网易新闻推荐：深度学习排序系统及模型

薛海霞 深度学习团队

2019.6.15

信息流场景中个性化推荐的形态

文章

视频

图像

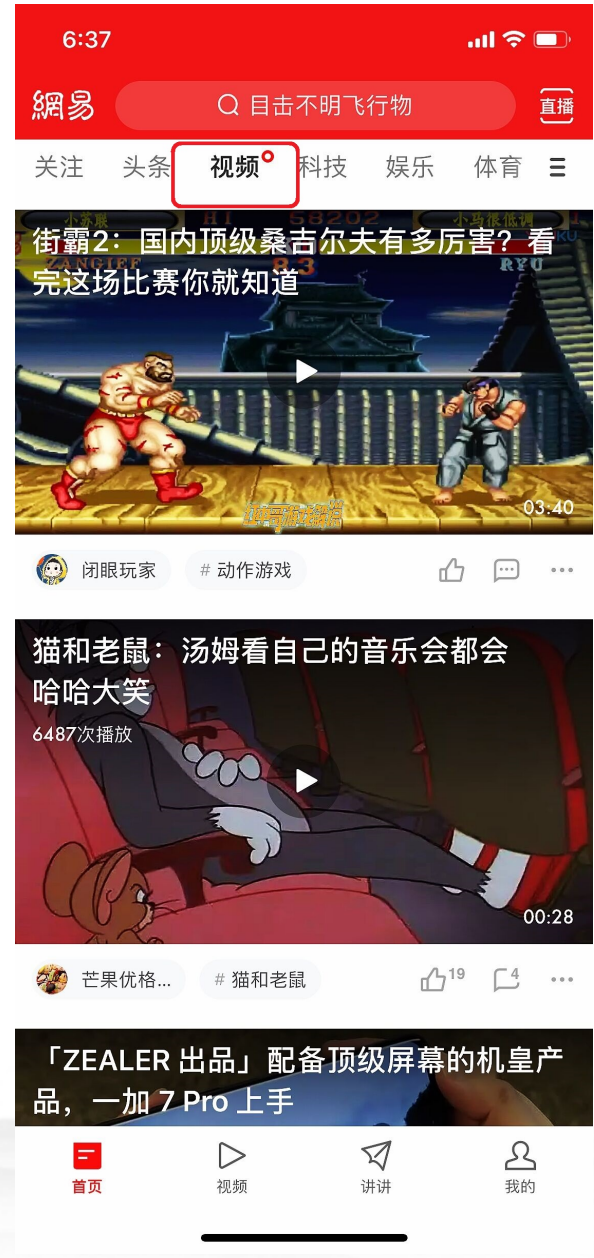
用户画像

场景

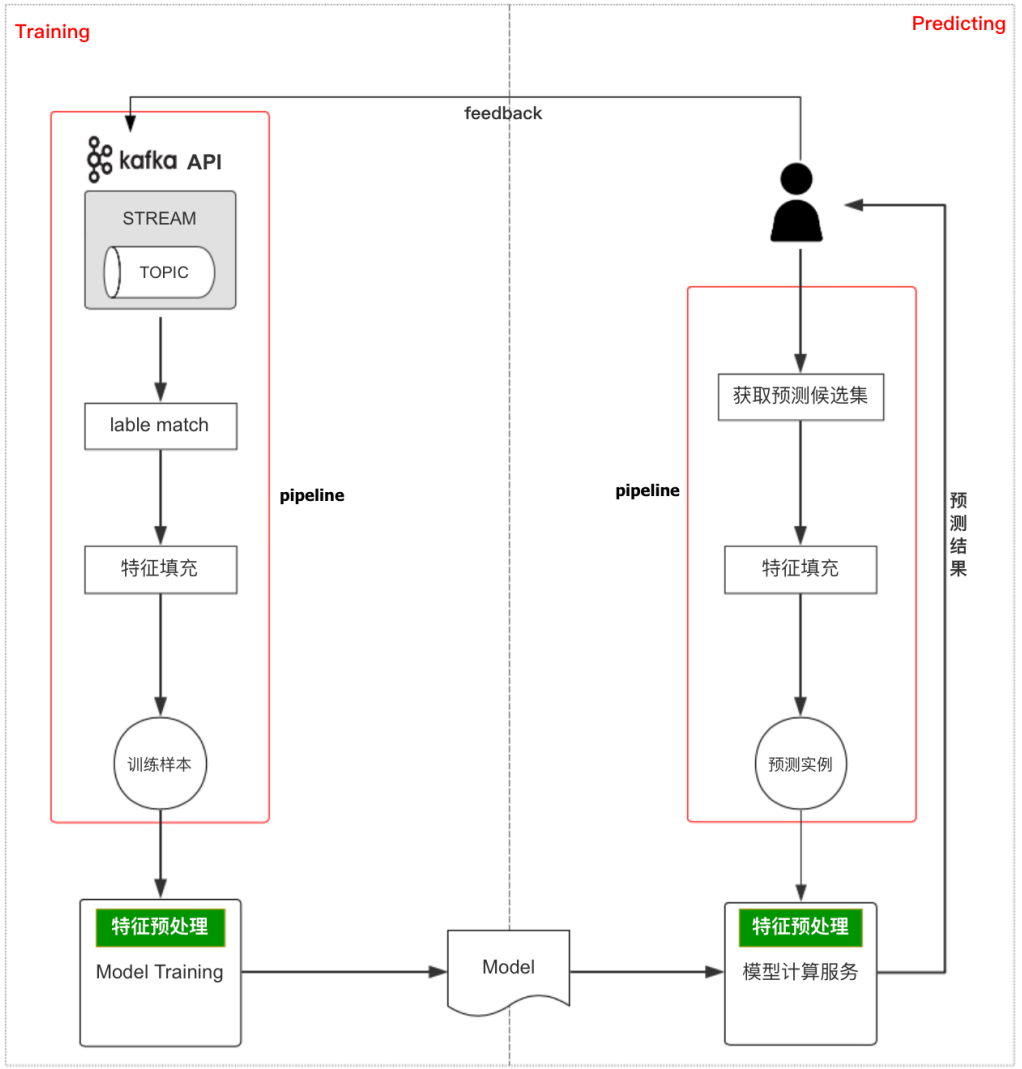
关注关系

.....

