EXTR: Click-Through Rate Prediction with Externalities in E-Commerce Sponsored Search

Chi Chen*
cc274542@alibaba-inc.com
Alibaba Group
Beijing, China

Hui Chen*
chenh19@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Kangzhi Zhao
Junsheng Zhou
kangzhi.zkz@alibaba-inc.com
chuanyan.zjs@alibaba-inc.com
Alibaba Group
Beijing, China

Li He
Hongbo Deng
hl121322@alibaba-inc.com
dhb167148@alibaba-inc.com
Alibaba Group
Beijing, China

Jian Xu
Bo Zheng
xiyu.xj@alibaba-inc.com
bozheng@alibaba-inc.com
Alibaba Group
Beijing, China

Yong Zhang
Chunxiao Xing
zhangyong05@tsinghua.edu.cn
xingcx@tsinghua.edu.cn
Tsinghua University
Beijing, China

ABSTRACT

Click-Through Rate (CTR) prediction, estimating the probability of a user clicking on items, plays a key fundamental role in sponsored search. E-commerce platforms display organic search results and advertisements (ads), collectively called items, together as a mixed list. The items displayed around the predicted ad, i.e. external items, may affect the user clicking on the predicted. Previous CTR models assume the user click only relies on the ad itself, which overlooks the effects of external items, referred to as external effects, or externalities. During the advertising prediction, the organic results have been generated by the organic system, while the final displayed ads on multiple ad slots have not been figured out, which leads to two challenges: 1) the predicted (target) ad may win any ad slot, bringing about diverse externalities. 2) external ads are undetermined, resulting in incomplete externalities. Facing the above challenges, inspired by the Transformer, we propose EXternality TRansformer (EXTR) which regards target ad with all slots as query and external items as key&value to model externalities in all exposure situations in parallel. Furthermore, we design a Potential Allocation Generator (PAG) for EXTR, to learn the allocation of potential external ads to complete the externalities. Extensive experimental results on Alibaba datasets demonstrate the effectiveness of externalities in the task of CTR prediction and illustrate that our proposed approach can bring significant profits to the real-world e-commerce platform. EXTR now has been successfully deployed in the online search advertising system in Alibaba, serving the main traffic.

*Both authors contributed equally to the paper

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '22, August 14–18, 2022, Washington, DC, USA

© 2022 Association for Computing Machinery.

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539053

CCS CONCEPTS

• **Information systems** → **Online advertising**.

KEYWORDS

Click-through rate prediction; Online advertising; Deep learning

ACM Reference Format:
Chi Chen, Hui Chen, Kangzhi Zhao, Junsheng Zhou, Li He, Hongbo Deng, Jian Xu, Bo Zheng, Yong Zhang, and Chunxiao Xing. 2022. EXTR: Click-Through Rate Prediction with Externalities in E-Commerce Sponsored Search. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3534678.3539053

1 INTRODUCTION

Sponsored search (a.k.a search advertising) provides a cost-efficient way for advertisers to promote their products to hundreds of thousands of online customers and has brought significant profits to e-commerce platforms (e.g., Taobao, JD, and Amazon). To optimize the satisfaction of both users and advertisers, numerous advertisements (ads) are ranked by the product of ad quality scores and advertiser bids to compete for the limited advertising space [25, 34]. As one of the most commonly used quality scores, click-through rate (CTR) measures the user clicking probability on the ad and plays a key fundamental role in advertising systems. By virtue of the strong ability of deep learning, many deep CTR models are widely used in practice. Traditional deep CTR models focus on the interactions between features, such as Wide&Deep [5], Deep FM [16], xDeepFM [24], etc. With the aim of obtaining richer user expressions, a series of deep neural networks are proposed to mine the interests of users from their huge historical behavioral data [27, 37, 38]. More recently, facing the highly skewed and sparse features in the real world, industrial approaches concentrate on encoding features more efficiently [15, 17, 20].

Although the existing efforts have demonstrated a remarkable performance, the isolated manner which assumes the user clicking probability only depends on the ad itself restricts the predictive power of the model because it overlooks other displayed items, called external items, in final presentations. In reality, users are

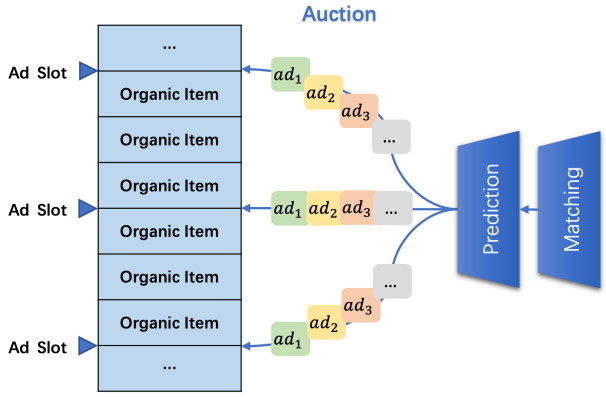

Figure 1: The advertising system follows a three-stage design with matching, prediction and auction.

accustomed to making a purchase online using various electronic devices, such as smartphones, and browse a batch of products on one screen comparatively. Other displayed items, especially on the same page, will affect the user clicking on the target ads [32]. For example, if the ad is placed around an extraordinarily creative neighbor, it will be inconspicuous and lead to unsatisfactory user feedback. Borrowing the concept from the fields of economics and mechanism design, we collectively refer to the effects of external items as external effects, or ***externalities***. Externalities are personalized since the user's sensitivity on good's attributes is full of diversity [4]. In other words, even if the surrounding items are exactly the same, the external effects may perform differently for individuals. Customers like students are likely to react sharply to the price difference, while economical users pay more attention to the subtle quality gap between items. In this paper, we are interested in the personalized externalities for CTR prediction.

However, designing and developing an industrial CTR model with externalities is a challenging task. The e-commerce platform returns a hybrid sequence of organic search results and ads in the specific UI layout. The organic list is firstly generated by the organic system, then the advertising system assigns the right ads on the right ad slots. Practical advertising systems follow a three-stage design with matching, prediction and auction as Figure 1 shows. The final display results of multiple ad slots will be decided after the three stages and thus are incapable of being acquired during the prediction, which results in two major challenges:

Diverse Externalities. The target ad may win any ad slot corresponding to the completely different surroundings. With the aim of predicting accurately, it is requisite to model the diverse externalities when the target ad is displayed on different ad slots. A straight-forward method is to learn the externalities in each exposure situation respectively, and then access the model online many times to obtain a set of pCTRs (predicted CTRs) for all exposure situations. Nevertheless, because of hundreds of millions of visits and limited user click data, it is time-consuming and hard to be conducted in industrial systems. A more efficient way is required to learn the diverse externalities.

