AiCon
全球人工智能与机器学习技术大会

知乎 News Feed 中的机器学习实践

张瑞

知乎个性化推荐及 Feed 业务技术负责人

主办方 Geekbang. 极客邦科技 InfoQ

极客时间

重拾极客精神·提升技术认知

下载极客时间App

获取有声IT新闻、技术产品专栏，每日更新

扫一扫下载极客时间App

人工智能基础课

“通俗易懂的人工智能入门课”

王天一
博士　副教授

扫一扫，免费试读

AI技术内参

你的360度人工智能信息助理

洪亮劼
Etsy 数据科学主管

扫一扫，免费试读

AI 前线

关注落地技术，探寻AI应用场景

- 14万AI领域垂直用户
- 8000+社群技术交流人员，不乏行业内顶级技术专家
- 每周一节干货技术分享课
- AI一线领军人物的访谈
- AI大会的专家干货演讲整理
- 《AI前线》月刊
- AI技能图谱
- 线下沙龙

扫码关注带你涨姿势

Geekbang. 极客邦科技 | InfoQ

ArchSummit 全球架构师峰会

2018 · 深圳站

从2012年开始算起，InfoQ已经举办了9场ArchSummit全球架构师峰会，有来自Microsoft、Google、Facebook、Twitter、LinkedIn、阿里巴巴、腾讯、百度等技术专家分享过他们的实践经验，至今累计已经为中国技术人奉上了近千场精彩演讲。

限时7折报名中，名额有限，速速报名吧！

2012.08.10-12 深圳站

2018.07.06-09 深圳站

会议：07.06-07.07

培训：07.08-07.09

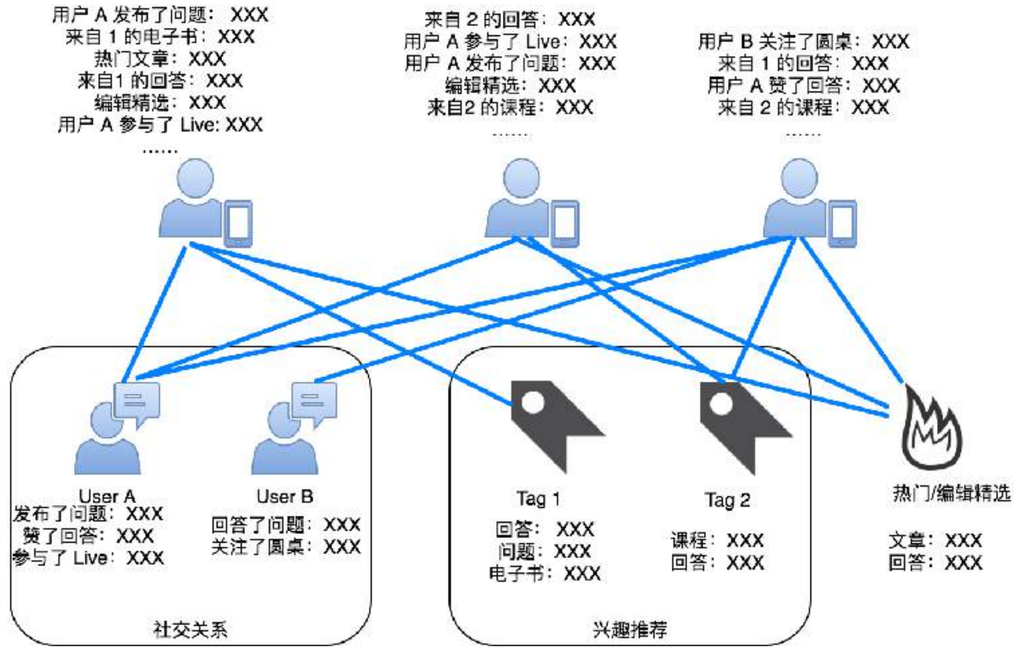
个人介绍

- 教育经历
  - 北京邮电大学网络技术研究院硕士
- 工作经历
  - 百度 NLP & KS
    - 高级研发工程师 - 机器学习方向
  - 豌豆荚 - 搜索部门
    - InAppSearch 搜索质量高级研发工程师及内容推荐技术负责人
  - 知乎
    - 组建知乎机器学习和数据挖掘部门，担任知乎机器学习团队技术负责人
    - 目前担任知乎个性化 Feed 与推荐业务技术负责人
- 研究领域和兴趣
  - 自然语言处理
  - 搜索及个性化推荐业务中的架构与策略

