Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 2503137, 17 pages
https://doi.org/10.1155/2020/2503137

Research Article

Abstraction and Association: Cross-Modal Retrieval Based on Consistency between Semantic Structures

Qibin Zheng,[1] **Xiaoguang Ren**,[2,3] **Yi Liu**,[2,3] **and Wei Qin**[2,3]

[1]*Army Engineering University of PLA, Nanjing, China*
[2]*National Innovation Institute of Defense Technology (NIIDT), Beijing, China*
[3]*Tianjin Artificial Intelligence Innovation Center (TAIIC), Tianjin, China*

Correspondence should be addressed to Yi Liu; albertliu20th@163.com

Received 17 December 2019; Revised 16 March 2020; Accepted 15 April 2020; Published 7 May 2020

Academic Editor: Francesco Lolli

Cross-modal retrieval aims to find relevant data of different modalities, such as images and text. In order to bridge the modality gap, most existing methods require a lot of coupled sample pairs as training data. To reduce the demands for training data, we propose a cross-modal retrieval framework that utilizes both coupled and uncoupled samples. The framework consists of two parts: *Abstraction* that aims to provide high-level single-modal representations with uncoupled samples; then, *Association* links different modalities through a few coupled training samples. Moreover, under this framework, we implement a cross-modal retrieval method based on the consistency between the semantic structure of multiple modalities. First, both images and text are represented with the semantic structure-based representation, which represents each sample as its similarity from the reference points that are generated from single-modal clustering. Then, the reference points of different modalities are aligned through an active learning strategy. Finally, the cross-modal similarity can be measured with the consistency between the semantic structures. The experiment results demonstrate that given proper abstraction of single-modal data, the relationship between different modalities can be simplified, and even limited coupled cross-modal training data are sufficient for satisfactory retrieval accuracy.

1. Introduction

Recent years have witnessed a surge of need in jointly analyzing multimodal data [1, 2]. As one of the fundamental problems of many multimodal applications, cross-modal retrieval aims to find semantically similar items from objects of different modalities (such as text, visual, or audio object) [3].

The modality gap is the main challenge of cross-modal retrieval [4, 5]. A common approach to bridge the modality gap is constructing a shared representation space where the multimodal samples can be represented uniformly [2]. However, it is not easy because it requires detailed knowledge of the content of each modality and the correspondence between them [6]. A variety of tools are used to construct the shared space, such as canonical correlation analysis (CCA) [1, 7–10], topic model [11–13], and hashing [14–18]. Among these methods, the deep neural network (DNN) has become the most popular one because of its strong learning ability [6, 19–24]. The performance of most of these methods, especially DNN-based methods, heavily depends on sufficient coupled cross-modal samples [25]. However, collecting coupled training data is labor-intensive and time-consuming.

Although it may not be explicitly announced, two types of relationships are essential considerations when constructing the shared representation space: the intermodal relation and the intramodal relation [5, 10]. They play critical roles in preserving the cross-modal similarity and the single-modal similarity, respectively [5, 10, 25]. Also, separate representation learning and shared representation learning in some existing works are preserving these two relationships [26, 27].

It should be noticed that the information to maintain these two relationships is different: the correspondence between cross-modal samples is essential to preserve

intermodal relations, while the similarity relation between single-modal samples is indispensable to preserve intramodal relations [10]. Most of the existing methods, such as [5, 25, 28–30], only use coupled cross-modal samples to preserve intermodal relations and intramodal relations; however, uncoupled single-modal samples are discarded.

In many cross-domain learning tasks, such as machine translation [31–33], unlabeled samples in the single domain are significant. As a typical cross-domain learning task, cross-modal retrieval should also benefit from uncoupled training samples. Besides, in contrast to coupled training samples, uncoupled ones are easier to obtain. Thus, it is necessary to introduce uncoupled training samples into the construction of the shared representation space, especially when coupled ones are insufficient.

Inspired by the discussion above, a two-stage cross-modal retrieval framework is proposed. As illustrated in Figure 1, the proposed *Abs-Ass* framework uses training samples in a different way. In existing methods, only coupled training samples are used to preserve intramodal and intermodal relations. However, in this framework, both coupled and uncoupled samples are used to maintain intramodal relations; only a few coupled cross-modal sample pairs are used to maintain intermodal relations. Thus, the process of constructing the shared representation space is divided into two subprocesses: *Abstraction* that preserves intramodal relations and *Association* that preserves intermodal relations.

The name *Abstraction* indicates that we need to consider the intramodal relation at the semantic level rather than the feature level. The name *Association* means that the process of preserving the intermodal relation is exactly finding the correlation between different modalities. *Abstraction* fully explores intramodal relations through uncoupled samples of each modality, which enables *Association* to recognize multimodal samples at a higher level; thus, *Association* can find the correlation between cross-modal samples much easier, even though only a few coupled training samples are available. In the ideal case, high-level representations of different modalities can be associated even with a linear transformation [34].

Moreover, following the framework above, we proposed a cross-modal retrieval method based on the reference-based representation and the correlation between the semantic structures of different modalities. Specifically, *Abstraction* is implemented by the reference-based representation [35], which represents multimodal objects through the semantic structure. The term semantic structure refers to all pairwise similarities of a set of n samples, for some similarity measure [36]. This representation scheme is modality-independent and can provide multimodal objects a relatively isomorphic representation space. Moreover, we prove that if the reference points of different modalities are one-to-one matched, the semantic structures of different modalities are naturally correlated. Thus, cross-modal similarity can be measured with the linear correlation between semantic structures of different modalities. In our implementation, the cross-modal relations have a fixed and straightforward form, and cross-modal sample pairs only play the role of the multimodal reference set; therefore, its performance has much lower dependence on coupled training samples.

Through this paper, we demonstrate the importance of uncoupled samples for preserving intramodal relations and the correlation between semantic structures of different modalities, which together provide the possibility of cross-modal retrieval with limited coupled training samples. The main contribution can be summarized as follows: (1) *Abs-Ass* cross-modal retrieval framework. We propose a two-stage framework consisting of the *Abstraction* and the *Association* that emphasizes different roles of coupled and uncoupled training samples. In contrast to the end-to-end learning model, the proposed framework separates the process of preserving intermodal and intramodal relations into two stages and uses uncoupled single-modal samples and coupled cross-modal samples to learn them, respectively. Compared with the existing methods, the *Abs-Ass* framework improved the using efficiency of training samples and has lower demands for coupled training data. (2) Semantic structure-based cross-modal retrieval method. Following the *Abs-Ass* framework, we propose a cross-modal retrieval method by introducing the reference-based representation to represent multimodal data at the semantic level and proving the positive correlation between the semantic structures of different modalities. Although some existing works also try to find the cross-modal correlation from the semantic view [1, 3], the correlation between semantic structures naturally exists and has a fixed and straightforward pattern. Therefore, even a few coupled training samples are enough to align semantic structures of different modalities. Besides, the proposed method is unsupervised because the reference-based representation scheme does not need class labels.

The remainder of this paper is organized as follows. Section 2 introduces the related works of the cross-modal retrieval task. Section 3 introduces the proposed implementation of the *Abs-Ass* framework. Section 4 tests the proposed method through the experiments on public data sets.

2. Related Work

2.1. CCA-Based Methods. To the best of our knowledge, the first well-known cross-modal correlating model may be the CCA-based model proposed by Hardoon et al. [7]. It learns a linear projection to maximize the correlation between the representation of different modalities in the projected space. Inspired by this work, many CCA-based models are designed for cross-modal analyzing [1, 8–10, 37]. Rasiwasia et al. [1] utilized CCA to learn two maximally correlated subspaces, and multiclass logistic regression was performed within them to produce the semantic spaces, respectively. Mroueh et al. [9] proposed a truncated-SVD based algorithm to compute the full regularization path of CCA for multimodal retrieval efficiently. Wang et al. [10] developed a new hypergraph-based canonical correlation analysis (HCCA) to project low-level features into a shared space where intrapair and interpair correlation is maintained simultaneously. Liang et al. [37] incorporated the group correspondence and CCA to cross-modal retrieval.

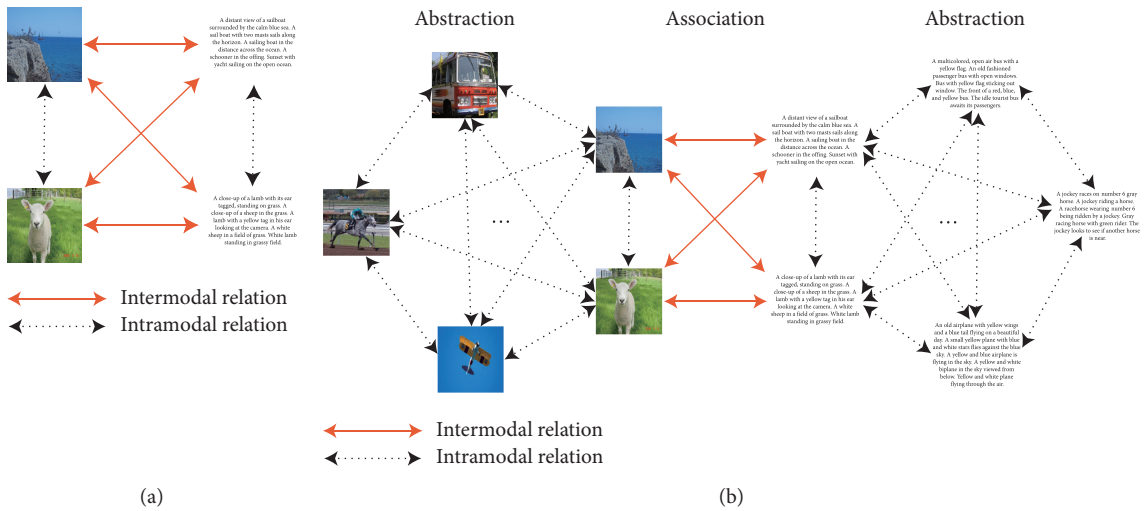

Figure 1: Contrast between the *Abs-Ass* framework and existing methods. (a) Existing methods. (b) *Abs-Ass* framework.

2.2. Topic Model Methods. The topic model is also helpful for uniform representing of multimodal data, assuming that objects of different modalities share some latent topics. Latent Dirichlet allocation- (LDA-) based methods establish the shared space through the joint distribution of multimodal data and the conditional relation between them [11, 12]. Roller and Walde [12] integrated visual features into LDA and presented a multimodal LDA model to learn joint representations for text and visual data. Wang et al. [13] proposed the multimodal mutual topic reinforce model (M^3R) to discover mutual consistent topics.

