分布式词和短语表示及其组合性

Tomas Mikolov Google Inc.
Mountain View
mikolov@google.com

Ilya Sutskever Google Inc.
山景城
ilyasu@google.com

Kai Chen
谷歌公司
山景城
kai@google.com

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

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

摘要

最近引入的连续 Skip-gram 模型是一种高效的方法，用于学习高质量的分布式向量表示，能够捕捉大量的精确的语法和语义词关系。在本文中，我们提出了几种改进，这些改进既提高了向量的质量，也加快了训练速度。通过频繁词的下采样，我们获得了显著的加速，并且学习到了更规则的词向量表示。我们还描述了一种与分层 softmax 相对简单的替代方法，称为负采样。

词表示的一个固有限制是它们对词序的忽视以及无法表示习语的能力。例如，“加拿大”和“航空”这两个词的意义无法简单地组合成“加拿大航空”。受到这个例子的启发，我们提出了一种简单的方法来在文本中找到短语，并展示了学习数百万个短语的良好向量表示是可能的。

1 引言

在向量空间中分布式表示单词有助于通过将相似的单词分组来使学习算法在自然语言处理任务中获得更好的性能。最早使用单词表示可以追溯到 1986 年，由 Rumelhart、Hinton 和 Williams 提出[13]。这一想法随后在统计语言建模中取得了显著成功[1]。后续工作包括将其应用于自动语音识别和机器翻译[14, 7]，以及广泛的语言处理任务[2, 20, 15, 3, 18, 19, 9]。

最近，Mikolov 等人[8]提出了 Skip-gram 模型，这是一种从大量非结构化文本数据中学习高质量单词向量的有效方法。与之前用于学习单词向量的大多数神经网络架构不同，Skipgram 模型的训练（见图 1）不涉及密集矩阵乘法。这使得训练极其高效：优化后的单机实现可以在一天内训练超过 1000 亿个单词。

使用神经网络计算出的词向量非常有趣，因为学习到的向量明确编码了许多语言规律和模式。令人惊讶的是，许多这些模式可以表示为线性变换。例如，向量计算 vec(“马德里”) - vec(“西班牙”) + vec(“法国”) 的结果与 vec(“巴黎”) 更接近，而不是其他任何词向量 [9, 8]。

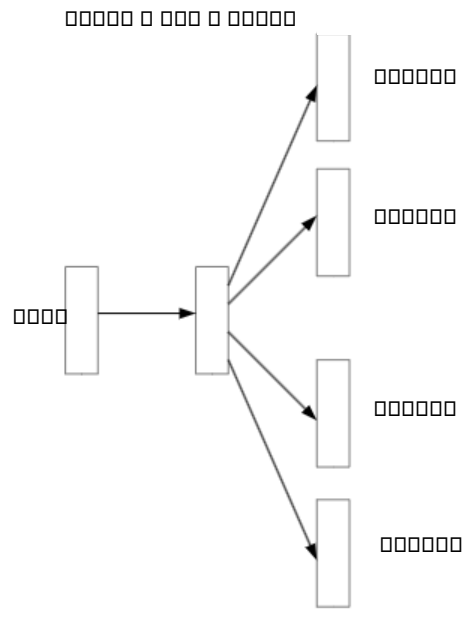

图 1: Skip-gram 模型架构。训练目标是学习能够很好地预测附近词的词向量表示。

在本文中，我们提出了原始 Skip-gram 模型的几个扩展。我们展示了在训练过程中对频繁词进行下采样可以显著提高速度（大约 2 倍 - 10 倍），并且可以提高不频繁词的表示准确性。此外，我们还提出了一种简化版的噪声对比估计（NCE）[4]，用于训练 Skip-gram 模型，这比之前工作中使用的复杂层次 softmax 更快，并且对于频繁词的向量表示更好。

