ATAL BIHARI VAJPAYEE INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, GWALIOR

MINI PROJECT

Neural Image Caption Generator

Presentation

Submitted To -
Dr Somesh Kumar

Submitted by -
Amit Kumar (2017-IMT-009)
Ankit Meena(2017-IMT-012)
Vishal Jain(2017-IMT-090)
Vishal Shivhare(2017-IMT-091)

Outline

- Introduction
- Objectives
- Components Used / Type of Coding
- Methodology/Circuit Diagram
- Importance for Society/Applications
- Progress So Far / Time Framework
- References

Introduction

Well some of you might say “A white dog in a grassy area”, some may say “White dog with brown spots” and yet some others might say “A dog on grass and some pink flowers”. Definitely all of these captions are relevant for this image and there may be some others also. But the point I want to make is; it’s so easy for us, as human beings, to just have a glance at a picture and describe it in an appropriate language. Even a 5 year old could do this with utmost ease. But for a Computer it becomes a tedious task.

Objective

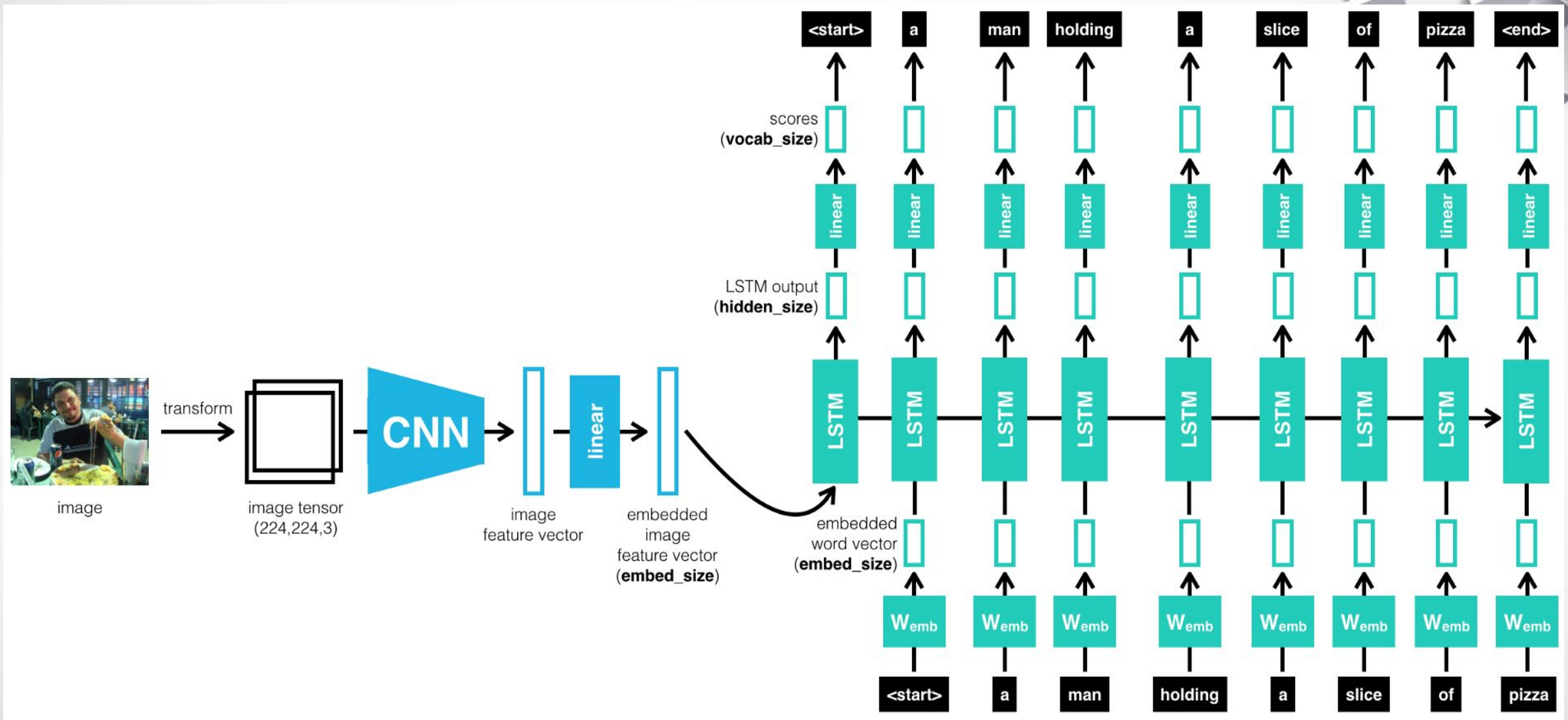
Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. A description must capture not only the objects contained in an image, but it also must express how these objects relate to each other as well as their attributes and the activities they are involved in.

Basically we are trying to build an application which describes the surroundings in an image with some functionalities.

Components Used/Type of Coding

- Deep Learning concepts like
- Multi-layered Perceptrons
- Convolution Neural Networks
- Recurrent Neural Networks
- Transfer Learning
- Backpropagation
- Overfitting
- Probability
- Text Processing
- Python syntax and data structures, Keras library

Methodology/Circuit Diagram



RESNET MODEL ARCHITECTURE

- Very deep networks using residual
- connections
- - 152-layer model for ImageNet
- - ILSVRC'15 classification winner
- (3.57% top 5 error)
- - Swept all classification and
- detection competitions in
- ILSVRC'15 and COCO'15!

