****

**自然语言处理课作业报告**

姓 名 ： 张弛

学 号 ： 2201913

2023 年 1 月 7 日

1. 背景

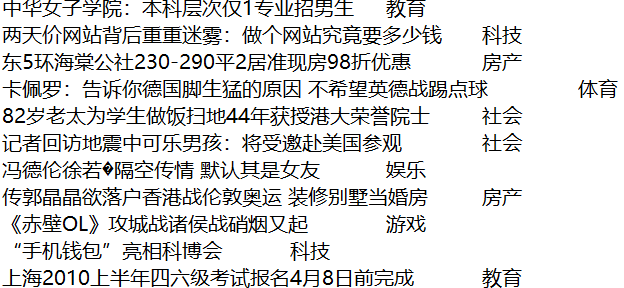
自然语言处理是人工智能领域的一个重要方向。自然语言即人类日常使用的语言。自然语言处理可以通过研制计算机系统或软件，研究自然语言。自然语言处理目前主要应用于机器翻译、摘要提取、文本分类、对话系统、文本语义对比。因个人研究方向与对话系统或情感分析相关，所以我本打算实现一个传统的情感分类，但因找到的数据集不是很合适，所以改为文本分类。本次简单采用了pytorch-bert实现了一个中文文本分类。最后训练的loss值为0.19，测试集上的准确率为94.30%。使用的是bert\_base\_chinese。其中共12层编码层，768个隐藏单元。

1. 实验环境

使用了python3.8，torch，tqdm，transformers, Linux服务器。

1. 数据集

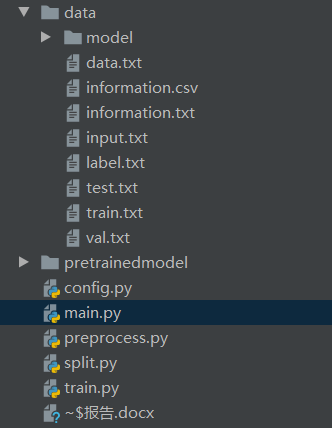
从THUCnews中获取到了20万条的新闻标题数据，其中包含是个类别：财经、房产、股票、教育、科技、社会、时政、体育、游戏、娱乐。每类各有两万条数据。存放在一个txt文件中。数据形式如下例。



每一条数据中，前段文字为标题文字，后段为类别标签。将20万条数据分为了训练集，验证集和测试集。其中训练集18万条新闻标题，每类别18000条，验证集1万条标题，每类别1000条，测试集一万条标题，每类别1000条。

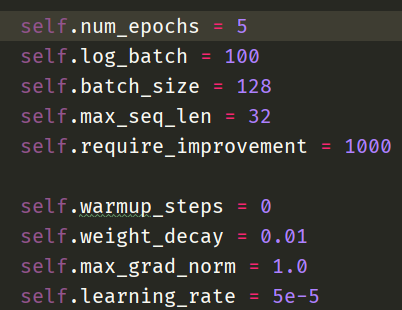
1. 实验内容

文件结构如下图所示：

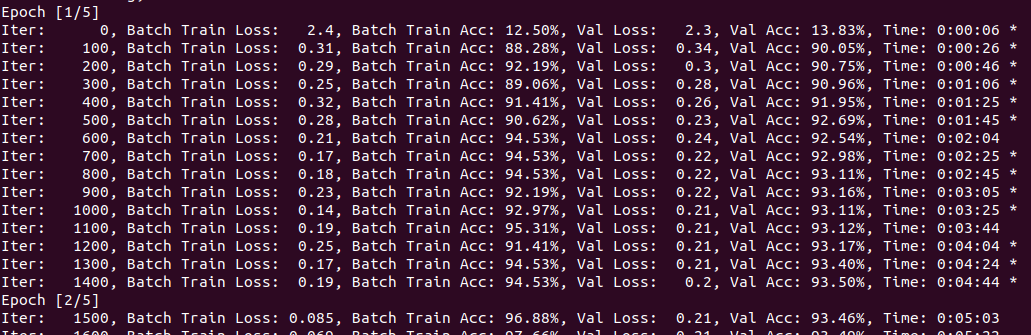


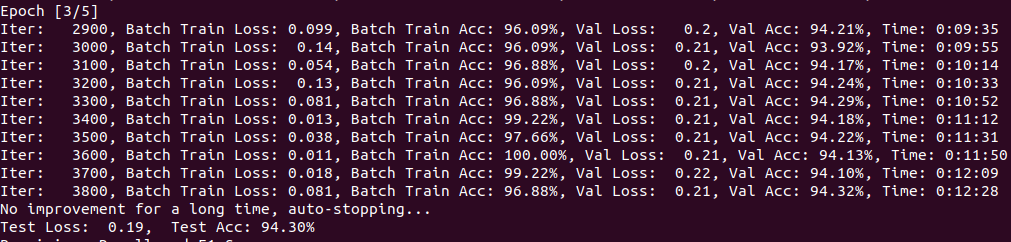
本次报告采用的是预训练bert模型，在huggingface（<https://huggingface.co/bert-base-chinese>）上下载bert\_base\_chinese模型权重、配置文件和词典。将这些文件放到pretrainedmodel文件夹下。使用sklearn中的train\_test\_split函数，将20万条新闻标题按照标签(label)划分数据集。分别保存为train.txt,val.txt和test.txt。

训练采用的优化器是 Adam 和 weight decay 相 结合的 AdamW。参数配置为epoch数5轮，batch-size为128，初始学习率设定为5e-5,另外设定了早停机制，当上次权重更新的迭代次数与当前迭代次数的差超过1000次时，停止训练。具体超参数设置如下。

