Information Sciences 546 (2021) 298–311

Contents lists available at ScienceDirect

Information Sciences

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

DRSL: Deep Relational Similarity Learning for Cross-modal Retrieval

Xu Wang [a,1], Peng Hu [a,e,1], Liangli Zhen [d], Dezhong Peng [a,b,c,*]

[a] Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu 610065, China
[b] Shenzhen Peng Cheng Laboratory, Shenzhen 518052, China
[c] College of Computer & Information Science, Southwest University, Chongqing 400715, China
[d] Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore 138632, Singapore
[e] Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore 138632, Singapore

ARTICLE INFO

Article history:
Received 7 May 2019
Received in revised form 27 July 2020
Accepted 5 August 2020
Available online 11 August 2020

Keywords:
Cross-modal retrieval
Relation network
Relational similarity learning
Heterogeneity gap

ABSTRACT

Cross-modal retrieval aims to retrieve relevant samples across different media modalities. Existing cross-modal retrieval approaches are contingent on learning common representations of all modalities by assuming that an equal amount of information exists in different modalities. However, since the quantity of information among cross-modal samples is unbalanced and unequal, it is inappropriate to directly match the obtained modality-specific representations across different modalities in a common space. In this paper, we propose a new method called Deep Relational Similarity Learning (DRSL) for cross-modal retrieval. Unlike existing approaches, the proposed DRSL aims to effectively bridge the heterogeneity gap of different modalities by directly learning the natural pairwise similarities instead of explicitly learning a common space. DRSL is a deep hybrid framework that integrates the relation networks module for relation learning, capturing the implicit nonlinear distance metric. To the best of our knowledge, DRSL is the first approach that incorporates relation networks into the cross-modal learning scenario. Comprehensive experimental results show that the proposed DRSL model achieves state-of-the-art results in cross-modal retrieval tasks on four widely-used benchmark datasets, *i.e.*, Wikipedia, Pascal Sentences, NUS-WIDE-10K, and XMediaNet.

1. Introduction

As an increasingly appealing research area, cross-modal retrieval aims to retrieve or search across heterogeneous media modalities (*e.g.*, image vs. text). It is generally believed that there potentially exist cross-modal correlations between different modalities. However, suffering from the heterogeneity gap triggered by the inconsistent distributions of distinct modalities, it is challenging to establish their association and measure their similarities.

Among most of the existing methods, a common strategy used to bridge the heterogeneity gap is to transform the data of different modalities into a common subspace. In such a space, the similarities among heterogeneous modalities can be directly measured by utilizing predefined distance metrics (*e.g.*, the Euclidean distance and the cosine similarity).

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

https://doi.org/10.1016/j.ins.2020.08.009
0020-0255/

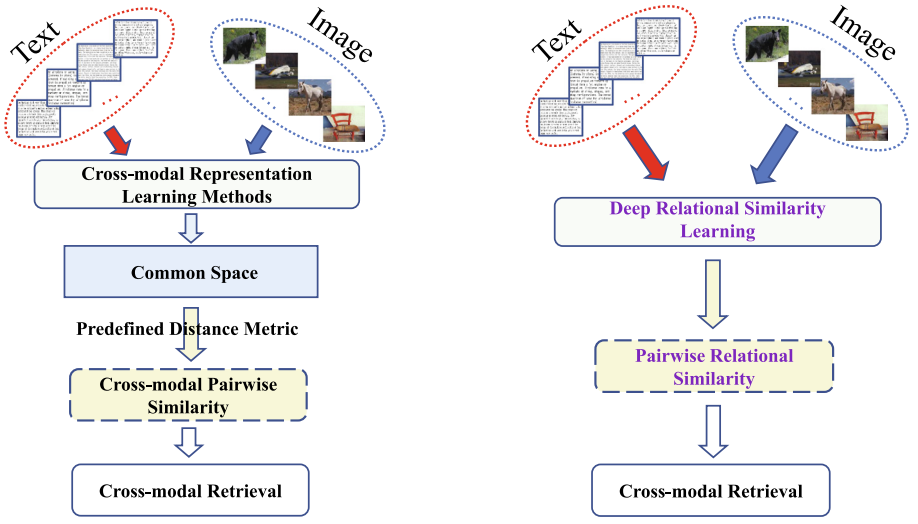
The traditional approaches [1,2] exploit pairwise statistical co-occurrence information (*e.g.*, correlations and covariances) to learn linear projections. However, these methods are based on shallow models so that they may fail in handling real data without linear subspace structure [3]. While a few kernel-based methods [4] are proposed, they suffer from the difficulties of selecting suitable kernel and handling large-scale problems, thus hindering their usage in many real applications. Profit from the great success of deep learning [5], a variety of deep learning-based methods [6–16] have emerged to bridge the heterogeneity gap between different modalities. They utilize the strong ability of deep neural networks (DNNs) to capture the nonlinear structure of data and have achieved encouraging performance.

One characteristic of the methods above is that almost all of them assume an equal amount of information can be acquired from different modalities during cross-modal learning, which is also argued in Ref. [9]. However, different modalities generally have complementary relationships, which means that they have an unbalanced and unequal information quantity in depicting the same scenario or semantics. In other words, the modality-specific features across different modalities cannot be exactly matched. In addition, these methods design different objective functions to learn a shared space, in which the commonly-used metrics can be adopted to measure similarities directly. In this way, however, one pitfall is likely to follow, *i.e.*, the loss functions are not specifically designed for the adopted metrics, thus leading to a potentially inaccurate measurement of the similarities and degrading the final retrieval performance.

In this work, we strive to address the above issues and propose a new method called Deep Relational Similarity Learning (DRSL). DRSL directly learns a pairwise relational similarity for cross-modal retrieval. Inspired by the recently proposed relation networks (RNs) [17–19], we exploit a relation networks module to learn the intrinsic pairwise similarities between image and text modalities. Specifically, DRSL learns the similarities for cross-modal retrieval through the following four steps: 1) Employing two convolutional neural networks (CNNs) to obtain two modality-specific deep nonlinear transformations, which project image and text into two modality-specific spaces, respectively; 2) Fusing the modality-specific representations to constitute pairwise samples; 3) Feeding the pairwise samples into RNs module to compute their relational similarity matrix; and 4) Ranking the similarities for cross-modal retrieval.

Fig. 1 illustrates the difference between existing cross-modal retrieval methods and DRSL. More specifically, DRSL directly generates cross-modal similarities in a deep learning architecture, capturing the nonlinear metric between image and text. Meanwhile, DRSL is a deep hybrid framework that integrates the relation networks module for relation learning. It effectively mitigates the heterogeneity gap between different modalities without explicitly learning the common space. To the best of our knowledge, DRSL is the first approach that incorporates RNs into the cross-modal learning scenario. Comprehensive experimental results show that our DRSL model achieves state-of-the-art results in cross-modal retrieval tasks on four widely-used cross-modal benchmarks, *i.e.*, Wikipedia, Pascal Sentences, NUS-WIDE-10K, and XMediaNet.

Notations: Throughout this paper, **Calligraphy letters** denote sets, **Lower-case bold letters** denote column vectors and **Upper-case bold letters** represent matrices, unless otherwise stated.



(a) Existing Cross-modal Representation Learning (b) Deep Relational Similarity Learning (proposed)

Fig. 1. The illustration of the comparison between the existing cross-modal representation learning methods and the proposed Deep Relational Similarity Learning (DRSL). (a) Existing cross-modal representation learning methods transform different modalities into a common space, where the similarities are measured by predefined distance metrics, *e.g.*, the Euclidean distance and the cosine similarity. In contrast, (b) the proposed DRSL directly learns a pairwise relational similarity matrix by using deep neural networks, without explicitly learning a common space.

2. Related work

In this section, we review the related methods from the following two aspects: common space learning for cross-modal retrieval and relation network.

2.1. Common space learning for cross-modal retrieval

There are many methods proposed to learn common representations to eliminate the heterogeneity gap between distinct modalities. We briefly introduce some representative ones of them with the following two classes: 1) traditional cross-modal representation learning methods, and 2) deep cross-modal representation learning methods.

