粗排

王树森

http://wangshusen.github.io/

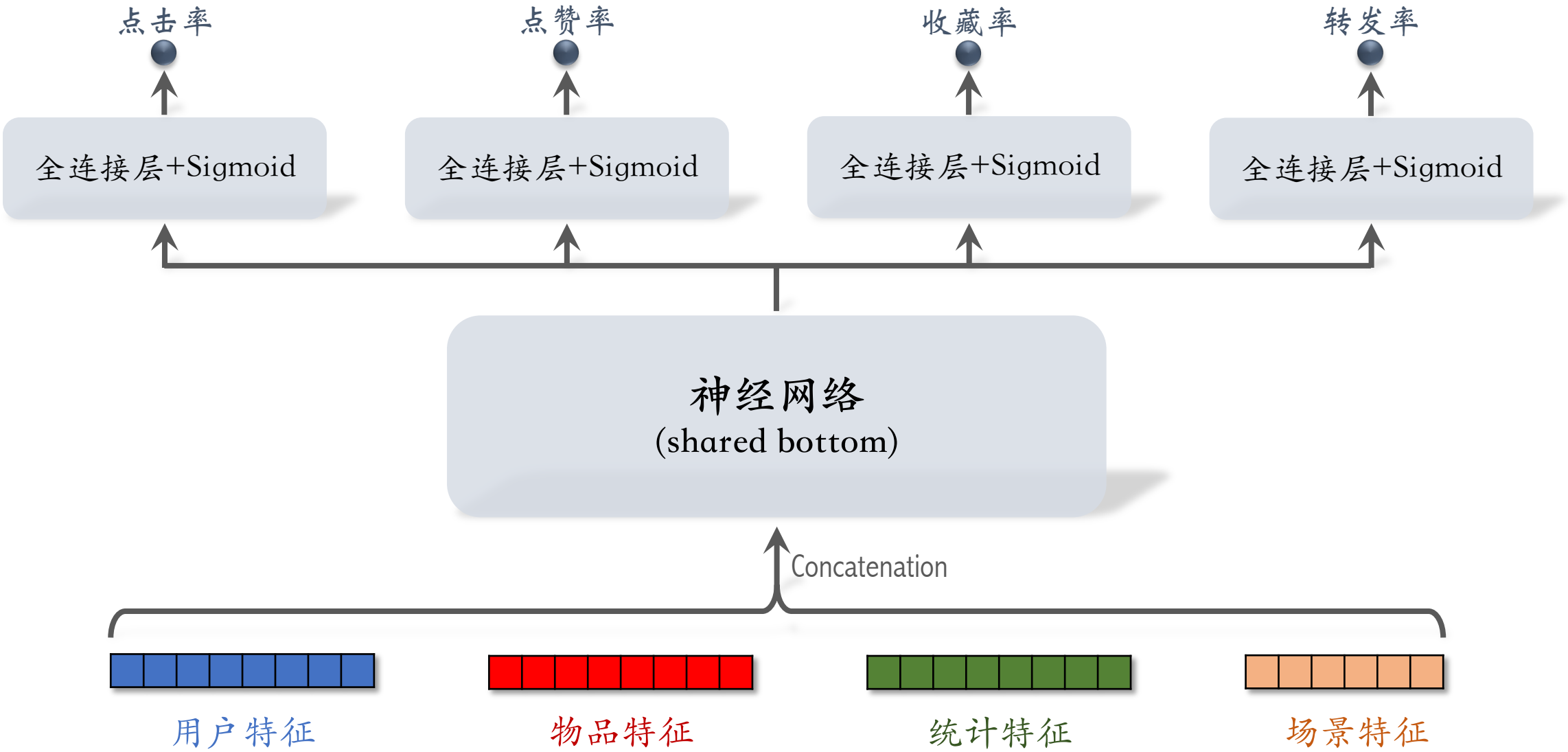
小红书

粗排 vs 精排

粗排	精排
• 给几千篇笔记打分。	• 给几百篇笔记打分。
• 单次推理代价必须小。	• 单次推理代价很大。
• 预估的准确性不高。	• 预估的准确性更高。

