Disentangled Multimodal Representation Learning for Recommendation

Fan Liu, *Member, IEEE*, Huilin Chen, Zhiyong Cheng, Anan Liu, Liqiang Nie, *Senior Member, IEEE* and Mohan Kankanhalli, *Fellow, IEEE*

Many multimodal recommender systems have been proposed to exploit the rich side information associated with users or items (e.g., user reviews and item images) for learning better user and item representations to improve the recommendation performance. Studies from psychology show that users have individual differences in the utilization of various modalities for organizing information. Therefore, for a certain factor of an item (such as *appearance* or *quality*), the features of different modalities are of varying importance to a user. However, existing methods ignore the fact that different modalities contribute differently towards a user's preference on various factors of an item. In light of this, in this paper, we propose a novel *Disentangled Multimodal Representation Learning* (DMRL) recommendation model, which can capture users' attention to different modalities on each factor in user preference modeling. In particular, we employ a disentangled representation technique to ensure the features of different factors in each modality are independent of each other. A multimodal attention mechanism is then designed to capture users' modality preference for each factor. Based on the estimated weights obtained by the attention mechanism, we make recommendations by combining the preference scores of a user's preferences to each factor of the target item over different modalities. Extensive evaluation on five real-world datasets demonstrate the superiority of our method compared with existing methods.

***Index Terms*—Multimodal, Modality Preference, Disentangled Representation, Recommendation**

I. Introduction

RECOMMENDER systems have been widely applied in modern web services, such as e-commerce platforms, news websites and social media sites. Recommendations not only assist users in quickly finding the desired content from a massive amount of information, but also help develop the providers (e.g., Amazon [1], eBay [2], TikTok [3]) services. Among various recommendation techniques, Collaborative Filtering (CF) based methods [1], [2] have been very successful, attributed to its simple idea of learning the user and item representations by exploiting the interaction data (e.g., *click* or *purchase*). However, a problem of only using the interaction data is that the performance will dramatically drop when the interactions between users and items are sparse. An effective way to overcome this problem is to leverage the side information associated with users or items, which provide rich information about user preference and item characteristics [3], [4], [5].

So far, there have been many approaches proposed to exploit various types of side information, such as attributes [4], [5], reviews [6], [7], [8], [9], [10] and images [3], [11]. Most existing methods often explore the capability of a single modality, like reviews [9], [10] or images [3], [11], on enhancing the recommendation performance. In the multimedia domain, it has been well-recognized that features from different modalities are complementary to each other and the use of multi-modal features can help improve the performance in general [12], [13], [14]. In recent years, researchers have also exploited the multimodal information for learning more comprehensive user and item representation in recommendation systems [15], [16], [17], [18], [19], [20], [21]. Based on how multimodal information is utilised in user and item embedding learning, we broadly classify the existing methods into two types. The first type integrates the features learned from different modalities via a linear [15], [19], [20] or non-linear method [16], [18] to obtain a joint representation for each user or item. For instance, JRL [16] concatenates the representations learned from ratings, reviews and images to form the final representation. The other type utilizes a regularization scheme to constrain the relations between representations of different modalities in a latent space [17], [18]. For example, CML [17] uses the multimodal features as a regularization term with a designed loss to facilitate the embedding learning of items and users.

It is well-known that users preferences on items are diverse. That is, the cause for a user to like an item are varied across different items. For example, a user may favor a product due to its *quality* and *brand*, while prefer another one just because of its *aesthetic appearance*. Because multimodal information provides rich features that directly describe different aspects and even evidences of users' opinion on those aspects (e.g., comments), it has also been exploited to model the diverse preferences of users towards various items. To achieve this, existing methods either learn an attention vector for different aspects with respect to the target user-item pair [16] or update the target item/user vector based on the multimodal features of the target items [18], [9], [10]. The success of the aforementioned methods is largely attributed to their effectiveness of exploiting multimodal information to learn

F. Liu and M. Kankanhalli are with School of Computing, National University of Singapore, Singapore. Email: liufancs@gmail.com, mohan@comp.nus.edu.sg

H. Chen is with the School of Computer Science and Engineering, Tianjin University of Technology, China. Email:ClownClumsy@outlook.com

Z. Cheng is with the Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences), China. Email: jason.zy.cheng@gmail.com

A. Liu is with the School of Electrical and Information Engineering, Tianjin University, China. Email:anan0422@gmail.com

L. Nie is a professor with Harbin Institute of Technology (Shenzhen), China. Email: nieliqiang@gmail.com

Fan Liu is the corresponding author.

[1]https://www.amazon.com

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

[3]https://www.tiktok.com

better user and item representations. We argue that they have not fully utilized the multimodal information. This is because they do not disentangle the multimodal features to mine the relative contributions of different modality features to users' preference for each factor of an item.

Firstly, the salient factors captured by different modalities are different. For example, user reviews contain more of users' opinions on factors such as *quality* and *comfort* [9], [10]. In contrast, item images capture more of user preferences concerning visual appearance (like *color* and *type*) [3], [22]. Existing multimodal CF models usually exploit the features of multimodal information without distinguishing the feature differences between different modalities [15], [16], [12]. For example, TranSearch [12] uses a deep neural network to fuse multimodal representations using a non-linear function. The previous studies have not considered the differing contributions of multimodal information at the factor-level for users' preference modeling.

In addition, users have individual differences in perceiving different modalities for organizing information as pointed out in psychology studies [23]. Accordingly, the features of different modalities are of different importance to a user's preference on a certain factor of a target item. In other words, users have different preferences for different modalities (or *modality preference*) in each factor of an item. For example, the *style* information can be directly viewed from the images or obtained from the descriptions in reviews when purchasing *clothes*. For this factor, some users directly view the images to check whether they like them, while some others might be more willing to refer to the opinions in reviews. However, existing multimodal recommendation methods have not considered the modality preferences of users in the utilization of information when exploiting different modalities, resulting in sub-optimal performance.

Finally, the entangled representations of different factors in each modality can hurt the performance [24], [25], [26]. For example, a user purchases a dress because of its *style* while caring less about *comfort*; otherwise, the user would have purchased the other dress considering both *style* and *comfort*. When the representations of different factors in each modality are entangled, it results in information redundancy and performance degradation in preference modeling. We would like to model user preference from different factors and each factor can be represented by the features of different modalities. Disentangled representation has been demonstrated to be more resilient to feature complexity [24], [25]. Recently, researchers have also applied disentangled representation learning techniques to model users' diverse intents [26]. Although some progress has been achieved, they only use the user-item interaction data and leave the rich information contained in multimodal data unused. In fact, the multimodal data provides a lot of useful information to help us discover users' intents on items. The problem is how to effectively disentangle and fuse the multimodal features to better serve our recommendation model.

