A Survey On Visual Questioning Answering : Datasets, Approaches And Models

Sunny Katiyar, M. S. Wakode

Abstract: Visual Questioning Answering is a new way of interacting with artificial systems. They are aimed at making interaction with machines similar to interaction with humans. Lots of VQA models are proposed over time which are focused on datasets such as VQA, MS-COCO, Flicker30k etc. These VQA systems are limited to the type of dataset used to trained the system. There are different techniques for different types of images like natural images, artificial images, mathematical plots, etc. The performance of the models also depend on the type of question asked. In this paper we will discuss about different datasets available for VQA tasks and techniques used to build VQA systems on those datasets. We will also discuss the state of art VQA system in detail in the later section of this paper.

Index Terms: Artificial Intelligence, Machine Learning, Natural Language Processing, Computer Vision, Object detection, Context awareness, Prediction Methods, Attention Models

1. INTRODUCTION

In this decade, there have been a lots of advancement in the field of computer vision tasks such as Image classification, Object Detection etc. Visual Question Answering (VQA) is one of the complex computer vision task which combines techniques from other computer vision tasks to improve interaction with the machines. For example we can provide an image and ask text-based questions regarding that image and VQA system will provide answer in natural language. Such questions can contain other sub-problems like Scene Classification (What is the weather in the image?), Object recognition (What animal is in the image?), Object detection - Are there any cats in the image?), Attribute classification (What color is the dog?), Counting Problems (How many dogs are in the image?), etc. Current VQA models can answer better than human for some type of questions like counting question or object detection but are far behind when some reasoning is required. The human accuracy on DAQUAR dataset which consist mostly indoor images has human accuracy of about only 50%. And models with far better accuracy have been developed. The models available for natural abstract images, does not work with graphical plots, as the questions asked are quite different and need different approach to find answers. So different models are needed to develop efficient models to work on graphical images that can provide satisfactory results.

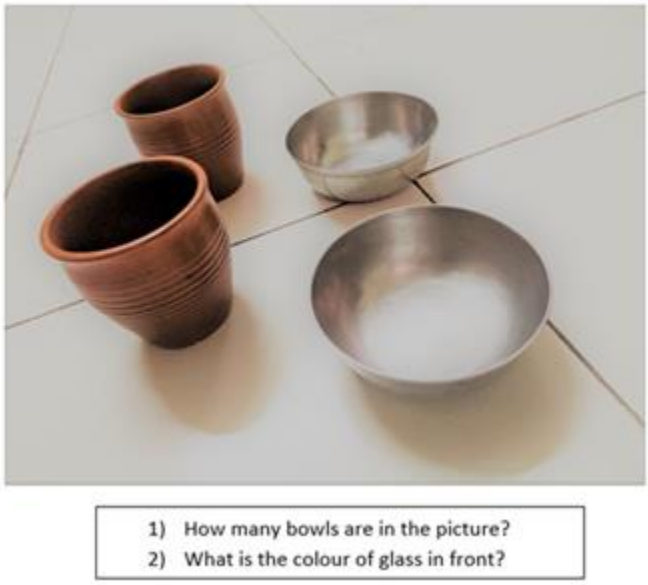

***Fig.** 1 VQA system example – image and question*

On the Figure-QA dataset the human accuracy is calculated around (91.21%) and it would be better to get anything near about it. The accuracy of relation network model[1] which is also the baseline model is 72.40% provided. The rest of the paper is presented as follows. In section II we describe the relevant literature. The common system architecture is given in Section III. Section IV discuss different datasets.Section V shows a common VQA System Analysis. And Section VI contains description about common evaluation measures. And Section VII talks about state of art pythia model. Last section concludes the paper and shed some light on potential future work.

