螺旋桨 RNA 结构预测竞赛第六名解决方案简介

队名：逍遥郎 1392
队员：谢自强

任务：对于给定的 RNA 碱基序列，要求构建模型预测 RNA 碱基不成对的概率。
工程环境：Python3.7+**PaddlePaddle2.0.2**，fork 官方基线系统；
Baseline 方案：lstm

1. 数据分析

初赛数据集包含训练集 4750 条，验证集 250 条，测试集 444 条，采用官方给定的数据进行训练，不再单独划分；

2. 方案分析

2.1 baseline

直接利用 baseline 自带的模型生成结果提交，验证了整个测试、结果生成和提交流程；
模型位置：work/model-0/model_dev=0.0772

2.2 训练 baseline 网络

直接无修改训练 baseline 网络，生成结果并提交；
模型位置：work/model/model_bl_self=0.0676

2.3 增大网络训练

将 baseline 网络增大一倍，layers 参数由 8 设置为 16，dmodel 参数由 128 修改设置为 256；
模型位置：work/model_x2/model_x2_dev=0.0660

2.4 增加训练 epoch

将训练 epoch，由 10 增加为 20，训练 baseline 网络，效果不理想，模型最终未保存；

2.5 修改激活函数

将网络自带的 Relu 激活函数修改为 swish 激活函数，该工程为 work2；
修改代码：work2/src/network.py
模型位置：work2/model/model_dev=0.0700

2.6 全数据训练

充足的训练数据才是模型效果的保障，因此最后考虑采用训练加验证合并训练的方法，同时为了使新增数据得到充分训练，将其扩充为 2 倍加入到训练集中，训练 baseline 网络；
数据文件：work/data/train_dev2.txt
模型位置：work/model/model_add_dev=0.0624

3. 模型融合

在上述方案中，如原 baseline 网络训练多次，很多效果不是很理想，仅保留了其中最好的结果的模型。然后采用多个模型结果融合的方法进一步提升指标。

模型融合采用 4 个模型，分别是：

1. baseline 模型（2.1）
2. baseline 训练模型（2.2）

3. 修改激活函数的模型（2.5）

4. 全数据训练模型（3.6）

融合方式采用线性加权的方式，为 4 个模型生成的概率值分别赋给不同的权重然后生成融合后的结果；

初赛中，通过分配不同的比例的权重进行多次的对比与测试，最终采用的融合比例为 0.8/0.1/0.05/0.05，初赛为第 20 名；

最终的结果：postprocess/predict.files.zip

复赛中，由于不同模型对复赛的测试集的效果与初赛有很大的不同，因此在分别测试了单个模型的结果后，需要对比例进行调整和验证，最终的比例为 0.45/0.01/0/0.45，复赛为第 6 名；

最终的结果：postprocess_2/predict.files.zip

4. 结果复现

环境准备：在 aistudio 上，fork 官方基线系统 https://aistudio.baidu.com/aistudio/projectdetail/1444108

删除或重命名自带的 work 文件夹，然后上传拷贝本方案代码至 work 文件夹位置；

4.1 利用模型生成可提交结果

进入 work 文件夹，执行：

python src/main.py test --model-path-base model-0/model_dev=0.0772

执行完成后会在当前目录生成 test_log.txt，然后在当前目录执行：

python create_result.py

生成 predict.files 文件夹即为可提交的结果，并重命名为 predict.files_bl；

按照上述方法，执行：

python src/main.py test --model-path-base model/ model_bl_self=0.0676

python create_result.py

生成 predict.files 并重命名为 predict.files_bl_self；

执行：

python src/main.py test --model-path-base model/ model_add_dev=0.0624

python create_result.py

生成 predict.files 并重命名为 predict.files_bl_add_dev；

然后进入 work2 文件夹，执行：

python src/main.py test --model-path-base model/ model_dev=0.0700

python create_result.py

生成 predict.files 并重命名为 predict.files_bl_swish；

4.2 生成模型融合结果

将 4.1 中生成的 4 个重命名的以 predict.files 开头的文件夹拷贝到 postprocess_2 文件夹，并执行：

python postprocess.py

完成后会生成 predict.files 文件夹，即是最终的结果；

最后执行：

zip –r predict.files.zip predict.files

生成最终提交的压缩文件 predict.files.zip。

说明：

1. 以上步骤为生成复赛结果的步骤；

2. 如果需要生成初赛的结果，需要修改 work/src/dataset.py 中 line40 和 line41 的测试集路径，将 B_board_112_seqs.txt 修改为 test_nolabel.txt， work2 同理；

3. 按照 4.1 步骤生成 4 个文件夹，然后拷贝到 postprocess 文件夹中，修改 postprocess.py 中的文件名，然后执行该文件即可；