CNS-决赛答辩文档

沈琢乔 中国海洋大学 大四
朱锐 YOHO 算法工程师

一、模型创建思路

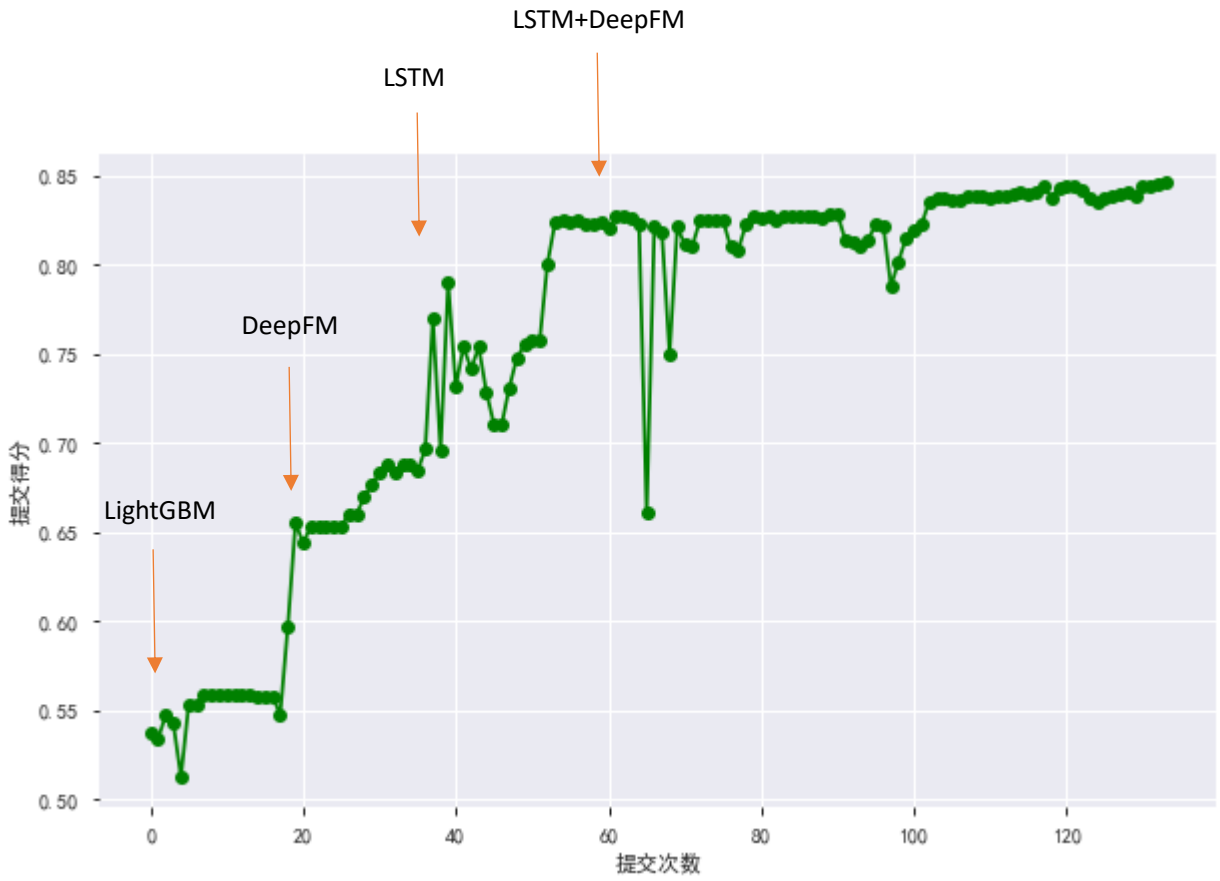

图 1 提交分数趋势

1. 最初拿到赛题，进行了初步的分析，简单的搭建了一个 LGB 模型，一股脑把特征塞进去，效果不理想，约 0.53 成绩。

2. 在已经建立好模型的基础上继续做 EDA，发现含有较多稀疏类别特征，而且训练集 deviceid,newsid 等特征跟测试集中有很大的的交叉。改用 DeepFM 建模，取得第二名的成绩，约 0.6。