Incomplete Externalities. With the purpose of responding immediately, advertising systems simultaneously generate pCTRs of all ads in a parallel way. The ads to be displayed around the target, i.e. external ads, are in processing and cannot be obtained during the prediction. Only excluding the effects of external ads will lead to the incomplete externalities and the inaccurate prediction. To maximize the platform revenue and user satisfaction, it is necessary to learn the allocation of potential external ads to complete the externalities.

To address the aforementioned challenges, we propose an efficient model to exploit personalized externalities in CTR prediction for the practical e-commerce sponsored search system. Inspired by the Transformer [30], which employs an attention-based query-key mechanism to allow for parallelization on input sequences, we design a new deep neural network, called EXternality TRansformer or EXTR, to jointly learn the diverse externalities for all ad exposure situations in a parallel way. Our network is composed of two kinds of Transformer layers: self-attention Transformer layers focus on the interactions of external items and heterogeneous Transformer layers are responsible for the externality extraction. The self-attention layer is a classic structure which has been widely used in many fields, while the heterogeneous layer designs a query-key attention mechanism for different objects to learn the diverse externalities. Furthermore, to tackle the incomplete externalities, we propose Potential Allocation Generator or PAG to learn the promising display slots for all candidate ads except the target. To validate the effectiveness of our approach, in this paper, we perform extensive experiments on real-world datasets from Alibaba. The experimental results demonstrate the effectiveness of externalities in the task of CTR prediction. The further online A/B tests illustrate that our proposed approach can result in a significant incremental revenue of sponsored search business. To sum up, the contributions of this paper include:

- To the best of our knowledge, this paper is the first to study the personalized externalities for CTR prediction in the field of online sponsored search.
- For developing an industrial CTR model with personalized externalities, we design an efficient framework EXTR to simultaneously infer pCTRs of target ads for diverse externalities. To tackle the unknown external ads, we propose PAG to estimate the potential display slots of candidate ads.
- We conduct experiments on real-world datasets. Results verify the necessity of externalities and effectiveness of our proposed approach. The proposed approach has been deployed in the commercial search advertising system in Alibaba, one of the world's largest advertising platforms, contributing significant improvement to the business.

2 RELATED WORK

With the successful development of deep learning in computer vision [18] and natural language processing [2], numerous scientists concentrate on exploiting deep neural network for CTR prediction. Most traditional efforts focus on learning the interactions between features [5, 16, 24, 28, 29]. For example, Wide & Deep [5] proposed by Google combines the benefits of memorization and generalization of feature interactions. Deep FM [16] and xDeepFM [24] directly derive an end-to-end learning model that emphasizes both low-order and high-order feature interactions. However, the user representation vector with a limited dimension in such Embedding&MLP methods restricts the expression of user interests. In

order to capture the user interests, a series of network structures are proposed: Deep Interest Network (DIN) [38] points out that user interests are diverse and vary with items and introduces the attention mechanism to capture the diverse user's interests. Deep Interest Evolution Network (DIEN) [37] proposes an auxiliary loss to capture the latent interests from concrete behavior and refines Gated Recurrent Unit (GRU) [7] to model evolution of interests. The co-design solution of UIC and MIMN is developed to handle the user interest modeling with the unlimited length of sequential behavior data [27]. More recently, facing highly skewed and sparse features in real-world recommendation systems, a number of industrial papers pay more attention to constructing efficient approaches to learn feature embeddings [15, 17, 20]. Previous prediction models only treat items in an isolated manner and lack the consideration of externalities.

Although externalities are overlooked by the existing CTR prediction methods, some other research for online recommendation and advertising has discussed related topics. Externality is originally a term of economics[1] and refers to the cost or benefit caused by a producer that is not financially incurred or received by that producer. Ghosh and Mahdian [10] initiate the study of externalities in the auction mechanism design for online advertising and point out that the Winner Determination Problem is solvable in polynomial time under some specific conditions. Subsequently, a collection of papers explore the keyword auctions with externalities in sponsored search theoretically [1, 8, 11, 12, 21] and empirically [13, 19]. These studies mostly focus on the design of auction mechanism. Recent prediction methods have not involved the externalities. Moreover, the mutual influences among products have been discussed by many ranking papers. A personalized re-ranking model [26] is proposed to learn the mutual influences between items by utilizing the Transformer, nevertheless is inadequate to capture the influence among the heterogeneous items. Zhang *et al.* [9] develop a two-step end-to-end model to solve the heterogeneous feed ranking problem where the solution pays more attention to the influence among types instead of items. A reinforcement learning framework [35] is built to jointly optimize the recommending and advertising strategies and the final display results of ads are determined based on rec-lists. The external effects of rec-lists are taken into account in the ranking model, however, the effects between ads are ignored since the method assumes that at most one ad can be inserted. Ranking tasks pay more attention to the order of items, but ignore the score accuracy. Thus, although ranking models consider the externalitites, they cannot be directly used for the prediction in advertising systems.

Substantial economics literature has presented a number of theories of consumer's context-dependent choice [3, 4, 32] which demonstrates the importance of taking into account externalities in the CTR prediction task. In this paper, we focus on modeling the externalities of both organic results and ads on the target ad for CTR prediction on real-world e-commerce sponsored search platforms.

3 PRELIMINARIES

As aforementioned, in this paper, we pay attention to externalities in the CTR prediction for sponsored search to achieve more accurate forecasting. Before stepping into the specific process, we first introduce the background, then make a more formal problem definition, finally describe the transformer architecture and its parallel processing in this section.

