【论文导读】2018阿里CTR预估模型---DIN（深度兴趣网络），后附TF2.0复现代码

原创 潜心学习的小透明 推荐算法的小齿轮 6月1日

收录于话题
#推荐系统论文与复现

21个

点击上方“潜心的Python小屋”关注我们，第一时间推送优质文章。

前言

大家好，我是潜心。今天分享一篇最近看的，阿里2018在KDD上发表的论文《Deep Interest Network for Click-Through Rate Prediction》。文章的核心就是使用一个局部激活单元（类似Attention机制）来提高与候选广告相关的历史信息的权重。当然文章还提到了两个在工业深度网络上的技术。总体来说这是一篇偏向工程方面的论文，非常值得一读。本文文末还附上了自己用TF2.0复现的代码，有详细注释【官方给的开源代码难以读懂切为TF1.4版本】。

本文约5k字，预计阅读15分钟。

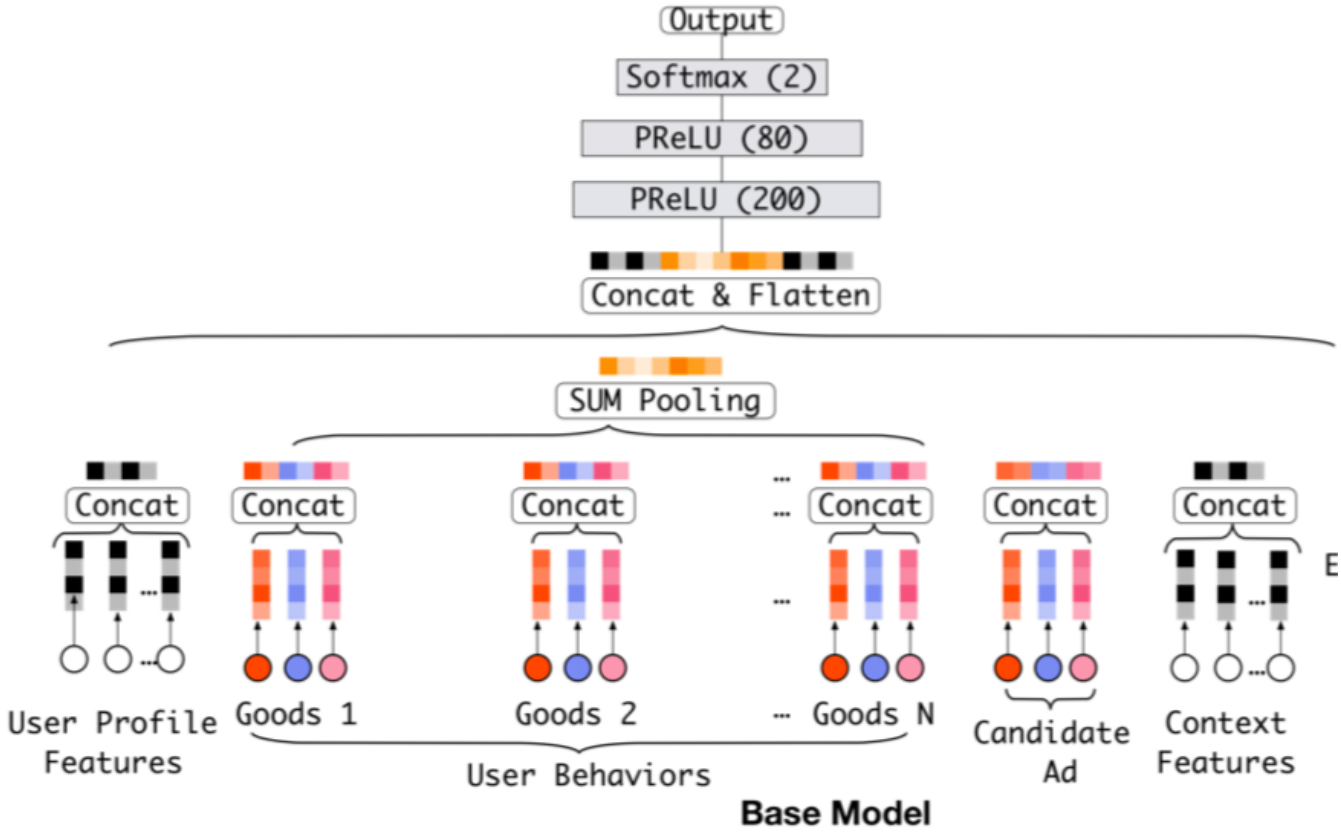
摘要与引言

背景

在工业领域例如在线广告上点击率（Click-through rate，CTR）预测是一个很重要的任务。在每次点击费用（cost-per-click，CPC）的广告系统中，广告按有效价格，即每千个有效成本（effective cost per mille，eCPM）排名，该价格是出价与CTR的乘积，而点击率则需要通过系统预估。因此，CTR预估模型的效果直接影响最终收益，并在广告系统中发挥关键作用。

现状（指2018之前）

如今，CTR预估广告有一个深度学习的Base Model，即**Embedding + MLP**。大规模稀疏输入特征首先被映射到低维Embedding向量中，然后转换为固定长度向量【sum pooling/avg pooling】，最后被连接在一起【concatenate】输入到MLP以学习特征之间的非线性关系。对比很多的逻辑回归模型，确实减少了很多的特征工程和增强模型的能力。常见的模型有Wide&Deep[1]、 DCN[2]、PNN[3]、DeepFM[4]等。



瓶颈

Embedding&MLP模型的瓶颈就是表达用户多样的兴趣。例如，在电子商务中，用户浏览电商网站时可以同时对多个不同的物品产生兴趣。在CTR预估问题中，就是从用户的浏览历史中去捕获用户的兴趣。而该方法不管候选广告是什么，都是将多个特征向量【Embedding】压缩到一个固定长度的表示向量来学习特定用户所有的兴趣表示，这限制了模型的能力，很难从历史行为中提取用户变化的兴趣。

