clayanddev的博客







# 基于卷积神经网络(CNN)的中文垃圾邮件检测

标签： cnn 垃圾邮件 深度学习

2017-04-25 16:05 7092人阅读 评论(9) 收藏 举报

分类： 人工智能（5）



目录(?) [+]

## 前言

跳过废话，直接看正文

文本分类任务是一个经久不衰的课题，其应用包括垃圾邮件检测、情感分析等。

传统机器学习的做法是先进行特征工程，构建出特征向量后，再将特征向量输入各种分类模型（贝叶斯、SVM、神经网络等）进行分类。

随着深度学习的发展以及RNN、CNN的陆续出现，特征向量的构建将会由网络自动完成，因此我们只要将文本的向量表示输入到网络中就能够完成自动完成特征的构建与分类过程。

就分类任务而言，CNN比RNN更为合适。CNN目前在图像处理方向应用最为广泛，在文本处理上也有一些的应用。本文将参考Denny Britz的WILDML教程 IMPLEMENTING A CNN FOR TEXT CLASSIFICATION IN TENSORFLOW 来设计一个简单的CNN，并将其应用于中文垃圾邮件检测任务。

## 正文

### 1 预备知识

1.1神经网络基础知识

如果你对深度学习或RNN、CNN等神经网络并不太熟悉，请先移步至这里寻找相关文章进行精读，这个博主写的每一篇文章都很好，由浅至深，非常适合入门。

1.2如何将CNN运用到文本处理

参考understanding-convolutional-neural-networks-for-nlp

1.3CNN网络结构和实现方法(必读)

此博文中的CNN网络结构和实现方法绝大部分是参考了 IMPLEMENTING A CNN FOR TEXT CLASSIFICATION IN TENSORFLOW 这篇文章的，CNN的结构和实现细节在这篇文章均有详述，在此我就不运相同的内容了，在请务必精读这篇文章。

### 2 训练数据

2.1 中文垃圾邮件数据集

说明：对TREC06C进行了简单的清洗得到，以utf-8格式存储

下载地址：百度网盘

2.2垃圾邮件

spam_5000.utf8

三种循环神经网络(RNN)算法... (0)
[工具代码]使|Textfield具有默... (0)
[笔记]2016阿里中间件性能挑... (0)
[笔记]2016阿里中间件性能挑... (0)
[笔记]2016阿里中间件性能挑... (0)
C++单元测试框架Gtest的配置... (0)
用Scrapy与Django一起搭建一... (0)
Spring MVC入门级项目示例 (0)

推荐文章

* CSDN日报20170429 ——《程序修行从"拔刀术"到"万剑诀"》
* 抓取网易云音乐歌曲热门评论生成词云
* Android NDK开发之从环境搭建到Demo级十步流
* 个人的中小型项目前端架构浅谈
* 基于卷积神经网络(CNN)的中文垃圾邮件检测
* 四无年轻人如何逆袭

最新评论

基于卷积神经网络(CNN)的中文垃圾邮件...
clayanddev ：@sinox2010p1:只要具有区分不同数据的能力，都能够直接作为cnn的输入

基于卷积神经网络(CNN)的中文垃圾邮件...
sinox2010p1 ：@clayanddev:假如这个特征列表只是一串数据，特征并不明显，也可以作为embedding向量...

基于卷积神经网络(CNN)的中文垃圾邮件...
sunjc2018 ：很有借鉴意义，谢谢楼主

基于卷积神经网络(CNN)的中文垃圾邮件...
clayanddev ：@yanguokai:按照README.md中步骤应该是能够正常运行的。请问具体遇到什么问题呢？运行...

基于卷积神经网络(CNN)的中文垃圾邮件...
YANGUOKAI ：运行代码时存在着一些问题，不知道大家有没有遇到呢？

基于卷积神经网络(CNN)的中文垃圾邮件...
clayanddev ：@sinox2010p1:传统机器学习分类时输入的是特征列表，这个列表包含哪些特征，这些特征该怎么计...

基于卷积神经网络(CNN)的中文垃圾邮件...
JinCheng_1978 ：可以的

基于卷积神经网络(CNN)的中文垃圾邮件...
ll19930428 ：非常有借鉴意义

基于卷积神经网络(CNN)的中文垃圾邮件...
sinox2010p1 ：对于 标签 特征列表 用机器学习似乎很方便。最后输入特征列表 可以获得分类。CNN也可以这么做吗？既...

```
1 有情之人，天天是节。一句寒暖，一线相喧；一句叮咛，一笺相传；一份相思， 一心相盼；一份爱意，一生相恋。 搜寻201:::http://201.855.com
2 我司是一家实业贸易定税企业；有余额票向外开 费用相对较低，此操作方式可以为贵公司（工厂） 节约部分税金。 公司本着互利互惠的原则,真i
3 本公司有部分普通发票（商品销售发票）增值税发票及海关代征增值税专用缴款书及其它服务行业发票, 公路、内河运输发票。可以以低税率为贵
4 来京记得找我啊 我社为您提供优惠的旅游价格，如酒店、机票、火车票、北京地接 在线咨询：QQ:305652179,欢迎您留言 MSN：yezikao8855@l
5 ......
```

