一点资讯技术编程大赛

CTR预估

西安交通大学

八月无烦恼

2021年9月

目　录

CONTENTS

团队介绍

西安交通大学
跨媒体知识融合与工程应用研究所

刘启东
自动化科学与技术

赵成成
计算机技术

马黛露丝
计算机科学与技术

目　录

CONTENTS

赛题理解

- 数据
  - 用户侧：年龄、性别、省市、设备信息…
  - 文章侧：标题、类别、关键词、发布时间…
  - 交互侧： 是否点击、消费时长、展现时间、刷新次数…
- 目标
  - 预测用户对文章的点击率（CTR）
- 评分标准

$$AUC = \frac{\sum_{ie\ positiveClass} rank_i - \frac{M(1+M)}{2}}{M \times N}$$

- 理解
  - 典型的推荐场景、CTR预估

目　录

CONTENTS

3.1 特征工程

数据分析

- 数据概况

名称	用户	文章	交互
数量	1,538,384	633,391	189,766,959

- 检查空缺值
  - 用户和文章数据有少量缺失
  - 训练集和测试集都不存在空缺值
- 检查冷启动
  - 验证发现，训练集和测试集内都不存在冷启动问题

3.1 特征工程

数据预处理

- 类别特征
  - One-hot：LabelEncode后输入到embedding层
  - Multi-hot：多个类别embedding加权平均
    - ✓ **定长**：用户的年龄、性别，均值填充空缺值
    - ✓ **不定长**：文章的关键词，零值填充空缺值，概率归一化
- 历史交互特征
  - 截取用户最近点击的15个文章

3.2 模型介绍

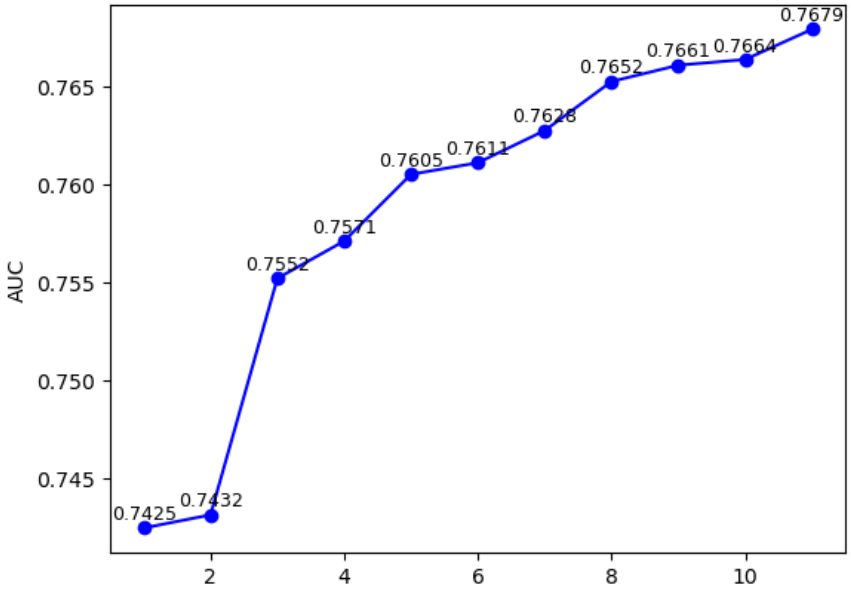
树模型

- 树模型（LightGBM）
  - base： 0.653927
  - 加入用户侧统计特征： 0.71894（+6.5个百分点）
- 缺点：
  - 全量训练
  - 内存消耗大（190G+）
  - 训练周期长（30h+）

