2017知乎看山杯ye组代码及复现说明

代码链接：https://mega.nz/#F!0iolCYQI!0SzIr1NEN6OyOGFIBTAj3w

1 运行环境

主要依赖的运行环境信息如表1所示。

表 1. 运行环境

|  |  |  |
| --- | --- | --- |
| **环境/库** | **版本** | |
| Ubuntu  python  jupyter notebook  tensorflow-gpu | | 14.04.5 LTS  2.7.12  4.2.3  1.2.1 |
| numpy  pandas  matplotlib  word2vec  tqdm | | 1.12.1  0.19.2  2.0.0  0.9.1  4.11.2 |

2 文件结构

|- zhihu-ye-6

| |- home1

| | |- data # 预处理好的数据

| | |- models-notebook # 模型的代码

| | | |- run\_all\_in\_home1.sh # 执行本目录中所有模型并保存预测矩阵

| | |-data\_helpers.py # 数据处理函数

| |- home2

| | |- data # 预处理好的数据

| | |- wd-data # 预处理好的数据

| | |- ch-data # 预处理好的数据

| | |- models-notebook # 模型的代码

| | | |- run\_all\_in\_home1.sh # 执行本目录中所有模型并保存预测矩阵

| | |-data\_helpers.py # 数据处理函数

| |- ckpt # 所有训练好的模型

| |- data # 预处理好的数据

| |- data\_process # 数据处理代码

| |- scores # 测试集预测概率矩阵

| |- notebook-old # 比赛中未经过整理的代码

| |- local\_ensemble.ipynb # 验证集模型融合

| |- ensemble.py # 测试集模型融合

| |- data\_helpers.py # 数据处理函数

| |- evaluator.py # 评价函数

在比赛过程中，因为硬盘存储不够所以将数据复制到了另外一个分区，在移动以后，我又重新对数据进行了处理，所以两部分的数据格式会有些差异。data\_process/ 文件夹下面的数据处理代码是后期整理过的代码，但是直接使用这些代码得到的数据对原先的模型输入会出错，因此我把代码分别放在 home1/ 和 home2/ 两个目录下，两个目录结构相同，但是输入数据不同。

zhihu-ye-6/ckpt/ 放置训练保存好的模型

zhihu-ye-6/scores/ 保存生成的测试集的预测概率矩阵

zhihu-ye-6/home/data/ 放置已经处理好的数据

zhihu-ye-6/home/models\_notebook/ 放置保存好的模型代码，执行本目录下的每个 py 文件都会生成一个测试集的预测概率矩阵，并自动保存到 scores/ 目录下。

ensemble.py 对所有模型生成的测试集预测概率矩阵进行线性加权，生成最后的预测结果。

3 复现步骤

步骤一：下载云盘上整个项目代码文件，并将里边所有的压缩包解压到当前文件夹。注意zhihu-ye-6/home1/data/ 文件夹中已经有一个 eval\_segs\_content.py，将解压的结果和这个文件夹合并即可。

步骤二：进入 zhihu-ye-6/home1/models-notebook/ 目录下，输入：

sh run\_all\_in\_home1.sh

运行结束后会在 zhihu-ye-6/scores/ 目录下生成 13 个npy文件。

步骤三：进入 zhihu-ye-6/home2/models-notebook/ 目录下，输入：

sh run\_all\_in\_home2.sh

运行结束后会在 zhihu-ye-6/scores/ 目录下再生成 23 个npy文件，现在zhihu-ye-6/scores/ 目录下应该有36 个 .npy 文件。

步骤四：在 zhihu-ye-6/ 目录下，输入：

python ensemble.py

运行该文件，结束后会在 zhihu-ye-6/ 目录下生成提交结果： ye-final36-result.csv

4 结果说明

由于时间较紧，而且需要处理的文件比较多。在比赛过程中，所有的代码我都是在 jupyter notebook 上面编写的，所以结构比较混乱，最终导出成 .py 文件也有可能出错。在比赛中，有些模型保存出错，其中有一个模型 p2-1-rnn-cnn-256-256 的结果保存出错没法运行，其他的36个模型我都重新跑过一遍，有些模型的最优模型可能删错了，通过练习赛提交的分数要比我最后提交的结果低了1个多千分点。现在我只保留着提交时每个模型生成的概率矩阵。如果有什么问题，请随时联系我：

Email: [yongye@bupt.edu.cn](mailto:yongye@bupt.edu.cn)

Phone: 18811384152