Adversarial Multimodal Representation Learning for Click-Through Rate Prediction

Xiang Li, Chao Wang, Jiwei Tan, Xiaoyi Zeng, Dan Ou, Bo Zheng
Alibaba Group, Hangzhou & Beijing, China
{leo.lx,xiaoxuan.wc,jiwei.tjw,yuanhan,oudan.od,bozheng}@alibaba-inc.com

ABSTRACT

For better user experience and business effectiveness, Click-Through Rate (CTR) prediction has been one of the most important tasks in E-commerce. Although extensive CTR prediction models have been proposed, learning good representation of items from multimodal features is still less investigated, considering an item in E-commerce usually contains multiple heterogeneous modalities. Previous works either concatenate the multiple modality features, that is equivalent to giving a fixed importance weight to each modality; or learn dynamic weights of different modalities for different items through technique like attention mechanism. However, a problem is that there usually exists common redundant information across multiple modalities. The dynamic weights of different modalities computed by using the redundant information may not correctly reflect the different importance of each modality. To address this, we explore the complementarity and redundancy of modalities by considering modality-specific and modality-invariant features differently. We propose a novel Multimodal Adversarial Representation Network (MARN) for the CTR prediction task. A multimodal attention network first calculates the weights of multiple modalities for each item according to its modality-specific features. Then a multimodal adversarial network learns modality-invariant representations where a double-discriminators strategy is introduced. Finally, we achieve the multimodal item representations by combining both modality-specific and modality-invariant representations. We conduct extensive experiments on both public and industrial datasets, and the proposed method consistently achieves remarkable improvements to the state-of-the-art methods. Moreover, the approach has been deployed in an operational E-commerce system and online A/B testing further demonstrates the effectiveness.

CCS CONCEPTS

• **Information systems** → **Content ranking**; **Online shopping**; **Recommender systems**; **Content analysis and feature selection**; **Data encoding and canonicalization**.

KEYWORDS

multimodal learning, adversarial learning, recurrent neural network, attention, representation learning, e-commerce search

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380163

ACM Reference Format:
Xiang Li, Chao Wang, Jiwei Tan, Xiaoyi Zeng, Dan Ou, Bo Zheng. 2020. Adversarial Multimodal Representation Learning for Click-Through Rate Prediction. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3366423.3380163

1 INTRODUCTION

Large E-commerce portals such as Taobao and Amazon are serving hundreds of millions of users with billions of items. For better user experience and business effectiveness, Click-Through Rate (CTR) prediction has been one of the most important tasks in E-commerce. As widely studied by both academia and industry, extensive CTR prediction models have been proposed. Nevertheless, it is also an effective way to improve the CTR prediction accuracy through better mining and leveraging the multimodal features of items, since an item in E-commerce usually contains multiple heterogeneous modalities including IDs, image, title and statistic. Therefore, in this paper we make our effort to improve the CTR prediction accuracy by learning better representations for items of multiple modalities.

To leverage the multiple modalities for better item representations, a straightforward way [20, 38, 39] is to concatenate the multiple modality features, which is equivalent to giving a fixed importance weight to each modality regardless of different items. However, the importance of a modality may be different according to different items, and ideal item representations should be able to weigh different modalities dynamically so that emphasis on more useful signals is possible. For example, in the clothing category, whether users will click an item or not is highly affected by observing the images, so greater importance should be given to the image feature. On the contrary, in the cell phone and grocery food categories, the statistic feature of items reflects the popularity of items, while there is little difference between the images of items. A conceivable improvement [14, 31] is to learn dynamic weights of different modalities and emphasis on more useful signals through technique like attention mechanism. However, a problem is that there usually exists common redundant information across multiple modalities. The dynamic weights of different modalities computed by using the redundant information may not correctly reflect the different importance of each modality.

Due to the above reason, our motivation is that the weight of completely independent information across modalities should be dynamic, and the weight of information shared by different modalities should be fixed. To address this, we divide multiple modality features into modality-specific (exist in one modality, and should have dynamic weights) and modality-invariant features (redundant in different modalities, and should have fixed weights). Take a dress which is displayed by an image with a title *Girls Ballet Tutu*

Zebra Hot Pink for example. The item involves a latent semantic feature of the material, such as yarn, which can be expressed by its image while not involved in its title, so the material feature is considered as the modality-specific (latent) feature for this example. The item also involves a common latent semantic feature of color (hot pink) in the subspace of both its image and title features, so the color feature is considered as the modality-invariant (latent) feature for this example. The key idea is that modality-specific features provide an effective way to explore dynamic contributions of different modalities, while modality-invariant features should have a fixed contribution and can be used as supplementary knowledge for comprehensive item representations.

To the best of our knowledge, this is the first work that learns multimodal item representations by exploiting modality-specific and modality-invariant features differently. To achieve this, we propose a Multimodal Adversarial Representation Network (MARN) to deal with this challenging task. In MARN, a modality embedding layer extracts embedding vectors from multiple modalities and decomposes each embedding vector into modality-specific and modality-invariant features. Then, a multimodal attention network calculates the weights of multiple modalities for each item according to its modality-specific features. Also, a multimodal adversarial network learns modality-invariant representations where a double-discriminators strategy is introduced. The double-discriminators strategy is designed to identify the potential modalities involving common features across modalities and drive knowledge transfer between modalities. Finally, we achieve multimodal item representations by combining both modality-specific and modality-invariant representations. The contributions of the paper are summarized as:

- The proposed MARN introduces a novel multimodal representation learning method for multimodal items, which can improve the CTR prediction accuracy in E-commerce.
- We explore the complementarity and redundancy of modalities by considering modality-specific and modality-invariant features differently. To achieve discriminative representations, we propose a multimodal attention fusion network. Moreover, to achieve common representations across modalities, we propose a double-discriminators multimodal adversarial network.
- We perform extensive experiments on both public and industrial datasets. MARN significantly outperforms the state-of-the-art methods. Moreover, the approach has been deployed in an operational E-commerce system and online A/B testing further demonstrates the effectiveness.

2 RELATED WORK

2.1 Multimodal Learning

Representing raw data in a format that a computational model can work with has always been a big challenge in machine learning [1]. Multimodal representation learning methods aim to represent data using information from multiple modalities. Neural networks have become a very popular method for unimodal representations [2, 7]. They can represent visual or textual data and are increasingly used in the multimodal domain [19, 22]. To construct multimodal representations using neural networks, each modality starts with several individual neural layers followed by a hidden layer that projects the modalities into a joint space [33]. The joint multimodal representations are then passed through extra multiple hidden layers or used directly for prediction.