Motivated by the above considerations, in this paper, we propose a *Disentangled Multimodal Representation Learning* (DMRL for short) model, which models user's preference by considering the different contributions of features from different modalities for each disentangled factor. In our model, a user's preference for an item is predicted by aggregating her weighted preferences on all modality features of different factors. To enable robust and independent representations for each factor, we employ a disentangled representation technique to ensure that the features of different factors in each modality are independent of each other. Based on the disentangled representations of different factors, we design a multimodal attention mechanism to capture users' modality preference on different modalities for each factor. Finally, for the preference prediction, given a user and a target item, we first estimate the user's preference score for each factor in each modality, and then linearly combine the scores of all the factors across different modalities using the estimated weights obtained by the attention mechanism. To this end, our model can profile users' personal preference based on disentangled factors represented by multimodalities with the consideration of their personal modality preference. Extensive experiments on five real-world datasets have been carried out to demonstrate the effectiveness of our method with comparison to several strong baselines, including the ones using multimodal information and disentangled learning techniques.

II. RELATED WORK

In this section, we briefly review the recent advances in model-based collaborative filtering methods, especially the ones based on multimodal information and disentangled representation learning, which are closely related to our work.

In CF models, users and items are represented as dense vectors (i.e., embeddings) in the same latent space. Based on the learned vectors, an interaction function is used to predict the preference of a user to an item. Take Matrix Factorization (MF) as an example, the user and item embeddings are learned by minimizing the error of re-constructing the user-item interaction matrix, and the dot product is used as the interaction function for prediction. This simple idea has achieved tremendous success, and many variants of MF have been developed [27], [28], [2], [29], [30]. The advent of deep learning has accelerated the development of model-based CF techniques. Due to the powerful capability of deep learning, it has been widely applied to learn better user and item embeddings [31], [32], [33] or model more complicated interactions between users and items [2], [34]. For example, NeuMF [2] models the complex interactions between users and items using nonlinear neural networks as the interaction function. More recently, Graph Convolution Network (GCN) techniques have also been deployed that have set a new standard for recommendation [35], [36], [37], [38]. The advantage of GCN-based recommendation models is attributed to its capability of explicitly modeling higher-order proximity between users and items. For example, NGCF [36] exploits higher-order proximity by propagating embeddings on the user-item interaction graph.

All of the aforementioned methods learn the representations of users and items merely relying on the interaction data. And their performance suffers when the interactions are sparse.

Besides the interaction data, the rich side information, such as reviews and images, provide valuable information of user preference and item characteristics, which has been widely used to alleviate the data sparsity problem in recommendation [7], [4], [8], [5].

A. Multimodal Collaborative Filtering

Various types of side information have been utilized in recommender systems, such as attributes [4], [5], reviews [7], [8], [9], [10] and images [3], [11]. Most existing models exploits different types of information individually, because of the challenges of fusing multimodal information. However, it is well-recognized that information from a different modality can provide complementary information, as demonstrated in [15], [16], [18], [20], [39], [40].

More recently, several deep models have been proposed to model user preference by using multimodal features. Zhang et al. [15] proposed a knowledge based method, which extracts the multimodal knowledge, unstructured textual and visual knowledge to jointly learn the latent representations of items within a collaborative filtering framework. In JRL [16], Zhang et al. first extracted user and item features from ratings, reviews, and images separately with deep neural networks, and then concatenated those features to form the final user and item representations for recommendation. For capturing fine-grained user preference, Liu et al. [18] proposed a metric learning based method, which models diverse user preferences by exploiting the item's multimodal feature. In recent years, Graph Convolution Networks (GCNs) have attracted increasing attention in multimedia recommendation due to the powerful capability on representation learning from non-Euclidean structures [19], [20], [41]. MMGCN [19] learns the model-specific user preference to the content information via direct information interchange between user and item in each modality. Later on, Wei et al. [20] proposed a structure-refined GCN model for multimedia recommendation.

Despite the progress, the factors behind multimodal features are entangled in representations, resulting in sub-optimal performance. In this paper, in order to generate robust multimodal representations, we employ the disentangled representation technique to encourage the independence of different features.

B. Disentangled Representation Learning

Disentangled representation learning has gained considerable attention, in particular in the field of image representation learning [24]. It aims to identify and disentangle the underlying explanatory factors behind the data. Such representations have demonstrated to be more resilient to the complex variants [25].

In recent few years, modeling user's diverse intent for liking items have attracted increasing attention. Cheng et al. [9] applied topic model on reviews to analyze user preferences on different aspects of items, which are used to model a user's diverse intents to different items. Liu et al. [18] presented a metric-learning based recommendation model, which employs an attentive neural network to estimate user attention on different aspects of the target item by exploiting the item's multimodal features (e.g., review and image).

In fact, the factors behind features in the aforementioned methods are highly entangled, which results in sub-optimal performance of recommendation. For learning robust and independent representations from user-item interaction data, disentangled representation are increasingly being valued in recommendation [42], [43], [26], [44]. Ma et al. [42] captured user preferences regarding the different concepts associated with user intentions separately. Wang et al. [43] proposed a disentangled heterogeneous graph attention network, which learns disentangled user/item representations from different aspects in a heterogeneous information network. For studying the diversity of user intents on adopting the items, Wang et al. [26] presented a GCN-based model, which yields disentangled representations by modeling a distribution over intents for each user-item interaction.

The above methods apply the disentangled learning techniques on interaction data to model users' diverse intents on adopting items without exploiting any side information. We claim that side information, especially the combination of multimodal information (e.g., reviews and images), contain rich evidence about users' preferences on items. In addition, we claim that for a specific factor, the features of different modalities convey different information. Inspired by this consideration, in this paper, we apply the disentangled learning technique on multimodal information to learn a users' preference on each factor of an item. Moreover, we design a multimodal attention network to capture a user's modality preferences on different modalities for each factor.

III. Our Model

A. Preliminaries

Before describing our model, we would like to introduce the problem setting first. Given a user set $\mathcal{U}$ and an item set $\mathcal{I}$, we use $\boldsymbol{R}^{N_u \times N_i}$ to denote the user-item interaction matrix, in which a nonzero entry $r_{u,i} \in \boldsymbol{R}$ indicates that user $u \in \mathcal{U}$ has interacted with item $i \in \mathcal{I}$ before; otherwise, the entry is zero. Notice that the interactions can be implicit (e.g., click) or explicit (e.g., rating). N_u and N_i are the numbers of users and items, respectively. In our setting, each user and each item is assigned a unique ID, and it is represented by a trait vector. Besides, each item is associated with multi-modal information, e.g., reviews and images. And for each item, its associated review and image are represented as a textual feature vector and a visual feature vector, respectively. It should be noted that item ID, image and review are regarded as three types of modality in our setting. Our goal is to recommend a user $u \in \mathcal{U}$ with suitable items which the user did not purchase before and will find appealing.