$$Pr(click|item, user, context)$$

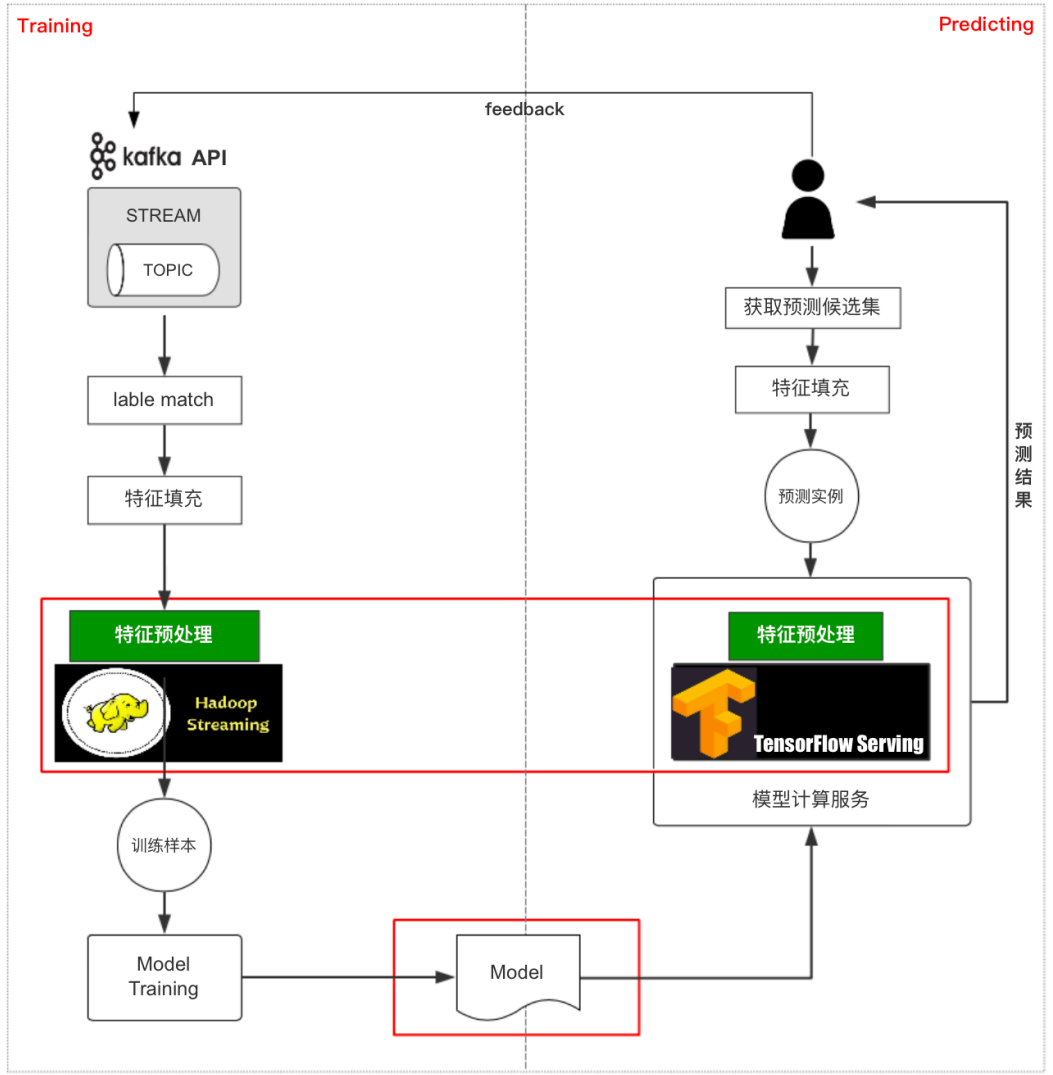


排序系统的基本流程



排序系统的三部分：

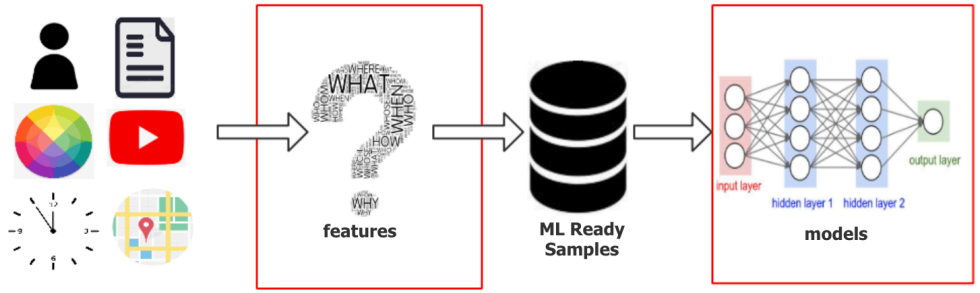
- Pipeline
- 排序模型
- 模型计算服务



当前排序系统面临的核心问题

1. pipeline：

- 线下训练和线上计算的性能问题
- 保证线上线下一致性，尤其是特征处理



2. 深度学习排序模型：

- 通用性 & 可扩展性，以快速支持不断增加的业务场景需求，支持模型的快速迭代
- 灵活性，根据业务的需求对模型做定制化

pipeline优化 — 特征处理框架

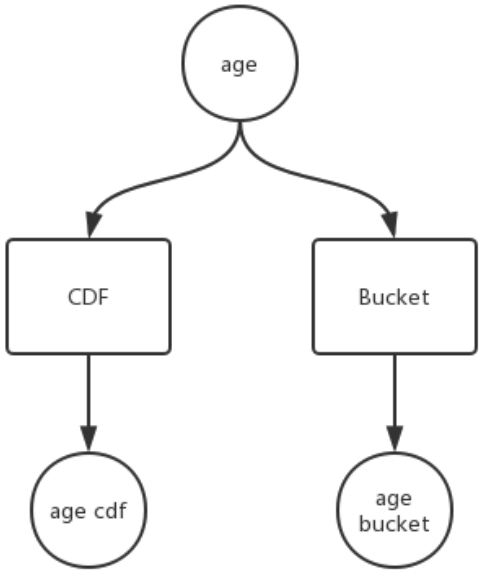
1. 特征处理库：

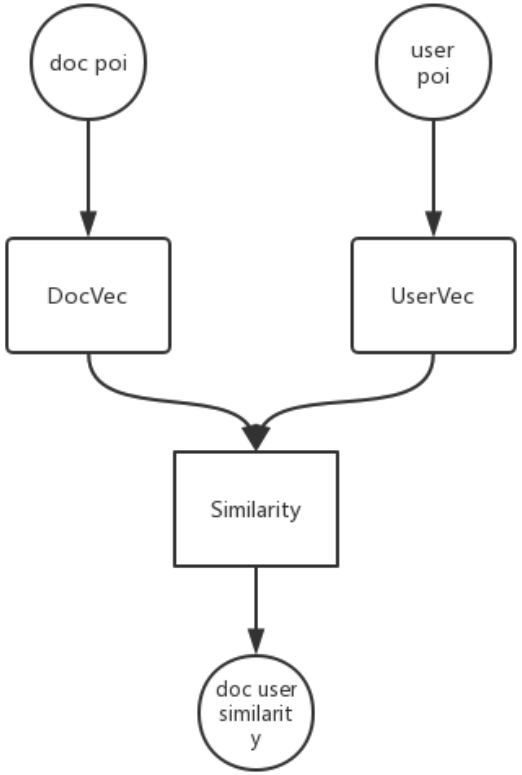
- 将训练阶段中需要跑在CPU上的特征处理任务转移到pipeline阶段的大规模CPU计算集群上，以此避免在模型训练阶段对GPU计算的影响，同时也避免了训练阶段的重复计算。
- 将特征处理库（内含一系列特征处理算子）独立出来，线下部分在pipeline阶段调用，线上部分在模型中调用此库

