Modality Matches Modality: Pretraining Modality-Disentangled Item Representations for Recommendation

Tengyue Han[*]
Beijing University of Posts and Telecommunications
hantengyue@bupt.edu.cn

Pengfei Wang[*][†]
Beijing University of Posts and Telecommunications
wangpengfei@bupt.edu.cn

Shaozhang Niu[*]
Beijing University of Posts and Telecommunications
szniu@bupt.edu.cn

Chenliang Li[†]
Wuhan University
cllee@whu.edu.cn

ABSTRACT

Recent works have shown the effectiveness of incorporating textual and visual information to tackle the sparsity problem in recommendation scenarios. To fuse these useful heterogeneous modality information, an essential prerequisite is to align these information for modality-robust features learning and semantic understanding. Unfortunately, existing works mainly focus on tackling the learning of common knowledge across modalities, while the specific characteristics of each modality is discarded, which may inevitably degrade the recommendation performance.

To this end, we propose a pretraining framework **PAMD**, which stands for **P**retr**A**ining **M**odality-**D**isentangled Representations Model. Specifically, PAMD utilizes pretrained VGG19 and Glove to embed items' both visual and textual modalities into the continuous embedding space. Based on these primitive heterogeneous representations, a disentangled encoder is devised to automatically extract their modality-common characteristics while preserving their modality-specific characteristics. After this, a contrastive learning is further designed to guarantee the consistence and gaps between modality-disentangled representations. To the best of our knowledge, this is the first pretraining framework to learn modality-disentangled representations in recommendation scenarios. Extensive experiments on three public real-world datasets demonstrate the effectiveness of our pretraining solution against a series of state-of-the-art alternatives, which results in the significant performance gain of 4.70%-17.44%.

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

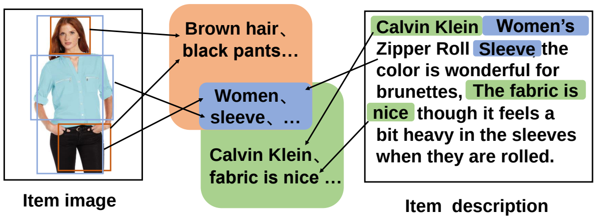

Brown hair、 black pants…
Women、 sleeve、 …
Calvin Klein、 fabric is nice …
Calvin Klein Women's Zipper Roll Sleeve the color is wonderful for brunettes, The fabric is nice though it feels a bit heavy in the sleeves when they are rolled.
Item image
Item description

Figure 1: An example to illustrate the modality alignment problem in recommendation scenario. The characteristics in blue background are aligned modality features, while characteristics in orange and green background are modality-specific features of item image and description respectively.

KEYWORDS

Pretraining, Disentangle Encoder, Contrasive Learning, Modality-Disentangled Representation

ACM Reference Format:
Tengyue Han, Pengfei Wang, Shaozhang Niu, and Chenliang Li. 2022. Modality Matches Modality: Pretraining Modality-Disentangled Item Representations for Recommendation. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3485447.3512079

1 INTRODUCTION

Recommender system is a widespread application of artificial intelligence techniques. Such system plays a central role in shopping scenarios by sifting the items a user prefers from huge corpora [21, 33, 34]. To generate a high-quality recommendation, fusing different modalities relevant with items (*i.e.,* textual or visual information) to alleviate the sparsity problem now is becoming an attractive topic [2, 3, 5, 13, 14, 24].

Previous works have demonstrated the effectiveness of learning uni-modality item representations in recommendation scenarios, *e.g.,* using textual or image data [12, 20, 24]. However, exploiting uni-modality information may not produce a comprehensive feature learning for each item than combining multiple heterogeneous information of different modalities together. Generally, to fuse these heterogeneous information together, an essential prerequisite is to align these cross-modality data precisely for facilitating both modality-robust feature learning and better preference estimation. Unfortunately, existing works mainly focus on learning common

[*] Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia
[†] Corresponding author.

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9096-5/22/04…$15.00
https://doi.org/10.1145/3485447.3512079

knowledge across modalities [4, 6, 11], while the specific characteristics of each modality is usually discarded, which could be very useful for recommendation.

In Fig. 1, we choose a simple example to explain this point further. We can see that both the image and the textual description of an item can enrich semantics that are beneficial to the recommendation task. Specifically, both the textual description and the image share some common knowledge and also provide modality-specific knowledge towards the item respectively. Hence, the key challenge is to design appropriate matching metrics between these modalities for alignment without explicit supervision. Though previous works [9, 41] utilize encoder-decoder models to align their common characteristics (*i.e.,* features in blue background), the modality-specific characteristics (*i.e.,* the features in either green or orange background) are always neglected. This information loss phenomenon may inevitably degrade the recommendation performance. A simple solution is to directly fuse heterogeneous information of different modalities with some simple operation like vector concatenation. However, this straightforward fusion approach encodes both modality-common and modality-specific features into a single vector representation and therefore complicates the model learning process, leading to inferior performance. How to align common characteristics of multi-modal data and preserve their specific characteristics, is an interesting yet challenging problem.

To this end, a pretraining modality disentangled representation framework, named PAMD is proposed to support multi-modality driven item recommendation. PAMD contains two components: a *disentangled encoder* to decompose multi-modal information into modality-common and modality-specific features, and a *contrastive learning* with a self-supervised objective to guarantee the precise alignment and separation between these decomposed features. Specifically, given both the visual and textual modality that an item is associated with, PAMD utilizes the off-the-shelf VGG19 [36] and Glove [30] to embed modalities into the continuous embedding space respectively. Based on these primitive heterogeneous representations, PAMD first feeds them into a disentangled encoder, where each modality data is decomposed into two representations: a modality-common representation that shares semantics across modalities, and a modality-specific representation that contains the unique characteristics in this modality channel. After this, a contrastive learning is further designed to ensure precise alignment and separation between these decomposed representations. Owing to the benefit of self-supervised training, our PAMD can automatically learn modality-disentangled item representations without supervised signals, and are proven to be more effective than the existing state-of-the-art recommendation models. To summarize, the main contributions of this paper are listed as follows:

- We propose a pretraining framework to learn modality-disentangled item representations. To the best of our knowledge, this is the first attempt in leveraging multi-modal information in recommendation scenarios in a self-supervised pretraining paradigm.
- In our proposed PAMD, we design a disentangled encoder to decompose representations, and a contrastive learning to guide the alignment and separation between these decomposed representations.
- Empirical results on three real world datasets demonstrate that the proposed pretraining scheme is feasible and our PAMD achieves significantly better recommendation performance than many state-of-the-art baselines.

