Deep Supervised Cross-modal Retrieval

Liangli Zhen* Peng Hu* Xu Wang Dezhong Peng†
Machine Intelligence Laboratory, College of Computer Science
Sichuan University, Chengdu, 610065, China
llzhen@outlook.com, {penghu.ml, wangxu.scu}@gmail.com, pengdz@scu.edu.cn

Abstract

Cross-modal retrieval aims to enable flexible retrieval across different modalities. The core of cross-modal retrieval is how to measure the content similarity between different types of data. In this paper, we present a novel cross-modal retrieval method, called Deep Supervised Cross-modal Retrieval (DSCMR). It aims to find a common representation space, in which the samples from different modalities can be compared directly. Specifically, DSCMR minimises the discrimination loss in both the label space and the common representation space to supervise the model learning discriminative features. Furthermore, it simultaneously minimises the modality invariance loss and uses a weight sharing strategy to eliminate the cross-modal discrepancy of multimedia data in the common representation space to learn modality-invariant features. Comprehensive experimental results on four widely-used benchmark datasets demonstrate that the proposed method is effective in cross-modal learning and significantly outperforms the state-of-the-art cross-modal retrieval methods.

1. Introduction

Cross-modal retrieval aims to enable flexible retrieval across different modalities (e.g., texts vs. images) [30]. It takes one type of data as the query to retrieve relevant data of another type. The provided search results across various modalities can be helpful to the users to obtain comprehensive information about the target events or topics. With the rapid growth of different types of media data such as texts, images, and videos on the Internet, cross-modal retrieval becomes increasingly important in real-world applications [32]. Recently, cross-modal retrieval has attracted the considerable attention of the researchers from both academia and industry. The challenge of cross-modal retrieval is how to measure the content similarity between different types of data since they , which is referred to as the heterogeneity gap [32].

A common approach to bridge the heterogeneity gap is representation learning. It tries to find a function to transform the data samples from different modalities into a common representation space in which the similarity between them can be measured directly. A variety of cross-modal retrieval methods [20] have been developed, which propose different learning ways for finding the common space. The traditional ones use the statistical correlation analysis to learn linear projections by optimising target statistical values. For example, Canonical Correlation Analysis (CCA) [8] is one of the most representative works, which learns the common space by maximising the pairwise correlations between two sets of heterogeneous data. However, the correlation of multimedia data in the real world is too complex to be fully modelled only by applying linear projections. Then, some kernel-based methods [1, 34] have been developed to address this issue, but how to select a suitable kernel function for particular cross-modal learning application is still an unsolved problem.

Inspired by the great success of deep neural networks in representation learning [14], a large number of deep learning-based approaches [2, 33, 19, 36, 21, 25, 7] have been proposed to learn a common presentation space for multimedia data. For instance, Ngiam *et al.* [18] propose a bimodal deep auto-encoder to learn the cross-modal correlation as well as preserve the reconstruction information and apply a Restricted Boltzmann Machine (RBM) to learn the common space for cross-modal retrieval. Different from [18], which learns common representations in an unsupervised way, some supervised deep cross-modal learning approaches have been proposed to learn more discriminative representations. They are potentially able to provide a much better separation between classes in the common representation space. In this class of methods, Jiang *et al.* [9] propose to use the label information to learn the discriminative information between samples from inter-modalities. In addition, the cross-modal similarity is preserved by enforcing the representations of each image-text pair to be close to

*First two authors contributed equally to this work.
†D. Peng is the corresponding author.

each other in a common Hamming space. In [35], Wang *et al.* propose a Multi-modal Deep Neural Network (MDNN) based on a deep Convolutional Neural Network (CNN) and a Neural Language Model (NLM) to learn mapping functions for the image modality and the text modality, respectively. The (labels of the samples) classification information is used to learn intra-modal semantics for image and text. The Euclidean distance is used to measure the difference between the representations for an image-text pair to guide the cross-modal learning. In [30], the classification information is also used to learn intra-modal discrimination in data during the feature projection. It is notable that even though the classification information has been used in these approaches, the classification information is only used to learn discriminative features within each modality or between intermodalities. The semantic information is not fully exploited in these cross-modal learning approaches.

In this paper, we present a novel cross-modal retrieval method, called Deep Supervised Cross-modal Retrieval (DSCMR). It aims to preserve the discrimination among the samples from different semantic categories and eliminate the cross-modal discrepancy as well. To achieve this goal, it minimises the discrimination loss of the samples both in the label space and the common representation space to supervise our model learning discriminative features. Furthermore, it simultaneously minimises the modality invariance loss and uses a weight sharing strategy to learn modality-invariant features in the common representation space. Following this learning strategy, both the pairwise label information and the classification information are as fully exploited as possible to ensure the learned representation to be both discriminative in semantic structure and invariant across modalities.

The main contributions of this work can be summarised as follows:

- A deep supervised cross-modal learning architecture is proposed to bridge the heterogeneity gap between different modalities. It can effectively learn the common representations for the heterogeneous data by preserving the semantic discrimination and modality invariance simultaneously in an end-to-end manner.

- Two sub-networks with weight sharing constraint at the last layers are developed to learn the cross-modal correlation between image and text modalities. Furthermore, the modality-invariance loss is directly formulated into the objective function to eliminate the cross-modal discrepancy.

