Coupled CycleGAN: Unsupervised Hashing Network for Cross-Modal Retrieval

Chao Li,[1] **Cheng Deng,**[1*] **Lei Wang,**[1] **De Xie,**[1] **Xianglong Liu**[2†]
[1]School of Electronic Engineering, Xidian University, Xi'an 710071, China
[2]State Key Lab of Software Development Environment, Beihang University, Beijing 100191, China
li_chao@stu.xidian.edu.cn, {chdeng.xd, lwang.xidian, xiede.xd}@gmail.com, xlliu@nlsde.buaa.edu.cn

Abstract

In recent years, hashing has attracted more and more attention owing to its superior capacity of low storage cost and high query efficiency in large-scale cross-modal retrieval. Benefiting from deep leaning, continuously compelling results in cross-modal retrieval community have been achieved. However, existing deep cross-modal hashing methods either rely on amounts of labeled information or have no ability to learn an accuracy correlation between different modalities. In this paper, we proposed Unsupervised coupled Cycle generative adversarial Hashing networks (UCH), for cross-modal retrieval, where outer-cycle network is used to learn powerful common representation, and inner-cycle network is explained to generate reliable hash codes. Specifically, our proposed UCH seamlessly couples these two networks with generative adversarial mechanism, which can be optimized simultaneously to learn representation and hash codes. Extensive experiments on three popular benchmark datasets show that the proposed UCH outperforms the state-of-the-art unsupervised cross-modal hashing methods.

Introduction

Multi-modal data, such as text and image, exhibit heterogeneous properties, making it difficult for users to search for information of interest effectively and efficiently. Cross-modal retrieval, which aims to search the images (resp. texts) that are relevant to a given text (resp. image) query, has been studied in the past decade (Zhang and Wang 2016; Yang et al. 2017b; Gu et al. 2018). However, most of these methods suffer from high computation burden because of large volumes and high dimensions of multimedia data. Hashing based cross-modal retrieval, mapping multi-modal data into compact binary codes and conducting retrieval via fast bit-wise XOR operation, has become a hot topic. The fundamental challenges in cross-modal hashing retrieval lie in capturing the correlation between different modalities and learning reliable hash codes.

Cross-modal hashing methods can be generally categorized into two groups: supervised and unsupervised. Supervised hashing methods (Liu et al. 2017; Deng et al. 2016; Wu et al. 2014; Masci et al. 2014; Yang et al. 2017a; Li et al. 2018) learn hash codes under some supervision such as label information. However, label information collection consumes massive time and labor, making it infeasible in real-world applications. Different from supervised scenario, unsupervised hashing methods (Bronstein et al. 2010; Kumar and Udupa 2011; Feng, Wang, and Li 2014; Yu, Liu, and Shao 2017; Yang et al. 2018b) without using semantic labels, usually depend on the criterion of maintaining the original relationship among the training data.

Recently, deep cross-modal hashing methods, leveraging the superior capacity of deep neural network (Krizhevsky, Sutskever, and Hinton 2012; Goodfellow et al. 2014; Zhu et al. 2017) to capture the correlation between different modalities, have illustrated that deep cross-modal hashing methods are usually more effective than shallow structure based counterparts. Deep Cross-Modal Hashing (DCMH) (Jiang and Li 2017), Triplet based Deep Hashing (TDH) (Deng et al. 2018), Shared Predictive Deep Quantization (SPDQ) (Yang et al. 2018a), and Self-Supervised Adversarial Hashing (SSAH) (Li et al. 2018) are reported recently to encode individual modalities into their corresponding features by constructing two different pathways in deep networks. SPDQ constructs two specific network layers to learn modality-common and modality-private representations. DCMH and SSAH learn hash codes by preserving semantic correlation between different modalities which is constructed by label information. However, almost all of these supervised methods require amounts of label information. Compared with them, unsupervised deep hashing methods learn the modality correlation depending on correspondences between pairs of data (such as image-text pairs), making it more feasible in real word settings. Among these unsupervised deep hashing methods, Unsupervised Generative Adversarial Cross-modal Hashing (UGACH) (Zhang, Peng, and Yuan 2017) exploits graph manifold structure to learn modality correlation. Unsupervised Deep Cross-Modal Hashing (UDCMH) (Wu et al. 2018) equips with adaptive learning strategy to learn hash codes iteratively. Even so, there are still some common disadvantages hindering these methods. First, due to plentiful semantically-similar pairs existing in real datasets, these methods just simply adopt individual network to extract features for each modality, which cannot build accurate correlation between different modali-