精排模型 & 双塔模型

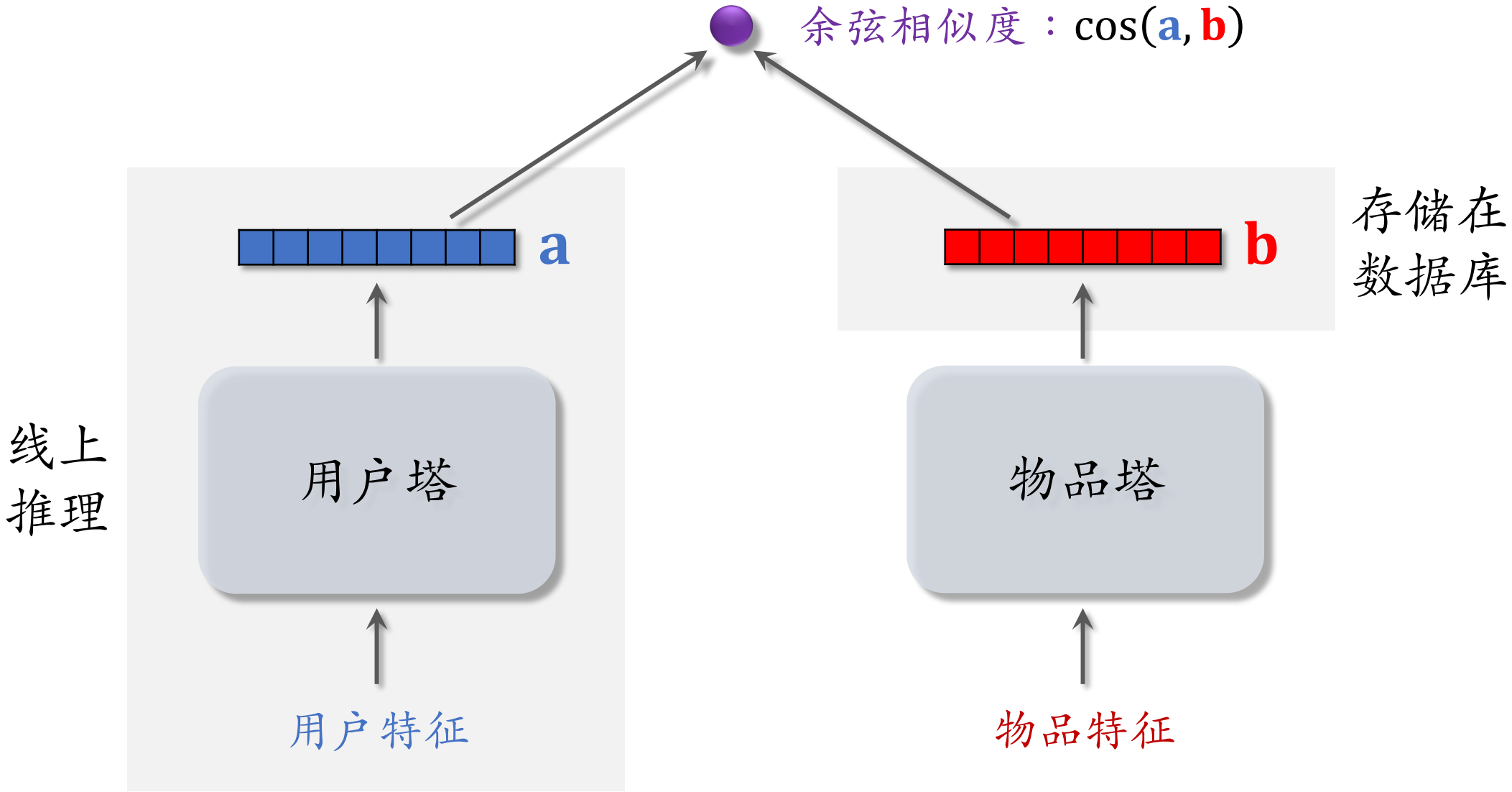
精排模型


点击率
点赞率
收藏率
转发率
全连接层+Sigmoid
全连接层+Sigmoid
全连接层+Sigmoid
全连接层+Sigmoid
神经网络
(shared bottom)
Concatenation
用户特征
物品特征
统计特征
场景特征

