Multi-Domain Deep Learning from a Multi-View Perspective for Cross-Border E-commerce Search

Yiqian Zhang[1*], Yinfu Feng[2*], Wen-Ji Zhou[2*], Yunan Ye[2], Min Tan[1], Rong Xiao[2], Haihong Tang[2], Jiajun Ding[1], Jun Yu[1†]

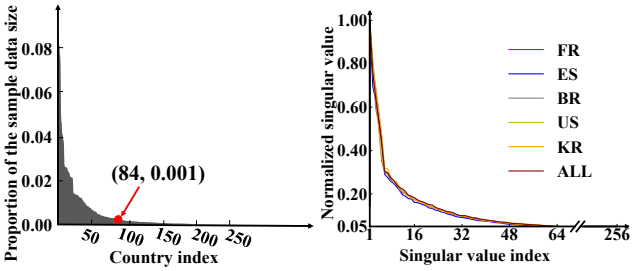
[1]Hangzhou Dianzi University
[2]Alibaba International Digital Commerce Group
yiqian.zyq@gmail.com, fyf200502@gmail.com, eric.zwj@alibaba-inc.com, yunan.yyn@alibaba-inc.com, tanmin@hdu.edu.cn, xiaorong.xr@taobao.com, tanghaihong@alibaba-inc.com, djj@hdu.edu.cn, yujun@hdu.edu.cn

Abstract

Building click-through rate (CTR) and conversion rate (CVR) prediction models for cross-border e-commerce search requires modeling the correlations among multi-domains. Existing multi-domain methods would suffer severely from poor scalability and low efficiency when number of domains increases. To this end, we propose a **D**omain-**A**ware **M**ulti-view m**O**del (DAMO), which is domain-number-invariant, to effectively leverage cross-domain relations from a multi-view perspective. Specifically, instead of working in the original feature space defined by different domains, DAMO maps everything to a new low-rank multi-view space. To achieve this, DAMO firstly extracts multi-domain features in an explicit feature-interactive manner. These features are parsed to a multi-view extractor to obtain view-invariant and view-specific features. Then a multi-view predictor inputs these two sets of features and outputs view-based predictions. To enforce view-awareness in the predictor, we further propose a lightweight view-attention estimator to dynamically learn the optimal view-specific weights w.r.t. a view-guided loss. Extensive experiments on public and industrial datasets show that compared with state-of-the-art models, our DAMO achieves better performance with lower storage and computational costs. In addition, deploying DAMO to a large-scale cross-border e-commence platform leads to 1.21%, 1.76%, and 1.66% improvements over the existing CGC-based model in the online AB-testing experiment in terms of CTR, CVR, and Gross Merchandises Value, respectively.

Introduction

Click-through rate (CTR) and conversion rate (CVR) prediction are two important tasks for e-commerce search, attracting increasing attention from both academia and industry in recent years (Cheng et al. 2016; Zhou et al. 2018; Tang et al. 2020). Cross-border scenarios, in particular, bring more challenges when building effective yet efficient CTR/CVR models. According to our statistics on a cross-border e-commerce platform which is active in more than 230 countries and available in 18 languages, the overlap rate among the top-1k exposed, clicked, and purchased products in Russia, Spain, the USA, and Brazil is only 31.8%, 17.1%, and even 5.1%, respectively. Statistics show that consumers from different countries naturally exhibit distinctive shopping preferences, leading to cruel multi-domain problems in e-commerce data. These problems, such as domain imbalance (refer to Fig 1(a)) and domain-specific modeling, become more severe in real-world applications where hundreds of countries can be involved.



(a) (b)

Figure 1: Statistical analysis on a large-scale cross-border e-commerce platform. (a) The proportion of user behavior data within two weeks among different countries. (b) Normalized singular values of multi-domain feature representations.

Multi-domain learning based methods (Zhu et al. 2022; Liu et al. 2018) are popular choices in terms of addressing the above-mentioned problems. Typically, multi-domain features are extracted and further decomposed into domain-specific and domain-invariant cues (Ma et al. 2018a; Sheng et al. 2021). Though providing satisfactory performance in experiments, existing methods suffer from poor scalability and low efficiency with an increasing number of domains due to their domain-number-variant design. Furthermore, the domain-specific parameters of some domains with limited data may also face the inadequate training problem. An engineering trick can mitigate this by working on data clusters rather than the original separated multi-domain data. However, their enhanced efficiency often accompanies a reduction in performance capabilities as these methods may disrupt domain correlations within multi-domain data.

In this work, we propose a **D**omain-**A**ware **M**ulti-view m**O**del (DAMO) to address the scalability and efficiency

*These authors contributed equally.
†Corresponding author

problem in cross-border e-commerce search. Our main idea is inspired by the low-rank property in multi-domain e-commerce data, reflecting the fact that multi-domain data can be well reconstructed by only a few bases in a multi-view feature space. This low-rank property is validated by our observation in Fig 1(b). Specifically, we randomly sample 0.1 million samples from the top-5 countries (*i.e.*, France(FR), Spain(ES), Brazil(BR), the United States(US), and Korea(KR)) of the platform. From the normalized singular values of feature representations, we notice that these multi-domain data are approximately low-rank. Therefore, we turn to low-rank multi-view space such that our model is domain-number-invariant, resulting in a more efficient representation without sacrificing the overall performance. More specifically, DAMO firstly extracts multi-domain features in an explicit feature-interactive manner. Next, we introduce a multi-view extractor to map the multi-domain features to view-invariant and view-specific spaces. Then a multi-view predictor inputs these two sets of features and outputs view-based predictions.

To impose the informativeness of multi-view bases, we introduce a lightweight view-attention estimator. This estimator aims to learn view-specific weights w.r.t. a view-guided loss that guides the models to learn differentiated views. In addition, it aids in an effective fusion of multi-view predictions, compared to conventional manual processes.

Our main contribution can be summarized as follows:

- A novel DAMO model that exploits the low-rank property of multi-domain data by decomposing multi-domain features into view-invariant and view-specific spaces.
- A lightweight view-attention estimator to enforce view-awareness and effective fusion of multi-view predictions.
- SOTA performances with much lower computational costs on public datasets as well as a large-scale cross-border e-commerce platform.

