GRAB
百度推荐广告
生成式排序模型技术实践

演讲人：陈少鹏 | 百度资深工程师

微信公众号：百度商业技术
干货满满欢迎大家关注！

目录

CONTENTS

01

推荐大模型化趋势

Large Recommendation Model

推荐大模型化趋势 | 百度传统广告模型瓶颈

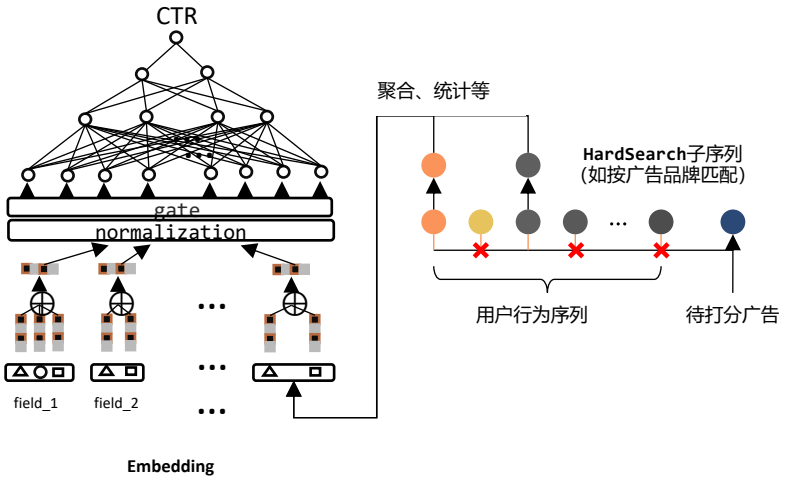
传统范式DLRMs：大规模离散特征+MLP

1、特征规模已至上限，难以迭代

2、序列表征有压缩，建模损失大

3、离散特征刻画，强记忆，弱推理

4、广告变化快，ID记忆下，序列建模时历史行为、历史ID的激活率比较低



推荐大模型化趋势 | 百度传统广告模型瓶颈

DLRMs难以Scaling！大模型的Scaling Law成为新希望！

传统范式DLRI

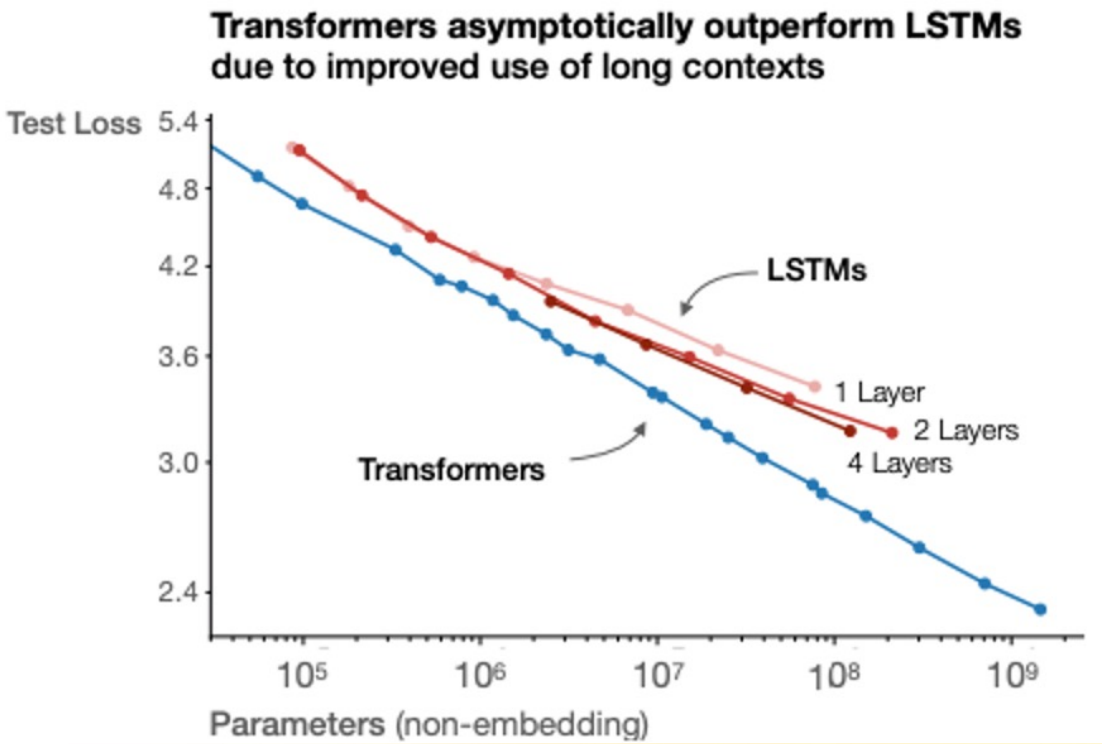
1、特征规模已

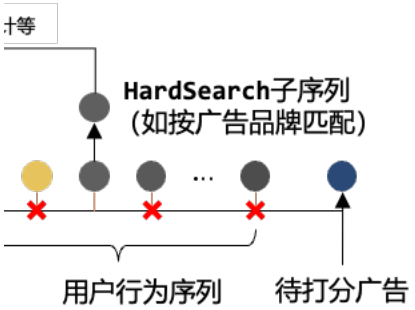
