

大模型在电商推荐的应用

张超 / 京东健康推荐团队负责人

2024-08-03

目录

- ① 大模型推荐技术发展回顾
- ② 健康电商推荐背景与挑战
- ③ 大模型推荐在电商场的落地实践

大模型推荐技术发展回顾

- 视角1: 对现有推荐的“改变”程度

16年 开始CTR大模型
存储密集到计算密集

- ① Behavior sequence – DIN, DIEN, SIM... life long
- ② Feature interaction – FM, DCN, Transformers, PPNET ...
- ③ Multimodal, Multi task, Multi objective... ESMM, PLE...
- ④ Global hash...

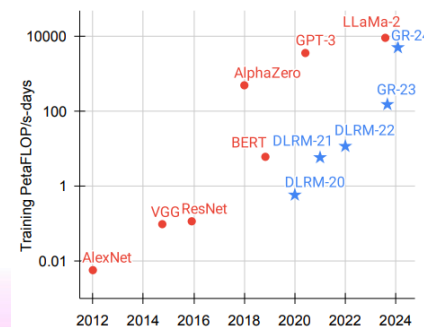
20年 LLM语言大模型
增强Recsys

- ① Data/Fe - LLM token, prompt engineer for text slot preprocess
- ② NN Representation - LLM embedding, for directly match; for pre-load + finetune, for u2i or ctr model

24年 技术范式

多级过滤判别到生成式

- ① LLM as Rec – P5
- ② Meta GR– all - milestone



大模型推荐技术发展回顾

- 视角2：大模型能力 如何解 推荐曾经发展中的“瓶颈” (1/2)

推荐曾经发展中的“瓶颈”

阶段1:

卷数据

阶段2:

卷网络表达

阶段3:

卷算力&卷Infra

- ✓ PC时代数据远不如今天 (volume, side info, label), 13年移动互联网爆发得到缓解
- ✓ 受限于产品设计 (强账号产品少) ID, label和feature均稀疏
- ✓ 13年毕设基于知识地图的学习路径推荐, 大量工作在标数据
- ✓ 发论文和打比赛讨论: 找数据洗数据 + 树模型打天下 (GBDT, 再到后来XGB, 竞赛利器)

- ✓ NN时代到来 - LR换DNN; 开始从人工交叉到model交叉
- ✓ 浅层时代CTR代表作 - 百度大规模离散LR, 奠定了后来发展的两个走向: **高效的计算框架、精细的特征工程**; show+click+mlp+ubm, join/update交替训练;
- ✓ NN时代: 特工从人交给网络结构, 开始了持续至今的模型卷特征交叉表达的时代

- ✓ transformer, 天然可并行的叠罗汉结构让算力有了卷的条件;
- ✓ 行为序列的丰富 (长*宽*高) 和fea interaction更复杂也给算力和infra提出了新要求
- ✓ 对话场景 VS 推荐场景 用户RT的容忍 对REC算力提出更高的要求

大模型推荐技术发展回顾

- 视角2：大模型能力 如何解 推荐曾经发展中的“瓶颈” （2/2）

大模型关键能力：Scaling Law 质变涌现 + World Knowledge 基础理解

	数据瓶颈	NN表达瓶颈	算力&infra 瓶颈
World knowledge	<p>显性知识：丰富样本行和列</p> <ul style="list-style-type: none">● 样本生成 Eg. 搜索相关性 – 相似query, 通过prompt engine生成hard/easy 样本; query suggestion, query rewrite;● 数据预处理 Eg. 文本类数据kw extract	<p>隐性知识：作为input or preload更好表达</p> <ul style="list-style-type: none">● Embedding for user or item understanding Eg. 召回直接用于i2i, or 作为新增的辅助slot放入 u2i or ctr模型输入● Embedding for model pre-load and finetune Eg. 冷启动, side info给一个相对不错的init	
Scaling law		<p>Scaling law for 建模表达丰富</p> <ul style="list-style-type: none">● 召回: deep i2i, 头部item精准到全量item精准; u2i, more ps more data get better performance● 排序: UBM的进化, 同质数据空间的长宽高到 异质数据全空间的统一生成式建模; feature interaction layer的不断叠加	<p>与scaling law同步适配进化的算力和infra</p> <ul style="list-style-type: none">● 设计适用于并行计算的结构+底层加速计算优化; Eg. 多头注意力等结构叠罗汉, Infer moe sparse 加速推理, 各种大模型发展过程带来的加速技巧● infra空间换时间, 实时换近线, 近线转离线;