Related Work

Multi-domain Learning

Multi-domain learning (MDL) can leverage multi-domain data to address the inherent data imbalance within each domain and improve performance. Roughly speaking, existing multi-domain CTR/CVR models can be divided into three categories. **Data sharing.** This kind of method (Jiang et al. 2022) usually uses the correlated samples from other domains and assigns them with pseudo labels for training a model in one domain. Recently, outputs from an MMOE-based deep model across different domains have been combined to exploit the correlation in the label space (Li et al. 2020). These methods show promising results for domains with limited labels. However, they may alter the original label distribution and adversely affect model performance in domains with abundant labels. **Domain adaptation.** This kind of method (Sheng et al. 2021; Jiang et al. 2022) aims to align the representation and distribution of data across diverse domains using domain-specific parameters, thereby reducing the diversity of multi-domain data. However, when the number of domains increases, especially in the case of a large number of domains, these methods may encounter the efficiency problem. **Parameter differentiation.** To capture both diversity and commonality in multi-domain data, some researchers proposed using shared and customized network structures as a means of learning and understanding these nuances (Jiang et al. 2022; Zhang et al. 2022a; Zou et al. 2022). Recently, some multi-task learning models (Ma et al. 2018a; Tang et al. 2020; Ding et al. 2021) have also been applied to deal with multi-domain data. However, when dealing with some domains with limited data, domain-specific structures in these methods may encounter optimization difficulties. To address this problem, Zhang *et al*. (Zhang et al. 2022a) proposed a meta-learning-based method to dynamically generate parameters by leveraging scenario knowledge. However, this method relies heavily on a large amount of training data to achieve model convergence and generalization, which would impose a significant cost. Moreover, the data imbalance problem in multi-domain data can reduce the effectiveness of this method.

Multi-view Learning

Multi-view learning (MVL) has gained significant attention and achieved practical success by exploiting complementary information from multiple features or modalities in multi-view data. It has also been applied to boost CTR/CVR prediction in the e-commerce scenario (Elkahky, Song, and He 2015; Tai et al. 2020; Wu et al. 2022). For example, Elkahky *et al*. (Elkahky, Song, and He 2015) extended a deep learning approach for content-based recommendation by jointly learning features of items from different domains and user features within a multi-view deep learning model. Tai *et al*. (Tai et al. 2020) proposed a recommendation model wherein items were represented from user and entity views. Li *et al*. (Li et al. 2022) modeled user multi-view preferences from knowledge, semantic, and consuming views when building the conversational recommender system. In our work, inspired by the low-rank property of e-commerce multi-domain data, we transform the original multi-domain features into multi-view features. Unlike existing methods, we use an implicit and data-learned view space for feature representation. The view weights of samples are learned via sparse coding with a learned compacted view dictionary on multi-domain data. These weights can be used to fuse multi-view prediction results optimally.

Proposed Method: DAMO

We first formulate the conventional multi-domain CTR/CVR prediction task and point out the undesirable strong correlation between model size and the number of domains. Then we describe DAMO in detail, including three proposed components and the objective function. The overall framework is shown in Fig 2.

Preliminary

Suppose $\mathbf{D} = \{\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, \cdots, \mathbf{D}^{(n)}\}$ is a multi-domain dataset collected from n domains. The dataset for domain i is represented by $\mathbf{D}^{(i)} = \{(\mathbf{x}_1^{(i)}, y_1^{(i)}), \cdots, (\mathbf{x}_{m_i}^{(i)}, y_{m_i}^{(i)})\}$, where m_i is the number of samples in domain i, $\mathbf{x}_j^{(i)} \in \mathbb{R}^{d \times 1}$

is the feature representation of the j-th sample in domain i, and $y_j^{(i)} \in \{0, 1\}$ is the associated click/conversion label.

Multi-domain CTR/CVR Prediction

Normally, a conventional multi-domain CTR/CVR prediction task is formulated by the following problem:

$$\begin{aligned} &\min_{\Theta} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \mathcal{L}\left(y_j^{(i)}, \hat{y}_j^{(i)}\right), \\ &s.t. \quad \hat{y}_j^{(i)} = \mathcal{F}\left(\mathbf{x}_j^{(i)}; \Theta\right), \forall i, j, \end{aligned} \tag{1}$$

where $\hat{y}_j^{(i)}$ is the predicted CTR/CVR result for $\mathbf{x}_j^{(i)}$ under the model $\mathcal{F}(\cdot)$ parameterized by Θ, and $\mathcal{L}(\cdot)$ is a loss function (*e.g.*, the cross-entropy loss function).

Different domains likely share some domain-invariant information, while each domain holds distinct information. To exploit both the shared and domain-specific information within multi-domain data, most existing deep models incorporate both shared and domain-specific parameters (Tang et al. 2020) by solving the following problem:

$$\begin{aligned} &\min_{\Theta_C, \Theta_R^{(i)}|_{i=1}^n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \mathcal{L}\left(y_j^{(i)}, \hat{y}_j^{(i)}\right), \\ &s.t. \quad \hat{y}_j^{(i)} = \mathcal{F}\left(\mathbf{x}_j^{(i)}; \Theta_C, \Theta_R^{(i)}\right), \forall i, j, \end{aligned} \tag{2}$$

where Θ_C and $\Theta_R^{(i)}$ are the shared and i-th domain-specific parameters, respectively. Consequently, there is a hard correlation between model parameters and the number of domains (*i.e.*, the number of domain-specific parameters $\Theta_R^{(i)}$ increases linearly with the number of domains n), which leads to the scalability and efficiency problem.

Architecture Overview

To tackle the aforementioned problems, we propose DAMO for multi-domain CTR/CVR prediction from a multi-view perspective. As shown in Fig 2, we develop a correlated multi-view extractor to obtain projection (view-invariant) and view-specific features, and devise a domain-aware multi-view predictor to generate view-based prediction results; finally, a lightweight view-attention estimator is adopted to dynamically learn the optimal view-specific weights w.r.t. a view-guided loss to impose the informativeness of different views. More details are described below.

