算法大佬看了流泪，为什么这么好的CTR预估总结之前没分享(下篇)

原创 Eric陈健锋 炼丹笔记 1周前

收录于话题

#搜索推荐前沿算法 46 #搜索推荐基础知识 14

↑↑↑关注后"星标"**炼丹笔记**

炼丹笔记干货

作者：特邀资深炼丹师Eric陈健锋

排版：十方导语

导语

在广告、推荐系统CTR预估问题上，早期的完全规则方法被过渡到以LR为代表的机器学习方法，为了充分发挥组合特征的价值，在相当长一段时间里，业界热衷于使用LR+人工特征工程。但人工组合特征成本高昂，在不同任务上也难以复用。2010年FM因子分解方法的出现解决了人工组合特征的困境，2014年Facebook提出的GBDT+LR也给出了一种利用树模型特点构建组合特征的思路。不过随着深度学习的崛起，2015年以后，借助非线性自动组合特征能力的深度模型，开始成为业内的主流。从经典DNN到结合浅层的Wide&Deep，用于CTR预估的深度模型在近些年间百花盛开，各种交叉特征建模方法层出不穷，Attention机制也从其他研究领域引入，帮助更好的适应业务，提升模型的解释性。在这进化路线之下，核心问题离不开解决数据高维稀疏难题，自动化组合特征，模型可解释。我们梳理了近些年CTR预估问题中有代表性的模型研究/应用成果，并对部分经典模型的实现原理进行详细剖析，落成文字作为学习过程的记录。

目录

04 引入注意力机制

直观上，注意力机制可以借用人类的视觉行为来进行解释，例如我们观看一张照片的时候，视觉系统会很自然的将注意焦点集中在某些区域，而自动忽略其他不相关的内容。这种机制被引入到机器学习/深度学习，其中一个著名事件是，2014年Bengio将其提出应用于机器翻译领域[17]，取得SOTA结果。正如上图所示的，横轴是英文原句，纵轴是法文翻译句，像素颜色越浅表示翻译某单词时对原句相关单词的注意力权重越大，机器翻译时的注意力权重分布是和人类的直觉认知相符的。相比黑盒模型，注意力机制增强了可解释性，并且这些注意力权重是在翻译时根据输入动态计算的，模型的自适应能力也获得提升。基于这

些优点，注意力机制在广告/推荐系统的排序模型上也有很多的探索和应用，下文列举一些比较经典的工作。

4.1 AFM模型

首先，承接上文未展开的AFM模型（2017）[15]，模型结构图如上，符号“⊙”表示向量的哈达玛积，不难看出模型实现的也是一种向量级vector-wise的特征交互。相比NFM模型将Bi-Interaction Pooling层两两向量交互后的结果直接求和压缩成一个向量，AFM引入了attention权重因子来对交互后的向量进行加权求和，以增强不同特征交互对的重要度区分：

NFM Bi-Interaction Pooling:

$$f_{BI}(\mathcal{V}_x) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} x_i \mathbf{v}_i \odot x_j \mathbf{v}_j,$$

AFM Attention-based Pooling:

$$f_{Att}(f_{PI}(\mathcal{E})) = \sum_{(i,j)\in\mathcal{R}_x} a_{ij}(\mathbf{v}_i \odot \mathbf{v}_j)x_i x_j$$

其中，attention权重因子由Attention Net计算得到，Attention Net的定义为：

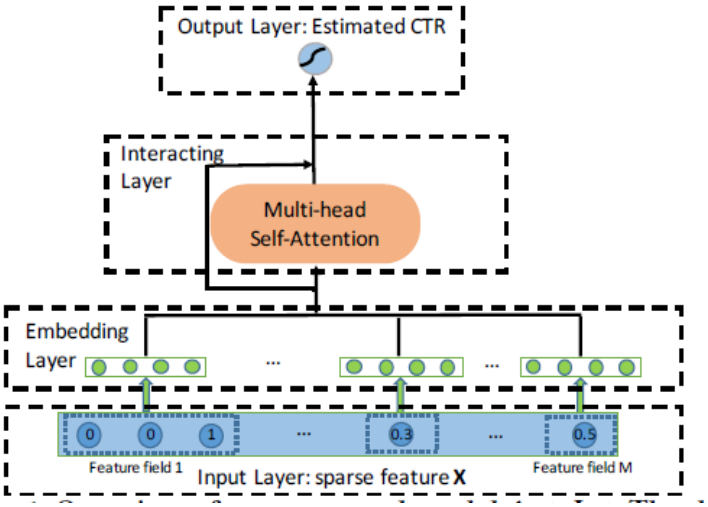
$$a'_{ij} = \mathbf{h}^T ReLU(\mathbf{W}(\mathbf{v}_i \odot \mathbf{v}_j)x_i x_j + \mathbf{b})$$
$$a_{ij} = \frac{\exp(a'_{ij})}{\sum_{(i,j)\in\mathcal{R}_x} \exp(a'_{ij})}$$

