International Journal of Future Generation Communication and Networking
Vol. 13, No. 3, (2020), pp. 3202–3208

A Review of Different Approaches for Visual Question-Answering Systems

Prof. K. P. Moholkar[1], Ajay Pisharody[2], Noorul Hasan Sayyed[3], Rakesh Samanta[4], Aadarsh Valsange[5]

JSPM's Rajarshi Shahu College of Engineering, Pune, India
[1]kavita.moholkar@gmail.com, [2]ajaypisharody715@gmail.com,
[3]mohdnoorsayyed@gmail.com, [4]samantarakesh83@gmail.com,
5aadarshvalsange99@gmail.com

Abstract

The ability of a computer system to be able to understand surroundings and elements and to think like a human being to process the information has always been the major point of focus in the field of Computer Science. One of the ways to achieve this artificial intelligence is Visual Question Answering. Visual Question Answering (VQA) is an AI-based system where for a given pair of question and image, the system attempts to give a relevant answer. This is usually a combination of image processing and understanding the question to match both to deliver the output. VQA can be efficiently used for scenarios where identification and classification of the images are needed. This can be done in different forms either as a run-time system constantly identifying its surroundings or a static application for medical sciences to find out important features like presence of cancerous cells, irregularities of an organ. VQA can be also used for surveillance systems, crime scene investigation. In this attempt, we try to build a model based on knowledge or additional information, with fine-tuned neural networks. In this paper, we have compared the results on popular image datasets.

Keywords: *VQA, CNN, RNN, AI, LSTM, Neural Networks, Image Processing*

1. Introduction

Visual Question Answering (VQA) is a system which predicts an answer for a given image and a question associated to it. This has been possible due to the advancements in the field of Artificial Intelligence notably in the Computer vision and Natural Language processing. Visual Question Answering has been a notable interest between researchers, universities and organizations as VQA provides vast applications due to its generic nature. With the help of specialized data and some modifications to the VQA models, it can be used for that particular situation. VQA can be efficiently used for scenarios where identification and classification of the images are needed. This can be done in different forms either as a run-time system constantly identifying its surroundings or a static application for healthcare and medical to find out important features like presence of cancerous cells, irregularities of an organ. VQA can be also used for surveillance systems, crime scene investigation. There has been a great progress in the development of the AI field throughout the decade, but applications like visual understandings, VQAs are still long away being perfect. Different models have been proposed and implemented such as an Attention-based model [4], Multi-modal approach [11], and Knowledge-based systems [7].

2. Literature Survey

Some of the proposed systems [7] [9] use pre-trained models like VGGNet16 and CIFAR10. These models lack the ability to be used in specialized situations and are more of a general approach towards VQA. Knowledge-based or Models with supporting captions and annotations prove to be better among others, but still lack the accuracy which is needed in high risk applications of medicinal science and other fields. There are some implementations which try to integrate multiple approaches like Attention with a knowledge base, Multi-modal approach and such. Such system seems to be a better alternative but does not provide adequate results in comparison to others. In many scenarios, the use of proper training datasets and testing dataset are less. But with the help of transfer learning and incremental learning approaches the models can be made with better accuracy and overall better answers.

Yuetan Lin et al (2016) in the paper titled "**Simple and Effective Visual Question Answering in a Single Modality"** proposes a Visual QA system based on a simple baseline format which uses textual modality to solve the main task. This method proves to be comparatively better as it inculcates image‟ s description as valuable input and due to this performs well than models based on attention.

Qi Wu et al (2017) in the paper titled "**Image Captioning and Visual Question Answering Based on Attributes and External Knowledge**" introduces a trainable Neural Network which is based on the image attributes using specific Convolutional and Recurrent Networks, and can be used for the problem of generating answers for wide range of questions. For DAQUAR dataset, the „Attention Captioning Selected knowledgebase LSTM‟ model proves to be the best among other models. This model outperforms the non-selected knowledgebase variant due to its efficient filtering of knowledge according to context. For Toronto dataset COCO-QA, the same model scores the perfect result. It exceeds the previously implemented methods around 10%, and outperforms the base model by nearly 20%. For VQA dataset, the proposed model produces the most suited results, by exceeding the baseline VggNet-LSTM by more than 10%. Future systems can be improved by filtering knowledge to its accurate context of the given data so that the extracted information fits the model well to generate relevant answers.