Correlated Multi-view Extractor As shown in Fig 2, the raw input features in E-commerce describe user, item, search query, and context information (Wang et al. 2015). They can be classified into three types: sparse, dense, and sequence features. Normally, the sparse and sequence features can be transformed into dense embedding features in the bottom feature network. To better exploit the correlation between all domain information and domain-specific information, we decompose these features into three parts: the global features $\mathbf{x}_{j_G}^{(i)}$, the local features $\mathbf{x}_{j_L}^{(i)}$ and others $\mathbf{x}_{j_O}^{(i)}$. The global features describe all domains, while the local features only describe one or a few domains. Take CTR for example, the CTR about an item on the whole e-commerce platform is a global feature, while that on a particular country site of the platform belongs to a local feature.

Effectively modeling correlation among global and local features is crucial for multi-domain CTR/CVR prediction in cross-border e-commerce scenarios. Unfortunately, due to the ignoring of some important local and global similarity/dissimilarity information, the conventional simple concatenating of different features together cannot make full use of multi-domain data (Dai et al. 2021). To address this issue, we propose to explicitly model global-local feature interactions by performing the outer product of these two vectors. Furthermore, we adopt a fully connected layer (FC) with a Rectified Linear Unit (ReLU) activation function to perform dimension reduction after feature interaction, to reduce the computation cost. Finally, we concatenate all features together, so an elaborated multi-domain feature representation $\bar{\mathbf{x}} \in \mathbb{R}^{d_0 \times 1}$ can be obtained as follows:

$$\bar{\mathbf{x}} = \mathbf{x}_L \oplus FC(\mathbf{x}_L \otimes \mathbf{x}_G) \oplus \mathbf{x}_G \oplus \mathbf{x}_O, \tag{3}$$

where $\oplus$ and $\otimes$ are the concatenation and outer product operators, $\mathbf{x}_G$, $\mathbf{x}_L$ and $\mathbf{x}_O$ are the global features, the local features, and the other features, respectively. Here we omit some subscripts and superscripts of $\mathbf{x}$ to ease the explanation. Similarly, we denote $\mathbf{x}_R$ as the embedding of the domain ID of a sample.

Inspired by the great success of signal decomposition and subspace learning (Mallat 1989; Li et al. 2015), we propose a correlated multi-view extractor to learn two types of derived features (*i.e.*, shared projection features $\mathbf{x}_p^{(t)}|_{t=1}^T \in \mathbb{R}^{d_1 \times 1}$ and view-specific features $\mathbf{x}_s^{(v)}|_{v=1}^V \in \mathbb{R}^{d_1 \times 1}$) from $\bar{\mathbf{x}}$ to enrich feature representation. The shared projection features are used to explicitly represent the sample from multi-view, while the view-specific features are used to keep some important view-specific information. Based on a learnable meta-weight matrix $\mathbf{W}_b \in \mathbb{R}^{d_1 \times d_0}$, we randomly initialize T projection matrices $\mathbf{W}_p^{(t)}|_{t=1}^T \in \mathbb{R}^{d_1 \times d_1}$ and the associated biases $\mathbf{b}_p^{(t)}|_{t=1}^T \in \mathbb{R}^{d_1 \times 1}$. Then an affine transformation operation is applied to $\bar{\mathbf{x}}$ to generate $\mathbf{x}_p^{(t)}|_{t=1}^T$. These shared projection features are visible to all view subnetworks, which are used in CTR/CVR prediction. Meanwhile, the view-specific features are extracted with V individual fully connected layer. Thus, the t-th projection feature $\mathbf{x}_p^{(t)}$ and the v-th view-specific feature are defined as below:

$$\mathbf{x}_p^{(t)} = \sigma\left(\mathcal{B}\left(\left(\mathbf{W}_p^{(t)}\mathbf{W}_b + \alpha\mathbf{W}_b\right)\bar{\mathbf{x}} + \mathbf{b}_p^{(t)}\right)\right), \tag{4}$$

$$\mathbf{x}_s^{(v)} = FC^{(v)}\left(\bar{\mathbf{x}}\right), \tag{5}$$

where α is a learnable scalar, $\sigma(\cdot)$ is an activation function (*e.g.*, the ReLU function), and $\mathcal{B}(\cdot)$ is batch normalization.

Domain-aware Multi-view Predictor With the shared projection features and the view-specific features, we convert the conventional multi-domain CTR/CVR prediction task into a multi-view CTR/CVR prediction task and propose a domain-aware multi-view predictor. The predictor module is described below in detail.

