ATTENTION-GUIDED SEMANTIC HASHING FOR UNSUPERVISED CROSS-MODAL RETRIEVAL

Xiao Shen[1], Haofeng Zhang[1,], Lunbo Li[1], Li Liu[2]*

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[2]Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates
Email:{shenxiao, zhanghf, lunboli}@njust.edu.cn, liuli1213@gmail.com

2021 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-3864-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICME51207.2021.9428330

ABSTRACT

Recently, due to the low storage consumption and high search efficiency of hashing methods and the powerful feature extraction capability of deep neural networks, deep cross-modal hashing has received extensive attention in the field of multimedia retrieval. However, existing methods tend to ignore the latent relationships between heterogeneous data when learning a common semantic subspace, and cannot retain more important semantic information when mining deep correlations. In this paper, an attention mechanism which focuses on the characteristics of the associated features is employed to propose an attention-aware semantic fusion matrix that integrates important information from different modalities. We introduce a novel network that can pass the extracted features through the attention module to efficiently encode rich and relevant features, and can also generate hash codes under the self-supervision of the proposed attention-aware semantic fusion matrix. Our experimental results and detailed analysis prove that our method can achieve better retrieval performance on the three popular datasets, compared with the recent unsupervised cross-modal hashing methods.

Index Terms— cross-modal hashing, attention mechanism, semantic similarity learning

1. INTRODUCTION

With the rapid growth of big data in recent years and the massive popularity of social media, cross-modal data is ubiquitous on the Internet [1]. Therefore, cross-modal retrieval, which aims to use different modal queries to search for relevant semantic examples, has received extensive attention in the situation of processing a large number of complex media data. As we know, instances of different modalities are heterogeneous data with different feature representations and spatial distributions, and how to effectively unify heterogeneous data into the common subspace and reduce their semantic gap is an important part of cross-modal retrieval [2].

The existing cross-modal retrieval methods can be roughly divided into two categories [3, 4]. The first is real-valued latent embedding learning from different modalities. However, directly learning real-valued latent embeddings has high computational complexity and low search efficiency. Therefore, a second method called cross-modal hashing is proposed, which converts the original high-dimensional heterogeneous data into compact binary hash codes in the common Hamming space. Benefiting from the low storage consumption and high search efficiency of hashing methods, cross-modal hashing has been widely used in the field of large-scale multimedia retrieval [5].

Cross-modal hashing methods can be simply divided into two categories: supervised and unsupervised. For supervised methods [6], the usage of semantic tags can help to capture the similarity between modal data, thereby learning more consistent and distinguishable hash codes [7]. The unsupervised cross-modal hashing methods only use the characteristic information of the input pair-wise data to maximize their similarity in the Hamming space [8]. Without the label information, the effect of the unsupervised methods is significantly inferior to the supervised ones. However, since it is difficult to obtain a large amount of labeled data, unsupervised methods are more practical and popular in realistic scenarios. The existing unsupervised methods mainly rely on inter-correlations and intra-correlations to guide hash code learning, and cannot explore the overall structure of each modality [9]. How to bridge the modal gap and learn a robust latent subspace is the focus of our work.

In this paper, we propose a novel unsupervised cross-modal hashing method named Attention-Guided Semantic Hashing (AGSH) to solve the problem mentioned above. This method ensures that the multi-modal semantic similarity and attended information are both preserved. The main contributions of this work can be summarized as follows:

- We introduce a cross-modal hashing method that incorporates an attention mechanism, which can strengthen

[*]Corresponding author. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants No. 61872187 and No. 62072246, in part by the Natural Science Foundation of Jiangsu Province under Grant No. BK20201306, and in part by the "111" Program under Grant No. B13022.