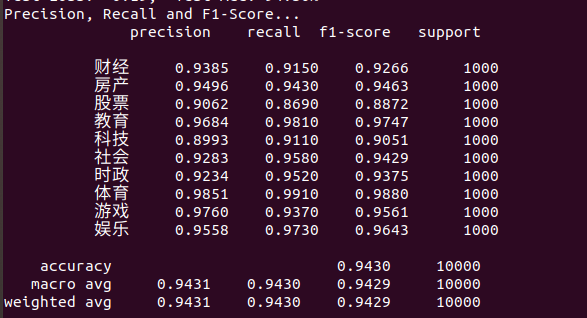


训练过程中的截图如下：



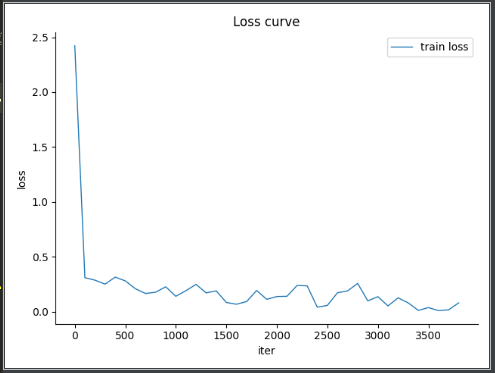


因上次更新loss的迭代数已经超过1000次，所以训练在第三个epoch停止。训练结束，经过测试集测试的结果如下：



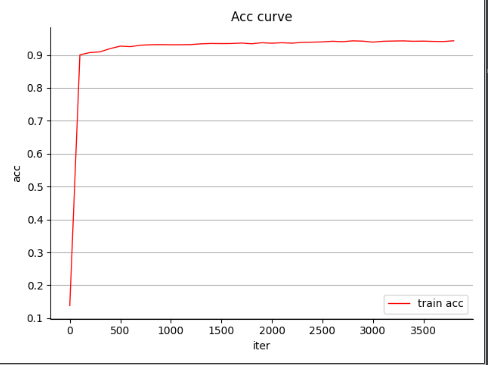
上图分类显示了精确度（precision），召回率（recall）和f1值。可以看见在股票和科技类别的精确率较低，对于预测体育和游戏类别的表现较好。对于整个测试集的1万条新闻标题的精确率为94.30%。将训练过程中的loss和acc值保存下来，绘制了loss和acc曲线。

主要代码在draw.py文件。



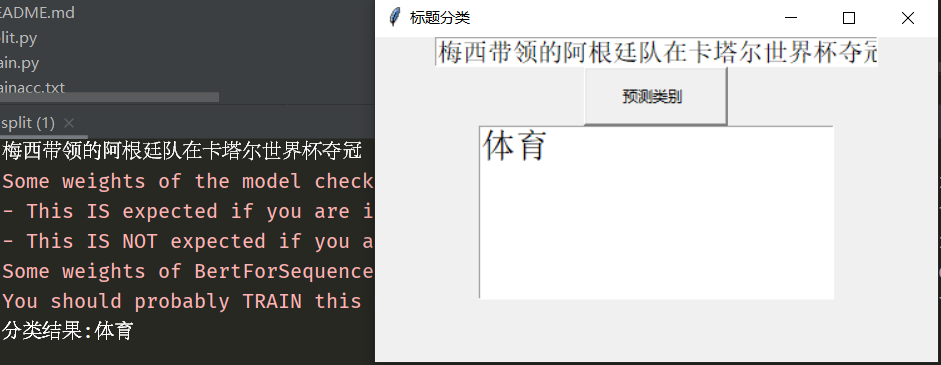
在loss曲线，可以看见在3900次迭代左右，训练提前停止。

下图为训练时的acc曲线，可以看见对于这种简单的文本分类任务，bert模型的效果非常好。我选取的新闻标题文字较短，不超过32个中文字，且往往标题的概括能力较强，所以训练效果很好。将训练好的模型权重保存到data/model下，以便测试时使用。

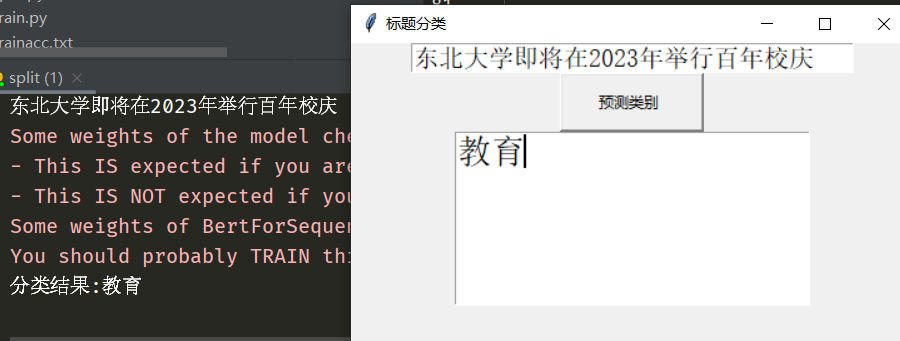


1. 实验结果

训练之后使用了python中的tkinter，设计了一个非常简单的窗口，实现输入新闻标题，后台预测，之后显示到窗口上的功能。主要代码在split.py。下面为示例：



在输入框中输入：梅西带领的阿根廷队在卡塔尔世界杯夺冠，预测结果为：体育。



在输入框中输入：东北大学即将在2023年举行百年校庆，预测结果为：教育。