

计算机科学与技术学院 大数据分析实践 课程实验报告

实验题目：BERT 环境配置与实验		学号：202100202121
日期：2023. 11. 15	班级：数据 21	姓名：李芷墨
Email：1621737438@qq. com		
<p>实验目的：</p> <p>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。</p>		
<p>实验软件和硬件环境：</p> <p>Linux 服务器，显卡配置如下</p> <div><p>GPU 2080 Ti-11G 数量： 1 显存： 11 GB</p><p>CPU Intel(R) Xeon(R) CPU E5-2683 v4 实例内存： 31G 核心： 6 核</p><p>实例存储 系统盘： 20G 数据盘： 50GB NVME</p><p>网络 上行带宽： 1000 Mbps/s 下行带宽： 1000 Mbps/s</p><p>费用 ￥0.90/小时 不可用代金券</p><p>机器ID MACHpWA4nJILLD9IMTuHGcJN</p><p>最高CUDA版本 11.6</p><p>显卡驱动版本 510.47.03</p></div>		
<p>实验原理和方法：</p> <p>一、BERT 环境配置</p> <p>1.1 服务器环境配置</p> <ol style="list-style-type: none">使用 SSH 连接远程服务器从 Anaconda 的官网下载 for Linux 的安装包，用 FTP 传输至服务器上想要的安装位置，并使用 bash 命令在该位置进行安装使用 conda create 命令创建 base 环境以外的虚拟环境使用 conda activate 命令进行虚拟环境的切换安装所需要的包 <p>1.2 SSH 连接服务器</p> <p>使用 vscode 下载 Remote-SSH，连接远程服务器</p>		

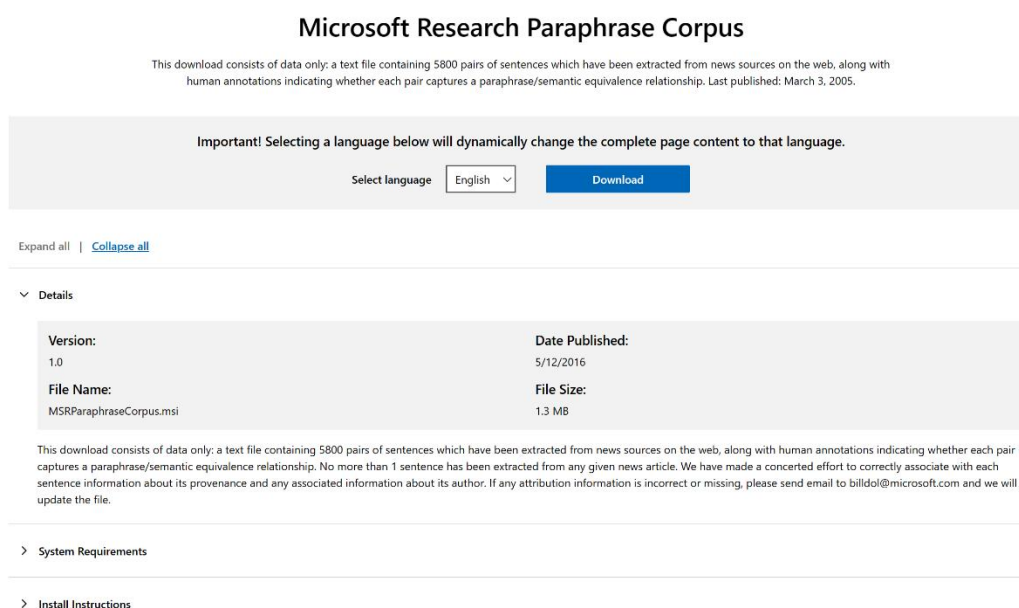
二、实验步骤：（不要求罗列完整源代码）BERT 实践

熟悉 PyTorch 框架下，利用预训练的 transformers 的预训练 BERT 模型对 MRPC 数据集进行同义预测的 pipeline。尝试理解数据是如何预处理，模型是怎么读入数据，是如何进行推理，如何进行评价的。

2.1 数据集

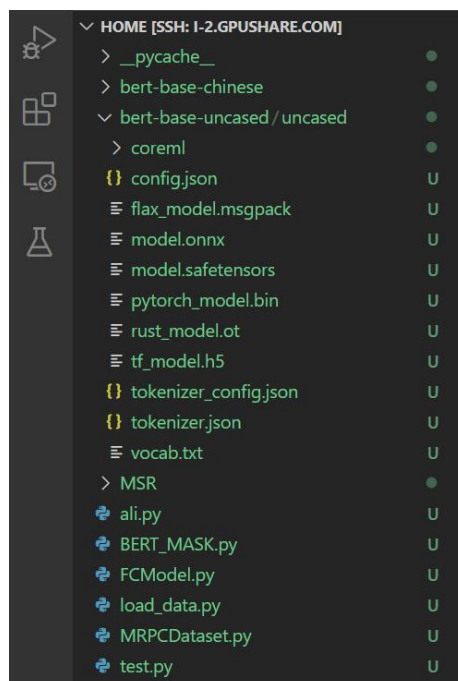
MRPC (Microsoft Research Paraphrase Corpus) 包含了 5800 个句子对，有的是同义的，有的是不同义的，是否同义由一个二元标签进行描述。

