CMPD: Using Cross Memory Network With Pair Discrimination for Image-Text Retrieval

Xin Wen, *Graduate Student Member, IEEE*, Zhizhong Han, and Yu-Shen Liu, *Member, IEEE*

***Abstract*— Cross-modal retrieval using deep neural networks aims to retrieve relevant data between the two different modalities. The performance of cross-modal retrieval is still unsatisfactory for two problems. First, most of the previous methods failed to incorporate the common knowledge among modalities when predicting the item representations. Second, the semantic relationships indicated by class label are still insufficiently utilized, which is an important clue for inferring similarities between the cross modal items. To address the above issues, we propose a novel cross memory network with pair discrimination (CMPD) for image-text cross modal retrieval, where the main contributions lie in two-folds. First, we propose the cross memory as a set of latent concepts to capture the common knowledge among different modalities. It is learnable and can be fused into each modality through attention mechanism, which aims to discriminatively predict representations. Second, we propose the pair discrimination loss to discriminate modality labels and class labels of item pairs, which can efficiently capture the semantic relationships among these modality labels and class labels. Comprehensive experimental results show that our method outperforms the state-of-the-art approaches in image-text retrieval.**

***Index Terms*— Retrieval, cross-modal retrieval, adversarial learning.**

I. Introduction

CROSS-MODAL retrieval aims at mining the semantic relationships between the items of multiple modalities from different multimedia sources [1]. It is a basic and important task related to many real-world multimedia applications such as image-text retrieval [2]–[4], image-video retrieval [5], and image-3D retrieval [6]. This task is challenging because the items from different modalities usually have different distributions in the latent feature subspace [7]. Such modality gap impedes directly assessing the semantic similarity between the items across modalities. Generally speaking, one common solution is to learn a common representation subspace for different modalities, where the similarities among the items from different modalities can be measured directly. Most previous studies [7], [8] focus on constructing a good common subspace so that the similar items from different modalities are close to each other in the common space. Without loosing generality, in this paper, we follow the previous work [2], [9] to mainly address the problem by focusing on learning the common representation subspace for the items from image and text modalities. Because it is the widely used multimedia forms and is adopted as the typical cross-modal resources in many previous multimedia studies [10], [11].

Although a lot progress has been achieved in this area, the performance of image-text retrieval still remains unsatisfactory. The problems lie in two-folds. The first one is the absence of common knowledge when predicting the representation for single modality. Usually, it is naturally desirable for network to notice the common information across modalities, when predicting the representation for certain one of these modalities. However, current approaches only establish the link to share information at the end of the network, where the element-wise spatial constraints (e.g. Euclidean or cosine distance) are used for describing the common information between the semantically similar items. As a result, such practice limits the representation learning ability of the network. The second one is the insufficient utilization of class relationship, where two items with the same class should be kept close, no matter which modalities they come from. The class relationships between pair of items (i.e. from the same class, or not from the same class) can be a helpful signal to decide the semantic similarity between items. However, only few work considers such relationship. Both SSAH [12] and DSCMR [13] try to preserve the class relationships in a latent label space. The problem is, such kind of method suffers from the information loss by indirectly measuring the class relationship between different modalities.

Therefore, in this paper, we propose a cross memory network with pair discrimination (CMPD) to address the above-mentioned issues. In order to share the learned common information across modalities when predicting representations for certain modality, we maintain a set of learnable latent concept representations called *cross memory*. For each item from image or sentence modality, we enable its visual or semantic feature to interact with these memories via attention mechanism. Benefit from the capability of attention in

Manuscript received July 2, 2020; accepted August 2, 2020. Date of publication August 17, 2020; date of current version June 4, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB0505400, Grant 2019YFB1405703, and Grant TC190A4DA/3; in part by the NSFC under Grant 61672307; and in part by the Tsinghua-Kuaishou Institute of Future Media Data. This article was recommended by Associate Editor Y. Qin. *(Corresponding author: Yu-Shen Liu.)*

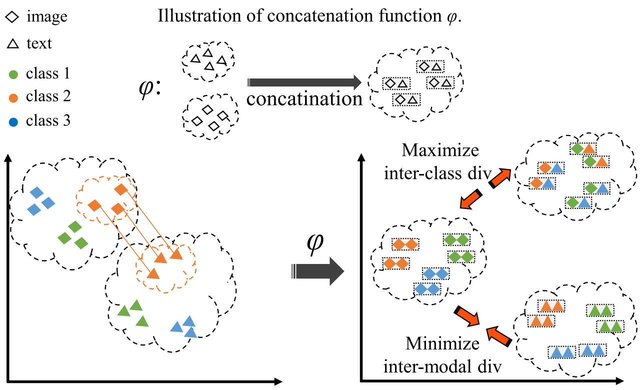
Xin Wen is with the School of Software, Tsinghua University, Beijing 100084, China (e-mail: x-wen16@mails.tsinghua.edu.cn).

Zhizhong Han is with the Department of Computer Science, University of Maryland at College Park, College Park, MD 20737 USA (e-mail: h312h@mail.nwpu.edu.cn).

Yu-Shen Liu is with the School of Software, BNRist, Tsinghua University, Beijing 100084, China (e-mail: liuyushen@tsinghua.edu.cn).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2020.3017344



(a) Common representation subspace. (b) Joint distribution subspace.

Fig. 1. The comparison of our pair discrimination in (b) with the element-wise constraints in (a). In (a), the element-wise constraints only attend to the paired relationships represented by the solid line. On the other hand, in (b), the proposed CMPD preserves the semantic structure of the entire distribution in the modality level (black dotted cloud) and the class level (orange dotted cloud) by considering the modal-pair and class-pair divergence in the subspace of paired items.

modeling relationships among items, we can explicitly share the representation learned from the relatedness between the modalities through the cross memory, which can further provide the evidence for deciding the similarity between item pairs.

Meanwhile, in order to construct a class-relationship-aware common representation subspace, we propose the pair discrimination method, where we incorporate the class relationships between a pair of items by means of adversarial learning. Specifically, we design two discriminators on top of the common representation subspace. One discriminator aims to distinguish whether the input pair of items come from the same modality or not. And the other discriminator is used to distinguish whether there exists a class relationship between them, i.e., whether they belong to the same class.

In CMPD, we propose two pair discrimination divergences for the discriminators. (1) The modal-pair divergence distinguishes the modality label of input pairs, in which the items of a pair have the same class labels. We minimize the modal-pair divergence in order to merge the text and image distribution together. (2) The class-pair divergence distinguishes the class label of input pairs, in which the discriminator should judge whether the items in the input pair have the same label. We optimize the class-pair divergence in order to separate the pairs containing different class labels away from the pairs containing the same class labels. Note that we use the term "divergence" instead of "loss" because we incorporate our CMPD with the WGAN [14] framework, and leverage the pair discrimination as minimizing or maximizing the divergences between different distributions, as shown in Fig. 1.

The proposed CMPD is evaluated on four widely used benchmarks. The main purpose of our work concerns the representation learning component in image-text retrieval task, and in particular regarding the effectiveness of the process and the objectives for learning a well structured common subspace. The contributions can be summarized as follows.

- In CMPD, the cross memory is proposed to capture the common knowledge between two modalities. By interacting with the representation from various modalities through attention mechanism, the model can efficiently learn the common knowledge among modalities, and refer to it when predicting the representations for items.
- By combing with adversarial learning framework, we design two discriminators on top of the common representation subspace, which can effectively integrate both the modality-level and class-level semantic relationships between the items across all modalities.
- Two pair discrimination divergences are proposed in CMPD to capture the semantic relationships between unpaired image and text items. Compared with element-wise spatial constraints, pair discrimination divergences can leverage the modality-level and class-level semantic relationships as a whole, which preserves a more discriminative structure of common space.