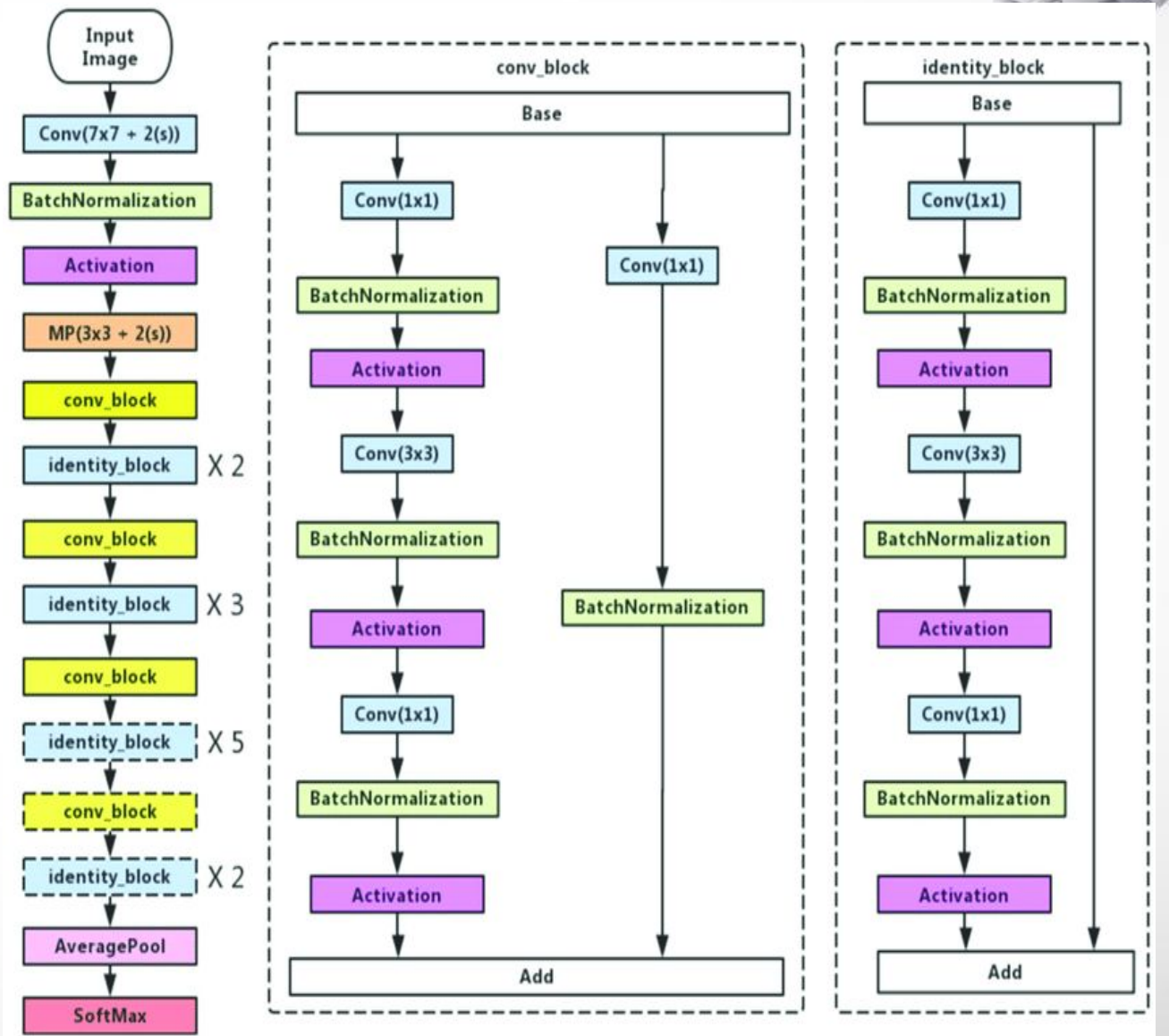
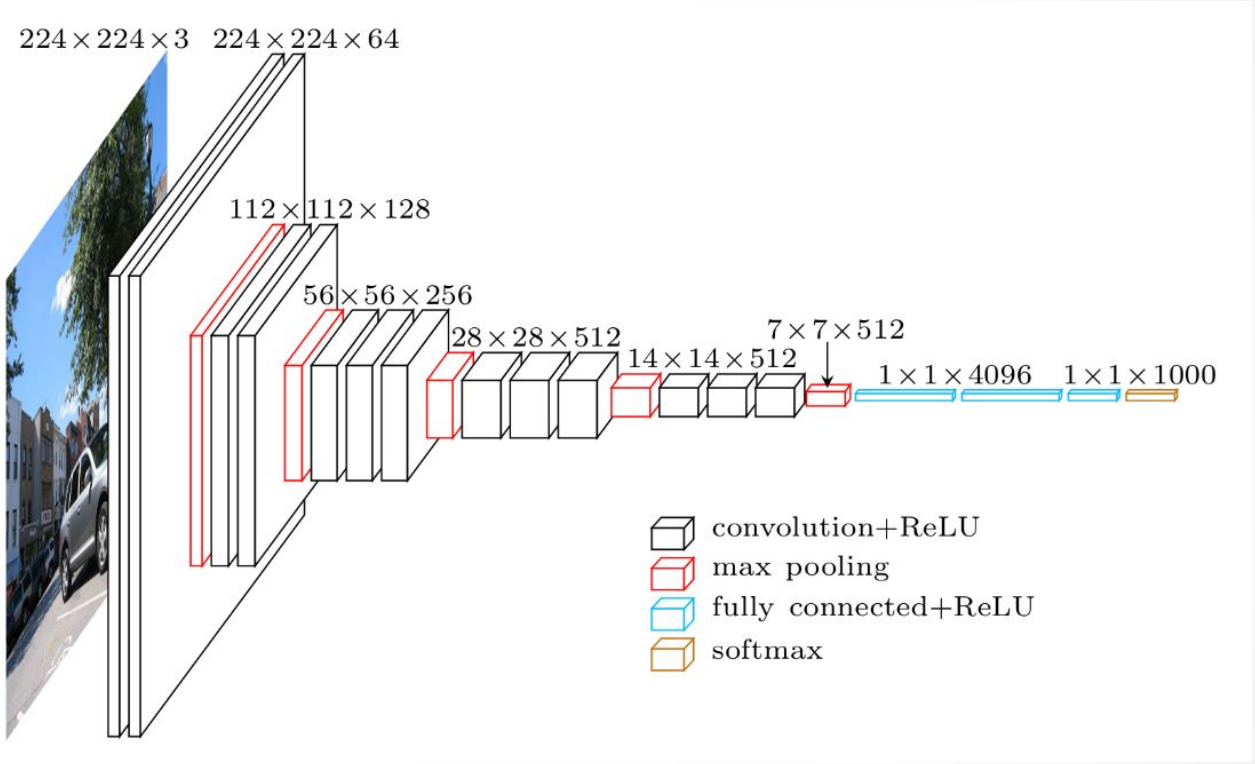
