【CTR】DIN：阿里深度兴趣网络

贪心科技 阿泽的学习笔记 6月24日

作者：阿泽

1.背景

今天我们学习下阿里的深度兴趣网络（Deep Interest Network，以下简称DIN），目前很多应用在推荐系统中的模型，都是以 Embedding & MLP 的方法结合，这种方法相对传统机器学习有较好的效果提升，但是在其还是存在一些缺点：

- 用户的兴趣通常是多种多样的，而 Embedding & MLP 方法中有限的向量维度会成为用户多样化兴趣的瓶颈，如果扩大向量维度会极大地增加学习参数和计算负荷，并增加过拟合风险；
- 不需要将用户的所有兴趣都压缩到同一个向量中。比如说：用户购买了泳镜并不是因为上周购买了鞋子，而是因为之前购买了泳衣；

针对这些问题，DIN 模型通过考虑给定候选广告的历史行为的相关性，自适应地计算用户兴趣的表示向量。通过引入局部激活单元，DIN 模型通过软搜索历史行为的相关部分来关注相关的用户兴趣，并采用加权总和池化来获取有关候选广告的用户兴趣的表示形式。与候选广告相关性更高的行为会获得更高的激活权重，并且支配着用户兴趣。这样用户 的兴趣表示向量就会随着广告的不同而变化，从而提高了模型在有限尺寸下的表达能力，并使得模型能够更好地捕获用户的不同兴趣。

- 训练具有大规模稀疏特征网络具有非常大的挑战，例如：基于 SGD 的优化方法可以采用 Mini-Batch 来更新参数，但加上 L2 正则化后其计算量会非常大，因为每个 Mini-Batch 都需要计算所有参数的 L2 范式；

作者提出了一个新颖的 Mini-Batch 感知正则化方法，可以只计算非零特征参数的 L2 范式；此外，作者还考虑数据输入分布，从而设计邻一个能够自适应数据的激活函数。

