



Modality-specific and shared generative adversarial network for cross-modal retrieval

Fei Wu^{a,*}, Xiao-Yuan Jing^b, Zhiyong Wu^a, Yimu Ji^c, Xiwei Dong^a, Xiaokai Luo^a, Qinghua Huang^d, Ruchuan Wang^c

^a College of Automation, Nanjing University of Posts and Telecommunications, Nanjing, China

^b School of Computer, Wuhan University, Wuhan, China

^c College of Computer, Nanjing University of Posts and Telecommunications, Nanjing, China

^d School of Science, Nanjing University of Posts and Telecommunications, Nanjing, China

ARTICLE INFO

Article history:

Received 10 June 2019

Revised 2 March 2020

Accepted 12 March 2020

Available online 14 March 2020

Keywords:

Cross-modal retrieval

Generative adversarial networks (GAN)

Modality-specific feature learning

Modality-shared feature learning

ABSTRACT

Cross-modal retrieval aims to realize accurate and flexible retrieval across different modalities of data, e.g., image and text, which has achieved significant progress in recent years, especially since generative adversarial networks (GAN) were used. However, there still exists much room for improvement. How to jointly extract and utilize both the **modality-specific** (complementarity) and **modality-shared** (correlation) features effectively has not been well studied. In this paper, we propose an approach named **Modality-Specific and Shared Generative Adversarial Network** (MS²GAN) for cross-modal retrieval. The network architecture consists of two sub-networks that aim to learn modality-specific features for each modality, followed by a common sub-network that aims to learn the modality-shared features for each modality. Network training is guided by the adversarial scheme between the generative and discriminative models. The generative model learns to predict the semantic labels of features, model the inter- and intra-modal similarity with label information, and ensure the difference between the modality-specific and modality-shared features, while the discriminative model learns to classify the modality of features. The learned modality-specific and shared feature representations are jointly used for retrieval. Experiments on three widely used benchmark multi-modal datasets demonstrate that MS²GAN can outperform state-of-the-art related works.

© 2020 Elsevier Ltd. All rights reserved.

1. Introduction

Nowadays, we are facing with big data in the form of multi-modal data, e.g., image, text, video, audio, etc., in Internet or our daily life. The presence of this huge repository of multi-modal data has greatly spurred the demand of cross-modal retrieval for search engine or digital library, such as searching concerned images with a text query or searching related videos with a query audio. Different from the traditional single-modal retrieval task [1,2], e.g., image retrieval, that requires the query and retrieval results to have the same modality, cross-modal retrieval [3–5] is a more flexible application that provides the query of any modality to find relevant information with the same or different modalities.

Since data of different modalities usually has inconsistent distributions and representations, the similarity of data from differ-

ent modalities cannot be directly measured. In other words, there exists a modality gap to be bridged [6,7]. In recent years, to deal with the modality gap, a lot of cross-modal retrieval methods have been presented. Traditional methods are mainly devoted to learning linear projections to explore the correlation in data of different modalities, like canonical correlation analysis (CCA)-based methods [8]. With the fast progress of deep learning technologies, deep neural network (DNN)-based methods [9–11] have been mainstream for bridging the gap across modalities. Recently, inspired by the advantages of generative adversarial networks (GAN) to model data distribution and learn discriminative representation, it has been introduced into the field of cross-modal retrieval, and some GAN-based works have been developed [12–16].

1.1. Motivation and contribution

Although several cross-modal retrieval methods have been presented, there still exists much room for improvement. Most of existing methods focus on modality-shared information exploration,

* Corresponding author.

E-mail addresses: wufei_8888@126.com (F. Wu), jingxy_2000@126.com (X.-Y. Jing).

and map data of different modalities into a common space to obtain the common representation, without taking the exploration and utilization of modality-specific information into consideration. Peng et al., [17] developed a method termed modality-specific cross-modal similarity measurement (MCSM), which pays attention to the issue of modality-specific characteristics exploration. However, how to effectively extract and utilize both the modality-specific (complementarity) and modality-shared (correlation) features jointly for retrieval has not been well studied.

Inspired by Hu et al. [18] that aims to learn deep metric from multi-view data, which learns not only an individual distance metric for each view to preserve its characteristics but also a sharable representation for multiple views to retain the common properties, in this paper, we propose a novel approach named Modality-Specific and Shared Generative Adversarial Network (MS²GAN) for cross-modal retrieval. The contribution of our study can be summarized as following three points:

- (1) To jointly learn modality-specific and shared feature representations, we adopt two feed-forward sub-networks to learn modality-specific features for each modality, followed by a common sub-network to learn modality-shared features for each modality. The learned modality-specific feature representations are combined with the shared feature representations for retrieval. **To our knowledge, this is the first work that leverages both modality-specific and shared features for cross-modal retrieval.**
- (2) We adopt the adversarial scheme for network training. The generative model learns to predict the semantic labels of the integration of modality-specific and shared features, model both the inter- and intra-modal similarity with label information, and ensure the difference between the modality-specific and modality-shared features, such that features are semantically discriminative in both inter- and intra-modal aspects and the complementarity information in multi-modal data can be effectively explored. The discriminative model learns to classify the modality of modality-shared features for promoting modality-invariance.
- (3) We evaluate MS²GAN on three widely used benchmark multi-modal datasets Wikipedia [8], NUS-WIDE-10k [9], and XMedia [7]. And the experimental results demonstrate that our approach can outperform related state-of-the-art works.

1.2. Organization

The rest of the paper is organized as follows. In Section 2, we briefly introduce the related work. We describe our MS²GAN approach in Section 3. The experimental results are reported in Section 4. And conclusions are drawn in Section 5.

2. Related work

2.1. Non-GAN-based cross-modal retrieval methods

Traditional cross-modal retrieval methods pursue the common representations of different modalities usually by learning linear projections [19–22]. Based on canonical correlation analysis (CCA) [23], a series of algorithms have been developed for the cross-modal retrieval. Rasiwasia et al. [8] incorporated the semantic labels into CCA to learn correlation between two modalities. Cross-modal factor analysis (CFA) [24] minimizes the distances between the common representations of pairwise data. Wang et al. [25] presented a method called learning coupled feature spaces (LCFS), which employs coupled linear regression to learn projections for mapping multi-modal data into a common space, and takes the issue of feature selection into consideration.

In recent years, due to the strong nonlinear representation learning ability, deep learning has made great advance for cross-modal retrieval. Correspondence autoencoder (Corr-AE) [9] combines autoencoder and correlation costs for jointly performing representation learning and correlation learning. Peng et al. [10] presented a cross-media multiple deep network (CMDN) algorithm, which models the intra- and inter-media information to get separate representation of each media type, and then hierarchically combines the inter- and intra-media representations to get the shared representations. Cross-modal correlation learning (CCL) [26] is an extension of CMDN, which learns separate representation by jointly optimizing the intra- and inter-modality correlation, and then adopts a multi-task learning strategy to balance the intra-modality semantic category constraints and the inter-modality pair-wise similarity constraints, and then performs multi-grained fusion. Wei et al. [27] provided a deep semantic matching (Deep-SM) algorithm and validated the superiority of convolutional neural network (CNN) visual features for cross-modal retrieval. Cross-modal hybrid transfer network (CHTN) [11] adopts a modal-sharing transfer sub-network to transfer knowledge to both two modalities and a layer-sharing correlation sub-network to further adapt to the retrieval task. Modality-specific cross-modal similarity measurement (MCSM) [17] learns independent semantic spaces for different modalities, where the modality-specific characteristics are explored by attention mechanism in the process of cross-modal correlation learning. Wu et al. [28] presented a cross-modal online similarity function learning algorithm, which provides an online similarity function learning framework that aims to learn the metric being able to well explore the cross-modal semantic relation. Taking the intra-class low-rank structure into consideration, Kang et al. [29] developed a deep semantic space learning model with intra-class low-rank constraint for cross-modal retrieval. Scalable deep multimodal learning (SDML) [30] predefines a common subspace, and then it independently trains a set of modality-specific networks for multiple modalities (one network for each modality) to transform data from different modalities into the predefined common subspace for realizing multimodal learning. To make the learned common representations be discriminative and modality-invariant, deep supervised cross-modal retrieval (DSCMR) [31] simultaneously minimizes the discrimination loss (in the common representation space and the label space) and the modality invariance loss.

There exist large differences between these works and our approach. These works (except MCSM) mainly focus on common representation learning, and MCSM mainly pays attention to the modality-specific characteristics exploration. Contrastively, our MS²GAN approach focuses on the exploration and utilization of both the modality-specific and shared features.