解决这个问题最简单的方法就是扩展向量的维度，但这样会增加学习的参数和在有限的数据中有过拟合的风险。

启发与创新

其实不必将某个用户所有的兴趣【用户的历史购买记录】全部压缩到向量中，因为只有用户部分的兴趣会影响当前行为（对候选广告点击或不点击）。例如，一位女游泳运动员会点击推荐的护目镜，这主要是由于购买了泳衣而不是上周购物清单中的鞋子。

受到上述启发，作者提出了**Deep Interest Network**模型[5]，它通过考虑【给定的候选广告】和【用户的历史行为】的相关性，来计算用户兴趣的表示向量。具体来说就是通过引入局部激活单元，通过软搜索历史行为的相关部分来关注相关的用户兴趣，并采用加权和来获得有关候选广告的用户兴趣的表示。与候选广告相关性较高的行为会获得较高的激活权重，并支配着用户兴趣。该表示向量在不同广告上有所不同，大大提高了模型的表达能力。

提出两个工业技术

训练具有大规模稀疏特征的工业深度网络是巨大的挑战。例如，基于SGD的优化方法仅更新出现在每个小批量中的稀疏特征【非零】的那些参数，因此需要加入正则化来降低过拟合的风险。但是，加上传统的l2正则化后，计算量过大，这需要为每个小型批处理在整个参数上计算L2范数。因此本文提出了一种工业技术：**mini-batch aware regularization**，仅出现在每个微型批处理中的非零特征参数才参与**L2-**范数的计算，从而使计算可接受。

另外还提出**data adaptive activation function**，通过输入的分布自适应调整修正点来推广常用的PReLU。

贡献

概括上述创新点，文章的贡献为：

- 指出使用固定向量来表示用户不同的兴趣的限制性和通过引入局部激活单元建立了一个新的模型DIN。
- 提出两个训练工业神经网络的技术：小批量感知正则化器（a mini-batch aware regularizer），它可以节省具有大量参数的深度网络上正则化的大量计算，并且有助于避免过度拟合；数据自适应激活函数（a data adaptive activation function），它通过考虑输入的分布来概括PReLU，并显示出良好的性能。
- 在对公共和Alibaba数据集进行了广泛的实验。结果证实了提出的DIN的有效性。