词表示受限于它们无法表示非单个词组成的习语短语。例如，“波士顿环球报”是一份报纸，而“波士顿”和“环球报”的含义组合并不自然。因此，使用向量来表示整个短语使得 Skip-gram 模型更具表达力。其他旨在通过组合词向量来表示句子含义的技术，如递归自编码器 [15]，也会从使用短语向量而不是词向量中受益。

从基于词的模型到基于短语的模型的扩展相对简单。首先，我们使用数据驱动的方法识别大量短语，然后在训练过程中将短语视为单独的标记。为了评估短语向量的质量，我们开发了一组类比推理任务的测试集，其中包含单词和短语。我们测试集中的一个典型类比对是“蒙特利尔”：“蒙特利尔加拿大人队”::“多伦多”：“多伦多枫叶队”。如果 vec(“蒙特利尔加拿大人队”) - vec(“蒙特利尔”) + vec(“多伦多”)的最近表示是 vec(“多伦多枫叶队”)，则认为其回答正确。

最后，我们描述了 Skip-gram 模型的另一个有趣特性。我们发现简单的向量加法通常可以产生有意义的结果。例如，vec(“俄罗斯”) + vec(“河”)接近于 vec(“伏尔加河”)，vec(“德国”) + vec(“首都”)接近于 vec(“柏林”)。这种组合性表明，通过使用单词向量表示的基本数学运算可以获得一定程度的语言理解。

2 Skip-gram 模型

Skip-gram 模型的训练目标是找到有用的词表示，用于预测句子或文档中的周围词。更正式地说，给定一个训练词序列 w, w, w, . . . , w，Skip-gram 模型的目标是最大化平均对数概率

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p(w|w) \tag{1}$$

其中 c 是训练上下文的大小（可以是中心词 w 的函数）。c 越大，训练样本越多，因此可能会提高准确性，但代价是

训练时间。基本的 Skip-gram 公式使用 softmax 函数定义 p(w|w)：

$$p(w|w) = \sum_{w=1}^{} \frac{\exp(vv)}{\exp} \left(\frac{}{vv} \right) \quad (2)$$

其中 v 和 v 分别是 w 的“输入”和“输出”向量表示，W 是词汇表中的单词数。这种公式是不切实际的，因为计算∇ log p(w|w)的成本与 W 成正比，而 W 通常很大（10 到 10 的 10 次方项）。

2.1 分层 softmax

全 softmax 的一个计算效率更高的近似是分层 softmax。在神经网络语言模型的背景下，它最早由 Morin 和 Bengio [12] 提出。主要优势在于，它不需要评估神经网络中的 W 个输出节点来获得概率分布，只需要评估大约 log(W)个节点即可。

分层 softmax 使用一个以 W 个单词为叶子节点的二叉树表示输出层，并且对于每个节点，显式地表示其子节点的相对概率。这些定义了一个随机游走，将概率分配给单词。

更精确地说，每个单词 w 可以通过从树根到 w 的适当路径到达。设 n(w, j)为从根到 w 的路径上的第 j 个节点，设 L(w)为该路径的长度，因此 n(w, 1) = 根且 n(w, L(w)) = w。此外，对于任何内部节点 n，设 ch(n)为 n 的一个任意固定子节点，并设[[x]]为 x 为真时为 1，否则为-1。然后分层 softmax 定义 p(w|w)如下：

$$p(w|w) = \prod_{j=1}^{L(w)-1} \sigma \left([[n(w, j+1) = \mathrm{ch}(n(w, j))]] \cdot vv \right) \quad (3)$$

其中 $\sigma(x) = 1/(1+\exp(-x))$。这意味着计算 $\log p(w|w)$ 的成本与 L(w) 成正比，平均来说不超过 $\sum_{w=1}^{W} \log W$。此外，与标准 softmax Skip-gram 表述中为每个词 w 分配两个表示 v 和 v 不同，分层 softmax 表述为每个词 w 只有一个表示 v，为二叉树的每个内部节点 n 也有一个表示 v。