- Scaling Law条件：工业界海量数据优势+transformer等高并行网络结构+model算力提升+搜推广infra不断拉高quota
- Scaling Law 在NLP的成功 再次坚定 搜推广的尝试方向

目录

- ① 大模型推荐技术发展回顾
- ② 健康电商推荐背景与挑战
- ③ 大模型推荐在电商场的落地实践

健康电商推荐系统背景与挑战

京东健康推荐覆盖京东App垂直频道、京东健康APP、健康ABC小程序等多个场域的商品、内容、榜单、店铺、服务等多素材的个性化推荐，承接80+场域的个性化推荐分发。



京东健康APP-综合推荐



京东APP-首页feed流



京东APP-频道推荐



健康中心-综合推荐



互联网医院医生推荐

- 场景特点:
- ✓ 刚需/知识 VS 兴趣
 - ✓ 标品 VS 非标
 - ✓ 低频稀疏冷启比高
 - ✓ 多场景差异大

目录

- ① 大模型推荐技术发展回顾
- ② 健康电商推荐背景与挑战
- ③ 大模型推荐在电商场的落地实践

大模型结合推荐在JDH落地工作概述

生成式推荐召回工作

- 召回: LLM4CB

解决问题: 稀疏行为健康新/低频用户的标品召回

解决思路:

第一, 利用LLM的世界知识与健康推荐刚需&知识驱动的特点, 增强LLM对用户和商品的理解;

第二, 利用推荐系统的样本数据, 拉齐世界知识和垂域场知识的gap;

第三, LLM推理无法满足召回的RT要求, 离线/近线处理, 损失覆盖率;

技术挑战:

- ID物料表示: 兼容语义和区分度
- 任务对齐: 用推荐样本与LLM 精调任务对齐
- Inference: online性能扛不住, 精度&覆盖率 trade off

传统推荐在scale up上的工作

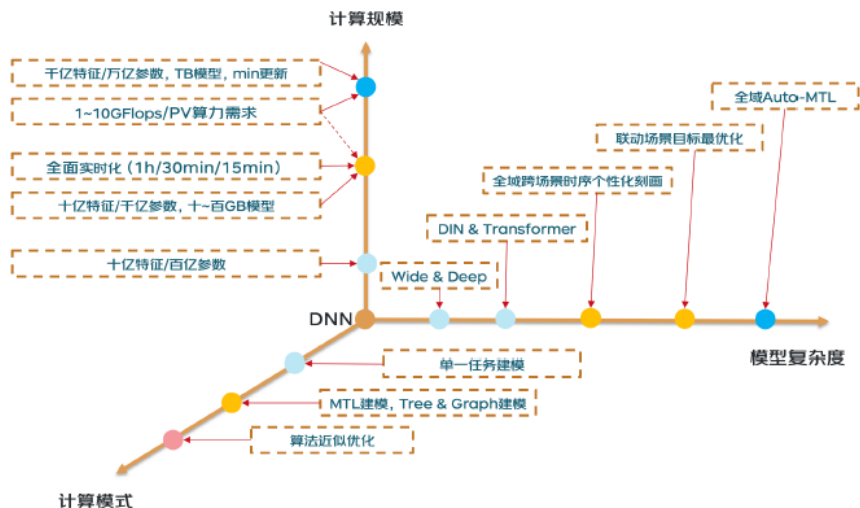
- 召回: Deep I2I

wider and deeper. (word2vec, Dssm, BGE based pre-train...)

一路I2I打通全场景 (中长尾场景也获得一个相对不错的召回)

