Applied Intelligence (2022) 52:33–54
https://doi.org/10.1007/s10489-021-02308-3

Intra-class low-rank regularization for supervised and semi-supervised cross-modal retrieval

Peipei Kang[1,2] · Zehang Lin[3] · Zhenguo Yang[1] · Xiaozhao Fang[4] · Alexander M. Bronstein[2] · Qing Li[3] · Wenyin Liu[1]

Accepted: 3 March 2021 / Published online: 22 April 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract
Cross-modal retrieval aims to retrieve related items across different modalities, for example, using an image query to retrieve related text. The existing deep methods ignore both the intra-modal and inter-modal intra-class low-rank structures when fusing various modalities, which decreases the retrieval performance. In this paper, two deep models (denoted as ILCMR and Semi-ILCMR) based on intra-class low-rank regularization are proposed for supervised and semi-supervised cross-modal retrieval, respectively. Specifically, ILCMR integrates the image network and text network into a unified framework to learn a common feature space by imposing three regularization terms to fuse the cross-modal data. First, to align them in the label space, we utilize semantic consistency regularization to convert the data representations to probability distributions over the classes. Second, we introduce an intra-modal low-rank regularization, which encourages the intra-class samples that originate from the same space to be more relevant in the common feature space. Third, an inter-modal low-rank regularization is applied to reduce the cross-modal discrepancy. To enable the low-rank regularization to be optimized using automatic gradients during network back-propagation, we propose the rank-r approximation and specify the explicit gradients for theoretical completeness. In addition to the three regularization terms that rely on label information incorporated by ILCMR, we propose Semi-ILCMR in the semi-supervised regime, which introduces a low-rank constraint before projecting the general representations into the common feature space. Extensive experiments on four public cross-modal datasets demonstrate the superiority of ILCMR and Semi-ILCMR over other state-of-the-art methods.

Keywords Cross-modal retrieval · Deep neural networks · Intra-class low-rank · Supervised learning · Semi-supervised learning

1 Introduction

Cross-modal retrieval, which is an increasingly important topic in the multimedia field, uses data from one modality as a query to retrieve relevant information from other modalities. Due to the cross-modal discrepancy in the data representations and distributions, multi-modal data cannot be compared to measure their similarity directly. Therefore, the main task of cross-modal retrieval is to learn an appropriate multi-modal data representation space in which the cross-modal discrepancy can be eliminated, allowing data from different modalities to be compared directly.

In recent years, a number of methods for fusing cross-modal data and reducing their discrepancies have been proposed by researchers. All these methods learn a common feature space based on the assumption that a latent common space exists for data with similar semantics [27]. Some methods [23] learn this common space by linearly transforming the modality-specific data. During this transformation, certain representations of the original data are used as inputs, such as the histogram feature of images. However, this linear transformation degrades the retrieval performance because the intrinsic nonlinear characteristics cannot be completely modelled by the linear transformation. Recently, the common space has been obtained using deep neural networks (DNNs) [28, 55], which can learn more complex data representations than their linear counterparts. However, many of the frameworks used in these DNN-based methods are separated into two stages that first learn data representation, and then learn the common

✉ Zhenguo Yang
zhengyang5-c@my.cityu.edu.hk

Extended author information available on the last page of the article.

space. On one hand, although sub-optimal results can be obtained in each stage, the previously learned features may not be optimally compatible with the second stage. On the other hand, when separating the cross-modal task into two independent stages, the cross-modal correlation is not considered during the data representation learning stage. Furthermore, most methods consider only a single regularization term during DNN training, which neglects some inherent data structures, such as the low-rank property.

The term 'rank' is a mathematical property of a matrix defined in linear algebra. The rank of a matrix reflects the relevance between its rows or columns, i.e., when many rows are linearly-dependent, the matrix has a low rank, and vice versa. The theory of low-rank representation (LRR) has been applied in a number of different real-world tasks [5, 45, 48]. In subspace learning, the low-rank property is often considered to enhance subspace [19] discriminability. Because data from the same category are related in the original space, they should similarly have relevant representations in the subspace. Mathematically, data with the same category are expected to form a low-rank matrix in the subspace, which is achieved by regularizing the rank of the matrix. However, to the best of our knowledge, the low-rank structure is considered only on a single modality but ignored on cross-modal data due to their discrepancies. Meanwhile, low-rank structures have been explored only under the circumstance of linear transformations and have been entirely neglected by the DNN-based models, mainly because rank minimization is a non-differentiable function.

This paper is an extension of our earlier publication that proposed a deep semantic space with an intra-class low-rank constraint for cross-modal retrieval (DSSIL) [20]. DSSIL includes two subnetworks that transform cross-modal representations into a common feature space. DSSIL exploits semantic consistency regularization to fuse cross-modal data and applies inter-modal intra-class low-rank regularization to reduce the cross-modal discrepancy. DSSIL also provides the rank-r approximation to implement the low-rank constraint. In this paper, we extend DSSIL from four aspects. First, we impose the intra-modal intra-class low-rank regularization on each single modality to encourage data within the same class in the original space to be more relevant in the common space. Second, we provide the explicit gradients of the rank-r approximation. Third, we propose a semi-supervised model that can make full use of both limited labelled data and unlabelled data. Finally, we conduct extensive experimental analyses and visualizations.

More specifically, we propose intra-class low-rank regularizations for deep supervised and semi-supervised cross-modal retrieval (denoted as ILCMR and Semi-ILCMR, respectively). In addition to semantic consistency, ILCMR considers both the intra- and inter-modal intra-class low-rank structures. The intra-modal constraint encourages data from the same class to be more relevant to each other for every modality, while the inter-modal constraint reduces the cross-modal discrepancy. Moreover, we provide the explicit gradients for the rank-r approximation allowing the networks to be optimized in an end-to-end manner. Based on ILCMR, Semi-ILCMR further constrains the low-rank structure before the projection layers. Specifically, before projecting the general data representations into common space, Semi-ILCMR regularizes the rank of the data matrix as the number of categories. In this way, the limited labelled data can guide representation learning, while the abundant unlabelled data can enhance the learning process by exploring the low-rank structure. We conduct extensive experiments to verify the effectiveness of ILCMR and Semi-ILCMR, including an ablation study, an intra-class low-rank analysis, an inter-modal fusion analysis, a discussion of the batch strategy, and so forth. In summary, this paper makes the following contributions.

- An end-to-end ILCMR is proposed for cross-modal retrieval that constrains the semantic consistency and the intra-class low-rank structure of intra- and inter-modal data.
- For the non-differentiable rank minimization problem, the rank-r approximation method and its explicit gradients are presented.
- The Semi-ILCMR model is proposed by further regularizing the low-rank structure of general data representations.
- Extensive experimental analyses are conducted on four public datasets.

The remainder of this paper is organized as follows. We review related works on cross-modal retrieval together with low-rank representation in Section 2. Section 3 presents the proposed ILCMR and Semi-ILCMR, including the general framework, the objective function and the optimization algorithm. Section 4 presents extensive experimental analyses, followed by a conclusion in Section 5.

2 Related works

2.1 Cross-modal retrieval

In the context of cross-modal retrieval, queries and retrieval databases stem from different modalities. On one hand, they come from distinct data distributions and representations; thus, their similarity cannot be calculated

directly. On the other hand, cross-modal data may share the same semantics and be related to each other. One common approach is to transform the data from different representation spaces to a common space where they can be compared directly. The traditional linear methods learn a pair of linear transformation matrices that project the cross-modal data into a common space. Canonical correlation analysis (CCA) [15] learns the transformation matrices by maximizing the correlations among cross-modal data in a common space. Some variations of CCA [15] exist. For example, Kernel-CCA [15] transforms samples from the original space to a feature space through a nonlinear projection and then applies the linear canonical correlation analysis algorithm in the feature space. Explicit construction of the feature space is avoided using the kernel trick. Other works [23] have taken advantage of the pairwise information and learned the linear transformation matrices using the rank rule, i.e., in the common space, objects relevant to the query are expected to rank higher than irrelevant ones. Some researchers have projected cross-modal data to a label semantic space and then minimized the inter-modal invariance [50]. These works usually adopt partially feature representations as data inputs, such as the histogram feature of images. However, these partial features are insufficient to describe the characteristics of the original data well. Moreover, the linear transformation matrices cannot fully extract the intrinsic structure of the data, which is expected to be nonlinear.

With the growing popularity of DNNs, DNN-based cross-modal retrieval methods [44] have become mainstream. Deep-CCA [47] is a deep version of CCA that adopts deep networks instead of linear transformations. The correspondence autoencoder (Corr-AE) [13] is a two-stage deep model that first acquires the modality-specific representations and then performs data reconstruction and cross-modal correlation mining through two subnetworks. The cross-media multiple deep network (CMDN) [26] learns the feature representation for each modality through multiple networks and then obtains the common representation using a stacked network. Hybrid representation learning (HRL) [1] learns modality-friendly and modality-mutual representations first, based on which the hybrid representation is obtained, and then achieves shared representation using stacked bi-modal autoencoders. Adversarial cross-modal retrieval (ACMR) [39] utilizes adversarial training to fuse two modalities. DLA-CMR [36] adopts dictionary learning to reconstruct discriminative common features. GSS-SL [52] constructs a local similarity graph to make use of unlabelled data and learn features. The graph representation learning method (GRL) [3] builds a cross-modal graph to reconstruct the original data and their relationships, which is then used to learn a common representation space. The deep relational similarity learning method (DRSL) [41] utilizes pairwise similarity to bridge the heterogeneity gap. Cross-modal correlation learning (CCL) [28] considers the inter-modal relationship and the fine-grained problem during two-stage learning. Some researchers have extracted general data representations using pre-trained networks and projected them to a label space in the second stage [18, 54]. Because the general cross-modal representations lack interaction and may not be optimal for the subsequent stage, the two-stage strategies are unable to learn the most appropriate representations to meet the final objective for the cross-modal task [17]. In contract, the end-to-end cross-modal methods learn the modality-specific and common representations simultaneously; then, all the networks can be optimized under the final objective [17]. Deep semantic matching (deep-SM) [42] transforms the cross-modal retrieval problem into a classification problem, which treated the label distribution as the semantic representation. Similarly, Situ et al. [38] learned a deep semantic space (DSS) for cross-modal event retrieval. However, these methods consider only a single regularization term during DNN training and neglect the inherent low-rank data structure. Recent works tend to combine multiple regularizations. Deep adversarial metric learning (DAML) [46] reduces the heterogeneous gap through adversarial learning, and constrains the discrimination as well as the correlation of features simultaneously. Semantic ranking structure preserving (SRSP) [24] proposes a common space in which the dependency relationships between labels and the correlation rankings of the representations are preserved. There are also some hashing works [32, 37, 40] that represent samples as binary codes and calculate similarity using the XOR operator. For example, UCH [22] proposed the coupled cycleGAN for cross-modal retrieval: the outer GAN is used to learn a common representation, and the inner GAN is used to generate hash codes. This type of binary representation accelerates the retrieval process but reduces the retrieval quality due to quantization errors.

In summary, the current cross-modal retrieval approaches suffer from the following shortcomings. 1) They often use partially described features that insufficiently represent the original data. 2) In the two-stage learning strategies, the first-stage network cannot be jointly tuned with the retrieval loss function. 3) The intrinsic properties underlying the data, such as low-rank structures, have been neglected. 4) Most of the methods are supervised and thus require labour-intensive labelling operations.

2.2 Low-rank representation

Mathematically, the rank of a matrix is defined as the number of its linearly-independent vectors. When many vectors are maximally dependent, a matrix is low-rank, and

vice versa. The concept of LRR has been applied in many tasks [11, 12, 43, 51]. These methods are usually based on dictionary learning or graph learning, and they use the reconstruction coefficient matrix as the data representation or the edge weights in the data graph. To make similar data more relevant in subspaces, the coefficient reconstruction matrix must be low-rank. Wen et al. [43] applied LRR to discover the intrinsic subspace structure and adaptively construct graphs for clustering. Fei et al. [12] introduced the pairwise similarity relationship into LRR; thus, it can capture both the global low-rank structure and the local similarity structure of data, which is helpful in learning a more discriminative data graph for classification. Ding et al. [8] implemented LRR in the context of zero-shot learning. Other methods have tended to directly constrain the low-rank structure in the latent subspace. The original data within the same class are similar; thus, in the subspace, the intra-class representations constitute a matrix whose rank is expected to be low. Kang et al. [19] transformed all the data to a subspace, and minimized the rank of the intra-class data. Qiu et al. [33] minimized the rank of the intra-subspace data and maximized the separation of inter-subspace data. Furthermore, Ding et al. [7] applied LRR to cross-domain learning. Lezama et al. [21] proposed a plug-and-play loss term for deep networks that explicitly reduced the intra-class variance and simultaneously enforced the inter-class margin. However, most of the proposed approaches focus solely on single-modal data.

