Image Classification and Captioning using Deep Neural Networks

EECE 5644 Final Project

Aditya Patgaonkar
Alan Jacob
Basil Mir
Kiran Tulsulkar

Problem Statement

- Generating textual description of an image using Deep Neural Network and Natural Language Processing(NLP)
- Most of the already existing image captioning models were implemented using KERAS whereas our model is implemented using PyTorch.
- Few real world examples where solution to this problem can be used include:
    - Generating relevant captions for images taken by CCTV cameras
    - Generate an aide for visually impaired person which will guide them in travelling on roads without the support of someone else
    - Can be used in autonomous vehicles by properly captioning the surrounding of the car

Data Set Description :

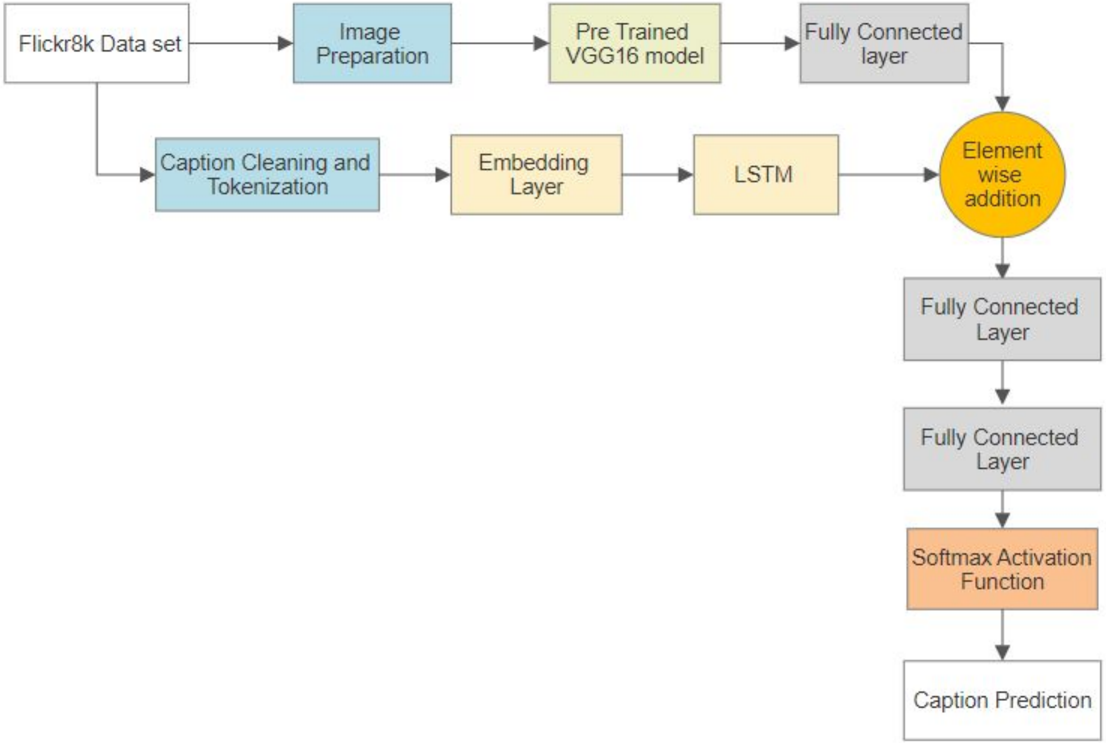
ImageNet : For multi-class classification

- Number of images: 14,197,122
- Number of classes: 1000
- Number of high level categories: 27

Flickr8k : For image captioning

- Number of images: 8091
- Resolution of images: Variable sizes.
- Number of captions per image: 5
- Partitioning of data:
    - Training: 60%
    - Validation:20%
    - Testing :20%