目录

- 知乎 News Feed 产品简介
- 知乎 News Feed 后端策略演进
    - Edge Rank
    - Learning to Rank
    - DNN Based Recommendation
- 总结

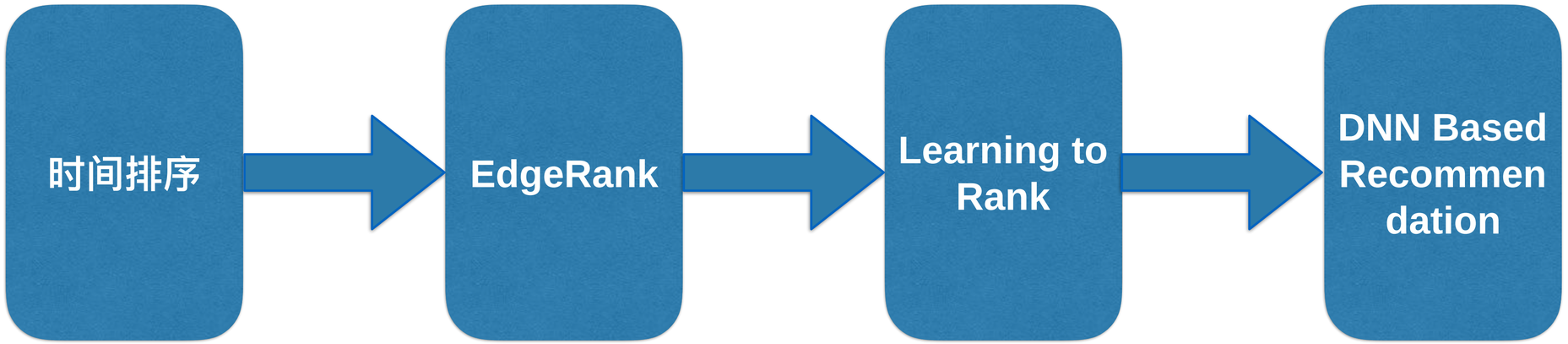
知乎 News Feed 产品简介



多种来源：主动（关注），被动（推荐），编辑精选

多种内容：问题/回答/文章/专栏/Live/电子书/视频/……

知乎 News Feed 后端策略演进



知乎 News Feed 后端策略演进

Edge Rank: 打破「时间序」的尝试

- 背景
  - 用户增长：信噪比增大，信息过载
  - 时效性?相关度?
- 优点
  - 足够简单，可解释性强
  - 易于对「相关度」「时效性」「内容类型」等进行权衡和调整
- 缺点
  - 过于简单，对数据的使用能力差
  - 调整参数没有有效的指导

$$Score = \sum_{e \in \Lambda edges} s_e w_e d_e$$

d: 时效性指标

$$d(t) = e^{\alpha(t-t_0)}$$

s: 来源相关性

```
def affinity_score(display,
                   up,
                   down,
                   previous_score):
  // 根据用户对内容来源的交互更新
  Affnity Score
```

