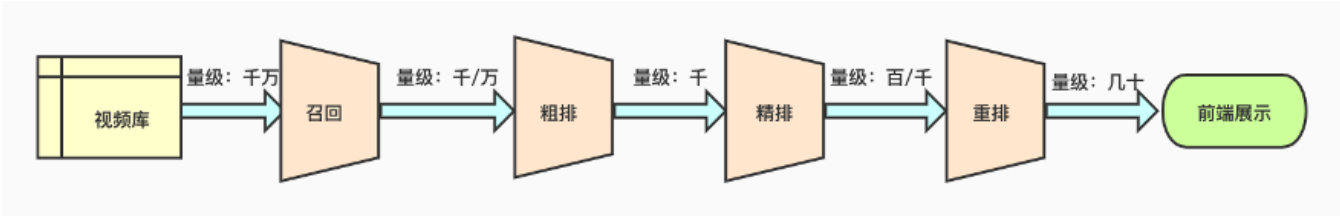
爱奇艺短视频推荐：粗排篇

随刻基础推荐团队 DataFunTalk 2月27日

导读 工业界的推荐系统通常包括召回、粗排、精排以及重排四个阶段，如图一所示，每个阶段都像是一个漏斗，从海量的物品集合中过滤出用户最有可能感兴趣的物品。其中粗排模型发挥的主要作用是统一计算和过滤召回结果，在尽量保证推荐准确性的前提下减轻精排模型的计算压力。本文主要介绍爱奇艺随刻基础推荐团队在短视频推荐业务的粗排模型优化上落地的一系列实践方案。



图一：推荐整体流程架构图

背景

工业界在做粗排模型选型时，性能通常是一个很重要的考量因素。按照工业界选型粗排模型的发展历程，大致可以将粗排模型分为以下几大类：

1. 最早也是最简单的粗排过滤方法，直接根据召回计算的得分做截断，控制输入给精排模型的物品候选数量，或者根据**全局的ctr**等统计指标做统一截断。

2. 以LR/决策树为代表的，结构比较简单又有一定个性化表达能力的**机器学习模型，**统一对召回候选集做打分截断。

3. 当前工业界应用最广泛的粗排模型—基于向量内积的**双塔DNN模型，**两侧分别输入用户特征和物品特征，经过深度网络计算后，分别产出用户向量和物品向量，再通过向量相似度等计算得到排序分数。

爱奇艺短视频推荐业务**最初采用的粗排模型**可以归为上述第二类选型模型，是一个基于各个纬度统计特征的GBDT模型。统计特征维度主要包括下面几个维度：

1. 不同属性的用户群体对不同类型视频（分标签、创作者和视频本身等）的消费统计特征。

2. 视频维度累积的消费统计特征，如视频的点击率、时长消费中位数和均值等；创作者up主的消费统计特征以及视频标签的消费统计特征等。

3．用户历史消费的视频内容统计特征，如用户历史消费的类型标签统计、消费的创作者内容统计等。

在业务的精排模型优化升级为**wide&deep模型**后，我们对粗排模型和精排模型的预估结果做了详细的统计和分析，发现粗排模型预估为top的头部视频和精排模型预估的头部视频有很大的差异。归咎原因主要是以下两方面的原因：

1．特征集合的差异：粗排GBDT模型中主要是一些稠密类统计特征，而精排wide&deep模型中发挥重要作用的特征主要是用户长短期消费的**视频id、视频tag、up主id**等以及视频本身的**id、tag和up主id**等稀疏类型特征。

2．模型结构的差异：树型结构模型和DNN模型的优化和拟合数据时的侧重点还是有很大的差异的。

除了预估结果和**精排wide&deep模型**有比较大的差异性外，**GBDT模型**在特征处理和挖掘方面还需要投入大量的人力。综合以上分析，为了尽量弥补粗排模型和精排模型的Gap，缩小粗排模型和精排模型预估结果的差异性，并节省大量特征统计和挖掘的人力成本，我们对粗排模型进行了**一系列的升级和优化。**

双塔DNN粗排模型

综合计算性能和实验效果，我们最终也选择了目前主流的双塔DNN模型作为我们最终选型的粗排模型。我们双塔DNN模型结构如下图二所示。用户侧和Item侧分别构建**三层全联接DNN模型，**最后分别输出一个多维（512）的embedding向量，作为用户侧和视频侧的低维表征。