2 RELATED WORKS

There have been many representative efforts of exploiting different modality data for recommendation. In this section, we provide a brief overview of the relevant works from two perspectives: modality representation learning and the pretraining techniques.

2.1 Multi-modal Representation Learning

Multi-modal representation learning which aims to learn a comprehensive representation by considering heterogeneous information from different modalities jointly, now is becoming one of the most important problem in multi-modal applications [38, 43]. Currently, there are two dominating learning paradigms for modality fusion [1]: joint representation learning and coordinated representation learning.

The joint representation learning aims to project uni-modal representations together into a shared semantic subspace for modality fusion, which is usually used in tasks where multi-modal data is present during both training and inference stages. Previous works usually concatenate single-modal feature directly to obtain the final representations [22, 28]. Due to the recent revive of neural models, many works attempt to learn a better function for modalities fusion with deep neural networks. For example, Jiang *et al.* [19] proposed a novel unified framework that jointly exploits the feature relationships and the class relationships in a high-level semantics. Qi *et al.* [31] captured cross-modal correlation by bidirectional translation training in both visual and textual feature spaces.

Instead of projecting the modalities together into a joint space, the coordinated representation leaning aims to learn each modality data separately, and coordinates them with constraints. This learning approach is suitable for applications where only one modality is present at inference stage, such as multi-modal retrieval, translation and zero-shot, etc. For example, Yan *et al.* [42] matched images and captions in a joint latent space built by using deep canonical correlation analysis. Huang *et al.* [18] proposed a multi-modal unsupervised image-to-image translation framework to learn the conditional distribution of corresponding modality in the target domain. Wu *et al.* [40] presented a general framework for the zero-shot event detection using only textual descriptions.

Recently, multi-modal fusion approach is also introduced into recommendation system to alleviate the data sparsity problem. Researchers have developed a series of hybrid approaches of incorporating the items' content information and the collaborative filtering effects for recommendation [25]. For example, Chen et al. [4] explored the fine-grained user preference on the items and introduced a novel attention mechanism to model item- and component-level feedbacks in multimedia recommendation. Wei et al. [39] devised a novel GCN-based framework to leverage information exchange between users and micro-videos for better recommendation.

2.2 Pretraining Models

Pretraining has been intensively studied in the past years. A series of studies have gradually pushed the frontier of model performance and proven its worth in a wide range of tasks [26, 29, 44].

For example, Devlin *et al.* [8] introduced masked language modelling, which aims to encode the contextual semantics in terms of interactions between left and right context words. Lewis *et al.* [23] designed a denoising autoencoder to pretrain sequence-to-sequence models. Dahl et al. [7] found that pretraining with deep belief networks can well improve the feedforward acoustic models. Mert et al. [35] introduced image-conditioned masked language modeling as a proxy task to learn visual representations over image-caption pairs.

Recently, some recommendation works [32, 45] also employed the pretraining technique by learning better hidden representations to enhance the recommendation performance [27]. For example, Sun *et al.* [37] utilized the Transformer and the Cloze objective to learn bidirectional contextual representations for the user interactions. Zhou *et al.* [46] designed several auxiliary self-supervised tasks to enhance the data representations for sequence data to improve sequential recommendation. Qiu *et al.* [32] designed a pretraining framework that utilized two self-supervision tasks to leverage the abundant reviews in other domains to model user preferences for recommendation.

To the best of our knowledge, we are the first work of leveraging multi-modal information in a pretraining framework for recommendation. We introduce a novel pretraining framework, which respectively models the common part and specific part of characteristics, to resolve the above mentioned problem.

3 PRELIMINARY

Let $\mathcal{I}$ denote the set of items, for each item $i \in \mathcal{I}$, we use v_i and t_i to denote its visual and textual modality respectively. We aim to align the common characteristics of v_i and t_i, while preserving their specific characteristics for the downstream recommendation task. For simplicity, we describe the algorithm for a single item, and drop the subscript of i in the notations to keep conciseness.

4 OUR APPROACH

In this section, we introduce our pretraining framework in detail. The architecture of PAMD is shown in Fig.2, which consists of two main modules: a disentangled encoder to decompose modalities to obtain their common and specific representations respectively, and a contrastive learning is further applied on modality-disentangled representations to generate a contrastive loss to guide the model optimization.

4.1 Disentangled Encoder

Given the modalities of an item, we first project them into a low dimensional space. Specifically, for visual modality, we use pretrained VGG19 [36] to derive the visual features and generate its primitive representation via a MLP layer, denoted as $\mathbf{e}_v$. As for textual modality, we utilize Glove [30] to obtain the word embeddings. Then these word embeddings are concatenated and fed into a MLP layer to derive the textual primitive representation, denoted as $\mathbf{e}_t$. Based on the representations $\mathbf{e}_v$ and $\mathbf{e}_t$, the proposed disentangled encoder uses the following function to extract their common characteristics as follows:

$$\mathbf{e}_t^c = MLP(\mathbf{e}_t; \Theta_t); \quad \mathbf{e}_v^c = MLP(\mathbf{e}_v; \Theta_v) \tag{1}$$

where Θ_t and Θ_v are all the parameters for the MLP layer respectively. Given the primitive and extracted common representations of each modality, we then utilize a subtraction operation to obtain the specific representations of both textual and visual modality as follows:

$$\mathbf{e}_t^s = \mathbf{e}_t - \mathbf{e}_t^c; \quad \mathbf{e}_v^s = \mathbf{e}_v - \mathbf{e}_v^c \tag{2}$$

According to Eq. 1 and Eq. 2, we decompose each modality to two distinct representations. To guarantee the resultant modality-common and -specific representations capture different aspects of the modalities, we further add constraints on the decomposed representations. Specifically, for modality-common representations, we minimize their discrepancy to force them align together, while for modality-specific characteristics, we apply an orthogonality constraint to force these modality-specific representations share none common characteristics. The loss of disentangled encoder is written as follows:

$$\mathcal{L}_{en} = ||\mathbf{e}_t^c - \mathbf{e}_v^c||^2 + ||\mathbf{e}_t^s \cdot \mathbf{e}_v^s||^2 \tag{3}$$

where Θ represents the learnable parameters (*i.e.,* Θ={Θ_t, Θ_v}).

