天才报大腿


TAAC
腾讯广告算法大赛
Tencent Advertising Algorithm Competition

Contents

目录

团队介绍

团队介绍

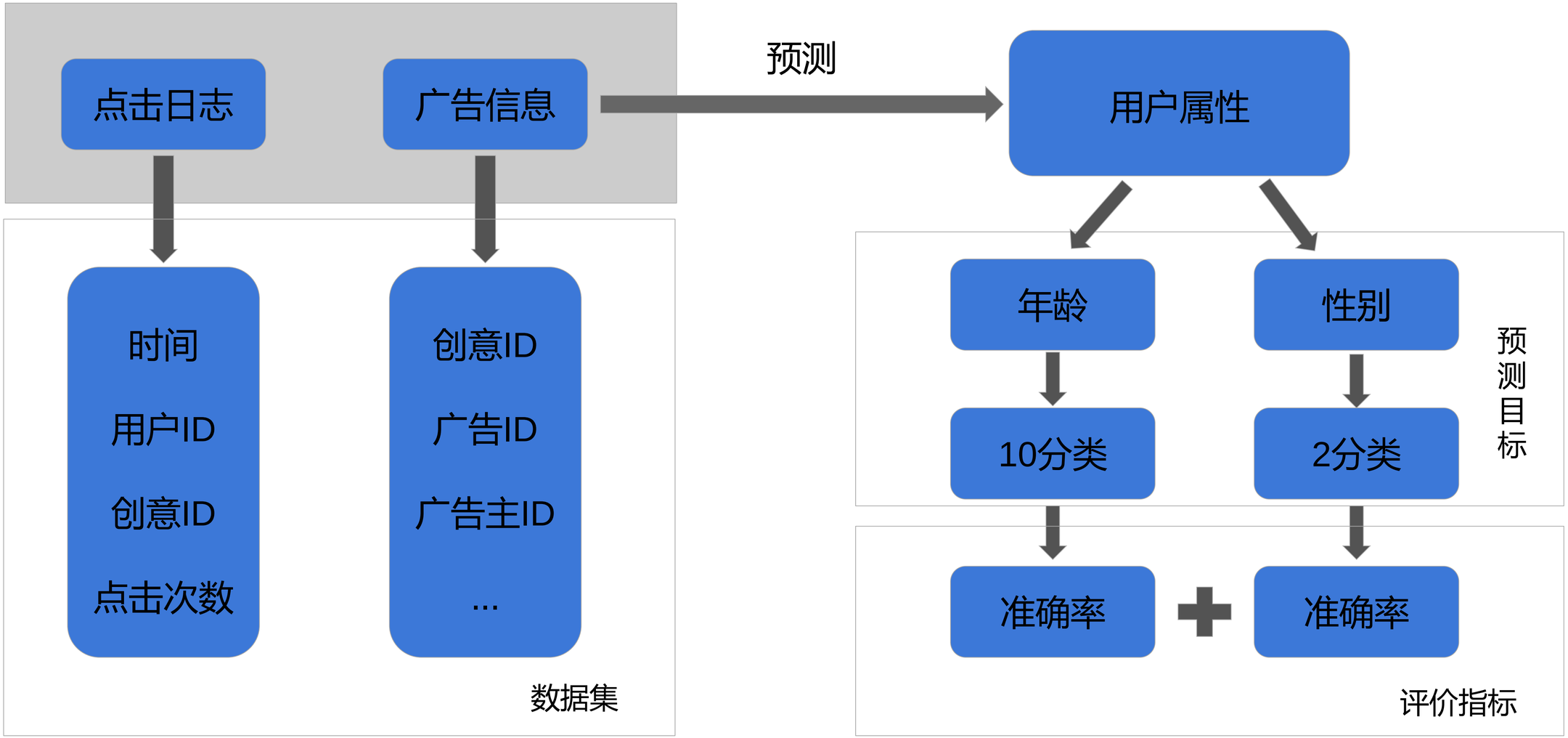
陶超
华中科技大学
统计学硕士
算法工程师

程唯
南京大学
计算机硕士
算法工程师

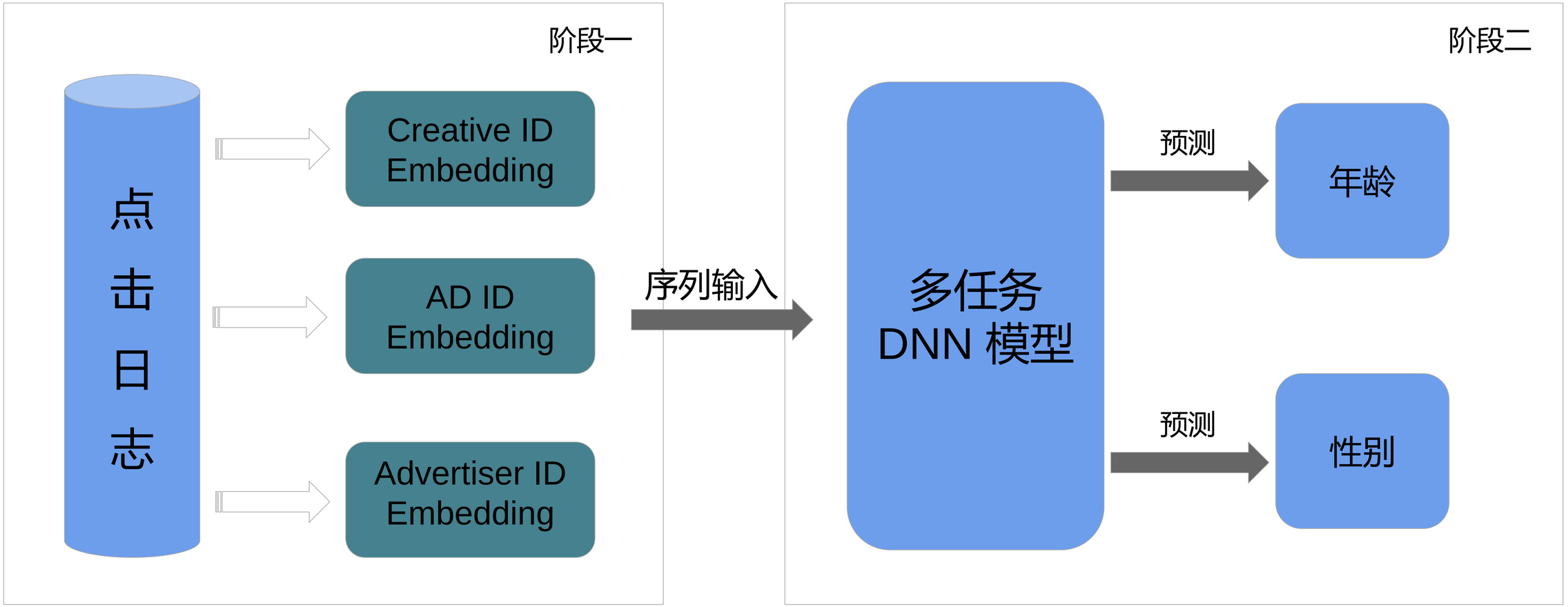
罗欣晨
四川大学
计算机本科
算法工程师

赛题理解

赛题理解


点击日志
广告信息
预测
用户属性
时间
用户ID
创意ID
点击次数
创意ID
广告ID
广告主ID
...
数据集
年龄
性别
预测目标
10分类
2分类
准确率
准确率
评价指标

