electronics

MDPI

Article

MultArtRec: A Multimodal Neural Topic Modeling for Integrating Image and Text Features in Artwork Recommendation

Jiayun Wang [1,*], Akira Maeda [2,*] and Kyoji Kawagoe [2]

[1] Graduate School of Information Science and Engineering, Ritsumeikan University, Shiga 525-8577, Japan
[2] College of Information Science and Engineering, Ritsumeikan University, Shiga 525-8577, Japan
* Correspondence: jiayunwong@hotmail.com (J.W.); amaeda@is.ritsumei.ac.jp (A.M.)

Abstract: Recommender systems help users obtain the content they need from massive amounts of information. Artwork recommender systems is a topic that has attracted attention. However, existing art recommender systems rarely consider user preferences and multimodal information at the same time, while utilizing all the information has the potential to help make better personalized recommendations. To better apply recommender systems to the artwork-recommendation scenario, we propose a new neural topic modeling (NTM)-based multimodal artwork recommender system (MultArtRec), that can take all the information into account at the same time and extract effective features representing user preferences from multimodal content. Also, to improve MultArtRec's performance on monomodal feature extraction, we add a novel topic loss term to the conventional NTM loss. The first two experiments in this study compare the performances of different models with different monomodal inputs. The results show that MultArtRec can improve the performance with image modality inputs by up to 174.8% compared to the second-best model and improve the performance with text modality inputs by up to 10.7% compared to the second-best model. The third experiment is conducted to compare the performance of MultArtRec with monomodal inputs and multimodal inputs. The results show that the performance of MultArtRec with multimodal inputs can be improved by up to 15.9% compared to monomodal inputs. The last experiment preliminarily tests the versatility of MultArtRec on a fashion recommendation scenario that considers clothing image content and user preferences. The results show that MultArtRec outperforms the other methods across all the metrics.

Keywords: multimodal; recommender system; neural topic modeling

Citation: Wang, J.; Maeda, A.; Kawagoe, K. MultArtRec: A Multimodal Neural Topic Modeling for Integrating Image and Text Features in Artwork Recommendation. *Electronics* **2024**, *13*, 302. https://doi.org/10.3390/electronics13020302

Received: 20 November 2023
Revised: 21 December 2023
Accepted: 8 January 2024
Published: 10 January 2024

1. Introduction

Recommender systems are designed to assist users in discovering items of interest in a vast sea of choices. However, no recommender system can be universally applied to diverse contexts, and as such, recommender systems are often designed to adapt to specific scenarios and the accompanying datasets.

Artwork recommender system research is a topic of significant academic and commercial value. The artwork recommender systems not only allow users to better discover their favorites from many artworks but also contribute to the enhancement of art-education purposes in public institutions and the sales performance of art-selling websites. There already exist artwork-recommendation applications in numerous public institutions, such as Europeana (https://www.europeana.eu/en, accessed on 7 January 2024), The Metropolitan Museum of Art (https://www.metmuseum.org/), Rijksmuseum (https://www.rijksmuseum.nl/en/rijksstudio, accessed on 7 January 2024), Brooklyn Museum (https://www.brooklynmuseum.org/), etc., and commercial websites, such as Art.com (https://www.art.com/) and Artsy (https://www.artsy.net/). On these websites, the recommender systems provide relevant items that may be of interest to the users. For example, in Europeana, after clicking on a digital heritage item, the user will see items

that have similar metadata, such as contributors and publishers, to the current digital heritage item on the detailed page. For another example, in Artsy, if a user clicks on a certain artwork item, artworks created by the same author and artworks in the same gallery will appear at the bottom of the detailed page. Many of the artwork recommender system applications have a very simple recommendation mechanism—text matching—which is to find similar items by matching the same authors, publisher, gallery, etc. This cannot meet the needs of users in many cases, such as exploring similar items of visual styles or cultural backgrounds.

Researchers have noticed the above problem and proposed more advanced recommender systems and content-based (CB) recommendation methods [1,2] for the artwork-recommendation scenario. CB methods are used to extract the features of the artwork items and make recommendations based on the similarities between the items. However, CB methods may not provide recommendations that are well-suited to individual user preferences because users do not always like the same content. There are two main categories of recommendation methods: CB methods and collaborative filtering (CF) methods. CB methods leverage features of item contents (image, text, metadata, etc.) to calculate similarities between items and then make recommendations based on the descending order of content similarities. On the other hand, CF methods leverage user–item interaction matrices to compute similarities between users or items. User–item interaction matrices are two-dimensional matrices created by users' historical preferences for items, where the values in the matrices represent the user's preference for the item. After the similarities are calculated, CF methods will recommend preferred items that similar users visited or directly recommend similar items. In general, compared to the CB methods, CF methods possess an advantage in that they can provide recommendations that align more closely with user preferences, as they take into account user preference data. Furthermore, considering the increasing availability of diverse sources for obtaining users' historical preferences, CF methods are promising to have a more pronounced impact in the field of art recommender systems. Considering the potential possibility of art recommendation applications in the future, the methods that utilize CF information are more suitable as a proposal method.

In recent years, image-based recommender systems that take into account users' historical preferences have attracted significant attention, such as in picture recommendations [3] and fashion recommendations [4]. These methods have been designed with specific considerations for image features in particular contexts. For instance, picture recommendations take into account the user's preferences for the color, layout, and concept of the picture, while fashion recommendations focus on large-scale or small-scale clothing styles. Compared with the general image-based recommendation, in the field of art recommendation, a user's preference is notably affected by the text features of the items (e.g., artwork titles and descriptions). In this study, we aim to propose a method for an art recommendation scenario, so we consider using both the image features and text features of the items.

Multimodal technologies are widely used in the computer vision field [5,6] and natural language processing field [7]. An object in the real world usually does not only have monomodal information (only image information or only text information). Leveraging information from more than one modality can enhance the representation of objects in high-dimensional vector spaces, therefore enabling algorithms to perform more effectively across various machine-learning tasks. For instance, animals that are visually similar in images can be more accurately identified by incorporating text that describes their distinguishing features. Multimodal recommender systems have also attracted increasing attention [8]. One of the reasons is that, in many cases, items have multimodal information, which can aid recommender systems in more effectively extracting user preferences and creating more accurate user profiles. Another reason is that the maturation of models that extract monomodal features [9,10] contributes to the potential enhancement of downstream task performance. Our proposed method is inspired by multimodal technology, using image features and text features of artworks to better represent users and artwork items and extract the relationship between them to provide better recommendation results.

In this study, the issues to be addressed are stated as follows:

- To help users find the items they are interested in when browsing websites providing artwork content, propose a recommendation method that can: (1) take into account multimodal data (e.g., image, text, video, etc.), which is increasingly abundant in those websites and helpful for extracting user preferences, (2) consider the user–item interactions for providing precise recommendations.
- Design a part in the proposed method to effectively extract multimodal features of items. Among the existing multimodal recommender systems, there are already methods that can effectively extract one type of multimodal feature (e.g., image-only or text-only) at one time. Based on these existing methods, the purpose is to propose a method that can effectively extract multiple multimodal features at the same time and can best extract better features from monomodal input.

To address these issues, we propose a multimodal artwork recommender system, MultArtRec. The source code of MultArtRec can be downloaded from GitHub (https://github.com/blueorris/MultArtRec). Our research has the following novelties:

- In contrast to typical multimodal models utilizing only image + CF information or text + CF information, MultArtRec can take into account multimodal information and CF information at the same time. Therefore, MultArtRec is considered to be more effective and flexible than conventional multimodal models.
- To our best knowledge, we are the first to adopt neural topic modeling (NTM) for multimodal feature extraction in recommendation tasks. We also propose to add a novel topic loss term to the conventional NTM loss function that both ensures the symmetry of the encoder–decoder architecture and better extracts monomodal or multimodal features from the inputs.

This paper is structured as follows. Section 2 states the work related to our research. Section 3 introduces the background knowledge to understand the NTM model. Section 4 states the methodology of our proposed MultArtRec. Section 5 introduces the experiments, and Section 6 concludes the paper.

2. Related Work

This study involves a wide range of backgrounds. We introduce some representative studies in this section. They are divided into 3 categories: (1) background of recommender systems, (2) recommender systems based on monomodal or multimodal information, and (3) the encoder–decoder structured recommender systems that are most similar to our proposed MultArtRec.