2. REVIEW OF LITERATURE

A VQA system is studied under domain of computer vision. In last few years the popularity of the of VQA system has increased many folds. We have come a long way from digit recognition[11] to the state of art Pythia[12] VQA system by facebook.A lot of work has been done in the field. Various approaches are proposed like Multimodal fusion, , Compositional approaches,Question-Aware models, etc. A VQA task

- *Sunny Katiyar is currently pursuing masters degree program in computer engineering in P.I.C.T, Pune University, Maharashtra,India, PH-7007978375.*
  E-mail: sunnykatiyar50@gmail.com
- *Prof.M.S.Wakode is currently appointed as professor in Computer Engineering Department in P.I.C.T Pune.*

consists of multiple tasks such as object detection, scene detection, object classification, image cap- tioning [4]. In [10], they focused on how human recognize objects or scenes by and how human attention works. Later, many attention based models were proposed like bidirectional flow of attention in [15], bottom up top-down attention [5] model which later won 2017 VQA challenge. After image classification and object detection, image captioning became the main focus of experts. In [9] , it is described how visual attention can be used for the purpose of image captioning. VQA systems use models from both image processing and language processing community. These models are combined or concatenated using certain methods to make them provide answers according to image. In [8] proposed a multimodal pooling method to concatenate text and image models. Deep Convolutional Neural Network are preferred over other techniques for image captioning and VQA systems, as they contain various layers which can represent various details of an image. In [16] they proposed how these networks can be used for huge number of images. After image caption generation, interacting with those images was the next popular task which resulted into VQA systems. Using questions and answers along-with images while training gives the model capability for answering when asked questions. Attention mechanism was the base of most of the models [13] [3] for developing VQA systems. dataset play a major role when developing a system as complex as VQA systems. Initially VQA v1 dataset was used, but models exploited the dataset and bias their answer according to most used answer or unique answer for unique question for an image. So in [2] a new dataset was introduced VQA v2 which contains twice number of images and every question having at-least two different answers. Later other models like [6] [7] and [13] were also introduced which support passing relational facts along-with answers to the model while training. This helped making system more semantically capable and answer question containing why other than what, how, which etc. The VQA systems were not that great for answering graph- ical questions to images and plots which are formed using numeric data collected by sensors or from markets. Because question asked mostly are relational, so models require rea- soning and relating capability. The questions cane be one to one , one to many or many to many [4]. So Models require a different approach to train for such images and more over require a balanced dataset for training such models. In [1], They have introduced a dataset called Figure-QA dataset along-with a baseline models.

2.1 APPROACHES FOR SOLVING VQA TASK

1) COMPOSITIONAL VQA MODELS :

In compositional approach the questions are interpreted as composition of multiple sub- tasks. For example - what color shirt the boy is wearing? can be interpreted as composition of finding the boy and locating the shirt and then naming the color of the shirt. There are different frameworks available for processing the tasks in sub-steps. [22] and [24] describes two frameworks using neural networks to process tasks in sub-steps.

2) ATTENTION BASED MODELS:

Attention mechanism form the basis of lot of models out there. Attention models[?] are popular as they are based on the fact that certain regions in image and certain words in text are important than others. But for attention models to work, dataset must include the bounding-box annotations that contain information about the different objects in different regions in the image. There can be two ways of encoding local feature in an image. One is to divide image in uniform grid-size and then specify grid-wise features and the other is by defining bound- ing box annotation around different objects in the images. The VQA models which use annotation box as local feature encoding needs to pass image and the question along-with annotation box while training.In[25], a Focused Dynamic Attention model was introduces which suggest to use only those bounding box whose labels match with the words in the question. In this way a lot of redundant processing is saved with a little extra effort of matching the labels. Later the matched labels are classifed using the ResNet[26].

3) BAYESIAN AND QUESTION-AWARE MODELS:

This approach is not very good for using in systems that answer questions about the image. Because the model based on this approach does not even consider looking at the image and instead predict answer on the basis of bayesian model by finding probabilities of the words in answers in the dataset. This can give biased result in favour of the most or frequent occurring word in answer-set. Similarly there are question-aware models which pre-decides a set of answer words by looking at the question. For example - what is the color of the shirt? In this question without looking at the image, model concludes that the answer is one of the color name only. And thus, it applies the probabilistic model over the set of all the colors in the dataset.

