ParsVQA-Caps: A Benchmark for Visual Question Answering and Image Captioning in Persian

Shaghayegh Mobasher*, **Ghazal Zamaninejad***, **Maryam Hashemi, Melika Nobakhtian, Sauleh Eetemadi**
Iran University of Science and Technology
{sh_mobasher, gh_zamaninejad, m_hashemi94, m_nobakhtian@comp.iust.ac.ir, sauleh@iust.ac.ir

Abstract

Despite recent advances in vision-and-language tasks, most progress is still focused on resource-rich languages such as English. Furthermore, widespread vision-and-language datasets directly adopt images representative of American or European cultures resulting in bias. Hence we introduce ParsVQA-Caps, the first benchmark in Persian for Visual Question Answering and Image Captioning tasks. We utilise two ways to collect datasets for each task, human-based and template-based for VQA and human-based and web-based for image captioning.

1 Introduction and Related Works

Language and vision tasks such as Image Captioning and Visual Question Answering (VQA) have recently gained popularity. However, many datasets have been released and show promising results; most datasets reflect American and European cultures and support resourceful languages like English (Antol et al., 2015; Johnson et al., 2017; Sharma et al., 2018; Young et al., 2014). There are a few works in image captioning in Persian that we mention in the following. Persian Image Captioning Dataset (Lashkaryani, 2021) consists of images related to a news article, with about 1500 news articles corresponding to images. Images and news articles are crawled from Tasnim News Agency's website. COCO-Flickr Farsi (Navid Kanaani, 2021) is another image captioning dataset in Persian that uses images and captions from COCO and flickr datasets translated to Persian. To the best of our knowledge, no VQA dataset is published in Persian and some use translation of VQA v1 and v2 datasets. Translation may seem suitable to create datasets in Non-English languages, but this approach has some problems. Language translation tools are usually imperfect and make mistakes. These datasets do not represent destination languages well, since they do not have language-specific and cultural phrases. To fill this gap, we introduce ParsVQA-Caps, the first benchmark in Persian for Visual Question Answering and Image Captioning.

*These authors contributed equally.

2 Datasets Collection

In this section, we describe how we collect image captioning and VQA datasets.

2.1 Image Captioning

We determine seven concepts related to Persian language and culture for collecting images. Given the list of image categories, we harvest about 100 images for each category from the web manually to ensure the relevance of the images gathered to the selected category[1]. We hire four native annotators to write captions for each image in Persian through a web-based tool we have developed according to detailed guidelines and examples. We review and correct these captions and clean them using the HAZM normalizer tool (Roshan-AI, 2022).

Image Category	#Images	#Human Captions
people	101	404
indoor	100	400
food	101	404
sport	102	408
ceremonies	100	400
cars	106	424
outdoor	103	0
Total	**713**	**2440**

Table 1: Number of images and captions collected manually based on their categories.

Another part of the Image Captioning dataset was created automatically by crawling a Persian Vector images and graphics website [2] . Some of the images on this website have Persian annotations suitable for being used as captions. This part of the dataset contains about 7000 images, each with one caption.

2.2 Visual Question Answering

To collect images for template-based VQA, we randomly select 15,000 images from MS COCO (Lin

[1]We searched for pages related to Persian culture on websites such as Pinterest.

[2]We would like to thank Motahareh Mirzayi and Mohammad Javad Pirhadi for their contribution to crawling data.

et al., 2014) dataset. We design over 100 templates to automatically generate formal and informal questions by considering ten different categories shown in Table 2. We ask native annotators using our web-based tool to provide short answers (1 to 3 words) to questions according to prepared guidelines. Since questions are generated automatically, some ambiguities may arise, like an object mentioned in a question has multiple correspondences in an image. Therefore, we allow annotators to select the "useless" option and ignore these questions. We utilise short answers to generate long answers (sentence answers) with templates.

