

论文链接

[Deep & Cross Network for Ad Click Predictions`](https://arxiv.org/abs/1708.05123)

<https://arxiv.org/abs/1708.05123>

[Wide & Deep Learning for Recommender Systems](https://dl.acm.org/citation.cfm?id=2988454)

<https://dl.acm.org/citation.cfm?id=2988454>

用神经网络的视角看FM：嵌入后再进行内积

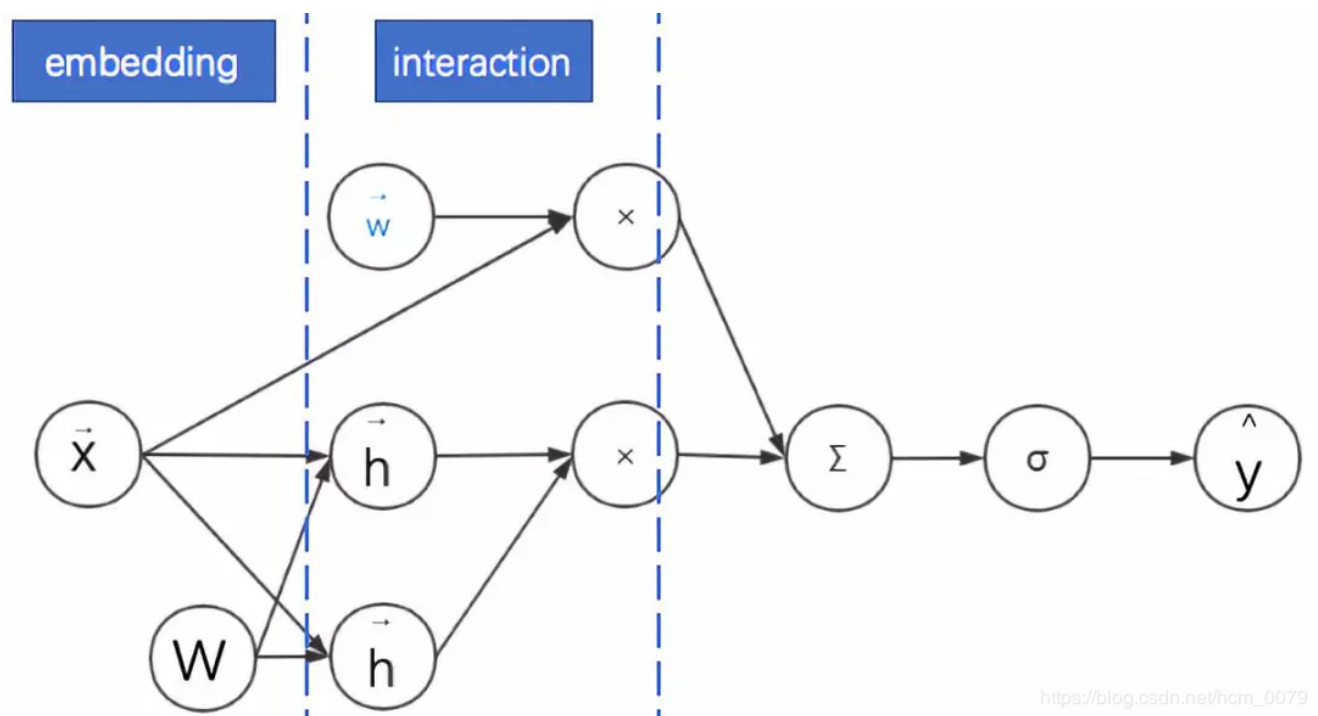
我们观察FM公式的矩阵内积形式：

$$y_{FM} = \sigma(\langle \vec{w}, \vec{x} \rangle + \langle W \cdot \vec{x}, W \cdot \vec{x} \rangle)$$

发现 $W \cdot x$ 部分就是将离散系数特征通过矩阵乘法降维成一个低维稠密向量。这个过程对神经网络来说就叫做嵌入（embedding）。所以用神经网络视角来看：

