Attention-aware Deep Adversarial Hashing for Cross-Modal Retrieval

Xi Zhang[1,2][0000−0002−9173−4119], Hanjiang Lai[1,2 ⋆][0000−0001−8057−6744], and Jiashi Feng[3][0000−0001−6843−0064]

[1] School of Data and Computer Science, Sun Yat-Sen University, GuangZhou, China
[2] Guangdong Key Laboratory of Big Data Analysis and Processing
zhangx368@mail2.sysu.edu.cn, laihanj3@mail.sysu.edu.cn
[3] Department of Electrical and Computer Engineering,
National University of Singapore, Singapore, Singapore
elefjia@nus.edu.sg

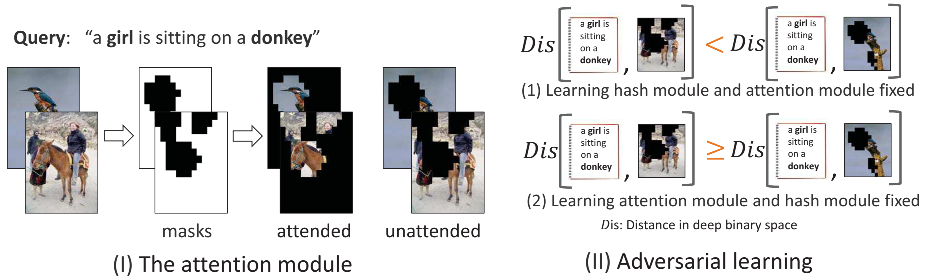
Abstract. Due to the rapid growth of multi-modal data, hashing methods for cross-modal retrieval have received considerable attention. However, finding content similarities between different modalities of data is still challenging due to an existing heterogeneity gap. To further address this problem, we propose an adversarial hashing network with an attention mechanism to enhance the measurement of content similarities by selectively focusing on the informative parts of multi-modal data. The proposed new deep adversarial network consists of three building blocks: 1) the feature learning module to obtain the feature representations; 2) the attention module to generate an attention mask, which is used to divide the feature representations into the attended and unattended feature representations; and 3) the hashing module to learn hash functions that preserve the similarities between different modalities. In our framework, the attention and hashing modules are trained in an adversarial way: the attention module attempts to make the hashing module unable to preserve the similarities of multi-modal data w.r.t. the unattended feature representations, while the hashing module aims to preserve the similarities of multi-modal data w.r.t. the attended and unattended feature representations. Extensive evaluations on several benchmark datasets demonstrate that the proposed method brings substantial improvements over other state-of-the-art cross-modal hashing methods.

Keywords: Hashing, Adversarial Learning, Attention Mechanism, Cross Modal Retrieval

1 Introduction

Due to the rapid development of the Internet, different types of media data are also growing rapidly, e.g., texts, images, and videos. Cross-modal retrieval, which takes one type of data as the query and returns the relevant data of

⋆ Corresponding author: Hanjiang Lai


Query: "a girl is sitting on a donkey"
masks
attended
unattended
Dis
a girl is sitting on a donkey
< Dis
(1) Learning hash module and attention module fixed
≥ Dis
(2) Learning attention module and hash module fixed
Dis: Distance in deep binary space

(I) The attention module

(II) Adversarial learning

Fig. 1. Attention-aware deep adversarial hashing. To learn the attention masks, we train the attention module and the hashing module in an adversarial way (II): (1) the hashing module learns to preserve the similarities of multi-modal data, while (2) the attention module attempts to generate attention masks that make the hashing module unable to preserve the similarities of the unattended features.

another type, is increasingly receiving attention since it is a natural way to search for multi-modal data. The solution methods can be roughly divided into two categories [33]: real-valued representation learning and binary representation learning. Because of the low storage cost and fast retrieval speed of the binary representation, we only focus on cross-modal binary representation learning (i.e., hashing [31, 17]) in this paper.

To date, various cross-modal hashing algorithms [41, 40, 8, 19, 15, 3, 36] have been proposed for embedding correlations among different modalities of data. In the cross-modal hashing procedure, feature extraction is considered the first step for representing all modalities of data, and then, these multi-modal features can be projected into a common Hamming space for future searches. Many methods [8, 40] use a shallow architecture for feature extraction. For example, collective matrix factorization hashing (CMFH) [8] and semantic correlation maximization (SCM) [40] use the hand-crafted features. Recently, deep learning has also been adopted for cross-modal hashing due to its powerful ability to learn good representations of data. The representative work of deep-network-based cross-modal hashing includes deep cross-modal hashing (DCMH) [15], deep visual-semantic hashing (DVSH) [3], pairwise relationship guided deep hashing (PRDH) [36], etc.

In parallel, the computational model of "attention" has drawn much interest due to its impressive result in various applications, e.g., image caption [34]. It is also desired for cross-modal retrieval problems. For example, as shown in Fig. 1, given a query *girl sits on donkey*, if we can locate the more informative regions in the image (e.g., the black regions), a higher degree of accuracy can be obtained. To the best of our knowledge, the attention mechanism has not been well-explored for cross-modal hashing.

In this paper, we propose an attention mechanism for cross-modal hashing. The model first decides where (i.e., which region of multi-modal data) it should attend to; then, the attended region should be favoured for retrieval. Based on this, an *attention module* is proposed to find the attended regions and a *hashing module* is to learn the similarity-preserving hash functions. In the attention module, the adaptive attention mask is generated for each data, which divides the data into attended and unattended regions. Ideally, well-learned attention masks should locate discriminative regions, which means that the unattended regions of data are uninformative and difficult to preserve the similarities. Hence, the attention module undergoes learning to make the hashing module unable to preserve the similarities of the unattended regions of data. However, the learned hash functions should preserve the similarities for both the attended (which can be viewed as easy examples) and unattended (hard examples) regions of data to enhance the robustness and performance. Thus, the hashing module undergoes learning to preserve the similarities of both the unattended and attended regions of data. Note that the attention module and the hashing module are trained in an adversarial way: the attention module attempts to find the unattended regions in which the hashing module fails to maintain the similarities, whereas the hashing module aims to preserve the similarities of the multi-modal data.

