From Abstract to Details: A Generative Multimodal Fusion Framework for Recommendation

Fangxiong Xiao*
xiaofangxiong20@mails.ucas.edu.cn
Search and Recommendation Platform Department, JD.COM

Lixi Deng*
denglixi@jd.com
Search and Recommendation Platform Department, JD.COM

Jingjing Chen†
chenjingjing@fudan.edu.cn
School of Computer Science, Fudan University

Houye Ji
Search and Recommendation Platform Department, JD.COM

Xiaorui Yang
Search and Recommendation Platform Department, JD.COM

Zhuoye Ding
Search and Recommendation Platform Department, JD.COM

Bo Long
Search and Recommendation Platform Department, JD.COM

ABSTRACT

In E-commerce recommendation, Click-Through Rate (CTR) prediction has been extensively studied in both academia and industry to enhance user experience and platform benefits. At present, most popular CTR prediction methods are concatenation-based models that represent items by simply merging multiple heterogeneous features including ID, visual, and text features into a large vector. As these heterogeneous modalities have moderately different properties, directly concatenating them without mining the correlation and reducing the redundancy are unlikely to achieve the optimal fusion results. Besides, these concatenation-based models treat all modalities equally for each user and overlook the fact that users tend to pay unequal attention to information of various modalities when browsing items in the real scenario. To address the above issues, this paper proposes a generative multimodal fusion framework (GMMF) for CTR prediction task. To eliminate the redundancy and strength the complementary of multimodal features, GMMF generates the new visual and text representations by a Difference-Set network (DSN). These representations are non-overlapping with the information conveyed by ID embedding. Specifically, DSN maps ID embedding into visual and text modalities and depicts the difference between multiple modalities based on their properties. Besides, GMMF learns unequal weights to multiple modalities with a Modal-Interest network (MIN) modeling users' preference on heterogeneous modalities. These weights reflect the usual habits and hobbies of users. Finally, We conduct extensive experiments on both public and collected industrial datasets, and the results show that GMMF greatly improves performance and achieves state-of-the-art performance.

CCS CONCEPTS

• **Information systems → Social recommendation**.

KEYWORDS

multimodal, generative learning, recommender systems

ACM Reference Format:
Fangxiong Xiao, Lixi Deng, Jingjing Chen, Houye Ji, Xiaorui Yang, Zhuoye Ding, Bo Long. 2022. From Abstract to Details: A Generative Multimodal Fusion Framework for Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3548366

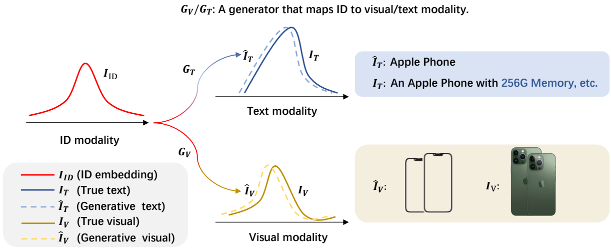

Figure 1: Recommender systems contain heterogeneous modalities: ID, visual, and text. ID embedding I_{ID} represents abstract information while visual I_V and text I_T contain detailed information. To better extract the complementary information between them, I_{ID} is firstly mapped into visual and text spaces marked as $\hat{I}_V$, $\hat{I}_T$, respectively.

1 INTRODUCTION

In recent years, B2C e-commerce platforms, such as JD.COM, Amazon and Alibaba are developing rapidly. To enhance user experience and platform benefits, clickthrough rate (CTR) prediction has been

[*] Work done at JD.com. The first two authors make equal contribution to this work.
[†] To whom correspondence should be addressed.

MM '22, October 10–14, 2022, Lisboa, Portugal
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548366

extensively studied in both academia and industry [6, 11, 19, 23, 30, 47, 48]. Most popular recommender systems such as YoutubeNet, DIN, and DIEN, learn ID embedding that is related to the interaction between users with the items [6, 47, 48]. However, as multimodal information is becoming increasingly vital in recommender systems, these ID embeddings only provide abstract information and lack the expression of multimodal information. As a result, more recent works [18, 25] try to improve the CTR prediction accuracy by utilizing multimodal information. It has been demonstrated in [19], representing items with multimodal features is effective in improving the recommendation accuracy. Therefore, we make efforts to explore the complementarity of multimodal information and improve CTR prediction of multimodal recommender systems in this paper.

It is a dilemma how to leverage multimodal information from different sources for learning compact and discriminative multimodal representations. A simple and straightforward way is to merge features from different modalities, e.g., text and image, into a large vector to form the multimodal representations for recommendation [1, 40]. However, this method is likely to strengthen the redundancy and weaken the complementarity between multiple modalities. To efficiently fuse the multimodal information, numerous studies have been presented. They are broadly divided into two paradigms: attention-based methods [7, 12, 17, 18, 20, 25, 43, 44] and out-product methods [34, 35]. Attention mechanism is proposed to emphasize similarity and capture dependencies between sequences of unimodal entities in Natural Language Processing tasks [2]. It is effective in describing the correlation and connection between different entities. Unlike attention, out-product methods establish the connection between different dimensions of multiple modalities to obtain second-order interaction features [34, 35]. It characterizes more fine-grained properties and is beneficial for mining potential association rules in data. However, two factors are not considered in the aforementioned approaches. First, the different representative level between ID embeddings and visual/text features is not taken into account. To be specific, ID embeddings and visual/text features have different properties and semantic meanings. The former are abstract representations learned through the CTR task while the latter, which are extracted from pre-trained models, are specific and contain details. Both attention and out-product based methods treat abstract and detailed features as equal, resulting in the complementary details being easily overwhelmed by abstract information. Second, these works do not explore the interest of users in different modalities during fusion and overlook the fact that users show inequitable preference due to their habits or other factors.