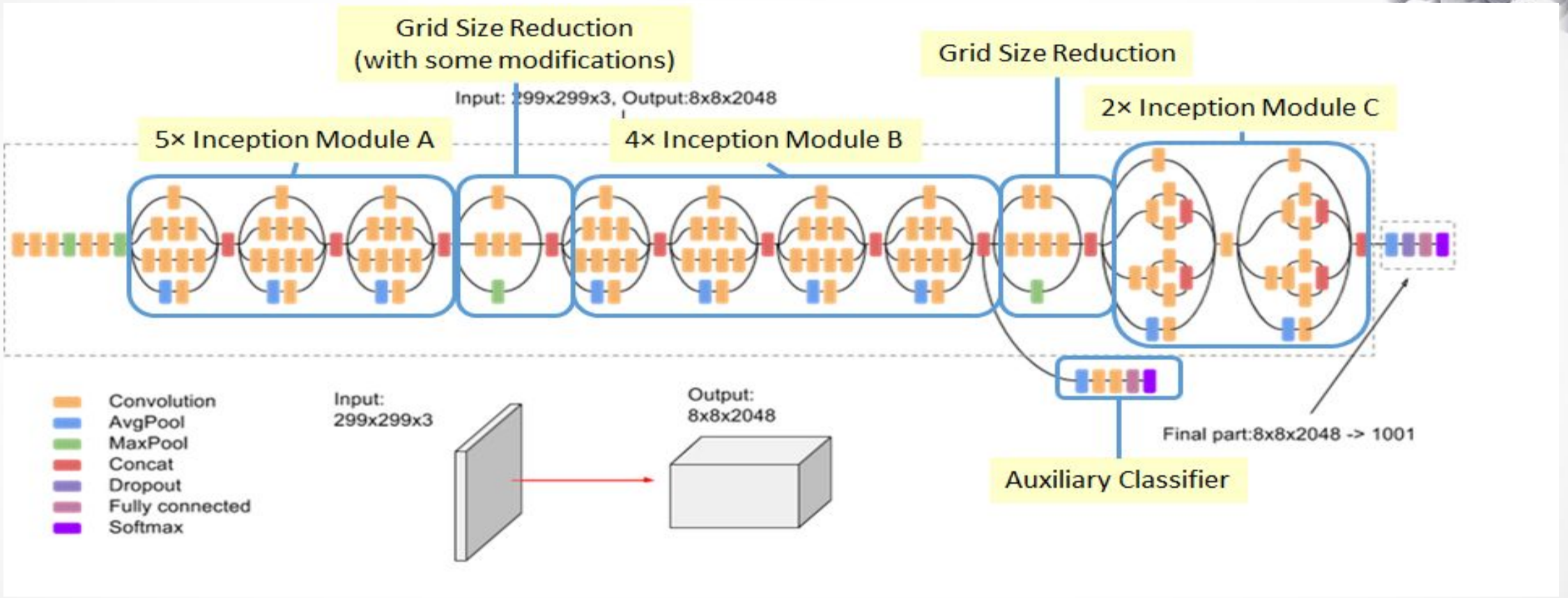