- FM首先是对离散特征进行嵌入。
- 之后通过对嵌入后的稠密向量进行内积来进行二阶特征组合。
- 最后再与线性模型的结果求和进而得到预估点击率。

其示意图如下。为了表述清晰，我们绘制的是神经网络计算图而不是网络结构图——在网络结构图中增加了权重 W 的位置。



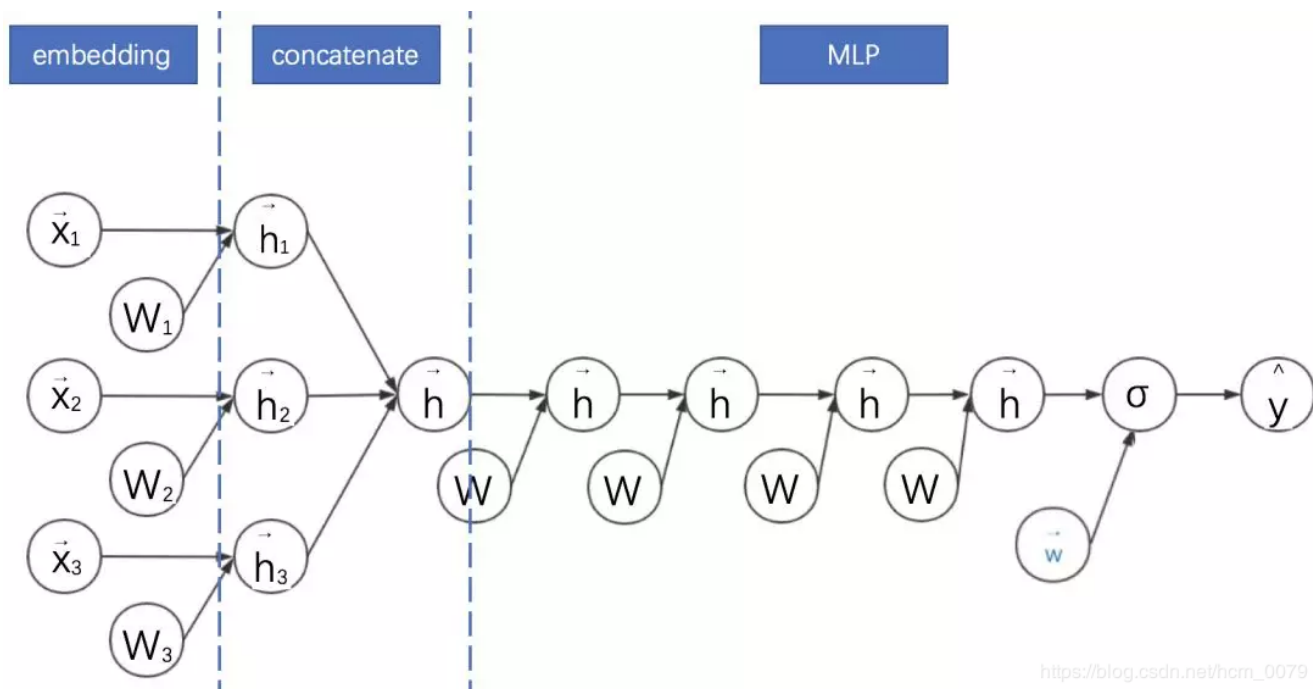
embedding+MLP：深度学习CTR预估的通用框架

embedding+MLP是对于分领域离散特征进行深度学习CTR预估的通用框架。深度学习在特征组合挖掘(特征学习)方面具有很大的优势。比如以CNN为代表的深度网络主要用于图像、语音等稠密特征上的学习，以W2V、RNN为代表的深度网络主要用于文本的同质化、序列化高维稀疏特征的学习。CTR预估的主要场景是对离散且有具体领域的特征进行学习，所以其深度网络结构也不同于CNN与RNN。

