Article

Self-Attention and Adversary Guided Hashing Network for Cross-Modal Retrieval

Shubai Chen [1,*], Li Wang [2] and Song Wu [1,*]

[1] College of Computer and Information Science, Southwest University, Chongqing 400715, China; chansuba@email.swu.edu.cn

[2] College of Electronic and Information Engineering, Southwest University, Chongqing 400715, China; t123456t@email.swu.edu.cn

* Correspondence: chansuba@email.swu.edu.cn; songwuswu@swu.edu.cn

Abstract: Recently deep cross-modal hashing networks have received increasing interests due to its superior query efficiency and low storage cost. However, most of existing methods concentrate less on hash representations learning part, which means the semantic information of data cannot be fully used. Furthermore, they may neglect the high-ranking relevance and consistency of hash codes. To solve these problems, we propose a Self-Attention and Adversary Guided Hashing Network (SAAGHN). Specifically, it employs self-attention mechanism in hash representations learning part to extract rich semantic relevance information. Meanwhile, in order to keep invariability of hash codes, adversarial learning is adopted in the hash codes learning part. In addition, to generate higher-ranking hash codes and avoid local minima early, a new batch semi-hard cosine triplet loss and a cosine quantization loss are proposed. Extensive experiments on two benchmark datasets have shown that SAAGHN outperforms other baselines and achieves the state-of-the-art performance.

Keywords: adversarial learning; deep cross-modal hashing; self-attention mechanism

1. Introduction

With the explosive growth of multimedia data in social networks and search engines, including texts, audio, videos and images, efficiently retrieval information between different modalities types has become a hot spot. For example, given an image, one may want to obtain all semantic related captions from the database. However, due to the inconsistent distribution and representation of different modalities, it has posted new challenges to unify different modalities and bridge their semantic gaps efficiently and effectively.

Existing cross-modal retrieval methods can be roughly divided into two categories. The first one is latent embedding learning methods and the second is hashing based methods. Many real-valued latent embedding methods have been proposed, including topic models [1,2], subspace learning [3–5] and deep models [6–9]. While the high computing complexity and low search efficiency have become another problem. Another universal solution to cross-modal retrieval is hashing based methods, which use the Manifold Learning to transfer binary codes from high dimensional features. Many cross-modal hashing methods have been proposed to bridge the "heterogeneity gap" by finding a latent space's representation (i.e. Hamming), in which the similarities between features can be directly estimated with same distance metric.

Up to now, many cross-modal hashing methods has been proposed including unsupervised and supervised manners. Unsupervised cross-modal hashing methods utilize the original multi-modal features to explore the latent data structure and distribution. To name a few, Collective Matrix Factorization Hashing (CMFH) [10], Inter-Media Hashing (IMH) [11], Unsupervised Generative Adversarial Cross-modal Hashing (UGACH) [12], and Latent Semantic Sparse Hashing (LSSH) [13]. Supervised methods, on the other hand, try to leverage semantic labels as supervised information to learn hash coding functions. In this way, supervised cross-modal hashing methods have better performance than unsupervised cross-modal hashing methods. Representative supervised cross-modal hashing methods include Semantic Preserving Hashing (SePH) [14], Semantic Correlation Maximization (SCM) [15], Semi-Relaxation Supervised Hashing (SRSH) [16] and Dictionary Learning Cross-modal Hashing (DLCMH) [17]. However, almost all these existing cross-modal hashing methods are based on shadow architectures and hand-crafted features. One shortcoming is that the feature extraction and hash function learning procedure are independent, which means the two steps might not be optimally compatible with each other.

As the tremendous progress of deep learning recently, Deep Convolution Neural Network (DCNN) has achieved great success in many computer vision applications [18–22]. The latest cross-modal retrieval methods based on deep learning construct an end-to-end architecture which can simultaneously learn both binary representations and hash function. Moreover, deep model based cross-modal hashing methods have shown superior performance over the traditional hashing methods, such as Deep Cross-Modal Hashing (DCMH) [23], Correlation Hashing Network (CHN) [24], Pairwise Relationship Guided Deep Hashing (PRDH) [25], Collective Deep Quantization (CDQ) [26], Self-Supervised Adversarial Hashing (SSAH) [27], Adversary Guided Asymmetric Hashing (AGAH) [28], and Triplet-based Deep Hashing (TDH) [29]. However, there are still some issues that need to be further considered in current DCNN-based cross-modal hashing methods. Firstly, a two-stream structure is employed to learn projection functions separately so that multimodality features can be mapped into Hamming space. However, the heterogeneous gap is neglected because of the inconsistent representation and distribution. Furthermore, label constraint imposed on the end of this structure fails to ensure either higher-ranking correlation among item pairs which share same labels or multi-label semantics in hash codes. Moreover, the hash representations are usually extracted from high-level layers of deep neural networks. However, representations from high-level layers mostly encode semantic information, while information from intermediate layers which is more modality discriminative have been disposed.

To address issues mentioned above, in this paper, a novel and efficient Self-Attention and Adversary Guided Hashing Network (SAAGHN) is proposed. As illustrated in Fig. 1, the proposed method jointly learns hash representations and hash codes in an end-to-end way. In feature learning model, we use a CNN model to extract multi-level information. In order to take advantages of rich semantic and spatial information, a self-attention based module is proposed. In this module, features from each layers are fused together into a richer features pool, and the final hash representations are produced from features pool in a self-attention way. In order to eliminate the inconsistent distribution of hash codes across modalities, an adversarial guided method is introduced in hashing model. Besides, to better use the multi-label semantic knowledge and avoid earlier local minima, a batch semi-hard triplet-margin constraint is adopted to preserve higher-ranking relationship. In addition, a hamming intra constraint is proposed to integrate label and feature information. Finally, a modality-invariance constraint is adopted to minimize the gap between image codes and text codes. The main contributions of this work are summarized as follows:

An effective and efficient self-attention based module is proposed to rich the extracted information from each modality. This module transfers the learned features from different layers to a features pool, and the final hash representations are produced by a self-attention mechanism. In this way, different levels of hidden representations in the networks are fused to get discriminative and balanced hash representations. Besides, integrating with self-attention helps to find global, long-range dependencies within internal representations of different modalities efficiently and effectively.

