

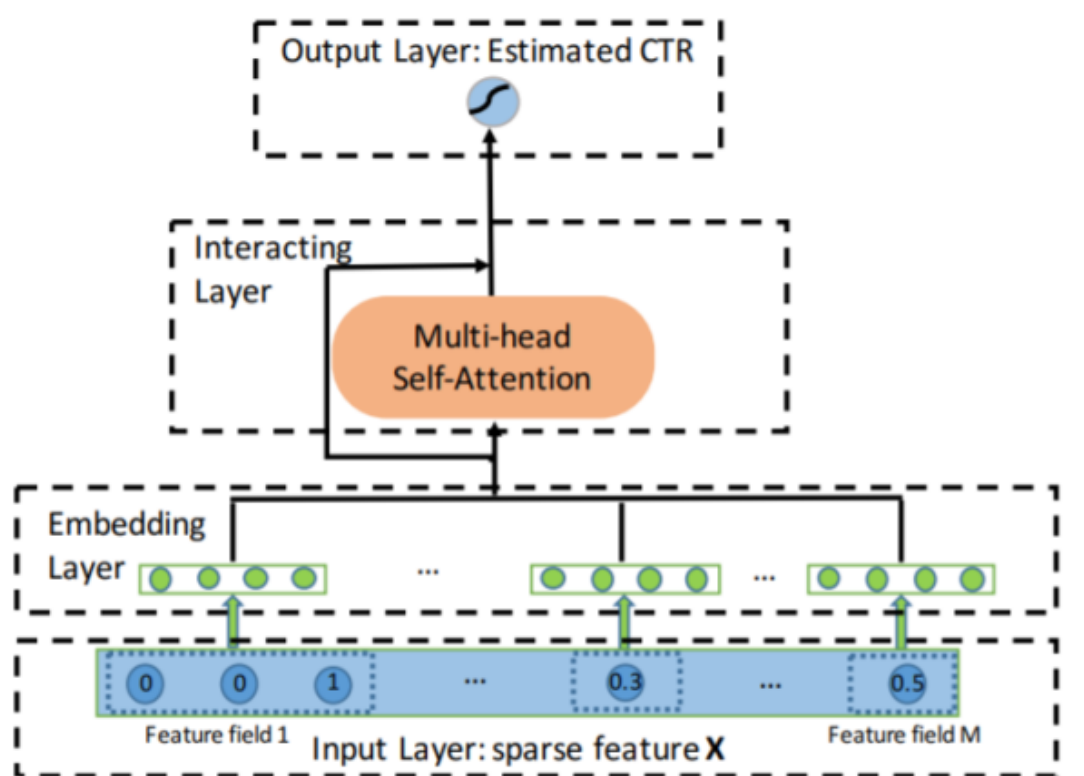
AutoInt

这篇文章还是解决**高阶显式交叉特征**的，使用了Transformer中的Multi-head self attention+**残差连接**来显式地捕捉不同阶的高阶交叉特征。

1. 模型思路

1.1 Embedding

首先，还是将**高维稀疏**的特征做成embedding（这也是FM最重要的思想），不然那么高维的特征，会导致**过拟合**的（只有记忆没有泛化）。然后，就要考虑如何去捕捉显式高阶交叉特征的问题了。



1.2 显式高阶交叉特征

使用DNN来隐式的捕捉交叉特征自然无可厚非，例如PNN、FNN、NFM在二阶特征交叉的基础上又叠加了DNN来捕捉高阶交叉特征；WDL、DeepFM的Deep端也是用来隐式的捕捉交叉特征。但是，DNN在每层之间是以**"相加"**的方式来进行特征交叉的。这个在讲到PNN的时候说过，**相加**的方式在捕捉交叉特征的任务上不如**相乘**的方法来得好。而且，DNN的特征交叉是**bit-wise**的，而不是**vector-wise**的，没有对不同field的embedding进行区分，同一个field中的不同元素也可以互相影响。最后，DNN**隐式**捕捉交叉特征，其可解释性不强，我们甚至都不知道最后DNN拟合的函数完成了几阶交叉。

... The final function learned by DNNs can be arbitrary, and there is no theoretical conclusion on what the maximum degree of feature interaction is. -- xDeepFM

