Literature Review of Visual Question Answering and Exploration.

Kehan Guo

Email: gkehan@bu.edu

***Abstract*— Historically, the establishment of a system that can answer natural language questions about images has always been recognized as a promising direction. Visual Question Answering (VQA) , as a new computer science brach which combines multiple fields ranging from natural language processing (NLP) , computer vision and other branches of artificial intelligence, has made attractive breakthrough these years. In this paper, we will give detailed background knowledge of VQA, providing a bunch of developing directions of the field and do a specific case study conducted by Antol[1].We will do the evaluation by going through the technical details of the model, analyzing its advantages a Finally, this paper will provide a prediction on VQA as well as exploring potential developing directions in this field.**

Keywords - Visual Question Answering(VQA), Computer Vision.

I. INTRODUCTION

The VQA system needs to use pictures and problems as input, combining these two parts of information to produce a human language as output. Convolutional Neural Network(CNN) and Recurrent Neural Network(RNN) are the most common methods that scientists implement to do image processing. CNN takes an images as input and process them based on their shared-weights architecture and translation invariance characteristics[2], which significantly reduces the computational cost because some weights are shared in some parameters. For Recurrent Neural Network(RNN), where a sequence of problems are more important than single piece of word, it is usually used to deal with natural language processing(NLP). These tools are wildly implemented in most VQA studies. Apart from that, question design and dataset selection also plays a significant role in VQA study. Previous research has shown its limitation regarding these aspects as they set the question in a certain situation or picking up the answers in a given dataset [1]. VQA today tends to generate answers that are open-ended and of diversity, which involves robust AI capacity to guarantee its precision.

II. RELATED WORK

A. Convolutional Neural Network(CNN)

In deep learning, a convolutional Neural Network is a class of deep neural networks, most commonly applied to analyzing visual imagery[2]. In VQA field, CNN are mostly commonly used to shape the size of image by assigning weigh to various objects in the image and be able to differentiate it from one another. Its working mechanism can be shown in figure 1. In a study presented by Rui Hou, a 3D deep convolutional neural network (3D CNN) to evaluate video quality without reference by generating spatial/ temporal deep features within different video clips 3D CNN is designed by collaboratively and seamlessly integrating the features output from VGG-Net on video frames[3].

B. Recurrent Neural Network(RNN)

Recurrent Neural Network is a model designed to solve a sequence of message.In RNNs, inputs can be stored in a internal memory to do the next-step processing(shown in figure 2a) It is firstly implemented to language modeling by Mikolov, and speech recognition by Graves et al[4]. However, in real world applications, RNN doesn't compile well with feedforward neural network, which results in a fact that current RNNs are very small as a tiny increase could cause the system to overfit. Usually, RNN works with a Long Short-term Memory Units(LSTM), (figure2b), and in the study by Wojciech Zaremba, they demonstrated how great would the overfitting be reduced when using RNN correctly.[5]

C. Dataset Selection

Choosing the right dataset in VQA is of great importance. Currently, the most popular two dataset for VQA study are Microsoft Common Objects in Context (MS COCO) Dataset and DAtaset for QUestion Answering on Real-world images (DAQUAR). In [1], Antol conducted their experiment based on images from MS COCO not only in real image(pictures taken in real world), but also images

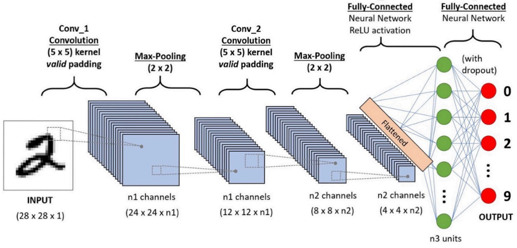


Figure1: A CNN sequence to classify handwritten digits[5]

in cartoon or computer generated to enrich the diversity of the research. "The more diverse our collection of images, the more diverse, comprehensive, and interesting the resultant set of questions and their answers" [1]