Block Diagram Representation of our Model



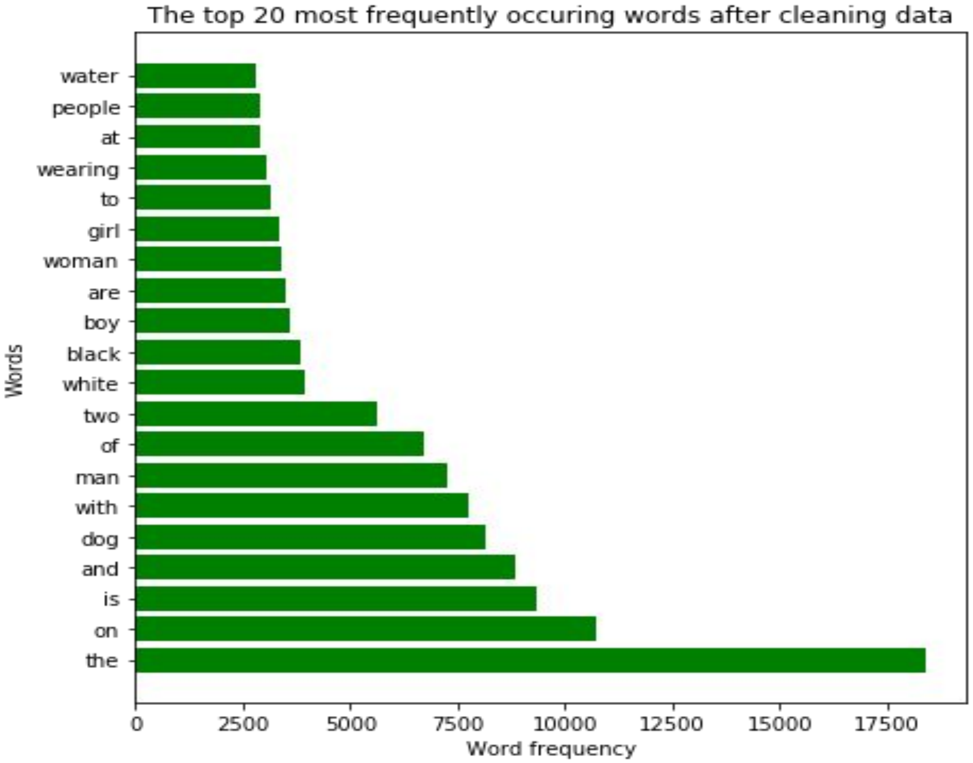
Data Preparation

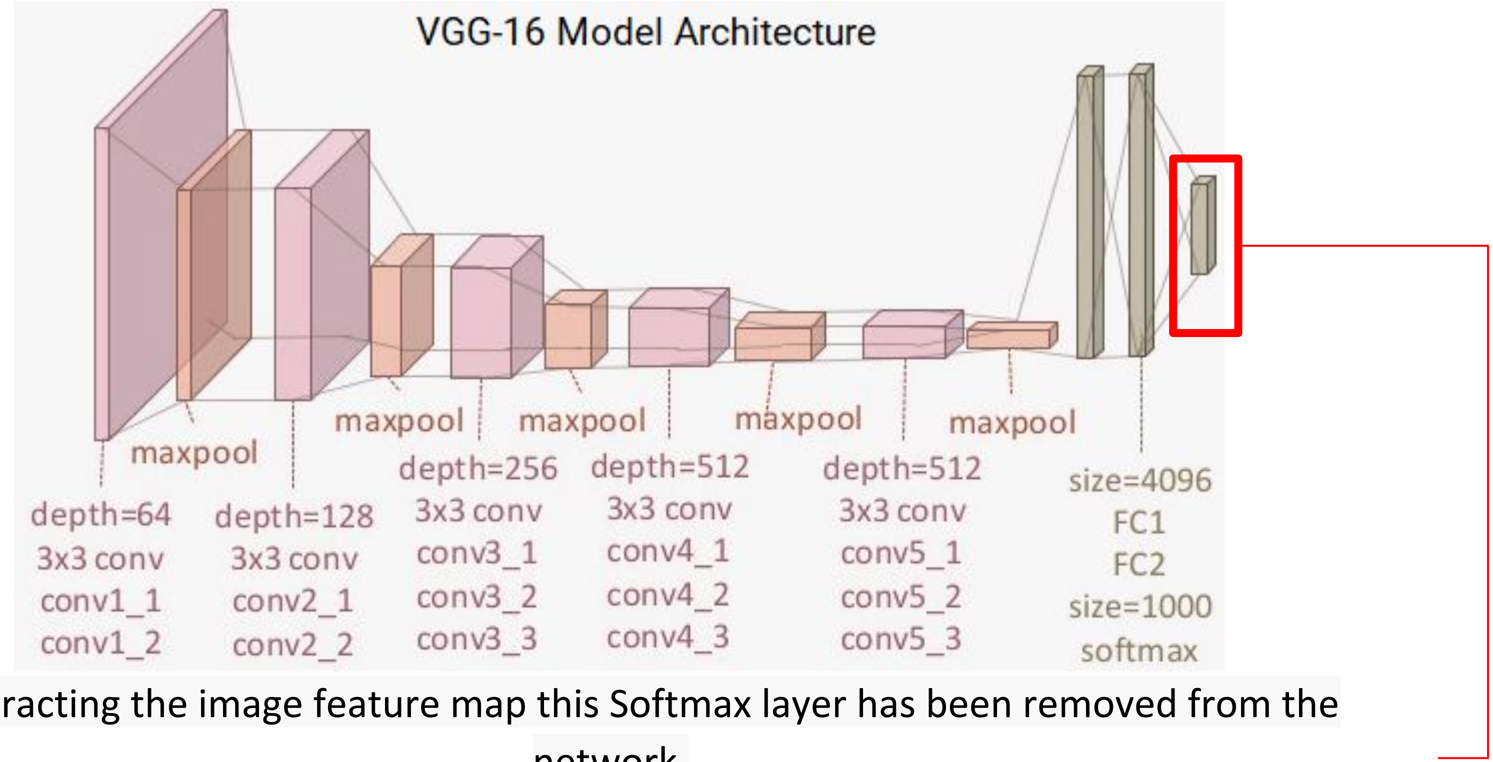
Text Preparation(Captions):

