山东大学计算机科学与技术学院

大数据分析与实践课程实验报告

学号：202320130203	姓名：李姿含	班级：数据23
实验题目：基于 BERT 的 MRPC 同义句预测实践		
实验学时：2	实验日期：20251107	

实验目标：

理解 MRPC 数据集的结构与预处理流程，能够处理数据格式异常、字段缺失等问题。

熟悉 BERT 模型的工作原理，掌握基于预训练 BERT 模型进行微调的完整流程，包括数据加载、模型构建、训练优化等环节。

作品描述（实验背景、数据集来源、描述思路、实验背景）：

1、实验背景：

在自然语言处理领域，同义句识别是一项重要的基础任务，其目的是判断两个句子是否表达相同或相近的语义。该任务在机器翻译、信息检索、问答系统等多个场景中都有广泛应用。BERT（Bidirectional Encoder Representations from Transformers）模型作为一种预训练语言模型，能够捕捉上下文的双向语义信息，在同义句识别等自然语言处理任务中表现优异。本次实验基于 MRPC 数据集，利用 BERT 模型进行同义句预测，旨在深入理解深度学习在自然语言处理任务中的应用流程。

2、数据集来源：

本次实验使用的数据集为 MRPC（Microsoft Research Paraphrase Corpus），该数据集包含 5800 个句子对，每个句子对都带有二元标签，用于指示两个句子是否为同义句。数据集可通过微软官方链接下载：https://www.microsoft.com/en-us/download/details.aspx?id=52398。下载后得到的数据集文件为 msr_paraphrase_train.txt（训练集）和 msr_paraphrase_test.txt（测试集）。

3、实验思路：

环境配置：在本地 Windows 系统中创建 Python 虚拟环境，安装 PyTorch 2.0.0+cpu、Transformers 4.28.1、Pandas 1.5.3、NumPy 1.24.3 等相关库，解决库版本依赖冲突问题。

数据预处理：编写 MRPCDataset 类，实现数据集的读取、异常行处理、多余字段合并、无效数据过滤等功能，确保输入模型的数据格式正确。

模型构建：基于 Transformers 库加载预训练的 bert-base-uncased 模型和对应的分词器，构建包含 BERT 特征提取层和全连接分类层的完整模型。

训练优化：设置合适的训练参数（批次大小、学习率、训练轮次等），使用 AdamW 优化器和 BCELoss 损失函数进行模型训练，实时监控训练过程中的损失和准确率。

