Effective Use of Word Order for Text Categorization with Convolutional Neural Networks

Abstract

Text Categorization is one of the long lasting research in the field of natural language processing. By the advent of many machine learning approach the accuracy of such categorization task have improved drastically. The text categorization process have many applications which includes sentimental analysis, spam detection etc. The text categorization task can be summarize as follow, given a sequence of text it should able to assign the text into predefined class labels.

For any Natural Language Processing task the initial stage is to represent the text document in some meaningful manner, then classify the document using some model. The most common approach to represent the text document is the simple Bag-of-Words models, although it is a fairly simple and straightforward approach it lacks word ordering thus it limits the semantic representation of text. For classification the mostly classifiers like Support Vector Machines are used but the quality of categorization of these models are highly depended on handcrafted features.

By the advent of Deep Neural Networks the classification tasks are further improved by a great margin. Convolutional Neural Networks(CNN) for example, originally developed for computer vision problems, have successfully applied to many Natural Language Processing Tasks and has achieved remarkable results. In this project in order to benefit from word ordering for text categorization a Convolutional Neural Network is used. CNN works by extracting features from a small region in the input and later combines to get the global features. At the end there is a final linear classifier for doing the final classification, the input to the classification layer is the features that extracted by the CNN layers.

Muhammad Ijaz k
MCA S6
LTVE17MCA067