DEEP & ATTENTIONAL CROSSING NETWORK FOR CLICK-THROUGH RATE PREDICTION

Youming Zhang, Ruofei Zhu, Zhengzhou Zhu*, Qun Guo, Lei Pang

School of Software and Microelectronics, Peking University, Beijing, China

ABSTRACT

The problem of Click-through rate(CTR) prediction is the core issue to many real-world applications such as online advertising and recommendation systems. An effective prediction relies on high-order combinatorial features, which are often hand-crafted by experts. Limited by human experience and high implementation costs, combinatorial features cannot be manually captured thoroughly and comprehensively. There have been efforts in improving hand-crafted features automatically by designing feature-generating models such as FMs, DCN, and so on. Despite the great success of these structures, most of the existing models cannot differentiate the high-quality feature interactions from the huge amount of useless feature interactions, which can easily impair their performance. In this paper, we propose a Higher-Order Attentional Network(HOAN) to select high-quality combinatorial features. HOAN is a hierarchical structure, the multiple crossing layers can learn feature interactions of any order in an end-to-end manner. Inside the crossing layer, each interaction item has its unique weight with consideration of global information to eliminate useless features and select high-quality features. Besides, HOAN also maintains the integrity of individual feature embedding and offers interpretive feedback to the calculating process. Furthermore, we combine DNN and HOAN, proposing a Deep & Attentional Crossing Network (DACN) to comprehensively model feature interactions from different perspectives. Experiments on sufficient real-world data show that HOAN and DACN outperform state-of-the-art models.

KEYWORDS

Click-through rate prediction, Feature interaction networks, Attention mechanism, Hybrid model

1. INTRODUCTION

The click-through rate prediction has a wide range of application scenarios, such as recommendation systems and online advertising, which can directly affect the company's commercial revenue [1] [2]. Under certain business circumstances, thousandth improvements can bring huge economic benefits, thus click-through rate prediction is a very inspiring research direction both in industry and academia.

Effective prediction relies on combinatorial feature implemented by experts. However, it is difficult to achieve the desired effectiveness completely based on manual development. Firstly, the benefits of specific features rely on the repeated appearance of the same feature, and it can be seriously suffered from data sparsity [3]. Especially for high-order crossing features, they require more resources to develop but have lower occurrence, which, consequently, makes benefits fluctuating. Secondly, it is difficult to capture potential high-qualityfeature interactions with human experience as experts have strong limitations in designing combinatorial features that they have little knowledge of. And third, the number of feature interactions increases exponentially

David C. Wyld et al. (Eds): SPPR, NECO, GridCom, ICCSEA, SCAI, UBIC, SEMIT, MLDS - 2020
pp. 197-207, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101916

with its crossing degree [4]. Simply developed by humans requires an extremely heavy workload. However, a specific crossing model with proper design can achieve feature interactions with limited complexity. Considering the abovementioned limitations of artificial features, replacing or improving hand-craft engineering in an automatic way can lead to better performance and effectiveness.

The idea of automatically capturing feature interactions shows its superiority in some traditional machine learning models, one of the most representative model is Factorization Machines [5] and models based on FM such as AFM [6], HOFM [7]. Nowadays deep learning has provided a new perspective for a click-through rate prediction. One of the most widely used structures is Deep Neural Networks(DNN), DNN is very successful in condensing information as to its powerful capability in expression. Several state-of-the-art models choose DNN to learn feature expressions, but unfortunately, DNN has obvious limitations in modeling feature interactions. First, DNN calculates in a bit-wise way, but features are often projected into a vector in the Embedding & MLP paradigm which is widely used in click-through rate prediction models (that is, first, mapping each feature into a low-dimension and dense vector through an embedding layer, and then learn a specific structure to fit the target). Splitting original expression of the features may introduce incomplete information and be considered to be harmful. Second, DNN learns interactions in an implicit way. In CTR prediction, to meet the strict requirements on model efficiency, sometimes models need to provide feedback on the effect of features for selecting appropriate combinations, that interpretability is what DNN lacks. However, there are lots of successful structures modeling feature interactions. For example, [8] proposes Cross Network modeling high-order interactions in an efficient way. [9] introduced a multi-layered self-attention mechanism to learn cross features, maintaining the integrity of vector calculations. And [10] proposed the Compressed Interaction Network (CIN) introducing the convolutional neural network (CNN) mechanism to achieve feature crossover at any order. Despite their achievements, we find that most of the existing models lack the ability to select high-quality feature interactions. As there are a huge amount of useless interactions in all feature interactions, introducing all the interactions indiscriminately may seriously impair the performance of prediction.