2.网络结构

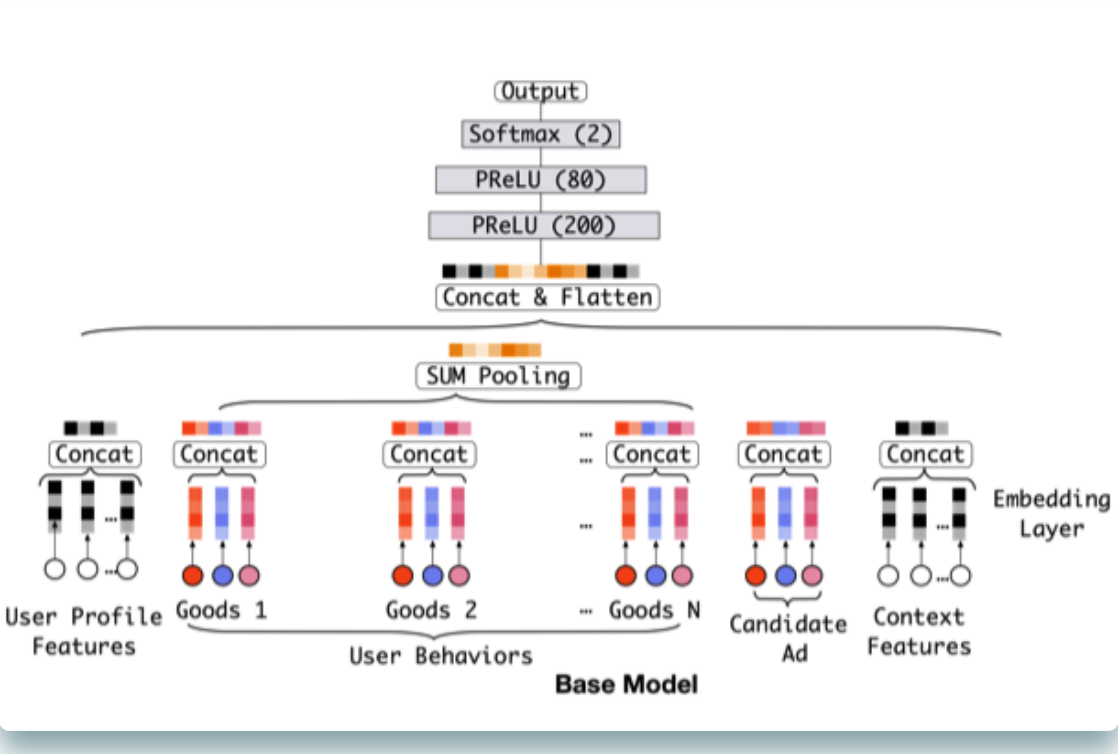
2.1 特征构建

下图为阿里构建的特征，主要为四个方面：用户特征、用户行为特征、广告特征、背景特征。

Category	Feature Group	Dimemsionality	Type	#Nonzero Ids per Instance
User Profile Features	gender	2	one-hot	1
	age_level	~ 10	one-hot	1
	...	...	...	...
User Behavior Features	visited_goods_ids	~ 10^9	multi-hot	~ 10^3
	visited_shop_ids	~ 10^7	multi-hot	~ 10^3
	visited_cate_ids	~ 10^4	multi-hot	~ 10^2
Ad Features	goods_id	~ 10^7	one-hot	1
	shop_id	~ 10^5	one-hot	1
	cate_id	~ 10^4	one-hot	1
	...	...	...	...
Context Features	pid	~ 10	one-hot	1
	time	~ 10	one-hot	1
	...	...	...	...