4.2 Contrastive Learning

Recall that in disentangled encoder, though we force the disentangled representations containing different aspects of modalities, we cannot guarantee each representation containing the desired properties (*i.e.,* the common representation really keeps the common characteristics over modalities). Note that it is hard to build the supervision signals to guide the above decomposition process. Even feasible, any label for the modality-common and modality-specific features are expensive to obtain. Here, we introduce a contrastive objective to optimize our pertraining model. In detail, we firstly aim to perform the cross-modality alignment over these representations by taking the primitive representations $\mathbf{e}_v$ and $\mathbf{e}_t$ as anchor. Specifically, for modality-common representations, we aim to learn a representation mapping network across modalities as follows:

$$\hat{\mathbf{e}}_t^c = MLP(\mathbf{e}_v^c; \Phi_v); \quad \hat{\mathbf{e}}_v^c = MLP(\mathbf{e}_t^c; \Phi_t) \tag{4}$$

where $\hat{\mathbf{e}}_v^c$ and $\hat{\mathbf{e}}_t^c$ represent the transformed representations obtained by using $\mathbf{e}_t^c$ and $\mathbf{e}_v^c$ respectively, Φ_t and Φ_v are learnable parameters. This network is expected to precisely realize the representation mapping for modality-common features, which can only be achieved when both $\mathbf{e}_v$ and $\mathbf{e}_t$ contains no noisy information (*e.g.,* the modality-specific features). By minimizing cross-modality gap $\mathcal{L}_{de}^c$ as follows, we can ensure the correctness of extracting common characteristics to its maximum:

$$\mathcal{L}_{de}^c = ||\mathbf{e}_t - \hat{\mathbf{e}}_t^c||^2 + ||\mathbf{e}_v - \hat{\mathbf{e}}_v^c||^2 \tag{5}$$

However, in the downstream task, we find the recommendation performance is quite unstable. The reason still lies in the decomposing process of the disentangled encoder. As lack of supervised signals, PAMD has a large freedom to reach a local optimum for modality decomposition. For example, a wrong decomposition (*i.e.,*

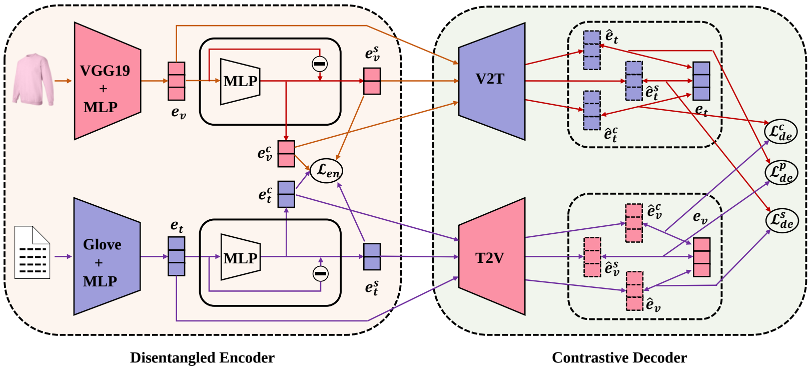


Figure 2: Over architecture of our proposed Pre-trAining Modality-Disentangled representations model. PAMD contains two components: a disentangled encoder to decompose modalities into common and specific representations, and a contrastive learning generates a contrastive loss to guide the model optimization.

assigning common features to the specific counterpart) will lose informative aspects that are shared across modalities (ref. Eq. 5).

Therefore, we introduce a contrastive loss by further feeding modality-primitive and -specific representations into optimization. Specifically, similar with Eq. 4, we feed primitive and specific modality representations to obtain their transformed representations as follows:

$$\hat{\mathbf{e}}_v = MLP(\mathbf{e}_t; \Phi_t); \quad \hat{\mathbf{e}}_t = MLP(\mathbf{e}_v; \Phi_v) \tag{6}$$

$$\hat{\mathbf{e}}_v^s = MLP(\mathbf{e}_t^s; \Phi_t); \quad \hat{\mathbf{e}}_t^s = MLP(\mathbf{e}_v^s; \Phi_v) \tag{7}$$

Based on these transformed representations, we then measure the gap between the original representation and the corresponding transferred representation as follows:

$$\mathcal{L}_{de}^p = ||\mathbf{e}_t - \hat{\mathbf{e}}_t||^2 + ||\mathbf{e}_v - \hat{\mathbf{e}}_v||^2 \tag{8}$$

$$\mathcal{L}_{de}^s = ||\mathbf{e}_t - \hat{\mathbf{e}}_t^s||^2 + ||\mathbf{e}_v - \hat{\mathbf{e}}_v^s||^2 \tag{9}$$

Note that parameters Φ_t and Φ_v are introduced solely to map the modality-common features across the two modalities. Since the gaps measured in $\mathcal{L}_{de}^c$, $\mathcal{L}_{de}^p$ and $\mathcal{L}_{de}^s$ all use the primitive representation $\mathbf{e}_t$ and $\mathbf{e}_v$ as the anchor. It is expected that $\mathcal{L}_{de}^c$ should be smaller than $\mathcal{L}_{de}^p$, since $\mathbf{e}_t^c$ and $\mathbf{e}_v^c$ are the more desired inputs for the mapping than $\mathbf{e}_t$ and $\mathbf{e}_v$ respectively. Similarly, since we expect $\mathbf{e}_t^s$ and $\mathbf{e}_v^s$ to contain no common characteristics, $\mathcal{L}_{de}^p$ should be smaller than $\mathcal{L}_{de}^s$. Hence, an inherent contrastive learning objective can be formulated as follows:

$$\mathcal{L}_{de} = \arg\min_{\Phi} \log\left(\sigma(\mathcal{L}_{de}^c - \mathcal{L}_{de}^p)\right) + \log\left(\sigma(\mathcal{L}_{de}^p - \mathcal{L}_{de}^s)\right) \tag{10}$$

where Φ represents all parameters used in the contrastive learning, $\Phi = \{\Phi_t, \Phi_v\}$.

