# 知乎推荐页 Ranking 经验分享

单厚智

2018.12.15

# 主要内容

- 知乎推荐页场景和 Ranking 历程介绍
- 深度学习在 Ranking 中的尝试和应用现状
- Ranking 面临问题和未来研究方向

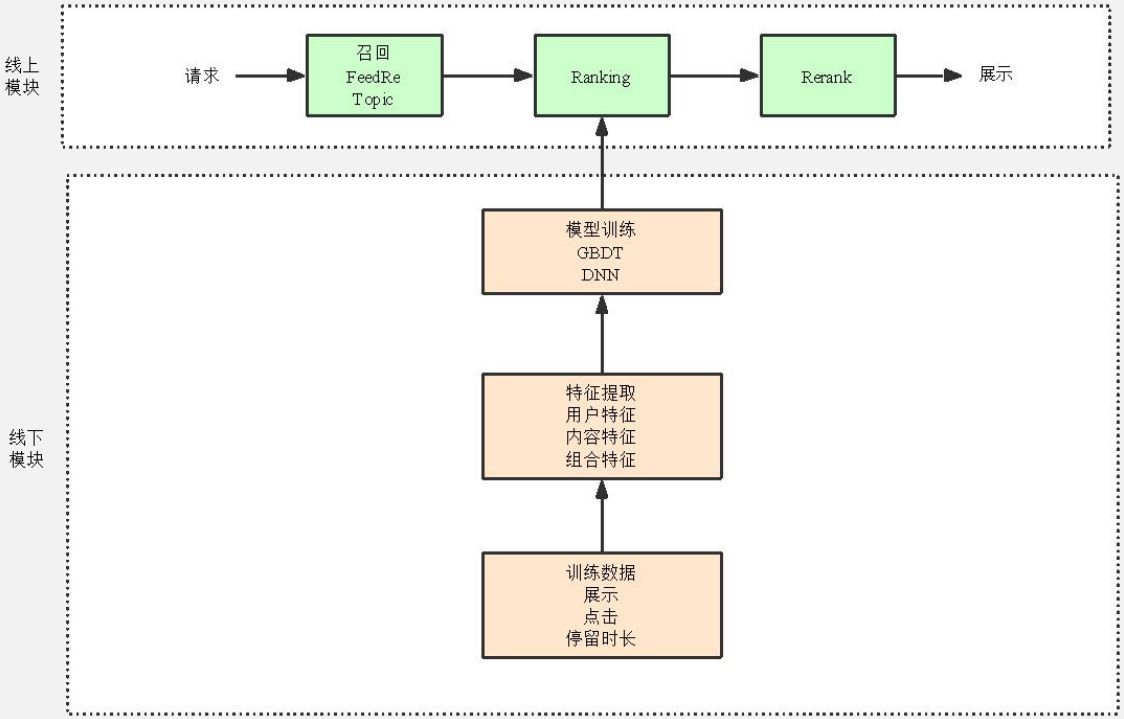
# 知乎推荐页场景


23:20
熊晓鸽的投资方法
提问
关注
推荐
热榜
视频
为什么电动汽车电池不能精确显示剩余电量?
赵玮炜: 首先回答楼主的问题"为什么不能像手机一样显示剩余电量百分比"这个问题的答案是: 能显示百分比, ...
17 赞同
靠工资在北京买房现实吗?
羊迪: 当然现实了, 不需要年入百万, 月入一万都没多大问题, 来, 我教你。首先, 确认下购房条件是否满足, 如...
388 赞同 · 136 评论
后悔当初没有考上高校? 环球网校轻松考取MBA, 点击报名
名校梦想无法实现? 现在机会来了! 环球网校轻松考取MBA, 帮你实现名校学历, 职场转型利器, 薪资爆炸式提升
环球网校 的广告 · 查看详情
据说机器学习长于预测, 计量经济学长于解释。有什么具体例子是机器学习完成了很好的预测却在解...
颢卿: 机器学习之于计量经济学 — 随机森林算法大战最小二乘法 今天这篇正好借花献佛, 来和大家从解释和预测两个角度来和大家简单讲讲机器学习和传统计量方法...
242 赞同 · 21 评论


23:19
熊晓鸽的投资方法
提问
关注
推荐
热榜
视频
Deeplearning.ai卷积神经网络( 1.6 如何做三维卷积)
忆臻: 0、前言 打算对吴恩达老师的Deeplearning课程做一个笔记回顾, 大体整理思路为吴老师课堂内容和个...
Convolutional Neural Networ
Computer vis
0 赞同
提升学历很难吗? 来环球网校考在职研究生, 学时短, 含金量高
学历是求职的敲门砖, 也是升职加薪的垫脚石。高考失利没关系! 在职研是你的又一次机会。来环球网校, 帮你高效学习, 提升学历!
环球网校 的广告 · 查看详情
电动车续航破600, 燃料电池汽车出路何在?
玖牛咨询: 导读 电动汽车与燃料电池汽车谁是新能源汽车未来? 比亚迪唐EV、现代NEXO、丰田MIRAR续航里程接近。失去续航里程高的优势, 氢燃料电池汽车出路何...
1 赞同 · 20 评论
海洋里有哪些有趣的哺乳动物?
一个男人在流浪: 谢邀。因为题目里说到了是海洋哺乳动物, 这就把我们的可选范围限制在了一定很小的面里——...
1,783 赞同 · 90 评论


