IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

VISUAL QUESTION ANSWERING

Sonali Vishwanath Thakur[1], Kalyani Dattatrey Shinde[2], Mohini Mahadev Pokale[3], Rupali Mangesh sathe[4]

[123]Student, [4]Professor

Department Of Information Technology Engineering.

Pillai Hoc College of Engineering & Technology, Rasayani, India.

Abstract: - Visual question answering (VQA) is a difficult task achieved by both image processing and natural language processing. The main goal is to combine convolutional neural networks and recurrent neural networks to get the common feature of image and question. We review available information on datasets to train and test VQA systems. Different datasets contain questions of different complexity that require different skills. We thoroughly examine the question-and-answer pairs and assess the suitability of the scene-based machine vision framework for VQA and also calculate the accuracy of the VQA model**.**

Index Terms: **- Convolutional Neural Network, Recurrent Neural Network (especially LSTM), Natural Language Processing, Activation Functions.**

I. INTRODUCTION

Artificial intelligence (AI) technology has been widely used to build elements of larger systems since the late 1990s, with applications in various fields. Advances in deep learning have led to an exponential growth in the number of AI applications. can be considered a subset of artificial intelligence. It is used in a variety of artificial intelligence tasks such as machine understanding, machine translation, computer vision, object recognition, etc.

VQA can be seen as an extension of the concept of automatic understanding. It is also a multidisciplinary artificial intelligence activity that combines advanced image processing and natural language techniques.

The VQA system takes an image and a natural language question on that image as input generates a natural language response. VQA requires an understanding of image as well as text. The main objective of VQA is that the reasoning part is done in the context of the image. So, if we have an image with the corresponding question, the system must be able to understand the image well to generate an appropriate answer. For example, if the question is the colour of the banana, the system must be able to detect the colour of the banana. Many of these common problems like face detection, binary object classification (yes or no), object detection, etc.

II. LITERATURE SURVEY

VQA System gain lots of interest in resent year. Before advance VQA model, there was existing system which analyses text and that can able to give answer a query about a text. Now, we come up with advance approaches, where all approaches involve below Extract feature from image, Extract feature form question, combine both image and question feature using algorithm and generate predictions.

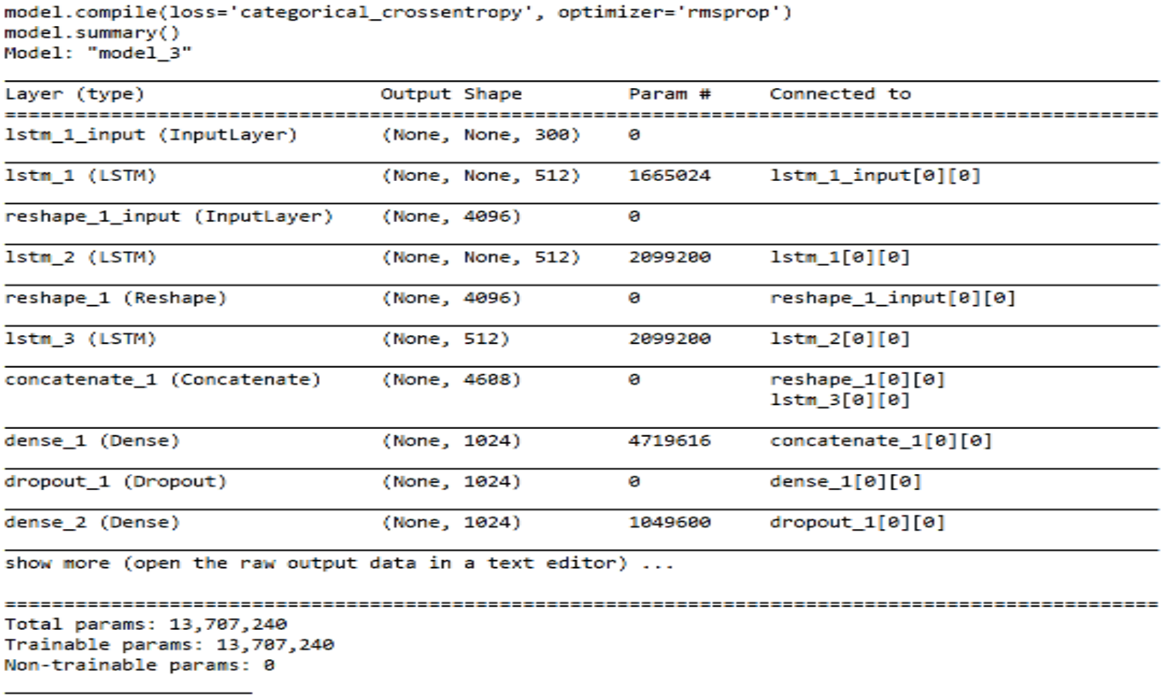
In 2019, VQA (Visual Inquiry Answering) is a relatively new activity that requires algorithms to reason about the visual content of a picture in order to respond to a natural language question. We compare the performance of state-of-the-art VQA algorithms on various VQA benchmarks in this paper. Each benchmark performs better at testing VQA algorithms at various levels. To begin, there's the joint embedding method, which focuses on how to map visual and textual data into a single embedding space. Second, attention-based strategies that concentrate on important aspects of the image or issue. Finally, there are compositional models, which deal with putting together a model from smaller components. Finally, they discuss algorithms that are based on external knowledge and require external sources to retrieve information.

