

中医药天池大数据竞赛 比赛总结及答疑

导师：向右



关注公众号
获取第一手干货咨询








添加小享
获得Baseline&课件


















比赛相关

baseline-代码框架图

Baseline思路、框架1

- ▶  coding
- ▶  data
- ▶  submit
-  README.txt
-  requirements12.txt

Baseline思路、框架2

- ▶  bert
- ▶  cache
- ▶  data
- ▶  model_saved
- ▶  tf_utils
-  conf.py
-  data_utils.py
-  eval_metrics.py
-  infer-cv.py
-  infer.py
-  label2json.py
-  model.py
-  optimization.py
-  train-cv.py
-  utils.py



比赛相关

代码设计思路

Baseline, 线上0.765, 整体代码设计思路:

1. 简单定义即可修改网络结构:
支持采用原生bert最后一层 或 最后多层进行融合, 也可自行设计;
支持修改bert+不同网络结构 (BILSTM、CNN) 进行encoding, 也可自行尝试新的结构。
2. 严格按照构建验证集方式, 记录实验结果:
支持模型训练过程中保存每个epoch下验证集对应的准确率、召回率、F1值, 用于挑选最优模型;
支持设置交叉验证, 同步记录实验结果。
3. 训练样本目标构造方式上, 采用IOBS方式, 如想修改设计思路, 可自行修改, 其他代码复用。
4. 解码方式采用crf, 可以根据需要修改为softmax与sigmoid等, 代码框架可复用。
5. 如果没有其他设计思路, 可以利用该整合版本代码跑不同的实验结果, 进行模型融合。

比赛相关

上分点

1. 调参, lr | batch_size | dropout | bert最后一层, 还是最后多层 | bert+cnn? 抑或 bert+rnn?
2. bert输出的动态加权
3. 模型融合 (很重要的上分点) | PS: 多实验, 多记录过程, 实时保存最好的模型文件, 最后进行模型融合。
a) 概率融合; b) 投票融合。
4. 半监督迁移学习 【网上找公开的医学相关数据, 最好数据分布差异小】
5. 使用chinese_roberta_wwm_large_ext_L-24_H-1024_A-16 (对机器要求高)
6. state of the art 模型使用 (flat)
7. 过滤不合理的解码输出 【crf如何控制】

目录

- 1/ 知识点总结(25分钟)
- 2/ 其他比赛方案(25分钟)
- 3/ 面试相关问题(5分钟)
- 4/ 互动时间(10分钟)