两阶段建模

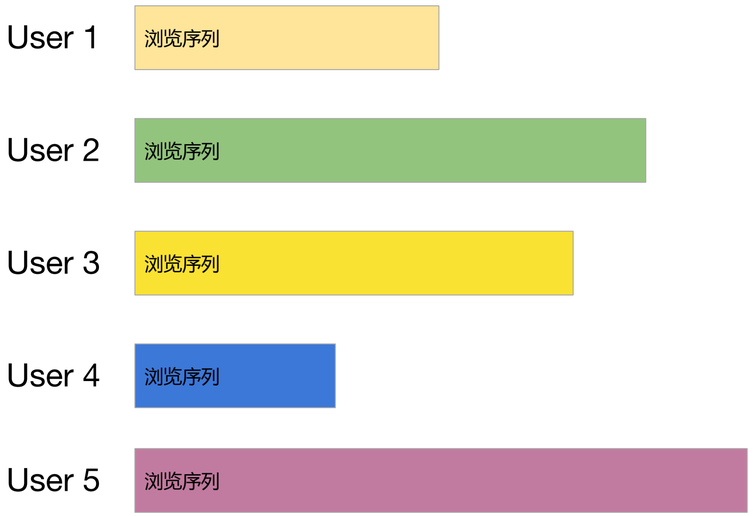

阶段一
点击日志
Creative ID Embedding
AD ID Embedding
Advertiser ID Embedding
序列输入
阶段二
多任务 DNN 模型
预测
年龄
预测
性别