精排模型

- 前期融合：先对所有特征做 concatenation，再输入神经网络。
- 线上推理代价大：如果有 n 篇候选笔记，整个大模型要做 n 次推理。

双塔模型



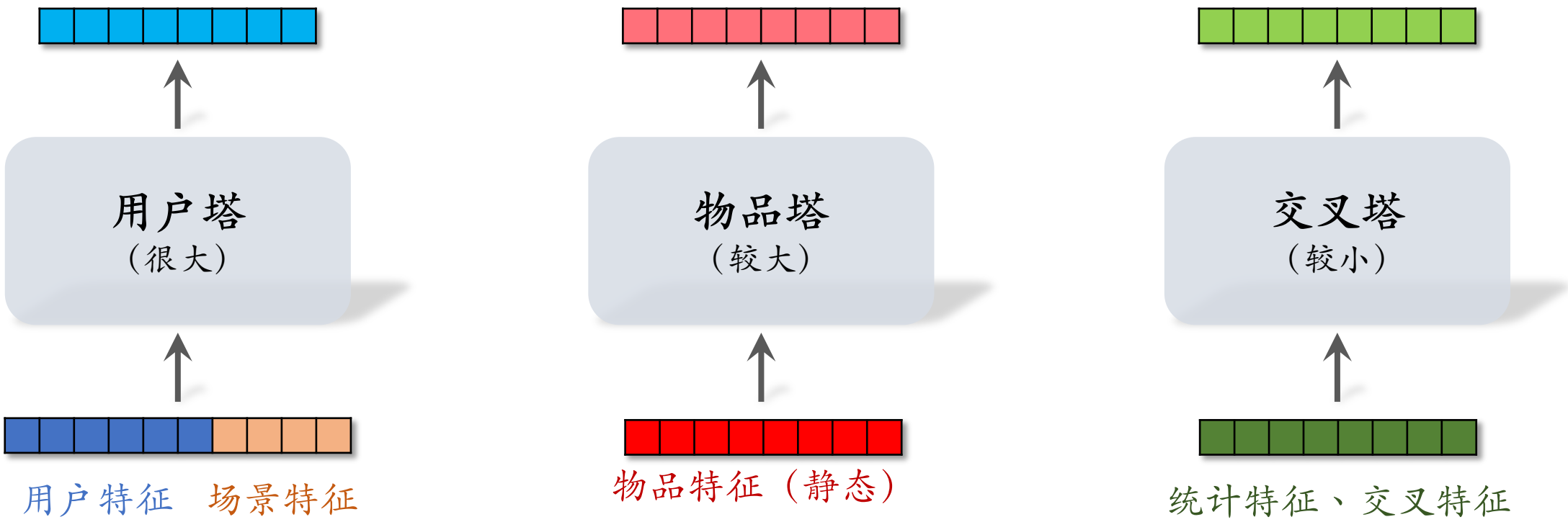
双塔模型

