推荐搜索学习进化篇

原创 尘沙杰少 kaggle竞赛宝典 6月14日

阅读本文前，如果你饿了不知道吃什么，欢迎关注**微信视频号《吃货的推荐》**，程序员推荐的吃货分享，周周有特色，每周都要吃好吃饱哦。

言归正传，吃饱了就该好好学习了，在之前我们已经基本了解了传统的推荐算法，从LR，Poly2，FM，FFM，这些方法基本都还是基于二阶交叉，随着DNN的发展，有了词的embedding，我们开始尝试用堆叠网络层的方案来获取深度的特征交叉，于是在中间几年更加多推荐搜索相关的网络考虑在之前二阶交叉的基础上尝试使用更加深层次的交叉，出现了Deep&Wide，二阶加深度的DeepFM，NFM，以及人为控制高阶交叉的DCN和xDeepFM等。

这些都是18年及之前探索的工作方向，这些网络层的改进可以类比到我们比赛中的特征工程，也就是一阶，二阶以及高阶的组合特征，主要的不同之处在于，之前这些特征都是我们人为设计挖掘的。而神经网络是通过网络结构设计来完成的，我们再想想，光靠这些简单的一阶二阶和高阶的统计特征在推荐类的竞赛中能拿到前10吗，答案肯定是否定的，所以我们的网络还欠缺什么呢？亦或者说我们在推荐类的竞赛中除了这些一阶二阶以及高阶的特征工程，还需要做哪些？

如果这些想明白了，那么最近几年推荐类的论文也就很容易明白了，没有错，除了传统的这些一阶二阶和高阶的统计特征之外，我们发现推荐搜索类的比赛最重要的一类特征就是①.用户历史行为特征挖掘，如果想要找类似的例子，可以参考Inscart竞赛中金牌选手的开源，IEEE竞赛中金牌选手所说的encoding技术，还有kaggle手机欺诈预测的Top选手的分享等。上面是最重要的一点，还有一点，就是所有竞赛都会考虑的，②.特征冗余度的问题，不管是我们简单的二阶交叉还是人为设计的可控的高阶交叉，大部分都是枚举式的，而这毫无疑问会带来巨大的冗余，如何能在网络学习的过程中自动进行特征筛选，也是网络设计可以研究的方向。

上面两点做好了，肯定能为网络效果带来更好地效果，但是如果从特征构建的角度来看，似乎还有些许特征没有被处理好，因为看的论文较少，还没有看到类似的处理该类信息的网络，等看到了再补充到后面的文章中，介绍每个网络结构和不同类特征的关联性。

本文我先分享三篇上周读得文章，基本就是上面所述的两点相关的，后面会补充上其他经典文章。

继交叉特征之后加入特征筛选（交叉特征的扩展）

2.6 AutoInt

- 参考文献：AutoInt: Automatic Feature Interaction Learning via Self-Attentive Neural Networks(CIKM,19)

深度网络帮助我们能够挖掘更加深层次的特征交叉，但是因为深度网络的不可解释性，而且从平时大量的实验中，我们发现这样的挖掘并不是最优的，后来大家从FM那边借鉴思想，显示的构建二阶特征的交叉(DeepFM, NFM)，取得了非常不错的效果，但是二阶的交叉明显不是最优的，显示的高阶交叉（DCN，xDeepFM)带来了更加多的帮助，但是这些模型却存在一个较大的问题，空间和时间复杂度太高，虽然人为控制了高阶的特征交叉，但是还是太冗余了。

2.6.1 AutoInt简介

AUtoInt利用了Multihead Self-Attention的方式，对所有的embedding特征进行学习，这么做究竟在干什么呢?