Inspired by Self-Attention, a popular mechanism in natural language processing, this paper proposes a novel structure named Higher-Order Attention Network(HOAN) with the purpose of selecting high-quality feature interactions. Specifically, HOAN is a hierarchical structure, the multiple interacting layers can implement feature interactions of any order in an end-to-end manner. Within the interacting layer, each interaction has its unique weight with considering global information, which gives HOAN the ability to select high-quality interactions and eliminate useless ones, the particular design reducing the exponential complexity to an acceptable level. In addition, HOAN also maintains the integrity of individual feature vector and good interpretability in calculating process. We further combine the DNN and the proposed HOAN to learn feature interactions from low-order to high-order and propose a hybrid model named Deep & Attentional Crossing Network(DACN). To summarize, in this paper we make the following contributions:

- We propose a novel structure inspired by an attention mechanism named HOAN to select high-quality feature interactions through the crossing pro- cess. The hierarchical design of the network makes it possible to perform feature interactions of any order and keep an acceptable complexity. Furthermore, HOAN also has the characteristics of computational integrity and good interpretability.
- We take the HOAN as a core part to propose a hybrid model named DACN, utilizing DNN for generalization in an implicit way, and combining HOAN to learn feature interactions for memorization in an explicit way. The model does not need artificial feature engineering and captures more comprehensive interactions than HOAN.

- We conduct experiments on a sufficient real-world data set and evaluates the model from multiple aspects. The results show that HOAN and DACN gain superior performance than other state-of-the-art models.
  The code is available in https://github.com/meRacle-19/HighOrderAttention.

2. PRELIMINARIES

2.1. Click-Through Rate Prediction

CTR estimation has a wide range of applications, and its general form canbe defined as follows. Given $x \in R^N$ as input features, including user profile f_u and features about the item to be predicted f_t, as well as contextual features f_c, where N represents the dimensions of the feature vector. When the feature is encoded as a one-hot vector, N is the number of values of all features. Then the CTR estimate can be defined as the probability that a specific user clicks on a specific item in a given context.

Since features under business circumstances are often very high-dimensional and sparse, raw features can easily lead to overfitting. An intuitive method is to transform feature vectors like one-hot encoding into a low-dimensional continuous space, such as the embedding layer in deep networks does. Moreover, another effective method to overcome this problem is to combine the original features called a combinatorial feature, which has shown excellent results in many works.

2.2. Combinatorial features

Many high-quality work has appeared in the field of combined features, as well as different definitions of higher-order combinatorial features. We study in detail these state-of-the-art works and give definition of high-order feature interactions as Equation (1). Supposing $p_n(x)$to be high-order combinatorial features of degree n with the input feature $x \in R^N$, n-th order interactions can be written as:

$$p_n(x) = \left\{\sum_{\alpha} \omega_\alpha \cdot g_\alpha(x_1, x_2, \cdots, x_n) | 0 \leq |\alpha | \leq k^n\right\} \tag{1}$$

Where w is the weight of the combinatorial feature, k represents a number of feature values and g(·) is a non-additive combination function, such as dot product and Hadamard product. For n-order combinatorial features, it has $O(k^n)$ inter- actions including useful and useless features. For example, supposing f_g represents a user gender feature, $f_{v,m}$ and $f_{v,w}$ represent the duration of men and women watching videos respectively, second-order interaction $f^2(f_g = \text{man}, f_{v,m})$ is obviously more effective than $f^2(f_g = \text{man}, f_{v,w})$. Moreover, the latter may introduce noise which is harmful to prediction. Unfortunately, most of the existing approaches set w to a constant one, which ignores this point. One of our goals is to give each interaction unique weights to distinguish useful and useless features in an efficient way.

2.3. Embedding layer

Not like nature language processes and computer vision that their dense data can be directly fed to DNNs, data in CTR prediction is usually suffered from serious sparsity. Because data in CTR prediction is collected from a different source, showing less spatial or temporal correlation, single-value and multi-value features, as well as continuous feature all usually are converted to one-hot feature to enhance the generalization. For example, one instance {gender = male, age = 18, interests = basketball&music} will be converted to one-hot encoding {[1,0],[0,...,1,0,...,0],[0,...,1,0,1...,0]}.

