Summary

A neural probabilistic language model
(Original Word2Vec)

https://arxiv.org/pdf/1301.3781.pdf

Md. Nafis Faiyaz

May 16, 2021

1 Research Objectives

Introduce techniques to train high quality word vectors from a huge corpus with a low computational cost. They came up with continuous bag of words and skip-gram model.

2 Dataset

Training: Google News corpus with 6B tokens. First evaluated models trained on subsets of the training data, with vocabulary restricted to the most frequent 30k words.

Testing: 8869 semantic and 10675 syntactic questions were created in two steps: first, a list of similar word pairs was created manually. Then, a large list of questions is formed by connecting two word pairs. e.g. a list of 68 large American cities and the states they belong to, and formed about 2.5K questions by picking two word pairs at random. Multi-word entities (such as New York) are not present

3 Methodology

Make a graph / pipeline (lucid chart)

4 Experimental analysis

Setup and results

5 Discussion

Contribution and future work