A new deep adversarial hashing for cross-modal retrieval is illustrated in Fig. 2. It consists of three major components: (1) a feature learning module that uses CNN or MLP to extract high level semantic representations for the multi-modal data; (2) an attention module that generates the adaptive attention masks and divides the feature representations into the attended and unattended feature representations; and (3) a hashing module that focuses on learning the binary codes for the multi-modal data. The adversarial retrieval loss and the cross-modal loss are proposed to obtain good attention masks and powerful hash functions.

The main contributions of our work are three-fold. First, we propose an attention-aware method for the cross-modal hashing problem. It is able to detect the informative regions of multi-modal data, which is helpful for identifying content similarities between different modalities of data. Second, we propose a deep adversarial hashing for learning effective attention masks and compact binary codes simultaneously. Third, we quantitatively evaluate the usefulness of attention in cross-modal hashing, and our method yields better performances in comparison with several state-of-the-art methods.

2 Related Work

2.1 Cross-modal Hashing

According to the utilized information for learning the common representations, cross-modal hashing can be categorized into three groups [33]: 1) the unsupervised methods [29], 2) the pairwise-based methods [21, 41] and 3) the supervised methods [39, 4]. The unsupervised methods only use co-occurrence information to learn hash functions for multi-modal data. For instance, cross-view hashing (CVH) [27] extends spectral hashing from uni-modal to multi-modal scenarios. The pairwise-based methods use both the co-occurrence information and similar/dissimilar pairs to learn the hash functions. Bronstein et al. [11] proposed cross-modal similarity sensitive hashing (CMSSH), which learns the hash functions to ensure that if two samples (with different modalities) are relevant/irrelevant, their corresponding binary codes are similar/dissimilar. The supervised methods exploit label information to learn more discriminative common representation. Semantic correlation maximization (SCM) [40] uses a label vector to obtain the similarity matrix and reconstruct it through the binary codes. Xu et al. [35] proposed discrete cross-modal hashing (DCH), which directly learns discriminative binary codes with the discrete constraints. Most of these works are based on hand-crafted features.

The deep learning with neural networks has shown that this approach can effectively discover the correlations across different modalities. The deep cross-modal hashing (DCMH) [15] integrates feature learning and hash-code learning into the same framework. Cao et al. [3] proposed deep visual-semantic hashing (DVSH), which utilizes both a convolutional neural network (CNN) and long short-term memory (LSTM) to separately learn the common representations for each modality. Pairwise relationship guided deep hashing (PRDH) [36] also adopts deep CNN models to learn feature representations and hash codes simultaneously.

2.2 Generative Adversarial Network

Recently, generative adversarial networks (GANs) [10] have received a lot of attention and achieved impressive results in various applications, including image-to-image translation [42], image generation [23, 1] and representation learning [24, 22]. GANs have also been applied to retrieval problem. IRGAN [32] is a recently proposed method for information retrieval, in which the generative retrieval focuses on predicting relevant documents and the discriminative retrieval focuses on predicting relevancy given a query document pair. IRGAN is designed for uni-modal retrieval. While we focus on cross-modal retrieval in this paper.

Very recently, Wang et al. [28] present an adversarial cross-modal retrieval (ACMR) method to seek an effective common subspace based on adversarial learning: the modality classifier distinguishes the samples in terms of their modalities, and the feature projector generates modality-invariant representations that confuse the modality classifier. Both the ACMR and the proposed method use the adversarial learning, the main difference is that ACMR seeks to learn common subspace for the multi-modal data, while the adversarial learning in the proposed method is tailored to explicitly handle the attention-aware networks for cross-modal hashing. In addition, the ACMR falls into the category of real-valued approaches, while our method belongs to binary approaches. Further, Li et al. [18] present a self-supervised adversarial hashing (SSAH) for cross-modal retrieval.

To the best of our knowledge, the attention mechanism has not been well-explored for cross-modal hashing. The attention mechanism has been proved to be very powerful in many applications, such as image classification [2], image caption [34], image question answering [38], video action recognition [25] and