1、知识点总结

Summary of knowledge



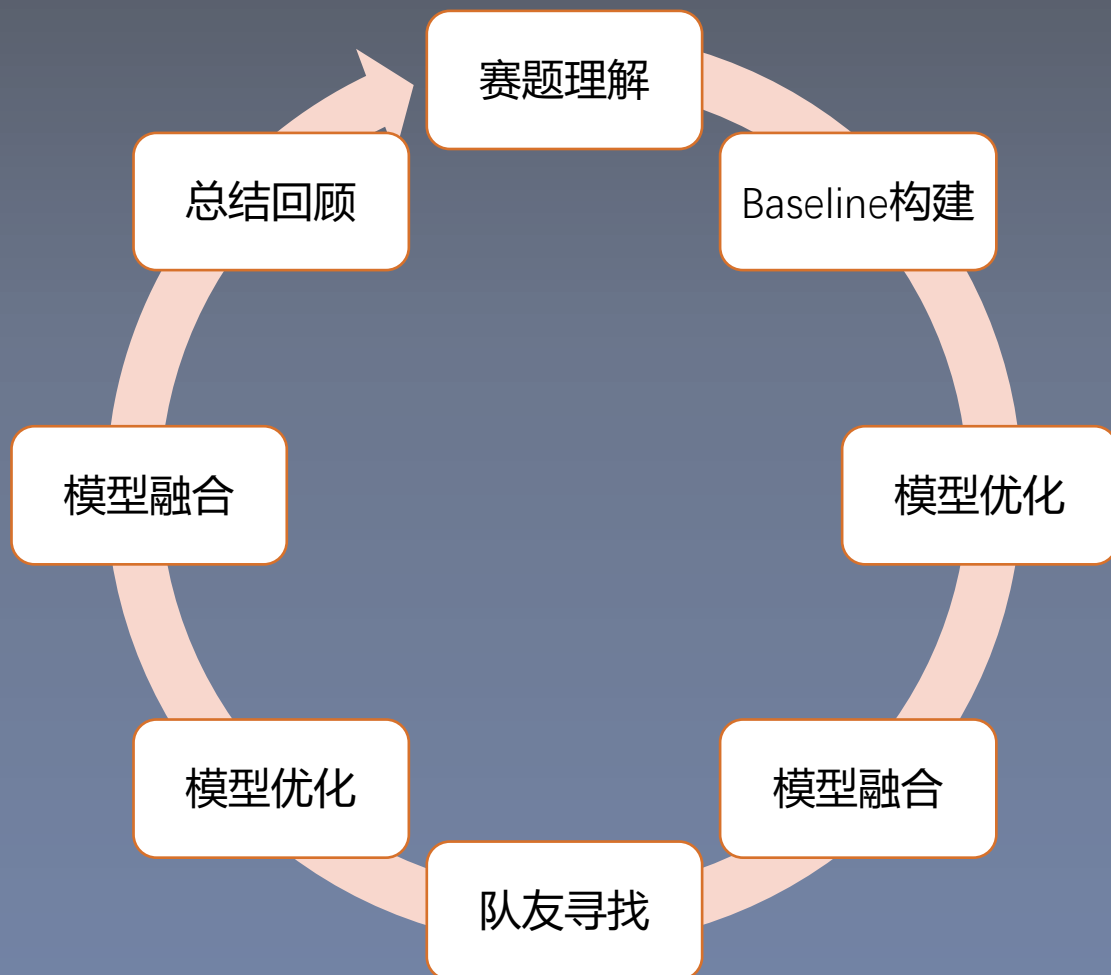
原价1498比赛会员
现在立减100
仅限20张!



添加小享
获得Baseline&课件

我们如何带大家学习这门课

学习流程



- 01 ➤ 赛题理解、Baseline解析
- 02 ➤ Transformer、Bert、XLNET、Roberta等深度学习预训练模型
- 03 ➤ softmax、CRF解码介绍
- 04 ➤ 模型训练与验证、数据扩增、过拟合与欠拟合、模型融合
- 05 ➤ 介绍NER任务的state-of-the-art方法(模型)
- 06 ➤ 比赛思路全复盘



知识点总结

Summary of knowledge

Baseline构建

迁移学习/**Transformer**/预训练模型介绍

哪些步骤没有做?