具体来说，embedding+MLP的过程如下：

- 对不同领域的one-hot特征进行嵌入（embedding），使其降维成低维度稠密特征。
- 然后将这些特征向量拼接（concatenate）成一个隐含层。
- 之后再不断堆叠全连接层，也就是多层感知机(Multilayer Perceptron, MLP，有时也叫作前馈神经网络)。
- 最终输出预测的点击率。

其示意图如下：



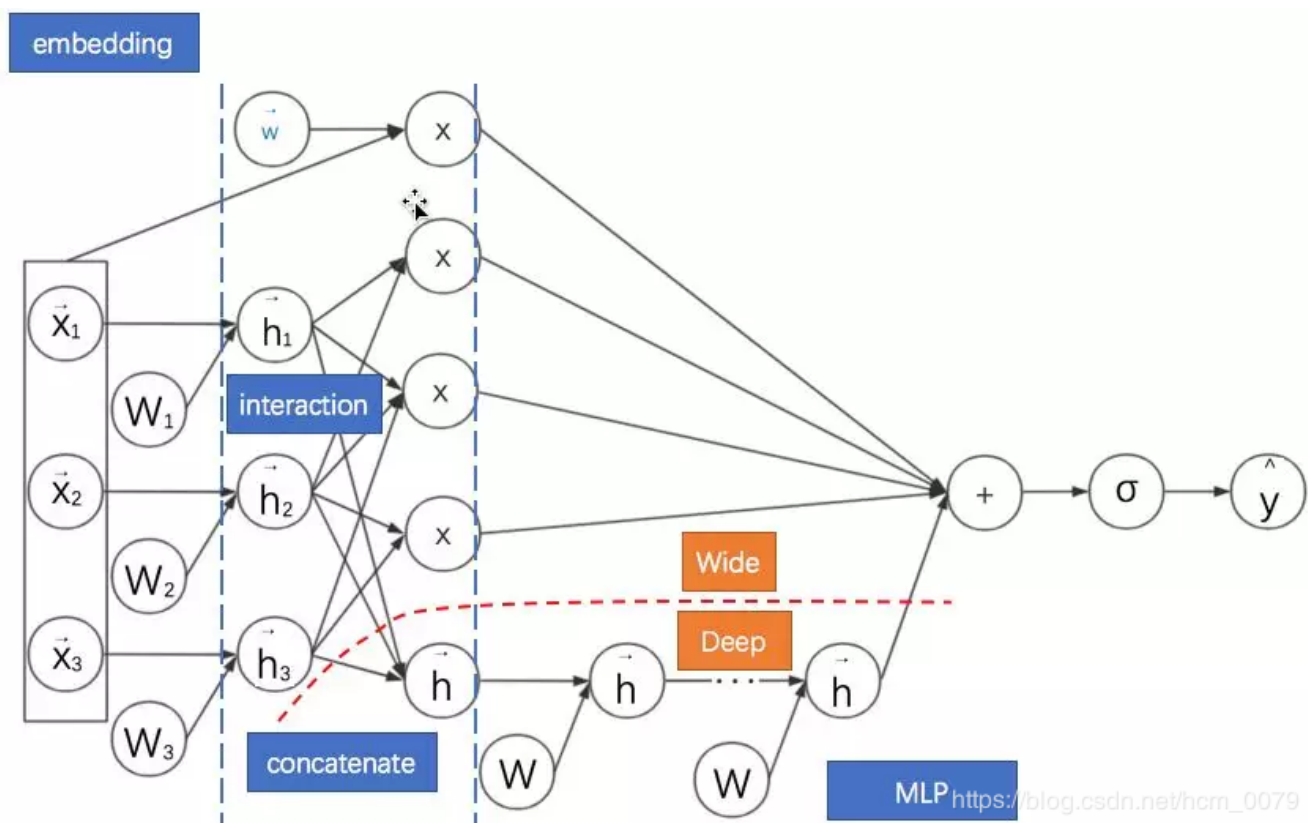
embedding+MLP的缺点是只学习高阶特征组合，对于低阶或者手动的特征组合不够兼容，而且参数较多，学习较困难。