Adversarial learning is introduced in hashing learning part, which can ensure invariance and semantic similarity of modality. Besides, a batch semi-hard cosine triplet loss is introduced to preserve multi-label high-level semantic information as well as keep the higher-ranking similarities.

The SAAGHN is evaluated on two benchmark datasets. The experimental results show its outperformance over current state-of-the-art methods.

2. Related Work

In this section, we briefly review some typical related work on hashing based cross-modal retrieval.

2.1. Deep Cross-Modal Hashing

Since the previous cross-modal hashing method are built on shallow architecture, they cannot explore the complex nonlinear correlation across different modalities, thus, deep cross-modal hashing retrieval has attracted increasing attention. One of the most representative work is deep cross modal hashing (DCMH) [23] which simultaneously learns features and hash codes in an end-to-end deep learning framework. Correlation hashing network (CHN) [24] adopts a hybrid deep architecture to jointly learn good data representation tailored to hash coding and formally control the quantization error. Pairwise relation guided deep hashing (PRDH) [25] jointly performs inter-modality and intra-modality pairwise constraints to enhance the semantic preservation. Although these method has achieved appealing performance, they all lose the rich information from the intermediate layers and higher-ranking correlation relationship among inter-modality pairs.

2.2. Generative Adversarial Networks (GAN)

Generative adversarial networks (GAN) [30] have achieved remarkable progress in image generation, representation learning, super resolution and other fields since it was proposed. Many deep cross-modal hashing methods based on GAN were proposed. Adversarial cross-modal retrieval (ACMR) [31] adds discriminators to distinguish different modalities features, which is the first method to apply GAN in cross-modal hashing. Self-supervised adversarial hashing (SSAH) [27] leverages two adversarial networks to maximize the semantic relevance and consistency across modalities. In comparison of SSAH, label network is replaced by label information to directly bridge modality gap, and adversarial module is added at the hash codes learning step to ensure consistency of hash codes across modalities.

2.3. Attention Mechanism

Recently attention based deep networks has drawn increasing interest because of the appealing performance in various tasks, like image classification [32], visual question answering [33] and person re-identification [34] etc. Attention-aware deep adversarial hashing (DAH) [35] divides the hash representations into the attended and the unattended in an adaptive method. Adversarial guided asymmetric hashing (AGAH) [28] adopts a multi-label attention map to preserve semantic information. In spite of this progress, self-attention has yet been explored in the context of cross-modal hashing. Thus, a self-attention constraint module is proposed to enhance preservation of global, long-range dependencies within internal representations of different modalities.

3. Proposed Method

3.1. Problem Definition

We use uppercase letters represent matrices, such as X, and lowercase letters represent vectors, such as y. And G^T denotes the transpose of G, the Frobenius norm is denoted as $\|\cdot\|_F$. The sign$(\cdot)$ is an element-wise sign function defined as follows:

$$\text{sign}(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases} \tag{1}$$

Without losing generality, assuming there are N instances of image-text features pairs. Let $X = \{x_i\}_{i=1}^N$ and $Y = \{y_i\}_{i=1}^N$ denote the feature vector of image modality and text modality, where $x_i \in R^{d_x}$ denotes

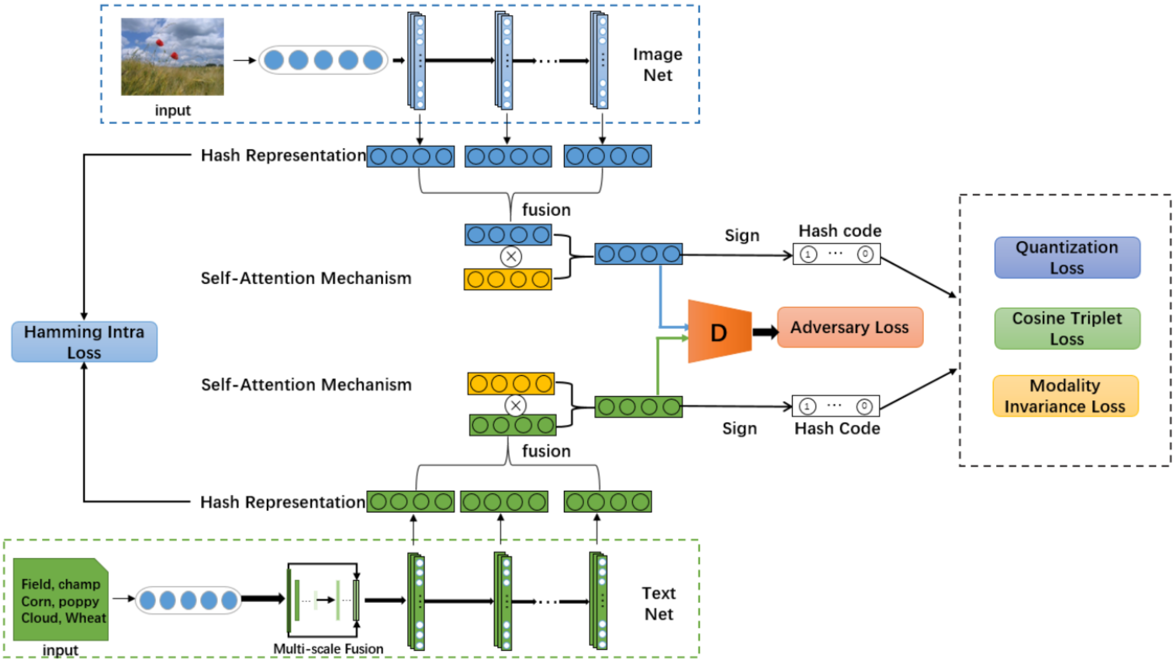

Figure 1. The framework of our proposed SAAGHN.