在构建粗排模型特征集合时，为了控制粗排模型参数的复杂度，我们对粗排的特征集合做了大量的裁剪，用户侧和视频侧都只采用了少部分精排模型的特征子集。其中，用户侧特征主要选取了下面**几维特征：**

1．用户基础画像特征、上下文特征如手机系统、型号、地域等。

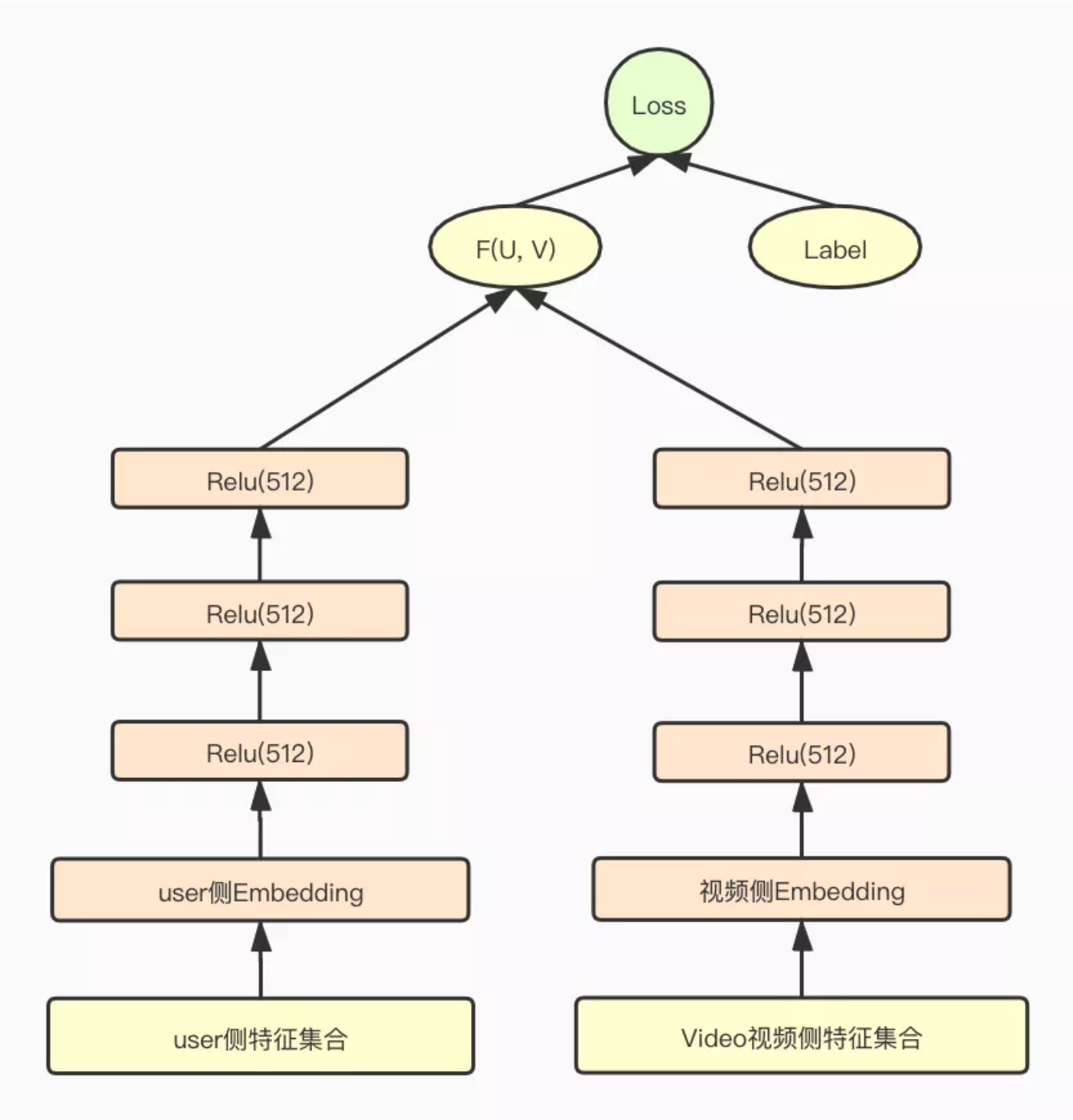
2．用户的历史行为特征，如用户观看的视频ID、up主ID，以及观看视频的关键词tag等，以及**用户session内的行为特征**等。

