Convolutional neural network for modelling sentences and its uses in sentiment analysis

Arpit Shrivastava
shriap@iitk.ac.in, 12161

Jayesh Kumar Gupta
jayeshkg@iitk.ac.in, 11337

March 23, 2015

Problem and motivation

The ability to accurately represent sentences is central to language understanding. Semantic modelling is at the core of many tasks involving a degree of natural language comprehension. These tasks include sentiment analysis, paraphrase detection, entailment recognition, summarisation, discourse analysis etc. Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. [2]. Existing approaches to sentiment analysis can be grouped into four main categories: keyword spotting, lexical affinity, statistical methods, and concept-level techniques.
In this project we will implement the convolutional architecture named Dynamic Convolutional Neural Network (DCNN) described in paper [1] for the semantic modelling of sentences. We will then test our implementation on the problem of sentiment prediction using distant supervision on twitter tweets/ movie reviews.

Method

Semantic modelling

We will define a **convolutional neural network architecture** and apply it to the semantic modelling of sentences. This network will handle input sequences of varying length. The layers of the DCNN will be formed by a convolution operation followed by a pooling operation. The layers in the network interleave one-dimensional convolutional layers and dynamic k-max pooling layers. Dynamic k-max pooling is a generalisation of the max pooling operator.

Sentiment analysis

We plan to train the models on a large dataset of tweets, where a tweet is automatically labelled as positive or negative depending on the emoticon that occurs in it. The tweets are preprocessed following the procedure described in paper [3].
The network will then be trained on this data with mini-batches by back-propagation and the gradient-based optimisation will be performed using the Adagrad update rule.
The performance of this model will then be compared with that of the non-neural n-gram based classifiers.

Datasets

The dataset will consist of a large number of tweets, where a tweet is automatically labelled as positive or negative depending on the emoticon that occurs in it. The test set will contain some hand-annotated tweets.
Another corpus that we can use is the dataset of tweets with sentiment labels available at [4].

References

[1] Phil Blunsom, Edward Grefenstette, and Nal Kalchbrenner, "A convolutional neural network for modelling sentences". Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014.

[2] Sentiment Analysis, `http://en.wikipedia.org/wiki/Sentiment_analysis`. Accessed on 20 March, 2015

[3] Alec Go, Richa Bhayani, and Lei Huang, "Twitter sentiment classification using distant supervision". Processing, pages 1-6, 2009.

[4] Sentiment140 (formerly known as "Twitter Sentiment"), `http://help.sentiment140.com/home`. Accessed on 20 March, 2015