DeepFM: FM与MLP的并联结合

针对FNN需要预训练的问题，Huifeng Guo等提出了深度因子分解机模型（Deep Factorisation Machine, DeepFM, 2017）。该模型的特点是：

- 不需要预训练。
- 将考虑领域信息的FM部分与MLP部分并联起来（借用初中电路的术语），其实就是对两个模型进行联合训练。
- 考虑领域信息的FM部分的嵌入向量拼接起来作为MLP部分的输入特征，也就是是两个模型共享嵌入后的特征。

其计算图如下所示：

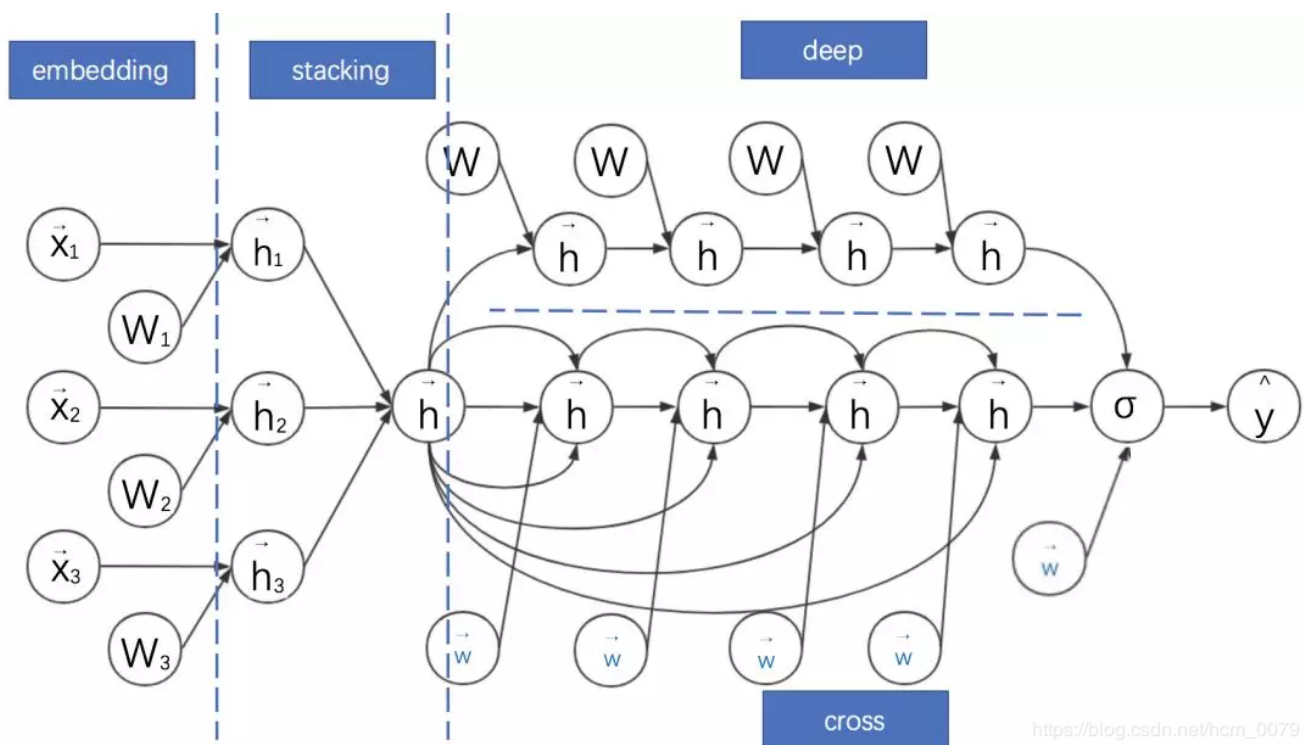


通过观察DeepFM的计算图可以看出红色虚线以上部分其实就是FM部分，虚线以下就是MLP部分。

DCN:高阶FM的降维实现

以上的FM推广形式，主要是对FM进行二阶特征组合。高阶特征组合是通过MLP实现的。但这两种实现方式是有很大不同的，FM更多是通过向量embedding之间的内积来实现，而MLP则是在向量embedding之后一层一层进行权重矩阵乘法实现。

