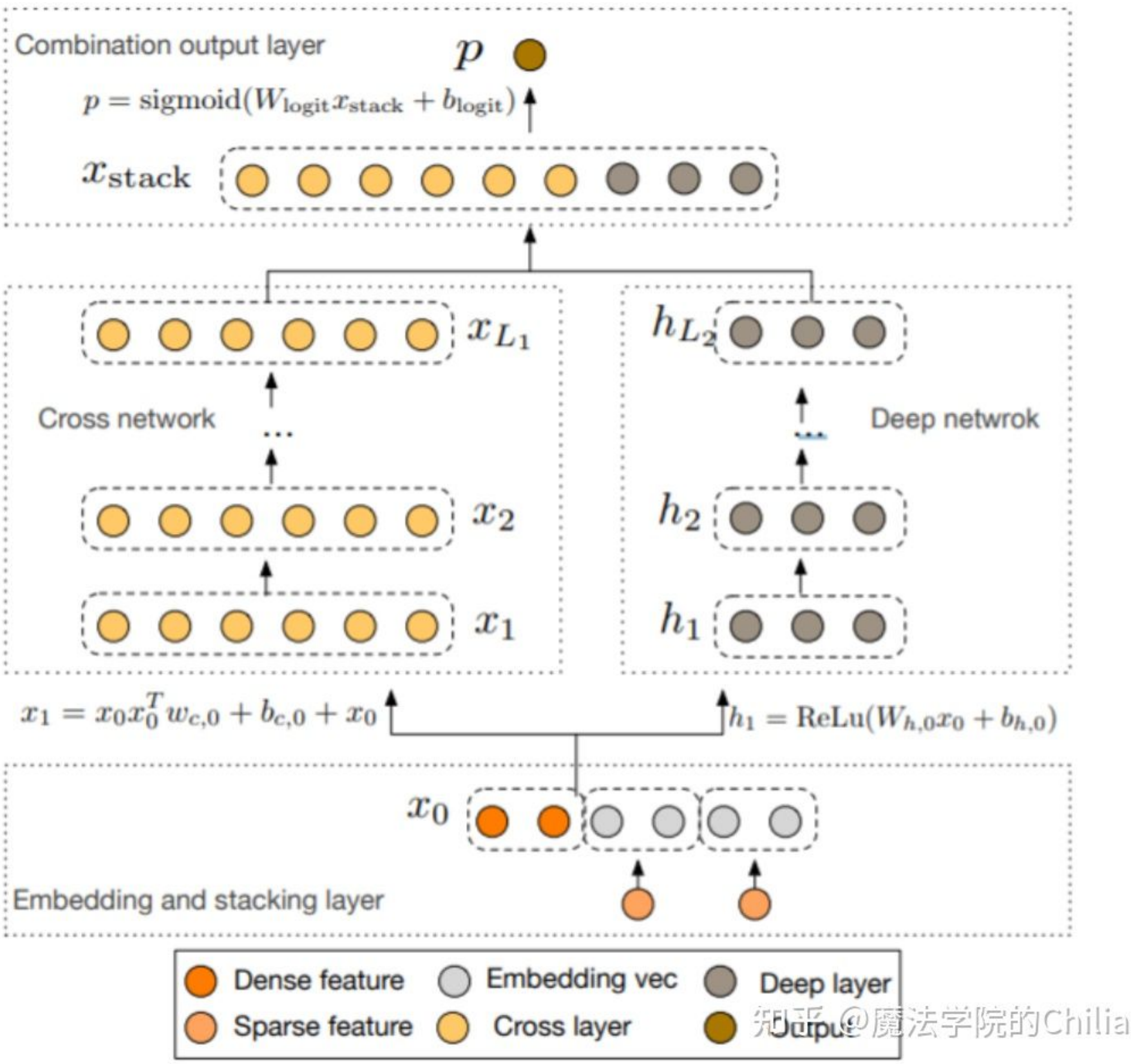
Deep & Cross [2017] -- 高阶交叉空许约

出自论文 *Deep & Cross Network for Ad Click Predictions*。该模型主要特点在于提出Cross network，用于高阶特征的自动化显式交叉编码。这是因为传统DNN对于高阶特征的提取效率并不高，我们甚至不知道DNN能否构造出交叉特征来。而Cross Network通过调整结构层数能够**显式构造出有限阶（bounded-degree）交叉特征**，可以提高了模型的表征能力。同时，DCN引入了**残差结构**的思想，使得模型能够更深，同时把握不同阶的信息。

回忆WDL中的Wide端只能**人工构造**交叉特征，并不能够自动化；FM中虽然能够自动构建交叉特征，但是只能构造二阶。那么能不能显式的构造任意有限阶交叉特征呢？

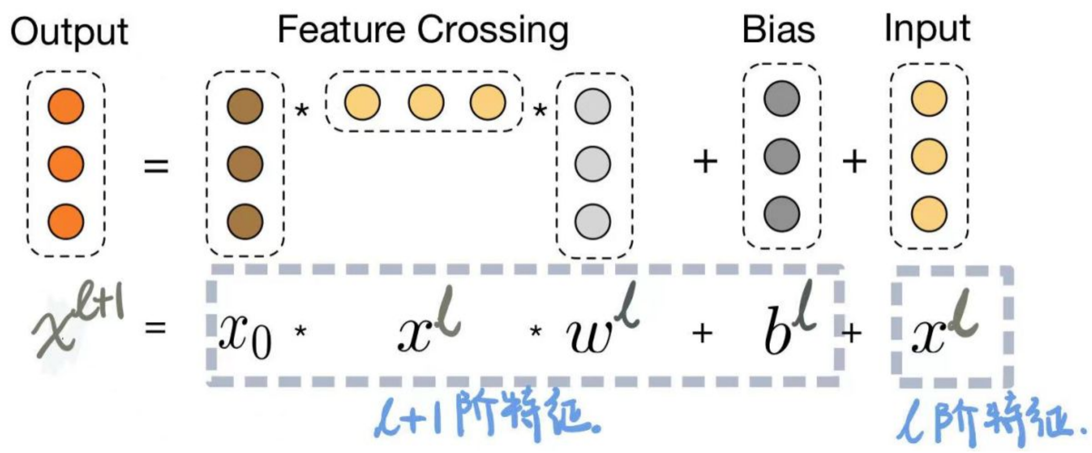
0x01. 模型结构



还是分为两路的结构，其中Deep端就是一个普通的MLP，不必多言；Cross端是模型的核心，数学表达式如下：

$$X_{l+1} = X_0 X_l^T W_l + b_l + X_l = f(X_l, W_l, b_l) + X_l$$

其中 $X_l, X_{l+1} \in \mathbb{R}^d$ 分别代表Cross Network的第 $l, l+1$ 层的输出，$W_l, b_l \in \mathbb{R}^d$ 分别为第l层的参数与偏置项，是需要学习的参数。x0为一开始输入特征的embedding的拼接，它会在每一层都参与运算。再使用ResNet的思想，将 X_l 直接输入下一层，可以将原始信息在CrossNet中进行传递。这样，到第k层的时候，我们就有了从1阶~k阶所有阶的交叉信息(hopefully). 结构上可以用下面的图来辅助理解：



利用 x_0 与 x^l **向量外积**得到embedding中所有的**元素**的交叉组合（注意不是每个feature的交叉组合，而是embedding元素的交叉组合，这也就是DCN中所谓"交叉"和FM中"交叉"的不同之处！），层层叠加之后便可得到任意有界阶组合特征，当cross layer叠加到 l 层，交叉最高阶可以达到 l+1 阶。下面详细地计算一下：

令

$$X_0 = \begin{bmatrix} x_{0,1} \\ x_{0,2} \end{bmatrix}$$

那么第一层有一阶、二阶"交叉特征"：

$$\begin{aligned}
X_1 &= X_0 X_0' W_0 + X_0 \\
&= \begin{bmatrix} x_{0,1} \\ x_{0,2} \end{bmatrix} [x_{0,1} x_{0,2}] \begin{bmatrix} w_{0,1} \\ w_{0,2} \end{bmatrix} + \begin{bmatrix} x_{0,1} \\ x_{0,2} \end{bmatrix} \\
&= \begin{bmatrix} x_{0,1}^2, x_{0,1} x_{0,2} \\ x_{0,2} x_{0,1}, x_{0,2}^2 \end{bmatrix} \begin{bmatrix} w_{0,1} \\ w_{0,2} \end{bmatrix} + \begin{bmatrix} x_{0,1} \\ x_{0,2} \end{bmatrix} \\
&= \begin{bmatrix} w_{0,1} x_{0,1}^2 + w_{0,2} x_{0,1} x_{0,2} \\ w_{0,1} x_{0,2} x_{0,1} + w_{0,2} x_{0,2}^2 \end{bmatrix} + \begin{bmatrix} x_{0,1} \\ x_{0,2} \end{bmatrix} \\
&= \begin{bmatrix} w_{0,1} x_{0,1}^2 + w_{0,2} x_{0,1} x_{0,2} + x_{0,1} \\ w_{0,1} x_{0,2} x_{0,1} + w_{0,2} x_{0,2}^2 + x_{0,2} \end{bmatrix}
\end{aligned} \tag{3}$$

