山东大学计算机科学与技术学院

大数据分析与实践课程实验报告

学号：202300130205	姓名：李尚远	班级：23 级数据班
实验题目：机器学习实践		
实验学时：2	实验日期：2025.11.7	
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.		

一、dataset 代码

```
#MRPCDataset
class MRPCDataset(Dataset):
    def __init__(self, data_dir="/root/autodl-tmp/data6",split="train"):
        self.data_dir = data_dir
        self.split = split
        if self.split == "train":
            self.file_path=f"{data_dir}/msr_paraphrase_train.txt"
        else:
            self.file_path=f"{data_dir}/msr_paraphrase_test.txt"
        self.data=pd.read_csv(
            self.file_path,
            sep="\t",
            header=0,
            encoding="utf-8",
            on_bad_lines="skip"
            )
    def __len__(self):
        return len(self.data)
    def __getitem__(self, idx):
        item=self.data.iloc[idx]
        sen1=str(item['#1 String'])
        sen2=str(item['#2 String'])
        label=int(item['Quality'])
```

```
        return (sen1,sen2),label
def collate_fn(batch):
    sen1_list=[sample[0][0] for sample in batch]
    sen2_list=[sample[0][1] for sample in batch]

    labels=torch.tensor([sample[1] for sample in batch],dtype=torch.float32)
    return (sen1_list,sen2_list),labels
```

二、全连接层代码

```
#全连接层
class FCModel(nn.Module):
    def __init__(self, input_size=768,
hidden1_size=512,hidden2_size=256,output_size=1):
        super(FCModel, self).__init__()
        self.fc1 = nn.Linear(input_size, hidden1_size)
        self.fc2 = nn.Linear(hidden1_size, hidden2_size)
        self.fc3 = nn.Linear(hidden2_size, output_size)
        self.relu = nn.ReLU()
        self.sigmoid = nn.Sigmoid()
    def forward(self, X):
        out = self.fc1(X)
        out = self.relu(out)
        out = self.fc2(out)
        out = self.relu(out)
        out = self.fc3(out)
        out = self.sigmoid(out)
        return out
```

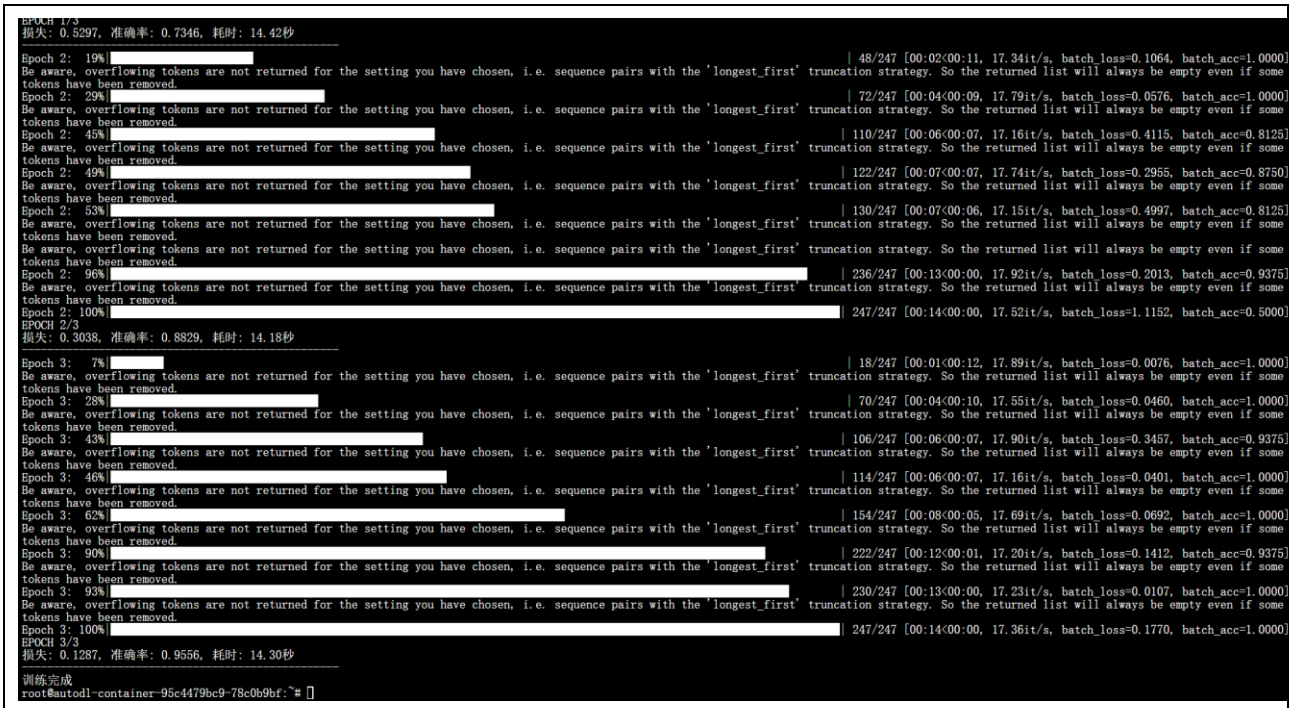
三、数据加载

```
  # 1. 数据加载
    print("正在加载数据集...")
    mrpc_dataset = MRPCDataset()
    train_loader = DataLoader(
        dataset=mrpc_dataset,
        batch_size=16,
        shuffle=True,
        num_workers=4,  # 使用多进程加载数据，加快速度
        collate_fn=collate_fn
    )
```

