**Distributed Representations of Sentences and Documents**

**Quoc Le**

**Tomas Mikolov**

QVL@GOOGLE.COM

TMIKOLOV@GOOGLE.COM

Google Inc, 1600 Amphitheatre Parkway, Mountain View, CA 94043

**Abstract**

Many machine learning algorithms require the input to be represented as a fixed-length feature vector. When it comes to texts, one of the most common fixed-length features is bag-of-words. Despite their popularity, bag-of-words features have two major weaknesses: they lose the ordering of the words and they also ignore semantics of the words. For example, “powerful,” “strong” and “Paris” are equally distant. In this paper, we propose Paragraph Vector, an unsupervised algorithm that learns fixed-length feature representations from variable-length pieces of texts, such as sentences, paragraphs, and documents. Our algorithm represents each document by a dense vector which is trained to predict words in the document. Its construction gives our algorithm the potential to overcome the weaknesses of bag-of- words models. Empirical results show that Para- graph Vectors outperform bag-of-words models as well as other techniques for text representations. Finally, we achieve new state-of-the-art results on several text classification and sentiment analysis tasks.

**摘要**

许多机器学习算法要求输入表示为固定长度的特征向量。当涉及到文本时, 最常见的固定长度特征之一就是词袋。尽管它们很受欢迎, 但词袋特征有两个主要缺点: 失去了单词的顺序, 而且还忽略了单词的语义。例如, "强大"、"强" 和 "巴黎" 在该向量空间中的距离同样遥远。本文提出了段落向量, 一种非监督算法，从文本的可变长度片段 (如句子、段落和文档) 中学习固定长度特征表示。我们的算法通过一个稠密向量来表示每个文档, 该向量被训练以预测文档中的单词。它的构造使我们的算法有能力克服词袋模型的弱点。实验结果表明, 在文本表示中, 段落向量优于词袋模型及其他技术。最后, 我们在几个文本分类和情绪分析任务中也实现了新的最先进的结果。