Canadian AI 2023
5—9 June 2023, Montréal

Image Captioning Dataset in Persian Using Flicker30k Images

Shima Baniadamdizaj[†], Alexander Breuer[†]
[†] Fredrich Schiller University Jena

Abstract

Image Captioning is the task of describing an image with a textual description in natural language. It makes use of each Natural Language Processing and Computer Vision to generate the captions. For this task, it is necessary to identify objects, actions, their relationship, and some silent feature that may be missing in the image. The final result must be a syntactically and semantically correct generated, relevant and brief description of the image. The use cases are computer vision and natural language processing (NLP). It is not an easy task for a machine to generate or imitate the human brain's ability for image captioning, although research in this field has shown great achievements. Deep learning techniques are capable to handle such problems using CNN and LSTM. There are many datasets in this case, but mostly contain captions in English, whereas datasets with captions described in other languages are scarce. In this paper, we introduce a new and at the moment first image captioning dataset in Persian. The images are collected from the Flickr30k dataset and have five different references per image. Also, both the mean and variance of reference sentence length are high, which makes this dataset challenging due to its linguistic aspect.

Keywords: Image Captioning, Image Captioning Dataset, Image-To-Text, Computer Vision, Natural Language Processing

1. Introduction

Automatically describing image content using natural sentences is known as image captioning. It is still a big challenge because understanding the semantic relationship of the objects in the image, their attributes, and the actions are required to generate a natural correct sentence. Furthermore, machines need to verbalize the semantic relations in addition to visual interpretation. There are so many image sources on television, the internet, news, and many other sources but most images do not have descriptions. A human can interpret an image without having a description whereas this is a tough task for machines to describe the content of an image naturally. These days generating caption and natural description of an image is a research concept.

This research topic is crucial for various reasons for example locating or understanding the type of activities. Image captioning need to identify objects, action, and the relationship among objects in the image and their activities. This is more difficult than identifying the objects because it also must pay attention to the relationships. For example, in an image with a train and people in a station, the caption can be that people are getting on the train or are waiting on the platform for a train. The generated sentence must be syntactically and semantically correct. The methods that are used for image captioning are broadly based on deep learning-based methods [1–4].

The most important benefit of labeled datasets with various captions, like Flicker30k [5] is that deep convolutional neural networks (CNN) are efficacious. A method that automatically generates a sentence description has enthralled more research focus on artificial intelligence, it has played a significant role in computer vision, i.e., allowing computer systems to recognize images, that can be helpful for various purposes. The image captioning task became easier for a machine because of the existence of the large amount of annotated

*shima.bani@uni-jena.de

data, like Flickr8k [6], Flickr30k [5], and MS-COCO Captions [7]. In addition to this many other large-scale datasets have been created [8], most of them contain only English captions. In contrast, datasets with captions described in other languages are scarce.

Translating existed datasets from English to other languages is a cheap way to train models to generate non-English captions. Using this kind of non-English data has already shown that it introduces noise that can affect the performance of models. Particularly, Xue et al. [9] shows that the model performance is harmed when translated datasets are used instead of using originally annotated datasets in the target language.

Hence, we introduce the new dataset with images and Persian descriptions. As far as we know, this is the first dataset proposed for the Image Captioning problem with captions in Persian. This dataset has 10180 labels for images from Flicker that are randomly selected and captioned manually by different people. Each image has five different reference descriptions. Also, the average reference length is 13.3 and the standard deviation is 5.5. These values are considerably high in comparison to other datasets on this subject. It should be considered that "mi" is also considered an independent word. These characteristics make this dataset challenge.

2. **Literature Review**

The image captioning task became easier for a machine because of the existence of a large amount of annotated data, like Flickr8k [6], Flickr30k [5], and MS-COCO Captions [7]. Microsoft Common Objects in COntext (MS-COCO) is a dataset created from the images in MS-COCO [7] and human-generated captions. The first version of the MS-COCO dataset contains 164K images split into training (83K), validation (41K) and test (41K) sets. MS-COCO Captions dataset includes around 160k images from Flickr, distributed in over 80 categories, with five captions per image. Its images are annotated using Amazon Mechanical Turk (AMT). As a result, the descriptions' average sentence length is about 10 words.

Many large-scale datasets have been created [11-17]. One example of a dataset that follows this approach is the Conceptual Captions dataset [8] which has more than 3.3M pairs of images and English captions. It was created by crawling web pages and extracting images and its alt-text HTML attribute. Agrawal et al. proposed nocaps [18] using real-world images which is a benchmark that consists of validation and test set with 4500 and 10,600 images with 11 human-generated captions per image. Moreover, it has more objects per image than MS-COCO, and it has 600 object categories. This dataset is created by selecting images from the Open Images V4 dataset [19].

Recently, Gurari et al. proposed the VizWiz-Captions dataset [20] focused on the real use case for visually impaired like blind people. This dataset includes 39,181 images that are taken by blind people, each image has five captions using the AMT platform. This dataset also has metadata that indicates the image issues. The overlap between VizWiz and MS-COCO in the case of the caption is about 54%, which shows a significant bias in photography situations. InstaPIC-1.1M [21] and #PraCegoVer [22] are created by collecting posts from Instagram. InstaPIC-1.1M includes 721,176 pairs of image-caption from 4.8k users. Based on the 270 selected hashtags. The captions in the InstaPIC-1.1M dataset are the contents that Instagram users wrote about their posts, which can be quite irrelevant to the content of the image. #PraCegoVer approach also is based on Instagram posts, but in contrast to InstaPIC-1.1M and Conceptual Captions, it collects only #PraCegoVer tagged captions. This dataset contains captions with 40 words on average, while those in MS-COCO Captions have only ten words, and the variance of sentence length in our dataset is also more significant.

In case of dataset in non-English languages also there is a dataset in Korean [23]. It contains Korean tourist spots (KTS) and focused on Korean multi-model deep-learning research. The KTS dataset has four modalities (image, text, hashtags, and likes) and consists of 10 classes related to Korean tourist spots. All data were extracted from Instagram and preprocessed. The dataset includes 10 different classes with 1000 instances per class (total of 10,000 instances).

3. Dataset Collection and Analysis

3.1. Dataset Collection

3.2. Dataset Characteristics

4. Experiment

5. Conclusion and Future Works

In this paper, we introduced a new dataset for image captioning with Persian annotations for multi-modal tasks in the field of machine learning. This dataset is s designed for research with Persian texts, and it consists of images from Flicker. The dataset can be used for multi-modal tasks, such as image captioning, multi-modal classification, and recommendation systems, as well as single-modal tasks, like image classification. In the future, we plan to increase the size of the dataset by adding more images and more captions per image. Also, we planned to categorize the images into different classes such as landscape, human, animals, and so on.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (2014).

[2] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *arXiv preprint arXiv:1406.1078* (2014).

[3] I. Sutskever, O. Vinyals, and Q. V. Le. "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems* 27 (2014).

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[5] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2641–2649.

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

[7] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. "Microsoft coco captions: Data collection and evaluation server". In: *arXiv preprint arXiv:1504.00325* (2015).

[8] P. Sharma, N. Ding, S. Goodman, and R. Soricut. "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 2556–2565.

[9] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. "mT5: A massively multilingual pre-trained text-to-text transformer". In: *arXiv preprint arXiv:2010.11934* (2020).