

Cross-media retrieval via fusing multi-modality and multi-grained data

Z. Liu^{a,b,*}, S. Yuan^{a,b}, X. Pei^{a,b}, S. Gao^{a,b}, and H. Han^{a,b}

a. School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, 250014, Shandong, China.

b. Shandong Provincial Key Laboratory of Digital Media Technology, Shandong University of Finance and Economics, Jinan, 250014, Shandong, China.

Received 25 January 2022; received in revised form 1 October 2022; accepted 9 May 2023

KEYWORDS

Cross-media retrieval;
 Multi-modality data;
 Multi-grained data;
 Multi-margin triplet
 loss;
 Margin-set.

Abstract. Traditional cross-media retrieval methods mainly focus on coarse-grained data that reflect global characteristics while ignoring the fine-grained descriptions of local details. Meanwhile, traditional methods cannot accurately describe the correlations between the anchor and the irrelevant data. This paper aims to solve the abovementioned problems by proposing to fuse coarse-grained and fine-grained features and a multi-margin triplet loss based on a dual-framework. (1) Framework I: A multi-grained data fusion framework based on Deep Belief Network, and (2) Framework II: A multi-modality data fusion framework based on the multi-margin triplet loss function. In Framework I, the coarse-grained and fine-grained features fused by the joint Restricted Boltzmann Machine are input into Framework II. In Framework II, we innovatively propose the multi-margin triplet loss. The data, which belong to different modalities and semantic categories, are stepped away from the anchor in a multi-margin way. Experimental results show that the proposed method achieves better cross-media retrieval performance than other methods with different datasets. Furthermore, the ablation experiments verify that our proposed multi-grained fusion strategy and the multi-margin triplet loss function are effective.

© 2023 Sharif University of Technology. All rights reserved.

1. Introduction

Cross-media retrieval refers to the task in which the query and the retrieval items belong to different forms of media. Due to the famous “heterogeneous gap”, which denotes that feature representations of multi-modality data are inconsistent [1], multi-modality data analysis and retrieval are facing great challenges. Therefore, cross-media retrieval aims to utilize effective

representation learning to eliminate the “heterogeneous gap”. That is to say, the representations of different modalities that describe the same object should get close to each other in the high-level semantic space [2].

Early works are mainly based on Canonical Correlation Analysis (CCA) [3,4]. Such methods can only construct data of two modalities, and then some works further introduce graph regularization to constrain the traditional subspace learning [5,6]. More recently, with the development of neural networks, Auto-Encoder has been widely applied in representation learning [7–9]. The nonlinear learning ability of neural networks achieves great performance improvement.

However, these approaches mainly focus on coarse-grained data at the global level. Later on, researchers pay more attention to the fine-grained data that contain rich information at the local level [10,11].

*. Corresponding author. Tel.: +86 13066035106
 E-mail addresses: Lzh_48@126.com (Z. Liu);
 ysj_sd@163.com (S. Yuan); pxl1998418@163.com (X. Pei);
 gsszxy@aliyun.com (S. Gao); 944915627@qq.com (H. Han)
 ORCID(s): 0000-0002-6846-5782 (Z. Liu);
 0000-0003-2072-3620 (S. Yuan)

Subsequent works optimize the relationship between multi-modality data from the perspective of metric learning [12,13], and the triplet loss [14] has become the mainstream loss function used in cross-media retrieval.

Despite the success of previous methods, two problems still need to be solved. First, most existing works only use one type of data granularity while ignoring the complementarity between the coarse-grained and the fine-grained data. Second, most studies on the triplet loss function use a simple binary strategy to optimize the distances between anchors and relevant/irrelevant instances, aiming to distinguish instances in the same semantic category and different semantic categories. However, it fails to accurately capture quantitative relationships between the anchors and the instances.

To address the problems mentioned above, in this work, we design a dual-framework model to fuse the multi-modality and multi-grained data effectively. We innovatively propose the **Multi-margin triplet loss** to improve the traditional triplet loss and use the **Margin-set**. As the irrelevant samples have different distances from the anchor, our proposed Multi-margin triplet loss aims to distinguish the irrelevant samples in a multi-margin way, thus achieving better optimization results. An illustration of “separating relevant and irrelevant in a multi-margin way” is shown in Figure 1.

The main contributions of this paper lie in the following aspects:

- **Fusing multi-grained data:** In Framework I, the fine-grained data emphasize important local details and make a supplement to the original coarse-grained data. To make the cross-media relevance more accurate, we use two symmetric Deep Belief

Networks for coarse-grained and fine-grained feature learning, then fuse them together;

- **Multi-margin triplet loss function:** In Framework II, we design a multi-margin triplet loss function and innovatively propose the concept of “Margin-set”, which contains a series of margins. The “Margin-set” differentiates distances between the anchor and irrelevant samples in a multi-margin way.

The remaining parts of the paper are organized as follows. Section 2 gives a brief review of the related works. Sections 3 and 4 explain the details of our proposed multi-grained data fusion framework and multi-modality data fusion framework, respectively. Experiments are conducted in Section 5, and Section 6 summarizes the whole paper and discusses further research.

2. Related works

Existing cross-media retrieval methods are roughly divided into two main categories: (1) common subspace learning method and (2) non-common subspace learning method.

2.1. Common subspace learning

This category is subdivided into three categories: (1) projection matrix, (2) deep learning, and (3) semantic matching.

2.1.1. Projection matrix

CCA [3] learns a linear projection matrix to maximize the pairwise correlations between heterogeneous data in the subspace. Rasiwasia et al. [4] use CCA to map the features of images and texts into the same space. Deep Canonical Correlation Analysis (DCCA) [15] can not only solve the problem of data scalability but also learn better correlation representations.

More recently, Zhai et al. [6] proposed a Joint Graph Regularization Heterogeneous Metric Learning (JGRHML), and they also proposed Joint Representation Learning (JRL) to learn common spaces with semi-supervised regularization [16]. Wang et al. [5] apply the graph regularization to preserve the inter-modality and intra-modality relationships in the common subspace learning. Peng et al. [17] proposed a Semi-Supervised framework with the Unified Patch Graph (S2UPG) regularization to learn cross-media features.

2.1.2. Deep learning

Frome et al. [18] proposed a Deep Visual-Semantic Embedding model (Devise) to learn the correlations between the image and the textual data. Socher et al. [19] used the dependency trees to map the sentence into the image representation space in the same semantic category and then measure the distances between these two modalities of data in the image

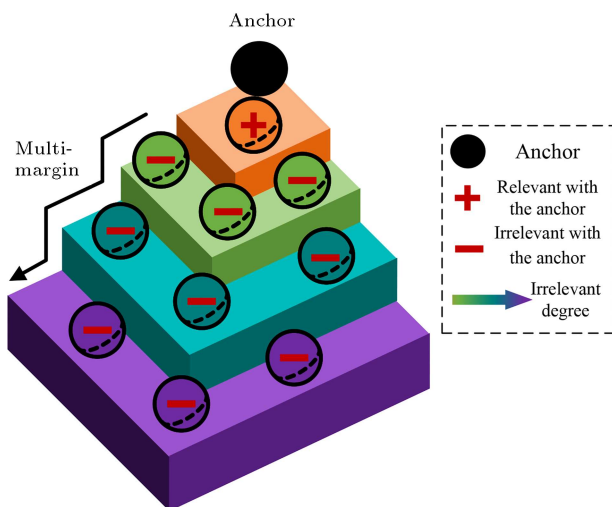


Figure 1. Illustration of the key idea: separating relevant and irrelevant in a multi-margin way.

representation space. Wang et al. [12] proposed a two-branch neural network to learn the joint embedding of image and textual data. Vendrov et al. [20] proposed a partial order structure of the hierarchical structure model to optimize feature learning. In addition, the attention mechanism [11,21] and the Graph Neural Networks [22] have been used in cross-media retrieval.

2.1.3. Semantic matching

To infer the alignment between sentences and image regions, Lee et al. [11] mapped words and image regions into a common embedding space based on the attention mechanism. Wang et al. [23] studied two-branch neural networks to learn the similarities between images and texts. Xu et al. [24] proposed a Cross-modal Attention with Semantic Consistency (CASC) for image-text matching.

2.2. Non-common subspace learning methods

This category is divided into: (1) Auto-Encoder, (2) Probabilistic Graphical Models (PGMs), and (3) Generative Adversarial Network (GAN).

2.2.1. Auto-Encoder

Unsupervised data are also used to construct the pre-train model. Ngiam et al. [7] extended the Auto-Encoders to multi-modality. Silberer and Lapata [25] proposed using a multi-modality Auto-Encoder for semantic concept prediction. Correspondence Auto-Encoder (Corr-AE) [8] uses a Deep Belief Network to generate representations for a single modality.

2.2.2. Probabilistic Graphical Models (PGMs)

Deep Boltzmann Machine (DBM) [26] has been widely used to train data representations. The network uses Restricted Boltzmann Machine (RBM) [27] as a block. Peng et al. [28] concentrated on the fine-grained data, and designed a multi-task learning strategy to adaptively balance the constraints of semantic category.

2.2.3. Generative Adversarial Network (GAN)

Inspired by adversarial learning, Zhang et al. [29] proposed a semi-supervised cross-media generative confrontation hashing method, which uses GANs to recognize the confused samples. Zhang et al. [30] proposed an unsupervised cross-media generative counter-hashing method, which effectively models the manifold structure of data for hash code generation. Wang et al. [31] proposed an adversarial cross-media retrieval framework to mine cross-media relationships with adversarial learning strategies.

There are still some shortcomings in existing methods:

1. Some existing works only use fine-grained data to focus on semantic matching; however, they ignore

the fact that coarse-grained information also plays an important role in cross-media retrieval;

2. The fine-grained data has been used in some traditional methods. However, due to the limitations of traditional frameworks, these methods cannot effectively deal with the complex correlations between fine-grained data.

Therefore, this paper proposes a dual-framework cross-media model (shown in Figure 2), which contains: (1) a multi-grained data fusion framework based on Deep Belief Network and (2) a multi-modality data fusion framework integrated with the multi-margin triplet loss.

3. Framework I: Multi-grained data fusion framework based on the deep belief network

In this section, we first introduce the strategy of the visual and the textual fine-grained data division and present the framework of the symmetric network and its optimization strategy. In particular, important notations and descriptions of our work are listed in Table 1.

3.1. Image and text fine-grained division strategy

The division strategies for the fine-grained data are also different for various modalities and datasets. For the image modality, we adopt two strategies: (1) Dividing an image into patches evenly and (2) using the selective search algorithm to select image blocks from candidate bounding boxes. There are also two strategies for the text modality: (1) If the textual data are composed of multiple sentences for the dataset, each sentence is regarded as fine-grained data. (2) If the dataset uses tag lists as the textual data, we take each tag of them as the fine-grained data.

Both the coarse-grained and fine-grained image features are extracted from the Convolutional Neural Network, and the dimension of the image feature is 4096, that is, $d_1 = 4096$. In addition, Bag-of-Words (BoWs) features are used as both the coarse-grained and fine-grained features for the text modality. The details of feature representations are illustrated in Section 5.2.

3.2. Multi-grained data fusion symmetric network

The multi-grained data fusion symmetric network has two parts: (1) Cross-modality joint optimization and (2) multi-grained data fusion.

3.2.1. Cross-modality joint optimization

First of all, we introduce the symmetric Gaussian RBM for both the coarse-grained and the fine-grained data of the image modality. The probability assigned to the visible vector \mathbf{v}^I is calculated as follows:

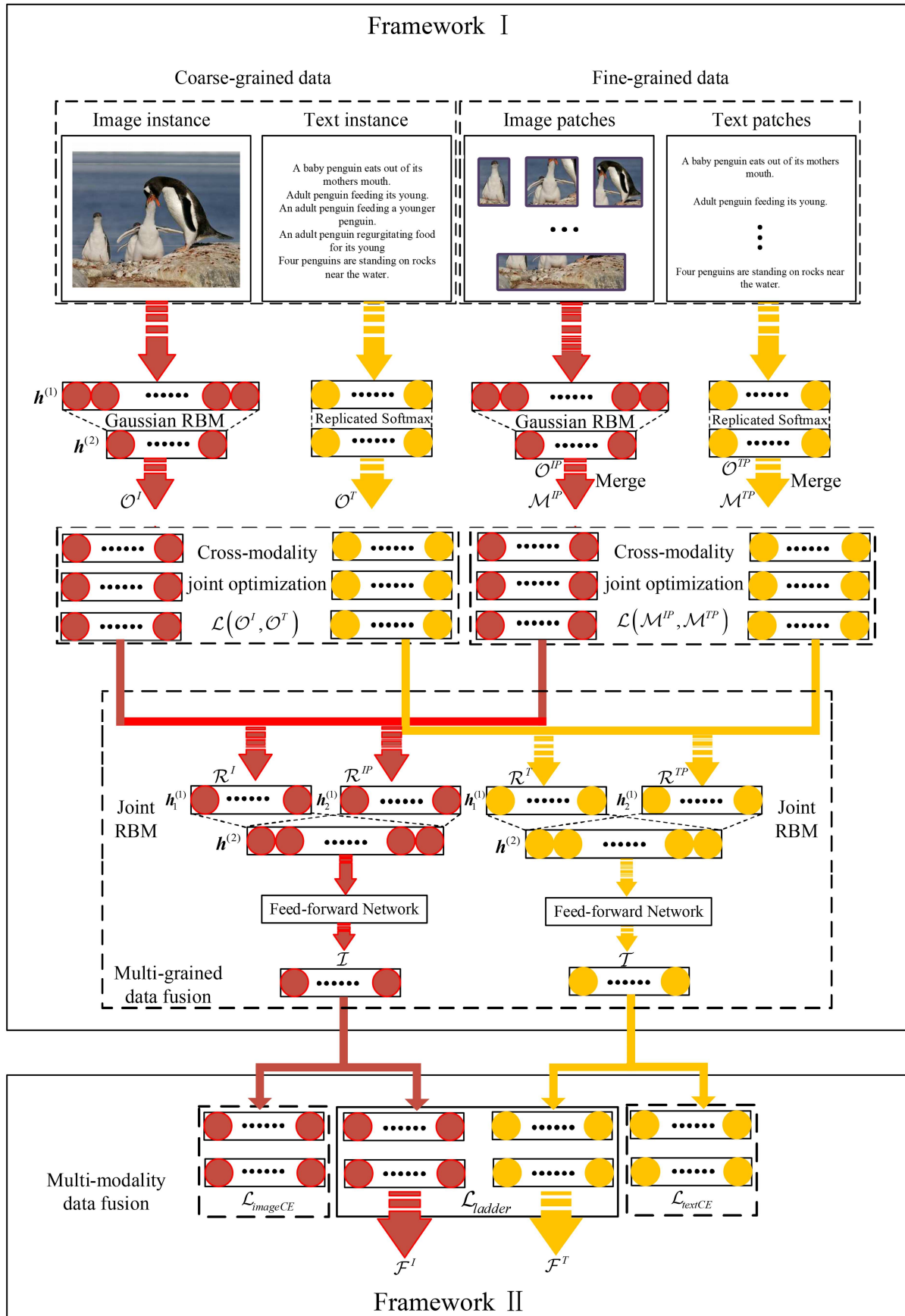


Figure 2. Cross-media retrieval framework via fusing multi-modality and multi-grained data.