Base Model与DIN Model

特征表示（Feature Representation）

工业点击率预测任务中的数据大多采用多组类别的形式，如下图所示，这通常要通过转化为高阶稀疏二元特征【one-hot或者multi-hot】。

$$\underbrace{[0,0,0,0,1,0,0]}_{\text{weekday=Friday}}\quad\underbrace{[0,1]}_{\text{gender=Female}}\quad\underbrace{[0,..,1,...,1,...0]}_{\text{visited\_cate\_ids=\{Bag,Book\}}}\quad\underbrace{[0,..,1,...,0]}_{\text{ad\_cate\_id=Book}}$$

所有的特征为：

Table 1: Statistics of feature sets used in the display advertising system in Alibaba. Features are composed of sparse binary vectors in the group-wise manner.

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
	...	...	...	...

基本模型（Embedding&MLP）

Embedding layer

输入是高维稀疏二元向量，Embedding层将其转化为低维密集表示。

对于第i个特征组t_i，用$W^i = [w_1^i, \dots, w_j^i, \dots, w_{K_i}^i] \in \mathbb{R}^{D \times K_i}$表示第$i$个embedding字典，$K_i$表示第$i$个特征组的维度。$w_j^i \in R^D$是D维度的embedding向量。Embedding操作后伴随着一个表的查找机制：

- 如果t_i是一个one-hot向量$t_i[j] = 1$，则t_i的embedding表示是一个单一的embedding向量$e_i = w_j^i$
- 如果t_i是一个multi-hot向量$t_i[j] = 1, j \in \{i_1, i_2, \dots i_k\}$，那$t_i$的embedding表示是一个embedding向量列表$\{e_{i_1}, e_{i_2}, \dots, e_{i_k}\} = w_{i_1}^i, w_{i_2}^i, \dots, w_{i_k}^i$

Pooling layer 和 Concat layer

不同的用户有不同数量的历史行为，即multi-hot行为特征的向量会导致所产生的embedding向量列表的长度不同，而全连接需要固定长度的输入。一个通用的方法去转化embedding向量列表是通过Pooling层去获得固定长度：

$$\mathbf{e}_i = pooling(\mathbf{e}_{\mathbf{i_1}}, \mathbf{e}_{\mathbf{i_2}}, \dots, \mathbf{e}_{\mathbf{i_k}})$$