最后，模型的表达式：

$$\hat{y}_{AFM}(\mathbf{x}) = w_0 + \sum_{i=1}^{n} w_i x_i + \mathbf{p}^T \sum_{i=1}^{n} \sum_{j=i+1}^{n} a_{ij}(\mathbf{v}_i \odot \mathbf{v}_j)x_i x_j,$$

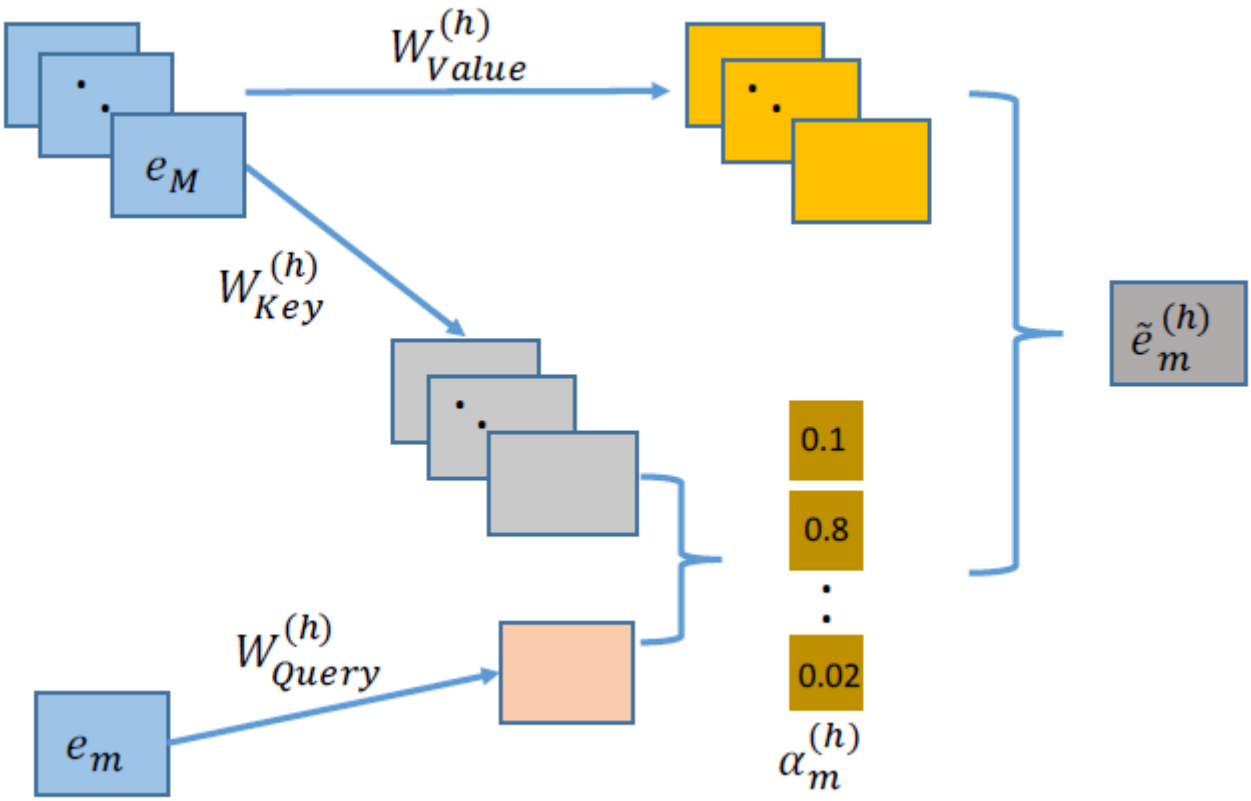
不过，如上文所说，AFM的不足是模型限制在二阶交叉特征的表达上，缺少更高阶信息的建模。

4.2 AutoInt模型



2018年提出的AutoInt模型[18]是一个使用多头自注意力机制[19]增强模型解释性，同时又具备高阶交叉建模能力的模型。模型结构如上图，主要创新点在interacting layer，因此下面我们也只围绕interacting layer讨论，其他模块层不再赘述。下面我们来看interacting layer是怎么利用注意力机制生成有价值的特征组合的。

