ISSN 2616-5775 Vol. 7, Issue 1: 18-23, DOI: 10.25236/AJCIS.2024.070103

Reweighted Attention with Auto Feature Select Gate for CTR Prediction

Chengxu He[1,a], Jing Chen[1,b,*]

[1]Guangdong University of Technology, 100 Waihuan Xi Road, Higher Education Mega Center, Guangzhou, 510000, China
[a]2112115165@mail2.gdut.edu.cn, [b]jchen125@gdut.edu.cn
[*]Corresponding author

Abstract: *CTR (Click-Through-Rate) prediction is an important part of today's recommendation scenarios. Its purpose is to predict whether the user will click on the relevant item.The more mainstream research includes feature interaction and user history behavior interest modeling. However, in some current CTR feature interaction methods, only bit-wise or vector-wise feature interaction is used, or the two-tower approach is used to simply add the output of the two-tower structure at the prediction layer. These methods cannot more comprehensively represent the interaction between features, thus losing a lot of potential feature interaction information.In this paper we propose two novel method:1)Reweighted Attention Network(RAN),which employs vector re-weight after capturing the explicit feature interaction .This module can help model capture the high-order feature potential information more effectively.2)Auto Feature Select Gate(AFSG),which mining potential interactive information of shallow features and higher-order features On the basis of avoiding information loss.Experiments on three public datasets show that our method performs better than the current mainstream CTR prediction models.*

Keywords: *Data mining, CTR prediction, Feature interaction, Gate network*

1. Introduction

CTR is very important for industrial recommendation systems, and the estimated click-through rate will directly affect some subsequent advertising placement decisions and arrangement of advertising positions. Therefore, the impact of click-through rate prediction on platform revenue and user experience is very critical [4]. The data is usually in the form of row data as input to the model, eg{male, user_id, student, blue}.

Many models have been proposed and applied in this field,some shallow networks such as logistic regression(LR), Factorization Machines(FM)[8], field-aware factorization machine(FFM).And with the development of deep neural networks,many deep learning base models have been proposed ,including PNN, FNN with serial structure and parallel structure generally composed of two components, such as DeepFM[5], DCN[1], xDeepFM[6] and other models.It is obvious that the combination of explicit feature interaction module and DNN module has achieved excellent results in academic fields and industrial scenarios.Then, in the electronic shopping landscape, DIN[2],DIEN and other models of feature interaction are proposed for different types of data and scenarios.However, in many dual-tower models based on Wide deep model architecture, the MLP part and the independent feature interaction part of the model only add logits at the last output layer, which results in the independent feature output of the two parts.Therefore, we propose Reweighted Attention as an independent module to obtain high-order interactions of features by using attention and MLP layers .Our method is inspired by the Dual FEN Layer of DIFM[11] ,which is aim to get a reweight representation of the original feature.However,Taking the higher-order feature interaction as the weight of the original feature, the process includes linear transformation and reshape, which will lead to the loss of too much higher-order feature interaction information.Therefore, in this paper, Multi-Head Self-Attention is directly used to represent the higher-order interaction of features, while bit-wise feature interaction information is extracted through MLP, and the two are fused to obtain the higher-order interactive representation of features, which enable obtain a more fine grained feature representation while preventing information loss.On this basis, we use AFSG to better extract the feature importance relationship information, and as a gating structure, effectively combine the high-level interaction information with the original information, and further avoiding the problem of information loss and the phenomenon of weak gradient.

ISSN 2616-5775 Vol. 7, Issue 1: 18-23, DOI: 10.25236/AJCIS.2024.070103

Our main contributions are as follows:

• This paper constructs a novel high-order representation method for feature interaction, which combines bit-wise and vector-wise feature interaction without information loss.

• Solve the combination of high-order feature interaction and low-order interaction in CTR prediction via AFSgate.

• The two components proposed by us can be flexibly combined with other methods, and the combination of FM layers has achieved the best effect in our experiment.

