# 基于深度学习的训练词向量和文本分类

邵栩辉
（陕西省渭南市韩城市象山中学，陕西渭南， 715400）

**摘要：**随着互联网的普及，人们在网上“冲浪”的同时产生了海量的文本数据，而对文本数据进行分析和归类则是近些年来的研究热点，其中一个研究方向就是文本分类。本文主要介绍了基于深度学习，运用CBOW算法来训练词向量，以及利用fasttext来进行文本分类的实验。

**关键词：**fasttext；词向量；文本分类；分词原理

DOI:10.16589/j.cnki.cn11-3571/tn.2020.20.034

## 0 引言

随着信息时代的到来，计算机与人们生活的联系也不断地紧凑起来，发挥着越来越大的作用。对于使用汉语的我国来说，中文信息处理技术已在我国的计算机科学方面占有一席地位。中文信息处理主要对中文进行转换、传输、存储、分析等加工，它是人类跨向认知智能所面临的最大技术难点之一，人类需要教会计算机理解人类自然语言，从而让机器会思考以及推理。

人工智能、信息检索、机器翻译和自动文摘等领域突破的关键，长期以来一直困扰着这一研究领域的许多专家学者，目前我国研究人员在自然语言处理上取得了斐然的成绩[3]，每年自然语言处理的相关会议投稿者，有一半是华裔学者，这也显示了在中文自然语言处理上我国的巨大优势，但是仍然有很多领域需要去突破，比如分词技术。

## 1 文本分类历史及主流用法

随着中文信息处理研究的深入，中文文本分词问题已经引起了相当程度的重视，成为中文信息处理的一个前沿问题，对此，人们设计了许多的文本分词分类的方法，经典的分类算法有朴素贝叶斯（Naive Bayes NB）、K- 近邻（K-Nearest Neighbor KNN）、决策树（Decision Tree DTree）、算数平均质心（Arithmetical Average Centroid AAC）、支持向量机（Support Vector Machine SVM）。

我们采用的文本分类工具是 fasttext，fasttext 是 facebook 开源的一个词向量与文本分类工具，于 2016 年开源，典型应用的场景是“带监督的文本分类问题”。它是一种简单并且高效的文本分类以及表征学习的方法，性能要比深度学习好并且速度也更加快。fasttext 将自然语言处理以及机器学习当中最成功的理想相结合。这其中又包括了使用词袋和 n-gram 表征语句，并通过隐层表征学习在类别间共享信息。fasttext 主要用于进行文本分类。随着技术的突破，fastText 的学习速度也变得非常快，效果也在不断地变好。fasttext 特别适用于分类类别非常大而且数据集足够多的情况，然而当分类类别比较小或者数据集比较小时，会容易过拟合。

同时本文基于文本数据，训练了一版词向量，词向量就是将词用向量的方式来表示。词向量主要包括 One-hot、词的分布式表达等。One-hot 用来表示词向量非常的简单，但它存在许多问题。其一就是向量的维度会随着句子中的词的数量类型增大而增大。比如，众所周知词汇表通常较为庞大，如果达到了百万级别，每个词就都会用百万维的向量来表示，随之而来的就会是一场内存灾难。再一个就是任何两个词之间都是分开的，无法在语义层面上表示词汇之间的相关信息。

以前 one-hot 的表示方法就是取消词的语义，再将其符号化。那应该怎么将语义融入到词表示中？针对这个问题， Harris 提出了“分布式假说”，为这个设想提供了基础：假如上下文的词差不多相似，那么就可以认为它们的意思也差不多相似。而分布式假说被 Firth 在 1957 年进行了更进一步的阐述和明确：词的语义由它的上下文决定。以 CBOW 为例，如果一个句子“the children play computer games in their house”，在训练的时候将“the children play computer games in their”作为输出，则预测出最后一个词是“house”。分布式表示的较大优点在于它具有非常强大的表征能力，比如说 n 维向量每维有 k 个值，那么就可以表示为 k 的 n 次方个概念。所以，当我们要完成某个任务时，我们就可以将其拆分为两个步骤：

