词语分布式表达及其组合性

托马斯米科洛夫
谷歌公司
山顶风光
mikolov@google.com

Ilya Susskever
谷歌公司
山顶风光
ilyasu@google.com

陈凯
谷歌公司山景城
kai@google.com

格雷格科拉多
谷歌公司
山顶风光
gcorrado@google.com

杰弗里迪恩
谷歌公司
山顶风光
jeff@google.com

摘要

最近推出的连续Skip-gram模型是学习高质量分布式矢量表示的有效方法，可以捕获大量精确的句法和语义单词关系。在本文中，我们提出了几个扩展，可以提高向量的质量和训练速度。通过对频繁单词进行二次取样，我们获得了显着的加速，并且还学习了更多的常规单词表示。我们还描述了一种称为负采样的分层softmax的简单替代方案。

单词表示的固有局限性是它们对单词顺序的漠不关心以及它们无法表示惯用语。例如，“加拿大”和“空气”的含义不能轻易组合以获得“加拿大航空”。在这个例子的推动下，我们提出了一种在文本中查找短语的简单方法，并表明可以为数百万个短语学习好的矢量表示。

1 介绍

向量空间中的单词的分布式表示有助于学习算法通过对相似的单词进行分组来在自然语言处理任务中实现更好的性能。由于Rumelhart，Hinton和Williams [13]，最早使用单词表示的一个可以追溯到1986年。这个想法已经应用于统计语言建模并取得了相当大的成功[1]。后续工作包括自动语音识别和机器翻译的应用[14,7]，以及各种NLP任务[2,20,15,3,18,19,9]。

最近，Mikolov等人。[8]介绍了Skip-gram模型，这是一种从大量非结构化文本数据中学习单词的高质量矢量表示的有效方法。与大多数先前使用的用于学习单词向量的神经网络架构不同，Skip-gram模型的训练（参见图1）不涉及密集矩阵乘法。这使得培训非常有效：优化的单机实施可以在一天内培训超过1000亿个单词。

使用神经网络计算的单词表示非常有趣，因为学习的向量明确地编码许多语言规则和模式。有些令人惊讶的是，许多这些模式可以表示为线性翻译。例如，向量计算vec（“Madrid”）- vec（“Spain”）+ vec（“France”）的结果更接近vec（“Paris”）而不是任何其他单词向量[9,8]。

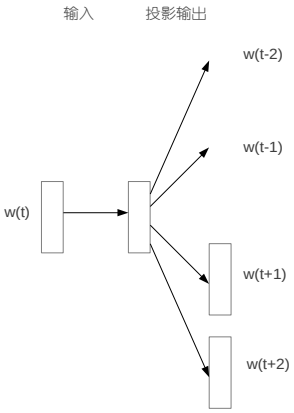

图1：Skip-gram模型架构。培训目标是学习擅长预测附近单词的单词矢量表示。

在本文中，我们提出了原始Skip-gram模型的几个扩展。我们表明，在训练期间频繁词的子采样导致显着的加速（大约2x-10x），并且提高了频率较低的词的表示的准确性。此外，我们提出了一个简化的噪声对比度估计（NCE）变体[4]，用于训练Skip-gram模型，与先前使用的更复杂的分层softmax相比，可以更快地训练和更好的矢量表示频繁的单词。工作[8]。

单词表示受限于它们无法表示不是单个单词的组成的惯用短语。例如，“波士顿环球报”是一份报纸，因此它不是“波士顿”和“环球报”含义的自然组合。因此，使用向量来表示整个短语使得Skip-gram模型更具表现力。旨在通过组合单词向量来表示句子含义的其他技术，例如递归自动编码器[15]，也将受益于使用短语向量而不是单词向量。