Image feature Extraction using Transfer learning

1. *Resnet model*
2. *Input size : 224x224x3*
3. *Activation vector size: 7x7x512*
4. *res5c_branch2c (Conv2D) : (None, 7, 7, 2048)*
5. *output Gap layer : (None, 2048)*

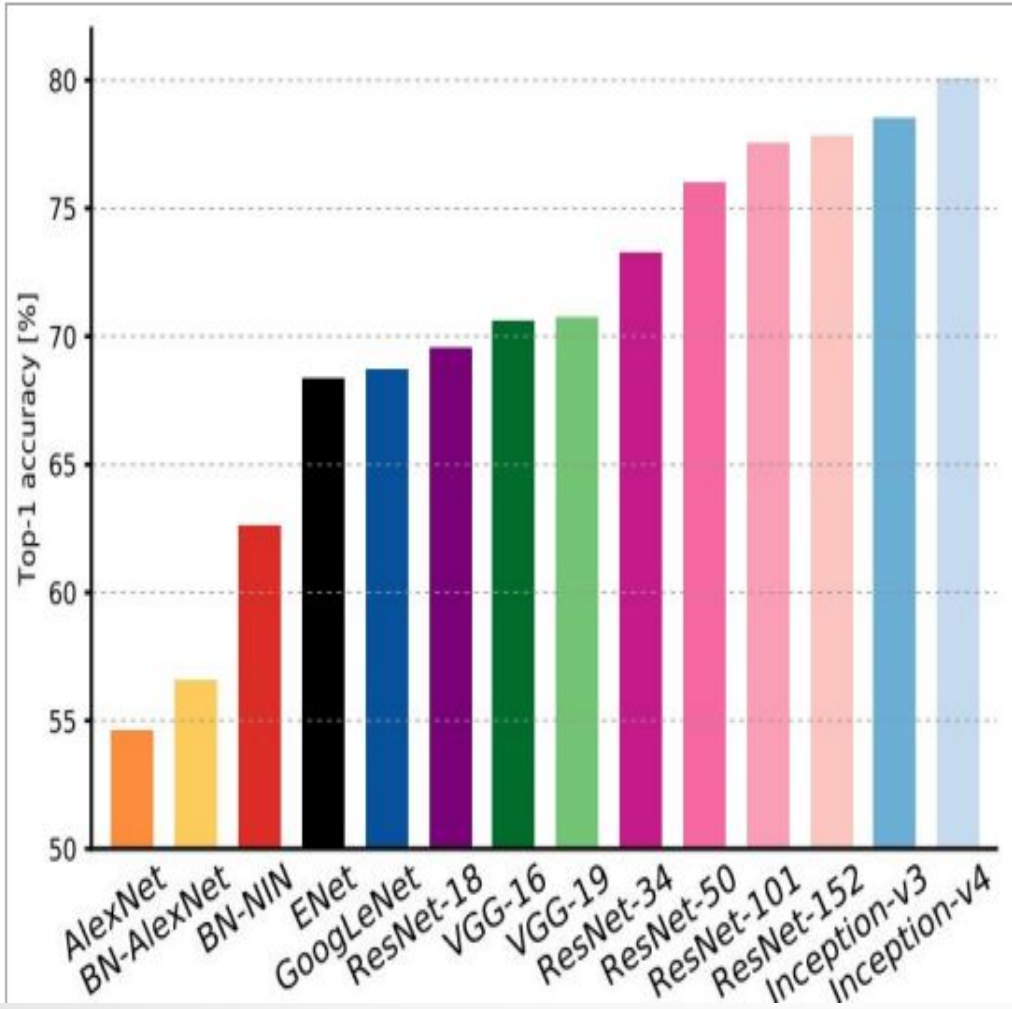


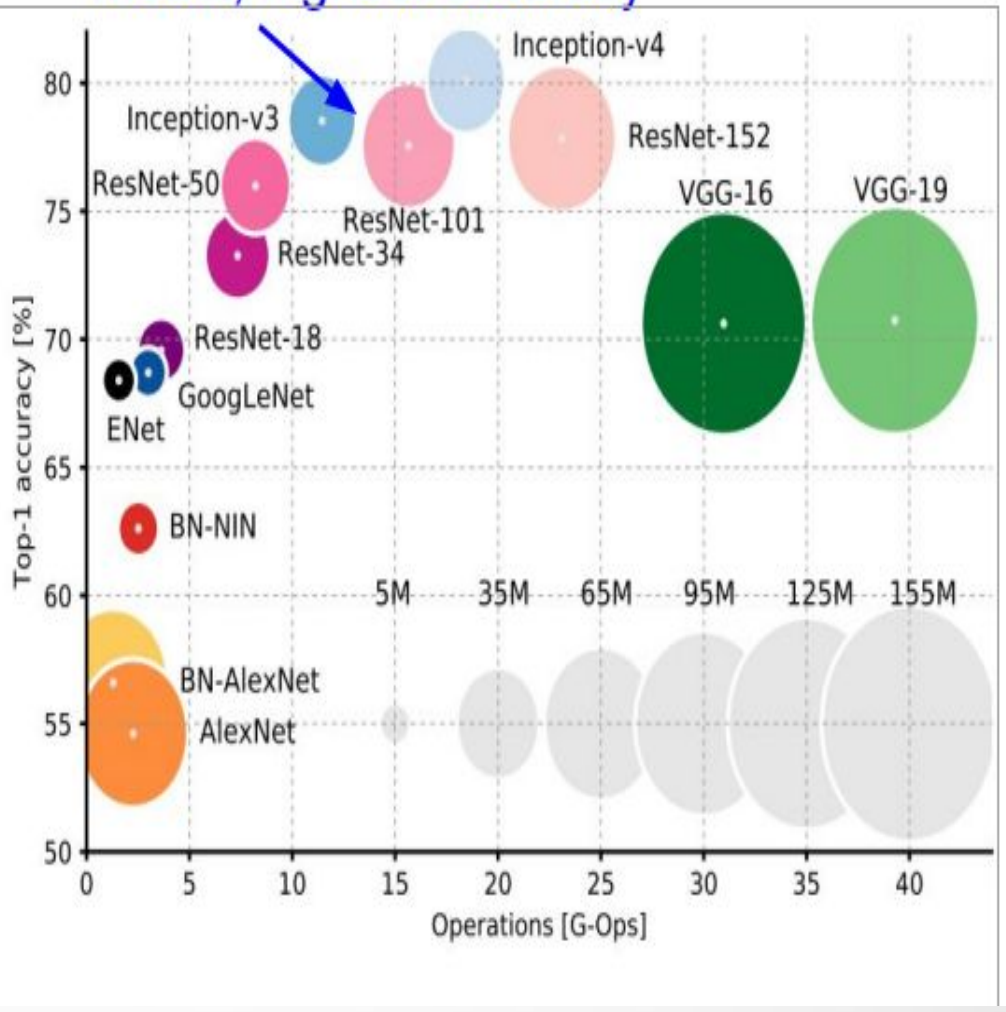
INCEPTION MODEL ARCHITECTURE



Comparing complexity...

ResNet:
Moderate efficiency depending on model, highest accuracy





Inception-v4: Resnet + Inception!

VGG: Highest memory, most operations

