# Visual Recognition : Image Captioning

Project Report

Hrithik Sharma (IMT2018029) hrithik.sharma@iiitb.ac.in
Prajwal Agarwal (IMT2018056) prajwal.agarwal@iiitb.ac.in
Sahaj Vaghasiya (IMT2018060) sahaj.kuman@iiitb.ac.in
Team Name : HPS

## 1 Abstract

In this project, we experimented with CNN LSTM model, typically a hybrid neural network model that uses less memory. . Our end goal was to create a model that can caption images with a high accuracy. We implemented our own model to caption the images. We were expected to submit a report about the approach, methods we tried to implement and the algorithms used. This is the final report of our team for this project.

## 2 Introduction

Image Captioning is the process of generating a textual description for given images. It has been a very important and fundamental task in the Deep Learning domain. Image captioning has a huge amount of application. Image captioning can be regarded as an end-to-end sequence to sequence problem, as it converts images, which is regarded as a sequence of pixels to a sequence of words. For this purpose, we need to process both the language or statements and the images. To implement image captioning successfully we use a special class of hybrid neural networks : CNN LSTM. These are a class of models that are both spatially and temporally deep and sit at the boundary of Computer Vision and Natural Language Processing. Image Captioning using neural networks involves a generic architecture as shown below.

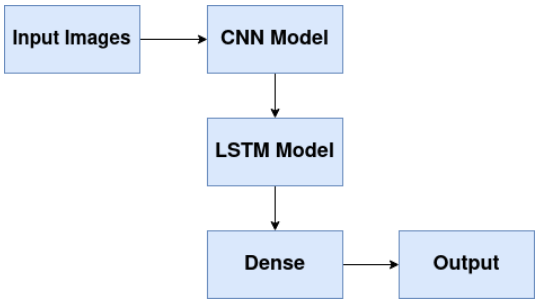



Figure 1. Generic Model Architecture - CNN LSTM

## 3 Problem Statement

Given a set of images, our goal is to extract the features of the images and produce a suitable caption for the image. The feature extraction carried out on the image uses a pre-trained Convolution Neural Network and the LSTM (Long Short Term Memory) system to generate the captions based on extracted features. We had to design a combination of the Convolutional Neural Network and the LSTM system for an automated captioning of images.

## 4 About the Dataset

**Dataset** : Flickr8K (Provided with problem statement)

**Image :**
**Dataset Size :** 8092 images (6000 : training, 1000 : test, 1000 : development)
**Description :** Different images in JPEG format of covering a wide range of situations and scenarios.

**Text :**
**Dataset Size :** 5 text files.
**Description :** Number of files containing different sources of descriptions for the photographs. Each image is provided with 5 unique descriptions.

## 5 System Modules

### 5.1 Feature Extraction

Feature Extraction is a Neural Network model that takes an image as input and outputs the features of the images in the form of fixed length vector. A Convolution Neural Network is used as the feature extraction model. We tried the following pre-trained models:

#### 5.1.1 **VGG16** a pre-trained convolutional neural network model takes in (224, 224) RGB images and converts them into features.

#### 5.1.2 **Inception** Version 3 a widely used image recognition model for feature extraction.

**Implementation :**
We have used the pre-trained VGG16 for feature extraction. As the model is pre-trained we just have to execute the model for each image and store the features.

### 5.2 Language Model

For generating captions for images a LSTM (Long short term memory) system based model is used. The model generates a caption sequence of words for any given image based on the features generated by the convolutional neural network model (VGG16).
**Implementation :**
The Model trains on the already processed features generated by the CNN model, descriptions provided with the the data set and the Vocabulary of words generated from the descriptions. The description provided with the dataset first needs to be processed to be used in the model. The loaded descriptions are all converted to lower case and all punctuation, all single letter words and words with numbers are removed. Once the descriptions are cleaned the vocabulary for the model can be defined. The vocabulary is defined by the words in the descriptions. The vocabulary is then reduced by placing a restriction on the frequency as a model with reduced frequency provides better results. Once all the requirements are complete the model is trained for 10 epochs.

## 6 Problems and Issues

- **Training time**
  Training time significantly increase as we tried to make the model more complex.
- **Vocabulary**
  Size of vocabulary seems to affect the accuracy of model alot. One of the problems that we faced was that if the size of vocabulary is large, then the accuracy of model is very low. However if we size of vocabulary is taken to be less then it shows increase in accuracy of the model.

## 7 Model Specification

- **Total Vocabulary Size** : 8763
- **Used Vocabulary Size** : 3472
- **Description Length** : 34
- **Total parameters** : 3,421,072
- **Trainable parameters** : 3,421,072

Below is the representation of our model.

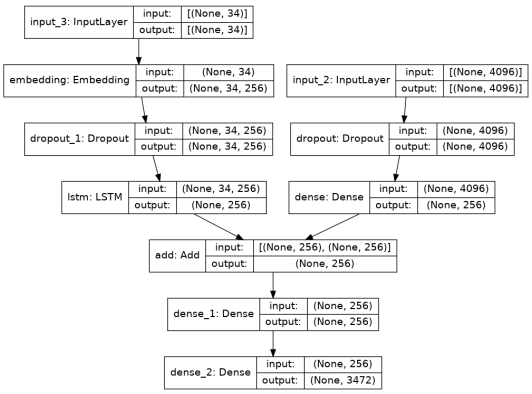



Figure 2. Model Representation

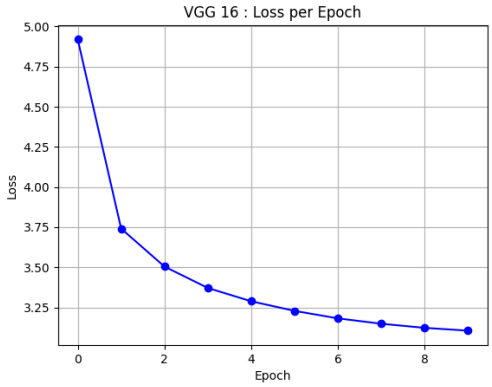
## 8 Result

**BLEU Scores :**

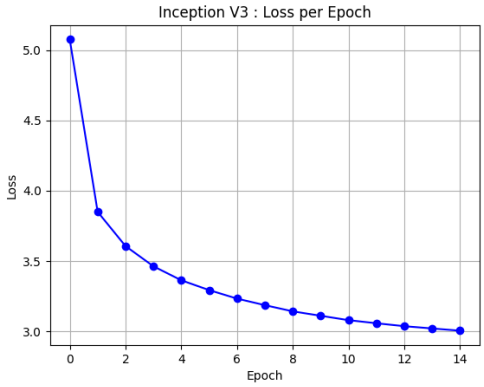
- **BLEU-1**: 0.540916
- **BLEU-2**: 0.284391
- **BLEU-3**: 0.190238
- **BLEU-4**: 0.084603

**Loss Graphs :**

- VGG16




- Inception v3




**Captions:**

**Image 1**

**Caption :** man in red shirt is sitting on the street

**Image 2**

**Caption :** man in black shirt and black pants is standing on the street

**Image 3**

**Caption :** man in black shirt and hat stands in front of crowd

**Image 4**

**Caption :** man in red shirt is standing on the street

**Image 5**

**Caption :** two children are playing in the grass

## 9 Conclusion

This is our first project, where we were able to combine Computer Vision and Natural Language Processing together using memory efficient models. Overall this project proved to be a good learning task for us. We learnt different algorithms involved in image captioning in detail. We learnt how deep learning especially LSTM models are used in image captioning.

## References


1) Wikipedia
2) LSTM
3) CNN LSTM
4) Transfer Learning