登峰杯

论文类别	☐ 学术作品—自然科学类　☐ 学术作品—人文社科类 ☐ 数学建模竞赛　☐ 数据挖掘竞赛　☐ 艺术创意设计竞赛
论文题目	133703492 深度学习方法训练词向量

清华大学教育研究院
中国高等教育学会学习科学研究分会

摘要：

词向量是词的一种特征表示，是许多自然语言处理任务的基础步骤。本文介绍了一种深度学习训练词向量的方法，训练完成的词向量具有维度低、保留语义相似性的特点。词向量训练包括如下过程：用统计语言模型进行中文分词、去停用词、构造词库、CBOW 模型训练词向量。本文在清华大学新闻分类语料库上，利用 fastText 实现了一版词向量，并且训练得到一个正确率为 94.1%的分类器。

关键词：中文分词 词向量 CBOW fastText

一、深度学习和词向量

1.1 深度学习

随着 21 世纪电子信息技术的发展以及数据资源的爆炸，机器学习逐渐由浅层学习发展为深度学习，其动机在于建立模拟人脑进行分析学习的神经网络。简单而言，深度学习指基于深层神经网络的机器学习模型，即深层次学习。深度学习与浅层学习最重要的区别和优点在于突出了特征学习的重要性。它能够以海量的训练数据为背景，通过构建深度学习模型自主抽取样本的特征，极好地避免了人工选取的错误与不足。目前深度学习的主流模型包括：深度信念网络、深度玻尔兹曼机、自动编码器、卷积神经网络、循环神经网络以及长短时记忆模型等。

1.2 词向量

多种模型的搭建使得深度学习的应用越来越广泛，其中在自然语言处理中的应用尤为重要。但由于语言文字是人类认知过程中产生的较高级抽象信息，因此其应用进展较为缓慢，目前取得最基础也是最有趣的一个成果即为词向量。利用基于深度学习方法得到词的分布式语义表示——词向量，这种连续、低维的数值型向量可以非常方便地应用到许多任务中，例如词性解析、命名实体识别、语言模型等。并且能够在海量的信息中有效地提取有用信息，从而更加方便了人们的生活。

二、研究现状

随着计算机计算能力增强和数据资源指数型增长，深度学习在近几年得到迅速发展，并成为各大公司和高校的研究热点。1943 年，Warren McCulloch 在单层神经网络的基础上首次提出深度神经网络——前向多层神经网络；目前深度学习已经在语音识别和图像识别领域取得了突破性进展。例如，在语音任务中， Frank 使用基于深度神经网络的隐马尔可夫模型将语音到词语的转换错误率从 27.4%降低到 18.5%；图像任务中，深度学习模型大大降低了手写数字识别任务的识别错误率，将错误率从 1.4%下降到 0.39%。

深度学习在自然语言处理任务中取得的主要成就是用深度神经网络训练词向量。1986 年，Hinton 发表论文提出分布式表示的词向量；2003 年，Benjor 提出用一种利用统计语言模型构造神经网络训练词向量的方法——NNLM

模型，首次将深度学习方法较好的应用到词向量训练中，取得了较好的结果；此后，在 NNLM 模型的基础上，关于利用深度学习训练词向量的研究还有 2008 年 Ronan Collobert 和 Jason weston C&W 方法、Google 的 Word2Vector 模型、Mikolov 的 RNNLM 模型。

三、深度学习方法训练词向量

在自然语言处理任务中，词是表达语义的最小单位。词向量是词的一种特征表示，能够在保留词的语义信息的基础上方便的被计算机处理。因此词向量成为很多自然语言处理任务的基础步骤。中文词向量训练主要包括以下两个步骤：1）中文分词；2）词向量训练。

3.1 中文分词

本文介绍一种基于概率统计的中文分词方法，这种分词方法基于一个非常简单的思想：一个句子是否合理，就看它的可能性大小，而这种可能性可以用概率来衡量。从数学上描述如下：