2. 自定义样本读取和数据处理模块：

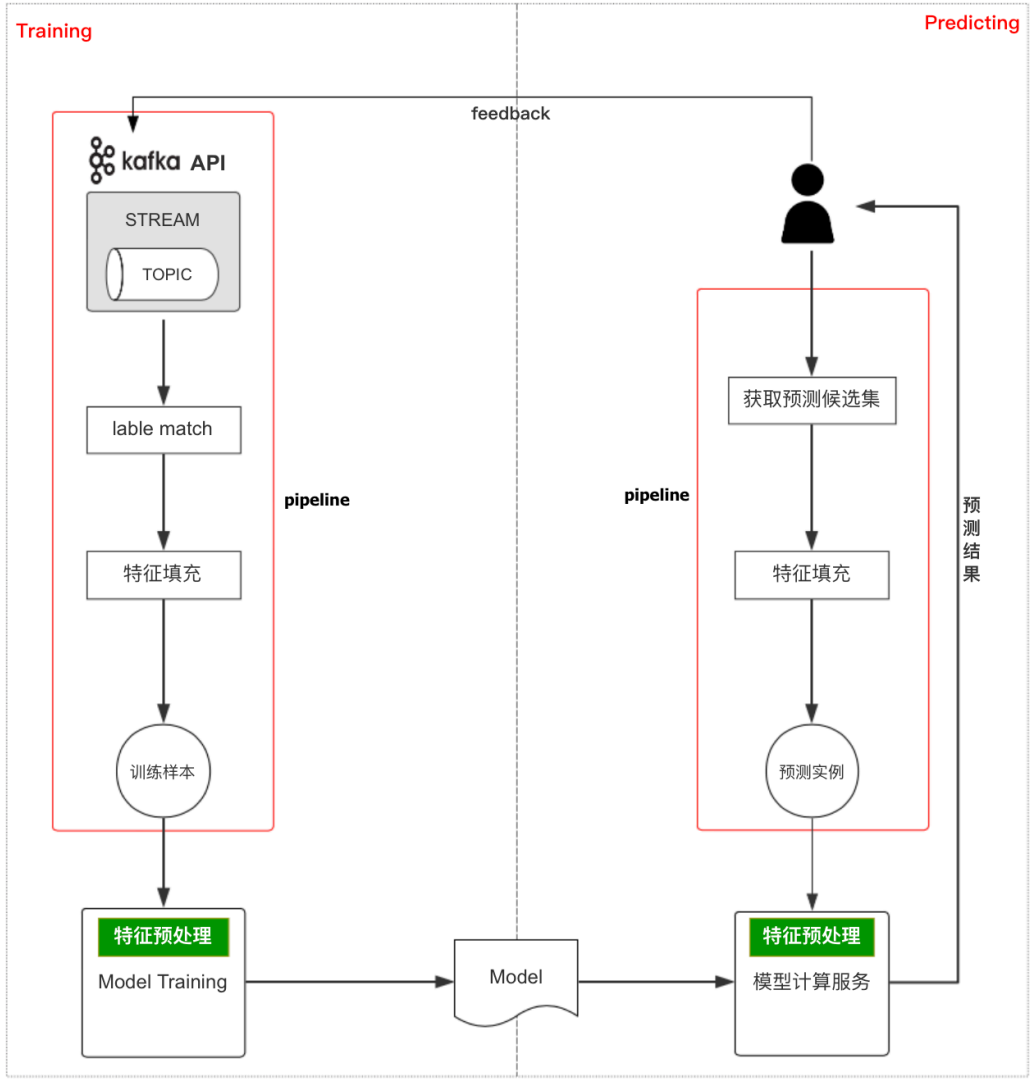
- 原生Tensorflow通常使用tf.data的接口进行数据读取，通过feature_column进行数据预处理，但是在大规模数据场景中，存在性能瓶颈，因此我们重新实现了数据读取和预处理模块，优化了性能
- 为了支持多值带权的特征，我们使用了自定义的样本格式，原生接口对样本格式的解析也并不友好，所以自定义解析模块
- 想要手动融合一些操作或TensorFlow原生不支持的一些操作

