汇报 0608

朱睿涵　李宸亦

项目三

2022 年 6 月 8 日

目录

上周工作

1. 将模型中单一 score 指标增加到 AVERAGE,F1 score 和 EM。主要根据 F1score 进行评估
2. 尝试将观点类问题和 no answer 问题加入模型中，但是效果较差。

{"AVERAGE": "10.601", "F1": "17.878", "EM": "3.324", "TOTAL": 3219, "SKIP": 0, "FILE": "output/dev_predictions.json"}

```
"question" : ["抽烟有益于健康"]
"context" : ["抽烟对于人体健康有着极大危害"]

'answer' : '危害'
```

预计工作

- 确保基础事实性阅读理解的效果，并进行封装，将主要注意力集中在封装 API 上。
- 继续尝试提高观点类问题回答的效果。

目录

问题

在实际运行新的多任务模型时发现效果非常差（F1 仅为 0.07 左右）
暂时没有确定出现 bug 的原因，目前猜测的问题：

- loss：目前用的是 $L = \alpha L_g + (1-\alpha)L_c$ 的总 loss，$\alpha = 0.5$，有可能这两个损失函数减小的速度差距悬殊，答案生成主任务反而被完全掩盖了。
- eval：计算 F1 score 的时候只评判预测 answer 和测试集 answer，实际上并没有将 yes no 观点纳入考虑。
- 问题分类任务是否真的对答案生成任务有帮助？

预计工作

目前打算：

1. debug，对多任务的两个任务加入自适应的学习率，让新的多任务更加均衡，至少不能低于 baseline
2. eval 依旧打算采用 F1 score
3. 如果新模型跑成功，包装 API