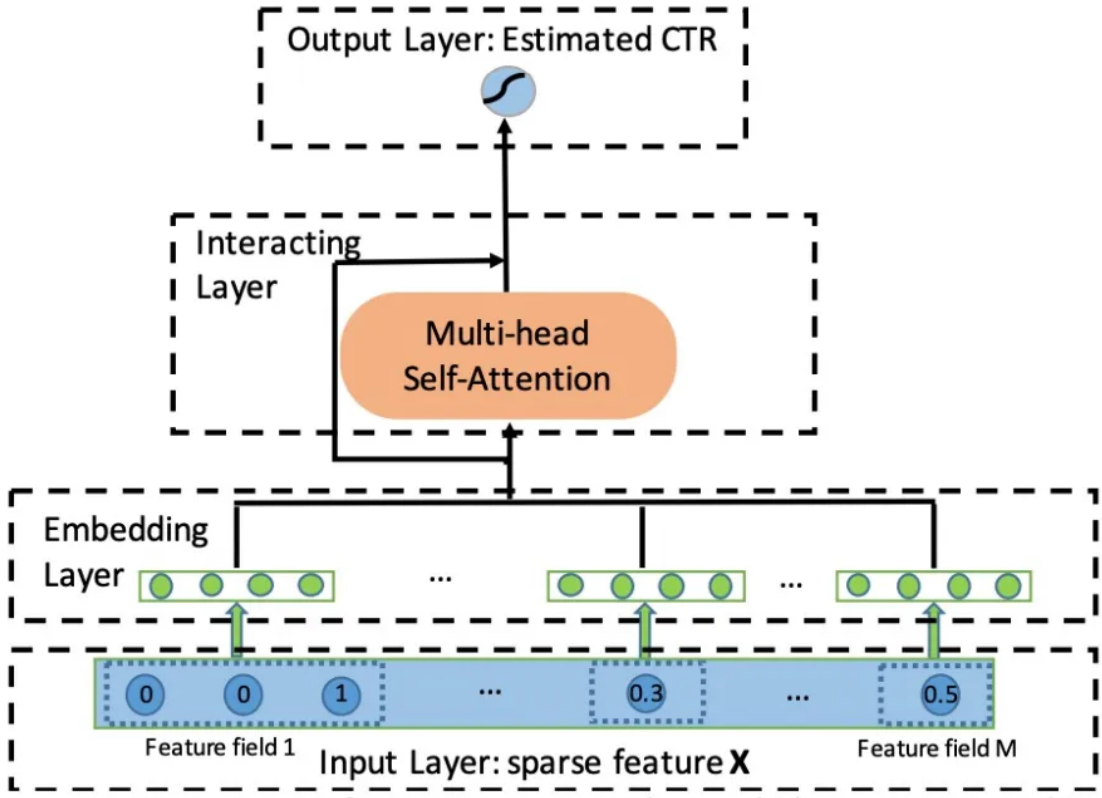
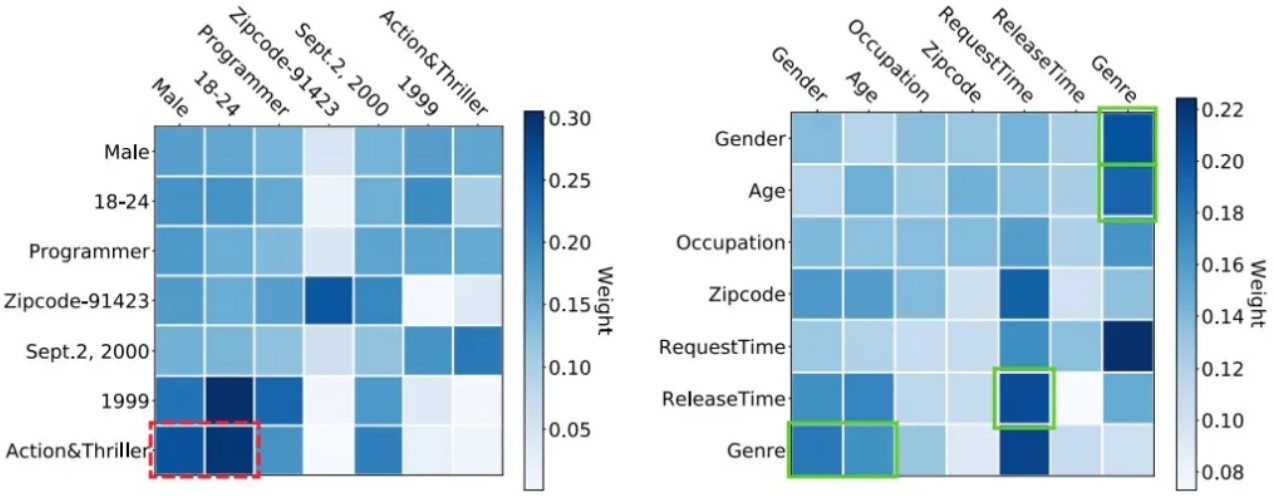


Figure 1: Overview of our proposed model AutoInt. The details of embedding layer and interacting layer are illustrated in Figure 2 and Figure 3 respectively.

