小太阳2020战队

Contents

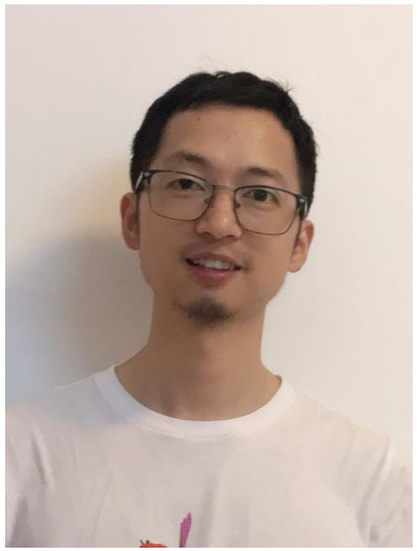
目录

团队介绍

团队介绍

队长

- 数据科学爱好者
- Kaggle Master
- 北京交通大学硕士



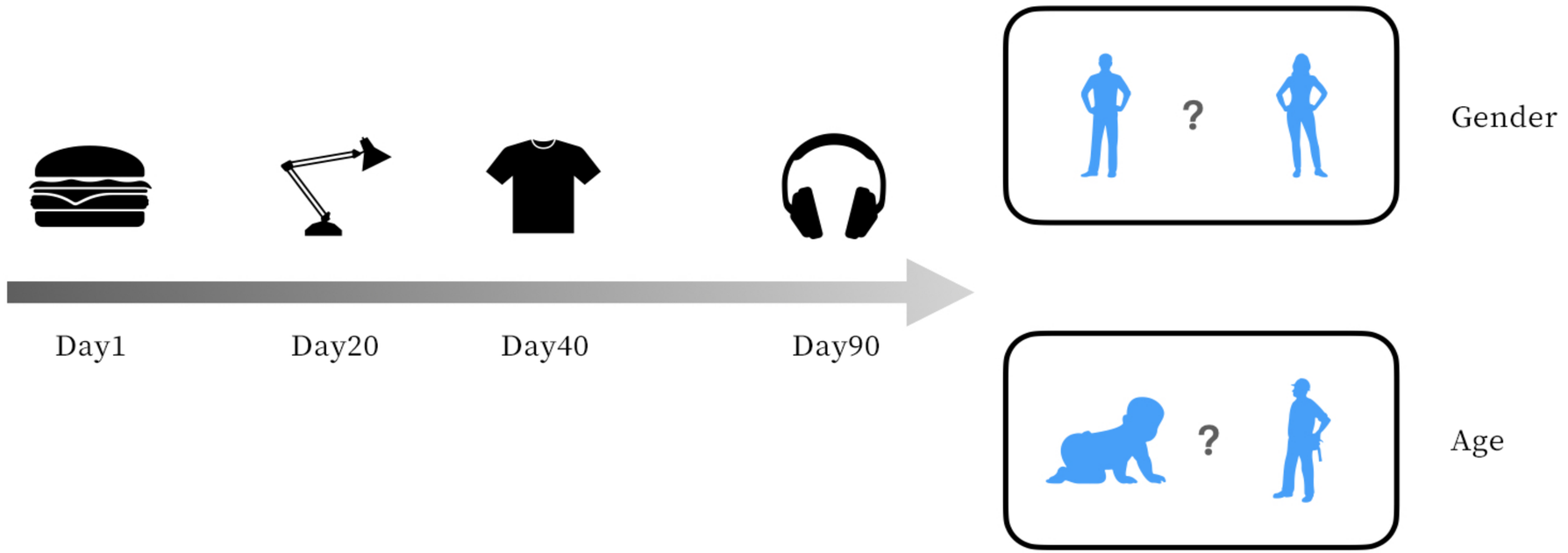
队员

- 数据科学爱好者
- Kaggle学习者
- Kaggle比赛中获5块银牌
- KDD2020中Top20

02

赛题理解

赛题理解



使用广告点击数据，逆向预测用户性别和年龄段

解题思路

- **LightGBM+特征工程：**target encode，tfidf，计数特征，w2v向量特征等。
- **深度学习-序列建模：**将用户点击过的广告按时间顺序组成序列，利用LSTM、Transformer等序列建模方法挖掘序列信息来推测用户属性。

特征工程

特征工程

- **Target encode：**首先统计每个广告的用户群体的target encode(对于age，我们将其onehot以后的十维属性分别进行统计)，然后对用户下所有广告(忽略低频广告)的target encode求mean作为用户的target encode特征。
- **Tfidf：**将用户点击过的广告组成序列计算tfidf特征。
- **W2V特征：**将预训练词向量作为特征。
- **计数特征：**统计用户点击过的广告数量，广告nunique值，统计用户对高频低频广告的偏好程度等等。