- 排序: CTR大模型

CTR model变大, 从存储型的大往计算型的大演进



LLM4CB – LLM解稀疏行为用户召回

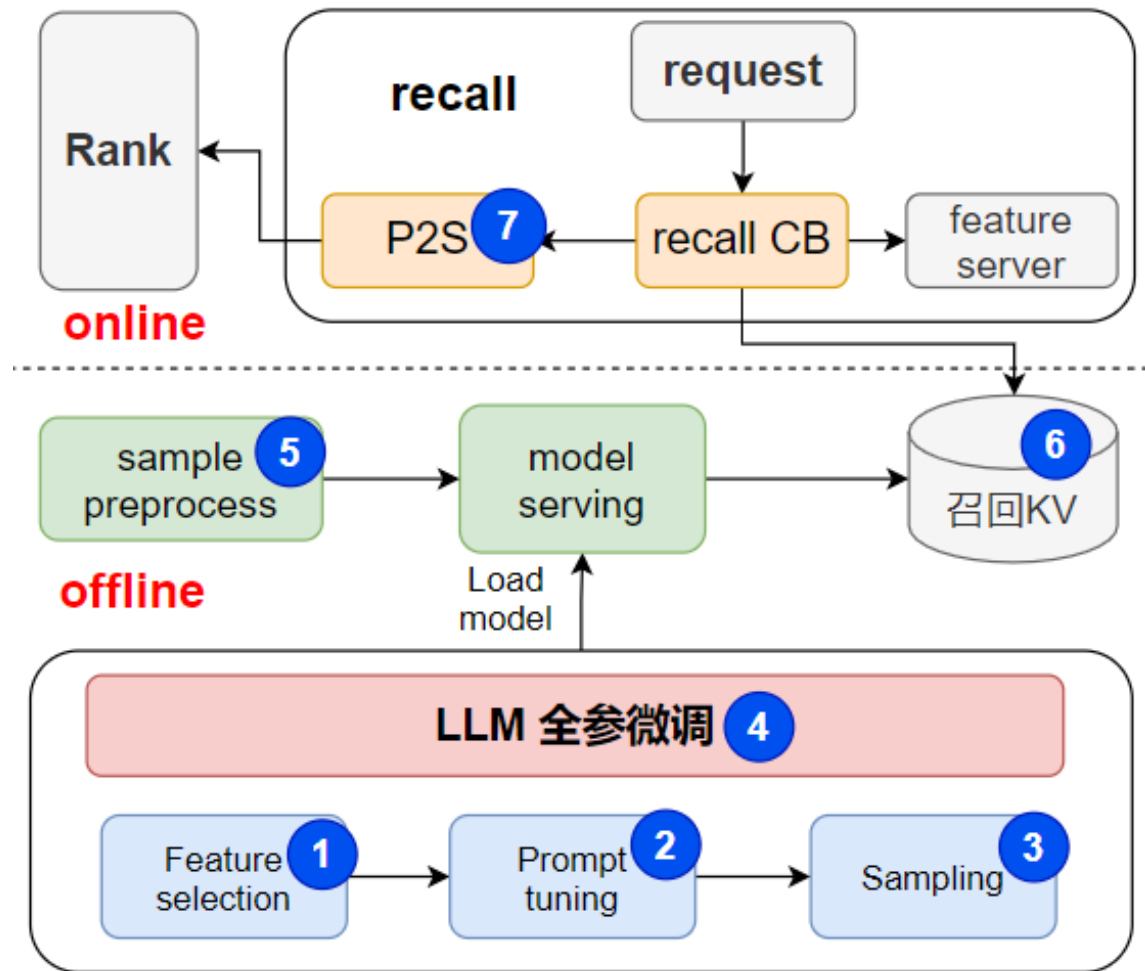
数学建模及系统实现pipeline

$$S_u^p = \{sku \mid EE(p_u, sku) \geq \varepsilon\}$$

$$p_u = \{G(u, cxt) \mid (f(P, G)) \geq \theta\}$$

- ✓ S_u^p 表示给定标品p, 召回u感兴趣的sku list
- ✓ $EE(p_u, sku)$ 是用EE的方式筛选得分高的sku
- ✓ p_u 表示给定用户u, 生成式的方式产出“可能”的标品集合
- ✓ $f(P, G)$ 表示生成商品和真实商品的匹配度
- ✓ $G(u, cxt)$ 基于LLM做任务对齐, loss function 最大化序列的似然, 负对数似然, 实现时等效于一个vocabulary size softmax的交叉熵损失

$$loss = - \sum_{c=1}^n y_{ic} \log p_{ic}$$



图：LLM4CB pipeline

LLM4CB – LLM解稀疏行为用户召回

JDH 京东健康

Prompt Engineering

- ✓ Instruction tuning for feature selection
- ✓ Trade off Efficiency and cover rate for online user
- ✓ Search a better performance of LLM

Automatic Prompt Engineering

感谢登录: zhangchao64

Prompt辅助生成/Meta Prompt

Prompt优化迭代/Prompt optimize evaluation

批量测试/Test multiple lines data

Prompt竞技场/Prompt arena

任务

请描述你的任务

你是一位健康商品电商平台的商品推荐员，请根据用户基础信息和用户行为，为用户推荐下一个可能感兴趣的产品

变量

请输入你的变量，每个变量占一行，如男没有，也可以不填

user_profile
user_action

生成Prompt/Generate Prompt

Prompt生成/Prompt Template Generated

You are an AI assistant working as a product recommender for a health product e-commerce platform. Your task is to recommend the next product that a user might be interested in based on their profile and past behavior.