分层 softmax 使用的树结构对性能有很大影响。Mnih 和 Hinton 探索了多种构建树结构的方法，并研究了这些方法对训练时间和最终模型准确率的影响 [10]。在我们的工作中，我们使用二叉 Huffman 树，因为它为频繁词分配了较短的代码，从而加快了训练速度。之前的研究表明，通过频率对词进行分组作为神经网络语言模型的一种非常简单的加速技术效果很好 [5, 8]。

2.2 负采样

层次 softmax 的一个替代方法是噪声对比估计（NCE），它由 Gutmann 和 Hyvarinen [4] 提出，并由 Mnih 和 Teh [11] 应用于语言建模。NCE 假设一个好的模型应该能够通过逻辑回归区分数据和噪声。这类似于 Collobert 和 Weston [2] 使用的 hinge loss，他们通过将数据排名在噪声之上来训练模型。

虽然 NCE 可以证明大约最大化 softmax 的对数概率，但 Skipgram 模型只关心学习高质量的向量表示，因此我们可以简化 NCE，只要向量表示保持其质量。我们通过目标定义负采样（NEG）

$$\log \sigma(vv) + \sum_{i=1}^{k} E \left[\log \sigma(-vv) \right] \quad (4)$$

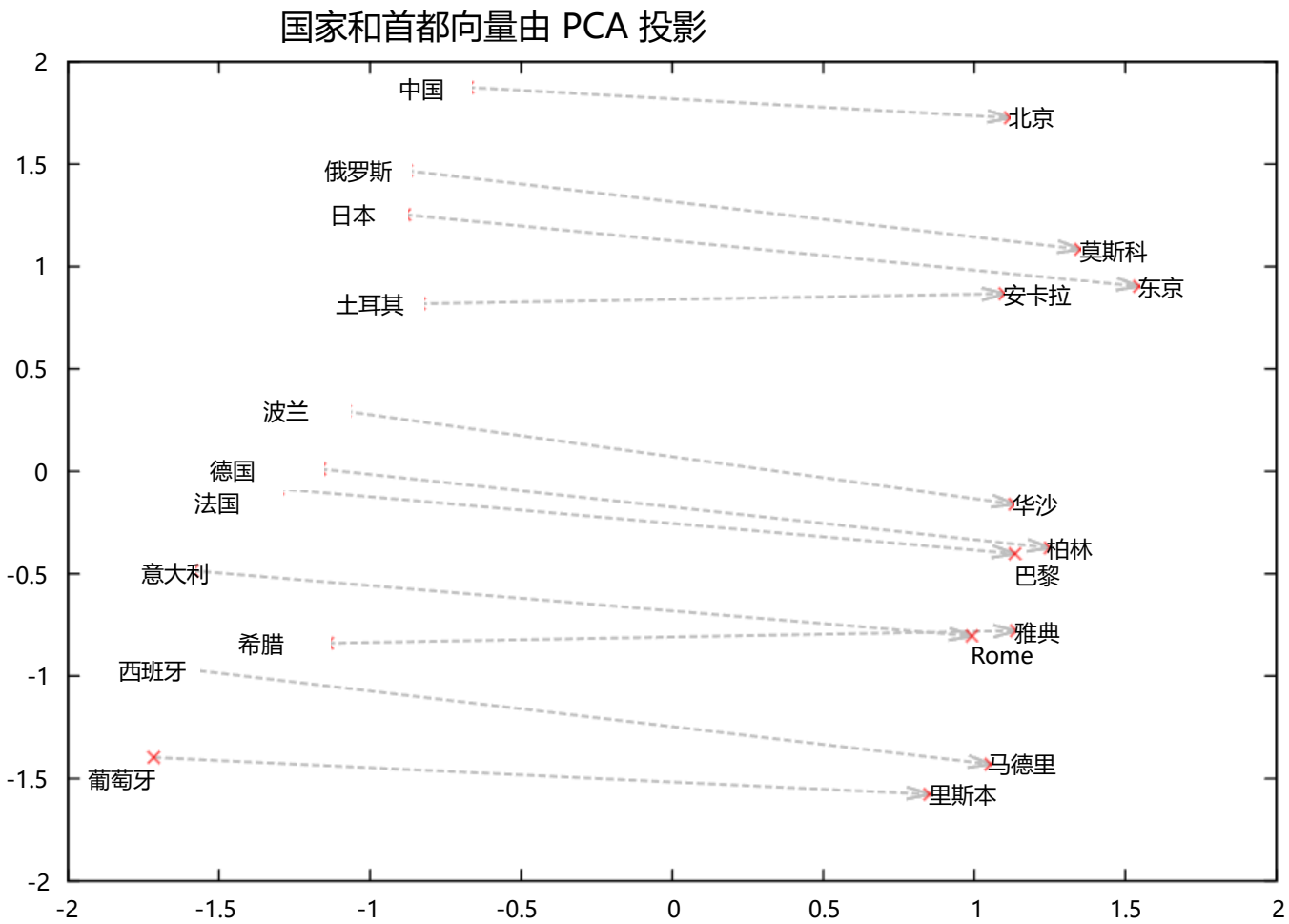

图 2：1000 维 Skip-gram 向量的国家及其首都城市的二维 PCA 投影。该图展示了模型自动组织概念并隐式学习它们之间关系的能力，在训练过程中我们并未提供任何关于首都城市含义的监督信息。

该公式用于替换 Skip-gram 目标中的每个 log P (w|w)项。因此，任务是使用逻辑回归区分目标词 w 与噪声分布 P(w)的抽样，其中每个数据样本有 k 个负样本。我们的实验表明，对于小训练数据集，k 值在 5-20 之间是有用的，而对于大数据集，k 值可以小至 2-5。与 NCE 相比，负采样的主要区别在于 NCE 需要样本和噪声分布的数值概率，而负采样仅使用样本。虽然 NCE 大约最大化了 softmax 的对数概率，但这一特性对我们应用来说并不重要。