Generally, the intra-class low-rank structure is explored merely on a single modality, and its application to cross-modal data to address their discrepancies is overlooked. Second, the low-rank structure has been explored only when using linear transformations not DNNs. This is mainly because of the non-differentiable property of the rank minimization problem, which makes it unable to be optimized in a gradient-friendly manner.

3 Methodology

In this section, we first formulate the cross-modal retrieval problem. Then, we present the framework and detail the objective function of ILCMR, followed by an alternative optimization strategy. Finally, the semi-supervised ILCMR is descibed.

3.1 Notation

Without loss of generality, we focus on two modalities, image and text, to illustrate the problem of cross-modal retrieval. We denote the image (v) and text (t) databases as $D^m, m \in \{v, t\}$, $D^m = D_L^m \cup D_U^m$, where D_U^m are the unlabelled data, $D_L^m = \{d_{Li}^m\}_{i=1}^{n_L^m}$ are the labelled data, and n_L^m is the number of labelled samples in modality m. The labels are one-hot encoded as $Y^m = \{y_i^m\}_{i=1}^{n_L^m}$. The labelled data can also be described categorically as $D_L^m = \{D_L^{m(k)}\}_{k=1}^{C}$, where C is the number of categories.

Because images and text have different representations, it is difficult to compare them directly to enable cross-modal retrieval. To solve this problem, cross-modal methods first extract general representations from each modality using neural networks $f^m, m \in \{v, t\}$ and then project these representations to a common space through projection layers $g^m, m \in \{v, t\}$. We denote the general representations as $Z^m = f^m(D^m, \theta^m)$, where θ^m are the network parameters. After executing the projection layers, we obtain a common representation $X^m = g^m(f^m(D^m, \theta^m), \omega^m)$, where ω^m are the projection layer parameters. Specifically, $X^m = X_L^m \cup X_U^m$. $X_L^m = \{x_{Li}^m\}_{i=1}^{n_L^m}$ are the labelled common representations, which can also be represented as $X_L^m = \{X_L^{m(k)}\}_{k=1}^{C}$.

3.2 The ILCMR framework

The ILCMR framework is illustrated in Fig. 1. It transforms data from the distinct image and text spaces to a common space in an end-to-end manner. The framework uses three supervised regularization terms, including semantic consistency, intra-modal intra-class low-rank structure, and inter-modal intra-class low-rank structure during common space learning to induce discriminative representations. In the following, we detail the ILCMR ingredients.

3.2.1 Image network

The image network consists of a pretrained ResNet50 and a projection layer. Many models suitable for image feature learning exist; in this paper, we chose ResNet50 as the backbone for the following reasons. First, ResNet [16] models have demonstrated impressive performances for image classification in the past few years, and they have a light parametric burden, which is advantageous for fine tuning during network training. Second, as an extension of DSSIL, ILCMR focuses more on exploring the low-rank structure to fuse cross-modal data; therefore, we adopted the same network as DSSIL to ensure a fair comparison. As shown in Fig. 1, raw images are input into the network. The input layer is followed by a fully-connected projection layer whose output dimension equals the number of categories.

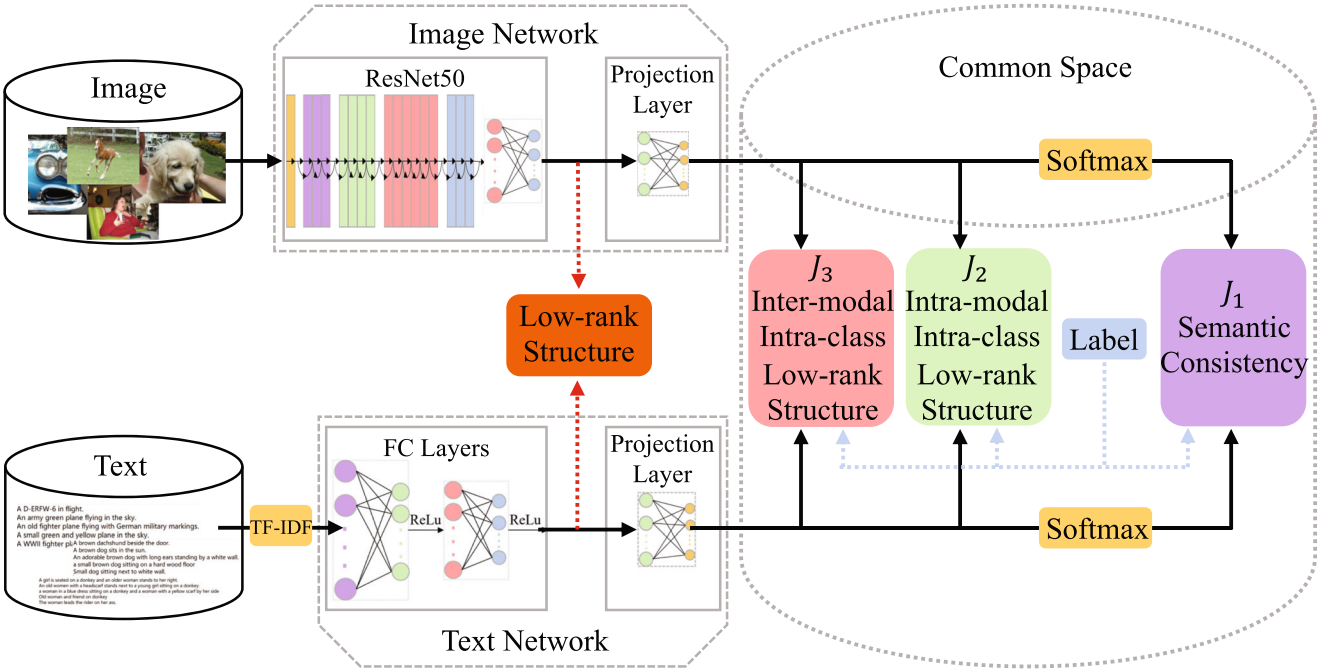


Fig. 1 ILCMR and Semi-ILCMR frameworks. In ILCMR, J_1, J_2 and J_3 in common space are constrained. Semi-ILCMR further adds a low-rank structure constraint before the projection layers

Note that all the parameters in the image network are learned during training.

3.2.2 Text network

Text representation is more discriminative than image repensentation, and mapping from a text feature space to the common semantic space is easier to be accomplish [42]. Fully-connected layers are generally adopted in text networks [18, 20, 38, 42, 49, 54]; to ensure fair comparisons, we designed a three-layer fully-connected network (FC layers) as the text network. This network maps the text feature space to the semantic space and learns common text representations. Each of the first two layers is followed by a nonlinear rectified linear unit activation function (ReLU). Similar to the image network, the last FC layer is a projection layer. For text, we extract the term frequency-inverse document frequency (TF-IDF) features and use them as input to the network. Alternatively, some pretrained models such as BERT [6] and ELMo [29] can also be used to extract text features. The parameters in the text network are denoted collectively as θ_T and are learned during training.

3.2.3 Semantic consistency

Cross-modal data are related and have the same semantic labels. To maintain the semantic consistency between cross-modal data, their common representations are converted to probability distributions over categories through a softmax function. Therefore, we utilize the cross-entropy between the predicted category distribution and the one-hot encoded ground truth as the semantic consistency loss.

3.2.4 Intra-modal intra-class low-rank structure

The rank of a matrix is defined as the maximal number of its linearly independent vectors. In one sense, rank describes the informativeness of the matrix: when the vectors are correlated, the matrix is low-rank, and vice versa. For a single modality, the original data within the same class are similar; thus, the intra-class representations in matrix form are expected to be low-rank, forming a low-dimensional subspace.

In ILCMR, we propose an intra-modal intra-class low-rank regularization that induces high relevance between intra-class samples. Ideally, we expect that the intra-class representations are more similar to each other. Consequently, we treat each representation as a row vector, and construct a matrix using the intra-class vectors. Additionally, we minimize the informativeness for each intra-class matrix, i.e., the rank of the matrix is promoted to be low. As a result, the network is able to extract the interacting features for samples in the same class.

3.2.5 Inter-modal intra-class low-rank structure

ILCMR also regularizes the inter-modal intra-class low-rank structure for the current tasks. Similar to the intra-modal

scenario, we constrain the matrix, which consists of cross-modal data samples within the same category to be low-rank. It is worth noting that low-rank regularization have rarely been integrated into the DNN framework due to its non-smooth nature. We introduce the rank-r approximation algorithm to solve the rank minimization problem as detailed in the next section.

3.3 Objective function

ILCMR learns a common space for cross-modal retrieval. In this common space, samples from the same semantic category should be similar, while samples from different semantic categories should be dissimilar. To obtain this semantic consistency between cross-modal data, their representations are converted to a probability distribution over the semantic categories through a softmax function. Furthermore, we use cross-entropy loss between the predicted category distributions and the one-hot encoded labels.

Let the vector $x \in R^C$ be a common representation (which could be an image or a text) with a label $y \in R^C$. We can predict its category distribution through a softmax function

$$\hat{y}_k = \frac{\exp(x_k)}{\sum_{j=1}^{C} \exp(x_j)}, \tag{1}$$

where x_k is the k-th element of x. Given the predicted category distribution, we then utilize the cross-entropy loss.

$$L_{ce}(\hat{y}, y) = -\sum_{k=1}^{C} y_k * \log(\hat{y}_k) \tag{2}$$

To maintain the semantic consistency between the image and the text, we align them in the semantic space, leading to

$$J_1 = \sum_{m}^{\{v,t\}} \sum_{i=1}^{n_L^m} L_{ce}(\hat{y}_{Li}^m, y_i^m). \tag{3}$$

Then, to boost the relevance of the intra-class samples, we impose the intra-modal intra-class low-rank structure on each individual modality simultaneously. For each modality, we aggregate the common representation within the k-th category and construct a matrix $X_L^{m(k)} \in R^{(n_L^{m(k)} \times C)}$, whose rank is minimized. We constrain this type of structure for both the image and the text modalities, through the intra-modal intra-class low-rank loss term (4):

$$J_2 = \sum_{m}^{\{v,t\}} \sum_{k=1}^{C} rank(X_L^{m(k)}), \tag{4}$$

where $rank(\cdot)$ denotes the rank. Furthermore, we regularize the inter-modal intra-class low-rank structure to reduce the cross-modal discrepancy. We group the cross-modal samples within the same class in the common space into a matrix, and minimize its rank. The inter-modal intra-class low-rank loss is defined as follows

$$J_3 = \sum_{k=1}^{C} rank([X_L^{v(k)}, X_L^{t(k)}]). \tag{5}$$

By combining (3), (4) and (5), we obtain the final objective function of ILCMR:

$$J_{ILCMR} = \lambda_1 J_1 + \lambda_2 J_2 + (1 - \lambda_1 - \lambda_2) J_3, \tag{6}$$

where λ_1 and λ_2 are coefficients that reflect the relative contribution of each term.

3.4 Optimization

The objective function (6) is difficult to optimize during network back-propagation because the rank is non-differentiable, and rank minimization cannot be automatically derived. Therefore, we provide the rank-r approximation to optimize the rank minimization problem, allowing the entire objective to be optimized automatically by using stochastic gradient descent (SGD).

Let A be a rank-r matrix in which r is expected to be small. We solve the non-differentiable rank minimization problem by the following rank-r approximation. First, A is decomposed using singular value decomposition (SVD). Then, its singular values are sorted in descending sequence. Finally, we minimize the last $(s-r)$ singular values, in other words, we minimize the error between 1 and the largest r singular values as follows:

$$rank(A) \approx \rho(A, r) = \frac{\sum_{i=r+1}^{s} \delta_i}{\sum_{i=1}^{s} \delta_i} = 1 - \frac{\sum_{i=1}^{r} \delta_i}{\sum_{i=1}^{s} \delta_i}, \tag{7}$$

where $\{\delta_i\}_{i=1}^{s}$ represents the sorted singular values of A, and s is the count of these singular values. We can see that $\rho(A, r)$ promotes the rank of A to be r.

