Deep Interest Network for Click-Through Rate Prediction

Guorui Zhou, Chengru Song, Xiaoqiang Zhu
Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, Kun Gai
Alibaba Group
{guorui.xgr,chengru.scr,xiaoqiang.zxq,zhuhan.zh,fanying.fy,maxiao.ma,yanghui.yyh,junqi.jjq,lihan.hl,jingshi.gk}@alibaba-inc.com

PAN XIANG

2018.10.30

FLOW

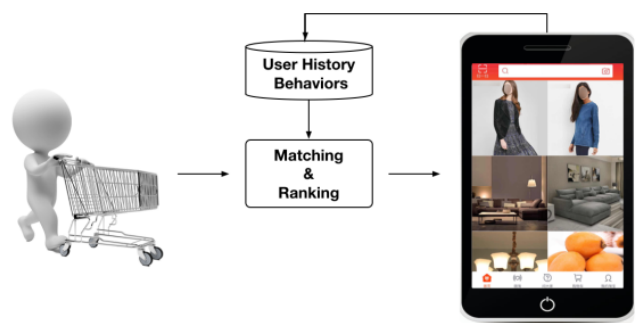

Figure 1: Illustration of running procedure of display advertising system in Alibaba, in which user behavior data plays important roles.

- check history
- matching model
- ranking model
- CTR rank
- response -> label

Embedding&MLP paradigm

usual way

- embedding
- transformed into fixed-length vectors in a group-wise manner
- fully connected layers(multilayer perceptron, MLP)

problem

- user interests are diverse
  - embedding vectors of user behaviors into a fixed-length vector
  - large?

CONTRIBUTION

DIN

- the related user interests by soft-searching for relevant parts of historical behaviors

- takes a weighted sum pooling to obtain the representation of user interests

Two novel techniques to help training industrial deep networks

- a mini-batch aware regularizer
- a data adaptive activation function
  - generalizes PReLU by considering the distribution of inputs and shows well performance

TEST ON Alibaba datasets

Embedding

[weekday=Friday, gender=Female,visited_cate_ids={Bag,Book}, ad_cate_id=Book]

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	$\sim 10^9$	multi-hot	$\sim 10^3$
	visited_shop_ids	$\sim 10^7$	multi-hot	$\sim 10^3$
	visited_cate_ids	$\sim 10^4$	multi-hot	$\sim 10^2$
Ad Features	goods_id	$\sim 10^7$	one-hot	1
	shop_id	$\sim 10^5$	one-hot	1
	cate_id	$\sim 10^4$	one-hot	1
	...	...	...	...
Context Features	pid	~ 10	one-hot	1
	time	~ 10	one-hot	1
	...	...	...	...

Figure 1:

four groups of features are illustrated as:

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

Figure 2:

Base Model

Embedding layer

pooling layer and Concat layer

- If $\boldsymbol{t}_i$ is one-hot vector with j-th element $\boldsymbol{t}_i[j] = 1$, the embedded representation of $\boldsymbol{t}_i$ is a single embedding vector $\boldsymbol{e}_i = w^i_j$.
- If $\boldsymbol{t}_i$ is multi-hot vector with $\boldsymbol{t}_i[j] = 1$ for $j \in \{i_1, i_2, ..., i_k\}$, the embedded representation of $\boldsymbol{t}_i$ is a list of embedding vectors: $\{\boldsymbol{e}_{i_1}, \boldsymbol{e}_{i_2}, ...\boldsymbol{e}_{i_k}\} = \{w^i_{i_1}, w^i_{i_2}, ...w^i_{i_k}\}$.

Figure 3:

$$\boldsymbol{e}_i = \text{pooling}(\boldsymbol{e}_{i_1}, \boldsymbol{e}_{i_2}, ...\boldsymbol{e}_{i_k}).$$

Figure 4:

MLP

LOSS

Loss. The objective function used in base model is the negative log-likelihood function defined as:

$$L = -\frac{1}{N} \sum_{(\boldsymbol{x},y)\in\mathcal{S}} (y \log p(\boldsymbol{x}) + (1-y)\log(1-p(\boldsymbol{x}))), \qquad (2)$$

where $\mathcal{S}$ is the training set of size N, with $\boldsymbol{x}$ as the input of the network and $y \in \{0, 1\}$ as the label, $p(\boldsymbol{x})$ is the output of the network after the softmax layer, representing the predicted probability of sample $\boldsymbol{x}$ being clicked.