To address the above issues, we purpose a two-level network termed Generative Multimodal Fusion Framework (GMMF). GMMF consists of two novel modules – Difference-Set network (DSN) and Modal-Interest network (MIN). DSN aims to eliminate the redundancy of multimodal information and is placed in the bottom of our GMMF framework. Meanwhile, MIN is proposed to model the modality specific user's preference for CTR prediction and is placed in the top of the framework. Based on generative adversarial networks, DSN firstly projects the ID embedding vector into sub-space of visual and text modalities. It not only generates a new vector that simplifies the interaction between ID embeddings and other multimodal features, but also preserves the semantics consistency among the generated and the original embeddings. Then redundant information between ID embedding and visual/text features are explicitly eliminated inspired by the set theory, leading the details are extracted for improving the performance of recommender system. Besides, to establish the users' preference for modalities, MIN firstly integrates historical behavior as the modality-specific embedding for each modality. Then MIN calculates the attention weights for each modality to control the intake of information.

The contributions of this paper are summarized as follows:

- We propose a novel Generative Multimodal Fusion Framework (GMMF) for multimodal recommendation. It improves the recommendation accuracy and robustness by utilizing multimodal information from both items and users. Extensive experiments show that the GMMF achieves the state-of-the-art performances on both public and industrial datasets.
- We explore the interactions between ID embeddings, visual, and text features and propose a generative multimodal representation learning module named as DSN. By considering the different properties of ID and visual/text modalities, our method can eliminate the redundancy in multimodal information for forming compact and discriminative multimodal representations.
- We investigate user preference in specific modalities and propose a MIN module. It captures users' diverse interest of content in different modalities through attention mechanism.

2 RELATED WORK

2.1 Multimodal Learning

Currently, multimodal learning has made significant progress in speech recognition, emotion classification, video retrieval, etc [8, 28, 32, 45]. The key problem in multimodal learning is how to fuse multimodal information and learn discriminative multimodal representation, which has attracted the numerous researcher attentions [3]. To obtain good representations with the properties of smoothness, space-time invariance, sparseness, and natural clustering [4], several deep-based multi-modal representation learning methods have been proposed [16, 39]. For example, Liu et al. [21] and Pu et al. [29] introduce category or topic information to mine and update existing pre-trained features, and finally generate natural clustering representation. Li et al. [19, 44, 45] learn modality representation including Modal-Special embedding and Modal-Invariant embedding rather than unique forms. In short, these works take attention mechanism as the basic unit to mine the interactive information between multiple modalities [18, 25, 43]. Another small part of the works fuse multiple modalities according to the matrix principle [34, 35]. They take mathematical reasoning as the main rule and design several interactive principles. Then they regard the interactive features as images and use convolutional network to learn deep semantics. However, these methods ignore that the interactive features have no similar spatial properties. Wei et al. [41] design a graph network to explore the users' preference from sub-graph for each modality but ignore the fusion between multiple modalities.

In practical, the methods mentioned ignore the fact that multiple modalities have different properties and need to be dealt with differently. Therefore, we propose a framework to mine complementary information from visual and text with ID information, which distinguishes the abstract and details.

2.2 Cross-modal Learning

Variational autoencoder (VAE) and generative adversarial nets (GAN) are applied to capture the data distribution for cross-modal tasks [10, 15]. The GAN restricts the generated distribution by modifying the discriminator while VAE approximates the original distribution in the process of self-supervised learning. Mirza et al. [24] add an extra input in GAN to get a distribution under certain conditions, such as getting a picture of a cat by adding the word "cat" to the input. Currently, Pandeva et al. [26] propose a novel GAN extension for multimodal distribution learning to find a disconnected data representation in the latent space. Peng et al. [27] model the joint distribution over the data of different modalities to correlate such heterogeneous data. Zhu et al. [49] propose the R2GAN with multiple loss functions to restrict the representation of intermediate latent variables.

These methods prefer to project modality features into a common latent space, which do not explicitly limit the consistency of cross-modal semantics. Different to these works, we design a special discriminator to detect whether the generated information keeps same semantically consistent with the input.

2.3 Recommender Systems

In recent years, deep-based methods have been proposed [47, 48] increasingly in recommender systems, and they are divided into Attention, RNN, CNN and GNN models. Among them, the attention-based approaches has become the basis for most academic or industrial research. Deep interest network [48] weights the attention of the historical behavior sequence, and better captures the diverse characteristic of user interests. Behavior Sequence Transformer [5] captures the relationship between each item in the behavior sequence. Comparing to other methods, attention-based recommender systems have good performance in both speed and effect, and go into the industry applications.

In general, these methods directly merge all the features together without comprehensively mining multimodal information. Considering the importance of modalities in both users and items, we extract complementary information from items and design a mechanism that reflects user selectivity for modalities.

3 THE PROPOSED APPROACH

The CTR task is aimed to predict the probability $\mathbb{P}(click|user, item)$ whether the user clicks on the given item. Specifically, given a user u, his/her historical behavior sequence $u_h = \{x_1, x_2, x_3, \ldots, x_{N-1}, x_N\}$, and the recommended item x_{N+1}, the recommender systems will calculate the probability that u clicks on x_{N+1}. Generally, the item x and user u are represented by unique ID and recommender systems learn embeddings for them. But in our multimodal recommendation task, the item x is represented by a tuple (ID, V, T), where V denotes the visual features and T denotes text features. For the convenience of representation, we denote the ID modality as E. Considering heterogeneous modalities have moderately different properties, we propose a generative multimodal framework to fuse these modalities. As described in Figure 2, the GMMF is mainly composed of two parts: Difference-Set and Modal-Interest modules. The DSN includes Auto-Encoder, Cross-modal Generative Adversarial Net (CGAN), and Auto-Difference. It projects abstract ID embedding into visual and text modalities, and extracts effective differences for capturing the complementary information. After obtaining differentiated multimodal information, the MIN models the users’ preference for various modalities.

3.1 Multi-Modal Embedding Initialization

Next, we describe the user representation, item representation, and the way to process for them in our task.

