

Bridging multimedia heterogeneity gap via Graph Representation Learning for cross-modal retrieval

Qingrong Cheng, Xiaodong Gu*

Department of Electronic Engineering, Fudan University, 200433, Shanghai, China



ARTICLE INFO

Article history:

Received 2 March 2020

Received in revised form 10 November 2020

Accepted 23 November 2020

Available online 28 November 2020

Keywords:

Cross-modal retrieval

Common space learning

Cross-modal graph

Graph representation learning network

Feature transfer learning network

Graph embedding

ABSTRACT

Information retrieval among different modalities becomes a significant issue with many promising applications. However, inconsistent feature representation of various multimedia data causes the “heterogeneity gap” among various modalities, which is a challenge in cross-modal retrieval. For bridging the “heterogeneity gap,” the popular methods attempt to project the original data into a common representation space, which needs great fitting ability of the model. To address the above issue, we propose a novel Graph Representation Learning (GRL) method for bridging the heterogeneity gap, which does not project the original feature into an aligned representation space but adopts a cross-modal graph to link different modalities. The GRL approach consists of two subnetworks, Feature Transfer Learning Network (FTLN) and Graph Representation Learning Network (GRLN). Firstly, FTLN model finds a latent space for each modality, where the cosine similarity is suitable to describe their similarity. Then, we build a cross-modal graph to reconstruct the original data and their relationships. Finally, we abandon the features in the latent space and turn into embedding the graph vertexes into a common representation space directly. During the process, the proposed Graph Representation Learning method bypasses the most challenging issue by utilizing a cross-modal graph as a bridge to link the “heterogeneity gap” among different modalities. This attempt utilizes a cross-modal graph as an intermediary agent to bridge the “heterogeneity gap” in cross-modal retrieval, which is simple but effective. Extensive experiment results on six widely-used datasets indicate that the proposed GRL outperforms other state-of-the-art cross-modal retrieval methods.

© 2020 Elsevier Ltd. All rights reserved.

1. Introduction

Multimedia data have greatly enriched people's life with the dramatic development of multimedia devices and information transmission technology. In recent years, people have witnessed explosive growth in various types of multimedia data, such as image, text, audio, video, and 3D model. These different types of multimedia data usually deliver similar semantic information during the transportation of social networks, which is consistent with the research of cognitive science. For example, our human brain can apperceive the outside environment by multiple sensory organs, such as the ear and eye (McGurk & MacDonald, 1976). Therefore, it is of great realistic significance to analyze the multi-modal data and exploit the semantic relationship among various modalities. Cross-modal retrieval becomes a highlighted research topic to analyze the relationship between different types of multimedia data. The early single-modal retrieval performs the retrieval task in a single modality, such as retrieving image

by an image (Datta, Joshi, Li, & Wang, 2008), while the cross-modal retrieval allows the modality of the retrieved results and the query to be different, such as retrieving image by a text or retrieving text by an image (Peng, Huang, & Zhao, 2017). However, there exists a vast “heterogeneity gap” among different modalities, which brings great difficulties for cross-modal retrieval. The “heterogeneity gap” means that the feature spaces of different modalities are different, so the similarity of different modal instances cannot be measured directly.

For bridging the “heterogeneity gap”, most existing cross-modal retrieval methods aim at mapping the features of different modalities into a feature-shared subspace. This idea is based on a hypothesis that there exists a common semantic space, where the semantic similarity measurement is suitable for all instances of all modalities. Therefore, many cross-modal retrieval methods have been proposed to learn the aligned representation of various modalities. For example, early works, such as Canonical Correlation Analysis (CCA) (Hotelling, 1936) and Cross-modal Factor Analysis (CFA) (Li, Dimitrova, Li, & Sethi, 2003), mainly adopt linear projection to fitting the aligned representation learning process. This branch of cross-modal retrieval methods is classified as traditional method. However, the aligned representation

* Corresponding author.

E-mail addresses: qrcheng17@fudan.edu.cn (Q. Cheng), xdgu@fudan.edu.cn (X. Gu).

learning process is highly complex and nonlinear, which is hard for traditional methods.

In recent years, inspired by the great success of deep learning, many Deep Neural Networks (DNNs) based methods have been proposed to solve cross-modal analysis issue. Compared with traditional cross-modal retrieval methods, these approaches show superior performance in cross-modal retrieval tasks because of the enormous capacity of nonlinear learning. Besides, adversarial learning gradually obtains great concern since Goodfellow, Pouget-Abadie, Mirza, et al. (2014) proposed Generative Adversarial Networks (GANs). Motivated by the progress of GANs, many researchers attempt to apply the adversarial learning mechanism into DNNs based cross-modal retrieval models. Adversarial learning-based methods (Xu et al., 2018; Zhang & Peng, 2019) achieve excellent retrieval accuracy compared with other approaches. Furthermore, introducing both adversarial learning and DNNs into aligned representation learning becomes a popular strategy in cross-modal retrieval. However, because adversarial learning is not suitable for graph embedding strategy, we do not introduce this idea in the proposed method. Besides, hashing coding-based methods also achieve remarkable cross-modal retrieval performance, such as Deng et al. (2019), Deng, Yang, Liu, and Tao (2019), Yang et al. (2018) and Zhang, Lai, and Feng (2018). Their main idea is to learn deep representations for various modalities and then encode the features into a semantic shared hamming space by hashing coding. Various technologies can be taken into account in learning deep representations, such as adversarial learning, and generative network. Hashing coding-based methods are an essential branch in large-scale information retrieval.

The graph model shows high potential capacity in machine learning because of its characteristic data structure, which can present the connections among different data. For example, the recommendation system in social networks aims at mining the preferred contents by studying the users' previous action networks. Therefore, some researchers attempt to introduce graph model into the aligned representation learning. For example, Wu, Wang, and Huang (2018) introduce a semantic graph as additional information to preserve the local and global semantic structure during the aligned representation learning process. Besides, Yu et al. (2018) realize the cross-modal retrieval by modeling the text with graph convolutional network, and Wang et al. (2018) introduce graph regularization into objective function in modality-independent feature learning for cross-modal retrieval. These methods adopt graph as an additional constraint for aligned representation learning, which also improves the ability of shared representation learning.

Inspired by their works, in this paper, we propose a Graph Representation Learning (GRL) approach to accomplish shared representation learning for cross-modal retrieval. On the constructed cross-modal graph, each vertex represents a multimedia instance, and the connections among different vertexes indicate the similarity information of the original multimedia data. The GRL views the cross-modal graph as a bridge to link the "heterogeneity gap" among different modalities and then adopts a graph-embedding layer to project different vertexes to low dimensional representations. The framework of the proposed model is presented in Fig. 1. Specifically, we adopt the pre-trained models to extract the features of different modalities. Then, an FTLN model for each modality is applied to transfer the original features into a latent space. This step is designed to make the features suitable for graph construction. The FT LN models are optimized by the triplet loss function with cosine distance. At last, we adopt a graph-embedding layer to learn the modality-invariant features for all instances. In optimizing of the proposed GRLN, the overall objective function contains two parts, supervised learning loss

and unsupervised learning loss. Category constraint supervises our model to learn more discriminating features in the label space. During the unsupervised learning process, the graph local structure optimization loss is adopted to keep the neighboring instances close in the embedded space.

The previous works aim to project the original multimedia data into aligned representation space by linear learning or deep neural networks. However, the proposed method does not directly process the original data but reconstructs the data and their relationships into a cross-modal graph and then utilizes the graph structure to learn the aligned representation. The proposed GRL utilizes a cross-modal graph as a bridge to link the "heterogeneity gap". Compared with previous works, the proposed method is more straightforward but more effective, which does not need plenty of computing resources and converges rapidly.

Compared with previous researches, the main contributions of our work can be summarized as follows.

- A novel cross-modal retrieval method named Graph Representation Learning is proposed to bridge the vast "heterogeneity gap" among various modalities. We represent the original multimedia data by a cross-modal graph, in which each vertex denotes an original multimedia instance, and the connections indicate the semantic similarity information. Then, we adopt a node-to-vector strategy to embed every node into a low-dimensional vector. Because we do not process the original data, the proposed method can effectively project the original various modal data into an aligned representation space, which is an effective attempt. Sufficient experimental results indicate that the proposed method obtains the best performance and makes significant improvements compared to previous works.
- We propose a pre-aligning framework named Feature Transfer Learning Network before constructing the cross-modal graph, aiming to transfer the original feature to an aligned latent space. In the latent space, cosine similarity is adopted to construct the k-nearest neighbor graph. In the experiment, we adopt different Feature Transfer Learning Networks to verify the feasibility of the proposed pre-aligning strategy. The experimental results show that this is an efficient mechanism for improving cross-modal retrieval performance.
- For optimizing the GRLN model, adopting category constraint with AMSoftmax function (Wang, Cheng, Liu, & Liu, 2018) ensures the embedded representation is discriminative in the label space. Also, a simple but effective unsupervised objective function is proposed to optimize the local graph structure. This function adopts Laplacian Eigenmaps as the graph optimization term, which aims at keeping the embedding of two nodes close if they are similar in the original feature space.

The structure of this paper is organized as follows. The second section draws the related work in cross-modal retrieval. The third section presents the details of the proposed GRL method. The detailed experimental results and analyzes are shown in the fourth section. An overall conclusion of the paper is in the final section.

2. Related work

In this section, we review the related works of cross-modal retrieval. As mentioned before, the most challenging problem of cross-modal retrieval is how to bridge the "heterogeneity gap" among different modalities. Therefore, the mainstream of cross-modal retrieval approaches is common space learning, which can represent these different modal data in semantic shared

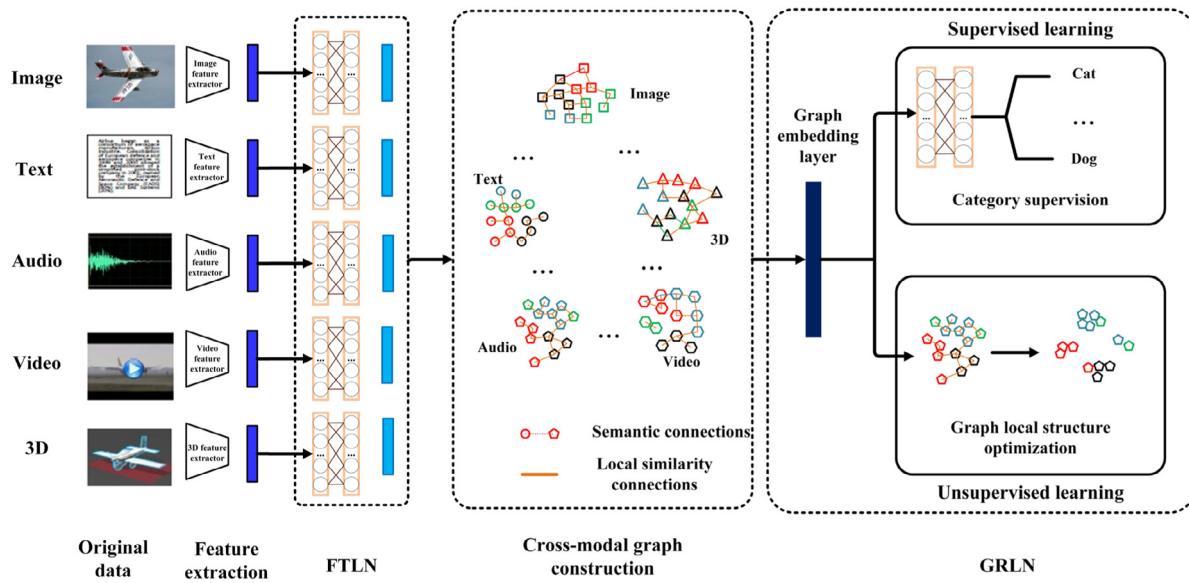


Fig. 1. The overall framework of the proposed Graph Representation Learning method for cross-modal retrieval.

representation space. The similarity among different instances can be calculated directly by common similarity measurements, such as cosine distance and Euclidean distance. Existing cross-modal retrieval methods include traditional methods, DNN-based methods, and graph regularization based methods.

2.1. Traditional methods

Traditional methods mainly aim at learning a linear projection matrix for feature shared space, which is the basic paradigm and precedent of common space learning. For instance, Canonical Correlation Analysis (CCA) (Hotelling, 1936) proposed by Hardoon et al. is one of the most representative works in cross-modal retrieval. CCA learns a shared space by maximizing the pairwise correlations between different modal data. As an early classical work, CCA has many variants such as Kernel CCA (KCCA) (Hardoon, Szedm'ak, & Shawe-Taylor, 2004), Semantic Correlation Matching (SCM) (Pereira et al., 2013), Deep CCA (DCCA) (Andrew, Arora, Bilmes, & Livescu, 2010), Multi-view CCA (Gong, Ke, Isard, & Lazebnik, 2014), and multi-label CCA (ml-CCA) (Ranjan, Rasiwasia, & Jawahar, 2015), which are the most popular baseline models in cross-modal retrieval. To be specific, the original CCA is an unsupervised method for cross-modal retrieval, which does not utilize the semantic information. Therefore, the researchers attempt to introduce semantic supervision into the common representation learning model, such as Semantic Correlation Matching (SCM) and multi-label CCA (ml-CCA) (Ranjan et al., 2015). Pereira et al. propose SCM, which combines traditional correlation matching (CM) and semantic matching (SM). Multi-label CCA (Hotelling, 1936), an extended version of CCA, is designed to deal with multiple label cross-modal retrieval problem. Multi-view CCA (Gong et al., 2014) takes ground-truth semantic labels as the third view of additional supervision. Analogously, another classical statistical correlation analysis for cross-modal retrieval is Cross-modal Factor Analysis (CFA) (Li et al., 2003), which directly minimizes the Frobenius norm of different modalities in the aligned representation space. Wang, He, Wang, Wang, and Tan (2015) solve the cross-modal retrieval problem by introducing a joint feature selection and subspace learning (JFSL) method, combining subspace learning for different modalities and norms for coupled features selection. In recent years, researchers try to incorporate extensive information into aligned representation learning in order to learn a more

robust linear projection, such as Partial Least Squares (Rosipal & Kramer, 2006) and adaptive semantic hierarchy (Kang, Xiang, Liao, Xu, & Pan, 2015). Deng, Tang, Yan, Liu, and Gao (2015) proposed discriminative dictionary learning with a common label alignment based method to realize cross-modal retrieval. Due to the perfection of linear projection based common representation learning, traditional methods have made significant progress. However, due to the limited fitting capacity of linear projection, the problematic “heterogeneity gap” cannot be solved directly by linear projection.