所以，AutoInt使用了multi-head self-attention, 每一层都让每一个feature去和其他所有的特征进行交互，然后自动的为每个特征分配权重、进行融合。“multi-head”就说明不同的“头”将feature映射到了**不同的特征空间**，因此在不同的特征空间可以捕捉到不同的特征交互。同时，在每个multi-head self-attention层中间还增加了残差连接，这是为了可以捕捉不同阶的交叉特征。

1.3 高阶特征的分析

首先，先定义一下什么叫“p阶交叉特征”：

交叉特征 $g(x_{i_1}, \dots, x_{i_p})$ ，其中每个feature都来自不同的field，且 $g(\cdot)$ 做的是non-additive combination（例如乘法、内积、外积都算），那么 $g(x_{i_1}, \dots, x_{i_p})$ 就叫p阶交叉特征。

AutoInt使用multi-head self-attention来捕捉高阶交叉特征。和Transformer中的计算方法一样，每个feature都对应Q,K,V三个向量，然后用该feature对应的Q向量去和其他所有feature的K向量相乘，得到权重系数对所有feature的V向量进行加权求和。为了保留之前层计算而得的交叉特征，使用残差连接。

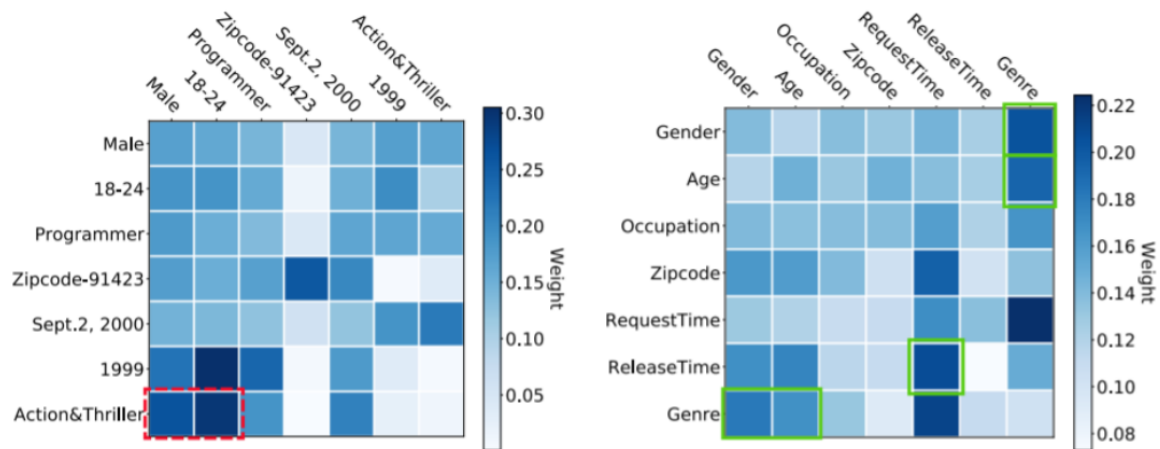
时间复杂度为 $O(Mdd' + M^2d')$ 。



例如，假设只有四个特征 x_1, x_2, x_3, x_4 ，那么在第一个self-attention层，每个feature都和其他的特征做了交互，得到二阶交叉特征，如 $g(x_1, x_2), g(x_1, x_3), g(x_2, x_3)$ 。（这是因为每层的每个feature对应的Q向量去乘以及其他所有feature的K向量时，就引入了**乘法**，构成了交叉特征。）在第二层，由于还有第一层的**残差连接**，所以可以捕捉三阶、四阶的交叉特征，例如 $g(x_1, x_2, x_3), g(x_1, x_2, x_3, x_4)$ 。

1.4 可解释性

因为使用了attention，所以提供了一定的可解释性。



(a) Label=1, Predicted CTR=0.89 (b) Overall feature interactions

Figure 7: Heat maps of attention weights for both case- and global-level feature interactions on MovieLens-1M. The axes represent feature fields $\langle \text{Gender}, \text{Age}, \text{Occupation}, \text{Zipcode}, \text{RequestTime}, \text{ReleaseTime}, \text{Genre} \rangle$. We highlight some learned combinatorial features in rectangles.

左图是某个case的可解释性分析，发现 $\langle \text{Male}, \text{Action\&Thriller} \rangle$ 交叉特征和 $\langle 18-24, \text{Action\&Thriller} \rangle$ 交叉特征很重要。右图是所有field的相互交叉重要度（是一个平均值）。