International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 06 | June 2018 www.irjet.net p-ISSN: 2395-0072

Image Question Answering: A Review

Abhishek Chidambar[1], Sagar Karambelkar[2], Kaushik Prabhu[3], Vishal Darade[4], S. S. Sonawani[5]

[1, 2, 3, 4] *Student, Department of Computer Engineering, Maharashtra Institute of Technology, Pune*
[5] *Professor, Department of Computer Engineering, Maharashtra Institute of Technology, Pune*

Abstract - *Part of the renewed excitement about research in image and video captioning, which is a combination of Computer Vision, Natural Language Processing, and Knowledge Representation and Reasoning stems from a belief that it is a drastic step towards solving AI. The problems at the intersection of Computer Vision and the processing of natural language are of significant importance to challenging research questions and for the rich set of applications they enable. Given an input image and an open-ended question in natural language, it needs an understanding of visual elements contained in the image and common-sense knowledge to provide a useful response. In this review, we examine multiple approaches to this problem - where images and questions are mapped to a common feature space - along with the datasets available to train and evaluate these systems. We also discuss future scope and promising directions in which this field could be headed towards.*

Key Words: **Artificial Intelligence, Neural Networks, Deep Learning, Natural Language Processing, Image Processing**

1. INTRODUCTION

Computer vision deals with how computers can process, analyze and understand images. Whereas, natural language processing deals with how computers can interact with humans by understanding and synthesizing various forms of speech. Image Question Answering is a fairly recent topic of interest in the AI community and has therefore amassed significant amount of interest due to its multi-modal nature to connect visual and textual modalities. Recent research in the machine learning/deep learning domain has witnessed a lot of progress in the areas of image processing and natural language processing that involve complex tasks like image classification [1], object detection [2], semantic analysis [3], [4], [5], and much more.

In most forms of image question answering tasks, the system is presented with an image and a textual question related to that image. The system must attempt to determine the answers to that image. The type of answers given by the system usually include a few words, a short phrase or, in some sophisticated systems, a longer sentence. The questions framed could be in the form of multiple choice, binary (yes/no) and open-ended settings. Unlike conventional object detection systems, solutions to the Visual Question Answering (VQA) problem are diverse and arbitrary in the sense that they not only answer questions to detect an object in the image ("What is there in the image?"), but also such that they answer questions to recognize the object ("Is the detected object a cat?"), classify it based on its attributes ("What is the colour of the cat?") & the scenario ("Is it day or night?") and count multiple occurrences of the object in the image ("How many cats are there in the image?"). Beyond this there are many common-sense reasoning questions that a model might find very challenging to answer, nevertheless, should be ideally capable of. For example - for an image with a blanket of snow over the terrain - "Is it cold outside?". The questions proposed to the system are only known to the system at run-time.

A Turing Test is a test that can tell computers and humans apart. Similarly, a Visual Turing Test (VTT) is a test that can determine if a computer is artificially intelligent enough to 'understand' images as perceived by humans. Passing a test like this would require the computer to perform complex image processing and visual understanding tasks.

2. DATASETS

2.1 Available Datasets

There have been quite a few datasets released in the recent past that focus on solving the VQA task. Majority of these datasets contain several images and question-answer pairs for each image. These datasets help Image Question Answering systems to be trained on and evaluated

2.1.1 DAQUAR

The DAtaset for QUestion Answering on Realworld images (DAQUAR) [6] is one of the first datasets released specifically for the VQA task. This dataset is built upon the NYU-Depth V2 dataset [7]. It contains scenes that are exclusively taken in indoor environments. Question-answer pairs for the DAQUAR dataset were collected from human annotators for the NYU-Depth V2 dataset. DAQUAR is a relatively small dataset that consists of 6795 training and 5673 testing QA pairs based on images from the NYU-DepthV2 Dataset. There are two evaluation metrics proposed for this dataset.

2.1.2 COCO-QA

In COCO-QA, question-answer pairs are produced for images using a Natural Language Processing (NLP) algorithm that derives them from the COCO image captions [19]. For example, using the image caption Two children are playing with a ball, it is possible to create the question "What are the

children playing?" or "How many children are there?" with 'ball' and 'two' as the answer respectively. COCO-QA contains 78,736 training and 38,948 testing QA pairs. The algorithm used for generating questions and answers from the image captions, however, does not generate the questions properly. It has difficulty generating the pairs with changes in the grammar of the sentence.