3.2 模型介绍

模型概览

- 深度模型
  - DeepFM
  - xDeepFM
  - xDeepFM + DIN
  - xDeepFM + ESMM



主要改进方法	效果
DeepFM baseline	0.74250
加入性别和年龄特征	+6.7个万分点
调整batch size 和学习率	+1.2个百分点
加入文章二级分类	+1.9个千分点
将DeepFM替换为xDeepFM	+4.5个千分点
加入ESMM和keywords	+1.2个千分点
调整Embedding和cin的大小	+2.5个千分点
加入用户历史行为（10）	+8.2个万分点
xDeepFM & DIN	+2.9个万分点
模型融合	+1.5个千分点
--	0.76791

3.2 模型介绍

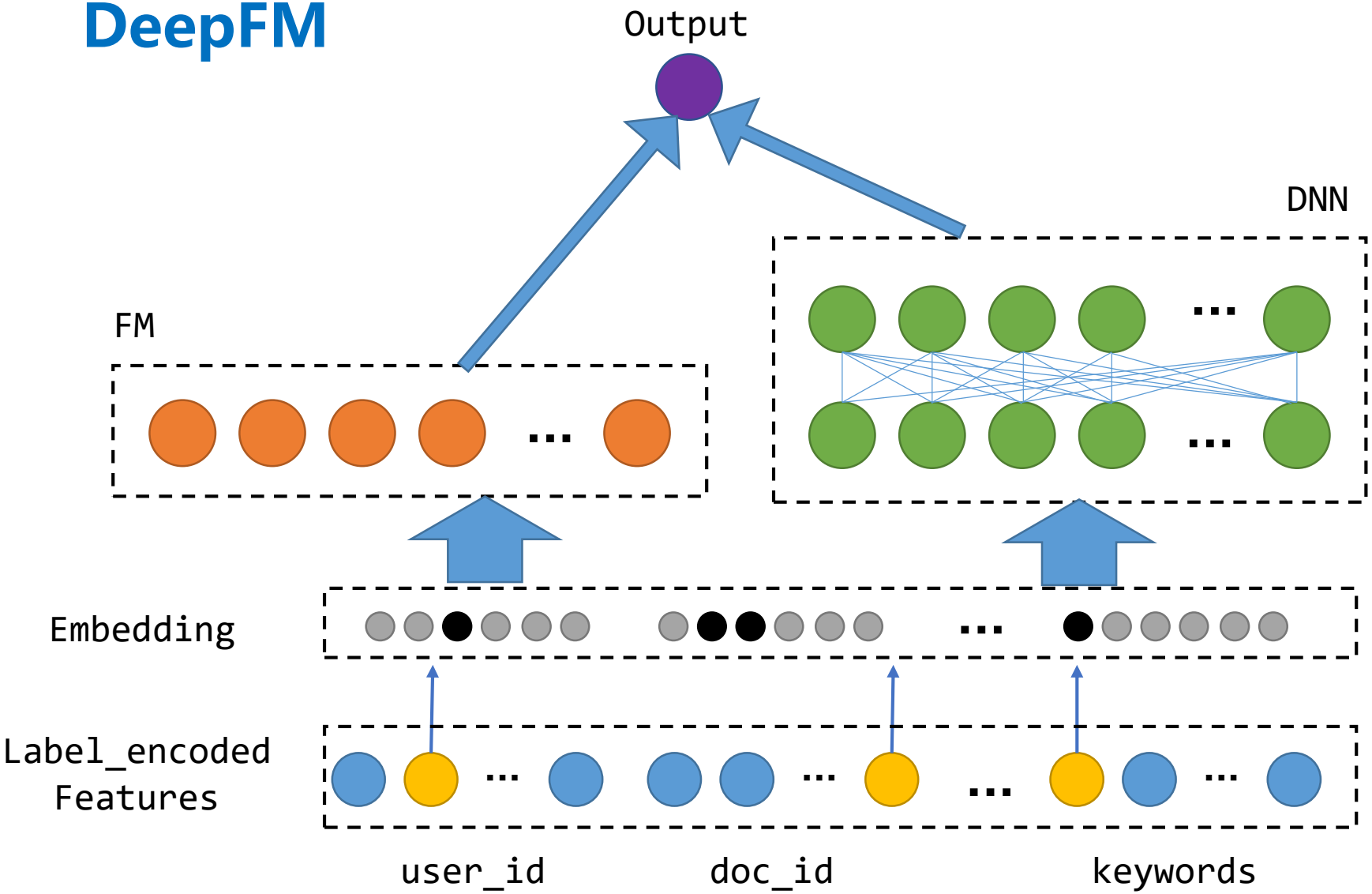
DeepFM

- 动机：
  - 易于实现
  - 不需要大量人工特征工程
  - 可以同时学习低阶和高阶的组合特征
  - FM模块和Deep模块共享Embedding层，可以使模型更快收敛

