**A New Multi-Model Image Dataset with Persian captions**

Shima Baniadamdizaj1[0000-0003-1678-5108], Alexander Breuer1

1 Friedrich Schiller University Jena, Jena 07743, Germany

{shima.bani, alex.breuer}@uni-jena.de

**Abstract.** Image Captioning is describing images automatically using natural sentences. For the task of image captioning, 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 achievement. Deep learning techniques are enough capable to handle such problems using CNN and LSTM. There are many datasets in this case, but most contain only English captions, whereas datasets with captions described in other languages are scarce. In this paper, we introduce the first image captioning dataset in Persian. The images are based on 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 in Persian; 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 captions and natural descriptions 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 Flickr30k [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 a large amount of annotated data, like Flickr8k [6], Flickr30k [5], and MS COCO Captions [7]. In addition to this, many other large-scale datasets have been created [8], and most of them contain only English captions. In contrast, datasets with captions described in other languages are scarce.

Translating existing 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] and Rosa et al. [10] show 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 Flickr 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 “می” is also considered an independent word. These characteristics make this dataset a challenge.

1. **Related Work**

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 sets 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 the case of the dataset in non-English languages, there is also a dataset in Korean [23]. It contains Korean tourist spots (KTS) and is focused on Korean multimodel 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 (a total of 10,000 instances).

1. **Dataset Description**

This section describes the process of data collection and analysis of the collected dataset. All the images in this dataset are collected from Flickr. Flickr is an American image hosting service and community that is a popular way for amateur and professional photographers to share their high-resolution photos. Flickr is a great source for finding photos. Most open-source image sites use the Flickr API to curate images onto their site.

3.1 Dataset collection

In this paper, we collected images randomly from the Flickr webpage but also, but we noticed that all images in our dataset are also included in the Flickr dataset [5]. So, based on this feature, there is this possibility that uses this dataset as a bilingual dataset for different image captioning or images description tasks.

Intending to have a diverse set of images, we rely on images from Flickr. Specifically, we use the same subset of images as in the Flickr dataset [5]; these images have been verified to contain various images in different situations and also have human annotators. Using the same images as Flickr additionally allows multitask learning and transfer learning scenarios between two languages in image captioning tasks. The images in both datasets, our collected and Flickr30k, are annotated by human annotators.

Annotators were asked to describe images. This description can be just some words, one or more complete sentences. It is just enough that the content of the caption describes the image in the right way and at least points to the most important part of the image.

After the completion of five captions for each image, to check if the captions in the dataset provide correct content, we observed each caption for each image to be sure that the content is related to the image. For this step, we asked some people to be evaluators of the captions. Evaluators were asked to vote yes/no on whether the caption that is written in the first step of captioning delights the following features:

* Is correct for the specific image.
* Consists of at least one word.
* Is grammatically correct.
* Does not contain any foreign language content.

Five independent captions were created for each image. Additional sixth or more captions were collected only if one of the captions in the evaluation step does not meet the basic requirements of a valid caption. The annotators did not see previously collected captions for a particular image and did not see the same image twice.

Any image in the dataset is annotated five times, so there are five different captions per image. These captions can be similar but are completely different and if the captions were the same for an image, we deleted that one and asked again to get a new caption. After captioning task, we have 9696 captions for 1939 different images.

The maximum length of a caption in collected dataset is for image number 119364730 with 55 words and the caption is “مردی با ظاهری مصمم صورت ، چکش کوچک چوبی را در دست دارد ، در حالی که یک جیب را نشان می دهد و خرچنگ قرمز بزرگی را نشان می دهد ، سخت پوستان قرمز دیگری در دست او نگه داشته می شود در حالی که مرد دیگری نشسته و پشت سر خود مقداری غذا نگه می دارد”. for counting the length of a caption, we counted all the prepositions and postpositions like “به”,” و”, “می” and other ones. Also, the minimum length of a caption is for image 1370615506 with just two words and the caption is “نوزادی میخندد” (See Fig. 1 and 2).

| | یک مرد با لباس مشکی و شلوار جین مشکی و یک پسر کوچک در حال بررسی داخل موتور یک کامیون سبز قدیمی هستند در حالی که این ماشین در یک زمین چمن پارک شده و یک مخزن بنزین در جلوی آن قرار دارد |  | | --- | --- | | یک مرد با لباس مشکی و شلوار مشکی روی موتور یک ماشین قدیمی و عتیقه سبز کار میکند و یک مخزن بنزین زرد روی چمن قرار دارد | | یک جوان و یک کودک موتور یک کامیون قدیمی را نگاه میکنند | | یک مرد و یک بچه روی موتور یک کامیون قدیمی کار میکنند | | مرد و کودک روی کامیون قدیمی کار میکند | |
| --- | --- | --- | --- | --- | --- | --- |

**Fig. 1.** Image with maximum caption and its other captions *in Persian and English*

As is visible in Fig. 1 the captions of an image vary in length. Therefore, it is a more challenging dataset and can help to get high accuracy in the training and testing of deep learning methods. The average length of captions is 13.3 and the standard deviation is 5.5.

