

毕业设计说明书

|  |  |  |  |
| --- | --- | --- | --- |
| 作 者: | 赵斯蒙 | 学 号： | 913106840231 |
| 学 院: | 计算机科学与工程学院 | | |
| 专业(方向): | 计算机科学与技术 | | |
| 题 目: | 基于卷积神经网络的文本的情感分析 | | |
|  |  | | |

|  |  |
| --- | --- |
| 指导者： | 贾修一 副教授 |
|  |  |
| 评阅者： | \*\*\* 副教授 |

2017 年 5 月

声 明

我声明，本毕业设计说明书及其研究工作和所取得的成果是本人在导师的指导下独立完成的。研究过程中利用的所有资料均已在参考文献中列出，其他人员或机构对本毕业设计工作做出的贡献也已在致谢部分说明。

本毕业设计说明书不涉及任何秘密，南京理工大学有权保存其电子和纸质文档，可以借阅或网上公布其部分或全部内容，可以向有关部门或机构送交并授权保存、借阅或网上公布其部分或全部内容。

学生签名：

年 月 日

指导教师签名：

年 月 日

# 绪论

卷积神经网络作为近几年再次兴起的机器学习算法，受到了学术界的广泛关注，随之而来的深度学习热潮以及其在各个领域上的应用，更是为人类掀起了了技术革命的新篇章

## 工程背景及问题

短文本的分类一直是机器学习领域一个比较困难的问题，由于短文本自身所具有的特点：上下文信息匮乏、字词的歧义性，致使短文本的分类本身就具有一定的不确定性，而微博文本，微博评论就是典型的短文本，对此类短文本进行情感分析以及更进一步的情感分类，对于舆情研究，满意度反馈，证券投资[[1]](#endnote-1)等领域，具有十分巨大的潜力。对于此类文本，传统文本分析的方法无法提供足够的精确度，面对微博文本这种时效性和口语性较强的文本类型时，更是无法给出让人足够满意的结果，所以，利用机器学习方法对此类短文本并进行分类就成为了一个潜力巨大的新方向。

## 相关技术的现状

目前，实现文本分类的机器学习技术主要有支持向量机(SVM)[[2]](#endnote-2)，朴素贝叶斯[[3]](#endnote-3)，以及本文中提到的神经网络等，在实际使用中，也有结合了传统的词典分类以及情感标签[[4]](#endnote-4)的分类方法，都取得了一定的成效，随着机器学习领域的不断进步以及近几年卷积神经网络和深度学习的发展，对于短文本分析的精度被不断刷新；除了分类器(classifier)的算法为文本分类提供了有力支撑，随着词向量[[5]](#endnote-5)的出现，文本数据有了更好的特征选择和特征表示方法；中文文本分词技术的提高则更好的利用了词向量所带来的语义上的优势；此外，计算能力的提升数据源的增加也为卷积神经网络在文本分类上的应用的成功做出的不可忽视的贡献。