视频侧特征只保留了三维：

1．视频ID

2．up主ID

3．视频标签



图二：粗排双塔DNN模型结构图

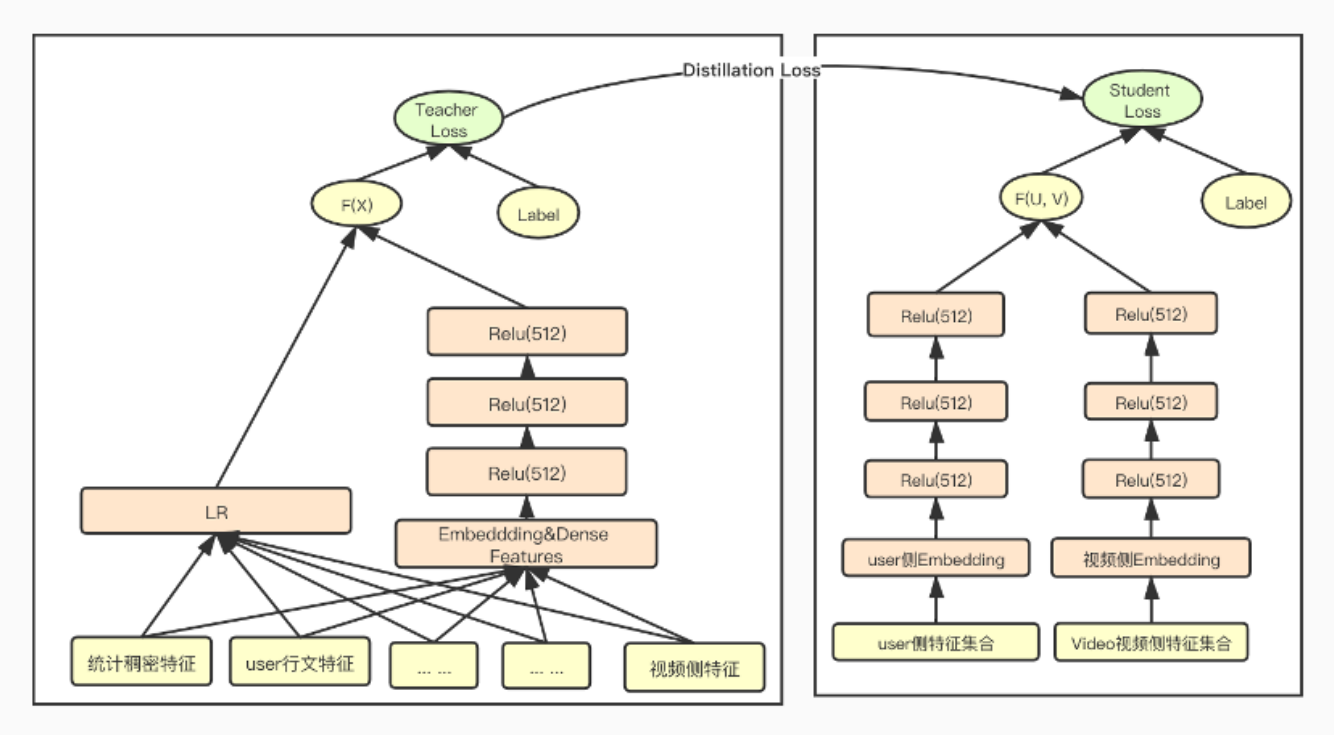
从GBDT到双塔DNN，模型的复杂度和参数量级还是有**爆炸性增长**的，为了不损失粗排模型的精度同时满足上线的性能指标要求，我们从下面几个方面做了很多优化工作：

1．知识蒸馏

为了弥补特征裁剪带来的损失，保证裁剪后粗排模型的精度，我们在训练粗排模型时，采用了**模型压缩**常用的方法-**知识蒸馏**来训练粗排模型。

知识蒸馏是一种模型压缩常见方法，在**teacher-student框架**中，将复杂、学习能力强的网络学到的特征表示“知识蒸馏”出来，传递给**参数量小、学习能力弱**的网络。从而我们会得到一个**速度快，能力强**的网络。

将粗排模型和精排模型放到知识蒸馏的**teacher-student框架**中，以蒸馏训练的方式来训练粗排模型，以精排模型为teacher来指导粗排模型的训练，从而得到一个**结构简单，参数量小，但表达力不弱**的粗排模型，蒸馏训练示意图如图三所示。