- Create a data-frame containing each word and its frequency in the captions.
- Clean the data- frame by
  1. Removing punctuation
  2. Removing single character
  3. Removing numeric characters
- Add start and end sequence tokens
- Change character vector to integer vector using Tokenizer

Image Preparation:

Images are passed through a pre-trained VGG16 model for feature extraction



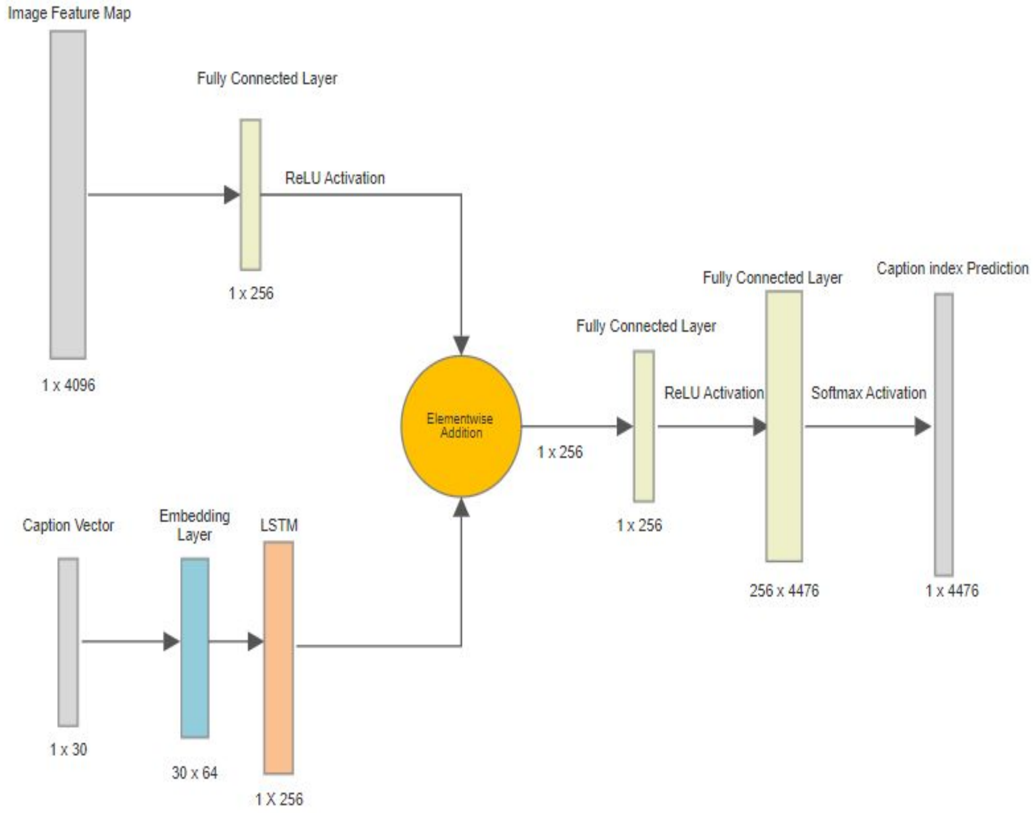


For extracting the image feature map this Softmax layer has been removed from the network.

