深度学习在搜狗无线搜索广告中的演化之路

搜狗无线搜索研发部　舒鹏

2016年11月

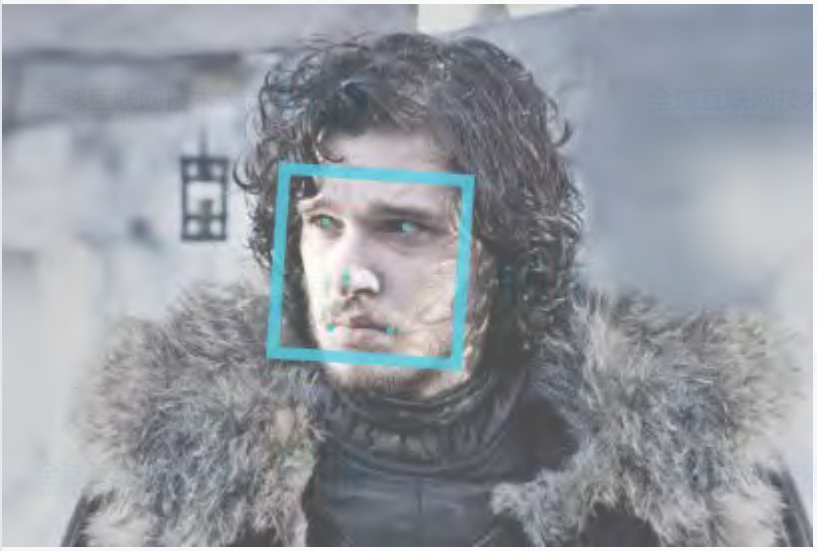
目录
CONTENTS

背景知识

语音识别

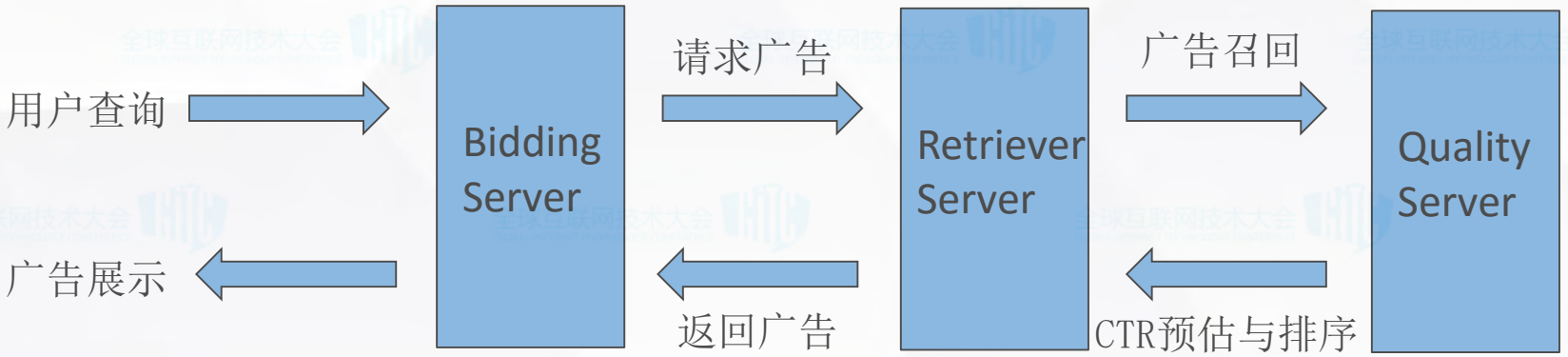
博弈

人脸识别



深度学习获得极大成功的应用领域

背景知识



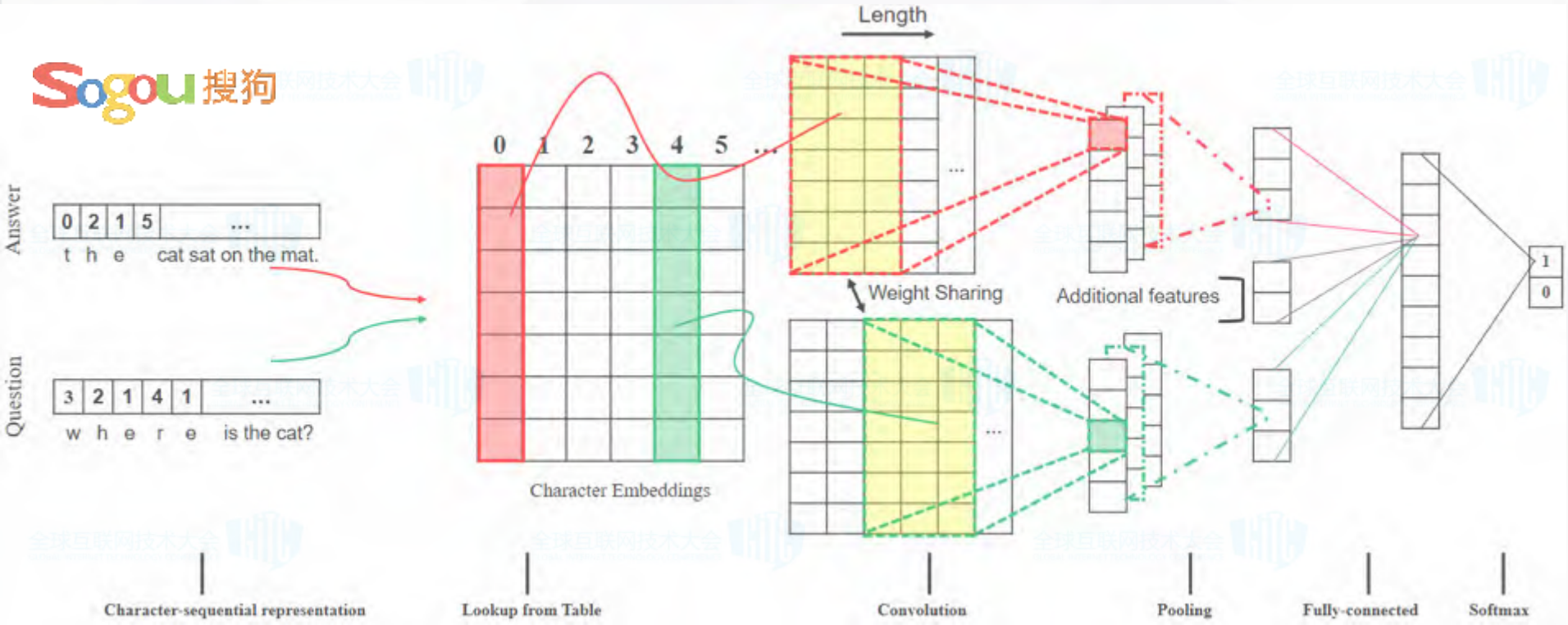
基本的搜索广告处理流程

重要环节	作用	关键点
广告召回（Retrieve）	选出和查询词最相关的广告，减少后续模块负载	文本相关性
广告排序（Rank）	预估点击率，和出价结合后按RPM排序	CTR预估

02

无需分词的问答系统设计

无需分词的问答系统设计



无需分词：基于字符粒度表达的问答系统设计

L.X Meng, Y.Li, M.Y Liu, P Shu. Skipping Word: A Character-Sequential Representation based Framework for Question Answering. In *Proceedings of The 25th ACM International Conference on Information and Knowledge Management(CIKM2016)*, pages 1869-1872, 2016. Sogou Inc