## 需解决的工程问题

要实现文本分类，首先想到的就是如何提取特征，使要输入的文本变成可以被模型所识别的数值形式，鉴于近年来词向量在文本分类领域的优异表现，本实验中决定采用的特征提取方式就是词向量，但是从源文本转化成词向量还需经历两个步骤，分别是“词”和“向量”，显然，中文文本和英文文本有着明显的区别英文以及其他诸多语言都有着天然的分词模式，而中文的分词则没有这种特点，有些时候甚至会出现一词多义、一句多分的情况，这就需要在特征选择阶段对文本进行分词，随着时代的变化，更多新名词的出现，分词的工具也需要不断更新迭代；在分词结束后，要给每一个分好的词映射对应的向量值，根据词向量5的解释，在获得实际向量值之前还需要另外训练一个词向量的模型，从而得到每一个词对应的向量值；在得到了一个句子的每一个词的向量值后，就可以近似的把这个句子作为一个图片，至此我们就得到了一个句子的数值矩阵形式，而且这种形式能很好的反映出这个句子所具有的语义特征；得到了可以输入的数据之后，接下来的工作就是把数据置入模型进行训练，但是要进行训练现在的数据维度还是太大，为了解决这个问题，我们引入了卷积操作，在合适的卷集合的大小之下，目标文本的序列特征将得到足够的体现。但是在进入神经网络模型训练前，还需要进行一个步骤，因为神经网络对于输入数据的大小有确定性要求，所以需要进行池化(pooling)操作，使每一个卷积操作的结果只有一个会被传输到输入端。在上述步骤完成后，到达神经网络的输入端的数据一定是确定数量的卷积核的输出值，这个确定的数量就是卷积核的数量；之后，要搭建适合的神经网络，在搭建神经网络的过程中，要选择适当的训练算法和学习率等超参数，以较好的适应学习的情景。在训练的过程结束之后，就需要进行验证的步骤，对模型的准确率进行评估，所以，在训练之前还要注意原始数据集的划分工作，划分成用于训练的训练集和用于测试的测试集。

## 论文章节安排

本篇论文主要介绍了借助卷积神经网络对微博文本进行情感分析的理论基础和系统流程，

# 卷积神经网络原理

神经网络算法最早源于上世纪40年代，其主要思想是模仿人类神经元的工作特性，每一个输出值都与每一个输入值呈线性关系，通过增加模型的复杂度和参数的数量，增加模型的精确性和可靠性，随着计算能力的增加，可以训练的神经网络的规模也越来越大，在各个机器学习的应用领域取得了越来越好的结果。

## 神经网络的基本原理

人工神经网络(artificial neural network)ANN，简称神经网络，最基本的单位是神经元：[[6]](#endnote-6)如图2-1所示

图 2‑1

……

f

其中，神经元的每一个输入都乘以相应的权重并相加，与神经元自身的阈值进行对比，这个过程需要一个激活函数来进行，决定是否向下一层传播，从而得到用于输出的，成为模型的输出或者是进入下一层继续训练，用数学公式表达起来就是如下形式：

理想中的激活函数应当是如下图2-2的阶跃函数，但是由于在进行后续操作的过程中，会出现求导，求积分等运算，使用阶跃函数就无法很好的适应这样的计算，因为显然，该函数在0处不可导，而且在0处不光滑，这就导致在涉及到求导以及积分的时候必须要进行分类，增加了模型的复杂度。所以，当前主要采用的激活函数石是另一种更加平滑，连续的函数，这就是被称为sigmoid函数的激活函数，典型的sigmoid函数如图2-3所示，这种函数也被称为挤压函数，因为其目的就是尽可能的模拟阶跃函数的性质，尽可能的减少在函数值中间区域（0-1）的输入的范围，从而实现激活的效果。

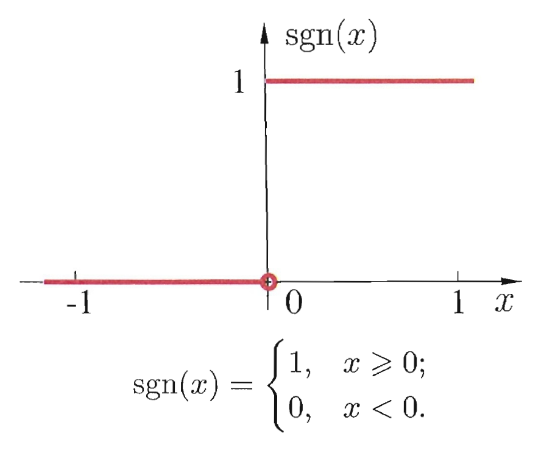
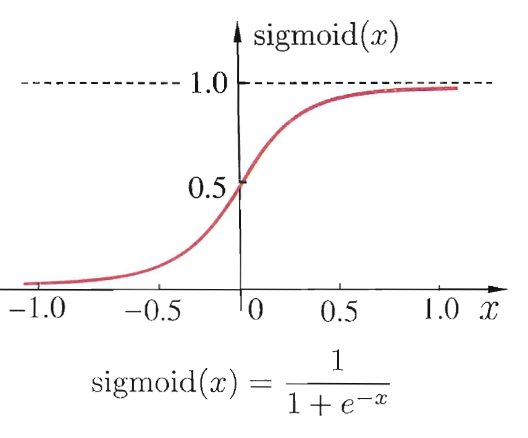
由多个神经元按一定的层次结构组合起来的模型就成为神经网络，简单来说，神经网络可以看做由若干线性模型嵌套并列组成的数学模型。