2.2. DNNs-based methods

With the advancement of deep learning, deep neural networks (DNNs) have shown significantly superiority in many challenging tasks, such as image classification (Ciresan, Meier, Masci, Maria Gambardella, & Schmidhuber, 2011) and object detection (Ren, He, Girshick, & Sun, 2015). DNNs have a great non-linear fitting capacity, which makes it successfully applied in cross-modal retrieval. For example, Andrew et al. (2010) introduce DNNs into the traditional CCA, named DCCA, to improve the original CCA's performance. Srivastava and Salakhutdinov (2012) introduce a deep belief network (Multimodal DBN) into cross-modal retrieval architecture for learning a joint representation of multimodal data. Wei et al. (2017) fulfill cross-modal retrieval by adopting DNN-based visual features extracted by the pre-trained CNN model large-scale image dataset, which shows the superiority of deep feature in cross-modal retrieval. Peng, Huang, and Qi (2016) propose a two-stage learning strategy, cross-media multiple deep networks (CMDN), to exploit the complex correlation of different modalities. The first stage generates the separate representation of each modal instance, and the second stage learns the shared representation by a deep two-level network in a hierarchical manner. The cross-modal hybrid transfer network (CHTN) (Huang, Peng, & Yuan, 2017) introduces transfer learning to improve cross-modal retrieval performance. Besides, MNiL (Zhang, Ma, Li, Huang, & Tian, 2017) is a DNN-based method that combines LSTM and ResNet for large-scale cross-modal retrieval.

The DNNs based methods mainly contain two types in cross-media retrieval. The first type directly makes use of the great non-linear ability of projection to realize common space learning, while the second type utilizes the capacity of DNNs and

combines other learning strategies, such as adversarial learning. As is known, adversarial learning has been successfully applied in various tasks with the development of generative adversarial networks. Under this circumstances, many researchers attempt to apply this effective mechanism in the cross-modal retrieval model, such as Gu, Cai, Joty, Niu, and Wang (2018), Huang, Peng, and Yuan (2020), Tzeng, Hoffman, Saenko, and Darrell (2017) and Wang, Yang, Xu, Hanjalic, and Shen (2017). Specifically, Wang et al. (2017) combine the idea of adversarial learning and triplet constraint to learn the modality-invariant representations for different modal data, named Adversarial Cross-modal Retrieval (ACMR). The modal-adversarial hybrid transfer network (MHTN) proposed by Huang et al. (2020) is also based on adversarial learning. MHTN consists of two subnetworks, modal-sharing knowledge transfer subnetwork and modal-adversarial semantic learning subnetwork. Adversarial discriminative domain adaptation (ADDA) (Tzeng et al., 2017) combines discriminative modeling and GAN-based loss for domain adaptation task, which shows more excellent performance than other methods. Gu et al. (2018) incorporate bi-modal generative models, text-to-image generative, and image-to-text generative models into the cross-modal feature embedding, which can learn both the global abstract feature and the local features over texts and images. Peng et al. propose cross-modal generative adversarial networks (CM-GANs) (Peng & Qi, 2019) to learn discriminative common representation for bridging the heterogeneity gap. Generally, combining DNNs and GANs becomes the most popular strategy for shared space learning in cross-modal retrieval.

2.3. Graph related methods

In cross-modal retrieval, another notable strategy is introducing a graph into the model construction process. Graph regularization is an effective mechanism in semi-supervised learning, which considers the problem as labeling a partially labeled graph. In literature Zhai, Peng, and Xiao (2014), Zhai et al. propose a joint representation learning (JRL) method that can integrate the sparse and semi-supervised regulation for different models into a unified optimization problem. Specifically, they construct separate graphs for up-to five media types, in which the edge weights denote affinities of labeled and unlabeled data of the same media type. Peng, Zhai, Zhao, and Huang (2016) propose a semi-supervised cross-modal feature learning method with unified patch graph regularization (S²UPG). S²UPG constructs the cross-media graph for all types of media data, which contributes to exploiting multi-level correlations of cross-media data. Many recent works (Liang, Li, Cao, He, & Wang, 2016; Wang et al., 2018) introduce graph regularization as an essential part of their cross-modal retrieval models, which can preserve the intra-modal and inter-modal affinity relationships. Besides, Xu, Li, Yan, et al. (2019) combine graph convolutional networks with hashing for cross-modal retrieval.

The above summarization indicates that current cross-modal retrieval methods are mainly based on common subspace learning. Many researchers have proposed various approaches to achieve this goal, including traditional methods, DNNs based methods, and graph-related methods. The graph-related methods only use the graph regulation term as an additional constraint to obtain better retrieval performance. Compared with their methods, the proposed GRL adopts the cross-modal graph as a bridge to link different modalities, in which the graph structure indicates the affinity relationship among different modal instance. Then, embedding each vertex into a feature vector plays the role of learning the shared representations of all instances.

3. The proposed method

3.1. Problem formulation

For each dataset, it consists of N instances belonging to T_0 modalities, denoted as $\{o^{M_0}, \dots, o^{M_p}, \dots, o^{M_{T_0}}\}$, where M_p is the modal type such as text, image, audio, video, etc. A one-hot vector of label $\mathbf{y}_i = [y_1^i, y_2^i, \dots, y_C^i] \in \mathbb{R}^C$, where C is the number of all categories in each dataset, is assigned to each media instance. All instances of each modality are divided into three sub-datasets randomly, N_0 pairs of instances for training, N_1 pairs of instances for validating, and N_2 pairs of instances for testing. The goal of the proposed method is to find an aligned representation space for all instances, where the similarities of instances can be denoted by Euclidean distance or cosine similarity. The basic idea of the proposed model is shown in Fig. 2. To be specific, the proposed method consists of two main subnetworks, Feature Transfer Learning Network (FTLN) and Graph Representation Learning Network (GRLN). The FTLN aims to find a latent space for every modality before constructing the cross-modal graph, where the cosine similarity is suitable to represent the similarity of the instances. Then, the GRLN formulates a graph-embedding model for embedding all vertexes into a low-dimensional latent space, where each vertex represents a multimedia instance.

Specifically, the FTLN consists of several networks, which is determined by the number of modalities. The FTLN has various implementations, such as multiple fully-connected layers and transfer learning. For better presentation, we adopt multi-layer fully-connected layers to transfer the original feature to a latent space. As follows,

$$(H^M, \dots, H^{M_p}, \dots, H^{M_{T_0}}) \\ = f(o^{M_0}, \dots, o^{M_p}, \dots, o^{M_{T_0}}; \theta_0, \dots, \theta_p, \dots, \theta_{T_0}), \quad (1)$$

where H^{M_p} is the adjusted feature of the M_p modality in latent space and θ_p is the network learnable parameters of the M_p modality. The cross-modal graph is constructed from the adjusted features H^{M_p} by cosine similarity.

During graph embedding learning, each vertex of the constructed graph is projected to a specific vector. The similarity of instances among various modalities can be represented by cosine similarity. Then, the retrieval results of a query can be obtained by ranking the similarities. The following formula abstracts these sub-processes.

$$(\mu^{M_0}, \dots, \mu^{M_p}, \dots, \mu^{M_{T_0}}) = f(H^{M_0}, \dots, H^{M_p}, \dots, H^{M_{T_0}}; \theta, \hat{\theta}), \quad (2)$$

where μ^{M_p} is the hidden representation of M_p modality, θ is the learnable parameter matrix of the embedding layer and $\hat{\theta}$ is the learnable parameters of the semantic classification sub-network.

3.2. Feature Transfer Learning Network

As mentioned before, we need to construct a k -nearest neighbor cross-modal graph by cosine similarity. However, as a matter of experience, the cosine similarity may not be the most proper similarity measure metric for the original BOW feature and the original VGG19 feature. Therefore, we adopt a pre-aligning strategy named FTLN to adjust the distribution of the original feature. In theory, the FTLN can have various types of networks to learning the latent representation. To illustrate accurately, we adopt fully connected layer as the FTLN, as shown in Fig. 3. To ensure the latent space features be suitable for cosine similarity measurement, we aim to project these features into a new latent space. To be specific, taking image modality as an example, we first obtain the original high-level representation o_i^j . Then, the model adopts

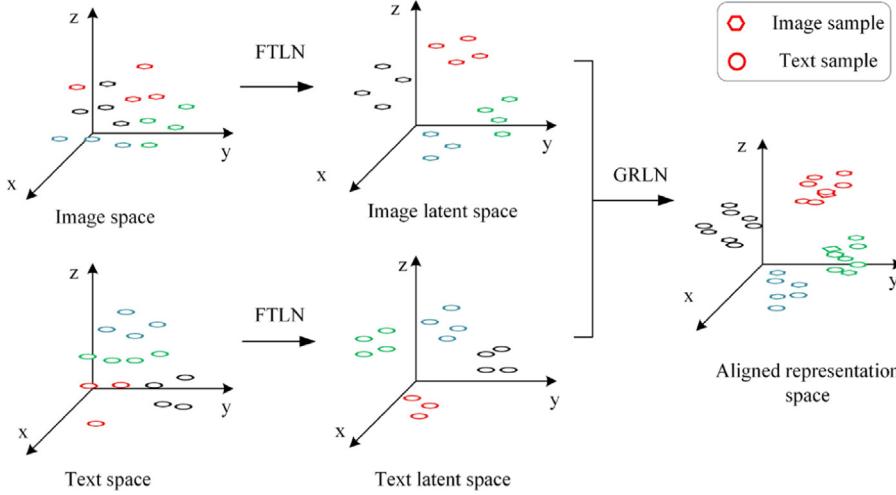


Fig. 2. The general idea of the proposed approach is to achieve modality-invariant embedding for different modality instances with the help of a latent space.

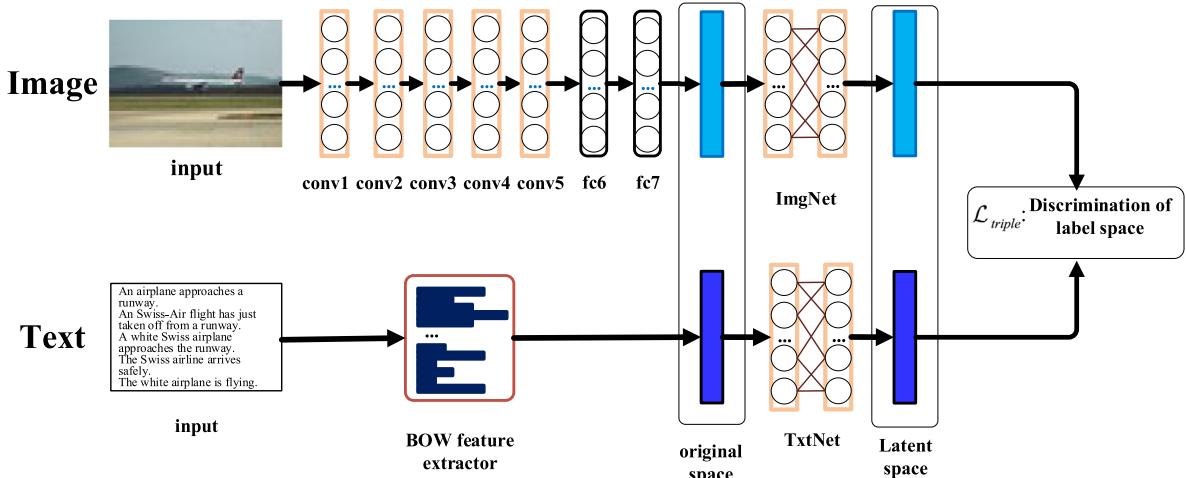


Fig. 3. The detailed framework of the Feature Transfer Learning Network.

the triplet ranking loss function to maximize the cosine similarity of different categories and minimize the cosine similarity of the instances belonging to the same class. Then, we can obtain the hidden feature H^l of image modality.

As is known to all, the triplet ranking loss needs a distance metric in its function expression. It should be noted that all distances between the mapped latent representations H^l are calculated by cosine distance. The reason for choosing cosine similarity instead of Euclidean distance is that cosine similarity considers the angle of two vectors while Euclidean distance considers the spatial distance. Compared with Euclidean distance, cosine similarity plays a vital role in many tasks, such as image classification and information retrieval, because of its robustness. What is more, we construct the cross-modal graph by cosine similarity in the next step. Therefore, the distance of triplet loss adopts cosine distance in the FTLN model, as follows.

$$\begin{aligned} d(v_i, v_j) &= \cos\langle h_i^l, h_j^l \rangle \\ &= f_v(o_i^l, \theta_v)^T \cdot f_v(o_j^l, \theta_v) \\ &= \|f_v(o_i^l, \theta_v)\| \cdot \|f_v(o_j^l, \theta_v)\|. \end{aligned} \quad (3)$$

As mentioned before, we adopt triplet constraint to optimize the FTLN model. Therefore, we need to construct triplets which consist of an anchor, a positive sample, and a negative sample. Firstly, for each modality, we build positive pairs by randomly

selecting two instances with the same label. Specifically, we built positive pairs $\{(v_i, v_j^+)\}_{i,j}$, where v_i is selected as an anchor while v_j^+ with same label is assigned as a positive match. Secondly, we select negative samples having different semantic label to build negative pair $\{(v_i, v_k^-)\}_{i,k}$. Combining positive pair and negative pair, we construct a set of triplets $\{(v_i, v_j^+, v_k^-)\}_{i,j,k}$. Finally, we compute the semantic invariance loss using the following expression that takes the triplets as input.

$$\mathcal{L}_{triplet} = \sum_{i,j,k} \max(\delta - d(v_i, v_j^+) + d(v_i, v_k^-), 0), \quad (4)$$

where δ is margin value. In addition, we introduce the regularization term to prevent the parameters from overfitting, as follows.

$$\mathcal{L}_{reg} = \frac{1}{2} \sum_{k=1}^K \left(\|\mathbf{W}^{(k)}\|_F^2 \right), \quad (5)$$

where F denotes the Frobenius norm and \mathbf{W} is the parameters of DNNs. Therefore, the overall loss function of FTLN model is

$$\mathcal{L}_{FTLN} = \mathcal{L}_{triplet} + \mathcal{L}_{reg}. \quad (6)$$

3.3. Cross-modal graph construction

At first, we give a brief introduction to the graph model.

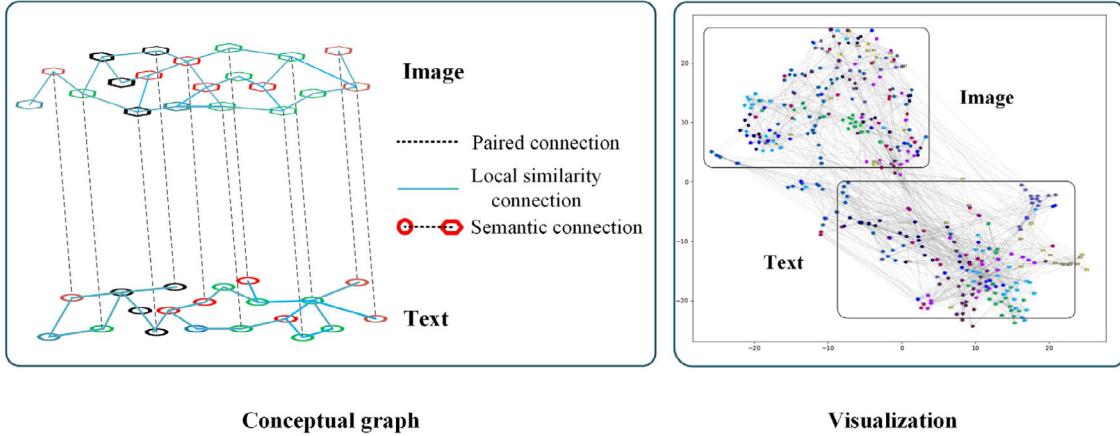


Fig. 4. The detailed information of the cross-modal graph construction. The visualization figure is drawn by PCA, which projects the features from high dimension to two dimension. Then, the nodes and edges are drawn in the graph. In the visualization, the neighbor number k is set as 10.