特征工程

Embedding构建/数据增广

Embedding构建

通过点击日志构造每个用户的浏览序列

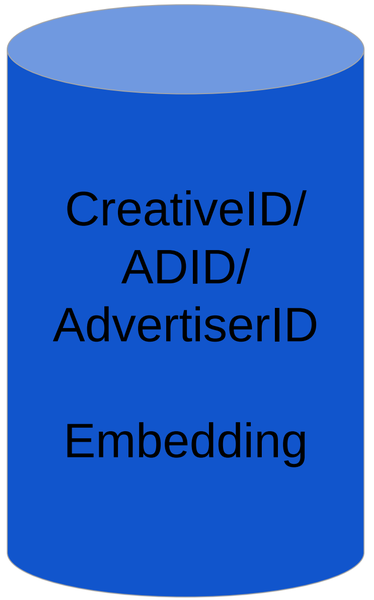


支持对数据在线shuffle

Word2Vec
CBOW

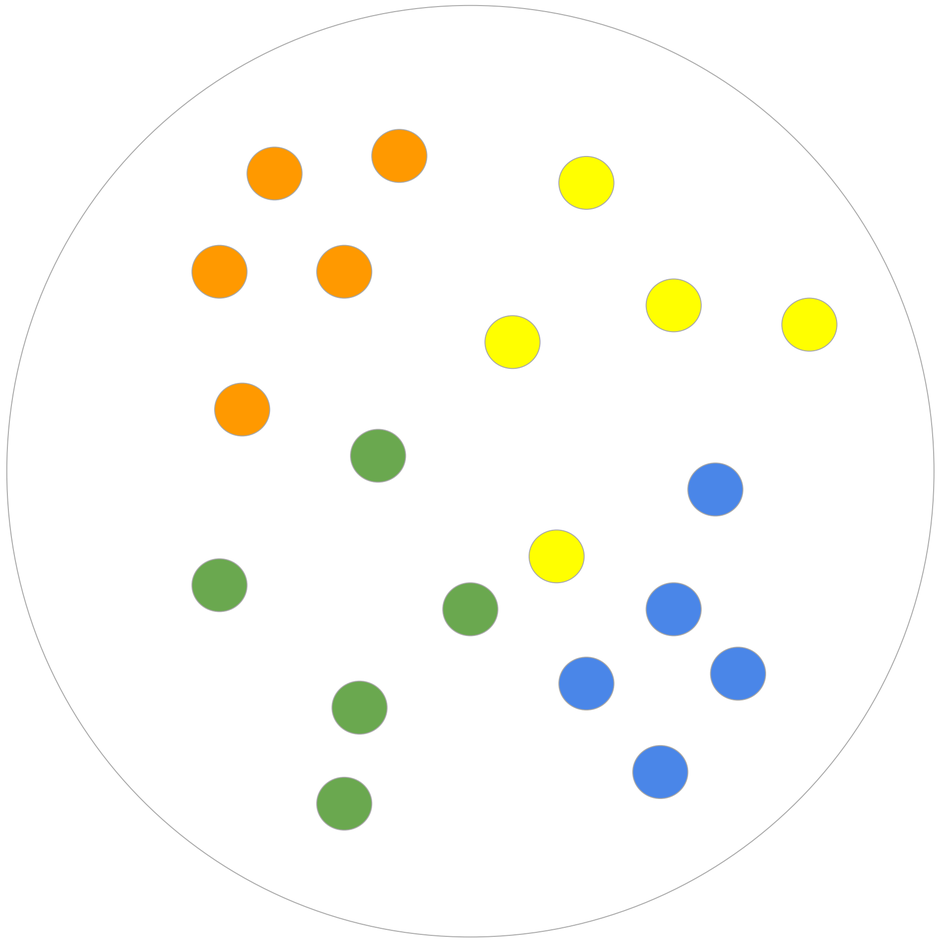
Word2Vec
Skip-Gram

GloVe

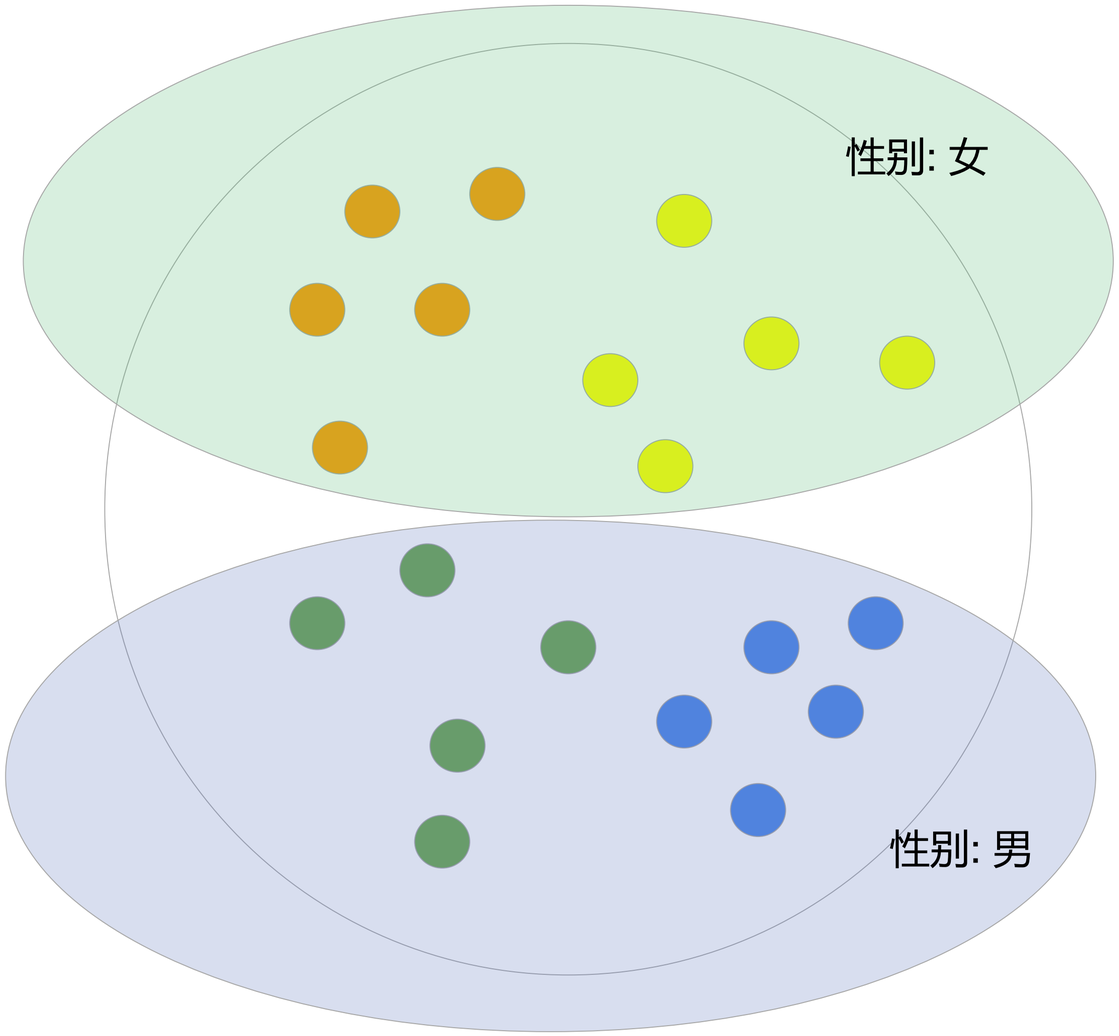


Embedding构建-混合标签

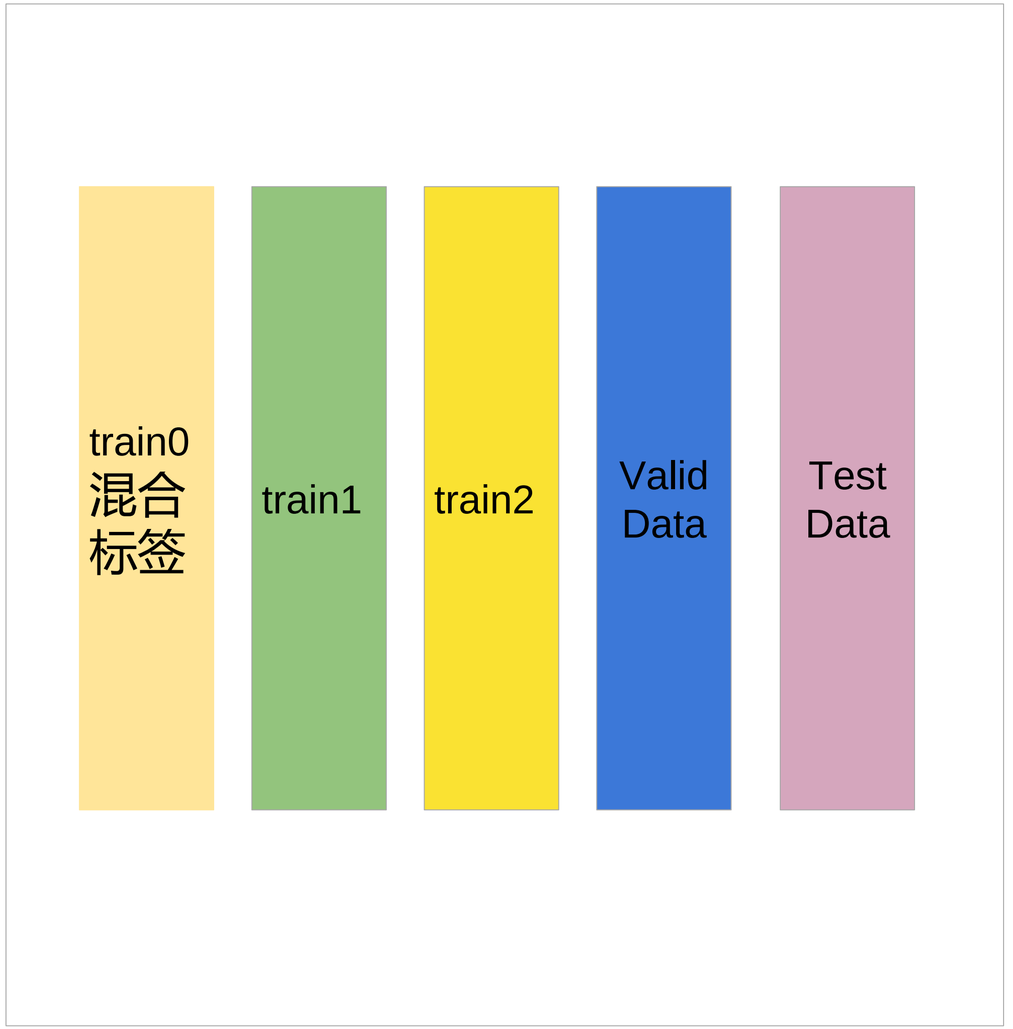
通过Word2Vec/GloVe 可以让相似的广告 聚类在一起，包含一定的语义信息