*Corresponding author
†Corresponding author

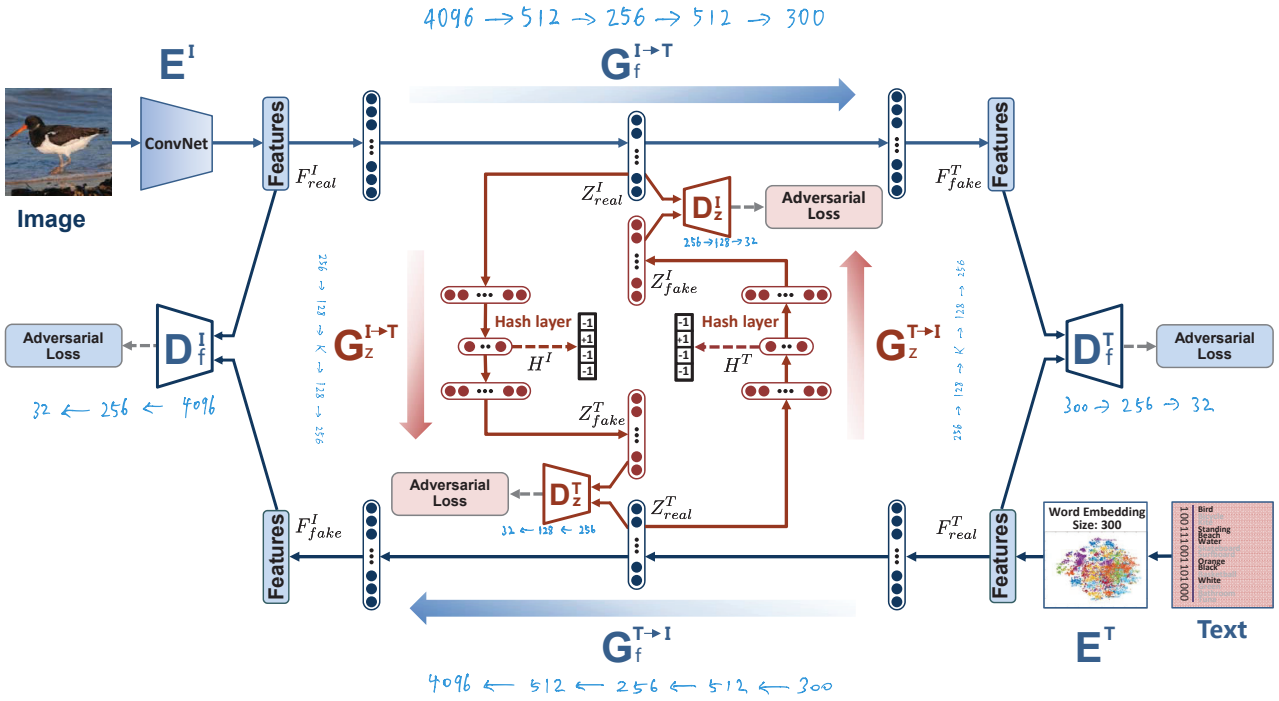

Figure 1: The proposed unsupervised coupled cycle generative adversarial deep cross-modal hashing learning framework. The main framework mainly consists of two modules: representation learning (in color blue) and hashing learning (in color red). The representation learning is constructed with E^I, E^T, $G_f^{I\to T}$, $G_f^{T\to I}$, D_f^I, and D_f^T, where $G_f^{I\to T}$, D_f^T, $G_f^{T\to I}$, and D_f^I format the outer-cycle GAN. $G_z^{I\to T}$, D_z^T, $G_z^{T\to I}$, and D_z^I, formating the inner-cycle GAN, are combined into the hashing learning.

ties. Second, most of these methods, separating hash codes generation from common representation learning, greatly decreases accuracy of the learned hash codes.

In this paper, we propose a novel unsupervised coupled cycle generative adversarial hashing network called UCH, for cross-modal retrieval. Specifically, we devise pair-coupled generative adversarial networks to build two cycle networks in an unified framework, where outer-cycle network is constructed to learn powerful representations for individual modality data to capture accurate correlation between different modalities and inner-cycle network is designed to generate reliable compact hash codes depending on the learned representations. The highlights of our work can be summarized as follows:

- We design an unsupervised coupled cycle generative adversarial hashing network for cross-modal retrieval, where powerful common representation and reliable hash codes can be learned simultaneously in an unified framework.
- By utilizing the proposed coupled cycle networks, common representation and hash codes learning can interact with each other and reach the optimum when network is convergence in the meanwhile.
- Extensive experiments conducted on three popular benchmark datasets demonstrate that our proposed UCH outperforms the state-of-the-art unsupervised cross-modal hashing methods.

The rest of this paper starts with a review of the most related work. Then the proposed unsupervised coupled cycle generative adversarial deep cross-modal hashing method is presented in detail, which is followed by the experiments and conclusion.

Related Work

In this section, the most related work on topic of unsupervised cross-modal hashing methods are reviewed, both traditional shallow structure based methods and deep structure based methods.

For shallow structure based unsupervised cross-modal hashing methods, co-occurrence information is uniformly utilized to exploit modality correlation since image-text pair that occurs simultaneously tend to be of similar semantic. In (Rasiwasia et al. 2010), canonical component analysis (CCA) is adopted to map image-text pairs into a latent space where their similarities can be measured. Cross-view hashing (CVH) (Kumar and Udupa 2011) considers the cross-modality similarity preservation during hashing learning process. In (Rastegari et al. 2013), Predictable Dual-View Hashing (PDH) learns hashing function in virtue of self-taught learning algorithm. In Collective Matrix Factorization Hashing (CMFH) (Ding, Guo, and Zhou 2014), unified hash codes are learned by collective matrix factorization from different views. Latent Semantic Sparse Hashing (LSSH) (Zhou, Ding, and Guo 2014) is proposed to jointly learn latent features from images and texts with sparse coding. In Fusion Similarity Hashing (FSH) (Liu et al. 2017), modality correlation is captured by fusing similarity across

different modalities.

For deep structure based unsupervised cross-modal hashing methods, deep networks pretrained on other public dataset are usually utilized to predict individual modality representation and hash codes further can be learned based on the achieved representation. UGACH presents that a similarity relationship matrix depending on its extracted representation is first constructed and then modality correlation is learned by exploiting manifold structure between different modalities. However, there are two problems indwelling in UGACH. One is that an accurate similarity matrix cannot be obtained through only one calculation, the other is that time consuming in calculating similarity matrix cannot be ignored. UDCMH is another representative unsupervised deep hashing method, which applies matrix factorization on the extracted deep features to generate hash codes and then uses the obtained hash codes as supervision to update feature extractor network. However, hash codes usually have short bit representations, whose supervised information is insufficient to learn discriminative representations for exploiting the modality correlation. In contrast, our UCH can effectively build modality correlation by directly learning bidirectional transformation between different modalities. Additionally, with the proposed coupled cycle framework, hash codes can be updated iteratively, which are more reliable.