The approximation in (7) is reasonable. The rank of a matrix is calculated by counting the number of its nonzero singular values. Intuitively, we can approximate this problem by either minimizing the smallest singular values to 0 or by maximizing the largest singular values to 1 as shown in (7). Theoretically, we prove the rank-r approximation formula as follows.

Proposition 1 *Suppose we have a matrix A with singular values $\{\delta_i\}_{i=1}^{s}$. We then have*

$$min \quad rank(A) \approx \rho(A, r) = \sum_{i=r+1}^{s} \delta_i, \tag{8}$$

where $r < s$ is the rank-r approximation parameter.

Proof A full-rank matrix A can be decomposed by the SVD method, i.e., $A = U\Sigma V^T$. Here, $\Sigma = diag(\delta_1, \delta_2, \ldots, \delta_s)$, where $\{\delta_i\}_{i=1}^s$ are the singular values whose numerical values are sorted by Formula (9) [9]:

$$\delta_1 \geq \delta_2 \geq \cdots \geq \delta_s > 0. \tag{9}$$

The rank of a matrix is calculated by counting its nonzero singular values; thus, the rank of A is s.

Eckart-Young-Mirsky theory [14] assumes that a lower-rank matrix better approximates the original matrix. Assume that the rank-r matrix B is the lower-rank approximation of matrix A, and $\{\xi_i\}_{i=1}^s$ are the singular values of B, i.e.,

$$\xi_1 \geq \xi_2 \geq \cdots \geq \xi_r > 0 = \xi_{r+1} = \xi_{r+2} = \cdots = \xi_s, \tag{10}$$

where $r < s$ is the rank of B.

It is known that the singular values of a matrix are fixed [53]. Thus, if B is the lower-rank approximation of matrix A, we can derive (11).

$$\delta_i \approx \xi_i, i = 1, \ldots, s. \tag{11}$$

Furthermore, we have $\delta_i \approx \xi_i = 0, i = r+1, r+2, \ldots, s$. Consequently, we can minimize the smallest singular values,

$$min \sum_{i=r+1}^{s} \delta_i. \tag{12}$$

By normalizing (12), we can acquire the approximation from (7).

This approximation is differentiable, and the derivative in terms of A can be calculated by the chain rule as follows:

$$\frac{\partial rank(A)}{\partial A} \approx \frac{\partial \rho(A, r)}{\partial A} = \sum_{i=1}^{s} \frac{\partial \rho}{\partial \delta_i} \cdot \frac{\partial \delta_i}{\partial A}. \tag{13}$$

Specifically,

$$\frac{\partial \rho}{\partial \delta_i} = \begin{cases} -\frac{\sum_{j=r+1}^{s} \delta_j}{\left(\sum_{j=1}^{s} \delta_j\right)^2}, & i = 1, \ldots, r \\ \frac{\sum_{j=1}^{r} \delta_j}{\left(\sum_{j=1}^{s} \delta_j\right)^2}, & i = r+1, \ldots, s \end{cases}, \tag{14}$$

and

$$\frac{\partial \delta_i}{\partial A} = 1./(u_i {v_i}^T), \tag{15}$$

where u_i and v_i are the i-th vectors of A's left and right singular matrices respectively. The symbol '$./$' denotes the element division operation. A detailed derivation can be found in Appendix A. □

3.5 Semi-supervised ILCMR

ILCMR is a supervised approach in which all the training samples should be labelled to guide model training. However, in real-world situations, most of the data may be unlabelled. Thus, being able to use this abundance of unlabelled data is critical. To this end, we also propose a semi-supervised ILCMR (Semi-ILCMR), which utilizes unlabelled data during training. As mentioned in ILCMR, regularizing the low-rank structure of the data matrix can promote similar samples to obtain similar feature representations. In Semi-ILCMR, we regularize the low-rank structure of the entire data matrix-including both the labelled and unlabelled data. As a result, the labelled samples can be used to supervise model training, while the unlabelled samples can help in learning feature representations similar to those labelled samples.

Specifically, as shown in Fig. 1, we can obtain the labelled and unlabelled general representations from ResNet for images and from the FC layers for text. Because ILCMR is a supervised method that projects labelled data to a common space, Semi-ILCMR utilizes the unlabelled data before projecting the general representations to the common space. In more detail, we assemble both the labelled and unlabelled general representations into a matrix, and then regularize its rank to match the number of categories. Consequently, the labelled intra-class representations and their similar unlabelled representations will be close. Similar to ILCMR, we also regularize the intra- and inter-modal low-rank structure. For the intra-modal low-rank structure, we group all the intra-modal general representations together, and constrain the rank to match the number of categories, which is formulated as follows:

$$J_4 = \sum_{m}^{\{v,t\}} \rho(Z^m, C), \tag{16}$$

where $\rho(\cdot)$ is defined in (7). For the inter-modal low-rank structure, we stack all the data into a matrix, and the regularization is denoted as:

$$J_5 = \rho([Z^v, Z^t], C). \tag{17}$$

By further utilizing the labelled data as in ILCMR, the overall objective of Semi-ILCMR is as follows:

$$J_{Semi-ILCMR} = J_{ILCMR} + \lambda_2 J_4 + (1 - \lambda_1 - \lambda_2) J_5. \tag{18}$$

4 Experimental evaluation

To verify the effectiveness of ILCMR, we conducted experiments on four public cross-modal datasets. We compared ILCMR with the state-of-the-art cross-modal retrieval methods, and conducted further analyses, including ablation, parameter sensitivity, optimizer analysis, text embedding discussion, and convergence studies. We also examined the effectiveness of the intra-class low-rank constraint and the fusion of the cross-modal data.

4.1 Datasets

The details of the four public datasets are as follows; their statistics are summarized in Table 1.

Pascal Sentence dataset [34]. The Pascal Sentence dataset is a widely used cross-modal dataset that includes 1,000 pairs of images and text documents. Each document includes 5 text sentences describing the corresponding image. The image-document pairs are labelled into 20 categories, with 50 pairs in each category. Following the dataset partition strategy [28], we randomly selected 40 pairs per class to form a training set, used 5 of the remaining pairs per class as a validation set, and used the remainder as a test set. As shown in Table 1, the 5 sentences in each document are treated as 5 instances.

Wikipedia dataset [35]. The wikipedia dataset provides a document that lists 2,866 image-text pairs. The pairs are labelled into 10 categories, and we follow the same dataset partitioning strategy used in [28], by randomly selecting 2,173 pairs as the training set, 231 pairs as the validation set, and 462 pairs as the test set.

NUS-WIDE-10k dataset [4]. The NUS-WIDE dataset includes 269,648 images with tags; these images are classified into 81 categories. By following [13] to compose the NUS-WIDE-10k dataset, we randomly selected 1,000 instances that exclusively belong to each of the 10 largest categories. The 10 categories include 'animal', 'clouds', 'flowers', 'food', 'grass', 'person', 'sky', 'toy', 'water' and 'window'. For each category, we randomly selected 800 instances for training, and 100 instances each for validation and testing. Noted that the tags are treated as the text modality in this dataset.

Table 1 Statistics of the datasets used in the experiments, where '*/*/*' in the 'Instance' column means the number of instances in the training/validation/test subset

Dataset	Label	Modality	Instance
Pascal Sentence	20	image	800/100/100
		text	4,000/500/500
Wikipedia	10	image	2,173/231/462
		text	2,173/231/462
NUS-WIDE-10k	10	image	8,000/1,000/1,000
		text	8,000/1,000/1,000
Wiki-Flickr Event	82	image	17,269/4,318/7,238
		text	7,176/1,794/2,990

Wiki-Flickr Event dataset [49]. The event-related Wiki-Flickr Event dataset contains 82 event categories, each of which is composed of a variety of texts and images from different news media platforms comprising 11,960 texts and 28,825 images. The dataset does not include pair information (the texts are not exact descriptions of the images, but both describe the same event via different representations); instead, the Wiki-Flickr Event dataset was released as a weakly aligned multi-modal dataset. Following the data partition in [49], we randomly selected 60% of the texts and images from each category as a training set, 15% of the texts and images from each category as a validation set and used the rest as the test set.

4.2 Implementation details

4.2.1 Neural networks

The text network of the proposed ILCMR includes 2 FC layers (1000, 100), each with a ReLU activation. The output dimensionality of ResNet50 in the image network is set to 100. The projection layer projects general representations to the common feature space ($100 \rightarrow C$). In the model training stage, we optimize the network parameters using SGD and gradient clipping; the learning rate is set to 0.1. For the rank-r optimization, we set the parameter $r = 1$ in (7) for two reasons: 1) The intra-class common representations are expected to be relevant to each other, and the ideal situation is $r = 1$; 2) Our previous research [20] experimentally proved that the $r = 1$ approximation performs better. In this paper, we conduct two retrieval tasks: using an image as a query to retrieve text (Img2Txt) and using a text as a query to retrieve images (Txt2Img). For the evaluation, we adopted the model parameters that obtained the best average mean average precision (MAP) for Img2Txt and Txt2Img on the validation dataset.

4.2.2 Data preparation

We augmented the image input data using the the cropping, flipping, resizing, and normalization operations. We first cropped and horizontally flipped images randomly and then

resized them to 224×224 and normalized them by mean and standard deviation. For the text inputs, we extracted their TF-IDF features as the raw inputs.

4.2.3 Batch strategy

During training, the batch size in each modality was set to 32. Because ILCMR introduces the inter-modal intra-class low-rank constraint, intra-class samples are expected to be included within each batch. For a single modality, we randomly selected a number of samples per class (denoted as spc) ranging within the batchsize factor set, i.e., {1, 2, 4, 8, 16, 32}. We will subsequently discuss the impact of different spc. The categories in each batch are selected randomly.

4.2.4 Evaluation metrics

Following the distance measurement used by [28], we adopted the cosine distance to calculate the similarity in the common space. To assess the retrieval performance, we applied MAP to the retrieved list. which is the mean of the average precision (AP) over all queries. The AP integrates both the precision and the ranking information of the retrievals for evaluation. Suppose that there are n instances in the retrieval database, and m query instances. The average precision of the $q-th$ query is defined as follows:

$$AP(q) = \frac{1}{R}\sum_{p=1}^{n}\frac{R_p}{p} \times rel(p), \tag{19}$$

where R is the number of relevant instances in the retrieval database (instances that share the same ground truth label as the query q); R_p is the number of relevant instances in the first p retrieved samples; $rel(p) = 1$ if the $p-th$ sample is relevant to q; and $rel(p) = 0$ otherwise. Finally, we obtain the MAP by averaging the APs of all the queries.

$$MAP = \frac{1}{m}\sum_{q=1}^{m} AP(q). \tag{20}$$

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

To verify the effectiveness of the proposed ILCMR, we compare it with the state-of-the-art methods, including linear transformation-based methods, two-stage DNN-based methods, and end-to-end DNN methods. Note that some of the presented results are cited from other works [18, 28, 49, 54].

4.3.1 MAP comparison on the Pascal Sentence dataset

Table 2 reports the MAP results on the Pascal Sentence dataset. We make the following observations.

- The DNN-based methods perform significantly better than the traditional linear retrieval methods by benefiting from the fact that DNNs can learn the nonlinear characteristics of cross-modal data, while linear methods are unable to explore these characteristics.
- Most of the end-to-end learning methods are superior to the two-stage methods. During the end-to-end training, both the feature representation and common representation learning are optimized, while the two-stage strategy optimize only the second-stage network.
- Among all the compared methods, our approach achieves the best retrieval results. One reason may be that in the Pascal Sentence dataset, the samples from different categories are more distinct, i.e., the semantic labels are more accurate. In addition, ILCMR is implemented with three constraints under the supervision of categorical labels, allowing it to better use the semantic label information. This result indicates the effectiveness of the semantic consistency constraint and the intra-class low-rank constraint.

