Name: Avirat Belekar

CWID: 10454332

Course Name: CS584-A Natural Language Processing

Assignment 4: Convolution Networks

1. Document Classification using Convolution Neural Networks

Problem Statement:

Classify text paragraph into three categories using Convolution Neural Network.

Model:

The neural network model is described in such a manner by which one feature is extracted from one filter. The model uses multiple filters (with varying window sizes) to obtain multiple features. These features form the penultimate layer and are passed to a fully connected SoftMax layer whose output is the probability distribution over labels.

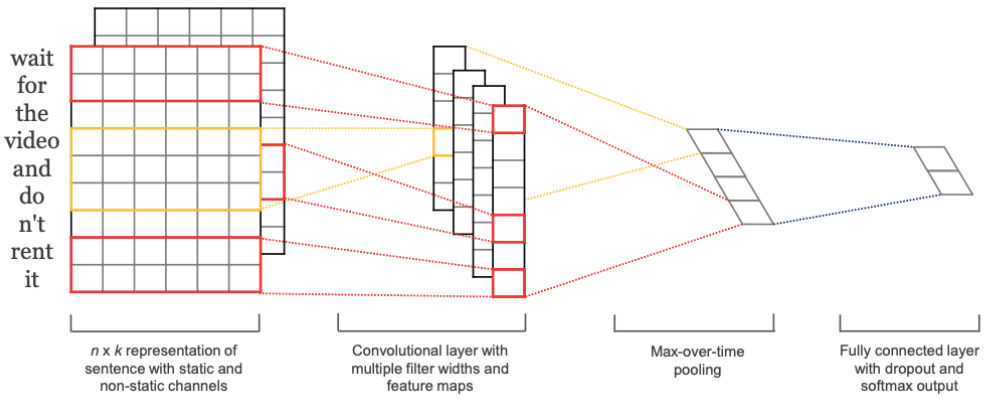


Fig 2 : Document Classification using Convolution Network Architecture

Model Summary:

First Layer: Embedding layer

First Convolution layer-

No of filters: 16

Kernel Size:16

Activation Function: Relu

Second Convolution layer-

No of filters: 20

Kernel Size: 20

Activation Function: Relu

Learning Rate = 1e-3

Optimizers: Adams Optimizer

Loss: Category Cross-Entropy

No of epochs = 5

Dataset and Experimental setup:

Data was collected from the Project Gutenberg file which consisted for three text files. This raw data was preprocessed using regex and split into paragraphs each of length 25. These paragraphs from different text files were labelled 0, 1 and 2 and data frame was created. This dataset was then split up into 70 % of training data and rest as testing data. Both the training and testing data were tokenized for feature extraction.

Dataset summary:

Dataset shape: 11449,2

No of tokens: 113

Length of Vocabulary:25815

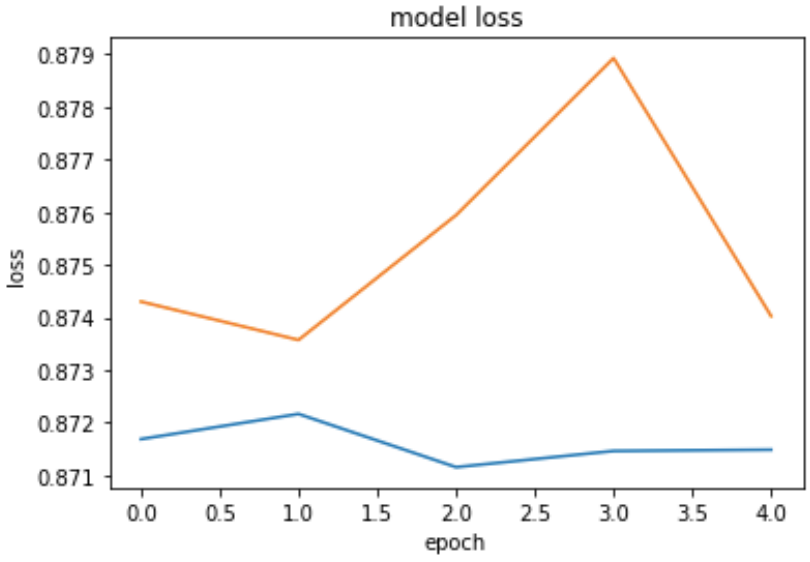
Training Shape: 8014,804

Testing Shape: 3425,804

Results:

The model gives an accuracy of 64.25 % using the sklearn accuracy score metrics. Using a random text, I tried to predict the category of this new random text. According to the prediction it belonged category 0 which a part of book written by Jane Austen or part of text file pg31100.txt.

The graph validation loss vs model loss is shown in the graph below



2. Sentimental Analysis using Convolution Neural Networks

Problem Statement:

Creating a convolution neural network model to analyze positive or negative reviews.

Model:

The neural network model is described in such a manner by which one feature is extracted from one filter. The model uses multiple filters (with varying window sizes) to obtain multiple features. These features form the penultimate layer and are passed to a fully connected SoftMax layer whose output is the probability distribution over labels.

