

12月第二次汇报

刘寰硕, 2022/12/9



目录

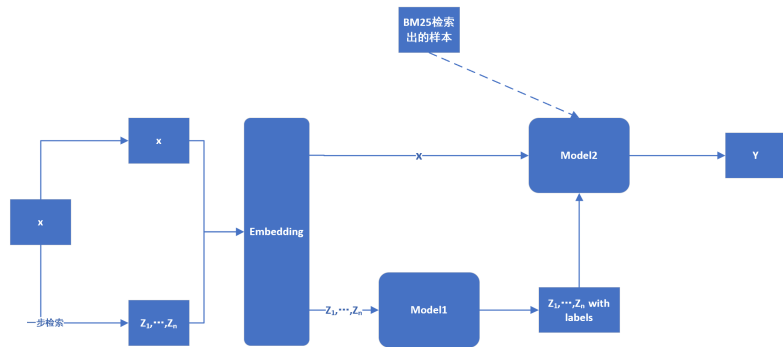
1. 课题背景
2. 原方案
3. 下一步计划

课题背景

1. 缩短RIM的检索时间使之落地
2. 引入更精细的特征建模方式提升效果

原方案

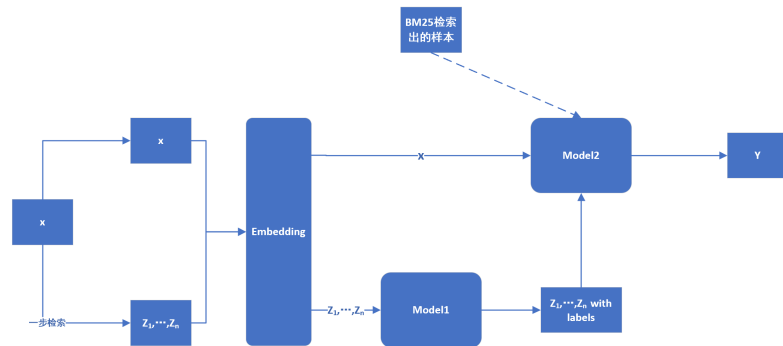
1. 参考: GAN, Actor Critic, 知识蒸馏
2. 思想: 我们不去检索相似的样本, 而是利用user, item本身的性质去构造样本
 - 假定来了一条instance (u, i)
 - 找到 u 曾经访问过的所有item \mathcal{I} (取10个)
 - 找到所有访问过 i 的user \mathcal{U} (取10个)
 - 将 $\forall (u, i) \in \mathcal{U} \times \mathcal{I}$ 作为 (u, i) 的邻居丢入Gen
 - 取Gen中分数排前25%和后25%的以及实际出现过的作为丢入Dis中的“邻居”
3. 网络:
 1. Dis: RIM(一个能够Aggregate邻居特征的网络)
 2. Gen: DeepFM(一个足够简单, 足够快能够产生label(置信度)的网络)



原方案

遇到的问题

1. 处理流程非常复杂且不智能
 1. u 或 i 一步检索的对象不够预设的数量甚至为空时，需要对user, item本身进行padding
 2. 一个user id或者一个item id会有对应多个特征的情况。比如一个物品属于多个类(不合理但是有这种情况)，需要从类中采样
 3. 构造样本时需要进行笛卡尔积，同时会出现真实label和预测label混在一起的情况
2. 最重要：数据量非常大，即使是使用稀疏矩阵存储也只能减少存储空间而无法加快运算
 1. 放在dataset里训练非常慢
 2. 参考RIM预存数据和我们想要的结果会有些不太一样



