Alibaba Group 阿里巴巴集团

粗排技术体系与最新进展

王哲

阿里定向广告算法团队

大纲

- ✓ 粗排发展历史
- ✓ 粗排最新进展
- ✓ 总结与展望

Alibaba Group
阿里巴巴集团

粗排发展历史

背景介绍

✓ 大型工业排序系统一般采用多阶段级联架构，包含：

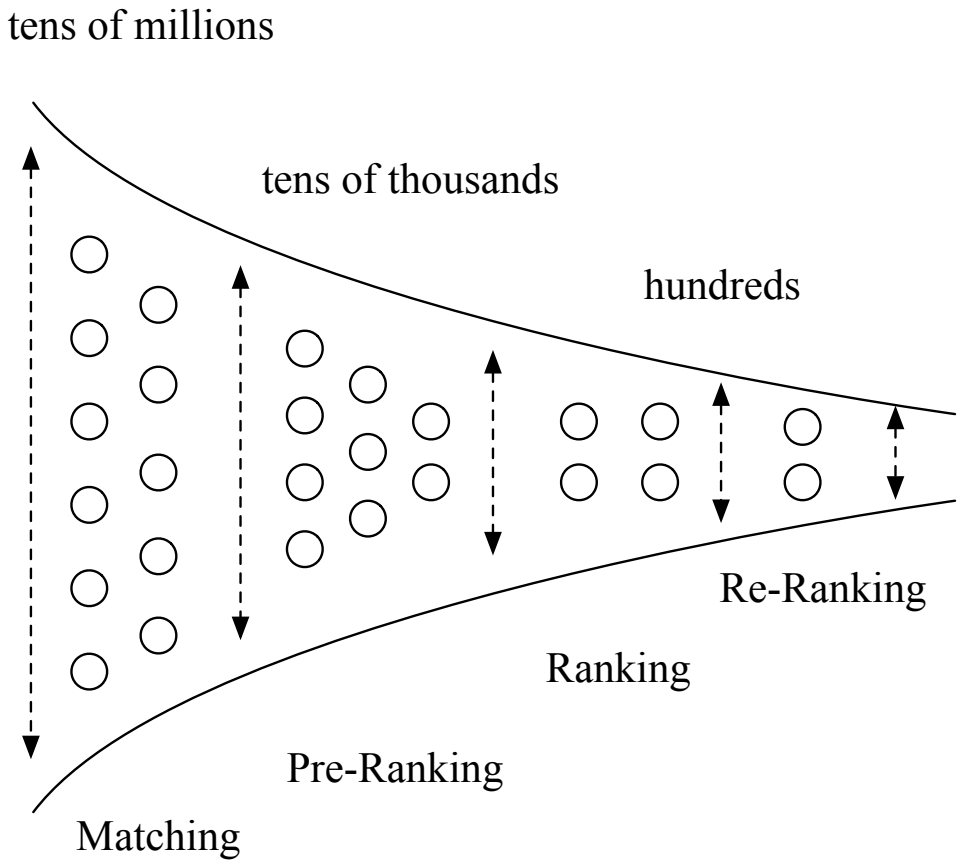
- 召回：1000W+
- 粗排：1W+
- 精排：上百
- 重排：上百

✓ 粗排目标：

- 在满足算力rt约束的情况下，选出满足后链路需求的集合。

✓ 粗排与精排的比较：

- 算力rt约束：粗排打分量远高于精排，同时有较严格的延迟约束：10-20ms
- 解空间问题：粗排线上打分的候选集更大，面临更严重的选择偏差问题。



粗排的两大技术路线

集合选择技术

- ✓ 以集合为建模目标，选出满足后链路需求的集合
- ✓ 依赖对后链路的学习，可控性较弱
- ✓ 算力消耗一般较小
- ✓ 代表技术：
  - 多通道
  - Listwise，如LambdaMART
  - 序列生成算法
    - 集合评估器
    - 集合生成器

精准值预估技术

- ✓ 以值为建模目标，直接对最终系统目标进行精确值预估
- ✓ 可控性更强
- ✓ 算力消耗一般较大
- ✓ 代表技术：
  - Pointwise

粗排的前深度学习时代（2016年以前）

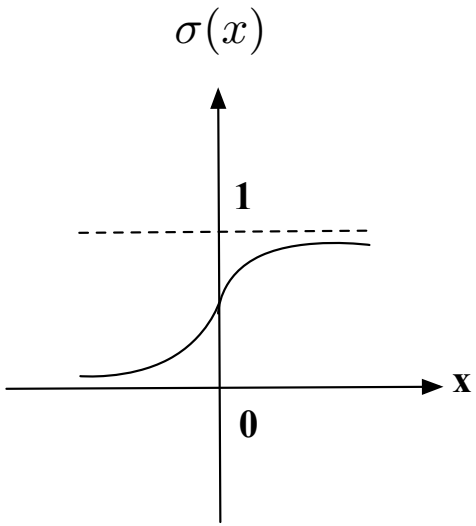
✓ 质量分
- 基于广告的历史平均CTR，只使用了广告侧的信息
- 表达能力有限
- 实时性强