Peng Wang et al (2017) in the paper titled "**FVQA: Fact-based Visual Question Answering**" proposes a specialized dataset which contains additional content about the image data. With this new dataset, the authors were able to extract the specific relevant content in order to correctly match with the extracted features of the images. Even though this approach potentially has good result, the task is tedious and requires extra resources and computing power. Issues may persist of incorrect feature extractions, irrelevant content matching or even over-fitting of the model. The question-to-query mapping (via LSTM) is found to be less accurate as well.

Zhou Yu et al (2018) in the paper titled "**Beyond Bilinear: Generalized Multimodal Factorized High-Order Pooling for Visual Question Answering** " proposes a co-attention architecture that is developed to mutually understand both the image and the question features, which reduces the features which are irrelevant and obtain more accurate features for image and question representations. This approach achieves more effective inculcation of the features generated from the image and the text features from the question. This approach can significantly outperform existing pooling approaches due its effectiveness in fusing both the extracted features from image and question.

Iqbal Chowdhury et al (2017), in the paper titled "**A Cascaded Long Short-Term Memory (LSTM) Driven Generic Visual Question Answering (VQA)**" propose VQA architecture with Cascaded LSTM. For the purpose of extracting image features a pre-

trained model named VGGNet-19 is used. LSTM is used for understanding the text features in a way that it is mapped to its POS tags at the same time. This method achieves a better result in comparison to other methods based on DAQUAR data as this method is loosely structured making it versatile while understanding data features. The main outcome of this model is its ability to cover a wide range of topics due to its unique learning method of associating text directly to its image features. But due to this, it is a daunting task to train maximum possibilities of image-question pairs. For solving this problem, a generalized way of model training is to be developed which can handle the general natural language. With future developments, different datasets related to this task can be infused to create an even better architecture capable of real life applications. Such a model can even be evolved into a QA system of videos.

Geonmo Gu et al (2017), in the paper titled "**Adaptive Attention Fusion Network for Visual Question Answering**" proposes a VQA model based on Attention of context in a given image or scenario which can adapt itself for different categories of images. An attention map of the image is collected at each level of word embedding. This produces accurate information about the features as well as the context of the features in a given image. Due to this, the model can train efficiently by adapting to different attentions of different data on the basis of word, phrase and question levels.
The proposed method achieved 66.9% accuracy for Object related questions, 52.6% for Number related questions, 63.1% for Color based questions, and 57.1% accuracy Location based questions. And subsequently achieved an overall accuracy of 64.7% which is comparatively better than most of the VQA models.

Dongchen Yu et al (2018) in the paper titled "**Structured Semantic Representation for Visual Question Answering**" introduces a model which uses a Divide and Conquer strategy by decomposing questions into smaller chunks to extract only the relevant reasoning and context of the question with the help of Tree-LSTM architecture. A negative sample is generated to utilize the VQA 2.0‟ s complementary image data. A dual path network is developed for using this new feature of the VQA 2.0 dataset. The model achieves a training accuracy of 53.16% using this dual path network and a baseline accuracy of 51%. With improvements of the network, and better representations of features, the model can achieve a better result.

Nelson Ruwa et al (2018) in the paper titled, "**Affective Visual Question Answering Network**" (AVQAN) introduces an attention model which focuses on the image features, text features and also the mood present in the question, making it a three-way attention model producing precise answers. All these three data parts are fused to generate a faster LSTM module using only a single LSTM block and a soft-max classifier for processing the features. The AVQAN model was evaluated on the Visual 7W dataset using a specialized mood based attention module to establish the relations of images to question on the basis of mood. The achieved accuracy of 54% does show some potential in this approach but still needs to be refined in a way that only the relevant parts of the image dataset is considered for the task.

