2022-2023学年秋季学期东北大学

《自然语言处理》课程作业

**姓 名** 卢全宇

**专 业**  计算机科学与技术

**学 号**  2201843

**学 院**  计算机科学与工程学院

**任课教师** 肖桐

**2023 年 1 月 15 日**

1 介绍

1.1 任务内容

在python 3.7+pytorch 1.1+tqdm+sklearn+tensorboardX环境下实现了用TextCNN对文本进行分类。

1.2 数据集

在[THUCNews](http://thuctc.thunlp.org/)中抽取了20万条新闻标题，文本长度在20到30之间。一共10个类别，每类2万条。

类别：财经、房产、股票、教育、科技、社会、时政、体育、游戏、娱乐。

数据集划分：

数据集|数据量

--|--

训练集|18万

验证集|1万

测试集|1万

1.3 任务来源

出自对Github一博主的学习过程中，对其提出的问题及解决方案学习改动后研究而来。

2 实验过程

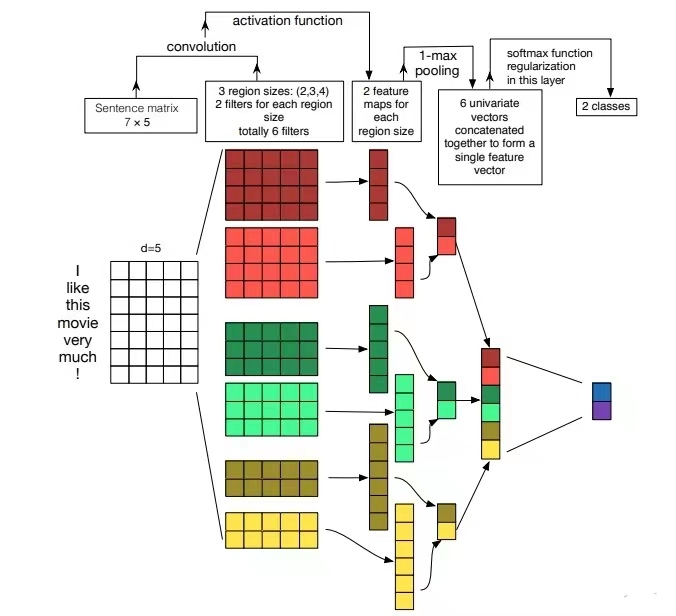


图1 TextCNN

数据处理：所有句子padding成一个长度：seq\_len

2.1 模型输入  
 **[batch\_size, seq\_len]**

2.2 经过embedding层：加载预训练词向量或者随机初始化, 词向量维度为embed\_size：  
 **[batch\_size, seq\_len, embed\_size]**

2.3 卷积层：NLP中卷积核宽度与embed-size相同，相当于一维卷积。3个尺寸的卷积核：(2, 3, 4)，每个尺寸的卷积核有100个。卷积后得到三个特征图： **[batch\_size, 100, seq\_len-1]**  
 **[batch\_size, 100,** **seq\_len-2]**  
 **[batch\_size, 100,** **seq\_len-3]**

2.4 池化层：对三个特征图做最大池化  
 **[batch\_size, 100]**  
 **[batch\_size, 100]**  
 **[batch\_size, 100]**

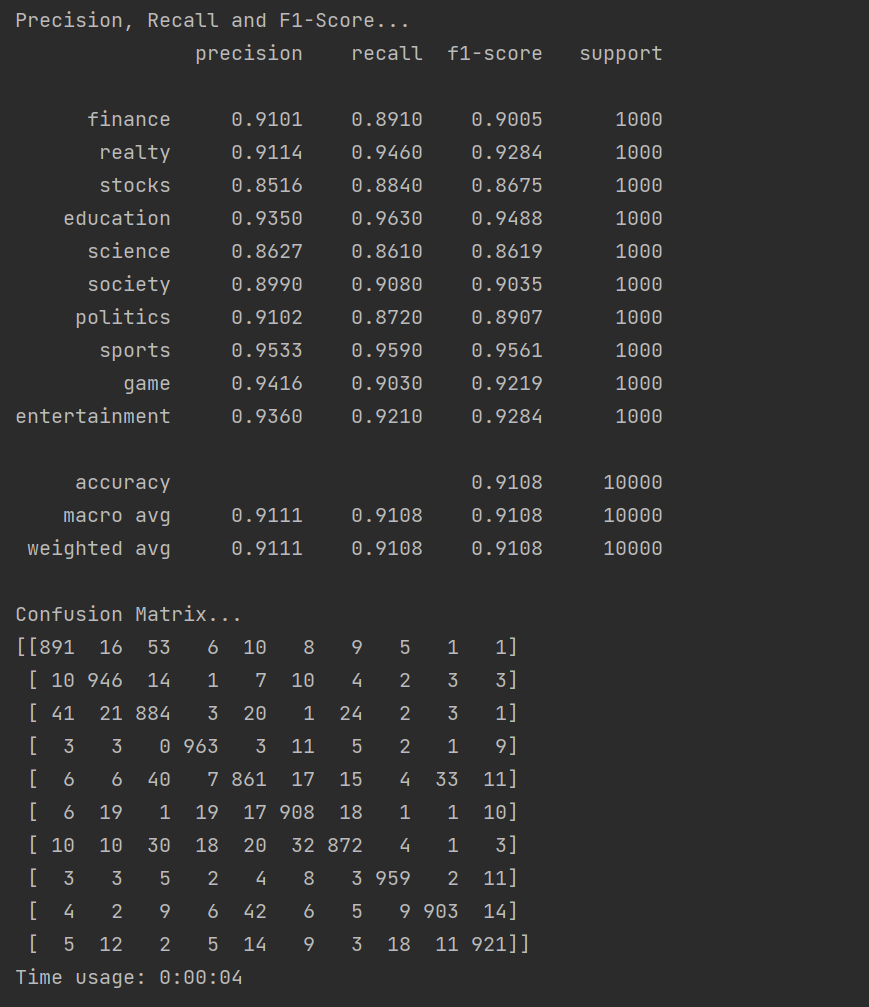
2.5 拼接：  
 **[batch\_size, 300]**

2.6 全连接：num\_class是预测的类别数  
 **[batch\_size, num\_class]**

2.7 预测：softmax归一化，将num\_class个数中最大的数对应的类作为最终预测  
 **[batch\_size, 1]**

2.8 分析：卷积操作相当于提取了句中的2-gram，3-gram，4-gram信息，多个卷积是为了提取多种特征，最大池化将提取到最重要的信息保留。

3 实验结果



4 结论

采用经典的CNN文本分类可对新闻类别进行较为准确、效果较好的分类。在搜集资料学习的过程中也尝试了一些其他方法，除少数模型预测效果较差外，其余对于文本分类问题皆有不俗的效果，希望在今后的学习中能不断尝试和研究。