Assessing Few-shot and Zero-shot Learning with CLIP Model for Visual Question Answering

Ghazal Zamaninejad[1], **Shaghayegh Mobasher**[2], **Sauleh Eetemadi**[1]

[1]Iran University of Science and Technology
[2]Sharif University of Technology

gh_zamaninejad@comp.iust.ac.ir, mobashershaghayegh@gmail.com, sauleh@iust.ac.ir

Abstract

In this study, we evaluate the efficacy of few-shot and zero-shot learning techniques within a Transformer-based model for Visual Question Answering (VQA). Leveraging the CLIP model, we conducted experiments across both English and Persian languages using the ParsVQA-Caps dataset. Our findings highlight the effectiveness of prompt-based strategies in enhancing VQA performance and provide valuable insights into their application in multilingual contexts.

1 Introduction And Related Works

In recent years, learning-based methods for recognizing objects and events in images have seen significant progress. Numerous models have emerged across the fields of Computer Vision and Natural Language Processing, demonstrating impressive accuracy in image categorization, discovery, and recognition. However, these models typically rely on extensive labeled data to achieve high performance. Despite their data-hungry nature, acquiring large datasets, particularly in certain domains and languages, can be very expensive. To the best of our knowledge, there exists a notable deficiency in models tailored for Persian language and culture across various Computer Vision and NLP domains. Additionally, Persian datasets in these domains remain notably limited. Consequently, current research is increasingly focused on leveraging limited data for training, and it seems that the future of artificial intelligence is aimed at reducing the amount of data (Wilson et al.).

Few-shot and Zero-shot learning represent learning approaches that demand minimal data for handling previously unseen categories (Rahman et al., 2017). Although some recently introduced models using these approaches have achieved notable outcomes, the endeavor to improve the performance of the models is still ongoing.

In this paper, we leverage a Transformer-based model called CLIP (Radford et al., 2021) through VQA using Few-shot and Zero-shot learning approaches (Shen et al., 2021) (Song et al., 2022). Our experiments encompasses both English and Persian languages, utilizing the ParsVQA-Caps dataset (Mobasher et al., 2022) which is a benchmark for VQA and Image Captioning in Persian.

2 System Overview

Considering that CLIP model has only text encoder and image encoder, to adapt this model for VQA task, we need to utilize prompt-based learning. We employed several prompt templates in order to compare their results. The architecture of our model is shown in Fig.1. The Persian prompts employed in our study are as follows:

- پرسش: [Q] پاسخ: [A]
- [Q] [A]
- [A] پاسخ [Q] است.

And the English prompts are:

- Retrieve [A] from [Q]
- [Q] [A]
- Question: [Q] Answer: [A]

During the training process, we construct prompts by pairing each question with its corresponding correct answer. These prompts are then used as inputs for the text encoder of the CLIP model. The image is fed into CLIP image encoder. CLIP employs a contrastive learning approach, which is a technique that

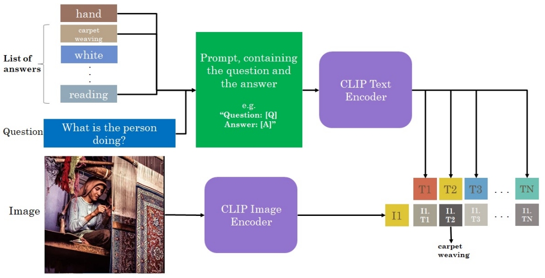

Figure 1: Overview of our model

encourages the model to bring similar pairs (encoded vector of text and encoded vector of image) closer together in the shared latent space while pushing dissimilar pairs further apart. This is achieved by minimizing a loss function that quantifies the cosine similarity between the encoded representations.

$$Cosine(I, T_i) = \frac{I \cdot T_i}{|I||T_i|} \tag{1}$$

On one hand, for English, first we translated the Persian data to English using Google Translate API from deep-translator tool[1]. Then we employed CLIP with VIT-B/32 as its image encoder. On the other hand, for Persian, we utilized the CLIPfa model (Sajjad Ayoubi, 2022) with VIT-B/32 as its image encoder. It is worth highlighting that we employed Zero-shot and Few-shot learning techniques to adapt the model to our specific task, without prior pre-training.

