Cross-Modal Retrieval via Similarity-Preserving Learning and Semantic Average Embedding

TAO ZHI, YINGCHUN FAN, AND HONG HAN, (Member, IEEE)
School of Artificial Intelligence, Xidian University, Xi’an 710071, China

Corresponding author: Hong Han (hanh@mail.xidian.edu.cn)

This work was supported by the Open Fund of Shaanxi Key Laboratory of Integrated and Intelligent Navigation under Grant SKLIIN-20180102 and Grant SKLIIN-20180107.

ABSTRACT Cross-modal retrieval takes one modality data as the query to search related data from different modalities (e.g., images vs. texts). As the heterogeneous gap exists between different media data, mainstream methods focus on reducing modality gap using common space learning. However, the heterogeneous media gap is big and it is too hard to be eliminated completely. Besides this, the representations of the same modality are diverse, which is important but is ignored by most existing methods. In this paper, we propose a novel cross-modal retrieval via Similarity-preserving Learning and Semantic Average Embedding (SLSAE) method. There are two key ideas in our method, one is to reduce modality gap by similarity-preserving learning, the other is to use semantic average embedding to weaken the impact of diversity existing in the common space. The similarity-preserving learning process will push embeddings from the same category together and pull embeddings from different categories apart. Eliminating the influence of embeddings diversity can improve performance and robustness, which is more friendly to real-world cross-modal retrieval applications. The model of proposed method is concise, and can be extended to multimodal cross-modal retrieval situation flexibly. Comprehensive experimental results show that our method significantly outperforms the state-of-the-art methods in bimodal cross-modal retrieval, and it also achieves excellent performance in multimodal retrieval scenarios.

INDEX TERMS Common space learning, cross-modal retrieval, multimodal retrieval, similarity-preserving learning.

I. INTRODUCTION

With the development of Internet and digital media technology, we have entered the era of big data. People generate amounts of data every day, such as images, videos, audios and so on. These data are diverse and disorganized, and people have the demand to retrieve them. Under this situation, cross-modal retrieval has become a key research topic, which is proposed to perform retrieval across various modalities [6]. Cross-modal retrieval is different from traditional single-modal retrieval, where the query data and retrieval results are from different modalities, like images and texts. For example, when a horse image is given as query, retrieve related text from text modality, and vice versa. Cross-modal retrieval has a wide range of applications, and it can be used in intelligence searching engine and multimedia data management.

However, it is still a challenging problem due to the modality gaps between different media data.

The features in different modalities are related to its media data structure, such as images including edges and corners, while texts including words and punctuation. As the big media gap exists in different modalities data, the features and their distributions are quite different, which make it impossible to compare and retrieve in original feature space directly. Naturally, the mainstream methods focus on eliminating heterogeneous gap. These methods tried to reduce media gap by projecting different modalities data to the common subspace. The main idea of common space learning is that the data belonging to the same category has latent correlations. If we transform features from original space to the common space, the embeddings of all modalities can be gathered together according to its semantic label. The embedding in the common space represents its semantic feature information of different modalities so it can be compared and retrieved directly.
FIGURE 1. The overall framework of the proposed SLSAE. Pair inputs will be transformed to common space through modality-specific embedding network. Inner-modality and inter-modality loss concentrates on semantic similarity in the same modal and cross modal, respectively. The joint loss contains label classifier supervised loss, pair loss, inner-modality and inter-modality semantic similarity distance constraint in the common space. Semantic average embedding is produced according to the label classifier prediction and used for better retrieval performance.

Although the previous cross-modal retrieval methods perform well, there are still two problems to be solved: 1) the big modality heterogeneous gap cannot be removed completely ideally, and 2) there is also diversity existing in embeddings of the same modality, even though they belong to the same object. The modality heterogeneous gap influences retrieval results and the diversity of embeddings will also impact retrieval performance and robustness heavily.

We propose a novel method which is called cross-modal retrieval via similarity-preserving learning and semantic average embedding (SLSAE) to address those issues mentioned above. As illustrated in Figure 1, our method is built around semantic similarity-preserving learning and semantic average embedding. The highlights of SLSAE framework is reducing modality gap by distance constraint, and moreover, the influence of embedding diversity existing in the common space is weakened with semantic average embedding.

In our SLSAE, the modality-specific embedding network transforms the features from original space to the common space. The semantic similarity-preserving learning controls the distance between embeddings, which will push the embeddings belonging to the same object closer and pull different object embeddings further apart, therefore the embeddings are easier to be distinguished from others. Since there is also diversity existing in embeddings belonging to the same object, the average embedding of all embeddings belonging to the same label could be utilized to retrieve. Furthermore, the model of SLSAE is designed so light and concise that can be extended to multimodal situation flexibly and easily. Our contributions can be summarized as follows.

1) A novel semantic similarity-preserving learning method is proposed, which tend to gather different modalities data in the common space. The embeddings of the same category are clustered together, meanwhile the different categories embeddings are dispersed, which is easier to retrieve.

2) The semantic average embedding is used to retrieve instead of original query embedding. The embedding of the same object is still diverse, especially some outliers, heavily affect the retrieval performance. The semantic average embedding can reduce embedding diversity significantly, and also improve retrieval effectiveness and robustness. In the real-world cross-modal retrieval applications, the distribution of query data is often quite different from training data, thus the utilization of
semantic average embedding is more friendly to real-world applications.

3) The entire model is concise and can be extended to multimodal retrieval situation flexibly. Our proposed method is simple and it only contains modality-specific embedding networks and a common label classifier network, therefore it is of high maneuverability and has the actual application value.