- A linear classifier is applied to classify the samples in the common representation space. In this way, DSCMR minimises the discrimination loss in both the label space and the common representation space, which makes the learned common representations be significantly discriminative.

- Extensive experiments on widely-used benchmark datasets have been conducted. The results demonstrate that our method outperforms current state-of-the-art methods for cross-modal retrieval, which indicates the effectiveness of the proposed method.

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

2. Related Work

The cross-modal learning methods aim to learn a common representation space, where the similarity between the samples from different modalities can be measured directly. A variety of approaches have been proposed to learn such a common representation space, which can be roughly divided into two categories: 1) binary-valued representation learning [9, 3, 41], also called as cross-modal hashing, and 2) real-valued representation leaning [30, 19, 21]. The binary-valued approaches are more geared towards computational efficiency and map the heterogeneous data into a common Hamming space, in which the cross-modal retrieval would be fast. Since the representations are encoded to binary codes, the retrieval accuracy generally decreases slightly due to the loss of information [20].

The proposed method in this paper is the one in the category of real-valued representation learning approaches. This category includes unsupervised approaches [2, 5, 33], pairwise approaches [38, 39, 31] and supervised approaches [32, 28]. The unsupervised methods only use co-occurrence information (co-exist in a multimedia document) to learn common representations for different types of data. The methods of CCA, Deep CCA (DCCA) [2], Correspondence Auto-encoder (Corr-AE) [5] and Deep Canonically Correlated Auto-encoder (DCCAE) [33] are representative ones of this subclass. The pairwise-based methods utilise more similar pairs to learn a meaning metric for comparing samples from different modalities. The representative methods of this subclass include the Multi-view Metric Learning with Global consistency and Local smoothness (MVML-GL) method [38], the Joint Graph Regularised Heterogeneous Metric Learning (JGRHML) method [39] and the Modality-Specific Deep Structure (MSDS) method[31].

To learn more discriminative common representations, supervised methods exploit label information to distinguish the samples from different semantic categories. The

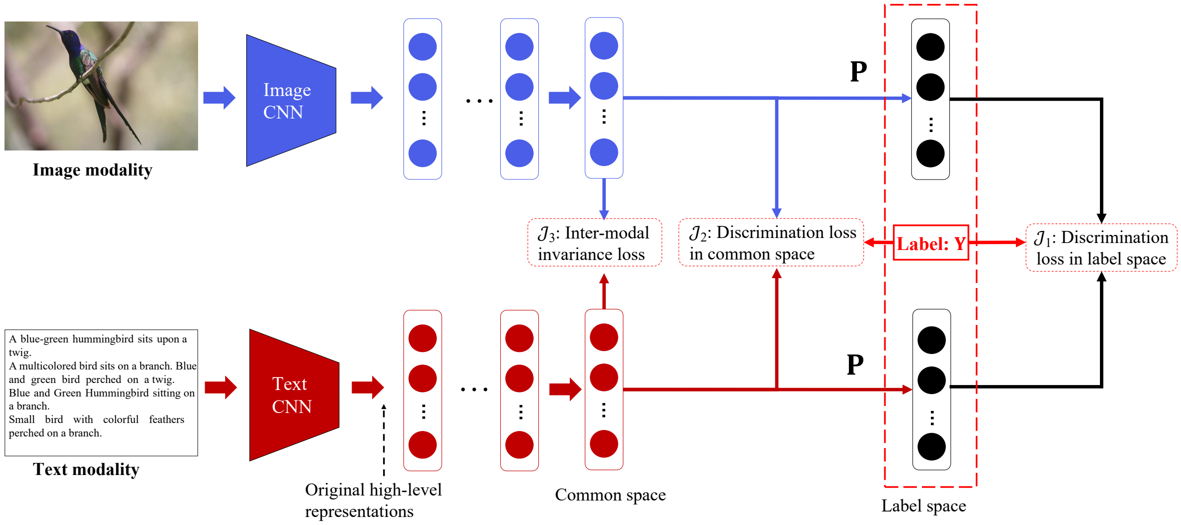


Image modality
Image CNN
P
$\mathcal{J}_3$: Inter-modal invariance loss
$\mathcal{J}_2$: Discrimination loss in common space
Label: Y
$\mathcal{J}_1$: Discrimination loss in label space
A blue-green hummingbird sits upon a twig.
A multicolored bird sits on a branch. Blue and green bird perched on a twig.
Blue and Green Hummingbird sitting on a branch.
Small bird with colorful feathers perched on a branch.
Text modality
Text CNN
P
Original high-level representations
Common space
Label space

Figure 1. The general framework of the proposed DSCMR method. The images and the text are inputed into an image CNN [13] and a text CNN [37], respectively, to obtain original high-level semantic representations. Then, a number of fully-connected layers are separately added on the top of them to map the samples from different modalities into a common representation space. Finally, a linear classifier (with parameters in $\mathbf{P}$) is used to predict the category of each sample to supervise the network to learn cross-modal transformation functions $f(\cdot)$ and $g(\cdot)$.

supervised methods enforce different-category samples to be transformed far apart while the same-category samples lie as close as possible. To obtain such a common space, Sharama *et al.* [28] proposed a supervised extension of CCA, named as Generalised Multi-view Analysis (GMA), by using the semantic category labels to guide the learning of common representations. The recently proposed methods in [9], [22] and [30] also exploited semantic category labels to learn discriminative features for cross-modal retrieval. In [22] and [30], the adversarial learning [6] has been employed to improve the performance of cross-modal learning as well. They both have achieved promising performance on cross-modal retrieval tasks.