一层interacting layer的计算过程：



其中，M表示特征field的总数量，e_m表示第m个特征field的embedding向量。所有特征field都将进行两两embedding向量的交互组合，而图中展示的是第m个field与其他各个field进行交互的示例。第m个特征field与第k个特征field（k=1~M）组合的权重由以下的attention因子决定，

$$\alpha_{m,k}^{(h)} = \frac{exp(\psi^{(h)}(e_m, e_k))}{\sum_{l=1}^{M} exp(\psi^{(h)}(e_m, e_l))}$$

$$\psi^{(h)}(e_m, e_k) = \langle W_{Query}^{(h)} e_m, W_{Key}^{(h)} e_k \rangle$$

参数W_{Query}，W_{Key}，W_{Value}表示对原始embedding向量先进行空间R^d到$R^{d'}$的转换，$\psi^{(h)}(\cdot,\cdot)$是attention函数，论文中将这个函数定义为向量点积，即<·，·>。

最后，$\tilde{e}_m^{(h)}$表示在第h个子空间下，第m个特征field经过interacting layer与其他特征field进行交互后的输出，

$$\tilde{e}_m^{(h)} = \sum_{k=1}^{M} \alpha_{m,k}^{(h)} (W_{Value}^{(h)} e_k)$$

这里的第h个子空间，也就是multi-head中的第h个head，使用多个子空间能丰富特征组合的语义表达。同时，由于进行attention交互的query和key都来自同一张量，即[e_1，e_2，…，e_M]，因此interacting layer的核心原理为多头自注意力机制（Multi-head Self-Attention）。

AutoInt将并行的H个head的输出进行拼接，得到一个考虑组合信息后的特征向量$\tilde{e}_m$（combinatorial feature），

$$\tilde{e}_m = \tilde{e}_m^{(1)} \oplus \tilde{e}_m^{(2)} \oplus \cdots \oplus \tilde{e}_m^{(H)}$$

除此之外，AutoInt也加入了其他的trick，例如残差连接，以实现更深的网络层数，

$$e_m^{Res} = ReLU(\tilde{e}_m + W_{Res} e_m)$$

最后，模型表达式如下，

$$\hat{y} = \sigma(w^T(e_1^{Res} \oplus e_2^{Res} \oplus \cdots \oplus e_M^{Res}) + b)$$

其中，e_m^{Res}表示经过若干interacting layer后的第m个考虑组合信息后的特征向量，“⊕”是向量拼接符号。通过级联多层interacting layer，AutoInt可以以显式向量级交互方式实现高阶特征交叉。

附AutoInt论文中的实验结果：

Model Class	Model	Criteo		Avazu		KDD12		MovieLens-1M	
		AUC	Logloss	AUC	Logloss	AUC	Logloss	AUC	Logloss
First-order	LR	0.7820	0.4695	0.7560	0.3964	0.7361	0.1684	0.7716	0.4424
Second-order	FM [23]	0.7836	0.4700	0.7706	0.3856	0.7759	0.1573	0.8252	0.3998
	AFM[36]	0.7938	0.4584	0.7718	0.3854	0.7659	0.1591	0.8227	0.4048
High-order	DeepCrossing [29]	0.8012	0.4513	0.7643	0.3889	0.7715	0.1591	0.8453	0.3814
	NFM [12]	0.7957	0.4562	0.7708	0.3864	0.7515	0.1631	0.8357	0.3883
	CrossNet [34]	0.7907	0.4591	0.7667	0.3868	0.7773	0.1572	0.7968	0.4266
	CIN [16]	0.8009	0.4517	**0.7758**	0.3829	0.7800	0.1566	0.8286	0.4108
	HOFM [4]	0.8005	0.4508	0.7701	0.3854	0.7707	0.1586	0.8304	0.4013
	AutoInt (ours)	**0.8061**	**0.4454**	0.7752	**0.3823**	**0.7881**	**0.1545**	**0.8460**	**0.3784**

Model	Criteo		Avazu		KDD12		MovieLens-1M		Avg. Changes	
	AUC	Logloss	AUC	Logloss	AUC	Logloss	AUC	Logloss	AUC	Logloss
Wide&Deep (LR)	0.8026	0.4494	0.7749	0.3824	0.7549	0.1619	0.8300	0.3976	+0.0292	-0.0213
DeepFM (FM)	0.8066	0.4449	0.7751	0.3829	0.7867	0.1549	0.8437	0.3846	+0.0142	-0.0113
Deep&Cross (CN)	0.8067	0.4447	0.7731	0.3836	0.7869	0.1550	0.8446	0.3809	+0.0199	-0.0164
xDeepFM (CIN)	0.8070	0.4447	0.7768	0.3832	0.7820	0.1560	0.8467	0.3800	+0.0068	-0.0068
AutoInt+ (ours)	**0.8080**	**0.4437**	**0.7771**	**0.3811**	**0.7892**	**0.1544**	**0.8486**	**0.3757**	+0.0019	-0.0014