2、序列表征有

3、离散特征刻

4、广告变化快

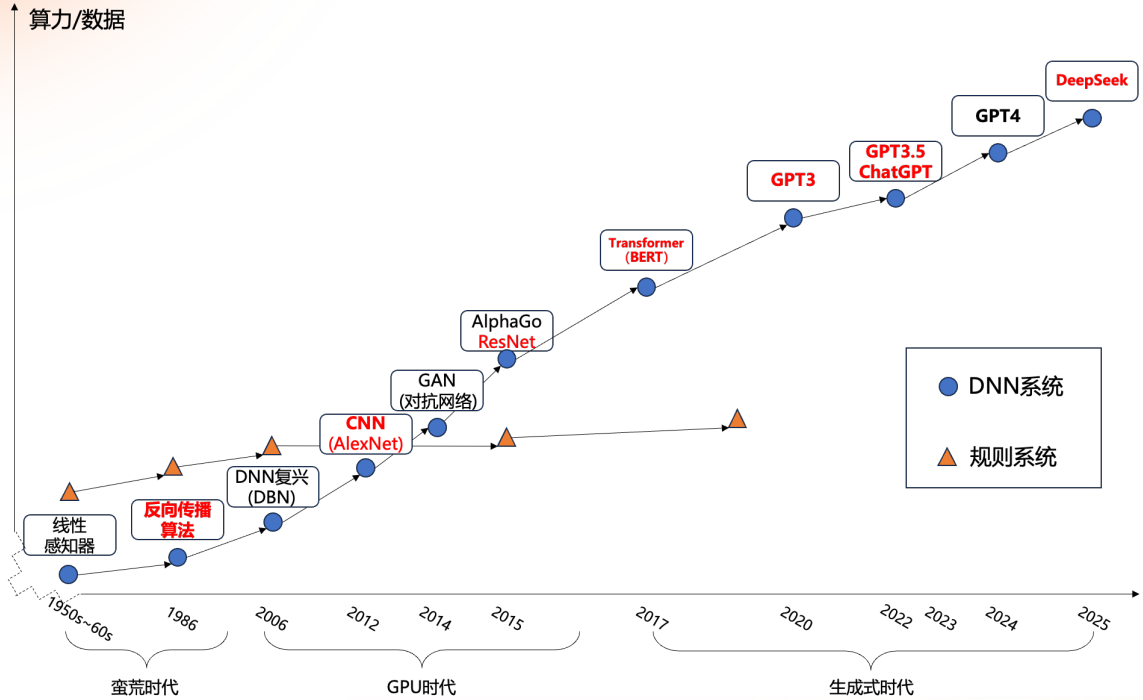
行为、历史ID的





推荐大模型化趋势 | 再谈LLM的Scaling Law(i)

数学原理在牛顿时期已经形成，在300年后的今天LLM发挥出价值



- **数据供给**：数字信息时代、互联网时代
- **算力跃升**：并行计算、异构硬件
- **算法创新**：
  - BP算法（训练）
  - ResNet（深度）
  - Transformer（序列+并行）

推荐大模型化趋势 | 再谈LLM的Scaling Law(i)

数学原理在牛顿时期已经形成，在300年后的今天LLM发挥出价值

- **数据供给**：数字信息时代、互联网时代
- **算力跃升**：并行计算、异构硬件
- **算法创新**：
  - BP算法（训练）
  - ResNet（深度）
  - Transformer（序列+并行）

推荐大模型化趋势 | 再谈LLM的Scaling Law(ii)

