Research Article

Visual Question Answering using Convolutional Neural Networks

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

[1,2,3,4,5]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, [5]aadarshvalsange99@gmail.com,

Article History: Received: 10 November 2020; Revised: 12 January 2021; Accepted: 27 January 2021; Published online: 05 April 2021

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 a trained system which can answer the questions associated to a given image in Natural Language. VQA is a generalized system which can be used in any image-based scenario with adequate training on the relevant data. This is achieved with the help of Neural Networks, particularly Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). In this study, we have compared different approaches of VQA, out of which we are exploring CNN based model. With the continued progress in the field of Computer Vision and Question answering system, Visual Question Answering is becoming the essential system which can handle multiple scenarios with their respective data.

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

1. Introduction

Artificial Intelligence (AI) has always been seen as a robotic system having the ability to think like a human, but AI can be technically distributed into parts such as Natural Language Processing (NLP), Computer Vision, Image Processing, and Text Processing. For a system to be called Human-like, it should be able to understand like a human and respond to a stimulus in a similar way that of a human. This is quite a challenging task, as the actual working of a human's way of thinking is still unknown. But, there have been progress in the development of such systems. Visual Question Answering is such a system, where it can understand a given image and answer the question asked upon the image. This is done in two parts; the system understands the features of a given input image and also analyzes the given question to find the importance in it and association between the words in the question and features of the image. Finally, an answer is generated in Natural Language.

As this task consists of two different parts of processing, individual processing of image and question and image-feature mapping must be done accurately to achieve the desired result. This is particularly dependent on the way of training the dataset and the choice of properly fine-tuned Neural Networks.

In this research, we have compared previous approaches of VQA by studying their model-training, accuracies, feature-extraction methods and use of dataset. We also propose an approach of implementing VQA with the help of Convolutional Neural Networks and Recurrent Neural Networks with the inclusion of external knowledge of the images of the dataset. The use of external knowledge helps the system to properly map the image information with its corresponding question-answer pair by providing additional details of the features in the image. This helps in decreasing random answers irrelevant of the image or question.

1.2. Related Work

There have been some approaches to tackle the challenge of VQA, mainly with the help of Artificial Neural Networks specifically Convolutional Neural Network (Qi Wu 2017) and Recurrent Neural Network (Iqbal Chowdhury *et.al*). We compared different approaches on the basis of their test accuracies of answering the question correctly. We found out that initial approaches such as (Yuetan Lin *et.al.*2016) were based on a smaller image and question dataset. Models which were based on external input or information (Qi Wu 2017) proved to be a better approach. As such an approach was not truly based on the given image data and an associated question-answer pair but consisted of additional information of the image which explains in detail about the extracted features of the image.

With more developments, models based on attention (Peter Anderson 2018) were developed. These models focus on the feature extraction phase of the image set. Feature extraction using VGGnet (Yuetan Lin *et.al.*2016) only accepts the image in a format shape of 224*224, and with attention models rigorously finding features in an image by either top-down attention model (Peter Anderson 2018) or adaptive attention (Geonmo Gu 2017) gave the feature extraction phase a better perspective by focusing on the major features of the image.

2. Proposed System

The main model is based on CNN (Convolutional Neural Networks) and LSTM (Long Short-term memory) for the task of image model and question-answer model (Kavita Moholkar *et.al*). The system is 4 modules: Image feature extraction module, question-answer vocabulary, image model and prediction module. The handling of the VQA task starts with processing and understanding the given image dataset. The image dataset used in this system is MSCOCO 2014 (Tsung-Yi Lin *et. al*). The processing is done by extracting the features of the images with the help of CNN pre-trained image model name VGG16, which is a 16 layered convolutional neural network which takes image input of size 224 by 224 in RGB (Red, Green, Blue). The processing of the images is done in batches of 10 images per batch until all the images are completed. This processing results in a feature list of shape (82568,4096) which is stored as a feature file in h5py format (h5). Along with this feature list file an image identification list (size of 82568) of associated image features is created. Both these files are then used as inputs for creating the image model.

The second stage of processing is associated to the questions and answers. Here, the annotations of the MSCOCO images are broken down in a format where the questions and answers are separated into two different files as question-answer and vocab file. The question-answer file is required as an input while training the model and the vocab file is used during the prediction stage to understand the given input question.

Image-id: COCO120001
Annotations:
<
Image-Id: COCO120001,
Question-Id:121001,
Question: What is this?
Answers: Zebra, confidence: yes,
Animal, confidence: maybe,
Horse, confidence: no.
>

Image Model

We are using VGG16 network to train the MSCOCO 2014 dataset, VGG16 is a pre-trained CNN network which comprises of 13 convolutional layers, 5 Max Pooling layers and 3 Dense layers. The image features extracted previously in the processing stage is given as input to this network to create a trained model. Along with the image feature, the question-answer vocabulary is given as input which is handled by the LSTM network.