However, these high-dimension feature encodings are very sparse and can not be directly used for deep networks. One particular solution is adopting Embedding & MLP diagram [11] [12] [13] [14]. As structures evolving, MLP has been replaced by more powerful deep networks, but the embedding layer is still adopted in most deep structures to compress one-hot encodings to relative low-dimension and dense-information vectors. For single-value feature, one-hot encoding is directly projected into a dense vector. As for multi-value features, they are first projected into several vectors, then added to one dense vector. The embeddingis calculated as follows:

$$e_i = \begin{cases} Wf_i \\ \sum_j Wf_{ij} \end{cases} \tag{2}$$

Where f_i is one-hot encoding, $e_i \in R^d$ and d is the length of dense embedding. In this paper, we feed the dense embedding to HOAN and DACN for the abovementioned reasons, also adopting fixed length for each feature to eliminate influence to feature crossing model.

3. Our Proposed Model

3.1. Higher-Order Attention Network

HOAN contains multiple crossing layers, the hierarchical structure is shown as Figure 1.

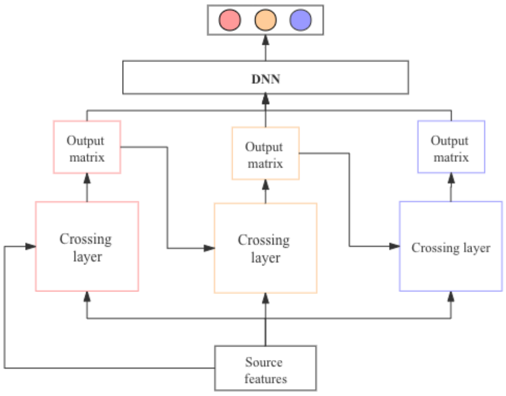


Figure 1: Hierarchical structure of HOAN.

For the i-th crossing layer, where i> 0, the input data consists of two particular parts. One part is the matrix produced by the (i−1)-th crossing layer noted as $M_{c,i-1}$, involving feature interactions of specific orders assuming as k. The other is the matrix produced by the embedding layer, involving densevectors of original features, considered to represent the first-order features, noted as M_s. After crossing by i-th layer, M_s and $M_{c,i-1}$ are merging into one matrix and $M_{c,i}$. With the assumption that $M_{c,i-1}$ denotes the k-order feature interactions, $M_{c,i}$ contains the (k+1)-order crossing features consequently, which is the sum order of M_s and $M_{c,i-1}$. The details of the crossing process will be discussed in the next phase. Then $M_{c,i}$ both can be re-crossing in next layer for higher-order and be processed by DNNs to produce layer output for final CTR prediction. One must pay attention to that, $M_{c,i-1}$ is actually M_s at the first crossing layer.

Within the crossing layer, the detail process is shown as Figure 2. The total calculation is:

$$M_{c,i_{p*}} = \sum_{q=1}^{n_f} \left[\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)_{i*} \cdot \left(V_i \circ \left\{B_{*1}^T, \dots, B_{*n_f}^T\right\}\right)\right]_{pq*} \tag{3}$$

Where $0 < p, q \le n_f$, n_f is a number of features, d is the length of feature embeddings, and $\circ$ denotes the Hadamard product like $\langle a_1, a_2, a_3\rangle \circ \langle b_1, b_2, b_3\rangle = \langle a_1b_1, a_2b_2, a_3b_3\rangle$. $Q, K, V, B \in R^{n_f \times d}$ are converted from the input data M_c and M_s respectively $Q = M_s w_q, B = M_s w_b$ and $K = M_c w_k, V = M_c w_v$, the projection is non-linear transformation. In fact, there are two fundamental elements in the formula, which are weights and values as shown in Figure 2. Weights are merged from Q and K by matrix multiplication as softmax(QK^T) to differentiate high-quality feature interactions from the useless ones. Values, a 3-dimension matrix, are transferred from V and B by Hadamard Product as $V_i \circ \left\{B_{*1}^T, \dots, B_{*n_f}^T\right\}$, including each crossing item of M_c and M_s. From a moredetailed perspective, weight and value of a couple of features, f_i and f_j, are both calculated from the corresponding dense feature vector. Supposing e_i and e_j are vectors of f_i and f_j, then the f_i related weight $W_{i,j}$ and value $V_{i,j}$ are:

$$W_{i,j} = \frac{e^{g\left(e_i e_j^T\right)}}{\sum_{k=0}^{n_f} e^{g\left(e_i e_k^T\right)}} \tag{4}$$

$$V_{i,j} = e_i \circ e_j \tag{5}$$

Where $g(\cdot)$ is non-linear transformation such as $Sogmoid$ or $Tahn$. Particularly, $W_{i,j}$ has different values in f_i related and f_j related calculation, as the denominator changes. For example, Equation (4) gives f_i related weight and f_j related weight's denominator is $\sum_{k=0}^{n_f} e^{g\left(e_k e_j^T\right)}$.