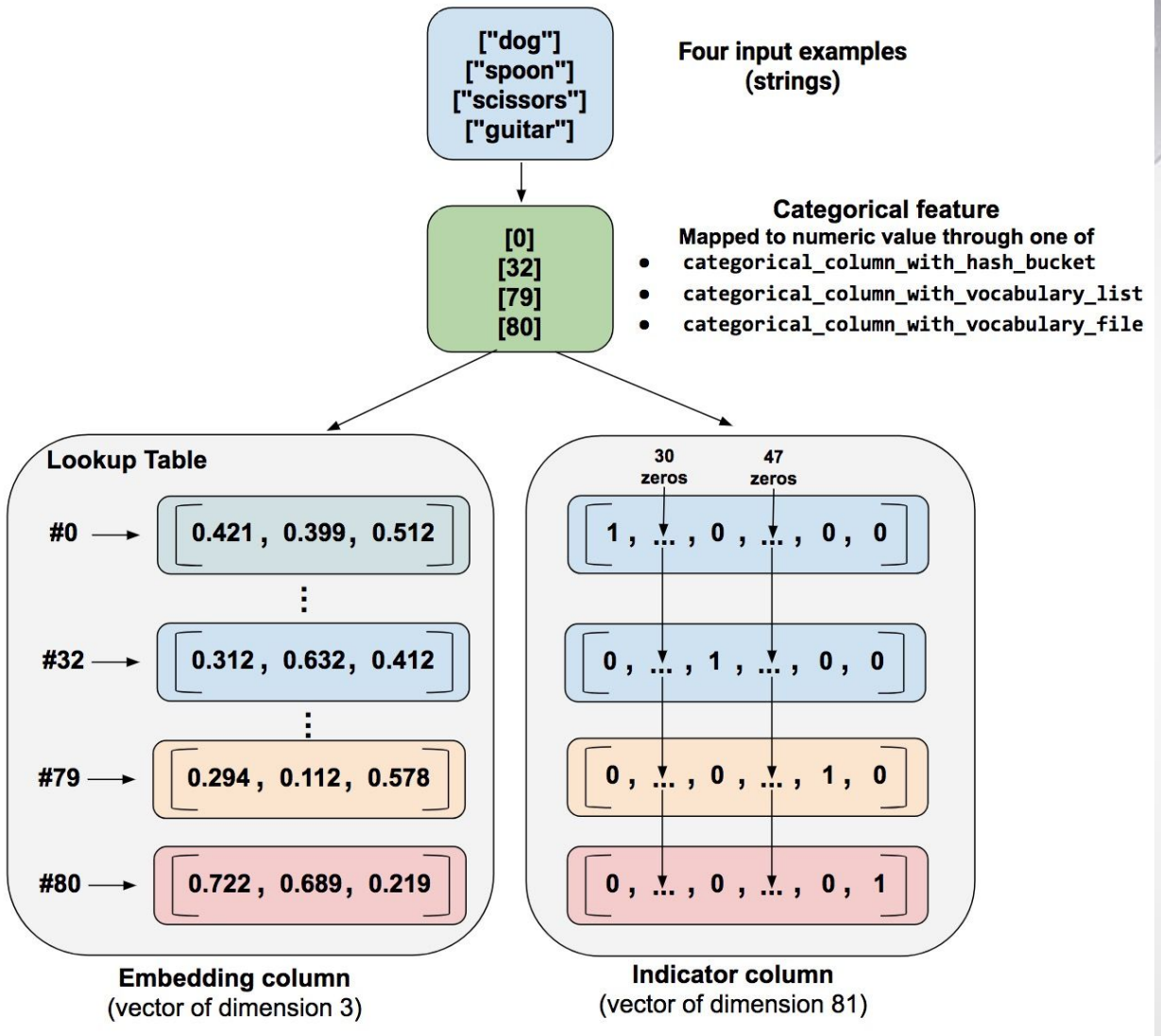
GoogLeNet: most efficient

AlexNet: Smaller compute, still memory heavy, lower accuracy
ResNet:Moderate efficiency depending on model, highest accuracy

Text Feature Engineering using Glove vector

- For a given image (Train) : 5 sentences
- Total Words 373837

- Total unique word (vocab):8424
- Reducing vocab size (Thresholding):1652+2
- glove.6B.50d or any other
- embedding matrix shape: (1654, 200)
- Max length of sentence : 34
- Further using Keras fuctional api
- (34x200)