图 2‑2

图 2‑3

## 训练算法

假设现有一层神经网络模型，输入为x，正确的输出结果应当是y，而我们的模型输出的结果是，结果不正确，需要对模型的参数进行调整，使：

其中η表示学习率，在学习率合适的时候，模型会向着给定输入输出的模式拟合。

但是，这只是在一层网络的情况下，在多层网络的情况之中，这样的训练方式就无法使用了，需要引入梯度下降算法。

误差逆传播算法

给定训练集即输入端含有d个属性，输出端由l个属性，而且这个神经网络由输入层，隐层和输出层三个层组成，便可以得到下图的神经网络图，d个输入层，q个隐层，l个输出层，其中输出岑第i个神经元的阈值由表示，隐层第h个神经元的阈值由表示，输入层第j个神经元和隐层第h个神经元之间的权值用表示，隐层第h个神经元和输出层第i个神经元之间的连接权值用表示，可以得到隐层第h个神经元的输出，输出层第i个神经元的输出,其中为隐层的第h个神经元的输出，假设隐层和输出层的神经元都使用了sigmoid函数,假定神经网络在上述训练集中的输出为即：

由y与所产生的均方误差，就称为神经网络之中的损失函数，其值为：

与一层的模型类似，每一层的权重值都需要进行更新操作，更新的估计式为：

接下来就是要计算的值了，这就要使用到梯度下降算法，可以看做由模型中所有待训练参数作为自变量的未知函数的值，所以，由梯度的概念可知，在当前位置求出所有权重参数关于损失函数的梯度值表示的是损失函数在这一个自变量上的偏导数：

由于损失函数表示训练模型和实际值之间的差别，所以训练的目的也就是让的值越来越小，所以，当目前的梯度值是正的，就表示随着的增加，也会不断增加，而反之，如果梯度值是负的，那么，随着的增加，则会不断变小，所以为了让不断变小，就有：

由于首先影响的是第i个输出层的神经元的输入值再影响到的，所以有：

显然

另外，sigmoid函数具有一个很好的性质：

于是，可以推出：

类似的，可以得到隐层和输入层权值的变化量，隐层阈值，输出层阈值的表达式，分别为：

## 卷积

卷积是利用两种函数生成第三个函数的一种数学算子，其主要分为两种，分别是对于连续函数和离散函数进行运算的形式，由于神经网络之中所有的输入源都是离散的，所以下面只对卷积操作的离散形式进行介绍。

对于定义在整数上的函数f，g我们定义卷积函数为：

其中，m表示卷积的窗口大小，默认为，而n表示进行卷积操作的基本点对于所有域外的取值，都设成零，当的支撑集为有限长度M时，上式就变成了对有限数量的函数进行求和：

在具体情境中，会视情况改变卷积核的大小，而且在二维形式下，公式会有变化一般的，有：

其中，k表示卷积核（kernel），m表示原始矩阵，而卷积核的大小则是a\*b，该式表示的值是在[x,y]处的卷积操作值，对所有的原始矩阵元素进行操作得到的新的矩阵，就是卷积操作所期望得到的结果。

