec. 2.1). Then, we clean the captions to obtain the descriptions, and we extract images and text features. We reduce the dimensionality of image feature vectors to optimize processing. Next, we cluster the images to analyze the data and to remove duplicates. Finally, we split the dataset into train, validation, and test sets, considering the proportion 60%, 20%, 20%, respectively, and we use this split to train the models.

We highlight in Fig. 1 the loss in each step of the pipeline. During the posts collection, about 2.3% of posts are lost because of profiles that become private during this process. Besides, 9.6% of posts have malformed captions. They do not follow the main pattern, which consists of the "#pracegover" followed by the description. Thus it is tough to extract the actual caption from the whole text. Therefore we remove them. Finally, 44.9% of the total amount of posts have duplicated either caption or image, which may easily overfit the models, then we also remove these cases from the dataset. In total, about 56.8% of data are lost or removed. It is worth noting that there is a loss of data inherent to the data source and the process.

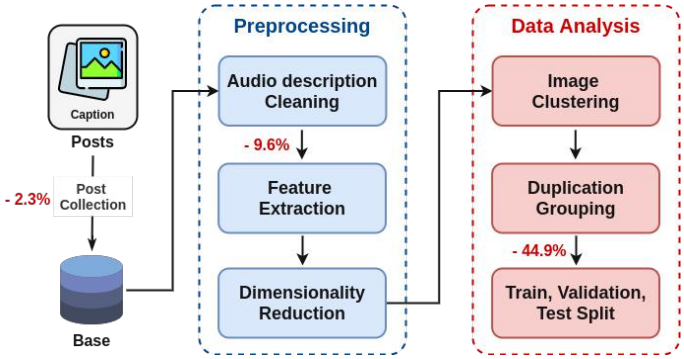


Figure 1: From data collection to dataset split. First, we collect the data, clean the captions to obtain the audio descriptions, and extract image and text features. Finally, we analyze the data to remove duplicates and split the dataset into train, validation, and test sets. We highlight the percentage of posts lost in each step.

2.1 Data Collection

In this work, we have collected data from Instagram because it is focused on image sharing and allows us to filter the posts by hashtag. This platform limits the filter by hashtag to posts published in the last seven days, therefore to overcome this obstacle, we first search for posts related to the hashtag *#PraCegoVer* and save just the profiles that have posted them. Next, we visit these profiles looking for posts with the hashtag. Inside the profile pages, the posts are not limited by post date or quantity. Thus, we have access to all images published by that user (they are public).

We execute this process daily and incrementally, storing the images, width, height, captions, and shortcode, which is a post identifier, post owner id, post date, and the collection date. In this way, we collect posts published any time ago, instead of up to the past seven days as restricted in the hashtag page. We ensure that the robot never accesses posts from private profiles, which would require an acceptance. However, there may be profiles that became private after we had collected their posts. So far, we have collected more than 400 thousand posts; thus, our dataset is larger than many available datasets, even in English.

2.2 Data Preprocessing

The captions of the collected posts, in general, contain other texts, emoticons, hashtags, profile mentions, and URL link marks beside the description; thus, to extract, we need to preprocess it. We identified patterns in those texts by thoroughly reading many of them, and we used regular expressions to find the description within the text. In general, a description comes right after the hashtag *#PraCegoVer*, so we first crop the caption keeping just the text after this hashtag. Then, we use regular expressions to remove emoticons, hashtags, URL links, and profile marks, but it might lead to miss punctuations at the end of the texts that we also remove. Some captions have a mark that indicates the

end of the description, such as "Fim da descrição"; thus, we also use it as an end delimiter. Fig. 2 shows a real example of caption. After cleaning, the final text is "Várias siglas de partidos e suas logomarcas misturadas juntas.".

Câmara aprova anistia a partidos que não investiram mínimo exigido em campanhas femininas.

#pracegover: Várias siglas de partidos e suas logomarcas misturadas juntas. Fim da descrição.

📸 Foto: Reprodução

👉 Confira mais informações através dos nossos Stories.

#brasil #joaopessoa #jornalismo #news #noticia #paraiba

Figure 2: An example of real caption tagged with the hashtag *#PraCegoVer*. Observe that there are emoticons and hashtags in the original caption, but we remove them.

Moreover, many profiles are used to re-post images from others, with just a few details changed. Fig. 3 (a) illustrates a cluster where there are images duplicated. Note that we consider a duplicate image that differs from each other by an image filtering, the addition of a logo, etc. Since the images are quite similar, we have to remove one of the duplicates because they might overfit. On the other hand, images of similar objects and scenarios (Fig. 3 (b)), may enrich our dataset because they represent the adverse situations that a model can face in a real-life application.

((a)) Sample from a cluster whose majority of the posts are related to perfumes. We highlighted the duplicated images, such that the ones with borders in the same color and line are considered duplicates.

((b)) Sample from a cluster that groups images of airplanes. We observe the variety of positions of the airplanes. Some images show just part of them, such as wings, turbines.

Figure 3: Samples of images from two different clusters.

We illustrate in Fig. 4 the pipeline used to group duplicates. First, we extract feature vectors from images using MobileNetV2 [3], then reduce the dimensionality from 1280 to 900 through UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) [4], and we cluster the images considering in this embedding using HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) [5]. Then, we compute the distance matrix in each cluster of images using the cosine distance, reducing the memory space needed to store the distances.

Regarding the captions, we first preprocess them, extracting just the audio description parts. Then, we convert the texts into lower case, remove stopwords, and transform them into TF-IDF vectors [6], and we compute the cosine distance to generate the distance matrix of captions related to images within the same cluster. Finally, we construct the similarity graph based on both distance matrices of images and captions for each cluster. The similarity graph is an undirected graph such that each vertex represents a post, and there is an edge between u and v if either text or image similarity of posts P_u and P_v is greater than text and image thresholds, respectively. Each connected component of this graph represents a group of post duplicates. Thus we can remove them, keeping just one example of the duplicates.

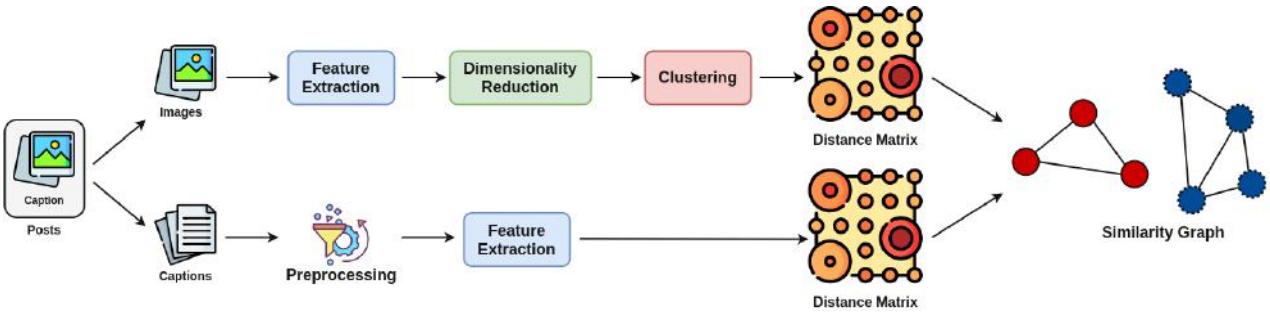
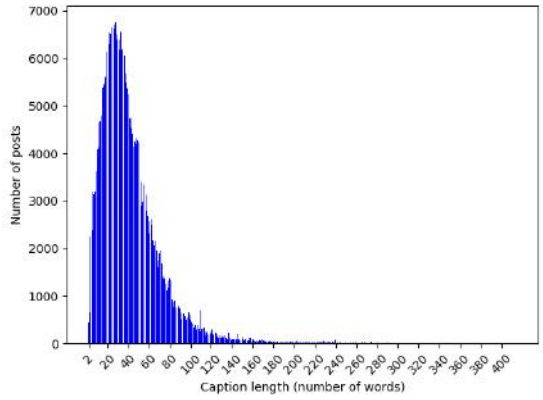


Figure 4: Duplicate grouping pipeline.

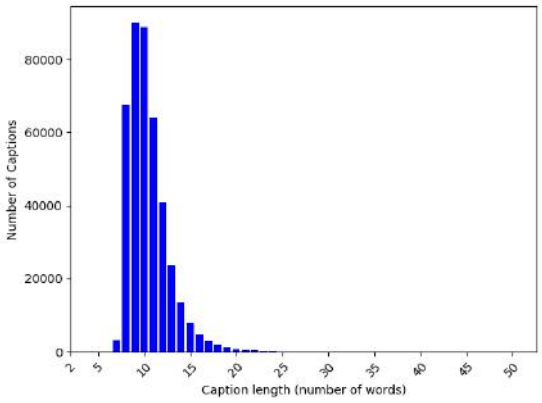
2.3 Statistics

We have collected 470.461 posts from 13.389 different profiles on Instagram; thus, our dataset contains a wide variety of images, such as images of animals, cars, airplanes in many positions, flowers, and advertisements. Such variety is essential when we train models to make sense of real-world images on the Internet.

