MULTIMODAL DISENTANGLED REPRESENTATION FOR RECOMMENDATION

Xin Wang, Hong Chen* and Wenwu Zhu*

{xin_wang, wwzhu}@tsinghua.edu.cn; h-chen20@mails.tsinghua.edu.cn

*Equal Contributions

2021 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-3864-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICME51207.2021.9428193

ABSTRACT

Discovering useful information formed by various hidden factors in multimodal data has been of great importance for recommender systems to improve both model performance and recommendation explainability. These factors hidden in multimodal data are highly entangled in a complex manner, posing great challenges in uncovering their entanglement during representation learning for recommendation. However, existing literature on disentangled representation learning only pays attention to unimodal data, failing to uncover the complex and entangled factors in multimodal data. In this paper, we study the problem of multimodal disentangled representation for recommendation in a weakly supervised manner, to the best of our knowledge, for the first time. We propose a multimodal disentangled recommendation (MDR) model that can learn well-disentangled representations carrying both complementary and common information from different modalities, such that both recommendation accuracy and representation explainability can be increased. Experimental results demonstrate the superiority of our MDR model in terms of both recommendation performance and explainabilty on various real-world datasets.

1. INTRODUCTION

One of the most widely adopted pipelines in recommender systems is to learn unique representations for both users and items. Recommendations are then generated based on similarity or relevance between candidate items and target users. A large number of existing algorithms [1, 2, 3, 4, 5] have learned various representations of users and items for recommendation. The huge success of deep neural networks (DNN) has further increased the varieties of representations for recommendation [6, 3, 7]. These methods learn both user and item representations in a highly entangled manner, neglecting the significance of item attribute factors for capturing user intension, and therefore failing to disentangle them for recommendations. The learned representations may mistakenly carry confounding factor information, which may result in non-robustness, low explainability, and even bad accuracy. The key to enhancing recommendation accuracy and explainability, two important aspects that can help increase user satisfaction, lies in discovering useful information formed by various factors in multimodal data. These factors, however, are entangled in a complex way, generating great difficulties in learning disentangled representations for more accurate and explainable recommendation.

Disentangled representation learning [8] aims to learn disentangled representations uncovering the hidden explanatory factors carried in observable data, which has received considerable attention. Disentangled representation benefits in several advantages such as robustness, i.e., less sensitive to misleading information in sparse data, and explainability, i.e., able to tell what different factors indicate in vectorized representations. These advantages essentially make disentangled representation find its direct applications in explainable recommendation [9, 10]. Nevertheless, most existing literature on disentangled representation learning mainly lies in the field of computer vision [11, 12, 13], natural language processing [14] and graph representation [15]. Works on disentangled representation learning for recommendation like [16] merely discovers user intention from sparse user purchase behavior, which is hard to explain and fails to handle multimodal scenarios where items have multiple modalities, each of which may carry either complementary or common information with others.

In this paper, we study the problem of multimodal disentangled representation learning for recommendation in a weakly supervised manner, to the best of our knowledge, for the first time. However, there are three challenges in learning multimodal disentangled representations for recommendation:

- One straightforward way to generate multimodal disentangled representation is directly using unimodal disentangled methods to obtain unimodal disentangled representation and concatenate them together. However, this will surely make the learned representation redundant and not perform well in downstream tasks, because different modalities may share some factor information in common, e.g., color, and not all factor information will be useful to recommendation.
- Disentangled factors with explainable semantic meanings in latent representations cannot be easily fixed to particular dimensions in existing unsupervised representation learning methods, and attribute information is hard to align between different modalities.

978-1-6654-3864-3/21/$31.00 ©2021 IEEE

- It is difficult to conduct multimodal fusion while preserving useful information from each modality and ensuring disentanglement in each group of factors simultaneously.