两个最常用的池化层是求和池化（sum pooling，各个对应元素进行累加）和平均池化（average pooling，各个对应元素求平均）。

然后将所有向量连接在一起（concatenate），以获得实例的总体表示向量。

MLP

给出连接后的稠密表示向量，利用全连通层自动学习特征的组合。设计MLP的结构，以更好地提取信息。

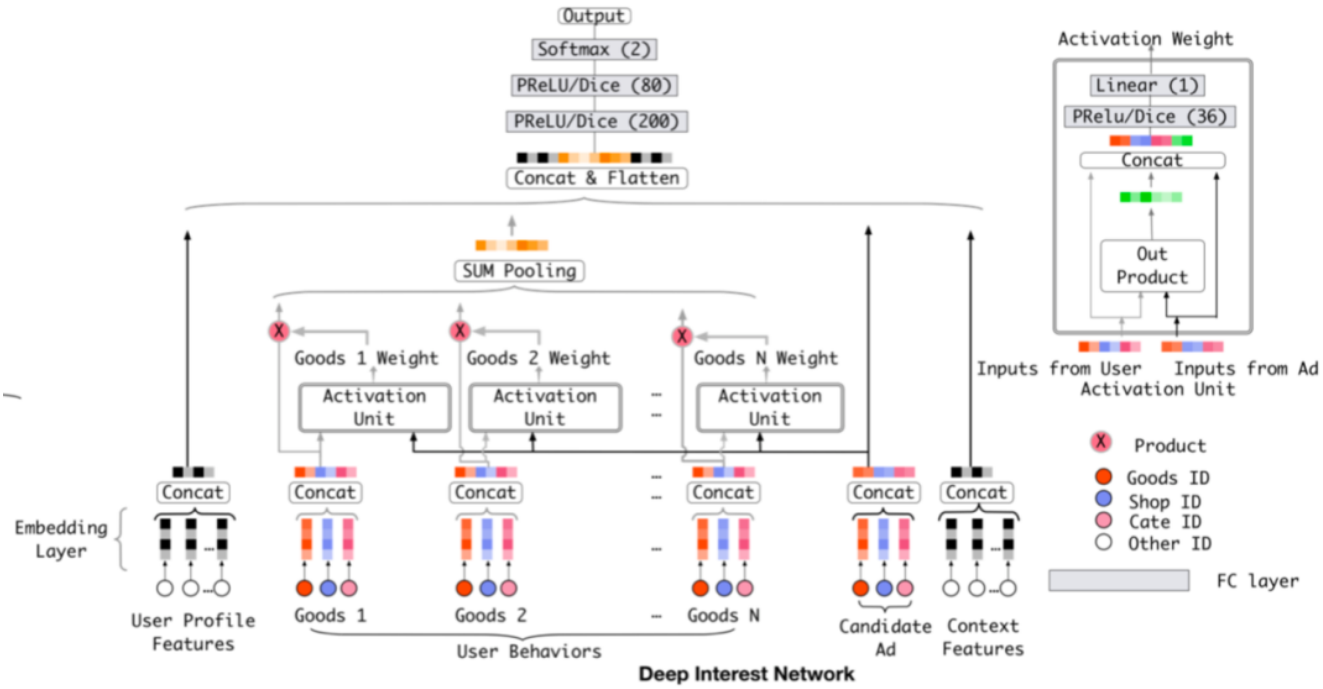
LOSS

负对数似然函数（the negative log-likelihood function）：

$$L = -\frac{1}{N}\sum_{(x,y)\in\mathcal{S}} (y\log p(x) + (1-y)\log(1-p(x)))$$

$\mathcal{S}$代表训练的集合，$p(x)$为模型的输出，即点击候选广告x的概率。

Deep Interest Network



基本模型是获得一个固定长度的用户的表示向量，但不管候选广告是什么，此表示向量对于给定用户均保持不变。这样，维度受限的用户表示向量将成为表达用户多样化兴趣的瓶颈。

与展示广告相关的行为极大地影响了点击操作。**DIN**通过给定一个候选广告，然后去注意与该广告相关的局部兴趣的表示来模拟此过程。DIN不会通过使用同一向量来表达所有用户的不同兴趣，而是通过考

虑历史行为的相关性来自适应地计算用户兴趣的表示向量（对于给定的广告）。该表示向量随不同广告而变化。

相比于基础模型，DIN引入了一种新颖设计的局部激活单元，并保持其他结构不变。具体而言，将激活单元应用于用户行为表示v_U，将其作为加权累加和池来执行，以在给定候选广告A的情况下自适应地计算用户表示：

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

其中$\{e_1, e2, ..eH\}$是用户U长度为H的embedding向量列表，v_A是广告A的embedding向量。$a(\cdot)$是一个前向传播网络，输出为激活权重。

局部激活单元与attention方法类似。但不同的是，$\sum_i w_i = 1$的约束被放宽了，为了存储更为强烈的用户兴趣。因此，在$a(\cdot)$输出后的用来归一化的softmax函数也舍弃了。例如，如果一个用户的历史行为包含90%的衣服和10%的电子产品。给定T恤和电话的两个候选广告，T恤会激活大多数属于衣服的历史行为，并且可能比电话获得更大的v_U（更高的兴趣强度）价值。传统的注意力方法是通过对$a(\cdot)$的输出进行归一化而失去v_U的数值规模的决心。【这里给出的开源代码却用了softmax】

作者尝试了LSTM对用户历史行为数据进行建模。但这并没有改善。因为用户历史行为的序列可能包含多个并发兴趣。这些兴趣的快速跳跃和突然结束导致用户行为的序列数据似乎很嘈杂。但这是一个研究的方向。【2019年的DIEN】

