BANJITINO

张琦、郑凌瀚、王黎翔

2020-08-03

Contents

目录

1 团队介绍

2 赛题理解

3 特征工程

3.1 词表处理

3.2 序列构造

3.3 Embedding

4 模型介绍

4.1 单模型

4.2 Ensemble

5 总结与思考

5.1 上分之路

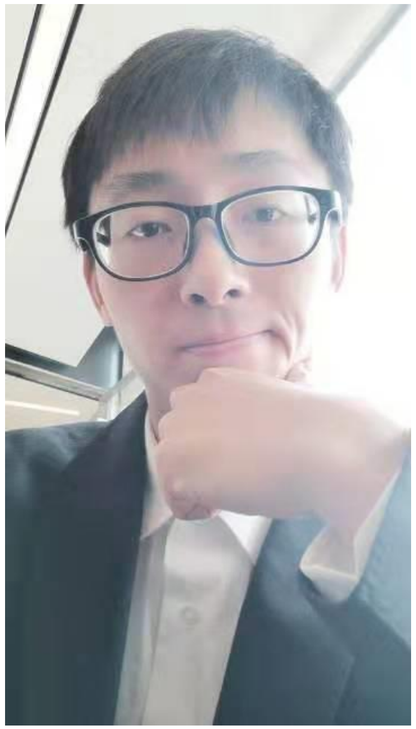
5.2 总结与思考

团队介绍

团队介绍

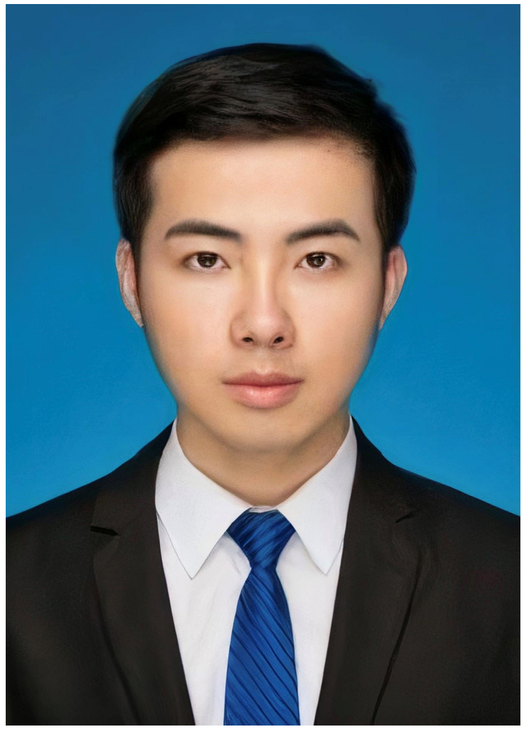
张琦
中国科学院大学 硕士
算法工程师

负责Pipeline构建、Embedding及单模型开发与融合



郑凌瀚
同济大学 硕士
算法工程师

负责EDA、Embedding及单模型开发与融合



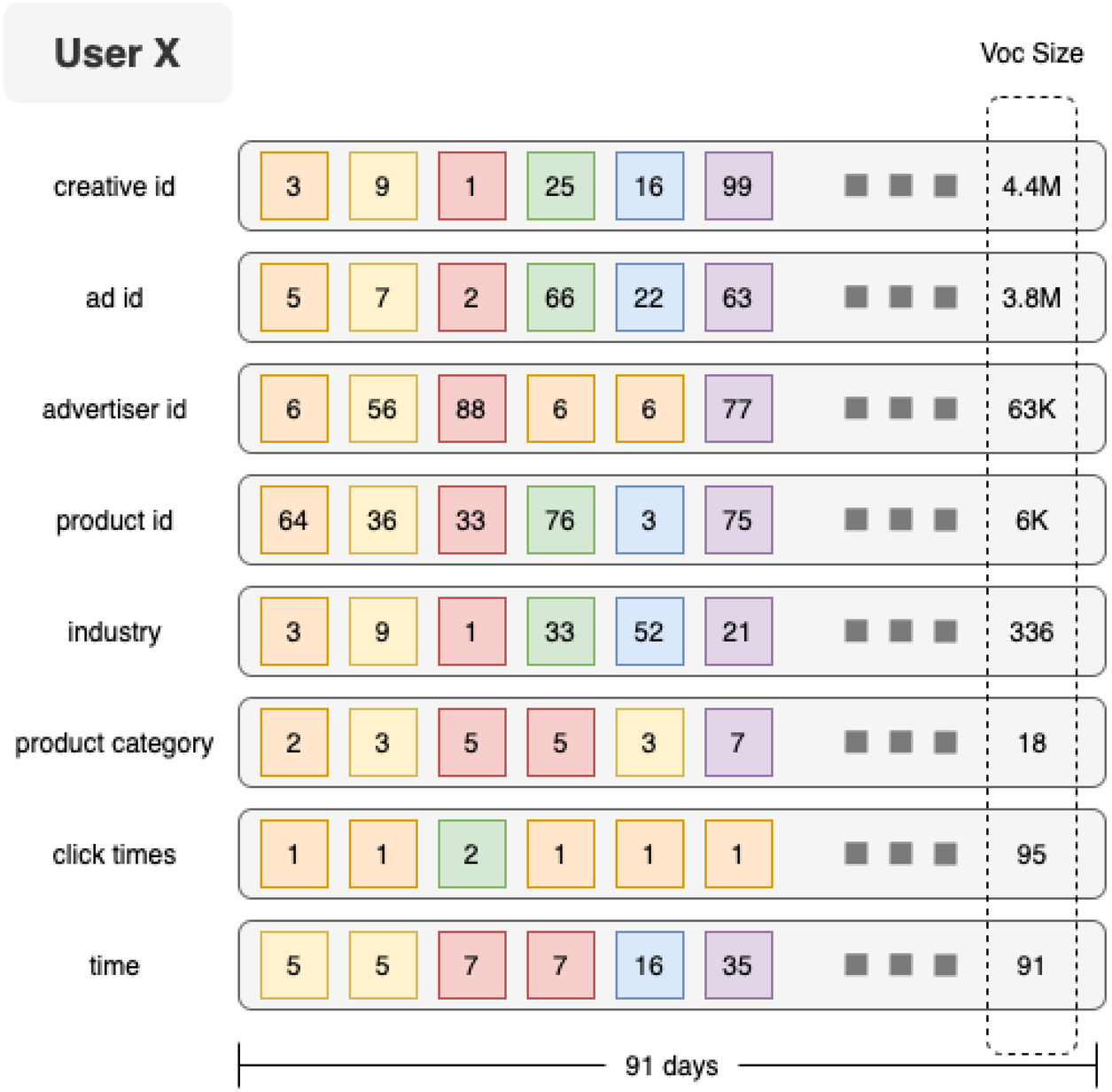
王黎翔
东南大学 研一在读
研究方向：图像理解/检索

负责CNN系单模的开发和训练

02

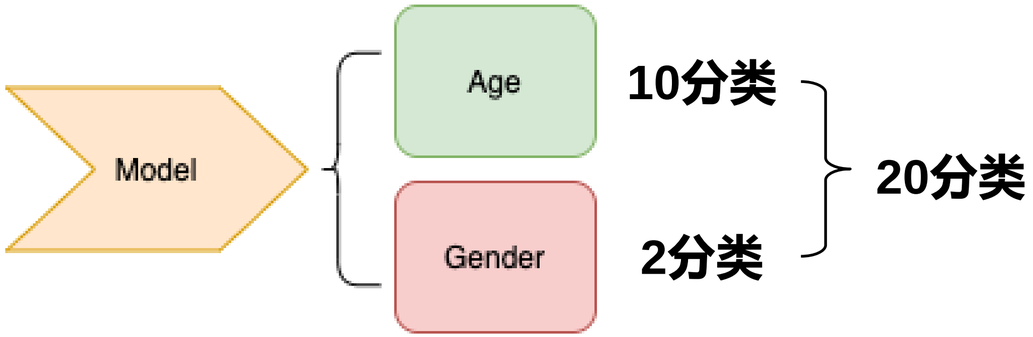
赛题理解

赛题理解



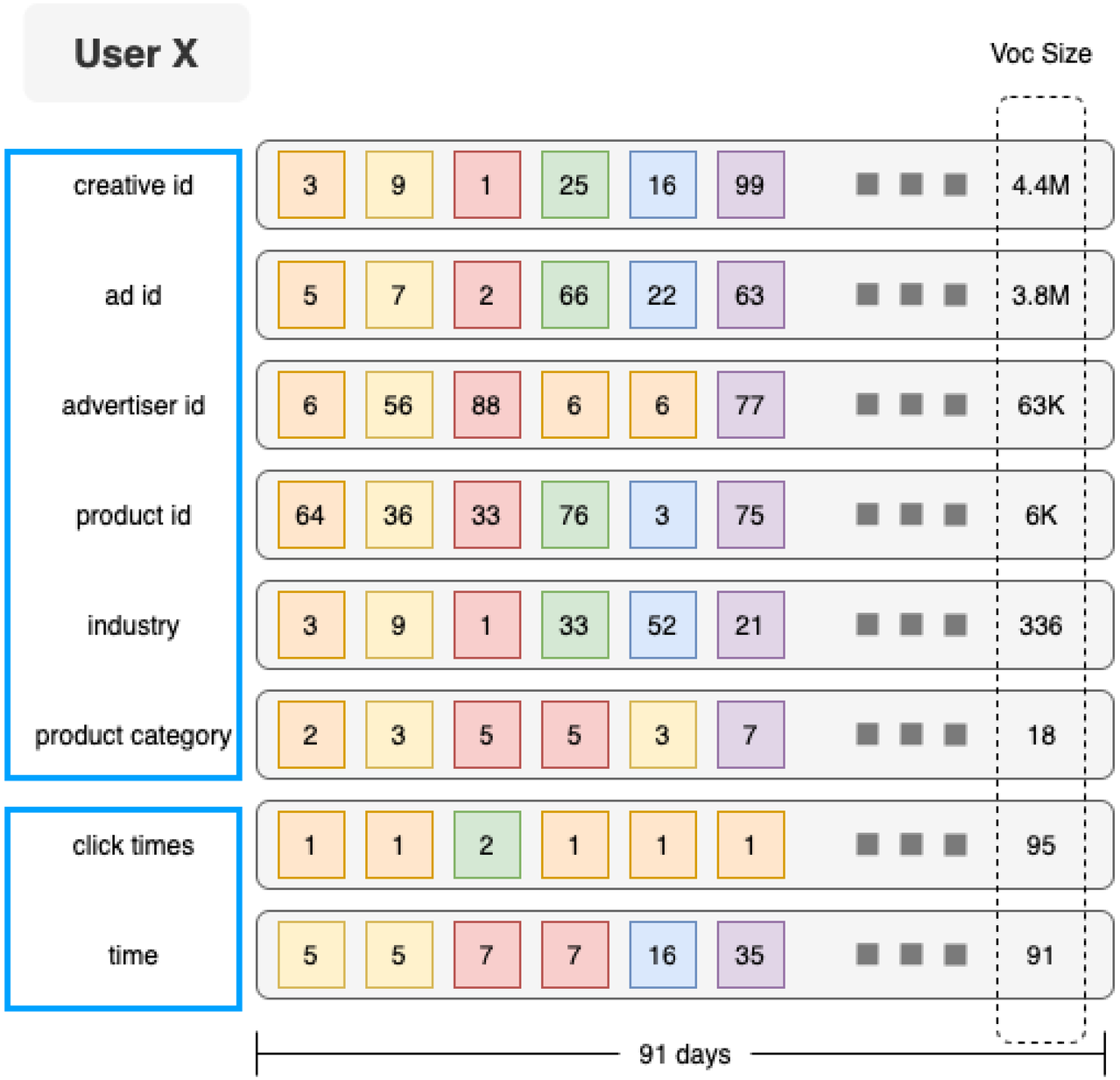
用户行为序列预测用户人口统计学属性

91天用户行为序列 => Age & Gender
Train: 300W用户　Test: 100W用户



类似于NLP中的文本分类问题

赛题理解



主键

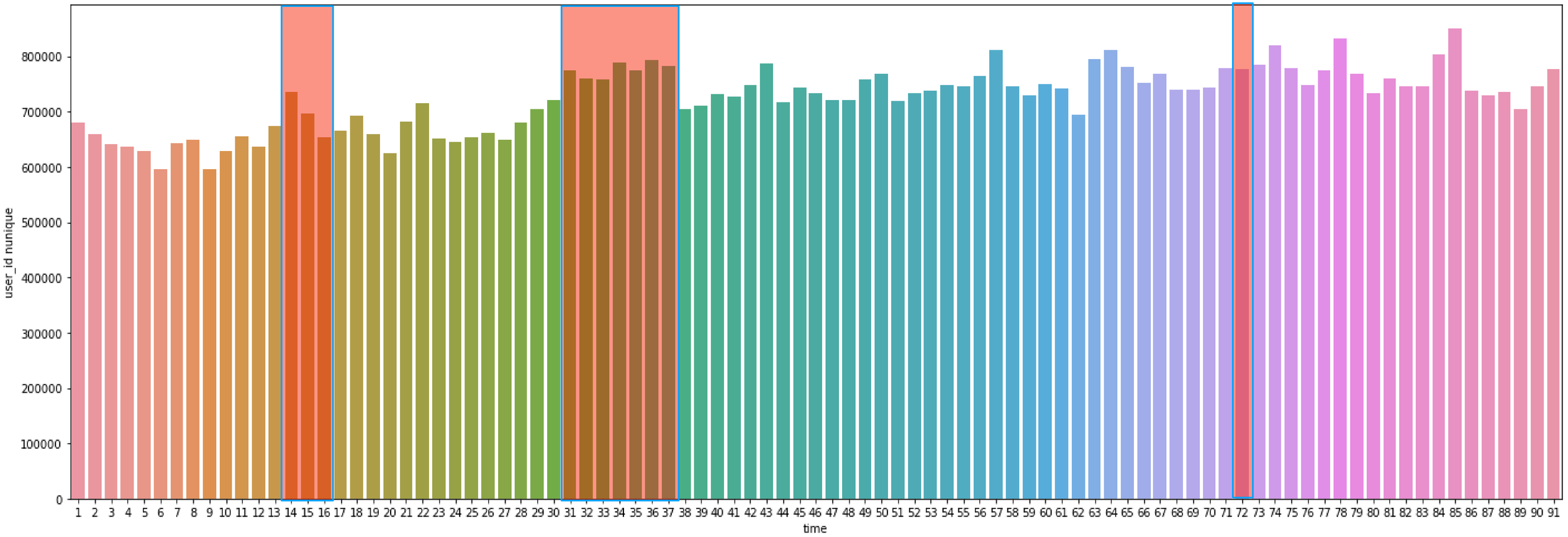
广告素材的相关属性
6个基本输入序列

Huge vocabulary size
Hard to make it end2end
Pretrain embedding …

可反应用户兴趣，可用于增强点击序列信息

时序，用户行为先后/ID出现的先后可进行分析和反脱敏

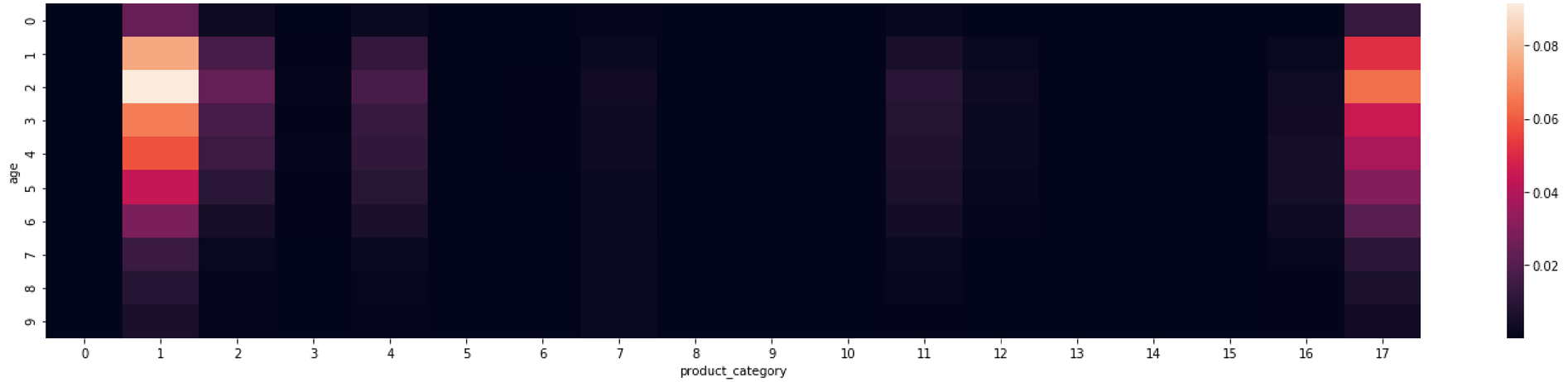
赛题理解



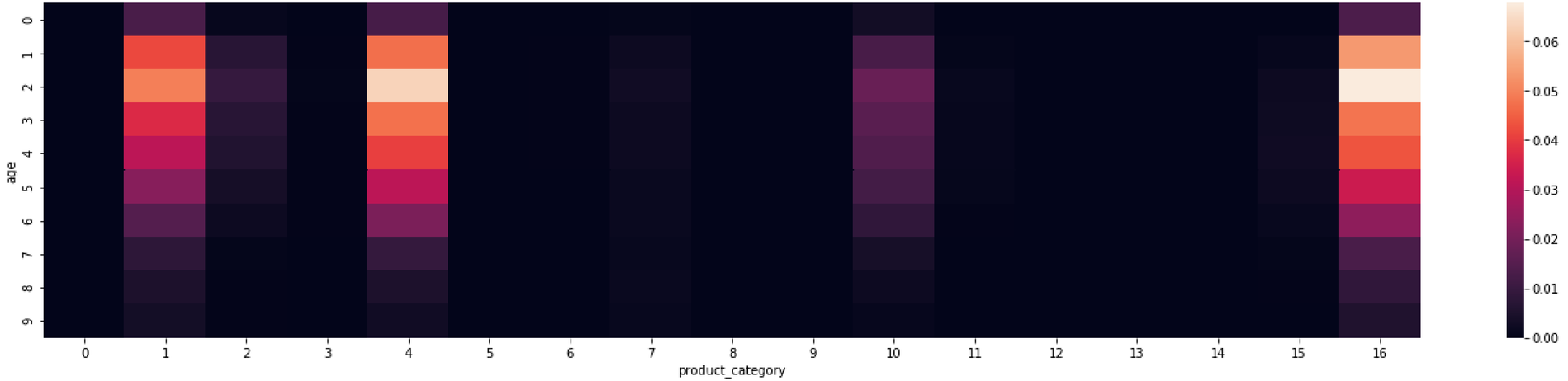
统计日活用户数量进行分析，反脱敏得出time的分布应为2019年9月1日到2019年11月30日期间有中秋节，国庆节，重阳节，学生节，双十一，感恩节等诸多特殊节假日

赛题理解

总体数据内不同age(Y轴) 用户点击不同product category(X轴)分布



双11 当天不同age(Y轴) 用户点击不同product category(X轴)分布



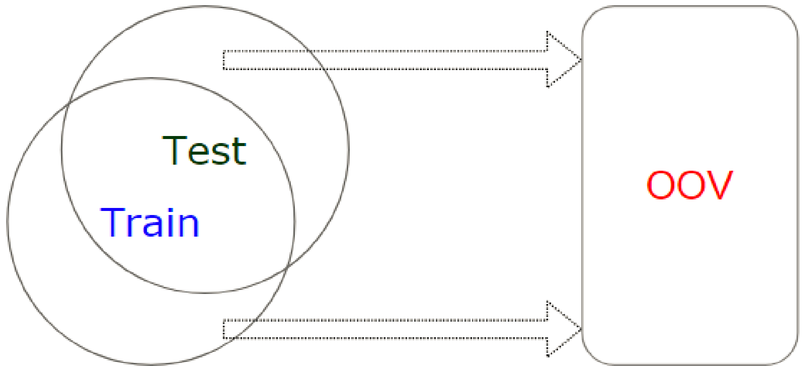
节假日用户点击广告素材的信息更为丰富，区分度更大
不同特殊日期内，用户年龄性别与用户广告点击行为的联合分布是不同的
构建用户行为序列中必须考虑到发生行为对应的日期，以及该日期内用户行为的特性

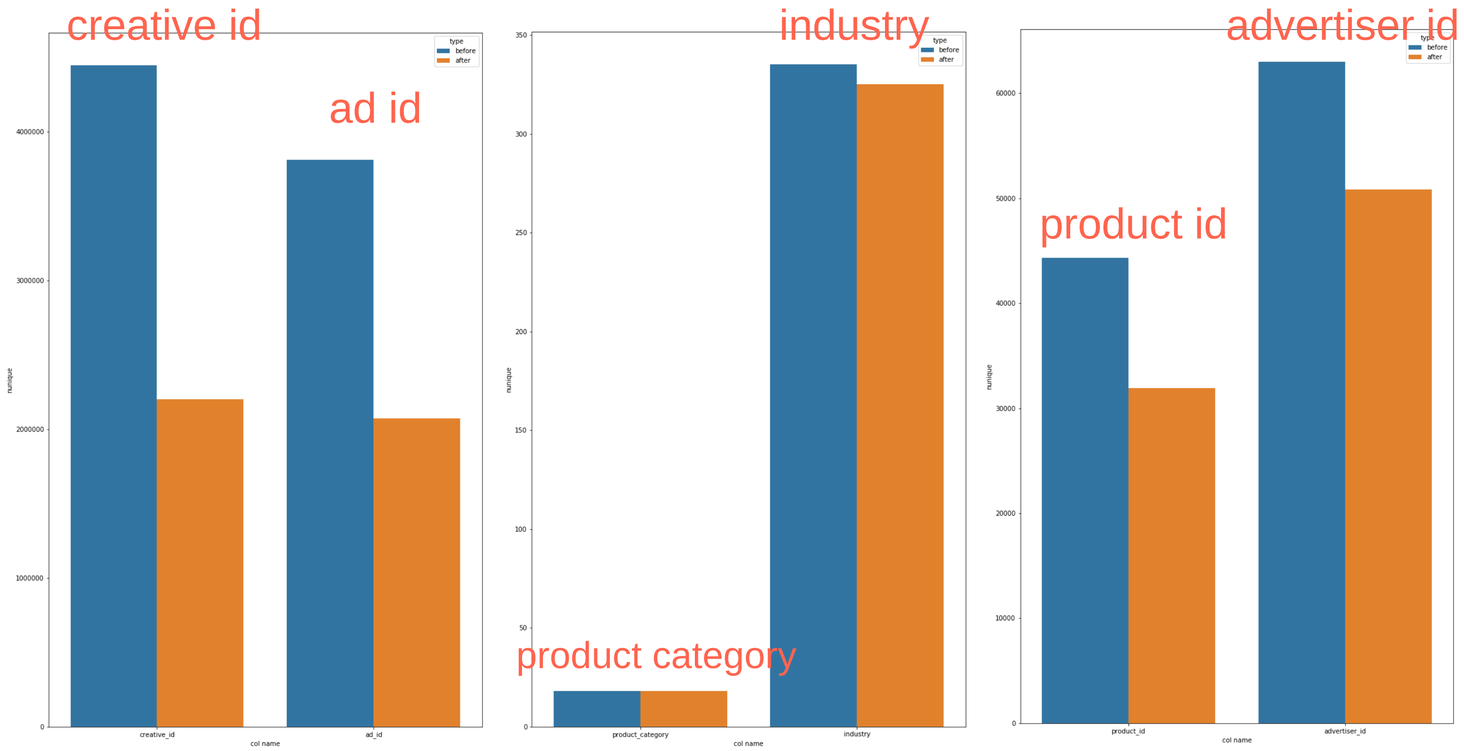
特征工程

词表处理/序列构造/Embedding

特征工程-词表处理

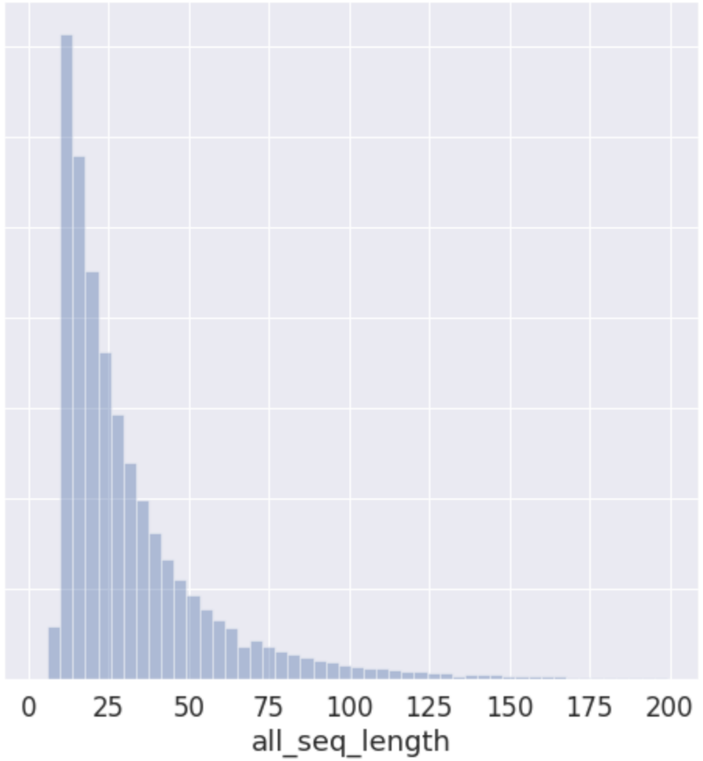
词表内仅在train或者test出现过1次的词，视为OOV，用统一的ID替代

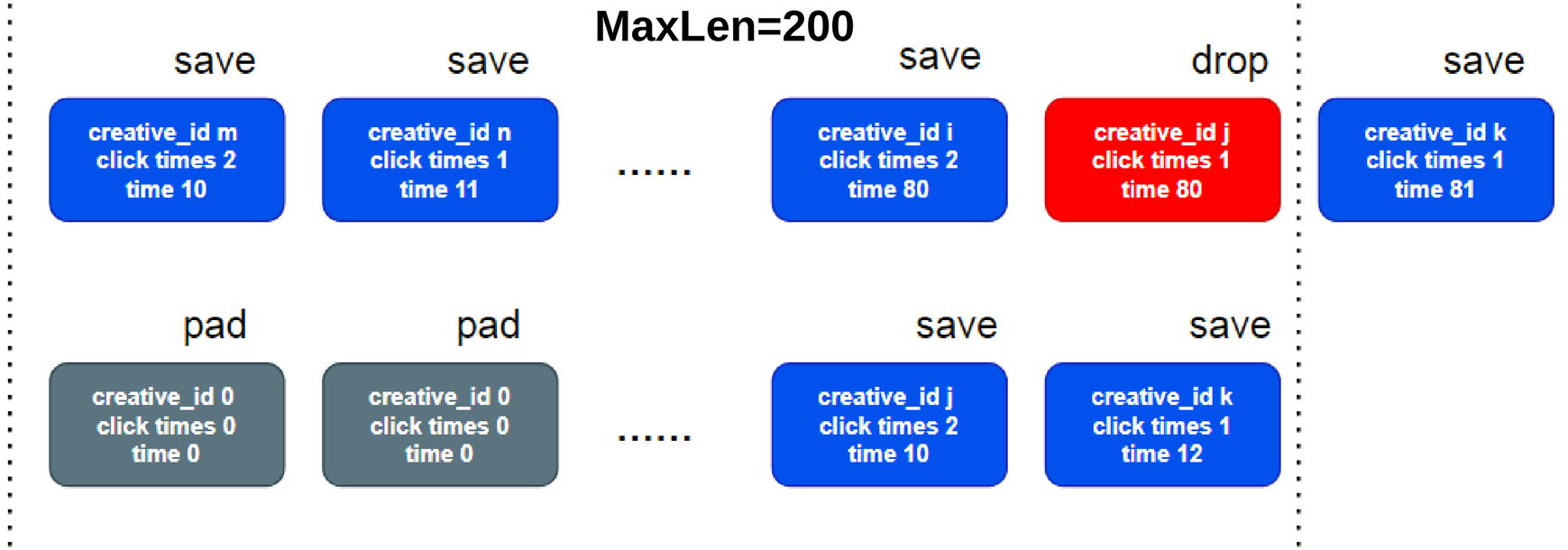




词表大幅缩小，embedding矩阵大幅缩小
节约内存，提高运算速度

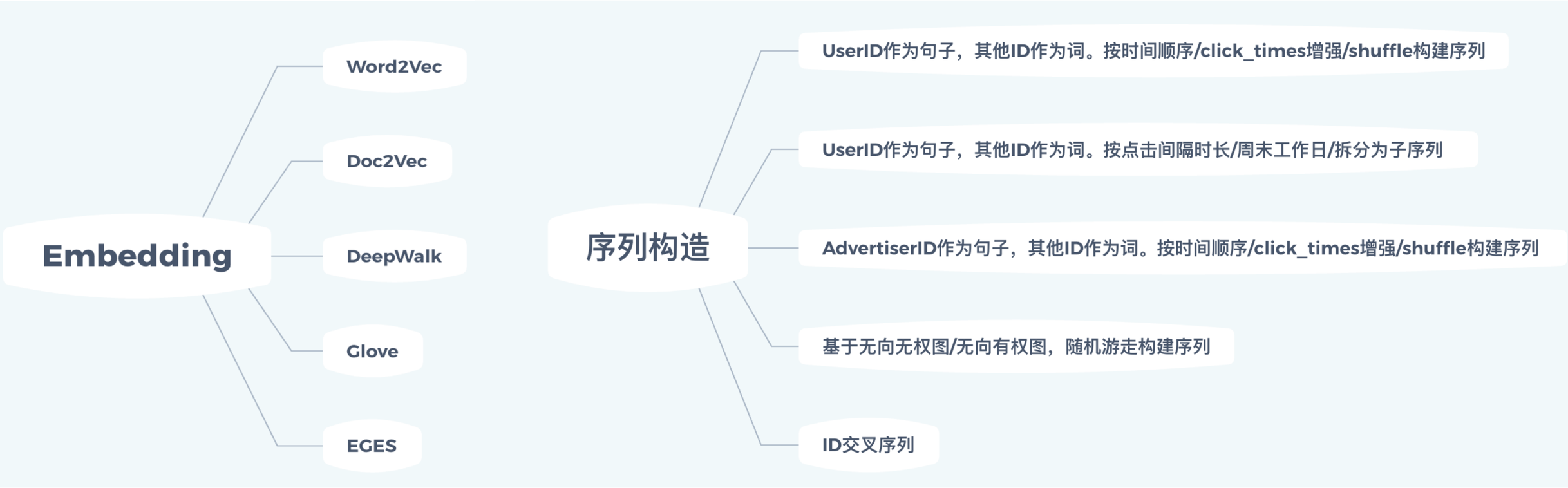
特征工程-序列构造



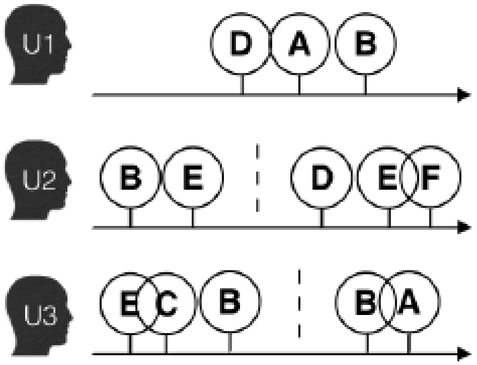


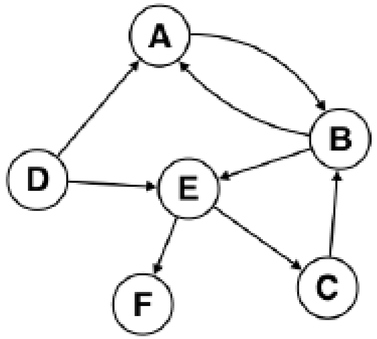
- 序列Max Len=200 [99%:157 | 99.5%:192 | 99.9%:294]
- Padding & Truncating: Pre
- 优先保留同一天内点击次数较多的广告素材
- 序列增强1：Shuffle+逆序+截断+Skip
- 序列增强2：TF-IDF低的随机drop

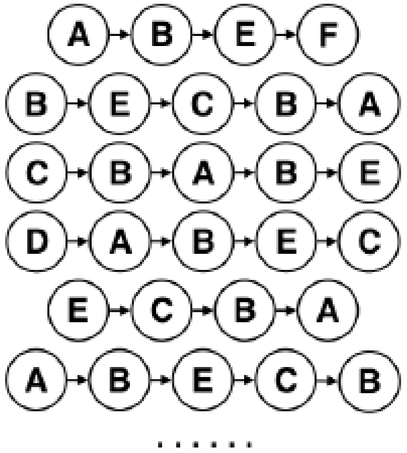
特征工程-Embedding

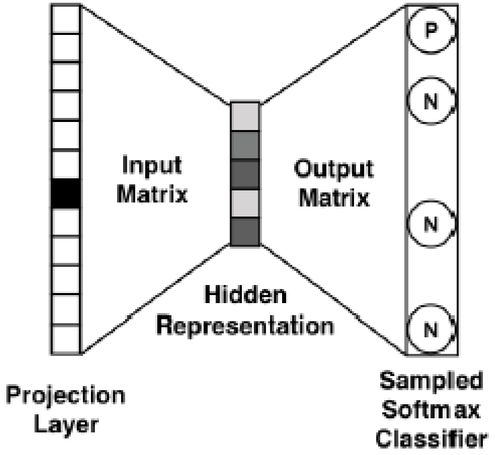


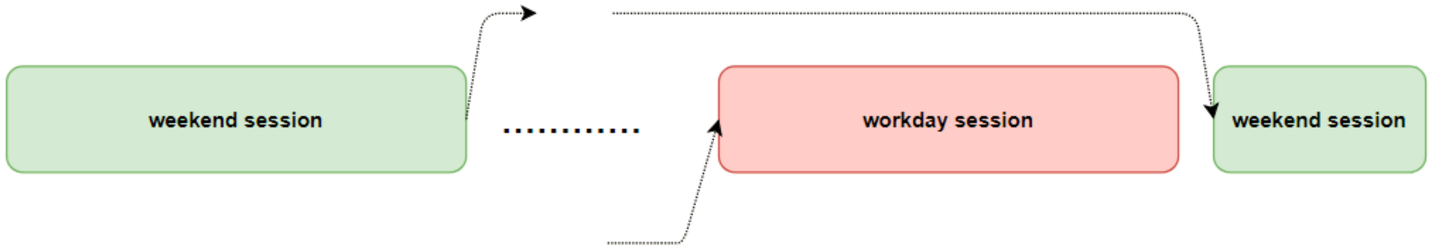
特征工程-Embedding

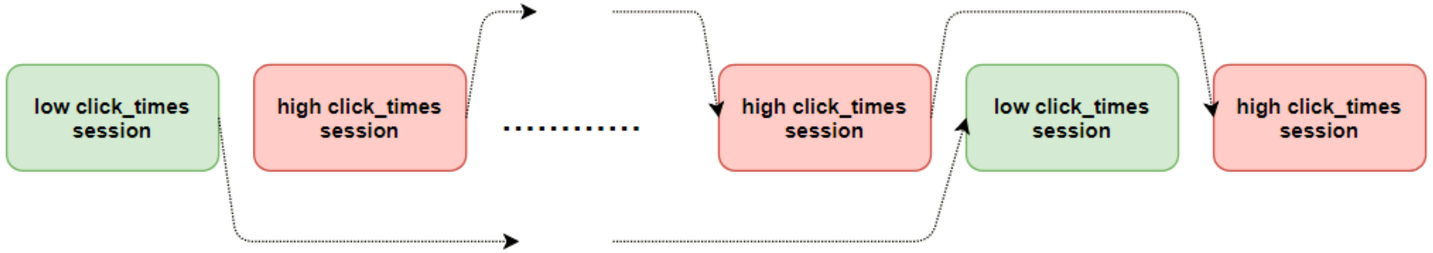










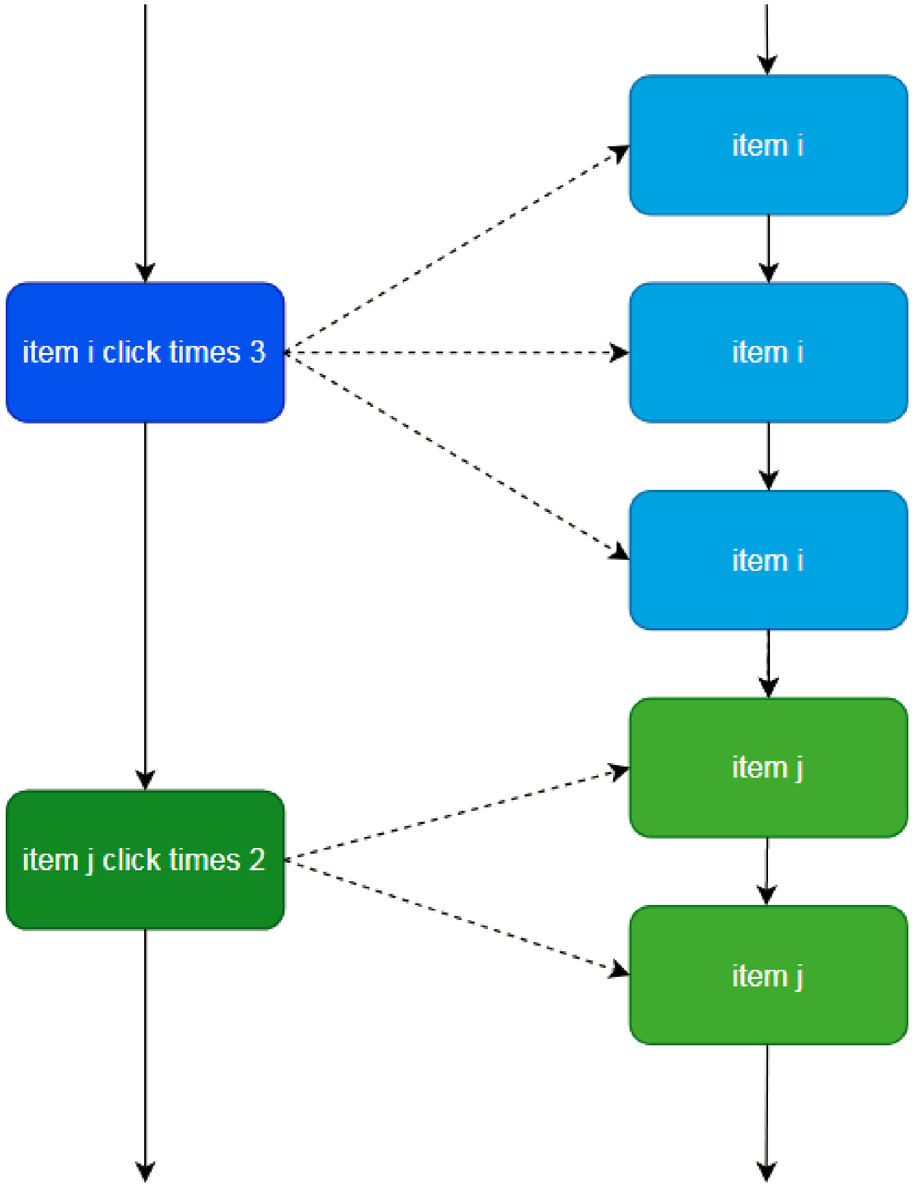


差异化预训练序列构建

- 按照工作日休息日划分
- 按照点击次数高低划分
- 按照点击行为时间间隔长短划分

https://dl.acm.org/doi/10.1145/3219819.3219869

特征工程-Embedding

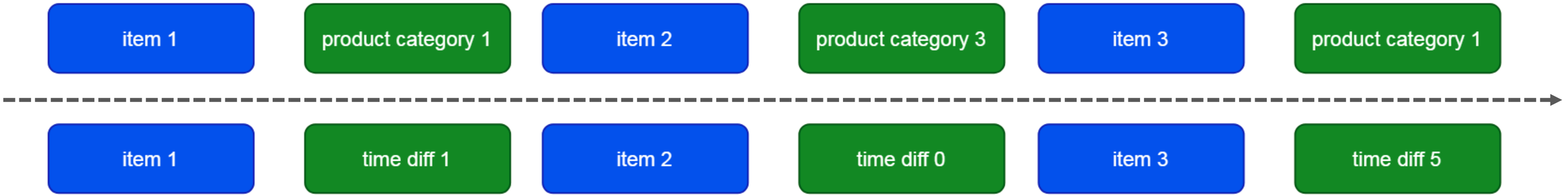


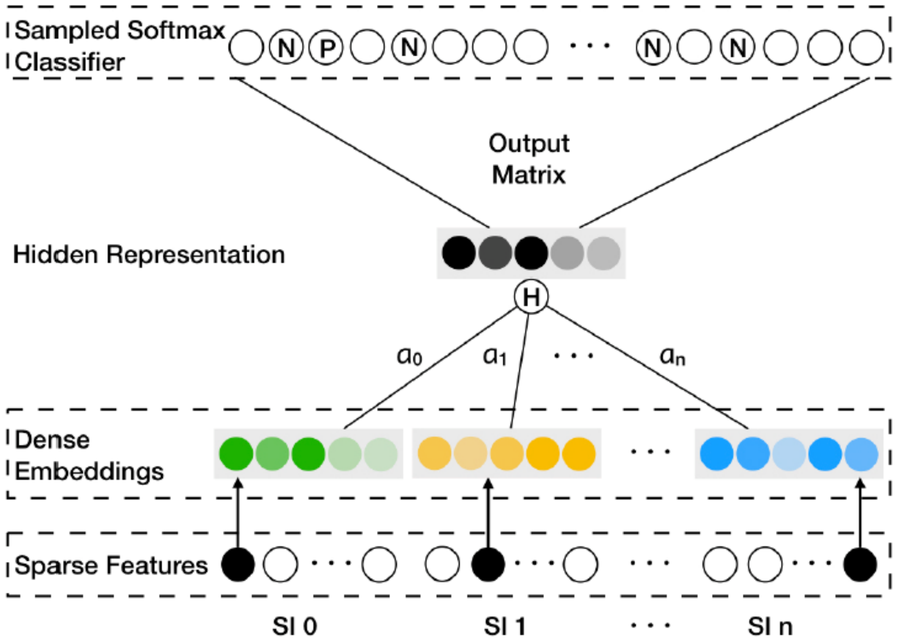
增强序列

点击次数反应用户的兴趣，根据用户对广告素材的点击次数对数据进行增强（复制+Shuffle）

使得总点击日志从1.34亿行增幅到2.5亿行。对增幅后的点击日志做词向量预训练可以在共现信息中包含用户点击广告素材的频率信息

特征工程-Embedding





EGES

将广告product category、industry及距离下次点击发生的时间差time diff 作为side information，用weighted SG训练得到embedding

https://dl.acm.org/doi/10.1145/3219819.3219869

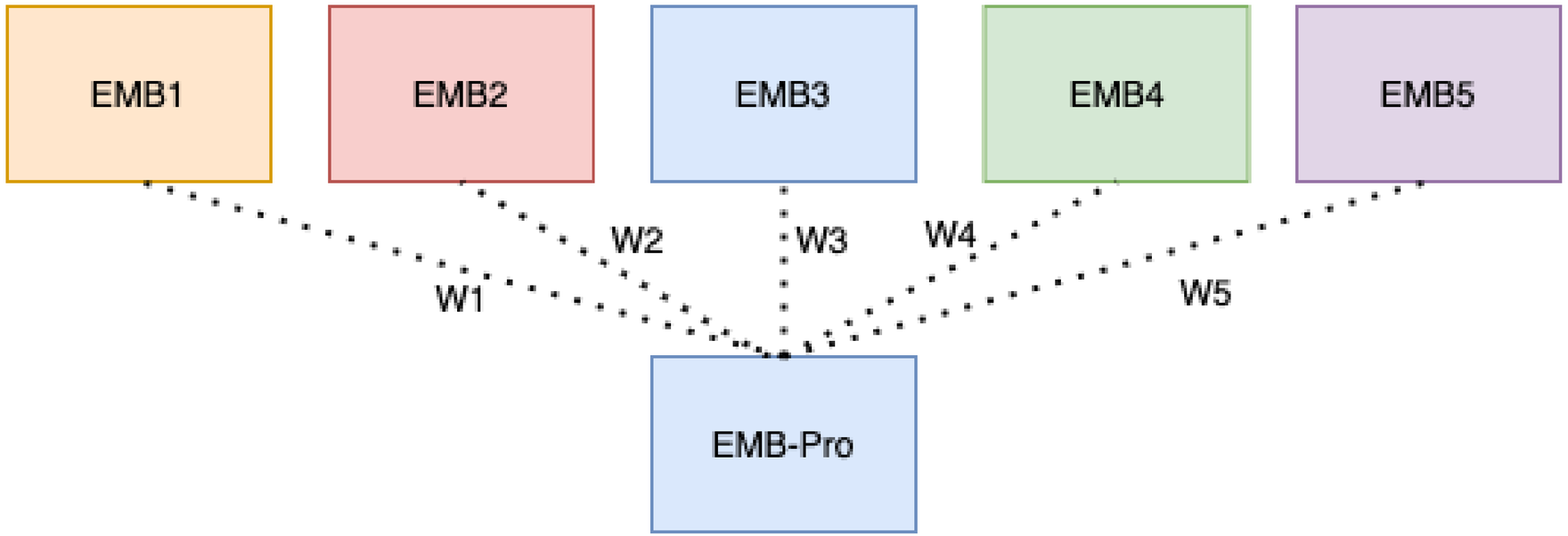
特征工程-Embedding

Time-> Entity embedding

提取的实体包括月份，当月第几周，是否是周末，星期，是否是中秋节，是否是国庆节等29类实体，用onehot的形式进行表征

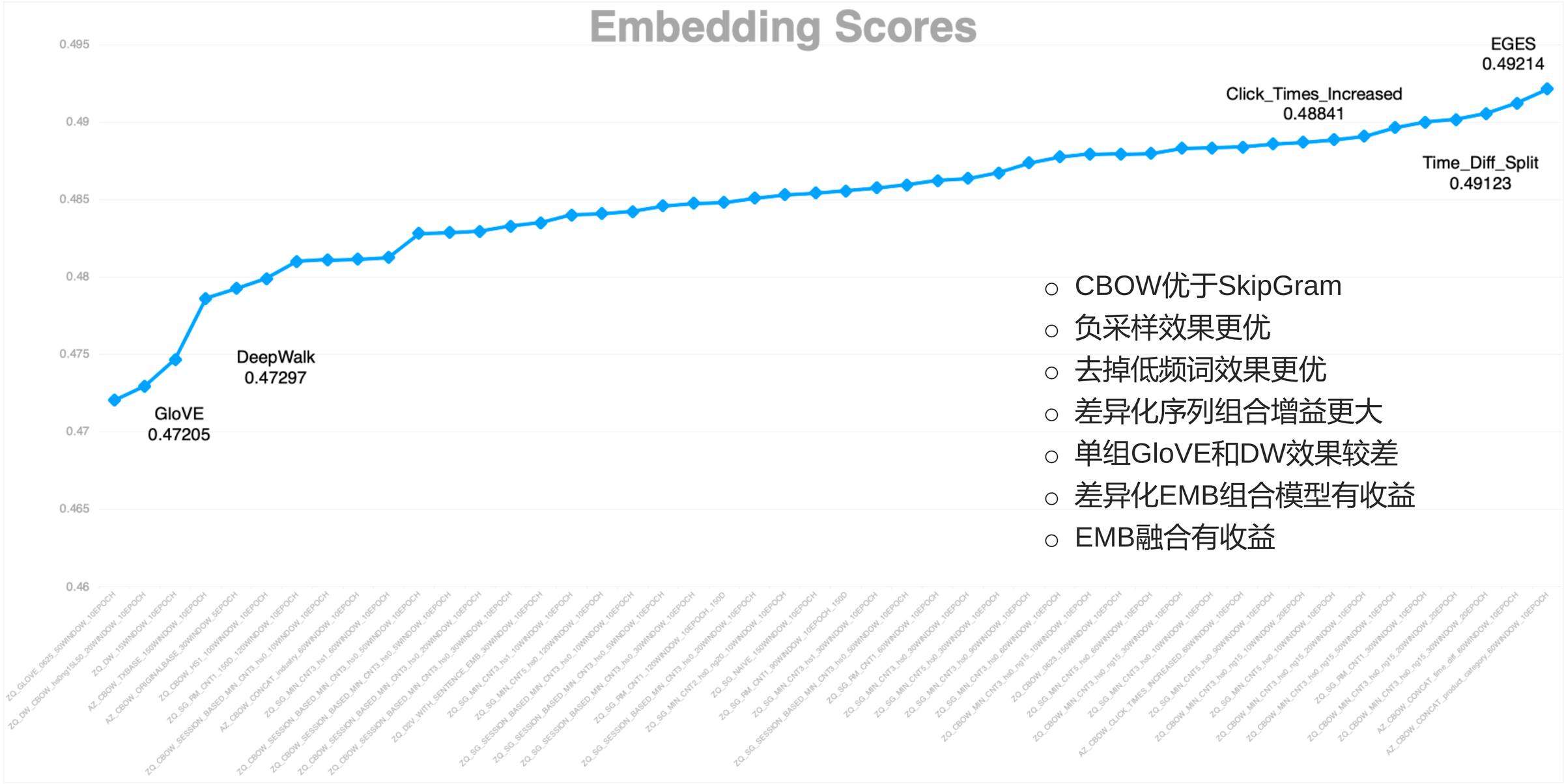
该部分作为输入后的embedding层，在训练中不进行权重更新

特征工程-Embedding



Embedding先进行一轮加权融合
Glove+W2V+DW，不同Window

特征工程-Embedding

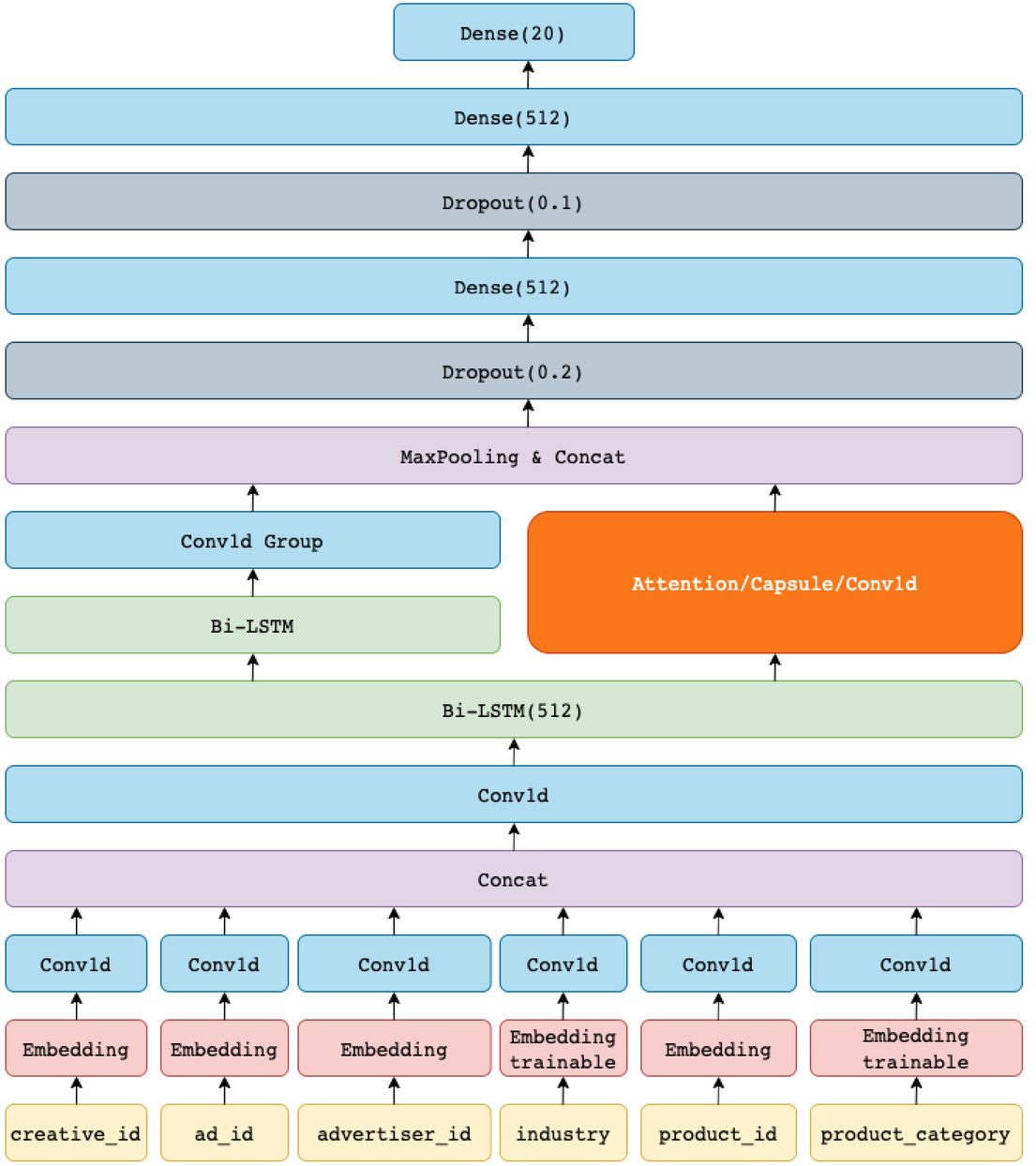


- CBOW优于SkipGram
- 负采样效果更优
- 去掉低频词效果更优
- 差异化序列组合增益更大
- 单组GloVE和DW效果较差
- 差异化EMB组合模型有收益
- EMB融合有收益

模型介绍

单模型/Ensemble

模型介绍-model1



- Bi-LSTM+Conv1d+Capsule+Attention

- Score: 1478~1.479

实验细节：

- 5 Folds Bagging
- 6输入 + 4~6组Embedding
- Batch Size: 512
- 优化器: AdamW
- LR Schedule+ReduceLROnPlateau

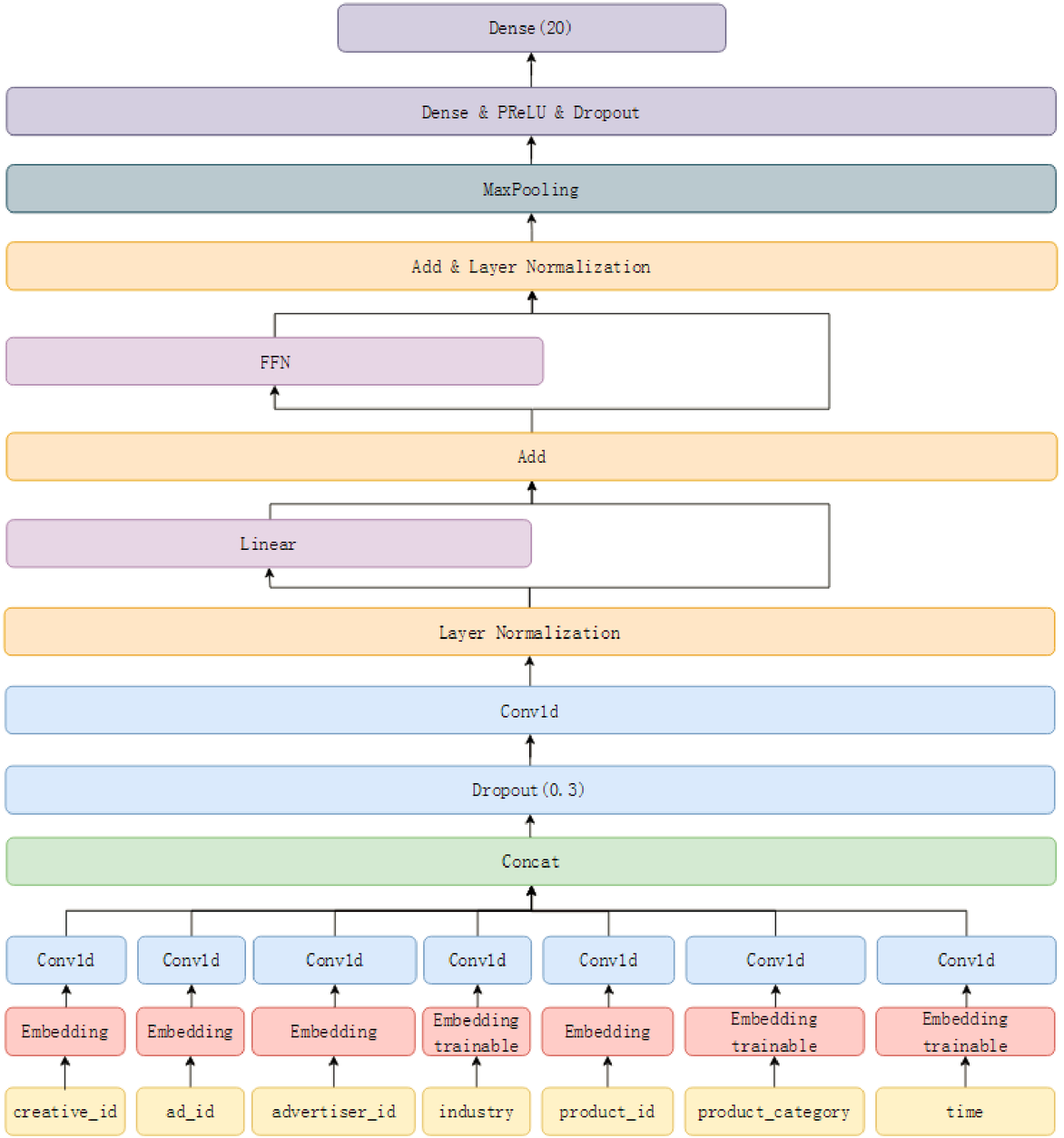
0~5　epoch 1e-3

6~12 epoch 5e-4

训练细节：

- 4*V100数据并行

模型介绍-model2



- DNN+Conv1d + Pre-LN
- Score: 1477~1.478

实验细节：

- 5 Folds Bagging
- 7 输入 + 6组Embedding
- Batch Size: 512
- 优化器: Adam
- LR Schedule

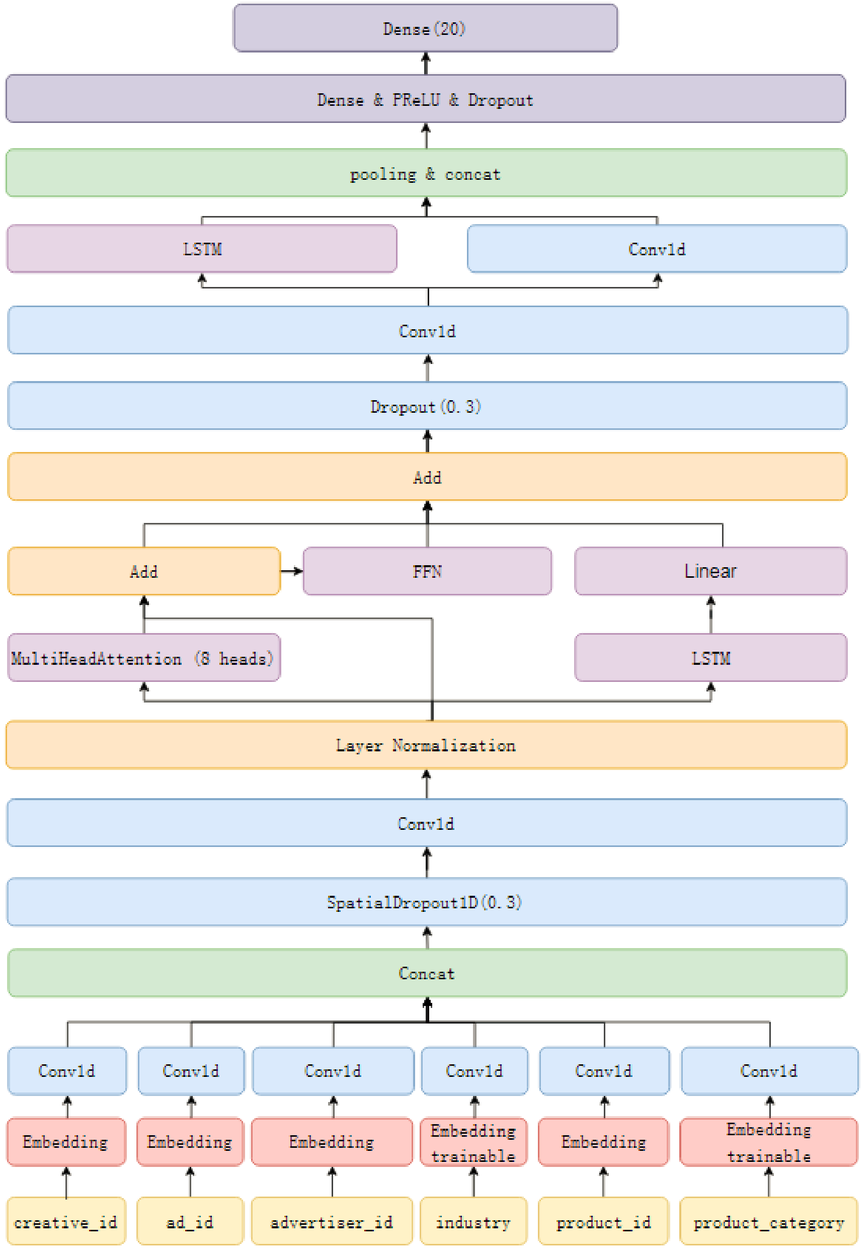
0~15 epoch 5e-4

16~22 epoch 1.25e-4

训练细节：

- 4*V100数据并行

模型介绍-model3



- Pre-LN+Transformer+LSTM+Conv1d
- Score:1477~1.478

实验细节：

- 5 Folds Bagging
- 6 输入 + 5组Embedding
- Batch Size: 512
- 优化器: Adam
- LR Schedule

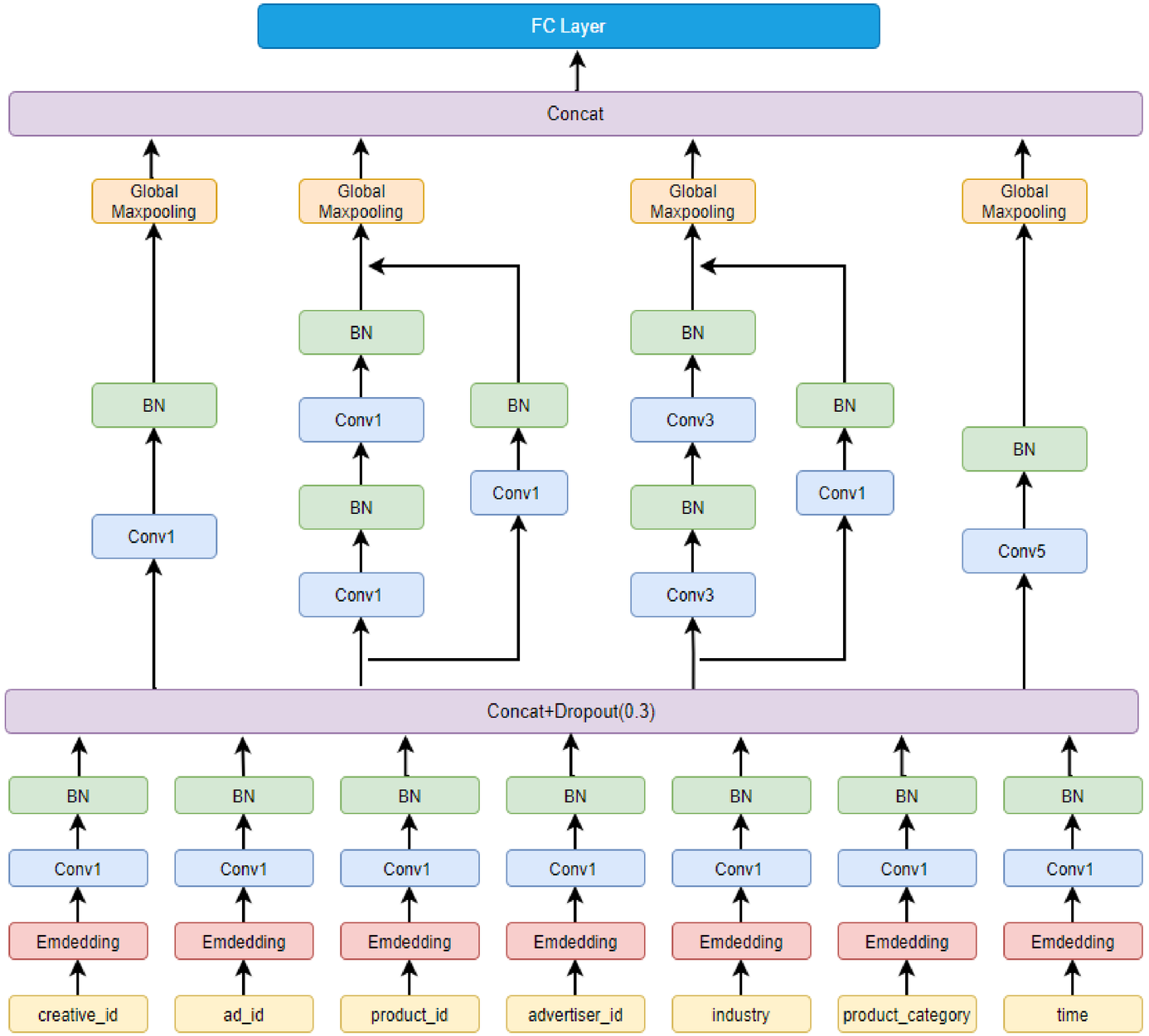
0~7　epoch 5e-4

8~13 epoch 1.25e-4

训练细节：

- 4*V100数据并行

模型介绍-model4



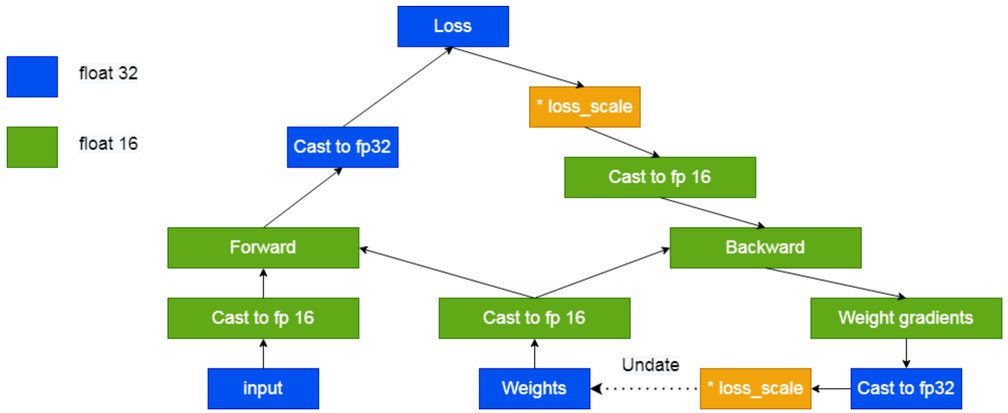
- Inception + Resnet
- Score: 1478~1.479

实验细节：

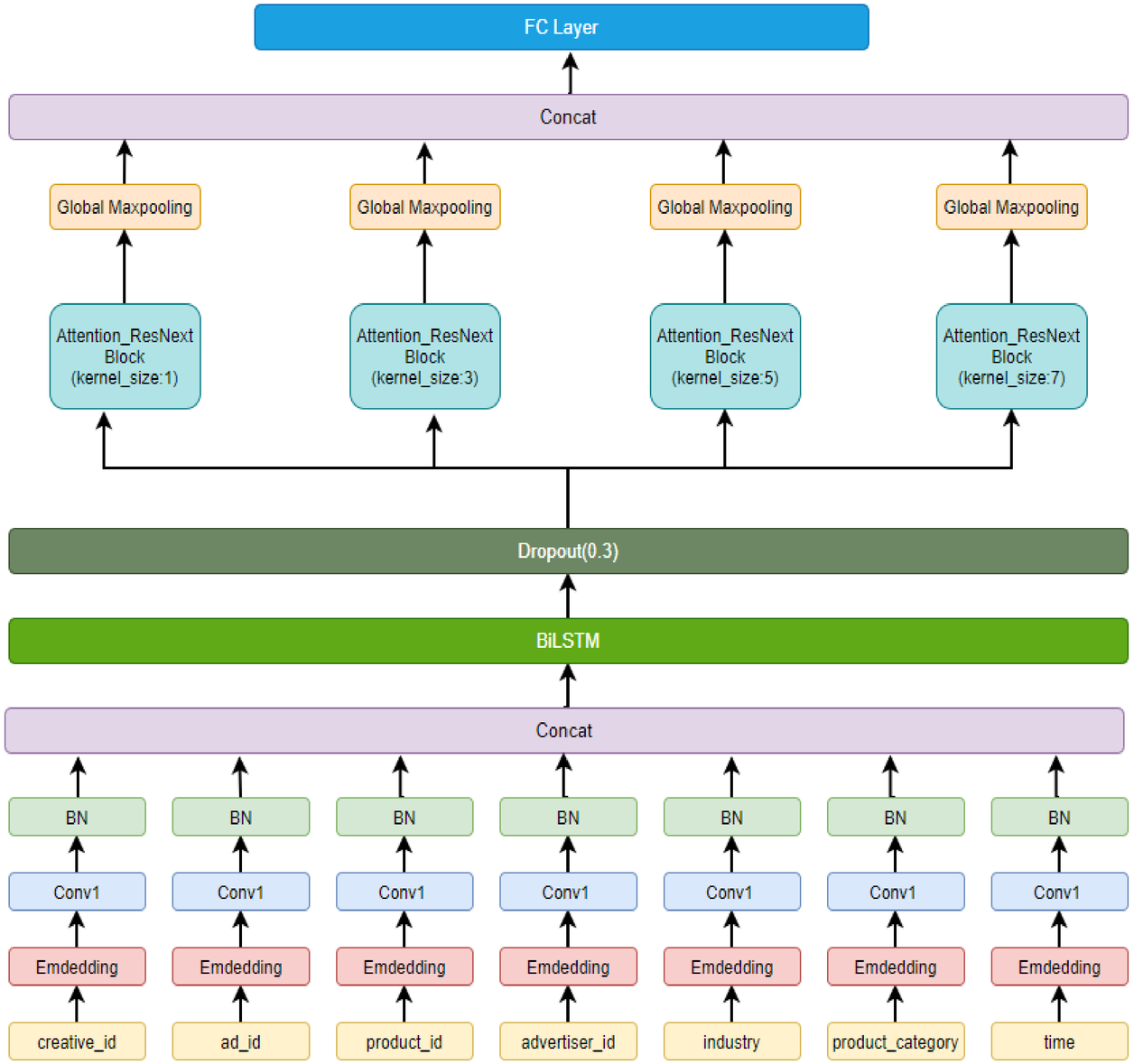
- 5 Folds Bagging
- 7 输入 + 5组Embedding
- Batch Size: 512
- 优化器: AdamW
- LR Scheduler: ReduceLROnPlateauLR
- Label_smoothing 0.03

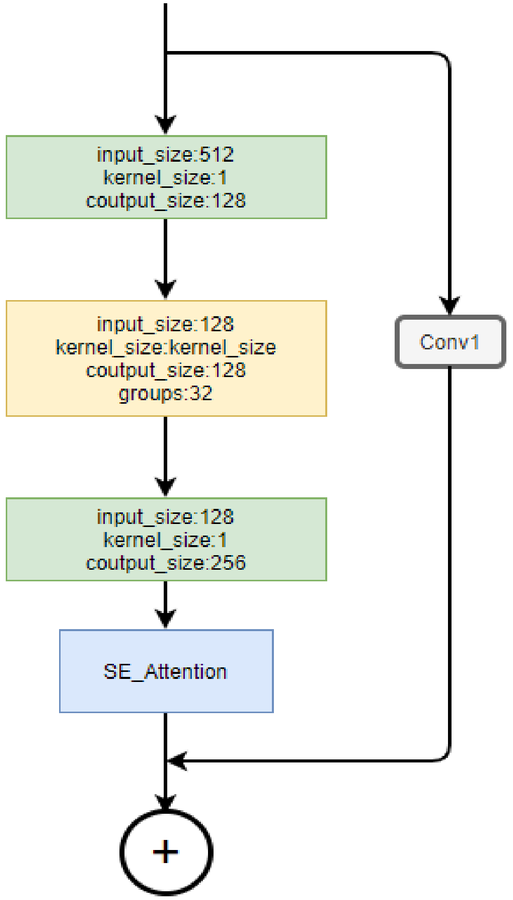
训练细节：

- 4*V100数据并行
- 混合精度



模型介绍-model5





BiLSTM + ResNeXt + Attention

Score: 1478~1.479

实验细节：

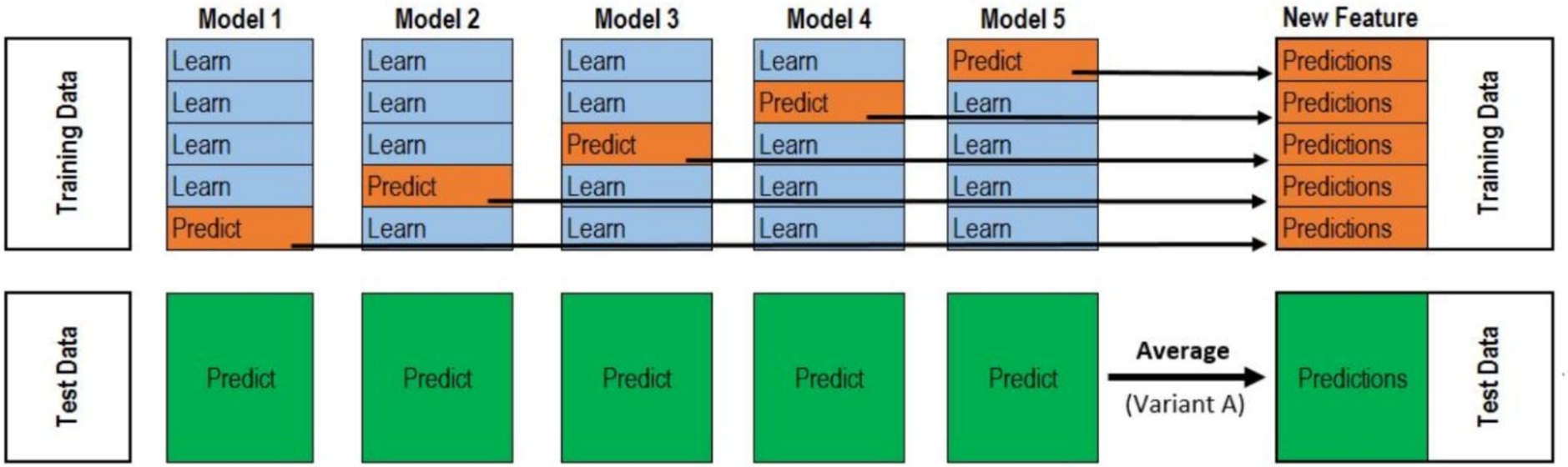
- 5 Folds Bagging
- 7输入+5组Embedding
- Batch Size: 512
- ResNeXt Groups: 32
- 优化器: AdamW
- ReduceLROnPlateauLR
- Label_smoothing 0.03

训练细节：

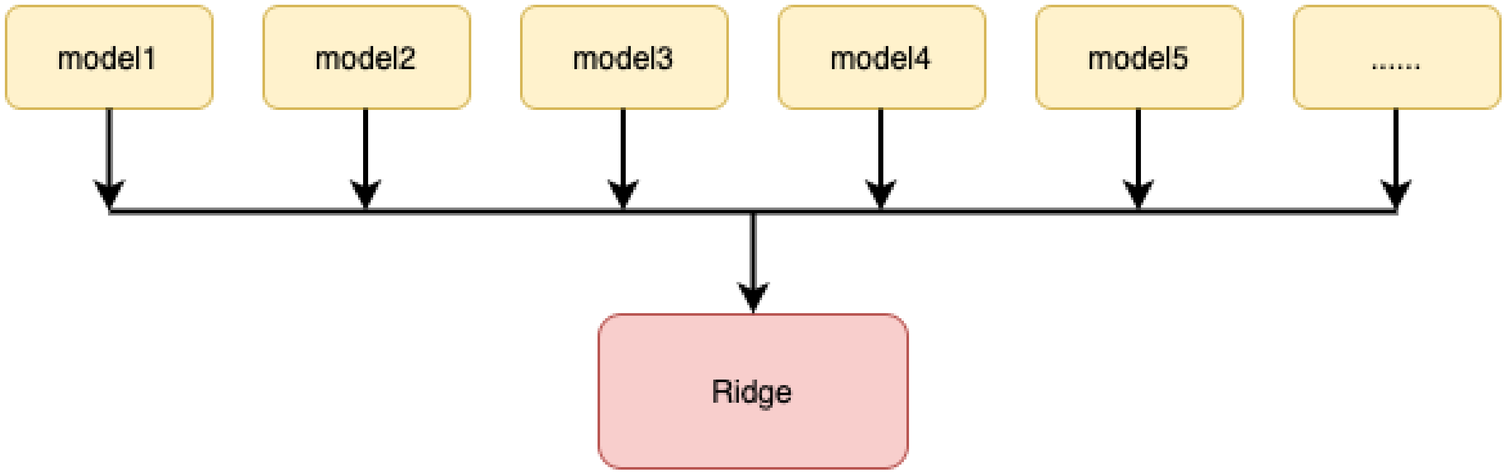
- 4*V100数据并行
- 混合精度

https://arxiv.org/abs/1611.05431

模型介绍- Ensemble



Stacking



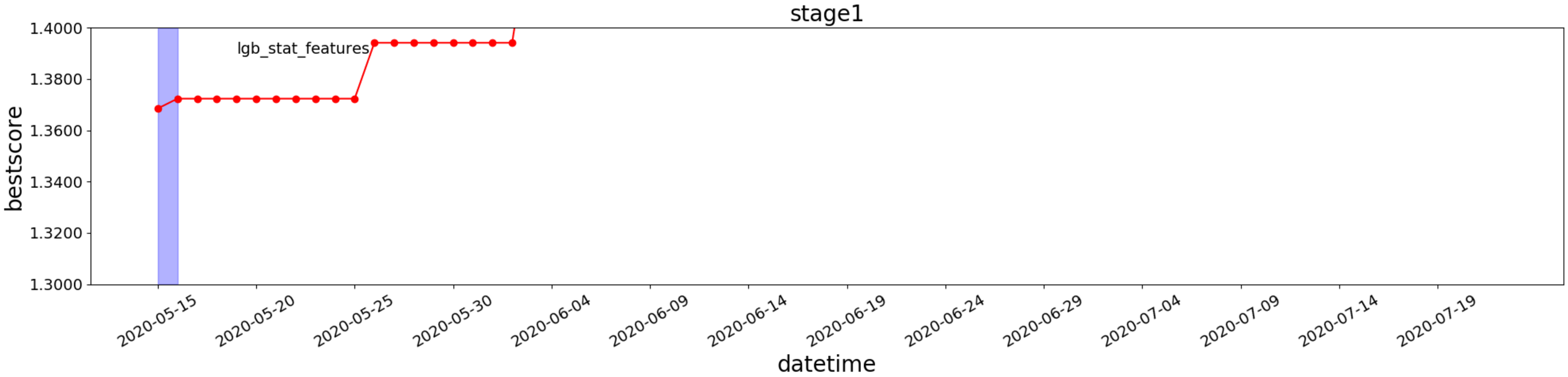
Ridge Classifier as Stacker
Final Score: 1.482612

总结与思考

上分之路/总结与思考

总结与思考- 上分之路

团队总计有效提交次数132次



几次主要的提升

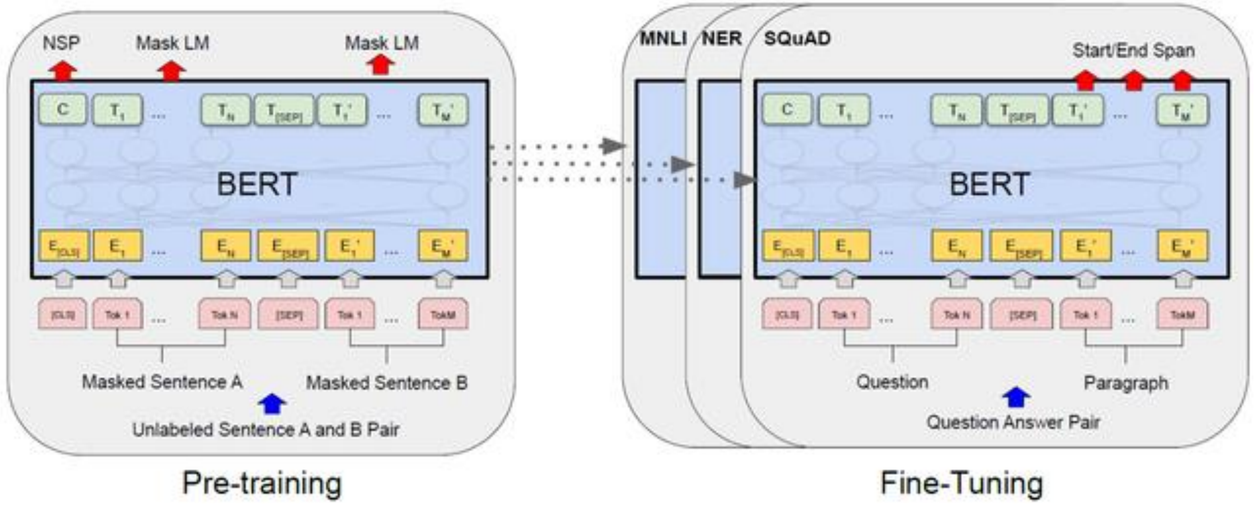
- 采用预训练+BILSTM建模（Top 20）
- Concat不同意义的预训练特征，并采用特征抽取结构进行降维提取（3-4K Top10）
- 数据增强（3-4K Top 5）

其他提升

- 联合概率转为边缘概率进行融合提交0.5K
- 模型融合4K

尝试未果

1、NN中加入统计特征/TF-IDF/User Embedding
2、后处理 argmax(w*logits)
3、ELMO/YouTube-Net Embedding
4、伪标签 PU-Learning
5、 数据增强：序列TOP-K近邻做随机替换
6、Embedding: Freeze => Unfreeze => Finetune
7、Lookahead、CyclicLR、Cosine LR Decay等
8、CTR类模型

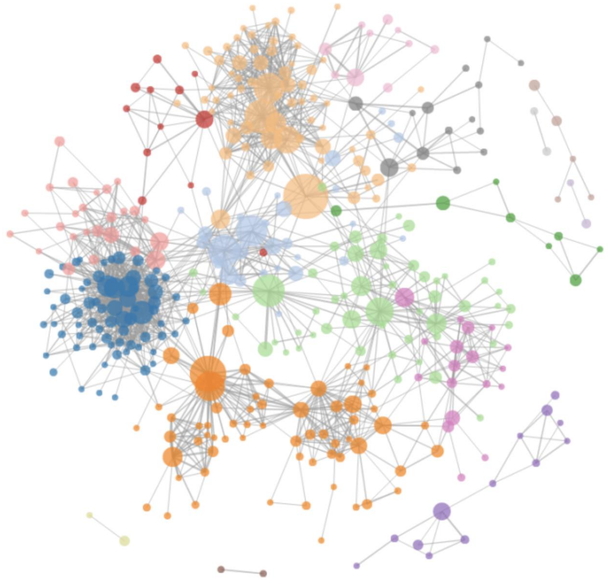


反思

1、Exploration可以更多， Exploitation可以少些
2、把握好节奏，循序渐进

未来展望

1、Bert
2、GNN



THANKS