注：AutoInt+表示AutoInt + DNN (2-layer MLP)

4.3 FiBiNET模型

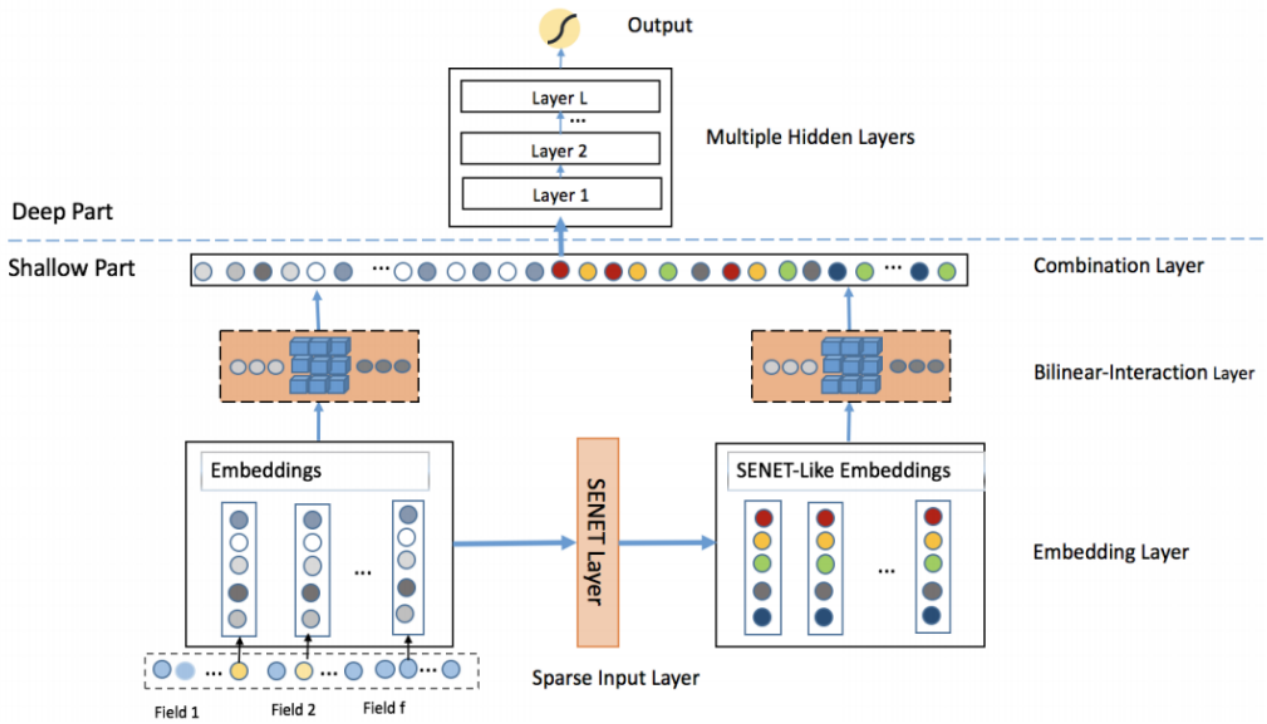


Figure 1: The architecture of our proposed FiBiNET

FiBiNET[20]，是结合特征重要性和双线性特征交互的CTR预估模型，由新浪微博机器学习团队发表在RecSys19。FiBiNET的整体模型结构如上图，相比传统深度学习模型，主要新颖点是加入了SENET

Layer和Bilinear-Interaction Layer。

1）SENET Layer：

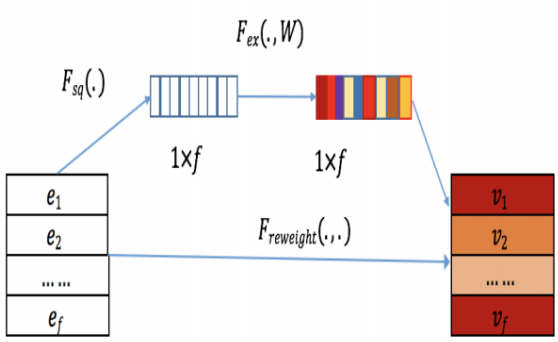


Figure 2: The SENET Layer

SENET Layer的主要作用是学习不同特征的重要度，对不同特征向量进行加权。即该Layer的输入为特征组embedding矩阵 E = [e_1, e_2, ..., e_m]，Layer内部学习出每组特征的重要度 A = [a_1, a_2, ..., a_m]，最后输出的“SENET-Like Embeddings”， V = A * E = [a_1*e_1, a_2*e_2, ..., a_m*e_m]。SENET Layer本质上也是Attention机制，具体包括三个步骤：

① Sequeen，将每个特征组embedding向量e_i压缩成标量z_i，用z_i表示第i个特征组的全局统计信息，m个特征组得到压缩后的统计向量 Z = [z_1, z_2, ..., z_m]。压缩的方法可以是mean-pooling或max-pooling等，不过原文表示mean-pooling效果要优于max-pooling。

$$z_i = F_{sq}(e_i) = \frac{1}{k}\sum_{t=1}^{k} e_i^{(t)}$$

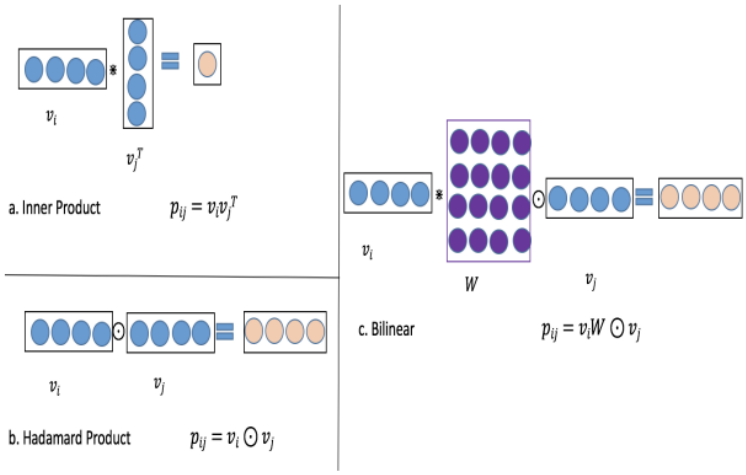
② Excitation，基于压缩后的统计向量Z，学习特征组的重要度权重A，原文使用的是两层神经网络，第一层为维度缩减层，第二层为维度提升层，可形式化表示为：

$$A = F_{ex}(Z) = \sigma_2(W_2\sigma_1(W_1 Z))$$

③ Re-Weight，顾名思义，通过前一步得到的重要度权重A对原始特征组embedding向量进行重新赋权，得到SENET-Like Embeddings：

$$V = F_{ReWeight}(A, E) = [a_1 \cdot e_1, \cdots, a_f \cdot e_f] = [v_1, \cdots, v_f]$$

2）Bilinear-Interaction Layer：



原文作者认为传统的内积或哈达玛积形式难以有效对稀疏数据进行交叉特征建模，因此提出了一种双线性特征交叉方式（这种形式前文介绍双线性FFM时也有提到），如上图c，通过引入额外参数矩阵W，先对vi和W进行内积，再与vj进行哈达玛积，增强模型表达能力来学习特征交叉。如前文提到，引入参数W有3种形式，第一种是所有特征共享一个矩阵W，原文称为Field-All Type；第二种是一个Field一个W_i，称为Field-Each Type；第三种是一个Field组合一个W_{ij}，称为Field-Interaction Type。

回到FiBiNet模型整体结构，原始特征组embedding向量p和经过SENET层输出的embedding向量q，经过Combination Layer，得到拼接后的向量c：

$$c = F_{concat}(p, q) = [p_1, \cdots, p_n, q_1, \cdots, q_n] = [c_1, \cdots, c_{2n}]$$

原文实验了两种应用方式，一种是直接对向量c的元素求和并经过sigmoid输出，得到一个浅层CTR预估模型。另一种是对向量c后接一个DNN深度神经网络，得到一个深层CTR预估模型。实验结果表明，同浅层/深层模型下，FiBiNET要优于其他模型，并且深层模型要优于浅层模型。

