* 将<https://github.com/Winnie-Qi/Continuous-Bag-of-Words-CBOW/blob/main/Continuous_Bag_of_Words_CBOW.py> 部署到本地，数据frankenstein\_with\_splits.csv在学习通-数据。
* 使用Pycharm的Debug断点调试功能
* Pycharm Debug断点调试（文字）：<https://mbd.baidu.com/newspage/data/dtlandingsuper?nid=dt_3789150325423933107&sourceFrom=search_a>
* Pycharm Debug断点调试（视频）：

<https://mbd.baidu.com/newspage/data/videolanding?nid=sv_6336502760740205272&sourceFrom=qmj>

* 理解代码的关键类、方法、训练流程，回答下方25个问题，并将答案写入README中，与文档07的word2vec\_gensim.ipynb一同上传至GitHub。代码中各模块将在下次作业中复用，请充分理解。文档07、08为作业四全部内容。

### Vocabulary类

1、在Vocabulary类中，mask\_token对应的索引通过调用add\_token方法赋值给self.\_\_ mask\_index\_\_属性。

2、lookup\_token方法中，如果self.unk\_index >=0，则对未登录词返回\_\_unk\_index\_\_。

3、调用add\_many方法添加多个token时，实际是通过循环调用\_\_add\_token\_\_方法实现的。

### CBOWVectorizer类

4、vectorize方法中，当vector\_length < 0时，最终向量长度等于\_\_indices\_\_的长度。

5、from\_dataframe方法构建词表时，会遍历DataFrame中\_\_context\_\_和\_\_target\_\_两列的内容。

6、out\_vector[len(indices):]的部分填充为self.cbow\_vocab.\_\_ mask\_index\_\_。

### CBOWDataset类

7、\_max\_seq\_length通过计算所有context列的\_\_context\_\_的最大值得出。

8、set\_split方法通过self.\_lookup\_dict选择对应的\_\_索引列表\_\_和\_\_目标列表\_\_。

9、\_\_getitem\_\_返回的字典中，y\_target通过查找\_\_target\_\_列的token得到。

### 模型结构

10、CBOWClassifier的forward中，x\_embedded\_sum的计算方式是embedding(x\_in).\_\_ sum\_\_(dim=1)。

11、模型输出层fc1的out\_features等于\_\_num\_embeddings\_\_参数的值。

### 训练流程

12、generate\_batches函数通过PyTorch的\_\_DataLoader\_\_类实现批量加载。

13、训练时classifier.train()的作用是启用\_\_Dropou\_\_和\_\_Batch Normalization\_\_模式。

14、反向传播前必须执行\_\_optimizer\_\_.zero\_grad()清空梯度。

15、compute\_accuracy中y\_pred\_indices通过\_\_y\_pred.argmax(dim=1)\_\_方法获取预测类别。

### 训练状态管理

16、make\_train\_state中early\_stopping\_best\_val初始化为\_\_float('inf')\_\_。

17、update\_train\_state在连续\_\_args.patience\_\_次验证损失未下降时会触发早停。

18、当验证损失下降时，early\_stopping\_step会被重置为\_\_0\_\_。

### 设备与随机种子

19、set\_seed\_everywhere中与CUDA相关的设置是\_\_torch.cuda\_\_.manual\_seed\_all(seed)。

20、args.device的值根据\_\_torch.cuda\_\_.is\_available()确定。

### 推理与测试

21、get\_closest函数中排除计算的目标词本身是通过continue判断word == \_\_target\_word: continue\_\_实现的。

22、测试集评估时一定要调用\_\_classifier.eval()\_\_方法禁用dropout。

### 关键参数

23、CBOWClassifier的padding\_idx参数默认值为\_\_0\_\_。

24、args中控制词向量维度的参数是\_\_args.embedding\_size\_\_。

25、学习率调整策略ReduceLROnPlateau的触发条件是验证损失\_\_增加\_\_（增加/减少）。