II. Related Work

The research for image-text retrieval task is a specific problem of cross-modal retrieval, which can be roughly divided into two mainstreams. The first is the *cross-modal hashing methods* that mainly focus on the retrieval efficiency by mapping the items of different modalities into a common binary Hamming space [12], [15]–[19]. The second is the joint representation learning methods, where four subclasses can be further distinguished as unsupervised methods [20]–[23], pairwise methods [24], [25], ranking-based methods [26] and supervised [7], [27] methods. Our model belongs to the subclass of supervised cross-modal retrieval methods, where the class labels of items are provided during training.

On the other hand, our model is also related to the generative adversarial networks (GANs) [28] in terms of joint distribution matching. Recently, several studies have been proposed to introduce the adversarial learning into cross-modal retrieval tasks. For instances, both ACMR [7] and UCAL [29] introduce a modality classifier to distinguish the modality of the input items. By applying an additional label prediction loss and triplet loss to preserve the cross-modal semantic structure in the common representation subspace, ACMR achieves the state-of-the-art performance on supervised cross-modal retrieval task. CM-GANs [30] introduces an intra-modal discriminator to judge original features apart from the reconstructed features. For unsupervised learning, the CM-GANs proposed a two-way convolutional autoencoder structure with weight sharing constraints for the generator. It is one of the few work in this area that explores the way to share the learned common semantic features between two generators. By combining the modality discriminator and the transfer learning strategy, the MHTN [31] method takes one step further to support the cross-modal learning of more than two modalities at the same time. Different from the previous work, the CMPD focuses more on exploring the adversarial learning framework's ability of learning the diverse semantic relationships between distributions both in modalities level and classes level.

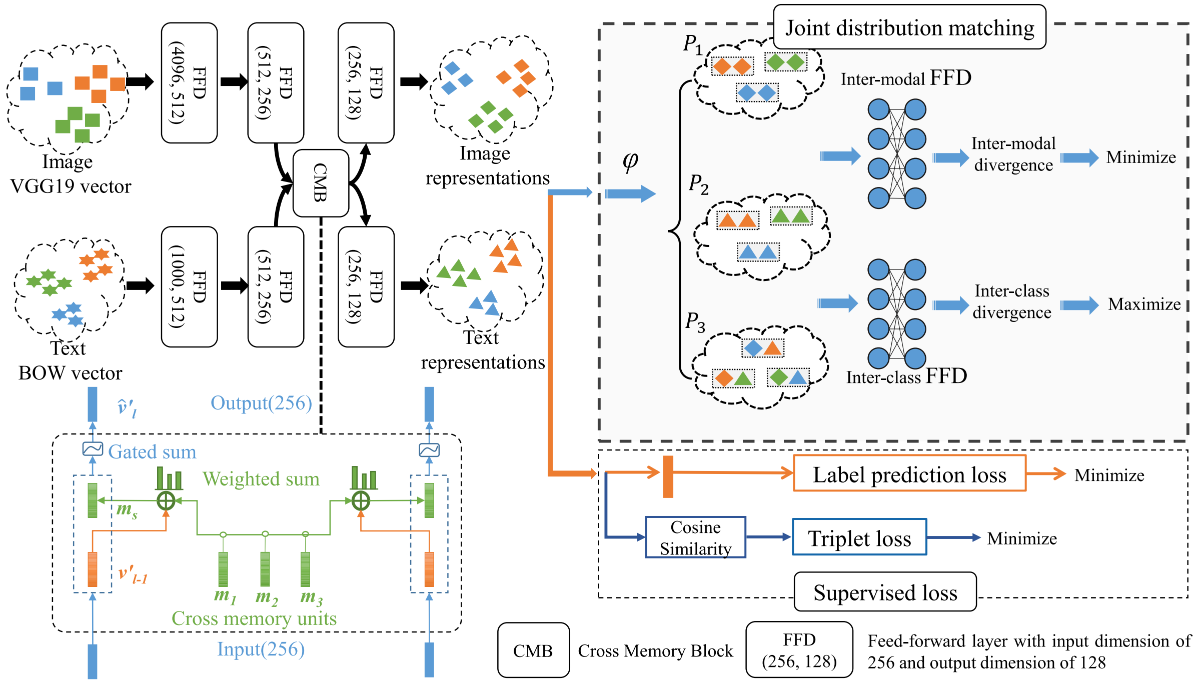


Fig. 2. The framework of our proposed CMPD. Our model consists of three parts. The first part is the representation generator, in which two feed-forward networks are applied to transfer the input images and texts feature vectors into the representation vectors. The second part is the discriminator, where the representation vectors are transferred from the common representation subspace into the joint distribution subspace by the concatenation mapping ϕ. Two divergences are calculated and optimized by two feed-forward networks. The third part is the supervised loss, where the label loss and triplet loss are adopted by following [7].

Our work is theoretically inspired by the WGAN [14], in which the Wasserstein distance (WD) is introduced to stabilize the training of GANs. The WD's property to measure two completely disjoint distributions is crucial for the joint distribution matching, where the situations of handling the disjoint distributions occur frequently. For discriminator, WGAN provides a reliable framework and training methodology to force the discriminators to constantly approach the WD between two distributions during training. Therefore, the optimization goal of the generator is then simplified by minimizing or maximizing the distance between two distributions.

III. The CMPD Model

In this section, we first introduce some important notations and background. Then, we detail the framework of our proposed CMPD. An overview of our model is shown in Fig. 2.

A. Background

1) Problem Formulation: Without losing generality, in this paper we formulate our description for cross modal retrieval problem, where the dataset consists of a collection of image-text pairs. Let $\{(\boldsymbol{v}_i^c, \boldsymbol{t}_i^c)\}_{i=1}^n \sim \mathcal{D}$ denote the discrete distribution of n image-text pairs with m classes, where $\boldsymbol{v}_i^c$ is the image feature vector and $\boldsymbol{t}_i^c$ is the text feature vector in class $c \in \{1, \ldots, m\}$, respectively. Then, the image feature vector distribution and the text feature vector distribution can be written as $\{\boldsymbol{v}_i^c\} \sim \mathcal{V}$ and $\{\boldsymbol{t}_i^c\} \sim \mathcal{T}$, respectively.

Then, the problem of cross-modal retrieval is to find two projections $\boldsymbol{v}'^c_i = f_{\mathcal{V}}(\boldsymbol{v}_i^c)$ and $\boldsymbol{t}'^c_i = f_{\mathcal{T}}(\boldsymbol{t}_i^c)$ that map the image and text feature vectors into a common representation subspace S, where the similarities between the cross-modal items can be directly assessed. The distributions of image and text representation vectors in the common representation subspace are denoted as $\{\boldsymbol{v}'^c_i\} \sim \mathcal{V}'$ and $\{\boldsymbol{t}'^c_i\} \sim \mathcal{T}'$, respectively.

