scientific reports

Check for updates

OPEN

Session interest model for CTR prediction based on self-attention mechanism

Qianqian Wang[1,3], Fang'ai Liu[2]✉, Xiaohui Zhao[2] & Qiaoqiao Tan[2]

Click-through rate prediction, which aims to predict the probability of the user clicking on an item, is critical to online advertising. How to capture the user evolving interests from the user behavior sequence is an important issue in CTR prediction. However, most existing models ignore the factor that the sequence is composed of sessions, and user behavior can be divided into different sessions according to the occurring time. The user behaviors are highly correlated in each session and are not relevant across sessions. We propose an effective model for CTR prediction, named Session Interest Model via Self-Attention (SISA). First, we divide the user sequential behavior into session layer. A self-attention mechanism with bias coding is used to model each session. Since different session interest may be related to each other or follow a sequential pattern, next, we utilize gated recurrent unit (GRU) to capture the interaction and evolution of user different historical session interests in session interest extractor module. Then, we use the local activation and GRU to aggregate their target ad to form the final representation of the behavior sequence in session interest interacting module. Experimental results show that the SISA model performs better than other models.

It is a critical problem to predict the probabilities of users clicking on ads or items for many applications such as online advertising or recommender systems[1,2]. Cost per click (CPC)[3] model is often used in advertising system. The accuracy of click-through rate (CTR) can influence the final revenue in CPC model. At the same time, displaying suitable advertisement to user can enhance their experience. Therefore, both academia and industry are concerned about how to design the CTR prediction models.

Modeling feature interaction is very critical in CTR prediction tasks. Recently, some effective methods ignore to capture user interest. User interest has an important influence on CTR prediction. In the fields with rich internet-scale user behavior data, such as online advertising, user's sequential behaviors reflect user evolving interests. However, some models based on user interest overlook the intrinsic structure of the sequences. Multiple sessions make up a sequence. A session is a list of user behaviors that occur within a given time frame. The user behavior in each session is highly homogeneous, and the user behavior in different sessions is heterogeneous. Grbovic et al.[4] found the session division principle that there is a time interval of more than 30 min. For example, the user mainly browses the T-shirt in the first half an hour as session 1, and browses the sneakers in the second half an hour as session 2. User shows different interests in session 1 and session 2.We can know the fact that people has a clear and unique intent at a session, but the interest usually changes when user start a new session.

Through the above observation, we propose Session Interest Model via Self-Attention (SISA) for CTR prediction, which uses multiple historical sessions to simulate the user's sequential behavior in the CTR prediction task. At session division module, we naturally divide the user sequential behavior into sessions. At session interest extractor module, a self-attention mechanism with bias coding is applied to model each session. Self-attention mechanism gets the internal relationship of each session behavior, then, extracts the user interest in each session. Since different session interest may be related to each other or follow a sequential pattern, we use Gated Recurrent Unit (GRU) to capture interaction and evolution of user different historical session interests at session interacting module. Because different session interests have different effects on the target item, we utilize attention mechanism to achieve local activation and use GRU to aggregate their target ad to get the final representation of the behavior sequence.

The main contributions of this paper are as follows:

[1]Shandong Women's University, Jinan, China. [2]Shandong Normal University, Jinan, China. [3]Shandong Provincial Key Laboratory of Network Based Intelligent Computing, Jinan, China. ✉email: lfa_sdnu@163.com

1. The user behavior in each session is highly homogeneous, and the behavior of user in different sessions is heterogeneous. We focus on user multiple session interest and propose a novel session interest model via self-attention (SISA). We can get more expressions of interest and more accurate prediction results.
2. We specially divide the user sequential behavior into sessions and design session interest extractor module. We employ a self-attention network with bias coding to obtain an accurate expression of interest for each session. Then we use GRU to capture the interaction of different session interests. At the same time, we exploit GRU with attentional update gate (AUGRU) to aggregate their target ad to find the influences of different session interests at session interacting module.
3. The experimental results show that our proposed model has great improvements over other models. At same time, the influence of key parameters and different variants is also explored, which proves the validity of the SISA model.

The rest of the paper is organized as follows. We discuss the related work in "Related work» and introduce the detailed architecture of proposed SISA model in "Material and methods". Section "Experiments" verify the prediction effectiveness of the proposed model through experiments, and analyse the results. Furthermore, in "Conclusion", we summarize the model presented in this paper and introduce the direction of future work.

Related work

The CTR prediction problem is normal formulated as a binary classification problem. Logistic regression (LR)[5] is a linear model that is used in the industry. Kumar et al.[6] used logistic regression to establish a model for CTR prediction. McMahan et al.[7] proposed a method to solve the Google's ad problem and got better performance. Multiple features are used as input data such as ad information and keywords. Chapelle et al.[8] used LR to solve the problem of prediction for Yahoo's. The linear model is easy, but it can not capture feature interaction. To overcome the limitation, Factorization Machine[9] (FM) and its variants[10] are used to capture feature interactions and get better results. The field-aware factorization machines (FFM) introduced field aware latent vectors to capture feature interaction. However, the Factorization Machine model is relatively weak in obtaining high-order feature interaction. He et al.[11] utilized decision trees and LR to improve the result. However, these models use shallow layer that have limited representation power of feature interactions.

Recently, deep neural networks have achieved great success in many research fields such as in computer vision[12,13], image processing[14,15] and natural language processing[16,17]. Therefore, researchers have proposed many CTR prediction models based on deep learning. How to effectively model feature interaction is an important problem in most models. Zhang et al.[18] proposed the Factorization Machine based Neural Network (FNN). The model uses FM to pre-train the embedding layer based on forward neural network. FNN model has a better performance on capturing high-order feature interactions. Cheng et al.[19] combines the linear and the deep neural network to capture feature interactions. The wide part of the model is still needed feature engineering. This means that feature interaction also needs to be designed manually. To solve the problem, DeepFM model[20] uses FM to replace the wide part, and shared the same input. DeepFM model is considered to be the more advanced model in the field of CTR estimation. Product-based Neural Networks (PNN) model[21] is used for user response prediction. The model utilizes a product layer and gets feature interaction. Lian et al.[22] proposed a CIN model, and captured feature interactions at the vector-wise level. Deep and Cross Network (DCN)[23] efficiently learns feature crossing and no manual feature engineering. Shan et al.[24] found the relationship behind the user behavior based on the residual neural networks and proposed the deep crossing model[25]. In addition, some models also are proposed based on Convolutional Neural Networks (CNN). Kim et al.[26] designed a multi-array CNN Model for ad CTR Prediction. This method can capture local feature information based on CNN. Wang et al.[27] used CNN based on attention to find the different features. Ouyang[28] considered each target ad independently and proposed MA-DNN model that achieved a better result.