w: 类型权重

用户关注问题	3.0
用户收藏答案	2.0
用户创建回答	1.0
用户赞同文章	0.8
……	……

知乎 News Feed 后端策略演进

Learning to Rank：流程与要素

流程

要素

数据

训练数据收集
清洗无效数据
对数据进行采样

特征

特征选择
特征变换

目标

根据业务目标设计需要优化的目标函数，确定 loss

模型

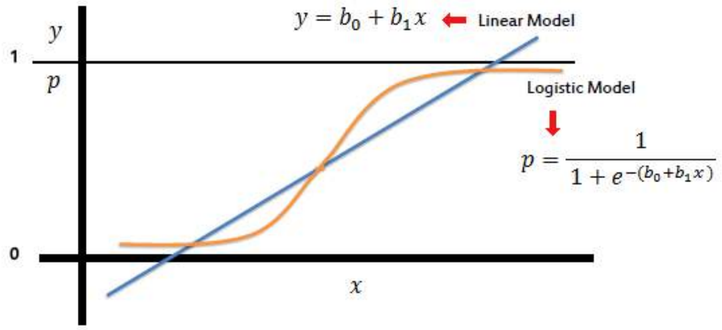
根据业务场景、数据情况、开发周期等选择合适的模型

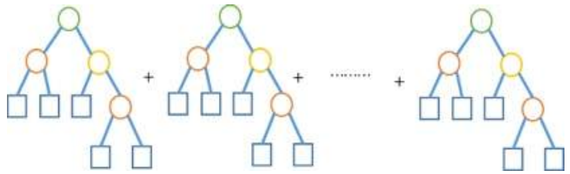
效果

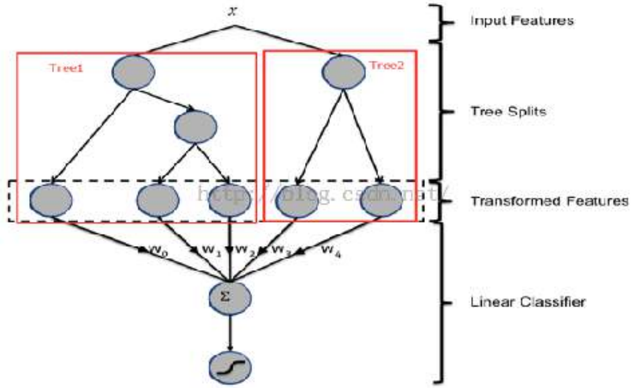
在线/离线效果评估，确认模型效果和收益

知乎 News Feed 后端策略演进

Learning to Rank：模型选择







LR

- 广义线性模型：简单，可解释性好
- 训练和推断效率高；并行化及 Online Update 简单方便
- 特征工程繁琐：离散化和特征组合等

GBDT

- 能够自动、有效地捕捉非线性
- 计算复杂度高；模型容量有限
- 没有支持特别好的并行化和在线学习版本
- 对稀疏特征的效果不好

GBDT+LR

- 融合 GBDT 和 LR 的优点
- 实际场景中的效果需要验证

知乎 News Feed 后端策略演进

Learning to Rank：特征体系

用户特征

- 用户类别：阅读型/写作性用户，新/老用户
- 人口学特征：地域，性别，学历，职业……
- 历史偏好：历史点击率，阅读文章的平均长度……

上下文特征

- 访问时间
- 机型/平台
- 非严格重复（同 Session 中同类内容条数）
- ……

交叉特征

- 是否关注内容源
- 是否关注创作者
- 对内容 Tag 的兴趣度（Max/Min/Avg）
- 用户对该类型内容的历史点击率
- ……

内容特征

- 历史交互信息：历史 CTR，获得赞同/反对数……
- 文本特征：类型，长度，格式(html 标签数，img 标签数，分段)……
- 作者信息：权威度，People Rank，关注者数量……

知乎 News Feed 后端策略演进

Learning to Rank：训练样本收集

- 主要问题
  - 负样本比例和正样本比例不平衡：展现但不点击的样本数倍于有点击的样本
  - 维度特征占比过低：部分特征维度上，非缺省值样本占比较低
  - 实时特征的易变性：在收集样本时候得到的特征和线上预测时使用的特征发生变化
  - 用户行为的时间相关性：易受到短期行为波动的影响
- 解决方案
  - 负样本欠采样，正负样本 1:1 分布；可以为 AUC 带来 2% 左右的提升
  - 对某些重要且有特征样本占比低的特征，增强非特征缺失的样本在训练集中的比例
  - 后端预抽样落地实时特征
  - 随机选择较大的时间范围（2w - 1m）抽取数据

