浙江大学院级大学生创新创业训练计划

中期检查表

项目编号： Y202104225

项目名称： 基于自然语言处理的新冠相关推特属性判断

项目负责： 段皞一　　学号： 3190105359

院（系）： 计算机科学与技术学院

联系电话： 19883161889　　电子邮件： 1031004722@qq.co

指导教师： 汤斯亮　　职　　称： 副教授

浙江大学本科生院教务处

2021年11月19日

项目名称		基于自然语言处理的新冠相关推特属性判断			
立项经费		700	起止时间	2021-03-16至2022-05-31	
负责人	学号	姓名	所在院系、专业	联系电话	E-mail
	3190105359	段皞一	计算机科学与技术学院、混合班	19883161889	1031004722@qq.com
参加成员	3190105301	杨浩峰	竺可桢学院、混合班	15779509886	1183045207@qq.com
参加成员	3190104555	翟智超	计算机科学与技术学院、计算机科学与技术	19883145932	3190104555@zju.edu.cn
导师	姓名	汤斯亮	院系:	职称	副教授
	E-mail	0012010@zju.edu.cn		联系电话	13588196277

一、项目研究进展情况（含项目研究已取得阶段性成果和收获）（800字内）

（1）项目研究进展情况

2021年4-5月：

了解了机器学习的基础知识，使用Keras搭建了Imdb情感分析神经网络，初步了解了神经网络结构；

2021年6-7月：

使用Pytorch，MXNet搭建了TextCNN模型的Imdb情感分析神经网络模型；

2021年8月：

简要了解了TextCNN和TextRNN模型背后的算法知识；

2021年9-11月：

使用MXNet搭建了CNN和RNN两种神经网络，并编写了读入相关训练集和标签的接口；在对原有的基于CNN和RNN的神经网络模型进行多次修改后仍乏善可陈，在测试集上的正确率达到了60%，小组讨论后认为对于CNN和RNN这种基础的神经网络已经是比较良好的结果，如果想要进一步提升正确率，可能需要从更加先进的自然语言处理模型入手，因此团队决定下一步尝试复现Bert以及衍生模型。另一方面由于mxnet的生态环境较差，难以找到比较先进的神经网络模型的学习资料，故而将整个项目迁移到pytorch。

2021年11月：

由于跑数据的速度较慢，尝试配置MXNet的Cuda版本，最终以失败告终。发现原因是MXNet官方已经不再更新和维护，没有支持Cuda11.0及以上的版本；

2021年11月：

配置好Pytorch的Cuda环境，为项目后期将环境转变为Pytorch做准备。

（2）项目研究已取得阶段性成果和收获

在对原有的基于CNN和RNN的神经网络模型进行多次修改后仍乏善可陈，在测试集上的正确率达到了60%，小组讨论后认为对于CNN和RNN这种基础的神经网络已经是比较良好的结果，如果想要进一步提升正确率，可能需要从更加先进的自然语言处理模型入手，因此团队决定下一步尝试复现Bert以及衍生模型。另一方面由于mxnet的生态环境较差，难以找到比较先进的神经网络模型的学习资料，故而将整个项目迁移到pytorch。

二、项目研究存在的主要问题分析及应对思路与措施（500字内）

1.由于运营维护的问题，MXNet的Cuda版本过于落后，无法支持Cuda11.0及以上，所以后期需要迁移到Pytorch进行训练,Pytorch有着良好的；

2.目前部分题目是对事实性陈述进行进一步的分类，第一问中非事实性陈述就在这部分问题中无关，标签是“nan”,因此预测结果需要依赖于第一问的预测结果。目前，我们对如何处理这种关联性还存在一定的问题，一种办法是忽略这种关联性，把这些问题当作三分类任务；另一种是我们所目前实现的，考虑这种关联性，将这些问题的预测结果在第一问的基础上进行过滤，从而变成二分类问题。

3.目前使用的模型是比较传统的CNN和RNN，并且没有对参数进行过多的分析，之后需要分析参数对预测结果的影响，并且还要再引入一些额外的模型和工具进行优化，比如BERT等；使用其他的神经网络进行训练。

4.目前模型训练时间较长，后期考虑使用并行之类的方法优化模型的计算性能。

5.当前的嵌入层的编码方式更侧重单个词能体现出的特征，从而忽视了句子结构对问题相关属性的影响。

三、项目研究下阶段主要任务及时间进程安排（500字内）

2021年11-12月：

将当前的TextCNN和TextRNN迁移到Pytorch上，使用Pytorch的Cuda初步跑出一些预测结果；

2021年12月-2月：

学习BERT模型，掌握其原理，并进行实现。

2021年2月-4月：

对模型进行具体优化。预处理上，考虑上下文之间的关系、句子的结构，能够对否定、转折等特殊结构的句子具有更好的预测效果。模型上，通过调整参数、配置优化器和损失函数的方法得出更优质的预测结果；性能上，从并行的角度尝试对模型的计算性能进行优化。

四、项目组成员个人分工所承担和完成研究内容情况（100字内）

负责人所承担和完成研究内容情况汇报：了解机器学习的基础知识，使用Keras搭建了Imdb情感分析神经网络，初步了解了神经网络结构；配置好Pytorch的Cuda环境，为项目后期将环境转变为Pytorch做准备；训练网络。

杨浩峰所承担和完成研究内容情况汇报：了解机器学习的基础知识，使用MXNet搭建了CNN和RNN两种神经网络，并编写了读入相关训练集和标签的接口；训练网络；简要了解了TextCNN和TextRNN模型背后的算法知识。

翟智超所承担和完成研究内容情况汇报：了解机器学习的基础知识，进行MXNet向Pytorch的迁移；训练网络，调整TextCNN和TextRNN的参数和网络结构尝试进行预测结果的优化。

五、项目经费使用情况 （说明购置材料、资料、调研、交通等已开支经费数额）（100字内）

购买了Deep Learning、Python机器学习等书籍进行深度学习相关知识的学习。

六、指导教师意见 （从研究内容和进展、阶段性成果、存在问题等方面加以评价）（180字内）

新冠相关推特属性判断是一个应景而且有意义的实际问题，本项目针对该任务开展一系列调研与探索，进度安排合理，取得了一定的进展。

签　名：汤斯亮

2021年11月21日

七、院（系）评审意见（100字内）

签名盖章

年　月　日