4.3.2 MAP comparisons on the Wikipedia dataset

Table 3 shows the MAP comparison on the Wikipedia dataset. From Table 3, we make the following observations:

Table 2 MAP comparison of ILCMR and other methods on the Pascal Sentence dataset. Best results are highlighted in bold

Method	Img2Txt	Txt2Img	Average
CCA [15]	0.169	0.151	0.160
KCCA [15]	0.361	0.325	0.343
DCCA [47]	0.456	0.462	0.459
Corr-AE [13]	0.489	0.484	0.487
JRL [50]	0.563	0.505	0.534
CMDN [26]	0.534	0.534	0.534
ACMR [39]	0.538	0.544	0.541
Deep-SM [42]	0.560	0.539	0.550
CCL [28]	0.566	0.560	0.563
DSS [38]	0.545	0.574	0.560
CBT [31]	0.602	0.583	0.592
S^3CA [49]	0.588	0.607	0.598
DSSIL [20]	0.630	0.652	0.641
ILCMR	**0.632**	**0.654**	**0.643**

Table 3 MAP comparison of ILCMR and other methods on the Wikipedia dataset. Best results are highlighted in bold

Method	Img2Txt	Txt2Img	Average
CCA [15]	0.258	0.250	0.254
KCCA [15]	0.215	0.214	0.215
DCCA [47]	0.440	0.390	0.415
Corr-AE [13]	0.402	0.395	0.399
JRL [50]	0.343	0.376	0.330
CMDN [26]	0.488	0.427	0.458
ACMR [39]	0.479	0.426	0.452
GSS-SL [52]	0.466	0.413	0.440
Deep-SM [42]	0.478	0.422	0.450
CCL [28]	0.504	0.457	0.481
DSS [38]	0.516	0.461	0.489
CBT [31]	0.516	0.464	0.490
DSCMR [54]	0.521	0.478	0.499
SDML [18]	0.522	0.488	0.505
S^3CA [49]	0.551	0.485	0.518
DSSIL [20]	0.563	0.515	0.539
ILCMR	**0.564**	**0.518**	**0.541**

- Similar to the results on the Pascal Sentence dataset, ILCMR outperforms the other methods on the Wikipedia dataset.
- The results on the Wikipedia dataset are generally lower than those on the Pascal Sentence dataset; however, the performances are reasonable. The categories in the Pascal Sentence dataset are concrete objects such as birds, cars, and sofas, which are easier to distinguish, while the Wikipedia dataset categories are abstract concepts, such as art, history, and music, which are more difficult to differentiate.

4.3.3 MAP comparison on the NUS-WIDE-10k dataset

Table 4 shows the MAP results on the NUS-WIDE-10k dataset. Although DSCMR obtains higher MAP on the Txt2Img task, ILCMR still achieves the best average retrieval performance on this dataset.

4.3.4 MAP comparison on the Wiki-Flickr Event dataset.

Table 5 presents the retrieval comparison on the Wiki-Flickr Event dataset. We found that

- ILCMR obtains the best cross-modal retrieval MAP compared with the other end-to-end deep models. This result illustrates the effectiveness of the intra-class low-rank structure constraints.

Table 4 MAP comparison of ILCMR and other methods on the NUS-WIDE-10k dataset. Best results are highlighted in bold

Method	Img2Txt	Txt2Img	Average
CCA [15]	0.378	0.394	0.386
DCCA [47]	0.532	0.549	0.540
JRL [50]	0.586	0.598	0.592
CMDN [26]	0.492	0.515	0.504
CCL [28]	0.506	0.535	0.521
ACMR [39]	0.588	0.599	0.593
DSCMR [54]	0.611	**0.615**	0.613
DSSIL [20]	0.632	0.600	0.616
ILCMR	**0.640**	0.600	**0.620**

- The Wiki-Flickr Event dataset is a weakly aligned dataset. Although the MAP scores of the other methods generally decrease, ILCMR still obtains relatively good results. On one hand, this reflects the difficulty of cross-modal retrieval on this weakly aligned dataset. On the other hand, it shows the effectiveness of ILCMR on a weakly aligned dataset.

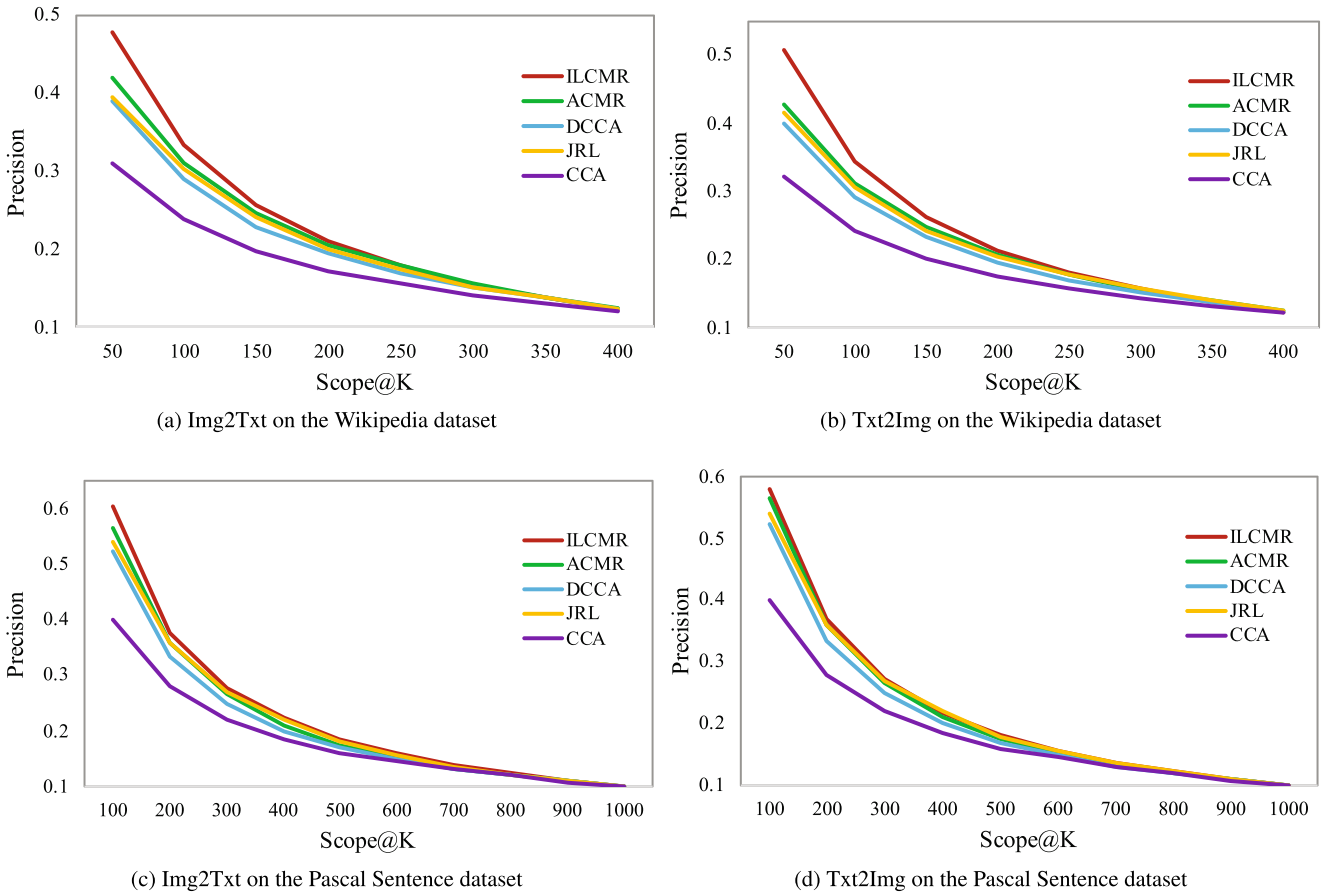
In addition to MAP, we report the precision-scope curves of the approaches on Wikipedia and NUS-WIDE-10k datasets. A precision-scope curve illustrates the variation in precision that occurs with the number of retrieved samples, as shown in Fig. 2. Like the MAP comparison, the proposed ILCMR achieves the best performance.

4.4 Ablation study

The objective function of ILCMR consists of three regularization terms, i.e., the semantic consistency term (J_1), the intra-modal intra-class low-rank term (J_2), and the inter-modal intra-class low-rank term (J_3). To investigate the impact of all three terms, we conducted ablation experiments on the Wikipedia dataset. The experimental results are shown in Table 6, where ILCMR_1 represents ILCMR with only J_1 regularization, ILCMR_1_2 is the ILCMR without J_3 regularization, ILCMR_1_3 is

Table 5 MAP comparison of ILCMR and other methods on the Wiki-Flickr dataset. Best results are highlighted in bold

Method	Img2Txt	Txt2Img	Average
CCA [15]	0.244	0.218	0.231
Deep-SM [42]	0.373	0.386	0.380
DSS [38]	0.578	0.570	0.574
S^3CA [49]	0.608	0.611	0.610
DSSIL [20]	0.619	0.625	0.622
ILCMR	**0.621**	**0.629**	**0.625**



(a) Img2Txt on the Wikipedia dataset

(b) Txt2Img on the Wikipedia dataset

(c) Img2Txt on the Pascal Sentence dataset

(d) Txt2Img on the Pascal Sentence dataset

Fig. 2 The precision-scope curves on the Wikipedia and Pascal Sentence datasets, where K is the number of retrieved samples

the ILCMR without J_2 regularization (also known as the DSSIL), ILCMR_2_3 is the ILCMR without J_1 regularization, and ILCMR includes all the regularizations.

From Table 6, we can make the following observations.

- The combination of all three regularizations obtains the best retrieval performance.
- Semantic consistency is critical because using this regularization alone achieves good performance, and the performance decreases dramatically when this regularization is removed. This observation is consistent with that of the trade-off parameter analysis in Section 4.6.1, *Trade-off parameter analysis*.
- Both the inter-modal and the intra-modal intra-class low-rank constraints are helpful in improving the performance on the retrieval task. The inter-modal and intra-modal intra-class low-rank constraints achieve similar performances because they have some overlap when minimizing the rank of the data matrix. However, combining them obtains better results because it makes the intra-class samples from the same original space more relevant in the common space and simultaneously reduces the inter-modal discrepancy.

4.5 Semi-supervised ILCMR

From the ablation study, we know that using the intra-class low-rank constraint alone will cause model collapse because all the representations are optimized to be relevant as 0 to meet the constraint of intra-class rank minimization. Therefore to further explore the effectiveness of the intra-class low-rank regularization, we adopt partial supervision instead of full supervision (i.e., we create a semi-supervised version of ILCMR). During model training, it is assumed that only parts of the training data are labelled, while other

Table 6 The MAP comparison of ILCMR and its variations on the Wikipedia dataset. Best results are highlighted in bold

Method	Img2Txt	Txt2Img	Average
ILCMR_1	0.557	0.513	0.535
ILCMR_1_2	0.555	**0.521**	0.538
ILCMR_1_3	0.563	0.515	0.539
ILCMR_2_3	0.136	0.126	0.131
ILCMR	**0.564**	0.518	**0.541**

parts are unlabelled. Table 7 shows a MAP comparison of the results of ILCMR and Semi-ILCMR.

In Table 7, 'Labelled' denotes the percentage of labelled data over all the training data. ILCMR_1 was introduced in Table 6 and trained with only labelled data. We find that an increase in labelled training data also causes the retrieval performance to rise. In addition, it is obvious that ILCMR can obtain competitive MAP score even when only 20% of the training data are labelled, which shows the effectiveness of semi-supervised ILCMR and the low-rank regularizations.

4.6 Parameter analysis

4.6.1 Trade-off parameters

The objective function of ILCMR includes two trade-off parameters (i.e., λ_1 and λ_2), that control the contribution of each regularization. λ_1 reflects the importance of the semantic consistency loss, which ranges from $[0, 1]$, and λ_2 reflects the importance of intra-modal intra-class low-rank loss, which ranges from $[0, 1 - \lambda_1]$. As a result, the importance of the inter-modal intra-class low-rank loss is $1 - \lambda_1 - \lambda_2$. During the experiments, we chose the values of these parameters using a grid-search strategy. Figure 3 shows the MAP results under different parameter combinations on the Pascal Sentence dataset.