3.1 Background

Practical e-commerce search platforms, such as Taobao, typically recommend a number of products to online users by two separate systems, i.e., organic search system (OS) and search advertising system (AS), which are optimized by different metrics: OS aims at optimizing the user experience or engagement, while AS maximizes the platform revenue and returns on investment (ROI) from advertisers. In order to present a hybrid list of search and advertised items, the organic result list is firstly generated by OS, then AS assigns the right ads on the right ad slots. In this paper, we mainly focus on AS which follows a three-stage design with matching, prediction and auction. Specifically, numerous ads related to user queries are firstly retrieved and subsequently sent to the prediction models to estimate the various ad quality metrics, such as CTR and CVR (Conversion Rate). In the auction stage, these candidate ads are ranked by the generated metrics and advertiser bids. The winning ads after the auction will be finally presented to consumers along with the organic search results. As Figure 1 shows, during the prediction period, previous independent CTR models, which infer the pCTRs for target ads in an isolated manner, lack the consideration of other displayed items, and send the same prediction results for one ad to completely distinct slots. For the purpose of providing more accurate prediction results, it is necessary to infer the click probability with diverse externalities in multiple different exposure situations.

3.2 Problem Formulation

Now we formulate the CTR prediction with externalities. Let $O = \{o_1, ..., o_N\}$ represent the sequence of organic results with the length N which has been generated by OS. The M candidate ads need to be predicted by AS represented by $ad_1, ..., ad_M$. We use $S \subset O \cup \{...ad_{t-1}, ad_{t+1}, ...\}$ to denote the external items of the target ad_t which are engaged in two types, i.e., the organic and the advertised. Given the user profile U, the search query Q, the external items S and a target ad ad, we aim to train a CTR model which can estimate the user click probability $\mathbb{P}(click = 1|ad, Q, U, S)$ on ad. The predicted click probability is denoted by $\hat{c}$. The real user click is denoted by $c \in \{0, 1\}$. Comparatively, the prediction results of independent methods, which merely focus on the target ad ignoring the external items, can be formulated as $\mathbb{P}(click = 1|ad, Q, U)$.

As aforementioned, it is difficult to find out the exact surrounding items because the exposure of target ads is still unknown during the prediction. Therefore, we attempt to infer the click probability in all possible exposure situations and use $\hat{c}_1, ..., \hat{c}_z, ..., \hat{c}_Z$ to denote the prediction results. $\hat{c}_z$ represents the pCTR when the target ad is finally exposed on the z-th ad slot. Z is the number of ad slots. $\hat{c}_z$ can be expressed more formally as $\mathbb{P}(click_z = 1|ad, Q, U, S; z)$.

Machine learning based approaches for CTR prediction usually discover various effective patterns from historical ad impression logs and user clicks collected by AS. The impression logs provide a large number of features which can be divided into three major categories: user-related features $X_u \in \mathbb{R}^{d_u}$, query-related features $X_q \in \mathbb{R}^{d_q}$ and ad-related features $X_{ad} \in \mathbb{R}^{d_{ad}}$. In this paper, to

[1] https://www.investopedia.com/terms/e/externality.asp

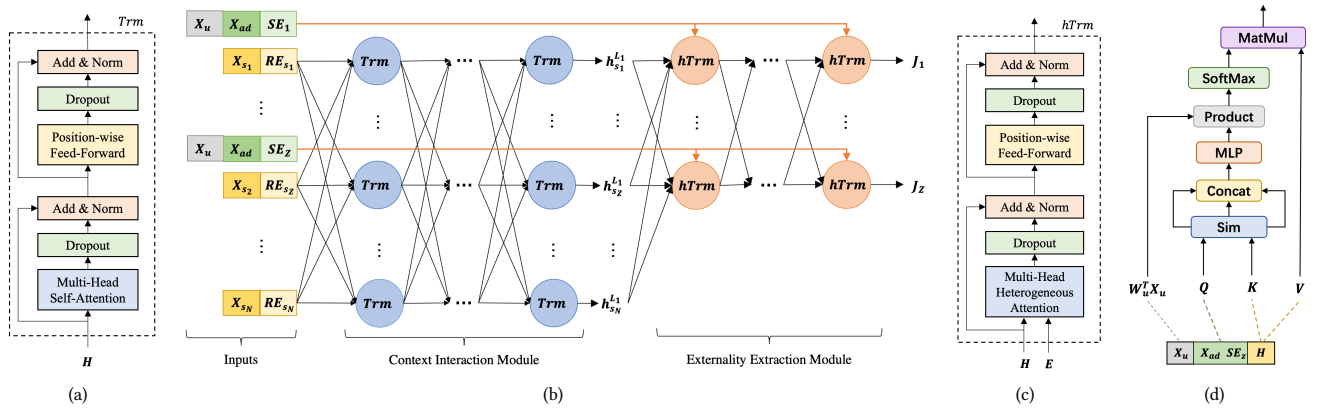

Figure 2: The whole model architecture of EXTR is shown in (b) which consists of two kinds of Transformer layers. (a) displays the self-attention Transformer layer Trm applied by Context Interaction Module. (c) shows the heterogeneous attention Transformer layer $hTrm$ employed by Externality Extraction Module. (d) depicts the heterogeneous attention of $hTrm$.

involve the externalities, we additionally introduce the organic search impressions where the features of organic search item $X_o \in \mathbb{R}^{d_o}$ can be recognized. Different from the independent methods regarding the isolated ad, the accurate externalities require more detailed labels $c_z \in \{0, 1\}$, i.e. the user clicks on the ad when it is exposed on the z-th ad slot. The probability of a user click can be estimated through fitting the training data by the negative log-likelihood function defined as:

$$\mathcal{L}_{ctr} = -\frac{1}{|D|}\sum(clog\hat{c} + (1-c)log(1-\hat{c})) \tag{1}$$

where D is the training set and $|D|$ represents the size of D.

3.3 Transformer Architecture

Transformer, which is initially proposed by Vaswani et al. [30] as a new attention-based building block for machine translation, has played a leading role in many fields, such as neural language translation and computer vision, etc. The technical core of the transformer is the attention mechanism [2, 22] which selectively concentrates on relevant things and ignores others. Transformers introduce self-attention layers which scan through each element of a sequence and update it by aggregating information from the whole sequence. One of the main advantages of attention-based models is their global computations and perfect memory, which makes them more suitable than RNNs on sequences [6, 14]. Inspired by the parallelization of transformers, we design an attention-based query-key network to simultaneously model the diverse external effects for all ad exposure situations.

