Group 6 - Project Proposal
DATS 6203(10)
Machine Learning II

By: Hemanth K., Madhuri Y., Sharmin K.

Problem Statement

This project aims to perform Image Captioning using Deep Learning techniques. It is the process of generating textual descriptions (words, sentences, or paragraphs) of images and falls under the domain of Computer Vision and utilizes Deep Learning and Natural Language Processing techniques and algorithms. The applications include:

- ❏ Image Search in Search Engines
- ❏ Image Segregation and Classification
- ❏ Text-to-Speech to aid visually impaired individuals
- ❏ Automatic Image annotation in Facial Recognition, E-commerce, etc.

Dataset

The data used for this project is the Flickr30k dataset. It is a standard dataset to generate sentence captions for image description. There are about 32,000 images, hence the data is large enough to be trained using deep neural networks.
Link - [Flickr30k Image Dataset]

Network Architecture & Framework

We are planning to use the bottom-up approach mentioned in reference [1] for the project. The bottom-up approach generates items observed in an image and then attempts to combine the items identified into a caption. CNN-LSTM model can be used for this approach.

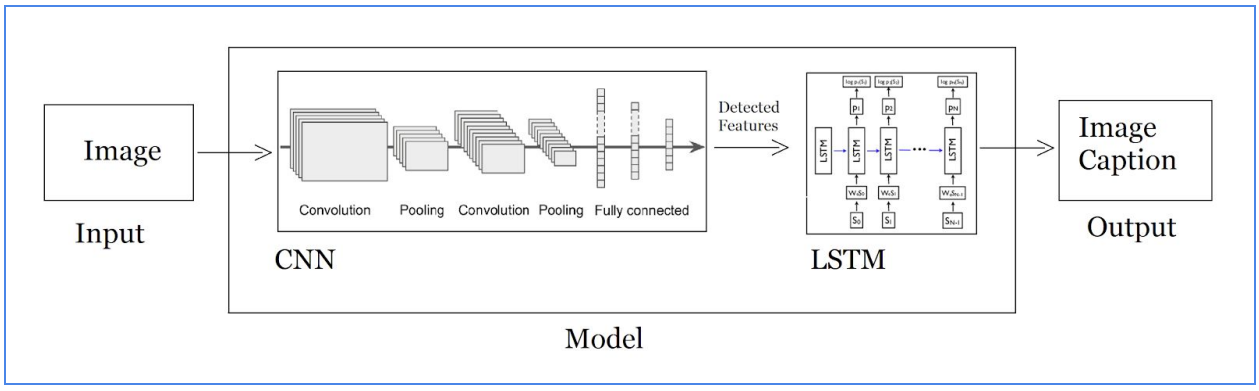


Figure 1: Model-Architecture

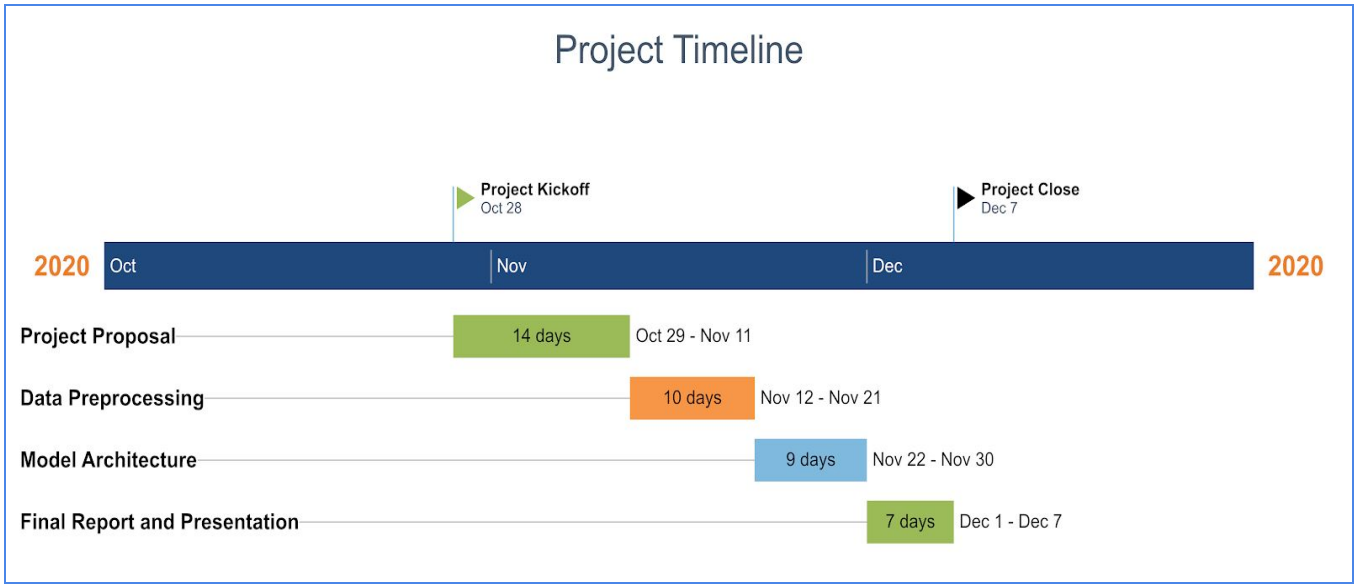
As shown in figure 1 The deep convolutional neural network generates a vectorized representation of an image that can be fed into a Long-Short-Term Memory (LSTM) network, which then generates captions.

To achieve the proposed model we plan to use OpenCV, Keras, and TensorFlow frameworks. We also plan to experiment with Pytorch (deep learning framework).

Performance Metrics

Since **BLEU** is considered one of the first metrics and considering its consistent popularity we plan to measure the performance of our model using BLEU. We also plan to report our score on recently devised metrics **METEOR (**unigram matching based on their surface forms**)** and **SPICE (**includes objects, attributes, and relations in the candidate caption for a better score**)**

Schedule



References

1. Learning CNN-LSTM Architectures for Image Caption Generation
2. Image Captioning
3. Show and Tell: A Neural Image Caption Generator
4. A Hierarchical Approach for Generating Descriptive Image Paragraphs
5. VizSeq: A Visual Analysis Toolkit for Text Generation Tasks
6. Recurrent Neural Network Regularization
7. Image Captioning - A Deep Learning Approach

The following references have been used for network architecture and definitions:

1. https://hagan.okstate.edu/NNDesign.pdf
2. https://kharshit.github.io/blog/2019/01/11/image-captioning-using-encoder-decoder
3. https://towardsdatascience.com/image-captioning-in-deep-learning-9cd23fb4d8d2