**Efficient Estimation of Word Representations in**

**Vector Space**

**Tomas Mikolov**

Google Inc., Mountain View, CA

tmikolov@google.com

**Kai Chen**

Google Inc., Mountain View, CA

kaichen@google.com

**Greg Corrado**

Google Inc., Mountain View, CA

gcorrado@google.com

**Jeffrey Dean**

Google Inc., Mountain View, CA

jeff@google.com

**Abstract**

We propose two novel model architectures for computing continuous vector repre- sentations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

**摘要**

本文提出了两种新的用于计算超大数据集的连续的词向量表示的模型体系结构。这些词表示的质量在一个词相似性任务中进行度量, 并将其结果与以前的最佳技术，即基于不同类型神经网络的技术进行了比较。我们以更低的计算成本得到大幅提高的精度, 也就是说, 从16亿字数据集学习高质量的词向量需要不到一天的时间。此外, 本文还表明, 这些向量在用于测量句法和语义词的相似性的测试集上表现了最先进的性能。