单词和短语的分布式表示以及它们的构成

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

Ilya Sutskever
Google Inc.
Mountain View
ilyasu@google.com

Kai Chen
Google Inc.
Mountain View
kai@google.com

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

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

Abstract

最近引入的连续Skip-gram模型是一种一种学习高质量分布式向量表示的有效方法捕捉大量精确的句法和语义词汇人际关系。本文提出了几种改进两者的扩展向量的质量和训练速度。通过对频繁单词进行二次采样，我们获得了显著的速度提升并学习更正规的单词表示。我们还简单描述了一下分层softmax的替代方案称为负采样。

单词表示的一个固有限制是它们的无差异语序和他们不能表示习惯短语的问题。为了例如，“加拿大”和“空气”的含义就不容易理解合并获得‘加拿大航空’。动机: 在这个例子中，我们给出了一个简单的查找方法短语在文本中，并显示良好的学习向量数百万个短语的表示是可能的。

1 简介

词在向量空间中的分布式表示帮助学习算法实现通过分组在自然语言处理任务中取得更好的性能类似的词。最早使用的单词表示法之一追溯到1986年，由于Rumelhart, Hinton和Williams [13]。这一思想被应用到统计语言建模中，取得了相当大的效果成功[1]。后续工作包括自动语音识别和机器翻译的应用[14, 7]，以及广泛的NLP任务[2, 20, 15, 3, 18, 19, 9]。

最近，Mikolov等人[8]引入了Skip-gram 模型，一种学习高质量向量的有效方法大量非结构化文本数据中单词的表示。与之前使用的大多数神经网络架构不同对于学习词向量，可以训练Skip-gram模型(参见图1) 不涉及稠密矩阵乘法。这使得训练极其高效:优化的单机实现可以进行训练一天超过一千亿个单词

使用神经网络计算的单词表示是非常有趣，因为学习的向量很明确编码许多语言规律和模式。有些令人惊讶的是，这些模式中的许多都可以表示作为线性变换。例如，向量计算的结果vec(” Madrid ”) - vec(” Spain ”) + vec(” France ”)更接近vec(” Paris’)比任何其他词向量[9, 8]。

本文给出了对原始的Skip-gram模型。我们展示了频繁抽样在训练过程中，单词的速度显著提高(约2倍- 10倍)，并有所提高较少出现的单词表示的准确性。此外，本文还提出了一种简化的噪声对比方法估计(NCE) [4]用于训练Skip-gram模型导致更快的训练和更好的向量表示频繁的单词，相比更复杂的分层softmax 在之前的工作中使用[8]。

单词表示因其无能而受到限制表示不是由个人组成的习惯用语文字。例如，“波士顿环球报”是一份报纸，所以它不是“波士顿”和“环球”意思的自然结合。因此，使用向量来表示整

个短语使Skip-gram模型更加丰富富有表现力。其他旨在表示句子含义的技巧通过组合词向量，如递归自动编码器[15]，也将受益于使用用短语向量代替词向量。

从基于词的模型到基于短语的模型的扩展相对简单。首先，我们确定大量的使用数据驱动方法的短语，然后我们将短语视为训练期间的个人token。评价...的质量短语向量，我们开发了一套类比推理任务的测试集包含单词和短语。这是我们测试集中典型的类比是" Montreal ": " Montreal Canadiens ":: " Toronto ": '多伦多枫叶"。它被认为是正确的回答，如果vec(" Montreal Canadiens')的最近代表- vec(" Montreal') + vec(" Toronto ")是vec("多伦多枫叶')。

最后，我们描述了Skip-gram的另一个有趣的性质模型。我们发现简单的向量加法通常可以产生有意义的结果结果。例如，vec(" Russia ") + vec(" river ") 离vec(伏尔加河)很近，还有vec(" Germany ") + vec(' capital ")接近vec(' Berlin ")。这种组合性表明不明显的程度语言理解可以通过使用基础数学来获得对词向量表示的操作。

2 Skip-gram模型

Skip-gram模型的训练目标是找到单词表示，对于预测句子中周围的单词很有用或者一份文件。更正式地说，给定一个训练单词序列$w_1, w_2, w_3, \ldots, w_T$, 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(可以是一个函数中心词w_t)。c越大，结果越多训练样本，从而可以导致更高的精度，在培训时间的花费。基本的Skip-gram公式定义$p(w_{t+j}|w_t)$使用softmax函数:

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

其中v_w和v'_w是'输入'和'输出'向量表示w的数量，W是词汇表中的单词数量。这个公式是不切实际的，因为计算$\nabla\log p(w_O|w_I)$的成本与W成比例，而往往很大(10^5—10^7术语)。

2.1 分层Softmax

计算上有效的完整softmax近似是分层softmax。在神经网络语言模型的背景下，它是第一个由Morin和Bengio介绍[12]。主要的优点是不用在神经网络中评估W输出节点来获得概率分布，只需要评估$\log_2(W)$节点。

分层softmax使用二叉树表示输出层以W字为叶，为每一个字Node，显式地表示其子节点的相对概率节点。它们定义了一个随机游走，为单词分配概率。

更准确地说，每个单词w都可以通过适当的路径到达从树的根。的j -th节点为$n(w,j)$ 从根到w的路径，设$L(w)$为这条路径的长度，所以$n(w,1) = \text{root}$和$n(w,L(w)) = w$。此外，对于任何内部节点n，让ch(n)为的任意固定子节点n，如果x为真，则$[\![x]\!]$为1，否则为-1。然后分层softmax定义$p(w_O|w_I)$如下:

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

其中$\sigma(x) = 1/(1+\exp(-x))$。可以证明$\sum_{w=1}^{W} p(w|w_I) = 1$。这意味着计算$\log p(w_O|w_I)$和$\nabla\log p(w_O|w_I)$的成本与$L(w_O)$成正比，平均而言并不更大比$\log W$。此外，与标准的softmax公式的Skip-gram不同哪个分配两个表示v_w和v'_w到每个单词w, the 分层softmax公式具有一个代表v_w为每个单词w和一个代表v'_n 对于二叉树的每个内部节点n。

分层softmax使用的树的结构具有对性能有相当大的影响。Mnih和Hinton 探讨了构建树状结构的几种方法以及对训练时间和结果模型精度的影响[10]。在我们的工作中，我们使用二叉霍夫曼树，因为它将短代码分配给频繁的单词这导致了快速训练。在将单词组合在一起之前就已经观察到了通过它们的频率可以很好地作为神经网络的一种非常简单的加速技术基于网络的语言模型[5, 8]。

2.2 负采样

分层softmax的替代方案是噪声对比估计(NCE)，由Gutmann和Hyvarinen提出[4] 并应用于Mnih和Teh [11]的语言建模。NCE假设一个好的模型应该能够利用逻辑回归进行数据