假定S表示一个特定的有意义的句子，由一连串特定顺序排列词$w_1, w_2, w_3, \dots, w_n$组成，其中 n 是句子的长度，计算句子可能性即计算句子出现的概率，可以表示成如下：

$$P(S) = P(w_1, w_2, \dots, w_n) \tag{1}$$

利用条件概率公式，我们可以将句子出现的概率展开如下：

$$P(w_1, w_2, \dots, w_n) = P(w_1) * P(w_2|w_1) * P(w_3|w_1, w_2) * \dots * P(w_n|w_1, w_{2,\dots}, w_{n-1}) \tag{2}$$

其中$P(w_1)$表示词w_1出现的概率，$P(w_2|w_1)$表示在词w_1出现的条件下w_2出现的概率，不难看出第$i, (i = 2, 3, \dots, n)$个词出现的与前面$i-1$个词相关。从计算上面来看$P(w_1)$计算起来很简单，$P(w_2|w_1)$也不难，但是随着n的增大，$w_1, w_2, \dots, w_n$的可能性呈指数增长，无法估算。

俄国一位叫马尔科夫的科学家提出了一种有效的方法：假设每个词出现只跟前面一个词出现有关系。这种假设在数学上被称为马尔科夫假设，加入了马尔科夫假设后，S出现的概率可以按照如下方式计算：

$$P(w_1, w_2, \dots, w_n) = P(w_1) * P(w_2|w_1) * P(w_3|w_2) * \dots * P(w_n|w_{n-1}) \tag{3}$$

增加马尔科夫假设的语言模型我们称之为二元模型。在一个大型的语料库中，根据大数定理，我们可以用频率值近似地描述概率值从而计算得到句子的概率值。这种计算按照某种词序列组成的句子的概率的模型我们称之为统计语言模型。

利用上述的统计语言模型，我们可以将中文分词用几个简单的数学公式阐述如下：假设一个句子有几种分词方法，这里我们假设有三种，分别表示如下：

$$S = A_1, A_2, \dots, A_k \tag{4}$$

$$S = B_1, B_2, \dots, B_m \tag{5}$$

$$S = C_1, C_2, \dots, C_n \tag{6}$$

其中，$A_1, A_2, \ldots, B_1, B_2, \ldots, C_1, C_2, \ldots$ 等等就是汉语的词，不同的分词结果可以产生不同的词串，k, m, n 三个不同的下标表示句子在不同的分词下词串的数量。那么按照统计概率的思想最好的分词方式就是分词后保证这个句子出现的概率最大。如果按照上述的二元语言模型，我们计算得到不同分词方式的概率，如果有：