This paper is dedicated to fully exploit the classification information to guide the model learning more discriminative and modal-invariant representations for the data of different types and bridge the heterogeneity gap, and thus improving the cross-modal retrieval accuracy.

3. The Proposed Method

In this section, we first introduce the formulation of the cross-modal retrieval problem. Then, we present the proposed method to learn the common presentations of data from different modalities. At last, we provide more implementation details of the proposed method.

3.1. Problem Formulation

Without losing generality, we focus on cross-modal retrieval for bimodal data, *i.e.*, for images and text. We assume that there is a collection of n instances of image-text pairs, denoted as $\Psi = \{(\mathbf{x}_i^\alpha, \mathbf{x}_i^\beta)\}_{i=1}^n$, where $\mathbf{x}_i^\alpha$ is the input image sample and $\mathbf{x}_i^\beta$ is the input text sample of the ith instance. Each pair of instances $(\mathbf{x}_i^\alpha, \mathbf{x}_i^\beta)$ has been assigned a semantic label vector $\mathbf{y}_i = [y_{1i}, y_{2i}, \ldots, y_{ci}] \in \mathbb{R}^c$, where c is the number of the categories. If the ith instance belongs to the jth category, $y_{ji} = 1$, otherwise $y_{ji} = 0$.

Since the image feature vectors and text feature vectors typically have different statistical properties and lie in different representation spaces, they cannot be directly compared against each other for cross-modal retrieval [30]. Cross-modal learning is to learn two functions for two modalities: $\mathbf{u}_i = f(\mathbf{x}_i^\alpha; \Upsilon_\alpha) \in \mathbb{R}^d$ for the image modality and $\mathbf{v}_j = g(\mathbf{x}_j^\beta; \Upsilon_\beta) \in \mathbb{R}^d$ for the text modality, where d is the dimensionality of the representation in the common representation space, and Υ_α and Υ_β are the trainable parameters of the two functions. It makes the samples can be compared directly even though they come from different modalities, and in the common space, the similarity of the samples from the same category is larger than the similarity of the samples from the different categories. Therefore, the relevant samples of different data types in the data set can be returned for one query of any data type. In the following, the image representation matrix, the text representation matrix and the label matrix for all instances in Ψ are denoted as $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n]$, $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$ and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]$ with $\mathbf{u}_i$ be the learned image representation for the ith instance and $\mathbf{v}_j$ be the learned text representation for the jth instance in the common representation space.

3.2. Framework of DSCMR

The general framework of the proposed method is shown in Figure 1, from which we can see that it includes two sub-networks, one for image modality and another for text modality, and they are trained in an end-to-end manner. The convolutional layers of the deep neural network for image modality are the same as those in 19-layer VGGNet [29], which is pre-trained on the ImageNet. We generate $4,096$-dimensional feature vector from fc7 layer as the original high-level semantic representation for image, denoted as h_i^{α}. Then, several fully-connected layers conduct the common representation learning to obtain the common representation for each image, denoted as $\mathbf{u}_i$. To perform common representation learning for text, we employ the Word2Vec model [16], which is pre-trained on billions of words in Google News, to represent each network as a k-dimensional feature vector first. Thus, each text can be represented as a matrix with each column as a k-dimensional feature vector. Then, the text matrix is feed to the convolutional layers as same the configuration as sentence CNN [37] to generate the original high-level semantic representation for text, denoted as h_i^{β}. In a similar way, a number of fully-connected layers are followed to learn the common representation for text, denoted as $\mathbf{v}_i$. To ensure the two sub-networks to learn a common representation space for image and text modalities, we enforce these two sub-networks to share the weights of their last layers. This is intuitively to generate as similar as possible representations for the image and text samples from the same category.

Finally, based on the assumption that the common representations in the common space are ideal for classification, a linear classifier with the parameter matrix $\mathbf{P}$ is connected to these two sub-networks to learn discriminative features by exploiting the label information. Therefore, the cross-modal correlation could be well learned and the discriminative features can be simultaneously exacted.

3.3. Objective Function

The goal of DSCMR is to learn the semantic structure of the data, *i.e.*, to learn a common space where the samples from the same semantic category should be similar, even though these data may come from different modalities, and the samples from different semantic categories should be dissimilar. To learn discriminative features of the multimedia data, we propose to minimise the discrimination loss in both the label space and the common representation space. Simultaneously, we minimise the distance between the representations of each image-text pair to reduce the cross-modal discrepancy as well. In the following, we present more details about the objective function of our DSCMR.

To preserve the discrimination of samples from different categories after the feature projection, we assume that the common representations are ideal for classification and use a linear classifier to predict the semantic labels of the samples projected in the common representation space. Specifically, a linear layer is connected on the top of the image modal network and the text modal network. This classifier takes the representations of the training data in the common space and generates a predicted label of c-dimensional vector for each sample. We propose the following objective function to measure the discrimination loss in the label space:

$$\mathcal{J}_1 = \frac{1}{n}\|\mathbf{P}^T\mathbf{U} - \mathbf{Y}\|_F + \frac{1}{n}\|\mathbf{P}^T\mathbf{V} - \mathbf{Y}\|_F, \qquad (1)$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $\mathbf{P}$ is the projection matrix of the linear classifier.