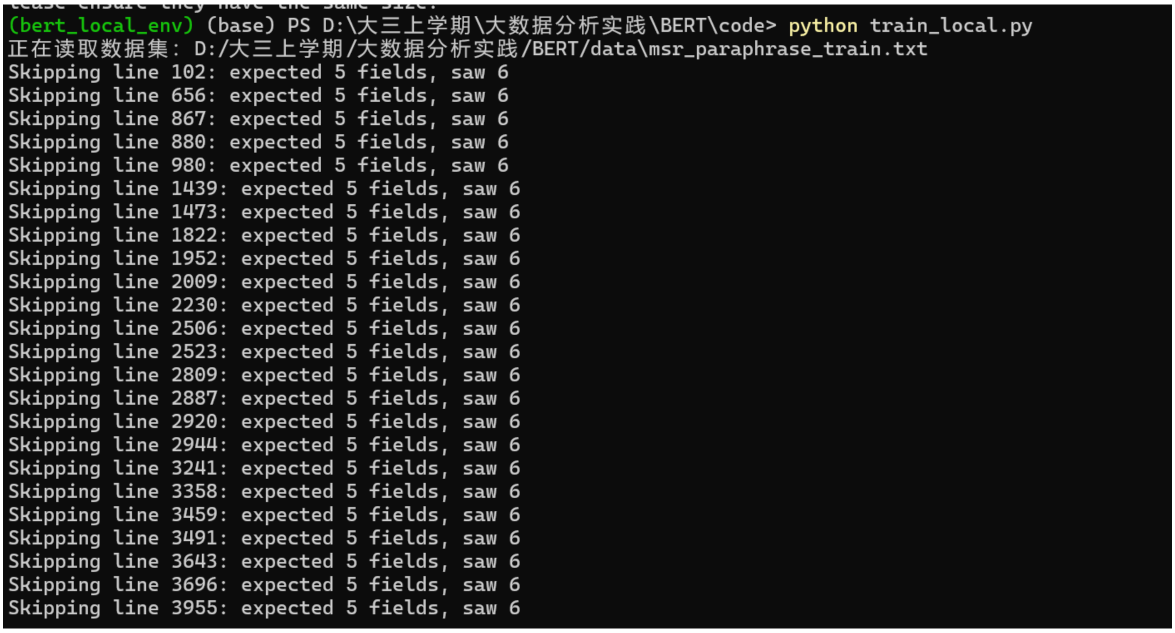
结果保存：训练完成后，保存模型参数，以便后续复用和推理。

4、实验过程：

① 环境配置：首先创建虚拟环境 bert_local_env 并激活，通过指定国内镜像源安装所需库，避免网络问题导致的下载失败。针对 PyTorch 版本兼容问题，选择 2.0.0+cpu 版本，确保与其他库协同工作。安装过程中解决了 NumPy 与 Pandas 的二进制接口不兼容问题，通过指定兼容版本组合（NumPy 1.24.3 + Pandas 1.5.3）确保数据处理模块正常运行。

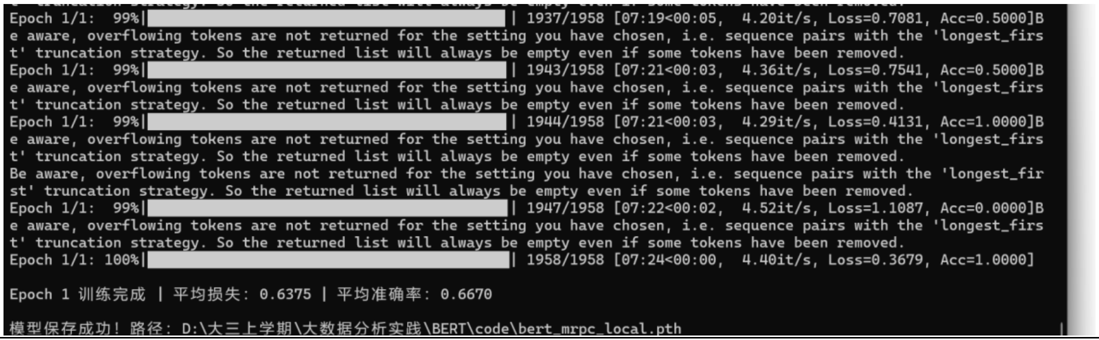
② 数据预处理：编写 MRPCDataset 类处理数据集：首先读取原始数据，跳过表头，对字段数异常的行进行警告处理；合并因句子中包含额外制表符而拆分出的多余字

段，确保每个句子对仅包含 s1 和 s2 两个字符串；过滤空值、无效字符串等数据，截取样本数为批次大小的整数倍，避免训练时出现维度不匹配问题。最终成功加载有效样本。



```
lease ensure they have the same size.
(bert_local_env) (base) PS D:\大三上学期\大数据分析实践\BERT\code> python train_local.py
正在读取数据集: D:/大三上学期/大数据分析实践/BERT/data\msr_paraphrase_train.txt
Skipping line 102: expected 5 fields, saw 6
Skipping line 656: expected 5 fields, saw 6
Skipping line 867: expected 5 fields, saw 6
Skipping line 880: expected 5 fields, saw 6
Skipping line 980: expected 5 fields, saw 6
Skipping line 1439: expected 5 fields, saw 6
Skipping line 1473: expected 5 fields, saw 6
Skipping line 1822: expected 5 fields, saw 6
Skipping line 1952: expected 5 fields, saw 6
Skipping line 2009: expected 5 fields, saw 6
Skipping line 2230: expected 5 fields, saw 6
Skipping line 2506: expected 5 fields, saw 6
Skipping line 2523: expected 5 fields, saw 6
Skipping line 2809: expected 5 fields, saw 6
Skipping line 2887: expected 5 fields, saw 6
Skipping line 2920: expected 5 fields, saw 6
Skipping line 2944: expected 5 fields, saw 6
Skipping line 3241: expected 5 fields, saw 6
Skipping line 3358: expected 5 fields, saw 6
Skipping line 3459: expected 5 fields, saw 6
Skipping line 3491: expected 5 fields, saw 6
Skipping line 3643: expected 5 fields, saw 6
Skipping line 3696: expected 5 fields, saw 6
Skipping line 3955: expected 5 fields, saw 6
```