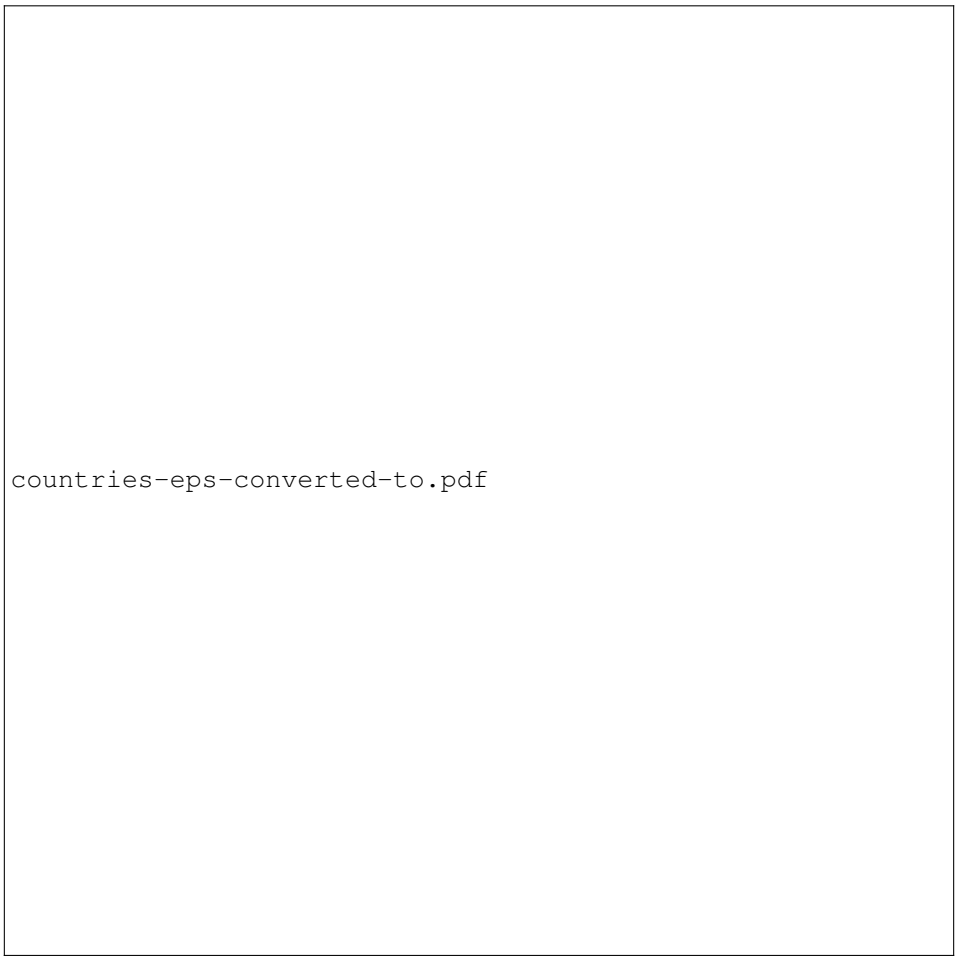

Figure 2: 二维PCA投影的国家及其首都的1000维Skip-gram向量城市。该图说明了模型自动组织的能力概念并隐式学习它们之间的关系，就像在培训我们没有提供任何监督信息关于什么资本城市意味着。

与噪声的区分。这是类似于Collobert和Weston [2]使用的铰链损失该模型通过将数据置于噪声之上进行排序。

而NCE可以被证明近似最大化日志softmax的概率，Skip-gram模型只与学习有关高质量的向量表示，因此我们可以将NCE简化为只要向量表示保持它们的质量。我们定义负采样(NEG)按目的

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

它用于替换Skip-gram目标中的每个$\log P(w_O|w_I)$术语。因此，任务是区分目标词w_O from从噪声分布中提取$P_n(w)$使用逻辑回归，哪里有k阴性每个数据样本的样本。实验结果表明:k在5- 20的范围内，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。负采样和NCE之间的主要区别是NCE 需要样本和噪声分布的数值概率，而负采样仅使用样本。而NCE近似最大化对数概率对于softmax，这个属性对我们的应用程序并不重要。

NCE和NEG的噪声分布均为$P_n(w)$ as 一个自由参数。我们调查了$P_n(w)$的一些选择发现单字分布$U(w)$上升到$3/4$ rd Power(即$U(w)^{3/4}/Z$)的性能明显优于unigram 以及我们尝试的每个任务上的NCE和NEG的均匀分布包括语言建模(这里没有报告)。

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
The following results use 10^{-5} subsampling				
NEG-5	14	61	58	60
NEG-15	36	61	61	**61**
HS-Huffman	21	52	59	55

Table 1: 精度各种Skip-gram 300维模型类比推理任务定义在[8]。NEG- k表示负采样k 每个正样本对应负样本;NCE是噪声对比估计的缩写HS-Huffman代表基于频率的Huffman的分层Softmax 代码。

2.3 频繁单词的二次采样

在非常大的语料库中，最频繁的单词很容易出现上亿次(例如，" in"，" the"和" a")。通常这样的词提供的信息价值比生僻词少。例如，当Skip-gram模型受益于观察到" France"和"巴黎"，它从频繁出现的"法国"中获益甚微还有"the"，因为几乎每个单词都经常在一个句子中同时出现用" the"。这个想法也可以用在相反的地方方向;频繁单词的向量表示不会改变在经过数百万个示例的训练后，效果明显。

为了解决罕见词和频繁词之间的不平衡，我们使用了一个简单的下采样方法:训练集中的每个单词w_i是用公式计算出的概率丢弃

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

其中$f(w_i)$是单词w_i的频率，t是一个选择阈值，通常在10^{-5}附近。我们选择了这个二次采样公式，因为它对频率为的单词进行了积极的子采样大于t，同时保留频率的排序。我们发现，虽然这种下采样公式是启发式选择的该方法在实际应用中效果良好。它加速学习，甚至显著提高学习到的罕见词向量的准确性，将在以下几节中展示。