从基于单词的模型到基于短语的模型的扩展相对简单。首先，我们使用数据驱动方法识别大量短语，然后在训练期间将短语视为单独的标记。为了评估短语向量的质量，我们开发了一组包含单词和短语的类比推理任务。我们测试集中的典型类比对是“蒙特利尔”：“蒙特利尔加拿大人队”::“多伦多”：“多伦多枫叶队”。如果对vec（“蒙特利尔加拿大人”）- vec（“蒙特利尔”）+ vec（“多伦多”）的最近代表是vec（“多伦多枫叶队”），则认为已经正确回答。

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（“河流”）接近vec（“伏尔加河”），vec（“德国”）+ vec（“大写”）接近vec（“柏林”）。这种组合性表明，通过对单词矢量表示使用基本数学运算，可以获得非显而易见的语言理解程度。

2 Skip-gram模型

Skip-gram模型的训练目标是找到对预测句子或文档中的周围单词有用的单词表示。更正式地，给定一系列训练单词w_1，w_2，w_3，...，w_t，Skip-gram模型的目标是最大化平均对数概率

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c,\ j\neq 0}\log p(w_{t+j}|w_t) \tag{1}$$

其中c是训练上下文的大小（可以是中心词w_t的函数）。较大的c导致更多的训练样例，因此可以导致更高的准确性，但代价是

训练时间。基本的Skip-gram公式使用softmax函数定义$p(w_{t+j}|w_t)$：

$$p(w_O|w_I) = \frac{\exp(v'^{\top}_{w_O} v_{w_I})}{\sum_{w=1}^{W} \exp(v'^{\top}_{w} v_{w_I})} \tag{2}$$

其中v_w和v'_w是w的“输入”和“输出”矢量表示，W是数字词汇中的单词。由于计算成本，这种表述是不切实际的$\nabla \log p(w_o|w_i)$与W成正比，W通常很大（10^5-10^7项）。

2.1 分层Softmax

完全softmax的计算上有效的近似是分层softmax。在神经网络语言模型的背景下，它首先由Morin和Bengio引入[12]。主要优点是，不是评估神经网络中的W输出节点来获得概率分布，而是仅需要评估$\log_2(W)$节点。

分层softmax使用输出层的二叉树表示，其中W字作为其叶，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，它将概率分配给单词。

更确切地说，每个单词w可以通过来自树根的适当路径到达。设$n(w, j)$为从根到w的路径上的第j个节点，让$L(w)$为该路径的长度，因此$n(w, 1) = root$和$n(w, L(w)) = w$。另外，对于任何内部节点n，令$ch(n)$为n的任意固定子节点，如果x为真，则令$[x]$为1，否则为-1。然后，分层softmax定义$p(w_o|w_i)$如下：

$$p(w|w_I) = \prod_{j=1}^{L(w)-1} \sigma([n(w, j+1) = ch(n(w, j))] \cdot v'^{\top}_{n(w,j)} v_{w_I}) \tag{3}$$

其中$\sigma(x) = 1/(1+\exp(-x))$。可以证实$\sum_{w=1}^{W} p(w|w_i) = 1$.这意味着计算$\log p(w_o|w_i)$和$\nabla \log p(w_o|w_i)$的成本。$w_i$）与$L(w_o)$成比例，其平均不大于$\log W$。此外，不同于Skip-gram的标准softmax配方为每个单词w分配两个表示v_w和v'_w，分层softmax公式具有每个字w的一个表示v_w和每个字的每个内部节点n的一个表示v'_n二叉树。

分层softmax使用的树的结构对性能具有相当大的影响。Mnih和Hinton探索了许多构建树结构的方法以及对训练时间和模型精度的影响[10]。在我们的工作中，我们使用二进制霍夫曼树，因为它为频繁的单词分配短代码，从而导致快速训练。之前已经观察到，通过它们的频率将单词分组在一起很好地作为基于神经网络的语言模型的非常简单的加速技术[5,8]。

