IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

IJARSCT
Impact Factor 6.252

Volume 2, Issue 1, July 2022

Visual Question Answering

Dr. Sai Madhavi D[1], Durga Shreya M[2], Manasa A[3], Pooja U Joshi[4],
Professor and HOD, Department of Computer Science and Engineering (AI & ML)[1]
Students, Department of Computer Science and Engineering[2,3,4]
Rao Bahadur Y Mahabaleshwarappa Engineering College, Ballari, Karnataka, India

Abstract: *We propose the task of free-form and open- ended Visual Question Answering (VQA). Given an image and a natural language question about the image, the task is to provide an accurate natural language answer. Mirroring real-world scenarios, such as helping the visually impaired, both the questions and answers are open-ended. Visual questions selectively target different areas of an image, including background details and underlying context. As a result, a system that succeeds at VQA typically needs a more detailed understanding of the image and complex reasoning than a system producing generic image captions. Moreover, VQA is amenable to automatic evaluation, since many open- ended answers contain only a few words or a closed set of answers that can be provided in a multiple choice format. We provide a dataset containing ~0.25M images, ~0.76M questions, and ~10M answers and discuss the information it provides. Numerous baseline for VQA are provided and compared with human performance. In this model, we have exclusively introduced a feature of voice to text using Speech recognition, Google Text-to-Speech and pygame module.*

Keywords: Visual Question Answering

I. INTRODUCTION

Visual Question Answering is a research area about building a computer system to answer questions presented in an image and a natural language. First of all, let's examine three datasets in Visual Question Answering. Visual Question Answering (VQA) is an AI-complete task lying at the intersection of computer vision (CV) and natural language (NLP). Given an image-question pair, our model generates not only an answer, but also a set of reasons (as text) and visual attention maps. 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.

II. RELATED WORK

J. Pennington [2014]
J. Pennington R. Socher and C. Manningon "Glove: Global vectors forword representation" Which presents the techniques for classification such as CNN, and determining the CNN rule. Glove: Global vectors for word representation Presents the techniques for classification such as CNN, and determining the CNN rule. It is published in Proceedings of the 2014 conference on empirical methods in natural language processing(EMNLP),2014.

2. L.MA[2016]: L. Ma, Z. Lu, and H. Li, on "Learning to answer questions from image using convolutional neural network" which Provides an overview of different aspects of Visual Question Answering & algorithms used for the same The proposed method is Learning to answer questions from image using convolutional neural network Provides an overview of different aspects of Visual Question Answering & algorithms used for the same & the publication was on Thirtieth AAAI Conference on Artificial Intelligence,2016

3. P.GAO[2019]: P.Gao, Z.Jiang, H.You, P.Lu,
S.C.Hoi, X.Wang, and H.Lion "Dynamic fusion with intra-and inter-modality attention flow for visual question answering" which Gives information about Natural Language Processing & aspects of NLP used in Visual Question Answering and the publication was IEEE Conference on Computer Vision and PatternRecognition,2019.

DOI 10.48175/IJARSCT-5763

III. EXISTING SYSTEM AND DRAWBACKS

CQ-VQA: Visual Question Answering on Categorized Questions. The CQ-VQA is a novel two-level hierarchical but end-to-end model to solve the task of visual question answering (VQA). The first level of CQ-VQA, referred to as Question Categorizer (QC). The Question Categorizer uses attended and fused features of the input question and image. The second level, referred to as Answer Predictor. This model is evaluated on the Task Directed Image Understanding Challenge (TDIUC) dataset and is benchmarks against state-of-art approaches. Results indicate a competitive or better performance of CQ-VQA.

3.1 Limitations

- A VQA algorithm is a given a text- based question and an image, and it must produce a text-based answer.
- It combines problems from multiple areas of computer vision.