Both NCE 和 NEG 都有噪声分布 P(w)作为自由参数。我们调查了 P(w)的几种选择，并发现四分之三次幂的单词分布 U(w)（即 U(w)/Z）在我们尝试的所有任务中，包括语言建模（未在此报告）中，显著优于单词分布和均匀分布，无论是 NCE 还是 NEG 都表现得更好。

2.3 频繁词下采样

在非常大的语料库中，最常见的词可能会出现数亿次（例如，“in”，“the”和“a”）。这类词通常提供的信息价值低于罕见词。例如，虽然 Skip-gram 模型可以从观察“France”和“Paris”的共现中受益，但它从观察“France”和“the”的频繁共现中受益较少，因为几乎每个词在句子中都会频繁地与“the”共现。这个想法也可以反过来应用；频繁词的向量表示在训练几百万个例子后不会显著变化。

为了平衡罕见词和常见词之间的不平衡，我们使用了一个简单的下采样方法：训练集中的每个词以由公式计算的概率被丢弃

$$P(w) = 1 - \sqrt{\frac{t}{f(w)}} \tag{5}$$

Method	Time [min]	Syntactic [%]	Semantic [%]	Total accuracy [%]
NEG-5	38	63	54	59
NEG-15	97	63	58	61
HS-Huffman	41	53	40	47
NCE-5	38	60	45	53
以下结果使用了 10 次下采样				
NEG-5	14	61	58	60
NEG-15	36	61	61	61
HS-Huffman	21	52	59	55

表 1：在[8]中定义的类比推理任务上，各种 Skip-gram 300 维模型的准确性。NEG-k 表示每个多正样本有 k 个负样本的负采样；NCE 表示噪声对比估计，HS-Huffman 表示基于频率的霍夫曼编码的分层 softmax。