Here are the inputs you will use:
<USER_PROFILE>
{\$USER_PROFILE}
</USER_PROFILE>

<USER_ACTION>
{\$USER_ACTION}
</USER_ACTION>

Follow these steps to make a recommendation:

1. Analyze the user's profile information provided in USER_PROFILE. This may include details such as age, gender, health goals, dietary preferences, and any other relevant information.
2. Review the user's past actions and behaviors provided in USER_ACTION. This may include their browsing history, past purchases, items added to the cart, and any other interactions with the platform.
3. Based on the analysis of the user's profile and actions, identify patterns or preferences that can help you determine the most suitable product recommendation.
4. Select a product that aligns with the user's interests and needs. Ensure that the product is relevant to their health goals and preferences.
5. Write your recommendation inside <recommendation> tags. Include a brief explanation of why you chose this product based on the user's profile and actions.

Here is the format for your response:

```
1 {  
2     "instruction": "你是一位电商平台商品推荐员，请根据用户信息和用户行为，为用户推荐他下一个可能感兴趣的产品", //指令型数据  
3     "input": "该用户，性别女，年龄36~40岁，婚姻状况未婚，居住城市广东省，职业为未知，学历为专科，未开通平台会员，用户价值中，收入水平中高，购买力中，属于小白领人群，并对箱包皮具品类感兴趣，喜欢在21:00~22:00点浏览商品。",  
4     "output": "龙胆泻肝片",  
5     "history": { }
```

Automatic Prompt Engineering

感谢登录: zhangchao64

Prompt辅助生成/Meta Prompt

Prompt优化迭代/Prompt optimize evaluation

批量测试/Test multiple lines data

Prompt竞技场/Prompt arena

请输入原始的Prompt

recommendation]
[Brief explanation of why this product is suitable for the user]
</recommendation>

Additional rules:
- Be polite and professional in your recommendation.
- Do not recommend products that are not related to health or wellness.
- Ensure that the recommended product is available on the platform.

Your goal is to provide a personalized and relevant product recommendation that enhances the user's shopping experience.

[可选]输入需要转换的变量

\$USER_PROFILE: 性别女，年龄36~40岁，婚姻状况未婚，居住城市广东省，职业为未知，学历为专科，未开通平台会员，用户价值中，收入水平中高，购买力中，属于小白领人群。
\$USER_ACTION: 近期点击了箱包皮具品类，喜欢在21:00~22:00点浏览商品

替换后的结果

product that a user might be interested in based on their profile and past behavior.

Here are the inputs you will use:
<USER_PROFILE>
{\$USER_PROFILE}
</USER_PROFILE>

<USER_ACTION>
{\$USER_ACTION}
</USER_ACTION>

Follow these steps to make a recommendation:

请输入需要评估的Prompt

You are an AI assistant working as a product recommender for a health product e-commerce platform. Your task is to recommend the next product that a user might be interested in based on their profile and past behavior.

Here are the inputs you will use:
<USER_PROFILE>
{\$USER_PROFILE}
</USER_PROFILE>

<USER_ACTION>
{\$USER_ACTION}
</USER_ACTION>

[可选]输入需要转换的变量

\$USER_PROFILE: 性别女，年龄36~40岁，婚姻状况未婚，居住城市广东省，职业为未知，学历为专科，未开通平台会员，用户价值中，收入水平中高，购买力中，属于小白领人群。
\$USER_ACTION: 近期点击了箱包皮具品类，喜欢在21:00~22:00点浏览商品

替换后的结果

You are an AI assistant working as a product recommender for a health product e-commerce platform. Your task is to recommend the next product that a user might be interested in based on their profile and past behavior.