原方案

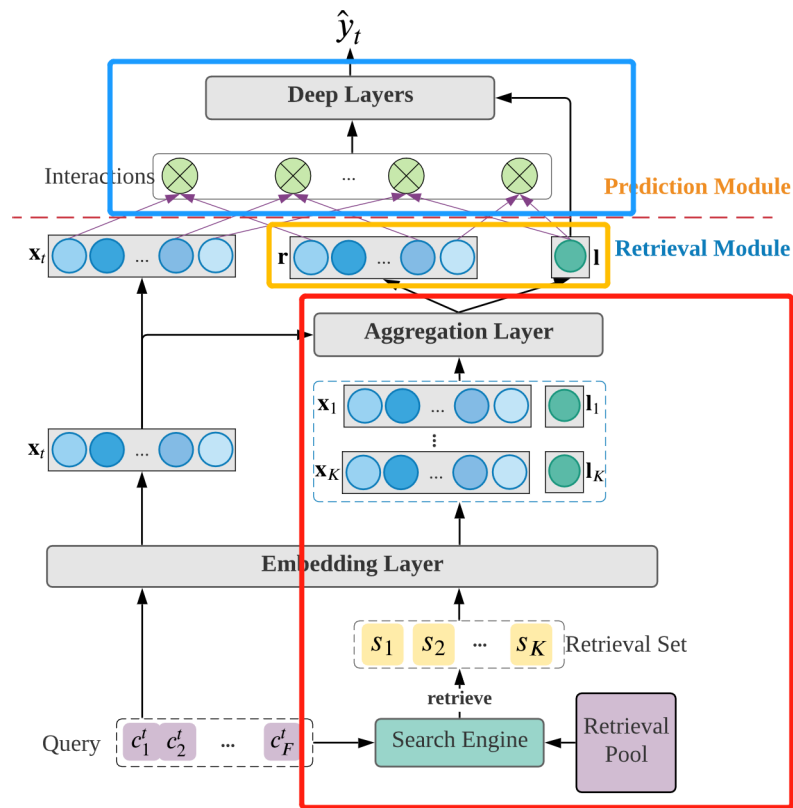
新方案

1. 思路：直接学习RIM的聚合特征，将这个任务转换成一个预训练任务，最终希望得到预训练好的embedding和聚合特征，兼具性能和其他模型的兼容性

1. 利用一个简单的网络代替红框的部分(检索+聚合)去学习黄色框中的向量
2. 将蓝色框中的向量换成任何一种模型 (DeepFM,xDeepFM,DIN,DIEN)

2. 步骤：

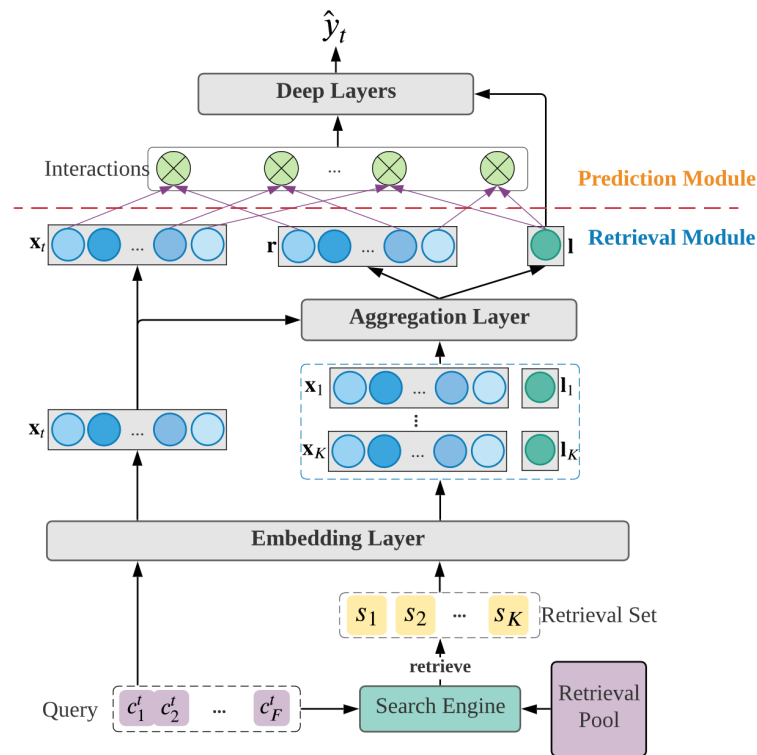
1. Pretrain
2. Further Pretrain
3. Finetune



新方案(图片来自RIM原文)