其中 f(w) 是单词 w 的频率，t 是一个选定的阈值，通常约为 10。我们选择了这个下采样公式，因为它会大量下采样频率大于 t 的单词，同时保持频率的排名。尽管这个下采样公式是通过启发式方法选择的，但我们发现它在实践中效果很好。它加速了学习过程，并且甚至显著提高了稀有词学习向量的准确性，这将在后面的章节中展示。

3 实验结果

在本节中，我们评估了分层 softmax（HS）、噪声对比估计、负采样以及训练词的子采样。我们使用了 Mikolov 等人[8]引入的类比推理任务。该任务包括类似“Germany” : “Berlin” :: “France” : ?这样的类比，通过找到一个向量 x，使得 vec(x)与 vec(“Berlin”) - vec(“Germany”) + vec(“France”)根据余弦距离最接近来解决。我们从搜索中丢弃输入词。如果 x 是“Paris”，则认为这个特定的例子被正确回答。该任务分为两大类：语法类比（如“quick” : “quickly” :: “slow” : “slowly”）和语义类比，例如国家与其首都的关系。

为了训练 Skip-gram 模型，我们使用了一个包含各种新闻文章的大数据集（一个内部 Google 数据集，包含一亿个单词）。我们从词汇表中丢弃了在训练数据中出现次数少于 5 次的所有单词，这导致词汇表的大小为 692K。表 1 报告了各种 Skip-gram 模型在词类比测试集上的性能。表格显示，Negative Sampling 在类比推理任务上的表现优于 Hierarchical Softmax，并且甚至在 Noise Contrastive Estimation 上稍微好一些。频繁词的下采样可以大大提高训练速度，并使词表示更加准确。

有人认为 Skip-gram 模型的线性性使其向量更适合此类线性类比推理，但 Mikolov 等人[8]的结果也表明，标准的 Sigmoid 递归神经网络（高度非线性的）在训练数据量增加时在该任务上的表现显著提高，这表明非线性模型也倾向于词表示的线性结构。

4 学习短语

如前所述，许多短语的意义并不是其各个单词意义的简单组合。为了学习短语的向量表示，我们首先找到经常一起出现的单词，并且在其他上下文中不经常出现。例如，“纽约时报”和“多伦多枫叶队”在训练数据中被替换为唯一的标记，而双词短语“this is”则保持不变。

[1] code.google.com/p/word2vec/source/browse/trunk/questions-words.txt

报纸	
纽约 《纽约时报》 巴尔的摩 《巴尔的摩太阳报》 圣何塞 圣何塞水星报 辛辛那提 《辛辛那提邮报》	
NHL 球队	
波士顿 波士顿棕熊 蒙特利尔 蒙特利尔加拿大人	凤凰城 亚利桑那郊狼 纳什维尔 纳什维尔掠夺者
NBA 球队	
底特律 底特律活塞 多伦多 多伦多猛龙 奥克兰金州勇士队 膜城 膜城灰熊队	
航空公司	
奥地利 奥地利航空 西班牙 西班牙航空 比利时	布鲁塞尔航空 希腊爱琴海航空
公司高管	
史蒂夫·鲍尔默 微软 拉里·佩奇 谷歌 萨缪尔·J·帕米萨诺	国际商业机器公司 文森特·沃吉尔斯 亚马逊

表 2：短语类类比推理任务的示例（完整测试集包含 3218 个示例）。目标是使用前三者计算出第四个短语。我们的最佳模型在该数据集上的准确率为 72%。

这样，我们可以在不大幅增加词汇量的情况下形成许多合理的短语；理论上，我们可以使用所有 n-克 gram 训练 Skip-gram 模型，但那将非常耗费内存。许多技术已经开发出来用于识别文本中的短语；然而，超出我们工作的范围去比较它们。我们决定使用一种简单的数据驱动方法，根据单克 gram 和双克 gram 的计数形成短语，使用

$$\text{score}(w, w) = \frac{\text{count}(ww) - \delta}{\text{count}(w) \times \text{count}(w)}. \tag{6}$$