2) Wassertein Distance and WGAN: In most of the previous studies, the discriminator D is adopted as the modality classifier and takes the cross entropy loss of binary predictions as the optimization goal, which can be formulated as below

$$\mathcal{L}_D = \mathbb{E}_{\boldsymbol{v}\sim\mathcal{V}} \log D(f_{\mathcal{V}}(\boldsymbol{v})) + \mathbb{E}_{\boldsymbol{t}\sim\mathcal{T}} \log[1 - D(f_{\mathcal{T}}(\boldsymbol{t}))]. \quad (1)$$

According to [14], the optimization of $\mathcal{L}_D$ is equivalent to approach the Jesen-Shannon Divergence (JSD) between the image representation distribution $\mathcal{V}$ and the text representation distribution $\mathcal{T}$. However, the JSD will raise a constant value for distributions that have no overlap with each other, which is usually the case in the training process of GAN. This will cause the problem of gradient diminishing and hinder the optimization of discriminator. Therefore, the discriminator could not be fully optimized to approach the actual value of the JSD. This situation deviates from the discriminator's training goal of our model, which is to obtain a precise estimation of the divergence between two distributions. Compared with JSD, the Wasserstein distance (WD) introduced in WGAN [14] theoretically avoids the problem of raising constant value when measuring two non-overlap distributions. Therefore, it can be adopted as the divergence measurement in our model. According to [32], the optimization can force the discriminator constantly approach the WD of two distributions by simply applying a gradient penalty term in the discriminator's loss.

B. Cross Memory Networks

In this paper, the feed-forward network is adopted as the basic framework for both image generator $f_{\mathcal{V}}$ and text generator $f_{\mathcal{T}}$. Note that, since the model focuses on learning the common representation of image and text items, the pre-trained feature vectors for both images and texts are used as the input of generators. The overall structure of cross memory network consists of three feed-forward layers with one cross memory block (CMB) between two top layers, as shown in left half of Fig. 2. For simplicity, we only introduce the image generator. The structure of text generator is the same as image generator except for the input dimensions of the first feed-forward layer.

Let $\boldsymbol{v'}_l \in \mathbb{R}^{d_l}$ denote the output of the l-th layer for image generator. For the bottom two layers, the output is given as

$$\boldsymbol{v'}_l = f(W_l^{\boldsymbol{v'}}\boldsymbol{v'}_{l-1} + \boldsymbol{b}_l^{v'}), \tag{2}$$

where $W_l^{\boldsymbol{v'}} \in \mathbb{R}^{d_{l-1}\times d_l}$ and $\boldsymbol{b}_l^{v'} \in \mathbb{R}^{d_l}$ are exclusive parameters for image generator.

Between the top two layers, the cross memory block (CMB) maintains a set of cross memory units (CMUs) $M = \{\boldsymbol{m}_1, \boldsymbol{m}_2, \boldsymbol{m}_3, \ldots, \boldsymbol{m}_k\}$, in which $\boldsymbol{m}_i \in \mathbb{R}^{d_{l-1}}$ represents the learned common feature between image and text items, as shown in the left bottom part of Fig. 2. For each input $\boldsymbol{v'}_{l-1}$ from the previous layer, the cross memory block determines the importance of each cross memory unit to the input vectors, by assigning a weight factor w_i for each unit $\boldsymbol{m}_i$ according to the current input $\boldsymbol{v'}_{l-1}$. The importance is calculated as a probability between 0 and 1 through sigmoid function as below

$$w_i = sigmoid(\boldsymbol{m}_i^T\boldsymbol{v'}_{l-1}). \tag{3}$$

Then the shared feature representation $\boldsymbol{m}_s$ is obtained by the weighted sum of all the units in the cross memory block

$$\boldsymbol{m}_s = \sum_{i=1}^{i=k} w_i \cdot \boldsymbol{m}_i. \tag{4}$$

To integrate the information of original input $\boldsymbol{v'}_{l-1}$ and the shared feature representation $\boldsymbol{m}$, the cross memory block uses a gate structure to determine how much information of $\boldsymbol{m}$ is remained for combination and how much information is removed from the input $\boldsymbol{v'}_{l-1}$, and the ratio of remaining information is calculated but sigmoid function as a logit between 0 and 1, as shown in Eq. (5).

$$p = sigmoid([\boldsymbol{m_s} : \boldsymbol{v'}_{l-1}]W_g), \tag{5}$$

where $W_g \in \mathbb{R}^{2d_{l-1}\times 1}$ and ":" denotes the concatenation of two matrixes. Then the combination representation of shared and modality-specific features $\hat{\boldsymbol{v}}'_l$ is given as

$$\hat{\boldsymbol{v}}'_l = (1-p)\cdot\boldsymbol{v'}_{l-1} + p\cdot\boldsymbol{m_s}. \tag{6}$$

The output of the l-th feed-forward layer is formulated as follow

$$\boldsymbol{v'}_l = relu(W_l^{\hat{\boldsymbol{v}}'}\hat{\boldsymbol{v}}'_l + \boldsymbol{b}_l^{\hat{v}'}), \tag{7}$$

where $W_l^{\hat{\boldsymbol{v}}'} \in \mathbb{R}^{d_{l-1}\times d_l}$ and $\boldsymbol{b}_l^{\hat{v}'} \in \mathbb{R}^{d_l}$.

Before training, the cross memory units are first initialized as a set of random vectors sampled from a random distribution. We maintain these cross memory units by regarding them as trainable weights and updating them along with the other weights in the network during training. The common representation in the cross memory units is learned through the attention mechanism, where the mechanism calculates the importance (which is the weight) of each unit to the input data. These units are fixed during testing phase but not in training phase. During testing, these units are treated as ordinary network parameters to calculate representations.

For representation learning, L2 normalization (L2-norm) has been proved to be efficient for improving the performance of image classification [33], [34]. And for text classification and retrieval, cosine similarity is more generally used as a metric for measuring the distance when working with text feature vectors represented by word counts. Inspired by the above two aspects, we follow [34] to introduce L2-constraint on the generated representation vectors and use cosine similarity as distance measurement. Specifically, for generated image and text representation vector $\boldsymbol{v'}$ and $\boldsymbol{t'}$, we regularize them to the unit hypersphere in the common representation subspace as $\tilde{\boldsymbol{v}} = \frac{\boldsymbol{v'}}{\|\boldsymbol{v'}\|}$ and $\tilde{\boldsymbol{t}} = \frac{\boldsymbol{t'}}{\|\boldsymbol{t'}\|}$. Then the cosine similarity between $\tilde{\boldsymbol{v}}$ and $\tilde{\boldsymbol{t}}$ can be formulated as

$$sim(\tilde{\boldsymbol{v}}, \tilde{\boldsymbol{t}}) = \frac{\tilde{\boldsymbol{v}}\cdot\tilde{\boldsymbol{t}}}{\|\tilde{\boldsymbol{v}}\|\|\tilde{\boldsymbol{t}}\|}. \tag{8}$$

Note that the similarity function is actually redundant, since the representations are eventually normalized to the unit length in our model. In practice, we only use the dot product for triplet loss during training and the retrieval task during evaluation.

C. Joint Distribution Divergences

The joint distribution matching is a variant of adversarial learning, but focusing more on exploring the discriminator's ability of measuring divergence between distributions. Solutions label prediction loss and triplet loss [7] may help to relief the problem of preserve the inter-class structure of the items. However, label prediction loss cannot disperse the distributions of different classes far away in the common representation subspace, which we will discuss in Fig. 4, while the triplets sampled from mini-batch data cannot traverse the entire collection of triplets for its expensive computational cost [7], both of the two methods cannot fully preserve the inter-class structure of the items in practice.