2.2 负抽样

分层softmax的替代方案是噪声对比度估计（NCE），由Gutmann和Hyvarinen [4]引入，并应用于Mnih和Teh的语言建模[11]。NCE认为一个好的模型应该能够通过逻辑回归将数据与噪声区分开来。这类似于Collobert和Weston [2]使用的铰链损耗，他通过将数据排在噪声之上来训练模型。

虽然NCE可以显示为近似最大化softmax的对数概率，但Skip-gram模型仅关注学习高质量向量表示，因此只要向量表示保持其质量，我们就可以自由地简化NCE。我们通过目标定义负抽样（NEG）

$$\log \sigma(v'^{\top}_{w_O} v_{w_I}) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} \log \sigma(-v'^{\top}_{w_i} v_{w_I}) \tag{4}$$

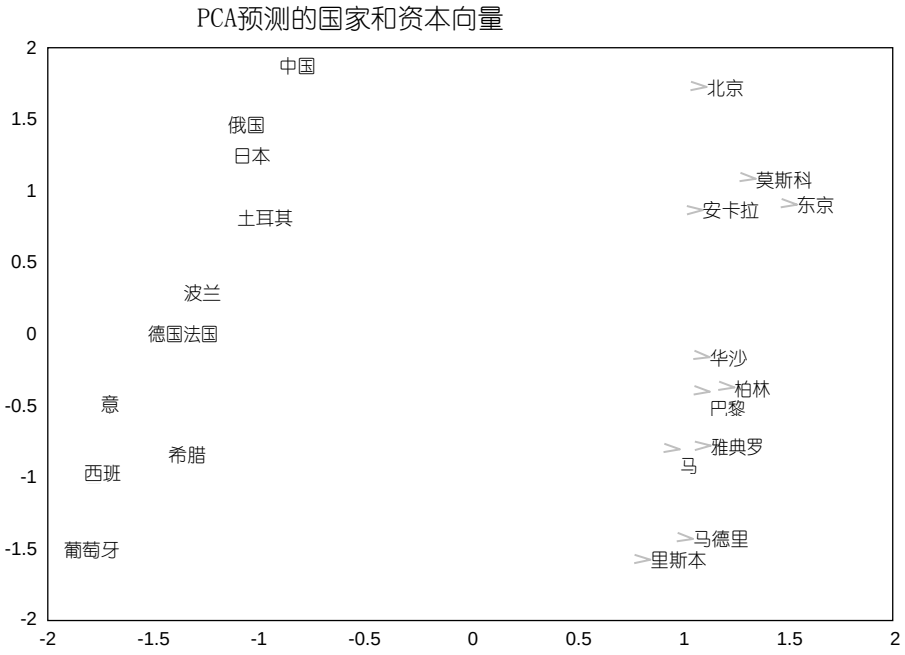

图2：国家及其首府城市的1000维Skip-gram矢量的二维PCA投影。该图说明了模型自动组织概念和隐含地学习它们之间关系的能力，因为在培训期间我们没有提供有关首都城市意义的任何监督信息。

用于替换Skip-gram目标中的每个log P（w_o| w_i ）项。因此，任务是使用逻辑回归将目标词w_o与来自噪声分布P_n（w）的绘制区分开，其中对于每个数据样本存在k个负样本。我们的实验表明了价值观
5-20范围内的k对于小型训练数据集是有用的，而对于大型数据集，k可以小到2-5。负抽样和NCE之间的主要区别在于NCE需要样本和噪声分布的数值概率，而负抽样仅使用样本。虽然NCE近似最大化softmax的对数概率，但这个属性对我们的应用并不重要。

NCE和NEG都具有噪声分布P_n（w）作为自由参数。我们研究了P_n（w）的许多选择，发现单字母分布U（w）上升到3/4幂（即U（w）$^{3/4}$/ Z）显着优于单字组和NCE和NEG在我们尝试的每项任务中均匀分布，包括语言建模（此处未报告）。