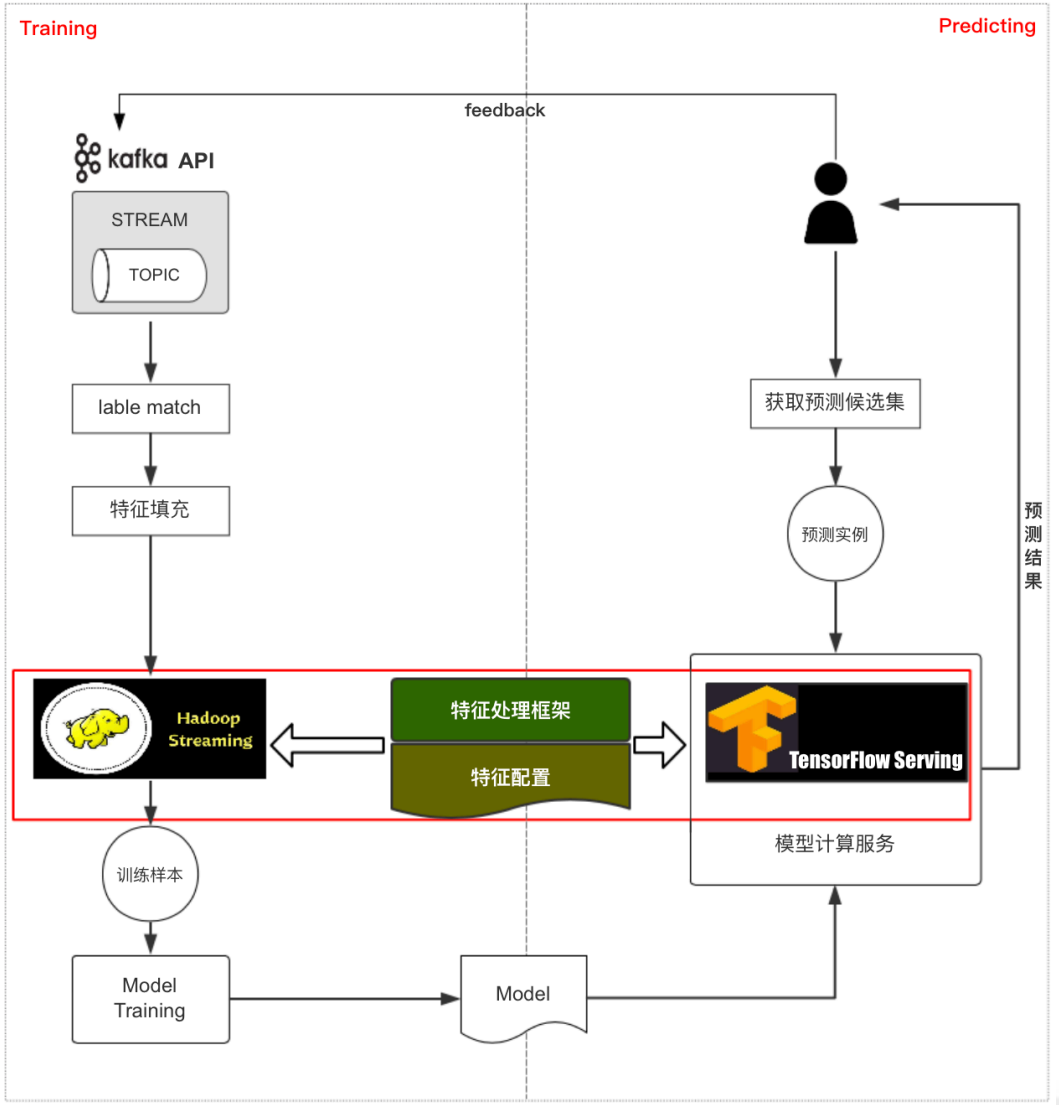
可配置化


age
CDF
Bucket
age cdf
age bucket


doc poi
user poi
DocVec
UserVec
Similarity
doc user similarit y

优化后的pipeline





模型框架面临的问题

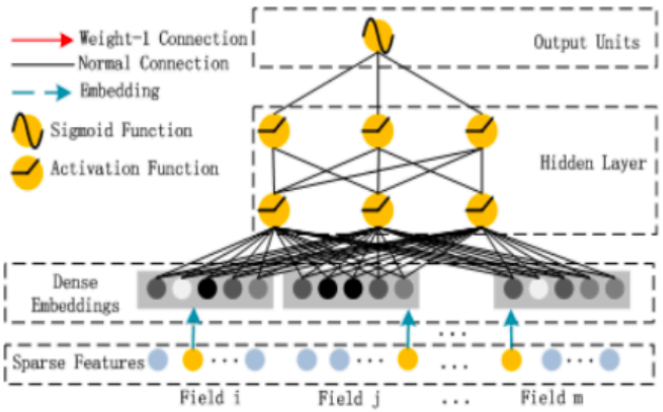
1. 深度学习排序模型框架：

- 通用化 & 可扩展的推荐算法库框架，以支持新的业务场景的快速落地，以及模型迭代
- 灵活性，根据业务的需求对模型做定制化

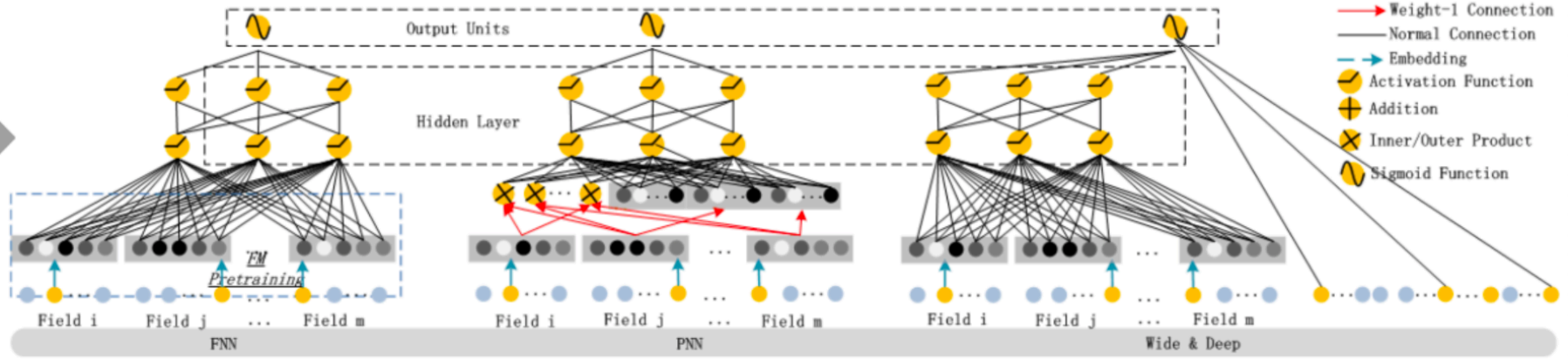
2. 可配置化：

- 对共通部分进行模块化后，框架通过读取配置文件来构建模型，避免了直接编写tensorflow代码的工作

深度学习排序模型的网络结构



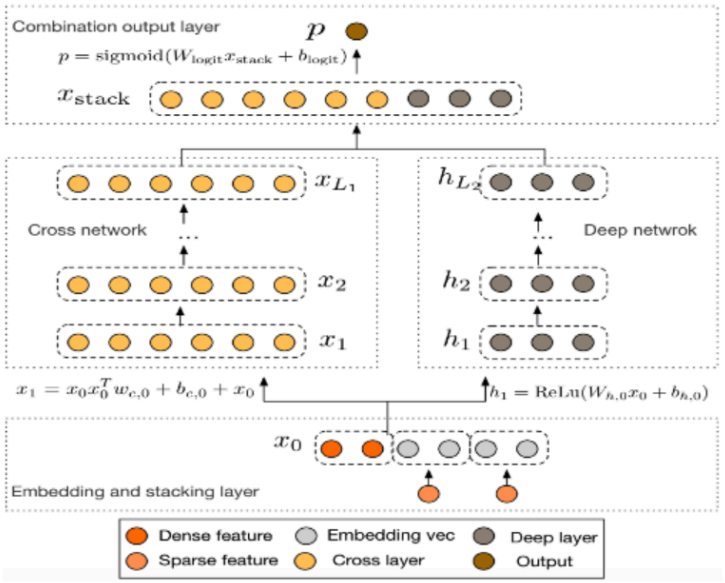
The architecture of DNN



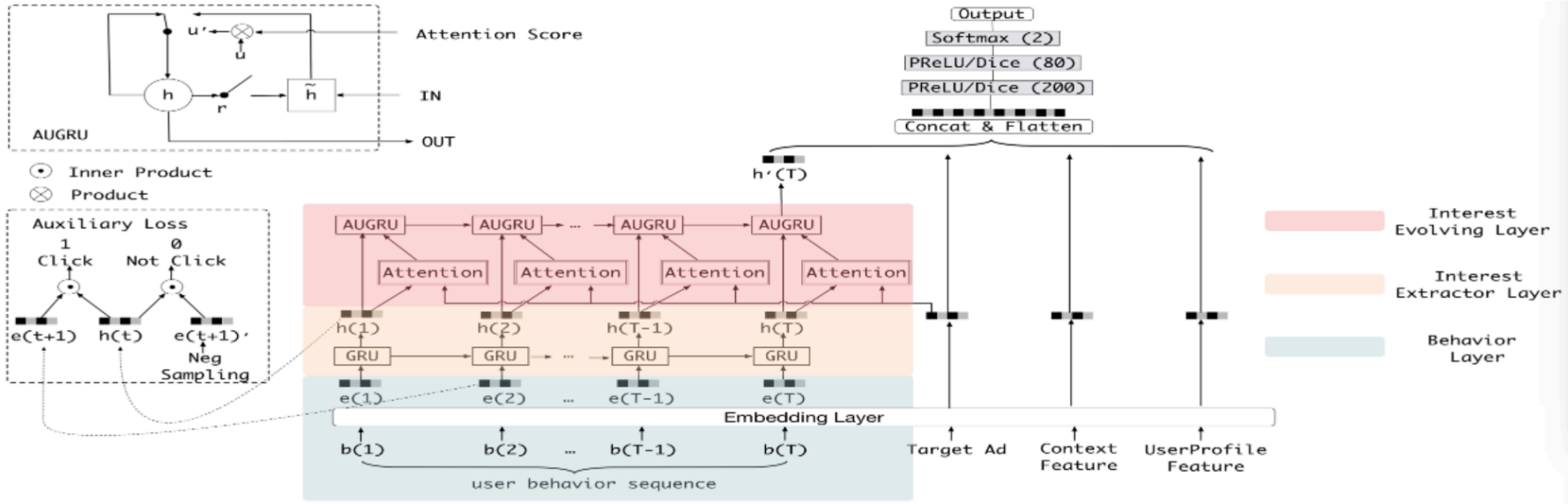
The architectures of deep models for CTR prediction: FNN, PNN, Wide & Deep Model