Training Technique

在阿里巴巴的广告系统中，商品和用户数量达到了数亿。实际上，训练具有大规模稀疏输入特征的工业深度网络是巨大的挑战。

Mini-batch Aware Regularization

模型的过拟合一般来说需要正则化来进行抑制，但对于工业数据集来说，直接应用传统的正则化方法是不实际的，在训练网络上有稀疏输入和上百万的参数。以l2正则化为例。在基于SGD的优化方法的情况下，仅需要更新每个微型批处理中出现的非零稀疏特征的参数，而无需进行正则化。但是，当添加l2正则化时，需要为每个小批量计算整个参数的L2-范数，这将导致计算量极大，并且参数扩展到数亿个是不可接受的。

作者介绍了一种有效的小批量处理感知型正则化器，它仅针对每个微型批处理中出现的稀疏特征的参数计算**L2-范数**，从而使计算成为可能。$\mathbf{W} \in \mathbb{R}^{D \times K}$定义了整个embedding字典的参数，$D$是embedding的维度，$K$是特征空间的维度。

$$L_2(\mathbf{W}) = \|\mathbf{W}\|_2^2 = \sum_{i=1}^{K} \|\boldsymbol{w}_j\|_2^2 = \sum_{(x,y)\in S} \sum_{i=1}^{K} \frac{I(\boldsymbol{x}_j \neq 0)}{n_j} \|\boldsymbol{w}_j\|_2^2$$

其中$w_j \in \mathbb{R}^D$是第j个embedding向量，$I(x_j \neq 0)$是x含有特征id$j$，n_j代表所有样本中特征id$j$出现的数量。

进一步转化：

$$L_2(\mathbf{W}) = \sum_{j=1}^{K} \sum_{m=1}^{B} \sum_{(\boldsymbol{x},y)\in\mathcal{B}_m} \frac{I(\boldsymbol{x}_j \neq 0)}{n_j} \|\boldsymbol{w}_j\|_2^2$$

其中B定义了mini-batches的数量，$\mathcal{B}_m$定义了第m个mini-batch。

可以被近似为：

$$L_2(\mathbf{W}) \approx \sum_{j=1}^{K} \sum_{m=1}^{B} \frac{\alpha_{mj}}{n_j} \|\boldsymbol{w}_j\|_2^2$$

特征j的embedding权重的梯度为：

$$\boldsymbol{w}_j \leftarrow \boldsymbol{w}_j - \eta \left[\frac{1}{|\mathcal{B}_m|} \sum_{(\boldsymbol{x},y)\in\mathcal{B}_m} \frac{\partial L(p(\boldsymbol{x}),y)}{\partial \boldsymbol{w}_j} + \lambda \frac{\alpha_{mj}}{n_j} \boldsymbol{w}_j \right]$$

Data Adaptive Activation Function

PRelu是常用的激活函数：

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

PReLU采用值为0的硬修正点（a hard rectified point），当每层的输入遵循不同分布时，这可能不适合。作者设计了一个新的激活函数Dice：

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

其中$E[s]$和$Var[s]$为均值和方差，ϵ是一个常量，为10^{-8}

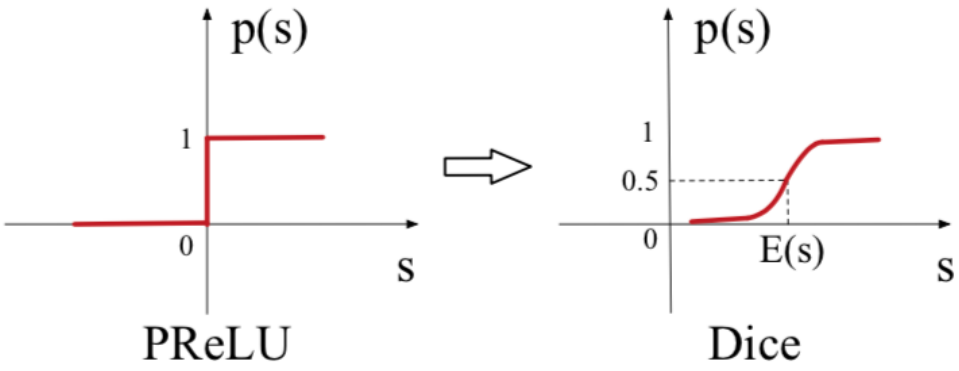


Figure 3: Control function of PReLU and Dice.