我们知道Self-Attention会先计算每个embedding向量和其他向量的相关系数，然后再用所有向量的系数乘上自己的embedding作为下一层的输入，所以我们可以计算得到每个向量和其他向量的关系，再将相关性强的特征进行组合，作为新的特征，,个人这样的组合最大的意义我们可以找到经常一起出现的特征。



从作者最后的组合特征的可视化中，我们可以看到，应该是这些特征经常一起出现，所以关系较近，被捕捉到了。

2.6.2 AutoInt的优缺点

1. 1.优点

> 找到意思相近的embedding并进行组合，形成一些可解释性较强的组合特征；
> 大量的实验也验证这种方式的高阶交叉组合的优势；

2. 缺点

> 个人感觉还是未能充分挖掘有意义的高阶交叉特征；此处的组合只是找到了关系相近的特征，关系相近的特征进行组合并不一定是合适的方式，也就是说multi-head selfattention能做到有意义的特征组合，但却不能说明关系不相近的特征的意义就不大。

加入用户历史序列（新结构的引入&扩展）

2.7 DIN

- 参考文献: Deep Interest Network for Click-Through Rate Prediction(KDD,18)

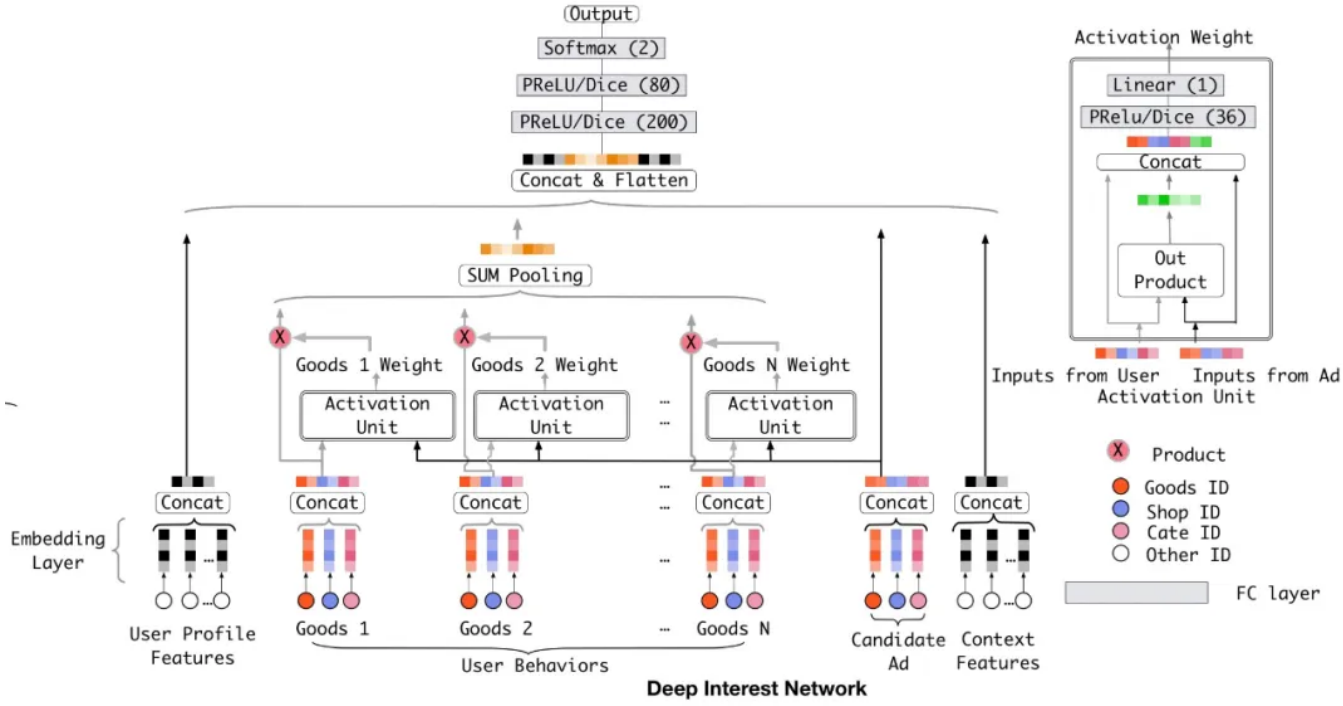
在上面的文章中,我们一直专注于交叉特征对构建,包括二阶对交叉(DeepFM,NFM),人为控制高阶的交叉等(DCN,xDeepFM)等,继续模拟特征的交叉带来模型方面的提升会受限，所以此处我们将方向转向数据的使用方面，传统的模型在建模时考虑了数据的统计信息,但是在序列的信息挖掘方面却不是非常好,而序列中又包含了非常多的用户信息,例如通过用户的历史购买点击序列，我们就可以很好地捕捉到用户的多维不同的兴趣。