04

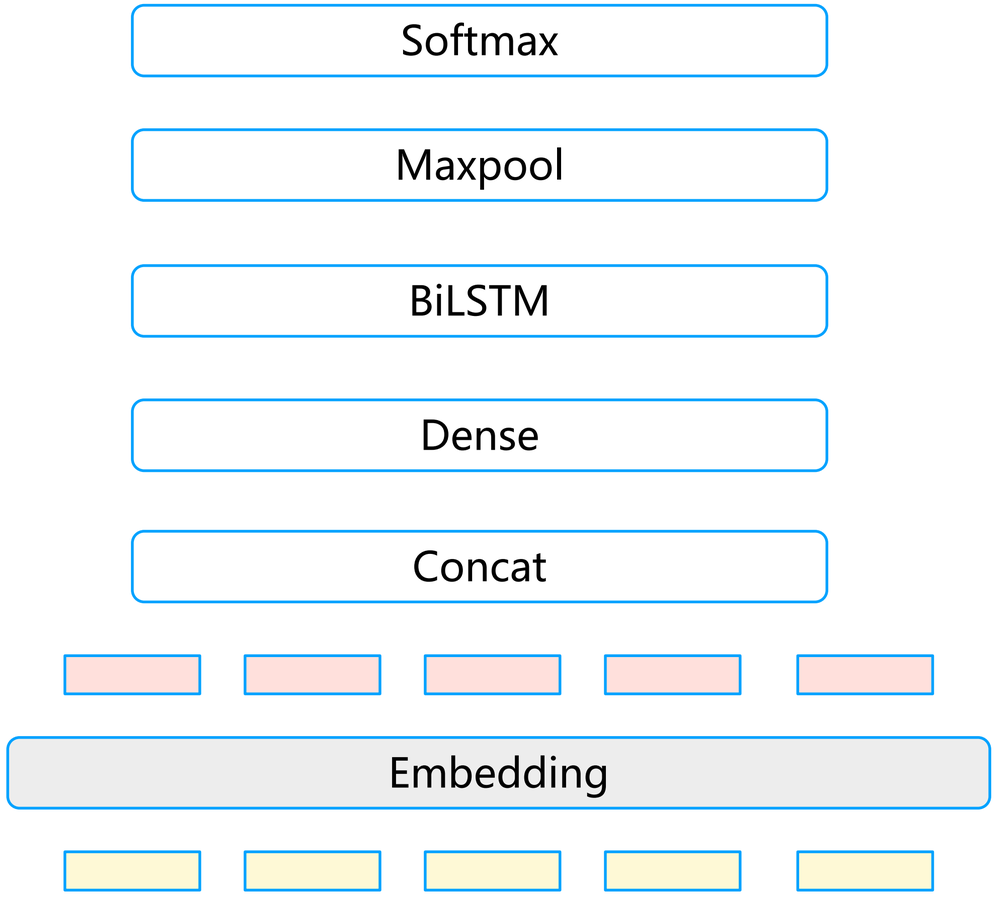
模型介绍

时间步上的Dropout

- 用户对广告的点击行为是一种随机性比较强的行为，可能因为不小心会点击自己不感兴趣的广告，对自己感兴趣的广告也不一定每次都点击。
- 基于这样的前提，我们假设，两个广告点击序列，如果只是少数广告不同，那么我们可以认为这两个广告点击序列背后的用户属性是相同的。
- 所以，我们可以对数据集中用户的点击序列进行小比例(0.2)的随机丢弃来进行数据增强。
- 在深度学习模型的训练过程中，在Embedding层后，我们通过在时间步上对序列做Dropout来实现数据增强，大大提高模型的泛化能力(提升3k）。

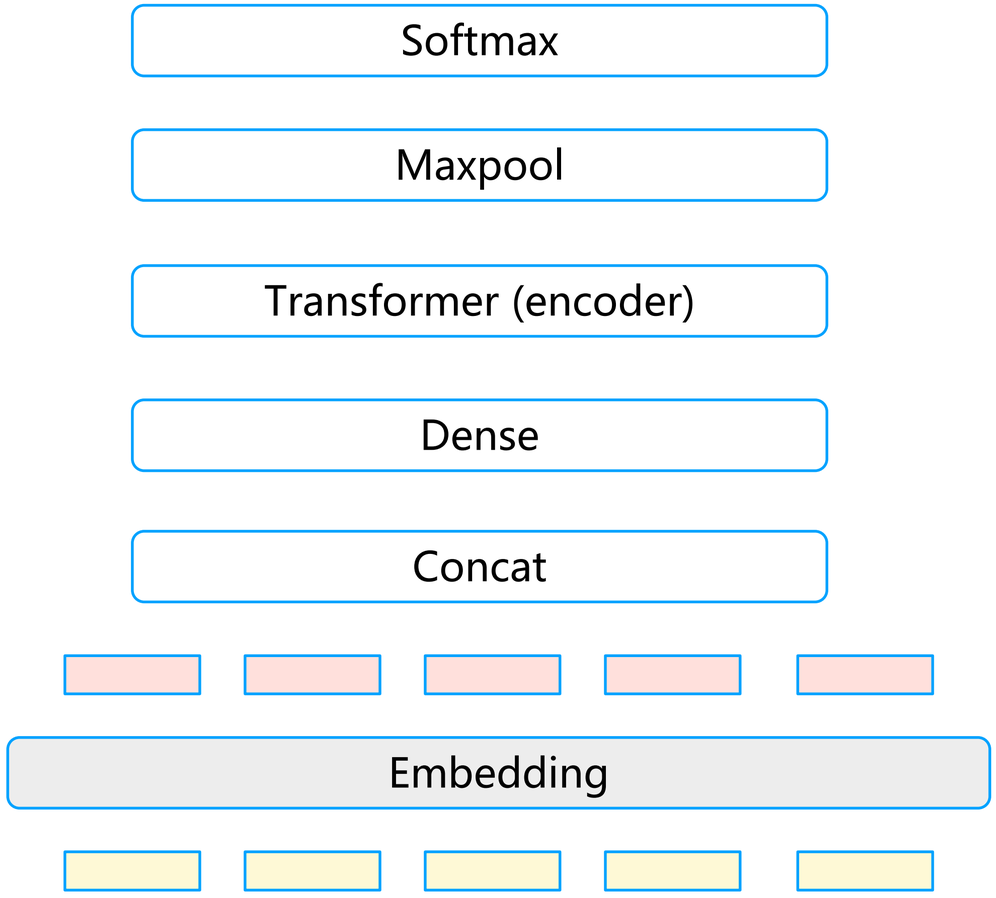
LSTM

- 将用户点击过的广告按时间顺序组织成序列，使用Word2Vec和Glove预训练Embedding，同时还输入有物品的target encode，使用BiLSTM(一层)挖掘序列中信息。
- 我们使用了ad_id、 advertiser_id、industry_id、product_id、product_category这五个广告属性组成五个id序列作为模型的输入。
- id数目大，数据稀疏，Word2Vec预训练很重要。
- Dropout带来的提升很大，我们在Dense中、BiLSTM中、Maxpool前都使用了Dropout，总的提升超过一个百分点。
- 大的LSTM隐向量size带来不错的收益，我们使用了1024的size。



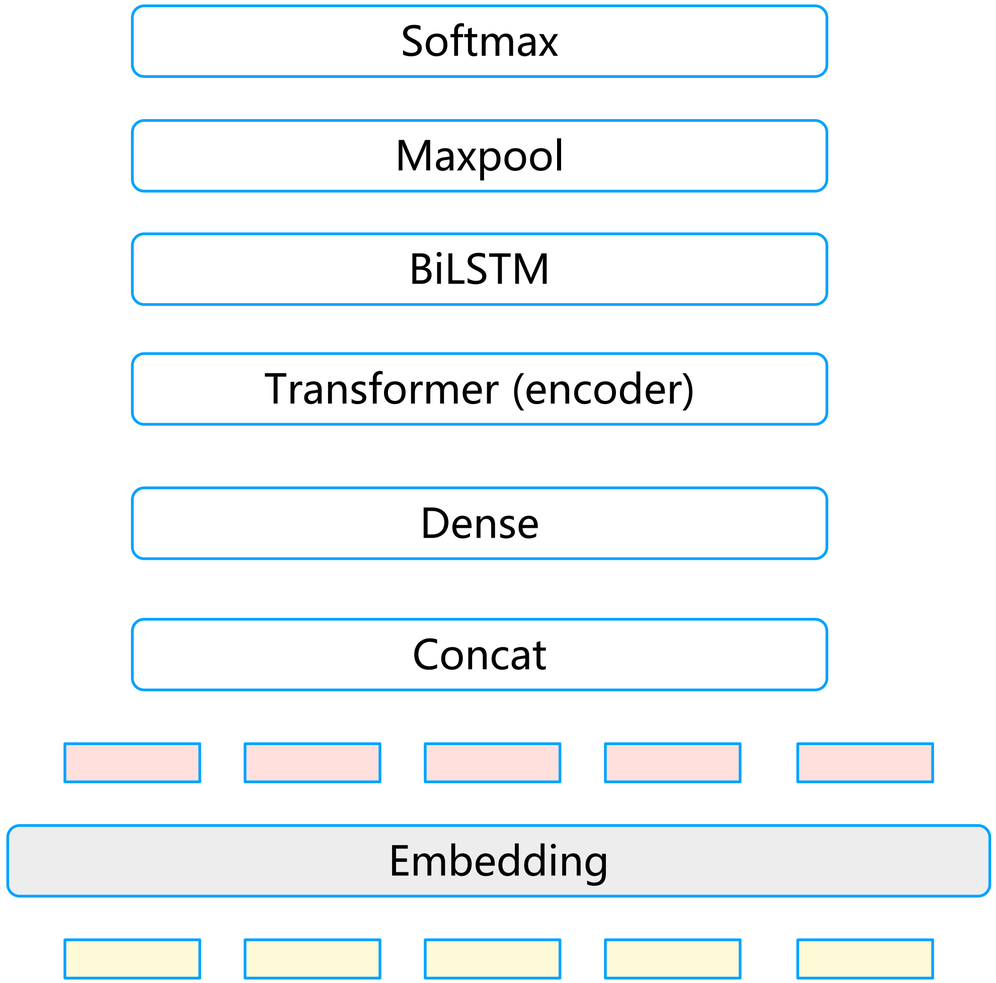
Transformer

- 同样的，我们使用了ad_id、 advertiser_id、industry_id、product_id、product_category这五个广告属性组成五个id序列和它们的target encode作为模型的输入。
- 使用Transformer的encoder部分来挖掘用户点击序列中的信息。
- 相比LSTM，Transformer能够进行序列内全局的交互，能够挖掘到更多的能够区分用户属性的广告组合。
- 我们对比了多头自注意力和单头自注意力，在此赛题任务下，两个精度差别不大。
- 同样的，Dropout的使用至关重要。
- 对于Transformer，较小的学习率能够获得更好的精度。我们使用了3e-4的起始学习率，然后在训练过程中衰减。



Transformer and BiLSTM

- 我们组合了Transformer和BiLSTM来挖掘用户点击序列中的信息。在一层Tansformer的encoder上接一层BiLSTM。
- 相比单独使用Transformer和BiLSTM，这种组合结构能够带来一定的提升(2k)。

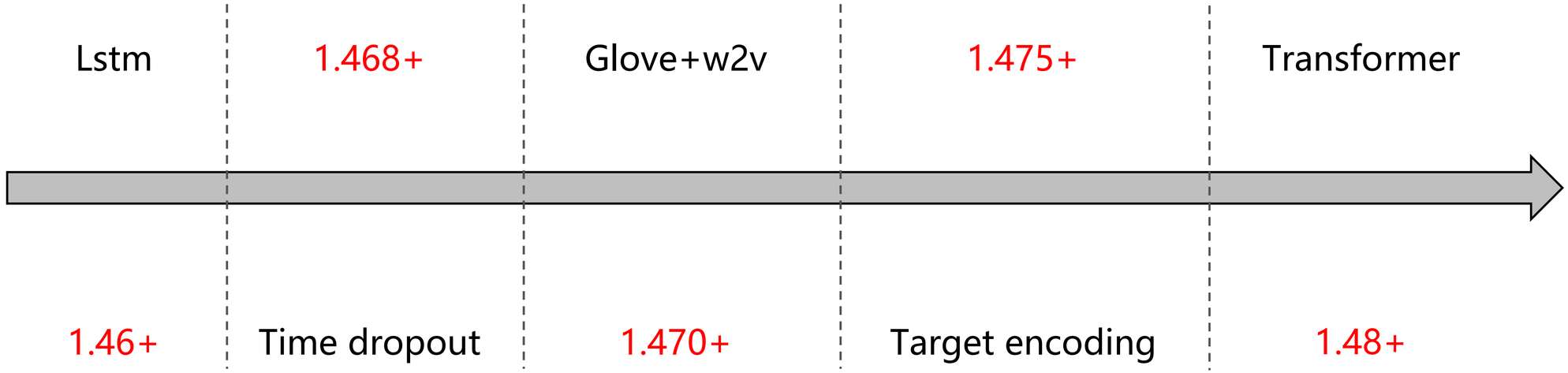


模型融合

- 基于手工特征工程的LightGBM模型精度低，最终方案未使用。
- 对于前面介绍的BiLSTM,Transformer和Transfomer+BiLSTM三种模型，我们使用了十折的方式来训练模型。
- 对于三种模型的预测概率，我们使用加权融合的方式来生成最后的测试集预测结果。

05 总结与思考

Big Improvement



★★ Target encoding

★★ Time dropout

★ Transformer

总结与思考

- 在这个赛题任务下，我们尝试了LightGBM+特征工程的方案和深度学习-序列建模的方案。最终我们选择了深度学习-序列建模的方案。深度学习确实适合这种序列建模、离散特征多的场景。
- 在训练深度学习模型时，预训练词向量、Dropout搭配大的模型容量、合理的学习率非常重要。
- 我们在这个赛题任务下，尝试一些方法，取得了一定的效果。还有一些值得探索的方向没有尝试，比如graph embedding，更细致的特征工程等。

THANKS