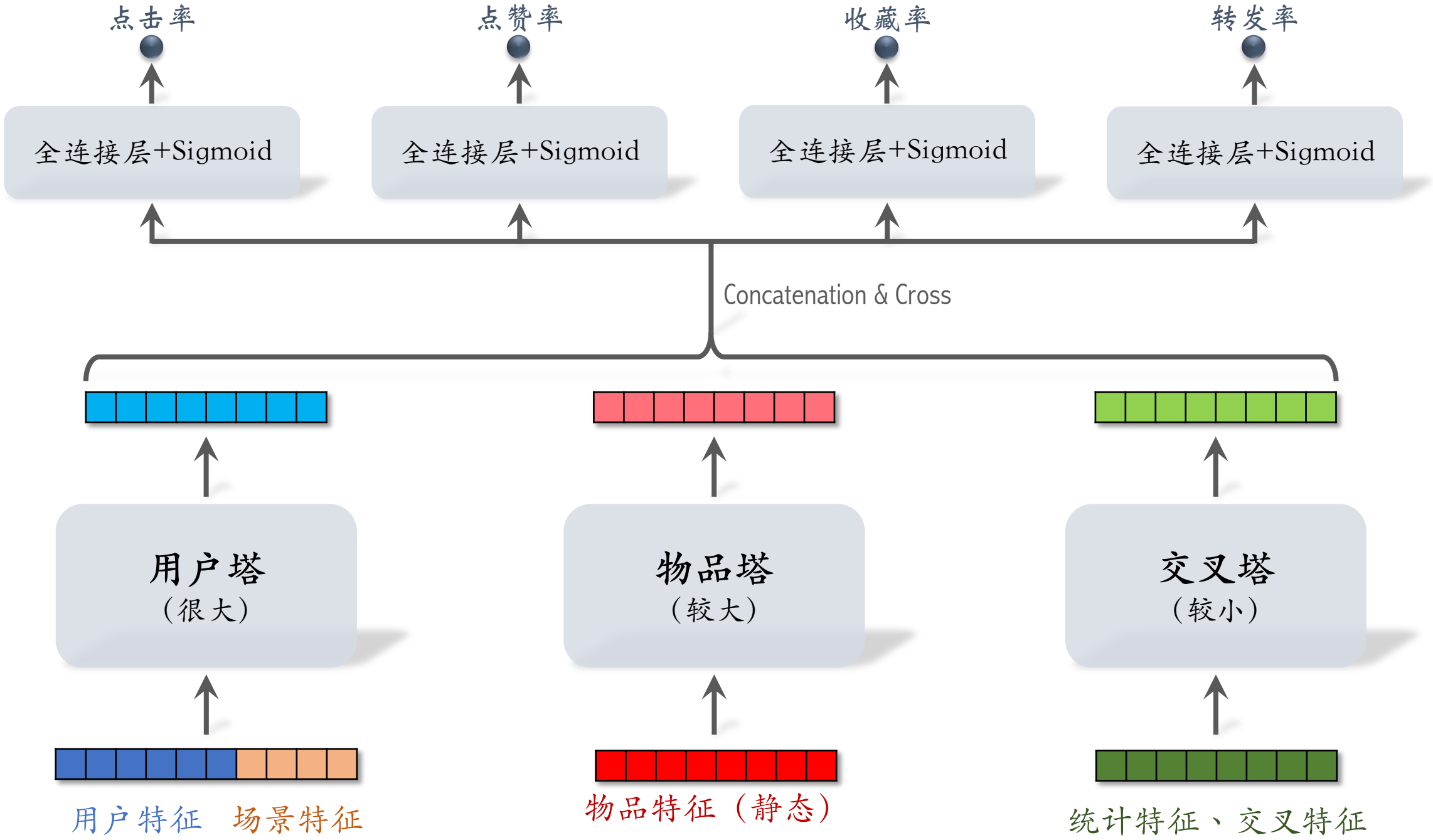
- 后期融合：把用户、物品特征分别输入不同的神经网络，不对用户、物品特征做融合。
- 线上计算量小：
  - 用户塔只需要做一次线上推理，计算用户表征 **a**。
  - 物品表征 **b** 事先储存在向量数据库中，物品塔在线上不做推理。
- 预估准确性不如精排模型。