In 2020, created a VQA system in which challenge is to offer an accurate natural language answer to a given question concerning an image. This article employs deep learning-based architecture to solve the challenge of visual question answering. Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) are two popular neural network designs used in the suggested model (LSTM). For encoding the given image and word embeddings for encoding the questions and feature extraction, they employed CNN. In addition, we employed LSTM to interpret the question. Finally, the results are compared to models such as the Multi-Layer Perceptron (MLP) and the Stacked Attention Network (SAN). The proposed model's effectiveness has also been compared.

VQA system was developed in 2021 to propose a method for solving the task of image-based question answering utilising EfficientDet and Bidirectional LSTM (BiLSTM). Image recognition and Natural Language Processing (NLP) techniques are used in Visual Question Answering (VQA). For image processing, EfficientDet is used, and for question processing, BiLSTM is used. Because of effective image processing, the model performs well. The model takes an image and a question as input and analyses them separately before fusing the results and predicting the answer to the query.

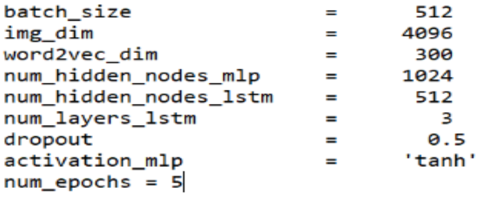
III. MODEL ARCHITECTURE

The model consists of two parts of input i.e., an image and a question that related to image the Functional API is used because cannot use the Sequential API of the Keras library which allows creating many models and finally merge them. As shown below diagram high-level architecture of neural network of submodules. After connecting the two separate models the into fully connected layer it will look like the shown in below following,



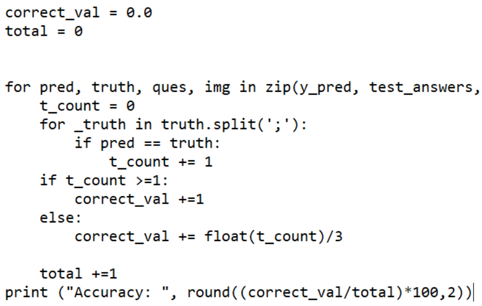
```
model.compile(loss='categorical_crossentropy', optimizer='rmsprop')
model.summary()
Model: "model_3"
_____________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
=============================================================================================
lstm_1_input (InputLayer)       (None, None, 300)    0
_____________________________________________________________________________________________
lstm_1 (LSTM)                   (None, None, 512)    1665024     lstm_1_input[0][0]
_____________________________________________________________________________________________
reshape_1_input (InputLayer)    (None, 4096)         0
_____________________________________________________________________________________________
lstm_2 (LSTM)                   (None, None, 512)    2099200     lstm_1[0][0]
_____________________________________________________________________________________________
reshape_1 (Reshape)             (None, 4096)         0           reshape_1_input[0][0]
_____________________________________________________________________________________________
lstm_3 (LSTM)                   (None, 512)          2099200     lstm_2[0][0]
_____________________________________________________________________________________________
concatenate_1 (Concatenate)     (None, 4608)         0           reshape_1[0][0]
                                                                 lstm_3[0][0]
_____________________________________________________________________________________________
dense_1 (Dense)                 (None, 1024)         4719616     concatenate_1[0][0]
_____________________________________________________________________________________________
dropout_1 (Dropout)             (None, 1024)         0           dense_1[0][0]
_____________________________________________________________________________________________
dense_2 (Dense)                 (None, 1024)         1049600     dropout_1[0][0]
_____________________________________________________________________________________________
show more (open the raw output data in a text editor) ...

=============================================================================================
Total params: 13,707,240
Trainable params: 13,707,240
Non-trainable params: 0
_____________________
```