2.2 Base Model

我们先来看 Base Model，其网络结构主要由 Embedding 和 MLP 构成，如下图所示。



Embedding 层： 输入的是高维二值化的稀疏向量，输出是低维的高密度向量；

Pooling 层： 由于不同用户会有不同数量的行为，所以该层输入不同数量的 Embedding 向量，输出为固定大小的向量；

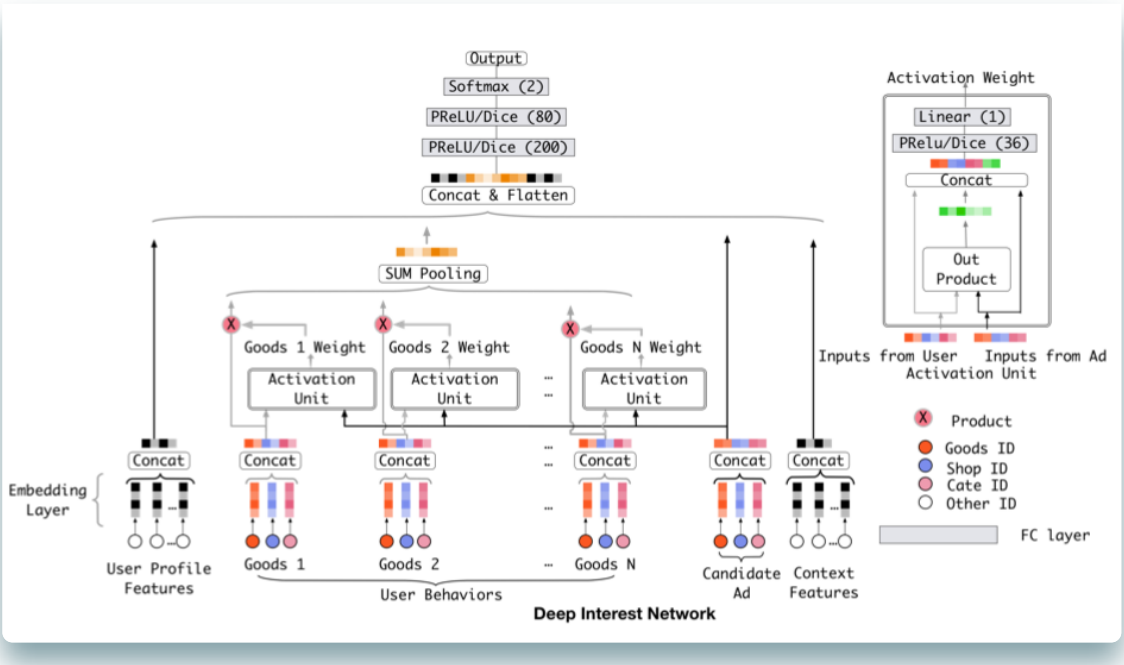
MLP： 采用激活单元为 PReLU 的全连接网络；

损失函数： 交叉熵损失函数。

Base Model 有一个非常大的缺点——**用户兴趣多样化表达受限**。Base Model 直接通过池化所有的 Embedding 向量来获的一个定长的表示向量。由于表示向量长度固定，所以其表达也会受到一定的限制，假设最多可以表示 k 个独立的兴趣爱好，如果用户兴趣广泛不止 k 个则表达受限[2]，极大的限制了用户兴趣多样化的表达，而如果扩大向量维度则会带来巨大的计算量和过拟合的风险；

2.3 DIN Model

基于 DIN 的网络结构如下图所示，作者认为用户行为在电商应用场景中至关重要，所以在建模时更加关注于用户行为。



DIN 针对 Base Model 的缺点提出了**局部激活单元**，其目的在于：**在有限的特征空间中表达用户复杂的兴趣。** DIN 将用户兴趣刻画为用户分布表示，而不是固定的一个点，这样即使在 k 维空间也可以获得超过 k 维的表达能力。

电商的用户行为特点往往是多需求并发（Diversity）的，行为序列是多个需求子序列的并集，而当用户注意到某个商品时，其决定通常只与其一个或者部分需求有关[2]。

在阿里的广告系统中，当用户点击商品，候选广告将通过软搜索用户的历史行为并挖掘其最近浏览过的类似商品，从而满足用户当前的相关兴趣。也就是说 DIN 是通过考虑用户的当前点击行为与历史行为的相关性来自适应地计算用户兴趣的表示向量，用户的历史行为的权重依赖于正在看的商品。

DIN 网络根据目标商品，反向激活和过滤历史行为来自适应计算用户的表示向量，公式如下：

$$v_U(A) = f(V_A, e_1, e_2, \ldots, e_H) = \sum_{j=1}^{H} a(e_j, v_A)e_j = \sum_{j=1}^{H} w_j e_j$$

其中，$v_U(A)$为用户 U 所有的广告 A 的 Embedding 向量，e_j为用户历史行为的 Embedding 向量。$a(\cdot)$是前馈神经网络，输出为激活权重。

NLP 中的注意力机制需要对输出进行归一化操作，而这里放宽了 $\sum w_i = 1$ 的约束。目的是保留用户兴趣的强度。举个例子：一个用户的历史行为 90% 为衣服，10% 为电子产品，对于 T 恤和手机两个候选广告，T 恤会激活大多数属于衣服的历史行为，并且可以获得更高的兴趣强度。

上图右侧部分为**局部激活单元（Local Activation Unit）**，除了两个输入的向量外，还将他们的点乘也加入到网络中，主要通过显示知识帮助进行关联建模（简单的多层全连接无法学出内积，相当于做了一次特征工程）。

下图为 DIN 中的自适应激活的示意图， 与候选广告高度相关的行为具有较高的激活权重。

3.训练技巧

在工业界，特别是商品用户都过亿的情况下，训练有大量的稀疏输入特征的是一个巨大的挑战。接下来主要介绍两个重要的技术用来加速训练。

3.1 Mini-Batch 感知正则化

再上一节我们看到阿里的特征构建中，直接使用了上亿维度的商品 id、用户 id、广告 id 等，这些大规模的稀疏特征和数亿个参数无法直接应用传统的 L2 正则化。以 L2 正则化为例，每次 Mini-Bacth 都需要计算整个参数的 L2 范式，这将导致计算量极大增加，并使得训练速度急速下降。

