目标：

1.小论文撰写

2.课外阅读：每天半小时

3.cs224n

4.英语

5.leetcode

计算机基础课程学习：网络编程、软件工程、编译原理 shell编程，

犯困：leetcode

收尾：阅读，琐事，课外书

茫然：cs224n

英语学习：口语纠正45min 阅读30min

Leetcode

1，看懂题目

2，分析，推导解法

3，将思路转换为代码。

从某个类型开始 25min 写不出来直接题解 ＋笔记记录

9~10

完成开题报告撰写，掌握研究方向情况，由2~3篇相关论文，创新点有思路。有进行简单的实验。

8.30~9.3：截至周六

紧急重要：1.找几篇nlp相关安全论文 完成阅读

紧急不重要：整理下周计划，项目内容填充

重要不紧急:英语，运动，

不重要不紧急：课外阅读，leetcode 题

9.6~9.10：截至周六

紧急重要：1.related work中相关论文 5篇，方向包括威胁情报来源信任分析 和非结构化识别

紧急不重要：整理下周计划，做表单，做表格，画ppt

重要不紧急：英语，运动

不重要不紧急：课外阅读，leetcode

9.13~9.18

紧急重要：1.按标准模板书写开题报告格式，开始整理内容 2.看完几篇综述

紧急不重要：整理下周计划，核酸检测，补牙

重要不紧急：英语，运动

不重要不紧急：课外阅读，leetcode

9.22~9.26

紧急重要：1.顶会三年内相关遍历 2.确定论文主要内容 数据集有无 相关性

紧急不重要：整理下周计划

不重要不紧急：课外阅读，leetcode

9.29~10.10

紧急重要：1.针对社交媒体的有毒内容检测 和恶意软件分析 进行评估。

紧急不重要：整理下周计划

不重要不紧急：课外阅读，leetcode，深水证

10.11~10.17

紧急重要：1.社交媒体检测 2.专利转论文

重要不紧急：1.运动 2.口语

紧急不重要：整理下周计划

不重要不紧急：课外阅读，leetcode

10.11：洗衣服、买衣服，内存卡，记账。

运动：了解一下饮食，健身房计划，姿势纠正资料。

课外阅读：心流

论文：委婉语检测 两部分1.委婉语检测 2.委婉语鉴别

10.12：买外套（牛仔外套，普通外套），记账、羽绒服（双十一）、内衣当垫

外套：levis牛仔外套（想凑单）

论文：已记录

10.13：普通外套，记账、羽绒服（双十一）、写微博

10.14 预约牙医，论文转换完毕，健身房，信息会议，

10.15 自由泳教程 找完related work

委婉语含义：没有明确研究的 只有词义发现【60】【61】 词义消歧【62】~【65】

地下论坛黑话识别：

已记录

10.17 自由泳 论文 表达和求助

地下论坛黑话检测

已记录

情绪分析来寻找委婉语

已记录

## 10.18~10.24

紧急重要：1.相关论文阅读 2.论文内容填充

重要不紧急：1.运动 2.口语

紧急不重要：整理下周计划

不重要不紧急：课外阅读，leetcode

10.18情绪分析寻找委婉语

记录

10.19 1.找李老师 2.看自由泳基础 x 3.专利补充

断网消失

10.22 ：论文转英语两章；两篇论文阅读；确定蛋糕；

10.25 论文一章

## 10.26 ~10.31，，

紧急重要：1. 论文ppt 2. 创新点 3.整理完之前看的三篇论文

重要不紧急：1.运动 2.口语

紧急不重要：整理下周计划

不重要不紧急：课外阅读，leetcode

10.26 day1

10.27 day2 论文改成了模板

10.28 精读地下黑市论文，读完另一篇地下黑市工具论文

10.29 做好ppt

10.30

地下市场分析工具：

已记录

识别地下市场产品：识别地下市场产品：

使用词嵌入网络确定仇恨发言中的编码词

已记录

## 11.1~11.7

紧急重要：1. 项目任务2. 创新点 3.ai顶会上的相关论文 以及找到的剩余论文

重要不紧急：1.运动 2.口语 3.周末整理一遍内容

紧急不重要：整理下周计划

不重要不紧急：课外阅读，leetcode

11.1：项目格式改动，讨论班，看完kdes（没啥感想）

11.2

项目书再检查一遍，仔细考察一遍需要提交内容，确定进度安排，是否需要将论文填入。

课题一、二不要。四只需要基于区块链的安全分发与溯源功能模块