图三：粗排模型蒸馏训练示意图

在蒸馏训练的过程中，为了使粗排模型输出的logits和精排模型输出的logits分布尽量对齐，训练优化的目标从原来单一的粗排模型的logloss调整为如公式一所示的三部分loss的加和，包括**student loss**（粗排模型loss）、**teacher loss**（精排模型loss）和**蒸馏loss**三部分组成。

其中蒸馏loss我们线上采用的是粗排模型输出和精排模型输出的最小平方误差，为了调节蒸馏loss的影响，我们在该项loss前又加了一维**超参lamda**，我们设置超参lamda随着训练步数迭代逐渐增大，增强蒸馏loss的影响，在训练后期使得粗排模型预估值尽**量向精排模型对齐**，lamda随着训练step的变化趋势如图四所示。

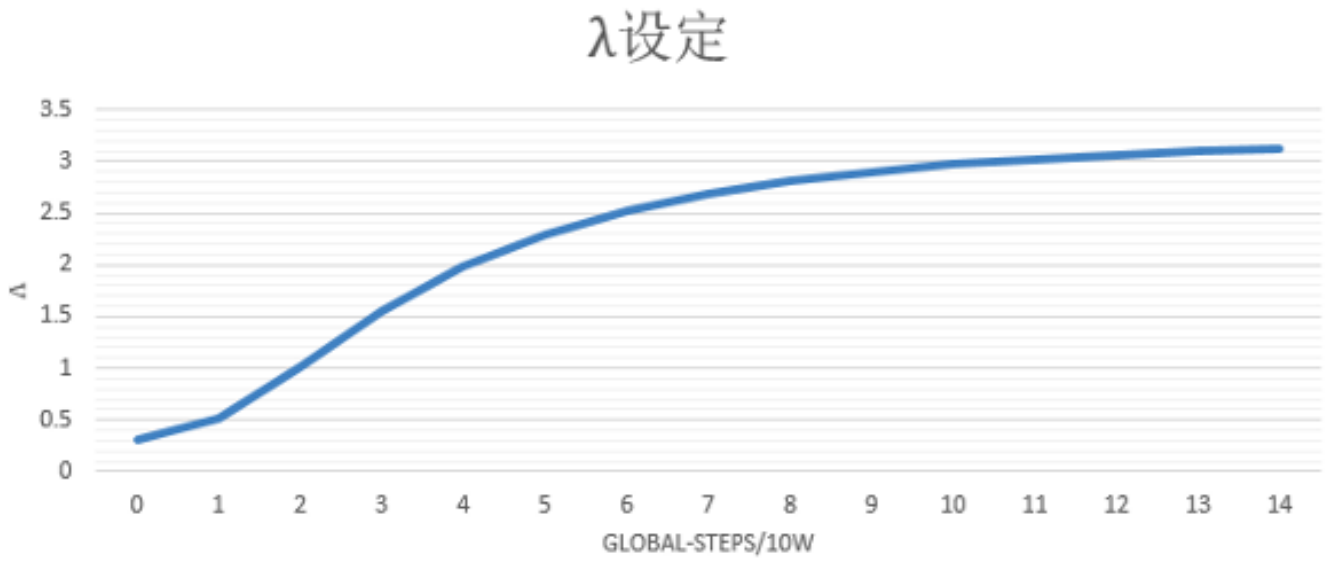
$$minL_s(y, f(X; W_s)) + \lambda * L_d(f(X, X^*; W_t), f(X; W_s)) + L_t(y, f(X, X^*; W_t))$$

图四：Distillation loss超参lamda设定

2. Embedding参数优化

为了进一步减少参数量级，压缩模型大小，提高模型传输速度，同时减少加载模型时的内存消耗，我们在训练粗排模型时，将模型**embedding**参数的优化器替换为稀疏解优化神器FTRL，其他层的参数依然还用AdaGrad。这一步调整不仅离线Auc有小幅提升，训练得到的粗排模型的embedding参数全为0的比率也**高达49.7%**之多。我们在做模型导出时，裁剪掉了全为0的