Definition 1 (Graph). A graph is denoted as $G = (V, E)$, where $V = (v_1, \dots, v_n)$ is the vertexes and $E = \{e_{i,j}\}_{i,j=1}^n$ is the edges. Each edge $e_{i,j}$ is associated with a weight $\omega_{i,j}$, which indicates the strength of the connection between two vertexes.

In our experiment, we only adopt binary value weight for the connections, as follows.

$$\omega_{i,j} = \begin{cases} 0, & \text{if } i\text{th vertex is not connected with } j\text{th vertex,} \\ 1, & \text{if } i\text{th vertex is connected with } j\text{th vertex.} \end{cases} \quad (7)$$

In our work, we construct the multi-modal graph by three types of connections. The first type is semantic connections for all training data. The second type is pairing connections for all training data. The last type is the k -nearest neighbor connections for all instances. To be specific, the connections of any instance $h_i^{M_p}$ (M_p is the modality of this instance) contains three types. The semantic connection denotes the instance $h_i^{M_i}$ is connected with $h_j^{M_k}$ of another modality if $h_i^{M_i}$ and $h_j^{M_k}$ belong to the same category. Because the semantic connection crosses different types of multimedia data, the semantic connection is also called inter-modal connection, which indicates the category information of training instances. The local similarity connection denotes the k -nearest similar neighbors calculated by cosine similarity. The local similarity connections only exist in the same modality, so it is also called intra-modal connections. In reality, the multimedia data usually present in a pairing manner, such as Wikipedia dataset, Pascal Sentence dataset, and NUS-WIDE-10k dataset. Therefore, the constructed cross-modal graph of these datasets has another connection, pairing connection. Although the inter-modal semantic connections already contain the pairing connections, we still introduce this type of connection in constructing the cross-modal graph. The reason is that the pair connections are crucial for aligning the image modality and text modality on account of the pair connections can denote the instances' relationships in the same category. In the training process, the pairing connection and semantic connection are optimized by the different objective functions.

Cosine distance considers the angle size of two feature vectors, which is widely used in information retrieval. Therefore, our experiment adopts cosine similarity to construct the k -nearest similar local structure. For each vertex, k vertex indexes N_{bor}^i is assigned to indicate its neighborhoods if these vertexes in the k -nearest similar neighborhood. In the cross-modal graph: an instance of multimedia data is represented by a vertex; the intra-modal connection is the k -nearest cosine similarity neighbor; the inter-modal connection is the various modal data with semantic information. Their connection status represents the weight

between any two vertexes. Besides, some datasets have pairing connections if their multimedia data present in a pairing manner. Therefore, the original data relationships are represented by the constructed graph, as shown in Fig. 4. The left figure is the concept map of the constructed cross-modal graph, while the right figure is the visualization of a part of the cross-modal graph on the Wikipedia dataset. In both concept graph and visualization graph, the color of nodes denotes the category label. The visualization of cross-modal graph shows that the image modality and text modality are distributed in two different subspaces before graph representation learning. Nonetheless, in same modality, the instances are relatively clustered in different colors. For intuitively observing the constructed graph, Fig. 5 shows several examples selected from the cross-modal graph on Wikipedia dataset. Finally, graph representation learning aims to map the graph vertexes into a low dimensional space, where a low-dimensional vector denotes each vertex.

3.4. Graph Representation Learning Network

The framework of the proposed GRLN model is shown in Fig. 6. As mentioned before, graph representation learning is a node-to-vector embedding problem, instantiated with graph embedding. Graph embedding is the same as the problem of the word-to-vector model in the natural language process, which refers to embedding a one-hot high-dimensional vector into a continuous vector with a low dimension in aligned representation space.

The following content presents the definition of graph embedding.

Definition 2 (Graph Embedding). Given a graph $G = (V, E)$, a graph embedding is a mapping $f: v_i \rightarrow y_i \in \mathbb{R}^d$, $\forall i \in [1, \dots, n]$ such that $d \ll |n|$ and the function f preserves some proximity measure defined on graph G .

An embedding maps each node to a low-dimensional feature vector and tries to preserve the connection strengths between vertices. Adopting global semantic supervision and graph local context learning optimizes the proposed GRL, which can keep the global semantic consistent and preserve the local graph structure, respectively. To be specific, for semantic connections, we adopt global semantic supervised loss function as objective. For local similarities connections, we adopt unsupervised graph local context learning loss function as objective. We also adopt the unsupervised graph local context learning loss function for these datasets with pairing connections.

| Node | Neighbors | | | | | | | | | | Pairing Instance |
|------------|------------|-----------|---------|-----------|-----------|---------|------------|-----------|------------|-----------|--|
| | Music | Warfare | Warfare | Royalty | Media | Music | Literature | Art | Music | Music | After more than 15 years away from the theatre, Sorkin found himself still writing plays, including a play about ploughing in 2003 when he took to revising his play "A Few Good Men" for a revival at the London West End theatre, the Haymarket. ... |
| Media | Music | Warfare | Warfare | Royalty | Media | Music | Literature | Art | Music | Music | Media |
| | Sport | Music | Music | Music | Warfare | Music | Art | Geography | Music | Music | On February 28, 2007, Hanasaki released "I Am...," a pair of compilation albums containing songs from "I Am ..." to "Me" understood". The two versions, "white" and "black", debuted at the first and second positions... |
| Music | Sport | Music | Music | Music | Warfare | Music | Art | Geography | Music | Music | Music |
| | Geography | Geography | Warfare | Biology | Warfare | Warfare | Warfare | Geography | Art | Geography | Both "Kirishima" and "Yamashiro" were scuttled and sunk by 03:55, November 15. Morison. The engagement was one of only two battleship against aircraft surface battles fought in the entire Pacific campaign of World War II... |
| Warfare | Geography | Geography | Warfare | Biology | Warfare | Warfare | Warfare | Geography | Art | Geography | Warfare |
| | Geography | Biology | Biology | History | Geography | Biology | Music | Biology | Literature | Biology | The Blue Iguana Recovery Programme grew from a small project started within the National Trust for the Cayman Islands in 1996. It is now a partnership, linking the Trust with the Cayman Islands Department... |
| Biology | Geography | Biology | Biology | History | Geography | Biology | Music | Biology | Literature | Biology | Biology |
| | Literature | Sport | History | History | History | Sport | Sport | Warfare | Sport | History | On November 26, 1861, Governor Johnson issued an address to the citizen of the Commonwealth blaming abolitionists for the breaking of the Union. Johnson asserted his belief that the Union and Confederacy were forces of equal strength. |
| History | History | Art | History | Geography | History | Meid | Geography | History | Art | History | |
| | History | Art | History | History | History | Meid | Geography | History | Art | History | History |
| History | History | History | History | History | History | Meid | Geography | History | Art | History | History |
| | History | History | History | History | History | Meid | Geography | History | Art | History | |
| Literature | History | History | History | History | History | Meid | Geography | History | Art | History | History |
| | History | History | History | History | History | Meid | Geography | History | Art | History | |
| Literature | History | History | History | History | History | Meid | Geography | History | Art | History | History |
| | History | History | History | History | History | Meid | Geography | History | Art | History | |
| Literature | History | History | History | History | History | Meid | Geography | History | Art | History | History |
| | Media | Media | Music | Music | Music | Media | Music | Music | Music | Music | |
| Music | Media | Media | Music | Music | Music | Media | Music | Music | Music | Music | Music |
| | Art | Warfare | Warfare | History | Warfare | History | Warfare | History | Warfare | Warfare | |
| Warfare | Art | Warfare | Warfare | History | Warfare | History | Warfare | History | Warfare | Warfare | Warfare |

Fig. 5. Several examples of the cross-modal graph.

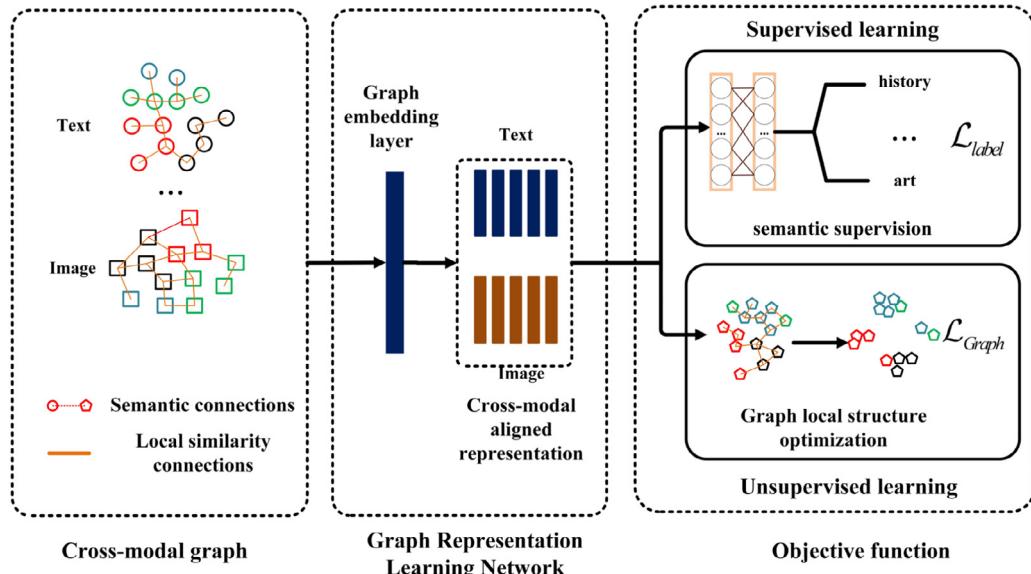


Fig. 6. The detailed framework of the Graph Representation Learning Network.

3.4.1. Global semantic supervised loss

The category supervision is adopted to preserve the global semantic information, where can bring instances together if they are with the same semantic connections. Specifically, a fully-connected layer predicts the category in our model. Then, using cross-entropy loss function optimizes the class distribution, as follows.

$$\mathcal{L}_s = \frac{1}{N} \sum_{k=1}^N y_i \log(\hat{y}_i), \quad (8)$$

where \hat{y}_i is the predicted category. Instead of traditional softmax function, a more advanced Additive Margin softmax function (AMSoftmax) (Wang, Cheng, Liu, & Liu, 2018) is introduced to formulate the probability. The AMSoftmax function is

$$\hat{y}_i = \frac{e^{s(z_i - m_0)}}{e^{s(z_i - m_0)} + \sum_{j=1, i \neq j}^C e^{s z_j}} \quad (9)$$

where s is a scale factor and m_0 is the margin size. Besides, z_i is the output of the fully connected layer, as follows.

$$z_i = \mathbf{W}^T \mu_i + \mathbf{b} \quad (10)$$

where μ_i the embedded vector of the i th instance. The supervised loss only applies to the training data. To prevent overfitting, we introduce \mathcal{L}_2 -norm regularization term in category supervised loss function, as follows.

$$\mathcal{L}_{reg} = \frac{1}{2} \sum_{k=1}^K \left(\|\mathbf{W}^{(k)}\|_F^2 \right), \quad (11)$$

where F is Frobenius norm and \mathbf{W} is the parameters of semantic predict layer.

Eventually, the loss function of category supervision is

$$\mathcal{L}_{label} = \mathcal{L}_s + \mathcal{L}_{reg}. \quad (12)$$

3.4.2. Unsupervised graph local context learning loss

For local structure of the graph, the connections indicate their affinity relationships. Optimizing loss function of graph context learning is unsupervised learning, as follows.

$$\mathcal{L}_{Graph} = E_{(\mu_i, \mu_j, \omega_{i,j}) \sim N_{bor}^i} \log(1 + e^{-\omega_{i,j}(\mu_i, \mu_j)}), \quad (13)$$

where N_{bor}^i is the connected neighbors of the i th node. If the dataset has pairing connections, N_{bor}^i contains pairing instances. According to Jenson inequality, the graph context learning loss can be rewritten as

$$\begin{aligned} \mathcal{L}_{Graph} &= E_{(\mu_i, \mu_j, \omega_{i,j}) \sim N_{bor}^i} \log(1 + e^{-\omega_{i,j}(\mu_i, \mu_j)}) \\ &\geq \log(1 + e^{-E_{(\mu_i, \mu_j, \omega_{i,j}) \sim N_{bor}^i} (\omega_{i,j}(\mu_i, \mu_j))}). \end{aligned} \quad (14)$$

We adopt the widely used Laplacian Eigenmaps (LE) (Zhai et al., 2014) as the graph optimization term. Laplacian Eigenmaps (LE) aims to keep the embedding of two nodes close if the weight $\omega_{i,j}$ is large. Therefore, the graph optimization function is defined as follows.

$$\begin{aligned} \mathcal{L}_{Graph} &= \log(1 + e^{-E_{(\mu_i, \mu_j, \omega_{i,j}) \sim N_{bor}^i} (\omega_{i,j}(\mu_i, \mu_j))}), \\ &= \log(1 + e^{-\mathcal{L}_{LE}}) \end{aligned} \quad (15)$$

where,

$$\begin{aligned} \mathcal{L}_{LE} &= \frac{1}{2} \sum_{i,j} \omega_{i,j} \|f(o_i; \theta) - f(o_j; \theta)\|_2^2 \\ &= \frac{1}{2} \sum_{i,j} \omega_{i,j} \|\mu_i - u_j\|_2^2 \\ &= \text{tr}(\mathbf{\mu}^T L \mathbf{\mu}), \end{aligned} \quad (16)$$

Here, L is the Laplacian of the graph G . The loss function \mathcal{L}_{Graph} plays the function of preserving the local graph structure, which assures that the neighboring instances are also close in the embedded space. It should be noted that both the pairing connections and local graph connections are optimized by \mathcal{L}_{Graph} .

Finally, the overall loss function is

$$\mathcal{L}_{mix} = \alpha_0 \mathcal{L}_{label} + \alpha_1 \mathcal{L}_{Graph}, \quad (17)$$

where α_0 and α_1 are balancing factors. For optimizing the proposed model, the goal is to minimize the loss function by gradient descent. During the training process, the two loss functions are optimized alternately. It is important to note that both kinds of loss functions are indispensable for achieving the desired performance. Therefore, the unsupervised loss function and the supervised loss function have the same contribution in improving the cross-modal retrieval performance.