✓ LR为代表的传统机器学习模型
- 结构简单，有一定的个性化表达能力
- 可以在线更新，在线服务

$$y = \frac{\#\ \text{clicks(ad)}}{\#\ \text{impressions(ad)}}$$

$$y = f(x_a)$$

Generation 1
Ad-wise statistical score

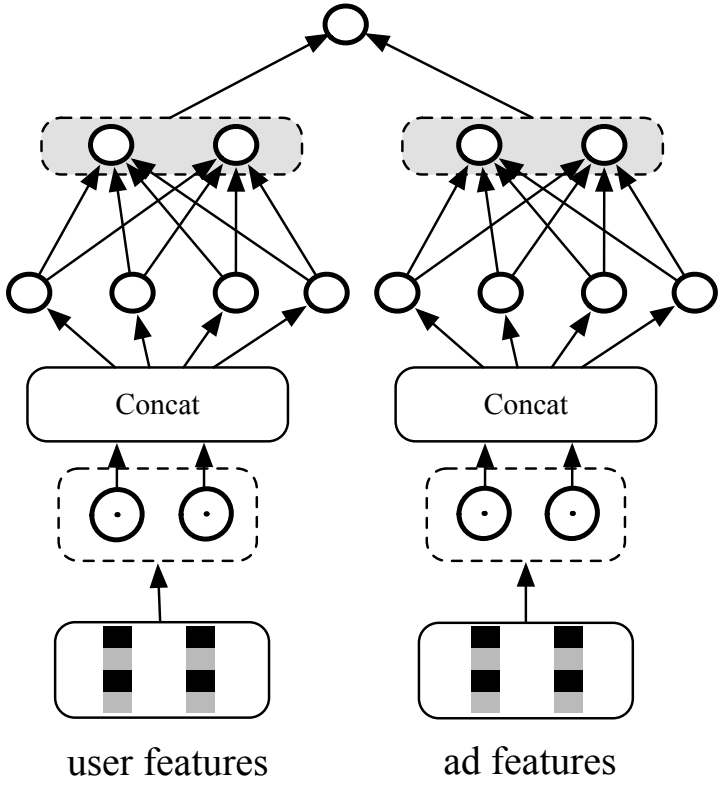


$$y = \sigma(\theta^T x)$$
$$x = \text{concat}(x_u, x_a, x_{ua})$$

Generation 2
Logistic Regression

粗排的深度时代-向量内积模型（2016）

- ✓ 双塔结构，两侧分别输入user特征和ad特征，经过DNN变幻后分别产出user向量和ad向量
- ✓ user侧网络可以引入transformer等复杂结构对用户行为序列进行建模
- ✓ 优点：
  - 内积计算简单，节省线上打分算力
  - user向量和ad向量离线计算产出，因此可以做的非常复杂而不用担心rt问题



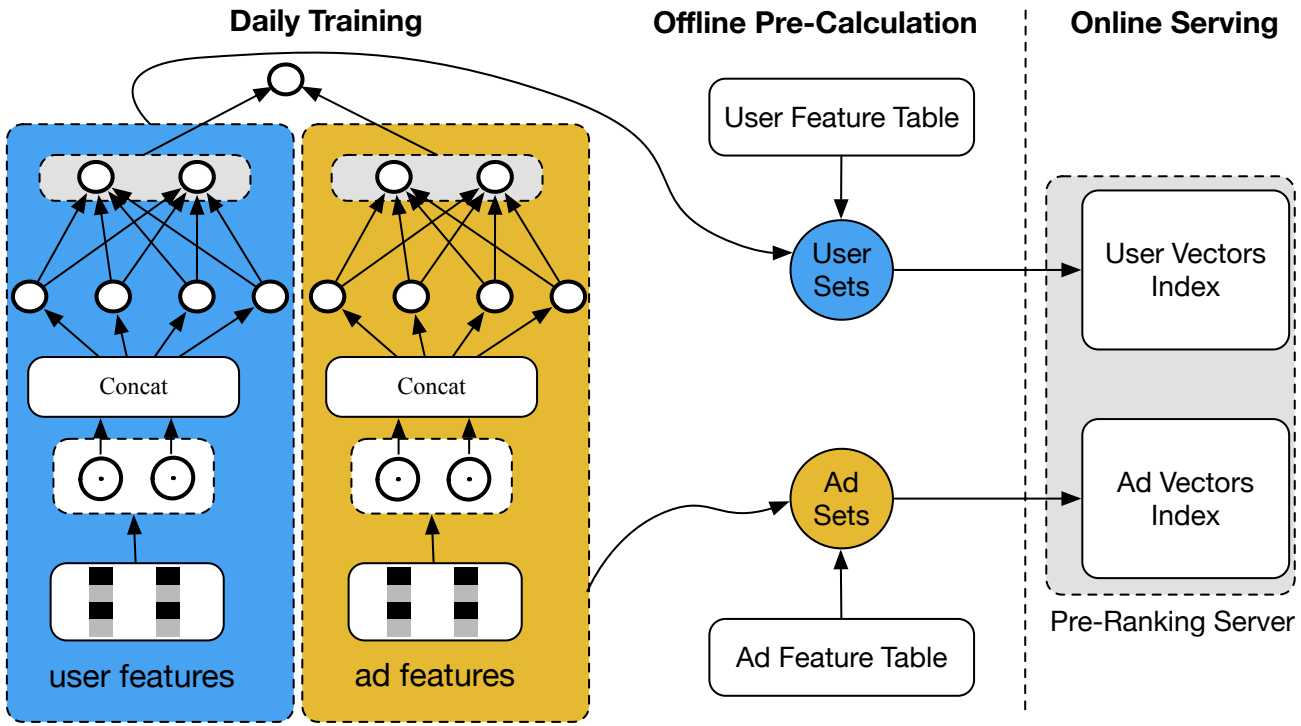
$$y = \sigma(\mathrm{FC}(e_u), \mathrm{FC}(e_a))$$

Generation 3
Vector-Product based DNN

Covington P, Adams J, Sargin E. Deep Neural Networks for YouTube Recommendations. RecSys. 2016.