The proposed SLSAE method has been evaluated on four widely used benchmark datasets (three for two modalities and one for five modalities) and compared with several existing methods. The experimental results show that SLSAE outperforms other state-of-the-art methods significantly.

The rest of the paper is organized as follows: In section II, a brief introduction of related work is given. And the detailed description of SLSAE is given in section III. Then the experimental results and further analysis are provided in section IV. Section V just concludes this paper.

II. RELATED WORK

The existing cross-modal retrieval methods can be simply divided into two categories, real-valued and binary-valued representation learning methods [1]. The binary-valued representation learning methods is also named as hashing based methods [7], [15], [17], [18], [28], [33], [34], [37], which aim at mapping different modalities data into the binary Hamming space. In the Hamming space, all representations are encoded by 0 and 1. The advantage of this kind methods is effective and time-saving, but its retrieval accuracy is not as good as other methods. Like the binary-valued methods, the real-valued approaches also have subspace, but the representation is real-valued and can be compared directly by widely used distance metric, such as Euclidean distance and cosine distance etc.

The common space learning based approaches is a typical subcategory of real-valued methods. Until now, a variety of common space learning methods [8], [11], [13], [22], [23] for cross-modal retrieval have been proposed. The canonical correlation analysis (CCA) [8] is one of the most representative traditional statistical methods. CCA aims at learning a subspace that can maximize the pairwise correlations between two modalities data. However, the correlation of real world multimodal data is too sophisticated to be captured by linear projections. CCA is an early method, and there are also some variant methods based it. Generalized multiview analysis (GMA) [19] extended CCA with semantic information, and some works extended CCA to nonlinear models with the kernel tricks, such as Kernel CCA (KCCA) [2]. However, because the kernel is predetermined, the representation learned from modalities data is limited. With the rapid development of deep learning, deep neural network (DNN) has important applications in related fields. Deep canonical correlation analysis (DCCA) [14] is another nonlinear extension of CCA based on DNN.

As DNN has strong feature extraction capabilities, a large number of DNN-based methods have been proposed and achieved excellent performance with common space leaning. Peng et al. [11] proposed cross-media multiple deep networks (CMDN) with a hierarchical architecture. CMDN generated two kinds of separated but complementary representations for each modality, and combined the representations hierarchically to learn the common space. Corr-AE [13] proposed the correspondence autoencoders, which considered the reconstruction errors and the correlation losses with autoencoders. The comprehensive distance-preserving autoencoders (CDP) [10] is an unsupervised method, which aims to preserve the respective distances between the representations in common space, so that they are consistent with the distances in their original media spaces. The domain adaptation with scene graph (DASG) [39] transfers knowledge from the source domain to target domain with scene graph.

Goodfellow et al. [20] proposed generative adversarial networks (GAN), which trained a generative and discriminative model simultaneously via an adversarial process. There are also some GAN-based methods wanted to learn the common space in an adversarial way. The adversarial cross-modal retrieval (ACMR) [5] combined adversarial learning and triplet loss to preserve modality level semantic embedding structure in the common space, which achieved a higher performance improvement. In ACRE [5], two feature projectors and the modality classifier acted as generator and discriminator respectively, and worked together for adversarial learning. If the modality classifier cannot distinguish the embeddings modality in the common space, the feature projectors generate modality-invariant embeddings, so it can reduce the modality gap. The cross-modal similarity transfering (CMST) [38] tries to learn the single-modal similarities and transfer them to the common subspace with adversarial learning. The modal-adversarial hybrid transfer network (MHTN) [25] aims to transfer knowledge from a single-modal source domain to a cross-modal target domain and learn the cross-modal common representation. The adversarial cross-modal retrieval based on dictionary learning algorithm (DLA-CMR) [36] uses dictionary learning and adversarial learning to reconstruct discriminative features and mine the statistical characteristics respectively. The deep adversarial metric learning approach (DAML) [40] maximizes the correlations between modalities and introduces adversarial learning as additional regularization. The ternary adversarial networks with self-supervision (TANSS) [41] uses thee subnetworks to tackle zero-shot retrieval problem. The self-supervised network in TANSS can leverage the word vectors of both seen and unseen categories as guidance to supervise the feature extraction network. Based on the fact that learning the low-level feature is the basis of learning the high-level representation, the unified binary generative adversarial network (BGAN+) [45] converts images to binary codes for both image compression and retrieval tasks in a multi-task fashion and an unsupervised way.

Compared to other methods, these GAN-based methods can improve cross-modal retrieval performance to some
III. PROPOSED METHOD

In this section, we first introduce the formulation of cross-modal retrieval problem. And then, the design of semantic similarity-preserving learning and how to use semantic average embedding for cross-modal retrieval is proposed. How to extend this method to multimodal is introduced. More information about the implementation details are also given.

A. PROBLEM FORMULATION

Though SLSAE can be extended to multimodal retrieval flexibly, without losing generality, here we mainly focus on bimodal retrieval problem (images and texts). Multimodal retrieval is similar to bimodal problem. We assume that there is a collection of images and texts. Let $V = \{v_1, v_2, \ldots, v_n\}$ be the collection of images, and $T = \{t_1, t_2, \ldots, t_n\}$ be the collection of texts, $v_i \in R^{d_i}, t_i \in R^{d_t}$ where $d_i$ and $d_t$ are the input dimensions of the image and text features respectively, and $n$ is the number of instance in the collection. Pairs of image-text input $(v_i, t_j)$ could be formed from the image collection $V$ and text collection $T$. Each image-text pair is also assigned with a semantic label vector $y_i = [y_{i1}, y_{i2}, \ldots, y_{ic}]$ where $c$ denotes the number of categories in the collection. If the $i$-th pair belongs to the $j$-th category, $y_{ij} = 1$, otherwise $y_{ij} = 0$.