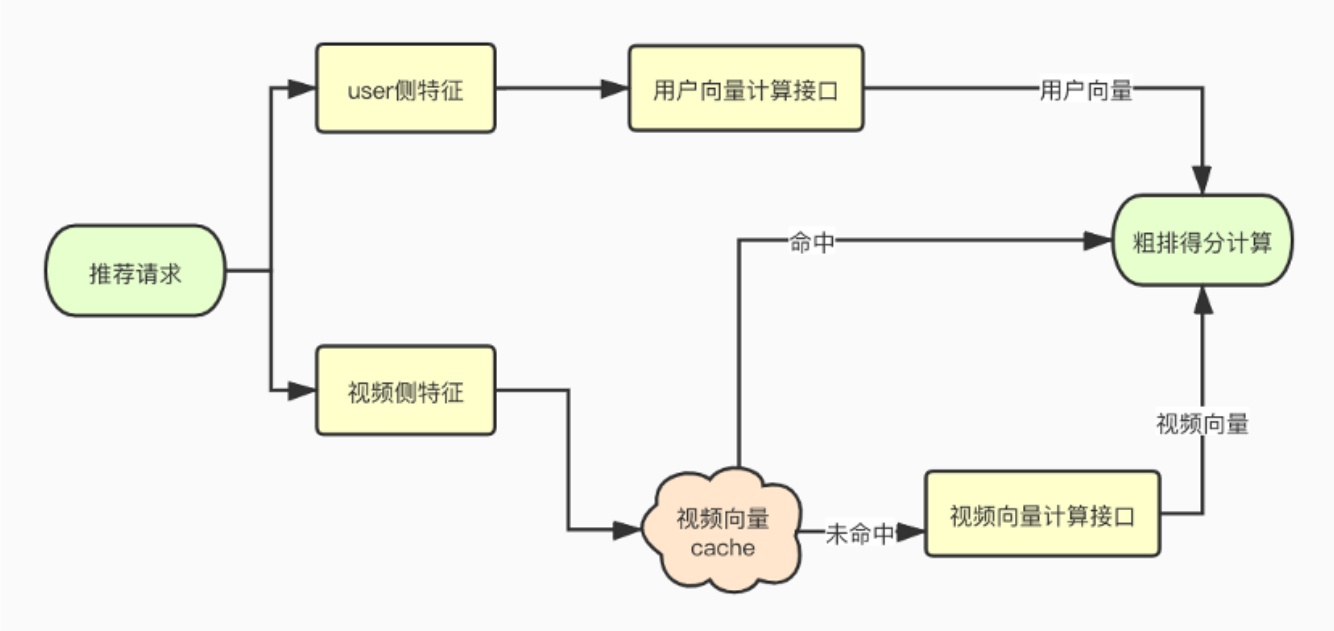
embedding参数，粗排模型大小**减小了46.8%**，使得模型的传输速度也有近一倍的提升，同时线上加载模型的内存消耗也**降低了100%**。



3.线上Inference优化

除了离线训练时的一系列优化，我们对双塔粗排模型的线上inference计算也做了很多**计算去重和优化**。首先在计算user侧embedding时进行了合并计算，对同一个用户+千级别候选视频pair，在计算user侧embedding时，将user侧的特征独立拆分，**只过一次user侧NN的inference**。这个优化使得粗排模型打分计算服务p99的计算耗时**减少了19ms**左右。

此外，基于视频推荐时特别长尾的分布，以及如前文所述，粗排模型视频侧的特征全部是静态特征（视频id确定，特征也是确定的），我们对高频视频的embedding进行了缓存。视频侧embedding优先从缓存里查询，未命中缓存时再进行inference计算。优化后的粗排打分服务架构如下图五所示：