2. Related work

2.1. Feature interaction in CTR

In recent years, there have been many deep learning based CTR prediction models that have achieved remarkable results in different directions. Among them, feature interactions are the most important in improving CTR forecast indicators.

In 2016, google proposed the wide&deep model, which kicked off the large-scale application of Deep learning in the field of ctr prediction. The Wide&Deep model is a hybrid model composed of a single-layer Wide part and a multi-layer deep part. Later, DeepFM proposed to replace the wide part of Wide&Deep with FM; DCN proposed Cross Network to learn higher-order feature interaction; DCN-V2 replaced vector parameters with matrix parameters to increase the expression capability of the network; xDeepFM[6] proposed CIN structure,a vector-wise interaction approach to construct more interpretable higher-order feature interactions. AutoInt[7] employs multi-head self-attention layers to represent feature interactions,FINT[9] design Field-aware Interaction Layer,which effectively extract field—award information,its also vector-wise level interaction method.PHN[3] combine three parallel towers (FFN, Cross layer, Field Interaction layer) to improve the expression ability of CTR prediction model.

2.2. Gating mechanism

Gated Networks are a commonly used neural Network structure that uses gating mechanisms to control the flow of information and selectively integrate information from different sources, such as Highway Networks in computer vision, which utilize conversion gates and carry gates to represent how much output is generated by converting inputs and carry outputs, respectively. In the field of NLP, such as LSTM[12], GRU can automatically learn the importance of different modules by using gate.

In the recommendation system method, MMOE proposed to use gate network to control the weights of different expert networks, and achieved remarkable results on the multi-task model. In DIN[2], gated weighting is used to dynamically adjust the weight of different users' historical behavior. In GateNet[10], it is proposed that the embedding gate should employ a MLP gate at the embedding layer and employ a MLP gate after MLP. It is proposed that information can be shared between deep feature representations and shallow feature representations for simultaneous use in CTR prediction.

3. Methods

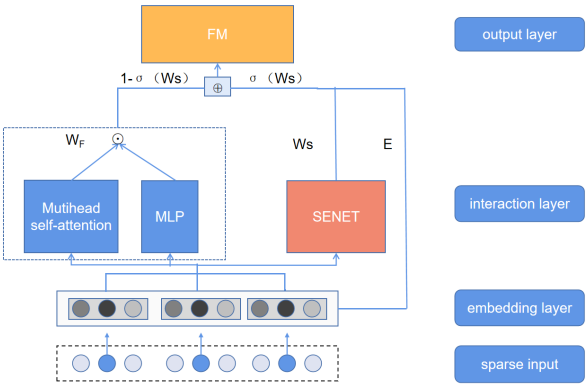

Figure 1: The overall structure

ISSN 2616-5775 Vol. 7, Issue 1: 18-23, DOI: 10.25236/AJCIS.2024.070103

The traditional CTR model can be divided into three parts: Embedding layer, interaction layer, output layer, and our approach follows this structure. The overall structure is shown in Figure 1

3.1. Multi-head Self-Attention

Before transformer's Multi-head Self-Attention was proposed, various attention mechanisms had emerged.Transformer is inspired by the idea of CNNS using multiple kernels in the same convolutional layer. In Multi-head Self-Attention, the input vector is first transformed into QKV, then the QKV is split into several parts respectively, and the self-attention is calculated separately. Finally, the obtained results are joined together, so that the model can learn independent information in different subspaces. In short, the purpose of multi-head design is to make each attention head focus on only one subspace, independent of each other.

This multi-dimensional approach to consider the relationship between input information coincides with the need for different context representations of CTR. Through the multi-head mechanism, it is possible to consider the interaction between features in multiple subspaces when each field feature is represented in different contexts with other fields.