向量内积模型的问题

- ✓ 模型表达能力受限
  - 难以很好的利用交叉特征
- ✓ 实时性较差
  - user向量和item向量一般需要提前计算好，这种提前计算会拖慢系统更新速度，难以对数据分布快速变化做出响应，例如双十一
  - 冷启动问题，对新广告不友好
- ✓ 迭代效率
  - user向量和item向量的版本同步影响迭代效率



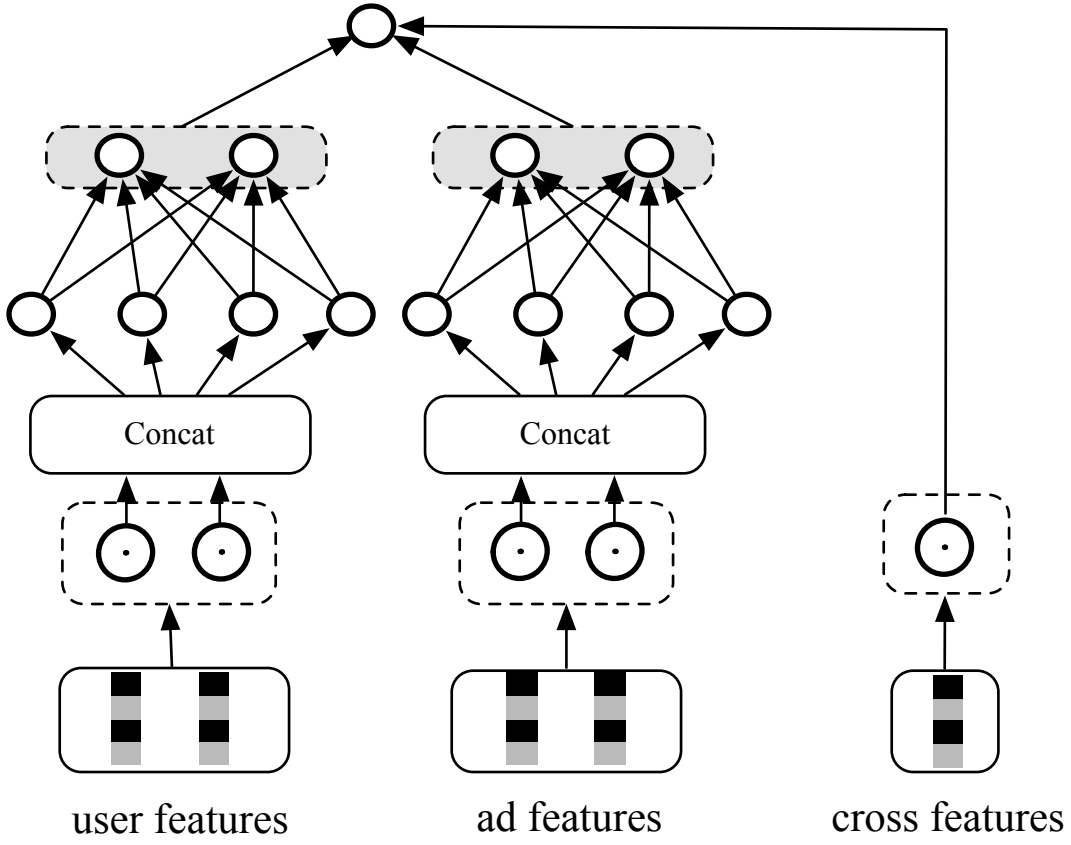
向量内积模型的改进-向量版Wide&Deep模型(2019)