In addition, there exist several works that focus on the task of visual question answering (VQA) [32]. Given an image and a corresponding open-ended, natural language question about the image, VQA tries to provide an accurate natural language answer. Anderson et al. [33] presented a combined bottom-up and top-down attention mechanism, and applied this mechanism to image captioning and VQA. Zhao et al. [34] studied the multi-turn video question answering (VideoQA) problem, and developed the hierarchical attention context reinforced network for this task. Focusing on the problem of VideoQA, Jin et al. [35] designed a knowledge-based progressive spatial-temporal attention network. Fan et al. [36] presented a heterogeneous memory enhanced multimodal attention mechanism for the VideoQA task. Zhao et al. [37] adopted a dynamic hierarchical reinforced network for the challenging long-form VideoQA application. There exist differences between VQA/VideoQA and cross-modal retrieval tasks: (1) for VQA, besides the visual content, the additional knowledge is needed for generating accurate reasoning, while for cross-modal retrieval, the effort

is mainly made to explore the discriminant and common content across modalities; (2) VQA may focus more on the local features, while cross-modal retrieval mainly focuses on the global features in different modalities. In this paper, we mainly study the problem of retrieval issue across different modalities.

2.2. Generative adversarial network (GAN)-based methods

The framework of GAN presented by Goodfellow et al. [38] includes two parts: a generator and a discriminator. These two parts have opposite training goals. A generator aims to generate samples that cannot be distinguished by the discriminator, while a discriminator is learned to be able to discriminate the generated samples. Nowadays, some GAN-based methods have arisen for cross-modal retrieval. Adversarial cross-modal retrieval (ACMR) [12] trains the network with a minimax game. Specifically, ACMR requires the feature projector to generate modality-invariant and discriminative representations and requires the modality classifier to detect the modality of feature representation. Cross-modal generative adversarial networks (CM-GANs) [16] presents the cross-modal convolutional autoencoders as the generative model and uses two discriminative models to conduct intra- and inter-modality discrimination. Modal-adversarial semantic learning network (MASLN) [13] adopts the conditional auto-encoder to realize cross-modal reconstruction, and adopts the adversarial learning mechanism to capture semantic correlation and discrimination of classes in the common space. Deep adversarial metric learning (DAML) [14] maps data of different modalities into a shared nonlinear subspace and uses a modality classifier to predict the modality of transformed features. Modal-adversarial hybrid transfer network (MHTN) [15] adopts a modal-sharing knowledge transfer subnetwork to transfer knowledge from single-modal source domain to cross-modal target domain, and uses a modal-adversarial semantic learning subnetwork to construct an adversarial mechanism between common representation generator and modality discriminator.

As analyzed in Section 1.1, these methods mainly focus on modality-shared information exploration and neglect the exploration and utilization of modality-specific information. Different from them, our approach extracts and utilizes both the modality-specific and modality-shared features jointly. In addition, recently, a bundle of hashing methods [39–44] has been developed for cross-modal retrieval. **Different from these works, our study mainly aims to learn useful real-valued feature representations rather than the binary code.**

3. Our approach

3.1. Problem formulation

Let $O = \{o_n, l_n\}_{n=1}^N$ be a collection of N instances of image-text pairs, where each instance $o_n = (i_n, t_n)$ includes an image feature vector $i_n \in \mathbb{R}^{d_i}$ and a text feature vector $t_n \in \mathbb{R}^{d_t}$, usually $d_i \neq d_t$. The feature matrices for the image and text modalities are represented as $I = [i_1, \dots, i_N]$ and $T = [t_1, \dots, t_N]$, respectively. $l_n = [l_{n1}, \dots, l_{nC}]^T$ denotes the semantic label vector corresponding to o_n , where C is the number of semantic categories. If o_n is from the c^{th} class, $l_{nc} = 1$, otherwise $l_{nc} = 0$. To explore the modality-specific characteristics, we use two **3-layer feed-forward sub-networks** to learn nonlinear feature representations $I_f = f_I(I; \theta_I) = [i_n^f]_{n=1}^N \in \mathbb{R}^{d_f \times N}$ and $T_f = f_T(T; \theta_T) = [t_n^f]_{n=1}^N \in \mathbb{R}^{d_f \times N}$ for image and text modalities, where $f_I(i_n; \theta_I)$ and $f_T(t_n; \theta_T)$ are the mapping functions of two modalities. To discover the correlation between modalities, we further adopt a common 2-layer sub-network to map I_f and T_f into a common space for learning modality-shared feature representations, by $I_s = s(I_f; \vartheta) = [i_n^s]_{n=1}^N \in \mathbb{R}^{d_s \times N}$ and $T_s = s(T_f; \vartheta) = [t_n^s]_{n=1}^N \in$

$\mathbb{R}^{d_s \times N}$, where $s(o_n^f; \vartheta)$ is the common mapping function, and $d_f = d_s$.

Generally, we have the following three objectives: (1) **the modality-specific and shared features can be effectively explored and used**; (2) **the modality gap of pairwise modality-shared features from different modalities should be reduced**; and (3) **the learned features should be semantically discriminative**. We train the network with the adversarial mechanism between the generative and discriminative models. Fig. 1 shows the overall architecture of our MS²GAN approach.

3.2. Generative model

3.2.1. Label prediction

To make the generated features to be semantically discriminative, a feed-forward 1-layer sub-network activated by Softmax is deployed as a classifier, such that when the concatenation $[i_u^f; i_u^s]$ or $[t_u^f; t_u^s]$ is input, the corresponding probability distribution of semantic categories, i.e., $\hat{p}_u([i_u^f; i_u^s])$ or $\hat{p}_u([t_u^f; t_u^s])$, can be output. Based on the probability distribution, we define the following **semantic discrimination loss**

$$L_{sd}(\theta_I, \theta_T, \vartheta, \nu) = -\frac{1}{U} \sum_{u=1}^U (l_u (\log \hat{p}_u([i_u^f; i_u^s]) + \log \hat{p}_u([t_u^f; t_u^s]))) \quad (1)$$

where ν denotes the parameters of the classifier, and U represents the number of instances in each mini-batch.

3.2.2. Inter- and intra-modal semantic similarity modeling

Motivated by the idea of deep metric learning [45], we compute the **Euclidean distance** between features to measure their similarity, and require that the similarity of features with the same semantic category is enhanced and that of features with different semantic categories is reduced. For each pair of instances o_u and o_v , the distance between their features is defined as

$$d_c(u, v) = \|i_u^f - i_v^f\|_2^2 + \|t_u^f - t_v^f\|_2^2 + \|i_u^s - i_v^s\|_2^2 \quad (2)$$

which not only describes the intra-modal distance between modality-specific features but also characterizes the inter-modal distance between modality-shared features. We thus provide the following **contrastive loss**

$$L_c(\theta_I, \theta_T, \vartheta) = \frac{1}{|E|} \sum_{(u,v) \in E} h(d_c(u, v) - \tau) + \frac{1}{|D|} \sum_{(u,v) \in D} h(\tau - d_c(u, v)) \quad (3)$$

where $h(x) = \max(0, x)$ is the hinge loss function, $E = \{(u, v)\} (D = \{(u, v)\})$ is a pairwise index set of feature pairs having the same (different) semantic labels in each mini-batch, $|\cdot|$ represents the size of the set, and τ is a positive threshold. The objective (3) imposes constraint on the distance $d_c(u, v)$, such that the distance of features for a within-class pair is smaller than the threshold τ , while the distance for a between-class pair is larger than τ .

3.2.3. Modality-specific and shared features distinguishment

There should exist differences between the modality-specific feature and the corresponding modality-shared feature for a specific image or text. We provide the following large margin loss to perform discrimination of these two types of features, such that the complementarity information in data of different modalities can be effectively learned

$$L_{lm}(\theta_I, \theta_T, \vartheta) = \frac{1}{U} \sum_{u=1}^U (h(\zeta - d_{lm}^i(u, u)) + h(\zeta - d_{lm}^t(u, u))) \quad (4)$$