4.3 Pretraining and Recommendation

Pretraining. Recall that we also have an encoder loss defined in Eq. 3, by incorporating it with the contrastive loss formulated in Eq. 10, the final loss objective is written as follows:

$$\mathcal{L}_{pre} = \mathcal{L}_{de} + \mathcal{L}_{en} \tag{11}$$

It is obvious to see that this loss objective is self-supervised and requires no explicit supervision. We can perform the pretraining with Eq. 11 on any dataset and plug the resultant modality-common and -specific representations into many recommendation models. We choose stochastic gradient descent (SGD) to minimize the loss function.

Recommendation. After the pertaining stage, each item is associated with four modality representations. To leverage these representations directly, we adopt pairwise ranking optimization framework (BPR) [33] as the downstream recommendation model. Specifically, we utilize an attention mechanism to calculate attention weights of each modality representation as follows:

$$[a_t^c, a_c^c, a_t^s, a^v] = softmax(\frac{\mathbf{e}[\mathbf{e}_t^c, \mathbf{e}_v^c, \mathbf{e}_t^s, \mathbf{e}_v^s]^T}{\sqrt{d}}) \tag{12}$$

where $\mathbf{e}$ is the learnable item representation, d represents the embedding dimension size. We then fuse all representations according to these weights to obtain a refined item representation:

$$\mathbf{e} = \mathbf{e} + a_t^c\mathbf{e}_t^c + a_v^c\mathbf{e}_v^c + a_t^s\mathbf{e}_t^s + a_v^s\mathbf{e}_v^s \tag{13}$$

Note that for BPR model learning, both the user embeddings and item embeddings (*i.e.*, $\mathbf{e}$) are initialized and updated via the Adam optimizer. The parameters Θ of PAMD are further fine-tuned during this recommendation learning phase. Given a user and an item, we compute the preference score with the inner product. We than rank the items according to their scores, and select the top-N results as the final recommendations.

5 EXPERIMENTS

In this section, we evaluate our proposed PAMD[1] over three real-world datasets. We firstly describe the experimental settings, and then discuss the results.

[1] https://github.com/hantengyue/PAMD

Table 1: Statistics of datasets for experiments.

Datasets	#Users	#Items	#Feedbacks
Clothing, Shoes & Jewelry	39,387	23,033	278,677
Yelp	10,457	8,937	90,301
Allrecipes	149,672	39,213	2,307,996

5.1 Dataset

We evaluate different recommendation algorithms over three datasets, including one e-commerce recommendation dataset, one review dataset, and one diet rating dataset.

- **Clothing, Shoes & Jewelry**[2] is a category of Amazon, which comprises a large corpora of item descriptions and images on more than $20,000$ items.
- **Yelp**[3] dataset is a subset of Yelp's businesses, reviews, and user data. Here, we use the data spanning from Jan 1, 2020 to Dec 1, 2020.
- **Allrecipes**[4] is a diet dataset crawled from Allrecipes.com, a recipe-sharing platform for western diet.

We perform the following preprocessing on the three datasets. Since our focus is to leverage multi-modal information for recommendation, we first remove items without textual information or images in each dataset. Then, we choose to retain the 5-core users and items such that each user or item is associated with at least 5 interactions, which is a common setting in the relevant literature [21]. The statistics of three datasets after preprocessing are reported in Table 1.

5.2 Evaluation Metrics

In order to present a comprehensive evaluation, we choose to apply 5-fold cross validation where the ratio of training, validation and testing is $3 : 1 : 1$. Following the recommended setting in [10], we choose to apply sampled metric for performance evaluation. Specifically, for each user, we randomly pair $1,000$ negative items, and rank these items with the ground-truth items. Two widely adopted metrics are used in the experiments: Recall@K and NDCG@K ($k = 5/10$). Here, Recall@K measures the percentage of target items appearing in the top-K results. NDCG@K further takes the ranking position in the top-K list into account, such that the higher rankings for the ground-truth items produce larger NDCG scores, *i.e.,* a better recommendation performance. We also perform statistical significance test by using the paired t-test. The differences are considered statistically significant when the p-value is lower than 0.05.

5.3 Baselines

To evaluate the effectiveness of our approach, we compare PAMD against the following competitive methods, including one traditional recommender, three modality-enhanced recommenders, and three multi-modal recommenders. For traditional recommnder, we consider BPR, which is also used as the built-in recommender in our proposed PAMD:

- **BPR** [33]: Bayesian personalized ranking algorithm with matrix factorization, which is a well-known item recommendation method.

For modailty-enhanced recommenders, we consider the enhanced versions of BPR, NFM, and NeuMF, which treat modality information as extra contextual features for recommendation:

- **BPR++**: This is an enhanced version of standard BPR model, which leverages both visual and textual modality information of items as its representations, and uses the matrix factorization (MF) framework to reconstruct the historical interactions between users and items. In the experiments, we use the concatenation of multi-modal features as the content information for recommendation.
- **NFM++**: The neural factorization machine (NFM) [15] is a deep architecture for effective feature interaction modeling. For fair comparison, we enhance original NFM by including both the image and description as contextual features.
- **NeuMF++**: By concatenating each item with its visual and textual representations, we adopt NeuMF [16] to model their complex interactions.