2.1. Recommender Systems

Presently, recommender systems have been widely used in various commercial applications, such as e-commerce website Amazon (https://www.amazon.com/), food delivery platform Uber Eats (https://www.ubereats.com/), music streaming service Spotify (https://open.spotify.com/), etc. These applications employ recommender systems to provide convenience to the users and have gained popularity.

As well as being applied as commercial applications, recommender systems are also active in various fields. Some research focuses on making people's lives more convenient. For example, literature [11] states the approaches for food recommendation. Research [12] proposed a method to recommend tourist destinations. The article [13] concentrates on recommending people to users in the social media and career development context.

Some studies focus on contributing to public utilities. For example, literature [14] shows the implementations of recommender systems that provide support to educational endeavors. Literature [15] summarizes research that contributes to healthcare, which makes recommendations for lifestyle, nutrition, and health information. In the COVID-19 era, recommender systems also show their viability of facilitating human beings in overcoming the challenge. Research [16] proposed a method for medical staff recommendation under

the scenario of limited resources, and research [17] provides a method for recommending vaccines.

These applications and studies show that recommender systems have been widely used. It is considered that the recommender systems have the potential to participate in a wider range of fields.

2.2. Recommender Systems Based on Monomodal or Multimodal Information

In this section, we introduce various recommender systems that recommend images or artworks, and we classify them based on the information type they use. Table 1 shows the information types that are designed to be used in each method.

Table 1. Different information types that are designed to be used by different methods in related work.

Method	Semantic	Image	Text	CF
Wang et al. [18]	✓			
Deladiennee et al. [19]	✓			
Strezoski et al. [1]		✓		
Messina et al. [2]		✓		
Frost et al. [20]		✓		
Qiu et al. [21]		✓		✓
Messina et al. [22]		✓	✓	
Yilma et al. [23]		✓	✓	
Wang et al. [24]		△	△	✓
Li et al. [25]		△	△	✓
MultArtRec		✓	✓	✓

(1) The symbol "✓" indicates that all the information types can be used at the same time. (2) The symbol "△" indicates that the information types cannot be used at the same time (e.g., either image + CF or text + CF).

2.2.1. Semantic-Based Recommender Systems

A semantic network is a graphical representation of relationships between entities (concepts). It is often used to represent the relationship among the metadata of artworks. Therefore, many of the studies in the field of artwork recommendation concentrate on taking semantic information into account.

Wang et al. [18] is one of the first efforts to contribute to artwork recommender systems. They implemented a recommendation framework, CHIP, for the Rijksmuseum. CHIP utilizes the semantic network of cultural heritages and includes a cultural heritage recommender system, a tour wizard, and a tour guide. This study shows that the proposed recommender system using semantic networks can deal with the cold-start and sparsity problems and can explain the recommendation results.

Deladiennee et al. [19] introduced a semantic-based recommender system that relies on an ontological formalization of knowledge. The study also noted the abundant multimodal data in museums, such as painting, video, and audio, which can describe and contextualize an item by its history and creator. This work is mainly to solve the problem of information overload while helping visitors find interesting content from a large amount of content and helping museums target users' interests when introducing their exhibits.

Although semantic information does often exist in artwork data, there are also many artworks, especially newly created artworks, that do not have semantic information. Furthermore, website visitors who are interested in semantic information often have professional knowledge to some degree, but many of the visitors do not have that knowledge. Considering these factors, we do not consider semantic information in our proposed artwork recommender system for this time.

2.2.2. Image-Based Recommender Systems

Strezoski et al. [1] proposed TindART, aimed at analyzing users' tastes in visual art. They employed a multi-task learning deep neural network to extract content features from artworks. Through real-time artwork preference selections, users can express and explore their art style preferences.

Messina et al. [2] proposed CuratorNet, a model that utilizes ResNet [26] to extract image content features and employs a triplet training strategy to learn users' visual preferences. The triplet training strategy optimizes item representations by ensuring high similarity between artworks liked by a user and low similarity between liked and disliked artworks. This can intuitively capture the representation of artwork image contents and is able to recommend items for users who come to the system for the first time (cold-start case).

As mentioned in Section 1, the two studies above utilized CB methods, which will fail to provide precise recommendations to the user preferences.

Frost et al. [20] proposed a method that focuses on the image features of artworks that the users do not like and only avoids recommending items that are similar to those disliked ones. This can give the recommendations a broader range. This method can effectively prevent the filter bubble phenomenon. The filter bubble phenomenon refers to the fact that always providing similar recommendations to users will prevent them from accessing other perspectives of information. This study did very meaningful work, but the purpose of our work is to help users precisely obtain the content in which they are interested; therefore, it is not consistent with the purpose of our research.

Qiu et al. [21] proposed a recommender system, CausalRec, for fashion products. It introduced an observation that users, when purchasing clothes in online shops, are influenced not only by the products' visual factor but also by other non-image factors, such as brand and material, when buying clothes. To address this issue, they employed causal graphs to identify and analyze the visual bias of some existing methods, and they proposed a debiased visually aware recommender system, CausalRec. This method is designed to eliminate this spurious relationship that misleads the prediction of the user's real preference. The proposed method outperforms methods with visual biases [27,28] on the Amazon product datasets. In addition to image features, CausalRec also takes image information and CF information into account. CausalRec is used in the comparative experiments with image modality input in this research.

2.2.3. Multimodal Recommender Systems

Messina et al. [22] proposed a CB method that uses metadata and image features of artworks. Some of the image features were created manually, and some were extracted using a neural network. However, comparative experiments showed that the features extracted using neural networks were far superior to the features extracted automatically and manually in the task of recommending works of art. They also indicated that the hybrid approach combining visual features and textual attributes (e.g., artist, title, style, etc.) yields a performance improvement and that using only image features yields better results than using only metadata. This study provides a very valuable conclusion for our research, i.e., (1) using both image and text information to recommend artworks is effective, and (2) using neural networks to extract image features is effective.

Yilma et al. [23] adopted neural networks to extract image features and text features (e.g., title, artist, description, etc. of artworks) and used the combination of them to create embeddings of artworks. They also created combination embeddings for each user to represent their preferences. Finally, the similarities between the artworks and the users are calculated, and recommendations are made. This study stated that user evaluation data are very important. This research also provides valuable experience for our research, i.e., using both image and text features is effective in art recommendation.

Both of the above studies used multimodal data, especially image and text data, and obtained good results, which gave us great inspiration. However, they did not con-

sider the user–item interaction matrix, so they considered it to be difficult to provide personalized recommendations.

2.2.4. Encoder–Decoder Architectured Multimodal Recommender Systems

In the field of multimodal recommender systems, there is a class of methods based on the encoder–decoder architecture. Since such architecture can be flexibly applied to various data and modalities, it is very suitable for our recommendation scenarios.

Wang et al. [24] proposed collaborative deep learning for recommender systems (CDL). The CDL model is designed to utilize auxiliary information to alleviate the sparsity problem and cold-start problem, which are common problems in recommender systems. The CDL model extracts the latent features of the items' auxiliary information with an encoder–decoder architectured deep-learning neural network part and then integrates the latent features into a CF probabilistic framework. This model is considered to effectively capture the similarity and implicit relationship between items and users. In addition, the CDL model has a denoise layer in it, thus having better effects on sparse data. The deep-learning part that extracts latent features can be applied to any auxiliary information that can be represented as vectors, therefore ensuring that the CDL model has flexibility for various data modalities.

Li et al. [25] proposed a collaborative variational autoencoder for recommender systems (CVAE). Similar to CDL, CVAE also utilizes auxiliary information to alleviate the sparsity problem as well as the cold-start problem, extracts item latent features, and then integrates them into a CF framework. The CVAE model extracts item latent features by the variational autoencoder (VAE) [29]. As mentioned in this study, the main reason CVAE's feature extraction structure is better than CDL is that the denoising layer added to CDL makes the features lose their Bayesian nature, which is difficult for Bayesian inference or requires high computational cost, while CVAE avoids this problem. This study also provided variants of the CVAE model, allowing it to have room for improvement with image input or sequential input.

2.3. Neural Topic Modeling

Among neural network models with encoder–decoder, the neural topic models (NTMs) are considered to be capable of extracting topics from input vectors. This provides a promising method for neural networks to understand the focus of features.