Proposed UCH

Fig. 1 shows the overview of our proposed UCH for cross-modal retrieval, which mainly consists of two modules: representation learning (in color blue) and hashing learning (in color red). Specifically, the representation learning part consists of the E^I, E^T, $G_f^{I\to T}$, $G_f^{T\to I}$, D_f^I, and D_f^T, where $G_f^{I\to T}$, D_f^T, $G_f^{T\to I}$, and D_f^I form a generative adversarial cycle network, namely outer-cycle GAN. By using E^I and E^T, real image features F_{real}^I and real text features F_{real}^T are extracted from original image and text. To learn powerful cross-modal common representations Z_{real}^I and Z_{real}^T, each modality data is first mapped into a shared common space, and then to generate different modality data. Additionally, hashing learning part consisting of $G_z^{I\to T}$, D_z^T, $G_z^{T\to I}$, and D_z^I are combined to form a generative adversarial cycle network, namely inner-cycle GAN. To further build the projection from original image and text to hash codes, we feed the learned Z_{real}^I and Z_{real}^T into $G_z^{I\to T}$ and $G_z^{T\to I}$ to generate Z_{fake}^T and Z_{fake}^I recorded as fake image and fake text representation. Two hashing layers are devised within the inner-cycle GAN to generate hash codes directly. The inner-cycle GAN and outer-cycle GAN are trained jointly. Overall, through the outer-cycle GAN network of cross-modal feature learning, we hope to learn powerful cross-modal common representations, and through the inner-cycle GAN network of hashing learning, we hope to learn reliable hash codes. In the following, we first give some definitions and then introduce the proposed method in detail.

Problem Definition

Let $O = \{o_i\}_{i=1}^n$ denote one cross-modal dataset, $o_i = (v_i, t_i)$ is an image-text pair, where v_i and t_i are raw visual and textual information describing the ith instance. The goal of cross-modal hashing is to generate reliable hash codes for image and text: $B^* \in \{-1, 1\}^{n\times K}, * = \{v, t\}$, where K is the length of binary codes. In our UCH, the outputs of hashing layers are defined as $H^*, * \in \{v, t\}$ for image and text respectively, binary hash codes B^* are generated by applying a sign function to H^*:

$$B^* = sign(H^*), * \in \{v, t\}, \tag{1}$$

Representation Learning

Our goal is to obtain reliable hash codes for image and text. However, in view of the huge modality gap between two modalities caused by their heterogeneous structures, modality correlation cannot be learned directly without using any supervision information. To build the relationship between different modalities, we construct outer-cycle GAN with two generator networks for image and text, making it possible to generate image (resp. text) modality data from text (resp. image) modality data. By training these two networks, powerful individual representation can be learned for different modalities and thus modality gap can be bridged effectively. Given image-text pair: v and t, F_{real}^I and F_{real}^T are extracted through $E^I(v, \theta^I)$ and $E^t(t, \theta^T)$, where θ^I and θ^T are network parameters. Within the outer-cycle GAN, the generation can be formulated as:

$$\begin{aligned} F_{fake}^T &= G_f^{I\to T}(F_{real}^I, \eta^{I\to T}), \\ F_{fake}^I &= G_f^{T\to I}(F_{real}^T, \eta^{T\to I}), \end{aligned} \tag{2}$$

where $G_f^{I\to T}$ and $G_f^{T\to I}$ are two generation functions, $\eta^{I\to T}$ and $\eta^{T\to I}$ are network parameters.

Moreover, to judge the quality of the generative data, we further devise two discriminators $D_f^I(\cdot, \mu^I)$ and $D_f^T(\cdot, \mu^T)$ to provide adversarial loss for generators, where μ^I and μ^T are network parameters. For different modalities image and text, their individual adversarial losses $\mathcal{L}_{adv\_f}$ can be written as:

$$\mathcal{L}_{adv\_f} = \mathcal{L}_{adv\_f}^I + \mathcal{L}_{adv\_f}^T, \tag{3}$$

where $\mathcal{L}_{adv\_f}^I$ and $\mathcal{L}_{adv\_f}^T$ are formulated as follows:

$$\begin{aligned} \mathcal{L}_{adv\_f}^I = \min_{\eta^{T\to I}} \max_{\mu^I} \; & E_{F_{real}^I \sim P(F_{real}^I)} \left[\log D_f^I(F_{real}^I)\right] \\ & + E_{F_{fake}^I \sim P(F_{fake}^I)} \left[\log(1 - D_f^I(F_{fake}^I))\right], \end{aligned} \tag{4}$$

$$\begin{aligned} \mathcal{L}_{adv\_f}^T = \min_{\eta^{I\to T}} \max_{\mu^T} \; & E_{F_{real}^T \sim P(F_{real}^T)} \left[\log D_f^T(F_{real}^T)\right] \\ & + E_{F_{fake}^T \sim P(F_{fake}^T)} \left[\log(1 - D_f^T(F_{fake}^T))\right]. \end{aligned} \tag{5}$$

Furthermore, to improve the generative capacity of the outer-cycle GAN networks, we feed the generated F_{fake}^T and F_{fake}^I back into the generative networks to reconstruct the original data and minimize the reconstruction loss, which

can be formulated as:

$$\begin{aligned}\mathcal{L}_{rec\_f} &= \min\ \mathcal{L}^I_{rec\_f} + \mathcal{L}^T_{rec\_f} \\ &= \min \sum_{i=1}^{n} E_{F^I_{real}\sim P(F^I_{real})}\left[\|F^I_{real} - G^{T\to I}(F^T_{fake})\|_2^2\right] \\ &\quad + \sum_{i=1}^{n} E_{F^T_{real}\sim P(F^T_{real})}\left[\|F^T_{real} - G^{I\to T}(F^I_{fake})\|_2^2\right].\end{aligned} \tag{6}$$

Finally, we hope to generate similar representations Z^I_{real} and Z^T_{real} for semantically-similarity paired cross-modal data. l_2 loss function is adopted in the shared common space to regulate the learned representation to be similar. Thus, the similarity loss can be achieved by:

$$\mathcal{L}_{sim\_f} = \min \sum_{i=1}^{n} \|Z^I_{real} - Z^T_{real}\|_2^2. \tag{7}$$

As to the representation learning, the whole loss function to learn powerful individual representation for different modalities is made up of adversarial loss, reconstruction loss, and similarity loss, which can be defined as:

$$\mathcal{L}_f = \mathcal{L}_{adv\_f} + \mathcal{L}_{rec\_f} + \mathcal{L}_{sim\_f}. \tag{8}$$

Through training the outer-cycle GAN to minimize (8), discriminative representations can be achieved.