the d_x-dimensional feature vector of the image modality, and $y_i \in R^{d_y}$ denotes the d_y-dimensional feature vector of the text modality. The similarity matrix $S = \{S^{xx}, S^{yy}\}$ is used to measure the label relevance of two intra-instances, where $S_{ij}^{xx} = 1$ or $S_{ij}^{yy} = 1$ if two intra-modal instances share at least one label, and $S_{ij}^{xx} = 0$ or $S_{ij}^{yy} = 0$ otherwise. Given the training data information X, Y and S, the goal of SAAGH is to learn two modality-specific hashing functions $h^{(x)}(\mathbf{x})$ and $h^{(y)}(\mathbf{y})$ which project feature vectors of image modality and text modality into binary code $B^X \in \{-1, +1\}^c$ and $B^Y \in \{-1, +1\}^c$, where c is the length of the hash code. The framework of cross-modal hashing can be divided into hash representation learning part and hash-function learning part. $F = \{f_{x_i} \mid i = 1, 2, \cdots, N\} \in R^{N \times c}, G = \{g_{y_i} \mid i = 1, 2, \cdots, N\} \in R^{N \times c}$ are denoted as learned hash representation from image modality and text modality. Finally, the hash code can be calculated by applying a sign function on hash representation F and G: $B^X = sign(F), B^Y = sign(G)$.

3.2. Network Architecture of SAAGH

The feature extractor is composed of two deep neural networks for image-modality feature learning and text-modality feature learning. Inspired by SCAHN, a self-attention constraint module is proposed to guide the intermediate layers as the additional supervision. The framework of self-attention constraint module is shown in Fig. 2.The self-attention constraint module fuses the features extracted from intermediate into a features pool. The features pool integrates with self-attention module to get the weighted features.

The whole SAAGH framework is shown in Fig 1, which learns hash representations and hash codes in an end-to-end architecture. The adopted neural network for both ImageNet and TextNet is ResNet. Image-text pairs are used as input data for the whole architecture. Inspired by SSAH [27], the text instances are represented by bag-of-words (BoW) to fuse multiple scale features. Multiple scale features are generated by multiple average pooling layers and they are fused by a up-sampling operation. In this way, the multiple semantic relevance for text modality can be kept, and the same module can be applied on both text and image modalities.

Inspired by SCAHN [36], we create the side branch from intermediate layers to make a robust features. Without losing generality, we assume there are K blocks in ResNet, which corresponds to K side branches. As shown in Fig. 2 ,the global averaging pooling and a 1x1 convolution layer are added. All blocks are fused into a features pool. The features pool will be transferred to final hash representation by going through a self-attention module. Finally, the hash representations for image modality and text modality can be expressed as f_{x_i} and g_{y_j}.

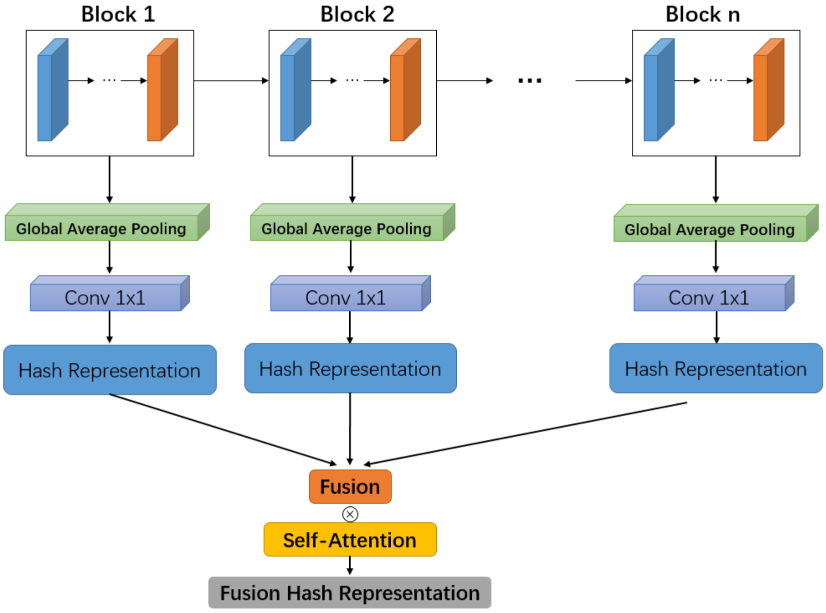

Figure 2. The architecture of self-attention module

3.3. Hash Function Learning

In order to generate high-quality and high-ranking relevance hash codes, several loss functions are proposed by exploring the semantic relevance in intra-modality and inter-modality.

Adversarial Loss. To generate modality-invariant hash codes across modalities, the adversarial learning is introduced in hashing function learning part. We design two discriminators of GAN (D_x and D_y) for image modality and text modality, which are composed of two-layer feed-forward neural networks with parameters θ_{D_x} and θ_{D_y}. For text modality discriminator D_y, the image network is regraded as generator for text hash codes. The hash codes generated from text modality are real text hash codes, while hash codes from image network are considered as fake codes. This discriminator is designed to distinguish whether the text hash code is real or fake. The image modality discriminator acts the same way.