Table 1. Important notations and descriptions.

Notation	Description
\mathbf{v}^I	$\mathbf{v}^I \in \mathbb{R}^{d_1}$ is the feature vector of coarse-grained image instance as the input of Framework I
\mathbf{v}^{IP}	$\mathbf{v}^{IP} \in \mathbb{R}^{d_1}$ is the feature vector of fine-grained image patch as the input of Framework I
\mathbf{v}^T	$\mathbf{v}^T \in \mathbb{N}^K$ is the feature vector of coarse-grained text instance as the input of Framework I
\mathbf{v}^{TP}	$\mathbf{v}^{TP} \in \mathbb{N}^K$ is the feature vector of fine-grained text patch as the input of Framework I
\mathcal{O}^I	$\mathcal{O}^I = \{\mathbf{o}_1^I, \mathbf{o}_2^I, \dots, \mathbf{o}_s^I\}$ is the feature set of coarse-grained image instances, which is the outputs from Gaussian DBN
\mathcal{O}^{IP}	$\mathcal{O}^{IP} = \{\mathbf{o}_1^{IP}, \mathbf{o}_2^{IP}, \dots, \mathbf{o}_n^{IP}\}$ is the feature set of fine-grained image patches, which is the outputs from Gaussian DBN
\mathcal{O}^T	$\mathcal{O}^T = \{\mathbf{o}_1^T, \mathbf{o}_2^T, \dots, \mathbf{o}_s^T\}$ is the feature set of coarse-grained text instances, which is the outputs from Replicated Softmax
\mathcal{O}^{TP}	$\mathcal{O}^{TP} = \{\mathbf{o}_1^{TP}, \mathbf{o}_2^{TP}, \dots, \mathbf{o}_u^{TP}\}$ is the feature set of fine-grained text patches, which is the outputs from Replicated Softmax
\mathcal{M}^{IP}	$\mathcal{M}^{IP} = \{\mathbf{m}_1^{IP}, \mathbf{m}_2^{IP}, \dots, \mathbf{m}_s^{IP}\}$ is the feature set of fine-grained image patches after merging
\mathcal{M}^{TP}	$\mathcal{M}^{TP} = \{\mathbf{m}_1^{TP}, \mathbf{m}_2^{TP}, \dots, \mathbf{m}_s^{TP}\}$ is the feature set of fine-grained text patches after merging
\mathcal{R}^I	$\mathcal{R}^I = \{\mathbf{r}_1^I, \mathbf{r}_2^I, \dots, \mathbf{r}_s^I\}$ is the feature set of coarse-grained image instances after joint optimization
\mathcal{R}^T	$\mathcal{R}^T = \{\mathbf{r}_1^T, \mathbf{r}_2^T, \dots, \mathbf{r}_s^T\}$ is the feature set of coarse-grained text instances after joint optimization
\mathcal{R}^{IP}	$\mathcal{R}^{IP} = \{\mathbf{r}_1^{IP}, \mathbf{r}_2^{IP}, \dots, \mathbf{r}_s^{IP}\}$ is the feature set of merged fine-grained image patches after joint optimization
\mathcal{R}^{TP}	$\mathcal{R}^{TP} = \{\mathbf{r}_1^{TP}, \mathbf{r}_2^{TP}, \dots, \mathbf{r}_s^{TP}\}$ is the feature set of merged fine-grained text patches after joint optimization
\mathcal{I}	$\mathcal{I} = \{\mathbf{i}_1, \mathbf{i}_2, \dots, \mathbf{i}_s\}$ is the feature set of the image modality, which is the outputs from Framework I
\mathcal{T}	$\mathcal{T} = \{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_s\}$ is the feature set of the text modality, which is the outputs from Framework I
s	The number of instances.
n	The number of image patches.
u	The number of text patches.
C	The number of semantic categories.
K	The vocabulary size.
\mathcal{F}^I	$\mathcal{F}^I = \{\mathbf{f}_1^I, \mathbf{f}_2^I, \dots, \mathbf{f}_s^I\}$ is the final feature set of the image modality, which is the outputs from Framework II
\mathcal{F}^T	$\mathcal{F}^T = \{\mathbf{f}_1^T, \mathbf{f}_2^T, \dots, \mathbf{f}_s^T\}$ is the final feature set of the text modality, which is the outputs from Framework II

$$\begin{aligned}
P(\mathbf{v}^I) &= \sum_{\mathbf{h}^{(1)}} P(\mathbf{h}^{(1)} \mathbf{v}^I) \\
&= \sum_{\mathbf{h}^{(1)}} P(\mathbf{h}^{(1)}) P(\mathbf{v}^I | \mathbf{h}^{(1)}), \quad (1)
\end{aligned}$$

$$P(\mathbf{h}^{(1)}) = \sum_{\mathbf{h}^{(2)}} P(\mathbf{h}^{(2)}, \mathbf{h}^{(1)}), \quad (2)$$

$$P(\mathbf{v}^I) = \sum_{\mathbf{h}^{(1)}, \mathbf{h}^{(2)}} P(\mathbf{h}^{(2)}, \mathbf{h}^{(1)}) P(\mathbf{v}^I | \mathbf{h}^{(1)}), \quad (3)$$

where $\mathbf{v}^I \in \mathbb{R}^{d_1}$ denotes the feature vector of image instance, $\mathbf{h}^{(1)}$ and $\mathbf{h}^{(2)}$ are represented as two hidden layers of the DBN. Eq. (3) is the probability of the coarse-grained data of the image modality. We can derive Eq. (3) from Eq. (1) and Eq. (2). Similarly, the probability function of the vector \mathbf{v}^{IP} can be obtained:

$$P(\mathbf{v}^{IP}) = \sum_{\mathbf{h}^{(1)}, \mathbf{h}^{(2)}} P(\mathbf{h}^{(2)}, \mathbf{h}^{(1)}) P(\mathbf{v}^{IP} | \mathbf{h}^{(1)}), \quad (4)$$

where $\mathbf{v}^{IP} \in \mathbb{R}^{d_1}$ is the feature vector of fine-grained image patch, and \mathbf{v}^{IP} is used as the input of Framework I.

Next, the symmetric Replicated Softmax networks are used to model the coarse-grained and fine-grained data in text modality. The probability functions are defined as follows:

$$P(\mathbf{v}^T) = \sum_{\mathbf{h}^{(1)}, \mathbf{h}^{(2)}} P(\mathbf{h}^{(2)}, \mathbf{h}^{(1)}) P(\mathbf{v}^T | \mathbf{h}^{(1)}), \quad (5)$$

$$P(\mathbf{v}^{TP}) = \sum_{\mathbf{h}^{(1)}, \mathbf{h}^{(2)}} P(\mathbf{h}^{(2)}, \mathbf{h}^{(1)}) P(\mathbf{v}^{TP} | \mathbf{h}^{(1)}), \quad (6)$$

where $\mathbf{v}^T \in \mathbb{N}^K$ and $\mathbf{v}^{TP} \in \mathbb{N}^K$ refer to coarse-grained and fine-grained BoWs features respectively.

The outputs of the above-mentioned symmetric

Gaussian RBM and the symmetric Replicated Softmax are represented as follows:

$\mathcal{O}^I = \{\mathbf{o}_1^I, \mathbf{o}_2^I, \dots, \mathbf{o}_s^I\}$ is the feature set of coarse-grained image instances, and $\mathcal{O}^{IP} = \{\mathbf{o}_1^{IP}, \mathbf{o}_2^{IP}, \dots, \mathbf{o}_n^{IP}\}$ refers to the feature set of fine-grained image patches. In addition, $\mathcal{O}^T = \{\mathbf{o}_1^T, \mathbf{o}_2^T, \dots, \mathbf{o}_s^T\}$ denotes the feature set of coarse-grained text instances, and $\mathcal{O}^{TP} = \{\mathbf{o}_1^{TP}, \mathbf{o}_2^{TP}, \dots, \mathbf{o}_u^{TP}\}$ represents the feature set of fine-grained text patches.

Next, a three-layers fully connected network is used to reconstruct the coarse-grained features of images \mathcal{O}^I and features \mathcal{O}^T of texts. The reconstructed coarse-grained image feature set $\mathcal{R}^I = \{\mathbf{r}_1^I, \mathbf{r}_2^I, \dots, \mathbf{r}_s^I\}$ and coarse-grained text feature set $\mathcal{R}^T = \{\mathbf{r}_1^T, \mathbf{r}_2^T, \dots, \mathbf{r}_s^T\}$ are generated by minimizing the loss function in Eq. (7):

$$\mathcal{L}(\mathcal{O}^I, \mathcal{O}^T) = \|\mathcal{O}^I - \mathcal{R}^I\|^2 + \|\mathcal{O}^T - \mathcal{R}^T\|^2 + \|\mathcal{O}^I - \mathcal{O}^T\|^2, \quad (7)$$

where $\|\cdot\|^2$ refers to the squared Euclidean distance. For both the intra-modality and inter-modality, the loss function $\mathcal{L}(\mathcal{O}^I, \mathcal{O}^T)$ aims to minimize the distances between before-reconstructed features and reconstructed features as small as possible.

The operation of “merging” on \mathcal{O}^{IP} and \mathcal{O}^{TP} to get fine-grained image feature set $\mathcal{M}^{IP} = \{\mathbf{m}_1^{IP}, \mathbf{m}_2^{IP}, \dots, \mathbf{m}_s^{IP}\}$ and fine-grained text feature set $\mathcal{M}^{TP} = \{\mathbf{m}_1^{TP}, \mathbf{m}_2^{TP}, \dots, \mathbf{m}_s^{TP}\}$ are defined as follows:

$$\mathbf{m}_q^{IP} = \frac{1}{N_q^{IP}} \sum_{k=1}^{N_q^{IP}} \mathbf{o}_k^{IP}, q \in \{1, \dots, s\},$$

$$\mathbf{m}_q^{TP} = \frac{1}{N_q^{TP}} \sum_{k=1}^{N_q^{TP}} \mathbf{o}_k^{TP}, q \in \{1, \dots, s\}, \quad (8)$$

where \mathbf{o}_k^{IP} and \mathbf{o}_k^{TP} are the outputs of the Gaussian RBM and the Replicated Softmax respectively. N_q^{IP} and N_q^{TP} are the number of patches in the q^{th} image or text respectively. Afterward, the joint optimization function for fine-grained features is denoted as:

$$\mathcal{L}(\mathcal{M}^{TP}, \mathcal{M}^{IP}) = \|\mathcal{M}^{IP} - \mathcal{R}^{IP}\|^2 + \|\mathcal{M}^{TP} - \mathcal{R}^{TP}\|^2 + \|\mathcal{M}^{IP} - \mathcal{M}^{TP}\|^2, \quad (9)$$

where $\mathcal{R}^{IP} = \{\mathbf{r}_1^{IP}, \mathbf{r}_2^{IP}, \dots, \mathbf{r}_s^{IP}\}$ and $\mathcal{R}^{TP} = \{\mathbf{r}_1^{TP}, \mathbf{r}_2^{TP}, \dots, \mathbf{r}_s^{TP}\}$ refer to the reconstructed features of image patches and text patches by the joint optimization. The purpose of the loss function in Eq. (9) is similar to the loss function in Eq. (7).

3.2.2. Multi-grained data fusion

The distributions of multi-grained joint RBM for the

image modality and the text modality are defined as follows:

$$P(\mathbf{r}^I, \mathbf{r}^{IP}) = \sum_{\mathbf{h}_1^{(1)}, \mathbf{h}_2^{(1)}, \mathbf{h}^{(2)}} P(\mathbf{h}_1^{(1)}, \mathbf{h}_2^{(1)}, \mathbf{h}^{(2)}) \times \sum_{\mathbf{h}_1^{(1)}} P(\mathbf{r}^I | \mathbf{h}_1^{(1)}) \times \sum_{\mathbf{h}_2^{(1)}} P(\mathbf{r}^{IP} | \mathbf{h}_2^{(1)}), \quad (10)$$

$$P(\mathbf{r}^T, \mathbf{r}^{TP}) = \sum_{\mathbf{h}_1^{(1)}, \mathbf{h}_2^{(1)}, \mathbf{h}^{(2)}} P(\mathbf{h}_1^{(1)}, \mathbf{h}_2^{(1)}, \mathbf{h}^{(2)}) \times \sum_{\mathbf{h}_1^{(1)}} P(\mathbf{r}^T | \mathbf{h}_1^{(1)}) \times \sum_{\mathbf{h}_2^{(1)}} P(\mathbf{r}^{TP} | \mathbf{h}_2^{(1)}), \quad (11)$$

where $\mathbf{r}^I, \mathbf{r}^{IP}, \mathbf{r}^T, \mathbf{r}^{TP}$ are the outputs of the previous network, which are regarded as the inputs for the joint RBM. $\mathbf{h}_1^{(1)}$ and $\mathbf{h}_2^{(1)}$ refer to the first hidden layers that connect to the coarse-grained and fine-grained features. $\mathbf{h}^{(2)}$ denotes the second layer that connects to the $\mathbf{h}_1^{(1)}$ and $\mathbf{h}_2^{(1)}$. Then the outputs of joint RBM are connected to the three-layer feed-forward network. Finally, the coarse-grained and fine-grained information are fused together to obtain the image features $\mathcal{I} = \{i_1, i_2, \dots, i_s\}$ and $\mathcal{T} = \{t_1, t_2, \dots, t_s\}$.

4. Framework II: Multi-modality data fusion framework based on the multi-margin triplet loss function

In this section, we propose a multi-margin triplet loss function to learn a better common feature representation by constraining inter-modality relevance.

4.1. Inter-modality data correlation constraint

As shown in Figure 3, triplet loss [32] cannot accurately describe the similarity between the anchor and irrelevant samples, leading to poor performance. To solve this problem and inspired by the idea of [33], we propose a modified **Multi-margin triplet loss** (Figure 4) in which we innovatively propose the concept of **Margin-set** (shown in Definition 1). The margin-set makes the irrelevant candidates get away from the anchor with different distances according to their relevance degrees. Unlike the values of margins changing with a chain way in [33], we use the Euclidean distance to evaluate the relationships between samples, and the margins are monotonically increasing.