After executing code, there are structure of the pre-processing directory as follow,



```
batch_size                   =      512
img_dim                      =     4096
word2vec_dim                 =      300
num_hidden_nodes_mlp         =     1024
num_hidden_nodes_lstm        =      512
num_layers_lstm              =        3
dropout                      =        0.5
activation_mlp               =     'tanh'
num_epochs = 5
```

Evaluation of model as following.



```
correct_val = 0.0
total = 0


for pred, truth, ques, img in zip(y_pred, test_answers, test_questions, test_image_id):
    t_count = 0
    for _truth in truth.split(';'):
        if pred == truth:
            t_count += 1
    if t_count >=1:
        correct_val +=1
    else:
        correct_val += float(t_count)/3

    total +=1
print ("Accuracy: ", round((correct_val/total)*100,2))
```

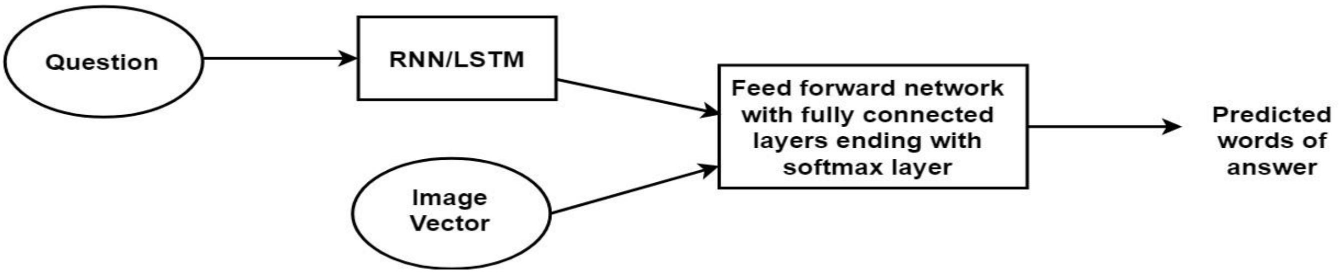

Question
RNN/LSTM
Feed forward network with fully connected layers ending with softmax layer
Image Vector
Predicted words of answer

