Pattern Recognition 112 (2021) 107734

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Cross-modal discriminant adversarial network

Peng Hu[a,b], Xi Peng[a], Hongyuan Zhu[b], Jie Lin[b], Liangli Zhen[c], Wei Wang[a], Dezhong Peng[a,d,e,*]

[a] College of Computer Science, Sichuan University, Chengdu 610065, China
[b] Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore
[c] Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore
[d] Shenzhen Peng Cheng Laboratory, Shenzhen 518052, China
[e] College of Computer & Information Science, Southwest University, Chongqing 400715, China

ARTICLE INFO

Article history:
Received 3 June 2020
Revised 2 October 2020
Accepted 29 October 2020
Available online 5 November 2020

Keywords:
Adversarial learning
Cross-modal representation learning
Cross-modal retrieval
Discriminant adversarial network
Cross-modal discriminant mechanism
Latent common space

ABSTRACT

Cross-modal retrieval aims at retrieving relevant points across different modalities, such as retrieving images via texts. One key challenge of cross-modal retrieval is narrowing the heterogeneous gap across diverse modalities. To overcome this challenge, we propose a novel method termed as Cross-modal discriminant Adversarial Network (CAN). Taking bi-modal data as a showcase, CAN consists of two parallel modality-specific generators, two modality-specific discriminators, and a Cross-modal Discriminant Mechanism (CDM). To be specific, the generators project diverse modalities into a latent cross-modal discriminant space. Meanwhile, the discriminators compete against the generators to alleviate the heterogeneous discrepancy in this space, *i.e.*, the generators try to generate unified features to confuse the discriminators, and the discriminators aim to classify the generated results. To further remove the redundancy and preserve the discrimination, we propose CDM to project the generated results into a single common space, accompanying with a novel eigenvalue-based loss. Thanks to the eigenvalue-based loss, CDM could push as much discriminative power as possible into all latent directions. To demonstrate the effectiveness of our CAN, comprehensive experiments are conducted on four multimedia datasets comparing with 15 state-of-the-art approaches.

1. Introduction

In numerous real-world applications, it is highly expected to retrieve the "similar" samples across various modalities for a given query because an object is usually described by multiple modalities, *e.g.*, text, video, image, and voice [1–3]. The key of such a so-called cross-modal retrieval problem is narrowing the heterogeneous gap/discrepancy which is caused by the fact of different modalities might lie in completely distinct spaces [4].

During past decades, many cross-modal approaches have been proposed to alleviate the heterogeneous discrepancy, which aim at projecting diverse modalities into a single unified space via shallow [5–7] or deep models [8–10]. In brief, the pioneer methods attempt to project multimedia data into a latent single unified space using shallow modality-specific transformations, which could be further categorized into unsupervised [11,12] and supervised methods [13–16]. The unsupervised models eliminate the heterogeneous discrepancy by maximizing the correlations between cross-modal pairwise samples [17]. Alternatively, the supervised methods employ the semantic information to boost the performance by preserving the discrimination into the common space [4,18]. Although the traditional shallow approaches have achieved promising performance, most of them are linear approaches and may be unable to capture the high-level semantics of real-world multimodal data that are highly nonlinear. To alleviate the issue, some kernel extensions [17] were proposed. However, it is still a daunting task to choose an appropriate kernel function as pointed out in [19].

To adaptively capture the nonlinearity of data, several recent works proposed using Deep Neural Network (DNN) for cross-modal analysis [8,20]. Among these works, [21–23] adopt some variants of Generative Adversarial Network (GAN) [24,25] to eliminate the modality discrepancy, which shows promising performance in practice. Although GAN-based methods have achieved promising performance, the redundant information may be preserved in the generated representations due to adversarial learning. To be specific, the GAN-based methods consist of discriminators and generators as shown in 1(a). The discriminators try to distinguish the generated representations from the real ones. Meanwhile, the gen-

* Corresponding author.
E-mail address: pengdz@scu.edu.cn (D. Peng).

https://doi.org/10.1016/j.patcog.2020.107734
0031-3203/

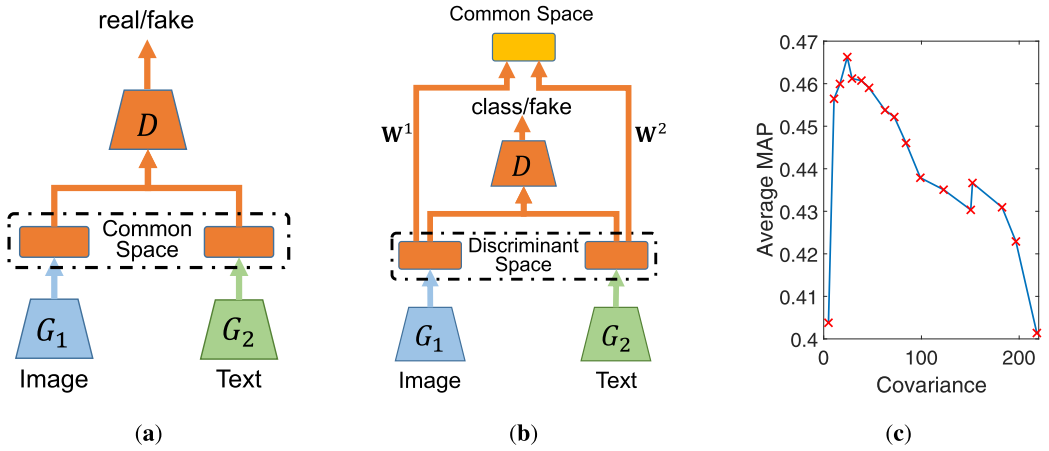


Fig. 1. A visual comparison between existing GAN-based cross-modal learning models including Adversarial Cross-Modal Retrieval (ACMR) [26] and ours. (a) The learning paradigm of most existing GAN-based methods. (b) Our CAN. In the figures, G_1 and G_2 represent the image and text generators, and D represents the discriminators. The major difference between the common paradigm Fig. 1(a) and ours Fig. 1(b) is that our generator does not learn a common space. Instead, it learns a discriminant space which could facilitate the cross-modal retrieval performance. In our method, $\mathbf{W}^1$ and $\mathbf{W}^2$ denote the modality-specific linear discriminant transformations, which are utilized to refine the discrimination from the generated features (*i.e.* the discriminant space). (c) Our observation on the relationship between the redundancy of the representation and the retrieval performance. To be exact, we use the covariance to measure the redundancy following [27], and show the average Mean Average Precision (MAP) achieved by ACMR on the Wikipedia database w.r.t. the corresponding covariance. One could see that the MAP of ACMR first increases with increasing covariance and then continuously decreases. In other words, ACMR needs a sufficiently high dimensional representation to contain the latent information. After that, a high dimensional representation will incorporate the redundancy and weaken the weight of discrimination, thus decreasing the MAP. More details on this experiment could refer to 4.5.

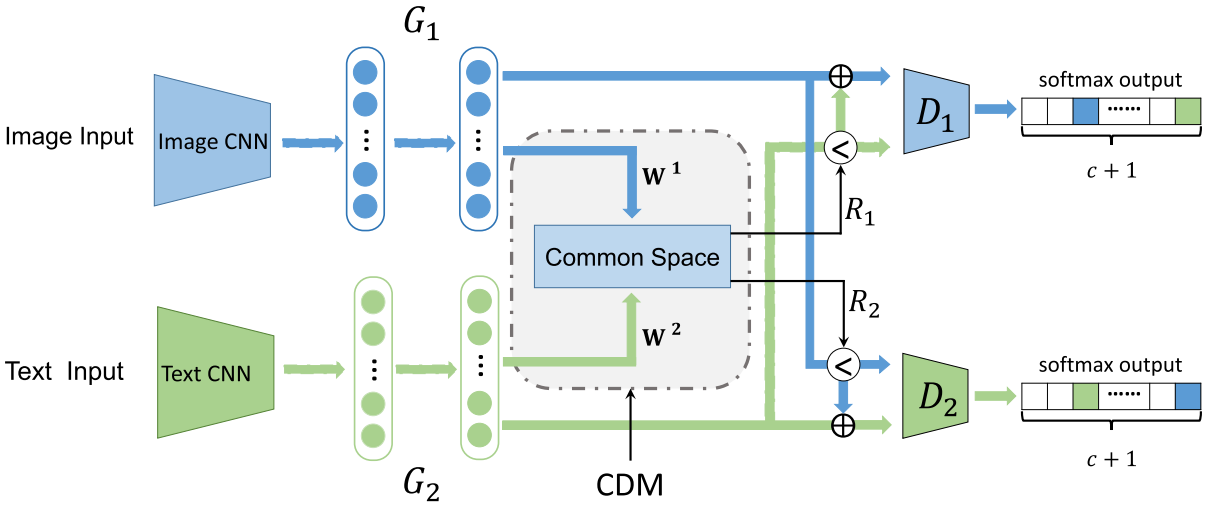


Fig. 2. The framework of our CAN, taking bimodal data (*i.e.*, image and text) as a showcase. CAN consists of three parts: two modality-specific generators (G_1 and G_2), two modality-specific discriminators (D_1 and D_2), and a cross-modal discriminant mechanism (CDM). In the figure, R_1 and R_2 are the feedback for D_1 and D_2, which are computed on the common space by our CDM and reflect the performance of CDM for cross-modal retrieval. c denotes the number of classes. The generators and discriminators compete with each other to eliminate the cross-modal discrepancy, while preserving the discrimination in the generated features. To further remove the redundancy and refine the discrimination, our CDM projects the generated features into a common space through $\mathbf{W}^1$ and $\mathbf{W}^2$. Note that $\mathbf{W}^1$ and $\mathbf{W}^2$ could be solved analytically from the outputs of the generators through Eq. (10).

erators try to generate the samples to confuse the discriminators. In such a way, the generators indirectly alleviate the heterogeneous discrepancy with the help of the discriminators, instead of explicitly capturing the distribution by the generators themselves only. Therefore, some redundant information will be remained during the adversarial process, which may decrease the performance of cross-modal retrieval as shown in 1(c).

Based on the aforementioned observations, we propose a Cross-modal discriminant Adversarial Network (CAN) which aims to eliminate the redundancy during adversarial learning. The proposed CAN consists of three modules, namely, two parallel modality-specific generators and discriminators accompanying a novel Cross-modal Discriminant Mechanism (CDM) (see 2). As shown in 1, the major differences of our method with existing ones are given in the following two aspects. On one hand, most existing paradigms directly generate the common representations, whereas our CAN employs CDM to project the generated features (*i.e.*, the discriminant space) into a latent common space via $\mathbf{W}^{1,2}$. Thanks to our CDM, the redundancy could be alleviated and the discrimination is refined from the generated features. In other words, the cross-modal discrepancy could be eliminated while incorporating the discrimination, thus improving the retrieval performance. On the other hand, our discriminators not only distinguish real/fake samples but also simultaneously perform classification on the real samples using the available label information as shown in 1. In consequence, the generators can encapsulate the label-induced discriminative information into the common space to improve retrieval performance. The main contributions of this work are summarized as follows:

- CAN is proposed to learn a latent discriminant space for multimedia data, which is with a novel network structure and a novel learning mechanism (CDM). In brief, CDM projects the generated features of all modalities into a latent common space and gives the positive/negative feedback to adversarial learning. Therefore, our method could reduce the modality discrepancy, while preserving the discriminative information into the common space.
- To improve our CDM, a novel objective function is presented to learn a latent unified space wherein the within-class points could be compacted and the between-class points could be scattered. Furthermore, the transformations of the CDM can be analytically solved from the generated features, thus escaping from the trap of local minimal.
- To avoid the trivial solutions of directly optimizing the CDM objective function, a novel logarithmic eigenvalue-based loss is proposed. Another advantage of the proposed loss is that it could push as much discrimination as possible into all latent directions of CDM's transformations instead of only the dominant ones.

2. Related work

To bridge the heterogeneous gap, numerous cross-modal learning approaches are proposed in recent years. In this section, we will briefly review some shallow and deep cross-modal representation learning methods.

2.1. Traditional cross-modal approaches

Most traditional methods [12,28] usually learn modality-specific transformation matrices to project each modality into a latent single unified space wherein the correlation of different modalities is maximized. To utilize the semantic labels for cross-modal representation learning, [4,29,30] were proposed to learn a latent single unified space with Fisher's criterion. To further utilize available label information, [29] proposed Generalized Multiview Analysis (GMA) that simultaneously utilizes the label-induced discriminative information and the pairwise-modality relationship to learn the common space. Moreover, [15] proposed Generalized Semi-supervised Structured Subspace Learning (GSS-SL) which models the correlations between different modalities by taking the label space as a linkage. To better capture the nonlinearity of data, several kernel-based approaches have been proposed to nonlinearly model the cross-modal correlation, *e.g.*, Kernel Canonical Correlation Analysis [17,31]. However, it is a daunting task to determine the kernel function since there is no golden criterion to choose the kernel function [19,32]. Moreover, the kernel methods are shallow models which might be with the limited capacity of modeling nonlinearity.

2.2. Deep cross-modal approaches

During past years, DNN has been widely utilized to model the correlation between different modalities [21,33]. On one hand, some unsupervised pioneers attempt to progressively learn some modality-specific nonlinear transformations to project the corresponding modalities into a latent unified space, while maximizing the correlation between these modality-specific representations in the learned space [34,35]. On the other hand, some approaches propose to learn the common discriminant representations by jointly modeling the intra- and inter-modality correlations [20,21]. In [8], a Multi-view Deep Network (MvDN) approach is proposed to learn one discriminative single unified space by adopting the Fisher's criterion into a feedforward DNN.

Recently, inspired by the great success achieved by the Generative Adversarial Nets (GAN) [24] in capturing data distribution, some GAN-based works propose to seek an effective representations by using adversarial learning [22,36,37]. In [36], Duan *et al.* creatively utilize numerous easy negatives to generate potential hard negatives as complements with adversarial learning to learn the representations. Moreover, Wang *et al.* proposed an Adversarial Cross-Modal Retrieval (ACMR) approach to learn one single unified space by utilizing adversarial learning, which incorporates a feature extractor, a modality classifier, and a triplet constraint [26]. In [21], a GAN-based approach, termed Cross-modal Generative Adversarial Networks (CM-GANs), is proposed to correlate the multimedia data across distinct modalities. In [22], a generative cross-modal feature learning framework is proposed, called GXN, which applies generative-adversarial processes into the cross-modal representation learning. As mentioned above, some redundant information may be preserved in the learned common space during their adversarial learning process.

3. Cross-modal discriminant GAN

The overall pipline of our approach is shown in 2. For the discriminative model, the modality-specific discriminators aim to discriminate the real and fake points. For the generative model, the modality-specific generators attempt to generate the modality-invariant and discriminative representations to confuse our discriminator. The discriminators and generators compete with each other so that the multimodal data is projected into a latent linear discriminant space. Such a process could preserve as much discrimination as possible, while narrowing the heterogeneous gap of multi-modal data. Moreover, to remove redundancy, our CDM projects these generated representations into a latent common space.

3.1. Overview of the framework

For ease of presentation, we first give some definitions as below. We denote k-th labeled modality as $\mathcal{X}^k = \{(\mathbf{x}_{ij}^k, \ell_i) | i = 1, 2, \cdots, c;\ j = 1, 2, \cdots, N_i^k;\ k = 1, \cdots, m\}$, where $\mathbf{x}_{ij}^k$ denotes the j-th point from the k-th modality of the i-th category, c is the number of categories, N_i^k is the number of points from the k-th modality of the i-th category, and $m = 2$ denotes the total number of modalities. ℓ_i is a one hot vector with the length of $c+1$, where the i-th entry is with the value of 1 (see 2 for an example). Note that, each of the first c entires represent the category of the corresponding individual, and the $(c+1)$-th entry indicates the fake/real label with the value of 1/0. The extra class entry of the one-hot vector is valid that corresponds to fake label, denoted as ℓ_f. Namely,

$$\ell_f = [\underbrace{0, \cdots, 0}_{c}, \overset{\text{Extra class}}{\overset{\downarrow}{1}}]. \tag{1}$$

As shown in 2, the proposed CAN consists of two generators and two discriminators. The generators aim to extract modality-invariant discriminative features with help of the discriminators. The generator and discriminator from the k-th view are respectively represented as nonlinear functions $G_k(\cdot; \Theta_G^k)$ and $D_k(\cdot; \Theta_D^k)$, where Θ_G^k and Θ_D^k are their parameters. Mathematically, our objective function is formulated as below:

$$\begin{aligned} &\arg\min_{G_1, \cdots, G_m \atop D_1, \cdots, D_m} \ (\mathcal{L}_G + \mathcal{L}_D + \lambda \mathcal{L}_W), \end{aligned} \tag{2}$$

where $\mathcal{L}_G$, $\mathcal{L}_D$ and $\mathcal{L}_W$ denote the losses of generators, discriminators, and cross-modal discriminant analysis, respectively. More specifically, the generators G_1 and G_2 aim at mapping the samples from the corresponding modalities into the cross-modal discriminant representations via $\mathbf{y}_{ij}^k = G_k(\mathbf{x}_{ij}^k)$. The generated representations are expected to linearly map into a latent common space with modality invariance and discrimination via $\mathbf{z}_{ij}^k = (\mathbf{W}^k)^T\mathbf{y}_{ij}^k$, where $\mathbf{W}^k$ is the k-th modality-specific transformation. Furthermore, the generators are pitted against the corresponding discriminators. Taking the bi-modal data as a showcase, the generators G_1 and G_2 play the following roles: 1) enforcing the generated representation to be separable and modality-invariant, 2) fooling the discriminators D_1 and D_2, and 3) making the common representation discriminative. In the following section, we will elaborate our method.

3.2. Cross-modal discriminant mechanism

The generator G_k aims to transform an input sample $\mathbf{x}_{ij}^k$ into a latent linear discriminant space as follows:

$$\mathbf{y}_{ij}^k = G_k(\mathbf{x}_{ij}^k). \tag{3}$$

To eliminate the redundancy and further utilize the discrimination, the generated representation $\mathbf{y}_{ij}^k$ is projected into one latent single unified space via $\mathbf{z}_{ij}^k = (\mathbf{W}^k)^T\mathbf{y}_{ij}^k$ by achieving within-class compactness and between-class scatter. To the end, the within-class compactness matrix is defined as

$$\begin{aligned}\hat{\mathbf{S}}_W &= \sum_{i=1}^{c}\sum_{k=1}^{m}\sum_{j=1}^{N_i^k}\left(\mathbf{z}_{ij}^k - \boldsymbol{\mu}_i\right)\left(\mathbf{z}_{ij}^k - \boldsymbol{\mu}_i\right)^T \\ &= \mathbf{W}^T\mathbf{S}_W\mathbf{W},\end{aligned} \tag{4}$$

$\mathbf{S}_W$ is defined as below:

$$\begin{bmatrix}(\mathbf{S}_W)_{11} & \cdots & (\mathbf{S}_W)_{1v}\\ \vdots & \ddots & \vdots \\ (\mathbf{S}_W)_{v1} & \cdots & (\mathbf{S}_W)_{vv}\end{bmatrix}, \tag{5}$$

where the (k,l)-th sub-matrix is in the form of

$$(\mathbf{S}_W)_{kl} = \sum_{i=1}^{c}\left((k==l)\sum_{j=1}^{N_i^k}\mathbf{y}_{ij}^k(\mathbf{y}_{ij}^l)^T - \frac{1}{N_i}\mathbf{s}_i^k(\mathbf{s}_i^l)^T\right), \tag{6}$$

and $k==l$ is a Boolean equation whose value is 1 if $k=l$ and 0 otherwise. $\mathbf{s}_i^k = \sum_{j=1}^{N_i^k}\mathbf{y}_{ij}^k$ is the sum of all generated representations of the i-th category for the k-th modality.

Like the within-class matrix, the between-class scatter matrix could be formulated as follows:

$$\begin{aligned}\hat{\mathbf{S}}_B &= \sum_{i=1}^{c} N_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T \\ &= \mathbf{W}^T\mathbf{S}_B\mathbf{W},\end{aligned} \tag{7}$$

where $\boldsymbol{\mu}_i = \frac{1}{N_i}\sum_{k=1}^{m}\sum_{j=1}^{N_i^k}\mathbf{z}_{ij}^k$ is the mean of all obtained unified features of the i-th class from all modalities, N_i^k is the point number of the i-th category in the k-th modality, $\boldsymbol{\mu} = \frac{1}{N}\sum_{k=1}^{m}\sum_{i=1}^{c}\sum_{j=1}^{N_i^k}\mathbf{z}_{ij}^k$ is the mean of all obtained unified features in all modalities, $N_i = \sum_{k=1}^{m}N_i^k$ is the total point number of the i-th class in the mini-batch, and $\mathbf{W}^T = \left[(\mathbf{W}^1)^T\ \cdots\ (\mathbf{W}^m)^T\right]^T$ is a combined matrix, which consists of all the modality-specific transformations.

Similarly, $\mathbf{S}_B$ is a partitioned matrix with

$$(\mathbf{S}_B)_{kl} = \sum_{i=1}^{c}\frac{1}{N_i}\mathbf{s}_i^k(\mathbf{s}_i^l)^T + \frac{1}{N}\mathbf{s}_i^k(\mathbf{s}_i^l)^T, \tag{8}$$

where N is the number of points in the mini-batch.

With the above defined $\mathbf{S}_W$ and $\mathbf{S}_B$, we have

$$\begin{aligned}\mathbf{W}^* &= \arg\max_{\mathbf{W}}\mathrm{Tr}\left(\frac{\mathbf{W}^T\mathbf{S}_B\mathbf{W}}{\mathbf{W}^T\mathbf{S}_W\mathbf{W}}\right) \\ &= \arg\max_{\mathbf{W}}\frac{|\mathbf{W}^T\mathbf{S}_B\mathbf{W}|}{|\mathbf{W}^T\mathbf{S}_W\mathbf{W}|},\end{aligned} \tag{9}$$

where $\mathrm{Tr}(\cdot)$ and $|\cdot|$ are the trace and determinant operators, respectively. Eq. (9) could be equivalently reformulated as the generalized eigenvalue decomposition (GED) problem:

$$\mathbf{S}_B\mathbf{w}_i = \lambda_i\mathbf{S}_W\mathbf{w}_i, \tag{10}$$

where λ_i and $\mathbf{w}_i$ $(i = 1, 2, \cdots, m)$ are the i-th largest eigenvalue and the corresponding eigenvector of the generalized eigenvalue decomposition problem, respectively. $\mathbf{w}_i$ is i-th column vector of the matrix $\mathbf{W}$, and $m \leqslant c-1$ is the objective reduction dimension. Note that, the upper bound of m is $c-1$ since there are at most $c-1$ nonzero generalized eigenvalues [38]. Therefore, the common representation of a given sample $\mathbf{x}_{ij}^k$ can be obtained by

$$\mathbf{z}_{ij}^k = (\mathbf{W}^k)^T G_k(\mathbf{x}_{ij}^k). \tag{11}$$

As the common space (Eq. (11)) cannot compute the gradient to optimize the neural network, we use the common representations to produce the positive/negative feedback and give it back to train the networks. To obtain the feedback, we perform cross-modal matching among the computed common representations. Specifically, the class centers of the k-th modality are used to predict the cross-modal matching probability for the l-th $(k \neq l)$ modality. The probability of matching $\mathbf{z}_{ij}^l$ to the k-th modality is defined as

$$p(\mathbf{z}_{ij}^l|r;k) = \frac{e^{-\|\mathbf{z}_{ij}^l - \boldsymbol{\mu}_r^k\|_2^2}}{\sum_{r=1}^{c}e^{-\|\mathbf{z}_{ij}^l - \boldsymbol{\mu}_r^k\|_2^2}}, \tag{12}$$

where $\boldsymbol{\mu}_i^k = \sum_{j=1}^{N_i^k}\mathbf{z}_{ij}^k$ is the average vector of the common representations of the i-th category from the k-th modality.

With $p(\mathbf{z}_{ij}^l|r;k)$, we could compute the matched set via $\breve{\mathbb{U}}^{kl} = \left\{\mathbf{x}_{ij}^l | p(\mathbf{z}_{ij}^l|i;k) \geqslant \sigma\right\}$, where $\frac{1}{c} < \sigma < 1$ is a positive threshold and we empirically fix $\sigma = 0.9$. Similarly, the mismatched set can be obtained by $\hat{\mathbb{U}}^{kl} = \left\{\mathbf{x}_{ij}^l | p(\mathbf{z}_{ij}^l|i;k) < \sigma\right\}$. Obviously, $\breve{\mathbb{U}}^{kl}$ and $\hat{\mathbb{U}}^{kl}$ satisfy $\breve{\mathbb{U}}^{kl}\bigcup\hat{\mathbb{U}}^{kl} = \mathcal{X}^l$. This matching information could give the generators and discriminators feedback to push more discrimination into the common space, *i.e.*, R_1 and R_2. More details about the feedback mechanism will be given in the following sections.

3.3. Modality-specific generators

For a given data point $\mathbf{x}^k$ of the k-th modality, our method aims to enforce that the generated representations are as similar as possible across different modalities, and meantime the discriminators could classify the corresponding representation into the correct category.

To be specific, in the latent common space, the k-th modality is matched with another modality to produce the matched samples $\breve{\mathbb{U}}^{kl}$ $(l \neq k)$ and the mismatched samples $\hat{\mathbb{U}}^{kl}$ $(l \neq k)$, where $\breve{\mathbb{U}}^{kl}$ and $\hat{\mathbb{U}}^{kl}$ are the sets of the matched and mismatched points in the l-th modality for the k-th modality respectively. Clearly, $\mathcal{X}^l = \breve{\mathbb{U}}^{kl}\cup\hat{\mathbb{U}}^{kl}$. Note that, the discriminators not only performs classification on the generated representations, but also judges the representations matched or mismatched. Formally, the loss function of the generators could be formulated as

$$\mathcal{L}_G = \sum_{k=1}^{m}\sum_{i=1}^{c}\sum_{\mathbf{x}_{ij}^k\in\overline{\mathbb{V}}^k}\mathcal{H}\left(D_k\left(G_k(\mathbf{x}_{ij}^k)\right), \ell_i\right) \tag{13}$$

where $\overline{\mathbb{V}}^k = \bigcup_{l\neq k}^{m} \hat{\mathbb{U}}^{kl}$ denotes the fake samples containing all mismatched samples from the other modalities, and $\mathcal{H}(\cdot,\cdot)$ is a cross-entropy loss used in the *softmax* layer.

3.4. Modality-specific discriminators

The goals of the discriminator are twofold. On one hand, it will distinguish the real samples from fake ones. The real samples include all samples generated by the corresponding generator, and the matched samples from the other modality-specific generators. In contrast, the false samples are the mismatched samples generated by the other modality-specific generators. On the other hand, our discriminators attempt to distinguish generated features from the real points into corrected categories.

Specifically, the discriminator D_k attempts to regard both the samples from its own modality $\mathcal{X}^k$ and the matched samples from other modalities $\check{\mathbb{U}}^{kl}$ $(l \neq k)$ as real individuals and categorize them as their real classes. Moreover, it also distinguishes the mismatched samples $\hat{\mathbb{U}}^{kl}$ from other modalities as fake ℓ_f, where $\check{\mathbb{U}}^{kl}$ and $\hat{\mathbb{U}}^{kl}$ $(l \neq k)$ are the sets of matched and mismatched samples from the l-th modality for the k-th one, respectively. In summary, we can formulate the loss function for our discriminators as

$$\begin{aligned} &\mathcal{L}_D = \sum_{k=1}^{m}\sum_{i=1}^{c}\sum_{j=1}^{N_i^k} \mathcal{H}\Big(D_k\big(G_k(\mathbf{x}_{ij}^k)\big), \ell(\mathbf{x}_{ij}^k)\Big), \\ &s.t. \quad \ell(\mathbf{x}_{ij}^k) = \begin{cases} \ell_i & if\ \mathbf{x}_{ij}^k \in \mathbb{V}^k, \\ \ell_f & otherwise. \end{cases} \end{aligned} \tag{14}$$

where $\mathbb{V}^k = \bigcup_{l\neq k}^{m} \check{\mathbb{U}}^{kl} \bigcup \mathcal{X}^k$ is the real set containing the k-th modality and the corresponding matched samples from other modalities. The positive/negative (matched/mismatched) feedback is used to bridge discriminators and the common space by cross-modal matching.

3.5. Logarithmic eigenvalue-based constraint

To further push more discriminative information into the latent common space, the cross-modal discriminant analysis is introduced into the adversarial learning process, which can enhance the performance of our CDM. The loss of the cross-modal discriminant analysis is defined as

$$\mathcal{L}_W = -\ln(J_W) \tag{15}$$

where $\ln(\cdot)$ is the natural logarithm operator and $J_W = |\hat{\mathbf{S}}_B^{-1}\hat{\mathbf{S}}_W|$ is a criterion function in Eq. (9). However, directly optimizing this equation or Eq. (9) will produce some problems as pointed out in [39,40]. To better discuss these problems, we rewrite Eq. (15) as an eigenvalue-based formulation through the following theorem.

Theorem 1. *Let λ_i $(0 < \lambda_1 < \cdots < \lambda_{c-1})$ and $\mathbf{w}_i$ are respectively the i-th largest positive eigenvalue and the corresponding eigenvector of the GED problem Eq. (10), then $\ln(J_W)$ is equivalent to:*

$$\sum_{i=1}^{c-1} \ln(\lambda_i). \tag{16}$$

Proof. Firstly, the Eq. (9) can be rewritten as:

$$J_W = |\hat{\mathbf{S}}_B^{-1}\hat{\mathbf{S}}_W| = \left|\mathbf{W}^{-1}\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{W}\right| = \left|\mathbf{S}_W^{-1}\mathbf{S}_B\right| \tag{17}$$

Obviously, $\lambda_i|_{i=1}^{c-1}$ are also the eigenvalues of $\mathbf{S}_W^{-1}\mathbf{S}_B$. Then, the $\ln(J_W)$ is equivalent to:

$$\ln(\prod_{i}^{c-1} \lambda_i). \tag{18}$$

Therefore, $\ln(J_W)$ can be rewritten as

$$\sum_{i=1}^{c-1} \ln(\lambda_i), \tag{19}$$

□

Based on Theorem 1, the loss of the cross-modal discriminant analysis can be formulated as:

$$\mathcal{L}_W = -\frac{1}{c-1}\sum_{i=1}^{c-1} \ln(\lambda_i). \tag{20}$$

However, it would result in trivial solutions if we directly optimize the above problem, *e.g.*, the optimizer will overemphasize the largest eigenvalues that will produce higher rewards in back-propagation than the minor ones. Specifically, the overemphasis problem make the optimizer focus on mainly maximizing the large between-class distance between apart points for classification, leading big overlap among neighboring categories [39–42]. Intuitively, to tackle this issue, we should weaken the weight of dominant eigenvalues in the optimization process, and maximize the minor ones'. To achieve this goal, we maximize the n lowest eigenvalues and filter the dominant ones to optimize the generators. Thus, the amount of each eigenvector direction (*i.e.*, the value of the corresponding eigenvalue) could be maximized during the generator optimization. The corresponding loss function is formulated as follows:

$$\begin{aligned} &\mathcal{L}_W = -\frac{1}{n}\sum_{i=1}^{n} \ln(\lambda_i), \\ &s.t. \quad n = \lceil q(c-1) \rceil, \end{aligned} \tag{21}$$

where $\lceil x \rceil$ is a mathematic operator to round each decimal value of x to equal its nearest integer, and q $(0 < q < 1)$ is a positive balance parameter to obtain the n lowest eigenvalues. From the formulation, one could see that the bottom of eigenvalues will be maximized without overemphasizing dominant ones during the network optimization process. Therefore, the generators could encapsulate discriminative variances among the $c-1$ representation directions (*i.e.*, eigenvectors) to the common space as much as possible, avoiding the overemphasis problem. That is to say, the optimizer pushes as much discrimination as possible into the unified space. Furthermore, Eq. (21) could be optimized with an end-to-end manner.

From the above, all losses of CAN have been formulated in Eq. (2). The generators and discriminators are optimized in an adversarial learning manner, which could ensure the generated representations are modality-invariant and the discriminative information is preserved in the latent common space. Compared with other adversarial learning approaches, the supervised labels for optimizing our generators and discriminators are different as shown in 1, which are adopted to optimize the generators and discriminators in an alternative manner. The optimization process of our CAN is summarized in 1.

4. Experimental study

To comprehensively evaluate the effectiveness of our CAN, some experiments are conducted on four widely-used databases, namely Wikipedia [43], Pascal Sentence [44], NUS-WIDE-10K [45] and XMediaNet [46] datasets.

4.1. Datasets

Here we briefly introduce four multimedia databases utilized in our experiments, including Wikipedia, Pascal Sentence, NUS-WIDE-10K, and XMediaNet databases. We split each dataset as 3 subsets, namely training, validation, and test sets. The statistics of four

Algorithm 1 Optimization procedure of CAN.

Input: The training data $\mathcal{X}^k|_{k=1}^m$, the dimensionality of the generated representations d, batch size N_b, positive balance parameters λ and q, positive integer N_s, and learning rate α.

while not converge **do**
  Calculate the modality-specific linear transformations $\{\mathbf{W}^i\}_{i=1}^m$ by Eqs. (3−10) on the training set.
  Compute the mean of the common representations of the i-th class for each modality by $\boldsymbol{\mu}_i^k = \sum_{j=1}^{N_i^k} \mathbf{z}_{ij}^k$.
  for N_s steps **do**
    Randomly select N_b samples for each modality from $\{\mathcal{X}^i\}_{k=1}^m$ to construct the corresponding modality-specific minibatch.
    Cross-modal matching is conducted on all modality-specific minibatches to obtain the matched and mismatched sets by Eq. (12).
    Update the parameters of the discriminators by minimizing $\mathcal{L}_D$ in Eq. (14) with descending their stochastic gradient:
$$\Theta_D^k = \Theta_D^k - \alpha \frac{\partial \mathcal{L}_D}{\partial \Theta_D^k} \ (k = 1, \cdots, m)$$
    Update the parameters of the generators by minimizing $\mathcal{L}_G + \lambda \mathcal{L}_W$ in Eq. (13) and Eq. (21) with descending their stochastic gradient:
$$\Theta_G^k = \Theta_G^k - \alpha \left(\frac{\partial \mathcal{L}_G}{\partial \Theta_G^k} + \lambda \frac{\partial \mathcal{L}_W}{\partial \Theta_G^k} \right) (k = 1, \cdots, m)$$
  end for
end while

Output: Optimized CAN model.

Table 1
General statistics of the four adopted datasets in our experiments, where "*/*/*" in the "Instances" column denotes the sample number of training/validation/test subsets, c is the number of categories, D_{img} and D_{txt} denote the dimensionality of image and text modalities, respectively.

Dataset	Instances	c	D_{img}	D_{txt}
Wikipedia	2,173/231/462	10	4,096D	$2,891 \times 300$D
Pascal Sentence	800/100/100	20	4,096D	102×300D
NUS-WIDE-10K	8,000/1,000/1,000	10	4,096D	60×300D
XMediaNet	32,000/4,000/4,000	200	4,096D	849×300D

databases are summarized in 1. The details of these datasets are given as follows.

4.1.1. Wikipedia dataset [43]

Wikipedia is a widely-used benchmark dataset to evaluate the effectiveness of multimodal approaches for cross-modal retrieval. It contains 2866 image-text pairs. Each pair (*i.e.*, an image and a text document) is classified into a label from 10 semantic categories, *e.g.*, art, history, etc. We split the dataset into 3 subsets: 2,173, 231 and 462 pairs for training, validation and test sets, respectively, following [47].

4.1.2. Pascal sentence dataset [44]

This dataset contains 1000 image-text pairs. Each image is generated by the 2008 PASCAL development kit, and its corresponding text is generated by annotating Amazon Mechanical Turk. Each text contains 5 independent sentences from diverse annotators. Moreover, each pair is classified into 20 classes. For a fair comparison, this dataset is also divided to 3 subsets following [47], *i.e.*, 800 pairs (40 samples per class), 100 pairs (5 samples per class), 100 pairs (5 samples per class) in the training, validation, and test sets, respectively.

4.1.3. NUS-WIDE-10K dataset [45,47]

This dataset is a subset sampled from the NUS-WIDE dataset [47], which is evenly selected from the 10 largest classes of NUS-WIDE, *e.g.*, animal, cloud, etc. NUS-WIDE-10K contains 10,000 image-text pairs. Like the aforementioned datasets, this dataset is also divided to 3 subsets following [47]: 8,000 pairs, 1000 pairs, and 1000 pairs for training, validation, and test sets, respectively.

4.1.4. XMediaNet dataset [46,48]

XMediaNet is a large-scale multimodal dataset that contains five different multimedia types, *i.e.*, image, text, video, audio, and 3D model. In the experiments, only image and text are utilized to evaluate the effectiveness of tested methods, *i.e.*, 40,000 image-text pairs. Each pair of the dataset is classified into 200 categories, *e.g.*, owl, elephant, etc. Like other datasets, XMediaNet is also split to training, validation and test sets which respectively contain 32,000 pairs, 4000 pairs and 4000 pairs following [21,46].

4.2. Experiment setting

For a fair comparison, all tested approaches utilize the same cross-modal features in our experiments. Specifically, 4,096-dimensional image features are extracted by fc7 layer of 19-layer VGGNet [49], which is a pretrained PyTorch model on the ImageNet dataset. The words of text documents are represented as 300-dimensional feature vectors that are extracted by a pretrained Word2Vec model [50] on Google News. Then each text is represented by an $300 \times s$ feature matrix, where s the maximal word number of all texts in the dataset, and zero-padding is utilized for the short texts. For our method, sentence CNN is used to handle the text feature matrices, but other compared methods cannot directly deal with the matrix inputs. Thus, each text feature matrix could be reshaped as a $300s$-dimensional feature vector to feed to the other methods. However, the dimensionality of text feature vectors is too high to be handled for Wikipedia and XMediaNet by other approaches on our devices, since the datasets have much long text documents, *e.g.*, 869,100 for Wikipedia. Following [15], 300-dimensional mean vectors of the $300 \times s$ feature matrices are computed to represent the texts. For the other datasets (*i.e.*, Pascal Sentence and NUS-WIDE-10K), each text could be represented as a $300s$-dimensional feature vector for the compared methods.

For the shallow compared methods (*i.e.*, CCA, MCCA, PLS, GMA, MvDA, and MvDA-VC), the objective dimensionality is determined by the best accuracy of the corresponding methods on the validation set traversing $[1 : 250]$ for all dataset, and the other parameters use the default ones provided by the authors. For the proposed CAN, the sentence CNN architecture is with the same configuration of [51]. Moreover, the dimensionality of the generated representations is 32 for all datasets except for the XMediaNet dataset which uses 200. The ReLU activation function is used on all layers except for the final one that uses a linear activation function. q and λ are respectively set as 0.8 and 1 in all the experiments on all datasets. Learning rate α is set as 0.0002 in all the experiments on all datasets. Finally, the mean average precision (MAP) scores (calculating on all returned results) of the cross-modal retrieval tasks are utilized to evaluate the effectiveness of the tested approaches. Note that the results of MCSM, CMDN, CCL, CBT and CM-GANs are reported by the authors with 4,096-dimensional fine-tuned VGGNet features and 300-dimensional pretrained sentence CNN features in the same datasets.

4.3. Evaluation metric

We adopt cross-modal retrieval on four benchmark datastets (*i.e.*, Wikipedia, Pascal Sentence, NUS-WIDE-10K and XMediaNet)

Table 2
Performance comparison of cross-modal retrieval in terms of MAP scores on Wikipedia.

Method	Img2Txt	Txt2Img	Average
CCA [28]	0.141	0.139	0.140
MCCA [11]	0.224	0.233	0.229
PLS [12]	0.303	0.225	0.264
GMA [29]	0.201	0.154	0.177
MvDA [30]	0.315	0.281	0.298
MvDA-VC [4]	0.374	0.345	0.359
GSS-SL [15]	0.504	0.461	0.483
DCCA [34]	0.361	0.326	0.343
DCCAE [35]	0.372	0.335	0.354
GXN [22]	0.318	0.280	0.299
ACMR [26]	0.493	0.462	0.478
MCSM [46]	0.516	0.458	0.487
CMDN [52]	0.487	0.427	0.457
CCL [20]	0.504	0.457	0.481
CM-GANs [21]	0.521	0.466	0.494
CAN	**0.540**	**0.474**	**0.507**

to evaluate the performance of all tested approaches. There are two kinds of retrieval tasks:

- **Image query text** (Img2Txt): retrieving relevant text samples in the test set ranked by calculated image-text similarity, using an image query.
- **Text query image** (Txt2Img): retrieving relevant image samples in the test set ranked by computed image-text similarity, using a text query.

The cosine similarity metric is utilized to compute the similarity scores across different modalities. A widely-used evaluation metric, *i.e.*, Mean Average Precision (MAP) score, is adopted to investigate the performance of all tested approaches on the datasets. MAP is the mean value of Average Precision (AP) of all queries. The definition of AP for the i-th query is

$$AP_i = \frac{1}{R(n)} \sum_{k=1}^{n} \frac{R(k)}{k} \times P(k) \tag{22}$$

where n is the number of retrieving points, and $R(k)$ counts the number of the related points in the top k returned results. $P(k)$ is a Boolean function that equals 1 if the returned sample of the rank k is a related point, and zero otherwise. In the experiments, the MAP scores of all experiments are calculated on the all returned results following [26]. Besides MAP, some precision-recall curves are plotted to visually evaluate the performance of the proposed method and its counterparts.

4.4. Comparisons with 15 state-of-the-art approaches on the datasets

In this section, we evaluate our CAN with some related methods, including CCA [28], MCCA [11], PLS [12], GMA [29], MvDA [30], MvDA-VC [4], GSS-SL [15], DCCA [34], DCCAE [35], MCSM [46], CMDN [52], CCL [20], GXN [22], CM-GANs [21] and ACMR [26]. All experimental results are presented on our CAN as well as all the tested methods, on the following four datasets, *i.e.* Wikipedia, Pascal Sentence, NUS-WIDE-10K and XMediaNet datasets. The MAP scores of the retrieval tasks (Img2Txt and Txt2Img) and their average scores on the four datasets are shown in 2, 3, 4 and 5, respectively. From the experimental results, one could see that our CAN achieves the best performance comparing with its 15 counterparts on the tested datasets. Among all the tested methods, we could see that the GAN-based cross-modal approaches (GXN, CM-GANs and ACMR) achieve promising results on the four datasets owing to the power of the adversarial learning. However, they can not completely outperform all traditional methods on all datasets. The reason may be their generated representations have some redundancy

Table 3
Performance comparison of cross-modal retrieval in terms of MAP scores on Pascal Sentence.

Method	Img2Txt	Txt2Img	Average
CCA [28]	0.170	0.165	0.168
MCCA [11]	0.571	0.574	0.573
PLS [12]	0.459	0.384	0.421
GMA [29]	0.544	0.565	0.554
MvDA [30]	0.583	0.591	0.587
MvDA-VC [4]	0.568	0.582	0.575
GSS-SL [15]	0.624	0.623	0.623
DCCA [34]	0.520	0.504	0.512
DCCAE [35]	0.527	0.534	0.531
GXN [22]	0.550	0.542	0.546
ACMR [26]	0.606	0.567	0.587
MCSM [46]	0.598	0.598	0.598
CMDN [52]	0.544	0.526	0.535
CCL [20]	0.576	0.561	0.569
CM-GANs [21]	0.603	0.604	0.604
CAN	**0.697**	**0.691**	**0.694**

Table 4
Performance comparison of cross-modal retrieval in terms of MAP scores on NUS-WIDE-10K.

Method	Img2Txt	Txt2Img	Average
CCA [28]	0.114	0.114	0.114
MCCA [11]	0.114	0.114	0.114
PLS [12]	0.346	0.273	0.309
GMA [29]	0.284	0.110	0.197
MvDA [30]	0.499	0.506	0.503
MvDA-VC [4]	0.445	0.465	0.455
GSS-SL [15]	0.542	0.557	0.550
DCCA [34]	0.435	0.441	0.438
DCCAE [35]	0.435	0.449	0.442
GXN [22]	0.280	0.310	0.295
ACMR [26]	0.531	0.536	0.533
CCL [20]	0.506	0.535	0.521
CMDN [52]	0.492	0.515	0.504
CAN	**0.562**	**0.573**	**0.568**

Table 5
Performance comparison of cross-modal retrieval in terms of MAP scores on XMediaNet.

Method	Img2Txt	Txt2Img	Average
CCA [28]	0.343	0.351	0.347
MCCA [11]	0.361	0.374	0.368
PLS [12]	0.100	0.062	0.081
GMA [29]	0.454	0.479	0.466
MvDA [30]	0.502	0.491	0.496
MvDA-VC [4]	0.467	0.431	0.449
GSS-SL [15]	0.505	0.493	0.499
DCCA [34]	0.289	0.311	0.300
DCCAE [35]	0.290	0.310	0.300
GXN [22]	0.120	0.129	0.125
ACMR [26]	0.479	0.519	0.528
MCSM [46]	0.540	0.550	0.545
CMDN [52]	0.485	0.516	0.501
CCL [20]	0.537	0.528	0.533
CM-GANs [21]	0.567	0.551	0.559
CAN	**0.670**	**0.661**	**0.665**

for cross-modal retrieval. With our proposed cross-modal discriminant mechanism, the generated representations are projected into a single unified space to further reduce the modality discrepancy and preserve more discriminative information for cross-modal retrieval. Therefore, our CAN outperforms all tested cross-modal methods. Although some traditional methods (*e.g.*, GSS-SL) can achieve satisfactory performance on small datasets (*i.e.* Wikipedia, Pascal Sentence and NUS-WIDE-10K datasets), they are not so good at handling the big dataset (*i.e.* XMediaNet dataset) comparing with the supervised deep methods. In conclusion, our CAN out-

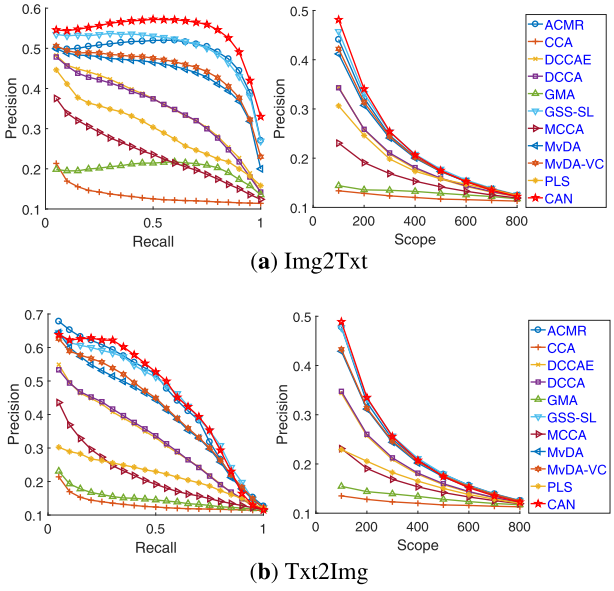


Fig. 3. Precision-recall and precision-scope curves for Img2Txt and Txt2Img on Wikipedia.

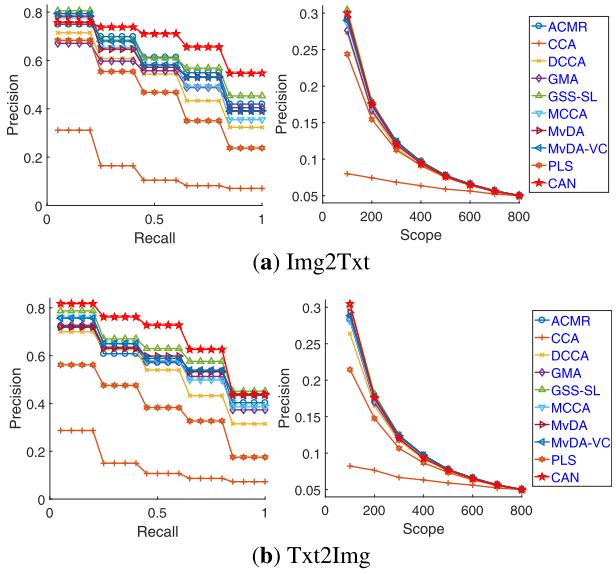


Fig. 4. Precision-recall and precision-scope curves for Img2Txt and Txt2Img on Pascal Sentence.

performs all the tested approaches on both small and big datasets, which indicates that our CAN is a good cross-modal learning approach for cross-modal retrieval. More detail analysis is given as follows.

4.4.1. Wikipedia dataset

The experimental results of performance comparison are shown in 2, which is in terms of MAP scores for cross-modal retrieval. From this table, we could see that our CAN achieves the best retrieval results comparing with 15 compared approaches. As can be seen, our CAN has improved the best competitor, CM-GANs, by 2.63% on average. CM-GANs also learn discriminative common representations by utilizing GAN to eliminate the modality discrepancy. Thus, this performance improvement clearly shows the advantage of applying CDM with GAN to extract more discrimination. On one hand, GSS-SL achieves the best cross-modal retrieval MAP score among the compared traditional approaches, which is closer to one DNN-based method MCSM. On the other hand, the 8 DNN-based approaches achieve greatly different retrieval MAP results, while several of them are outperformed by some traditional methods, *e.g.*, the average MAP scores of DCCA, DCCAE, GXN, CMDN and ACMR are lower than GSS-SL, and the average MAP scores of DCCA, DCCAE and GXN are also lower than traditional method MvDA-VC. In detail, the proposed method outperforms GSS-SL by 4.97%, MCSM by 4.11%, CM-GANs by 2.63% and ACMR by 6.07%, indicating that the proposed approach is a good multimodal learning method for cross-modal retrieval. In addition to evaluating the performance of the tested approaches from MAP scores, some precision-recall and precision-scope curves are plotted for visual comparison as shown in 3. The evaluated results of the precision-recall and precision-scope curves are consistent with the MAP results of cross-modal retrieval, where our CAN achieves the best performance comparing with its counterparts.

4.4.2. Pascal sentence dataset

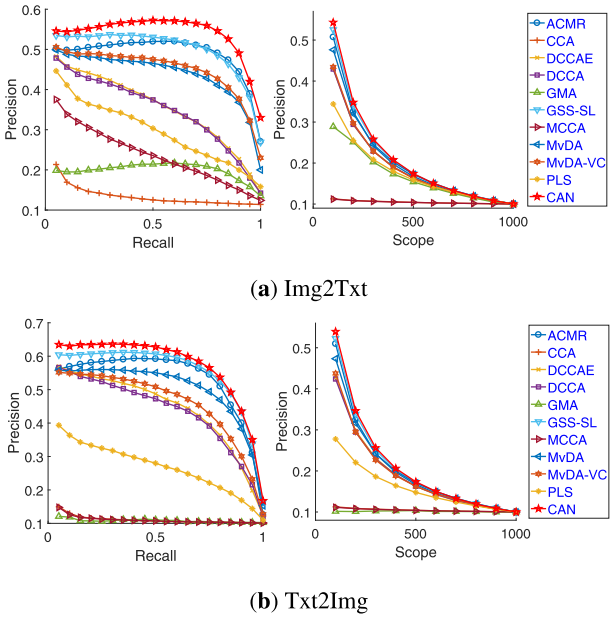
To evaluate the performance of our CAN, we also conduct experiments on another widely-used dataset, *i.e.*, Pascal Sentence. First, the retrieval MAP scores of all tested methods are shown in 3. From the experimental results, we could see that our proposed CAN outperforms all the other 15 compared state-of-the-art methods. As we can see, our CAN has improved the best competitor, GSS-SL, by 11.40% on average. While GSS-SL also learns a discriminative common subspace using the label information, our performance improvement clearly shows the advantage of applying discriminative adversarial learning for cross-modal retrieval. Moreover, the traditional method GSS-SL outperforms all the other compared DNN methods since the Pascal Sentence is too small (only 800 pairs in training set) to boost the performance of these DNN methods, which is to say that our proposed DNN framework can achieve better performance on small training data. To show more visualized results, we draw the cross-modal precision-recall and precision-scope curves in 4. From these figures, we could see that the visualized results are consistent with the retrieval MAP results, where our CAN outperforms all counterparts.

4.4.3. NUS-WIDE-10K dataset

The cross-modal retrieval results in terms of MAP scores are shown in 4. From the table, one could see that our CAN outperforms its 11 state-of-the-art counterparts on the NUS-WIDE-10K dataset. As shown in 4, our CAN has improved GSS-SL from 0.550 to 0.568 in terms of the average MAP score on the NUS-WIDE-10K dataset. It should be noted that GSS-SL achieves the best retrieval results comparing other compared methods. That is to say, our method can remarkedly improve cross-modal retrieval performance. In detail, the proposed method outperforms GSS-SL by 3.27%, CCL by 9.02%, CMDA by 12.70% and ACMR by 6.57% on average, indicating that the proposed approach is a good multimodal approach for cross-modal retrieval. For additional comparison, the precision-recall and precision-scope curves are plotted in 5. The experimental results are consistent with the retrieval MAP results in 4, where our CAN achieves the best performance.

4.4.4. XMediaNet dataset

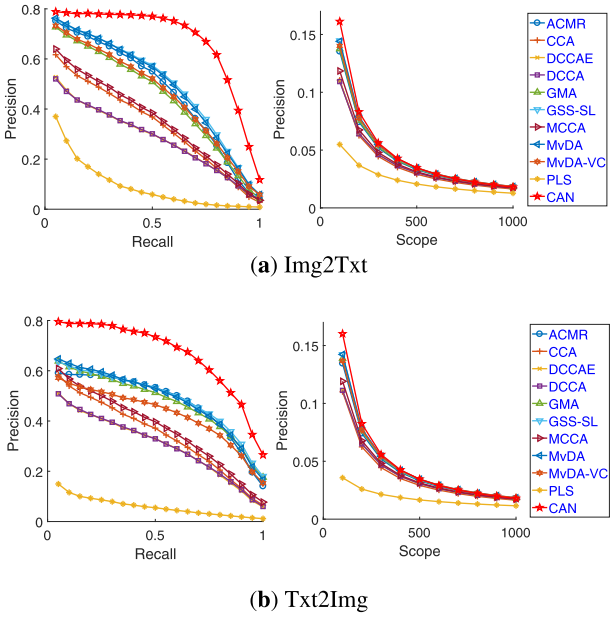
We also evaluate our CAN on the XMediaNet dataset for cross modal retrieval. The retrieval MAP results are shown in 5. From the experimental results, one could see that the proposed method achieves the best retrieval MAP scores comparing with its 15 counterparts. From the results, we could see that our CAN has improved the best competitor, CM-GANs, by 18.96% on average. Although some traditional methods (*e.g.*, GSS-SL) outperform some DNN-based approaches, DNN-based ones still maintain great advantages



(**a**) Img2Txt

(**b**) Txt2Img

Fig. 5. Precision-recall and precision-scope curves for Img2Txt and Txt2Img on NUS-WIDE-10K.



(**a**) Img2Txt

(**b**) Txt2Img

Fig. 6. Precision-recall and precision-scope curves for Img2Txt and Txt2Img on XMediaNet.

in handling the dataset. Specifically, except for DCCA and DCCAE, the other DNN methods all outperform the traditional approaches, which show the advantage of DNN to handling large-scale data. In detail, the proposed method outperforms GSS-SL by 33.27%, MCSM by 22.02%, CM-GANs by 18.96% and ACMR by 25.95% on average, indicating that our method is a good cross-modal learning method for cross-modal retrieval on the XMediaNet dataset. As the previous experiments, the precision-recall and precision-scope curves are plotted to investigate the effectiveness of different approaches for cross-modal retrieval on the XMediaNet dataset as shown in 6. From the visualized results, one could see that our CAN also outperforms all the compared approaches, which is consisted with the retrieval MAP results.

Table 6
Performance comparison of cross-modal retrieval in terms of MAP scores and covariance with GAN-based methods.

Dataset	Method	Covariance	Img2Txt	Txt2Img	Average
Wikipedia	GXN [22]	480.862	0.318	0.280	0.299
	ACMR [26]	17.236	0.493	0.462	0.478
	CAN	**1.541**	**0.540**	**0.474**	**0.507**
Pascal Sentence	GXN [22]	313.247	0.550	0.542	0.546
	ACMR [26]	108.415	0.606	0.567	0.587
	CAN	**3.161**	**0.697**	**0.691**	**0.694**
NUS-WIDE	GXN [22]	713.829	0.280	0.310	0.295
	ACMR [26]	15.457	0.531	0.536	0.533
	CAN	**1.047**	**0.562**	**0.573**	**0.568**
XMediaNet	GXN [22]	794.343	0.120	0.129	0.125
	ACMR [26]	23.654	0.479	0.519	0.528
	CAN	**6.705**	**0.670**	**0.661**	**0.665**

4.5. Redundancy analysis for GAN-based methods

In the section, we discuss the influence of the redundancy on the performance of cross-modal retrieval for GAN-based methods. We utilize the absolute magnitude of the covariance to measure the degree of redundancy among different dimensions following [27]. The experimental results are shown in 6. From the table, one could see that our method can efficiently remove the redundant information from the generated features (the discriminant latent space) and improve retrieval performance. Note that, both the reconstruction process and the adversarial learning will lead to redundancy to GXN [22]. Although ACMR [26] try to extract discrimination, the redundant information is still preserved and thus decreases the performance. By reducing the redundancy from the generated features, our CAN achieves the best performance.

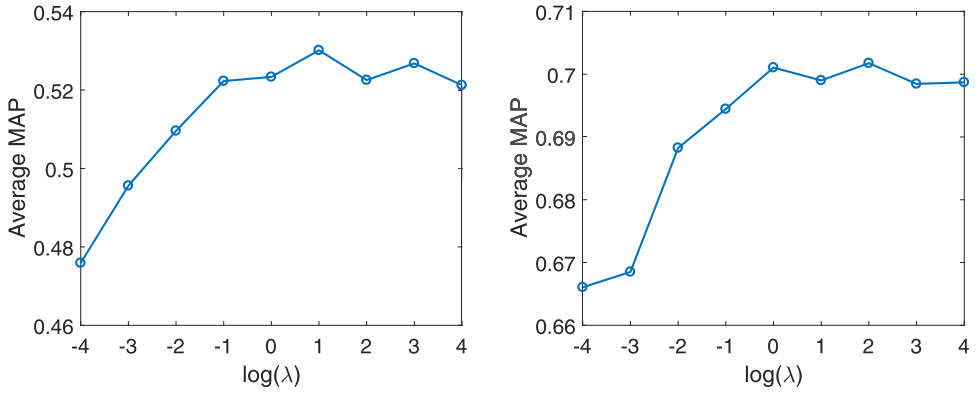
4.6. Parameter analysis

In the section, the impact of the parameter λ is investigated for our CAN in terms of the average MAP scores of cross-modal retrieval on the Pascal Sentence and Wikipedia datasets. In our experiments, the validation sets of these datasets are used to tune the hyper parameters. 7 shows the average retrieval MAP scores *vs.* different value of λ. From the experimental results, one could see that the average retrieval MAP scores of our CAN are impacted just marginally (by about 1%) when λ is within a scale of about four orders of magnitude (1 to 10,000). This verifies that the performance of our CAN is insensitive to λ in a wide range. Furthermore, one could see that the average MAP scores are much higher than other results at $\lambda = 0.0001$, which indicates that $\mathcal{L}_W$ is an important term to improve the performance of our CAN. By default, we set $\lambda = 1$ in our experiments.

Furthermore, we also investigate the influence of the threshold σ to the performance of cross-view retrieval in terms of the average MAP scores on the validation set of Pascal Sentence. 8 illustrates the average retrieval MAP scores *versus* different value of σ. From the figure, one could see that the average MAP scores achieve stable results from $\sigma = 0.5$ to $\sigma = 0.9$. That is to say, the average retrieval MAP scores of our CAN are insensitive to σ in a suitable range (*e.g.*, $[0.5, 0.9]$).

4.7. Convergence analysis

We also investigate the convergence of the proposed approach on Wikipedia and Pascal Sentence datasets.Fig. 9(a) plots the losses of generators and discriminators *versus* average MAP for cross-modal retrieval with the number of epochs increasing on the Wikipedia dataset. Similarly, Fig. 9(b) shows the losses *versus* the average MAP on the Pascal Sentence dataset. From 9, we see that the proposed method converges in $200 \sim 400$ epochs and the



(**a**) Wikipedia (**b**) Pascal Sentence

Fig. 7. Parameter analysis. Average MAP scores *versus* different values of λ for the cross-modal retrieval on the validation set of Wikipedia and Pascal Sentence, respectively.

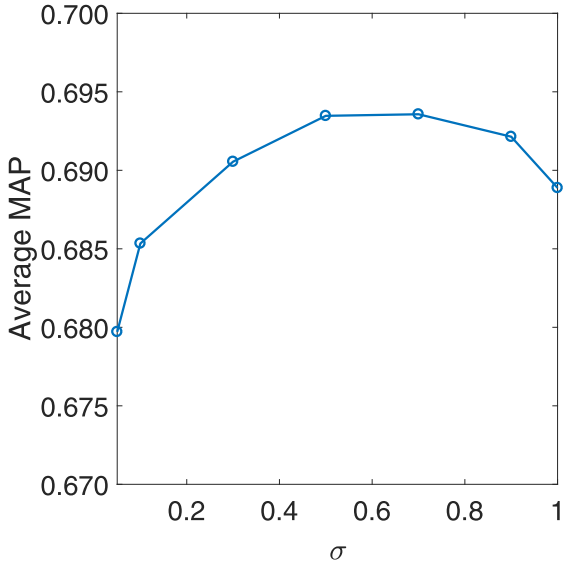


Fig. 8. Parameter analysis. Average MAP scores *versus* different values of σ for the cross-modal retrieval on the validation set of the Pascal Sentence dataset.

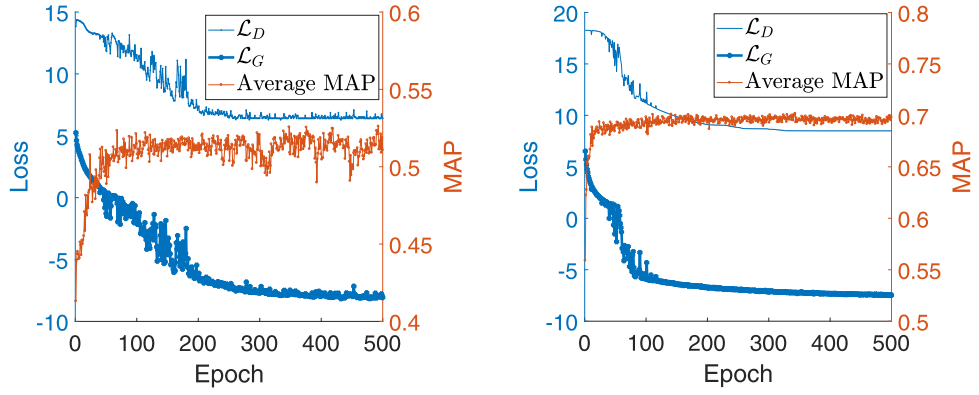
changing rates are much faster before the 200-th epoch than later epochs in these figures. Moreover, the proposed method achieves satisfactory cross-modal MAP scores before the 200-th epoch on the validations of Pascal Sentence and Wikipedia as shown in Fig. 9(a) and Fig. 9(b). Therefore, we set the maximum epoch as 200 in all experiments to reduce the training time.

4.8. Ablation study

In this section, we investigate the contribution of each component for cross-modal retrieval. We define the following three alternative baselines to study the influence of different components:

- **CAN-1** is one variant of our CAN without $\mathcal{L}_W$, which is used to investigate the effectiveness of the proposed $\mathcal{L}_W$ to improve the performance of our CDM.
- **CAN-2** trains the model using only $\mathcal{L}_W$, which is adopted to investigate the effectiveness of GAN for cross-modal retrieval.
- **CAN-3** trains the model using the ratio trace in Eq. (9), which is adopted to investigate the effectiveness of the proposed logarithmic eigenvalue-based constraint.
- **CAN-4** is one variant of our CAN to directly learn a shared common space without the proposed CDM, which is used to investigate the effectiveness of our CDM.

The optimization procedure and network architectures of these variations are as same as our CAN. 7 shows the performance of CAN and its four variations on the Wikipedia and Pascal Sentences datasets. We can see that both the discriminant GAN and cross-modal discriminant analysis contribute to the final retrieval scores, indicating that simultaneously optimizing the $\mathcal{L}_G$ and $\mathcal{L}_W$ in the proposed model performs better than optimizing only one of them.



(**a**) Wikipedia (**b**) Pascal Sentence

Fig. 9. Losses *versus* average MAP on Wikipedia and Pascal Sentence.

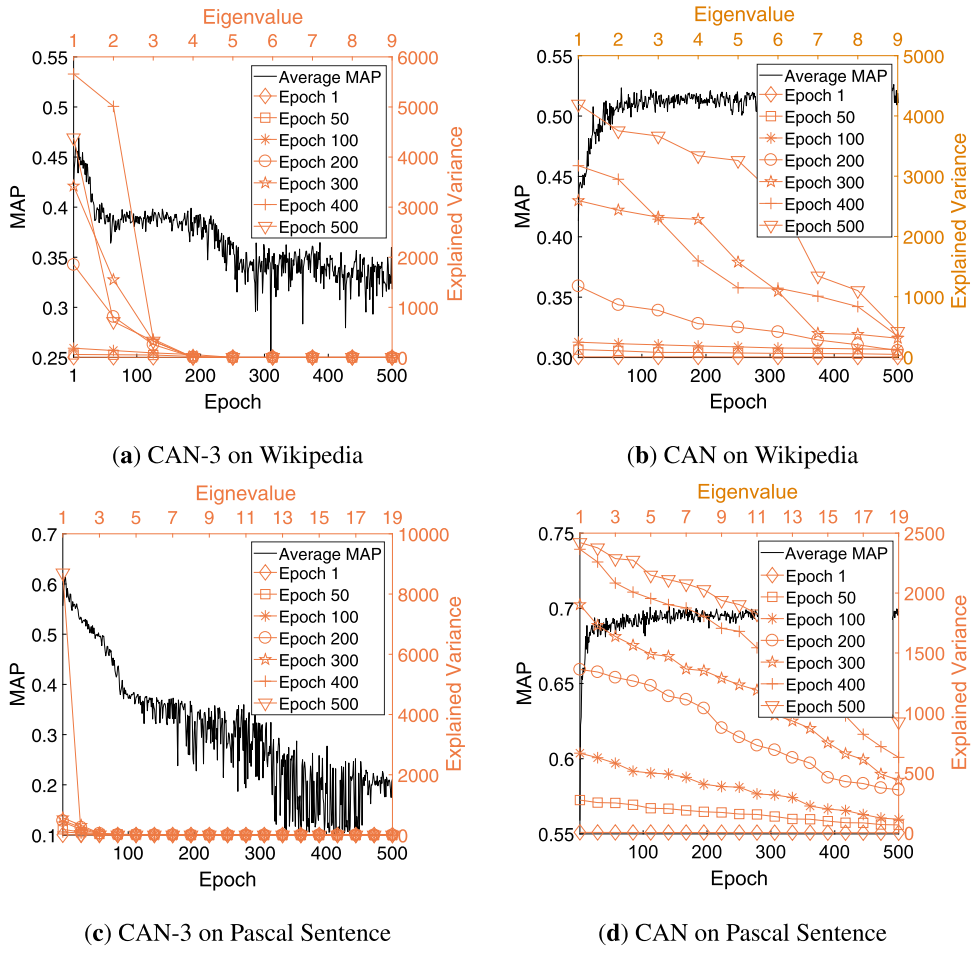


Fig. 10. Individual eigenvalues *versus* average MAP. The figure shows the evolution of individual eigenvalues and average MAP on the training set of the Wikipedia and Pascal Sentence datasets during the training stage.

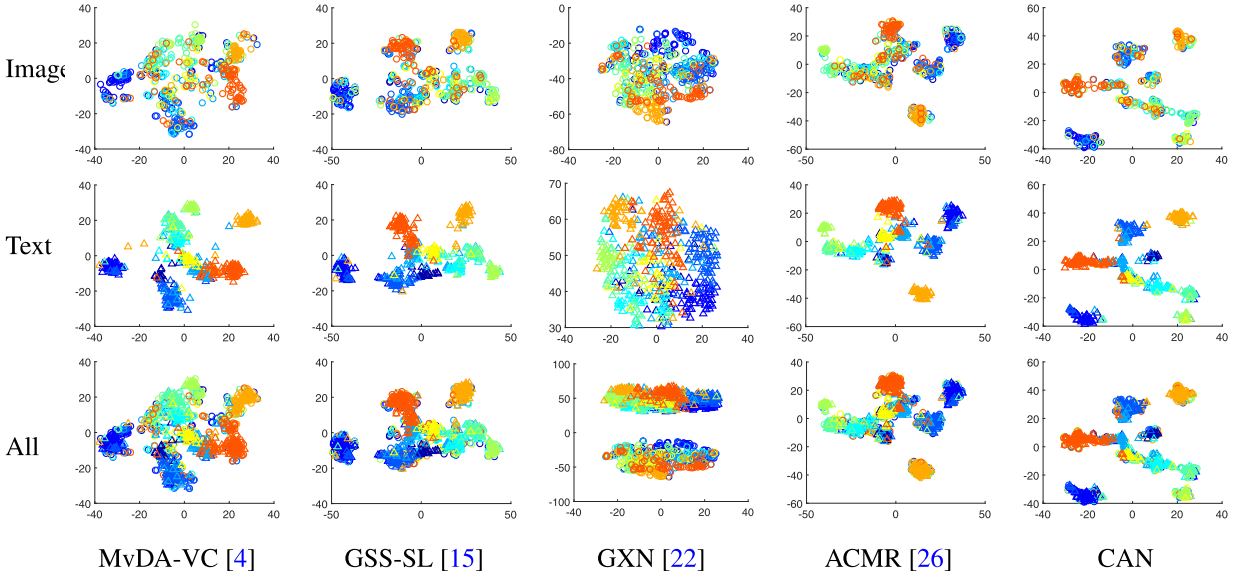


Fig. 11. The visualization for the test data on the Wikipedia dataset by using the t-SNE method [53]. In this figure, the different shape of markers represents its corresponding view, and the different colors denote their corresponding classes, respectively.

Table 7
Performance comparison of cross-modal retrieval in terms of MAP scores utilizing each component on the validations of Wikipedia and Pascal Sentence.

Dataset	Method	Img2Txt	Txt2Img	Average
Wikipedia	CAN-1	0.467	0.467	0.448
	CAN-2	0.532	0.463	0.497
	CAN-3	0.476	0.438	0.457
	CAN-4	0.411	0.374	0.392
	CAN	**0.540**	**0.474**	**0.507**
Pascal Sentences	CAN-1	0.649	0.666	0.658
	CAN-2	0.687	**0.699**	0.693
	CAN-3	0.579	0.631	0.605
	CAN-4	0.452	0.458	0.455
	CAN	**0.697**	0.691	**0.694**

Table 8
Inference time comparison on the test set of Pascal Sentence.

Method	Average Inference Time
GXN [22]	1.337s
ACMR [26]	0.008s
CAN	0.065s

From the experimental results, the eigenvalue-based discriminant term $\mathcal{L}_W$ obtains better performance than the discriminant GAN, since $\mathcal{L}_W$ aims to directly push discriminative information into the common space and the GAN achieves this goal with an indirect way (*i.e.*, an adversarial manner). Moreover, the performance of CAN-3 is worse than our CAN, which indicates that overemphasizing problem produces poor discrimination in the learned representations. Since CAN-4 achieves the worst performance compared with the other variations and our CAN, we can draw the conclusion that the proposed CDM can effectively extract more discriminative information from the generated rough features (the discriminant latent space) and improve the performance of the discriminant GAN.

To illustrate the relationship between eigenvalues and performance for cross-modal retrieval, we draw the curves about individual eigenvalues *versus* average MAP in 10. From the experimental results, we can see that the average MAP of cross-modal retrieval firstly increases as the eigenvalues become larger and larger. However, CAN-3 then overemphasizes the largest eigenvalues and produces worse performance, when the eigenvalues exceed a certain value. Conversely, the proposed logarithmic eigenvalue-based loss can avoid overemphasizing the dominant eigenvalues and ignoring the discriminative variances in the minor eigenvalues, so our CAN can achieve promising performance as shown in 10, which is consistent with the evaluations in 7. Therefore, our CAN can push as much discrimination as possible to the common space without overemphasizing the dominant eigenvalues.

4.9. Inference time comparison

To investigate the speed of the method in the real-world application scenario, we have conducted some comparison experiments with prior GAN-based methods (*i.e.*, ACMR [26] and GXN [22]) on the Pascal Sentence dataset. We run the trained model of each method to compute the inference time for 100 times on the test set of Pascal Sentence on 1 NVIDIA GeForce RTX 2080TI, respectively. The average inference time of each method is reported in 8. From the experimental results, we can see that our CAN costs a bit more time (*i.e.*, 0.057s) than ACMR and much less time (*i.e.*, 1.272s) than GXN, indicating that our method can achieve better performance with comparable efficiency. Therefore, our CAN can achieve sufficient efficiency in application scenario.

4.10. Visualization of the learned representation

To visually investigate the discrimination of common representations learned by different cross-modal methods, we adopt the *t*-SNE approach [53] to embed the samples from the Wikipedia dataset into a two-dimensional space as shown in 11. From this figure, we could see that the learned representations of these cross-modal methods from different modalities can overlap with each other indicating that they can project diverse modalities into one latent unified space, expect GXN. GXN is an unsupervised method that ignores the discriminative information in the cross-modal samples. Therefore, the discrimination in the multimodal data is important for cross-modal retrieval. From 11, we could see that these methods attempt to project diverse modalities into one unified space and separate the samples of different classes from each other. Obviously, our CAN can make the different classes more scattered and the same ones more compact. That is to say, the proposed CAN can obtain more discriminative information from the cross-modal data, which is consistent with the retrieval MAP scores of Img2Txt and Txt2Img.

5. Conclusion

In this paper, we propose CAN to eliminate the modality discrepancy for cross-modal retrieval and alleviate the redundancy of adversarial learning. The advantages of our CAN are three-fold: 1) the generated features are modality-invariant, 2) the common representations are cross-modal matching consistent, and 3) the discriminative information could be preserved while alleviating the redundancy. On the other words, our CAN can project the cross-modal data into a latent common space in which the discrimination can be preserved as much as possible with eliminating the cross-modal discrepancy. With these advantages, our method can own the advantage of the adversarial learning with alleviating its redundancy, and thus improve the performance of the cross-modal retrieval. Extensive experimental results show that our method achieves superiority over recently proposed methods on four widely-used datasets. Some extensive analysis of our CAN demonstrates the effectiveness of the proposed components. Recently, machine learning is creatively utilized to address many problems in healthcare (*e.g.*, pulmonary disease [54–56] and brain activity [57]), and more and more researchers are attracted to the community. In the future, we will explore how to adopt cross-modal learning in healthcare. Moreover, our CAN is a supervised method, which needs all the training data to be labeled. However, it is time- and cost-prohibitive to obtain well-labeled data, which will be more serious for multiple modalities. Therefore, as future work, we plan to investigate how to apply our CAN in a semi-supervised setting under which only a few samples are labeled.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grants U19A2078, 61971296, 61625204, 61836011, and 61806135; Sichuan Science and Technology Planning Projects under Grant 2020YFG0319, 2020YFH0186, and 2019YFG0495; the Fundamental Research Funds for the Central Universities under Grant YJ201949; and the Agency for Science, Technology and Research(A*STAR) under its AME Programmatic Funds (Project no. A1892b0026).

References

[1] Z. Kang, G. Shi, S. Huang, W. Chen, X. Pu, J.T. Zhou, Z. Xu, Multi-graph fusion for multi-view spectral clustering, Knowl. Based Syst. (2019) 105102.
[2] C. Zhang, H. Fu, Q. Hu, X. Cao, Y. Xie, D. Tao, D. Xu, Generalized latent multi--view subspace clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (1) (2018) 86–99. 1–1
[3] C. Lu, S. Yan, Z. Lin, Convex sparse spectral clustering: single-view to multi--view, IEEE Trans. Image Process. (TIP) 25 (6) (2016) 2833–2843.
[4] M. Kan, S. Shan, H. Zhang, S. Lao, X. Chen, Multi-view discriminant analysis, IEEE Trans. Pattern Anal. Mach. Intell. 38 (1) (2016) 188–194.
[5] X. Shen, W. Liu, I.W. Tsang, Q.-S. Sun, Y.-S. Ong, Multilabel prediction via cross-view search, IEEE Trans. Neural Netw. Learn. Syst. 29 (9) (2018) 4324–4338, doi:10.1109/TNNLS.2017.2763967.
[6] M. Yang, C. Deng, F. Nie, Adaptive-weighting discriminative regression for multi-view classification, Pattern Recognit. 88 (2019) 236–245.
[7] X. Peng, Z. Huang, J. Lv, H. Zhu, J.T. Zhou, COMIC: Multi-view clustering without parameter selection, in: Proceedings of the 36th International Conference on Machine Learning, 97, PMLR, 2019, pp. 5092–5101.
[8] M. Kan, S. Shan, X. Chen, Multi-view deep network for cross-view classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4847–4855.
[9] X. Liu, L. Huang, C. Deng, B. Lang, D. Tao, Query-adaptive hash code ranking for large-scale multi-view visual search, IEEE Trans. Image Process. 25 (10) (2016) 4514–4524, doi:10.1109/TIP.2016.2593344.
[10] P. Hu, H. Zhu, X. Peng, J. Lin, Semi-supervised multi-modal learning with balanced spectral decomposition, in: Proceedings of the AAAI Conference on Artificial Intelligence, 34, 2020, pp. 99–106.
[11] J. Rupnik, J. Shawe-Taylor, Multi-view canonical correlation analysis, in: Conference on Data Mining and Data Warehouses (SiKDD 2010), 2010, pp. 1–4.
[12] A. Sharma, D.W. Jacobs, Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 593–600.
[13] X. Xu, F. Shen, Y. Yang, H.T. Shen, L. He, J. Song, Cross-modal retrieval with label completion, in: Proceedings of the 24th ACM international conference on Multimedia, 2016, pp. 302–306.
[14] V.E. Liong, J. Lu, Y.-P. Tan, Cross-modal discrete hashing, Pattern Recognit. 79 (2018) 114–129.
[15] L. Zhang, B. Ma, G. Li, Q. Huang, Q. Tian, Generalized semi-supervised and structured subspace learning for cross-modal retrieval, IEEE Trans. Multimed. 20 (1) (2018) 128–141.
[16] Y. Fang, Y. Ren, Supervised discrete cross-modal hashing based on kernel discriminant analysis, Pattern Recognit. 98 (2020) 107062.
[17] W. Wang, K. Livescu, Large-scale approximate kernel canonical correlation analysis, in: International Conference on Learning Representations (ICLR), 2016.
[18] T. Yao, G. Wang, L. Yan, X. Kong, Q. Su, C. Zhang, Q. Tian, Online latent semantic hashing for cross-media retrieval, Pattern Recognit. 89 (2019) 1–11.
[19] X. Peng, S. Xiao, J. Feng, W. Yau, Z. Yi, Deep subspace clustering with sparsity prior, in: Proceedings of the 25 International Joint Conference on Artificial Intelligence, New York, NY, USA, 2016, pp. 1925–1931.
[20] Y. Peng, J. Qi, X. Huang, Y. Yuan, CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network, IEEE Trans. Multimed. 20 (2) (2018) 405–420, doi:10.1109/TMM.2017.2742704.
[21] Y. Peng, J. Qi, CM-GANs: Cross-modal generative adversarial networks for common representation learning, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 15 (1) (2019) 22.
[22] J. Gu, J. Cai, S. Joty, L. Niu, G. Wang, Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7181–7189.
[23] P. Hu, D. Peng, X. Wang, Y. Xiang, Multimodal adversarial network for cross–modal retrieval, Knowl Based Syst 180 (2019) 38–50.
[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.
[25] P. Ge, C.-X. Ren, D.-Q. Dai, J. Feng, S. Yan, Dual adversarial autoencoders for clustering, IEEE Trans. Neural. Netw. Learn. Syst. (2019) 1–8, doi:10.1109/TNNLS.2019.2919948.
[26] B. Wang, Y. Yang, X. Xu, A. Hanjalic, H.T. Shen, Adversarial cross-modal retrieval, in: Proceedings of the 2017 ACM on Multimedia Conference, ACM, 2017, pp. 154–162.
[27] Shlens J. A tutorial on principal component analysis, arXiv preprint arXiv:1404.1100, 2014.
[28] H. Hotelling, Relations between two sets of variates, Biometrika 28 (3/4) (1936) 321–377.
[29] A. Sharma, A. Kumar, H. Daume, D.W. Jacobs, Generalized multiview analysis: A discriminative latent space, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 2160–2167.
[30] M. Kan, S. Shan, H. Zhang, S. Lao, X. Chen, Multi-view discriminant analysis, in: European Conference on Computer Vision, 2012, pp. 808–821.
[31] S. Akaho, A kernel method for canonical correlation analysis, in: International Meeting of Psychometric Society, 2001, pp. 263–269.
[32] X. Peng, J. Feng, S. Xiao, W.Y. Yau, J.T. Zhou, S. Yang, Structured autoencoders for subspace clustering, IEEE Trans. Image Process. 27 (10) (2018) 5076–5086, doi:10.1109/TIP.2018.2848470.
[33] J. Hu, J. Lu, Y.-P. Tan, Sharable and individual multi-view metric learning, IEEE Trans. Pattern Anal. Mach. Intell. 40 (9) (2017) 2281–2288.
[34] G. Andrew, R. Arora, J. Bilmes, K. Livescu, Deep canonical correlation analysis, in: International Conference on Machine Learning, 2013, pp. 1247–1255.
[35] W. Wang, R. Arora, K. Livescu, J. Bilmes, On deep multi-view representation learning, in: International Conference on Machine Learning, 2015, pp. 1083–1092.
[36] Y. Duan, W. Zheng, X. Lin, J. Lu, J. Zhou, Deep adversarial metric learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2780–2789.
[37] Y. Wei, S. Liu, W. Zhao, J. Lu, Conditional single-view shape generation for multi-view stereo reconstruction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9651–9660.
[38] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces *vs.* fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720, doi:10.1109/34.598228.
[39] A. Stuhlsatz, J. Lippel, T. Zielke, Feature extraction with deep neural networks by a generalized discriminant analysis, IEEE Trans. Neural Netw. Learn. Syst. 23 (4) (2012) 596–608.
[40] M. Dorfer, R. Kelz, G. Widmer, Deep linear discriminant analysis, in: International Conference on Learning Representations (ICLR), 2016.
[41] L. Wu, C. Shen, A. van den Hengel, Deep linear discriminant analysis on fisher networks: a hybrid architecture for person re-identification, Pattern Recognit. 65 (2017) 238–250.
[42] P. Hu, D. Peng, Y. Sang, Y. Xiang, Multi-view linear discriminant analysis network, IEEE Trans. Image Process. 28 (11) (2019) 5352–5365.
[43] 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: Proceedings of the international conference on Multimedia, ACM, 2010, pp. 251–260.
[44] C. Rashtchian, P. Young, M. Hodosh, J. Hockenmaier, Collecting image annotations using amazon's mechanical turk, in: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Association for Computational Linguistics, 2010, pp. 139–147.
[45] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y.-T. Zheng, Nus-wide: A real-world web image database from national university of singapore, in: Proc. of ACM Conf. on Image and Video Retrieval (CIVR'09), Santorini, Greece., 2009.
[46] Y. Peng, J. Qi, Y. Yuan, Modality-specific cross-modal similarity measurement with recurrent attention network, IEEE Transactions on Image Processing 27 (11) (2018) 5585–5599. 1–1
[47] 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.
[48] Y. Peng, X. Huang, Y. Zhao, An overview of cross-media retrieval: concepts, methodologies, benchmarks and challenges, IEEE Trans. Circuits Syst. Video Technol. (2017).
[49] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, International Conference on Learning Representations (2015).
[50] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, 2013, pp. 3111–3119.
[51] Y. Kim, Convolutional neural networks for sentence classification, Empirical Methods in, Natural Language Processing (2014).
[52] Y. Peng, X. Huang, J. Qi, Cross-media shared representation by hierarchical learning with multiple deep networks, in: IJCAI, 2016, pp. 3846–3853.
[53] L.v.d. Maaten, G. Hinton, Visualizing data using t-sne, J. Mach. Learn. Res. 9 (Nov) (2008) 2579–2605.
[54] G. Altan, Y. Kutlu, A. Gokcen, Chronic obstructive pulmonary disease severity analysis using deep learning, Turk. J. Elec. Eng. Comp. Sci. 28 (5) (2020) 2979–2996, doi:10.3906/elk-2004-68.
[55] G. Altan, Y. Kutlu, N. Allahverdi, Deep learning on computerized analysis of chronic obstructive pulmonary disease, IEEE J. Biomed. Health Inform. 24 (5) (2019) 1344–1350.
[56] G. Altan, Y. Kutlu, Hessenberg elm autoencoder kernel for deep learning, J. Eng. Technol. Appl. Sci. 3 (2) (2018) 141–151.
[57] G. Altan, Y. Kutlu, Generative autoencoder kernels on deep learning for brain activity analysis, Nat. Eng. Sci. 3 (3) (2018) 311–322.

Peng Hu received the B.Eng. degree in computer science and technology from the Southwest University of Science and Technology in 2013, and the M.Sc. and Ph.D. degree in computer science and technology from Sichuan University, China, in 2016 and 2019. He is currently a Research Scientist at the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore. His current research interests include multi-view learning, cross-media analysis, and deep learning.

Xi Peng received the Ph.D. degree in computer science from Sichuan University, Chengdu, China, in 2013. He currently is a Research Professor with the College of Computer Science, Sichuan University. His current research interests include unsupervised representation learning and differentiable programming, as well as their applications in computer vision and image processing. In these areas, he has authored over 40 papers.

Hongyuan Zhu is currently a Research Scientist with the Institute for Infocomm Research, A*STAR, Singapore. His research interests include multimedia content analysis and segmentation, specially image segmentation/cosegmentation, object detection, scene recognition, and saliency detection.

Jie Lin received the B.S. and Ph.D. degrees from the School of Computer Science and Technology, Beijing Jiaotong University, Beijing, China, in 2006 and 2014, respectively. He is currently a Research Scientist with the Institute of Infocomm Research, A*STAR, Singapore. He was previously a visiting student in the Rapid-Rich Object Search Laboratory, Nanyang Technological University, Singapore, and the Institute of Digital Media, Peking University, Beijing, China, from 2011 to 2014. His research interests include deep learning, feature coding and large-scale image/video retrieval. His work on image feature coding has been recognized as core contribution to the MPEG-7 Compact Descriptors for Visual Search (CDVS) standard.

Liangli Zhen received the PhD degree in Computer Science from Sichuan University, China, in 2018. He is currently a Research Scientist with the Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore. His research focuses on representation learning and evolutionary optimisation, including deep neural networks, multimodal learning, multiobjective optimisation, and their applications in subspace clustering, cross-modal retrieval, and search-based software engineering.

Wei Wang received the B.Sc. degree in information and electronic engineering from Zhejiang University, Hangzhou, China, and the M.Sc. degree in signal and informa-Chengdu, China, in 1997 and 2004, respectively. Currently he is a D.Eng. candidate in the College of Computer Science at Sichuan University. His main research interests include topic modeling and neural networks.

Dezhong Peng received the B.Sc. degree in applied mathematics, and the M.Sc. and Ph.D. degrees in computer software and theory from the University of Electronic Science and Technology of China, Chengdu, China, in 1998, 2001, and 2006, respectively. From 2001 to 2007, he was with the University of Electronic Science and Technology of China as an Assistant Lecturer and a Lecturer. He was a Post-Doctoral Research Fellow with the School of Engineering, Deakin University, Australia, from 2007 to 2009. He is currently a Professor with the Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, China. His research interests include blind signal processing and neural networks.