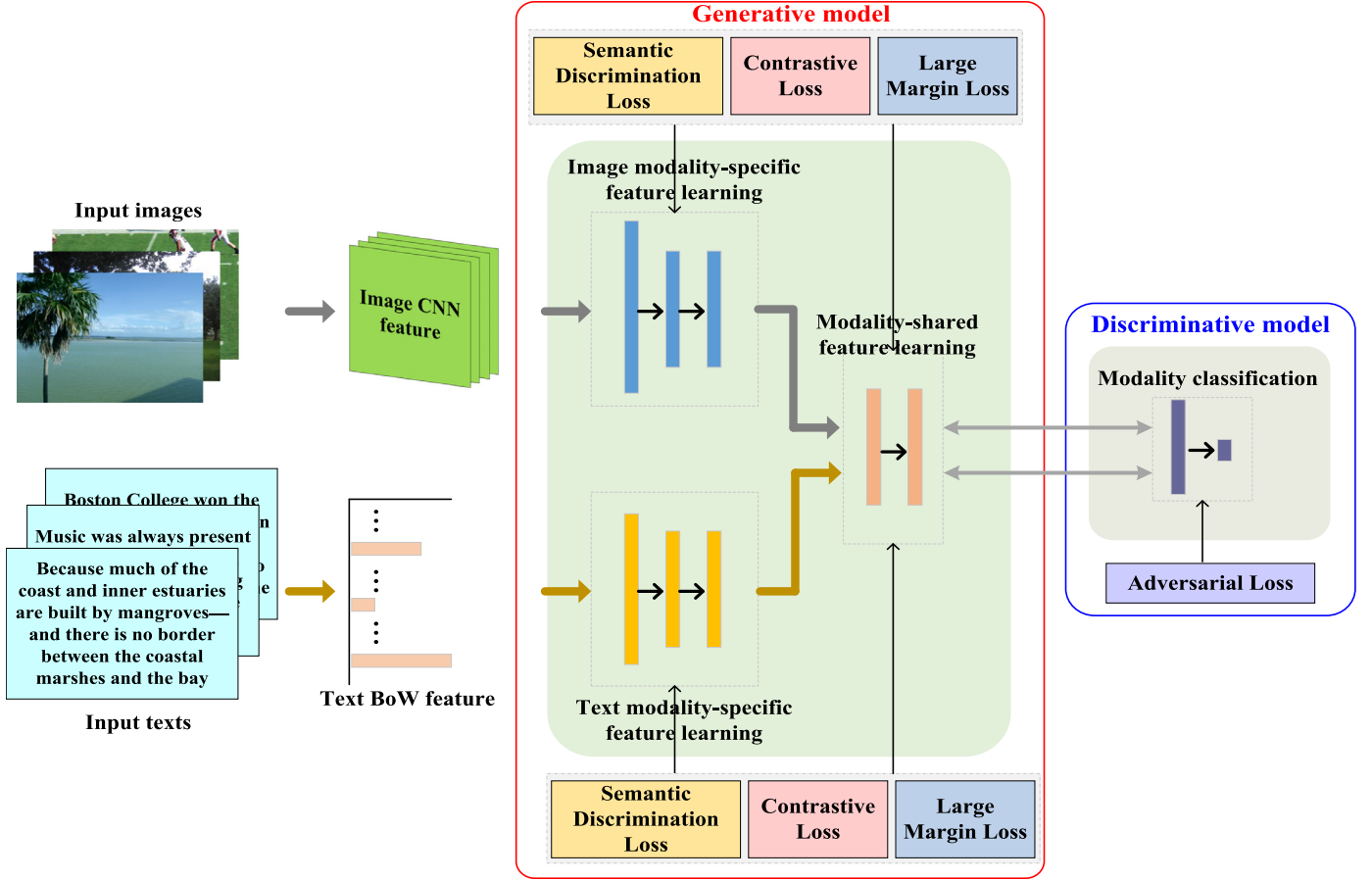


Fig. 1. The overall framework of our MS²GAN approach. Two feed-forward sub-networks are employed to learn modality-specific features for image and text modalities, followed by a common sub-network to learn the modality-shared features for each modality. Adversarial training is performed between the generative and discriminative models. The generative model embeds the semantic label information (with the semantic discrimination loss), models inter- and intra-modal semantic similarity (with the contrastive loss), and promotes differences between the modality-specific and modality-shared features (with the large margin loss), for generating discriminative modality-shared (modality-invariant) and modality-specific representations, while the discriminative model distinguishes modality-shared features with respect to the modalities. The learned modality-shared representations corresponding to the image and text modalities will be input into the modality classifier for recognizing the modality information, and the discriminative model will direct/influence the feature learning of the generative model.

where $h(x) = \max(0, x)$, ζ is a positive threshold, distances $d_{lm}^i(u, u)$ and $d_{lm}^t(u, u)$ are defined as

$$d_{lm}^i(u, u) = \|i_u^f - i_u^s\|_2^2 \text{ and } d_{lm}^t(u, u) = \|t_u^f - t_u^s\|_2^2 \quad (5)$$

Eq. (4) imposes constraint on the distances $d_{lm}^i(u, u)$ and $d_{lm}^t(u, u)$ with the threshold ζ , requiring the distances $d_{lm}^i(u, u)$ and $d_{lm}^t(u, u)$ to be larger than ζ , which acts as a feature component discriminator, i.e., distinguishing the modality-specific features from the modality-shared features.

By combining the semantic discrimination loss, the contrastive loss and the large margin loss, we can obtain the total loss function of the generative model

$$L_{gm}(\theta_I, \theta_T, \vartheta, \nu) = L_{sd} + \alpha L_c + \beta L_{lm} \quad (6)$$

where α and β balance three terms.

3.3. Discriminative model

To reduce the modality gap, we build a modality classifier with a 2-layer sub-network, as shown in Fig. 1, which acts as an adversary. The goal of the classifier is to recognize the modality information of an unknown modality-shared feature representation. We assign each item of an instance a one-hot modality label vector to indicate which modality it belongs to. The adversarial loss is

formulated as

$$L_{adv}(\theta_A) = -\frac{1}{U} \sum_{u=1}^U (g_u (\log A(i_u^s; \theta_A) + \log(1 - A(t_u^s; \theta_A)))) \quad (7)$$

where $A(\cdot; \theta_A)$ is the generated modality probability of feature representation $i_u^s(t_u^s)$, θ_A denotes the parameters of the modality classifier, and g_u is the ground-truth modality label for each item (i_u or t_u) of instance o_u . With this cross-entropy loss term, we can explicitly reduce heterogeneity gap between modalities.

3.4. Optimization

The optimal feature representations can be obtained by jointly minimizing the loss functions of generative and discriminative models. Considering that the optimization goals of these two models are opposite, the min-max game [38] is employed for optimization with the following two concurrent sub-processes

$$(\hat{\theta}_I, \hat{\theta}_T, \hat{\vartheta}, \hat{\nu}) = \arg \min_{\theta_I, \theta_T, \vartheta, \nu} (L_{gm}(\theta_I, \theta_T, \vartheta, \nu) - L_{adv}(\hat{\theta}_A)) \quad (8)$$

$$(\hat{\theta}_A) = \arg \max_{\theta_A} (L_{gm}(\hat{\theta}_I, \hat{\theta}_T, \hat{\vartheta}, \hat{\nu}) - L_{adv}(\theta_A)) \quad (9)$$

The min-max game can be implemented through stochastic gradient descent algorithm. Following [12,14], we also add a Gradient Reversal Layer (GRL) before the first layer of the modality classifier

Algorithm 1 Optimization of MS²GAN.

1. **Input:** Image and text features $\{i_n\}_{n=1}^N$ and $\{t_n\}_{n=1}^N$, and the corresponding semantic label and modality label sets $\{l_n\}_{n=1}^N$ and $\{g_n\}_{n=1}^N$.
2. **Update until convergence**
 - (a) Separately update $\theta_I, \theta_T, \vartheta, \nu$ by descending their stochastic gradients with the learning rate of r :

$$\theta_I \leftarrow \theta_I - r \cdot \nabla_{\theta_I} \frac{1}{U} (L_{gm} - L_{adv}),$$

$$\theta_T \leftarrow \theta_T - r \cdot \nabla_{\theta_T} \frac{1}{U} (L_{gm} - L_{adv}),$$

$$\vartheta \leftarrow \vartheta - r \cdot \nabla_{\vartheta} \frac{1}{U} (L_{gm} - L_{adv}),$$

$$\nu \leftarrow \nu - r \cdot \nabla_{\nu} \frac{1}{U} (L_{gm} - L_{adv}).$$
 - (b) Update θ_A by ascending its stochastic gradients with GRL:

$$\theta_A \leftarrow \theta_A + r \cdot \nabla_{\theta_A} \frac{1}{U} (L_{gm} - L_{adv}).$$
3. **Output:** Optimized parameters $\theta_I, \theta_T, \vartheta$, modality-specific features $\{i_n^f\}_{n=1}^N$ and $\{t_n^f\}_{n=1}^N$ and the modality-shared features $\{i_n^s\}_{n=1}^N$ and $\{t_n^s\}_{n=1}^N$.

to perform min-max optimization. Algorithm 1 briefly summarizes the optimization process.

4. Experiments

4.1. Datasets and features

In this paper, we evaluate our approach on three widely used datasets, i.e., Wikipedia [8], NUS-WIDE-10k [9], and XMedia [7].

- Wikipedia consists of 2866 image/text pairs, and each pair is labeled with one of 10 categories. Following [9,15,17], the dataset is divided into 3 parts: 2,173 pairs used for training, 231 pairs for validation, and 462 pairs for testing.

- NUS-WIDE-10k is a subset of the NUS-WIDE dataset [46]. NUS-WIDE-10k consists of 10,000 image/text pairs that are evenly selected from the 10 largest categories of NUS-WIDE. Following [9,15,17], this dataset is also split into three subsets: the training set with 8000 pairs, the validation set with 1000 pairs, and the testing set with 1000 pairs.

- XMedia is a cross-modal dataset with five modalities, i.e., text, image, video, audio and 3D model. There are 20 classes, and 250 image/text pairs, 25 videos, 50 audio clips, and 25 3D models for each class. Following [15], the dataset is split into three parts: training set with 9600 instances, validation set with 1200 instances, and testing set with 1200 instances. In a specific cross-modal retrieval task, the instances of the corresponding modalities are used. For example, when only image and text modalities are used in experiment, following [15], the training set with 4000 image-text pairs, validation set with 500 pairs, and the testing set with 500 pairs are used.