base units: Linear cross Deep … complex model: FNN PNN DeepFM Wide&Deep …

深度学习排序模型的网络结构



The architecture of Deep & Cross Network



The structure of DIEN. At the behavior layer, behaviors are sorted by time, the embedding layer transforms the one-hot representation $\mathbf{b}[t]$ to embedding vector $\mathbf{e}[t]$. After the behavior layer, interest extractor layer extracts each interest state $\mathbf{h}[t]$ with the help of auxiliary loss. At interest evolving layer, AUGRU models the interest evolving process that is relative to target item. The final interest state $\mathbf{h}'[T]$ and embedding vectors of remaining feature are concatenated, and fed into MLR for final CTR prediction.

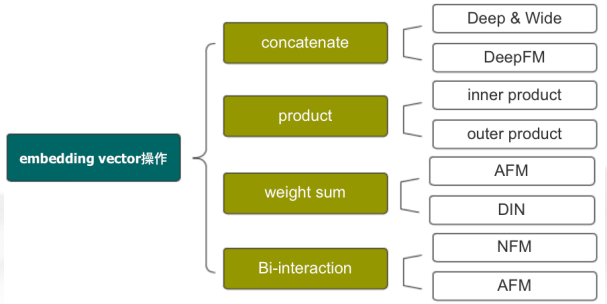
base units: Linear cross Deep … → complex model: Deep & Cross Network DIEN…

深度学习排序模型的通用范式

网络结构	是否需要人工特征	组成	需要预训练	低阶特征表达	高阶特征表达
LR	是	LR	否	是	否
FM	否	FM	否	是	是
DNN	否	MLP	否	否	是
FNN	否	FM + MLP	是	否	是
PNN	否	FM + product + MLP	否	否	是
Deep & Wide	是	LR + Embedding + MLP	否	是	是
DeepFM	否	FM + Embedding + MLP	否	是	是
NFM	否	FM + Embedding + MLP	否	是	是
AFM	否	FM + Embedding + attention + MLP	否	是	是
DCN	否	Embedding + cross + MLP + LR	否	是	是
DIN	否	Embedding + attention + MLP	否	是	是