The above multimodal learning methods typically treat different modalities equally by concatenation, which is equivalent to giving the fixed weight for each modality. In practical, item representation learning methods should be capable of assigning dynamic importance weights to each modality according to different items. Therefore, we propose a multimodal attention network to learn different weights of multiple modalities for each item, so that better item representations can be achieved by the weighted combination.

2.2 Adversarial Transfer Learning

Features of multiple modalities may contain redundant information, which should be eliminated when computing the different contributions of modalities. To exploit the redundant information, the common subspace across different modalities should be exploited. Adversarial transfer learning [4, 8] is inspired by Generative Adversarial Nets [9], which enables domain adaptation in deep networks that can be trained on a large amount of labeled data from the source domain and unlabeled data from the target domain. Wang et al. [30] propose an adversarial cross-modal retrieval method, which seeks an effective common subspace based on adversarial learning. Pei et al. [23] capture multimodal structures to enable fine-grained alignment of different data distributions based on multiple domain discriminators. Unlike previous adversarial transfer methods that solely match distribution at domain-level, Xie et al. [34] propose to match distribution at class-level and align features semantically without any target labels. Yu et al. [36] propose a novel Dynamic Adversarial Adaptation Network to learn domain-invariant representations while quantitatively evaluate the relative importance of the marginal and conditional domain distributions.

The above adversarial learning methods are still a one-to-one adversarial paradigm, although the items in E-commerce involve multiple modalities. Moreover, different modality features involve different degrees of common features. Therefore, we propose a novel double-discriminators multimodal adversarial network to learn common latent subspace across multiple modalities.

2.3 Context Aware Personalization Model

In recent years, there have been growing numbers of researches on the personalization based on deep neural networks for recommending music [29], news [21], videos [5], and jobs [3].

Personalization for search and recommendation is typically based on user behaviors, where the RNN-based model is a good choice due to the sequential characteristic of user behaviors. Hidasi et al. [11] apply the RNN model to infer users' future intentions based on their previous click behavior sequence. Tan et al. [28] present several extensions to the basic RNN model to enhance the performance of recurrent models. Ni et al. [20] adopt LSTM and the attention mechanism to model the user behavior sequence. Compared to sequence-independent approaches, these methods can significantly improve the CTR prediction accuracy and most of these techniques have been deployed in real-world applications [20, 32, 38, 39].

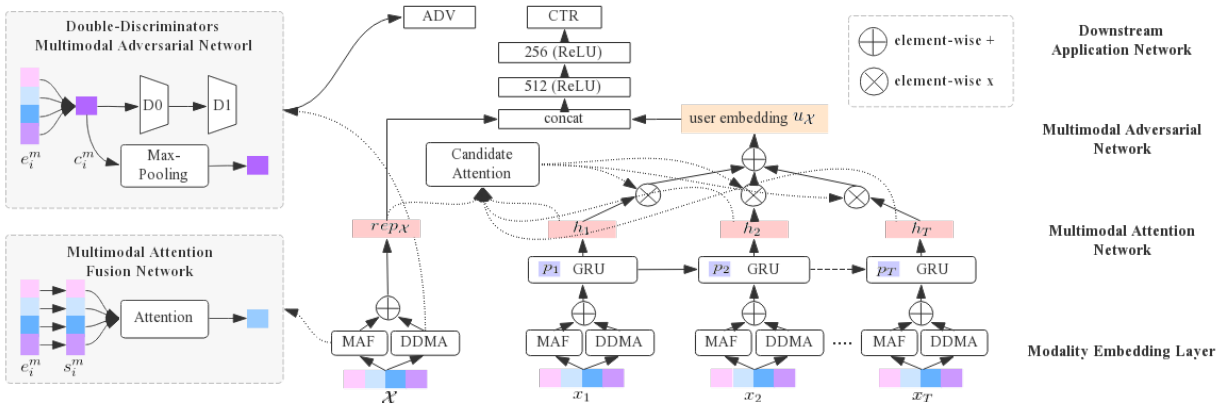


Figure 1: The architecture of MARN.

3 OUR PROPOSED METHOD

The typical framework of CTR prediction is to take a user behavior sequence $u = \{x_1, x_2, \ldots, x_T\}$ and the candidate item $\mathcal{X}$ as inputs, and aims to learn the probability that u clicks $\mathcal{X}$. Items are mostly represented by ID to learn its representation through embedding. When considering multiple modalities, previous works either first extract multimodal embeddings and fuse these embeddings through concatenation [20, 38, 39], or dynamically give different weights for each modality through an attention mechanism [14, 31]. Differently, we propose to divide multiple modality features into modality-specific and modality-invariant features. The overall architecture of the proposed MARN is presented in Figure 1. In our work, *Multimodal Attention Network* learns the dynamic weights of multiple modalities for each item according to its modality-specific features. *Multimodal Adversarial Network* learns common representations across multiple modalities for more comprehensive item representations. Besides the above two components, the *Modality Embedding Layer* and *Downstream Application Network* of our model are similar to related CTR prediction models.

3.1 Modality Embedding Layer

Each item in the user behavior sequence is associated with a behavior property and forms a user-item interaction pair $< x_i, p_i >$. The modality embedding layer is applied upon the sparse features of each item in the user behavior sequence and its behavior property to compress them into low-dimensional vectors.

3.1.1 Item Modality Embedding. An item x_i is represented by the multimodal information: i) **IDs**, unordered discrete feature including item ID, shop ID, brand ID and category ID; ii) **image**, pixel-level visual information; iii) **title**, word sequence; iv) **statistic**, historical exposure, click, transaction order/amount.

IDs: The IDs feature is represented as $[x_i^{id(1)}, \ldots, x_i^{id(F)}]$, which is a multi-hot vector ($[\cdot,\cdot]$ denotes vector concatenation). F is the number of IDs feature, and $x_i^{id(f)}$ is the f^{th} feature which is a one-hot vector representing an ID like item ID or shop ID. The embedding layer transforms the multi-hot vector into a low-dimensional vector e_i^{id} with an embedding lookup table, as shown in Equation 1:

$$e_i^{id} = [W_{emb}^1 x_i^{id(1)}, \ldots, W_{emb}^F x_i^{id(F)}], W_{emb}^f \in R^{d_{emb}^f \times V_f} \quad (1)$$

where d_{emb}^f is the dimension and V_f is the vocabulary size.

Image: Recent progress in computer vision shows that the learned semantic embeddings from the pixel-level visual information for classification tasks have good generalization ability [10, 27]. Thus for an input image, the image embedding is the output of pre-trained VGG16 [27] model with the last two layers for classification purpose removed, which result in a 4096-dimensional vector.