2.3. Hashing-Based Methods. For the rapid growth of data volume, the cost of finding the nearest neighbors cannot be dismissed. Hashing is a salable method for finding nearest neighbors approximately [14]. It projects data into a Hamming space, where the neighbor search can be performed efficiently. In order to improve the efficiency of finding similar multimodal objects, many cross-modal hashing methods have been proposed [14–18, 38, 39]. Kumar and Udupa [15] proposed a cross-view hashing method to generate such hash codes that minimized the distance in a Hamming space between similar objects and maximized that between dissimilar ones. Yi et al. [16] used a coregularization framework to generate such binary code that the hash codes from different modalities were consistent. Ou et al. [17] constructed a Hamming space for each modality and built the mapping between them with logistic regression. Wu et al. [18] proposed a sparse multimodal hashing method for cross-modal retrieval. Song et al. [38] proposed Self-Supervised Video Hashing (SSVH), which outperforms the state-of-the-art methods on unsupervised video retrieval. Ye and Peng [39] proposed Multiscale Correlation Sequential Cross-modal Hashing Learning (MCSCH) to utilize multiscale features of cross-modal data. Liu et al. [40] proposed the Matrix Tri-Factorization Hashing (MTFH) that discards the unified Hamming space to obtain higher representation scalability.

2.4. Deep Learning Methods. Due to the strong learning ability of the deep neural network, many deep models have been proposed for cross-modal learning, such as [6, 19–24, 26, 27, 41, 42]. Ngiam et al. [19] presented an autoencoder model to learn joint representations for speech audios and videos of the lip movements. Srivastava and Salakhutdinov [20] employed the restricted Boltzmann machine to learn a shared space between data of different modalities. Frome et al. [22] proposed a deep visual-semantic embedding (DeViSE) model to identify the visual objects using the information from the labeled image and unannotated text. Andrew et al. [21] introduced deep canonical correlation analysis to learn such nonlinear mapping between two views of data that the corresponding objects are linearly related in the representation space. Jiang et al. [23] proposed a real-time Internet cross-media retrieval method, in which deep learning was employed for feature extraction and distance detection. Due to the powerful representing ability of the convolutional neural network visual feature, Wei et al. [24] employed it coupled with a deep semantic matching method for cross-modal retrieval. Peng et al. [26, 27] proposed two-stage frameworks to learn the separate representation and the shared representation, which are implemented by cross-media multiple deep networks (CMDN) and cross-modal correlation learning (CCL), respectively. Song et al. [43] proposed multimodal stochastic RNNs (MS-RNN) for the video caption task, which solved a critical deficiency of the existing methods based on the encoder-decoder framework. Recently, the attention mechanism is playing an important role in maintaining the intermodal and intramodal relations. Qi et al. [41] proposed a visual-language relation attention model to explore the intermodal and intramodal relation between fine-grained patches, as well as the cross-media multilevel alignment to boost precise cross-media correlation learning. Gao et al. [42] proposed hierarchical LSTMs with adaptive attention for visual captioning.

Although these methods have achieved great success in multimodal learning, most of them need a mass of training data to learn the complex correlation between objects from

different modalities. To reduce the demand for training data, some methods have been proposed from different views. Gao et al. [25] proposed an active similarity learning model for cross-modal data. Nevertheless, without extra information, improvement is limited. Chowdhury et al. [44] introduced additional web information to cross-modal retrieval.

3. Proposed Approach

The cross-modal retrieval can be formalized as follows. The multimodal data set $D(X, Y)$ consists of $X = \{x_1, \ldots, x_n\} \in \mathbb{R}^{n \times d}$ and $Y = \{y_1, \ldots, y_m\} \in \mathbb{R}^{m \times e}$. Given a query set Q of any modality, the goal of cross-modal retrieval is to calculate similarity between each query and a set T of all the targets of the other modalities and retrieve the similar samples by ranking all the target samples according to the similarity [30]. We assume the availability of a small training set $\mathrm{Tr} = \{(x_{\mathrm{tr}}, y_{\mathrm{tr}}) \mid (x_{\mathrm{tr}} \approx y_{\mathrm{tr}})\}$, where $x_{\mathrm{tr}} \approx y_{\mathrm{tr}}$ means x_{tr} and y_{tr} are similar. Because this work focuses on unsupervised and few-coupled cross-modal retrieval, class labels of both modalities' samples are not available, and the size of the training set is much smaller than the whole data set.

The process of our proposed method can be described as equations (1) ~ (5). First, extract visual and text features through tools in Section 3.1:

$$\mathscr{M}_1 : X \longrightarrow \mathscr{X}, \tag{1}$$

$$\mathscr{M}_2 : Y \longrightarrow \mathscr{Y}. \tag{2}$$

Second, represent the to-be-matched objects (the nonred points in Figure 2) of $\mathscr{X}$ and $\mathscr{Y}$ by the distributed representation in Section 3.2:

$$\mathscr{M}_3 : \mathscr{X} \longrightarrow \mathfrak{R}^{\mathscr{X}}, \tag{3}$$

$$\mathscr{M}_4 : \mathscr{Y} \longrightarrow \mathfrak{R}^{\mathscr{Y}}. \tag{4}$$

As illustrated in Section 3.3, $\mathfrak{R}^{\mathscr{X}}$ and $\mathfrak{R}^{\mathscr{Y}}$ have been well abstracted and are highly isomorphic in semantic and form. Therefore, in Section 3.4, the representation space of different modalities can be easily aligned by the coupled training samples:

$$\mathscr{M}_5 : \mathfrak{R}^{\mathscr{X}} \longrightarrow \mathfrak{R} \leftarrow \mathfrak{R}^{\mathscr{Y}}, \tag{5}$$

and the similarity between cross-modal samples can be measured with general similarity metrics.

3.1. Feature Extraction for Text and Images. In the early research on cross-modal learning, the weak-effectiveness of low-level feature extraction is one of the main factors that limits the retrieval accuracy. The application of the CNN visual feature has significantly improved the accuracy of cross-modal retrieval [4, 24]. In contrast to the visual feature, some works still take the BoW (bag-of-word) as the default tool to extract text features [5, 29], which is not effective enough to model intramodal relations in text modality. The consistency of the semantic structure is beneficial to transferring learning tasks, including cross-modal retrieval [45]; thus, we take the pretrained CNN model and the sentence embedding with the pretrained word vector for feature extraction of images and text, respectively.

3.1.1. Pretrained Convolutional Neural Network for Feature Extraction of Images. CNN has demonstrated outstanding performance for various computer vision tasks, such as image classification and object detection. Wei et al. proposed to utilize the pretrained CNN for visual feature extraction in cross-modal retrieval [24], which performs much better than the low-level feature. Because we aim to reduce the dependency on training data, we directly take the pretrained VGG19 [46] (not fine-tuned) to extract the feature of images, namely, the mapping in equation (1).

3.1.2. Sentence Embedding for Feature Extraction of Text. The advancement of NLP techniques provides us powerful tools for text feature extraction. Given enough supervised information, a good end-to-end model can automatically extract the most important features; however, with limited training data, it is hard to train such a model. Instead, we take the pretrained word embedding and an unsupervised text embedding method for the feature extraction of the text, which is the mapping in equation (2).

Many text embedding methods for general NLP tasks can be helpful; among them, smooth inverse frequency (SIF) is a simple but a powerful sentence embedding method [47]. With the pretrained word vector (such as Glove [48]), SIF provides a completely unsupervised method to embed sentences into the semantic space, which can be summarized as equations (6) and (7). Given a sentence s, each word in s is represented as its word vector v_w; then, the sentence s is represented as the weighted average of all word vectors:

$$v_s' = \frac{1}{|s|} \sum_{w \in s} \frac{a}{a + p(w)} v_w, \tag{6}$$

where $p(w)$ is the probability of a word w which is emitted in the sentence s. The computation of parameter a is complex, which can be found in [47]. Let X be a matrix whose columns are $[v_s : s \in S]$ and u be the first singular vector of X that can be computed by singular vector decomposition (SVD); the final sentence embedding vector is obtained by

$$v_s = v_s' - uu^T v_s'. \tag{7}$$

3.2. Semantic Structure-Based Representation for Single-Modality Data. Although the extracting tools above provide more accurate features for image and text data, it is still hard to directly perform retrieval tasks on them, especially when only limited coupled training samples are available. Therefore, feature-level representations $\mathscr{X}$ and $\mathscr{Y}$ need further abstraction, i.e., representation learning in (3) and (4). Besides, we mainly consider unsupervised representation learning because training samples are not always labeled in real-world applications. In this way, we introduce an

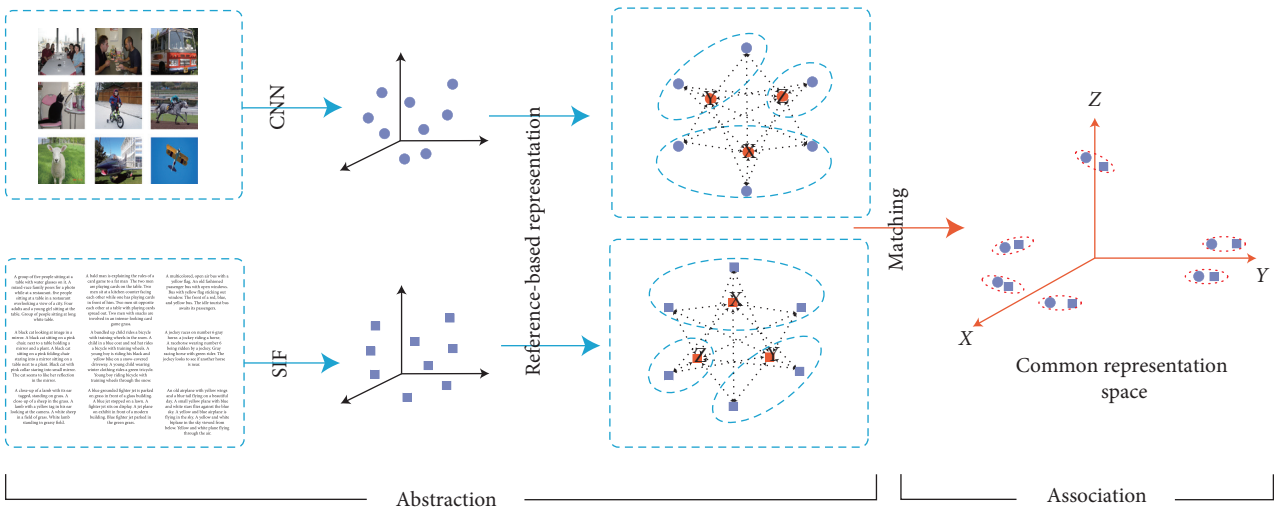

Figure 2: The semantic structure-based implementation of the *Abs-Ass* framework.

unsupervised representation scheme to represent image and text data that can preserve intramodal relations.

In the unsupervised setting, the semantic structure-based representation (also named space structure-based representation, SSR) is a simple but effective way to preserve intramodal relations, as some unsupervised learning methods did [49, 50]. Given a set of samples $\mathcal{X}$, in the semantic structure-based representation, each sample $x_i \in \mathcal{X}$ is represented as the similarity vector:

$$x_i^r = \left[x_{i1}^r, x_{i2}^r, \ldots, x_{ij}^r, \ldots, x_{in}^r\right], \tag{8}$$

where x_{ij}^r is the similarity between x_i and x_j. In this paper, it is computed with the cosine similarity:

$$x_{ij}^r = \frac{x_i x_j}{\left|x_i\right| \cdot \left|x_j\right|}. \tag{9}$$

The reason for choosing cosine similarity lies in two aspects: on the one hand, the cosine similarity is normalized, and its value range is always $[-1, 1]$, which helps us to measure the consistency between semantic structures easier in Section 3.3; on the other hand, both image and text are high-dimensional data, where cosine similarity performs good on both accuracy and efficiency.