3.5. Optimization

For optimizing the proposed method, we need to optimize the two subnetworks, FTLN and GRLN. By minimizing the overall loss \mathcal{L}_{FTLN} and \mathcal{L}_{mix} as a function of parameter $\{\theta_0, \dots, \theta_p, \dots, \theta_{T_0}\}$ and $\{\theta, \hat{\theta}\}$ respectively, we can optimize the aforementioned model. Therefore, our goal is to find the parameters $\{\theta_0, \dots, \theta_p, \dots, \theta_{T_0}\}$ and $\{\theta, \hat{\theta}\}$ for getting the excellent performance, as follows.

$$(\theta_0, \dots, \theta_p, \dots, \theta_{T_0}) = \arg \min_{\theta_0, \dots, \theta_p, \dots, \theta_{T_0}} \mathcal{L}_{FTLN}, \quad (18)$$

$$(\theta, \hat{\theta}) = \arg \min_{\theta, \hat{\theta}} \mathcal{L}_{mix}. \quad (19)$$

To be specific, the essential step is to calculate the derivative of the two loss functions. Based on the two equations, we can update the parameters as follows:

$$\theta_p \leftarrow \theta_p - u_0 \frac{\partial \mathcal{L}_{FTLN}}{\partial \theta_p}, \quad (20)$$

$$\theta \leftarrow \theta - u_1 \frac{\partial \mathcal{L}_{mix}}{\partial \theta}, \quad (21)$$

$$\hat{\theta} \leftarrow \hat{\theta} - u_1 \frac{\partial \mathcal{L}_{mix}}{\partial \hat{\theta}} \quad (22)$$

where u_0 and u_1 denote the learning rates for the two subnetworks respectively. The parameter update of the two equations can be realized by RMSprop optimization algorithm. Algorithm 1 shows the details of the optimization process. It should be noted that the proposed method takes all of the data into account during the optimization, including the testing data. The testing data are optimized by the unsupervised learning process. For newly added data, updating the cross-modal graph and fine-tuning the optimized GRLN model can obtain the aligned representation of these data.

4. Experimental results and evaluation

In this section, we do extensive experiments and in-depth analyses on several widely-used datasets to objectively and fully verify the effectiveness of the proposed GRL in cross-modal retrieval task.

4.1. Datasets and features

The cross-modal retrieval experiments are conducted on the widely used datasets, namely Wikipedia dataset (Rasiwasia et al., 2010), NUS-WIDE-10k dataset (Chua et al., 2009), Pascal Sentences dataset (Rashtchian, Young, Hodosh, & Hockenmaier, 2010),

Algorithm 1: The training algorithm of the proposed GRL

Input: m : mini-batch size; u_0 : Learning rate of Feature Transfer Learning Network; u_1 : learning rate of Graph Representation Learning Network; α_0, α_1 : trade-off factors for \mathcal{L}_{Graph} ;

Output: weight parameters $\{\theta_0, \dots, \theta_p, \dots, \theta_n\}$ and $\{\theta, \hat{\theta}\}$ of FTLN and GRLN respectively.

begin

 Initialize FTLN model with random weights.

 repeat

 Sample mini-batch size triplet tuples based on their semantic information

 Update parameters of the FTLN model

$$\bullet \theta_p \leftarrow \theta_p - u_0 \frac{\partial \mathcal{L}_{FTLN}}{\partial \theta_p}$$

 end until FTLN model converges

 Obtain the H^P latent feature for all every modality

 Construct the cross-modal graph

 Initialize GRLN model with random weights.

 repeat

 Sample mini-batch size m instances from cross-modal graph

 Update the parameters of the GRLN model

$$\bullet \theta \leftarrow \theta - u_1 \frac{\partial \mathcal{L}_{mix}}{\partial \theta}, \hat{\theta} \leftarrow \hat{\theta} - u_1 \frac{\partial \mathcal{L}_{mix}}{\partial \hat{\theta}}$$

 end until GRLN model converges

 Obtain the aligned representations for each modality

end

PKU XMedia dataset (Peng et al., 2017), PKU XMediaNet dataset (Peng, Qi, & Yuan, 2018), and MSCOCO dataset (Lin et al., 2014) respectively. It should be noted that both PKU XMedia dataset and PKU XMediaNet dataset consist of five types of media data. For the five datasets, the image feature is 4096-dimensional VGG-19 feature (Simonyan & Zisserman, 2014) and the text feature is 3000-dimensional BOW feature, if we do not explain specially. The detailed statistic information of the evaluated datasets are shown in Table 1. The next content presents the detailed information about these five datasets.

Wikipedia dataset (Raswasia et al., 2010) is constructed from the “featured article” of Wikipedia, which consists of 10 high-level semantic categories such as art and history. This dataset contains 2866 image-text pairs and each image-text pair only belongs to one class. The experiment details of the Wikipedia dataset strictly follow the partition shown in McGurk and MacDonald (1976). We randomly choose 2173 image-text pairs as training set, 231 pairs of instances as validating set, and 462 image-text pairs as testing set.

NUS-WIDE-10k dataset (Chua et al., 2009) is a part of a well-known large-scale dataset, named NUS-WIDE dataset, which contains 269,648 images and several corresponding semantic tags for each image. Following Huang et al. (2017), we choose the 10 largest categories of images with a unique label to conduct the experiments. In total 10,000 image-text pairs, we choose 8000 pairs of instances for training and 1000 pairs of instances for testing. The text feature is 1000-dimensional BOW feature.

Pascal Sentence dataset (Rashtchian et al., 2010) contains 1000 images of 20 categories. Each image is described by five sentences, which is viewed as a document. Our experiment splits the dataset into three sub-datasets, 800 pairs for training, 100 pairs for validating and 100 pairs for testing.

PKU XMedia dataset (Peng et al., 2017) is constructed for multi-modal retrieval, which contains five modalities, image, text, audio, video, and 3D model. This dataset contains 4000 images, 4000 texts, 500 videos, 1000 audios, and 500 3D models. In our experiment, the dataset partition follows the previous

work (Huang et al., 2020). The adopted features are as follows: video feature, 4096-dimensional CNN feature; Audio, 29-dimensional MFCC (Han, Chan, Choy, & Pun, 2006); 3D model, 4700-dimensional LightField feature (Chen, Tian, Shen, & Ouhyoung, 2003).

PKU XMediaNet dataset (Peng, Qi, & Yuan, 2018) is a large-scale dataset of texts, images, videos, audios, and 3D models for cross-modal retrieval, which consists of five types of media data with more than 100,000 instances. This dataset consists of 40,000 texts, 40,000, 10,000 videos, 10,000 audios, and 2000 3D models. The total number of all multimedia instances exceeds 100,000. The data partition is shown in Table 1. The video feature is extracted by C3D model (Tran, Bourdev, Fergus, Torresani, & Paluri, 2015), which is pre-trained on Sports1M (Karpathy et al., 2014). The audio feature is extracted by jAudio (McKay, Fujinaga, & Depalle, 2005) using its default setting. The same to PKU XMedia dataset, the 3D model is represented by 4700-dimensional vector of a LightField descriptor set.

MSCOCO dataset (Lin et al., 2014) is a large-scale dataset for various tasks, such as object recognition, image classification, and cross-modal retrieval. The dataset contains about 120,000 instances of image-text pairs. In this dataset, five sentences describing one image. In our experiment, the five sentences are considered as a document. It should be noted that each image-text pair of MSCOCO dataset is associated with multiple class labels. Following previous works (Peng, Huang, & Qi, 2016), we adopt 66,226 instances for training, 16,557 for validating, and 16,557 for testing. The 4096-dimensional image feature and 2000-dimensional text feature are extracted by the VGG19 model and BoW model, respectively.

4.2. Evaluation metric and implementation details

4.2.1. Evaluation metric

We adopt the widely used metric in cross-modal retrieval, Mean Average Precision (MAP), to evaluate the performance of the proposed method. Mean Average Precision (MAP) score, the mean value of ground-truth matchings in the retrieved results for all queries, is a very popular evaluation metric in retrieval tasks. For every query, the Average Precision is computed as the following formula.

$$AP = \frac{1}{R} \sum_{j=1}^N \frac{R_j}{j} \times rel_j, \quad (23)$$

where rel_j is denoted by 0 or 1, and R_j is the sequence number of the relevant instances among the top- j retrieved results. Then, we obtain the eventual MAP by calculating the mean value of all the average precision values, as follows.

$$MAP = \frac{1}{N} \sum_{j=1}^N AP_j, \quad (24)$$

In our experiments, the mean average precision is calculated by all the retrieved results not top-50 results (Wang et al., 2017) if we do not explain specially. Besides, it should be noted that the MAP scores are presented in percentage type in our paper.

Besides, the top returned instances are usually more important in evaluation of information retrieval. As is common in information retrieval, we also measure the performance by recall at K (R@K) defined as the fraction of queries for which the correct item is retrieved in the closest K points to the query. The calculation formula is as follows,

$$R@k = \frac{1}{N} \sum_{j=1}^N Re_j, \quad (25)$$

where Re_j is 1 if the correct instance is returned in the top- k results, otherwise it is 0.

Table 1

The detailed statistic information of the evaluated datasets.

| Dataset | Category | Modality number | Image feature | Text feature | Train | Valid | Test | Total |
|-----------------|----------|-----------------|---------------|--------------|--------|--------|--------|---------|
| Wikimedia | 10 | 2 | 4096d VGG19 | 3000d BoW | 2,173 | 231 | 462 | 2,635 |
| NUS-WIDE-10k | 10 | 2 | 4096d VGG19 | 1000d BoW | 8,000 | 1,000 | 1,000 | 10,000 |
| Pascal Sentence | 20 | 2 | 4096d VGG19 | 3000d BoW | 800 | 100 | 100 | 1,000 |
| PKU XMedia | 20 | 5 | 4096d VGG19 | 3000d BoW | 9,600 | 1,200 | 1,200 | 12,000 |
| PKU XMediaNet | 200 | 5 | 4096d VGG19 | 3000d BoW | 81,600 | 10,200 | 10,200 | 102,000 |
| MSCOCO | 80 | 2 | 4096d VGG19 | 2000d BoW | 66,226 | 16,557 | 16,557 | 99,340 |

4.2.2. Implementation details

We implement the proposed method on a desktop computer with an Intel Core i7-6700K CPU, NVIDIA GeForce GTX1080ti for acceleration and 64 GB of memory. In our experiment, the details of subnetworks are as follows: For the FTLN model, it consists of two fully connected layers respectively and is optimized by triplet loss. For the GRLN model, every node index of the cross-modal graph is embedded into a 512-dimensional vector, which aims to learn the aligned feature. Besides, we adopt a fully connected layer with AMSoftmax activation function for label prediction in the supervised learning. The weight parameters of the embedding layer are shared in the global semantic objective function optimization and graph local structure optimization. The proposed method is implemented on Keras and optimized by RMSprop (Tieleman & Hinton, 2012) optimizer with backpropagation. The learning rate of RMSprop optimizer is set as 0.001 for both two losses. The trade-off hyper-parameters of various losses are set as 5 and 2, respectively. The value of m_0 and s are 0.2 and 15 respectively. All the hyper-parameters are determined by experiments. For different datasets, the values of hyper-parameters have variations.

4.3. Experimental results and evaluations

Two types of cross-modal retrieval tasks are conducted for objectively evaluating the proposed method, as follows.

- **Bi-modal cross-modal retrieval.** We choose the most widely used image modality and text modality to conduct the experiment.
- **Multi-modal cross-modal retrieval.** The experiments are conducted on all five modalities (i.e., image, text, audio, video, and 3D model) of the PKU XMedia dataset and PKU XMediaNet dataset.

For each cross-modal retrieval, there are two kinds of evaluating experiments. The first is bi-directional retrieval, which performs retrieval between two modalities, such as image retrieve text (image → text), and text retrieve image (text → image). The second is all-modal retrieval, which performs retrieval within all modalities, such as image retrieve all modalities (image → all), and text retrieve all modalities (text → all).

The compared methods contain: **traditional cross-modal retrieval methods**, such as CCA (Hotelling, 1936), KCCA (Hardoon et al., 2004), CCA-3V (Gong et al., 2014), CFA (Li et al., 2003), JFSSL (Wang et al., 2015), SCM (Pereira et al., 2013), JRL (Zhai et al., 2014), LGCFL (Kang et al., 2015), and **DNN-based approaches**, such as DCCA (Andrew et al., 2010), Bimodal AE (Ngiam et al., 2011), Multimodal DBN (Srivastava & Salakhutdinov, 2012), Corr-AE (Feng, Wang, & Li), Deep-SM (Wei et al., 2017), CMDN (Peng, Huang, & Qi, 2016), CM-GANs (Peng & Qi, 2019), MScm (Peng, Qi, & Yuan, 2018), ACMR (Wang et al., 2017), MHTN (Huang et al., 2020), CMST (Wen, Han, Yin, & Liu, 2019), CCL (Peng, Qi, Huang, & Yuan, 2018), TPCKT (Huang & Peng, 2019). Among DNN-based approaches, MHTN (Huang et al., 2020), ACMR (Wang et al., 2017), CM-GANs (Peng & Qi, 2019), CMST (Wen et al.,

2019) and TPCKT (Huang & Peng, 2019) are adversarial learning based methods. Next, we briefly introduce some approaches, which are not included in the previous content.

Bimodal AE (Tran et al., 2015) adopts deep networks to learn features of different modalities and then learns a shared representation between different modalities.

Corr-AE (Feng et al.) introduces a correspondence autoencoder to realize cross-modal retrieval. It models the correlation and reconstruction learning error with two different multi-modal autoencoders.

CMST (Wen et al., 2019) first employs an unsupervised strategy to learn the endogenous semantic relationships and then transfers the learned relationships to the common representation subspace.

CCL (Peng, Qi, Huang, & Yuan, 2018) realizes cross-modal correlation learning with multi-grained fusion by a hierarchical network, which can deeply explore the intra-modality semantic supervision and inter-modality pair-wise similarity constraints.

TPCKT (Huang & Peng, 2019) proposes the approach of two-level progressive cross-media knowledge transfer, which transfers knowledge from large-scale cross-media data, to boost the retrieval accuracy on cross-media data of another domain.

MScm (Xu et al., 2019) attempts to construct the independent semantic space for each modality instead of learning the common representation of different modalities, which can generate the similarity of different instances by an end-to-end framework.