粗排的三塔模型

参考文献：

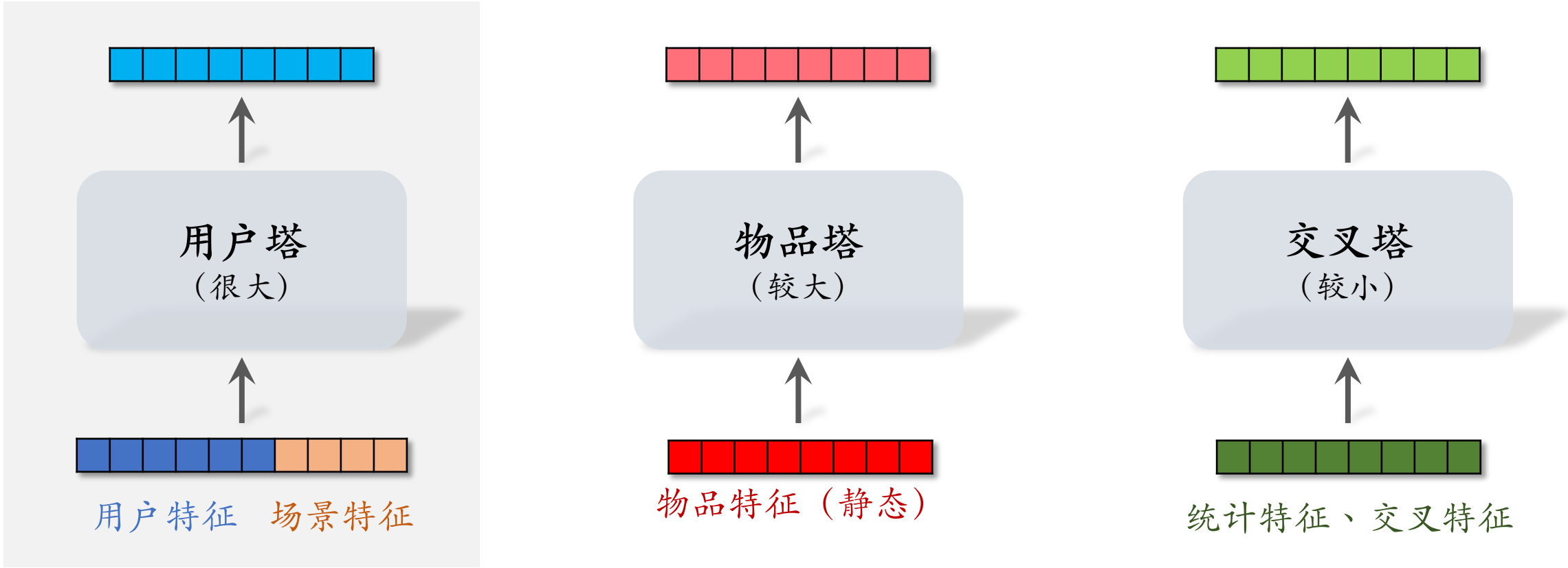
- Zhe Wang et al. COLD: Towards the Next Generation of Pre-Ranking System. In *DLP-KDD*, 2020.


