EasyChair Preprint

№ 11137

Unsupervised Joint-Semantics Autoencoder Hashing for Multimedia Retrieval

Yunfei Chen, Jun Long, Yinan Li, Yanrui Wu and Zhan Yang

October 23, 2023

Unsupervised Joint-Semantics Autoencoder Hashing for Multimedia Retrieval

Yunfei Chen, Jun Long*, Yinan Li, Yanrui Wu, and Zhan Yang*

Big Data Institute, School of Computer Science, Central South University, Changsha Hunan, China
{yunfeichen,junlong,liyinan,8209200520,zyang22}@csu.edu.cn

Abstract. Cross-modal hashing has emerged as a prominent approach for large-scale multimedia information retrieval, offering advantages in computational speed and storage efficiency over traditional methods. However, unsupervised cross-modal hashing methods still face challenges in the lack of practical semantic labeling guidance and handling of cross-modal heterogeneity. In this paper, we propose a new unsupervised cross-modal hashing method called Unsupervised Joint-Semantics Autoencoder Hashing(UJSAH) for multimedia retrieval. First, we introduce a joint-semantics similarity matrix that effectively preserves the semantic information in multimodal data. This matrix integrates the original neighborhood structure information of the data, allowing it to better capture the associations between different modalities. This ensures that the similarity matrix can accurately mine the underlying relationships within the data. Second, we design a dual prediction network-based autoencoder, which implements the interconversion of semantic information from different modalities and ensures that the generated binary hash codes maintain the semantic information of different modalities. Experimental results on several classical datasets show a significant improvement in the performance of UJSAH in multimodal retrieval tasks relative to existing methods. The experimental code is published at https://github.com/YunfeiChenMY/UJSAH.

Keywords: Cross-modal Hashing · Multimedia Retrieval · Joint-Semantics · Dual Prediction.

1 Introduction

With the continuous advancements in science and technology, network data size and variety are rapidly expanding. Traditional information retrieval methods that use original data for computation suffer from high computational complexity. Therefore, achieving high retrieval efficiency while requiring minimal storage space has become a crucial research direction. Hash learning has gained significant attention in large-scale multimodal data retrieval due to its efficient computation speed and low storage requirements [19]. Hash learning involves a

* Corresponding authors.

hash function to map high-dimensional raw data into a low-dimensional binary hash code. The mainstream hashing methods primarily rely on classical hash functions based on features. At the same time, some recent works [21] have integrated semantic information extraction, hash function training, and hash code generation into the same framework.

Given the inherent heterogeneity among different modal data, directly calculating their similarity becomes difficult. Cross-modal hashing methods aim to map the original data from different modalities into a shared binary space, which enables the similarity calculations between different modalities using the Hamming distance. The data of real application scenarios are mostly unlabeled, severely hindering supervised hashing development. The unsupervised deep cross-modal hashing method utilizes deep networks' powerful feature extraction capability to fully extract deep semantic features from the raw data, enabling the generation of binary hash codes rich in semantic information. It [4] mainly integrates features of different modalities' raw data to construct similarity matrices and employs deep neural networks to construct hash functions for the generation of hash codes. Although unsupervised deep cross-modal hashing has gone well, there is still significant room for progress in constructing the similarity matrix of raw data and cross-modal feature heterogeneity. Traditional autoencoder hashing methods only consider the decoding and reconstruction of intra-modal semantic information and lack the mining of cross-modal semantic information.

We propose a new unsupervised multimedia hashing method called Unsupervised Joint-Semantics Autoencoder Hashing to address the aforementioned challenge. First, the UJSAH method design joint-semantics similarity matrices to comprehensively explore similarity relationships between multimodal data. Second, the UJSAH method uses an encoding module to generate hash codes for a given data, a decoding module to reconstruct the raw data to ensure that the resulting hash codes preserve the complete semantic information contained in the raw data, and a dual prediction module is designed in the autoencoder that explores the deep correlation relationships between different modal data. The core work of UJSAH is as follows:

1. The joint-semantic similarity matrix is constructed to explore multi-modal similarity relations and improve the hash function training guidance. The similarity matrix is constructed considering the similarity within each modality and the similarity between different modalities.
2. An autoencoder established on a dual prediction network is designed to generate hash codes. The method employs an autoencoder to ensure that the resulting hash codes preserve the complete semantic information contained in the raw data and uses a dual prediction network to achieve the exploration of semantic association relationships between multi-modal data.
3. Comprehensive experiments conducted on the MIRFlickr and NUS-WIDE datasets verify that the UJSAH method significantly exceeds the mainstream baseline methods.