http://dl.acm.org/citation.cfm?id=2983861&CFID=859921406&CFTOKEN=71449114

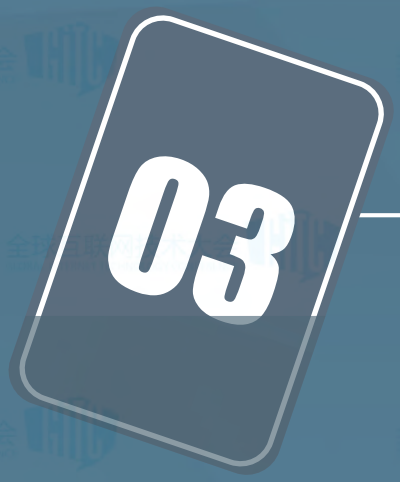
无需分词的问答系统设计

该算法同样可用于计算文本相关性，基于搜狗数据集的评测结果如下

Model	ACC	AUC
WE	Base	Base
CSR	+1%	**+0.3%**

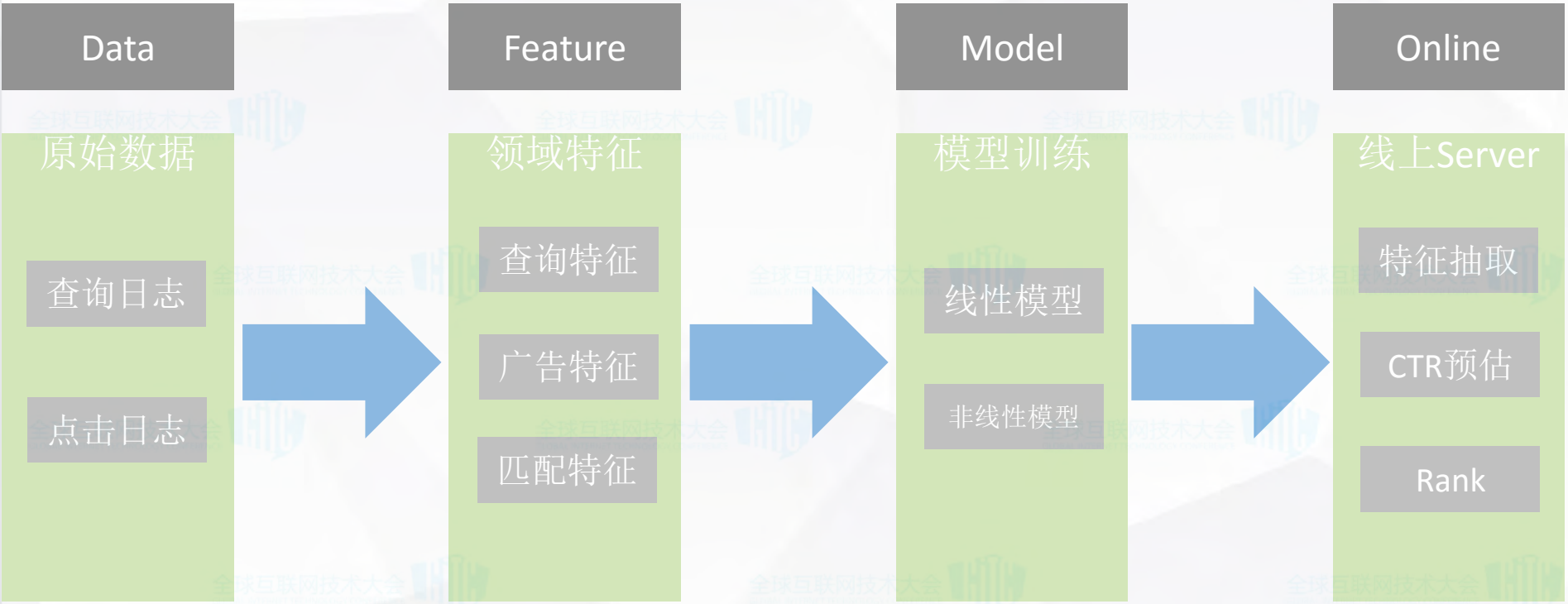
WE: Word Embedding
CSR: Character-Sequential Representation