ID. For unique user and item IDs, we first map them to the one-hot vectors. Then we extract the corresponding feature vectors from the embedding matrix according to these one-hot vectors.

Visual. Following Li et al. [19], the 4096-dimensional visual feature vectors are extracted by the VGG16 [14, 31] model that drops the last two layers for classification tasks.

Text. Bidirectional Encoder Representations from Transformers (BERT) learns representations from unlabeled text and works great on downstream tasks[13, 33]. Therefore, we input text into the BERT model, then obtain a 768-dimensional feature vector with the comprehensive information at the position of "CLS".

3.2 Difference-Set Network.

The DSN is designed to eliminate redundancy and strengthen complementarity of multimodal information. It consists of 3 submodules: Auto-Encoder, CGAN, and Auto-Difference modules. First, Auto-Encoder is adopted to reduce the huge dimension gap of multimodal features, while keeping their original information. Then, the CGAN module projects the ID embedding vector into the sub-space of the visual and text modalities through adversarial learning. Finally, the Auto-Difference network calculates the similarity weights between original and generated features and then subtracts redundant information from the weights.

Auto-Encoder reduces the dimension of original visual and text features. The unreduced high-dimensional multimodal features will overwhelm other features because of the huge dimensional gap [46]. Moreover, high-dimensional multimodal features increase computational complexity and reduce the time efficiency of the recommender systems. To reduce the dimension of features, linear dimensionality reduction methods such as Principal Component Analysis and Canonical Correlation Analysis are proposed, but they rely heavily on the distribution of data [37]. In our scenario, the distribution of visual and text is extremely complex, and linear dimensionality reduction will cause considerable errors. Therefore, we adopt Auto-Encoder [36] to reduce the dimension of visual and text features while preserving the original information maximally. Formally, the Auto-Encoder takes the true modality features I_m, as input of each modality $m \in \{V, T\}$, and returns the low-dimensional vectors H_m. The encoder is defined as:

$$H_m = f_m(I_m), \tag{1}$$

where f_m is nonlinear projection layers and reduces the dimension of the input from d_I^m to d_H^m for modality m. The input I_m is the features of the modality m extracted by pre-trained models. And

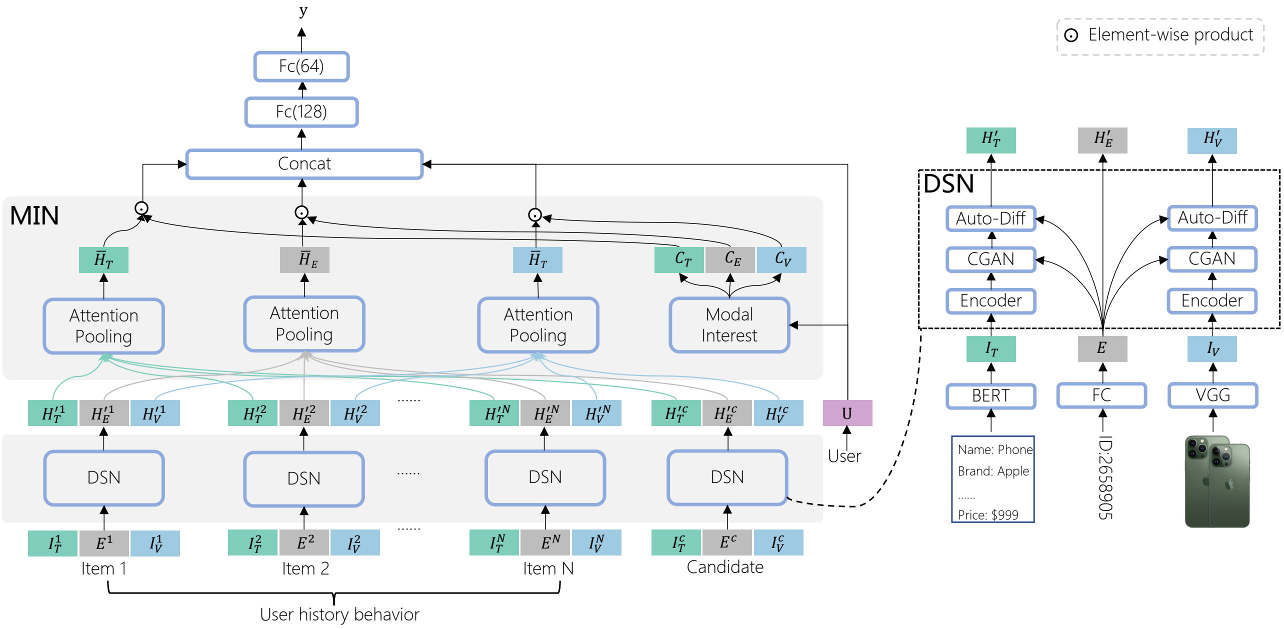

Figure 2: The architecture of GMMF. We first input the features of items into the Difference-Set network to mine the details and complementary information. Next, the Modal-Interest network synthesizes information in different modalities and interacts with user preferences on modalities. Last, all features are merged and sent into the downstream network for CTR prediction.

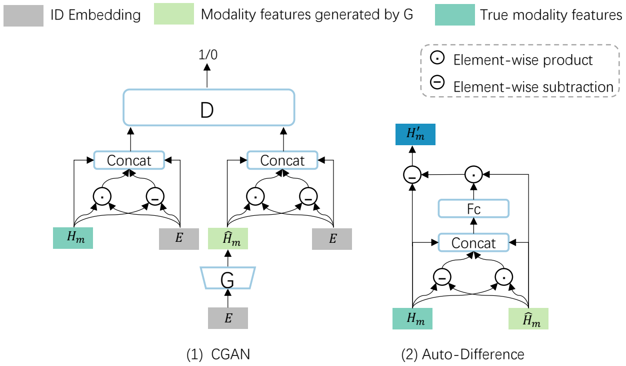

Figure 3: The left part is the CGAN and the right part is Auto-Difference.