下载链接 <https://www.microsoft.com/en-us/download/details.aspx?id=52398>



2.2 模型下载

可以在阿里云的相关镜像上下载模型，及其配置，并上传到服务器上。



2.3 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。

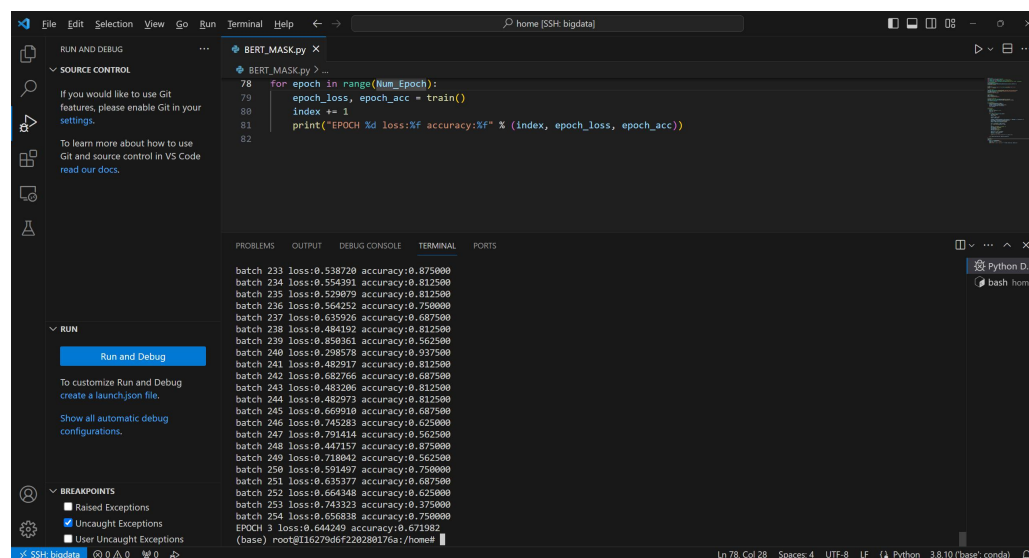
结论分析与体会:

一共进行了三个 epoch,

EPOCH 1 loss:0.650114 accuracy:0.655545

EPOCH 2 loss:0.650712 accuracy:0.658734

EPOCH 3 loss:0.644249 accuracy:0.671982



The screenshot shows a VS Code editor with a Python file named `BERT_MASK.py`. The code is a training loop for BERT. The terminal output shows the results of the training for epochs 1, 2, and 3, along with batch-level loss and accuracy.

```
78 for epoch in range(num_epochs):
79     epoch_loss, epoch_acc = train()
80     index += 1
81     print("EPOCH %d loss:%f accuracy:%f" % (index, epoch_loss, epoch_acc))
82
```

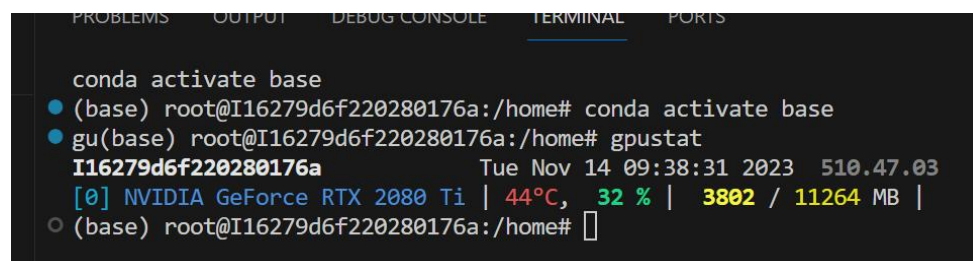
Terminal Output:

```
batch 233 loss:0.538720 accuracy:0.875000
batch 234 loss:0.554391 accuracy:0.812500
batch 235 loss:0.529879 accuracy:0.812500
batch 236 loss:0.564252 accuracy:0.750000
batch 237 loss:0.639206 accuracy:0.687500
batch 238 loss:0.484192 accuracy:0.812500
batch 239 loss:0.858361 accuracy:0.562500
batch 240 loss:0.298578 accuracy:0.937500
batch 241 loss:0.482917 accuracy:0.812500
batch 242 loss:0.682766 accuracy:0.687500
batch 243 loss:0.483286 accuracy:0.812500
batch 244 loss:0.482973 accuracy:0.812500
batch 245 loss:0.669918 accuracy:0.687500
batch 246 loss:0.745283 accuracy:0.625000
batch 247 loss:0.791414 accuracy:0.562500
batch 248 loss:0.447157 accuracy:0.875000
batch 249 loss:0.718842 accuracy:0.562500
batch 250 loss:0.591497 accuracy:0.750000
batch 251 loss:0.635377 accuracy:0.687500
batch 252 loss:0.664188 accuracy:0.625000
batch 253 loss:0.743323 accuracy:0.375000
batch 254 loss:0.656838 accuracy:0.750000
EPOCH 3 loss:0.644249 accuracy:0.671982
(base) root@I16279d6f220280176a:/home#
```

就实验过程中遇到和出现的问题, 你是如何解决和处理的, 自拟 1—3 道问答题:

1、服务器并不能连接 huggingface, 麻烦, 之前的使用我已经自己电脑下载了 bert-chinese, 但这次使用 bert-uncased, 遂让队员在阿里云下载模型。

2、显存占用不高, 可以提升 batch size



The screenshot shows a terminal window with the following commands and output:

```
conda activate base
(base) root@I16279d6f220280176a:/home# conda activate base
gu(base) root@I16279d6f220280176a:/home# gpustat
I16279d6f220280176a Tue Nov 14 09:38:31 2023 510.47.03
[0] NVIDIA GeForce RTX 2080 Ti | 44°C, 32 % | 3802 / 11264 MB |
(base) root@I16279d6f220280176a:/home#
```