Expanding Flicker30k: a Novel Dataset for Image Captioning in Persian

Shima Baniadamdizaj

Abstract—Image captioning, the challenge of describing images through natural language, particularly benefits from deep learning techniques which need 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.

Index Terms—Image Captioning in Persian Computer Vision Natural language processing image-to-text.