✓ 模型结构：

- Deep部分仍然为向量内积结构
- 通过Wide部分引入交叉特征

✓ 特点：

- 一定程度上克服了内积模型无法使用交叉特征的问题
- Wide部分是线性的，表达能力仍然受到限制



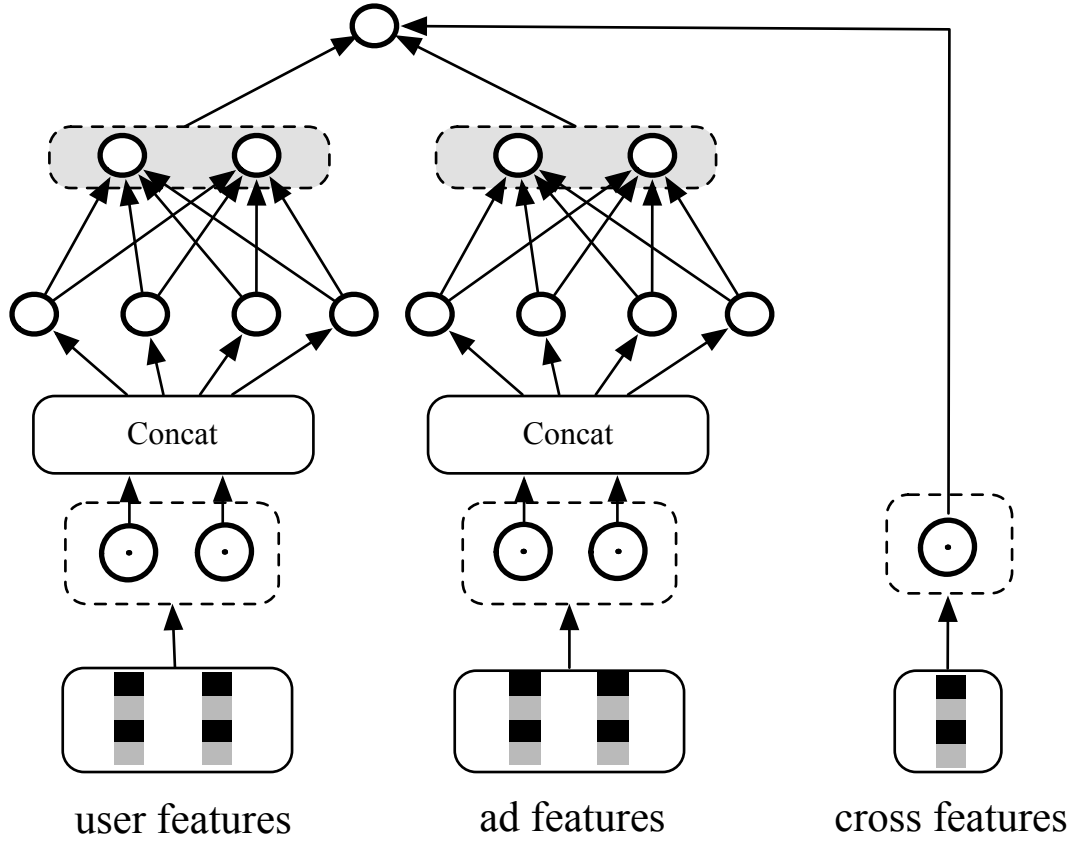
$$y = \sigma(\mathrm{FC}(e_u), \mathrm{FC}(e_a))$$

Generation 3.2
Wide&Vector-Product based DNN

Cheng H-T, Koc L, Harmsen J, et al. Wide & Deep Learning for Recommender Systems. 2016

向量内积模型的改进-实时化(2019)

- ✓ User向量通过线上打分实时产出
- ✓ Ad向量仍然离线产出，但是更新频次加快
- ✓ 特点：
  - 通过实时打分，可以引入实时信息，实时性加强
  - 实时打分使向量内积模型的RT和算力优势减弱
  - 引入新的打分模型和ad向量版本一致性问题



$$y = \sigma(\mathrm{FC}(e_u), \mathrm{FC}(e_a))$$

Generation 3.2
Wide&Vector-Product based DNN

Alibaba Group
阿里巴巴集团

粗排最新进展

COLD：新一代粗排框架（2019）

✓ COLD : Computing power cost-aware Online and Lightweight Deep pre-ranking system

- 基于算法-系统Co-Design视角设计，算力作为一个算力与模型进行联合优化
- 模型结构没有限制，可以任意使用交叉特征
- 工程优化解决算力瓶颈
- 在系统实时系统，实时训练，实时打分，以应对线上分布快速变化

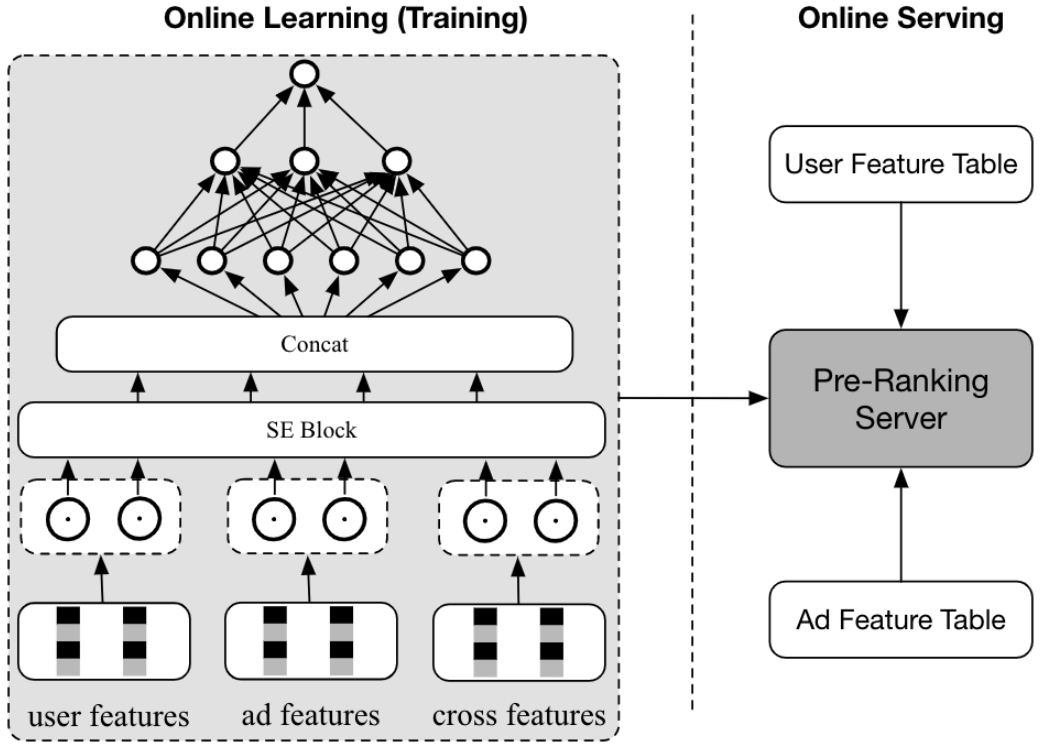


Figure 7: Infrastructure of fully online infrastructure of COLD pre-ranking system.