子课题任务书（就是整体的任务书截取你们研究这一部分），课题研究方案

四月份要开始原型，七月份要交报告。

论文：找新的年份的委婉语检测的论文，可以重点关注一下含义鉴别。

已记录

游泳：自由泳踢腿姿势练习20min左右

英语口语：半小时 x

中午阅读：半小时

11.3

逼着自己输出

项目：图改完 任务书写完

论文：相关：词级语义相似度的无监督模型

两篇检测论文 浅读即可

英语口音

阅读：

王德峰心学

Transforms 进行识别+towards codewords

11.4

项目、论文、阅读、英语

算法总结：

11.5

索要数据信：

11.6

比赛：如何根据字形和同义进行预处理

论文：精读核心论文主要思想（实验内容）、写要数据的信、俚语识别相关论文（与论文场景的不同之处）

游泳：看两个小节去实践

比赛思考：1.同形、同拼音角度检查 2.针对其他类型错误，联系鲁棒性高的解决方案 3.增加数据集数目 4.模型融合 。先自己探索为主，再使用已有方案。

论文：

已记录

11.7

论文：立项依据、研究内容、研究方案、课题特色

## 11.8~11.16

11.9

王德峰：

开题报告：要最后一份数据、观察数据情况、观察中文语境下的情况、明天约老师的时间

代码：

健身：需要锻炼肱三头肌

11.10：

王德峰：

论文：下午确定创新点，再约李老师时间。对无监督方法进行更细粒度的考察，再看中文语境下的telegram论文。

代码：notebook 如何调用其他py文件里面的函数

### 11.11

论文：用bert做自监督的相关论文，俚语辨识相关论文，Zero-shot Knowledge Transfer

Bert的mlm是否能替换为其他掩码模型

代码：写出qa 模型，最好能试试训练

方向：查询工程方向

王德峰：

### 11.12

代码：qa调试、leetcode一道

论文：俚语鉴别、自监督学习、

实验：筛选有禁用词的句子、

开题报告：缺少鉴别一部分的思路，参考词义发现的文章和之前的几种方法

工程方向 r&d d

### 11.13

1.整理一下目前的想法

2.试着补之前没做好的内容 Python 类基础知识 、正则表达式 俚语鉴别和自监督学习论文

3.完成代码等 qa调试

4.leetcode

Python:z字符串前缀 f 相当于format函数 r停止python解析器，防止将/当成转义字符，b 字符串byte类型 ；u 字符串以unicode 编码

### 11.14

Adamw

整理一下收藏夹

## 11.16 ~11.22

紧急重要： 1.项目任务 2.pytorch的优化方法 3.github练手项目寻找 4.开题的框架和报告撰写

5.阅读论文：风格识别、半监督

重要不紧急：1.运动 2.周末整理一遍内容 3.每天记录内容，并进行输出，

4 优化方法复习

紧急不重要：1 整理下周计划 2 leetcode

不重要不紧急：心流阅读完毕，

* pytorch优化方法：swa 四类 fine-tune 7类 efficiently 3类 加速5类

github练手项目寻找：中文nlp框架、比赛

* 心流阅读：51% 一天 10%

理出一个简单的流程来

* Leetcode：
* 开题的框架和报告撰写
  + 模型框架确定，可行性预估
  + 确定工作内容、创新点包装
* 阅读论文：风格识别、半监督
* 优化方法复习：

### 11.16

修笔，拿咖啡

Leetcode：

王德峰：金刚经、

优化方法：swa、question

框架：

待办不紧急：Paddle 上能不能用pytorch 怎么用、有道云笔记 怎么默认python

、有道云笔记内容格式调整、怎么合理规划任务时间、docker如何使用

paddle 上能不能用pytorch 怎么用：不能

怎么合理规划任务时间：

优化方法：

### 11.17

|  |  |  |  |
| --- | --- | --- | --- |
| 任务排序 | 意愿得分 |  | 番茄 |
| 框架撰写，开题报告重叙 | 4 | 5 | 5.5 |
| 论文阅读 | 5 | 7 |  |
| 优化方法复习 | 1 | 4 | 4 |
| pytorch 优化方法 | 6 | 10 |  |
| leetcode | 3 | 8 |  |
| 王德峰，心流 | 2 | 8 |  |