Fig. 5 (a) shows a histogram with the distribution of posts by the number of words in the descriptions after noise removal. The concentration of posts whose captions have from 2 to 100 words is noticeable, and the average number of words in a cleaned caption is close to 40. Our dataset has captions with a varied number of words. On average, ut has 40 words whereas MS COCO [2] dataset has, on average, about 10 words per caption (Fig. 5 (b)), and also it has a little variation in the number of words among the captions; thus, our dataset tends to be more challenging.



((a)) Histogram of the number of captions, in our dataset, by their length in terms of the number of words.



((b)) Histogram of the number of captions, in MS COCO dataset, by the number of words in them.

Figure 5: Distribution of caption length by number of words.

3 Preliminary Results

We based our model on Huang et al.'s work [7]. We use their implementation of Attention on Attention for Image Captioning. Initially, we trained the original model in our dataset, but we notice that it repeats the same word to generate a caption with a length similar to the reference. Thus, we proposed a change in the training using the Self-Critical Sequence [8] approach to consider a modified CIDEr [9] to take into account the varied number of words of our dataset. We relaxed the penalty in case of a different number of words in the generated sentence and the reference. However, this approach still produces a significant number of repetitions in the generated captions. Thus, we also include a term to penalize the generated sentences with repeated words, and this approach worked better in terms of the semantic quality of the generated captions.

We used a subset with 100k posts of our current dataset. We trained the models considering this subset split into

train, validation, and test. Fig. 6 shows two examples of images and the captions generated by our model. Note that the generated texts have several words similar to the original caption. Also, the model avoids word repetitions and generates captions that describe the overall scene.

((a)) **Original**: "Foto de mulher sorrindo sentada na calçada com confetes e serpentinas com um pandeiro nas mãos, calçando e vestindo Havaianas."; **Generated**: "Fotografia de uma mulher sentada em um tapete de madeira. No canto inferior esquerdo, o texto: "Unknown"."

((b)) **Original**: "Na imagem, duas Bases fortalecedoras aparecem entre vários girassóis."; **Generated**: "Na imagem, o produto está em cima de flores.".

Figure 6: Examples of two images followed by their original captions and the captions generated by our model. We have brushed face of the woman and suppressed her name in original caption to anonymize her.

4 Conclusions

In this work, we collected and analyzed more than 470k posts from Instagram. We introduced the largest dataset to image captioning in Portuguese, having captions with 40 words on average and various lengths, making the problem more challenging since most of the metrics used to address this problem work better for small captions with a similar number of words. Furthermore, we modified the metric used to train the Attention on Attention model, avoid word repetitions, and deal with the captions with a varied number of words.

For future work, we will keep collecting data from Instagram to increase our dataset size, and we will explore the techniques proposed in Literature in the context of #PraCegoVer. We will especially address the problem of describing eventual texts presented in the images, which is common in advertisements, for example. Moreover, we will repeat the experiments considering the total dataset instead of a subset. Finally, we intend to create a framework for image captioning so that it can be used to describe images for visually impaired people.

Acknowledgments

G. Oliveira is funded by FAPESP (2019/24041-4). S. Avila is partially funded by FAPESP (2013/08293-7) and Google LARA. LaRoCS and RECOD Labs. are supported by projects from FAPESP, CNPq, and CAPES.

References

[1] Web para Todos, "Criadora do projeto #PraCegoVer incentiva a descrição de imagens na web." http://mwpt.com.br/criadora-do-projeto-pracegover-incentiva-descricao-de-imagens-na-web, 2018.

[2] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *ECCV*, 2014.

[3] M. Sandler *et al.*, "MobileNetv2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018.

[4] L. McInnes *et al.*, "UMAP: Uniform manifold approximation and projection for dimension reduction," *JOSS*, 2018.

[5] L. McInnes *et al.*, "HDBSCAN: hierarchical density based clustering," *JOSS*, 2017.

[6] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, 1972.

[7] L. Huang *et al.*, "Attention on attention for image captioning," in *ICCV*, 2019.

[8] S. J. Rennie *et al.*, "Self-critical sequence training for image captioning," in *CVPR*, 2017.

[9] R. Vedantam *et al.*, "CIDEr: Consensus-based image description evaluation.," in *CVPR*, 2015.