To tackle these challenges, we propose a multimodal disentangled recommendation (MDR) model capable of learning well-disentangled representations which carry both complementary and common information from different modalities. More concretely, we adopt a tailored encoding, fusion, and decoding structure with specially designed disentangled methods to obtain multimodal disentangled representation. During the encoding process, the input image and text are encoded into unimodal representations with particular encoders respectively. Then the obtained unimodal representations are fused in a disentangled manner to obtain a multimodal representation. During the decoding process, reconstruction loss is used to preserve adequate input information, and minimize mutual information among facotrs will guarantee disentanglement, where weak supervision is introduced to further improve representation explainability. We utilize regularized information constraints to ensure good explainability of the learned multimodal disentangled representations for recommendation. We further apply the multimodal disentangled representation to recommendation tasks, and experimental results against several state-of-the-art methods on various real-world datasets demonstrate that our model outperforms baseline approaches in terms of both accuracy and explainability.

Our main contributions are summarized as follows:

- We are the first to study the problem of multimodal disentangled representation for recommendation in a weakly supervised manner.
- We propose a multimodal disentangled recommendation (MDR) model which can learn well-disentangled representations carrying both complementary and common information from different modalities, where regularized information constraints are employed to increase recommendation explainability.
- We utilize our learned representation to recommendation tasks and conduct extensive experiments to show the superiority of our proposed model against several state-of-the-art baseline methods by validating its effectiveness in both recommendation accuracy and explainability on various real-world datasets.

2. RELATED WORK

Multimodal Recommendation. Several multimodal recommendation works [6, 17, 3] aim to learn deep representations for both users and items from human statistics like shop ids, item categories, etc. He et al. [1] extracts visual features of items through deep neural networks for recommendation, considering the fact that people generally purchase an item upon visual cognition. Li et al. [18] makes recommendations based on the information of reviews from other customers. Chen et al. [7] propose a VECF model utilizing both images and text reviews for learning multimodal item representations. Yang et al. [19] also takes advantage of multimodal information for recommendation through learning weights for different modalities.

Disentangled Representation. Kingma and Welling [11] propose to utilize Bayesian posterior inference and variational estimation to learn the controllable factors hidden in the observed data. Higgins et al. [12] propose $\beta - VAE$ by setting a weight β for the KL divergence to improve representation disentanglement learned in [11]. Kim and Mnih [13] borrow the idea of InfoGAN to optimize the loss function of VAE Jian et al. [20] utilize methods based on triplets to learn aspect representations from sentences. To further improve the degree of disentanglement in representation learning, some weakly supervised models are introduced [21, 22, 23]. Ma et al. [15] apply the idea of disentanglement in training of graph convolutional networks. They later learn both macroscopic and microscopic unimodal disentangled representations for users in recommendation [16].

3. MULTIMODAL DISENTANGLED RECOMMENDATION

3.1. Multimodal Disentangled Representation learning

In this work, we assume there are two modalities, i.e., image and text, describing each item in recommendation. However, we remark that our proposed MDR model is applicable to multiple modalities. The overall architecture of our proposed MDR model is shown in Figure 1. **Image Encoding**. We adopt a VAE structure to learn the image disentangled representation. The *Image Encoder* outputs a distribution of disentangled factors behind the input image, whose mean vector is f_i and standard deviation vector is λ_i.

$$ImageEncoder(x_i) = [f_i, \lambda_i], \tag{1}$$

where $f_i, \lambda_i \in R^N$. We denote $s_i \sim \mathcal{N}(f_i, \lambda_i)$ as a group of controllable latent factors following a Guassian Distribution parameterized by f_i and λ_i. Let $s_i = [u_{1i}, u_{2i}, \cdots, u_{ji}, \cdots, u_{mi}]$ where u_{ji} is the j-th attribute with explainable semantic meanings (e.g., color or shape) and $f_i = [c_{1i}, c_{2i}, \cdots, c_{mi}]$, $\lambda_i = [\delta_{1i}, \delta_{2i}, \cdots, \delta_{mi}]$, then $u_{ji} \sim \mathcal{N}(c_{ji}, \delta_{ji})$, $\forall j \in [1, m]$, where i, j are all integers. Assuming there are totally m attributes, then each attribute occupies $\frac{N}{m}$ dimensions of s_i.