4.3.1. Bi-modal cross-modal retrieval

In this subsection, we compare the proposed GRL with many state-of-the-art methods by the cross-modal retrieval accuracy to evaluate the performance of the proposed GRL approach. For a fair comparison, all the compared methods adopt the same MAP metric and the same extracted features. The experimental results of MAP(@all) are shown in Tables 2 and 4, which show the results of bi-directional retrieval and all-modal retrieval, respectively. Besides, we also calculate the MAP(@50), which is widely adopted by many methods such as ACMR and CMST, to show the retrieval performance in the top-50 retrieved results. Table 3 shows the MAP(@50) scores of various approaches. These experimental results show that our proposed GRL method achieves the best retrieval accuracy compared with all the state-of-the-art models on all of the evaluated datasets. Besides, our proposed GRL makes significant improvements on all the datasets. For example, on PKU XMedia dataset, the MAP(@all) score has been improved from 84.8 to 92.0. On the PKU XMediaNet dataset, the MAP(@all) has been improved from 55.9 to 61.8. On Pascal Sentence dataset, the MAP(@50) score has been improved from 60.4 to 72.7. On all-modal retrieval, Table 4 shows the same results that our proposed method outperforms the state-of-the-art methods on the evaluated datasets.

Among the compared methods, we can observe that most DNNs based methods, such as MScm, CCL, CM-GAN, TPCKT, can obtain better performance than traditional approaches. These results also verify the great fitting capacity of DNNs, which contributes to this superior performance. Furthermore, among these DNNs based methods, the hybrid strategy that combines DNNs

Table 2

The MAP (@all) comparison of the cross-modal retrieval performance on Wikipedia dataset, NUS-WIDE-10k Dataset, Pascal Sentence dataset, PKU XMedia dataset, and PKU XMediaNet dataset, calculated on all returned results.

| Dataset | Task | CCA | CFA | DCCA | Bimodal AE | Multimodal DBN | Corr-AE | JRL | LGCFL | CMDN | Deep-SM | ACMR | MHTN | CCL | CM-GANs | MCSM | TPCKT | GRL |
|-----------------|--------------|------|------|------|------------|----------------|---------|------|-------|------|---------|------|------|------|---------|------|-------|------|
| Wikipedia | Image → Text | 38.4 | 39.6 | 40.9 | 30.1 | 20.4 | 37.3 | 40.8 | 41.6 | 40.9 | 45.8 | 43.9 | 51.4 | 50.5 | 52.1 | 51.6 | 54.8 | 54.7 |
| | Text → Image | 36.7 | 37.3 | 35.5 | 26.7 | 14.5 | 35.7 | 36.0 | 36.0 | 36.4 | 34.5 | 36.1 | 44.4 | 45.7 | 46.6 | 45.8 | 48.9 | 50.0 |
| | Average | 36.5 | 38.4 | 38.2 | 28.4 | 17.5 | 36.5 | 38.8 | 38.8 | 38.7 | 40.2 | 40.0 | 47.9 | 48.1 | 49.4 | 48.7 | 51.9 | 52.3 |
| NUS-WIDE-10k | Image → Text | 15.9 | 29.9 | 38.4 | 23.4 | 17.8 | 30.6 | 41.0 | 40.8 | 41.0 | 38.9 | 44.5 | 52.0 | 48.1 | – | – | 57.5 | 57.9 |
| | Text → Image | 18.9 | 30.1 | 38.2 | 37.6 | 14.4 | 34.0 | 44.4 | 37.4 | 45.0 | 49.6 | 47.3 | 53.4 | 52.0 | – | – | 58.9 | 59.4 |
| | Average | 17.4 | 30.0 | 38.3 | 30.5 | 16.1 | 32.3 | 42.7 | 39.1 | 43.0 | 44.3 | 45.9 | 52.7 | 50.1 | – | – | 58.2 | 58.7 |
| Pascal Sentence | Image → Text | 11.0 | 34.1 | 31.2 | 40.4 | 43.8 | 41.1 | 41.6 | 38.1 | 45.8 | 44.0 | 43.4 | 49.6 | 57.6 | 60.3 | 59.8 | 58.6 | 71.6 |
| | Text → Image | 11.6 | 30.8 | 31.1 | 44.7 | 36.3 | 47.5 | 37.7 | 43.5 | 44.4 | 41.4 | 41.6 | 50.0 | 56.1 | 60.4 | 59.8 | 59.5 | 70.9 |
| | Average | 11.3 | 32.5 | 31.1 | 42.6 | 40.1 | 44.3 | 39.7 | 40.8 | 45.1 | 42.7 | 42.5 | 49.8 | 56.9 | 60.4 | 59.8 | 59.1 | 71.3 |
| PKU XMedia | Image → Text | 25.7 | 29.2 | 47.2 | 59.8 | 9.3 | 45.0 | 77.0 | 74.4 | 79.4 | 82.2 | 70.4 | 85.3 | – | – | – | – | 91.3 |
| | Text → Image | 34.1 | 28.3 | 46.6 | 64.2 | 12.0 | 43.7 | 80.0 | 80.4 | 80.5 | 80.7 | 71.0 | 84.3 | – | – | – | – | 92.7 |
| | Average | 20.0 | 28.6 | 41.0 | 26.7 | 12.3 | 24.2 | 54.8 | 38.7 | 36.3 | 65.7 | 55.5 | 84.8 | – | – | – | – | 92.0 |
| PKU XMediaNet | Image → Text | 21.2 | 25.2 | 42.5 | – | – | 46.9 | 48.8 | 44.1 | 48.5 | 39.9 | – | – | 53.7 | 56.7 | 54.0 | – | 61.2 |
| | Text → Image | 21.7 | 40.0 | 43.3 | – | – | 50.7 | 40.5 | 50.9 | 51.6 | 34.2 | – | – | 52.8 | 55.1 | 55.0 | – | 62.3 |
| | Average | 21.5 | 32.6 | 42.9 | – | – | 48.8 | 44.7 | 47.5 | 50.1 | 37.1 | – | – | 53.3 | 55.9 | 54.5 | – | 61.8 |

and adversarial learning can achieve better retrieval accuracy, such as CM-GAN, CMST, and TPCKT. ACMR is the first attempt to introduce adversarial learning into cross-modal retrieval. Among the adversarial learning-based methods, TPCKT and CM-GANs obtain the second and third-best performance. CM-GANs introduces not only the cross-modal generative adversarial networks but also a new cross-modal adversarial mechanism in cross-modal retrieval model. What is more, both the CM-GANs and TPCKT adopt the pre-trained text CNN model, which is pre-trained on billions of words in Google News, to extract the deep text feature instead of adopting the shallow BoW feature. Although the proposed method adopts shallow BoW feature for text, the experimental results show that the GRL has several advantages over CM-GANs and TPCKT. It should also be noted that, among these traditional methods, JRL and LGCFL obtain competitive performance by comparing with DNNs based methods.

The MSCOCO is a standard dataset for cross-modal retrieval, widely used in image-text matching task. We do experiments on the MSCOCO dataset to evaluate the performance of the proposed GRL. Several recently proposed methods, such as CCA (Klein, Lev, Sadeh, & Wolf), DVSA (Karpathy & Fei-Fei), m-RNN (Mao et al., 2014), m-CNN (Ma, Lu, Shang, & Li), DSPE (Wang, Li, & Lazebnik), and ACMR (Wang et al., 2017), are included to compare with the proposed method. For a fair evaluation of the MSCOCO dataset, we follow the experimental protocol and quote the MAP results obtained by ACMR (Wang et al., 2017). The MAP scores of the GRL and the compared methods are shown in Table 5. The results show that the proposed method obtains the best MAP score, which is 93.5 with FTNL and 92.3 without FTNL. On the MSCOCO dataset, the proposed method obtains significant improvement (from 90.5 to 93.5) compared with the performance of the second-best method, ACMR. The reason is graph representation learning increase the learning ability of shared representation learning.

For information retrieval, the top returned instances are usually more essential to obtain highly related items. Therefore, the precision of the top returned instances is included to evaluate the proposed method's effectiveness. The quantitative evaluation results of R@k are shown in Table 6. Table 6 shows that the proposed method achieves excellent performance on precision at the top-k returned instances. Besides, we can observe that the MAP score metric is highly related to the R@k evaluation metric. Specifically, the proposed GRL obtains higher R@k precision value on Pascal Sentence dataset and PKU Xmedia dataset, which are consistent with the results of the MAP score evaluation metric.

For visual comparison, we draw the precision-recall curves of cross-modal retrieval tasks on the Wikipedia dataset, Pascal Sentence dataset, NUS-WIDE-10k dataset, and PKU XMediaNet dataset, as shown in Figs. 7, 8, 9, and 10. The visual comparison shows that the proposed GRL has a noticeable improvement compared with state-of-the-art methods. Besides, for better observing effectiveness of the proposed method, we draw the visualized feature distribution by t-SNE visualization (Maaten & Hinton, 2008) for the Wikipedia dataset, Pascal Sentence dataset, NUS-WIDE-10k dataset, and PKU XMedia dataset, as shown in Fig. 11. The numbers in the legend indicate the category labels in each dataset. The visualization of PKU XMediaNet dataset is drawn in Fig. 12, which contains up to 200 classes. The t-SNE visualizations of the original feature and embedded feature indicate that the original feature distribution is chaotic, while the embedded feature distribution is ordered. The embedded feature distribution presents many clusters, which denote different categories of both image and text.

To intuitively observe the cross-modal retrieval results, we choose some retrieval examples, as shown in Fig. 13. In addition to successful cases, Fig. 8 also shows some failure examples. For

instance, the retrieved texts of a query image of the 'History' category are almost 'warfare' texts. The reason is 'history' and 'warfare' are indistinguishable in high-level semantic representation, even for our human beings. Therefore, the algorithm also makes similar mistakes when it faces to this challenging issue.

4.3.2. Multi-modal cross-modal retrieval

We do experiments on the PKU XMedia dataset and PKU XMediaNet dataset to verify the effectiveness of the proposed GRL on multi-modal cross-modal retrieval tasks. As mentioned before, both PKU XMedia dataset and PKU XMediaNet dataset have five modalities, image, text, audio, video, and 3D model, respectively. The experiment results of multi-modal bi-directional retrieval are shown in Table 7. Compared with bi-modal retrieval, multi-modal retrieval is more difficult because of the various feature types. Table 7 indicates that the proposed method also can obtain excellent performance on multi-modal retrieval tasks, especially in the PKU XMedia dataset. However, the experiment results on PKU XMediaNet dataset show that cross-modal retrieval is more complicated than the PKU XMedia dataset. The reason is that the PKU XMedia dataset only contains 20 categories, while the PKU XMediaNet dataset contains 200 categories. The experiment results show that the audio modality related MAP scores are very low. As mentioned before, the audios are represented by 78-dimensional shallow feature extracted by the jAudio model. However, the category of audio modality is up to 200, which leads to vast differences between the original feature dimension and the target label dimension. Therefore, the audio modality related retrieval tasks have poor results. The proposed method can achieve excellent performance on the other modalities, such as image, text, video, and 3D model. In future work, improving the retrieval results on PKU XMediaNet dataset focuses on enhancing the feature quality.

4.3.3. Experimental analysis

In this subsection, we make in-depth analyses about the experimental results of our proposed GRL. Besides, we make more experiments to analyze the advantages of the proposed method and the contributions of each part of the proposed model.

To analyze each category's specific retrieval performance, we draw these histograms to show the MAP score of each category, as shown in Fig. 14. The category information is extremely high-level semantic and abstract on the Wikipedia dataset, such as 'art' and 'history', which are also difficult for our human beings to distinguish. The experiment results show that some categories obtain relatively low MAP scores, such as 'history', 'art', 'royalty', and 'literature', which are consistent with our intuition. On the NUS-WIDE-10k dataset, the categories 'sky' and 'clouds' obtain the lowest MAP score. Compared with other categories, the differences between these two categories are relatively semantic indistinctness. On the Pascal Sentence dataset, most categories can obtain relatively high MAP scores. However, a few categories have low performance, such as 'cow' and 'sheep'. Therefore, in future work, we need to extract more distinguishing features to improve the cross-modal retrieval, especially for similar categories.

We also show the plots of the two sub-loss functions and their total sum loss values of the GRLN model, as shown in Fig. 15. We can observe that the value of category supervision loss reduces quickly with the growth of training iteration. On the Wikipedia dataset, the loss \mathcal{L}_{label} converges about 200 iterations, while the loss \mathcal{L}_{Graph} converges about 600 iterations. In our experiment, we find that the retrieval accuracy on the validation dataset obtains the highest value at about 300 iterations, where the plots of graph loss generally become stable.

In the proposed GRL, the extracted features are critical to the cross-media graph construction. To verify the effectiveness

Table 3

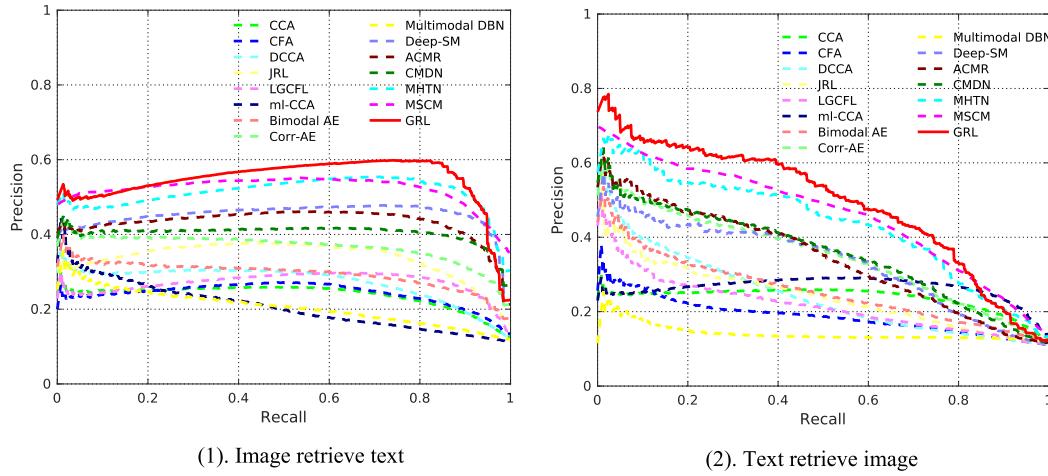
The MAP (@50) comparison of the cross-modal retrieval performance on Wikipedia dataset, NUS-WIDE-10k dataset, and Pascal Sentence dataset, calculated on the top-50 returned results.

