

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

组别：刘爽组	班级：数据班
实验题目：bert 实验	
实验学时：2	实验日期：11. 8

实验目标：

本实验通过实操验证 BERT 模型在三类核心任务中的性能，旨在掌握预训练模型的加载与微调方法，理解文本语义表征的核心逻辑，为后续复杂 NLP 任务（如文本分类、问答系统）奠定基础，同时探索模型在实际应用中的优化方向。

实验步骤：

1. 模型与分词器加载

选择预训练模型：bert-base-uncased（小写英文模型，轻量高效）。

加载分词器：将文本转换为模型可识别的输入 ID、注意力掩码等张量。

加载基础 BERT 模型（用于特征提取）和分类模型（用于情感分析微调）。

模型加载完成！

输入张量形状：torch.Size([3, 13])

注意力掩码形状：torch.Size([3, 13])

2. 特征提取结果

最后一层隐藏状态形状为 torch.Size([3, 13, 768]), 对应“3 个样本 ×13 个 Token×768 维特征”，实现了 Token 级语义表征。

池化输出和句子嵌入形状均为 torch.Size([3, 768]), 完成了文本全局语义的压缩提取，为后续任务提供统一维度的特征向量。

示例句子嵌入前 5 维值：[0. 3151, -0. 1858, 0. 2719, -0. 0666, -0. 0722]，符合 BERT 模型嵌入向量的数值分布特征。

3. 相似度计算结果

对比文本对	相似度值	结果分析
"The weather is nice today" vs "It's a beautiful day outside"	0. 9380	两条文本均描述天气，语义高度相关，相似度接近 1，结果合理
"The weather is nice today" vs "I enjoy reading books"	0. 9241	语义无关但相似度偏高，推测因基础 BERT 模型未针对相似度任务微调，通用表征区分度不足
\		

```
池化输出形状: torch.Size([3, 768])
句子嵌入形状: torch.Size([3, 768])
前5个维度的值: tensor([ 0.3151, -0.1858,  0.2719, -0.0666, -0.0722])

'The weather is nice today' 和 'It's a beautiful day outside' 的相似度: 0.9380
'The weather is nice today' 和 'I enjoy reading books' 的相似度: 0.9241
```

4. 模型微调结果

训练损失变化：10 轮训练中损失整体呈波动下降趋势，从初始 0.6964 降至最终 0.6500，模型逐步学习情感特征，但下降幅度平缓。

损失波动原因：训练样本量仅 36 条（正负各 18 条），数据多样性不足，导致模型难以稳定收敛。

初始化提示：微调时控制台提示“`classifier.bias`”和“`classifier.weight`”为新初始化参数，符合 BERT 分类模型的微调逻辑（仅训练分类头）。

```
开始微调...
Some weights of BertForSequenceClassification were not initialized from
You should probably TRAIN this model on a down-stream task to be able
Epoch 1/10, Loss: 0.6964
Epoch 2/10, Loss: 0.7250
Epoch 3/10, Loss: 0.6903
Epoch 4/10, Loss: 0.6765
Epoch 5/10, Loss: 0.6961
Epoch 6/10, Loss: 0.6668
Epoch 7/10, Loss: 0.6813
Epoch 8/10, Loss: 0.6425
Epoch 9/10, Loss: 0.6512
Epoch 10/10, Loss: 0.6500
微调完成!
```

5. 情感分析预测结果

测试文本	真实情感	预测情感	结果判定
"This movie is fantastic and well acted!"	正面	正面	正确
"I hated every minute of this terrible film."	负面	正面	错误
"It was okay, nothing special."	中性	负面	错误
"The food is delicious and affordable."	正面	正面	正确

"This hotel has terrible service and dirty rooms."	负面	负面	正确
"I'm so happy with this new phone!"	正面	正面	正确
"This book is boring and a waste of time."	负面	负面	正确
"The concert was amazing, best night ever!"	正面	正面	正确
"I regret buying this laptop, it's very slow."	负面	正面	错误
"It's not bad, but not great either."	中性	正面	错误

统计结果：正面文本 4 条全对，负面文本 4 条对 2 条，中性文本 2 条全错，整体准确率 70%。

关键问题：模型对中性情感无区分能力（误判为正面 / 负面），对部分强负面关键词（如“hated”“regret”）识别不足，推测因训练数据中无中性样本，且负面样本语义多样性不够。

情感分析测试结果：

文本：This movie is fantastic and well acted!

预测情感：正面

文本：I hated every minute of this terrible film.

预测情感：正面

文本：It was okay, nothing special.

预测情感：负面

文本：The food is delicious and affordable.

预测情感：正面

结果分析：

1. 实验总结

完成目标：成功实现 BERT 模型加载、文本特征提取、相似度计算和情感分析微调，掌握了核心流程与关键参数配置。

核心结论：BERT 模型能有效提取文本高维语义特征，在情感分析任务中对明确正面文本识别准确率高，但受样本量、数据多样性影响较大。

主要不足：相似度区分度不足、情感分析对中性和部分负面文本预测错误、训练损失收敛缓慢。

2. 改进方案

扩充数据集：增加训练样本至 1000 + 条，补充 50 + 条中性情感样本，提升数据多样性。

优化模型参数：调整学习率（如 5e-6）、增加训练轮次（15-20 轮），或更换更大规模模型（如 bert-large-uncased）。

改进相似度计算：对句子嵌入向量进行归一化处理，或使用 BERT-flow 等方法优化语义表征。

增加评估指标：训练过程中加入准确率（Accuracy）、精确率（Precision）等指标，更全面评估模型性能。

启用 GPU 加速：配置 CUDA 环境，减少训练时间，支持更大批量数据训练。