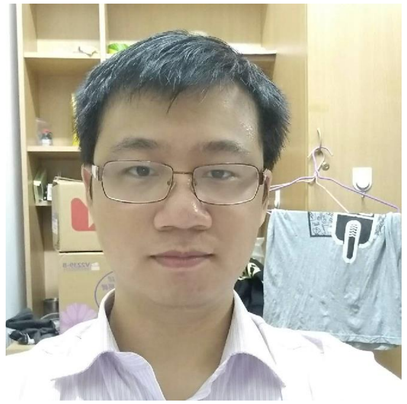
广告算法大赛

旁一百队

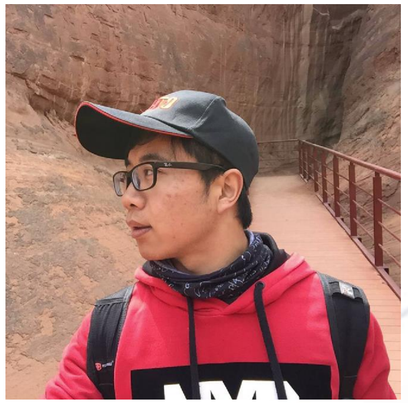
目录

团队介绍

王宇晟
北京航空航天大学
计算机学院
研二



李冈嶷
北京大学
信息科学技术学院
研二



包梦蛟
北京航空航天大学
计算机学院
研一

赛题回顾

赛题回顾

- 意义：相似人群扩展，基于广告主提供的一个种子人群，自动计算出相似的人群
- 问题转化：预测用户点击广告的可能性
- 赛题特征
  - 用户：用户ID、年龄、性别、学历、兴趣、消费能力等等
  - 广告：广告ID、广告主ID、推广计划ID、广告类目、商品类型等等
- 正负样本比例：1:20
- 数据规模：初赛约为800万条，复赛约为4500万条
- 评估指标：$\frac{1}{m}\sum_{i=1}^{m} AUC_i$

解决思路与算法

框架

数据处理

- 负采样
- 缺失值填充
- 低频过滤
- 等频归一化

特征构造

- 基础特征
- 一阶类别统计
- 交叉类别统计
- 非类别统计
- One-Hot
- TF-IDF

模型

- GBDT模型
- ResNFFM
- ResNFM

模型融合

- 自融合
- 不同Loss融合
- 不同模型融合

11729072

数据处理

负采样（GBDT）

- Why?
  - 数据量太大，95%以上都是负样本
  - 机器有限
  - AUC指标只和相对排序有关
- How?
  - 对负样本进行1/10负采样
- Result?
  - 节省模型训练速度，防止Memory error
  - 提高特征维度

11729072

数据处理

低频过滤

- 稀疏的离散特征会造成过拟合问题
- 统一替换出现次数在K次以下的特征值为“other”

等频归一化

- 不同特征的取值分布、相同维度特征值差异很大，一些特征长尾分布
- 对特征值排序，按照分位点进行等频分桶，分成100个桶

缺失值补充

- 对不同类型的数据填充不同类型的值

11729072

特征构造

基础特征

- 原始用户、广告特征

One-Hot

- 将类别特征离散化
- 广告ID、广告联盟ID、广告类别ID、产品ID、产品类型、地理位置、年龄等等

TF-IDF

- 将多个取值的特征向量化
- 如用户app安装记录、用户兴趣、用户主题、用户关键词等等

11729072

特征构造

类别/交叉类别统计特征

- 最高四折交叉
- 对uid及其交叉做贝叶斯平滑
- 五折交叉统计

用户类别

广告类别

用户类别 + 广告类别

用户类别A + 用户类别B

广告类别A + 广告类别B

...