978-1-6654-3864-3/21/$31.00 ©2021 IEEE

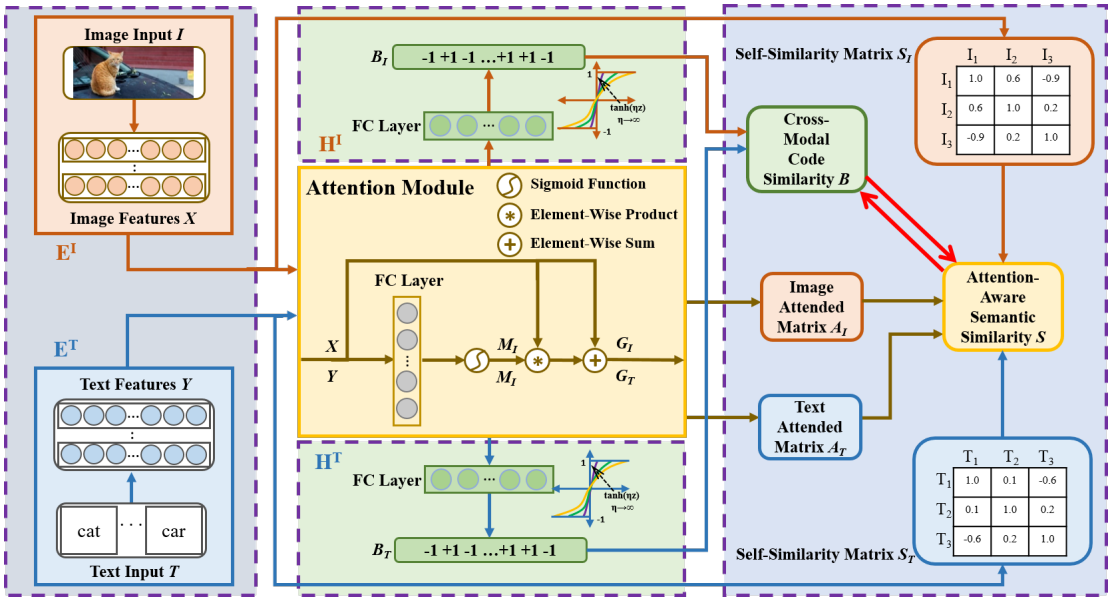

Fig. 1. The framework of our proposed **A**ttention-**G**uided **S**emantic **H**ashing(AGSH) for unsupervised cross-modal retrieval. It contains four main parts: the left part is feature extraction module, the center part is attention module, the upper part and the part below are hashing module, and the right part is the optimization part.

important information and suppress unimportant information, thereby reducing information loss in the process of data dimensionality reduction.

- We construct an attention-aware semantic fusion matrix, which combines the original self-similarity matrix with the learned self-attention matrix, to boost the semantic representation of similarity matrix, and ensure both the cross-modal invariability and the semantic distinctiveness of learned features.
- Experimental results on three benchmark datasets show the priority of the proposed model compared with other state-of-the-art deep unsupervised cross-modal hashing methods.

2. RELATED WORK

2.1. Unsupervised Cross-Modal Hashing

Earlier cross-modal hashing work mainly uses hand-crafted features. For example, Collective Matrix Factorization Hashing (CMFH) [10] learns unified hash codes through collective matrix factorization of latent factor models with different modal data. Inter-Media Hashing (IMH) [11] extends Spectral Hashing [12] to multi-modal scenes, with the goal of maintaining semantic consistency within and between modalities in a common space. However, these methods are hard to be promoted, because the hand-crafted features contain insufficient information and the extraction process of them is time-consuming and laborious. The powerful feature extraction capabilities of deep neural networks alleviate this trouble. Deep Cross-Modal Hashing (DCMH) [7] shows the superiority of deep neural networks in eliminating the modal gap. Unsupervised methods usually use the co-occurrence information of the input pair-wise modal data to maximize their correlation in the Hamming space. Unsupervised Cross-Modal Hashing (UDCMH) [13] combines matrix decomposition and Laplacian restriction during network training, and explicitly constrains the hash codes to retain the domain characteristics of the original data, so as to obtain better performance. Unsupervised Generative Adversarial Cross-modal Hashing (UGACH) [14], narrows the gap between modalities through the minimax game. Deep Joint-Semantics Reconstructing Hashing (DJSRH) [15] fuses the semantic similarities into a unified affinity matrix to capture the latent relevance for the input multi-modal instances. Although these methods have achieved acceptable results, they mostly neglect to preserve the attended information among all pairs.

2.2. Attention Mechanism

Recently, the attention mechanism based on neural networks has attracted widespread attention due to its satisfactory performance in different applications, such as image and video caption, image classification, and visual question answering. The attention mechanism can be trained to recognize the places where should be paid attention to in a particular task. [16] shows a residual attention network to learn more discriminative features in image classification tasks. [17] represents self-attention thoughts can be used in aggregated data. However, only a few papers discuss the attention mechanism based cross-modal retrieval tasks.

In this work, we try to bridge the semantic gap using the attention mechanism, and further capture various seman-

tic relevance and representation consistency through the self-supervision of the proposed attention-aware semantic fusion matrix.