DataFun. | 百度商业技术

大模型的Scaling Law在百度推荐场景复现——推荐大模型化

1、**大语言模型推荐**：借用LLM的记忆、推理能力，场景数据适配到LLM

2、**大语言模型表征**：使用LLM的表征增强推荐模型，提升表征、泛化能力

3、**生成式序列化建模**：借用LLM上对于Transformer、 Long-context等序列化建模技术，在推荐场景进行改造应用

推荐大模型化趋势 | 再谈LLM的Scaling Law(ii)

DataFun. | 百度商业技术

大模型的Scaling Law在百度推荐场景复现——推荐大模型化

1、**大语言模型推荐**：借用LLM的记忆、推理能力，场景数据适配到LLM

指标差了在百分位以上

直接迁移广告数据到通用LLM，缺失离散特征、模型能力限制、LLM与本场景差异大

2、**大语言模型表征**：使用LLM的表征增强推荐模型，提升表征、泛化能力

可以提升泛化性、补充多模态理解能力

但偏长期，且价值受长尾、冷启用户的商业价值限制

3、**生成式序列化建模**：借用LLM上对于Transformer、Long-context等序列化建模技术，在推荐场景进行改造应用

Meta提出GRs，革新推荐模型范式。百度商业已于2025年4月份推全GRAB方案

按照用户序列来获得推荐模型的Scaling能力，得到了广泛验证

01

GRAB整体设计

Design of GRAB

GRAB整体设计 | 设计理念&目标

1、**“历史序列与目标行为”分离到统一：**统一表征历史序列与目标行为，在同一空间进行行为建模；

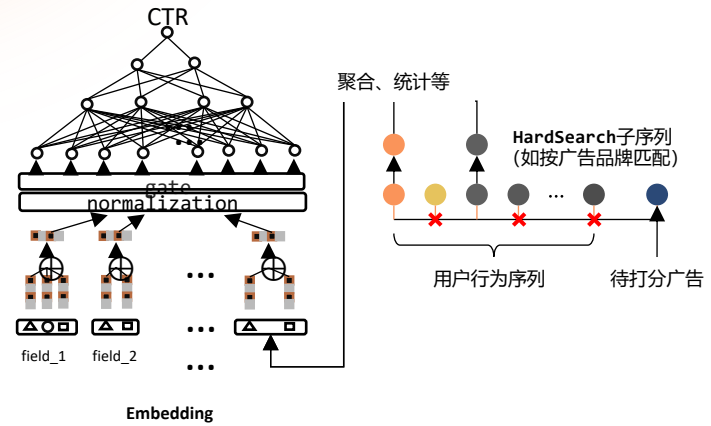
2、**“平铺”到“结构化”：**将用户行为表征重构为结构化表征，实现不定长、不定宽、层级化等的序列信号输入；