图 2　t 时刻 gap 序列（ts 为曝光时间戳）

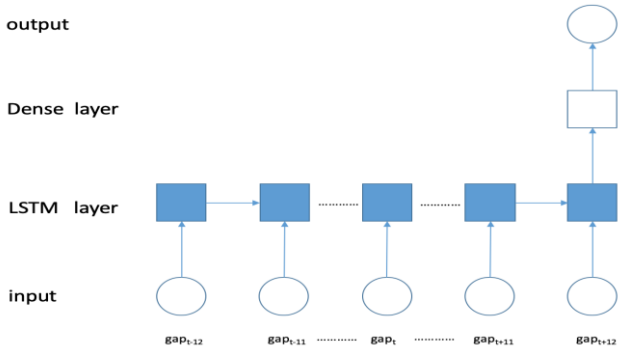

图 3　LSTM 模型

3. 继续分析数据，发现训练集、测试集的曝光时间戳都提供了。这是一个穿越特征。曝光时间戳之间的 gap（后一个 ts 减去当前 ts）很能反映用户是否点击观看了。我们进行了简单的处理，添加一个时间穿越特征后达到 0.65，继续修改特征并调整模型后达到 0.69，再次思考为什么第一名的 baseline 能够有这么高的分数，肯定是某种强特，我们对曝光时间戳 gap 挖掘的还不够深。那可否用 LSTM 自动挖掘序列特征呢？遂依据曝光时间戳 gap 开始构造时序数据集，每条特征为 gap 序列（当前记录的前后 12 个 gap，见图 2），并构建 LSTM 模型（见图 3），线上分数为 0.76+。

4.

图 4　app 推荐列表

图 5　pos、gap 时序特征

5. 下载官方 app 使用后发现（图 4），app 上显示的是有新闻标题或者图片的，曝光时间戳 gap 不能完全反应用户行为，有可能停留了一段时间看了会新闻标题或者图片。使用 app 发现点击了新闻后，pos 是会变化的，所以结合 pos 特征能很好的反应用户行为。我们把 pos embedding（embedding 长度为 8）和 gap 进行拼接(图 5)，组成序列特征，模型跟图 3 相似，不同的是 input 拼接了 pos 的 Embedding（Embedding 长度为 8）。线上达到了 0.82+。

6. 之前准备将原始类别特征（比如 newsid 等等）embedding 拼接 LSTM 的输出再经过几层 MLP 进行点击预测的。理想分数应该有所提升，但是没有，反而下降了。单独使用 LSTM，或者直接使用原始类别特征 embedding 是没有问题。可能问题出现在了网络。通过分析，我们网络有块地方有问题，使用了 dropout 后直接进行了 BatchNormal，这样会有问题，随机 dropout 后会导致 BatchNormal 的输出分布不稳定。解决了这个 bug 后，进一步增加了使用 DeepFM 处理类别特征，LSTM 的输出直接拼接到 DeepFM 的 DNN 模块的输入进行训练，线上分数达到了 0.837+。

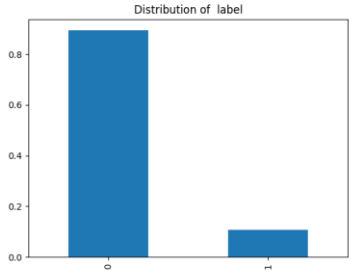

图 6 正负样本分布

7. 由于正负样本相差太大（图 6），模型中加入了 focal loss。线上分数达到了 0.83862。

8. 这时候服务器 20g 内存已经吃紧了（即使了开了虚拟内存，但是运行太慢了），支撑不了过多的特征，比如 CTR 特征，交叉特征等。考虑到内存，并且新增特征收益还要大，只能继续构造时序特征。首先将当前记录的前后 12 个 gap 改为当前记录的前后 14 个 gap，并且将 newsid 的 embedding 加入序列特征中。线上达到了 0.845+。

二、模型说明

2.1 特征：

1. gap（dense feature）、pos（sparse feature）、newsid（sparse feature）组成的序列特征，作为 LSTM 的 input。

