TGCN: A Novel Deep Learning Model for Text Classification

Yuan Li
New York University
Brooklyn, NY
yl6606@nyu.edu

Dongzi Qu
New York University
Brooklyn, NY
dq394@nyu.edu

1. Problem Statement

Text classification is a fundamental and important problem in natural language processing (NLP). A necessary intermediate step during text classification is text representation. There are several traditional methods for text presentation such as IDF-TF vectors, bag-of-words, n-grams and so on. Recently, recurrent neural networks (RNN) such as long short term memory (LSTM) have received wide attention from scientists. They can capture semantic and syntactic information in a sequence. However, these deep learning models may have a poor performance in some long-term or non-consecutive sequences analysis. Meanwhile, RNN models usually need numerous training data to reach high accuracy, which cause constraints in industrial applications. So we want to research if there is a better method for text classification problem with less data and training time.

2. Related Work

There are different methods to solve text classification problem. Traditional studies primarily focus on feature engineering. They use feature engineering to generate well designed features, such as the bag-of words, n-grams and graph features[2]. However, these methods need set up text representations manually, which will cause training them more difficult.

Besides traditional methods, deep learning methods are also widely used in text classification problem. One representative deep network is RNN. Liu[1] used LSTM, one specific kind of RNN model, to solve text classification problem. Recently, graph neural networks (GNN) have been initially used in deep learning problems, as they use vast relational architecture to preserve a better global information of a graph or long-distance feature than LSTM. Yao[3] provided a novel graph neural network for text classification, using smaller data set and training time, but receiving a better performance.

3. Goal

We intend to follow the graph neural network proposed by Yao[3] and reproduce the results on several different datasets. In this paper, the authors took advantage of two-layer graph convolutional network (GCN) and constructed the graph using both words (from vocabulary) and documents. Instead of using pre-trained word-embeddings, this model generates the embeddings automatically while making the classifications. Specifically, there are three goals in our project:

- Reproduce the results provided in the original paper.
- Make some modifications on the original model, for instance, train with more layers or represent the weight using different method rather than TF-idf (word to document) and PMI (word to word). Experiment on modified models and compare the new results with previous ones.
- Since the original paper achieve the outstanding performance mainly on English document classification, we plan to test this model's performance on Chinese document classification task.

References

[1] P. Liu, X. Qiu, and X. Huang. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*, 2016.

[2] Y. Luo, Ö. Uzuner, and P. Szolovits. Bridging semantics and syntax with graph algorithms—state-of-the-art of extracting biomedical relations. *Briefings in bioinformatics*, 18(1):160–178, 2017.

[3] L. Yao, C. Mao, and Y. Luo. Graph convolutional networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7370–7377, 2019.