编码器与解码器串讲、数据增强、伪标签

标签缺失、标签错误、模型融合

state-of-art 论文分享

Transformer框架

Transformer

- **Left: Encoder**
- **Right: Decoder**

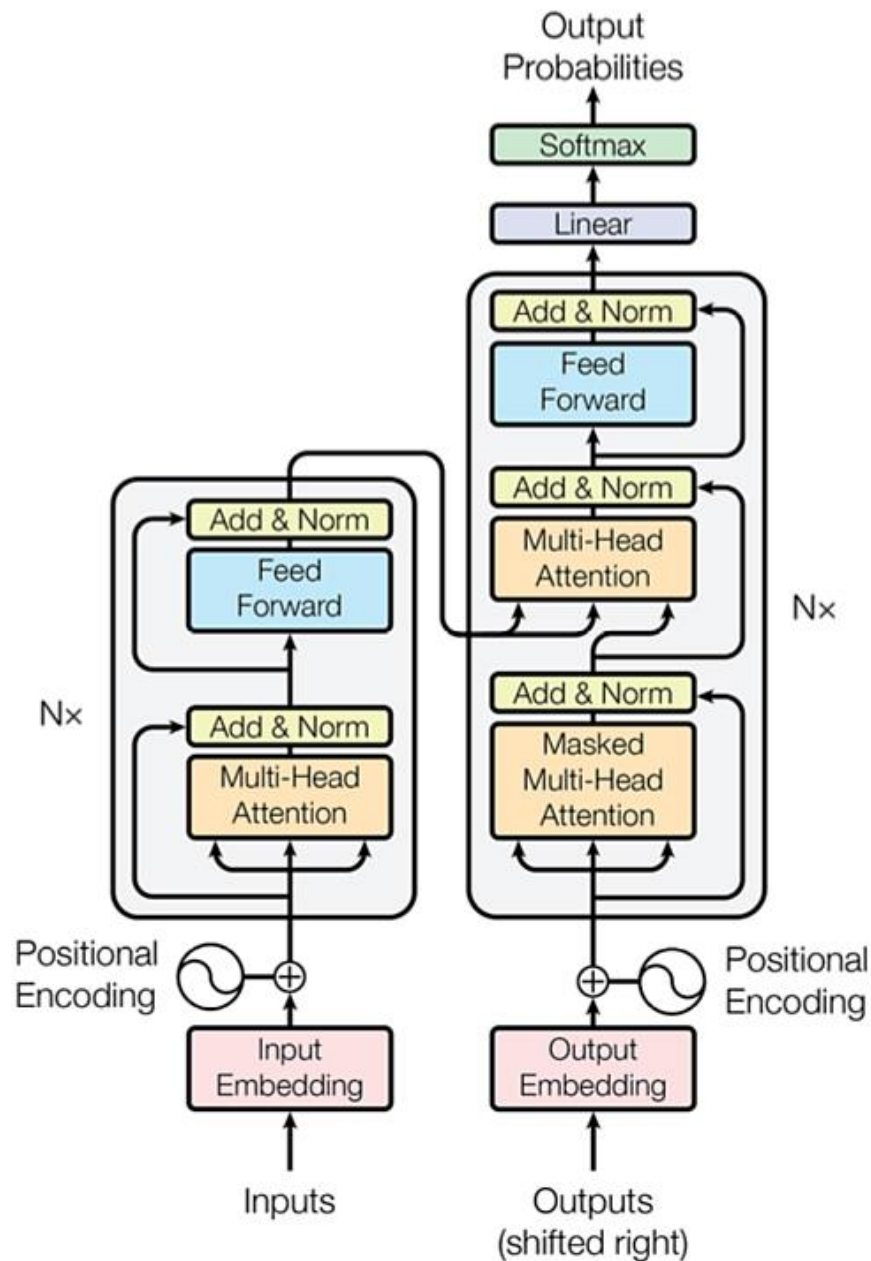


Figure 1: The Transformer - model architecture.



深度之眼
deepshare.net

编码器扩展

- BERT-CRF
- BERT-BILSTM-CRF
- BERT-IDCNN-CRF
(<https://arxiv.org/pdf/1702.02098.pdf>)
- BERT多层表示的动态权重融合
- 不同预训练模型替代BERT模型(详见专题一)