H_m is the low-dimensional representation of modality m. Similar to the encoder, the decoder is defined as follows:

$$\hat{I}_m = \phi_m(H_m), \tag{2}$$

where ϕ_m is nonlinear projection layers and raises the dimension of the input from d^m_H to d^m_I. And $\hat{I}_m$ is the reconstructed representation of modality m.

CGAN is the core of Cross-modal in our method, which transform heterogeneous modalities into homologous modalities. Despite dimensionality reduction scales down the complexity of interaction among ID embedding and visual/text features, they are still in different modal spaces. Directly comparing the differences among them would hurt the integrity of multimodal information. As ID embedding represents the abstract information and visual/text features represent specific information, the specific information will be submerged in abstract information and converged to abstract information without any constraint. Thus, we propose the CGAN that maps the ID embedding into the sub-space of specific features. In detail, the CGAN consists of a group of generators $\{\mathcal{G}_t, \mathcal{G}_v\}$ and discriminators $\{\mathcal{D}_t, \mathcal{D}_v\}$. These generators map the ID embedding to text and visual modalities spaces respectively. Specifically, the generator $\mathcal{G}_m$ receives ID embedding E as the prior information that controls the distribution and returns a vector $\hat{H}_m$. It is calculated as follows,

$$\hat{H}_m = \mathcal{G}_m(E), \tag{3}$$

where E is the ID embedding and $\mathcal{G}_m$ is multi-layer nonlinear networks, which projects E into modal space of m. $\hat{H}_m$ represents the cross-modal features corresponding to E for modality m. In addition, the corresponding discriminator $\mathcal{D}_m$ determines the semantics of generated $\hat{H}_m$. Not only do we want the generated $\hat{H}_m$ to be in the cross-modal space, but we also want the generated data to be semantically related to the original ID embedding. Therefore, we use $\{E, H_m\}$ as positive samples, compared to the original generator that uses H_m as positive samples and only force to generate data similar to H_m. And the essence of the discriminator is a binary classifier:

$$\hat{y}_m = \mathcal{D}_m(E, H_m/\hat{H}_m), \tag{4}$$

where $\mathcal{D}_m$ is a multi-layer nonlinear network with a *sigmoid* activation function as output. The expected output of $\mathcal{D}_m$ is 1 if the input is <E, H_m>, and 0 if the input is <$E, \hat{H}_m$>. Finally, the whole

CGAN is optimizing a min-max game as:

$$\min_{\mathcal{G}_m}\max_{\mathcal{D}_m} \mathcal{L}^m_{GAN}, \tag{5}$$

where L^m_{GAN} is obtained as:

$$\mathcal{L}^m_{GAN} = \mathbb{E}[\log \mathcal{D}_m(E, H_m)] + \mathbb{E}[1 - \log \mathcal{D}_m(E, \mathcal{G}_m(E))], \tag{6}$$

Optimizing L^m_{GAN} is equivalent to optimizing Jensen-Shannon divergence. For fixed $\mathcal{G}_m$, $\mathcal{D}_m$ has an optimal solution is [10]:

$$\mathcal{D}^*_m = \frac{p(E, H_m)}{p(E, H_m) + p(E, \mathcal{G}_m(E))}, \tag{7}$$

where p is a probability distribution. In the case of $p(E, H_m) = p(E, \mathcal{G}_m(E))$, $\mathcal{L}^m_{GAN}$ arrivals the global optimal goal.

Auto-Difference automatically learns the difference rule among vectors $\hat{H}_m$ generated by CGAN and the original multimodal features. Inspired by the difference set, we aim to take true modality features minus the intersection of true modality and cross-modal features as complementary information. As the traditional difference methods are according to certain rules, they are lack of diversity and difficult to design. Therefore, we design a network named Auto-Difference to automatically learn difference rule. Given the original modality features H_m and generated modality features $\hat{H}_m$, the Auto-Difference network subtracts redundant information in $\hat{H}_m$ from H_m as follows:

$$H'_m = \begin{cases} H_m - W^s_m \cdot \hat{H}_m, m \in \{V, T\}, \\ E, m \in \{E\}. \end{cases} \tag{8}$$

where W^s_m is the similar weights of $\hat{H}_m$ and H_m. Specifically, W^s_m is obtained by the function τ as follows:

$$\begin{aligned} W^s_m &= \tau(H_m, \hat{H}_m) \\ &= W_m \times Concat[H_m, \hat{H}_m, H_m \odot \hat{H}_m, H_m - \hat{H}_m], \end{aligned} \tag{9}$$

where W_m are parameters learned by a fully connect layer and $\odot$ represents element-wise product operation. It is worth mentioning $H_m \odot \hat{H}_m$ and $H_m - \hat{H}_m$ are also input into the network to enhance additional information while the former measures their similarity and the latter measuring their degree of difference.

3.3 Modal-Interest Network

After processing by the above modules, the representations of each item in the network becomes <H'_E, H'_V, H'_T>. The representations removes redundancy and mines relevant details in visual and text modalities. The current mainstream methods directly concatenate these representations together and deliver them to the downstream network but ignore that users pay different attention to different modalities. To capture user interest in modalities, we first generate the modality-special embedding from the user behaviors and candidate item c. With attention-pooling mechanism, three modality-special embeddings are generated as follows:

$$\bar{H}_m = \sum_{i=1}^{N} Relu\left(\frac{H'^c_m H'^i_m}{\sum_{j=1}^{N} H'^c_m H'^j_m}\right) H'^i_m, m \in \{E, V, T\}. \tag{10}$$

where c is the mark of the candidate item and N is the length of historical behavior. Besides, we design a gate for each modality to control how much modality information flows in. Assuming user interest in a modality is across the modality space, we quantify them to vectors the same dimension as modal-special features as follows:

$$C_m = Softmax(\rho_m(U)), \tag{11}$$

where U is user ID embedding and ρ_m is a fully connect network. Finally, we merge all modality-special features with user interest and calculate prediction result:

$$y_{CTR} = sigmoid(W \times Concat[C_E \odot \bar{H}_E, C_V \odot \bar{H}_V, C_T \odot \bar{H}_T, U]), \tag{12}$$

where W is parameters and y_{CTR} is the prediction probability that the user clicks on the given item.