δ 用作折扣系数，防止形成太多由非常罕见的词组成的短语。得分高于所选阈值的双词短语随后被用作短语。通常，我们在训练数据上运行 2-4 次迭代，每次迭代阈值减小，允许由多个词组成的更长的短语形成。我们使用一种新的类比推理任务来评估短语表示的质量，该任务涉及短语。表 2 显示了此任务中使用的五类类比的示例。此数据集已在网上公开。

4.1 短语 Skip-Gram 结果

与之前的实验使用相同的新闻数据，我们首先构建了基于短语的训练语料库，然后使用不同的超参数训练了几种 Skip-gram 模型。如前所述，我们使用向量维度 300 和上下文大小 5。这种设置已经在短语数据集上实现了良好的性能，并允许我们快速比较负采样和分层 softmax，包括对频繁词进行子采样的情况。

结果总结在表 3 中。

结果显示，即使在 k = 5 的情况下，Negative Sampling 也能达到不错的准确率，而使用 k = 15 则能实现显著更好的性能。令人惊讶的是，虽然我们在不使用下采样时发现 Hierarchical Softmax 的性能较低，但在对频繁词汇进行下采样后，它反而成为了表现最好的方法。这表明下采样不仅可以加快训练速度，还能提高准确率，至少在某些情况下是这样的。

[2] code.google.com/p/word2vec/source/browse/trunk/questions-phrases.txt

方法	维度	无下采样 [%]	10 下采样 [%]	NEG-5
	300	24		27
NEG-15	300	27		42
HS-Huffman	300	19		47

表 3：在短语类比数据集上 Skip-gram 模型的准确率。模型在新闻数据集约十亿单词上进行了训练。

NEG-15 with 10subsampling Lingsugur 意大利探险家 贝加尔湖 大裂谷 阿拉尔湖 阿兰·贝恩 里贝卡 纳奥米 月球漫步者 爱奥尼亚海 勒根 爱奥尼亚群岛 国际象棋大师 国际象棋特级大师 加里·卡斯帕罗夫	HS with 10subsampling	Vasco de Gama 朗基

表 4：使用两种不同模型对给定短语的最接近实体示例。

捷克 + 币种越南 + 首都德国 + 航空公司俄罗斯 + 河流法国 + 女演员				
克朗 胡志明市 航空公司 Lufthansa 莫斯科 杰丽特·比诺切 检查克朗 胡志明市航空公司 Lufthansa 伏尔加河 凡妮莎·帕拉迪丝 波兰 黄色 越南国旗航空公司 Lufthansa 沿河夏洛特·甘斯布 CTK 越南航空公司 Lufthansa 俄罗斯 西莉尔·德				

表 5：向量组合性使用逐元素相加。显示了使用最佳 Skip-gram 模型的两个向量之和的四个最接近的标记。

为了在短语类比任务中最大化准确性，我们通过使用包含约 330 亿个词的数据集增加了训练数据量。我们使用了分层 softmax，维度为 1000，以及整个句子作为上下文。这导致模型的准确率达到 72%。当我们把训练数据集的大小减少到 6 亿个词时，准确率下降到 66%，这表明大量的训练数据至关重要。

为了进一步了解不同模型学习的不同表示之间有多大差异，我们手动检查了不同模型中稀见短语的最近邻。在表 4 中，我们展示了这种比较的一个样本。与之前的成果一致，似乎使用分层 softmax 和采样的模型学习到的短语表示是最好的。

5 加性组合性

我们证明了 Skip-gram 模型学习到的词和短语表示具有线性结构，使得可以通过简单的向量算术来进行精确的类比推理。有趣的是，我们发现 Skip-gram 表示还表现出另一种线性结构，使得可以通过词向量表示的元素级相加来有意义地组合词。这一现象在表 5 中得到了说明。