Dice的关键思想是根据输入数据的分布来自适应地调整修正点，其值设置为输入的平均值。当$E[s] = 0, Var[s] = 0$则退化为PRelu。

实验

数据集

Table 2: Statistics of datasets used in this paper.

Dataset	Users	Goods[a]	Categories	Samples
Amazon(Electro).	192,403	63,001	801	1,689,188
MovieLens.	138,493	27,278	21	20,000,263
Alibaba.	60 million	0.6 billion	100,000	2.14 billion

[a] For MovieLens dataset, goods refer to be movies.

Amazon Dataset：包含来自Amazon的产品评论和元数据。选取电商类子集包含192403用户，63001物品，801个种类，和1689188个样本。特征包括： `goods_id` 、 `cate_id` 、 `goods_id_list` 、 `cate_id_list` 。

MovieLens Dataset：选用20M，包含138493用户，27278电影，21种类和20000263样本数。为了适应CTR预估任务，将其转化为2元分类数据---评分4～5为正样本，其余为负样本。特征为： `movie_id` 、 `movie_cate_id` 、 `user rated movie_id_list` , `movie_cate_id_list` 。

Alibaba Dataset：从阿里巴巴的在线展示广告系统收集了流量日志，其中两个星期的样本用于训练，第二天的样本用于测试。训练和测试集的规模分别约为20亿和1.4亿。embedding维度为12对于所有的16个组来说。

Baseline

LR：在深度学习网络之前应用非常广泛；

BaseModel；

Wide&Deep：wide：手工设计低阶特征的交叉，deep：自动提取高阶非线性特征

PNN：embedding层之后引入乘积层捕获高阶特征交互。

DeepFM：将Wide&Deep模型中的wide部分改为FM。

Metrics

AUC：在原有基础上引入了用户加权AUC的变化形式，它通过对用户AUC进行平均来衡量用户内部订单的优劣，并且显示出与展示广告系统中的在线效果更为相关。计算公式如下：

$$\text{AUC} = \frac{\sum_{i=1}^{n} \#\text{impression}_i \times \text{AUC}_i}{\sum_{i=1}^{n} \#\text{impression}_i}$$

其中n表示用户的数量。

RelaImpr：衡量模型的相对改进。对于随机猜测者，AUC的值为0.5。因此，RelaImpr的定义如下：

$$RelaImpr = \left(\frac{\text{AUC(measured model)} - 0.5}{\text{AUC(base model)} - 0.5} - 1 \right) \times 100\%$$

Result

Table 3: Model Coparison on Amazon Dataset and MovieLens Dataset. All the lines calculate RelaImpr by comparing with BaseModel on each dataset respectively.

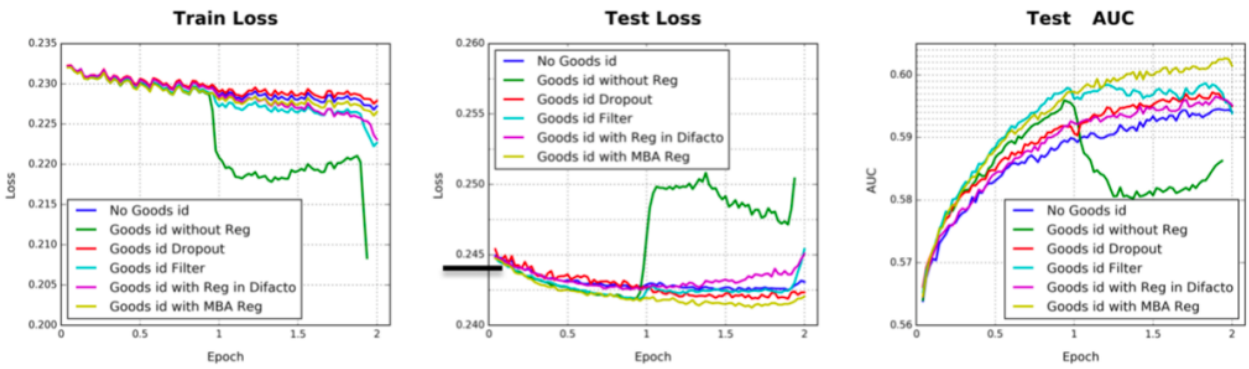
Model	MovieLens.		Amazon(Electro).	
	AUC	RelaImpr	AUC	RelaImpr
LR	0.7263	-1.61%	0.7742	-24.34%
BaseModel	0.7300	0.00%	0.8624	0.00%
Wide&Deep	0.7304	0.17%	0.8637	0.36%
PNN	0.7321	0.91%	0.8679	1.52%
DeepFM	0.7324	1.04%	0.8683	1.63%
DIN	**0.7337**	**1.61%**	**0.8818**	**5.35%**
DIN with Dice[a]	**0.7348**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