In the process of adversarial learning, the text network attempts to generate image hash codes to cheat the image discriminator, and the image network tries to generate text hash codes to cheat the text discriminator. The adversarial loss in hash function learning part is defined as $\mathcal{L}_{adv}$, which is a cross-entropy loss. The formulation is defined as follows:

$$\begin{aligned}\mathcal{L}_{adv} =&\mathcal{L}^{x}_{adv} + \mathcal{L}^{y}_{adv}\\ =&-\frac{1}{n}\left[\sum_{i=1}^{n}\left(\log\left(D_x\left(\mathrm{b}_i^x;\theta_{D_x}\right)\right)+\log\left(1-D_x\left(\mathrm{b}_i^y;\theta_{D_x}\right)\right)\right)\right.\\ &\left.+\sum_{i=1}^{n}\left(\log\left(D_y\left(\mathrm{b}_i^y;\theta_{D_y}\right)\right)+\log\left(1-D_y\left(\mathrm{b}_i^x;\theta_{D_y}\right)\right)\right)\right]\end{aligned} \tag{2}$$

Specially, $\mathcal{L}^{*}_{adv}$ denotes the image modality or text modality adversarial loss and $\mathrm{b}_i^*, * \in \{x, y\}$ means the hash code of i-th instance of image or text.

Batch Semi-hard Cosine Triplet Loss. In order to preserve higher-ranking relationships between modalities, pairwise loss is usually adopted in many other methods like [25]. While pairwise cannot fully explore the discriminability between instances. So we proposed a batch semi-hard cosine triplet loss which is inspired by the cosine triplet loss used in AGAH [28].

Take the image modality for example, the form of triplet in our method is $\left(\mathbf{b^X}_i, \mathbf{b^Y}_j^+, \mathbf{b^Y}_k^-\right)$ where text hash code $\mathbf{b^Y}_j^+$ is similar to image hash code $\mathbf{b^X}_i$ while the $\mathbf{b^Y}_k^-$ is another way around. Cosine triplet loss means that cosine distance is considered as surrogate of Hamming distance because of the gap between Hamming distance and inner product for continuous representations which is mentioned in CHN [24]. The cosine triplet loss for image modality can be formulated as follows:

$$\mathcal{L}_{I\to T} = \sum_{i,j,k} \max\left(\cos\left(\mathbf{b^X}_i, \mathbf{b^Y}_k^-\right) - \cos\left(\mathbf{b^X}_i, \mathbf{b^Y}_j^+\right) + m, 0\right) \tag{3}$$

where m is the margin parameter.

In the same way, the cosine triplet loss for the text modality can be formulated as:

$$\mathcal{L}_{tri}^{y} = \sum_{i,j,k} \max\left(\cos\left(\mathbf{b^Y}_i, \mathbf{b^X}_k^-\right) - \cos\left(\mathbf{b^Y}_i, \mathbf{b^X}_j^+\right) + m, 0\right) \tag{4}$$

Although cosine triplet loss in [28] has achieved appealing results, it still suffers from some disadvantages. On the one hand, as the dataset gets large, the number of triplets grows rapidly which makes the training inefficiently. On the other hand, in the process of hard triplet loss learning, selecting the hardest negatives will in practice lead to bad local minima early, specifically it can result in a collapsed model.

To solve the problem mentioned above, inspired by FaceNet [37] and Batch Hard triplet loss, the semi-hard triplet selection and batch hard triplet loss form are used. The principle of semi-hard selection is defined as follows:

$$d\left(x_q, x_i^+\right) < d\left(x_q, x_i^-\right) \tag{5}$$

where $d(.)$means the distance between two instances. In this way, the negatives can be far away from the anchor than the positive example, but still in a hard way to explore the relevance between instances.

Moreover, the loss function can be formulated in a batch-hard manner based on semi-hard selection, which is formulated as follows:

$$\mathcal{L}_{I\to T} = \sum_{i,j,k} \max\left(\mathbf{min}cos\left(\mathbf{b^X}_i, \mathbf{b^Y}_k^-\right) - \mathbf{max}cos\left(\mathbf{b^X}_i, \mathbf{b^Y}_j^+\right) + m, 0\right) \tag{6}$$

$$\mathcal{L}_{T\to I} = \sum_{i,j,k} \max\left(\mathbf{min}cos\left(\mathbf{b^Y}_i, \mathbf{b^X}_k^-\right) - \mathbf{max}cos\left(\mathbf{b^Y}_i, \mathbf{b^X}_j^+\right) + m, 0\right) \tag{7}$$

Finally, combining (6) and (7) the total triplet loss can be formulated as:

$$\mathcal{L}_{tri} = \mathcal{L}_{T\to I} + \mathcal{L}_{I\to T} \tag{8}$$

Hamming Distance Intra-Loss. Although the adversarial loss is proposed to generate modality-invariant hash codes and batch semi-hard cosine triplet loss is proposed to preserve higher-ranking relationship, they almost concentrate on inter-modality, neglecting the similarity between intra-modality instances. The Hamming distance intra-loss is introduced to solve this problem. The similarity of learned hash representation f_{x_i} and g_{y_j} can be represented by $S = \{S^{xx}, S^{yy}\} \in \{0, 1\}$, which is evaluated by their inner product. $S_{ij}^* = 0$ means that the two intra-modality instances are dissimilar, consequently the inner product should be small, $S_{ij}^* = 1$ otherwise. The likelihood function can be defined as follows:

$$p\left(S_{ij}^* \mid M_{*_i}, M_{*_j}\right) = \begin{cases} \sigma\left(\theta_{ij}\right), & S_{ij}^* = 1 \\ 1 - \sigma\left(\theta_{ij}\right), & S_{ij}^* = 0 \end{cases} \tag{9}$$

where $\theta_{ij} = \alpha M_{*_i} M_{*_j}^T$, $M_* = F_*$ or G_*, $F_*$and G_* denote the image modality and text modality learned hash representations, α is the control parameter under different hash bit, which is set to $\alpha = 2^{-\log_2^{(c/32)}}$, and $\sigma\left(\theta_{ij}\right) = \frac{1}{1+e^{-\theta_{ij}}}$. We use the negative log likelihood loss as Hamming distance loss, which can be formulated as follows:

$$
\begin{aligned}
\mathcal{L}_{intra-image} &= \sum_{k=1}^{K+1}\left(-\sum_{i,j=1}^{N}\log p\left(S_{ij}^{xx} \mid f_{x_i}^k, f_{x_j}^k\right)\right) \\
&= -\sum_{k=1}^{K+1}\sum_{i,j=1}^{N}\left(S_{ij}^{xx}\theta_{x_i^k x_j^k} - \log\left(1+e^{\theta^k_{x_i^k x_j}}\right)\right) \\
\mathcal{L}_{intra-text} &= \sum_{k=1}^{K+1}\left(-\sum_{i,j=1}^{N}\log p\left(S_{ij}^{yy} \mid g_{y_i}^k, g_{y_j}^k\right)\right) \\
&= -\sum_{k=1}^{K+1}\sum_{i,j=1}^{N}\left(S_{ij}^{yy}\theta^k_{y_i^k y_j} - \log\left(1+e^{\theta^k_{y_i^k y_j}}\right)\right)
\end{aligned}
\tag{10}
$$

where $\theta_{x_i^k *_j} = \alpha f_{x_i}^k F_{*j}^T$, and $\theta_{y_i^k *_j} = \alpha g_{y_i}^k G_{*j}^T$. Then the intra-modality Hamming distance loss can be defined as:

$$
\mathcal{L}_{intra} = \mathcal{L}_{\text{intra-image}} + \mathcal{L}_{\text{intra-text}} \tag{11}
$$

Quantization Loss Although the cosine triplet loss can preserve high-ranking correlationship, such a surrogate may suffer from the disagreement between small cosine distance and large Hamming distance. As shown in Fig. 3,their cosine distance is small while they lie on different side of hyper-plane, which makes their Hamming distance still large. Thus, we design the cosine quantization loss as follows:

$$
\begin{aligned}
\mathcal{L}_{cos} &= \mathcal{L}_{cos}^x + \mathcal{L}_{\cos}^y \\
&= \sum_i \frac{1}{1+\log\left(\exp\left(\cos\left(\left|\mathbf{u}_i\right|, 1\right)\right)\right)} \\
&+ \sum_i \frac{1}{1+\log\left(\exp\left(\cos\left(\left|\mathbf{v}_i\right|, 1\right)\right)\right)}
\end{aligned}
\tag{12}
$$

Modality Invariance Loss. To obtain a lower quantization loss when projecting the hash representations $f_{x_i} \in R^c$ and $g_{y_i} \in R^c$ to hash codes $\boldsymbol{b}_i^x = \operatorname{sgn}\left(\boldsymbol{f}_i\right) \in \{-1,1\}^c$ and $\boldsymbol{b}_i^y = \operatorname{sgn}\left(\boldsymbol{g}_i\right) \in \{-1,1\}^c$, specially in some unbalanced outlier codes, we proposed modality invariance loss to minimize the distance between the hash codes of image-text pairs. The modality invariance loss are formulated as follows:

$$
\mathcal{L}_m = \frac{1}{c}\|\mathbf{B}^{\mathbf{x}} - \mathbf{B}^{\mathbf{y}}\|_F \tag{13}
$$

3.4. Optimization

By assembling the above five loss functions together, the final overall loss function is given as follows:

$$
\begin{aligned}
\min_{B,\theta_x,\theta_y,\theta_{D_x},\theta_{D_y}} \mathcal{L} &= \mathcal{L}_{tri} + \mathcal{L}_{intra} + \mathcal{L}_{cos} + \mathcal{L}_m + \widehat{\mathcal{L}}_{adv} \\
\widehat{\mathcal{L}}_{adv} &= \max_{\theta_{G_x},\theta_{Gy}} \mathcal{L}_{adv}
\end{aligned}
\tag{14}
$$

where θ_x, θ_y are the parameters of the image and text networks expect the discriminators in the adversarial learning part. Note that $\theta_{G_x}, \theta_{Gy}$ are part of θ_x, θ_y. An alternating optimization strategy is employed to optimize equation 14. At the end of each epoch, some parameters will be optimized while others are fixed. The whole optimization algorithm for SAAGHN is outlined in Algorithm 1.

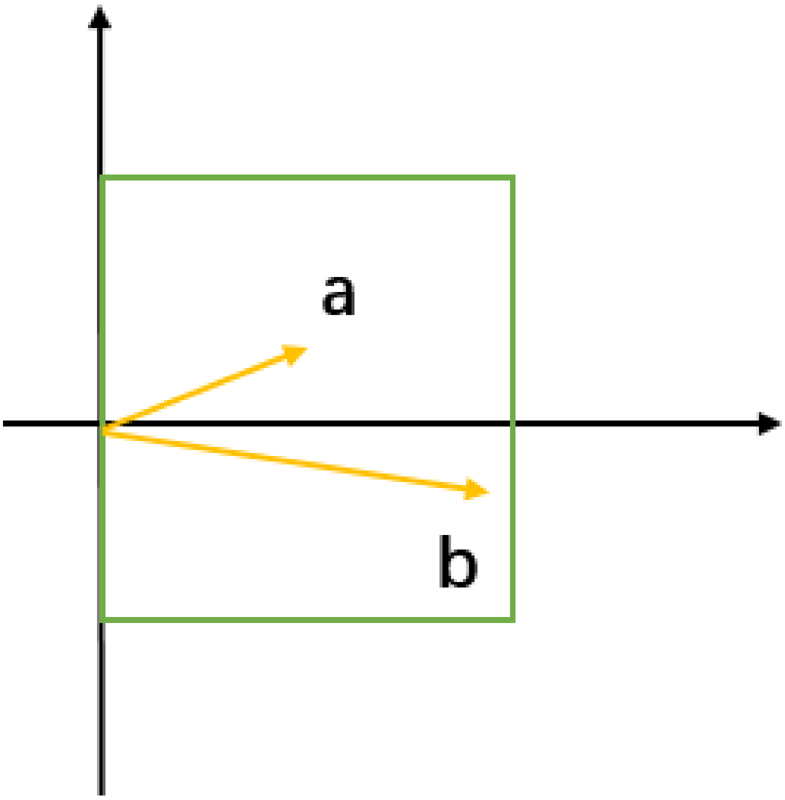

Figure 3. Example of gap between cosine distance and Hamming diatance.

4. Experiment and Discussion

In the experiment, SAAGH was executed on several popular cross-modal retrieval benchmark datastes MIRFLICKR-25K [38] and NUS-WIDE [39]. The experiments results show the superiority of our method.

4.1. Datasets

MIRFLICKR-25K [38] is a common used dataset which consists of 25,000 image-text pairs collected from Flickr. Each image is annotated by its associated textual tags. The experiment protocols of DCMH [23] is followed by our experiment. All image-text pairs have at least 24 unique labels. The text for each pair is represented as a 1,386-dimensional Bag-of-Words(BoW) vector.

NUS-WIDE [39] dataset contains 269,468 image-text pairs, and each of which is annotated by at least 81 labels. The text for each instance is represented as a 1000-dimensional BoW vector. All of the instances without labels are removed. We select 190,421 image-text pairs labeled by 21 most frequently used concepts.

Besides, 10,000 and 10,500 instances are randomly selected as the training set for MIRFLICKR-25K and NUS-WIDE. And 2,000 and 2,100 additional instances are randomly chosen as the query set of MIRFLICKR-25K and NUS-WIDE. Moreover, the remaining data are used as the retrieval set after picking the query data. Table 1 summarize the statistics of instances of the datasets.

4.2. Implementation Details

For both image and text networks in our SAAGHN, Resnet50 with four blocks is employed as the basic neural network to extract the fused representations. Furthermore, pooling size of 1, 5, 10, 15, 30 and 50 of BoW representations are generated as the text-modality input. Resnet50 for image

Algorithm 1: Optimization algorithm of SAAGHN.

```
Input: Training set $\{x_i, y_i, l_i\}_{i=1}^{N}$, cross-modal similarity matrix $S$;
Output: Optimized parameters $\theta_x$, $\theta_y$, $\theta_{D_x}$, $\theta_{D_y}$ of networks, and binary codes $B$;
Initialization: Initialize the parameters of network, set the mini-batch size to $n_x = n_y = 128$,
  initialize hash representations of each modality: F and G, set iteration number iter and other
  hyper-parameters.
for t=1 to iter do
    Update $\theta_{G_*}$ by ascending their gradients:
        $$\theta_{G_*} \leftarrow \theta_{G_*} + \eta \cdot \nabla_{\theta_{G_*}} \frac{1}{n} \mathcal{L}_{adv}, * \in \{x, y\}$$
    Update $\theta_{D_*}$ by descending theur gradients:
        $$\theta_{D_*} \leftarrow \theta_{D_*} - \eta \cdot \nabla_{\theta_{D_*}} \frac{1}{n} \mathcal{L}_{adv}, * \in \{x, y\}$$
    Update $\theta$ by BP algorithm:
        $$\theta \leftarrow \theta - \eta \cdot \nabla_{\theta} \frac{1}{n} (\mathcal{L} - \mathcal{L}_{adv})$$
end
Update binary codes $B$
Until a fixed number of iterations or convergence;
```

Dataset	Total	Train/Test
MIRFLCKR-25K	20,015	$10,000/2,000$
NUS-WIDE	190,421	$10,500/2,100$

Table 1. Statistics of instances of the dataset

network is initialized by the pre-trained model on the ImageNet [40] dataset, while the initialization of Resnet50 in text network is based on the Gaussion distribution $N\left(\mu, \sigma^2\right)$ with $\mu = 0$ and $\sigma = 0.1$. We implemented our SAAGHN in Pytorch [41] with two NVIDIA TITAN Xp GPU. The mini-batch size is set to 128 and the learning rate is 0.0003.

4.3. Metrics and Baselines

4.3.1. Evaluation criteria

Hamming ranking is widely used as retrieval procedure of hashing-based retrieval methods. The Hamming distance is sorted based on hamming distance and Mean Precision (MAP) [42] is a common used metric. In our experiment, MAP is adopted as the evaluation criteria for SAAGHN.

4.3.2. Baselines

Seven state-of-the-art cross-modal hashing methods are adopted to compare with SAAGHN, including Semantic Correlation Maximization (SCM) [15], Deep Cross-modal Hashing (DCMH) [23], Cross-modal Hamming Hashing (CMHH) [43], Self-supervised Adversarial Hashing (SSAH) [27], Pairwise Relationship Guided Deep Hashing (PRDH) [25], Correlation Hashing Network (CHN) [24], Semantic-preserving Hashing (SePH) [14] and Self-Constraining and Attention-based Hashing Network (SCAHN) [36].