3.4 Loss Function

The overall loss function includes reconstruction loss $\mathcal{L}_r$, CTR loss $\mathcal{L}_{CTR}$, and CGAN loss $\mathcal{L}_{GAN}$. First, the reconstruction loss has made strict constraints on the dimension reduction of visual and text information to ensure that the original information is not damaged as much as possible. We adopt the mean squared error loss [9] to measure the Auto-Encoder approximation error. To adjust the strictness of the constraints, we can adjust the coefficient λ_m in $\mathcal{L}_r$. In equation 14, this criterion minimizes the reconstruction loss $\mathcal{L}_r$ by measuring the difference between $\hat{I}_m$ and I_m. Second, the CTR loss $\mathcal{L}_{CTR}$ is our final optimization goal, cross-entropy loss [9] is used to improve prediction accuracy. Since our task is a binary classification task, $\mathcal{L}_{CTR}$ can be further written as binary-cross entropy loss as equation 15. These two losses constitute the loss of the backbone prediction network as follows:

$$\mathcal{L}_{pred} = \mathcal{L}_r + \mathcal{L}_{CTR}, \tag{13}$$

$$\mathcal{L}_r = \sum_{m \in V,T} \lambda_m MSELoss(\hat{I}_m, I_m), \tag{14}$$

$$\mathcal{L}_{CTR} = BCELoss(y_{CTR}, y), \tag{15}$$

where y is the true label and y_{CTR} is predicted by the model. Besides, the $\mathcal{L}_{GAN}$ is a set of losses including $\mathcal{L}_{\mathcal{G}_m}$ and $\mathcal{L}_{\mathcal{D}_m}$. We iteratively optimize $\mathcal{L}_{\mathcal{G}_m}$ and $\mathcal{L}_{\mathcal{D}_m}$. Both of them optimize the parameters of the network through binary-cross entropy loss:

$$\mathcal{L}_{\mathcal{G}_m} = \sum_{m \in V,T} BCELoss(\mathcal{D}_m(E, \hat{H}_m), 1), \tag{16}$$

$$\begin{aligned} \mathcal{L}_{\mathcal{D}_m} = &\sum_{m \in V,T} BCELoss(\mathcal{D}_m(E, H_m), 1) \\ &+ \sum_{m \in V,T} BCELoss(\mathcal{D}_m(E, \hat{H}_m), 0). \end{aligned} \tag{17}$$

In total, the backbone loss $\mathcal{L}_{pred}$ and generative network loss $\mathcal{L}_{GAN}$ are trained iteratively. After obtaining a stable item ID embedding, we start the training of the CGAN network. The training process is detailed in Algorithm 1.

4 EXPERIMENT

4.1 Dataset

We perform our experiments on two datasets. One is the public dataset Amazon, and the other one is a collected industrial E-commerce multimodal recommendation dataset named ECM. The above datasets are detailed in Table 1.

Table 1: Statistics of Amazon and ECM datasets.

#Dataset	#Users	#Items	#Images	#Texts	#Categories	#Samples
Electronics	192403	63001	57387	47817	801	2801167
Home	66519	28237	27811	24134	890	903807
Cell Phones	27879	10429	10012	8368	126	305241
ECM	577042	710153	690377	626722	3078	5118048

Algorithm 1 Algorithm to optimize GMMF

```
Require: CGAN Network, GMMF Network, the number of iterations $N$, the threshold value $K$, learning rate $\eta$.
  for each $epoch$ in $N$ do
      if $epoch$ > K then
          Optimize CGAN Network parameters.
      end if
      Update GMMF Network parameters $\theta$ by minimizing:
          $\theta = \theta - \eta \bigtriangledown (\mathcal{L}_r + \mathcal{L}_{CTR})$
  end for
```

Amazon dataset: We use three subsets (Electronics, Home, Cell Phones) of the Amazon dataset while multimodal information covers at least 75 percent of the items of each subset. We set the training data and testing data following Zhou et al. [48] with the given user behavior sequence $u_i = \{x_1, x_2, x_3, \ldots, x_{N-1}, x_N\}$.

ECM dataset: ECM dataset is a large-scale industrial E-commerce content dataset. We sample 5118048 samples from the offline database, each sample contains user ID, item ID, and label for the CTR task. In the training phase, we use the samples across 30 days. In the testing phase, we use the sample of the next day. In this dataset, there are 577042 users and 710153 items but we only have 5118048 samples. The sparse property of user behavior and incompleteness of multimodal make this problem tricky.

4.2 Baselines

To verify the performance of GMMF, we compare 9 baseline methods, which are divided into three categories: traditional methods with single modality information (LR, FM), concatenation-based multimodal methods (DeepFM, YoutubeNet, DIN), and the methods based on the various multimodal fusion technologies (LMF, MTFN, NAML, MARN).