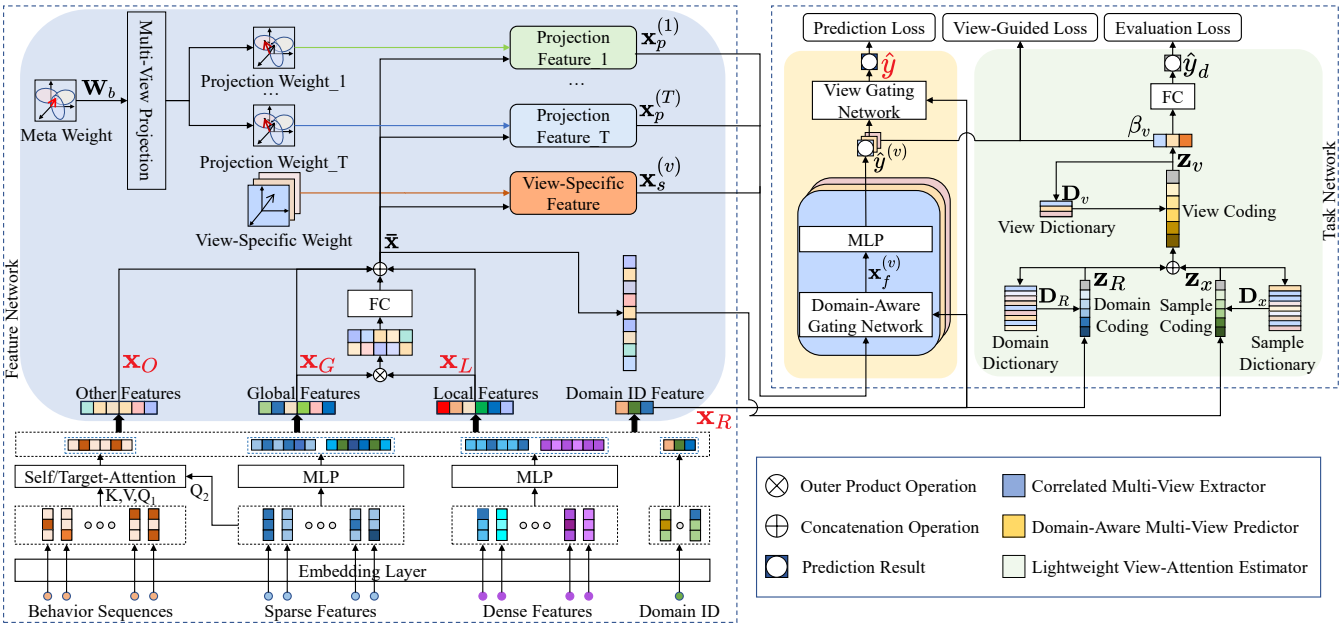


Figure 2: The architecture of DAMO can be divided into two parts, i.e., the feature network on the left and the task network on the right. The feature network includes a bottom feature network for feature engineering and an upper Correlated Multi-view Extractor. The task network attributes in a Domain aware Multi view Predictor and a Lightweight View Attention Estimator.

To jointly use both the shared projection features and view-specific features, we first employ a view-based domain-aware gating network to generate the view-level fusion feature $\mathbf{x}_f^{(v)}|_{v=1}^{V} \in \mathbb{R}^{d_1 \times 1}$ as follows:

$$\begin{aligned} &\mathbf{x}_f^{(v)} = [\mathbf{x}_s^{(v)}, \mathbf{x}_p^{(1)}, \cdots, \mathbf{x}_p^{(T)}] \times g_1(\mathbf{x}_R)^{(v)}, \\ &s.t.\ g_1(\mathbf{x}_R)^{(v)} = \mathcal{S}(\mathcal{B}(\mathbf{W}_{g_1}^{(v)} \mathbf{x}_R)), \end{aligned} \tag{6}$$

where $\mathbf{W}_{g_1}^{(v)} \in \mathbb{R}^{(1+T)\times d_2}$, d_2 is the dimension of $\mathbf{x_R}$ and $\mathcal{S}(\cdot)$ is the softmax function. Since $\mathbf{x}_R$ is the embedding of the domain ID of the sample, we call $g_1(\cdot)^{(v)}$ as a domain-aware gating network.

Afterward, each view sub-network independently makes a CTR/CVR prediction (*i.e.*, $\hat{y}^{(v)}|_{v=1}^{V}$) using a two layers multilayer perceptron (MLP) as below:

$$\hat{y}^{(v)} = MLP^{(v)}(\mathbf{x}_f^{(v)}). \tag{7}$$

Since the domain information has been encoded in $\mathbf{x}_f^{(v)}$, the multi-view predictor in Eq(7) is domain-aware. Meanwhile, we can minimize the loss $\ell\left(y, \hat{y}^{(v)}\right)$ to enforce all multi-view predictors giving estimates as accurate as possible.

To avoid the imbalanced learning of a gating network in the training stage (Shazeer et al. 2017) and to enhance the prediction accuracy, we use a domain-aware view gating network (*i.e.*, $g_2(\cdot)$) to fuse all prediction results as follows,

$$\begin{aligned} &\hat{y} = [\hat{y}^{(1)}, \hat{y}^{(2)}, \cdots, \hat{y}^{(V)}] \times g_2(\mathbf{x}_R), \\ &s.t.\ g_2(\mathbf{x}_R) = \mathcal{S}(\mathcal{B}(\mathbf{W}_{g_2} \mathbf{x}_R)), \end{aligned} \tag{8}$$

where $\mathbf{W}_{g_2} \in \mathbb{R}^{V \times d_2}$ is the learnable parameters of $g_2(\cdot)$. So, we can minimize the weighted multi-view prediction loss $\ell\left(y, \hat{y}\right)$ to make full use of the underlying correlation between different views.

Lightweight View-Attention Estimator Considering the big gap across domains, it is unwise to assign equal weight to each view-guided loss $\ell\left(y, \hat{y}^{(v)}\right)$. Differentiated weights need to be introduced to enforce view-awareness and benefit effective fusion of multi-view predictions. Particularly, we devise a lightweight view-attention estimator to calculate view-specific weights, to adaptively and optimally fuse all view-guided losses. The whole estimator consists of two stages, *i.e.*, view data coding and view weight learning.

In the coding stage, we apply the deep dictionary learning (DL) method (Rodríguez-Domínguez and Dalmau 2020) to learn two over-complete dictionaries, *i.e.*, domain dictionary $\mathbf{D}_R$ and sample dictionary $\mathbf{D}_x$; leveraging the dictionaries, we can calculate the corresponding domain coding $\mathbf{z}_R$ and feature coding $\mathbf{z}_x$ from the domain embedding $\mathbf{x}_R$ and the elaborated multi-domain feature representation $\bar{\mathbf{x}}$, respectively. The DL model can be formulated as follows:

$$\min_{\mathbf{D},Z} \frac{1}{2}\|\mathbf{DZ} - \mathbf{X}\|_2^2 + \alpha_1\psi_1(\mathbf{Z}) + \alpha_2\psi_2(\mathbf{D}) \tag{9}$$

where $\mathbf{X}$ is a set of training samples (*e.g.*, $\mathbf{x}_R$), $\mathbf{Z}$ is the coding of $\mathbf{X}$ related to dictionary $\mathbf{D}$; ψ_1 and ψ_2 denote two regularization terms on $\mathbf{Z}$ and $\mathbf{D}$; α_1 and α_2 are two regularization parameters. Note that, in most DL methods, the dictionary is learned by two alternative steps, *i.e.*, fix $\mathbf{D}$ and perform sparse coding to compute $\mathbf{Z}$, and update $\mathbf{D}$ with fixed $\mathbf{Z}$. To reduce the computation cost, we apply a dimension reduction processing with a fully connected layer on $\bar{\mathbf{x}}$ before dictionary learning. After dictionary learning pro-