23:55
毕赣邀你下周日见面聊爱情
提问
关注
推荐
热榜
视频
怎样评价一个演员的台词功底?
我们还是想走
We wanted to leave, but couldn't.
01:33
大酱: 我当时以为这是真的杀人犯采访
2,757 赞同 · 447 评论
有什么事是外国人来了中国之后才知道的?
笑笑木: 很多国家是不"拼假"的。比如, 今年国庆假期之前的周六日正常上班上学, 然后连着放七天, 这就是"拼假"。这种模式现在被很多人吐槽。我上大学的时候, 隔...
216 赞同 · 77 评论
汽车的操控性到底是指什么?
翰德林曼: 汽车的操控性到底是指什么? 说到操控性, 在中国和欧美还是有些差异, 在中国, 普通消费者理解的操控包含了动力加速性能, 驾驶性, 转向 ,极限操控 ,甚至...
25 赞同
如何克服一个人独处时莫名的空虚寂寞感? 或许, Soul里有要的答案

# 推荐页请求流程


线上
模块
请求
召回
FeedRe
Topic
Ranking
Rerank
展示
线下
模块
模型训练
GBDT
DNN
特征提取
用户特征
内容特征
组合特征
训练数据
展示
点击
停留时长

# 推荐页模块详解

- 召回：负责把用户可能感兴趣的内容提取出来，重点是全
  - 基于话题：关注，行为挖掘
  - 基于内容：协同
- 排序：负责对召回的内容进行打分，可以理解为感兴趣程度，重点是准
  - 基于规则：时间顺序，线性加权
  - 基于模型：GBDT，DNN
- 重排序：出于产品或者业务的考虑，对排序后的内容进行重排，最终展示
  - 提权：比如对视频进行一定的提权
  - 隔离：相似内容隔开
  - 强插：高质量的新内容流通

# 推荐页Ranking历程

# 模型选择

深度学习大趋势

能处理更高维度特征，
如几十万话题量级，
GBDT 无法完成训练

非线性模型，不必做大
量的特征组合

使用GPU + HDFS结构，
可以训练更多的样本

# 特征介绍

- 用户画像
  - 用户属性特征：性别等
  - 统计特征，用户点赞数等
- 内容画像
  - 固有特征：文章长度，关键词等
  - 统计特征：历史点赞数等
- 交叉特征
  - 用户与内容的交叉特征：比如用用户感兴趣的话题和当前待推荐内容的话题交叉

# 特征介绍

- 特征类别
  - 数值特征：文章长度，点赞数
  - Onehot：内容类型
  - Multihot：内容多个话题 id
  - Onehot with value：用户对单类型内容的感兴趣程度
  - Multihot with value ：用户对各话题的感兴趣程度

# 特征设计

## ●设计原则

- 特征尽量全：从现有的数据中提取尽可能多的特征
- 特征原始值全：比如加历史CTR 特征的时候，可以把 pv 和 click 都带上
- 覆盖率大：去掉一些覆盖率很低的特征，这些特征影响影响范围小，大部分是缺失值
- 线上线下一致：覆盖率和取值分布尽可能接近

## ●新特征方向

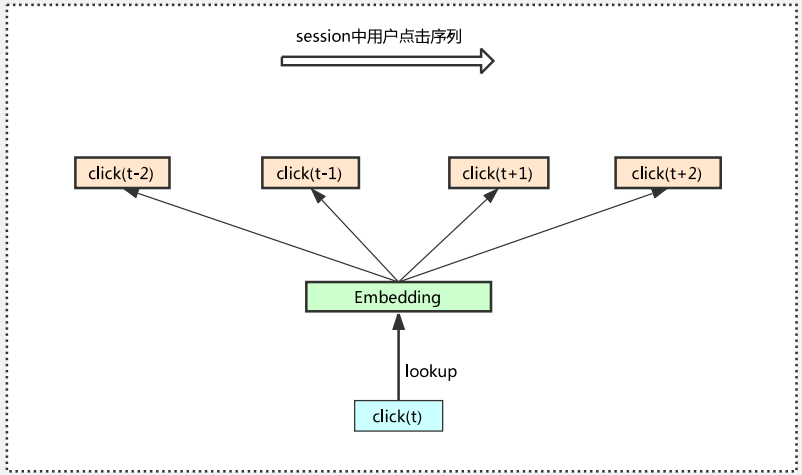
- 显式交叉特征：DNN 能学习特征的非线性能力，增加交叉特征可以降低模型搜索的空间，在训练数据一定的情况下可以提升效果，如用户的话题兴趣和当前话题的均值和最大值，效果提升明显
- 出于业务考虑：需要对业务有一定的理解，把自己当做用户，考虑什么情况下点击率会大，什么样的内容更容易被用户点，比如视频在 wifi 下更容易被点，视频点击率高的人更喜欢视频
- 数据挖掘特征：如内容 Embedding 特征

