Visual Questioning Answer

Xin Wang

Now a day, with the computer technology and hardware source improvement, Artificial Intelligence gets wildly used in different areas, such as Internet, Economy, Medication, and Society. There are many different aspects in AI, like Computer Vision (CV), Nature Language Processing (NLP), and Knowledge Representation and Reasoning (KR). Also the combination research and project among the aspects increasing significantly these years. One welcomed project among them is named as Visual Questioning Answering (VQA), which gets a nature language and open-ended question and a question related image as input and returns a correctly nature language answers as output. How to achieve this goal is a multi-discipline hard task, and requires Computer Vision and Nature Language Processing technology helping the deal with the image information and question information at the same time. With the help of machine learning, the model will figure out the final relationship between questions and images to generate the answer. The application of VQA is also wildly used. VQA can help visual-impaired person to explore the world around them more conveniently. For information statistician, they can easily obtain the target information by entering a example question.

There are many papers and open-source correlative with VQA published and available.

VQA is a giant project. And there are many aspects that can be focused on. Starting from processing the data sets, in the paper “VQA: Visual Questioning Answering” by A Agrawal, J Lu, S Antol, M Mitchell, L Zitnick, D Batra, and D Parikh, there are two types of image data: real images and abstract scenes which are generated by cartoon graphs and are used to explore the high-level reasoning problem because of their simplicity and fewer noise properties. There are 123,287 training and validation images and 81,434 test images from the Microsoft Common Object in context (MS COCO) and 50K abstract images generated by the research team. Each images is questioned by three nature-language questions and answered by people to create the huge bunch of question, image and answer paired data sets.

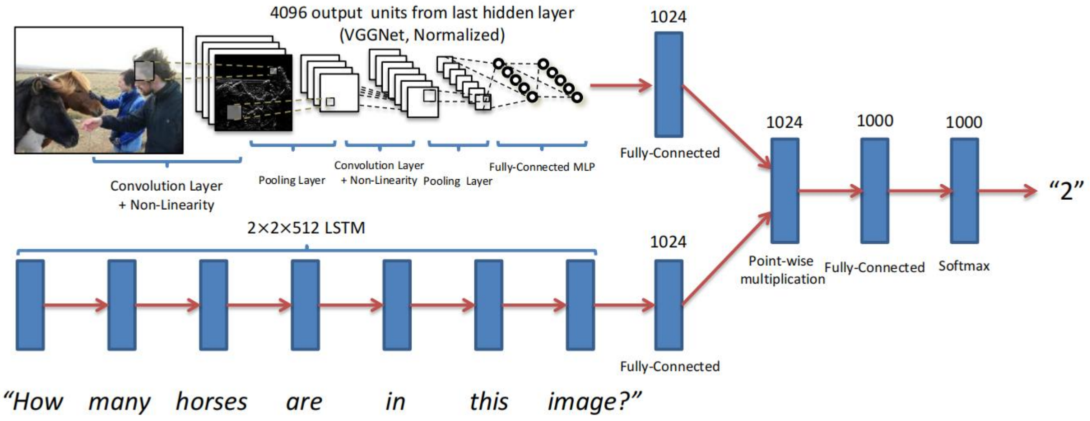
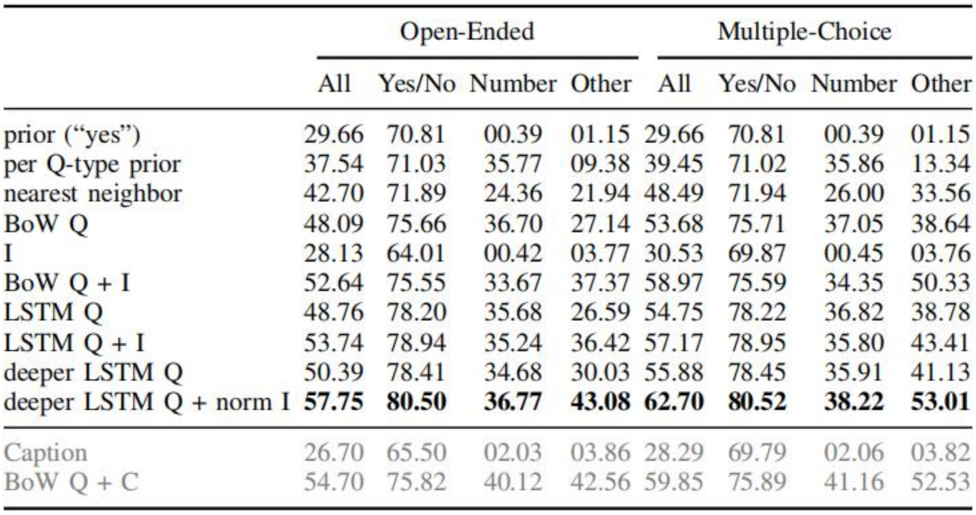