Image Source:https://towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a

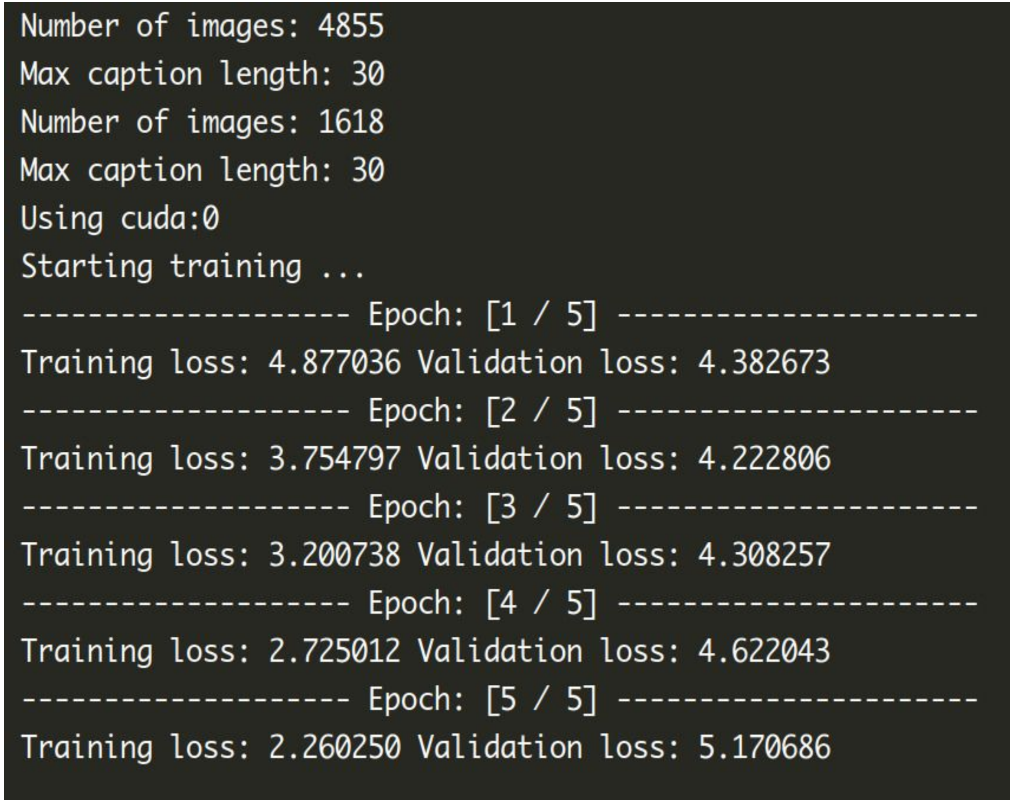
Neural Network Structure:

- Partial caption vector is passed through an embedding and LSTM layer with 256 hidden states.
- The input feature map is transformed non-linearly to a 256 dim vector.
- The 256 dim vectors from caption LSTM and image feature map respectively added elementwise and transformed via 2 fully connected hidden layers to predict the next word of the caption.