NTMs [30] are a kind of topic modeling technique that includes various neural network structures. One of the basic NTM model structures is mentioned in [31] (in the subsequent text, the basic NTM model is denoted as NTM). The neural network structure of NTM is very similar to VAE, which is also developed with a neural variational inference framework. The NTM model has been verified to be able to effectively extract topics from documents.

Compared with the classic topic modeling method latent Dirichlet allocation (LDA) [32], NTM is easier to combine with other neural networks, which can easily achieve more complex tasks, such as recommendation tasks. LDA assumes that documents are mixtures of topics and words are generated from these topics with a certain probability distribution. It was at first used to extract document topics, but later, LDA also showed superiority in extracting image topics for image classification task [33]. In [34], Hörster et al. also pointed out that the topic modeling method, latent semantic analysis, has a strong ability to extract coherent topics from image features. Inspired by these studies, we propose to use the NTM model to extract the latent topics from the image and text contents of the artwork items.

3. Introduction of Basic Models

In this section, we aim to introduce the technical background that is related to our proposed NTM-based multimodal recommendation method to better explain the rationality and effectiveness of its application in art recommendation.

3.1. Introduction of Variational Autoencoder (VAE)

Dimensionality reduction is a technique in machine learning and statistics that aims to reduce the number of input features while retaining as much relevant information as possible. High-dimensional data, where each data point has many features, can pose challenges such as increased computational complexity, increased risk of overfitting, and the curse of dimensionality.

An autoencoder is a typical model with encoder–decoder architecture that is designed to learn efficient representations of data, often for dimensionality reduction. The structure of an autoencoder is shown in Figure 1.

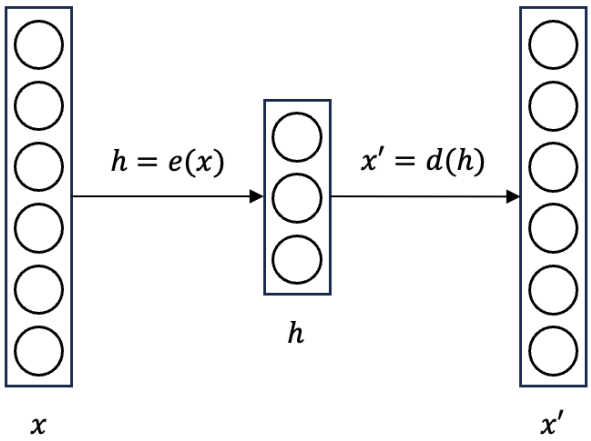

Figure 1. The model structure of an autoencoder.

The encoder of the autoencoder is denoted as $e(\cdot)$, and the decoder is denoted as $d(\cdot)$ in the figure. The latent features that are obtained after dimensionality reduction are denoted as h. The loss function of the autoencoder, $\mathcal{L}_{recon}(x, x')$, is to measure the difference between the input x and the reconstructed x', which is also called the reconstruction error. Common loss functions for calculating reconstruction error include mean squared error (MSE) and binary cross-entropy.

It is difficult for an autoencoder to produce new content. The reason is that the autoencoder does not learn the distribution of the input in the latent space, so it cannot sample from the latent space to generate new content. Moreover, autoencoders will easily cause overfitting problems because they have not learned the real distribution of the latent space, thus affecting the quality of latent features.

VAE assumes the latent space as a standard multivariate normal distribution, making the model better express the latent space. The model structure of VAE is shown in Figure 2.

In the VAE model, the input x is at first encoded into a multivariate normal distribution with mean μ and covariance matrix $\sum$, which are obtained through $e_1(\cdot)$ and $e_2(\cdot)$, respectively. Next, the model samples latent variables h from the distribution $\mathcal{N}(\mu, \sum)$. Finally, it decodes h to reconstruct x' through $d(\cdot)$, which is the same as the autoencoder.

VAE's loss function includes a reconstruction term and a regularization term. The reconstruction term is the same as that in the autoencoder. The regularization term encourages the distribution of the latent space $\mathcal{N}(\mu, \sum)$ to be close to a standard normal distribution $\mathcal{N}(0, I)$. It is expressed as the Kullback–Leibler divergence (KL divergence). In summary, VAE's loss function $\mathcal{L}_{VAE}$ can be denoted as:

$$\mathcal{L}_{VAE} = \mathcal{L}_{recon}(x, x') + KL(\mathcal{N}(\mu, \sum), \mathcal{N}(0, I)). \quad (1)$$

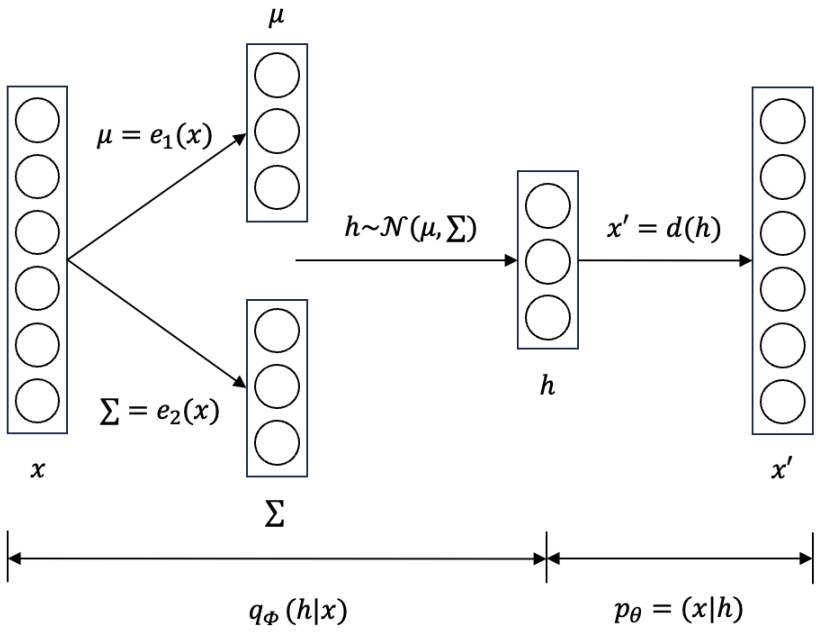

Figure 2. The model structure of VAE.

Because the KL divergence is often intractable to compute directly, VAE uses evidence lower bound (ELBO) to approximate the KL loss. The equation of ELBO is:

$$\text{ELBO} = \mathbb{E}_{q_\phi(h|x)}[\log p_\theta(x \mid h)] - \text{KL}\big(q_\phi(h \mid x) \| p(h)\big). \tag{2}$$

with this design, VAE can model more expressive latent space than the autoencoder, which is an advantage for expressing rich item content.

3.2. Introduction of Neural Topic Modeling (NTM)

NTM is a model designed based on VAE. The NTM's encoder q_Φ and decoder p_θ structure is shown in Figure 3. In this model, the encoder infers a doc–topic distribution that encodes the original documents into a latent topic space z. The decoder infers a topic-doc distribution that generates documents from the latent topics.

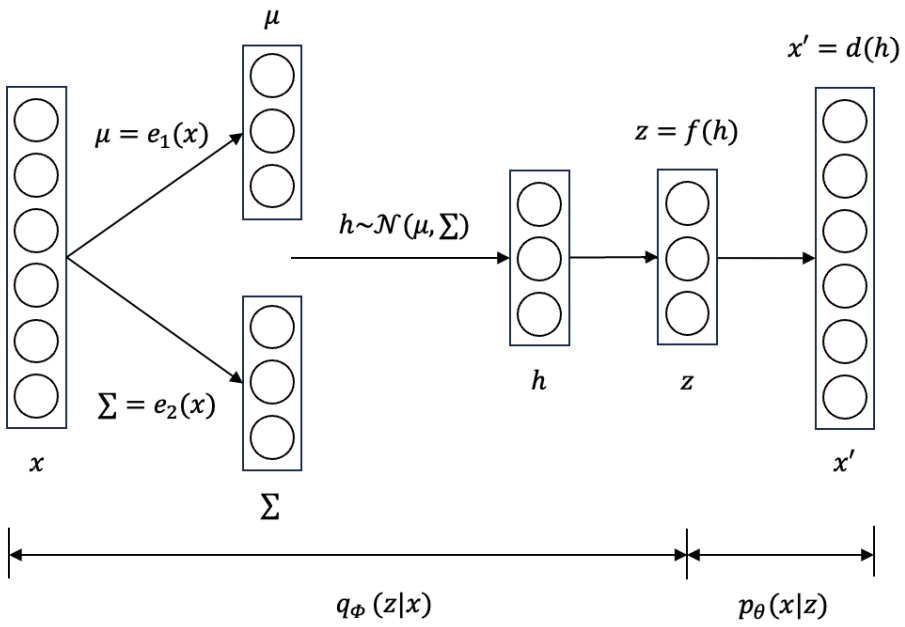

Figure 3. The model structure of NTM.