Figure1. the deeper LSTM Q and norm I model structure. [1]

The method in the paper VQA: Visual Questioning Answering, as the figure 1 shows, is to process the image and question singly at first. Then fusing two layers through point-wise multiplication to build a combination. And finally passing through a fully connected layer with a softmax layer to obtain the answer. There are two ways to process image: Use VGGNet (I) and normalize VGGNet (I_norm). There are three ways to process question-information: Bag-of-Words Question (BoW Q), LSTM Q, and deeper LSTM Q.

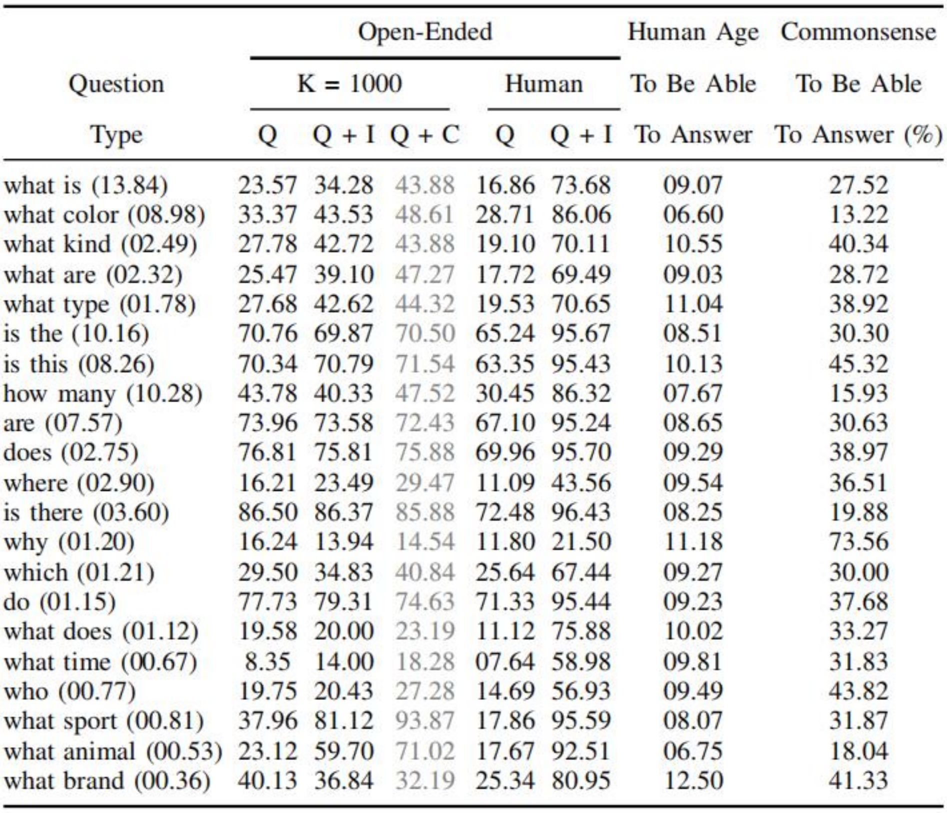
The research group has tried to choose each one of the methods from image channel and question channel to build the model and test all data sets. The evaluation of the model is the accuracy of the predict answer that model generated. And the accuracy is chose from the minimum of the average of the answer that is provided by human and 1.Their research results are as following.



	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

Figure2. Accuracy of the methods for open-ended and multiple-choice tasks on the VQA test-dev for real images.

The figure 2 shows the accuracy results of different models. The first three lines are baselines which are the result of answer by human. And the conclusion 1 is that the image-alone model which ignores the text information in questions provides really bad accuracy. The conclusion 2 is that the text-alone model which ignores the image information provides kind of good accuracy, which the researchers thought that may because of the model has learned some reasoning relationship between questions and answers. So the model can answer the question such as “What is the color of banana?” without seeing the image. And the conclusion 3 is that the best performance model combines the normal I and deeper LSTM Q. But it still can't catch up to the performance of human being.

