# 词向量：一项调查

费利佩-阿尔梅达 杰拉尔多
里约热内卢联邦大学计算机和系统工程项目
巴西，里约热内卢
{falmeida,xexeo}@cos.ufrj.br

## 摘要

这项调查基于分布式假说，列举并描述了最近为单词构建固定长度、密集和分布式表示的主要策略。这些表示现在通常被称为词嵌入，除了编码非常好的与语法和语义信息外，还被证明在许多下游 NLP 任务中作为额外功能很有用。

## 1 简介

表示单词和文档的任务是大多数自然语言处理(NLP)任务的一部分。一般来说，将它们表示为向量是很有用的，具有比较直观的解释，可以作为有用操作的主题(例如：加减、距离度量等)，并很好地应用于许多机器学习(ML)算法和策略中。

向量空间模型(VSM)，起源于信息检索(IR)社区，由 Salton 等人于 1975 年提出，可以说是最成功和最有影响力的将单词和文档编码为向量的模型。

当然，基于自然语言的解决方案的另一个非常重要的部分是对语言模型的研究。语言模型是语言使用的统计模型。它主要侧重于在给定多个先前单词的情况下预测下一个单词。这非常有用，例如，在语音识别软件中，即使信号质量很差或存在大量背景噪声，也能正确判断说话者所说的词是什么。

这两个看似独立的领域可以说是通过最近对神经网络语言模型（NNLMs）的研究结合到了一起，Bengio 等人于 2003 年开发了第一个基于神经网络的大规模语言模型。

他们的想法是将其重新定义为无监督学习问题。该方案的一个关键特点是，原始词向量首先被投影到所谓的嵌入层，然后再被输入到网络的其他层。除其他原因外，这被认为有助于缓解维度诅咒对语言模型的影响，并有助于泛化（Bengio 等，2003）。

随着时间的推移，词嵌入已成为一个研究主题，人们意识到它们良好的句法和语义词关系的事实(Mikolov 等，2013)，并且可以在许多 NLP 任务中用作独立特征(Turian 等，2010)。

最近，出现了其他创建嵌入的方法，它们不依赖神经网络和嵌入层，而是利用词上下文矩阵来获得词的向量表示。在最有影响力的模型中，我们可以引用 GloVe 模型(Pennington 等，2014)。

这两种类型的模型有一些共同之处，即它们依赖于具有相似上下文的单词(其他单词)具有相同含义的假设。这被称为分布式假设，由 Harris 等人于 1954 年提出。

这让我们想到了我们将在本文中使用的词嵌入的定义，正如文献所建议的（例如 Turian 等，2010；Blacoe 和 Lapata，2012；Schnabel 等，2015），根据这个定义，词嵌入是密集的、分布

式、固定长度的词向量，根据分布式假设，使用词的共现性统计建立。

源自神经网络语言模型的嵌入模型(Baroni 等，2014)被称为基于预测的模型，因为它们通常利用语言模型，根据上下文预测下一个单词。其他基于矩阵的模型被称为基于计数的模型，因为它们考虑了全局单词上下文共现计数来得出词嵌入。下面将对这些进行描述。

该调查的结构如下:在第 2 节中，我们描述了统计语言建模的起源。在第 3 节中，我们概述了由所谓的基于预测的模型和基于计数的方法生成的词嵌入。在第 4 节中，我们进行了总结，在第 5 节中，我们指出了一些未来的研究方向。

### 1.1 动机

据我们所知，目前还没有关于词嵌入的全面调查，更别说包括这一领域的现代发展的调查了。此外，鉴于词嵌入在各种随之而来的 NLP 任务中的有用性(Turian 等，2010)以及在这些向量中编码的非常准确的语义信息，(Mikolov 等，2013a)我们认为这样的工作是有价值的。

