Knowledge-Based Systems 180 (2019) 38–50

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Multimodal adversarial network for cross-modal retrieval

Peng Hu [a], Dezhong Peng [a,b,c,*], Xu Wang [a], Yong Xiang [d]

[a] Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu 610065, China
[b] Chengdu Sobey Digital Technology Co., Ltd., Chengdu 610041, China
[c] Shenzhen Cyberspace Laboratory, Shenzhen 518055, China
[d] School of Information Technology, Deakin University, Victoria 3125, Australia

ARTICLE INFO

Article history:
Received 17 January 2019
Received in revised form 10 May 2019
Accepted 10 May 2019
Available online 14 May 2019

Keywords:
Cross-modal retrieval
Latent common space
Adversarial learning
Multimodal discriminant analysis
Multimodal representation learning

ABSTRACT

Cross-modal retrieval aims to retrieve the pertinent samples across different modalities, which is important in numerous multimodal applications. It is challenging to correlate the multimodal data due to a large heterogeneous gap between distinct modalities. In this paper, we propose a Multimodal Adversarial Network (MAN) method to project the multimodal data into a common space wherein the similarities between different modalities can be directly computed by the same distance measurement. The proposed MAN consists of multiple modality-specific generators, a discriminator and a multimodal discriminant analysis (MDA) loss. With the adversarial learning, the generators are pitted against the discriminator to eliminate the cross-modal discrepancy. Furthermore, a novel MDA loss is proposed to preserve as much discrimination as possible into all available dimensions of the generated common representations. However, there are some problems in directly optimizing the MDA trace criterion. To be specific, the discriminant function will overemphasize 1) the large distances between already separated classes, 2) and the dominant eigenvalues. These problems may cause poor discrimination of the common representations. To solve these problems, we propose a between-class strategy and an eigenvalue strategy to weaken the largest between-class differences and the dominant eigenvalues, respectively. To the best of our knowledge, the proposed MAN could be one of the first works to specifically design for the multimodal representation learning (more than two modalities) with adversarial learning. To verify the effectiveness of the proposed method, extensive experiments are carried out on four widely-used multimodal databases comparing with 16 state-of-the-art approaches.

1. Introduction

With the rapid growth of multimodal data such as image, text and audio on the Internet, there are increasing demands on developing cross-modal methods for a variety of applications. Among them, cross-modal retrieval is an important application, which aims to retrieve interested contents across different modalities [1]. However, the samples from different modalities may lie in completely disparate spaces due to the large gap between different modalities [2–5]. Therefore, it is still challenging to deal with multimodal data.

To address the above mentioned cross-modal retrieval problem, numerous cross-modal methods were proposed to project multimodal data into a common space. These approaches can be roughly classified into two categories: traditional approaches [6–12] and deep models [13–18]. The traditional approaches usually learn modality-specific transformations to map the multimodal data into a latent common space. One typical scheme is to maximize the correlations between all possible pairwise modalities [6–8]. To utilize the label information, some semi-supervised and supervised multimodal methods were proposed to preserve the discrimination into the common representations [9–11,19]. However, all of them are linear methods and may be incapable of capturing the high-level nonlinear semantics of real-world data. Although they can be easily extended to nonlinear models with the kernel trick [20,21], the learned representation is limited due to the predetermined kernel. In addition, it is still an open issue to choose a suitable kernel function [22]. To overcome the aforementioned problems, several recent works attempted to use Deep Neural Network (DNN) to nonlinearly learn common representations across different modalities in an unsupervised [13,14] or supervised manner [15–17]. Very recently, inspired by the strong ability of the Generative Adversarial Nets (GAN) [23] in modeling data distribution, the adversarial learning strategy is introduced to model the joint distribution over the multimodal data to learn a latent common space [16,24–26]. However, they only focus on two modalities and no trial proves their ability to project the multimodal data (more than two modalities) into a single common space.

* Corresponding author at: Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu 610065, China.
E-mail address: pengdz@scu.edu.cn (D. Peng).

https://doi.org/10.1016/j.knosys.2019.05.017
0950-7051/

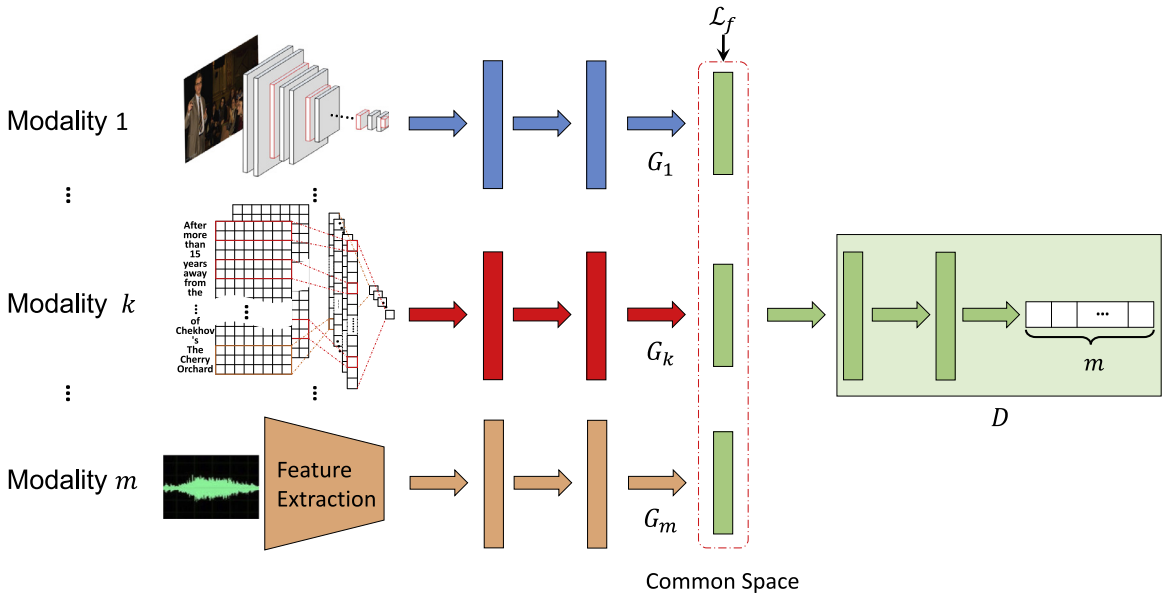


Fig. 1. The framework of our method. G_k is a modality-specific generator for the kth modality and D is the modality discriminator. $\mathcal{L}_f$ is the multimodal discriminant analysis loss, which aims to preserve the discrimination of multimodal data in the common space.

Based on the above observations, we propose a Multimodal Adversarial Network (MAN) for cross-modal retrieval. Different from existing two-modality GAN-based methods [16,24,25], MAN consists of multiple modality-specific generative subnetworks and one discriminative subnetwork as shown in Fig. 1. The generators and discriminator are pitted against each other to obtain modality-invariant representations. Meanwhile, a novel multimodal discriminant analysis (MDA) is combined with adversarial learning to embed discrimination into the common representations. To the best of our knowledge, the proposed MAN could be one of the first works to specifically design for the multimodal representation learning (more than two modalities) with adversarial learning. Through our proposed model, multiple modalities can be projected into a latent common space. In this space, the within-class samples can be compacted and the between-class samples can be scattered.

Besides the contribution to novel multimodal adversarial network, we also contribute to a general ratio trace optimization. To be exact, our objective needs directly optimizing the MDA ratio trace, but the discriminant function will overemphasize (1) the large distances between already separated classes, (2) and the dominant eigenvalues of the ratio trace. These problems may cause poor discrimination of the common representations. Therefore, we propose two simple but effective strategies to solve the problems. By our proposed strategies, the proposed model can push as much discrimination as possible into all the available dimensions of the common space. The main contributions of this paper are summarized as follows:

- A multimodal generative adversarial network is proposed to eliminate the cross-modal discrepancy. With adversarial learning, the modality-specific generators try their best to generate modality-invariant representations.
- A novel between-class strategy is proposed to reduce the influence of large distances between the different classes. Therefore, the discriminant function will not overemphasize the large distances between already separated classes, and therefore more discrimination can be preserved in the common space.
- A novel eigenvalue strategy is proposed to weaken the dominant eigenvalues and emphasize the minor ones, *i.e.*, all of them can be maximized without overemphasizing only the largest ones. Therefore, the model can push as much discriminative variance of the eigenvalues as possible into all the common dimensions.

Specifically, the multimodal adversarial learning is used to eliminate the cross-modal discrepancy, and MDA is used to extract the discrimination from the multimodal data. Moreover, the two proposed strategies are used to address the problems in the ratio trace of MDA for more discrimination. Overall speaking, these three contributions are a unified whole of our supervised model to extract modality-invariant and discriminative representations from the multiple modalities. With the learned common representations, the cross-modal data can be correlated by a common distance metric. To evaluate the performance of our MAN, extensive experiments are conducted on cross-modal retrieval which aims to retrieve the pertinent samples across different modalities on the learned common representations. The effectiveness of the proposed method is verified by extensive experiments carried out on the widely-used Reuters, PKU XMedia, Wikipedia and Pascal Sentence datasets, in comparison with 16 state-of-the-art approaches.

The remainder of this paper is organized as follows: Section 2 introduces related works, Section 3 details the proposed method, Section 4 evaluates the proposed method on four different datasets, and Section 5 concludes this paper.

2. Related works

In this section, the related works, which are most close to our work, are briefly reviewed from the following two aspects: traditional multimodal representation learning methods and deep multimodal representation learning methods.

2.1. Traditional multimodal representation learning methods

One typical method is the well-known Canonical Correlation Analysis (CCA) [6] which maximizes the cross-modal correlation to learn a common space. Another unsupervised cross-modal method is Partial Least Squares (PLS) [8] which attempts to learn two linear transformations by maximizing the covariance of two

modalities. However, they are specially designed for two modalities and cannot handle multimodal data. To overcome this limitation, Multiset CCA (MCCA) [7,27] was proposed to learn a common space by maximizing the correlations between all possible pairwise modalities. To utilize the label information, several semi-supervised and supervised methods were proposed to learn the discriminative common representations through different ways, such as classification [19,28], Frobenius norm [29] and Fisher's criterion [30–32]. In [19], Generalized Semi-supervised Structured Subspace Learning (GSS-SL) was proposed to project multimodal data into a discriminative common space by taking the label space as a linkage to model the cross-modal correlations. With the well-known Fisher's criterion, some approaches were proposed to learn a discriminative common space by maximizing the between-class variations and simultaneously minimizing the within-class variations [9–11]. However, they are linear methods. To extend them to nonlinear models, some kernel methods are proposed in decades, such as Kernel CCA (KCCA) [33], Kernel Nonlinear Orthogonal Iterations (KNOI) [21].

2.2. Deep multimodal representation learning methods

Over the past few years, the deep neural network (DNN) has achieved great success in numerous single-modality problems such as object detection, image classification and clustering [34]. Furthermore, DNN has also widely been utilized to project multimodal data into a common space [13–16,35]. In [13], Deep Canonical Correlation Analysis (DCCA) was proposed to learn complex nonlinear transformations of two-modality data so that the resulting representations are highly linearly correlated. Inspired by both DCCA and reconstruction-based objectives, Deep Canonically Correlated Autoencoders (DCCAE) was proposed in [14]. Alternatively, DCCAE can be seen as adding an autoencoder regularization term to DCCA. However, they are unsupervised methods and ignore some semantic information, such as the class label. To use this information, Multi-view Deep Network (MvDN) [15] was proposed to learn modality-invariant representation by introducing the Fisher's loss into a feedforward neural network. In [36], a Cross-Media Multiple Deep Network (CMDN) approach was proposed to jointly model the intra- and inter-modality correlation in both separate representation and common representation learning stages with multiple deep networks. Moreover, Deep Coupled Metric Learning (DCML) [37] adopts a coupled feedforward neural networks to learn two sets of hierarchical nonlinear transformations (one set for each modality) so that cross-modal samples are nonlinearly mapped into a shared latent feature subspace. In [38], a Cross-modal Correlation Learning (CCL) approach was proposed to fully explore both intra- and inter-modality correlation simultaneously with multi-grained and multi-task learning. In [39], a Cross-modal Bidirectional Translation (CBT) approach was proposed to treat images as a special kind of language to provide visual descriptions, so that translation can be conduct between bilingual pair of image and text to effectively explore cross-modal correlation. Moreover, CBT further explores the utilization of reinforcement learning to improve the translation process.

In [40], Generative Adversarial Networks (GAN) were proposed to estimate the generative model via an adversarial process, which consists of two models: a generative model G and a discriminative model D. This framework corresponds to a minimax two-player game, in which D aims to discriminate the generated data as false from training data, while G is expected to generate the data to "fool" D as the real training data. Recently, some works were proposed to generate common representation using GAN. In [24], Wang et al. presented a novel Adversarial Cross-Modal Retrieval (ACMR) method to seek an effective common subspace based on adversarial learning, which consists of a feature projector, a modality classifier and a triplet constraint. Moreover, Cross-modal Generative Adversarial Networks (CM-GANs) approach was proposed to learn discriminative common representation for bridging the heterogeneity gap by utilizing the power of GAN to model the cross-modal joint distribution in [16]. In other words, CM-GANs approach aims to effectively correlate existing large-scale heterogeneous data of different modalities. In [25], Gu et al. proposed to incorporate generative processes into the cross-modal feature embedding, through which it is able to learn not only the global abstract features but also the local grounded features. However, these GAN-based cross-modal methods are all specifically designed for the case of two modalities and cannot handle the multiple (more than two) modalities.

3. Multimodal adversarial network

The overall framework of the proposed method is shown in Fig. 1. The model has multiple modality-specific generators and a multimodal discriminator. The generators are pitted against the discriminator to eliminate the cross-modal discrepancy. Furthermore, MDA is proposed to combine with adversarial learning to preserve the discrimination into the common space.

3.1. Problem formulation

For clear description, we first give some definitions. The kth modality data are denoted as $\mathcal{X}^k = \{\mathbf{x}_{ij}^k | i = 1, 2, \ldots, c; j = 1, 2, \ldots, N_i^k; k = 1, \ldots, m\}$, where $\mathbf{x}_{ij}^k$ denotes the jth sample from the kth modality of the ith class; c is the number of classes; N^k is the number of samples from the kth modality; and m is the total number of modalities. The kth generator and the discriminator are denoted as nonlinear functions $G_k(\cdot; \Theta_{G_k})$ and $D(\cdot; \Theta_D)$, where Θ_{G_k} and Θ_D are their parameters, respectively.

Our proposed MAN is equipped with m generators and one discriminator as shown in Fig. 1. The generators $G_k|_{k=1}^m$ attempt to map a sample from the corresponding modality into a latent common representation $\mathbf{y}_{ij}^k = G_k(\mathbf{x}_{ij}^k)$, which is expected to be modality-invariant and discriminative. On the other hand, the discriminator D is pitted against the multiple modality-specific generators to constrain the generated common representations to be modality-invariant. Specifically, the discriminator classifies the generated representations $\mathbf{y}^k$ as the corresponding modality k, meanwhile the generators attempt to generate modality-invariant representations to fool the discriminator. Therefore, the model can produce common representations from this adversarial learning process. To preserve the discrimination of the multimodal data, we employ multimodal discriminant analysis (MDA) on the generated representations combining with adversarial learning. Different from the minmax game of the traditional GANs, the overall objective function of our MAN is formulated as follows:

$$\arg\min_{\Theta_D, \Theta_{G_1}, \ldots, \Theta_{G_m}} \left(\mathcal{L}_G + \mathcal{L}_D + \beta \mathcal{L}_f\right), \tag{1}$$

where $\mathcal{L}_G$, $\mathcal{L}_D$ and $\mathcal{L}_f$ are respectively the loses of generators, discriminator and MDA, which are introduced in detail in the following sections.

3.2. Multimodal adversarial learning

Firstly, the label of kth modality is denoted as a one-hot vector

$$l_k = [\ \overbrace{\underbrace{0, \ldots, 0}_{k-1}, 1, \underbrace{0, \ldots, 0}_{m-k}}^{m}\ ]. \tag{2}$$

The generator G_k attempts to transform an input sample $\mathbf{x}_{ij}^k$ into a latent common representation

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

On the other hand, the discriminator aims to classify the generated representation $\mathbf{y}_{ij}^k$ as its corresponding modality l_k. Meanwhile, the modality-specific generators attempt to fool the discriminator. Different from the minmax game of the traditional GANs, the kth fake modality label $\hat{l}_k$ is defined to guide the adversarial learning. $\hat{l}_k$ is denoted as follows:

$$\hat{l}_k = \frac{1}{m-1}[\overbrace{\underbrace{1,\ldots,1}_{k-1},0,\underbrace{1,\ldots,1}_{m-k}}^{m}]. \tag{4}$$

Overall speaking, the discriminator attempts to classify the generated representation $\mathbf{y}_{ij}^k$ as its corresponding modality l_k. Then the loss function of the discriminator D can be formulated as:

$$\mathcal{L}_D = \sum_{k=1}^{m} \frac{1}{2} \mathbb{E}_{\mathbf{x}^k \sim p_{\mathbf{x}^k}(\mathbf{x}^k)} \left[\left\| D(G_k(\mathbf{x}^k)) - l_k \right\|_2^2 \right], \tag{5}$$

where $\|\cdot\|_2$ is ℓ_2-norm. Meanwhile, the modality-specific generators aim to generate modality-invariant representations to fool the discriminator. Therefore, the loss function of the generators can be formulated as follows:

$$\mathcal{L}_G = \sum_{k=1}^{m} \frac{1}{2} \mathbb{E}_{\mathbf{x}^k \sim p_{\mathbf{x}^k}(\mathbf{x}^k)} \left[\left\| D(G_k(\mathbf{x}^k)) - \hat{l}_k \right\|_2^2 \right]. \tag{6}$$

3.3. Multimodal discriminant analysis

The common representations are expected to be discriminative and modality-invariant, *i.e.*, the same classes are compacted and the different classes are scattered in the latent common space. In order to formulate criteria for class separability, we should construct such objective function $\mathcal{L}_f$, which should be smaller when the between-class scatter is larger and the within-class scatter is smaller [41]. With the Fisher's criterion, one typical trace criterion can be formulated as:

$$\widehat{\mathcal{L}}_f = -\mathrm{Tr}\left(\frac{\widehat{\mathbf{S}}_B}{\mathbf{S}_W} \right), \tag{7}$$

where $\mathrm{Tr}(\cdot)$ is the trace operator, $\mathbf{S}_W$ is the within-class scatter matrix, and $\widehat{\mathbf{S}}_B$ is the between-class scatter matrix. Moreover, the within-class scatter matrix is defined as

$$\mathbf{S}_W = \sum_{i=1}^{c} \sum_{k=1}^{m} \sum_{j=1}^{N_i^k} \left(\mathbf{y}_{ij}^k - \boldsymbol{\mu}_i \right) \left(\mathbf{y}_{ij}^k - \boldsymbol{\mu}_i \right)^T, \tag{8}$$

and the between-class scatter matrix is defined as

$$\widehat{\mathbf{S}}_B = \sum_{i=1}^{c} N_i \left(\boldsymbol{\mu}_i - \boldsymbol{\mu} \right) \left(\boldsymbol{\mu}_i - \boldsymbol{\mu} \right)^T, \tag{9}$$

where $\boldsymbol{\mu}_i = \frac{1}{N_i} \sum_{k=1}^{m} \sum_{j=1}^{N_i^k} \mathbf{y}_{ij}^k$ is the mean of all samples in the ith class from the kth modality; N_i^k is the number of samples from the kth modality of the ith class; $\boldsymbol{\mu} = \frac{1}{N} \sum_{k=1}^{m} \sum_{i=1}^{c} \sum_{j=1}^{N_i^k} \mathbf{y}_{ij}^k$ is the mean of all samples from all modalities; $N_i = \sum_{k=1}^{m} N_i^k$ is the total number of samples of the ith class across all modalities; and N is the total number of samples of all classes across all modalities.

However, directly optimizing $\widehat{\mathcal{L}}_f$ will overemphasize the large distances between already separated classes causing a large overlap and thus a poor discrimination, which was already discussed by some single-modal methods [42,43]. Different from them, we propose a simple yet effective between-class strategy to reduce the large distances between different classes. This is achieved by an increasing differentiable function $g(\cdot)$, which is used to reduce the difference between the class center and the global center. This function is defined as:

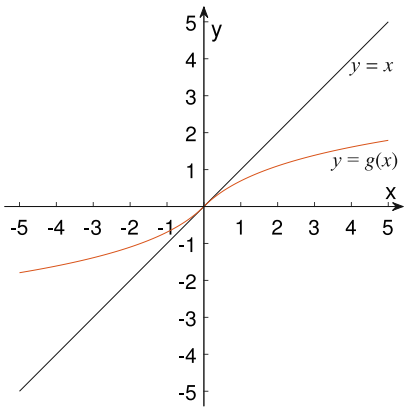


Fig. 2. Graphs of $y = g(x)$ and $y = x$.

$$g(x) = \mathrm{sgn}(x) \ln(|x| + 1), \tag{10}$$

where $\mathrm{sgn}(\cdot)$ is the sign function that extracts the sign of a real number and $\ln(\cdot)$ is the natural logarithm function. $g(x)$ is a smoothly increasing function, where the rate of change is decaying with x leaving apart from the origin comparing to $y = x$, see Fig. 2, where $y = x$ illustrates the figure without using $g(\cdot)$. More specifically, $|g(x)|$ will increase more and more slowly as $|x|$ increases. For the difference $|x|$ between the centers, $g(x)$ tends to emphasize smaller difference over larger one. Therefore, $g(x)$ can be used to reduce the large differences between already separated classes and improve the discriminative ability of the model. Then, the between-class scatter matrix $\widehat{\mathbf{S}}_B$ can be rewritten as:

$$\mathbf{S}_B = \sum_{i=1}^{c} \left(g\left(\boldsymbol{\mu}_i - \boldsymbol{\mu} \right) \right) \left(g\left(\boldsymbol{\mu}_i - \boldsymbol{\mu} \right) \right)^T. \tag{11}$$

Furthermore, let $\mathbf{w}_i$ is the generalized eigenvector of $\mathbf{S}_B$ and $\mathbf{S}_w$ corresponding to the ith largest generalized eigenvalue, *i.e.*,

$$\mathbf{S}_B \mathbf{w}_i = \lambda_i \mathbf{S}_W \mathbf{w}_i. \tag{12}$$

Note that there are at most $c - 1$ nonzero generalized eigenvalues [44]. Therefore, we can rewrite Eq. (7) as follows:

$$\widetilde{\mathcal{L}}_f = -\sum_{i=1}^{c-1} \lambda_i. \tag{13}$$

However, directly solving the above problem would yield trivial solutions *e.g.* maximizing only the largest eigenvalue since this will produce the highest reward of back-propagation. In terms of classification, this means that the discriminant function overemphasizes the discrimination of the dominant eigenvalues but ignores the discriminative information of the minor eigenvalues. Some previous works attempted to solve the problems for single-modal data, such as maximizing the lower bound of the eigenvalues with a threshold [43]. However, the eigenvalue threshold strategy may ignore some discriminative information in the largest eigenvalues. To address this problem, each eigenvalue, which measures the amount of discriminative information at the corresponding dimension, should be considered during the optimization process. Therefore, an increasing differentiable concave function $f(x)|_{x>0}$ is used to decay the dominant eigenvalues and emphasize the minor eigenvalues. Moreover, the nonpositive eigenvalues are considered as nonsignificant components such

as noise. Then only the positive eigenvalues $\lambda_i|_{i=1}^{n}$ are used to train the model, where $n \leqslant c-1$ is the number of all positive eigenvalues. Finally, we can obtain the loss function:

$$\mathcal{L}_f = -\frac{1}{n}\sum_{i=1}^{n} f(\lambda_i). \tag{14}$$

In this paper, the natural logarithm function is selected as $f(\cdot)$ to achieve the goals. With this eigenvalue strategy, the largest eigenvalues can be reduced and the minor eigenvalues can be emphasized in the optimization process. Moreover, this formulation allows to train our model with back-propagation in end-to-end fashion (see Appendix A for a derivative of $\mathcal{L}_f$).

3.4. Optimization

The process of learning the optimal representation is conducted by jointly minimizing $\mathcal{L}_D$, $\mathcal{L}_G$ and $\mathcal{L}_f$, as obtained in Eqs. (5), (6) and (14), respectively. Since the optimization goals of these two objective functions are opposite, the process runs the two concurrent sub-processes:

$$\Theta_D^* = \arg\min_{\Theta_D} \mathcal{L}_D, \tag{15}$$

$$\left(\Theta_{G_1}^*, \ldots, \Theta_{G_m}^*\right) = \arg\min_{\Theta_{G_1},\ldots,\Theta_{G_m}} \left(\mathcal{L}_G + \beta\mathcal{L}_f\right), \tag{16}$$

With the obtained objective function in Eq. (1), the proposed MAN is iteratively optimized in an adversarial manner. Like all adversarial learning methods, the real and fake modality labels are different as shown in Eqs. (2) and (4). The parameters of generators are fixed during the discriminator training stage and vice versa. This can eliminate the discrepancy of the generated representations by the m modality-specific generators. Therefore, the overall network can be optimized using a stochastic gradient descent optimization algorithms, like Adam [45]. The detailed optimization process is shown in Algorithm 1.

Algorithm 1 Optimization procedure of MAN

Input: The training data $\mathcal{X}^k|_{k=1}^m$, the dimensionality of the generated representations d, batch size N_b, positive balance parameter β, learning rate α

1: **while** not converge **do**
2: Randomly select N_b samples for each modality from $\mathcal{X}^k|_{k=1}^m$ to construct a multimodal mini-batch.
3: Compute the generated representations by the corresponding modality-specific generators for all modalities of the mini-batch.
4: Compute the between- and within-class scatter matrices $\mathbf{S}_W$ and $\mathbf{S}_B$ using Eqs. (8) and (11).
5: Solve the generalized eigenvalue decomposition (GED) problem defined in Eq. (12) and obtain n positive eigenvalues.
6: Update the parameters of the discriminator D by minimizing $\mathcal{L}_D$ in Eq. (5) with descending their stochastic gradient:
$\Theta_D = \Theta_D - \alpha\frac{\partial\mathcal{L}_D}{\partial\Theta_D}$
7: Update the parameters of the generators $G_k|_{k=1}^m$ by minimizing $\mathcal{L}_G + \beta\mathcal{L}_f$ in Eqs. (6) and (14) with descending their stochastic gradient:
$\Theta_{G_k} = \Theta_{G_k} - \alpha\left(\frac{\partial\mathcal{L}_G}{\partial\Theta_{G_k}} + \beta\frac{\partial\mathcal{L}_f}{\partial\Theta_{G_k}}\right)\ (k = 1, \cdots, m)$
8: **end while**

Output: Optimized MAN model.

4. Experiments

To evaluate the proposed methods, we conduct experiments on four datasets, namely, the Reuters dataset [2,46], the PKU XMedia dataset [5,47], the Wikipedia dataset [3], and the Pascal Sentence dataset [4]. In the following experiments, we first compare our MAN with 16 state-of-the-art cross-modal methods to verify its effectiveness. Then the additional evaluations are conducted to investigate the performance of MAN in more detail.

4.1. Datasets

Here we briefly introduce 4 multimodal datasets adopted in the experiments, including Reuters, PKU XMedia, Wikipedia and Pascal Sentence datasets.

4.1.1. Reuters dataset[5] [2,46]

This dataset consists of documents that are written in five different languages and their translations. We use the subset that is written in English (EN) and all their translations in all the other 4 languages, *i.e.* French (FR), German (GE), Spanish (SP) and Italian (IT) following [46]. All the documents are categorized into 6 classes, and there are 18,758 instances in this subset. Moreover, the dataset is randomly split into three subsets: the training set with 10,000 instances, the validation set with 4000 instances and the testing set with 4758 instances. To make the data be more tractable on our computer, principal component analysis (PCA) [48] is first applied to map each high dimensional sample (higher than 10,000 dimensions) into a 1024-dimensional feature vector.

4.1.2. PKU Xmedia dataset[6] [5,47]

This dataset consists of 5000 texts, 5000 images, 1143 videos, 1000 audio clips and 500 3D models. The dataset is evenly split into 20 categories, which are insect, bird, wind, dog and so on. The PKU XMedia dataset was split to 2 part by the authors: the training set has 10,169 instances (with 4000 texts, 4000 images, 969 videos, 800 audio clips and 400 3D models), and the testing set has 2474 instances (with 1000 texts, 1000 images, 174 videos, 200 audio clips and 100 3D models). For pairwise multimodal methods, we construct a pairwise training set by selecting 400 instances from each modality according to the labels of the 400 3D models.

4.1.3. Wikipedia dataset[7] [3]

It is the most widely-used dataset for cross-modal retrieval. The dataset consists of 2866 image-text pairs where each pair consists of an image and the corresponding complete text article annotated with a label from 10 semantic classes (*i.e.* art, biology, history, etc.). For a fair comparison, we also exactly follow the data partition strategy of [49] to divide the dataset into 3 subsets: 2173 pairs in training set, 231 pairs in validation set and 462 pairs in the testing set.

4.1.4. Pascal sentence dataset[8] [4]

This dataset contains 1000 images, which is generated from the 2008 PASCAL development kit. Each image is annotated via Amazon Mechanical Turk by crowdsourcing to generate 5 independent sentences from different annotators, which forms on a document. This dataset is categorized into 20 categories, and following [49], 800 for training (40 cases per category), 100 for validation (5 cases per category) and 100 for testing (5 cases per category).

[5] This dataset is available at https://github.com/yeqinglee/mvdata.
[6] This dataset is available at http://www.icst.pku.edu.cn/mipl/XMediaNet/.
[7] This dataset is available at http://www.svcl.ucsd.edu/projects/crossmodal/.
[8] This dataset is available at http://vision.cs.uiuc.edu/pascal-sentences/.

4.2. Compared methods

To verify the effectiveness of our proposed methods, we compare 16 state-of-the-art methods in the experiments, including 6 traditional multimodal methods, namely MCCA [7], PLS [7], GMLDA [9], MvDA [10], MvDA-VC [11] and GSS-SL [19], as well as 9 DNN-based methods, namely DCCA [13], DCCAE [14], CMDN [36], MvDN [15], ACMR [50], CM-GANs [16], MCSM [17], CCL [38], CBT [39]. We briefly introduce these compared methods in the following paragraphs.

- MCCA [7] learns multiple modality-specific linear transformations to map the different modalities into a common space by maximizing the correlations between all possible pairwise modalities.
- PLS [7] maximizes the covariance of the two modalities to learn a latent common space.
- GMLDA [9] uses the discriminative information of each individual modality and the pairwise relationship of any two modalities to project the samples from different modalities into a common space.
- MvDA [10] uses Fisher's criterion on all modalities to obtain multiple modality-specific linear transforms to project the multimodal data into a common space.
- MvDA-VC [11] introduces a constraint into MvDA to enforce the modality consistency of the multiple linear transforms based on the observations from different modalities shared similar data.
- GSS-SL [19] takes the label space as a linkage to model the correlations among different modalities so that a discriminative common subspace is learned.
- DCCA [13] adopts a similar objective function with CCA on the top of two separate subnetworks to maximize the correlation between the two modalities.
- DCCAE [14] learns the common representations of two modalities by adding an autoencoder regularization terms to DCCA.
- CMDN [36] jointly models the intra- and inter-modality correlation in both separate representation and common representation learning stages with multiple deep networks.
- MvDN [15] introduces Fisher's loss into a feedforward neural network to learn modality-invariant representations of the multiple modalities.
- ACMR [24] seeks an effective common subspace based on adversarial learning, which consists of a feature projector, a modality classifier and a triplet constraint.
- MCSM [17] adopts an end-to-end framework to directly generate modality-specific cross-modal similarity without explicit common representation.
- CM-GANs [16] learns discriminative common representation for bridging the heterogeneity gap by utilizing the power of GAN to model the cross-modal joint distribution. In other words, CM-GANs aim to effectively correlate existing large-scale heterogeneous data of different modalities.
- GXN [25] incorporates generative processes into the cross-modal feature embedding to learn not only the global abstract features but also the local grounded features.
- CCL [38] fully explores both intra- and inter-modality correlation simultaneously with multi-grained and multi-task learning.
- CBT [39] treats images as a special kind of language to provide visual descriptions, so that translation can be conduct between bilingual pair of image and text to effectively explore cross-modal correlation.

4.3. Implementation detail

For all traditional methods, the results are reported on the testing subsets when the corresponding methods get the best performance on the validation for all datasets except for the PKU XMedia dataset. Because PKU XMedia does not have the validation subset, we report the best results on the testing subset for all methods. The parameters of all compared methods are provided by the authors in all experiments. Furthermore, the same features for all modalities are adopted in all compared approaches and our MAN for a fair comparison. Specifically, an image is represented by a 4096-dimensional feature vector, which is extracted from the fc7 layer in 19-layer VGGNet [51] pre-trained on the ImageNet. Similarly, each word is represented as a 300-dimensional feature vector extracted by Word2Vec model [52] which is pre-trained on billion of words in Google News.[9] Therefore, each document can be represented by a $p \times 300$ feature matrix, where p is the maximum word number of the documents, and zero-padding is adopted for other documents beneath this limit. However, because the documents have so many words for the Wikipedia dataset, the text feature dimension is too high to be handled by other methods (such as MCCA, PLS, GMLDA, MvDA, MvDA-VC, GSS-SL, DCCA, DCCAE, ACMR and CMDN) on our computer, *e.g.*, the text feature dimension is 869,100 in the Wikipedia dataset. Therefore, we calculate the 300-dimensional mean vector of the $p \times 300$ feature matrix, following [19]. Furthermore, for the Pascal Sentence dataset, the text feature matrix is reshaped as a $1 \times 300p$ vector as the input of the other methods since the dimension of text features is computable on our computer. For our MAN, the text CNN architecture is applied for the $p \times 300$ text feature matrix with the same configuration of [53]. Moreover, the output dimension of the generators is the number of classes for each dataset. All the layers use ReLU activations except for the final layer which uses a linear activation.

4.4. Evaluation metric

Cross-modal retrieval is performed in the learned common space on the above datasets with $m(m-1)$ kinds of cross-modal retrieval tasks. The testing is conducted in a pairwise manner, *i.e.*, the samples from one modality are used as database while the ones from another modality are used as queries, which is defined as follows.

- $\mathcal{X}^k$-**query**-$\mathcal{X}^l$ ($\mathcal{X}^k \rightarrow \mathcal{X}^l$, $k \neq l$): for a query from the kth modality, relevant instances from the lth modality are retrieved in the testing set ranked by the calculated cross-modal similarity in the common space.

The similarity is computed by cosine distance in the common space. Taking two-modality case (image and text) for an example, there are two kinds of retrieval tasks, *i.e.*, image-query-text (Image $\rightarrow$ Text) and text-query-image (Text $\rightarrow$ Image). To further evaluate the performance of multimodal methods for more than two modalities, the all-modal retrieval is performed on the Reuters and PKU XMedia datasets following [16], which is defined as follows.

[9] The model is available at https://code.google.com/archive/p/word2vec/.

- $\mathcal{X}^k$-**query-All** ($\mathcal{X}^k \rightarrow$ all): for a query from the kth modality, relevant instances from all modalities are retrieved in the testing set ranked by the calculated similarity in the common space.

Mean Average Precision (mAP) score is adopted as the evaluation metric on the four datasets. mAP is the mean value of Average Precision (AP) scores for each query. Moreover, AP is defined as follows:

$$\mathrm{AP} = \frac{1}{R}\sum_{k=1}^{n}\frac{R_k}{k} \times \mathrm{rel}(k) \tag{17}$$

where the testing set contains R relevant instances and n instances. R_k is the number of relevant instances in the top k returned results. rel(k) is an indicator function which is equal to 1 if the returned result of rank k is a relevant instance, zero otherwise. mAP score considers the ranking of returned retrieval results as well as precision simultaneously, which is extensively adopted in cross-modal retrieval tasks, such as [17]. It should be noted that the mAP score is calculated on the top 50 and all returned results following [19] in our experiments.

4.5. Comparisons with state-of-the-art methods

In this subsection, we evaluate the effectiveness of the proposed MAN by comparing with 16 state-of-the-art methods on 4 widely-used multimodal datasets. Because most compared methods can handle only two modalities, there are just five multimodal approaches (MCCA, GMLDA, MvDA, MvDA-VC and MvDN) that can be used to compare with our MAN on the Reuters and PKU XMedia datasets. Each of the two datasets has five modalities. In Table 1, the mAP scores of 20 retrieval tasks and their average results are used to evaluate the performance of all methods for cross-modal retrieval across different language. As shown in Table 1, our method achieves the best results in cross-modal retrieval tasks on the Reuters dataset. From Table 1, we can see that all the DNN-based methods (MvDN and our MAN) outperform the traditional multimodal methods. This demonstrates the effectiveness of DNN to capture the nonlinear semantic information from large scale data. Because more discrimination can be extracted from the multimodal data using class labels, all supervised methods achieve better results than the unsupervised MCCA. Although our MAN cannot outperform GMLDA in some tasks, MAN achieves the best performance compared to its counterparts in the most tasks and the average results. As shown in Table 1, our proposed MAN has improved the average mAP score from 0.777 to 0.819 in R = ALL on the Reuters dataset. Like the Reuters dataset, the mAP scores of 20 retrieval tasks and their average results are computed to evaluate the effectiveness of our MAN on the PKU XMedia dataset. However, not all the modalities have the same number of samples on the PKU XMedia dataset and some compared multimodal methods need any two modalities to be pairwise. Therefore, we compare these pairwise multimodal methods on our constructed pairwise training set of the PKU XMedia dataset. As shown in Table 2, our proposed approach has improved the average mAP score from 0.350 to 0.444 in R = ALL on the pairwise training set. From Table 2, we can see that all supervised approaches outperform the unsupervised MCCA as the results of Table 1. However, because the pairwise training set is small-scale (only 400 samples for each modality from 20 categories), the MvDN cannot be trained as a well-performing deep model and its performance is not as convincing as for the Reuters dataset. Unlike MvDN, our proposed MAN can also achieve satisfactory results on the small-scale training set. We believe that this is due to the proposed between-class and eigenvalue strategies can preserve as much discrimination as possible into all the dimensions of the common representations. Since more available samples can participate in training the models on the full training set, MvDN and our MAN have achieved better performance than that on the pairwise training set. However, the imbalance modalities make the performance of MvDA worse. This shows that deep learning has more advantages over the traditional methods in the multimodal case. Finally, our proposed MAN outperforms the best competitor, MvDA-VC, by 53.14% on the average mAP scores. To further evaluate the performance of multimodal methods for more than two modalities, all-modal retrieval is performed on the Reuters and PKU XMedia datasets as shown in Tables 3 and 4, respectively. The all-modal retrieval evaluations are consistent with the mAP scores for cross-modal retrieval tasks, where our MAN outperforms all the compared methods. The results of Tables 1–4 indicate that our MAN is an effective multimodal representation learning method for cross-modal retrieval on more than two modalities.

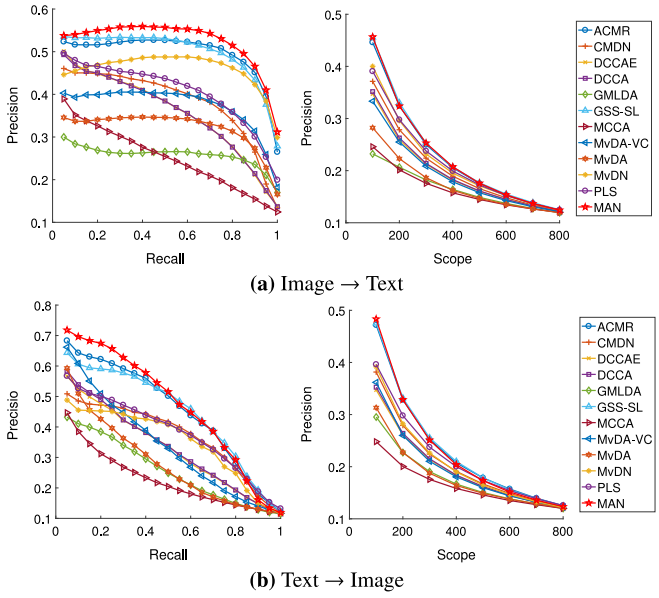
Cross-modal retrieval across image and text is evaluated on the Wikipedia and Pascal Sentence datasets. The results are shown in Tables 5 and 6, respectively. As shown in these tables, our proposed MAN shows advantage over 16 state-of-the-art methods on the two datasets. In Table 5, the MAN achieves improvements of 1.34% for image-query-texts, 3.00% for text-query-images, and 2.02% for average compared with the best results of counterparts (*i.e.*, CM-GANs) in R = ALL. Moreover, most DNN-based methods outperform the traditional ones for the strong power of nonlinear correlation modeling. For some supervised methods, such as GSS-SL and ACMR, their performance can be boosted by the discriminative information of the class labels. However, the discriminative information cannot be fully utilized to preserve the discrimination into the common representations by some supervised approaches, such as GMLDA and MvDA. The similar trends can be seen on the Pascal Sentence dataset. As shown in Table 6, the performance of the compared DNN-based methods is limited on the Pascal Sentence dataset because that dataset is small-scale (only 800 image-text pairs from 20 categories). On the contrary, some traditional methods, such as GSS-SL and MvDA-VC, show more satisfactory performance on the small-scale dataset. As expected, our MAN achieves the best performance compared with all the counterparts. Especially, the MAN outperforms the best counterpart (*i.e.*, GSS-SL) by 7.93% for image-query-texts, 10.94% for text-query-images, and 9.35% for average in R = ALL. The results of Tables 5 and 6 indicate that our MAN is an effective multimodal representation learning method for cross-modal retrieval across image and text.

In addition to the evaluation in terms of the mAP score, the precision–recall curves and precision–scope curves are drawn for additional comparison. On the Wikipedia dataset, the precision–recall and precision–scope curves of the image-query-texts and text-query-images are plotted in Figs. 3(a) and 3(b), respectively. Similarly, the precision–recall and precision–scope curves of the image-query-texts and text-query-images on the Pascal Sentence dataset are respectively displayed in Figs. 4(a) and 4(b). The scope (*i.e.*, the top K retrieved samples) of the precision–scope varies from K = 100 to 800 as [19]. Figs. 3 and 4 show the curves of our MAN and 12 state-of-the-art cross-modal methods. The precision–recall and precision–scope evaluations are consistent with the mAP scores for cross-modal retrieval tasks, where our MAN outperforms all the compared methods.

From all the experimental results, we can see that our MAN and other cross-modal GAN-based methods, such as ACMR and CM-GANs, can extract modality-invariant representations with adversarial learning. Thus, we can conclude that adversarial learning has some advantages in multimodal representation learning. Furthermore, it can be seen that the performance of MvDN is unsatisfactory from Tables 1 to 6, because the discriminant

Table 1
Performance comparison in terms of mAP@R scores on the Reuters dataset. LA, EN, FR, GE, IT and SP are denoted as language, English, French, German, Italian and Spanish, respectively.

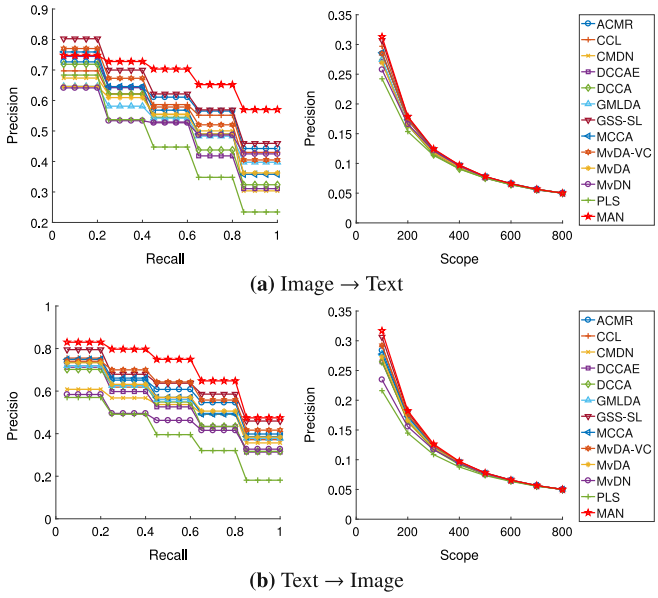
Method	Query	$R = 50$						$R = ALL$					
		EN	FR	GE	IT	SP	Avg.	EN	FR	GE	IT	SP	Avg.
MCCA [7]	EN	–	0.779	0.776	0.779	0.775	0.777	–	0.409	0.408	0.409	0.407	0.408
	FR	0.776	–	0.771	0.775	0.771	0.773	0.410	–	0.405	0.407	0.404	0.406
	GE	0.774	0.771	–	0.772	0.769	0.771	0.408	0.405	–	0.405	0.403	0.405
	IT	0.775	0.773	0.769	–	0.770	0.772	0.409	0.406	0.405	–	0.404	0.406
	SP	0.772	0.770	0.768	0.771	–	0.770	0.407	0.404	0.402	0.404	–	0.404
	Avg.	0.774	0.773	0.771	0.774	0.772	0.773	0.409	0.406	0.405	0.406	0.404	0.406
GMLDA [9]	EN	–	0.852	0.848	0.851	0.849	0.850	–	0.765	0.756	0.761	0.756	0.760
	FR	**0.848**	–	0.840	0.843	0.842	0.844	0.766	–	0.747	0.752	0.747	0.753
	GE	**0.843**	0.839	–	0.837	0.835	0.838	0.759	0.749	–	0.744	0.740	0.748
	IT	**0.842**	0.839	0.835	–	0.837	0.838	0.761	0.751	0.742	–	0.742	0.749
	SP	**0.839**	**0.839**	0.833	**0.838**	–	0.837	0.758	0.748	0.739	0.743	–	0.747
	Avg.	**0.843**	0.842	0.839	0.842	0.841	0.841	0.761	0.753	0.746	0.750	0.746	0.751
MvDA [10]	EN	–	0.816	0.812	0.811	0.809	0.812	–	0.654	0.621	0.652	0.611	0.634
	FR	0.808	–	0.798	0.807	0.802	0.804	0.655	–	0.627	0.638	0.628	0.637
	GE	0.789	0.791	–	0.791	0.777	0.787	0.616	0.623	–	0.615	0.597	0.613
	IT	0.801	0.802	0.801	–	0.796	0.800	0.650	0.632	0.615	–	0.615	0.628
	SP	0.783	0.792	0.785	0.785	–	0.786	0.608	0.621	0.599	0.613	–	0.610
	Avg.	0.795	0.800	0.799	0.799	0.796	0.798	0.632	0.632	0.616	0.629	0.613	0.624
MvDA-VC [11]	EN	–	0.821	0.784	0.820	0.821	0.812	–	0.689	0.672	0.660	0.662	0.671
	FR	0.805	–	0.794	0.814	0.811	0.806	0.686	–	0.625	0.657	0.652	0.655
	GE	0.795	0.779	–	0.792	0.803	0.792	0.671	0.621	–	0.626	0.588	0.626
	IT	0.770	0.790	0.810	–	0.803	0.793	0.653	0.656	0.631	–	0.646	0.646
	SP	0.794	0.791	0.760	0.794	–	0.785	0.655	0.649	0.580	0.642	–	0.632
	Avg.	0.791	0.795	0.787	0.805	0.810	0.798	0.666	0.654	0.627	0.646	0.637	0.646
MvDN [15]	EN	–	0.828	0.830	0.829	0.831	0.829	–	0.788	0.787	0.787	0.787	0.787
	FR	0.821	–	0.818	0.817	0.821	0.819	0.789	–	0.775	0.775	0.775	0.779
	GE	0.819	0.812	–	0.810	0.815	0.814	0.787	0.774	–	0.772	0.772	0.776
	IT	0.819	0.813	0.814	–	0.814	0.815	0.788	0.774	0.773	–	0.772	0.777
	SP	0.811	0.803	0.803	0.801	–	0.804	0.780	0.766	0.765	0.764	–	0.769
	Avg.	0.817	0.814	0.816	0.814	0.820	0.816	0.786	0.776	0.775	0.774	0.776	0.777
MAN	EN	–	**0.853**	**0.853**	**0.852**	**0.854**	**0.853**	–	**0.829**	**0.822**	**0.825**	**0.825**	**0.825**
	FR	0.845	–	**0.846**	**0.845**	**0.849**	**0.846**	**0.828**	–	**0.817**	**0.820**	**0.820**	**0.821**
	GE	0.839	**0.840**	–	**0.838**	**0.841**	**0.839**	**0.822**	**0.818**	–	**0.815**	**0.814**	**0.817**
	IT	0.838	**0.841**	**0.841**	–	**0.841**	**0.840**	**0.824**	**0.819**	**0.813**	–	**0.815**	**0.818**
	SP	0.837	**0.839**	**0.838**	0.837	–	**0.838**	**0.822**	**0.817**	**0.811**	**0.814**	–	**0.816**
	Avg.	0.840	**0.843**	**0.844**	**0.843**	**0.847**	**0.843**	**0.824**	**0.821**	**0.816**	**0.818**	**0.818**	**0.819**



(a) Image → Text

(b) Text → Image

Fig. 3. Precision–recall curves and precision–scope curves for the image-query-texts and text-query-images experiments on the Wikipedia dataset.



(a) Image → Text

(b) Text → Image

Fig. 4. Precision–recall curves and precision–scope curves for the image-query-texts and text-query-images experiments on the Pascal Sentence dataset.

function overemphasizes the largest between-class distances and the dominant eigenvalues. With the proposed between-class and eigenvalue strategies, our MAN can effectively avoid overemphasizing the largest distances between already separated classes and

Table 2
Performance comparison in terms of mAP@R scores on the PKU XMedia dataset.

Training set	Method	Query	$R = 50$						$R = ALL$					
			Audio	3D	Text	Image	Video	Avg.	Audio	3D	Text	Image	Video	Avg.
Pairwise	MCCA [7]	Audio	–	0.181	0.248	0.250	0.147	0.206	–	0.153	0.125	0.140	0.113	0.133
		3D	0.208	–	0.207	0.215	0.156	0.196	0.149	–	0.099	0.129	0.119	0.124
		Text	0.200	0.173	–	0.222	0.143	0.185	0.149	0.145	–	0.129	0.109	0.133
		Image	0.221	0.194	0.239	–	0.161	0.204	0.165	0.167	0.124	–	0.126	0.145
		Video	0.155	0.158	0.145	0.149	–	0.151	0.110	0.134	0.080	0.098	–	0.105
		Avg.	0.196	0.177	0.210	0.209	0.152	0.189	0.143	0.150	0.107	0.124	0.117	0.128
	GMLDA [9]	Audio	–	0.260	0.267	0.286	0.203	0.254	–	0.246	0.209	0.268	0.169	0.223
		3D	0.281	–	0.486	0.532	0.329	0.407	0.233	–	0.408	0.474	0.290	0.351
		Text	0.299	0.489	–	0.662	0.338	0.447	0.246	0.451	–	0.616	0.300	0.403
		Image	0.356	0.525	0.702	–	0.415	0.499	0.295	0.494	0.588	–	0.369	0.436
		Video	0.220	0.337	0.340	0.399	–	0.324	0.179	0.316	0.288	0.373	–	0.289
		Avg.	0.289	0.403	0.449	0.470	0.321	0.386	0.238	0.377	0.373	0.433	0.282	0.341
	MvDA [10]	Audio	–	0.283	0.289	0.316	0.212	0.275	–	0.267	0.235	0.285	0.169	0.239
		3D	0.260	–	0.444	0.479	0.337	0.380	0.218	–	0.368	0.438	0.282	0.327
		Text	0.306	0.470	–	0.665	0.360	0.450	0.245	0.419	–	0.604	0.297	0.391
		Image	0.356	0.527	0.719	–	0.434	0.509	0.288	0.468	0.596	–	0.367	0.430
		Video	0.199	0.313	0.305	0.325	–	0.285	0.163	0.272	0.247	0.313	–	0.249
		Avg.	0.280	0.398	0.439	0.446	0.336	0.380	0.229	0.356	0.361	0.410	0.279	0.327
	MvDA-VC [11]	Audio	–	0.251	0.248	0.264	0.196	0.240	–	0.236	0.207	0.238	0.171	0.213
		3D	0.230	–	**0.487**	**0.547**	**0.365**	0.407	0.199	–	**0.413**	**0.501**	**0.319**	0.358
		Text	0.266	0.496	–	0.682	0.414	0.465	0.218	0.458	–	0.631	0.366	0.418
		Image	0.306	0.549	0.740	–	0.491	0.521	0.249	0.508	0.622	–	0.437	0.454
		Video	0.216	0.339	0.396	0.453	–	0.351	0.170	0.312	0.330	0.414	–	0.307
		Avg.	0.255	0.408	0.468	0.486	0.367	0.397	0.209	0.378	0.393	0.446	0.323	0.350
	MvDN [15]	Audio	–	0.172	0.363	0.339	0.258	0.283	–	0.143	0.316	0.338	0.218	0.254
		3D	0.075	–	0.095	0.138	0.124	0.108	0.070	–	0.094	0.075	0.091	0.083
		Text	0.256	0.130	–	0.403	0.259	0.262	0.215	0.121	–	0.367	0.235	0.235
		Image	0.410	0.241	0.735	–	0.499	0.471	0.346	0.209	0.623	–	0.436	0.404
		Video	0.254	0.198	0.393	0.461	–	0.327	0.219	0.175	0.334	0.427	–	0.289
		Avg.	0.249	0.185	0.397	0.335	0.285	0.290	0.212	0.162	0.342	0.302	0.245	0.253
	MAN	Audio	–	**0.398**	**0.513**	**0.525**	**0.344**	**0.445**	–	**0.364**	**0.480**	**0.506**	**0.301**	**0.413**
		3D	**0.378**	–	0.457	0.464	0.349	**0.412**	**0.301**	–	0.412	0.452	0.309	**0.369**
		Text	**0.556**	**0.517**	–	**0.788**	**0.472**	**0.583**	**0.461**	**0.471**	–	**0.729**	**0.414**	**0.519**
		Image	**0.583**	**0.562**	**0.820**	–	**0.517**	**0.621**	**0.502**	**0.521**	**0.765**	–	**0.459**	**0.562**
		Video	**0.383**	**0.406**	**0.455**	**0.464**	–	**0.427**	**0.322**	**0.380**	**0.446**	**0.451**	–	**0.400**
		Avg.	**0.475**	**0.471**	**0.561**	**0.560**	**0.421**	**0.498**	**0.396**	**0.434**	**0.526**	**0.535**	**0.371**	**0.452**
Full	MvDA [10]	Audio	–	0.102	0.113	0.070	0.107	0.098	–	0.097	0.089	0.065	0.081	0.083
		3D	0.088	–	0.261	0.265	0.159	0.193	0.073	–	0.205	0.189	0.117	0.146
		Text	0.097	0.256	–	0.444	0.166	0.241	0.081	0.226	–	0.303	0.117	0.182
		Image	0.082	0.233	0.445	–	0.181	0.235	0.070	0.203	0.305	–	0.125	0.176
		Video	0.089	0.148	0.116	0.132	–	0.121	0.085	0.143	0.105	0.100	–	0.108
		Avg.	0.089	0.185	0.234	0.228	0.153	0.178	0.078	0.167	0.176	0.164	0.110	0.139
	MvDN [15]	Audio	–	0.149	0.342	0.271	0.242	0.251	–	0.141	0.318	0.294	0.196	0.237
		3D	0.083	–	0.100	0.092	0.106	0.095	0.068	–	0.086	0.110	0.094	0.090
		Text	0.230	0.171	–	0.417	0.272	0.272	0.194	0.165	–	0.382	0.229	0.242
		Image	0.456	0.346	0.766	–	0.548	0.529	0.394	0.315	0.710	–	0.509	0.482
		Video	0.296	0.244	0.394	0.466	–	0.350	0.252	0.229	0.385	0.452	–	0.330
		Avg.	0.266	0.227	0.400	0.312	0.292	0.300	0.227	0.212	0.375	0.310	0.257	0.276
	MAN	Audio	–	**0.443**	**0.580**	**0.559**	**0.401**	**0.496**	–	**0.382**	**0.600**	**0.571**	**0.351**	**0.476**
		3D	**0.405**	–	**0.546**	**0.518**	**0.403**	**0.468**	**0.334**	–	**0.547**	**0.506**	**0.353**	**0.435**
		Text	**0.677**	**0.642**	–	**0.938**	**0.637**	**0.723**	**0.599**	**0.546**	–	**0.901**	**0.559**	**0.651**
		Image	**0.644**	**0.612**	**0.892**	–	**0.622**	**0.692**	**0.577**	**0.525**	**0.891**	–	**0.550**	**0.636**
		Video	**0.421**	**0.437**	**0.527**	**0.521**	–	**0.477**	**0.364**	**0.392**	**0.553**	**0.530**	–	**0.460**
		Avg.	**0.537**	**0.533**	**0.636**	**0.634**	**0.516**	**0.571**	**0.469**	**0.461**	**0.648**	**0.627**	**0.453**	**0.531**

Table 3
Performance comparison in terms of mAP@ALL scores for all-modal retrieval on the Reuters dataset.

Method	English → All	French → All	German → All	Italian → All	Spanish → All	Avg.
MCCA [7]	0.411	0.408	0.407	0.408	0.406	0.408
GMLDA [9]	0.762	0.753	0.746	0.748	0.745	0.751
MvDA [10]	0.641	0.642	0.619	0.628	0.616	0.629
MvDA-VC [11]	0.675	0.660	0.618	0.649	0.618	0.644
MvDN [15]	0.790	0.778	0.775	0.775	0.767	0.777
MAN	**0.834**	**0.830**	**0.821**	**0.822**	**0.825**	**0.826**

Table 4
Performance comparison in terms of mAP@ALL scores for all-modal retrieval on the PKU XMedia dataset.

Training set	Method	Audio → All	3D → All	Text → All	Image → All	Video → All	Avg.
Pairwise	MCCA [7]	0.137	0.113	0.123	0.155	0.098	0.125
	GMLDA [9]	0.201	0.391	0.499	0.586	0.305	0.396
	MvDA [10]	0.233	0.375	0.496	0.602	0.250	0.391
	MvDA-VC [11]	0.180	**0.426**	0.531	0.634	0.350	0.424
	MvDN [15]	0.243	0.079	0.248	0.493	0.296	0.272
	MAN	**0.443**	0.263	**0.655**	**0.699**	**0.413**	**0.495**
Full	MvDA [10]	0.268	0.288	0.091	0.180	0.113	0.188
	MvDN [15]	0.219	0.080	0.311	0.579	0.328	0.304
	MAN	**0.534**	**0.404**	**0.857**	**0.816**	**0.500**	**0.622**

Table 5
Performance comparison in terms of mAP@R scores on the Wikipedia dataset.

Method	Tasks					
	$R = 50$			$R = ALL$		
	Image → Text	Text → Image	Avg.	Image → Text	Text → Image	Avg.
MCCA [7]	0.346	0.380	0.363	0.256	0.234	0.245
PLS [8]	0.465	0.512	0.488	0.412	0.387	0.400
GMLDA [9]	0.280	0.410	0.345	0.262	0.268	0.265
MvDA [10]	0.327	0.472	0.400	0.325	0.289	0.307
MvDA-VC [11]	0.387	0.541	0.464	0.378	0.347	0.363
GSS-SL [19]	0.511	0.591	0.551	0.503	0.457	0.480
DCCA [13]	0.451	0.500	0.476	0.365	0.341	0.353
DCCAE [14]	0.453	0.491	0.472	0.364	0.336	0.350
CMDN [36]	0.441	0.477	0.459	0.389	0.375	0.382
MvDN [15]	0.439	0.439	0.455	0.462	0.356	0.409
ACMR [24]	0.503	0.617	0.560	0.502	0.462	0.482
MCSM [17]	0.496	0.590	0.543	0.516	0.458	0.487
GXN [25]	0.406	0.384	0.395	0.319	0.275	0.297
CCL [38][a]	0.490	0.613	0.551	0.504	0.457	0.481
CM-GANs [16][a]	0.500	0.621	0.561	0.521	0.466	0.494
CBT [39][a]	–	–	–	0.516	0.464	0.490
MAN	**0.517**	**0.655**	**0.586**	**0.528**	**0.480**	**0.504**

[a]The results are reported by the authors.

Table 6
Performance comparison in terms of mAP@R scores on the Pascal Sentence dataset.

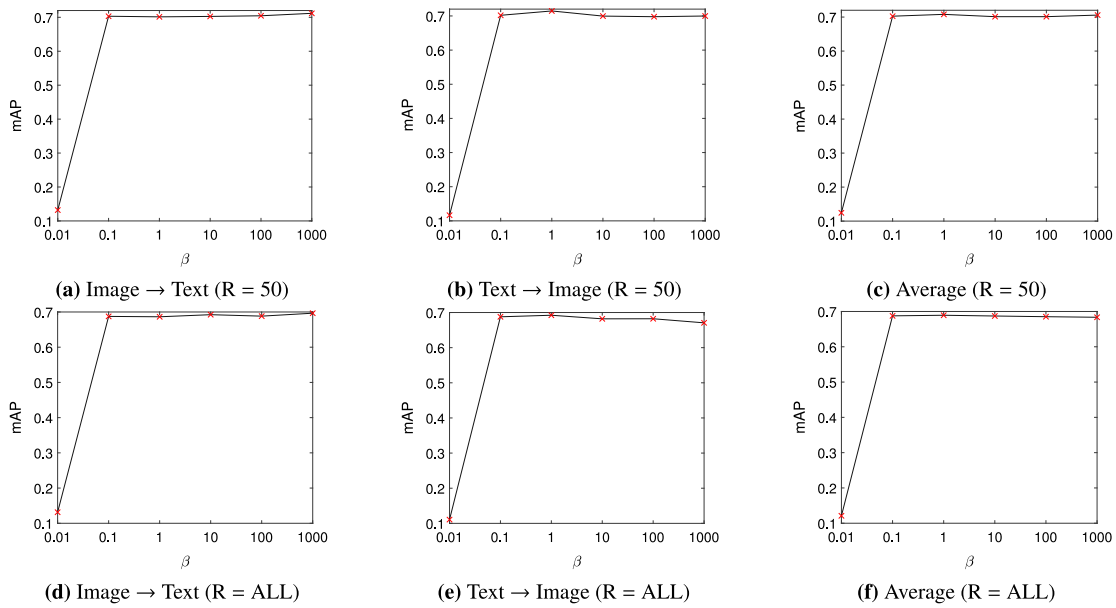
Method	Tasks					
	$R = 50$			$R = ALL$		
	Image → Text	Text → Image	Avg.	Image → Text	Text → Image	Avg.
MCCA [7]	0.572	0.586	0.579	0.564	0.572	0.568
PLS [8]	0.473	0.419	0.446	0.450	0.392	0.421
GMLDA [9]	0.542	0.567	0.554	0.530	0.556	0.543
MvDA [10]	0.546	0.576	0.561	0.532	0.565	0.549
MvDA-VC [11]	0.599	0.620	0.609	0.589	0.611	0.600
GSS-SL [19]	0.634	0.643	0.639	0.630	0.631	0.631
DCCA [13]	0.553	0.545	0.549	0.529	0.526	0.528
DCCAE [14]	0.552	0.534	0.543	0.529	0.516	0.523
CMDN [36]	0.544	0.524	0.534	0.544	0.526	0.535
MvDN [15]	0.524	0.462	0.493	0.523	0.457	0.490
ACMR [24]	0.604	0.598	0.601	0.598	0.588	0.593
MCSM [17]	0.657	0.659	0.658	0.640	0.621	0.630
GXN [25]	0.566	0.552	0.559	0.550	0.542	0.546
CCL [38]	0.591	0.577	0.584	0.577	0.561	0.569
CM-GANs [16][a]	0.612	0.610	0.611	0.603	0.604	0.604
CBT [39][a]	–	–	–	0.602	0.583	0.592
MAN	**0.683**	**0.709**	**0.696**	**0.680**	**0.700**	**0.690**

[a]The results are reported by the authors.

the dominant eigenvalues. Therefore, more discrimination can be preserved in the generated common representations.

4.6. Parameter analysis

To investigate the impact of the parameter β, we analyze the performance of our MAN with different values of β on the validation subset of the Pascal Sentence dataset as shown in Fig. 5. These figures show the results of two tasks and their average. They show the mAP scores of MAN versus different values of β for image-query-texts, text-query-images and the average results. From the figures, we can see that MDA loss is very important to preserve discrimination into common representations. When β is small, such as 0.01, no sufficient discrimination can be extracted from the multimodal data. With MDA playing a more important role in the proposed model, sufficient discrimination can be pushed into the common space. Furthermore, our MAN is insensitive to β when β is sufficiently large. Therefore, β is set as 1 for all experiments.



(a) Image → Text (R = 50) **(b)** Text → Image (R = 50) **(c)** Average (R = 50)

(d) Image → Text (R = ALL) **(e)** Text → Image (R = ALL) **(f)** Average (R = ALL)

Fig. 5. Cross-modal retrieval performance of MAN in terms of mAP@R with different values of β on the validation subset of the Pascal Sentence dataset.

4.7. Effect of adversarial learning

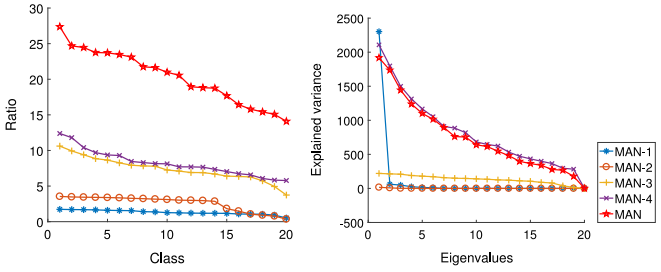
In order to investigate the effect of adversarial learning, we compare MAN with its a variation: MAN (with $\mathcal{L}_f$ only), which only optimizes the generators with $\mathcal{L}_f$ and without adversarial learning. This comparison is conducted on the Wikipedia and Pascal Sentence datasets as shown in Table 7. From the experimental results, we can see that MAN outperforms the MAN ($\mathcal{L}_f$ only) on the two datasets. Thanks to adversarial learning, the generators can eliminate the modality discrepancy in the common space and generate modality-invariant representations. Therefore, adversarial learning is an effective technology to eliminate the discrepancy between different modalities.

4.8. Influence analysis of different strategies

We investigate the contributions of different strategies in our proposed model. We define the following five alternative baselines to study the importance of different proposed strategies in our common space learning model:

1. MAN-1: without the between-class and eigenvalue strategies.
2. MAN-2: with the between-class strategy only.
3. MAN-3: with eigenvalue threshold strategy [43].
4. MAN-4: with the eigenvalue strategy only.

For a fair comparison, all these baselines and our proposed MAN have the same network architecture and settings. The differences between them are just different strategies to compute the loss functions. Table 8 shows the cross-modal mAP scores of MAN and the other four alternative variations on the Pascal Sentence dataset. From the experimental results, we can see that the between-class strategy improves MAN-1 by 3.19% and MAN-4 by 2.22% on average, which indicates that the proposed strategy is effective. Similarly, the eigenvalue strategy improves MAN-1 by 53.76% and MAN-2 by 52.32% on average, which indicates that the proposed eigenvalue strategy can preserve more discrimination into the common space. Furthermore, in Table 8, our proposed eigenvalue strategy is more effective than the threshold strategy of [43], which maximizes the lower bound



(a) Ratio r_i vs. i-th Class **(b)** Individual Eigenvalues

Fig. 6. Ratio-class and individual eigenvalues curves on the Pascal Sentence dataset.

of eigenvalues and ignores the largest eigenvalues. Therefore, our proposed eigenvalue strategy can preserve more discriminative information from all eigenvalues into all dimensions of common representations.

In additional to the evaluation in terms of the mAP scores, the ratio-class and individual eigenvalues curves are drawn to investigate the performance of the two proposed strategies as shown in Fig. 6, respectively. Firstly, the mean within-class and between-class distances for the ith class is respectively defined as follows:

$$D_W^{(i)} = \frac{1}{N_i} \sum_{k=1}^{m} \sum_{j=1}^{N_i^k} \|\mathbf{y}_{ij}^k - \boldsymbol{\mu}_i\|_2 \tag{18}$$

and

$$D_B^{(i)} = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}\|_2, \tag{19}$$

where $\|\cdot\|_2$ denotes the ℓ_2-norm. Then, we can use the ratio of the between-class distance to the mean within-class distance to evaluate the discriminative variance for each class. The ratio of the ith class is defined as:

$$r_i = \frac{D_W^{(i)}}{D_B^{(i)}}. \tag{20}$$

Table 7
Performance comparison in terms of mAP@R scores for the effect of adversarial learning on the Wikipedia and Pascal Sentence datasets.

Dataset	Method	Tasks					
		$R = 50$			$R = ALL$		
		Image → Text	Text → Image	Avg.	Image → Text	Text → Image	Avg.
Wikipedia	MAN($\mathcal{L}_f$ only)	0.512	0.551	0.532	0.517	0.354	0.435
	MAN	**0.517**	**0.655**	**0.586**	**0.528**	**0.480**	**0.504**
Pascal Sentences	MAN($\mathcal{L}_f$ only)	0.651	0.686	0.668	0.648	0.655	0.651
	MAN	**0.683**	**0.709**	**0.696**	**0.680**	**0.700**	**0.690**

Table 8
Performance comparison in terms of mAP@R scores using the different components on the Pascal Sentence dataset.

Method	Tasks					
	$R = 50$			$R = ALL$		
	Image → Text	Text → Image	Avg.	Image → Text	Text → Image	Avg.
MAN-1	0.432	0.463	0.447	0.430	0.448	0.439
MAN-2	0.462	0.456	0.459	0.461	0.445	0.453
MAN-3	**0.698**	0.679	0.688	0.668	0.603	0.635
MAN-4	0.678	0.698	0.688	0.670	0.680	0.675
MAN	0.683	**0.709**	**0.703**	**0.680**	**0.700**	**0.690**

From Fig. 6(a), MAN-1 has poor performance, because the overemphasis problem causes a large overlap. MAN-2 and MAN obtain more discrimination than MAN-1 and MAN-4 with the proposed between-class strategy, respectively. The improvement indicates our proposed strategy is effective. Furthermore, MAN-2 still cannot get satisfactory performance, because the eigenvalue overemphasis problem ignores some useful information in the minor eigenvalues. From the experimental results, the proposed eigenvalue strategy can also effectively push more discriminative information into the common representations. Fig. 6(a) also demonstrates that the eigenvalue overemphasis problem causes poor discrimination for MAN-1 and MAN-2. Our proposed eigenvalue strategy can maximize all eigenvalues to push as much discriminative information into all dimensions of the common space.

Overall speaking, the two proposed strategies are effective to preserve more discriminative information into the generated common representations.

5. Conclusion

In this paper, we propose a Multimodal Adversarial Network (MAN) to project multiple modalities into a latent common space for cross-modal retrieval. The proposed model is equipped with three parts, *i.e.*, multiple modality-specific generators, a multimodal discriminator and a multimodal discrimination analysis (MDA) loss. Moreover, we propose a simple but effective algorithm to optimize the MDA trace criterion like discriminator loss to avoid overemphasizing the largest distances between separated classes and the dominant eigenvalues. Comprehensive experimental results have demonstrated the effectiveness of the proposed strategies to prevent the discrimination function from overemphasizing the largest distances between the separated classes and the dominant eigenvalues. Furthermore, experiments on 4 widely-used multimodal datasets verify the effectiveness of our proposed method compared with 16 state-of-the-art methods. As for future work, we attempt to investigate how to transfer knowledge from external databases to further boost the performance of cross-modal retrieval.

Acknowledgments

This work is supported by National Key R&D Program of China under contract No. 2017YFB1002201 and partially supported by the National Natural Science Foundation of China (Grants No. 61836006, U1831121), SCU-LuZhou Sciences and Technology Coorperation Program (Grant No. 2017CDLZ-G25), Sichuan Science and Technology Planning Projects (Grants No. 18PTDJ0085, 2019YFH0075, 2018GZDZX0030), and Graduate Student's Research and Innovation Fund of Sichuan University, China (Grant No. 2018YJSY010).

Appendix A. Derivative of $\mathcal{L}_f$

To train with back-propagation we provide the partial derivatives of optimization target $\mathcal{L}_f(\mathbf{Y}^1, \mathbf{Y}^2, \ldots, \mathbf{Y}^m)$ proposed in Eq. (14) with respect to the generated representations $\mathbf{Y}^k|_{k=1}^m$ (respectively represent the common features of the corresponding modalities). As the reminder, $\mathcal{L}_f$ focuses on maximizing the n positive eigenvalues $\lambda_i|_{i=1}^n$ of the GED problem. The derivative of eigenvalue λ_i with respect to the generated representations $\mathbf{Y}^k$ of the kth modality is defined in [54] as:

$$\frac{\partial \lambda_i}{\partial \mathbf{Y}^k} = \mathbf{w}_i^T \left(\frac{\partial \mathbf{S}_B}{\partial \mathbf{Y}^k} - \lambda_i \frac{\partial \mathbf{S}_W}{\partial \mathbf{Y}^k} \right) \mathbf{w}_i \tag{A.1}$$

The partial derivative of $\mathcal{L}_f$ with respect to the representations $\mathbf{Y}^k$ of the kth modality is then defined as:

$$\begin{aligned} \frac{\partial \mathcal{L}_f}{\partial \mathbf{Y}^k} &= -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial f(\lambda_i)}{\partial \mathbf{Y}^k} \\ &= -\frac{1}{n} \sum_{i=1}^{n} \mathbf{w}_i^T \left(f'(\lambda_i) \left(\frac{\partial \mathbf{S}_B}{\partial \mathbf{Y}^k} - \lambda_i \frac{\partial \mathbf{S}_W}{\partial \mathbf{Y}^k} \right) \right) \mathbf{w}_i, \end{aligned} \tag{A.2}$$

where $f'(\cdot)$ is the derivative function of $f(\cdot)$. In this paper, $f(\cdot)$ is set as the natural logarithm function $\ln(\cdot)$. Then Eq. (A.2) can be rewritten as:

$$\frac{\partial \mathcal{L}_f}{\partial \mathbf{Y}^k} = -\frac{1}{n} \sum_{i=1}^{n} \mathbf{w}_i^T \left(\frac{1}{\lambda_i} \frac{\partial \mathbf{S}_B}{\partial \mathbf{Y}^k} - \frac{\partial \mathbf{S}_W}{\partial \mathbf{Y}^k} \right) \mathbf{w}_i \tag{A.3}$$

References

[1] J. Zhang, Y. Peng, M. Yuan, SCH-GAN: Semi-supervised cross-modal hashing by generative adversarial network, IEEE Trans. Cybern. (2018) 1–14, http://dx.doi.org/10.1109/TCYB.2018.2868826.

[2] M. Amini, N. Usunier, C. Goutte, Learning from multiple partially observed views-an application to multilingual text categorization, in: Advances in Neural Information Processing Systems, 2009, pp. 28–36.

[3] 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.
[4] 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.
[5] X. Zhai, Y. Peng, J. Xiao, Learning cross-media joint representation with sparse and semisupervised regularization, IEEE Trans. Circuits Syst. Video Technol. 24 (6) (2014) 965–978.
[6] H. Hotelling, Relations between two sets of variates, Biometrika 28 (3/4) (1936) 321–377.
[7] J. Rupnik, J. Shawe-Taylor, Multi-view canonical correlation analysis, in: Conference on Data Mining and Data Warehouses, SiKDD 2010, 2010, pp. 1–4.
[8] 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.
[9] 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.
[10] M. Kan, S. Shan, H. Zhang, S. Lao, X. Chen, Multi-view Discriminant Analysis, in: European Conference on Computer Vision, 2012, pp. 808–821.
[11] 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.
[12] P. Hu, D. Peng, J. Guo, L. Zhen, Local feature based multi-view discriminant analysis, Knowl.-Based Syst. 149 (2018) 34–46.
[13] G. Andrew, R. Arora, J. Bilmes, K. Livescu, Deep canonical correlation analysis, in: International Conference on Machine Learning, 2013, pp. 1247–1255.
[14] W. Wang, R. Arora, K. Livescu, J. Bilmes, On deep multi-view representation learning, in: International Conference on Machine Learning, 2015, pp. 1083–1092.
[15] 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.
[16] Y. Peng, J. Qi, Y. Yuan, CM-GANs: Cross-modal generative adversarial networks for common representation learning, ACM Trans. Multimed. Comput. Commu. Appl. (2018).
[17] Y. Peng, J. Qi, Y. Yuan, Modality-specific cross-modal similarity measurement with recurrent attention network, IEEE Trans. Image Process. (2018) 1.
[18] L. Zhen, P. Hu, X. Wang, D. Peng, Deep supervised cross-modal retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.
[19] 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.
[20] D. Lopez-Paz, S. Sra, A. Smola, Z. Ghahramani, B. Schölkopf, Randomized nonlinear component analysis, in: International Conference on Machine Learning, 2014, pp. 1359–1367.
[21] W. Wang, K. Livescu, Large-scale approximate kernel canonical correlation analysis, in: International Conference on Learning Representations, ICLR, 2016.
[22] 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.
[23] 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.
[24] 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.
[25] 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.
[26] X. Wang, D. Peng, P. Hu, Y. Sang, Adversarial correlated autoencoder for unsupervised multi-view representation learning, Knowl.-Based Syst. 168 (2019) 109–120, http://dx.doi.org/10.1016/j.knosys.2019.01.017.
[27] A.A. Nielsen, Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data, IEEE Trans. Image Process. 11 (3) (2002) 293–305.
[28] C. Zhang, J. Bi, S. Xu, E. Ramentol, G. Fan, B. Qiao, H. Fujita, Multi-imbalance: An open-source software for multi-class imbalance learning, Knowl.-Based Syst. 174 (2019) 137–143, http://dx.doi.org/10.1016/j.knosys.2019.03.001.
[29] X. Peng, C. Lu, Y. Zhang, H. Tang, Connections between nuclear norm and frobenius norm based representation, IEEE Trans. Neural Netw. Learn. Syst. 29 (1) (2018) 218–224, http://dx.doi.org/10.1109/TNNLS.2016.2608834.
[30] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Hum. Genet. 7 (2) (1936) 179–188.
[31] Z. Huang, H. Zhu, J.T. Zhou, X. Peng, Multiple marginal Fisher analysis, IEEE Trans. Ind. Electron. (2018) 1, http://dx.doi.org/10.1109/TIE.2018.2870413.
[32] P. Hu, D. Peng, Y. Sang, Y. Xiang, Multi-view linear discriminant analysis network, IEEE Trans. Image Process. (2019) 1–14, http://dx.doi.org/10.1109/TIP.2019.2913511.
[33] S. Akaho, A kernel method for canonical correlation analysis, in: International Meeting of Psychometric Society, 2001, pp. 263–269.
[34] 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, http://dx.doi.org/10.1109/TIP.2018.2848470.
[35] P. Hu, L. Zhen, D. Peng, P. Liu, Scalable deep multimodal learning for cross-modal retrieval, in: International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019.
[36] Y. Peng, X. Huang, J. Qi, Cross-media shared representation by hierarchical learning with multiple deep networks, in: IJCAI, 2016, pp. 3846–3853.
[37] V.E. Liong, J. Lu, Y.-P. Tan, J. Zhou, Deep coupled metric learning for cross-modal matching, IEEE Trans. Multimed. 19 (6) (2017) 1234–1244.
[38] 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, http://dx.doi.org/10.1109/TMM.2017.2742704.
[39] J. Qi, Y. Peng, Cross-modal bidirectional translation via reinforcement learning, in: IJCAI, 2018, pp. 2630–2636.
[40] 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.
[41] K. Fukunaga, Introduction to Statistical Pattern Recognition, Elsevier, 2013.
[42] 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.
[43] M. Dorfer, R. Kelz, G. Widmer, Deep linear discriminant analysis, in: International Conference on Learning Representations, ICLR, 2016.
[44] 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, http://dx.doi.org/10.1109/34.598228,
[45] D. Kinga, J.B. Adam, A method for stochastic optimization, in: International Conference on Learning Representations, ICLR, vol. 5, 2015.
[46] Y. Li, F. Nie, H. Huang, J. Huang, Large-Scale multi-view spectral clustering via bipartite graph, in: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.
[47] Y. Peng, X. Zhai, Y. Zhao, X. Huang, Semi-supervised cross-media feature learning with unified patch graph regularization, IEEE Trans. Circuits Syst. Video Technol. 26 (3) (2016) 583–596.
[48] I. Jolliffe, Principal component analysis, in: International Encyclopedia of Statistical Science, Springer, 2011, pp. 1094–1096.
[49] 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.
[50] 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.
[51] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[52] 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.
[53] Y. Kim, Convolutional neural networks for sentence classification, 2014, arXiv preprint arXiv:1408.5882.
[54] J. de Leeuw, Derivatives of Generalized Eigen Systems with Applications, 2011.