3、**“手动设计”到“自适应”：**用户序列直接输入，实现自适应的信息输入，避免人工设计的特征带来的信息折损；

4、**“高效注意力”替代序列检索：**通过高效的注意力机制，替代序列检索，并实现全域、全周期的用户序列建模。

GRAB整体设计 | 整体框架

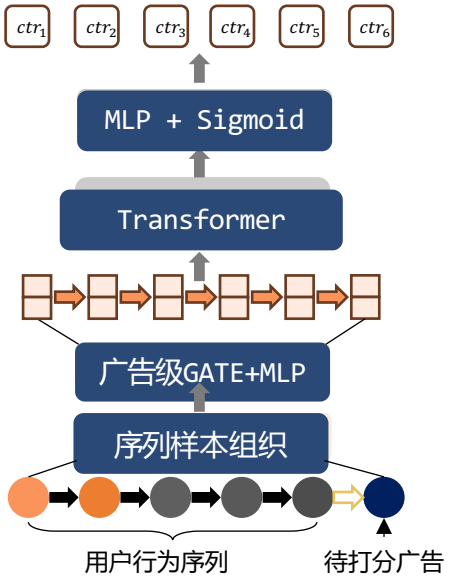
旧范式：单次行为预估点击率



SIM/ETA/TWIN等传统序列化建模技术

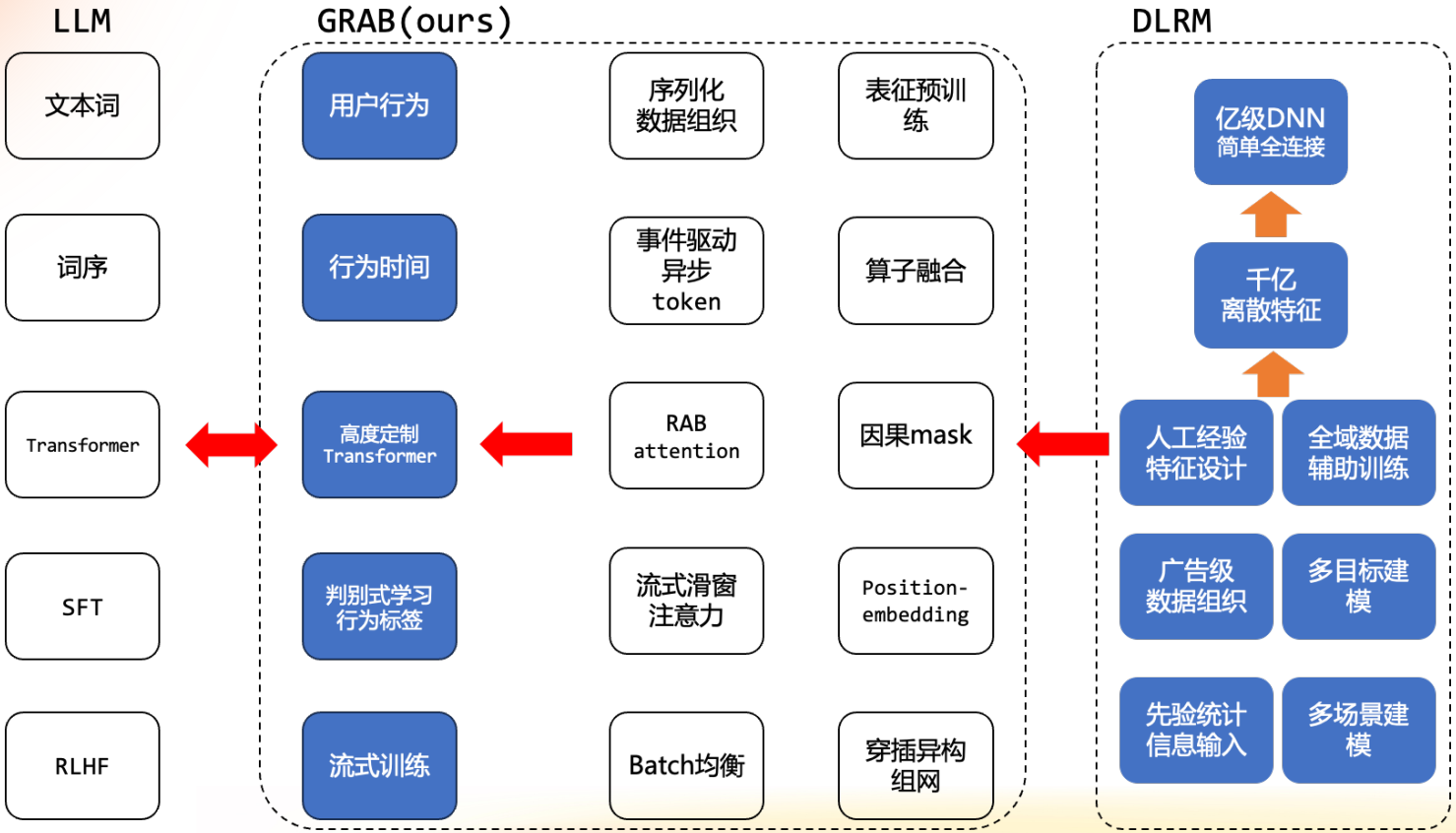
新范式： GRAB-百度商业生成式推荐排序模型

Generative Ranking for Ads at Baidu



- 序列化点击率预估 **端到端序列化建模**
- 交互感知因果注意力 **生成式用户序列推理**
- 序列与目标同序列组织 **行为目标统一序列**
- 曝光序列广告全信息 **输入信息无损**