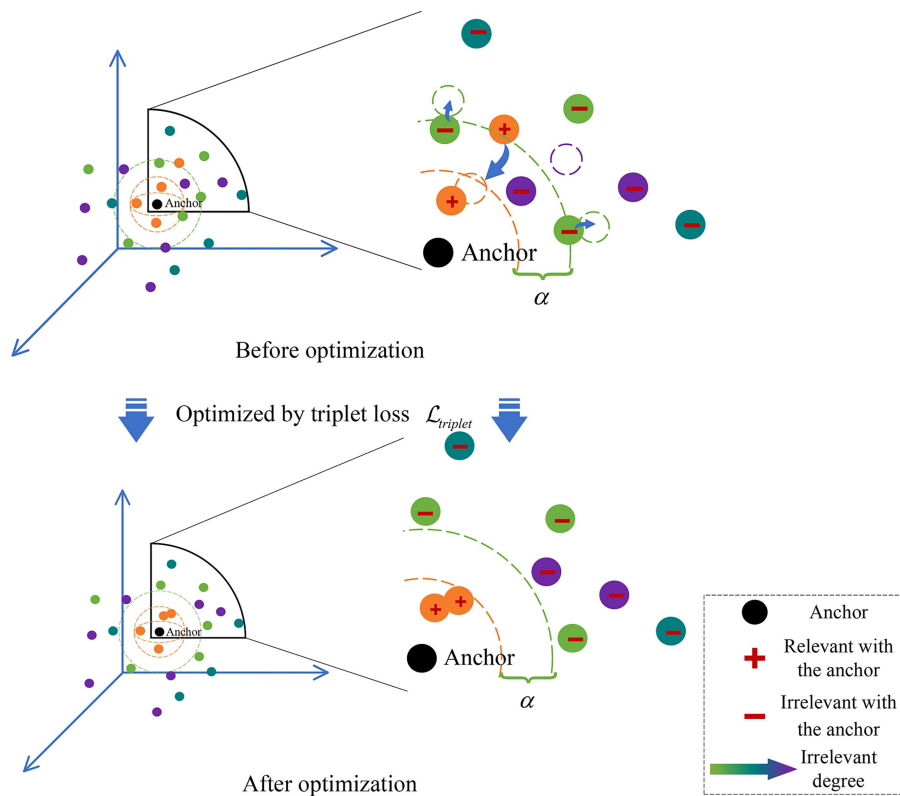


Figure 3. Illustration of the optimization result of the triplet loss function.

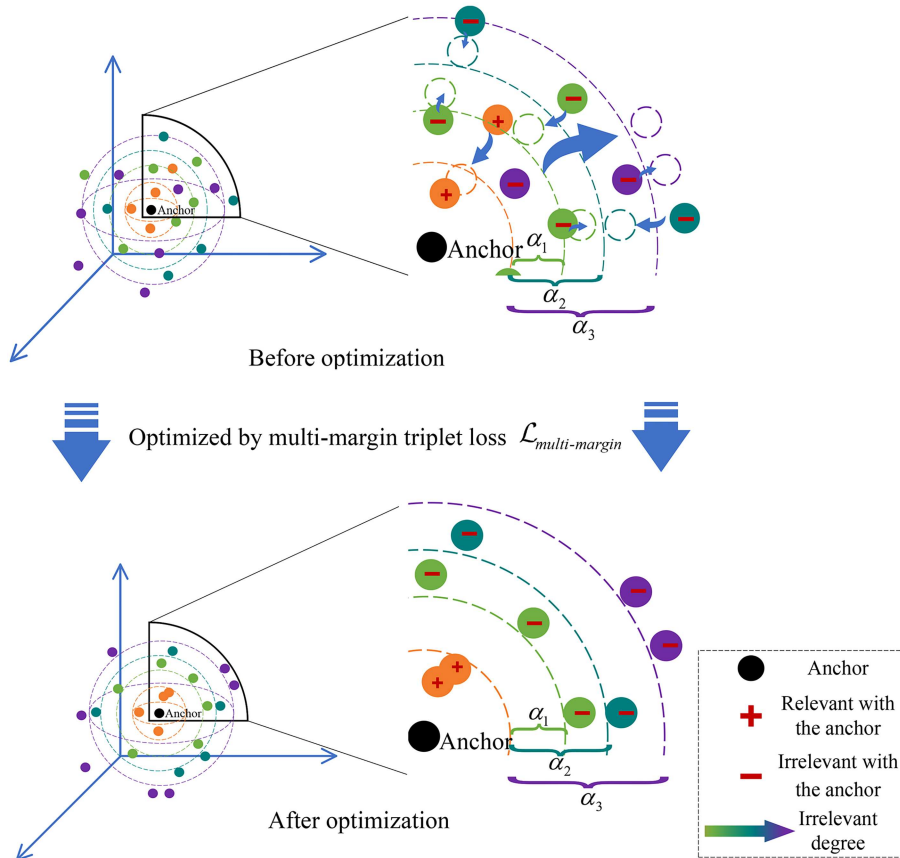


Figure 4. Illustration of the optimization results of the multi-margin triplet loss function.

Definition 1. Margin-set: It is a series of pre-defined margins between relevant and irrelevant samples, and the values of the elements in the margin-set are ascending, that is, Margin-set = $\{\alpha_1, \alpha_2, \dots, \alpha_w\}$. The elements should satisfy the condition of $\alpha_1 < \alpha_2 < \dots < \alpha_w$.

In Figure 4, we set $w = 3$ as an example to illustrate the effect of feature optimization results with our proposed Multi-margin triplet loss. For the optimization of multi-modality features, the principle of multi-margin triplet loss is as follows:

$$\alpha_1 + \|f(i^a) - g(t^+)\|^2 < \|f(i^a) - g(t^-)\|_{\alpha_1}^2$$

$$\dots$$

$$\alpha_w + \|f(i^a) - g(t^+)\|^2 < \|f(i^a) - g(t^-)\|_{\alpha_w}^2, \quad (12)$$

$$\alpha_1 + \|g(t^a) - f(i^+)\|^2 < \|g(t^a) - f(i^-)\|_{\alpha_1}^2$$

$$\dots$$

$$\alpha_w + \|g(t^a) - f(i^+)\|^2 < \|g(t^a) - f(i^-)\|_{\alpha_w}^2, \quad (13)$$

where i is the original image features, and t denotes the original text features. $f(\cdot)$ and $g(\cdot)$ represent nonlinear mapping features for the image and text modalities, respectively. $\|\cdot\|_{\alpha}^2$ refers to squared Euclidean distance that satisfies the selected α condition in Eq. (15). Particularly, i^+ and t^+ are relevant samples, while i^- and t^- refer to irrelevant samples.

Multi-margin triplet loss picks different margins according to the distances between the anchor and irrelevant samples. Consequently, the value of α is determined by the squared distance normalization Eq. (14) and piece-wise function Eq. (15):

$$dis_{norm}(i_p, t_q) = \frac{\|f(i_p) - g(t_q)\|^2 - \text{Min}(\|f(i) - g(t)\|^2)}{\text{Max}(\|f(i) - g(t)\|^2) - \text{Min}(\|f(i) - g(t)\|^2)}, \quad (14)$$

where we use one batch of data (the batch size in the experiment is set to 80) to calculate $\|f(i_p) - g(t_q)\|^2$ ($\forall p, q \in [1, 80]$), then normalize the value in $[0, 1]$. According to the $dis_{norm}(i_p, t_q)$, we further divide multiple sub-ranges. Consequently, the value of α is defined as follows:

$$\alpha = \begin{cases} \alpha_1, 0 \leq dis_{norm}(i_p, t_q) \leq e_1 \\ \alpha_2, e_1 < dis_{norm}(i_p, t_q) \leq e_2 \\ \dots \\ \alpha_w, e_{w-1} < dis_{norm}(i_p, t_q) \leq 1 \end{cases} \quad (15)$$

where e_1, \dots, e_{w-1} are piece-wise parameters, that are used to divide the distance range between image and text features into $[0, e_1], (e_1, e_2], \dots, (e_{w-1}, 1]$. Following this, the multi-margin triplet loss is defined as:

$$\mathcal{L}_{multi-margin} = \frac{1}{N} \left\{ \sum_{N_1} \left[l_{image}^{multi-margin^1}(i^a, t^+, t^-) + l_{text}^{multi-margin^1}(t^a, i^+, i^-) \right] + \dots + \sum_{N_w} \left[l_{image}^{multi-margin^w}(i^a, t^+, t^-) + l_{text}^{multi-margin^w}(t^a, i^+, i^-) \right] \right\}. \quad (16)$$

Especially, $\mathcal{L}_{multi-margin}$ contains w similar terms. N_1, \dots, N_w are the number of triplet pairs that satisfy w margins in Eq. (15), and $N = N_1 + N_2 + \dots + N_w$.

$$l_{image}^{multi-margin^k}(i^a, t^+, t^-) = \text{Max} \left(0, \alpha_k + \|f(i^a) - g(t^+)\|^2 - \|f(i^a) - g(t^-)\|^2 \right), \quad (17)$$

where $k(k \in [1, w])$ is a constant. The anchor in Eq. (17) refers to the sample i^a of the image modality. In this paper, i^+ and i^a are regarded as relevant samples when they belong to the same semantic category. Moreover, i^- and i^a are not in the same semantic category and $dis_{norm}(i^a, t^-) \in (e_{k-1}, e_k]$, hence, the value of the margin is set to α_k . Similar to Eq. (17), we define the loss function that the anchor is from the text modality as follows:

$$l_{text}^{multi-margin^k}(t^a, i^+, i^-) = \text{Max} \left(0, \alpha_k + \|g(t^a) - f(i^+)\|^2 - \|g(t^a) - f(i^-)\|^2 \right), \quad (18)$$

where the anchor t^a is a text, t^a and i^+ belong to the same semantic category, and t^a and irrelevant sample i^- belong to the different semantic categories. The value of the margin is set to α_k according to $dis_{norm}(i^-, t^a) \in (e_{k-1}, e_k]$.

Following this, we can calculate the partial derivations of loss function Eq. (17) and Eq. (18) with respect to $f(i^a)$ and $g(t^a)$, and the derivations are calculated as follows:

$$\frac{\partial l_{image}^{multi-margin^k}}{\partial f(i^a)} = 2[f(i^a) - g(t^+)] - 2[f(i^a) - g(t^-)], \quad (19)$$

$$\frac{\partial l_{text}^{multi-margin^k}}{\partial g(t^a)} = 2 [g(t^a) - f(i^+)] - 2 [g(t^a) - f(i^-)]. \quad (20)$$

4.2. Intra-modality semantic category constraint

For the intra-modality semantic constraint, we minimize the cross-entropy loss functions for two modalities as follows:

$$\begin{aligned} \mathcal{L}_{imageCE} &= -\frac{1}{B} \sum_{s=1}^B \sum_{j=1}^C y_{js}^i \log(\hat{y}_{js}^i), \\ \mathcal{L}_{textCE} &= -\frac{1}{B} \sum_{s=1}^B \sum_{j=1}^C y_{js}^t \log(\hat{y}_{js}^t), \end{aligned} \quad (21)$$

where C is the number of semantic categories, \hat{y}_{js}^i and \hat{y}_{js}^t are the probability distributions after the Softmax functions, and y_{js}^i and y_{js}^t are ground truth semantic category probability distributions. B is the batch size ($B = 80$ in our experiment). Ultimately, the total loss function is defined as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{multi-margin} + \lambda_2 \mathcal{L}_{imageCE} + \lambda_2 \mathcal{L}_{textCE}, \quad (22)$$

where λ_1 and λ_2 are balance parameters, $\mathcal{L}_{multi-margin}$ is obtained by Eq. (16), and $\mathcal{L}_{imageCE}$ and \mathcal{L}_{textCE} are denoted in Eq. (21). After optimizing with the loss function \mathcal{L} , the outputs of the network are the final feature representations, that are $\mathcal{F}^I = \{\mathbf{f}_1^I, \mathbf{f}_2^I, \dots, \mathbf{f}_s^I\}$ and $\mathcal{F}^T = \{\mathbf{f}_1^T, \mathbf{f}_2^T, \dots, \mathbf{f}_s^T\}$.

5. Experiment

In this section, we conduct comprehensive experiments to verify the effectiveness of our proposed framework on three widely used datasets. Several state-of-the-art methods are selected to compare with our method.

5.1. Details of the network

In Framework I, Gaussian RBM has 2048 dimensions in the first hidden layer $\mathbf{h}^{(1)}$ and 1024 dimensions in the second hidden layer $\mathbf{h}^{(2)}$. Replicated Softmax has 1024 dimensions in both layers. Then, the dimension of the three cross-modality joint optimization layers is 1024. Next, the dimensions of joint RBM are 1024 and 2048 for $\mathbf{h}^{(1)}$ and $\mathbf{h}^{(2)}$, respectively. The three-layer feed-forward network on the top of the joint RBM has 1024 dimensions, and the soft-max loss is used for optimization. In Framework II, the dimension number of the three-layer fully-connected network is 1024.

5.2. Datasets and feature representations

In this paper, we have carefully selected three highly regarded datasets for our study. Furthermore, we have conducted a comprehensive analysis of the feature representations used for these datasets.

- **Pascal Sentence:** It contains 1,000 images, which are equally divided into 20 semantic categories. Each image has five matching sentences, and these sentences form a document. This dataset is split into three subsets: 800 for training, 100 for validation, and 100 for testing.
- **Corel 5K:** It contains a total number of 5,000 images, which covers 50 semantic categories. Each semantic category contains 100 images, the number of tags in the dictionary is 260, and each image has 1–5 tags. We execute data cleaning and remove eight images without tags, then further divide this dataset into three parts: 4493 for training, 250 for validation, and 249 for testing.
- **NUS-WIDE 5K:** It consists of 269,648 images and their corresponding tags collected from Flickr, with a total number of 5,018 unique tags, 81 concepts, and the number of tags in the dictionary is 560. NUS-WIDE 5K is a subset of NUS-WIDE; it has 4996 image/text pairs selected from 6 concepts: animal, clouds, person, sky, water, and window. The dataset is split into three subsets: 4,500 pairs for training, 248 for validation, and 248 for testing. For NUS-WIDE 5K, each image/text pair only belongs to a single concept.

A selective search algorithm is used to generate candidate bounding boxes for each image in Pascal Sentence. If the value of IoU of two areas is larger than 0.7, we pick the bigger bounding box region as selected fine-grained image data. The coarse-grained and fine-grained image features of the Pascal Sentence are output from FC7 of the VGG network, which are regarded as the inputs of Framework I. Each image of Corel 5K and NUS-WIDE 5K is evenly divided into 3*3 patches. We use the 4096-dimension features that output from FC7 of Alexnet as the coarse-grained and fine-grained image features.