Hashing Learning

With the similar idea that different modality data can be inherited by the common representations in representation learning part, we learn hash codes within the inner-cycle GAN constructed by $G_z^{I\to T}(\cdot, \xi^{I\to T})$, $D_z^T(\cdot, \varepsilon^T)$, $G_z^{T\to I}(\cdot, \xi^{T\to I})$, and $D_z^I(\cdot, \varepsilon^I)$, where $\xi^{I\to T}$, ε^T, $\xi^{T\to I}$, and ε^I are network parameters. Similarly, given Z^I_{real} and Z^T_{real}, the generation can be formulated as:

$$\begin{aligned} Z^T_{fake} &= G_z^{I\to T}(Z^I_{real}, \xi^{I\to T}) \\ Z^I_{fake} &= G_z^{T\to I}(Z^T_{real}, \xi^{T\to I}). \end{aligned} \tag{9}$$

Further, the adversarial loss for hashing learning $\mathcal{L}_{adv\_z} = \mathcal{L}^I_{adv\_z} + \mathcal{L}^T_{adv\_z}$, where $\mathcal{L}^I_{adv\_z}$ and $\mathcal{L}^T_{adv\_z}$ are designed as :

$$\begin{aligned}\mathcal{L}^I_{adv\_z} = \min_{\xi^{T\to I}} \max_{\varepsilon^I}\ & E_{Z^I_{real}\sim P(Z^I_{real})}\left[\log D^I_z(Z^I_{real})\right] \\ & + E_{Z^I_{fake}\sim P(Z^I_{fake})}\left[\log(1 - D^I_z(Z^I_{fake}))\right].\end{aligned} \tag{10}$$

$$\begin{aligned}\mathcal{L}^T_{adv\_z} = \min_{\xi^{I\to T}} \max_{\varepsilon^T}\ & E_{Z^T_{real}\sim P(Z^T_{real})}\left[\log D^T(Z^T_{real})\right] \\ & + E_{Z^T_{fake}\sim P(Z^T_{fake})}\left[\log(1 - D^T(Z^T_{fake}))\right].\end{aligned} \tag{11}$$

Next, reconstruction loss $\mathcal{L}_{rec\_z}$ is constructed as:

$$\begin{aligned}\mathcal{L}_{rec\_z} &= \min\ \mathcal{L}^I_{rec\_z} + \mathcal{L}^t_{rec\_z} \\ &= \min \sum_{i=1}^{n} E_{Z^I_{real}\sim P(Z^I_{real})}\left[\|Z^I_{real} - G^{T\to I}(Z^T_{fake})\|_2^2\right] \\ &\quad + \sum_{i=1}^{n} E_{Z^T_{real}\sim P(Z^T_{real})}\left[\|Z^T_{real} - G^{I\to T}(Z^I_{fake})\|_2^2\right].\end{aligned} \tag{12}$$

Algorithm 1 Optimizing process of the proposed UCH

Require: Image-Text pairs dataset;
Ensure: Optimal code matrix B
Initialization
Initialize network parameters θ^*, μ^*, ε^*, $\eta^{I\to T}$, $\eta^{T\to I}$, $\xi^{I\to T}$, and $\xi^{T\to I}$, where $* = \{v, t\}$, batch size: $N^{I,T} = 128$, maximum iteration number: T_{max}.
repeat
  Update μ^I and μ^T by (16) with BP algorithm;
  Update θ^I, θ^T, $\eta^{I\to T}$, and $\eta^{T\to I}$ by (17) with BP algorithm;
  Update ε^I and ε^I by (18) with BP algorithm;
  Update $\xi^{I\to T}$ and $\xi^{T\to I}$ by (19) with BP algorithm;
  Update hash codes matrix B by $B = sign(H^I + H^T)$;
until maximum iteration number T_{max}.

Finally, similarity loss $\mathcal{L}_{sim\_z}$ is utilized to promote our UCH to produce uniformed hash codes, which can be written as:

$$\mathcal{L}_{sim\_z} = \min \sum_{i=1}^{n} \|H^I - H^T\|_2^2. \tag{13}$$

Therefore, the final loss for hashing learning within inner-cycle GAN can be written as:

$$\mathcal{L}_z = \mathcal{L}_{adv\_z} + \mathcal{L}_{rec\_z} + \mathcal{L}_{sim\_z}. \tag{14}$$

Overall, the whole loss for the entire framework mainly consists of two parts: representation learning loss and hashing learning loss, which is written as:

$$\mathcal{L}_{Total} = \mathcal{L}_f + \mathcal{L}_z. \tag{15}$$

The whole alternating learning algorithm for the proposed UCH is briefly outlined in Algorithm 1.

Optimization

Considering the gradient vanishing problem caused by the minimax loss for generative adversarial networks, we optimize our UCH separately.

In order to guarantee the learning efficiency of the proposed UCH, the outer-cycle GAN is trained to produce powerful representations firstly. Update μ^I and μ^T, with the other parameters fixed:

$$\mu^I, \mu^T = \arg\max_{\mu^I, \mu^T} \mathcal{L}_{adv\_f}. \tag{16}$$

Update θ^I, θ^T, $\eta^{I\to T}$, and $\eta^{T\to I}$ by minimizing (8) with the other parameters fixed:

$$\theta^I, \theta^T, \eta^{I\to T}, \eta^{T\to I} = \arg \min_{\substack{\theta^I, \theta^T \\ \eta^{I\to T}, \eta^{T\to I}}} \mathcal{L}_{rec\_f} + \mathcal{L}_{sim\_f}. \tag{17}$$

Then, train the inner-cycle GAN to generate reliable hash codes. Update ε^I and ε^T with other parameters fixed:

$$\varepsilon^I, \varepsilon^T = \arg\max_{\varepsilon^I, \varepsilon^T} \mathcal{L}_{adv\_z}. \tag{18}$$

Finally, update $\xi^{I\to T}$ and $\xi^{T\to I}$ with other parameters fixed:

$$\xi^{I\to T}, \xi^{T\to I} = \arg \min_{\xi^{I\to T}, \xi^{T\to I}} \mathcal{L}_{rec\_z} + \mathcal{L}_{sim\_z}. \tag{19}$$

Table 1: Comparison in terms of MAP scores of two retrieval tasks on MIRFlickr-25K, IAPR TC-12, and COCO datasets with different lengths of hash codes. The best accuracy are shown in boldface.