3 实证结果

在本节中，我们评估了分层Softmax (HS)，噪声对比估计，负采样和训练词的下采样。我们用过类比推理任务[1] 由Mikolov等人介绍[8]。该任务由类比组成，如"德国":"柏林"::"法国":?这些都是通过找到一个向量**x** 使vec(**x**)最接近vec('柏林')- vec('德国')+ vec('法国')根据余弦距离(我们丢弃搜索中的输入单词)。这个具体的例子被认为是如果**x**是"巴黎"正确回答。任务有两大类:句法类比(如" quick ": " quickly ":: ' slow ": ' slow ")和语义类比，例如作为国家与首都城市的关系。

为了训练Skip-gram模型，我们使用了一个大数据集由各种新闻文章组成(一个包含10亿个单词的内部谷歌数据集)。我们从词汇表中删除了所有出现过的单词在训练数据中不到5次，这导致了692K大小的词汇表。各种Skip-gram模型在单词上的性能类比测试集报告在表1中。下表显示了负采样在类比上优于分层Softmax 推理任务，甚至比噪声对比估计的性能稍好。对频繁词进行二次采样，使训练速度提高了几倍并使单词表示变得更加准确。

可以认为，skip-gram模型的线性构成了它的向量这样的线性比较适合类比推理，但结果不一样Mikolov等人[8]也表明，通过标准的s形递归神经网络(高度非线性) 随着训练数据量的增加，显著改善此任务，这表明非线性模型也偏好线性模型单词表示的结构。

4 学习短语

如前所述，许多短语都有a 意思不是一个简单的组成意思的个体文字。为了学习短语的向量表示，我们首先找一些经常一起出现，或者不经常一起出现的词在其他情况下。例如，"纽约时报"和"多伦多枫叶"被训练数据中的独特标记所取代，而二元语法'this is '将保持不变。

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

Newspapers			
New York	New York Times	Baltimore	Baltimore Sun
San Jose	San Jose Mercury News	Cincinnati	Cincinnati Enquirer
NHL Teams			
Boston	Boston Bruins	Montreal	Montreal Canadiens
Phoenix	Phoenix Coyotes	Nashville	Nashville Predators
NBA Teams			
Detroit	Detroit Pistons	Toronto	Toronto Raptors
Oakland	Golden State Warriors	Memphis	Memphis Grizzlies
Airlines			
Austria	Austrian Airlines	Spain	Spainair
Belgium	Brussels Airlines	Greece	Aegean Airlines
Company executives			
Steve Ballmer	Microsoft	Larry Page	Google
Samuel J. Palmisano	IBM	Werner Vogels	Amazon

Table 2: 的例子短语的类比推理任务(完整的测试集有3218个示例)。我们的目标是使用前三个短语计算第四个短语。我们最好的模型在这个数据集上达到了72%的准确率。

这样，我们可以形成许多合理的短语，而不会大大增加短语的大小词汇;理论上，我们可以训练Skip-gram模型使用所有的n元语法，但那样会内存过于密集。许多技术都是以前开发的识别文本中的短语; 但是，对它们进行比较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，在其中形成短语基于一元分词和二元分词计数，使用

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

δ用作折扣系数，防止太多由非常不常用的词组成的短语。分数高于选定阈值的二元分词就被用作短语。通常，我们对训练数据进行2-4遍历，次数递减阈值，允许由多个单词组成的较长的短语。用一种新的类比来评估短语表示的质量涉及短语的推理任务。表2显示本任务中使用的五类类比的例子。这个数据集是公开可用的在网站[2]上。

4.1 短语**Skip-Gram**结果

从之前的实验中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，然后训练了几个短语使用不同超参数的Skip-gram模型。和之前一样，我们使用vector 维度300，上下文大小为5。这个设置已经在短语上取得了很好的性能数据集，并允许我们快速比较负采样以及分层Softmax，包括二次采样和不二次采样频繁的标记。结果总结在表3中。