Here are the inputs you will use:
<USER_PROFILE>
{\$USER_PROFILE}
</USER_PROFILE>

<USER_ACTION>
{\$USER_ACTION}
</USER_ACTION>

替换原始prompt中的变量

选择OpenAI模型
gpt-4o

替换修订prompt中的变量

选择Claude模型
anthropic.claude-3-sonnet-20240229-v1:0

Execute prompt

OpenAI Output

<recommendation>
美容保健品类的电动护肤保健品，如维生素E胶囊。这款产品有助于美容和延缓衰老，符合您作为职业小白领的需求。可以在繁忙的工作之余保持健康和美丽。近期浏览箱包皮具品类，显示您对高品质生活用品的兴趣，而维生素E胶囊同样具备高品质属性，能够补充生活必需，同时适合在夜间浏览时了解和购买。希望这款产品能为您的健康和美容提供进一步的支持。
</recommendation>

Claude Output

<recommendation>
根据您的个人资料和浏览记录，我建议您可以考虑购买一款时尚优雅的真丝丝巾或围巾。作为一位36~40岁的都市白领女性，一款优雅大方的丝巾不仅能彰显您的气质和品位，还能为日常着装增添一抹亮丽的色彩。丝巾材质柔软亲肤，既实用又能起到保暖作用，非常适合作为秋季的穿搭。此外，丝巾还能用作包包装饰，为您的整体造型增添一份时尚感。

我推荐这款产品是因为它符合您作为都市白领女性对时尚和品质的追求，同时也很实用。无论是上班还是休闲，一款优雅的丝巾都能让您焕发独特魅力。它不仅满足您对箱包皮具的兴趣，还能为您的生活增添一份优雅和品味。
</recommendation>

DataFun.