### 1.2 范围

我们选择了基于引用计数和对新模型的报告影响的混合的文章或策略。

## 2 背景:向量空间模型和统计语言模型

为了了解词嵌入产生和发展背后的原因，我们认为向量空间模型和统计语言模型是两个最重要的课题。

向量空间模型很重要，因为它是 NLP 大部分工作的基础;它允许使用成熟的数学理论（如线性代数和统计）来支持我们的工作。此外，向量表示也是广泛的机器学习算法和方法所需要的，这些算法和方法被用来帮助解决 NLP 任务。

现代词嵌入的研究(特别是基于预测的模型)在某种程度上是基于提高语言建模效率和准确性的尝试。事实上，词嵌入被视为语言模型的副产品，直到一段时间后(可以说是在 Collobert 和 Weston(2008)之后)，词嵌入的构建才从语言模型的任务中分离出来。

下面我们将对这两个主题进行简要介绍。

### 2.1 向量空间模型

人们尝试将分析方法应用于文本数据时，遇到的第一个问题可能是如何以一种适合于相似性、组合等操作的方式来表示它。

信息检索(IR)领域提出了最早实现这一目标的方法之一，由 Salton 等人于 1975 年提出了一种编码过程，其中集合中的每个文档都由一个 t 维向量表示，每个元素表示该文档中包含的一个不同的术语。这些元素可以是二进制数或实数，可以选择使用加权方案（例如 TF-IDF）进行归一化，以说明每个术语提供的信息差异。

有了这样一个向量空间，就可以继续利用这些向量进行有用的工作，例如计算文档向量之间的相似性(甚至使用简单的操作，如它们之间的内

积)，为搜索结果评分(将搜索词视为一个伪文档)，等等。

Turney 和 Pantel(2010)对利用 VSM 的不同方法进行了彻底的调查，同时解释了最适合他们的特定应用。

### 2.2 统计语言模型

统计语言模型是一种语言中词汇分布的概率模型。例如，在给定上下文的情况下，它们可以被用来计算下一个单词的可能性。它们最早的用途之一是在语音识别领域(Bahl 等，1983)，以帮助正确识别在声音信号中受到噪声和故障通道影响的单词或短语。

在文本数据领域，此类模型可用于许多的 NLP 任务以及其他相关任务，例如信息检索。

虽然一个包含所有可能在语言中出现的单词上下文的每个单词的可能性的完全概率模型显然是难以处理的，但根据经验观察，使用小至 3 个单词的上下文可以获得令人满意的结果(Goodman，2001)。这种窗口大小为 T 的 n-gram 模型的简单数学公式如下:

$$P(w_1^T) = \prod_{t=1}^{T} P(w_t|w_1^{t-1})$$

其中$w_t$为第 t 个单词，$w_i^T$为$w_i$到$w_T$的单词序列，即($w_i$，$w_{i+1}$，$w_{i+2}$，…，$w_T$)。$P(w_t|w_1^{t-1})$表示$w_t$出现在序列$w_1^{t-1}$之后的次数。在给定上下文下一个单词的实际预测是通过对词汇中的所有单词进行最大似然估计(MLE)来完成的。

这些模型报告的一些问题是(Bengio 等，2003)计算词汇量在 10 万个单词的离散联合分布时涉及到高维，以及将模型推广到训练集中不存在的单词序列时的困难。

早期缓和这些影响的尝试，特别是那些与泛化到看不见的短语相关的尝试，包括使用平滑，例如，假装每个新序列在训练集中都有一个计数，而不是 0(这被称为拉普拉斯平滑。此外，当较长的上下文无法使用时，要转而使用越来越短的上下文(Katz，1987)。另一种减少所需计算量并有助于泛化的策略是在所谓的类中对单词进行聚类(参见现在著名的 Brown clustering Brown 等，1992)。

最后，神经网络(Bengio 等，2003；Bengio 和 Senécal，2003）Collobert 和 Weston，2008）和对数线性模型（Mnih 和 Hinton，2007；Mikolov 等，2013b,c）也被用来训练语言模型（产生了所谓的神经语言模型)，提供更好的结果，以困惑度来衡量。