向量的加法性质可以通过检查训练目标来解释。词向量与 softmax 非线性输入之间存在线性关系。随着词向量被训练以预测句子中的相邻词，这些向量可以被视为表示词出现的上下文分布。这些值与输出层计算的概率成对数关系，因此两个词向量的和与两个上下文分布的乘积有关。这里的乘积作为 AND 函数工作：由两个词向量都赋予高概率的词将具有高概率，而其他词将具有低概率。因此，如果“伏尔加河”经常与“俄罗斯”和“河”一起出现在同一句子中，这两个词向量的和将产生一个特征向量，该向量接近“伏尔加河”的向量。

6 与已发表的词表示方法比较

Many authors who previously worked on the neural network based representations of words have published their resulting models for further use and comparison: amongst the most well known authors are Collobert and Weston [2], Turian et al. [17], and Mnih and Hinton [10]. We downloaded their word vectors from the web. Mikolov et al. [8] have already evaluated these word representations on the word analogy task, where the Skip-gram models achieved the best performance with a huge margin.

[3] http://metaoptimize.com/projects/wordreprs/

Model Redmond Havel 忍术 喷漆 放弃（训练时间）					
Collobert (50d)	conyers plauen reiki cheesecake abdicate (2 个月)		lubbock dzerzhinsky kohona gossip accede		
	keene	奥地利	空手道	dioramas rearm	
Turian (200d)	McCarthy Jewell - 枪火 (几周)		Alston Arzu - 情绪 -		
	Cousins Ovitz - 逍遥法外 -				
Mnih (100d)	Podhurst Pontiff - 麻醉剂	Mavericks (7 天)	Harlang Pinochet - 猴子计划		
	Agarwal	Rodionov	-	犹犹豫豫	
Skip-Phrase 红蒙蒙华盛顿 Vaclav Havel 暗夜行者 喷漆 投降（1000d，1 天） 红蒙蒙华盛顿 总统 Vaclav Havel 武术 喷涂 投降					
	微软	Velvet Revolution 剑术标签者投降			

表 6：给定各种知名模型和基于短语训练的 Skip-gram 模型，最接近的令牌示例。空单元格意味着该词不在词汇表中。

为了更深入地了解学习向量质量的差异，我们通过表 6 中不常见词的最近邻来提供实证比较。这些示例表明，基于大型语料库训练的大 Skip-gram 模型在学习表示的质量上明显优于其他所有模型。这在一定程度上可以归因于该模型训练的数据量约为 300 亿个词，比以往工作中的典型数据量大两个到三个数量级。有趣的是，尽管训练集更大，但 Skip-gram 模型的训练时间只是之前模型架构所需时间复杂度的一小部分。

7 结论

这项工作有几个关键贡献。我们展示了如何使用 Skip-gram 模型训练词和短语的分布式表示，并证明了这些表示具有线性结构，使得精确的类比推理成为可能。本文介绍的技术也可以用于训练[8]中介绍的连续词袋模型。

我们成功地在比之前发布的模型多几个数量级的数据上训练了模型，这得益于计算效率高的模型架构。这导致了学习到的词和短语表示的质量有了很大的提高，尤其是在稀有实体方面。我们还发现，对频繁词进行采样不仅加快了训练速度，而且显著提高了不常见词的表示质量。我们论文的另一个贡献是负采样算法，这是一种极其简单的训练方法，特别适用于学习频繁词的准确表示。

训练算法的选择和超参数选择是任务特定的决策，因为我们发现不同的问题有不同的最优超参数配置。在我们的实验中，对性能影响最大的决策是模型架构的选择、向量的大小、子采样率以及训练窗口的大小。

这项工作的一个非常有趣的结果是，仅仅使用简单的向量加法就可以对词向量进行一定程度的有意义的组合。这篇论文中提出的另一种学习短语表示的方法是简单地用一个标记来表示短语。将这两种方法结合起来，提供了一种强大而简单的表示较长文本的方法，同时具有最小的计算复杂度。因此，我们的工作可以被视为与现有的试图使用递归矩阵-向量操作来表示短语的方法互补[16]。

