Image Caption Generator

Sanjana Ramprasad(50170374)
Nikhil Shekhar(50169106)

Problem Description:

The aim of this project is to generate captions for the given images in English. Being able to automatically describe the content of an image using properly formed English sentences is a very challenging task. This task is significantly harder, for example, than the well-studied image classification or object recognition tasks, which have been a main focus in the computer vision community. Indeed, a description must capture not only the objects contained in an image, but it also must express how these objects relate to each other as well as their attributes and the activities they are involved in. Moreover, the above semantic knowledge has to be expressed in a natural language like English, which means that a language model is needed in addition to visual understanding.

Data:

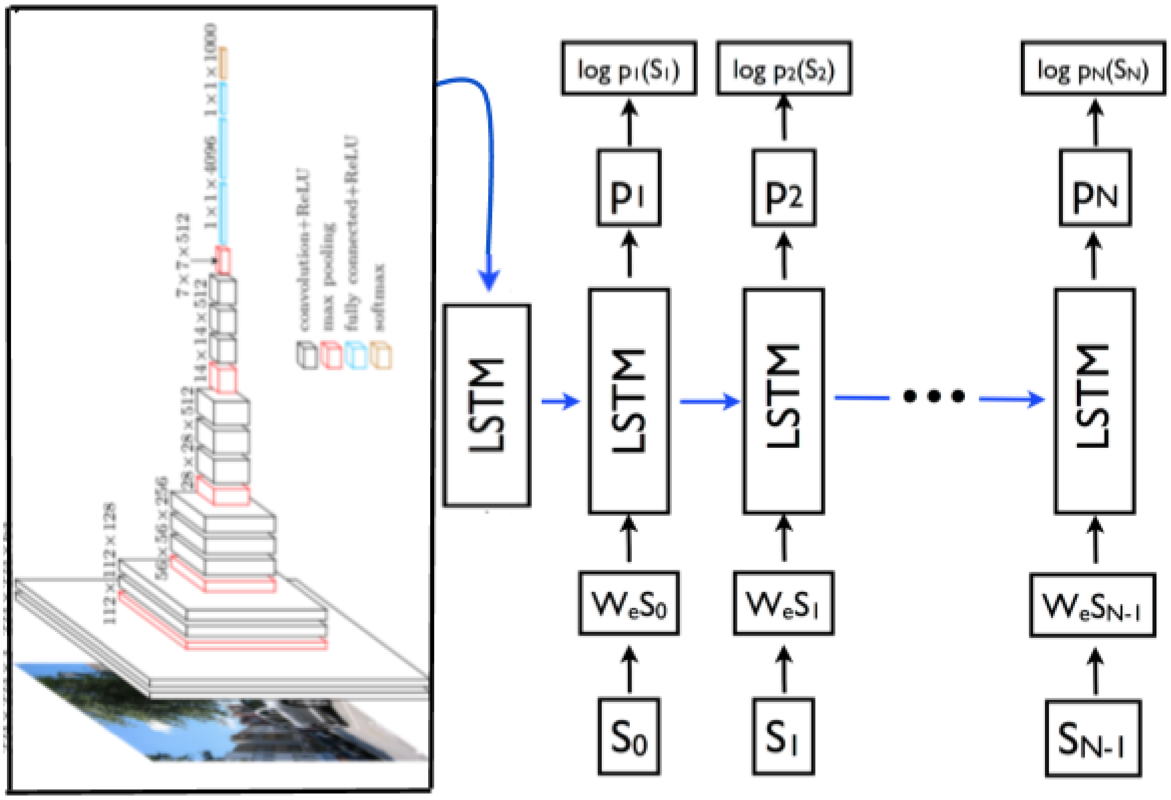
The project will make use of the Flickr 30k datasets.The data has 30K images and 150K descriptive captions which will be used for training the CNN and LSTM network. This is a relatively small dataset compared to ImageNet to be used for training a CNN on, but given the compute power and timelines, this dataset seems practical. The data can be accesses through the links below.
http://shannon.cs.illinois.edu/DenotationGraph/
http://web.engr.illinois.edu/~bplumme2/Flickr30kEntities/

Approach:

The idea is to implement a deep recurrent architecture that automatically produces short descriptions of images. Our model comprises of:

- A CNN - which was pre-trained on VGG Network, to obtain images features.
- PCA - To reduce the computational overhead of the network, we intend to reduce the dimensionality of the output of the CNN by using PCA before passing it to the LSTM model.
- RNN/LSTM - We then feed the feature outputs from the CNN features into either a vanilla RNN or a LSTM network to generate a description of the image in valid English.

Architecture:



VGG Net

LSTM layers

Related Work:

Great progress in image classification has been made over the last couple of years, especially with the use of deep learning techniques. Generating sentences that describe the content of images has already been explored. Several works attempt to solve this task by finding the image in the training set that is most similar to the test image and then returning the caption associated with the test image. Links to the few of the works which we will be using for our implementation is listed below:

- We will be making use of a layer from the pre-trained model of VGG net. The paper regarding the same can be accessed here - https://arxiv.org/pdf/1409.1556.pdf
- The paper from google, on image captioning gave us the idea about our architecture. The same can be accessed here - https://arxiv.org/abs/1609.06647
- An implementation of image captioning in tensorflow based on the Inception model was released a month ago which can be accessed through the link below https://github.com/tensorflow/models/tree/master/im2txt