Text Encoding. Given an input sentence, we employ *gensim* tools [1] to map each word in the sentence to its corresponding word embedding. As such, the input sentence can be represented as $sentence_i = [w_{e1}, w_{e2}, \cdots, w_{ej}, \cdots, w_{en}]$, where w_{ej} is the embedding of the j-th word of the sentence. The whole

[1] https://radimrehurek.com/gensim/auto_examples/core/run_core_concepts.html

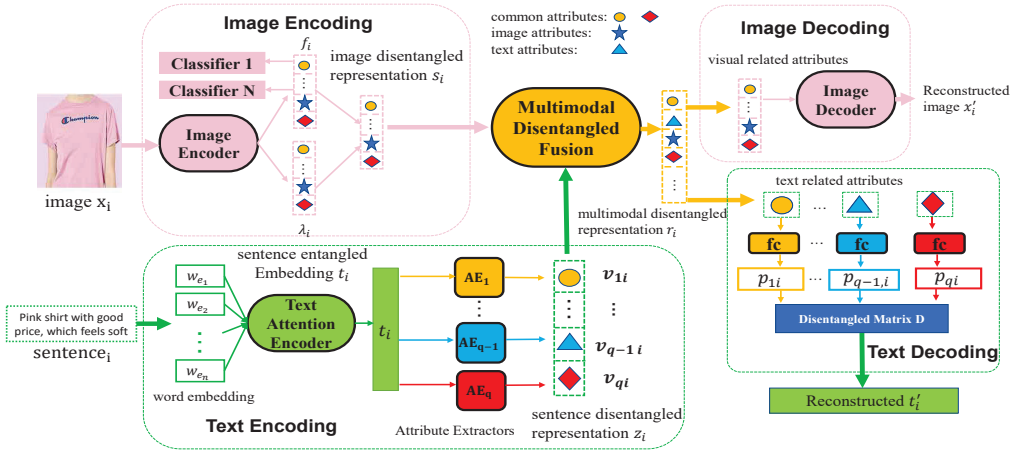


Fig. 1. MDR: the overall architecture of multimodal disentangled representation learning model

sentence representation for $sentence_i$, denoted as t_i, can be obtained by an attention based encoder as follows,

$$a_j = w_{ej}^T M\,(\frac{1}{n}\sum_{t=1}^{n} w_{et}), b_k = \frac{e^{a_k}}{\sum_{j=1}^{n} e^{a_j}}, t_i = \sum_{k=1}^{n} b_k \cdot w_{ek}, \quad (2)$$

where M is the attention matrix that needs to be learned.We design a group of attribute extractors for each explainable attribute to obtain corresponding disentangled latent factor from the highly entangled vector t_i.

$$v_{ji} = AE_j(t_i), j = 1, 2, \cdots, q. \quad (3)$$

where v_{ji} is the text attribute factors, and the disentangled text representation is obtained by concatenating all v_{ji} , which is $z_i = [v_{1i}, v_{2i}, \cdots, v_{qi}]$.

Algorithm 1 Multimodal Disentangled Fusion

Input:
1: visual disentangled representation for $item_i$, $s_i = [u_{1i}, u_{2i}, \cdots, u_{mi}]$
2: text disentangled representation for $item_i$, $z_i = [v_{1i}, v_{2i}, \cdots, v_{qi}]$
Output: multimodal disentangled representation for $item_i$, r_i.
3: For visual-dominant explainable attributes that visual and text signals share in common, e.g., color and shape, $r_{ji} = u_{ji} + fc(fc([u_{ji}, \alpha * v_{ji}]))$, where $\alpha \in [0, 1]$ serves as a hyperparameter and fc denotes a fully connected layer.
4: For other text-dominant explainable attributes shared by both visual and text signals, $r_{ji} = v_{ji} + fc(fc([\alpha * u_{ji}, v_{ji}]))$.
5: For attributes only possessed by visual signals, $r_{ji} = u_{ji}$.
6: For attributes only possessed by text signals, $r_{ji} = v_{ji}$.
7: For the remaining attributes that could be hardly explained, we handle them in a fully black-box manner, $r_{blackbox,i} = fc(fc([s_{blackbox,i}, z_{blackbox,i}]))$, where $s_{blackbox,i}$ and $z_{blackbox,i}$ denote the unexplainable factors in visual and text signals, respectively.
8: $r_i = [r_{1i}, r_{2i}, \cdots, r_{Ki}, r_{blackbox,i}]$.

