山东大学计算机科学与技术学院

大数据分析实验报告

学号: 202300130050

姓名: 王睿

班级: 数据 23

实验题目: bert 实践

实验学时: 2

实验日期: 2025.11.8

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.

实验环境

实验 4 配置的 BERT 服务器环境

操作系统

Linux (AutoDL 容器环境) Python 版本：3.9.11

深度学习框架

PyTorch 2.4.1+cu121 (CUDA 12.1) Transformers 4.46.3

数据处理库

Datasets 3.1.0

实验步骤

1. 配置环境与参数
在 VS CODE 中通过 Remote - SSH 插件连接实验服务器。导入需要的库，定义训练过程中所需的超参数。

2. 定义模型与工具函数创建用于分类的全连接层模型 FCModel，以及用于计算准确率的 binary_accuracy 函数、处理文本数据的 preprocess_function 和在验证集上评估模型的 evaluate 函数。

3.加载数据和模型使用 datasets 库加载 MRPC 数据集，自动分割为训练集和验证集。

随后加载预训练的 BERT 模型和分词器，并将模型移动到可用的计算设备（在服务器上使用 GPU）。

4. 配置训练组件定义优化器（Adam）、损失函数（二元交叉熵损失 BCELoss）。对训练集和验证集进行预处理，创建 DataLoader 以支持批量数据加载。

5. 执行训练与验证循环执行训练过程。在每个 Epoch 中，首先在训练集上进行模型训练，计算并打印每批数据的损失和准确率。训练结束后，立即在验证集上评估模型当前的泛化能力，并输出验证结果。

```
正在载入MRPC数据集...
数据集载入完成，训练集：3668 个样本，验证集：408 个样本
使用设备：cuda
正在加载BERT模型...
Some weights of the model checkpoint at bert-base-uncased were not used when initializing BertModel: ['cls.predictions.transform.LayerNorm.bi
as', 'cls.predictions.transform.LayerNorm.weight', 'cls.predictions.transform.dense.bias', 'cls.predictions.bias', 'cls.predictions.transform
.dense.weight', 'cls.seq_relationship.bias', 'cls.predictions.decoder.weight', 'cls.seq_relationship.weight']
- This IS expected if you are initializing BertModel from the checkpoint of a model trained on another task or with another architecture (e.g
. initializing a BertForSequenceClassification model from a BertForPreTraining model).
- This IS NOT expected if you are initializing BertModel from the checkpoint of a model that you expect to be exactly identical (initializing
 a BertForSequenceClassification model from a BertForSequenceClassification model).
BERT模型加载完成
正在创建全连接层模型...
全连接层模型创建完成
正在预处理训练集...
正在预处理验证集...
数据集预处理完成
开始训练...

===== Epoch 1/3 =====
Batch 10/459 - Loss: 0.6471, Acc: 0.7500
Batch 20/459 - Loss: 0.6702, Acc: 0.6250
Batch 30/459 - Loss: 0.6307, Acc: 0.7500
Batch 40/459 - Loss: 0.6393, Acc: 0.7500
Batch 50/459 - Loss: 0.7845, Acc: 0.3750
Batch 60/459 - Loss: 0.5426, Acc: 0.8750
Batch 70/459 - Loss: 0.6685, Acc: 0.6250
Batch 80/459 - Loss: 0.6139, Acc: 0.6250
Batch 90/459 - Loss: 0.5774, Acc: 0.7500
Batch 100/459 - Loss: 0.5376, Acc: 0.7500
Batch 110/459 - Loss: 0.4992, Acc: 0.8750
```

结论分析：

模型训练效果显著，训练集准确率从 70.69%升至 92.86%，验证集达 85.78%，证明 BERT 微调能有效捕捉语义匹配特征。第 3 轮验证集损失略升，因数据集小、模型复杂，出现轻微过拟合，整体性能符合任务预期。

体会：

通过本次实验，实现了利用机器学习方法（BERT 微调）分析 MRPC 文本数据的目标，深入理解了数据预处理、模型训练与评估的全流程。同时，掌握了通过 VS CODE 连接服务器，实现在远程环境下调试代码。