GRAB整体设计 | 技术对比



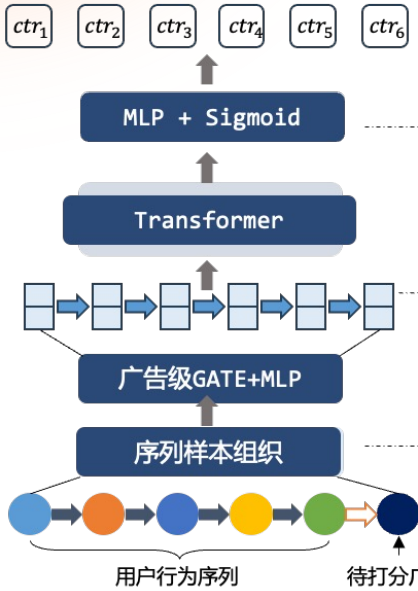
01

挑战及解法

Challenges and Solutions

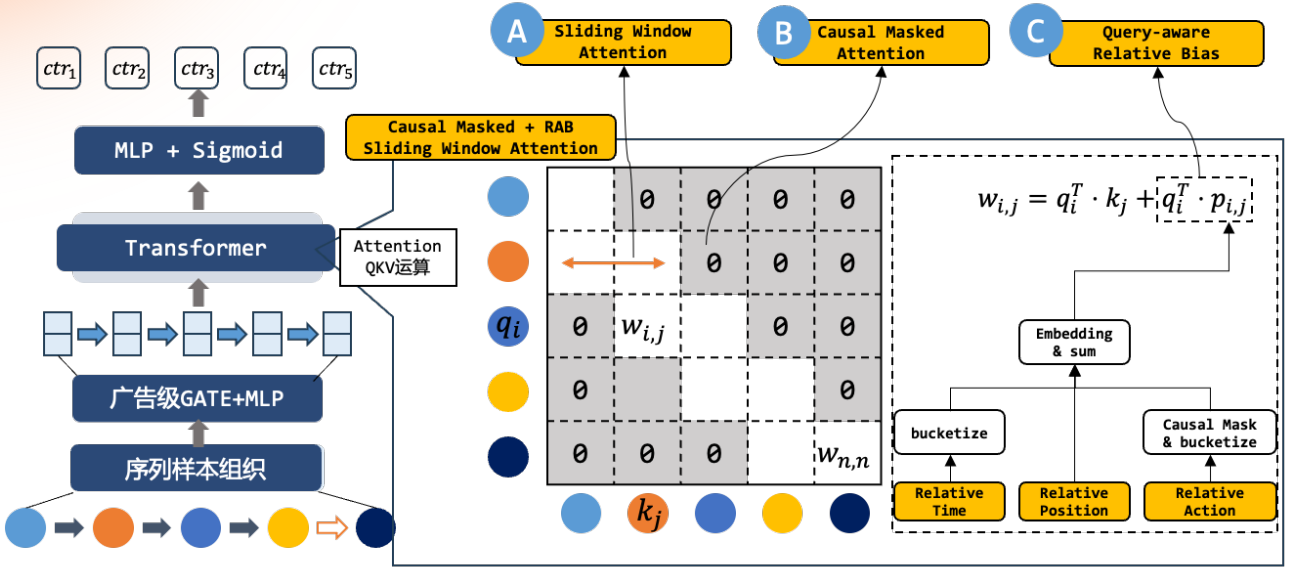
挑战及解法｜注意力、训练、表征、推理、系统

DataFun. | 百度商业技术



设计目标	路径拆解	8个挑战	解法
推荐广告 生成式排序模型	**推荐系统 生成式推理**	◆ 在线推理成本**阶次上涨** ◆ 推荐系统整体生成式序列化重构	✓ **KV-Cache、M-Falcon** ✓ 系统序列化重构改造，行为实时化感知
生成式序列 生成式用户序列推理	**用户行为 注意力机制**	◆ 面向推荐广告用户序列推理的**高效注意力机制** ◆ 生成式框架兼容推荐模型流式训练	✓ **Q-Aware RAB因果注意力** ✓ **双滑窗**注意力机制
行为表征 行为目标统一序列	**端到端 序列化训练**	◆ 序列化组织样本训练，成本/性能**+1个数量级** ◆ 序列化端到端训练，模型严重过拟合	✓ **变长序列零冗余**算法 ✓ **Sequence then Sparse**训练算法
离散特征 输入信息无损	**离散特征 兼容继承**	◆ 老汤模型纯冷启**完全打不平** ◆ 差异化捕捉用户序列中的变与不变化	✓ **双Loss 机制+离散**特征热启 ✓ **异构Token**表达长期属性

