Utilizing Word Order in Text Categorization by Convolutional Neural Network

Chang Liu Ning Zhang

November 6, 2015

1 Background

Text categorization is a very interesting topic in NLP field. Currently most of the work focus on using Bag-Of-Words method, which ignores the word order and only take use of the word count feature. In previous research, researchers find that losing the word order information will more or less influence the accuracy when predicting the text category.

In this experiment, we want to explore the convolutional neural network's (CNN) potential in spatial feature when classifying the text. Especially, as CNN can utilize the image's spatial covariance, we want to know how this characters can be applied in text processing field. According to the most recent work in[1], Rie has done some excellent work in applying CNN in text and achieved state-of-art accuracy. We want to use their method as the baseline and improve the representation ability by enriching the network structure or trying other deep learning methods.

2 Progress

We've gone through their papers for the past weeks and get a good understanding of how their algorithm works, then we configured the environment to reproduce their result. After some efforts, we get the CUDA code running, as their software packages can only be running on GPU using CUDA. And then we review their CUDA code and find that it's quite complex, they've implemented a very complex network and libraries for future extension, but for some basic operations in NLP, we need to extract the simple task and find the core processing using CNN.

We are working hard on this field and hope to quickly adapt to their framework. As we know if we want to implement similar work it may take too long, so our focus in next stage should be apply our ideas in their framework using CUDA and try to replace their CNN structure with ours. I think it should be major part in the next step.

If we can do it successfully, we will get some new result and then verify our algorithm's effectiveness. Otherwise, we will think of our methods in CNN field to improve the training model.

3 Data set

We've used their database:
1) IMBD: movie revies
2) Elec: electronics product reviews
3) RCV1: topic categorization dataset of Reuters news articles.

4 Plan

Potential working directions to explore:

1.Bow-CNN outperforms seq-CNN in topic categorization even though it discard the local word order. This is interesting, we are wondering if this is brought by the calculation efficiency, since from the common sense word order should always provide positive influence on performance.

2.Parallel CNN, this is kind of "multi-resolution" processing in NLP domain. Also such kind of structure reminds us of multimodal learning. If we take different pathways as independent models and finally convergent at one shared layer. Therefore, it is possible to use CNN purely as a feature extraction tool. We can use the extracted features as input for some successful multimodal learning model such as Multimodal DBM.

5 Reference

[1] Rie Johnson and Tong Zhang. Effective use of word order for text categorization with convolutional neural networks. NAACL-HLT 2015.

[2] Rie Johnson and Tong Zhang. Semi-supervised convolutional neural networks for text categorization via region embedding. To appear in NIPS 2015.

[3] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. ACL, 2011.

[4] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. RecSys, 2013.