[1] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: A factorization-machine based neural network for CTR prediction. In Proceedings of the IJCAI. 2782--2788.

3.2 模型介绍

DeepFM



3.2 模型介绍

DeepFM

- 模型效果
  - base：0.74250
  - 加入用户性别和年龄：0.74317（+6.7个万分点）
  - 调整batch size和学习率：0.75522（+1.2个百分点）
  - 加入文章类别：0.75712（+1.9个千分点）

3.2 模型介绍

xDeepFM

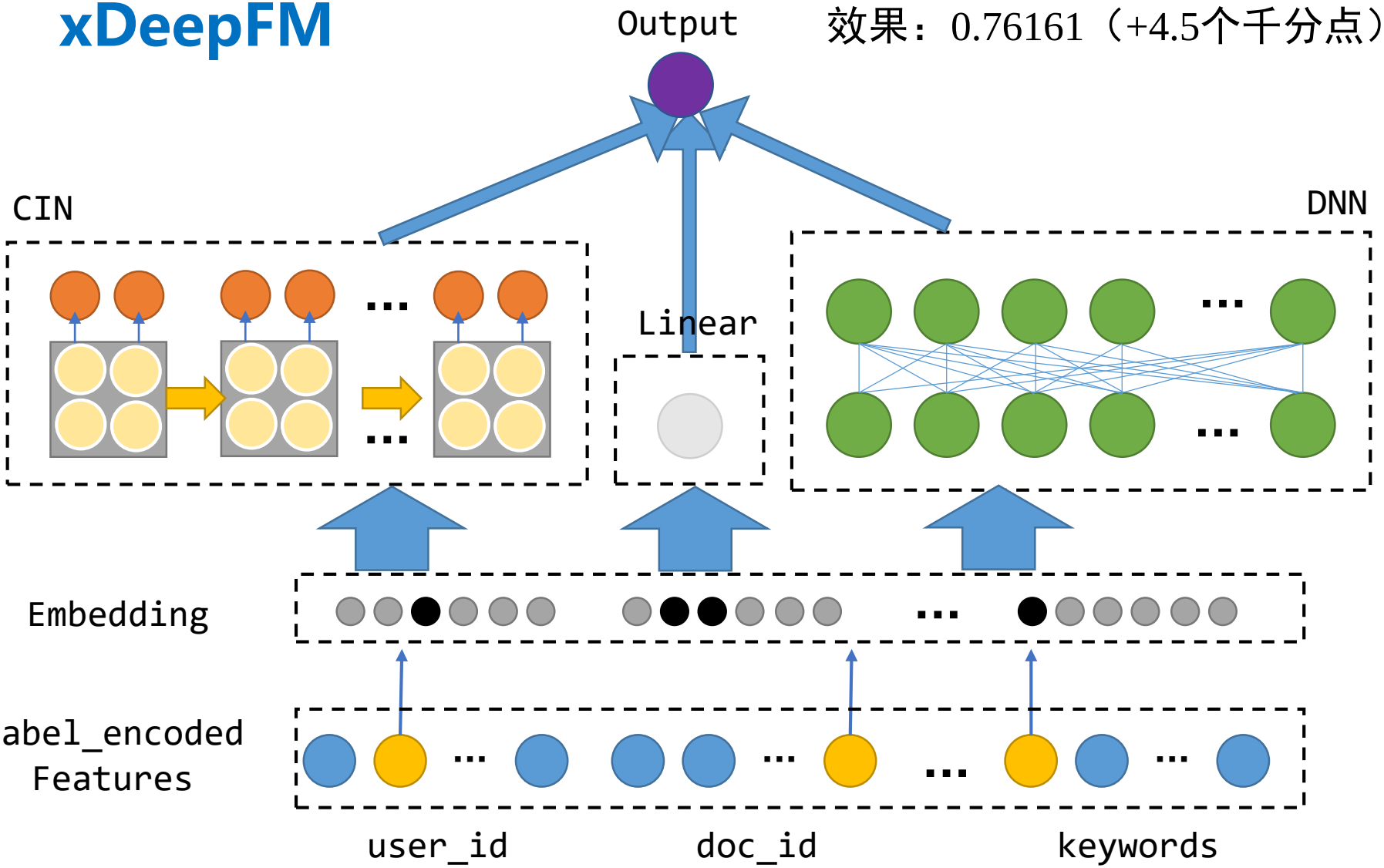
- 动机：
  - 可以学习显式高阶特征交互（CIN）
  - 有效结合了显式高阶交互模块、隐式高阶交互模块
  - vector-wise level的特征组合（相比bit-wise）