4) BI-LINEAR POOLING METHODS:

Every VQA system includes combining extracted features from images and question sep- arately. But how to combine those features also affects the type of model is being developed. Simple concatenation or element-wise product are simplest methods. More complex and effective methods are suggested by other. [8] proposed a multi-linear bi-pooling method which sug- gest an outer product between the two extracted entities which combines them more effectively than other simple methods like concatenation or element-wise products. The outer product may be very complex therefore, the product is done in low dimension to reduce compelxity. The winner of 2016 VQA challange [27] used this multi-linear bi-pooling with soft- attention mechanism on COCO dataset.

3. DATASETS

3.1. DAQUAR

DAQUAR stands for DAtaset for QUestion Answering on Real-world images [18]. It was the first and the smallest dataset to be introduced in the field. It consist of only 6795 training and 5673 testing QA pairs based on images. Other small dataset with 37 object categories was also given and called as DAQUAR-37. It has only 3825 training and 297 testing QA pairs. Also to evaluate performance ground truth answers were also provided with the dataset. The images were mostly indoor images and were very difficult recognize even by humans and the accuracy was only 52%.

3.2 COCO-QA

COCO-QA[19] contains machine-generated questions- answer pairs using the caption of the images. For example if the

caption is - the dog is eating then question would be - what is the dog doing and answer can be - eating. COCO-QA has about 78,736 training and 38,948 testing QA pairs Most questions were about the object in the image (70%), with the other questions being about color (16%), about counting (7%) and about location (6%). The main flaw in the auto-generation of questions was the grammatical errors in lengthy questions and questions that contains phrases.

3.3. VQA (v1 v2)

VQA v1 [20] dataset is one of the biggest dataset which used images from COCO dataset and added some artificial images to make it more complete. There are three question per image and about ten answers per question. There are about 82000 training images,40000 validation images and about 81000 testing images and about 600 thousand QA-pairs, in which there are more than 200 thousand training and testing QA pairs and more than 100 thousand validation QA-pairs. The answer contains most plausible answer, correct answer, popular answers and random answers.Different independent group of workers were assigned tasks of generating questions and answers. VQAv2 [2] was introduced to remove any type of biasness in case there is only one answer to one question for a particular image. This was removed by adding other image with same question but differnt answer. Thus the size of the new dataset was doubled.

3.4 Figure-QA

Figure-QA [1] is the most recently introduced dataset in the family but, unlike others it contains synthetic images generated from numeric data. These images consist of pie charts, line plots, dot-line plots, histograms, etc. So the questions asked to such images are different. The questions are relational and can be any of one to one, one to many or many to many type. The dataset selected for training is Figure-QA dataset which consist about 100,000 images and 1,300,000 question answer pairs.It consist of five categories of mathematical graphs and plots vertical and horizontal bar graphs, pie charts, line graphs and dot graphs. Nearly 100 colors are selected which are at sufficient distance from background color. Annotations are also included with the images and questions which can be used to apply attention mechanism.

3.5 Visual GNOME

Visual GNOME is the largest dataset ever for VQA.It has about consists of 108,249 images that occur in YFCC100M [23] and COCO [mre] images. It has about 1.7 million qa- pairs at an average of about 17 questionsper image. Not a lot of models have been developed on Visual GNOME as it has large no. of images and hence will take lot of time training and also it is newer than other VQA datasets discussed. It contains question of types - what, where, when, who, how and why. people were asked to ask any type of question freely while preparing the dataset. There are also region based questions where region is defined by bounding-box annotation.

