VISUAL QUESTION ANSWERING(VQA)

ABSTRACT

Task of visual question answering is to automatically answer natural language questions with reference to a given image. It deals with image segmentation and object recognition. Real world scenarios include predicting answers for open ended questions. Visual questions selectively target different areas of an image including background details and underlying context. VQA is an ongoing research project where accuracy obtained is around 70. The task of Visual Question Answering (VQA) is receiving increasing interest from researchers in both the computer vision and natural language processing fields. Tremendous advances have been seen in the field of computer vision due to the success of deep learning. Inspired by the recent success of text-based question answering, VQA is proposed to automatically answer natural language questions with the reference to a given image.

INTRODUCTION

We are witnessing a renewed excitement in multi-discipline Artificial Intelligence (AI) research problems. Research in image captioning that combines Computer Vision (CV), Natural Language Processing (NLP), and Knowledge Representation & Reasoning (KR) has dramatically increased in the past year. Automatically answering questions about visual content is one of the highest goals of Artificial Intelligence. Visual question answering (VQA) involves an image and a related text question, to which the machine must determine the correct answer. This task spans the fields of computer vision and natural language processing, since it requires both the comprehension of the question and parsing the visual elements of the image. VQA has attracted extensive attention recently, since VQA is considered approaching towards the milestone of “AI-complete” that enables a machine to reason across language and vision as humans. Compared with text-based QA system in natural language processing (NLP), VQA takes one step further, which can answer a natural language question by considering the correspondence between a question and a reference image. The capability of automatic VQA can significantly promote the mutual understanding between language and vision, and further benefit a variety of applications.

A second parallel motivation for the study of VQA is its utility. A system capable of answering questions about images has direct practical applications, such as a personal assistant, or in robotics as aids for the visually impaired. Note, however, that current VQA data sets do not directly address this setting, because questions are typically collected in a non-goal-oriented setting. Realistic, motivated questions would likely require information not present in the image and involve rare words and concepts. In comparison, most questions in current data sets are purely visual and centred on common concepts. VQA particularly embodies this confidence in achieving high-level image understanding. It has also attracted significant interest in the past few years and can be compared to VQA as they both combine vision and language. The two tasks are complementary as they evaluate different capabilities. Captioning requires mostly descriptive capabilities that involve

almost purely visual information. VQA, in comparison, often requires reasoning with common sense and with other information not present in the given image. In this respect, VQA constitutes an AI-complete task since it requires multimodal knowledge beyond specific domains. This reinforces the motivation for research on VQA, as it provides a proxy to evaluate progress toward general AI, with systems capable of advanced reasoning combined with deep image and language understanding.

PROBLEM DEFINITION

An instance of VQA consists of an image and a related question given in plain text. The task for the machine is to determine the correct answer, which is, in current data sets, typically a few words or a short phrase. Two practical variants are usually considered, an open-ended and a multiple-choice setting. In the latter, a set of candidate answers are proposed. This makes the evaluation of a generated answer easier than in the open-ended setting, where the comparison between the machine's output and a ground truth (i.e., human provided) answer faces issues with synonyms and paraphrasing. In this sense, VQA more closely reflects the challenges of general image understanding. For example, compare the phrase “a red hat” with the multitude of its representations that one could picture, e.g., with many different styles and details that cannot be described in a short phrase.

Let us mention the relation of VQA with the task of automatic image captioning. Captioning requires mostly descriptive capabilities that involve almost purely visual information. VQA, in comparison, often requires reasoning with common sense and with other information not present in the given image. In this respect, VQA constitutes an AI-complete task since it requires multimodal knowledge beyond specific domains.

PROBLEM IMPLEMENTATION

1) We have imported vgg16 predefined model from keras.applications. It is a convolutional neural network model with max pooling. It has 16 layers in which 13 belongs to CNN and rest are fully connected networks

2) Using Numpy, we have loaded the labels.

3) Pre trained weights taken from image net are loaded into vgg16 model and chosen image from file directory will be connected into vectors and referred as reshape layer