4 EXTERNALITY MODELING

Previous independent models, such as Wide & Deep, xDeepFM, DIN, etc., employ various deep neural networks (DNNs) $\mathcal{F}$ to learn the user click behavior and generate $pCTR$s based on user, query and target ad related features, which can be formulated by $\hat{c} = \mathcal{F}(X_q, X_u, X_{ad})$. Although the independent approaches have demonstrated a remarkable performance, the isolated manner overlooks the external effects and limits the predictive power of the model. It is quite valuable to incorporate the external items into the current CTR framework to pursue more accurate prediction. Without loss of generality, we abstract the prediction model with externalities as $\hat{c} = \mathcal{F}(X_q, X_u, X_{ad}, \mathcal{G}(*))$. $\mathcal{G}$ is an externality extractor and generates externality representations using various inputs, especially external items.

In the following sections, we start to explore how to extract the externalities by $\mathcal{G}$. Section 4.1 gives a class of methods as base models, which capture the externalities by pooling the embeddings of external items and target ads. Facing the diverse externalities in multiple ad exposure situations, we propose a novel framework EXTR and present the detailed description in Section 4.2. To pursue more accurate prediction, we further introduce the PAG unit to help model the effects of external ads in Section 4.3.

4.1 Base model (Embedding&Pooling)

With the objective to extract externalities, the target ads and external items are essentially required. A class of methods are to transform the features of them into low dimensional dense representations by embedding, then generate the externality representations by pooling the representations. These approaches can be formulated by

$$J = \mathcal{G}(X_s, X_{ad}) \tag{2}$$

where J denotes the extracted externality representations. X_s denotes the feature vector of the external item list. The particular $\mathcal{G}$ can be any neural network, such as MLP, LSTM, Transformer, etc. During the advertising prediction, only the organic external items are already to known. The external ads are also being processed along with the target. Accordingly, the external items can only be acquired partially in practice. Specifically, $X_s = X_{o_1}, ..., X_{o_N}$.

4.2 Externality Transformer (EXTR)

As described in Section 3.1, the externalities are completely different for distinct ad slots. To capture the accurate externalities, it is essential to take all ad slots into account. The following shows a formulation to learn the externalities for each slot,

$$J_z = \mathcal{G}(X_s, X_{ad}, z) \tag{3}$$

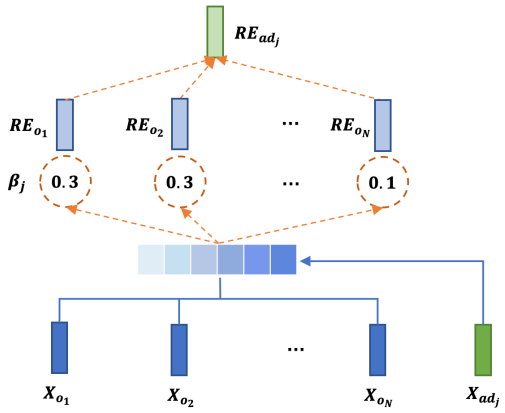

Figure 3: Potential Allocation Generator

where J_z represents the externalities when the target ad is displayed on the z-th ad slot. A straightforward method to realize $\mathcal{G}$ in Equation. (3) is to learn the externalities for each ad slot separately and access the model online many times to obtain the pCTRs. However, due to hundreds of millions of visits and limited user click data, the serial approach is time-consuming and hard to be conducted in the industrial system. Here, we introduce an efficient prediction model called EXTR, which extracts the externality representations of target ads across all ad slots in a parallel way. The whole architecture of EXTR is shown in Figure 2. EXTR mainly consists of two parts which are a Context Interaction Module and an Externality Extraction Module. The context interaction module learns about the mutual influence among the external items. The externality extraction module attempts to model the diverse externalities for all exposure situations simultaneously.

Context Interaction Module. Before extracting the externalities, we allow the external items to interact with each other in advance. Considering the user's sequential browsing, it is appropriate to adopt the sequence network on the external item list S. Here we feed the external item embedding X_{s_i} and the ranking encoding RE_{s_i} into the L_1-layer self-attention Transformer [30]. The hidden representations $h^l_{s_i}$ at each layer l are iteratively computed, while all s_i in the same layer l are processed simultaneously. we stack $h^l_{s_i} \in \mathbb{R}^d$ together into matrix $H^l_s \in \mathbb{R}^{T_s \times d}$. The self-attention Transformer layer Trm can be formulated as,

$$H^l_s = Trm(H^{l-1}_s), \quad \forall l \in [1, ..., L_1] \tag{4}$$

The external items include the organic and the advertised. The organic items are already generated by OS, however, the external ads are still under processing. To tackle the unknown external ads, we propose PAG and unfold the detailed description in Section 4.3.

Externality Extraction Module. The goal of the extraction layer is to capture the diverse externalities efficiently for the target ad when it is exposed on the distinct ad slots. To achieve this goal, inspired by the parallelization of the Transformer, we take the target ad with ad slot as the query, and external items as the key to simultaneously extract the diverse externalities in a parallel way. The extraction module consists of L_2 heterogeneous Transformer Layers ($hTrm$), each of which contains two sub-layers, i.e., a *Heterogeneous Attention* layer and a *Feed-Forward Network* (FFN) layer. As Figure 2(a), 2(c) show, the only difference between Trm and $hTrm$ is the attention mechanism.

Before introducing the specific structure of heterogeneous attention, we first construct the query inputs of the heterogeneous attention by combining features of target ad $X_{ad} \in \mathbb{R}^{d_{ad}}$ and ad slot encoding $SE \in \mathbb{R}^{d_{se}}$, i.e., formally,

$$\boldsymbol{E} = \begin{bmatrix} X_{ad}; SE_1 \\ X_{ad}; SE_2 \\ ... \\ X_{ad}; SE_Z \end{bmatrix} \tag{5}$$

where ; is the concatenate operator. SE_z denotes the encoding of the z-th ad slot. The query $\boldsymbol{E}$ contains the target ad representations with respect to all ad slots. H^l_e denotes both key and value of the heterogeneous attention at each Transformer layer l. The key&value of the first heterogeneous layer are the outputs of the interaction module, i.e., $H^0_e = H^{L_1}_s$. Multi-head heterogeneous attention can be realized as

$$\begin{aligned} MH &= Concat(head_1, ..., head_h)W^O \\ head_i &= Attention(EW^Q, H_eW^K, H_eW^V) \end{aligned} \tag{6}$$

where $W^Q \in \mathbb{R}^{(d_{ad}+d_{se})\times d}$, $W^K, W^V \in \mathbb{R}^{d\times d}$. $W^O \in \mathbb{R}^{hd\times d}$ is the projection matrix. h is the number of headers. The query-key attention network feeds different objects for the query and the key. To extract the externalities, we develop the particular attention function by