In the view above, we propose to solve the problem in a joint distribution subspace, where the item's modality and class information can be considered simultaneously. Before further discussion, we first introduce some complementary notations for convenience. Given the regularized representation vectors distribution $\{\tilde{\boldsymbol{v}}_i^c\}_{i=1}^n$ and $\{\tilde{\boldsymbol{t}}_i^c\}_{i=1}^n$, we consider the following three joint distributions that contain both the modality and class information of the data, which are text-text pairs with texts have the same class label in each pair as Eq. (9), image-image pairs with images have the same class label in each pair as Eq. (10), and image-text pairs with the items have different label in each pair as Eq. (11).

$$P_1 = \{(\tilde{\boldsymbol{v}}_i^c, \tilde{\boldsymbol{v}}_j^c)\}_{c=1, i \neq j}^{c=m}, \tag{9}$$

$$P_2 = \{(\tilde{\boldsymbol{t}}_i^c, \tilde{\boldsymbol{t}}_j^c)\}_{c=1, i \neq j}^{c=m}, \tag{10}$$

$$P_3 = \{(\tilde{\boldsymbol{v}}_i^c, \tilde{\boldsymbol{t}}_j^d)\}_{c \neq d, i \neq j}. \tag{11}$$

By introducing the above three distributions, we transform the simple modality matching problem into a joint representation space, where both the inter-modal and inter-class structure of the items can be fully preserved during adversarial training.

1) Inter-Modal Divergence: In order to preserve the inter-modal structure, we consider to minimize the following inter-modal divergence ($\mathcal{L}_{imd}$) based on Wasserstein distance (W). When this divergence takes the minimum value, the distributions of P_1 and P_2 are completely overlapped with each other, which means the modality distributions of image and text representation vectors overlap in the common representation subspace. Besides, since the image and text pairs of P_1 and P_2 consist of items that have the same class label, minimizing this divergence will bring representation vectors with the same class labels close to each other.

To stabilize the training, following the practice of [32], we use the gradient penalty ($\mathcal{P}_{imd}$) as the constraints and formulate objective function for discriminator and generator as follow:

$$\mathcal{L}_{imd}(D) = \mathbb{E}_{x \sim P_1}[D_{imd}(x)] - \mathbb{E}_{x \sim P_2}[D_{imd}(x)] + \lambda_{gp}\mathcal{P}_{imd}, \tag{12}$$

$$\mathcal{P}_{imd} = \sum_{i \in \{1,2\}} \mathbb{E}_{x \sim P_i}[(\|\nabla_x D_{imd}(x)\|_2 - 1)^2], \tag{13}$$

$$\mathcal{L}_{imd}(G) = -\mathbb{E}_{x \sim P_1}[D_{imd}(x)] + \mathbb{E}_{x \sim P_2}[D_{imd}(x)], \tag{14}$$

where λ_{gp} is the preset weight of gradient penalty term. Note that the generator G in the above function consists of text representation generator $f_{\mathcal{T}}$ and image representation generator $f_{\mathcal{V}}$.

2) Inter-Class Divergences: The main purpose of inter-class divergence is to pull items of different classes away from each other in the common representation subspace. Solutions like directly measuring the divergence between every class pairs in the common representation subspace may also achieve the same effect. However, it usually comes at the expense of the training time, especially for dataset with large number of classes.

Same as the inter-modal divergence, we have the following objective functions for discriminator and generator, where $\mathcal{L}_{icd}(D)$ in Eq. (15) aims to measure the divergence between P_1 and P_3, and $\mathcal{L}_{icd}(G)$ in Eq. (17) aims to minimize the divergence between P_1 and P_3.

$$\mathcal{L}_{icd}(D) = \mathbb{E}_{x \sim P_1}[D_{icd}(x)] - \mathbb{E}_{x \sim P_3}[D_{icd}(x)] + \lambda_{gp}\mathcal{P}_{icd}, \tag{15}$$

$$\mathcal{P}_{icd} = \sum_{i \in \{1,3\}} \mathbb{E}_{x \sim P_i}[(\|\nabla_x D_{icd}(x)\|_2 - 1)^2], \tag{16}$$

$$\mathcal{L}_{icd}(G) = \mathbb{E}_{x \sim P_1}[D_{icd}(x)] - \mathbb{E}_{x \sim P_2}[D_{icd}(x)]. \tag{17}$$

The total adversarial training objective functions for generator and discriminator are formulated as

$$\mathcal{L}_{adv}(G) = \mathcal{L}_{imd}(G) + \lambda_{icd}\mathcal{L}_{icd}(G), \tag{18}$$

$$\mathcal{L}_{adv}(D) = \mathcal{L}_{imd}(D) + \mathcal{L}_{icd}(D). \tag{19}$$

3) Simplified Gradient Penalty: The gradient penalty performed between the input distributions is used to force the discriminator to meet the Lipschiz constraint [14]. It is usually performed in the space between the input distributions according to WGAN-GP [32]. In CMPD, there are three joint distributions for two discriminator networks, which yield four time-consuming penalty terms to be calculated over P_1, P_2 and P_3 during training. Therefore, we consider to combine the DRAGAN [44] and the L2-norm to reduce the calculation redundance. To reduce the computation redundance, by applying L2-norm, we project all the vectors to a unit hyper-sphere in the common representation subspace, restricting all the distributions close to each other. Under such circumstances, one gradient penalty for one discriminator can cover all the distributions on the hypersphere. In practice, gradient penalty is applied over P_1, which is shared in both $\mathcal{L}_{imd}$ and $\mathcal{L}_{icd}$. Then, Eq. ((13)) and Eq. ((16)) are reformulated according to DRAGAN [44] as

$$\mathcal{P}_{imd} = \mathbb{E}_{x \sim P_1}[(\|\nabla_x D_{imd}(x)\|_2 - 1)^2], \tag{20}$$

$$\mathcal{P}_{icd} = \mathbb{E}_{x \sim P_1}[(\|\nabla_x D_{icd}(x)\|_2 - 1)^2]. \tag{21}$$

D. Supervised Loss and Total Loss Function

1) Supervised Loss: Since the joint distribution matching is mainly proposed for mining the semantic relationships between unpaired items, constraints between paired items are still necessary for constructing the semantic structure in the common representation subspace. In this paper, we follow [7] to adopt the class prediction loss and the triplet loss as the supervised optimization goals for generator.

For class prediction, we add a classifier consisting of the feed-forward network activated by softmax on top of the representation generators $f_{\mathcal{V}}$ and $f_{\mathcal{T}}$, respectively. The corresponding objective function is

$$\mathcal{L}_{cla}(G) = \frac{1}{m}\sum_{i=1}^{m}(y_i \cdot (\log p_i(\tilde{\boldsymbol{v}}_i) + \log p_i(\tilde{\boldsymbol{t}}_i))), \tag{22}$$

where y_i is 1 for correct class label and 0 for wrong class label, respectively.

For triplet loss, we follow [7] to sample negative items that have different class label with a given text-image pair from the mini-batch to construct the triplets $(\tilde{\boldsymbol{v}}_i^c, \tilde{\boldsymbol{t}}_i^c, \tilde{\boldsymbol{t}}_j^d)$ and $(\tilde{\boldsymbol{t}}_i^c, \tilde{\boldsymbol{v}}_i^c, \tilde{\boldsymbol{v}}_j^d)$. The objective function is formulated as

$$\mathcal{L}_{tri,\mathcal{T}}(G) = \sum_{i,j,c,k}(\tilde{\boldsymbol{v}}_i^c \cdot \tilde{\boldsymbol{t}}_i^c + \max(0, \mu - \tilde{\boldsymbol{v}}_i^c \cdot \tilde{\boldsymbol{t}}_j^k)), \tag{23}$$

$$\mathcal{L}_{tri,\mathcal{V}}(G) = \sum_{i,j,c,k}(\tilde{\boldsymbol{t}}_i^c \cdot \tilde{\boldsymbol{v}}_i^c + \max(0, \mu - \tilde{\boldsymbol{t}}_i^c \cdot \tilde{\boldsymbol{v}}_j^k)), \tag{24}$$

where μ is a preset threshold for training.