# 特征设计

- 内容 Embedding 介绍
  - Embedding 目的：把内容映射到低纬空间，相似内容距离较近，可以当做内容特征
  - 文本角度：tfidf，关键词进行word2vec 等
  - 行为角度：考虑用户在知乎的行为，搜索内容相关性较好，依据搜索点击序列设计 Embedding
- 实现细节
  - 数据：将搜索行为进行session切分，组织成类似于 sentence 序列
  - 样本：85亿
  - 模型：skip-gram
  - loss：nce-loss

# 特征设计

- skip-gram 结构




- embedding 效果

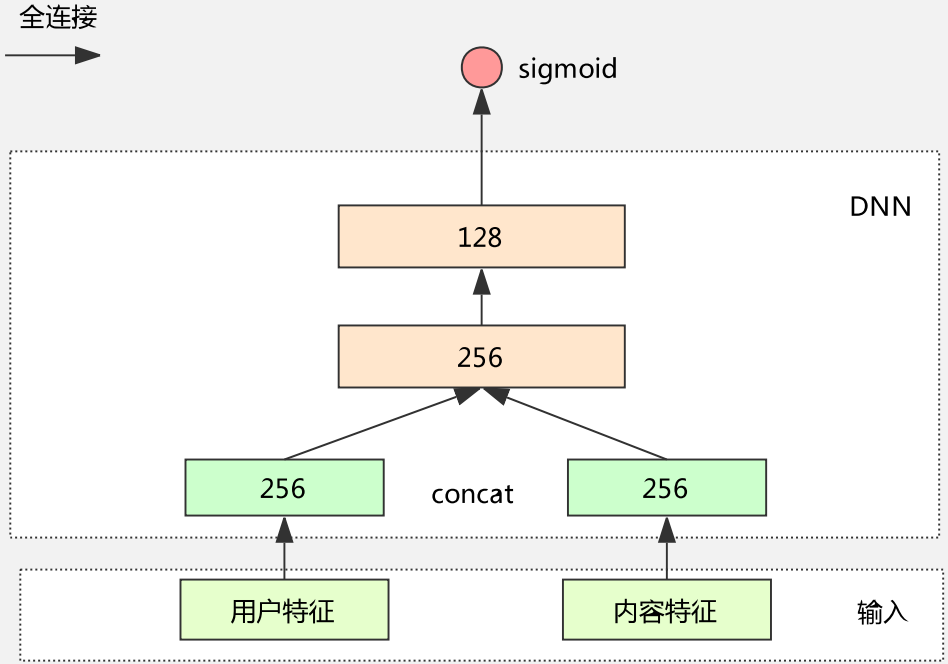
| 类型 | ID | 标题 | 摘要 |
|---|---|---|---|
| 回答 | 97421005 | 回答: 像vb vc ve这样的保健品，可以天天吃吗? | 谢邀，这些保健品VB是非常容易缺乏的，可以补充。维生素C和维生素E不太容易缺乏，多吃蔬菜水 |
| 回答 | 67962958 | 回答: 开始拼命地吃维生素b族 vb3 vb6 vc ve 望能治疗 | 答主：怎样私信你？我也遇到同样问题，痘痘长了大概有十一年了，今年25周岁，期间试过各种方 |
| 回答 | 61376154 | 回答: 我是22岁的女生，请问日常吃什么营养品比较好 | 我高三的时候被我妈逼迫吃葡萄籽、vc和各种保健品，但是葡萄籽我最讨厌吃，因为真的很难吃， |
| 回答 | 2927800 | 回答: 继续用综合维生素还是只服用vc vb？ | 题主你好！综合维生素的问题之前我也一直在关注，首先陈述我的观点1.综合维生素按时服用后， |
| 回答 | 66482293 | 回答: 吃维生素b族 vb3 vb6 vc ve 望能治疗我的痘痘 | 我也吃过小瓶的VB6坚持吃了三个月好像 基本上就是按说明上的吃的 吃了10瓶 吃完后确实能控制 |
| 回答 | 93548384 | 回答: 看微博里说纽斯特的VC+VE+葡萄籽可以美白， | 本人黑，双11买的纽斯特的美白套餐，就是题主说的那个。到现在四个月了。首先说白，没感觉变 |
| 回答 | 83179992 | 回答: 吃药店的小瓶vc和吃自然堂，汤臣倍健的vc片功 | 没有差别，因为你摄入的维生素剂量是一样的，这是标准剂量，只是自然堂、汤臣倍健等保健品品 |
| 回答 | 94153709 | 回答: 想利用服用VC和VE的胶囊美白，这科学吗? VC | 两种维生素都有抗氧化的功能，至于能不能美白，我只能说有待商榷。维生素ADEK属于脂溶性维生 |
| 回答 | 18055293 | 回答: 因为长痘痘，可以VB.VC.VE一起服用一段时间吗 | 我当时也是长过很多痘痘，在杭州看的中医调理一段时间确实好了，吃中药的同时医生配了vc vb综 |
| 回答 | 93241351 | 回答: 吃富含vc的水果和直接喝vc片有什么区别? | 谢邀！当然会有区别！片剂有提取的和合成的两种片剂含量多少不一片剂中营养素单一水果中富含 |
| 回答 | 33467535 | 回答: 普通人长期每天吃 Nature Made莱萃美的vc和ve | 谢邀，首先抱歉的回答，因为对您购买的保健品并不了解，无法正面回答您的问题，抱歉。其次， |