③ 模型训练：加载预训练模型：通过 Hugging Face 国内镜像源下载 bert-base-uncased 模型和分词器，确保模型文件完整且能正常加载。数据加载：使用 DataLoader 和 BatchSampler 加载数据集，设置批次大小为 2，丢弃最后一个不足批次大小的样本，保证每个批次的样本数一致。训练过程：在 CPU 环境下进行训练，冻结 BERT 模型权重以加快训练速度，仅训练全连接分类层。训练过程中实时打印每个批次的损失和准确率，监控模型训练状态。针对训练初期出现的维度不匹配问题，通过强制裁剪 / 填充预测结果，确保预测值与标签形状一致。



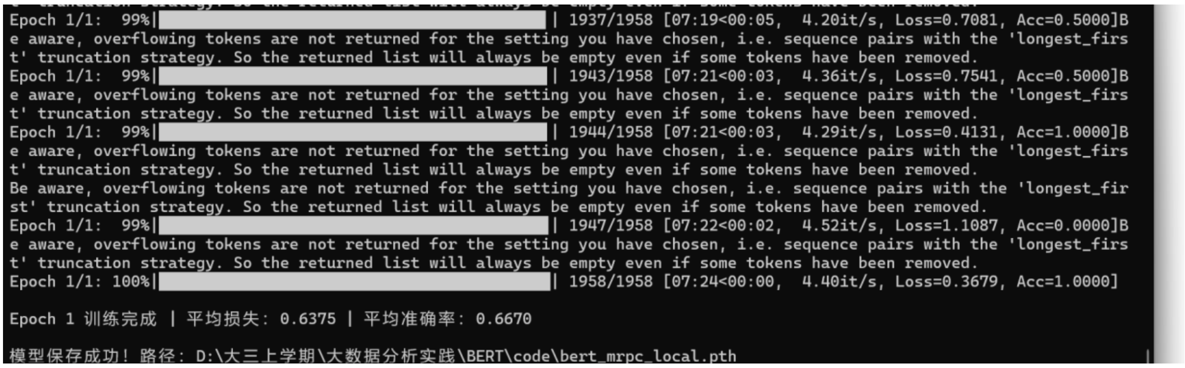
```
Epoch 1/1:  99%|████████████████████████████████████████| 1937/1958 [07:19<00:05,  4.20it/s, Loss=0.7081, Acc=0.5000]B
e aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_firs
t' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Epoch 1/1:  99%|████████████████████████████████████████| 1943/1958 [07:21<00:03,  4.36it/s, Loss=0.7541, Acc=0.5000]B
e aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_firs
t' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Epoch 1/1:  99%|████████████████████████████████████████| 1944/1958 [07:21<00:03,  4.29it/s, Loss=0.4131, Acc=1.0000]B
e aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_firs
t' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Be aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_fir
st' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Epoch 1/1:  99%|████████████████████████████████████████| 1947/1958 [07:22<00:02,  4.52it/s, Loss=1.1087, Acc=0.0000]B
e aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_firs
t' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Epoch 1/1: 100%|████████████████████████████████████████| 1958/1958 [07:24<00:00,  4.40it/s, Loss=0.3679, Acc=1.0000]

Epoch 1 训练完成 | 平均损失: 0.6375 | 平均准确率: 0.6670

模型保存成功! 路径: D:\大三上学期\大数据分析实践\BERT\code\bert_mrpc_local.pth
```

结果图片：

训练集有效样本数经过滤后符合批次大小要求，共 1958 个批次。最终模型平均损失为 0.6375，平均准确率为 0.6670，模型参数成功保存至 bert_mrpc_local.pth 文件中。



```
Epoch 1/1:  99%|████████████████████████████████████████| 1937/1958 [07:19<00:05,  4.20it/s, Loss=0.7081, Acc=0.5000]B
e aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_firs
t' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Epoch 1/1:  99%|████████████████████████████████████████| 1943/1958 [07:21<00:03,  4.36it/s, Loss=0.7541, Acc=0.5000]B
e aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_firs
t' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Epoch 1/1:  99%|████████████████████████████████████████| 1944/1958 [07:21<00:03,  4.29it/s, Loss=0.4131, Acc=1.0000]B
e aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_firs
t' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Be aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_fir
st' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Epoch 1/1:  99%|████████████████████████████████████████| 1947/1958 [07:22<00:02,  4.52it/s, Loss=1.1087, Acc=0.0000]B
e aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_firs
t' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Epoch 1/1: 100%|████████████████████████████████████████| 1958/1958 [07:24<00:00,  4.40it/s, Loss=0.3679, Acc=1.0000]

Epoch 1 训练完成 | 平均损失: 0.6375 | 平均准确率: 0.6670

模型保存成功! 路径: D:\大三上学期\大数据分析实践\BERT\code\bert_mrpc_local.pth
```