Literature Review of Visual Question Answering (September 2021)

Fei Gao

***ABSTRACT*—In recent years, deep learning technology has greatly advanced the field of artificial intelligence (AI), including natural language processing (NLP) and computer vision (CV). People hope that machines can think and communicate like humans. Humans can make decisions by seamlessly combining sound, sight, language, and many other multiple forms of stimulation together. So it becomes an important next step for artificial intelligence to involve multiform stimulation, especially language and vision[1]. Visual question answering (VQA) is a new area that combines applications of artificial intelligence for computer vision and natural language processing. It aims to answer different types of questions expressed in natural language and targeted at any image[2]. This paper focuses on introducing VQA and reviewing a project based on VQA.**

Index Terms—Visual question answering

I. Introduction

Visual Question Answering (VQA) is a natural language Question and Answer for Visual images. As a research direction of Visual Understanding, it connects vision and language[3]. Most current tasks on images do not require a complete understanding of the information contained in the image. Such as image classification, object detection, action recognition, etc. Solving the VQA problem requires a complete understanding of the image. The VQA system takes images and any form of open-ended natural language questions about images as input and natural language answers as output. This goal-driven task is suitable for situations where visually impaired users or intelligence analysts actively acquire visual information. Here are some examples of visual question answering.

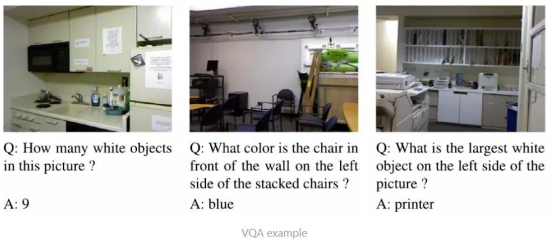

Figure 1: VQA example[3]

To better understand visual question answering. I reviewed a project about visual question answering. This project call Where To Look: Focus Regions for Visual Question Answering. It was completed by Kevin J. Shih, Saurabh Singh, and Derek Hoiem and published in 2015. In that project, they propose a method for learning to answer visual questions by selecting image regions associated with text-based queries[4].

II. Literature review

A. Goal

This project is focused on a key problem for VQA and other visual reasoning tasks: knowing where to look. The goal is to get the right answers to natural language questions, such as “What color is the walking light?” or “Is it raining?” This is difficult because it requires vision and learning to identify objects, use relationships, and determine correlations. For example, rain can be determined by detecting the presence of puddles, gray skies, or umbrellas in the scene, while the color of the walking light requires focusing on the light itself. When answering questions like this, the system should know the type of answer expected and wherein the image its response should be based[4].

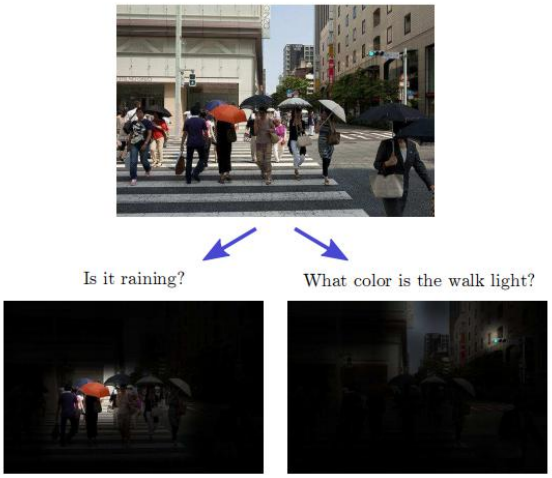

Figure 2: An example of attention regions[4].

B. Works

To achieve the goal, contributors set an image region selection mechanism to learn and recognize the image regions related to the problem. And they propose a learning framework for solving multi-choice visual QA based on marginal loss, with significantly better performance than the baseline provided. Besides, they also provide detailed comparisons with various baselines to highlight exactly when our region selection model improves VQA performance. And the method is to put the text problem and set of visual image regions into a potential space in which the internal product produces a correlation weighting for each region[4].