Structurally, NTM has one more latent variable z than VAE, which intuitively encodes the input into a space where each dimension represents a topic word. It is obtained by:

$$z = ReLU(h). \tag{3}$$

In the loss function, both NTM and VAE have a reconstruction term and a regularization term. The difference lies in that NTM utilizes the invariance of KL divergence under deterministic mapping between h and z, that is:

$$\mathrm{KL}\big(q_\phi(z \mid x) \| p(z)\big) = \mathrm{KL}\big(q_\phi(h \mid x) \| p(h)\big). \tag{4}$$

The NTM is superior in its ability to describe documents well and to produce topics that carry coherent semantic meaning, as well as offer a flexible framework to integrate into other neural networks.

4. Methodology

This section introduces our proposed NTM-based method, MultArtRec. The structure of MultArtRec is illustrated in Figure 4. Table 2 explains the notations in the structure figure. This model is divided into two parts: the NTM part and the CF part. The NTM part is used to extract topics of an item's content. Since the NTM part cannot predict user–item ratings, the CF part is adopted to do that. The CF part is the same as that in the CVAE method and CDL method. After combining these two parts, we can also easily compare our proposed MultArtRec with two classic multimodal recommendation models, CVAE and CDL.

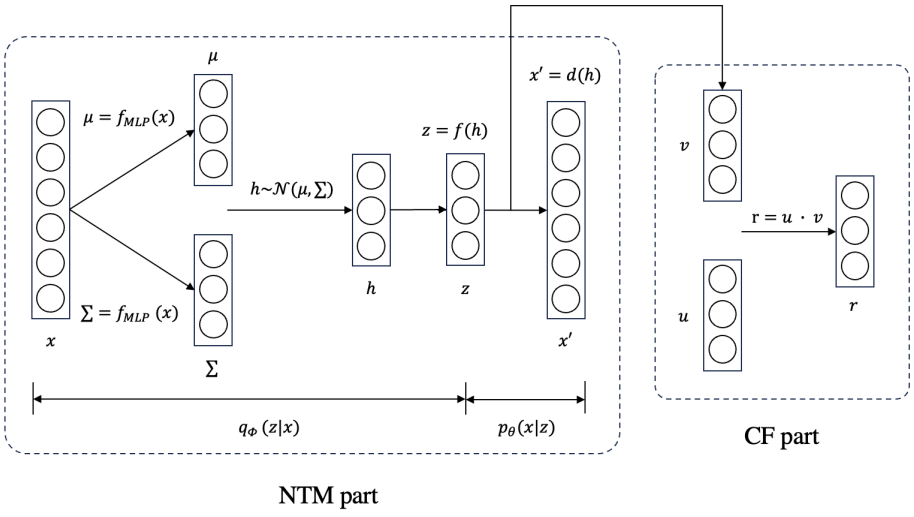

Figure 4. The model structure of our proposed method MultArtRec.

Table 2. Notations for MultArtRec.

Symbol	Description
x	input
μ	mean vector of latent features
$\sum$	covariance matrix of latent features
h	latent features
z	latent topics
x'	reconstructed x
u	embedding of users
v	embedding of items
r	rating predictions
f_{MLP}	multi-layer perceptron

In the NTM part, Let $x \in \mathbb{R}^N$ be the input vector. It can be the embedding of any item content, and N is the dimension of the input features. Let $z \in \mathbb{R}^K$ be the latent topic variable, where K is the number of topics. The encoder of the NTM part is represented as $q_\phi(z \mid x)$, and the decoder of the NTM part is represented as $p_\theta(x \mid z)$. In the encoder, f_{MLP} refers to a multi-layer perceptron that can include any number of layers. Generally, an input with a larger number of dimensions requires more layers of MLP for better feature extraction. In the CF part, let $v \in \mathbb{R}^K$ be the embeddings of items, and $u \in \mathbb{R}^K$ be the embeddings of users. The rating prediction is calculated by the inner product of u and v.

In MultArtRec, item embeddings are represented as v in the CF part. They are obtained by first sampling from $\mathcal{N}(\mu, \sum)$ to obtain h, then encoding h into a space z where each dimension represents an image or text topic, and finally, through the linear transformation between z and v. User embeddings are represented as u in the CF part. They are randomly initialized at first. The rating r is the inner product of item embeddings and user embeddings. Both item embedding v and user embedding u are updated during the training phase so that they have better expression in the corresponding tasks. The specific tasks are specified in the loss function.

The loss function of our proposed MultArtRec $\mathcal{L}_{MultArtRec}$ is constructed with five losses: reconstruction loss $\mathcal{L}_{recon}$, regularization loss $\mathcal{L}_{regu}$, item loss $\mathcal{L}_{item}$, rating loss $\mathcal{L}_{rating}$, and topic loss $\mathcal{L}_{topic}$. By minimizing the loss, all the parameters in the neural network, including user embedding u, will be updated. The loss function can be denoted as:

$$\mathcal{L}_{MultArtRec} = \lambda_n \cdot \mathcal{L}_{recon} + \lambda_r \cdot \mathcal{L}_{regu} + \lambda_v \cdot \mathcal{L}_{item} + \lambda_c \cdot \mathcal{L}_{rating} + \lambda_t \cdot \mathcal{L}_{topic}, \tag{5}$$

where $\lambda_n, \lambda_r, \lambda_v, \lambda_c, \lambda_t$ are manually set parameters that are used to control the proportion of each loss.

To be more specific, reconstruction loss is calculated as the squared error between the input and the reconstruction. It can ensure that the extracted topic still retains the main features of the input after dimension reduction so that it can be reconstructed. The formula of the reconstruction loss is:

$$\mathcal{L}_{recon} = \sum_{i=1}^{N} ||x_i - x_i'||^2. \tag{6}$$

Regularization loss is the same as that in VAE and NTM, adopting ELBO to minimize the divergence between the latent feature distribution and a standard multivariate normal distribution. It satisfies the assumption that h is sampled from a standard multivariate normal distribution. The formula of the regularization loss is:

$$\mathcal{L}_{regu} = \sum \log p_\theta(x \mid z) - KL(q_\phi(h \mid x)||p_\theta(h)). \tag{7}$$

The item loss is adopted in the CVAE's and CDL's CF parts, i.e., the squared error between the latent topic z and the item embedding v. It extracts more specific features from topic z for rating prediction. The formula for the item loss is:

$$\mathcal{L}_{item} = \sum_{i=1}^{K} ||z_i - v_i||^2. \tag{8}$$

The rating loss is also adopted in CVAE's and CDL's CF part, i.e., the squared error between the predicted ratings r and the ground truth ratings $\hat{r}$. This is an important term that helps to obtain item embeddings and user embeddings that can represent user preferences. The formula for rating loss is:

$$\mathcal{L}_{rating} = \sum_{i=1}^{N} ||r_i - \hat{r}_i||^2. \tag{9}$$

The topic loss is a new improvement we propose, which calculates the squared error between latent features h and latent topics z:

$$\mathcal{L}_{topic} = \sum_{i=1}^{K} ||h_i - z_i||^2. \tag{10}$$

The reason we propose topic loss is stated below. In VAE practical implementations, symmetrical matrices in the model are sometimes set to be the same to ensure the symmetry of encoder–decoder architecture. CVAE is also implemented in this way. However, CVAE shows limited performance in recommendation tasks. We assume that the reason for this is because of the inappropriate symmetrical design, and that is why we propose the topic loss. Topic loss can be intuitively understood as reducing the difference between the middle symmetry layers.

Previous content in this section introduces the model structure and loss function for monomodal modality input. When using multimodal data as input, the model can be understood as having multiple NTM parts. Each NTM part extracts latent topics from different modalities. In this case, the item loss consists of multiple terms that calculate differences between latent topics of different modalities and the item embedding v, and each term has a bias. The following formula defines the item loss of a model with two NTM parts:

$$\mathcal{L}_{item-multimodal} = \lambda_a \sum_{i=1}^{K} ||z_a^i - v_i||^2 + \lambda_b \sum_{i=1}^{K} ||z_b^i - v_i||^2. \tag{11}$$

where the bias λ_a and λ_b can be manually set, z_a and z_b are the latent topics extracted from different modalities, v stands for item embeddings. Except for item loss, other loss terms are the same as that used in the monomodality input case.