2.1.3 VQA

VQA is a new dataset containing open-ended questions about images [5]. These questions require an understanding of vision, language and common-sense knowledge to answer. Questions and answers in this dataset are generated from crowd-sourced workers. The crowd-sourcing of this dataset has brought about a comparatively large number of questions – a total of 614,163 – including the training, validation and testing counterparts. About 10 answers are obtained for each question from 10 independent annotators respectively. The most frequent answers by these annotators for a particular question is concluded to be the 'correct' answer. In addition to that, there are 10 ground truth answers and 3 plausible answers for every question. The answers typically consist of a word or a short phrase. The VQA dataset consists of a total of 265,016 images. This includes images from the COCO dataset and other abstract (cartoon) scenes. The subsection of the VQA dataset that contains these abstract scenes is commonly known as SYNTH-VQA.

2.1.4 FM-IQA

FM-IQA stands for Freestyle Multilingual Image Question Answering. The images in this dataset is based on the MS-COCO dataset and is created to train and assess the mQA model discussed in [20] of its performance. Answers can be words, phrases or full sentences. The dataset contains 158,392 images and 316,193 questions.

The difference between the FM-IQA and the COCO-QA datasets is that in the FM-IQA dataset, the human annotators could input their own questions as long as they were related to the contents of the image. The question/answer pairs were originally written in Chinese and then translated to English. The evaluation of this dataset is done manually by humans via a Visual Turing Test. The manual evaluation of the FM-IQA is a contributing factor as to why the dataset hasn't gained a lot of popularity.

Table 1: Comparison of Existing Datasets

Comparison of Existing Datasets				
	DAQUAR	COCO-QA	VQA	FM-IQA
No. of Images	1,449	123,287	204,721	120,360
Collection Method	Human	Automatic	Human	Human
QA Pairs	12,468	117684	614163	250569
Avg. Answer Length	1.2 words	1.0 words	1.1 words	N/A

3. METHODS

3.1 Algorithms

Several solutions for the VQA challenge have been proposed in the recent past. Majority of the existing methods carry out the process of first extracting features from the image, understanding the meaning of the question and then using an algorithm that combines the two to then generate an appropriate answer (or list of answers).

Two simple standards for the VQA problem is a linearly modeled or a Multi-Layer Perceptron (MLP) classifier [5]. The features of the image are extracted from the last hidden layer of the CNN and then concatenated with the word embeddings of the question. This approach performs almost as well as many complex systems involving co-attention modules or spatial attention. The CNNs in these approaches are usually pre-trained on the ImageNet or COCO datasets. Some of these include the popular VGGNet [18] and the Residual Network CNNs [1].

As for the question feature extraction, the bag-of-words (BOW) approach has been widely used. Recurrent Neural Networks (RNNs) composed of Long Short Term Memory (LSTMs) units or Gated Recurrent Units (GRUs), used in conjunction with CNNs has been gaining a lot of popularity and has proved to significantly outperform the BOW approach.

Most VQA systems use a softmax classifier in order to generate answers. The answers to be predicted are predetermined and treated as separate classes.

3.2 Baseline Models

Baseline models give us a starting point on which we can evaluate our own approaches. A baseline approach for VQA can be as simple as guessing the most frequent answers in the dataset. The baseline used in [5] is a Multi-Layer Perceptron (MLP) which uses a single vector as a combination of image and question features and then fed as input into the MLP. Similar procedures have been observed in [9] and [10] in which features are combined using

pointwise multiplication or stacking the vectors on top of one another.

In [10], the image features from GoogLeNet are concatenated with a simple BOW using One Hot Encoding techniques. These features are then forwarded into a classifier and the answer is predicted using softmax. It outperformed the then baseline model which used LSTMs although this was a weaker model in theory.

In [11], every word in the input question was successively encoded with its respective CNN feature appended to it and then fed into the LSTM.

The approach used in [13] makes use of three custom CNNs – one to encode the features of the image, one for the question representation and the third to learn the joint representations to be able to predict an answer from candidate answers. This model outperformed the then state-of-the-art on the DAQUAR and COCO-QA datasets.