Model Training

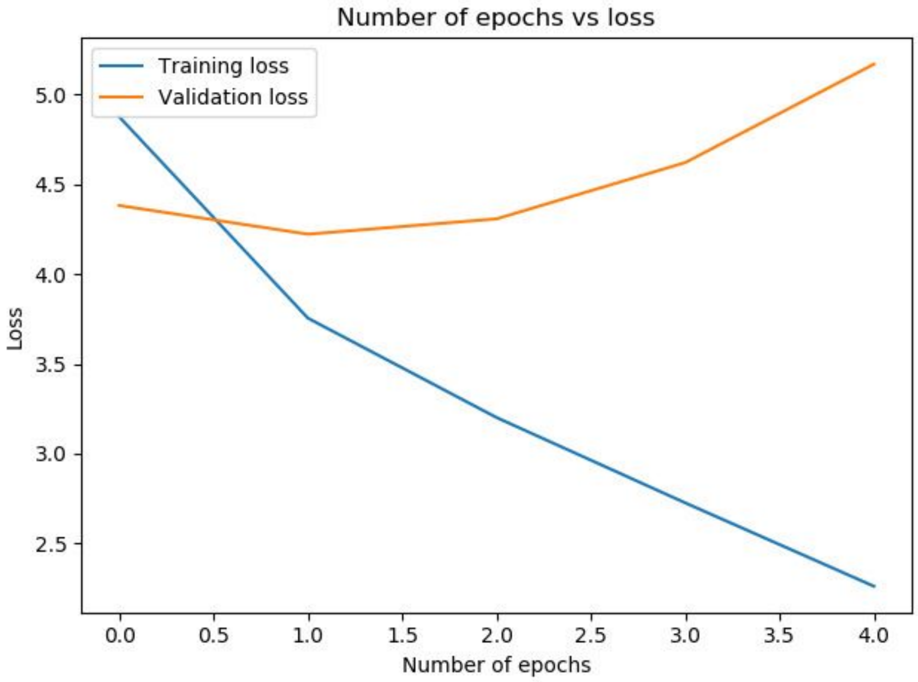
- We train the model over 5 epochs with the following parameters:
  - Learning Rate: 0.001
  - Batch Size : 64
  - Optimizer: Adam
- In the list of Caption Maximum Length of Caption can be 30.
- Split of Images :
  - Total = 8091
  - Training = 4855
  - Validation = 1618
  - Test = 1618



```
Number of images: 4855
Max caption length: 30
Number of images: 1618
Max caption length: 30
Using cuda:0
Starting training ...
-------------------- Epoch: [1 / 5] ---------------------
Training loss: 4.877036 Validation loss: 4.382673
-------------------- Epoch: [2 / 5] ---------------------
Training loss: 3.754797 Validation loss: 4.222806
-------------------- Epoch: [3 / 5] ---------------------
Training loss: 3.200738 Validation loss: 4.308257
-------------------- Epoch: [4 / 5] ---------------------
Training loss: 2.725012 Validation loss: 4.622043
-------------------- Epoch: [5 / 5] ---------------------
Training loss: 2.260250 Validation loss: 5.170686
```

Validation Loss and Training Loss over epochs

As Epoch increases the training loss tends to 0 due to overfitting which can be clearly seen by the increase in Validation Loss after a certain point in the graph.

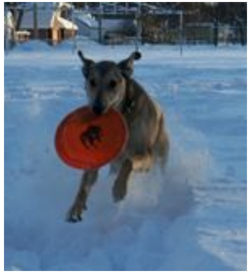


Prediction

Prediction: startseq dog is catching frisbee

Target: startseq dog catching frisbee endseq

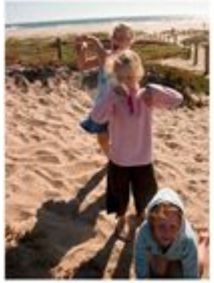
BLEU Score: 0.833



Prediction: startseq boy and girl are playing in the sand endseq

Target: *startseq three children are playing in sand near to the beach endseq*

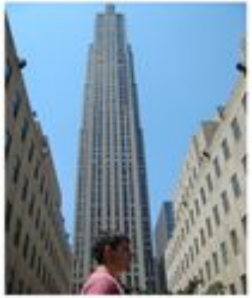
BLEU Score: 0.518



Prediction: startseq man in black jacket is standing on railing endseq

Target: startseq man is standing in front of skyscraper endseq

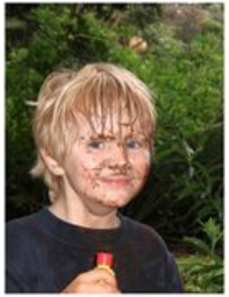
BLEU Score: 0.666



Prediction: startseq boy in blue shirt is playing with his arms crossed endseq

Target: startseq boy in blue shirt with dirt on his face endseq

BLEU Score: 0.666



References

- https://machinelearningmastery.com/develop-a-deep-learning-caption-generation-model-in-python/
- https://fairyonice.github.io/Develop_an_image_captioning_deep_learning_model_using_Flickr_8K_data.html#Visualization-of-the-VGG16-features
- https://github.com/ZhenguoChen/Neural-Network-Image-Captioning
- https://pytorch.org/

Code :

https://github.com/adiRpatgaonkar/MS_ECE/tree/master/EECE5644/Project