As Fig. 3 shows, the retrieval obtains the best MAP results on both the Img2Txt and Txt2Img tasks when λ_1 equals 0.1 instead of 1, indicating the critical importance of the intra-class low-rank constraints. However, we can also find that the MAP scores approach 0 when λ_1 equals 0, indicating that the unique intra-class low-rank regularization is ineffective. This is because intra-class low-rank regularization alone will result in a trivial solution. In this case, all the representations are constrained to be relevant and finally become 0. Nevertheless, this phenomenon also indicates that the semantic consistency constraint contributes less to the overall objective. Regardless, a semantic consistency constraint is necessary; it separates samples into different semantics insofar as possible, avoiding model collapse. We can also see that λ_2 has a limited effect on the retrieval results, because both the inter-modal and intra-modal constraints are based on the intra-class low-rank constraint, and there is some overlap. Minimizing the ranks of two modalities separately or concurrently can achieve similar performances, but minimizing them simultaneously leads to better results, because it not only unites the intra-class data in each single modality but also reduces the inter-modal discrepancy.

Table 7 The MAP comparison of ILCMR with and without intra-class low-rank regularizations, when only part of the training data are labeled on Wikipedia dataset. Best results are highlighted in bold

Labeled	Method	Img2Txt	Txt2Img	Average
20%	ILCMR_1	0.438	0.352	0.395
	ILCMR	0.453	0.373	0.413
	Semi-ILCMR	**0.457**	**0.435**	**0.446**
40%	ILCMR_1	0.470	0.366	0.418
	ILCMR	**0.474**	0.387	0.431
	Semi-ILCMR	0.465	**0.465**	**0.465**
60%	ILCMR_1	0.472	0.386	0.429
	ILCMR	0.499	0.411	0.455
	Semi-ILCMR	**0.503**	**0.473**	**0.488**
80%	ILCMR_1	0.492	0.429	0.460
	ILCMR	**0.536**	0.441	0.488
	Semi-ILCMR	0.533	**0.496**	**0.515**

4.6.2 Batch parameter

In each batch, the parameter *spc* denotes the number of samples of each class in that batch. In our experiments, we set the batch size to 32, and to ensure the same number of samples in each class, we set *spc* to a factor of 32 (i.e., 1, 2, 4, 8, 16, or 32). We varied the value of *spc* to investigate the effect of *spc* on the retrieval task. Table 8 shows the MAP results under different *spc* values on the Pascal Sentence dataset. As these results show, the average MAP is the highest when spc equals 4. However, when *spc* equals 32, the results are nearly 0. This is because when *spc* equals 32, each batch is trained for single-class data samples, and the semantic consistency cannot differentiate among the interclass samples, leading to a decrease in the retrieval performance. Simultaneously, this phenomenon also indicates the sensitivity of selecting an appropriate *spc* value.

4.7 Convergence analysis

Figure 4 shows the convergence speed of ILCMR on both the Pascal Sentence dataset and the Wikipedia dataset. We can observe that the objective values descend quickly along with the iterations. Specifically, the objective values decline rapidly over the first 50 iterations, and become stable after 300 iterations on the Pascal Sentence dataset and after 200 iterations on the Wikipedia dataset. These figures illustrate the excellent convergence of ILCMR and support the feasibility of the proposed derivable rank-*r* approximation.

4.8 Optimizer analysis

It is known that non-differentiable rank minimization can be replaced by the rank-*r* approximation, allowing the ILCMR

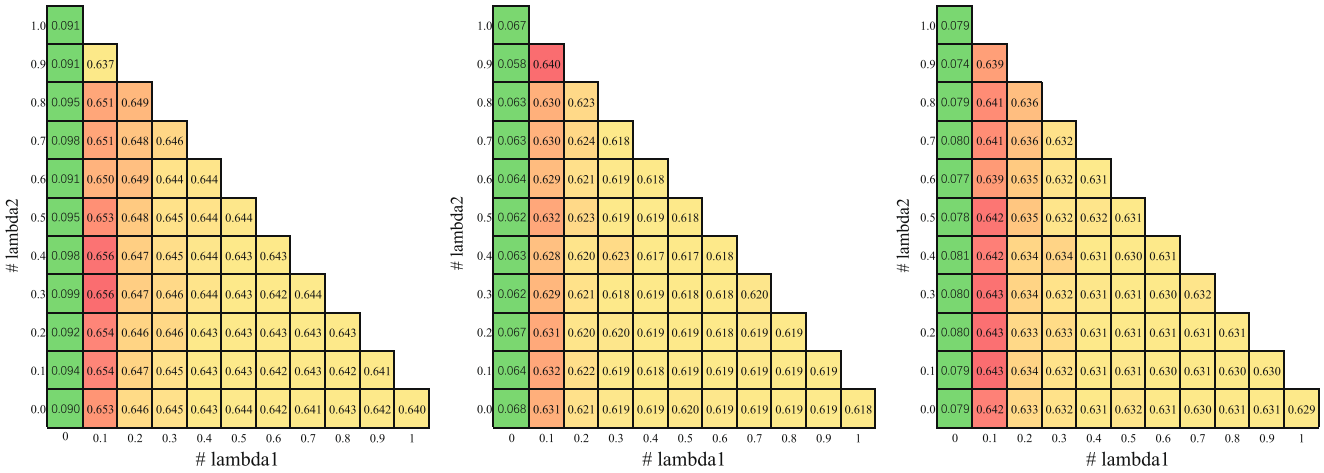


Fig. 3 The MAP versus different trade-off parameter combinations on the Pascal Sentence dataset

objective to be optimized with an automatic gradient. We compared the performances on the Wikipedia dataset by adopting different optimizers including SGD, Adam, Adagrad, Adamax, and RMSProp. As shown in Fig. 5, the results on the Img2Txt task are generally higher than those on the Txt2Img task for all the optimizers, which is consistent with the observations in Table 3. Besides, SGD achieves the best MAP on both tasks.

4.9 Comparison of text embedding models

Many natural language processing (NLP) technologies are available that can generate text embeddings for downstream tasks [2, 10, 30]. In this subsection, we compare the performances when utilizing different NLP models for text embedding, including the TF-IDF, ELMo [29], and BERT [6]. The results are listed in Table 9.

TF-IDF: is the abbreviation for 'term frequency - inverse document frequency', which is a bag-of-words method. The importance of each word is proportional to its count but inversely proportional to its frequency in the documents.

ELMo: is the abbreviation for 'embeddings from language models'. ELMo learns contextualized word embedding using long contexts. It trains a deep bi-directional LSTM and uses all the layers for prediction. We utilized a medium pre-trained ELMo model pre-trained on the 1 Billion Word Benchmark for this experiment.

BERT: is a transformer model pre-trained on a large corpus of English data in a self-supervised fashion. It was trained by using two masked language modelling and next sentence prediction objectives. We utilized the BERT-base-uncased model for this experiment.

As Table 9 shows, the TF-IDF feature obtains the best MAP results on the Pascal Sentence dataset for the current tasks. However, the performance obtained by using pre-trained BERT is not as satisfactory as we expected, possibly because the pre-trained models are generic and may not fit the new corpus. In contrast, the TF-IDF representation is generated based on the current dataset.

Table 8 The MAP results for different *spc* on the Pascal Sentence dataset. Best results are highlighted in bold

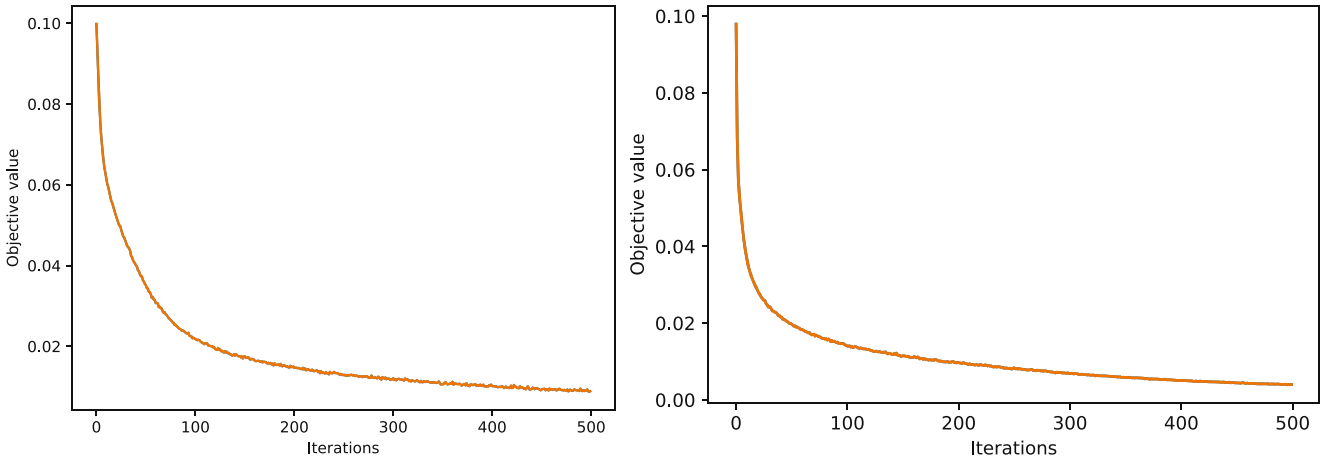
spc	Img2Txt	Txt2Img	Average
1	0.635	0.632	0.633
2	**0.643**	0.638	0.641
4	0.632	**0.654**	**0.643**
8	0.624	0.642	0.633
16	0.598	0.626	0.612
32	0.074	0.151	0.112

4.10 Retrieval examples

We provide some representative correct and incorrect retrieved results on the Pascal Sentence dataset in Fig. 6, which visualizes the top-5 retrieved results for each query in the two tasks (Img2Txt and Txt2Img). The failure cases mainly occur because some intersections exist between categories. For example, for the text query of 'Sofa', the pictures tagged as incorrect also contain sofas. Another example is the image query for 'Person', where the incorrecyly retrieved images include dining tables and bottles, while dining tables and bottles are present in the content of this image query as well.

4.11 Intra-class low-rank analysis

ILCMR constrains the intra-class low-rank structure and provides the rank-r approximation for optimization. To verify



(a) Pascal Sentence dataset (b) Wikipedia dataset

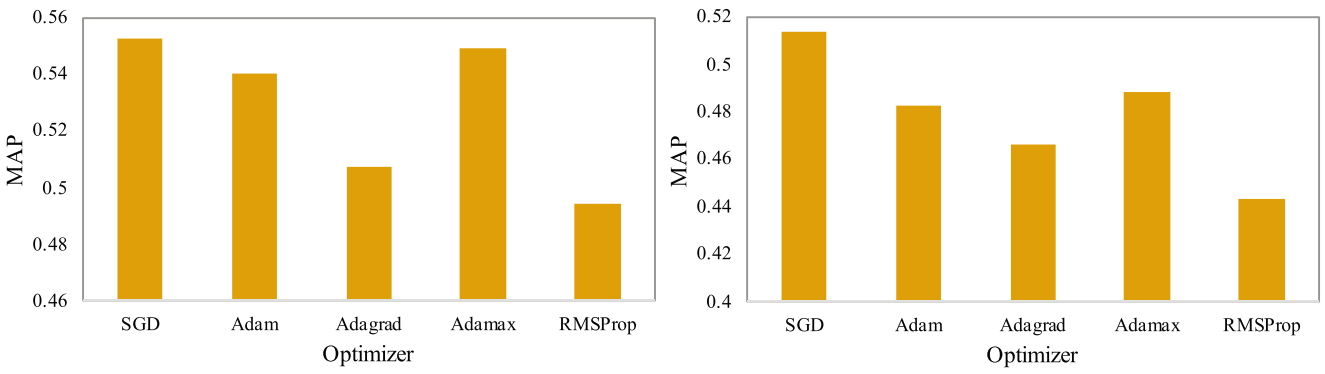
Fig. 4 The objective value versus the training iterations on the Pascal Sentence dataset and the Wikipedia dataset

the effectiveness of the rank-r approximation, we show the rank information of the inter-modal intra-class sample matrix. The rank of a matrix is calculated by counting the number of its nonzero singular values. Accordingly, we aggregate the inter-modal intra-class common representations to construct a matrix, and visualize its singular values as shown in Fig. 7.

Figure 7 shows the singular value comparisons on the Pascal Sentence dataset by ILCMR with and without (w/o) the intra-class low-rank constraint. The first 20 subfigures show the class-specific singular values from 20 categories, and the last subfigure shows the average singular values of all categories. In each subfigure, the green bars denote the singular values from the ILCMR without the intra-class low-rank constraint and the red bars denote the singular values from the ILCMR with the intra-class low-rank constraint. Note that the first singular value is expected to be larger and the remaining singular values are expected to be smaller, indicating a better low-rank approximation of the matrix.