## 逻辑函数

\begin{align}
h_\theta(x) = \frac{1}{1+\exp(-\theta^Tx)},
\end{align}在进行分类的过程中，需要将模型的输出属性确定为若干个输出，其中，最简单的分类应当是二分类问题，在进行二分类问题之前，首先假设，在输出层的假设函数（激活函数）为：


\begin{align}
J(\theta) = -\frac{1}{m} \left[ \sum_{i=1}^m y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log (1-h_\theta(x^{(i)})) \right]
\end{align}
而代价函数为：

根据梯度下降算法可以计算出每一个参数的调整值，逻辑回归的本质上是线性回归，只不过根据二分类的问题调整了代价函数以及激活函数的形式。

# 算法设计

卷积神经网络由两部分组成，分别是卷积层和神经网络层，其中，神经网络的设计源于人类神经元的突触，激活的方式，让各个神经元之间彼此相连，通过调整参数来实现模型的拟合，而卷积层的操作可近似作为一个参数稀疏的神经网络层，其主要功能是降维和挖掘相邻像素(在本例中是词汇)之间的关系。

## 特征选择与预处理

在进行神经网络训练之前，首先要做的是将文本信息转化成有意义的数字信息，目前常见的编码方式有词袋模型（BOVW），独热码（one hot encoding）以及词向量（word2vec），也有使用随机分配的方式进行赋值的方法，经过比对，还是包含了语音信息的词向量的效果最好。

### word to vector词向量模型

文本处理不同于图像处理以及其他直接得到的数值处理任务，文本本身的计算机编码是不具有任何意义的，相邻的数据所具有的含义可能较大，从模型的层面上，相近的输入得到的输出结果应该是相似的，但是原文本所包含的含义却截然不同，对预测结果势必有一定的影响，对于这一点，英文的影响要明显大于汉语的影响，因为无论是utf-8还是各种传统的汉字编码，基本上都是按照文字的偏旁部首排序的，由于汉字本身所具备的特点，偏旁部首都是有一定的意义的，所以相近的字在语义上也存在着一定的相似性，而英文的情况就有所不同了，因为英文本身的字母不具有任何含义，拼成的词汇也很难发现其内在逻辑，所以，找到一种合适的表示文本特征的方法就是提高神经网络模型计算准确度的思路之一。

word embedding是在2003年由Bengio等人提出5，旨在从大量的数据源样本之中找到词元的分布特性，进而量化其语义特性，在一定条件下还可以进行分类，在本实验中采用的方法是由Tomas Mikolov团队所提出的Word2Vec方法 ，下面将对其进行适当介绍。

* Skip-gram 模型

Skip-gram模型是基于这样一个假设：在有足够大数量的训练文本的情况下，根据特定的上下文语境，在任意空缺的地方填入一个词，所有的单词中，可能性最高的几个单词之间一定具有相近的含义；这样的假设有一定的实际道理，比如，在上下文文本“一只白色的\_\_\_\_穿过了马路”，我们一般都会联想到应该填入“狗”，“猫”之类的词，而如果在训练集之中存在这样类似的文本，那么，由上下文所影响的关于单词的属性值就能反映出这个单词所具有的含义。所以，为了得到这种映射，首先要对这个映射进行初始化，给每一个单词对应的向量进行随机赋值，接下来，就可以看成一个神经网络的训练问题，每一个词的向量值可以看作是神经网络之中的权值信息，而损失函数则是由下式表示：

