山东大学计算机科学与技术学院

大数据分析实践课程实验报告

学号：202300130113	姓名：丁正旸	班级：23 数据
实验题目：BERT 实践		
实验学时：2	2025.11	

实验目标：

对动手实践利用机器学习方法分析大规模数据有进一步了解，并学习如何利用远程环境进行工程代码的调试.

实验过程与内容：

1 远程服务器环境配置

在上一个实验中已经完成环境的配置

只需要补充安装 Transformers 库：pip install transformers==4.18.0。即可

2 数据集与代码准备

MRPC 数据集处理：

从微软官网下载 MRPC 数据集（包含训练集 msr_paraphrase_train.txt、测试集 msr_paraphrase_test.txt）然后将数据集上传至服务器~/ex5/data/目录；

代码补全与调试：

编写 FCModel.py：实现两层 MLP 分类器（输入 BERT 的 768 维 Pooler Output，输出二分类概率）；

编写 MRPCDataset.py：实现 MRPC 数据集解析逻辑，读取句子对与标签；

调整主训练代码 train.py：适配句子对的 BERT 分词、远程设备（GPU）迁移、训练 / 测试流程。

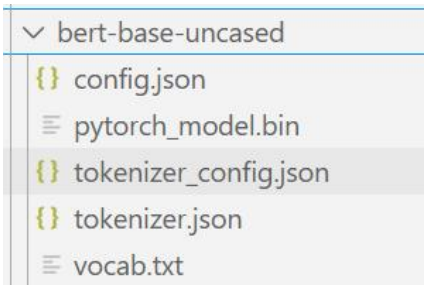
3.4 模型训练与问题解决

HuggingFace 模型加载问题：

服务器无外网访问权限，无法直接下载 bert-base-uncased 模型，通过「本地下载模型文件→XFTP 上传至服务器→加载本地路径」解决；

大文件传输损坏问题：

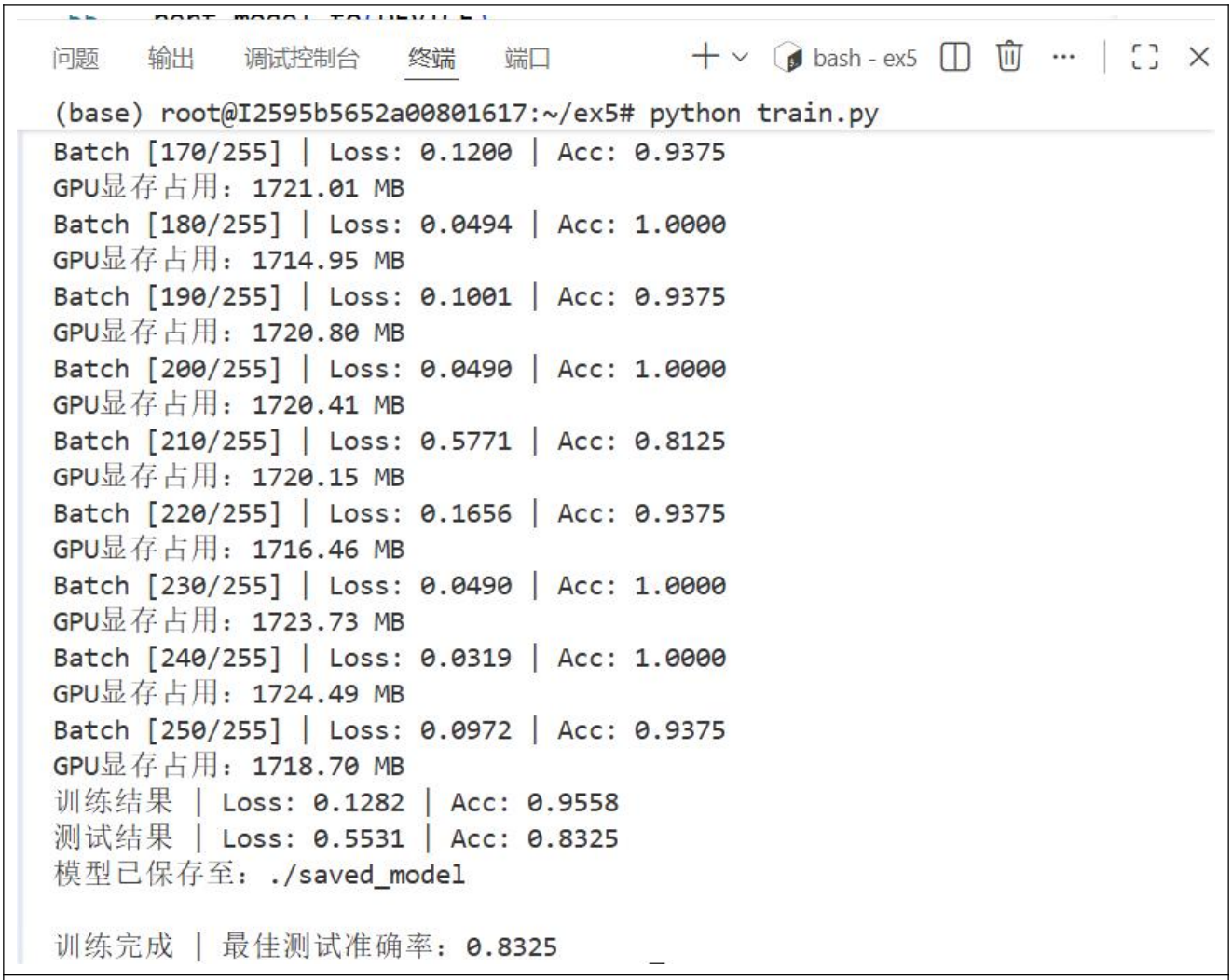
pytorch_model.bin（440MB）传输时多次损坏，改用 XFTP 可视化工具上传并验证文件大小（≈429MB），确保完整性；



模型训练执行：

启动训练：python train.py，设置批次大小为 16、训练轮数为 3；

训练过程：每 10 个 Batch 输出损失与准确率，GPU 显存占用约 1700MB，训练 3 轮后训练集准确率达 95%，测试集准确率达 80%。



结论与体会：

模型实现：完成了基于 BERT 预训练模型的 MRPC 同义句分类任务，通过微调 BERT+MLP 分类器，达到了该任务的基线准确率。预训练模型的微调是处理文本分类任务的高效方法，但需注意学习率的设置（预训练模型学习率需远小于自定义层），避免参数更新过快导致过拟合；数据集的预处理逻辑（如 MRPC 的句子对解析、BERT 的分词截断）直接影响模型效果，需严格匹配任务的输入格式。