## 3 词向量

如前所述，词嵌入是单词的固定长度向量表示。有多种方法可以获得这样的表示，本节将探讨各种不同的方法来训练词嵌入、详细描述它们的工作原理，以及它们之间的区别。

词嵌入根据(Baron 等，2014；Pennington 等，2014；Li 等，2015)诱导策略通常分为两种类型。利用局部数据(例如一个单词的上下文)的方法被称为基于预测的模型，通常会让人想起神经语言模型。另一方面，使用全局信息(通常是语料库范围的统计信息，如单词计数和频率)的方法称为基于计数的模型。我们接下来描述这两种类型。

### 3.1 基于预测的模型

基于预测的嵌入式模型的发展历史与神经语言模型(NNLMs)的发展历史密切相关，因为这是它们最初产生的方式。如前所述，词嵌入就是将原始单词向量投影到这些模型的第一层，即所谓的嵌入层。

NNLMs 的历史始于第一个大型神经语言模型(Bengio 等，2003)，它主要是一个逐步提高效率的过程，偶尔会在复杂模型和能训练更多数据的简单模型之间进行取舍。

尽管早期的结果(以困惑度衡量)清楚地表明，神经语言模型确实比之前的基于 n-gram 的模型更擅长建模语言，但长时间的训练(有时长达数天或数周)经常被认为是阻碍此类模型发展的主要因素。

在 Bengio 等人(2003)的开创性论文发表后不久，人们为提高这些模型的效率和性能做出了许多贡献。

Bengio 和 Senècal（2003）指出，计算成本的主要来源之一是 softmax 输出层所需的分区函数或归一化因子，例如神经网络语言模型（NNLMs）中的那些。他们使用一个叫做重要性抽样的概念（Doucet (2001)），设法绕过了昂贵的归一化因子的计算，而是使用一个辅助分布（例如旧的 n-gram 语言模型）来估计神经网络的梯度，并从词汇表中随机抽样。报告称，与前一个模型相比，他们的训练时间增加了 19 倍，分数相似 (以困惑度衡量)。

稍后，Morin 和 Bengio (2005)提出了另一种方法来加快训练和测试时间，使用层次化的 Softmax 层。他们意识到，如果将输出的词按层次二叉树的结构进行排列，就可以使用在

| 文章 | 战略概述 | 架构 | 笔记 |
|---|---|---|---|
| Bengio 等，2003 | 嵌入是源于神经网络语言模型的副产品 | 神经网络 | 通常被称为第一个神经网络工作语言模型 |
| Bengio 和 Senecal，2003 | 通过使用蒙特卡洛方法估计梯度，绕过高计算成本的分区函数，对上一篇论文做了改进 | 神经网络 | 与 2003 年 Bengio 等人的研究相比，训练时间减少了 19 倍 |
| Morin 和 Bengio，2005 | 完全的 softmax 预测被一种更有效的二叉树所取代，在这种方法中，只需要在通往目标词的每个节点上进行二元决策 | 神经网络，分层 softmax | 报告说，与 2003 年 Bengio 和 Senecal 的研究相比，速度有所提高（训练期间速度是 3 倍多，测试期间速度为 100 倍），但得分（困惑度）略低 |
| Mnih 和 Hinton，2007 | 在其他模型中，这里介绍的是对数线性模型，对数线性模型是具有单一的、线性的、隐藏层的神经网络 | 对数线性模型 | 首次出现的对数线性模型，是一个更简单的模型，速度更快，并略微超过了 Bengio 等人 2003 的模型 |
| Mnih 和 Hinton，2008 | 正如 Morin 和 Bengio（2005）所建议的那样，作者使用分层的 softmax 来训练对数线性模型，但词树是学习而不是从外部获得的 | 对数线性模型，分层 softmax | 报告称其速度是以前对数线性模型的 200 倍 |
| Collobert | 多任务神经网络不仅使用无监 | 深度神经 | 第一次建立一个主要是为了 |