我们希望预训练的Embedding表达语义对下游分类任务更友好



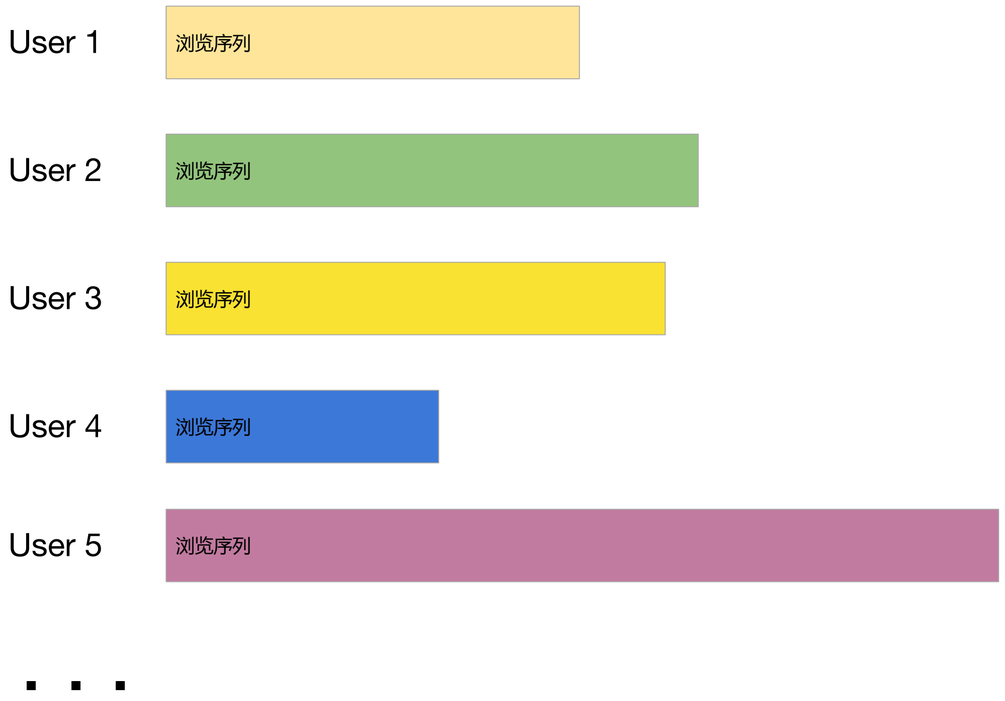
Embedding构建-混合标签



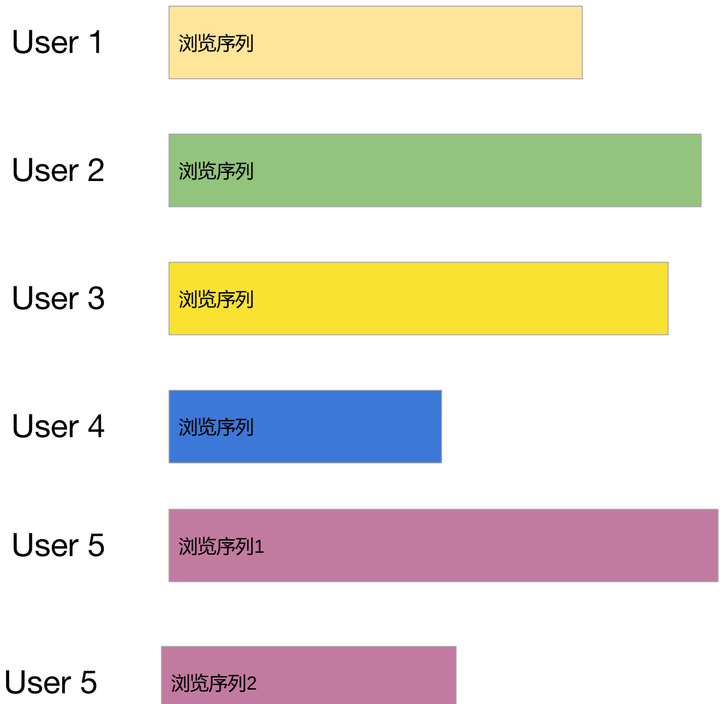
数据增广

拆分超长浏览序列

通过点击日志构造每个用户的浏览序列

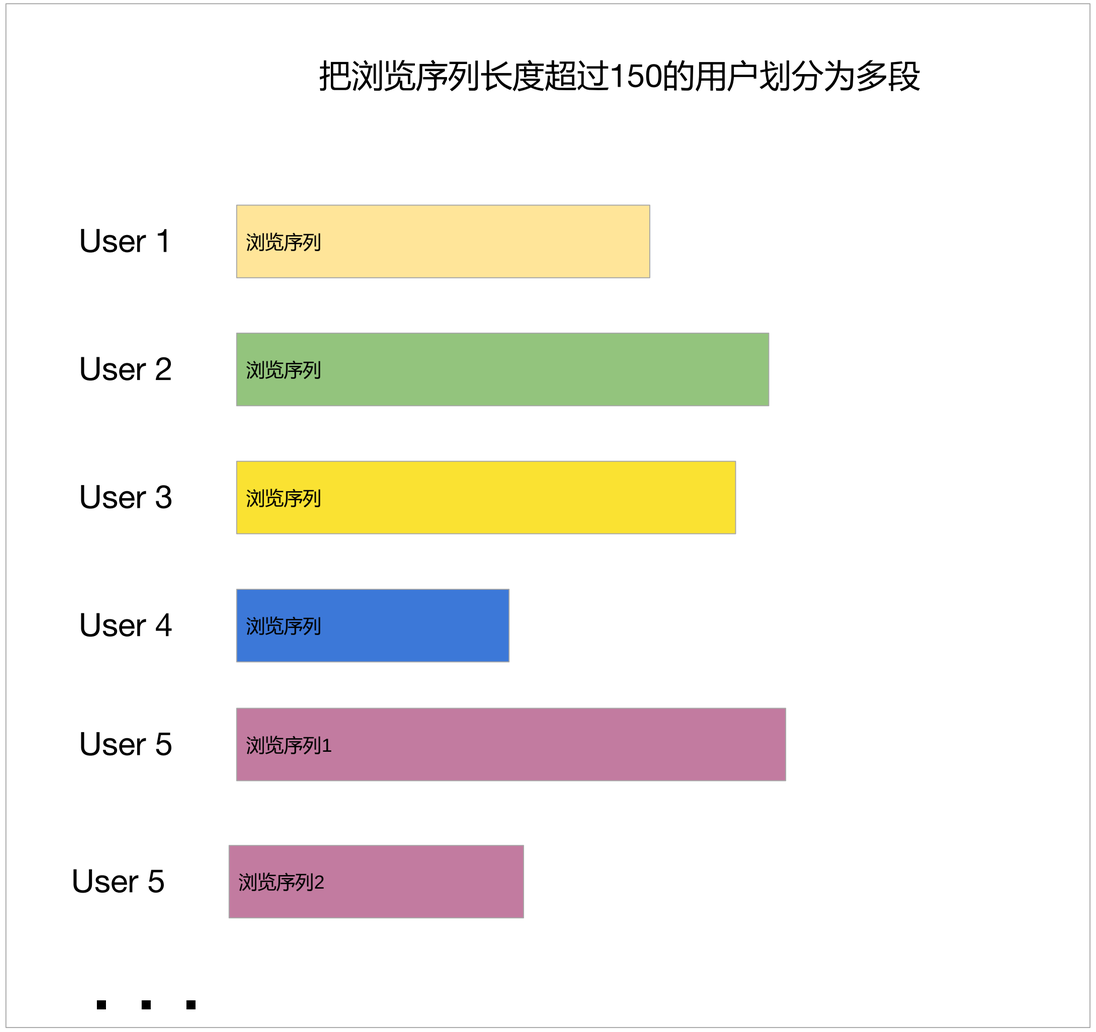


把浏览序列长度超过150的用户划分为多段



数据增广

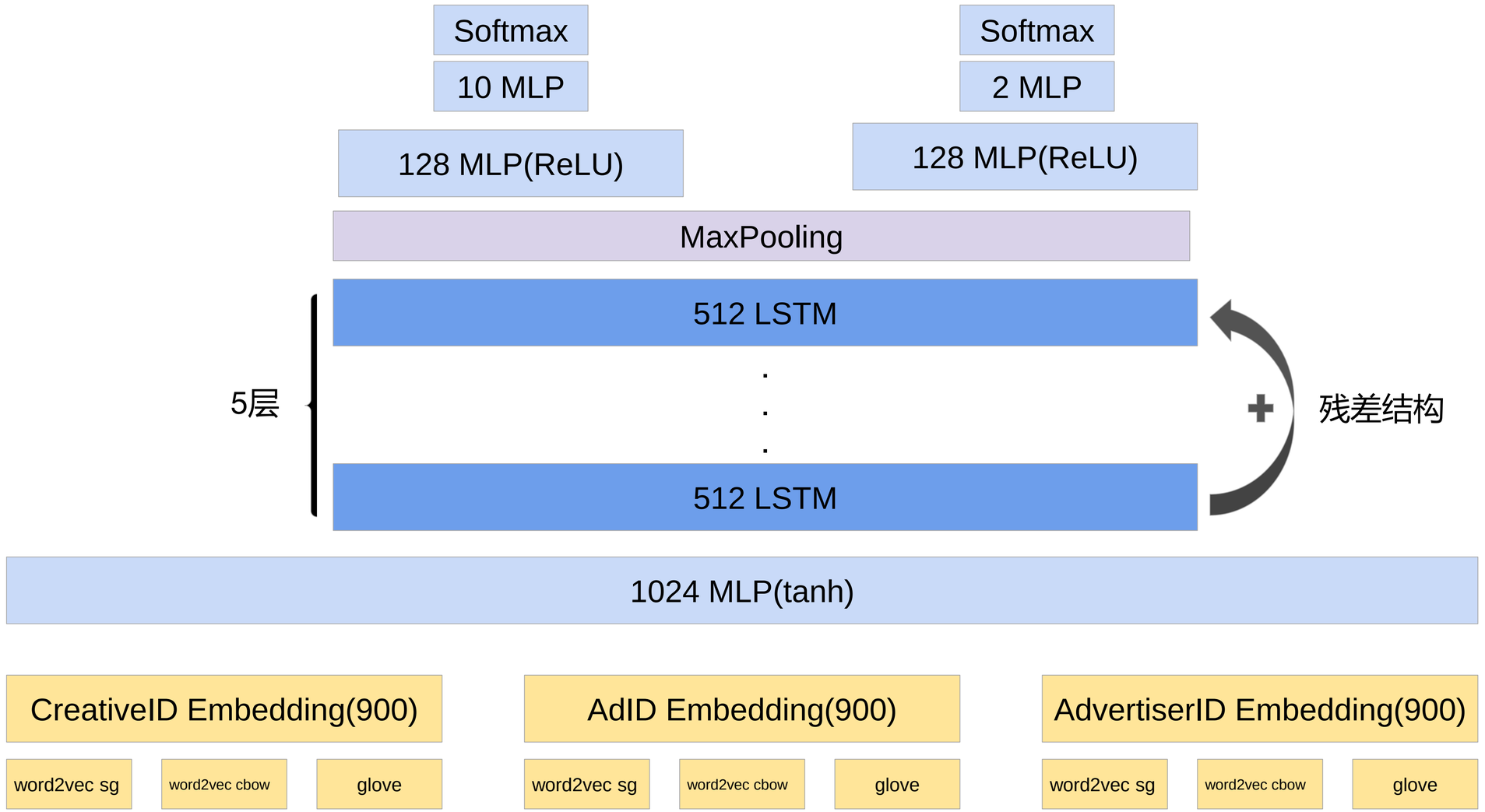
在线数据增广：shuffle 并截断数据



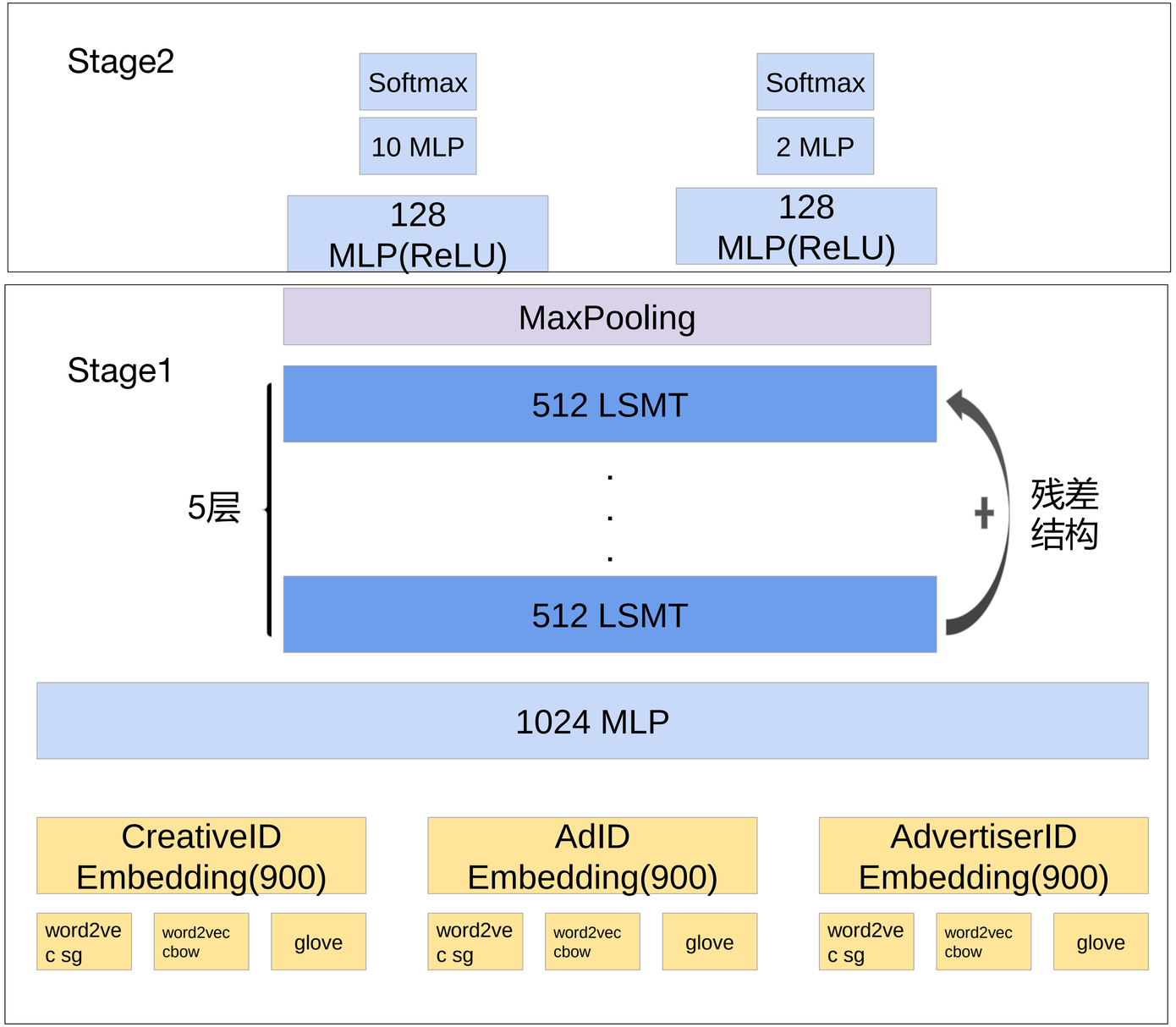