2.3 频繁词的子采样

在非常大的语料库中，最频繁的单词很容易发生数亿次（例如，“in”，“the”和“a”）。这些词通常提供的信息价值低于稀有词。例如，虽然Skip-gram模型受益于观察“法国”和“巴黎”的共同出现，但是通过观察“法国”和“the”的频繁共现而得益更少，因为几乎每个词都是co - 经常在句子中出现“the”。这个想法也可以在相反的方向上应用:经过数百万个例子的训练后，频繁词的矢量表示没有显着变化。

为了抵消罕见和频繁词之间的不平衡，我们使用了一种简单的子采样方法：训练集中的每个词w_i 被丢弃，概率由公式计算

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

方法	时间[分钟]	句法[%]	语义[%]	总准确度[%]
负5	38	63	54	59
负15	97	63	58	**61**
HS-霍夫曼	41	53	40	47
NCE-5	38	60	45	53
以下结果使用10^{-5}子采样				
负5	14	61	58	60
负15	36	61	61	**61**
HS-霍夫曼	21	52	59	55

表1：各种Skip-gram 300维模型对[8]中定义的类比推理任务的准确性。NEG-k代表负取样，每个阳性样品有k个阴性样品;NCE代表噪声对比度估计，HS-Huffman代表具有基于频率的霍夫曼码的Hierarchical Softmax。

其中$f(w_i)$是字w_i的频率，t是选择的阈值，通常约为10^{-5}。我们选择了这个子采样公式，因为它积极地对频率大于t的单词进行子采样，同时保留频率的排名。虽然这个子样本 - 启发式选择了穆拉，我们发现它在实践中运作良好。它可以加速学习，甚至可以显着提高稀有单词的学习向量的准确性，如以下部分所示。

3 实验结果

在本节中，我们评估训练字的分层Softmax（HS），噪声对比度估计，负采样和子采样。我们使用Mikolov等人介绍的类比推理任务[1]。[8]。任务包括类似“德国”：“柏林”::“法国”：？，通过找到向量x使得vec（x）最接近vec（“柏林”） - vec（“德国” “） + vec（“France”）根据余弦距离（我们丢弃搜索中的输入词）。如果x是“Paris”，则认为该特定示例已被正确回答。该任务有两大类：句法类比（如“快速”：“快速”::“慢”：“慢”）和语义类比，如国家与首都关系。

为了训练Skip-gram模型，我们使用了一个由各种新闻文章组成的大型数据集（一个包含十亿字的内部Google数据集）。我们从词汇表中丢弃了在训练数据中发生少于5次的所有单词，这导致了大小为692K的词汇。表1中报告了各种Skip-gram模型在单词类比测试集上的表现。该表显示负抽样在类比推理任务上优于Hierarchical Softmax，并且甚至比噪声对比估计具有更好的性能。频繁词的子采样几次提高了训练速度，使得词表示更加准确。

可以认为，skip-gram模型的线性使得它的向量更适合于这种线性类比推理，但是Mikolov等人的结果。[8]也表明，标准的S形回归神经网络（高度非线性）学习的向量在训练数据量增加时显着改善了这项任务，这表明非线性模型也偏好于单词表示的线性结构。

4 学习短语

如前所述，许多短语的含义不是其单个词的含义的简单组合。为了学习短语的矢量表示，我们首先找到经常出现在一起的单词，而不是偶尔出现在其他情境中。例如，“纽约时报”和“多伦多枫叶队”在训练数据中被独特的令牌所取代，而“这是”的二元组将保持不变。

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

纽约圣何塞	纽约时报圣何塞水星报	巴尔的摩辛辛那提	巴尔的摩太阳辛辛那提询问者
	报纸		
	NHL团队s		
波士顿	波士顿棕熊队	蒙特利尔	蒙特利尔加拿大人队
凤凰	凤凰郊狼	纳什维尔	纳什维尔掠夺者
	NBA球队s		
底特律	底特律活塞队	多伦多	多伦多猛龙队
奥克兰	金州勇士队	孟菲斯	孟菲斯灰熊队
	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
	公司高管		
史蒂夫鲍尔默	微软	拉里佩奇	谷歌
明盛	IBM	沃纳 威格尔	亚马逊