挑战及解法 | 注意力设计



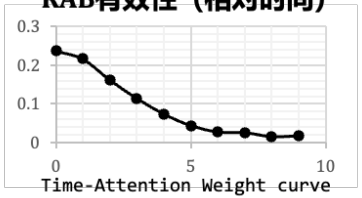
Naive Relative Bias

$$w_{i,j} = \mathrm{q_i^T} \cdot \mathrm{k_j} + \mathrm{Pos}_{i,j} + \mathrm{Action}_{i,j} + \mathrm{Time}_{i,j}$$

Query-aware Relative Attention Bias

$$w_{i,j} = q_i^T \cdot (k_j + Pos_{i,j} + Action_{i,j} + Time_{i,j})$$

RAB有效性（相对时间）

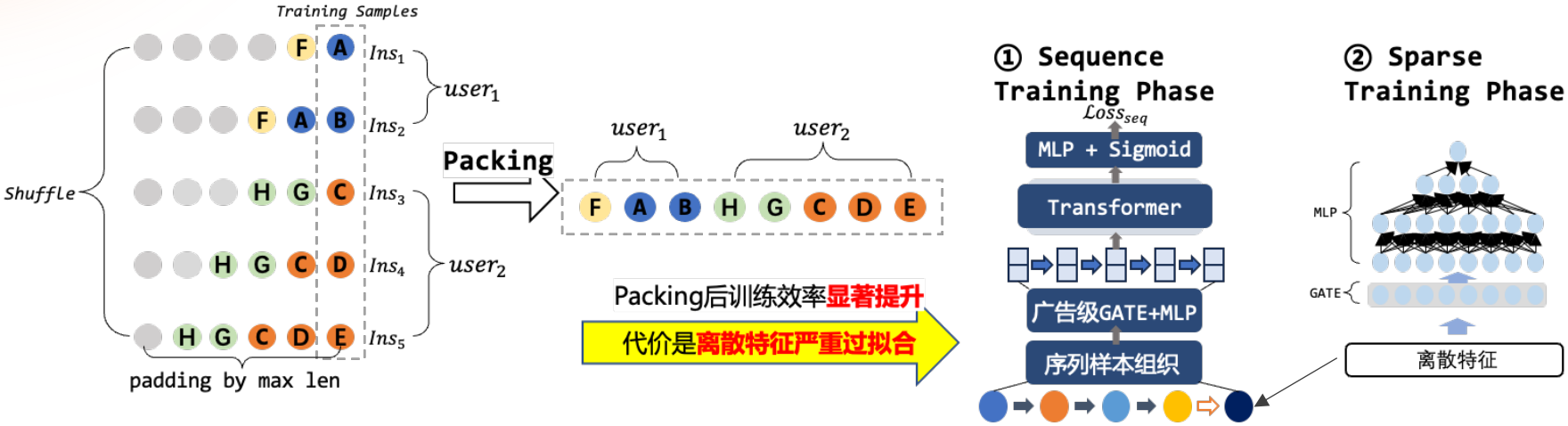


[1] 预算查表技巧： $RAB_{i,j} = \mathrm{q_i^T} \cdot P_{i,j} = q_i^T \cdot \mathbb{E}^P[\mathrm{p}_{i,j}] = (q_i^T \cdot \mathbb{E}^P)[\mathrm{p}_{i,j}]$ ， $\mathbb{E}$为P对应Emb Weight