Peter Anderson et al (2018) in the paper titled, "**Bottom-up and top-down attention for image captioning and visual question answering**" introduces a bottom-up and top-down attention technique that enables attention to be workable at object level and other image regions levels. The bottom-up approach proposes important image regions, each with a related vector feature, while top-down approach controls feature weightings. The test results of the VQA dataset achieve the highest accuracy of 70% which still leads the VQA challenge leader-board.

3. Proposed System

We propose a VQA system based on external knowledge-base [13] as we have found it provides better accuracy and overall gives relevant answers to the questions. The MSCOCO [1] dataset comes with annotations related to the images with its related feature regions. This makes it easy for the model and improves accuracy. We try to feed the model a knowledge base to improve the understanding of the image. The knowledge-base[6] [13] consists of facts about common objects; these facts will be mapped in the model building process. We would also like to try transfer learning as to improve the system's knowledge with the increasing development in the datasets and knowledge-bases and also to refine the performance by some extent. Transfer learning gives the model itself an ability to shift from one specialization to other which will be beneficial for implementing applications in different domains with different datasets. The proposed model is based on multi-layered CNN and RNN (LSTM). The Convolutional networks would handle the image processing whereas the RNN (LSTM) would help in understanding the questions and the knowledge-base.

We start with preparing our data; training images from the MSCOCO [1] dataset is resized into a fix size and converted into gray-scale as this makes it easy for the neural networks to handle the data evenly. This data is then converted into vectors for feeding it to the neural network. This saved data consist the features vector of the image.

With the model, we define a CNN+LSTM network to which the saved data is given as input. The CNN layers will take the images and features to build an image model. RNN is used to build the model for the understanding the question and the knowledge-base [13] [6] by tokenizing the questions and the knowledge base and then extracting its features and using word embedding to form the vectors of the associated data. After building the individual models for image, question and knowledge-base [13] [6] the models are merged to form a single model and which is then fed to the network and forming a trained model. This trained model can be used in the application where an image and its associated question are taken as inputs which results in an output answer. Figure 1 shows the flow and working of the implemented system.