LLM4CB – LLM解稀疏行为用户召回

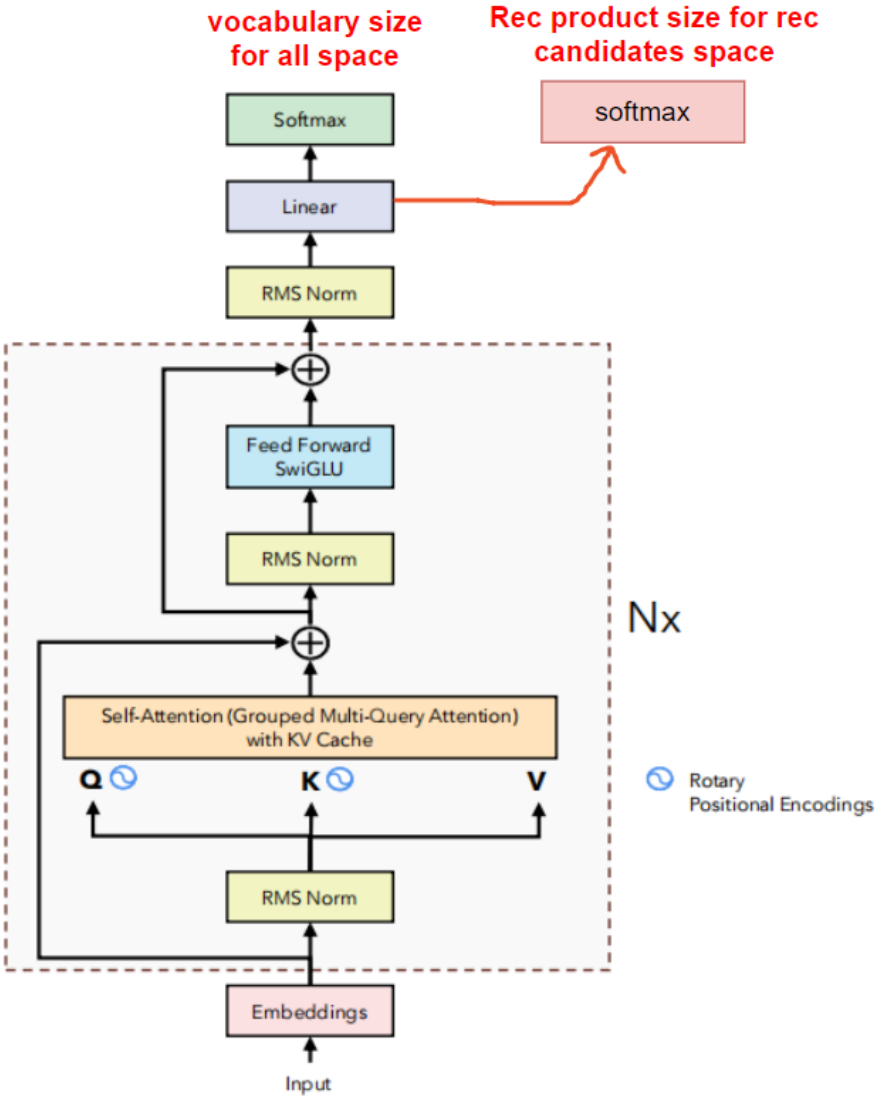
LLM Finetune

- ✓ 保留LLM 全量调参VS加候选池size softmax转多分类任务
- ✓ 小样本+保留世界知识 VS 大样本+更对齐推荐任务

Fine tune 选型原因：

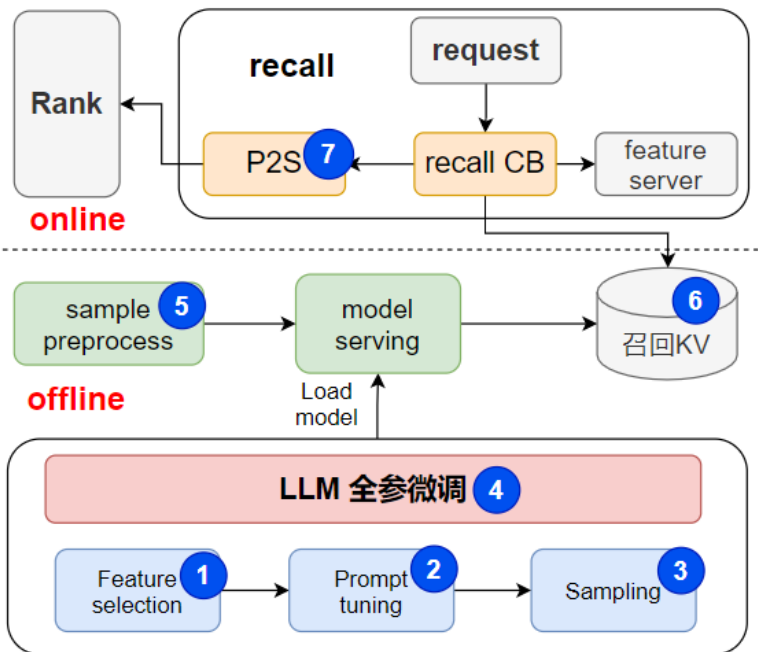
	LLM 全量调参	换softmax frozen LLM
优势	<p>能力保留度高：可以充分利用LLM的强大表达能力和大规模参数</p> <p>灵活度高：能够适应不同的推荐任务和数据格式</p>	<p>简化问题：将推荐任务转化为多分类问题，简化了模型的输出和目标。</p> <p>计算效率高：只需在LLM基础上增加一层softmax，计算开销较小</p>
劣势	<p>计算资源消耗大：全量调参需要大量计算资源和时间，尤其是对于大规模LLM。</p> <p>复杂性：需要处理tokenization和序列化，可能增加实现复杂性。</p>	<p>信息丢失：可能无法充分利用LLM的上下文理解能力，特别是对于复杂的推荐场景。</p> <p>扩展性差：如果候选池item数量变化，模型需要重新训练或调整。</p>