Furthermore, we also measure the discrimination loss of all samples from both modalities in the common representation space directly:

$$\begin{aligned}\mathcal{J}_2 = & \underbrace{\frac{1}{n^2}\sum_{i,j=1}^{n}(log(1+e^{\Gamma_{ij}}) - S_{ij}^{\alpha\beta}\Gamma_{ij})}_{\text{inter-modalities}} \\ & + \underbrace{\frac{1}{n^2}\sum_{i,j=1}^{n}(log(1+e^{\Phi_{ij}}) - S_{ij}^{\alpha\alpha}\Phi_{ij})}_{\text{image modality}} \\ & + \underbrace{\frac{1}{n^2}\sum_{i,j=1}^{n}log(1+e^{\Theta_{ij}}) - S_{ij}^{\beta\beta}\Theta_{ij}),}_{\text{text modality}}\end{aligned} \qquad (2)$$

where $\Gamma_{ij} = \frac{1}{2}\cos(\mathbf{u}_i, \mathbf{v}_j)$, $\Phi_{ij} = \frac{1}{2}\cos(\mathbf{u}_i, \mathbf{u}_j)$, $\Theta_{ij} = \frac{1}{2}\cos(\mathbf{v}_i, \mathbf{v}_j)$, $S_{ij}^{\alpha\beta} = \mathbf{1}\{\mathbf{u}_i, \mathbf{v}_j\}$, $S_{ij}^{\alpha\alpha} = \mathbf{1}\{\mathbf{u}_i, \mathbf{u}_j\}$, $S_{ij}^{\beta\beta} = \mathbf{1}\{\mathbf{v}_i, \mathbf{v}_j\}$, $\cos(\cdot)$ is the cosine function used to compute the similarity between two input vectors, and $\mathbf{1}\{\cdot\}$ is an indicator function, whose value is 1 if the two elements are the representations of intra-class samples, otherwise 0. The first term of Equation (2) is the negative log likelihood of the inter-modal sample similarities with the likelihood function defined as follows:

$$p(S_{ij}^{\alpha\beta}|\mathbf{u}_i, \mathbf{v}_j) = \begin{cases} \delta(\Gamma_{ij}), & \text{if } S_{ij}^{\alpha\beta} = 1; \\ 1 - \delta(\Gamma_{ij}), & \text{otherwise,} \end{cases} \qquad (3)$$

where $\delta(\Gamma_{ij}) = \frac{1}{1+e^{-\Gamma_{ij}}}$ is the sigmoid function. It is easy to find that minimising this negative log likelihood function is equivalent to maximising the likelihood. We can also see that, the larger the similarity (cosine similarity $\cos(\mathbf{u}_i, \mathbf{v}_j)$) is, the larger $p(1|\mathbf{u}_i, \mathbf{v}_j)$ will be, which implies that should be classified as similar, and vice versa. Likely, the second and the third terms measure the similarities of the image samples and the text samples, respectively. Therefore, Equation (2) is a reasonable similarity measure for common representations and is a well criterion for learning discriminative features.

To eliminate the cross-modal discrepancy, we propose to minimise the distance between the representations of all image-text pairs. Technically, we formulate the modality invariance loss as follows:

$$\mathcal{J}_3 = \frac{1}{n}\|\mathbf{U} - \mathbf{V}\|_F. \tag{4}$$

Combining Equations (1), (2) and (4), we obtain the objective function of the proposed method DSCMR as:

$$\mathcal{J} = \mathcal{J}_1 + \lambda\mathcal{J}_2 + \eta\mathcal{J}_3, \tag{5}$$

where the hyper-parameters λ and η control the contributions of the last two components, and n is the number of the input instances. The objective function of DSCMR in Equation (5) can be optimised using a stochastic gradient descent optimisation algorithm [12]. The details of the optimisation procedure are summarised in Algorithm 1.

Algorithm 1 Optimisation procedure of the proposed DSCMR

```
Input: The training data set $\Psi = \{(\mathbf{x}_i^\alpha, \mathbf{x}_i^\beta)\}_{i=1}^n$, the label matrix $\mathbf{Y}$, the dimensionality of the common representation space $d$, the batch size $n_b$, the learning rate $\tau$, the maximal number of epochs $\aleph$, and the hyper parameters $\lambda$ and $\eta$.
Output: The optimised parameters in the two sub-networks $\Upsilon_\alpha$, $\Upsilon_\beta$.
Randomly initialise the parameters of the two sub-networks $\Upsilon_\alpha$, $\Upsilon_\beta$ and the parameters of the linear classifier $\mathbf{P}$.
for $t = 1, 2, \ldots, \aleph$ do
  for $\ell = 1, 2, \ldots, \lfloor \frac{n}{n_b} \rfloor$ do
    Randomly sample $n_b$ image-text pair samples from $\Psi$ to construct a mini-batch.
    Compute the representations $\mathbf{u}_i$ and $\mathbf{v}_j$ for the samples in the mini-batch by forward-propagation.
    Calculate the result of the objective function in Equation (5).
    Update the parameters of the linear classifier $\mathbf{P}$ by minimising $\mathcal{J}$ in Equation (5) with:
      $\mathbf{P} = (\mathbf{U}\mathbf{U}^T)^{-1}\mathbf{U}^T\mathbf{Y} + (\mathbf{V}\mathbf{V}^T)^{-1}\mathbf{V}^T\mathbf{Y}$.
    Update the parameters of the sub-networks, $\Upsilon_\alpha$ and $\Upsilon_\beta$, by minimising $\mathcal{J}$ in Equation (5) with descending their stochastic gradient:
      $\Upsilon_\alpha = \Upsilon_\alpha - \tau\frac{\partial\mathcal{J}}{\partial\Upsilon_\alpha}$; $\Upsilon_\beta = \Upsilon_\beta - \tau\frac{\partial\mathcal{J}}{\partial\Upsilon_\beta}$.
  end for
end for
```