The size of LSTM nodes is set to 512 for 3 layers. The image feature length is 4096 which is pre-determined by the VGG16 network. The dropout rate for both image and word embedding are set to 0.5. While training this model the batch size is set to 200 for 100 epochs or iterations with a learning rate of 0.001. This model combines the extracted image features with the associated question-answer air. With the given training amount and resources, the model achieves an accuracy of 60%.

The time taken to train this model is approximately 80 hours on an intel i7-8750h processor with 8gb of RAM.

Prediction Model

At this stage, we have two trained models of image and question data. These models are merged or combined to form a model which will generate the relevant answer, represented in Figure 1. The output from the image model is the classes or classifications of the features obtained from the images. As we've used VGG16 (Yuetan Lin *et. Al 2016*), 1000 classes are generated. The question model consists the importance of a question, from which the answer is generated by comparing the resultant answer classes and the important words of the question. Finally, the answers with highest percentage of confidence are displayed as the final answer. We are displaying the top 5 predicted classes with the highest percentage of confidence as answers. For our application we only choose one answer with highest confidence.

The architecture represented in Fig. 1 shows the connection of layers of the network of our model which is based on VGG16 (Yuetan Lin *et. Al 2016*), with inclusion of 3 layers of LSTM which handles the question-answer pair.

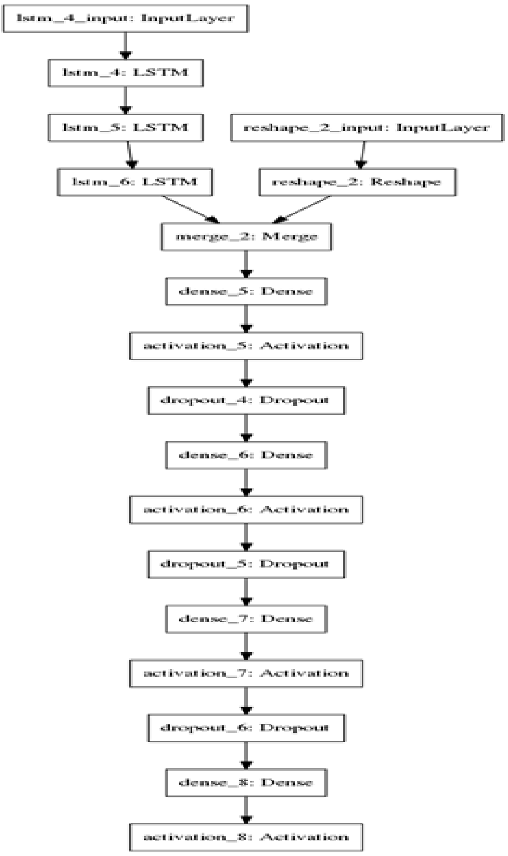

Figure 1. Architecture of connection of layers

At the end of the network, we have used Relu as our activation function as it provided better result overall for our model.

Transfer Learning

Transfer learning is the method in which a neural network is trained on data, and the generated weights of the data is then used for another neural network rather than training the new data, in the sense to transfer the gained knowledge from the previously trained data. This is done to make the system more efficient for handling new data.

Our system is trained on the MSCOCO (Tsung-Yi Lin *et. al*) dataset of over 87000 training images, but to increase the capabilities the model is capable to accept a new training dataset to re-train the model with this new dataset on the basis of the existing data. This makes learning procedure of the system natural and improves the abilities of answering different situations or images. the weights of the pretrained dataset MSCOCO (Tsung-Yi Lin *et. al*) is given to another neural network. In this manner, our system can handle any dataset by giving the previously trained dataset.

3. Datasets

With the increasing developments, there has been an increase in the datasets. Datasets such as VQA v1 and v2 (Yash Goyal *et.al*), MSCOCO (Tsung-Yi Lin *et.* al), KVQA (Yuetan Lin *et. Al 2016*), Visual-Genome (Ranjay Krishna *et. al*) and Flickr (Peng Wang *et. al* 2017) provide large amounts of images. These datasets also improve as the VQA models improve.

a) MSCOCO (Tsung-Yi Lin *et. al*)
COCO is a large-scale object detection, segmentation, and captioning 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 (Yash Goyal *et.al*)
VQA is a new dataset containing open-ended questions about images. These questions require an understanding of vision, 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 evaluation metric.

c) Visual Genome (Ranjay Krishna *et. al*)
Visual Genome is a dataset, a knowledge base, an ongoing effort to connect structured image concepts to language. The data comprises of 108,077 images and 5.4 million region descriptions. There are 1.7 million visual question answers with 3.8 million object instances, 2.8 million attributes and 2.3 million relationships.

d) Flickr8k (Peng Wang *et. al* 2017)
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.