Task	Method	MIRFlickr-25K			IAPR TC-12			COCO		
		16	32	64	16	32	64	16	32	64
Image Query v.s. Text Database	CVH	0.596	0.585	0.582	0.397	0.385	0.377	0.474	0.475	0.459
	STMH	0.575	0.582	0.625	0.374	0.387	0.400	0.405	0.414	0.406
	LSSH	0.589	0.604	0.634	0.443	0.456	0.460	0.468	0.473	0.485
	FSH	0.580	0.583	0.591	0.400	0.420	0.422	0.453	0.489	0.494
	CMFH	0.601	0.605	0.610	0.442	0.450	0.463	0.475	0.490	0.524
	CMSSH	0.598	0.593	0.600	0.390	0.386	0.377	0.503	0.516	0.517
	SCM	0.635	0.636	0.639	0.402	0.412	0.419	0.442	0.461	0.486
	OURS	**0.654**	**0.669**	**0.679**	**0.447**	**0.471**	**0.485**	**0.521**	**0.534**	**0.547**
Text Query v.s. Image Database	CVH	0.604	0.592	0.577	0.405	0.393	0.384	0.470	0.474	0.456
	STMH	0.586	0.594	0.632	0.381	0.406	0.429	0.391	0.422	0.449
	LSSH	0.583	0.588	0.601	0.409	0.416	0.421	0.456	0.460	0.463
	FSH	0.575	0.576	0.583	0.412	0.429	0.444	0.471	0.509	0.514
	CMFH	0.627	0.636	0.639	0.435	0.445	0.456	0.468	0.478	0.509
	CMSSH	0.563	0.571	0.566	0.385	0.383	0.393	0.431	0.434	0.469
	SCM	0.652	0.657	0.659	0.437	0.450	0.458	0.426	0.442	0.461
	OURS	**0.661**	**0.667**	**0.668**	**0.446**	**0.469**	**0.488**	**0.499**	**0.519**	**0.545**

Experiments

Datasets

Three popular benchmark datasets in cross-modal retrieval: *MIRFlickr-25K* (Huiskes and Lew 2008), *NUS-WIDE* (Chua et al. 2009), and *Microsoft COCO* (Lin et al. 2014) are adopted to validate our proposed method.

MIRFlickr-25K dataset consists of 25,000 images collected from Flickr website. Each image is associated with multiple textual tags and manually annotated with at least one of the 24 unique labels. Leaving out data without labeled information, 20,015 image-text pairs are used in our experiment totally, where 2,000 image-text pairs are randomly selected as query set and the rest are regarded as retrieval set. The textual information for each image is represented as a 1386-dimensional bag-of-words vector. For supervised baselines, 5,000 image-text pairs are selected from retrieval set to construct training set.

IAPR TC-12 dataset contains 20,000 images with corresponding sentence descriptions. These image-text pairs present various semantics such as action and people categories. In our experiment, each sentence is represented as a bag-of-words vector with 2,000-dimension. We prune all images without any labeled information and finally 18,571 image-text pairs are remained. Each image-text pair belongs to the top 22 frequent labels from the 275 concepts. We randomly select 5,000 image-text pairs to construct training set for supervised methods and the rest are retrieval set.

Microsoft COCO dataset totally contains 82,783 training images and 40,504 validation images. Each image is described with five different sentences and labeled with at least one of 80 unique labels. In our experiment, we represent each sentence with a 2,000-dimension bag-of-words vector. After deleting the image-text pairs without any textual information, finally, 122,218 image-text pairs are left to formulate the dataset used in our experiment. 2,000 image-text pairs are randomly selected as query set and the left 120,218 pairs are regarded as retrieval set. 6,000 image-text pairs from retrieval set are randomly selected to construct training set for supervised methods. Additionally, it should be noted that all retrieval set are used as training set for all unsupervised methods.

Baselines and Evaluation

To illustrate the effectiveness of our proposed UCH, there are seven state-of-the-art cross-modal hashing baselines compared in experiments, including five unsupervised methods CVH (Kumar and Udupa 2011), CMFH (Ding, Guo, and Zhou 2014), STMH (Wang et al. 2015), LSSH (Zhou, Ding, and Guo 2014), and FSH (Liu et al. 2017) and two supervised methods CMSSH (Bronstein et al. 2010) and SCM (Zhang and Li 2014), which are all shallow structure based cross-modal retrieval hashing methods. We additionally compare our UCH with the recently proposed UGACH (Zhang, Peng, and Yuan 2017), which is a deep structure based cross-modal retrieval hashing method. For fair comparison, we use the same deep network CNN-F to extract deep features for all shallow structure based methods.

Following the previous methods, three common used protocols: Mean Average Precision (MAP), Precision-Recall curves (PR curves), and Precision of the top N curves (Precision@N), are adopted to evaluate the retrieval performance of all methods in our experiments.

Implementation Details

In this subsection, we will introduce the implementation details about the proposed UCH and settings of some parameters. UCH is implemented via *TensorFlow* and executed on a server with two NVIDIA TITAN X GPUs.

Representation Learning: $\mathbf{E^I}$, $\mathbf{E^T}$, $\mathbf{G_f^{I\to T}}$, $\mathbf{G_f^{T\to I}}$, $\mathbf{D_f^I}$,