2. netmodel，device_vendor，device_version，app_version，deviceid，newsid，pos，这些特征都是 sparse features，作为 DeepFM 的 input。

3. 相同特征名共享 Embedding。

2.2 模型：

见图 7：

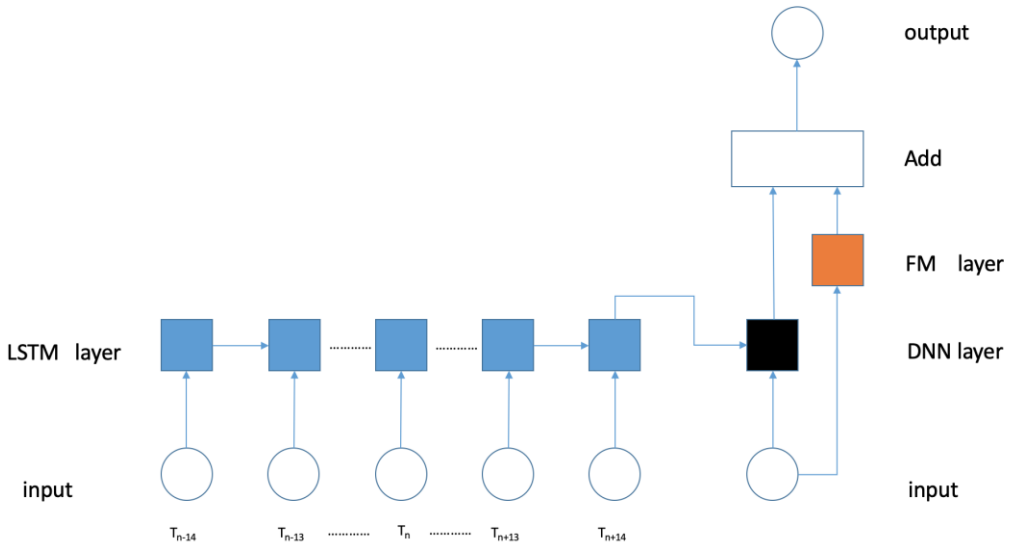

图 7 神经网络结构

首先左边是 LSTM 模块，输入是 pos+gap+newsid 组成的 29 长度（14+14+1）的序列，shape 是 batchsize*29*17（8+1+8：pos，newsid 的 Embedding 长度都为 8，gap 为 1），LSTM 的输出直接拼接 DeepFm DNN 模块的输入，最后把 DeepFM 的 DNN logit 输出和 FM logit 输出相加作为最终输出。感谢浅梦大神的 DeepCTR 框架，优雅高效，提供了各种 CTR 模型。我们的 DeepFM 模型直接 copy DeepCTR 的相关代码并进行了部分修改。

三、相关经验技巧总结

1、 寻找相似的过往比赛的代码分享来学习，多看论文

2、 当遇见有较多 ID 重复出现的数据是请思考是否可以做统计特征和时序特征

3、 对以 F1 为优化指标的题目，阈值的选择也很重要。

4、 由于内存的限制很多方案没有尝试。

模型上比如：单层 LSTM 增加为多层 LSTM；DeepFM 改成 xDeepFM，NFM，NFFM 等等；sparse feature 的 Embedding 可能还没学习到足够的语义，可以增大 Embedding size（目前使用的是 8）或者用 w2v（甚至使用 deepwalk 构造更多数据）来预热 Embedding。

特征方面也有不少可以尝试，比如未使用的 user 表跟 app 表中的特征可以很好的刻画用户画像，还有之前提到的 CTR 特征、交叉特征等等。模型融合也值得尝试，相信神经网络跟树模型融合收益还是很高的。

Pseudo label 在图像及文本类比赛中得到了不错的效果，对于本次比赛也可以尝试下。

参考链接

1、 DeepFM: A Factorization-Machine based Neural Network for CTR Prediction ：https://arxiv.org/abs/1703.04247

2、 DeepCTR：https://github.com/shenweichen/DeepCTR

3、 2019 腾讯广告算法大赛入门-Part1（竞赛小白晋升之路）：https://zhuanlan.zhihu.com/p/63718151