解码器串讲

- Softmax
- **CRF**
- MEMM
- HMM
- Sigmoid

原理参考：

<https://kexue.fm/archives/5542>

《统计机器学习》

<https://kexue.fm/archives/7213>



标签缺失、标签错误

知识蒸馏：

训练集有一定缺漏和不规范的，因此，可以尝试一种类似知识蒸馏的方式来重新整理训练集，改善训练集质量。

比如

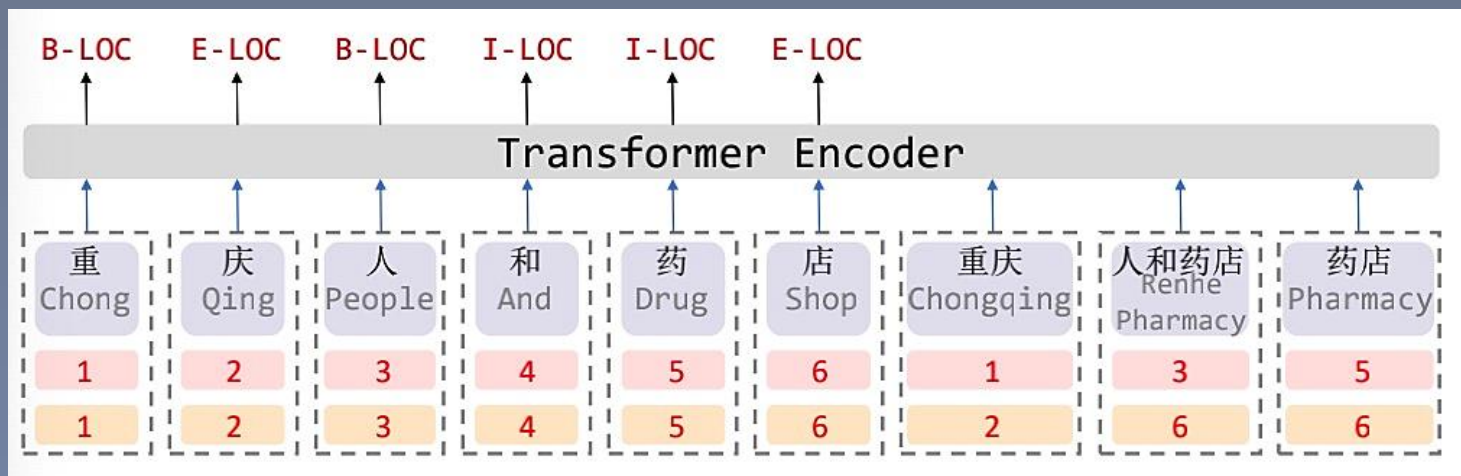
首先，使用原始训练集加交叉验证的方式，得到了8个模型，然后用这8个模型对训练集进行预测，得到关于训练集的8份预测结果。如果某个样本的某个标签同时出现在8份预测结果中但没有出现在训练集的标注中，那么就将这个标签补充到该样本的标注结果中；如果某个样本的某个标签在8份预测结果中都没有出现但却被训练集标注了，那么将这个标签从该样本的标注结果中去掉。



NER

FLAT

从Transformer的position representation得到启发，作者给每一个span(字、词)增加了两个位置编码，分别表示该span在sentence中开始(head)和结束(tail)的位置，对于字来说，head position和tail position是相同的。





NER

FLAT

通过这种方式，可以从这样的标签序列中无损地重建Lattice结构。同时，这样扁平的结构允许我们使用Transformer Encoder，其中的self-attention机制允许任何字符和词汇进行直接的交互。

$$\text{Att}(\mathbf{A}, \mathbf{V}) = \text{softmax}(\mathbf{A})\mathbf{V}, \quad (1)$$

$$\mathbf{A}_{ij} = \left(\frac{\mathbf{Q}_i \mathbf{K}_j^T}{\sqrt{d_{\text{head}}}} \right), \quad (2)$$

