**基于TextCNN剧评情感预测**

刘 专

2019104208

**【摘要】**

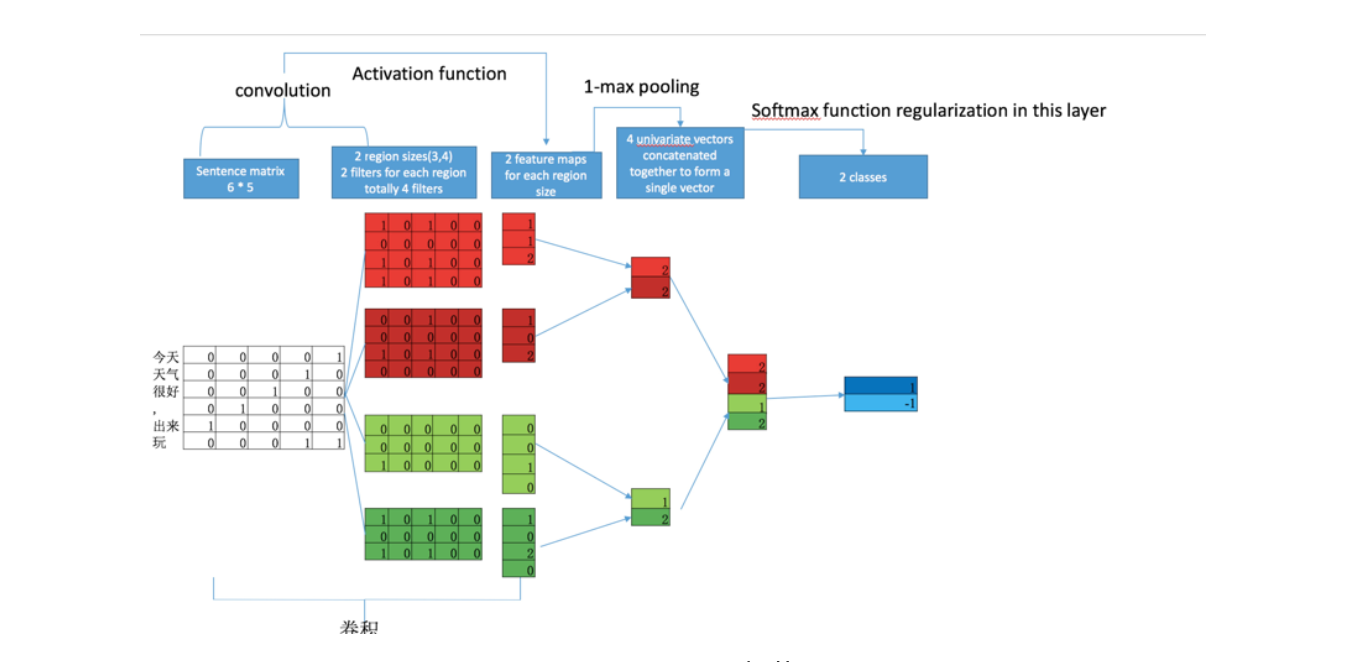
随着社交网络的兴起,更多人选择在网络上发表自己对影视作品的观点,这为影视投资人了解观众对电影的反馈提供了更方便的途径.例如,豆瓣影评中包含了海量用户或积极或消极的情感观点,而分析豆瓣影评的情感倾向能够辅助投资人进行决策,提升作品质量.大量数据分析必须借助计算机技术手段完成,其中情感分析是自然语言处理(natural languageprocessing, NLP)的一个方向,常用来分析判断文本描述的情绪类型,因此也被称为情感倾向分析。本次实验，我们使用TextCNN来对影评进行情感预测，预测任务实际上是文本的二分类，积极或者消极，在豆瓣采集的电视剧影评数据上进行训练预测，达到了75%的正确率。

# 引言

情感分析又称情感倾向性分析,是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程。例如,从各类评论中识别观众对电影的褒贬评价,判断用户对产品各种属性的使用评价,甚至预测股票走势等。情感分析的主要目的是识别用户对事物或人的看法和态度,而目前的研究对象以电商购物网站用户对商品的评价为主。一般来说,用户对物品的态度相对固定,但观众对电影作品的评价是对演员、导演、编剧的评价,是对人的评价,而人对人的评价相对更具有多变性。因此,对影评进行情感分析难度更大。本次实验，使用TextCNN进行剧评的二分类。

# 实验原理

Yoon Kim针对CNN的输入层做了一些变形，提出了文本分类模型textCNN，在此之前，CNN一般应用于CV领域，TextCNN的结构如下，包括一层卷积，一层maxpooling，一层softmax，网络结构比较简单。



**图一** TextCNN的基本结构

相对于图像中的CNN网络，TextCNN最大的区别在于输入数据的不同

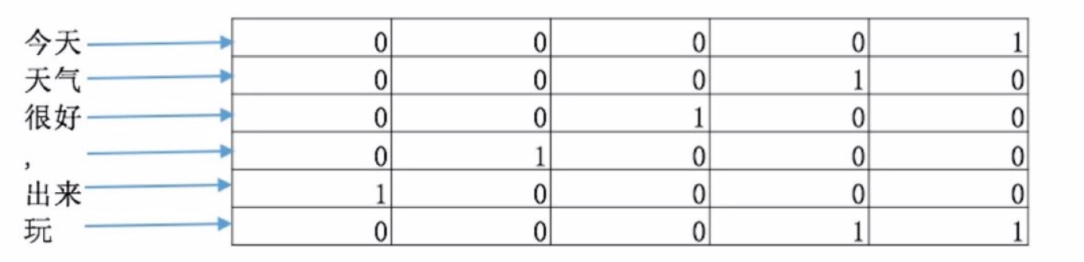
1. 在CV领域，图像的卷积核是从左到右，从上到下的方式滑动进行特征抽取。
2. 自然语言是1维数据，即使通过word-embedding的方式生成了2维向量，但是对词向量从左到右的方式滑动卷积是没有意义的，比如“今天”对应的词向量[0,0,0,0,1]，按窗口大小为 1\* 2 从左到右滑动得到[0,0], [0,0], [0,0], [0, 1]这四个向量, 对应的都是"今天"这个词汇。