| | | | |
|---|---|---|---|
| 和 Weston，2008 | 督数据进行训练，还使用监督数据（如 SRL 和 POS 注释）进行训练。 该模型联合优化了所有这些任务，但目标只是学习嵌入。 | 网络，负向取样 | 第一次构建模型主要是为了输出嵌入。半监督模型（语言模型+NLP 任务） |
| Mikolov 等，2013b | 介绍了两个新的模型，即 CBOW 和 SG，两者都是对数线性模型，使用两步训练过程。CBOW 预测给定上下文的目标词，SG 预测给定目标词的上下文词 | 对数线性模型，分层 softmax | 在 DistBelief 上训练，他是 Tensorflow 的前驱（Abadi 等，2015） |
| Mikolov 等，2013c | 对 CBOW 和 SG 的改进，包括使用负向取样代替分层 softmax 和频繁词的二次取样 | 对数线性模型，负向取样 | SGNS（带负样的跳格），是 word2vec 的最佳性能变体，在此被引入 |
| Bojanowski 等，2016 | 嵌入是在 n-gram 层面上训练的，以帮助对未见过的数据进行归纳，特备是对那些形态学起重要作用的语言 | 对数线性模型，分层 softmax | 报告的结果比 SGNS 更好。据报道，嵌入也有利于组合（成句子，文档嵌入）。 |

表 1：为嵌入构建基于预测的模型的策略概述

每个指向该词的节点上选择正确路径的概率，作为计算每个词的完整分布的代理。由于二叉树在一组单词上的高度是 |V |/ log(|V |)，这可以产生指数级的加速。在实践中，收益并不明显，但与使用重要性抽样的模型相比，他们仍能在训练时间和测试时间上分别获得 3 倍和 100 倍的收益。。

Mnih 和 Hinton（2007）可能是第一个提出对数线性模型（LBL）的作者，该模型在后来的工作中也具有很大的影响力。

Mnih 和 Hinton(2008)的另一篇文章可以看作是 LBL (Mnih 和 Hinton(2007))模型的扩展，使用 Morin 和 Bengio(2005)提出的分层 softmax 方案的一个轻微修改版本，产生了所谓的分层对数双线性模型(HLBL)。Morin 和 Bengio(2005)使用了从 WordNet 中预先构建的单词树，而 Mnih 和 Hinton(2008)专门为手头的任务学习了这样的树。除了其他小的优化，他们报告了比以前的 LBL 模型(速度是以前的 200 倍)有很大的改进，并得出结论，使用专门构建的单词树是获得这些结果的关键。

与刚才提到的作品有些相似的是，Collobert 和 Weston(2008)从一个稍微不同的角度来研究这个问题;他们是第一个以学习嵌入为目的设计模型的人。在以前的模型中，嵌入只是作为主要任务(通常是语言模型)的一个有趣的副产品。除此之外，他们还引入了两个值得一提的改进:他们使用单词的完整上下文(之前和之后)来预测中心词。或许最重要的是,他们引入了一个更聪明的方法来利用未标记的数据生产良好的嵌入:而不是训练语言模型(这里不客观),他们用错误或否定的例子扩展了数据集，并简单的训练了一个模型，可以从错误的例子中分辨出肯定的（实际发生的）例子。

在这里,我们应该提到 Mikolov 等人（2009；2010）的两个具体贡献，这

两个贡献已在后来的模型中使用。在第一项工作中，（Mikolov 等，2009）提出了一种引导 NNLM 的两步方法，即使用单个单词作为上下文训练第一个模型。然后，训练完整的模型（具有更大的上下文），将第一步发现的模型作为初始嵌入。

在（Mikolov 等，2010）中，首次提出了使用递归神经网络（RNN）来训练语言模型的想法；其论点是 RNN 将状态保持在隐藏层中，帮助模型记住任意长的上下文，并且不需要事先决定在任意一侧使用多少单词作为上下文。

