Visual Question Answering using combination of LSTM and CNN: A Survey

Riddhi N. Nisar[1], Devangi P. Bhuva[2], Prof. Pramila M. Chawan[3]

[1]B.Tech Student, Dept of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India
[2]B. Tech Student, Dept of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India
[3]Associate Professor, Dept of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India

---***---

Abstract – *In this article, we will study the concept of Visual Question Answering (VQA) using novel long short-term memory (LSTM) and convolutional neural network (CNN) attention models that combine the local image features and the question from corresponding specific parts or regions of an image to produce answers for the questions posed by making use of a pre-processed image dataset. Here, the word attention means using techniques that allow the model to only focus on those parts of the image that are relevant to both the object and the key words in the question. The irrelevant parts of the image are not taken into consideration and thus the classification accuracy is improved by reducing the chances of predicting wrong answers. We use the Keras python package with the backend of Tensorflow and the NLTK Python libraries for extracting image features using CNN, the language semantics using NLP and finally we use the multi-layer perceptron to combine the results from the image and question.*

Key Words: convolutional neural network, long short-term memory, word vector, keras, statistical bias, and multilayer perceptron

1. INTRODUCTION

The issue of solving visual question answering goes past the ordinary issues of image captioning and natural language processing, as it is a blend of both the strategies which makes it a perplexing system. Since language has a complex compositional structure, the issue of taking care of vision and language becomes a tough task.

It is very simple to get a decent superficial exhibition of accuracy when the model disregards the visual content on account of the predisposition that is available in the dataset. For this situation, the model doesn't genuinely comprehend the information embedded in the picture and just focuses on the language semantics, which is not what we need. For example, in the VQA dataset, the most widely recognized game answer "cricket" is the right response for 41% of the inquiries beginning with "What game is", and "white" is the right response for half of the inquiries beginning with "What shading is". These dialects can make a bogus impression of accomplishing precision. It is very conceivable to get cutting edge results with a moderately low comprehension of the picture. This should be possible by misusing the factual inclinations as well, which are present in the datasets. They are commonly obvious in standard language models as well. Presently, we need language to posture difficulties including the visual comprehension of rich semantics. The frameworks should not have the option to get rid of ignoring the visual data.

In this work, we propose to counter these language biases and make the role of image understanding in VQA more impactful by utilizing the underlying image features and the corresponding semantics of language.

1.1 Neural Networks

Neural networks are an approach to perform fast and efficient machine learning where there are normally numerous layers. These layers are full of various interconnected nodes and every node is initiated by an activation function. In the 'input layer' we typically input the pixels for images or words for natural language processing features for other machine learning problems, which in turn connects to one or more 'hidden layers' where the actual processing is done using activation functions and dependencies. The output of a neural network is calculated by training the network using forward propagation and back propagation steps, and then the model can be used for further analysis or testing purposes.

1.2 Supervised Learning

Supervised learning can be considered as a type of machine learning task in which a function can be learnt which will map an input to an output whenever a training set is provided which consists of these input-output pairs. Here we are using an unbalanced dataset of several types of images (for example: clever dataset, graphical dataset) and using supervised deep learning for image classification and feature extraction.

1.3 Recurrent Neural Networks

Neural networks have proven to be largely successful in the domain of computer vision. The convolutional neural networks (CNNs) are expert systems in taking images and extracting relevant features from them by using small windows that travel over the image. Similar to this, for text data we use recurrent neural networks (RNN). This kind of network is designed for sequential data and applies the same function to the words or characters of the text. These models have been very useful in translation (Google Translate), speech recognition (Cortana) and language generation.

2. LITERATURE REVIEW

2.1 Convolutional Neural Networks

In neural networks, convolutional neural network (ConvNets or CNNs) are one of the major categories to do the recognition and classification of various types of images. Computers see an input image as an array of pixels and the size of the array depends on the image resolution. It decides the pixels depending on the image resolution, for example we will see h x w x d (h = Height, w = Width, d = Dimension). Similarly, an image of 10 x 10 x 3 array of matrix of RGB (3 refers to RGB values) and an image of 8 x 8 x 1 array of matrix of gray scale image.

Deep learning is used in CNN patterns to prepare and test, where each information picture is gone by a progression of convolution layers which includes of channels (Kernels), completely associated layers, and all the more completely associated layers, and later a Softmax capacity is completed to group the item. Essentially, the picture is encoded into a vector, which has encoded in it all the basic highlights of the model to be needed for handling further in the figure for answer expectation.

2.2 Long Short Term Memory:

Long Short Term Memory networks are usually known as "LSTMs" and these are special kinds of recurrent neural networks which are intelligent enough to learn long-term dependencies. Long short-term memories have been designed for the sole purpose of avoiding long-term dependencies. They use gates and filters to remember old information. Their basic behaviour is to recollect pieces of information for long periods of time rather than for short periods, as is the case of vanilla recurrent neural networks.

2.3 Multilayer Perceptron (MLP)

A multilayer perceptron (MLP) is portrayed fundamentally as a neural system which is profound, which has shrouded layers present in it. As intended by the name, it has various perceptions. A multilayer perceptron contains info layer, managed to transfer the essential information highlights and a yield layer, which give out the conclusive outcome of the counterfeit neural system. In the middle of these two layers, we have at least one number of shrouded layers which are the policy units of the multilayer perceptron that figure the initiation capacities and go about as the cerebrums of the system.

3. PROPOSED SYSTEM

3.1 Problem statement

"To find the correct answer to a question posed based on an image using the technique of combination of language and vision via Keras, long short term memory, convolutional neural network and multilayer perceptron."

3.2 Problem Elaboration

This has been observed that finding the answers to questions based on an image correctly without inherent statistical bias on the dataset is a bit difficult. This leads to answers based on the dataset bias, which give quite accurate results, but without considering the features of the image. Its solution can be addressed with the help of methodologies that include image feature extraction using CNN and question semantic recognition and understanding using LSTMs.

To carry out the process of finding the correct answer to a question posed based on an image, we implement a python script using Keras, that encodes the question and image into vectors and then concatenates the two using MLP.

3.3 Proposed Methodology

There are different methods in the language+vision domain to find the answer to the question posed based on the input image. But each of the methodologies has their pros and cons. To work effectively with the proposed framework and after an exhaustive comprehension of the given writing review, the proposed approach appears to be a reasonable fit for accomplishing best in class exactnesses.

To use the multilayer perceptron model with respect to visual question answering, all the input questions need to be transformed into image feature vectors of a fixed length and similarly for the question posed. Then we combine both these results and apply element wise addition in the multilayer perceptron to get the final answer result. The following steps are proposed:

1. The first step will be the word transformation. For the question, we will convert each word to its word vector, and then sum up all the vectors. We have to make sure that the length of this feature vector matches the length of a single word vector, and these word vectors are also called embeddings.

2. In the next step, these word vectors will be sent sequentially to the long short-term memory network, respective to the tokens in the question. The representation of the input question is

actually the vector coming from the output gate of the long short-term memory.

3. Coming to the image, it is sent through a Deep Convolutional Neural Network (from the well-known VGG Architecture), and the image features are extracted from the activation of the second last layer (that is, the layer before the softmax function).

4. To combine the features from the image and the word vector, we use a multilayer perceptron consisting of fully connected layers. The final layer consists of the Softmax activation function, and thus we get a probability distribution over all the possible outputs. The output with the highest probability is our answer to the question posed based on the image.

3.4 Proposed System Architecture

The proposed system workflow is as given as shown in diagram:

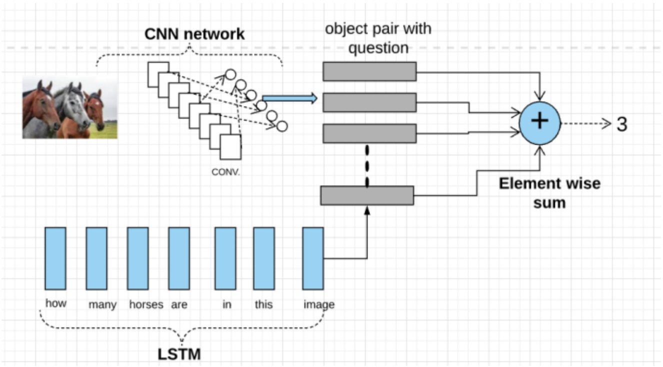

Fig - 1: Workflow

The above block diagram illustrates the architecture that we propose for the Visual Question Answering system. Here we import the **Keras** library to create a Convolutional Network layer for particular image and extract the required image features. LSTM as a part of RNN is used for natural language processing to convert the question into word vector, understanding its semantics. We merge the extracted image features and word vector and using MLP we create an (object, question) pair. Then the element-wise summation of these vectors results in the answer to the given example i.e. number of horses in the image.

4. CONCLUSION

Visual Question Answering is a research topic that covers parts of both computer vision and natural language processing and it requires a system to do much more than simple task-specific algorithms, such as object recognition and object detection. To build an algorithm that will be able to answer random questions about images would be a huge achievement in the artificial intelligence domain, and would have a large potential to benefit visually impaired users. In this paper, we proposed methodologies for visual question answering using the novel ideas of long short term memories, which are a better alternative to the vanilla recurrent neural networks for question semantic understanding and using convolutional neural networks for image feature extraction into vectors. We discussed the challenges of evaluating answers generated by standard algorithms which use unbalanced biased datasets. We also described how biases and other problems make existing datasets less informative for real-world test images.

REFERENCES

[1] Yin and Yang: Balancing and Answering Binary Visual Questions (CVPR 2016)

[2] VQA: Visual Question Answering by Aishwarya Agrawal , Jiasen Lu , Stanislaw Antol ,Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh

[3] Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, Devi Parikh, Virginia Tech, Army Research Laboratory, Georgia Institute of Technology.

[4]https://towardsdatascience.com/deep-learning-and-visual-question-answering-c8c8093941bc

[5]https://visualqa.org/

[6] https://tryolabs.com/blog/2018/03/01/introduction-to-visual-question-answering/

[7] Image Captioning and Visual Question Answering Based on Attributes and External Knowledge Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, Anton van den Henge

[8] Learning to Reason: End-to-End Module Networks for Visual Question Answering ICCV 2017 Ronghang Hu Jacob Andreas Marcus Rohrbach Trevor Darrell Kate Saenko

[9] Graph-Structured Representations for Visual Question Answering CVPR 2017 Damien Teney, Lingqiao Liu, Anton van den Hengel

AUTHOR'S PROFILES

Riddhi N. Nisar, Final Year B. Tech Student, Department of Computer Engineering and IT, VJTI, Mumbai, Maharashtra, India.

Devangi P. Bhuva, Final Year B. Tech Student, Department of Computer Engineering and IT, VJTI, Mumbai, Maharashtra, India.

Prof. Pramila M. Chawan is working as an Associate Professor in the Computer Engineering Department of VJTI, Mumbai. She has done her B. E. (Computer Engineering) and M.E (Computer Engineering) from VJTI COE, Mumbai University. She has 27 years of teaching experience and has guided 75+ M. Tech. projects and 100+ B. Tech. projects. She has published 99 papers in the International Journals, 21 papers in the National and International conferences/symposiums. She has worked as an Organizing Committee member for 13 International Conferences, one National Conference and 4 AICTE workshops. She has worked as NBA coordinator of Computer Engineering Department of VJTI for 5 years. She had written proposal for VJTI under TEQIP-I in June 2004 for creating Central Computing Facility at VJTI. Rs. Eight Crore (Rs. 8,00,00,000/-) were sanctioned by the World Bank on this proposal.