挑战及解法 | 训练算法

同用户前缀共享，跨用户序列拼接

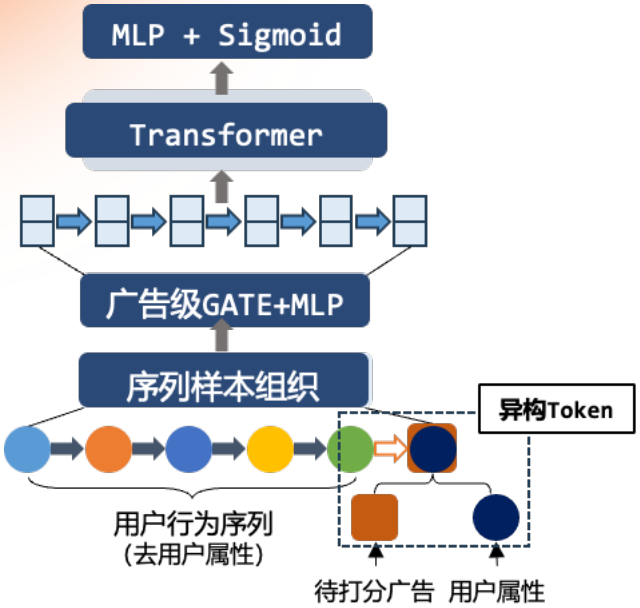
同序列内行为强相似，分布严重倾斜



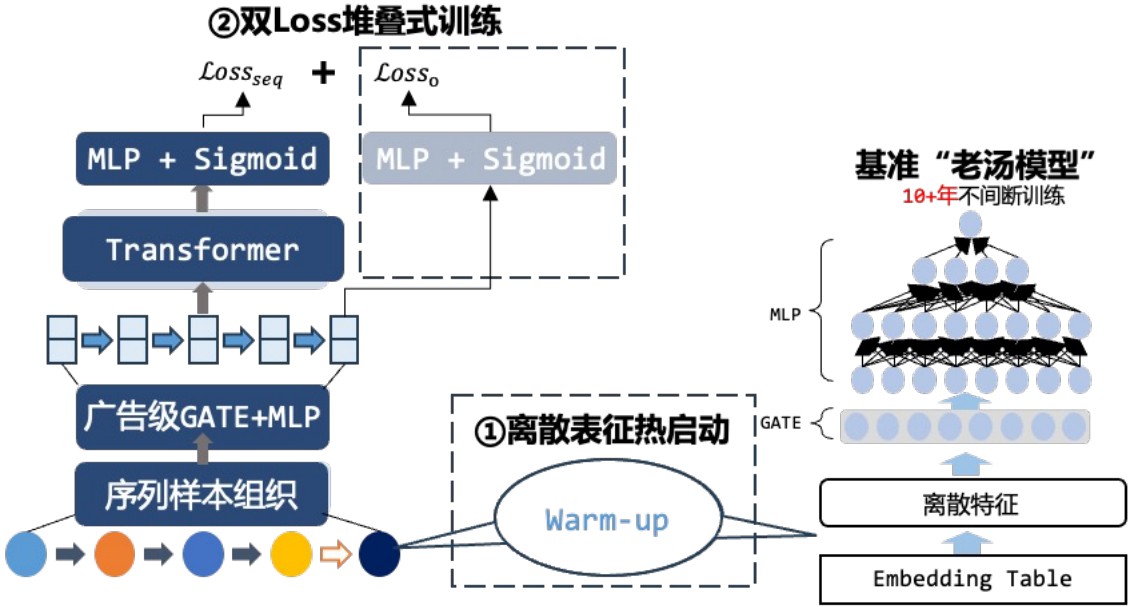
变长序列零冗余算法

STS(Sequence then Sparse)两阶段训练方式

挑战及解法｜表征——异构Token&老汤热启

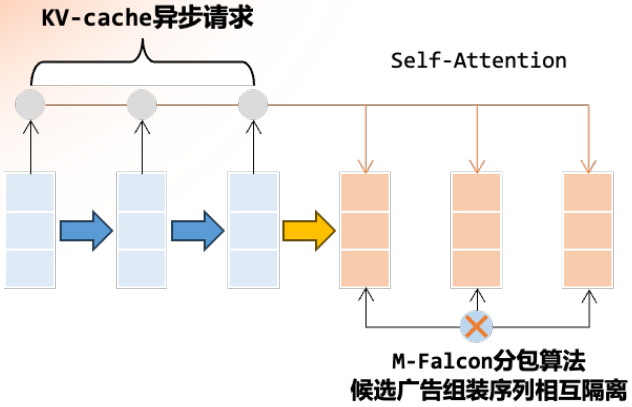


异构Token：特异刻画用户不变属性

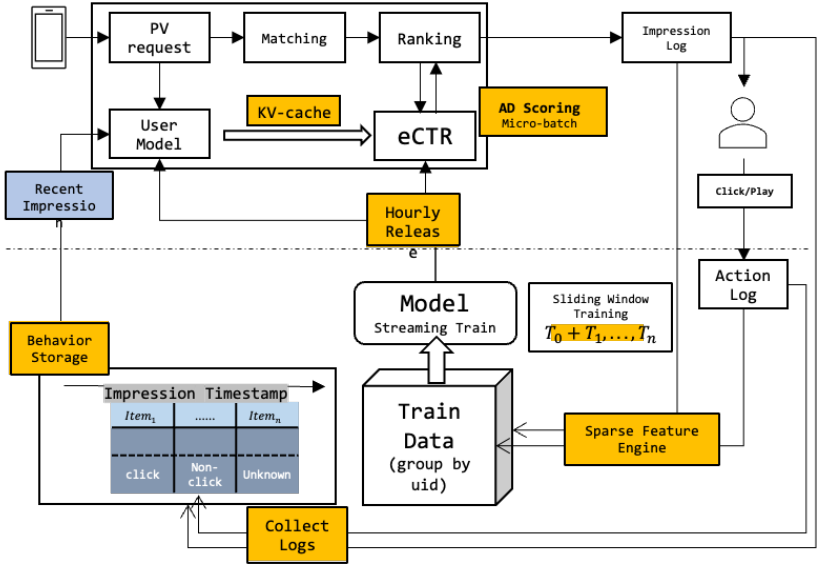


1、**热启表征**：离散及表征层热启自基准
2、**堆叠训练**：收敛前保留原组网及$Loss_o$，之后逐步去掉

挑战及解法 | 推理及系统



1、单向因果注意力，KV无后效性
2、Mask机制隔离候选AD间注意力
3、算子融合、低精度等



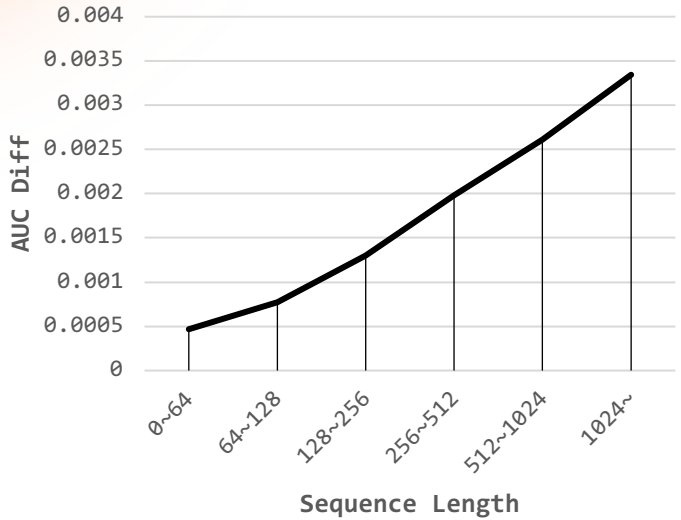
1、序列时效性做到实时
2、通过Mask机制对齐保证离在线序列一致性

总结及展望

Conclusion And Future Work

总结及展望(i) | GRAB收益及长期展望

落地百度推荐广告CTR
AUC ~ 3千分位



GRAB架构刻画能力：
序列长度每提升一个阶次，**模型指标获得线性增长。**

工作展望：长序列、全域、用户模型基座