Algorithm 1 CMPD, Our Proposed Algorithm. All Experiments in the Paper Used the Default Values Learning Rate $\alpha_D = 0.0005$ for Discriminator and $\alpha_G = 0.0001$ for Generators, Batch Size $N = 64$, and Iteration Steps for Discrminator $n_D = 3$

Require: Pre-trained image feature vectors $\{\boldsymbol{v}_i^c\}_{i=1}^n$, text feature vectors $\{\boldsymbol{t}_i^c\}_{i=1}^n$ and hyperparameters λ_{adv}, λ_{icd}, λ_{gp} and λ_{tri} for objective functions;

while model has not converged **do**
  for $t = 0, \ldots, n_D$ **do**
    update the discriminator's parameters θ_D to obtain the estimation of the inter-modal and inter-class divergence:
    $grad_D \leftarrow \nabla_D \frac{1}{N}\mathcal{L}(D)$
    $w \leftarrow \theta_D + \alpha_D \cdot \mathrm{Adam}(\theta_D, grad_D)$
  end for
  update the generator's parameters θ_G according to the estimation of the discriminator and the supervised optimization goals:
  $grad_G \leftarrow \nabla_G \frac{1}{N}\lambda(G)$
  $\theta_G \leftarrow \theta_G + \alpha_G \cdot \mathrm{Adam}(\theta_G, grad_G)$
end while

Thus, we have the total supervised objective functions for generator as

$$\mathcal{L}_{sup}(G) = \mathcal{L}_{cla}(G) + \lambda_{tri}[\mathcal{L}_{tri,\mathcal{V}}(G) + \mathcal{L}_{tri,\mathcal{T}}(G)]. \quad (25)$$

The optimization of our proposed model basically consists of two parts. First, the discriminator is optimized for several steps to obtain the estimation of divergences between distributions, then the generator is optimized according to the estimated divergence along with the supervised optimization goals. The total optimization goals for generator and discriminator are given below

$$\mathcal{L}(G) = \lambda_{adv}\mathcal{L}_{adv}(G) + \mathcal{L}_{sup}(G), \quad (26)$$

$$\mathcal{L}(D) = \mathcal{L}_{adv}(D). \quad (27)$$

The pseudo code of the training process is shown in Algorithm 1.

IV. Experiments and Evaluation

In this section, we first evaluate the performance of our proposed CMPD on four widely-used cross-modal datasets: Wikipedia dataset [45], NUS-WIDE-10k dataset [46], Pascal Sentence dataset [47], and MSCOCO dataset [48]. Then, we further conduct a visualization analysis on our proposed model and discuss the effect of some key parameters.

A. Experiment Setup

We compare the performance of our model with other existing methods in terms of the mean average precision (mAP). The evaluation is conducted on both directions, i.e. retrieving text using image query (Img2txt) and retrieving image using text query (Txt2img). For fair comparison, we follow [7], [22] to compute top-50 mAP for the results on each dataset.

For the representation generators, *relu* activation is applied to the bottom two layers and *tanh* activation is applied to the top layer. The learning rate is fixed to 10^{-4} for both text and image generators in all the experiments. The number of feature representations maintained in cross memory unit is set to 128. The label classifier contains only one fully-connected layer with no activation function. The weight factor λ_{adv} and λ_{tri} are set to 1 and 0.01 for all the experiments, respectively.

The discriminator consists of three feed-forward layers with *tanh* activation for the bottom two layers and no activation for the output layer. The number of the output units from the bottom to the top layer is set to 64, 32 and 1, respectively. The learning rate for discriminator is fixed to 5×10^{-4} and the weight factor λ_{gp} for gradient penalty is set to 10. To train the model, we use Adam Optimizer [49] with hyper parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$.

B. Quantitative Comparison

In this subsection, we comprehensively compare our proposed CMPD with three kinds of retrieval methods: (1) traditional methods including CCA [35], CCA-3V [36], LSFS [37], JRL [38] and JFSSL [27]; (2) DNN-based methods including Multi-DBN [39], Bimodal-AE [40], Corr-AE [21], DCML [41] and CMDN [42]; (3) GAN-based methods including UCAL [29], CM-GAN [30], MHTN [31], ACMR [7], CMST [43] and DSCMR [13]. Note that **CMPD(Ours)** is the results obtained under the above-mentioned circumstances, while **CMPD(Ours-textcnn)** is an additional version for fair comparison with DSCMR [13], which will be detailed later. The experimental results are shown in Table I, and the observations can be concluded as follows. The experiments are carried out on four widely-used cross-modal retrieval datasets, which are listed in below.

- **Wikipedia dataset** [45] has 2866 image-text pairs labelled by 10 categories. For fair comparison, the dataset is split into 2 subsets, which are training set with 1300 pairs and testing set with 1566 pairs, following [7].
- **Pascal Sentence dataset** [47] consists of 1000 image-text pairs with 20 categories. For each image in the dataset, there are 5 corresponding short sentence descriptions. The whole dataset is divided into 2 groups, namely 800 for training and 200 for testing.
- **NUS-WIDE-10K dataset** [46] is a dataset that contains 9000 image-text pairs with 350 categories. We select 8000 pairs from the dataset for training and leaving the 1000 pairs for testing.
- **MSCOCO dataset** [48] is a large scale dataset compared to the previous three datasets. It contains 123,000 images, with each images annotated by 5 descriptions. Following [7], we use the 66,226 image-text pairs for training, and 16,557 image-text pairs for testing, which all come from the training set of MSCOCO.

In experiments, we exactly follow the partition and feature extraction strategies of [7] for fair comparison. Image features are taken from the fc7 layer of a pre-trained VGGNet-19 (VGG19) model while text features are computed by the classical Bag-of-Words features with tf-idf weighting. The experiments results are shown in Table I, and the observations can be concluded as follows.

TABLE I
COMPARISON WITH EXISTING IMAGE-TEXT RETRIEVAL METHODS IN ASPECT OF MEAN AVERAGE PRECISION (MAP)

Category	Methods	Wikipedia			Pascal Sentences			NUS-WIDE-10K		
		Img2txt	Txt2Img	Avg.	Img2txt	Txt2Img	Avg.	Img2txt	Txt2Img	Avg.
Traditional methods	CCA [35]	0.267	0.222	0.245	0.363	0.219	0.291	0.189	0.188	0.189
	CCA-3V [36]	0.437	0.383	0.410	0.316	0.270	0.293	-	-	-
	LCFS [37]	0.455	0.398	0.427	0.442	0.357	0.400	0.383	0.346	0.365
	JRL [38]	0.453	0.400	0.426	0.504	0.489	0.496	0.426	0.376	0.401
	JFSSL [27]	0.428	0.396	0.412	-	-	-	-	-	-
DNN-based methods	Multi-DBN [39]	0.204	0.183	0.194	0.477	0.424	0.458	0.201	0.259	0.230
	Bimodal-AE [40]	0.314	0.290	0.302	0.456	0.470	0.458	0.327	0.369	0.348
	Corr-AE [21]	0.402	0.395	0.398	0.489	0.444	0.467	0.366	0.417	0.392
	DCML [41]	0.554	0.538	0.546	-	-	-	0.385	0.405	0.395
	CMDN [42]	0.488	0.427	0.458	0.534	0.534	0.534	0.492	0.515	0.504
GAN-based methods	UCAL [29]	0.263	0.273	0.268	0.448	0.325	0.387	-	-	-
	CM-GAN [30]	0.521	0.466	0.494	0.603	0.604	0.604	-	-	-
	MHTN [31]	0.514	0.444	0.479	0.496	0.500	0.498	0.520	0.534	0.527
	ACMR [7]	0.619	0.489	0.554	0.535	0.543	0.539	0.544	0.538	0.541
	CMST [43]	0.632	0.505	0.569	0.621	0.586	0.604	**0.628**	0.562	0.595
	DSCMR [13]	0521	0.478	0.499	0.710	0.722	0.716	0.611	0.615	0.613
	CMPD (Ours)	**0.641**	**0.513**	**0.577**	0.606	0.652	0.629	0.581	**0.668**	**0.625**
	CMPD (Ours-textcnn)	-	-	-	**0.713**	**0.736**	**0.725**	-	-	-