$$P(A_1, A_2, \ldots, A_k) > P(B_1, B_2, \ldots, B_m) \quad (7)$$

$$P(A_1, A_2, \ldots, A_k) > P(C_1, C_2, \ldots, C_n) \quad (8)$$

我们就认为$S = A_1, A_2, \ldots, A_k$是最好分词方法。

如果我们采用穷举的方式计算所有可能的句子组成方式，计算量会非常大，现在一般是把它看做成为一个动态规划问题，可以采用目前很成熟的维特比算法快速的找到最佳分词。

采用深度学习方法训练词向量的时候，首先对语料库中所有文档进行中文分词，去掉停用词，收集所有不重合的词组成词库，在词库上训练词向量。

3.2 深度学习

深度学习是机器学习算法的一种，主要是多层神经网络，神经网络结构如下图所示：

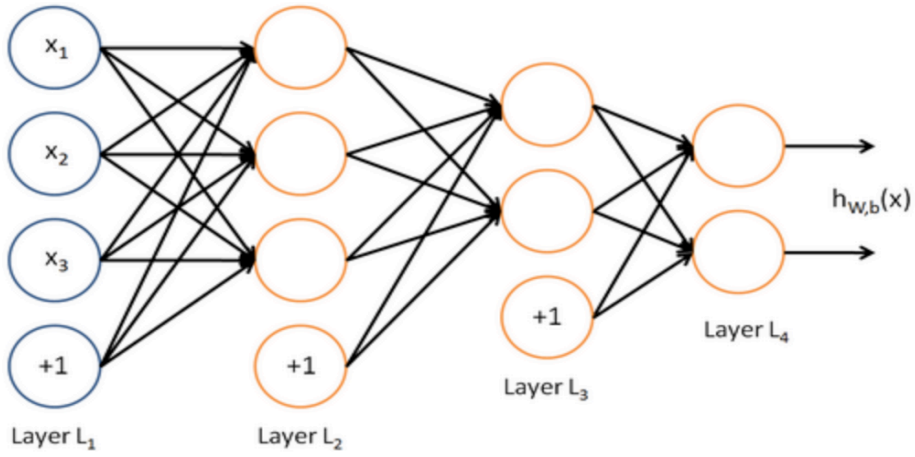

图 1：神经网络结构

有如下结构特点：

（1）层级结构

网络层数大于等于三层，其中第一层为输入层，最后一层为输出层，中间所有层统称为隐含层。这种层级结构使得神经网络能够在前层提取的特征基础上进行特征提取，通过多层特征提取，神经网络能够学习到高级抽象的特征。

（2）多神经元

输入层多神经元能够使得神经网络能够允许多个输入变量，一个输入层神经元对应一个输入变量，输出层多神经元使得神经网络能够允许多个预测值，一个输出层神经元对应一个输出变量，输入层和输出层多神经元的结构使得神经网络模型结构灵活，能够广泛的应用到各种问题。隐含层多个神经元使得神经网络能够有可能学习到多个维

度的特征，进一步提高了神经网络模型的特征表达能力。

（3）线性运算和非线性运算相结合

L+1 层的每个神经元和 L 层是全连接，L+1 层每个神经元单独和前层所有神经元构成一个感知机结构，其神经元值是 L 层所有神经元值的加权加和再通过一个非线性的激活函数，这种线性运算和非线性运算结合的方式使得神经网络具有很强的表达能力。

上述三个特点使得神经网络模型容量很大，学习能力很强，在数据足够的基础上，能够学习到很好的特征表示，因此当前训练词向量的方式主要是采用深度学习的方式。

3.3 词向量训练

词向量有两种重要表示，即独热表示与分布式表示。独热表示将单词表示为维度与词典大小一致并且只有单词索引位置对应维度为 1，其余均为 0 的特征向量，one-hot 是 bow 的基本形式，但仍有巨大缺点：1、维度大，易发生维度灾难；2、词汇之间有鸿沟，无法捕捉词与词之间相似度。而随着深度学习的发展应用，建立起来的分布式表示则很好地解决了上述缺点，即将单词表示为共现矩阵，更加灵活地捕捉上下文信息。因此伴随着计算机硬件提升和优化算法改进，基于神经网络的深度学习技术逐渐占据了词向量表示的主导地位。