Title: The title containing h words (padded where necessary) is represented as an $h \times d_{emb}^{term}$ matrix, where each word is represented as a d_{emb}^{term}-dimensional vector. As suggested by Word2Vec [18], we set $d_{emb}^{term} = 300$. We design a fast convolutional network for title embeddings which uses multiple filters with varying window sizes $n = 2, 3, 4$ to obtain the n-gram features, following Kim [13].

Statistic: We found it difficult in learning a good embedding directly on continuous statistic feature. Following Yuanfei et al. [37], we adopt multi-granularity discretization, which discretizes each numerical feature into two, rather than only one, categorical features, each with a different granularity, ranging from 0 to 9 and 0 to 99 according to the number of items. We then perform categorial feature lookups to obtain two 8-dimensional vectors.

3.1.2 Behavior Property Embedding. The behavior property p_i describes the type and time of items in the user behavior sequence. Behavior type is a one-hot vector representing click, add-to-cart or purchase. Behavior time is a statistic feature indicating the seconds from the time it happens to the current recommendation time. We transform each behavior property into embedding $p_i = e_i^p$ and treat the the behavior property embedding p_i as a strong signal to reflect the importance of each behavior.

3.2 Multimodal Attention Network

Utilizing multiple modality features is often effective to improve the performance of CTR tasks. A straightforward way [20, 38, 39] is to concatenate the multiple modality features, which is equivalent to giving a fixed importance weight to each modality regardless of

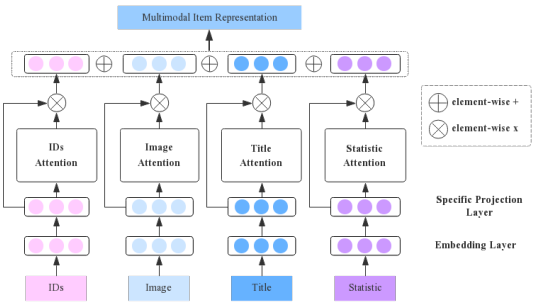


Figure 2: Illustration of multimodal attention fusion.

different items. A conceivable improvement [14, 31] is to dynamically distinguish the contributions of different modalities through an attention mechanism. However, features of multiple modalities may contain redundant information, which should be eliminated when computing the different contributions of modalities.

Due to the above reason, we propose to explore the complementarity and redundancy of modalities by considering modality-specific and modality-invariant features differently. More precisely, the distributions of modality-specific features are separated, whereas the distributions of modality-invariant features are close to each other. To address this, we decompose each item embedding vector into modality-specific and modality-invariant features, i.e., s_i^m, c_i^m. More specifically, we project each item modality embedding into vector e_i^m of the same dimension, and then apply two nonlinear projection layers to obtain two 256-dimensional vectors according to $s_i^m, c_i^m = \mathcal{S}_m(e_i^m), \mathcal{I}(e_i^m)$. $\mathcal{S}_m(\cdot)$ is an independent projection matrix of each modality m for extracting modality-specific features, while the projection matrix $\mathcal{I}(\cdot)$ is shared across modalities for learning modality-invariant features.

Therefore, we propose a multimodal attention fusion (MAF) network shown in Figure 2, to learn dynamic weights according to its modality-specific features s_i^m and then obtain the modality-specific representation s_i by the weighted summation:

$$s_i = \sum_{m=1}^{M} atten_i^m \odot s_i^m \tag{2}$$

$$atten_i^m = tanh(W_m^\top \cdot s_i^m + b_m) \tag{3}$$

where $atten_i^m$ controls the attention weight of modality m, and $\odot$ denotes element-wise multiplication. The parameters $W_m^\top$ is matrix with the size of $d \times d$, and b_m is a vector with the size of $d \times 1$, where $d = 256$. The vector attention adjusts the importance weights of each dimension for the modality-specific features.

3.3 Multimodal Adversarial Network

Since the weights of the $\mathcal{S}_m(\cdot)$ and the $\mathcal{I}(\cdot)$ are only supervised by the label of the CTR task, we cannot expect that MARN has successfully learned the modality-specific and modality-invariant features. The distributions of modality-specific features extracted by $\mathcal{S}_m(\cdot)$ should be kept as far as possible, meanwhile, the distributions of modality-invariant features extracted by $\mathcal{I}(\cdot)$ should be drawn as close as possible. To address this, we design a multimodal adversarial network, which makes the modality-invariant feature extractor compete with a modality discriminator, in order to learn a common subspace across multiple modalities. Moreover, we propose a modality-specific discriminator to supervise the modality-specific feature extractor, so that the learned multimodal information has very good separability, leading to the modality-specific features.

3.3.1 Cross-Modal Adversarial. The general cross-modal adversarial method [30] seeks an effective common subspace across modalities, which is built around a minimax game, as:

$$\begin{aligned} \min_{\mathcal{I}} \max_{\mathcal{D}} \mathcal{L}_D = & \mathbb{E}_{x \sim px^1} [log(\mathcal{D}(\mathcal{I}(x)))] \\ & + \mathbb{E}_{x \sim px^2} [log(1 - \mathcal{D}(\mathcal{I}(x)))] \end{aligned} \tag{4}$$

where px^1 and px^2 are distributions of two modalities.

For $\mathcal{I}$ fixed, the optimal discriminator D is:

$$D^*_{\mathcal{I}}(x) = \frac{px^1(x)}{px^1(x) + px^2(x)} \tag{5}$$

Given the optimum D^*, the minimax game of Equation 4 is:

$$\mathcal{L}_D = -log(4) + 2 \cdot JSD(px^1 || px^2) \tag{6}$$

Since the Jensen-Shannon divergence (JSD) between two distributions is always non-negative and achieves zero only when they are equal, $\mathcal{L}_D$ obtains the global minimum on $px^1 = px^2$ [9], which is the goal of cross-modal adversarial learning.

3.3.2 Double-Discriminators Multimodal Adversarial. The above cross-modal adversarial method focuses on a one-to-one adversarial paradigm, while the items in E-commerce involve multiple modalities. Moreover, different modality features involve different degrees of common latent features, and the degrees can be treated as guidance to confuse the modality discriminator towards better discrimination performance. For example, the item ID embedding should be assigned relatively small weights due to the uniqueness. On the contrary, the image and title embeddings that involve much potential common subspace should be emphasized for further confusing the modality discriminator.

To address this, we propose a novel double-discriminators multimodal adversarial (DDMA) network. More precisely, we bring two contributions. First, we expand the original one-to-one adversarial paradigm to multimodal scenarios. Second, we introduce a double-discriminators strategy. The first discriminator identifies modality-invariant features that are potentially from the common latent subspace across multiple modalities and also emphasizes the identified modality-invariant features for further confusing the second discriminator. Meanwhile, the second discriminator drives knowledge transfer between modalities so as to learn common latent subspace across multiple modalities.