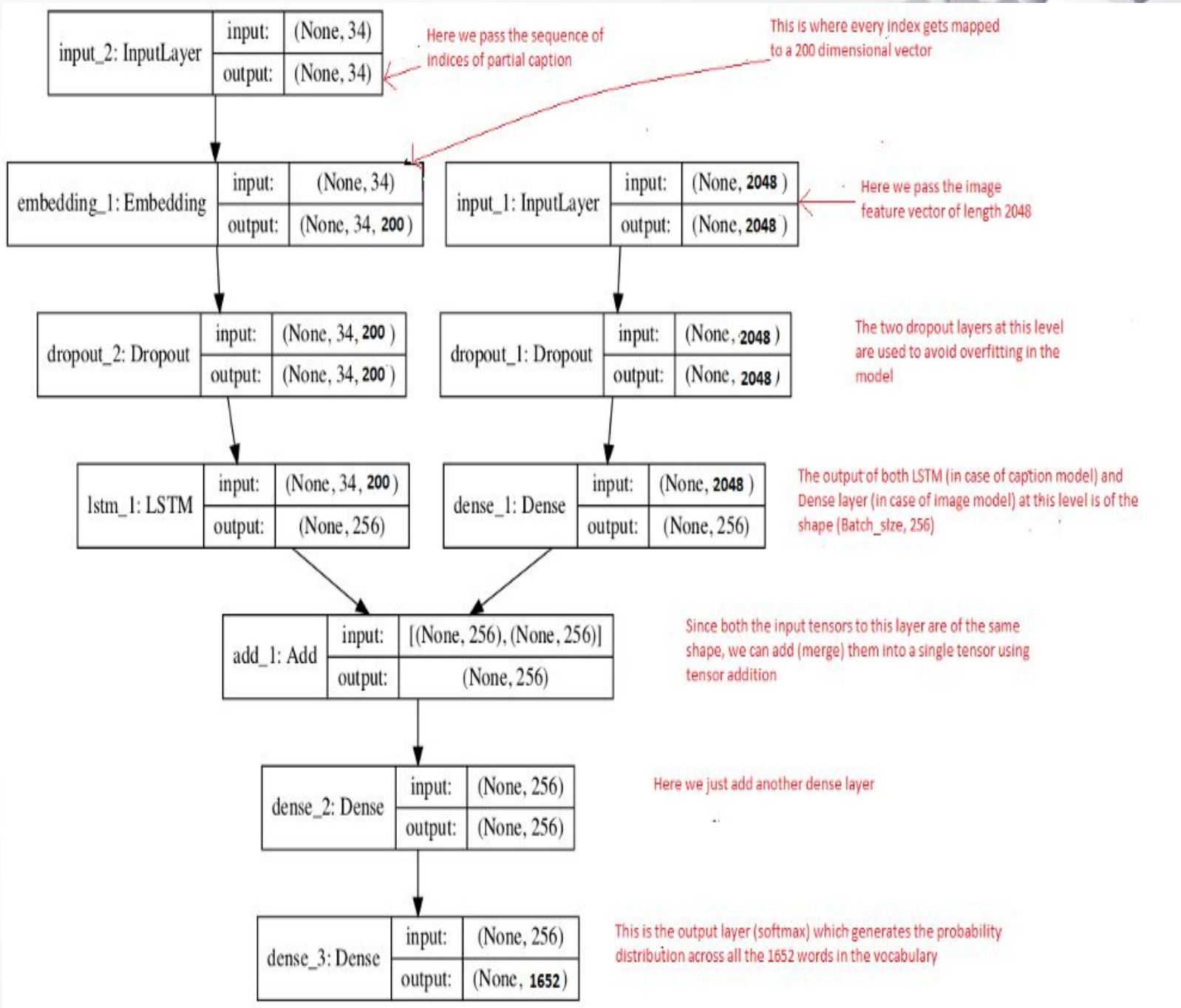


Visualizing the structure of the network and better understand the two streams of input:

- For text
- Total Words: 373837
- Filtering words : threshold = 10
- vocab size : 1652