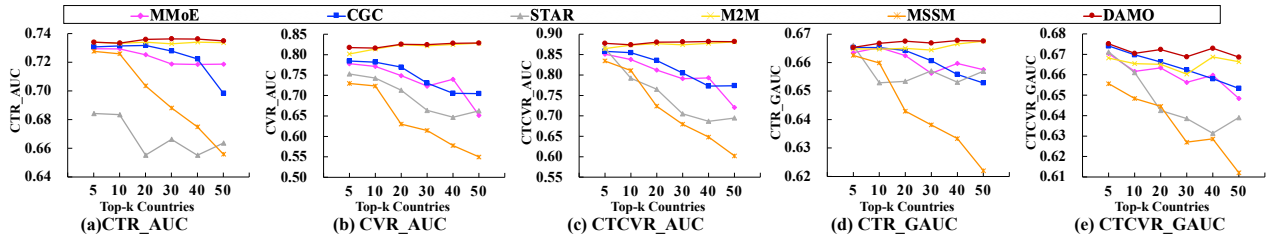


Figure 3: Accuracy comparison under different sampled industrial sub-datasets from top-K domains with varying K.

cessing, we have

$$\mathbf{x}_R \approx \mathbf{D}_R \mathbf{z}_R,\ FC(\bar{\mathbf{x}}) \approx \mathbf{D}_x \mathbf{z}_x. \tag{10}$$

Afterward, we concatenate $\mathbf{z}_R$ and $\mathbf{z}_x$ to generate the following sample representation to learn a view dictionary:

$$(\mathbf{z}_R \oplus \mathbf{z}_x) = \mathbf{D}_v \mathbf{z}_v, \tag{11}$$

wherein $\mathbf{D}_v$ is the learned view dictionary and $\mathbf{z}_v$ is the view coding (representation coefficient vector). Note that $\mathbf{D}_v$ is an undercomplete dictionary, which is different from $\mathbf{D_R}$ and $\mathbf{D_x}$. Besides, we try to enforce that each column vector d_v in $\mathbf{D}_v$ represents a view sub-network, and the sparse coding $\mathbf{z}_v$ will be further utilized to generate the view-specific weight vector β of all views by the softmax function $\mathcal{S}$:

$$\beta = \mathcal{S}(\mathbf{z}_v). \tag{12}$$

Particularly, to make sparse coding result consistent with that of the CTR/CVR prediction result, we generate an additional evaluation loss $\ell(y, \hat{y}_d)$ by a FC layer FC as below:

$$\min \ell(y, \hat{y}_d), \tag{13}$$

$$s.t. \quad \hat{y}_d = FC(\mathbf{z}_v). \tag{14}$$

Objective Function

We learn the proposed DAMO by minimizing the overall loss function as below:

$$\mathcal{L}_{DAMO} = \ell\left(y, \hat{y}\right) + \lambda_1 \sum_{v=1}^{V} \left(\beta_v \times \ell\left(y, \hat{y}^{(v)}\right)\right) + \lambda_2 \ell\left(y, \hat{y}_d\right), \tag{15}$$

wherein λ_1 and λ_2 are two hyperparameters. β_v is the v-th element of view-specific weight vector β obtained in Eq 12. In Eq 15, the first two terms are used to ensure that all view sub-networks can give out relatively accurate prediction results, while the last one is used to guide training the lightweight view-attention estimator. $\ell(\cdot)$ can be formulated as the cross-entropy loss or other task-related functions.

Experiments

In this section, we will introduce our experiments on one publicly available dataset Ali-CCP (Ma et al. 2018b) and one industrial dataset. More importantly, we further report the online performance of DAMO on a real-world e-commerce platform. Results demonstrate that DAMO is superior to existing methods in terms of both efficacy and scalability.

Dataset and Experimental Setup

Datasets **The industrial dataset** is a billion-level industrial dataset collected from 238 countries, where the training and testing sets are created by user click and purchase logs from a specific 14-day period and the subsequent day, respectively. As described before, data originating from every country can be regarded as constituting a distinct domain. We introduce two settings on these domains during training where the *Setting 1* focus on the top 50 of them w.r.t. their data sizes, instead of working on all 238 domains, to facilitate testing and comparison. Furthermore, we uniformly random sample 1% of data out of these top domains. And *Setting 2* works on the 5 clusters generated by performing k-means on all 238 domains. In contrast, **Ali-CCP** (Alibaba Click and Conversion Prediction) is a public dataset collected from real-world click/purchase logs of the recommendation system in Taobao [1]. We follow the same data split as (Xi et al. 2021). In both datasets, the majority of down-sampling (He and Garcia 2009) is applied to deal with data-imbalance problems of negative and positive samples. Their overall statistics can be found in the appendix.

Baselines We compare our DAMO with the existing SOTA methods, including MMoE (Ma et al. 2018a), CGC (Tang et al. 2020), STAR (Sheng et al. 2021), M2M (Zhang et al. 2022a) and MSSM (Ding et al. 2021). See more details about these methods in the appendix.

Implementation Details All experiments are conducted on Tesla A100×60 with Tensorflow framework and Adam optimizer. As suggested in (Zhang et al. 2022b), we set the number of epochs to 1 in all methods to prevent overfitting. All methods, including both DAMO and baselines, treat CTR and CVR as two independent tasks. Thus results of CTR and CVR are from two individual models for each method. More details about the network architecture and hyperparameter settings can be found in the appendix.

Evaluation Metrics AUC and GAUC (Zhou et al. 2018) are popular evaluation metrics in e-commerce literature. Specifically, the former evaluates both intra-user and inter-user orders by ranking all the items with predicted CTR. And the latter focuses more on intra-user order by averaging AUC over users, which proves to be more relevant to online performance (Zhou et al. 2018) (see more details in the appendix). In this work, we mainly report the averaged results of five metrics, including CTR_AUC, CVR_AUC,

[1] https://tianchi.aliyun.com/dataset/408

Method	AUC			GAUC	
	CTR	CVR	CTCVR	CTR	CTCVR
MMoE	0.7233	0.7354	0.8009	0.6607	0.6599
CGC	0.7237	0.7461	0.8166	0.6607	0.6639
STAR	0.6680	0.6970	0.7516	0.6566	0.6472
M2M	<u>0.7336</u>	<u>0.8191</u>	<u>0.8746</u>	<u>0.6655</u>	<u>0.6657</u>
MSSM	0.6960	0.6375	0.7165	0.6431	0.6360
DAMO	**0.7350**	**0.8233**	**0.8794**	**0.6669**	**0.6714**

Table 1: Average performance of different methods on top5/10/20/30/40/50 domains on the industrial dataset [2].