During the test phase, we define the entire set of answers in the test dataset as the valid set, denoted as Z. For each test data, we generate a set of prompts, where we incorporate all possible answers from Z. These completed prompts are then passed through the text encoder, and their representations are compared to the image feature vector using cosine similarity. The label associated with the prompt that yields the highest similarity score is chosen as the correct answer for that particular test instance.

3 Results

We conducted multiple experiments on both Persian and English. Our Few-shot evaluations were conducted with varying shot counts, specifically 2, 4, 8, and 16 shots and the results are shown in Fig.2 and Fig.3.

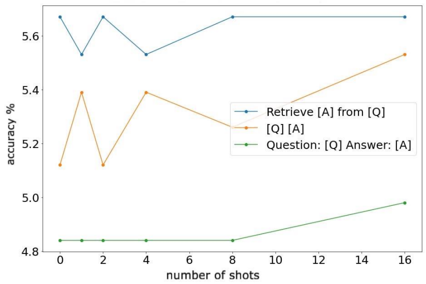

Figure 2: CLIP accuracy on VQA - English

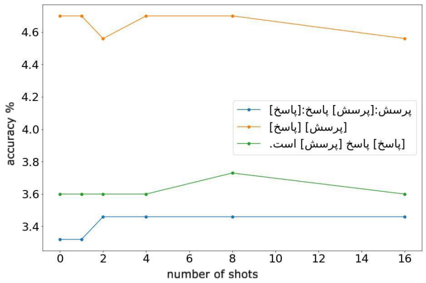

Figure 3: CLIP accuracy on VQA - Persian

Based on the charts, it is evident that the "[Q][A]" prompt yielded the highest accuracy in Persian, while the "Retrieve [A] from [Q]" prompt performed best in English. To provide a more precise assessment, it's important to note that in the English dataset, there are 188 distinct labels. Therefore, the random chance level would be 0.53%, and our top-performing model achieved an accuracy of 5.6%, which is over 10 times better than chance. In contrast, in the Persian dataset, there are 192 distinct labels. Consequently, the random chance level would be 0.52%, and our best model achieved an accuracy of 4.6%, which represents an improvement of approximately 8.8 times over chance. While it was anticipated that increasing the number of shots in Few-shot learning would lead to higher model accuracy, the diagrams for both languages reveal fluctuations in performance.

4 Conclusion

Our study leveraged the CLIP model for Zero-shot and Few-shot learning using a dataset showing Persian culture. We conducted comprehensive experiments with a range of prompts, spanning both English and Persian languages.

[1] https://pypi.org/project/deep-translator/

References

Shaghayegh Mobasher, Ghazal Zamaninejad, Maryam Hashemi, Melika Nobakhtian, and Sauleh Eetemadi. 2022. Parsvqa-caps: A benchmark for visual question answering and image captioning in persian.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Shafin Rahman, Salman Hameed Khan, and Fatih Porikli. 2017. A unified approach for conventional zero-shot, generalized zero-shot and few-shot learning. *CoRR*, abs/1706.08653.

Amir Ahmadi Sajjad Ayoubi, Navid Kanaani. 2022. Clipfa: Connecting farsi text and images. `https://github.com/SajjjadAyobi/CLIPfa`.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2021. How much can CLIP benefit vision-and-language tasks? *CoRR*, abs/2107.06383.

Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. 2022. Clip models are few-shot learners: Empirical studies on vqa and visual entailment.

H. James Wilson, Paul R. Daugherty, and Chase Davenport. The future of ai will be about less data, not more. https://hbr.org/2019/01/the-future-of-ai-will-be-about-less-data-not-more. Accessed: 2023-09-09.