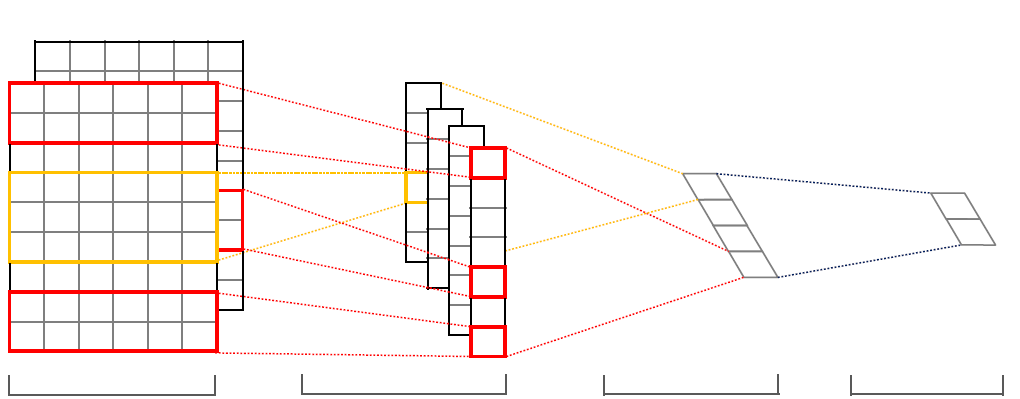
其中表示在出现的情况下之后第j个单词出现的概率，而该模型训练的目的就是让这个损失的值尽可能的小，接下来就要计算。

其中，代表对应w的向量值，而分别表示的输入单词和输出单词，对于word embedding 中所训练的神经网络的具体细节，在此就不做过多的展开，只需知道由这个操作得到的词向量映射包含一定的语音信息，而且具有确定的长度，这对于卷积神经网络的输入是非常重要的。

## 模型搭建

卷积神经网络的模型如下图所示，首先是卷积核进行的卷积，然后是pooling的汇合操作，接着就是数据进入隐含层，最终到输出层要决定是采用多个二分类分类器还是softmax分类器，根据学习问题的情景不同，做出的选择也有所不同，计算出衰减函数后就可以应用梯度下降进行训练，但是利用传统的均方误差学习方法又存在一些问题，因此引入交叉熵和权重衰减操作，更好的规范模型。

图 3‑1



文本的词向量序列，其中每个单词映射出一行，一个句子由一个矩阵表示

每一个卷积核都对全部的词矩阵进行操作，得到对应的一维值序列

Pooling操作，每一个卷积核得到的值序列，都只取其中的一个值，以保证输入神经网络的数据大小是确定的

全连接层，在神经网络中计算之后得到用于分类的值数据

### 卷积神经网络

经过上一部分的工作，假设我们已经得到了所有需要的单词的词向量，而且每个单词的词向量长度为k，现在需要输入一个句子，令：作为与句子中第i个单词有关的单词向量，一个长度为n的句子（必要时需要进行填补）就可以表示为：

其中是并置算符，就是让作为单词的并置。同时一个卷积操作还应当包括一个卷积核，这个卷积核的作用在上面已经描述过，就是在卷积操作中的第一个函数，注意，这个卷积核的大小一般是h\*k，其中，k是向量的长度，而h则是小于n的任意正整数，这就表明，每一次卷积操作都会包含若干个单词的所有向量信息，其含义就是应用于一个规模为h的局部语块，从而产生一个新的特征。举个例子是一个从单词中生成的新特征，那么的生成公式为：

其中：是运算偏项，f是一个非线性函数（比如说双曲正切函数），这个滤波器应用于句子中的每一个局部语块，从而产生一个特征映射

### Pooling汇合操作

可以看出上式得到的结果c的大小与源向量长度n有关，这对于神经网络的输入来说是难以接受的，一般的，我们可以采用扩展或者是压缩的办法来统一输入的大小，但一般都是应用于图像处理的课题中，使用了拉伸和压缩，是否对语义没有影响，还难以有确切的定论，而且本文主要讨论的是卷积神经网络的应用价值，所以在汇合的过程之中就采用了最简单也是目前比较主流的max over time池化操作，对每一个卷积核生成的向量