新增待做：Adadelta深究、adam详解、牛顿法、费曼学习法作用、

苏剑林博客有关bert

风格识别做prompt

向上需求：与人交互、理想自我的投射（热爱、专注）、认识自我、

情绪需求：琐碎、倾诉、

优化方法：

GD 优缺点：简单、更新次数多

动量优化法：

### 11.18

|  |  |  |  |
| --- | --- | --- | --- |
| 任务排序 | 意愿得分 |  | 番茄 |
| 框架撰写，开题报告重叙 | 5 | 6 | 1 |
| 论文阅读 | 6 | 7 |  |
| 竞赛 | 1 | 4 | 3个番茄 |
| pytorch 优化方法 | 2 | 6 | 4 |
| leetcode | 3 | 8 |  |
| 王德峰，心流 | 4 | 10 |  |

代码：

王德峰：

佛性：

数据情况：

### 11.19

|  |  |  |  |
| --- | --- | --- | --- |
| 任务排序 | 意愿得分 |  | 番茄 |
| 竞赛 | 3 | 4 | 1 |
| 框架撰写，开题报告重叙 | 4 | 6 |  |
| Pytorch优化 | 2 | 5 |  |
| 海森堡 | 1 | 5 |  |
| leetcode | 5 | 10 |  |
| 王德峰，心流 | 6 | 11 |  |

滴答清单 子任务 roc voc

MarginRankingLoss

### 11.20~21

优化方法

晚上写总结

## 11.22~11.28

紧急重要： 1.开题ppt撰写 2.项目

重要不紧急： 1.竞赛 2.运动 3.leetcode 4.pytorch优化方法

紧急不重要：1.周末整理一遍内容 2.每天记录内容，并进行输出

不重要不紧急：心流重新看一遍并整理

### 11.22

|  |  |  |  |
| --- | --- | --- | --- |
| 任务排序 | 意愿得分 |  | 番茄 |
| 开题报告 | 2 | 3 |  |
| 项目 | 3 | 5 |  |
| 竞赛 | **1** | 4 |  |
| Pytorch 优化方法 | 4 | 8 |  |
| 王德峰,阅读 | 5 | 10 |  |
| leetcode | 6 | 12 |  |

Leetcode刷题：

### 11.24

上午：论文问题并整理

下午：外勤

晚上：kaggle

方向word该怎么使用：

王德峰 心即理 第一讲

### 11.25

上午：发现问题 下午：摆烂

竞赛：单个模型优化、badly的使用、数据分离同时提交

### 11.26

本地cv 、代码整理

思考关于该怎么做

基于术语特性进行后处理。黑暗术语是人为刻意选择，具有一些可被利用的特性，如名词、与禁用词不相干、不同语料下的词义不同。通过术语特性来对MLM模型结果进行后处理，对结果进行筛查，提高准确率同时也避免影响模型本身的鲁棒性。

### 11.27

ppt制作、自由泳练习

烦恼的事：当朋友：peer pressure 就当是考验 必须去克服。

tqdm、python 格式化、lighting

一周的总结 代码整理

## 11.29~12.5

紧急重要： 1.开题ppt撰写 2.项目

重要不紧急： 1.竞赛 2.运动 3.leetcode 4.pytorch优化方法

紧急不重要：1.周末整理一遍内容 2.每天记录内容，并进行输出

不重要不紧急：心流重新看一遍并整理

|  |  |  |  |
| --- | --- | --- | --- |
| 任务排序 | 意愿得分 |  | 番茄 |
| 开题报告 |  |  |  |
| 项目 |  |  |  |
| 竞赛 |  |  |  |
| Pytorch 优化方法 |  |  |  |
| 王德峰,阅读 |  |  |  |
| leetcode |  |  |  |