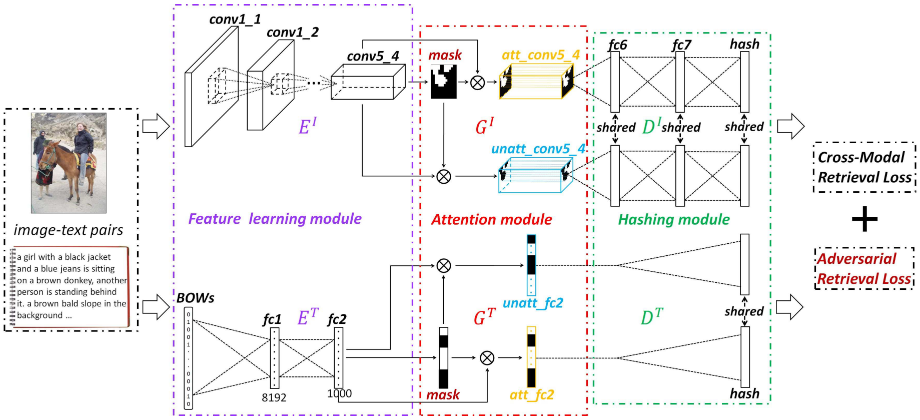


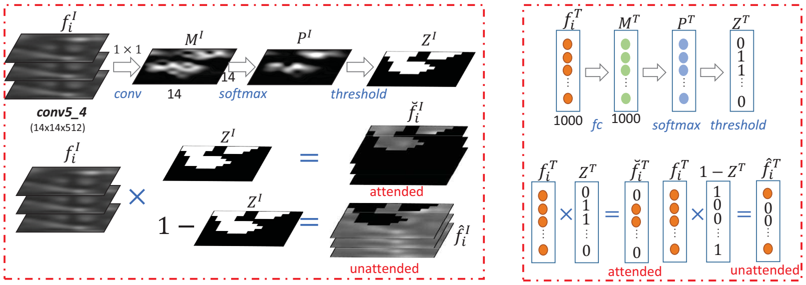
Fig. 2. Overview of our method. Above is the image modality branch, and below is the text modality branch. Each branch is divided into three parts: the feature learning module (including E^I and E^T), the attention module (G^I and G^T) and the hashing module (D^I and D^T). The feature learning module maps the input multi-modal data into the high-level feature representations. Then, the attention module learns the attention masks to divide the features representations into the attended and unattended features. Finally, the hashing module encodes all features into binary codes and learn similarity-preserving hash functions. We train the attention module and the hashing module alternately.

etc. Inspired by that, in this paper, we carefully design an attention-aware deep adversarial hashing network for cross-modal hashing.

3 Deep Adversarial Hashing for Cross-modal Retrieval

3.1 Problem Definition

Suppose there are n training samples, each of which is represented in several modalities, e.g., audio, video, image, and text. In this paper, we only focus on two modalities: text and image. Note that our method can be easily extended to other modalities. We denote the training data as $\{I_i, T_i\}_{i=1}^n$, where I_i is the i-th image and T_i is the corresponding text description of image I_i. We also have a cross-modal similarity matrix S, where $S(i,j) = 1$ means that the i-th image and the j-th text are similar, while $S(i,j) = 0$ means that they are dissimilar. The goal of cross-modal hashing is to learn two mapping functions to transform images and texts into a common binary codes space, in which the similarities between the paired images and texts are preserved. For instance, if $S(i,j) = 1$, the Hamming distance between the generated binary codes of the i-th image and the j-th text should be small. When $S(i,j) = 0$, the Hamming distance between them should be large.



(I) The attention module for image: G^I

(II) The attention module for text: G^T

Fig. 3. The attention module. It first generates the attention masks Z^I and Z^T. Then, each feature is divided into the attended and the unattended two parts.

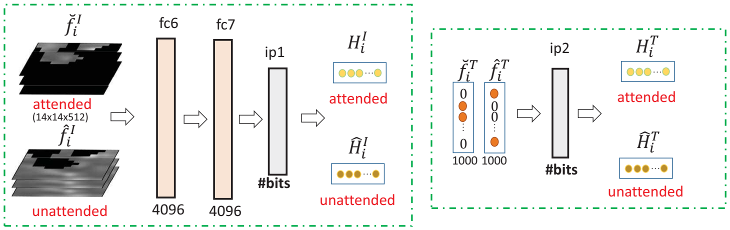
3.2 Network Architecture

The proposed deep adversarial hashing network contains three components: 1) the feature learning module to obtain the high-level representations of the multi-modal data; 2) the attention module to generate the attention masks, and 3) the hashing module to learn the similarity-preserving hash functions.

Feature Learning Module: E^I and E^T For the image modality, a convolutional neural network is used to obtain the high-level representation of images. Specifically, we use VGGNet [26] to extract the image feature maps, i.e., conv5_4 in VGGNet. For representing text instances, we use a well-known bag-of-words (BOW) vector. Then, we utilize the two-layer feed-forward neural network (BOW → 8192 → 1000) to obtain the semantic text features. Let $f_i^I = E^I(I_i)$ and $f_i^T = E^T(T_i)$ denote the image feature maps and the text feature vector, respectively.

Attention Module: G^I and G^T With the powerful image feature maps f^I and the text feature vector f^T, we first feed them into a one-layer neural network, i.e., a convolutional layer with a 1 × 1 kernel size for image feature maps and a fully connected layer for the text feature vector, followed by softmax and threshold functions to generate the attention distribution over the regions of the multi-modal data. Then, the attention masks are used to divide the feature representations into the attended and unattended feature representations.

More specifically, the detailed pipeline for processing the image modality is shown on the left side of Fig. 3. Suppose $f_i^I \in \mathbb{R}^{H \times W \times C}$ represents the feature maps for the i-th image, where H, W and C are the height, weight and channels, respectively. In the first step, we first use a convolutional layer to compress the feature maps f_i^I to a matrix $M_i^I = Conv(f_i^I)$, where $M_i^I \in \mathbb{R}^{H \times W}$. In the second step, the matrix M_i^I goes through a *softmax* layer, and the output is the probability matrix P_i^I. In the third step, we add a *threshold* layer to obtain the



(I) The hashing module for image: D^I

(II) The hashing module for text: D^T

Fig. 4. The hashing module for image modality D^I and text modality D^T.

attention mask, which is defined as

$$Z_i^I(h,w) = \begin{cases} 1 & P_i^I(h,w) \geq \alpha \\ 0 & P_i^I(h,w) < \alpha, \end{cases} \tag{1}$$

where α is a predefined threshold. We set $\alpha = \frac{1}{H \times W}$ in our experiment. The output of the threshold layer is a binary mask. Based on the binary mask, we can calculate the attended and unattended feature maps for the i-th image by multiplying the binary mask in element-wise, which is formulated as

$$\begin{aligned} \breve{f}_i^I(h,w,c) &= Z_i^I(h,w) \times f_i^I(h,w,c), \quad \textbf{(attended)} \\ \hat{f}_i^I(h,w,c) &= \Big(1 - Z_i^I(h,w)\Big) \times f_i^I(h,w,c), \quad \textbf{(unattended)} \end{aligned} \tag{2}$$

for all h, w and c. For ease of representation, we denote the whole procedures as $[\breve{f}_i^I, \hat{f}_i^I] = G^I(f_i^I)$.