TABLE II
COMPARISON WITH EXISTING METHODS ON MSCOCO IN ASPECT OF mAP

Models	MSCOCO		
	Img2txt	Txt2Img	Avg.
CCA (FV HGLMM)[36]	0.791	0.765	0.778
CCA (FV GMM+HGLM)[36]	0.809	0.766	0.788
DVSA [51]	0.777	0.748	0.777
m-RNN [52]	0.835	0.770	0.803
m-CNN [53]	0.841	0.828	0.835
DSPE [26]	0.892	0.869	0.881
ACMR [7]	0.932	0.871	0.902
CMPD(Ours)	**0.942**	**0.932**	**0.937**

On Wikipedia dataset, the proposed CMPD outperforms the state-of-the-art method ACMR by 1.3% and 1.2% in Img2txt and Txt2Img retrieval tasks, and it achieves a more convincing improvement (9.2% in Img2txt and 3.7% in Txt2Img) on the NUS-WIDE-10K dataset. We believe there are two reasons for that CMPD achieves a better results on NUS-Wide-10K. First, because NUS-WIDE-10K contains a larger scale of data, CMPD can be fully optimized during training. Second, benefited from the larger number of classes contained in the NUD-WIDE-10K, joint distribution matching's advantage of learning the semantic relationships between unpaired items is further magnified. On Wikipedia and NUS-Wide-10K, the effectiveness of ACMR model comes from the element-wise triplet loss and the adversarial learning framework. Compared to CMPD, the adversarial learning in ACMR only learns the semantic relationships of unpaired items between modalities. It fails to capture the more detailed relationship between classes, which is the advantage of joint distribution matching.

For Pascal Sentence dataset, our model outperforms the state-of-the-art method of CM-GANs by 0.3% and 4.8% in Img2txt and Txt2Img retrieval task, respectively. We believe the our CMPD beats CM-GANs due to two reasons. First, CM-GANs use the weight sharing constraints to force the two generators to learn the shared common features of image and text. On the contrary, the cross memory unit used in our CMPD is a more complicated mechanism which automatically learns a combination of shared and modality-specific feature. Second, the reconstruction loss introduced in CM-GANs is still a constraint that considers the semantic relationships between paired items, although it can learn a better feature representations compared to element-wise constraints. The experimental results of CM-GANs and CM-Net justify the importance of learning relationships between unpaired items for constructing the semantic structure in the common representation subspace, as well as the effectiveness of joint distribution matching and cross memory unit proposed in this work.

We specially note that, the results of DSCMR is quoted from its original paper, in which the input text feature is extracted using the TextCNN [50] instead of simple BOW method. Even though, our CMPD can still outperform the DSCMR on Wikipedia and NUS-WIDE-10K dataset. And we use the preprocessed data publicized by DSCMR on Pascal Sentence to obtain the results of CMPD(Ours-textcnn), which also outperforms the DSCMR. Compared to CMPD(Ours), the higher performance of CMPD(Ours-textcnn) on Pascal Sentence proves the better quality of TextCNN features. Therefore, the better results of CMPD on NUS-WIDE and Wikipedia dataset compared to DSCMR is much more convincing, because our CMPD achieves better performance using only BOW text features. This comparison result further proves the superiority of CMPD over DSCMR.

In Table II, we show the performance of CMPD on MSCOCO, which is a large scale cross-model retrieval dataset. Note that since MSCOCO is an unlabeled dataset, we follow [7] to evaluate CMPD on the MSCOCO dataset. We quote the mAP results obtained by the baselines as refer to [7].

C. Model Analysis

1) Analysis of Cross Memory Block: In Table V, we analyze the influence of the gate structure in CMB and the influence

TABLE III
THE ANALYSIS OF JOINT DISTRIBUTION DIVERGENCE (JD) AND CROSS MEMORY UNITS (CMU) ON THE PROPOSED CMPD MODEL (FULL)

Models	Wikipedia			Pascal			NUS-WIDE-10K		
	Img2txt	Txt2Img	Avg.	Img2txt	Txt2Img	Avg.	Img2txt	Txt2Img	Avg.
baseline	0.602	0.480	0.546	0.536	0.540	0.538	0.540	0.532	0.536
CMPD-JD	0.632	0.501	0.567	0.593	0.626	0.610	0.575	0.636	0.606
CMPD-CMU	0.604	0.510	0.559	0.564	0.614	0.589	0.552	0.641	0.597
CMPD-Full	**0.641**	**0.513**	**0.577**	**0.606**	**0.652**	**0.629**	**0.581**	**0.668**	**0.625**

TABLE IV
EFFECTS OF THE DIFFERENT λ_{icd} IN TERMS OF THE MAP ON THE WIKIPEDIA, PASCAL AND NUS-WIDE-10K

λ_{icd}	Wikipedia			Pascal			NUS-WIDE-10K		
	Img2txt	Txt2Img	Avg.	Img2txt	Txt2Img	Avg.	Img2txt	Txt2Img	Avg.
10	0.622	0.496	0.559	0.593	0.627	0.610	0.558	0.601	0.580
1.	0.621	0.491	0.556	0.593	0.612	0.603	0.551	0.616	0.584
0.1	**0.641**	**0.513**	**0.577**	**0.606**	**0.652**	**0.629**	**0.581**	**0.668**	**0.625**

TABLE V
ANALYSIS OF CROSS MEMORY BLOCK

Models	Wikipedia		
	Img2txt	Txt2Img	Avg.
raw-input	0.590	0.489	0.540
no-gate	0.620	0.509	0.565
original	**0.641**	**0.513**	**0.577**

of CMS's position. We design two different variations of CMPD. The *raw-input* variation removes the feed-forward layers before CMB and directly applies CMB on the raw input of image and text feature. The *no-gate* variation removes the gate structure in CMB, and directly add the shared feature representation $\boldsymbol{m}_s$ in Eq. (4) with $\boldsymbol{v}'_l$ in Eq. (2). And the *original* model denotes the original version of CMPD, which serves as the baseline for comparison. The best performance is achieved by the original version, which proves the effectiveness of preprocessing the raw input feature with feed-forward layers and the gate structure in the CMB module.

2) Ablation Study: CMPD is mainly realized and improved as a combination of two parts, the cross memory unit based generator and the joint distribution matching constraints. In order to further investigate more detailed effect of each part, we developed and evaluated two different variations of our CMPD model: (1) **CMPD-JD**, the model with joint distribution divergence only. The cross memory unit is removed from the feed-forward based network. (2) **CMPD-CMU**, the model with cross memory unit only. The joint distribution divergence is removed from the model. The constraints only consist of the triplet loss and the classification loss. Besides, we also provide a **baseline** model from which both joint distribution matching and cross memory unit are removed. All of the training strategies and the hyper-parameter settings are the same as the full CMPD model (**CMPD-Full**). In Table III, we show the performance of the above four models on the Wikipedia, Pascal and NUS-Wide-10K dataset in terms of mAP.