For these three datasets, they are randomly split into the training, validation and testing sets. To eliminate the affect of randomness, we performed 10 random running to report the average results.

On these three datasets, the image features extracted with convolutional neural network (CNN) are employed to represent images, and the bag of words (BoW) features are used to represent texts. Specifically, for image modality, following [12,14,26], 4,096-dimensional features extracted by the fc7 layer of VGGNet are used for Wikipedia and NUS-WIDE-10k; and the 4,096-dimensional CNN feature provided by the original authors are used for XMe-

dia¹. For text, following [12,15], the 3,000-dimensional, 1,000-dimensional, and 3,000-dimensional BoW features are separately used for Wikipedia, NUS-WIDE-10k, and XMedia datasets. For video on XMedia, following [15], the features of 4096 dimensions extracted by the C3D model [47] pre-trained on Sports1M [48] are used. For audio on XMedia, the 78-dimensional features extracted by the jAudio [49] are used. And for 3D model on XMedia, the 4,700-dimensional vectors of a LightField descriptor set [50] are used.

4.2. Compared methods and evaluation metric

We compare our approach with three types of related methods (totally 12 state-of-the-art methods): 1) traditional cross-modal retrieval methods: CCA [8], CFA [24], and LCFS [25]; 2) deep learning-based (non-GAN-based) methods: Corr-AE [9], CMDN [10], Deep-SM [27], MCSM [17], SDML [30], and DSCMR [31]; 3) GAN-based methods: ACMR [12], DAML [14], and MHTN [15]. For compared methods with totally the same experimental setting as in our experiment, we reported their experimental results with the published results in the original papers to obtain a fair evaluation. In the same time, we also run the methods with the codes provided by the authors or our own implementation to validate the reported results. We were able to obtain the retrieval results matching the published results. For compared methods that were run with different settings, we run the codes provided by the authors or our own implementation with the setting in our experiment to report their results.

In experiment, we mainly focus on two cross-modal retrieval tasks, namely image-to-text (img2txt) and text-to-image (txt2img) retrieval. We compute the cross-modal similarity between features of each image (text) and all texts (images), and evaluate the ranking list by mean average precision (MAP). MAP can be calculated by computing the mean value of average precision (AP) for all queries. For a query of a specific modality, the AP value can be defined as

$$AP = \frac{1}{R} \sum_{k=1}^{N_{te}} \left(\frac{R_k}{k} \times rel_k \right) \quad (10)$$

where N_{te} denotes the number of instances in the test set, R is the number of relevant instances in test set, and R_k denotes the number of relevant instances in the top k returned results. $rel_k = 1$ if the k^{th} returned result is relevant, otherwise, $rel_k = 0$. MAP jointly considers the issues of precision and the ranking of returned retrieval results, which is a widely used evaluation metric for cross-modal retrieval.

For our MS²GAN approach, when the modality-specific and shared feature representations for images and texts in the test set are obtained, we concatenate these two types of features for each image (text) for retrieval.

4.3. Implementation details

The details of each layer and the corresponding dimension are as follows: we separately use a 3-layer sub-network (with three fully connected layers) with dimension [1024, 512, 128] to learn modality-specific feature representations for the image and text modalities. To learn modality-shared feature representations, a 2-layer sub-network with dimension [128, 128] is used. For label prediction, one layer with dimension of the number of semantic categories is used. And the activation function of *tanh* function is used after each layer. For modality classification, two layers with dimension [64, 2] activated by the *LReLU* function are used. In addition,

¹ <http://59.108.48.34/tiki/XMediaNet/>.

Table 1

Cross-modal retrieval results in terms of MAP (mean \pm standard deviation) of compared methods and our approach on three datasets.

Method	Wikipedia			NUS-WIDE-10k		
	img2txt	txt2img	ave	img2txt	txt2img	ave
CCA [8]	0.258 \pm 0.008	0.250 \pm 0.007	0.254	0.202 \pm 0.007	0.220 \pm 0.006	0.211
CFA [24]	0.334 \pm 0.007	0.297 \pm 0.009	0.316	0.400 \pm 0.009	0.299 \pm 0.008	0.350
LCFS [25]	0.455 \pm 0.012	0.398 \pm 0.010	0.427	0.383 \pm 0.012	0.346 \pm 0.010	0.365
Corr-AE [9]	0.402 \pm 0.010	0.395 \pm 0.013	0.399	0.366 \pm 0.009	0.417 \pm 0.010	0.392
CMDN [10]	0.488 \pm 0.007	0.427 \pm 0.008	0.458	0.492 \pm 0.011	0.515 \pm 0.009	0.504
Deep-SM [27]	0.458 \pm 0.016	0.345 \pm 0.014	0.402	0.389 \pm 0.012	0.496 \pm 0.015	0.443
MCSM [17]	0.516 \pm 0.009	0.458 \pm 0.008	0.487	0.543 \pm 0.005	0.541 \pm 0.005	0.542
SDML [30]	0.522 \pm 0.006	0.488 \pm 0.005	0.505	0.551 \pm 0.003	0.538 \pm 0.005	0.545
DSCMR [31]	0.521 \pm 0.025	0.478 \pm 0.019	0.499	0.552 \pm 0.002	0.542 \pm 0.005	0.547
ACMR [12]	0.518 \pm 0.009	0.412 \pm 0.009	0.465	0.544 \pm 0.008	0.538 \pm 0.008	0.541
DAML [14]	0.559 \pm 0.014	0.481 \pm 0.013	0.520	0.512 \pm 0.010	0.534 \pm 0.009	0.523
MHTN [15]	0.541 \pm 0.008	0.461 \pm 0.009	0.501	0.552 \pm 0.002	0.541 \pm 0.005	0.547
MS ² GAN	0.601 \pm 0.008	0.500 \pm 0.006	0.551	0.556 \pm 0.001	0.548 \pm 0.003	0.552

Method	XMedia		
	img2txt	txt2img	ave
CCA [8]	0.257 \pm 0.005	0.341 \pm 0.006	0.299
CFA [24]	0.292 \pm 0.006	0.283 \pm 0.005	0.288
LCFS [25]	0.529 \pm 0.009	0.475 \pm 0.010	0.502
Corr-AE [9]	0.450 \pm 0.010	0.437 \pm 0.008	0.444
CMDN [10]	0.794 \pm 0.004	0.805 \pm 0.005	0.800
Deep-SM [27]	0.822 \pm 0.006	0.807 \pm 0.005	0.815
MCSM [17]	0.876 \pm 0.005	0.872 \pm 0.005	0.874
SDML [30]	0.889 \pm 0.002	0.906 \pm 0.004	0.898
DSCMR [31]	0.876 \pm 0.003	0.902 \pm 0.003	0.889
ACMR [12]	0.863 \pm 0.006	0.860 \pm 0.005	0.862
DAML [14]	0.810 \pm 0.007	0.836 \pm 0.006	0.823
MHTN [15]	0.853 \pm 0.008	0.843 \pm 0.008	0.848
MS ² GAN	0.894 \pm 0.003	0.911 \pm 0.003	0.903

Softmax activation is added after the last layer of label prediction and modality classification parts.

In the training process of our MS²GAN approach, the batch size is set as 128 on three datasets. We tune the hyper-parameters (the thresholds τ and ζ in (3) and (4), and balance factors α and β in (6)) using grid search. The search range for τ and ζ is [1, 10] with the step length of 1, and the search range for α and β is [0.01, 100] with 10 times per step. Specifically, they are set as $\tau = \zeta = 4$, $\alpha = 1$, and $\beta = 0.1$.

We implement our approach with the TensorFlow framework. And the experiments are performed on a PC with a i7-8700k 3.7GHz CPU and a single NVIDIA GeForce GTX 1080Ti GPU.