针对这个问题，该文引入了一个有效的 Mini-Batch 感知算法。

回顾 DIN 网络模型结构，我们可以看到大部分的参数源于 Embedding 层。所以我们另 $W \in R^{D \times K}$ 表示整个嵌入字典的参数，其中 D 为嵌入向量的维数，K 为特征空间的维度。W 的 L2 正则化公式为：

$$L_2(W) = ||W||_2^2 = \sum_{j=1}^{K} ||w_j||_2^2 = \sum_{(x,y) \in S} \sum_{j=1}^{K} \frac{I(x_j \neq 0)}{n_j} ||w_j||_2^2$$

其中，$w_j \in R^D$ 表示第 j 个embedding 向量；$I(x_j \neq 0)$ 为示性函数，表示样本 x 的第 j 个特征是否为 0；n_j 表示特征 j 出现的次数。上式可以改写为：

$$L_2(W) = \sum_{j=1}^{K} \sum_{m=1}^{B} \sum_{(x,y) \in B_m} \frac{I(x_j \neq 0)}{n_j} ||w_j||_2^2$$

其中 B 为 Mini-Batch 的次数，B_m 表示第 m 次 Mini-Batch。我们令 $a_{mj} = max_{(x,y) \in B_m} I(x_j \neq 0)$ 表示第 m 批次中含有特征 j 的样本至少出现过一次。我们继续改写，用最大值来代替求和，用速度换取精度：

$$L_2(W) \approx \sum_{j=1}^{K} \sum_{m=1}^{B} \frac{a_{mj}}{n_j} ||w_j||_2^2$$

此时，我们便得到了一个 L2 正则化的近似计算解。下式为参数更新：

$$w_j \leftarrow w_j - \eta [\frac{1}{|B_m|} \sum_{(x,y) \in B_m} \frac{\partial L(p(x), y)}{\partial w_j} + \lambda \frac{a_{mj}}{n_j} w_j]$$

其中，n_i 表示第 i 维度特征的非零频次，区别于传统的梯度更新方式：

- 过滤频次为 0 的特征，仅计算出现在第 m 次 Mini-Batch 中的特征参数；
- 考虑特征出现的频率，频次越高，单次正则压制越小；频次越低，单次正则压制越大。

3.2 数据自适应激活函数

PReLU 是一个常用的激活函数：

$$f(s) = \begin{cases} s & if\ s > 0 \\ \alpha s & if\ s \leq 0 \end{cases} = p(s) \cdot s + (1 - p(s)) \cdot \alpha s$$

其中 s 是激活函数 $f(\cdot)$的输入的一维，而 $p(s) = I(s > 0)$ 是示性函数，α 是学习率。我们称 $p(s)$ 为控制函数。

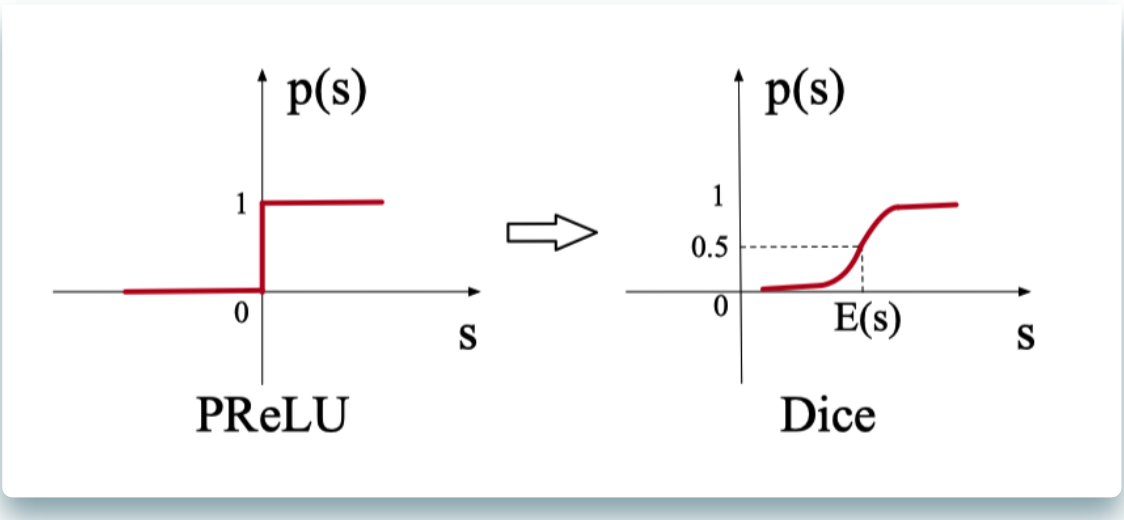
由于每层的输入遵循不同的分布，直接把参数设为 0 不太合适，所以我们设计了一个新的激活函数 Dice：