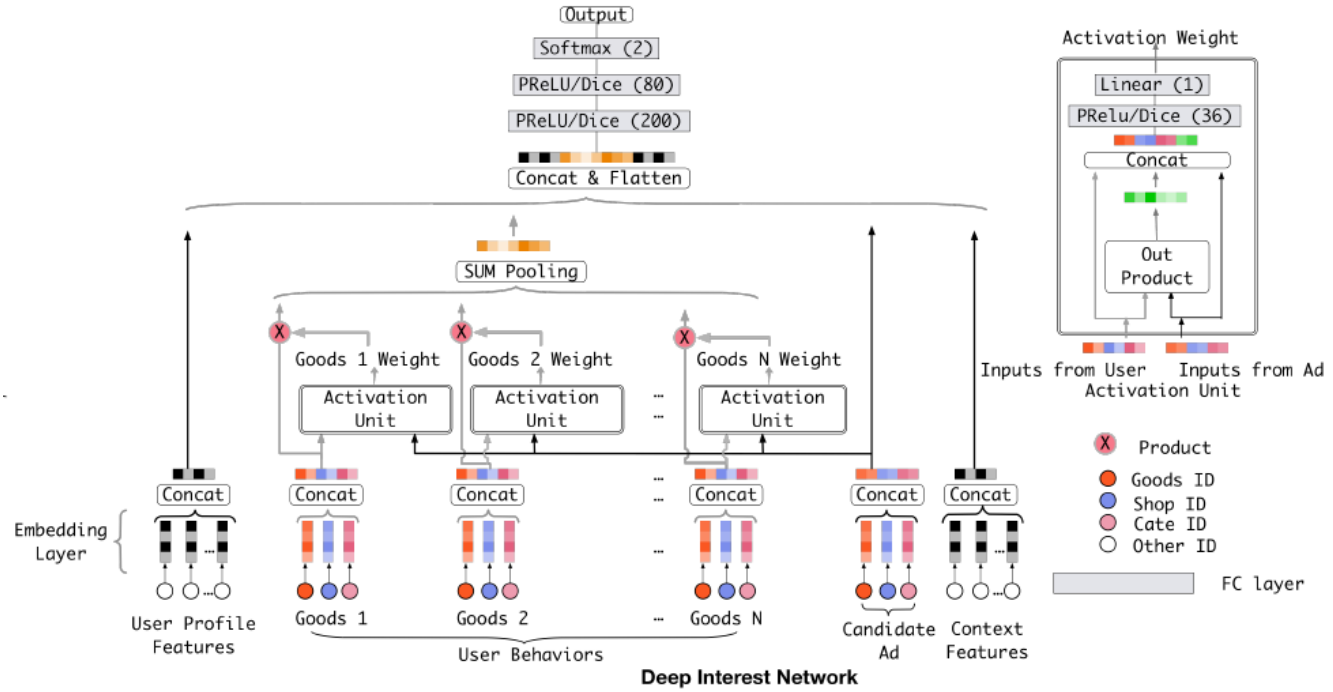
Table 1: The overall performance of shallow models on Criteo and Avazu datasets. The SE-FM-ALL denotes the shallow model with the Field-All type of Bilinear-Interaction layer.

	Criteo		Avazu	
Model	AUC	Logloss	AUC	Logloss
LR	0.7808	0.4681	0.7633	0.3891
FM	0.7923	0.4584	0.7745	0.3832
FFM	0.8001	0.4525	0.7795	0.3810
AFM	0.7965	0.4541	0.7740	0.3839
SE-FM-All	**0.8021**	**0.4495**	**0.7803**	**0.3800**

4.4 DIN模型

Deep Interest Network（DIN）[21]是阿里妈妈广告算法团队在2017年提出的，DIN是一个工业应用性很强的方案，特别是在电商领域。在介绍模型创新之前，我们先简单回顾一下DIN的研究动机。在线上广告系统或者其他的推荐场景，为了提升用户体验，同时最大化平台方的流量价值，要求我们的排序模型能精准识别用户兴趣，充分利用用户兴趣信息，对<user, ad/item>的匹配度进行打分。对于在线应用，用户兴趣主要来源于对用户行为数据的刻画，由多种多样的行为数据所表达出来的用户兴趣也应该具有多样性。但这里的一个挑战是，传统DNN模型会将这些用户行为embedding向量通过固定的形式（sum/max/mean等）pooling成一个定长向量作为用户兴趣表示，所有用户的兴趣表示都被映射在一个相同的固定空间，而由于空间维度（即向量维度）的限制，用户兴趣的多样性特点无法得到充足的表达。

举个例子，一个用户在电商网站上的浏览历史中，有30%是美妆相关，30%育儿相关，20%运动健身相关，剩下20%是其他，直观地看，这个用户的兴趣多样性分布是很明显的，但如果我们将这些历史行为向量（embedding vector）用上述的形式pooling成一个向量，恐怕得到的兴趣表示就很模糊，谁都不好解释，甚至模型也难以理解。



而DIN对这个问题的解决方案是“局部激活”（Local Activation），即根据候选广告来自适应的为历史行为向量分配激活权重，以attention-based pooling的方式将其合并作为用户在该候选广告下的兴趣表示，使得与候选广告更相关的历史行为在用户兴趣表示中占更主导的作用。

上面这段话可用如下数学形式进行表示，用 $\boldsymbol{v_U}$ 表示用户的兴趣向量，在候选广告A下，用户的兴趣向量 $\boldsymbol{v_U}(A)$ 的表达式为：

$$\boldsymbol{v}_U(A) = f(\boldsymbol{v}_A, \boldsymbol{e}_1, \boldsymbol{e}_2, .., \boldsymbol{e}_H) = \sum_{j=1}^{H} a(\boldsymbol{e}_j, \boldsymbol{v}_A)\boldsymbol{e}_j = \sum_{j=1}^{H} w_j \boldsymbol{e}_j$$

其中，$\boldsymbol{v_A}$ 是候选广告A的向量表示，$\boldsymbol{e_1}, \boldsymbol{e_2}, .., \boldsymbol{e_H}$ 是用户历史行为向量表示，a(·, ·)是attention函数，wj是最后计算得到的行为向量激活权重。

从特征组合的角度来看，DIN实际上是在建模广告与用户历史行为的交叉特征信息，而“局部激活”的背后原理也正是注意力机制。

除了“局部激活”这一重杀器，DIN论文中还给出了一些非常实用的深度模型技巧，例如，批感知的正则化方法（Mini-batch Aware Regularization）、数据自适应的激活函数（Dice），不过限于篇幅，这里不再展开介绍。

下表是DIN在阿里电商广告数据集中的离线实验结果，DIN模型AUC绝对值比Base模型（Embedding&MLP）高0.0059，同时也要优于Wide&Deep、PNN、DeepFM等模型。作者同时也给了

在阿里进行的线上A/B testing实验结果，和Base模型相比，DIN*获得了10% CTR和3.8% RPM (Revenue Per Mille)的提升。

* with MBA Reg. and Dice

Model	AUC	RelaImpr
LR	0.5738	- 23.92%
BaseModel[a,b]	0.5970	0.00%
Wide&Deep[a,b]	0.5977	0.72%
PNN[a,b]	0.5983	1.34%
DeepFM[a,b]	0.5993	2.37%
DIN Model[a,b]	**0.6029**	**6.08%**
DIN with MBA Reg.[a]	**0.6060**	**9.28%**
DIN with Dice [b]	**0.6044**	**7.63%**
DIN with MBA Reg. and Dice	**0.6083**	**11.65%**

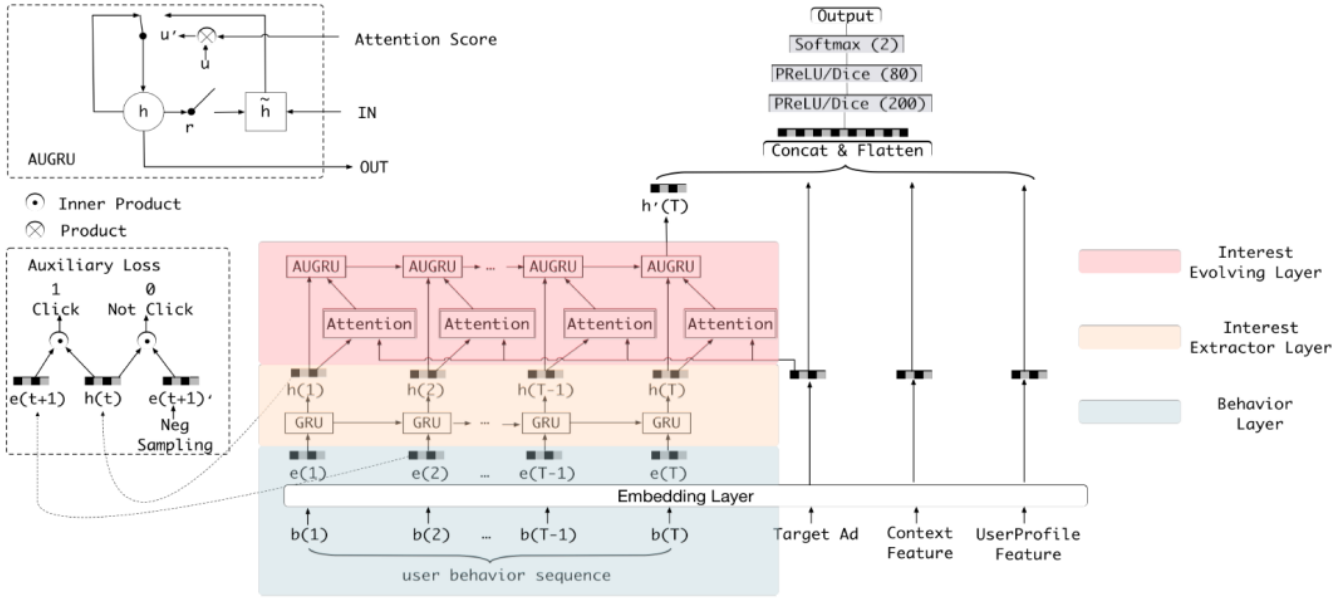
[a] These lines are trained with PReLU as the activation function.
[b] These lines are trained with dropout regularization.