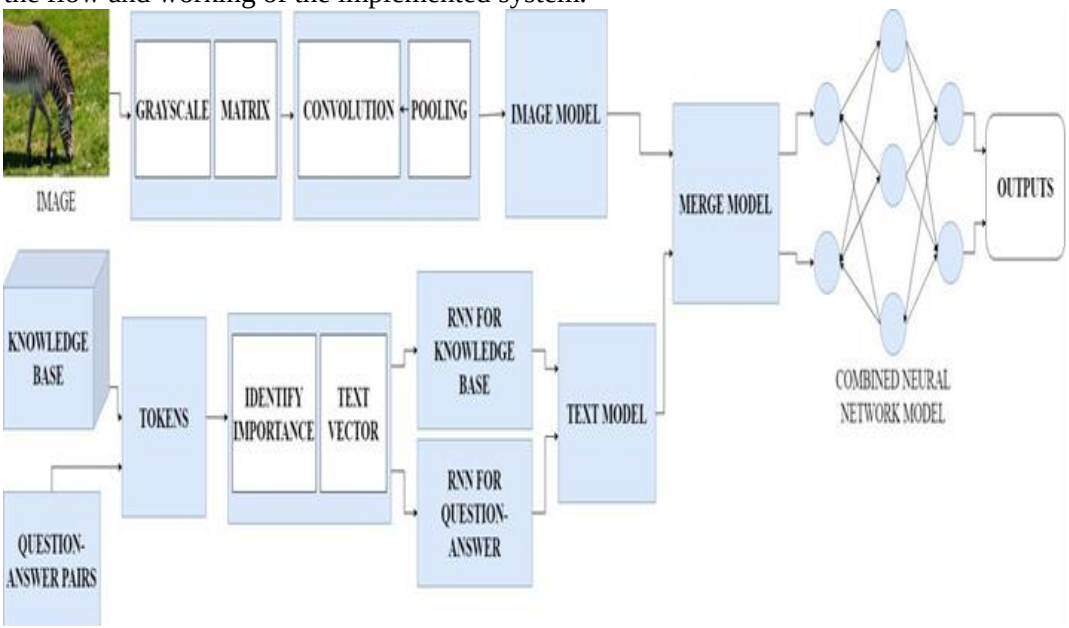

Figure 1. System Architecture

4. Datasets

With the increasing developments, there has been an increase in the datasets. Datasets such as VQA v1 and v2 [2], MSCOCO [1], KVQA [14], Visual-Genome [3] and Flickr[6]

provide large amounts of images. These datasets also improve as the VQA models improve.

a) MSCOCO [1]

MSCOCO is a large-scale object detection, super-pixel stuff segmentation, and 330,000 images dataset. With features including Object Segmentation, content recognition, super-pixel stuff segmentation. The dataset comprises of 330,000 images out of which more than 200,000 images are labeled with over 1.5 million object instances. MSCOCO Dataset consists of different categories such as 80 object categories and 91 stuff categories. Each image present in the dataset comes with 5 captions associated to the respective image.

b) VQA dataset [2]

VQA is a recently dataset which has open-ended questions about images. These textual questions require an understanding of perspective, language and common-sense knowledge to answer. The VQA dataset consist up to 270000 images (COCO and abstract scenes) with at least 3 questions per image and supporting ground truth answers per question (3 plausible answers per question). The dataset comes with automatic metric evaluation.

c) Visual Genome [3]

Visual Genome is a dataset, a knowledge base, which connect structured image concepts to language and defines a new way to generate more accurate features. The data comprises of 108,077 images and 540,000 million region descriptions. There are 170,000 visual question answers with 380,000 million object instances, 280,000 attributes and 230,000 million relationships.

d) Flickr8k [6]

Flickr dataset is a dataset comprised of images and image description in a sentence. There are two variants Flickr8k and Flickr30k which consist 8000 images and 30,000 images with its respective descriptions respectively.

5. Challenges

The main issue with generalized systems like VQA is the resources required for the task. Resources such as high performance processors, high speed memory and storage are needed for the training process of the models.
The dataset is another factor that needs to be considered during the implementation of VQA. Proper datasets with enough images and question-answer pairs are difficult to create or obtain as it is tedious task. Standard datasets for the VQA task is MSCOCO dataset and VQA v1, v2 dataset.

6. Conclusion

In figure 2, accuracies of the approaches on their respective datasets are compared. The chart shows that approaches [11] [14] on VQA achieves the highest accuracy of 70%. This shows that approaches based on knowledge-bases with attention provides higher accuracy.

Table 1. Results of different approaches

Author	Dataset	Accuracy (in %)
Yuetan Lin [4]	Toronto COCO-QA	59.66
Qi Wu [5]	VQA dataset	59.50
Peng Wang [6]	MSCOCO	56.91
Iqbal Chowdhury [7]	DAQUAR	43.05
Geonmo Gu [9]	MSCOCO	64.7
Dongchen Yu [10]	VQA 2.0	53.16
Zhou Yu [11]	VQA 1.0, VQA 2.0,	69.2 70.92
Nelson Ruwa [12]	Visual7W	54.7
Peter Anderson [14]	VQA 2.0	70.34

With this analysis, we find out that even the highest-achieved accuracy is of 70% and there is a good scope of improvement.