### 11.29

ppt撰写：

wb测试：wb教程写，

leetcode：

模型优化： scheduler了解

模型训练完输出一些 验证的文件 方便在本地对bad end进行分析

### 11.30

竞赛：数据增强

Leetcode：

Pandas .map .apply 使用上的区别 axis 详解

List 使用 pop 出 append进 菜鸟教程

Python 精进 ，numpy、pandas、torch系统学习

怎么拍落日

粗细分类器：粗粒度负责全局特征， 细粒度负责身体部位等进行动作识别

### 12.1~2

Ppt：检查

竞赛：错误分析、论文

列表对顺序敏感

在pands 中用 多组条件筛选列时 应该 用（）并使用＆，|，〜进行条件连接

错误分析：关注错的 和与score 偏离太多的

拆成对的 和错的，在对的中分析 找出正确率与score 偏离太多的

画相关关系图作为可视化

Validation 这部分数据有正确答案

不要落入猜对所有pair\_True 陷阱中

如何解决fold情况下 模型对同一文本的重复评估 （严重干扰了我们的cv估计）

同时cv 要真实反映模型能力

### 12.3

Cv策略1：每个fold 单独编号 评估计算准确性 ，

Cv 策略2：每个fold中文本不单独评估 ，重复评估使用平均值计算准确性

.copy .view 区别 什么时候会造成错误

竞赛：

错误分析:分析错误的最厉害的 也就是value\_gap最大的

以及在

在正确评估中

在错误评估中

在所有样本中

## 12.6~12.12

紧急重要： 1.开题ppt修改 2.项目

重要不紧急： 1.竞赛 2.运动 3.leetcode 4.pytorch优化方法

紧急不重要：1.周末整理一遍内容 2.每天记录内容，并进行输出

不重要不紧急：心流重新看一遍并整理

|  |  |  |  |
| --- | --- | --- | --- |
| 任务排序 | 意愿得分 |  | 番茄 |
| 开题报告 |  |  |  |
| 项目 |  |  |  |
| 竞赛 |  |  |  |
| Pytorch 优化方法 |  |  |  |
| 王德峰,阅读 |  |  |  |
| leetcode |  |  |  |

### 12.6

方法已经有了，怎么确定目标，怎么选择目标呢。

竞赛：如何做错误分析

开会

Ppt整合

论文提交

专利修改

在错误样例中 1397：224 大部分错误和评估把握不大有关

### 12.8

上午：开题报告撰写

下午：可能要开会，不行就看论文

定个时间集中去想 12：05~12：40

王德峰：理与心的关系：理学与心学的对立，理学认为先学理后有心。心学认为先有心后有理。心就是生命情感，生命情感来自人与人的关系当中。中国人的价值落实在五伦（君父妇兄朋）中。

致良知：良知受到私欲的遮蔽和逻辑的干扰。

知行合一：去实践中去听良知。如同种树，花也好、果也好都只是结果，最重要的时把根种好。

蛋白质摄入不够？

拉伸学习

瘦腿

Cosine Annhealing Learning Rate、Cosine Annealing Warm Restarts

Lighting 用于减少重复性工作，暂时用不到

参数参考 adamw

Swa效果好，weight decay 1e-6->1e-3 eoslion 1e-8->1e-6 贝塔2 0.999

Fp16 可以开启

三阶段工作

使用fp16 时间对比

方向：数据增强、过去方案、tf-idf集成、错误分析

两到三个模型开始：从非nn开始、对base进行优化尝试、新的bert、

验证方法1 额外数据 训练 val 验证 2.val 做fold训练 3.剩余10k加入训练

### 12.9

Lightgbm学习

### 12.11~12

王德峰：

哪些实验要做

Swa\_use 的bug ： 降低学习率到1e-4 swa\_use cv 0.70257 lb 0.820

fp16没有加速效果原因、

调参、 wd、eps调整后 cv 0.746512 lb 0.795

调参+swa\_use cv 0.6228577 lb 0.781

Tf-idf 上：两个notebook 的学习使用结合、 新旧数据的使用策略、

Ensenmble of 3 cv 0.69138 lb 0.83

Fold策略

37min fp16 不用fp16 31

25s 26s 可能原因是对p100不进行支持 以及batch较小的情况下 效果一般

Cv策略：对单个comment 进行fold （但是会损失一部分标注）

Cv1：对所有评估句子对fold之后形成comment 评估，在总comments上求均值

哪个更贴近在leaderboard上

Fp16 和 swa\_use 的用法总结

总结做好

相信独立个体是存在的、更好掌控自我、度过人生空虚。做选择 一定是选择做好

## 12.13~12.19

紧急重要： 1.项目 2.竞赛

重要不紧急： 2.运动 3.leetcode 4.pytorch优化方法

紧急不重要：1.周末整理一遍内容 2.每天记录内容，并进行输出

不重要不紧急：心流重新看一遍并整理

|  |  |  |  |
| --- | --- | --- | --- |
| 任务排序 | 意愿得分 |  | 番茄 |
| 竞赛 |  |  |  |
| 项目 |  |  |  |
| Pytorch 优化方法 |  |  |  |
| 王德峰,阅读 |  |  |  |
| leetcode |  |  |  |