Multimodal Disentangled Fusion. As is shown in Algorithm 1, we fuse explainable attributes shared by both visual and text signals with residual structure and concatenate explainable attributes existing in only one modality directly to the fusion representation. In order to preserve information from each modality as much as possible, we take care of the attributes which can be hardly explained in a black-box way to generate $r_{blackbox,i}$ through utilizing two fc layers to fuse these unexplainable latent factors from each modality together. This process may decrease the degree of disentanglement in multimodal fusion because the disentanglement of $r_{blackbox,i}$ can not be guaranteed. To tackle this issue, we set regularization constraints (which we call information constraints) to prevent these black-box factors from containing too much information, therefore leading the disentangled factors to contribute more to the final recommendation performance. More concretely, $L2$ regularization is set for all the parameters of the two fully-connected layers to restrict the values of $r_{blackbox,i}$. These regularization terms will deteriorate the ability of the black-box factors to carry unexplainable information by pushing the values of more parameters of the black-box factors towards zero, thus forcing the black-box factors to contribute less to later recommendation while explainable factors contribute more. The loss introduced by the $L2$ regularization can be written as follows,

$$loss_{reg_i} = ||W||_2, \quad (4)$$

where W refers to all the parameters used to fuse $s_{blackbox,i}$ and $z_{blackbox,i}$ into $r_{blackbox,i}$. With above fusion strategies, we obtain the multimodal disentangled representation r_i.

Decoding and Disentangling. For the visual related attributes contained in r_i shown in Figure 1, they should contain enough information to reconstruct the input image x_i. Mathematically, the marginal log-likelihood of the observed data x_i given r_i in expectation over the distribution of r_i given x_i and $sentence_i$ needs to be maximized:

$$\max_{\phi,\theta} \mathbb{E}_{p_\theta(r_i|x_i,sentence_i)} log(q_\phi(x_i|r_i)), \quad (5)$$

where θ and ϕ denote parameters in the encoding-fusion and decoding process respectively. Besides, some mutual information constraints should be set so that the visual related attributes could be disentangled. Since the visual related attributes are obtained from s_i and z_i in an attribute-level fusion manner, once factors in s_i and z_i are disentangled, the visual

related attributes after fusion will be well disentangled. We first discuss disentanglement of s_i and leave the text related attributes disentanglement later. We need $p_\theta(s_i|x_i)$, the posterior distribution of s_i, to be close to the normal Gaussian distribution, $p(s_i) = N(0, I)$, where factors are independent and mutual information among factors are naturally zero. Hence, the constraints could be formulated as follows,

$$D_{KL}(p_\theta(s_i|x_i)||p(s_i)) < \sigma, \tag{6}$$

where σ is positive and the left term is the KL Divergence. Therefore, the optimal objective could be written as follows with KKT conditions,

$$\max_{\phi,\theta} \mathbb{E}_{p_\theta(r_i|x_i,sentence_i)} log(q_\phi(x_i|r_i)) - \beta * D_{KL}(p_\theta(s_i|x_i)||p(s_i)), \tag{7}$$

The first term guarantees r_i containing rich input information and the second term makes s_i more disentangled. About the second term, by denoting that the actual distribution of the disentangled representation s_i equals to $q(s_i)$, it can be reformulated as follows,

$$\mathbb{E}_{p(x_i)} D_{KL}(p_\theta(s_i|x_i)||p(s_i))) = I(s_i; x_i) + D_{KL}(q(s_i)||p(s_i)). \tag{8}$$

This indicates that the KL-divergence in Eq (6) can be written as the sum of mutual information between x_i and s_i and the KL-divergence between the $q(s_i)$ and the preset prior distribution $p(s_i)$. Penalization on the latter term will give rise to more closeness between the actual distribution and the independent prior distribution, thus increasing disentanglement.