Fig.2 VQA Model Architecture

IV. IMPLEMENTED TECHNIQUE

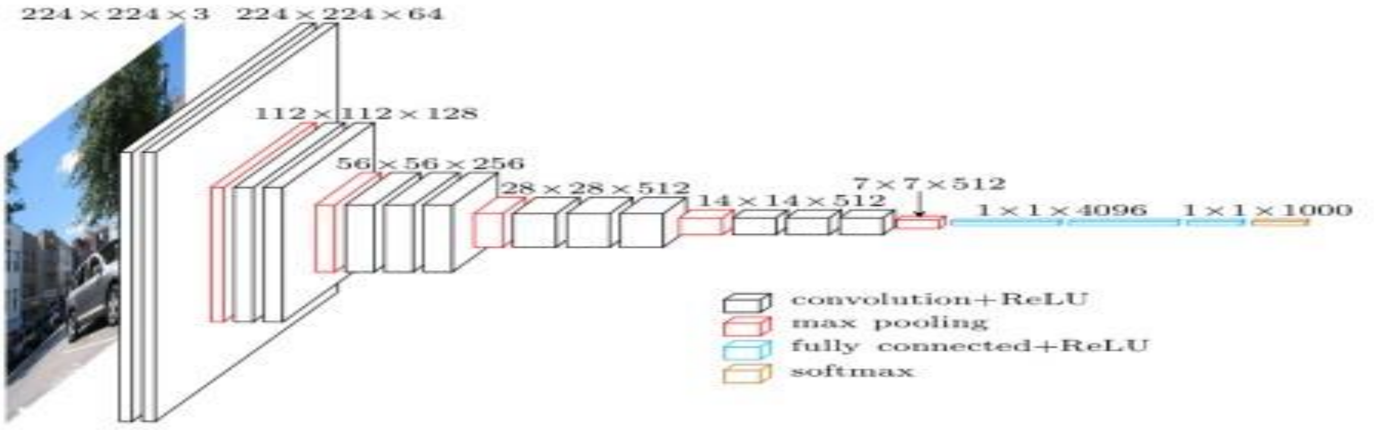
1. Pre-processing Data

The data are question related to image and answers.it is in the JSON format. All extracted data is stored in the .txt file format. After executing code, there are structure of the pre-processing directory as follow,

Answer.txt	// Answer for question
Images_id.txt	//image id for image
ques.txt	// question about images
ques_id.txt	//question id for each question
question_len.txt	// length of each question

2. Data Pre-processing-Image

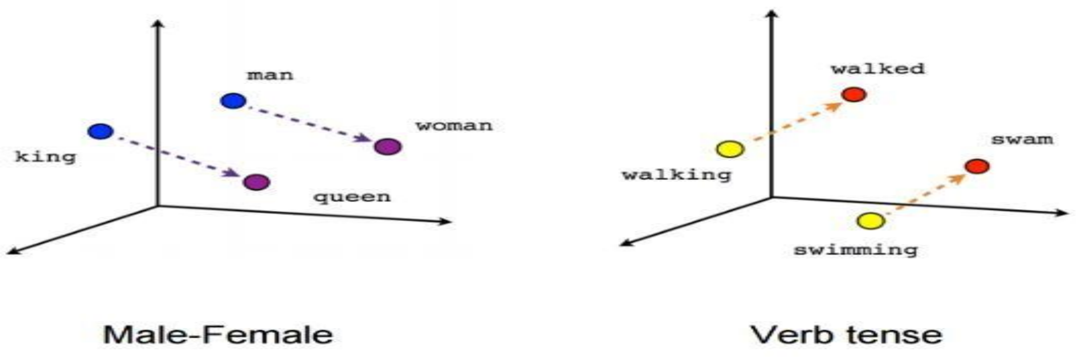
Image is input which is given to model. Before giving image to model it should be converted into vector form. therefore, every image is converted into vector form and then it is pass to neural network. Image is passed through VGG-16 model architecture is trained on the ImageNet dataset to classify the image into one of 1000 classes. The task image but to get the important features from image. Hence after removing the SoftMax layer, a 4096-dimensional in form of vector for each image is obtained.



For the VQA dataset, the images are from the COCO (Common Object in Context) dataset and each image has unique id associated with it. All these images are fed to the VGG-16 and their vector representation is stored in the “. mat” file along with id of image. So, in actual, we need not have to implement VGG-16 instead we just do look up into file with the id of image and we will get a 4096-dimensional vector representation for image.

