Expanding Flicker30k: a Novel Dataset for Persian Language Captions

Shima Baniadamdizaj[1][0000−0003−1678−5108], Alexander Breuer[1]

Friedrich Schiller University Jena, Jena, Germany
{shima.bani,alex.breuer}@uni-jena.de

Abstract. *********** TODO

Keywords: Image Captioning in Persian · Computer Vision · Natural language processing · image-to-text.

1 Introduction

Image captioning refers to the automated process of describing the content of an image using natural language sentences. While humans can interpret images without explicit descriptions, machines struggle with the task of generating natural and informative captions. They need to understand the semantic relationships between objects in an image, their attributes, and the actions in order to generate accurate and coherent descriptions, which is an effortless task for humans. Additionally, machines need to translate these semantic relations into correct literature. For instance, a suitable caption for an image which has a train and people at a station, may involve describing people either boarding the train or waiting on the platform.

The prevalent methods employed for image captioning predominantly rely on deep learning techniques [1–4]. Deep learning-based image captioning relies on extensive image datasets to train models effectively. The availability of a large volume of diverse images is crucial for capturing complex situations. Although there is a vast array of image sources available through television, the internet, news, and other platforms, the majority of these images lack accompanying descriptions. Various captioned image datasets, such as Microsoft COCO [5], Flickr8k [6], Flickr30k [7] and IAPR TC-12 [?], offer valuable resources for training deep learning models in image captioning tasks. These datasets contain diverse images with human-generated captions, enabling the development of accurate and meaningful captioning models.

The discussed datasets consist of image captions in the English language. However, there exists a scarcity of large-scale image captioning datasets available for languages other than English. For instance, the Multi30k dataset [8], encompassing English-German captions. Translating existing English datasets into other languages offers an inexpensive approach for training models to generate non-English captions. Studies [10–12] have specifically shown that employing

translated datasets, rather than using originally annotated datasets in the target language, can adversely affect model performance.

In light of these considerations, there is a scarcity of available image captioning datasets specifically for the Persian language. However, to address this gap, we have developed a dataset based on the widely used Flickr30k dataset [7] for Persian language image captioning. To the best of our knowledge, this dataset represents the first proposed solution for the worldwide Image Captioning problem with Persian captions. The dataset consists of 10,180 image labels randomly selected from Flicker30k and manually captioned by different individuals. Each image is associated with five distinct reference descriptions. The average length of descriptions is 13.3 words, with a standard deviation of 5.5 words. Additionally, it should be noted that the Persian prepositions are considered as an independent word in this dataset and are separated from the next word by space.

*********** TODO add chapters

2   Motivation and Related Works

The machine's ability to perform image captioning has been greatly facilitated by the availability of large-scale annotated datasets. Some well known examples of such datasets in English language are Microsoft COCO [5], Flickr8k [6] and Flickr30k [7]. Despite the existence of non-English image captioning datasets, such as Multi30k [8] and VIST [9] for English-German and English-Chinese captions, respectively, there is still a noticeable scarcity of Persian language datasets. The availability of large-scale captioning datasets specifically in Persian is limited, which presents a challenge for training models to generate captions in this language.

2.1   English Captioned Datasets

MS COCO (Microsoft Common Objects in COntext) [5] is an extensively used computer vision dataset. It contains a collection of over 164k images, sourced from diverse platforms. The dataset offers comprehensive annotations, including precise bounding boxes for object detection and segmentation masks. The training set consists of approximately 83k images, while the validation and test sets contain around 41k images each. Consequently, the average sentence length of the descriptions is approximately 10 words. The dataset is widely utilized in various computer vision tasks, including object recognition, scene understanding, and image captioning.

Flickr8k [6] is another notable computer vision dataset widely used in the field. It is specifically designed for image captioning tasks and consists of a diverse collection of 8k high-quality images sourced from the popular photo-sharing platform Flickr. Each image in the dataset is accompanied by five human-generated captions, providing rich textual descriptions of the visual content. The dataset aims to capture a wide range of scenes, objects, and activities,

showcasing the diversity of real-world imagery. Flickr8k has become a resource for training and evaluating image captioning models, enabling researchers to develop algorithms that generate accurate and contextually relevant descriptions for images. The availability of Flickr8k has played a role in advancing the field of image captioning and in natural language understanding and computer vision.

In addition to the Flickr8k dataset, another notable dataset is the Flickr30k dataset. The Flickr30k [7] dataset is an extension of Flickr8k, with a larger collection of images and relevant captions. While Flickr8k comprises 8k images, Flickr30k consists of approximately 30k images, making it substantially larger in scale.Both datasets share similarities in terms of their origin, as they are sourced from the popular photo-sharing platform Flickr. They aim to represent diverse scenes, objects, and activities, capturing images in their natural context. Both datasets are accompanied by human-generated captions that provide textual descriptions of the visual content. However, the increased size of the Flickr30k dataset brings several advantages. The increased diversity of content enhances model robustness and generalization in computer vision tasks. The larger number of images in Flickr30k provides a wider representation of real-world scenarios and concepts. This enhanced coverage allows researchers to tackle more complex challenges in object recognition, scene understanding, and image captioning tasks.