In practical applications, different predictors usually have different predictive capabilities. Features that have a greater contribution to the prediction results should be given greater weights. As we all know, the attention mechanism[29] has a powerful function in distinguishing importance of features. Wang et al.[30] improves FM based on the attention mechanism to find the different importance of different features. Cao et al.[31] proposed a Meta-Wrapper model that utilized the attention mechanism and capture the user interested items in historical behaviors. Xiao et al.[32] builds Attentional Factorization Machine (AFM) model, the model can mine feature interaction based on neural attention network. However, the model ignored the important of user behavior for CTR prediction.

In summary, the high-order expression and interaction of features significantly improves the expression ability of features and the generalization ability of the models. However, in the process of capturing feature interactions, the influence of user interest is often ignored. Constructing a model to capture the user's dynamics and evolving interests from the user's sequential behavior has been widely proven effective in CTR prediction tasks. At the same time, Dynamic Quality of Service (QoS) prediction for services is currently a hot topic and a challenge for research in the fields of service recommendation and composition. Jin et al.[33] addresses the problem with a Time-aWare service Quality Prediction method (named TWQP). Deep Interest Network (DIN)[34] introduced the influence of user interests and found user interests based on user behaviors. DIN can capture the diversity characteristic of user interests and improve the performance of CTR prediction. In order to capture the dynamic evolution of user interests, Deep Interest Evolution Network (DIEN)[35] was proposed. DIEN gets interest features and finds interest evolving process. Wang et al.[36] presented a Trust-based Collaborative Filtering (TbCF) algorithm to perform basic rating prediction in a manner consistent with the existing CF methods. The algorithm employs multiple perspectives to extract proper services and achieves a good tradeoff between the robustness, accuracy, and diversity of the recommendation. Liu et al.[37] proposed an attention-based bidirectional

gated recurrent unit (GRU) model for point-of-interest (POI) category prediction (ABG_poic). They regard the user's POI category as the user's interest preference because the fuzzy POI category is easier to reflect the user interest than the POI. By modeling the user's sequential behavior, the feature representation is enriched, and the prediction accuracy is significantly improved.

The concept of session often appears in sequential recommendation, but it is rarely seen in CTR prediction tasks. Session-based recommendation achieves good results via user dynamic interest evolving. Neural Attentive Recommendation Machine (NARM)[38,39] used an attention mechanism to capture the user purpose in the current session. Zhang et al.[40] analyzes the current session information from multiple aspects and improves user satisfaction. However, most existing studies for CTR prediction ignore that the sequences are composed of sessions. Upon all these perspectives, we introduce a novel session interest model via self-attention (SISA) to get a better result for CTR.

Material and methods

We propose Session Interest Model via Self-Attention (SISA) for CTR prediction, which uses multiple historical sessions to simulate the user's sequential behavior in the CTR prediction task. The SISA model includes five modules, we describe session interest model via self-attention in this section. We first introduce feature representation and embedding in "Feature representation and embedding". Next, "Session division module" illustrates the session division module. Then, we describe the session interest extractor module in "Session interest extractor module" and session interacting module in "Session interest interacting module". Finally, "The overall architecture of SISA model" presents the structure of SISA model.

Feature representation and embedding. We use four groups of features (User Profile, Scene Profile, Target Ad, User Behavior) as input data for the model. The encoding vector of the feature group can be expressed by $E \in \mathbb{R}^{M \times d_{\text{model}}}$, where d_{model} is the embedding size and M is the size of sparse features. Through feature embedding, User Profile can be represented by $X^U \in \mathbb{R}^{N_u \times d_{\text{model}}}$, where N_u is the number of User Profile sparse features. Similarly, both Scene Profile and Target Ad can be expressed as $X^S \in \mathbb{R}^{N_s \times d_{\text{model}}}$,$X^I \in \mathbb{R}^{N_i \times d_{\text{model}}}$, where N_s and N_i are the number of Scene Profile and Target Ad sparse features, respectively. User Behavior is represented by $X=[x_1; ...; x_i; ...x_N] \in \mathbb{R}^{N \times d_{\text{model}}}$, where N is the number of user historical behaviors and x_i is the embedding of the i-th behavior.

Session division module. In order to get the user session interests, we divide the user behavior sequences X into sessions S, where the k-th session $S_k = [x_1; ...; x_i; ...x_T] \in \mathbb{R}^{T \times d_{\text{model}}}$,$T$ is the number of behaviors in each session and b_i is user i-th behaviors in current session. Many behaviors that are more than 30 min apart into user sessions follow by Grbovic's[4] method.

Session interest extractor module. On the one hand, the behaviors in the same session are closely related to each other, on the other hand, the user random behavior in the session deviates from the original expression of session interest. In order to capture the inner relationship between behaviors in the same session and find the impact of those irrelevant behaviors, a multi-head self-attention mechanism[41] is used for each session. At the same time, we apply positional encoding to the embedding based on self-attention mechanism.

So as to take advantage of order relationship of the sequence, self-attention mechanism is used to positional encoding to the input embedding. Also, we capture the sequence of sessions and the bias in different representation subspaces. So, we define bias encoding BE as follows:

$$\text{BE}(u,k,t) = W_u^U + W_k^K + W_t^T \tag{1}$$

where $W^U \in \mathbb{R}^{d_{\text{model}}}$ is the bias vector of the unit position in the behavior embedding, and u is the index of the unit in the behavior embedding. $W^K \in \mathbb{R}^K$ is the bias vector of session, k is the index of session. $W^T \in \mathbb{R}^T$ is the bias vector of the position in the session. So user behavior sessions can be represented as follows via bias encoding:

$$\text{S} = \text{S} + \text{BE} \tag{2}$$

As we all know, the user's click behavior is influenced by many factors, such as color, shape, and price. Multi-head self-attention can get the relationship in different representation subspaces. We use $S_k = [S_{k1}; ...; S_{kn}; ...S_{kN}]$, where $S_{kn} \in \mathbb{R}^{T \times d_n}$ is the n-th head of S_k.N is the number of heads, $d_n = \frac{1}{n} d_{\text{model}}$.Through these representations, we can calculate the output of $head_n$ as follows:

$$head_n = Attention(S_{kn}W^Q, S_{kn}W^K, S_{kn}W^V) \tag{3}$$

$$Attention(Q,K,V) = softmax\left(\frac{S_{kn}W^Q W^{K^T} S_{kn}^T}{\sqrt{d^{model}}}\right) S_{kn}W^V \tag{4}$$

where W^Q, W^K, W^V are weight matrices. Then bias encoding-based feedforward neural network can further improve the nonlinear ability:

$$I_k^S = FNN(Concat(head_1, ...head_N)W^O) \tag{5}$$

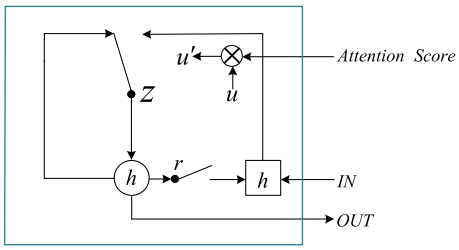

Figure 1. The architecture of AUGRU.