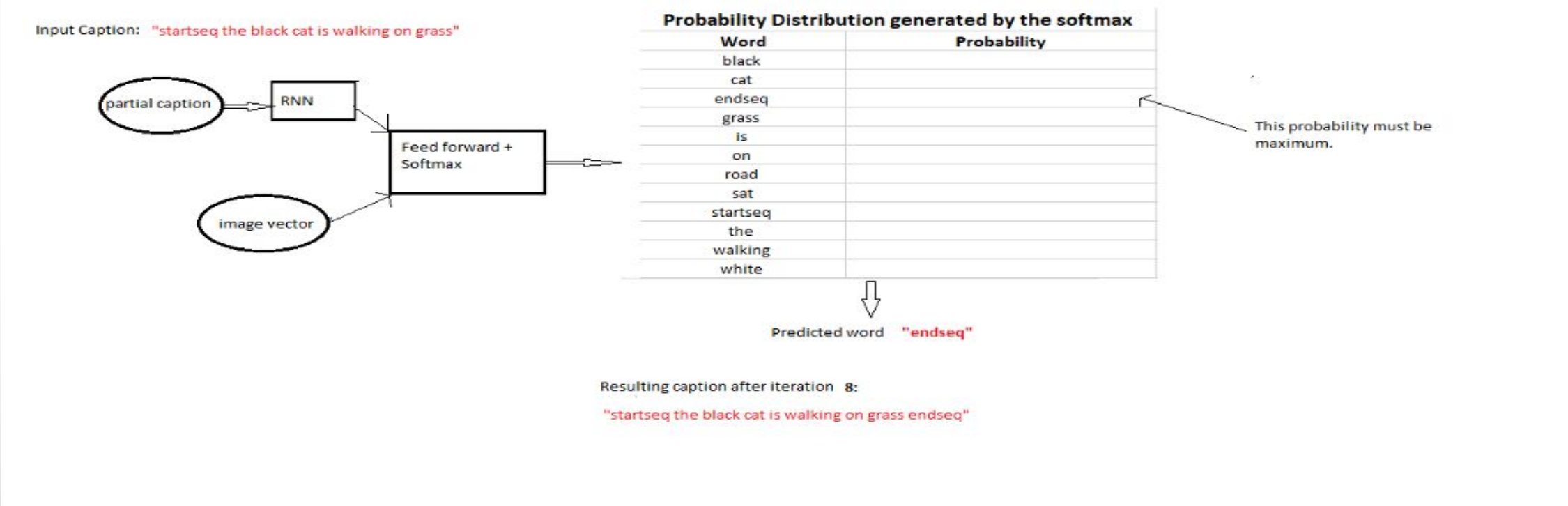
- For image
- Resnet feature extraction
- Input size : 224X224X3
- avg_pool (GlobalAveragePooling2 (None, 2048)

- Maximum Likelihood Estimation (MLE)
- Greedy Sampling



Caption Generation

- Maximum Likelihood Estimation (MLE) i.e. we select that word which is most likely according to the model for the given input. And sometimes this method is also called as Greedy Search, as we greedily select the word with maximum probability.



Importance for Society and expected outcomes/applications

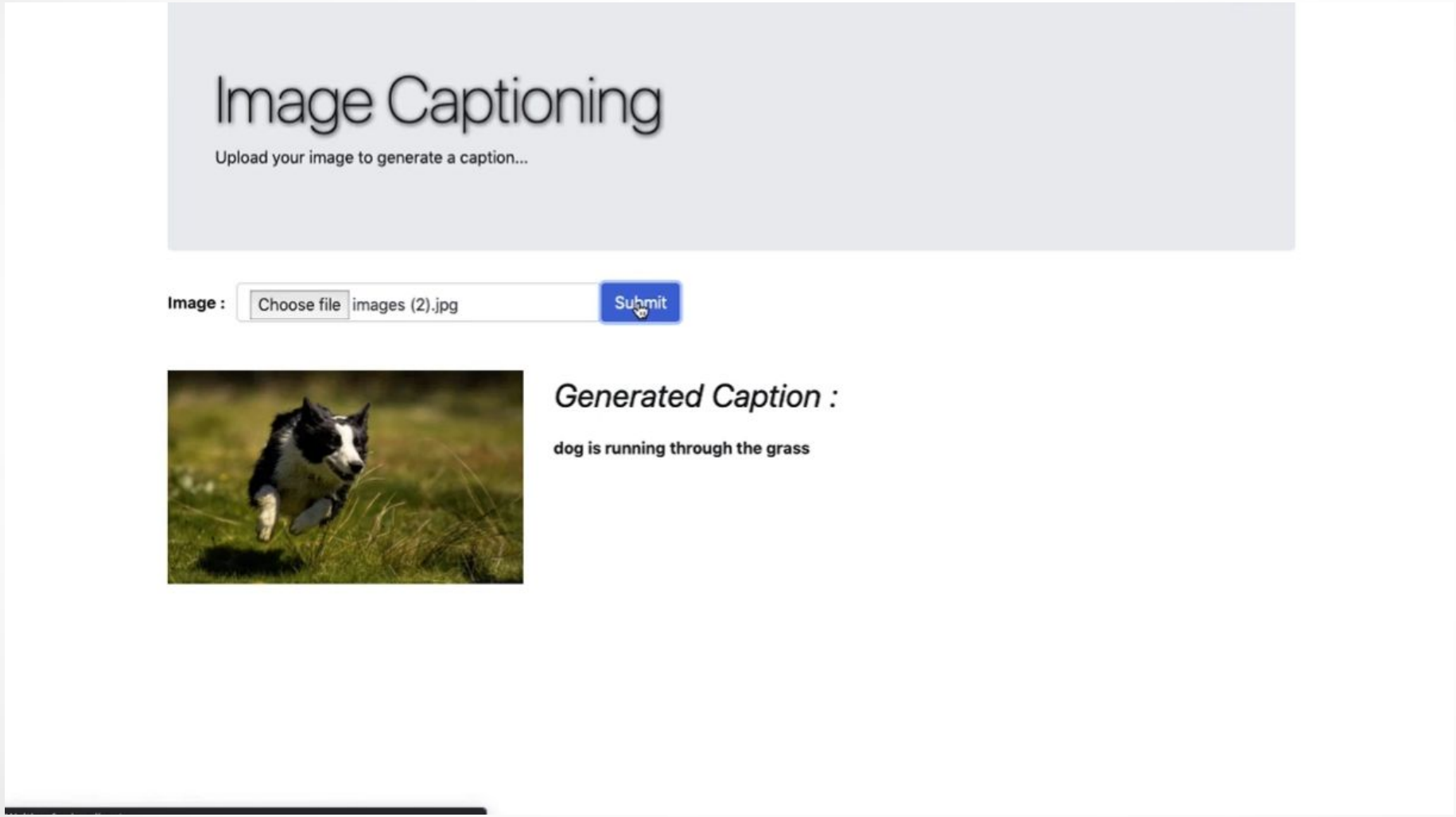
Help the Visually Impaired : This technology can be integrated with a headband or glasses, which a visually impaired person would wear, the headband/glasses are connected to a camera. The camera sends the video stream of the surroundings to the model and then the model predicts the surrounding.This prediction can be heard by the person.