知乎 News Feed 后端策略演进

Learning to Rank：优化目标选择

- 交互：点击（阅读），分享/收藏
  - 0-1 目标
  - 套用 CTR 预估框架
- DWellTime：分级目标
  - Normalized Dwell Time
    - 使用 z-value 压缩 Dwell Time 的范围，进行归一化 $z_i = \frac{\log(t_i) - \mu_{C_i}}{\sigma_{C_i}}$
    - 对 z-value 进行分桶：Long/Normal/Short/NoClick
  - 修改 Pesudo Response 而不是 sample 权重

$$\text{pseudo\_response}(x) = -g_m(\mathbf{x}_i) \times \text{scale}(label)$$

知乎 News Feed 后端策略演进

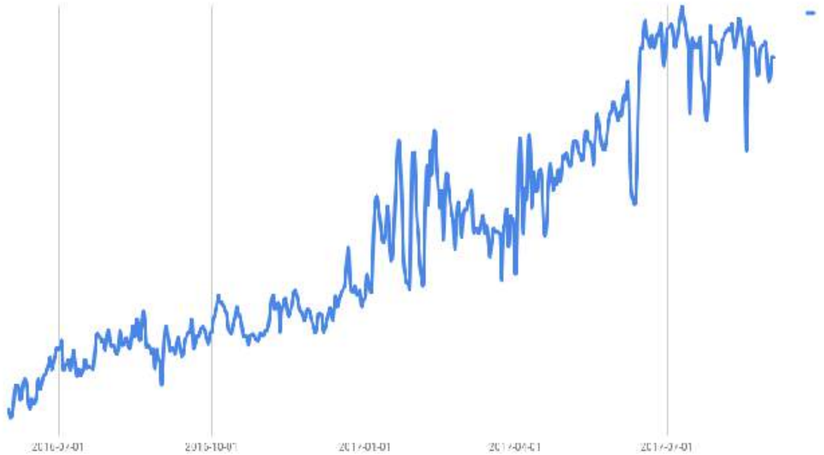
Learning to Rank：离线效果评估

- AUC
  - 用于 0 - 1 分类（CTR 预估）场景下的离线效果评估
  - AUC 的变化趋势和线上效果正相关，但难以估计提升量
- DCG Gain
  - DCG: $\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)}$
  - DCG Gain：$DCG_{10,reranked} / DCG_{10,online} - 1.0$
  - DCG Gain 和线上数据指标（尤其是 CTR）表现正相关，并且相对量上基本对应

知乎 News Feed 后端策略演进

Learning to Rank：线上实验及效果

- 总体效果
  - CTR Increase： 100%+
  - Duration Increase： 40%+
  - Views： 15%+
- 实验节点



	效果	Remark
EdgeRank -> GBDT LTR	CTR +13%	和 Edge Rank 使用特征类似，加入了内容维度统计特征和 Context 信息
引入实时兴趣	CTR + 15%，duration + 7%	无
拟合 Dwell Time	duration: 3% +, CTR 1.5%-	展示结果中长文本比例增加

知乎 News Feed 后端策略演进

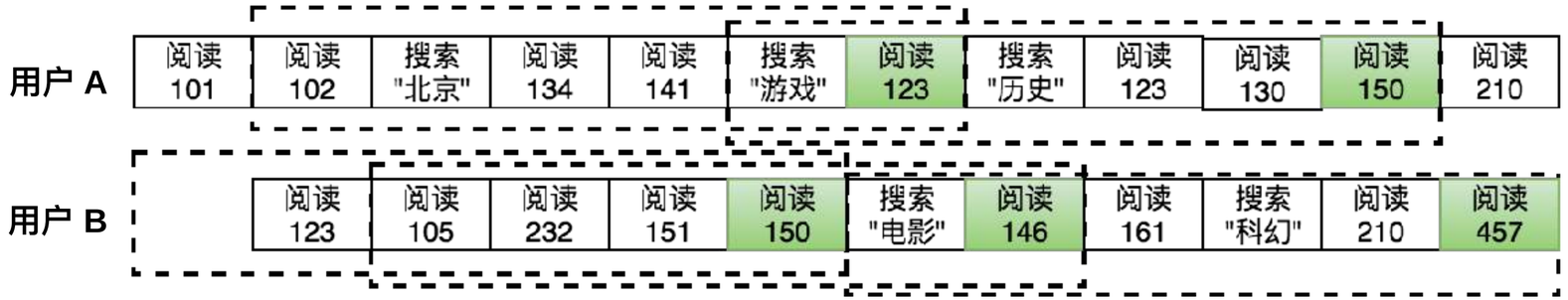
DNN Based Recommendation：Why and How?