3.4. Implementation Details

In this work, there are two sub-networks, one for image modality and the other for text modality. The convolutional layers have the same configuration with 19-layer VGGNet [29] for image sub-network and the sentence CNN [37] for text sub-network as mentioned in Section 3.2. Then two fully-connected layers with Rectified Linear Unit (ReLU) [17] active function are followed in each subnetwork. The numbers of the hidden units for the two layers are $2,048$ and $1,024$, respectively. The weights of the second fully-layers of the two sub-networks are shared to learn the correlation of two different modalities.

The entire network is trained on a Nvidia GTX 1080 Ti GPU in PyTorch. For training, we employ the ADAM [12] optimiser with a learning rate of 10^{-4} and set the maximal number of epochs as 500.

4. Experiments

To verify the effectiveness of the proposed method, we conduct experiments on four widely-used benchmark datasets: the Wikipedia dataset [24], the Pascal Sentence dataset [26], the NUS-WIDE-10k dataset [4] and the X-MediaNet dataset [20, 23]. In the experiments, we firstly compare the proposed DSCMR method with the state-of-the-art methods to evaluate its performance. Then, we provide further analysis of the DSCMR method. It includes the convergency investigation, the visualisation of the learned representation in the common representation space and the impact of different components in Equation (5).

4.1. Datasets and Features

In our experiments, we follow the dataset partition and feature exaction strategies from [22, 25]. We adopt a 19-layer VGGNet [29] to learn the representations of the samples and obtain a $4,096$-dimensional representation vector outputted by the fc7 layer of the VGGNet for each image. For representing text samples, we use the sentence CNN [37] to learn a 300-dimensional representation vector for each text. The statistical results of the three datasets are summarised in Table 1. It is notable that all the compared methods adopt the same CNN features as for both image and text obtained by the CNN architectures used in our method.

4.2. Evaluation Metric

We evaluate the compared methods by using the mean Average Precision (mAP) score for all returned results with cosine similarity on all the four datasets. The mAP metric jointly considers the ranking information and precision, which is a widely-used performance evaluation criterion in the research on cross-modal retrieval [32, 19, 30]. In our experiments, we report the mAP scores of the compared

Table 1. Statistical results of the four benchmark datasets used in our experiments, where n_{train} and n_{test} stand for the numbers of training and test image-text pairs, respectively. The symbol c is the number of categories, d_i and d_t are the dimensionalities of the image and text features obtained by VGGNet [29] and sentence CNN [37], respectively.

Dataset	n_{train}	n_{test}	c	d_i	d_t
Wikipedia	2,173	462	10	4,096	300
Pascal Sentence	800	100	20	4,096	300
NUS-WIDE-10k	8,000	1,000	10	4,096	300
XMediaNet	32,000	4,000	200	4,096	300

methods for two different cross-modal retrieval tasks: 1) retrieving text samples using image queries (Image2Text) and 2) retrieving images using text queries (Text2Image).

4.3. Comparison with State-of-the-art Methods

To verify the effectiveness of our proposed methods, we compare the proposed method with ten state-of-the-art methods in the experiments, including five traditional methods, namely CCA [8], MCCA [27], MvDA [10], MvDA-VC [11] and JRL [40], as well as five deep learning-based methods, namely CMDN [19], CCL [21], DCCA [2], DCCAE [33] and ACMR [30].

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

Method	Image2Text	Text2Image	Average
CCA [8]	0.134	0.133	0.134
MCCA [27]	0.341	0.307	0.324
MvDA [10]	0.337	0.308	0.323
MvDA-VC [11]	0.388	0.358	0.373
JRL [40]	0.449	0.418	0.434
CMDN [19]	0.487	0.427	0.457
CCL [21]	0.504	0.457	0.481
DCCA [2]	0.444	0.396	0.420
DCCAE [33]	0.435	0.385	0.410
ACMR [30]	0.477	0.434	0.456
Ours	**0.521**	**0.478**	**0.499**

Tables 2-5 report the mAP scores of the proposed DSCMR and the compared methods on the four benchmark datasets (the mAP score results of CCL [21] and CMDN [19] are provided by their authors), from which we have the following observations:

- DSCMR significantly outperforms both the traditional peer methods and the deep learning-based methods on all of the four datasets. Specifically, DSCMR outperforms the second-best methods with an improvement of $0.018, 0.038, 0.020$ and 0.050 in terms of average mAP scores on the Wikipedia, Pascal Sentence, NUS-WIDE-10k and XMediaNet datasets, respectively.

Table 3. Performance comparison in terms of mAP scores on the Pascal Sentence dataset. The highest score is shown in boldface.