- input->embedding
  把大规模的稀疏特征用embedding操作映射为低维稠密的embedding向量

- embedding层向量
  concat，sum，average pooling，product等操作，大部分CTR模型在该层做改造



- embedding->output
  通用的DNN全连接框架

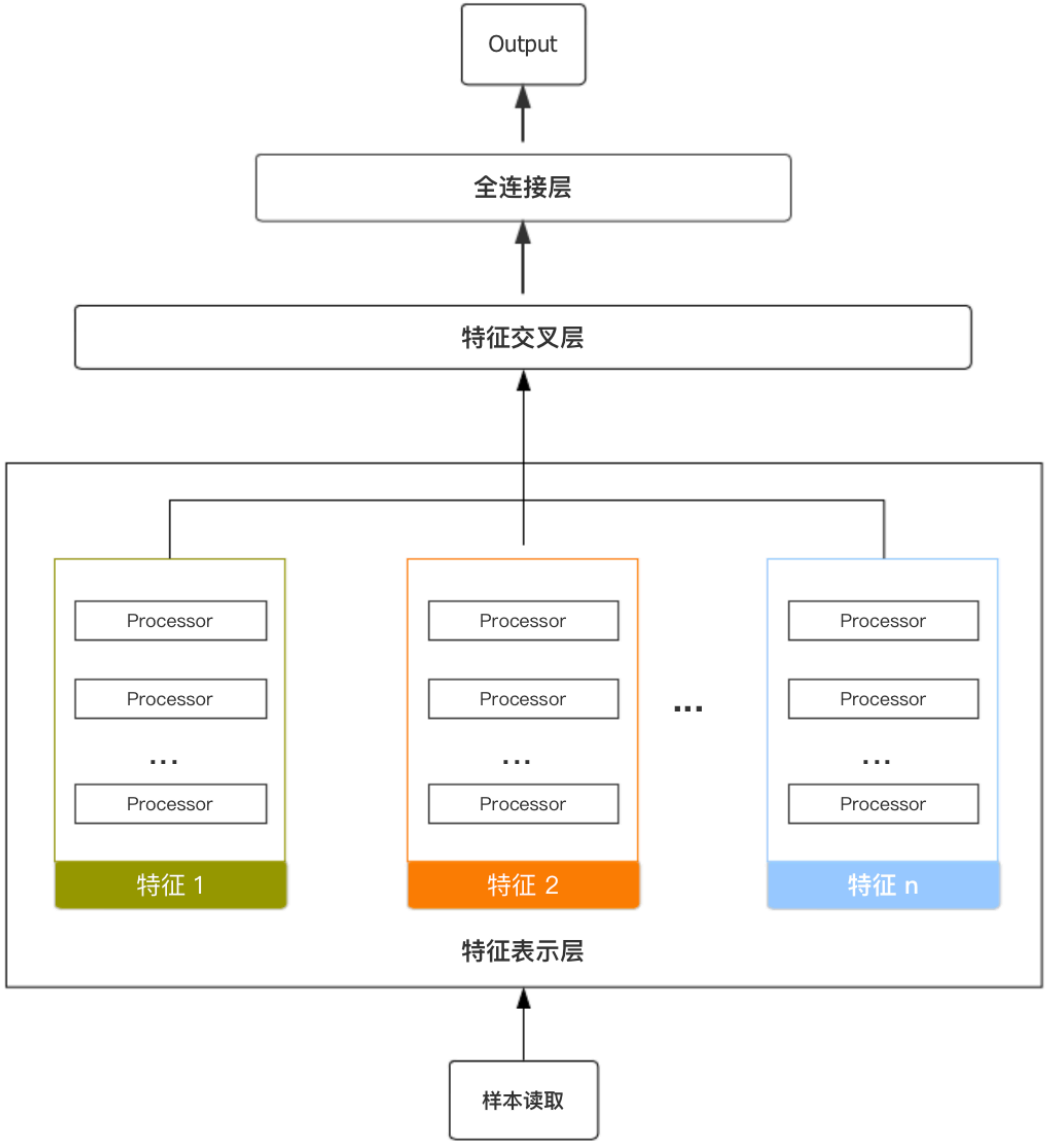
通用化的模型框架

- 特征处理层
  会对每个单独的特征做一些处理。例如对离散特征做embedding，对于多值的离散特征，经过embedding层后对结果进行average pooling。另外一些特征可能需要reshape等等。把这些操作封装成processor，可以指定每个特征要依次经过哪些processor。

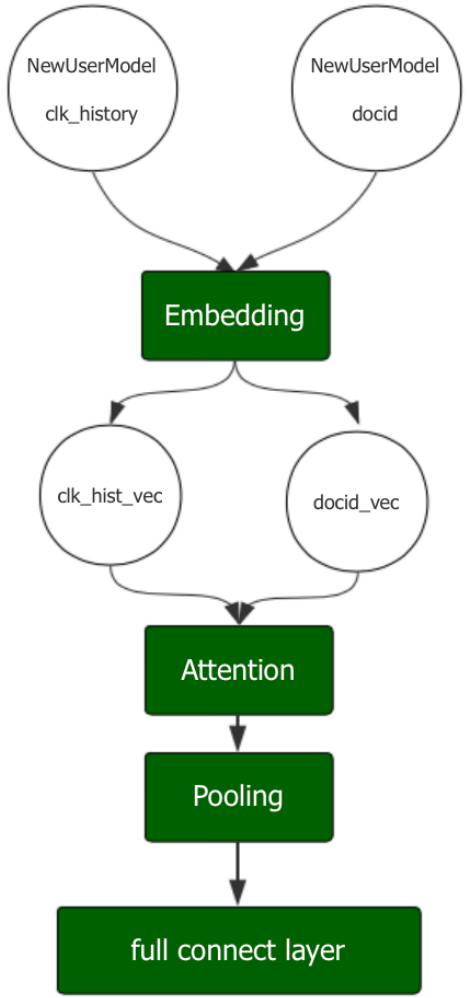
- 特征交叉层
  包括特征间进行attention计算或者显式的特征交叉等等。这种类型的操作模式简单，一般都发生在embedding层后，全连接层前。我们把这一类特征交互的操作分离出来，作为一个单独的模块。业务方可以根据自己对特征的理解，手动添加一些交叉项，提高模型的效果。