共5000行，每一行对应一封邮件

2.3正常邮件

ham_5000.utf8

```
1 讲的是孔子后人的故事。一个老领导回到家乡，跟儿子感情不和，跟贪财的孙子孔为本和睦。 老领导的弟弟魏宗万是赶马车的。 有个洋妞大概是
2 不至于吧，离开这个破公司就没有课题可以做了？ 谢谢大家的关心，她昨天晚上睡的很好。MM她自己已经想好了。见机行事吧，拿到相关的能i
3 生一个玩玩，不好玩了就送人 第一，你要知道，你们恋爱前，你爹妈对她是毫无意义的。没道理你爹妈就要求她生孩子，她就得听话。换句话说
4 微软中国研发啥？本地化？ 新浪科技讯 8月24日晚10点，微软中国对外宣布说，在2006财年(2005年7月-2006年6月)，公司将在中国招聘约800名
5 ......
```

共5000行，每一行对应一封邮件

## 3 预处理

3.1输入

- 上述两个文件 ( spam_5000.utf8 ham_5000.utf8)
- embedding_dim (word embedding的维度， 即用多少维度的向量来表示一个单词)

3.2 输出：

- max_document_length (最长的邮件所包含的单词个数)
- x (所有邮件的向量表示， 维度为[所有邮件个数，max_doument_length, embedding_dim])
- y (所有邮件对应的标签，[0, 1]表示正常邮件，[1, 0]表示垃圾邮件，y的维度为[所有邮件个数， 2])

3.3 主要流程：

- 3.3.1 过滤字符

  为了分词的方便，示例程序中去除了所有的非中文字符，你也可以选择保留标点符号，英文字符，数字等其他字符，但要在分词时进行一定的特殊处理

- 3.3.2 分词

  为了训练Word2Vec 模型，需要先对训练文本进行分词。这里为了方便起见，直接对每个中文字符进行分隔，即最后训练处的word2vec 的向量是对字的embedding, 效果也比较不错

- 3.3.3 对齐

  为了加快网络的训练过程，需要进行批量计算，因此输入的训练样本需要进行对齐（padding）操作，使得其维度一致。这里的对齐就是把所有的邮件长度增加到max_document_length (最长的邮件所包含的单词个数)，空白的位置用一个指定单词进行填充(示例程序中用的填充单词为"PADDING")

- 3.3.4 训练word2vec

  在对文本进行分词和对齐后，就可以训练处word2vec模型了，具体的训练过程不在此阐述，程序可以参考项目文件中的word2vec_helpers.py。

## 4 定义CNN网络与训练步骤

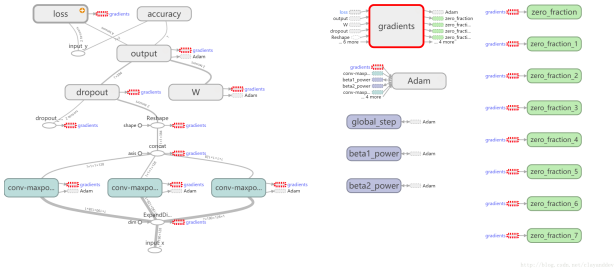
4.1 网络结构

此博文中的CNN网络结构和实现方法绝大部分是参考了 IMPLEMENTING A CNN FOR TEXT CLASSIFICATION IN TENSORFLOW 这篇文章的，CNN的结构和实现细节在这篇文章均有详述。重复的地方不再说明，主要说说不同的地方。

那篇文章中实现的CNN是用于英文文本二分类的，并且在卷积之前，有一层embedding层，用于得到文本的向量表示。

而本博文中实现的CNN在上面的基础上略有修改，用于支持中文文本的分类。CNN的结构的唯一变化是去掉了其中的embedding层，改为直接将word2vec预训练出的embedding向量输入到网络中进行分类。

网络结构图如下图所示：



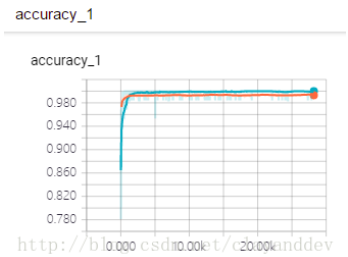

4.2 训练步骤

在预处理阶段得到了x和y, 接下来将x 和 y 按照一定比例分成训练集train_x, train_y和测试集dev_x, dev_y。

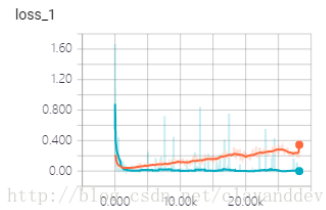
接着按照batch_size分批将train_x输入至网络TextCNN中进行训练，经过三个卷积层的卷积和max-pool之后，合并得到一个向量，这个向量[illegible]积层学到的关于训练数据的某些特征，最后将这个向量输入到一个单层的神经网络并用softmax分类，得到最终的分类结果，计算损失（交[illegible]后向传播，执行批量梯度下降来更新网络参数。

## 5 结果

准确率：




误差：




因为数据集并没有标准的训练集和测试集，本文只是按照0.1的比例进行了简单的分割，且并没有对一些重复的文档进行筛选，所以准确率能够达到99%左右。如果用比较标准的数据集，并加入交叉验证等方法，相信准确率会降低一些，但相信准确率仍能够超过绝大部分用传统机器学习的方法写出的分类器。

## 后记

这两天又在这个网络的基础上，尝试完成一个文本生成的任务（之前大多数人都是用RNN来做的），发现效果并不太好，生成的句子并不通顺。CNN对于分类任务较为擅长，但是卷积和pool丢失了一些上下文信息，因此在对上下文把握方面，RNN还是更甚一筹。当然，有研究表明CNN能够通过调整卷积层的各种参数来获取上下文的特征，但限于我现有的知识水平，并不清楚该怎么调参。只有等将来对CNN和RNN理解更进一步时再继续完善这个任务吧。

附：本文所实现的基于CNN进行垃圾邮件检测的项目下载地址




- 上一篇 用JAVA写一个视频播放器

我的同类文章

人工智能（5）

- [代码]基于RNN的文本生成算法 2016-12-31 阅读 239
- 利用gpu加速神经网络算法 2016-12-31 阅读 266
- 三种循环神经网络(RNN)算法的实现(Fro... 2016-12-30 阅读 157
- RNN在自然语言处理中的应用 2016-12-31 阅读 206
- 基于python实现一个简单的神经网络 2016-12-31 阅读 123



参考知识库

猜你在找

- Python编程基础视频教程(第六季)
- Hadoop生态系统零基础入门
- 零基础学HTML 5实战开发(第一季)
- Java之路
- 反编译Android应用
- 6卷积神经网络CNN
- CNN笔记通俗理解卷积神经网络
- 语音学习笔记11------卷积神经网...
- 语音学习笔记14------卷积神经网...
- DeepLearning tutorial4CNN卷积...



查看评论

sunjc2018 5楼 前天 13:53发表
很有借鉴意义，谢谢楼主

YANGUOKAI 4楼 前天 09:09发表
运行代码时存在着一些问题，不知道大家有没有遇到呢？

clayanddev Re: 前天 09:49发表
回复YANGUOKAI：按照README.md中步骤应该是能够正常运行的。请问具体遇到什么问题呢？运行环境是什么？

JinCheng_1978 3楼 4天前 16:00发表
可以的

ll19930428 2楼 4天前 15:46发表
非常有借鉴意义

sinox2010p1 1楼 4天前 14:01发表
对于 标签 特征列表 用机器学习似乎很方便。最后输入特征列表 可以获得分类。
CNN也可以这么做吗？既然传统机器学习就可以做了，何必用cnn,cnn计算更复杂，准确度更高？
Word2Vec处理的数据，所在文件名就表示他的分类是spam还是ham,所以训练结果会生成标签。
Word2Vec应该是针对文本的cnn训练，不同图像的cnn

clayanddev Re: 3天前 13:55发表
回复sinox2010p1：传统机器学习分类时输入的是特征列表，这个列表包含哪些特征，这些特征该怎么计算，这些都需要人工去定义和并编写特定程序提取，这就是特征工程。
CNN的输入不是特征列表，而是原始数据（在图像中通常是各个像素的值，而在文本方向则通常是用embedding向量）。讲这些原始数据直接输入 CNN，由卷积层自己计算出特征列表，然后进行分类。
特征相当于是由网络自我学习得到的，虽然特征的意义对人而言并不明确，但是比人自己定义的特征要更全面，更利于区分不同类型的样本。
深度学习的优势就在于训练数据越多，得到的特征也就更好，分类准确率也就越高。

关闭

sinox2010p1 Re: 前天 16:22发表
回复clayanddev：假如这个特征列表只是一串数据，特征并不明显，也可以作为embedding向量直接输入cnn

clayanddev Re: 前天 19:27发表
回复sinox2010p1：只要具有区分不同数据的能力，都能够直接作为cnn的输入