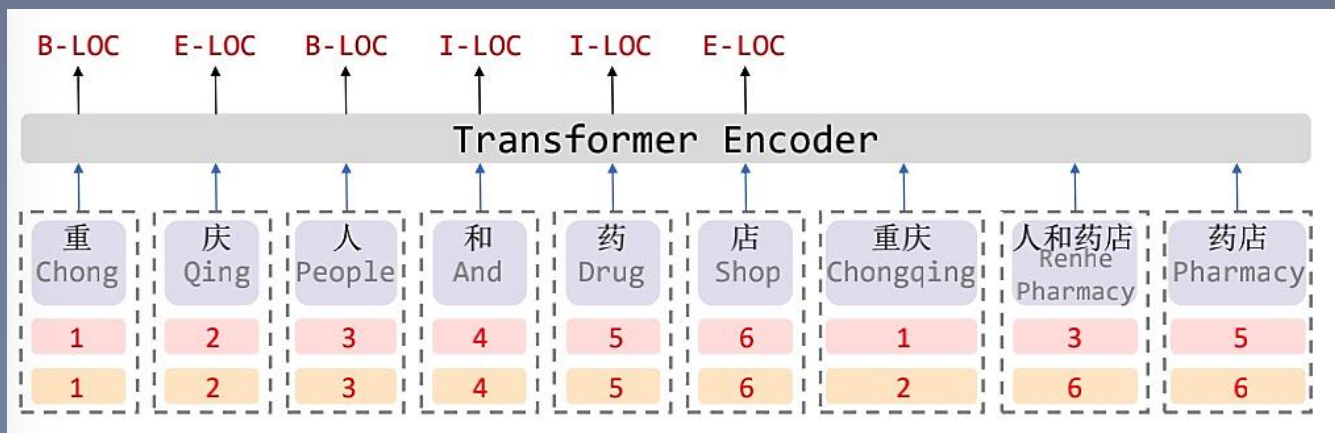
$$[\mathbf{Q}, \mathbf{K}, \mathbf{V}] = E_x[\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v], \quad (3)$$



NER

FLAT

span是字符和词汇的总称，span之间存在三种关系：交叉、包含、分离，然而作者没有直接编码这些位置关系，而是将其表示为一个稠密向量。作者用 $head[i]$ 和 $tail[i]$ 表示span的头尾位置坐标，并从四个不同的角度来计算 x_i 和 x_j 的距离：