$$I_k = Avg(I_k^S) \tag{6}$$

where W^O is the weight matrix. $FNN(\cdot)$ is the feedforward neural network. Avg$(\cdot)$ is the average pooling. I_k is the user k-th session interest.

Session interest interacting module. The different session interest may be related to each other or follow a sequential pattern, GRU performs better at capturing sequential relationships, so we use GRU[42,43] to capture the interaction and evolution of user different historical session interests.

The activation h_t of the GRU at time t is a linear interpolation between the previous activation h_{t-1} and the candidate activation $\tilde{h}_t$:

$$h_t = (1 - z_t)h_{t-1} + z_t\tilde{h}_t \tag{7}$$

The update gate is represented by:

$$z_t = \sigma(W_z\mathrm{I}_t + U_z h_{t-1} + b_z) \tag{8}$$

Candidate activation $\tilde{h}_t$ is calculated as follows:

$$\tilde{h}_t = \tanh(WI_t + u(r_t \odot h_{t-1}) + b_h) \tag{9}$$

where r_t is a set of reset gates and $\odot$ is an element-wise multiplication. When off (r_t close to 0), the reset gate makes it forget the state of the previous calculation.

At the same time, the r_t is represented as follows:

$$r_t = \sigma(W_r\mathrm{I}_t + U_r h_{t-1} + b_r) \tag{10}$$

where σ is a sigmoid function, I_t is the input of GRU, W and b are the parameters that are trained.

The hidden state h_t can capture the dependency between session interests. However, the user's session interest related to the target ad has a greater impact on whether the user will click on the target ad. So the weight of the user's session interest needs to be reassigned to the target ad. We use attention mechanism for local activation and use GRU to model the representation of session interests and target ad.

We use I'_t,h'_t represent the input and hidden state of GRU. The input of second GRU is corresponding state in the part of capturing session interest interaction: I'_t=h_t. The attention function we used can be formulated as:

$$a_k^I = \frac{\exp(I_k W^I X^I)}{\sum_k^K \exp(I_k W^I X^I)} \tag{11}$$

where W^I has the corresponding shape, attention score can reflect the relationship between target ad X^I and input h_t, and the more relevant ones will get more attention weight.

We combine attention mechanism and GRU and use the GRU with attentional update gate (AUGRU) to consider influences of between session interests and the target ad:

$$\tilde{u}'_t = a_k^I * u'_t \tag{12}$$

$$h'_t = (1 - \tilde{u'}_t) \circ h'_{t-1} + \tilde{u'}_t \circ \tilde{h'}_t \tag{13}$$

where u'_t is the original update gate of AUGRU, $\tilde{u}'_t$ is the attentional update gate that we use in AUGRU. h'_t, h'_{t-1} and $\tilde{h'}_t$ are the hidden states of AUGRU. Figure 1 is the framework of GRU application attention mechanism (AUGRU).

By using AUGRU, we retain the original dimension information of the update gate. We measure all dimensions of the update gate by using attention score and consider the impact of different session interests on the target ad.

The overall architecture of SISA model. The SISA model includes five modules, and the structure is shown in Fig. 2.

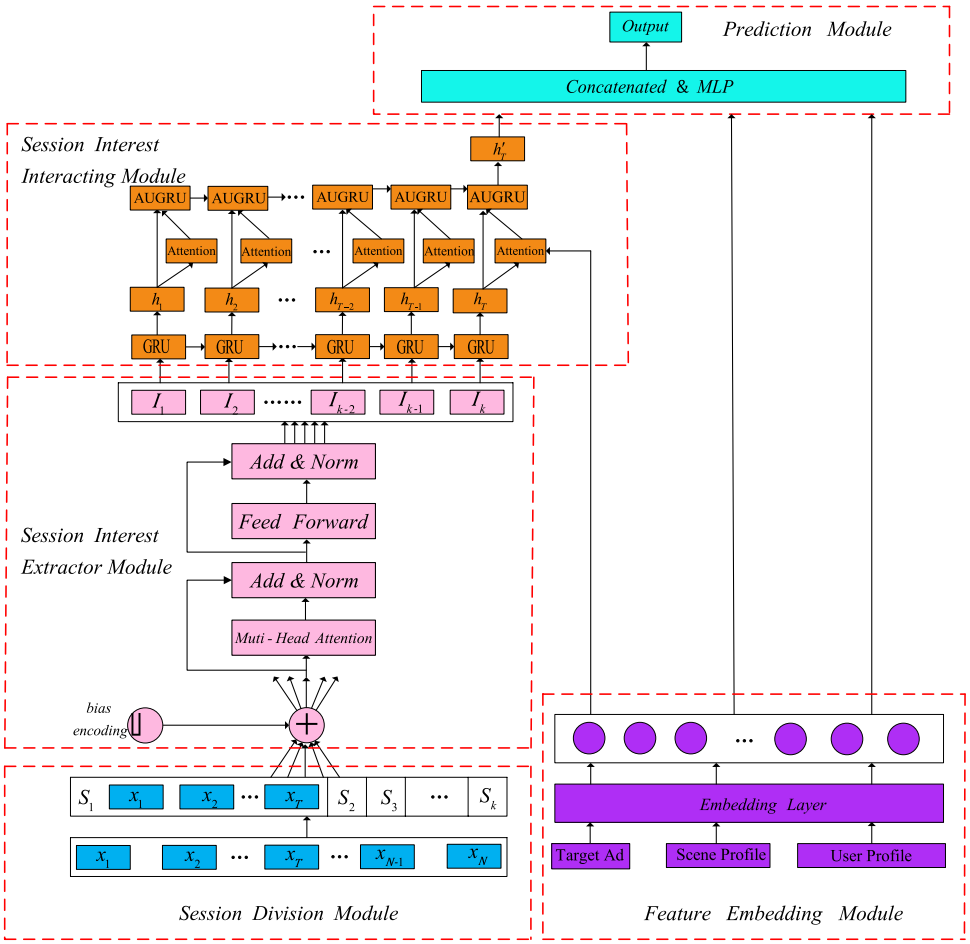

Figure 2. The structure of SISA.