Study [31] mentioned that the NTM model can extract coherent topics from inputs. Considering that, the NTM part in the proposed MultArtRec can be understood as extracting coherent topics from multimodal features, and item embeddings are further generated from the latent topics. In the CF part, the rating predictions are calculated for recommendation tasks that obtain the inner product of the item embeddings and user embeddings.

5. Experiments

To verify the effectiveness of our proposed method, we conducted different comparative experiments on a public dataset. This section introduces the experimental dataset, experimental settings, and experimental results in detail.

5.1. Dataset

We use two datasets to evaluate our proposed method: (1) WikiArt Emotions Dataset [35], and (2) Amazon Clothing Dataset. Since we do not use all the contents of these two datasets, below we describe how we use these two datasets separately.

5.1.1. WikiArt Emotions Dataset

The dataset we use to evaluate our proposed method is the WikiArt Emotions Dataset. This dataset was created after a survey that asked 307 annotators to annotate 4105 artworks (mainly paintings). The selected artworks are famous artworks selected from 22 categories from the WikiArt.org collection. Each of the categories contains 200 artworks. Each artwork was annotated by at least 10 annotators. The survey asks the annotators to rate (the score is -3 to 3) and tag emotions to the artworks when shown the pictures, the titles, and both pictures and titles. One of the annotation examples is shown in Figure 5 and Table 3.

Table 3. The emotion tags and rating from one annotator for "Dance at the Spring" in original WikiArt Emotions Dataset.

Emotion Tag for Image	Emotion Tag for Title	Emotion Tag for Both Image and Title	Rating for the Artwork
happiness, anticipation	happiness	happiness	1

Figure 5. The image of artwork "Dance at the Spring" in original WikiArt Emotions Dataset.

WikiArt Emotions Dataset includes images and text information about the artworks, as well as the emotion tags and ratings obtained from the annotators. Some studies use WikiArt Emotion Dataset for emotion recognition [36] or style and color representation analysis [37]. Our purpose is to propose a method that can be broadly applicable to artwork recommendation and use the WikiArt Emotion Dataset to evaluate the proposed method. Since emotion tags are rarely collected in existing online artwork services, and emotion tags are not easy to obtain, we do not include emotion tags for the experiments. In addition, our research focuses on the particularity of artwork images and titles. In order to keep the initial experiment simple, we do not include other text information (e.g., artist, genre, year) in the WikiArt Emotion Dataset but only use titles of the artworks. Therefore, we created a new dataset based on the WikiArt Emotions Dataset. It includes 63,425 user–item ratings and 4105 artwork images and titles. An example of the created dataset is shown in Tables 4 and 5.

Table 4. An example of user–item rating in the created dataset for experiments.

Annotator ID	Artwork ID	Rating
18	577284a7edc2cb3880fe813a	1

Table 5. An example of artwork image and title in the created dataset for experiments.

Artwork Image	Artwork Title
	Dance at the Spring

5.1.2. Amazon Clothing Dataset

The Amazon Clothing Dataset is created by the recommender system evaluation framework Cornac [38] (https://github.com/PreferredAI/cornac, accessed on 7 January 2024) based on Amazon product data (http://jmcauley.ucsd.edu/data/amazon/, accessed on 7 January 2024), and can be obtained through functions built in Cornac. This makes

it easier to evaluate and compare the recommendation methods. The original Amazon product data includes various information on the users' purchasing activity. The Amazon Clothing Dataset makes sure that all the items have three types of auxiliary data: text, image, and context (item-item relationships), including data of the user–item interactions, product review text, product image features, item-item relationships, including 5377 users, 3393 products and 13,689 user–item ratings.

5.2. Input Feature Extraction Models

We use pre-trained models or finetuned models to extract image features and text features of the artwork item content in the WikiArt Emotion Dataset, then use these features as the input for the recommendation models. The information on the feature extraction models is summarized in Table 6.

Table 6. Pre-trained models and finetuned models for input feature extraction.

Notation	Base Model	Modality	Feature Dimension	Pre-Training or Finetuning Dataset	Model URL
BERT-base	BERT [10]	Text	768	BooksCorpus and Wikipedia	https://huggingface.co/bert-base-uncased
BERT-emotion	BERT	Text	768	Twitter Sentiment Analysis	https://huggingface.co/bhadresh-savani/bert-base-uncased-emotion
BERT-poem	BERT	Text	768	Poem Sentiment	https://huggingface.co/nickwong64/bert-base-uncased-poems-sentiment
ResNet	ResNet50 [26]	Image	2048	ImageNet	https://www.tensorflow.org/api_docs/python/tf/keras/applications/resnet50/ResNet50
VGG	VGG16 [9]	Image	4096	ImageNet	https://www.tensorflow.org/api_docs/python/tf/keras/applications/vgg16/VGG16

All the model URLs are accessed on 7 January 2024.

For text feature extraction, we chose the commonly used BERT-base model, which is trained on a large corpus and can fit various tasks. The BERT-emotion model and the BERT-poem model are finetuned on sentiment-related short-text datasets, so they are considered better for extracting the title features of artworks. For image feature extraction, we chose the commonly used ResNet model and VGG model, which can fit various tasks.

5.3. Evaluation Metrics

In the comparative experiments, precision@k, recall@k, and NDCG@k are used to evaluate the methods. These metrics are often used to measure whether the recommendation results satisfy the users. Among them, NDCG@k considers the ranking order of the recommendation results. The metrics are stated below:

$$\text{Precision@k} = \frac{\text{hits}}{\text{k}}, \tag{12}$$

$$\text{Recall@k} = \frac{\text{hits}}{\text{relevant items}}, \tag{13}$$

$$\text{NDCG@k} = \frac{\sum_{i=1}^{k\ (\text{actual order})} \frac{\text{hits}}{\log_2(i+1)}}{\sum_{i=1}^{k\ (\text{ideal order})} \frac{\text{hits}}{\log_2(i+1)}}, \tag{14}$$

where k refers to the number of top-ranked recommended items that are retrieved. In the comparative experiments, $\text{k} = \{10, 20, 50\}$. The evaluation metrics emphasize the quality of top-k recommendation results because, in many practical scenarios, the users will only browse part of (often few of) the results provided by the recommendation algorithm. In all

the metrics, the term hits refers to the number of correctly recommended items. In recall@k, the term relevant items refers to the number of liked items for each user in the test data. In NDCG@k, the term actual order refers to the actual recommendation order. The term ideal order refers to the best recommendation order.

5.4. Comparative Experiments

In this section, we conduct experiments on two aforementioned datasets. The results are illustrated in Section 5.4.1 and Section 5.4.2, respectively.

5.4.1. Experiments on WikiArt Emotions Dataset

We use the recommendation algorithms provided in the comparative framework Cornac to compare with our proposed MultArtRec on the WikiArt Emotions Dataset.

For all the experiments in this section, when splitting the training data and the test data, we used the RatioSplit method in Cornac to set the ratio of the training set to the test set as 8:2 and set the same splitting random seed for each experiment. Finally, the training set includes 307 users, 4105 artworks, and 49,603 user–item ratings. The test set includes 307 users, 3751 artworks, and 12,611 user–item ratings. Other experimental settings are stated before each experiment.

The rest of this section illustrates the experiments that use image modality, text modality, and multimodal modality as input, respectively.

Image Modality Input

To verify the effectiveness of MultArtRec on image modality, we compared it with CausalRec, CVAE, and CDL. The input of each model is the artwork image features extracted by ResNet and VGG. The experiment results of using VGG features and ResNet features are shown in Table 7 and Table 8, respectively.

For MultArtRec, CVAE, and CDL, we set the rating threshold as 0.5, which means that only the items with rating predictions larger than 0.5 can be recommended. Because the CausalRec method must use a dataset with ratings of 1 to 5, for CausalRec, we first normalize the ratings from -3 to 3 to 1 to 5 and then set the rating threshold to 3, which is equivalent to 0.5 in the -3 to 3 scope. Because MultArtRec has a similar structure to CVAE and CDL, for comparison, the size of the latent topic layer K in MultArtRec is set as 50, which is the same as the size of latent feature layers in CVAE and CDL. The number of training epochs is set as 300. The remaining parameters remain the default in the Cornac examples.

