2018 贝贝网·种子杯初赛解题报告

队名：写的代码都

队长：潘翔

队员：林子涵、徐光磊

2018 年 10 月 22 号

1. 使用语言及运行环境

使用语言：python 3.6

运行环境：

Keras 2.2

Tensorflow 1.10

2. 代码接口及运行方式

先运行 seedpro.ipynb 文件进行数据预处理，然后至今运行任意一个.py 文件即可，如：python rcnn.py，本队共用到了四个模型，分别对应四个 py 文件：AttentionGru.py、rcnn.py、textcnn.py、TextInception.py，以及两个用于模型融合的文件：getPro.py、MergePro.py，依次运行以上 6 个文件后运行 MakeUploadResult.py 文件完成三层预测值的整合并输出结果的 txt 文件。其中对于三层的预测是通过模型内部的不同参数体现的，这部分在参数选择部分详细讲解。

3. 数据特征提取思路

因为一条数据包含已经分词后的标题与描述部分，考虑到描述较长不易于训练，且实际情况中标题大多数情况下已经包含了有用的分类信息，所以我们舍弃了部分描述，只选择标题和一些描述来进行训练，又因为数据分为字和词两个部分，这里我们选择的词作为训练内容，后续考虑对字也进行训练以进行模型融合。

4. 预测模型选取

第一层：TextCNN 三路版本

第二层：rcnn 并行、attention+双向 gru、TextCNN 三路版本

第三层：rcnn 并行、attention+双向 gru、TextInception、TextCNN 三路版本

5. 模型参数的选择和优化思路

第一层模型：embedding 层词典大小设置为总词数：353717，

embedding 结果维度设置为 256
dropout=0.5
输出层：softmax 维度：10
loss= crossentropy
optimizer= adam
metrics= f1
batchsize = 64

第二层模型：

Textcnn： 在第一层模型的前提下只修改输出层维度为 64

Rcnn：embedding 层词典大小设置为总词数：353717
embedding 结果维度设置为 120
输出层：softmax 维度：64
dropout=0.2
Gru 单元：256
loss= crossentropy
optimizer= adam
metrics= f1
batchsize = 32

Attention+gru：
embedding 层词典大小设置为总词数：353717
embedding 结果维度设置为 120
输出层：softmax 维度：64
Gru 单元数量：100
loss= crossentropy
optimizer= adam
metrics= f1
batchsize = 32

第三层模型：

Textcnn： 在第二层模型的前提下只修改输出层维度为 125

rcnn： 在第二层模型的前提下只修改输出层维度为 125

Attention+gru： 在第二层模型的前提下只修改输出层维度为 125

TextInception：embedding 层词典大小设置为总词数：353717
embedding 结果维度设置为 120
输出层：softmax 维度：125
loss= crossentropy
optimizer= adam
metrics= f1
batchsize = 32

模型融合方法：

对于二三层，都需要对多个模型的预测结果进行融合，这里我们直接将多个模型的预测概率向量直接相加，再取最大值下标，得到融合后的预测结果。

优化思路：对于第一层预测，textcnn 网络最后表达能力不够，所以我们添加了一个 fc 层，降低了 dropout 来提升结果。

对于第二三层，单独应用 textcnn 结果还不够，所以我们又考虑了时序的信息，将 attention 和 rcnn 模型与 textcnn 相融合，以提高泛化能力，取得好的预测结果。