GCAE 实验总结

1. 数据集

- 原 GCAE 代码中使用的数据集为：

```
rest_train = {16: '../SemEval2016Task5/ABSA16_Restaurants_Train_SB1_v2.xml',
              15: '../SemEval2015/ABSA15_Restaurants_Train_Final.xml',
              14: '../SemEval2014Task4/Restaurants_Train_v2.xml'}
rest_test = {16: '../SemEval2016Task5/EN_REST_SB1_TEST.xml.gold',
             15: '../SemEval2015/ABSA15_Restaurants_Test.xml',
             14: '../SemEval2014Task4/Restaurants_Test_Gold.xml'}

laptop_train = {16: '../SemEval2016Task5/ABSA16_Laptops_Train_SB1_v2.xml',
                15: '../SemEval2015/ABSA15_Laptops_Train_Data.xml',
                14: '../SemEval2014Task4/Laptop_Train_v2.xml'}
laptop_test = {16: '../SemEval2016Task5/EN_LAPT_SB1_TEST_.xml.gold',
               15: '../SemEval2015/ABSA15_Laptops_Test.xml',
               14: '../SemEval2014Task4/Laptops_Test_Gold.xml'}
ds_train = {'r': rest_train, 'l': laptop_train}
ds_test = {'r': rest_test, 'l': laptop_test}
ds_yelp = '../data/yelp/review.json'

my_acsa_train = '../acsa-restaurant-large/acsa_train.json'
my_acsa_test = '../acsa-restaurant-large/acsa_test.json'
my_atsa_train = '../atsa-restaurant/atsa_train.json'
my_atsa_test = '../atsa-restaurant/atsa_test.json'
```

选择数据集的文件为：`GCAE\model_files\getsemeval.py`

- 杜老师给出的数据集为：`GCAE\ABSA\acsa-restaurant-large`，其数据类型为 `csv` 类型文件，其中含有三个文件，，依次是：
  - acsa_hard_test
  - acsa_test
  - acsa_train

  这三个文件中含有相同的三个属性：
  - sentence
  - aspect
  - polarity：有 positive、negative、none 三类属性

- 由于原 `GCAE` 实验中使用的是 `json` 类型的文件，所以我对给出的测试 `csv` 文件进行处理，将其转为了 `json` 类型的文件，由于 `json` 类型的文件单个属性是通过双引号包括的以及给定字符的转义问题，所以就存在一些问题：
  - 比如原句子 he says:"he is very powerful"，转成 `json` 类型文件后，该句子变为""sentence" : "he says:"he is very powerful"""，这时候这个句子就有问题了，因为一个属性中只能存在一个双引号对，所

以"he is very powerful" 这句话的双引号就破坏了 `json` 类型，所以我在这里对双引号进行了转义，也即加上了 \ 符号，不知道这是否会对数据集的处理造成破坏，但是这样处理之后数据集可以正常进行使用；
    - 比如原句子中存在中文字符，在转义之后变成未知符号 @$ 诸如此类

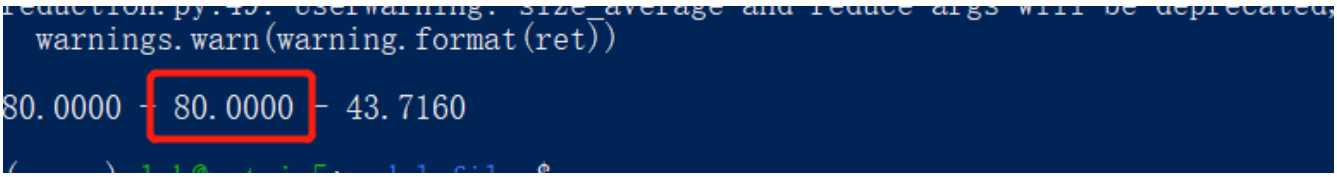
2. 实验结果

- 运行代码的指令为：

进入 `model_files` 文件夹下：

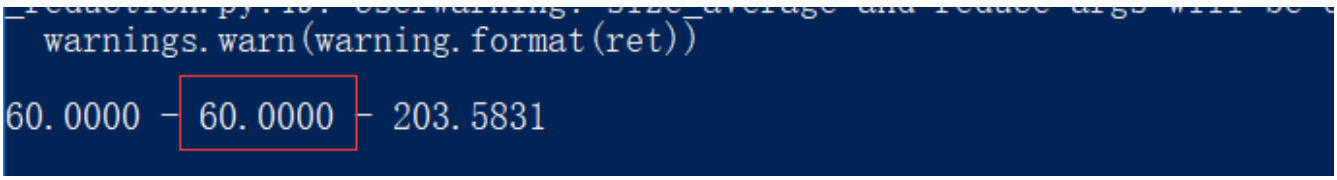
```
python run.py -lr 5e-3 -batch-size 128 -verbose 1 -model CNN_Gate_Aspect -embed_file glove
-r_l r -year 14 -epochs 6  -l2 0 -gpu_id 0
```

- GCAE 原实验代码中使用的数据集，在运行之后的实验结果为：

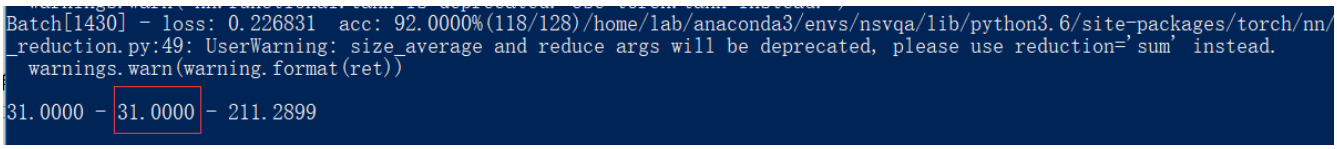


注：从左到右三个数字意义为：校正集准确率、测试集准确率、运行时间

- GCAE 使用 `acsa_train` 作为训练集、`acsa_test` 作为测试集的结果，在运行之后的实验结果为：



- GCAE 使用 `acsa_train` 作为训练集、`acsa_hard_test` 作为测试集的结果，在运行之后的实验结果为：



3. 在服务器上运行代码的流程

服务器上目前 `run.py` 文件使用的数据集为原实验代码使用的数据集，并不是经过转换的 `json` 类型的数据集，所以需要改动数据集的话，在本地对 `GCAE\model_files\getsemeval.py` 文件中红框标注的代码进行修改即可，随后更新到服务器上即可，服务器上已经上传有完整的待测试数据集与完成可运行的代码；

同时服务器上的代码位于 `A305\GCAE` 位置；