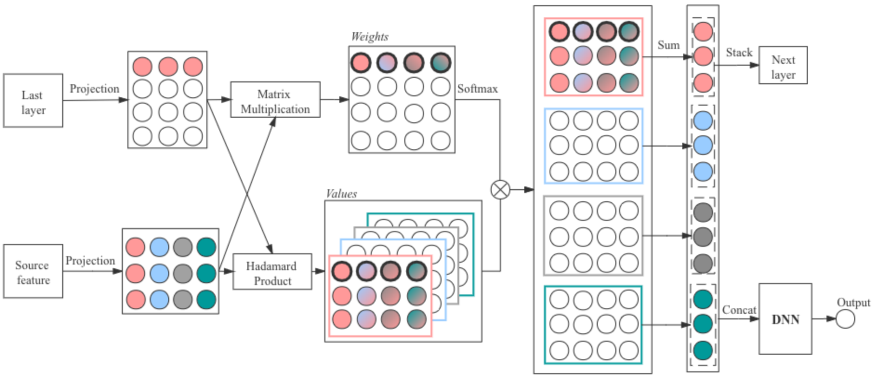


Figure 2: Internal structure of crossing layer

It is interesting to point out that Equation (3) has a strong connection with the well-known Self-attention in Natural Language Processing shown as Equation (6).[15] Q and K in Self-attention is the response to give unique weight to corresponding value, thus select high-quality feature values. Specifically, we add a base matrix B to introduce original feature for attention process in

HOAN, expanding V in Self-attention from original order to added order of two input matrices. At the same time, the base item of V corresponds to a vector instead of a single value, maintaining the integrity of the feature vector.

$$\text{Attention }(Q,K,V)=\text{softmax}\left(\frac{QK^T}{\sqrt{x}}\right)V \tag{6}$$

Figure 2 also gives the output of k-th crossing layer. After been crossing within layer, $M_{c,i}$ is first line up to one long embedding $e_s^+ \in R^{\sum_i H_i}$ with $H_i$denoting length of $M_{c,i}$, and then feed to DNNs to produce one single output:

$$y_i^{\text{hoan}}=\frac{1}{1+\exp\{e_s^{+T}w\}} \tag{7}$$

3.2. HOAN Analysis

3.2.1. Space Complexity

The k-th layer contains input data M_s and M_c, as well as MLP in projection and DNNs. M_s and M_c both occupy $O(n_f d)$ space. Supposing projecting output dimension is d_o, there are $O(dd_o)$ parameters in projection. Then $W\ eights$ and $V\ alues$ are both transformed from M_s and M_c, it doesn't introduce new parameters, but the $V\ alues$ itself contains $O(n_f^2 d)$elements. As for DNNs, it is related to depth d_p and width d_w, thus space complexity is $O(d_p d_w)$. To sum up, one single crossing layer has total $O(n_f^2 d + dd_o + d_p d_w)$ space complexity. Usually d_o, d are less than 10, can be treated as a constant and $n_f \gg d_o, d_w \gg d_o$, so simplified space complexity can be $O(n_f^2 d)$.

3.2.2. Time Complexity

Time complexity is discussed according to a sequence of forward propagation. The first is a projection, it has $O(n_f dd_o)$ calculations. Then $Weights$ and $V\ alues$ are produced with $O(n_f^2)$ calculations for each element and $O(n_f^2 d_o)$ for total time consumption. The next is crossing between $Weights$ and $V\ alues$, it is easy to know that each element in $Weights$ and interacts with corresponding vectors of $V\ alues$ for $O(d)$ times, and the total amount of $Weights$ is $O(n_f^2)$. Besides, the sum-pooling and DNNs inference can be ignored comparing to the abovementioned items. Even though, the total time complexity of one single layer still reaches $O(n_f^2 d_o)$, which is the major drawback of HOAN.