CTCVR_AUC, CTR_GAUC, and CTCVR_GAUC, over three runs to eliminate the fluctuations caused by distributed training strategy. Their standard deviations (STD) are also reported when necessary.

Offline Performance

Accuracy and efficiency under Setting 1 of the industrial dataset: We report the overall accuracy of all methods under setting 1 of the industrial dataset in Fig. 3. Unsurprisingly, DAMO outperforms existing methods with a noticeable margin in terms of the best performance under all five evaluation metrics (refer to Tab. 1). Moreover, the superiority of DAMO is consistent among all settings where all methods are required to work on top-k domains and $k \in \{5, 10, 20, 30, 50\}$. While most baseline methods' performances decline with the increasing number of domains, DAMO maintains high performance with minimum fluctuation under all evaluation metrics (see Fig. 3).

Besides high performance, DAMO is noticeably less computational-intense. We report the overall computational cost of all methods in terms of FLOPs and parameters in Fig. 4. Unlike most of the methods whose computational cost increases w.r.t. number of domains (*e.g.*MSSM, CGC, STAR), that of DAMO remains invariant. Moreover, DAMO almost always has the minimum FLOPs and parameters, *e.g.* 6.28B FLOPs and 3.19M parameters, among all methods. Interestingly, though M2M is the second-best in terms of accuracy, we observe that it requires three times the computational resources compared to DAMO.

In conclusion, we can see that DAMO is capable of beating the SOTA methods with a consistent yet lowest computation cost. Our experiments under setting 1 of the industrial dataset clearly validate the advantage of DAMO in terms of efficacy, efficiency, and scalability.

Accuracy under Setting 2 of the industrial dataset: We further report the overall performances of all methods under setting 2 of the industrial dataset. Please note that the overall data size under setting 2 is much larger than that of setting 1, which proves the scalability of DAMO in another dimension. Although clustering enables other methods to predict CTR/CVR with an acceptable efficiency, it is not an optimal solution due to disrupting domain correlations. As can be found in Tab. 2, our DAMO outperforms five baselines under all evaluation metrics, demonstrating that DAMO is still desired under a limited number of domains (Here we treat one cluster as one domain). In addition, we also observe the lower standard deviation of DAMO compared to other baselines, reflecting the stability of DAMO.

[2]The best and second-best results in tables are highlighted in boldface and underlined, respectively.

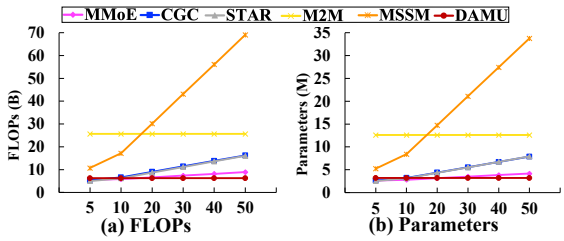


Figure 4: FLOPs and parameters of different methods on top 5/10/20/30/40/50 domains of sampled industrial data.

We also introduce two variations of DAMO, namely DAMO-clusterID and DAMO-countryID. Compared to our final DAMO, we observe that removing either country or cluster ID leads to inferior performances. This is understandable as the former inevitably would lose domain information in the original space during clustering. While the latter may encounter cold start issues from the ID embeddings of some data-limited countries.

Together with our previous results under setting 1, DAMO is undoubtedly the optimal choice in various aspects, including reliability and cost-effectiveness.

Accuracy on Ali-CCP Ali-CCP contains three domains only. GAUC is statistically meaningless in Ali-CCP as there is only one sample for each user in this dataset. We report the overall performance on Ali-CCP in Tab. 4. Again, our DAMO surpasses all baselines under all three evaluation metrics. We would like to further highlight that the performance gap on the extremely sparse domain (*i.e.*, the second domain) is more significant than that of other domains. We refer the readers to the appendix for more of the full table.

Though M2M performs well on the industrial dataset, it behaves poorly on Ali-CCP. Overall, DAMO gives the best performance under various datasets, showcasing its superiority in terms of generalizability.

Ablation Study

Finally, we perform ablation studies on DAMO[3]. We introduce three variations of DAMO: 1) w/o **C**orrelated **M**ulti-view **E**xtractor (CME), which is replaced by shared and specific experts in CGC; 2) **D**omain-aware **M**ulti-view **P**redictor (DMP), replaced by a gating network followed by MLP (*i.e.*, the views number is 1); and 3) w/o **L**ightweight **V**iew-attention **E**stimator (LVE) with $\beta_v = 0.2, \forall v$.

As can be found in Tab. 3, removing any of the three components of DAMO leads to performance drops, demonstrating the effectiveness of all components. Interestingly, these

[3]More analysis such as hyperparameter sensitivity experiments and model interpretability can be found in the appendix.

Method	CTR_AUC	CVR_AUC	CTCVR_AUC	CTR_GAUC	CTCVR_GAUC
MMoE	0.7691 ± 1.4e-4	0.8537 ± 9.6e-4	0.9164 ± 8.3e-4	0.6752 ± 4.3e-4	0.6997 ± 5.9e-4
CGC	0.7689 ± 1.2e-5	0.8538 ± 1.3e-3	0.9157 ± 7.2e-4	0.6744 ± 4.9e-5	0.6992 ± 6.0e-4
STAR	0.7689 ± 2.4e-4	0.8524 ± 8.7e-4	0.9163 ± 7.9e-4	0.6751 ± 2.0e-4	0.6985 ± 9.3e-4
M2M	0.7692 ± 9.9e-5	0.8533 ± 1.2e-4	0.9160 ± 6.5e-5	0.6749 ± 3.3e-4	0.6997 ± 5.8e-4
MSSM	0.7687 ± 1.1e-4	0.8530 ± 1.4e-3	0.9145 ± 1.2e-3	0.6745 ± 5.8e-5	0.7011 ± 2.9e-4
DAMO-clusterID	0.7694 ± 4.1e-4	0.8540 ± 1.2e-3	0.9166 ± 1.0e-3	0.6753 ± 2.8e-5	0.7011 ± 4.1e-4
DAMO-countryID	0.7692 ± 1.3e-5	0.8540 ± 4.3e-5	0.9160 ± 3.9e-4	0.6756 ± 3.5e-4	0.7013 ± 2.5e-4
DAMO	**0.7698 ± 1.1e-4**	**0.8543 ± 4.2e-5**	**0.9174 ± 2.6e-4**	**0.6757 ± 3.9e-6**	**0.7023 ± 3.9e-4**