模型介绍

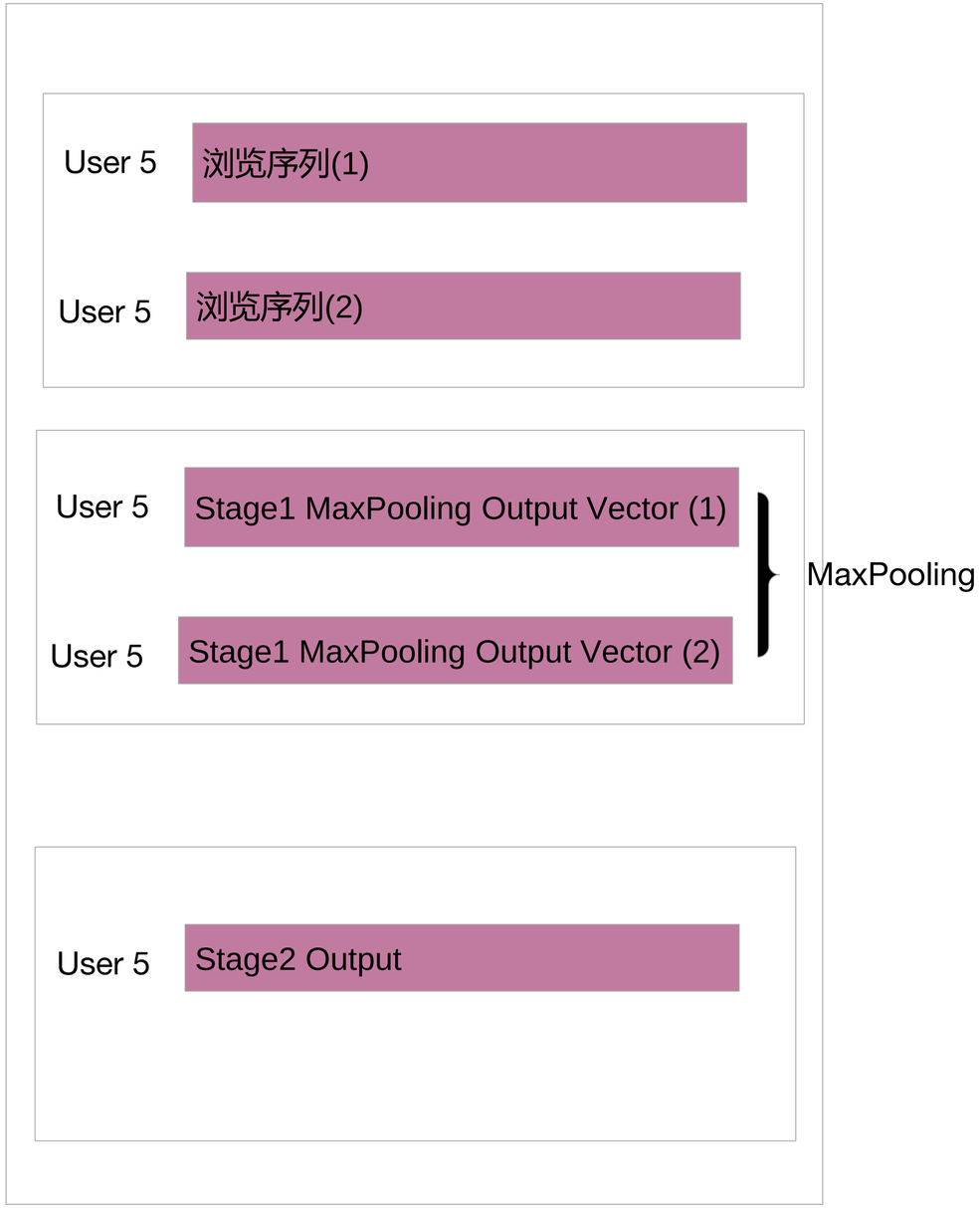
模型结构/模型训练/模型预测

模型结构 - 既要宽更要深


Softmax
10 MLP
Softmax
2 MLP
128 MLP(ReLU)
128 MLP(ReLU)
MaxPooling
512 LSTM
5层
残差结构
512 LSTM
1024 MLP(tanh)
CreativeID Embedding(900)
AdID Embedding(900)
AdvertiserID Embedding(900)
word2vec sg
word2vec cbow
glove
word2vec sg
word2vec cbow
glove