Since the statistical properties and distributions of image features $V$ and text features $T$ are quite different, they cannot be compared directly for cross-modal retrieval task. Our proposed method SLSAE aims at finding a common space, which can preserve semantic similarity for comparisons and retrieval.

B. FRAMEWORK OF SLSAE

The general framework of proposed SLSAE is shown in Figure 1. It includes two pairs of inputs $(v_i, t_i)$ and $(v_j, t_j)$, feature embedding networks, label classifier network, joint loss and the process of obtaining semantic average embedding for retrieval.

First, two batch image-text pair inputs are fed to the embedding network to produce embeddings in common space. Image modality and text modality have their own embedding network individually. Embedding network plays the role of transforming input features from quite different original space to the same common space. Next, the label classifier is deployed to predict semantic labels of the embeddings in the common space. It is noted that there is only one classifier network. It can classify all embeddings no matter whether the embedding belongs to image or text modality. And then, the model is trained with joint loss, including label classifier supervised loss, pair inputs loss and semantic similarity distance constraint in inner-modality and inter-modality. The inner-modality loss reflects the embeddings of the same object in the same modality, and get them closer. The inter-modality loss reflects the similarity of embeddings from different modalities containing the same and different objects. It will make the similarity of embeddings of the same object bigger in different modalities, on the contrary, reduce the similarity of different objects from different modalities.

Finally, we get the semantic average embeddings, and explore the hidden information of average embeddings for improving retrieval performance.

C. SIMILARITY-PRESERVING LEARNING

The embedding networks transform features to the common space, which will make embeddings have some common statistical properties and unique characteristics. In the common space, the distance of embeddings represents the similarity, the smaller distance, the more similar, and vice versa.

The best ideal situation is that the embeddings with the same label gather to the same point in the common space, and the embeddings are far apart if they belong to different semantic labels. If the embeddings distribution of query data satisfies the above ideal situation, it can be retrieved perfectly. However, the fact is that we cannot achieve the best ideal
situation, and these embeddings are easily messy and mixed together, which makes it hard to retrieve accurately.

The SLSAE designs the loss function from the perspective of embedding distance constraint. Our method uses cosine distance to measure the similarity of embeddings in the common space. It is given as:

$$s(x, y) = \cos(x, y) = \frac{x \cdot y}{\|x\| \|y\|}$$  \hspace{1cm} (1)$$

where, $x$ and $y$ are embeddings of two instances. Euclidean distance can also be used to measure similarity, but it is absolute distance of points, and directly related to the position coordinate of each point. The cosine distance measures the angle of the embeddings, which reflects the influence of difference in direction than the position, therefore it is more suitable for retrieval task.

We use $D(x, y)$ to measure the distance of two embeddings, which is used in objective function, and it is given as:

$$D(x, y) = 1 - s(x, y) = 1 - \frac{x \cdot y}{\|x\| \|y\|}.$$  \hspace{1cm} (2)$$

As we can see, $D(x, y) \in [0, 2]$. It stands for dissimilarity, the greater distance, the less similar.

We use $E$ as the embedding network transforming process. $E(v_i), E(t_i), E(v_j), E(t_j)$ represent the embeddings of $v_i, t_i, v_j,$ and $t_j$ in the common space respectively.

In each iteration, we have two batches of image-text pair inputs. The distances of pairwise inputs ($v_i, t_i$) and ($v_j, t_j$) are expected to be as small as possible. The pairwise loss reflects the embeddings of the same semantic label for different modalities, and it is defined as:

$$L_{pair} = D(E(v_i), E(t_i)) + D(E(v_j), E(t_j)).$$  \hspace{1cm} (3)$$

The inner-modality loss mainly focuses on the embedding distance of the same object in the same modality. The ideal situation is that all embeddings gather together if they belong to the same category. Therefore, we use Euclidean distance as the objective function to push the embeddings belonging to one object closer in the same modality.

$$L_{inner} = \|E(v_i) - E(v_j)\|_2 \cdot p(v_i, v_j) + \|E(t_i) - E(t_j)\|_2 \cdot p(t_i, t_j) \hspace{1cm} (4)$$

where $p(x, y) = 1$, if $x$ and $y$ belong to the same category, otherwise 0, no matter the modality of $x$ and $y$.

The inter-modality loss works on the embeddings distance of different modalities, both for the same object and different objects. For the same object part, it is similar to the pairwise loss. As for the different object part, we limit them with a margin.

$$L_{inter} = D(E(v_i), E(t_j)) \cdot p(v_i, t_j) + D(E(t_i), E(v_j)) \cdot p(t_i, v_j) + \max(0, \mu - D(E(v_i), E(t_j))) \cdot (1 - p(v_i, t_j)) + \max(0, \mu - D(E(t_i), E(v_j))) \cdot (1 - p(t_i, v_j)).$$  \hspace{1cm} (5)$$

Here the first two items are applicable for the same object part, and the last two items are applicable for different objects meanwhile, where $\mu$ is the margin.

The inter-modality loss will reduce the embedding distance of the same semantic label of different modalities, and tries to keep the embedding distance of different label with a boundary, which makes them easier to distinguish. The inner-modality loss enforces the embeddings belonging to the same label gather together. The pairwise loss focuses on reducing the embedding distance of pair inputs.