The first discriminator D_0 is an M-class classifier and is given by $D_0(x) = softmax(\mathcal{D}_0(x))$, where $x = c = \mathcal{I}(e)$, e is the modality embedding, $\mathcal{I}$ is the invariant projection layer and c is the projected modality-invariant feature. Suppose that D_0 has converged to its optimum D_0^* and x belongs to modality i, $D_0^i(x)$ gives the likelihood of x in terms of modality i. If $D_0^{i*}(x) \approx 1$, then x highly involves few common features across modalities, since it can be perfectly discriminated from other modalities. Thus, the degree of x as $w^i(x)$

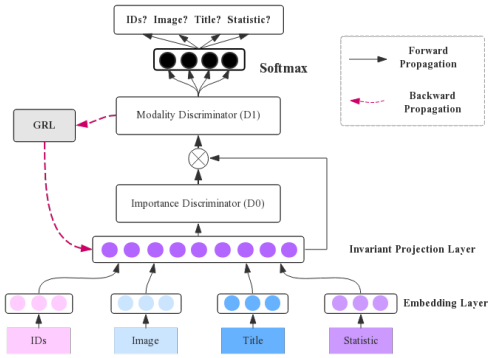


Figure 3: Illustration of double-discriminators multimodal adversarial network.

contributing to the modality-invariant feature should be inversely related to $D_0^i(x)$ according to $w^i(x) = 1 - D_0^i(x)$.

The second discriminator D_1 is also an M-class classifier, which is applied to reduce the JSD between modalities. After adding the degrees to the modality features for the discriminator D_1, the objective function of weighted multimodal adversarial network is:

$$\min_{\mathcal{I}} \max_{\mathcal{D}_1} \mathcal{L}_{D_1} = \mathbb{E}_{e \sim pe^m} \sum_{m=1}^{M} \left[w^m(x) log(\mathcal{D}_1(\mathcal{I}(e))) \right] \tag{7}$$

where $w^m(x) = w^m(\mathcal{I}(e))$ is a constant and independent of $\mathcal{D}_1$.

For fixed $\mathcal{I}$ and $\mathcal{D}_0$, the optimal output for D_1^i is:

$$D_1^{i*}(x) = \frac{w^i(x) px^i(x)}{\sum_{m=1}^{M} w^m(x) px^m(x)} \tag{8}$$

Given the optimum $D_1^*(x)$, the minimax game of Equation 7 is:

$$\begin{aligned}
\mathcal{L}_{D_1} &= \sum_{i=1}^{M} \int_x w^i(x) px^i(x) log(\frac{w^i(x) px^i(x)}{\sum_{m=1}^{M} w^m(x) px^m(x)}) dx \\
&= \sum_{i=1}^{M} \int_x w^i(x) px^i(x) log(\frac{w_i(x) px^i(x)}{\sum_{m=1}^{M} \frac{w^m(x) px^m(x)}{M}}) dx - M log(M) \\
&= \sum_{i=1}^{M} KL\left(w^i(x) px^i(x) \parallel \sum_{m=1}^{M} \frac{w^m(x) px^m(x)}{M} \right) - M log(M) \\
&= M \cdot JSD_{\frac{1}{M}, \ldots, \frac{1}{M}} \left(w^1(x) px^1(x), \cdots, w^M(x) px^M(x) \right) - M log(M)
\end{aligned} \tag{9}$$

As deduced in Equation 9, the minimax game of Equation 7 is essentially reducing JSD between M weighted modality features and obtains the optimum on $w^1(x)px^1(x) = \cdots = w^M(x)px^M(x)$.

Since $w^m(x) = w^m(\mathcal{I}(e^m))$ and $px^m(x) = c^m = \mathcal{I}(e^m)$, we can achieve modality-invariant representation through a max-pooling operation over the weighted modality-invariant features and take the maximum value, as $c_i = max\{w^1 c^1, \ldots, w^M c^M\}$.

3.3.3 Modality-Specific Discriminator. To encourage the modality-specific features s_i^m to be discriminated between multiple modalities, we propose a modality-specific modality discriminator D_s on the modality-specific features,where the discriminating matrix of D_s is shared with the second discriminator D_1.

Finally, we achieve the complementarity and redundancy of multiple modalities, resulting in the modality-specific representations s_i and modality-invariant representations c_i. Then, the multimodal item representation rep_i is produced by the element-wise summation of modality-specific and modality-invariant representations.

3.4 Downstream Application Network

3.4.1 Behaviors Integrating Layer. With the proposed item representation learning model, we can integrate the user behaviors with various kinds of RNN-based model. In this paper, we choose GRU [11] to model the dependencies between the user behaviors. However, GRU lacks of capturing two characteristics: i) different behavior type and time contribute differently to extracting users' interests; ii) whether to click a candidate item or not highly depends on the most related behaviors with respect to the candidate item. Moreover, extracting the key information in the complex behaviors further contributes towards learning better item representations.

Therefore, we propose an Attentional Property GRU (APGRU) layer. First, we modify the update gate of GRU to extract the most interested behavior state according to $u_t = \sigma(W^u rep_t + P^u p_t + U^u h_{t-1} + b^u)$, where p_t is the behavior property, rep_t is the t^{th} item representation, and σ is the sigmoid function.

Second, we concatenate the hidden state h_t and the candidate item rep_χ and apply a Multilayer Perceptron as the candidate attention network. Then we calculate the normalized attention weights, and the user embedding u_χ can be calculated as the weighted summation of the output hidden states.

3.4.2 CTR Prediction Layer. In this paper, we set the prediction layer to be a deep network with point-wise loss for the CTR task. The overall objective functions are:

$$\begin{aligned}
&\min_{\mathcal{F}, \mathcal{I}, \mathcal{S}_m, \mathcal{E}} \mathcal{L}_{CTR} = -\frac{1}{N} \sum_{n=1}^{N} [y_n log(\mathcal{F}(u_n, rep_n)) \\
&\qquad + (1 - y_n) log(1 - \mathcal{F}(u_n, rep_n))] \\
&\min_{\mathcal{D}_s, \mathcal{S}_m} \mathcal{L}_{D_s} = \lambda \left(-\mathbb{E}_{e, y \sim pe^{e,y}} \sum_{m=1}^{M} \mathbb{I}_{[m=y]} log \mathcal{D}_s(\mathcal{S}_m(e)) \right) \\
&\min_{\mathcal{D}_0} \mathcal{L}_{D_0} = -\mathbb{E}_{c, y \sim pc^{c,y}} \sum_{m=1}^{M} \mathbb{I}_{[m=y]} log \mathcal{D}_0(c) \\
&\min_{\mathcal{I}} \max_{\mathcal{D}_1} \mathcal{L}_{ADV} = \lambda \left(\mathbb{E}_{e \sim pe^m} \sum_{m=1}^{M} \left[w^m(e) log(\mathcal{D}_1(\mathcal{I}(e))) \right] \right)
\end{aligned} \tag{10}$$

where u_n denotes the user embedding of the n^{th} instance. y_n denotes the label, which is set to 1 if user clicked item and 0 otherwise. λ is a tradeoff hyper-parameter that controls the learning procedure of modality-specific and modality-invariant representations.

