Ikhlas Attarwala
05-08-19 / Week #6
Improving Efficiency with NNLMs / Paper 1

Efficient Estimation of Word Representation in Vector Space
by Mikolov, Tomas et al.

This paper was presented by Mikolov in 2013, and is widely recognized as the famous word2vec paper. They proposed representing words in a continuous vector space, such that differences between semantics and syntax would be preserved. This allowed for better computation of analogies via vector addition and cosine similarity. Using the assumption that words with similar meanings tend to appear in similar locations, the method they propose also uses unsupervised learning to minimize computational complexity, such as with NNLMs and RNNLMs, and then covering log linear CBOW (continuous bag-of-words) and skip-gram. Computational complexity is proportional to the number of training epochs, times the number of words in the training set, times a value Q depending on the type of model.

With feedforward NNLMs, the input is projected to a projection layer P with dimensionality N*D, with a hidden layer of size H that computes the probability distribution across all words. A vocabulary size of V is encoded as a 1-of-V input layer. This method reduces Q by using hierarchical softmax & Huffman's binary tree. The RNNLM is similar, but it does not have a projection layer. A lot of complexity falls on nonlinear hidden layers. The CBOW model is similar to feedforward NNLMs, but there is no nonlinear hidden layer, and log-linear classifiers use a window of word to predict middle words. Finally, the continuous skip-gram model is similar to CBOW but uses the middle word of the window to predict the remaining words within that same window. More distant words from the middle are given less weight as they are sampled from less distant words. The results showed that skip-grams had the greatest semantic accuracy when compared to NNLMs and CBOW, CBOW had the greatest syntactic accuracy of the three, and skip-gram + RNNLMs had greater accuracy than all other architectures or combinations of.