	MIRFLICKR-25K						NUS-WIDE					
Method	Image query Text			Text query Image			Image query Text			Text query Image		
	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
SCM[15]	0.6354	0.5618	0.5634	0.6340	0.6458	0.6541	31.21	31.11	31.21	42.61	43.72	44.78
SePH[14]	0.6740	0.6813	0.6830	0.7139	0.7258	0.7294	0.4797	0.4859	0.4906	0.6072	0.6280	0.6291
DCMH[23]	0.7316	0.7343	0.7446	0.7607	0.7737	0.7805	0.5445	0.5597	0.5803	0.5793	0.5922	0.6014
CHN[24]	0.7504	0.7495	0.7461	0.7776	0.7775	0.7798	0.5754	0.5966	0.6015	0.5816	0.5967	0.5992
PRDH[25]	0.6952	0.7072	0.7108	0.7626	0.7718	0.7755	0.5919	0.6059	0.6116	0.6155	0.6286	0.6349
SSAH[27]	0.7745	0.7882	0.7990	0.7860	0.7974	0.7910	0.6163	0.6278	0.6140	0.6204	0.6251	0.6215
SCAHN[36]	0.8168	0.8311	0.8348	0.8034	0.8105	0.8193	0.6429	0.6510	0.6635	0.6501	0.6575	0.6685
CMHH[43]	0.7334	0.7281	0.7444	0.7320	0.7183	0.7279	0.5530	0.5698	0.5924	0.5739	0.5786	0.5889
SAAGHN	**0.8146**	**0.8321**	**0.8392**	**0.8133**	**0.8201**	**0.8276**	**0.6542**	**0.6765**	**0.6917**	**0.6485**	**0.6743**	**0.6843**

Table 2. Mean Average Percision (MAP) comparison results

4.4. Performance Evaluation

The MAP results of SAAGHN and other baselines for different hash code lengths are shown in Table 2 . From the table, we have the following observations.

- SAAGHN outperforms all the other state-of-the-art methods, which demonstrates the superiority of our method in cross-modal retrieval. The superiority is partly because SAAGHN explores rich fused hash representations as well as the high-ranking correlation preservation in cosine triplet loss.
- In some baselines, the results of Image-query-Text greatly outperform the Text-query-Image, while SAAGHN can achieve an almost equal performance. This may because the adversarial learning part and modality invariance loss can fully preserve the modality-invariability.
- Deep cross-modal hashing methods (DCMH, CMHH, SSAH, PRDH, CHN and SACHN) have an improved performance than the other hand-crafted cross-modal hashing methods(SCM and SePH). This proves that the deep hash representations learned from datasets are more robust than hand-crafted representations.

4.5. Ablation Study

The impact of different modules of SAAGHN are verified in this section. To test the effectiveness of our contributions, three ablation experiments are designed as follows:

- SAAGHN-COS: This experiment is designed to replace batch semi-hard cosine triplet loss by cosine triplet loss used in AGAH.
- SAAGHN-ATT: This experiment is designed based on SAAGHN without self-attention fusion module, which means the hash representations are generated from the last layer of ImageNet and TextNet.
- SAAGHN-ADV: This experiment means that SAAGHN is used without adversarial learning part, which means the discriminator are removed as well as the adversarial loss.

Table 3 shows the results of ablation study. Firstly, we can observe that the designed batch semi-hard cosine triplet loss can achieve better results than cosine triplet loss. This may partly because the batch semi-hard cosine triplet loss can avoid early local minima. Secondly, we can observe that the performance improves a lot by using self-attention fusion module. This is probably because that the self-attention fusion module can explore more semantic relevant information and global dependencies. Finally, the improvement of adversarial learning part is competitive.

5. Conclusion

In this paper, a novel deep hashing method for cross modal retrieval (SAAGHN) is proposed, which not only learns rich modality-invariant hash representations but also generates discriminative

	MIRFLICKR-25K		NUS-WIDE	
Method	Image query Text	Text query Image	Image query Text	Text query Image
SAAGHN-COS	0.8094	0.8101	0.6595	0.6536
SAAGHN-ATT	0.8157	0.8012	0.6693	0.6625
SAAGHN-ADV	0.8217	0.8160	0.6834	0.6728
SAAGHN	0.8392	0.8276	0.6917	0.6843

Table 3. MAP on two datasets for ablation study.

hash codes for cross-modal retrieval in an end-to-end framework. To enhance the feature learning part, a self-attention mechanism is introduced to explore global, long-range dependencies of hash representations among modalities. Furthermore, adversarial learning and batch semi-hard cosine triplet loss is adopted to learn the consistent and high-ranking hash codes. Moreover, we add a modality-invariance loss to minimize the gap between hash codes of image modality and text modality. Due to the contributions mentioned above, the optimal hash codes can be learned at the end of learning process. Extensive experiments demonstrate that SAAGHN outperforms the current hashing-based cross-modal retrieval methods.

6. *acknowledgements

The authors appreciate helpful comments from the reviewers on improving our work. This work was supported by the National Natural Science Foundation of China (61806168), Fundamental Research Funds for the Central Universities (SWU117059), and Venture & Innovation Support Program for Chongqing Overseas Returnees (CX2018075).

References

1. Wang, Y.; Wu, F.; Song, J.; Li, X.; Zhuang, Y. Multi-modal mutual topic reinforce modeling for cross-media retrieval. Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 307–316.
2. Jia, Y.; Salzmann, M.; Darrell, T. Learning cross-modality similarity for multinomial data. 2011 International Conference on Computer Vision. IEEE, 2011, pp. 2407–2414.
3. Gong, Y.; Ke, Q.; Isard, M.; Lazebnik, S. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision* **2014**, *106*, 210–233.
4. Zhuang, Y.; Wang, Y.; Wu, F.; Zhang, Y.; Lu, W. Supervised coupled dictionary learning with group structures for multi-modal retrieval. Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, 2013, pp. 1070–1076.
5. Mao, X.; Lin, B.; Cai, D.; He, X.; Pei, J. Parallel field alignment for cross media retrieval. Proceedings of the 21st ACM international conference on Multimedia, 2013, pp. 897–906.
6. Karpathy, A.; Joulin, A.; Fei-Fei, L.F. Deep fragment embeddings for bidirectional image sentence mapping. Advances in neural information processing systems, 2014, pp. 1889–1897.
7. Wang, J.; He, Y.; Kang, C.; Xiang, S.; Pan, C. Image-text cross-modal retrieval via modality-specific feature learning. Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, 2015, pp. 347–354.
8. Wang, W.; Yang, X.; Ooi, B.C.; Zhang, D.; Zhuang, Y. Effective deep learning-based multi-modal retrieval. *The VLDB Journal* **2016**, *25*, 79–101.
9. Wang, Y.; Wu, F.; Song, J.; Li, X.; Zhuang, Y. Multi-modal mutual topic reinforce modeling for cross-media retrieval. Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 307–316.
10. Ding, G.; Guo, Y.; Zhou, J. Collective matrix factorization hashing for multimodal data. Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 2075–2082.
11. Song, J.; Yang, Y.; Yang, Y.; Huang, Z.; Shen, H.T. Inter-media hashing for large-scale retrieval from heterogeneous data sources. Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, 2013, pp. 785–796.