We optimize the four objective functions, i.e, $\mathcal{L}_{CTR}$, $\mathcal{L}_{D_s}$, $\mathcal{L}_{D_0}$, and $\mathcal{L}_{ADV}$ to update different layers in MARN simultaneously. First, we minimize $\mathcal{L}_{CTR}$ to optimize the overall prediction performance. Second, we minimize $\mathcal{L}_{D_s}$ to learn the modality-specific projection layers $\mathcal{S}_m$, while the gradient of $\mathcal{D}_s$ will not be back-propagated for updating the embedding layer $\mathcal{E}$. Then, we minimize the first discriminator $\mathcal{L}_{D_0}$ to identify the potential common subspace across

multiple modalities. Since $\mathcal{D}_0$ are learned on un-weighted modality-invariant features and would not be a good indicator, the gradient of $\mathcal{D}_0$ will not be back-propagated for updating invariant projection layer $\mathcal{I}$. Finally, the second discriminator $\mathcal{D}_1$ plays the minimax game with the modality-invariant feature for updating $\mathcal{I}$. To solve the minimax game between $\mathcal{I}$ and $\mathcal{D}_1$, we adopt an end-to-end architecture through a gradient reversal layer (GRL) [8], so the sub-networks are trained jointly and boost each other.

4 EXPERIMENTS

In this section, we present our experiments in detail, including dataset, evaluation metric, experimental setup, model comparison, and the corresponding analysis. Experiments on a public dataset with user behaviors and a dataset collected from Taobao Search system demonstrate MARN achieves remarkable improvements to the state-of-the-art methods on the CTR task.

4.1 Datasets and Experimental Setup

Amazon Dataset: We use the Amazon dataset in Zhou et al. [39], which comprises several subsets of the amazon product data [16] and have sequential user behavior sequence. For a behavior sequence $\{x_1, x_2, \ldots, x_j, \ldots, x_T\}$, in the training phase, we predict the $j+1^{th}$ behavior with the first j behaviors, where $j = 1, 2, \ldots, T-2$. In the test phase, we use the first $T-1$ behaviors to predict the last one. The feature set we use contains item ID, category ID, image feature (extracted using pre-trained VGG16 model), statistic, together with 300-dimensional GloVe vectors [24] for the terms of the title. The statistics of Amazon dataset is shown in Table 1.

Taobao Dataset: We collect 8×10^9 samples from the daily exposure and click logs in Taobao Search system, which consist of the user behaviors and the labels (i.e., exposure or click) for the CTR task. The user behavior type consists of click, add-to-cart and purchase. The feature set of an item contains item ID, shop ID, brand ID, category ID, image feature (extracted using pre-trained VGG16 model), statistic and title, also with a behavior type and time. For the production scenario, we use the samples across 7 days for training and evaluate on the samples of the next day. The statistics of Taobao dataset is shown in Table 1.

We set the hyper-parameter configuration as follows. For Amazon and Taobao datasets, the dimensions of the item ID, shop ID, brand ID and category ID are set to be the same as in [20], which are 32, 24, 24 and 16, respectively. The vocabulary sizes of the embeddings are set according to the statistics of the datasets. Since we focus on multimodal item representation learning in this paper, we make the hidden units of the prediction layer as fixed values for all models, with [128,64] for the Amazon dataset and [512,256] for the Taobao dataset. The activation functions are ReLU. The model is trained on 32-sized batches for Amazon dataset and 1024-sized batches for Taobao dataset. We use AdaGrad [6] as the optimizer with the learning rate of 0.1. We gradually change λ from 0 to λ_0 following the schedule $\lambda = \lambda_0(\frac{2}{1+\exp(-\gamma \cdot p)} - 1)$, where γ is set to 10 in all experiments (the schedule was not optimized/tweaked), and p linearly changes from 0 to 1 in the training progress. The hyper-parameter λ_0 is tuned ranging from 0.01 to 1.8 and MARN achieves the best performance at $\lambda_0 = 0.05$.

4.2 Offline Comparison of Different Methods

To show the effectiveness of our method, we compare our method with three groups of nine baselines. The first group consists of models before deep networks for the CTR prediction task.

***LR* [17]:** Logistic regression (LR), which is a widely used shallow model before deep networks for CTR prediction task.

***FM* [25]:** Factorization Machine, which models both first-order feature importance and second-order feature interactions.

The second group contains concatenation-based mutlimodal fusion methods for mutlple modalities of items, which are the mainstream CTR prediction models.

***YoutubeNet* [5]:** A deep model proposed to recommend videos in YouTube, which gets user representations by simply averaging the item embeddings in the user behavior sequence.

***xDeepFM* [15]:** A compressed interaction network, which aims to automatically learn high-order feature interactions in both explicit and implicit fashions.

***DUPN* [20]:** A general user representation learning method for E-commerce scenarios, which adopts LSTM and attention mechanism to model the user behavior sequence.

***DIEN* [38]:** A two-layer RNN structure with attention mechanism. It uses the calculated attention values to control the second RNN layer to model drifting user interests.

The third group is formed of the state-of-the-art multimodal methods, which learn unified representations from multiple modalities and utilize variety of fusion techniques to improve the performance of the CTR prediction task.

***DMF* [12]:** A hierarchically multimodal joint network, which densely correlates the representations of different modalities layer-by-layer, where the shared layer not only models the correlation in the current level but also depends on the lower one.

***MMSS* [14]:** A modality-based attention mechanism model with image filters to selectively use visual information.

***NAML* [31]:** An attentive multi-view learning model, which incorporates titles, bodies and categories as different views of news, and applies attention mechanism to news encoder to select important views for learning informative news representations.

Note that, the inputs of LR and FM contains the IDs feature (item ID, shop ID, brand ID and category ID). All compared neural network models are fed with the same features as MARN for fair comparison. Finally, we conduct the significance test to verify the statistical significance of the performance improvement of our model against the baseline models.