Pretrain

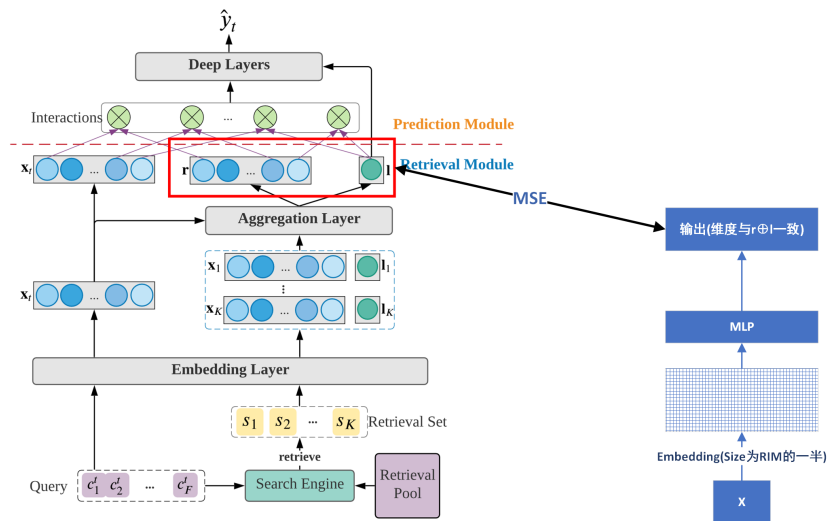
预训练一个RIM



Pretrain(图片来自RIM原论文)

Further Pretrain

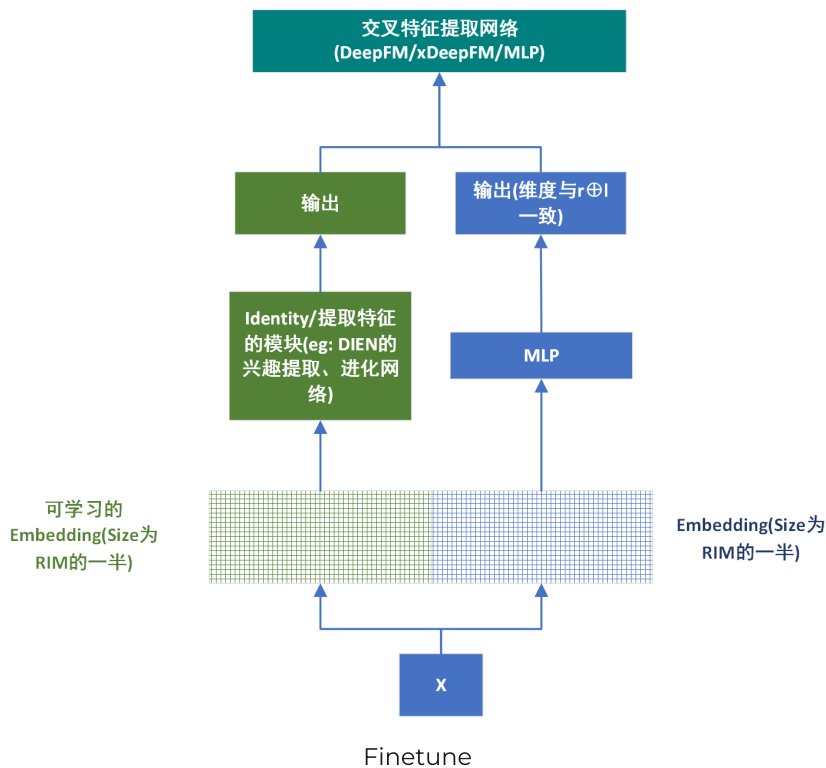
1. 用Embedding+MLP去逼近RIM中的聚合特征(红框的部分)
2. Embedding Size是原生RIM的一半
3. 可以引入对比学习、无法引入GAN



Pretrain(图片来自RIM原论文)

Finetune

1. 最终的Embedding Layer是由可学习的Embedding(Size为RIM的一般)和Further Pretrain中预训练出的Embedding组合而成的
2. 右边蓝色的网络是固定的，左边绿色的部分是可学习的；分别建模短期和长期(检索)特征
3. 最终预期的推理和训练时间不会比普通模型多太多



下一步计划

1. 完成不同模型的效果测试
2. 读更多文献看看是否有方案可以借鉴