[2] Jianxun Lian, Xiaohuan Zhou, Fuzheng Zhang, Zhongxia Chen, Xing Xie, and Guangzhong Sun. 2018. xDeepFM: Combining explicit and implicit feature interactions for recommender systems. arXiv preprint arXiv:1803.05170 (2018).

3.2 模型介绍

xDeepFM

效果：0.76161（+4.5个千分点）



3.2 模型介绍

xDeepFM + ESMM

➢ 动机：

- 如何使用用户消费时长特征
- 消费时长目标同点击目标之间存在关联关系→多任务
- 点击和消费时长的关系与点击和转化的关系是相同的

$$\underbrace{p(y=1,z=1|\boldsymbol{x})}_{pCTCVR} = \underbrace{p(y=1|\boldsymbol{x})}_{pCTR} \times \underbrace{p(z=1|y=1,\boldsymbol{x})}_{pCVR}$$

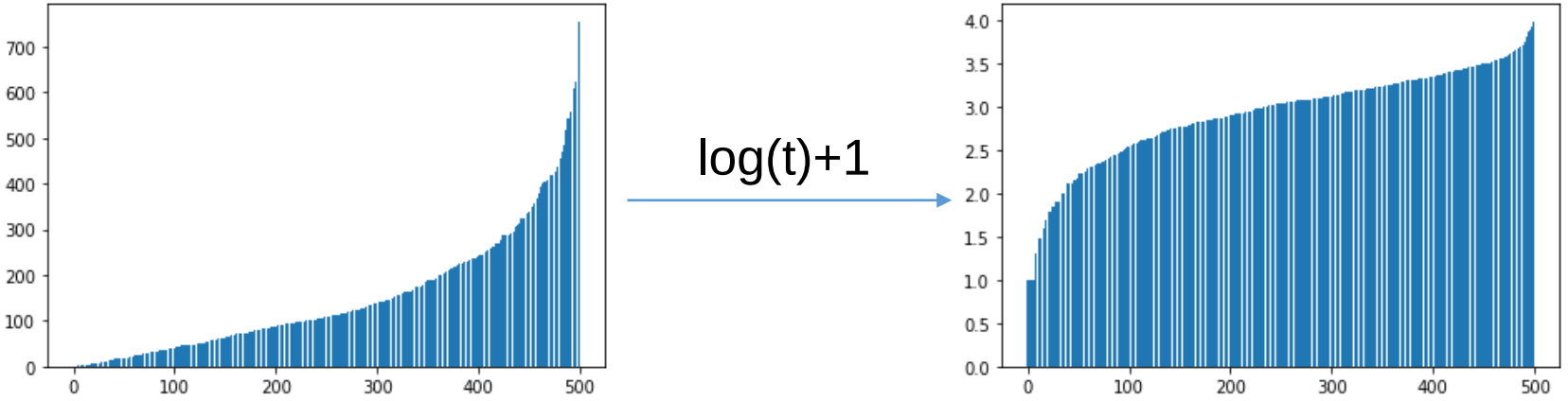
[3] Xiao Ma, Liqin Zhao, Guan Huang, Zhi Wang, Zelin Hu, Xiaoqiang Zhu, and Kun Gai. 2018. Entire space multi-task model: An effective approach for estimating post-click conversion rate. In SIGIR.