本篇文章，我们通过用户的历史行为结合attention机制，来描述用户的多峰的兴趣，通过计算当前的商品或者广告与用户多个兴趣的关系来描述用户点击该商品的可能性。DIN的设计对于工业的帮助比较大，因为上线会受到内存等的影响，我们的user的vector往往不能很大，这个时候user的vector的表示能力会大打折扣，很难表示用户的多峰兴趣，而DIN基于用户的历史行为加入attention机制很好地缓解了用户表示的问题。

> zhouguorui:用户的兴趣是多峰的不是单峰的，DIN的attention机制部分是为了用一个fix length的vector刻画用户面对不同的商品展现出不同的兴趣,这个点看起来很简单,但是传统的预估方法在一个user面对不同商品时用一个同样的vector来表达这个user,如果在这种情况下要想表达多样的兴趣,最简单的方案是增加user vector的维度,然而这会带来overfitting和计算压力

2.7.1 DIN简介

我们先看DIN的网络结构，



从上面的结构中，我们发现DIN与传统的模型最大的不同之处在于中间多了一条序列，这部分序列是用户的历史行为序列，这么做的好处是什么呢？我们发现DIN会用当前商品与用户在历史上的每个商品计算得到一个权重，用历史商品与该商品的计算得到权重再与该商品的外积作为一个indicator（暗示该用户对此类商品的兴趣），最后再把多个indicators做sum pooling并concat起来，用于最后的训练。

关于DIN的两个实践技巧:

1. Mini Batch reg：更加适用于工业界，每次仅仅更新在Mini-Batch采样的稀疏特征，这也使得我们的模型能在大规模数据情况下完成L2 norm的更新。
2. Dice: Dice可以看做是PRelu的一种扩展， $f(s) = p(s) \dot s + (1 - p(s)) \dot \alpha s, p(s) = \frac{1}{1+ e^{- \frac{s - E[s]}{\sqrt(Var[x] + \epsilon)}}}$

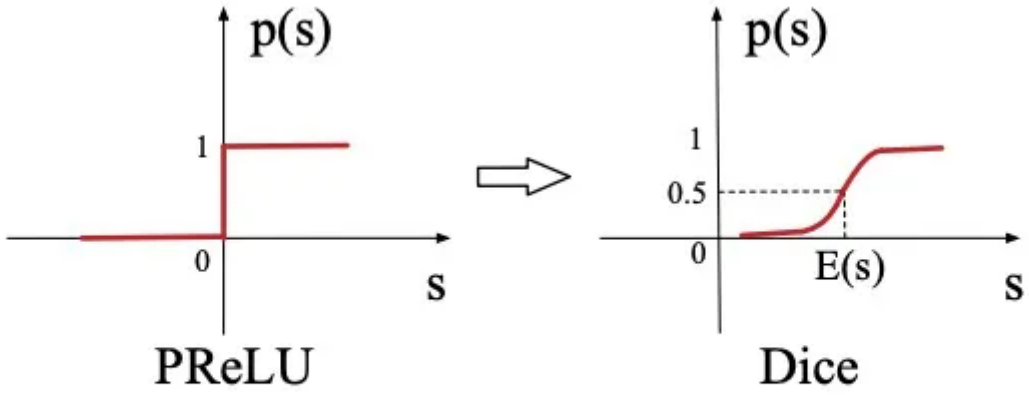


Figure 3: Control function of PReLU and Dice.

2.7.2 DIN的优缺点

1. 1.优点

> 挖掘用户的历史行为信息，用到了用户的序列信息,既可从中挖掘用户的兴趣信息，捕获到兴趣的动态进化性；同时这样的操作还可以弥补user vector的限制,缓解了传统扩充user vector长度的overfitting和计算压力大的问题；
> 文章中的Mini-Batch Aware Reg和Dice在模型防止过拟合的过程中也帮助非常多,带来了较多的帮助。

