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

姓名：吴宇轩、李江涛、陆云、张珈恺、崔嘉铭	班级：数据 23
实验题目：机器学习实践	
实验学时：2	实验日期：2025.10.31

实验目的：

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：

远程带 GPU 的服务器，且 CUDA 版本大于 10.0

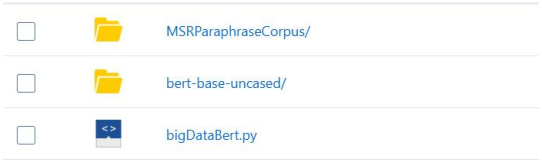
torch==1.7.0

transformers==4.18.0

实验步骤与内容：

（1）配置环境

将数据集、bert 模型和代码挂载到服务器上。



使用阿里云中的 Deep Learning Containers（DLC）产品进行模型的云上训练。

（2）PyTorch 框架下，利用预训练 BERT 模型对 MRPC 数据集进行同义预测

```
torch.cuda.memory_allocated():  1787616256
batch 248 loss:0.625683 accuracy:0.687500
torch.cuda.memory_allocated():  1799613952
batch 249 loss:0.696826 accuracy:0.562500
torch.cuda.memory_allocated():  1785527296
batch 250 loss:0.726771 accuracy:0.500000
torch.cuda.memory_allocated():  1790032896
batch 251 loss:0.692947 accuracy:0.562500
torch.cuda.memory_allocated():  1790400000
batch 252 loss:0.632694 accuracy:0.687500
torch.cuda.memory_allocated():  1788265472
batch 253 loss:0.717117 accuracy:0.500000
torch.cuda.memory_allocated():  1789401600
batch 254 loss:0.694246 accuracy:0.545455
EPOCH 1 loss:0.643563 accuracy:0.667975
```

（3）数据集

MRPC (Microsoft Research Paraphrase Corpus) 包含了 5800 个句子对，有的是同义的，有的是不同义的，是否同义由一个二元标签进行描述。

（4）代码逻辑

对 BERT 进行微调，每个句子对用 BERT 指定分隔符 [SEP] 连接后，通过 BERT 得到合成句子的 representation。再通过一个两层的多层感知机得到分类结果。这里预训练 BERT 模型使用的是 HuggingFace 的 bert-base-uncased。

结论分析与体会：

在本次实验中掌握了如何在远程环境中进行代码的调试和运行，初步了解如何利用 bert 模型处理自然语言任务。