### 12.13

现有的成绩会阻碍你的思路

更新fold策略 在新号上进行cv测试

人像钻研书

水果采购

np.aroud 、StratifiedKFold

### 12.14

Jc 数据有重复 但是主要体现在 test 里面的未标注数据

### 12.15

Bert：pretrain 方法、输出模型

回归问题转分类问题

将过去比赛数据转换为 less和more的形式

### 12.16

保存字典solve过去 gbm保存调用方式

确定clean 使用 数据集的抉择 r是否可替代 写inference

data\_name model 写好字典保存 factor 不传了拿去过拟合 //都不传了

### 12.18~19

RealToxicityPrompts dataset (2020): contains 100k sentence snippets from the web with Perspective API toxicity scores for studying the risk of neural toxic degeneration in language models."

"OLID (Offensive Language Identification Dataset; 2019): contains 14,100 English tweets, annotated according to the three-level taxonomy as described here. "

评论被标注为使另一个人脱离对话的概率。有毒被定义为任何大于0.5的概率。

试试看去掉关键字 对各个模型的影响

Stack+blend的方法

当模型足够多样化时 考虑将其纳入 如果模型的 Pearson 相关性 > 0.99 且 Kolmogorov-Smirnov 统计量 < 0.01，则混合它们不太可能获得太多好处。一个非常宽松的经验法则：您需要 Pearson < 0.95（< 0.9 更好）和 K-S 统计 > 0.05（> 0.1 更好）的模型。

计算结果脚本在这

<https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/discussion/50827>

实践一下 翻译增强方法 在ridge 上 和在bert上 加油 submission 别全用掉了（防止在不同fold上翻译造成的data\_leake）

新的消除量纲方法在 ridge 上效果

有没有类似量纲的情况 干扰了 gbm和 juc 的值

Parser

线性模型再优化

## 12.20~12.26

紧急重要： 1.项目 2.竞赛

重要不紧急： 2.运动 3.leetcode 4.pytorch优化方法

紧急不重要：1.周末整理一遍内容 2.每天记录内容，并进行输出

不重要不紧急：心流重新看一遍并整理

|  |  |  |  |
| --- | --- | --- | --- |
| 任务排序 | 意愿得分 |  | 番茄 |
| 竞赛 |  |  |  |
| 项目 |  |  |  |
| Pytorch 优化方法 |  |  |  |
| 王德峰,阅读 |  |  |  |
| leetcode |  |  |  |

### 12.20

专利改完、leetcode和百面机器学习\

### 12.21

Linear model优化、翻译增强bert、ppt

代码更加高效（多写测试、读一遍代码逻辑、写之前考虑周全）

使用过去数据作为cv

寻找线性模型cv 变高原因 ，线性组合方面使用更多清洗、组合

针对单个模型进行 fold 参数等的调优

linear 模型中 再训练一个 val 不计算val 的cv

进行翻译增强的bert

Python 编程快捷键 python 如何高效且出错少的方法撰写

if \_\_name\_\_ == '\_\_main\_\_': 用于确认执行的主程序，确保其余py文件被引用时不会被执行

记录线性过程中的实验数据

### 12.23

Swa French 4epoch 7 hour \

整理完linear 后对 代码进行一个仔细的 融合 。。。。。。 修改ml\_inference

### 12.24

在linear model上对比实验验证 哪些有improve

重写ml inference 本地一份 linear 一份 inference 保证能复现结果为止

试试 fr 2swa 效果如何

实在把握不住 提升在哪里就放弃

按行处理的速度比较慢 如果可以直接索引会快很多

0.48841081, 0.21194874, 0.39584097, 0.40030452, 0.46899059, 0.42384744])