结果表明，虽然负采样取得了不错的效果准确度即使使用$k = 5$，使用$k = 15$达到相当好的效果表演令人惊讶的是，虽然我们发现分层的Softmax 在不进行二次采样的情况下进行训练时，会获得较低的性能，当我们对频繁出现的单词进行降采样。这说明了二次采样可以加快训练速度，也可以提高准确性，至少在某些情况下是这样。

为了最大限度地提高短语类比任务的准确性，我们增加了训练数据量，使用大约330亿个单词的数据集。我们使用分层softmax，维度为1000和整个句子的上下文。这使得模型的准确率达到了**72**%。我们达到了较低的精度66%当我们将训练数据集的大小减少到6B个单词时，这表明大量的训练数据至关重要。

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

Method	Dimensionality	No subsampling [%]	10^{-5} subsampling [%]
NEG-5	300	24	27
NEG-15	300	27	42
HS-Huffman	300	19	**47**

Table 3: Skip-gram模型在短语类比数据集上的准确性。模型是从新闻数据集中训练了大约10亿个单词。

	NEG-15 with 10^{-5} subsampling	HS with 10^{-5} subsampling
Vasco de Gama	Lingsugur	Italian explorer
Lake Baikal	Great Rift Valley	Aral Sea
Alan Bean	Rebbeca Naomi	moonwalker
Ionian Sea	Ruegen	Ionian Islands
chess master	chess grandmaster	Garry Kasparov

Table 4: 最接近给定短语的实体的例子，使用两个不同的模型。

Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

Table 5: 使用元素加法实现向量的组合性。四个最接近和的符号使用最好的Skip-gram模型给出了两个向量中的一个。

为了进一步了解不同的人学习到的表示有多么不同模型是，我们确实手动检查不常见短语的最近邻居使用各种模型。在表4中，我们展示了这种比较的一个示例。与之前的结果一致，似乎的最佳表示短语由一个具有分层softmax和下采样的模型学习。

5 加性合成性

我们演示了通过Skip-gram学习到的单词和短语表示模型呈现出线性结构，使得执行成为可能基于简单向量算法的精确类比推理。有趣的是，我们发现Skip-gram表示法展示了另一种线性结构使有意义的组合成为可能单词由其向量表示的元素相加。这种现象如表5所示。

向量的可加性可以通过考察培养目标。词向量与输入呈线性关系到softmax非线性。当词向量被训练时预测句子中周围的单词，向量可以被视为代表一个单词在其中的上下文分布出现了。这些值与概率呈对数关系由输出层计算，因此两个词向量的和与两个上下文分布的乘积。产品在这里的作用是和功能:是由两个词向量分配的高概率将具有高概率，并且其他单词的概率很低。因此，如果“伏尔加河”频繁出现在同一个句子中对于单词” Russian’和’ river’，这两个词向量的总和将产生这样一个特征向量，与‘伏尔加河’的向量非常接近。

6 与已发表的单词表示形式的比较

许多以前研究基于神经网络的单词表示的作者已经发表了他们的结果供进一步使用和比较的模型:来自最著名的作者是Collobert和Weston [2], Turian等人[17]，Mnih和Hinton [10]。我们下载了它们的词向量网站[3]。Mikolov等人[8]已经在单词类比任务中评估了这些单词表示，其中，Skip-gram模型以巨大的利润实现了最佳性能。

让我们更深入地了解学者素质的差异向量，通过显示不频繁的最近邻居来进行实证比较表6中的单词。这些例子表明，大型Skip-gram模型在大型数据集上训练在学习到的表示质量方面，语料库明显优于所有其他模型。这可以部分归因于这个模型已经训练了大约300亿个单词，这大约比之前工作中使用的典型尺寸。有趣的是，尽管训练集更大，Skip-gram模型的训练时间只是一小部分降低了之前模型架构所需的时间复杂度。

7 结论

这项工作有几个关键贡献。我们展示了如何训练分布式用Skip-gram模型表示单词和短语，并证明这些表征表现出线性结构，可以进行精确的类比推理有可能。本文介绍的技术也可以用于训练在[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，得益于计算高效的模型架构。这极大地提高了学习到的单词和短语表示的质量，特别是对于稀有实体。我们还发现，子采样的频率单词既能加快