In feature representation and embedding module, informative features such as user profile, scene profile and target ad are transformed into dense vectors by using an embedding layer. In session division module, we divide user's behavior sequences into sessions and get use's session sequence. In session interest extractor module, we capture the inner relationship between behaviors in the same session. We employ multi-head self-attention to reduce the influence of unrelated behaviors. We also apply positional encoding to the input embedding based on self-attention mechanism and get order relations of the sequence. In session interest interacting module, we use GRU to capture the interaction and evolution of user different historical session interests. The user's session interest related to the target ad has a greater impact on whether the user will click on the target ad. So we use AUGRU to model the representation of session interests and target ad. In prediction module, embedding of sparse features and session interests that we capture are concatenated and then imported into MLP. Finally, the softmax function is used to get probability that people click on the ad.

The loss function is a negative log-likelihood function and is usually expressed as:

$$L = -\frac{1}{N}\sum_{(x,y)\in D}^{N}(y\log p(x) + (1-y)\log(1-p(x))) \tag{14}$$

where D denotes the training size N, $p(x)$ denotes the probability that the user clicks on an ad.

Experiments

Experiments setting. *Datasets.* In order to verify the effectiveness of SISA model, we conducted the experiments on two subsets of Amazon dataset[44]: Books and Electronics and two public datasets: Avazu and Criteo. In CTR model evaluation, most researchers often use Criteo dataset. The result of the different dataset is showed in Table 1. The datasets are randomly divided into three parts: training set (80%), validation set (10%) for adjusting hyperparameters, and test set (10%).

Evaluation metrics. In our experiment, we employ three metrics: AUC (Area Under ROC), Logloss and RMSE (Root Mean Square Error).

Dataset	Users	Items	Features	Samples
Books	53,126	46,783	48,632	297,659
Electronics	32,359	36,191	39,715	203,114
Avazu	80,724	71,473	127,694	854,261
Criteo	30,023	33,871	46,723	210,342

Table 1. Basic statistics of the datasets.

AUC: The area under the ROC curve is a more commonly used indicator for evaluating classification problems (such as CTR prediction)[45]. The value of AUC is larger, the result is better.

Logloss: Logloss is applied to calculate the distance in a binary classification problem. The value of logloss is smaller, the performance of the model is the better.

RMSE: RMSE[46] can be defined as follows:

$$RMSE = \sqrt{\frac{1}{|T|}\sum_{i}(y_i - \hat{y}_i)^2} \tag{15}$$

where y_i^t is the observed scores and $\hat{y}_i^t$ is the value of prediction, T is the testing set. The score of RMSE is small, the result of the model is great.

Parameter settings. We employ dropout to prevent over-fitting in the neural networks and the value of dropout rate is 0.4. We set the size of the hidden state in the GRU is 56. At the same time, we use $10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}$ as learning rates to test. Also, different number of neurons from 100 to 800 is employed.

Comparisons with different models. To verify the efficiency of the SISA that we proposed, we compare SISA with some mainstream CTR prediction model. It shows the results of the different models for AUC in Fig. 3. Tables 2 and 3 show the value with logloss and RMSE, respectively. Through comparison, many aspects can be seen.

1. Wide&Deep[19] is a model that has wide part and deep part. The linear part uses manually designed cross-product feature for better interactions. However, the wide part can not get the features automatically. So the Wide&Deep model does not have better performance in all the models.
2. FNN[18] is a model that can get high-order features and uses deep learning to automatically learn feature interactions. Also, it improves the FM. However, when use the DNN, the model needs to train FM, so it has limitations.
3. AFM[32] can find the importance of different feature interactions and use neural network to capture the feature interactions. As we all know, different feature interaction has different useful for results. The performance of AFM is better. This can be verified that using the attention mechanism can enhance performance of the model.
4. DeepFM[20] is a new network framework that combines the FM and deep neural networks. It can model low-order feature interactions like FM and model high-order feature interactions like deep neural networks. DeepFM can be trained without any feature engineering. So the model outperforms both Wide&Deep and FNN.
5. DIN[34] is a model for CTR prediction that exploits the rich historical behavior data to extract user interest. The model captures the feature interactions based on neural networks. It is based on the attention mechanism to get the representation of the user behavior and target ad. DIN has the better performance than DeepFM.
6. ADI[35] captures interest evolving processes from user behaviors and gets higher prediction accuracy. However, the SISA model performs better than others. In SISA, we partition user behavior sequences into multiple sessions, user session interests follow a sequential pattern and more suitable for modeling. We can see that SISA model based on user session interest improves accuracy in all datasets. As we can see that in Avazu dataset the SISA increased by 1.8% compared to other models.

Sensitivity analysis of the model parameters. We explore the influence of different parameters for the results of SISA model, such as the epoch, the number of neurons per layer, and the dropout rate β.

Dropout is the probability of neurons remaining in the network. First we set β to be 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8. As shown in Fig. 4, the β can help SISA learn powerful features. The β is properly set (from 0.5 to 0.8), the SISA model is able to reach its best performance at all datasets. However, with an increasing of the value of β, the performance of SISA shows a downward trend. So we choice β=0.6 in the following experiment.

When other factors remain the same, we study the effect of different number of neurons. In Fig. 5, it is not that the higher the number of neurons, the better the result. When the number of neurons is 500, 600 or 700, the performance of SISA is stably and even worse in all datasets. It is because that the model is overfit. So we select 400 in the experiment.