In experiments, the first term in Eq (7) corresponds to reconstruction of x_i from r_i, and we set regularizers as follows to maximize it,

$$loss_{rec,i} = ||x_i - x_i'||_2, x_i' = ImageDecoder(r_i). \tag{9}$$

As for the second KL Divergence term, it could be written as follows with the assumption that $p(s_i) = N(0, I)$,

$$loss_{KL_i} = -\frac{1}{N} \sum_{k=0}^{N-1} 0.5 * \left(1 + \lambda_i[k] - f_i[k]^2 - e^{\lambda_i[k]}\right), \tag{10}$$

where $\lambda_i[k]$ and $f_i[k]$ denote the k-th element in vectors λ_i and f_i respectively. And the whole regularizer for visual attributes could be written as:

$$loss_{visual,i} = loss_{rec,i} + \beta * loss_{KL,i}. \tag{11}$$

Furthermore, weak supervision is introduced to improve disentanglement of s_i and locate explainable attributes with semantic meanings within our expected dimensions of the latent factors. We label a small number of image attributes (such as the color of an item) by human annotations and design several attribute classifiers for the latent factors as shown in Figure 1. Given $item_i$, when expecting u_{ji} to represent color, we set a color classifier $classifier_j$ for c_{ji} whose output is the predicted color category, like red or yellow, etc. Mathematically,

$$att_{ji} = classifier_j(c_{ji}),\ j = 1, 2, \cdots, m, \tag{12}$$

where $att_{ji} \in R^n$ and n is the total number of categories for attribute j. Suppose $item_i$ has a label in terms of attribute j, which is denoted as a one-hot vector $label_{ji} \in R^n$, then the weakly supervised loss from classifiers can be written as follows:

$$loss_{ws_i} = \sum_{j \in M} CE(label_{ji}, att_{ji}), \tag{13}$$

where M is the set of labeled attributes for item i and CE is the cross-entropy loss.

The text decoding process is shown in Figure 1. Similar to the situation within image modality, r_i should also contain disentangled text related attributes r_{ji} that could reconstruct the input sentence. The reconstruction process can be described as follows,

$$p_{ji} = fc_j(r_{ji}), \quad t_i' = (p_i \cdot D)^T = \sum_{j=1}^{q} p_{ji}\, d_j, \tag{14}$$

where p_{ji} is the calculated importance weight for each factor and $D = [d_1, d_2, \cdots, d_j, \cdots, d_q]^T$, each d_j is a normalized column vector which corresponds to a disentangled latent representation for each attribute, and the total attribute number is q. The parameters in D are trainable. Moreover, we introduce two regularizers to guarantee adequate input information and disentanglement of textual factors in r_i, which can be described as follows:

$$loss_{text_i} = \sum_{j=1}^{g} max(0, 1 - t_i^T\, t_i' + n_j^T t_i') + ||D * D^T - I||, \tag{15}$$

where n_j in the first term represents the mean vector of the negative sentence randomly sampled from text corpus, and g is negative sample number. The first term is the reconstruction loss targeting at pushing t_i' to have closer semantic meanings with current input sentence compared to other negative samples, while the second term pushes $D * D^T$ closer to an identity matrix I, targeting at making each row d_{ji}^T in D to disentangle from others because the inner product of different rows are induced to be zero. Therefore, the corresponding weights p_{ji} and text related attributes r_{ji} that have information only related to d_{ji}^T are disentangled. To this end, text related fusion factors in r_i are disentangled.

3.2. Recommendation

The structure of our recommendation framework is shown in Figure 2. Assuming user A has purchased $[item_1, \cdots, item_j]$, we need to predict whether she would buy $item_i$. The whole process can be formulated as follows:

$$h_k = MDR(sentence_k, x_k), k = 1, 2, \cdots, j \tag{16}$$

$$h_i = MDR(sentence_i, x_i), \tag{17}$$

$$wh_k = DIN\_attentive(h_i, h_k), k = 1, 2, \cdots, j \tag{18}$$

$$u_A = \sum_{k=1}^{j} wh_k h_k, \tag{19}$$

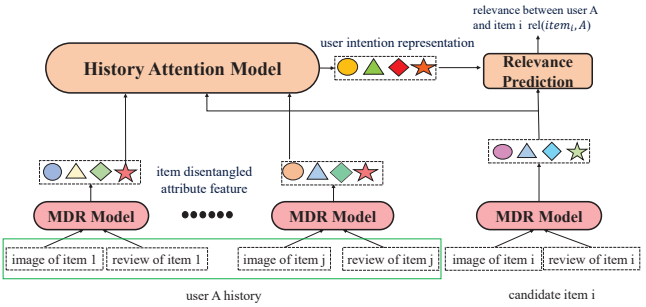


Fig. 2. The structure of our recommendation framework based on the learned disentangled representation

