Weekly Update for NLP Project #2

Chang Liu

September 17, 2015

1 Introduction

This document is used to record the weekly update for me in my NLP project. This is the second update for our project in 3rd week.

2 Identify your project topic

My project topic is to study CNN on **text categorization** to exploit 1-D structure of text data for accurate prediction. More specifically, I want to learn how to use CNN to better represent the word order, so that we can make full use of this information for text categorization. In the last step, I want to apply this framework to *sentiment classification* and *topic classification* tasks.

3 Explore related literature

(please include 1-2 paragraph summaries of the papers you read),

The paper talks about the use of Convolutional Neural Network(CNN) to make use of the word order for text categorization. It's noted that the loss of word order caused by bag-of-word vectors(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 CNN to solve this problem.

Specifically, for text, treat each word as a pixel, and then given a document $D = (w1, w2, ...)$ 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 method is the n-gram method that use bag-of-words method, by comparing these two representations and experiment results, the paper reached a conclusion that CNN has a better representation power.

4 Identify relevant data sets

(please include links) Dataset: http://riejohnson.com/cnn_data.html

5 Set up and execute the annotation task (if applicable)

6 Adapt external libraries

NA

7 Implement your system

NA

8 Evaluate your system performance

NA

9 Reference

Effective use of word order for text categorization with convolutional neural networks:http://arxiv.org/pdf/1412.1058v2.pdf

Project **ConTEXT**: http://riejohnson.com/cnn_download.html

Data: http://riejohnson.com/cnn_data.html

Code: http://riejohnson.com/software/conText-v1.01.tar.gz

10 Summary

For this week, I just start searching papers and find the paper that actually did my previous plan. I read the paper about applying CNN on text categorization, and find some very interesting points in the work that are novel according to my understanding of the traditional bag-of-n-gram approach. I roughly reviewed their project website,dataset and source code as well, and find many related resources to explore in the next week.

My plan for the next week is to try to set up the environment of their program, and then run some basic experiments in the examples. Then if time is abundant, I will try to read their implementation of some of key points that combine the CNN with text processing, and I will try to find some work the worth further exploring.