本质是在垂域空间和世界知识做选择，当然，更牛逼的是都要，那就是我本身的任务就是同LLM量级的样本和网络，完成一个完美的垂域世界知识



LLM4CB – LLM解稀疏行为用户召回

Other tips



Evaluate

- ✓ AB – 点击下单均有相对百分比个位数增长
- ✓ Hitrate/请求覆盖率 - 离线优化调参

1 Feature selection

- ✓ Low cardinality
- ✓ 贴近自然语言表达的特征
- ✓ 覆盖率高

3 Training sampling

- ✓ 多样性: Multi type of product label
- ✓ 不同类型样本比例 – 实验
- ✓ 样本量 – 实验

5 Inference sampling

- ✓ 覆盖率: 拉长样本, N-1对第N天覆盖
- ✓ 性能: infer样本去重

6 Infra methods

- ✓ 时效: 实时, 离线, 近线

7 Strategy methods

$$hits_{P_o}^{P_g} = f(P_g, P_o)$$

$$simi_{sku}^p = EE(p_u, sku)$$

- ✓ 商品 map 候选池产品名
- ✓ 产品名 到 sku的策略 EE优化

Deepl2l – scale up makes i2i deeper and wider

解决什么问题

- ✓ 一路i2i满足各种长尾场景
- ✓ 低频item充分建模
- ✓ Trigger sku 的相似sku 充分拉长

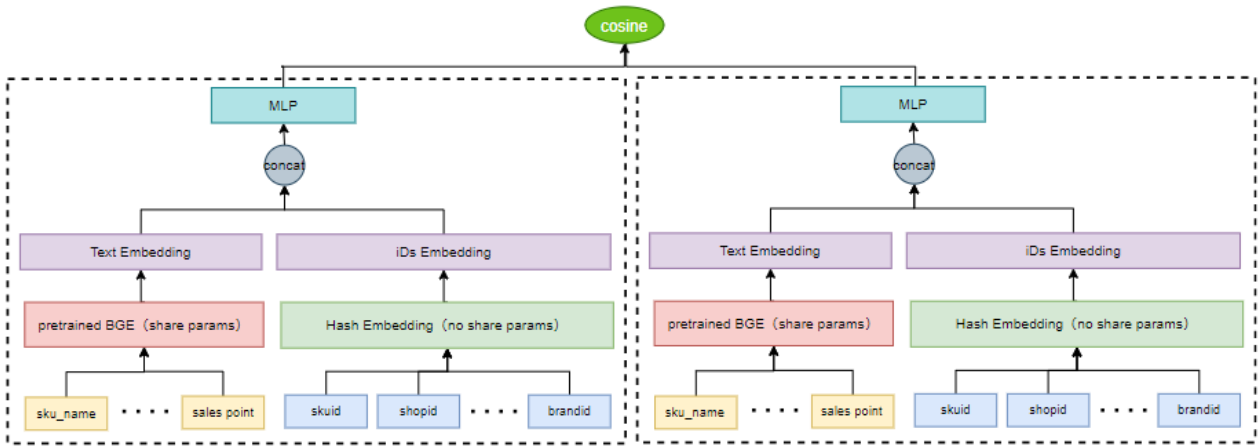