4. Discussion and Conclusion

In table 1, we have compared the accuracies of different approaches. Only few approaches reach the accuracy mark of 70% whereas rest of them lie in the range of 50-60%. Our model of VGG networks and LSTM resulted in training accuracy of 60% on MSCOCO 2014 (Tsung-Yi Lin *et. al*) training dataset comprising of 82000 images. The testing accuracy of this model reached 57% over the same dataset. In conclusion, we have found that artificially implementing a human nature of question-answering is a challenging but rewarding task. As VQA is a generalized system, its application is endless with the exception of availability of the scenario related dataset.

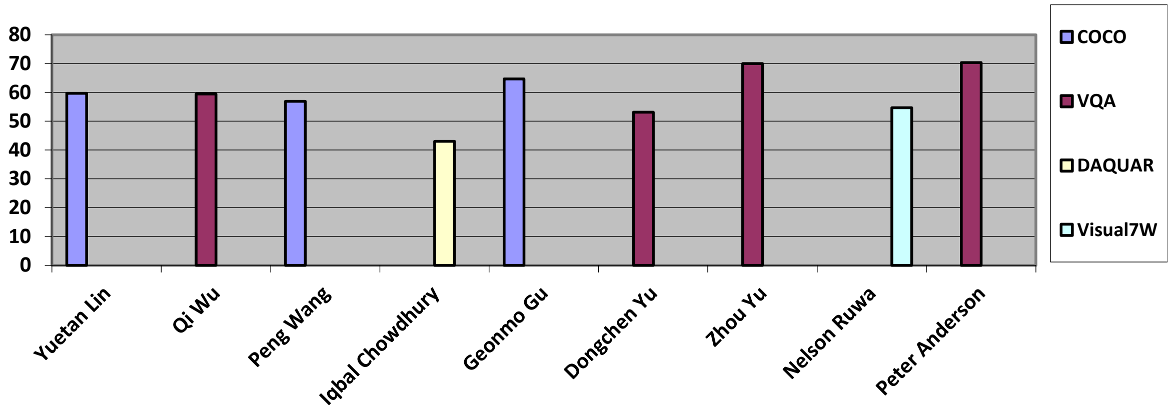

Figure 2. Comparative Analysis of approaches' accuracies on VQA, COCO, DAQUAR and Visual7W dataset

In Fig. 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 the different approaches

Author	Dataset	Accuracy (in %)
Yuetan Lin (2016)	Toronto COCO-QA	59.66
Qi Wu (2017)	VQA dataset	59.50
Peng Wang (2017)	MSCOCO	56.91
Iqbal Chowdhury (2017)	DAQUAR	43.05
Geonmo Gu (2017)	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. 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.

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 (2016) "Simple and Effective Visual Question Answering in A Single Modality", IEEE International Conference of Image Processing
5. Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, Anton van den Hengel, (2017) "Image Captioning and Visual Question Answering Based on Attributes and External Knowledge",IEEE Transactions on Pattern Analysis and Machine Intelligence 2017.
6. K. P. Moholkar, S.H. Patil, (2019) "A Question Answer System: A survey", International Journal of Computer Sciences and Engineering, Vol.7, Issue.3, pp.441-447, 2019.
7. Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, Anton van den Hengel (2017) "FVQA: Fact-based Visual Question Answering", IEEE Transactions on Pattern Analysis and Machine Intelligence 2017.
8. Iqbal Chowdhury, Kien Nguyen, Clinton Fookes, Sridha Sridharan, (2017) "A Cascaded Long Short-Term Memory (LSTM) Driven Generic Visual Question Answering (VQA)", IEEE International Conference of Image Processing 2017.
9. Hongyang Xue, Zhou Zhao, and Deng Cai, (2017) "Unifying the Video and Question Attentions for Open-Ended Video Question Answering", IEEE Transactions on Image Processing 2017.
10. Geonmo Gu, Seong Tae Kim, Yong Man Ro (2017) "Adaptive Attention Fusion Network for Visual Question Answering", IEEE International Conference on Multimedia and Expo (ICME) 2017.
11. Dongchen Yu, Xing Gao, Hongkai Xiong (2018) "Structured Semantic Representation for Visual Question Answering", IEEE International Conference of Image Processing 2018.
12. Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao (2018) "Beyond Bilinear: Generalized Multimodal Factorized High-Order Pooling for Visual Question Answering", IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS 2018.
13. Nelson Ruwa, Qirong Mao, Liangjun Wang, Ming Dong(2018) "Affective Visual Question Answering Network", IEEE Conference on Multimedia Information Processing and Retrieval 2018.
14. Sanket Shah, Anand Mishra, Naganand Yadati and Partha Pratim Talukdar. (2019) "KVQA: Knowledge-Aware Visual Question Answering", AAAI 2019.
15. Peter Anderson, Xiadong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, Lei Zhang (2018) "Bottom-up and top-down attention for image captioning and visual question answering", 2018.
16. K. P. Moholkar, S.H. Patil,(2020) "Multiple Choice Question Answer System using Ensemble Deep Neural Network" Second International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2020), 5-7 March 2020, Scopus Indexed ISBN : 978-1-7281-4167 https://ieeexplore.ieee.org/document/9074855