3. PROPOSED MODEL

3.1. Problem Definition

Suppose there are N training samples, each of which is represented in several modalities, e.g., image, text, video and audio. In this paper, we only focus on two modalities: image and text, which are represent as $\{I_i, T_i\}_{i=1}^{N}$, where I_i denotes the i-th image and T_i represents the corresponding text description of image I_i. The goal of our algorithm is to learn k-bit binary representations $\boldsymbol{B}_I = \{\boldsymbol{b}_{I_1}, ..., \boldsymbol{b}_{I_i}, ..., \boldsymbol{b}_{I_N}\} \in \{-1, +1\}^{k \times N}$ and $\boldsymbol{B}_T = \{\boldsymbol{b}_{T_1}, ..., \boldsymbol{b}_{T_i}, ..., \boldsymbol{b}_{T_N}\} \in \{-1, +1\}^{k \times N}$ for I and T. Two effective hash functions are expected to be learned to project the input data of each modality to the common Hamming space. In this paper, for the matched cross-modal data pair (I_i, T_i), it is expected that their corresponding binary codes $\boldsymbol{b}_{I_i}$ and $\boldsymbol{b}_{T_i}$ has smaller Hamming distance than the unmatched pairs.

3.2. Proposed Architecture

The proposed attention-guided semantic hashing network consists of three components: 1) the feature extraction module to obtain feature representations of high-dimensional multi-modal data; 2) the attention module to generate features filtered by the attention mechanism; 3) the hashing module to learn the semantic-preserving hash functions.

Feature Extraction Module E^I and E^T: We follow the recent convention to use the fc7 features of pre-trained VGG-16 as the image feature input and the universal sentence encoder features [18] for text input. Denote $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_i, ..., \boldsymbol{x}_N\} \in \mathbb{R}^{d \times N}$ and $\boldsymbol{Y} = \{\boldsymbol{y}_1, ..., \boldsymbol{y}_i, ..., \boldsymbol{y}_N\} \in \mathbb{R}^{d \times N}$ as the outputs of the image network and text network respectively, where $\boldsymbol{x}_i = E^I(I_i)$ and $\boldsymbol{y}_i = E^T(T_i)$. We use the pre-trained feature extraction networks and fix them during training.

Attention Module: When deep neural network is used to process a large amount of input samples, referring to the attention mechanism of the human brain, only some key input information is selected for processing to improve the efficiency of deep neural network and the accuracy of targeted tasks. Inspired by the latest work of [16], we employ the attention module to learn the mask weights of the extracted cross-modal features. The mask weights $\boldsymbol{M}_I \in \mathbb{R}^{d \times N}$ and $\boldsymbol{M}_T \in \mathbb{R}^{d \times N}$ of image feature $\boldsymbol{X}$ and text feature $\boldsymbol{Y}$ are as follows:

$$\begin{cases} \boldsymbol{M}_I = Sigmoid(\boldsymbol{W}\boldsymbol{X} + \boldsymbol{d}), \\ \boldsymbol{M}_T = Sigmoid(\boldsymbol{W}\boldsymbol{Y} + \boldsymbol{d}), \end{cases} \tag{1}$$

where, Sigmoid means the Sigmoid activation function, and $\boldsymbol{W}$ and $\boldsymbol{d}$ are the parameters of the shared fully connected layer.

The attention module is shown in the center of Fig.1. Firstly, the fully connected layer is used to perform nonlinear transformation of the feature vector and learn the global semantic information, so as to achieve the purpose of grabbing richer semantic information and discarding irrelevant information. Then, in order to control the mask weights in the range of [0,1], we use the Sigmod activation function to process the weight values.

In addition, the weight sharing mechanism can embed visual attention and text attention into the same attention space. This indicates that the features of one modality can learn to activate the corresponding and most relevant features of other modalities. Because the features are mapped to the same attention space, the gap between the multiple modalities can also be reduced. The pair-wise modal data can retain the same attention features, which is more conducive to the preservation of semantic similarity; and if the features of one modality cannot be fully distinguished, the same attention space can produce and enhance more distinguishing features.