# TextCNN的流程

## 构建分词向量

如图2所示, textCNN 首先将 "今天天气很好,出来玩" 分词成"今天/天气/很好/，/出来/玩, 通过word2vec的embedding 方式将每个词成映射成一个5维(维数可以自己指定)词向量, 如

"今天" -> [0,0,0,0,1], "天气" ->[0,0,0,1,0], "很好" ->[0,0,1,0,0]

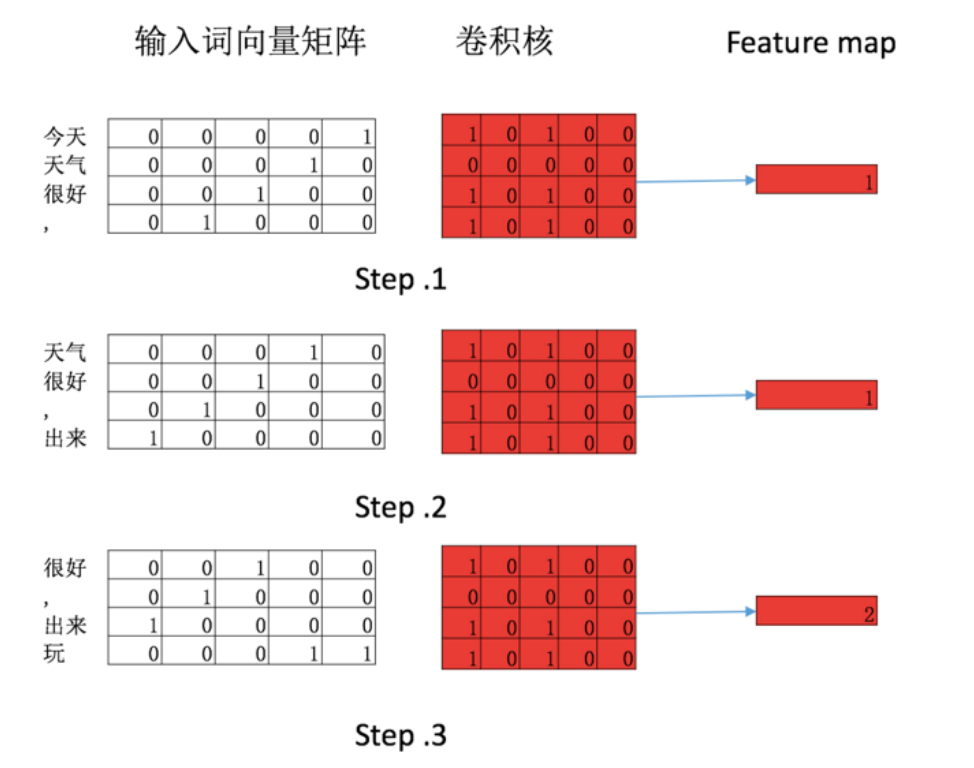


**图2** word embedding

这样做的好处主要是将自然语言数值化，方便后续的处理。从这里也可以看出不同的映射方式对最后的结果是会产生巨大的影响, nlp 当中目前最火热的研究方向便是如何将自然语言映射成更好的词向量。我们构建完词向量后，将所有的词向量拼接起来构成一个6\*5的二维矩阵，作为最初的输入。

## 卷积

卷积的步骤和流程如图3



**图3** 卷积示意图

## Channel

在CNN 中常常会提到一个词channel,每个4\*5 矩阵与输入矩阵做一次卷积操作得到一个feature map. 在计算机视觉中，由于彩色图像存在 R, G, B 三种颜色, 每个颜色便代表一种channel。

根据原论文作者的描述, 一开始引入channel 是希望防止过拟合(通过保证学习到的vectors 不要偏离输入太多)来在小数据集合获得比单channel更好的表现，后来发现其实直接使用正则化效果更好。

不过使用多channel 相比与单channel, 每个channel 可以使用不同的word embedding, 比如可以在no-static(梯度可以反向传播) 的channel 来fine tune 词向量，让词向量更加适用于当前的训练。

对于channel在textCNN 是否有用, 从论文的实验结果来看多channels并没有明显提升模型的分类能力, 七个数据集上的五个数据集 单channel 的textCNN 表现都要优于多channels的textCNN。

# 实验结果

数据来源是通过后羿采集器在豆瓣爬取的电视剧评论信息，爬取包括了一些热门电视剧的评论，数据约10000条数据，其中7990条作为我们模型的训练数据，1900条作为测试数据。



**图4** 模型测试结果

如图4，从实验结果来看，我们的模型可以达到约75%的预测准确率，还有可提升的空间，我们这次实验的目的在于了解TextCNN的处理方法，所以爬取训练的数据不是特别充足。

1. Kim Y . Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014.