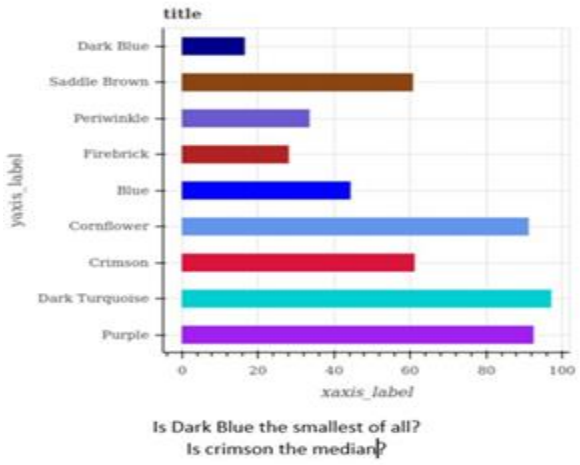

***Fig. 2**. Example from FigureQA Dataset. [1]*

4. COMMON SYSTEM ARCHITECTURE

The system architecture consist of feature extraction from both image and the question. The image features are extracted using a 5 Convolutional layer Model and question features are extracted using the 2 layer LSTM. The architecture of these models can be chosen differently. We have chosen these models because they are the most used models in their respective community and are simple to implement providing satisfactory results. After the features are extracted, the image embeddings are mapped with the annotations provided with the dataset. These annotations mark the different regions in the image. After mapping, above embedding is concatenated with the question embedding. This allocates different weights to different re- gions in the image according to the context of a region in the question. Higher the weight, higher is the context of the region to the question.

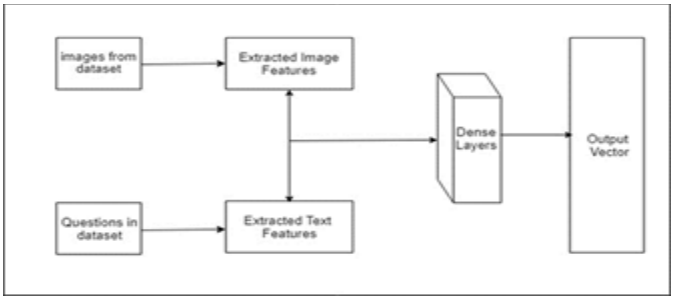

***Fig. 3**. Most Common VQA System Architecture*

The concatenated embeddings are passed through fully connected layers which classifies the answers with different probabilities into output vectors. The answer with the highest probability is the final answer.

5. COMPARATIVE ANALYSIS

There are lots of model developed for VQA system on different datsets like COCO-QA, VQA , Visual GNOME, etc. VQA challenge is one of the biggest competition in which different developers submit their models to test accuracy on VQA test set. Accuracy on different types of question categories like Yes/No, Numbers, etc are measured before announcing a winner.

Year	Model	Team / Developer	Accuracy on test split
2018 (winner)	Pythia	Facebook AI Research A-STAR Team	72.27%
2017 (winner)	Bottom-Up Top-Down Attention Mechanism	Adelaide-Teney ACRV MSR	70.34%
2016 (winner)	MCB to efficiently and expressively combine multimodal features	UC Berkeley & Sony	66.9%

***Fig. 4**. Best VQA Models year-wise*

Pythia from Facebook AI Research is the winning model in 2018 with overall accuracy of 72.27%. It used Bottom-up Top-Down Attention Mechism model which is the winner of 2017 challenge as the base and improved accuracy from previous accuracy of 70.34%. In 2016 the model wusing Multi-linear bi-linear pooling as the concatenation method won the challenge. So the best accuracy available for VQA v2 is 72.27%.For VQA dataset, the Relation Network (RN) model which is also the baseline model for dataset gives accuracy of about 72%.

6. EVALUATION METRICS FOR VQA

6.1 Simple Accuracy

Simple Accuracy is measured by comparing predicted answer with the provided ground truth answer. But it consider answer either only right or wrong. So it is good for small number of unique answers like multiple choice questions. But simple accuracy cannot handle incomplete or less incorrect or similar answers. For example - cat in place of cats is considered as wrong as dog in place of cats. Similarly half correct answer is also considered wrong.

6.3 Modified WUPS

Wu-Palmer similarity is better option than simple accuracy as it considers lexical similar words and doesn't requires exact match. But to measure prediction in VQA systems modified Wu-Palmer is used with a applied threshold and a scaling factor. The answer with WUPS score less than a threshold is scaled down by a scaling factor. This was used as standard evaluating measure in DAQUAR and COCO-QA. But problems appears in modified WUPS when two words with lexical similarity but very different meaning appears, for example name of colors. Also WUPS cannot be used for answers with more than one word.