Method	Image2Text	Text2Image	Average
CCA [8]	0.225	0.227	0.226
MCCA [27]	0.664	0.689	0.677
MvDA [10]	0.594	0.626	0.610
MvDA-VC [11]	0.648	0.673	0.661
JRL [40]	0.527	0.534	0.531
CMDN [19]	0.544	0.526	0.535
CCL [21]	0.576	0.561	0.569
DCCA [2]	0.678	0.677	0.678
DCCAE [33]	0.680	0.671	0.675
ACMR [30]	0.671	0.676	0.673
Ours	**0.710**	**0.722**	**0.716**

Table 4. Performance comparison in terms of mAP scores on the NUS-WIDE-10K dataset. The highest score is shown in boldface.

Method	Image2Text	Text2Image	Average
CCA [8]	0.378	0.394	0.386
MCCA [27]	0.448	0.462	0.455
MvDA [10]	0.501	0.526	0.513
MvDA-VC [11]	0.526	0.557	0.542
JRL [40]	0.586	0.598	0.592
CMDN [19]	0.492	0.515	0.504
CCL [21]	0.506	0.535	0.521
DCCA [2]	0.532	0.549	0.540
DCCAE [33]	0.511	0.540	0.525
ACMR [30]	0.588	0.599	0.593
Ours	**0.611**	**0.615**	**0.613**

Table 5. Performance comparison in terms of mAP scores on the XMEDIANET dataset. The highest score is shown in boldface.

Method	Image2Text	Text2Image	Average
CCA [8]	0.598	0.595	0.597
MCCA [27]	0.620	0.616	0.618
MvDA [10]	0.651	0.639	0.645
MvDA-VC [11]	0.650	0.627	0.638
JRL [40]	0.586	0.578	0.582
CMDN [19]	0.485	0.516	0.501
CCL [21]	0.537	0.528	0.533
DCCA [2]	0.583	0.596	0.590
DCCAE [33]	0.594	0.606	0.600
ACMR [30]	0.639	0.639	0.639
Ours	**0.697**	**0.693**	**0.695**

- The nonlinear transformations in the deep learning-based methods can be helpful to improve the performance of the traditional methods, e.g., DCCA outperforms CCA with a significant margin on the first three datasets.

- The traditional methods with the deep features could also potentially be able to achieve a high mAP score on cross-modal retrieval. For example, the linear methods CCA, MCCA, MvDA, MvDA-VC and JRL obtained promising results (average mAP of 0.597, 0.618, 0.645, 0.638 and 0.582) on the XMediaNet dataset. This may be contributed to that the image CNN and the text CNN have transformed the input image and text samples into approximately linear subspaces, which significantly reduced the difficulty of the original cross-modal learning task.

4.4. Further Analysis on DSCMR

4.4.1 Convergency

Figure 2 shows the value of the objective function of our method versus the different number of training epochs on the Pascal Sentence dataset. From the result, we can see that during the entire training procedure, the value of the objective function decreases almost monotonously and converges smoothly. The value of the objective function of DSCMR becomes stable after 500 epochs, which illustrates that the proposed method can be efficiently trained by using the stochastic gradient descent optimisation algorithm Adam [12].

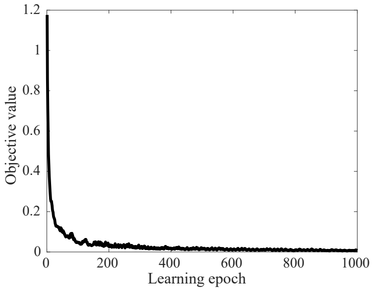


Figure 2. The value of the objective function of DSCMR versus the different number of training epochs on the Pascal Sentence dataset.

4.4.2 Visualisation of the Learned Representation

To visually investigate the effectiveness of the proposed DSCMR, we adopt the t-SNE approach to embed the representations of the image and text samples (in the common representation space) into a two-dimensional visualisation plane. The results of the original images represented by the $4,096$-dimensional (VGGNet [29]) features and the text samples represented by the 300-dimensional (sentence CNN [37]) features (after the embedding process) are displayed in Figure 3(d) and Figure 3(e), respectively. We can see that the distributions of the image modality and the text modality in the Wikipedia dataset are largely different and the samples are hard to be classified in the original input space.

Figure 3(a) and Figure 3(b) show the two-dimensional distributions of the image and text representations in the common space. From the results, we can see that the formulation of the discrimination loss in both the common space and the label space is able to model the discrimination between the samples from different semantic categories, and effectively separates the representations into several semantically discriminative clusters. We can also find that a small number of the representations from different semantic categories are mixed together, which makes DSCMR returns some irrelevant results for a query. These results are in accordance with the retrieval results shown in Table 2. Furthermore, the distributions of image modality and text modality in Figure 3(c) are well mixed together and are difficult to be separated from each other. It means that the cross-modal discrepancy is largely reduced by using the proposed method.