Wang, Zhe et al. COLD: Towards the Next Generation of Pre-Ranking System. DLP-KDD 2020

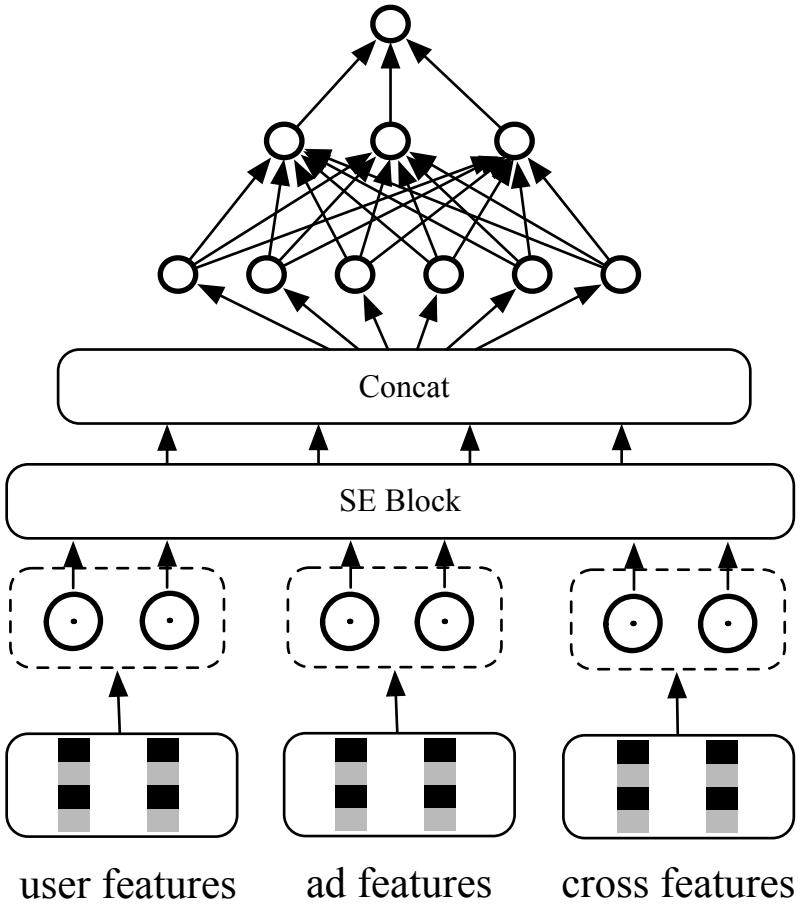
COLD：模型结构

✓ 特征筛选

- 特征重要性计算：基于Se Block，先将M个输入特征的embedding e_i 转拼接在一起，送进全连接网络处理以后，得到M维向量，代表每个特征的重要性得分。特征重要性得分再乘到对应的特征embedding上：

  $$s = \sigma(W[e_1, ..., e_m] + b)$$

- 筛选：对所有特征按重要性得分排序，在满足RT和QPS约束的情况下，选择GAUC最高的特征组合，作为最终使用特征，以灵活的平衡算力和效果
- Se Block仅用于特征筛选阶段，线上模型不包含该结构



$$y = \sigma(\mathrm{FC}(e_u, e_a, e_{ua}))$$

COLD：模型结构

✓ 基于scaling factor的结构化剪枝：

- 在每个神经元的输出后面乘上一个gamma，然后对gamma进行稀疏惩罚，当某一神经元的gamma为0时，此时该神经元的输出为0，对其后的模型结构不再有任何影响，即视为该神经元被剪枝
- iterative pruning的方式，每隔t轮训练会对gamma为0的神经元进行mask，这样可以保证整个剪枝过程中模型的稀疏率是单调递减的

$$\min_{w,\lambda} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i, W, \gamma)) + R_s(\gamma)$$