用户
超长序列

全域
序列刻画

预训练+实时微调+偏好对齐

用户大模型基座

总结及展望(ii) | 推荐大模型长期思考

假如不考虑资源成本，目前能想象的推荐系统应是什么？

观点：

1、**广域知识储备、推演能力**

- 具备基本的世界知识
- 理解概念以及掌握一般规律，有推理演绎能力

2、**推荐域知识全覆盖**

- 内容全模态输入（图文、视频、声音等多模态）
- 交互行为全时空刻画（电影式行为回放记忆）

3、**快速学习适配能力**

- 可以无限细分刻画每一个用户、每一个细微场景，保证无损
- 能够快速响应适配输入变化（客户投放变化、流行趋势等）

总结及展望(ii) | 推荐大模型长期思考

假如不考虑资源成本，目前能想象的推荐系统应是什么？

观点：

1、**广域知识储备、推演能力**

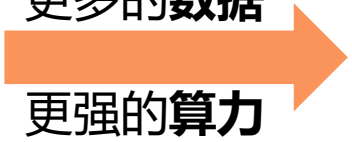
- 具备基本的世界知识
- 理解概念以及掌握一般规律，有推理演绎能力

2、**推荐域知识全覆盖**

- 内容全模态输入（图文、视频、声音等多模态）
- 交互行为全时空刻画（电影式行为回放记忆）

3、**快速学习适配能力**

- 可以无限细分刻画每一个用户、每一个细微场景，保证无损
- 能够快速响应适配输入变化（客户投放变化、流行趋势等）



规则系统->混合系统->生成式混合系统->生成式系统

Q&A

演讲人：陈少鹏 | 百度资深工程师

微信公众号：百度商业技术
干货满满欢迎大家关注！