表2：短语的类比推理任务的示例（完整测试集具有3218个示例）。目标是使用前三个计算第四个短语。我们的最佳模型在此数据集上的准确率达到了72%。

这样，我们可以形成许多合理的短语，而不会大大增加词汇量;从理论上讲，我们可以使用所有n-gram来训练Skip-gram模型，但这会过于记忆密集。先前已经开发了许多技术来识别文本中的短语;但是，比较它们超出了我们的工作范围。我们决定使用一种简单的数据驱动方法，使用unigram和bigram计数形成短语

$$\text{score}(w_i, w_j) = \frac{\text{计数}(w_i w_j) - \delta}{\text{计数}(w_i) \times \text{计数}(w_j)}. \tag{6}$$

δ 用作贴现系数，并且防止形成由非常罕见的单词组成的太多短语。然后将得分高于所选阈值的双胞胎用作短语。通常，我们对训练数据进行2-4次传递，阈值降低，允许形成由多个单词组成的较长短语。我们使用涉及短语的新类比推理任务来评估短语表示的质量。表2显示了此任务中使用的五类类比的示例。该数据集可在网上公开获取[2]。

4.1 短语Skip-Gram结果

从与之前实验相同的新闻数据开始，我们首先构建了基于短语的训练语料库，然后我们使用不同的超参数训练了几个Skip-gram模型。和以前一样，我们使用了矢量维度300和上下文大小5.这个设置已经在短语数据集上实现了良好的性能，并且允许我们快速比较负抽样和分层Softmax，无论是否有频繁令牌的子采样。结果总结在表3中。

结果表明，尽管负采样即使k = 5也能达到可观的精度，但使用k = 15可以获得相当好的性能。令人惊讶的是，虽然我们发现Hierarchical Softmax在没有子采样的情况下训练时可以达到较低的性能，但是当我们对频繁的单词进行下采样时，它成为表现最佳的方法。这表明，子采样可以导致更快的训练并且还可以提高准确性，至少在某些情况下如此。

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

方法	维数	没有子采样[%]	10^{-5} 子采样[%]
负5	300	24	27
负15	300	27	42
HS-霍夫曼	300	19	**47**

表3：短语类比数据集上的Skip-gram模型的准确性。这些模型接受了来自新闻数据集的大约10亿个单词的训练。

	NEG-15与10^{-5} 子采样	HS与10^{-5} 子采样
瓦斯科德加马	林格苏古尔	意大利探险家
贝加尔湖	东非大裂谷	咸海
艾伦比恩	Rebbeca Naomi	月球漫步者
爱奥尼亚海	吕根岛	爱奥尼亚群岛
象棋大师	国际象棋大师	加里卡斯帕罗夫

表4：使用两种不同模型的给定短语的最接近实体的示例。

捷克+货币	越南+资本	德国+航空公司	俄罗斯+河流	法国+女演员
克朗	河内	汉莎航空公司	莫斯科	朱丽叶 比诺什
检查表冠	胡志明市	汉莎航空公司	伏尔加河	凡妮莎帕拉迪斯
波兰zolty	越南	旗舰航空公司汉莎航空	上游	夏洛特 盖恩斯堡
CTK	越南	德国汉莎航空公司	俄国	塞西尔德

表5：使用逐元素添加的矢量组成。使用最佳Skip-gram模型显示了四个最接近两个向量之和的令牌。

为了最大限度地提高短语类比任务的准确性，我们通过使用大约330亿字的数据集来增加训练数据的数量。我们使用了分层softmax，维数为1000，以及整个句子用于上下文。这导致模型达到72%的准确度。当我们将训练数据集的大小减小到6B字时，我们实现了66%的较低准确度，这表明大量的训练数据是至关重要的。