Existing multimodal methods often represent each modality (i.e., ID, review, and image) of each item as a holistic vector. As mentioned earlier, we would like to model users' diverse preference on different factors, and more specifically, their attention to different modalities of the same factor. To achieve this, in our model, the vector of each modality is composed of K chunks, with the assumption that each chunk represents

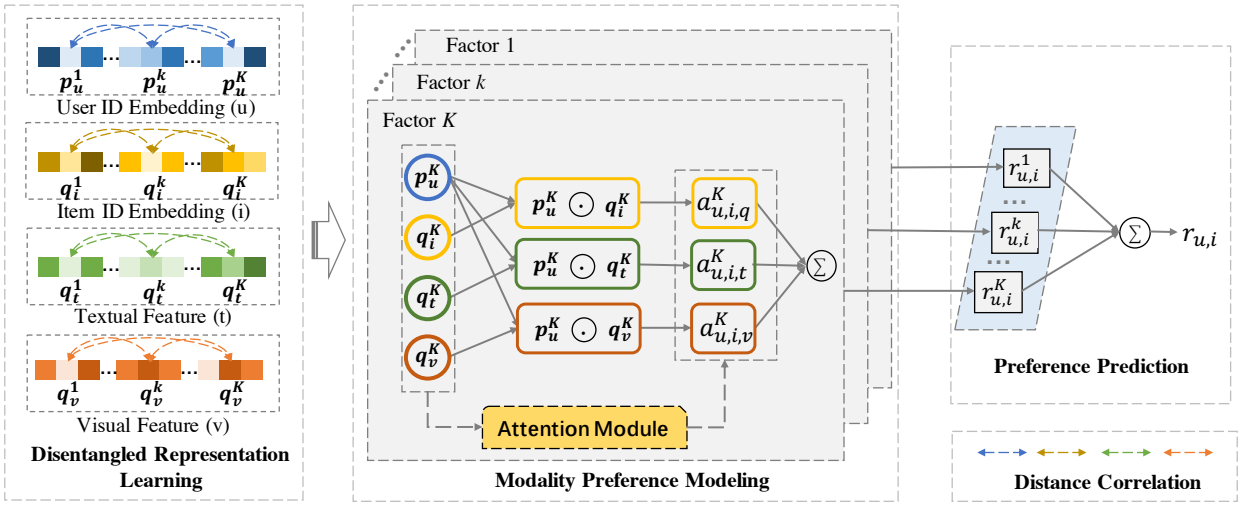


Fig. 1: Overview of our DMRL model. Best viewed in color.

a factor (such as appearance or quality). Without loss of generality, we use the user ID embedding as an example, it is initialized as:

$$\boldsymbol{p_u} = (\boldsymbol{p_u^1}, \boldsymbol{p_u^2}, \cdots, \boldsymbol{p_u^K}), \tag{1}$$

where $\boldsymbol{p_u} \in \mathbb{R}^d$ is ID embedding to capture intrinsic characteristics of user u; $\boldsymbol{p_u^k} \in \mathbb{R}^{\frac{d}{k}}$ is u's chunked representation of the k-th factor. Analogously, $\boldsymbol{q_i} = (\boldsymbol{q_i^1}, \boldsymbol{q_i^2}, \cdots, \boldsymbol{q_i^K})$, $\boldsymbol{q_t} = (\boldsymbol{q_t^1}, \boldsymbol{q_t^2}, \cdots, \boldsymbol{q_t^K})$ and $\boldsymbol{q_v} = (\boldsymbol{q_v^1}, \boldsymbol{q_v^2}, \cdots, \boldsymbol{q_v^K})$ are defined as the embeddings of item i's ID, textual feature and visual feature, respectively.

The embeddings of user ID and item ID are randomly initialized in our model, and the initial embeddings of textual and visual features of items are extracted from reviews and images, which are described in the following.

Feature Extraction. In our implementation, the BERT [45] model is utilized to extract the text feature. Specifically, BERT learns continuous distributed vector representations for textual documents in an unsupervised manner. BERT tries to preserve the semantic information and also considers the sequential information in the word sequence. It takes text documents as inputs and outputs their vector representations in a latent semantic space. The visual feature $\boldsymbol{F}_v$ is extracted by the ViT [46] model that uses the transformer encoder with 32 layers. In this work, we take the output of the transformer encoder, resulting in a 1024-D feature vector as the visual features for each item [4].

The extracted textual and visual features are then fed into two separate two-layer neural networks, which are used as refinement networks to learn valuable features oriented for recommendation [5]. Specifically, the two-layer neural network is:

$$\boldsymbol{q}_x = \sigma(\boldsymbol{W}_x^0 \sigma(\boldsymbol{W}_x^1 \boldsymbol{F}_x + \boldsymbol{b}_x^1) + \boldsymbol{b}_x^0), \tag{2}$$

where $\boldsymbol{F_x}$ is the input feature vector of the x-modality (i.e., visual or textual). $\boldsymbol{W}_x^0$, $\boldsymbol{W}_x^1$ and $\boldsymbol{b}_x^0$, $\boldsymbol{b}_x^1$ are the corresponding weight matrices and bias vectors for two layers, respectively. $\sigma(\cdot)$ is the activation function and LeakyReLU is employed because of its biologically plausibility and non-saturated property [36]. The textual feature $\boldsymbol{F_t}$ and visual feature $\boldsymbol{F_v}$ are processed by two separate networks to obtain the embeddings of $\boldsymbol{q_t}$ and $\boldsymbol{q_v}$ respectively for subsequent learning. Noted that our focus in this paper is to study whether the disentangled representation of factors can help us enhance the recommendation performance. The above textual and visual extraction methods can be replaced by any other more advanced methods. In the following, we give the details of our disentangled multimodal representation learning recommendation model.

B. Our Model

As shown in Figure 1, our *Disentangled Multimodal Representation Learning* (DMRL) consists of three key components: 1) **disentangled representation learning**, which employs distance correlation [47], [48] as a regularizer to encourage the independence of the features of different factors in each modality; 2) **modality preference modeling**, which captures user modality preference on different modalities for each factor; 3) **preference prediction**, which predicts the preference of a user to the target item by aggregating the learned disentangled features from different modalities with assigned attention weights. Next, we introduce each component in sequence.

1) Disentangled Representation Learning

We would like to divide the feature representations of each modality into k chunks. And each chunk represents a latent factor which impacts user's preference on the item [6]. For simplicity, the features from each modality are evenly divided into k continuous chunks. As a result, the features of different factors will be entangled. This will cause information redundancy and reduce the efficacy of preference modeling based on those factors. To avoid this problem, we need to first

[4]In this paper, each item is associated with one image.

[5]Experiments show that the neural network with two layers could achieve better performance.

[6]Note that we use the semantic factors, such as *appearance* and *quality*, as examples for ease of understanding of our model. Latent factors are implicit and unexplainable.

disentangle the features of different factors in each modality. To achieve this goal, we apply distance correlation [48], [47] to characterize independence of any two-paired embeddings of different factors. Its coefficient is zero if and only if these vectors are independent. For a feature vector y, we formulate it as:

$$L_y = \sum_{k=1}^{K} \sum_{k'=k+1}^{K} dCor\left(\boldsymbol{y}^k, \boldsymbol{y}^{k'}\right), \tag{3}$$

where $dCor(\cdot)$ is the function of distance correlation and it is defined as:

$$dCor\left(\boldsymbol{y}^k, \boldsymbol{y}^{k'}\right) = \frac{dCov\left(\boldsymbol{y}^k, \boldsymbol{y}^{k'}\right)}{\sqrt{dVar\left(\boldsymbol{y}^k\right) dVar\left(\boldsymbol{y}^{k'}\right)}}, \tag{4}$$

where $\boldsymbol{y}^k$ and $\boldsymbol{y}^{k'}$ are respectively the k and k'-th chunks of the vector y; $dCov(\cdot)$ represents the distance covariance between two matrices; $dVar(\cdot)$ is the distance variance of each matrix. In this way, we can define the regularization losses L_p, L_q, L_t, L_v to encourage the feature independence of the factors in user ID, item ID, textual feature and visual feature, respectively.