Table III shows that both CMPD-JD and CMPD-CMU outperform the baseline model, which proves the effectiveness of joint distribution matching and cross memory unit, respectively. By combining both of the two parts, the CMPD-Full achieves the best results over three datasets, which demonstrates that both of the two parts contribute to the final retrieval performance.

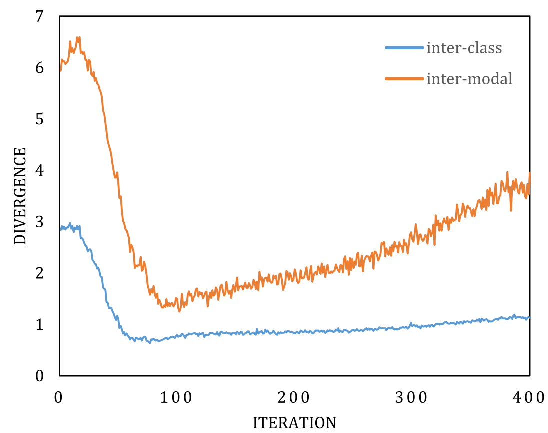


Fig. 3. The training curves for inter-modal divergence and inter-class divergence.

3) Visualization of the Learned Representations: To explore the influence of balance factor λ_{icd} on the proposed CMPD, as well as demonstrate the effectiveness of joint distribution matching, we use t-SNE to visualize the representation vectors learned by our CMPD on the Wikipedia dataset, as shown in Fig. 4. The top row is colored according to the class labels, and the bottom row is colored according to the modality.

From the left to the right, we first show the visualization results of representation vectors learned with balance factor $\lambda_{icd} = 10$, 1 and 0.1, respectively. The smaller balance factor is, the less influence of the inter-class divergence has on the model. For comparison, we also provide the visualization results of model trained without joint distribution divergences (Without JD) in Fig. 4. The basic experiment settings are the same as full CMPD.

The visualization results demonstrate that the semantic structure in the common subspace learned by CMPD is quite sensitive to the balance factor λ_{icd}. A large λ_{icd} tends to split the image and text distributions apart from each other.

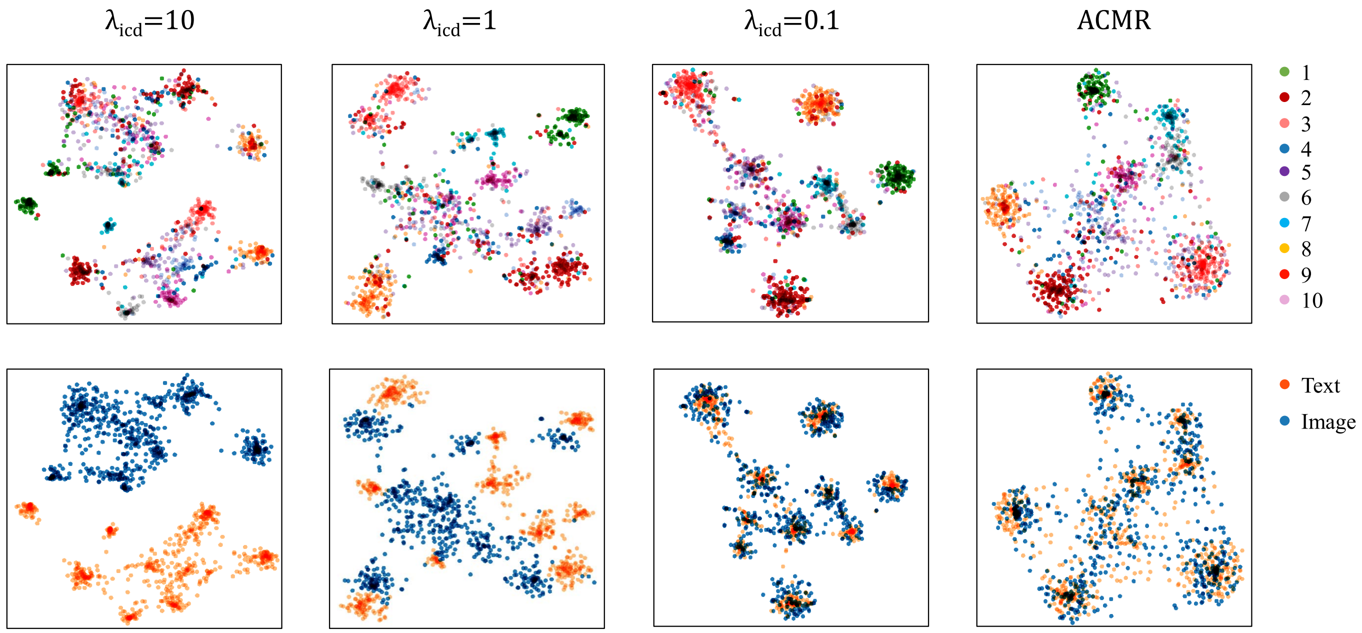


Fig. 4. The visualization results on Wikipedia dataset. Top row is colored according to the class label and the bottom row is colored according to the modality label.

TABLE VI
THE ABLATION STUDY FOR THE NUMBER OF THE CMPD UNITS

Number of units	Wikipedia			Pascal			NUS-WIDE-10K		
	Img2txt	Txt2Img	Avg.	Img2txt	Txt2Img	Avg.	Img2txt	Txt2Img	Avg.
16	0.623	0.476	0.550	0.573	0.607	0.590	0.517	0.569	0.543
32	0.632	0.506	0.569	0.576	0.611	0.594	0.558	0.591	0.575
64	**0.641**	0.513	**0.577**	0.606	**0.652**	**0.629**	**0.581**	0.668	**0.625**
128	0.640	**0.514**	0.577	**0.607**	0.632	0.620	0.575	**0.670**	0.622

It is a side effect for learning all the semantic relationships across different class labels as a whole. For $\lambda_{icd} = 0.1$, the model provides the best results. It tends to merge the items' distributions very tightly both in terms of modality and class labels. Compared to the model trained without joint distribution matching, our model establishes a more discriminative semantic structure in terms of the class labels, where the items of the same classes are gathered more tightly and the items with the different classes are pulled further.

In Table IV, we also show the accuracy of image-text retrievals in Wikipedia dataset, Pascal and NUS-WIDE-10K dataset in terms of different balance factors. From Table IV, we find that the best result of our model is achieved by $\lambda_{icd} = 0.1$, which is in accordance with the visualization results shown in Fig. 4.

4) Analysis of Joint Distribution Divergences: In this subsection, we will go into the training process to explore the interactions and effects between the two divergences. As shown in Fig. 3, we plot the training curves of the inter-class and the inter-modal divergences.

The training process can be divided into two stages according to the plotted curves. In the first stage, the inter-modal divergence plays the key role, bringing the image and the text distributions together in the common subspace. As a side effect, the inter-class divergence also decreases during this stage, because the inter-modal divergence will only bring the items with the same modality and class label together, regardless of the semantic structure between different classes. In the second stage of training, the inter-class divergence begins to work and distribute the items that have different class labels away from each other. We can find that in this stage, the inter-modal divergence stops falling and shows a slightly upward trend, which is the side effect of optimizing the inter-class divergence. But overall, the inter-modal divergence remains relatively stable during the second stage and successfully preserves the inter-modal structure to a certain degree when the inter-class structure is constantly shaped during training.