Table 7. Comparison of our proposed model MultArtRec with CDL, CVAE, and CausalRec using only VGG image features as input.

Metric	CasualRec	CDL	CVAE	MultArtRec
NDCG@50	0.0747	0.0716	0.0708	**0.0773** *
NDCG@20	0.0487	0.0533	0.0554	**0.0617** *
NDCG@10	0.0384	0.0466	0.0508	**0.0588** *
Precision@50	**0.0558** *	0.0284	0.0288	0.0305
Precision@20	**0.0484** *	0.0365	0.0398	0.0427
Precision@10	0.0415	0.0406	0.0462	**0.0515**
Recall@50	0.0975	0.1446	0.1379	**0.1463** *
Recall@20	0.0280	0.0575	0.0556	**0.0618**
Recall@10	0.0129	0.0205	0.0219	**0.0285** *

(1) Bold and underlined results show the best performance for each metric when inputting VGG image features.
(2) Results with asterisks show the best performance for each metric across all kinds of image feature inputs.

Table 8. Comparison of our proposed model MultArtRec with CDL, CVAE, and CausalRec using only ResNet image features as input.

Metric	CasualRec	CDL	CVAE	MultArtRec
NDCG@50	0.0524	0.0693	0.0709	**0.0766**
NDCG@20	0.0349	0.0547	0.0553	**0.0607**
NDCG@10	0.0286	0.0513	0.0508	**0.0572**
Precision@50	**0.0418**	0.0277	0.0288	0.0314
Precision@20	0.0346	0.0386	0.0396	**0.0431**
Precision@10	0.0320	0.0482	0.0462	**0.0521** *
Recall@50	0.0607	0.1312	0.1380	**0.1425**
Recall@20	0.0226	0.0562	0.0556	**0.0621** *
Recall@10	0.0114	0.0224	0.0219	**0.0257**

(1) Bold and underlined results show the best performance for each metric when inputting ResNet image features.
(2) Results with asterisks show the best performance for each metric across all kinds of image feature inputs.

From the results shown in the tables above, we can see that CausalRec+VGG performs the best in precision@50 and precision@20 when using VGG features, and CausalRec+ResNet performs the best in precision@50 when using ResNet features, MultArtRec is optimal in the remaining metrics. However, CausalRec+ResNet performs worse than either MultArtRec+VGG or MultArtRec+ResNet on average. This shows that MultArtRec is more robust than CausalRec in extracting effective features from different image features. Moreover, most of the best results across all metrics are achieved when using VGG features, indicating that VGG features are more appropriate for this recommendation task. Compared with CausalRec, MultArtRec can at most improve 174.8% with ResNet features on recall@20.

Text Modality Input

As for the verification of MultArtRec's effectiveness on text modality, we compare it with CVAE and CDL. The input of each model is the text features extracted from artwork titles by BERT-base, BERT-emotion, and BERT-poem. The experiment results of using three kinds of features are shown in Table 9, Table 10, and Table 11, respectively. The setting of the rating threshold, the latent topic layer size K, and the training epochs are the same as the image-only modality experiments.

Table 9. Comparison of our proposed model MultArtRec with CDL and CVAE using only BERT-base text features as input.

Metric	CDL	CVAE	MultArtRec
NDCG@50	0.0746	0.0711	**0.0780** *
NDCG@20	**0.0620** *	0.0555	0.0609
NDCG@10	0.0550	0.0509	**0.0575** *
Precision@50	0.0304	0.0289	**0.0316**
Precision@20	**0.0427**	0.0398	0.0416
Precision@10	0.0465	0.0462	**0.0512**
Recall@50	0.1393	0.1386	**0.1442** *
Recall@20	**0.0637**	0.0556	0.0605
Recall@10	**0.0238**	0.0219	0.0230

(1) Bold and underlined results show the best performance for each metric when inputting BERT-base text features.
(2) Results with asterisks show the best performance for each metric across all kinds of text feature inputs.

From the results shown in the tables above, we can see that sometimes CDL outperforms MultArtRec, especially when using BERT-emotion features as the input. However, MultArtRec has the best performance across all kinds of text feature inputs. This illustrates that compared with CDL, MultArtRec can extract better feature extraction capabilities

when using appropriate inputs. Compared with CDL, MultArtRec can, at most, improve 10.7% with BERT-poem features on precision@50.

Table 10. Comparison of our proposed model MultArtRec with CDL and CVAE using only BERT-emotion text features as input.

Metric	CDL	CVAE	MultArtRec
NDCG@50	**0.0737**	0.0712	0.0720
NDCG@20	**0.0607**	0.0556	0.0596
NDCG@10	**0.0575**	0.0509	0.0553
Precision@50	**0.0300**	0.0290	0.0292
Precision@20	**0.0404**	0.0399	0.0403
Precision@10	**0.0479**	0.0462	0.0475
Recall@50	0.1269	**0.1387**	0.1338
Recall@20	0.0596	0.0557	**0.0640** *
Recall@10	**0.0290**	0.0219	0.0283

(1) Bold and underlined results show the best performance for each metric when inputting BERT-emotion text features. (2) Results with asterisks show the best performance for each metric across all kinds of text feature inputs.

Table 11. Comparison of our proposed model MultArtRec with CDL and CVAE using only BERT-poem text features as input.

Metric	CDL	CVAE	MultArtRec
NDCG@50	0.0718	0.0712	**0.0740**
NDCG@20	**0.0584**	0.0554	0.0583
NDCG@10	**0.0573**	0.0509	0.0554
Precision@50	0.0289	0.0290	**0.0320** *
Precision@20	0.0403	0.0396	**0.0436** *
Precision@10	**0.0518** *	0.0462	0.0505
Recall@50	0.1287	0.1387	**0.1406**
Recall@20	**0.0576**	0.0556	0.0544
Recall@10	**0.0300** *	0.0219	0.0245

(1) Bold and underlined results show the best performance for each metric when inputting BERT-poem text features. (2) Results with asterisks show the best performance for each metric across all kinds of text feature inputs.

From the above two experiments, we can see that an obvious drawback of the CVAE model is that for the same modality, no matter what input features are used, its performances are almost the same. This is also the limitation we found and mentioned in the last section. By introducing a new topic loss to ensure the symmetric structure in MultArtRec, the model obviously overcomes this limitation.

Multimodal Modality Input

In the above experiments, we revealed the comparison results of our proposed MultArtRec and several existing models when taking in monomodal input. Next, we test the performance of the proposed MultArtRec on multimodal data. The experiment results for NDCG@k, precision@k and recall@k are shown in Table 12, Table 13 and Table 14, respectively. In order to compare the inputs of different modalities, the previous experiments on image features and text features are also placed in the corresponding tables.

The setting of the recommendation threshold and the latent topic layer size K are the same as the image-only modality and text-only modality experiments. As mentioned in Section 4, Equation (11), when using multimodal data as input, there are multiple NTM parts in the proposed framework. Each NTM part extracts latent topics of one modality. In this experiment, we set the bias of the text modality item loss as one and the bias of the image modality item loss as 0.2.

Table 12. The performance of the proposed MultArtRec in text modality, image modality, and multimodality on NDCG@k metric.

Input Feature	Modality	NDCG@50	NDCG@20	NDCG@10
VGG	image	0.0773	0.0617	0.0588
ResNet	image	0.0739	0.0609	0.0547
BERT-base	text	**0.0780**	0.0609	0.0575
BERT-emotion	text	0.0720	0.0596	0.0553
BERT-poem	text	0.0740	0.0583	0.0554
VGG+BERT-base	image + text	0.0704	0.0607	0.0572
VGG+BERT-emotion	image + text	0.0777	0.0617	0.0572
VGG+BERT-poem	image + text	0.0726	0.0583	0.0559
ResNet+BERT-base	image + text	0.0717	0.0607	0.0589
ResNet+BERT-emotion	image + text	0.0768	**0.0624**	**0.0634**
ResNet+BERT-poem	image + text	0.0744	0.0619	0.0569

Bold and underlined results show the best performance on NDCG@k when inputting different features to MultArtRec.

Table 13. The performance of the proposed MultArtRec in text modality, image modality, and multimodality on precision@k metric.