For the text modality, the coarse-grained textual data of the Pascal Sentence is a document composed of five sentences, each regarded as fine-grained textual data. The coarse-grained textual data of the Corel 5K and NUS-WIDE 5K are tag lists, and the fine-grained textual data is a single tag. All the coarse-grained and fine-grained textual data use the BoWs model to represent text features.

5.3. Evaluation metrics

We employ three different metrics to evaluate performance from various perspectives, aiming to obtain a more objective performance analysis:

- **MAP.** Is used to calculate the Mean Average Precision (MAP) of all queries. The larger the value of MAP, the better the performance.

- **PR curve.** Shows the changing of retrieval precision at all recall values. The larger the curve encloses the area, the better the performance.
- **Top N -precision curve.** Is defined as the average precision ratio according to the top N retrieval items. In our experiments, N is set to 50.
- **Recall@ K .** Is the fraction of queries for which the correct items are retrieved in the top K positions of the ranking list. The larger the value of Recall@ K is, the better performance it represents.

5.4. Parameters analysis

We analyze the influence of different parameters on each dataset. Particularly, $I \rightarrow T$ denotes that images query texts, and $T \rightarrow I$ means the reverse retrieval direction.

5.4.1. Influence of λ_1 and λ_2 on MAP

As shown in Figure 5, we set λ_1 and λ_2 ranging from 0.01 to 100, and experiment results reveal that three datasets are not very sensitive to λ_1 and λ_2 . MAP values fluctuate around 0.02; the best performance is achieved when $\lambda_1 = 1$ and $\lambda_2 = 1$, which indicates that multi-margin triplet loss and cross-entropy loss play equally important roles in optimization. In the following part of the parameters analysis, we set $\lambda_1 = 1$ and $\lambda_2 = 1$.

5.4.2. Influence of margin-set values on MAP

We design three types of strategies: (1) the small step increment, (2) the large step increment, and (3) combining (1) and (2). The values of the small step increment and the large step increment for different datasets are shown in Table 2.

Experimental results on different datasets are provided as follows:

- **Experimental results on Pascal sentence.** Figure 6(a) is the experimental results of the small step increment on Pascal Sentence. While $\alpha = 0.1$ (the first 4 bars), although the average of MAP improves with the increasing of margin values, we unexpectedly find that the trends of MAP changing for $I \rightarrow T$ and $T \rightarrow I$ are not consistent. As shown in Figure 6(a), in the small-step increment strategy, the MAP reaches its peak value when $\{\alpha_1 = 0.1, \alpha_1 = 0.5, \alpha_1 = 0.7\}$. After that, if we increase margin values, the MAP decreases. This result reveals there is an upper bound to move irrelevant data away from the anchor feature. Once a certain upper bound is exceeded, the MAP will decrease.

Figure 6(b) shows the results of the large step increment. It can be easily observed that when we fix α_1 and α_2 and only change α_3 ; the most irrelevant features cannot be too far away from the anchor feature. If α_1 and α_3 are fixed and only change α_2 , MAP increases. The best performance is achieved when parameters are $\{\alpha_1 = 1, \alpha_2 = 3, \alpha_3 = 7\}$. In general, it seems that Pascal Sentence can be affected by the values of the margin and the increment of step; MAP fluctuates between 0.55–0.59. Although the multi-margin triplet loss improves the MAP of Pascal Sentence, this dataset shows low sensitivity with the margins changing.

- **Experimental results on Corel 5K.** Compared with the Pascal Sentence dataset, Corel 5K has more semantic categories and a larger data scale. After doing extensive experiments, we find that when $\alpha \in (0, 1)$ or $\alpha \in (10, +\infty)$, the performance on Corel 5K

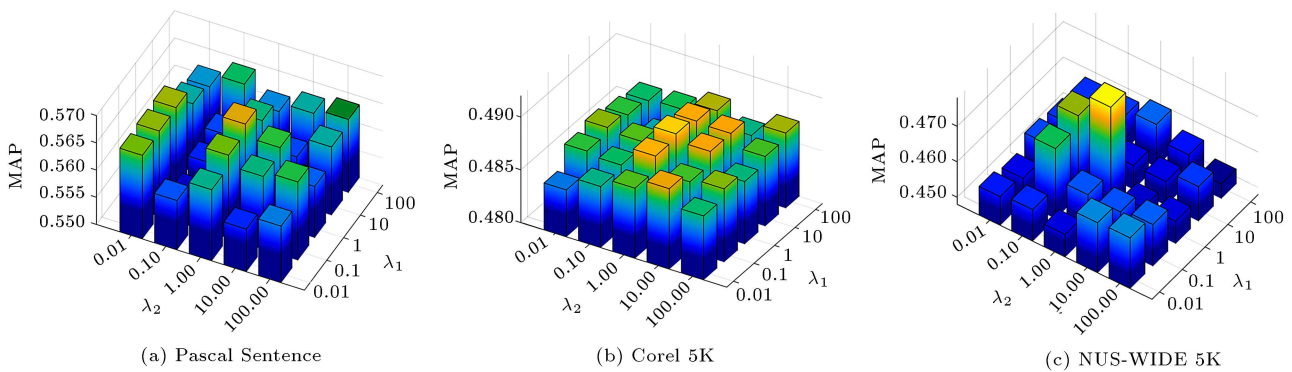


Figure 5. Influence of λ_1 and λ_2 on different datasets.

Table 2. Parameter settings of small step increment and the large step increment for different datasets.

Dataset	Small step increment	Large step increment
Pascal Sentence	<1	≥ 1 and <10
Corel 5K	0.5	1 or 1.5
NUS-WIDE 5K	0.1	0.5

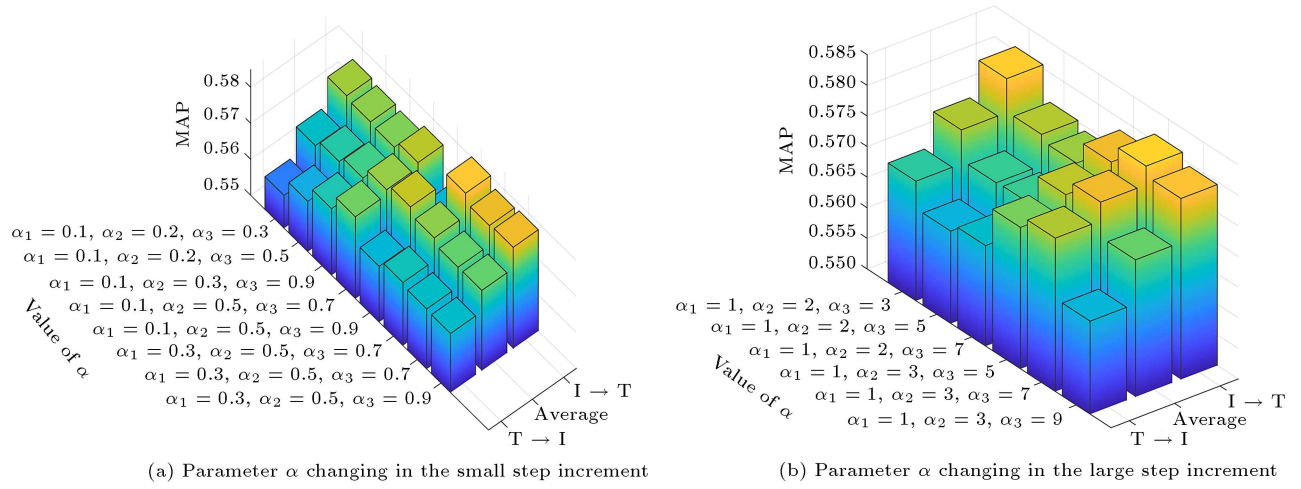
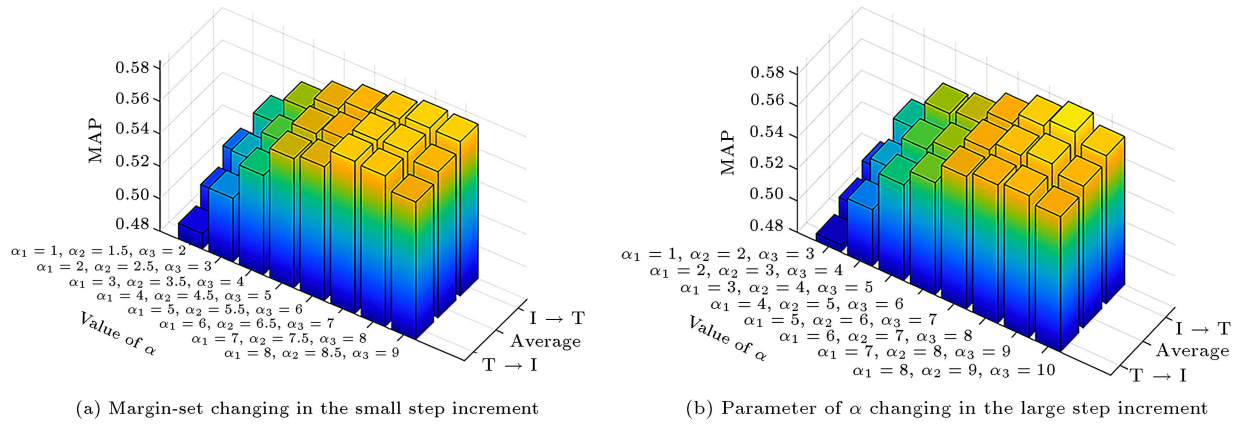
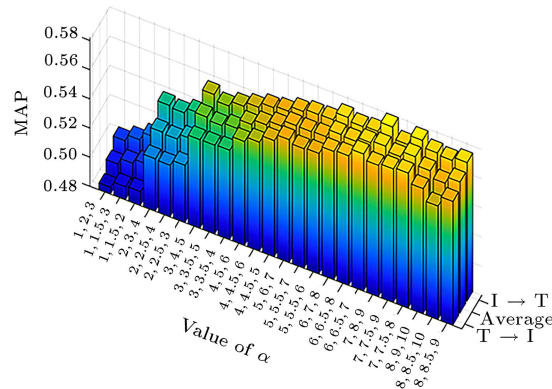


Figure 6. Influence of margin-set values for Pascal Sentence.



(a) Margin-set changing in the small step increment (b) Parameter of α changing in the large step increment



(c) Parameter of α changing that combines the small and the large step increment

Figure 7. Influence of margin-set values for Corel 5K.

is poor. Therefore, in the following experiment analysis, margins are in $[1, 10]$, the small step increment is 0.5, and the large step increment is 1. Next, we observe the results of combining the small step and the large step increment. We try to fix two margin values and see the effect of one changing margin.

As shown in Figure 7(a), MAP rises when the values of margin increase, and the trends of $I \rightarrow$

T and $T \rightarrow I$ are consistent. When the margin-set is set to $\{\alpha_1 = 7, \alpha_2 = 7.5, \alpha_3 = 8\}$, MAP reaches the peak and then decreases, which indicates that the irrelevant features cannot get away from the anchor feature too much. As for the large step increment shown in Figure 7(b), we can get the same conclusion, and the maximum MAP is obtained when the parameters are $\{\alpha_1 = 7, \alpha_2 = 8, \alpha_3 = 9\}$.

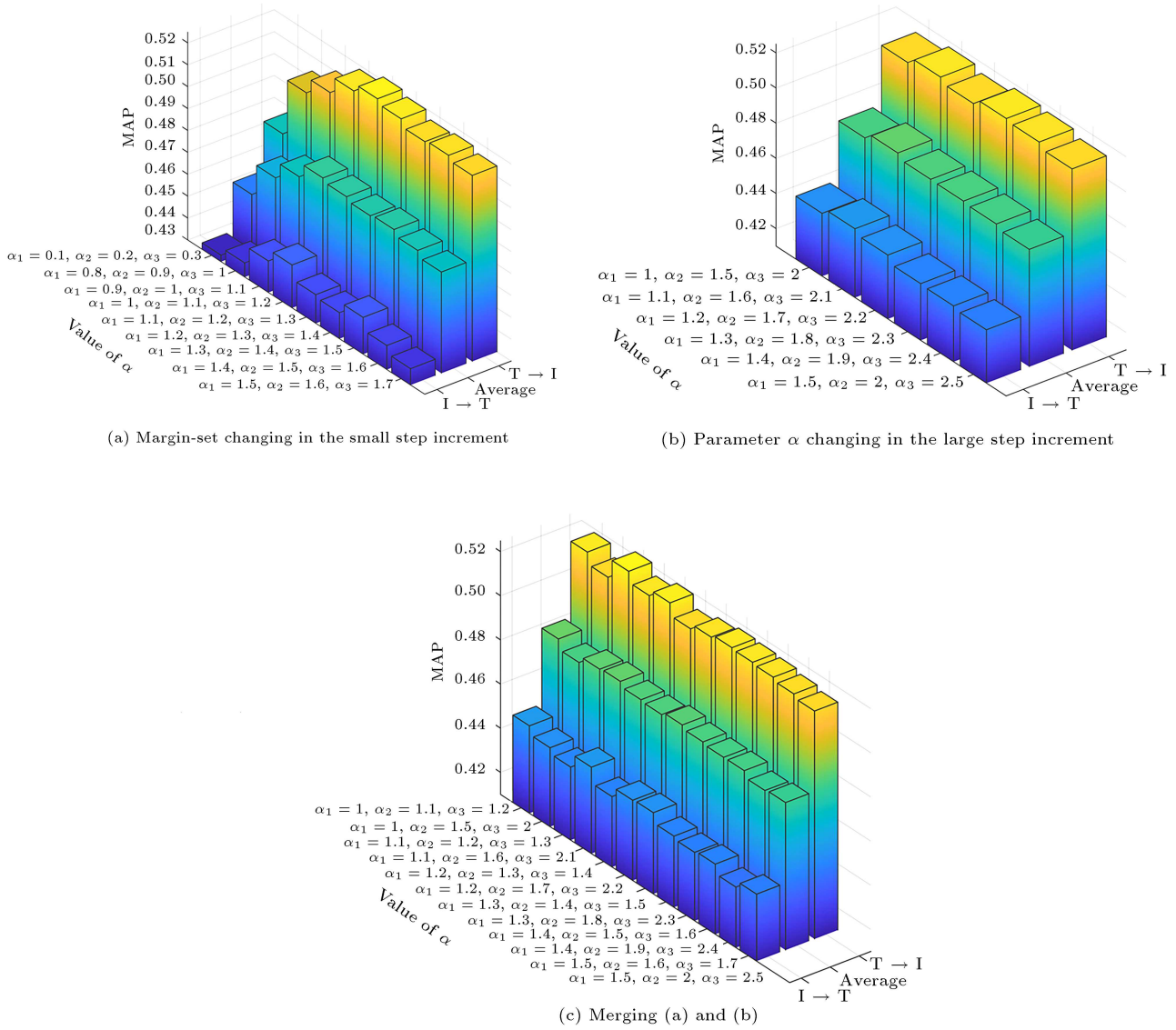


Figure 8. Influence of margin-set values for NUS-WIDE 5K.