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax(\ A^{QK}\ )\boldsymbol{V} \tag{7}$$

where the input of softmax function is A^{QK} which represents the strength of each query-key pair and can be written as

$$A^{QK} = \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & ... & \alpha_{1,T_e} \\ \alpha_{2,1} & \alpha_{2,2} & ... & \alpha_{2,T_e} \\ ... & ... & ... & ... \\ \alpha_{Z,1} & \alpha_{Z,2} & ... & \alpha_{Z,T_e} \end{bmatrix}$$

The real matrix $A^{QK} \in \mathbb{R}^{Z\times T_e}$ is in charge of asking the external effects on the target ad for different ad slots in a parallel way where T_e denotes the length of keys. We integrate query, key and their similarities together in the calculation of α. More formally,

$$\alpha_{i,j} = MLP([Q_i; K_j; \text{sim}\left\langle Q_i, K_j \right\rangle]) \tag{8}$$

In practice, we use the cosine similarity to estimate the vector's similarity. That is, $\text{sim}\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}^\top \boldsymbol{v} / \|\boldsymbol{u}\| \|\boldsymbol{v}\|$. The $\alpha_{i,j}$ shown in Equation. (8) models the objective externalities based on the platform presentation, nevertheless, have the lack of consideration of user's subjective sensitivity. Similar to the preference, the user's sensitivity on good's attributes is full of diversity. For example, customers like students are likely to react sharply to the price difference, while economical users pay more attention to the subtle quality gap between items. It is quite beneficial to join the user characteristics and learn the personalized externalities. Therefore, we modify the $\alpha_{i,j}$ to be

$$\alpha_{i,j} = MLP([Q_i; K_j; \text{sim}\left\langle Q_i, K_j \right\rangle]) \odot W^T_u X_u \tag{9}$$

where $\odot$ represents the element product (i.e. Hadamard product). The outcome of the linear transformation $W^T_u X_u$ implies the sensitivity of the user u, which can strengthen or weaken the objective

externalities individually. Figure 2(d) depicts the heterogeneous attention. To endow the model with nonlinearity and interactions between different dimensions, we stack the point-wise Feed-Forward Network on the heterogeneous attention sub-layer in each layer. The whole Heterogeneous Transformer Layer can be formulated by

$$H_e^l = hTrm(\boldsymbol{E}, H_e^{l-1}), \quad \forall l \in [1, ..., L_2] \tag{10}$$

The final outputs of the extraction module express the externalities in all exposure situations. $\boldsymbol{J_z}$ in Equation. (3) corresponds to the vector $H_e^{L_2}[(z-1)d : zd]$ which is the slice of $H_e^{L_2}$ from the $(z-1)d$-th to the zd-th dimension.

4.3 Potential Allocation Generator (PAG)

In this section, we introduce unit PAG in EXTR for the purpose of tackling the incomplete externalities. As described in Section 4.2, the context interaction module tries to learn the mutual influence among all external items regardless of their particular type. The organic items and their relative rankings are able to be acquired from the OS, while the advertised ones are still undetermined. Only excluding the effects of external ads will lead to a less accurate prediction. Therefore, we collect the candidate ads except the target and estimate their potential ranking encodings. Figure 3 shows the architecture of PAG. To unify the ranking encoding space, we create the ranking encodings of external ads from the organic encodings, i.e., formally,

$$RE_{ad_j} = \sum_{i=1}^{N} \beta_j^i RE_{o_i} \tag{11}$$

where RE_{ad_j} and RE_{o_i} denote the ranking encodings of candidate ad_j and organic item O_i respectively. The $\beta_j^i \in \mathbb{R}$ estimates how close ad_j and O_i are if ad_j can be exposed in the final display list. β_j^i can be generated by

$$\beta_j = \text{softmax}(MLP[X_{ad_j}; W_p^T X_O]) \tag{12}$$

where $X_O = [X_{o_1}; ...; X_{o_N}]$ and $W_p^T \in \mathbb{R}^{d_{ad} \times N d_o}$. The MLP tries to learn the allocation of candidate ad_j. Furthermore, we introduce an auxiliary loss which supervises β_j by using the real placement k of ad_j,

$$r_j^i = \frac{e^{-(k+0.5-i)^2/\tau}}{\sum_i e^{-(k+0.5-i)^2/\tau}} \tag{13}$$

where $k \in [0, ..., N+1]$ is the number of organic items before ad_j in the real exposure. In order to be compatible with unexposed candidate ads, we add the end item [EOS] for the organic sequence denoted by O_{N+1}. The ads that have not been exposed yet correspond to the label with $k = N+1$ implying it will be displayed on the following pages. τ is the temperature that is normally set to 2. Using a higher value for τ produces a softer probability distribution over classes.

We use the Kullback-Leibler divergence to optimize the potential allocation, i.e., $\mathcal{L}_{aux} = \text{KL}(r_j \| \beta_j)$, then combine the auxiliary loss with the cross-entropy loss into a unified objective function:

$$\mathcal{L} = \mathcal{L}_{ctr} + \lambda \mathcal{L}_{aux} \tag{14}$$

where λ is the hyper-parameter which balances potential encoding generation and CTR prediction.

5 EXPERIMENT

5.1 Datasets

To the best of our knowledge, there is no publically available click dataset with context information for CTR prediction. Therefore, we constructed a real-world dataset by collecting one-week impression logs and user clicks in November, 2021 from Taobao[2], which is one of the world's largest e-commerce platforms owned by Alibaba. The dataset contains large-scale records including 35.4 million users, 277.8 million queries, and nearly one million sellers, covering more than twelve thousand item categories such as clothing, electronic products, and fresh products. A total of 13.9 million ads and 67,901 million organic items are involved in the dataset. The feature description of the dataset is shown in Table 1. The external items are characterized by the item features, such as item ID and shop ID. Besides item features, ad creatives are also included in the target ad feature set.

Table 1: Features in Taobao dataset.