## 12.28~1.2

### 12.28

写简历并投递

Python 赋值 情况

面试 还有两周 完成百面

正则表达式\1 匹配第一个括号分组内的内容 replace不具备正则表达式 pd.str.replace 可以

### 12.29

### 12.30

考虑代码可能的越界情况

层学习率 与优化器学习率的交互

设计合理的val函数再去 进行batch 的评估

Noleak 方法的理解

将所有的比较做成链，作为以个group，同一group不拆开分到 train /dev里面

Rud 4epoch 0.6861963 0.775

Rud 3epoch 0.6863624

Rud 3epoch batcheva 0.6860 0.757

Rud 4epoch 10eva 0.6818 0.768

Rud 4epoch 10eva <0 0.683 0.755

Rud 4epoch 翻译增强 0.6855 0.761

Rud 4epoch hatebert 0.67012 0.724

val noleake 3epoch 4层输出结构 0.698385811 0.797

val noleake 3epoch origin 0.6949 0.789

val 3epoch origin 50eva 0.698

Linear

Jc\_fr 65.99

Jc\_de 65.88

Jc\_es 66.5

每个batch 进行调优

Swa debug

**Rud**

不选择将支持度置为0

翻译增强 在rud上用法

不同模型 hatebert 尝试

Swa

用cv 作为验证部分 Robert hatebert

**Jc**

新的cv策略 在baseline1上

Swa 使用 参数… 用到原baseline1上

改了train函数 增加了import 改了相应CONFIG 调用

from torch.optim.swa\_utils import AveragedModel, SWALR

baseline1 的train函数改写 使用swa

3epoch 可能刚好 使用origin model 使用 多个batch方式

使用juc 数据集 使用jc数据集

**Linear model**

Linear 再努努力 使用翻译数据 单独扔出来试试看

Pd.na() 用于判断没有isna 的类型是否包含 pandas定义的nan

## 1.3~1.10

紧急重要： 1.刷题 2.竞赛 3.面经

重要不紧急： 2.运动

紧急不重要：1.周末整理一遍内容 2.每天记录内容，并进行输出

不重要不紧急：心流重新看一遍并整理

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Val | roberta | 更多epoch和学习率以及warm up配合调整、更多fold。策略组合组装roberta |
| Other | Linear | 1.模型整理，将相同类型的进行整合，排除比例干扰、2.检索两个notebook 有效性 3.赋值的bias 性 |
|  |  |  |
|  |  |  |
|  |  |  |

组装顺序：fold选择、epoch和学习率、dropout0、结构选择、翻译增强、warm up 、swa

Large

### 1.3

高性能计算需要的内容

Spark 、hadoop、flink

ETL工具的典型代表有:Informatica、Datastage、OWB、微软DTS等。

具备分布式训练下切割模型、调参、调试调优的能力。

数据挑选

### 1.4

剑指offerr100

编程规范

单调栈用于 动态表示递增信息 记得是否要清栈 计算所有结果

BERTweet (Nguyen et al., 2020)

– XLM-T (Barbieri et al., 2021)

Electric

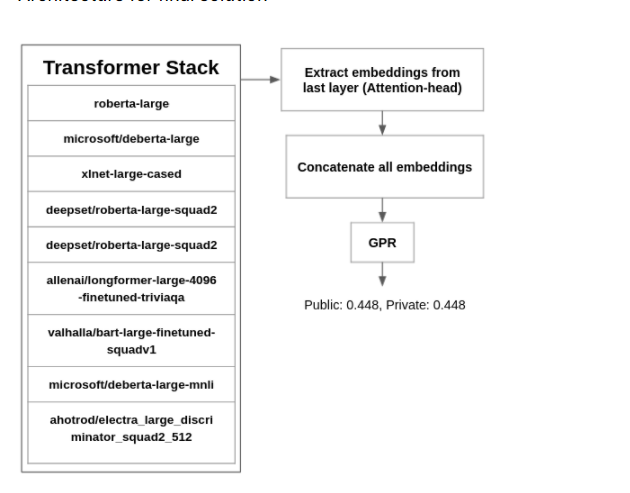
Gpt 2

Earily stop

Train 2-3 epochs with Warmup Linear Schedule

Replace model head with 2 layer DNN with BN

1.5



### 1.5

1.照相纸。请按照LNCS格式准备好相机准备纸（PDF版本）。LNCS模板可从以下位置下载：https://www.springer.com/gp/computer-science/lncs/conference-proceedings-guidelines.鼓励在论文中加入作者的兽人。

2.论文的源文件。请准备论文的源文件，包括word文件或Latex文件包。

3.版权表格。请准备版权表并签字。版权表格模板（Springer copyright form.doc）可从以下网址下载：https://www.icaisconf.com/registration/.