2 Related Work

Current hashing can be broadly categorized into two categories: supervised hashing and unsupervised hashing.

2.1 Supervised Hashing

In supervised hashing, the semantic information present in the labels is utilized to guide the learning of semantic information in the hash codes. SASH [14] adaptively learns the similarity matrix and saves the association information in the labels to the data features to extract the label relevance for optimizing the previously mentioned matrix. [25] proposes incorporating probabilistic code balance constraints into deep supervised hashing, which enforces a discrete uniform distribution for each hash code. To guarantee that the binary hash codes generated by the model align with the semantic information classification think in the original data, DSDH [9] proposes a deeply supervised discrete hashing algorithm. Literature [17] presents an end-to-end model that effectively extracts key features and generates hash codes with precise semantic information. DPN [3] applies differentiable bit hinge-like losses to the network's output channels, ensuring their values deviate from zero. SHDCH [22] accepts hash codes by explicitly exploring hierarchical tags. DSH [11] introduces a novel approach to deep supervised hashing, aiming to retain compact hash codes that maintain similarity for large-scale image data.

Although supervised hashing methods have made significant progress in information retrieval, most data is unlabeled in real-world scenarios. In contrast, unsupervised hashing methods are more suitable for handling real-world application scenario data and reduce the expensive cost of the manual labeling process due to its property of not relying on labeled information.

2.2 Unsupervised Hashing

Unsupervised hashing fully uses the semantic information between the raw data to guarantee its maintenance in the binary hash code. To address the issue of ignoring neighboring instances and label granularity, DCH-SCR [12] digs deeper into the semantic similarity information within multimedia data. CAGAN [10] proposes an adaptive attention network model to retrieve massive multimodal data efficiently. In order to protect the privacy and security of data, [23] proposes a data-centric multimedia hash learning approach. To efficiently retrieve cross-modal remote sensing images, DACH [5] uses generative adversarial network hashing to extract fine-grained feature information in remote sensing images. To alleviate the limitations in similarity supervision and optimization strategies, DAEH [15] uses discriminative similarity matrix and adaptive self-updating optimization strategies to generate hash codes and train hash functions.

Existing cross-modal hash retrieval methods have significantly progressed around data feature extraction and cross-modal association mining. However, there is still much room for improvement in dealing with cross-modal heterogeneity and deep data feature association extraction.

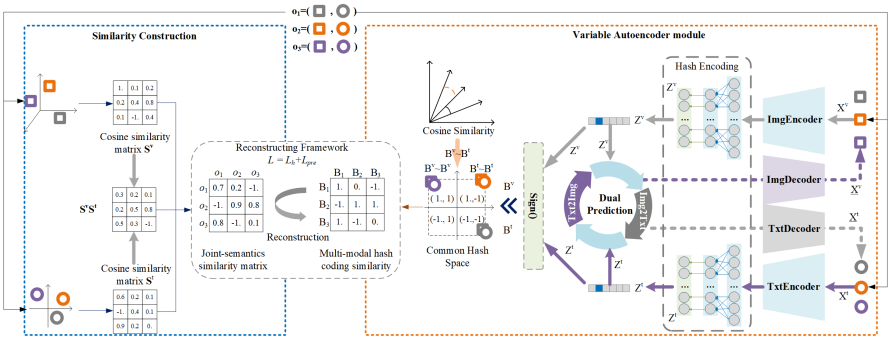


Fig. 1. The basic framework of the Unsupervised Joint-Semantics Autoencoder Hashing.

3 The Proposed Method

This section describes UJSAH method, including notations, architecture, objective function, and extensions. The framework consists of an encoding network, a dual prediction network, and a decoding network. The encoding network maintains consistency in multimedia information, the dual prediction network focuses on reconstructing the different modal data, and the decoding network combines the original data with the generated hash codes.

3.1 Notations

This paper uses bold uppercase and lowercase letters to represent matrices and vectors. Given a dataset $\mathbf{X} = \{\mathbf{X}^1, ..., \mathbf{X}^m\}|_{m=1}^{\mathrm{M}}$ of M modalities, where $\mathbf{X}^m = \{x_1, ..., x_n\} \in \mathbb{R}^{d_m \times n}$, d_m is the dimensionality of data modality $\mathbf{X}$, n represents the size of the dataset. We use two modalities, image and text, to verify the effectiveness of the proposed method, and we set $\mathbf{X} = \{\mathbf{X}^v, \mathbf{X}^t\}$, where $\mathbf{X}^v$ and $\mathbf{X}^t$ denote the image and text feature matrices. The proposed UJSAH method is to generate compact hash code $\mathbf{B} \in \{-1, 1\}^{k \times n}$, where k represents the length of the hash code.