| Dataset | Task | CCA | Multimodal DBN | Bimodal-AE | Corr-AE | JRL | CCA-3V | JFSSL | CMDN | ACMR | CMST | CCL | CM-GANs | GRL |
|-----------------|--------------|------|----------------|------------|---------|------|--------|-------|------|------|------|------|---------|------|
| Wikipedia | Image → Text | 26.7 | 20.4 | 31.4 | 40.2 | 45.3 | 43.7 | 42.8 | 48.8 | 61.9 | 63.2 | 49.0 | 50.0 | 53.3 |
| | Tex → Image | 22.2 | 18.3 | 29.0 | 39.5 | 40.0 | 38.3 | 39.6 | 42.7 | 48.9 | 50.5 | 61.3 | 62.1 | 65.0 |
| | Average | 24.5 | 19.4 | 30.2 | 39.8 | 42.6 | 41.0 | 41.2 | 45.8 | 55.4 | 56.9 | 55.1 | 56.1 | 59.2 |
| NUS-WIDE-10k | Image → Text | 18.9 | 20.1 | 32.7 | 36.6 | 42.6 | 40.8 | 38.9 | 49.2 | 54.4 | 62.1 | – | – | 67.9 |
| | Text → Image | 18.8 | 25.9 | 36.9 | 41.7 | 37.6 | 37.4 | 49.6 | 51.5 | 53.8 | 58.6 | – | – | 65.8 |
| | Average | 18.9 | 23.0 | 34.8 | 39.2 | 40.1 | 39.1 | 44.3 | 50.4 | 54.1 | 60.4 | – | – | 66.9 |
| Pascal Sentence | Image → Text | 24.7 | – | – | 48.9 | 50.4 | 31.6 | – | 53.4 | 53.5 | 62.1 | 59.2 | 61.2 | 73.2 |
| | Text → Image | 24.8 | – | – | 44.4 | 48.9 | 27.0 | – | 53.4 | 54.3 | 58.6 | 57.6 | 61.0 | 72.3 |
| | Average | 24.8 | – | – | 46.7 | 49.6 | 29.3 | – | 53.4 | 53.9 | 60.4 | 57.9 | 61.1 | 72.7 |

Table 4

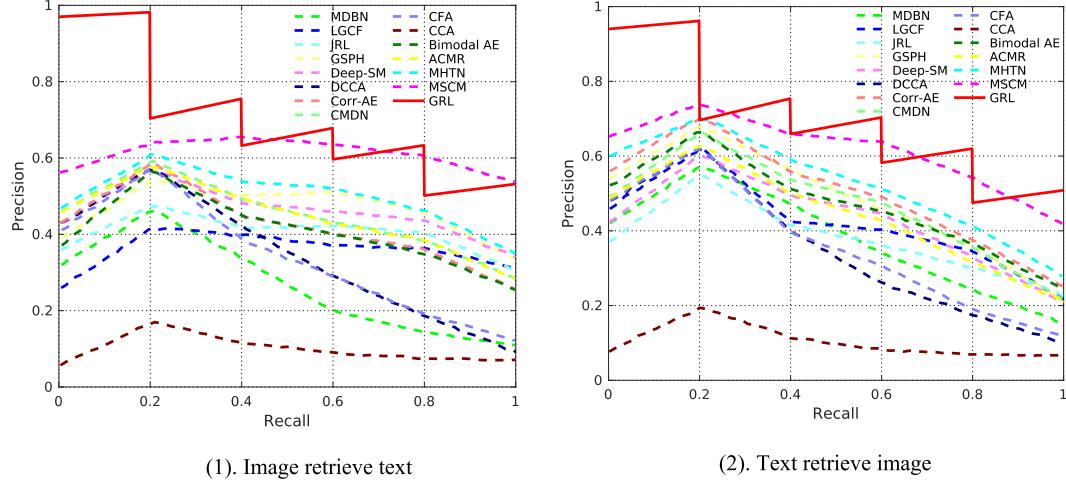
The MAP (@all) comparison of the cross-modal retrieval performance on Wikipedia dataset, Pascal Sentence dataset, and PKU XMediaNet dataset, calculated on the all returned results.

| Dataset | Task | CCA | CFA | KCCA | DCCA | JRL | LGCFL | Deep-SM | CMDN | ACMR | CCL | CM-GANs | GRL |
|-----------------|-------------|------|------|------|------|------|-------|---------|------|------|------|---------|------|
| Wikipedia | Image → All | 26.8 | 27.9 | 35.4 | 37.1 | 40.4 | 39.2 | 39.1 | 40.7 | 39.9 | 42.2 | 43.4 | 45.3 |
| | Tex → All | 37.0 | 34.1 | 51.8 | 56.0 | 59.5 | 59.8 | 59.7 | 61.1 | 59.5 | 65.2 | 66.1 | 68.5 |
| | Average | 31.9 | 31.0 | 43.6 | 46.6 | 50.0 | 49.5 | 49.4 | 50.9 | 49.7 | 53.7 | 54.8 | 56.9 |
| Pascal Sentence | Image → All | 23.8 | 47.0 | 34.6 | 55.6 | 56.1 | 38.5 | 55.5 | 49.6 | 56.5 | 57.5 | 58.4 | 69.2 |
| | Text → All | 30.1 | 49.7 | 42.9 | 65.3 | 63.1 | 42.0 | 65.3 | 62.7 | 62.5 | 63.2 | 69.8 | 70.6 |
| | Average | 27.0 | 48.4 | 38.8 | 60.5 | 59.6 | 40.3 | 60.5 | 56.2 | 59.5 | 60.4 | 64.1 | 69.9 |
| PKU XMediaNet | Image → All | 25.4 | 31.8 | 29.9 | 43.3 | 50.8 | 31.4 | 35.1 | 50.4 | 57.0 | 55.2 | 58.1 | 68.1 |
| | Text → All | 25.2 | 20.7 | 18.6 | 47.5 | 50.5 | 54.4 | 33.8 | 56.3 | 54.8 | 57.8 | 59.0 | 59.0 |
| | Average | 25.3 | 26.3 | 24.3 | 45.4 | 50.7 | 42.9 | 34.5 | 53.4 | 55.9 | 56.5 | 58.6 | 63.6 |



(1). Image retrieve text

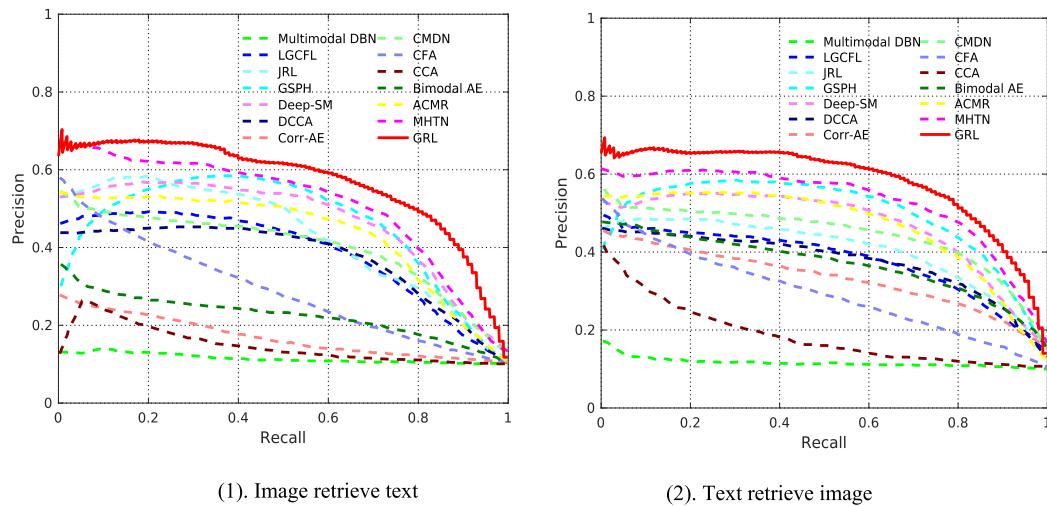
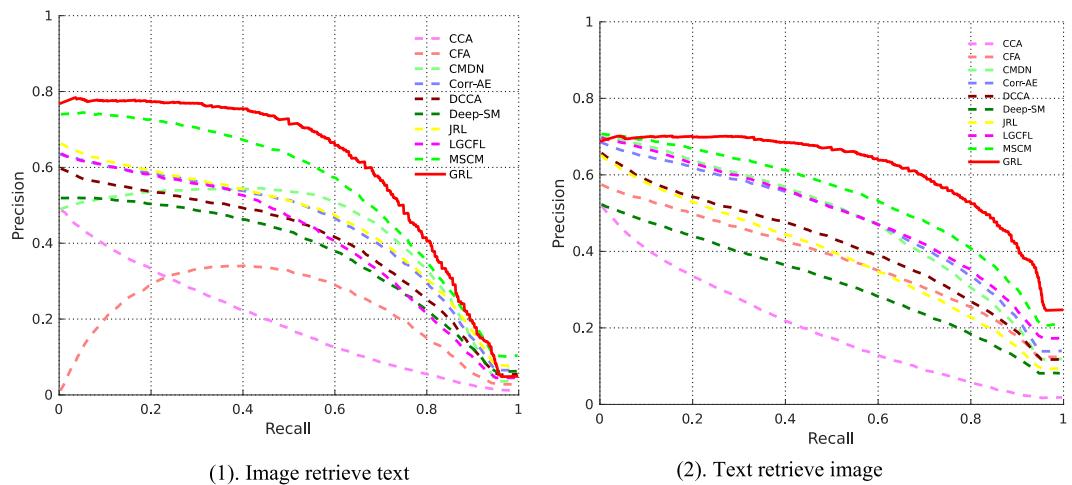
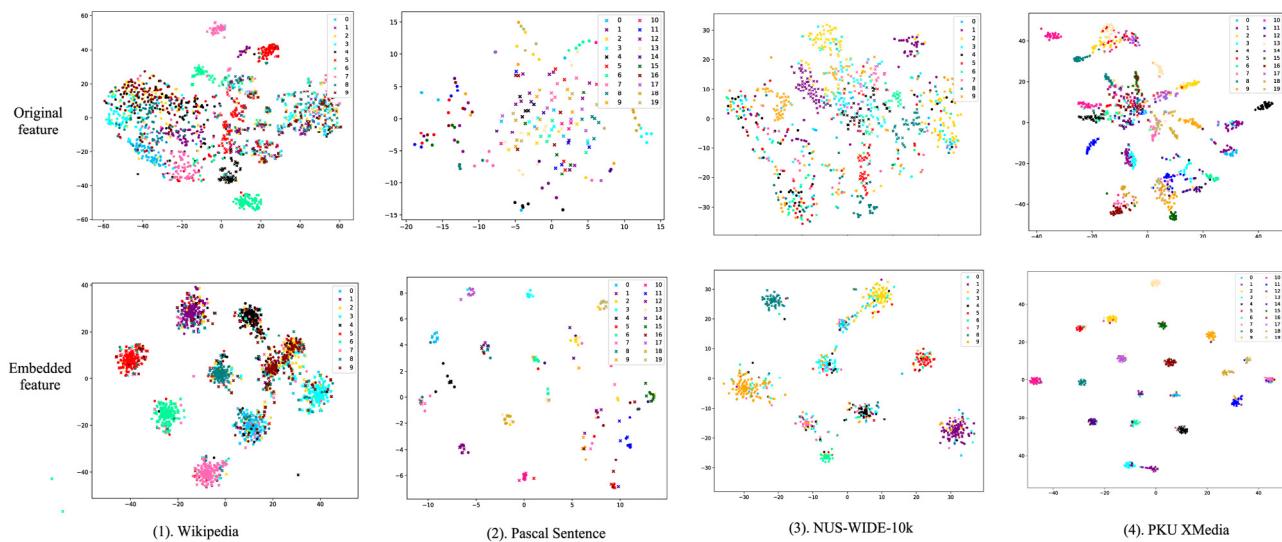
(2). Text retrieve image

Fig. 7. Precision recall curves of cross-modal retrieval on Wikipedia dataset.

(1). Image retrieve text

(2). Text retrieve image

Fig. 8. Precision recall curves of cross-modal retrieval on Pascal Sentence dataset.

**Fig. 9.** Precision recall curves of cross-modal retrieval on NUS-WIDE-10k dataset.**Fig. 10.** Precision recall curves of cross-modal retrieval on PKU XMediaNet dataset.**Fig. 11.** The feature visualization by using t-SNE on Wikipedia dataset, Pascal Sentence dataset, NUS-WIDE-10k dataset, and PKU XMedia dataset.

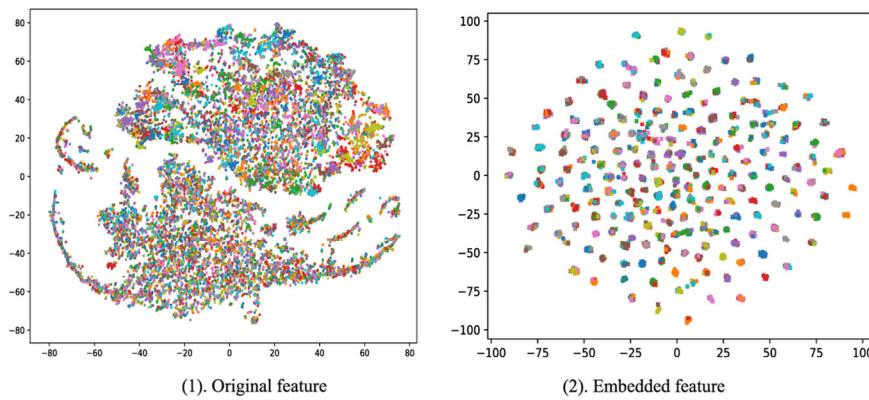


Fig. 12. The feature visualization by t-SNE on PKU XMediaNet dataset.

| Task | Query | Top 5 Results | | | | | Performance | |
|--------------------|-----------|---------------|-----------|-----------|-----------|-----------|-------------|------|
| Image → Text | geography | | geography | geography | geography | geography | geography | Good |
| | History | | history | warfare | warfare | warfare | history | Fair |
| | royalty | | warfare | history | history | geography | art | Bad |
| Text → Image | art | | | | | | art | Fair |
| | royalty | | | | | | warfare | Bad |
| | warfare | | | | | | warfare | Good |

Fig. 13. Some cross-modal retrieval examples on Wikipedia dataset. The instances in green border are successful cases, while the instances in red border are failures. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5

The MAP comparison of the cross-modal retrieval performance on MSCOCO dataset.

| Methods | Image-text retrieval | | |
|--------------------|----------------------|--------------|---------|
| | Image → text | Text → image | Average |
| CCA (FV HGLMM) | 79.1 | 76.5 | 77.8 |
| CCA (FV GMM+HGLM) | 80.9 | 76.6 | 78.8 |
| DVSA | 80.5 | 74.8 | 77.7 |
| m-RNN | 83.5 | 77.0 | 80.3 |
| m-CNN | 76.7 | 81.3 | 82.1 |
| DSPE | 89.2 | 86.9 | 88.1 |
| ACMR | 93.2 | 87.1 | 90.2 |
| GRL (without FTLN) | 94.2 | 90.0 | 92.1 |
| GRL (Full) | 95.7 | 91.2 | 93.5 |

of the feature transfer learning process, we set up comparing experiments of original features and transferred features by different models. The MAP scores of this ablation study, as shown in [Table 8](#). We can observe that the GRL method's performances with the transferred features are better than with the original extracted features. For a large dataset, the FTLN model plays a more critical function in the cross-modal retrieval. For example, on the PKU XMediaNet dataset, the GRL with the original feature can only achieve the MAP score of 53.5, while the GRL with the features in the latent space can obtain significant progress, which improves the MAP score from 53.5 to 61.8. The experimental results of various FTLN model are shown in [Table 9](#). The comparisons of various models show that a deeper network is not certain to obtain better performance. For example, the MAP scores of three fully-connected layers are lower than those of one

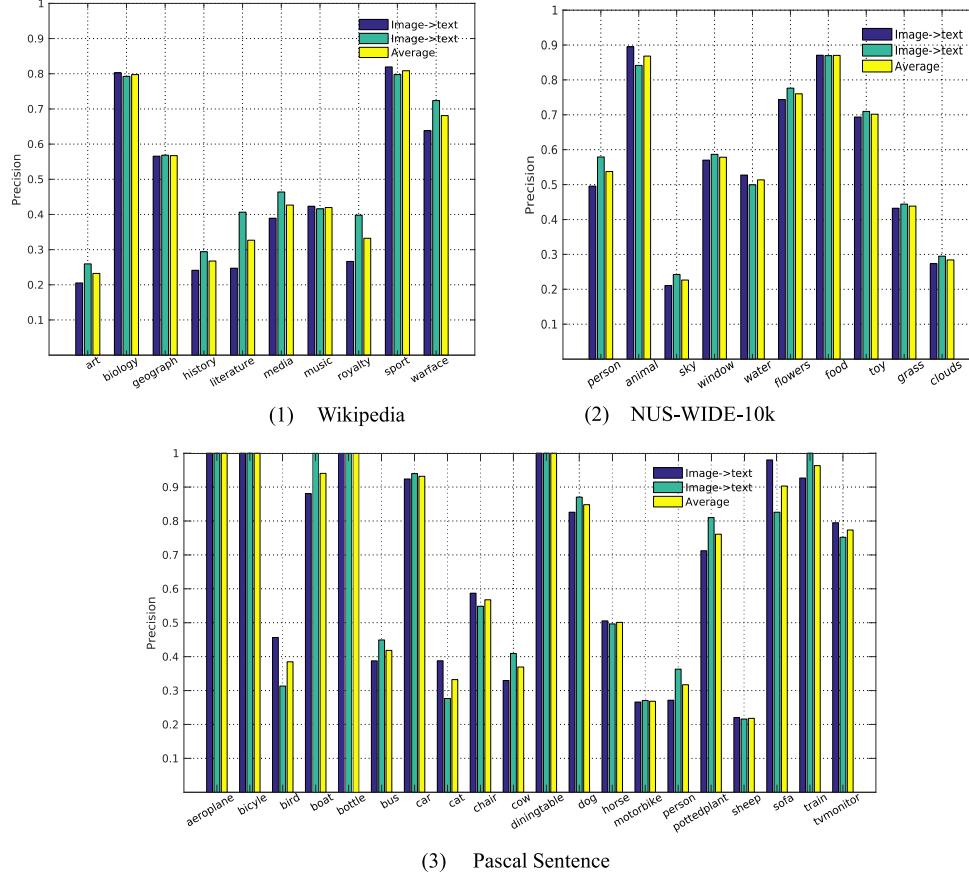


Fig. 14. MAP scores of different categories on Wikipedia dataset, NUS-WIDE-10k dataset, and Pascal Sentence dataset.

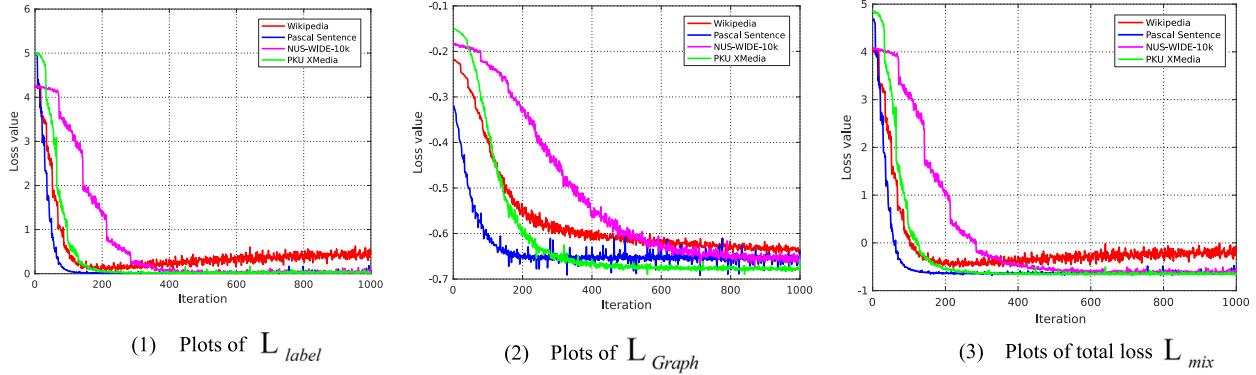


Fig. 15. The plots of different losses on Wikipedia dataset, NUS-WIDE-10k dataset, Pascal Sentence dataset, and PKU XMedia dataset.

or two fully-connected layers. The reason is that deep networks may lead to overfitting, which will decrease the performance of the testing dataset. We can conclude that a simple model can obtain remarkable performance for the features extracted from the pre-trained VGG19 and BoW models.

Furthermore, because the simple k -nearest neighbor algorithm is adopted to construct the cross-modal graph, we need to study the influence of the value of k and then find the most proper k for different datasets. The experimental results with various values of k are shown in Table 10. We can observe from the table that it cannot obtain the optimal cross-modal retrieval accuracy if the value of k is too large or too small. Besides, we plot the MAP curves of different k , as shown in Fig. 16. For the Wikipedia dataset, we can conclude that the most proper value of variable k is about 15. Because the size of the Pascal Sentence dataset

is relatively small, the most proper value of k is about 5. For the other three datasets, NUS-WIDE-10k dataset, PKU XMedia dataset, and PKU XMediaNet dataset, the most proper values of k are 20, 50, and 50, respectively. In general, for a larger scale dataset, the value of k should be bigger.

During the optimization of GRLN, the training balance of the supervised loss and unsupervised loss is vital for the performance of the proposed approach. We do more parameter analysis experiments to explore the impacts of these parameters α_0 and α_1 on Wikipedia dataset. All of the two parameters vary from 0 to 9 (0, 1, 3, 5, 7, 9). The experimental results of parameter sensitivity analysis are shown in Fig. 17. The first sub-figure shows that the proposed method can achieve relatively better performance only with the unsupervised loss. We can observe from the second sub-figure that the performance is relatively robust when varies

Table 6

Quantitative evaluation results of the cross-modal retrieval on Wikipedia, NUS-WIDE-10k, Pascal Sentence, PKU XMedia, PKU XMediaNet, and MSCOCO dataset in terms of Recall@K (R@K).

| Dataset | Image-text retrieval | | | Text-image retrieval | | | |
|-----------------|----------------------|-------|-------|----------------------|-------|-------|-------|
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@sum |
| Wikipedia | 48.9 | 60.8 | 62.9 | 73.4 | 94.1 | 97.0 | 437.2 |
| NUS-WIDE-10k | 61.6 | 79.6 | 84.1 | 65.7 | 75.9 | 79.1 | 446.0 |
| Pascal Sentence | 97.0 | 100.0 | 100.0 | 94.0 | 100.0 | 100.0 | 591.0 |
| PKU XMedia | 91.2 | 92.6 | 95.4 | 95.4 | 96.0 | 96.0 | 564.0 |
| PKU XMediaNet | 76.7 | 81.3 | 82.1 | 68.7 | 71.7 | 72.3 | 452.9 |
| MSCOCO | 90.0 | 95.0 | 96.1 | 95.4 | 98.6 | 99.1 | 574.3 |

Table 7

The MAP (@all) scores of the multi-modal cross-modal retrieval performance (bi-directional retrieval) on PKU XMedia dataset and PKU XMedianet Ddataset.

| Dataset | PKU XMedia | PKU XMediaNet |
|---------------|------------|---------------|
| Image → Text | 89.6 | 56.5 |
| Image → Video | 55.5 | 32.7 |
| Image → Audio | 48.9 | 19.1 |
| Image → 3D | 72.2 | 29.7 |
| Text → Image | 91.8 | 56.9 |
| Text → Video | 59.8 | 27.8 |
| Text → Audio | 51.2 | 16.8 |
| Text → 3D | 75.6 | 26.5 |
| Video → Image | 55.6 | 34.4 |
| Video → Text | 60.0 | 27.9 |
| Video → Audio | 37.0 | 14.4 |
| Video → 3D | 54.9 | 22.5 |
| Audio → Image | 49.4 | 20.1 |
| Audio → Text | 53.3 | 16.5 |
| Audio → Video | 37.6 | 14.7 |
| Audio → 3D | 47.6 | 16.3 |
| 3D → Image | 64.5 | 29.8 |
| 3D → Text | 68.5 | 24.4 |
| 3D → Video | 46.2 | 19.8 |
| 3D → Audio | 37.1 | 15.2 |
| Average | 57.8 | 26.1 |

the value of α_1 . In addition, the MAP scores are higher than the results of the control groups. The third sub-figure indicates that the model cannot be optimized without the unsupervised loss \mathcal{L}_{Graph} . The last sub-figure shows that the cross-modal retrieval performance decreases if enlarge the weight of unsupervised loss function (from 1 to 9). Comparing the first and the third sub-figures, we can draw a conclusion that the unsupervised loss can optimize the model independently while the supervised loss cannot. In general, the large weight of the supervised loss and small weight of the unsupervised have better robust and higher evaluation scores.

The experimental results of cross-modal retrieval indicate that the proposed GRL outperforms the state-of-the-art approaches. From the experimental results and the above analyses, the reasons can be concluded as the following: (1) The GRL bypasses the heterogeneity gap by reconstructing the original feature and their relationships into a cross-modal graph, which is the crucial part of the proposed method. Then, the GRL adopts a node-to-vector strategy to embedding the vertex on the graph into

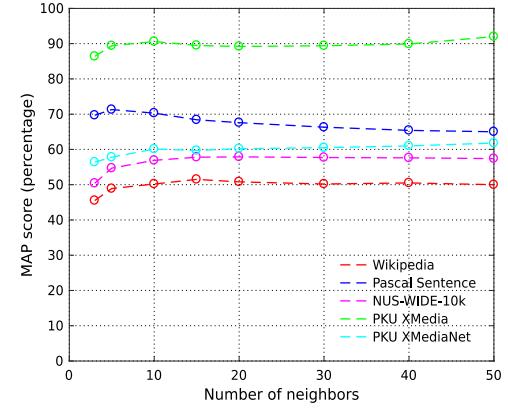


Fig. 16. The MAP scores of different number of neighbors on Wikipedia dataset, NUS-WIDE-10k dataset, Pascal Sentence dataset, and PKU XMedia dataset.

a low-dimensional vector. Compared with traditional common space learning methods, the GRL processes the relationships of original data instead of processing the extremely complex original features. Precisely because of the complexity of original features, there is a huge heterogeneity gap among different modalities, which improves the difficulty of cross-modal retrieval. Therefore, embedding the vertexes into a shared space is simpler than mapping the original data into an aligned representation space. (2) Before constructing the cross-modal graph, we first project the originally extracted features by five FTNL models, which consists of fully connected networks respectively. This step aims at transferring the original features into a latent space, where the use of cosine similarity to construct the cross-modal graph is reasonable. Therefore, the constructed graph can more reasonably represent the original data and their relationships. (3) During the graph optimization process, we adopt a distance minimization strategy if the two instances are in the k -nearest neighbors to learn the graph context. After sufficient game between different neighbors, the unlabeled graph vertexes will be embedded into a certain label space. The experimental results indicate that the value of variable k influences retrieval accuracy. In general, the value of k is determined by the size of the dataset.

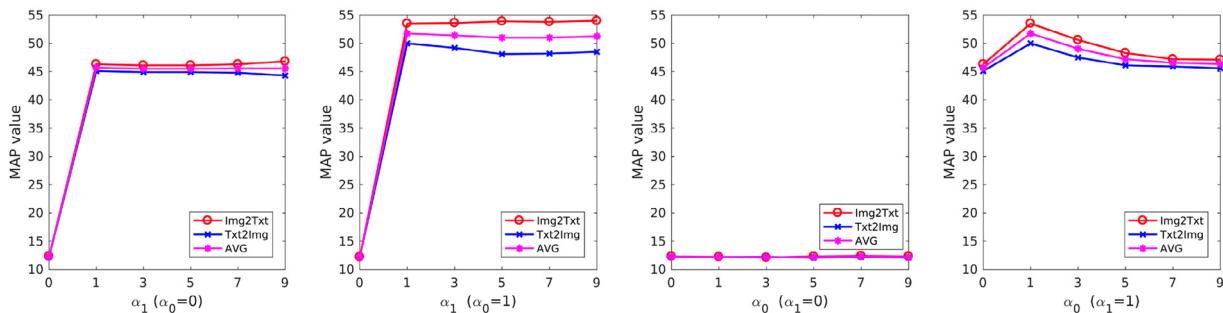


Fig. 17. A sensitivity analysis of the balancing parameters.

Table 8

MAP (@all) comparisons of ablation study (without FTLN and full model) for cross-modal retrieval (bi-directional retrieval and all-modal retrieval) on the five datasets.

| Dataset | Bidirectional retrieval | Without FTLN | Full | All-modal retrieval | Without FTLN | Full |
|-----------------|-------------------------|--------------|------|---------------------|--------------|------|
| Wikipedia | Image → Text | 53.2 | 54.7 | Image → all | 45.0 | 45.3 |
| | Text → Image | 48.2 | 50.0 | Text → all | 64.3 | 68.5 |
| | Average | 50.7 | 52.3 | Average | 54.7 | 56.9 |
| NUS-WIDE-10k | Image → Text | 54.4 | 57.9 | Image → all | 58.2 | 59.4 |
| | Text → Image | 55.2 | 59.4 | Text → all | 52.9 | 57.3 |
| | Average | 54.8 | 58.7 | Average | 55.6 | 58.4 |
| Pascal Sentence | Image → Text | 66.4 | 71.6 | Image → all | 66.9 | 69.2 |
| | Text → Image | 66.4 | 70.9 | Text → all | 68.1 | 70.6 |
| | Average | 66.4 | 71.3 | Average | 67.5 | 69.9 |
| PKU XMedia | Image → Text | 88.4 | 91.3 | Image → all | 87.4 | 89.1 |
| | Text → Image | 88.7 | 92.7 | Text → all | 89.9 | 94.4 |
| | Average | 88.5 | 92.0 | Average | 88.7 | 91.7 |
| PKU XMediaNet | Image → Text | 53.4 | 61.2 | Image → all | 63.1 | 68.1 |
| | Text → Image | 53.6 | 62.3 | Text → all | 48.6 | 59.0 |
| | Average | 53.5 | 61.8 | Average | 55.9 | 63.6 |

Table 9

MAP (@all) comparisons of various type of FTLN models for cross-modal retrieval (bi-directional retrieval) on the five datasets.

| Dataset | Bidirectional retrieval | Without FTLN | Fine-tune VGG19 | 1 FC layer | 2 FC layers | 3 FC layers |
|-----------------|-------------------------|--------------|-----------------|------------|-------------|-------------|
| Wikipedia | Image → Text | 53.2 | 53.7 | 53.5 | 54.7 | 54.3 |
| | Text → Image | 48.2 | 48.2 | 48.6 | 50.0 | 48.5 |
| | Average | 50.7 | 50.9 | 51.0 | 52.3 | 51.4 |
| NUS-WIDE-10k | Image → Text | 54.4 | 57.6 | 57.2 | 57.9 | 55.8 |
| | Text → Image | 55.2 | 58.0 | 58.5 | 59.4 | 58.4 |
| | Average | 54.8 | 57.8 | 57.9 | 58.7 | 57.0 |
| Pascal Sentence | Image → Text | 66.4 | 67.6 | 71.1 | 71.6 | 68.2 |
| | Text → Image | 66.4 | 67.4 | 69.8 | 70.9 | 67.7 |
| | Average | 66.4 | 67.5 | 70.5 | 71.3 | 68.0 |
| PKU XMedia | Image → Text | 88.4 | – | 88.9 | 91.3 | 86.8 |
| | Text → Image | 88.7 | – | 90.4 | 92.7 | 87.3 |
| | Average | 88.5 | – | 89.7 | 92.0 | 87.1 |
| PKU XMediaNet | Image → Text | 53.4 | – | 61.2 | 60.8 | 57.6 |
| | Text → Image | 53.6 | – | 62.3 | 60.4 | 58.4 |
| | Average | 53.5 | – | 61.8 | 60.6 | 58.0 |

Table 10

The MAP (@all) comparisons of different number of k on the five datasets.

| Dataset | Bidirectional retrieval | Number of neighbors | | | | | | | |
|-----------------|-------------------------|---------------------|------|------|------|------|------|------|------|
| | | 3 | 5 | 10 | 15 | 20 | 30 | 40 | 50 |
| Wikipedia | Image → Text | 48.7 | 52.5 | 53.1 | 54.3 | 53.3 | 53.6 | 53.3 | 52.5 |
| | Text → Image | 42.4 | 45.7 | 47.4 | 48.8 | 48.4 | 46.9 | 47.8 | 47.5 |
| | Average | 45.5 | 48.9 | 50.2 | 52.3 | 50.8 | 50.2 | 50.5 | 50.0 |
| NUS-WIDE-10k | Image → Text | 49.6 | 54.3 | 56.3 | 57.1 | 57.9 | 57.0 | 57.1 | 56.9 |
| | Text → Image | 51.2 | 55.2 | 57.5 | 58.4 | 59.4 | 58.4 | 58.0 | 57.9 |
| | Average | 50.4 | 54.7 | 56.9 | 57.8 | 58.7 | 57.7 | 57.6 | 57.4 |
| Pascal Sentence | Image → Text | 69.8 | 71.6 | 71.2 | 68.3 | 67.4 | 65.5 | 65.5 | 64.0 |
| | Text → Image | 69.6 | 70.9 | 69.4 | 68.5 | 67.8 | 67.0 | 65.2 | 66.0 |
| | Average | 69.7 | 71.3 | 70.3 | 68.4 | 67.6 | 66.3 | 65.4 | 65.0 |
| PKU XMedia | Image → Text | 85.7 | 89.6 | 90.5 | 89.1 | 88.7 | 89.2 | 89.0 | 91.3 |
| | Text → Image | 87.1 | 89.1 | 90.6 | 89.9 | 89.4 | 89.5 | 90.7 | 92.7 |
| | Average | 86.4 | 89.4 | 90.6 | 89.5 | 89.2 | 89.4 | 89.9 | 92.0 |
| PKU XMediaNet | Image → Text | 55.9 | 57.8 | 60.1 | 59.8 | 60.2 | 60.2 | 60.5 | 61.2 |
| | Text → Image | 56.8 | 57.7 | 60.0 | 59.6 | 60.2 | 60.7 | 61.5 | 62.3 |
| | Average | 56.4 | 57.8 | 60.1 | 59.7 | 60.2 | 60.5 | 61.0 | 61.8 |

5. Conclusion

This paper presents a simple but effective approach called Graph Representation Learning (GRL) for cross-modal retrieval. This method inherits the previous cross-modal retrieval idea that maps the different types of media data features into aligned representation space. However, we realize this idea by graph embedding strategy, which first reconstructs the original data into a cross-modal graph and then embeds each vertex into a low-dimensional vector in common representation space. Besides, to

make the cross-modal graph more reasonable for representing original features and their relationships, we design an FTLN model to transfer the original feature into latent hidden feature. The graph connections consist of three parts, semantic connection and pairing connection of training set, and similar local connections for all instances. By embedding the nodes into unique vectors, the features of various modalities are located in the aligned representation space. Compared with the state-of-the-art methods, the experimental results show that the proposed method obtains excellent performance.

Declaration of competing interest

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

Acknowledgments

This research is supported by the National Natural Science Foundation of China under grants 61771145 and 61371148.

References

- Andrew, G., Arora, R., Bilmes, J., & Livescu, K. (2010). Deep canonical correlation analysis. In *International conference on machine learning* (pp. 3408–3415).
- Chen, D.-Y., Tian, X.-P., Shen, Y.-T., & Ouhyoung, M. (2003). On visual similarity based 3D model retrieval. *Computer Graphics Forum*, 22(3), 223–232.
- Chua, T., 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 *ACM international conference on image and video retrieval*.
- Ciresan, D. C., Meier, U., Masci, J., Maria Gambardella, L., & Schmidhuber, J. (2011). Flexible, high performance convolutional neural networks for image classification. In *Twenty-second international joint conference on artificial intelligence: Vol. 22*, (1), (p. 1237).
- Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2), 5.
- Deng, C., Tang, X., Yan, J., Liu, W., & Gao, X. (2015). Discriminative dictionary learning with common label alignment for cross-modal retrieval. *IEEE Transactions on Multimedia*, 18(2), 208–218.
- Deng, C., Yang, E., Liu, T., Li, J., Liu, W., & Tao, D. (2019). Unsupervised semantic-preserving adversarial hashing for image search. *IEEE Transactions on Image Processing*, 28(8), 4032–4044.
- Deng, C., Yang, E., Liu, T., & Tao, D. (2019). Two-stream deep hashing with class-specific centers for supervised image search. *IEEE Transactions on Neural Networks Learning Systems*.
- Feng, F., Wang, X., & Li, R. (2014). Cross-modal retrieval with correspondence autoencoder. In *ACM international conference on multimedia* (pp. 7–16).
- Gong, Y., Ke, Q., Isard, M., & Lazebnik, S. (2014). A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision*, 106(2), 210–233.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).
- Gu, J., Cai, J., Joty, S. R., Niu, L., & Wang, G. (2018). Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7181–7189).
- Han, W., Chan, C. F., Choy, C. S., & Pun, K. P. (2006). An efficient MFCC extraction method in speech recognition. In *2006 IEEE international symposium on circuits and systems* (pp. 4-pp).
- Hardoon, D. R., Szedm'ak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12), 2639–2664.
- Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4), 321–377.
- Huang, X., & Peng, Y. (2019). TPCKT: Two-level progressive cross-media knowledge transfer. *IEEE Transactions on Multimedia*, 21(11), 2850–2862.
- Huang, X., Peng, Y., & Yuan, M. (2017). Cross-modal common representation learning by hybrid transfer network. In *Twenty-Sixth international joint conference on artificial intelligence*.
- Huang, X., Peng, Y., & Yuan, M. (2020). MHTN: Modal-adversarial hybrid transfer network for cross-modal retrieval. *IEEE Transactions on Cybernetics*, 50(3), 1047–1059.
- Kang, C., Xiang, S., Liao, S., Xu, C., & Pan, C. (2015). Learning consistent feature representation for cross-modal multimedia retrieval. *IEEE Transactions on Multimedia*, 17(3), 370–381.
- Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (3128–3137).
- Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *IEEE conference on computer vision and pattern recognition* (pp. 1725–1732).
- Klein, B., Lev, G., Sadeh, G., & Wolf, L. (2015). Associating neural word embeddings with deep image representations using fisher vectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4437–4446).
- Li, D., Dimitrova, N., Li, M., & Sethi, I. K. (2003). Multimedia content processing through cross-modal association. In *ACM international conference on multimedia* (pp. 604–611).
- Liang, J., Li, Z., Cao, D., He, R., & Wang, J. (2016). Self-paced cross-modal subspace matching. In *Proc. int. ACM SIGIR conference on research and development in information retrieval* (pp. 569–578).
- Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755).
- Ma, L., Lu, Z., Shang, L., & Li, H. (2015). Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE international conference on computer vision* (pp. 2623–2631).
- Maaten, L. V. D., & Hinton, G. (2008). Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, 2579–2605.
- Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., & Yuille, A. (2014). Deep captioning with multimodal recurrent neural networks (m-RNN). arXiv preprint arXiv: 1412.6632.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746–748.
- McKay, C., Fujinaga, I., & Depalle, P. (2005). jaudio: An feature extraction library. In *International conference on music information retrieval* (pp. 600–603).
- Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In *International conference on machine learning* (pp. 689–696).
- Peng, Y., Huang, X., & Qi, J. (2016). Cross-media shared representation by hierarchical learning with multiple deep networks. In *International joint conference on artificial intelligence* (pp. 3846–3853).
- Peng, Y., Huang, X., & Zhao, Y. (2017). An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9), 2372–2385.
- Peng, Y., & Qi, J. (2019). CM-GANs: Cross-modal generative adversarial networks for common representation learning. *ACM Transactions on Multimedia Computing Communications and Applications*, 15(1).
- Peng, Y., Qi, J., Huang, X., & Yuan, Y. (2018). CCL: Cross-modal correlation learning with multi-grained fusion by hierarchical network. *IEEE Transactions on Multimedia*, 20(2), 405–420.
- Peng, Y., Qi, J., & Yuan, Y. (2018). Modality-specific cross-modal similarity measurement with recurrent attention network. *IEEE Transactions on Image Processing*, 27(11), 5585–5599.
- Peng, Y., Zhai, X., Zhao, Y., & Huang, X. (2016). Semi-supervised cross-media feature learning with unified patch graph regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(3), 583–596.
- Pereira, J. C., Covello, E., Doyle, G., Rasiwasia, N., Lanckriet, G. R., Levy, R., et al. (2013). On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3), 521–535.
- Ranjan, V., Rasiwasia, N., & Jawahar, C. V. (2015). Multi-label cross-modal retrieval. In *Proc. IEEE international conference on computer vision* (pp. 4094–4102).
- Rashtchian, C., Young, P., Hodosh, M., & Hockenmaier, J. (2010). Collecting image annotations using Amazon's mechanical turk (pp. 139–147). North American Chapter of the Association for Computational Linguistics.
- Rasiwasia, N., Costa Pereira, J., Covello, E., Doyle, G., Lanckriet, G. R., Levy, R., et al. (2010). A new approach to cross-modal multimedia retrieval. In *ACM international conference on multimedia* (pp. 251–260).
- Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91–99).
- Rosipal, R., & Kramer, N. (2006). Overview and recent advances in partial least squares. In *Proc. statistical optimization perspectives workshop: Subspace, latent struct. feature selection* (pp. 34–51).
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
- Srivastava, N., & Salakhutdinov, R. (2012). Learning representations for multi-modal data with deep belief nets. In *International conference on machine learning*.
- Tieleman, T., & Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 4(2), 26–31.
- Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In *IEEE international conference on computer vision* (pp. 4489–4497).
- Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7167–7176).
- Wang, F., Cheng, J., Liu, W., & Liu, H. (2018). Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7), 926–930.
- Wang, K., He, R., Wang, L., Wang, W., & Tan, T. (2015). Joint feature selection and subspace learning for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10), 2010–2023.

- Wang, L., Li, Y., & Lazebnik, S. (2016). Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5005–5013).
- Wang, B., Yang, Y., Xu, X., Hanjalic, A., & Shen, H. T. (2017). Adversarial cross-modal retrieval. In *ACM international conference on multimedia* (pp. 154–162).
- Wang, L., Zhu, L., Dong, X., Liu, L., Sun, J., & Zhang, H. (2018). Joint feature selection and graph regularization for modality-dependent cross-modal retrieval. *Journal of Visual Communication and Image Representation*, 54, 213–222.
- Wei, Y., Zhao, Y., Lu, C., Wei, S., Liu, L., Zhu, Z., et al. (2017). Cross-modal retrieval with CNN visual features: A new baseline. *IEEE Transactions on Cybernetics*, 47(2), 449–460.
- Wen, X., Han, Z., Yin, X., & Liu, Y. S. (2019). Adversarial cross-modal retrieval via learning and transferring single-modal similarities. arXiv preprint arXiv: 1904.08042.
- Wu, Y., Wang, S., & Huang, Q. (2018). Learning semantic structure-preserved embeddings for cross-modal retrieval. In *2018 ACM multimedia conference on multimedia* (pp. 825–833).
- Xu, R., Li, C., Yan, J., et al. (2019). Graph convolutional network hashing for cross-modal retrieval. In *Proceedings of the twenty-eighth international joint conference on artificial intelligence* (pp. 10–16).
- Xu, X., Song, J., Lu, H., Yang, Y., Shen, F., & Huang, Z. (2018). Modal-adversarial semantic learning network for extendable cross-modal retrieval. In *Proceedings of the 2018 ACM on international conference on multimedia retrieval* (pp. 46–54).
- Yang, E., Deng, C., Li, C., Liu, W., Li, J., & Tao, D. (2018). Shared predictive cross-modal deep quantization. *IEEE Transactions on Neural Networks Learning Systems*, 29(11), 5292–5303.
- Yu, J., Lu, Y., Qin, Z., Zhang, W., Liu, Y., Tan, J., et al. (2018). Modeling text with graph convolutional network for cross-modal information retrieval. In *Pacific rim conference on multimedia* (pp. 223–234).
- Zhai, X., Peng, Y., & Xiao, J. (2014). Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(6), 965–978.
- Zhang, X., Lai, H., & Feng, J. (2018). Attention-aware deep adversarial hashing for cross-modal retrieval. In *Proceedings of the European conference on computer vision* (pp. 591–606).
- Zhang, L., Ma, B., Li, G., Huang, Q., & Tian, Q. (2017). Multi-networks joint learning for large-scale cross-modal retrieval. In *Proceedings of the 25th ACM international conference on multimedia* (pp. 907–915).
- Zhang, J., & Peng, Y. (2019). Multi-pathway generative adversarial hashing for unsupervised cross-modal retrieval. *IEEE Transactions on Multimedia*.