Domain	Feature	Type	Dimension Size
Query	query encoding	Category	128
	page	Category	8
	...	...	...
User	user ID	Category	128
	gender	Category	2
	historical average click price	Raw	1
	...	...	...
Item	item ID	Category	128
	shop ID	Category	64
	brand ID	Category	64
	category ID	Category	32
	price	Raw	1
	...	...	...
Ad	creative ID	Category	128
	...	...	...

5.2 Evaluation Metrics

In our experiments, we evaluate the prediction performance with various metrics including AUC, COPC and Logloss. The detailed description of the metrics is elaborated as follows.

- **AUC:** AUC is a widely used metric in the CTR prediction task, which measures the ranking ability of the model for the whole sample. It represents the probability that the model ranks the positive sample higher than the negative sample for any pair of positive and negative samples.
- **COPC:** COPC metric measures the deviation between pCTR and true CTR. The closer the value is to 1, the more accurate the model prediction will be. It is calculated as follows:

$$COPC = \frac{\sum_{i=0}^{n} \text{click}_i}{\sum_{i=0}^{n} pCTR_i} \tag{15}$$

  where *click* represents the actual click of the sample.
- **Logloss:** Logloss is the value of Equation. (1) on the test set and is used to evaluate the consistency of pCTR with the sample's actual CTR, the smaller the better.

[2]https://www.taobao.com/

Table 2: Performance of compared methods on CTR prediction.

Model	AUC	COPC	Logloss
CAN [36]	0.6863	0.9659	0.2345
MLP	0.6891	0.9727	0.2289
PNN [28]	0.6895	1.0298	0.2254
Wide&Deep [5]	0.6902	0.9777	0.2255
DCN [31]	0.6907	0.9815	0.2253
XDeepFM [24]	0.6912	0.9849	0.2257
Transformer [30]	0.6918	1.0179	0.2243
EXTR_org (Ours)	0.6967	0.9895	0.2194
EXTR (Ours)	0.7011	0.9922	0.2181

5.3 Compared Methods

The recent efforts on CTR prediction assume the user click only relies on the ad itself regardless of the external effects. To evaluate the effectiveness of externalities, we employ CAN [36], the state-of-the-art CTR model, as an independent baseline. Furthermore, to compare the performance of externality extractors $\mathcal{G}$, we conduct the experiments on our proposed model and a couple of Embedding&Pooling methods which use the definite organic results sequences and can be abstracted by Equation. (2). The outputs of the externality extractors will be integrated into CAN for the final CTR prediction. A detailed description of compared methods is shown below.

- **CAN** [36]. CAN is the state-of-the-art independent prediction method, which proposes a Co-action Unit to model the co-action between user interest and target item.
- **MLP**. MLP is a commonly used neural network which processes input features by multiple fully connected layers.
- **PNN** [28]. PNN model introduces a product layer after embedding layer to capture high-order feature interactions. In practice, we regard organic results and target ads as two types of fields, the same as below.
- **Wide&Deep** [5]. Wide&Deep model is composed of the "wide" module and the "deep" module. The deep part is the same as DNN, and the wide part adopts a linear model.
- **DCN** [31]. DCN can be viewed as an improved variant of Wide&Deep, which achieves efficient high-order feature interactions by introducing a cross network to replace the linear model in the wide part.
- **xDeepFM** [24]. In addition to the traditional Wide and Deep parts, xDeepFM model uses a compressed interaction network to model high-order cross-features vector-wise.
- **Transformer** [30]. We transform the features of organic sequences into hidden representations by a classic Transformer, then generate the externality representations by pooling the hidden representations.
- **EXTR**. Our proposed model learns the external effects of both organic items and external ads. EXTR can model the diverse externalities in multiple exposure situations for target ads in a parallel way.
- **EXTR_org**. To align the input features with compared MLP&Pooling models, this method applies the EXTR framework on only external organic items.

Table 3: Impact of each component in EXTR.

Personalization	Candidate Ads	Auxiliary Loss	AUC	COPC	Logloss
			0.6946	1.0118	0.2189
✓			0.6967	0.9895	0.2194
✓	✓		0.7005	0.9916	0.2172
✓	✓	✓	0.7011	0.9922	0.2181

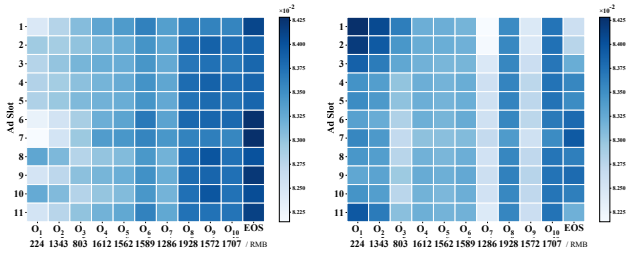

Figure 4: Visualization of attention map in EXTR. (a) on the sample of a high-spending user, (b) on the sample with a low consumption user. Each cell at row k represents the effect strength α of organic results if the target ad, which is priced at RMB 1,599, is exposed before the k-th organic item.

5.4 Experimental Setup

We use data from the first six days as the training set and split the last day's data into validation set (20%) and test set (80%). Due to the huge size of data, we set the mini-batch size to be 2048 and use Adam [23] as the optimizer. We apply exponential decay, in which the learning rate starts at 0.001 and the decay rate is set to 0.9. The Transformer layer numbers in the context interaction module and externality extraction module are both set to 1. In order to capture richer information from different representation subspaces, the multi-head attention mechanism assigns the number of attention heads as 8 for all Transformers. we employ grid search to find the optimal hyperparameters that can result in maximized AUC on the validation sets for EXTR [33]. Specifically, we search the scaling factor λ within {0.1, 0.5, 1} and tune the temperature τ within {1, 2, 4, 6}. $\lambda = 1, \tau = 2$ perform best on our dataset.

5.5 Overall Performance

Table 2 presents a comprehensive analysis of the effectiveness of all compared methods in terms of AUC, COPC and LogLoss for the CTR prediction task. From the table, we have several following observations: All externality extractors can significantly improve the independent baseline which demonstrates the necessity of externalities. Among the various Embedding&Pooling externality methods, PNN is superior to the basic MLP model by modeling higher-order features. The models that combine "wide" module and "deep" module such as Wide&Deep, DCN and xDeepFM perform better than deep-only models. DCN and xDeepFM benefit from the fine design of higher-order feature interaction modules and thus achieve a higher AUC. Transformer models the organic results as a sequence which gains better performance than other compared Embedding&Pooling methods. EXTR_org demonstrates

Table 4: Results from Online A/B testing.