（1）通过某种方式来描述其上下文；

（2）构建一个模型来分析目标词与它的上下文之间的关系。

目前业界比较常用的词向量表示是一种分布式的表示，它是基于神经网路训练得到的，它的核心仍然是通过某种方式描述其上下文以及构建模型描述上下文之间的关系。

在此之上我们了解了 One-hot 编码的维度过大的缺点。所以进行了如下的改进：

（1）将 vector 当中的元素由整形改为浮点型，并将其调整为用整个实数范围表示；

（2）将之前的巨大维度压缩到一个更小的维度空间中。

词向量的本质是训练神经网络时的隐藏层参数，或者说是矩阵值。而 CBOW(Continuous Bag-of-Words) 和 skip-



(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

gram 是两种用来训练词向量的方法。CBOW 模型的训练是将某一个特征词与上下文相关的词的词向量输入进去，这样输出的就是这个特征词的词向量。比如：“Our English teacher taught us a long time”。我们的上下文大小取值为 4，那么特定的这个值就是“taught”，也就是我们要输出的词向量，那么我们的输入就是上下文对应的那 8 个词的词向量。然而 skip-gram 模型与 CBOW 的思路相反，简而言之为输入某一个特定的词的词向量，输出的是那个特定的词对应的上下文的词向量。

## 2 技术原理

CBOW 与 Skip-gram 模型互为镜像，其算法流程如图一所示，CBOW 算法是利用周围的词预测中心词，而 Skip-garm 算法则是用中心词去预测周围的词。

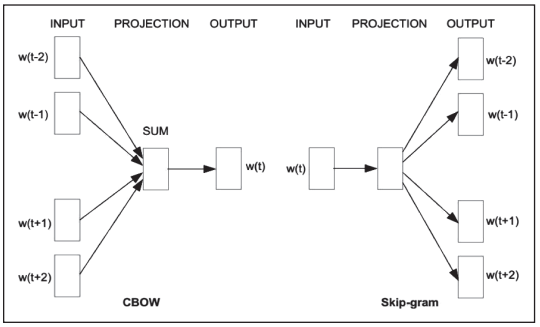


图 1　CBOW 和 Skip 的算法流程

(1) 分词的原理

分词就是把连续的字序列用一定的规律重新排列，然后将它们重新组合成词序列的过程。中文分词就是将一个汉字序列进行切分，得到一个个单独的词。表面上看，分词其实就是这样的，但是分词效果好不好对信息检索、实验结果还是影响很大的，同时分词背后还涉及各种各样的算法。然后我们根据这些算法的特点可以将他们分为三大类，分别是基于规则的分词、基于统计的分词、基于语义的分词。

基于统计的分词方法是我们的主要研究对象，这种分词方法的定义是把每个词看作是由词的最小单位及各个字组成的，如果相连的字在不同的文本当中出现的次数非常多，那么这些相连的字很可能就是一个词。这样的话就可以利用它们出现的频率来推算出成词的可靠度，统计材料中相邻共现的各个字的组合的频度，当组合频度高于某一个临界值时，我们便可以认为此字组可能会构成一个词语。 这其中有以下几个统计模型：N 元文法模型 (N-gram)、隐马尔可夫模型、最大熵模型、条件随机场模型等。

这里主要介绍 N-gram 算法，N-gram 模型基于一个假设，该假设认为第 x 个词出现只与第 x-1 个词有关，与其他任何词都没有关系，而整个句子的概率就是各个词出现的概率的乘积。假设一个 n 个词组成的句子 $S=(w1,w2,\ldots,wn)$，如何衡量它的概率呢？让我们假设，每一个单词 wi 都要依赖于从第一个单词 w1 到它之前一个单词 wi-1 的影响：

$$P(S)=p(w_1w_2\ldots w_n)=p(w_1)p(w_2|w_1)\ldots p(w_n\mid w_{n-1\ldots}w_2w_1) \quad (1)$$

这个衡量方法确实简单，但是存在以下两个缺陷：①参数空间过大，概率 p(wn|wn-1…w2w1)p(wn|wn-1…w2w1) 的参数有 O(n) 个；②数据稀疏严重，词同时出现的情况可能没有，组合阶数高时尤其明显。

为了解决这个问题，学者们又提出了马尔可夫假设，该假设认为一个词的出现只需要依赖于它前面出现的有限的一个或者几个词。如果一个词的出现只需要依赖于在它前面出现的某个词，我们就可以称它为 bigram。即：

$$P(S)=p(w_1w_2\ldots w_n)=p(w_1)p(w_2|w_1)\ldots p(w_n\mid w_{n-1\ldots}w_2w_1)$$
$$\approx p(w_1)p(w_2|w_1)\ldots p(w_n\mid w_{n-1}) \quad (2)$$

S 代表着一个句子，w 代表着每个词

（2）训练词向量的原理

词向量顾名思义，就是用一个向量的形式来表示一个词，这样一来，词和词之间就可以用定量来度量它们之间的关系，从而得到词与词之间的联系。One-hot 是最早用来表示词向量的方法，在这种方法中，每个词被表示为一个实数向量，其长度为字典大小，每个维度对应一个字典里的每个词，除了这个词对应的维度上是 1，其他元素都是 0。但是 One-hot 他也有两个缺点：

(1) 若句子中词的数量类型不断地增大，则向量的维度也会随着不断地增大；

(2) 最大的缺点就是任何两个词之间全都是孤立的，不能将语义层面上词汇之间的相关信息表示出来。

因为 One-hot 的缺点，我们又引出了神经网络训练词向量的方法，神经网络训练的算法主要分为两种：CBOW 和 skip-gram 算法。 CBOW 模型的训练是将某一个特征词与上下文相关的词的词向量输入进去，这样一来输出的就是这个特征词的词向量。

（3）文本分类

FastText 是 facebook 开源的一个词向量与文本分类工具，它主要提供简单并且高效的文本分类方法以及表征学习的方法，且性能可与深度学习相比而且速度变得更快。Fasttext 将自然语言处理和机器学习中最成功的理念相互结合。这其中包括了 N-gram 表征语句和词袋，然后又通过隐层的表征学习共享信息。在这里我们采用了另一个 softmax 层级来为运算过程提速。同时这些概念用于两个不同的任务，



(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

分别是有监督学习，进行有效的进行文本分类；无监督学习，来学习词向量表征。

而 fasttext 原理主要包含三个部分：

-- 模型架构

-- 层次 softmax

--N-gram 特征

在这里我们重点研究的是 fasttext 的模型结构，fasttext 的架构和 word2vec 中的 CBOW 的架构类似，因为他们的作者都是 facebook 的科学家 Tomas Mikolov, 而且确实 fasttext 也算是 word2vec 衍生出来的。fasttext 模型它的输入是一句话或者一个文本，然后输出这句话或者这个文本属于各个类别的概率。它的模型具体结构可以分为三层，第一层是序列中的词与词组成的特征向量，然后通过一些线性变换映射到中间层，中间层再通过一些变换映射到输出层，也就是其对应的各个类别的概率。在输出层使用的是非线性激活函数，而在中间层使用的是线性激活函数。

显然 fasttext 的模型结构与 CBOW 训练词向量的模型结构很相似。其中最大的不同就是，我们用 fasttext 预测其属于各个类别的概率，而 CBOW 模型是运用上下文词的词向量来预测目标词的词向量。当类别过多时，fasttext 为了加速，它还使用了一个分层分类器，同时使用了一个层次 softmax 技巧。当然，fasttext 也可用来进行文本分类和句子分类。不管是当中的哪一个，我们经常用的是词袋模型，词袋模型通俗的理解就是把词无序的放在一个“袋子”中，这个“袋子”其实就是我们构建的词典。但是词袋模型又不能考虑词序，所以 fasttext 就加入了 N-gram 特征。为了提高效率，我们需要过滤掉低频率的 N-gram。

## 3 实验过程

### ■ 3.1 数据来源

THUCNews 是由新浪新闻频道在 2005 年至 2011 年间的历史数据生成的，这其中包含有 74 万篇的新闻（2.19GB），均为 UTF-8 纯文本格式。我们在原始的新浪新闻分类体系基础上，重新整合和划分出 14 个分类类别，分别是家居、教育、科技、社会、时尚、时政、体育、星座、财经、彩票、房产、股票、游戏和娱乐。在发布这个数据集时，清华的实验室还发布了他们在这个数据集上的分类效果为 87.5%。

### ■ 3.2 分词的实验过程

在这里我们采用的是 jieba 分词工具包。我们分词是用 jieba 分词工具包对一个个的文档进行分词，分完词后，词与词之间用空格隔开，此外我们还进行了去停用词的操作，停用词表来源于 github 上面公开的百度用的停用词表，将一些常见的词从分词结果中去掉。

### ■ 3.3 训练词向量过程

我们先获得上一步分词后得到的实验结果，把他们都导入内存中，然后调用 python 包 gensim 提供的 word2vec 接口，用 CBOW 算法来训练词向量。

### ■ 3.4 文本分类

我们要先按照 1:1 的比例划分训练集和测试集，然后我们文本分类时用训练集来训练模型，用测试集来验证模型，我们用的是 facebook 开源的 fasttext 来训练文本分类模型，然后我们设置训练模型时的 epoch 为 10（即将模型训练 10 次），损失函数是 softmax。训练之后我们将模型压缩后保存到本地，便于后面进行验证。

## 4 实验结果分析

（1）训练词向量结果的好坏：我们实验的时候验证了与“伤心”这个词最接近的词有：心痛、难过、内疚、惋惜、沮丧、痛心、伤感、悲伤、懊悔、自责。

```
搜索词为：  伤心
('心痛', 0.8967168927192688)
('难过', 0.8825957775115967)
('内疚', 0.8368224501609802)
('惋惜', 0.8177362680435181)
('沮丧', 0.8176010847091675)
('痛心', 0.8137685060501099)
('伤感', 0.8044317364692688)
('悲伤', 0.799864649772644)
('懊悔', 0.7980804443359375)
('自责', 0.7836451530456543)
```

图 2　实验结果

（2）文本分类的实验结果：我们用测试集 418039 个新闻，来测试模型预测的准确率，最后准确率为 94.0%。这个结果与当初清华实验室发布数据集时的分类效果 87.5% 对比，我们的实验还是很有优势的。

我们的文本的准确率很高，通过分析后，这可能是我们的新闻文本较长，所以模型针对长文本进行分类时准确率比较高，后续我们可以针对短文本进行一些实验。

综上所述，我们基于 THUCNews 自己训练了一版词向量，并且实现了同义词的词向量在向量空间中距离很近，同时，我们利用 fasttext 在训练集上训练了一个分类模型，并在测试集上验证了模型的效果，最终为 94.0%，该实验结果与这个数据发布时的效果 87.5% 对比，有很大的提升。

## 5 未来展望

过去一段时间里，中文信息处理技术发展的非常快，逐渐满足了人们对中文信息处理的基本需求，但从长远的角度看，它的自动分词技术还是最为重要的一部分，所以对它之后的发展，我们还是要不断地完善之前的技术。为了让汉语

（下转第 100 页）



(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

得出高频系数 D 的区域方差 FD, 从而得出归一化区域方差 DFD；

得出高频系数 E 的区域方差 FE, 从而得出归一化区域方差 DFE。

我们构建一致性约束规则，构建时运用归一化区域方差作为辅助，我们可以得出 3×3 这一区域归一化区域方差的值，它们分别是 DFEj 和 DFDj。

接下来，我们把归一化区域方差 DFEj、归一化区域方差 DFDj 全部赋值给这一区域的像素的最核心点。我们利用刚刚得到的融合系数，进行 HSV 逆变换和快速离散 curvelet 逆变换 , 这样我们就得到了经过融合后的图像 [6]。

图 3　融合后的图像

## 5 结束语

上述文中，我们得到 MS 图像亮度分量的最佳方法，为采用 HSV 变换 , 这样能较完整的保留图像的光谱信息 ; 而我们得到图像分解系数的最佳方法，为快速离散 curvelet 变换，该方式降低了算法的难度 , 提高了计算效率。利用图像归一化区域能量的特性 , 采用不同的方法对低频系数进行融合 , 得到更具空间特性和光谱信息的融合图像。高频系数的重要特性之一便是归一化区域方差特性 , 该特征能对构建一致性约束规则有着不可替代的作用 , 正是这一点很好地约束了像素点的一致性 , 使融合图像的效果更好。实验发现 , 文中所提算法融合效率与融合质量效果良好。


**参考文献**

＊ [1] 王欲，傅艺伟，群殴诶等 . 数字图像的融合 [M]. 西安 : 西安交通大学出版社 ,2010,40-46

＊ [2] 谭毅天 , 王久天 . 一种基于 curvelet 变换的改进型图像融合的算法 [J]. 工程学报，2010，19(5)：152-158

＊ [3] 李章程，彭大卫，王一查等 .curvelet 智能图像处理技术 [M]. 电子工业出版社，2009：250-274

＊ [4] 陈书海，傅录祥 . 实用 curvelet 数字图像处理 . 北京 : 科学出版社 [M],2012.1-7，140-145

＊ [5] 章毓晋 .curvelet 图像处理和分析基础 . 北京 : 高等教育出版社 [M],2015.7，1-118

＊ [6] 王海晖 , 彭嘉雄 , 吴巍等 . 多源遥感 curvelet 图像融合效果评价方法研究 [J]. 计算机工程与应用 ,2016，20:34-37


（上接第 82 页）

文化逐渐散播到全球，我们必须解决汉语的分词问题，从而使更多的人们在使用汉语时更加的方便。所以在未来，我们可将研究方向转向更方便的操作和分词等方面，优化现在的操作模式，提高未来的分词的准确度，以便使计算机能够对汉语文本进行自由处理，增强汉语的生命力，使其适应时代发展的需求。


**参考文献**

＊ [1] 王明文，付雪峰等．网页与文本自动分类综述．南昌工程学院学报，2005，24(3)：20—25.

＊ [2] 孙丽华．中文文本自动分类的研究．哈尔滨工程大学，2002.

＊ [3] 张巍．流形学习算法在中文问题分类中的应用研究．计算机应用与软件，2014，31(8)：269-287；河北工业大学学报，2014，43(2)：1—7.

＊ [4] 石陆魁，王歌，杨璐等．基于特征词相交和流形学习的文本分类方法 [J]. 河北工业大学学报 . 2014 年 02 期 第 1-7 页

＊ [5] 郭淑妮 . 中文信息处理中自动分词的研究和展望 . 内蒙古民族大学计算机科学与技术学院，2015.

＊ [6] 文庭孝、邱均平、侯经川 . 汉语自动分词研究展望 . 武汉大学中国科学研究评价中心，2004.

＊ [7] 刘迁、贾惠波 . 中文信息处理中自动分词的研究和展望 . 清华大学精密仪器与机械学系，清华大学光盘国家工程研究中心 .2006.


（上接第 93 页）


＊ [3] 李梦诗 , 余达 , 陈子明 , 等 . 基于深度置信网络的风力发电机故障诊断方法 [J]. 电机与控制学报 ,2019,23(02):118-126.

＊ [4] 刘建林 . 基于压缩感知的异步电机故障诊断数据压缩与重构 [J]. 湖南师范大学自然科学学报 ,2018,041(005):88-94.




(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net