4.3 Evaluation Metrics

Following [15, 20, 35, 38], we adopt AUC as the evaluation metric, which is widely used in CTR prediction tasks. Average AUC is evaluated as following:

$$AUC = \frac{1}{\left|U^{Test}\right|} \sum_{u \in U^{Test}} \frac{1}{\left|I_u^+\right| \left|I_u^-\right|} \sum_{i \in I_u^+} \sum_{j \in I_u^-} \delta(\hat{p_{u,i}} > \hat{p_{u,j}}) \quad (11)$$

where $\hat{p_{u,i}}$ is the predicted probability that a user $u \in U^{Test}$ may click on the item i in the test set and $\delta(\cdot)$ is the indicator function. I_u^+ is the item set clicked by the user, and I_u^- is the item set that the user does not click.

Dataset	# Users	# Items	# Shops	# Brands	# Categories	# Images	# Titles	# Samples
Electro. (Amazon)	192403	63001	-	-	801	63001	63001	2993570
Clothe. (Amazon)	39387	23033	-	-	484	23033	23033	477862
Taobao	0.25 billion	0.8 billion	17 million	1.0 million	20 thousand	0.8 billion	0.8 billion	64 billion

Table 1: Statistics of Amazon and Taobao datasets.

Group	Method	Electro. AUC	Clothe. AUC	Taobao AUC
1	*LR*	0.7272	0.7143	0.7011
	FM	0.7369	0.7224	0.7091
2	*YoutubeNet*	0.7675	0.7593	0.7241
	xDeepFM	0.7762	0.7691	0.7287
	DUPN	0.7904	0.7774	0.7348
	DIEN	0.7923	0.7793	0.7363
3	*DMF*	0.7924	0.7797	0.7371
	MMSS	0.7934	0.7806	0.7380
	NAML	0.7958	0.7825	0.7398
	MARN	**0.8034***	**0.7909***	**0.7486***

Table 2: AUC on Amazon and Taobao datasets. Here, * indicates statistical significance improvement compared to the best baseline (NAML) measured by t-test at p-value of 0.05.

4.4 Results on Amazon and Taobao Datasets

All experiments are repeated 5 times and averaged metrics are reported in Table 2. The influence of random initialization on AUC is less than 0.0002. From Table 2, we can tell that MARN improves the performance significantly against all the baselines and achieves state-of-the-art performance.

In the majority of the cases, non-neural network models in Group 1, i.e., LR and FM perform worse than neural network models, which demonstrates the power of deep learning.

Comparing with the concatenation-based methods of Group 2, multimodal item learning methods of Group 3 outperforms all of them, which reflects the benefits of well designed multimodal fusion methods for learning good representation of items.

By comparing MARN with the models in Group 3, i.e., DMF, MMSS and NAML, although all the baselines are proposed to deal with multimodal information, MARN has better performance on the multimodal items in E-commerce. Since DMF was proposed to hierarchically joint multimodal information, it has the same issue as the concatenation-based methods which may ignore the different contributions of multiple modalites. MMSS was proposed to selectively use visual information with image filters, which may not utilize all image information effectively. Though the attention mechanism of NAML improves the performance, it has not explored the complementarity and redundancy of modalities by considering modality-specific and modality-invariant features differently.

Finally, the improvement of MARN over the best baseline model NAML are 0.76%, 0.84% and 0.88% AUC scores. As reported in [39], compared with YoutubeNet, DIN improves AUC scores by 1.13% and the improvement of online CTR is 10.0%, which means a small improvement in offline AUC is likely to lead to a significant increase in online CTR. Considering the massive size of Taobao Search system, the lift of AUC scores by MARN may bring considerable CTR improvement and additional income each year.

4.5 Ablation Study

In this section, we design ablation experiments to study how each component in MARN contributes to the final performance.

BaseModel*:** A model with unimodal item features which uses GRU to integrate the behavior embeddings. We conduct ***BaseModel+IDs with IDs feature (item ID, shop ID, brand ID and category ID) and ***BaseModel+IMAGE*** with image feature.

***BaseModel+CONC*:** A concatenation model, where concatenation is performed in the fusion layer by element-wise summing up all the projected modality features (IDs, image, title and statistic).

***BaseModel+MAF*:** A sub-model of MARN with the multimodal attention fusion network (MAF).

***BaseModel+MAF+ADV*:** A sub-model of MARN with both MAF and the original adversarial transfer network (ADV)[30].

***BaseModel+MAF+DDMA*:** A sub-model of MARN with both MAF and the proposed double-discriminators multimodal adversarial network (DDMA).

***MARN*:** The entire proposed model with MAF, DDMA and Attentional Property GRU (APGRU).

As shown in Table 3, removing any component in MARN leads to a drop in performance. Further than that, we have the following observations. First, with the help of bringing in more modalities (i.e., image, title, and statistic), CONC outperforms the unimodal baselines by 0.35%, 0.63% and 0.65% in AUC. Second, MAF outperforms CONC about 0.32%, 0.35% and 0.39% in terms of AUC, which reflects that learning dynamic contributions of different modalities for multimodal item representations further leads to better performance. Third, both the original adversarial network and the proposed double-discriminators multimodal adversarial network can boost the CTR task performance, which reveals the benefits of identifying the potential common subspace across multiple modalities by multimodal adversarial learning. Moreover, the double-discriminators strategy (DDMA) outperforms ADV by 0.39%, 0.51% and 0.52% AUC scores. It demonstrates that emphasizing the identified modality-invariant features by the first discriminator and further confusing the second discriminator are both crucial for learning better common latent subspace across multiple modalities. APGRU further provides AUC gain of 0.2%, which results in the final performance.

4.6 Item Representations Validation

4.6.1 Generalization on Unpopular Items. To verify what kinds of items benefit more from MARN, we uniformly divide all items in the test set of Taobao dataset into 10 equal frequency popularity levels according to their frequencies of interactions. The higher thepopularity level, the more popular items. We then compute the

Method	Electro. AUC	Clothe. AUC	Taobao AUC
BaseModel+IDs	0.7897	0.7729	0.7295
BaseModel+IMAGE	0.7631	0.7564	0.6833
BaseModel+CONC	0.7929	0.7792	0.7360
BaseModel+MAF	0.7961	0.7827	0.7399
BaseModel+MAF+ADV	0.7974	0.7838	0.7411
BaseModel+MAF+DDMA	0.8013	0.7889	0.7463
MARN (*MAF+DDMA+APGRU*)	**0.8034**	**0.7909**	**0.7486**

Table 3: AUC on Amazon and Taobao datasets.

AUC at each level and show the results in Table 4. It shows that both methods perform very well at high popularity levels from 7 to 9, while the performances decrease at low popularity levels from 0 to 3. For those less popular items, the unimodal model may not have enough training instances to capture the co-occurrence of the items. Fortunately, with the help of representing items with multimodal information, such as image and title features, MARN achieves extremely remarkable improvements at low popularity levels from 0 to 3. It demonstrates that *BaseModel+IDs* cannot deal with unpopular items well, while *MARN* boosts the generalization of the item representations and achieves better results.