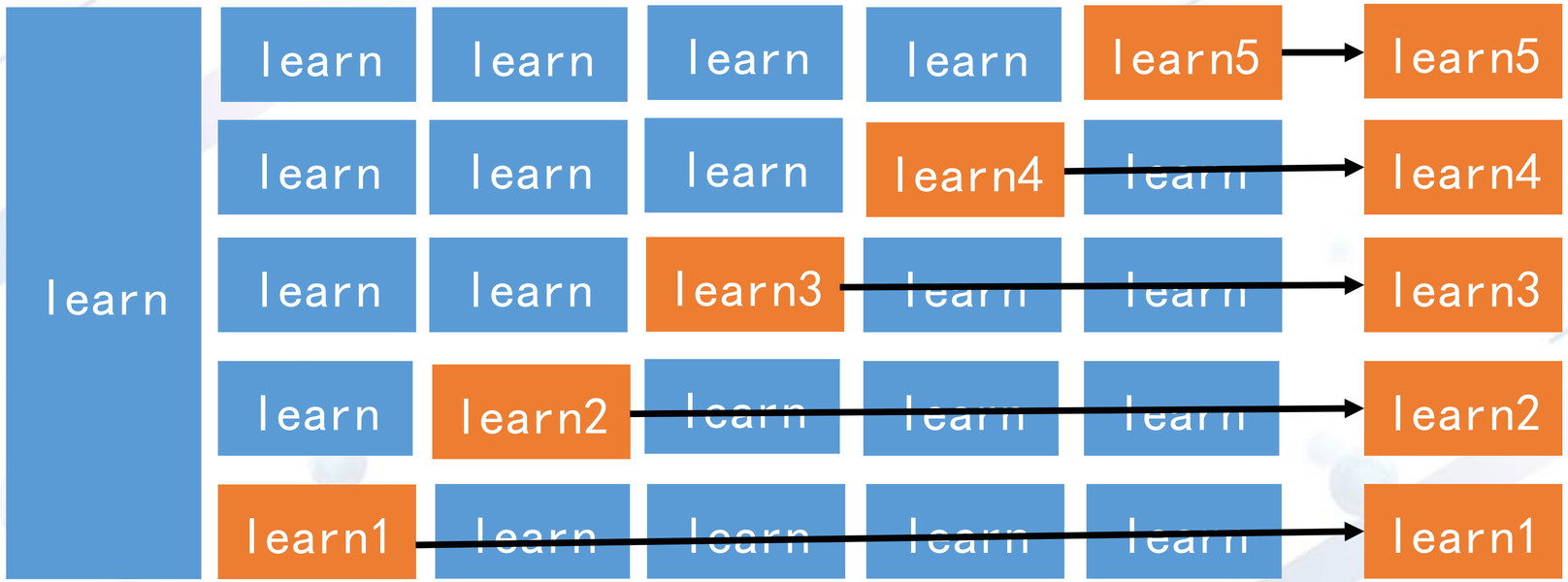
点击

展现

点击率

11729072

特征构造



val

test

五折交叉构造统计特征：

- 训练集每次用四份数据作为一份数据的统计
- 验证集和测试集使用所有数据作为统计

11729072

特征构造

非类别统计特征

- 针对一些有多个值的特征，例如用户兴趣、app安装记录、用户关键词等
- 假设kw2为“11 22 33”，表示用户在kw2下有11、22、33三个关键词
- 分别计算11、22、33的点击率，假设为t_{11}、 t_{22}、 t_{33}
- 特征值：$\log((1-t_{11})(1-t_{22})(1-t_{33}))$

