# Efficient Estimation of Word Representations in Vector Space

## 1 Review

Point of view [1] about this paper, The authors provide a clear motivation for the work, explaining the limitations of previous approaches to word representation and the potential benefits of distributed representations. They also present a thorough evaluation of their algorithms, comparing them to other state-of-the-art methods on a range of tasks. One of the key strengths of this paper is the clarity of the exposition. The authors provide detailed explanations of the Skip-gram and CBOW models, including the underlying mathematical concepts and the optimization algorithms used to learn the word vectors. They also provide guidance on how to choose hyperparameters and how to interpret the learned embeddings. Another strength of this paper is the empirical evaluation, which demonstrates the effectiveness of the proposed algorithms on a range of tasks, including word similarity, analogy completion, and named entity recognition. The authors compare their results to other state-of-the-art methods and show that their algorithms are competitive or superior in many cases. However, one potential limitation of this paper is that the evaluation is mostly focused on word-level tasks, and it may not generalize as well to larger-scale tasks such as document classification or machine translation.

From [2] The skip-gram and continuous bag-of-words (CBOW) models propose a simple single-layer architecture based on the inner product between two words vectors. For simplicity, robustness and observation many NLP systems and techniques treat words as atomic units where no notation of similarities between words and indices in a vocabulary. The main vision of this paper is to introduce techniques for measuring the quality of the resulting vector representations that words can have multiple degrees of similarity and can be used for learning high-quality word vectors from huge data sets with billions of words, and with millions of words in the vocabulary. Author claimed none of the previously architecture has been successfully trained on more than a few hundred of millions of words, with a modest dimensionality of the word vectors between 50 - 100. They try to maximize accuracy of vector operations by developing new model architectures that preserve the linear regularities among words and also design a new comprehensive test set for measuring both syntactic and semantic regularities.

For estimating continuous representations of words author proposed Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). Previously it was shown that distributed representations of words learned by neural networks perform significantly better than LSA for preserving linear regularities among words and LDA moreover becomes computationally very expensive on large data sets. First define the computational complexity of a model as the number of parameters that need to be accessed to fully train the model and then try to maximize the accuracy, while minimizing the computational complexity.

In the models, using hierarchical softmax where the vocabulary is represented as a Huffman binary tree. To train models on huge data sets, implemented several models on top of a large-scale distributed framework called DistBelief, including the feedforward NNLM and Adagrad (mini-batch asynchronous gradient descent with an adaptive learning rate procedure). As most of the complexity is caused by the non-linear hidden layer so they propose Log-linear Models for learning distributed representations of words that try to minimize computational complexity where the non-linear hidden layer is removed and the projection layer is shared for all words, and trained in two steps: first, continuous word vectors are learned using simple model, and then the N-gram NNLM is trained on top of these distributed representations of words.

The first architecture is continuous bag-of-words (CBOW) model where all words get projected into the same position. The second architecture is Continuous Skip-gram Model, it tries to maximize classification of a word based on another word in the same sentence. To measure quality of the word vectors, defining a comprehensive test set that contains five types of semantic questions, and nine types of syntactic questions and there are 8869 semantic and 10675 syntactic questions. The questions in each category were created in two steps: first, a list of similar word pairs was created manually and then, a large list of questions is formed by connecting two words pairs. Author included single token words in test set. By using a Google News corpus for training the word vectors and corpus contains about 6B tokens to estimate the best choice of model architecture for obtaining

as good as possible results quickly, evaluated models trained on subsets of the training data, with vocabulary restricted to the most frequent 30k words. Firstly, compare different model architectures for deriving the word vectors using the same training data and using the same dimensionality of 640 of the word vectors. Then trained a feedforward NNLM with the same number of 640 hidden units using the DistBelief parallel training and trained recurrent neural network language model that took about 8 weeks to train on a single CPU. For Semantic-Syntactic Word Relationship test set CBOW and Skip-gram, Semantic accuracy and Syntactic accuracy are 24,55 and 64,59 respectively. For MSR Word Relatedness test set score are 61 and 56. For the experiments report three training epochs using with stochastic gradient descent and backpropagation where starting learning rate 0.025 and decreased it linearly. For Comparison of publicly available word vectors on the Semantic-Syntactic Word Relationship test set, and word vectors from their models the Syntactic accuracy of NNLM [6B training words] is 64.5 percent, total 50.8 persent, and Semantic accuracy of Skip-gram [783M training data] is 50 percent and the total accuracy is 53.3 persent.

Moreover used 50 to 100 model replicas during the training and report the results of several models trained on the Google News 6B data set, with mini-batch asynchronous gradient descent and the adaptive learning rate procedure called Adagrad. Comparison of models trained using the DistBelief distributed framework the total accuracy of NNLM, CBOW, and Skip-gram are 50.8, 63.7, and 65.6 persent. Comparison and combination of models on the Microsoft Sentence Completion Challenge the accuracy of Skip-gram + RNNLMs is 58.9 persent. The Skip-gram model itself does not perform better than LSA similarity, the scores from this model are complementary to scores obtained with RNNLMs, and a weighted combination leads to a new state of the art result 58.9 persent accuracy (59.2 persent on the development part of the set and 58.7 persent on the test part of the set). Examples of the word pair relationships, using the best word vectors Skip-gram model trained on 783M words with 300 dimensionality score about 60 persent. The observation of improvement of accuracy of their best models by about 10 persent absolutely on the semantic-syntactic test. Through this paper it is possible to train high quality word vectors using very simple model architectures, compared to the popular neural network models (both feedforward and recurrent) and the word vectors can be successfully applied to automatic extension of facts in Knowledge Bases, and also for verification of correctness of existing facts.

# References


[1] OpenAI, "Chatgpt," 2022.

[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.