$$rel(item_i, A)) = DIN\_predict(u_A, h_i). \quad (20)$$

First, we use MDR model to learn multimodal disentangled representation for user history items and candidate item i and obtain h_k and h_i, respectively. After that, we apply the DIN attentive mechanism [3] to calculate the importance of each history item and get a scalar wh_k and regard the weighted sum of all h_k as the user intention representation. Finally, both the user intention u_A and candidate $item_i$ representation is sent to the DIN prediction module to predict whether user A would buy $item_i$. Denoting $rel(item_i, A)$ as $y_i{}'$, the prediction loss is formulated as follows:

$$loss_{A,pred_i} = -y_i log y_i' - (1 - y_i) log(1 - y_i'), \quad (21)$$

where $y_i \in \{0, 1\}$ is the ground truth and $y_i = 1$ means candidate $item_i$ is in the set of items that are purchased by users. $y_i' \in (0, 1)$ indicates the predicted relevance between $item_i$ and the target user.

The Overall Objective. The overall objective is to minimize the following loss:

$$loss_{A,i} = \gamma * loss_{A,pred_i} + \sum_{k \in M_A} (loss_{visual,k} + loss_{text_k} + \eta * loss_{reg_k} + \mu * loss_{ws_k}), \quad (22)$$

where γ, μ, β, η are hyperparameters and $loss_i$ indicates the loss of recommending $item_i$ to the target user, M_A represents the set of history items and the candidate item for user A.

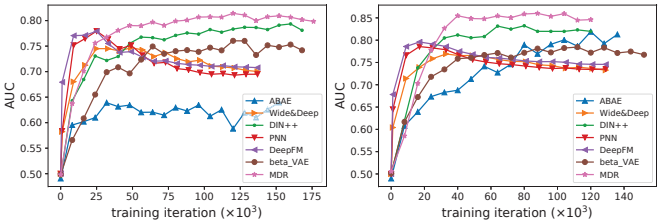
4. EMPIRICAL EXPERIMENTS

Datasets. We conduct experiments on three Amazon product datasets [24], i.e., Amazon Cloth, Toys and Office, containing rich information about users and items, whose detailed information is shown in Table 1.

Comparative Approaches. We compare the proposed MDR model with state-of-the-art methods (i.e., **Wide&Deep** [6], **PNN** [25], **DeepFM** [17]) using sparse features such as one-hot embedding as well as other baselines with unimodal disentangled representations (i.e., β**-VAE** [12], **ABAE** [14]) and concatenated representations of β**-VAE** and **ABAE**, denoted as **DIN++**.

Table 1. Dataset statistics

Dataset	users	goods	categories	behaviors	sparsity(%)
Cloth	39387	23033	484	278677	0.031
Toys	19412	11924	366	167597	0.072
Office	4905	2420	279	53258	0.449



(a) Cloth (b) Toys

Fig. 3. AUC for Wide&Deep, PNN, DeepFM, DIN++, β-VAE, ABAE, and MDR on Amazon Cloth and Toys datasets

Model Performance. We adopt the weighted AUC in [3] and the widely adopted NDCG metrics to evaluate the recommendation performance. Figure 3 depicts the AUCs for various comparative methods on Amazon Cloth and Toys datasets. We observe that our proposed MDR model significantly beats other baselines in all three datasets. Table 2 shows the best result of AUC and NDCG@10 metrics for different models on three datasets, illustrating the advantages of our proposed MDR model.

Table 2. AUC and NDCG@10 for different models

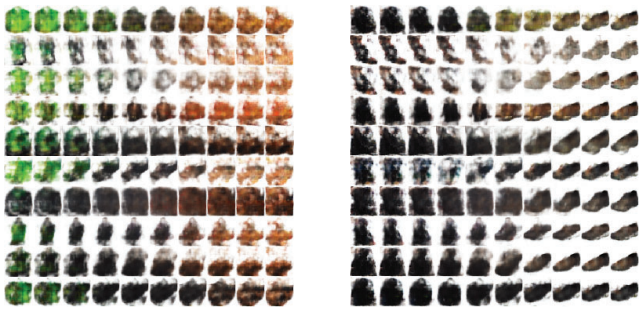
Model	**Cloth**		**Toys**		**Office**	
	AUC	NDCG	AUC	NDCG	AUC	NDCG
Wide&Deep	0.7547	0.2278	0.7728	0.2445	0.8030	0.2434
PNN	0.7794	0.2591	0.7945	0.2837	0.8235	0.2705
DeepFM	0.7846	0.2568	0.8046	0.2949	0.8290	0.2673
β-VAE	0.7671	0.2213	0.7849	0.1971	0.8372	0.2740
ABAE	0.6750	0.1399	0.8359	0.1191	0.8157	0.1231
DIN++	0.7938	0.2396	0.8336	0.2427	0.8671	0.3420
MDR (ours)	**0.8070**	**0.2598**	**0.8675**	**0.3048**	**0.8738**	**0.3955**

