Image-based Natural Language Question Answering

Bhargav Mangipudi
mangipu2

Pramod Srinivasan
psrnvsn2

Vishaal Mohan
vmohan9

Abstract

The aim of this project is to propose accurate natural language answers for natural language questions based on images. There are two parts to this project; the first is when the algorithm is given options to choose from and the second being the learner coming up with grammatically and semantically correct answers by itself. We attempt to use the state-of-art deep learning algorithms such as Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks.

1 Introduction

Deep learning has risen as the answer to many unsolved or partially solved machine learning problems. More specifically, image, video and audio analysis which posed a challenge because of the complex nature of the feature/representation to be learned, have become easier due to the reduced effort on the part of the user in feature selection.
Image-based question answering has been the grand challenge on the fronts of natural language understanding and computer vision. We plan to address this problem using deep learning models. (Hinton, Geoffrey E., 2006)

2 Related Work

Driven by the newly emerging data, multi-modal research involves leveraging techniques in Computer Vision (CV) and Natural Language Processing (NLP). The newly developed large-scale vision and language datasets have not only played a vital role in defining work in this area but also served as a foundation to set benchmarks for evaluating system performance. In addition, factors such as crowd-sourcing and large image datasets such as those provided by Flickr have enabled researchers to propose methods for vision and language tasks alongside an accompanying dataset. (Francis Ferraro, 2015)
Recent research has focused on real-world scenarios such as aiding the visually impaired where answers are sought to open-ended questions about the image. Often, such questions are about specific information about the image, the answers requires commonsense knowledge as well as visual understanding of the scene. (Haoyuan Gao and Junhua Mao, 2015) (Mengye Ren and Ryan Kiros, 2015). There is previous work with similar goals of answering questions based on images but with poor results.

2.1 Proposed Work

Due to the vast nature of this problem and given the time constraints, we plan to work on the first task of answering questions given multiple-choices and a constricted domain. For example, we will first address questions like, "How many fingers is he holding up?" and "How many wheels does the vehicle have?" and then move on to more complicated questions and diverse domains.
The initial phase of the project will involve a survey of the work that has been done to this end and choosing the dataset (based on the nature of question to be answered and the ease of the task on the chosen dataset). We plan to then build a CNN/LSTM-based deep network to achieve this and then iterate over these two steps, each time increasing the complexity of the question and the diversity of the pictures considered.

References

Hinton, Geoffrey E. and Osindero, Simon and Teh, Yee Whye. 2006. *A fast learning algorithm for deep belief nets*, Neural computation.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Mar-

garet Mitchell, Dhruv Batra, C. Lawrence Zitnick and Devi Parikh 2015. *VQA: Visual Question Answering*, CoRR.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang and Wei Xu 2015. *Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering*, CoRR.

Mengye Ren and Ryan Kiros and Richard S. Zemel 2015. *Image Question Answering: A Visual Semantic Embedding Model and a New Dataset*, CoRR.

Francis Ferraro, Nasrin Mostafazadeh, Ting-Hao Huang, Lucy Vanderwende, Jacob Devlin, Michel Galley and Margaret Mitchell 2015. *On Available Corpora for Empirical Methods in Vision & Language*, CoRR.