2012 年，Mnih 和 Teh 建议进一步提高 NNLMs 的培训效率。利用噪声对比估计（NCE）。NCE（Gutmann and Hyvarinen（2010））是一种通过对真/假例子的二元决策来估计概率分布的方法。这使得作者能够进一步减少 NNLMs 的训练时间。除了更快的训练时间，他们还报告了以前的神经语言模型的困惑评分更好。

可以说，2013 年，随着 Mikolov 等人（2013a；2013b；2013c）的参与，NLP 社区再次（主要的另一个例子是 Collobert 和 Weston（2008））将单词嵌入作为一个值得研究的话题。这些作者分析了通过训练递归神经网络工作模型（Mikolov 等人（2010））获得的嵌入，以期找到可能编码在向量中的句法规律。

也许令人惊讶的是，对于作者自己来说，他们在数据中不仅发现了语法上的规律，而且还发现了语义上的规律。许多常见的关系，如男女关系、单复数关系等，实际上对应于一个人可以对词向量执行的算术运算（例如，参见图 1）。

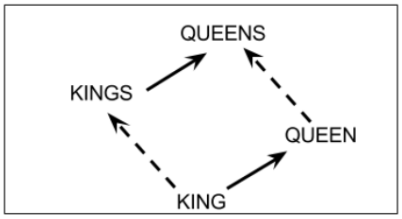


图 1：高维词嵌入（通过 RNN 语言模型获得）在 2D 中的投影：高级词嵌入对词之间的多个关系进行编码； 这里显示：单复数（虚线）和男女（实线）关系。 改编自 Mikolov 等人。 （2013a）。

不久之后，在 2013b 和 2013c，Mikolov 等人引入了两种学习嵌入的模型，即连续词袋（CBOW）和跳过语法（SG）模型。这两种模型都是对数线性模型（如前所示），并使用两步过程（Mikolov 等，2009）进行训练。CBOW 和 SG 之间的主要区别在于用于更新模型的损失函数；虽然 CBOW 训练了一个模型，旨在根据上下文预测中心词的模型，但在 SG 中，角色被颠倒了，而中心词被用来预测上下文中出现的每个词。

CBOW 和 SG 的第一个版本(Mikolov 等人(2013b))使用分层的 softmax，而 Mikolov 等人(2013c)提出的变体则使用负采样代替。此外，变体引入了对频繁词的二次采样，以减少因过度频繁词而产生的噪音量，并加速训练。这些变体表现得更好，训练时间更快。

在基于预测的单词嵌入模型的最新贡献中，我们可以引用两篇文章(Bojanowski 等人(2016)和 Joulin 等人(2016))，这两篇文章通常被引用为 Facebook 公司提供的 FastText 工具包的来源。他们提出了对 Mikolov 等人(2013 年 3 月)的 skip-gram 模型的

改进,即学习的不是单词嵌入,而是 n-gram 嵌入(可以组成单词)。这一决定背后的基本原理是，那些严重依赖词法和构词法的语言(如土耳其语、芬兰语和其他高度不灵活的语言)本身就有一些编码在单词部分的信息，这些信息可以用来帮助概括不可见的单词。尤其是在德语、法语和西班牙语等语言中(Mikolov 等人(2013 年 c))，表现出了更好的效果。

表 1 对构建词嵌入的基于预测的模型进行了结构化比较。

### 3.2 基于计数的模型

如前所述，基于计数的模型是产生词嵌入的另一种方式，不是通过训练算法预测给定上下文的下一个单词（就像语言建模中的情况），而是通过利用语料库中的单词上下文共现计数。它们通常被表示为词上下文矩阵（Turney and Pantel（2010））。

利用词上下文矩阵产生词嵌入的最早相关例子当然是潜在语义分析(LSA) (Deerwester 等人(1990))，其中 SVD 应用于 term-document 矩阵。这个解决方案最初是用来帮助信息检索的。虽然人们可能对 IR 中的文档向量更感兴趣，但也可以通过这种方式获得单词向量;我们只需要看分解后的矩阵的行(而不是列)。

不久之后,Lund 和 Burgess（1996）引入了语义存储模型（HAL）。他们的策略可以描述如下：对于词汇表中的每个单词，分析它出现在其中的所有上下文，并计算目标单词和每个上下文单词之间的共现计数，与上下文单词到目标单词的距离成反比。作者称结果良好（通过类比任务衡量），最佳上下文窗口大小为 8。

最初的 HAL 模型没有对发现的单词共现计数进行任何标准化。因此，像“贡献”这样非常常见的词，与所有与它们同时出现的词不成比例。Rohde 等人（2006 年）发现这是一个问题，并介绍了 COALS 方法，引入了规范化策略来消除单词中的频率差异。而不是使用原始计数，他们建议最好考虑有条件的共现，即单词 A 与单词 B 共同出现的可能性要大于与词汇中的一个随机词共发生的程度。他们报告的结果比以前的方法要好，使用的是奇异值分解变量。

| 文章 | 概述 | 笔记 |
| --- | --- | --- |
| Deerwester 等，1990 | 引入 LSA。奇异值分解(SVD) 应用于术语文档矩阵 | 主要用于 IR，但也可以用于词嵌入 |
| Lund 和 Burgess，1996 | 介绍了 HAL 方法，每次扫描整个语料库中的一个词，在该词周围有一个上下文窗口,收集加权的词共现次数,建立一个词共现矩阵 | 报告称最佳语境大小为 8 |
| Rohde 等，2006 | 作者介绍了 COALS 方法，它是 HAL 的改进版，使用归一化程序来阻止非常常见的术语过度影响共同出现次数 | 最佳的变体使用了 SVD 因子化。报告称比 HAL、LSA 和其他方法更有优势 |
| Dhillon 等，2011 | 介绍了 LR-MVL。使用左右语境之间的 CCA 来诱导单词嵌 | 报告称,在许多 NLP 任务中,比 C&W 嵌入、HLBL 和其他方 |

| | | |
|---|---|---|
| | 入 | 法都有收益 |
| Lebret 和 Collobert，2013 | 将修改后的主成分分析（称为 Hellinger PCA）应用于词上下文矩阵。 | 在实际 NLP 任务使用之前，可以对嵌入进行调整。还称在许多 NLP 任务中比 C&W 嵌入、HLBL 和其他方法的收益 |
| Pennington 等，2014 | 引入了 GloVe，这是一种经过训练的对数线性模型，用于将单词之间的语义关系编码为学习向量空间中的向量偏移，使用共现率而不是原始计数是词义的实际传送器的洞察力。 | 报告称，在多个 NLP 任务中，比之前所有基于计数的模型和 SGNS 都有进步 |

表 2：为嵌入构建基于计数的模型的策略概述

Dhillon 等人（2011）提出了一种稍有不同的替代方案，他们在其中引入了低阶多视角学习（LR-MVL）方法。简而言之，它是一种迭代算法，其中嵌入是通过利用给定单词的左右上下文之间的典型相关分析 (CCA) (Hotelling (1935)) 得出的。该模型的一个有趣特性是，当嵌入用于下游 NLP 任务时，它们也与上下文单词的嵌入连接在一起，从而产生更好的结果。作者列举了在许多 NLP 任务中，与其他矩阵分解方法以及神经嵌入相比的优势。

Lebret 和 Collobert(2013)也对基于计数的模型做出了贡献，他们建议将 Hellinger PCA 转换应用于单词上下文矩阵。据报道，结果比之前的基于计数的模型，如 LR-MVL 和神经嵌入，如 Collobert 和 Weston(2008)和 HLBL Mnih 和 Hinton(2008)的模型更好。

