EC601Project1 BU ID:U96545272. Name :Songlan Wang

Report of VQA

Visual Question Answering is one of the multi-discipline AI problems. People pay much attention to such problems because they can combine the most advanced technologies like Computer Vision (CV), Natural Language Processing (NLP) or Knowledge Representation. People have a better understanding of the technologies. A VQA system takes as input an image and a free-form, open-ended, natural-language question about the image and produces a natural-language answer as the output.[1] VQA is widely used in our daily life. For people with visual impairment, they can hardly see things clearly. With the help of VQA, they can get the correct answer from the VQA system. When they cross the road, they can ask VQA what the current signal is. So they can cross the road safely. Moreover, researchers will also be much more efficient than before. When researchers need to deal with a large number of image-related questions, they can use the VQA system to automatically generate answers to these image questions. With the widespread use of VQA, there will be a significant increase in the awareness of people with disabilities in society. Overall, there is a great deal of research to be done in the area of VQA.

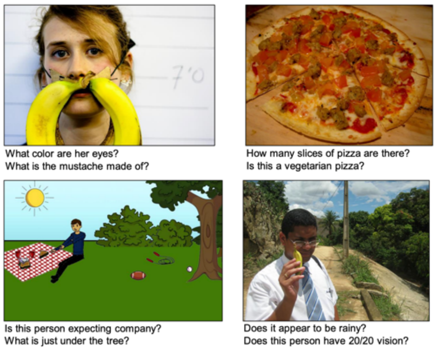

What color are her eyes?
What is the mustache made of?
How many slices of pizza are there?
Is this a vegetarian pizza?
Is this person expecting company?
What is just under the tree?
Does it appear to be rainy?
Does this person have 20/20 vision?

Fig. 1: Examples of free-form, open-ended questions collected for images via Amazon Mechanical Turk. Note that commonsense knowledge is needed along with a visual understanding of the scene to answer many questions.

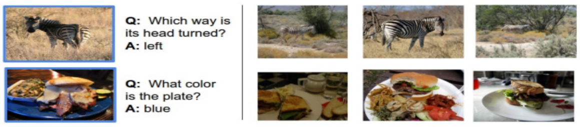


Figure 5: Three counter-example or negative explanations (right three columns) generated by our model, along with the input image (left), the input question Q and the predicted answer A.

Since VQA is a huge theme, we can approach it in two ways. We divide questions to open-ended answering task and a multiple-choice answering task. [2][3] The multiple-task are required to choose a best answer from a list of possible answers. VQA requires very large data sets. Commonly, the researcher tends to use MS COCO dataset. The MS COCO is able to support large-scale object detection, segmentation, key-point detection and captioning dataset.[4] Why do they choose this dataset? It is because we can not only find the so many images but also it contains 50000 newly created scenes. These are the main factors that allow us to ask questions. We don't just want to ask basic, low-level questions, but questions that require common sense to answer. VQA system not only needs to look at the pictures to get the answers, but also needs to add commonsense to answer the questions. Apart from the question, it is also important to reflect on the answer. For the main part of the questions, simple yes or no. The most important part is the question that require the short phrase and it is likely that several answers is correct to the same question. In order to solve the problem, the researchers proposed that they can pick 10 answers for each question from unique workers. After that comes the very important question of whether we need images to solve the problem. Using commonsense knowledge is definitely critical to the VQA and people must take advantages of using the image. Only the commonsense information is far from sufficient.

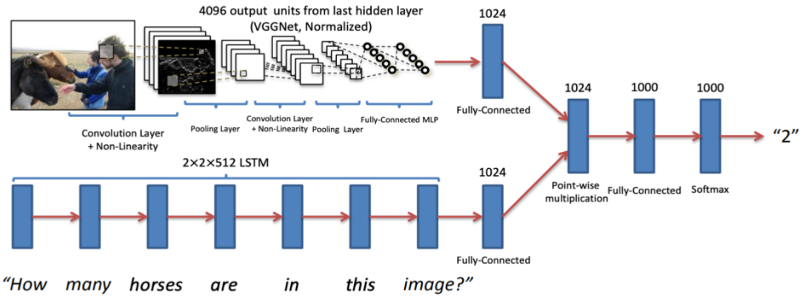
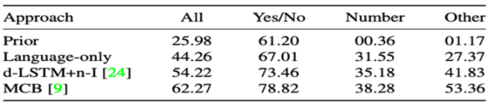


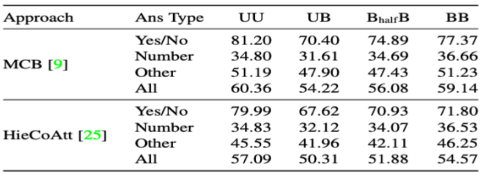
Fig. 8: Our best performing model (deeper LSTM Q + norm I). This model uses a two layer LSTM to encode the questions and the last hidden layer of VGGNet [48] to encode the images. The image features are then ℓ_2 normalized. Both the question and image features are transformed to a common space and fused via element-wise multiplication, which is then passed through a fully connected layer followed by a softmax layer to obtain a distribution over answers.

To begin the research, we need to choose the answer randomly. Moreover, 'yes' is always the needed choice to be put in the answer set. Q-type prior and KNN algorithm are also important. The researchers propose the method that use a 2-channel vision(image) + language (question) model. When it comes to Image channel, we divide the it to 2 types. The former one is I and latter one is norm I. When it comes to question channel, the research proposes the BoW Q, LSTM Q, deeper LSTM Q. According to the result of the completed research, the best model is deeper LSTM Q and norm I. After consulting the paper about the VQA, LSTM and norm I are both preferred algorithms. For a better adaptation of balanced dataset, we find that the use of MCB and HieCoAtt algorithm are better than LSTM and norm I. But the most valuable thing is how to we balance the dataset. The researchers counter the image bias. They identify an image I' that is similar to I, but results in the answer to the question Q to become A0 (which is different from A). [4] The researchers proposed the 'counter-example explanation' which is different from the conventional VQA way by retrieving images that are similar to I but not the same as the answer Q to the question. Because of this, researchers are able to address the strong language and elevate the role of image understanding.[4]



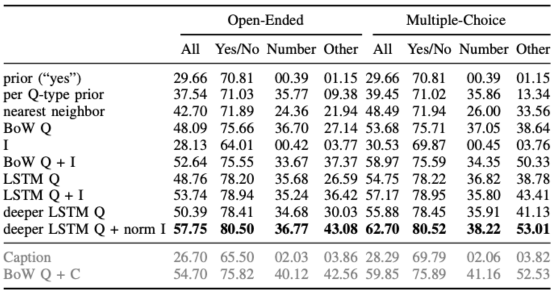
Approach	All	Yes/No	Number	Other
Prior	25.98	61.20	00.36	01.17
Language-only	44.26	67.01	31.55	27.37
d-LSTM+n-I [24]	54.22	73.46	35.18	41.83
MCB [9]	62.27	78.82	38.28	53.36

Table 2: Performance of VQA models when trained on VQA v2.0 train+val and tested on VQA v2.0 test-standard dataset.



Approach	Ans Type	UU	UB	$B_{half}B$	BB
MCB [9]	Yes/No	81.20	70.40	74.89	77.37
	Number	34.80	31.61	34.69	36.66
	Other	51.19	47.90	47.43	51.23
	All	60.36	54.22	56.08	59.14
HieCoAtt [25]	Yes/No	79.99	67.62	70.93	71.80
	Number	34.83	32.12	34.07	36.53
	Other	45.55	41.96	42.11	46.25
	All	57.09	50.31	51.88	54.57

Table 3: Accuracy breakdown over answer types achieved by MCB [9] and HieCoAtt [25] models when trained/tested on unbalanced/balanced VQA datasets. UB stands for training on **U**nbalanced train and testing on **B**alanced val datasets. UU, $B_{half}B$ and BB are defined analogously.



	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). See text for details.

Other aspect of the VQA is also tackled. For instance, some researchers focused on solving the binary question of the VQA problem. They propose the approach that extracts a concise summary of the question in the tuple, identifying the region that it ought to focus on and verified the existence of the visual concept of the question.

All in all, there is a long way to fully resolve the VQA problem. From the several papers referenced, they have given a basic implementation of the research VQA method and have each optimized the solution from a different perspective. There are also many web resources that I can go to for reference. I was prompted to at least look at problem and dataset optimization to solve the VQA problem. I'm looking forward to going through VQA with the group!

[1] M. Richardson, C. J. Burges, and E. Renshaw. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In EMNLP, 2013. 1, 2

[2]X. Lin and D. Parikh. Don’t Just Listen, Use Your Imagination: Leveraging Visual Common Sense for Non-Visual Tasks. In CVPR, 2015. 1

[3]Lin, TY. et al. (2014). Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer, Cham.

[4]Goyal, Y., Khot, T., Agrawal, A. et al. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. Int J Comput Vis 127, 398–414 (2019). https://doi.org/10.1007/s11263-018-1116-0

[5] Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., & Parikh, D. (2016). Yin and Yang: Balancing and Answering Binary Visual Questions. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5014-5022.