# CTR 模型

- 选择 CTR 模型原因：
  - 推荐页排序目标是把用户推荐感兴趣的内容排在前面，可有下面两个学习目标
    - 停留时长：适合用回归问题来解决，最后会偏向于长文章
    - 点击率：二分类问题，知乎的问答一般不长，更加合适
  - 分类问题相比回归问题，目标类别少，相对准确率高
  - 分类问题场景业界应用较广，可交流空间大
  - 分类问题最后会输出一个概率分，方便与多目标结合
- 损失函数

$$Loss = -\sum_{i=1}^{N} (y_i \log(p_i) + (1-y_i)\log(1-p_i))$$

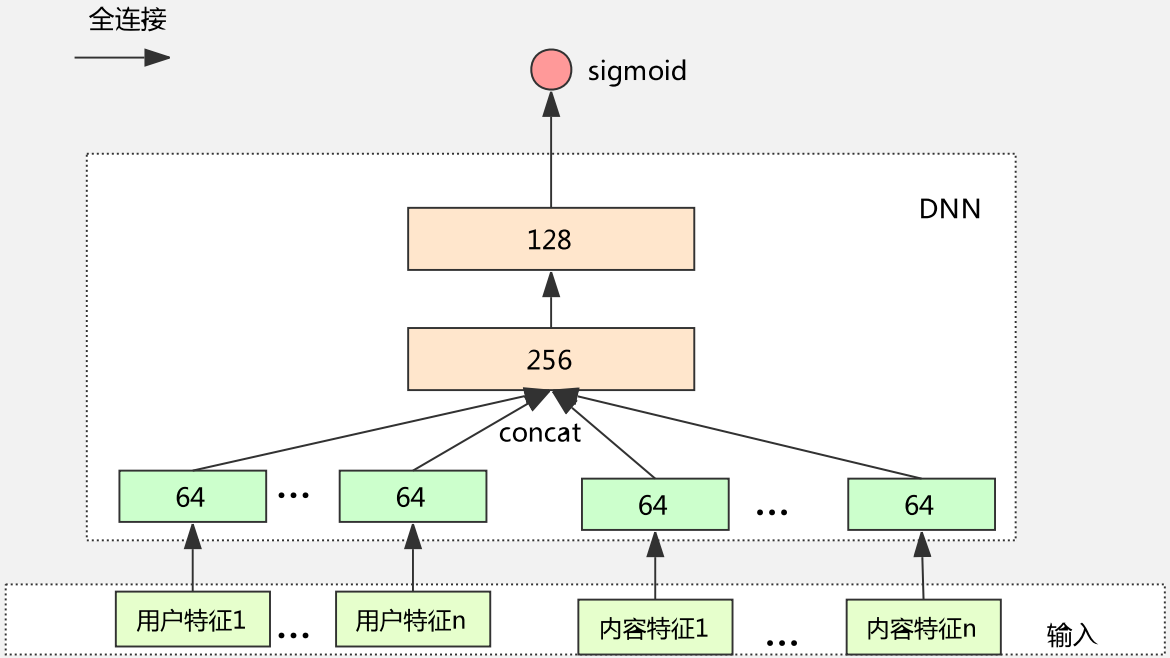
# 最初DNN 结构




1. 将输入特征分为用户和内容两块
2. 经过特征映射后分别通过全连接与两个独立的隐含层连接
3. 两个独立的隐含层 cancat 后再经过两个全连接层
4. 最后输出 sigmoid 与交叉熵损失作为 loss