The multi-modal recommenders include:

- **VECF** [4]: VECF is a multi-modal attention network that jointly leverages image region-level features and the auxiliary textual information for fashion-aware recommendation.
- **Corr-AE** [9]: A cross-modal learning model utilizes correspondence autoencoder to incorporate both representation learning and correlation learning across different modalities. Here, Corre-AE is adopted to pretrain textural and visual representations, and apply BPR on them for recommendation.
- **MMGCN** [39] : Multi-modal Graph Convolution Network (MMGCN)] is a state-of-the-art multi-modal approach, which builds user-item bipartite graph for each modal, then uses GCN to train each bipartite graph. Then, MMGCN merges these multi-modal information for recommendation.

5.4 Parameter Settings

For all recommendation methods in comparison, we initialize the embedding vectors in the range of $(0, 1)$. To enable a fair comparison, we select the best learning rate for each method in the range of $\{0.0001, 0.001, 0.01, 0.05, 0.1\}$, and the dimension size is tuned in the range of $\{100, 150, 200, 250, 300, 350\}$. Moreover, we tune all parameters of baselines according to the validation set.

For the modality-enhanced models including BPR++, NFM++, and NeuMF++, we use VGG-19 to obtain the visual representations, and Glove for textual representations, which is the same as PAMD. Specifically, as to PAMD, the learning rate is set to 10^{-3}, the batch size is 100, and the embedding dimension is set to 150. In the recommendation phase, the learning rate is set to 10^{-4}.

5.5 Performance Comparison

The overall performance of our proposed PAMD and all the baselines are reported in Table 2. Here, we have the following observations:

[2] http://deepyeti.ucsd.edu/jianmo/amazon/index.html
[3] https://www.yelp.com/dataset
[4] https://www.kaggle.com/elisaxxygao/foodrecsysv1

Table 2: Performance comparison between baselines and our model PAMD for recommendation. The best performance of each column is highlighted in boldface. Symbol $*$ denotes the best baseline. Symbol ▲ denotes the relative improvement of our results against the best baseline, which are consistently significant at 0.05 level.

Dataset	Metrics	BPR	BPR++	NFM++	NeuMF++	VECF	Corr-AE	MMGCN	PAMD	▲(%)
Clothing, Shoes & Jewelry	Recall@5	0.0273	0.0601	0.0678	0.0731	0.0825	0.0847	0.0932*	**0.1066**	14.38
	Recall@10	0.0416	0.0927	0.1034	0.1101	0.1189	0.1186	0.1321*	**0.1451**	9.84
	NDCG@5	0.0212	0.0459	0.0520	0.0556	0.0575	0.0680	0.0732*	**0.0874**	19.40
	NDCG@10	0.0267	0.0580	0.0654	0.0695	0.0705	0.0805	0.0866*	**0.1017**	17.44
Yelp	Recall@5	0.0445	0.0633	0.0700	0.0763	0.0853	0.1077	0.1238*	**0.1341**	8.32
	Recall@10	0.0715	0.0852	0.1025	0.1055	0.1306	0.1522	0.1702*	**0.1895**	11.34
	NDCG@5	0.0345	0.0540	0.0561	0.0668	0.0764	0.0836	0.1127*	**0.1229**	9.05
	NDCG@10	0.0453	0.0631	0.0709	0.0763	0.1171	0.1181	0.1232*	**0.1442**	17.05
Allrecipes	Recall@5	0.1002	0.1244	0.1357	0.1402	0.1469	0.1488	0.1527*	**0.1637**	7.2
	Recall@10	0.1624	0.1891	0.2110	0.2181	0.2216	0.2274	0.2427*	**0.2541**	4.70
	NDCG@5	0.0860	0.1096	0.1178	0.1212	0.1358	0.1361	0.1391*	**0.1460**	4.96
	NDCG@10	0.1135	0.1382	0.1606	0.1652	0.1712	0.1754	0.1783*	**0.1895**	6.28

(1) By incorporating multi-modal information, we can see that BPR++ performs better than BPR. This verifies the advantage of integrating different modalities for recommendation. It is reasonable since the different modality information could generate more comprehensive representation for an item, which in turn enhances user preferences learning.

(2) We see that Corr-AE performs consistently better than VECF over all the three datasets. The reason might be that though both Corr-AE and VECF consider the alignment problem over visual and textual modalities, VCEF only utilizes the fine-grained visual characteristics. This treatment could inevitably introduce a significant loss of textual information, which could adversely affect the recommendation performance. Comparing with VCEF and Corr-AE, MMGCN utilizes a graph convolutional network to refine the high-order collaborative user-item interactions based on the modality information, and obtains a better performance.

(3) Finally, by leveraging both modality-common and -specific item representations, our proposed PAMD achieves significantly better recommendation performance across the three datasets in terms of both Recall and NDCG. Take the *Clothing, Shoes & Jewelry* dataset as an example, when compared with the best baseline (*i.e.,* MMGCN), the relative performance improvement obtained by PAMD is about 9.84% and 17.4% on Recall@10 and NDCG@10 respectively. It is worthwhile to highlight that our proposed PAMD performs recommendation with a simple BPR module. As we observed in Table 2, either BPR and BPR++ achieves much worse performance than the other recommendation models. The superior performance obtained by PAMD suggests the effectiveness of pretraining modality-disentangled item representations.

5.6 Ablation Study on PAMD

PAMD is a pretraining model aiming at learning modality-common and -specific item representations. These representations are further fine-tuned for the downstream task. In this section, we investigate the impact of different design choices via ablation study.

Table 3: Performance comparison of PAMD and its different variants on three datasets. The best performance is highlighted in boldface.

