**Distributed Representations of Words and Phrases**

**and their Compositionality**

**Tomas Mikolov** Google Inc.Mountain Viewmikolov@google.com

**Ilya Sutskever** Google Inc.Mountain Viewilyasu@google.com

**Kai Chen** Google Inc.Mountain Viewkai@google.com

**Greg Corrado** Google Inc. Mountain View gcorrado@google.com

**Jeffrey Dean** Google Inc. Mountain View jeff@google.com

**Abstract**

The recently introduced continuous Skip-gram model is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships. In this paper we present several extensions that improve both the quality of the vectors and the training speed. By subsampling of the frequent words we obtain significant speedup and also learn more regular word representations. We also describe a simple alternative to the hierarchical softmax called negative sampling.

An inherent limitation of word representations is their indifference to word order and their inability to represent idiomatic phrases. For example, the meanings of “Canada” and “Air” cannot be easily combined to obtain “Air Canada”. Motivated by this example, we present a

simple method for finding phrases in text, and show that learning good vector representations for millions of phrases is possible.

**摘要**

最近的Skip-gram模型对于学习高质量的分布式向量表示, 从而捕获大量精确的句法和语义词关系而言是一种有效的方法。本文提出了几种提高向量质量和训练速度的扩展方法。通过对高频词采样, 我们获得了显著的加速, 也可以学习到更多的常规词表示。我们还描述了一个简单的分层 softmax 替代，即负采样。

词表示的局限性是他们对语序的忽视和无法表示惯用短语。例如, "加拿大" 和 "空气" 的含义不能很容易地结合起来, 以获得 "加拿大航空"。在这个例子的激励下, 我们提出了一种简单的方法来查找文本中的短语, 并表明用学习好的向量表示数以百万计的短语是可能的。