# Project One Report: Visual Question Answering Investigation

Zihang Jiang
jzh15@bu.edu

## I. INTRODUCTION

Visual question answering (VQA) is a field that combines computer vision and natural language processing (NLP). For a common VQA task, with an image and a textual question about the image given as inputs, the goal is to give a correct answer to the question, often in one or a few words.

Compared to image caption [2]–[4] which is also a combination of computer vision and NLP, VQA is much more difficult to deal with as image caption only need to describe the content of the image in text while VQA usually needs external knowledge like common sense or specific knowledge base to answer the question not merely limited to the range of the image.

Compared to traditional CV tasks, VQA raises a higher demand on image information extraction. For other CV tasks such as segmentation or object detection, the single question that they need to answer is predetermined and only the input image changes [1]. However, the question to be answered in VQA is unknown until execution, which means that features of the image have to be extracted as much as possible as VQA models don't know what is needed when answering the question.

As for the use cases of VQA, I think VQA is a field much closer to artificial intelligence (AI) in the future that people always imagine in science fiction and movies than other machine learning tasks. How exciting it is that we can talk to machines and they can answer your questions based on merely a picture! At present, VQA can be applied to helping blind people understanding the world like VizWiz challenge [5]. For blind people who are not able to see the world in person, real world photos can be sent to VQA models and they can ask the models what the world look like. Except for that, counting the number of items in the image is also a field that VQA can be applied to. We human can easily count items in the image when the number of items is small like no more than 10. But when the number of things that need to be counted is very large, for example, giving an answer about the number of buildings over ten stories tall in a super high-resolution city image, it's quite hard for us human to count one by one, while a VQA model can give the correct answer easily and fast. Background recognition is another very interesting use case for VQA which has great potential. Imagine a scene that a criminal take a selfie in flight and post it on social network. Police can ask VQA model where the criminal is based on the background of the selfie and VQA model can give the correct answer based on external knowledge base like google map data, which helps police to grasp the criminal very fast. These scenes can be often seen in the movies, but with VQA models, it seems that they will come true very soon!

## II. RELATED WORK

### A. Datasets

A great deal of effort of research community has gone into designing proper datasets for VQA and there are a variety of datasets put forward since 2014. A VQA dataset is developed for training and evaluating VQA models, so for different VQA use cases different datasets need to be put forward. The size of dataset is a very important metric to measure the quality of the dataset. An ideal VQA dataset has to be large enough to capture the variability within questions, images, and concepts that occur in real world scenarios [6]. But when the dataset become too large, it's hard to develop and VQA models will spend a lot of time training. So having a proper size is of importance to VQA datasets. Here I will introduce several common-used VQA datasets.

**DAQUAR:** The DAtaset for QUestion Answering on Real-world images (DAQUAR) [7] is the first dataset put forward for VQA. It contains 1449 images from NYU-Depth v2 dataset [8] and 12468 question-answer pairs, which is quite small so that it can't be used to train complex VQA models. Another disadvantage is that NYU-Depth v2 dataset contains mainly indoor scenes, leading to low variety of DAQUAR.

**VQA Dataset:** VQA dataset [9] is one of the largest and most widely used datasets which contains two parts: VQA-real containing 204721 images from MS COCO dataset [10] and VQA-abstract containing 50000 abstract scenes. VQA-real is the portion that most works work on consisting of three questions per image with ten answers per question, which has 614163 questions and 7984119 answers in total. Though VQA dataset has a great variety of questions, there is a disadvantage that many questions can be answered without using the image due to language biases, which causes some simple image-blind algorithms achieving good performance on it [6].

**CLEVR:** CLEVR [11] is a synthetic dataset with only three object cylinder, sphere and cube to generate images. These objects have two different sizes, two different materials and eight different colors. CLEVR has 100000 images and 853554 questions in total, which is also quite large. Compared to real-world datasets like DAQUAR and VQA-real, synthetic datasets focus on high-level information instead of visual details to study connections between computer vision and NLP.