继续计算

$$X_2$$

，第二层有一阶、二阶、三阶"交叉特征"：

$$
\begin{aligned}
X_2 &= X_0 X_1^{l} W_1 + X_1 \\
&= \begin{bmatrix} x_{0,1} \\ x_{0,2} \end{bmatrix} \begin{bmatrix} w_{0,1}x_{0,1}^2 + w_{0,2}x_{0,1}x_{0,2} + x_{0,1}, & w_{0,1}x_{0,2}x_{0,1} + w_{0,2}x_{0,2}^2 + x_{0,2} \end{bmatrix} \begin{bmatrix} w_{1,1} \\ w_{1,2} \end{bmatrix} \\
&+ \begin{bmatrix} w_{0,1}x_{0,1}^2 + w_{0,2}x_{0,1}x_{0,2} + x_{0,1} \\ w_{0,1}x_{0,2}x_{0,1} + w_{0,2}x_{0,2}^2 + x_{0,2} \end{bmatrix} \\
&= \begin{bmatrix} w_{0,1}x_{0,1}^3 + w_{0,2}x_{0,1}^2x_{0,2} + x_{0,1}^2, & w_{0,1}x_{0,2}x_{0,1}^2 + w_{0,2}x_{0,2}^2x_{0,1} + x_{0,2}x_{0,1} \\ w_{0,1}x_{0,1}^2x_{0,2} + w_{0,2}x_{0,1}x_{0,2}^2 + x_{0,1}x_{0,2}, & w_{0,1}x_{0,2}^2x_{0,1} + w_{0,2}x_{0,2}^3 + x_{0,2}^2 \end{bmatrix} \begin{bmatrix} w_{1,1} \\ w_{1,2} \end{bmatrix} \\
&+ \begin{bmatrix} w_{0,1}x_{0,1}^2 + w_{0,2}x_{0,1}x_{0,2} + x_{0,1} \\ w_{0,1}x_{0,2}x_{0,1} + w_{0,2}x_{0,2}^2 + x_{0,2} \end{bmatrix} \\
&= \begin{bmatrix} w_{0,1}w_{1,1}{\color{red}x_{0,1}^3} + w_{0,2}w_{1,1}{\color{red}x_{0,1}^2x_{0,2}} + w_{1,1}{\color{red}x_{0,1}^2} + w_{0,1}w_{1,2}{\color{red}x_{0,2}x_{0,1}^2} + w_{0,2}w_{1,2}{\color{red}x_{0,2}^2x_{0,1}} + w_{1,2}{\color{red}x_{0,2}x_{0,1}} \\ w_{0,1}w_{1,1}{\color{red}x_{0,1}^2x_{0,2}} + w_{0,2}w_{1,1}{\color{red}x_{0,1}x_{0,2}^2} + w_{1,1}{\color{red}x_{0,1}x_{0,2}} + w_{0,1}w_{1,2}{\color{red}x_{0,2}^2x_{0,1}} + w_{0,2}w_{1,2}{\color{red}x_{0,2}^3} + w_{1,2}{\color{red}x_{0,2}^2} \end{bmatrix} \\
&+ \begin{bmatrix} w_{0,1}{\color{red}x_{0,1}^2} + w_{0,2}{\color{red}x_{0,1}x_{0,2}} + {\color{red}x_{0,1}} \\ w_{0,1}{\color{red}x_{0,2}x_{0,1}} + w_{0,2}{\color{red}x_{0,2}^2} + {\color{red}x_{0,2}} \end{bmatrix}
\end{aligned} \tag{4}
$$

由以上公式可知，当cross layer叠加l层时，"交叉"最高阶可以达到$l+1$阶，并且包含了所有阶的交叉特征。

总结：Cross端相对于DNN端是非常轻量级的，每层只需要w^l, b^l这两个参数，这两个参数都是向量。所以整个模型的复杂度还是由DNN端主导的。也正是因为Cross端的参数量很少，所以模型capacity不够，这才需要用Deep端来进行模型capacity的补充。

0x02. DCN真的做了特征交叉吗?

到现在为止，一切看起来都很美好。但是转念一想，高阶特征交叉明明是一个**指数级**的操作，而DCN用简单的矩阵乘就"做到了"，这会不会太过廉价了?

考虑只有一层的DCN，我们能让它去还原出FM的形式吗?

还是令

$$X_0 = \begin{bmatrix} x_{0,1} \\ x_{0,2} \end{bmatrix}$$

，那么

$$x_0 x_0^T w = \begin{bmatrix} x_{0,1} \\ x_{0,2} \end{bmatrix} \begin{bmatrix} x_{0,1}, x_{0,2} \end{bmatrix} w = \begin{bmatrix} x_{0,1}^2, x_{0,1}x_{0,2} \\ x_{0,1}x_{0,2}, x_{0,2}^2 \end{bmatrix} w$$

要是按照FM中的交叉特征定义（即内积），必须拿出上面 $x_0 x_0^T$ 矩阵的上三角或者下三角（不包含对角线）的所有元素加到一起才可以，这时候怎么解出 w 呢？我发现不论如何设计 w 其实都是做不到的（本质原因还是因为w是个向量，势单力薄）。所以，这里的交叉最后和我们见到的FM以及类似模型中的交叉已经不是一个东西了!

当交叉变成了embedding中**元素的乘法**，而不是原来**整个embedding合起来内积**，就不存在embedding泛化性的保证了，那么交叉的意义还剩下多大呢?

对于上面的问题，xDeepFM中相当于是做了一个归纳：**DCN的本质实际上是给 x_0 乘了一个系数**!

结合上面的图， x^l 和w乘起来就是一个数字，也就是说，最后一层一层迭代完了，只得到一个 x_0 的倍数。你要说没有交叉吧，系数其实还是和 x_0 有关系的，你要说有交叉吧，又不是我们FM，PNN，ONN等等网络中讲得这么回事。

这么看下来，DCN给我们的大体上是一个空头支票。它的交叉也不一定是我们想要的交叉。

参考：https://zhuanlan.zhihu.com/p/422368322