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

1、深度学习网络打败了LR，证明了深度学习提取高阶特征的能力。

2、PNN和DeepFM效果比Wide＆Deep更好。DIN在所有竞争对手中表现最好。特别是在具有丰富用户行为的Amazon Dataset上，DIN表现突出。这归功于DIN中局部激活单元结构的设计。DIN通过软搜索与候选广告相关的部分用户行为来关注局部相关的用户兴趣。通过这种机制，DIN获得了用户兴趣的自适应变化表示，与其他深度网络相比，极大地提高了模型的表达能力。此外，带Dice的DIN带来了对DIN的进一步改进，从而验证了所提出的Dice的有效性。

正则化参数



由于Amazon数据集和Movielens数据集的功能维度都不高（约10万），因此所有深度模型（包括我们提出的DIN）都不会遇到严重的过拟合问题。但是，当涉及包含较高维度稀疏特征的在线广告系统中的Alibaba数据集时，过度拟合将是一个很大的挑战。例如，当训练具有细粒度特征的深层模型（例如，表1中尺寸为6亿个goods_ids的特征）时，不加正则化会在第一个epochs之后会发生严重的过度拟合，这会导致模型性能迅速下降。因此，检验几种常用正则化的性能：

- Dropout：在每一个样本中随机丢弃50%的特征id；
- Filter：按样本中的出现频率过滤访问的goods_id，仅保留最频繁的那些。剩下的前2000万个goods_id；
- Regularization in DiFacto：与频繁特征相关的参数不太会被过度正则化；
- MBA：Mini-Batch Aware regularization method；

Dropout可快速防止过拟合，但会降低收敛速度。DiFacto中的正则化会以较高的频率对goods_id设置更大的惩罚，其效果要比Filter差。MBA效果最好。

Result of Alibaba Dataset

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.

online A/B testing

2017年5月至2017年6月在阿里巴巴的展示广告系统中进行了在线A/B测试。在将近一个月的测试中，与在线服务模型的最新版本BaseModel相比，接受了Regularizer和激活函数的DIN贡献了高达10.0％的点击率和3.8％的RPM（每千收入）促销。这是一项重大改进，证明了提出的方法的有效性。

原文：

“值得一提的是，每天都有成千上万的用户访问我们的系统，对工业深层网络进行在线服务并非易事。更糟糕的是，在流量高峰时，我们的系统每秒为超过100万用户提供服务。需要以高吞吐量和低延迟进行实时CTR预测。例如，在我们的真实系统中，我们需要在不到10毫秒的时间内为每个访问者预测数百个广告。”