- GBDT 模型的局限性
  - 容量有限
  - 无法有效利用 ID 类特征
- 业务的发展
  - 「被动式」信息需求越来越多；新用户及低频用户缺少统计信息
  - CF 推荐：数据过于稀疏；计算复杂性高
  - E&E 算法：模型过于简单，E&E 的效果受限于内容池划分、超参数等一系列问题
- DNN Based Recommendation
  - 对用户和内容进行特征 embedding 表示
  - 使用 LSH + KNN 召回

知乎 News Feed 后端策略演进

Deep Learning Based Recommendation: 样本收集

- Hold-Out 收集样本



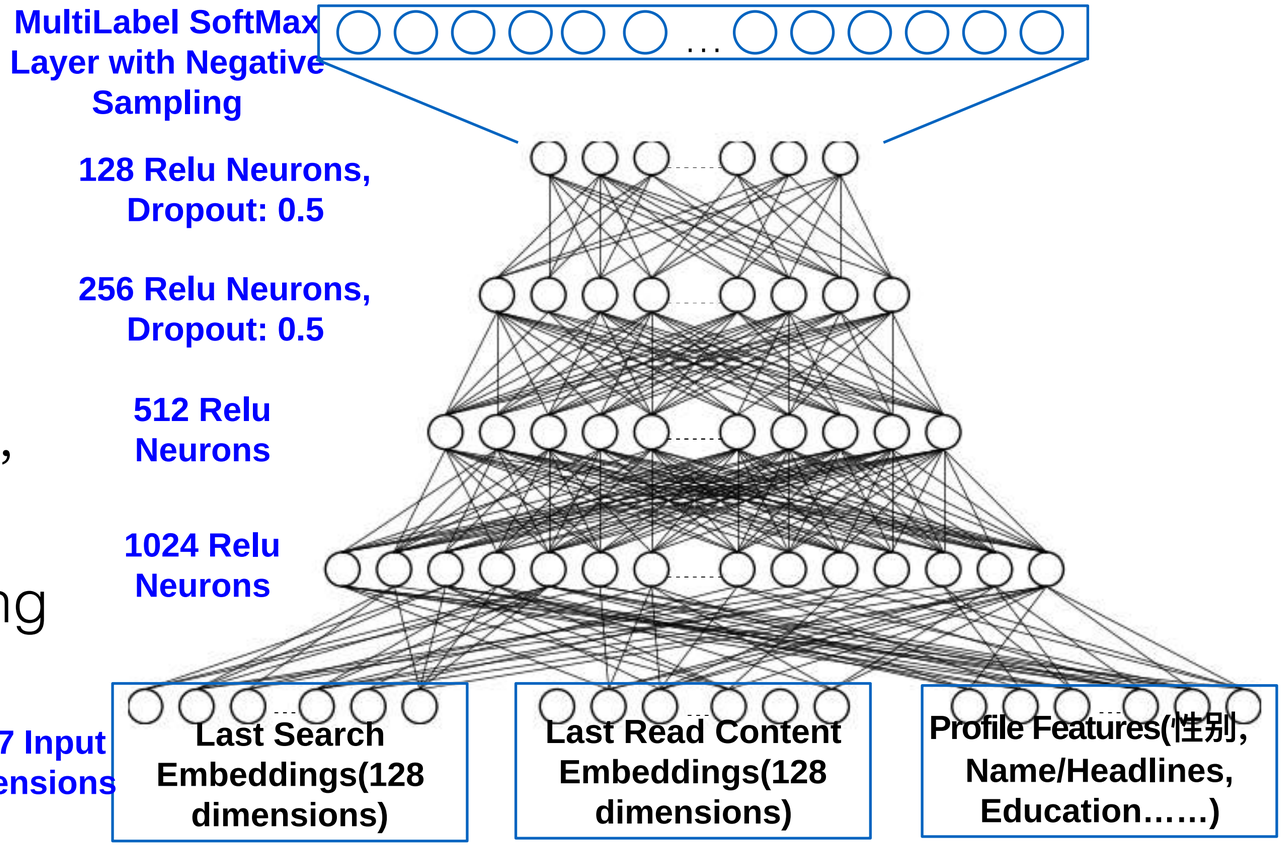
id	feature	label
1	Read: 102, 134, 141 Search: “北京”, “游戏”	123
2	Read: 123, 123, 130 Search: “游戏”, “历史”	150
3	Read: 123, 105, 232, 151 Search: Null	150
4	Read: 105, 232, 151, 150 Search: “电影”	146
5	Read: 146, 161, 210 Search: “电影”, “科幻”	457

