CS5590 – Deep Learning – Lab Assignment 2 Report

Multi-class Text Classification using CNN

Class Id 22: Avni Mehta

# Introduction

Text classification is a very common problem, where the goal is to classify the sentences into predefined categories. A convolutional neural network (CNN) is very effective for image recognition or text classification.

For the last In-Class exercise, we performed text classification using CNN where we trained the model for fashion and lifestyle data and then we evaluated the model based on accuracy and loss. For this lab assignment, we are required to perform a multi-class Text classification using CNN.

The dataset used is Consumer Finance Complaints from Kaggle, where the input is the complaint narrative and the output is product (has 11 classes).

# Objectives

The objective is to train a CNN model that classifies the complaints into 11 different classes. The performance is evaluated by calculating the accuracy, number of correct predictions and loss while predicting the test data.

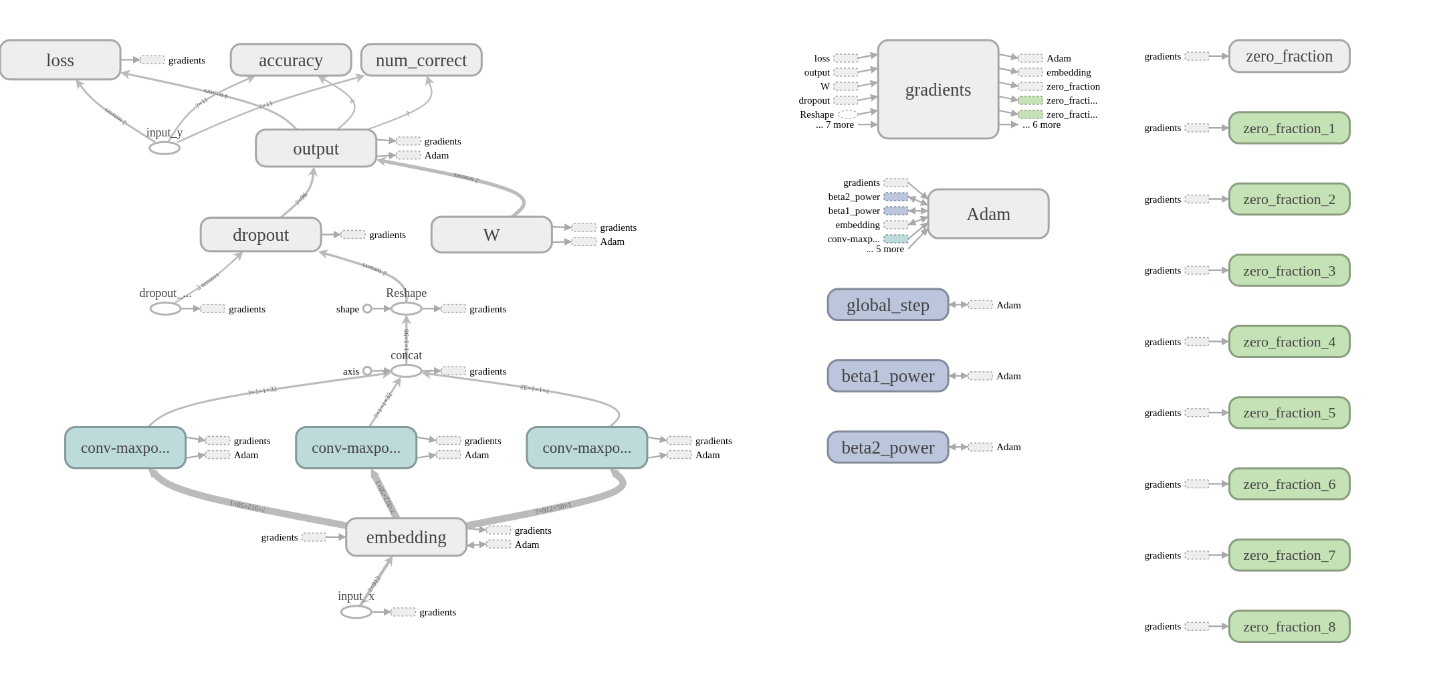
# Approaches / Methods

The steps involved in building this model for Customer Complaints dataset are:

1. Load the data from csv file.
2. Perform data-preprocessing:
3. Drop (unwanted) columns except complaint narrative and product.
4. Shuffle the new dataframe.
5. Map the labels to one-hot encoding.
6. Define parameters for the model viz. dropout rate, number of epochs, etc
7. Split the data into 90% training data and 10% testing data.
8. The initial layer embeds words into low-d vector. The next layers are convolutional layers for multiple filter sizes (3,4,5). Then we maxpool the result of cnn layer, add dropout and classify using softmax.
9. AdamOptimizer, an extension to schotastic gradient descent is used for training the model inorder to minimize the cost.
10. For training the model, training data is used. Total loss for is calculated for every epoch by running the optimizer.
11. Test data is used for testing the newly built model. The predicted classification is compared with actual value and the accuracy is calculated based on number of correct predictions.

# Workflow

The entire network for CNN model can be visualized using Tensorboard as follows:



# Datasets

The dataset used is US Consumer Finance Complaints dataset from Kaggle. This data set consists of thousands of consumer complaints for financial products, sent by CFPB to Kaggle. The csv file contains many columns but only consumer\_complaint\_narrative is used as input and product is used as output. The purpose is to classify the complaint data into 11 product classes.

# Parameters

The parameters used for building the model are:

* num\_epochs: 1
* batch\_size: 37
* num\_filters: 32
* filter\_sizes: 3,4,5
* embedding\_dim: 50
* l2\_reg\_lambda: 0.0
* evaluate\_every: 200
* dropout\_keep\_prob: 0.5

# Evaluation & Discussion

For just one epoch and with the dropout rate of 0.5, we got accuracy of about 75%. If the number of epochs were to be increased, the accuracy would increase drastically. Also, by adjusting the dropout rate, we could further improve the accuracy.

# Conclusion

Convolution Neural Network again proves to be an efficient model for standard problems like multi-class text classification. We achieved about 75% accuracy for this dataset with just one epochs. When the number of epochs is increased to 150, and when the dropout rate is adjusted, the accuracy would drastically be improved. This dataset proved good for classification problems.