

# 搜索排序

## 精排模型迭代路线图

2020.8 基本特征、训练 数据积累阶段 2020.10 有一定的点击样本, 关注用户兴趣变化 2020.12 样本量足够,使用id 类特征做复杂交互 2021.02 模型复杂度权衡、 学习的有效性? 2021.07 关注用户多兴趣 建模、CVR探索

DeepFM (CTR)

DeepFM+DIN (CTR)

DeepFM+TransFormer (CTR)

DeepFM
TransFormer?
CAN?
(CTR)

ESMM?
MMOE?
PLE?
(CTR-CVR)
LTR?

- 特征交叉保证模型表。达能力
- · 适合搜索起步阶段, 具备一定的个性化能 力

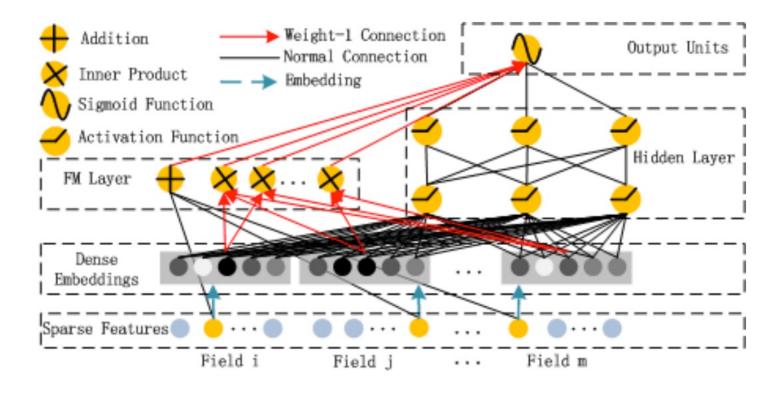
类目维度行为行为 用户行为序列进一 步提升搜索个性化 程度

- 使用稀疏id类特征细化 特征表达
- · <u>Encoder关注序列内部</u> 交互
- 模型表达能力明显提升
- id embedding有效性
- 线上推理速度