For the text modality, we imitate the pipeline of the image modality, which is shown on the right hand of Fig. 3:

$$\begin{aligned} M_i^T &= \text{fc}(f_i^T), \\ P_i^T &= \text{softmax}(M_i^T), \\ Z_i^T &= \text{threshold}(P_i^T), \\ \breve{f}_i^T(j) &= Z_i^T(j) \times f_i^T(j), \quad \textbf{(attended)} \\ \hat{f}_i^T(j) &= \Big(1 - Z_i^T(j)\Big) \times f_i^T(j), \quad \textbf{(unattended)} \end{aligned} \tag{3}$$

where fc is a fully connected layer, and $Z(j)$ is the j-th value of the vector Z. We denote $[\breve{f}_i^T, \hat{f}_i^T] = G^T(f_i^T)$ as the attended and unattended features for the i-th text.

Directly taking the derivative of the *threshold* function is incompatible with the back-propagation in training. To address this issue, we follow the idea proposed in [7], which uses the straight-through estimator to propagate the gradients of the *threshold* function.

Hashing Module: $\boldsymbol{D^I}$ and $\boldsymbol{D^T}$ For the image modality, since we adopt VGGNet as our basic architecture, we also use the last fully connected layers, i.e., fc6 and fc7 [4]. Then, we add a fully connected layer with q dimensional features and a tanh layer that restricts the values in the range $[-1, 1]$ as shown on the left side of Fig. 4. Let the outputs of the discriminator be 1) the attended features $H_i^I = D^I(\breve{f}_i^I)$ and 2) the unattended features $\hat{H}_i^I = D^I(\hat{f}_i^I)$.

For the text modality, we also add a fully connected layer and a tanh layer to encode the text features into q bits as shown on the right side of Fig. 4. The outputs are 1) the attended features $H_i^T = D^T(\breve{f}_i^T)$ and 2) the unattended features $\hat{H}_i^T = D^T(\hat{f}_i^T)$.

3.3 Hashing Objectives

Our objectives contain two terms: 1) the cross-modal retrieval loss that corresponds to learning to preserve the similarities between different modalities of data and 2) the adversarial retrieval loss that corresponds to the hashing module aiming to preserve the similarities of the unattended binary codes, while the attention module tries to make the hashing module fails to maintain the similarities of the unattended binary codes.

Cross-modal Retrieval Loss The aim of the cross-modal loss function is to keep the similarities between images and texts. The inter-modal ranking loss and the intra-modal ranking loss are used to preserve the similarities. That is, the hash codes from different modalities should preserve the semantic similarities, and the hash codes from the same modality should also preserve the semantic similarities. Hence, the cross-modal retrieval loss can be formulated as

$$\min \mathcal{F}_{T \to I} + \mathcal{F}_{I \to T} + \mathcal{F}_{I \to I} + \mathcal{F}_{T \to T}, \tag{4}$$

where the first two terms are used to preserve the semantic similarities between different modalities, and the last two terms are used to preserve the similarities in their own modality. The symbol $A \to B$ denotes the A modality is taken as the query to retrieve the relevant data of the B modality, where $A \in \{T, I\}$ and $B \in \{T, I\}$. $\mathcal{F}_{A \to B}$ is the loss function for the A modality as the query and B modality as the database, which is defined as

$$\begin{aligned} \mathcal{F}_{A \to B} = & \sum_{\langle i,j,k \rangle} \max\{0, \varepsilon + ||H_i^A - H_j^B|| - ||H_i^A - H_k^B||\} \\ s.t. \qquad & \forall \langle i,j,k \rangle, \; S(i,j) > S(i,k), \end{aligned} \tag{5}$$

where $\langle i, j, k \rangle$ is the triplet form and ε is the margin. The objective is the triplet ranking loss [16], which shows effectiveness in the retrieval problem.

[4] The last fully connected layer (i.e., fc8) is removed since it is for classification problems.

Adversarial Retrieval Loss Inspired by the impressive results of the generative adversarial network, we adopt it to generate the attention distributions and learn the binary codes. Take the *text* → *image* as an example, which is also shown in Fig. 1. Given a query H_i^T, the hashing and the attention modules are trained in an adversarial way: 1) the hashing module preserves the semantic similarity between the query and the unattended features of the image modality, that is H_i^T is closer to $\hat{H}_j^I$ than to $\hat{H}_k^I$ when $S(i,j) > S(i,k)$; 2) the attention module tries to find the unattended regions of the images in which the hashing module fails to preserve the similarities, that is H_i^T is closer to $\hat{H}_k^I$ but not to $\hat{H}_j^I$. The objective can be defined as $\mathcal{F}_{T \to \hat{I}} = \sum_{\langle i,j,k \rangle} \max\{0, \varepsilon + ||H_i^T - \hat{H}_j^I|| - ||H_i^T - \hat{H}_k^I||\}$. The hashing module tries to minimize the objective, while the attention module tries to maximize it. The same process for the *image* → *text*. Thus, the loss can be expressed as

$$\begin{aligned} \mathcal{F}_{T \to \hat{I}} + \mathcal{F}_{I \to \hat{T}} = & \sum_{\langle i,j,k \rangle} \max\{0, \varepsilon + ||H_i^T - \hat{H}_j^I|| - ||H_i^T - \hat{H}_k^I||\} \\ & + \sum_{\langle i,j,k \rangle} \max\{0, \varepsilon + ||H_i^I - \hat{H}_j^T|| - ||H_i^I - \hat{H}_k^T||\}, \end{aligned} \tag{6}$$

where $\hat{H}^T$ and $\hat{H}^I$ are the unattended features defined in Subsection 3.2. The first term corresponds to taking the text modality as the query to retrieve the unattended features of the image modality. The second term corresponds to the image modality being taken as the query to retrieve the unattended features of the text modality. G^I, G^T attempt to maximize the loss and D^I, D^T to minimize the objective:

$$\min_{D^I, D^T} \max_{G^I, G^T} \mathcal{F}_{T \to \hat{I}} + \mathcal{F}_{I \to \hat{T}}. \tag{7}$$

Full Objective Our full objective is

$$\begin{aligned} \mathcal{F}(E^I, E^T, G^I, G^T, D^I, D^T) = & \ \mathcal{F}_{T \to \hat{I}} + \mathcal{F}_{I \to \hat{T}} \\ & + \mathcal{F}_{T \to I} + \mathcal{F}_{I \to T} + \mathcal{F}_{I \to I} + \mathcal{F}_{T \to T}. \end{aligned}$$

We train our model alternatively. The parameters in G^I and G^T are fixed, while the other parameters are trained:

$$\min_{E^I, E^T, D^I, D^T} \mathcal{F}(E^I, E^T, G^I, G^T, D^I, D^T). \tag{8}$$

Then E^I, E^T, D^I, and D^T are fixed and the attention models are updated:

$$\max_{G^I, G^T} \mathcal{F}_{T \to \hat{I}} + \mathcal{F}_{I \to \hat{T}}. \tag{9}$$

4 Experiments

In this section, we evaluate the performance of our proposed methods on three datasets and compare it to the performance of several stage-of-the-art algorithms.

4.1 Experimental Settings

Datasets. We choose three benchmark datasets: IAPR TC-12 [9], MIR-Flickr 25K [13] and NUS-WIDE [6].

- **IAPR TC-12** [9]: This dataset consists of 20,000 images taken from locations around the world. Each image is associated with a text caption, e.g., a sentence. The image-text pairs are annotated using 255 labels. For the text modality, each sentence is represented as a 2,912-dimensional bag-of-words vector [5].
- **MIR-Flickr 25K** [13]: This dataset contains 25,000 multi-label images downloaded from the Flickr [6] photo-sharing website. Each image is associated with several textural tags. For a fair comparison, we follow the settings in [15] to use the subset of the image-text pairs with at least 20 textual tags. For the text modality, the textural tags are represented as a 1,386-dimensional bag-of-words vector.
- **NUS-WIDE** [6]: This dataset consists of 269,648 images collected from Flickr. Each image is associated with one or multiple textural tags in 81 semantic concepts. We evaluate the performance on 195,834 image-text pairs belonging to the 21 most frequent labels, as suggested by [15]. The text is represented as a 1,000-dimensional bag-of-words vector.

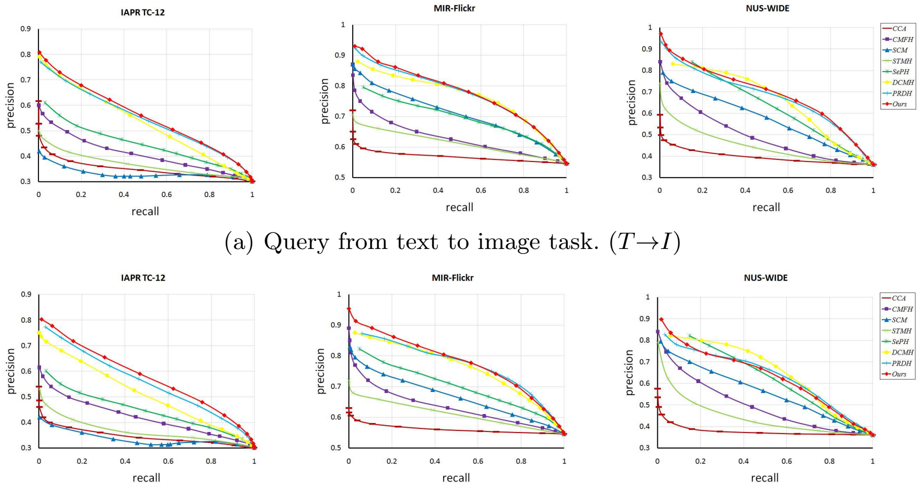
We follow the settings of DCMH [15] to construct the query sets, training sets, and retrieval databases. The randomly sampled 2,000 image-text pairs are constructed as the query set for IAPR TC-12 and MIR-Flickr 25K. For the NUS-WIDE dataset, we randomly sample 2,100 image-text pairs as the query set. For all datasets, the remaining image-text pairs are used as the databases for retrieval. For all supervised methods, we also sample 10,000 pairs from the retrieval set as the training set for IAPR TC-12 and MIR-Flickr 25K, as well as 10,500 pairs from the retrieval set as the training set for NUS-WIDE.

Note that the representations of text are not the focus of this paper. Since the most related works, e.g., DCMH [15], use bag-of-words, we also use bag-of-words for a fair comparison.

Implementation details. We implement our codes based on the open source *caffe*[14] framework. In training, the networks are updated alternatively through the stochastic gradient solver, i.e., ADAM ($\alpha = 0.0002$, $\beta_1 = 0.5$). We alternate between four steps of optimizing E, D and one step of optimizing G. For the image modality, the weights of VGGNet are initialized with the pre-trained model that learns from the ImageNet dataset. For text modality, all parameters are randomly initialized with a Gaussian with mean zero and standard deviation 0.01. The batch size is 64, and the total epoch is 100. The base learning rate is 0.005, and it is changed to one-tenth of the current value after every 20 epochs. In testing, we use only the attended features of the data to construct the binary codes.

[5] We follow the settings of DCMH [15] for a fair comparison

[6] www.flickr.com



(a) Query from text to image task. ($T \rightarrow I$)

(b) Query from image to text task. ($I \rightarrow T$)

Fig. 5. Precision-recall curves on three datasets. The length of hash code is 16.

Evaluation Measures. To evaluate the performance of hashing models, we use two metrics: the mean average precision (MAP) [20] and precision-recall curves. The MAP is a standard evaluation metric for information retrieval.

4.2 Comparison with State-of-the-art Methods

The first set of experiments is to evaluate the performance of the proposed method and compare it with the performance of several state-of-the-art algorithms [7]: CCA [12], CMFH [40], SCM [8], SMTH [30], SePH [19], DCMH [15], and PRDH [37]. The results of CCA, CMFH, SCM, STMH, SePH and DCMH are directly cited from [15] published in CVPR17 [8]. Since the experimental settings of PRDH in [37] are different from those of the proposed method, we carefully implement PRDH using the same CNN network and the same settings for a fair comparison.

The comparison results of the search accuracies on all three datasets are shown in Table 1. We can see that our method outperforms other baselines and achieves excellent performance. For example, on IAPR TC-12, the MAP of our method is 0.5439, compared to the value of 0.5135 for the second best algorithm (PRDH), on 64 bits when taking the image as the query to retrieve text. The precision-recall curves are also shown in Fig. 5. It can be seen that our method shows comparable performance to the existing baselines.

[7] Note that IRGAN is designed for uni-modal retrieval. ACMR is a cross-modal retrieval method that falls in the category of real-valued approaches. In this paper, we only focus on cross-modal hashing.

[8] Table 4 in http://openaccess.thecvf.com/content_cvpr_2017/papers/Jiang_Deep_Cross-Modal _Hashing_CVPR_2017_paper.pdf

Table 1. MAP of Hamming ranking w.r.t. different numbers of bits on three datasets.

Task		IAPR TC-12			MIR-Flickr 25k			NUS-WIDE		
		16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
Text ↓ Image	CCA	0.3493	0.3438	0.3378	0.5742	0.5713	0.5691	0.3731	0.3661	0.3613
	CMFH	0.4168	0.4212	0.4277	0.6365	0.6399	0.6429	0.5031	0.5187	0.5225
	SCM	0.3453	0.3410	0.3470	0.6939	0.7012	0.7060	0.5344	0.5412	0.5484
	STMH	0.3687	0.3897	0.4044	0.6074	0.6153	0.6217	0.4471	0.4677	0.4780
	SePH	0.4423	0.4562	0.4648	0.7216	0.7261	0.7319	0.5983	0.6025	0.6109
	DCMH	0.5185	0.5378	0.5468	0.7827	0.7900	0.7932	0.6389	0.6511	0.6571
	PRDH	0.5244	0.5434	0.5548	0.7890	0.7955	0.7964	0.6527	0.6916	0.6720
	Ours	**0.5358**	**0.5565**	**0.5648**	**0.7922**	**0.8062**	**0.8074**	**0.6789**	**0.6975**	**0.7039**
Image ↓ Text	CCA	0.3422	0.3361	0.3300	0.5719	0.5693	0.5672	0.3742	0.3667	0.3617
	CMFH	0.4189	0.4234	0.4251	0.6377	0.6418	0.6451	0.4900	0.5053	0.5097
	SCM	0.3692	0.3666	0.3802	0.6851	0.6921	0.7003	0.5409	0.5485	0.5553
	STMH	0.3775	0.4002	0.4130	0.6132	0.6219	0.6274	0.4710	0.4864	0.4942
	SePH	0.4442	0.4563	0.4639	0.7123	0.7194	0.7232	0.6037	0.6136	0.6211
	DCMH	0.4526	0.4732	0.4844	0.7410	0.7465	0.7485	0.5903	0.6031	0.6093
	PRDH	0.5003	0.4935	0.5135	0.7499	0.7546	0.7612	0.6107	**0.6302**	0.6276
	Ours	**0.5293**	**0.5283**	**0.5439**	**0.7563**	**0.7719**	**0.7720**	**0.6403**	0.6294	**0.6520**

Since the code of DVSH is not publicly available and it is difficult to re-implement the complex algorithm, we utilize the same experimental settings used in DVSH for our method. The results of DVSH are directly cited from [3] for a fair comparison. The top-500 MAP results on IAPR TC-12 are shown in Table 2. Moreover, we make a comparison with DCMH under the same settings. Please note that DVSH adopts the LSTM recurrent neural network for text representation, while DCMH and our method only use bag-of-words. From the table, we can see that our methods can achieve better performance than the baselines in most cases, even we use the weak representations of text.

Table 2. The comparison results w.r.t. the top-500 MAP on the IAPR TC-12 dataset.