- **LR[23]:** Logistic regression (LR) is a shallow machine learning algorithm for solving binary classification problems to estimate the likelihood.
- **FM[30]:** Factorization Machines (FM) is based on matrix decomposition, which is to solve the problem of feature combination in the large-scale sparse matrix.
- **DeepFM[11]:** DeepFM combines the advantages of FM and DNN models, and simultaneously learns low-order feature combinations and high-order feature combinations.
- **YoutubeNet[6]:** YoutubeNet is a deep model proposed for video recommendation. It is also widely used in commercial recommender systems.
- **DIN[48]:** Deep Interest Network (DIN) is a deep model that adaptively learns the representation of user interests from historical behaviors by a designed local activation unit with attention mechanism.
- **LMF[22]:** Low-rank Multimodal Fusion (LMF) leverages low-rank weight tensors to make efficient multimodal fusion without compromising on performance.
- **MTFN[38]:** Multi-modal Tensor Fusion Network (MTFN) incorporates the advantages of both embedding-based and classification-based methods, and explicitly learns an accurate modalities similarity function with rank-based tensor fusion rather than seeking a common embedding.
- **NAML[42]:** Neural News Recommendation Approach With Attentive Multi-view Learning(NAML) uses attention mechanism to select important modalities for learning informative news representations.
- **MARN[19]:** Multimodal Adversarial Representation Network (MARN) explores complementarity by considering modality-specific and modality-invariant features for CTR prediction task.

In all the above methods, LR and FM use recommended item ID information only. DeepFM, YoutubeNet, and DIN concatenate ID, visual, and text features into a large vector but DeepFM does not utilize historical behavior sequences. LMF, MTFN, NAML, and MARN integrate multiple modalities to learn deeper interactive information.

4.3 Experimental Settings

We refer to the previous work [19] for setting the parameters. We set the ID embedding dimension of the user, item, category to 32, 32

Method	Electronics	Home	Cell Phones	ECM
LR	0.7212	0.6624	0.6199	0.5973
FM	0.7418	0.6743	0.6385	0.5961
DeepFM	0.7697	0.6870	0.6877	0.6082
YoutubeNet	0.7954	0.7373	0.7818	0.6367
DIN	0.7965	0.7397	0.8149	-
LMF	0.8047	0.7261	0.7702	0.6618
MTFN	0.8145	0.7368	0.7509	0.6613
NAML	0.8133	0.7233	0.6562	0.6626
MARN	0.8034	0.6922	0.7521	0.6630
GMMF	**0.8198**	**0.7484**	**0.8300**	**0.6632**

Table 2: AUC on Amazon and ECM dataset. We design nine baseline experiments to compare performance. Hint: - indicates that this method can not experiment in this dataset.

and 16, respectively. After dimensionality reduction, the dimension of visual and text features are the same as the ID embedding of the item. Adam optimizer is used to train CGAN and GMMF at different learning rates of 10^{-5} and 10^{-4}, respectively. In order to compare the performance, we adopt AUC as the evaluation metric [19]:

$$AUC = \frac{1}{|U|}\sum_{u\in|U|}\frac{1}{|I_u^+||I_u^-|}\sum_{i\in|I_u^+|}\sum_{j\in|I_u^-|}\delta(p_i > p_j), \tag{18}$$

where I^+ is positive samples, I^- is negative samples and δ is indicator function.

4.4 Comparison Result

We compare our approach with the baselines mentioned in the section 4.2 on Amazon and ECM datasets. Table 2 lists the performance comparison, and from the results, we have the following observations. First, the performance of GMMF is better than all baselines and achieves the state-of-the-art performance in both public and industrial datasets. GMMF improves AUC scores by 0.53%, 0.87%, 1.51%, and 0.02% respectively on each dataset over the best baseline method. The result demonstrates that our model is more effective in solving the multimodal recommendation problem by learning better representations than other methods. Second, compared to the traditional methods, multimodal methods with deep learning greatly enhance performance, which demonstrates the effectiveness of the multimodal information in the recommender system. Third, on Electronics and ECM datasets, the third group methods which focus on multimodal fusion perform better than other methods. However, when the amount of training data is relatively small in Home and Cell Phones datasets, the second group performs better than the third group. None of them perform perfectly on all datasets showing that these methods are heavily dependent on the size of the data. And our method achieves consistently best performance, which shows the stability and robustness of our method. Last but not least, on the ECM dataset, we find that the variance of the performance of multimodal methods is small. The reason for this is that the ECM dataset is a real large-scale industrial dataset and the data is relatively sparse. Besides, the distribution of training set and testing set can not be guaranteed to be independent and identically distributed. These facts easily lead to overfitting of the models. But our models still increase by 0.03% relative percentage on AUC score than the best performance.

4.5 Ablation Study

In this section, we design experiments to investigate the contribution of different modules to the recommender systems.

- **Base:** A model concatenates ID embedding, original visual and text features as input.
- **Base+Auto-Encoder:** Base model adds Auto-Encoder to encode the original visual and text features.
- **Base+CGAN+Auto-Difference:** Base model adds CGAN and Auto-Difference modules, as the two modules should combined together while the input of Auto-Difference is the output of CGAN.
- **Base+MIN:** Base model adds the MIN.
- **Base+DSN:** Base model adds the whole DSN.

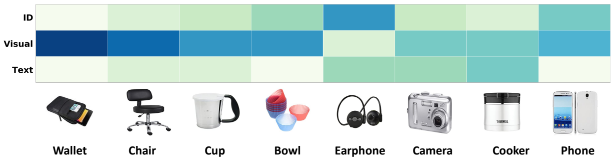

Figure 4: The heat map of modalities importance of randomly sampled items. The darker the color, the greater the importance.

- **Base+Auto-Encoder+MIN:** Base model adds Auto-Encoder and MIN.
- **Base+CGAN+Auto-Difference+MIN:** The GMMF eliminates the Auto-Encoder.
- **GMMF:** The entire proposed model with DSN and MIN.

