Semi-supervised Multi-modal Learning with Balanced Spectral Decomposition

Peng Hu,[1] **Hongyuan Zhu,**[1] **Xi Peng,**[2] **Jie Lin**[1*]
[1]Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore
[2]College of Computer Science, Sichuan University, Chengdu 610065, China

Abstract

Cross-modal retrieval aims to retrieve the relevant samples across different modalities, of which the key problem is how to model the correlations among different modalities while narrowing the large heterogeneous gap. In this paper, we propose a Semi-supervised Multimodal Learning Network method (SMLN) which correlates different modalities by capturing the intrinsic structure and discriminative correlation of the multimedia data. To be specific, the labeled and unlabeled data are used to construct a similarity matrix which integrates the cross-modal correlation, discrimination, and intra-modal graph information existing in the multimedia data. What is more important is that we propose a novel optimization approach to optimize our loss within a neural network which involves a spectral decomposition problem derived from a ratio trace criterion. Our optimization enjoys two advantages given below. On the one hand, the proposed approach is not limited to our loss, which could be applied to any case that is a neural network with the ratio trace criterion. On the other hand, the proposed optimization is different from existing ones which alternatively maximize the minor eigenvalues, thus overemphasizing the minor eigenvalues and ignore the dominant ones. In contrast, our method will exactly balance all eigenvalues, thus being more competitive to existing methods. Thanks to our loss and optimization strategy, our method could well preserve the discriminative and instinct information into the common neural space and embrace the scalability in handling large-scale multimedia data. To verify the effectiveness of the proposed method, extensive experiments are carried out on three widely-used multimodal datasets comparing with 13 state-of-the-art approaches.

Introduction

With the rapid growth of multimedia data such as image, text, and audio on the Internet, there are increasing demands on developing various applications to handle this data, such as classification (Guan et al. 2015), clustering (Xu et al. 2018; Peng et al. 2019; Xu et al. 2019), and retrieval (Deng et al. 2018; Hu et al. 2019b). Over the past decades, more and more attention has been attracted by retrieving the interested contents across different modalities, namely cross-modal retrieval (Peng, Qi, and Yuan 2018; Hu et al. 2019a). However, it is still challenging to correlate different modalities because distinct modalities lie in completely disparate spaces. In other words, the different modalities cannot be directly compared with each other due to the large cross-modal gap among them, *i.e.*, the so-called "heterogeneous gap".

To narrow the heterogeneous gap, some multimodal approaches proposed projecting multimedia data into a latent common space in which the similarity between any two samples from different modalities can be calculated. These methods can be roughly classified into three categories: unsupervised (Xu, Tao, and Xu 2015; Zhang et al. 2018a; Gu et al. 2018), supervised (Kan et al. 2016; Hu et al. 2019b), and semi-supervised (Zhai, Peng, and Xiao 2014; Zhang et al. 2018b) approaches. The unsupervised methods attempt to learn multiple modality-specific transformations by maximizing the correlations among different modalities, which ignore some semantic information in the multimedia data. To use the label information, some supervised and semi-supervised methods are proposed to preserve the discrimination into the latent common space (Kan et al. 2016; Zhang et al. 2018b). Although supervised cross-modal methods have achieved promising performance by utilizing the label information, they entirely rely on the labeled data and have faced two problems: 1) it is time and memory cost-prohibitive to collect well-annotated data. Especially, considering the multimedia data, such a task is more undesirable since multiple modalities will remarkably increase the labeling workload; 2) a large number of unlabeled data are much easier to obtain, but they cannot be used by the supervised approaches. Therefore, it is particularly important to boost the cross-modal retrieval performance by fully exploiting these unlabeled data. To the end, some semi-supervised cross-modal methods have been proposed to employ the intrinsic structure of the multimedia data and show promising performance in cross-modal retrieval (Zhai, Peng, and Xiao 2014; Zhang et al. 2018b). However, they have to compute

*Corresponding author: Jie Lin.

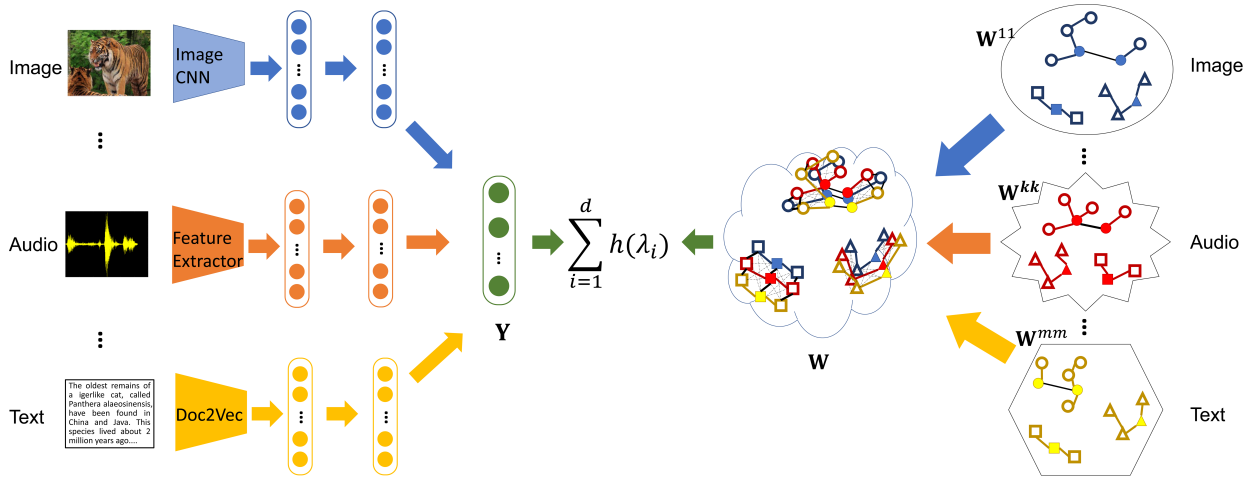

Figure 1: The framework of our SMLN for the data with m modalities which lie in distinct spaces. In the right part of the figure, the solid items represent labeled samples and the hollow items denote unlabeled points. Different shapes represent different classes, and different colors denote different modalities. $\mathbf{W}^{kk}$ is the intra-modality similarity matrix for the k-th modality. $\mathbf{W}$ and $\mathbf{Y}$ are the similarity and representation matrices for all modalities, respectively. Moreover, $\sum_{i=1}^{d} h(\lambda_i)$ is the proposed eigenvalue-based objective function, which aims to push as much discriminative information as possible into all the dimension of the common space by preserving the intrinsic structure and discrimination in the training stage. In the inference stage, the corresponding trained modality-specific network is directly used to extract the representation of a test sample.

the graph matrix based on the whole training dataset, which leads to high computational and space complexity, thus making difficulty in handling large-scale multimodal data. Furthermore, most existing semi-supervised approaches are linear methods and cannot handle the highly nonlinear complexity in real-world datasets. Although they can be extended to kernel variants, their performance is limited by the predefined kernel function and there lacks a golden criterion to choose a kernel function as pointed in (Peng et al. 2016).

To overcome these problems, we propose a Semi-supervised Multimodal Learning Network (SMLN) which correlates different modalities by capturing the intrinsic structure and discriminative correlation of the multimedia data as shown in Figure 1. In brief, the labeled and unlabeled data are used to construct a similarity matrix that encapsulates the cross-modal correlation, discrimination and intra-modal graph information of the multiple modalities.

Different from the existing graph regularization methods, our loss can be optimized in a batch-by-batch manner, thus being capable of handling large-scale multimedia data. Another major contribution of this work is proposing an effective ratio trace criterion optimizer. To be specific, our loss will involve solving a ratio trace criterion problem within a neural network. As shown in (Dorfer, Kelz, and Widmer 2016), directly optimizing the ratio trace criterion based losses including ours will lead an undesirable solution, namely, the dominant eigenvalues will be overemphasized and the information in the bottom eigenvalues will be ignored. The optimizer of (Dorfer, Kelz, and Widmer 2016) could solve this problem, but faces a new one, *i.e.*, the dominant eigenvalues will be ignored as it tries to maximize the lower bound of the eigenvalues over a given threshold. As shown in our ablation study, there is useful discriminative information in the directions of both the dominant and minor eigenvalues for cross-modal retrieval. It is therefore highly expected to consider all eigenvalues during optimizing. To the end, we propose a novel optimization strategy that could simultaneously weaken the dominant eigenvalues and emphasize the minor ones. Thanks to our optimization strategy, all the eigenvalues will be considered in the training stage and all the discriminative variances (*i.e.* eigenvalues) will be maximized without overemphasizing and ignoring any ones else. In other words, the discriminative and instinct information in all dimensions could be preserved into the learned common space. The main contributions of this paper are summarized as follows:

- A semi-supervised multimodal learning method is proposed to learn multiple nonlinear transformations by projecting multimedia data into a latent common space. As a result, the highly nonlinear cross-modal discrepancy could be eliminated and the common representations could be used to compute the similarity for the cross-modal retrieval.
- A cross-modal similarity matrix is proposed to measure the similarity between any two samples of all modalities by fully exploiting the discrimination of the labeled data, the intrinsic geometric structures, and correlation in the labeled and unlabeled data.
- A novel eigenvalue-based loss function is proposed to balance the eigenvalues instead of directly maximizing the ratio trace. Through the proposed method, the discriminative and instinct information in all dimensions can be preserved in the common space without overemphasizing the dominant eigenvalues and ignoring the minor or dominant ones.

Related Works

In this section, we will briefly review some related works which are close to this work from the following two aspects: unsupervised learning and supervised learning.

Unsupervised cross-modal methods attempt to project multimodal data into a latent common space by maximizing the correlations between different modalities. One typical method is the well-known Canonical Correlation Analysis (CCA) which maximizes the cross-modal correlation to learn two linear transformations (Hotelling 1936; Rupnik and Shawe-Taylor 2010). Similarly, another method, called Partial Least Squares (PLS), attempts to learn two linear transformations by maximizing the covariance of two modalities (Sharma and Jacobs 2011). To extend the CCA to deal with more than two modalities, Multiset CCA (MCCA) was proposed to learn a common space by maximizing the correlations between all possible pairwise modalities. Furthermore, some kernel methods are proposed to extend CCA to nonlinear models (Lopez-Paz et al. 2014; Wang and Livescu 2016). However, the predefined kernel function will limit the performance and are difficult to choose (Peng et al. 2018). Then, some researches extend CCA with Deep Neural Network (DNN). For example, (Andrew et al. 2013) proposed Deep Canonical Correlation Analysis (DCCA) to learn complex nonlinear transformations of two-modality data so that the learned representations are highly linearly correlated. Inspired by both DCCA and reconstruction-based objectives, Deep Canonically Correlation Autoencoders (DCCAE) was proposed in (Wang et al. 2015) by adding an autoencoder regularization term into DCCA.

Supervised cross-modal methods exploit the label information to learn the common representations, which are superior to unsupervised methods. With the well-know Fisher's criterion, some approaches were proposed to learn a discriminative common space by simultaneously maximizing the between-class variations and minimizing the within-class variations (Sharma et al. 2012; Kan et al. 2016). In (Wang et al. 2017), a novel Adversarial Cross-modal Retrieval method (ACMR) is presented to seek an effective common space based on adversarial learning, which consists of a feature projector, a modality classifier, and a triplet constraint. In (Hu et al. 2019b), a novel Scalable Deep Multimodal Learning method (SDML) is proposed to separately project different modalities into a predefined common space.

Although supervised methods can achieve considerable performance for cross-modal retrieval, they ignore the cross-modal correlation information in a large amount of unlabeled multimedia data. During past decades, some methods try to use the graph regularization to extract useful information in the unlabeled data (Zhai, Peng, and Xiao 2014; Zhang et al. 2018b; Liu et al. 2016). However, they have to compute the graph matrix based on the samples of the whole training dataset, which are with very high computation and space complexity, thus hindering them to handle large-scale multimodal datasets. Furthermore, they are linear methods and cannot handle the highly nonlinear complexity in many real-world applications. In this paper, we proposed a deep semi-supervised multimodal learning method, which can be trained in a batch-by-batch manner and tackle large-scale databases.

The Proposed Method

Notations

For ease of presentation, some definitions are given as below. For the training set $\mathcal{X}$ consisting of m modalities, *i.e.*, $\mathcal{X} = \{\mathcal{X}^1, \mathcal{X}^2, \cdots, \mathcal{X}^m\}$, let $\mathcal{X}^k$ denote the k-th modality. In practice, it is often time and space cost-prohibitive to annotate the large-scale multimedia data due to the heterogeneous semantic gap and a huge amount of data. Therefore, it is highly expected to utilize the labeled and unlabeled multimodal data together to boost the cross-modal retrieval performance. Consequently, the k-th modality $\mathcal{X}^k = \{\check{\mathcal{X}}^k, \hat{\mathcal{X}}^k\}$ includes both labeled data $\check{\mathcal{X}}^k = \{\check{\mathbf{x}}_1^k, \check{\mathbf{x}}_2^k, \cdots, \check{\mathbf{x}}_{\check{n}}^k\}$ and unlabeled data $\hat{\mathcal{X}}^k = \{\hat{\mathbf{x}}_1^k, \hat{\mathbf{x}}_2^k, \cdots, \hat{\mathbf{x}}_{\hat{n}}^k\}$, where $\check{\mathbf{x}}_i^k \in \mathbb{R}^{d_k \times p_k}$ and $\hat{\mathbf{x}}_i^k \in \mathbb{R}^{d_k \times p_k}$ denote the i-th labeled and unlabeled samples from the k-th modality with $d_k \times p_k$ dimensionality. Here, $\mathbf{x}_i^k$ is a vector input when $p_k = 1$, and $\check{n}$ and $\hat{n}$ are the numbers of the labeled and unlabeled samples for each modality, respectively.

For the k-th modality, the label of the i-th sample $\check{\mathbf{x}}_i^k$ is denoted as $\check{\mathbf{l}}_i^k \in \mathbb{R}^c$, where c is the class number of the dataset. All modalities share the same c categories but follow different distributions. If $\check{\mathbf{x}}_i^k$ belongs to the j-th category, $\check{l}_{ij}^k$ is set to 1, otherwise 0. For the single-label data, each label vector contains only one nonzero value. Moreover, for the multi-label data, each sample belongs to multiple classes and then its label vector contains more than one nonzero value.

As the aforementioned discussion, different modalities lie in different spaces, so it is impossible to directly calculate the similarity between two cross-modal samples. As shown in Figure 1, we aim at learning m modality-specific neural networks to project different modalities into a latent common space in which the heterogeneous samples can be compared with each other. The k-th modality-specific network can be denoted as a nonlinear function $f_k(\cdot, \Theta_k) \in \mathbb{R}^d$, where Θ_k denotes the parameters of the network and q is the dimensionality of the common space. Then, the common representation of the k-th modality is formulated as

$$\mathbf{y}_i^k = f_k(\mathbf{x}_i^k) \tag{1}$$

where $\mathbf{x}_i^k \in \mathcal{X}^k$ is the i-th samples from the k-th modality.

Objective Function

First, we use the heat kernel and labeled data to estimate the intra-modality similarity between any two samples, $\mathbf{x}_i^k$ and $\mathbf{x}_j^k$, in the same modality as follows

$$W_{ij}^{kk} = \begin{cases} g(\mathbf{x}_i^k, \mathbf{x}_j^k) & \left\{\mathbf{x}_i^k, \mathbf{x}_j^k\right\} \subset \check{\mathcal{X}}^k \\ e^{\frac{\|\mathbf{x}_i^k - \mathbf{x}_j^k\|^2}{\tau}} & \mathbf{x}_i^k \in \mathcal{N}_r(\mathbf{x}_j^k) \text{ or } \mathbf{x}_j^k \in \mathcal{N}_r(\mathbf{x}_i^k) \\ 1 & i = j \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\mathcal{N}_r(\mathbf{x}_j^k)$ and $\mathcal{N}_r(\mathbf{x}_i^k)$ denote the r nearest neighbors of $\mathbf{x}_j^k$ and $\mathbf{x}_i^k$, respectively. Moreover, $g(\cdot, \cdot)$ is a function

to compute the semantic similarity between the samples $\mathbf{x}_i^k$ and $\mathbf{x}_j^l$. Formally,

$$g(\mathbf{x}_i^k, \mathbf{x}_j^l) = \begin{cases} 1 & \ell(\mathbf{x}_i^k)^T \ell(\mathbf{x}_j^l) \geqslant 1 \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $\ell(\mathbf{x}_i^k)$ is the label vector of the point $\mathbf{x}_i^k$. For a labeled point $\check{\mathbf{x}}_i^k$, $\ell(\check{\mathbf{x}}_i^k) = \check{\mathbf{l}}_i^k$, otherwise $\ell(\hat{\mathbf{x}}_i^k) \in \mathbb{R}^c$ is a zero vector. Based on the intra-similarity matrices $\left\{\mathbf{W}^{kk}\right\}_{k=1}^m$, the inter-similarity between two samples from different modalities ($k \neq l$) can be computed by the following formulation:

$$W_{ij}^{kl} = \begin{cases} g(\mathbf{x}_i^k, \mathbf{x}_j^k) & \mathbf{x}_i^k \in \check{\mathcal{X}}^k \text{ and } \mathbf{x}_j^l \in \check{\mathcal{X}}^l \\ \frac{1}{2}(W_{ij}^{kk} + W_{ij}^{ll}) & \text{otherwise.} \end{cases} \tag{4}$$

Therefore, we can obtain the final similarity matrix $\mathbf{W}$ for the multimodal data as a block matrix, which is defined as following

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}^{11} & \mathbf{W}^{12} & \dots & \mathbf{W}^{1m} \\ \mathbf{W}^{21} & \mathbf{W}^{22} & \dots & \mathbf{W}^{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{W}^{m1} & \mathbf{W}^{m2} & \dots & \mathbf{W}^{mm} \end{bmatrix}. \tag{5}$$

The obtained similarity matrix $\mathbf{W}$ can be regarded as the weighted graph with edges connecting space- and semantic-nearby points to each other. Here, in the graph, the weight of the connected edge is the computed similarity W_{ij} between two corresponding points based on the labels and the input multimodal data. Based on the spectral graph theory (Belkin and Niyogi 2003), the connected points should be as close as possible in the latent common space. Following (Belkin and Niyogi 2003), a reasonable criterion for choosing the transformations is to minimize the following objective function:

$$\begin{aligned} [f_1^*, f_2^*, \cdots, f_m^*] &= \underset{f_1, f_2, \cdots, f_m}{\arg\min} \sum_{k=1}^{m} \sum_{l=1}^{n} \sum_{i=1}^{n} \sum_{j=1}^{n} (\mathbf{y}_i^k - \mathbf{y}_j^l)^2 W_{ij}^{kl} \\ &= \underset{\substack{f_1, f_2, \cdots, f_m \\ \mathbf{Y}^T\mathbf{D}\mathbf{Y}=\mathbf{I}}}{\arg\min} \operatorname{Tr}(\mathbf{Y}^T\mathbf{L}\mathbf{Y}) \\ &= \underset{f_1, f_2, \cdots, f_m}{\arg\min} \operatorname{Tr}\left(\frac{\mathbf{Y}^T\mathbf{L}\mathbf{Y}}{\mathbf{Y}^T\mathbf{D}\mathbf{Y}}\right) \end{aligned} \tag{6}$$

where $\operatorname{Tr}(\cdot)$ is the trace operator, $\mathbf{I} \in \mathbb{R}^{d \times d}$ is an identity matrix; $\mathbf{D}$ is a diagonal matrix whose entries are the row/column sums of $\mathbf{W}$, *i.e.*, $D_{ii} = \sum_j^{mn} W_{ij}$; $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the Laplacian matrix. Moreover, $\mathbf{Y}$ denotes the representation matrix for all modalities with the following formulation:

$$\begin{aligned} \mathbf{Y} &= [f_1(\mathcal{X}^1), f_2(\mathcal{X}^2), \cdots, f_m(\mathcal{X}^m)] \\ &= [\mathbf{y}_1^1, \cdots, \mathbf{y}_n^1, \mathbf{y}_1^2, \cdots, \mathbf{y}_i^k, \cdots, \mathbf{y}_n^m]. \end{aligned} \tag{7}$$

Obviously, Eq. (6) can be equivalent to solve the following problem:

$$\begin{aligned} [f_1^*, f_2^*, \cdots, f_m^*] &= \underset{f_1, f_2, \cdots, f_m}{\arg\max} \operatorname{Tr}\left(\frac{\mathbf{Y}^T\mathbf{D}\mathbf{Y}}{\mathbf{Y}^T\mathbf{L}\mathbf{Y}}\right) \\ &= \underset{f_1, f_2, \cdots, f_m}{\arg\max} \sum_{i=1}^{d} \lambda_i \end{aligned} \tag{8}$$

where λ_i is the i-th largest eigenvalue of the following generalized eigenvalue problem

$$\mathbf{Y}^T\mathbf{D}\mathbf{Y}\mathbf{w}_i = \lambda \mathbf{Y}^T\mathbf{L}\mathbf{Y}\mathbf{w}_i. \tag{9}$$

From Eq. (8), we can see that the objective function aims to maximize the individual eigenvalues. In particular, each eigenvalue λ_i quantifies the magnitude of the discriminative variance (separation) in the direction of the corresponding eigenvector $\mathbf{w}_i$. However, directly solving the problem in Eq. (8) (or Eq. (6)) would yield trivial solutions considering the plugged-in neural network, *e.g.*, maximizing only the largest eigenvalue since this will produce the highest reward of back-propagation. To solve the problem, (Dorfer, Kelz, and Widmer 2016) attempts to maximize the lower bound of the eigenvalues with a threshold, which will ignore some discriminative information in the dominant eigenvalues. Different from the aforementioned works, we propose a method that balances all the eigenvalues rather than the eigenvalues of lower bound as in (Dorfer, Kelz, and Widmer 2016). To refrain from overemphasizing the dominant eigenvalues, our method weakens the dominant eigenvalues and emphasizes the minor eigenvalues with the following rewritten objective function

$$[f_1^*, f_2^*, \cdots, f_m^*] = \underset{f_1, f_2, \cdots, f_m}{\arg\min} \sum_{i=1}^{d} h(\lambda_i), \tag{10}$$

where $h(x) = e^{-\alpha x}$ and α is a balance parameter. To eliminate and automatically adapt the hyper-parameter α, we set $\alpha = \frac{1}{\lceil \min\{\lambda_i | \lambda_i > 0;\ i=1,\cdots,d\} \rceil}$, where $\lceil \cdot \rceil$ rounds the element to the nearest integer that is not less than that element.

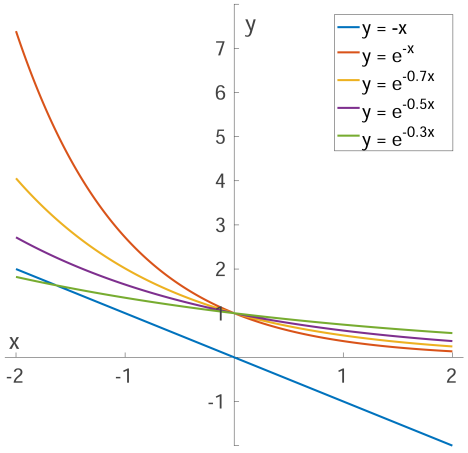

Figure 2: A toy example to illustrate how balanced spectral decomposition $h(x)$ works. The graph of $y = h(x)$ with respect to different α.

Figure 2 shows a toy example to illustrate why $h(x)$ could work well. From the figure, one could see that this function is a decreasing function as $y = -x$, which explains that it can achieve the same goal, *i.e.*, minimizing the objective. With the input value increasing, $h(x)$ is becoming smoother and reduces the differential coefficient. Moreover, when the

input value is small, it will produce a larger differential coefficient than the larger inputs. That is to say $h(x)$ can reduce the reward of back-propagation of the dominant eigenvalues and emphasize the minor eigenvalues. Therefore, the proposed method can address the overemphasizing problem of directly optimizing Eq. (8) (or Eq. (6)). We will further experimentally show its effectiveness in the following ablation study.

Algorithm 1 Optimization procedure of SMLN

Input: The training data $\mathcal{X}^k|_{k=1}^m$, the objective dimensionality of the common representation d, batch size N_b, the number of nearest neighbors r, and learning rate β.
Output: Optimized SMLN model.

1: **while** not converge **do**
2: Randomly select N_b samples from all modalities to construct a multimodal mini-batch.
3: Compute the intra- and inter-modality similarity matrix to construct the similarity matrix $\mathbf{W}$ for all modalities.
4: Compute the common representation $\mathbf{y}$ through its corresponding modality-specific network to construct the representation matrix $\mathbf{Y}$ for all modalities.
5: Compute the eigenvalues and eigenvectors by Eq. (9).
6: Update the parameters of the modality-specific networks by minimizing the objective function in Eq. (10) with descending their stochastic gradient:
$\Theta^k = \Theta^k - \beta \frac{\partial \sum_{i=1}^{d} h(\lambda_i)}{\partial \Theta^k}$ $(k = 1, \cdots, m)$
7: **end while**

Based on the new loss and the above discussions, the proposed model could be optimized in an end-to-end manner using any one stochastic gradient descent-based optimization algorithm. The detailed optimization process is summarized in Algorithm 1.

Experiment Study

In this section, we elaborate on the used datasets, implementation details, comparing methods as well as the experimental configurations.

Datasets and Features Three multimodal datasets are adopted in our experiments, including the Wikipedia dataset (Rasiwasia et al. 2010), the NUS-WIDE dataset (Chua et al. July 8 10 2009), and the XMediaNet dataset (Peng, Qi, and Yuan 2018; Peng, Huang, and Zhao 2017). The statistics of the three datasets are summarized in Table 1.

We randomly selected 5%, 10% and 30% samples from the training set as labeled data, and the rest samples as unlabeled data. Therefore, there are three groups for each dataset as shown in our experimental results.

The image features in our experiments are extracted from the fc7 layer of a 19-layer VGGNet (Krizhevsky, Sutskever, and Hinton 2012) with a dimension of $4,096$. The text representation is extracted by a Doc2Vec model[1] (Lau and Baldwin 2016) pre-trained on Wikipedia with a dimension of 300.

Dataset	Label	Modality	Instance	Feature
Wikipedia	10	Image	2,173/231/462	4,096D VGG
		Text	2,173/231/462	3,00D Doc2Vec
NUS-WIDE	10	Image	42,941/5,000/23,661	4,096D VGG
		Text	42,941/5,000/23,661	3,00D Doc2Vec
XMediaNet	200	Image	32,000/4,000/4,000	4,096D VGG
		Text	32,000/4,000/4,000	3,00D Doc2Vec

Table 1: General statistics of the three datasets used in the experiments, where "*/*/*" in the "Instance" column stands for the size of training/validation/test subsets.

Implementation Details The proposed method would train multiple modality-specific neural networks to handle the multimodal data. For each modality, the network has three fully-connected layers with each layer following a Rectified Linear Unit (ReLU) (Nair and Hinton 2010) active function except the last layer. We employ a four-layer feed-forward neural network to nonlinearly project different modalities into a latent common space, *i.e.*, $4096 \rightarrow 4096 \rightarrow 4096 \rightarrow c$ for image modality and $300 \rightarrow 4096 \rightarrow 4096 \rightarrow c$ for text modality. In the inference process, the outputs of the corresponding modality-specific networks are the common representations of the samples.

The batch size N_b is set to 128 for the Wikipedia and NUS-WIDE datasets, and 512 for the XMediaNet dataset. The number of nearest neighbors is set to 2, 3, and 3 for Wikipedia, NUS-WIDE, and XMediaNet, respectively. The dimensionality of the common space is set to c for all the datasets. The learning rate α is set to 10^{-4} in all the experiments on all datasets. For training, we employ the ADAM (Kingma and Ba 2014) optimizer with the maximal epochs of 200. Furthermore, for the supervised methods, only the corresponding percentage of the labeled data are used to train their models. Note that, for the semi-supervised and unsupervised methods, all the unlabeled and labeled multimedia data are used to train their models. The proposed model is trained on two Nvidia GTX 2080Ti GPUs in PyTorch.

Evaluation Metric and Compared Methods To evaluate the performance of the methods, we perform cross-modal retrieval tasks, *i.e.*, retrieving one modality by another modality query, such as retrieving text by image query (Img2Txt) and retrieving image by text query (Txt2Img). We adopt mean average precision (mAP) as the evaluation metric which is calculated on all returned results for a comprehensive evaluation following (Wang et al. 2017; Peng, Qi, and Yuan 2018). The cosine similarity is used to measure the distance between two samples in the obtained common space.

To demonstrate the effectiveness, we investigate the performance of 13 state-of-the-art cross-modal retrieval methods including six traditional cross-modal methods, namely MCCA (Rupnik and Shawe-Taylor 2010), GMLDA (Sharma et al. 2012), GMMFA (Sharma et al.

[1]The pre-trained Doc2Vec model is available at https://github.com/jhlau/doc2vec.

Method	5% labeled Data			10% labeled Data			30% labeled Data		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
MvDA (Kan et al. 2016)	0.208	0.205	0.206	0.219	0.208	0.213	0.287	0.265	0.276
MvDA-VC (Kan et al. 2016)	0.205	0.199	0.202	0.259	0.238	0.249	0.324	0.296	0.310
GMLDA (Sharma et al. 2012)	0.208	0.201	0.205	0.219	0.207	0.213	0.231	0.276	0.254
GMMFA (Sharma et al. 2012)	0.207	0.210	0.209	0.212	0.197	0.204	0.196	0.193	0.194
JRL (Zhai, Peng, and Xiao 2014)	0.229	0.225	0.227	0.297	0.280	0.289	0.377	0.340	0.359
GSS-SL (Zhang et al. 2018b)	0.258	0.244	0.251	0.307	0.274	0.291	0.345	0.306	0.326
MCCA (Rupnik and Shawe-Taylor 2010)	0.143	0.141	0.142	0.143	0.141	0.142	0.143	0.141	0.142
PLS (Sharma and Jacobs 2011)	0.353	0.334	0.344	0.353	0.334	0.344	0.353	0.334	0.344
DCCA (Andrew et al. 2013)	0.230	0.223	0.227	0.230	0.223	0.227	0.230	0.223	0.227
DCCAE (Wang et al. 2015)	0.260	0.250	0.255	0.260	0.250	0.255	0.260	0.250	0.255
ACMR (Wang et al. 2017)	0.219	0.212	0.216	0.294	0.271	0.282	0.390	0.354	0.372
SDML (Hu et al. 2019b)	0.267	0.233	0.250	0.347	0.291	0.319	0.450	0.393	0.421
FGCrossNet (He, Peng, and Xie 2019)	0.247	0.239	0.242	0.322	0.286	0.304	0.427	0.384	0.406
Ours	**0.389**	**0.359**	**0.374**	**0.407**	**0.362**	**0.385**	**0.459**	**0.413**	**0.436**

Table 2: Performance comparison in terms of mAP scores on the Wikipedia dataset. The highest score is shown in boldface.

Method	5% labeled Data			10% labeled Data			30% labeled Data		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
MvDA (Kan et al. 2016)	0.467	0.492	0.479	0.270	0.257	0.264	0.580	0.606	0.593
MvDA-VC (Kan et al. 2016)	0.530	0.556	0.543	0.324	0.310	0.317	0.575	0.604	0.590
GMLDA (Sharma et al. 2012)	0.392	0.430	0.411	0.252	0.247	0.249	0.531	0.512	0.521
GMMFA (Sharma et al. 2012)	0.504	0.526	0.515	0.271	0.236	0.253	0.260	0.227	0.243
JRL (Zhai, Peng, and Xiao 2014)	0.545	0.548	0.547	0.563	0.568	0.566	0.597	0.603	0.600
GSS-SL (Zhang et al. 2018b)	0.258	0.244	0.251	0.554	0.566	0.560	0.567	0.580	0.574
MCCA (Rupnik and Shawe-Taylor 2010)	0.556	0.562	0.559	0.556	0.562	0.559	0.556	0.562	0.559
PLS (Sharma and Jacobs 2011)	0.589	0.606	0.598	0.589	0.606	0.598	0.589	0.606	0.598
DCCA (Andrew et al. 2013)	0.451	0.451	0.451	0.451	0.451	0.451	0.451	0.451	0.451
DCCAE (Wang et al. 2015)	0.485	0.494	0.490	0.485	0.494	0.490	0.485	0.494	0.490
ACMR (Wang et al. 2017)	0.540	0.537	0.538	0.538	0.563	0.551	0.559	0.574	0.567
SDML (Hu et al. 2019b)	0.596	0.607	0.601	0.602	0.618	0.610	0.636	0.638	0.637
FGCrossNet (He, Peng, and Xie 2019)	0.559	0.563	0.561	0.588	0.586	0.587	0.616	0.620	0.618
Ours	**0.627**	**0.628**	**0.627**	**0.632**	**0.636**	**0.634**	**0.654**	**0.649**	**0.652**

Table 3: Performance comparison in terms of mAP scores on the NUS-WIDE dataset. The highest score is shown in boldface.

2012), JRL (Zhai, Peng, and Xiao 2014), MvDA (Kan et al. 2016), MvDA-VC (Kan et al. 2016) and GSS-SL (Zhang et al. 2018b), five DNN-based cross-modal methods, namely DCCA (Andrew et al. 2013), DCCAE (Wang et al. 2015), ACMR (Wang et al. 2017), SDML (Hu et al. 2019b), and FGCrossNet (He, Peng, and Xie 2019). For a fair comparison, all the compared methods adopt the same image and text features as our approach.

Comparisons with State-of-the-art Methods

In this section, the mAP score comparison between our proposed SMLN and other state-of-the-art methods on two retrieval tasks (Img2Txt and Txt2Img) on Wikipedia, NUS-WIDE and XMediaNet datasets are presented in Table 2, Table 3, and Table 4, respectively.

From the results, one can observe that our proposed approach achieves the highest retrieval accuracy on all datasets. Among all the compared methods, one can see that the supervised methods heavily rely on a large number of labeled data. When the size of the labeled data is small (*e.g.*, 5%), their performance is much worse than the unsupervised methods (*e.g.*, MCCA and PLS). On the other hand, the unsupervised methods can employ a large number of unlabeled data, hence can have better performance than the supervised methods trained with little labeled data. With the increasing size of labeled data, the supervised methods achieve more competitive performance than the unsupervised approaches.

Furthermore, the compared semi-supervised methods, *i.e.*, JRL and GSS-SL, can achieve good performance thanks they can use both labeled and unlabeled data. However, they are simple linear models without sufficiently exploiting the discrimination and instinct correlations in the labeled and unlabeled data, hence when labeled data is scarce ($\leq 30\%$), their performance is worse than the unsupervised methods.

In contrast, our SMLN can significantly exploit the nonlinear discrimination and correlation in the labeled and unlabeled data, and achieve the best performance. In conclusion, our proposed SMLN outperforms all the other methods on both small and big datasets, indicating that our method is a good multimodal learning method for cross-modal retrieval.

Ablation Study

In this section, we investigate the effectiveness of the similarity structure of neighbors and eigenvalue-based loss with the following two alternative baselines.

- SMLN-1 is a variant of the proposed method, which directly trains the networks with Eq. (8) instead of the proposed eigenvalue-based loss.

Method	5% labeled Data			10% labeled Data			30% labeled Data		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
MvDA (Kan et al. 2016)	0.125	0.132	0.129	0.132	0.137	0.135	0.360	0.350	0.355
MvDA-VC (Kan et al. 2016)	0.150	0.159	0.154	0.197	0.200	0.199	0.385	0.374	0.380
GMLDA (Sharma et al. 2012)	0.026	0.151	0.089	0.024	0.133	0.079	0.290	0.290	0.290
GMMFA (Sharma et al. 2012)	0.147	0.158	0.153	0.148	0.156	0.152	0.313	0.333	0.323
JRL (Zhai, Peng, and Xiao 2014)	0.130	0.130	0.130	0.249	0.236	0.242	0.347	0.323	0.335
GSS-SL (Zhang et al. 2018b)	0.184	0.191	0.187	0.256	0.251	0.254	0.326	0.311	0.319
MCCA (Rupnik and Shawe-Taylor 2010)	0.360	0.350	0.355	0.360	0.350	0.355	0.360	0.350	0.355
PLS (Sharma and Jacobs 2011)	0.277	0.266	0.271	0.277	0.266	0.271	0.277	0.266	0.271
DCCA (Andrew et al. 2013)	0.125	0.131	0.128	0.125	0.131	0.128	0.125	0.131	0.128
DCCAE (Wang et al. 2015)	0.121	0.127	0.124	0.121	0.127	0.124	0.121	0.127	0.124
ACMR (Wang et al. 2017)	0.140	0.162	0.151	0.197	0.225	0.211	0.289	0.335	0.312
SDML (Hu et al. 2019b)	0.271	0.303	0.287	0.382	0.401	0.391	0.500	0.538	0.519
FGCrossNet (He, Peng, and Xie 2019)	0.168	0.208	0.188	0.264	0.293	0.278	0.409	0.432	0.421
Ours	**0.613**	**0.612**	**0.612**	**0.618**	**0.617**	**0.617**	**0.619**	**0.620**	**0.620**

Table 4: Performance comparison in terms of mAP scores on the XMediaNet dataset. The highest score is shown in boldface.

- SMLN-2 is a variant that does not use the geometric structure.
- SMLN-3 is another one which maximizes the lower bound of the eigenvalues using the optimization strategy of (Dorfer, Kelz, and Widmer 2016).

For a fair comparison, all variants have the same network architecture and settings as our SMLN. The difference among them is the loss function. Table 5 shows the experimental results on the XMediaNet dataset. From the results, one could see that both geometric structure and eigenvalue-based loss contribute to the performance of our method. The geometric structure of each view can be used to preserve the instinct information in the common space and improve the semi-supervised performance. Furthermore, the eigenvalue-based method can be used to avoid the overemphasizing problem and push as much discriminative variance in the common space, thus facilitating the retrieval performance. Although the optimization strategy of (Dorfer, Kelz, and Widmer 2016) can address the overemphasizing problem in SMLN-1, some discriminative information in dominant eigenvalues is ignored. This information is still useful for cross-modal retrieval as shown in the comparison with our SMLN.

Labeled data	Method	Img2Txt	Txt2Img	Avg.
5%	SMLN-1	0.007	0.009	0.008
	SMLN-2	0.581	0.586	0.583
	SMLN-3	0.568	0.591	0.579
	SMLN	**0.613**	**0.612**	**0.612**
10%	SMLN-1	0.007	0.009	0.008
	SMLN-2	0.596	0.597	0.597
	SMLN-3	0.567	0.590	0.578
	SMLN	**0.618**	**0.617**	**0.617**
30%	SMLN-1	0.007	0.009	0.008
	SMLN-2	0.599	0.600	0.599
	SMLN-3	0.575	0.592	0.583
	SMLN	**0.619**	**0.620**	**0.620**

Table 5: Ablation study on the contributions of structure and eigenvalue-based loss using the XMediaNet dataset. The highest score is shown in boldface.

Conclusion

In this paper, we proposed a novel semi-supervised multimodal approach (SMLN) to fully employ the useful information in the input multimedia data, *i.e.*, instinct structure and semantic label. The labeled and unlabeled data are utilized to fully exploit the cross-modal correlation, discrimination, and graph information in the multiple modalities. Unlike the existing graph regularization methods, which need to compute the graph matrix on the whole training set, our proposed method can be trained in a batch-by-batch manner. Thus, our SMLN is more suitable to the large-scale multimedia data than the existing methods. Moreover, to solve the ratio trace problem caused by our loss, we present a novel eigenvalue-based objective function by maximizing all the eigenvalues instead of directly maximizing the ratio trace or the lower bound of eigenvalues. Therefore, our SMLN can fully use the information in the labeled and unlabeled multimedia data for cross-modal retrieval. Comprehensive experimental results on three widely-used multimedia benchmark datasets have verified the effectiveness of our SMLN by comparing with 13 state-of-the-art approaches. As for the future work, we attempt to extend our method to learn from very few labeled data, which is more difficult in cross-modal retrieval.

Acknowledgments

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funds (Project No.A1892b0026), by the Fundamental Research Funds for the Central Universities under Grant YJ201949 and 2018SCUH0070, and by the National Natural Science Foundation of China under Grant 61806135.

References

Andrew, G.; Arora, R.; Bilmes, J.; and Livescu, K. 2013. Deep canonical correlation analysis. In *International Conference on Machine Learning*, 1247–1255.

Belkin, M., and Niyogi, P. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation* 15(6):1373–1396.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y.-T. July 8-10, 2009. 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)*.

Deng, C.; Chen, Z.; Liu, X.; Gao, X.; and Tao, D. 2018. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Transactions on Image Processing* 27(8):3893–3903.

Dorfer, M.; Kelz, R.; and Widmer, G. 2016. Deep linear discriminant analysis. In *International Conference on Learning Representations (ICLR)*.

Gu, J.; Cai, J.; Joty, S.; Niu, L.; and Wang, G. 2018. 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*, 7181–7189.

Guan, Z.; Zhang, L.; Peng, J.; and Fan, J. 2015. Multi-view concept learning for data representation. *IEEE Transactions on Knowledge and Data Engineering* 27(11):3016–3028.

He, X.; Peng, Y.; and Xie, L. 2019. A new benchmark and approach for fine-grained cross-media retrieval. In *Proceedings of the 2019 ACM on Multimedia Conference*.

Hotelling, H. 1936. Relations between two sets of variates. *Biometrika* 28(3/4):321–377.

Hu, P.; Peng, D.; Sang, Y.; and Xiang, Y. 2019a. Multi-view linear discriminant analysis network. *IEEE Transactions on Image Processing* 28(11):5352–5365.

Hu, P.; Zhen, L.; Peng, D.; and Liu, P. 2019b. Scalable deep multimodal learning for cross-modal retrieval. In *Proceedings of the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, 635–644. New York, NY, USA: ACM.

Kan, M.; Shan, S.; Zhang, H.; Lao, S.; and Chen, X. 2016. Multi-view discriminant analysis. *IEEE transactions on pattern analysis and machine intelligence* 38(1):188–194.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Lau, J. H., and Baldwin, T. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Workshop on Representation Learning for NLP*, 78–86. Association for Computational Linguistics.

Liu, X.; Huang, L.; Deng, C.; Lang, B.; and Tao, D. 2016. Query-adaptive hash code ranking for large-scale multi-view visual search. *IEEE Transactions on Image Processing* 25(10):4514–4524.

Lopez-Paz, D.; Sra, S.; Smola, A.; Ghahramani, Z.; and Schölkopf, B. 2014. Randomized nonlinear component analysis. In *International Conference on Machine Learning*, 1359–1367.

Nair, V., and Hinton, G. E. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, ICML'10, 807–814. USA: Omnipress.

Peng, X.; Xiao, S.; Feng, J.; Yau, W.; and Yi, Z. 2016. Deep subspace clustering with sparsity prior. In *Proceedings of the 25 International Joint Conference on Artificial Intelligence*, 1925–1931.

Peng, X.; Feng, J.; Xiao, S.; Yau, W. Y.; Zhou, J. T.; and Yang, S. 2018. Structured autoencoders for subspace clustering. *IEEE Transactions on Image Processing* 27(10):5076–5086.

Peng, X.; Huang, Z.; Lv, J.; Zhu, H.; and Zhou, J. T. 2019. COMIC: Multi-view clustering without parameter selection. In *International Conference on Machine Learning*, 5092–5101.

Peng, Y.; Huang, X.; and Zhao, Y. 2017. An overview of cross-media retrieval: Concepts, methodologies, benchmarks and challenges. *IEEE Transactions on Circuits and Systems for Video Technology*.

Peng, Y.; Qi, J.; and Yuan, Y. 2018. Modality-specific cross-modal similarity measurement with recurrent attention network. *IEEE Transactions on Image Processing* 27(11):5585–5599.

Rasiwasia, N.; Costa Pereira, J.; Coviello, E.; Doyle, G.; Lanckriet, G. R.; Levy, R.; and Vasconcelos, N. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the international conference on Multimedia*, 251–260. ACM.

Rupnik, J., and Shawe-Taylor, J. 2010. Multi-view canonical correlation analysis. In *Conference on Data Mining and Data Warehouses (SiKDD 2010)*, 1–4.

Sharma, A., and Jacobs, D. W. 2011. Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 593–600. IEEE.

Sharma, A.; Kumar, A.; Daume, H.; and Jacobs, D. W. 2012. Generalized multiview analysis: A discriminative latent space. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2160–2167. IEEE.

Wang, W., and Livescu, K. 2016. Large-scale approximate kernel canonical correlation analysis. In *International Conference on Learning Representations (ICLR)*.

Wang, W.; Arora, R.; Livescu, K.; and Bilmes, J. 2015. On deep multi-view representation learning. In *International Conference on Machine Learning*, 1083–1092.

Wang, B.; Yang, Y.; Xu, X.; Hanjalic, A.; and Shen, H. T. 2017. Adversarial cross-modal retrieval. In *Proceedings of the 2017 ACM on Multimedia Conference*, 154–162. ACM.

Xu, C.; Guan, Z.; Zhao, W.; Niu, Y.; Wang, Q.; and Wang, Z. 2018. Deep multi-view concept learning. In *IJCAI*, 2898–2904. International Joint Conferences on Artificial Intelligence Organization.

Xu, C.; Guan, Z.; Zhao, W.; Wu, H.; Niu, Y.; and Ling, B. 2019. Adversarial incomplete multi-view clustering. In *IJCAI*, 3933–3939. International Joint Conferences on Artificial Intelligence Organization.

Xu, C.; Tao, D.; and Xu, C. 2015. Multi-view learning with incomplete views. *IEEE Transactions on Image Processing* 24(12):5812–5825.

Zhai, X.; Peng, Y.; and Xiao, J. 2014. Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology* 24(6):965–978.

Zhang, C.; Fu, H.; Hu, Q.; Cao, X.; Xie, Y.; Tao, D.; and Xu, D. 2018a. Generalized latent multi-view subspace clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1–1.

Zhang, L.; Ma, B.; Li, G.; Huang, Q.; and Tian, Q. 2018b. Generalized semi-supervised and structured subspace learning for cross-modal retrieval. *IEEE Transactions on Multimedia* 20(1):128–141.