To further explore the influence of each margin, we only change one margin parameter at a time. In Figure 7(c), for example, the value of $\alpha \{1, 2, 3\}$ refers to $\{\alpha_1 = 1, \alpha_2 = 2, \alpha_3 = 3\}$. If the α_1 and α_2 are fixed, the small step increment (0.5) between α_2 and α_3 performs better than the large step increment (1.5). If we fix α_1 and α_3 , the small step increment between α_1 and α_2 also gets better performance, which accords with early observation that the small step increment performs better.

Corel 5K is more sensitive to the variations of the margin parameters, and MAP fluctuates in a wide range between 0.48–0.59. Corel 5K has a large amount of data to train this model, and abundant semantic categories are useful to the constraint of multi-margin triplet loss, which are the reasons for the high sensitivity to the margin parameters. This proves that our proposed multi-margin triplet loss

has impressive effectiveness on Corel 5K.

Experimental results on NUS-WIDE 5K.

Figure 8(a)–(c) shows the experimental results of different margin sets on NUS-WIDE 5K. We limit the margin values and set the small step increment as 0.1 (shown in Figure 8(a)). In addition, we set the large step increment as 0.5 (shown in Figure 8(b)). To better compare the effect of the two strategies, we merge the above-mentioned experimental results in Figure 8(c).

In Figure 8(a), if the margins are too small like $\{\alpha_1 = 0.1, \alpha_2 = 0.2, \alpha_3 = 0.3\}$, the MAP is low. The highest MAP is obtained when $\{\alpha_1 = 1, \alpha_2 = 1.1, \alpha_3 = 1.2\}$. In Figure 8(b), the results of the large step strategy perform smoothly. Figure 8(c) shows that when we only fix α_1 , most of the small step increment margin-sets have higher scores

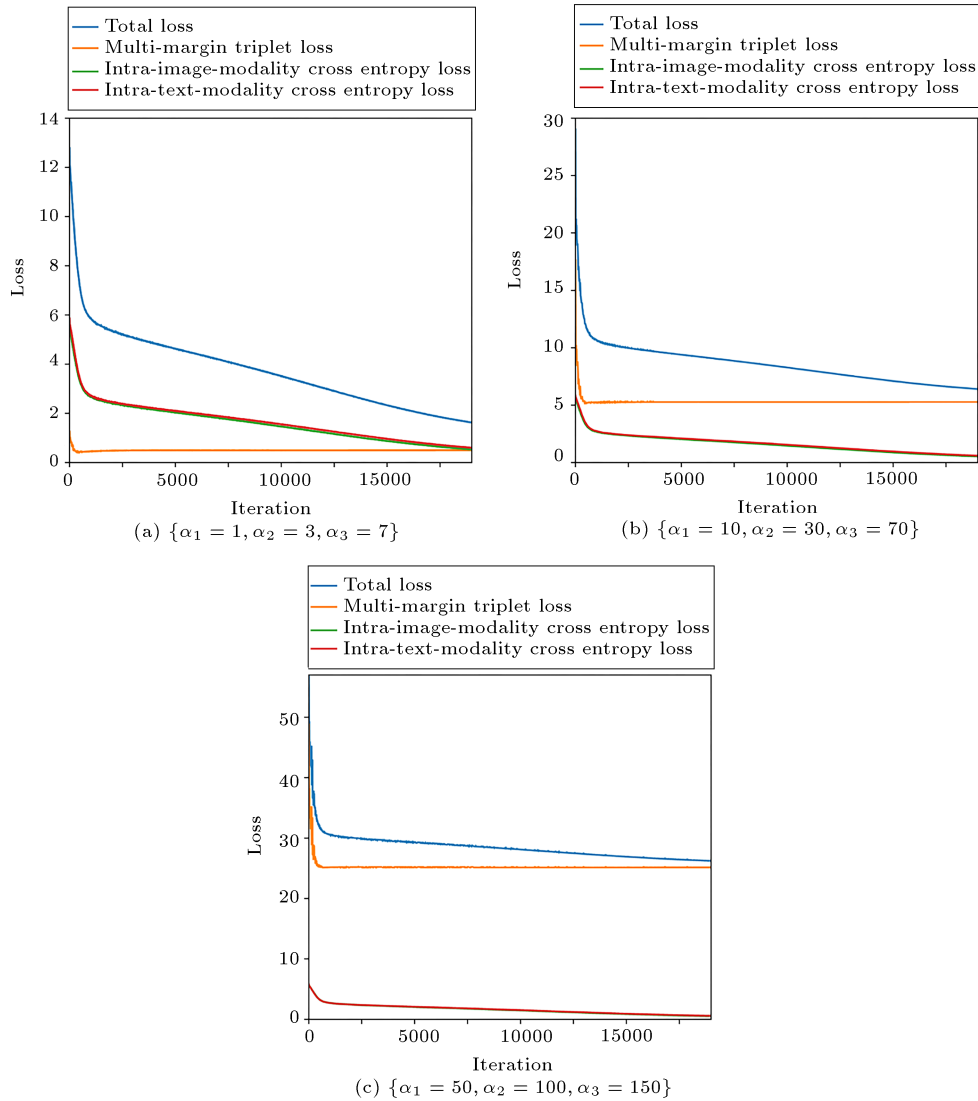


Figure 9. Influence of margin-set on loss function convergence for Pascal Sentence.

of MAP than the large step ones.

Different from the aforementioned two datasets, all the results in Figure 8(a)–(c) have an interesting phenomenon, that is, $T \rightarrow I$ performs much better than $I \rightarrow T$. This is because the image modality data of NUS-WIDE 5K contains more information than modality data, which makes it hard to retrieve precise textual results.

5.4.3. Influence of margin-set on loss function convergence

In this part, we fix $\lambda_1 = 1$ and $\lambda_2 = 1$, and we further study the influence of margin-set. The convergence of different datasets is shown in Figures 9–11.

From Figure 9(a)–(c), we can observe that when the value of the margin-set increases, the iteration times of the loss function increase. Besides, the final convergence value of the total loss function basically relies on the value of the multi-margin triplet loss.

Although Figure 9(c) takes an extreme example, the loss function converges at around 30.

However, Figure 10(a)–(c) is sensitive to the values of the margin-set. In Figure 10(c), the set of parameters with the highest MAP value reveals great volatility when the iteration times increase.

Due to the large initial values in Figure 11(a)–(c), we intercept part of the loss values at the 0 iteration, which are 282.779, 286.792, and 306.497 for these three parameter combinations. Then, we find the loss values decrease fast, and the curves show periodic fluctuations.

5.5. Experimental analysis

The following experiment analyses are based on the optimal parameter combination obtained from Section 5.4. In our experiments, the basic configurations of the devices are as follows: the CPU is Intel Core i5-9400f, the GPU is Nvidia GTX 2080, the graphic

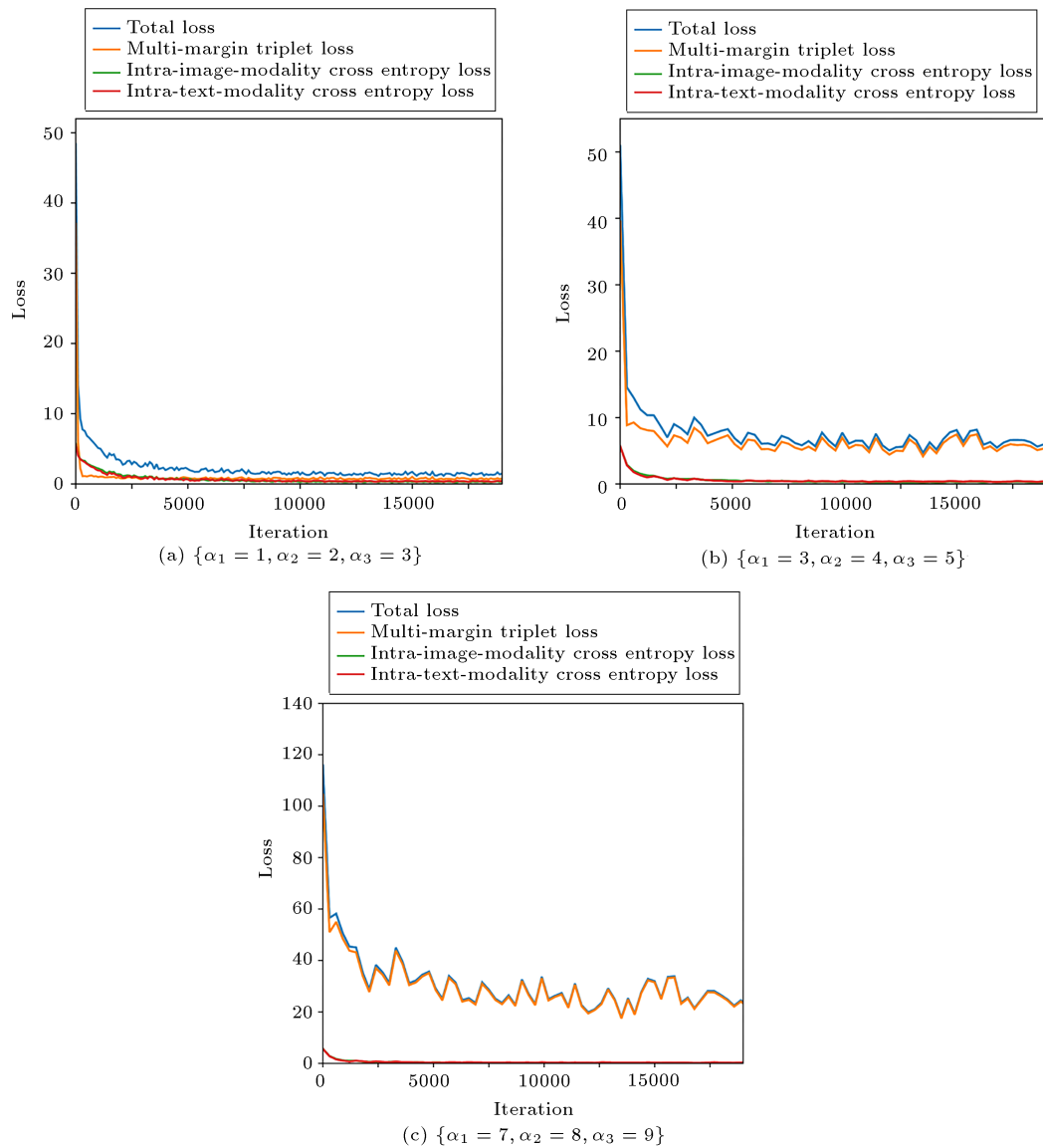


Figure 10. Influence of margin-set on loss function convergence for Corel 5K.

Table 3. Training time of different datasets.

Datasets	Framework I	Framework II	
		10000 iterations	20000 iterations
Pascal Sentence	3587 s	77 s	96 s
Corel 5K	2919 s	143 s	168 s
NUS-WIDE 5K	6261 s	80 s	132 s

memory is 8 G, and the RAM is 48 G. The training times of two frameworks for different datasets are shown in Table 3.

5.5.1. Comparison with the benchmarks and the state-of-the-art methods

In the following, different evaluation metrics are used to make performance comparison:

(1) MAP analysis

As Pascal Sentence has an uncertain number of patches for each image, the bag-of-words features of Corel 5K and NUS-WIDE 5K are sparse, some methods that cannot be compared with us. From Table 4, we can draw the following conclusions:

1. Based on CCA [3], SM [4] uses the semantic cate-

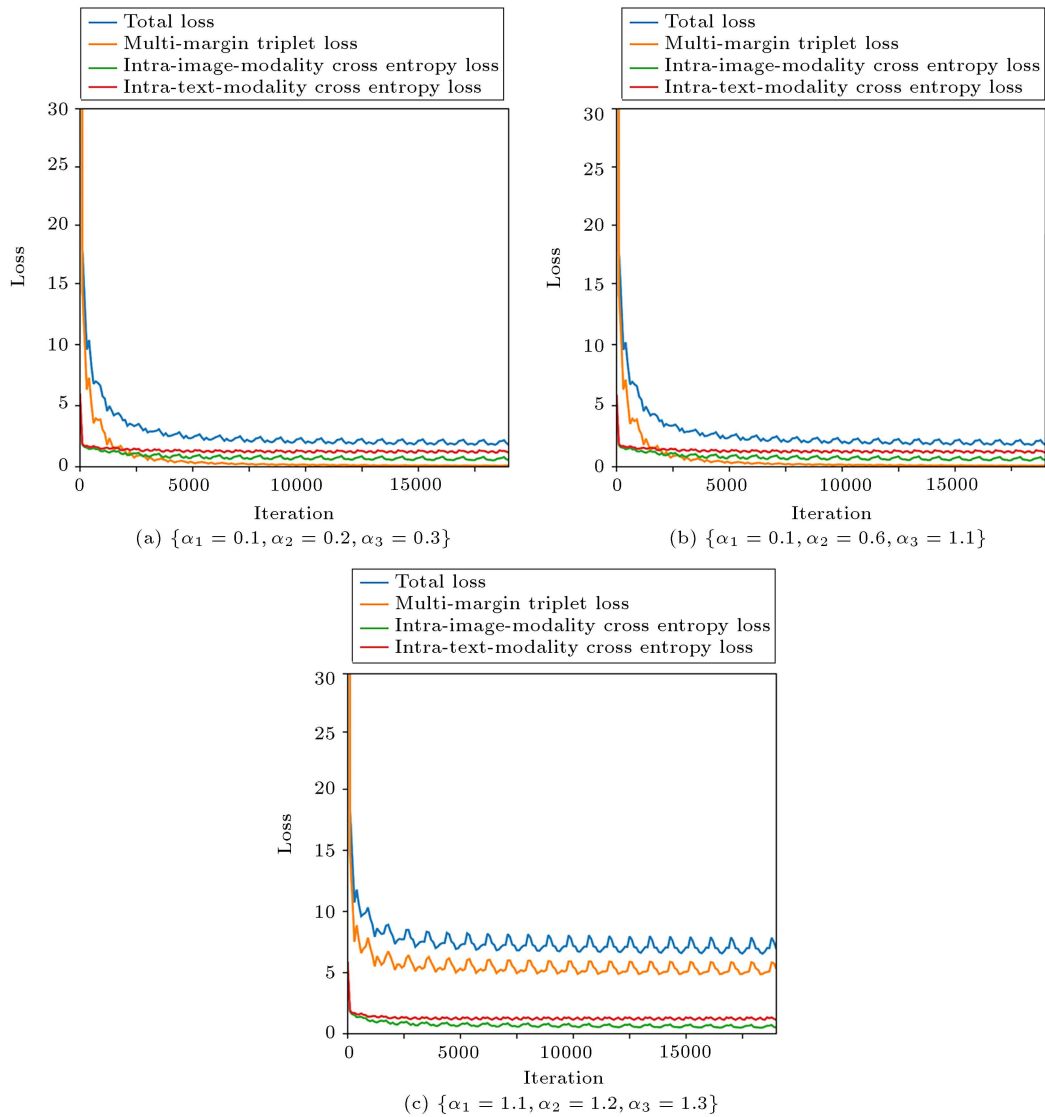


Figure 11. Influence of margin-set on loss function convergence for NUS-WIDE 5K.

Table 4. MAP of our method and the compared methods on three datasets.

Compared methods	Pascal Sentence			Corel 5K			NUS-WIDE 5K		
	I \rightarrow T	T \rightarrow I	Average	I \rightarrow T	T \rightarrow I	Average	I \rightarrow T	T \rightarrow I	Average
CCA	0.0928	0.0888	0.0908	0.1006	0.1129	0.1067	0.2146	0.2136	0.2141
SM	0.4401	0.4282	0.4341	—	—	—	—	—	—
DCCA	—	—	—	0.3107	0.3064	0.3086	—	—	—
CMCP	0.4717	0.4376	0.4546	0.3681	0.3638	0.3659	0.3995	0.4244	0.4120
HSNN	0.4136	0.3915	0.4025	0.3378	0.3572	0.3475	0.4768	0.4514	0.4641
JGRHML	—	—	—	0.3996	0.4097	0.4046	0.2085	0.2085	0.2085
JRL	0.1338	0.1338	0.1338	0.4081	0.4197	0.4139	0.3777	0.3108	0.3443
JFSSL	0.5073	0.4640	0.4856	0.4139	0.4139	0.4139	0.2242	0.2099	0.2170
S2UPG	—	—	—	0.4289	0.4249	0.4269	0.3713	0.3271	0.3492
CCL	0.5679	0.5633	0.5656	0.4354	0.4413	0.4383	0.4329	0.5036	0.4683
SCAN-AVG	—	—	—	0.3416	0.3458	0.3437	0.2818	0.2373	0.2595
SCAN-LSE	—	—	—	0.3405	0.3340	0.3372	0.2971	0.2087	0.2529
VSESC	—	—	—	0.3617	0.3549	0.3583	0.3356	0.2835	0.3095
Our method	0.5811	0.5750	0.5780	0.5801	0.5702	0.5751	0.4468	0.5162	0.4815

gory information and adds the correlations between textual space and visual space, which makes the performance better than the classical method CCA on Pascal Sentence;

2. CMCP [34] handles the relevant and the irrelevant features of different modalities, and HSN [35] further considers the correlations between nearest neighbors. However, fine-grained information is not considered in these methods, and the correlations between features are not comprehensive;
3. JGRHML [6] uses a joint graph regularization algorithm. JRL [16] further considers semantic information, semi-supervised regularization, and sparse regularization. However, the textual features are sparse in Pascal Sentence, which leads to poor performance. JFSSL [5] learns the projection matrix for each modality; meanwhile, it maintains the relationships between inter-modality and intra-modality. However, fine-grained information is not used in these methods, which leads to limited improvement;
4. S2UPG [17] and CCL [17] take both the coarse-grained and the fine-grained information into consideration; hence, the performances are improved. However, they haven't distinguished the irrelevant features in a multi-margin way;
5. SCAN [11] and VSESC [21] use the attention mechanism, which has a significant effect on matching the fine-grained features with different weights. However, they only use fine-grained data and ignore coarse-grained information. Furthermore, these methods haven't considered the irrelevant degrees between the anchor and the irrelevant samples.

Our method not only fuses coarse-grained and fine-grained information but also considers the degree of discrimination of irrelevant samples to achieve the best results. Especially our method can significantly increase MAP scores. Moreover, Map scores of Pascal sentence and NUS-WIDE 5K by our method increase at around 2% than the best-compared method.

(2) PR curve and TopN-precision curve analysis

Figure 12 contains PR curves for Pascal Sentence, Corel 5K, and NUS-WIDE 5K. We generate random curves and add them to these figures. As expected, the performances of the random curves are the worst.

From all these figures, we can observe that the areas enclosed by the curve of our method are larger than all compared methods. These PR curves reflect that our method has better performance than the benchmarks and the state-of-the-art methods.

Figure 13 contains the TopN-precision curve for Pascal Sentence, Corel 5K, and NUS-WIDE 5K. These curves reflect that our method achieves better per-

formance, especially in Corel 5K. Figure 13(e) and Figure 13(f) show that when the value of N is in $[0,20]$, our method performs much better in NUS-WIDE 5K. This indicates that the higher the sample ranks, the more precise the retrieval result.

(3) Recall@K analysis

We can observe from Table 5 that some of the $I \rightarrow T$ recall@K values of our method are not the highest. This indicates that the true positive items retrieved in the text modality are not accurate enough. In these three datasets, Corel 5K has the highest recall@1 values and the lowest recall@10, and the recall@10 of Pascal Sentence almost approaches 1. Moreover, HSN has a much better performance of the $I \rightarrow T$ recall@K on NUS-WIDE 5K. However, the better performance of CCL and HSN can't continuously maintain the effectiveness on all datasets.

5.5.2. Ablation study

As shown in Table 6, our ablation components are explained as follows: (1) "Coarse-grained" means we train the model using coarse-grained features, (2) "Fine-grained" indicates we use fine-grained features, and (3) "Multi-margin" denotes we use the proposed multi-margin triplet loss to optimize the model. If this item is not selected, it means we only use the traditional triplet loss. From ablation experiments in Table 6, we can observe that:

1. Row 1 and Row 4, Row 2 and Row 5 reflect that using only coarse-grained data performs better than only using fine-grained data. Although fine-grained features have become a research hotspot, coarse-grained data containing position information should not be ignored;
2. Row 3 and Row 6 reveal the importance of fusing the complementary information existing in the fine-grained data. Without the fine-grained features, the performances on three datasets degrade severely, which indicates that the details in fine-grained features can effectively improve the retrieval performance;
3. Row 1 and Row 3, Row 4 and Row 5 show the influence of multi-margin triplet loss with single granularity data. The multi-margin triplet loss improves the retrieval performance greatly on three datasets. Compared with Row 2 and Row 3, when there are more semantic categories, the loss function is more important than using fine-grained data. Thus, the MAP of Corel 5K is improved more obviously than Pascal Sentence and NUS-WIDE 5K;
4. In Row 2 and Row 6, when we only use the triplet loss function, the MAP values are a little lower than the multi-margin triplet loss function on Pascal

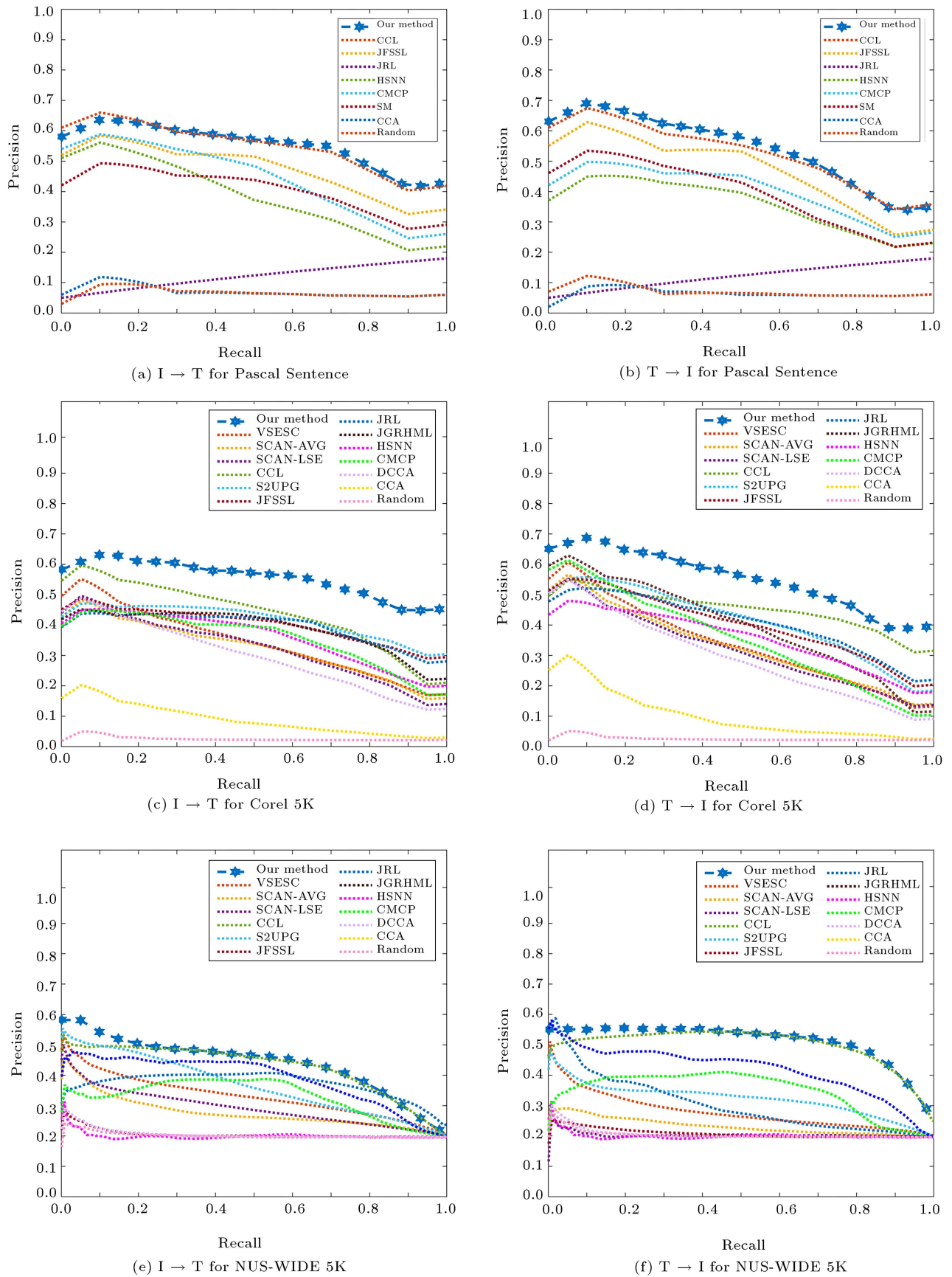


Figure 12. PR curve for all datasets.

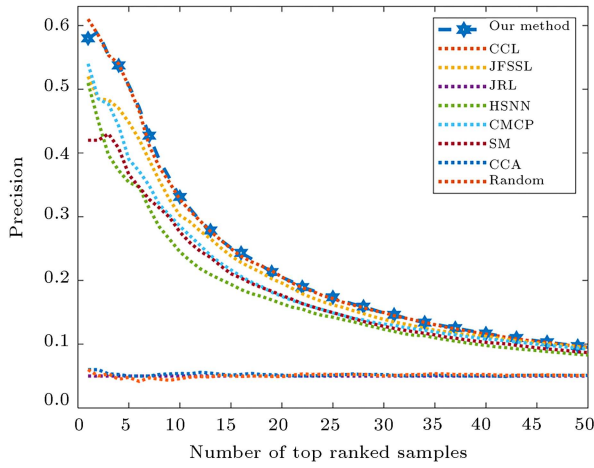
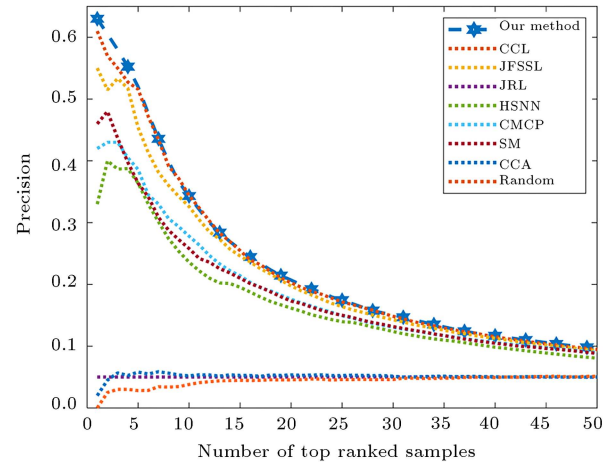
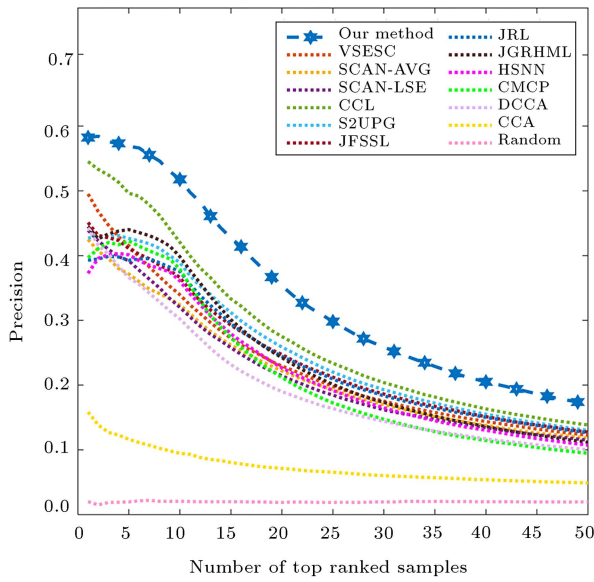
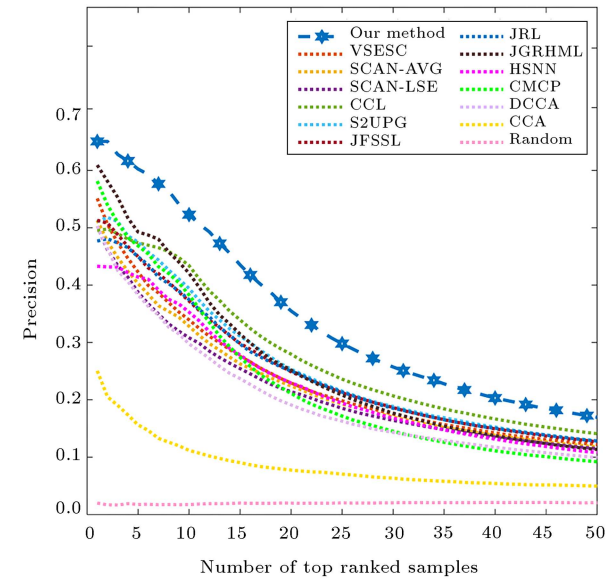
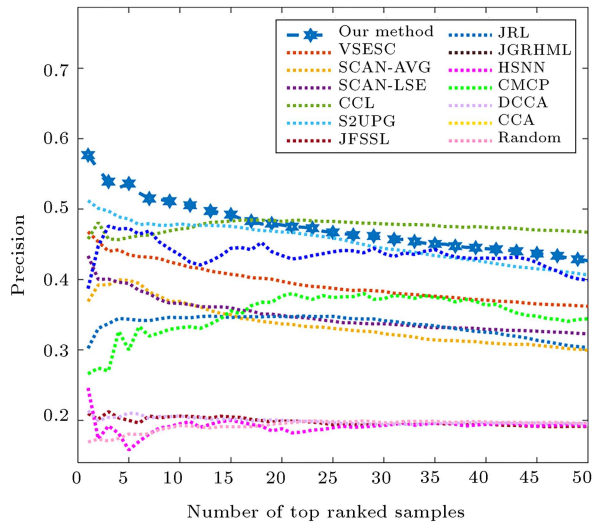
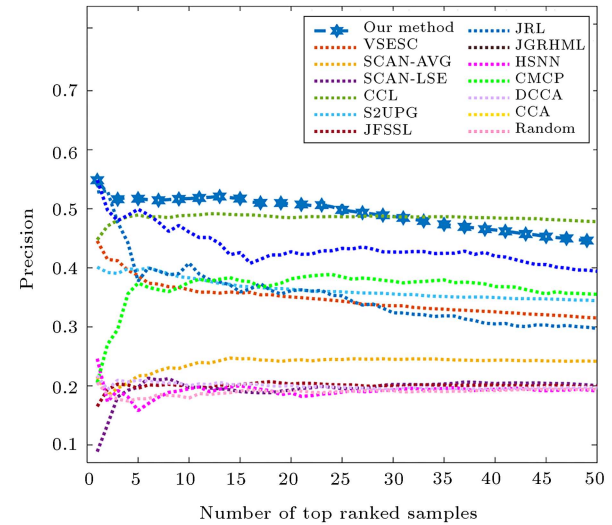
(a) $I \rightarrow T$ for Pascal Sentence(b) $T \rightarrow I$ for Pascal Sentence(c) $I \rightarrow T$ for Corel 5K(d) $T \rightarrow I$ for Corel 5K(e) $I \rightarrow T$ for NUS-WIDE 5K(f) $T \rightarrow I$ for NUS-WIDE 5K**Figure 13.** Top N-precision curve for all datasets.