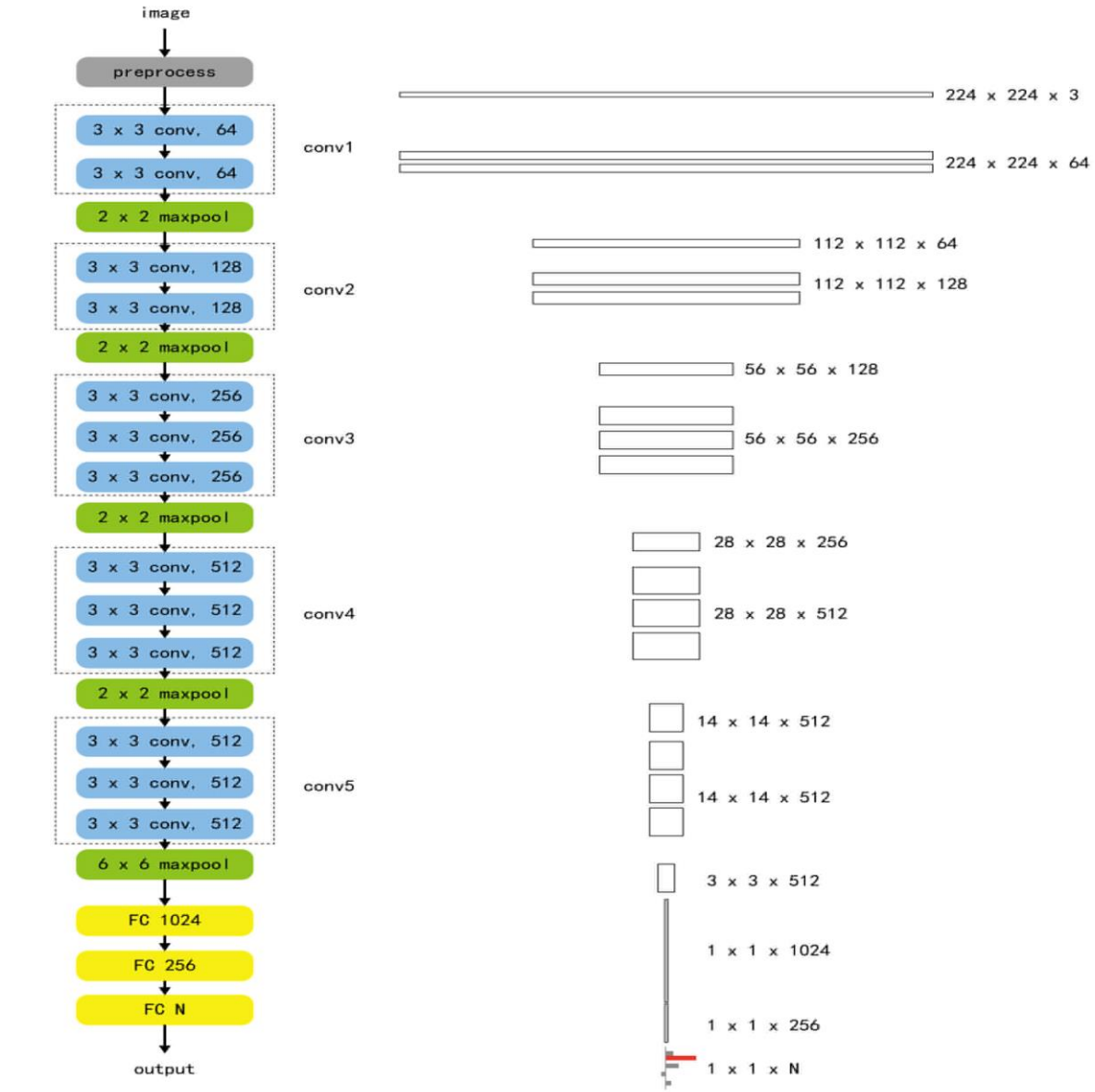
4) Given question is tokenised into words and considered as lstm (long short term memory) layer.

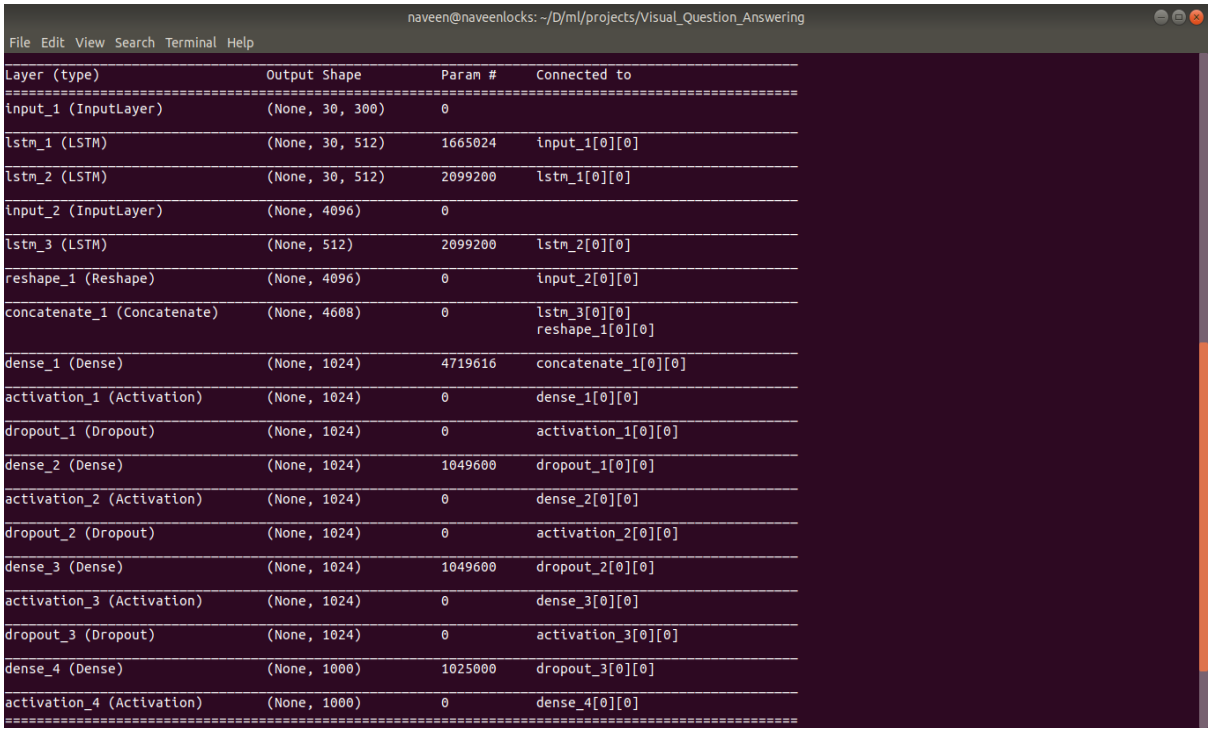
5) Above two layers are concatenated and feeded into CNN. Recursively false output label is given as an input by optimizing using SGD (stochastic gradient descent)