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

Model (training time)	Redmond	Havel	ninjutsu	graffiti	capitulate
Collobert (50d) (2 months)	conyers lubbock keene	plauen dzerzhinsky osterreich	reiki kohona karate	cheesecake gossip dioramas	abdicate accede rearm
Turian (200d) (few weeks)	McCarthy Alston Cousins	Jewell Arzu Ovitz	- - -	gunfire emotion impunity	- - -
Mnih (100d) (7 days)	Podhurst Harlang Agarwal	Pontiff Pinochet Rodionov	- - -	anaesthetics monkeys Jews	Mavericks planning hesitated
Skip-Phrase (1000d, 1 day)	Redmond Wash. Redmond Washington Microsoft	Vaclav Havel president Vaclav Havel Velvet Revolution	ninja martial arts swordsmanship	spray paint grafitti taggers	capitulation capitulated capitulating

Table 6: 最近的标记的例子给出了各种著名的模型和Skip-gram模型使用超过300亿个训练单词对短语进行训练。空牢房意味着这个词不在词汇表中。

训练速度，又能显著提高不常见词汇的表示能力文字。我们论文的另一个贡献是负采样算法，哪一种是极其简单的训练方法它可以学习准确的表示，特别是对频繁出现的单词。

训练算法的选择和超参数的选择是否是一个任务特定的决策，就像我们发现的不同问题一样不同的最佳超参数配置。在我们的实验中，影响性能的最关键的决策是选择模型架构、向量的大小、下采样率，以及训练窗口的大小。

这项工作的一个非常有趣的结果是词向量是否可以有意义地组合使用只是简单的向量加法。另一种学习表示的方法本文提出的短语的主要特点是用单个表示短语token。这两种方法的结合提供了一种强大而简单的方法如何表示较长的文本，同时具有最小的计算量复杂性。因此，我们的工作可以看作是对现有工作的补充试图使用递归表示短语的方法矩阵向量运算[16]。

我们基于这些技术编写了用于训练单词和短语向量的代码本文所描述的是一个开源项目[4]。

References

[1] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

[2] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[3] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, 513–520, 2011.

[4] Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *The Journal of Machine Learning Research*, 13:307–361, 2012.

[5] Tomas Mikolov, Stefan Kombrink, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. Extensions of recurrent neural network language model. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5528–5531. IEEE, 2011.

[6] Tomas Mikolov, Anoop Deoras, Daniel Povey, Lukas Burget and Jan Cernocky. Strategies for Training Large Scale Neural Network Language Models. In Proc. *Automatic Speech Recognition and Understanding*, 2011.

[7] Tomas Mikolov. Statistical Language Models Based on Neural Networks. *PhD thesis, PhD Thesis, Brno University of Technology*, 2012.

[8] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *ICLR Workshop*, 2013.

[9] Tomas Mikolov, Wen-tau Yih and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of NAACL HLT*, 2013.

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

[10] Andriy Mnih and Geoffrey E Hinton. A scalable hierarchical distributed language model. *Advances in neural information processing systems*, 21:1081–1088, 2009.

[11] Andriy Mnih and Yee Whye Teh. A fast and simple algorithm for training neural probabilistic language models. *arXiv preprint arXiv:1206.6426*, 2012.

[12] Frederic Morin and Yoshua Bengio. Hierarchical probabilistic neural network language model. In *Proceedings of the international workshop on artificial intelligence and statistics*, pages 246–252, 2005.

[13] David E Rumelhart, Geoffrey E Hintont, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[14] Holger Schwenk. Continuous space language models. *Computer Speech and Language*, vol. 21, 2007.

[15] Richard Socher, Cliff C. Lin, Andrew Y. Ng, and Christopher D. Manning. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, volume 2, 2011.

[16] Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. Semantic Compositionality Through Recursive Matrix-Vector Spaces. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2012.

[17] Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394. Association for Computational Linguistics, 2010.

[18] Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. In *Journal of Artificial Intelligence Research*, 37:141-188, 2010.

[19] Peter D. Turney. Distributional semantics beyond words: Supervised learning of analogy and paraphrase. In *Transactions of the Association for Computational Linguistics (TACL)*, 353–366, 2013.

[20] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three*, pages 2764–2770. AAAI Press, 2011.