Final Project Proposal

Ari Beller (abeller) and Nicholas Lum (nlum1)

Convolutional Neural Networks for Sentence Classification

Reference: https://arxiv.org/pdf/1408.5882.pdf

For our project we would like to replicate the text classification architecture using convolutional neural nets developed in this paper. The paper shows that their CNN implementation achieves "excellent results on multiple benchmarks". Words are represented as word embeddings, which are pre-trained (Google word2vec). The paper experiments with a variety of different training regimens, sometimes leaving the embeddings static and other times fine-tuning them in the training process.

The training architecture is described by the following diagram.

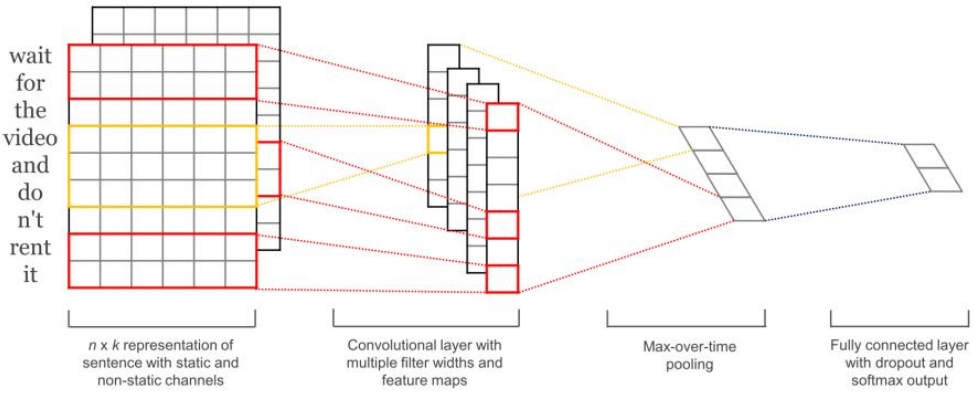

Figure 1: Training Architecture

Sentences are represented as embedding vectors of constituent words concatenated column-wise resulting in a matrix of size sentence length by embedding size. We then apply a convolutional layer, followed by a max-pooling layer, and lastly a fully connected layer with dropout and softmax. There are a number of minor variants on the model that are further detailed in the paper. Our aim is to build a system that allows us to modify the CNN such that we can replicate the experiments carried out in the paper.

Most of the datasets involved in the experiments from the paper are publicly available and linked to by the paper. Most of the classification tasks involve some kind of sentiment analysis. There is also a question type analysis, and a classification task geared at determining whether sentences are objective or subjective.