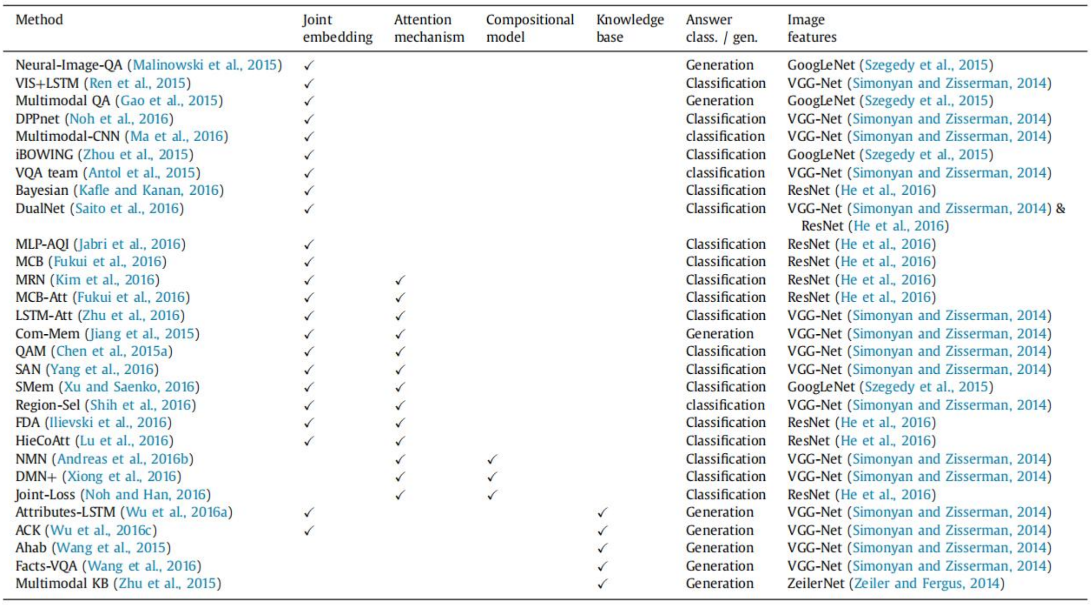


	Open-Ended					Human Age	Commonsense
Question	K = 1000			Human		To Be Able	To Be Able
Type	Q	Q + I	Q + C	Q	Q + I	To Answer	To Answer (%)
what is (13.84)	23.57	34.28	43.88	16.86	73.68	09.07	27.52
what color (08.98)	33.37	43.53	48.61	28.71	86.06	06.60	13.22
what kind (02.49)	27.78	42.72	43.88	19.10	70.11	10.55	40.34
what are (02.32)	25.47	39.10	47.27	17.72	69.49	09.03	28.72
what type (01.78)	27.68	42.62	44.32	19.53	70.65	11.04	38.92
is the (10.16)	70.76	69.87	70.50	65.24	95.67	08.51	30.30
is this (08.26)	70.34	70.79	71.54	63.35	95.43	10.13	45.32
how many (10.28)	43.78	40.33	47.52	30.45	86.32	07.67	15.93
are (07.57)	73.96	73.58	72.43	67.10	95.24	08.65	30.63
does (02.75)	76.81	75.81	75.88	69.96	95.70	09.29	38.97
where (02.90)	16.21	23.49	29.47	11.09	43.56	09.54	36.51
is there (03.60)	86.50	86.37	85.88	72.48	96.43	08.25	19.88
why (01.20)	16.24	13.94	14.54	11.80	21.50	11.18	73.56
which (01.21)	29.50	34.83	40.84	25.64	67.44	09.27	30.00
do (01.15)	77.73	79.31	74.63	71.33	95.44	09.23	37.68
what does (01.12)	19.58	20.00	23.19	11.12	75.88	10.02	33.27
what time (00.67)	8.35	14.00	18.28	07.64	58.98	09.81	31.83
who (00.77)	19.75	20.43	27.28	14.69	56.93	09.49	43.82
what sport (00.81)	37.96	81.12	93.87	17.86	95.59	08.07	31.87
what animal (00.53)	23.12	59.70	71.02	17.67	92.51	06.75	18.04
what brand (00.36)	40.13	36.84	32.19	25.34	80.95	12.50	41.33

Figure3. Open-ended test-dev results for different question types on real image.[1]

Furthermore the researchers add caption information which are image type labels generated by people to pursue the better performance of the model. As figure 3. some reason-level questions such as “What is” and “why” didn’t show any improvement. And for scene-level questions such as “What sport” and “What animal”, the result of model with adding caption information improve a lot.