- 全连接层
  这一层完成隐式的特征交叉。我们可以指定全连接层的层数和每层的神经元个数。

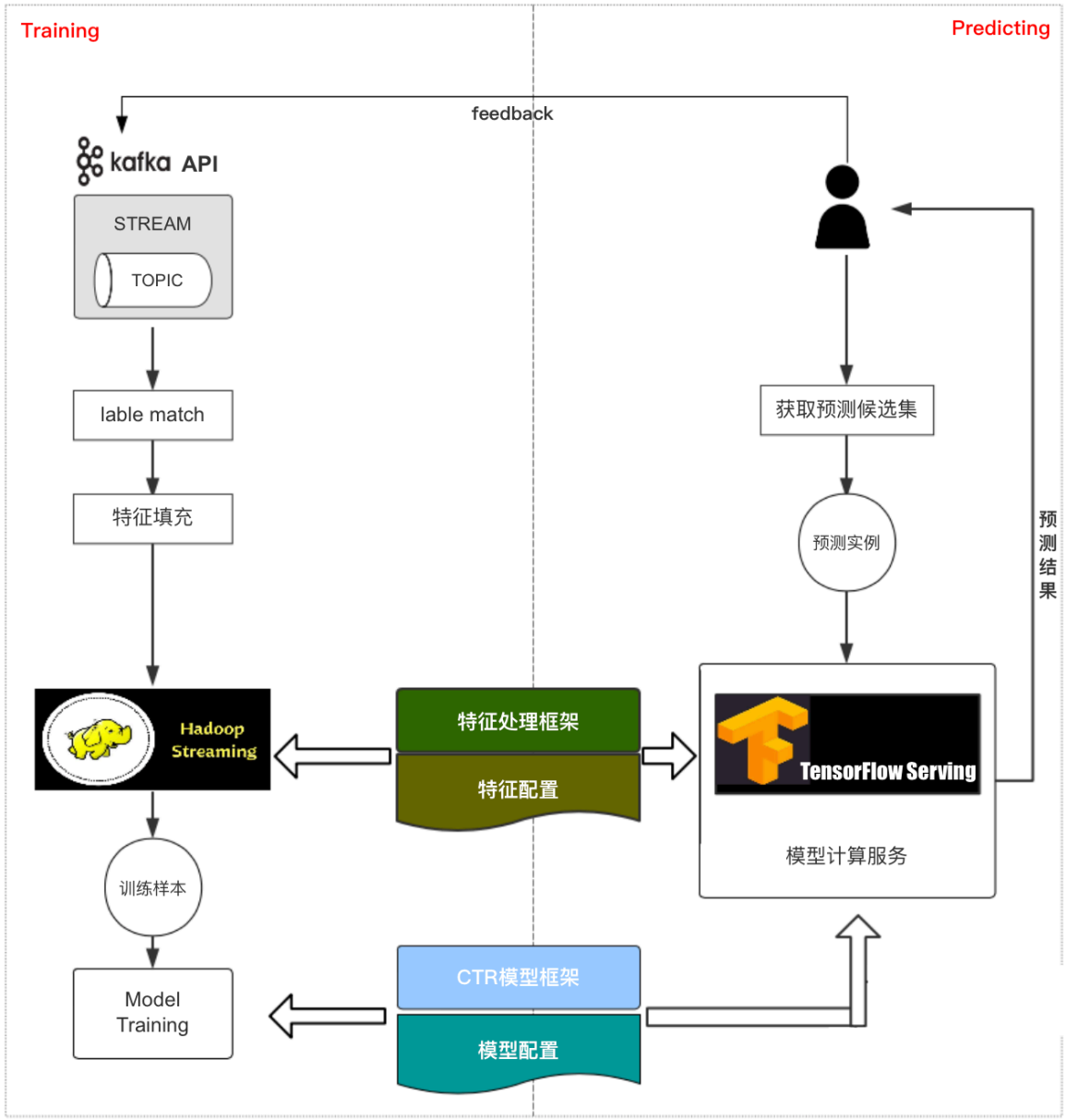
通用化的模型框架


Output
全连接层
特征交叉层
Processor
Processor
...
Processor
Processor
Processor
...
Processor
Processor
Processor
...
Processor
特征 1
特征 2
特征 n
特征表示层
样本读取

可配置化模型框架


NewUserModel
clk_history
NewUserModel
docid
Embedding
clk_hist_vec
docid_vec
Attention
Pooling
full connect layer

目前的排序系统架构


Training
Predicting
feedback
kafka API
STREAM
TOPIC
lable match
特征填充
Hadoop Streaming
训练样本
Model Training
特征处理框架
特征配置
CTR模型框架
模型配置
获取预测候选集
预测实例
TensorFlow Serving
模型计算服务
预测结果

模型效果

经过AB测试，深度学习排序系统目前已在网易新闻头条的多个推荐场景扩量上线。

在上线后一周，其中新用户排序业务的Deep Learning实验线相比baseline点击率提升9.854%，人均浏览时长提升8.0766%。