实验三

实验要求

使用 pytorch 或者 tensorflow 的相关神经网络库编写基于 BERT 的预训练语言模型，利用少量的训练数据，微调模型用于文本情感分类。并和直接用 RNN/Transformer 训练的文本分类器进行对比，研究训练数据量变化对性能带来的影响。

实验步骤

1. **网络框架**：要求选择 pytorch 或 tensorflow 其中之一，依据官方网站的指引安装包。这个实验还需要安装 transformers 库，方便调用预训练模型。（如果前面实验已经安装过，则这个可以跳过）
2. **数据准备：** 本次实验统一使用指定的 IMDB 公开数据集 "Large Movie Review Dataset"。该数据集分别包含 25,000 条电影评论作为训练集和测试集。任务为二分类任务。数据下载地址为：Sentiment Analysis (stanford.edu)
3. **数据预处理：** 你需要通过pytorch 或 tensorflow所提供的标准数据接口，将原始数据处理为方便模型训练脚本所使用的数据结构，如 torch.utils.data.Dataset 等。 这个数据集是非常常见的公开数据集，你可以参考一些公开代码片段。
4. **语言模型：** 搭建 BERT 模型并加载大语料库上预训练的模型参数，推荐的预训练参数来源为 BERT (huggingface.co)；RNN 模型；Transformer 模型。
5. **情感分类：** 情感分类模型包含一个语言模型和一个分类器（MLP）。首先，将一个句子中的每个单词对应的词向量输入语言模型，得到句子的向量表征。 然后将句向量作为分类器的输入，输出二元分类预测，然后进行 loss 计算和反向梯度传播训练 ，这里的 loss 是分类 loss ，如交叉熵 loss。
6. **研究训练数据量：** 对于 IMDB 中的 train 数据，采用不同比例的数据（如25%，50%，75%，100%）用于划分训练集和验证集以训练/微调情感分类模型。剩下的train中数据不使用。然后使用训练好的模型测试在 test 数据上的表现。

(所以在整个实验中需要调整的参数就是不同的**语言模型**和**训练数据的比例**)

实验提交

本次实验截止日期为 **5月 23 日 23:59:59**，需提交代码源文件及实验报告到邮箱：proton00@mail.ustc.edu.cn，具体要求如下：

1. 全部文件打包在一个压缩包内，压缩包命名为 学号- 姓名 - exp3.zip

2. 代码仅包含 .py 文件，请勿包含实验中间结果（例如中间保存的数据集等），如果有多个文件，放在 src/ 文件夹内。
3. 代码中提供一个可以直接运行的并输出结果的 main.py，**结果包括训练集损失、验证集损失随 epoch 改变的曲线（保存下来）和测试集的评价指标。**
4. 代码中提供一个描述所有需依赖包的 requirements.txt，手动列入代码中用到的所有非标准库及版本或者使用 `pip freeze > requirements.txt` 命令生成。
5. 实验报告要求 pdf 格式，要求包含姓名、学号。内容包括简要的**实验过程**和**关键代码**展示，对训练数据量和不同语言模型的**实验分析**。

参考资料

往届同学的实验代码和报告： https://github.com/hehaha68/USTC_2022Spring_Introduction-to-Deep-Learning

提供的 Lab4_demo.ipynb

实验数据下载链接：https://rec.ustc.edu.cn/share/923bdc50-ee3b-11ed-9e6b-95f1dcc044f2