Dataset	Metrics	PAMD$_{\neg s}$	PAMD$_{\neg c}$	PAMD
Clothing, Shoes & Jewelry	Recall@10	0.1347	0.1371	**0.1451**
	NDCG@10	0.0945	0.0959	**0.1017**
Yelp	Recall@10	0.1789	0.1827	**0.1895**
	NDCG@10	0.1374	0.1398	**0.1442**
Allrecipes	Recall@10	0.2391	0.2465	**0.2541**
	NDCG@10	0.1749	0.1792	**0.1895**

Modality-Common *vs.* Modality-Specific Firstly, we analyze the recommendation performance when only utilizing modality-common and -specific representations respectively. Specifically, in the recommendation phase, we remove the modality-specific embeddings and only utilize modality-common representations, and we denote this variant as PAMD$_{\neg s}$. Similarly, we choose to keep the modality-specific representations instead for recommendation, we denote it as PAMD$_{\neg c}$. Table 3 reports the performance comparison of these models.

An interesting observation is that PAMD$_{\neg c}$ performs slightly better than PAMD$_{\neg s}$. The same observation is also made in the previous work [43]. Moreover, it is clear that PAMD obtains better performance than these two variants. This suggests that each modality indeed has its own specific characteristics that cannot be aligned with another modality. Those unaligned characteristics also play significant contribution towards a better recommendation. By considering both modality-common and -specific characteristics, PAMD can offer robust and comprehensive feature learning for better recommendation.

Contrastive Learning *vs.* Non-Contrastive Learning. Another key novelty of our proposed PAMD lies on the proposal of a contrastive loss to guide the pretraining process (Eq.9). Here, we aim to check whether it is more effective than the other options.

Table 4: Performance comparison of PAMD when applying different learning strategies. The best performance is highlighted in boldface.

Dataset	Metrics	PAMD-CON	PAMD
Clothing, Shoes & Jewelry	Recall@10	0.1255	**0.1451**
	NDCG@10	0.0881	**0.1017**
Yelp	Recall@10	0.1727	**0.1895**
	NDCG@10	0.1352	**0.1442**
Allrecipes	Recall@10	0.2418	**0.2541**
	NDCG@10	0.1790	**0.1895**

Specifically, we can update the objective function of PAMD as follows:

$$\mathcal{L} = \mathcal{L}_{en} + \mathcal{L}_{de}^{c} \tag{14}$$

According to Eq. 14, PAMD pretrains multi-modal representations without a contrastive strategy. We refer to this new setting as PAMD-CON. Table 4 shows the performance of these two different learning strategies.

It is obvious that PAMD consistently outperforms PAMD-CON on three datasets. This suggests that the proposed contrastive learning strategy is more effective: PAMD can well handle the alignment and separation over modalities, and obtains robust modality representations. In addition, we observe that PAMD also converges faster in the pretraining phase. We take *Clothing, Shoes & Jewelry* as an example, comparing with PAMD-CON, the time saved by PAMD to reach convergence is around 23.6%.

Disentangled Encoder *vs.* Contrastive Learning. PAMD utilizes two components to pretrain modality representations: a disentangled encoder and a contrastive learning. In this section, we further analyze the impact of these two components respectively.

Specifically, we choose only the loss function defined in Eq. 3 and Eq. 10 to train our model respectively. In this sense, PAMD directly degrades to a disentangled encoder and a contrastive learning respectively. We refer to these two settings as PAMD$_{en}$ and PAMD$_{de}$ respectively. Note that when only utilizing Eq. 10 for optimization, as there is no correlation constraint for both modality-common and -specific representations, PAMD$_{de}$ is similar to an autoencoder model [17] to perform transfer between visual and textual modality.

Table 5 reports the performance comparison among PAMD and these two variants. We can see that PAMD$_{de}$ performs better than PAMD$_{en}$. This observation demonstrates the necessity of modality alignment in the learning process. Though PAMD$_{en}$ preserves more characteristics by cross-modality matching for the common knowledge. However, without a cross-modality alignment process, the matching process of disentangled encoder would easily fail to extract the real common characteristics. These unaligned common features bring noises and in turn adversely affect the final recommendation performance. By combining these two components together, PAMD can well align the common characteristics while preserving the specific characteristics for each modality, and obtain the best performance.

Table 5: Performance comparison of PAMD, PAMD $_{en}$ and PAMD $_{de}$ on the three datasets. Best performance is highlighted in boldface.

Dataset	Metrics	PAMD$_{en}$	PAMD$_{de}$	PAMD
Clothing, Shoes & Jewelry	Recall@10	0.1155	0.1284	**0.1451**
	NDCG@10	0.0781	0.0842	**0.1017**
Yelp	Recall@10	0.1427	0.1658	**0.1895**
	NDCG@10	0.1252	0.1374	**0.1442**
Allrecipes	Recall@10	0.2268	0.2467	**0.2541**
	NDCG@10	0.1720	0.1819	**0.1895**

5.7 Analysis on Pretraining Stage

Considering PAMD consists of a pretraining stage and a recommendation stage. In this section, we aim to analyze the recommendation performance when utilizing different pretraining settings. For simplicity, we only give the results on *Allrecipes* dataset due to the page limitation, and the results on other two datasets are quite similar.

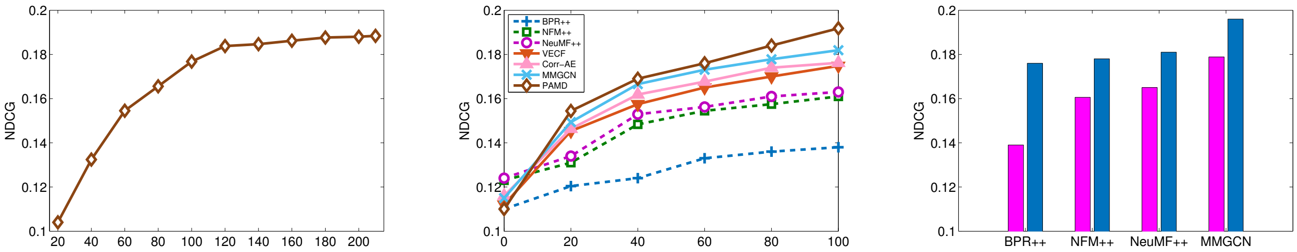
Recommendation Performance *w.r.t* the Amount of Pretraining Epoch. In the pretraining stage, PAMD can learn the enhanced representations from different modalities, the number of pretraining epochs thus is an important factor that affects the performance of the downstream recommendation task. To investigate this, we pretrain our model with a varying number of epochs and then finetune it on the recommendation task. Fig. 3(a) presents the results on *Allrecipes* dataset.