Table 5. Recall@K of our method and the compared methods.

Methods	Pascal Sentence					Corel 5K					NUS-WIDE 5K							
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10						
	I → T	T → I	I → T	T → I	I → T	T → I	I → T	T → I	I → T	T → I	I → T	T → I						
	I → T	T → I	I → T	T → I	I → T	T → I	I → T	T → I	I → T	T → I	I → T	T → I						
CCA	0.06	0.02	0.24	0.27	0.46	0.44	0.158	0.250	0.306	0.439	0.427	0.581	0.194	0.226	0.577	0.613	0.730	0.859
SM	0.42	0.46	0.71	0.76	0.88	0.85	-	-	-	-	-	-	-	-	-	-	-	-
DCCA	-	-	-	-	-	-	0.454	0.371	0.585	0.641	0.732	0.752	-	-	-	-	-	-
CMCP	0.54	0.42	0.78	0.77	0.91	0.89	0.398	0.565	0.561	0.715	0.623	0.773	0.387	0.548	0.798	0.669	0.883	0.855
HSNN	0.51	0.31	0.73	0.74	0.87	0.82	0.383	0.441	0.579	0.627	0.671	0.713	0.266	0.206	0.831	0.729	0.939	0.847
JGRHML	-	-	-	-	-	-	0.417	0.605	0.509	0.726	0.573	0.802	0.246	0.246	0.790	0.790	0.891	0.891
JRL	0.05	0.05	0.15	0.15	0.30	0.30	0.393	0.419	0.549	0.631	0.641	0.715	0.302	0.528	0.479	0.806	0.524	0.912
JFSSL	0.52	0.55	0.84	0.92	0.92	0.98	0.451	0.513	0.577	0.663	0.649	0.743	0.209	0.165	0.637	0.54	0.851	0.734
S2UPG	-	-	-	-	-	-	0.427	0.511	0.579	0.687	0.687	0.782	0.462	0.375	0.724	0.720	0.829	0.887
CCL	0.61	0.61	0.85	0.93	0.93	0.98	0.545	0.495	0.759	0.721	0.838	0.802	0.201	0.218	0.535	0.603	0.813	0.845
SCAN-AVG	-	-	-	-	-	-	0.506	0.304	0.696	0.589	0.792	0.746	0.089	0.343	0.728	0.597	0.823	0.732
SCAN-LSE	-	-	-	-	-	-	0.191	0.211	0.380	0.338	0.448	0.435	0.216	0.383	0.567	0.742	0.768	0.905
VSESC	-	-	-	-	-	-	0.554	0.423	0.790	0.692	0.881	0.835	0.446	0.385	0.702	0.752	0.899	0.909
Our method	0.58	0.63	0.84	0.94	0.92	0.99	0.602	0.639	0.771	0.807	0.855	0.879	0.548	0.577	0.637	0.855	0.698	0.919

Table 6. Ablation experiment results.

Ablation			MAP								
			Pascal Sentence			Corel 5K			NUS-WIDE 5K		
Coarse-grained	Fine-grained	Multi-margin	I → T	T → I	Average	I → T	T → I	Average	I → T	T → I	Average
✓			0.514	0.508	0.511	0.443	0.436	0.440	0.256	0.314	0.285
✓	✓		0.576	0.568	0.572	0.506	0.490	0.498	0.434	0.511	0.473
✓		✓	0.572	0.555	0.564	0.543	0.526	0.535	0.415	0.465	0.440
	✓		0.386	0.393	0.390	0.432	0.428	0.430	0.206	0.211	0.209
	✓	✓	0.417	0.419	0.418	0.452	0.446	0.449	0.407	0.463	0.435
✓	✓	✓	0.581	0.575	0.578	0.580	0.570	0.575	0.447	0.516	0.482

Sentence and NUS-WIDE 5K. As Corel 5K has more semantic categories, the multi-margin triplet loss shows better effectiveness. This means that the multi-margin triplet loss function with multiple margins is more effective than the triplet loss with a single margin.

Overall, it can be easily concluded that our method that combines these three components outperforms than leaving out any part.

5.5.3. Visualization

We illustrate the visualization of our results in two aspects as follows:

- **Visualization of the feature space.** As is shown in Figure 14, we reduce the dimensions of the image features into two-dimensional spaces and use circles to mark the features that belong to the same category.

From Figure 14(a) and (b), we can see that most of the circle areas on the bottom are smaller than those on the top, which indicates that our multi-margin triplet loss is more effective in optimizing the distributions of features in the same semantic category.

As shown in Figure 14(c) and (d), we could find that all the areas of the circles on the bottom are smaller than on the top. In all categories, the feature aggregation is better with the optimization by the multi-margin triplet loss function.

Figure 14(e) and (f) show the weakest optimization effect in these three datasets, which is in accordance with the MAP results. The clusters of the semantic categories are not obvious, and the areas of the circles on the bottom are a little bit smaller than on the top.

- **Retrieval precision of each semantic category.** It can be seen from Figure 15(a) and (b) that Pascal Sentence has the same trend in the precision of $I \rightarrow T$

and $T \rightarrow I$. The precision values of mutual retrieval are higher for large objects, while smaller items (such as bottles, chairs, dog, and potted plants) have lower precision.

As shown in Figure 15(c) and (d), since the 50 semantic categories of Corel 5K have no specific names, we observe that the semantic categories with lower MAP values usually have strong relatives in visual similarity. They are mostly scenery and have no specific objects, or there are many types of animals in the same semantic category. This leads to a low precision rate in some semantic categories.

As for NUS-WIDE 5K in Figure 15(e) and (f), the precision values of this dataset are generally lower than others. Especially, clouds and sky are highly relative, and the item of the window is a small object; these three categories have poor performance.

5.6. Examples of cross-media retrieval

As described in Figure 16, we provide several typical examples of cross-media retrieval by our method as well as two representative compared methods, CCL and JFSSL, on the Pascal Sentence dataset and display the top five results of $I \rightarrow T$ and $T \rightarrow I$ corresponding to a specific query. Particularly, we choose two queries under the semantic concepts of “Cat” and “Airplane” on $I \rightarrow T$ in Figure 16(a). Moreover, we select another two queries under the semantic concepts of “Car” and “Tvmonitor” on $T \rightarrow I$ in Figure 16(b). Besides, the true matches are marked in green rectangles with check marks, while red rectangles and cross marks indicate incorrect retrieval results.

In general, our method exhibits the best overall performance than CCL and JFSSL and only makes a mistake at the fifth position of the query “Cat” for the task of $I \rightarrow T$. Furthermore, compared with a nonDNN-based method JFSSL, CCL performs better on two tasks of cross-media retrieval, a DNN-based

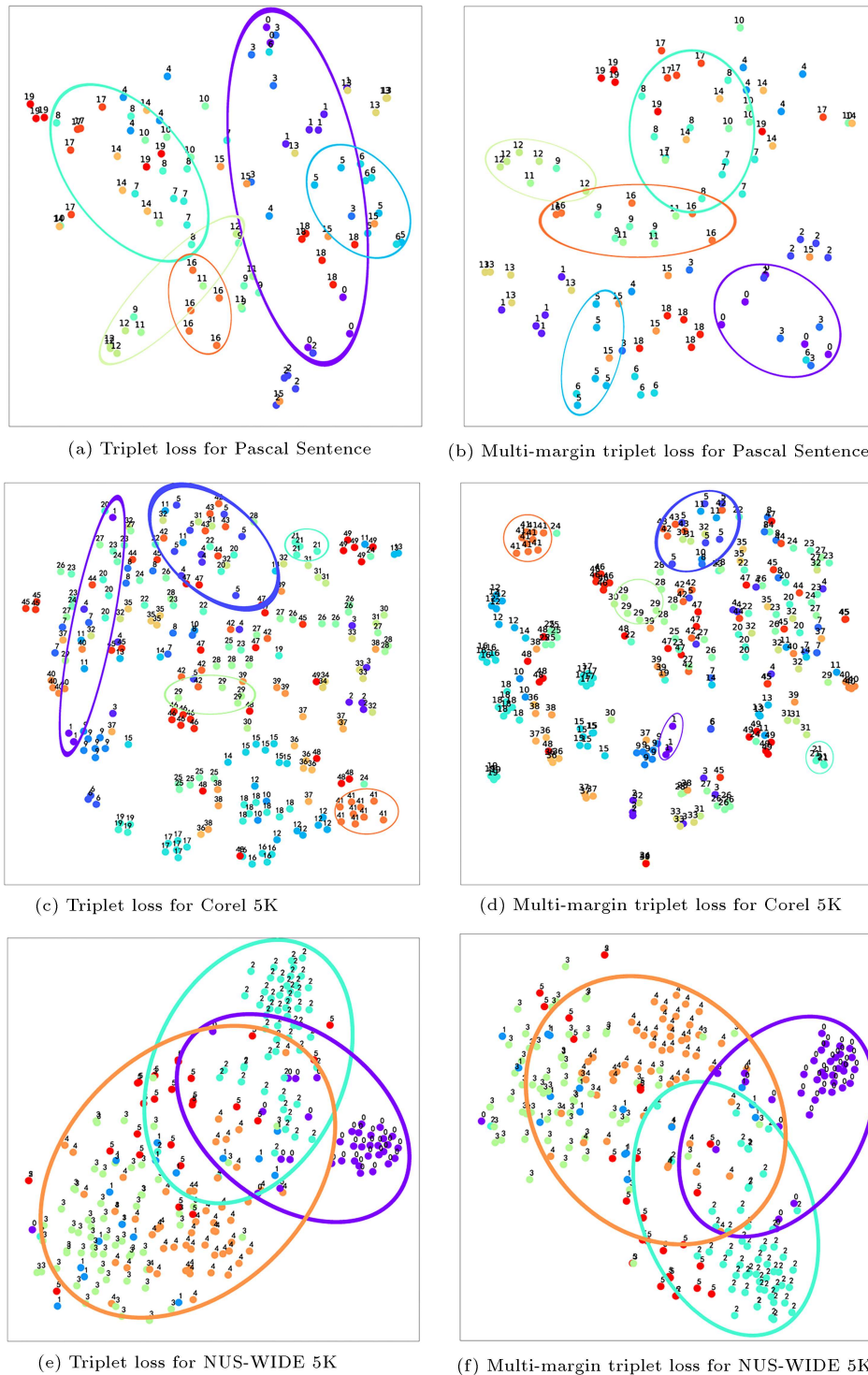


Figure 14. Visualization of image features.

method that takes advantage of fusing multi-grained information in cross-media correlation learning.