Traditional cross-modal representation methods aim to learn single-layer linear or nonlinear transformations [6,20,21]. As the most representative one, Canonical Correlation Analysis (CCA) [6] attempts to obtain the common subspace of two modalities by maximizing the cross-modal correlation. As another common unsupervised cross-modal approach, Partial Least Squares (PLS) [2] linearly transforms cross-modal data into a shared subspace in which heterogeneous modalities are highly correlated. They can be easily extended to nonlinear models with the kernel trick, *e.g.*, kernel canonical correlation analysis [4]. To utilize semantic information (*e.g.*, class label) and enhance the correlation of cross-modal data, a general multiview feature extraction approach called Generalized Multiview Analysis (GMA) [22] is presented to transform distinct feature spaces into a common one. Specifically, GMA utilizes the discriminative information of every modality and the pairwise relationship of any two modalities to transform the samples from different modalities into a common subspace. In Ref. [23], a Joint Representation Learning (JRL) is presented to jointly exploit the correlation and semantic information under a unified optimization framework. In Ref. [24], the authors proposed a Simple to Complex Cross-Modal learning to rank framework (SCCM) with diversity regularization to determine an optimal latent space, achieving encouraging performance. Additionally, in Ref. [25], Kan et al. presented Multiview Discriminant Analysis method (MvDA), which utilizes the Fisher criterion on all modalities to learn different view-specific linear projections, one for each view. Since observations from different views share similar data structures, a constraint is introduced in Ref. [20] to enforce the view consistency of the multiple linear transforms, obtaining a method called MvDA-VC.

Moreover, to utilize both labeled data and unlabeled data, several semi-supervised cross-modal methods have been developed recently. In Ref. [26], the authors proposed Generalized Semi-supervised Structured Subspace Learning (GSS-SL), which exploits the label space as a connection to establish the correlations among distinct modalities, learning a common discriminative subspace. In Ref. [27], an effective method is proposed to adaptively learn the labels for unlabeled data and ensure that the original feature distribution is consistent with the semantic distribution in the common subspace.

In the past several years, deep neural networks (DNNs) have come a long way in many single-modal problems (*e.g.*, object detection and image classification) [3,28,29] due to its strong power of nonlinear modeling. Benefited from the success, *deep cross-modal representation learning methods* have sprung up to model the cross-modal correlation [6,7,30–35]. In Ref. [6], the authors proposed Deep Canonical Correlation Analysis (DCCA) to learn the complex nonlinear projections of two modalities, so that the representations have high linear correlations. DCCA can be regarded as a deep extension of the linear CCA. In DCCA, two modality-specific subnetworks are used to respectively extract deep nonlinear features from the corresponding modalities by maximizing the canonical correlation between the features. By exploiting both DCCA and reconstruction-based objectives, Wang et al. [7] presented a model that optimizes the combination of two terms: 1) the canonical correlation between the learned bottleneck representations, 2) the reconstruction loss of the autoencoders. Though both DCCA and DCCAE learn the nonlinear projections in a latent space using DNN, some discriminative information might be discarded without employing the semantic information (*e.g.*, the class labels). To exploit the supervised information, Kan et al. proposed the Multi-view Deep Network (MvDN) [31], which learns modality-invariant representation via introducing Fisher's criterion into a DNN. Besides, in Ref. [32], the authors proposed Deep Coupled Metric Learning (DCML). It adopts coupled neural networks to obtain two deep transformations (one for each modality) to project cross-modal samples into a common feature space.

Overall speaking, all approaches mentioned above aim to learn common representations of distinct modalities for cross-modal retrieval, assuming that an equal amount of information exists in different modalities. However, the amount of knowledge among cross-modal samples generally is unbalanced and unequal. Therefore, it is inappropriate to directly match the obtained modality-specific representations across different modalities.

2.2. Relation network

It is generally considered that modeling relations between samples is helpful to object recognition. A relation network (RN) is a structural neural network module, which can mine the relation between objects. The design idea behind RNs is to constrain the functional form of neural networks to capture the core commonality of the pairwise relation between samples. From Ref. [36], the simplest form of RN is the composite function shown below:

$$\mathrm{RN}(\mathcal{O}) = f_{\Phi}\left(\sum_{ij} g_{\Theta}(\mathbf{o}_i, \mathbf{o}_j)\right), \tag{1}$$

where $\mathcal{O} = \{\mathbf{o}_i\}_{i=1}^{n}$ is a set of input "objects", $\mathbf{o}_i$ denotes the i-th object, and $f_{\mathbf{\Phi}}$ and $g_{\mathbf{\Theta}}$ respectively denote the functions with learnable parameters $\mathbf{\Phi}$ and $\mathbf{\Theta}$. The output of $g_{\mathbf{\Theta}}$ is a "relation" between the two input samples. Therefore, the role of $g_{\mathbf{\Theta}}$ is to infer in what way the two samples are related, or whether they are related. Furthermore, RNs have three significant advantages: 1) they learn to infer relationships, 2) they are data-efficient, and 3) they operate on a set of objects [18].

The existing RNs-based works mainly focus on the single-modal problem. One of the most popular applications is relational reasoning. In Ref. [36], Raposo et al. introduced a general-purpose neural network architecture, *i.e.*, relation networks, for learning to reason about or model objects and their relations. In Ref. [18], Santoro et al. explored RNs as a general solution to relational reasoning in neural networks. Furthermore, the authors also described how to use RNs as a simple plug-and-play module to solve problems that depend on relational reasoning fundamentally. Besides relational reasoning, Hu et al. [19] proposed an objective relation module, which processes a group of objects at the same time through the interaction between the appearance feature and geometry, allowing the relationship between them to be modeled. Sun et al. [37] modeled the spatio-temporal relations to capture the interactions between human actors, relevant objects, and scene elements essential to differentiate similar human actions. This approach is weakly supervised and mines the relevant elements automatically with an actor-centric relational network (ACRN). ACRN computes and accumulates pairwise relation information from actor and global scene features, and generates relation features for action classification. Sung et al. [17] proposed a two-branch RN that performs few-shot recognition by learning to compare query images against few-shot labeled sample images, where a classifier is used to recognize new classes which are given only a few examples from each.

Although RNs have achieved great success in the aforementioned single-modal applications, the methods mentioned above cannot directly handle heterogeneous multimedia data. There is no significant progress in exploiting modality relations for cross-modal learning. In this paper, we investigate how to apply RNs in cross-modal retrieval.

3. The proposed method

The pipeline of our DRSL is shown in Fig. 2. From the figure, we could see that two modality-specific networks and one relation network are adopted to compute the pairwise similarities in a DNN-based framework.

3.1. Problem formulation

Without losing generality, the bimodal case is only considered in the paper, *i.e.*, for image and text. First of all, we give some formal definition of bimodal data as follows. The image modality is denoted as $\mathcal{X}^i = \left\{\mathbf{X}_p^i\right\}_{p=1}^{n_i}$, where n_i is the number of image points, and $\mathbf{X}_p^i$ is the p-th image sample. Similarly, the text modality $\mathcal{X}^t = \left\{\mathbf{X}_q^t\right\}_{q=1}^{n_t}$ totally consists of n_t samples, and $\mathbf{X}_q^t$ is the q-th text instance. Besides, each sample is classified into its corresponding category, ant the labels of training set are denoted as $\mathcal{Y}^i = \left\{y_p^i\right\}_{p=1}^{n_i}$ and $\mathcal{Y}^t = \left\{y_q^t\right\}_{q=1}^{n_t}$ for image and text, respectively. It is notable that all modalities (*i.e.*, image and text) share the same categories.

