Expanding Flicker30k: a Novel Dataset for Persian Language Captions

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

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

Abstract. Image captioning, the challenge of describing images through natural language, particularly benefits from deep learning techniques which needs diverse datasets. While existing datasets like Microsoft COCO and Flickr30k offer valuable resources, they are predominantly in English. The Expanding Flicker30k dataset fills this void for the Persian language. With manually curated captions averaging 13.3 words, the dataset captures various visual scenarios. By addressing the absence of Persian captioning resources, this paper contributes to both image captioning and multi-lingual research, fostering improved image understanding and language generation capabilities. This paper presents a novel resource for Persian language image captioning. Built upon the Flickr30k dataset, the new collection comprises 51,000 captions corresponding to 10,200 distinct images, providing a crucial dataset for advancing image captioning research in the Persian language. Each picture has five captions in Persian, which helps overcome the lack of non-English captioning datasets and allows for detailed language analysis. The dataset includes columns for image names, comment numbers, Persian captions, and English translations, facilitating bilingual comparisons.

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

1 Introduction

Image captioning involves the automated process of describing image content using natural language sentences. While humans effortlessly understand images, machines struggle to generate informative captions. They must grasp semantic relationships between objects, attributes, and actions to create coherent descriptions, a task humans find easy. Additionally, machines must convert these semantic relations into accurate literature. For example, an appropriate caption for an image with a train and people at a station might describe people boarding the train or waiting on the platform.

Image captioning predominantly relies on deep learning techniques [1–4]. Successful deep learning-based captioning necessitates vast image datasets for effective model training. Diverse image sources such as TV, internet, and news lack accompanying descriptions. Captioned image datasets like Microsoft COCO

[5], Flickr8k [6], and Flickr30k [7] provide valuable resources for training. These datasets, comprising diverse images with human-generated captions, facilitate accurate captioning model development.

Despite abundant English datasets, scarcity prevails in large-scale non-English image captioning datasets.For instance, the Multi30k dataset [8], encompassing English-German captions. Translating English datasets into other languages offers a cost-effective approach for non-English captions. Studies [9–11] indicate that translated datasets can impact model performance negatively. Addressing the scarcity of Persian image captioning datasets, we present a dataset based on Flickr30k [7]. This dataset of 10,180 images with Persian captions is the first proposed solution for global image captioning in Persian. Each image has five reference descriptions, averaging 13.3 words with a standard deviation of 5.5 words. Notably, Persian prepositions are distinct words separated by spaces.

This paper comprises five chapters, each vital to exploring Persian image captioning. Chapter 1 introduces the research and stresses the need for a dedicated dataset. Chapter 2 explores motivation, related works, and existing datasets, emphasizing the gap for a Persian dataset. Chapter 3 details dataset collection, annotation, and preprocessing. Chapter 4 presents experiments, discussions, and analyses. Chapter 5 concludes and outlines future prospects. This work contributes a resource for Persian language visual understanding and bridges the gap in image captioning research.

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 [12] 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 a widely used computer vision dataset containing over 164k diverse images. Sourced from various platforms, it offers comprehensive annotations, including precise bounding boxes for object detection and segmentation masks. The dataset's training set comprises approximately 83k images, while both the validation and test sets contain around 41k images each. The average description length is about 10 words. It finds applications in object recognition, scene understanding, and image captioning, contributing significantly to advancements in computer vision research.

Flickr8k [6], another significant computer vision dataset, focuses on image captioning. Curated for this task, it boasts a collection of 8k high-quality images drawn from the popular photo-sharing platform Flickr. Each image is accompanied by five human-generated captions, creating a rich source of textual descriptions for the visual content. This dataset is carefully designed to encompass a wide range of scenes, objects, and activities, showcasing the diverse nature of real-world imagery. It serves as a crucial resource for training and evaluating image captioning models, playing a pivotal role in the development of algorithms that generate contextually relevant and accurate image descriptions. The availability of Flickr8k has played a key role in advancing the realms of image captioning, natural language understanding, and computer vision as a whole.

Alongside the Flickr8k dataset, there exists a more expansive counterpart called Flickr30k [7]. While Flickr8k contains 8,000 images, Flickr30k significantly surpasses it with around 30,000 images, offering a substantial augmentation in scale. Both datasets share Flickr as their source and aim to encompass diverse scenes and objects within their contextual milieu, accompanied by human-generated captions. However, Flickr30k's increased size brings several distinct advantages. Its content diversity bolsters the resilience and applicability of computational models across intricate computer vision tasks. The broader range of images covers an array of real-world scenarios and concepts, empowering researchers to tackle intricate challenges in object recognition, scene comprehension, and image captioning. In essence, Flickr30k's expansion not only enriches the dataset but also stimulates comprehensive research and innovative problem-solving.

"Nocaps" [13] emerges as a dedicated dataset for the domain of image captioning. It sets its focus on the realm of novel and diverse scenes, aiming to expand the horizons of captioning models in handling an even wider array of images and contexts effectively. The dataset encompasses a vast collection of images, each coupled with multiple human-generated captions. A grand total of 166,100 captions accompany 15,100 images, spanning over 500 unique object classes. These images are sourced from the Open Images V4 [14], a publicly available dataset tailored for large-scale multi-label and multi-class image classification tasks. Open Images V4 enriches the dataset with a diverse spectrum of visual concepts, guaranteeing a comprehensive representation of real-world objects and scenes. This broader scope is cemented by the dataset's impressive compilation of 1.9 million images, each capturing complex scenes.

Some datasets cater to specific use cases. The VizWiz-Captions dataset [15], for instance, is meticulously crafted to address the distinct challenges faced by individuals with visual impairments when engaging with visual content. The dataset boasts an assemblage of approximately 31k images, with each image complemented by a human-generated caption. Notably, the VizWiz and MS COCO datasets exhibit an overlap of about 54% concerning the provided captions. Images and captions for the VizWiz-Captions dataset stem from the VizWiz mobile application, a platform that empowers visually impaired users to capture images and inquire about content through questions. This unique dataset is indispens-

able for driving advancements in accessibility and inclusivity within the realm of computer vision and natural language processing.

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.

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 VisionECCV 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 amazons mechanical turk." Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazons 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. Xue, Linting, et al. "mT5: A massively multilingual pre-trained text-to-text transformer." arXiv preprint arXiv:2010.11934 (2020).
10. Zoph, Barret, and Kevin Knight. "Multi-source neural translation." arXiv preprint arXiv:1601.00710 (2016).
11. Rosa, Guilherme Moraes, et al. "A cost-benefit analysis of cross-lingual transfer methods." arXiv preprint arXiv:2105.06813 (2021).
12. 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.
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 VisionECCV 2020: 16th European Conference, Glasgow, UK, August 2328, 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.