In equation (8), the dimensionality is very high for large data sets, which leads to high computational complexity of the representation and the follow-up task. Given a data set $X \in \mathbb{R}^{n \times d}$, the representing complexity is $O(n^2)$ [35]. Besides, it is not true that all the samples are useful for the representation, and some of them may undermine the representing ability. Zheng et al. [35] believed that it is better to take some representative samples as the reference points rather than all of them and propose a lower-dimensional SSR—the reference-based representation. As illustrated in Figure 3, six purple points are represented as their similarities (the dotted line) to three reference points (the orange ones). With this representation scheme, an object x_i is represented as the distribution over some reference points:

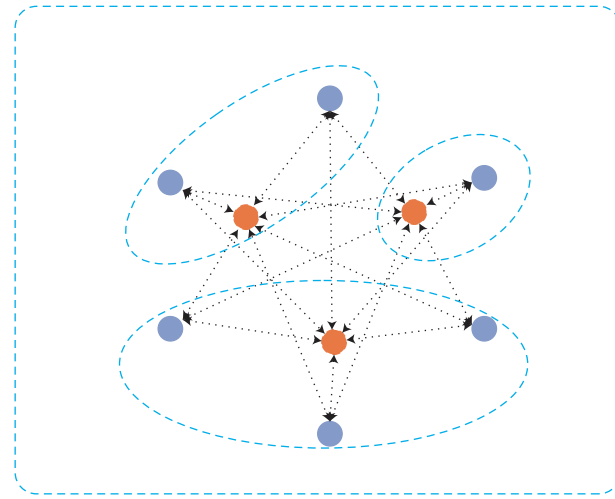

Figure 3: Reference-based representation.

$$x_i^r = \left[x_{i1'}^r, x_{i2'}^r, \ldots, x_{ij'}^r, \ldots, x_{ik'}^r\right], \tag{10}$$

where $x_{ij'}^r$ is the cosine similarity between x_i and the reference point $x_{j'} \in \mathcal{X}'$.

The reference set $\mathcal{X}'$ is a subset of $\mathcal{X}$, which is selected by a clustering-based strategy in [35]. As Figure 4 shows, $\mathcal{X}$ is divided into groups through clustering, and the center of each cluster is selected as a reference point.

The clustering method should generate cluster centers of the sample form because the reference point is the real samples of the data set. However, many popular clustering methods can only generate cluster centers in the form of prototypes, such as k-means. Therefore, we choose a simple and effective clustering method that can generate centers of the sample form, which is the k-medoids [51] method. Also, the clustering number is a significant consideration, which is directly related to the representation ability and the cost. Zheng et al. [35] provided two ways of deciding the cluster number: one by the canopy method [52], which can

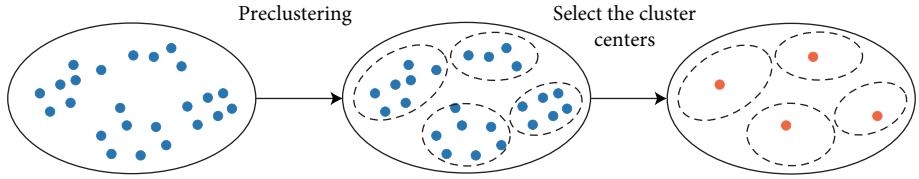

Figure 4: The process of selecting the reference points.

automatically give the number of clusters; the other is user-specifying, where users can balance the performance and the cost by themselves.

In the reference-based representation, $\mathfrak{R}^{\mathcal{X}}$ and $\mathfrak{R}^{\mathcal{Y}}$, image and text data are represented as their distribution over semantic prototypes; in this way, the correlation between them can be found more straightforward. In the next section, it is proved that semantic structures of different modalities are correlated, which can be used to measure the cross-modal similarity even with very limited coupled training samples.

3.3. Cross-Modal Similarity Computing Based on the Correlation between Semantic Structures. The semantic structure-based representation provides different modalities with a homogeneous representation scheme. Moreover, if the reference set $\mathcal{X}'$ and $\mathcal{Y}'$ is one-to-one matched, corresponding dimensions of $\mathfrak{R}^{\mathcal{X}}$ and $\mathfrak{R}^{\mathcal{Y}}$ have the same meanings—the similarity to a semantic prototype. In this way, the similarity between cross-modal samples can be computed according to the correlation between the corresponding dimensions of their reference-based representations.

This section proves that assuming $\mathfrak{R}^{\mathcal{X}}$ and $\mathfrak{R}^{\mathcal{Y}}$ share the reference points (i.e., the reference points of different modalities are one-to-one matched), values in the corresponding dimension of similar samples are positively correlated. Since a reference point in the reference-based representation is also a real sample, the assumption above holds if the semantic structures of different modalities are positively correlated. That is to say, if two images $x_i, x_j \in X$ are similar to each other, their corresponding text descriptions $y_i, y_j \in Y$ should be similar too and vice versa.

Although the assumption seems reasonable intuitively, it is hard to prove completely because the definition of the cross-modal similarity relation cannot be defined uniformly at the feature level. For simplicity, we discuss the case that similar cross-modal samples can be correlated through a linear transformation. The nonlinear case is not discussed because nonlinear mapping functions have much more complex and various forms; thus, it is difficult to discuss the nonlinear case comprehensively in a limited space. Besides, existing works [34, 36] have proved that nonlinear mapping functions have no obvious advantage over linear mapping in correlating cross-modal samples.

Following existing works [1, 25], we assume that similar cross-modal samples are correlated through a linear transformation:

$$y_i \approx x_i \longrightarrow y_i = x_i M, \tag{11}$$

where $M \in \mathbb{R}^{d \times e}$ is a mapping matrix. M is nonzero (not all elements are zero) because if M is zero, then y_i will always be zero, which is obviously unreasonable. The similarity between x_i and x_j and that between y_i and y_j, denoted as $s_{\mathcal{X}}(i,j)$ and $s_{\mathcal{Y}}(i,j)$, can be measured with their inner products:

$$\begin{aligned} s_{\mathcal{X}}(i,j) &= x_i x_j^T, \\ s_{\mathcal{Y}}(i,j) &= y_i y_j^T = x_i M M^T x_j^T. \end{aligned} \tag{12}$$

In this way, we have the following proposition:

Proposition 1. *If the similar samples in $\mathcal{X}$ and $\mathcal{Y}$ are linearly correlated to each other as equation (11), $s_{\mathcal{X}}(i,j)$ and $s_{\mathcal{Y}}(i,j)(i,j=1,2,\ldots,n)$ are positively correlated.*

Proof. We assume that $\mathcal{X}$ has already preprocessed by whitening [53, 54] and zero-centralization; thus, $x_{\_,\varphi}$ satisfies the I.I.D (independent and identically distributed) and zero-centered assumption, that is, $x_{\_,\varphi}(\varphi = 1,\ldots,d)$ are dependent random variables that are subject to the same distribution p_θ, and its expectation is zero. It should be noted that whitening and zero-centralization will not affect the similarity between samples.

The Pearson correlation coefficient is used to measure the correlation between $s_{\mathcal{X}}(i,j)$ and $s_{\mathcal{Y}}(i,j)$:

$$\rho_{s_{\mathcal{X}} s_{\mathcal{Y}}} = \frac{\mathrm{Cov}\left(s_{\mathcal{X}}(i,j), s_{\mathcal{Y}}(i,j)\right)}{\sqrt{D\left(s_{\mathcal{X}}(i,j)\right) \cdot D\left(s_{\mathcal{Y}}(i,j)\right)}}. \tag{13}$$

□

Step 1. Proving the denominator of equation (13) is positive.

Because neither $s_{\mathcal{X}}(i,j)$ nor $s_{\mathcal{Y}}(i,j)$ is constant, the variance of them is greater than zero. Thus, the denominator of equation (13) is greater than zero:

$$\sqrt{D\left(s_{\mathcal{X}}(i,j)\right) \cdot D\left(s_{\mathcal{Y}}(i,j)\right)} > 0. \tag{14}$$

Step 2. Factorizing the numerator of equation (13) by diagonalizing MM^T.

The covariance between $s_{\mathcal{X}}(i,j)$ and $s_{\mathcal{Y}}(i,j)$ is

$$\mathrm{Cov}\left(s_{\mathcal{X}}(i,j), s_{\mathcal{Y}}(i,j)\right) = \mathrm{Cov}\left(x_i x_j^T, x_i M M^T x_j^T\right). \tag{15}$$

Because MM^T is a real symmetric matrix, it can be diagonalized as

$$MM^T = P \Lambda P^T, \tag{16}$$

where $P = [p_1^T, \ldots, p_\gamma^T, \ldots, p_d^T] \in \mathbb{R}^{d\times d}$, p_γ^T is the γ–th eigenvector of MM^T, and Λ is a diagonal matrix whose diagonal elements are eigenvalues of MM^T. Thus, we have

$$x_i MM^T x_j^T = x_i P \Lambda P^T x_j^T = \sum_{\gamma=1}^{d} \lambda_\gamma \left(x_i p_\gamma^T\right)\left(x_j p_\gamma^T\right) \cdot \sum_{\gamma=1}^{d} \lambda_\gamma \sum_{\mu=1}^{d}\sum_{\nu=1}^{d} p_{\gamma\mu} p_{\gamma\nu} x_{i\mu} x_{j\nu}, \tag{17}$$

where λ_γ is the γ-th eigenvalue of MM^T.

Substituting equation (17) into equation (15),

$$\begin{aligned} &\mathrm{Cov}\left(s_{\mathcal{X}}(i,j), s_{\mathcal{Y}}(i,j)\right) \\ &= \mathrm{Cov}\left(\sum_{\varphi=1}^{d} x_{i\varphi}x_{j\varphi}, \sum_{\gamma=1}^{d}\lambda_\gamma \sum_{\mu=1}^{d}\sum_{\nu=1}^{d} p_{\gamma\mu}p_{\gamma\nu}x_{i\mu}x_{j\nu}\right) \\ &= \sum_{\gamma=1}^{d}\lambda_\gamma \sum_{\varphi=1}^{d}\sum_{\mu=1}^{d}\sum_{\nu=1}^{d} p_{\gamma\mu}p_{\gamma\nu}\mathrm{Cov}\left(x_{i\varphi}x_{j\varphi}, x_{i\mu}x_{j\nu}\right), \end{aligned} \tag{18}$$

where

$$\begin{aligned} \mathrm{Cov}\left(x_{i\varphi}x_{j\varphi}, x_{i\mu}x_{j\nu}\right) &= E\left(x_{i\varphi}x_{j\varphi}x_{i\mu}x_{j\nu}\right) \\ &\quad - E\left(x_{i\varphi}x_{j\varphi}\right)E\left(x_{i\mu}x_{j\nu}\right). \end{aligned} \tag{19}$$

Step 3. Computing the covariance through case-by-case discussion of equation (19).