Input Feature	Modality	Precision@50	Precision@20	Precision@10
VGG	image	0.0305	0.0427	0.0515
ResNet	image	0.0288	0.0413	0.0469
BERT-base	text	0.0316	0.0416	0.0512
BERT-emotion	text	0.0292	0.0403	0.0475
BERT-poem	text	0.0320	0.0436	0.0505
VGG+BERT-base	image + text	0.0302	0.0442	0.0495
VGG+BERT-emotion	image + text	0.0325	0.0432	0.0525
VGG+BERT-poem	image + text	0.0315	0.0417	0.0482
ResNet+BERT-base	image + text	0.0317	0.0437	**0.0531**
ResNet+BERT-emotion	image + text	**0.0329**	0.0437	0.0525
ResNet+BERT-poem	image + text	0.0316	**0.0452**	0.0502

Bold and underlined results show the best performance on precision@k when inputting different features to MultArtRec.

Table 14. The performance of the proposed MultArtRec in text modality, image modality, and multimodality on recall@k metric.

Input Feature	Modality	Recall@50	Recall@20	Recall@10
VGG	image	**0.1463**	0.0618	0.0285
ResNet	image	0.1384	**0.0652**	0.0247
BERT-base	text	0.1442	0.0605	0.0230
BERT-emotion	text	0.1338	0.0640	0.0283
BERT-poem	text	0.1406	0.0544	0.0245
VGG+BERT-base	image + text	0.1196	0.0568	0.0287
VGG+BERT-emotion	image + text	0.1433	0.0643	0.0270
VGG+BERT-poem	image + text	0.1310	0.0529	0.0278
ResNet+BERT-base	image + text	0.1214	0.0552	0.0206
ResNet+BERT-emotion	image + text	0.1356	0.0522	**0.0339**
ResNet+BERT-poem	image + text	0.1374	0.0620	0.0312

Bold and underlined results show the best performance on recall@k when inputting different features to MultArtRec.

From the experimental results above, we can see that for precision and all the @10 metrics, MultArtRec using multimodal data always performs the best. This demonstrates that the multimodal feature can express richer user preferences, allowing it to have a more accurate expression in vector space, and MultArtRec can effectively extract such richly

expressed features. MultArtRec using image monomodal data performs well on recall when k is large. This shows that in the WikiArt Emotions Dataset, when k increases, the image features extracted by MultArtRec are more representative of a user's general preferences. Compared with utilizing monomodal inputs, utilizing multimodal inputs can improve up to 15.9% on NDCG@10.

Moreover, we found that although MultArtRec performs better on many metrics when using the image features extracted by VGG than using image features extracted by ResNet, MultArtRec performs better on the multimodal feature that includes the image feature extracted by ResNet. This may be because the feature dimension of VGG is much larger than the dimension of the text feature, which makes the information unbalanced. One solution is to finetune the parameters to reduce the impact of the image modal input on the overall loss. Another solution is to design a better feature fusion module to solve this problem.

All the experiments show that the topic loss we propose is effective. When using monomodal input, compared with the CVAE model that has similar architecture but without topic loss, MultArtRec can better utilize appropriate monomodal features, thus obtaining better recommendation results. When using multimodal input, the loss function with our proposed topic loss can extract latent topics that are better than monomodal in recommendation tasks.

5.4.2. Experiments on Amazon Clothing Dataset

We also use the recommendation algorithms provided in the comparative framework Cornac to compare with our proposed MultArtRec on the Amazon Clothing Dataset. The purpose of conducting experiments with Amazon Clothing Dataset is to test the versatility of our proposed MultArtRec method on another recommendation scenario besides artwork recommendation. The preliminary experiments are conducted with image modality of product items, utilizing user–item interactions and product image features.

For the experiments in this section, we also use the RatioSplit method in Cornac with the same random seed, to split the training data and the test data with the ratio of 8:2. Finally, the training set includes 5081 users, 3326 products and 10,951 user–item ratings. The test set includes 1868 users, 1453 products, and 2200 user–item ratings.

In this experiment, the comparison models we selected include CDL and CVAE, which have a similar structure to our model, as well as CasualRec, visual matrix factorization (VMF) [39] and visual Bayesian personalized ranking (VBPR) [27], which perform well on image-based recommendation tasks. For all the models, the rating threshold is set as 3 (the ratings range from 1 to 5) and trained for 100 epochs. The latent feature layer of CDL and CVAE and the latent topic layer K in MultArtRec are all set to 200. The other parameters of CasualRec, VMF, and VBPR remain the default in the Cornac examples. The evaluation methods include precision@k, recall@k and NDCG@k, k $= \{10, 20, 50\}$. The experimental results are shown in Table 15.

Table 15. Comparison of our proposed model MultArtRec with other methods using Amazon Clothing Dataset's image features as input.

Metric	CasualRec	VMF	VBPR	CDL	CVAE	MultArtRec
NDCG@50	0.0065	0.0297	0.0611	0.0665	0.0619	**0.0685**
NDCG@20	0.0031	0.0245	0.0497	0.0587	0.0550	**0.0598**
NDCG@10	0.0018	0.0194	0.0416	0.0493	0.0473	**0.0515**
Precision@50	0.0006	0.0020	0.0037	0.0043	0.0039	**0.0044**
Precision@20	0.0005	0.0035	0.0061	0.0084	0.0079	**0.0086**
Precision@10	0.0004	0.0048	0.0084	0.0126	0.0122	**0.0134**
Recall@50	0.0245	0.0851	0.1583	0.1843	0.1695	**0.1918**
Recall@20	0.0081	0.0591	0.1032	0.1460	0.1367	**0.1496**
Recall@10	0.0031	0.0400	0.0717	0.1107	0.1075	**0.1181**

Bold and underlined results show the best performance for each metric.

From the results, we can see that MultArtRec ranks the best across all the metrics. This proves the versatility of MultArtRec. Besides artworks, it can also extract effective image features that represent user preferences from the pictures of the products.

6. Conclusions and Future Work

In this study, we proposed an NTM-based multimodal recommender system, MultArtRec, for rating prediction for user–item pairs. It can theoretically effectively extract latent topics for the recommendation task from any item content that can be expressed by vectors.

For the verification of the effectiveness of our proposed MultArtRec on real-world data, the public WikiArt Emotions Dataset and Amazon Clothing Dataset are adopted.

In the experiments conducted on the WikiArt Emotions Dataset, the first two experiments verify the performance of MultArtRec on monomodal item content. The results show that for monomodal input, MultArtRec has better results on many metrics than a state-of-the-art image-based recommender system, CausalRec, and two other similar architectured multimodal recommendation methods, CDL and CVAE. The third experiment explores whether MultArtRec can perform better when utilizing multimodal input rather than monomodal input. This experiment compares the performances of MultArtRec itself on two kinds of image modal features, three kinds of text modal features, and six combined multimodal features by the image modal features and text modal features. The experimental results show that MultArtRec can achieve better results when utilizing multimodal features than only utilizing monomodal features. All the experiment results prove that both the architecture and the loss function with topic loss of MultArtRec are effective.

In the experiments conducted on Amazon Clothing Dataset, the preliminary results show that MultArtRec outperforms those other models, which illustrates MultArtRec's versatility to some degree. However, more experiments on other modalities and datasets are needed to prove the real versatility of our proposed method.

Two main limitations of MultArtRec have also been discovered. The first limitation is that when utilizing monomodal input, the performance of MultArtRec is not always optimal; e.g., when utilizing the text monomodal input, CDL outperforms MultArtRec in some of the metrics. Understanding the reasons requires analyzing why CDL has good feature extraction effects for certain text features in future work. The second limitation is that when utilizing multimodal data, MultArtRec cannot make good use of the advantageous features that perform well in monomodal experiments, e.g., MultArtRec performs well with VGG features in monomodal experiments, but the experimental results of using VGG features in multimodal experiments are not ideal. In order to solve this problem, it is necessary to design a better model structure for feature fusion.

In the future, in addition to image modality and text modality, we will also consider verifying whether other modal information is effective in improving the performance of MultArtRec. We will first consider the tags of the artwork and the description text.

As stated in [40], the recommendation interface (e.g., the layouts of recommendations) is an important element for recommender systems. Article [41] emphasizes that the present aesthetics of the recommendation results will affect the recommendation effect. Study [42] proposed a novel deep-learning-based approach to both evaluate and optimize the recommendation interface. In our research, it is considered that an ideal artwork recommender system in practice is not only for finding artworks that match the users' interests but also to provide users with better access to information and aesthetic experience. Therefore, in the future development of the proposed artwork recommender system, we will focus on the interface design of the recommender system and use advanced technology to evaluate it.