- 在效果基本不变的情况下，粗排GPU的QPS提升20%

✓ 最终模型是7层全连接网络

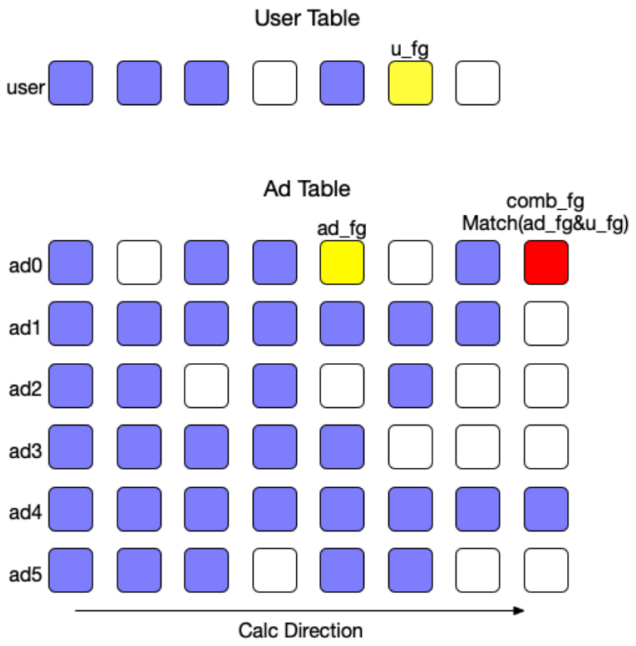
COLD：工程优化

✓ 并行优化：

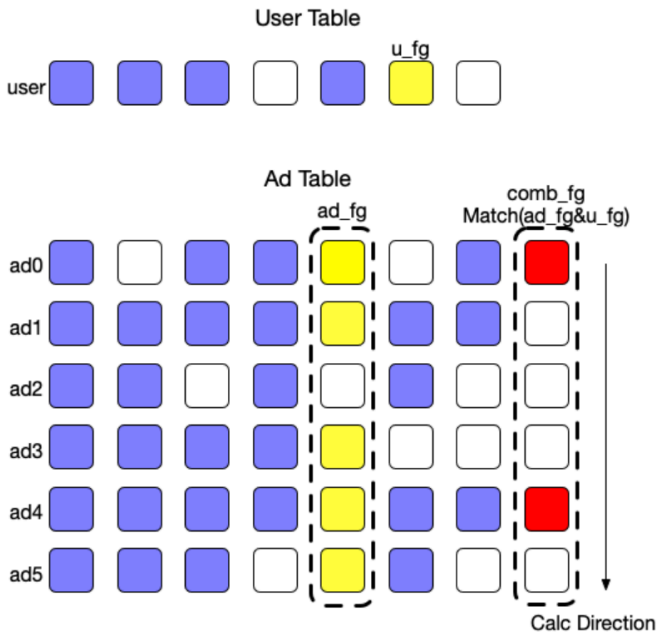
- 将打分请求拆包以后，在特征计算和模型计算的各个地方，尽可能进行多线程优化

✓ 列计算转换:

- 行计算方式：逐个广告计算在不同feature group下的特征，存在访存不连续的，有冗余遍历，查找的问题。
- 列计算方式：因为同一个feature group的计算方法相同，因此可以按列进行特征计算，对同一列上的稀疏数据进行连续存储，之后利用MKL优化单特征计算，使用SIMD (Single Instruction Multiple Data)优化组合特征算子，以打到加速的目的。



(A) Row based Feature Computation



(B) Column based Feature Computation

COLD：工程优化

✓ Float16加速：

- linear log trick

$$linear\_log(x) = \begin{cases} -log(-x) - 1 & x < -1 \\ x & -1 \leq x \leq 1. \\ log(x) + 1 & x > 1 \end{cases}$$

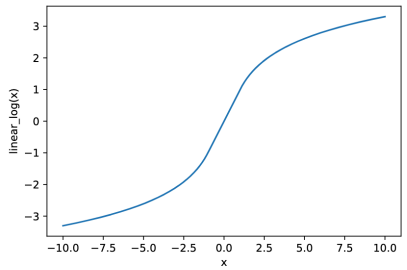


Figure 6: The linear_log function

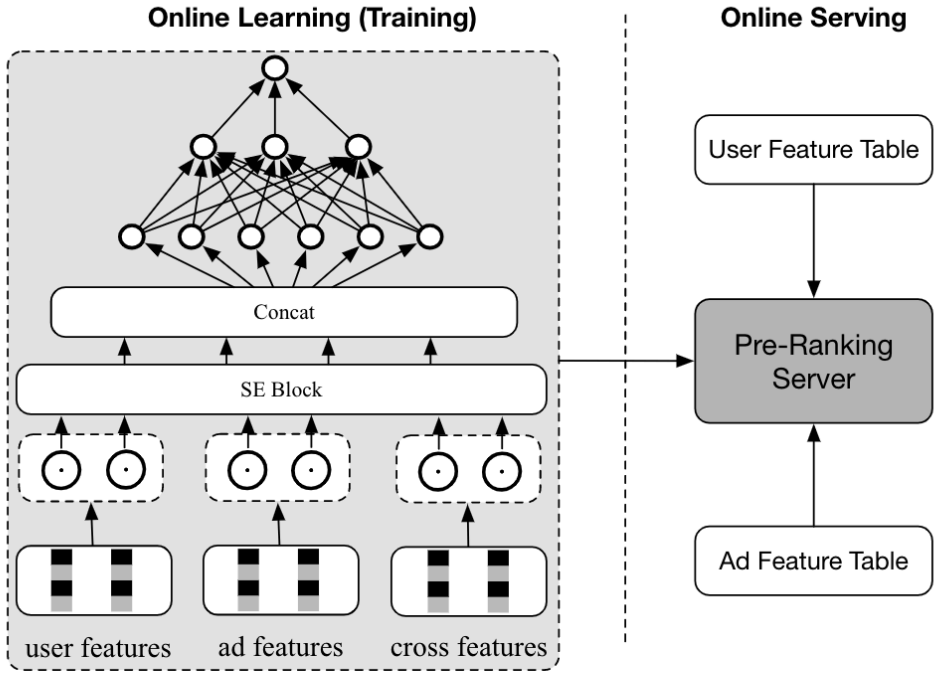


Figure 7: Infrastructure of fully online infrastructure of COLD pre-ranking system.