为了进一步了解不同模型所学习的表示方式的不同，我们使用各种模型手动检查了不常见短语的最近邻居。在表4中，我们显示了这种比较的样本。与先前的结果一致，似乎通过具有分层softmax和子采样的模型来学习短语的最佳表示。

5 添加剂组成

我们证明了Skip-gram模型学习的单词和短语表示呈现出线性结构，使得使用简单的矢量算法进行精确的类比推理成为可能。有趣的是，我们发现Skip-gram表示展示了另一种线性结构，这使得通过元素方式添加其矢量表示来有意义地组合单词成为可能。表5说明了这种现象。

可以通过检查训练目标来解释向量的附加属性。字向量与softmax非线性的输入成线性关系。当训练单词向量以预测句子中的周围单词时，可以将向量看作表示单词出现的上下文的分布。这些值与输出层计算的概率呈对数关系，因此两个字向量的总和与两个上下文分布的乘积有关。该产品在此作为AND功能工作：由两个单词向量分配高概率的单词具有高概率，而其他单词具有低概率。因此，如果“伏尔加河”与“俄语”和“河流”一起出现在同一个句子中，这两个词向量的总和将导致这样的特征向量接近“伏尔加河”的向量。

6 与已发表的Word表示法的比较

许多以前从事基于神经网络的单词表示的作者已经发表了他们的结果模型以供进一步使用和比较：最着名的作者之一是Collobert和Weston [2]，Turian等。[17]，和Mnih和Hinton [10]。我们从网上下载了他们的单词向量[3]。Mikolov等。[8]已经在单词类比任务上评估了这些单词表示，其中Skip-gram模型以极大的优势实现了最佳性能。

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

型号（培训时间）	雷德蒙	哈维尔	忍术	涂鸦	投降
科洛伯特（50d）	科尼尔斯	普劳恩	灵气	芝士蛋糕	退位
（2个月）	拉伯克 基恩	捷尔任斯基 奥斯特赖希	科霍纳 空手道	八卦 西洋镜	加入 重新武装
图里安（200d） （几个星期）	麦卡锡 阿尔斯通 卡曾斯	朱厄尔 阿尔苏 奥维茨	- - -	炮火 情感 有罪不罚	- - -
Mnih（100d） （7天）	波德赫斯特 哈朗 阿加瓦尔	教皇 皮诺切特 罗季奥诺夫	- - -	麻醉剂 猴子 犹太人	小牛 规划 犹豫
跳过短语 （1000d，1天）	雷德蒙德洗。 雷德蒙德华盛顿 微软	瓦茨拉夫哈维尔 总统瓦茨拉夫哈维尔 天鹅绒革命	忍者 武术 剑术	喷漆 涂鸦 标注器	投降 投降 投降

表6：给出各种众所周知的模型的最接近的令牌的示例和使用超过300亿个训练单词训练的短语的Skip-gram模型。空单元格表示该单词不在词汇表中。

为了更深入地了解学习向量的质量差异，我们通过显示表6中不常用词的最近邻居来提供经验比较。这些例子表明，在大型语料库上训练的大型Skip-gram模型明显优于所有其他模型在学习表征的质量。这可以部分归因于这个模型已经训练了大约300亿个单词，这比先前工作中使用的典型大小多了两到三个数量级。有趣的是，尽管训练集要大得多，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的贡献。我们展示了如何使用Skip-gram模型训练单词和短语的分布式表示，并证明这些表示呈现线性结构，使得精确的类比推理成为可能。本文介绍的技术也可用于训练[8]中介绍的连续词袋模型。

由于计算效率高的模型架构，我们成功地训练了比先前发布的模型多几个数量级的模型。这导致学习单词和短语表示的质量得到很大改善，特别是对于稀有实体。我们还发现，频繁词的子采样导致更快的训练和明显更好的不常见词的表示。我们论文的另一个贡献是负抽样算法，这是一种非常简单的训练方法，可以学习准确的表示，特别是对于频繁的单词。