Scaling law 启示解决i2i充分建模

- ✓ 样本 * 参数量 * epoch

wider

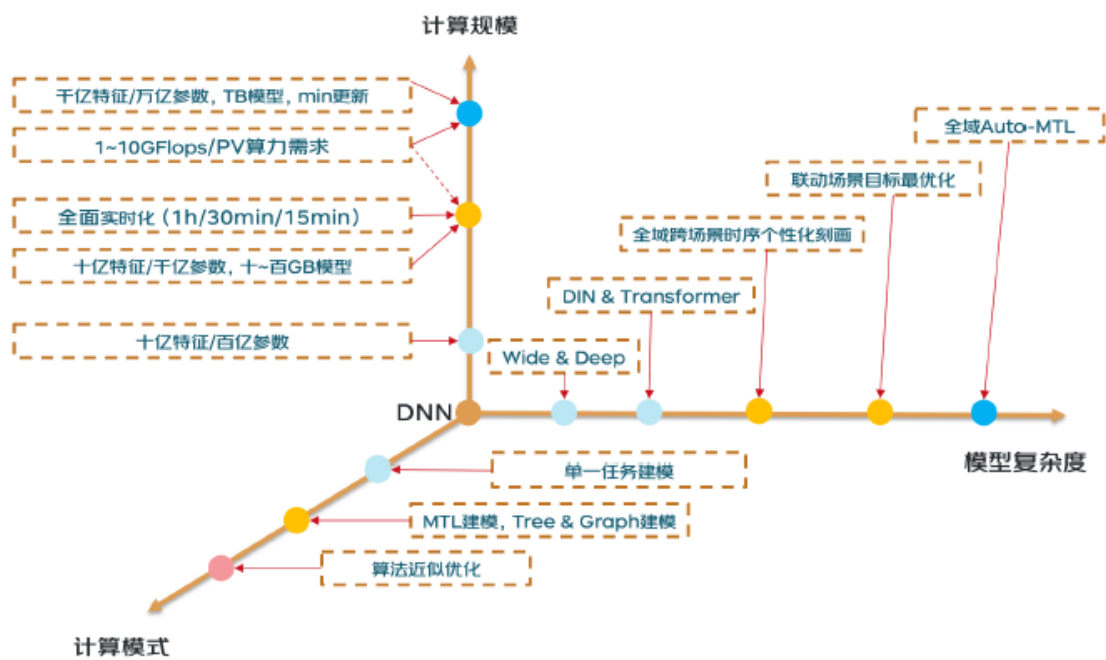
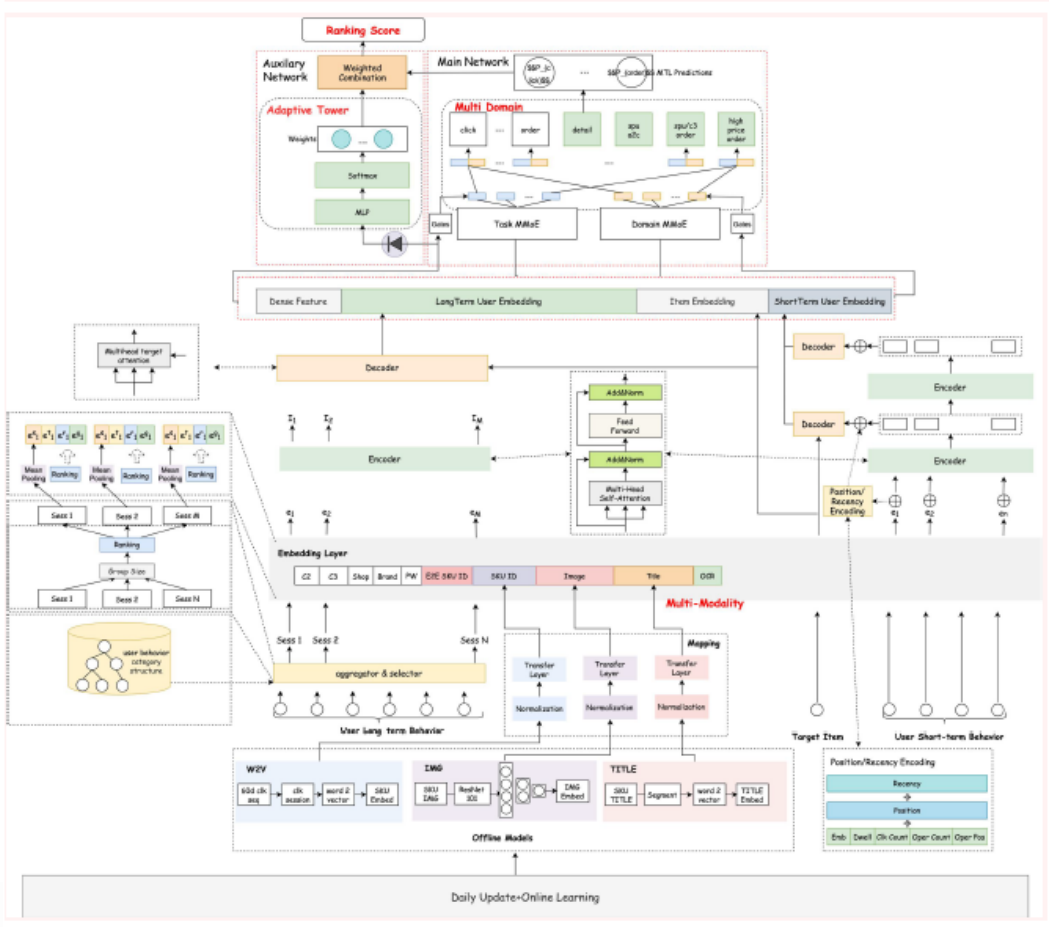
key	Value list			
Sku1	Sku_n	Sku_n+1		
sku2	Sku_m			
...
Sku_i	xx	xx	xx	xx

样本	1、原始日志 拓宽source*拉长时长 2、GNN 随机游走，构造低频item样本 3、数据增长，利用LLM
参数	1、IDs and text 均充分表达 2、扩side info 3、pre-trained model 4、dim, layer wider and deeper.



大模型CTR – 传统路子践行scaling law

- 精排：传统路子，ctr model变大
- 长序列、多模态、多目标、多专家、多xx



不展开讲，贴2个JD集团兄弟团队的图，不是今天的重点

生成式范式 取代 传统多级过滤+判别范式的大范围落地可行性？

- AI上一个技术范式之争：专家系统 VS data driven，树和马的例子 – 计算相似 vs 认知相似
- 电商推荐系统复杂性，非单一技术问题，供给、分发、营销种草和增长为导向的系统性工程
- Scaling law for xx 会成为未来一段时间的主流

THANKS

欢迎投递简历，团队现有推荐、
搜索、大模型方向的算法HC

联系方式：
zhangchao64@jd.com