word2vec sg
word2vec cbow
glove

模型结构 - 分段处理超长序列





模型训练 - 天下武功唯快不破

总共400万训练数据，200S 一个Epoch， 模型收敛到(线下age 0.52, gender 0.95) 需要30个epoch， 共1.7个小时

1. 高效的数据输入，TFRecord + tf.lookup.StaticHashTable Join 广告特征

2. 训练数据随机shuffle并截取前32条

对比150序列长度，无精度损失并提速4X

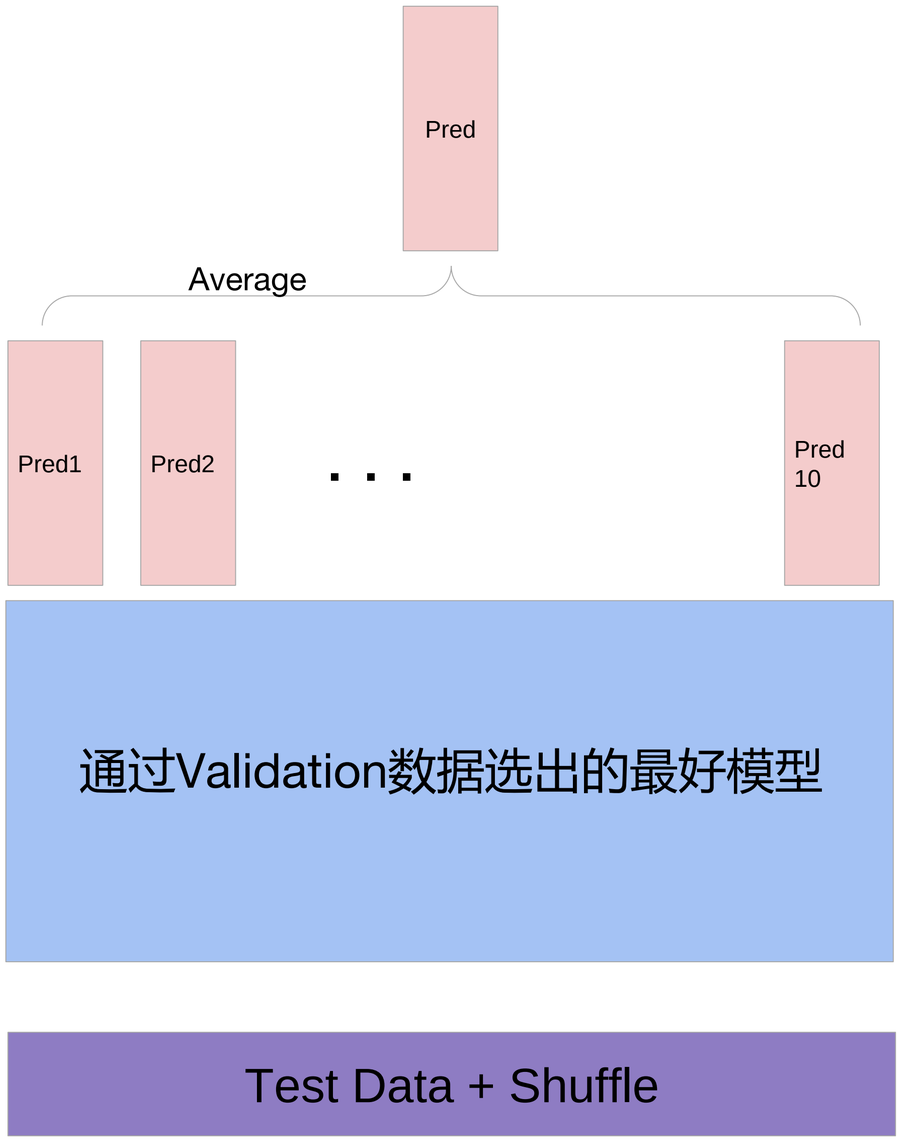
3. 过滤出现次数少于3次的ID，使用FP16存储Embedding，使得模型可以放进显存