相比传统的词向量算法，本算法效果持平或微涨，但无需分词，节省了计算步骤和内存耗用，更易于设计端到端的文本处理算法，具有一定的实用价值。



基于多模型融合的CTR预估

CTR预估流程


Data
Feature
Model
Online
原始数据
查询日志
点击日志
领域特征
查询特征
广告特征
匹配特征
模型训练
线性模型
非线性模型
线上Server
特征抽取
CTR预估
Rank

特征设计

离散特征

容易设计；刻画细致；特征稀疏；

特征量巨大；模型复杂度受限

连续特征

需要仔细设计；定长；特征稠密

特征量相对较小，可以使用多种模型训练

两者可以相互转换：DBN的使用

模型类别

非线性

- 能够学习特征间非线性关系
- 模型复杂、计算耗时
- DNN、GBDT

线性

- 简单、处理特征量大、稳定性好
- 需借助交叉特征
- Logistic Regression

模型融合

CTR bagging

- 将多个模型的输出CTR加权平均
- 实现方法简单，模型之间不产生耦合
- 可调参数有限，改进空间相对较小

模型融合

- 任一模型的输出作为另一模型的特征输入
- 实现方法复杂，模型之间有依赖关系
- 实验方案较多，改进空间较大

模型融合的工程实现

目标

- 可支持多个不同模型的加载和计算
- 可支持模型之间的交叉和CTR的bagging
- 可通过配置项随时调整模型融合方案
- 避免不必要的重复操作，减少时间复杂度