Self Driving Cars : All the self drive cars are using image/video processing with neural network to attain their goal.

Skin Cancer Prediction : This technology can be used in prediction of skin cancer by classifying the tumor as Benign or Malignant.

Classifying images in our Phone: This technology can be used to classify images in our Gallery for example classify in different categories like mountains, beach etc.

DEMO OF WEB APPLICATION



RESULT

In this project we have implemented a web app which describes what is going on in a image by giving it a suitable caption.

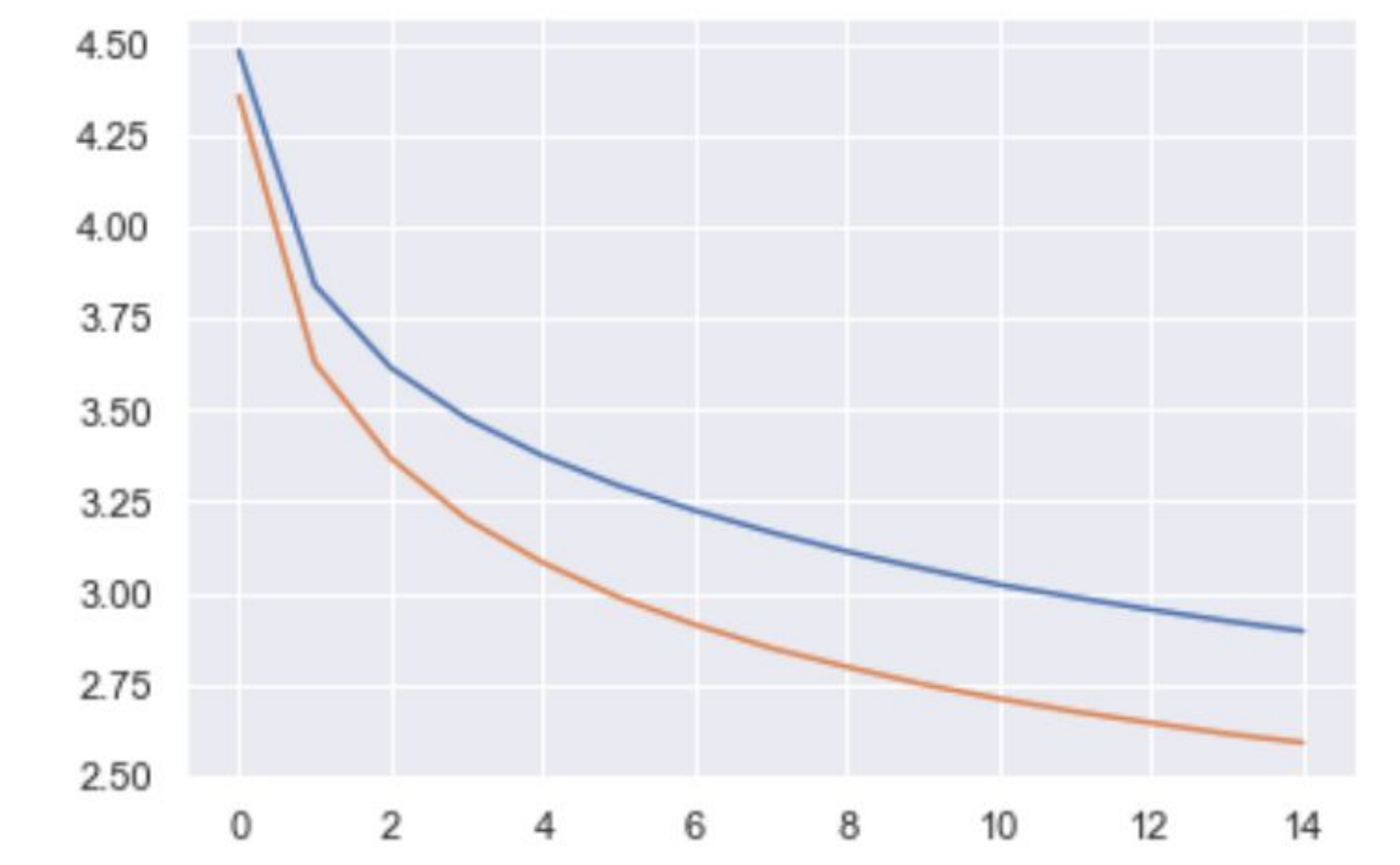
We used to Deep learning models Resnet 50 and Inception.

The evaluation metric was BLEU score.

Average BLEU score of Resnet 50 : **0.731262**
Average BLEU score of Inception : **0.69103**

RESULT

Comparison for 15 epoch of Loss function between Inception and Resnet50



References

- ***Shadab Hussain***, Flickr8K dataset kaggle.

- ***Faizan Shaikh***, Automatic Image Captioning using Deep Learning (CNN and LSTM) in Pytorch, Analytics Vidhya.

- ***Harshall Lamba***, Image Captioning with Keras, Towards Data Science.

Thank You