AUC：0.7618

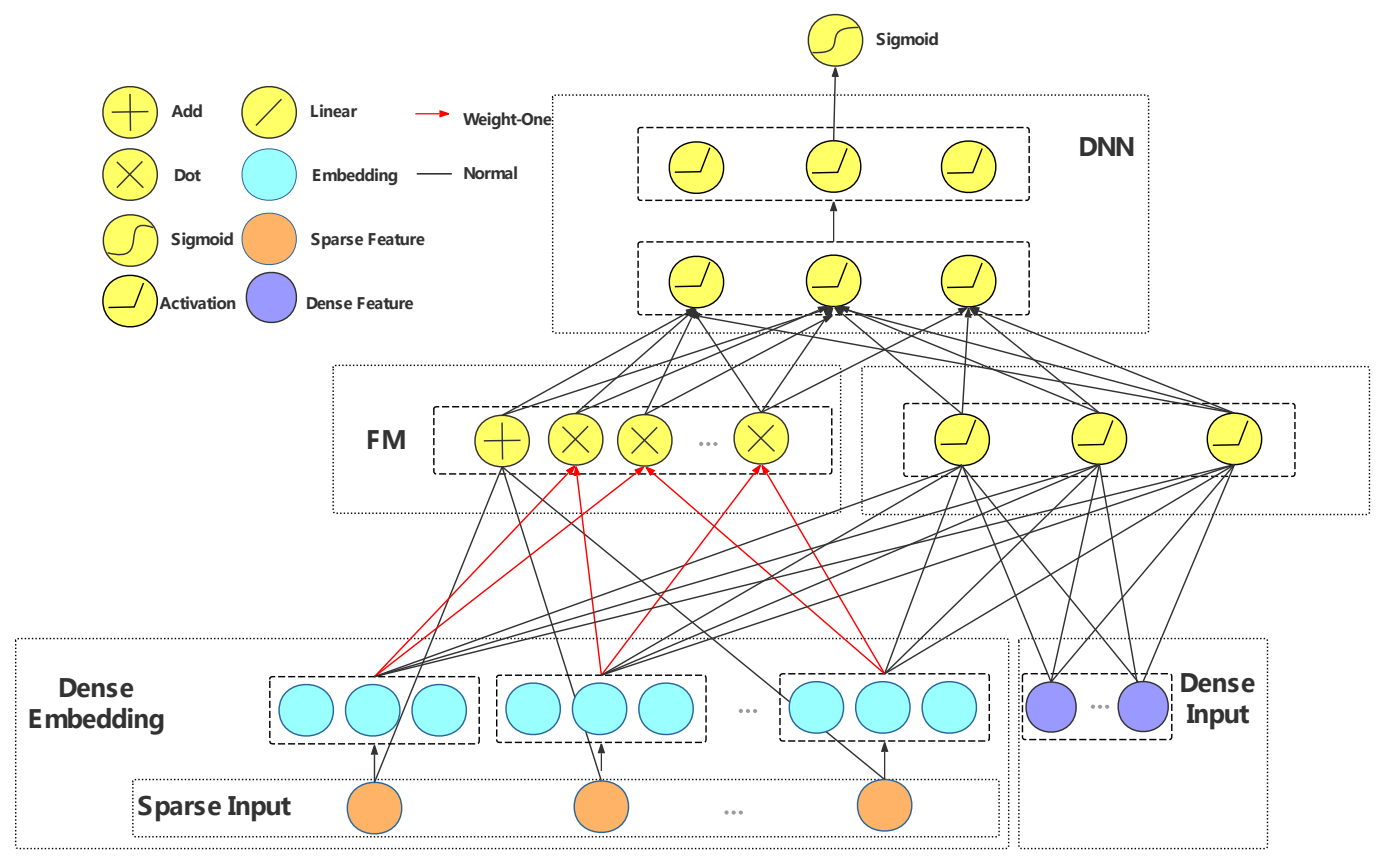
# 优化 DNN 结构




1. 将用户和内容的特征，分别按照内容的 field 分为不同的 block
2. 每个 block 先经过全连接到独立的隐含层
3. 将上面的隐含层 concat 再经过后面的 DNN 模型