本节将介绍的最后一个模型是 Pennington 等人（2014 年）的著名 GloVe。该模型从共现率开始，而非原始计数，编码关于这词对的实际语义信息。该关系用于为对数线性模型推导合适的损失函数，然后对其进行训练以最大化每个词对的相似性，如前面提到的共现率所测量的。作者称比其他基于计数的模型以及基于预测的模型（如 SGNS（Mikolov et al.（2013c））在单词类比和 NER（命名实体识别）等更好的结果。

表 2 给出了构建单词嵌入的基于计数模型的结构化比较

## 4 总结

单词嵌入已经被发现对许多 NLP 任务非常有用，包括但不限于 Chunking (Turian 等，2010)，Question answer (Tellex 等,2003)，句法分析和情感分析(Socher 等，2011)。

我们在这里概述了迄今为止用于推导这些嵌入的一些主要工作和方法，这两种方法都使用基于预测的模型，该模型对给定单词序列的下一个单词的概率进行建模（就像语言模型一样），以及基于计数的模型，该模型利用单词上下文矩阵中的全局共现统计信息。

文献中的许多建议已经被纳入广泛使用的工具包，如 Word2Vec、gensim、FastText,和 GloVe，从而产生了更加准确和快速的词嵌入，并准备用于 NLP

任务。

## 5 未来工作

单词表示(特别是单词嵌入)的研究仍然很活跃;我们认为最有前途的研究方向有:

### 5.1 为特定任务调整嵌入

Maas 等人(2011)、Labutov 和 Lipson(2013) 以及 Lebret 和 Collobert(2013)等人强调了针对特定任务调整嵌入时,NLP 任务的改进结果。

### 5.2 基于预测和基于计数的模型之间的联系

例如,Levy 和 Goldberg(2014)提出 SGNS 模型(Mikolov 等,2013 年)实际上相当于使用略微修改过的词上下文矩阵,使用 PMI(点态互信息)统计量加权。看看这两个模型之间的联系可能会在这两个领域产生更多的进展。

### 5.3 为更高层次的实体编写词嵌入

虽然关于如何组成单词向量来表示句子和文档等高级实体的研究并不完全是新的(通常以分布构成的名义),但最近的工作已经适应了专门用于神经词嵌入的解决方案:我们可以在这里引用段落 2vec (Le 和 Mikolov (2014)),Kiros 等人的 Skip-Thought V 载体(2015)以及 FastText 本身(Joulin 等人(2016)和 Bojanowski 等人(2016))。

## 对数线性模型与神经嵌入

对数线性模型是一种概率设备,可以用来对条件概率进行建模,就像单词上下文和目标单词之间的概率模型一样,这是语言模型的基本部分。

对数线性模型为每个输出单元订阅以下模板(Collins):

$$P(y|x;v) = \frac{\exp\left(v \cdot f(x,y)\right)}{\sum_{y' \in Y} \exp\left(v \cdot f(x,y')\right)}$$

当应用于语言建模任务时,使用神经嵌入:y 表示标签,即目标单词。x 表示一个词的上下文,即我们想要预测的目标词之前或周围的词。v 是一个学习过的参数,即共享权矩阵中的单行向量。

可以把上面的公式看作一个神经网络,它有一个单一的线性隐藏层,连接到 softmax 输出层。此外,类似于任何神经网络模型,它也可以用基于梯度的方法训练,可以扩展到包括正则化项,等等。

## 参考文献

Mart´ın Abadi, Ashish Agarwal, Paul Barham, et al., 2015. TensorFlow:

Large-scale machine learning on heterogeneous systems. Available at http://tensorflow.org/. Software available from tensorflow.org. L. R. Bahl, F. Jelinek, and R. L. Mercer.

March 1983. A maximum likelihood approach to continuous speech recognition, Marco Baroni, Georgiana Dinu, and German´ Kruszewski. June 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for

Computational Linguistics.