In order to make the cross-modal retrieval more effective, after obtaining the mask weights, we can use them to generate semantic embedding features. However, simple dot production will make the activation values small, so we add the original features to the attention module, which not only enables the network to embed more complex semantic information, but also preserves the original semantic features of each modality. In addition, the weighted mask can select more important and relevant regions to aggregate local features. The image and text outputs of the attention module are as follows:

$$\begin{cases} \boldsymbol{G}_I = \boldsymbol{X} + \boldsymbol{M}_I \cdot \boldsymbol{X}, \\ \boldsymbol{G}_T = \boldsymbol{Y} + \boldsymbol{M}_T \cdot \boldsymbol{Y}, \end{cases} \tag{2}$$

where, $\boldsymbol{G}_I \in \mathbb{R}^{d \times N}$ and $\boldsymbol{G}_T \in \mathbb{R}^{d \times N}$ are the outputs of the attention module of the intersection of image and text data. It is worth emphasizing is that multiply $\boldsymbol{M}_I\boldsymbol{X}$ and $\boldsymbol{M}_T\boldsymbol{Y}$ element by element.

Hashing Module H^I and H^T: Finally, $\boldsymbol{G}_I$ and $\boldsymbol{G}_T$ are fed into a fully connected layer, and then binarized to generate the final binary representations $\boldsymbol{B}_I$ of image modality and $\boldsymbol{B}_T$ of text modality, respectively.

3.3. Construct Attention-Aware Semantic Matrix

We try to use the attention-aware semantic fusion matrix as self-supervised information to better capture the semantic consistency between different modalities. Similar to the thinking of data fusion [15], the attention-aware semantic fusion matrix is to fuse the self-similarity matrix and the self-attention matrix from different modalities.

Firstly, after normalizing $\boldsymbol{X}$, $\boldsymbol{Y}$ to $\hat{\boldsymbol{X}}$, $\hat{\boldsymbol{Y}}$, we can calculate the feature self-similarity matrices to describe the original neighborhood structure for the input image data and text data

respectively. They are defined as follows:

$$\begin{cases} \boldsymbol{S}_I = \hat{\boldsymbol{X}}^T \hat{\boldsymbol{X}} \in [-1,+1]^{n\times n}, \\ \boldsymbol{S}_T = \hat{\boldsymbol{Y}}^T \hat{\boldsymbol{Y}} \in [-1,+1]^{n\times n}, \end{cases} \tag{3}$$

where, n is the batch size.

Secondly, after normalizing $\boldsymbol{G_I}$, $\boldsymbol{G_T}$ to $\hat{\boldsymbol{G}}_{\boldsymbol{I}}$, $\hat{\boldsymbol{G}}_{\boldsymbol{T}}$, we can calculate the attended matrices to describe the self-attenion similarity structure for image and text modality respectively. They are calculated as follows:

$$\begin{cases} \boldsymbol{A}_I = \hat{\boldsymbol{G}}_I^{\ T} \hat{\boldsymbol{G}}_I \in [-1,+1]^{n\times n}, \\ \boldsymbol{A}_T = \hat{\boldsymbol{G}}_T^{\ T} \hat{\boldsymbol{G}}_T \in [-1,+1]^{n\times n}. \end{cases} \tag{4}$$

Thirdly, we integrate the self-similarity matrices and the self-attention matrices, and represent as:

$$\boldsymbol{S}_{fuse} = \gamma \frac{\boldsymbol{S}_I \boldsymbol{A}_I^T}{n} + (1-\gamma)\frac{\boldsymbol{S}_T \boldsymbol{A}_T^T}{n}, \tag{5}$$

where, γ is the balancing coefficient to adjust the influence of different modalities. Calculating $\boldsymbol{S}_I \boldsymbol{A}_I^T$ and $\boldsymbol{S}_T \boldsymbol{A}_T^T$ respectively is to make the self-similarity matrix of each modality weighted with the corresponding self-attention matrix. It is worth noting that $\boldsymbol{S}_I \boldsymbol{A}_I^T \in [-1,+1]^{n\times n}$, $\boldsymbol{S}_T \boldsymbol{A}_T^T \in [-1,+1]^{n\times n}$.

Finally, according to Eq.5, the equation of the attention-aware semantic matrix is represented as follows:

$$\boldsymbol{S} = \lambda \boldsymbol{S}_{fuse} + (1-\lambda)\frac{\boldsymbol{S}_{fuse}\boldsymbol{S}_{fuse}^T}{n}, \tag{6}$$

where, λ is the trade-off parameter to control the balance of the two items. We calculate $\boldsymbol{S}_{fuse}\boldsymbol{S}_{fuse}^T$ to achieve a high order neighborhood description based on the principle that two semantic relevant instances should share the same similarity relations with other instances. It is also should be noted that $\boldsymbol{S}_{fuse}\boldsymbol{S}_{fuse}^T \in [-1,+1]^{n\times n}$.