Table 3 lists the contribution of Auto-Encoder, CGAN+Auto-Difference, MIN, and their combinations towards the performance improvement on Amazon and ECM datasets. From the results, we have the following observations. First, compared with Base model, each module of the proposed method improves the performance on all datasets. Second, Auto-Encoder improves by 0.41%, 0.04%, 3.22%, and 0.02% on Electronics, Home, Cell Phones and ECM datasets, respectively. The Auto-Encoder achieves better improvement on the Cell Phone dataset because the number of categories in the Cell Phone dataset is much less than that of other datasets. Specifically, this kind of situation leads features of items are more concentrated, so that the Auto-Encoder damages less information during dimensionality reduction. Third, only applying CGAN+Auto-Difference can improve the performance, but the performances are also unstable due to the instability of CGAN. Without Auto-Encoder, CGAN performs well on Home and ECM datasets, but improvements are slight on the other two datasets. Fourth, combining Auto-Encoder, the whole DSN improves 0.48%, 0.63%, 4.62% and 2.48% on Electronics, Home, Cell Phone, and ECM dataset, respectively. The results demonstrate that DSN is quite effective in improving the performance of recommmender systems. Fifth, with MIN, the performance improves 0.53%, 0.38%, 5.77% and 1.9% in terms of AUC, which verifies the MIN is also useful for recommender systems. Sixth, compared with using only MIN, the performance will be better with CGAN or Auto-encoder. It demonstrates MIN and the modules of DSN are complmentary to each other. Lastly, by combining the DSN and MIN, the model achieves the best recommendation performance on all dataset.

4.6 The Modality Specific User’s Interest

To explore the users’ interest of multiple modalities, we visualize the contribution of different modalities to the final prediction. In Figure 4, we randomly sample some user-item interactions from the Home dataset. We find that users tend to pay more attention to visual information when buying daily necessities such as bags and cups. However, when buying electronic products, users will pay less attention to visual information, but pay more attention to text information. As for items with the famous brand, ID embedding is given stronger attention, such as shock absorbing earphones and apple mobile phones in Figure 4.

Method					Electronics	Home	Cell Phones	ECM
		Auto-Encoder	CGAN+Auto-Difference	MIN				
With Multimodal	(a)				0.7999	0.7316	0.7613	0.6367
	(b)	✓			0.8040	0.7320	0.7935	0.6369
	(c)		✓		0.8012	0.7344	0.7644	0.6610
	(d)			✓	0.8052	0.7354	0.8190	0.6557
	(e)	✓	✓		0.8047	0.7379	0.8075	0.6615
	(f)	✓		✓	0.8077	0.7466	0.8255	0.6591
	(g)		✓	✓	0.8091	0.7444	0.8228	0.6617
	(h)	✓	✓	✓	**0.8198**	**0.7484**	**0.8300**	**0.6632**

Table 3: AUC on Amazon and ECM datasets in ablation experiments.

4.7 Qualitative Analysis

We statistics all samples when our model and other models make different predictions. From Figure 5, we can observe that our model overwhelmingly predicts more correct results than other models. Moreover, we notice that half of the cases where GMMF is better than baseline models are missing text modality. This phenomenon suggests that GMMF mitigates the negative effects of modality disappearance. Simultaneously, we sample some typical cases with complete modalities to further analyze.

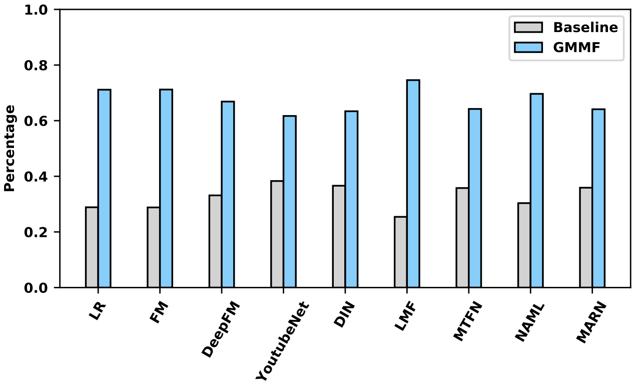

Figure 5: The proportion of samples correctly predicted by our model and other models when the predictions are inconsistent.

In Figure 6, we find that these cases have a common pattern: visual and text information are simple, and detailed information is scarce. The pattern causes ID embedding to be highly redundant with visual and text modalities. In this situation, LMF, MTFM and NAML focus on multimodal interactions leading to more serious information redundancy. MARN sums multimodal features with similar information, annihilating complementary information. Compared with them, our method GMMF removes the redundant information and mines sparse details from multimodal features.

4.8 Online Evaluation

We conduct the online A/B testing in our Content Recommender System for three weeks. For policy reasons, we only publish improvements relative to the base model. Compared with online base model which is mainly implemented as YoutubeNet, our model achieves 6.64% CTR relatively improvement. Considering the massing of the system, our method improves the online performance significantly, which shows the effectiveness of the proposed method in real industrial scenarios.

Visual	Text	Visual	Text
	SONY earphones.		**Crystal Glasses.**
	Apple charger.		**White sheets.**
	A black pencil.		**Toothpick.**

Figure 6: Some typical cases GMMF predicts correctly while all baseline models predict wrongly.

#Method	#CTR Improve
YoutubeNet	0%
GMMF	6.64%

Table 4: Online A/B testing.

5 CONCLUSIONS

This paper proposes a novel multimodal recommender framework to weaken the redundancy between heterogeneous modalities. To extract the differences between various modalities, we designs a DSN network based on generative adversarial learning. Besides, we take into account the differences between users and integrates historical behavior as modality-specific embedding for each modality. Moreover, we design the gating mechanism MIN to set unequal weights to different modalities. In order to increase confidence, we conduct a large number of experiments, the results show our model has achieved better performance to the state-of-the-art methods on both public and collected industrial datasets.

REFERENCES

[1] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems* 16, 6 (2010), 345–379.

[2] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.

[5] Qiwei Chen, Huan Zhao, Wei Li, Pipei Huang, and Wenwu Ou. 2019. Behavior sequence transformer for e-commerce recommendation in alibaba. In *Proceedings of the 1st International Workshop on Deep Learning Practice for High-Dimensional Sparse Data*. 1–4.

[6] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*. 191–198.

[7] Yi Ding, Alex Rich, Mason Wang, Noah Stier, Pradeep Sen, Matthew Turk, and Tobias Höllerer. 2021. Sparse Fusion for Multimodal Transformers. *arXiv preprint arXiv:2111.11992* (2021).