Table 2: Offline result on the industrial dataset. Note that due to a large amount of users and samples in our dataset, an improvement of 0.05% in AUC and GAUC in the offline evaluation is significant enough to bring online gains for the business.

Model	CTR_AUC	CVR_AUC	CTCVR_AUC	CTR_GAUC	CTCVR_GAUC
all	**0.7698 ± 1.1e-4**	**0.8543 ± 4.2e-5**	**0.9174 ± 2.6e-4**	**0.6757 ± 3.9e-6**	**0.7023 ± 3.9e-4**
w/o CME	0.7692 ± 4.3e-5	0.8539 ± 1.1e-3	0.9167 ± 6.6e-4	0.6749 ± 1.6e-4	0.7006 ± 1.1e-3
w/o DMP	0.7690 ± 2.1e-4	0.8541 ± 1.5e-3	0.9164 ± 1.6e-3	0.6747 ± 1.0e-3	0.6994 ± 9.3e-4
w/o LVE	0.7697 ± 3.5e-4	0.8536 ± 6.9e-4	0.9155 ± 7.3e-4	0.6755 ± 2.4e-4	0.7011 ± 7.7e-4

Table 3: Performance on the ablation study.

Method	CTR_AUC	CVR_AUC	CTCVR_AUC
MMoE	0.6092	0.6222	0.6145
CGC	0.6103	0.6258	0.6159
STAR	0.6108	0.6220	0.5987
M2M	0.6076	0.6193	0.6078
MSSM	0.6097	0.5923	0.5812
DAMO	**0.6129**	**0.6287**	**0.6170**

Table 4: Offline result on Ali-CCP.

Metric	CTR	CVR	GMV
Overall	+1.21%	+1.76%	+1.66%

Table 5: Improvement over base model in online A/B testing.

three components contribute to various aspects of DAMO. For instance, the absence of the CME weakens the sharing ability, leading to a significant performance decline compared to DAMO on CTR-related metrics. In contrast, LVE is more important to CVR as removing LVE brings worse performance on CVR-related metrics. This suggests that the LVE guides differentiated views through constraint coefficients, thereby enhancing differentiation ability. Note that although the contribution of LVE to CTR-related metrics is not as noticeable, it contributes to stable results (with a small standard deviation). Lastly, DMP seems to be favored in general as removing it produces the worst results across multiple metrics. This is reasonable because DMP describes the common and differentiated knowledge across views, and the collaboration and complementarity of multiple views are key factors of model stability.

Online A/B Test

We further deploy our DAMO on one of the largest e-commerce platforms. Compared to the offline setting, online tests use real-time, real-world user behavior data, covering the actual interaction of users. It can more realistically reflect the performance of methods in real applications.

Specifically, we conduct live online experiments in an A/B testing framework for two weeks. For the control group, 4.0% of users are randomly selected and presented with search results generated by the previous version of the online ranking model, which is a highly optimized deep model for CTR/CVR prediction. For the experiment group, 4.0% of users are randomly selected and presented with search results generated by the new version of the ranking model, which replaces the CTR/CVR prediction results with the help of DAMO. The metrics we evaluate online are online CTR, CVR, and Gross Merchandises Value(GMV). For security reasons, we only report the relative improvement of DAMO in Tab. 5. Clearly, DAMO improves CTR, CVR, and GMV by 1.21%, 1.76%, and 1.66%, respectively, showcasing DAMO's strong advantages in real-world applications.

Conclusion

This paper proposes a novel CTR/CVR prediction model namely DAMO to address the scalability and efficiency problem in cross-border e-commerce search scenarios. DAMO exploits the domain correlation from a multi-view perspective. By decoupling the strong correlation between model parameters and the number of domains, DAMO significantly reduces resource consumption. Experiment results on both industrial and public datasets indicate that DAMO outperforms existing models in terms of both CTR/CVR prediction accuracy and computational efficiency.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 62125201), Zhejiang Provincial Natural Science Foundation of China (No.LZ23F020007), National Natural Science Foundation of China (No.61972119, No. 62020106007), and Alibaba International Digital Commerce Group through Alibaba Innovative Research Program.

We appreciate the valuable contributions of Buyu Liu and Jian Zhang during the revision of this paper. Their insightful feedback, guidance, and constructive criticism greatly improved the quality and clarity of the manuscript. Their dedication to excellence and expertise in the field has been instrumental in shaping the final version of this work.

References

Cheng, H.-T.; Koc, L.; Harmsen, J.; Shaked, T.; Chandra, T.; Aradhye, H.; Anderson, G.; Corrado, G.; Chai, W.; Ispir, M.; Anil, R.; Haque, Z.; Hong, L.; Jain, V.; Liu, X.; and Shah, H. 2016. Wide & Deep Learning for Recommender Systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, DLRS 2016, 7–10. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-4795-2.

Dai, S.; Lin, H.; Zhao, Z.; Lin, J.; Wu, H.; Wang, Z.; Yang, S.; and Liu, J. 2021. POSO: Personalized Cold Start Modules for Large-scale Recommender Systems. arXiv:2108.04690.

Ding, K.; Dong, X.; He, Y.; Cheng, L.; Fu, C.; Huan, Z.; Li, H.; Yan, T.; Zhang, L.; Zhang, X.; and Mo, L. 2021. MSSM: A Multiple-Level Sparse Sharing Model for Efficient Multi-Task Learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, 2237–2241. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380379.