3.2.3. Polynomial Approximation

One of the most important properties of HOAN is high-order interactions. To examine it, we borrow the notations from [8] as shown in Equation (1). For simplicity, we simplify the HOAN by ignoring the details of $Weights$ calculation and concentrate on a single feature interaction. The simplified Equation of i-th layer can be:

$$x_c^i = W^i \cdot \left(x_c^{i-1} \circ x_s^0\right) \tag{8}$$

Where x^{i-1} is one dense feature vector produced by k-th crossing layer, and x_s^0 is original feature vector. There is no correspondence between x_c^{i-1} and some particular feature, the specific relation is hidden in W^i. Through this equation, $g(\cdot)$ in Equation (1) can be defined as $\circ$, thus HOAN can raise the order of feature interactions by personalized crossing. In addition, it also can be provedthat crossing order grows with the layer. The i + 1 layer can be written as:

$$x_c^i = W^{i+1} \cdot \left(x_c^i \circ x_s^0\right) \tag{9}$$

$$= W^{i+1} W^i \cdot \left(x_c^{i-1} \circ x_s^0 \circ x_s^0\right) \tag{10}$$

3.3. Deep Attentional Crossing Network

As discussed in Section 3.2, HOAN can add orders of input data. However, it at least only model second-order interactions in the first layer of HOAN, which lack the first-order feature information. To tackle this problem, we combine DNNs and HOAN to model feature interactions comprehensively. At the same time, a hybrid model can make amodel more robust like Wide & Deep. We name this model Deep attentional crossing network(DACN), the structure is shown in Figure 3.

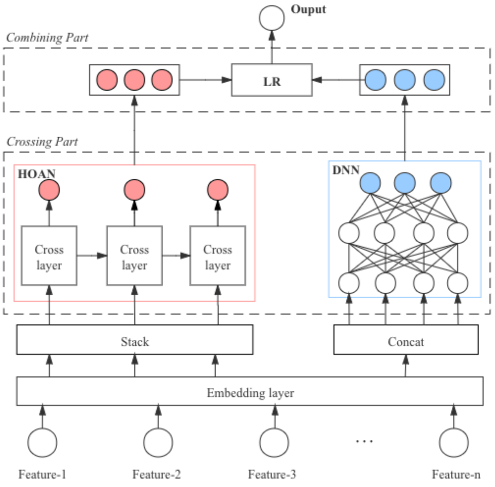


Figure 3: Structure of DACN.

DACN contains crossing part and combining part. Crossing part has HOAN modeling high-order feature interactions and DNNs modelingfirst-order interactions. In DNNs part, dense vectors from the embedding layer are the first contact to a long vector, and then feed into graph. And its output can be written as $y_d nn$. The combining part is the response to merging outputs of HOAN and DNNs, and produces a predicting score of CTR. We use LR in this part, the forward equation is shown as Equation (11).

$$\hat{y} = \sigma\left(W_{dnn}^T y_{dnn} + W_{hoan}^T \left[y_{hoan}^1, \dots, y_{hoan}^k\right]\right) \tag{11}$$

Where $y_d nn$ is output of the last layer in DNN, $y_h$oan^i is output of i-th layer in HOAN, $0 < i \leq k$ and k is the depth of HOAN.

4. EXPERIMENTS

4.1. Setup

4.1.1. Criteo Display Ads Data

The CriteoDisplay Ads[1] dataset is for ads click-through rate predicting. It contains 41 million records from a period of 7 days, each record has 13 integer features and 26 categorical features. Usually, a small improvement is considered as practically significant in ads CTR predicting. Especially for a large user base, a small improvement in prediction accuracy can potentially lead to a very large increase in a company's revenue. We randomly split the whole data to 10 folds, and use 8 folds for training, the rest averagely split for testing and validating.

4.1.2. Implementation Details

We briefly discuss some implementation details for training with DACN. As feature crossing is a property to be examined, we do not include any hand-crafted cross features. To keep concentration on model structure, we use fixed length 10 as feature embedding for all models. The learning rate is 0.001, and the batch-size is set to be 4096. We use L2 regularization with λ = 0.001 and dropout rate 0.1 in DACN. All other hyper parameters are tuned by grid-searching on the validation set, detailed settings is showed in the corresponding section. The code is available at http://labs.criteo.com/2014/02/kaggle-display-advertising-challenge-dataset/.

4.1.3. Baselines

To evaluate the performance of HOAN, we choose logistic regression(LR), Deep Neural Networks(DNN), Factorization Machine(FM), Wide and Deep Model (W&D), Deep & Cross Network(DCN) and eXtreme Deep Factorization Machine(xDeepFM) as baselines. Specifically, we compare HOAN with FM, DNN, CrossNet and Compressed Interaction Network(CIN), core part of DACN. DACN is compared with Integrated models including LR, FM, DNN, DCN, W&D andxDeepFM. All the baseline models are state-of-the-art models for the recommender system. In addition, they all are related to feature crossing. For example, LR models first-order interactions and FM models the second-order features, the other models like DNN, DCN and xDeepFM can model high-order interactions.

4.1.4. Metrics

We use AUC (Area Under the ROC curve) and Logloss (cross entropy) for model evaluation. AUC evaluates the possibility that one positive instance ranks higher than a negative instance. Thus higher AUC means a more suitable order in predicting instances. LogLoss measures how far a predicted score to a true label for each instance.

4.2. Experiment on Individual Crossing Networks(Q1)

We choose feature interacting structures for comparison with HOAN. LR and FM model specific order of combinatory features. Cross Net(CN), which is the core part of DCN, models high order with very few parameters. And Compressed Interacting Network(CIN) is a core part of xDeepFM, one particular advantage of CIN is that it models high order in an explicit way. All the structures

Table 1: Performance of individual models on the Criteo

model name	AUC	Logloss	Order
LR	0.7583	0.4806	-
FM	0.7727	0.4701	2
CN	0.7779	0.4655	4
CIN	0.7816	0.4642	4
HOAN	0.7847	0.4597	4

are shown in table 1. On the one hand, structures that model high order interactions such as CIN, CN and HOAN outperform FM and LR, which only can learn second order combinatory features. On the other hand, CIN and HOAN are in the same level of performance, it is probably because that they have similar complexity in space and time. In addition, our HOAN outperforms theother models, shows the superiority of selecting high-quality feature interactions.

4.3. Experiment on Hybrid Models(Q2)

DACN integrates HOAN and DNN into an end-to-end model. To match the properties of DACN, we compare HOAN with hybrid models that contain a crossing structure, and the results are shown in table 2. It can be seen that the hybrid model outperforms individual structures indicating that the combination indeed improves model performance. Besides, we are interested in how much does feature interaction layer improves. We observe that DCN, which contains a cross network for crossing features, and xDeepFM, which contains CIN for feature interactions, have better performance than those don't contain crossing network. It is probably because we haven't included artificial features, making more reliance on automatic feature crossing. And surprisingly, the results show that DACN still outperforms the other hybrid models.

Table 2: Performance of hybrid models on the Criteo

model name	AUC	Logloss	Sub-structures
DNN	0.7782	0.4651	-
Wide & Deep	0.7821	0.4701	DNN, LR
DCN	0.7833	0.4655	DNN, CN
xDeepFM	0.7879	0.4642	LR, DNN, CIN
DACN	0.7922	0.4597	HOAN, DNN

4.4. Explanation of HOAN(Q3)

The explanation is one of the most important properties of HOAN. To verify it, we first extract all weights of interactions and rank features by the sum of its weights, in which higher rank indicates higher contribution to prediction. Then we choose one trained model as a baseline. Furthermore, we remove five most valuable features shown in sort list as the Group 1 and drop five most useless features as Group 2. By retraining HOAN, we can find the results of the test set in table 3. Obviously, the performance of Group 2 has a very little downtrend comparing to baseline, but Group 1 has a certain decrease. This clearly shows that the feedback of HOAN is effective.

Table 3: Performance of re-trained models after filtering features.

Group	AUC	Trend
Baseline	0.7811	0.0%
Group-1	0.7806	-0.17%
Group-2	0.781	-0.01%

5. CONCLUSIONS

In this paper, we propose a novel network named Higher-Order Attention Networks, aiming at differentiating the high-quality feature interactions from the huge amount of useless feature interactions. HOAN can learn certain order of feature interactions. Besides, it also maintains the integrity of individual feature embedding and good interpretability through calculating process. Inspired by a popular combination diagram, we further incorporate a DNN and a HOAN in one end-to-end framework and named this hybrid model as Deep & Attentional Crossing Network. Thus DACN does not need extra artificial feature engineering and has superiorities of both generalization and memorization. We conduct experiments on sufficient public data and the results demonstrate that our model outperforms other models.

There are some directions for future work. First, as discussed in section 3.2.2, the high time complexity is one major downside of HOAN. As feature interaction $f_{i,j}$ is calculated twice in single inference, we are interested in exploit a better implementation like Matrix Decomposition and Factorization Machine do to reduce complexity. Second, with consideration of complexity, we simply use sum-pooling to produce the output matrix. Finding a more effective way is our next goal.

ACKNOWLEDGEMENTS

This paper was supported by the National Key Research and Development Program of China (Grant No.2017YFB1402400), the Ministry of Education "Tiancheng Huizhi" Innovation Promotes Education Fund (Grant No.2018B01004).

REFERENCES

[1] M. Richardson, E. Dominowska, R. Ragno, (2007) "Predicting clicks: estimating the click-through rate for new ads", Proceedings of the 16th international conference on World Wide Web, pp. 521–530.

[2] H. B. McMahan, G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov, D. Golovin, et al.,(2013) "Ad click prediction: a view from the trenches", Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1222–1230.

[3] Y. Shan, T. R. Hoens, J. Jiao, H. Wang, D. Yu, J. Mao, (2016)" Deep crossing: Web-scale modeling without manually crafted combinatorial features", Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 255–262.

[4] Y. Qu, H.Cai, K. Ren, W. Zhang, Y. Yu, Y. Wen, J. Wang, (2016)" Productbased neural networks for user response prediction", 2016 IEEE 16th International Conference on Data Mining (ICDM), IEEE, pp. 1149– 1154.

[5] S. Rendle,(2010)" Factorization machines", 2010 IEEE International Conference on Data Mining, IEEE, 2010, pp. 995–1000.

[6] J. Xiao, H. Ye, X. He, H. Zhang, F. Wu, T. S. Chua, " Attentional factorization machines: Learning the weight of feature interactions via attention networks. "

[7] M. Blondel, A. Fujino, N. Ueda, M. Ishihata, (2016)"Higher-order factorization machines", Advances in Neural Information Processing Systems, pp. 3351–3359.

[8] R. Wang, B. Fu, G. Fu, M. Wang, (2017)“Deep & cross network for ad click predictions”, Proceedings of the ADKDD’17, pp. 1–7.

[9] W. Song, C. Shi, Z. Xiao, Z. Duan, Y. Xu, M. Zhang, J. Tang,(2019) “Autoint: Automatic feature interaction learning via self-attentive neural networks”, Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 1161–1170.

[10] J. Lian, X. Zhou, F. Zhang, Z. Chen, X. Xie, G. Sun, (2018)“xdeepfm: Combining explicit and implicit feature interactions for recommender systems”, Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1754–1763.

[11] X. He, T.-S. Chua, (2017)“Neural factorization machines for sparse predictive analytics”,Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval,, pp. 355–364.

[12] W. Ouyang, X. Zhang, L. Li, H. Zou, X. Xing, Z. Liu, Y. Du, (2019)“Deep spatio- temporal neural networks for click-through rate prediction”, Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2078 – 2086.

[13] G. Zhou, N. Mou, Y. Fan, Q. Pi, W. Bian, C. Zhou, X. Zhu, K. Gai, (2019)“Deep interest evolution network for click-through rate prediction”, Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, pp. 5941– 5948.

[14] G. Zhou, X. Zhu, C. Song, Y. Fan, H. Zhu, X. Ma, Y. Yan, J. Jin, H. Li,K.Gai, (2018)“Deep interest network for click-through rate prediction”, Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1059–1068.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L . Kaiser, I. (2017)“Polosukhin, Attention is all you need”, Advances in neural information processing systems, pp. 5998–6008.

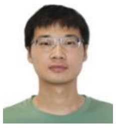
AUTHORS

Youming Zhang, Ph.D. candidate form Peking University. His current research interests include machine learning, MOOC adaptive learning and online educations.

Ruofei Zhu, Master degree graduate from Peking University. His current research interests include machine learning, commercial advertising and computing optimization.

Zhengzhou Zhu, Ph.D., associate professor in Peking University. His current research interests include Education big data, personalized recommendation. As the project leader presided over the National Natural Science Foundation and the Doctoral Fund of Ministry of education, Ministry of Education Key Laboratory of school funds and other national and provincial projects.

Qun Guo, Master degree graduate from Peking University. His current research interests include machine learning, model compressing and multi-modality.



Lei Pang, Master degree graduate from Peking University. His current research include dialogue system, information retrieval and onlineeducation.