For cross-modal retrieval, one sample from images $\mathcal{X}^i$ (resp. texts $\mathcal{X}^t$) are taken as a query to retrieve the correlative samples of texts $\mathcal{X}^t$ (resp. images $\mathcal{X}^i$). However, different modalities (images $\mathcal{X}^i$ and texts $\mathcal{X}^t$) are not directly comparable for cross-modal retrieval since they share distinct features and distributions. In addition, it is challenging to establish correlations and measure the similarities between distinct modalities. Most of existing cross-modal approaches assume an equal amount of information can be acquired from distinct modalities during multimodal representation learning. They attempt to project different modalities into a latent single shared space wherein the representations of distinct modalities are directly comparable. Then a widely-used distance metric is adopted to compute the similarities between different modalities for cross-modal retrieval. Different from existing cross-modal approaches, our DRSL could learn the cross-modal similarities in a DNN-based architecture that captures the nonlinear metric between distinct modalities. DRSL eliminates the cross-modal discrepancy without projecting all modalities into one common space but learns a satisfactory cross-modal similarity metric by the proposed deep model.

3.2. Framework

In this subsection, we introduce the detailed network architectures of two modality-specific networks and the relation network in the proposed DRSL approach. The architecture of DRSL is shown in Fig. 2.

3.2.1. Cross-modal networks module

We design two modality-specific convolutional subnetworks to form two parallel feature extraction models for the two modalities, denoted as $f_i(\mathbf{X}^i; \mathbf{\Theta}_i)$ for image $\mathbf{X}^i$ and $f_t(\mathbf{X}^t; \mathbf{\Theta}_t)$ for text $\mathbf{X}^t$, with parameters $\mathbf{\Theta}_i$ and $\mathbf{\Theta}_t$, respectively. Each cross-modal subnetwork contains several CNN layers (*i.e.*, image CNN or text CNN) and fully-connected layers to learn highly nonlinear semantic features from the corresponding modality.

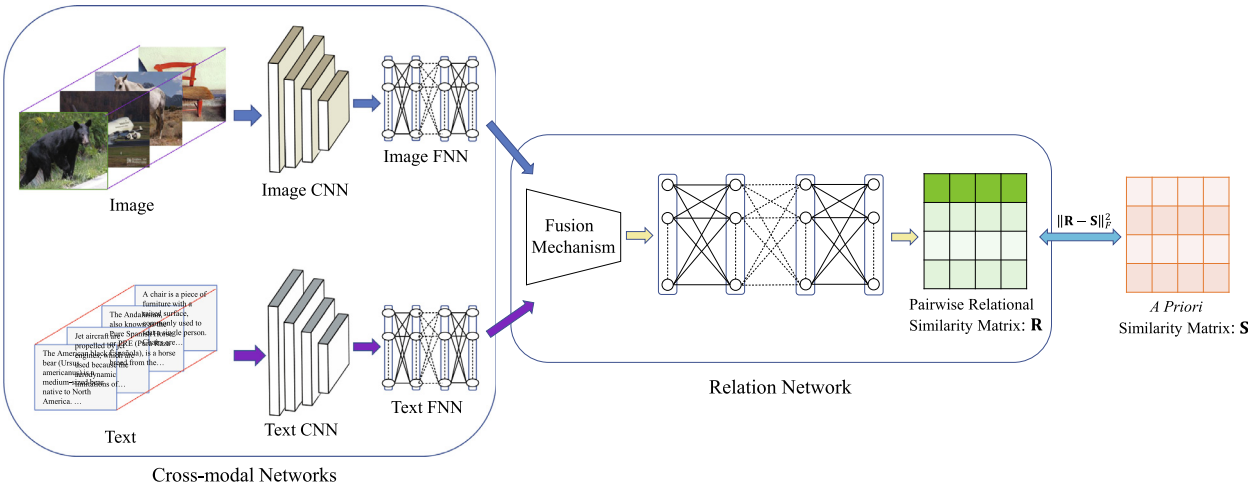


Fig. 2. The architecture of the proposed DRSL approach for cross-modal retrieval. DRSL consists of two main modules, *i.e.*, the cross-modal networks module and the relation network module. The top row of cross-modal networks represents the image feature extraction model, which consists of one convolutional neural network (Image CNN) and one fully-connected neural network (Image FNN). These two networks aim to extract high-level representations from images. The bottom row of cross-modal networks represents the text feature extraction model, which consists of one Text CNN and one Text FNN. These two networks aim to extract high-level representations from texts. Once the modality-specific representations are obtained, we fuse them to constitute pairwise samples through the fusion mechanism. Then the pairwise samples are fed into the relation network to obtain a pairwise relational similarity matrix **R**, which is expected to approximate an *a priori* similarity matrix **S**, by minimizing the square of the Frobenius norm of $(\mathbf{R}-\mathbf{S})$.

For the image modality, an image input $\mathbf{X}_p^i$ is resized to 256×256 and cropped as 224×224, and then fed into the corresponding cross-modal subnetwork to extract the high-level semantic features. Specifically, the CNN architecture of image subnetwork is the same as the VGG-19 [38], which is pretrained on ImageNet and finetuned on the training images $\mathscr{X}^i$ following Refs. [9,13]. Then several additional fully-connected layers are stacked on the fc7 layer of VGG-19 to extract high-level image representations, which will be used to compute the cross-modal similarities by the relation subnetwork. We can denote the obtained modality-specific feature vector for the p-th image $\mathbf{X}_p^i$ as

$$\mathbf{z}_p^i = f_i(\mathbf{X}_p^i). \tag{2}$$

For text modality, we assume that the q-th text instance with m' words could be represented as a matrix $\mathbf{X}_q^t \in \mathbb{R}^{k \times m'}$, where k is the dimensionality of a word feature. However, each text instance could not share the same word amount. Thus each text instance could be represented as a $k \times m$ matrix to obtain same-dimension inputs, where m is the maximal number of words in each text instances and vacant columns of the matrix are filled with zeros. In this paper, each word could be represented by a 300-dimensional vector, which is extracted by using the pretrained Word2Vec model [39] provided by the authors. Like image subnetwork, the text subnetwork also contains CNN and fully-connected layers. The text CNN layers are with the same configuration of Ref. [40]. Similarly, several additional fully-connected layers are stacked on the CNN layers to learn the high-level text representations, which will be used to calculate the cross-modal similarities with the image features by the relation subnetwork. We can also denote the obtained modality-specific feature vector for the q-th text $\mathbf{X}_q^t$ as

$$\mathbf{z}_q^t = f_t(\mathbf{X}_q^t). \tag{3}$$

3.2.2. Relation network module

The proposed relation network module consists of a fusion mechanism and a relation network. The fusion mechanism can be denoted as $g(\mathscr{Z}^i, \mathscr{Z}^t)$, where $\mathscr{Z}^i = \left\{\mathbf{z}_p^i\right\}_{p=1}^{n_i}$ and $\mathscr{Z}^t = \left\{\mathbf{z}_q^t\right\}_{q=1}^{n_t}$ are the image and text feature vector set, respectively. Similarly, the relation network is denoted as $k(\mathbf{v}; \mathbf{\Theta}_r)$, where $\mathbf{v}$ is the fused result and $\mathbf{\Theta}_r$ are the parameters of the relation subnetwork.

All cross-modal samples are fused in pairs by the fusion mechanism. Firstly, images and texts are fed into the cross-modal network module to produce the corresponding modality-specific features. Then any pairwise cross-modal samples (an image feature vector $\mathbf{z}^i$ and a text feature vector $\mathbf{z}^t$) are fused as $\mathbf{v}$. The fusion function can be a mathematical operation, such as a concatenation, multiplication, addition, or subtraction operation. We denote the fused feature vector set for the image feature set $\mathscr{Z}^i$ and the text feature set $\mathscr{Z}^t$ as

$$\mathscr{V} = g(\mathscr{Z}^i, \mathscr{Z}^t), \tag{4}$$

$\mathscr{V} = \{\mathbf{v}_{pq} | p = 1, \ldots, n_i; q = 1, \ldots, n_t\}$ is the fused feature vector set and $\mathbf{v}_{pq}$ is the fused feature vector of the p-th image and the q-th text.