From Fig 7, we can make the following discoveries.

- Compared with the scenario of the 'w/o constraint', the first singular values from the 'constrained' scenario tend to be larger in most cases (the rank of the inter-modal intra-class matrix will be lower). The reason is that there are two regularizations in ILCMR that induce the intra-modal and inter-modal intra-class data to be more relevant.
- It is easier to learn relevant intra-class common representations for some specific categories, such as '1', '13', and '19', where the rank of the intra-class matrix is nearly 1. However, the rank is even higher for other categories including '5', '7', '9', and '15', which correspond to the real categories of 'Bottle',



(a) Img2Txt task (b) Txt2Img task

Fig. 5 The performance of different optimizers on the Wikipedia dataset

Table 9 The MAP comparison by using different text embedding models on the Pascal Sentence dataset. Best results are highlighted in bold

Model	Img2Txt	Txt2Img	Average
ELMo	0.618	0.644	0.631
BERT	0.595	0.611	0.603
TF-IDF	**0.632**	**0.654**	**0.643**

'Car', 'Chair' and 'Person' respectively. The reason is that some overlaps occur in these categories leading to the high rank property, as was discussed in Section 4.10.

4.12 Inter-modal data fusion analysis

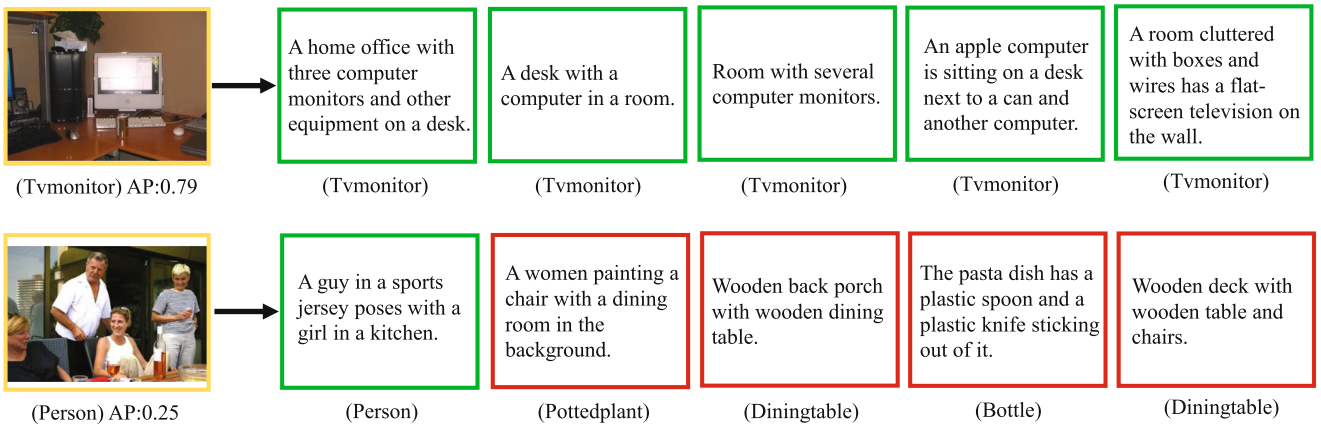
To reduce the cross-modal discrepancy, ILCMR constrains the intra-class low-rank structure of the intra-modal and inter-modal data. We visualize the common representations of the cross-modal data in Fig. 8 using t-SNE [25]. In Fig. 8, the first and third columns show visualizations of common representations learned by ILCMR without (w/o) intra-class low-rank regularization, while the second and fourth columns show visualizations of common representations learned by ILCMR with intra-class low-rank regularization. Among all the subfigures, a green '○' denotes an image sample, and a red '×' denotes a text sample. The number indicates the category. For example, 'Image_2' represents the test image samples from class 2, corresponding to the real class 'geography'.

From Fig. 8, we can make the following observations.

- For all the classes, the text and image within the same class are better fused by regularizing the intra-class low-rank structure, which boosts cross-modal data from the same category to be more relevant.
- The common representations for text learned by ILCMR both with and without constraints are almost identical.

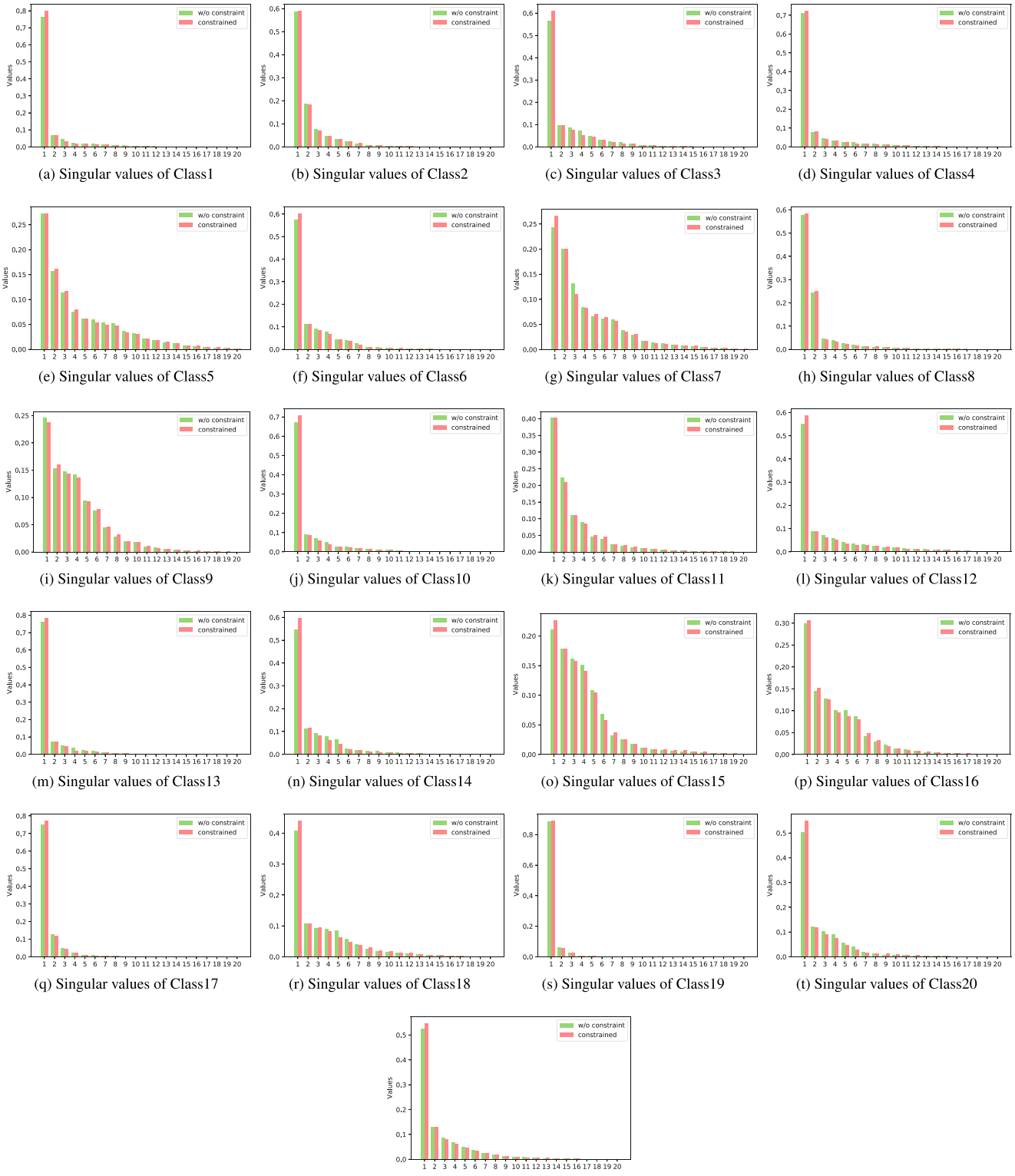


(a) Txt2Img



(b) Img2Txt

Fig. 6 Retrieval visualization of the top-5 returned results on Pascal Sentence dataset. The left hand of '→' are the queries with yellow rectangles, and the right hand are the corresponding top-5 retrieved results, where the right and wrong results are denoted by green and red rectangles, respectively. Their ground truth categories are given below the samples



(a) Singular values of Class1 (b) Singular values of Class2 (c) Singular values of Class3 (d) Singular values of Class4

(e) Singular values of Class5 (f) Singular values of Class6 (g) Singular values of Class7 (h) Singular values of Class8

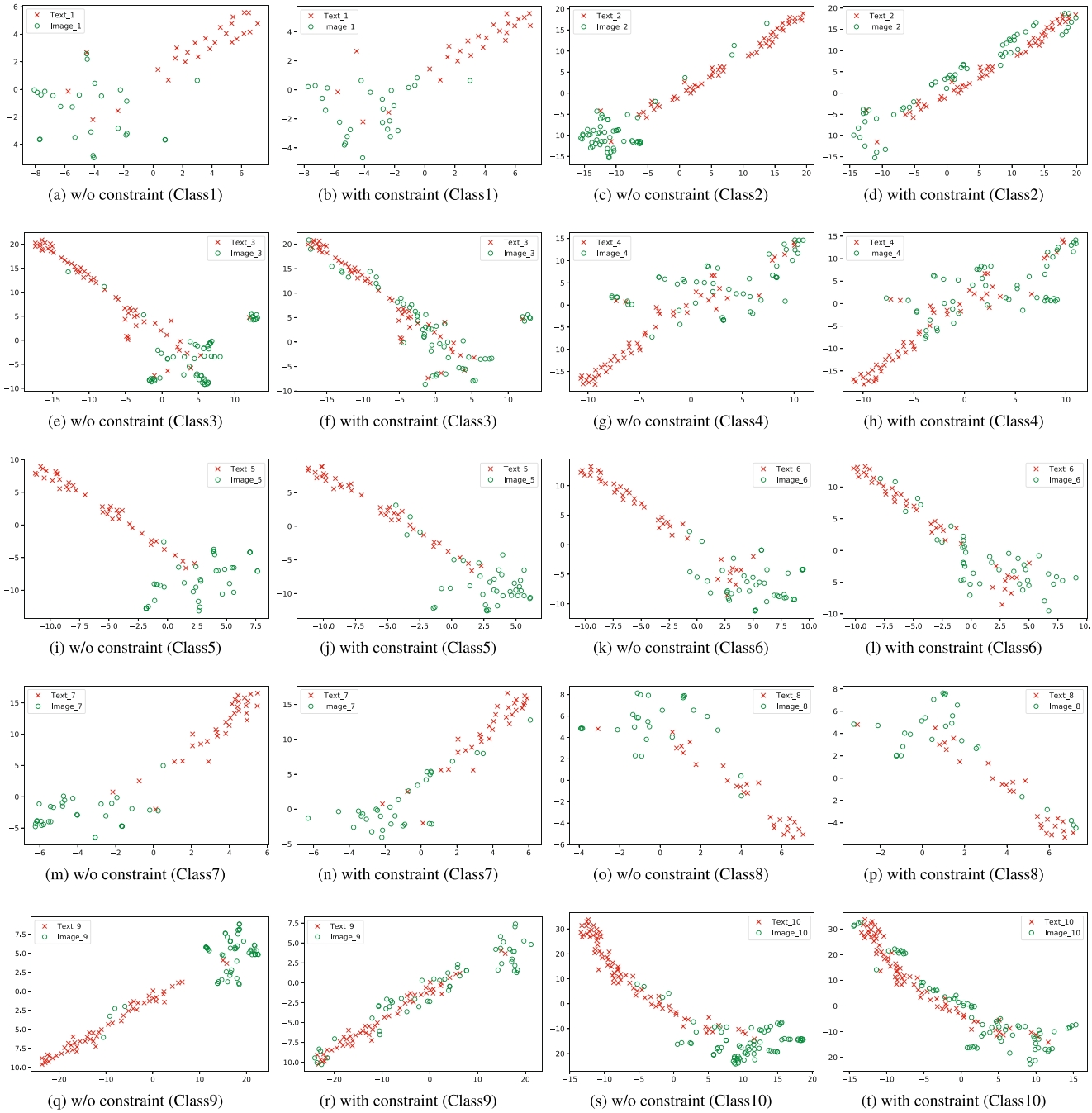
(i) Singular values of Class9 (j) Singular values of Class10 (k) Singular values of Class11 (l) Singular values of Class12

(m) Singular values of Class13 (n) Singular values of Class14 (o) Singular values of Class15 (p) Singular values of Class16

(q) Singular values of Class17 (r) Singular values of Class18 (s) Singular values of Class19 (t) Singular values of Class20

(u) Singular values of all classes

Fig. 7 The intra-class singular value comparison on test data of the Pascal Sentence dataset. The green bars are the result from ILCMR without (w/o) intra-class low-rank constraint, and the red bars are the result from ILCMR with constraint



(a) w/o constraint (Class1) (b) with constraint (Class1) (c) w/o constraint (Class2) (d) with constraint (Class2)

(e) w/o constraint (Class3) (f) with constraint (Class3) (g) w/o constraint (Class4) (h) with constraint (Class4)

(i) w/o constraint (Class5) (j) with constraint (Class5) (k) w/o constraint (Class6) (l) with constraint (Class6)

(m) w/o constraint (Class7) (n) with constraint (Class7) (o) w/o constraint (Class8) (p) with constraint (Class8)

(q) w/o constraint (Class9) (r) with constraint (Class9) (s) w/o constraint (Class10) (t) with constraint (Class10)

Fig. 8 The visualization of the class-specific test data on the Wikipedia dataset by using t-SNE. The first and third columns are the common representations learned by ILCMR without (w/o) intra-class low-rank constraint. The second and fourth columns are the common representations learned by ILCMR with intra-class low-rank constraint

This may occur because the text features are more explicit and have less uncertainty.