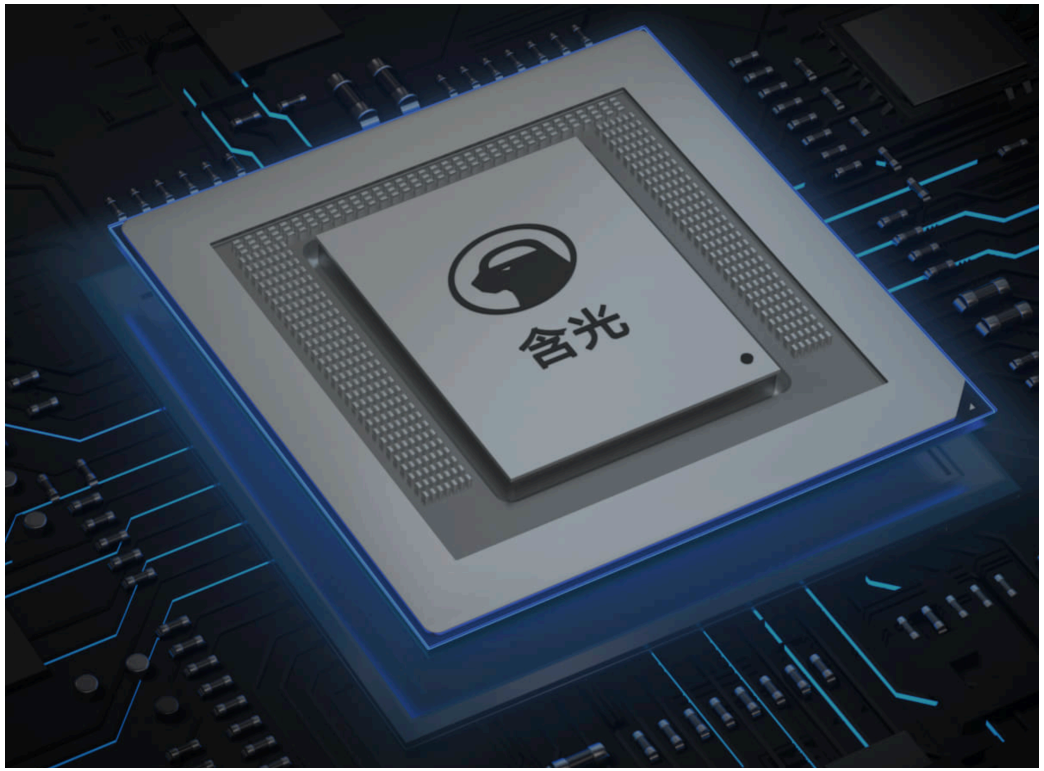
COLD：工程优化

✓ MPS（ Multi-Process Service):

- 解决kernel启动的开销

✓ NPU:

- 使用阿里自研的含光800 NPU专有硬件，替代原来的GPU，QPS进一步提升约1倍



COLD：在线服务架构

- ✓ 在线实时ODL训练
- ✓ 在线实时inference
- ✓ 更及时响应数据分布变化，对新广告更友好
- ✓ 实时架构对模型迭代和在线A/B测试都更有利

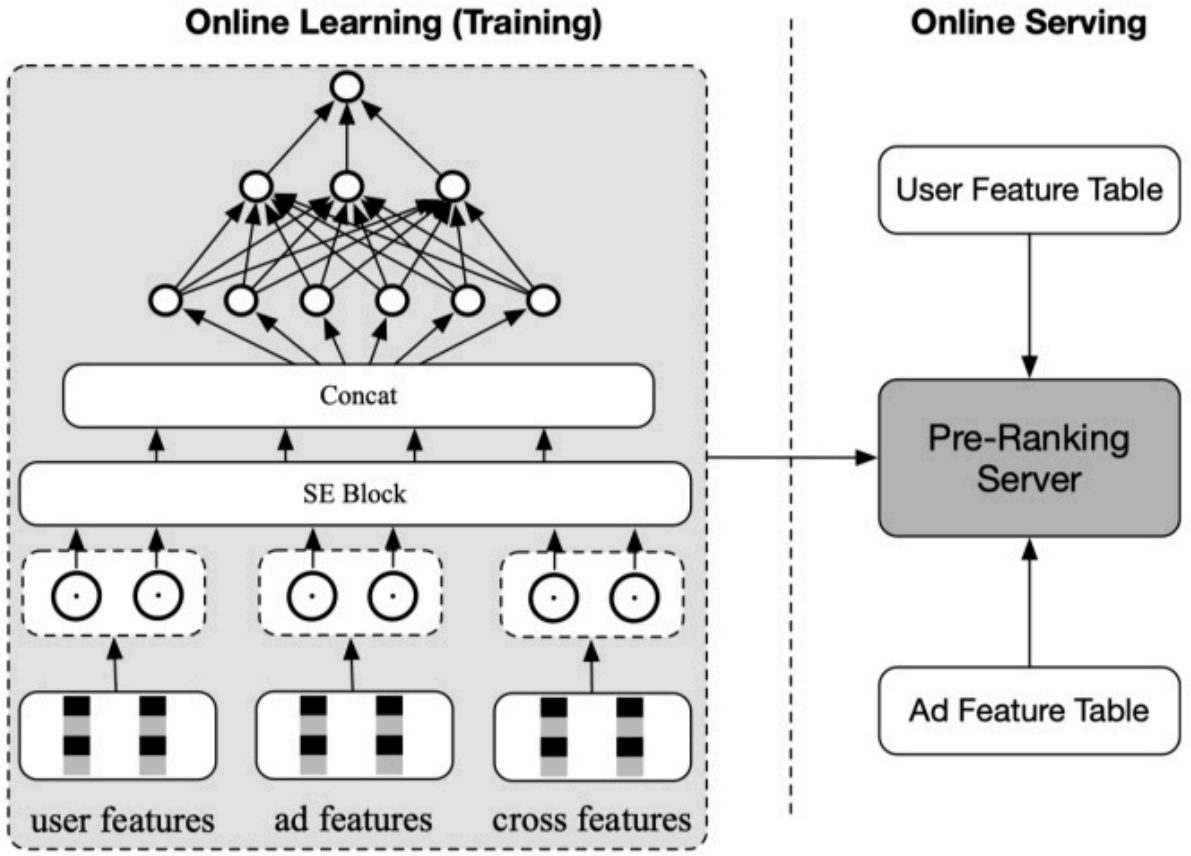


Figure 7: Infrastructure of fully online infrastructure of COLD pre-ranking system.