There is another dataset specifically created for the task of image captioning called "Nocaps" [13]. The dataset contains a large-scale collection of images and their corresponding captions. It emphasizes novel and diverse scenes, aiming to expand the range of images and situations that captioning models can handle effectively. Each image in the Nocaps dataset is associated with multiple human-generated captions. The dataset has a total of 166,100 human-generated captionsfor 15,100 images for over 500 unique object classes. The images of this dataset are sourced from the Open Images V4 [14]. The Open Images V4 dataset is a publicly available dataset for large-scale multi-label and multi-class image classification tasks. The OpenImages dataset offers a diverse range of visual concepts, ensuring a comprehensive representation of real-world objects and scenes.The dataset comprises 1.9 million images depicting complex scenes.

Some of the image captioning datasets are provided for a specific case. For example, the VizWiz-Captions dataset [15] is specifically designed to address the unique challenges faced by people with visual impairments in understanding visual content. The dataset contains approximately 31k images, each accompanied by a human-generated caption. There is an overlap about 54% between the VizWiz and MS COCO datasets in terms of the provided captions. The images and captions in the VizWiz-Captions dataset were collected through the VizWiz mobile application. This application enables visually impaired users to capture images and ask questions about the content.

The mentioned datasets have limitations as they provide captions for only English speakers. This lack of other language representation shows the necessity of the development of image captioning datasets for non-English speakers. This will enable a more comprehensive and inclusive environment for advancing

image captioning task,and provide a wider accessibility and applicability across different languages.

2.2   Non-English Image Captioning Datasets

As there is currently no specific Persian language image captioning dataset available, when discussing image captioning, the focus of this section is on non-English datasets instead of observing Persian ones. This shows the need for the development of Persian-specific datasets in the field of image captioning.

One of the well known examples of non-English image captioning datasets is the Multi30k dataset [8], which is a widely used benchmark dataset for image captioning and multimodal machine translation tasks.It is an extension of the Flickr30k dataset [7] and is aimed on generating English-German captions. Exactly like Flicker30k dataset each image has five reference caption but this time in German. Researchers rely on Multi30k to benchmark and advance the performance of models in generating accurate and meaningful captions in both English and German languages.

There is a Korean Tourist Spot (KTS) [16] which is a multiple modal dataset. It means that the dataset consists of images with textual descriptions, and additional metadata like hashtags and likes. This dataset is specifically tailored for studying Korean tourist spots and includes a total of 10 classes. The KTS dataset was collected from Instagram and followed by a preprocessing step. It comprises a substantial amount of data, with a total of 10k instances. Each of the 10 classes within the dataset is composed of 1k instances.

The ImageCLEF 2018 [17] dataset's images and their corresponding captions extracted from biomedical articles on PubMed Central and contains 232,305 pairs of image and caption, which were divided into separate training and test sets. The training set consists of 222,305 pairs, and the test set contains 10,000 pairs. The dataset includes multiple languages commonly used in biomedical research, such as English, Spanish, German, French, etc.

Datasets such as MS COCO and Flickr have multiple unofficial versions available for various languages, but no Persian. To address this gap, having a dataset in Persian could expand the resources for computer vision and image captioning tasks.

3   Dataset Collection

4   Experiments, Results and Discussions

5   Conclusion and Future Work

In this paper a novel dataset for image captioning presented with a specific focus on Persian annotations and sourced from Flicker30k.
          *********** TODO Add experiments and results

Moving forward, our future plans involve expanding the dataset by including a larger number of images and adding multiple captions per image. This expansion aims to enhance the dataset's diversity and provide researchers with a broader range of data to explore. Additionally, we intend to categorize the images into distinct classes, such as landscapes, humans, and animals, for targeted analysis.

References

1. Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
2. Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
3. Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. PMLR, 2015.
4. Luo, Jianjie, et al. "Semantic-conditional diffusion networks for image captioning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
5. Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014.
6. Rashtchian, Cyrus, et al. "Collecting image annotations using amazon's mechanical turk." Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk. 2010.
7. Young, Peter, et al. "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions." Transactions of the Association for Computational Linguistics 2 (2014): 67-78.
8. Elliott, Desmond, et al. "Multi30k: Multilingual english-german image descriptions." arXiv preprint arXiv:1605.00459 (2016).
9. Huang, Ting-Hao, et al. "Visual storytelling." Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies. 2016.
10. Xue, Linting, et al. "mT5: A massively multilingual pre-trained text-to-text transformer." arXiv preprint arXiv:2010.11934 (2020).
11. Zoph, Barret, and Kevin Knight. "Multi-source neural translation." arXiv preprint arXiv:1601.00710 (2016).
12. Rosa, Guilherme Moraes, et al. "A cost-benefit analysis of cross-lingual transfer methods." arXiv preprint arXiv:2105.06813 (2021).
13. Agrawal, Harsh, et al. "Nocaps: Novel object captioning at scale." Proceedings of the IEEE/CVF international conference on computer vision. 2019.
14. Krasin, Ivan, et al. "Openimages: A public dataset for large-scale multi-label and multi-class image classification." Dataset available from https://github.com/openimages 2.3 (2017): 18.
15. Gurari, Danna, et al. "Captioning images taken by people who are blind." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16. Springer International Publishing, 2020.
16. Jeong, Changhoon, et al. "Korean tourist spot multi-modal dataset for deep learning applications." Data 4.4 (2019): 139.

17. Ionescu, Bogdan, et al. "Overview of ImageCLEF 2018: Challenges, datasets and evaluation." Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings 9. Springer International Publishing, 2018.