3.2 Architecture

As shown in Fig.1, the proposed method UJSAH is an end-to-end framework with three main components, i.e., Encoding Network, Decoding Network, and Prediction Network, to process image and text data.

Similarity Construction

Efficient extraction of the underlying neighborhood structure and maintaining consistent hash code relationships with the original data are crucial in unsupervised cross-modal retrieval tasks. We employ cosine similarity to estimate the similarity between different data to achieve this. In this paper, we adopt the

idea of similarity cross-modal computation and combine the fusion computation of different modal data with improving the deep mining of the original data similarity. To fully explore the semantic relationships in the raw data, we normalize the raw data information of all modalities and integrate them, then calculate the similarity between the raw data.

We can compute the cosine similarity matrices $\mathbf{S}^v = c(\tilde{\mathbf{X}}^v, \tilde{\mathbf{X}}^v)$ and $\mathbf{S}^t = c(\tilde{\mathbf{X}}^t, \tilde{\mathbf{X}}^t)$ to represent the semantic association information for the images and texts. We first integrate $\mathbf{S}^v$ and $\mathbf{S}^t$ summed by weights as follows:

$$\tilde{\mathbf{S}} = \mu\mathbf{S}^v + (1-\mu)\mathbf{S}^t, s.t. \mu \in [0,1]. \tag{1}$$

Next, we consider $\tilde{\mathbf{S}}$ as a similarity relation between multimodal instances. The unified characterization of multimodal semantic relations can be achieved by computing $\mathbf{S}^v{\mathbf{S}^t}^\top$ to dig deeply into the semantic associations between different modal data to achieve a comprehensive representation of the associative relations between $\mathbf{S}^v$ and $\mathbf{S}^t$. We propose a joint-semantics similarity matrix $\mathbf{S} = J(\mathbf{S}^v, \mathbf{S}^t) \in [-1,+1]^{n\times n}$ to construct the semantic similarity between input instances $\mathbf{X}^v$ and $\mathbf{X}^t$. To introduce the hybrid function J, we finally define the joint-semantics similarity matrix $\mathbf{S}$ as follows:

$$\mathbf{S} = J(\mathbf{S}^v, \mathbf{S}^t) = (1-\eta)\tilde{\mathbf{S}} + \eta\frac{\mathbf{S}^v{\mathbf{S}^t}^\top}{n}, \tag{2}$$

where η is the trade-off parameter that regulates the similarity description.

Encoding Network:

Hash retrieval techniques mainly map the high-dimensional original feature into a low-dimensional information space while effectively preserving the original data's semantic information and semantic relationships. In the encoding stage, we use Image Encoder to transform image data $\mathbf{X}^v \in \mathbb{R}^{d_v\times\epsilon}$ into feature $\mathbf{Z}^v \in \mathbb{R}^{d_{ev}\times\epsilon}$ denote the $d_{ev}-$dimensional image feature vector with ϵ instances and further input hash layer to generate binary hash code $\mathbf{B}^v \in \{-1,1\}^{k\times\epsilon}$. Text Encoder extracts feature $\mathbf{Z}^t \in \mathbb{R}^{d_{et}\times\epsilon}$ represents the $d_{et}-$dimensional text feature vector with ϵ instances from the original text data $\mathbf{X}^t \in \mathbb{R}^{d_t\times\epsilon}$ and generates binary hash code $\mathbf{B}^t \in \{-1,1\}^{k\times\epsilon}$. The function of developing binary hash code is as follows:

$$\mathbf{Z}^m = f(\mathbf{X}^m;\theta_m), \mathbf{B}^m = sign(\mathbf{Z}^m), \tag{3}$$

where θ_v, θ_t denotes the parameter weights of the corresponding neural network.

Dual Prediction Network:

In the dual prediction stage, since different modalities of the same data have similar semantic information, we use the Image Prediction Network to convert the text information $\mathbf{Z}^t$ into image information $\bar{\mathbf{Z}}^v$. The Text Prediction Network extracts the information in image feature $\mathbf{Z}^v$ to generate the corresponding text feature $\bar{\mathbf{Z}}^t$. The dual prediction network can explore the semantic association of different modal information of the data. The process of dual prediction is defined

as follows:

$$\bar{\mathbf{Z}}^v = f(\mathbf{Z}^t; \theta_{pv}), \bar{\mathbf{Z}}^t = f(\mathbf{Z}^v; \theta_{pt}), \tag{4}$$

where θ_{pv}, θ_{pt} denote the parameter weights of the corresponding network.