用户塔
（很大）
物品塔
（较大）
交叉塔
（较小）
用户特征
场景特征
物品特征（静态）
统计特征、交叉特征


点击率
点赞率
收藏率
转发率
全连接层+Sigmoid
全连接层+Sigmoid
全连接层+Sigmoid
全连接层+Sigmoid
Concatenation & Cross
用户塔
（很大）
物品塔
（较大）
交叉塔
（较小）
用户特征
场景特征
物品特征（静态）
统计特征、交叉特征

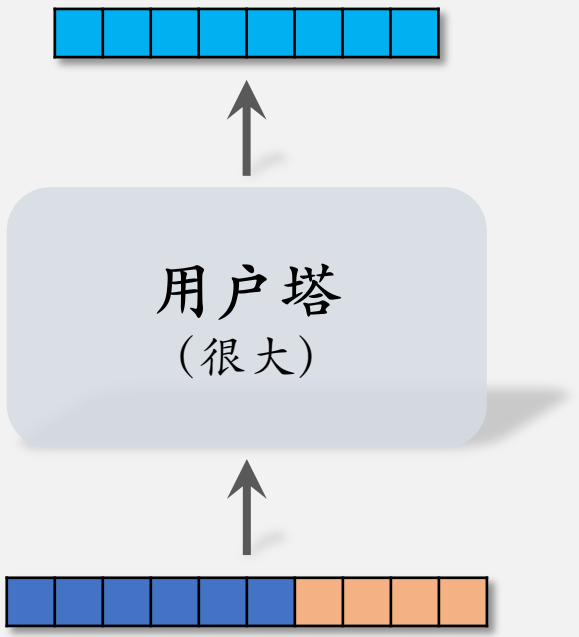
粗排的三塔模型

- 只有一个用户，用户塔只做一次推理。
- 即使用户塔很大，总计算量也不大。

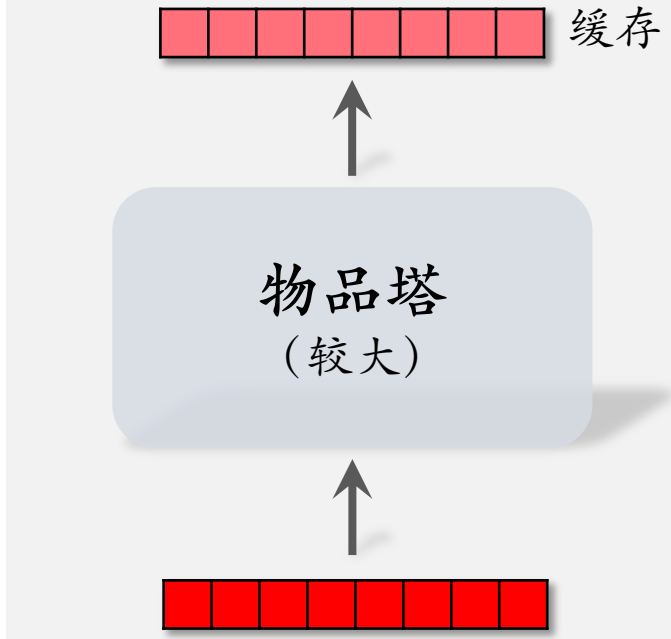


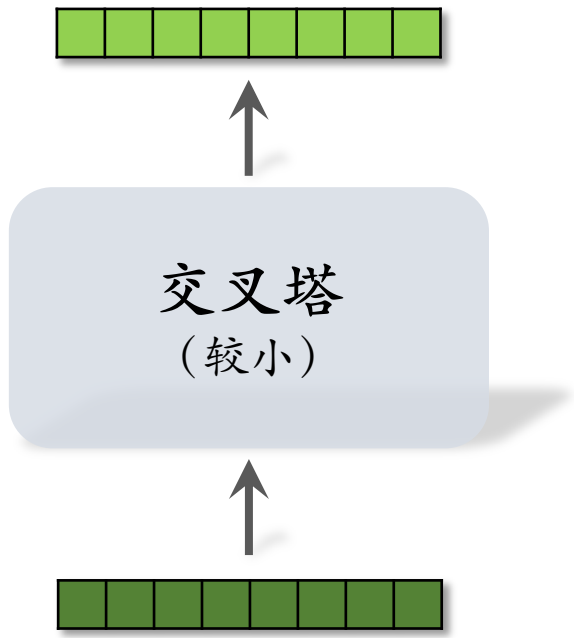
粗排的三塔模型

- 只有一个用户，用户塔只做一次推理。
- 即使用户塔很大，总计算量也不大。



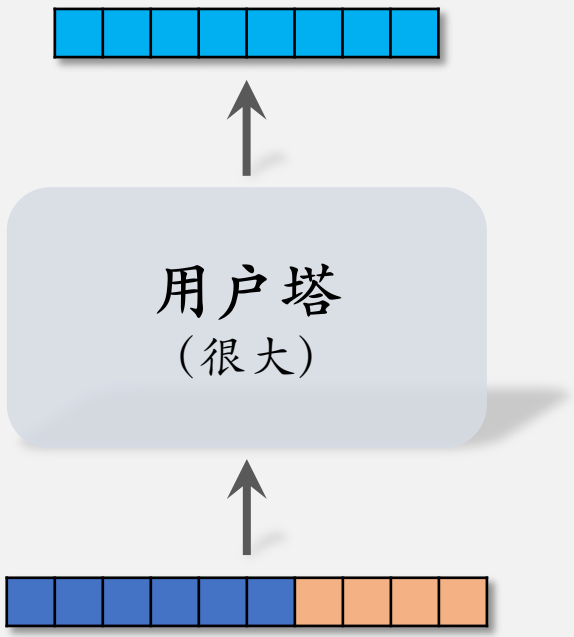
- 有 n 个物品，理论上物品塔需要做 n 次推理。
- PS 缓存物品塔的输出向量，避免绝大部分推理。