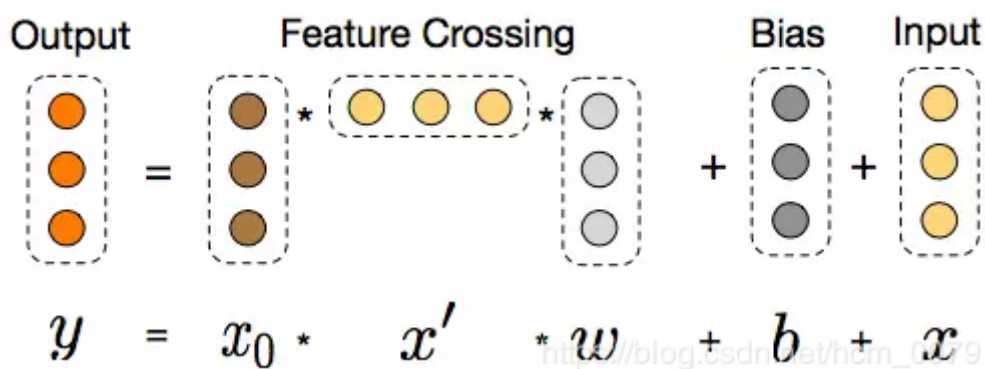
可否直接将FM的过程在高阶特征组合上进行推广？答案是可以的。Ruoxi Wang等在2017提出的深度与交叉神经网络（Deep & Cross Network, DCN）就是在这个方向进行改进的。DCN的计算图如下：



DCN的特点如下：

- Deep部分就是普通的MLP网络，主要是全连接。
- 与DeepFM类似，DCN是由embedding+MLP部分与cross部分进行联合训练的。Cross部分是对FM部分的推广。
- cross部分的公式如下：

$$\mathbf{x}_{l+1} = \mathbf{x}_0 \mathbf{x}_l^T \mathbf{w}_l + \mathbf{b}_l + \mathbf{x}_l = f(\mathbf{x}_l, \mathbf{w}_l, \mathbf{b}_l) + \mathbf{x}_l,$$



- 可以证明，cross网络是FM的过程在高阶特征组合的推广。完全的证明需要一些公式推导，可以直接参考原论文的附录。
- 而用简单的公式证明可以得到一个很重要的结论：只有两层且第一层与最后一层权重参数相等时的Cross网络与简化版FM等价。