四、模型加载

```
    # 3. 加载 BERT 模型和分词器
    print("正在加载 BERT 模型...")
    tokenizer = BertTokenizer.from_pretrained("bert-base-uncased")
    bert_model = BertModel.from_pretrained("bert-base-uncased")
    bert_model = bert_model.to(device)
    print("BERT 模型加载完成")
```

五、训练结果


EPOCH 1/3
损失: 0.5297, 准确率: 0.7346, 耗时: 14.42秒
Epoch 2: 19% | 48/247 [00:02<00:11, 17.34it/s, batch_loss=0.1064, batch_acc=1.0000]
Be aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_first' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Epoch 2: 29% | 72/247 [00:04<00:09, 17.79it/s, batch_loss=0.0576, batch_acc=1.0000]
Be aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_first' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Epoch 2: 45% | 110/247 [00:06<00:07, 17.16it/s, batch_loss=0.4115, batch_acc=0.8125]
Be aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_first' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Epoch 2: 49% | 122/247 [00:07<00:07, 17.74it/s, batch_loss=0.2955, batch_acc=0.8750]
Be aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_first' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Epoch 2: 53% | 130/247 [00:07<00:06, 17.15it/s, batch_loss=0.4997, batch_acc=0.8125]
Be aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_first' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Be aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_first' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Epoch 2: 96% | 236/247 [00:13<00:00, 17.92it/s, batch_loss=0.2013, batch_acc=0.9375]
Be aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_first' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Epoch 2: 100% | 247/247 [00:14<00:00, 17.52it/s, batch_loss=1.1152, batch_acc=0.5000]
EPOCH 2/3
损失: 0.3038, 准确率: 0.8829, 耗时: 14.18秒
Epoch 3: 7% | 18/247 [00:01<00:12, 17.89it/s, batch_loss=0.0076, batch_acc=1.0000]
Be aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_first' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Epoch 3: 28% | 70/247 [00:04<00:10, 17.55it/s, batch_loss=0.0460, batch_acc=1.0000]
Be aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_first' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Epoch 3: 43% | 106/247 [00:06<00:07, 17.90it/s, batch_loss=0.3457, batch_acc=0.9375]
Be aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_first' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Epoch 3: 46% | 114/247 [00:06<00:07, 17.16it/s, batch_loss=0.0401, batch_acc=1.0000]
Be aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_first' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Epoch 3: 62% | 154/247 [00:08<00:05, 17.69it/s, batch_loss=0.0692, batch_acc=1.0000]
Be aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_first' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Epoch 3: 90% | 222/247 [00:12<00:01, 17.20it/s, batch_loss=0.1412, batch_acc=0.9375]
Be aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_first' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Epoch 3: 93% | 230/247 [00:13<00:00, 17.23it/s, batch_loss=0.0107, batch_acc=1.0000]
Be aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_first' truncation strategy. So the returned list will always be empty even if some tokens have been removed.
Epoch 3: 100% | 247/247 [00:14<00:00, 17.36it/s, batch_loss=0.1770, batch_acc=1.0000]
EPOCH 3/3
损失: 0.1287, 准确率: 0.9556, 耗时: 14.30秒
训练完成
root@autodl-container-95c4479bc9-78c0b9bf:~#