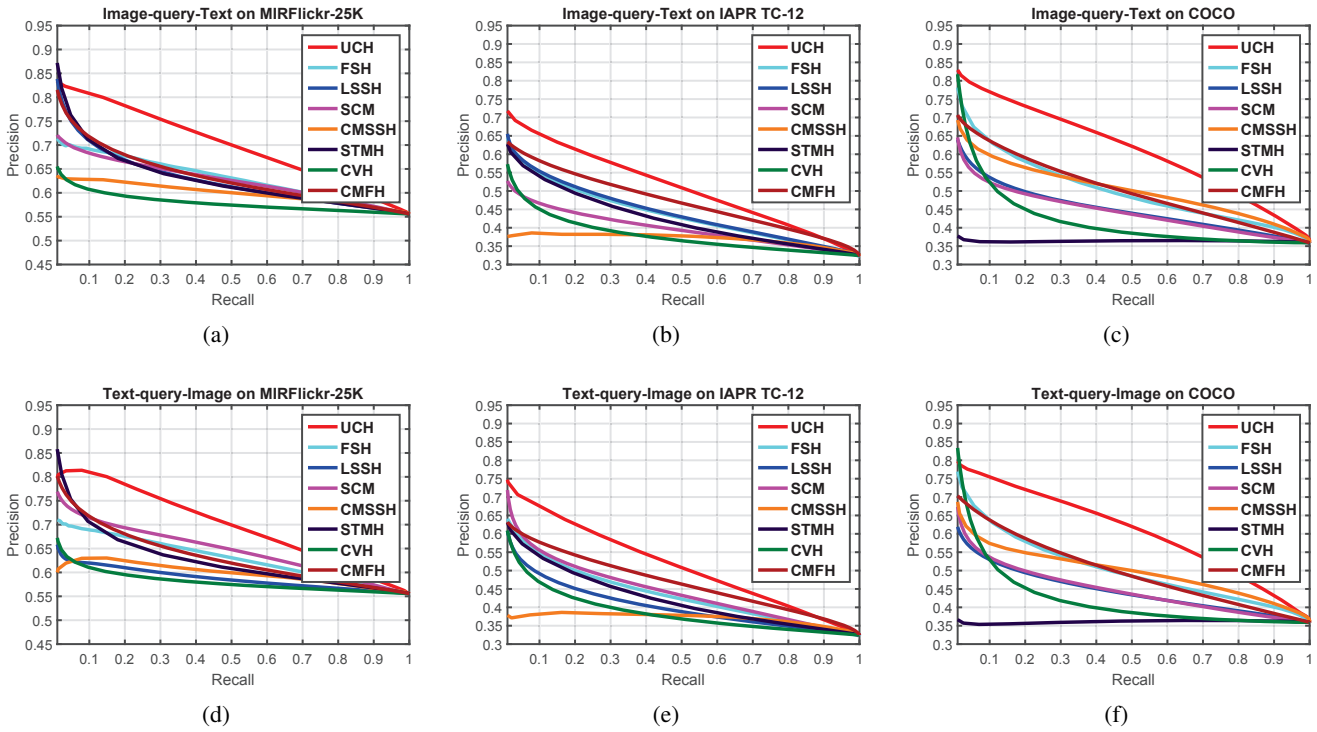

Figure 2: Precision-recall curves evaluated on MIRFlickr25k, IAPR TC-12, and COCO. The code length is 64.

and $\mathbf{D_f^T}$. Our UCH is a CNN-based method, where CNN-F (Chatfield et al. 2014) neural network pretrained on ImageNet dataset (Simonyan and Zisserman 2014) is adopted as image extractor. For image, the input raw image is resized into $224 \times 224 \times 3$ and feed into the CNN-F. We take the output of $fc7$ as real image features F^I_{real}. For text, we design an embedding network layer with 300 nodes following the text input layer to project raw bag-of-words vector into features with continuous values. We take the output of embedding layer as real text features F^I_{real}. $G^{I\to T}_f$ and $G^{T\to I}_f$ are constructed with two different deep networks with four full-connected layers, *e.g.*, ($G^{I\to T}_f$:$4096 \to 512 \to 256 \to 512 \to 300$ and $G^{T\to I}_f$: $300 \to 512 \to 256 \to 512 \to 4096$). F^I_{real} and F^I_{fake} are fed into $G^{I\to T}_f$ and F^T_{real} and F^T_{fake} are fed into $G^{T\to I}_f$. Meanwhile, two discriminator networks D^I_f and D^T_f are framed with two deep networks with two full-connected layers: D^I_f, *e.g.*, ($4096 \to 256 \to 32$) and D^T_f, *e.g.*, ($300 \to 256 \to 32$).

Hashing Learning: $\mathbf{G_z^{I\to T}}$, $\mathbf{G_z^{T\to I}}$, $\mathbf{D_z^I}$, and $\mathbf{D_z^T}$. $G^{I\to T}_z$ and $G^{T\to I}_z$ are uniformly framed with four full-connected layers, *e.g.*, ($256 \to 128 \to K \to 128 \to 256$), K is the queried code length. The discriminator networks D^I_z and D^T_z are framed with two full-connected layers, *e.g.*, ($256 \to 128 \to 32$). In all our experiments, the initial leaning rates of image and text networks are set to 10^{-4} and 10^{-2}. And batchsize and weight decay are set to 128 and 10^{-1}.

Experiment Results

Table 1 reports the MAP results for our proposed UCH and the compared the-state-of-art methods on MIRFlickr-25K, IAPR TC-12, and Microsoft COCO datasets. As shown in Table 1, we group these compared methods into two categories: supervised and unsupervised. From Table 1, some conclusions can be obtained: 1) Among the traditional methods, SCM and CMSSH depending on additional supervised information, achieve relatively good performance on both retrieval tasks on average. 2) CMFH, FSH, and LSSH can achieve comparable performance in general. 3) By comparing all these methods, our proposed UCH achieves the highest MAP score at all code lengths and significantly outperforms other competitors. This may be because that almost these compared methods learn common representations and hash codes dividually, which limits their retrieval accuracy. By contrast, the proposed unsupervised coupled cycle deep hashing network unifies common representation learning and hash codes learning together, which can be optimized in one framework. Therefore, more reliable hash codes can be achieved.