- 重采样
  - 用户阅读内容分布非常不平衡（label 频次不均衡），会造成推荐结果偏向于热门
  - 对流量大的内容进行降采样：ratio = sqrt(C/freq)

知乎 News Feed 后端策略演进

Deep Learning Based Recommendation: 网络结构

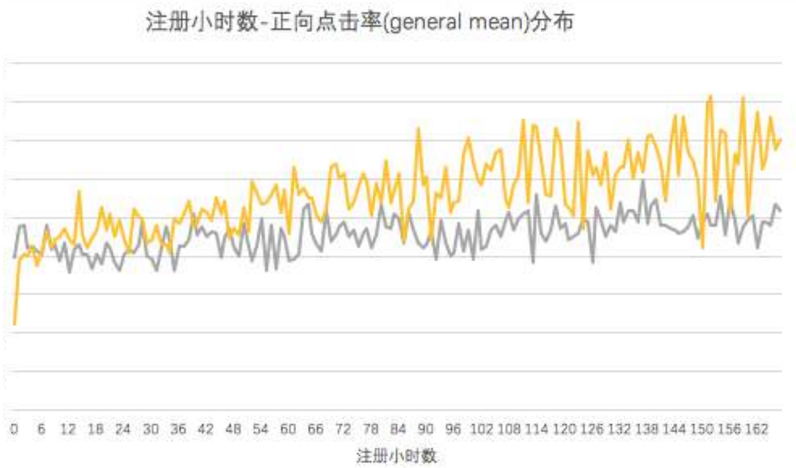
- 使用用户的 LastActions 及其 Profile 特征训练 User Embedding 网络，同时得到训练集中所有内容的 Embedding
- 使用训练好的 User Embedding 网络，固定网络连接的权重，在线推断出 User Embedding；同时使用单层 SoftMax 网络，得到已分发内容的 Embedding
- 利用 User Embedding 和 Feed Embedding 的结果进行内积运算，为用户推荐内容



知乎 News Feed 后端策略演进

Deep Learning Based Recommendation: 实验及改进（一）

- 整体效果
  - Top 100 ACC：0.26
  - 线上效果：初期持平 E&E 算法的效果；随用户行为量的增长，基于 DNN 的推荐展现出优势
- 时间衰减
  - Feed Embedding 的准确性随时间衰减，需要加入定期重训机制
- 网络复杂度对推荐效果的影响
  - 3 亿样本的情况下，4 层网络的效果较优



	2	3	4
2层 (1697-256-128)	0.173	0.181	0.177
3层 (1697-512-256-128)	0.194	0.207	0.205
4层 (1697-1024-256-128)	0.209	0.254	0.262
5层 (1697-1536-1024-512-256-128)	0.239	0.252	0.269