Finally, the obtained pairwise fused feature vectors are used to compute the corresponding similarities. Unlike existing methods, the pairwise similarities are computed by the relation network trained in a DNN-based fashion. Specifically, the fused feature vector $\mathbf{v}_{pq}$ of an image-text pair $< \mathbf{z}_p^i, \mathbf{z}_q^t >$ is fed into the relation network to compute their relational similarity. The similarity R_{pq} between $\mathbf{z}_p^i$ and $\mathbf{z}_q^t$ is denoted as

$$R_{pq} = k(\mathbf{v}_{pq}; \mathbf{\Theta}_r). \tag{5}$$

3.3. Objective function

From the description of Subsection 3.2, the image network transforms the image set $\mathscr{X}^i$ to the image feature vector set $\mathscr{Z}^i$ as follows:

$$\mathscr{Z}^i = f_i(\mathscr{X}^i). \tag{6}$$

Similarly, the text set $\mathscr{X}^i$ can be projected as the text feature vector set $\mathscr{Z}^t$ by the text network as follows:

$$\mathscr{Z}^t = f_t(\mathscr{X}^t). \tag{7}$$

Unlike existing cross-modal approaches, the obtained image/text features are not directly used to compute the image-text similarities by predetermined metrics. To obtain the pairwise similarities, we compute the fused feature vectors of the pairwise samples using the fusion mechanism, according to Eq. (4). Then the similarities are computed according to Eq. (5). We denote the whole process as a nonlinear function $h(\cdot, \cdot)$. Therefore, the pairwise similarities between the image set and the text set can be computed through the relation network module as

$$\mathbf{R} = h(\mathscr{Z}^i, \mathscr{Z}^t), \tag{8}$$

where $\mathbf{R} \in \mathbb{R}^{n_i \times n_t}$ is the similarity matrix of the pairwise samples between $\mathscr{Z}^i$ and $\mathscr{Z}^t$, and R_{pq} is the similarity between p-th image sample and q-th text sample. Intuitively, the similarities between the same classes are much larger than the similarities between different classes. We define an *a priori* similarity matrix $\mathbf{S} \in \mathbb{R}^{n_i \times n_t}$ for the image set $\mathscr{X}^i$ and the text set $\mathscr{X}^t$, where S_{pq} is the similarity between p-th image $\mathbf{X}_p^i$ and q-th text $\mathbf{X}_q^t$. Moreover, we define the value of intra-class similarity is 1, and the value of inter-class similarity is 0 as follows:

$$S_{pq} = \begin{cases} 1, & \text{if } y_p^i = y_q^t; \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

In the proposed model, the similarity matrix $\mathbf{R}$ is expected to approximate the *a priori* similarity matrix $\mathbf{S}$. Therefore, the loss function of our DRSL could be formulated as follows:

$$\mathscr{L} = \|\mathbf{R} - \mathbf{S}\|_F^2, \tag{10}$$

where $\|\cdot\|_F$ is the Frobenius norm. This loss function allows training DRSL networks with back-propagation in a DNN-based manner. Hence, the proposed model could be optimized using one stochastic gradient descent-based optimization algorithm, such as Adam [41]. The detailed optimization process of our DRSL is summarized in Algorithm 1.

Algorithm 1. Optimization procedure of DRSL

Input: The image training data $\mathscr{X}^i$, the text training data $\mathscr{X}^t$, the corresponding class label sets $\mathscr{Y}^i$ and $\mathscr{Y}^t$, the batch size N_b, and the learning rate α

1: **while** not converge **do**

2: Compute the image feature vectors $\mathscr{Z}_b^i$ for the b-th image batch $\mathscr{X}_b^i$ by the image cross-modal network according to Eq. (6) as:

$$\mathscr{Z}_b^i = f_i(\mathscr{X}_b^i)$$

3: Compute the text feature vectors $\mathscr{Z}_b^t$ for the b-th text batch $\mathscr{X}_b^t$ by the text cross-modal network according to Eq. (6) as:

$$\mathscr{Z}_b^t = f_i(\mathscr{X}_b^t)$$

4: Compute the pairwise similarity matrix $\mathbf{R}$ for the image-text batch by the relation network module according to Eq. (8) as:

$$\mathbf{R} = h(\mathscr{Z}_b^i, \mathscr{Z}_b^t).$$

5: Compute the *a priori* pairwise similarity matrix $\mathbf{S}$ for the batch image-text pairs according to Eq. (9).

(*continued on next page*)

Algorithm 1 (*continued*)

6: Update the parameters of the cross-modal networks module and relation network module by minimizing $\mathscr{L}$ in Eq. (10) with descending their stochastic gradient:

$$\Theta_i = \Theta_i - \alpha \frac{\partial \mathscr{L}}{\partial \Theta_i}$$

$$\Theta_t = \Theta_t - \alpha \frac{\partial \mathscr{L}}{\partial \Theta_t}$$

$$\Theta_r = \Theta_r - \alpha \frac{\partial \mathscr{L}}{\partial \Theta_r}$$

7: **end while**

Output: The optimized DRSL model.

3.4. Implementation details

Our DRSL is implemented by PyTorch,[2] which is a well-known machine learning library. The learning rate α is empirically set as 0.0001. The implementation details of the cross-modal network module and relation network module are introduced in the following paragraphs.

1. *Cross-modal networks module:* The cross-modal networks module consists of two convolutional networks (Image CNN and Text CNN *resp.*) for two modalities (image and text *resp.*), followed by two fully-connected neural networks (Image FNN and Text FNN *resp.*). The convolutional layers have the same configuration as the VGG-19 [38] for image subnetwork following Ref. [13]. As mentioned above, the used VGG-19 layers are pre-trained on the ImageNet and fine-tuned on the images of the training set $\mathscr{X}^i$. Similarly, the Text CNN layers have the same configuration with Ref. [40] and are pre-trained on the training text data. Then several fully-connected layers are utilized in each pathway, and each layer is followed by a ReLU activation function layer. For the reason that the size of the dataset is different, we design different models with different layer numbers according to a different dataset. For Wikipedia and NUS-WIDE-10K, the Image FNN and Text FNN are with 4096-1024-300 neurons and 300-1024-300 neurons, respectively. For Pascal Sentences, the Image FNN and Text FNN are with 4096-300 neurons and 300-300 neurons, respectively. While for XmediaNet, the two models are with 4096-1024-1024-300 neurons and 300-1024-1024-300 neurons, respectively. To stabilize the distributions of layer inputs, we also exploit batch normalization [42] for XmediaNet.
2. *Relation network module:* This module aims to compute the similarity between each image-text pair. Firstly, we use the fusion mechanism (concatenation in this paper) to obtain the pairwise fused feature vectors of image-text pairs, which will be fed into the relation network to compute the similarities. The relation network is a model with several fully-connected layers for pairwise samples, which projects each fused feature vector into the single-value predict similarity for the corresponding image-text pair, followed by a linear layer. Therefore, for image and text inputs, the cross-modal pairwise similarities can be obtained by the proposed relation network module to carry out cross-modal tasks, *e.g.*, cross-modal retrieval. For all datasets, the relation networks are with 600-1024-1 neurons.

3.5. Differences from related work

Differences from CCA [1], MCCA [43] and PLS [2]: CCA, MCCA and PLS attempt to learn some modality-specific linear transformations to project the different modalities into a latent shared space, where the distinct modalities could be highly correlated. However, the above methods are unsupervised and require that all cross-modal samples are pairwise. Some discriminative information (*e.g.*, class labels) could not be explicitly exploited to improve the performance of cross-modal retrieval. On the contrary, our DRSL is a deep supervised method and has no modality-pairwise limitation.

Differences from GMA [22], MvDA-VC [20] and JRL [23]: All of these methods attempt to learn the modality-invariant discriminative representations. Then the same predefined metric is adopted on the learned representations to calculate the similarity between image samples and text samples. More favorably, our DRSL employs the modality-specific networks and the relation module to learn the similarity between all possible pairs. Therefore, the learned similarity metric is more suitable for the unbalanced information of cross-modal data. On the other hand, these multimodal approaches are linear models that may not be suitable to capture the highly nonlinear high-level semantic features from real-world multimedia datasets. In order to tackle the nonlinear cases, the kernel trick is utilized to extend the linear models to nonlinear variants. However, how to select the predefined kernel function for a specific case is still an open problem [44]. In contrast, our DRSL could handle highly nonlinear cases with the power of deep neural networks.

Differences from CMML [45] and DCML [32]: Both of Refs. [45,32] are metric learning methods, which aims at learning a distance metric by maximizing the similarities between semantically similar cross-modal pairs and minimizing the similarities between semantically dissimilar pairs. Differently, DRSL attempts to enlarge the similarities between semantically

[2] The homepage of PyTorch is https://pytorch.org/.

similar pairs and reduce the similarities between dissimilar pairs. Moreover, these methods are also representation learning approaches as the aforementioned cross-modal methods. They need to select a suitable handcraft metric to measure the similarities between the obtained representations. Unlike them, the proposed DRSL directly computes the similarities between image-text pairs through the learned model. Therefore, the proposed method is more adaptive to cross-modal data.

Differences from DCCA [6], DCCAE [7], CCL [10] and MCSM [9]: All of these methods utilize DNN to nonlinearly learn a latent common space across distinct modalities. The methods above are cross-modal representation learning approaches that assume equal amounts of information can be acquired from different modalities during cross-modal learning. However, it is inconsistent with unbalanced information between multimodal data. Moreover, they are not specifically designed for the used handcraft metrics, thus leading to unsatisfactory performance. Differently, the pairwise similarities are directly calculated by the learned DRSL model instead of some handcraft metrics. Therefore, the learned similarities can solve the unbalance and unsatisfactory metrics for cross-nodal retrieval.

Differences from FGCrossNet [15]: FGCrossNet is a fine-grained cross-media learning method, which jointly considers three constraints (*i.e.*, classification, center, and ranking constraints) for common space learning. Whereas in the proposed DRSL, neither these constraints nor the explicit common space is needed.

Differences from GSS-SL [26] and ASFS [27]: Both of Refs. [26,27] are traditional semi-supervised cross-modal learning approaches, which combine both the label information and pairwise correlation to learn common space. Differently, the proposed DRSL is a deep supervised model that aims to directly learn the intrinsic similarities among samples from different views so that the heterogeneity gap can be effectively mitigated.

Differences from SCCM [24]: SCCM transforms different modalities into an optimal latent space gradually from easy rankings of diverse queries to more complicated ones. Although the learning to rank is similar to our relational similarity learning to some extent, the spirit of our DRSL approach is different in the respect that no latent space is explicitly learned.

4. Experiments

In this section, we evaluate the performance of our DRSL by comparing it with some state-of-the-art cross-modal retrieval approaches on four widely-used datasets. The source code and configurations are available at https://github.com/wangxu-scu/DRSL.

4.1. Datasets

To evaluate the performance of our DRSL, some experiments are conducted on four benchmark multimodal datasets, *i.e.*, Wikipedia [46], Pascal Sentence [47], NUS-WIDE-10K [48,49], and XMediaNet [9]. To comprehensively evaluate the performance of our DRSL, some comparison experiments are firstly conducted with 11 state-of-the-art approaches on four datasets to verify its effectiveness. Whereafter, some extensional experiments are used to evaluate the comprehensive performance of our DRSL.

4.1.1. Wikipedia [46]

This dataset is widely used to evaluate the performance of cross-modal retrieval. It contains 2866 image-text pairs. Each image or text sample is classified into ten semantic categories (*i.e.*, literature, media, music, etc.). Following the same data partition strategy of Ref. [49], we part Wikipedia into 3 subsets, *i.e.*, training set (with 2173 pairs), validation set (with 231 pairs), and test set (with 462 pairs).

4.1.2. Pascal sentences [47]

This dataset consists of 1,000 image-text pairs. Each image is obtained by using the 2008 PASCAL development kit, and the corresponding text sample has 5 independent sentences that are generated by distinct annotators of the Amazon Mechanical Turk annotating the image. Like Wikipedia, for a fair comparison, the image-text pairs of the dataset are randomly selected to constitute 3 sets: the training set with 800 image-text pairs (40 pairs per class), the validation set with 100 image-text pairs (5 pairs per class), and the testing set with 100 image-text pairs (5 pairs per class).

4.1.3. NUS-WIDE-10K [48,49]

This dataset is a subset of the NUS-WIDE dataset [48] provided by the authors of Ref. [49]. The authors evenly selected the image-text pairs (1000 pairs per class) from the pairs of the ten largest classes (*e.g.*, grass, person, sky, etc.) in NUS-WIDE to compose NUS-WIDE-10K. We also follow the data partition of Ref. [49] to part the dataset to 3 subsets: 8000 image-text pairs in the training set, 1000 image-text pairs in the validation set, and 1000 image-text pairs in the test set.

4.1.4. XMediaNet dataset [9]

XMediaNet is a large-scale multimedia dataset [9], which consists of 5 media types, *i.e.*, image, text, video, audio, and 3D model. Each sample of different modalities is classified in 200 independent classes. In this paper, we only focus on the bimodal case. Thus only images and texts are selected to conduct our experiments, which contains 40,000 image-text pairs. For a

fair comparison, we follow Refs. [13,9] to divide XMediaNet to 3 subsets, *i.e.*, the training set (with 32,000 image-text pairs), the validation set (with 4000 image-text pairs), and the testing set (with 4000 image-text pairs).

4.2. Evaluation criteria and compared approaches

In all experiments, to comprehensively evaluate the performance of the tested approaches, two widely-used cross-modal retrieval tasks are adopted, *viz.*, retrieving the relevant text samples by using a given image query (Image → Text), and 2) retrieving the relevant image samples by utilizing a given text query (Text → Image). As most of the previous retrieval approaches, the widely-used mean average precision (mAP) is used as an evaluation metric to investigate the performance of the tested methods, which is a standard evaluation criterion used in cross-modal retrieval. More specifically, *all returned retrieved results* are used to calculate mAP@ALL to quantify the retrieval performance in our experiments.

To evaluate the effectiveness of our DRSL, some comparison experiments are conducted with 11 state-of-the-art approaches. The compared approaches are: 1) traditional cross-modal retrieval approaches including CCA [1], MvDA [25], MvDA-VC [20], and JRL [23], and 2) deep learning-based methods including DCCA [6], DCCAE [7], ACMR [8], CMDN [12], MCSM [9], CM-GANs [13], and CCL [10]. For a fair comparison, it is notable that all the tested approaches use the same image features extracted from the fine-tuned VGG-19 and the same text features extracted from the trained sentence CNN in all the experiments.

4.3. Comparisons with the state-of-the-art

Tables 1–4 report the mAP scores of our DRSL and other tested approaches on the four benchmark datasets. Note that the results of CCL [10], CMDN [12], MCSM [9], and CM-GANs [13] are reported in original papers. From these tables, we could draw some following observations:

- It is clear that our DRSL is superior to both the traditional and DNN-based approaches on all benchmark datasets in terms of the mAP scores of two retrieval tasks and the average results. More specifically, our DRSL outperforms the best competitors by improvements of 0.005, 0.034, 0.031 and 0.053 on Wikipedia, Pascal Sentences, NUS-WIDE-10K, and XMediaNet, respectively.
- There is no doubt that the nonlinear transformations used in DNN-based approaches contribute to the improvements of performance, compared to traditional approaches, *e.g.*, DCCA surpasses CCA with distinct margins on four datasets.
- The traditional approaches such as CCA, MvDA, MvDA-VC and JRL are likely to benefit from the high-level semantic features obtained by CNN models, thus achieving favorable results.

In addition to the mAP scores, we also draw some P-R and P-S curves to visually evaluate the cross-modal retrieval performance for image-query-text and text-query-image tasks. The P-R and P-S curves are plotted in Fig. 3 on four datasets, from which one could see that our DRSL obtains the best retrieval performance on both image-query-text and text-query-image tasks. The results of P-R and P-S curves are consisted with the same observation of Tables 1–4. The scope (*i.e.*, the top R retrieved points) of the P-S curves vary from $R = 50$ to 400, $R = 10$ to 80, $R = 100$ to $R = 1000$ and $R = 500$ to 4000 on Wikipedia, Pascal Sentences, NUS-WIDE-10K and PKU XMediaNet, respectively.

Compared with the current state of the arts, the proposed DRSL achieves promising results owing to the following two facts: (1) DRSL directly establish the pairwise similarity model between different modalities, without considering the unequal amount of information across distinct modal. (2) The pairwise relational similarity outputted by relation network serves as the cross-modal retrieval metric, thereby introducing no extra error to the matching of the loss function and the retrieval metric.

4.4. Sensitivity analysis to fusion mechanism

The common fusion methods can be employed in the proposed DRSL. To investigate the impact of different fusion methods on the performance of the proposed method, we develop and evaluate four variants of DRSL: DRSL with concatenation fusion (DRSL-C), DRSL with multiplication fusion (DRSL-M), DRSL with addition fusion (DRSL-A) and DRSL with subtraction fusion (DRSL-S).

Tables 5 and 6 demonstrate the performance comparisons of the four variants of DRSL in terms of mAP scores on the Pascal Sentences and the NUS-WIDE-10K dataset. From the experimental results, we can observe that different fusion mechanisms have a certain impact on cross-modal retrieval performance. Specifically, DRSL-A and DRSL-C can obtain more satisfactory results than the others on the two datasets. On the other hand, the performance of multiplication fusion is not stable on the two datasets, *i.e.*, its performance is best on Pascal Sentence, but worst on NUS-WIDE-10K. Furthermore, all these fusion functions can achieve state-of-the-art performance. This is probably due to the fact that the nonlinear transformations of the relation network module capture the intrinsic pairwise similarities across different modalities. In future work, we will explore more effective fusion mechanisms and give an in-depth discussion.

Table 1
Cross-modal Retrieval mAP@ALL Comparison with State-of-the-Art Baselines on the Wikipedia dataset. The highest score is shown in boldface.

Method	Image → Text	Text → Image	Average
CCA [1]	0.357	0.326	0.341
MvDA [25]	0.425	0.384	0.404
MvDA-VC [20]	0.419	0.382	0.401
JRL [23]	0.478	0.436	0.457
DCCA [6]	0.452	0.411	0.431
DCCAE [7]	0.456	0.416	0.436
ACMR [8]	0.480	0.411	0.431
CMDN [12]	0.487	0.427	0.457
MCSM [9]	0.516	0.458	0.487
CM-GANs [13]	0.521	0.466	0.494
CCL [10]	0.505	0.457	0.481
FGCrossNet [15]	0.457	0.429	0.443
DRSL	**0.523**	**0.475**	**0.499**

Table 2
Cross-modal Retrieval mAP@ALL Comparison with State-of-the-Art Baselines on the Pascal Sentences dataset. The highest score is shown in boldface.

Method	Image → Text	Text → Image	Average
CCA [1]	0.225	0.227	0.226
MvDA [25]	0.592	0.622	0.607
MvDA-VC [20]	0.652	0.672	0.662
JRL [23]	0.627	0.658	0.642
DCCA [6]	0.606	0.633	0.620
DCCAE [7]	0.618	0.644	0.631
ACMR [8]	0.589	0.582	0.586
CMDN [12]	0.544	0.526	0.535
MCSM [9]	0.598	0.598	0.598
CM-GANs [13]	0.603	0.604	0.604
CCL [10]	0.576	0.561	0.569
FGCrossNet [15]	0.637	0.662	0.650
DRSL	**0.687**	**0.705**	**0.696**

Table 3
Cross-modal Retrieval mAP@ALL Comparison with State-of-the-Art Baselines on the NUS-WIDE-10K dataset. The highest score is shown in boldface.

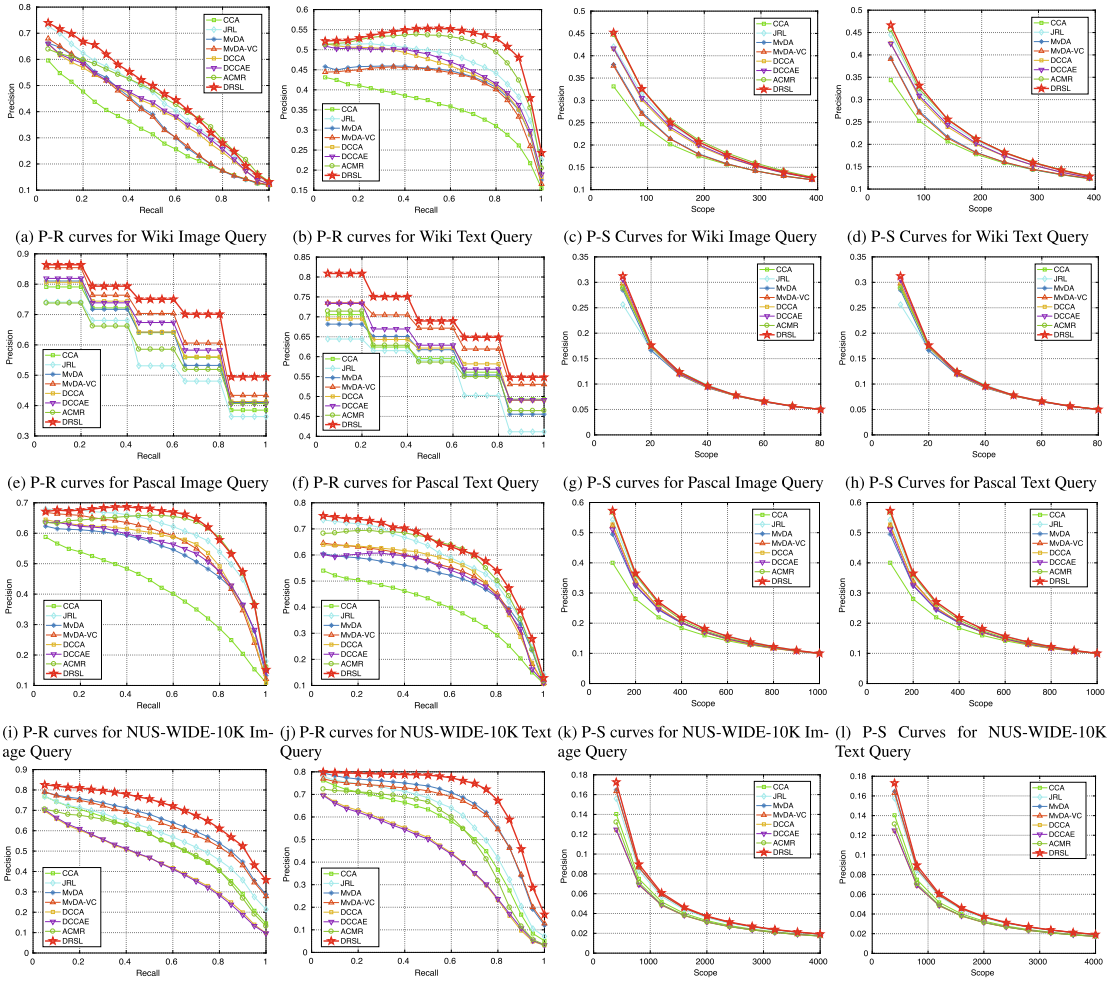
Method	Image → Text	Text → Image	Average
CCA [1]	0.398	0.414	0.406
MvDA [25]	0.501	0.526	0.513
MvDA-VC [20]	0.526	0.557	0.542
JRL [23]	0.580	0.599	0.589
DCCA [6]	0.532	0.549	0.540
DCCAE [7]	0.511	0.540	0.525
ACMR [8]	0.591	0.601	0.596
CMDN [12]	0.492	0.515	0.504
CCL [10]	0.506	0.535	0.521
FGCrossNet [15]	0.571	0.593	0.582
DRSL	**0.624**	**0.630**	**0.627**

4.5. Convergence

In this subsection, we investigate the convergence of the proposed DRSL using the XMediaNet dataset and the Wikipedia dataset. Fig. 4 shows the change trend of the objective function value *vs.* the number of epochs. This figure indicates that the objective function of DRSL becomes convergent after about 150 epochs and 100 epochs on XMediaNet and Wikipedia, respectively.