The dataset has 4 columns. The first column is the name or number of the image in the Flickr30k image dataset that is called “image\_name”. All images are in jpg format and the sizes of the images are different and the same size as the original image on the Flickr webpage. The second column is the number of captions that for each image from 0 to 4 then each image has 5 different captions and is called “comment\_number”. The next column is “comment-fa” and includes the comment in Persian for each image. And the last column is “comment-en” and the content of this column is the caption in English that is imported from the Flicker30k dataset like Table 1.

**Table 1.** The column names and orders of the dataset.

| image\_name | comment\_number | comment-fa | comment-en |
| --- | --- | --- | --- |

For the Final step, we did some simple and small preprocessing, except evaluation of each caption. For this goal, we looked at all captions and converted the numbers to alphabetical version of number like “4” to “چهار” and so on. Also, all verbs and other words that can be separated are re-written separately. or if some words have to be considered as a single word, like words with prefixes are written together. In this case, some suffixes should not be considered as a single word with the main part of the word. For example, the ones which show plural version of a word. Because we need this case. So “ها” is considered as a single word, although in real Persian grammar it is not a single word. We are not making a grammar correction network, but we want to train a model to be able to describe a picture.

1. **Dataset tests**
2. **Conclusion and Future Work**

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 Flickr. 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, animal, and so on.

We hope our dataset, available at [bit.ly/3PTGNNA](https://bit.ly/3PTGNNA), will encourage the community to design better image captioning models in Persian and address its technical challenges, especially for helping visually disabled or blind people.

**References**

1. Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).
2. Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).
3. Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems 27 (2014).
4. Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
5. Plummer, Bryan A., et al. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models." Proceedings of the IEEE international conference on computer vision. 2015.
6. Hodosh, Micah, Peter Young, and Julia Hockenmaier. "Framing image description as a ranking task: Data, models and evaluation metrics." Journal of Artificial Intelligence Research 47 (2013): 853-899.
7. Chen, Xinlei, et al. "Microsoft coco captions: Data collection and evaluation server." arXiv preprint arXiv:1504.00325 (2015).
8. Sharma, Piyush, et al. "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018.
9. Xue, Linting, et al. "mT5: A massively multilingual pre-trained text-to-text transformer." arXiv preprint arXiv:2010.11934 (2020).
10. Rosa, Guilherme Moraes, et al. "A cost-benefit analysis of cross-lingual transfer methods." arXiv preprint arXiv:2105.06813 (2021).
11. Farhadi, Ali, et al. "Every picture tells a story: Generating sentences from images." European conference on computer vision. Springer, Berlin, Heidelberg, 2010.
12. Elliott, Desmond, and Frank Keller. "Image description using visual dependency representations." Proceedings of the 2013 conference on empirical methods in natural language processing. 2013.
13. Kong, Chen, et al. "What are you talking about? text-to-image coreference." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.
14. Gan, Chuang, et al. "Stylenet: Generating attractive visual captions with styles." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
15. He, Sen, et al. "Human attention in image captioning: Dataset and analysis." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
16. Hsu, Ting-Yao, C. Lee Giles, and Ting-Hao'Kenneth Huang. "Scicap: Generating captions for scientific figures." arXiv preprint arXiv:2110.11624 (2021).
17. Sidorov, Oleksii, et al. "Textcaps: a dataset for image captioning with reading comprehension." European conference on computer vision. Springer, Cham, 2020.
18. Agrawal, Harsh, et al. "Nocaps: Novel object captioning at scale." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
19. Krasin, Ivan, et al. "Openimages: A public dataset for large-scale multi-label and multi-class image classification." Dataset is available from https://github. com/openimages 2.3 (2017): 18.
20. Gurari, Danna, et al. "Captioning images taken by people who are blind." European Conference on Computer Vision. Springer, Cham, 2020.
21. Chunseong Park, Cesc, Byeongchang Kim, and Gunhee Kim. "Attend to you: Personalized image captioning with context sequence memory networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
22. dos Santos, Gabriel Oliveira, Esther Luna Colombini, and Sandra Avila. "# PraCegoVer: A Large Dataset for Image Captioning in Portuguese." Data 7.2 (2022): 13.
23. Jeong, Changhoon, et al. "Korean tourist spot multi-modal dataset for deep learning applications." Data 4.4 (2019): 139.