IV. METHODOLOGY

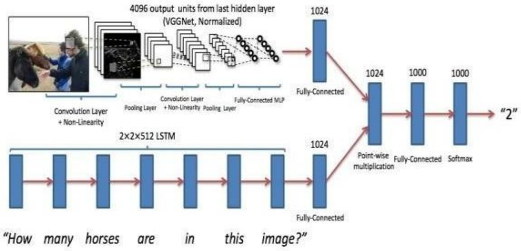


Figure: System Architecture of VQA

The learning architecture considers the problem, as a classification task, wherein, 1000 top answers are chosen as classes. Images are transformed by passing it through the VGG-19 Model that generates a 4096 dimensional vector in second last layer. The tokens in the question are first embedded into 300 dimensional vectors and then passed through 2 layer LSTMs. Both multimodal data points are then passed through a dense layer of 1024units and combined using point-wise multiplication. The new vector serves as input for a fully-connected model having a tanh and a final softmax layer.

V. ALGORITHM OVERVIEW

5.1 Convolution Neural Network (CNN)

Convolutional neural network is the special type of feed forward artificial neural network in which the connectivity between the layers are inspired by the visual cortex. Convolutional Neural Network (CNN) is a class of deep neural networks which is applied for analyzing visual imagery. They have applications in image and video recognition, image classification, natural language processing etc. Basically, the convolutional neural networks have 4 layers that is the convolutional layers, ReLU layer, pooling layer and the fully connected layer.

5.2 Convolutional Layer

In convolution layer after the computer reads an image in the form of pixels, then with the help of convolution layers we take a small patch of the images. These images or patches are called the features or the filters. By sending these rough feature matches is roughly the same position in the two images, convolutional layer gets a lot better at seeing similarities than whole image matching scenes. These filters are compared to the new input images if it matches then the image is classified correctly. Here lineup the features and the image and then multiply each image, pixel by the corresponding feature pixel, add the pixels up and divide the total number of pixels in the feature. We create a map and put the values of the filter at that corresponding place. Similarly, we will move the feature to every other position of the image and will see how the feature matches that area. Finally, we will get a matrix as an output.

5.3 RELU LAYER

ReLU layer is nothing but the rectified linear unit, in this layer we remove every negative value from the filtered images and replaces it with zero. This is done to avoid the values from summing up to zeroes. this is a transform function which activates a node only if the input value is above a certain number while the input is below zero the output will be 0

Pooling Layer In this layer we reduce or shrink the size of the image. Here first we pick a window size, then mention the required stride, then walk our window across our filtered images. Then from each window take the maximum values. This will pool the layers and shrink the size of the image as well as the matrix. The reduced size matrix is given as the input to the fully connected layer.

5.4 Fully Connected Layer

We need to stack up all the layers after passing it through the convolutional layer, ReLU layer and the pooling layer. The fully connected layer used for the classification of the input image. These layers need to be repeated if needed unless you get a 2x2 matrix. Then at the end the fully connected layer is used where the actual classification happens. Natural Language processing (NLP) refers to the branch of computer science and more specifically, the branch of artificial intelligence or AI concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

5.5 Recurrent Neural Network

Working of Recurrent Neural Network (RNN)

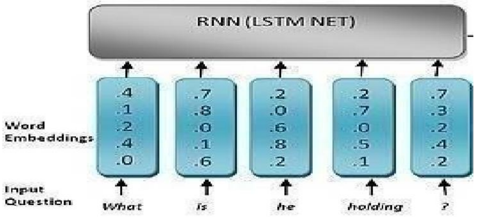


Figure: Recurrent Neural Network with specific weights

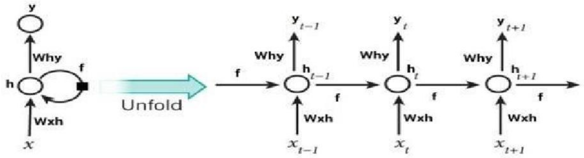


Figure: Working of RNN(Recurrent Neural Network

Here x_1, x_2, x_3,..,x_ it represents the ut words from the text, y_1, y_2, y_3,...,y_ itresent the predicted next words and h_0, ,h_2, h_3,..,h_t hold the information for thevious input words.