12. Zhang, J.; Peng, Y.; Yuan, M. Unsupervised generative adversarial cross-modal hashing. *arXiv preprint arXiv:1712.00358* **2017**.
13. Zhou, J.; Ding, G.; Guo, Y. Latent semantic sparse hashing for cross-modal similarity search. Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, 2014, pp. 415–424.
14. Lin, Z.; Ding, G.; Hu, M.; Wang, J. Semantics-preserving hashing for cross-view retrieval. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3864–3872.
15. Zhang, D.; Li, W.J. Large-scale supervised multimodal hashing with semantic correlation maximization. AAAI. Citeseer, 2014, Vol. 1, p. 7.
16. Zhang, P.F.; Li, C.X.; Liu, M.Y.; Nie, L.; Xu, X.S. Semi-relaxation supervised hashing for cross-modal retrieval. Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 1762–1770.
17. Xu, X.S. Dictionary learning based hashing for cross-modal retrieval. Proceedings of the 24th ACM international conference on Multimedia, 2016, pp. 177–181.
18. Yan, C.; Gong, B.; Wei, Y.; Gao, Y. Deep multi-view enhancement hashing for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2020**.
19. Girshick, R. Fast r-cnn. Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.
20. Guo, Y.; Liu, Y.; Oerlemans, A.; Lao, S.; Wu, S.; Lew, M.S. Deep learning for visual understanding: A review. *Neurocomputing* **2016**, *187*, 27–48.
21. Han, X.F.; Laga, H.; Bennamoun, M. Image-based 3D Object Reconstruction: State-of-the-Art and Trends in the Deep Learning Era. *IEEE transactions on pattern analysis and machine intelligence* **2019**.
22. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 2015, pp. 91–99.
23. Jiang, Q.Y.; Li, W.J. Deep cross-modal hashing. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3232–3240.
24. Cao, Y.; Long, M.; Wang, J.; Yu, P.S. Correlation hashing network for efficient cross-modal retrieval. *arXiv preprint arXiv:1602.06697* **2016**.
25. Yang, E.; Deng, C.; Liu, W.; Liu, X.; Tao, D.; Gao, X. Pairwise relationship guided deep hashing for cross-modal retrieval. Thirty-first AAAI conference on artificial intelligence, 2017.
26. Cao, Y.; Long, M.; Wang, J.; Liu, S. Collective deep quantization for efficient cross-modal retrieval. AAAI, 2017, Vol. 1, p. 5.
27. Li, C.; Deng, C.; Li, N.; Liu, W.; Gao, X.; Tao, D. Self-supervised adversarial hashing networks for cross-modal retrieval. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4242–4251.
28. Gu, W.; Gu, X.; Gu, J.; Li, B.; Xiong, Z.; Wang, W. Adversary guided asymmetric hashing for cross-modal retrieval. Proceedings of the 2019 on International Conference on Multimedia Retrieval, 2019, pp. 159–167.
29. Deng, C.; Chen, Z.; Liu, X.; Gao, X.; Tao, D. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Transactions on Image Processing* **2018**, *27*, 3893–3903.
30. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. Advances in neural information processing systems, 2014, pp. 2672–2680.
31. Wang, B.; Yang, Y.; Xu, X.; Hanjalic, A.; Shen, H.T. Adversarial cross-modal retrieval. Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 154–162.
32. Norouzi, M.; Fleet, D.J. Minimal loss hashing for compact binary codes. ICML, 2011.
33. Yang, Z.; He, X.; Gao, J.; Deng, L.; Smola, A. Stacked attention networks for image question answering. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 21–29.
34. Li, W.; Zhu, X.; Gong, S. Harmonious attention network for person re-identification. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2285–2294.
35. Zhang, X.; Zhou, S.; Feng, J.; Lai, H.; Li, B.; Pan, Y.; Yin, J.; Yan, S. HashGAN: attention-aware deep adversarial hashing for cross modal retrieval. *arXiv preprint arXiv:1711.09347* **2017**.
36. Wang, X.; Zou, X.; Bakker, E.M.; Wu, S. Self-Constraining and Attention-based Hashing Network for Bit-Scalable Cross-Modal Retrieval. *Neurocomputing* **2020**.
37. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.

38. Huiskes, M.J.; Lew, M.S. The MIR flickr retrieval evaluation. Proceedings of the 1st ACM international conference on Multimedia information retrieval, 2008, pp. 39–43.
39. 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. Proceedings of the ACM international conference on image and video retrieval, 2009, pp. 1–9.
40. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; others. Imagenet large scale visual recognition challenge. *International journal of computer vision* **2015**, *115*, 211–252.
41. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; others. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 2019, pp. 8026–8037.
42. Liu, Wei andimagenet Mu, C.; Kumar, S.; Chang, S.F. Discrete graph hashing. Advances in neural information processing systems, 2014, pp. 3419–3427.
43. Cao, Y.; Liu, B.; Long, M.; Wang, J. Cross-modal hamming hashing. Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 202–218.