5.作者名单应至少包括一家海外机构。

7.照相机就绪纸张的文件应命名为“Paper+Paper ID+Field

”，源文件应命名为““Source +Paper ID+Field

”，版权表格的文件应命名为“Copyright+Paper ID+Field

Encryption and Cybersecurity Cloud Computing and Security, Information Hiding, IoT Security

”。这些领域包括人工智能、大数据、云计算和安全、信息隐藏、物联网安全、多媒体取证、加密和网络安全。请在其中选择一个字段。

8.相机准备好的论文、论文源文件和版权表必须通过会议提交系统提交(https://cmt3.research.microsoft.com/User/Login?ReturnUrl=%2FICAIS2022%2F)2022年1月23日之前。否则，您的论文将不会在会议记录中发表。

### 1.6

生成0列表的注意事项

Python 跳过一个单词删除一个单词 操作快捷键

Github 提交了

### 1.7

检查fold num 代码正确性

106.121.184

106.101

10.112 我的

10.28.218

## 1.11~1.17

紧急重要： 1.刷题 2.竞赛 3.面经

重要不紧急：

紧急不重要：1.周末整理一遍内容 2.每天记录内容，并进行输出

不重要不紧急：

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Val | roberta | 更多epoch和学习率以及warm up配合调整、更多fold。策略组合组装roberta |
| Other | Linear | 1.模型整理，将相同类型的进行整合，排除比例干扰、2.检索两个notebook 有效性 3.取消第一届 积极toxic 4.增加更多数据集  5. 调整factor的比例 |
| Val | roberta | Swa 效果、层上drop取消、最早的swa 4设置 |
| Val | roberta | 复现start 、增加7fold 、分层、swa、翻译 |
| Val | roberta | Gpt2 robert large albert xlbert electbert |
|  |  |  |
|  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | cv | lb | loss |
| Roberta S2lastcls nod | 0.6906 |  | 0.3456 |
| Roberta 7f/3e origin | 0.6954 | 0.808 | 0.3438 |
| Roberta origin no drop config | 0.6894 | 0.788 | 0.3443 |
| Roberta maxpooling d | 0.6798 | 0.784 | 0.3573 |
| Roberta swa 7f 4e | 0.6917 |  | 0.3428 |
| Roberta 原设置 | 0.6974 | 0.791 | 0.34 |
| Roberta 3e层drop取消 | 0.697 | 0.782 | 0.343 |
| Roberta 2epoch | 0.6926 | 0.785 | 0.3452 |
| Roberta 3epoch 修复分层 | 0.6914 | 0.801 | 0.3459 |
| Roberta 3epoch 不要分层 | 0.6935 | 0.788 | 0.3423 |
| Roberta 3e 2e-5 | 0.6981 | 0.795 | 0.3387 |
| Roberta 3e 1e-5 | 0.6952 | 0.795 | 0.3401 |
| Roberta 3epoch 分层 2e-5 | 0.6995 | 0.800 |  |
| 复现starter | 0.65982 | 0.824 |  |
|  |  |  |  |
| Roberta large 2epoch |  | 0.803 | 0.3483 |
| Gpt 2 |  |  |  |
| Luke model | 0.6941 | 0.827 |  |
|  |  |  |  |
|  |  |  |  |

Jigsaw starter 0.816->0.806 0.3555

Starter 0.825 2eopch

|  |  |  |  |
| --- | --- | --- | --- |
|  | cv | lb | loss |
| Fasttext | 67.7361 |  |  |
| 不要test | 67.7162 | 0.880 |  |
| 全部数据 要test | 68.0284 | 0.880 |  |
| 0.5比例 全部数据 | 68.0384 | 0.876 |  |
| 全部数据 only fasttext | 65.4743 | 0.824 |  |

比赛：

无泄漏fold会带来更低的cv

合适的rank loss函数？

## 1.17 ~1.23

紧急重要： 1.刷题 2.竞赛 3.面经

重要不紧急：

紧急不重要：1.周末整理一遍内容 2.每天记录内容，并进行输出

不重要不紧急：

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Val | roberta | albert xlbert electbert |
| Val | roberta | 合适的rank loss函数？一致性计算 、本地cv、新的val方式 |
| Val | luke | Luke调整、尝试翻译 model,尝试在翻译时使用drop 在luke上实验, 检测不同翻译版本的预测结果上的差异性 ，cv上是否能有所提高 |
| linear |  | A= 0.5 a= 1 a=2 Baseliine 训练在服务器上 |
| 数据 |  | 检查 groupfold 是否考虑相同评估在一个fold里面 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | cv | lb | loss | 4：3.8 |
| Roberta fuxian 分层1e-4 | 0.6993 | 0.810 | 0.3403 | 0.75323 |
| Roberta 翻译 | 0.6663 | 0.821 |  | 0.741666 |
|  |  |  |  |  |
| Luke model | 0.6941 | 0.827 |  | 0.758846 |
| Albert v2 | 0.681 | 0.818 | 0.3724 | 0.7477 |
| Luke 2e nodrop | 0.6956 | 0.829 |  | 0.7605897 |
| Luke 3e | 0.6936 | 0.819 |  | 0.754692 |
| Xlnet | 0.6928 | 0.808 |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Jigsaw starter 0.816->0.806 0.3555