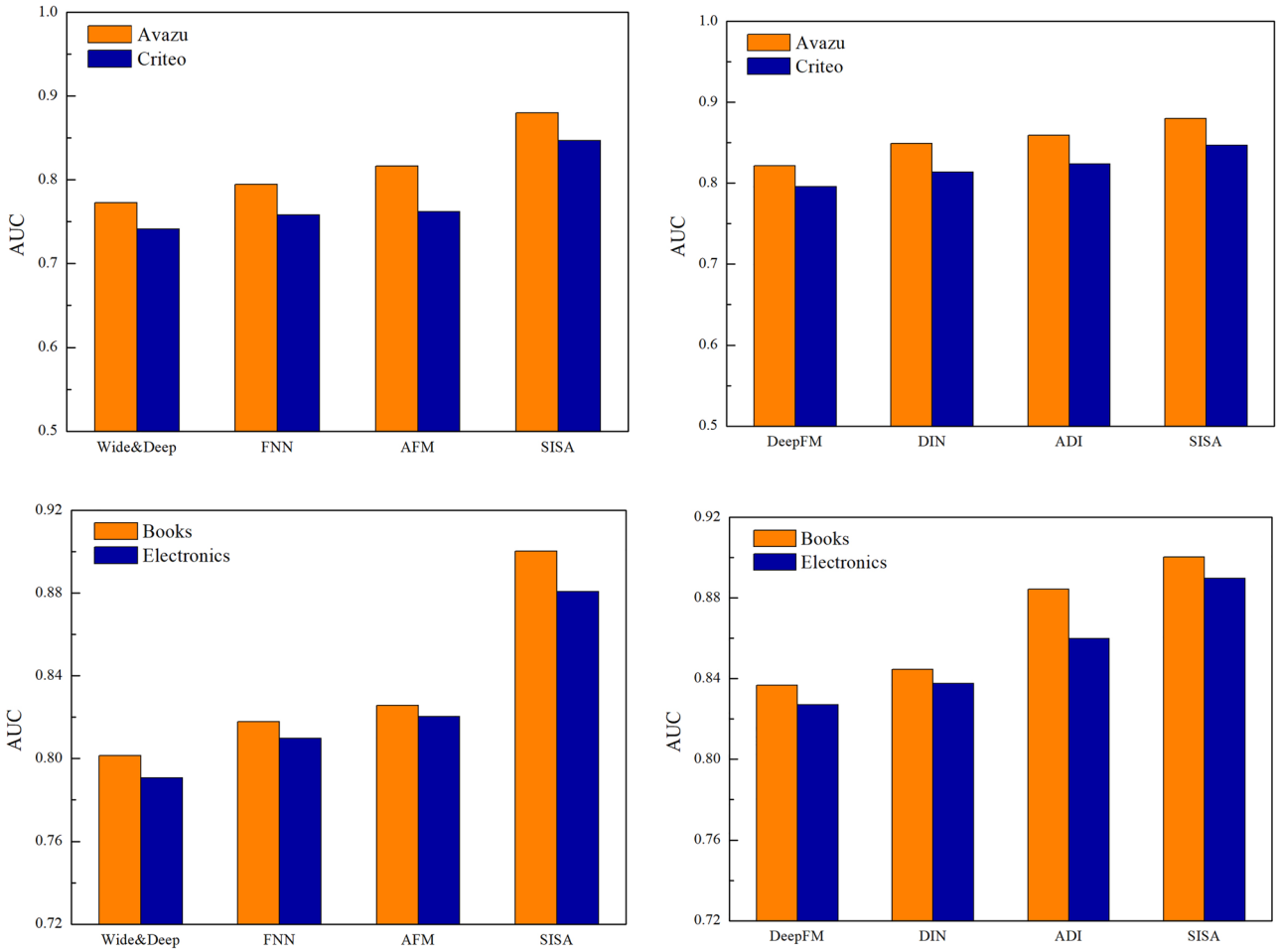

Figure 3. AUC performance comparison with other model.

Model	Logloss			
	Books	Electronics	Avazu	Criteo
Wide&Deep	0.1213	0.1279	0.2237	0.3543
FNN	0.1201	0.1268	0.2204	0.3529
AFM	0.1197	0.1235	0.2189	0.3482
DeepFM	0.1134	0.1192	0.2145	0.3438
DIN	0.1123	0.1146	0.2107	0.3407
ADI	0.1014	0.1077	0.2083	0.3372
SISA	0.0978	0.1006	0.1964	0.3286

Table 2. Overall CTR prediction for Logloss performance in different datasets.

Model	RMSE			
	Books	Electronics	Avazu	Criteo
Wide&Deep	0.5071	0.5224	0.5326	0.6072
FNN	0.5043	0.5202	0.5302	0.5986
AFM	0.4996	0.5138	0.5279	0.5823
DeepFM	0.4942	0.5079	0.5213	0.5761
DIN	0.4578	0.4693	0.5185	0.5625
ADI	0.3675	0.4267	0.5061	0.5485
SISA	0.3121	0.3783	0.4925	0.5306

Table 3. Overall CTR prediction for RMSE performance in different datasets.

We study the influence of different epoch values for CTR prediction through experiments. Can see from Fig. 6, the model performed poorly when the value of epoch is 0–5. This is because the number of iterations is too small to determine the appropriate parameters. As the value of epoch increases, the value of RMSE becomes smaller. Compared with other datasets, the value of RMSE in the Amazon dataset fluctuates relatively high. Because model needs different numbers of features to train on different datasets, and the diversity of the data will cause some errors. If the value of epoch is not suitable, the performance of the model will fluctuate greatly. At the same time, the model has better performance when the value of epoch is between 10 and 20 in Fig. 6. Therefore, in the experiment we set the value of epoch to 16.

Comparison among SISA variants. Although we have demonstrated strong empirical results, the results presented do not isolate the specific contributions of each component of SISA, so we conduct ablation experiments on SISA. In Table 4, SN stands for a network like the one used in FNN. IN stands for complex network that can get user session interest. AVG represents average pooling and MAX is maximum pooling strategies. The self-attention module uses AVG and MAX instead, and ATT-IN stands for the SISA model.

In Table 4, the IN-ATT model has the highest AUC value. FNN model uses neural networks to automatically capture feature interactions. The model of SN-AVG, -MAX, and -ATT ignore the impact of interest on click-through rate prediction, not as good as a model based on session interest. At the same time, some models do not distinguish the contribution of different features to the prediction results, so the performance of prediction is not good. SISA model has achieved better results by combining the user session interest with the self-attention mechanism.

Conclusion

In this paper, we propose a new model, namely SISA, to model user session interest for CTR prediction. First, we specially divide the user sequential behavior into sessions and design session interest extractor module. We employ a self-attention network with bias coding to obtain an accurate expression of interest for each session. Then, we use GRU to get the interaction of different session interests. We exploit GRU with attentional update gate (AUGRU) to aggregate their target ad to find the influences of different session interests at session interacting module. Next, embedding of sparse features and session interests that we capture are concatenated and fed into MLP in prediction module. Experiment results prove the efficiency of SISA on different datasets. In the feature, we will pay attention to use knowledge graph to capture user interests and model click-through rate predictions.

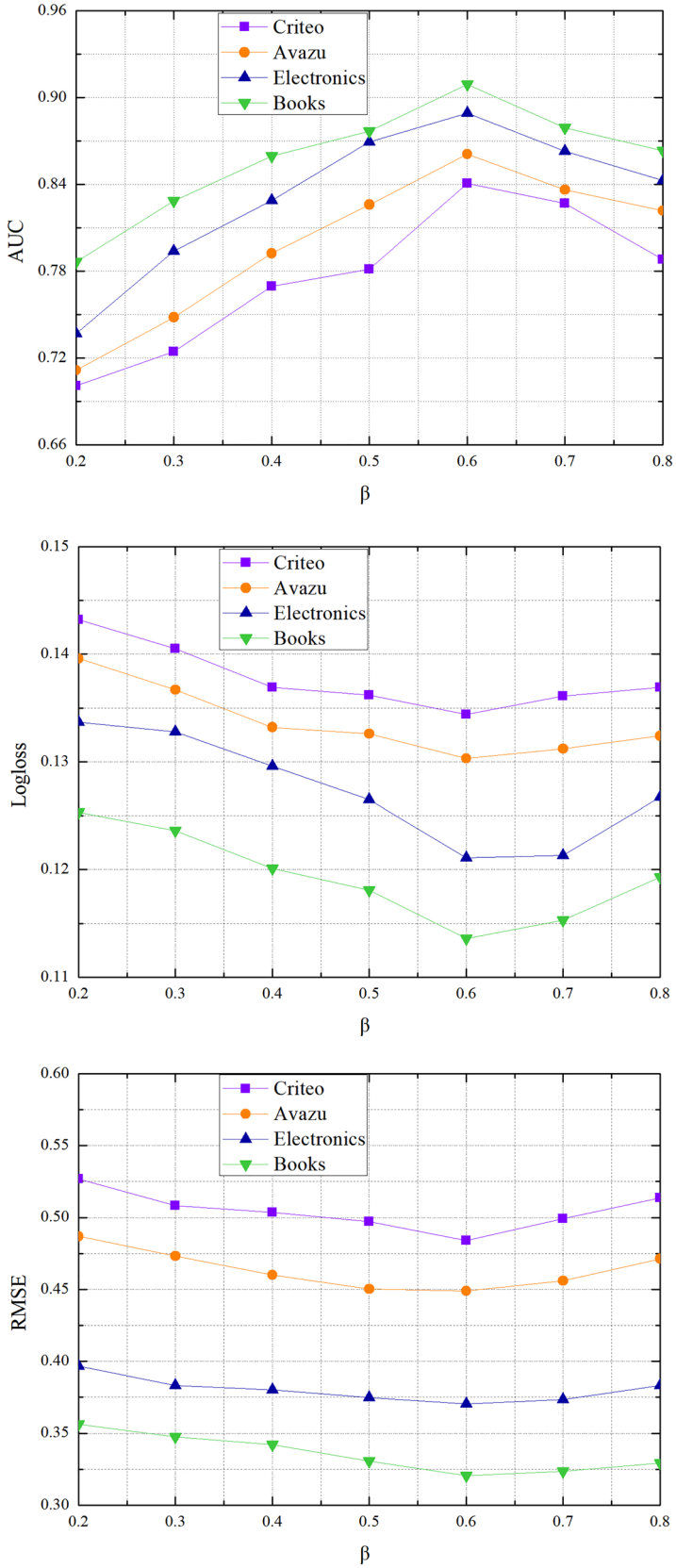

Figure 4. Performance comparisons w.r.t. the dropout rate β.

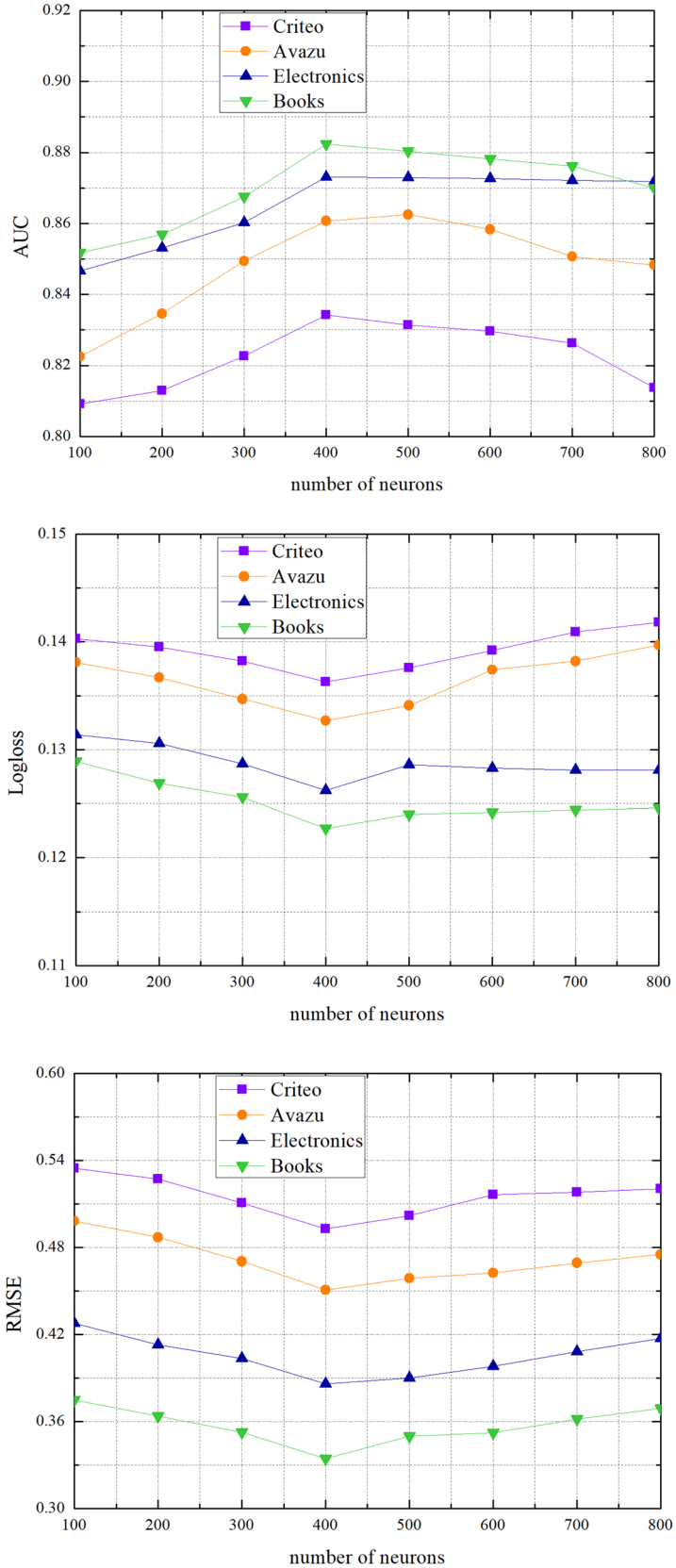

Figure 5. Performance comparisons w.r.t. the number of neurons.

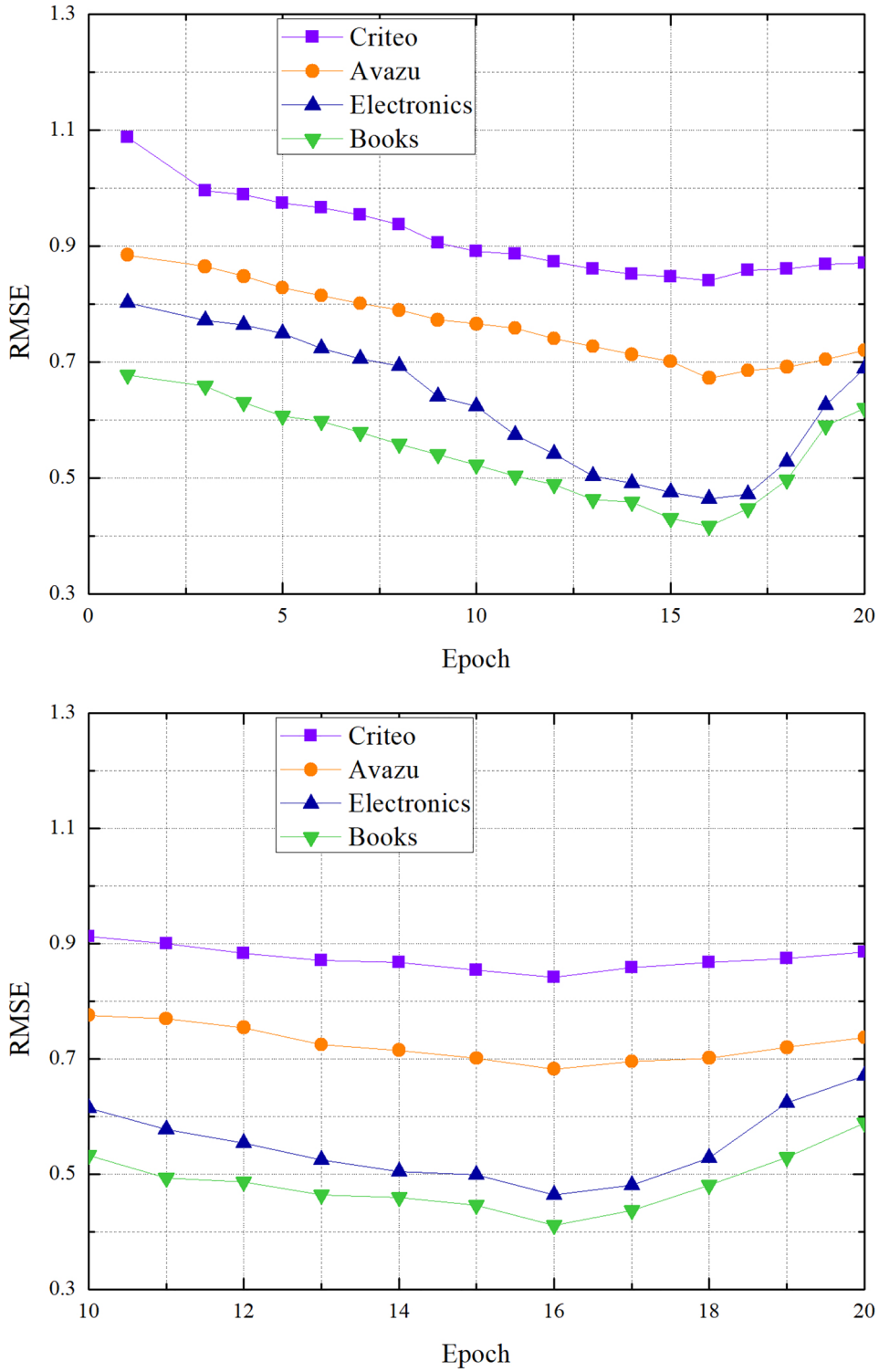

Figure 6. The effect of the epoch.

Datasets	FNN	AVG		MAX		ATT	
		SN	IN	SN	IN	SN	IN
Avazu	0.7825	0.7841	0.7937	0.7902	0.8016	0.8083	0.8324
Criteo	0.7758	0.7702	0.7814	0.7843	0.7938	0.8173	0.8244
Books	0.8026	0.8272	0.8395	0.8331	0.8405	0.8501	0.8965
Electronics	0.7938	0.8198	0.8306	0.8274	0.8317	0.8457	0.8772

Table 4. AUC of SISA variants in different datasets.

Received: 27 October 2021; Accepted: 10 December 2021
Published online: 07 January 2022

References

1. Cheng, H. T., Koc, L., Harmsen, J., et al. Wide and deep learning for recommender systems. In *The Workshop on Deep Learning for Recommender Systems*. ACM, pp. 7–10 (2016).
2. Graepel, T., Borchert, T. & Herbrich, R. Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's bing search engine. In *International Conference on International Conference on Machine Learning*. Omnipress, pp. 13–20 (2010).
3. Najafi-Asadolahi, S. & Fridgeirsdottir, K. Cost-per-click pricing for display advertising. *Manuf. Serv. Oper. Manag.* **16**(4), 482–497 (2014).
4. Grbovic, M. & Cheng, H. Real-time personalization using embeddings for search ranking at airbnb. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp 311–320 (2018).
5. Chapelle, O., Manavoglu, E. & Rosales, R. Simple and scalable response prediction for display advertising. *ACM Trans. Intell. Syst. Technol* **5**(4), 61 (2015).
6. Kumar, R., Naik, S. M., Naik, V. D., Shiralli, S., Sunil, V. G., & Husain, M. (2015). Predicting clicks: CTR estimation of advertisements using logistic regression classifier. In *2015 IEEE International Advance Computing Conference (IACC)* (pp. 1134–1138). IEEE.
7. McMahan, H. B., Holt, G. & Sculley, D., et al. Ad click prediction: A view from the trenches. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1222–1230 (2013).
8. Chapelle, O. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM (2014).
9. Rendle, S. Factorization machines. In *2010 IEEE International conference on data mining. IEEE*, pp. 995–1000 (2010).
10. Juan, Y., Zhuang, Y. & Chin, W. S., et al. Field-aware factorization machines for CTR prediction. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pp. 43–50 (2016).
11. He, X., Pan, J. & Jin, O., et al. Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, pp. 1–9 (2014).
12. Xu, S. *et al.* Computer vision techniques in construction: A critical review. *Arch. Comput. Methods Eng.* **2020**, 1–15 (2020).
13. Esteva, A. *et al.* Deep learning-enabled medical computer vision. *NPJ Digit. Med.* **4**(1), 1–9 (2021).
14. Bhattacharya, S. *et al.* Deep learning and medical image processing for coronavirus (COVID-19) pandemic: A survey. *Sustain. Cities Soc.* **65**, 102589 (2021).
15. Naranjo-Torres, J. *et al.* A review of convolutional neural network applied to fruit image processing. *Appl. Sci.* **10**(10), 3443 (2020).
16. Wolf, T., Chaumond, J. & Debut, L., et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45 (2020).
17. Galassi, A., Lippi, M. & Torroni, P. Attention in natural language processing. *IEEE Trans. Neural Netw. Learn. Syst.* **20**, 20 (2020).
18. Zhang, W., Du, T. & Wang, J. Deep learning over multi-field categorical data. In *European Conference on Information Retrieval*, pp. 45–57. (Springer, Cham, 2016).
19. Cheng, H. T., Koc, L. & Harmsen, J., et al. Wide and deep learning for recommender systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, pp. 7–10 (2016).
20. Guo, H., Tang, R. & Ye, Y., et al. Deepfm: An end-to-end wide and deep learning framework for CTR prediction. arXiv:1804.04950 (arXiv preprint) 2018.
21. Qu, Y., Cai, H. & Ren, K., et al. Product-based neural networks for user response prediction. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, pp. 1149–1154 (2016).
22. Lian, J., Zhou, X. & Zhang, F., et al. xdeepfm: Combining explicit and implicit feature interactions for recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1754–1763 (2018).
23. Wang, R., Fu, B. & Fu, G., et al. Deep and cross network for ad click predictions. In *Proceedings of the ADKDD'17*, pp. 1–7 (2017).
24. Huang, K. *et al.* Why do deep residual networks generalize better than deep feedforward networks?--A neural tangent kernel perspective. *Adv. Neural Inf. Process. Syst.* **20**, 33 (2020).
25. Shan, Y., Hoens, T. R. & Jiao, J., et al. Deep crossing: Web-scale modeling without manually crafted combinatorial features, In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 255–262 (2016).
26. Kim, T. S. Design of a multi-array CNN model for improving CTR prediction. *J. Korea Contents Assoc.* **20**(3), 267–274 (2020).
27. Wang, J., Liu, Q. & Liu, Z., et al. Towards accurate and interpretable sequential prediction: A cnn & attention-based feature extractor. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1703–1712 (2019).
28. Ouyang, W., Zhang, X. & Ren, S., et al. Click-through rate prediction with the user memory network. In *Proceedings of the 1st International Workshop on Deep Learning Practice for High-Dimensional Sparse Data*, pp. 1–4 (2019).
29. Chen, Y. *et al.* A novel deep learning method based on attention mechanism for bearing remaining useful life prediction. *Appl. Soft Comput.* **86**, 105919 (2020).
30. Wang, Q. *et al.* A new approach for advertising CTR prediction based on deep neural network via attention mechanism. *Comput. Math. Methods Med.* **20**, 18 (2018).
31. Cao, T. *et al.* Meta-wrapper: Differentiable wrapping operator for user interest selection in CTR prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 21 (2021).
32. Xiao, J., Ye, H. & He, X., et al. Attentional factorization machines: Learning the weight of feature interactions via attention networks. http://arxiv.org/abs/1708.04617(arXiv preprint) 2017.
33. Jin, Y., Guo, W. & Zhang, Y. A time-aware dynamic service quality prediction approach for services. *Tsinghua Sci. Technol.* **25**(2), 227–238 (2019).

34. Zhou, G., Zhu, X. & Song, C., et al. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1059–1068 (2018).
35. Zhou, G., Mou, N., Fan, Y., et al. (2019) Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), pp. 5941–5948.
36. Wang, F. *et al.* Robust collaborative filtering recommendation with user-item-trust records. *IEEE Trans. Comput. Soc. Syst.* **20**, 21 (2021).
37. Liu, Y. *et al.* Bidirectional GRU networks-based next POI category prediction for healthcare. *Int. J. Intell. Syst.* **20**, 21 (2021).
38. Li, J., Ren, P. & Chen, Z., et al. Neural attentive session-based recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 1419-1428 (2017).
39. Xu, C. *et al.* Graph contextualized self-attention network for session-based recommendation. *IJCAI* **19**, 3940–3946 (2019).
40. Zhang, Y. *et al.* Multi-aspect aware session-based recommendation for intelligent transportation services. *IEEE Trans. Intell. Transport. Syst.* **20**, 20 (2020).
41. Wu C, Wu F, Ge S, et al. Neural news recommendation with multi-head self-attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6389–6394 (2019).
42. Dey, R. & Salem, F. M. Gate-variants of gated recurrent unit (GRU) neural networks. In *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*. IEEE, pp. 1597–1600 (2017).
43. Haidong, S. *et al.* Enhanced deep gated recurrent unit and complex wavelet packet energy moment entropy for early fault prognosis of bearing. *Knowl. Based Syst.* **188**, 105022 (2020).
44. McAuley, J., Targett, C., Shi, Q., et al. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 43–52 (2015).
45. Bai, Z., Zhang, X. L. & Chen, J. Partial AUC optimization based deep speaker embeddings with class-center learning for text-independent speaker verification. In *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6819–6823 (2020).
46. Chen, W. *et al.* A new design of evolutionary hybrid optimization of SVR model in predicting the blast-induced ground vibration. *Eng. Comput.* **37**(2), 1455–1471 (2021).

Acknowledgements

This work was supported by the following Grants: National Natural Science Foundation of China (61772321); Natural Science Foundation of Shandong Province (ZR2021QF071, ZR202011020044); Opening Fund of Shandong Provincial Key Laboratory of Network based Intelligent Computing; Cultivation Fund of Shandong Women's University High-level Scientific Research Project (2020GSPSJ02); Discipline Talent Team Cultivation Program of Shandong Women's University (1904).

Author contributions

Q.W. wrote the main manuscript text and X.Z. prepared all figures . F.L. provided the some idea about model. Q.T. prepared the table. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.