– Comparing the 'constrained' with the 'w/o constraint' scenarios, the image distributions change considerably. This indicates that the fusion process results in more transformations in the image space than in the text space.

5 Conclusion and future work

In this paper, we proposed a deep supervised and a semi-supervised cross-modal retrieval algorithms, named ILCMR and Semi-ILCMR, respectively. Based on our previous work [20], ILCMR imposes an additional intra-modal intra-class low-rank constraint in the common feature space, making

the data within the same class from the original space more relevant in the common space. In terms of optimizations, ILCMR uses the explicit gradient of the rank minimization approximation. We proposed the Semi-ILCMR algorithm, which requires only slight label supervision, by constraining the low-rank structure before the projection layers. The experiments verified the effectiveness of ILCMR and Semi-ILCMR. The additional analyses, including the intra-class low-rank property, the cross-modal data fusion, the batch strategy, and the impact of each regularization, showed the effectiveness of the proposed approaches. In future work, we plan to extend the work in three aspects. First, the model will be applied to multi-modal data to explore multi-modal retrieval and classification. Second, we plan to explore the low-rank structure of incomplete multi-modal data, where some modalities can be missing. Third, the retrieval scenario with unseen categories is also in the plan. More common features are expected to learn by constraining the low-rank structure.

Acknowledgements This work is supported by the National Natural Science Foundation of China (No. 62076073), the Guangdong Basic and Applied Basic Research Foundation (No. 2020A1515010616), the Guangdong Innovative Research Team Program (No. 2014ZT05G157).

Declarations

Competing interests The authors declare that they have no conflict of interest.

Appendix A: Derivation of (13)

We acquired the approximation of rank calculation as shown in (7):

$$rank(A) \approx \rho(A,r) = \frac{\sum_{i=r+1}^{s}\delta_i}{\sum_{i=1}^{s}\delta_i} = 1 - \frac{\sum_{i=1}^{r}\delta_i}{\sum_{i=1}^{s}\delta_i},$$

and combine it with the SVD knowledge as shown in (21):

$$A = U\Sigma V^T = \delta_1 u_1 v_1^T + \delta_2 u_2 v_2^T + \cdots + \delta_s u_s v_s^T. \quad (21)$$

We acquire the derivative in terms of A by the chain rule as shown in (13):

$$\frac{\partial rank(A)}{\partial A} \approx \frac{\partial \rho(A,r)}{\partial A} = \sum_{i=1}^{s}\frac{\partial \rho}{\partial \delta_i}\cdot\frac{\partial \delta_i}{\partial A}.$$

A.1 $\frac{\partial \rho}{\partial \delta_i}$

For $i \in [1,r]$, we have the following derivation from (7):

$$\frac{\partial \rho}{\partial \delta_i} = \frac{0 * \sum_{j=1}^{s}\delta_j - 1 * \sum_{j=r+1}^{s}\delta_j}{\left(\sum_{j=1}^{s}\delta_j\right)^2} = -\frac{\sum_{j=r+1}^{s}\delta_j}{\left(\sum_{j=1}^{s}\delta_j\right)^2}, \quad (22)$$

and for $i \in [r+1,s]$, we have the following derivation from (7):

$$\frac{\partial \rho}{\partial \delta_i} = \frac{1 * \sum_{j=1}^{s}\delta_j - 1 * \sum_{j=r+1}^{s}\delta_j}{\left(\sum_{j=1}^{s}\delta_j\right)^2} = \frac{\sum_{j=1}^{r}\delta_j}{\left(\sum_{j=1}^{s}\delta_j\right)^2}. \quad (23)$$

Therefore, we can derive (14):

$$\frac{\partial \rho}{\partial \delta_i} = \begin{cases} -\frac{\sum_{j=r+1}^{s}\delta_j}{\left(\sum_{j=1}^{s}\delta_j\right)^2}, & i = 1,\ldots,r \\ \frac{\sum_{j=1}^{r}\delta_j}{\left(\sum_{j=1}^{s}\delta_j\right)^2}, & i = r+1,\ldots,s \end{cases}.$$

A.2 $\frac{\partial \delta_i}{\partial A}$

From (21), we have the following equation:

$$\begin{aligned} &\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1d} \\ a_{21} & a_{22} & \cdots & a_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nd} \end{pmatrix} \\ &= \delta_1 \begin{pmatrix} u_{11}v_{11} & u_{11}v_{12} & \ldots & u_{11}v_{1d} \\ u_{12}v_{11} & u_{12}v_{12} & \ldots & u_{12}v_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ u_{1n}v_{11} & u_{1n}v_{12} & \ldots & u_{1n}v_{1d} \end{pmatrix} \\ &+\delta_2 \begin{pmatrix} u_{21}v_{21} & u_{21}v_{22} & \ldots & u_{21}v_{2d} \\ u_{22}v_{21} & u_{22}v_{22} & \ldots & u_{22}v_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ u_{2n}v_{21} & u_{2n}v_{22} & \ldots & u_{2n}v_{2d} \end{pmatrix} \\ &+\cdots+\delta_s \begin{pmatrix} u_{s1}v_{s1} & u_{s1}v_{s2} & \ldots & u_{s1}v_{sd} \\ u_{s2}v_{s1} & u_{s2}v_{s2} & \ldots & u_{s2}v_{sd} \\ \vdots & \vdots & \ddots & \vdots \\ u_{sn}v_{s1} & u_{sn}v_{s2} & \ldots & u_{sn}v_{1d} \end{pmatrix} \end{aligned}. \quad (24)$$

Equivalently, we have

$$\begin{cases} a_{11} = \sum_{i=1}^{s}\delta_i u_{i1}v_{i1} \\ a_{12} = \sum_{i=1}^{s}\delta_i u_{i1}v_{i2} \\ \vdots \\ a_{nd} = \sum_{i=1}^{s}\delta_i u_{in}v_{id} \end{cases}. \quad (25)$$

Then, we can obtain the derivatives as follows:

$$\begin{cases} \frac{\partial a_{11}}{\partial \delta_i} = u_{i1}v_{i1} \rightarrow \frac{\partial \delta_i}{\partial a_{11}} = \frac{1}{u_{i1}v_{i1}} \\ \frac{\partial a_{12}}{\partial \delta_i} = u_{i1}v_{i2} \rightarrow \frac{\partial \delta_i}{\partial a_{12}} = \frac{1}{u_{i1}v_{i2}} \\ \vdots \\ \frac{\partial a_{nd}}{\partial \delta_i} = u_{in}v_{id} \rightarrow \frac{\partial \delta_i}{\partial a_{nd}} = \frac{1}{u_{in}v_{id}} \end{cases}. \quad (26)$$

Therefore, the derivative in terms of A can be obtained as follows:

$$\frac{\partial \delta_i}{\partial A} = 1./\begin{pmatrix} u_{i1}v_{i1} & u_{i1}v_{i2} & \dots & u_{i1}v_{id} \\ u_{i2}v_{i1} & u_{i2}v_{i2} & \dots & u_{i2}v_{id} \\ \vdots & \vdots & \ddots & \vdots \\ u_{in}v_{i1} & u_{in}v_{i2} & \dots & u_{in}v_{id} \end{pmatrix} = 1./u_i v_i^T, \quad (27)$$

where './' denotes the elementwise division operation.

References

1. Cao W, Lin Q, He Z, He Z (2019) Hybrid representation learning for cross-modal retrieval. Neurocomputing 345:45–57
2. Catelli R, Casola V, De Pietro G, Fujita H, Esposito M (2021) Combining contextualized word representation and sub-document level analysis through bi-lstm+ crf architecture for clinical de-identification. Knowl-Based Syst 213:106649
3. Cheng Q, Gu X (2020) Bridging multimedia heterogeneity gap via graph representation learning for cross-modal retrieval. Neural Networks
4. Chua TS, Tang J, Hong R, Li H, Luo Z, Zheng Y (2009) 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, pp 1–9. https://doi.org/10.1145/1646396.1646452
5. Deng T, Ye D, Ma R, Fujita H, Xiong L (2020) Low-rank local tangent space embedding for subspace clustering. Inf Sci 508:1–21
6. Devlin J, Chang MW, Lee K, Toutanova K (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp 4171–4186
7. Ding Z, Fu Y (2018) Deep transfer low-rank coding for cross-domain learning. IEEE Trans Neural Netw Learn Syst 30(6):1768–1779. https://doi.org/10.1109/TNNLS.2018.2874567
8. Ding Z, Shao M, Fu Y (2018) Generative zero-shot learning via low-rank embedded semantic dictionary. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2018.2867870
9. Eckart C, Young G (1939) A principal axis transformation for non-hermitian matrices. Bull Am Math Soc 45(2):118–121
10. Esposito M, Damiano E, Minutolo A, De Pietro G, Fujita H (2020) Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering. Inf Sci 514:88–105
11. Fang X, Han N, Wu J, Xu Y, Yang J, Wong WK, Li X (2018) Approximate low-rank projection learning for feature extraction. IEEE Trans Neural Netw Learn Syst 29(11):5228–5241. https://doi.org/10.1109/TNNLS.2018.2796133
12. Fei L, Xu Y, Fang X, Yang J (2017) Low rank representation with adaptive distance penalty for semi-supervised subspace classification. Pattern Recogn 67:252–262. https://doi.org/10.1016/j.patcog.2017.02.017
13. Feng F, Wang X, Li R (2014) Cross-modal retrieval with correspondence autoencoder. In: Proceedings of the 22nd ACM International Conference on Multimedia. ACM, pp 7–16. https://doi.org/10.1145/2647868.2654902
14. Golub GH, Hoffman A, Stewart GW (1987) A generalization of the eckart-young-mirsky matrix approximation theorem. Linear Algebra Appl 88:317–327
15. Hardoon DR, Szedmak S, Shawe-Taylor J (2004) Canonical correlation analysis: an overview with application to learning methods. Neural Comput 16(12):2639–2664. https://doi.org/10.1162/0899766042321814
16. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778
17. He Y, Xiang S, Kang C, Wang J, Pan C (2016) Cross-modal retrieval via deep and bidirectional representation learning. IEEE Trans Multimed 18(7):1363–1377. https://doi.org/10.1109/TMM.2016.2558463
18. Hu P, Zhen L, Peng D, Liu P (2019) Scalable deep multimodal learning for cross-modal retrieval. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 635–644
19. Kang P, Fang X, Zhang W, Teng S, Fei L, Xu Y, Zheng Y (2018) Supervised group sparse representation via intra-class low-rank constraint. In: Chinese Conference on Biometric Recognition. Springer, pp 206–213. https://doi.org/10.1007/978-3-319-97909-0_22
20. Kang P, Lin Z, Yang Z, Fang X, Li Q, Liu W (2019) Deep semantic space with intra-class low-rank constraint for cross-modal retrieval. In: Proceedings of the 2019 on International Conference on Multimedia Retrieval. ACM, pp 226–234. https://doi.org/10.1145/3323873.3325029
21. Lezama J, Qiu Q, Musé P, Sapiro G (2018) Ole: Orthogonal low-rank embedding-a plug and play geometric loss for deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 8109–8118
22. Li C, Deng C, Wang L, Xie D, Liu X (2019) Coupled cyclegan: Unsupervised hashing network for cross-modal retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 33, pp 176–183
23. Li K, Qi GJ, Ye J, Hua KA (2016) Linear subspace ranking hashing for cross-modal retrieval. IEEE Trans Pattern Anal Mach Intell 39(9):1825–1838. https://doi.org/10.1109/TPAMI.2016.2610969
24. Liu H, Feng Y, Zhou M, Qiang B (2020) Semantic ranking structure preserving for cross-modal retrieval. Appl Intell:1–11
25. Maaten Lvd, Hinton G (2008) Visualizing data using t-sne. J Mach Learn Res 9:2579–2605
26. Peng Y, Huang X, Qi J (2016) Cross-media shared representation by hierarchical learning with multiple deep networks. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, pp 3846–3853
27. Peng Y, Huang X, Zhao Y (2017) An overview of cross-media retrieval: concepts, methodologies, benchmarks, and challenges. IEEE Trans Circ Syst Video Technol 28(9):2372–2385. https://doi.org/10.1109/TCSVT.2017.2705068
28. Peng Y, Qi J, Huang X, Yuan Y (2017) Ccl: Cross-modal correlation learning with multigrained fusion by hierarchical network. IEEE Trans Multimed 20(2):405–420. https://doi.org/10.1109/TMM.2017.2742704
29. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp 2227–2237