4.6.2 Transfer from Seen Items to Unseen Items. New items usually cause the cold-start problem, and many methods cannot deal with them. The item-based CF (Collaborative Filtering) [26] cannot perceive the items with no historical records, so the result are nearly random items (*RANDOM*). *BaseModel+IDs* can relieve the cold-start items using some generic IDs feature such as brand ID and shop ID. MARN can further address the cold-start problem by constructing an item embedding vector from not only generic IDs feature, but also image and title features. To study item transferability on unseen items, we evaluate the click recall@top-N of the generated candidate set in the next day. The results are shown in Table 5. The classical item-based CF cannot deal with new items. Compared to *BaseModel+IDs*, MARN achieves the improvement of about 2% when N is less than 70 and about 1% when N is beyond 70. Although the effectiveness decreases as N increases, this is rare case because the number of items browsed by users is usually limited.

4.7 Case Study of Multimodal Attention

To verify the effectiveness of the multimodal attention in our model, we visualize several categories in Taobao dataset in Figure 4. We randomly select ten thousand items for each category and each item is represented by multiple modalities including IDs, image, title and statistic. We then calculate the average $l2$-norm of the vector attention of multiple modalities for each category. The heat map matrix represents the attention intensity of each modality with respect to different categories.

We can find that the image feature is highly weighted under categories such as clothing, shoes, and jewelry, while the statistic feature gains more weight in the cell phone and grocery food categories than that of the image feature. As a result, the multimodal attention mechanism in MARN can effectively learn dynamic contributions of different modalities for multimodal item representations.

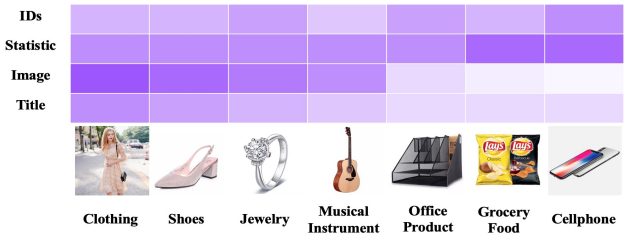


Figure 4: The heat map of attention weights with different item modalities of different categories.

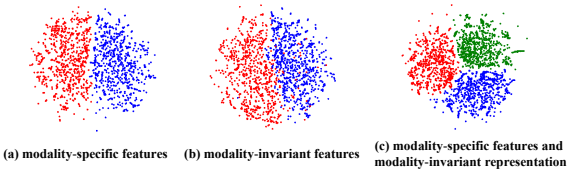
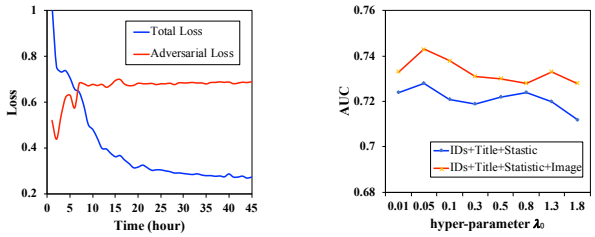


Figure 5: The t-SNE modality visualization. Red points denote the image modality feature, while blue points denote title modality feature. Green points denote the modality-invariant representation obtained by max-pooling.



(a) Convergence. (b) Parameter Sensitivity.

Figure 6: Convergence and parameter sensitivity study.

4.8 Empirical Analysis of Adversarial Network

4.8.1 Modality Visualization. To verify the effectiveness of the proposed multimodal adversarial network, we visualize the t-SNE features for ten thousand randomly selected items in Taobao dataset. The visualized features are modality-specific features (s_i^{image} and s_i^{title}), modality-invariant features (c_i^{image} and c_i^{title}) and modality-invariant representation (c_i). As shown in Figure5(a), s_i^{image} and s_i^{title} are almost completely separated, which reflects that the auxiliary modality discriminator can reduce the redundancy and increase the distinctiveness of the modality-specific features. As shown in Figure5(b), c_i^{image} and c_i^{title} are drawn close to each other, which reveals that the multimodal adversarial mechanism has the ability to decrease the modality gap and align distributions of different modalities. Figure5(c) illustrates that s_i^{image}, s_i^{title} and c_i are completely separated, which means MARN can effectively exploit the complementarity and redundancy of multiple modalities and achieve universal item representations by combining both modality-specific and modality-invariant representations.

Popularity-level	0	1	2	3	4	5	6	7	8	9
BaseModel+IDs	0.672	0.687	0.694	0.717	0.725	0.732	0.736	0.739	0.750	0.752
MARN	0.716	0.712	0.716	0.738	0.743	0.745	0.742	0.741	0.752	0.753
GAP	+4.4%	+2.5%	+2.2%	+2.1%	+1.8%	+1.3%	+0.6%	+0.2%	+0.2%	+0.1%

Table 4: AUC at different popularity levels on Taobao dataset.

Methods (top-N)	10	20	30	40	50	60	70	80	90	100	1000
RANDOM	0.01%	0.01%	0.02%	0.02%	0.03%	0.04%	0.05%	0.06%	0.07%	0.07%	0.59%
BaseModel+IDs	5.82%	6.45%	10.13%	11.21%	12.16%	12.81%	13.42%	14.24%	14.43%	14.54%	21.28%
MARN	6.95%	7.97%	12.31%	13.63%	14.86%	15.34%	15.59%	15.75%	15.86%	15.99%	21.99%

Table 5: The click recall @top-N on Taobao dataset.

4.8.2 Convergence and Parameter Sensitivity. As shown in Figure 6(a), along with the training procedure, the total loss of MARN on Taobao dataset converges to a stable state, while the adversarial loss gradually increases, which means the modality discriminator becomes incapable of discriminating between different modalities based on the learned modality-invariant features. Figure 6(b) shows the AUC of MARN on Taobao dataset with the hyper-parameter λ_0 ranging from 0.01 to 1.8. MARN achieves the best performance at $\lambda_0 = 0.05$, which is an appropriate balance.

4.9 Online Evaluation

4.9.1 Online Metrics. Careful online A/B testing in Taobao Search system was conducted for one month. The evaluation metrics are computed over all exposure items returned by the ranking system based on the CTR prediction models (i.e., YoutubeNet, DUPN and MARN). We report two metrics: CTR and GMV (Gross Merchandise Volume). CTR is computed as the ratio between the number of clicked items and the total number of exposure items. GMV is computed by the transaction number of items multiplied by the price of each item.

Method	CTR Improve	GMV Improve
YoutubeNet	0%	0%
DUPN	5.23%	3.17%
MARN	**10.46%**	**5.43%**

Table 6: Online A/B testing.