Meanwhile, we can regard the generated binary codes $\boldsymbol{B}_I$ and $\boldsymbol{B}_T$ as feature vectors of the hyper-cube vertex. From this perspective, adjacent vertices correspond to similar binary codes, and the distance between binary codes can be expressed by their angular distance. We calculate the pair-wise cosine similarity to describe the neighborhood structure of the Hamming space, the formula is defined as follows:

$$cos(\boldsymbol{b}_{I_i}, \boldsymbol{b}_{T_j}) = \frac{\boldsymbol{b}_{I_i}^T \boldsymbol{b}_{T_j}}{\|\boldsymbol{b}_{I_i}\|_2 \|\boldsymbol{b}_{T_j}\|_2} \in [-1,+1], \tag{7}$$

where, $\boldsymbol{b}_{I_i}$ represents the i-th row of $\boldsymbol{B}_I$, and $\boldsymbol{b}_{T_j}$ represents the j-th row of $\boldsymbol{B}_T$.

By minimizing the error between the attention-aware semantic fusion matrix $\boldsymbol{S}$ and the to-be-learned hash codes structure $cos(\boldsymbol{B}_I, \boldsymbol{B}_T)$, the semantic consistency between the different modalities is effectively preserved. The formula is defined as follows:

$$L_{cross} = \min_{\boldsymbol{B}_I, \boldsymbol{B}_T} \|\boldsymbol{S} - cos(\boldsymbol{B}_I, \boldsymbol{B}_T)\|_F^2. \tag{8}$$

Similarly, the intra-modal consistency in image and text modality is preserved by minimizing the error between the attention-aware semantic fusion matrix S and the to-be-learned hash codes structure $cos(\boldsymbol{B}_I, \boldsymbol{B}_I)$ and $cos(\boldsymbol{B}_T, \boldsymbol{B}_T)$:

$$\begin{cases} L_{img} = \min\limits_{\boldsymbol{B}_I} \|\boldsymbol{S} - cos(\boldsymbol{B}_I, \boldsymbol{B}_I)\|_F^2, \\ L_{txt} = \min\limits_{\boldsymbol{B}_T} \|\boldsymbol{S} - cos(\boldsymbol{B}_T, \boldsymbol{B}_T)\|_F^2. \end{cases} \tag{9}$$

The final training objective function of the proposed AGSH is as follows:

$$L = \alpha L_{cross} + \beta L_{img} + \mu L_{txt}, \tag{10}$$

where, α and β, μ are coefficients to balance the relationship between intra-modal semantic similarity and inter-modal semantic consistency.

3.4. Optimization

Eq.10 cannot be directly optimized with deep network due to the existence of $\boldsymbol{B}_I$ and $\boldsymbol{B}_T$. If the output of the last hidden layer of the network is used to generate binary codes by the sgn function, the gradient of the sgn function is zero for all non-zero inputs, during the process of backward propagation, which will destructively block the gradients back to the previous layers. This problem is often called gradient vanishing problem, and we use the scaled tanh function to solve it:

$$\boldsymbol{B} = tanh(\eta \boldsymbol{H}) \in [-1,+1]^{k\times N}, \eta \in \mathbb{R}^+, \tag{11}$$

where, $\boldsymbol{H}$ is the output of the fully connected layer after the feature matrix $\boldsymbol{G}$. As η increases, according to $\lim\limits_{\eta\to\infty} tanh(\eta x) = sgn(x)$, Eq.10 with the original tricky binary coding problem can be efficiently optimized.

4. EXPERIMENTS AND EVALUATION

4.1. Datasets

We select three typical image and text cross-modal retrieval datasets, including Wiki, MIRFlickr-25K and NUS-WIDE, to evaluate the performance of the proposed AGSH. For detailed experiment settings, please refer to [9] for Wiki, and [23] for MIRFlickr-25K and NUS-WIDE.

4.2. Evaluational Metrics and Settings

Focusing on two kinds of cross-modal retrieval tasks, namely image2text (I2T) and text2image (T2I), we apply the trained hash coding functions to binarize the extracted features of

Table 1. mAP results for various code lengths of text retrieval performance by image query (I2T) and image retrieval performance by text query (T2I).

