Iran University of Science and Technology
Deep Learning (Spring 2019)

Visual Question Answering in Persian Language

Maryam Sadat Hashemi
Department of Computer Engineering
Iran University of Science
and Technology
m_hashemi94@comp.iust.ac.ir

1 Problem definition

In recent years, there has been a lot of progress in artificial intelligence and deep learning problems at the intersection of Natural Language Processing (NLP) and Computer Vision (CV). One problem that has garnered a lot of attention recently is Visual Question Answering (VQA). Given an image and a question in natural language, the VQA system tries to find the correct answer to it using visual elements of the image and inference gathered from textual questions.

VQA is related to the task of textual question answering. Textual QA has been studied for a long time in the NLP community, and VQA is its extension to additional visual supporting information. The added challenge is significant as images are much higher dimensional and typically more noisy than pure text. Moreover, images lack the structure and grammatical rules of language. Finally, images capture more of the richness of the real world, whereas natural language already represents a higher level of abstraction [3].

In this task, the input is an image and a question based on the image, and the output is one or more words that answer the question.

There are numerous applications for VQA. One of the most significant ones among them is that VQA can be used as an aid for the visually impaired and blind.

To the best of our knowledge, no research has been done on VQA in the Persian language. Our aim in this project is to implement some state-of-the-art architectures designed for the task of the VQA and do experiments on the Persian dataset. For this purpose, no proper dataset is available hence we intend to use the VQA dataset and translate its questions and answers into Persian using existing APIs.

2 Dataset

The Visual Question Answering (VQA) dataset is one of the largest datasets collected from the MS-COCO dataset. The VQA dataset contains at least 3 questions per image with 10 answers per question. The dataset contains 614,163 questions in the form of open-ended and multiple choice(we are going to use only open-ended questions). As mentioned earlier, we intend to provide a Persian VQA dataset Thanks to the existing APIs designed for translation.

3 Evaluation metric

The experimental results will be presented in terms of the **accuracy** of the models over our provided dataset.

(LateX template borrowed from NeurIPS 2019)

Figure 1: Samples from VQA dataset

4 Baseline method

In this section, we describe the Vanilla VQA model [1] which considered as a benchmark for deep learning methods. The Vanilla VQA model uses CNN for feature extraction and LSTM or Recurrent networks for language processing. These features are combined using element-wise operations to a common feature, which is used to classify to one of the answers as shown in figure 2.

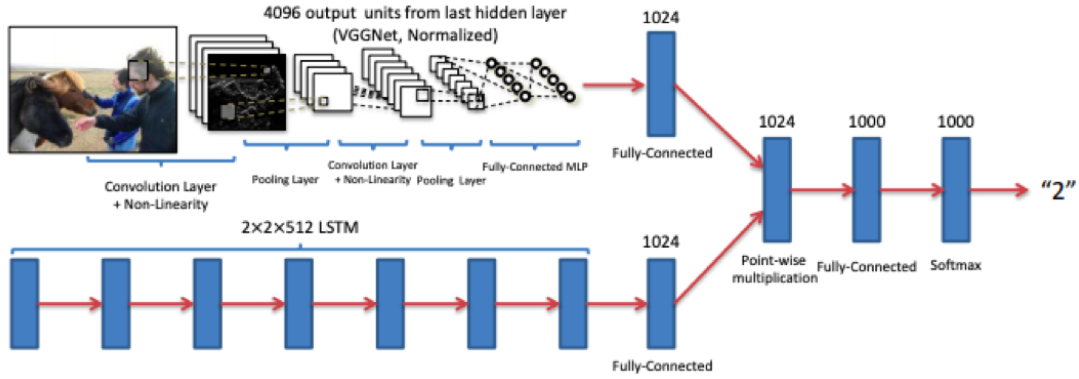

Figure 2: Vanilla VQA Network Model [1]

In this project, we try to implement the Vanilla VQA model and report the computation results across our provided dataset. In the following, According to our time and computational resources, we consider implementing one of the attention-based models such as Stacked Attention Networks [5], Pythia v1.0 [2] and the Differential Network [4](recent method proposed for VQA task that shows very promising performance over different datasets.)

References

[1] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra. Vqa: Visual question answering. *International Journal of Computer Vision*, 123(1):431, Nov 2016.

[2] Y. Jiang, V. Natarajan, X. Chen, M. Rohrbach, D. Batra, and D. Parikh. Pythia v0.1: the winning entry to the vqa challenge 2018, 2018.

[3] Y. Srivastava, V. Murali, S. R. Dubey, and S. Mukherjee. Visual question answering using deep learning: A survey and performance analysis. *ArXiv*, abs/1909.01860, 2019.

[4] C. Wu, J. Liu, X. Wang, and R. Li. Differential networks for visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8997–9004, 2019.

[5] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.