6. Conclusion

This paper proposes to separate relevant and irrelevant samples in a multi-margin way with a dual-

framework. (1) Framework I: a multi-grained data fusion framework based on Deep Belief Network, and (2) Framework II: a multi-modality data fusion framework using the multi-margin triplet loss function. These two frameworks aim to realize the multi-grained data fusion and optimize the multi-modality data by the multi-margin triplet loss, respectively. The experimental

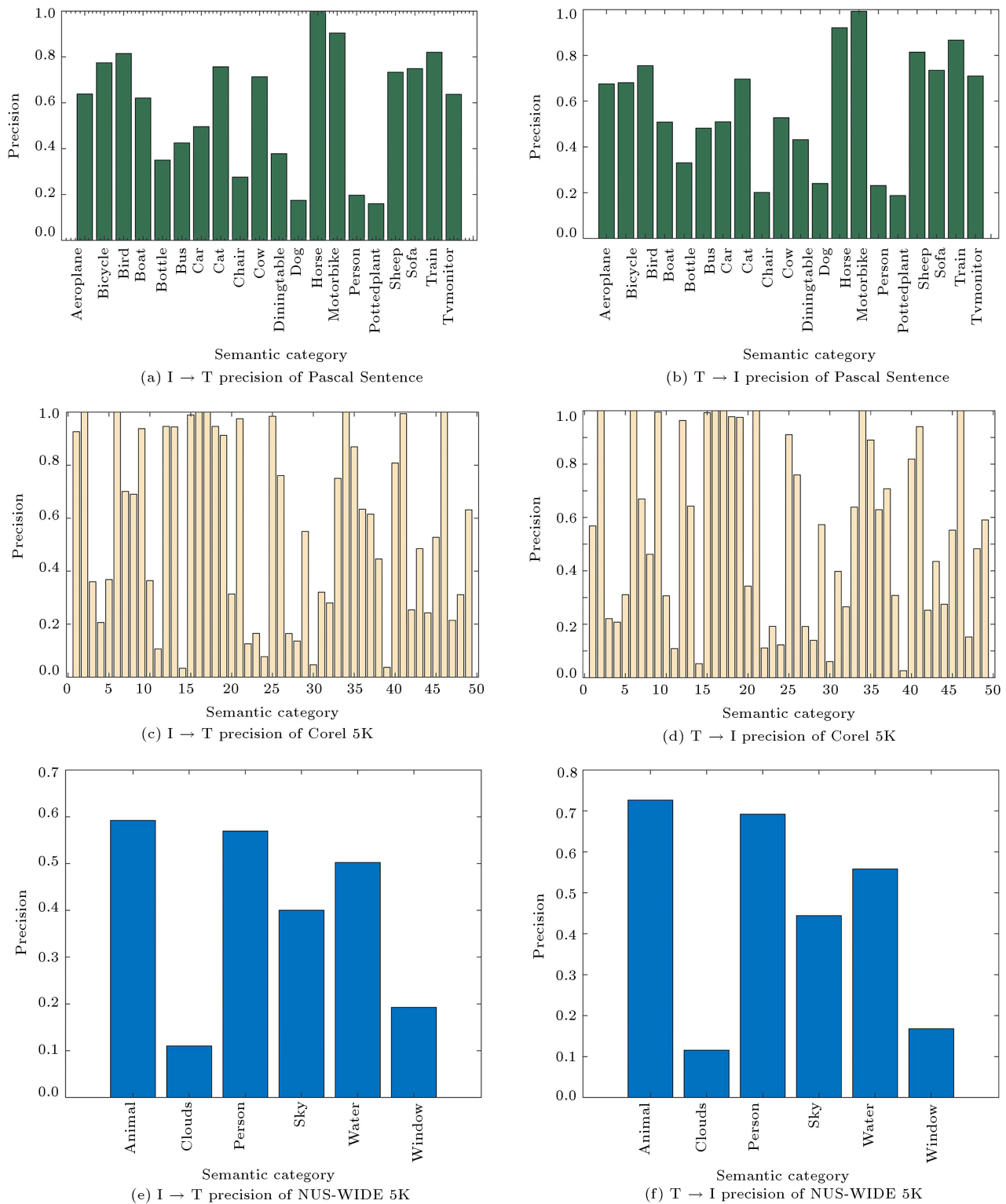

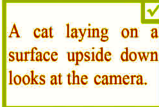
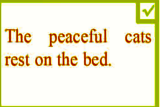
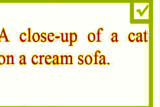
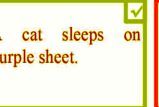
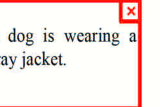
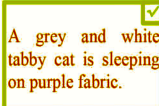
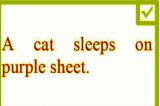

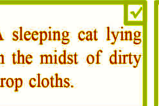
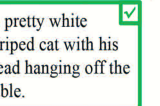
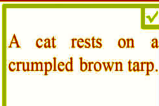
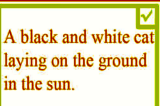
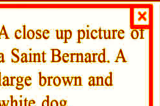
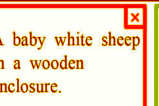
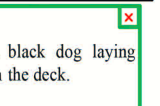


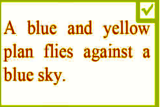

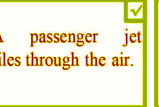
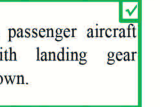
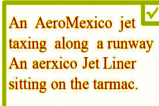
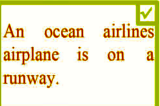

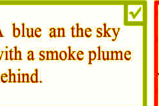
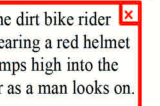
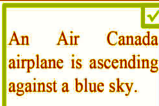
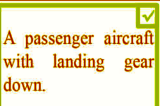
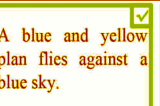
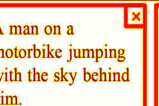



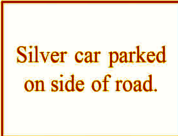















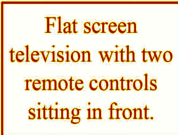









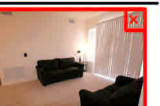





Figure 15. Single semantic category precision.

results show that our proposed method achieves better performance. Furthermore, the ablation experiments verify that the fine-grained information plays an important role in complementing the shortage of coarse-grained data, and the multi-margin triplet loss func-

tion is effective in optimizing the distributions of the features. Future research should consider the following aspects: (1) We could use advanced textual features from the natural language processing field, such as a Transformer; (2) We could get a better image division

Query	Method	Top 5 Results (I→T)				
 Cat	Our Method	 ✓	 ✓	 ✓	 ✓	 ✗
	CCL	 ✓	 ✓	 ✗	 ✓	 ✓
	JFSSL	 ✓	 ✓	 ✗	 ✗	 ✗
 Airplane	Our Method	 ✓	 ✓	 ✓	 ✓	 ✓
	CCL	 ✓	 ✓	 ✓	 ✓	 ✗
	JFSSL	 ✓	 ✓	 ✓	 ✗	 ✗

(a) (I→T)

Query	Method	Top 5 Results (T→I)				
 Car	Our Method	 ✓	 ✓	 ✓	 ✓	 ✓
	CCL	 ✓	 ✓	 ✗	 ✓	 ✗
	JFSSL	 ✓	 ✗	 ✓	 ✗	 ✗
 Tvmonitor	Our Method	 ✓	 ✓	 ✓	 ✓	 ✓
	CCL	 ✓	 ✓	 ✓	 ✓	 ✗
	JFSSL	 ✓	 ✓	 ✗	 ✓	 ✗

(b) (T→I)

Figure 16. Examples of retrieval results on Pascal Sentence dataset.

strategy from computer vision, such as faster-RCNN; and (3) we can implement the proposed model on other modalities, such as video or audio.

7. Acknowledgments

This research was funded by Humanities and Social Sciences Project of Education Ministry (20YJ-A870013), Natural Science Foundation of Shandong Province (ZR2019MF016, ZR2020MF037), NSFC-Zhejiang Joint Fund of the Integration of Informatization and Industrialization (U1909210), Scientific Research Studio in Colleges and Universities of Ji'nan City (202228105, 2021GXRC092), Introduction and Education Plan of Young Creative Talents in Colleges and Universities of Shandong Province, Research Project of Undergraduate Teaching Reform in Shandong Province (Z2020025), Key Research and Development Project of Shandong Province (2019GSF109112), Innovation Team of Youth Innovation Science and Technology Plan in Colleges and Universities of Shandong Province (2020KJN007).

References

- Peng, Y., Huang, X., and Zhao, Y. "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges", *IEEE Trans. Circuits Syst. Video Technol.*, **28**(9), pp. 2372–2385 (2018).
- Baltrušaitis T, Ahuja, C., and Morency, L.-P. "Multimodal machine learning: A survey and taxonomy", *IEEE Trans. Pattern Anal. Mach. Intell.*, **41**(2), pp. 423–443 (2019).
- Hotelling, H. "Relations between two sets of variates", *Biometrika*, **28**(3–4), pp. 321–377 (1936).
- Rasiwasia, N., Costa Pereira, J., Coviello, E., et al. "A new approach to cross-modal multimedia retrieval", *18th ACM Int. Conf. on Multimedia*, pp. 251–260 (2010).
- Wang, K., He, R., Wang, L., et al. "Joint feature selection and subspace learning for cross-modal retrieval", *IEEE Trans. Pattern Anal. Mach. Intell.*, **38**(10), pp. 2010–2023 (2016).
- Zhai, X., Peng, Y., and Xiao, J. "Heterogeneous metric learning with joint graph regularization for cross-media retrieval", *AAAI Conf. on Artif. Intell.*, **27**(1), pp. 1198–1204 (2013).
- Ngiam, J., Khosla, A., Kim, M., et al. "Multimodal deep learning", *28th Int. Conf. on Mach. Learn.*, pp. 689–696 (2011).
- Feng, F., Wang, X., and Li, R. "Cross-modal retrieval with correspondence autoencoder", *22nd ACM Int. Conf. on Multimedia*, pp. 7–16 (2014).
- Srivastava, N. and Salakhutdinov, R. "Learning representations for multimodal data with deep belief nets", *Int. Conf. on Mach. Learn. Workshop* (2012).
- Karpathy, A., Joulin, A., and Fei-Fei, L. "Deep fragment embeddings for bidirectional image sentence mapping", *27th Int. Conf. on Neural Inform. Process. Syst.*, pp. 1889–1897 (2014).
- Lee, K.-H., Chen, X., Hua, G., et al. "Stacked cross attention for image-text matching", *European Conf. on Comput. Vis.*, pp. 201–216 (2018).
- Wang, L., Li, Y., and Lazebnik, S. "Learning deep structurepreserving image-text embeddings", *IEEE Conf. on Comput. Vis. and Pattern. Recogn.*, pp. 5005–5013 (2016).
- Chen, H., Ding, G., Liu, X., et al. "IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval", *IEEE Conf. on Comput. Vis. and Pattern. Recogn.*, pp. 12652–12660 (2020).
- Hermans, A., Beyer, L., and Leibe, B. "In defense of the triplet loss for person re-identification", arXiv:1703.07737 (2017).
- Andrew, G., Arora, R., Bilmes, J., et al. "Deep canonical correlation analysis", *30th Int. Conf. on Mach. Learn.*, pp. 1247–1255 (2013).
- Zhai, X., Peng, Y., and Xiao, J. "Learning cross-media joint representation with sparse and semisupervised regularization", *IEEE Trans. Circuits Syst. Video Technol.*, **24**(6), pp. 965–978 (2014).
- Peng, Y., Zhai, X., Zhao, Y., et al. "Semi-supervised cross-media feature learning with unified patch graph regularization", *IEEE Trans. Circuits Syst. Video Technol.*, **26**(3), pp. 583–596 (2016).
- Frome, A., Corrado, G.S., Shlens, J., et al. "DeViSE: a deep visualesemantic embedding model", *AAAI 26th Int. Conf. on Neural Inform. Process. Syst.*, pp. 2121–2129 (2013).
- Socher, R., Karpathy, A., Le, Q.V., et al. "Grounded compositional semantics for finding and describing images with sentences", *Trans. of Assoc. for Comput. Ling.*, **2**, pp. 207–218 (2014).
- Vendrov, I., Kiros, R., Fidler, S., et al. "Order-embeddings of images and language", arXiv:1511.06361 (2015).
- Chen, H., Ding, G., Lin, Z., et al. "Cross-modal image-text retrieval with semantic consistency", *27th ACM Int. Conf. on Multimedia*, pp. 1749–1757 (2019).
- Misraa, A.K., Kale, A., Aggarwal, P., et al. "Multi-modal retrieval using graph neural networks", arXiv:2010.01666 (2020).
- Wang, L., Li, Y., Huang, J., et al. "Learning two-branch neural networks for image-text matching tasks", *IEEE Trans. Pattern Anal. Mach. Intell.*, **41**(2), pp. 394–407 (2019).

24. Xu, X., Wang, T., Yang, Y., et al. “Cross-modal attention with semantic consistence for image-text matching”, *IEEE Trans. Neural Networks Learn. Syst.*, **31**(12), pp. 5412–5425 (2020).
25. Silberer, C. and Lapata, M. “Learning grounded meaning representations with autoencoders”, *52nd Annu. Mtg. of the Assoc. for Comput. Ling.*, pp. 721–732 (2014).
26. Salakhutdinov, R. and Larochelle, H. “Efficient learning of deep Boltzmann machines”, *13th Int. Conf. on Artif. Intell. and Stats.*, pp. 693–700 (2010).
27. Hinton, G.E., Osindero, S., and Teh, Y.-W. “A fast learning algorithm for deep belief nets”, *Neural Comput.*, **18**(7), pp. 1527–1554 (2006).
28. Peng, Y., Qi, J., Huang, X., et al. “CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network”, *IEEE Trans. Multimedia*, **20**(2), pp. 405–420 (2018).
29. Zhang, J., Peng, Y., and Yuan, M. “SCH-GAN: semi-supervised cross-modal hashing by generative adversarial network”, *IEEE Trans. Cybern.*, **50**(2), pp. 489–502 (2020).
30. Zhang, J., Peng, Y., and Yuan, M. “Unsupervised generative adversarial cross-modal hashing”, *AAAI Conf. on Artif. Intell.*, **32**(1) (2018).
31. Wang, B., Yang, Y., Xu, X., et al. “Adversarial cross-modal retrieval”, *25th ACM Int. Conf. on Multimedia*, pp. 154–162 (2017).
32. Schroff, F., Kalenichenko, D., and Philbin, J. “FaceNet: a unified embedding for face recognition and clustering”, *2015 IEEE Conf. on Comp. Vis. Patt. Recog.*, pp. 815–823 (2015).
33. Zhou, M., Niu, Z., Wang, L., et al. “Ladder loss for coherent visualsemantic embedding”, *AAAI Conf. on Artif. Intell.*, pp. 13050–13057 (2020).
34. Zhai, X., Peng, Y., and Xiao, J. “Cross-modality correlation propagation for cross-media retrieval”, *2012 IEEE Int. Conf. on Acou. Sp. Sig. Proc.*, pp. 2337–2340 (2012).
35. Zhai, X., Peng, Y., and Xiao, J. “Effective heterogeneous similarity measure with nearest neighbors for cross-media retrieval”, *Int. Conf. on Adv. Multim. Model.*, pp. 312–322 (2012).

Biographies

Zheng Liu is a Professor and PhD supervisor with the School of Computer Science and Technology, Shandong University of Finance and Economics. He received the PhD degree from Shandong University in July 2011. His research interests mainly include cross-media analysis and retrieval, automatic image annotation, and machine learning.

Shaojing Yuan is a Master’s degree candidate at Shandong University of Finance and Economics. Her main research interests include cross-media retrieval and machine learning.

Xinlei Pei is a Master’s degree candidate at Shandong University of Finance and Economics. His main research interests include multimedia information retrieval and machine learning.

Shanshan Gao is a Professor and PhD supervisor at the School of Computer Science and Technology, Shandong University of Finance and Economics. Her current research interests include intelligent graphics and image processing, data mining, and visualization.

Huijian Han received the MSc and PhD degrees from Shandong University, Jinan, China, in 2003 and 2010, respectively. Since 2007, he has been a Professor at the School of Computer Science and Technology, Shandong University of Finance and Economics. He has authored or co-authored more than 50 papers in journals and conferences. His research interests include information management, cognitive intelligence, and visual decision.