Elkahky, A. M.; Song, Y.; and He, X. 2015. A Multi-View Deep Learning Approach for Cross Domain User Modeling in Recommendation Systems. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, 278–288. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. ISBN 9781450334693.

He, H.; and Garcia, E. A. 2009. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.*, 21(9): 1263–1284.

Jiang, Y.; Li, Q.; Zhu, H.; Yu, J.; Li, J.; Xu, Z.; Dong, H.; and Zheng, B. 2022. Adaptive Domain Interest Network for Multi-domain Recommendation. In Hasan, M. A.; and Xiong, L., eds., *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, 3212–3221. New York, NY, USA: Association for Computing Machinery.

Li, P.; Li, R.; Da, Q.; Zeng, A.-X.; and Zhang, L. 2020. Improving Multi-Scenario Learning to Rank in E-Commerce by Exploiting Task Relationships in the Label Space. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, 2605–2612. New York, NY, USA: Association for Computing Machinery. ISBN 9781450368599.

Li, S.; Xie, R.; Zhu, Y.; Ao, X.; Zhuang, F.; and He, Q. 2022. User-Centric Conversational Recommendation with Multi-Aspect User Modeling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, 223–233. New York, NY, USA: Association for Computing Machinery. ISBN 9781450387323.

Li, Z.; Liu, J.; Tang, J.; and Lu, H. 2015. Robust structured subspace learning for data representation. *IEEE transactions on pattern analysis and machine intelligence*, 37(10): 2085–2098.

Liu, A.-A.; Xu, N.; Nie, W.-Z.; Su, Y.-T.; and Zhang, Y.-D. 2018. Multi-domain and multi-task learning for human action recognition. *IEEE Transactions on Image Processing*, 28(2): 853–867.

Ma, J.; Zhao, Z.; Yi, X.; Chen, J.; Hong, L.; and Chi, E. H. 2018a. Modeling Task Relationships in Multi-Task Learning with Multi-Gate Mixture-of-Experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, 1930–1939. New York, NY, USA: Association for Computing Machinery. ISBN 9781450355520.

Ma, X.; Zhao, L.; Huang, G.; Wang, Z.; Hu, Z.; Zhu, X.; and Gai, K. 2018b. Entire Space Multi-Task Model: An Effective Approach for Estimating Post-Click Conversion Rate. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, 1137–1140. New York, NY, USA: Association for Computing Machinery. ISBN 9781450356572.

Mallat, S. G. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7): 674–693.

Rodríguez-Domínguez, U.; and Dalmau, O. 2020. Hierarchical Discriminative Deep Dictionary Learning. *IEEE Access*, 8: 142680–142690.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q. V.; Hinton, G. E.; and Dean, J. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Sheng, X.-R.; Zhao, L.; Zhou, G.; Ding, X.; Dai, B.; Luo, Q.; Yang, S.; Lv, J.; Zhang, C.; Deng, H.; and Zhu, X. 2021. One Model to Serve All: Star Topology Adaptive Recommender for Multi-Domain CTR Prediction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, 4104–4113. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384469.

Tai, C.-Y.; Wu, M.-R.; Chu, Y.-W.; Chu, S.-Y.; and Ku, L.-W. 2020. MVIN: Learning Multiview Items for Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information*

Retrieval, SIGIR '20, 99–108. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380164.

Tang, H.; Liu, J.; Zhao, M.; and Gong, X. 2020. Progressive Layered Extraction (PLE): A Novel Multi-Task Learning (MTL) Model for Personalized Recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems*, RecSys '20, 269–278. New York, NY, USA: Association for Computing Machinery. ISBN 9781450375832.

Wang, W.; Arora, R.; Livescu, K.; and Bilmes, J. 2015. On Deep Multi-View Representation Learning. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, 1083–1092. JMLR.org.

Wu, Y.; Xie, R.; Zhu, Y.; Ao, X.; Chen, X.; Zhang, X.; Zhuang, F.; Lin, L.; and He, Q. 2022. Multi-View Multi-Behavior Contrastive Learning In Recommendation. In *Database Systems for Advanced Applications: 27th International Conference, DASFAA 2022, Virtual Event, April 11–14, 2022, Proceedings, Part II*, 166–182. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-031-00125-3.

Xi, D.; Chen, Z.; Yan, P.; Zhang, Y.; Zhu, Y.; Zhuang, F.; and Chen, Y. 2021. Modeling the Sequential Dependence among Audience Multi-Step Conversions with Multi-Task Learning in Targeted Display Advertising. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, 3745–3755. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383325.

Zhang, Q.; Liao, X.; Liu, Q.; Xu, J.; and Zheng, B. 2022a. Leaving No One Behind: A Multi-Scenario Multi-Task Meta Learning Approach for Advertiser Modeling. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, 1368–1376. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391320.

Zhang, Z.; Sheng, X.; Zhang, Y.; Jiang, B.; Han, S.; Deng, H.; and Zheng, B. 2022b. Towards Understanding the Overfitting Phenomenon of Deep Click-Through Rate Models. In Hasan, M. A.; and Xiong, L., eds., *Proceedings of the 31st ACM International Conference on Information Knowledge Management*, CIKM '22, 2671–2680. New York, NY, USA: Association for Computing Machinery.

Zhou, G.; Zhu, X.; Song, C.; Fan, Y.; Zhu, H.; Ma, X.; Yan, Y.; Jin, J.; Li, H.; and Gai, K. 2018. Deep Interest Network for Click-Through Rate Prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, 1059–1068. New York, NY, USA: Association for Computing Machinery. ISBN 9781450355520.

Zhu, Y.; Sheng, Q.; Cao, J.; Nan, Q.; Shu, K.; Wu, M.; Wang, J.; and Zhuang, F. 2022. Memory-Guided Multi-View Multi-Domain Fake News Detection. *IEEE Transactions on Knowledge and Data Engineering*.

Zou, X.; Hu, Z.; Zhao, Y.; Ding, X.; Liu, Z.; Li, C.; and Sun, A. 2022. Automatic Expert Selection for Multi-Scenario and Multi-Task Search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, 1535–1544. New York, NY, USA: Association for Computing Machinery. ISBN 9781450387323.