本文主要介绍一种主流的深度学习训练词向量的方法——CBOW (Continuous Bag-of-Words)，模型结构如下图所示：

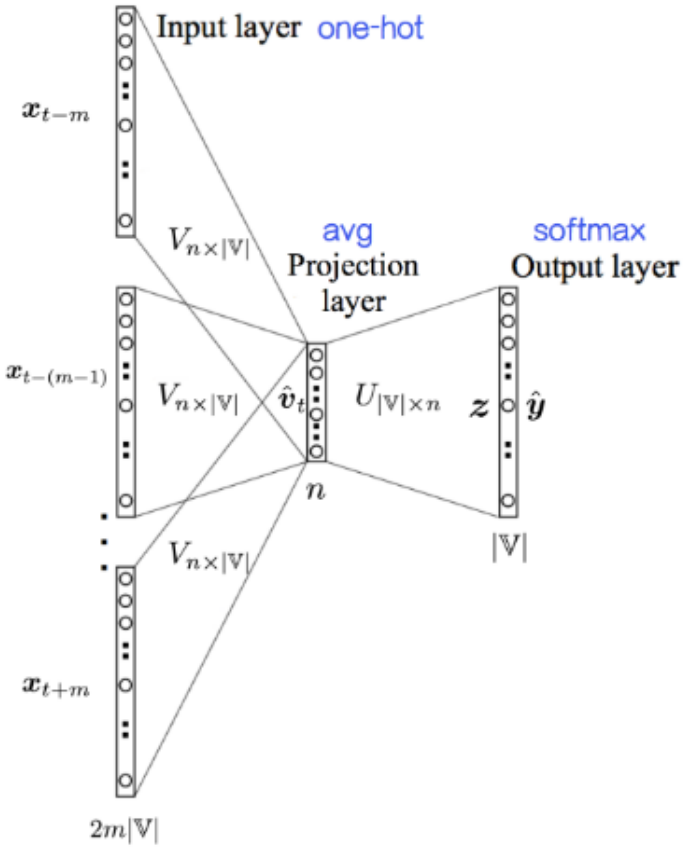

图 2：CBOW 模型

CBOW 模型是一个三层的神经网络结构，通过上下文来预测中心词，并且抛弃了词序信息：

输入层：n个节点，上下文共 2m个词的词向量的平均值；

输入层到输出层的连接边：输出词矩阵$U_{|V|\times n}$

输出层：|v|个节点。第 i 个节点代表中心词是词w_i的概率。

首先，将中心词w_t的上下文c_t: $w_{t-m},\dots,w_{t-1},w_{t+1},\dots w_{t+m}$ 由独热表示x_{t+j}转成输入词向量v_{t+j}:

$$v_{t+j} = Vx_{t+j}, j \in \{-m,\dots,m\}\backslash\{0\} \tag{9}$$

进而将上下文的输入词向量$v_{t-m},\dots,v_{t-1},v_{t+1},\dots v_{t+m}$ 求平均值，作为模型输入:

$$v_t = \frac{1}{2m}\sum_j v_{t+j} \ , j \in \{-m,\dots,m\}\backslash\{0\} \tag{10}$$

输出层采用 softmax 作为激活函数，用 logloss 作为损失函数，利用梯度下降带入训练数据训练词向量，文本中所有词向量存在于矩阵$V_{|V|\times n}$中。

四、实践应用

建立在上述理论基础上，通过利用网上提供的开源工具，我实现了一版词向量，并且在词向量的基础上，完成了一个文本分类的分类器，过程可以表示如下：

（1）获取语料

本文使用的语料来源于清华大学新闻分类语料库 THUCNews，THUCNews 是根据新浪新闻 RSS 订阅频道 2005~2011 年间的历史数据筛选过滤生成，包含 83 万篇新闻文档（3.9 GB），均为 UTF-8 纯文本格式。在原始新浪新闻分类体系的基础上，重新整合划分出 14 个候选分类类别：财经、彩票、房产、股票、家居、教育、科技、社会、时尚、时政、体育、星座、游戏、娱乐。

（2）语料预处理

解压后的新闻语料按照类别分布在不同目录的不同文件中，为了方便后续处理，将同一个目录下每个文件处理成一行，并且添加标签，存放到同一个文件中，文件以目录名称命名，得到 14 个类别的文件。

（3）中文分词

中文分词采用的是网上提供的开源工具 Ansj，用 Ansj 将语料中除了标签之外的文本进行分词，分词结果用空格分隔开，将文本中标签去掉，得到语料库中所有文本的中文分词结果，将 14 个文件合并到一个文件，得到最终训练集合。

（4）词向量训练

词向量训练采用的是网上提供的开源工具 fastText，fastText 是 Facebook 提供的一个词向量训练工具，速度很

快，能够在普通 CPU 上在几十分钟内快速训练得到词向量。在（3）中得到的训练集合中用 fastText，采用 CBOW 模型，训练词向量。训练用时 14 分钟，得到一个词库大小为 277959，向量维度为 100 维的词向量，人工检查发现向量相似性效果较好，下图是两个相似性例子：

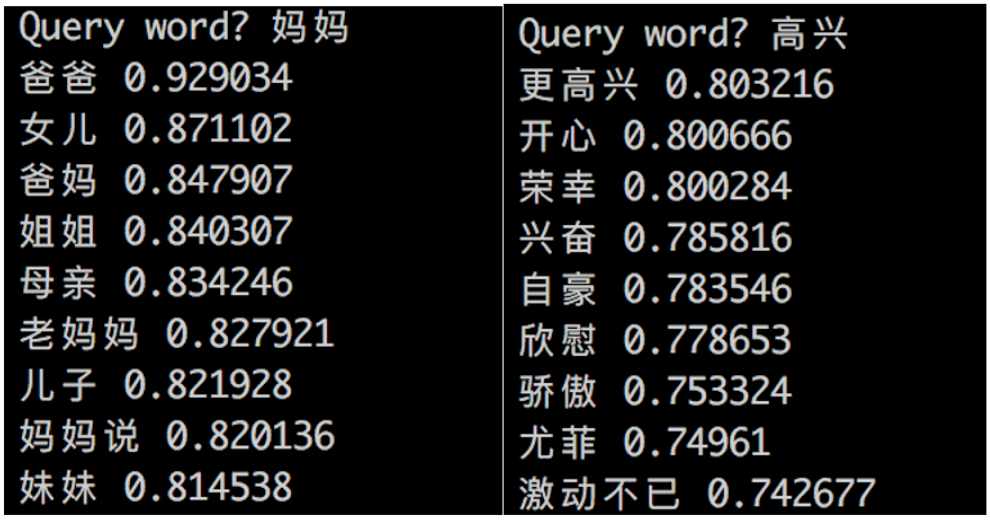

图 3：词向量相似性例子

（5）文本分类

利用训练好的词向量，根据中文分词的结果，将一个文档中所有词对应的词向量做一个加和平均可以得到句子的特征表示，用这种特征表示作为输入，（2）中标记好的标签作为输出，构造一个三层神经网络可以训练一个分类器。将数据按照 9：1 划分训练集和测试集，在训练集合上用 fastText 构造分类器，用时 2 分 14 秒，在测试集合上验证分类器正确率为 94.1%。

五、总结展望

本文首先介绍了进行词向量训练的一般过程，用 fastText 训练了一版词向量，并且用训练好的词向量做了一个文本分类的应用。在完成课题过程中得到如下结论：1）CBOW 模型能够快速训练出低维连续的词向量;2)训练好的词向量能够较好的保存语义相似性。于此同时，在完成课题的过程中我发现分词结果存在很多不足，例如“平凡的世界”的分词结果是“平凡/的/世界”，针对这种问题，查阅相关资料后我发现可以采用加入词库和加入 2gram 信息这两种方法解决；在完成文本分类应用的时候，发现分类正确率非常高，这可能是因为文本较长。后续我将调研这种分类模型在短文本上的分类效果。

六、参考文献

[1] 张志华. 基于深度学习的情感词向量及文本情感分析的研究[D]. 东北师范大学，2016.

[2] 闫琰. 基于深度学习的文本表示与分类方法研究[D]. 北京科技大，2016.

[3] 于政. 基于深度学习的文本向量化研究与应用[D]. 东北师范大学，2016.

[4] 周练. Word2Vec 的工作原理及应用探究[J] 科技情报开发与经济，2014，第 2 期.

[5] 吴军. 数学之美 [M]. 北京: 人民邮电出版社，2012.

[6] 清华大学新闻分类语料:http://thuctc.thunlp.org/ .

[7] 开源词向量训练工具 fasttext:https://github.com/facebookresearch/fastText .

[8] 开源分词工具 Ansj: https://github.com/NLPchina/ansj_seg

[9] Armand Joulin, et al. Bag of Tricks for Efficient Text Classification

CONTACT US

官方网站	www.dengfengbei.com
微信公众号	**Dengfengbeijingsai**
官方 QQ 群	（1）“登峰杯”学术作品学生 QQ 群 571526693 （2）“登峰杯”数学建模学生 QQ 群 571535826 （3）“登峰杯”机器人学生 QQ 群 571540979 （4）“登峰杯”结构设计学生 QQ 群 592858677 （5）“登峰杯”数据挖掘学生 QQ 群 144821810 （6）“登峰杯”艺术创意设计学生 QQ 群 318850726
官方邮箱	dengfengbei@126.com
联系电话	010-52909593，18310079788 （工作日 9:00~12:00，13:00~17:00）