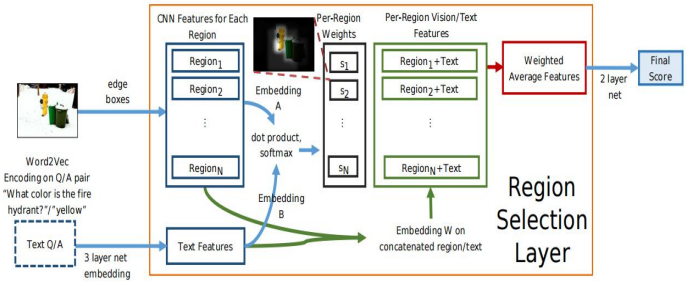

Figure 3: Overview of the question-answer pairing[4]

C. Experiment

The project evaluates the influence of the regional selection layer on the multiple-choice format of the MS COCO VQA dataset. And analyze how much accuracy improved in the selected region weights compared to the entire image or just using the language[4]. The first table shows a comparison of overall accuracy on validation sets. It can be found that the model they proposed is the best. The second table includes comparisons with the best performing problem - and image-based models in the VQA dataset paper, and here again, their proposed model is better than the others.

Model	Overall (%)
Word Only	53.98
Word+Whole Image	57.83
Word+Ave. reg.	57.88
Word+Sal. reg.	58.45
Word+Region Sel.	**58.94**
LSTM Q+I [1]	53.96

Model	All	Y/N	Num.	Others
test-dev				
LSTM Q+I [1]	57.17	**78.95**	35.80	43.41
Q+I [1]	58.97	75.97	34.35	50.33
iBOWIMG [24]	61.68	76.68	**37.05**	54.44
Word+Region Sel.	**62.44**	77.62	34.28	**55.84**
test-standard				
iBOWIMG [24]	61.97	76.86	**37.30**	54.60
Word+Region Sel.	**62.43**	**77.18**	33.52	**56.09**

Figure 4:Accuracy comparison on VQA test sets[4].

III. Conclusions

The project proposed a model that selects areas from the image to solve the visual problem and answer the question. VQA, which is between Vision and NLP, is an interesting and challenging question that requires visual understanding and reasoning ability. Its progress depends not only on the development of computer vision and the ability to process natural language but also on the understanding of images -- basic visual abilities such as recognition and detection, as well as the ability to learn knowledge and reason[3]. The study of visual question answering tasks has many practical applications, such as it can help the blind and visually impaired people to get more information on the Internet or in the real world, and even real-time human-computer interaction, which will greatly improve the living conditions and convenience of the blind and visually impaired people. It improves the way of human-computer interaction, can query visual content through natural language, expand the intelligent robot question and answer function. Visual question answering systems can also be used in the field of

image retrieval and basic AI problem[5]. Visual question answering belongs to the intersection of computer vision and natural language processing. Until then, however, computer vision and natural language processing developed separately. Although some achievements have been made in visual question answering research, there is still a long way to go in terms of the effect it can achieve so far. A visual question and answer model that truly understands images, learns knowledge and reasoning skills is the ultimate goal.

REFERENCES

[1] Ramakrishnan, S. K., Pal, A., Sharma, G., & Mittal, A. (2017). An empirical evaluation of visual question answering for novel objects. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4392-4401).

[2] Barra, S., Bisogni, C., De Marsico, M., & Ricciardi, S. (2021). Visual Question Answering: which investigated applications?. arXiv preprint arXiv:2103.02937.

[3] Xiao Feng Xiao Yu. https://www.jianshu.com/p/76d2e081e303. From Jane Book.

[4] Shih, K. J., Singh, S., & Hoiem, D. (2016). Where to look: Focus regions for visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4613-4621).

[5] Bao XG, Zhou CL, Xiao KJ, Qin B. Survey on Visual Question Answering. Journal of Software, 2021, 32(8): 2522-2544(in Chinese). http://www.jos.org.cn/1000-9825/6215.htm