DIN的可视化

显示了带有t-SNE的商品Embedding向量的可视化图。

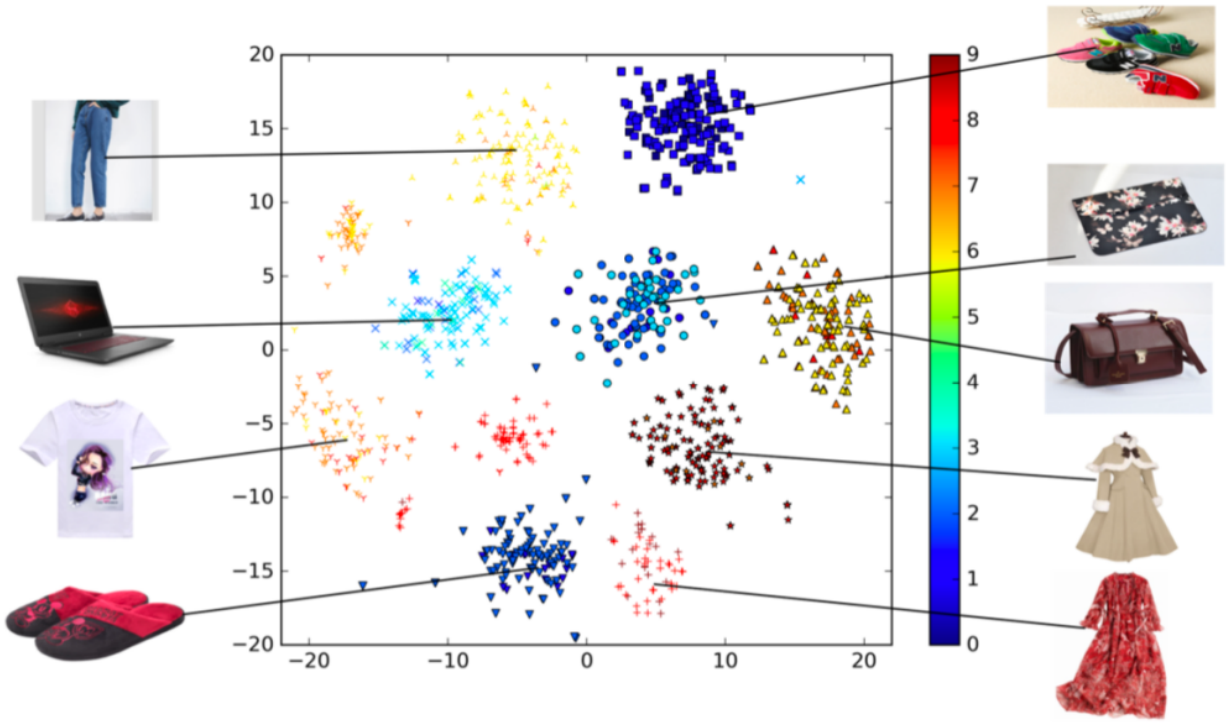


Figure 6: Visualization of embeddings of goods in DIN. Shape of points represents category of goods. Color of points corresponds to CTR prediction value.

总结

1、本文设计了DIN的新CTR预估模型来通过Local activation unit来获取针对不同广告而变化的用户兴趣的自适应表示向量。

2、还引入了两种新颖的技术（Mini-batch Aware Regularization、Data Adaptive Activation Function）来帮助培训工业深度网络并进一步提高DIN的性能。它们可以轻松地推广到其他行业深度学习任务。

代码复现---TF2.0

Github：https://github.com/BlackSpaceGZY/Recommended-System

数据集、代码详细介绍：https://zhuanlan.zhihu.com/p/144153291

参考文献

[1] Cheng H T, Koc L, Harmsen J, et al. Wide & deep learning for recommender systems[C]//Proceedings of the 1st workshop on deep learning for recommender systems. 2016: 7-10.
[2] Wang R, Fu B, Fu G, et al. Deep & cross network for ad click predictions[M]//Proceedings of the ADKDD'17. 2017: 1-7.
[3] Qu Y, Cai H, Ren K, et al. Product-based neural networks for user response prediction[C]//2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE, 2016: 1149-1154.
[4] Guo H, Tang R, Ye Y, et al. DeepFM: a factorization-machine based neural network for CTR prediction[J]. arXiv preprint arXiv:1703.04247, 2017.
[5] Zhou G, Zhu X, Song C, et al. Deep interest network for click-through rate prediction[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 1059-1068.

往期精彩回顾

第一次参赛---2020腾讯广告算法大赛Baseline思考与分析
Pandas笔记---通过比赛整理出的10条Pandas实用技巧
【论文导读】异构信息网络的Embedding进行推荐
【论文导读】MATRIX FACTORIZATION TECHNIQUES FOR RECOMMENDER SYSTEMS
机器学习笔记---信息熵

扫码关注更多精彩