Task	Method	Wiki			MIRFlickr-25K			NUS-WIDE		
		16bit	32bit	64bit	16bit	32bit	64bit	16bit	32bit	64bit
$I \rightarrow T$	IMH [19]	0.151	0.145	0.133	0.558	0.566	0.560	0.350	0.357	0.371
	CMFH [10]	0.173	0.169	0.184	0.580	0.573	0.555	0.382	0.430	0.417
	ACQ [9]	0.126	0.120	0.115	0.617	0.594	0.576	0.440	0.416	0.396
	PDH [20]	0.196	0.168	0.150	0.544	0.544	0.546	0.369	0.369	0.369
	QCH [21]	0.160	0.144	0.132	0.580	0.566	0.555	0.402	0.382	0.371
	DBRC [22]	0.203	0.215	0.238	0.615	0.631	0.639	0.485	0.506	0.523
	UGACH [23]	0.359	0.376	0.385	0.643	0.679	0.680	0.541	0.535	0.554
	OURS AGSH	**0.397**	**0.434**	**0.446**	**0.679**	**0.691**	**0.698**	**0.543**	**0.552**	**0.562**
$T \rightarrow I$	IMH [19]	0.236	0.237	0.218	0.561	0.569	0.563	0.350	0.356	0.372
	CMFH [10]	0.176	0.170	0.179	0.583	0.567	0.556	0.394	0.452	0.448
	ACQ [9]	0.344	0.291	0.247	0.628	0.602	0.581	0.445	0.420	0.399
	PDH [20]	0.345	0.293	0.251	0.544	0.544	0.546	0.366	0.367	0.367
	QCH [21]	0.341	0.289	0.245	0.585	0.567	0.557	0.406	0.385	0.373
	DBRC [22]	0.244	0.258	0.268	0.619	0.632	0.640	0.492	0.519	0.425
	UGACH [23]	0.337	0.367	0.380	0.656	0.682	0.689	0.542	0.554	0.565
	OURS AGSH	**0.431**	**0.443**	**0.453**	**0.674**	**0.689**	**0.693**	**0.543**	**0.567**	**0.570**

each instance in the database and the query set for evaluating the retrieval performance. We use two common retrieval metrics, namely mean Average Precision (mAP) and Precision-Recall (P-R) curves, to evaluate the retrieval performance of the proposed AGSH and the baselines. For multi-label datasets, any two instances are considered to be the ground-truth neighbors if they share at least one common tag.

We set the batch size as 64 and employ the SGD optimizer with momentum of 0.8 and weight decay of 0.0005. α, β and μ are empirically set to 1, 0.1 and 0.1 respectively for all three datasets. We set γ=0.3 and λ=0.9 for Wiki, γ=0.9 and λ=0.6 for MIRFlickr-25K, γ=0.6 and λ=0.6 for NUS-WIDE. Besides, the learning rate of image network is set to 0.001, and the learning rate of text network is set to 0.001.

4.3. Performance

We compare our proposed AGSH model with several representative baselines, including IMH [19], CMFH [10], ACQ [9], PDH [20], QCH [21], DBRC [22] and UGACH [23]. The first four methods are shallow, while DBRC, UGACH and ours are all deep approaches. We first compare the mAP results with the baselines, and the results are shown in Tab. 1. It can be seen that under various encoding lengths and datasets, our proposed AGSH is significantly better than the latest unsupervised cross-modal hashing methods. Specifically, compared to shallow methods but use deep features as their image feature representations, the deep baselines can achieve better performance because they can back propagate the gradients to the front network to learn more effective hash coding functions.

For a quantitative comparison, our I2T performance on Wiki obtains at least 22% improvement on 16 bits, 32 bits and 64 bits compared with the other non-deep algorithms while our T2I performance surpasses those methods more than 9%, 16% and 21% respectively for different lengths of binary codes. On the other two datasets, the improvements of our method are also obvious, especially compared with those unsupervised non-deep cross-modal hashing techniques.

Table 2. mAP of ablation study on Wiki and MIRFlickr-25K.