3.3 Attention Based Models

Attention based deep learning architectures have been used for the VQA task. These models focus on the nature of the questions since every question inquires about specifics of different regions in the image. The answer generation module in these models have question-guided attention, which means, the attention based CNNs used in these approaches determines respective regions in the image by finding the corresponding visual features in the feature maps with a "configurable convolution" operation. In the VQA problem, the attention should be focused on the parts of the image which are likely to give a correct answer. For example, for the question "What is the woman holding?", the focus should be the woman's hand/hands with which she's holding the respective object. Models of this architecture have been applied to several image processing tasks including object detection, object recognition, image captioning, etc, and has seen great success in those fields [14], [15]. Frameworks that use RNNs to predict attention regions based on previous attention region's location and features have been developed for object detection and recognition tasks. RNNs have also been used as decoders to extract a set of probable regions in an image and learn the weights of their attention.
[16] Proposed the Hierarchical Co-Attention model, in which both, the question and the image are reasoned with together at the same time with two separate information streams. The image representation in this model guides the attention of the question and vice-versa. Over and above the visual attention, this model 'hierarchically' encodes the questions at the word, phrase and question levels using the one-hot encoding technique, a tri-gram window and an LSTM at the end respectively.

Another interesting model has been discussed in [17] where jointly learned neural modules are used to first parse the question, co-ordinate with other modules based on those extracted features and in turn bring about attention, classification and other computational units to accordingly answer those questions with respect to the image.

3.4 Comparison of Existing Models

The following methods were assessed of their performance on the VQA dataset.

Table 2: Comparison of Existing Models

Comparison of Existing Models		
Method	Network Architecture	Accuracy %
[5] LSTM Q+I	VGGNet	54.1
[10] BOWIMG	GoogLeNet	55.9
[17] NMN	VGGNet	58.7
[9] HYBRID	ResNet	60.1
[13] Proposed CNN	Custom	58.4
[16] HieCoAtt	VGGNet	60.5
[16] HieCoAtt	ResNet	62.1

4. QUESTIONS TO BE ASKED

4.1 Are Binary Questions Enough?

Answering questions of the binary (yes/no) form has amassed a lot of interest in the VQA community. 'Yes' and 'No' are the most common responses in the VQA dataset generated by the human annotators. Although binary questions are easy to evaluate and can theoretically comprise of a wide range of tasks, there have been some opinions that argue the questions and answers generated by the human annotators are not complex or creative enough. Questions of this type can be especially difficult for algorithms to evaluate when it comes to the abstract scenes in the VQA dataset. Some datasets consist of questions that are overly biased towards 'Yes' responses. The COCO-QA is one example that contains significantly larger number of 'Yes' responses than 'No'. A VQA model that is capable of answering only binary questions, limits its own purpose of being of aid to the visually impaired or in any other general purpose application of the system.

4.2 How is Evaluation Impaired?

Dataset bias is an important talking point when it comes to solving the VQA challenge and evaluating the algorithms used for it. Questions relating to the presence of objects in an

image are far more common in most available datasets. Questions of this sort are tackled with ease by CNNs. Questions relating to the specifics of a scenario in an image, which are not objectively present in the image are a lot more challenging to answer. These questions occur rather very rarely. This includes questions that begin with 'Why' or 'How'. 'Yes' and 'No' responses are some of the most common responses (comprising of about 38% of all questions in the VQA dataset) with 'Yes' being a more biased response than the other (59%).

There are several limitations in majority of the datasets when it comes to the diversity in the types of the images and the annotated questions with their responses. For instance, images with scenes involving different cultures of the world, featuring people from different walks of life could propose a wider set of probable questions that could be annotated for those scenarios.

The impact that dataset bias has on algorithms solving the VQA task is substantial. For the VQA dataset, most models that improve accuracy on answering binary questions beginning with 'Is' and 'Are' by 15%, increase the general accuracy of the system by 5%. There is no symmetric effect observed when it comes to complex questions that begin with 'Why', 'How' and 'Where', which improves the general accuracy of the system by just 0.6%. The reason for this points towards the lack of these types of questions more so than the inability of these systems to be able to answer the questions correctly. A change in the approach to benchmark these algorithms by segregating the different types of questions, and then evaluating them in isolation, before taking the mean accuracy for every segregated category, could also be done. This is a viable option instead of calling for certain standards to be maintained in the quality of images and question-answer pairs available in future datasets.