Table 4
Cross-modal Retrieval mAP@ALL Comparison with State-of-the-Art Baselines on the XMediaNet dataset. The highest score is shown in boldface.

Method	Image → Text	Text → Image	Average
CCA [1]	0.544	0.546	0.545
MvDA [25]	0.651	0.639	0.645
MvDA-VC [20]	0.650	0.627	0.638
JRL [23]	0.586	0.578	0.582
DCCA [6]	0.583	0.596	0.590
DCCAE [7]	0.594	0.606	0.600
ACMR [8]	0.639	0.639	0.639
CMDN [12]	0.485	0.516	0.501
MCSM [9]	0.540	0.550	0.545
CM-GANs [13]	0.567	0.551	0.559
CCL [10]	0.537	0.528	0.533
FGCrossNet [15]	0.629	0.633	0.631
DRSL	**0.699**	**0.698**	**0.698**



(a) P-R curves for Wiki Image Query (b) P-R curves for Wiki Text Query (c) P-S Curves for Wiki Image Query (d) P-S Curves for Wiki Text Query

(e) P-R curves for Pascal Image Query (f) P-R curves for Pascal Text Query (g) P-S curves for Pascal Image Query (h) P-S Curves for Pascal Text Query

(i) P-R curves for NUS-WIDE-10K Image Query (j) P-R curves for NUS-WIDE-10K Text Query (k) P-S curves for NUS-WIDE-10K Image Query (l) P-S Curves for NUS-WIDE-10K Text Query

(m) P-R curves for XMediaNet Image Query (n) P-R curves for XMediaNet Text Query (o) P-S curves for XMediaNet Image Query (p) P-S curves for XMediaNet Text Query

Fig. 3. Precision-recall curves for the image-query-texts and text-query-images experiments on the Wikipedia, Pascal Sentences, NUS-WIDE-10K and XMediaNet datasets.

Table 5
Cross-modal Retrieval mAP@ALL Comparison of the four variants of DRSL on the Pascal Sentences dataset. The highest score is shown in boldface.

Method	Image → Text	Text → Image	Average
DRSL-C	0.687	0.705	0.696
DRSL-M	**0.699**	0.700	**0.699**
DRSL-A	0.687	**0.710**	**0.699**
DRSL-S	0.673	0.701	0.687

Table 6
Cross-modal Retrieval mAP@ALL Comparison of the four variants of DRSL on the NUS-WIDE-10K dataset. The highest score is shown in boldface.

Method	Image → Text	Text → Image	Average
DRSL-C	**0.624**	**0.630**	**0.627**
DRSL-M	0.618	0.617	0.618
DRSL-A	0.621	0.626	0.623
DRSL-S	0.617	0.623	0.620

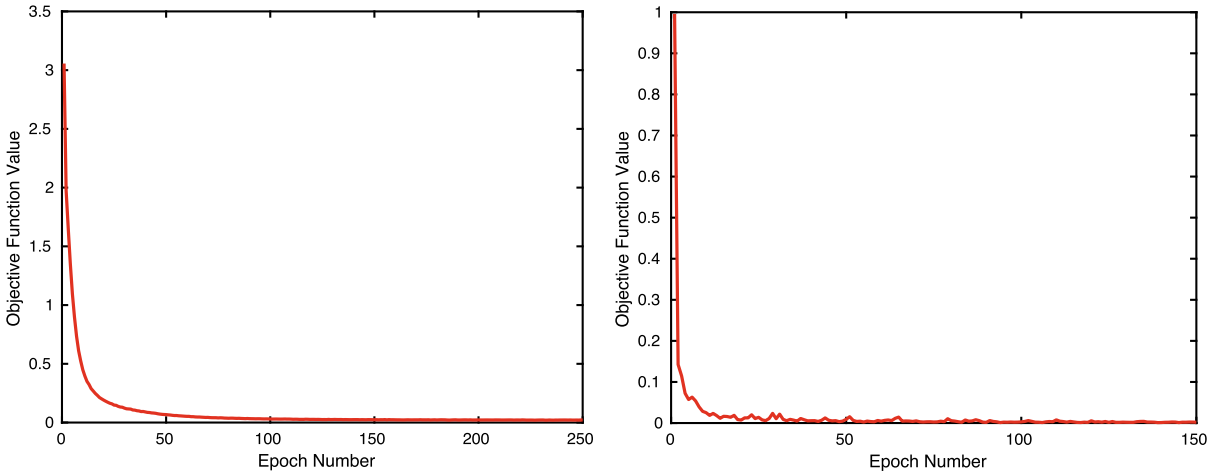


Fig. 4. The convergence curves of DRSL on two datasets: (a) XMediaNet; (b) Wikipedia.

Table 7
Training and testing time cost for the proposed DRSL on three datasets.

Dataset	Training time	Testing time
PKU XMediaNet	4m5.702 s	9.633 s
Wikipedia	17.975 s	0.483 s
Pascal Sentences	9.961 s	0.397 s

4.6. Time cost

To further investigate the time cost of the DRSL, we report the training as well as testing time on the PKU XMediaNet, Wikipedia and Pascal Sentences. We set the epochs to be 20 for each dataset. The experiments are conducted on one Geforce GTX 1080Ti graphics card and an Intel Xeon E5-2630 v4 @ 2.20 GHz CPU. The results are shown in Table 7.

5. Conclusion

This paper proposed a new method called Deep Relational Similarity Learning (DRSL) for cross-modal retrieval. The existing cross-modal methods need to learn a shared space, where pairwise similarities are computed between different modal-

ities. Different from these approaches, the proposed DRSL is supposed to directly learn a pairwise relational similarity matrix instead of explicitly learning a shared space. Hence it can refrain from the issue of unbalanced and unequal information across distinct modalities. Meanwhile, the pairwise relational similarity serves as the cross-modal retrieval metric, thereby introducing no extra error to the matching of the loss function and retrieval metric. Extensive comprehensive experiments on four public datasets show the promising performance of the proposed method in cross-modal retrieval task.

CRediT authorship contribution statement

Xu Wang: Conceptualization, Methodology, Data curation, Writing - original draft. **Peng Hu:** Methodology, Data curation, Writing - review & editing. **Liangli Zhen:** Investigation. **Dezhong Peng:** Supervision.

Declaration of Competing Interest

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

Acknowledgment

This work is supported by the National Key Research and Development Project of China under contract No. 2017YFB1002201 and partially supported by the National Natural Science Foundation of China (Grants Nos. 61971296, U19A2078, 61625204), the Ministry of Education & China Mobile Research Foundation Project (No. MCM20180405), Sichuan Science and Technology Planning Project (Nos. 2020YFH0186, 2020YFG0319, 2017GFW0097), and Scu-Luzhou Corporation Sci&Tech Research Project (No. 2019CDLZ-07). Xu Wang is grateful for the support from the program of China Scholarships Council (No. 201906240063).