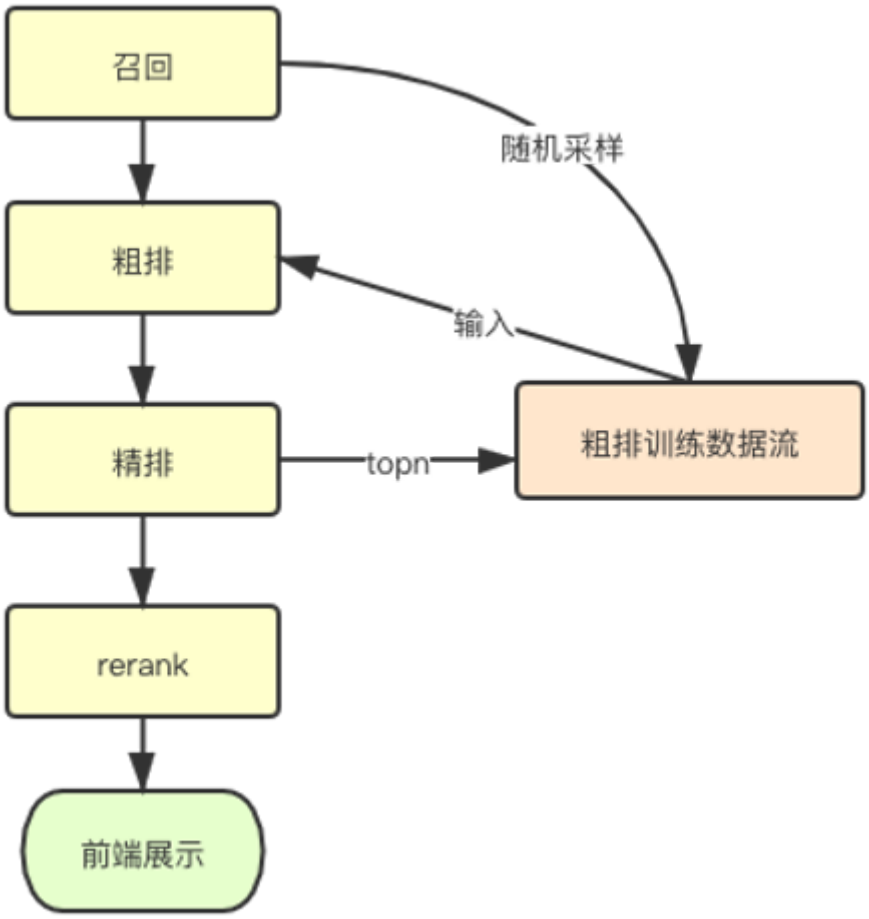
图五：粗排打分服务架构图

经过上述一系列的优化，粗排双塔DNN模型的线上计算性能对比之前**GBDT模型**基本可以持平，分别应用到了**爱奇艺热点频道**和**随刻首页推荐场景**中，两边的线上指标都有显著提升，其中随刻首页feed推荐的用户人均消费时长**涨了3%点**左右；爱奇艺热点频道的人均消费时长也**涨了1%**，ctr和人均消费视频数也都有**2%的涨幅**。

级联模型

随着后续业务的迭代和逐步深入，精排模型的学习目标也不断做调整，为了方便添加不同的学习目标，我们将线上的精排模型升级为了由Google提出的，目前也在工业界广泛使用的**MMOE多目标模型**。为了一劳永逸地解决粗排模型目标和精排模型目标一致性的问题，我们对粗排模型进行了又一次的迭代优化。通过升级为级联模型，使得粗排模型能够自适应地对齐精排模型目标的变化，同时也**节省了蒸馏训练的环节**，大大地**节省了训练资源消耗**。

从实践的角度，级联模型对模型结构以及模型输入的特征集没有做任何修改，只是调整了粗排模型训练样本的生成方式，升级后的粗排模型从学习线上真实曝光点击/播放样本，调整为**直接学习精排模型**的预估结果，将精排模型预估topn的结果作为粗排模型学习的正样本。级联模型的样本生成方式具体如图六所示。



图六：级联模型训练样本数据流

所以这一次升级粗排模型到级联模型只是简单地调整了模型的训练样本，却取得了很显著的收益，除了去掉了蒸馏学习的环节，大大地**缩减了粗排模型训练的资源和时间消耗**，实际上线也取得了很显著的收益，视频推荐场景的曝光点击率和人均有效观看视频数都**提升了3%左右**。同时各项互动指标也有显著提升，其中人均评论量有**12%的显著提升**。

未来规划

以上是我们近期在短视频推荐粗排模型上尝试的一系列优化，实践证明粗排模型和精排模型的一致性对线上效果还是有很大影响的。后续我们会从以下几个方面持续优化粗排模型：

1. 尝试面向下一代的粗排排序系统——**COLD**。

2. 持续优化粗排模型线上计算的性能，在性能允许的情况下，扩大召回的视频数量，同时添加更多在精排模型验证有效的特征到粗排模型，**提升**粗排模型的**准确性**。

3. 优化**user embedding**和**视频embedding**的相似度计算，考虑增加一个浅层网络来计算user和item的相似性，替换目前简单的cosine相似度计算。

参考文献

1.https://www.kdd.org/kdd2018/accepted-papers/view/modeling-task-relationships-in-multi-task-learning-with-multi-gate-mixture-

2.H.B.McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and L1 regularization. In AISTATS, 2011

3.https://arxiv.org/abs/1503.02531

4.https://arxiv.org/abs/2007.16122

今天的分享就到这里，谢谢大家。

在文末分享、点赞、在看，给个3连击呗~

团队介绍：

爱奇艺随刻事业部基础推荐团队，负责随刻app首页短视频feed和爱奇艺随刻热点feed的推荐策略优化。