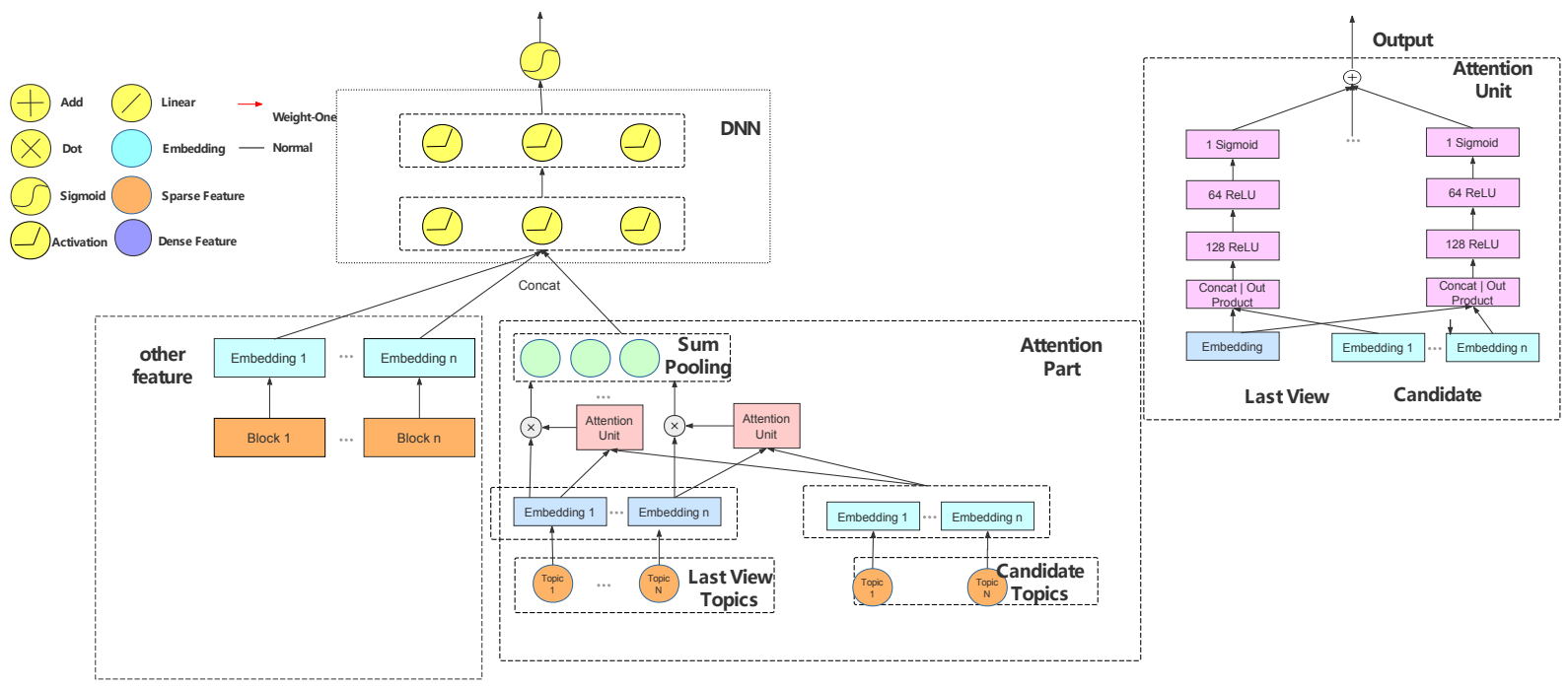
AUC：0.7678，提升0.6%

# Deep FM



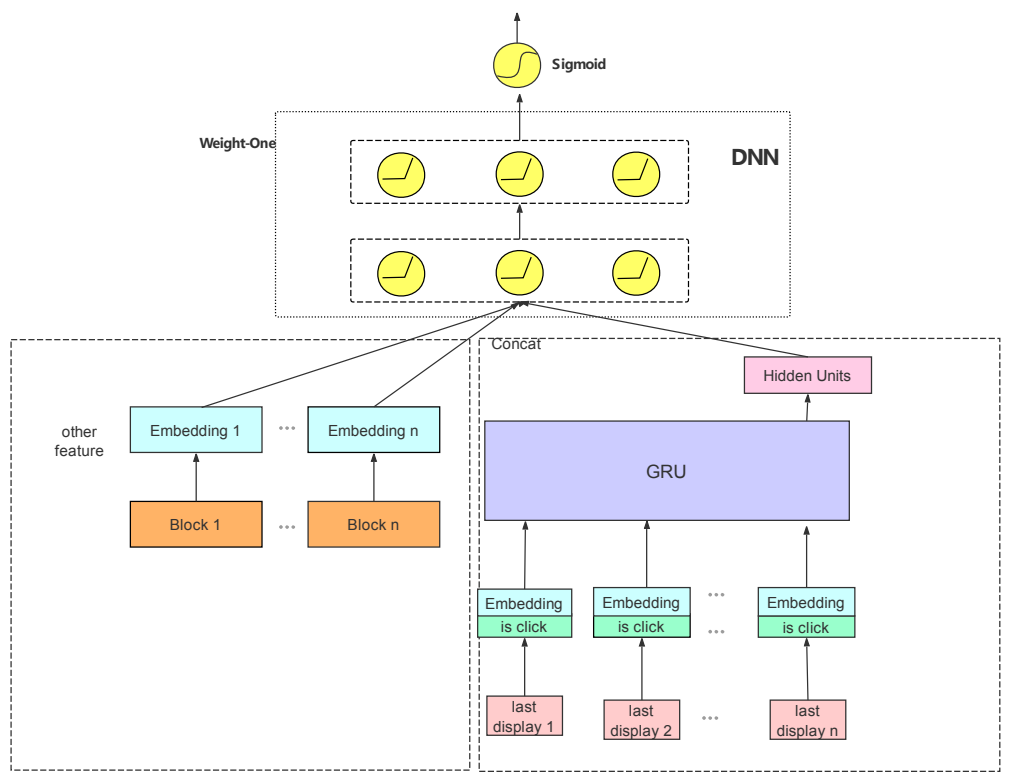

1. 增加了一阶和 FM 模块，FM 通过 block 之间的内积实现
2. AUC 提升 0.2%

# Last View + DIN



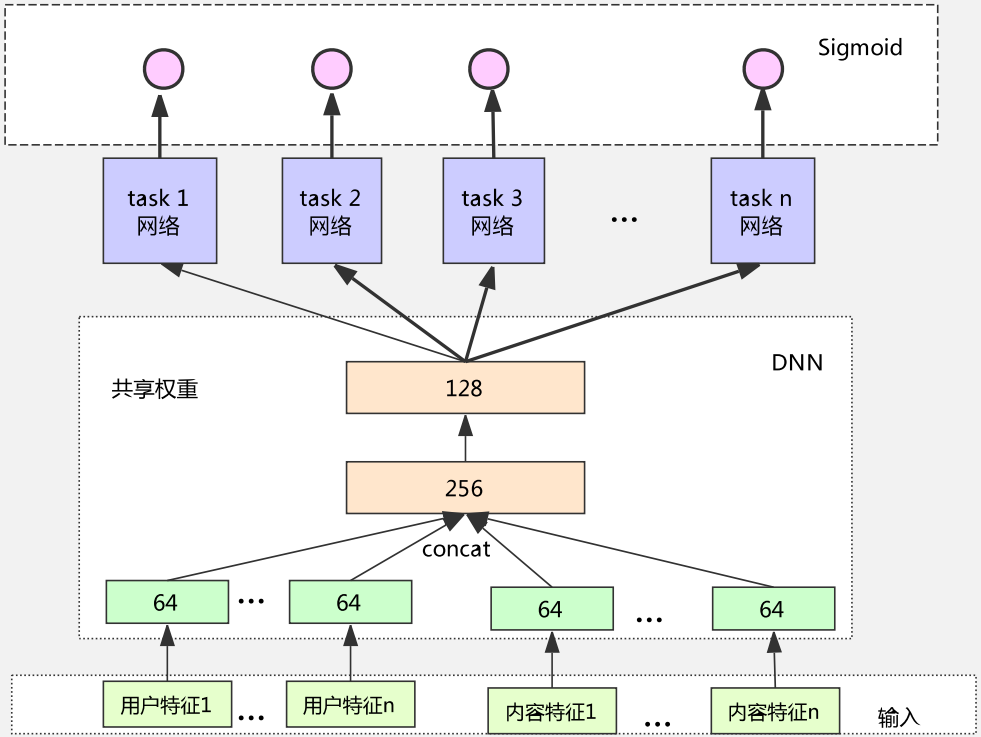

1. Last view topic 与当前内容的几个 topic 计算 Attention Score，再按权重进行 sum pooling
2. AUC 提升约 0.2%

# Last Display + GRU



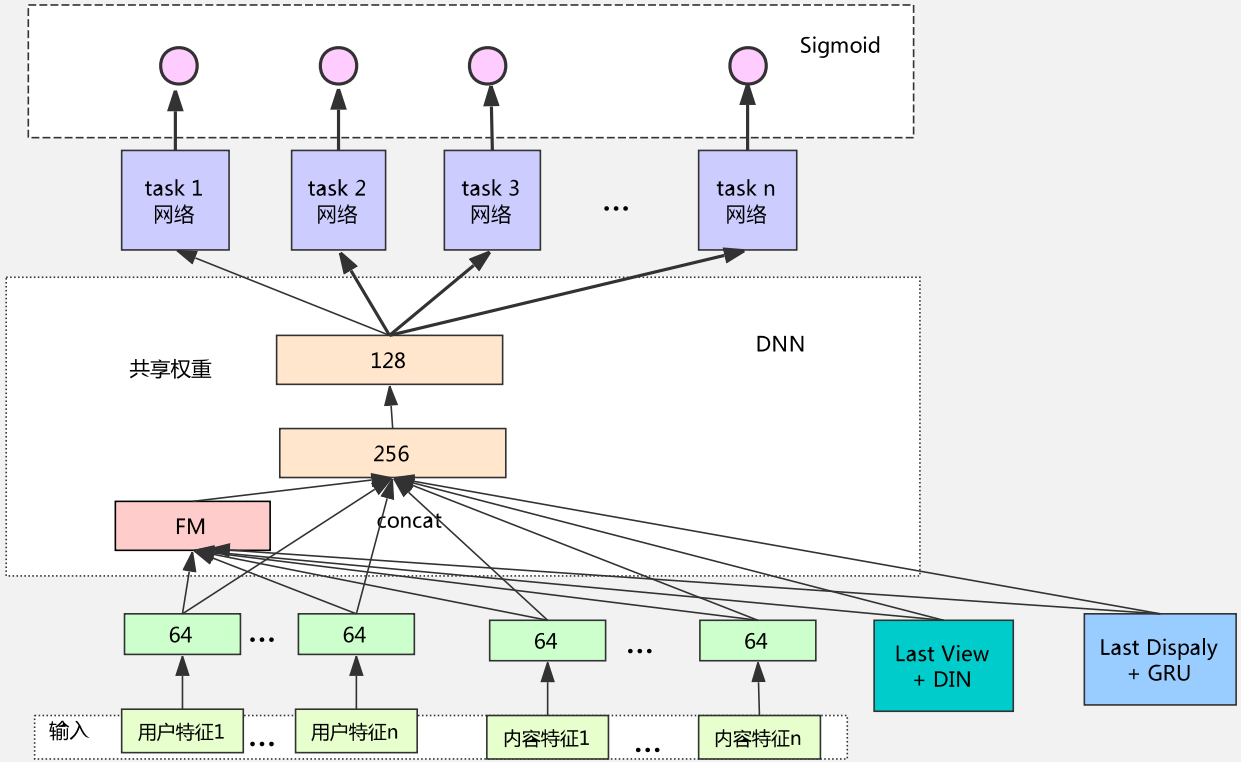

1. Last Display 经过 Embedding 后与是否点击结合，再进入 GRU 模块，最后状态当做 DNN 输入
2. AUC 提升约 0.4%

# 多目标




1. 每个 task 共享前面的几层权重，可以节省训练和预测的计算量
2. Loss 可以是几个 task 的 loss 做简单线性加权
3. 上线仍然要确定各个 ctr 的加权值，经验比较重要