11729072

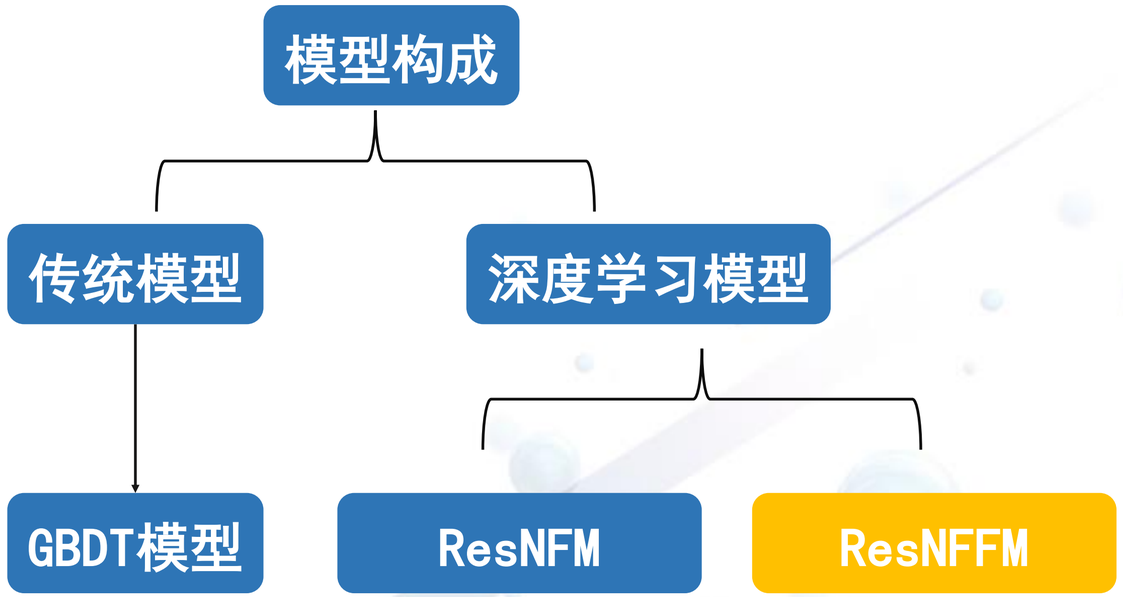
模型

排序模型的挑战：

- Memorazation
- Generalization

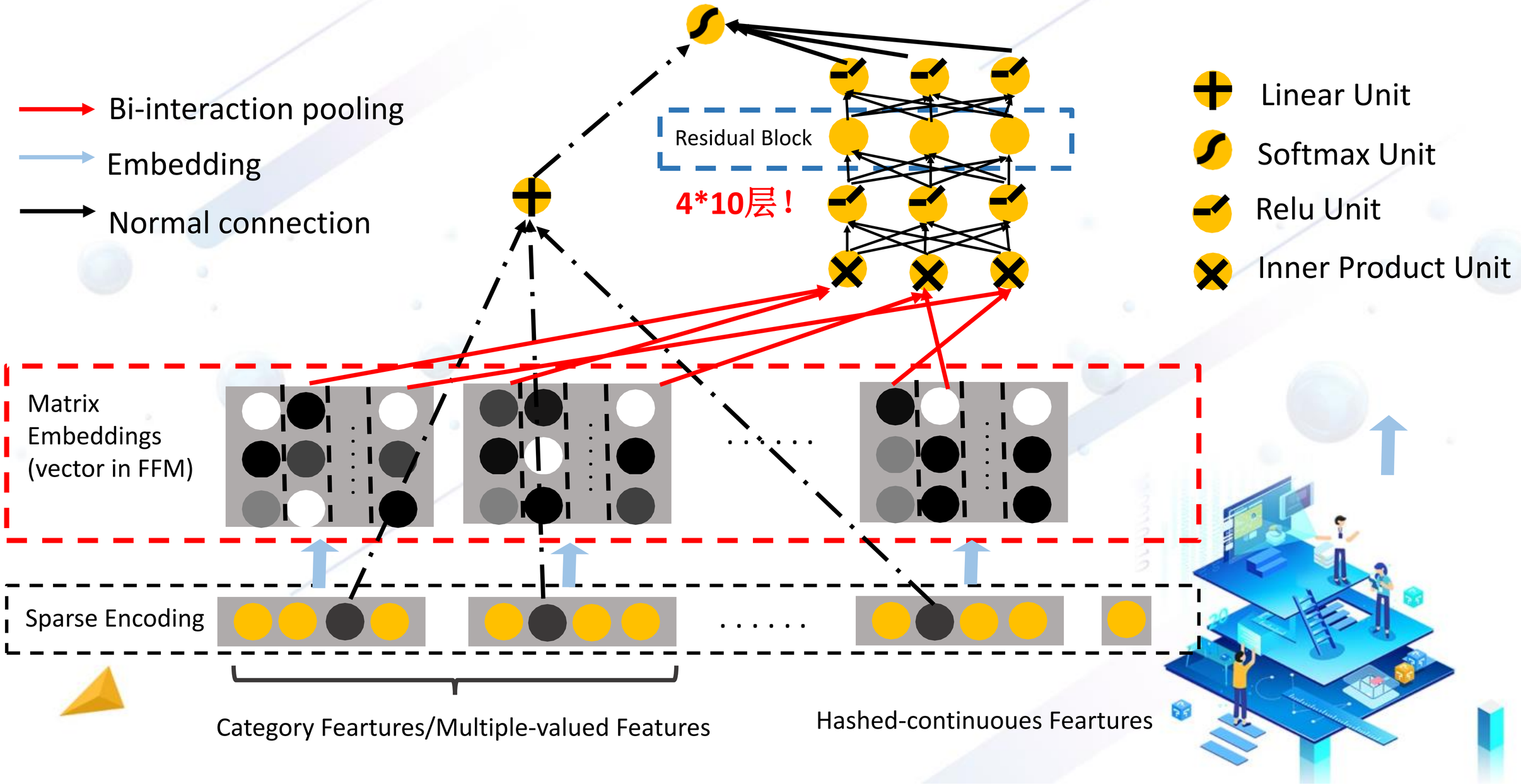
模型比较：

- 树模型Memorazation更强，记忆特征和标签相关特征组合能力强，因此在初赛小数据集上有很好的结果
- 深度模型Generalization更强，探索未出现的特征组合，在复赛大数据集上能取得更好的结果



11729072

模型改进 Deep Residual Networks on Field-aware Factorization Machine(ResNFFM)

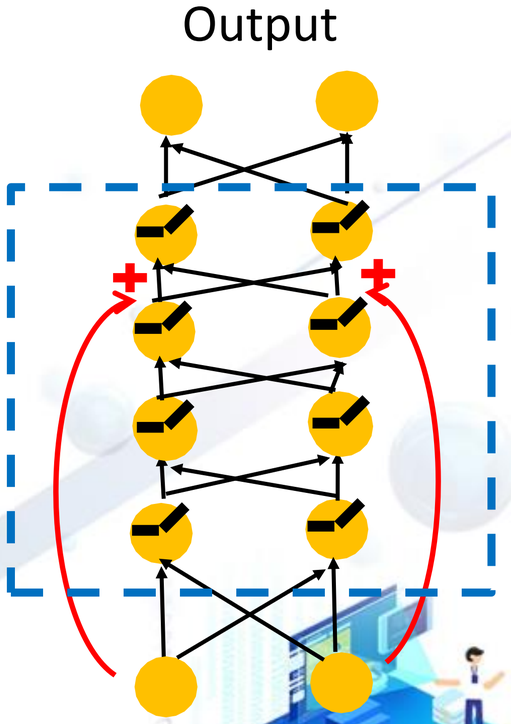


11729072

模型改进 Deep Residual Networks on Field-aware Factorization Machine(ResNFFM)

为什么要用ResNet结构？

- 沿着飞线，可以使得Embedding层更靠近Loss层，更好学得嵌入向量
- 使用ResNet结构捕捉更深层次的交叉
- 有更好的Generalization，探索训练集中很少出现的新特征组合



4*10层！

效果：

- 使用40层的网络，相同特征下比较NFFM提高3个千分点
- 加入十几维特征，单模型成绩为0.7768

$$y = f(x, \{W_i\}) + x$$

11729072

模型融合

同模型同Loss融合

- 一次训练中，将分数大于阈值的模型保留，取平均
- 节约模型训练时间

同模型不同Loss融合

- auc、auc_exp、auc_log、op_auc

不同特征模型融合

- 同模型输入不同特征
- 不同模型融合

思考和总结

11729072

思考

Wide or Deep？

- 初赛时，基本上top队伍的最优单模型都是树模型
- 数据量增加后，最优单模型基本上是NN结构
- 然而，在wide部分增加人工特征工程仍旧能提升模型上限…

更多数据?

- 在我们的实验中，迁移学习也取得了比较好的效果

个性化

- 用户id和其他特征的交叉强特征
- 在此基础上，设计个性化定制部分提升模型

11729072

总结

收获

- 在比赛中组建队伍、从一个人到三个人，收获友情
- 学习ctr预估知识，工业界实际大数据处理
- 跟着大佬们斗智斗勇…

遗憾

- 由于时间原因，特征没有完全加入到神经网络中
- 特征迭代效率较慢，需要更好的pipeline

Thanks