# Efficient Estimation of Word Representations in Vector Space

Author: Mikolov et al.

created by paper2slides

Uploaded to arxiv on 2013-01-16

https://arxiv.org/abs/1301.3781

# Executive Summary

- Introduces two novel architectures (CBOW and Skip-gram) for computing continuous vector representations from large datasets.
- Demonstrates quality improvements in word similarity tasks compared to previous models.
- Shows large accuracy improvements in syntactic and semantic word similarity at reduced computational costs.
- Vectors trained on a 1.6 billion words set achieve state-of-the-art performance on syntactic and semantic tasks.

# Introduction

- Traditional NLP systems treat words as atomic units without considering similarities.
- Simple models trained on large datasets often outperform complex models trained on smaller datasets.
- High-quality transcribed speech data is often limited, making simple techniques insufficient.
- Advanced techniques, such as distributed representations of words, outperform traditional models.

# Goals of the Paper

- Introduce techniques to learn high-quality word vectors from billions of words with large vocabularies.
- Use existing techniques to measure quality by ensuring word similarities and maintaining multiple degrees of similarity.
- Develop new model architectures that preserve linear regularities among words.
- Create a comprehensive test set for both syntactic and semantic regularities.

# Previous Work

- Representation of words as continuous vectors has a long history, involving various neural network models.
- Feedforward NNLM proposed by Bengio et al. uses linear projection and non-linear hidden layers.
- Recent models like CBOW and Skip-gram are simpler and computationally efficient.
- CBOW and Skip-gram architectures perform better at preserving syntactic and semantic relationships than older models.

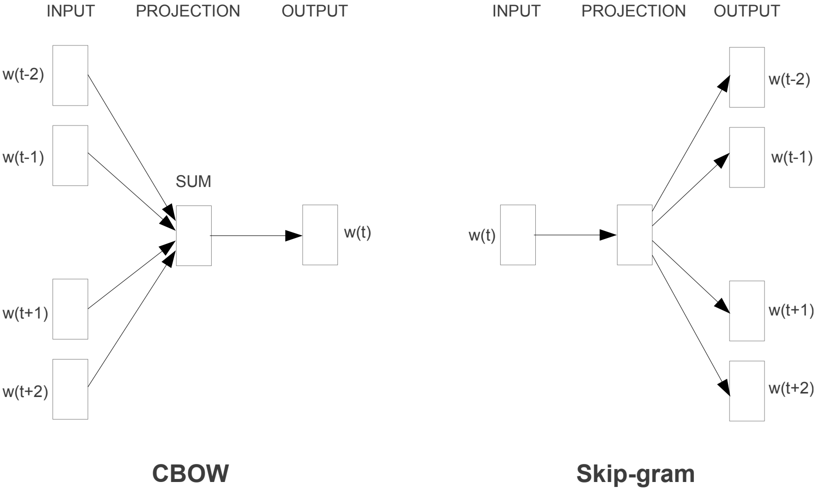


Figure: New model architectures: CBOW predicts a word based on context, while Skip-gram predicts context words given a word.

# Continuous Bag-of-Words Model (CBOW)

- Similar to feedforward NNLM but removes the non-linear hidden layer and shares the projection layer.
- Order of words in context does not matter; uses a bag-of-words approach.
- Predicts the current word based on surrounding context words.
- Training complexity is $Q = N \times D + D \times \log_2(V)$.

# Continuous Skip-gram Model

- Uses the current word to predict surrounding context words.
- Maximizes the classification of a context word given the current word within a certain range.
- Less computationally intensive and allows for large-scale training.
- Training complexity is $Q = C \times (D + D \times \log_2(V))$.

# Experimental Setup

- Used Google News corpus containing about 6 billion tokens and limited to the most frequent 1 million words.
- Evaluated models trained on subsets of data to find optimal configuration.
- Training used three epochs with stochastic gradient descent and backpropagation.
- Model accuracies evaluated on a comprehensive Semantic-Syntactic test set.

# Comparison of Architectures

- RNNLM, NNLM, CBOW, and Skip-gram architectures were compared using the same training data.
- Skip-gram model showed the best performance on semantic tasks.
- CBOW model performed well on syntactic tasks.
- RNNLM showed lower performance compared to CBOW and Skip-gram.

# Performance Results

- CBOW and Skip-gram were trained on a single CPU using the Google News corpus.
- Performance compared against publicly available word vectors.
- Skip-gram model achieved the highest total accuracy at 53.3%.
- Showed computational efficiency and scalability to larger datasets.

# Parallel Training with DistBelief

- Implemented models in DistBelief for large-scale distributed training.
- Used mini-batch asynchronous gradient descent and Adagrad for learning rate adaptation.
- Training employed 50–100 model replicas, each using many CPU cores.
- Results: improved accuracy significantly compared to single CPU training.

# Microsoft Research Sentence Completion Challenge

- Evaluated Skip-gram on MSR Sentence Completion Challenge.
- Combined Skip-gram and RNNLM for state-of-the-art performance of 58.9% accuracy.
- Demonstrated complementary scores for improved overall results.

# Examples of Learned Relationships

- France – Paris + Italy = Rome.
- big – bigger, small – larger, cold – colder.
- Einstein – scientist, Picasso – painter.
- Microsoft – Windows, Google – Android.

# Conclusion and Future Work

- Demonstrated that simple architectures (CBOW and Skip-gram) can efficiently create high-quality word vectors.
- Showed the effectiveness of these vectors on syntactic and semantic tasks.
- Future work includes scaling to even larger datasets and improving the training algorithms.