Method(Task)	Wiki			MIRFlickr-25K		
	16bit	32bit	64bit	16bit	32bit	64bit
AGSH-1(I2T)	0.354	0.365	0.371	0.625	0.648	0.656
AGSH-2(I2T)	0.372	0.387	0.409	0.637	0.659	0.661
AGSH(I2T)	**0.397**	**0.434**	**0.446**	**0.679**	**0.691**	**0.698**
AGSH-1(T2I)	0.388	0.389	0.402	0.625	0.638	0.657
AGSH-2(T2I)	0.390	0.395	0.413	0.634	0.644	0.673
AGSH(T2I)	**0.431**	**0.443**	**0.453**	**0.674**	**0.689**	**0.693**

Moreover, Fig.3 shows the Precision-Recall (P-R) curves among the compared methods, in which AGSH still significantly outperforms the state-of-the-art baselines on all three datasets, which further reveals the priority of our proposed scheme in the task of unsupervised cross-modal retrieval.

4.4. Ablation Studies

Component analysis: In this section, the structure of our model is slightly modified to check the significance of each component. In our experiments, we denote AGSH-1 to represent the architecture that uses the extracted features directly, and it does not pass through the attention module. AGSH-2 stands for the architecture using the semantic matrix directly without the multiplication of the self-attention matrices, *i.e.*, eq.5 is written as $\boldsymbol{S}_{fuse} = \gamma \boldsymbol{S}_I + (1-\gamma)\boldsymbol{S}_T$. The results of our ablation study are recorded in Tab. 2. We can observe that the union of attention module is very important, and the performance can be improved by at least 5% and 4% on Wiki and MIRFlickr-25K. From Tab. 2, we can also find that the both two components are indispensible in our method.

Hyper-parameter analysis: The hyper-parameters are also analyzed. In Fig. 2, we illustrate the influence of different loss penalties of α, β, μ and the effect of attention-

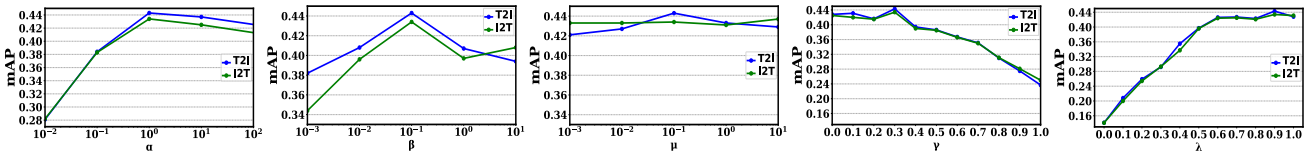

Fig. 2. Hyper-parameter analysis of α, β, μ, γ and λ for Wiki with 32-bit codes.

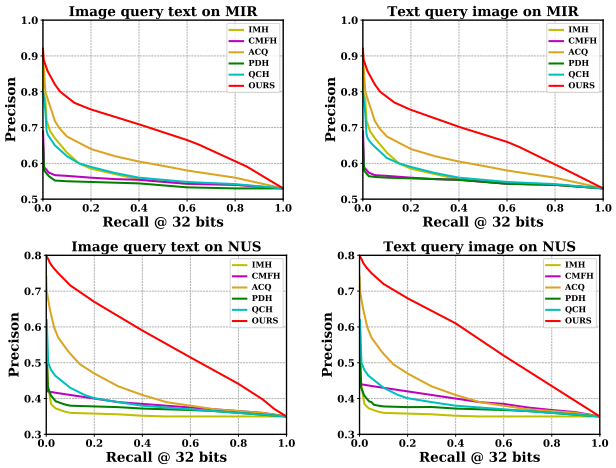

Fig. 3. Results of Precision-Recall curves of various cross-modal hashing methods on MIRFlickr-25K and NUS-WIDE with 32-bit codes.

aware semantic fusion matrix parameters γ, λ on Wiki dataset with 32-bit codes. It can be discovered that the influence of γ and λ is slightly larger than α, β and μ. What's more, it can be clearly seen that β and μ dominate the effect from image modality and text modality respectively. It is important to adjust β and μ appropriately, so that the two retrieval tasks I2T and T2I can achieve the superior overall performance. In addition, the figure of α suggests that the cross-modal loss L_{cross} has greater impact than image loss L_{image} and text loss L_{text}, which means inter-modal semantic similarity is more important than intra-modal semantic consistency.

5. CONCLUSION

In this paper, we have proposed an Attention-Guided Semantic Hashing (AGSH) method for unsupervised cross-modal retrieval. AGSH first used the attention module to filter the feature representations, so as to more accurately grasp the important semantic information of each modality. In order to improve the performance, AGSH minimizied the error between hash codes and our proposed attention-aware semantic matrix $\boldsymbol{S}$ to reserve the inter-modal and intra-modal semantic consistency. Extensive experiments have proved the effectiveness and superiority of AGSH, and we also have conducted detailed explanations and studies on each component.