Figure 5:

- overfitting
- computation and storage

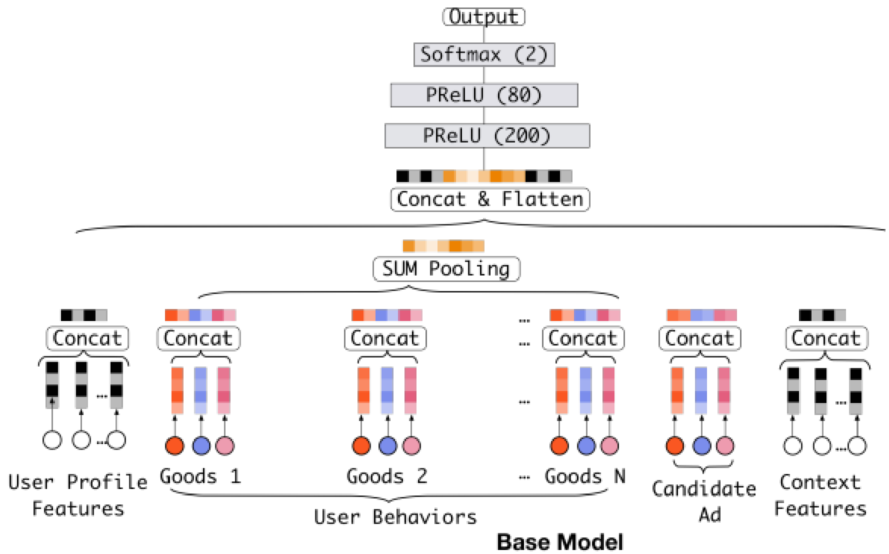

Figure 6:

---

behaviors related to displayed ad greatly contribute to the click action

ATTENTION?

POOLING

- We have tried LSTM to model user historical behavior data in the sequential manner. But it shows no improvement.
- Differentfrom text which is under the constraint of grammar in NLP task, the sequence of user historical behaviors may contain multiple concurrent interests
- Rapid jumping and sudden ending
  - special structure?->future

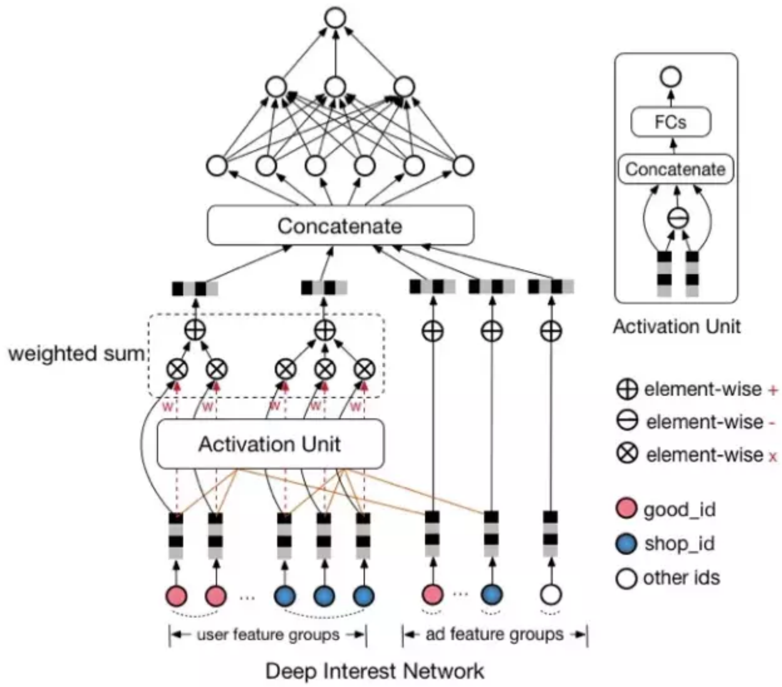

Figure 7:

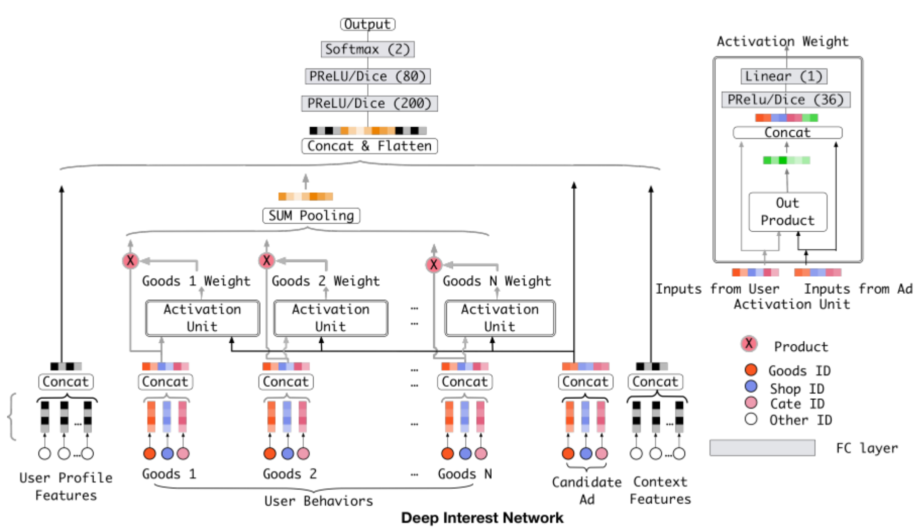

Figure 8:

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

Figure 9:

TRAINING TECHNIQUES

activation fun

Relu

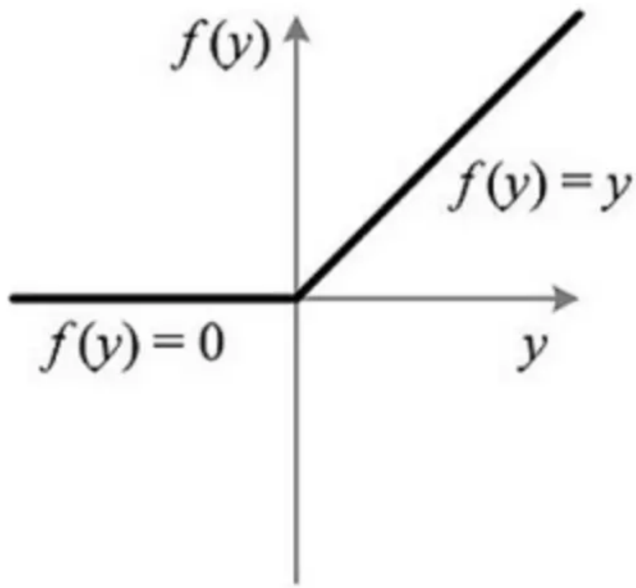

Figure 10:

PRelu(Leaky Relu)

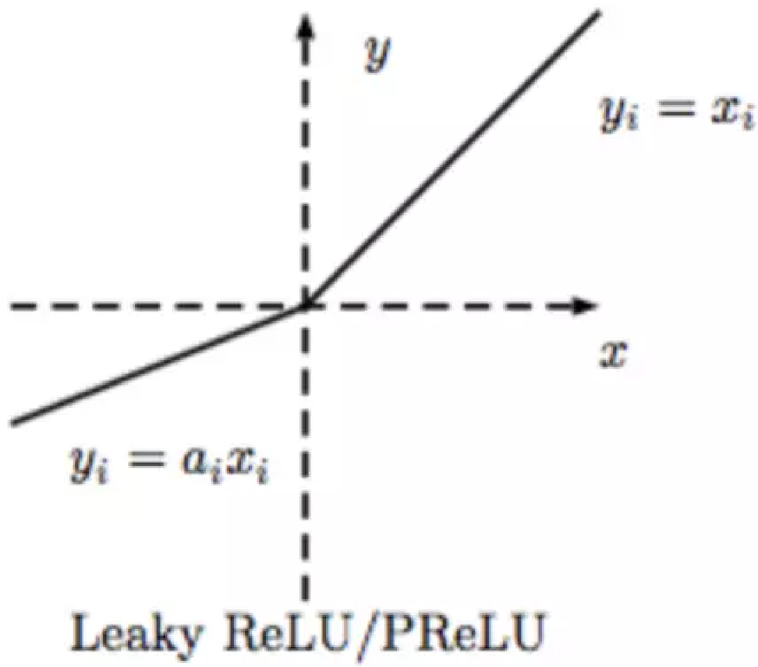

Figure 11:

DICE(Data Dependent Activation Function) - the division point should be decided by the data

$$y_i = a_i(1 - p_i)y_i + p_i y_i$$

$$p_i = \frac{1}{1 + e^{-\frac{y_i - E[y_i]}{\sqrt{Var[y_i] + \epsilon}}}}$$

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

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

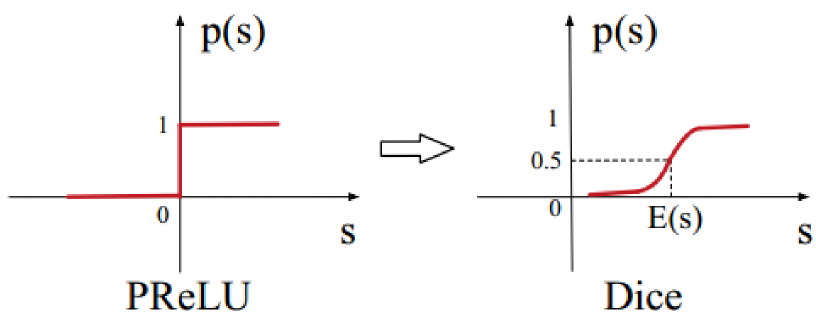

Figure 12:

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

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

Mini-batch Aware Regularization

Inspiration

- L1 L2 Dropout?
- long-tail law
  - many feature ids only appeared a few times
- Drop?
  - threshold(Hyperparameter)

---

principle

- according to the frequency of feature id adjusting the strength of regularization

- the more frequency the less strength of regularization

- the less frequency the more strength of regularization

$$I_i = \begin{cases} 1, & \exists (x_j, y_j) \in B, s.t. [x_j]_i \neq 0 \\ 0, & \text{other wises} \end{cases}$$

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

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

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

---

data set

http://jmcauley.ucsd.edu/data/amazon/
https://grouplens.org/datasets/movielens/20m/

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

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.

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%**

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

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%**

visualization

Figure 5: Illustration of adaptive activation in DIN. Behaviors with high relevance to candidate ad get high activation weight.

Figure 13:

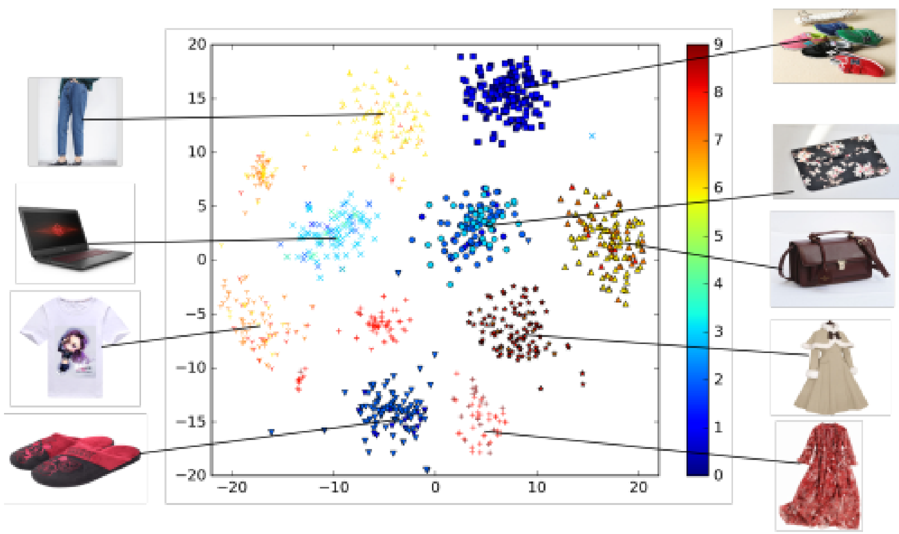

Figure 6: Visualization of embeddings of goods in DIN. Shape of points represents category of goods. Color of points corresponds to CTR prediction value.

Figure 14: