VQA with Yes or No Questions

A Project for COMP.5450 Machine Learning Spring 2016 at UMass Lowell

Joakim Monk (yodaman11@gmail.com)

May 4, 2016

Introduction

The goal of visual question answering is for the machine to take an image and a natural language question about the image and output a correct natural language answer. In this project, the questions are restricted to yes or no questions and the images are only of a certain abstract type, composed of clip-art style images. This project gets its data from the Visual Question Answering Challenge at <http://visualqa.org/>.

A machine that was good at visual question answering would be a huge step in machine learning and artificial intelligence. Being able to answer questions about what is seen is essential to measuring understanding in the human sense. One can imagine many practical applications of such a machine. One general idea would be that it can take a question and then search through images much faster than any human could in order to determine which images satisfy the question. Also, depending on the sophistication of the machine, it may be able to analyze an image better than almost any human.

Background

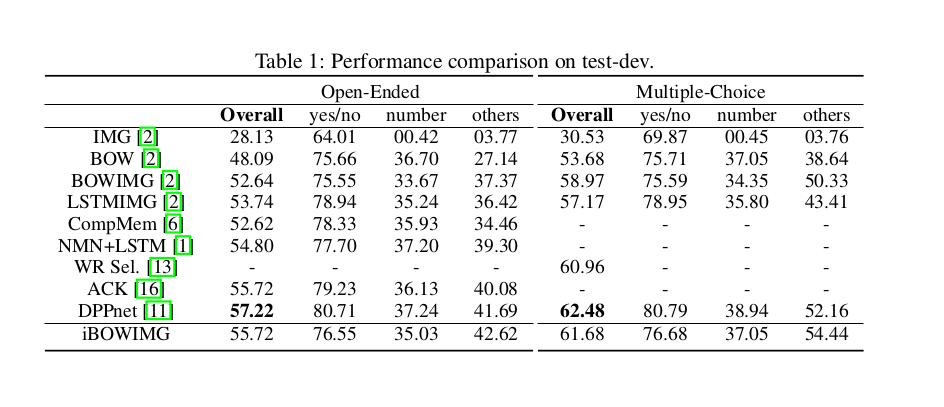
The creators of the challenge at visualqa.org have provided data and there own attempt at the challenge (See 1). In their attempt, they create a bag-of-words representation for the questions, and pass the image through VGGNet (See 2), a 19 weight layer convolutional neural network used for image recognition. They take the last hidden layer of VGGNet as their image feature. They then try two different approaches to get the final classification: a multi-layer perceptron neural network classifier with 2 hidden layers and tanh non-linearity, and an LSTM model followed by a softmax layer to generate the answer.

Another group used naive bag-of-words as the text feature, and the deep features from GoogLeNet as the visual features (See 3). In their paper, they have a comparison of many of the different techniques tried in the literature, shown below in Figure 1.

We can see that overall accuracy for general, open-ended questions is at most 57.22%, and for just yes/no questions is at most 80.71%. In general, it seems that the question is more important than the image in generating a correct answer, which shows a bias in the types of questions and answers the images are getting. Interestingly, for yes or no questions, using just the questions works just about as well as using the question and image in training.

Approach

I have adapted the code provided by (1) to apply to this specific problem. The approach of group (1) uses Python, Lua, NLTK, and Torch. Torch allows for training using the GPU.

Figure 1. Performance of various VQA techniques. From (3). IMG stands for image only, BOW for bag-of-words only, iBOWIMG is for this groups technique (simple naive bag-of-words) as opposed to the VQA challenge creators' BOWIMG technique.

Dataset

The data is a subset of the larger set from <http://visualqa.org/>. It contains 20,000 PNG images using an abstract clip-art style to depict various scenes, as well as questions to go along with the images. An accompanying annotations files lists those questions which have yes or no answers. An image from the dataset is given below.

FIgure 2. An image from the dataset.

Conclusion

A machine that could do visual question answering well would be an amazing achievement. The current methods used are not sufficient. I believe that in general there needs to be more clever concepts used in processing the images and questions in addition to the current large neural network approach.

References

1. Stanislaw Antol, et al. VQA: Visual Question Answering. International Conference on Computer Vision (ICCV), 2015.

2. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014.

3. Stanislaw Antol, et al. VQA: Visual Question Answering. International Conference on Computer Vision (ICCV), 2015.