CONVOLUTION NEURAL NETWORK

When programming a convolutional layer, each convolutional layer within a neural network should have the following attributes:

- Input is a tensor with shape (number of images) x (image width) x (image height) x (image depth).
- Number of convolutional kernels.
    - Width and height of kernels are hyper-parameters.
    - Depth of kernels must be equal to the image depth. Convolutional layers apply a convolution operation to the input, passing the result to the next layer. The convolution emulates the response of an individual neuron to visual stimuli.





POOLING

Convolutional networks may include local or global pooling layers. Pooling layers reduce the dimensions of the data by combining the outputs of neuron clusters at one layer into a single neuron in the next layer. Local pooling combines small clusters, typically 2 x 2. Global pooling acts on all the neurons of the convolutional layer. In addition, pooling may compute a max or an average. Max pooling uses the maximum value from each of a cluster of neurons at the prior layer

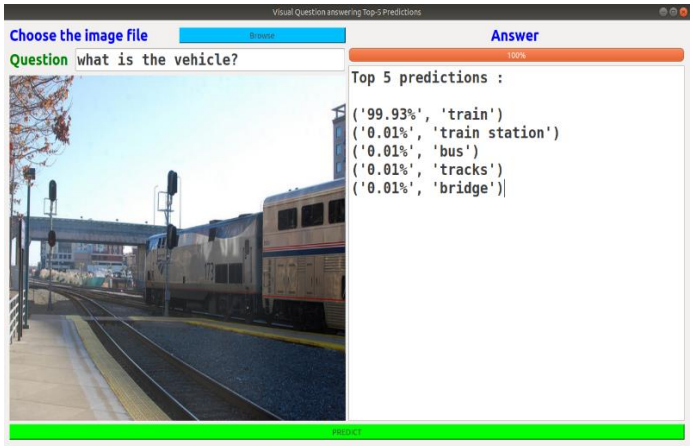
FULLY CONNECTED NETWORK

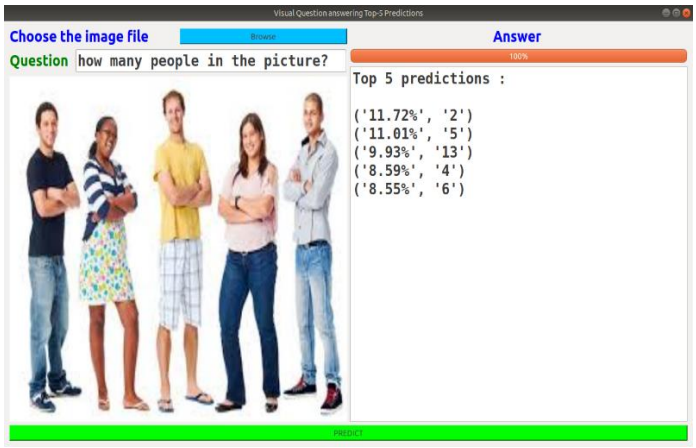
Neurons in a fully connected layer have full connections to all activations in the previous layer, as seen in regular Neural Networks. Their activations can hence be computed with a matrix multiplication followed by a bias offset.

DATA SET

We used pre trained vgg16 (Visual geometry group) model. Data used in vgg16 is 123,287 training and validation images and 81,434 test images from the newly-released Microsoft Common Objects in Context (MS COCO) dataset. Data labels and weights are taken from image net.

INPUT AND OUTPUT





Related answer will be ensured to be in top 5 predictions. Since our project is an on going research project, Maximum accuracy achieved for our VQA is around 70 percent with resource constraint we have.

SOFTWARE AND HARDWARE

Software Language: Python Distribution PIP, NumPy, Matplotlib, Scikit-learn, Kera's, Tensorflow

Operating system: Linux

Hardware Processor: Intel Core i5 Coffee Lake

REFERENCES

1) Visual Question Answering: A Tutorial

Authors: Damien Teney, Qi Wu, Anton van den Hengel

2) Image-Question-Linguistic Co-Attention for Visual Question Answering

Authors: Chenyue Meng, Yixin Wang, Shutong Zhang

3) VQA: Visual Question Answering

Authors: Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh