Scalable Deep Multimodal Learning for Cross-Modal Retrieval

Peng Hu*
Machine Intelligence Laboratory
College of Computer Science, Sichuan University,
Chengdu 610065, China

Liangli Zhen*
Institute of High Performance Computing
Agency for Science, Technology and Research (A*STAR)
Singapore 138632

Dezhong Peng†
Machine Intelligence Laboratory
College of Computer Science, Sichuan University,
Chengdu 610065, China

Pei Liu
Machine Intelligence Laboratory
College of Computer Science, Sichuan University,
Chengdu 610065, China

ABSTRACT

Cross-modal retrieval takes one type of data as the query to retrieve relevant data of another type. Most of existing cross-modal retrieval approaches were proposed to learn a common subspace in a joint manner, where the data from all modalities have to be involved during the whole training process. For these approaches, the optimal parameters of different modality-specific transformations are dependent on each other and the whole model has to be retrained when handling samples from new modalities. In this paper, we present a novel cross-modal retrieval method, called Scalable Deep Multimodal Learning (SDML). It proposes to predefine a common subspace, in which the between-class variation is maximized while the within-class variation is minimized. Then, it trains m modality-specific networks for m modalities (one network for each modality) to transform the multimodal data into the predefined common subspace to achieve multimodal learning. Unlike many of the existing methods, our method can train different modality-specific networks independently and thus be scalable to the number of modalities. To the best of our knowledge, the proposed SDML could be one of the first works to independently project data of an unfixed number of modalities into a predefined common subspace. Comprehensive experimental results on four widely-used benchmark datasets demonstrate that the proposed method is effective and efficient in multimodal learning and outperforms the state-of-the-art methods in cross-modal retrieval.

CCS CONCEPTS

• **Information systems → Multimedia and multimodal retrieval**.

KEYWORDS

Cross-modal retrieval, multimodal learning, representation learning.

ACM Reference Format:
Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. 2019. 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), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3331184.3331213

1 INTRODUCTION

Cross-modal retrieval takes one type of data as the query to retrieve relevant data of another type, which enables flexible retrieval across different modalities (e.g., texts vs. images) [38]. With the rapid growth of multimedia data including image, text, video, and audio on the Internet, cross-modal retrieval is becoming increasingly important for the search engine as well as multimedia data management [39]. However, it is well known that the inconsistent representation and distribution of distinct modalities, such as image, text, and audio, cause the heterogeneity gap, which makes cross-modal similarity cannot be directly computed [28]. Therefore, the challenge of cross-modal retrieval is how to effectively measure the similarity between the samples from different modalities.

To bridge the heterogeneity gap, most of the existing methods follow the idea of representation learning to find modality-specific transformations to project the data samples from different modalities into a common subspace. In this common subspace, the similarity between different modalities can be measured directly by adopting common distance metrics. Over the past decades, a large number of cross-modal retrieval methods [26] have been developed to eliminate the heterogeneity gap and learn the common representation by different learning models. These approaches can be categorized into two classes according to their distinct models as follows: the traditional approaches and the deep approaches. The traditional cross-modal approaches attempt to learn linear or nonlinear single-layer transformations to project different modalities into a common subspace. One of the typical works is to use the statistical correlation analysis to learn linear projections by optimizing target statistical values [7, 33, 34]. To utilize the semantic information, such as the class label, some semi-supervised and supervised cross-modal methods were proposed to learn the common representations by Fisher's criterion [8, 12, 35], label space [41, 43] and so on. However, their performance is limited by the linear

*First two authors contributed equally to this work.
†D. Peng (the corresponding author) is also with the Chengdu Sobey Digital Technology Co., Ltd., Chengdu 610041, China

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331213

models, which cannot capture the complex cross-modal correlation with high nonlinearity. Although they can be easily extended to nonlinear variants by the kernel trick, such as Kernel CCA [1], their performance is limited due to the predetermined kernel functions. In addition, it is still an open issue to select a suitable kernel function for particular cross-modal learning applications [24]. To overcome aforementioned problems, inspired by the great success of deep neural networks (DNN) in representation learning [18], several DNN-based approaches have been proposed to learn the complex nonlinear transformations for cross-modal retrieval in an unsupervised [2, 6, 40] or supervised [9, 10, 30, 38, 44] manner.

However, it is notable that the aforementioned cross-modal learning methods were proposed to learn the common representations of the multimodal data in a joint manner. Thus, all the modalities must dependently join in learning the common subspace in these methods. This learning paradigm has the following two disadvantages: 1) it cannot learn different modality-specific transformations separately; and 2) the whole model has to be retrained when handling samples from new modalities.

In this paper, we present a novel cross-modal retrieval method, called Scalable Deep Multimodal Learning (SDML), which solves the above two problems simultaneously. The framework of SDML is summarized in Figure 1. Specifically, we construct m modality-specific networks for m modalities (one network for each modality) to transform the multimodal data into a common subspace. Each modality-specific network consists of a supervised loss, a modality-specific encoder, and the corresponding decoder. The proposed supervised loss aims at pushing the encoder to preserve as much discrimination as possible into the predefined common subspace. Each modality-specific decoder is stacked on the corresponding encoder to preserve semantic consistency of the modality. These m modality-specific networks do not share any parameters and are independent of each other. Different from the existing cross-modal works [9, 10, 12, 38, 40, 44, 45], these modality-specific networks are able to be trained in parallel, and we only need to train a new modality-specific network to handle samples from a new modality. To the best of our knowledge, the proposed SDML could be one of the first works to independently learn common representations from the data of an unfixed number of modalities. The main contributions and novelty of this work can be summarized as follows:

- A novel deep supervised cross-modal learning architecture is proposed to bridge the heterogeneity gap between different modalities. Unlike most of the existing methods, our proposed method predefines a common subspace to constraint the modality-specific encoder networks, which makes our method be able to learn modality-specific transformations independently. This benefits SDML can be trained in parallel and is scalable to the number of modalities.
- Multiple modality-specific decoders are designed to stack on the corresponding encoders to reconstruct the inputs. They are helpful to the encoding networks to extract underlying features that preserve semantic consistency and facilitate accurate prediction of the input data. This is also been verified by our experimental studies.
- The novel cross-modal learning strategy makes SDML much efficient in terms of the GPU resource usage or the computational time cost. Furthermore, extensive experiments on four widely-used benchmark datasets have been conducted. The results demonstrate that our method outperforms current state-of-the-art methods for cross-modal retrieval.

The remainder of this paper is organized as follows. Section 2 reviews the related work in cross-modal learning. Section 3 presents the proposed method, includes the problem definition, the SDML model and implementation details. Section 4 provides the experimental results and analysis. Section 5 concludes this paper.

2 RELATED WORK

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 kind of methods attempts to project the multimodal data into a latent common subspace by maximizing the correlation between different modalities, such as Canonical Correlation Analysis (CCA) [7], Partial Least Squares (PLS) [34], and Multiset CCA (MCCA) [22, 33]. However, some useful discriminative information, e.g., class label, is not considered in their training stages. To employ the category information, many semi-supervised and supervised approaches were proposed to learn a latent discriminative common subspace for the multimodal data. In [43], a semi-supervised cross-modal method, called Generalized Semi-supervised Structured Subspace Learning (GSS-SL), was proposed to learn the common features of two-modality data by taking the label space as a linkage to model the cross-modal correlations. In [41], a joint representation learning (JRL) was proposed to explore jointly the correlation and semantic information in a unified optimization framework. With the well-known Fisher's criterion, some approaches attempt to learn common discriminative features from multiple modalities by maximizing the between-class variations and simultaneously minimizing the within-class variations [11, 12, 35]. Although these methods have achieved the state-of-the-art performance in cross-modal retrieval, most of them are linear models and may be incapable of capturing the high-level nonlinear information of real-world multimodal data. At the first thought, they can be easily extended to nonlinear models with the kernel trick, such as Kernel CCA (KCCA) [1]. However, the learned representation is limited due to the predetermined kernel.

2.2 Deep Multimodal Representation Learning Methods

Over the past several years, the deep neural network (DNN) has achieved great success in many applications, such as image classification, object detection, and clustering [23, 46]. To overcome the shortcomings of kernel trick, DNN is widely used to project multimodal data into a high-level common subspace [27]. Firstly, some works attempt to extend traditional models into deep cross-modal

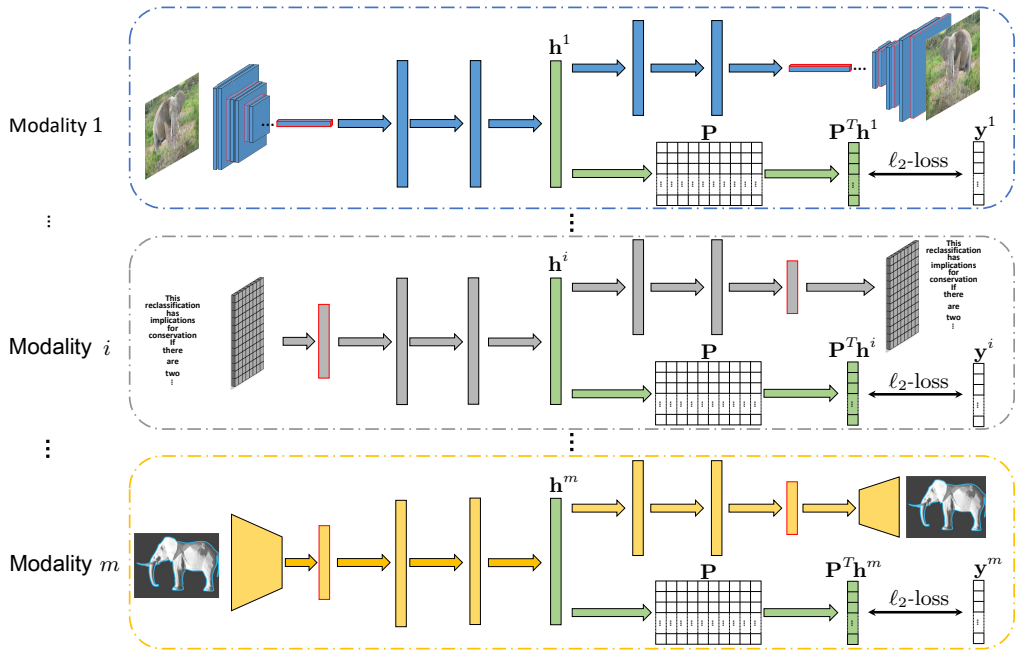


Figure 1: The general framework of the proposed SDML method. The m modality-specific neural networks (one network for each modality) can be trained separately since they do not share any trainable parameters.

models, such as Deep Canonical Correlation Analysis (DCCA) [2], Deep Canonically Correlated Autoencoders (DCCAE) [40], and Multi-view Deep Network (MvDN) [10]. DCCA uses two modality-specific subnetworks to nonlinearly project two modalities into a latent common subspace, where the resulting representations are highly linearly correlated. In [40], Wang *et al.* extent DCCA as DCCAE with adding an auto-encoder regularization term. However, some useful semantic information is ignored by these above approaches. To use the label information to boost the performance of DNNs, Kan *et al.* proposed MvDN to learn a common discriminative subspace by introducing the Fisher's criterion into a feed-forward neural network in [10]. Moreover, some works aim to utilize the inter- and intra-modality correlation to learn the common representations of multimodal data, such Cross-Media Multiple Deep Network (CMDN) [25] and Cross-modal Correlation Learning (CCL) [27]. Furthermore, some works attempt to project multimodal data into a latent common subspace by metric learning [20], cross-modal translation [30], and cross-modal hashing [5]. In [20], Deep Coupled Metric Learning (DCML) [20] adopts two DNNs to learn two sets of hierarchical nonlinear transformations (one subnetwork for each modality) so that multimodal samples are nonlinearly mapped into a shared latent feature subspace. In [30], a Cross-modal Bidirectional Translation (CBT) approach was proposed to translate one modality as a language into another modality, so that cross-modal translation can be conducted between two modalities to effectively explore cross-modal correlations with the utilization of reinforcement learning.

Unfortunately, most of the existing studies on cross-modal retrieval mainly focus on learning a latent common subspace by maximizing the correlation or discrimination of all modalities. Therefore, all modalities have to be dependently utilized to train the models. In contrast, our approach independently learns the common discriminative representations of each modality by the predefined common subspace.

3 OUR PROPOSED METHOD

In this section, we first introduce the problem definition for the multimodal learning. Then, we present the Scalable Deep Multimodal Learning (SDML) algorithm, which trains the models for different modalities independently. At last, we provide the implementation details of the proposed method.

3.1 Multimodal Learning Problem

Considering a collection of data from m modalities, we denote the j-th sample of the i-th modality as $\mathbf{x}_j^i$, and the set containing all the n_i samples of the i-th modality as $\mathcal{X}_i = \{\mathbf{x}_1^i, \mathbf{x}_2^i, \ldots, \mathbf{x}_{n_i}^i\}$. The corresponding label matrix of the i-th modality is denoted as $\mathbf{Y}_i = [\mathbf{y}_1^i, \mathbf{y}_2^i, \ldots, \mathbf{y}_{n_i}^i]$, and $\mathbf{y}_j^i = [y_{1j}^i, y_{2j}^i, \ldots, y_{cj}^i]^T \in \{0,1\}^c$ is a semantic label vector, where c is the number of semantic categories. If the sample belongs to the k-th category, $y_{kj}^i = 1$, otherwise $y_{kj}^i = 0$. Since the samples from different modalities typically have different statistical properties and lie in distinct representation spaces, we cannot directly compared the different modalities for cross-modal retrieval [38].

Multimodal learning is to learn modality-specific transformation functions for different modalities: $f_i(\mathbf{x}_j^i, \Theta_i) \in \mathbb{R}^d$, where d is the dimensionality of the representation in the common subspace, and Θ_i denotes the learnable parameters of the i-th modality-specific transformation function. In this way, the samples can be compared

in the common subspace directly even though they come from different modalities, and the similarity of the samples from the same category would be larger than the similarity of the samples from the different categories. Thus, the relevant samples of one modality can be returned for one query of another modality on cross-modal retrieval tasks.

3.2 Scalable Deep Multimodal Learning

The general flowchart of the proposed method is shown in Figure 1, from which we can see that there are m independent networks, one for each modality. They aim to learn the transform functions that can project the data of different modalities into a predefined common subspace, in which the between-class variation is maximized while the within-class variation is minimized. The key of SDML is that it predefines a common subspace by giving a fixed matrix to make the modality-specific neural networks be independent. The fixed matrix projects the representations of the samples from the common subspace into the label space, which makes the supervised information be transformed to the common subspace to supervise the learning of the modality-specific networks. Specifically, for each modality, we develop a deep supervised auto-encoder (DSAE) to transform the samples into a common subspace. DSAE is an auto-encoder (AE) with an addition of a supervised loss on the representation layer. The original AE is a neural network whose outputs are set to fitting its inputs. By learning to reconstruction the inputs, the AE can extract underlying features that facilitate accurate prediction of the input data [17]. The additional supervised loss is deduced from the label information to push as much discrimination as possible into the predefined common subspace. From Figure 1, we also can see that each supervised auto-encoder consists of three parts: an encoder, a decoder (on the upper right side), and a supervised label projection (on the lower right side).

Denoting the encoder as $\mathbf{h}_j^i = f_i(\mathbf{x}_j^i, \Theta_i)$ and the decoder as $\hat{\mathbf{x}}_j^i = g_i(\mathbf{h}_j^i, \Phi_i)$ for the i-th network (where Θ_i and Φ_i are their parameters, respectively), we formulate the objective function of the designed DSAE as follows:

$$\begin{aligned}\mathcal{J}^i &= \frac{1}{n_i}\sum_{j=1}^{n_i}[\lambda\mathcal{J}_r^i(\mathbf{x}_j^i) + (1-\lambda)\mathcal{J}_s^i(\mathbf{x}_j^i)] \\ &= \frac{1}{n_i}\sum_{j=1}^{n_i}[\lambda\|\hat{\mathbf{x}}_j^i - \mathbf{x}_j^i\|_2 + (1-\lambda)\|\mathbf{P}^T\mathbf{h}_j^i - \mathbf{y}_j^i\|_2],\end{aligned} \tag{1}$$

where λ is a balance parameter to trade-off between the reconstruction error $\mathcal{J}_r^i(\mathbf{x}_j^i)$ and the supervised loss $\mathcal{J}_s^i(\mathbf{x}_j^i)$, and $\mathbf{P}$ is a given matrix used to predefine the common subspace.

It is notable that the setting of the fixed matrix $\mathbf{P}$ is critical to the performance of the proposed method. It is desirable that in the common subspace, induced by the matrix $\mathbf{P}$, the similarity between the samples from the same class could be larger than the similarity between the samples from the different classes. Since the matrix $\mathbf{P}$ projects the common representations of the samples to the label space, we have that the ground truth label vector of each category has a corresponding (determined) vector in the common subspace. This determined vector would induce a one-dimensional subspace that can be used to guide the learning of modality-specific networks. Supposing the matrix $\mathbf{P}$ has u rows and v columns, we should set u the same as the number of output units of the encoder, and set v the same as the number of semantic categories, *i.e.*, c. In this work, we set the matrix $\mathbf{P}$ as an orthogonal matrix with orthonormal columns. This leads to the one-dimensional subspaces of different categories be orthogonal to each other, and makes the predefined common subspace more discriminative.

To achieve the multimodal learning, the proposed method needs to simultaneously minimize the objective function in Equation (1) for the m modality-specific networks as

$$\min \mathcal{J}^i \text{ for } i \in \{1, 2, \ldots, m\} \tag{2}$$

and obtains the transform functions for the m modalities.

The proposed method constructs one network for each modality and there are no any shared trainable parameters among all modality-specific networks. Thus, we can minimize the objective function of each modality-specific neural network $f_i(\mathbf{x}_j^i, \Theta_i)$ in Equation (2) separately, and train the neural networks in a paralleled manner. To minimize the objective function in Equation (1), we simply adopt the gradient descent algorithm to search the optimal parameters iteratively as

$$\begin{aligned}\Theta_i &\leftarrow \Theta_i + \alpha\frac{\partial\mathcal{J}^i}{\partial\Theta_i}; \\ \Phi_i &\leftarrow \Phi_i + \alpha\frac{\partial\mathcal{J}^i}{\partial\Phi_i},\end{aligned} \tag{3}$$

where α is the learning rate.

The objective function of SDML in Equation (2) can be optimized using a stochastic gradient descent optimization algorithm [14]. The details of the optimization procedure are summarized in Algorithm 1. The maximal number of training epochs $\aleph$ is taken as the termination condition in this work, and is typically set as 200.

Different from the existing multimodal learning approaches which learn $\mathbf{P}$ and the weight parameters $\Theta_1, \Theta_2, \ldots, \Theta_m$ simultaneously in a joint manner, the proposed method only need to learn Θ_i separately. Based on this learning strategy, SDML is scalable to the number of modalities, and all the modality-specific networks can be trained in parallel. Furthermore, SDML is efficiently to handle the samples from new modality, which only needs to train a new neural network for the new modality. However, the existing approaches have to combine the original data and the data of the new modality, and solving it as a totally new multimodal learning problem by retraining the whole model with the combined data set. Thus these approaches have a significantly higher computational complexity than our proposed method.

3.3 Implementation Details

The proposed method would train multiple modality-specific neural networks to handle the multimodal data. For each modality, the network has seven fully-connected layers with each layer following a Rectified Linear Unit (ReLU) [21] active function except the middle layer. The number of hidden units are 1024, 1024, 512, 1024, and 1024. The fixed matrix $\mathbf{P}$ is a randomly generated column orthogonal matrix. In the testing process, the decoder is ignored and the outputs of the encoder are the common representations of the samples. The cosine distance between these representations is taken as the similarity metric for cross-modal retrieval.

Algorithm 1 The optimization procedure of the proposed SDML

Input: The training data set of all modalities $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_m$, the corresponding label matrices $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$, the matrix $\mathbf{P}$, the dimensionality of the common representation space d, the batch size n_b, the learning rate α, and the hyper parameter λ.

Output: The optimal weight parameters $\Theta_1, \Theta_2, \ldots, \Theta_m$.

1: Randomly initialize the parameters $\Theta_1, \Theta_2, \ldots, \Theta_m$ and $\Phi_1, \Phi_2, \ldots, \Phi_m$.
2: **parfor** $i = 1, 2, \ldots, m$ **do** ▷ Parallel for loop
3: **while** not converge **do**
4: Randomly select n_b samples of the i-th modality $\mathcal{X}_i$ to construct a mini-batch.
5: Compute the representations $\mathbf{h}_j^i$ and $\hat{\mathbf{x}}_j^i$ for the samples in the mini-batch by forward-propagation.
6: Calculate the loss for the i-th modality-specific neural network with Equation (1).
7: Update the parameters of the i-th modality-specific neural network, Θ_i and Φ_i, by descending the stochastic gradient as follows:
$\Theta_i \leftarrow \Theta_i - \alpha \frac{\partial \mathcal{J}^i}{\partial \Theta_i}$;
$\Phi_i \leftarrow \Phi_i - \alpha \frac{\partial \mathcal{J}^i}{\partial \Phi_i}$.
8: **end while**
9: **end parfor**

The proposed model is trained on two Nvidia GTX 1080Ti GPUs with PyTorch. For training, we employ the ADAM [14] optimizer with a batch size of 100 and set the maximal number of epochs as 200.

4 EXPERIMENTAL STUDIES

To evaluate the proposed method, we conduct experiments on four datasets, namely, the PKU XMedia dataset [29, 42], the Wikipedia dataset [32], the NUS-WIDE dataset [4], and the MS-COCO dataset [19]. In our experiments, the true relevance between the samples is measured according to their semantic classes, by following the setting in a large number of cross-modal retrieval research works [27, 30, 38, 44]. In the following experiments, we first compare our SDML with 12 state-of-the-art cross-modal methods to verify its effectiveness. Then the additional evaluations are conducted to investigate the performance of SDML in more detail.

4.1 Datasets and Features

Four multimodal datasets are adopted in our experiments, including PKU XMedia, Wikipedia, NUS-WIDE and MSCOCO. The statistics of the four datasets are summarized in Table 1.

- The PKU XMedia dataset[1]: This dataset [29, 42] consists of 5,000 texts, 5,000 images, 1,143 videos, 1,000 audio clips and 500 3D models. It was divided into two parts by the authors: the training set has 10,169 instances (with 4,000 texts, 4,000 images, 969 videos, 800 audio clips and 400 3D models), and the testing set has 2,474 instances (with 1,000 texts, 1,000 images, 174 videos, 200 audio clips and 100 3D models). We further split the testing set as two subsets: the testing set has 1,237 instances and the validation set has 1,237 instances.
- The Wikipedia dataset[2]: It is the most widely-used dataset for cross-modal retrieval. The dataset [32] consists of 2,866 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 [6] to divide the dataset into three subsets: 2,173 pairs in training set, 231 pairs in validation set and 462 pairs in the testing set.
- The NUS-WIDE dataset[3]: This dataset [4] contains about 270,000 images with their tags categorized into 81 classes. Only the images exclusively belonging to one of the 10 largest categories in NUS-WIDE dataset are selected for experiments by following [27], and each image along with its corresponding tags is viewed together as an image/text pair with a unique class label. Finally, there are about 70,000 image/text pairs, where the training set consists of 42,941 pairs, the testing set consists of 23,661 pairs, and 5,000 pairs are in the validation set.
- The MS-COCO dataset[4]: This dataset [19] contains 123,287 images and their annotated sentences with their annotations categorized into 80 classes. After pruning images without category information, MSCOCO consists of 82,081 training images and 40,137 validation images, each of which is associated with five sentences. Similar to other datasets, we split the validation set as two parts: the validation subset has 10,000 pairs and the testing subset has 30,137 pairs.

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

Dataset	Label	Modality	Instance	Feature
PKU XMedia	20	Image	4,000/500/500	4,096D VGG
		Text	4,000/500/500	3,000D BoW
		Audio clip	800/100/100	29D MFCC
		3D model	400/50/50	4,700D LightField
		Video	969/87/87	4,096D C3D
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
MS-COCO	80	Image	82,081/10,000/30,137	4,096D VGG
		Text	82,081/10,000/30,137	3,00D Doc2Vec

In this work, the multimodal input features of the PKU XMedia dataset are provided by the authors. For images, each image is represented by a 4,096-dimensional CNN feature extracted from the fc7 layer of AlexNet [15], which is pre-trained on ImageNet. For text, the 3,000-dimensional BoW features are adopted as the text features.

[1]The PKU XMedia dataset is available at http://www.icst.pku.edu.cn/mipl/XMediaNet/.

[2]The Wikipedia dataset is available at http://www.svcl.ucsd.edu/projects/crossmodal/.

[3]The NUS-WIDE dataset is available at http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm.

[4]The MS-COCO dataset is available at http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm.

For video, C3D model [37] pre-trained on Sports-1M [13] is used to extract the 4,096-dimensional outputs from the fc7 layer as the video features. For audios, audio clips are represented by the 29-dimensional mel-frequency cepstral coefficients (MFCC) features. For 3D models, the models are represented by the concatenated 4,700-dimensional vectors of a LightField descriptor set [3]. For the other three datasets, i.e, Wikipedia, NUS-WIDE and MS-COCO, the image representation extractor has the same configuration with the 19-layer VGGNet [15] and the 4, 096-dimensional feature vector from the fc7 layer is extracted as the input image feature. The 300-dimensional text representation is extracted by a Doc2Vec model[5] [16], which is pre-trained on Wikipedia. The statistics of the four datasets are summarized in Table 1.

4.2 Evaluation Metric and Compared Methods

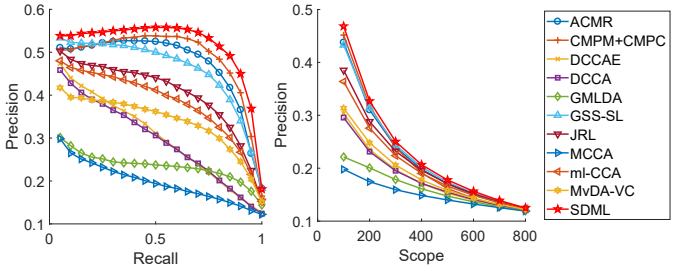
To evaluate the performance of the methods, we perform cross-modal retrieval tasks as retrieving one modality by another modality query, such as retrieving text by image query (Image → Text) and retrieving image by text query (Text → Image). We adopt mean average precision (mAP) as the evaluation metric, which is calculated on all returned results for a comprehensive evaluation.

The proposed approach is compared with 12 state-of-the-art cross-modal retrieval methods to demonstrate its effectiveness, including six traditional cross-modal methods, namely MCCA [33], GMLDA [35], JRL [41], ml-CCA [31], MvDA-VC [12] and GSS-SL [43], six DNN-based cross-modal methods, namely DCCA [2], DCCAE [40], ACMR [38], CMPM+CMPC [44], CCL [27] and CBT [30]. For a fair comparison, all the compared methods adopt the same CNN features for the image, which are extracted from the CNN architectures used in our approach. Specifically, the CNN feature for the image is extracted from the fc7 layer in the 19-layer VGGNet [36] with 4,096 dimensions. While the 300-dimensional text original representation is extracted by pre-trained Doc2Vec model [16] for Wikipedia, NUS-WIDE, and MS-COCO. On the other hand, all the features of PKU XMedia are provided by the authors. Moreover, the results of CCL and CBT are reported by their authors.

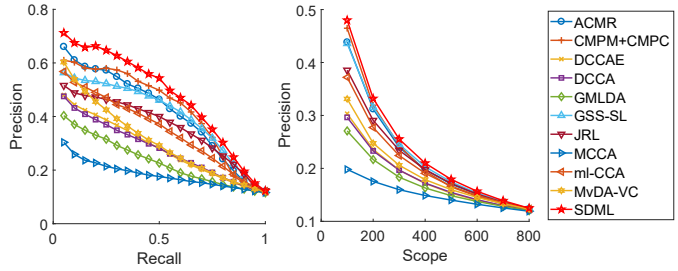
4.3 Comparison with State-of-the-art Methods

In this section, we evaluate the effectiveness of our proposed method by comparing with 12 state-of-the-art cross-modal methods on four widely-used multimodal datasets. Most compared methods can handle only two modalities, and four multimodal approaches (*i.e.*, MCCA, GMLDA, MvDA-VC and our SDML) can be directly used to learn a latent common subspace from multiple (more than two) modalities. To evaluate the performance of the two-modal method on the multimodal dataset, they should be respectively conducted in a pairwise manner on every two modalities. Specifically, the two-modality methods should be performed $\frac{m(m-1)}{2}$ times to learn the $\frac{m(m-1)}{2}$ common subspaces in pairwise manner on m modalities. Thus, the two-modal methods cost much more time to train $\frac{m(m-1)}{2}$ models than the multimodal methods which only need to train once to get a model for m modalities. There are five modalities in PKU XMedia, thus these two-modal methods must be performed 10 times on the 10 pairwise modalities. On the other hand, the multimodal methods are only conducted one time to obtain the common representations for all modalities. The mAP scores of the $5 \times 4 = 20$ cross-modal retrieval tasks and their average results are

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

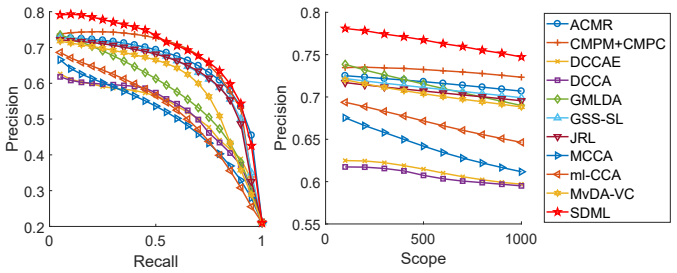


(a) Image → Text

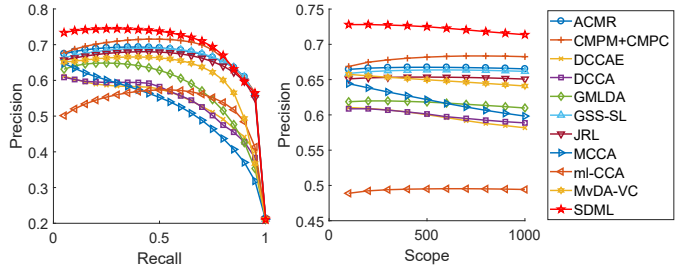


(b) Text → Image

Figure 2: 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

Figure 3: Precision-recall curves and precision-scope curves for the image-query-texts and text-query-images experiments on the NUS-WIDE dataset.

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

Method	Query	Image				Text				Audio				3D				Video				Avg.
	Target	Text	Audio	3D	Video	Image	Audio	3D	Video	Image	Text	3D	Video	Image	Text	Audio	Video	Image	Text	Audio	3D	
ml-CCA [31]*		0.597	0.241	0.284	0.377	0.613	0.242	0.243	0.355	0.202	0.210	0.159	0.176	0.228	0.165	0.130	0.203	0.374	0.307	0.166	0.252	0.276
JRL [41]*		0.770	0.296	0.521	0.376	0.788	0.279	0.477	0.348	0.260	0.233	0.302	0.214	0.522	0.365	0.211	0.403	0.399	0.272	0.172	0.366	0.379
GSS-SL [43]*		0.875	0.360	0.584	0.562	0.878	0.336	0.509	0.527	0.294	0.285	0.389	0.201	0.554	0.431	0.239	0.423	0.512	0.463	0.207	0.388	0.451
DCCA [2]*		0.869	0.264	0.186	0.463	0.871	0.306	0.221	0.406	0.218	0.265	0.221	0.157	0.169	0.169	0.177	0.155	0.433	0.369	0.167	0.163	0.312
DCCAE [40]*		0.868	0.278	0.195	0.492	0.878	0.288	0.244	0.450	0.229	0.273	0.184	0.227	0.150	0.164	0.190	0.141	0.442	0.427	0.185	0.194	0.325
CMPM+CMPC [44]*		0.897	0.544	0.637	0.641	0.896	0.500	0.669	0.675	0.568	0.551	0.481	0.397	0.578	0.666	0.388	0.506	0.583	**0.626**	**0.385**	0.497	0.584
ACMR [38]*		0.882	0.504	0.512	0.559	0.885	0.488	0.483	0.565	0.497	0.452	0.358	0.308	0.448	0.428	0.239	0.328	0.527	0.523	0.276	0.338	0.480
MCCA [33]		0.128	0.186	0.221	0.140	0.133	0.174	0.177	0.128	0.146	0.140	0.201	0.132	0.153	0.110	0.177	0.139	0.101	0.079	0.128	0.164	0.148
GMLDA [35]		0.608	0.186	0.513	0.414	0.629	0.170	0.470	0.332	0.267	0.211	0.239	0.170	0.487	0.420	0.150	0.319	0.368	0.282	0.121	0.329	0.334
MvDA-VC [12]		0.630	0.290	0.550	0.488	0.643	0.264	0.491	0.411	0.261	0.231	0.344	0.168	0.513	0.430	0.214	0.346	0.435	0.343	0.152	0.353	0.378
SDML		**0.899**	**0.552**	**0.690**	**0.659**	**0.917**	**0.572**	**0.722**	**0.686**	**0.576**	**0.604**	**0.501**	**0.425**	**0.668**	**0.694**	**0.428**	**0.533**	**0.587**	0.604	0.342	**0.514**	**0.609**

*These methods are two-modality methods.

Table 3: Performance comparison in terms of mAP scores on the Wikipedia dataset.

Method	Image → Text	Text → Image	Average
MCCA [33]	0.202	0.189	0.195
ml-CCA [31]	0.388	0.356	0.372
GMLDA [35]	0.238	0.240	0.239
JRL [41]	0.343	0.376	0.330
MvDA-VC [12]	0.397	0.345	0.387
GSS-SL [43]	0.466	0.413	0.440
DCCA [2]	0.301	0.286	0.294
DCCAE [40]	0.308	0.290	0.299
ACMR [38]	0.479	0.426	0.452
CMPM+CMPC [44]	0.493	0.438	0.466
CCL [27]	0.504	0.457	0.481
CBT [30]	0.516	0.464	0.490
SDML	**0.522**	**0.488**	**0.505**

Table 4: Performance comparison in terms of mAP scores on the NUS-WIDE dataset.

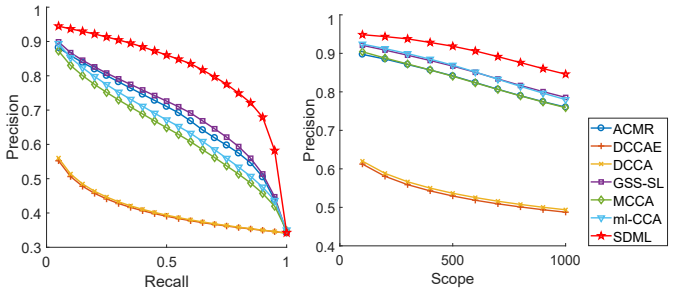
Method	Image → Text	Text → Image	Average
MCCA [33]	0.510	0.525	0.517
ml-CCA [31]	0.527	0.532	0.530
GMLDA [35]	0.582	0.577	0.580
JRL [41]	0.634	0.652	0.643
MvDA-VC [12]	0.604	0.616	0.610
GSS-SL [43]	0.640	0.659	0.650
DCCA [2]	0.524	0.541	0.533
DCCAE [40]	0.525	0.542	0.534
ACMR [38]	0.658	0.663	0.661
CMPM+CMPC [44]	0.669	0.675	0.672
CCL [27]	0.671	0.676	0.674
SDML	**0.694**	**0.699**	**0.697**

shown in Table 2 on the PKU XMedia dataset. From the experimental results, we can see that our proposed approach achieves the best performance compared to its counterparts. It outperforms the best competitor (*i.e.*, CMPM+CMPC [44]) by 4.28% on the average mAP scores. The results of Table 2 indicate that our SDML is an effective multimodal representation learning method for cross-modal retrieval on multiple modalities.

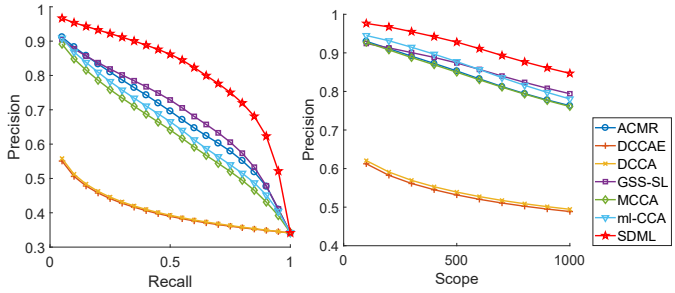
Table 5: Performance comparison in terms of mAP scores on the MS-COCO dataset.

Method	Image → Text	Text → Image	Average
MCCA [33]	0.646	0.640	0.643
ml-CCA [31]	0.667	0.661	0.664
GSS-SL [43]	0.707	0.702	0.705
DCCA [2]	0.415	0.414	0.415
DCCAE [40]	0.412	0.411	0.411
ACMR [38]	0.692	0.687	0.690
CMPM+CMPC [44]	0.759	0.748	0.753
SDML	**0.827**	**0.818**	**0.823**

On the other hand, the cross-modal retrieval across image and text is evaluated on other 3 cross-modal datasets, *i.e.*, Wikipedia, NUS-WIDE and MS-COCO. Similarly, the mAP scores of two cross-modal retrieval tasks (*i.e.*, image-query-texts and text-query-images) and their average results on the three datasets are shown in Table 3, Table 4 and Table 5, respectively. From these tables, we can see that our proposed approach also achieves the best performance. With utilizing label information, more discrimination can be extracted from the multimodal data. Therefore, the supervised methods outperform most of unsupervised methods. In Table 3, the SDML achieves the improvements of 1.16% for image-query-texts, 5.17% for text-query-images, and 3.06% for average compared with the best results of counterparts (*i.e.*, CBT [30]) on the Wikipedia dataset. In Table 4, our SDML outperforms the best competitor (*i.e.*, CMPM+CMPC) by 2.81% for image-query-texts, 2.77% for text-query-images, and 2.79% for average on the NUS-WIDE dataset. Furthermore, MS-COCO is a multi-label dataset and some methods can not handle this case, such as GMLDA, JRL and MvDA-VC. In Table 5, the proposed SDML achieves the improvements of 8.90% for image-query-texts, 9.51% for text-query-image, and 9.14% for average on the MS-COCO dataset. This indicates that our SDML is effective to handle multi-label case. In conclusion, the experimental results of Table 3, Table 4



(a) Image → Text



(b) Text → Image

Figure 4: Precision-recall curves and precision-scope curves for the image-query-texts and text-query-images experiments on the MS-COCO dataset.

and Table 5 indicate that our SDML is an effective multimodal representation learning approach for cross-modal retrieval across image and text.

In addition to the evaluation terms of the mAP score, we also draw precision-recall and precision-scope curves 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 Figure 2(a) and Figure 2(b), respectively. Similarly, the precision-recall and precision-scope curves of the image-query-texts and text-query-images on the NUS-WIDE dataset are respectively displayed in Figure 3(a) and Figure 3(b). Furthermore, the precision-recall and precision-scope curves of the image-query-texts and text-query-images on the MS-COCO dataset are plotted in in Figure 4(a) and Figure 4(b), respectively. The scope (*i.e.*, the top K retrieved samples) of the precision-scope varies from $K = 100$ to 800 on the Wikipedia dataset, and $K = 100$ to 1000 on the NUS-WIDE and MS-COCO datasets as [43]. Figure 2 and Figure 3 show the curves of our SDML and 10 state-of-the-art cross-modal methods. Since MS-COCO is a multi-label dataset, our SDML just compares with 6 state-of-the-art cross-modal methods in Figure 4. The precision-recall and precision-scope evaluations are consistent with the mAP scores for cross-modal retrieval tasks, where our SDML outperforms all the compared methods.

4.4 Parameter Analysis

To investigate the impact of the parameter λ, we analyze the performance of our SDML with different values of λ on the testing set of the Wikipedia dataset as shown in Figure 5. These plots show the mAP scores of SDML versus different values of λ for image-query-texts, text-query-images and the average results. The mAP scores is low When $\lambda = 1$, which indicates the importance of the supervised loss. In addition, we can see that the performance is not satisfactory without the auto-encoder, i.e., when λ is 0. At last, our SDML achieves the best result when $\lambda = 0.5$. Thus λ is set as 0.5 for SDML in all experiments.

4.5 Convergence Analysis

We also evaluated the convergence of our method on the Wikipedia dataset. Figure 6 plots the losses versus different numbers of epochs on the Wikipedia dataset. From the result, we can see that the proposed method converges in 100 $\sim$ 200 epochs. Thus we set maximum epoch as 200 for SDML in all experiments.

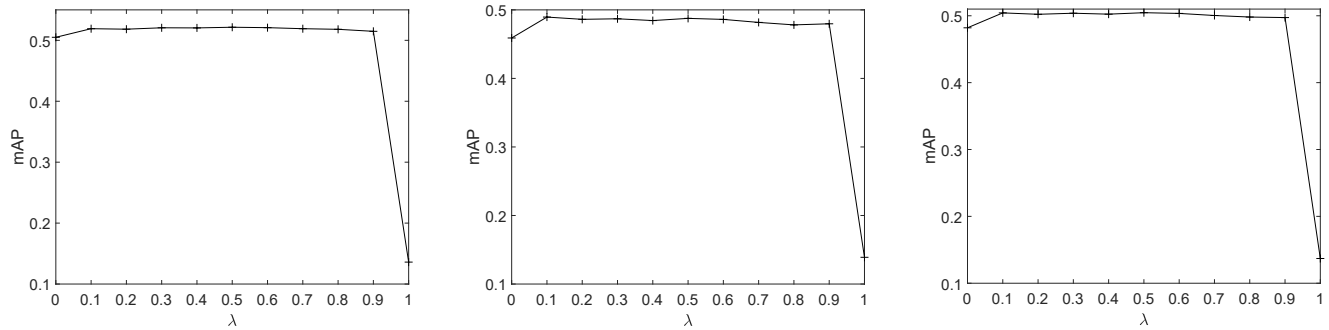
4.6 GPU Memory and Time Cost Analysis

In this section, we investigate the benefits of independently training different modality-specific networks. We define the following three alternative baselines to study the impact of independently training strategy:

- **SDML-1** is one variant of our SDML. All the modalities should be used to dependently train the networks like the existing cross-modal approaches.
- **SDML-2** serially trains each modality-specific network for the corresponding modality. At each time, just one modality-specific network is trained on the device.
- **SDML-3** parallelly trains each modality-specific model for the corresponding modality.

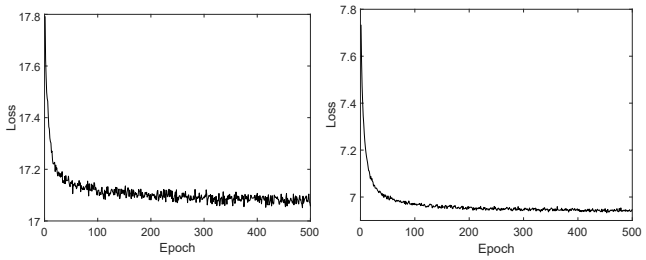
For a fair comparison, all these variants have the same network architecture and settings. The difference between them is the used training strategy. It is notable that SDML-3 can use multiple GPU devices to train different modality-specific networks in parallel and improve the training efficiency, *e.g.*, two Nvidia GTX 1080Ti GPUs are used in our experiments. Table 6 shows the comparison of GPU memory usages and training time costs among the three baselines for 200 training epochs on the four datasets. From the experimental results, we can see that independently training strategy reduces 7% $\sim$ 33% GPU memory usage or 29% $\sim$ 58% training time cost compared with the traditional joint strategy SDML-1. Note that there is a trade-off challenge between GPU memory usage and training time cost. On Wikipedia, NUS-WIDE, and MS-COCO, the serially training strategy can reduce 7.78% GPU memory usage with more training time (about 5%). However, for more modalities, *i.e.*, PKU XMedia, the serially training strategy can simultaneously reduce 33.45% GPU memory usage and 15.98% training time because of no inter-modality cost compared with the traditional joint strategy. Similarly, the parallel training can reduce much training time with more GPU memory usage. Overall, our SDML is much efficient in terms of the GPU resource usage and training time cost.

To further investigate the advantages of scalability, we compare our method with other multimodal methods in terms of the training GPU memory and time cost with the same runtime configuration on the PKU XMedia datasets. For new modalities, our method only trains the corresponding modality-specific networks instead of the whole model. However, the existing cross-modal methods should retrain the whole model. We report the time costs and the memory



(a) Image → Text (b) Text → Image (c) Average

Figure 5: Cross-modal retrieval performance of our proposed method in terms of mAP with different values of λ on the validation subset of the Wikipedia dataset.



(a) Image Loss (b) Text Loss

Figure 6: The losses vs. different numbers of epochs on the Wikipedia dataset.

Table 6: Comparison of GPU memory usages and training time costs among the three alternative baselines on the four datasets.

Method	Cost	PKU XMedia	Wikipedia	NUS-WIDE	MS-COCO
SDML-1	Memory	1,375MiB	797MiB	797MiB	797MiB
	Time	244.22s	45.10s	907.09s	1,756.46s
SDML-2	Memory	915MiB	735MiB	735MiB	735MiB
	Time	205.19s	47.41s	948.22s	1,835.60s
SDML-3	Memory	3,519MiB	1,326MiB	1,326MiB	1,326MiB
	Time	102.20s	31.88s	599.45s	1,160.66s

usages of the tested methods in Table 7. From the experimental results, we can see that our SDML can reduce 75.04% $\sim$ 99.75% training time compared with the traditional multimodal methods for a new modality.

5 CONCLUSION

In this paper, we proposed a novel approach (SDML) to independently learn the common representations for cross-modal retrieval. The proposed model is equipped with three parts, i.e, multiple independent modality-specific encoders, the corresponding modality-specific decoders, and a predefined label projection. The auto-encoder helps the model to extract underlying features that facilitate accurate prediction of each modality. The predefined common label projection allows us to project all modalities into the predetermined common subspace independently. Different from the existing methods, our SDML can train different modality-specific networks independently and be scalable to the number of modalities. Therefore, our SDML is more efficient and has broader application scenarios. Comprehensive experimental results on four widely-used multimodal datasets have demonstrated the effectiveness of our proposed method by comparing with 12 state-of-the-art methods. As for the future work, we attempt to extent our method to achieve the universal actual retrieval tasks, which usually do not have a predefined set of categories.

Table 7: Comparison of GPU memory usage and training time cost on the PKU XMedia datasets.

Method	Cost	Image	Text	Audio	3D	Video
MvDN [10]	Memory			1,065MiB		
	Time			347.80s		
MvLDAN [9]	Memory			1,065MiB		
	Time			2,681.42s		
SDML	Memory	773MiB	737MiB	599MiB	815MiB	773MiB
	Time	86.80s	69.36s	6.64s	10.64s	20.96s

ACKNOWLEDGMENT

This work is supported by National Key Research and Development Project 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), and Sichuan Science and Technology Planning Projects (Grants No. 18PTDJ0085, 2019YFH0075, 2018GZDZX0030), and Graduate Student's Research and Innovation Fund of Sichuan University (Grant No. 2018YJSY010).

REFERENCES

[1] Shotaro Akaho. 2001. A kernel method for canonical correlation analysis. In *International Meeting of Psychometric Society*. 263–269.
[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *International Conference on Machine Learning*. 1247–1255.
[3] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. 2003. On visual similarity based 3D model retrieval. In *Computer Graphics Forum*, Vol. 22. Wiley

Online Library, 223–232.
[4] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. July 8-10, 2009. NUS-WIDE: A Real-World Web Image Database from National University of Singapore. In *Proceedings of ACM Conference on Image and Video Retrieval (CIVR'09)*. Santorini, Greece.
[5] Cheng Deng, Zhaojia Chen, Xianglong Liu, Xinbo Gao, and Dacheng Tao. 2018. Triplet-Based Deep Hashing Network for Cross-Modal Retrieval. *IEEE Transactions on Image Processing* 27, 8 (2018), 3893–3903.
[6] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. 2014. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 7–16.
[7] Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika* 28, 3/4 (1936), 321–377.
[8] Peng Hu, Dezhong Peng, Jixiang Guo, and Liangli Zhen. 2018. Local feature based multi-view discriminant analysis. *Knowledge-Based Systems* 149 (2018), 34–46.
[9] Peng Hu, Dezhong Peng, Yongsheng Sang, and Yong Xiang. 2019. Multi-view Linear Discriminant Analysis Network. *IEEE Transactions on Image Processing* (2019), 1–14. https://doi.org/10.1109/TIP.2019.2913511
[10] Meina Kan, Shiguang Shan, and Xilin Chen. 2016. Multi-view deep network for cross-view classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4847–4855.
[11] Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. 2012. Multi-view Discriminant Analysis. In *European Conference on Computer Vision*. 808–821.
[12] Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. 2016. Multi-view discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 1 (2016), 188–194.
[13] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1725–1732.
[14] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. arXiv:cs.LG/1412.6980
[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. 1097–1105.
[16] Jey Han Lau and Timothy Baldwin. 2016. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*. Association for Computational Linguistics, 78–86. https://doi.org/10.18653/v1/W16-1609
[17] Lei Le, Andrew Patterson, and Martha White. 2018. Supervised autoencoders: Improving generalization performance with unsupervised regularizers. In *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 107–117.
[18] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436.
[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.
[20] Venice Erin Liong, Jiwen Lu, Yap-Peng Tan, and Jie Zhou. 2017. Deep coupled metric learning for cross-modal matching. *IEEE Transactions on Multimedia* 19, 6 (2017), 1234–1244.
[21] Vinod Nair and Geoffrey E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML'10)*. Omnipress, USA, 807–814.
[22] Allan Aasbjerg Nielsen. 2002. Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE transactions on image processing* 11, 3 (2002), 293–305.
[23] Xi Peng, Jiashi Feng, Shijie Xiao, Wei-Yun Yau, Joey Tianyi Zhou, and Songfan Yang. 2018. Structured AutoEncoders for Subspace Clustering. *IEEE Transactions on Image Processing* 27, 10 (Oct 2018), 5076–5086. https://doi.org/10.1109/TIP.2018.2848470
[24] Xi Peng, Shijie Xiao, Jiashi Feng, Wei-Yun Yau, and Zhang Yi. 2016. Deep Subspace Clustering with Sparsity Prior. In *Proceedings of the 25 International Joint Conference on Artificial Intelligence*. New York, NY, USA, 1925–1931. http://www.ijcai.org/Abstract/16/275
[25] Yuxin Peng, Xin Huang, and Jinwei Qi. 2016. Cross-Media Shared Representation by Hierarchical Learning with Multiple Deep Networks. In *IJCAI*. 3846–3853.
[26] Yuxin Peng, Xin Huang, and Yunzhen Zhao. 2017. An overview of cross-media retrieval: Concepts, methodologies, benchmarks and challenges. *IEEE Transactions on Circuits and Systems for Video Technology* (2017).
[27] Yuxin Peng, Jinwei Qi, Xin Huang, and Yuxin Yuan. 2018. CCL: Cross-modal Correlation Learning With Multigrained Fusion by Hierarchical Network. *IEEE Transactions on Multimedia* 20, 2 (Feb 2018), 405–420.
[28] Yuxin Peng, Jinwei Qi, and Yuxin Yuan. 2018. CM-GANs: Cross-modal Generative Adversarial Networks for Common Representation Learning. *ACM Transactions on Multimedia Computing, Communications, and Applications* (2018).
[29] Yuxin Peng, Xiaohua Zhai, Yunzhen Zhao, and Xin Huang. 2016. Semi-supervised cross-media feature learning with unified patch graph regularization. *IEEE Transactions on Circuits and Systems for Video Technology* 26, 3 (2016), 583–596.
[30] Jinwei Qi and Yuxin Peng. 2018. Cross-modal Bidirectional Translation via Reinforcement Learning.. In *IJCAI*. 2630–2636.
[31] Viresh Ranjan, Nikhil Rasiwasia, and CV Jawahar. 2015. Multi-label cross-modal retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*. 4094–4102.
[32] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the International Conference on Multimedia*. ACM, 251–260.
[33] Jan Rupnik and John Shawe-Taylor. 2010. Multi-view canonical correlation analysis. In *Conference on Data Mining and Data Warehouses (SiKDD 2010)*. 1–4.
[34] Abhishek Sharma and David W Jacobs. 2011. Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 593–600.
[35] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs. 2012. Generalized multiview analysis: A discriminative latent space. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2160–2167.
[36] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[37] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 4489–4497.
[38] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial cross-modal retrieval. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 154–162.
[39] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. 2016. A Comprehensive Survey on Cross-modal Retrieval. arXiv:cs.MM/1607.06215
[40] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015. On deep multi-view representation learning. In *International Conference on Machine Learning*. 1083–1092.
[41] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. 2014. Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology* 24, 6 (2014), 965–978.
[42] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. 2014. Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology* 24, 6 (2014), 965–978.
[43] Liang Zhang, Bingpeng Ma, Guorong Li, Qingming Huang, and Qi Tian. 2018. Generalized Semi-supervised and Structured Subspace Learning for Cross-Modal Retrieval. *IEEE Transactions on Multimedia* 20, 1 (2018), 128–141.
[44] Ying Zhang and Huchuan Lu. 2018. Deep Cross-Modal Projection Learning for Image-Text Matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 686–701.
[45] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. 2019. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[46] Joey Tianyi Zhou, Heng Zhao, Xi Peng, Meng Fang, Zheng Qin, and Rick Siow Mong Goh. 2018. Transfer hashing: From shallow to deep. *IEEE Transactions on Neural Networks and Learning Systems* (2018).