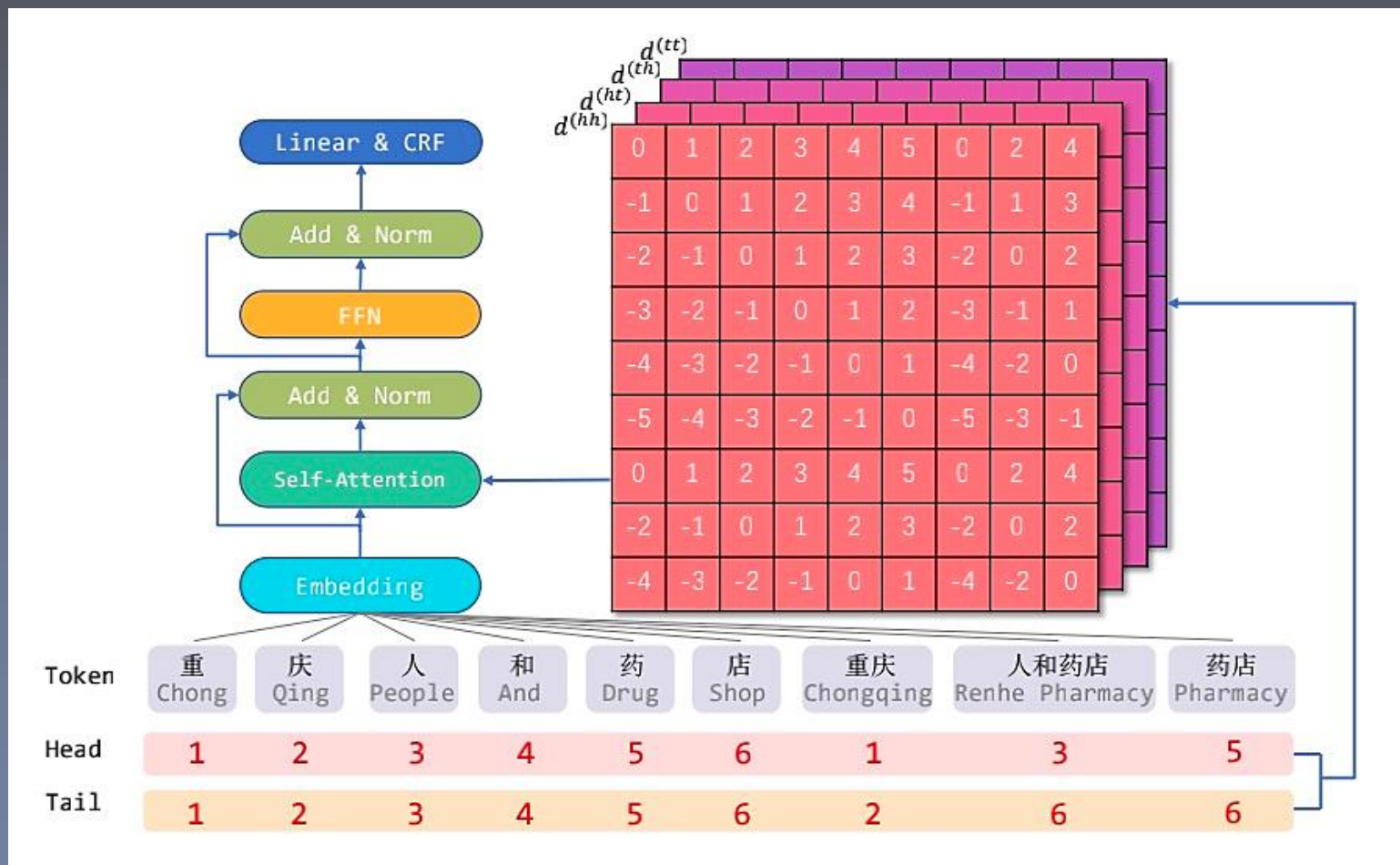


$$\begin{aligned}d_{ij}^{(hh)} &= head[i] - head[j] \\d_{ij}^{(ht)} &= head[i] - tail[j] \\d_{ij}^{(th)} &= tail[i] - head[j] \\d_{ij}^{(tt)} &= tail[i] - tail[j]\end{aligned}$$

NER

FLAT

此时得到四个相对距离矩阵：
 $d^{(hh)}$, $d^{(ht)}$, $d^{(th)}$, $d^{(tt)}$ ，其中
 $d_{ij}^{(hh)}$ 表示 x_i 的开始位置和 x_j 的
开始位置的距离。



The overall architecture of FLAT.

2、其他比赛方案

Other Competition Solutions



原价1498比赛会员
现在立减100
仅限20张!



添加小享
获得Baseline&课件



其他比赛方案

Other Competition Solutions

2019之江杯——电商评论观点挖掘

2st: https://mp.weixin.qq.com/s/AP_KNhHZ8ubuh_XbW3MqCw

其他比赛方案

Other Competition Solutions

两年前的文本分类比赛方案：

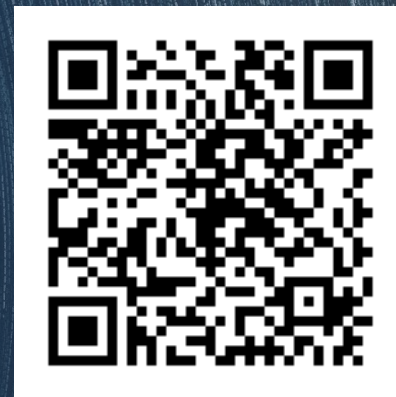
- 文本增强
- 伪标签
- CNN、RNN、Capsule、TFIDF+NBSVM等等

参考：① <https://zhuanlan.zhihu.com/p/34899693>

② <https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge>

3、面试相关问题

Interview Related Questions



原价1498比赛会员
现在立减100
仅限20张!



添加小享
获得Baseline&课件



面试相关问题

Interview Related Questions

项目介绍

原理介绍

Leetcode



面试相关问题

Interview Related Questions

- 请你描述一下竞赛背景，以及你在其中的工作/职责。
- 比赛方案的难点与痛点？你是用什么方法解决这个问题的？
- word2vec的词向量，知道它是如何训练的吗，有哪些trick。
- glove了解吗？elmo呢？
- xlnet和bert有啥不同。
- roberta，它和bert的区别？
- albert了解吗？



面试相关问题

Interview Related Questions

- LSTM你了解吗，它的大致结构是怎么样子的？
- 模型过拟合了，有哪些调整方法？
- Transformer的结构什么样的？
- xlnet和bert有啥不同。
- roberta，它和bert的区别？
- albert了解吗？



个人经验

Personal experiences

- 比赛要打
- 论文要刷
- leetcode要刷
-



原价1498比赛会员
现在立减100
仅限20张!



