自然语言处理课程实践报告

姓 名： 赵晗

学 号： 2272065

专 业： 计算机技术

学 院：计算机科学与工程学院

任 课 教 师： 肖桐

1. **实验内容**

本次实验的主要内容是使用卷积神经网络实现文本的分类问题。实验选用的数据是从网上找到的酒店的评论信息，该数据集已经被分成了积极（1），消极（0）两类。本次实验内容就是在该数据集的基础上，使用卷积神经网络实现文本的二分类任务。

文本分类任务也是自然语言处理任务之一，选择也是因为对于自己之后的研究方向有一些作用。由于对于transformer了解的还不够，所以权衡之下还是选择之前还算有过一点了解的卷积神经网络实现该项任务。但是其实这个模型做的也不是特别好，确实不是很擅长代码。

1. **具体细节**
2. 运行环境

anaconda3，python3.6.2，jupyter5.0.0，scikit-learn0.19.0

2.数据集划分

选用数据集是从网上找到的酒店的评论信息，该数据集已经被分成了积极（1），消极（0）两类。该数据集一共有7766条数据，其中积极的5322条，消极的2444条，实验中决定训练集和测试集的划分比例是8：2。划分完成后，由于积极、消极的数据是整齐排放的，因此还要进行一个打乱的操作。打乱之后，对数据集进行简单的处理：将label和文本分开，之后对文本信息进行分词并且去除停用词。分词使用jieba提供的现成函数进行实现，停用词词表使用从网上下载的，本人也在其中添加了一些停用词。

3.数据处理

包括对评论数据进行分词、词向量表示以及最为重要的通过文本卷积神经网络生成分类。对于数据的预处理部分，由于不像是图片能够很轻松的弄成机器可以识别的数据进行处理，文本还需要先转变为能够被机器试别的数字，才能进行之后的工作。所以需要先将文本处理成向量的形式。本次实验选用Word2Vec来实现改目的。

在文本分类卷积神经网络模型中顶替embedding层的是Word2Vec模型。Word2Vec是在神经网络的基础上变换而来的思想独特的特征抽取算法。包含两种模型：连续词袋模型和跳字模型。在实验中具体实现的时候，该模型使用的是gensim库中现成的Word2Vec模型。其中在Word2Vec具体模型的选择上，参考了一些资料之后选择使用跳字模型，输出词向量为300维，目标词前后分词的最大距离为5，当词小于5这个词频时不会对这个词进行操作。这些参数的调整、测试也是参考一些资料并且之后自行调试了一下得到的，应该不是最好的。