训练算法和超参数选择的选择是任务特定的决定，因为我们发现不同的问题具有不同的最优超参数配置。在我们的实验中，影响性能的最关键决策是模型体系结构的选择，向量的大小，子采样率和训练窗口的大小。

这项工作的一个非常有趣的结果是，使用简单的向量加法，单词向量可以在某种程度上有意义地组合。用于学习本文中呈现的短语表示的另一种方法是简单地用单个标记表示短语。这两种方法的结合提供了一种强大而简单的方法来表示较长的文本片段，同时具有最小的计算复杂度。因此，我们的工作可以被视为对现有方法的补充，该方法试图使用递归矩阵向量运算来表示短语[16]。

我们基于本文中描述的技术作为开源项目[4]制作了用于训练单词和短语向量的代码。

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

参考

[1] Yoshua Bengio, Re'jean Ducharme, Pascal Vincent和Christian Janvin。神经概率语言模型。The Machine of Machine Learning Research, 3: 1137-1155,2003。

[2] Ronan Collobert和Jason Weston。自然语言处理的统一架构：具有多任务学习的深度神经网络。在第25届机器学习国际会议论文集，第160-167页。ACM，2008。

[3] Xavier Glorot, Antoine Bordes和Yoshua Bengio。适用于大规模情感分类的领域适应：深度学习方法。在ICML，513-520,2011。

[4] Michael U Gutmann和AapoHyva¨rinen。非标准化统计模型的噪声对比估计，应用于自然图像统计。The Journal of Machine Learning Research，13：307-361,2012。

[5] Tomas Mikolov, Stefan Kombrink, Lukas Burget, Jan Cernocky和Sanjeev Khudanpur。递归神经网络语言模型的扩展。在声学，语音和信号处理（ICASSP），2011 IEEE国际会议，第5528-5531页。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训练策略。在Proc。自动语音识别和理解，2011。

[7] 托马斯米科洛夫。基于神经网络的统计语言模型。博士论文，博士论文，布尔诺理工大学，2012。

[8] Tomas Mikolov, Kai Chen, Greg Corrado和Jeffrey Dean。向量空间中词表示的有效估计。ICLR研讨会，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会议录，2013年。

[10] Andriy Mnih和Geoffrey E Hinton。可扩展的分层分布式语言模型。神经信息处理系统的进展，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 preprint arXiv：1206.6426,2012。

[12] Frederic Morin和Yoshua Bengio。分层概率神经网络语言模型。载于人工智能和统计国际研讨会论文集，第246-252页，2005年。

[13] David E Rumelhart, Geoffrey E Hintont和Ronald J Williams。通过反向传播错误来学习表示。Nature，323（6088）：533-536,1986。

[14] 霍尔格施文克。连续空间语言模型。计算机语音和语言，第一卷。2007年11月21日。

[15] Richard Socher, Cliff C. Lin, Andrew Y. Ng和Christopher D. Manning。用递归神经网络解析自然场景和自然语言。在第26届国际机器学习会议（ICML）会议录，第2卷，2011年。

[16] Richard Socher, Brody Huval, Christopher D. Manning和Andrew Y. Ng。通过递归矩阵 - 向量空间的语义成分。在2012年自然语言处理经验方法会议（EMNLP）会议录中，2012年。

[17] Joseph Turian, Lev Ratinov和Yoshua Bengio。单词表示：半监督学习的简单通用方法。“计算语言学协会第48届年会论文集”，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：语义的向量空间模型。在 *Journal of Artificial Intelligence Research, 37: 141-188, 2010。*

[19] 彼得 特尼。语言之外的分布语义：类比和释义的监督学习。在计算语言学协会（TACL）的交易中，353-366,2013。

[20] Jason Weston, Samy Bengio和Nicolas Usunier。Wsabie：扩大到大词汇量图像注释。在第二十二届人工智能国际联合会议论文集 - 第三卷，第2764-2770页。AAAI出版社，2011年。