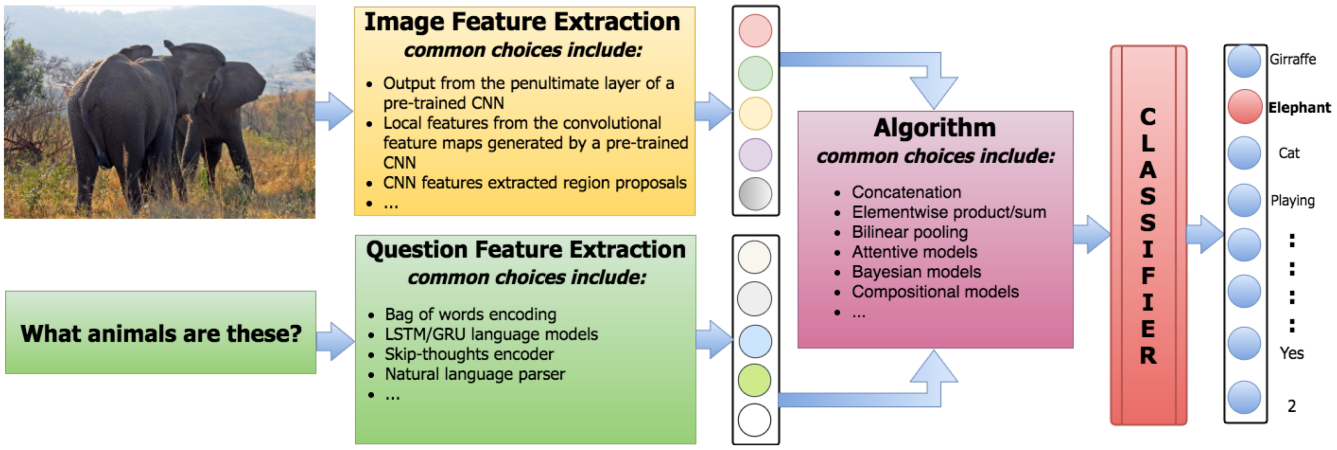


Fig. 1. Simplified illustration of the classification based framework for VQA [6]

**KVQA:** KVQA [12] is a knowledge based VQA dataset which contains 24000 images and 183100 question-answer pairs. Question in KVQA require multi-entity, multi-relation, and multi-hop reasoning over large knowledge graph (KG) to reach a solution [13], which are so-called commen-sense questions.

### *B. VQA Methods*

Though VQA datasets are of importance, they are still used to train and evaluate VQA systems, while VQA models and methods do solve the problem. So the main focus of research community is always on developing advanced VQA methods. Since the emergence of deep learning, both computer vision and NLP have gained great progress towards their separate goals, such as VGGNet [14], Inception [15], ResNet [16] in computer vision and RNN [17], LSTM [18] in NLP. As a combination of both fields, VQA has also seen great advances due to a marriage of efforts from both fields [19]. In this section I will introduce several major VQA methods.

**Joint Embedding Approach:** Almost all existing VQA models can be divided into three procedures: image feature extraction, question feature extraction and feature combination [6], which is called joint embedding approach as image embedding and question embedding are joined together. To reach an answer, the most common approach is to treat VQA as a classification problem from predefined answers [6], shown as Fig.1, while some other VQA models use LSTM to generate the answer one word at a time [20], [21]. For image feature extraction, a pre-trained CNN like VGGNet, ResNet and Inception can be used. For question feature extraction, a variety of question featurization methods like bag of words (BOW), LSTM and GRU [22] can be used. Though joint embedding approaches are quite straight forward to understand, there are disadvantages that they use the whole image containing lots of irrelevant information leading to performance loss and the answer generated from both classification and LSTM can't exceed the range of training dataset range.

**Attention Based Models:** If VQA models use the whole image to extract features, there will exist many irrelevant information like noises and other areas unrelated to the question, which makes the introduction of attention mechanism meaningful. Attention based models [23], [24] give each area of the image a corresponding weight that reflects the relevance to the question. In the running time the areas with low weight value will be abandoned, which helps the VQA system focus on the related image regions to improve performance. Though attention mechanisms improve the overall accuracy on all VQA datasets, closer inspection by question type show little or no benefit on binary (yes/no) questions [19].