We can see that the recommendation performance benefits mostly from the first 100 epochs. And after that, the performance improves slightly. Based on this observation, we can conclude that according a combination learning over two encoders, PAMD can obtain high-quality modality-disentangled representations by the self-supervised approach within a small number of epochs. Considering our PAMD is a pretraining model aiming to learn robust modality-disentangled representations for items, the run-time of model training linearly depends on the number of items.

Recommendation Performance *w.r.t* the Amount of Pretraining Data. As the cold start problem is a common challenge that recommendation system suffered in real-world applications. In this section, we consider the impact of data sparsity to our pretraining stage. Specifically, we simulate the data sparsity scenario by randomly masking different proportion of modality information from datasets, i.e., 20%, 40%, 60%, 80%, and 100%. Fig. 3(b) shows the recommendation performance on Allrecipes dataset.

(1) We see that when masking all modality information, the modality-enhanced approaches perform better than multi-modal approaches. This is conventional as all multi-modal approaches will degrade to common pair-wise ranking models when no modality information is included, while the modality-enhanced approaches can still benefit from modeling complex interactions over user and items.

(2) The performance of all models substantially increases when feeding more modality data. However, for modality-enhanced recommenders, the improvements turn to be not significant when considering more modality data. This observation demonstrates



(a) Different numbers of pretraining epochs.

(b) Different sparsity levels.

(c) Different models enhanced by PAMD.

Figure 3: Performance comparison of diffrent number of pretraining epochs (a), different sparsity levels (b), and the existing models enhanced by PAMD (c) on *Allrecipes* in terms of NDCG@10.

Table 6: Distribution of different representation types on three datasets.

Dataset	Common-feature	Specific-feature
Clothing, Shoes & Jewelry	47.85%	52.15%
Yelp	47.28%	52.72%
Allrecipes	46.18%	53.82%

our assumption: It is non-trivial to perform multi-modal information fusion to enhance the recommendation performance. While for the multi-modal recommenders, according to the attention or cross-modality tricks, then can still enjoy a better improvement when fusing more modality data.

(3) Finally, PAMD consistently performs better than baselines in most cases, this observation implies that PAMD is able to make better use of the data according to the designed matching technique, which can well alleviate the influence of data sparsity problem for recommendation to some extent.

5.8 Adapting PAMD to Other Recommendation Models

In PAMD, the modality-common and -specific representations can be learned in the pretraining phase, these high-quality representations can also be plugged into other recommendation models. In this section, we conduct experiments to check whether PAMD can bring further improvements to other models.

Specifically, based on the well pretrained modality representations on *Allrecipes* dataset, we directly apply them on BPR++, NFM++, NeuMF++, and MMGCN, by concatenating these four different representations together. We do not consider BPR and VCEF as when feeding the pretrained representations to BPR, BPR is the same with BPR++, while for VCEF, its architecture does not support the pretraining objectives. Note that we also fine-tune the hidden dimension size over the validation set to enable the fair comparison.

The results of NDCG@10 on *Allrecipes* dataset are illustrated in Figure 3(c). We can see that after pretraining by PAMD, all the baselines can achieve better performance. It demonstrates the effectiveness of our pertaining strategy, which can learn robust modality-disentangled item representations.

5.9 Visualizing Attention

In this section, we analyze the impact of different types of modality representations over each dataset, so as to understand which type of disentangled representations are more useful in the recommendation stage.

Specifically, for each user-item pair that PAMD recommends correctly in the testing stage, we identify the type of representations (the specific and common type) with the highest-attention score. We then analyze the distribution of the two types. For example, if there are 10 correctly recommended pairs, the modality-specific representation was learned as the highest attention score for 3 times, then the percentage of specific type will be 0.3. This distribution on three datasets is reported in Table 6.

As we can see, the modality-specific information plays a relatively more important role on the three datasets. This observation is interesting and consistent with the previous experiments (in Section **Modality-Common *vs.* Modality-Specific**). This further demonstrates the importance of preserving both the modality-common and -specific representations for better recommendation.

6 CONCLUSION

In this work, we propose a pretraining framework (named PAMD), which aims to learn modality-common and modality-specific representations for recommendation. Specifically, PAMD contains a disentangled encoder and a contrastive learning. The disentangled encoder aims to automatically extract their modality-common characteristics while preserving their modality-specific characteristics. The contrastive learning aims to guarantee the consistence and gaps between modality-disentangled representations instead.

Both modality-common and modality-specific have their own contributions to the improvement of recommendation performance. We hope this work can provide a new perspective on multi-modal representation learning for recommendation. To the best of our knowledge, this is the first pretraining framework to learn modality-disentangled representations in recommendation scenarios.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No.61802029, 61872278, U1536121, 61370195).

REFERENCES

[1] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 2 (2019), 423–443.

[2] Yang Bao, Hui Fang, and Jie Zhang. 2014. TopicMF: Simultaneously Exploiting Ratings and Reviews for Recommendation. In *AAAI 2014*. AAAI Press, 2–8.

[3] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive Collaborative Filtering: Multimedia Recommendation with Item- and Component-Level Attention. In *SIGIR 2017*. ACM, 335–344.

[4] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized Fashion Recommendation with Visual Explanations based on Multimodal Attention Network: Towards Visually Explainable Recommendation. In *SIGIR*. ACM, 765–774.

[5] Zhiyong Cheng, Ying Ding, Lei Zhu, and Mohan S. Kankanhalli. 2018. Aspect-Aware Latent Factor Model: Rating Prediction with Ratings and Reviews. In *WWW 2018*. ACM, 639–648.

[6] Qiang Cui, Shu Wu, Qiang Liu, Wen Zhong, and Liang Wang. 2020. MV-RNN: A Multi-View Recurrent Neural Network for Sequential Recommendation. *IEEE Trans. Knowl. Data Eng.* 32, 2 (2020), 317–331.

