作业六. PyTorch下的神经网络训练用于AGAC的实体识别（BERT+CRF）

生信1802-余思克-2018317220208

实验目的

PyTorch一个开放源代码的机器学习框架，可加快从研究原型到生产部署的过程。本实验将采用BERT+CRF模型，使用python中的PyTorch框架搭建的神经网络进行模型的训练从而实现AGAC语料库的实体识别。

实验步骤

1. PyTorch框架的的安装

PyTorch安装指引：https://pytorch.org/get-started/locally/

根据自己电脑的系统、语言环境、CUDA版本等属性，选择合适的PyTorch版本进行安装。

```
pip3 install torch==1.8.1+cu111 torchvision==0.9.1+cu111 torchaudio===0.8.1 -f
https://download.pytorch.org/whl/torch_stable.html # 安装Windows版本的torch 1.8
```

通过进入python环境，加载`torch`框架，使用简单的torch方法验证安装是否成功。

```
import torch
x = torch.rand(5, 3)
print(x) # 验证torch是否安装成功
# tensor([[0.8168, 0.6275, 0.7394],
#         [0.4408, 0.5465, 0.7166],
#         [0.0079, 0.2844, 0.5777],
#         [0.9364, 0.2016, 0.0551],
#         [0.9504, 0.6274, 0.3595]])
```

2. 实验数据及代码的获取

通过`git clone`命令从GitHub上获取本次实验所需的AGAC数据以及相关代码文件。

```
git clone https://github.com/bionlp-hzau/BERT-CRF-for-BioNLP-OST2019-AGAC-
Task1.git # 获取实验数据及代码文件
```

3. 相关python依赖包的安装

为了完成本实验的分析任务，除了PyTorch框架之外，还需要安装一系列python包（`numpy`、`transformers`、`TorchCRF`），它们的名称以及版本罗列在了`requirements.txt`文件中，使用`pip install`命令即可从该文件中提供的信息安装相应的依赖包。

```
pip3 install -r requirements.txt # 安装python依赖包
```

4. 模型的训练与评估

安装好所需的框架与包之后，运行主函数即可进行模型的训练与评估。其中，`label.txt`包含数据集中涉及的所有标签，以及与[CLS]，[SEP]和[Padding]对应的标签；`train_input.txt`，`test_input.txt`，`train_input.txt`，`test_input.txt`文件包含BIO格式的训练数据和测试数据。模型评估使用的是`Conlleval.pl`脚本。由于笔记本算力不足，耗时较长，固修改了模型训练的参数，将epoch修改为20。

```
self.num_train_epochs = 5 # 对config.py第44行的epoch参数修改为10
```

```
python3 main.py # 模型的训练与评估
```

实验结果

1. 第一轮训练结果

第一轮训练之后，正确预测的标签数目为184个，准确度为93.45%，非-O标注下的准确度为8.47%，FB1值为19.32%。

```
  processed 43128 tokens with 1610 phrases; found: 295 phrases; correct: 184.
  accuracy:   8.47%; (non-O)
  accuracy:  93.45%; precision:  62.37%; recall:  11.43%; FB1:  19.32%
                CPA: precision:   0.00%; recall:   0.00%; FB1:   0.00  0
            Disease: precision:   0.00%; recall:   0.00%; FB1:   0.00  0
            Enzyme: precision:   0.00%; recall:   0.00%; FB1:   0.00  0
              Gene: precision: 100.00%; recall:   0.32%; FB1:   0.64  1
        Interaction: precision:   0.00%; recall:   0.00%; FB1:   0.00  0
                MPA: precision:  28.57%; recall:   2.90%; FB1:   5.26  14
             NegReg: precision:  74.00%; recall:  38.14%; FB1:  50.34  50
            Pathway: precision:   0.00%; recall:   0.00%; FB1:   0.00  0
             PosReg: precision:  25.00%; recall:   3.92%; FB1:   6.78  8
            Protein: precision:   0.00%; recall:   0.00%; FB1:   0.00  0
                Reg: precision:   0.00%; recall:   0.00%; FB1:   0.00  0
                Var: precision:  63.06%; recall:  35.99%; FB1:  45.83  222
Epoch: 0, Epoch-Average Loss: 42.508985713470814
ACC_non_O: 8.4661, ACC_inc_O: 93.4544, Precision: 62.3729, Recall: 11.4286, F1-score: 19.3176
Updated model, best F1-score: 19.3176
```

2. 第五轮训练结果

第五轮训练之后，正确预测的标签数目为924个，准确度为95.49%，非-O标注下的准确度为63.71%，FBI值为53.23%。

```
  processed 43128 tokens with 1610 phrases; found: 1862 phrases; correct: 924.
  accuracy:  63.71%; (non-O)
  accuracy:  95.49%; precision:  49.62%; recall:  57.39%; FB1:  53.23%
                CPA: precision:  11.11%; recall:  16.67%; FB1:  13.33  63
            Disease: precision:  46.76%; recall:  63.86%; FB1:  53.99  340
            Enzyme: precision:   0.00%; recall:   0.00%; FB1:   0.00  0
              Gene: precision:  57.37%; recall:  68.37%; FB1:  62.39  373
        Interaction: precision:  25.00%; recall:  40.00%; FB1:  30.77  8
                MPA: precision:  22.94%; recall:  28.26%; FB1:  25.32  170
             NegReg: precision:  74.44%; recall:  69.07%; FB1:  71.66  90
            Pathway: precision:  28.57%; recall:  33.33%; FB1:  30.77  7
             PosReg: precision:  72.22%; recall:  76.47%; FB1:  74.29  54
            Protein: precision:   6.67%; recall:   2.50%; FB1:   3.64  15
                Reg: precision:  64.17%; recall:  57.68%; FB1:  60.75  240
                Var: precision:  47.81%; recall:  61.70%; FB1:  53.87  502
Epoch: 4, Epoch-Average Loss: 6.89190846127133
ACC_non_O: 63.7118, ACC_inc_O: 95.4948, Precision: 49.6241, Recall: 57.3913, F1-score: 53.2258
Updated model, best F1-score: 53.2258
```

讨论

通过不同轮数的训练数据的对比可以发现，神经网络实现的BERT+CRF模型能够快速迭代优化并显著提升模型预测准确度，仅仅在五轮训练之后，便达到了95.49，且非-O准确度也提升至63.71%，该模型的预测效果已经显著高于Wpiti工具的预测能力了，说明神经网络的出现对曾经的序列标注工具提出了很大的挑战。但是，神经网络也有一定的缺点，其进行模型训练时非常耗时，且需要大量的计算资源，训练出更优质的模型，获取更好的预测效果意味着付出大量的时间与算力成本，综合来说，神经网络是打破传统的一种新颖且优质的实体识别方法，可以广泛应用于各种领域的学习问题之中。

本次实验在进行模型训练时，可能会对model问题进行报错，原因是文件夹中没有相应的model文件。解决方法是通过网站：https://huggingface.co/dmis-lab/biobert-base-cased-v1.1/tree/main下载model文件（`congfig.json`、`vocab.txt`、`pytorch_model.bin`三个文件），并把它们放置在一个文件夹中，然后在`config.py`脚本中修改`self.model_name`参数，将其指定为model文件所在的文件夹地址，即可解决该问题。