4.3 Open Ended v/s Multiple Choice

It has been suggested that instead of open-ended answers, the system must select from multiple possible answers to make evaluation easier. In a modified version of the VQA dataset, the system could choose from a number of possible answers. A major drawback with this approach, is that the system will learn to pick the correct answer, and will focus on the list of answers instead of the image and the question. One of the objectives of the VQA challenge is to develop systems that can combine image and question features, and incorporating answers into the task is counterproductive to this end. Systems which used answers as features ended up giving very similar results with approaches that used attention models. For a true Visual Turing Test, the system must be able to generated open-ended answers.

5. WHAT WORKS BEST

Concluding what's the best method from the wide variety of available methods is not a straightforward task. The factors which influence the performance include the dataset, the architecture of the CNN, the complexity of the language model used, and whether an attention model was used. In general, ResNet outperforms the VGGNet and GoogLeNet models by 2%.Replacing VGG-19 by Resnet-152 increases performance by 2.3%. This is because the ResNet architecture does not degrade as rapidly with the addition of new layers, like the other architectures do.
Attention models also have a positive effect on the performance in comparison to models without attention, but it does not seem to be very substantial. Evaluation of these models for uncommon question types has a minor influence on the general performance score, which therefore makes it challenging to realize its advantages.

6. CONCLUSION

The VQA problem plays an important part in designing a relatively stringent Visual Turing Test. It requires systems to combine various facets of artificial intelligence like object detection, question understanding and determining the relationship between the objects and the question. Successful Image Question Answering on the whole has been widely known in the community and otherwise to be a momentous breakthrough among artificially intelligent systems. A system that can build a relationship between visual and textual modalities and give matching responses on arbitrary questions, could very strongly mark such a breakthrough.
In this paper, we have examined the various datasets and approaches used to tackle the VQA problem. We also examined some of the challenges VQA systems might face in regard to multiword answers and the types of questions asked. Future scope relies on creation of datasets that rectify bias in existing datasets, introduction of better evaluation metrics for the algorithms carrying out the VQA task to determine whether the algorithm is performing well on the VQA task in general and not only on that dataset. In addition to that, the inclusion of more varied, diverse scenes in the catalogue of images and more realistic, complex questions with a longer, descriptive set of answers would be highly valuable.

REFERENCES

[1] 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

[2] Joseph Redmon, Ali Farhadi "YOLO9000:Better, Faster, Stronger", arXiv:1612.08242v1, 25 Dec 2016

[3] Manning, Christopher D., Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J., and McClosky, David. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in Advances in Neural Information Processing Systems (NIPS), 2015.

[5] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "VQA: Visual Question Answering", in Proceedings of the IEEE International Conference on Computer Vision, pp. 24252433, 2015.

[6] M. Malinowski, Mario Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input ", NIPS, 2014.

[7] N. Silberman, D. Hoiem, et al, "Indoor segmentation and support inference from rgbd images," in European Conference on Computer Vision (ECCV), 2012.

[8] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, Ram Nevatia, "ABC-CNN: An attention based convolutional neural network for visual question answering", arXiv preprint arXiv:1511.05960, April 2016.

[9] K. Kafle and C. Kanan, "Answer-type prediction for visual question answering," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[10] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering," arXiv preprint arXiv:1512.02167, 2015

[11] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in The IEEE International Conference on Computer Vision (ICCV), 2015

[12] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in Advances in Neural Information Processing Systems (NIPS), 2015.

[13] Lin Ma, Zhengdong Lu, Hang Li, "Learning to answer questions from image using convolutional neural network" arXiv preprint arXiv:1506.00333

[14] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in International Conference on Machine Learning (ICML), 2015

[15] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," in International Conference on Learning Representations (ICLR), 2015

[16] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in Advances in Neural Information Processing Systems (NIPS), 2016

[17] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Deep compositional question answering with neural module networks," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations (ICLR), 2015

[19] M. Ren, R. Kiros, R. Zimel, "Image Question Answering: A Visual Semantic Embedding Model and a New Dataset", arXiv preprint arXiv:1505.02074v1 [cs.LG] 8 May 2015

[20] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? Dataset and methods for multilingual image question answering," in Advances in Neural Information Processing Systems (NIPS), 2015