Model	CTR Gain	RPM Gain	QPS
Independent Model	0%	0%	4.2K
Serial Model	-	-	0.9K
EXTR_org	+ 1.54%	+ 1.83%	3.9K
EXTR	+ 2.19%	+ 2.58%	3.5K

its superiority in modeling externalities compared to the above methods with the same external inputs. The promotion derives from two aspects. On the one hand, multi-slot modeling allows the model to learn externalities more accurately. On the other hand, the attention mechanism can adaptively adjust the weight α for each exposure situation which is beneficial to the externality extraction. In addition, EXTR achieves more accurate CTR predictions by modeling the effects of all external items, which indicates that the impact between ads is also non-negligible. Overall, our proposed EXTR brings 0.0148 absolute AUC gain and 0.0263 absolute COPC gain over the SOTA independent baseline which is a significant improvement to our system. In summary, all these results suggest that external items can help CTR models make more accurate predictions, moreover, it is necessary to model the externalities on different ad slots.

5.6 Ablation Studies

We perform ablation experiments over several key components of EXTR in order to better understand their impacts, including the personalization, the candidate ads, and the auxiliary loss. As Table 3 shows, personalized externality modeling can bring 0.002 AUC improvement. To illustrate the effects of personalization more intuitively, we conduct a case study and visualize the attention map in the heterogeneous attention layer. As shown in Figure 4, The subgraph on the left shows the attention map on a sample with a high-spending user when he searches for "pure cotton bedding" and the predicted ad is priced at RMB 1,599. We pick a low consumption user from the dataset and only replace the user-related features in the above sample. The attention map on this virtual sample is shown on the right. As the figure shows, cheaper products, such as the top two organic items, have a greater impact for the low consumption user, while playing an inconsequential role for the high-spending user. It indicates that our proposed model can capture the price sensitivity of users. For a price-sensitive user, the externalities of low-priced items may be more significant. As shown in Table 3, the EXTR with all contexts achieves 0.0044 higher AUC compared to only the organic. It demonstrates that it is inadequate to merely involve the organic results and the candidate ads are valuable. Excluding the effects of external ads would lead to an imprecise prediction. In addition, the model with auxiliary loss performs better than the one without. The auxiliary loss can help PAG generate more accurate ranking encodings of candidate ads to solve the challenge of unknown external ads.

5.7 Online A/B Testing

The online experiments are conducted on our search advertising platform over two weeks from Nov. 14th, 2021 to Nov. 28th, 2021. As shown in Table 4, compared to the independent model, EXTR has improved CTR by 2.19% and RPM by 2.58%. This demonstrates

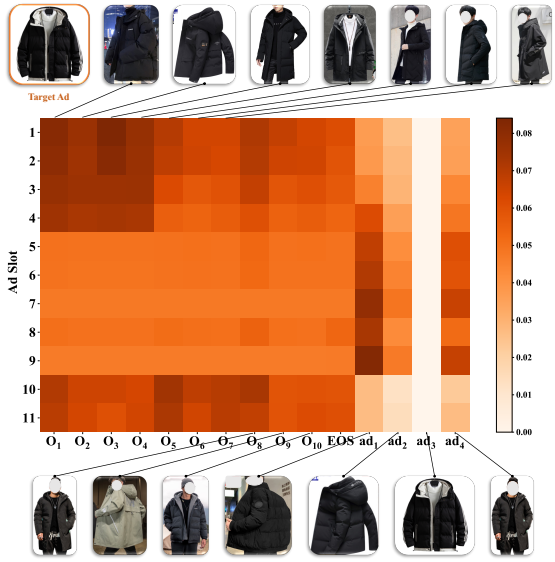

Figure 5: Visulization of attention map in EXTR.

that our proposed approaches are effective and can bring significant revenue to the platform. In addition, we compare EXTR with the online serving independent method and a serial model in terms of online efficiency. The serial model takes the combination of target ad and slot embeddings as inputs and has to be accessed multiple times per query to obtain the externalities on all ad slots. As Table 4 shows, the serial method only has 0.9K QPS (Query Per Second). Thus it is hard to be conducted in our system. In contrast, the parallel EXTR increases the QPS capacity of a single machine by more than 4 times. Compared to the online running independent method without externalities, EXTR slightly reduces the QPS, which is acceptable to the system. Now, EXTR has been deployed online and serves the main traffic, and contributes to significant business revenue growth.

5.8 Visualization of EXTR

We conduct a case study to reveal the inner structure of EXTR. The attention map in the heterogeneous attention can indicate the learned external effects by EXTR and is visualized in Figure 5. The 10 organic items returned by OS and the top 4 candidate ads from AS are all about the user query "men's winter coat". ad_3 is the predicted ad and ad_2 is unexposed. From the figure, we can observe the varying attention weights for ad slots. Specifically, the top four organic items have a significant impact on the target ad if it is placed in the front of the page, while ad_1 makes a major contribution when in the middle. When the ad is exposed at the bottom, it is mostly influenced by the organic results and each of which has an equal weight. There are a couple of explanations for the above observations: 1). The top four organic products are more similar to the target ad which are all thick coats and more relevant for the search term "winter". Thus, the top four organic products are the main competitors for the target. If the target is advertised near them, the influence is relatively apparent. 2). Similar to the top four organic products, ad_1 is also a thick coat which results in a vital contribution when the target is displayed around. 3). At the bottom

of the page, thick coats are far away. Hence all organic results bring equal externalities. The visualization results demonstrate that the appropriate externality weights can be captured by our proposed model.

6 CONCLUSIONS

In this paper, we propose an efficient framework EXTR to exploit the personalized externalities in the practical e-commerce sponsored search system for CTR prediction. We first design a new Transformer based model to jointly learn the diverse externalities in all possible ad exposure situations in parallel. Then we propose the Potential Allocation Generator unit to learn the promising exposure slots for all candidate ads except the target. Extensive experimental results on real-world datasets show that our model outperforms state-of-the-art baselines. EXTR has now been successfully deployed in the Alibaba online search advertising system.

ACKNOWLEDGMENTS

Chi Chen, Yong Zhang, and Chunxiao Xing are supported by National Key R&D Program of China 2020AAA0109603.