Because $x_{i\varphi}(\varphi = 1, 2, \ldots, d)$ are independent of each other and from the same distribution p_θ, we have the following conclusions.

If $\mu \neq \varphi$ and $\nu \neq \varphi$, $x_{i\varphi}$, $x_{j\varphi}$, $x_{i\mu}$, and $x_{j\nu}$ are dependent from each other; then, the covariance in equation (19) equals zero:

$$\begin{aligned} \mathrm{Cov}\left(x_{i\varphi}x_{j\varphi}, x_{i\mu}x_{j\nu}\right) &= E\left(x_{i\varphi}x_{j\varphi}x_{i\mu}x_{j\nu}\right) \\ &\quad - E\left(x_{i\varphi}x_{j\varphi}\right)E\left(x_{i\mu}x_{j\nu}\right) \\ &= E\left(x_{i\varphi}\right)E\left(x_{j\varphi}\right)E\left(x_{i\mu}\right)E\left(x_{j\nu}\right) \\ &\quad - E\left(x_{i\varphi}\right)E\left(x_{j\varphi}\right)E\left(x_{i\mu}\right)E\left(x_{j\nu}\right) \\ &= 0. \end{aligned} \tag{20}$$

If $\mu = \varphi$ and $\nu = \varphi$, the covariance in equation (19) is

$$\begin{aligned} \mathrm{Cov}\left(x_{i\varphi}x_{j\varphi}, x_{i\mu}x_{j\nu}\right) &= \mathrm{Cov}\left(x_{i\varphi}x_{j\varphi}, x_{i\varphi}x_{j\varphi}\right) \\ &= D\left(x_{i\varphi}x_{j\varphi}\right) \\ &> 0, \end{aligned} \tag{21}$$

where $D(x_{i\varphi}x_{j\varphi})$ is larger than zero because $x_{i\varphi}x_{j\varphi}$ is not a constant.

If $\mu = \varphi$ and $\nu \neq \varphi$ (or $\mu \neq \varphi$ and $\nu = \varphi$), the covariance in equation (19) is zero because $x_{\_,\varphi}$ is zero-centered:

$$\begin{aligned} \mathrm{Cov}\left(x_{i\varphi}x_{j\varphi}, x_{i\mu}x_{j\nu}\right) &= E\left(x_{i\varphi}x_{j\varphi}x_{i\varphi}x_{j\nu}\right) \\ &\quad - E\left(x_{i\varphi}x_{j\varphi}\right)E\left(x_{i\varphi}x_{j\nu}\right) \\ &= E\left(x_{i\varphi}^2\right)E\left(x_{j\varphi}\right)E\left(x_{j\nu}\right) \\ &\quad - E\left(x_{i\varphi}\right)^2 E\left(x_{j\varphi}\right)E\left(x_{j\nu}\right) \\ &= 0. \end{aligned} \tag{22}$$

From equations (20)–(22), we know that the covariance in equation (19) is not zero only when $\mu = \varphi$ and $\nu = \varphi$; then, the covariance between $s_{\mathcal{X}}(i,j)$ and $s_{\mathcal{Y}}(i,j)$ is

$$\mathrm{Cov}\left(s_{\mathcal{X}}(i,j), s_{\mathcal{Y}}(i,j)\right) = \sum_{\gamma=1}^{d}\lambda_\gamma \sum_{\varphi=1}^{d} p_{\gamma\varphi}^2 D\left(x_{i\varphi}x_{j\varphi}\right). \tag{23}$$

Step 4. Proving the covariance is positive.

Because MM^T is a positive semidefinite matrix, all λ_γ are greater than or equal to zero:

$$\lambda_\gamma(\gamma = 1, \ldots, d) \geq 0. \tag{24}$$

The sum of eigenvalues equals the sum of the diagonal elements of MM^T, which is larger than zero because M is a nonzero matrix:

$$\sum_{\gamma=1}^{d}\lambda_\gamma = \sum_{\gamma=1}^{d} m_{\gamma\gamma} > 0, \tag{25}$$

where $m_{\gamma\gamma}$ refers to the γ-th diagonal element of MM^T.

Equations (24) and (25) show that there exists at least one λ_γ which is greater than zero:

$$\exists\lambda_\gamma(\gamma = 1, \ldots, d) > 0. \tag{26}$$

Because the eigenvector p_γ is nonzero, from equation (21), we have

$$\sum_{\varphi=1}^{d}\left(p_{\gamma\varphi}\right)^2 D\left(x_{i\varphi}x_{j\varphi}\right) > 0. \tag{27}$$

Finally, from equations (26) and (27), the covariance between $s_{\mathcal{X}}(i,j)$ and $s_{\mathcal{Y}}(i,j)$ is greater than zero:

$$\mathrm{Cov}\left(s_{\mathcal{X}}(i,j), s_{\mathcal{Y}}(i,j)\right) > 0. \tag{28}$$

Step 5. Proving the Pearson coefficient is positive.

From equations (14) and (28), the Pearson correlation coefficient is larger than zero:

$$\rho_{s_{\mathcal{X}}s_{\mathcal{Y}}} > 0. \tag{29}$$

In conclusion, for any $x_i, x_j \in \mathcal{X}$ and $y_i, y_j \in \mathcal{Y}$, if $x_i \approx y_i$ and $x_j \approx y_j$, then $s_{\mathcal{X}}(i,j)$ is positively correlated to $s_{\mathcal{Y}}(i,j)$.

In the proposition and its proof, apart from being nonzero, we have no requirement for the mapping matrix M. However, some properties of M may lead to stronger conclusions. For example, low correlation between the columns of M is beneficial for high correlation between

$s_{\mathcal{X}}(i,j)$ and $s_{\mathcal{Y}}(i,j)$. In the most extreme case, if M is an orthogonal matrix, we have $s_{\mathcal{X}}(i,j) = s_{\mathcal{Y}}(i,j)$.

From Proposition 1, it can be inferred that $s_{\mathcal{X}}(i,j)$ and $s_{\mathcal{Y}}(i,j)$ are positively correlated if $x_i \approx y_i$ and $x_j \approx y_j$:

$$x_i \approx y_i,\ x_j \approx y_j \longrightarrow s_{\mathcal{X}}\left(x_i, x_j\right) \propto s_{\mathcal{Y}}\left(y_i, y_j\right). \tag{30}$$

Because reference points are also real samples in $\mathcal{X}$ and $\mathcal{Y}$, the conclusion above also holds between reference points and nonreference points. Therefore, representing multimodal samples x_i and y_i as equation (10), if the reference points x_{j_l} and y_{j_l} are matched, the values of similar cross-modal samples in the corresponding dimensions should be positively correlated:

$$x_{j'} \approx y_{j'} \longrightarrow \left(x_i \approx y_i \longrightarrow x^r_{ij} \propto y^r_{ij}\right). \tag{31}$$

Thus, if all the reference points of $\mathcal{X}$ and $\mathcal{Y}$ are one-to-one matched, the similarity between cross-modal samples can be measured according to the linear correlation between their reference-based representations:

$$S_{\mathcal{X},\mathcal{Y}}(i,j) = \frac{\left(x^r_i - \overline{x^r}\right) \cdot \left(y^r_j - \overline{y^r}\right)}{\left|x^r_i - \overline{x^r}\right| \left|y^r_j - \overline{y^r}\right|}, \tag{32}$$

where $\overline{x^r}$ and $\overline{y^r}$ are mean vectors of x^r_i and y^r_j.

Moreover, we have $\overline{x^r} = 0$ and $\overline{y^r} = \mathbf{0}$ because the cosine similarity is normalized; then, the reference-based representation of both modalities can be considered homogeneous. Therefore, the cross-modal similarity $S_{\mathcal{X},\mathcal{Y}}(i,j)$ can be directly computed with the cosine similarity:

$$S_{\mathcal{X},\mathcal{Y}}(i,j) = \frac{x^r_i \cdot y^r_j}{\left|x^r_i\right| \left|y^r_j\right|}. \tag{33}$$

Although the analysis above is somewhat lengthy, the cross-modal similarity computation based on which is quiet simple. The core of similarity computation is a multimodal reference set $R(\mathcal{X},\mathcal{Y}) = \{(x_{1'}, y_{1'}), \ldots, (x_{i'}, y_{i'}), \ldots, (x_{k'}, y_{k'})\}$, where x_{i_l} and y_{i_l} are matched cross-modal samples. However, the reference selection method in Section 3.2 only suits single-modal data, and we cannot expect that the reference sets generated, respectively, will be one-to-one matched.

3.4. Multimodal Reference Selection Based on Active Learning. The multimodal reference set $R(\mathcal{X},\mathcal{Y})$ plays two roles in *Abstraction* and *Association*, respectively: on the one hand, the reference set is the guarantee of satisfactory abstract representation for a modality; on the other hand, the correspondence relation between reference points is the basis of aligning the single-modal representations. We must comprehensively consider these two roles because both are crucial for accurate similarity computation.

In this section, we design an active learning-based strategy for the selection of the multimodal reference set $R(\mathcal{X},\mathcal{Y})$, based on which the similarity computation in equation (33) can be achieved.

From the analysis in Section 3.2, there exists a positive correlation between semantic structures of different modalities. Hence, the neighbor structure of different modalities should be similar. Therefore, if x_i is selected as a reference point of $\mathcal{X}$, its correspondent y_i can also be the reference point of $\mathcal{Y}$:

$$x_i \in \mathcal{X}',\ x_i \approx y_j \longrightarrow y_j \in \mathcal{Y}'. \tag{34}$$

Thus, we select the reference points for one single modality as in Section 3.2; then, the corresponding samples in the other modality are used as the reference points of this modality, which are obtained by asking the oracle. It is also important to choose the reference point from which modality. It is recommended to choosing one that has a clear group structure, which can bring better performance. Also, the cost of matching should also be considered. For example, the cost of querying images from text and querying text from images is different.

Finally, combining the similarity computing method in Section 3.3 with the reference selecting method above, we propose the semantic structure matching with the active learning (SSM-AL) method in Algorithm 1.

First, the multimodal reference set $R(\mathcal{X},\mathcal{Y})$ is generated in Steps 1 ~ 5: divide $\mathcal{X}$ (or $\mathcal{Y}$) into clusters by clustering, and take the centers of all clusters as the reference set $\mathcal{X}'$; then, query the corresponding sample y_j of each $x_i \in \mathcal{X}'$ and take it as the reference set $\mathcal{Y}'$. Thus, $\mathcal{X}$ and $\mathcal{Y}$ can be represented as equation (10) with reference sets $\mathcal{X}'$ and $\mathcal{Y}'$. Finally, the cross-modal similarity matrix can be computed directly according to the linear correlation between reference-based representations as equation (33).