多兴趣用户模型 探索多任务学习模型,考 虑CTR和CVR联合学习

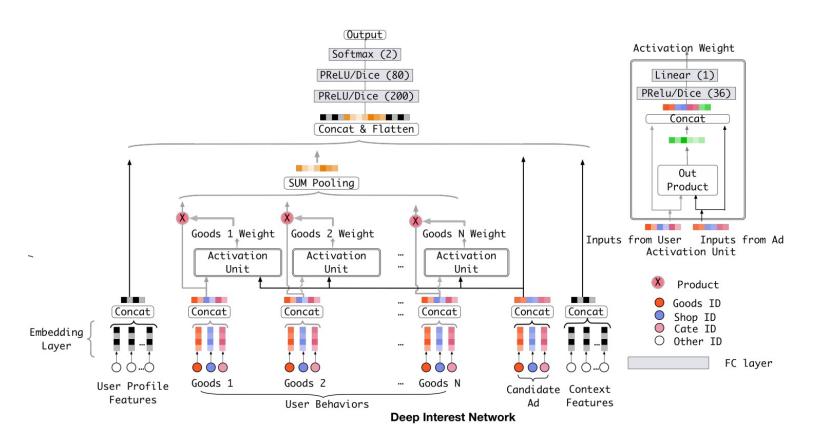
CTR、CVR、GPV多目标 导向

# **DeepFM**



- > 训练数据量少
- ▶ 用户行为不丰富
- > Base model

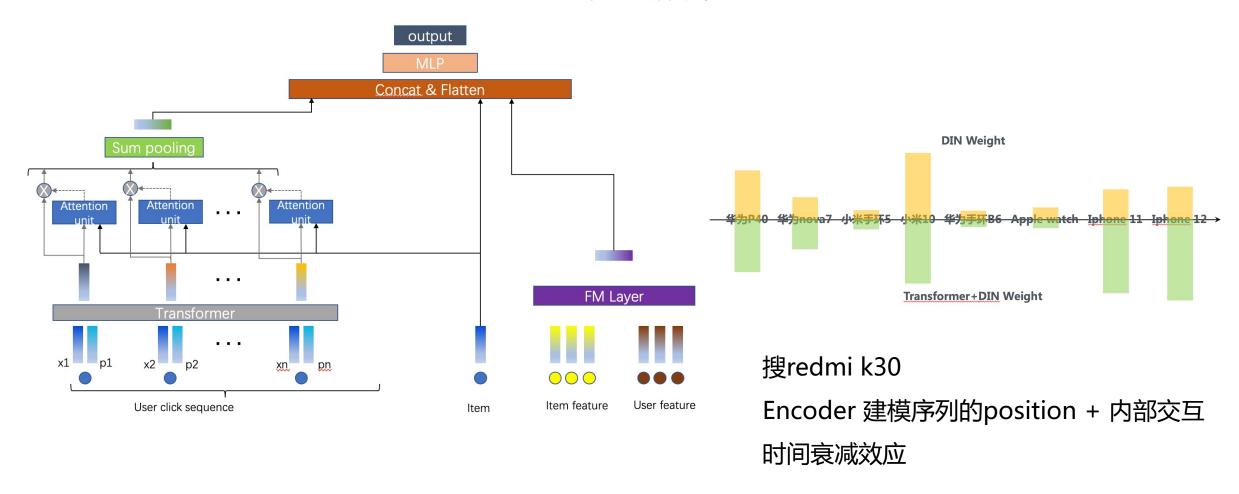
#### DIN



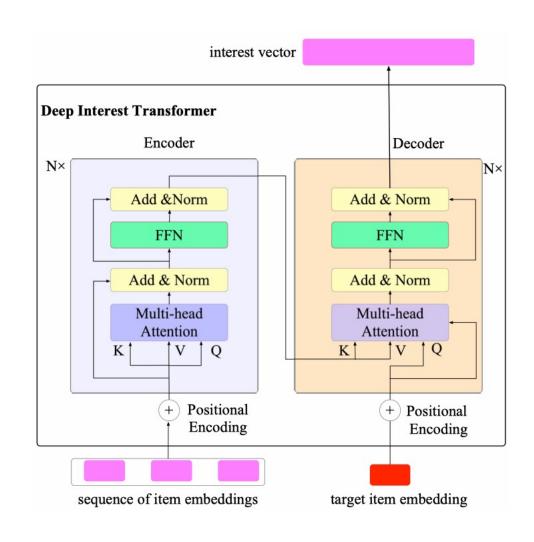
- > 三级类目序列 -> 商品id序列
- > 行为序列有丰富的用户兴趣表达
- > pvctr + 9.81%

#### **TransFormer**

#### 序列内部的互相影响? With encoder

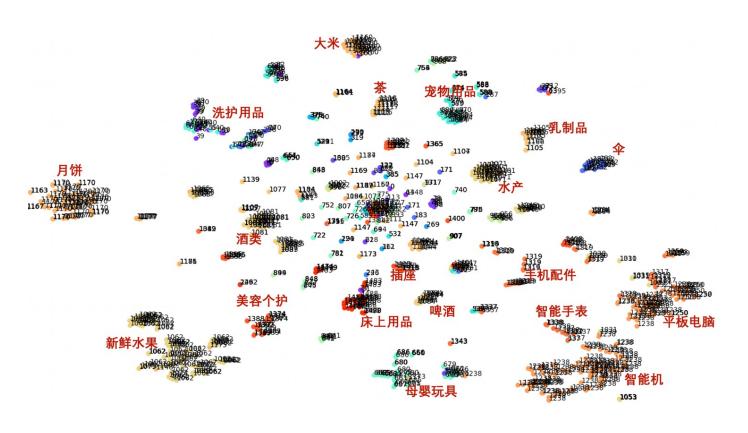


#### **TransFormer**



- ▶ 序列对目标商品的交互?
  - ✓ Encoder + decoder
- ▶ 不同序列类型?
  - 1. 点击 短期偏好
  - 2. 成交、收藏 长期偏好