6.3 Consensus Metrics

Alternative of considering semantics of answers was to have multiple independent ground truth answers per question. This helps to consider similar answer for the question as correct. This method was used in VQA datset. Every question has nearly ten ground truth answers. And if an answer is agreed upon by more than 3 annotators then it is given full score. But problem with this is there can be more than one answer for a question. And also creating such consensus for a dataset requires lots of efforts and workforce.

6.4 Manual Evaluation

Manual Evaluation involves using humans to decide whether the answer is correct or not. This removes problem incorpo- rated in previously discussed evaluation metrics, as humans can understand the semantics of the answer well and thus reduce the error of considering wrong answer to the minimum. But this induces new problem of personal opinion of humans while considering the answer. The other problem of this procedure is using humans is a resource extensive problem. Also it becomes difficult to improve system by a changing algorithm if the output is given by humans

7. STATE OF THE ART MODEL

A. Pythia from Facebook AI Research A-STAR Team[17]

Pythia [17] is a model given by Facebook AI Research Team and is the winner of VQA 2018 challenge. It gave overall accuracy of about 72.27% on VQA v2 dataset. Pythia is based on 2017 VQA-winner model based on Bottom Up Top-Down Attention Mechanism, which gave accuracy of about 70%.Pythia team used 300D GloVe [11] vectors to initialize the word embeddings and pass it to GRU network and a question attention module to extract text features. Features are concatenated using element-wise product and the hidden size used is 5000. The results , after above modifications, improved from 65.32% to 66.91% on VQA v2.0 test-dev. The dataset is increased by adding images from Visual Gnome and Visual dialog. The answer to a question are repeated to match VQA's 10 answer per question. Some images are mirrored and words like left are changed to right and vice-versa to introduce some variation in dataset.

8. CONCLUSION

Developing VQA Systems for various domains or areas can really be helpful and speed the tasks in those areas. Developing a VQA system to process synthetic images that represent huge amount of numeric data can help getting the conclusions about the raw data instantly. These synthetic images dataset are still getting attention and the proposed model is result of that. We are applying attention mechanism to the model which will help to improve the performance while training and later increase the accuracy of the model.

REFERENCES

[1] Kahou, Samira Ebrahimi and Michalski, Vincent and Atkinson, Adam and Kadar, Akos and Trischler, Adam and Bengio, Yoshua. "Fig- ureqa: An annotated figure dataset for visual reasoning". arXiv preprint arXiv:1710.07300, 2017.

[2] Goyal, Yash and Khot, Tejas and Summers-Stay, Douglas and Batra, Dhruv and Parikh, Devi. "Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),6325- 6334, 2017.

[3] Quanzeng and Jin, Hailin and Wang, Zhaowen and Fang, Chen and Luo, Jiebo."Image captioning with semantic attention". Proceedings of the IEEE conference on computer vision and pattern recognition.4651– 4659.2016

[4] Kafle, Kushal and Kanan, Christopher. "Visual question answering: Datasets, algorithms, and future challenges". Computer Vision and Image Understanding. vol. 163, pages 3–20,2017, Elsevier.

[5] Anderson, Peter and He, Xiaodong and Buehler, Chris and Teney, Damien and Johnson, Mark and Gould, Stephen and Zhang, Lei. "Bottom-up and top-down attention for image captioning and visual question answering". CVPR,volume 3,page 6, 2018.

[6] Lu, Pan and Ji, Lei and Zhang, Wei and Duan, Nan and Zhou, Ming and Wang, Jianyong. "R-VQA: Learning