There are many other papers have different focus about doing research on VQA. The paper “Ying and Yang: Balancing and Answering Binary Visual Questions” use [Object, relation, subobject] pair structure to build model and find the relationship between objects to answer the question.[2] The paper “Where To Look: Focus Region for Visual Question Answering" focus the algorithm to let the model to find the specific area that the question related.[3] The paper "Visual question answering: A survey of methods and datasets" analyzes different methods of VQA models shown in Figure 4. [4]



Method	Joint embedding	Attention mechanism	Compositional model	Knowledge base	Answer class. / gen.	Image features
Neural-Image-QA (Malinowski et al., 2015)	✓				Generation	GoogLeNet (Szegedy et al., 2015)
VIS+LSTM (Ren et al., 2015)	✓				Classification	VGG-Net (Simonyan and Zisserman, 2014)
Multimodal QA (Gao et al., 2015)	✓				Generation	GoogLeNet (Szegedy et al., 2015)
DPPnet (Noh et al., 2016)	✓				Classification	VGG-Net (Simonyan and Zisserman, 2014)
Multimodal-CNN (Ma et al., 2016)	✓				classification	VGG-Net (Simonyan and Zisserman, 2014)
iBOWING (Zhou et al., 2015)	✓				Classification	GoogLeNet (Szegedy et al., 2015)
VQA team (Antol et al., 2015)	✓				classification	VGG-Net (Simonyan and Zisserman, 2014)
Bayesian (Kafle and Kanan, 2016)	✓				Classification	ResNet (He et al., 2016)
DualNet (Saito et al., 2016)	✓				Classification	VGG-Net (Simonyan and Zisserman, 2014) & ResNet (He et al., 2016)
MLP-AQI (Jabri et al., 2016)	✓				Classification	ResNet (He et al., 2016)
MCB (Fukui et al., 2016)	✓				Classification	ResNet (He et al., 2016)
MRN (Kim et al., 2016)	✓	✓			Classification	ResNet (He et al., 2016)
MCB-Att (Fukui et al., 2016)	✓	✓			Classification	ResNet (He et al., 2016)
LSTM-Att (Zhu et al., 2016)	✓	✓			Classification	VGG-Net (Simonyan and Zisserman, 2014)
Com-Mem (Jiang et al., 2015)	✓	✓			Generation	VGG-Net (Simonyan and Zisserman, 2014)
QAM (Chen et al., 2015a)	✓	✓			Classification	VGG-Net (Simonyan and Zisserman, 2014)
SAN (Yang et al., 2016)	✓	✓			Classification	VGG-Net (Simonyan and Zisserman, 2014)
SMem (Xu and Saenko, 2016)	✓	✓			Classification	GoogLeNet (Szegedy et al., 2015)
Region-Sel (Shih et al., 2016)	✓	✓			classification	VGG-Net (Simonyan and Zisserman, 2014)
FDA (Ilievski et al., 2016)	✓	✓			Classification	ResNet (He et al., 2016)
HieCoAtt (Lu et al., 2016)	✓	✓			Classification	ResNet (He et al., 2016)
NMN (Andreas et al., 2016b)		✓	✓		Classification	VGG-Net (Simonyan and Zisserman, 2014)
DMN+ (Xiong et al., 2016)		✓	✓		Classification	VGG-Net (Simonyan and Zisserman, 2014)
Joint-Loss (Noh and Han, 2016)		✓	✓		Classification	ResNet (He et al., 2016)
Attributes-LSTM (Wu et al., 2016a)	✓			✓	Generation	VGG-Net (Simonyan and Zisserman, 2014)
ACK (Wu et al., 2016c)	✓			✓	Generation	VGG-Net (Simonyan and Zisserman, 2014)
Ahab (Wang et al., 2015)				✓	Generation	VGG-Net (Simonyan and Zisserman, 2014)
Facts-VQA (Wang et al., 2016)				✓	Generation	VGG-Net (Simonyan and Zisserman, 2014)
Multimodal KB (Zhu et al., 2015)				✓	Generation	ZeilerNet (Zeiler and Fergus, 2014)

Figure 4 Overview of existing approach. [4]

Reference:

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh (2015). VQA: Visual Question Answering. In International Conference on Computer Vision (ICCV).

[2] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, Devi Parikh (2016). Yin and Yang: Balancing and Answering Binary Visual Questions. In Conference on Computer Vision and Pattern Recognition (CVPR).

[3]Shih, K., 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 (CVPR).

[4]Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, Anton van den Hengel (2017). Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding, 163, 21-40.