

# 计算机科学与技术学院大数据分析与实践课程实验报告

实验题目：机器学习方法选择与可视化设计实践		学号：202322130197
日期：	班级：数据	姓名：崔嘉铭
Email：cjm13969665900@gmail.com		

## 实验目的：

本实验旨在通过实际动手操作，加深对机器学习方法选择与应用流程的理解，并结合可视化设计实践，培养对数据分析结果进行表达和解释的能力。

### 具体目标包括：

- 熟悉基于深度学习的自然语言处理任务流程，理解预训练模型在下游任务中的使用方式；
- 掌握利用 PyTorch 与 Transformers 框架进行模型微调的基本方法；
- 通过实际训练任务，理解数据预处理、模型训练与评估的完整 pipeline；
- 通过可视化设计实践，锻炼将数据与问题转化为合适可视化表达形式的能力。

## 实验软件和硬件环境：

### 硬件环境：

- 远程服务器（支持 GPU 加速）
- NVIDIA GPU（CUDA 可用）

### 软件环境：

- 操作系统：Linux
- Python 版本：3.x
- PyTorch：1.7.0
- Transformers：4.18.0
- CUDA：10.0 及以上
- 通过 Conda 创建并管理虚拟环境

## 实验原理和方法：

### 一) 机器学习方法选择原理

本实验选用 BERT (Bidirectional Encoder Representations from Transformers) 作为预训练模型，对 MRPC (Microsoft Research Paraphrase Corpus) 数据集进行同义句判别任务。

MRPC 是一个二分类任务，模型需要判断给定的两个句子在语义上是否表达相同含义。

实验中采用的基本思想为：

- 使用预训练的 BERT 模型对句子对进行编码；
- 利用 [CLS] 标记对应的语义表示 (pooler\_output) 作为句子对的整体表示；
- 通过全连接层输出二分类结果；
- 使用二分类交叉熵损失函数进行训练。

### (二) 可视化设计方法

在可视化设计实践中，通过提出多个候选可视化选题，对不同信息表达方式进行

比较，最终选取最合适的方案进行草图设计。设计重点放在信息编码方式的合理性和可读性上，而非具体实现细节。

实验步骤：(不要求罗列完整源代码)

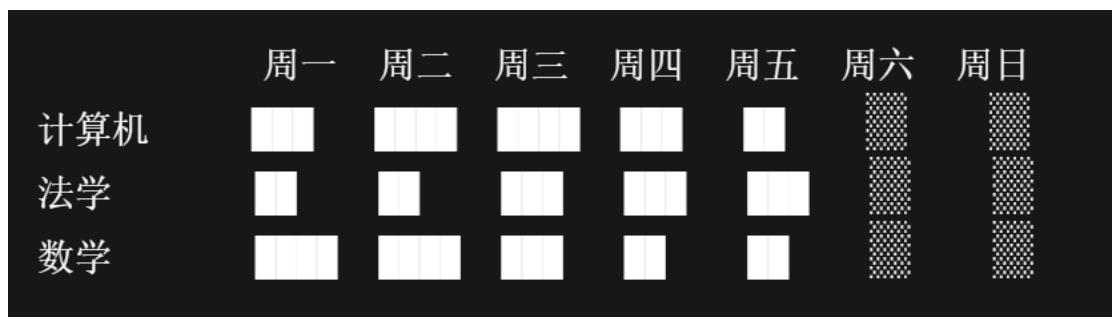
### (一) 机器学习实践部分

1. 在远程服务器上配置 Python 与深度学习运行环境，并安装 PyTorch 与 Transformers 相关依赖；
2. 下载并整理 MRPC 数据集，使用官方提供的 msr\_paraphrase\_train 与 msr\_paraphrase\_test 文件作为训练集与测试集；
3. 使用 tokenizer 对句子对进行编码，并构建 DataLoader 以支持批量训练；
4. 加载 BERT-base-uncased 预训练模型，并构建简单的全连接分类层；
5. 使用 BCEWithLogitsLoss 作为损失函数，对模型进行微调训练；
6. 在训练过程中记录 loss 与 accuracy，并在测试集上进行效果评估。

```
model.safetensors: 43%|  
Epoch 1 Step 0001 | loss=0.9981 | acc=0.1875  
Epoch 1 Step 0020 | loss=0.6931 | acc=0.5813  
Epoch 1 Step 0040 | loss=0.6567 | acc=0.6281  
Epoch 1 Step 0060 | loss=0.6447 | acc=0.6448  
Epoch 1 Step 0080 | loss=0.6321 | acc=0.6555  
Epoch 1 Step 0100 | loss=0.6162 | acc=0.6675  
Epoch 1 Step 0120 | loss=0.6079 | acc=0.6781  
Epoch 1 Step 0140 | loss=0.6007 | acc=0.6835  
Epoch 1 Step 0160 | loss=0.5840 | acc=0.6961  
Epoch 1 Step 0180 | loss=0.5753 | acc=0.7042  
Epoch 1 Step 0200 | loss=0.5646 | acc=0.7122  
Epoch 1 Step 0220 | loss=0.5490 | acc=0.7207  
Epoch 1 Step 0240 | loss=0.5490 | acc=0.7190  
[EPOCH 1] train loss=0.5460 acc=0.7206  
[EPOCH 1] test loss=0.4628 acc=0.7554  
[OK] Training finished.
```

### (二) 可视化设计实践部分

1. 围绕学习、生活和行为数据提出 10 个可视化设计候选题目；
2. 对比不同方案的信息表达效果，选择“不同专业学生一周课程强度对比”作为最终设计主题；
3. 采用热力图作为可视化形式，设计对应的横轴、纵轴及颜色编码方式；
4. 绘制可视化草图，并说明设计逻辑。



### 结论分析与体会：

在 MRPC 数据集上对 BERT 模型进行 1 个 epoch 的微调训练后，模型训练过程稳定，loss 随训练逐步下降，accuracy 持续提升。实验结果表明，模型能够有效学习句子对之间的语义关系。

在测试集上，模型准确率达到约 75%，说明基于预训练模型的方法在小规模数据集上依然具有良好的泛化能力。本实验的重点在于完整跑通训练与评估流程，而非追求最优性能，因此该结果符合实验预期。