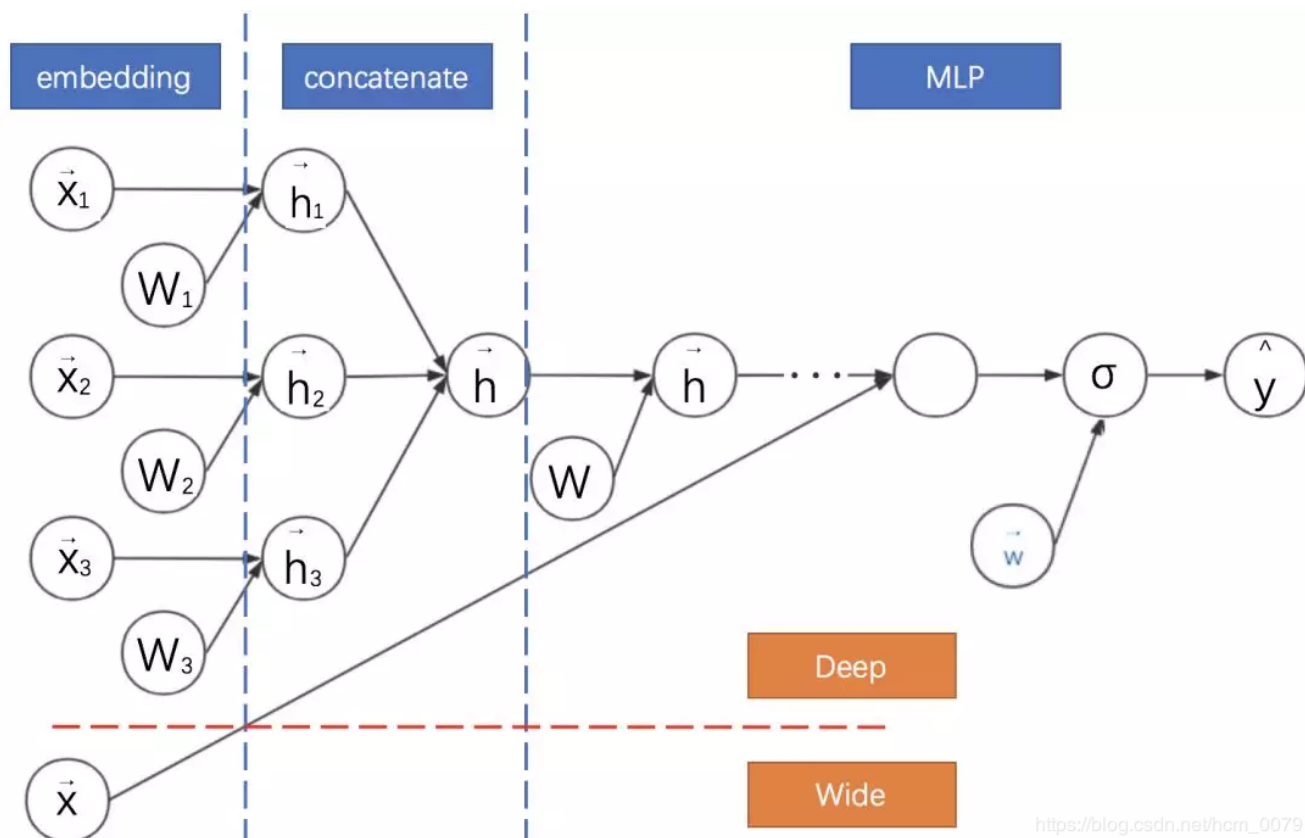
- 此处对应简化版的FM视角是将拼接好的稠密向量作为输入向量，且不做领域方面的区分（但产生这些稠密向量的过程是考虑领域信息的，相对全特征维度的全连接层减少了大量参数，可视为稀疏链接思想的体现）。而且之后进行embedding权重矩阵 W 只有一列——是退化成列向量的情形。
- 与MLP网络相比，Cross部分在增加高阶特征组合的同时减少了参数的个数，并省去了非线性激活函数。

Wide&Deep: DeepFM与DCN的基础框架

开篇已经提到，本文思路有两条主线。到此为止已经将基于FM的主线介绍基本完毕。接下来将串讲从embedding+MLP自身的演进特点的CTR预估模型主线，而这条思路与我们之前的FM思路同样有千丝万缕的联系。

Google在2016年提出的宽度与深度模型（Wide&Deep）在深度学习CTR预估模型中占有非常重要的位置，它奠定了之后基于深度学习的广告点击率预估模型的框架。

Wide&Deep将深度模型与线性模型进行联合训练，二者的结果求和输出为最终点击率。其计算图如下：



我们将Wide&Deep的计算图与之前的模型进行对比可知：

- Wide&Deep是前面介绍模型DeepFM与DCN的基础框架。这些模型均采用神经网络联合训练的思路，对神经网络进行并联。
- DeepFM、DCN与Wide&Deep的Deep部分都是MLP。
- Wide&Deep的Wide部分是逻辑回归，可以手动设计组合特征。
- DeepFM的Wide部分是FM，DCN的Wide部分是Cross网络，二者均不强求手动设计特征。但此时都与字面意义上的Wide有一定差异，因为均共享了降维后的嵌入特征。

参考资料：[从FM推演各深度CTR预估模型\(附代码\)](https://www.jiqizhixin.com/articles/2018-07-16-17)
<https://www.jiqizhixin.com/articles/2018-07-16-17>