$$f(s) = p(s) \cdot s + (1 - p(s)) \cdot \alpha s, \quad p(s) = \frac{1}{1 + exp(-\frac{s - E[s]}{\sqrt{Var[s] + \epsilon}})}$$

Dice 的关键思想是根据输入数据的分布自适应地进行调整控制函数，使得其值为输入的均值。$E(s)$ 和 $Var(s)$ 的更新方式如下：

$$E[s]_{t+1}{}' = E[s]_t{}' + \alpha E[s]_{t+1} Var[s]_{t+1}{}' = Var[s]_t{}' + \alpha Var[s]_{t+1}$$

当 $E(s)$ 和 $Var(s)$ 都为 0 时，退化为 PReLU。两种控制函数下的激活函数的对比如下图所示：



3.3 度量

在 CTR 预测领域，AUC是一个广泛常用的的度量方式。这里引入了一种基于用户加权的 AUC 计算方式：

$$GAUC = \frac{\sum_{i=1}^{n} \#impressions_i \times AUC_i}{\sum_{i=1}^{n} \#impressions_i}$$

其中 n 为用户数量，$\#impression_i$ 和 AUC_i 分别表示第 i 个用户的展示数量和 AUC 值。

这样做的好处就是消除不同用户的偏差对模型的影响，更能反应出广告系统的在线性能。

4.实验

简单看一下实验部分

首先是在 Amazon 数据集和 Movie-Lens 数据集上不同模型的表现：

Model	MovieLens.		Amazon(Electro).	
	AUC	RelaImpr	AUC	RelaImpr
LR	0.7263	-1.61%	0.7742	-24.34%
BaseModel	0.7300	0.00%	0.8624	0.00%
Wide&Deep	0.7304	0.17%	0.8637	0.36%
PNN	0.7321	0.91%	0.8679	1.52%
DeepFM	0.7324	1.04%	0.8683	1.63%
DIN	**0.7337**	**1.61%**	**0.8818**	**5.35%**
DIN with Dice[a]	**0.7348**	**2.09%**	**0.8871**	**6.82%**

[a] Other lines except LR use PReLU as activation function.

然后是不同规则下的表现：

Regularization	AUC	RelaImpr
Without goods_ids feature and Reg.	0.5940	0.00%
With goods_ids feature without Reg.	0.5959	2.02%
With goods_ids feature and Dropout Reg.	0.5970	3.19%
With goods_ids feature and Filter Reg.	0.5983	4.57%
With goods_ids feature and Difacto Reg.	0.5954	1.49%
With goods_ids feature and MBA. Reg.	**0.6031**	**9.68%**

在阿里巴巴数据集中的表现：

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.

DIN 的可视化，可以看到模型能够捕捉用户的相关历史行为：

5.总结

- 提出了 Deep Interest Network（DIN）模型，将 NLP 中的注意力机制引入到 CTR 预估中，通过设计局部激活单元来自适应学习用户的兴趣；
- 设计出小批量感知正则化和自适应激活函数，以便与大规模稀疏数据的工业训练；
- 引入基于用户加权的 GAUC 计算方式来代替传统的 AUC 计算。

6.引用

1. 《Deep Interest Network for Click-Through Rate Prediction》
2. 《互联网数据下的模型探索 - 盖坤》

喜欢此内容的人还喜欢