3.2 模型介绍

xDeepFM + ESMM

用户消费时长特征：

- 时间跨度大，消费时长分布呈长尾分布状
- 取对数加一： $t = \log(t) + 1$

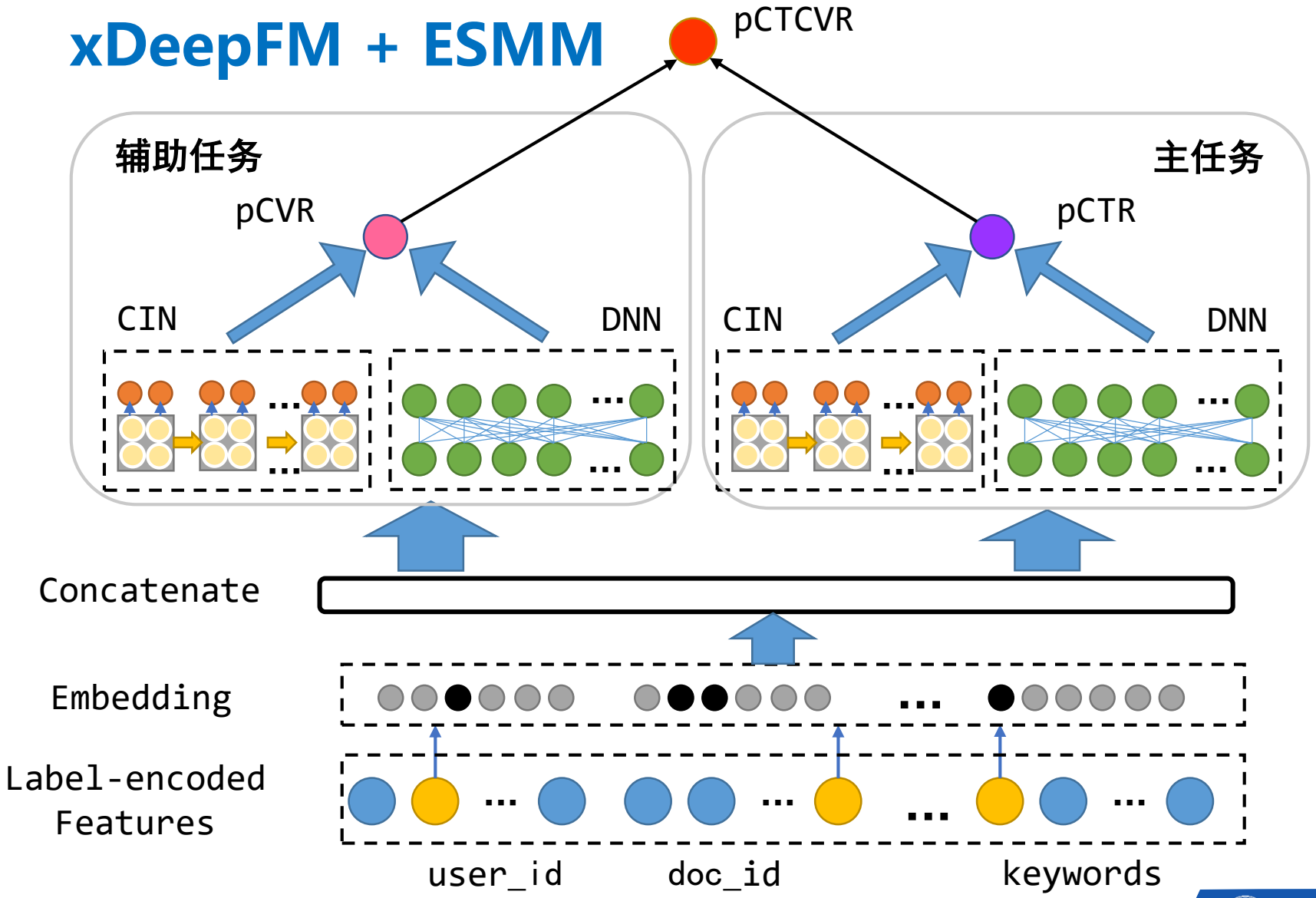


采样数据原始的消费时长分布

采样数据处理后消费时长分布

3.2 模型介绍

xDeepFM + ESMM



3.2 模型介绍

xDeepFM + ESMM

- 模型效果
  - 加入ESMM和keywords： 0.762752（+1.2个千分点）
  - 调整embedding size和cin size： 0.765252（+2.5个千分点）

3.2 模型介绍

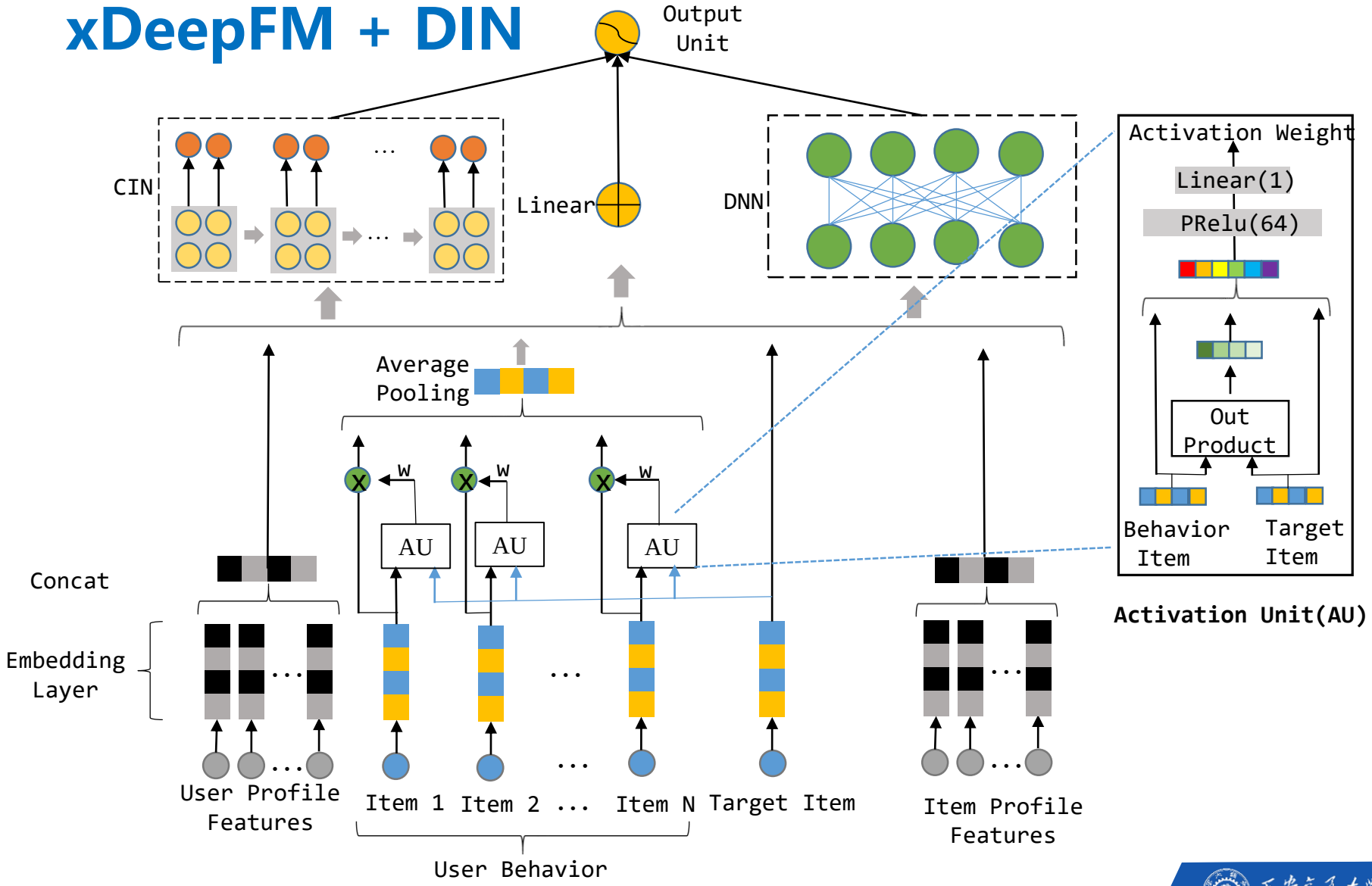
xDeepFM + DIN

- 动机：
  - 用户的兴趣是多元化的
  - 用户是否点击，取决于历史行为数据中的小部分，而不是全部

[4] Guorui Zhou, Chengru Song, Xiaoqiang Zhu, Xiao Ma, Yanghui Yan, Xingya Dai, Han Zhu, Junqi Jin, Han Li, and Kun Gai. 2017. Deep Interest Network for Click-Through Rate Prediction. arXiv preprint arXiv:1706.06978 (2017).

3.2 模型介绍

xDeepFM + DIN



3.2 模型介绍

xDeepFM + DIN

- 模型效果
  - 加入用户行为序列 (10)：0.766084（+8.3个万分点）
  - 加入Attention (15)：0.766377（+3个万分点）

注：对于没有点击行为序列的用户，在Embedding词典中保留一列词向量，学习该类用户的历史行为

3.3 超参调优

GridSearch

- 对于模型中的超参数的所有可能取值，利用组合的方式，找到效果最好的的超参数值组合
- Batch size：

  1024, 2048, 4096, **8192**, 16384, 32768
- Learning rate：

  0.01, 0.005, **0.001**, 0.0005
- Embedding size：

  16, 32, **64**, 128, 256, 512
- CIN size：

  [10]*3, [10]*5, **[20]*3**, [20]*5

3.4 其他

- 构建稳定的线下验证
  - 线下训练集：前11天的交互记录
  - 线下验证集：第12天交互记录中随机采样5万条
  - 线下验证结果：5次随机采样预测结果的平均
- 内存优化
  - 使用子类型优化数值列以降低内存消耗
  - batch内拼接用户特征和文章特征
- 模型融合
  - 均值融合法

目　录

CONTENTS

4.1 工作总结

- 构建稳定的线下验证
- 优化内存占用空间
- 年龄、性别、关键词当作mutil-hot特征使用
- 使用xDeepFM完成vector-wise的特征交叉
- 将消费时长的预估作为辅助任务辅助CTR任务
- 在xDeepFM基础上融入DIN，实现用户多样化兴趣的局部激活

4.2 尝试未果

- 将文章展现时间细化为星期、小时
- 添加统计特征，如用户刷新次数均值、用户消费时长均值、用户最多点击的文章一级类别、用户最多点击的文章二级类别等
- 使用BERT预训练模型得到文章标题的Embedding，然后使用PCA降维

4.3 展望

- 构造更有用的特征
- 增加用户历史行为序列的长度
- 使用DIEN对用户兴趣演化进行建模

谢谢！