## **TransFormer**

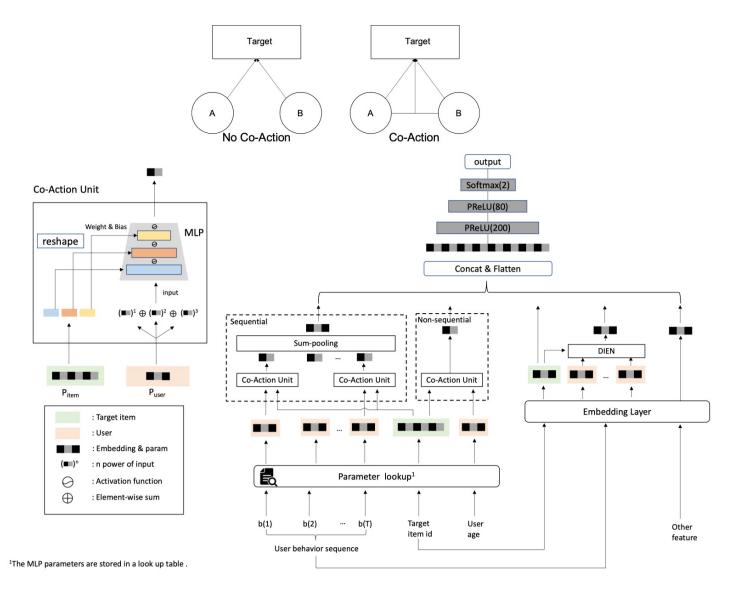


id embedding 有效性?

DeepFM + Encoder + Decoder pvctr + 5.51%

64d embedding TSNE

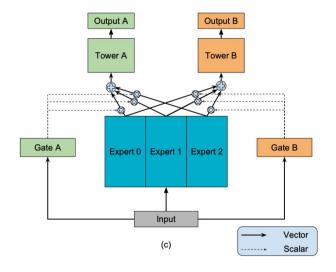
#### CAN

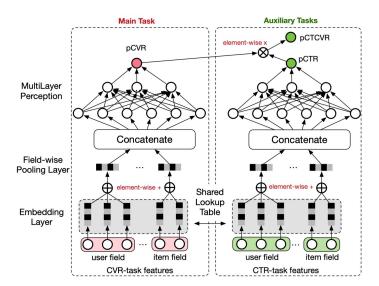




- ➤ DIN类模型对序列的aggregation只 是做了标量的修正
- ➤ 笛卡尔积作为新特征 (数量爆炸)
- ➤ Co-action特征独立学习,自由度更高,类似一个容量巨大的记忆模型
- ➤ CAN折中,隐式交互,一端作为输入、 一端作为MLP参数

# 多目标建模





- ▶ 样本不均衡,点击样本1.8%,购买样本0.1%
- ➤ CVR预估的问题
  - SSB问题,点击数据作为训练集,点击转化为 正样本、点击未转化为负样本,inference面 对整个曝光样本空间
  - 点击样本稀少,训练不充分
- > 基本思路
  - MMOE / PLE + loss Reweight
  - ESMM 预估pctr和pctcvr=pctr\*pcvr,面对整个曝光样本空间,隐式得到pcvr
  - 线上打分公式调整 ctr \* sigmoidCvr

$$sigmoidCvr = rac{1}{alpha + e^{-beta*cvr}}$$

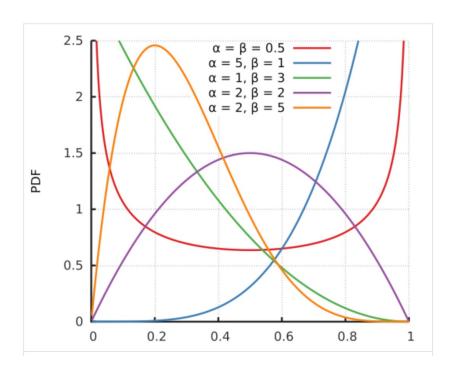
#### 新品冷启排序

#### **Bandit**

#### 模型打分 Bandit打分



- > 新品特征不足,模型难以做到准确的预估
- ➤ 强插保证曝光,独立排序,EE策略,交给用户决策
- ➤ 汤普森采样, beta分布, alpha=收益 beta=尝试
- ➤ 一个新品作为一个臂, alpha=F(点击, 购买), beta=F(曝光, 点击, 购买), beta分布结果作为得分



## 后期规划

- > 特征精细化
- point wise -> pair wise / list wise
- ▶ 重排模型化
- ▶ 与平台目标做更紧密的结合 -> 分人群 / 分商品