$$\text{AUC} = \frac{\sum_{i=1}^{n} \#impression_i \times \text{AUC}_i}{\sum_{i=1}^{n} \#impression_i} \quad RelaImpr = \left(\frac{\text{AUC(measured model)} - 0.5}{\text{AUC(base model)} - 0.5} - 1 \right) \times 100\%$$

注：gAUC（用户级别的group AUC），好处是可兼顾用户自身内部的广告排序优劣，消除用户偏差

从DIN论文实验的可视化结果来看，和候选广告相关性高的历史行为，确实得到了更高的激活权重，模型获得很好的可解释性。

4.5 DIEN模型



DIEN，全称Deep Interest Evolution Network[22]，也是来自阿里DIN的提出团队，是DIN的进化版，旨在挖掘用户行为背后的更高抽象语义的用户兴趣演化规律。这个模型我们不再阐述细节，概括DIEN的两个主要创新点：

1. 兴趣抽取层（Interest Extractor Layer），利用GRU序列建模，从用户历史行为中抽取兴趣状态表示h(t)，并且引入辅助loss利用下一时刻的行为信息帮助指导当前时刻h(t)的学习，更准确的建模用户兴趣向量。

2. 兴趣演化层（Interest Evolving Layer），引入AUGRU（GRU with attentional update gate）建模用户兴趣的演化过程，在AUGRU的隐状态更新中引入目标广告和兴趣向量的attention因子，目的是使最终输出给DNN网络的兴趣终态h'(T)和目标广告有更高的相关度。

最后，在论文的生产环境数据集上，DIEN模型的离线AUC和线上A/B testing实验结果，均优于作者较早之前提出的DIN模型。DIEN作为序列模型结合注意力机制建模用户兴趣演化过程，优化业务问题的创造性工作，值得学习和借鉴。

Model	AUC
BaseModel (Zhou et al. 2018c)	0.6350
Wide&Deep (Cheng et al. 2016)	0.6362
PNN (Qu et al. 2016)	0.6353
DIN (Zhou et al. 2018c)	0.6428
Two layer GRU Attention	0.6457
BaseModel + GRU + AUGRU	0.6493
DIEN	**0.6541**

Model	CTR Gain	PPC Gain	eCPM Gain
BaseModel	0%	0%	0%
DIN (Zhou et al. 2018c)	+ 8.9%	- 2.0%	+ 6.7%
DIEN	+ 20.7%	- 3.0%	+ 17.1%

05 总结

本文第一部分主要介绍CTR预估问题中的线性模型的演变，演变的主要思路是以因子分解的方法解决高维稀疏数据问题，以更有效的实现低阶交叉特征。第二部分介绍了非线性模型GBDT+LR，它实现了自动化特征工程。第三、第四部分属于深度模型的范畴，贯穿其中的是一个基本的DNN（Deep）框架，即sparse input layer -> embedding layer -> interaction layer -> output layer，其中interaction layer被用于以各种形式构建高阶交叉特征的信息表达。而Wide&Deep的提出，弥补了Deep对低阶信息利用的缺失，同时也丰富了模型架构的可能性。除了Deep部分的interaction layer可被不同形式地改进，Wide部分也被不断探索和优化，例如以FM、CrossNet替代LR，甚至引入一些新的子网络，例如CIN。在这些基础上，注意力机制的引入，也使模型能够自主充分地挖掘特征的重要性，有更好的自适应能力以及可解释性。

看完这篇是不是没过瘾，是不是心中还有问号？是不是想直接找Eric当面沟通交流？现在直接给你个机会和Eric大佬一起并肩作战你要不要？鹅厂招人啦！还是腾讯核心的排序组，对腾讯广告排序框架、Auction机制设计以及排序推荐策略算法优化感兴趣的小伙伴（T2/T3 and 以上），可邮简历至作者 eric77ch@163.com，帮忙内推直达！

参考文献

[1] http://www.cs.cmu.edu/~wcohen/10-605/2015-guest-lecture/FM.pdf

[2] https://www.csie.ntu.edu.tw/~b97053/paper/Rendle2010FM.pdf

[3] https://www.csie.ntu.edu.tw/~cjlin/papers/ffm.pdf

[4] 张俊林：FFM及DeepFFM模型在推荐系统的探索

[5] 美团技术团队：深入FFM原理与实践

[6] Practical Lessons from Predicting Clicks on Ads at Facebook, KDD 2014

[7] Efficient Estimation of Word Representations in Vector Space, ICLR Workshop 2013

[8] Deep Learning over Multi-field Categorical Data: A Case Study on User Response Prediction, ECIR 2016