The computational complexity of SSM-AL is analyzed as follows. Given the retrieval problem between $\mathcal{X} = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^{n \times d}$ and $\mathcal{X} = \{y_1, y_2, \ldots, y_m\} \in \mathbb{R}^{m \times e}$, the complexity of k-medoids clustering on $\mathcal{X}$ is $n \times d \times k$. The complexity of computing the representation of $\mathcal{X}$ and $\mathcal{Y}$ is $d \times k \times n$ and $e \times k \times m$. The complexity of cross-modal similarity is $n \times m \times k$. Considering d and e are constant, k is much smaller than m and n [35]; then, the complexity of the SSM-AL method is $O(dkn + dkn + ekm + nmk) = O(mn)$.

4. Experiments

In this section, we perform some experiments to evaluate the performance of the proposed method.

4.1. Data Sets. We use four benchmark data sets to evaluate the performance of the proposed method: Pascal-Sentences [55], Wikipedia [1], XMedia [56], and MSCOCO [57].

Pascal-Sentences: a subset of Pascal VOC, which contains 1,000 pairs of images and the corresponding text description from twenty categories.

Wikipedia: a data set containing 2,866 pairs of images and text from ten categories. Each pair of image and text is extracted from Wikipedia's articles [1].

XMedia: a publicly available data set consisting of five media types (text, image, video, audio, and 3D model).

Require: Two data set $\mathcal{X}$, $\mathcal{Y}$, and reference size k
Ensure: Cross-modal similarity matrix $S_{\mathcal{X},\mathcal{Y}}$
(1) Divide $\mathcal{X}$ into k clusters
(2) $\mathcal{X}' \longleftarrow$ the cluster centers of $\mathcal{X}$
(3) **for all** $x_i \in \mathcal{X}'$ **do**
(4) $\mathcal{Y}' \longleftarrow$ find one $y_i \in \mathcal{Y}$ that $x_i \approx y_i$ by asking *oracle*
(5) **end for**
(6) $\mathfrak{R}^{\mathcal{X}} \longleftarrow$ represent $\mathcal{X}$ with $\mathcal{X}'$ as equation (10)
(7) $\mathfrak{R}^{\mathcal{Y}} \longleftarrow$ represent $\mathcal{Y}$ with $\mathcal{Y}'$ as equation (10)
(8) **for all** $x_i^r \in \mathfrak{R}^{\mathcal{X}}, y_j^r \in \mathfrak{R}^{\mathcal{Y}}$ **do**
(9) $S_{\mathcal{X},\mathcal{Y}}(i, j) \longleftarrow$ compute similarity between x_i^r and y_j^r as equation (33)
(10) **end for**

Algorithm 1: SSM-AL.

We only use the image and text data in this paper, i.e., 5,000 pairs of images and text from twenty categories.

MSCOCO: a large data set containing 123,287 images and their annotated sentences. Each image is annotated by five independent sentences.

Following the existing works [24, 30], we take 20% samples as the testing set for Wikipedia, Pascal-Sentences, and XMedia. The testing set of MSCOCO is split as [58, 59]. The scale of training sets is set small because we aim to test performance with insufficient training samples.

4.2. Evaluation Protocol. We compare the retrieval performance of the proposed method with eight baselines:

CCA [1]: with canonical correlation analysis (CCA), a shared space is learned for different modalities where they are maximally correlated.

HSNN [60]: the heterogeneous similarity is measured by the probability of two cross-modal objects belonging to the same semantic category, which is achieved by analyzing the homogeneous nearest neighbors of each object.

JRL [56]: through semisupervised regularization and sparse regularization, JRL learns a common space using semantic information.

JFSSL [61]: a multimodal graph regularization is used to preserve the intermodality and intramodality similarity relationships.

CMCP [62]: a novel cross-modal correlation propagation method considering both positive relation and negative relation between cross-modal objects.

JGRHML [63]: a joint graph-regularized heterogeneous metric learning method, which integrates the structure of different modalities into a joint graph regularization.

VSEPP [59]: a learning visual-semantic embedding technique for cross-modal retrieval, which introduces a simple change to common loss functions used for multimodal embeddings.

GXN [34]: a cross-modal feature embedding method that incorporates generative processes, which can well match images and sentences with complex content.

SSM-AL: the proposed method has two settings: reference selection based on text clustering, denoted as SSM-AL^T and reference selection based on image clustering, denoted as SSM-AL^I. Each cluster corresponds to a coupled training sample; then, the cluster number is manually specified as the training sample N.

Among these methods, CCA, VSEPP, GXN, and our proposed SSM-AL are unsupervised methods that do not use class labels completely; HSNN, JFSSL, and CMCP are supervised methods where class labels are necessary; and JRL and JGRHML are semisupervised methods that need class labels of some samples.

For Pascal-Sentences, Wikipedia, and XMedia, a query item and a target item are considered actually similar if they share the same class label [30]. Mean average precision (MAP) is used to evaluate the performance in these data sets, which is a widely used metric of information retrieval [64]:

$$\mathrm{MAP} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \mathrm{AP}(i), \tag{35}$$

where Q is the query set (for example, in the image-to-text retrieval task, Q refers to all the images in the testing set, regardless of the class) and AP(i) is the average precision of query sample i. For the query x_i, the average precision can be computed as

$$\mathrm{AP}(i) = \frac{1}{L_i} \sum_{j=1}^{|T|} \mathrm{P}(j)\delta(j), \tag{36}$$

where L_i denotes the number of target samples that are actually similar to the i–th query (for these three data sets, L_i is also the number of target items that share the same class label with the query), T is the set of all target items, P(j) considers the position of the ranked target list and can be computed as $1/j$, $\delta(j) = 1$ if the j-th sample is similar to x_i, and $\delta(j) = 0$, otherwise. In cross-modal literature [1, 4, 62], two samples are considered similar if they share the same label. The MAP score can comprehensively reflect the quality

of ranked target list of all queries. Both MAP scores of bidirectional retrieval (image-to-text and text-to-image) are reported, and higher MAP indicates better performance of a method.

In contrast to the other three data sets, MSCOCO has no definite class labels. Following [28, 30], we considered a query and a target are actually similar only if they are the coupled image-text pair from the data set, and take the score of Recall @K instead of MAP as the performance metric:

$$\text{Recall @ } K = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\left(\frac{1}{L_i}\sum_{j=1}^{K}\phi(i,j)\right), \tag{37}$$

where $\phi(i,j) = 1$ if the j–th item in the ranked target list is actually similar to the i-th query; otherwise,. L_i is the number of targets that are actually similar to the i-th query (for the MSCOCO data set, $L_i = 1$ because each query only has one similar target). Another metric Precision@K is not reported because it is closely related to Recall @K in this data set. More specifically, because each query in this data set has one similar target, Precision @K = Recall @K/K [28]. We only report Recall @K of unsupervised methods (SSM-AL, CCA, VSEPP, and GXN) because supervised and semisupervised methods cannot be conducted on MSCOCO.

4.3. Retrieval Performance Comparisons. In this section, we compare the bidirectional (image-to-text and text-to-image) retrieval performance of SSM-AL and the baselines. The number of coupled cross-modal training samples (also the number of clusters and reference points in SSM-AL) is denoted as N.

MAP values of some methods with small training sets are not reported in Tables 1–3 because they cannot finish with such limited training data, which are marked with "—." Besides, to evaluate the impact of the number of reference points on the retrieval performance, we draw the MAP-N curves of the best-performing SSM-AL method and three representative baselines that have regular trends in Figures 5–7.

(1) The result of Pascal-Sentences: in Table 1, SSM-AL^I outperforms all the baselines in the retrieval task of both directions, including supervised, semi-supervised, and unsupervised methods. When $N = 10$, the MAP values of SSM-AL^T are lower than and SSM-AL^I but higher than baselines. The MAP values of JRL, JFSSL, JGRHML, and HSNN are not reported because they cannot finish normally. CCA and CMCP perform worse than SSM-AL but better than VSEPP and GXN. When $N = 50$, SSM-AL^T and SSM-AL^I still perform the best. JFSSL and HSNN perform the second-best in the image-to-text task and the text-to-image task, respectively. When N increases to 100, SSM-AL^I is still the best-performing method in both directions. The performance of CMCP increases a lot and exceeds SSM-AL^T in the image-to-text retrieval but still worse than SSM-AL^I. The MAP values of JRL, JFSSL, VSEPP, GXN, and CCA are obviously lower than other methods.

From Figure 5, in general, more reference points always bring higher performance in both retrieval tasks. The increasing speed is high when $N < 50$ but then slows down. The MAP value of CMCP decreases first and then increases fast when $N > 25$. Both performances of VSEPP and HSNN also increase with increasing N, but the speed of the former is much slower than the others.

(2) The result of Wikipedia: in Table 2, MAP values of SSM-AL^T and SSM-AL^I are higher than others, and SSM-AL^T performs best in both retrieval tasks. When $N = 10$, the retrieval performance of two SSM-AL methods is obviously higher than the other four. When $N = 50$, SSM-AL^T and SSM-AL^I also outperform the eight baselines. JRL performs better than the other baselines but is obviously poorer than SSM-AL^T and SSM-AL^I. CMCP, JGRHML, and HSNN have similar MAP values in both retrieval tasks. The MAP values of CCA, VSEPP, and GXN in the text-to-image retrieval task are similar, while the MAP value of CCA in the image-to-text task is higher than the other two. JFSSL method performs worst in all the methods. When $N = 100$, the MAP values of CMCP, JGRHML, and HSNN increase obviously but still lower than SSM-AL^T and SSM-AL^I. The performance of CCA, JRL, JFSSL, VSEPP, and GXN does not show a significant improvement.

In Figure 6, the MAP values of SSM-AL^T in both tasks also increase with the number of reference points. The performance gain of more reference points in the image-to-text task is less obvious than that in the text-to-image task. The MAP value of CMCP also decreases first and then increases fast. The performance of VSEPP in two tasks does not show visible changes as N is increasing.

(3) The result of XMedia: in Table 3, SSM-AL^T and SSM-AL^I outperform all the baselines, and SSM-AL^T performs better than SSM-AL^I. When $N = 40$, the performance of CMCP is worse than that of SSM-AL^T and SSM-AL^I but is obviously better than that of CCA, JGRHML, HSNN, VSEPP, and GXN. When $N = 200$, the MAP values of SSM-AL^T and SSM-AL^I in both retrieval tasks are still higher than the baselines. The MAP values of CMCP and JGRHML in the image-to-text retrieval task increase obviously and are higher than the other baselines; also, the MAP values of JFSSL, CMCP, and JGRHML in the text-to-image retrieval task are higher than the other baselines. When $N = 400$, SSM-AL^T and SSM-AL^I still perform the best. With larger N, MAP values of JFSSL and JGRHML do not show significant improvement.

In Figure 7, the MAP value of SSM-AL^T increases along with the increasing reference point number, especially when the number N is small. More reference points bring obvious performance gain for SSM-AL when $N < 100$; when N is larger than 100, the speed of performance gaining is much slower.

Table 1: MAP of the bidirectional retrieval task on Pascal-Sentences.

N	Task	CCA	SSM-AL^T	SSM-AL^I	JRL	JFSSL	CMCP	JGRHML	HSNN	VSEPP	GXN
10	Image to text	0.1275	0.2161	**0.2263**	—	—	0.1372	—	—	0.0875	0.0769
	Text to image	0.1275	0.2005	**0.2043**	—	—	0.1193	—	—	0.0814	0.0799
50	Image to text	0.1252	0.3207	**0.3484**	0.1275	0.2103	0.1485	0.1275	0.1271	0.1215	0.1186
	Text to image	0.0831	0.2999	**0.3174**	0.1275	0.2169	0.1359	0.1275	0.1636	0.1207	0.1148
100	Image to text	0.0899	0.3354	**0.3764**	0.1275	0.1010	0.3441	0.2407	0.2625	0.1331	0.1429
	Text to image	0.0704	0.3189	**0.3482**	0.1275	0.1058	0.2981	0.2932	0.2724	0.1322	0.1455

Table 2: MAP of the bidirectional retrieval task on Wikipedia.

N	Task	CCA	SSM-AL^T	SSM-AL^I	JRL	JFSSL	CMCP	JGRHML	HSNN	VSEPP	GXN
10	Image to text	0.1215	**0.2103**	0.1868	—	—	0.1761	—	—	0.1223	0.1201
	Text to image	0.1215	**0.1848**	0.1703	—	—	0.1222	—	—	0.1212	0.1208
50	Image to text	0.1678	**0.2575**	0.2268	0.2092	0.1163	0.1904	0.1993	0.2084	0.1238	0.1229
	Text to image	0.1138	**0.2273**	0.2051	0.2092	0.1095	0.1506	0.1612	0.1438	0.1266	0.1259
100	Image to text	0.1222	**0.2621**	0.2243	0.2092	0.1121	0.2592	0.2153	0.2321	0.1275	0.1270
	Text to image	0.1059	**0.2438**	0.2197	0.2092	0.1098	0.2063	0.1767	0.1817	0.1285	0.1289

Table 3: MAP of the bidirectional retrieval task on XMedia.

N	Task	CCA	SSM-AL^T	SSM-AL^I	JRL	JFSSL	CMCP	JGRHML	HSNN	VSEPP	GXN
40	Image to text	0.0569	**0.2827**	0.2781	—	—	0.2317	0.0569	0.0569	0.0691	0.0689
	Text to image	0.0569	**0.3485**	0.3350	—	—	0.2590	0.0569	0.0737	0.0510	0.0501
200	Image to text	0.0566	**0.6454**	0.6421	0.0572	0.1297	0.2403	0.2018	0.0800	0.0683	0.0644
	Text to image	0.0568	**0.7365**	0.7355	0.0572	0.4161	0.4139	0.3345	0.1161	0.0505	0.0517
400	Image to text	0.0574	**0.6902**	0.6729	0.4573	0.0669	0.2971	0.1410	0.1801	0.0663	0.0671
	Text to image	0.0580	**0.7775**	0.7629	0.4869	0.0832	0.5038	0.1626	0.2101	0.0481	0.0489

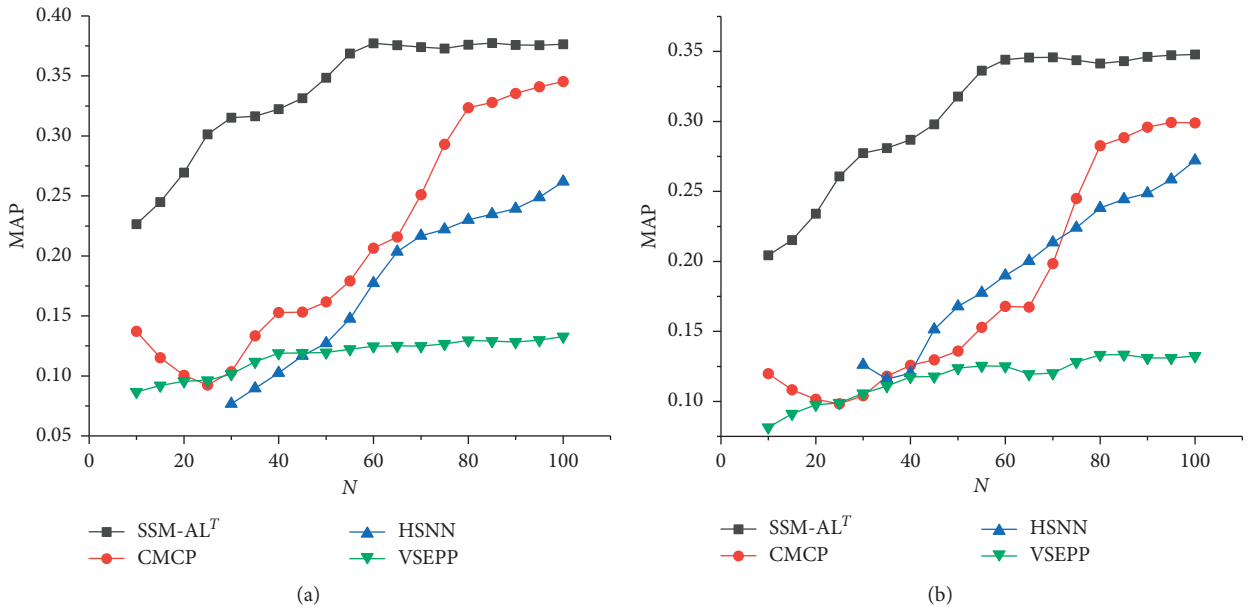

Figure 5: MAP values of Pascal-Sentences change along with N: (a) image to text; (b) text to image.

Although MAP values of CMCP and HSNN also increase as N increasing, compared to SSM-AL^T, the speed is much slower. The performance of VSEPP still does not show significant changes.

(4) The result of MSCOCO: in Figure 8, SSM-AL^T and SSM-AL^I outperform three baselines with different K values in the image-to-text retrieval task, and the gap between the SSM-AL methods and the others is quite large. In general, the performance of SSM-AL^T, SSM-AL^I, VSEPP, and GXN improves as the number of training samples increases, while that of CCA is the lowest and shows no visible change along with

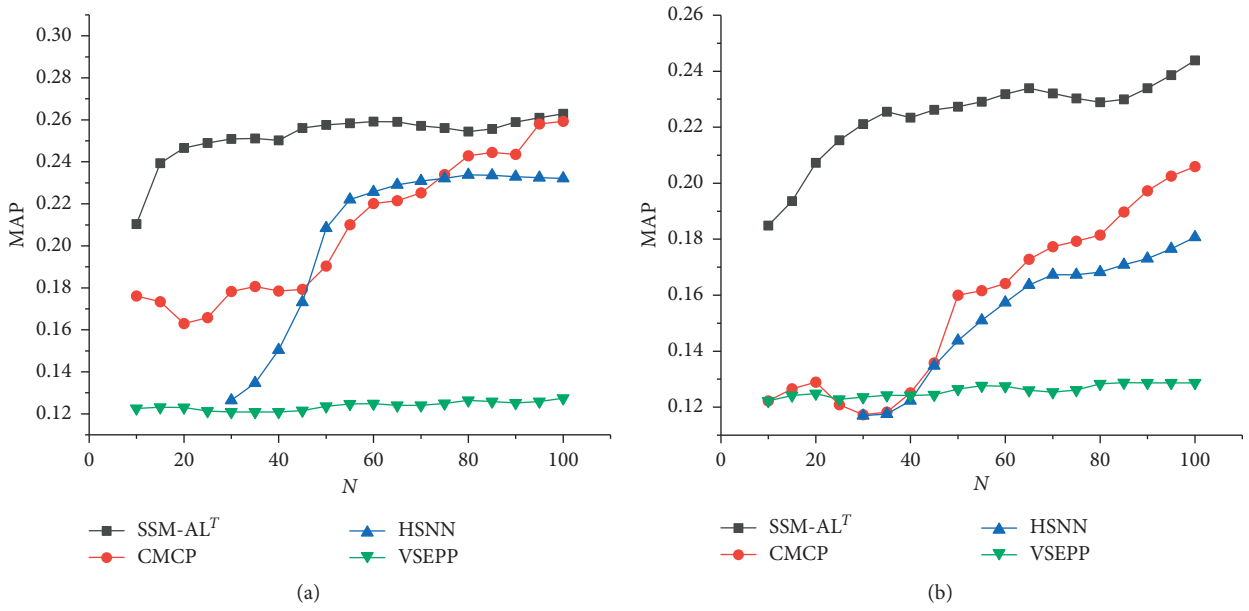

Figure 6: MAP values of Wikipedia change along with N: (a) image to text; (b) text to image.

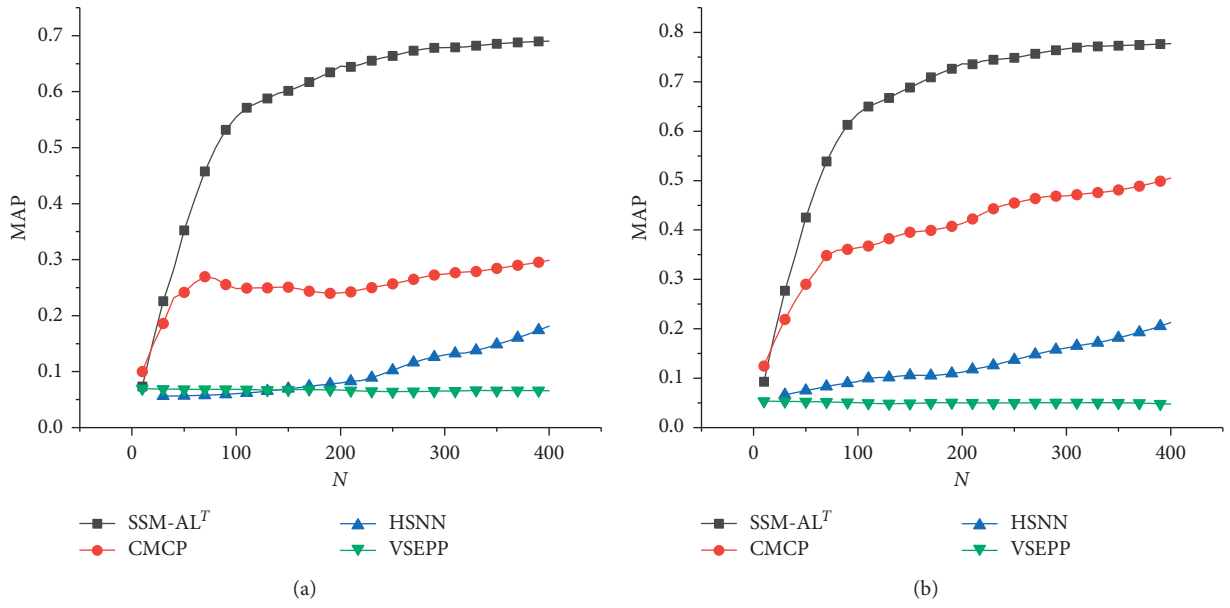

Figure 7: MAP values of XMedia change along with N: (a) image to text; (b) text to image.

the number of training samples. Although, in Figure 8(a), Recall @1 of SSM-AL^T and SSM-AL^I increases firstly then decreases slightly, it is still higher than Recall @1 of three baselines.

In Figure 9, Recall @K of SSM-AL^T, SSM-AL^I, VSEPP, and GXN increases as the number of the training samples increases, and SSM-AL methods still perform best in the text-to-image retrieval task. Although in four figures, Recall @K of VSEPP and GXN increases rapidly but is always lower than that of two SSM-AL methods. Recall @K of CCA is the lowest when $K = 1, 5, 10$, and 50 and shows no visible change along with the number of training samples. Moreover, SSM-AL^T and SSM-AL^I get similar Recall @K scores in general; however, Recall @K of SSM-AL^I is slightly higher when $K \geq 5$.

It can be concluded that the proposed SSM-AL^T and SSM-AL^I outperform all the baselines when matched training samples are insufficient, even though they do not use any label information. Experiment results prove the importance of the intramodal relation learning with uncoupled samples and the simple correlation between high-

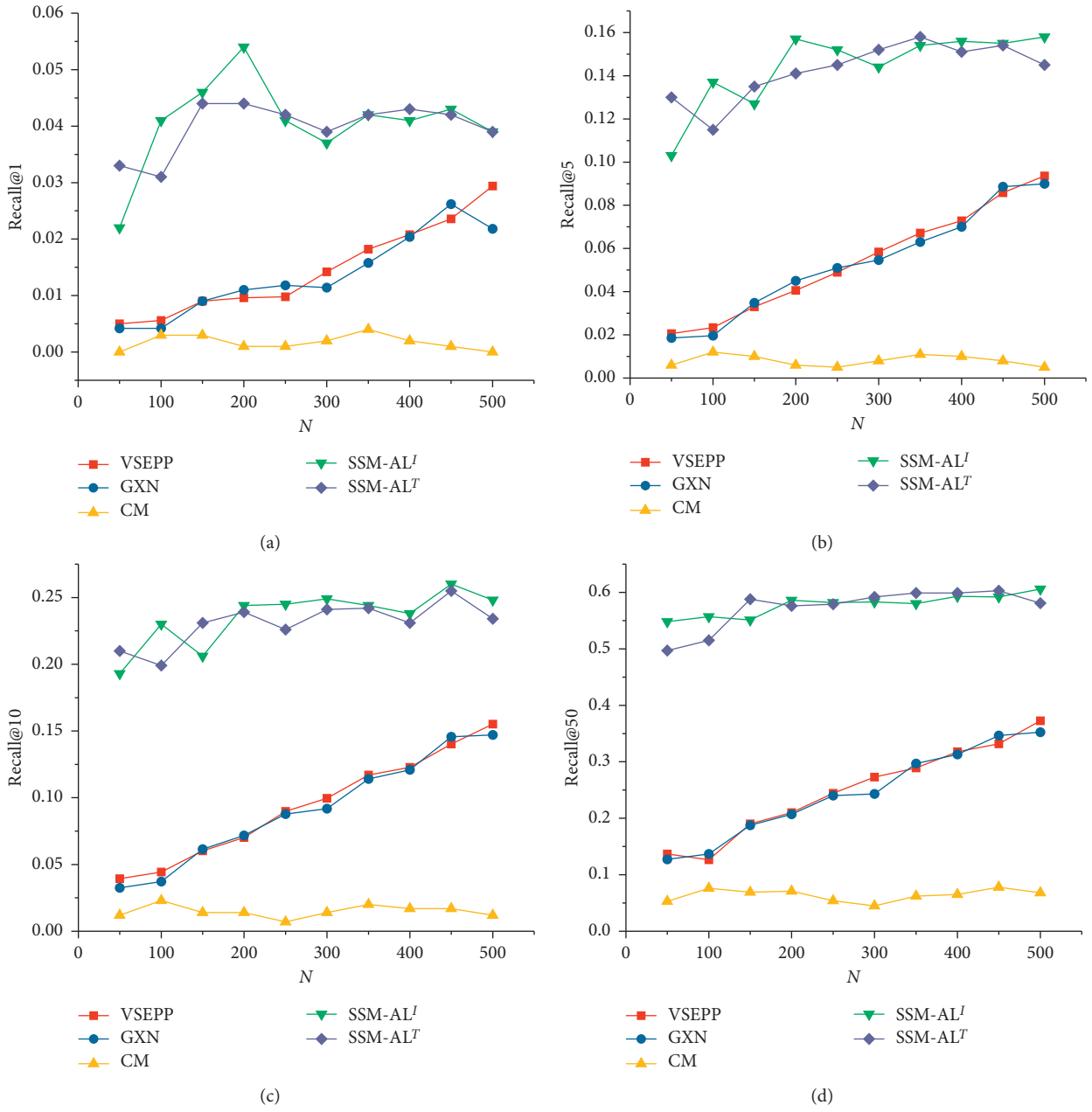

Figure 8: Recall @K of image-to-text retrieval on MSCOCO: (a) Recall @1, (b) Recall @5, (c) Recall @10, and (d) Recall @50.

level cross-modal concepts, as well as the effectiveness of the *Abs-Ass* framework. The performance gap between the proposed method and the baselines decreases with the increase of coupled training data; however, the proposed method still is an ideal choice with limited coupled training data. In addition, the DNN-based methods (VSEPP and GXN) do not show an advantage over the traditional ones when training samples are not sufficient.

Overall, more reference points are beneficial to the retrieval performance of SSM-AL: in most cases, the performance of SSM-AL improves with more reference points. However, it does not hold that the more the reference points, the better. On the one hand, more reference points mean higher costs for matching cross-modal samples; on the other hand, the performance gaining becomes limited when the number of reference points is high. In practice, the number of reference points should be decided according to the performance demand and the cost.

The above results also show that the retrieval performance is different when using different data as the basis of reference point selection. SSM-AL^T performs better than SSM-AL^I in the Wikipedia and XMedia data set, while SSM-AL^I performs better than SSM-AL^T in the Pascal-Sentences and MSCOCO data set.

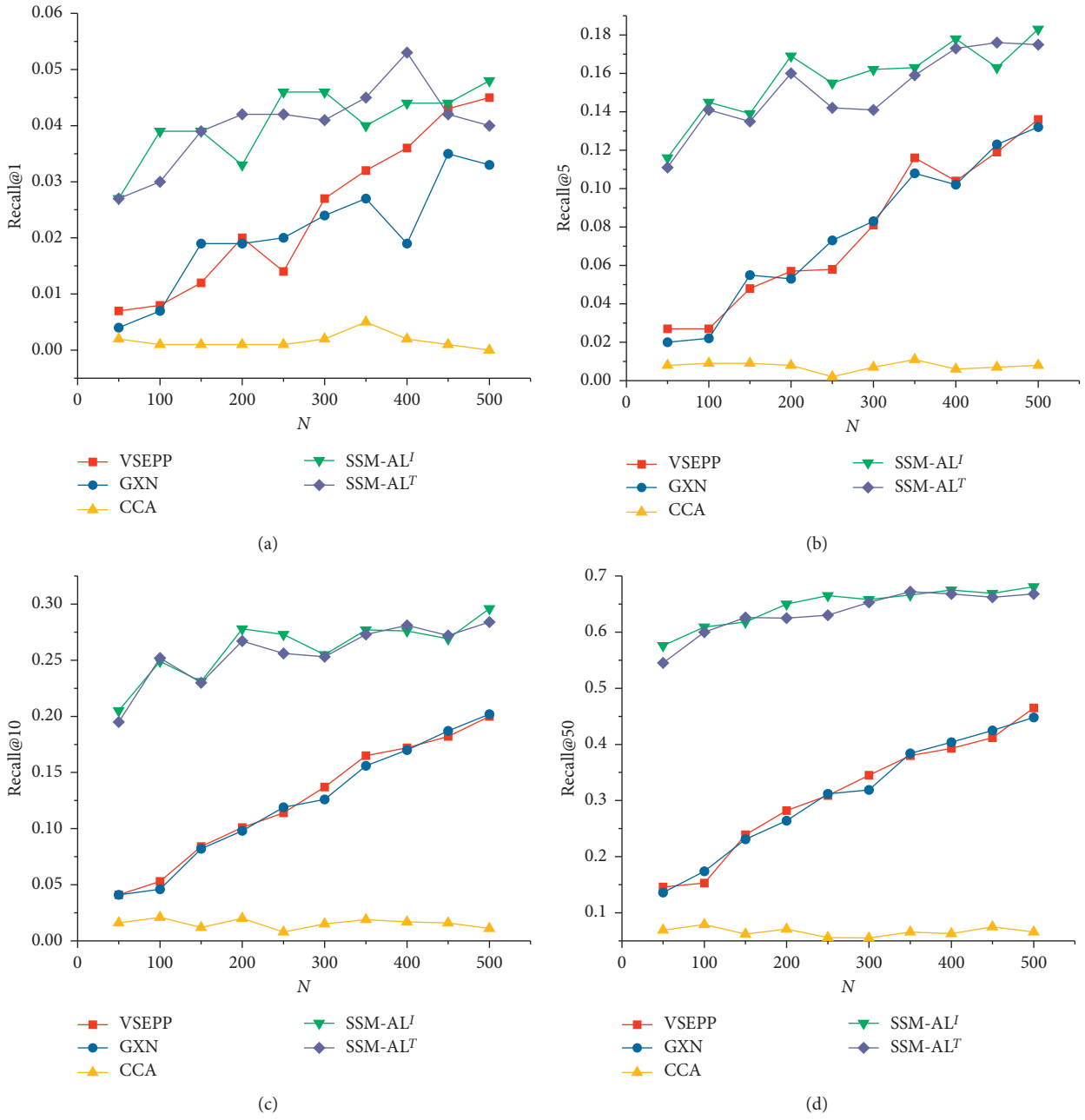

Figure 9: Recall @K of text-to-image retrieval on MSCOCO: (a) Recall @1, (b) Recall @5, (c) Recall @10, and (d) Recall @50.

5. Conclusion

In this paper, we try to improve the performance of cross-modal retrieval when training data are insufficient. Different from existing works, our proposed framework and its implementation emphasize the intramodal relation learning from data itself; no additional information (such as class labels and annotations from the web) is used as supplementary. The idea of this work is meaningful, especially when coupled training samples are insufficient; thus, it can be very helpful when applying cost is an essential consideration. Also, it can be incorporated in other methods to solve the cold-start problem of the cross-modal retrieval task.

The future work lies in two folds. On the one hand, this work can be improved by incorporating the class labels of a few samples when aligning the semantic structure of different modalities. On the other hand, we attempt to extend this work to other modalities, such as video and audio.

Data Availability

The multimodal data supporting this work are from previously reported studies and data sets, which have been cited. All of them are open access and available at the Internet.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was partially funded by the National Natural Science Foundation of China (nos. 91648204 and 61532007), the National Key Research and Development Program of China (nos. 2017YFB1001900 and 2017YFB1301104), the National Science Foundation for Young Scientists of China (Grant no. 61802426), and the National Science and Technology Major Project.

References

[1] N. Rasiwasia, J. C. Pereira, E. Coviello et al., "A new approach to cross-modal multimedia retrieval," in *Proceedings of the International Conference on Multimedia*, pp. 251–260, Firenze, Italy, 2010.

[2] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: concepts, methodologies, benchmarks, and challenges," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2372–2385, 2018.

[3] P. J. Costa, E. Coviello, G. Doyle et al., "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 36, no. 3, pp. 521–535, 2014.

[4] Y.-x. Peng, W.-w. Zhu, Y. Zhao et al., "Cross-media analysis and reasoning: advances and directions," *Frontiers of Information Technology & Electronic Engineering*, vol. 18, no. 1, pp. 44–57, 2017.

[5] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proceedings of the ACM International Conference on Multimedia*, ACM, Mountain View, CA, USA, pp. 154–162, 2017.

[6] A. Karpathy, A. Joulin, and F. F. Li, "Deep fragment embeddings for bidirectional image sentence mapping," in *roceedings of the Advances In Neural Information Processing Systems*, pp. 1889–1897, Montreal, Canada, 2014.

[7] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: an overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[8] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: a discriminative latent space," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2160–2167, Providence, RI, USA, June 2012.

[9] Y. Mroueh, E. Marcheret, and V. Goel, "Multimodal retrieval with asymmetrically weighted truncated-svd canonical correlation analysis," 2015, http://arxiv.org/abs/1511.06267.

[10] L. Wang, W. Sun, Z. Zhao, and F. Su, "Modeling intra- and inter-pair correlation via heterogeneous high-order preserving for cross-modal retrieval," *Signal Processing*, vol. 131, pp. 249–260, 2017.

[11] Y. Jia, M. Salzmann, and T. Darrell, "Learning cross-modality similarity for multinomial data," in *Proceedings of the International Conference on Computer Vision*, pp. 2407–2414, Barcelona, Spain, November 2011.

[12] S. Roller and S. S. I. Walde, "A multimodal lda model integrating textual, cognitive and visual modalities," in *Proceedings of the Conference on Empirical Methods In Natural Language Processing*, pp. 1146–1157, Seattle, WA, USA, 2013.

[13] Y. Wang, F. Wu, J. Song, X. Li, and Y. Zhuang, "Multi-modal mutual topic reinforce modeling for cross-media retrieval," in *Proceedings of the ACM International Conference on Multimedia*, pp. 307–316, New York; NY, USA, 2014.

[14] J. Wang, S. Kumar, and S. Chang, "Sequential projection learning for hashing with compact codes," in *Proceedings of the International Conference on Machine Learning*, pp. 1127–1134, Haifa, Israel, June 2010.

[15] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1360–1365, Barcelona, Spain, July 2011.

[16] Z. Yi and D. Y. Yeung, "Co-regularized hashing for multimodal data," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 1376–1384, Lake Tahoe, NV, USA, December 2012.

[17] M. Ou, P. Cui, F. Wang, J. Wang, W. Zhu, and S. Yang, "Comparing apples to oranges: a scalable solution with heterogeneous hashing," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 230–238, Chicago, IL, USA, August 2013.

[18] F. Wu, Z. Yu, Y. Yang, S. Tang, Y. Zhang, and Y. Zhuang, "Sparse multi-modal hashing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 427–439, 2014.

[19] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the International Conference on Machine Learning*, pp. 689–696, Bellevue, WA, USA, June 2011.

[20] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 2222–2230, Lake Tahoe, NV, USA, December 2012.

[21] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proceedings of the International Conference on Machine Learning*, pp. 1247–1255, Atlanta, GA, USA, June 2013.

[22] A. Frome, G. S. Corrado, J. Shlens et al., "A deep visual-semantic embedding model," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 2121–2129, Lake Tahoe, NV, USA, December 2013.

[23] B. Jiang, J. Yang, Z. Lv, K. Tian, Q. Meng, and Y. Yan, "Internet cross-media retrieval based on deep learning," *Journal of Visual Communication and Image Representation*, vol. 48, pp. 356–366, 2017.

[24] Y. Wei, Y. Zhao, C. Lu et al., "Cross-modal retrieval with cnn visual features: a new baseline," *IEEE Transactions on Cybernetics*, vol. 47, no. 47, pp. 449–460, 2017.

[25] N. Gao, S.-J. Huang, Y. Yan, and S. Chen, "Cross modal similarity learning with active queries," *Pattern Recognition*, vol. 75, pp. 214–222, 2018.

[26] Y. Peng, X. Huang, and J. Qi, "Cross-media shared representation by hierarchical learning with multiple deep networks," in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 3846–3853, New York, NY, USA, June 2016.

[27] Y. Peng, J. Qi, X. Huang, and Y. Yuan, "CCL: cross-modal correlation learning with multigrained fusion by hierarchical network," *IEEE Transactions on Multimedia*, vol. 20, no. 2, pp. 405–420, 2018.

[28] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3441–3450, Boston, MA, USA, June 2015.

[29] Y. Zhan, J. Yu, Z. Yu, R. Zhang, D. Tao, and Q. Tian, "Comprehensive distance-preserving autoencoders for cross-modal retrieval," in *Proceedings of the ACM International Conference on Multimedia*, pp. 1137–1145, ACM, Amsterdam, The Netherlands, June 2018.

[30] Y. Peng, J. Qi, and Y. Yuan, "Modality-specific cross-modal similarity measurement with recurrent attention network," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5585–5599, 2018.

[31] A. Klementiev, I. Titov, and B. Bhattarai, "Inducing cross lingual distributed representations of words," in *Proceedings of the International Conference on Computational Linguistics*, pp. 1459–1474, Mumbai, India, December 2012.

[32] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," 2013, http://arxiv.org/abs/1309.4168.

[33] S. Gouws, Y. Bengio, and G. Corrado, "Bilbowa: fast bilingual distributed representations without word alignments," in *Proceedings of the International Conference on Machine Learning*, pp. 748–756, Lille, France, July 2015.

[34] J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang, "Look, imagine and match: improving textual-visual cross-modal retrieval with generative models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7181–7189, Salt Lake, UT, USA, June 2018.

[35] Q. Zheng, X. Diao, J. Cao et al., "From whole to part: reference-based representation for clustering categorical data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 3, pp. 927–937, 2020.

[36] G. Collell and M.-F. Moens, "Do neural network cross-modal mappings really bridge modalities?" in *Proceedings 56th of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, July 2018.

[37] J. Liang, R. He, Z. Sun, and T. Tan, "Group-invariant cross-modal subspace learning," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 1739–1745, New York, NY, USA, July 2016.

[38] J. Song, H. Zhang, X. Li, L. Gao, M. Wang, and R. Hong, "Self-supervised video hashing with hierarchical binary auto-encoder," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3210–3221, 2018.

[39] Z. Ye and Y. Peng, "Multi-scale correlation for sequential cross-modal hashing learning," in *Proceedings of the ACM International Conference on Multimedia*, ACM, Amsterdam, The Netherlands, pp. 852–860, June 2018.

[40] X. Liu, Z. Hu, H. Ling, and Y.-m. Cheung, "MTFH: a matrix tri-factorization hashing framework for efficient cross-modal retrieval," 2018, http://arxiv.org/abs/1805.01963.

[41] J. Qi, Y. Peng, and Y. Yuan, "Cross-media multi-level alignment with relation attention network," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 892–898, AAAI Press, Stockholm, Sweden, July 2018.

[42] L. Gao, X. Li, J. Song, and H. T. Shen, "Hierarchical LSTMS with adaptive attention for visual captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 5, pp. 1112–1131, 2019.

[43] J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, and H. T. Shen, "From deterministic to generative: multimodal stochastic rnns for video captioning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 10, pp. 3047–3058, 2019.

[44] N. C. Mithun, R. Panda, E. E. Papalexakis, and A. K. Roy-Chowdhury, "Webly supervised joint embedding for cross-modal image-text retrieval," in *Proceedings of the ACM International Conference on Multimedia*, New York, NY, USA, 2018.

[45] Y. Li, D. Wang, H. Hu, Y. Lin, and Y. Zhuang, "Zero-shot recognition using dual visual-semantic mapping paths," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5207–5215, Honolulu, HI, USA, July 2017.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, http://arxiv.org/abs/1409.1556.

[47] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," in *Proceedings of the International Conference on Learning Representations*, Toulon, France, April 2017.

[48] J. Pennington, R. Socher, and C. Manning, "Glove: global vectors for word representation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, October 2014.

[49] Y. Qian, F. Li, J. Liang, B. Liu, and C. Dang, "Space structure and clustering of categorical data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 10, pp. 2047–2059, 2016.

[50] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: analysis and an algorithm," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 849–856, Vancouver, Canada, December 2002.

[51] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3336–3341, 2009.

[52] A. McCallum, K. Nigam, and L. H. Ungar, "Efficient clustering of high-dimensional data sets with application to reference matching," in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 169–178, Kyoto, Japan, April 2000.

[53] A. J. Bell and T. J. Sejnowski, "The "independent components" of natural scenes are edge filters," *Vision Research*, vol. 37, no. 23, pp. 3327–3338, 1997.

[54] A. C. Koivunen and A. B. Kostinski, "The feasibility of data whitening to improve performance of weather radar," *Journal of Applied Meteorology*, vol. 38, no. 6, pp. 741–749, 1999.

[55] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *Proceedings of the NAACL Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 139–147, Los Angeles, CA, USA, June 2010.

[56] X. Zhai, Y. Peng, and J. Xiao, "Learning cross-media joint representation with sparse and semisupervised regularization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 6, pp. 965–978, 2014.

[57] T. Y. Lin, M. Maire, S. Belongie et al., "Microsoft coco: common objects in context," in *Proceedings of the European Conference on Computer Vision*, Springer, Zürich, Switzerland, pp. 740–755, 2014.

[58] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137, Boston, MA, USA, June 2015.

[59] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: improving visual-semantic embeddings with hard negatives," in

Proceedings of the British Machine Vision Conference (BMVC), Tyne, UK, 2018.

[60] X. Zhai, Y. Peng, and J. Xiao, "Effective heterogeneous similarity measure with nearest neighbors for cross-media retrieval," in *Proceedings of the International Conference on Multimedia Modeling*, Springer, Klagenfurt, Austria, pp. 312–322, January 2012.

[61] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2010–2023, 2016.

[62] X. Zhai, Y. Peng, and J. Xiao, "Cross-modality correlation propagation for cross-media retrieval," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2337–2340, Kyoto, Japan, March 2012.

[63] X. Zhai, Y. Peng, and J. Xiao, "Heterogeneous metric learning with joint graph regularization for cross-media retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1198–1204, Bellevue, WA, USA, July 2013.

[64] N. Rasiwasia, P. J. Moreno, and N. Vasconcelos, "Bridging the gap: query by semantic example," *IEEE Transactions on Multimedia*, vol. 9, no. 5, pp. 923–938, 2007.