5) Analysis of Cross Memory Unit: In Table VI, we show our model's performance under different number of cross memory units. We find the model achieves the best results at 64 cross memory units. The performance of the model drops quickly when the number of units decreases, while increasing the number of units from 64 to 128 does not yield a better result. The explanations lie in two-folds: (1) for small number of units, the representational ability of the network is limited which cannot learn enough information from the input; (2) for large number of units, there may exist information redundancy which can hinder the network distinguish the learned features of different classes.

V. CONCLUSION

In this paper, we present a novel model for image-text retrieval, named CMPD. Our main work is to introduce the cross memory unit to share the learned features between image and text generators, and optimize the model using a novel joint distribution matching constraint for learning the inter-modal and the inter-class relationships. This joint distribution matching idea effectively resolve the element-wise constraint's issues which cannot preserve the relationships between unpaired items in current methods. The experiments on four widely used datasets demonstrate that CMPD can learn better representations and achieve the state-of-the-art results in class-level image-text retrieval tasks.

REFERENCES

[1] Q. Huang, A. Puri, and Z. Liu, "Multimedia search and retrieval: New concepts, system implementation, and application," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 5, pp. 679–692, Aug. 2000.

[2] M. Katsurai, T. Ogawa, and M. Haseyama, "A cross-modal approach for extracting semantic relationships between concepts using tagged images," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1059–1074, Jun. 2014.

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

[4] K. Ding, B. Fan, C. Huo, S. Xiang, and C. Pan, "Cross-modal hashing via rank-order preserving," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 571–585, Mar. 2017.

[5] A. Araujo and B. Girod, "Large-scale video retrieval using image queries," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 6, pp. 1406–1420, Jun. 2018.

[6] A.-A. Liu, W.-Z. Nie, and Y.-T. Su, "3D object retrieval based on multi-view latent variable model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 3, pp. 868–880, Mar. 2019.

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

[8] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, "Triplet-based deep hashing network for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3893–3903, Aug. 2018.

[9] D. Hu, F. Nie, and X. Li, "Deep binary reconstruction for cross-modal hashing," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 973–985, Apr. 2019.

[10] E. Chang, K. Goh, G. Sychay, and G. Wu, "CBSA: Content-based soft annotation for multimodal image retrieval using Bayes point machines," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 1, pp. 26–38, Jan. 2003.

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

[12] C. Li, C. Deng, N. Li, W. Liu, X. Gao, and D. Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4242–4251.

[13] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep supervised cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10394–10403.

[14] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*. [Online]. Available: http://arxiv.org/abs/1701.07875

[15] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, and X. Gao, "Pairwise relationship guided deep hashing for cross-modal retrieval," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 1618–1625.

[16] X. Xu, F. Shen, Y. Yang, H. T. Shen, and X. Li, "Learning discriminative binary codes for large-scale cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2494–2507, May 2017.

[17] Y. Cao, M. Long, J. Wang, and S. Liu, "Collective deep quantization for efficient cross-modal retrieval," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 3974–3980.

[18] D. Wang, Q. Wang, and X. Gao, "Robust and flexible discrete hashing for cross-modal similarity search," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2703–2715, Oct. 2018.

[19] Z.-D. Chen, C.-X. Li, X. Luo, L. Nie, W. Zhang, and X.-S. Xu, "SCRATCH: A scalable discrete matrix factorization hashing framework for cross-modal retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 2262–2275, Jul. 2020.

[20] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[21] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2014, pp. 7–16.

[22] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3441–3450.

[23] J. Gu, J. Cai, S. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7181–7189.

[24] R. Hong, Y. Yang, M. Wang, and X.-S. Hua, "Learning visual semantic relationships for efficient visual retrieval," *Technol. Behav. Data*, vol. 1, no. 4, pp. 152–161, Dec. 2015.

[25] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, and W. Lu, "Supervised coupled dictionary learning with group structures for multi-modal retrieval," in *Proc. AAAI Conf. Artif. Intell.*, 2013, pp. 1070–1076.

[26] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5005–5013.

[27] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2010–2023, Oct. 2016.

[28] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[29] L. He, X. Xu, H. Lu, Y. Yang, F. Shen, and H. T. Shen, "Unsupervised cross-modal retrieval through adversarial learning," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 1153–1158.

[30] Y. Peng and J. Qi, "CM-GANs: cross-modal generative adversarial networks for common representation learning," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 1, pp. 1–24, Feb. 2019.

[31] X. Huang, Y. Peng, and M. Yuan, "MHTN: Modal-adversarial hybrid transfer network for cross-modal retrieval," *IEEE Trans. Cybern.*, vol. 50, no. 3, pp. 1047–1059, Mar. 2020.

[32] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 5767–5777.

[33] A. Hasnat, J. Bohne, J. Milgram, S. Gentric, and L. Chen, "DeepVisage: Making face recognition simple yet with powerful generalization skills," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 1682–1691.

[34] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 212–220.

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

[36] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling Internet images, tags, and their semantics," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, 2014.

[37] K. Wang, R. He, W. Wang, L. Wang, and T. Tan, "Learning coupled feature spaces for cross-modal matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2088–2095.

[38] X. Zhai, Y. Peng, and J. Xiao, "Learning cross-media joint representation with sparse and semisupervised regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 965–978, Jun. 2014.

[39] N. Srivastava and R. Salakhutdinov, "Learning representations for multimodal data with deep belief nets," in *Proc. Int. Conf. Mach. Learn. Workshop*, 2012, pp. 1–8.

[40] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[41] V. E. Liong, J. Lu, Y.-P. Tan, and J. Zhou, "Deep coupled metric learning for cross-modal matching," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1234–1244, Jun. 2017.

[42] Y. Peng, X. Huang, and J. Qi, "Cross-media shared representation by hierarchical learning with multiple deep networks," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 3846–3853.

[43] X. Wen, Z. Han, X. Yin, and Y.-S. Liu, "Adversarial cross-modal retrieval via learning and transferring single-modal similarities," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 478–483.

[44] N. Kodali, J. Abernethy, J. Hays, and Z. Kira, "On convergence and stability of GANs," 2017, *arXiv:1705.07215*. [Online]. Available: http://arxiv.org/abs/1705.07215

[45] J. Costa Pereira *et al.*, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 521–535, Mar. 2014.
[46] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world Web image database from National University of Singapore," in *Proc. ACM Int. Conf. Image Video Retr. (CIVR)*, 2009, pp. 1–9.
[47] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's mechanical turk," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics*, 2010, pp. 139–147.
[48] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[50] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1746–1751.
[51] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3128–3137.
[52] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (M-RNN)," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–17.
[53] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2623–2631.

Xin Wen (Graduate Student Member, IEEE) received the B.S. degree in engineering management from Tsinghua University, China, in 2012, where he is currently pursuing the Ph.D. degree with the School of Software. His research interests include deep learning, shape analysis and pattern recognition, and NLP.

Zhizhong Han received the Ph.D. degree from Northwestern Polytechnical University, China, in 2017. He is currently a Post-Doctoral Researcher with the Department of Computer Science, University of Maryland at College Park, College Park, USA. He is also a Research Member of the BIM Group, Tsinghua University, China. His research interests include machine learning, pattern recognition, feature learning, and digital geometry processing.

Yu-Shen Liu (Member, IEEE) received the B.S. degree in mathematics from Jilin University, China, in 2000, and the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2006. From 2006 to 2009, he was a Post-Doctoral Researcher with Purdue University. He is currently an Associate Professor with the School of Software, Tsinghua University. His research interests include shape analysis, pattern recognition, machine learning, and semantic search.