NLP HW2 - 基于方面的情感分析

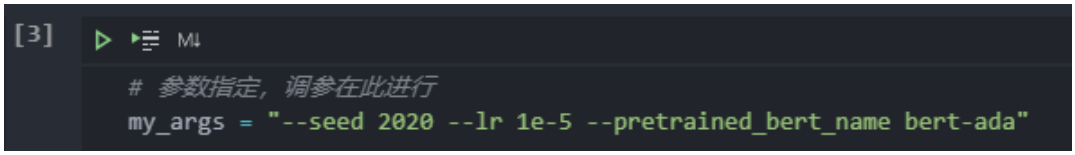
丁豪 南京大学 人工智能学院
181220010@smail.nju.edu.cn

一、实验描述

本次实验为基于**方面**的情感分析，即对某句话中某个方面词、短语进行{-1,0,1}的情感极性判别。在同一个句子中，不同的方面词很有可能对应不同的情感。本次实验的数据集由两部分给出，第一部分训练集由1880条三元组构成，每组数据中第一行为含有方面词的语句，其中方面词由T包围，第二行上述省略的反面词，第三行为方面词在这句话中的极性，总共1880*3=5640行。第二部分测试集，由650条二元组构成，与训练集相比少了第三行极性。实验允许使用外部预训练语言模型，在使用预训练语言模型的情况下满分阈值为accuracy=0.9。

二、运行方法

- 实验依赖文件已经描述在 **requirements.txt** 当中，执行`pip install -r requirements.txt`即可
- 本实验依赖的 pretrained model 下载方式参见 **参考资料** 第一条第三小点
- 本次实验使用 **jupyter notebook** 完成，建议在 **google-colab** 或同等GPU配置环境下运行以保障最佳性能。
  - 项目运行所需的文件有 `hw2.ipynb` 以及 `bert-ada` 文件夹下的全部内容（由于大小要求，bert-ada并未上传，如需运行请前往**参考资料2**的网址下载并放在同一级目录下）
  - 模型参数在第三个cell起始位置指定，可以根据个人需要选择合适的参数进行训练

  - 运行时直接点击 **jupyter notebook** 的 **Run All Cells** 按钮即可完成训练、预测、生成文件的完整过程。最终会在`logs/`中生成训练的过程记录，在`state_dict/`中保存每次训练中产生的最好模型，并在`datasets/181220010.txt`中生成最终预测结果。

三、实现方法

- 学习器被封装为 **Instructor** 类，他内嵌了一个下述的**Network**，并实现了参数指定、网络训练、评估、test_y生成等功能。
  - 参数由上述my_args变量指定，在初始化的过程中，指定神经网络、训练、数据及等超参数。
  - 网络的训练过程使用**adam优化器**优化指定batch_size下每一个batch的**CrossEntrophy损失**，不断训练指导达到最大轮数，或者持续多轮不能生成更好的模型。
  - 网络性能评估，采用accuracy和sklearn的f1_score函数来判断validation set上的正确率与f1，并在产生验证集上性能更好的模型时将其保存。
- 预测模型封装成 **Network** 类。其结构为 bert + dropout + linear，其中bert层初始化为预训练好的`bert-ada`（详见**参考材料**）。
  - 第一层bert层：输入为 text_bert_indices 与 bert_segments_ids，分别为完整句子和方面词的positional embeddings输入。由特殊标记'[CLS]'标记序列的第一个token，并用'[SEP]'将不同的tokens隔开。

    - 第二层dropout：按照指定的dropout率，在每次训练过程中随机失活一定比例的神经元，一次来达到 1：减轻过拟合　2：以类似集成学习的模式增强鲁棒性 的目的。
    - 第三层linear：整合dropout之后的输入，输出为3维，代表方面词在这局话中被判别为{-1,0,1}的相对可能性。
- 其余数据预处理的函数和方法包括
    - **set_random_seed(seed)**：设置全局所有使用到的包的随机种子，以保证复现性
    - **normal_sequence(sequence, maxlen, dtype='int64', value=0)**：将句子处理为标准长度，过长截断，果断则补充value指定的值
    - **MyTokenizer**：扩展了bert模型自带的tokenizer，使其可以和上面的normal_sequence配合，将文本转化为bert字典的id序列，且规范化到指定长度。
    - **MyDataset**：扩展了pytorch的Dataset，实现了从训练集原始文本文件读取数据，进行数据预处理，最终生成可以被pytorch直接调用的Dataset类的功能。

四、实验结果

- 在使用预训练bert-ada模型的情况下，测试集acc：0.9092。
- 训练时间在colab上Tesla T4显卡条件下约为5分钟，本地Geforce 2060情况下约为10分钟，使用何种显卡不影响模型收敛性以及最终acc。
- 如需复现实验结果，请不要修改my_args参数，并运行 `hw2.ipynb`。

五、参考资料

- **bert模型**
    - 通过查阅paperswithcode网站，选取了restruant数据集上泛化性能较好的**bert-ada**

      https://paperswithcode.com/sota/aspect-based-sentiment-analysis-on-semeval
    - 由paperswithcode网站分享的github链接，直接下载了bert-ada官方模型

      https://github.com/deepopinion/domain-adapted-atsc
    - 最终在这个网站上提供的外链下载了与训练的 bert-ada 模型压缩包

      https://drive.google.com/file/d/1DmVrhKQx74p1U5c7oq6qCTVxGIpgvp1c/view

      并将下载下来的文件解压缩后，文件夹命名为 **bert-ada**，放在项目目录下
- 使用pytorch基于bert进行基于方面情感分析

  https://github.com/songyouwei/ABSA-PyTorch
- 此外，还需要感谢 181220031 李惟康 以及 181220056 王宸旭 同学在实验中与我进行的交流讨论，他们对于我理解实验内容以及开展实验有着积极的推动作用。

五、实验总结

　通过本次实验，我对于基于“方面”的文本情感分析有了一定程度的认识。通过了解、使用预训练bert模型，对当前前言的业界语言模型的基本原理和使用方法有了粗浅的了解。