[7] George E. Dahl, Dong Yu, Li Deng, and Alex Acero. 2012. Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. *IEEE Trans. Speech Audio Process.* 20, 1 (2012), 30–42.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*. Association for Computational Linguistics, 4171–4186.

[9] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. 2014. Cross-modal Retrieval with Correspondence Autoencoder. In *ACM Multimedia*. ACM, 7–16.

[10] Guibing Guo, Shichang Ouyang, Xiaodong He, Fajie Yuan, and Xiaohua Liu. 2019. Dynamic Item Block and Prediction Enhancing Block for Sequential Recommendation. In *IJCAI*. 1373–1379.

[11] Ruining He, Chunbin Lin, Jianguo Wang, and Julian J. McAuley. 2016. Sherlock: Sparse Hierarchical Embeddings for Visually-Aware One-Class Collaborative Filtering. In *IJCAI*. IJCAI/AAAI Press, 3740–3746.

[12] Ruining He and Julian J. McAuley. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In *WWW 2016*. 507–517.

[13] Ruining He and Julian J. McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *AAAI 2016*. AAAI Press, 144–150.

[14] Xiangnan He, Tao Chen, Min-Yen Kan, and Xiao Chen. 2015. TriRank: Review-aware Explainable Recommendation by Modeling Aspects. In *CIKM 2015*. ACM, 1661–1670.

[15] Xiangnan He and Tat-Seng Chua. 2017. Neural Factorization Machines for Sparse Predictive Analytics. In *SIGIR*. ACM, 355–364.

[16] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *WWW*. ACM, 173–182.

[17] G. E. Hinton and R. R. Salakhutdinov. 2006. Reducing the Dimensionality of Data with Neural Networks. *Science* 313 (2006).

[18] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. 2018. Multimodal Unsupervised Image-to-Image Translation. In *ECCV (3) (Lecture Notes in Computer Science, Vol. 11207)*. Springer, 179–196.

[19] Yu-Gang Jiang, Zuxuan Wu, Jun Wang, Xiangyang Xue, and Shih-Fu Chang. 2018. Exploiting Feature and Class Relationships in Video Categorization with Regularized Deep Neural Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 2 (2018), 352–364.

[20] Wang-Cheng Kang, Chen Fang, Zhaowen Wang, and Julian J. McAuley. 2017. Visually-Aware Fashion Recommendation and Design with Generative Image Models. In *ICDM 2017*. IEEE Computer Society, 207–216.

[21] Wang-Cheng Kang and Julian J. McAuley. 2018. Self-Attentive Sequential Recommendation. In *ICDM 2018*. 197–206.

[22] Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining Language and Vision with a Multimodal Skip-gram Model. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*. 153–163.

[23] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*. Association for Computational Linguistics, 7871–7880.

[24] Chenliang Li, Xichuan Niu, Xiangyang Luo, Zhenzhong Chen, and Cong Quan. 2019. A Review-Driven Neural Model for Sequential Recommendation. In *IJCAI 2019*. 2866–2872.

[25] Hao Liu, Jindong Han, Yanjie Fu, Jingbo Zhou, Xinjiang Lu, and Hui Xiong. 2020. Multi-Modal Transportation Recommendation with Unified Route Representation Learning. *Proc. VLDB Endow.* 14, 3 (2020), 342–350.

[26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019).

[27] Zhiwei Liu, Ziwei Fan, Yu Wang, and Philip S. Yu. 2021. Augmenting Sequential Recommendation with Pseudo-Prior Items via Reversely Pre-training Transformer. In *SIGIR*. ACM, 1608–1612.

[28] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal Deep Learning. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*. 689–696.

[29] Aditya Pal, Chantat Eksombatchai, Yitong Zhou, Bo Zhao, Charles Rosenberg, and Jure Leskovec. 2020. PinnerSage: Multi-Modal User Embedding Framework for Recommendations at Pinterest. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*. 2311–2320.

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

[31] Jinwei Qi and Yuxin Peng. 2018. Cross-modal Bidirectional Translation via Reinforcement Learning. In *IJCAI*. ijcai.org, 2630–2636.

[32] Zhaopeng Qiu, Xian Wu, Jingyue Gao, and Wei Fan. 2021. U-BERT: Pre-training User Representations for Improved Recommendation. In *AAAI*. AAAI Press, 4320–4327.

[33] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI*. AUAI Press, 452–461.

[34] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized Markov chains for next-basket recommendation. In *WWW 2010*. ACM, 811–820.

[35] Mert Bülent Sariyildiz, Julien Perez, and Diane Larlus. 2020. Learning Visual Representations with Caption Annotations. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VIII*. 153–170.

[36] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

[37] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. In *CIKM*. ACM, 1441–1450.

[38] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP/IJCNLP (1)*. Association for Computational Linguistics, 5099–5110.

[39] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video. In *ACM Multimedia*. ACM, 1437–1445.

[40] Shuang Wu, Sravanthi Bondugula, Florian Luisier, Xiaodan Zhuang, and Pradeep Natarajan. 2014. Zero-Shot Event Detection Using Multi-modal Fusion of Weakly Supervised Concepts. In *CVPR*. IEEE Computer Society, 2665–2672.

[41] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML (JMLR Workshop and Conference Proceedings, Vol. 37)*. JMLR.org, 2048–2057.

[42] Fei Yan and Krystian Mikolajczyk. 2015. Deep correlation for matching images and text. In *CVPR*. IEEE Computer Society, 3441–3450.

[43] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis. In *AAAI*. AAAI Press, 10790–10797.

[44] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. 2021. Multimodal Contrastive Training for Visual Representation Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. 6995–7004.

[45] Zheni Zeng, Chaojun Xiao, Yuan Yao, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. 2021. Knowledge Transfer via Pre-training for Recommendation: A Review and Prospect. *Frontiers Big Data* 4 (2021), 602071.

[46] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-Rec: Self-Supervised Learning for Sequential Recommendation with Mutual Information Maximization. In *CIKM*. ACM, 1893–1902.