Finally, the regularization loss for disentangled learning is formulated as:

$$L_d = L_p + L_q + L_v + L_t. \tag{5}$$

Note that in this equation, the weights of different regularization losses (L_p, L_q, L_v, L_t) can be empirically tuned. Here, we set them to be equal in the learning process for simplicity.

*2) **Modality Preference Modeling***

Typically, for a certain factor, users often value its features exhibited by different modalities differently. For example, for the *appearance* of a product, a user will take the visual features (from images) more seriously than the textual features (from reviews); and value the textual features more than the visual features for the *quality* of the product. To capture the complicated user modality preferences on various modalities for each factor, we design a weight-shared multimodal attention mechanism in DMRL. Specifically, for an item i, a user u's specific attention $a^k_{u,i}$ to different modalities for the k-th factor is computed as:

$$\boldsymbol{h}^{\boldsymbol{k}}_{\boldsymbol{u,i}} = Tanh(\boldsymbol{W}[\boldsymbol{p}^k_u; \boldsymbol{q}^k_i; \boldsymbol{q}^k_t; \boldsymbol{q}^k_v] + \boldsymbol{b}), \tag{6}$$

$$\hat{\boldsymbol{a}}^{\boldsymbol{k}}_{\boldsymbol{u,i}} = \mathbf{W}_v \boldsymbol{h}^{\boldsymbol{k}}_{\boldsymbol{u,i}}, \tag{7}$$

where $\boldsymbol{W}$ and $\boldsymbol{b}$ are respectively the weight matrix and bias vector of the neural network. In order to enhance efficiency of learning, we use shared weight under all factors. $\boldsymbol{h}^{\boldsymbol{k}}_{\boldsymbol{u,i}}$ is an output of of the hidden layer; $\mathbf{W}_v$ is a transformation matrix that projects the hidden layer into an output attention weight vector. $[\boldsymbol{p}^k_u; \boldsymbol{q}^k_i; \boldsymbol{q}^k_t; \boldsymbol{q}^k_v]$ denotes the concatenation of $\boldsymbol{p}^k_u$, $\boldsymbol{q}^k_i$, $\boldsymbol{q}^k_t$ and $\boldsymbol{q}^k_v$. *Tanh* is used as the activation function.

Following the standard procedure of attention mechanism, there is a subsequent step to normalize $\hat{\boldsymbol{a}}^{\boldsymbol{k}}_{\boldsymbol{u,i}}$ with the softmax function, which converts the attention weights to a probability distribution. In our model, the final attention weight for the textual feature of the k-th factor is computed as:

$$a^k_{u,i,t} = \frac{exp\left(\hat{a}^k_{u,i,t}\right)}{exp\left(\hat{a}^k_{u,i,q}\right) + exp\left(\hat{a}^k_{u,i,t}\right) + exp\left(\hat{a}^k_{u,i,v}\right)}. \tag{8}$$

Similarly, the weight for item i's ID and visual feature $a^k_{u,i,q}$ and $a^k_{u,i,v}$ can be computed in the same way.

*3) **Preference Prediction***

User's preference for an item should be estimated based on her preferences on different modalities of all factors for this item. Hence, with the representations of different modalities for different factors, given a user u and a target item i with their representations, the user preference on the k-th factor according to the features of the x-modality is estimated as:

$$r^k_{u,i,x} = a^k_{u,i,x} \cdot \sigma(\boldsymbol{p}^{\boldsymbol{k}}_{\boldsymbol{u}} \odot \boldsymbol{q}^{\boldsymbol{k}}_{\boldsymbol{x}}). \tag{9}$$

where $\odot$ denotes dot product. Notice that the attention weight $a^k_{u,i,x}$ represents the importance of the x-modality to the k-th factor for the user's preference to the item; and the dot product between $\boldsymbol{p}^{\boldsymbol{k}}_{\boldsymbol{u}}$ and $\boldsymbol{q}^{\boldsymbol{k}}_{\boldsymbol{x}}$ estimates how the features of the x-modality fits this user's tastes on the k-th factor. The integration of both terms can evaluate the user's preference to the item on the k-th factor according to the features of x-modality more comprehensively. $\sigma(\cdot)$ is the activation function and softplus is used to ensure that the resultant score is positive. Similarly, we can compute $r^k_{u,i,q}$, $r^k_{u,i,t}$ and $r^k_{u,i,v}$ in the same way, which represent the user preferences on the k-th factor from different modalities, i.e., item ID, reviews, and images, respectively. And the final score of user preference on the k-th factor of the target item is predicted by aggregating the scores from different modalities:

$$r^k_{u,i} = r^k_{u,i,q} + r^k_{u,i,t} + r^k_{u,i,v}. \tag{10}$$

Finally, the predicted scores of all the factors are integrated together to predict user preference to the target item, which is:

$$r_{u,i} = \sum_{k=1}^{K} r^k_{u,i}. \tag{11}$$

C. Model Learning

1) Objective function

We target at the top-n recommendation, namely, to recommend a set of n top-ranked items which match the target user's preferences. Similar to other rank-oriented recommendation studies [16], [36], we use the pairwise-based learning method for optimization. The loss function is defined as:

$$L_{BPR} = \sum_{(\mathbf{u},\mathbf{i}^+,\mathbf{i}^-)\in\mathcal{O}} -\ln \phi(r_{u,i} - r_{u,i^-}), \tag{12}$$

where $\mathcal{O} = \{(u, i^+, i^-)|(u, i^+) \in \mathcal{R}^+, (u, i^-) \in \mathcal{R}^-\}$ denotes the training set; $\mathcal{R}^+$ indicates the observed interactions between user u and i^+ in the training dataset, and $\mathcal{R}^-$ is the sampled unobserved interaction set. For each positive pair (u, i^+), we first randomly sample n negative ones from the items that a user hasn't purchased before as candidate

TABLE I: Basic statistics of the used datasets.

Dataset	#user	#item	#interactions	sparsity
Office Products	4,874	7,279	52,957	99.85%
Baby	12,637	18,646	121,651	99.95%
Clothing	18,209	35,526	150,889	99.98%
Toys Games	18,748	30,420	161,653	99.97%
Sports	21,400	36,224	982,618	99.97%

items [7]. From the candidate items, we select the one which is most similar to the user u based on the dot product of their embeddings, as a hard negative sample to construct the negative pair (u, i^-).

2) Optimization

With the consideration of all the regularization terms, the final object function of our DMRL is,

$$loss = L_{BPR} + \lambda_\theta L_\theta + \lambda_d L_d, \tag{13}$$

where $L_\theta = \|\Theta\|_2^2$ represents the L_2 regularization for the parameters Θ of the model; λ_θ and λ_d are hyperparameters that control the weights of L_2 regularizer and independence regularizer, respectively. The optimization is quite standard and the stochastic gradient descent (SGD) algorithm is adopted. In implementation, the Adam optimizer [49] is adopted to tune the learning rate.

IV. Experiments

To validate the effectiveness of our model, we conducted extensive experiments on five public datasets. In the next, we first introduce the experimental setup, and then report and analyze the experimental results.