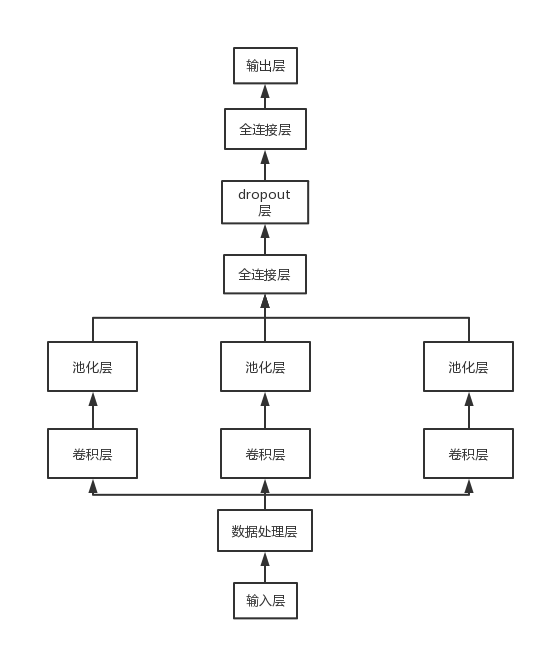
4.模型搭建

图1

在搭建卷积神经网络时，一共包括了输入层、嵌入层、卷积层、池化层、全连接层、分类器六个部分：

输入层：将原始新闻看点处理为分词词组；

嵌入层：将前面处理好的分词文本改写成词向量；

卷积层：利用卷积抽取看点的特征，卷积核个数是经由对看点的突出特点进行设定；

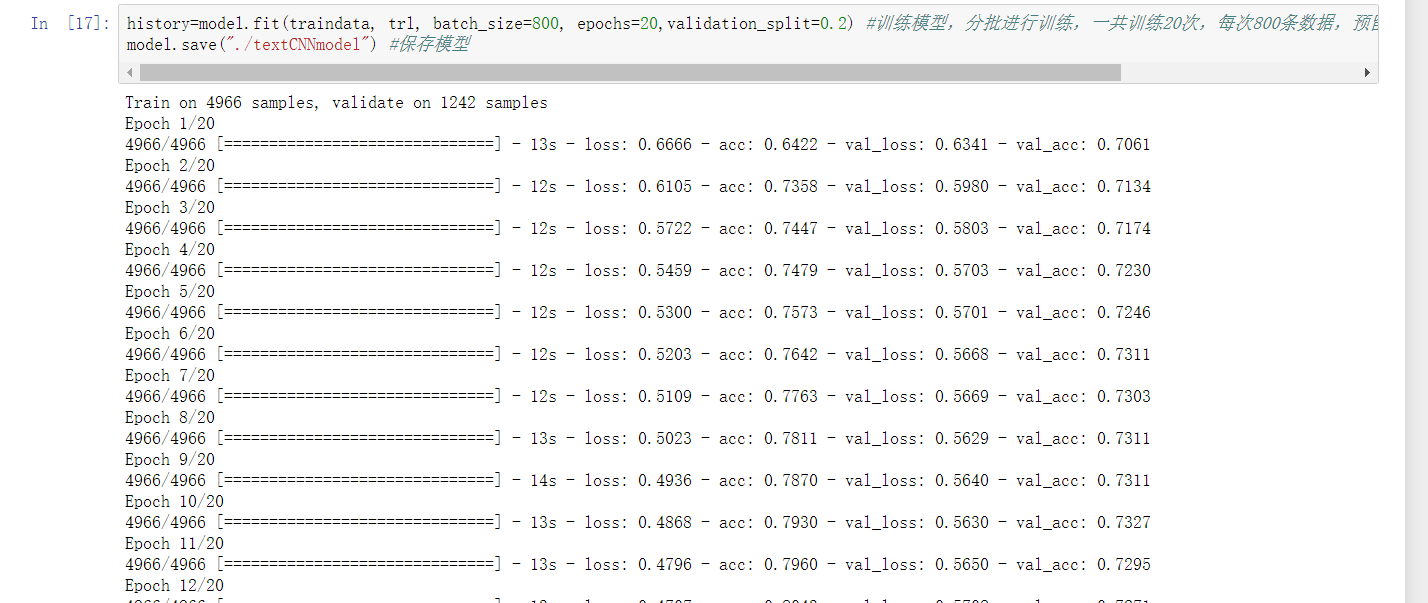
池化层：从卷积层中的所有通道得到的向量进行池化后得到多个标量，然后将这些标量拼接起来组成一个向量；

全连接层：将池化层的数据集进行全连接；

分类器层：根据前面获取的特征，尤这些结果综合判断看点的分类。

实际实现的时候，选择使用三层卷积层和三层池化层。总的来说，在最后实现的时候，文本分类卷积神经网络的第一层使用Word2Vec生成词向量，即作为卷积神经网络的嵌入层。之后是三次卷积和池化，卷积和池化结束之后，拼接结果，进入全连接层，为了防止过拟合，进行一次dropout，最后全连接得到输出结果。如图X所示，分别选择使用3，4，5卷积核来提取文本特征。模型结构如图1所示。代码上面都有比较具体的注释。

5.训练与测试

下图是在搭建完TextCNN使用训练集对模型进行的训练和测试。图中在训练模型的时候，设置参数batch\_size为800，参数epochs值为20，参数validation\_split值为0.2。意思是在进行梯度下降时每个batch包含800个样本，一共进行20次训练，最后的0.2表示选择百分之二十的训练集用来作为验证用数据，然后将这些数据用来对模型进行损失函数、精确度的计算，以此来进一步反应模型的优劣程度。训练结果如图2、3所示。

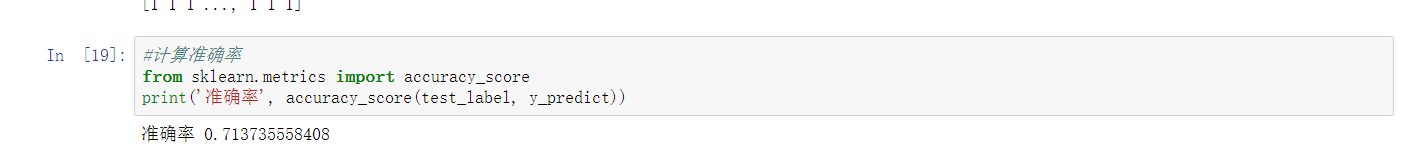
图2

图3

1. **总结**

本次实验其实完成的并不是很好，本人也不是很擅长写代码，但是感觉过程还是有趣的。通过本门课程，我学习到了很多有关自然语言处理的模型，并且这些模型还能用于解决其他的问题。加上我之后可能的一个研究方向是对话系统，本门课中教授的一些模型对于我以后的研究方向有很大的帮助。老师以及学长学姐们的讲课方式和讲课内容都十分的好，很庆幸能选到这门课。