Additionally, we provide the Precision-Recall curves (PR curves) and Precision of the top 1,000 curves (Precision@1000) in Fig. 2 and Fig. 3, respectively. PR curve is obtained by varying the Hamming radius from 0 to 64 with a stepsize 1. Precision@1000 presents the precision for the top 1,000 retrieved instances. Fig. 2 shows PR curves of all state-of-the-art methods with 64-bit hash codes on three benchmark datasets. From Fig. 2 and Fig. 3, similar conclu-

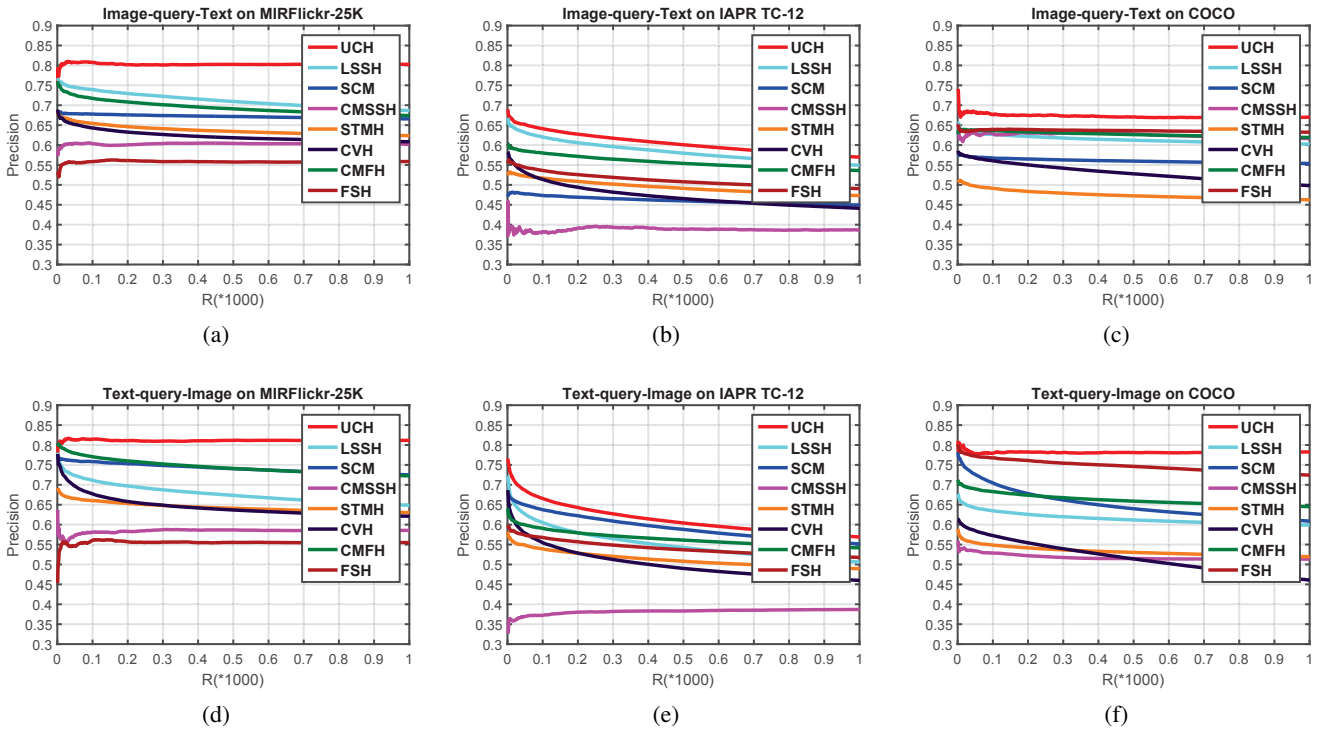

Figure 3: Precision curves with top 1,000 retrieved instances on MIRFlickr25k, IAPR TC-12, and COCO with 64 bits.

Table 2: Compared the proposed UCH with UGACH. MAP score evaluated on MIRFlickr25K.

Task	Method	MIRFlickr25K		
		16	32	64
Image→Text	UGACH	0.603	0.607	0.616
	UCH	**0.654**	**0.669**	**0.679**
Text→Image	UGACH	0.590	0.632	0.642
	UCH	**0.661**	**0.667**	**0.668**

Table 3: Compared the proposed UCH with UGACH. MAP score evaluated on IAPR TC-12.

Task	Method	IAPR TC-12		
		16	32	64
Image→Text	UGACH	0.439	0.454	0.479
	UCH	**0.447**	**0.471**	**0.485**
Text→Image	UGACH	0.433	0.456	0.480
	UCH	**0.446**	**0.469**	**0.488**

sions mentioned above can be achieved.

Comparison UCH with UGACH. We additionally compare our proposed UCH with UGACH, which is a representative deep learning based method proposed recently. For fair comparison, the same CNN-F network is adopted to extract deep features for UGACH. Table 2 and Table 3 show the results of comparison between UCH and UGACH in term of MAP values on MIRFlickr and IAPR TC-12 datasets. It is obvious that our proposed UCH outperforms UGACH with different code lengths in all cases. The main reason may be that UGACH just calculates relationship once among instances with deep features in the data preprocessing rather than update it iteratively, which causes that the built correlation is lack of accuracy and thus the retrieval performance is constrained. With no need to build similarity matrix, our UCH exploiting modality correlation by generating modality data bi-directionally between two different modalities can learn more powerful representations. Therefore, more reliable hash codes can be achieved with the proposed UCH.

Conclusions

In this paper, we proposed a novel unsupervised coupled cycle generative adversarial hashing network, for large-scale cross-modal retrieval. The uniqueness of our method is that powerful common representations and reliable hash codes can be learned in an unified framework without using any label information. Moreover, common representation learning and hashing learning, interacting with a coupled manner, can achieve optimal performance at the same time. The extensive experiments on three widely-used datasets show that our proposed model achieves state-of-the-art performance in cross-modal retrieval tasks.

Acknowledgments

Our work was supported in part by the National Natural Science Foundation of China under Grant 61572388 and 61703327, in part by the Key R&D Program-The Key Industry Innovation Chain of Shaanxi under Grant 2017ZDCXL-GY-05-04-02, 2017ZDCXL-GY-05-02 and 2018ZDXM-GY-176, and in part by the National Key R&D Program of China under Grant 2017YFE0104100.

References

[Bronstein et al. 2010] Bronstein, M. M.; Bronstein, A. M.; Michel, F.; and Paragios, N. 2010. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 3594–3601.

[Chatfield et al. 2014] Chatfield, K.; Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Return of the devil in the details: Delving deep into convolutional nets. *arXiv:1405.3531*.

[Chua et al. 2009] Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. Nus-wide: a real-world web image database from national university of singapore. In *Proc. ACM Conf. Image Video Retr.*, 48.