实验结果

✓ 离线实验

Method	GAUC	Recall
Vector-Product based DNN Model	0.6232	88%
COLD	0.6391	96%
DIEN	0.6511	100%

✓ 在线效果

Time	CTR lift	RPM lift
Normal Days	+6.1%	+6.5%
Double 11 Event	+9.1%	+10.8%

✓ 2019年以来，COLD已经在阿里妈妈定向广告各主要业务线落地，并取得了可观的线上效果提升。

实验结果

✓ 不同模型结构的性能表现

Model	QPS	RT
Vector-Product based DNN Model	60000+	2ms
COLD	6700	9.3ms
DIEN	629	16.9ms

Model	QPS	RT	GAUC
COLD (No Cross Features)	6860	8.6ms	0.6281
COLD *	6700	9.3ms	0.6391
COLD (All Features)	2570	10.6ms	0.6467

COLD的进一步发展-与精排更深度的整合

✓ 背景

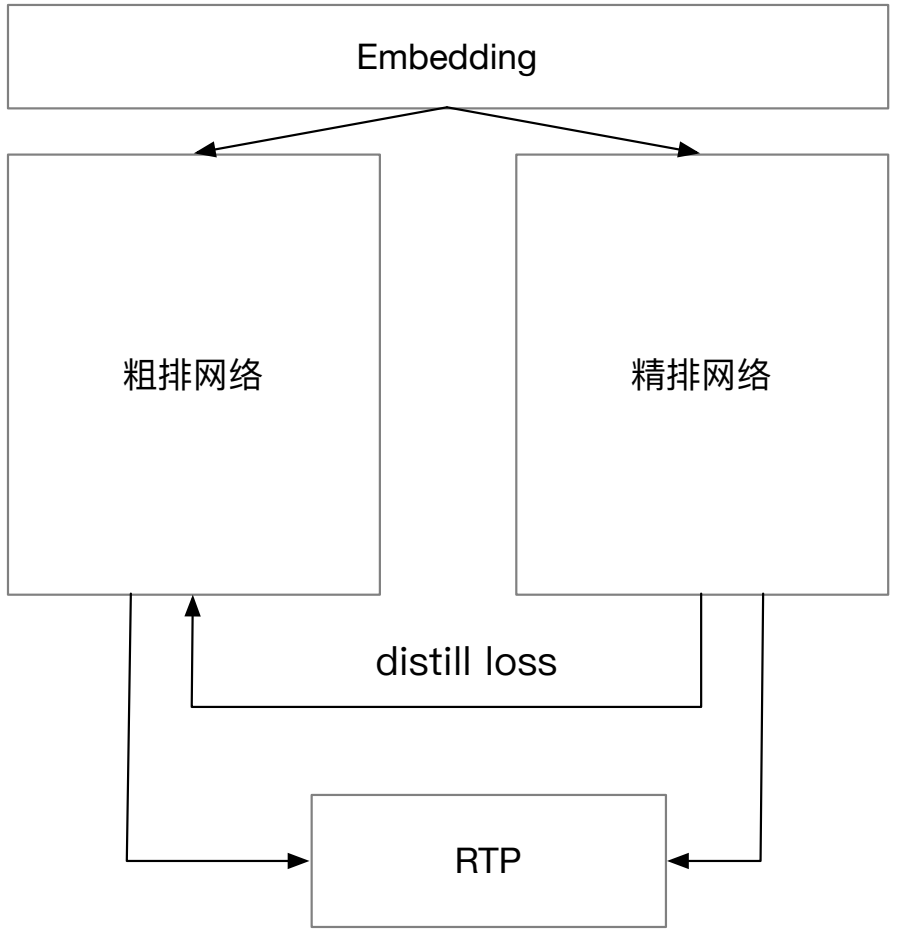
- 粗排精排独立迭代，存在前后不一致造成的链路损耗问题
- 粗排精排两套训练流程，维护成本较高
- 实时化的COLD，使粗排和精排进行更深度的联动成为可能。

✓ 技术方案

- 粗排精排联合训练，共享部分参数，精排得分用于对粗排的优势特征蒸馏和优势结构蒸馏
- 引入精排参竞日志，对于未展现样本借助精排得分进行辅助学习

✓ 优点

- 粗排精排模型一起训练，一起产出，提升对齐程度
- 引入精排参竞日志，缓解粗排解空间问题
- 降低运维成本，减少训练资源，提升迭代效率



Alibaba Group 阿里巴巴集团

总结与展望

总结

- ✓ 粗排目前已经全面迈向深度学习时代
- ✓ 深度学习时代的粗排目前存在向量内积和COLD两种主流技术路线
- ✓ 没有最好的算法，只有最合适的算法

展望

✓ 粗排未来发展的两种可能：

- 粗排精排化：
  - 精排技术持续向粗排迁移，粗排和精排的界限逐渐模糊，走向更深层次的整合和一体化。
  - 算力作为一个变量参与优化，精排存在多个不同算力版本的子模型，粗排只是其中一个，跟随精排自动升级和迭代，从而实现全链路算力和效果的平衡。
- 回归集合选择的本质：
  - 以产出符合后链路需要的集合为目标，真正以集合为对象进行建模。

Thanks