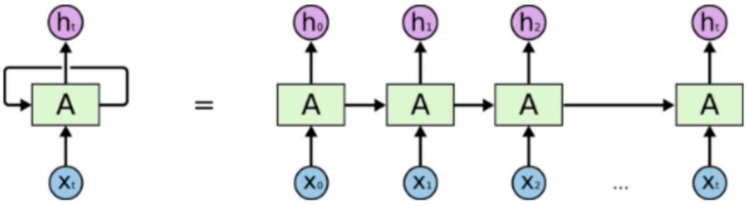


Figure 2a: unrolled recurrent neural network [6]

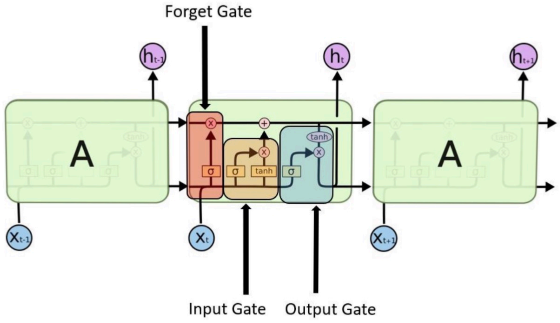


Figure 2b:LSTM gates[6]

III. VQA EXPERIMENT DESIGN BY ANTOL

Before Antol published the paper "VQA:Visual Question Answering", several related research regarding image recognition and natural language were going on, however, few of them are designed to generate open-ended and free-form answers[1]. In the research, the design of Antol's experiment could be roughly divided into 4 parts:

A. *Image selection*

MS COCO dataset is selected to conduct the experiment. The experiment includes not only real world images but also abstract scenes. Those abstract scenes are created by using 20 "paperdoll" models ranging from different background such as gender, age and race[1]. And the structure of the models are adjustable and a use of clipart in implemented to allow a creation of different situation such as indoor or outdoor. Examples are shown in figure 3.

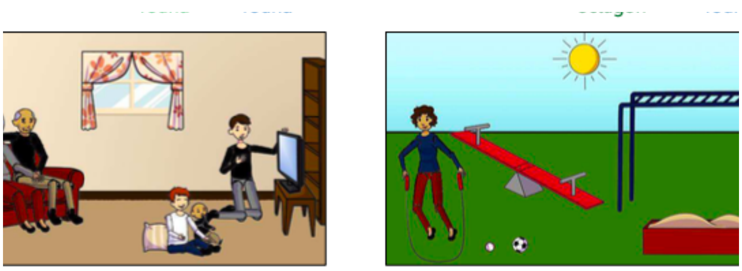


Figure 3: abstract scene example[1].

B. *Image pre-processing*

For real images, they have already been pre-labelled into classified categories in MS COCO Dataset at the time they are captured. In each categories, the images are verified and segmented[7]. An imperative property of the dataset is that it endeavors to discover non-iconic pictures containing objects in their normal setting, which provides more objects in image and increases the diversity of the VQA experiment.

For abstract images, according to [1], they create standard splits, separating the scenes into 20K/10K/20K for train/val/test splits, respectively.

C. Question design

Design questions in VQA is very challenging because the questions set should contain not only simply questions such as "what is the color of the tree", but also questions that will even be considered hard to human, like "what sound does this animal make"[1]. Also, question should interact with images to answer instead of merely using a commonsense, like in figure 4, the VQA system should know the mustache refers to bananas.

