SMAI - Project Proposal Document Team 44

Team Details

Team Number: 44

Team Name: KARS

Team Members: Karthik Viswanathan (2019113015), Abhishekh Sivakumar (2019101014), Rohan Asokan (2019101031), Sathyakameshwarao (2021900005)

Project Assigned: Character-level Convolutional Networks for Text Classification.

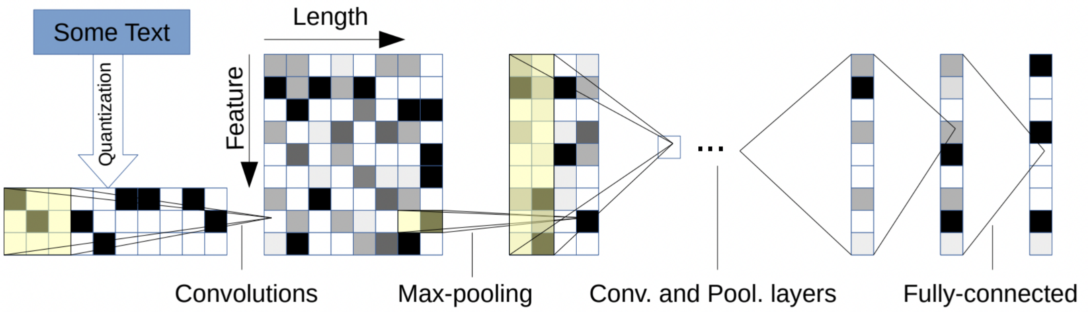
Problem Statement

Given a list of sequences of words and a classification task, train a Deep Learning Network to classify the list of sequences into their respective classes. Test and store the metrics of the classifier and compare it with other novel methods which exist.

Goals and Approach

The goal of this project is to automate text classification with minimal errors. We do so using a novel method described in the paper known as temporal convolutions. Temporal convolutions involve mapping the input feature set into a d-dimensional vector. The value of d depends on the stride length and the kernel size. Temporal convolutions are necessarily 1-D convolutions aimed to create a separable learning boundary between two different sequences logically. The paper's authors also propose a temporal pooling module, which takes the maximum pool from a given input vector. To quantify the input features, the authors tokenize each character as a one-hot encoding. They also take into consideration special characters.

The model involves text quantization followed by six temporal convolution layers and three fully connected layers to learn from text embeddings depending on the task. The authors also insert two dropout layers for regularization. Weights initialized have a Gaussian distribution. The illustration of the model is as follows:



Dataset

We plan on using the AG's News dataset for our project. Training NLP tasks using ConvNets takes time, and hence, if time permits, we plan to test our project on the Sogou News dataset. Unlike images, where data augmentation is done using signal transformations, the authors augment the data for this NLP task by replacing words with their appropriate synonyms.

Benchmarking

After training, we plan on benchmarking the model with some traditional and novel methods. These include:

1. Traditional methods include Bag of Words and Bag of n-grams. In bag of words, we take the k most frequent words and calculate the counts for each of these words. In bag of n-grams, we choose phrases of length n and calculate their document frequency. We will train each of these with a Naive Bayes classifier and evaluate its metrics.
2. Novel approaches include LSTM's. We will use a pre-trained word2vec model with an embedding size of 300.

Deliverables

The deliverables of this project include the working model, the performance of this model, and the comparison of this model with other benchmarks. We will also provide proper documentation and report for this project.

Expected Timeline

Milestone	Timeline
Dataset Preparation	@November 12, 2021
Working Model	@November 18, 2021
Benchmarking with traditional Approaches	@November 22, 2021
Benchmarking with LSTM	@November 28, 2021
Free Buffer in case of delays	@December 2, 2021
Report and Documentation	@November 5, 2021

Note

Please note that there might be minor changes attributing to feasibility. In such a situation, we will let our mentor TA know about the change.

References

1. Xiang Zhang and Junbo Zhao and Yann LeCun, 2016. Character-level Convolutional Networks for Text Classification, pp.1509.01626