Effectiveness of Information Constraints. We conduct ablation study on the effectiveness of our proposed information constraints. w/o IC in Table 3 show the performance of MDR model *without* information constraints ,validating its effectiveness in our MDR model.

Visual Disentanglement. We change dimensions of color and shape attributes in multimodal disentangled representation, and use the representation to reconstruct corresponding images in Figure 4. Each row has 10 pictures whose representations differ only in dimensions representing colors or shapes. Figure 4 (a) shows the changes of visual signals with respect to dimensions controlling color attribute on the Amazon Cloth dataset, while Figure 4 (b) shows the changes with respect to dimensions controlling the shape attribute.

Table 3. Ablation Study on Effectiveness of IC

Model	Cloth		Toys		Office	
	AUC	NDCG	AUC	NDCG	AUC	NDCG
MDR	**0.8070**	**0.2598**	**0.8675**	**0.3048**	**0.8738**	**0.3955**
w/o IC	0.7893	0.2407	0.8002	0.2336	0.8428	0.2762



(a) Amazon Cloth *Color* (b) Amazon Cloth *Shape*

Fig. 4. Visual disentanglement on Amazon Cloth dataset

Text Disentanglement. We try to find out the semantic meaning of attribute $v_{1i}, \cdots, v_{qi}$ in Figure 1. For Amazon Cloth dataset, we can easily observe from Figure 5 that attribute 5 stands for *purchase purpose*, attribute 11 refers to *color*, which shares the same dimensions as that of Figure 4, attribute 13 indicates *size* and attribute 9 implies *material*.

Aspect	Related Words
0	['compartment', 'card', 'purse', 'bag', 'wallet', 'carry', 'luggage', 'pocket', 'laptop', 'phone']
1	['mentioned', 'many', 'complained', 'comment', 'agree', 'suggest', 'commented', 'personal', 'know', 'noted']
2	['moisture', 'muy', 'push', 'pad', 'protect', 'de', 'foam', 'protection', 'e', 'el']
3	['muy', 'de', 'la', '1086', 'e', 'que', 'para', '1077', 'lo', 'el']
4	['loop', 'velcro', 'hole', 'sewn', 'broke', 'seam', 'attached', 'pulled', 'closure', 'hook']
5	['birthday', 'christmas', 'gift', 'niece', 'party', 'sister', 'friend', 'granddaughter', 'mother', 'daughter']
6	['seller', 'refund', 'delivery', 'shipped', 'service', 'arrived', 'shipping', 'sent', 'delivered', 'return']
7	['tummy', 'bust', 'bra', 'waist', 'tank', 'panty', 'chest', 'thigh', 'butt', 'belly']
8	['errand', 'beach', 'pool', 'outdoor', 'hiking', 'weather', 'yard', 'winter', 'morning', 'office']
9	['cotton', 'soft', 'wash', 'fabric', 'material', 'moisture', 'warm', 'fleece', 'washing', 'washed']
10	['watch', 'dial', 'ring', 'casio', 'battery', 'invicta', 'seiko', 'alarm', 'timex', 'stainless']
11	['rich', 'bright', 'vibrant', 'blue', 'darker', 'pink', 'color', 'match', 'brown', 'purple']
12	['klein', 'name', 'owned', 'superior', 'brand', 'similar', 'loom', 'compared', 'champion', 'fan']
13	['medium', 'xl', 'lb', 'large', 'chart', '11', 'pound', 'size', 'med', 'junior']
14	['arch', 'foot', 'toe', 'plantar', 'shoe', 'heel', 'insole', 'cushioning', 'sandal', 'sole']

Fig. 5. Text disentanglement on Amazon Cloth dataset

In general, the multimodal disentangled representation obtained by our proposed MDR model shows excellence at disentanglement and explainability. Particular dimensions of the learned multimodal disentangled representation are able to directly indicate particular explainable attributes with semantic meanings. Besides, the information obtained from visual and text signals is complementary, for example, visual signals do not contain information about material and brand which existing in textual information, validating the importance and necessity of learning disentangled representations for multimodal information.