For human-based VQA, we use the same images collected for image captioning to reflect Persian culture and language concepts. In this approach, annotators are asked to provide creative and challenging questions by determining categories and writing short and long answers for each question according to the guideline. To ensure that provided questions do not match template questions, we developed some validators that check similarity and prevent annotators from generating similar questions into templates. We also ensure correct question-category assignment by allowing annotators to create questions in a new category named "other". Table 2 shows the number of images, questions, and answers for each question category in our VQA dataset.

Category	#Images		#Formal Questions		#Informal Questions		#Answers	
	template	human	template	human	template	human	template	human
Object Presence	5346	161	2627	130	2720	48	5347	178
Sport Recognition	668	42	312	25	356	17	668	42
Positional Reasoning	3129	192	1676	158	1453	59	3129	217
Sentiment Understanding	548	20	284	14	264	6	548	20
Color Attributes	4102	328	2127	213	1976	166	4103	379
Counting Object	5171	373	2552	303	2619	144	5171	447
Activity Recognition	1491	81	771	57	720	26	1491	83
Object Detection	2767	157	1167	111	1600	58	2767	169
Object Material	2134	113	1051	72	1083	46	2134	118
Gender Recognition	1079	114	562	92	517	25	1079	117
Other	0	330	0	286	0	120	0	406
Total	**10170**	**723**	**13129**	**1461**	**13308**	**715**	**26437**	**2176**

Table 2: Number of images, questions and answers by category in our proposed VQA dataset.

3 Datasets Analysis

In this section, we analyze the captions, questions, and answers in image captioning and VQA.

3.1 Image Captioning

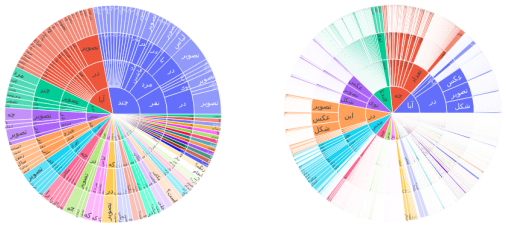
We quantify the captions by their length. The captions are between 4 and 27 words long. As shown in Figure 1a, on average, the web-crawled captions are shorter than captions collected by annotators. Furthermore, we apply Persian HAZM lemmatizer on all words of captions and visualize Word Cloud(see Figure 1b) without considering stop words to see the most frequent words.

(a) Distribution of caption length. (b) Word Cloud for words in human-generated captions.

Figure 1: Statistics for image captioning dataset.

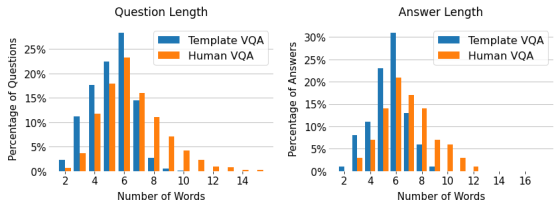
3.2 Visual Question Answering

We use a sunburst diagram to visualize the first four words used in questions. According to Figure 2a and 2b, ("In this image, how many people"), ("How many people in") are the most repeated question for human and template VQA respectively. Although Figure 2 shows more diverse questions in template-based VQA, the ratio of the number of unique questions based on the first four words to the total questions is 0.94 for human-based and 0.88 for template-based. In Figure 3, we show the percentage of question and long answer words. We see that most questions and answers range from four to seven words.



(a) Human-based VQA. (b) Template-based VQA.

Figure 2: Number of questions by their first four words.



(a) Distribution of question length. (b) Distribution of long answer length.

Figure 3: Statistics for VQA dataset

4 Conclusion

We present a benchmark in Persian for VQA and Image Captioning named ParsVQA-Caps, including images that depict Persian concepts and culture. The image captioning dataset consists of over 7.5k images and about 9k captions. The VQA dataset consists of almost 11k images and 28.5k question and answer pairs with short and long answers usable for both classification and generation VQA.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. Inferring and executing programs for visual reasoning.

Arman Malekzadeh Lashkaryani. 2021. Persian image captioning dataset.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Sajjad Ayoubi Navid Kanaani. 2021. Coco-flickr farsi.

Roshan-AI. 2022. hazm. `https://github.com/sobhe/hazm`.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.