基于本文描述的技术，我们提供了训练词和短语向量的代码作为开源项目。

[4] code.google.com/p/word2vec

参考文献

[1] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 一种神经概率语言模型。《机器学习研究杂志》，3:1137–1155，2003. [2] Ronan Collobert and Jason Weston. 统一架构用于自然语言处理：深度神经-

多任务学习的递归神经网络。在第 25 届国际机器学习会议论文集，第 160-167 页。ACM，2008。

[3] Xavier Glorot, Antoine Bordes, 和 Yoshua Bengio. 大规模情感分类的领域适应：一种深度学习方法。在 ICML，513–520，2011。[4] Michael U Gutmann 和 Aapo Hyv¨arinen. 未规范化统计模型的噪声对比估计及其在自然图像统计中的应用。《机器学习杂志》，13:307–361，2012。[5] Tomas Mikolov, Stefan Kombrink, Lukas Burget, Jan Cernocky, 和 Sanjeev Khudanpur. 扩展

递归神经网络语言模型。在 2011 年 IEEE 国际声学、speech 和信号处理会议（ICASSP）上，页码 5528–5531。IEEE，2011。

[6] Tomas Mikolov, Anoop Deoras, Daniel Povey, Lukas Burget 和 Jan Cernocky. 大规模神经网络语言模型的策略。在 2011 年自动语音识别与理解会议论文集上发表。[7] Tomas Mikolov. 基于神经网络的统计语言模型。博士论文，布拉格技术大学博士论文，2012 年。[8] Tomas Mikolov, Kai Chen, Greg Corrado 和 Jeffrey Dean. 在向量空间中高效估计词表示。ICLR 工作坊，2013 年。[9] Tomas Mikolov, Wen-tau Yih 和 Geoffrey Zweig. 连续空间词表示中的语言规律。在 NAACL HLT 会议上发表，2013 年。

[10] Andriy Mnih 和 Geoffrey E Hinton. 一种可扩展的分层分布式语言模型。神经信息处理系统进展，21:1081–1088，2009。[11] Andriy Mnih 和 Yee Whye Teh. 一种快速简单的神经概率语言模型训练算法。arXiv 预印本 arXiv:1206.6426，2012。[12] Frederic Morin 和 Yoshua Bengio. 分层概率神经网络语言模型。在国际人工智能与统计研讨会论文集上，页码：246–252，2005。[13] David E Rumelhart, Geoffrey E Hintont, 和 Ronald J Williams. 通过反向传播误差学习表示。自然，323(6088):533–536，1986。

[14] Holger Schwenk. 连续空间语言模型. 计算机语音和语言，第 21 卷，2007 年。[15] Richard Socher, Cliff C. Lin, Andrew Y. Ng, and Christopher D. Manning. 解析自然场景和

自然语言处理中的递归神经网络。在第 26 届国际机器学习会议（ICML）论文集，第 2 卷，2011 年。

[16] Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 语义组合性
通过递归矩阵-向量空间。在 2012 年会议论文集计算语言学实证方法研讨会（EMNLP）上，2012 年。

[17] Joseph Turian, Lev Ratinov, 和 Yoshua Bengio. 词表示：一种简单而通用的方法
半监督学习。在第 48 届计算语言学协会年会论文集上，第 384-394 页。计算语言学协会，2010 年。

[18] Peter D. Turney 和 Patrick Pantel. 从频率到意义：语义向量空间模型。在人工智能研究杂志上，第 37 卷：141-188，2010 年。

[19] Peter D. Turney. 分布式语义超越单词：基于类比和同义词的监督学习。
在计算语言学协会事务（TACL）上，第 353-366 页，2013 年。

[20] Jason Weston, Samy Bengio, 和 Nicolas Usunier. Wsabie：大规模图像标注的扩展。
在第二十二届国际联合人工智能会议论文集-卷三，第 2764-2770 页。AAAI 出版社，2011 年。