4.4. Results

Table 1 reports the cross-modal retrieval results in terms of MAP score of our approach and 12 competing methods on three datasets, where “ave” denotes the mean value of img2txt and txt2img results. From the table, generally, the GAN-based methods including ACMR, DAML, MHTN and our approach, achieve relatively better retrieval results than most of the other compared traditional cross-modal retrieval methods and early deep learning-based (non-GAN-based) methods. On all three datasets, our MS²GAN approach achieves the highest MAP scores on both retrieval tasks. Specifically, MS²GAN improves the MAP scores at least by 0.042=(0.601-0.559) for img2txt and 0.012=(0.500-0.488) for txt2img on Wikipedia, by 0.004=(0.556-0.552) for img2txt and 0.006=(0.548-0.542) for txt2img on NUS-WIDE-10k, and by 0.005=(0.894-0.889) for img2txt and 0.005=(0.911-0.906) for txt2img on XMedia. The reasons for the performance improvement mainly lie in the following two aspects: (1) modality-specific (complementarity) and modality-shared (correlation) features are jointly explored and leveraged effectively for the retrieval task, (2)

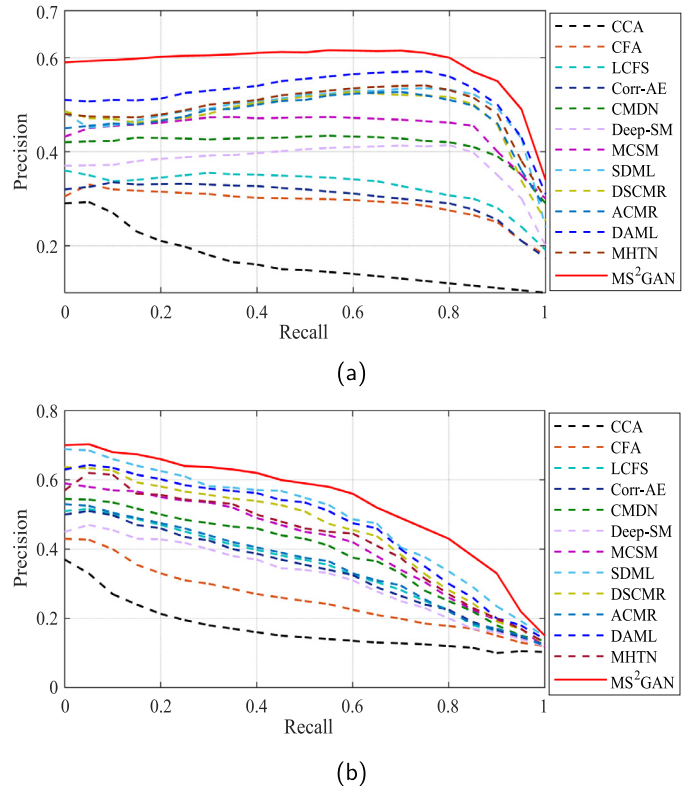


Fig. 2. Precision-recall curves of cross-modal retrieval on Wikipedia dataset, where (a) and (b) separately denote the Image \rightarrow Text and Text \rightarrow Image retrieval tasks.





Task	Query	Top 5 Results				
Image to Text	 Class: Biology	Specimens of "Archaeopteryx" were most notable for their well-developed flight feathers. They were markedly asymmetrical and showed the structure of flight feathers in modern birds, with vanes given stability by a barb-barbule-barbicle arrangement.	"Banksia spinulosa" var. "spinulosa" was introduced into cultivation in the United Kingdom in 1788 by Joseph Banks who supplied seed to Kew, Cambridge Botanic Gardens and Woburn Abbey among others; var. "collina" followed in 1800 and var. "cunninghamii" in 1822.	Seabirds, by virtue of living in a geologically depositional environment (that is, in the sea where sediments are readily laid down), are well represented in the fossil record. They are first known to occur in the Cretaceous Period, the earliest being the Hesperornithiformes	Fruiting bodies typically grow in clusters, and are found on dead or decaying wood, or on woody fragments in cow or horse dung. Dung-loving (coprophilous) species include "C. stercoreus", "C. costatus", "C. fimicola", and "C. pygmaeus". Brodie pp. 102–103.	Based on fossil and biological evidence, most scientists accept that birds are a specialised sub-group of theropod dinosaurs. More specifically, they are members of Maniraptora, a group of theropods which includes dromaeosaurs and oviraptorids, among others.
	 Class: Sport	The 2006–07 season represented a watershed year for the Raptors franchise. The roster was overhauled, including the selection of 2006 NBA Draft number one pick Andrea Bargnani, the acquisition of point guard T. J. Ford in exchange for Charlie Villanueva	The race started at 14:00 China Standard Time (UTC+8), and was scheduled to last until 17:30. The weather was unseasonably cool. It was cloudy and the roads were dry at the start of the race. The clouds brought steady rainfall midway through the race, making conditions challenging.	At the 1984 Winter Olympics in Sarajevo, Yugoslavia, the Soviet Union won its sixth gold medal. Czechoslovakia and Sweden won the silver and bronze medals. The 1988 Winter Olympics were held in Calgary, Alberta, Canada, where the Soviet team captured its seventh and final gold medal.	Because Trott's illness precluded his selection for the 1899 Australian team to England, the Australian captaincy passed to Joe Darling. On 8 May 1899, Trott was committed to the Kew Asylum, a psychiatric hospital in the eastern suburbs of Melbourne.	The 1995 race featured 22 mushers, of whom 13 finished.Saari, Matias. , "Fairbanks Daily News-Miner". February 6, 2008. Accessed February 27, 2009. Budget problems caused the first prize to drop by 25% to \$15,000
Text to Image	Dakota tribes, mostly the Mdewakanton, as early as the 16th century were known as permanent settlers near their sacred site of St. Anthony Falls. New settlers arrived during the 1850s and 1860s in Minneapolis from New England, New York, and Canada, and during the mid-1860s, Scandinavians from Sweden, Finland, Norway and Denmark began to call the city home. Class: Geography					
	"Constitution" entered dry dock in 1992 for what had been planned as an inspection and minor repair period but turned out to be her most comprehensive structural restoration and repair since she was launched in 1797. Over the 200 years of her career, as her mission changed from a fighting warship to a training ship and eventually a receiving ship, multiple refittings removed most of her original construction components and design. Class: Warfare					

Fig. 3. Examples of retrieval results by using our MS²GAN on the Wikipedia dataset. Each row shows the top five matches for an image/text query.

the inter- and intra-modal discrimination and the inter-modal invariance is elaborately modeled.

Fig. 2 shows the precision-recall curves of the img2txt and txt2img tasks on Wikipedia. From the figure, our MS²GAN can keep clear advantage as compared with other competing methods on all recall levels. Similar results can be also found on the other two datasets. These results indicate the superiority of our proposed approach.

Statistical test is a mathematical tool for analyzing the difference between the results of different methods. To statistically analyze the detailed results (MAP results of 10 random running) corresponding to Table 1, we also perform the Wilcoxon sign-rank test [51] at a confidence level of 95% with a Bonferroni correction factor, i.e., dividing the significance level by the number of compared methods. The Wilcoxon sign-rank test [51] is a non-parametric test, which ranks the differences in performances of two methods, ignoring the signs, and compares the ranks for the positive and the negative differences. We make a hypothesis that there is no significant difference between our MS²GAN and per competing method. Then, if the p-value is below 0.004(=0.05/12), the performance difference between our MS²GAN and each compared method can be considered to be statistically significant. Here, "12" denotes the number of compared methods. The statistical test results (p-value) are given in Table 2. From the table, almost all the p-values are smaller than 0.004, which indicates that MS²GAN makes a significant difference comparing with other competing methods.

Fig. 3 shows the examples of image(text) queries and the top five texts(images) retrieved by our MS²GAN approach. We can see that our approach can find the corresponding matches of the text/image modality with the same semantic label for the given image(text) query.

4.5. Discussion

4.5.1. Evaluation of components in MS²GAN

In this subsection, we evaluate the important components of MS²GAN. We separately call the version of MS²GAN without the semantic discrimination loss as MS²GAN-sd, the version of MS²GAN without the contrastive loss as MS²GAN-c, and the version of MS²GAN without the large margin loss as MS²GAN-lm. In addition, when the modality-specific and shared features of test instances are obtained, we perform cross-modal retrieval with only the modality-specific features or modality-shared features. We call these two cross-modal retrieval manners as MSGAN1 and MSGAN2. Table 3 shows the result comparison.

From the table, the MAP scores of MS²GAN-sd, MS²GAN-c and MS²GAN-lm are obviously inferior to those of the complete version of MS²GAN on three datasets. These results mean that the designed semantic discrimination loss, contrastive loss and the large margin loss promote semantically discriminative feature learning from both intra- and inter-modality aspects, and are useful for the cross-modal retrieval task. Furthermore, the MAP scores of MSGAN2 is slightly inferior to those of MS²GAN and the MAP scores of MSGAN1 are not so bad, which means that the modality-specific features are in fact beneficial to retrieval task to some extent.

4.5.2. Evaluation of retrieval performance with other modalities

Besides the image and text modalities, we also evaluate the retrieval performance of our approach with other modalities, i.e., video, audio, and 3D model, on the XMedia dataset [15]. reports the state-of-the-art retrieval results on XMedia dataset with all five modalities. And thus, MHTN [15] is chosen for comparison. Table 4 tabulates the cross-modal retrieval results in terms of MAP score of our approach and MHTN on XMedia dataset. We can see from the

Table 2

Statistical test results between MS²GAN and other methods for detailed cross-modal retrieval results on three datasets.

Method	Wikipedia		NUS-WIDE-10k		XMedia	
	img2txt	txt2img	img2txt	txt2img	img2txt	txt2img
CCA [8]	0.002	0.002	0.002	0.002	0.002	0.002
CFA [24]	0.002	0.002	0.002	0.002	0.002	0.002
LCFS [25]	0.002	0.002	0.002	0.002	0.002	0.002
Corr-AE [9]	0.002	0.002	0.002	0.002	0.002	0.002
CMDN [10]	0.002	0.002	0.002	0.002	0.002	0.002
Deep-SM [27]	0.002	0.002	0.002	0.002	0.002	0.002
MCSM [17]	0.002	0.002	0.002	0.002	0.002	0.002
SDML [30]	0.002	0.002	0.0059	0.0039	0.002	0.002
DSCMR [31]	0.002	0.0039	0.002	0.002	0.002	0.002
ACMR [12]	0.002	0.002	0.002	0.002	0.002	0.002
DAML [14]	0.002	0.002	0.002	0.002	0.002	0.002
MHTN [15]	0.002	0.002	0.002	0.002	0.002	0.002

Table 3

MAP scores of variants of MS²GAN.

Method	Wikipedia			NUS-WIDE-10k		
	img2txt	txt2img	average	img2txt	txt2img	average
MS ² GAN-sd	0.205	0.184	0.195	0.195	0.181	0.188
MS ² GAN-c	0.447	0.386	0.417	0.429	0.413	0.421
MS ² GAN-lm	0.559	0.459	0.509	0.511	0.500	0.506
MSGAN1	0.569	0.418	0.494	0.513	0.454	0.484
MSGAN2	0.596	0.483	0.540	0.548	0.534	0.541
MS ² GAN	0.601	0.500	0.551	0.556	0.548	0.552

Method	XMedia		
	img2txt	txt2img	average
MS ² GAN-sd	0.244	0.320	0.282
MS ² GAN-c	0.819	0.825	0.822
MS ² GAN-lm	0.860	0.874	0.867
MSGAN1	0.837	0.855	0.846
MSGAN2	0.882	0.902	0.892
MS ² GAN	0.894	0.911	0.903

Table 4

Cross-modal retrieval results (in terms of MAP) of MHTN and our MS²GAN with five modalities.

Task	MHTN	MS ² GAN
Image → Text	0.853	0.894
Image → Video	0.753	0.749
Image → Audio	0.730	0.738
Image → 3D Model	0.803	0.807
Text → Image	0.843	0.911
Text → Video	0.696	0.707
Text → Audio	0.689	0.696
Text → 3D Model	0.733	0.734
Video → Image	0.725	0.716
Video → Text	0.699	0.702
Video → Audio	0.632	0.624
Video → 3D Model	0.659	0.653
Audio → Image	0.694	0.698
Audio → Text	0.667	0.673
Audio → Video	0.599	0.590
Audio → 3D Model	0.614	0.613
3D Model → Image	0.697	0.705
3D Model → Text	0.678	0.684
3D Model → Video	0.589	0.588
3D Model → Audio	0.607	0.600
Average	0.698	0.704

table that our MS²GAN can achieve better retrieval results in most cases, leaving a small number of cases where the retrieval results of our approach are slightly inferior to (but are still comparable to) those of MHTN.

4.5.3. Visualization of the learned representation through adversarial learning

To further investigate the effectiveness of the learned feature representations, we employ the t-SNE tool [52] to embed the samples/features of image and text modalities into the two-dimensional space for visualization. Fig. 4 (a), (b) and (c) separately show the distributions of original image samples, i.e., the samples represented by 4,096-dimensional VGGNet features for image, the distribution of original text samples, i.e., the samples represented by 3,000-dimensional BoW features for text, and the distribution of learned feature representations, i.e., the learned modality-shared feature representations, on the Wikipedia dataset. We can observe that the distributions of image and text modalities are largely different and the samples with different class labels are not well separated in the original sample space. On the contrary, the distribution of learned feature representations shows that the distributions of image and text modalities are better mixed together, and the representations are generally separated into several semantically clusters. There also exist a small number of feature representations from different semantic classes being mixed together, which may bring irrelevant retrieval results. As a summary, this comparison indicates that the adversarial loss can effectively reduce the gap between modalities and the designed loss in generative model can make the features be semantically discriminative generally.

4.5.4. Convergence analysis

Fig. 5 shows the development of the loss values of generative and the discriminative models in the training process on the Wikipedia dataset. From the figure, the loss of generative model

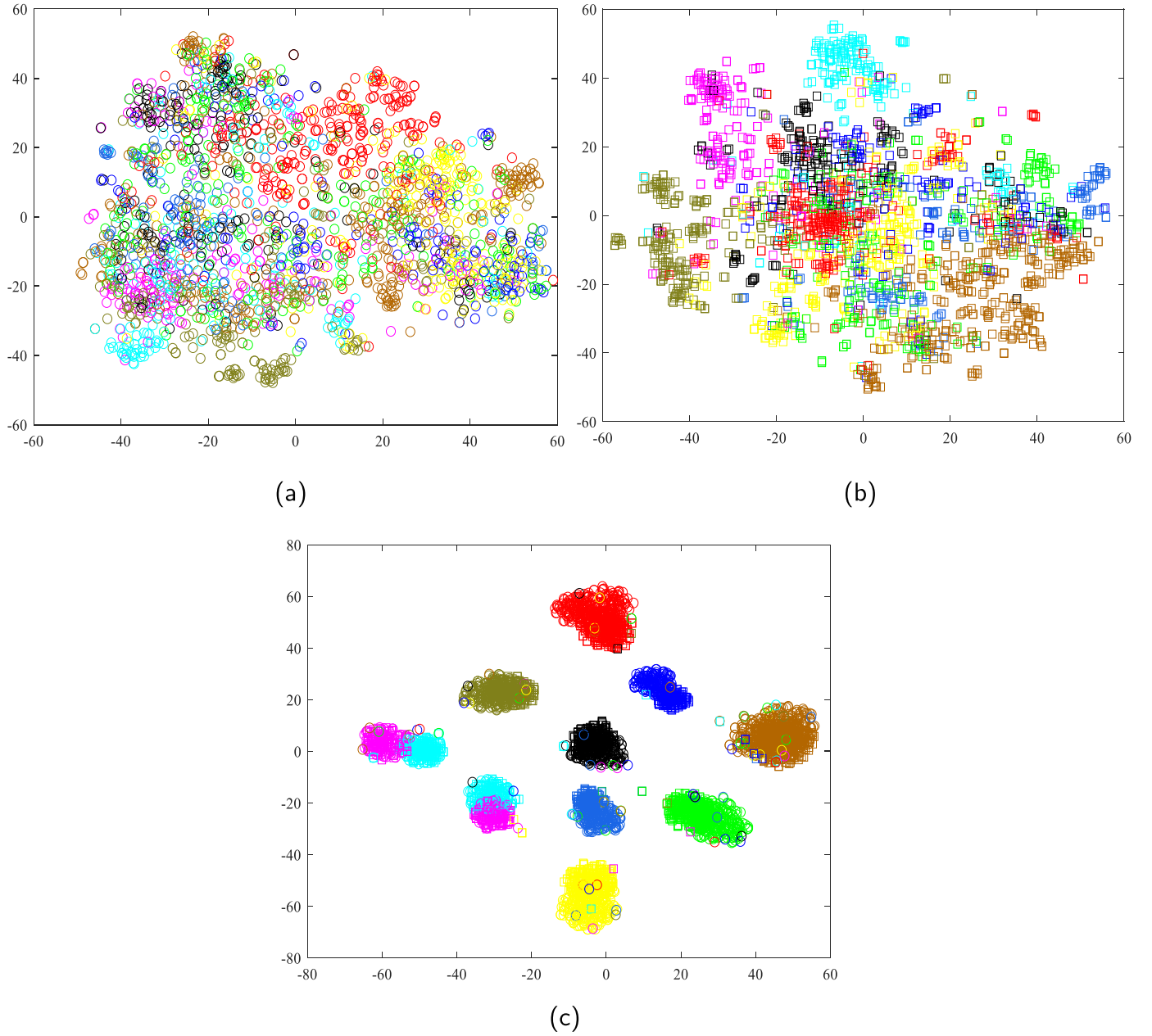


Fig. 4. t-SNE visualization of data on Wikipedia dataset, where (a) and (b) show the distributions of original image and text samples, and (c) shows the distribution of learned feature representations. In the figure, circles and squares separately denote features of image and text modalities. And different colors denote features from different semantic classes.

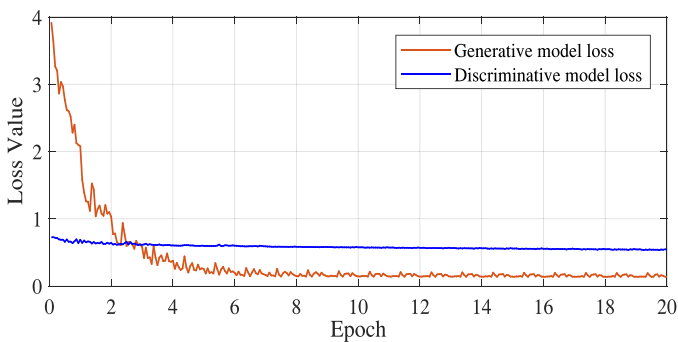


Fig. 5. Convergence curves of the generative and discriminative model losses in MS²GAN on the Wikipedia dataset.

shows a downward trend and can converge with about 10 epochs. The adversarial loss experiences slight fluctuation in the initial 4 epochs and then stabilizes. These results are in accordance with our expectation, since if the adversarial loss value continuously keeps increase, the modality classifier does not work well and it fails to direct/influence the feature learning process of the generator; and if the adversarial loss continuously decreases to zero, it means that the discriminator wins the minimax game and the generator fails to generate the modality-invariant feature representations. On the NUS-WIDE-10k and XMedia datasets, the similar convergence results can also be found.

4.5.5. Parameter analysis

Lastly, we discuss the sensitivity of our approach to different values of hyper-parameters τ , ζ , α and β . Fig. 6 shows the average MAP scores (the mean value of img2txt and txt2img results) versus

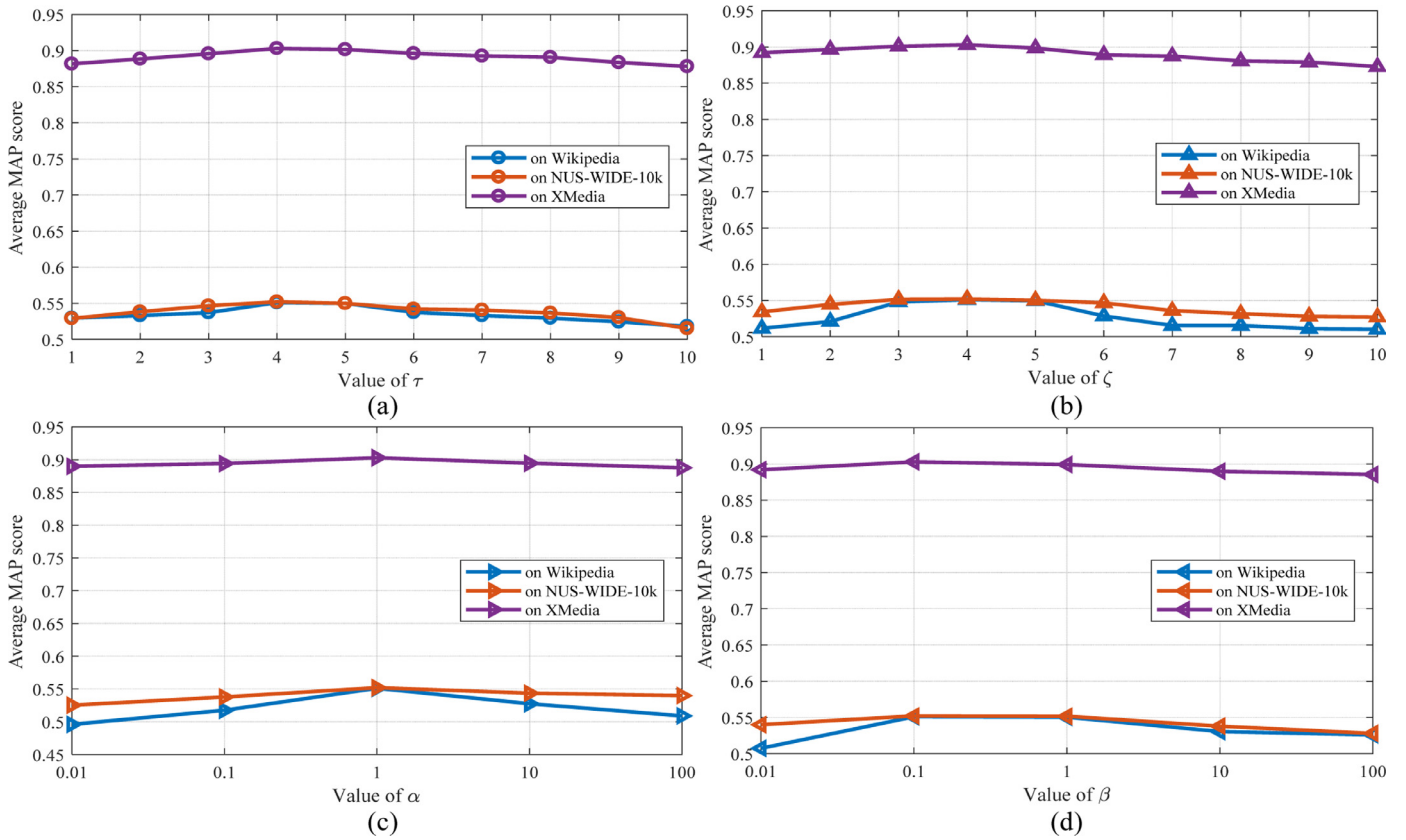


Fig. 6. Average cross-modal retrieval performances of MS²GAN with different values of (a) τ , (b) ζ , (c) α and (d) β on three datasets.

different values of hyper-parameters on the Wikipedia, NUS-WIDE-10k, and XMedia datasets. When one hyper-parameter is evaluated, the others are fixed. From the figure, our MS²GAN is not sensitive to the choice of τ in the range [4,5], ζ in the range [3,5], and β in the range [0.1,10]. In addition, from the figure, the best results can be achieved when $\alpha = 1$. For simplicity, we set $\tau = \zeta = 4$, $\alpha = 1$, and $\beta = 0.1$ for all three datasets.

5. Conclusion

In this paper, we propose a novel cross-modal retrieval approach named MS²GAN. The inter-modal invariance and the inter- and intra-modal discrimination is well modeled. Furthermore, the modality-specific and modality-shared features are jointly explored and leveraged, such that the complementarity and correlation information is effectively used for the retrieval task.

Comprehensive experiments on three widely used datasets with five modalities demonstrate that MS²GAN is able to outperform state-of-the-art cross-modal retrieval methods. Experimental results also validate the effectiveness of the important components, the adversarial learning scheme, and convergence of the proposed approach.

In this paper, we only evaluate the cross-modal retrieval performance of our approach on the public and extensively used datasets. For the future work, we will adopt more practical cross-modal data to validate the generalization ability of our approach and further improve the robustness of our approach. In addition, we will try to extend our approach to hashing method for further improving the retrieval efficiency in practical applications.

Declaration of Competing Interest

The authors have no declaration of interest to report.

Acknowledgment

The authors want to thank the editor and anonymous reviewers for their constructive comments and suggestions. The work in this paper was supported by the National Natural Science Foundation of China (No. 61702280), Natural Science Foundation of Jiangsu Province (No. BK20170900), National Postdoctoral Program for Innovative Talents (No. BX20180146), NSFC-Key Project of General Technology Fundamental Research United Fund (No. U1736211), China Postdoctoral Science Foundation (No. 2019M661901), Jiangsu Planned Projects for Postdoctoral Research Funds (No. 2019K024), CCF-Tencent Open Fund WeBank Special Funding (No. CCF-WebankRAGR20190104), and Scientific Research Starting Foundation for Introduced Talents in NJUPT (NUPTSF, No. NY217009).

References

- [1] H. Zhang, L. Liu, Y. Long, L. Shao, Unsupervised deep hashing with pseudo labels for scalable image retrieval, *IEEE Trans. Image Process.* 27 (4) (2018) 1626–1638.
- [2] W. Zhou, H. Li, J. Sun, Q. Tian, Collaborative index embedding for image retrieval, *IEEE Trans. Pattern Anal. Mach. Intell.* 40 (5) (2018) 1154–1166.
- [3] V.E. Liong, J. Lu, Y.-P. Tan, Cross-modal discrete hashing, *Pattern Recognit.* 79 (2018) 114–129.
- [4] D. Hu, F. Nie, X. Li, Deep binary reconstruction for cross-modal hashing, *IEEE Trans. Multimedia* 21 (4) (2019) 973–985.
- [5] Q.-Y. Jiang, W.-J. Li, Discrete latent factor model for cross-modal hashing, *IEEE Trans. Image Process.* 28 (7) (2019) 3490–3501.
- [6] L. Zhang, B. Ma, G. Li, Q. Huang, Q. Tian, Generalized semi-supervised and structured subspace learning for cross-modal retrieval, *IEEE Trans. Multimedia* 20 (1) (2017) 128–141.
- [7] Y. Peng, X. Huang, Y. Zhao, An overview of cross-media retrieval: concepts, methodologies, benchmarks, and challenges, *IEEE Trans. Circuits Syst. Video Technol.* 28 (9) (2018) 2372–2385.
- [8] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: *Pro-*

- ceedings of the 18th ACM International Conference on Multimedia, ACM, 2010, pp. 251–260.
- [9] F. Feng, X. Wang, R. Li, Cross-modal retrieval with correspondence autoencoder, in: Proceedings of the 22nd ACM International Conference on Multimedia, ACM, 2014, pp. 7–16.
 - [10] Y. Peng, X. Huang, J. Qi, Cross-media shared representation by hierarchical learning with multiple deep networks, in: Proceedings of the 25th International Joint Conference on Artificial Intelligence, 2016, pp. 3846–3853.
 - [11] X. Huang, Y. Peng, M. Yuan, Cross-modal common representation learning by hybrid transfer network, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017, pp. 1893–1900.
 - [12] B. Wang, Y. Yang, X. Xu, A. Hanjalic, H.T. Shen, Adversarial cross-modal retrieval, in: Proceedings of the 25th ACM International Conference on Multimedia, ACM, 2017, pp. 154–162.
 - [13] X. Xu, J. Song, H. Lu, Y. Yang, F. Shen, Z. Huang, Modal-adversarial semantic learning network for extendable cross-modal retrieval, in: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, ACM, 2018, pp. 46–54.
 - [14] X. Xu, L. He, H. Lu, L. Gao, Y. Ji, Deep adversarial metric learning for cross-modal retrieval, *World Wide Web* 22 (2) (2019) 657–672.
 - [15] X. Huang, Y. Peng, M. Yuan, Mhtn: modal-adversarial hybrid transfer network for cross-modal retrieval, *IEEE Trans. Cybern.* 50 (3) (2020) 1047–1059.
 - [16] Y. Peng, J. Qi, Cm-gans: cross-modal generative adversarial networks for common representation learning, *ACM Trans. Multimedia Comput. Commun. Appl.* 15 (1) (2019) 22:1–22:24.
 - [17] Y. Peng, J. Qi, Y. Yuan, Modality-specific cross-modal similarity measurement with recurrent attention network, *IEEE Trans. Image Process.* 27 (11) (2018) 5585–5599.
 - [18] J. Hu, J. Lu, Y.-P. Tan, Sharable and individual multi-view metric learning, *IEEE Trans. Pattern Anal. Mach. Intell.* 40 (9) (2018) 2281–2288.
 - [19] Y.T. Zhuang, Y.F. Wang, F. Wu, Y. Zhang, W.M. Lu, Supervised coupled dictionary learning with group structures for multi-modal retrieval, in: Proceedings of the 27th AAAI Conference on Artificial Intelligence, 2013, pp. 1070–1076.
 - [20] Y. Gong, Q. Ke, M. Isard, S. Lazebnik, A multi-view embedding space for modeling internet images, tags, and their semantics, *Int. J. Comput. Vis.* 106 (2) (2014) 210–233.
 - [21] F. Wu, X. Lu, J. Song, S. Yan, Z.M. Zhang, Y. Rui, Y. Zhuang, Learning of multi-modal representations with random walks on the click graph, *IEEE Trans. Image Process.* 25 (2) (2016) 630–642.
 - [22] K. Wang, R. He, L. Wang, W. Wang, T. Tan, Joint feature selection and subspace learning for cross-modal retrieval, *IEEE Trans. Pattern Anal. Mach. Intell.* 38 (10) (2016) 2010–2023.
 - [23] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: an overview with application to learning methods, *Neural Comput.* 16 (12) (2004) 2639–2664.
 - [24] D. Li, N. Dimitrova, M. Li, I.K. Sethi, Multimedia content processing through cross-modal association, in: Proceedings of the 11th ACM International Conference on Multimedia, ACM, 2003, pp. 604–611.
 - [25] K. Wang, R. He, W. Wang, L. Wang, T. Tan, Learning coupled feature spaces for cross-modal matching, in: Proceedings of the 14th IEEE International Conference on Computer Vision, 2013, pp. 2088–2095.
 - [26] Y. Peng, J. Qi, X. Huang, Y. Yuan, Ccl: cross-modal correlation learning with multigrained fusion by hierarchical network, *IEEE Trans. Multimedia* 20 (2) (2018) 405–420.
 - [27] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, S. Yan, Cross-modal retrieval with CNN visual features: a new baseline, *IEEE Trans. Cybern.* 47 (2) (2017) 449–460.
 - [28] Y. Wu, S. Wang, G. Song, Q. Huang, Online asymmetric metric learning with multi-layer similarity aggregation for cross-modal retrieval, *IEEE Trans. Image Process.* 28 (9) (2019) 4299–4312.
 - [29] P. Kang, Z. Lin, Z. Yang, X. Fang, Q. Li, W. Liu, Deep semantic space with intra-class low-rank constraint for cross-modal retrieval, in: Proceedings of the 2019 International Conference on Multimedia Retrieval, ACM, 2019, pp. 226–234.
 - [30] P. Hu, L. Zhen, D. Peng, P. Liu, Scalable deep multimodal learning for cross-modal retrieval, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2019, pp. 635–644.
 - [31] L. Zhen, P. Hu, X. Wang, D. Peng, Deep supervised cross-modal retrieval, in: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10394–10403.
 - [32] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, D. Parikh, Vqa: visual question answering, in: Proceedings of the 2015 IEEE International Conference on Computer Vision, 2015, pp. 2425–2433.
 - [33] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6077–6086.
 - [34] Z. Zhao, Z. Zhang, X. Jiang, D. Cai, Multi-turn video question answering via hierarchical attention context reinforced networks, *IEEE Trans. Image Process.* 28 (8) (2019) 3860–3872.
 - [35] W. Jin, Z. Zhao, Y. Li, J. Li, J. Xiao, Y. Zhuang, Video question answering via knowledge-based progressive spatial-temporal attention network, *ACM Trans. Multimedia Comput. Commun. Appl.* 15 (2s) (2019) 52:1–52:22.
 - [36] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, H. Huang, Heterogeneous memory enhanced multimodal attention model for video question answering, in: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1999–2007.
 - [37] Z. Zhao, Z. Zhang, S. Xiao, Z. Xiao, X. Yan, J. Yu, D. Cai, F. Wu, Long-form video question answering via dynamic hierarchical reinforced networks, *IEEE Trans. Image Process.* 28 (12) (2019) 5939–5952.
 - [38] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Proceedings of the 28th Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
 - [39] Q.-Y. Jiang, W.-J. Li, Deep cross-modal hashing, in: Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3232–3240.
 - [40] Z. Cao, M. Long, C. Huang, J. Wang, Transfer adversarial hashing for hamming space retrieval, in: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018, pp. 6698–6705.
 - [41] C. Deng, Z. Chen, X. Liu, X. Gao, D. Tao, Triplet-based deep hashing network for cross-modal retrieval, *IEEE Trans. Image Process.* 27 (8) (2018) 3893–3903.
 - [42] C. Li, C. Deng, N. Li, W. Liu, X. Gao, D. Tao, Self-supervised adversarial hashing networks for cross-modal retrieval, in: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4242–4251.
 - [43] Q.-Y. Jiang, X. Cui, W.-J. Li, Deep discrete supervised hashing, *IEEE Trans. Image Process.* 27 (12) (2018) 5996–6009.
 - [44] L. Wu, Y. Wang, L. Shao, Cycle-consistent deep generative hashing for cross-modal retrieval, *IEEE Trans. Image Process.* 28 (4) (2019) 1602–1612.
 - [45] J. Lu, J. Hu, J. Zhou, Deep metric learning for visual understanding: an overview of recent advances, *IEEE Signal Process. Mag.* 34 (6) (2017) 76–84.
 - [46] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, Nus-wide: a real-world web image database from national university of singapore, in: Proceedings of the 8th ACM International Conference on Image and Video Retrieval, ACM, 2009, p. No.48.
 - [47] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, in: Proceedings of the 2015 IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.
 - [48] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.
 - [49] C. McKay, I. Fujinaga, P. Depalle, jaudio: a feature extraction library, in: Proceedings of the 16th International Conference on Music Information Retrieval, 2005, pp. 600–603.
 - [50] D.-Y. Chen, X.-P. Tian, Y.-T. Shen, M. Ouhyoung, On visual similarity based 3D model retrieval, in: *Computer Graphics Forum*, 22, Wiley Online Library, 2003, pp. 223–232.
 - [51] J. Demšar, Statistical comparisons of classifiers over multiple data sets, *Mach. Learn. Res.* 7 (2006) 1–30.
 - [52] L.v.d. Maaten, G. Hinton, Visualizing data using t-sne, *J. Mach. Learn. Res.* 9 (2008) 2579–2605.

Fei Wu received the Ph.D. degree in computer science from Nanjing University of Posts and Telecommunications, China, in 2016. He is currently with the College of Automation in Nanjing University of Posts and Telecommunications. He has authored over forty scientific papers. His research interests include pattern recognition, artificial intelligence, and computer vision.

Xiao-Yuan Jing received the Doctoral degree of Pattern Recognition and Intelligent System in the Nanjing University of Science and Technology, 1998. Now he is a Professor with the School of Computer, Wuhan University, China. His research interests include pattern recognition and artificial intelligence.

Zhiyong Wu is pursuing the Master degree in computer technology from Nanjing University of Posts and Telecommunications, China. His research interests include computer vision and image processing.

Yimu Ji is a professor in Nanjing University of Posts and Telecommunications, China. His research mainly focuses on intelligent driving and data processing.

Xiwei Dong received the Ph.D. degree in computer science from Nanjing University of Posts and Telecommunications, China, in 2018. His research interests include machine learning, pattern recognition, and computer vision.

Xiaokai Luo is pursuing the Master degree in computer technology from Nanjing University of Posts and Telecommunications, China. His research interests include pattern recognition and computer vision.

Qinghua Huang is a lecturer in Nanjing University of Posts and Telecommunications, China. Her research mainly focuses on intelligent system and pattern recognition.

Ruchuan Wang is a professor in Nanjing University of Posts and Telecommunications, China. His research mainly focuses on information security and wireless sensor networks.