A. Experimental Setup

1) Datasets

The public Amazon review dataset [8] [7], which has been widely used for recommendation evaluation in previous studies, is used for evaluation in our experiments. This dataset contains user interactions (review, rating, helpfulness votes, etc.) on items as well as the item metadata (descriptions, price, brand, image features, etc.) on 24 product categories. Five product categories in this dataset are selected and all the reviews and item images are kept as side information for items. We pre-processed the dataset to keep only the items and users with at least 5 interactions. The basic statistics of the five datasets are shown in Table I. For each observed user-item interaction, we treated it as a positive instance, and then paired it with negative items which are randomly sampled from items that the user has not purchased before.

In this work, we focus on the top-n recommendation task, which aims to recommend a set of top-n ranked items that will be appealing to the target user. For each dataset, we randomly selected 80% of the interactions from each user to construct the training set, and the remaining 20% for testing. From the training set, we randomly selected 10% of interactions as a validation set to tune hyper-parameters.

[7]In our implementation, we empirically set $n = 4$ to balance efficiency and effectiveness.

[8]http://jmcauley.ucsd.edu/data/amazon.

2) Baselines and Evaluation Metrics

We compare our DMRL model with the following baselines, including both deep (NeuMF [2], JRL [16]) MF and metric learning (CML [17], MAML [18]) based models, as well as the graph based methods (NGCF [36], DGCF [26], MMGCN [19], GRCN [20]).

For each user in the test set, we treat all the items that the user did not interact with as negative items. Two widely used metrics for top-n recommendation are adopted in our evaluation: *Recall* and *Normalized Discounted Cumulative Gain* (NDCG) [50]. For each metric, the performance is computed based on the top 20 results. Notice that the reported results are the average values across all the testing users.

In experiments, for all the baselines, we use the codes they released for evaluation. And, we put great efforts to tune hyperparameters of these methods and reported their best performance.

3) Parameter Settings

We implemented our model with Tensorflow [9] and carefully tuned the key parameters. The embedding size is fixed to 128 for all methods and the embedding parameters are initialized with the Xavier method [51]. We optimized our method with Adam [49] and used the default learning rate of 0.0001 and default mini-batch size of 1024. The L_2 regularization coefficient λ_θ and the distance correlation λ_d are searched in the range of $\{1e^{-5}, 1e^{-4}, \cdots, 1e^{+1}\}$. We search the number of the factors being disentangled in $\{1, 2, 4, 8\}$. Besides, model parameters are saved in every 10 epochs. The early stopping strategy [36] is performed, $i.e.$, premature stopping if recall@20 does not increase for 50 successive epochs. In addition, our codes are released to facilitate the replication of our experiments [10].

B. Performance comparison

The results of our model and all the competitors over the five datasets are reported in Table II. Overall, it can be seen that our method outperforms all the competitors consistently across all the datasets in terms of different metrics. By grouping all the methods into two categories based on whether they use multimodal information, we make some interesting observations.

The methods in the first block only use the user-item interactions. NeuMF models the non-linear interactions between users and items by using deep neural networks and achieves better performance than the traditional MF methods using linear interactions [2]. NGCF exploits high-order connectivities between users and items through embedding propagation over the graph structure. Hence it obtains better user and item representations than NeuMF. DGCF outperforms NGCF across all the datasets. This is attributed to the utilization of disentangled representation to learn robust and independent user and item embeddings by considering users' diverse intents. However, the aforementioned methods all use the dot product as the interaction function. As pointed out in [17], dot product does not obey the triangle inequality and thus cannot model fine-grained user preferences well. By using a metric-based

[9]https://www.tensorflow.org.

[10]https://github.com/liufancs/DMRL.

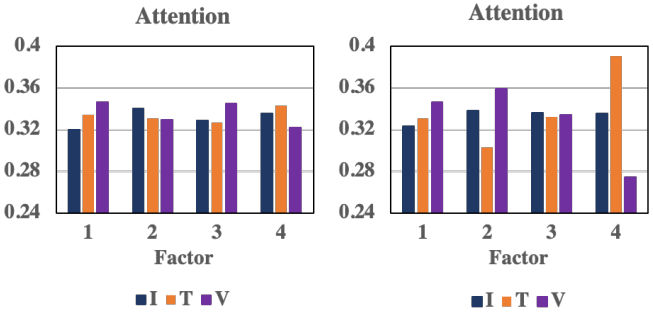
TABLE II: Performance of our DMRL model and the competitors over five datasets. The best results are highlighted in bold.

Datasets	Office Products		Baby		Clothing		Toys Games		Sports	
Metrics	Recall	NDCG	Recall	NDCG	Recall	NDCG	Recall	NDCG	Recall	NDCG
NeuMF	0.0605	0.0337	0.0502	0.0224	0.0189	0.0080	0.0253	0.0128	0.0330	0.0157
CML	0.1229	0.1395	0.0674	0.0444	0.0409	0.0224	0.1227	0.1160	0.0982	0.0640
NGCF	0.0814	0.0479	0.0694	0.0313	0.0466	0.0216	0.0970	0.0587	0.0707	0.0337
DGCF	0.1206	0.1105	0.0788	0.0465	0.0745	0.0405	0.1262	0.1085	0.1026	0.0629
JRL	0.0656	0.0365	0.0579	0.0266	0.0233	0.0124	0.0472	0.0413	0.0368	0.0214
MMGCN	0.1173	0.1231	0.0814	0.0496	0.0573	0.0312	0.1171	0.1056	0.0913	0.0572
MAML	0.1333	0.1412	0.0867	0.0521	0.0782	0.0387	0.1183	0.1117	0.1029	0.0676
GRCN	0.1351	0.1524	0.0883	0.0541	0.0861	0.0452	0.1336	0.1236	0.1065	0.0693
DMRL	**0.1563**	**0.1842**	**0.0906**	**0.0561**	**0.0917**	**0.0511**	**0.1434**	**0.1331**	**0.1111**	**0.0711**

learning approach, CML outperforms both NeuMF and NGCF by a large margin. This observation is also consistent with the results reported in [18].

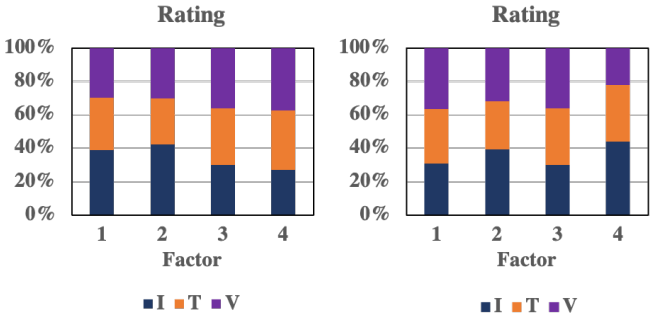
All the methods in the second block exploit both textual and visual features besides the interaction information. In general, these methods yield better performance than those without using multimodal features, demonstrating the effectiveness of leveraging side information on modeling user preference. Owing to the valuable information in multimodal features, JRL outperforms NeuMF accross all the datasets. By exploiting user-item interactions to guide the representation learning in different modalities, MMGCN yields better performance over NGCF on all the datasets. However, MMGCN underperforms DGCF on four datasets, which is because the entangled multimodal representations limit the capability of representation learning. MAML outperforms MMGCN across all the datasets, owing to the adoption of metric-based learning approach and attention mechanism to capture user diverse preferences. By discovering and pruning potential false-positive edges, GRCN obtains better performance across all datasets.