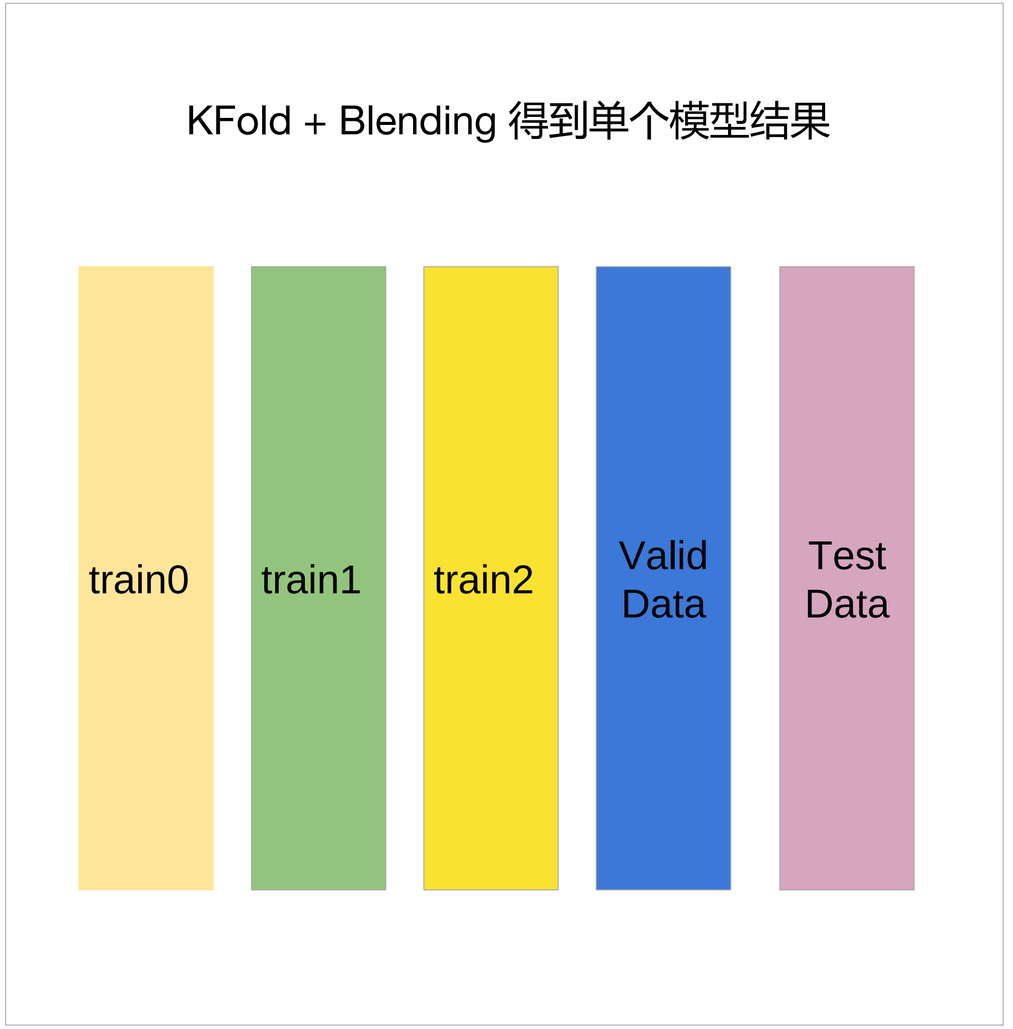
对比Embedding放在CPU，提速1.3X

4. 开启混合精度训练，使用TensorCore的能力

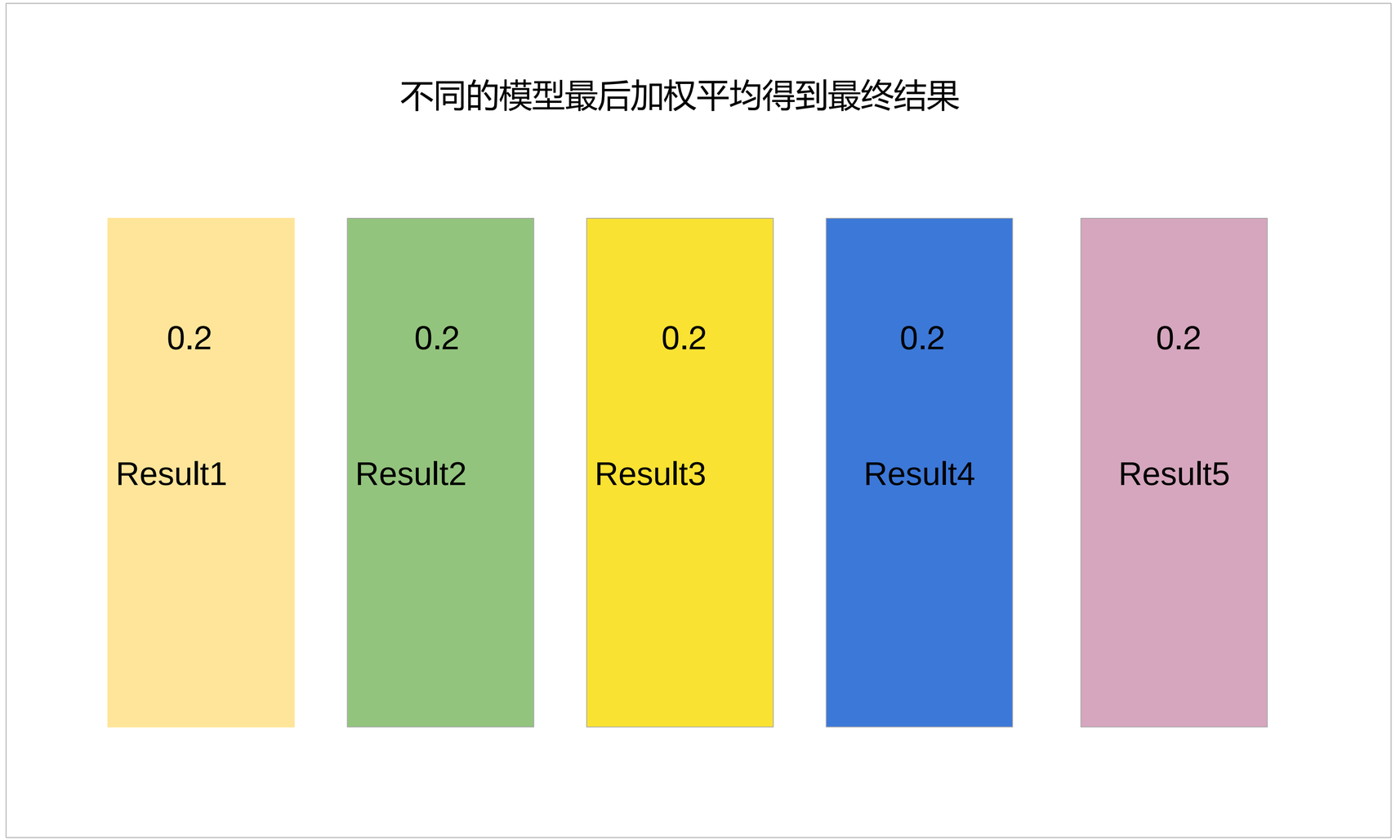
对比FP32训练，提速2.5X

模型预测


Pred
Average
Pred1
Pred2
. . .
Pred
10
通过Validation数据选出的最好模型
Test Data + Shuffle

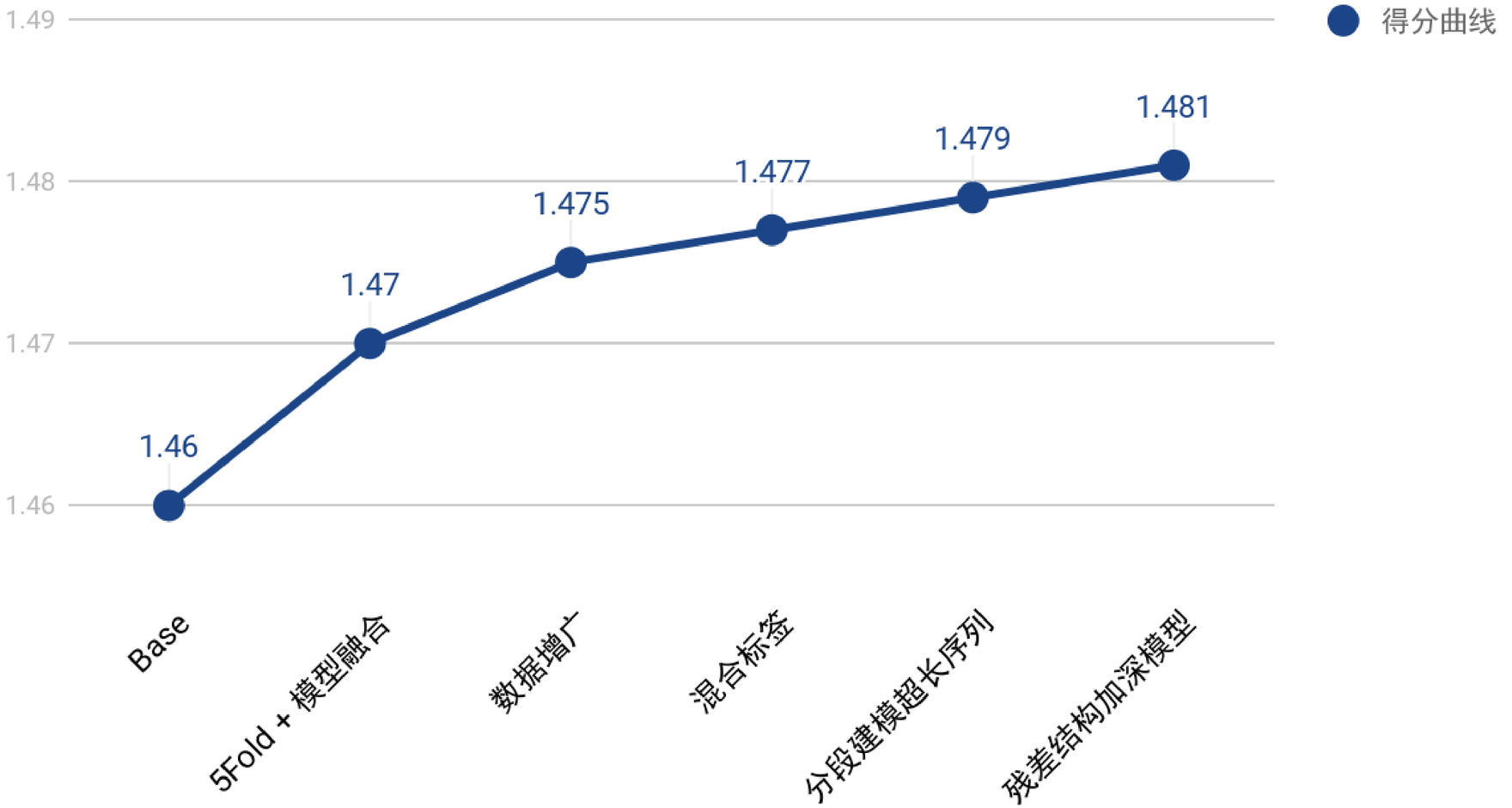

KFold + Blending 得到单个模型结果
train0
train1
train2
Valid
Data
Test
Data

模型融合


不同的模型最后加权平均得到最终结果
0.2
Result1
0.2
Result2
0.2
Result3
0.2
Result4
0.2
Result5

总结与思考

模型得分


1.49
1.48
1.47
1.46
得分曲线
1.46
1.47
1.475
1.477
1.479
1.481
Base
5Fold + 模型融合
数据增广
混合标签
分段建模超长序列
残差结构加深模型

总结与思考

主要创新

1. 提出了混合标签构建Word2Vec训练数据，训练更适合下游任务的Embedding

1. 提出了分阶段建模解决数据中的超长序列，训练数据随机采样到32条，在显著提高训练速度的同时不丢失模型精度

问题思考

1. 如何通过CTR日志得到优秀的Embedding表达是解题的关键，除了Word2Vec/GloVe等方式外使用更为复杂的FM/Wide&Deep等CTR模型构建AD和用户的Embedding表达或许可以进一步的提高结果。

1. 我们一直没有利用到时序特征，如何利用上时序特征是另一个可能的得分点

THANK S