Image Caption Generator

Shree Murthy, Dylan Inafuku, and Rahul Sura

2023-11-06

Table of Contents

1. Introduction

For our CPSC 393 Final Project, our group is planning on creating an Image Caption Generator built using CNN and LSTM architectures. The goal of this project is to create a model that can take in an image and generate a caption that describes the image. The dataset used for this proposal is the Flickr8k dataset. This dataset contains 8000 images that are paired with 5 different and unique captions that describe its respective image. For example, below are a few lines derived from the `captions.txt` file:

Image	Caption
1000268201_693b08cb0e.jpg	A child in a pink dress is climbing up a set of stairs in an entry way .
1000268201_693b08cb0e.jpg	A girl going into a wooden building .

2. Goal of the model

The model should be able to create accurate and readable (i.e a sentence that represents the image provided). The caption should be roughly 35 characters long (subject to change).

A successful model should have a high level readability and should mimic near-human performance when describing an image.

3. Architecture/Techniques

As of right now we are planning on using a CNN and LSTM architecture. The CNN will be used to extract features from the image and the LSTM will be used to generate the caption. There will be two components: an Image Encoder and Caption Generator.Below is a Picture of the Architecture:

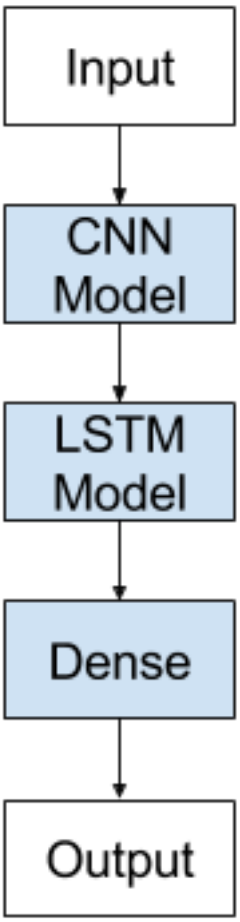

Figure 1: Architecture - Credit to Clairvoyant

3.1 Image Encoder

The Image Encoder can leverage a pre-trained CNN like VGG16 or ResNet to extract the main features/key topics of the images. We will have to remove the fully connected layers of the pre-trained model and use the output features from a self-created layer as the image representation. The output is will be a vector of the main features that will be fed into the LSTM for caption generation (i.e the decoder).

3.2 Caption Generator (Decoder)

The Caption Generator is an LSTM model that takes the image representation as its input and generates a caption. To train the LSTM's vocabulary we are planning on using Word2Vec

because of it's large vocabulary and the ability to create word embeddings. Furthermore, the captions from the dataset will be preprocessed and added ontop of the Word2Vec to create a robust vocabulary that the LSTM can leverage.

3.3 Measuring Statistics

To evaluate if the model is performing well we would want to compare the candidate, predicted, and reference, actual, captions. To measure this we can leverage one of three metrics: BLEU, METEOR, and ROGUE. These metrics are used to measure the similarity between two sentences. The higher the score the more similar the sentences are.

4. Is Deep Learning Required?

Deep learning is indeed required for this project, the dataset requires us to use Deep Learning to learn features from the images and to automatically generate captions. There are no basic models (i.e KNN, SVM, etc.) that can be used to generate captions. Furthermore, this is an intensive problem so using Deep Learning, especially transfer learning models, can help us save time and resources when training the model. Moreover, it will create a more accurate model that can learn image features and generate captions.

6. Ethical Concerns

There are no ethical concerns with this project. The dataset is a public dataset that is used for research purposes. Furthermore, the dataset does not contain any sensitive information and is not biased. The dataset has been used by others previously for various projects.

7. References

- Flickr8k Dataset
- Measuring Captions Generated
- Technical Report on Image Caption Generator
- Image Reference