将其中的最大值作为与特定滤波器绑定的特征值。仅仅保留每一个映射的最重要的特征——具有最高值。这种池化策略自然地可以处理可变长度的句子。

在一定程度上也减少了参数数量和计算的复杂度

### 正则化

为了防止函数的过拟合，需要在进行pooling操作之前进行正则化操作，正则化操作的方法为：

### 归一化指数函数（softmax）

由于之前对神经网络的基本原理进行了详细的解释，所以隐藏层的结构就不做详细介绍了，但是对于输出函数的选择还是需要解释一下。

在监督学习中，常见的分类模式有二分类，n-分类，二分类可以看作n-分类的特殊形式，任何n-分类问题也都可以由最多二分类问题解决，如使用卷积神经网络求解的有名的手写数字识别问题，其中分类器所分出的可能的类别有10种，且这十个类别之间互斥，那么对于特定输入，就需要得到划分在每一个类别上的概率，概率最大的结果将是最有可能的结果，与之对应的，若果每一个类别之间不是互斥的，那么使用n个二分类器就可以的到关于每一个类别的分类结果。所以，在前文中，我们提到了二分类问题采用的逻辑函数回归，而softmax回归就是这种回归在多分类问题上的推广，由于本文中讨论的情感分析系统仅仅进行的是二分类问题，所以对于softmax分类不做详细介绍。

### 交叉熵与权重衰减

在介绍神经网络的时候，我们采用的是来L2范式作为损失函数：

但是许多实验证明，用L2范式作为损失函数会出现一个问题；首先，对该函数求导，可以得到：

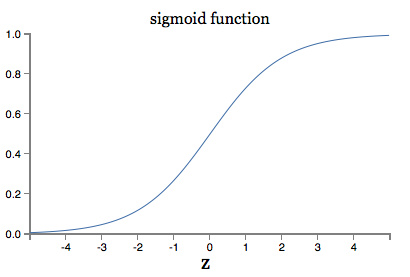
由sigmoid函数的图像（见图3-2），可以看出该函数所具有的性质，在z的值较大时，趋于平缓，而对应的的值也就趋于0，上式所表示的梯度值也就非常小。这就是为什么有的时候尽管模型的损失函数值很大，但学习速率还是很慢的原因。

图 3‑2

在这样的情境下，就需要引入新的损失函数，以谋求改善梯度值收敛缓慢的问题，在这样的背景下，用交叉熵表示的损失函数出现了。

其中，a表示训练数据的结果，而y表示实际的标签，这个加和式覆盖了所有的训练输入x，根据上式的处理方法，对此式进行处理，用带入上式，得到：

化简之后得到：

利用sigmiod函数的定义：，进行若干步推倒可以将上式化简为：

由此可以发现，梯度值与成线性关系，也就是目标值与实际值的差距越大，进行参数学习时的步长也就越大，也就从根本上避免了收敛缓慢的情况。

## 算法实现

# 系统实现以及测试

## 框架搭建

采用python下的tensorflow框架，（简要介绍运行环境，tensorflow 的工作方式，）

## 数据获取

## 数据预处理

1. 陆宇杰. 中文微博情感分析及其应用[D]. 华东师范大学, 2013. [↑](#endnote-ref-1)
2. Mccallum A, Nigam K. A Comparison of Event Models for Naive Bayes Text Classification[J]. IN AAAI-98 WORKSHOP ON LEARNING FOR TEXT CATEGORIZATION, 1998, 62(2):41--48. [↑](#endnote-ref-2)
3. Tong S, Koller D. Support vector machine active learning with applications to text classification[J]. Journal of Machine Learning Research, 2002, 2(1):45-66. [↑](#endnote-ref-3)
4. 孙建旺, 吕学强, & 张雷瀚. (2014). 基于词典与机器学习的中文微博情感分析研究. *计算机应用与软件,* *31*(7), 177-181. [↑](#endnote-ref-4)
5. Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. Journal of machine learning research, 2003, 3(Feb): 1137-1155. [↑](#endnote-ref-5)
6. Zhou, Z. (n.d.). 机器学习. 1st ed. 清华大学出版社, pp.98-99. [↑](#endnote-ref-6)