Since plain text cannot be used in a neuralork, we need to encode the words into tors. The best approach is to use word beddings but for this model, we will go for one-hot-encoded vectors. There are (V,1) tors (V is the number of words in our abulary) where all the values are 0, expect one at the i-th position. For example, if our abulary is apple, banana,.., king and the d is banana, then the vector is [0,1,..,0].

Define the equations needed for training:

$$1)\quad h_t = f(W^{(hh)}h_{t-1} + W^{(hx)}x_t)$$

$$2)\quad y_t = softmax(W^{(S)}h_t)$$

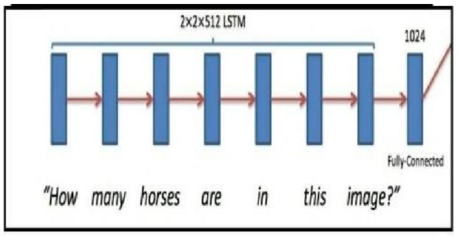
$$3)\quad J^{(t)}(\theta) = \sum_{i=1}^{|V|}(y'_{t_i} log\, y_{t_i})$$

1. Holds information about theprevious words in the sequence. As you can see, *h_t* is calculated using the previous *h_(t-*1*)*vector and current word vector *x_t*. We also apply an on-linear activation function *f* (usually tanh or sigmoid)to the final summation. It is acceptable to assume that *h_0* is a vector of zeros.
2. Calculates the predicted word vector at a given time step *t*. We use the softmax function to produce a *(V,1)*vector with all elements summing upto1.This probability distribution givesus the index of the most likely next word from the vocabulary.
3. Uses the cross-entropy loss functionat each time step *t* to calculate the error between the predicted and actual word.

VI. LSTM (LONG SHORT TERM MEMORY)

In the first step in our LSTM is to decide what information we are going to throw away from the cell state. This decision is made by a sigmoid layer called "forget gate layer". It looks at h_{t-1} and x_t and outputs a number between 0 and 1 for each number in the cell state C_{t-1}. 1 represents "completely keep this" while a 0 represents "completely get rid of this."

The next step is to decide what new information we are going to store in the cell state. This has two parts. First, a sigmoid layer called the "input gate layer" decides which values we'll combine these two create an update to the state. Finally we need to decide what we are going to output. This output will be based on our cell state, but will be a filtered version. First, we run a sigmoid layer which decides what parts of the cell state we are going to output. Then, we put the cell state through tanh (to push the values to be between -1 and 1) and multiply it by output of the sigmoid gate, so that we only output the parts we decided to.



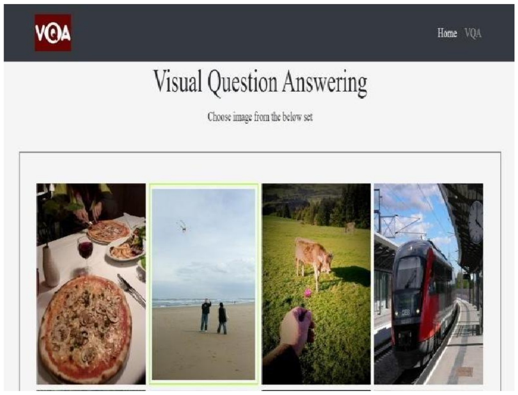
6.1 Multi- Layer Perceptron (MLP)

The image and question embeddings are combined via point-vise multiplication to obtain a single embedding. These combined embeddings are passed through as dense layer with 1024 units and tanh activation. This is followed by dropout and a softmax layer with 1000 nodes. The entire model is learned with cross-entropy loss with RSMProp optimizer. For training, we use 215359 training samples and report the performance using the answer that has the highest activation from potential multiple choice answers in the validation set which has 121512 samples.