Author Contributions: Conceptualization, J.W., A.M., and K.K.; methodology, J.W.; software, J.W.; writing—original draft preparation, J.W.; writing—review and editing, A.M. and K.K.; supervision, A.M. and K.K.; funding acquisition, A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by JSPS KAKENHI Grant Numbers 20K12567 and 23K11780.

Data Availability Statement: Provide original public dataset URL, and re-created dataset for the experiments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Strezoski, G.; Fijen, L.; Mitnik, J.; László, D.; Oyens, P.D.M.; Schirris, Y.; Worring, M. TindART: A personal visual arts recommender. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020.
2. Messina, P.; Cartagena, M.; Cerda-Mardini, P.; del Rio, F.; Parra, D. Curatornet: Visually-aware recommendation of art images. *arXiv* **2020**, arXiv:2009.04426.
3. Pal, A.; Eksombatchai, C.; Zhou, Y.; Zhao, B.; Rosenberg, C.; Leskovec, J. Pinnersage: Multi-modal user embedding framework for recommendations at pinterest. In Proceedings of the 26th ACM SIGKDD, Virtual, 6–10 July 2020.
4. Deldjoo, Y.; Nazary, F.; Ramisa, A.; Mcauley, J.; Pellegrini, G.; Bellogin, A.; Di Noia, T. A review of modern fashion recommender systems. *arXiv* **2022**, arXiv:2202.02757.
5. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 1 July 2021.
6. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-shot text-to-image generation. In Proceedings of the International Conference on Machine Learning, Virtual, 1 July 2021.
7. OpenAI. GPT-4 technical report. *arXiv* **2023**, arXiv:2303.08774v3.
8. Truong, Q.T.; Lauw, H. Multimodal review generation for recommender systems. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019.
9. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
10. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
11. Trattner, C.; Elsweiler, D. Food recommender systems: Important contributions, challenges and future research directions. *arXiv* **2017**, arXiv:1711.02760.
12. Abbasi-Moud, Z.; Vahdat-Nejad, H.; Sadri, J. Tourism recommendation system based on semantic clustering and sentiment analysis. *Expert Syst. Appl.* **2021**, *167*, 114324. [CrossRef]
13. Guy, I. People recommendation on social media. In *Social Information Access: Systems and Technologies*; Springer: New York, NY, USA, 2018; pp. 570–623.
14. Urdaneta-Ponte, M.C.; Mendez-Zorrilla, A.; Oleagordia-Ruiz, I. Recommendation systems for education: Systematic review. *Electronics* **2021**, *10*, 1611. [CrossRef]
15. De Croon, R.; Van Houdt, L.; Htun, N.N.; Štiglic, G.; Vanden Abeele, V.; Verbert, K. Health recommender systems: Systematic review. *J. Med. Internet Res.* **2021**, *23*, e18035. [CrossRef]
16. Sayeb, Y.; Jebri, M.; Ghezala, H.B. A graph based recommender system for managing COVID-19 Crisis. *Procedia Comput. Sci.* **2022**, *196*, 348–355. [CrossRef]
17. Adday, B.N.; Shaban, F.A.J.; Jawad, M.R.; Jaleel, R.A.; Zahra, M.M.A. Enhanced vaccine recommender system to prevent COVID-19 based on clustering and classification. In Proceedings of the IEEE International Conference on Engineering and Emerging Technologies (ICEET), Istanbul, Turkey, 27–28 October 2021.
18. Aroyo, L.M.; Wang, Y.; Brussee, R.; Gorgels, P.; Rutledge, L.W.; Stash, N. Personalized museum experience: The Rijksmuseum use case. In Proceedings of the Museums and the Web, San Francisco, CA, USA, 11–14 April 2007.
19. Deladiennee, L.; Naudet, Y. A graph-based semantic recommender system for a reflective and personalised museum visit. In Proceedings of the 12th IEEE International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Bratislava, Slovakia, 9–10 July 2017.
20. Frost, S.; Thomas, M.M.; Forbes, A.G. Art I don't like: An anti-recommender system for visual art. In Proceedings of the Museums and the Web, Boston, MA, USA, 2–6 April 2019.
21. Qiu, R.; Wang, S.; Chen, Z.; Yin, H.; Huang, Z. Causalrec: Causal inference for visual debiasing in visually-aware recommendation. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021.
22. Messina, P.; Dominguez, V.; Parra, D.; Trattner, C.; Soto, A. Content-based artwork recommendation: Integrating painting metadata with neural and manually-engineered visual features. *User Model. User-Adapt. Interact.* **2019**, *29*, 251–290. [CrossRef]
23. Yilma, B.A.; Leiva, L.A. The Elements of Visual Art Recommendation: Learning Latent Semantic Representations of Paintings. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023.
24. Wang, H.; Wang, N.; Yeung, D.Y. Collaborative deep learning for recommender systems. In Proceedings of the 21th ACM SIGKDD, New York, NY, USA, 10–13 August 2015.
25. Li, X.; She, J. Collaborative variational autoencoder for recommender systems. In Proceedings of the 23rd ACM SIGKDD, Halifax, NS, Canada, 13–17 August 2017.

26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE CVPR, Las Vegas, NV, USA, 26 June–1 July 2016.
27. He, R.; McAuley, J. VBPR: Visual bayesian personalized ranking from implicit feedback. In Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
28. Tang, J.; Du, X.; He, X.; Yuan, F.; Tian, Q.; Chua, T.S. Adversarial training towards robust multimedia recommender system. *IEEE Trans. Knowl. Data Eng.* **2019**, *5*, 855–867. [CrossRef]
29. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
30. Wu, X.; Nguyen, T.; Luu, A.T. A survey on neural topic models: Methods, applications, and challenges. *Res. Sq. Prepr.* **2023**. [CrossRef]
31. Ding, R.; Nallapati, R.; Xiang, B. Coherence-Aware Neural Topic Modeling. *arXiv* **2018**, arXiv:1809.02687.
32. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
33. Zheng, Y.; Zhang, Y.J.; Larochelle, H. Topic modeling of multi-modal data: An autoregressive approach. In Proceedings of the IEEE CVPR, Columbus, OH, USA, 23–28 June 2014.
34. Hörster, E.; Lienhart, R.; Slaney, M. Image retrieval on large-scale image databases. In Proceedings of the 6th ACM International Conference on Image and Video Retrieval, Amsterdam, The Netherlands, 9 July 2007.
35. Mohammad, S.; Kiritchenko, S. Wikiart emotions: An annotated dataset of emotions evoked by art. In Proceedings of the 11th International Conference on Language Resources and Evaluation, Miyazaki, Japan, 7–12 May 2018.
36. Tashu, T.M.; Hajiyeva, S.; Horvath, T. Multimodal emotion recognition from art using sequential co-attention. *J. Imaging* **2021**, *7*, 157. [CrossRef]
37. Srinivasa Desikan, B.; Shimao, H.; Miton, H. WikiArtVectors: Style and color representations of artworks for cultural analysis via information theoretic measures. *Entropy* **2022**, *24*, 1175. [CrossRef]
38. Truong, Q.T.; Salah, A.; Lauw, H. Multi-modal recommender systems: Hands-on exploration. In Proceedings of the 15th ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September–1 October 2021.
39. Park, C.; Kim, D.; Oh, J.; Yu, H. Do "also-viewed" products help user rating prediction? In Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 3–7 April 2017.
40. Fayyaz, Z.; Ebrahimian, M.; Nawara, D.; Ibrahim, A.; Kashef, R. Recommendation systems: Algorithms, challenges, metrics, and business opportunities. *Appl. Sci.* **2020**, *10*, 7748. [CrossRef]
41. Sulikowski, P.; Kucznerowicz, M.; Bąk, I.; Romanowski, A.; Zdziebko, T. Online Store Aesthetics Impact Efficacy of Product Recommendations and Highlighting. *Sensors* **2022**, *22*, 9186. [CrossRef]
42. Sulikowski, P.; Zdziebko, T. Deep learning-enhanced framework for performance evaluation of a recommending interface with varied recommendation position and intensity based on eye-tracking equipment data processing. *Electronics* **2020**, *9*, 266. [CrossRef]