[8] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. 2021. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[9] John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research* 12, 7 (2011).

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[11] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 1725–1731.

[12] Zhiqi Huang, Fenglin Liu, Xian Wu, Shen Ge, Helin Wang, Wei Fan, and Yuexian Zou. 2021. Audio-oriented multimodal machine comprehension via dynamic inter-and intra-modality attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 13098–13106.

[13] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.

[14] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. 2020. A survey of the recent architectures of deep convolutional neural networks. *Artificial intelligence review* 53, 8 (2020), 5455–5516.

[15] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[16] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* (2014).

[17] Sangho Lee, Youngjae Yu, Gunhee Kim, Thomas Breuel, Jan Kautz, and Yale Song. 2020. Parameter efficient multimodal transformers for video representation learning. *arXiv preprint arXiv:2012.04124* (2020).

[18] Chenyi Lei, Yong Liu, Lingzi Zhang, Guoxin Wang, Haihong Tang, Houqiang Li, and Chunyan Miao. 2021. SEMI: A Sequential Multi-Modal Information Transfer Network for E-Commerce Micro-Video Recommendations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 3161–3171.

[19] Xiang Li, Chao Wang, Jiwei Tan, Xiaoyi Zeng, Dan Ou, Dan Ou, and Bo Zheng. 2020. Adversarial multimodal representation learning for click-through rate prediction. In *Proceedings of The Web Conference 2020*. 827–836.

[20] Junyang Lin, An Yang, Yichang Zhang, Jie Liu, Jingren Zhou, and Hongxia Yang. 2020. Interbert: Vision-and-language interaction for multi-modal pretraining. *arXiv preprint arXiv:2003.13198* (2020).

[21] Hu Liu, Jing Lu, Hao Yang, Xiwei Zhao, Sulong Xu, Hao Peng, Zehua Zhang, Wenjie Niu, Xiaokun Zhu, Yongjun Bao, et al. 2020. Category-Specific CNN for Visual-aware CTR Prediction at JD. com. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2686–2696.

[22] Zhun Liu and Ying Shen. 2018. Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

[23] H Brendan McMahan, Gary Holt, David Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, et al. 2013. Ad click prediction: a view from the trenches. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1222–1230.

[24] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[25] Gabriel de Souza P Moreira, Sara Rabhi, Ronay Ak, Md Yasin Kabir, and Even Oldridge. 2021. Transformers with multi-modal features and post-fusion context for e-commerce session-based recommendation. *arXiv preprint arXiv:2107.05124* (2021).

[26] Teodora Pandeva and Matthias Schubert. 2019. Mmgan: Generative adversarial networks for multi-modal distributions. *arXiv preprint arXiv:1911.06663* (2019).

[27] Yuxin Peng and Jinwei Qi. 2019. CM-GANs: Cross-modal generative adversarial networks for common representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 1 (2019), 1–24.

[28] R Gnana Praveen, Eric Granger, and Patrick Cardinal. 2021. Cross Attentional Audio-Visual Fusion for Dimensional Emotion Recognition. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 1–8.

[29] Shi Pu, Yijiang He, Zheng Li, and Mao Zheng. 2020. Multimodal Topic Learning for Video Recommendation. *arXiv preprint arXiv:2010.13373* (2020).

[30] Steffen Rendle. 2010. Factorization machines. In *2010 IEEE International conference on data mining*. IEEE, 995–1000.

[31] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale visual recognition. *arXiv preprint arXiv:1409.1556* (2014).

[32] Xue Song, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. 2021. Spatial-temporal graphs for cross-modal text2video retrieval. *IEEE Transactions on Multimedia* (2021).

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[34] Sunny Verma, Chen Wang, Liming Zhu, and Wei Liu. 2019. Deepcu: Integrating both common and unique latent information for multimodal sentiment analysis. In *International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization.

[35] Sunny Verma, Jiwei Wang, Zhefeng Ge, Rujia Shen, Fan Jin, Yang Wang, Fang Chen, and Wei Liu. 2020. Deep-HOSeq: Deep higher order sequence fusion for multimodal sentiment analysis. In *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 561–570.

[36] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research* 11, 12 (2010).

[37] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. 2016. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215* (2016).

[38] Tan Wang, Xing Xu, Yang Yang, Alan Hanjalic, Heng Tao Shen, and Jingkuan Song. 2019. Matching images and text with multi-modal tensor fusion and re-ranking. In *Proceedings of the 27th ACM international conference on multimedia*. 12–20.

[39] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015. On deep multi-view representation learning. In *International conference on machine learning*. PMLR, 1083–1092.

[40] Wenjie Wang, Ling-Yu Duan, Hao Jiang, Peiguang Jing, Xuemeng Song, and Liqiang Nie. 2021. Market2Dish: Health-aware food recommendation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17, 1 (2021), 1–19.

[41] 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 *Proceedings of the 27th ACM International Conference on Multimedia*. 1437–1445.

[42] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with attentive multi-view learning. *arXiv preprint arXiv:1907.05576* (2019).

[43] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. MM-Rec: Multimodal News Recommendation. *arXiv preprint arXiv:2104.07407* (2021).

[44] Yang Wu, Zijie Lin, Yanyan Zhao, Bing Qin, and Li-Nan Zhu. 2021. A Text-Centered Shared-Private Framework via Cross-Modal Prediction for Multimodal Sentiment Analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 4730–4738.

[45] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 10790–10797.

[46] Rizgar Zebari, Adnan Abdulazeez, Diyar Zeebaree, Dilovan Zebari, and Jwan Saeed. 2020. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends* 1, 2 (2020), 56–70.

[47] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 5941–5948.

[48] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1059–1068.

[49] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao. 2019. R2gan: Cross-modal recipe retrieval with generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11477–11486.