4. 上线后线上表现：点击率基本不变，而其他的几个指标，比如点赞，收藏大幅提升

# 最终模型结构


Sigmoid
task 1
网络
task 2
网络
task 3
网络
...
task n
网络
共享权重
128
DNN
256
FM
concat
64
...
64
64
...
64
Last View
+ DIN
Last Dispaly
+ GRU
输入
用户特征1
...
用户特征n
内容特征1
...
内容特征n

# 经验分享

- 对于随时间变化的统计特征，比如用户和内容画像的统计值，线上 service 应当纪录请求时的值，生成训练样本时直接从纪录的日志里获取，避免特征穿越问题；
- 如果发现线下效果好，比如 AUC 和 NDCG 提升明显，但上线效果不显著，很可能是特征不一致导致的，可重点排查；
- 线上线下最好使用同一套特征抽取框架，只需使用的相同特征配置便可保证一致性，我们 Global Ranking 使用同一套 proto 结构和特征抽取模块实现；
- 做特征归一化操作，发现有特别大的值，比如几万或者几十万，要先取 log ，不然会导致这个特征大部分值都趋向0，相当于特征失效；
- 输入特征要做非法检查，防止出现 inf，nan，而导致模型训练出现异常的参数；
- 对于线上的每次请求，用户特征都是一样的，可以只计算一遍用户特征相关的 block，避免冗余运算；
- 训练数据量要尽可能大，可以使用 FlatBuffer 结构把训练数据存放在 HDFS 上，训练时直接从 HDFS 读取，边读取边训练；
- 线上模型要能自动更新，过老的模型效果下降严重；

# 面临问题

- 推荐页与搜索页的特性不同
  - 搜索带着 query 来的，结果与之相关性越高越好，不用太关心结果的多样性
  - 推荐页用户没有明确的目的，但是有兴趣偏好和对结果的多样性需求，推荐既要准确又要多样化
- CTR 预估模型是 pointwise 模型，没有考虑单个内容与其他内容同时出现的影响
- 用户对感兴趣的东西会出现审美疲劳，要及时抓住这种特点，比如一个算法工程师看完几个机器学习文章后就不想再看了，这时候要能推荐一些其他话题的内容

# 未来方向

- 强化学习
  - Actor：根据用户过去的浏览和点击行为生成推荐页整屏结果
  - Critic：接收到点击或者其他正向行为作为 reward，同时训练 Critic 和 Actor 网络参数
- 优点
  - 能及时捕捉用户的反馈，从而避免对同一话题产生审美疲劳
  - 推荐整屏幕内容，避免 pointwise 方式下内容较为集中问题
- 缺点
  - 模型结构复杂，模型参数训练较困难


Thank you!