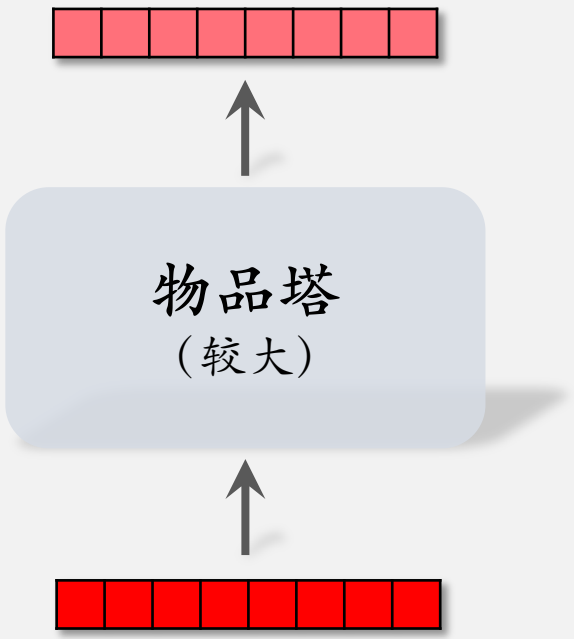


粗排的三塔模型

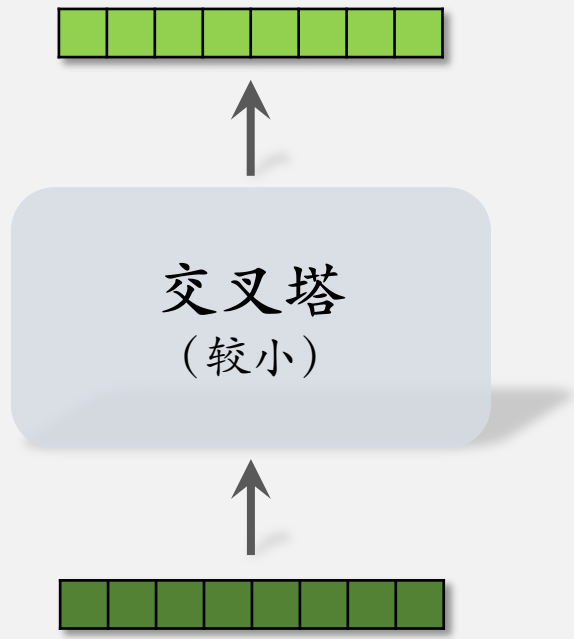
- 只有一个用户，用户塔只做一次推理。
- 即使用户塔很大，总计算量也不大。

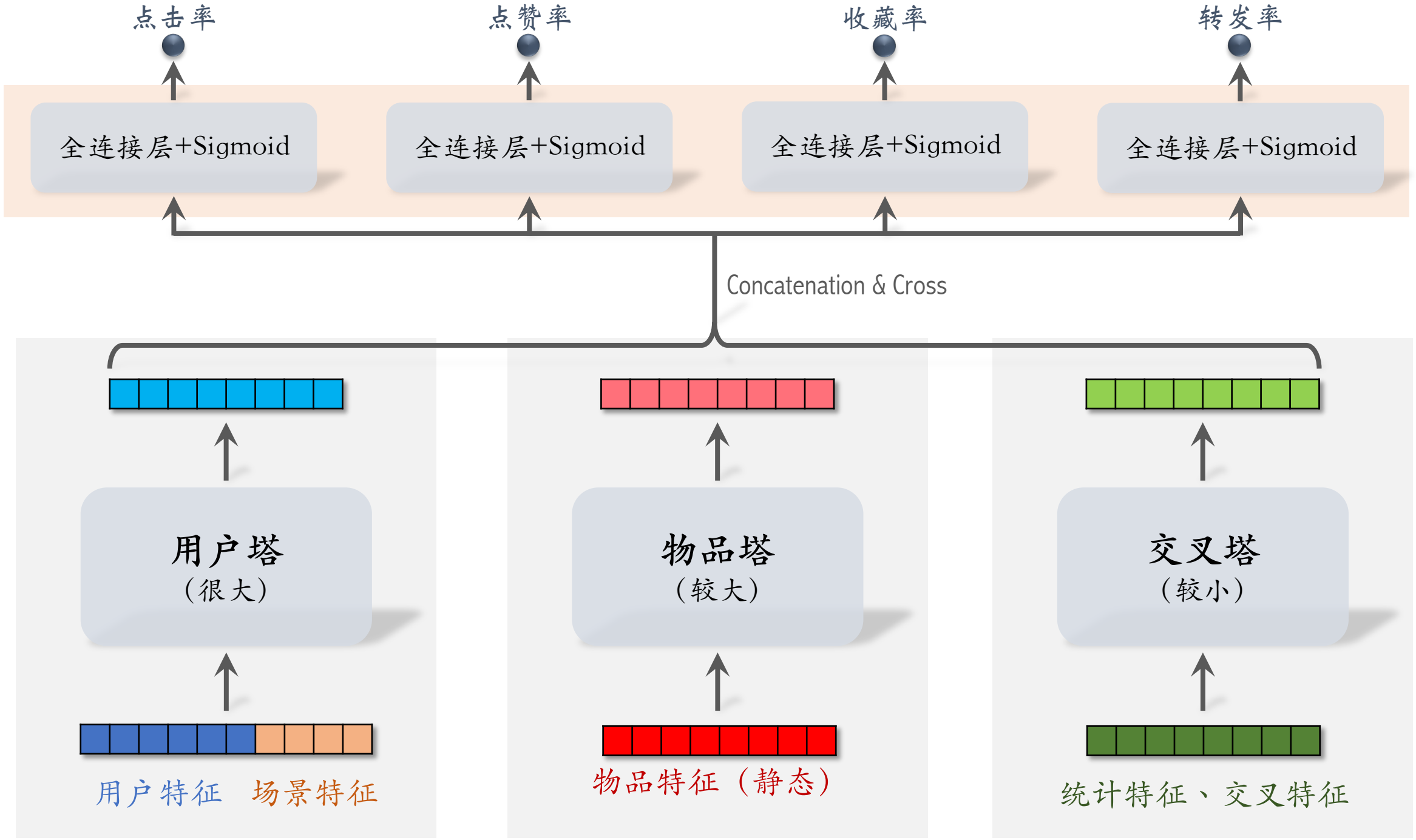


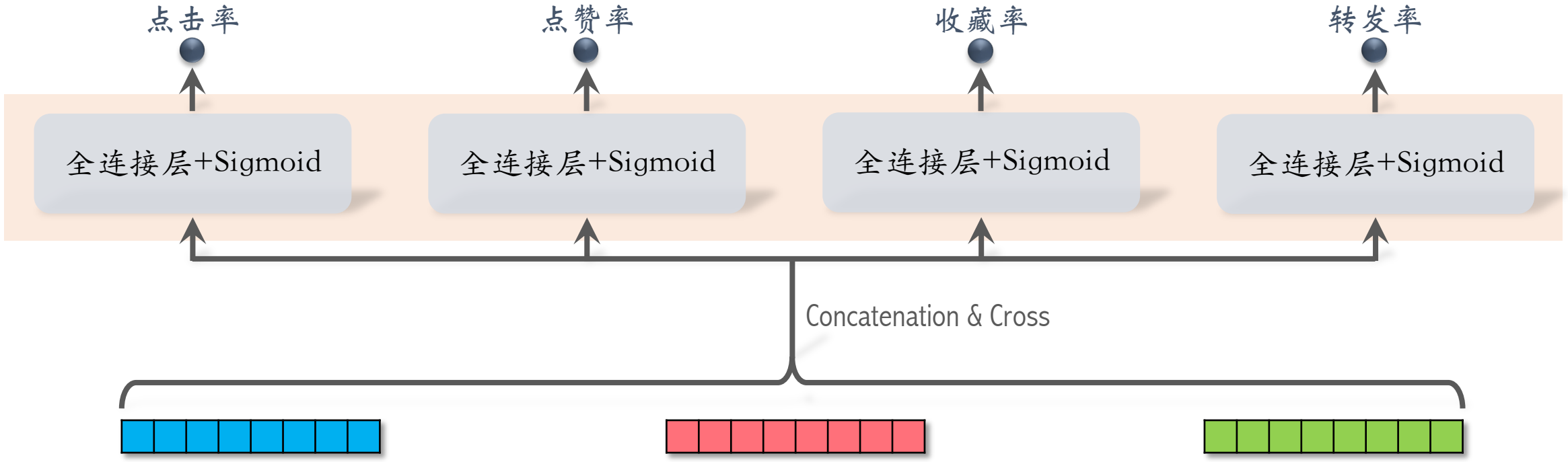
- 有 n 个物品，理论上物品塔需要做 n 次推理。
- PS 缓存物品塔的输出向量，避免绝大部分推理。



- 统计特征动态变化，缓存不可行。
- 有 n 个物品，交叉塔必须做 n 次推理。




点击率
点赞率
收藏率
转发率
全连接层+Sigmoid
全连接层+Sigmoid
全连接层+Sigmoid
全连接层+Sigmoid
Concatenation & Cross
用户塔
（很大）
物品塔
（较大）
交叉塔
（较小）
用户特征
场景特征
物品特征（静态）
统计特征、交叉特征



- 有 n 个物品，模型上层需要做 n 次推理。
- 粗排推理的大部分计算量在模型上层。

三塔模型的推理

- 从多个数据源取特征：
  - 1 个用户的画像、统计特征。
  - n 个物品的画像、统计特征。
- 用户塔：只做 1 次推理。
- 物品塔：未命中缓存时需要做推理。
- 交叉塔：必须做 n 次推理。
- 上层网络做 n 次推理，给 n 个物品打分。

Thank You!

http://wangshusen.github.io/