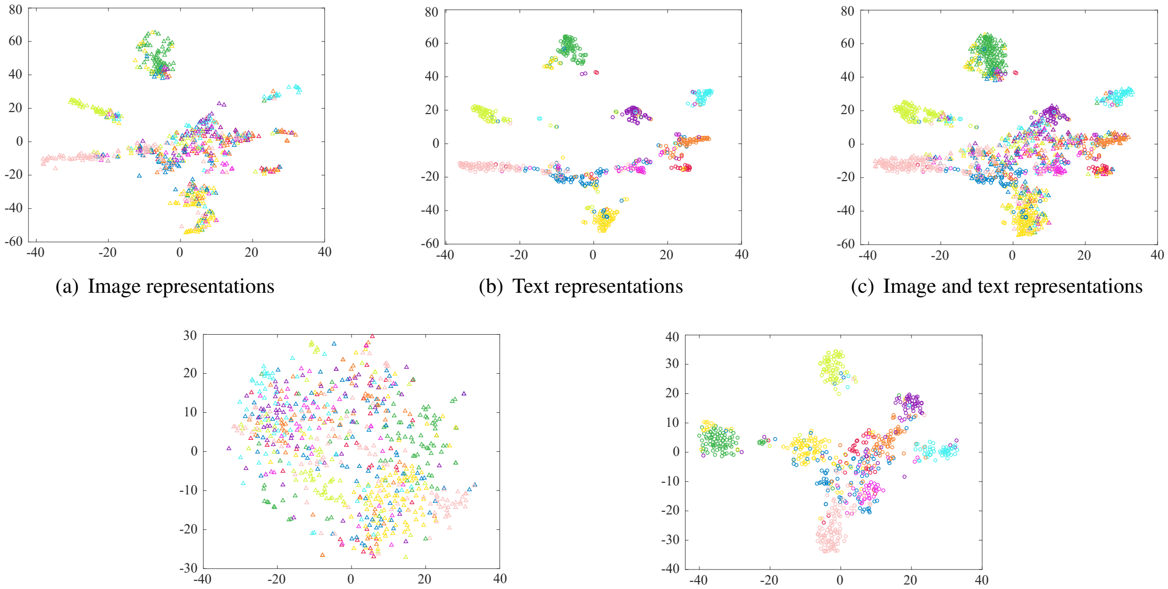
4.4.3 Impact of Different Components

The objective function of the proposed DSCMR combines three terms, which aim to minimise the discrimination loss in the label space, the discrimination loss in the common representation space, and the modality invariance loss in the common representation space, respectively. To investigate the impact of these terms on the performance of the proposed method, we developed and evaluated four variations of DSCMR: DSCMR without $\mathcal{J}_1$ (DSCMR1), DSCMR without $\mathcal{J}_2$ (DSCMR2), DSCMR without $\mathcal{J}_3$ (DSCMR3) and DSCMR only with $\mathcal{J}_1$ (DSCMR4). The optimisation procedure of these four cases is similar to the proposed DSCMR.

Table 6. Performance comparison of the proposed DSCMR and its four variations in terms of mAP scores on the Pascal Sentence dataset. The highest score is shown in boldface.

Method	Image2Text	Text2Image	Average
DSCMR1	0.583	0.631	0.607
DSCMR2	0.708	0.722	0.715
DSCMR3	0.691	0.683	0.694
DSCMR4	0.690	0.680	0.685
Full DSCMR	**0.710**	**0.722**	**0.716**

Table 6 and Table 7 show the performance comparisons of DSCMR and its four variations on the Pascal Sentence dataset and the NUS-WIDE-10K dataset. From the results, we can see that the full DSCMR performs best on both datasets, which indicates that all of the three terms in the



(a) Image representations

(b) Text representations

(c) Image and text representations

(d) Original image samples

(e) Original text samples

Figure 3. The visualisation for the test data in the Wikipedia dataset by using the t-SNE method [15]. The triangles denote the samples from image modality and the circles denote the samples from text modality. The samples come the same semantic category are marked with the same colour. (a) the image representations in the common representation space. (b) the text representations in the common representation space. (c) the image and text representations in the common representation space. (d) the original image samples represented by the $4,096$-dimensional (VGGNet [29]) features. (e) the original text samples represented by the 300-dimensional (sentence CNN [37]) features.

Table 7. Performance comparison of the proposed DSCMR and its four variations in terms of mAP scores on the NUS-WIDE-10K dataset. The highest score is shown in boldface.

Method	Image2Text	Text2Image	Average
DSCMR1	0.267	0.262	0.265
DSCMR2	0.610	0.612	0.611
DSCMR3	0.534	0.541	0.538
DSCMR4	0.527	0.520	0.524
Full DSCMR	**0.611**	**0.615**	**0.613**

objective function contribute to the final retrieval accuracy. We can also see that DSCMR outperforms DSCMR1 with a large margin, which demonstrates the importance of the first term (the discrimination loss in the label space). Furthermore, DSCMR4 (the variation only with the first term) obtained competitive results on both datasets. This also indicates that the importance of the first term for the model to learn modal-invariant discriminative features. However, DSCMR4 is still inferior to the DSCMR2 and DSCMR3, which demonstrates the significance of the second term and the third term of the proposed method. Based on the above analysis, we find that formulating both the discrimination loss and the inter-modal invariance loss in the objective function is a valuable strategy for multimodal learning.

5. Conclusion

In this paper, we proposed a new approach (DSCMR) to learn common representations for heterogeneous data. The learned common representations can be both discriminative and modality-invariant for cross-modal retrieval. DSCMR achieved this goal by minimising the discrimination loss (in the common representation space and the label space) and modality invariance loss simultaneously. Extensive experimental results on four widely-used benchmark datasets and the comprehensive analysis have demonstrated the effectiveness of the proposed cross-modal learning strategy, leading to superior cross-modal retrieval performance compared to state-of-the-art methods.