Besides the pairwise loss, inner-modality loss and intermodality loss, the label classifier also has a supervised loss. The label classifier predicts the semantic label of embeddings, which can adjust the embedding networks to produce better embeddings with semantic similarity-preserving. Furthermore, the label classifier also can be used to obtain semantic average embedding for retrieval. The classifier outputs a probability distribution of semantic categories for each embedding, and it is independent of the embedding modality. The supervised label classifier loss is given as:

$$L_{lab} = -[y_i \cdot (\log \hat{p}(E(v_i)) + \log \hat{p}(E(t_i))) + y_j \cdot (\log \hat{p}(E(v_j)) + \log \hat{p}(E(t_j)))] \hspace{1cm} (6)$$

where the $\hat{p}(E(x_i))$ is the probability distributions of $x_i$. Label loss is the cross-entropy loss of semantic category classification of embedding. The label classifier will promote the modality-specific embedding network to generate semantic-specific embedding clusters.

The joint loss consists of the four parts mentioned above.

$$L_j = \frac{1}{n} \sum_{i=1}^{n} (L_{pair} + \alpha \cdot L_{inner} + \beta \cdot L_{inter} + L_{lab}) \hspace{1cm} (7)$$

where $\alpha$ and $\beta$ are weight coefficients of the inner-modality loss and inter-modality loss, which can be adjusted; and $n$ is the number of pair inputs in one batch. The loss function is calculated in a mini-batch inputs.

### D. SEMANTIC AVERAGE EMBEDDING FOR RETRIEVAL

In section II, some GAN-based methods use adversarial learning and triplet constraint to reduce distance of embeddings from different modalities with the same semantic label. These methods have made some achievements in reducing modality gap in common space to a certain extent. But the embeddings still cannot be separated well, and there are some gaps between training set and test set. Our proposed SLSAE exploits the hidden information in semantic average embedding to retrieve from cross modality.

The most commonly used cross-modal retrieval method is described as follow. Firstly, given an instance (e.g. an image) in test data as the query; secondly, get its embedding with its corresponding embedding network; then, calculate the query embedding distances with all embeddings in another modality test data (e.g. text); finally, sort the distances in ascending order, the higher the ranking, the more similar.
If the test set and training set of all modalities have the same distribution, their embeddings will also have similar distribution. In this case, it can be retrieved well. But if there is a gap between the distribution of test set and training set in some modalities, the retrieval performance will get worse. Since the distribution difference of test set and training set exists in original feature space, the difference will further increase in common space, which will greatly influences retrieval results. Figure 2 shows the 2d t-SNE result for training set and test set in the Wikipedia dataset. We can see that, there exists difference of distribution in training set and test set exists in different modalities, as shown in Figure 3, we did not mix the average embeddings from different modalities together. The pipeline of retrieving with semantic average embedding is described as follow. Firstly, given a query (e.g. an image), its embedding is obtained with its corresponding embedding network; secondly, the label classifier prediction class could act as its category; then, the average embedding of another modality (e.g. text modality) is utilized, the retrieval performance could be improved greatly.

The semantic average embedding is obtained in training set. When retrieving, given a query sample, the label predicted by the label classifier is used as its true label; then its corresponding average embedding is applied to retrieve. It should be noted that, the semantic average embedding is obtained in image and text modality, individually. Because diversity exists in embeddings belonging to the same category in different modalities, as shown in Figure 3, we did not mix the average embeddings from different modalities together. The pipeline of retrieving with semantic average embedding is described as follow. Firstly, given a query (e.g. an image), its embedding is obtained with its corresponding embedding network; secondly, the label classifier prediction class could act as its category; then, the average embedding of another modality (e.g. text modality) is selected according to the label predicted by the label classifier; and finally, after a comparison of distance, the retrieved result is returned.

E. EXTENSION TO MULTIMODAL RETRIEVAL

The model in our method is concise, and it is easy to extend to multimodal cross-modal retrieval scene. For $M$ modalities, the two-modality retrieval methods must perform $M(M−1)/2$ times training process to retrieve from any modality. While in SLSAE, for $M$ modalities, we built $M$ modality-specific embedding network, and a common label classifier network. The model in SLSAE only needs to be trained once for multimodal retrieval. If there are $M$ modalities, $x^k$ represents $k$-th modality data. The pair loss, inner-modality and inter-modality loss can be extended to multimodal situation in a linear way.

\[
L_{\text{pair-multi}} = \sum_{k=1}^{M-1} L_{\text{pair}}(x^k, x^{k+1}).
\]

\[
L_{\text{inner-multi}} = \sum_{k=1}^{M-1} L_{\text{inner}}(x^k, x^{k+1}).
\]

\[
L_{\text{inter-multi}} = \sum_{k=1}^{M-1} L_{\text{inter}}(x^k, x^{k+1}).
\]
F. IMPLEMENTATION DETAILS

In this work, there are two embedding networks for image and text modality, and a common label classifier network. For embedding network, we use feed-forward neural network activated by tanh function, i.e. ($input \rightarrow 512 \rightarrow 128$). We set the dimension of embedding in common space to 128. A feed-forward neural network is deployed as the label classifier, i.e. ($128 \rightarrow 64 \rightarrow 10$). Furthermore, the softmax activation layer is added after the last layer of the label classifier. The CDPAE [10] shows that the features are accompanied by redundant noise. In order to reduce the negative influence of noises, some input components are set to zero randomly, similar to CDPAE. The margin of inter-modality loss $\mu$ is set to 0.4, and the hyper parameters $\alpha$ and $\beta$ was selected by grid searching.

In order to enhance the rich diversity of data, the non-matching data but with the same label is formed as the pairwise instance. The entire model is trained on an Nvidia GTX 1080Ti GPU with PyTorch and the ADAM [4] optimization method is employed with a learning rate of 0.0001.