Due to the policy restrictions, we do not report the absolute numbers for the online metrics. Instead, we report the relative numbers with respect to *YoutubeNet*. As shown in Table 6, compared to *DUPN*, the last version of our online serving model, *MARN* has improved CTR by 5.23% and GMV by 2.26%. Considering the massive size of Taobao Search system, such consistent online improvements are significant. MARN has already been deployed online, serving the main traffic for hundreds of million users with billions of items.

4.9.2 Online Serving. Online serving of industrial deep networks is not an easy job with hundreds of millions of users visiting our system everyday. Even worse, at traffic peak our system serves more than 200,000 users per second. It is required to make real-time CTR predictions with high throughput and low latency. Though MARN is complex, it has additional advantage that the multimodal item representations can be pre-extracted during the stage of model deploying. Since MARN automatically learns the weights of different modalities only according to the item itself, we can perform the subnetwork of MARN to extract the multimodal representation of each item and store the representation vector in the index of the search engine, thereby reducing the latency of online serving. Moreover, several important techniques are deployed for accelerating online serving of industrial deep networks: i) quantization for recurrent neural networks, which adopts multi-bit quantization strategies to accelerate online inference; ii) heterogeneous calculations including GPUs and ALI-FPGAs, which can accelerate matrix computations. In conclusion, optimization of these techniques doubles the QPS (Query Per Second) capacity of a single machine practically. Online serving of MARN also benefits from this.

5 CONCLUSIONS

This paper proposes a novel multimodal representation learning method for multimodal items, which can improve CTR and further boost GMV in E-commerce. We explore the complementarity and redundancy of multiple modalities by considering modality-specific and modality-invariant features differently. To achieve discriminative representations, we propose a multimodal attention fusion network. Moreover, to achieve a common latent subspace across modalities, we propose a double-discriminators multimodal adversarial network. We perform extensive offline experiments on Amazon and Taobao datasets to verify the effectiveness of the proposed method. MARN consistently achieves remarkable improvements to the state-of-the-art methods. Moreover, the approach has been deployed in an operational E-commerce system and online A/B testing further demonstrates the effectiveness.

ACKNOWLEDGMENTS

We thank colleagues of our team - Zhirong Wang, Tao Zhuang, Heng Li, Ling Yu, Lei Yang and Bo Wang for valuable discussions and suggestions on this work. We thank our search engineering team for the large scale distributed machine learning platform of both training and serving. We also thank scholars of prior works on multimodal representation learning and recommender system. We finally thank the anonymous reviewers for their valuable comments.

REFERENCES

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *TPAMI* (2018).

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *TPAMI* 35, 8 (2013), 1798–1828.

[3] Fedor Borisyuk, Liang Zhang, and Krishnaram Kenthapadi. 2017. LiJAR: A system for job application redistribution towards efficient career marketplace. In *SIGKDD*. 1397–1406.

[4] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. 2017. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, Vol. 1. 7.

[5] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Recsys*. 191–198.

[6] John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12, Jul (2011), 2121–2159.

[7] Ali Mamdouh Elkahky, Yang Song, and Xiaodong He. 2015. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *WWW*. International World Wide Web Conferences Steering Committee, 278–288.

[8] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised Domain Adaptation by Backpropagation. In *ICML*. 1180–1189.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*. 2672–2680.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*. 1026–1034.

[11] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based recommendations with recurrent neural networks. In *ICLR*.

[12] Di Hu, Chengze Wang, Feiping Nie, and Xuelong Li. 2019. Dense Multimodal Fusion for Hierarchically Joint Representation. In *ICASSP*. IEEE, 3941–3945.

[13] Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*.

[14] Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, Chengqing Zong, et al. 2018. Multi-modal sentence summarization with modality attention and image filtering. In *IJCAI*.

[15] 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 *KDD*. ACM, 1754–1763.

[16] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *SIGIR*. 43–52.

[17] H Brendan McMahan, Gary Holt, David Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, et al. 2013. Ad click prediction: a view from the trenches. In *SIGKDD*. 1222–1230.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*. 3111–3119.

[19] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *ICML*. 689–696.

[20] Yabo Ni, Dan Ou, Shichen Liu, Xiang Li, Wenwu Ou, Anxiang Zeng, and Luo Si. 2018. Perceive Your Users in Depth: Learning Universal User Representations from Multiple E-commerce Tasks. In *SIGKDD*. 596–605.

[21] Kyo-Joong Oh, Won-Jo Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2014. Personalized news recommendation using classified keywords to capture user preference. In *ICACT*. 1283–1287.

[22] Wanli Ouyang, Xiao Chu, and Xiaogang Wang. 2014. Multi-source deep learning for human pose estimation. In *CVPR*. 2329–2336.

[23] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. 2018. Multi-adversarial domain adaptation. In *AAAI*.

[24] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *EMNLP*. 1532–1543.

[25] Steffen Rendle. 2010. Factorization machines. In *ICDM*. IEEE, 995–1000.

[26] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *WWW*. 285–295.

[27] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

[28] Yong Kiam Tan, Xinxing Xu, and Yong Liu. 2016. Improved recurrent neural networks for session-based recommendations. In *Recsys*. 17–22.

[29] Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen. 2013. Deep content-based music recommendation. In *NIPS*. 2643–2651.

[30] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial cross-modal retrieval. In *ACMMM*. 154–162.

[31] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Attentive Multi-View Learning. In *IJCAI*.

[32] Sai Wu, Weichao Ren, Chengchao Yu, Gang Chen, Dongxiang Zhang, and Jingbo Zhu. 2016. Personal recommendation using deep recurrent neural networks in NetEase. In *ICDE*. 1218–1229.

[33] Zuxuan Wu, Yu-Gang Jiang, Jun Wang, Jian Pu, and Xiangyang Xue. 2014. Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In *ACMMM*. 167–176.

[34] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. 2018. Learning semantic representations for unsupervised domain adaptation. In *ICML*. 5419–5428.

[35] Ling Yan, Wu-jun Li, Gui-Rong Xue, and Dingyi Han. 2014. Coupled group lasso for web-scale ctr prediction in display advertising. In *ICML*. 802–810.

[36] Chaohui Yu, Jindong Wang, Yiqiang Chen, and Meiyu Huang. 2019. Transfer Learning with Dynamic Adversarial Adaptation Network. In *ICDM*.

[37] Luo Yuanfei, Wang Mengshuo, Zhou Hao, Yao Quanming, Tu WeiWei, Chen Yuqiang, Yang Qiang, and Dai Wenyuan. 2019. AutoCross: Automatic Feature Crossing for Tabular Data in Real-World Applications. In *KDD*.

[38] 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 *AAAI*, Vol. 33. 5941–5948.

[39] 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 *SIGKDD*. 1059–1068.