References

[1] H. Hotelling, Relations between two sets of variates, Biometrika 28 (3/4) (1936) 321–377.
[2] A. Sharma, D.W. Jacobs, Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 593–600.
[3] X. Peng, J. Feng, S. Xiao, W.-Y. Yau, J.T. Zhou, S. Yang, Structured autoencoders for subspace clustering, IEEE Transactions on Image Processing 27 (10) (2018) 5076–5086.
[4] P.L. Lai, C. Fyfe, Kernel and nonlinear canonical correlation analysis, International Journal of Neural Systems 10 (05) (2000) 365–377.
[5] L. Jin, S. Li, H.M. La, X. Luo, Manipulability optimization of redundant manipulators using dynamic neural networks, IEEE Transactions on Industrial Electronics 64 (6) (2017) 4710–4720.
[6] G. Andrew, R. Arora, J. Bilmes, K. Livescu, Deep canonical correlation analysis, in: International Conference on Machine Learning, 2013, pp. 1247–1255.
[7] W. Wang, R. Arora, K. Livescu, J. Bilmes, On deep multi-view representation learning, in: International Conference on Machine Learning, 2015, pp. 1083–1092.
[8] B. Wang, Y. Yang, X. Xu, A. Hanjalic, H.T. Shen, Adversarial cross-modal retrieval, in: Proceedings of the 2017 ACM on Multimedia Conference, ACM, 2017, pp. 154–162.
[9] Y. Peng, J. Qi, Y. Yuan, Modality-specific cross-modal similarity measurement with recurrent attention network, IEEE Transactions on Image Processing 27 (11) (2018) 5585–5599.
[10] Y. Peng, J. Qi, X. Huang, Y. Yuan, CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network, IEEE Transactions on Multimedia 20 (2) (2018) 405–420.
[11] L. Zhen, P. Hu, X. Wang, D. Peng, Deep supervised cross-modal retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 10394–10403.
[12] Y. Peng, X. Huang, J. Qi, Cross-media shared representation by hierarchical learning with multiple deep networks, in: IJCAI, 2016, pp. 3846–3853.
[13] Y. Peng, J. Qi, CM-GANs: Cross-modal generative adversarial networks for common representation learning, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 15 (1) (2019) 22.
[14] P. Hu, X. Wang, L. Zhen, D. Peng, Separated variational hashing networks for cross-modal retrieval, in: Proceedings of the 27th ACM International Conference on Multimedia, ACM, 2019, pp. 1721–1729.
[15] X. He, Y. Peng, L. Xie, A new benchmark and approach for fine-grained cross-media retrieval, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 1740–1748.
[16] X. Wang, P. Hu, P. Liu, D. Peng, Deep semisupervised class-and correlation-collapsed cross-view learning, IEEE Transactions on Cybernetics. doi:10.1109/TCYB.2020.2984489.
[17] F.S.Y Yang, L. Zhang, T. Xiang, Torr P.H., Hospedales T.M., Learning to compare: Relation network for few-shot learning, in: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018.
[18] A. Santoro, D. Raposo, D.G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, T. Lillicrap, A simple neural network module for relational reasoning, in: Advances In Neural Information Processing Systems, 2017, pp. 4967–4976.
[19] H. Hu, J. Gu, Z. Zhang, J. Dai, Y. Wei, Relation networks for object detection, in: Computer Vision and Pattern Recognition (CVPR), Vol. 2, 2018.
[20] M. Kan, S. Shan, H. Zhang, S. Lao, X. Chen, Multi-view discriminant analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (1) (2016) 188–194.
[21] P. Hu, D. Peng, J. Guo, L. Zhen, Local feature based multi-view discriminant analysis, Knowledge-Based Systems 149 (2018) 34–46.
[22] A. Sharma, A. Kumar, H. Daume, D.W. Jacobs, Generalized multiview analysis: a discriminative latent space, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 2160–2167.
[23] X. Zhai, Y. Peng, J. Xiao, Learning cross-media joint representation with sparse and semisupervised regularization, IEEE Transactions on Circuits and Systems for Video Technology 24 (6) (2014) 965–978.
[24] M. Luo, X. Chang, Z. Li, L. Nie, A.G. Hauptmann, Q. Zheng, Simple to complex cross-modal learning to rank, Computer Vision and Image Understanding 163 (2017) 67–77.
[25] M. Kan, S. Shan, H. Zhang, S. Lao, X. Chen, Multi-view discriminant analysis, in: European Conference on Computer Vision, 2012, pp. 808–821.

[26] L. Zhang, B. Ma, G. Li, Q. Huang, Q. Tian, Generalized semi-supervised and structured subspace learning for cross-modal retrieval, IEEE Transactions on Multimedia 20 (1) (2018) 128–141.
[27] E. Yu, J. Sun, J. Li, X. Chang, X.-H. Han, A.G. Hauptmann, Adaptive semi-supervised feature selection for cross-modal retrieval, IEEE Transactions on Multimedia 21 (5) (2018) 1276–1288.
[28] X. Peng, J. Feng, J. Lu, W.Y. Yau, Z. Yi, Cascade subspace clustering, in: Proc. of 31th AAAI Conf. on Artif. Intell., AAAI, SFO, USA, 2017, pp. 2478–2484.
[29] Q. Wang, M. Chen, F. Nie, X. Li, Detecting coherent groups in crowd scenes by multiview clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (1) (2018) 46–58, https://doi.org/10.1109/TPAMI.2018.2875002.
[30] X. Wang, D. Peng, P. Hu, Y. Sang, Adversarial correlated autoencoder for unsupervised multi-view representation learning, Knowledge-Based Systems 168 (2019) 109–120.
[31] M. Kan, S. Shan, X. Chen, Multi-view deep network for cross-view classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4847–4855.
[32] V.E. Liong, J. Lu, Y.-P. Tan, J. Zhou, Deep coupled metric learning for cross-modal matching, IEEE Transactions on Multimedia 19 (6) (2017) 1234–1244.
[33] P. Hu, D. Peng, Y. Sang, Y. Xiang, Multi-view linear discriminant analysis network, IEEE Transactions on Image Processing 28 (11) (2019) 5352–5365.
[34] P. Hu, D. Peng, X. Wang, Y. Xiang, Multimodal adversarial network for cross-modal retrieval, Knowledge-Based Systems 180 (2019) 38–50.
[35] P. Hu, H. Zhu, X. Peng, J. Lin, Semi-supervised multi-modal learning with balanced spectral decomposition, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 99–106.
[36] D. Raposo, A. Santoro, D. Barrett, R. Pascanu, T. Lillicrap, P. Battaglia, Discovering objects and their relations from entangled scene representations, arXiv preprint arXiv:1702.05068.
[37] C. Sun, A. Shrivastava, C. Vondrick, K. Murphy, R. Sukthankar, C. Schmid, Actor-centric relation network, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 318–334.
[38] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015.
[39] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: International Conference on Neural Information Processing Systems, 2013, pp. 3111–3119.
[40] Y. Kim, Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882, 2014.
[41] D. Kinga, J.B. Adam, A method for stochastic optimization, in: International Conference on Learning Representations (ICLR), vol. 5, 2015.
[42] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML'15), 37 (2015) 448-456.
[43] J. Rupnik, J. Shawe-Taylor, Multi-view canonical correlation analysis, in: Conference on Data Mining and Data Warehouses 2010, SiKDD, 2010, pp. 1–4.
[44] X. Peng, S. Xiao, J. Feng, W. Yau, Z. Yi, Deep subspace clustering with sparsity prior, in: Proceedings of the 25 International Joint Conference on Artificial Intelligence, New York, NY, USA, 2016, pp. 1925–1931.
[45] A. Mignon, F. Jurie, CMML: a new metric learning approach for cross modal matching, in: Asian Conference on Computer Vision, 2012.
[46] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: Proceedings of the 18th ACM International Conference on Multimedia, ACM, 2010, pp. 251–260.
[47] C. Rashtchian, P. Young, M. Hodosh, J. Hockenmaier, Collecting image annotations using amazon's mechanical turk, in: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Association for Computational Linguistics (2010) 139–147.
[48] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, NUS-WIDE a real-world web image database from national university of Singapore, in: Proceedings of the ACM international conference on image and video retrieval, ACM, 2009, p. 48.
[49] F. Feng, X. Wang, R. Li, Cross-modal retrieval with correspondence autoencoder, in: Proceedings of the 22nd ACM international conference on Multimedia, ACM, 2014, pp. 7–16.