Task	Methods	16 bits	32 bits	64 bits
Text→Image	DVSH	0.6037	0.6395	0.6806
	DCMH	0.6594	0.6744	0.6905
	Ours	**0.7018**	**0.6893**	**0.6941**
Image→Text	DVSH	0.5696	0.6321	**0.6964**
	DCMH	0.5780	0.6061	0.6310
	Ours	**0.6464**	**0.6373**	0.6668

We also explore the effects of small network architecture in the feature learning module for the image modality since VGGNet is a large deep network. In this experiment, we select CNN-F [5] as the basic network for the image modality. The comparison results are shown in Table 3. We can see that VGGNet

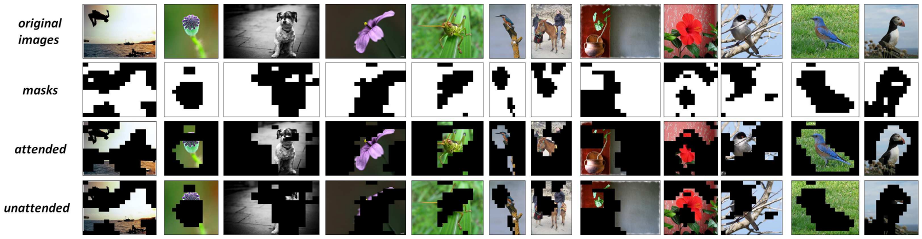
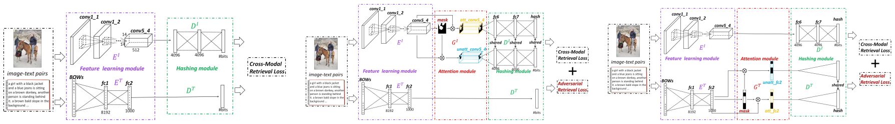


Fig. 6. Some image and mask samples. The first line represents the original images, the masks are shown in the second line, and the combinations are shown in the last two lines.



(a) No attention (b) Visual attention (c) Textural attention

Fig. 7. Different attention mechanisms.

performs better than CNN-F while our method using CNN-F also achieves good performance compared to other state-of-the-art baselines.

Table 3. MAP on IAPR TC-12 dataset with different networks.

Task	Networks	16 bits	32 bits	64 bits
Text→Image	VGG	0.5358	0.5565	0.5648
	CNN-F	0.5267	0.5459	0.5538
Image→Text	VGG	0.5293	0.5283	0.5439
	CNN-F	0.5211	0.5168	0.5208

The main reason for the good performance of our method is that we can obtain attended regions for the multi-modal data. Fig. 6 shows some examples of the image modality. Note that it is difficult to visualize the text modality (the networks for the text modality are the fully connected layers instead of the CNN. The order of words in the sentence are changed after going through the fully connected layers), thus, we do not show the masks learned in the text network.

4.3 Comparison with Different Attention Mechanisms

In this section, we present an ablation study to clarify the impact of each part of the attention modules on the final performance.

To provide an intuitive comparison of our method, we compare it with the following baselines. In the first baseline, we do not use any attention mechanism as shown on the left side of Fig. 7. It is also the traditional deep cross-modal hashing. In the second baseline, we only apply the visual attention mechanism as seen in the middle of Fig. 7. Similarly, the last baseline is to explore the textural attention mechanism as shown on the right side of Fig. 7. Note that all baselines, as well as our method, use the same network. The only differences are the use of the different attention mechanisms. These comparisons can show whether the proposed attention mechanism can contribute to the accuracy.

Table 4 shows the comparison results with respect to the MAP. The results show that our proposed attention mechanism can achieve better performance than the baselines that are lacking attention mechanisms. The main reason for this is that our method can focus on the most discriminative regions of the data.

Table 4. The comparison results for different attention mechanisms.

Task	Attn.	IAPR TC-12			MIR-Flickr 25k			NUS-WIDE		
		16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
Text ↓ Image	No	0.5039	0.5250	0.5258	0.7758	0.7801	0.7742	0.6476	0.6824	0.6733
	Visual	0.5294	0.5474	0.5576	0.7894	0.7925	0.7906	0.6723	0.6839	0.6984
	Textual	0.5334	0.5382	0.5469	0.7885	0.7867	0.7831	0.6648	0.6851	0.6867
	Both	**0.5358**	**0.5565**	**0.5648**	**0.7922**	**0.8062**	**0.8074**	**0.6789**	**0.6975**	**0.7039**
Image ↓ Text	No	0.4903	0.5001	0.5175	0.7347	0.7482	0.7495	0.6150	0.6178	0.6311
	Visual	0.5267	0.5173	0.5285	0.7466	0.7601	0.7584	0.6314	0.6260	0.6425
	Textual	0.5279	0.5232	0.5304	0.7520	0.7673	0.7717	0.6384	0.6227	0.6459
	Both	**0.5293**	**0.5283**	**0.5439**	**0.7563**	**0.7719**	**0.7720**	**0.6403**	**0.6294**	**0.6520**

5 Conclusion

In this paper, we proposed a novel approach called deep adversarial hashing for cross-modal hashing. The proposed method contains three major components: a feature learning module, an attention module, and a hashing module. The feature learning module learns powerful representations for the multi-modal data. The attention module and the hashing module are trained in an adversarial way, in which the hashing module tries to minimize the similarity-preserving loss functions, while the attention module aims to find the unattended regions of data that maximize the retrieval loss. We performed our method on three datasets, and the experimental results demonstrate the appealing performance of our method.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grants (61602530, U1611264, U1711262, 61472453, U1401256 and U1501252). This work is also supported by the Research Foundation of Science and Technology Plan Project in Guangdong Province (2017B030308007).