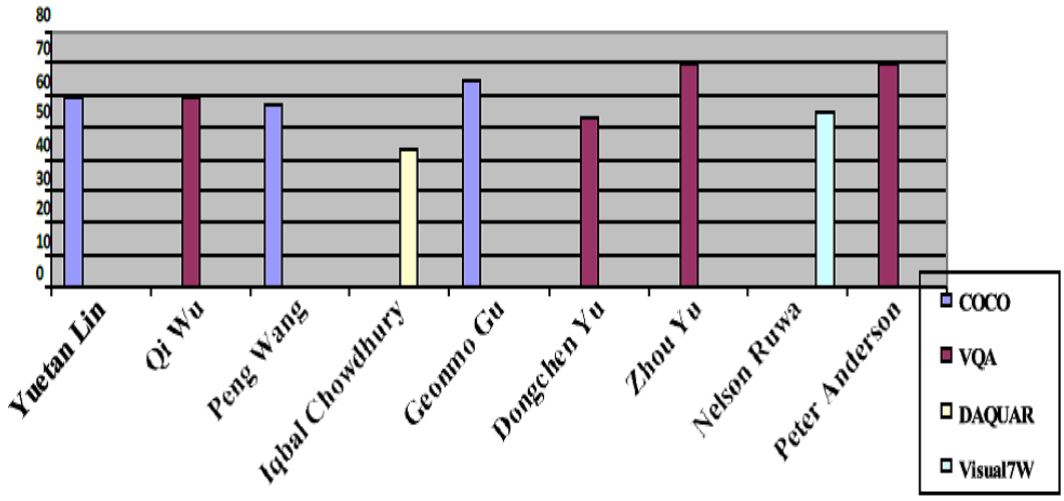

Figure 2. Comparative Analysis of Accuracies on different datasets

Models and approaches based on knowledge-base and attention achieves the goal of VQA with higher accuracies. With our proposed system, we try a similar approach with improvements of using transfer learning to increase the model's ability to answer relevantly to the asked question.

Acknowledgments

We are deeply grateful to our project guide Prof. K. P. Moholkar for the help in the field of Deep Learning and its approaches.

References

[1] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Doll´ar ”Microsoft COCO: Common Objects in Context”.

[2] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, Devi Parikh” Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering”.

[3] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, Li Fei-Fei” Visual Genome Connecting Language and Vision Using Crowdsourced Dense Image Annotations”.

[4] Yuetan Lin, Zhangyang Pang, Yanan Li, Donghui Wang” Simple and Effective Visual Question Answering in A Single Modality”, IEEE International Conference of Image Processing 2016.

[5] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, Anton van den Hengel,” Image Captioning and Visual Question Answering Based on Attributes and External Knowledge”,IEEE Transactions on Pattern Analysis and Machine Intelligence 2017.

[6] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, Anton van den Hengel”, FVQA: Fact-based Visual Question Answering”, IEEE Transactions on Pattern Analysis and Machine Intelligence 2017.

[7] Iqbal Chowdhury, Kien Nguyen, Clinton Fookes, Sridha Sridharan,” A Cascaded Long Short-Term Memory (LSTM) Driven Generic Visual Question Answering (VQA)”, IEEE International Conference of Image Processing 2017.

[8] Hongyang Xue, Zhou Zhao, and Deng Cai,” Unifying the Video and Question Attentions for Open-Ended Video Question Answering”, IEEE Transactions on Image Processing 2017.

[9] Geonmo Gu, Seong Tae Kim, Yong Man Ro” Adaptive Attention Fusion Network for Visual Question Answering”, IEEE International Conference on Multimedia and Expo (ICME) 2017.

[10] Dongchen Yu, Xing Gao, Hongkai Xiong” Structured Semantic Representation for Visual Question-Answering”, IEEE International Conference of Image Processing 2018.

[11] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao “Beyond Bilinear: Generalized Multimodal Factorized High-Order Pooling for Visual Question Answering”, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS 2018.

[12] Nelson Ruwa, Qirong Mao, Liangjun Wang, Ming Dong “Affective Visual Question Answering Network”, IEEE Conference on Multimedia Information Processing and Retrieval 2018.

[13] Sanket Shah, Anand Mishra, Naganand Yadati and Partha Pratim Talukdar “KVQA: Knowledge-Aware Visual Question Answering”, AAAI 2019.

[14] Peter Anderson, Xiadong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, Lei Zhang “Bottom-up and top-down attention for image captioning and visual question answering”, 2018.