REFERENCES

[1] Gagan Aggarwal, Jon Feldman, Shanmugavelayutham Muthukrishnan, and Martin Pál. 2008. Sponsored search auctions with markovian users. In *International Workshop on Internet and Network Economics*. Springer, 621–628.
[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
[3] Tom Blake, Sarah Moshary, Kane Sweeney, and Steve Tadelis. 2021. Price salience and product choice. *Marketing Science* 40, 4 (2021), 619–636.
[4] Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. 2013. Salience and consumer choice. *Journal of Political Economy* 121, 5 (2013), 803–843.
[5] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*. 7–10.
[6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).
[7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).
[8] Xiaotie Deng and Jiajin Yu. 2009. A new ranking scheme of the GSP mechanism with Markovian users. In *International Workshop on Internet and Network Economics*. Springer, 583–590.
[9] Zizhe Gao, Zheng Gao, Heng Huang, Zhuoren Jiang, and Yuliang Yan. 2018. An end-to-end model of predicting diverse ranking on heterogeneous feeds. In *eCOM@ SIGIR*.
[10] Arpita Ghosh and Mohammad Mahdian. 2008. Externalities in online advertising. In *Proceedings of the 17th international conference on World Wide Web*. 161–168.
[11] Arpita Ghosh and Amin Sayedi. 2010. Expressive auctions for externalities in online advertising. In *Proceedings of the 19th international conference on World wide web*. 371–380.
[12] Ioannis Giotis and Anna R Karlin. 2008. On the equilibria and efficiency of the GSP mechanism in keyword auctions with externalities. In *International workshop on internet and network economics*. Springer, 629–638.
[13] Renato Gomes, Nicole Immorlica, and Evangelos Markakis. 2009. Externalities in keyword auctions: An empirical and theoretical assessment. In *International workshop on internet and network economics*. Springer, 172–183.
[14] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. 2016. LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems* 28, 10 (2016), 2222–2232.
[15] Huifeng Guo, Bo Chen, Ruiming Tang, Weinan Zhang, Zhenguo Li, and Xiuqiang He. 2021. An Embedding Learning Framework for Numerical Features in CTR Prediction. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2910–2918.
[16] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv preprint arXiv:1703.04247* (2017).
[17] Wei Guo, Rong Su, Renhao Tan, Huifeng Guo, Yingxue Zhang, Zhirong Liu, Ruiming Tang, and Xiuqiang He. 2021. Dual Graph enhanced Embedding Neural Network for CTRPrediction. *arXiv preprint arXiv:2106.00314* (2021).
[18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.
[19] Przemyslaw Jeziorski and Ilya Segal. 2015. What makes them click: Empirical analysis of consumer demand for search advertising. *American Economic Journal: Microeconomics* 7, 3 (2015), 24–53.
[20] Wang-Cheng Kang, Derek Zhiyuan Cheng, Tiansheng Yao, Xinyang Yi, Ting Chen, Lichan Hong, and Ed H Chi. 2021. Learning to Embed Categorical Features without Embedding Tables for Recommendation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 840–850.
[21] David Kempe and Mohammad Mahdian. 2008. A cascade model for externalities in sponsored search. In *International Workshop on Internet and Network Economics*. Springer, 585–596.
[22] Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. 2017. Structured attention networks. *arXiv preprint arXiv:1702.00887* (2017).
[23] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[24] Jianxun Lian, Xiaohuan Zhou, Fuzheng Zhang, Zhongxia Chen, Xing Xie, and Guangzhong Sun. 2018. xdeepfm: Combining explicit and implicit feature interactions for recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1754–1763.
[25] Xiangyu Liu, Chuan Yu, Zhilin Zhang, Zhenzhe Zheng, Yu Rong, Hongtao Lv, Da Huo, Yiqing Wang, Dagui Chen, Jian Xu, et al. 2021. Neural Auction: End-to-End Learning of Auction Mechanisms for E-Commerce Advertising. *arXiv preprint arXiv:2106.03593* (2021).
[26] Changhua Pei, Yi Zhang, Yongfeng Zhang, Fei Sun, Xiao Lin, Hanxiao Sun, Jian Wu, Peng Jiang, Junfeng Ge, Wenwu Ou, et al. 2019. Personalized re-ranking for recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 3–11.
[27] Qi Pi, Weijie Bian, Guorui Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Practice on long sequential user behavior modeling for click-through rate prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2671–2679.
[28] Yanru Qu, Han Cai, Kan Ren, Weinan Zhang, Yong Yu, Ying Wen, and Jun Wang. 2016. Product-based neural networks for user response prediction. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 1149–1154.
[29] Ying Shan, T Ryan Hoens, Jian Jiao, Haijing Wang, Dong Yu, and JC Mao. 2016. 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*. 255–262.
[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
[31] Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. 2017. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD'17*. 1–7.
[32] Ruizhe Zhang, Xiaohui Xie, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2021. Constructing a Comparison-based Click Model for Web Search. In *Proceedings of the Web Conference 2021*. 270–283.
[33] Yuanxing Zhang, Langshi Chen, Siran Yang, Man Yuan, Huimin Yi, Jie Zhang, Jiamang Wang, Jianbo Dong, Yunlong Xu, Yue Song, et al. 2022. PICASSO: Unleashing the Potential of GPU-centric Training for Wide-and-deep Recommender Systems. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. IEEE.
[34] Zhilin Zhang, Xiangyu Liu, Zhenzhe Zheng, Chenrui Zhang, Miao Xu, Junwei Pan, Chuan Yu, Fan Wu, Jian Xu, and Kun Gai. 2021. Optimizing Multiple Performance Metrics with Deep GSP Auctions for E-commerce Advertising. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 993–1001.
[35] Xiangyu Zhao, Xudong Zheng, Xiwang Yang, Xiaobing Liu, and Jiliang Tang. 2020. Jointly learning to recommend and advertise. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3319–3327.
[36] Guorui Zhou, Weijie Bian, Kailun Wu, Lejian Ren, Qi Pi, Yujing Zhang, Can Xiao, Xiang-Rong Sheng, Na Mou, Xinchen Luo, et al. 2020. CAN: Revisiting Feature Co-Action for Click-Through Rate Prediction. *arXiv preprint arXiv:2011.05625* (2020).
[37] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 5941–5948.
[38] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1059–1068.