30. Pota M, Marulli F, Esposito M, De Pietro G, Fujita H (2019) Multilingual pos tagging by a composite deep architecture based on character-level features and on-the-fly enriched word embeddings. Knowl-Based Syst 164:309–323
31. Qi J, Peng Y (2018) Cross-modal bidirectional translation via reinforcement learning. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, pp 2630–2636
32. Qiang H, Wan Y, Liu Z, Xiang L, Meng X (2020) Discriminative deep asymmetric supervised hashing for cross-modal retrieval. Knowl-Based Syst 204:106188
33. Qiu Q, Sapiro G (2015) Learning transformations for clustering and classification. J Mach Learn Res 16(1):187–225
34. Rashtchian C, Young P, Hodosh M, Hockenmaier J (2010) 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, pp 139–147
35. Rasiwasia N, Costa Pereira J, Coviello E, Doyle G, Lanckriet GR, Levy R, Vasconcelos N (2010) A new approach to cross-modal multimedia retrieval. In: Proceedings of the 18th ACM International Conference on Multimedia. ACM, pp 251–260. https://doi.org/10.1145/1873951.1873987
36. Shang F, Zhang H, Zhu L, Sun J (2019) Adversarial cross-modal retrieval based on dictionary learning. Neurocomputing 355:93–104
37. Shen TH, Liu L, Yang Y, Xu X, Huang Z, Shen F, Hong R (2020) Exploiting subspace relation in semantic labels for cross-modal hashing. IEEE Trans Knowl Data Eng:1–1
38. Situ R, Yang Z, Lv J, Li Q, Liu W (2018) Cross-modal event retrieval: a dataset and a baseline using deep semantic learning. In: Pacific Rim Conference on Multimedia. Springer, pp 147–157. https://doi.org/10.1007/978-3-030-00767-6_14
39. Wang B, Yang Y, Xu X, Hanjalic A, Shen HT (2017) Adversarial cross-modal retrieval. In: Proceedings of the 25th ACM International Conference on Multimedia, pp 154–162. https://doi.org/10.1145/3123266.3123326
40. Wang D, Gao XB, Wang X, He L (2018) Label consistent matrix factorization hashing for large-scale cross-modal similarity search. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2018.2861000
41. Wang X, Hu P, Zhen L, Peng D (2021) Drsl: Deep relational similarity learning for cross-modal retrieval. Inf Sci 546:298–311. https://doi.org/10.1016/j.ins.2020.08.009
42. Wei Y, Zhao Y, Lu C, Wei S, Liu L, Zhu Z, Yan S (2016) Cross-modal retrieval with cnn visual features: a new baseline. IEEE Trans Cybern 47(2):449–460. https://doi.org/10.1109/TCYB.2016.2519449
43. Wen J, Xu Y, Liu H (2018) Incomplete multiview spectral clustering with adaptive graph learning. IEEE Transactions on Cybernetics. https://doi.org/10.1109/TCYB.2018.2884715
44. Wu F, Jing XY, Wu Z, Ji Y, Dong X, Luo X, Huang Q, Wang R (2020) Modality-specific and shared generative adversarial network for cross-modal retrieval. Pattern Recogn:107335
45. Xiao Q, Dai J, Luo J, Fujita H (2019) Multi-view manifold regularized learning-based method for prioritizing candidate disease mirnas. Knowl-Based Syst 175:118–129
46. Xu X, He L, Lu H, Gao L, Ji Y (2019) Deep adversarial metric learning for cross-modal retrieval. World Wide Web 22(2):657–672
47. Yan F, Mikolajczyk K (2015) Deep correlation for matching images and text. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3441–3450
48. Yang X, Jiang X, Tian C, Wang P, Zhou F, Fujita H (2020) Inverse projection group sparse representation for tumor classification: a low rank variation dictionary approach. Knowl-Based Syst 196:105768
49. Yang Z, Lin Z, Kang P, Lv J, Li Q, Liu W (2020) Learning shared semantic space with correlation alignment for cross-modal event retrieval. ACM Trans Multimed Comput Commun Appl 16(1):1–22. https://doi.org/10.1145/3374754
50. Zhai X, Peng Y, Xiao J (2013) Learning cross-media joint representation with sparse and semisupervised regularization. IEEE Trans Circ Syst Video Technol 24(6):965–978. https://doi.org/10.1109/TCSVT.2013.2276704
51. Zhan S, Wu J, Han N, Wen J, Fang X (2019) Unsupervised feature extraction by low-rank and sparsity preserving embedding. Neural Netw 109:56–66. https://doi.org/10.1016/j.neunet.2018. 10.001
52. Zhang L, Ma B, Li G, Huang Q, Tian Q (2017) Generalized semi-supervised and structured subspace learning for cross-modal retrieval. IEEE Trans Multimed 20(1):128–141. https://doi.org/10.1109/TMM.2017.2723841
53. Zhang X (2017) Matrix analysis and applications. Cambridge University Press
54. Zhen L, Hu P, Wang X, Peng D (2019) Deep supervised cross-modal retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 10394–10403
55. Zheng Z, Zheng L, Garrett M, Yang Y, Xu M, Shen YD (2020) Dual-path convolutional image-text embeddings with instance loss. ACM Trans Multimed Comput Commun Appl 16(2):1–23. https://doi.org/10.1145/3383184

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



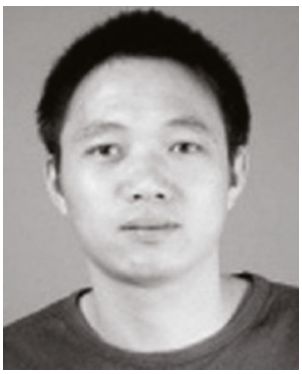
Peipei Kang received the M.S. degree in School of Computer Science and Technology from Guangdong University of Technology, Guangzhou, China. She is currently pursuing the Ph.D. degree in School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China. Her current research interests include pattern recognition, machine learning and multimedia retrieval.



Zehang Lin is a Research Assistant at The Hong Kong Polytechnic University Hong Kong. He received the master's degree from Guangdong University of Technology, Guangdong, China in 2019. His research interests include computer vision, deep learning, and transfer learning.



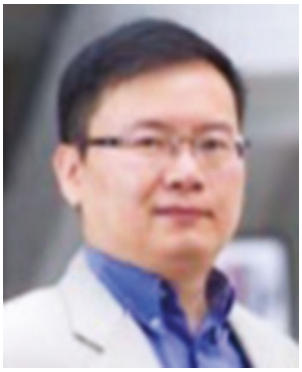
Zhenguo Yang is an Associate Professor at Guangdong University of Technology. He was a Postdoctoral Fellow at Guangdong University of Technology from 2017 to 2019. He received his Ph.D. degree in Computer Science from City University of Hong Kong, in 2017, and the M.E. degree in Computer Science from Zhejiang Normal University, China, in 2013, and the B.E. degree in Computer Science from Shandong Normal University, China, in 2010. His research interests include online event detection, and domain adaptation, etc.



Xiaozhao Fang received the M.S. degree and the Ph.D. degree in computer science and technology from the Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China, in 2008 and 2016, respectively. He is currently with the Department of Automation, Guangdong University of Technology, Guangzhou, China. His current research interests include pattern recognition, data mining, and machine learning.



Alexander M. Bronstein is a faculty in the Department of Computer Science at the Technion Israel Institute of Technology and a Principal Engineer at Intel. His research interests include numerical geometry, computer vision, and machine learning. Prof. Bronstein has authored over 100 publications in leading journals and conferences, over 30 patents and patent applications, the research monograph "Numerical geometry of nonrigid shapes", and edited several books. Highlights of his research were featured in CNN, SIAM News, Wired. Prof. Bronstein is a Fellow of the IEEE for his contribution to 3D imaging and geometry processing. In addition to his academic activity, he co-founded and served as Vice President of technology in the Silicon Valley start-up company Novafora (2005-2009), and was a co-founder and one of the main inventors and developers of the 3D sensing technology in the Israeli startup Invision, subsequently acquired by Intel in 2012. Prof. Bronstein's technology is now the core of the Intel RealSense 3D camera integrated into a variety of consumer electronic products. He is also a co-founder of the Israeli video search startup Videocites where he serves as Chief Scientist.



Qing Li is a Chair Professor at The Hong Kong Polytechnic University. He received the B.Eng. degree from Hunan University, China, and the M.Sc. and Ph.D. degrees from the University of Southern California, Los Angeles, USA, all in Computer Science. His research interests include database modeling, Web services, multimedia retrieval and management, and e-learning systems. He has authored over 300 papers in technical journals and international conferences. He is actively involved in the research community by serving as a journal reviewer, program committee chair/co-chair, and organizer/co-organizer of several international conferences. Prof. Li serves as the Chairman of the Hong Kong Web Society, a Councilor of the Database Society of China Computer Federation, and a Steering Committee Member of DASFAA, ICWL, WAIM, and International WISE Society.



Wenyin Liu is currently a Professor in School of Computer Science and Technology, Guangdong University of Technology. He was Deputy Director of Multimedia software Engineering Research Centre at the City University of Hong Kong from 2013 to 2016, an assistant professor in the Department of Computer Science at the City University of Hong Kong between from 2002 to 2012, and a full time researcher at Microsoft Research China/Asia from 1999 to 2001. His research interests include anti-phishing, Web identity authentication and management, multimedia information retrieval, question answering, graphics recognition, and performance evaluation. He has a BEng and MEng in computer science from Tsinghua University, Beijing and a DSc from the Technion, Israel Institute of Technology, Haifa. In 2003, he was awarded the International Conference on Document Analysis and Recognition Outstanding Young Researcher Award by the International Association for Pattern Recognition (IAPR). He had been TC10 chair of IAPR for 2006-2010. He had been on the editorial boards of the International Journal of Document Analysis and Recognition (IJDAR) from 2006 to 2011 and the IET Computer Vision journal from 2011-2012. He is a Fellow of IAPR and a senior member of IEEE.

Affiliations

Peipei Kang[1,2] · Zehang Lin[3] · Zhenguo Yang[1] · Xiaozhao Fang[4] · Alexander M. Bronstein[2] · Qing Li[3] · Wenyin Liu[1]

✉ Wenyin Liu
liuwy@gdut.edu.cn

Peipei Kang
ppkanggdut@126.com

Zehang Lin
zehlin@polyu.edu.hk

Xiaozhao Fang
xzhfang168@126.com

Alexander M. Bronstein
bron@cs.technion.ac.il

Qing Li
csqli@comp.polyu.edu.hk

1 School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China

2 Computer Science Department, Technion, Haifa, Israel

3 Department of Computing, Hong Kong Polytechnic University, Hong Kong, China

4 Department of Automation, Guangdong University of Technology, Guangzhou, China