Visual Relation Facts with Semantic Attention for Visual Question Answering". arXiv preprint arXiv:1805.09701, 2018.
[7] Wu, Qi and Teney, Damien and Wang, Peng and Shen, Chunhua and Dick, Anthony and van den Hengel, Anton."Visual question answering: A survey of methods and datasets". Computer Vision and Image Under- standing, volume163, pages 21-40, 2017, Elsevier.
[8] Fukui, Akira and Park, Dong Huk and Yang, Daylen and Rohrbach, Anna and Darrell, Trevor and Rohrbach, Marcus. "Multimodal compact bilinear pooling for visual question answering and visual grounding". arXiv preprint arXiv:1606.01847. 2016.
[9] Xu, Kelvin and Ba, Jimmy and Kiros, Ryan and Cho, Kyunghyun and Courville, Aaron and Salakhudinov, Ruslan and Zemel, Rich and Bengio, Yoshua. "Show, attend and tell: Neural image caption genera- tion with visual attention". Book International conference on machine learning, pages 2048–2057, 2015.
[10] Das, Abhishek and Agrawal, Harsh and Zitnick, Larry and Parikh, Devi and Batra, Dhruv. "Human attention in visual question answering: Do humans and deep networks look at the same regions". Computer Vision and Image Understanding, vol. 163, pages 90–100, 2017, Elsevier.
[11] Bottou, Le'on and Cortes, Corinna and Denker, John S and Drucker, Harris and Guyon, Isabelle and Jackel, Lawrence D and LeCun, Yann and Muller."Comparison of classifier methods: a case study in hand- written digit recognition". Pattern Recognition, 1994. Vol. 2- Conference B: Computer Vision Image Processing. Proceedings of the 12th IAPR International. Conference on vol. 2, pages 77–82, 1994.
[12] Jiang, Yu and Natarajan, Vivek and Chen, Xinlei and Rohrbach, Marcus and Batra, Dhruv and Parikh, Devi."Pythia v0. 1: the winning entry to the vqa challenge 2018, arXiv preprint arXiv:1807.09956, 2018.
[13] Xiong, Caiming and Merity, Stephen and Socher, Richard. "Dynamic memory networks for visual and textual question answering". Interna- tional conference on machine learning. pages 2397–2406, 2016.
[14]
[15] Wu, Qi and Shen, Chunhua and Wang, Peng and Dick, Anthony and van den Hengel, Anton. "Image captioning and visual question answering based on attributes and external knowledge". IEEE transactions on pattern analysis and machine intelligence, vol. 40,pages 1367-1381, 2018.
[16] Seo, Minjoon and Kembhavi, Aniruddha and Farhadi, Ali and Hajishirzi, Hannaneh. "Bidirectional attention flow for machine comprehension". arXiv preprint arXiv:1611.01603, 2016.
[17] Simonyan, Karen and Zisserman, Andrew. "Very deep convolu- tional networks for large-scale image recognition". arXiv preprint arXiv:1409.1556. 2014
[18] Jiang, Yu and Natarajan, Vivek and Chen, Xinlei and Rohrbach, Marcus and Batra, Dhruv and Parikh, Devi."Pythia v0. 1: the winning entry to the vqa challenge 2018",arXiv preprint arXiv:1807.09956,2018.
[19] M. Malinowski and M. Fritz."A multi-world approach to question answering about realworld scenes based on uncertain input," in Advances in Neural Information Processing Systems (NIPS), 2014.
[20] M. Ren, R. Kiros, and R. Zemel, Exploring models and data for image question answering," in Advances in Neural Information Processing Systems (NIPS), 2015.
[21] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and
[22] D. Parikh. "VQA:Visual question answering," in The IEEE International Conference on Computer Vision(ICCV), 2015.
[23] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen,
[24] Y. Kalantidis,L.-J. Li, D. A. Shamma, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations," International Journal of Computer Vision, vol. 123, 2017.
[25] 44 J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, Deep composi- tional question answering with neural module networks," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[26] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. "Yfcc100m: The new data in multimedia research". Communications of the ACM, vol. 59, no. 2, pp. 2016.
[27] [55] H. Noh and B. Han. "Training recurrent answering units with joint loss minimization for VQA". arXiv preprint arXiv:1606.03647, 2016.
[28] I. Ilievski, S. Yan, and J. Feng. "A focused dynamic attention model for visual question answering." arXiv preprint arXiv:1604.01485, 2016.
[29] K. He, X. Zhang, S. Ren, and J. Sun."Deep residual learning for image recognition." in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[30] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola."Stacked attention networks for image question answering." in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.