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. arXiv preprint arXiv:1701.07875 (2017)
2. Ba, J., Mnih, V., Kavukcuoglu, K.: Multiple object recongnition with visual attention. In: ICLR (2015)
3. Cao, Y., Long, M., Wang, J., Yang, Q., Philip, S.Y.: Deep visual-semantic hashing for cross-modal retrieval. In: KDD. pp. 1445–1454 (2016)
4. Cao, Y., Long, M., Wang, J., Zhu, H.: Correlation autoencoder hashing for supervised cross-modal search. In: ICMR. pp. 197–204 (2016)
5. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. Computer Science (2014)
6. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: Nus-wide: a real-world web image database from national university of singapore. In: ICIVR. p. 48 (2009)
7. Courbariaux, M., Bengio, Y.: Binarynet: Training deep neural networks with weights and activations constrained to +1 or -1. CoRR **abs/1602.02830** (2016)
8. Ding, G., Guo, Y., Zhou, J.: Collective matrix factorization hashing for multimodal data. In: CVPR. pp. 2075–2082 (2014)
9. Escalante, H.J., Hernndez, C.A., Gonzalez, J.A., Lpez-Lpez, A., Montes, M., Morales, E.F., Sucar, L.E., Villaseor, L., Grubinger, M.: The segmented and annotated iapr tc-12 benchmark. CVIU **114**(4), 419–428 (2010)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS. pp. 2672–2680 (2014)
11. He, R., Zheng, W.S., Hu, B.G.: Maximum correntropy criterion for robust face recognition. TPAMI **33**(8), 1561–1576 (2011)
12. Hotelling, H.: Relations Between Two Sets of Variates. Springer New York (1992)
13. Huiskes, M.J., Lew, M.S.: The mir flickr retrieval evaluation. In: ICMIR. pp. 39–43 (2008)
14. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
15. Jiang, Q.Y., Li, W.: Deep cross-modal hashing. In: CVPR (2016)
16. Lai, H., Pan, Y., Liu, Y., Yan, S.: Simultaneous feature learning and hash coding with deep neural networks. In: CVPR. pp. 3270–3278 (2015)
17. Lai, H., Yan, P., Shu, X., Wei, Y., Yan, S.: Instance-aware hashing for multi-label image retrieval. TIP **25**(6), 2469–2479 (2016)
18. Li, C., Deng, C., Li, N., Liu, W., Gao, X., Tao, D.: Self-supervised adversarial hashing networks for cross-modal retrieval. In: CVPR. pp. 4242–4251 (2018)
19. Lin, Z., Ding, G., Hu, M., Wang, J.: Semantics-preserving hashing for cross-view retrieval. In: CVPR. pp. 3864–3872 (2015)
20. Liu, W., Kumar, S., Kumar, S., Chang, S.F.: Discrete graph hashing. In: NIPS. pp. 3419–3427 (2014)
21. Masci, J., Bronstein, M.M., Bronstein, A.M., Schmidhuber, J.: Multimodal similarity-preserving hashing. TPAMI **36**(4), 824–830 (2014)
22. Mathieu, M.F., Zhao, J.J., Zhao, J., Ramesh, A., Sprechmann, P., LeCun, Y.: Disentangling factors of variation in deep representation using adversarial training. In: NIPS. pp. 5040–5048 (2016)
23. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)

24. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
25. Sharma, S., Kiros, R., Salakhutdinov, R.: Action recognition using visual attention. arXiv preprint arXiv:1511.04119 (2015)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
27. Sun, L., Ji, S., Ye, J.: A least squares formulation for canonical correlation analysis. In: ICML. pp. 1024–1031 (2008)
28. Wang, B., Yang, Y., Xu, X., Hanjalic, A., Shen, H.T.: Adversarial cross-modal retrieval. In: ACMMM. pp. 154–162 (2017)
29. Wang, D., Cui, P., Ou, M., Zhu, W.: Learning compact hash codes for multimodal representations using orthogonal deep structure. TMM **17**(9), 1404–1416 (2015)
30. Wang, D., Gao, X., Wang, X., He, L.: Semantic topic multimodal hashing for cross-media retrieval. In: ICAI. pp. 3890–3896 (2015)
31. Wang, J., Zhang, T., Sebe, N., Shen, H.T., et al.: A survey on learning to hash. TPAMI (2017)
32. Wang, J., Yu, L., Zhang, W., Gong, Y., Xu, Y., Wang, B., Zhang, P., Zhang, D.: Irgan: A minimax game for unifying generative and discriminative information retrieval models. arXiv preprint arXiv:1705.10513 (2017)
33. Wang, K., Yin, Q., Wang, W., Wu, S., Wang, L.: A comprehensive survey on cross-modal retrieval. arXiv preprint arXiv:1607.06215 (2016)
34. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML. pp. 2048–2057 (2015)
35. Xu, X., Shen, F., Yang, Y., Shen, H.T., Li, X.: Learning discriminative binary codes for large-scale cross-modal retrieval. TIP **26**(5), 2494–2507 (2017)
36. Yang, E., Deng, C., Liu, W., Liu, X., Tao, D., Gao, X.: Pairwise relationship guided deep hashing for cross-modal retrieval. In: AAAI. pp. 1618–1625 (2017)
37. Yang, E., Deng, C., Liu, W., Liu, X., Tao, D., Gao, X.: Pairwise relationship guided deep hashing for cross-modal retrieval. AAAI (2017)
38. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: CVPR. pp. 21–29 (2016)
39. Yu, Z., Wu, F., Yang, Y., Tian, Q., Luo, J., Zhuang, Y.: Discriminative coupled dictionary hashing for fast cross-media retrieval. In: SIGIR. pp. 395–404 (2014)
40. Zhang, D., Li, W.J.: Large-scale supervised multimodal hashing with semantic correlation maximization. In: AAAI. vol. 1, p. 7 (2014)
41. Zhen, Y., Yeung, D.Y.: Co-regularized hashing for multimodal data. In: NIPS. pp. 1376–1384 (2012)
42. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv preprint arXiv:1703.10593 (2017)