Acknowledgement

This work is supported by National Key R&D Program of China under contract No. 2017YFB1002201 and partially supported by the National Natural Science Foundation of China under Grants No. 61625204, 61836006 and U1831121, the SCU-LuZhou Sciences and Technology Coorperation Program (Grant No. 2017CDLZ-G25), and the Sichuan Science and Technology Planning Projects (Grants No. 18PTDJ0085, 2019YFH0075, 2018GZDZX0030).

References

[1] S. Akaho. A kernel method for canonical correlation analysis. In *Proceedings of the International Meeting of Psychometric Society*, pages 263–269, 2001. 1

[2] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *Proceedings of the International Conference on Machine Learning*, pages 1247–1255, 2013. 1, 2, 6

[3] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yu. Deep visual-semantic hashing for cross-modal retrieval. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1445–1454, New York, NY, USA, 2016. ACM. 2

[4] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and 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*, pages 48:1–48:9, New York, NY, USA, 2009. ACM. 5

[5] F. Feng, X. Wang, and R. Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 7–16, New York, NY, USA, 2014. ACM. 2

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pages 2672–2680, Cambridge, MA, USA, 2014. MIT Press. 3

[7] J. Gu, J. Cai, S. Joty, L. Niu, and G. Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7181–7189, 2018. 1

[8] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. 1, 6

[9] Q. Jiang and W. Li. Deep cross-modal hashing. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3270–3278. IEEE, 2017. 1, 2, 3

[10] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-view discriminant analysis. In *Proceedings of the European Conference on Computer Vision*, pages 808–821, 2012. 6

[11] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-view discriminant analysis. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):188–194, 2016. 6

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014. 5, 7

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the Advances in neural information processing systems*, pages 1097–1105, 2012. 3

[14] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436, 2015. 1

[15] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 8

[16] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 3111–3119, USA, 2013. Curran Associates Inc. 4

[17] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, ICML'10, pages 807–814, USA, 2010. Omnipress. 5

[18] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning*, pages 689–696, 2011. 1

[19] Y. Peng, X. Huang, and J. Qi. Cross-media shared representation by hierarchical learning with multiple deep networks. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3846–3853, 2016. 1, 2, 5, 6

[20] Y. Peng, X. Huang, and Y. Zhao. An overview of cross-media retrieval: Concepts, methodologies, benchmarks and challenges. *IEEE Transactions on Circuits and Systems for Video Technology*, 2017. 1, 2, 5

[21] Y. Peng, J. Qi, X. Huang, and Y. Yuan. CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network. *IEEE Transactions on Multimedia*, 20(2):405–420, Feb 2018. 1, 2, 6

[22] Y. Peng, J. Qi, and Y. Yuan. CM-GANs: Cross-modal generative adversarial networks for common representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2018. 3, 5

[23] Y. Peng, J. Qi, and Y. Yuan. Modality-specific cross-modal similarity measurement with recurrent attention network. *IEEE Transactions on Image Processing*, 27(11):1–1, Nov. 2018. 5

[24] J. C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):521–535, March 2014. 5

[25] J. Qi and Y. Peng. Cross-modal bidirectional translation via reinforcement learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 2630–2636, July 2018. 1, 5

[26] C. Rashtchian, P. Young, M. Hodosh, and 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*, pages 139–147, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. 5

[27] J. Rupnik and J. Shawe-Taylor. Multi-view canonical correlation analysis. In *Proceedings of the Conference on Data Mining and Data Warehouses*, pages 1–4, 2010. 6

[28] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2160–2167, June 2012. 2, 3

[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4, 5, 6, 7, 8

[30] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen. Adversarial cross-modal retrieval. In *Proceedings of the 2017*

ACM on Multimedia Conference, pages 154–162. ACM, 2017. 1, 2, 3, 5, 6

[31] J. Wang, Y. He, C. Kang, S. Xiang, and C. Pan. Image-text cross-modal retrieval via modality-specific feature learning. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 347–354, New York, NY, USA, 2015. ACM. 2

[32] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang. A comprehensive survey on cross-modal retrieval, 2016. 1, 2, 5

[33] W. Wang, R. Arora, K. Livescu, and J. Bilmes. On deep multi-view representation learning. In *Proceedings of the International Conference on Machine Learning*, pages 1083–1092, 2015. 1, 2, 6

[34] W. Wang and K. Livescu. Large-scale approximate kernel canonical correlation analysis. In *Proceedings of the International Conference on Learning Representations*, 2016. 1

[35] W. Wang, X. Yang, B. C. Ooi, D. Zhang, and Y. Zhuang. Effective deep learning-based multi-modal retrieval. *The VLDB Journal*, 25(1):79–101, Feb. 2016. 2

[36] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan. Cross-modal retrieval with cnn visual features: A new baseline. *IEEE Transactions on Cybernetics*, 47(2):449–460, Feb 2017. 1

[37] K. Y. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751, 2014. 3, 4, 5, 6, 7, 8

[38] D. Zhai, H. Chang, S. Shan, X. Chen, and W. Gao. Multiview metric learning with global consistency and local smoothness. *ACM Transactions on Intelligent Systems and Technology*, 3(3):53:1–53:22, May 2012. 2

[39] X. Zhai, Y. Peng, and J. Xiao. Heterogeneous metric learning with joint graph regularization for cross-media retrieval. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, pages 1198–1204. AAAI Press, 2013. 2

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

[41] F. Zheng, Y. Tang, and L. Shao. Hetero-manifold regularisation for cross-modal hashing. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1059–1071, 2018. 2