2. 缺点

> 此处的attention机制忽略了序列关系，所以虽然用到了历史的行为信息，但是在序列的前后关系挖掘上并不是非常理想。

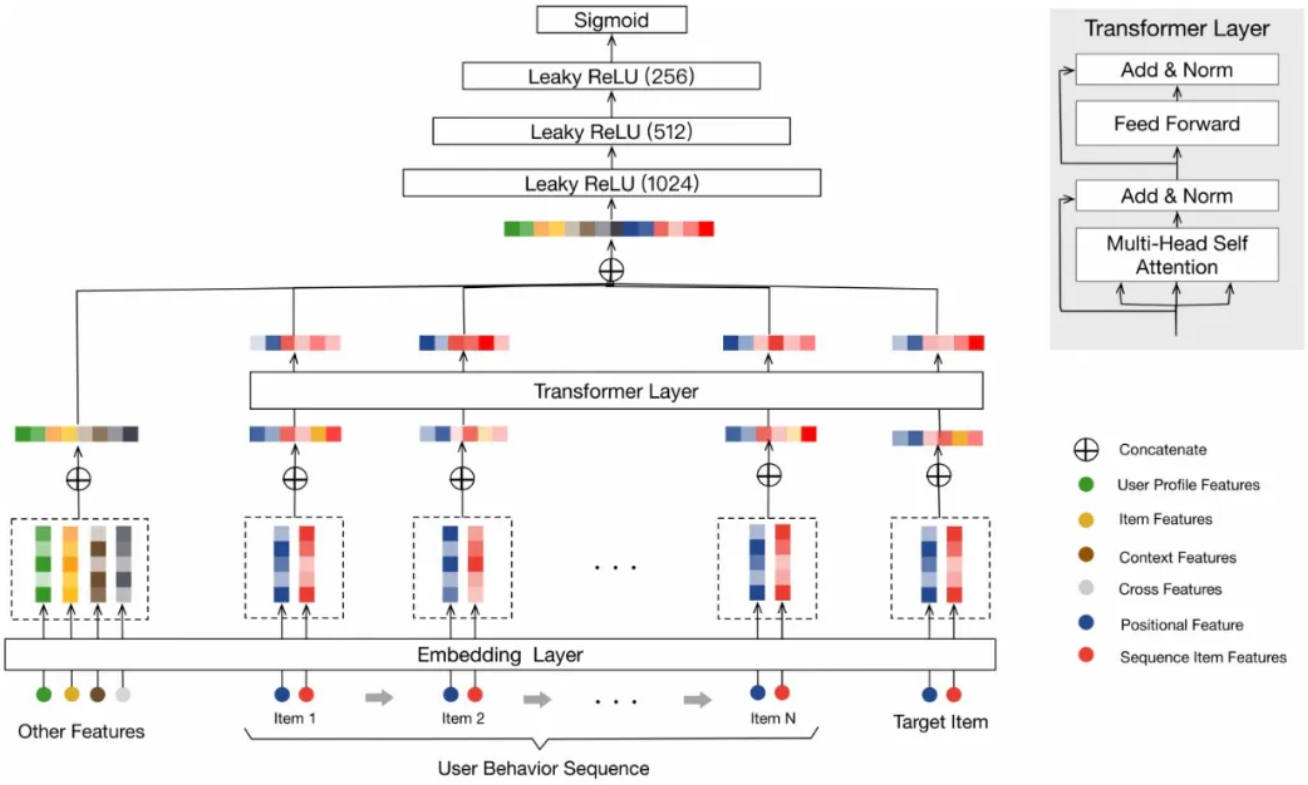
2.8 BST

- 参考论文：Behavior Sequence Transformer for E-commerce Recommendation in Alibaba

在DIN部分,我们开始关注到了用户的历史点击等行为信息，但是从DIN的设计中，我们发现DIN在处理序列问题时较好地考虑了用户的历史购买商品信息，也将现有的商品或者广告和历史点击的商品或广告进行attention来捕获用户的多兴趣，取得了非常不错的效果，但是DIN的设计忽略了用户兴趣的变化等信息，未能较好地捕捉序列信息。

2.8.1 BST简介

BST对用户对历史行为进行编码同时模仿原文将商品对应的时间信息也通过编码加入到模型中，我们看BST的网络结构，发现和DIN较大的不同之处，我们直接将用户对商品对行为序列以及当前对商品全部embedding之后concat之后经过transformer层输出，然后再进行concat经过mlp进行输出。



时间编码

- 和传统的编码方式不一致，在此处我们用推荐的时间和用户点击商品i的时间差（相对时间）来作为位置编码。

2.8.2 BST的优缺点

1. 优点

> 不仅仅是挖掘用户的多维度信息，同时尝试去挖掘用户的兴趣演变的信息。效果相较于DIN等也得到了不错的提升；

2. 缺点

> 该文章的贡献主要在于将transformer加进来，没有太多方法上的创新，而且时间序列的挖掘，并没有看到用户的内容，所以可能挖掘不充分。

因为本周时间原因，非常多的其他经典最新论文还未总结，会在后续更新，后面我们也会在kaggle的实践数据集上进行实验对比，欢迎有兴趣的小伙伴持续关注我们的公众号。

推荐搜索相关论文学习基础篇

基础篇：推荐搜索相关论文学习基础篇

强化篇：推荐搜索相关论文学习强化篇