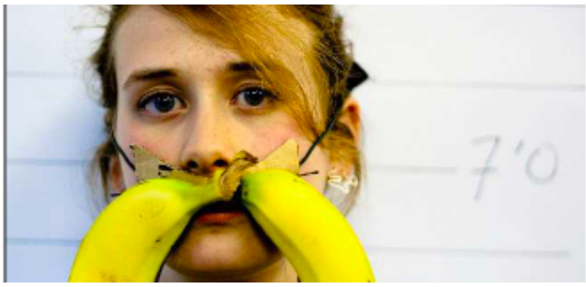


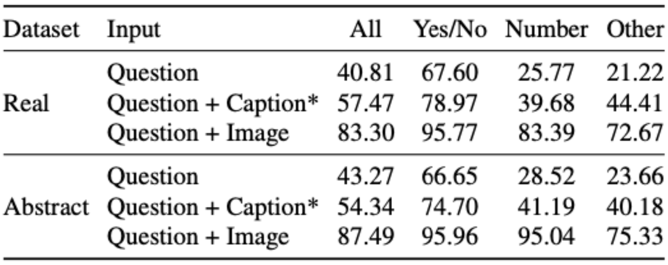
Figure 4: what is the mustache made of [1]

D. *Answering*

In this research, because we designed an open-ended question set, the answer to it could be various. As a result, two types of answers are defined: open answer and multiple choices. And the performance of the system is examined by comparing the answers with human reaction using the following accuracy metric: min (human that provided that answer /3,1), i.e., an answer is deemed 100% accurate if at least 3 workers provided that exact answer[1].

E. Performance

In Antol's research, accuracy were examined under 3 ciwcumasndances, questions , question with captions and question with images for both real images and abstract images. In [1], we find that the distributions of nouns, verbs, and adjectives mentioned in captions is statistically significantly different from those mentioned in our questions + answers (Kolmogorov- Smirnov test, $p < .001$) for both real images and abstract scenes. See supplementary material for details. The result are shown in figure 5.



Dataset	Input	All	Yes/No	Number	Other
Real	Question	40.81	67.60	25.77	21.22
	Question + Caption*	57.47	78.97	39.68	44.41
	Question + Image	83.30	95.77	83.39	72.67
Abstract	Question	43.27	66.65	28.52	23.66
	Question + Caption*	54.34	74.70	41.19	40.18
	Question + Image	87.49	95.96	95.04	75.33

Figure 5:Test-standard accuracy of human subjects when asked to answer the question without seeing the image (Question), see- ing just a caption of the image and not the image itself (Question + Caption), and seeing the image (Question + Image). Results are shown for all questions, “yes/no” & “number” questions, and other questions that are neither answered “yes/no” nor number. All answers are free-form and not multiple-choice. *These accuracies are evaluated on a subset of 3K train questions (1K images)[1].

IV. Path forward

A. Image recognition with higher accuracy. This would definitely increase the implementing value of this technology. For example, the VQA could be used in endangered animal detection and protection by simply getting images or video flows from the inhabitants. The VQA could save Bunch of time for human tracing and at the same time provide valid information to animal protectors.

B. Direct &Indirect commercial benefit

a) Searching things you don't know will be much easier. Image you saw someone wearing a new fancy shoes in the street and you failed to communicate due to whatever reason, the simplest way for you to figure it out is to take a picture with your phone and search with VQA. This technology would strongly boost company like amazon for it could provide customers with quick and convenient buying options. Or social medias like twitter where the back stage could label any image people shared in the platform, then for the system to list people to different categories and sold this information for adverting profits.

References

1. Antol, Stanislaw et al., 2015. VQA: Visual Question Answering. , pp.2425–2433.
2. https://en.wikipedia.org/wiki/Convolutional_neural_network
3. Hou, Rui, Zhao, YunHao, Hu, Yang, and Liu, Huan. "No-reference Video Quality Evaluation by a Deep Transfer CNN Architecture." Signal Processing. Image Communication 83 (2020): 115782. Web.K. Elissa, “Title of paper if known,” unpublished.
4. Mikolov. Statistical language models based on neural networks. PhD thesis, Brno Uni- versity of Technology, 2012.
5. https://towardsdatascience.com/understanding-rnn-and-lstm-f7cdf6dfc14e
6. https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53
7. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dolla´r, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014.