1. Introduction

   Visual Question Answering (VQA) is the system that combines Knowledge Representation & Reasoning, Computer Vision, and Natural Language Processing. [1] In this system, it takes images and natural language as input and gives the natural language answers as output. VQA could be very helpful when it is applied to scenarios of object detections or acquisition of information. However, VQA tasks could be challenging due to the size of datasets and some other factors.

2. Application

   Visual Question Answering has been widely used in different real-life areas. The most direct application is to help visually impaired persons; thus, they could ask questions by just taking photos. Also, it could be used in the medical or biological field by detecting X-rays or some other medical images to tell information.

3. Literature Review

   A lot of VQA papers and literature reviews suggest that it would be a multi-discipline problem to combines Knowledge Representation & Reasoning, Computer Vision, and Natural Language Processing to get the final distribution of answers.

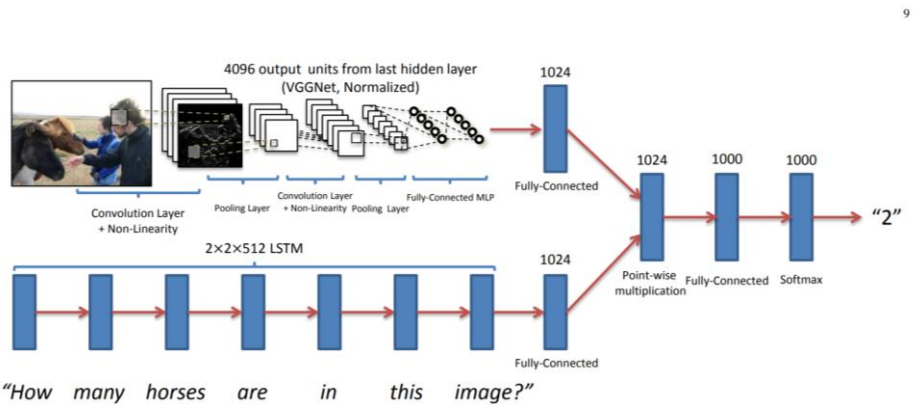
   3.1 Datasets

   Images: The image dataset is from Microsoft Common Objects in Context (MS COCO) datasets[2], which has images with multiple objects and wide contextual information. [1]

   Question and answers: In COCO dataset, it has already included questions and answers for each image, which are around 760,000 questions and 10,000,000 questions.

   3.2 Approaches outline in VQA

   According to Antol [1], there are two channels in the model, including image channel for embedding image features and question channel for embedding question features. Convolutional Neural Network (CNN) will be used in the image channel and Long Short-Term Memory (LSTM) will be used in the question channel. Then, two features will be fused by element-wise multiplication and transformed to a fully connected layer followed by a softmax layer.



4. Open-Source Research

   An example of CNN&LSTM combination is the “Deeper LSTM Q + Norm I” from tbmoon Github implemented by Pytorch [3]. Also, the combination of a LSTM channel for question features and a normalized VGGNet for image features shows the best performance. The accuracies of answers are shown as follows:

	Open-Ended				Multiple-Choice			
	All	Yes/No	Number	Other	All	Yes/No	Number	Other
prior (“yes”)	29.66	70.81	00.39	01.15	29.66	70.81	00.39	01.15
per Q-type prior	37.54	71.03	35.77	09.38	39.45	71.02	35.86	13.34
nearest neighbor	42.70	71.89	24.36	21.94	48.49	71.94	26.00	33.56
BoW Q	48.09	75.66	36.70	27.14	53.68	75.71	37.05	38.64
I	28.13	64.01	00.42	03.77	30.53	69.87	00.45	03.76
BoW Q + I	52.64	75.55	33.67	37.37	58.97	75.59	34.35	50.33
LSTM Q	48.76	78.20	35.68	26.59	54.75	78.22	36.82	38.78
LSTM Q + I	53.74	78.94	35.24	36.42	57.17	78.95	35.80	43.41
deeper LSTM Q	50.39	78.41	34.68	30.03	55.88	78.45	35.91	41.13
deeper LSTM Q + norm I	**57.75**	**80.50**	**36.77**	**43.08**	**62.70**	**80.52**	**38.22**	**53.01**
Caption	26.70	65.50	02.03	03.86	28.29	69.79	02.06	03.82
BoW Q + C	54.70	75.82	40.12	42.56	59.85	75.89	41.16	52.53

TABLE 2: Accuracy of our methods for the open-ended and multiple-choice tasks on the VQA test-dev for real images. Q = Question, I = Image, C = Caption. (Caption and BoW Q + C results are on val).

[1] Accuracy of different combinations of models

From the table above, it is obvious that the combination of deeper LSTM Q + norm I has the highest accuracies for both Open-Ended answers and Multiple-Choice answers.

REFERENCES:

[1] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 2425-2433).

[2] 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. 2, 3, 5, 21

[3]Tbmoon. “Pytorch VQA: Visual Question Answering”. Github, 2019. tbmoon/basic_vqa: Pytorch VQA : Visual Question Answering (https://arxiv.org/pdf/1505.00468.pdf) (github.com)