6. REFERENCES

[1] Weiwei Wang, Yuming Shen, Haofeng Zhang, Yazhou Yao, and Li Liu, "Set and rebase: determining the semantic graph connectivity for unsupervised cross-modal hashing," in *IJCAI*, 2020, pp. 853–859.

[2] Shifeng Zhang, Jianmin Li, and Bo Zhang, "Joint cluster unary loss for efficient cross-modal hashing," in *ACM ICMR*, 2019, pp. 212–216.

[3] V.E. Liong, Jiwen Lu, Yap-Peng Tan, and Jie Zhou, "Cross-modal deep variational hashing," in *ICCV*, 2017, pp. 4077–4085.

[4] Weiwei Wang, Yuming Shen, Haofeng Zhang, and Li Liu, "Semantic-rebased cross-modal hashing for scalable unsupervised text-visual retrieval," *Information Processing and Management*, vol. 57, no. 6, pp. 102374, 2020.

[5] Haomiao Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen, "Deep supervised hashing for fast image retrieval," in *CVPR*, 2016, pp. 2064–2072.

[6] Yuming Shen, Li Liu, Ling Shao, and Jingkuan Song, "Deep binaries: Encoding semantic-rich cues for efficienttextual-visual cross retrieval," in *ICCV*, 2017, pp. 4117–4126.

[7] Qingyuan Jiang and Wujun Li, "Deep cross-modal hashing," in *CVPR*, 2017, pp. 3270–3278.

[8] Ting Zhang and Jingdong Wang, "Collaborative quantization for cross-modal similarity search," in *CVPR*, 2016.

[9] Go Irie, Hiroyukai Arai, and Yukinobu Taniguchi, "Alternating co-quantization for cross-modal hashing," in *ICCV*, 2015, pp. 1886–1894.

[10] Guiguang Ding, Yuchen Guo, and Jile Zhou, "Collective matrix factorization hashing for multimodal data," in *CVPR*, 2014, pp. 2075–2082.

[11] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Hengtao Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *SIGMOD*, 2013, pp. 785–796.

[12] Yair Weiss, Antonio Torralba, and Rob Fergus, "Spectral hashing," in *NIPS*, 2008, pp. 1753–1760.

[13] Gengshen Wu, Zijia Lin, Jungong Han, Li Liu, Guiguang Ding, Baochang Zhang, and Jialie Shen, "Unsupervised deep hashing via binary latent factor models for large-scale cross-modal retrieval.," in *IJCAI*, 2018, pp. 2854–2860.

[14] Jian Zhang, Yuxin Peng, and Mingkuan Yuan, "Unsupervised generative adversarial cross-modal hashing," in *AAAI*, 2018.

[15] Shupeng Su, Zhisheng Zhong, and Chao Zhang, "Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval," in *ICCV*, 2020, pp. 3027–3035.

[16] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang, "Residual attention network for image classification," in *CVPR*, 2017.

[17] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Y.W Teh, "Set transformer: A framework for attention-based permutation-invariant neural networks," in *ICML*, 2019.

[18] Daniel Cer, Yinfei Yang, Shengyi Kong, Nan Hua, Nicole Limtiaco, R.S. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, and Chris Tar, "Universal sentence encoder," *arXiv:1803.11175*, 2018.

[19] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Hengtao Shen, "Inter-media hashing for largescale retrieval from heterogeneous data sources," in *ACM SIGMOD*, 2013, pp. 785–796.

[20] Rastegari M., Choi J., Fakhraei S., Hal D., and Davis L., "Predictable dual-view hashing," in *ICML*, 2013, pp. 1328–1336.

[21] Botong Wu, Qiang Yang, Weishi Zheng, Yizhou Wang, and Jingdong Wang, "Quantized correlation hashing for fast cross-modal search," in *IJCAI*, 2015, pp. 3946–3952.

[22] Di Hu, Feiping Nie, and Xuelong Li, "Deep binary reconstruction for cross-modal hashing," *IEEE TMM*, vol. 21, no. 4, pp. 973–985, 2018.

[23] Jian Zhang, Yuxin Peng, and Mingkuan Yuan, "Unsupervised generative adversarial crossmodal hashing," in *AAAI*, 2018, pp. 539–546.