DMRL outperforms all the baselines consistently over all the datasets. We credit this to the joint effects of the following four aspects. Firstly, the utilization of multimodal information on modeling user preference, which can be observed from the performance comparisons of methods between two blocks in Table III. Secondly, DMRL models user preference by considering the differing contributions of different modalities to each factor. Thirdly, an attention network is designed in our model to capture user's modality preference on different modalities for each factor. Finally, the use of the disentangled representation learning technique learns independent representations for different factors. As demonstrated in DGCF, disentangled representation can better model users' multiple intents and thus achieve better performance.



(a) User A (b) User B

Fig. 2: Attention weights of different modalities. I, T and V represent item ID, textual feature and visual feature, respectively.



(a) User A (b) User B

Fig. 3: Ratings from two users. I, T and V represent item ID, textual feature and visual feature, respectively.

C. *Effects of Modality Preference*

One of our assumptions in the model is that for each factor of an item, different users may value its features of one modality more than those of another modality. To capture this, we design an attention mechanism to compute the attention weights for each factor of different modalities. The other assumption is that different modalities contribute differently for a user's preference to an item. With this consideration, for each user-item pair, our model computes the rating given to each factor of an item across different modalities (see Eq. 9). In this section, we would like to validate the above two assumptions via visualization: 1) visualizing the learned attention weights to illustrate user's modality preferences on different modalities for each factor; and 2) visualizing the user preferences (ratings) to illustrate two users' preferences on different factors represented by different modalities.

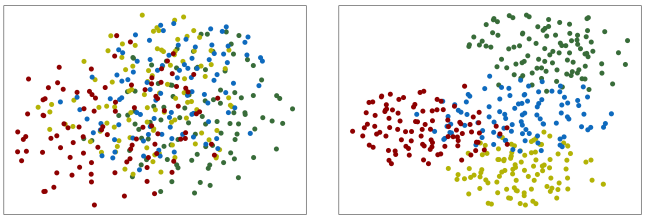
1) Modality Preferences (Attention Weights)

In order to study the user's modality preferences on different modalities for each factor, we compute the attention weights for different modalities of each factor (in the *Office* dataset) [11],which are visualized in Fig. 2. Specifically, I, T and

[11]We selected two users (0 and 197) who purchased the same item (1294) as an example.

TABLE III: Performance of our DMRL model and its variants over five datasets. The best results are highlighted in bold.

Datasets	Office		Baby		Clothing		ToysGames		Sports	
Metrics	Recall	NDCG	Recall	NDCG	Recall	NDCG	Recall	NDCG	Recall	NDCG
DGCF	0.1206	0.1105	0.0788	0.0465	0.0745	0.0405	0.1262	0.1085	0.1026	0.0629
DMRL$_t$	0.1517	0.1578	0.0783	0.0497	0.0766	0.0419	0.1258	0.1257	0.1071	0.0657
DMRL$_v$	0.1532	0.1612	0.0852	0.0534	0.0897	0.0464	0.1395	0.1287	**0.1122**	0.0682
DMRL$_{w/o\ a}$	0.1513	0.1454	0.0759	0.0479	0.0781	0.0392	0.1363	0.1262	0.0998	0.0597
DMRL$_{w/o\ u}$	0.1541	0.1559	0.0835	0.0513	0.0841	0.0427	0.1211	0.1186	0.1112	0.0691
DMRL	**0.1563**	**0.1842**	**0.0906**	**0.0561**	**0.0917**	**0.0511**	**0.1434**	**0.1331**	0.1111	**0.0711**



(a) Entangled Vector (b) Disentangled Vector

Fig. 4: Visualization of entangled and disentangled vectors extracted from textual features. Best viewed in color.

V represent item ID, textual feature and visual feature, respectively. The x-axis represents different factors. Comparing the results in Fig. 2(a) and Fig. 2(b), we can see that the weights of different modalities for the same factor are different between the two users. This verifies that for an item, the preferences of users to different modalities are different for each factor. In addition, it should be noted that the attention weight cannot represent the contribution of a modality to a user's preference to a factor of the target item.

*2) **User Preferences (Ratings)***

In order to further study the contributions of different modalities to a user's preference on the factors of the target item, we compute the ratings of the two different users given to different factors of the same item represented by different modalities. The ratings are then normalized across different modalities and visualized in Fig. 3. Overall, both Fig. 3(a) and Fig. 3(b) show that for the same factor, the features of different modalities have different contributions to user preference. For example, item ID yields a higher score for the second factor than the scores of both visual feature and textual feature. In other words, the textual and visual features contribute relatively less to the user's preference on this factor. What's more, take the results in the 4-th factor in Fig. 3(b) and Fig. 2(b) for example, we can find that although the user pays more attention to the textual feature according to the attention weight, the item ID actually contributes more to the user preference.

D. Effect of Disentanglement

In Fig. 4, we visualized the entangled and disentangled vectors extracted from textual features. The textual features is associated with the items from *Office* [12]. In both Fig. 4a and Fig. 4b, the dots with the same color (e.g., red) denote the vectors of the same factor. We use t-SNE [52] for clustering and visualization. The clustering results based on the vectors of different factors (right figure) can illustrate the effectiveness of using disentangled representation technique. In other words, it demonstrates that our proposed method could encourage independence of the vectors of different factors for modality features.

[12]We randomly sampled the items and their associated textual features.