[Deng et al. 2016] Deng, C.; Tang, X.; Yan, J.; Liu, W.; and Gao, X. 2016. Discriminative dictionary learning with common label alignment for cross-modal retrieval. *IEEE Trans. Multimed.* 18(2):208–218.

[Deng et al. 2018] Deng, C.; Chen, Z.; Liu, X.; Gao, X.; and Tao, D. 2018. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Trans. Image Process.* 27(8):3893–3903.

[Ding, Guo, and Zhou 2014] Ding, G.; Guo, Y.; and Zhou, J. 2014. Collective matrix factorization hashing for multimodal data. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2083–2090.

[Feng, Wang, and Li 2014] Feng, F.; Wang, X.; and Li, R. 2014. Cross-modal retrieval with correspondence autoencoder. 7–16.

[Goodfellow et al. 2014] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Proc. Adv. Neural Inf. Process. Syst.*, 2672–2680.

[Gu et al. 2018] Gu, J.; Cai, J.; Joty, S.; Niu, L.; and Wang, G. 2018. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 7181–7189.

[Huiskes and Lew 2008] Huiskes, M. J., and Lew, M. S. 2008. The mir flickr retrieval evaluation. In *Proc. Conf. Multimedia Inf. Retrieval*, 39–43.

[Jiang and Li 2017] Jiang, Q.-Y., and Li, W.-J. 2017. Deep cross-modal hashing. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*

[Krizhevsky, Sutskever, and Hinton 2012] Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Proc. Adv. Neural Inf. Process. Syst.*, 1097–1105.

[Kumar and Udupa 2011] Kumar, S., and Udupa, R. 2011. Learning hash functions for cross-view similarity search. In *Proc. Int. Joint Conf. Artif. Intell.*, volume 22, 1360.

[Li et al. 2018] Li, C.; Deng, C.; Li, N.; Liu, W.; Gao, X.; and Tao, D. 2018. Self-supervised adversarial hashing networks for cross-modal retrieval. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 4242–4251.

[Lin et al. 2014] Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proc. 13th Eur. Conf. Comput. Vis.*, 740–755.

[Liu et al. 2017] Liu, H.; Ji, R.; Wu, Y.; Huang, F.; and Zhang, B. 2017. Cross-modality binary code learning via fusion similarity hashing. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*

[Masci et al. 2014] Masci, J.; Bronstein, M. M.; Bronstein, A. M.; and Schmidhuber, J. 2014. Multimodal similarity-preserving hashing. *IEEE Trans. Pattern Anal. Mach. Intell.* 36(4):824–830.

[Rasiwasia et al. 2010] Rasiwasia, N.; Costa Pereira, J.; Coviello, E.; Doyle, G.; Lanckriet, G. R.; Levy, R.; and Vasconcelos, N. 2010. A new approach to cross-modal multimedia retrieval. In *Proc. ACM Int. Conf. Multimedia*, 251–260. ACM.

[Rastegari et al. 2013] Rastegari, M.; Choi, J.; Fakhraei, S.; Hal, D.; and Davis, L. 2013. Predictable dual-view hashing. In *Proc. 30th Int. Conf. Mach. Learn.*, 1328–1336.

[Simonyan and Zisserman 2014] Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*.

[Wang et al. 2015] Wang, D.; Gao, X.; Wang, X.; and He, L. 2015. Semantic topic multimodal hashing for cross-media retrieval. In *Proc. Int. Joint Conf. Artif. Intell.*, 3890–3896.

[Wu et al. 2014] Wu, F.; Zhou, Y.; Yang, Y.; Tang, S.; Zhang, Y.; and Zhuang, Y. 2014. Sparse multi-modal hashing. *IEEE Trans. Multimed.* 16(2):427–439.

[Wu et al. 2018] Wu, G.; Lin, Z.; Han, J.; Liu, L.; Ding, G.; Zhang, B.; and Shen, J. 2018. Unsupervised deep hashing via binary latent factor models for large-scale cross-modal retrieval. In *Proc. Int. Joint Conf. Artif. Intell.*, 2854–2860.

[Yang et al. 2017a] Yang, E.; Deng, C.; Liu, W.; Liu, X.; Tao, D.; and Gao, X. 2017a. Pairwise relationship guided deep hashing for cross-modal retrieval. In *AAAI*, 1618–1625.

[Yang et al. 2017b] Yang, Y.; Deng, C.; Tao, D.; Zhang, S.; Liu, W.; and Gao, X. 2017b. Latent max-margin multitask learning with skelets for 3-d action recognition. *IEEE Trans. Cybern.* 47(2):439–448.

[Yang et al. 2018a] Yang, E.; Deng, C.; Li, C.; Liu, W.; Li, J.; and Tao, D. 2018a. Shared predictive cross-modal deep quantization. *IEEE Trans. Neural Netw. Learn. Syst.* (99):1–12.

[Yang et al. 2018b] Yang, E.; Deng, C.; Liu, T.; Liu, W.; and Tao, D. 2018b. Semantic structure-based unsupervised deep hashing. In *IJCAI*, 1064–1070.

[Yu, Liu, and Shao 2017] Yu, M.; Liu, L.; and Shao, L. 2017. Binary set embedding for cross-modal retrieval. *IEEE Trans. Neural Netw. Learn. Syst.* 28(12):2899–2910.

[Zhang and Li 2014] Zhang, D., and Li, W. J. 2014. Large-scale supervised multimodal hashing with semantic correlation maximization. In *Proc. 28th AAAI Conf. Artif. Intell.*, 2177–2183.

[Zhang and Wang 2016] Zhang, T., and Wang, J. 2016. Collaborative quantization for cross-modal similarity search. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*

[Zhang, Peng, and Yuan 2017] Zhang, J.; Peng, Y.; and Yuan, M. 2017. Unsupervised generative adversarial cross-modal hashing. *arXiv:1712.00358*.

[Zhou, Ding, and Guo 2014] Zhou, J.; Ding, G.; and Guo, Y. 2014. Latent semantic sparse hashing for cross-modal similarity search. In *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 415–424. ACM.

[Zhu et al. 2017] Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2242–2251. IEEE.