Starter 0.825 2eopch

|  |  |  |  |
| --- | --- | --- | --- |
|  | cv | lb | loss |
|  |  |  |  |

比赛：

合理的本地cv：确定5%队数据会在多少对中出现 1.不重复评估的对 2.具有重复评估的队

Comments到相关句子对的工作

随机抽样5%句子对 做成句子 然后再找存在的句子 计算比例 重复多次求平均值

3：1

14k在train上 7k在val上

5% 句子对 能多大程度影响到 30%的句子对

Cv 分割比例 4：3.8 1：1

数据增强时需要drop 来限制发挥

<http://10.126.239.153:8888/tree>? BYR06215x

screen -S yourname 创建一个名为 yourname的screen

screen -ls 列出所有session

screen -r yourname 恢复到这个session

screen -d yourname 挂起某个session

ctrl a 新的终端

screen -list 查看session

screen -S yourname -X quit 结束session

tar -zxvf filename.tar.gz

unzip filename.zip

nonlocalpython解释器在处理变量的时候，如果没有出现会去上下文中寻找，如果不需要对这个变量赋值的话，这个变量就利用闭包的特性可以直接使用，如果需要赋值，那么变量会被解释器视作 局部变量

## 1.24~1.30

紧急重要： 1.刷题 2.竞赛 3.面经

重要不紧急：

紧急不重要：1.周末整理一遍内容 2.每天记录内容，并进行输出

不重要不紧急：

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Val | roberta |  |
| Val | Xlnet |  |
| Val | luke | 尝试更大的lukemodel？ |
| Val | Albert |  |
| linear |  | 输出val上预测结果方便做最后的cv， |
| 数据 |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | cv | lb | loss | 4：3.8 |
|  |  |  |  |  |

Jigsaw starter 0.816->0.806 0.3555

Starter 0.825 2eopch

|  |  |  |  |
| --- | --- | --- | --- |
|  | cv | lb | loss |
| Luke 0.7e | 0.6969 |  |  |
| Luke large | 0.707 |  |  |
| Albert 4epoch | 0.681 |  |  |
| Albert margin0.4 | 0.6935 |  |  |
|  |  |  |  |
|  |  |  |  |
| ruddit | 0.6965 | 0.860 |  |
| Ruddit 5sample | 0.6951 | 0.845 |  |
| Ruddit m=0.3 sample=5 | 0.69533 | 0.842 |  |
| Ruddit 3sample 0.4 | 0.6972 | 0.850 |  |
|  |  |  |  |
| Ruddit 3~8 9 | 0.6955 | 0.857 |  |
| Ruddit 3~8 9 2samples | 0.6977 | 0.855 |  |
| Ruddit 3~8 9 s=1 loss | 0.6938 | 0.847 |  |
| Ruddit 2~8 9 s=1 score | 0.6957 |  |  |
| Ruddit 1~8 9 s=1 score 0.67 | 0.6944 |  |  |
| Ruddit 3~7 9 s=1 score | 0.6947 |  |  |
| Ruddit 0~2 4 s=1 score | 0.6979 |  |  |
|  |  |  |  |
| Val 5 0~3 | 0.7053 | 0.834 |  |
| Val 5 0~3 | 0.70779 | 0.818 |  |
| Val |  |  |  |
|  |  |  |  |
| Subtask 28 9 3samples | 0.6947 | 0.825 |  |
|  |  |  |  |
| Distll roberta | 0.695008 |  |  |

差距较大模型得用不同文件保存

Luke 可以尝试较大模型 ? 其他模型有待商榷，但是luke既然本身效果好可以试试。

Luke 在kaggle上训练 luke large在服务器上进行

目前 luke 0.76058 albert 0.7477 xlnet 0.7489

数据增强尝试

Python append 的对象是浅拷贝

Series rename(new name)

67.52 66.06 67.71 67.89 62.61 65.65 67.8756

67.55 66.1 67.72 67.74 62.91 65.5 67.7793

67.47 65.85 67.67 67.74 63.05 65.36 67.7361

1 1 1 0.5 2 0.5