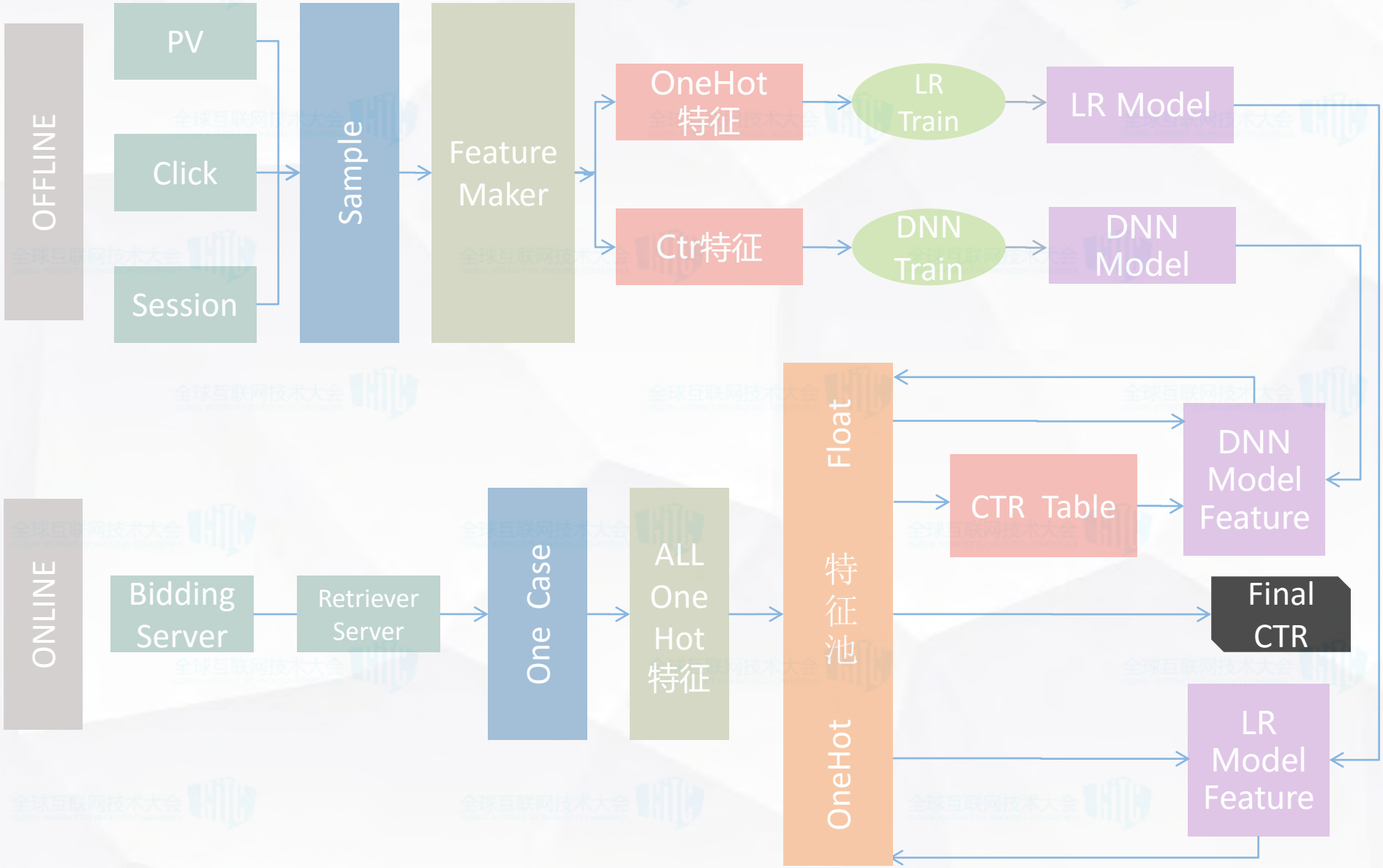
解决方案（引入ModelFeature的概念）

- 模型本身也看做一个抽象特征
- 模型特征依赖于其它特征，通过计算得到新的特征
- 模型特征输出可作为CTR，也可作为特征为其它模型使用
- 限定ModelFeature的计算顺序，即可实现bagging/模型交叉等功能

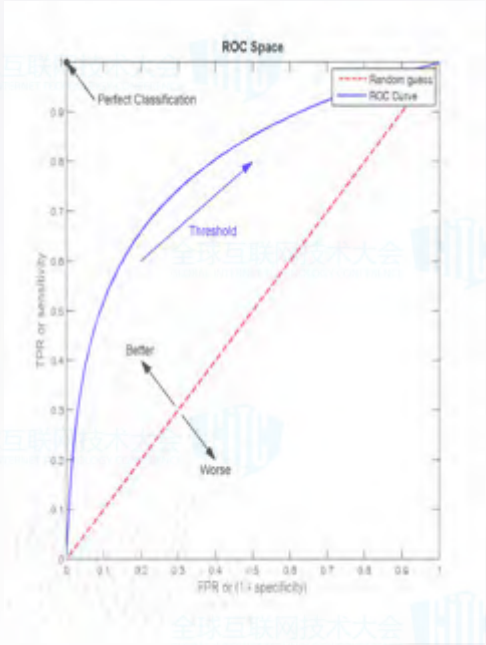
关键点

- 数据一致性
- 流程稳定性

模型融合


OFFLINE
PV
Click
Session
Sample
Feature Maker
OneHot 特征
LR Train
LR Model
Ctr特征
DNN Train
DNN Model
ONLINE
Bidding Server
Retriever Server
One Case
ALL One Hot 特征
Float
特征池
OneHot
CTR Table
DNN Model Feature
Final CTR
LR Model Feature

模型效果的评估



AUC

上线收益

Survival Bias
特征覆盖率

是否一致？

并行化训练

诉求

- 加大数据量，提升模型稳定性
- 加大数据量，提升模型收益

并行化训练

方案

- MxNet支持多机多卡, 使用成本低
- 构建多机多卡GPU集群，优化训练效率，提高加速比

04

若干思考

若干思考

1 DL的强项

输入不规整
结果确定
容易获取的海量训练数据

2 CTR预估

特征有明确含义
场景相关，以用户为导向
很难界定“Ground Truth”
训练样本“有限”

3 方向

特定业务场景
模型融合
提升效率，降低成本

THANKS YOU !

Q&A