**Bilinear Pooling Methods:** Many joint embedding approaches only use simple algorithms to join image features and question features, like concatenation, element-wise sum and point-wise multiplication, which aren't able to fuse the information from image and question fully. So researchers developed other more complex alternatives like outer-product. But using outer-product directly is very high-dimensional and computation-costly. Bilinear pooling methods like [25] proposed a multi-modal low-rank bilinear pooling (MLB) scheme that uses the Hadamard product and a linear mapping to achieve approximate bilinear pooling. Though MLB has reduce the computation and neural network parameters a lot, I think there is still room for further reduction.

## III. Reimplementation

I reimplemented MLB method using the open source codes from Cadene[1]. The reimplementation results and details can be seen in my VQA repository[2]. Cadene et al. put forward MUTAN [26] which is also a bilinear based VQA model

[1]https://github.com/Cadene/vqa.pytorch
[2]https://github.com/Zihang97/vqa.pytorch

and they use MLB as baseline to compare. I choose MLB to reimplement as MLB is one of the earliest bilinear pooling methods and has fairly good performance which is very suitable to serve as a baseline.

## IV. CONCLUSIONS

In this report I explain the definition and essence of VQA field and come up with my opinions about the use cases of VQA in section I. In section II I introduce and conclude the related works these years about VQA from two prospectives: datasets and VQA methods. Then I give a description of my reimplementation of MLB method in section III.

For the future paths of VQA, I think there are still two prospectives: datasets and methods. But these two directions are not separate from each other, in contrast, they have strong connections. When a new use case of VQA is discovered, the new VQA datasets and advanced VQA methods should be developed at the same time.

## REFERENCES


[1] G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in Plastics, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

[2] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2015.

[3] A. Karpathy, A. Joulin, and F. F. Li. Deep fragment embeddings for bidirectional image sentence mapping. In Proc. Advances in Neural Inf. Process. Syst., 2014.

[4] J. Mao, W. Xu, Y. Yang, J. Wang, and A. Yuille. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). In Proc. Int. Conf. Learn. Representations, 2015.

[5] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. CVPR.

[6] Kafle, K., Kanan, C. Visual question answering: Datasets, algorithms, and future challenges. Computer Vision and Image Understanding 163, 3–20 (2017)

[7] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In Proc. Advances in Neural Inf. Process. Syst., pages 1682–1690, 2014.

[8] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In Proc. Eur. Conf. Comp. Vis., 2012

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in IEEE CVPR, 2016, pp. 770–778.

[9] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In Proc. IEEE Int. Conf. Comp. Vis., 2015.

[10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In ECCV, 2014.

[11] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in IEEE CVPR, 2017, pp. 2901–2910.

[12] S. Shah, A. Mishra, N. Yadati, and P. P. Talukdar, "Kvqa: Knowledgeaware visual question answering," in AAAI, 2019.

[13] Yash Srivastava, Vaishnav Murali, Shiv Ram Dubey, and Snehasis Mukherjee. Visual question answering using deep learning: A survey and performance analysis. arXiv preprint arXiv:1909.01860, 2019.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in IEEE CVPR, 2016, pp. 2818–2826.

[17] L. R. Medsker and L. Jain, "Recurrent neural networks," Design and Applications, vol. 5, 2001.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Visual Question Answering: A Survey of Methods and Datasets. arXiv preprint arXiv:1607.05910, 2016a

[20] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? Dataset and methods for multilingual image question answering," in Advances in Neural Information Processing Systems (NIPS), 2015.

[21] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in The IEEE International Conference on Computer Vision (ICCV), 2015.

[22] K. Cho, B. Van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In Conference on Empirical Methods on Natural Language Processing (EMNLP), 2014.

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

[24] I. Ilievski, S. Yan, and J. Feng, "A focused dynamic attention model for visual question answering," arXiv preprint arXiv:1604.01485, 2016.

[25] J. Kim, K. On, J. Kim, J. Ha, and B. Zhang, "Hadamard product for low-rank bilinear pooling," arXiv preprint arXiv:1610.04325, 2016.

[26] H. Ben-younes, R. Cadene, M. Cord, and N. Thome. MUTAN: multimodal tucker fusion for visual question answering. In Proc. IEEE Int. Conf. Comp. Vis., 2017. 8, 13