First reshape the output of the resulting embedding layer to $E_o \in \mathbb{R}^{f \times d}$

$$\mathrm{E_o} = \mathrm{reshape(E)} \tag{1}$$

Through $\mathrm{W_{Q_i}}, \mathrm{W_{k_i}}, \mathrm{W_{v_i}}$ hree linear transformation matrix obtained $\mathrm{Q_i}, \mathrm{K_i}, \mathrm{V_i}$

$$\begin{aligned} \mathrm{Q_i} &= \mathrm{E_o W_{Q_i}} \\ \mathrm{K_i} &= \mathrm{E_o W_{K_i}} \\ \mathrm{V_i} &= \mathrm{E_o W_{V_i}} \end{aligned} \tag{2}$$

Where $\mathrm{W_{Q_i}}, \mathrm{W_{K_i}} \in \mathbb{R}^{\mathrm{d \times d_k}}, \mathrm{W_{V_i}} \in \mathbb{R}^{\mathrm{d \times d_k}}$, $\mathrm{d_k}, \mathrm{d_V}$ show the dimension of one attention head

$$\mathrm{Attention\ (Q_i, K_i, V_i) = softmax\ (\frac{Q_i K_i^T}{\sqrt{d_k}}) V_i} \tag{3}$$

We take the sum of the muti-head self-attention part and the original feature through residual connect, and then through the linear mapping of the last layer, re-project the dimension of the vector to d (the origin embedding dimension) to get $\mathrm{O_{vec}}$ Which can be expressed as

$$\mathrm{O_{vec}} = \mathrm{Relu(liner(\sigma(Multihead(E_o) + E_o W_r)))} \tag{4}$$

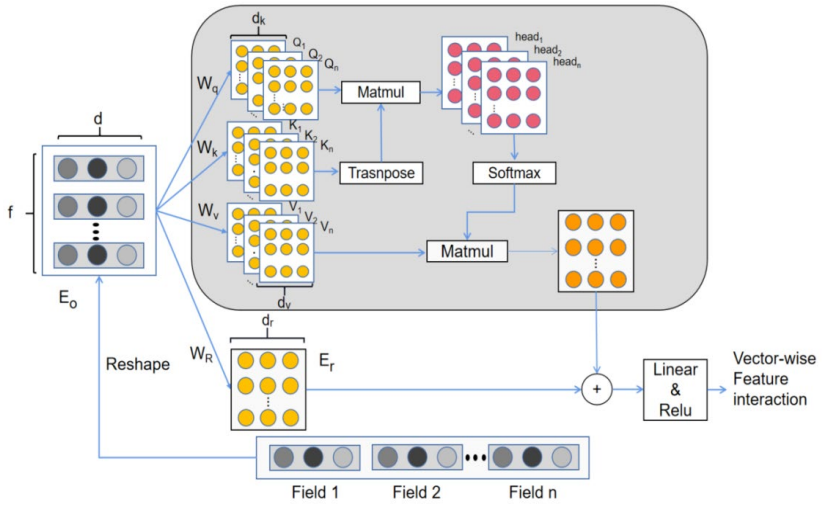

Figure 2: The network structure of the vector-wise part

ISSN 2616-5775 Vol. 7, Issue 1: 18-23, DOI: 10.25236/AJCIS.2024.070103

3.2. Vector reweight

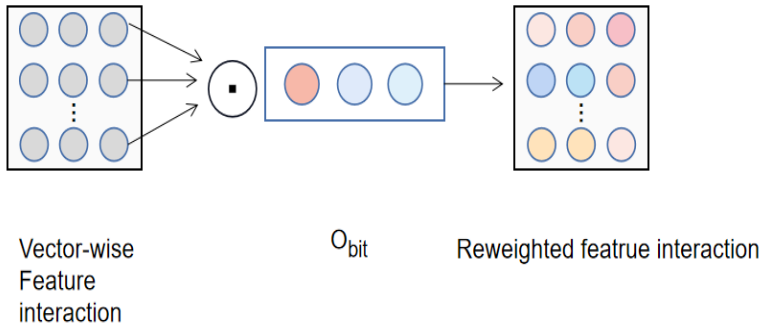

Figure 3: The network structure of vector reweight part

$$O_{bit} = \mathrm{PReLU}(W_L h_L + b_L) \tag{5}$$

Through the structure of Figure 2, vector-wise feature interaction is obtained. Then Figure 3 shows the process of vector reweight.which can be expressed as:

$$W_F = O_{vec} \odot O_{bit} \in \mathbb{R}^{f \times d} \tag{6}$$

The resulting final output $W_F$includes both the global attention and the bit-level fine-grained weights extracted by the MLP, resulting in a more comprehensive feature representation. Its dimensions are the same as the original embedding representation

3.3. Vector reweight Auto feature select gate

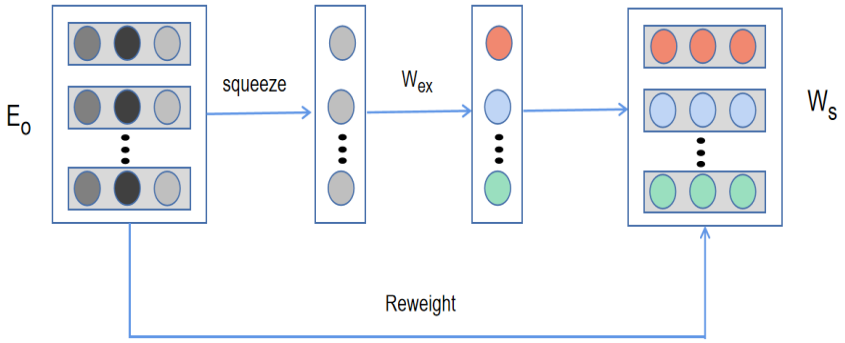

Figure 4: The network structure of SENet part

Different from existing CTR estimation methods, feature interaction is often transmitted directly to the output layer, or it is only a simple residual connection with the original feature. We believe that effective combination of original features and higher-order features can better represent the importance between features and the context connection. We use SENet to get this allocation weight, which can effectively extract the feature importance information of each field.

The structure of SENet is shown in Figure 4, and the final output of AFSgate can be expressed as(5).

$$E_{output} = \sigma(W_s) \odot E_o + (1 - \sigma(W_s)) \odot W_{F'} \tag{7}$$

ISSN 2616-5775 Vol. 7, Issue 1: 18-23, DOI: 10.25236/AJCIS.2024.070103

4. Experiments

4.1. Performance Comparison

Table 1: Overall accuracy comparison in the three datasets

	Cretio		Frappe		Avazu	
Model	AUC	Logloss	AUC	Logloss	AUC	Logloss
FM	0.8032	0.4510	0.9704	0.1937	0.7781	0.3828
DeepFm	0.8061	0.4443	0.9786	0.1811	0.7786	0.3810
DCN	0.8012	0.4602	0.9785	0.1816	0.7681	0.3940
xDeepFM	0.8088	0.4456	0.9792	0.1991	0.7808	0.3818
AutoInt	0.8067	0.4448	0.9751	0.2421	0.7752	0.3824
FiBiNET	0.8096	0.4432	0.9789	0.1822	0.7832	0.3786
DIFM	0.8087	0.4455	0.9783	0.1808	0.7822	0.3790
FRNet	0.8124	0.4408	0.9824	0.1586	0.7851	0.3602
Ours	0.8135	0.4388	0.9845	0.1573	0.7864	0.3623

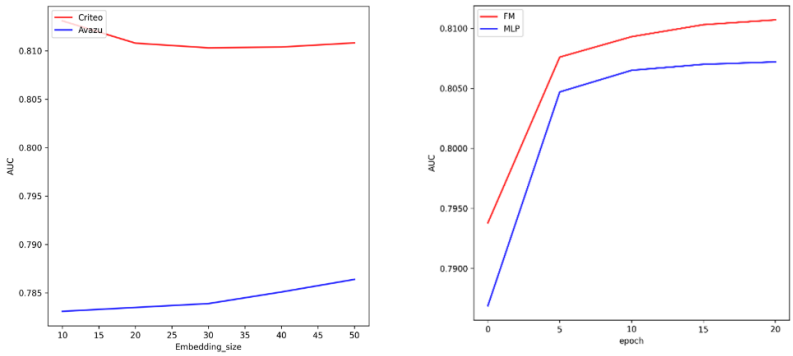

Figure 5: The performance of different embedding_size and output layers

Table 1 shows a comparison between our approach and eight other CTR models based on three common datasets: Criteo, Avazu, and Frappe. And all of our experiments are based on two metrics: AUC and Logloss.

Compared with DIFM, AUC increased by 0.48% and loss decreased by 0.67% in the Criteo dataset, which is already a considerable improvement in CTR prediction. This strongly proves that our method can extract more effective feature interaction information. Overall, our approach outperforms other models on all datasets

In the first figure in Figure 5, we try to adjust the embedding size range from 10 to 50, and the experiment shows that 10 is the best size for Criteo and 50 is the best size for Avazu. In the second figure, we try the performance of different output layers on Criteo, and the results show that FM is better than MLP, which indicates that the explicit feature interaction has better expression ability as output layer.

4.2. Ablation Study

After comparing the different models, we further study the contribution of each component to the overall model. As shown in Table2, RAN, AFSgate and FM layers all contribute to the performance of the overall model, among which the removal of Vector reweight from RAN has the greatest impact

Table 2: Ablation study on Criteo

	AUC	Logloss
BASE	0.8315	0.4388
NO-Reweight	0.8084	0.4426
NO-AFSgate	0.8103	0.4408
NO-FM	0.8115	0.4402

ISSN 2616-5775 Vol. 7, Issue 1: 18-23, DOI: 10.25236/AJCIS.2024.070103

5. Conclusions

This paper introduces two innovative models designed for click-through rate (CTR) prediction, namely the Reweighted Attention Net (RAN) and AFSgate. RAN can learn More fine grained representations of high-order feature interaction with the help of the the fusion of Multi-Head Self-Attention and MLP. In addition, AFSgate has made outstanding contributions to the fusion of higher-order feature interactions and original features. Experiments show that combining these two structures with FM performs better on three common datasets than the current mainstream models. Furthermore,the ablation experiment also verified the validity of each module in our method.

References

[1] WANG R,Deep & Cross Network for Ad Click Predictions,in: Proceedings of the ADKDD'17, 2017, pp.1-7.

[2] ZHOU G, Deep Interest Network for Click-Through Rate Prediction,in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018,pp. 1059–1068.

[3] SU R, Parallel heterogeneous network with soft gating for CTR prediction,in: Artificial Intelligence. CICAI 2022. Lecture Notes in Computer Science,2022.pp.413-424.

[4] Paul Covington, Deep neural networks for youtube recommendations, In Proceedings of the 10th ACM conference on recommender systems.2016,pp. 191–198.

[5] Huifeng Guo, DeepFM: a factorization-machine based neural network for CTR prediction, In Proceedings of the 26th International Joint Conference on Artificial Intelligence,2016, pp.1725–1731.

[6] LIAN J. xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems.in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018.pp.1754-1763.

[7] W. Song, Autoint: Automatic feature interaction learning via self-attentive neural networks, in: Proceedings of the 28thACM International Conference on Information and Knowledge Management, 2019, pp.1161-1170

[8] S.Rendle, Factorization machines, in: 2010 IEEE International conference on data mining. IEEE,2010,pp.995-1000

[9] ZHAO Z. FINT: Field-aware Interaction Neural Network for CTR Prediction. In: Cornell University - arXiv, 2021,doi: 2107.01999.

[10] HUANG T. GateNet: Gating-Enhanced Deep Network for Click-Through Rate Prediction.in: Cornell University - arXiv: Learning, 2020.doi: 2007.03519.

[11] LU W. A Dual Input-aware Factorization Machine for CTR Prediction.in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, 2021, pp.3139-3145.

[12] F. A. Gers, J. Schmidhuber and F. Cummins, Learning to Forget: Continual Prediction with LSTM: in Neural Computation, vol. 12, no. 10, pp. 2451-2471, 1 Oct. 2000, doi: 10.1162/ 089976600300015015.