添加小享
获得Baseline&课件

如何学习AI竞赛?

How to learn AI competition?



会员打包价仅需1498
领券还能立减100



深度之眼
deepshare.net

注：中药说明书实体识别指导班的同学
还可限时返学费!

Step0: 选修知识

数学基础

Python基础

图像基础

NLP基础

深度方向

Step1: 参加经典赛练习

四大方向+十三场经典赛

数据科学

NLP方向

CV方向

综合方向

Step2: 参加进行的新比赛

Kaggle



TIANCHI天池



Step3: 上TOP

拿奖金

奖励/内推/实习

PS欢迎来当讲师 (长期跪舔TOP大神)

解决**基础不牢固**
替你**查漏补缺**

按照个人学习能力和技术深度，设计了不同阶段课程，带你**层层提升**。

轻松入门 CV / NLP
扎实细分领域

<https://ai.deepshare.net/all/3279059>





深度之眼
deepshare.net

深度之眼本月正在开班中指导班

Kaggle-回答准确性预测大赛（传统机器学习方向）



某大厂互联网公司算法工程师

曾获多个比赛冠军

阿里云天池、DataFountain、京东零售科普讲师

讲师: Cookly

N次（N~10）竞赛Top3

仅列举一部分获奖

达观推荐1st 携程出行行行销售量 1st

携程预订房型 1st | 美年年健康 2nd

阿里里聚安全 3rd | 中国网网络对抗 8th



298元的课程
会员免费学!

Riiid! Answer Correctness Prediction

-----回答准确性预测

	课程专题	知识点	时间	讲师
1	开营仪式	赛题内容介绍 Baseline代码讲解	10/17 19:00	Cookly
2	结构化数据传统建模	数据处理、特征加工 特征筛选、模型训练与验证	10/18 19:00	Cookly
3	结构化数据深度建模	ctr类模型、 类别特征处理、 连续特征处理	10/24 19:00	Cookly
4	中期直播答疑	解答比赛过程中的问题	10/25 19:00	Cookly
5	比赛相关Paper讲解	相似领域Paper讲解	10/31 19:00	Cookly
6	传统Baseline模型进阶	拆解赛题、特征进阶 模型优化、模型融合	11/1 19:00	Cookly
7	深度Baseline模型进阶	深度模型结构设计技巧 深度模型tick	11/7 19:00	Cookly
8	比赛复盘	比赛思路全复盘 优胜选手方案分享 整理知识点和上分点 比赛经验的面试展现技巧 干货分享	根据赛程和开源情况定期	Cookly



我们如何带大家学习这门课

Kaggle-回答准确性预测大赛（传统机器学习方向）

- 1、实战Kaggle-riid比赛（kaggle大数据平台），提供**Lgb和Nffm的baseline方案**
- 2、**详尽介绍赛题解析，建模过程**，分别了解**传统建模阶段和深度建模阶段**
- 3、针对Baseline方法如何进一步优化，帮助理解每一个trick后的意义
- 4、**Ctr类点点点击预解决方案**异同点详解
- 5、如何在下一场类似的比赛中快速的取得好成绩
- 6、如何将实战中知识用于工作（面试）中



原价298！限时248！
扫码即可优惠价购买
限量50张！



298元的课程
会员免费学！

——结 语——

唯比赛可抵卓越漫长



4、互动时间



原价1498比赛会员
现在立减100
仅限20张!



添加小享
获得Baseline&课件



deepshare.net

深度之眼

联系我们：

电话：18001992849

邮箱：service@deepshare.net

Q Q：2677693114



公众号



客服微信