6.2 Audio Mechanism

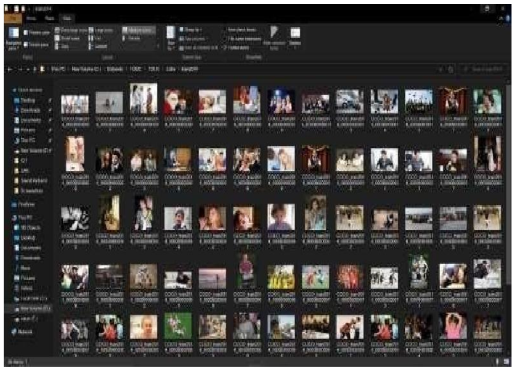
We used speech recognition in Python to convert the spoken words to text, to make a query. Speech recognition is a machine's ability to listen to spoken words and identify them. To get audio output directly we used Google Text-to-Speech Python Module. gtts is a tool that converts the text entered, into audio which can be saved as a mp3 file.

VII. RESULTS



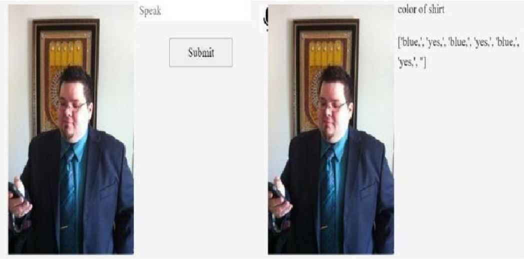
Snapshot 1: Snapshot of GUI

The above snapshot shows the snapshot of graphical user interface which consists of a header which reads Visual Question Answering and two buttons which are labeled choose file and submit respectively.



Snapshot 2 : Snapshot of Dataset

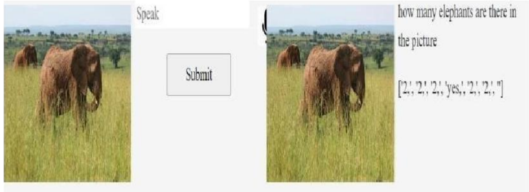
The above Snapshot shows the snapshot of the datasets, which here is a collection of various images belong to scenarios of day-to-day life.



color of shirt

['blue,', 'yes,', 'blue,', 'yes,', 'blue,', 'yes,', '']

Snapshot 3: Snapshot of output



how many elephants are there in the picture

['2,', '2,', '2,', 'yes,', '2,', '2,', '']

Snapshot 4: Snapshot of output

VIII. Conclusion

In this project work, a simple but powerful method is used to process images using CNN (Convolutional Neural Networks) architecture. In this project we focus on different methods for prediction and natural language processing techniques. We can modify available algorithms so as to obtain good accuracy while image processing. In this project we have introduced a feature of voice to text using Speech recognition module. This module helps visually impaired people to understand the image and also color blind people to differentiate between colors.

8.1 Future Improvement

One of the limitations of proposed approach is an inability to give accurate answer background objects of the picture. We plan to extend the VQA model for datasets with accuracy where the model would identify the background objects and its details.

REFERENCES

[1]. S. Antol, A. Agarwal, J. Lu, M. M. Mitchell, D. Batra, C. Lawrence Zitnick, and D.Parikh, “Vqa: Visual question answering,” in Proceedings of the IEEE international conference on computer vision, 2015,pp. 2425–2433.

[2]. A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra, “Vqa: Visualquestion answering,” International Journal ofComputer Vision, vol. 123,no. 1, pp. 4–31, May 2017.[Online].Available:https://doi.org/10.1007/s1 1263-016-0966-6.

DOI 10.48175/IJARSCT-5763

[3]. 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, 2016, pp.289–297.

[4]. H. Xu and K. Saenko, "Ask, attend and answer: Exploring question guideds patial attention for visual question answering," in European Conference on ComputerVision.Springer,2016, pp.451–466.

[5]. K. Kafle and C. Kanan, "Analysis of visual question answering algorithms," in Proceedings of the IEEE International Conference on Computer Vision,2017,pp.1965– 1973.

[6]. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition,2015,pp. 1–9