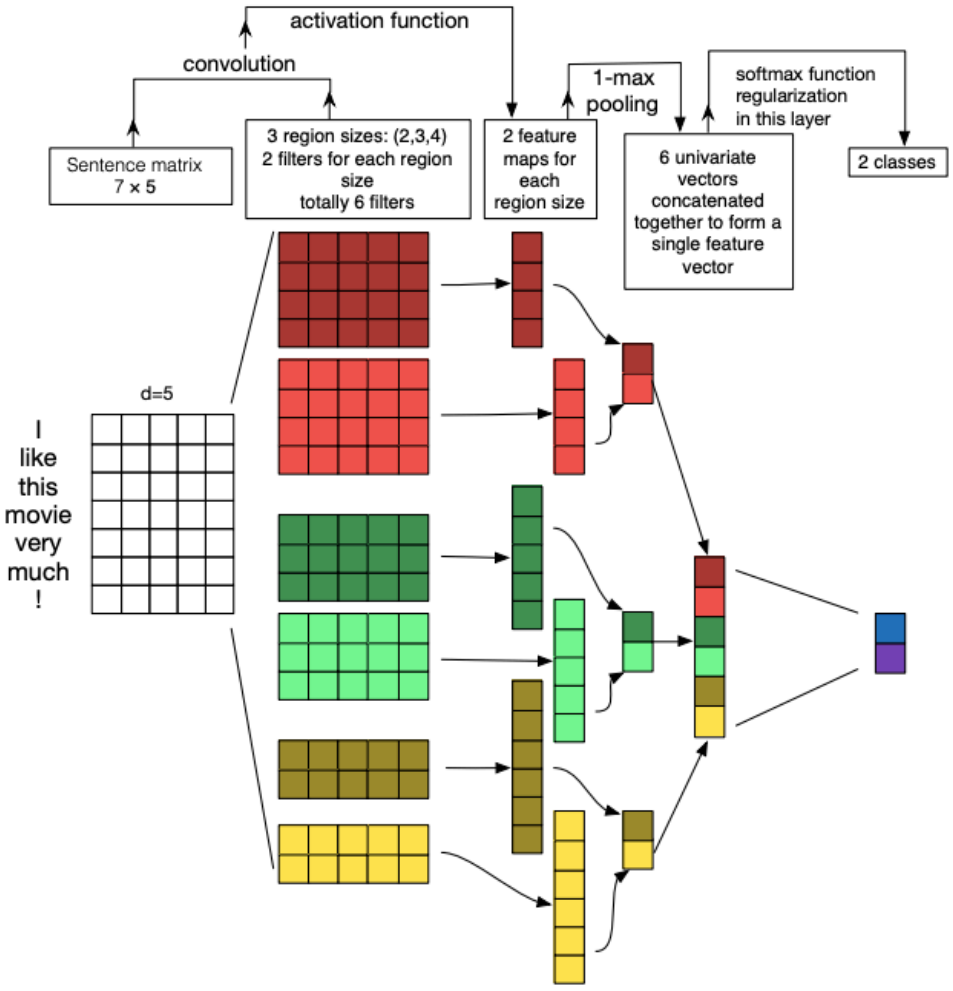


Fig 2 : Sentimental Analysis using Convolution Network Architecture

Model Summary:

First Layer: Embedding layer

First Convolution layer-

No of filters: 50

Kernel Size: 50

Activation Function: Relu

Second Convolution layer-

No of filters: 50

Kernel Size: 50

Activation Function: Relu

Learning Rate = 1e-3

Optimizers: Adams Optimizer

Loss: Category Cross-Entropy

No of epochs = 28

Dataset and Experimental setup:

Data was provided in the form of Xml file. This raw data was preprocessed using regex. These raw text files were then combined to form a data frame. The positive reviews were labelled as 1 and negative reviews were labelled as 0. The dataset was then split up into 90 % of training data and rest as testing data. Both the training and testing data were tokenized for feature extraction. The word2vec model was extracted from the Google word vectors which

can be downloaded online via the following link : //s3.amazonaws.com/dl4j-distribution/GoogleNews-vectors-negative300.bin.gz

Dataset summary:

Dataset shape: 2000,2

No of tokens: 113

Length of Vocabulary:23654

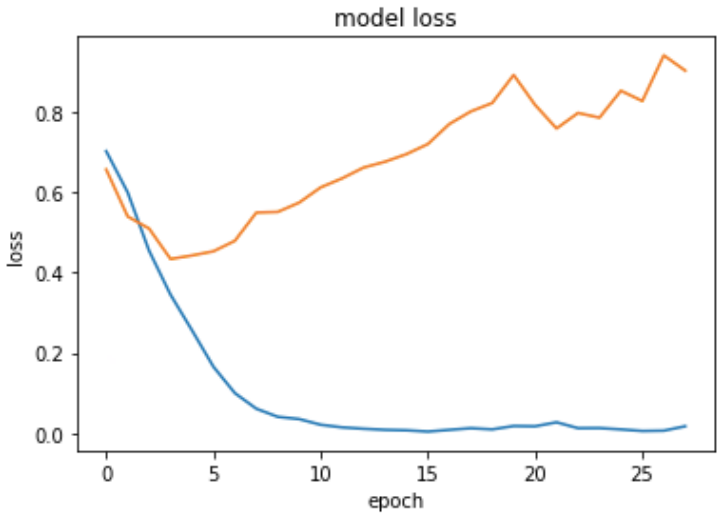
Training Shape with tokens: 23625,300

Results:

The model gives an accuracy of 81 % using the sklearn accuracy score metrics. Using the test dataset I have tried to predict whether the review was positive or negative. The confusion matrix for the model is given as follows:

1	101
0	99

The graph validation loss vs model loss is shown in the graph below



References:

1. Yoon Kim. Convolution neural Network for Sentence Classification.

2. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Network.