Department of Computer Science
Databases and Information Systems

Universitätsstr. 1 D–40225 Düsseldorf

Paper Summary

Distributed Representations of Words and Phrases and their Compositionality

Zeynep Boztoprak

The paper Distributed Representations of Words and Phrases and their Compositionality by Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado and Jeffrey Dean presents several extensions to the previously presented continuous Skip-Gram model. Further they introduced a simple data-driven approach to use vector representations of phrases to capture idiomatic phrases.

The Skip-gram model is used to find high-quality distributed vector representations. They use a simple neural network to predict the surrounding words from an input word. Therefore they define a hyperparameter C for the maximum number of words before and after that input word. These words are then defined as correct labels [1].

Their prior work on the Skip-gram model show that the learned vectors encode explicitly many linguistic regularities and patterns like syntactic and semantic word relationships. By using simple mathematical operations on the word vector representations meaningful results can be obtained. These pattern can be represented as linear translations.

Based on this model the authors modified the original Skip-gram model to improve both the quality of the vector representations and the training speed.

In the original Skip-gram model they used the hierarchical softmax which is an computationally more efficient approximation of the full softmax. In this paper they replaced it with negative sampling which is a simplified variant of the noise contrastive estimation [2]. This approach forces to only modify the weights of the k negative samples and for the current word rather than updating all weights. Selecting negative samples is done according to a uniform distribution.

They further make use of subsampling of frequent words for two reasons. First some of the most frequent words can provide less information than the rare words and second the vector representations of frequent words don't change significantly after training on million examples. The probability to keep a word in the vocabulary is related to its frequency.

With these extension the first evaluation of the word vectors was made on the analogical reasoning task containing syntactic and semantic analogies. The results show that negative sampling works best.

Next they extend the Skip-gram model from word based to phrase based by taking into account unigram and bigram counts. Bigrams above a choosen threshold are then used as phrases and replaced by one token. This is repeated 2-4 times over the dataset to construct phrases with more than two words. Evaluating these vector representations show that hierarchical softmax became the best performing method with subsampling frequent words.

Finally it should be said that the biggest factors influencing the model performance are the model architecture, the size of the vector representations, the subsampling rate used for frequent words and the size of the training window. Achieving the best performance on the word analogy task with a huge margin to all the other prior models is due to the fact that this model has been trained on about a huge amount of words. This was just possible because of the training speed gains.

References

[1] Mikolov, T., K. Chen, G. Corrado, and J. Dean
2013a. Efficient estimation of word representations in vector space.

[2] Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean
2013b. Distributed representations of words and phrases and their compositionality.