Efficient Estimation of Word Representations in Vector Space

Authors	Tomas Mikolov and team
Class	word2vec
Created	@Jan 3, 2021 10:40 PM
Materials	https://arxiv.org/pdf/1301.3781.pdf
Reviewed	☑
Status	Completed
Type	arXiv

Summary

The paper has introduced two models to represent words in continuous vector form. The first model is called CBOW-Continuous Bag Of Words and the second is Skip-gram. These models are trained on 1.6 billion English words. The model is built on the word frequency using a Huffman binary tree.

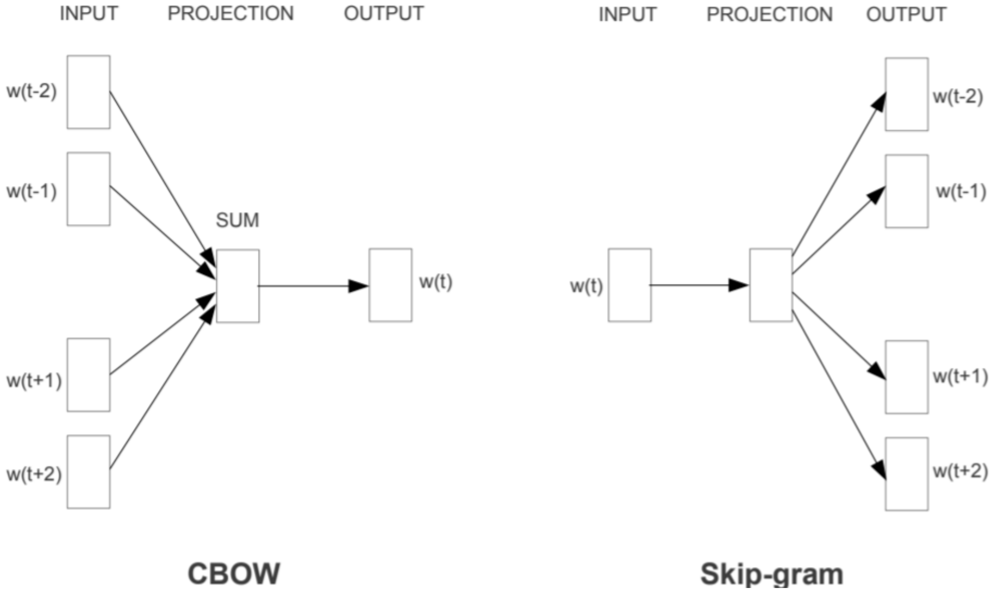
Architecture

1. Continuous Bag Of Words

   This is a simple feedforward neural network. There is a window size that can be set to specify the number of past words and future words to be looked at before predicting the center word. All the input words are one-hot encoded. The dimension of the input is the number of words in the training set. These are projected to reduce the dimension to a much lower level.

2. Continuous skip gram

   The second architecture is similar to CBOW, but instead of predicting the current word based on the context, it tries to maximize classification of a word based on another word in the same sentence. In this model, the words closer to the current word are given more weights and the words far from the current word are given lesser weights by sampling less from these words.



To get the probability of the words in the output, softmax is used.

Important Points

- The model is trained on google new data which had 6 billion tokens but only 1 million frequent tokens were used
- The performance of the model is measured on semantic and syntactic tasks