3. Data Pre-processing - Questions

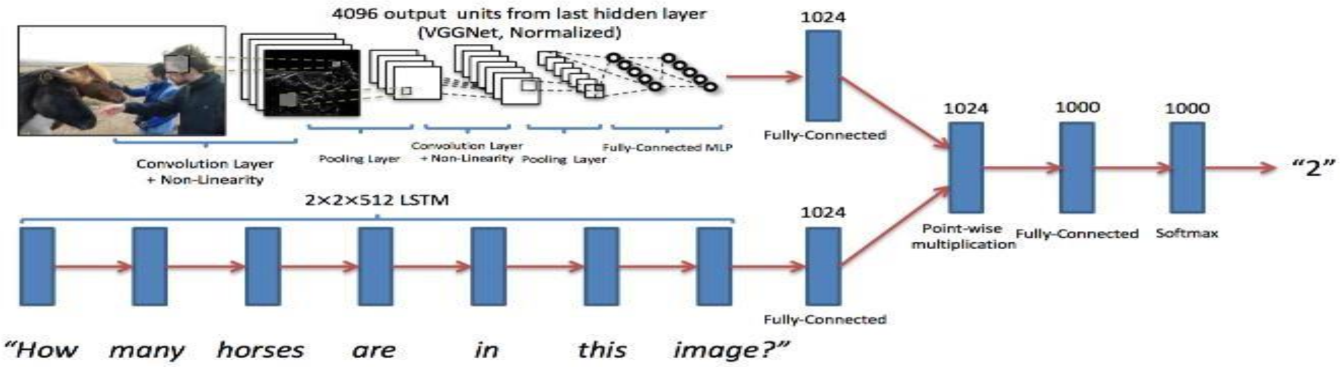
In the VQA dataset, the images are from COCO dataset and every image has it’s unique id. Each of these images are passed through the VGG-16 architecture and vector representation of image are stored in the “.mat” file along with there id. So, there is no need of implementing VGG-16 architecture, we just look up into a file with the id of the image and we will got a vector representation for the image.



This model is actually trained on millions of tokens. So we just need to call the of spaCy class for vector representation for word. For each question will get the 300-dimensional fixed representation for each word.

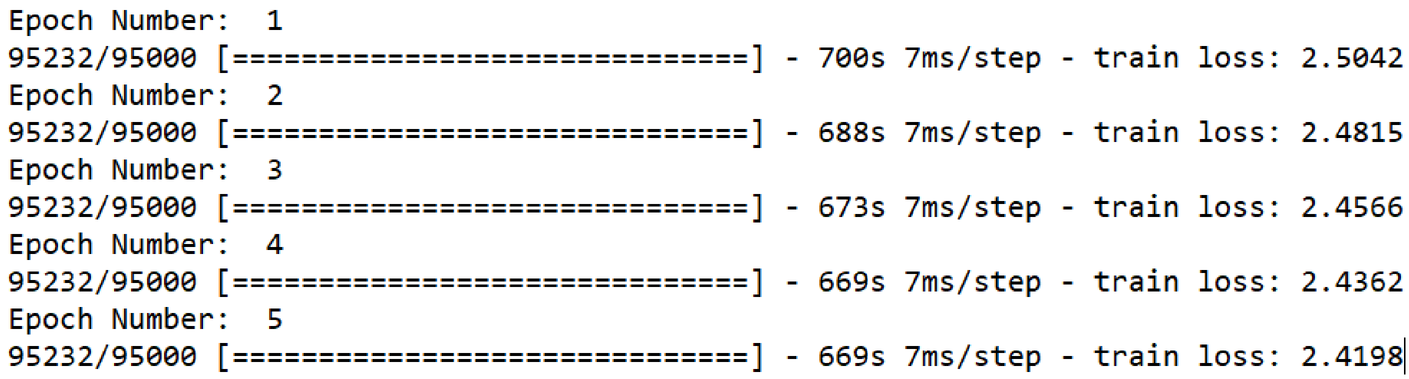
4. Model

VQA model is broken down into 2 broad concepts i.e., image and text i.e., question related to image. For this model we use the Convolutional Neural Network for image and Recurrent Neural Network for textual. we got the main features for images from Convolutional Neural Network and features for the text from Recurrent Neural Network and finally combine by passing them through some fully connected layers them to generate the answer.



5. Training Dataset

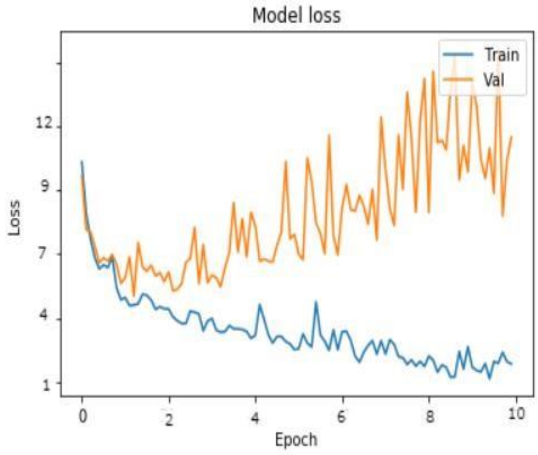
Using train-test split function split data into two subsets i.e. Training and Testing. In the training phase, we work on 9405 number of images, question and answer. The model is train up to 10 epochs. The number of epochs is directly proportion to accuracy of model i.e. if we increase number of epoch then accuracy of model is also increases. Our main aim is to make low loss as possible.



```
Epoch Number:  1
95232/95000 [==============================] - 700s 7ms/step - train loss: 2.5042
Epoch Number:  2
95232/95000 [==============================] - 688s 7ms/step - train loss: 2.4815
Epoch Number:  3
95232/95000 [==============================] - 673s 7ms/step - train loss: 2.4566
Epoch Number:  4
95232/95000 [==============================] - 669s 7ms/step - train loss: 2.4362
Epoch Number:  5
95232/95000 [==============================] - 669s 7ms/step - train loss: 2.4198
```

V. RESULT ANALYSIS

For this project, we're using the Google Collaboratory platform to write code.
The dataset had 105175 questions, 105175 responses, and 105175 photos. Given an image and a query, we presented it as a classification issue. We will anticipate the answer from the top answers in relation to the image. Basic preparation techniques were used to keep the data in the proper format.
We split the dataset using train-test split with train dataset having 95000 data points and test dataset having 10175 data points.
Using the last hidden layer of the pretrained VGG16 model, we created picture features and saved them to disc. We used to construct the train and test datasets, which loads the features batchwise. With no overfitting.



The model produces promising outcomes with 3.6089 log loss on train and test data and 63. 52 percent accuracy, we conclude that this model outperforms previous models.

VI. CONCLUSION

VQA system we use two model i.e. Computer vision and text there it is necessary to have good understanding of both model. VQA model having capability to give answer of any type of questions with accuracy. One of the most frequently asked and open-ended issues is if these VQA systems, which are trained and evaluated on questions, are reliable. can truly be regarded systems with multiple choice answers in the datasets in real-life settings with "good performance" In most datasets, the questions that need to be answered are: Crowdsourcing is utilised to acquire data for the model's training. Are these concerns valid? sufficient to describe all possible questions that the system will almost certainly face? The way questions are phrased has a significant impact on feature engineering and, as a result, on the system's overall predictive ability This individual model to combines together and it will improve performance of system. Taking about all existing system there are still research is going on about the techniques with which the model is trained and evaluated.

VII. REFERENCES

1. Theodoros Bozinis; NikolaosPassalis; Anastasios Tefas Recognition (ICPR) “Improving Visual Question Answering using Active Perception on Static Images”,2020 25th International Conference on Pattern Recognition (ICPR),2021
2. Ahmad Hasan Siregar;Dina Chahyati “Visual Question Answering for Monas Tourism Object using Deep Learning”,2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS),2020
3. Rahul Gupta;Parikshit Hooda;Sanjeev;Nikhil Kumar Chikkara “Natural Language Processing based Visual Question Answering Efficient: an EfficientDet Approach” ,2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS),2020
4. Liyana Sahir Kallooriyakath;Jithin M V;Bindu P V;Adith P P “Visual Question Answering: Methodologies and Challenges',2020 International Conference on Smart Technologies in Computing”, Electrical and Electronics (ICSTCEE),2020
5. Aakansha Mishra; Ashish Anand; Prithwijit Guh “CQ-VQA: Visual Question Answering on Categorized Questions” ,2020 International Joint Conference on Neural Networks (IJCNN),2020
6. Sarath.S; Amudha.J “Visual question answering models Evaluatıon”,2020 International Conference for Emerging Technology (INCET)