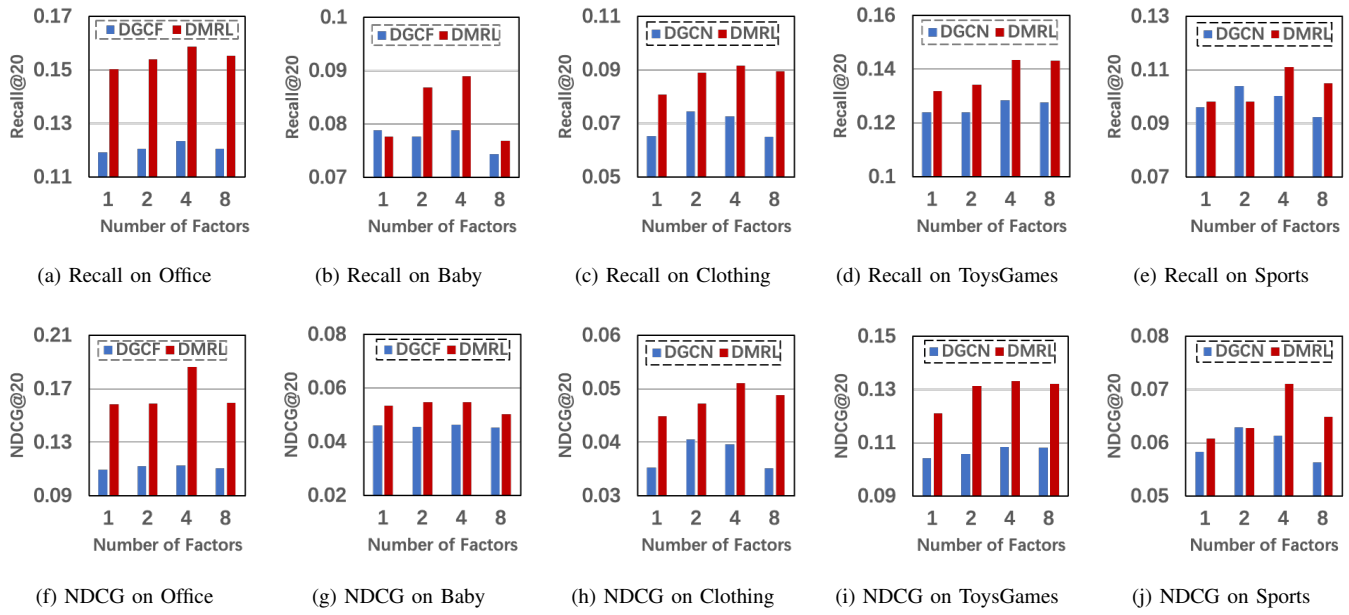
E. Ablation Study

In this section, we examine the contribution of different components to the performance of our model. Our analyses are based on the performance comparisons to the following methods.

- **DGCF [26]**: It is a disentangled representation based method which only uses item IDs.
- **DMRL$_t$**: It is a variant of our method which only uses item IDs and textual features.
- **DMRL$_v$**: It is a variant of our method which only uses item IDs and visual features.
- **DMRL$_{w/o\ a}$**: This variant removes the designed multimodal attention mechanism. It exploits the multimodal features of different factors indiscriminately.
- **DMRL$_{w/o\ u}$**: This variant removes user information in Eq. 6, namely, for the same item, the attention weights are all the same for different users. It is designed to investigate the effectiveness of modeling user modality preference on different modalities by comparing to our DMRL model.

The results of those variants and our method are reported in Table III. We group all the methods and variants into two groups based on whether the methods use all the modalities. The methods in the first block only use a part of three modalities. Both DMRL$_t$ and DMRL$_v$ outperform DGCF, demonstrating the effectiveness of leveraging side information to learn user preference. DMRL$_v$ outperforms DMRL$_t$ in terms of Recall@20 over all the datasets, which indicates the importance of visual features on user preference modeling. DMRL$_t$ performs better than DMRL$_v$ in terms of NDCG@20 on *ToysGames* and *Sports*. This is because textual features have a higher contribution for the two cases.

The variants in the second block use features of all three modalities. DMRL$_{w/o\ u}$ outperforms DMRL$_{w/o\ a}$ in almost all the cases, which indicates the importance of distinguishing different contributions of different modalities. More importantly, without the attention mechanism, DMRL$_{w/o\ a}$ even underperforms DMRL$_v$ and DMRL$_t$, which use our attention method. This further demonstrates the importance of modeling users' modality preferences. The further improvement of DMRL over DMRL$_{w/o\ u}$ validate the effectiveness of



(a) Recall on Office (b) Recall on Baby (c) Recall on Clothing (d) Recall on ToysGames (e) Recall on Sports

(f) NDCG on Office (g) NDCG on Baby (h) NDCG on Clothing (i) NDCG on ToysGames (j) NDCG on Sports

Fig. 5: Impact of the Factor Number (K).

distinguishing different users' attentions to various modalities for each factor.

It is surprising that the methods in the second block underperform the methods in the first block on $Sports$ in terms of Recall@20. This warns us that it is possible that the features of different modalities are inconsistent, which may exert negative influence on preference modeling. More advanced multimodal feature fusion methods are needed to tackle this problem. Besides this special case, DMRL can achieve superior performance over DMRL$_v$ and DMRL$_t$ consistently, which indicates the benefits of considering multimodal information in recommendation. The comprehensive ablation studies also demonstrate the positive utility of different components of our DMRL model.

F. Effects of the Factor Number

In this section, we study the influence of the factor number on the performance of our method. To compare the performance of our model and the baseline model with different factor numbers, we carry out experiments by tuning the number of factors in $\{1, 2, 4, 8\}$ on different datasets.

We compare our model to DGCF which also uses disentangled representations but did not use multimodal features. Fig. 5 shows the results in terms of Recall@20 and NDCG@20 on all five datasets. As we can see, with the increasing of the number of factors, both DMRL and DGCF achieve better performance and the best performance is obtained when the number is 2 or 4. This indicates that recommendation methods can benefit from robust and independent representation learning technique. Specifically, DGCF performs worse over *Office*, *Baby* and *Clothing* when $K = 1$ and $K = 2$, indicating that insufficient number of factors limits the capability of disentangled representation. However, the recommendation performance drops when when $K = 8$. The reason might be that the chunked representations with small embedding size have limited expressiveness in representation learning for each factor [26]. The large improvement over DGCF shows the effectiveness of our model on exploiting multimodal features.

V. Conclusion

In this work, we present a novel recommendation model called Disentangled Multimodal Representation Learning (DMRL), which models user's modality preference with respect to multimodal information on different factors of items. In DMRL, we employ disentangled representation learning technique to disentangle the representations of different factors in each modality. In addition, we design a multimodal attention mechanism to capture users' modality preferences to different modalities for each factor, so as to learn better representations for recommendation. Finally, we estimate a user's preference score to each factor of the target item based on its representation in each modality, and then combine the scores together across all modalities and factors with the computed attention weights. Extensive experiments on five publicly available datasets show that our model outperforms the state-of-the-art methods, demonstrating the superiority of our method in exploiting multimodal information. The ablation studies further validate the importance of modeling personal modality preference towards different modalities.

References

[1] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *IEEE Computer*, pp. 42–49, 2009.

[2] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proceedings of the 26th International Conference on World Wide Web*. IW3C2, 2017, pp. 173–182.

[3] J. McAuley, C. Targett, Q. Shi, and A. Van Den Hengel, "Image-based recommendations on styles and substitutes," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2015, pp. 43–52.

[4] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir *et al.*, "Wide & deep learning for recommender systems," in *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*. ACM, 2016, pp. 7–10.

[5] J. Chen, F. Zhuang, X. Hong, X. Ao, X. Xie, and Q. He, "Attention-driven factor model for explainable personalized recommendation," in *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2018, pp. 909–912.

[6] R. Catherine and W. Cohen, "Transnets: Learning to transform for recommendation," in *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 2017, p. 288–296.

[7] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: understanding rating dimensions with review text," in *Proceedings of the 7th ACM Conference on Recommender Systems*. ACM, 2013, pp. 165–172.

[8] Y. Tan, M. Zhang, Y. Liu, and S. Ma, "Rating-boosted latent topics: Understanding users and items with ratings and reviews," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, 2016.

[9] Z. Cheng, Y. Ding, L. Zhu, and M. Kankanhalli, "Aspect-aware latent factor model: Rating prediction with ratings and reviews," in *Proceedings of the 27th International Conference on World Wide Web*. IW3C2, 2018, pp. 639–648.

[10] J. Y. Chin, K. Zhao, S. Joty, and G. Cong, "ANR: Aspect-based neural recommender," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 2018, pp. 147–156.

[11] S. Wang, Y. Wang, J. Tang, K. Shu, S. Ranganath, and H. Liu, "What your images reveal: Exploiting visual contents for point-of-interest recommendation," in *Proceedings of the 26th International Conference on World Wide Web*. IW3C2, 2017, pp. 391–400.

[12] Y. Guo, Z. Cheng, L. Nie, X.-S. Xu, and M. Kankanhalli, "Multi-modal preference modeling for product search," in *Proceedings of the 26th ACM International Conference on Multimedia*. ACM, 2018, pp. 1865–1873.

[13] H. Fan, L. Zhu, Y. Yang, and F. Wu, "Recurrent attention network with reinforced generator for visual dialog," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 16, no. 3, 2020.

[14] H. Fan and Y. Yang, "Person tube retrieval via language description," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 10 754–10 761, 2020.

[15] F. Zhang, N. J. Yuan, D. Lian, X. Xie, and W.-Y. Ma, "Collaborative knowledge base embedding for recommender systems," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 353–362.

[16] Y. Zhang, Q. Ai, X. Chen, and W. B. Croft, "Joint representation learning for top-n recommendation with heterogeneous information sources," in *Proceedings of the 26th ACM on Conference on Information and Knowledge Management*. ACM, 2017, pp. 1449–1458.

[17] C.-K. Hsieh, L. Yang, Y. Cui, T.-Y. Lin, S. Belongie, and D. Estrin, "Collaborative metric learning," in *Proceedings of the 26th International Conference on World Wide Web*. IW3C2, 2017, pp. 193–201.

[18] F. Liu, Z. Cheng, C. Sun, Y. Wang, L. Nie, and M. Kankanhalli, "User diverse preference modeling by multimodal attentive metric learning," in *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 2019, p. 1526–1534.

[19] Y. Wei, Z. Cheng, X. Yu, Z. Zhao, L. Zhu, and L. Nie, "Personalized hashtag recommendation for micro-videos," in *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 2019, pp. 1446–1454.

[20] Y. Wei, X. Wang, L. Nie, X. He, and T.-S. Chua, "Graph-refined convolutional network for multimedia recommendation with implicit feedback," in *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 2020, p. 3541–3549.

[21] Y. Ding, Y. Ma, W. K. Wong, and T. Chua, "Modeling instant user intent and content-level transition for sequential fashion recommendation," *IEEE Transactions on Multimedia*, vol. 24, pp. 2687–2700, 2022.

[22] R. He and J. McAuley, "Vbpr: visual bayesian personalized ranking from implicit feedback," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, 2016, pp. 144–150.

[23] J. L. Derevensky, "Modal preferences and strengths: Implications for reading research," *Journal of Reading Behavior*, pp. 7–23, 1978.

[24] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2017.

[25] J. Ma, P. Cui, K. Kuang, X. Wang, and W. Zhu, "Disentangled graph convolutional networks," in *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 2019, pp. 4212–4221.

[26] X. Wang, H. Jin, A. Zhang, X. He, T. Xu, and T.-S. Chua, "Disentangled graph collaborative filtering," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2020, p. 1001–1010.

[27] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Proceedings of the 20th International Conference on Neural Information Processing Systems*. CAI., 2007, p. 1257–1264.

[28] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in *UAI*, 2009, pp. 452–461.

[29] W. Wang, F. Feng, X. He, H. Zhang, and T.-S. Chua, "Clicks can be cheating: Counterfactual recommendation for mitigating clickbait issue," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1288–1297.

[30] W. Wang, X. Lin, F. Feng, X. He, M. Lin, and T.-S. Chua, "Causal representation learning for out-of-distribution recommendation," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 3562–3571.

[31] H. Xue, X. Dai, J. Zhang, S. Huang, and J. Chen, "Deep matrix factorization models for recommender systems," in *Proceedings of the 26h International Joint Conference on Artificial Intelligence*. AAAI Press, 2017, pp. 3203–3209.

[32] Z. Cheng, F. Liu, S. Mei, Y. Guo, L. Zhu, and L. Nie, "Feature-level attentive icf for recommendation," *ACM Transactions on Information Systems*, vol. 40, no. 4, pp. 1–24, 2022.

[33] F. Liu, Z. Cheng, H. Chen, Y. Wei, L. Nie, and M. Kankanhalli, "Privacy-preserving synthetic data generation for recommendation systems," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2022, p. 1379–1389.

[34] C. Li, C. Quan, L. Peng, Y. Qi, Y. Deng, and L. Wu, "A capsule network for recommendation and explaining what you like and dislike," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2019, p. 275–284.

[35] R. van den Berg, T. N. Kipf, and M. Welling, "Graph convolutional matrix completion," in *2021 10th International Conference on Software and Computer Applications*. ACM, 2021.

[36] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, "Neural graph collaborative filtering," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2019, pp. 165–174.

[37] F. Liu, Z. Cheng, L. Zhu, C. Liu, and L. Nie, "An attribute-aware attentive gcn model for attribute missing in recommendation," *IEEE Transactions on Knowledge and Data Enginee*, vol. 34, no. 9, pp. 4077–4088, 2020.

[38] F. Liu, Z. Cheng, L. Zhu, Z. Gao, and L. Nie, "Interest-aware message-passing gcn for recommendation," in *Proceedings of the Web Conference 2021*. ACM, 2021, p. 1296–1305.

[39] W. Min, S. Jiang, and R. Jain, "Food recommendation: Framework, existing solutions, and challenges," *IEEE Transactions on Multimedia*, pp. 2659–2671, 2020.

[40] D. Cai, S. Qian, Q. Fang, and C. Xu, "Heterogeneous hierarchical feature aggregation network for personalized micro-video recommendation," *IEEE Transactions on Multimedia*, pp. 1–1, 2021.

[41] J. Zhang, Y. Zhu, Q. Liu, S. Wu, S. Wang, and L. Wang, "Mining latent structures for multimedia recommendation," in *Proceedings of the 29th ACM International Conference on Multimedia*. ACM, 2021, p. 3872–3880.

[42] J. Ma, C. Zhou, P. Cui, H. Yang, and W. Zhu, "Learning disentangled representations for recommendation," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. CAI., 2019.

[43] Y. Wang, S. Tang, Y. Lei, W. Song, S. Wang, and M. Zhang, "Disenhan: Disentangled heterogeneous graph attention network for recommendation," in *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*. ACM, 2020, p. 1605–1614.

[44] A. Li, Z. Cheng, F. Liu, Z. Gao, W. Guan, and Y. Peng, "Disentangled graph neural networks for session-based recommendation," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–13, 2022.

[45] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. ACL, 2019, pp. 4171–4186.

[46] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, and X. Zhai, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[47] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.

[48] G. J. Székely and M. L. Rizzo, "Brownian distance covariance," *The Annals of Applied Statistics*, vol. 3, no. 4, pp. 1236–1265, 2009.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[50] X. He, T. Chen, M.-Y. Kan, and X. Chen, "Trirank: Review aware explainable recommendation by modeling aspects," in *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*. ACM, 2015, pp. 1661–1670.

[51] G. Xavier and B. Yoshua, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*. JMLR, 2010, pp. 249–256.

[52] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.