5. CONCLUSION

In this paper, we study the problem of multimodal disentangled representation learning for recommendation for the first time. We propose a multimodal disentangled representation (MDR) model capable of learning well-disentangled multimodal representations which carry both common and complementary useful information across different modalities. Extensive experiments validates the superiority of the MDR model over existing literature.

Acknowledgment

This work is supported by the National Key Research and Development Program of China (No.2020AAA0107800, No.2020AAA0106300, No.2018AAA0102000) and National Natural Science Foundation of China No.62050110.

6. REFERENCES

[1] R. He et.al., "Vbpr: visual bayesian personalized ranking from implicit feedback," in *AAAI*, 2016.
[2] S. Shi et.al., "Attention-based adaptive model to unify warm and cold starts recommendation," in *CIKM*, 2018, pp. 127–136.
[3] G. Zhou et.al., "Deep interest network for click-through rate prediction," in *KDD*, 2018.
[4] X. Wang et.al., "Social recommendation with optimal limited attention," in *KDD*, 2019, pp. 1518–1527.
[5] C. Chen et.al., "An efficient adaptive transfer neural network for social-aware recommendation," in *SIGIR*, 2019, pp. 225–234.
[6] T. Cheng et.al., "Wide & deep learning for recommender systems," in *Proceedings of the 1st workshop on deep learning for recommender systems*, 2016.
[7] X. Chen et.al., "Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation," in *SIGIR*, New York, NY, USA, 2019.
[8] Y. Bengio et.al., "Representation learning: A review and new perspectives," *IEEE TPAMI*, vol. 35, no. 8, pp. 1798–1828, 2013.
[9] Y. Zhang et.al., "Explicit factor models for explainable recommendation based on phrase-level sentiment analysis," in *SIGIR*, 2014, pp. 83–92.
[10] X. He et.al., "Trirank: Review-aware explainable recommendation by modeling aspects," in *CIKM*, 2015, pp. 1661–1670.
[11] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
[12] T. Higgins et.al., "beta-vae: Learning basic visual concepts with a constrained variational framework.," *ICLR*, 2017.
[13] Hyunjik Kim and Andriy Mnih, "Disentangling by factorising," *arXiv preprint arXiv:1802.05983*, 2018.
[14] R. He et.al., "An unsupervised neural attention model for aspect extraction," in *ACL*, 2017.
[15] J. Ma et.al., "Disentangled graph convolutional networks," in *ICML*, 2019.
[16] J. Ma et.al., "Learning disentangled representations for recommendation," in *NeurIPS*, 2019.
[17] H. Guo et.al., "Deepfm: a factorization-machine based neural network for ctr prediction," *arXiv preprint arXiv:1703.04247*, 2017.
[18] P. Li et.al., "Neural rating regression with abstractive tips generation for recommendation," in *SIGIR*, 2017.
[19] B. Yang et.al., "Online video recommendation based on multimodal fusion and relevance feedback," in *Proceedings of the 6th ACM international conference on Image and video retrieval*, 2007.
[20] S. Jain et.al., "Learning disentangled representations of texts with application to biomedical abstracts," *arXiv preprint arXiv:1804.07212*, 2018.
[21] F. Locatello et.al., "Disentangling factors of variation using few labels," *arXiv preprint arXiv:1905.01258*, 2019.
[22] D. Kingma et.al., "Semi-supervised learning with deep generative models," in *NeurIPS*, 2014.
[23] Z. Feng et.al., "Dual swap disentangling," in *NeurIPS*, 2018.
[24] R. He et.al., "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *WWW*, 2016.
[25] Y. Qu et.al., "Product-based neural networks for user response prediction," in *ICDM*, 2016.