知乎 News Feed 后端策略演进

Deep Learning Based Recommendation: 实验及改进（二）

- 引入高频 ID 类特征及 FMPooling

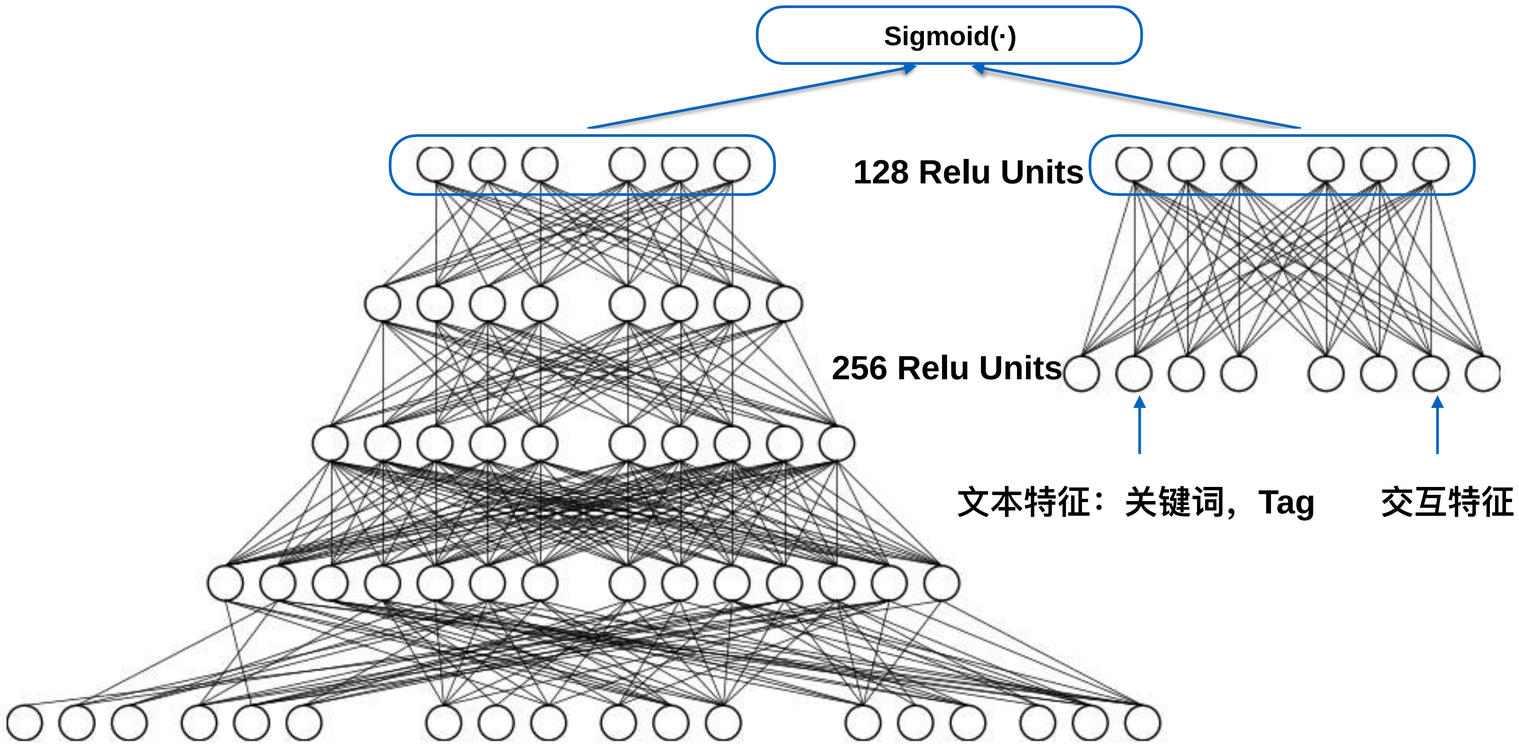
	原始网络	Top 2w QueryTag ID 化	Top 2w Query Tag ID 化+Top 2w 文章 Tag ID 化
Average	0.168	0.189	0.196
FMPooling	-	0.195	0.210

- 引入 Last - Skip 作为特征
  - 「展示但未读」可以做为负例也可以做为特征
  - Top 100 ACC: 0.26-0.29
- SoftMax 层改进
  - Negative Sampling 改成对 Skip 的内容采样

知乎 News Feed 后端策略演进

Deep Learning Based Recommendation: 实验及改进（三）

- 使用双神经网络实现对新内容的 Embedding 和推荐



Insights & Conclusions

建设业务友好的模型

- 从无到有
  - 架构和数据先行
  - 模型选择需要考虑可维护性、可扩展性及潜力
  - 模型殊途同归
- 从有到优
  - 从业务出发，设计目标、采样方式
  - 由合而分：由统一模型细化成针对各个目标的「小」模型
- 工程质量
  - 重视接入新特征和新数据的模型迭代速度

Thanks!