IV. EXPERIMENTS

To verify the proposed method, experiments were conducted on three widely-used benchmark datasets: the Wikipedia dataset, the Pascal Sentence dataset and the NUS-WIDE-10k dataset. It is also evaluated on the PKU XMedia dataset to extend its application on multimodal retrieval. In the experiments, the validity of our SLSAE is verified, from the results, it can be seen that it outperforms the state-of-the-art methods. And we conduct further analysis of the SLSAE to explore the role of different modules.

A. DATASETS AND FEATURES

Here, the four datasets and their features used in the experiments are introduced in brief, first three for bimodal retrieval and the last one for five modalities retrieval.

1) WIKIPEDIA

The Wikipedia dataset [3] is the most widely used dataset for cross-modal retrieval, which is collected from “Wikipedia featured articles”. It contains 2866 image-text pairs which belong to 10 semantic categories. There are 2173 samples in training set and 693 samples in test set. The Wikipedia dataset is available at http://www.svcl.ucsd.edu/projects/crossmodal/.

2) PASCAL SENTENCE

The Pascal sentence dataset [9] is randomly selected from the 2008 PASCAL Development Kit, which contains 1000 images. Each image has 5 sentences as its corresponding text description. It is divided into 20 categories and each category has 50 samples. There are 800 pairs in training set and 200 pairs in test set. The Pascal Sentence dataset is available at http://vision.cs.uiuc.edu/pascal-sentences/.

3) NUS-WIDE-10k

The NUS-WIDE-10k dataset is a subset of NUS-WIDE dataset [30], which has 270k images belonging to 81 categories. The NUS-WIDE-10k has 10k samples with 10 categories. Thera are 8000 pairs in training set and 2000 pairs in test set. The NUS-WIDE-10k dataset is available at https://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/nuswide/NUS-WIDE.html.

4) PKU XMedia

The PKU XMedia dataset [24], [32] is different from other datasets above. It has five modalities data, including image, text, video, audio and 3D models, which is built for multimodal retrieval. There are 5000 images, 5000 texts, 1143 videos, 1000 audio clips and 500 3D models and evenly split into 20 categories. The image and video feature is 4096d vector extracted by the fc7 layer of AlexNet [21], and the text represents is 3000d bag-of-words(Bow) feature. 3D model feature is 4700d vector of the Light Field descriptor. Audio clips are represented by 29d Mel Frequency Cepstral Coefficient (MFCC) feature. The PKU XMedia dataset is available at http://www.icst.pku.edu.cn/mipl/XMedia.

The image data features of the Wikipedia, Pascal sentence and NUS-WIDE-10k dataset is 4096d vector extracted by the fc7 layer of VGGNet [35]. The text instance is represented by BoW vector with the TF-IDF weighting scheme, 5000d for Wikipedia, 1000d for Pascal sentence and NUS-WIDE-10k dataset respectively. It is noted that, the shallow features of Wikipedia dataset is also applied to evaluate our proposed method, 128d SIFT features for image, and 10d Latent Dirichlet Allocation (LDA) features for text.

Table 1 shows the general statistics of the four datasets we used. As seen that, these datasets have very distinct properties including both deep features and shallow features.

B. EVALUATION METRICS

In the experiments, retrieval performance is evaluated by mean Average Precision (mAP). The mAP metric considers both the retrieval precision and the ranking information simultaneously, which is widely used in information retrieval field. There are two cross-modal retrieval tasks in two modalities, Img2Txt and Txt2Img. Img2txt is given an image as the query, retrieve from text modality; and Txt2Img is given a text as the query, retrieve from image modality. The average precision(AP) is defined as:

$$AP = \frac{1}{M} \sum_{k=1}^{R} \frac{M_k}{k} \cdot rel_k.$$

Here, $R$ represents the first $R$-top ranked retrieval data, $M$ is the total number of relevant data in retrieval results, and $M_k$ is the relevant item in the top $k$ returns. $rel_k = 1$ if the $k$-th retrieval data is relevant, otherwise it is 0. Besides this, top-$k$ precision was also used to compare our method with ACMR [5] and CMST [38].
### TABLE 1. General statistics of the four datasets used in the experiments.

| Dataset          | Label | Modality | Training/Test Features          |
|------------------|-------|----------|---------------------------------|
| Wikipedia [3]    | 10    | image    | 2173/893 4096d VGG + 128d SIFT |
|                  |       | text     | 2173/893 5000d BoW + 10d LDA   |
| Pascal sentence [9] | 20   | image    | 800/200 4096d VGG              |
|                  |       | text     | 800/200 1000d BoW              |
| NUS-WIDE-10k [30] | 10   | image    | 8000/2000 4096d VGG            |
|                  |       | text     | 8000/2000 1000d BoW            |
| PKU XMedia [24], [32] | 20 | image    | 4000/10000 4096d AlexNet       |
|                  |       | text     | 4000/10000 3000d BoW           |
|                  |       | video    | 969/176 4096d AlexNet          |
|                  |       | audio clip | 800/200 29d MFCC               |
|                  |       | 3D model | 400/100 4700d LightField       |

### TABLE 2. Comparisons in term of mAP in three datasets.

| Methods | Wikipedia [3] | Pascal Sentence [9] | NUS-WIDE-10k [30] |
|---------|---------------|---------------------|-------------------|
|         | Img2Tx | Txt2Img | Avg. | Img2Tx | Txt2Img | Avg. | Img2Tx | Txt2Img | Avg. |
| CCA [8] | 0.267 | 0.222 | 0.245 | 0.383 | 0.219 | 0.291 | 0.189 | 0.188 | 0.189 |
| LCFS [23] | 0.455 | 0.396 | 0.427 | 0.442 | 0.357 | 0.400 | 0.383 | 0.346 | 0.365 |
| JRL [24] | 0.453 | 0.400 | 0.426 | 0.504 | 0.489 | 0.496 | 0.426 | 0.376 | 0.401 |
| Corr-AE [13] | 0.402 | 0.395 | 0.398 | 0.489 | 0.444 | 0.467 | 0.366 | 0.417 | 0.392 |
| Deep-SM [12] | 0.458 | 0.345 | 0.402 | 0.440 | 0.414 | 0.427 | 0.389 | 0.496 | 0.443 |
| CMDN [11] | 0.488 | 0.427 | 0.458 | 0.534 | 0.534 | 0.534 | 0.492 | 0.515 | 0.504 |
| MHTN [25] | 0.514 | 0.444 | 0.479 | 0.496 | 0.500 | 0.498 | 0.520 | 0.534 | 0.527 |
| ACMR [5] | 0.619 | 0.488 | 0.546 | 0.535 | 0.543 | 0.539 | 0.544 | 0.538 | 0.541 |
| MASLN [26] | 0.623 | 0.499 | 0.561 | 0.544 | 0.547 | 0.545 | 0.550 | 0.552 | 0.551 |
| CMST [38] | 0.632 | 0.505 | 0.569 | 0.621 | 0.586 | 0.604 | 0.628 | 0.562 | 0.595 |
| **Ours** | **0.652** | **0.538** | **0.595** | **0.647** | **0.628** | **0.638** | **0.656** | **0.589** | **0.623** |

### TABLE 3. Comparisons in term of mAP on Wikipedia [3] with shallow feature.

| Methods | Img2Tx | Txt2Img | Avg. |
|---------|--------|---------|------|
| CCA [8] | 0.267  | 0.222  | 0.245|
| LCFS [23] | 0.279  | 0.214  | 0.246|
| Corr-AE [13] | 0.280  | 0.242  | 0.261|
| JRL [24] | 0.344  | 0.277  | 0.311|
| ACMR [5] | 0.366  | 0.277  | 0.322|
| CMST [38] | 0.371  | 0.306  | 0.339|
| **Ours** | **0.392** | **0.328** | **0.360** |

### C. COMPARISON WITH EXISTING METHODS

SLSAE was compared with 10 methods belonging to two categories, traditional statistical methods and DNN-based methods. Traditional cross-modal retrieval methods include CCA [8], LCFS [23] and JRL [24], and DNN-based methods include Corr-AE [13], CMDN [11], MHTN [25], Deep-SM [12], ACMR [5], MASLN [26] and CMST [38].

Table 2 shows the retrieval results on the Wikipedia [3], Pascal Sentence [9] and NUS-WIDE-10k [30] datasets. The best results on each dataset are highlighted in bold font. For a fair comparison, the mAP is calculated on top-50 retrieved results. Our proposed method achieves the best result in all three datasets, both in Img2Txt and Txt2Img. SLSAE significantly outperforms both the traditional and the DNN-based cross-modal retrieval methods. On the Wikipedia dataset [3], SLSAE improves the best method CMST [38] by 3.16% and 6.53% in Img2Txt andTxt2Img retrieval task, and 4.57% on average. On the Pascal Sentence dataset [9], SLSAE improves the best method CMST [38] by 4.19% and 7.17% in Img2Txt and Txt2Img retrieval task, and 5.63% on average. On the NUS-WIDE-10k dataset [30], SLSAE improves the best method CMST [38] by 4.46% and 4.80% in Img2Txt and Txt2Img retrieval task, and 4.71% on average.

In order to further verify the performance of our method, we conduct experiment on Wikipedia dataset [3] with the shallow features, which image is 128d SIFT feature and text is 10d LDA feature. Table 3 shows the results of shallow feature on Wikipedia dataset [3]. SLSAE significantly outperforms the compared methods. Compared to the CMST [38], our method improved 5.66%, 7.19%, and 6.19% in Img2Txt, Txt2Img and on average, although the dimension of shallow feature is small, harder to retrieve.

In addition, we also compare our method with ACMR [5] and CMST [38] methods on top-k precision on Wikipedia dataset. Table 4 shows the experiment results of top-k precision, where we use five k to examine the method performance. For k = 1, 5, our method outperforms CMST [38] by 8.85% and 7.43% on average top-k precision. The results show that...
SLSAE method can effectively improve the rank of retrieved data, which is vitally important for retrieval application.

In the subsection III-E, We extend the model to multimodal in a concise way. The experiments were deployed on PKU XMedia dataset [24], [32]. There are five modalities in the PKU XMedia dataset, image, text, video, audio clip and 3D model, which is created for multimodal retrieval. Our method was compared with two categories methods, two-modality retrieval methods including JRL [24], DCCA [14], CMPM+CMPC [16] and ACMR [5]; and multi-modal retrieval methods include MCCA [29], GMLDA [19] and SDML [31].

Table 5 shows the results of multimodal retrieval performance. It can be seen that our method is competitive compared to other multimodal retrieval methods, and achieves the best result on average. For multimodal, the model of SLSAE only needs to be trained once to perform mutual retrieval of any modalities. The experimental results proved that our method has excellent universality and generality for modalities. Unlike other methods, the model of SLSAE is quite concise where the modality networks are independent of each other without complicated coupling. Some methods have very complicated models and are limited to image and text modality, which cannot be transferred to other modalities. The experimental results demonstrated that the framework of SLSAE is more reasonable and superior.

D. FURTHER ANALYSIS ON SLSAE

In this subsection, we conduct several experiments to investigate the effective component of SLSAE. As we know, the semantic similarity-preserving learning process and the semantic average embedding play an important role in the method. We define the following five variant methods to explore the effectiveness of our method.

SLSAE-1 only deal with label loss without the pair loss, inner-modality and inter-modality loss.

SLSAE-2 performs retrieval task without semantic average embedding.

SLSAE-3 only has one batch semantic average embedding instead of modality-specific average embedding. In this way, only one set semantic average embedding is utilized no matter the modality number, and it is obtained in all modalities. While in original SLSAE, the semantic average embeddings are acquired in each modality separately.

SLSAE-4 obtains the query category by the embedding distance between the query and the semantic average embeddings in query modality. Which average embedding the query embedding is closest to, which category the query belongs to. In original SLSAE, we use the classifier prediction to get the category of query.

SLSAE-5 directly uses the groundtruth category of the query to retrieve. This method cannot be used in retrieval task. It is designed just for compare and explore the potential of our method.

The results of the five variant methods are shown in Table 6. As we can see, the performance of SLSAE-1 and SLSAE-2 has degraded 5.4% and 6.6% compared to the original SLSAE. SLSAE-1 only uses the label loss, without pair loss, the inner-modality and inter-modality loss. It shows that the distance constraint is reasonable and important in the design of joint loss. The main reason is that the joint

| Methods | Img2Txt | Txt2Img | Avg. |
|---------|---------|---------|------|
| SLSAE-1 | 0.622   | 0.503   | 0.563|
| SLSAE-2 | 0.614   | 0.498   | 0.556|
| SLSAE-3 | 0.633   | 0.516   | 0.575|
| SLSAE-4 | 0.650   | 0.534   | 0.592|
| SLSAE(original) | 0.652 | 0.538 | 0.595 |
| SLSAE-5 | 0.985   | 0.787   | 0.886|

| Methods | Img2Txt | Txt2Img | Avg. |
|---------|---------|---------|------|
| -       |         |         |      |

SLSAE-2 performs retrieval task without semantic average embedding.

SLSAE-3 only has one batch semantic average embedding instead of modality-specific average embedding. In this way, only one set semantic average embedding is utilized no matter the modality number, and it is obtained in all modalities. While in original SLSAE, the semantic average embeddings are acquired in each modality separately.

SLSAE-4 obtains the query category by the embedding distance between the query and the semantic average embeddings in query modality. Which average embedding the query embedding is closest to, which category the query belongs to. In original SLSAE, we use the classifier prediction to get the category of query.

SLSAE-5 directly uses the groundtruth category of the query to retrieve. This method cannot be used in retrieval task. It is designed just for compare and explore the potential of our method.

The results of the five variant methods are shown in Table 6. As we can see, the performance of SLSAE-1 and SLSAE-2 has degraded 5.4% and 6.6% compared to the original SLSAE. SLSAE-1 only uses the label loss, without pair loss, the inner-modality and inter-modality loss. It shows that the distance constraint is reasonable and important in the design of joint loss. The main reason is that the joint
loss can preserve semantic similarity in common space, which is beneficial to cross-modal retrieval. In SLSAE-2, the embedding of query is used to retrieve directly. It indicates that the utilization of semantic average embedding can significantly improve the retrieval performance. As the semantic average embedding contains the information of entire category, it can weaken the problem of data diversity.

In SLSAE-3, we mixed the average embedding belonging to different modalities. In original method, each modality has one set semantic average embedding. While in SLSAE-3, there is only one set semantic average embedding, which is obtained in all modalities. The performance of SLSAE-3 is worse than original SLSAE. It confirmed that the embeddings of different modalities in the common space are diverse, even though they have the same semantic label, as shown in Figure 3. Whereas the diversity of different modalities, it is better to deal with modalities separately in the common space. Our method looks squarely at the existence of the modality gap and weakens its influence.

As for SLSAE-4, its performance is similar to original SLSAE. The output of label classifier is the probability of the embedding belonging to each category. In original method, when given a query, its corresponding semantic average embedding is utilized instead of its own embedding to do the retrieval task. But the true category of the query is unknown, so the label classifier can be used to approximate its true semantic label. In this variant method, we obtain the query category by the distance among query embedding and the average embeddings in query modality. Which semantic average embedding the query is closest to, the corresponding category is the accuracy of query category predicted by the distance between query and semantic average embeddings.

As the semantic average embedding contains the information of entire category, it can weaken the problem of data diversity. Besides this, we also analyze the effectiveness of parameter $\alpha$ and $\beta$ in the joint loss, which control the contribution of inner-modality and inter-modality loss. In the experiments, we set $\alpha$ and $\beta$ range as \{0, 0.5, 1, 2, 4, 8\}. From Figure 7,

![Figure 4](image1.png)  ![Figure 5](image2.png)
FIGURE 6. t-SNE visualization of initial and convergence state on the Wikipedia dataset [3].

FIGURE 7. Retrieval performance of SLSAE with different weight coefficient of inner-modality and inter-modality loss on the Wikipedia dataset [3].

FIGURE 8. The retrieval examples on Pascal Sentence dataset [9].

V. CONCLUSION
In this paper, a new approach (SLSAE) for cross-modal retrieval is proposed, which can effectively learn a similarity-preserving common space with the joint loss. The embedding network transforms different modalities data into the same common space to retrieve. It can be observed that the distribution of training set and test set are different, and the embeddings belonging to the same category are diverse in common space, which often appear in real-world retrieval applications. Based on this, the semantic average embedding is used to improve retrieval performance and robustness. Our proposed method has superior retrieval performance compared to other state-of-the-art methods in three widely used benchmarks. Exhaustive experiments have been deployed to evaluate the effectiveness of our method. Experimental

it can be observed that SLSAE performs well while $\alpha$ and $\beta$ lie in the interval [0.5, 1].

The Figure 8 shows the top-5 retrieved results of image query text and text query image on the Pascal Sentence dataset. When the bicycle image is used as the query, the first four search results are correct and related to bicycles, and the last one is wrong which is related to motorcycles. When the cat text is used as th query, the first, second, and fourth retrieved results are correct; the third and fifth results are wrong. It should be noted that in Txt2Img the third and fifth results both contain cat, but their labels are bottle and dog respectively.
results demonstrate that modality gap still exists, which cannot be eliminated completely and the semantic average can reduce its impact and improve retrieval performance. It can be seen that our model is concise, but efficient, only containing modality-specific embedding network and a common label classifier merely. It is easy to extend to multimodal flexibly, and also performs well compared to other excellent multimodal retrieval methods. The experimental results demonstrate that the proposed method performs well in the aspects of modal generality and universality.

In the future, we intend to improve the work in the following two aspects. First, we will attempt to reduce the semantic gap in the common space with metric learning. Second, we will try to remove the label classifier in our method, exploiting the average embedding to improve performance in an unsupervised way.

REFERENCES

[1] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, “A comprehensive survey on cross-modal retrieval,” 2016, arXiv:1607.06215. [Online]. Available: http://arxiv.org/abs/1607.06215

[2] S. Akaho, “A kernel method for canonical correlation analysis,” 2006, arXiv:cs/0609071. [Online]. Available: https://arxiv.org/cs/0609071

[3] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos, “A new approach to cross-modal multimedia retrieval,” in Proc. 18th Int. Conf. Multimedia (MM), 2010, pp. 251–260.

[4] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

[5] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, “Adversarial cross-modal retrieval,” in Proc. 25th ACM Multimedia Conf. (MM), 2017, pp. 154–162.

[6] Y. Peng, X. Huang, and Y. Zhao, “An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges,” IEEE Trans. Circuits Syst. Video Technol., vol. 28, no. 9, pp. 2372–2385, Sep. 2018.

[7] X. Zhang, H. Lai, and J. Feng, “Attention-aware deep adversarial hashing for cross-modal retrieval,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 591–606.

[8] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, “Canonical correlation analysis: An overview with applications to learning methods,” Neural Comput., vol. 16, no. 12, pp. 2639–2664, Dec. 2004.

[9] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, “Collecting semantic annotations using amazon’s mechanical turk,” Proc. NAACL HLT Workshop Creating Speech Lang. Data Amazon’s Mech. Turk. Stroudsburg, PA, USA: Association Computational Linguistics, 2010, pp. 139–147.

[10] Z. Ye and Y. Peng, “Sequential cross-modal hashing learning via multi-scale correlation mining,” ACM Trans. Multimedia Comput., Commun., Appl., vol. 15, no. 4, pp. 1–20, Jan. 2020.

[11] Z. Ye and Y. Peng, “Unsupervised generative adversarial cross-modal hashing,” in Proc. 32nd AAAI Conf. Artif. Intell., 2018, pp. 539–546.

[12] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, “Adversarial cross-modal retrieval,” in Proc. 25th ACM Multimedia Conf. (MM), 2017, pp. 154–162.

[13] Y. Peng, X. Huang, and Y. Zhao, “An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges,” IEEE Trans. Circuits Syst. Video Technol., vol. 28, no. 9, pp. 2372–2385, Sep. 2018.

[14] X. Zhang, H. Lai, and J. Feng, “Attention-aware deep adversarial hashing for cross-modal retrieval,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 591–606.

[15] S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. Int. Conf. Machine Learn., 2014, pp. 2160–2167.
[43] X. Xu, T. Wang, Y. Yang, L. Zuo, F. Shen, and H. T. Shen, “Cross-modal attention with semantic consistence for image–text matching,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5412–5425, Dec. 2020.

[44] S. Wang, Y. Chen, J. Zhuo, Q. Huang, and Q. Tian, “Joint global and co-attentive representation learning for image-sentence retrieval,” in *Proc. 26th ACM Multimedia Conf. Multimedia Conf. (MM)*, 2018, pp. 1398–1406.

[45] J. Song, T. He, L. Gao, X. Xu, A. Hanjalic, and H. T. Shen, “Unified binary generative adversarial network for image retrieval and compression,” *Int. J. Comput. Vis.*, vol. 128, pp. 2243–2264, Feb. 2020.

[46] L. Gao, X. Li, J. Song, and H. T. Shen, “Hierarchical LSTMs with adaptive attention for visual captioning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1112–1131, May 2020.

**TAO ZHI** received the bachelor’s degree from Xidian University, Xi’an, China, in 2017. He is currently a Graduate Student with the School of Artificial Intelligence, Xidian University. His research interests include computer vision, face recognition, and cross-modal retrieval.

**HONG HAN** (Member, IEEE) was born in 1974. She received the Ph.D. degree in electrical engineering from Xi’an Jiaotong University, Xi’an, China, in 2003. She is currently a Senior Researcher with the School of Artificial Intelligence, Xidian University, Xi’an. Her research interests include computer vision, information fusion, and machine learning.

**YINGCHUN FAN** received the B.S. degree in intelligence science and technology from the School of Electronic Engineering, Xidian University, Xi’an, China, in 2014, and the M.S. degree in optical engineering from the School of Physics and Information Technology, Shaanxi Normal University, Xi’an, in 2017. He is currently pursuing the Ph.D. degree in pattern recognition and intelligent system with the School of Artificial Intelligence, Xidian University.

His research interests include computer vision, image processing, and simultaneously localization and mapping, with a focus on the SLAM in dynamic environment.

© 2020 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.