Yoshua Bengio and Jean-Sebastien Sen ´ ecal, 2003. ´ Quick training of probabilistic neural nets by importance sampling.

Yoshua Bengio, Jean Ducharme, Pascal Vincent, and Christian Janvin. March 2003. A neural probabilistic language model,

William Blacoe and Mirella Lapata. July 2012. A comparison of vector-based representations for semantic composition. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information,

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language,

John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word cooccurrence statistics: A computational study, Behavior Research Methods.

John A. Bullinaria and Joseph P. Levy. 2012. Extracting semantic representations from word cooccurrence statistics: stop-lists, stemming, and svd,

Michael Collins. Log-linear models. Available at http://www.cs.columbia.edu/ ˜mcollins/loglinear.pdf.

R. Collobert and J. Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In International Conference on Machine Learning, ICML.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis,

Paramveer Dhillon, Dean P Foster, and Lyle H. Ungar. 2011. Multi-view learning of word embeddings via cca. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 24. Curran Associates, Inc.

Arnaud Doucet. 2001. Sequential Monte Carlo methods in practice. Springer, New York.

Joshua Goodman. 2001. Classes for fast maximum entropy training,

Michael Gutmann and Aapo Hyvarinen, 2010. ¨ Noise-contrastive estimation: A new estimation principle for unnormalized statistical models.

Zellig S. Harris. 1954. Distributional structure,

Geoffrey E. Hinton. August 2002. Training products of experts by minimizing contrastive divergence,

Harold Hotelling. 1935. Canonical correlation analysis (cca), Journal of Educational Psychology.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification

Slava M. Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. In IEEE Transactions on Acoustics, Speech and Signal Processing.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov,

Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors,

Igor Labutov and Hod Lipson, 2013. Reembedding words.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents,

Remi Lebret and Ronan Collobert. 2013. Word ´ emdeddings through hellinger PCA,

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada.

Shaohua Li, Jun Zhu, and Chunyan Miao. 2015. A generative word embedding model and its low rank positive semidefinite solution,

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence,

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11. Association for Computational Linguistics. Risto Miikkulainen and Michael G. Dyer. 7 1991. Natural language processing with modular pdp networks and distributed lexicon,

Tomas Mikolov, Jiri Kopecky, Lukas Burget, Ondrej Glembek, and Jan Cernocky. 2009. Neural network based language models for highly inflective languages. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '09. IEEE Computer Society.

Tomas Mikolov, Martin Karafiat, Luk ´ as Bur- ´ get, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013a. Linguistic regularities in continuous space word representations. In Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, At□lanta, Georgia, USA.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Efficient estimation of word representations in vector space,

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013c. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.

Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. In ICML '07: Proceedings of the 24th international conference on Machine learning. ACM.

Andriy Mnih and Geoffrey Hinton. 2008. A scalable hierarchical distributed language

model. In In NIPS.

Andriy Mnih and Yee Whye Teh. 2012. A fast and simple algorithm for training neural probabilistic language models. In In Proceedings of the International Conference on Machine Learning.

Frederic Morin and Yoshua Bengio. 2005. Hierarchical probabilistic neural network language model. In Robert G. Cowell and Zoubin Ghahramani, editors, Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics. Society for Artificial Intelligence and Statistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. October 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics.

Douglas L. T. Rohde, Laura M. Gonnerman, and David C. Plaut. 2006. An improved model of semantic similarity based on lexical cooccurence,

G. Salton, A. Wong, and C. S. Yang. November 1975. A vector space model for automatic in□dexing,

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation meth□ods for unsupervised word embeddings. In Pro□ceedings of the Conference on Empirical Meth□ods in Natural Language Processing (EMNLP).

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP ’11. Association for Computational Linguistics.

Stefanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes, and Gregory Marton. 2003. Quantitative evaluation of passage retrieval algorithms for question answering. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, SIGIR ’03. ACM.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL ’10. Association for Computational Linguistics.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics