Utilizing Word Order in Text Categorization by Convolutional Neural Network

Chang Liu Ning Zhang

October 2, 2015

1 Background

In this project, we will focus on the problem of 'text categorization'. More specifically, 'text categorization' is the task of automatically assigning predefined categories to documents written in natural languages. The input for this problem is the text, and the output is the predicted category.

From recent reading, we get to know a bunch of methods for this problem. However, due to low use of word order, traditional natural language processing method like *bag-of-words* vectors doesn't achieve a good performance. As a result, a paper about using convolutional neural network is published and a new approach is proposed to improve the accuracy for this problem.

After reviewing their papers and other related work, we get to know the promising method and want to do some improvement for their algorithm to further enhance the representation ability for their network structure. The aim of our research is to explore the use of convolutional neural network on word order and by better understanding their effect and limitation, we will improve the network structure and adjust the parameters for the model to achieve better accuracy for the prediction.

2 Approach Review

In the original paper, it's noted that the loss of word order caused by *bag-of-word*(BoW) vectors is particularly problematic on sentiment classification. The remedy of using n-gram is not always effective. So in the paper, they tried another approach that use Convolutional Neural Network(CNN) to solve this problem.

Specifically, for text, treat each word as a pixel, and then given a document $D = (w_1, w_2, ...)$ with vocabulary V, treat D as if it were an image of $|D| \times 1$ pixels with $|V|$ channels, then represent each pixel as a $|V|$-dimensional vector, and get the representation of a document using a vector. Like the image, represent the region(like the sentences here) with concatenation of pixels, making a $p|V|$-dimensional vector. This is how CNN represent the document like an image.

Another baseline method is the n-gram method that use bag-of-words method, which is the state-of-art method at that time, with error rate of 8.13 by NB-LM with BoW3(MMRB14). By using the new method, they get an error rate of 7.67 for the same task, which is better by nearly 0.5%.

3 Our Approach

In our approach, we plan to examine their open source project and dataset, especially check their implementation of convolutional neural network, and improve their algorithm from the view of network architecture. Specifically, how can we better make the network more deeper and wider, with some rectified units(ReLUs) to empower the network capacity of word order. In this way, the system can have more accurate ability to utilize the word order.

Furthermore, in their work, they only use a 5-layer structure. We want to do more experiment about the influence of the network layer on the accuracy. We will try to adjust the parameters and layer size to explore its influence on the prediction result, and try to achieve the optimal structure and parameter for this task.

4 Overall Plan

For experiment, we will first try to set up the environment and reproduce the result that they proposed in the paper. Then we will fully utilize their code and add our network modules and training parameters to do more experiment about our hypothesis.

For evaluation, we will not only use the error rate that they use in the paper, we will further check the precision and recall, then draw the ROC curve to check the best classifier and network structure to avoid overfitting.

5 Dataset

Their code and dataset is available here: `http://riejohnson.com/cnn_download.html`