Decoding Network:

In the decoding stage, we use the potential features $\bar{\mathbf{Z}}^v$ which are generated by the dual prediction network as the input of the decoding network to generate the original data instances $\bar{\mathbf{X}}^v$, and input $\bar{\mathbf{Z}}^t$ to generate $\bar{\mathbf{X}}^t$ to achieve the decoding of different modalities' data, ensuring that the potential features $\bar{\mathbf{Z}}^m$ generated by the dual prediction network contain the comprehensive semantic information in the original data. Moreover, construct the similarity matrix between the original data instances to constrain the validity of the data features generated by the dual prediction network.

$$\bar{\mathbf{X}}^t = G(\bar{\mathbf{Z}}^t) = f(\bar{\mathbf{Z}}^t; \theta_{dt}), \bar{\mathbf{X}}^v = G(\bar{\mathbf{Z}}^v) = f(\bar{\mathbf{Z}}^v; \theta_{dv}), \tag{5}$$

where θ_{dv} and θ_{dt} denote the parameter weights of the corresponding networks.

3.3 Objective Function

To guarantee the quality of the resulting hash codes, we fully consider that the generated hash codes $\mathbf{B}^v$ and $\mathbf{B}^t$ maintain a similar relationship intra-modal and inter-modal of the original instances to improve the performance of retrieval further. Ultimately, the hash code generation loss $\mathcal{L}_h$ is defined as follows:

$$\begin{aligned} \min_{\theta_v, \theta_t} \mathcal{L}_h = \alpha \parallel \mathbf{S} - \mathbf{B}^v {\mathbf{B}^t}^\top \parallel_F^2 + \beta \parallel \mathbf{S} - \mathbf{B}^v {\mathbf{B}^v}^\top \parallel_F^2 \\ + \beta \parallel \mathbf{S} - \mathbf{B}^t {\mathbf{B}^t}^\top \parallel_F^2 + \gamma \parallel \mathbf{B}^v - \mathbf{B}^t \parallel_F^2, \end{aligned} \tag{6}$$

where θ_v, θ_t are the parameters of encoding network, and α, β, γ are the weighting factors. In order to ensure that the predicted generated data instances $\bar{\mathbf{X}}^m$ strictly maintain the similarity relationship between the raw data, the objective function is defined as follows:

$$\min_{\theta_{dv}, \theta_{dt}} \mathcal{L}_{pre} = \delta \sum_{m \in \{v,t\}} \parallel \bar{\mathbf{X}}^m - \mathbf{X}^m \parallel_F^2, \tag{7}$$

where δ is the weighting factor. The definition of the final objective function is given founded on the above several modular loss functions as follows:

$$\begin{aligned} \min_{\theta_m} \mathcal{L} = \mathcal{L}_h + \mathcal{L}_{pre} \\ = \alpha \parallel \mathbf{S} - \mathbf{B}^v {\mathbf{B}^t}^\top \parallel_F^2 + \beta \sum_{m \in \{v,t\}} \parallel \mathbf{S} - \mathbf{B}^m {\mathbf{B}^m}^\top \parallel_F^2 \\ + \gamma \parallel \mathbf{B}^v - \mathbf{B}^t \parallel_F^2 + \delta \sum_{m \in \{v,t\}} \parallel \bar{\mathbf{X}}^m - \mathbf{X}^m \parallel_F^2, \end{aligned} \tag{8}$$

where $\theta_m \in \{\theta_v, \theta_t, \theta_{pv}, \theta_{pt}, \theta_{dv}, \theta_{dt}\}$ denote the parameter weights of network.

3.4 Extensions

More modalities: The UJSAH method can accomplish the task of multimodal scenarios, and when there are multiple modalities, a new network model can be added for each modality with appropriate modifications to the objection function Eq.9 is shown below:

$$\begin{aligned} \min_{\theta_m} \mathcal{L} &= \mathcal{L}_h + \mathcal{L}_{pre} \\ &= \alpha \sum_{m_1,m_2\in \mathrm{G}} \| \mathbf{S} - \mathbf{B}^{m_1}\mathbf{B}^{m_2\top} \|_F^2 + \beta \sum_{m\in \mathrm{G}} \| \mathbf{S} - \mathbf{B}^{m}\mathbf{B}^{m\top} \|_F^2 \\ &\quad + \gamma \sum_{m_1,m_2\in \mathrm{G}} \| \mathbf{B}^{m_1} - \mathbf{B}^{m_2} \|_F^2 + \delta \sum_{m\in \mathrm{G}} \| \bar{\mathbf{X}}^{m} - \mathbf{X}^{m} \|_F^2 . \\ s.t. \mathrm{G} &= \{1, ..., \mathrm{M}\}, \mathrm{m}_1 \neq \mathrm{m}_2, \theta_{\mathrm{m}} \in \{\theta_1, ..., \theta_{\mathrm{M}}, \theta_{\mathrm{p1}}, ..., \theta_{\mathrm{pM}}, \theta_{\mathrm{d1}}, ..., \theta_{\mathrm{dM}}\}. \end{aligned} \tag{9}$$

Out-of-Sample: After the modal is fully trained, we can employ the trained model to develop binary hash codes for any sample of a new query. In detail, give a query data $x = \mathbf{X}^m \in \mathbb{R}^{d_m \times 1}$, we can obtain the hash code as follows:

$$b = sign(f(x; \theta_m)), s.t. m \in \{1, ..., \mathrm{M}\}. \tag{10}$$

3.5 Computational Complexity Analysis

This section examines the computational complexity of the UJSAH method, as shown in Algorithm 1. During the experiments, the primary time cost lies in Eq.(8). In each iteration of the model training, we calculate the function Eq.(8), which has a time complexity of $O(n/n_b(n_b^2 d_i + n_b^2 d_t)) = O(n(n_b d_i + n_b d_t))$. Generally, the computational complexity of the algorithm for each iteration is $O((n(n_b d_i + n_b d_t))t)$, where $n_b, d_i, d_t, k, t \ll n$ and t means the number of iterations required for the model training. This time complexity can be simplified to $O(n)$, linearly correlated to the dataset size.

4 Experiments

To validate the usefulness of our UJSAH method, we have executed comprehensive experiments on MIRFlickr [7] and NUS-WIDE [1] datasets.

Algorithm 1 Unsupervised Joint-Semantics Autoencoder Hashing

Input: The training data: $\{\mathbf{X}^v, \mathbf{X}^t\}$,max training epoch E min-batch size: ϵ, hash code length: c, balance parameters: $\alpha, \beta, \gamma, \delta$.
Output: Parameters of the network: θ_v, θ_t.
Procedure:
Random initialization of the neural network parameters θ_m;
Extraction of image and text features from the dataset and construct similarity matrix;
Repeat:

1: Select ϵ image-text pairs from the dataset in turn for training;
2: Construct a similarity matrix $\mathbf{S}$ for the selected data according to Eq.2;
3: Compute $\mathbf{Z}^v = f(\mathbf{X}^v; \theta_v), \mathbf{Z}^t = f(\mathbf{X}^t; \theta_t)$ for samples by forward-propagation;
4: Generate binary hash codes
5: Compute $\mathbf{B}^v, \mathbf{B}^t, \bar{\mathbf{Z}}^v, \bar{\mathbf{Z}}^t, \bar{\mathbf{X}}^v, \bar{\mathbf{X}}^t$ according to Eq.3, and Eq.4, Eq.5;
6: Calculate the loss $\mathcal{L}$ with the Eq.8;
7: Update the network parameter $\theta_v, \theta_t, \theta_{pv}, \theta_{pt}, \theta_{dv}, \theta_{dt}$ by using backpropagation;

Until convergent.
Return: θ_v, θ_t.

4.1 Datasets

MIRFlickr [7] contains 20,015 instances of image and text pairs and its semantic information can be classified into 24 label classes. The dataset is split into 18015 training data pairs and 2000 test data pairs, and we utilize all the available data for our experiments. **NUS-WIDE** [1] contains 270k image and text instance pairs, and this experiment selects 186,577 instance pairs from 10 of these labeled categories and 1867 image and text pairs as queries.

4.2 Baselines and Evaluation Metric

In our experiments, we experimentally analyze the proposed UJSAH method with advanced unsupervised cross-modal hash retrieval approaches, including CVH [8], IMH [16], LCMH [27], CMFH [2], LSSH [26], DBRC [6], RFDH [20], DJRH [18], AGCH [24], and DUCH [13]. All of these methods are evaluated for cross-modal retrieval, which includes retrieval of textual data by visual image information (Image-to-Text) and retrieval of visual image data by textual data (Text-to-Image). The evaluation metrics employed to estimate the retrieval accuracy of our UJSAH method and the baselines are mean average precision (mAP) and top-K accuracy. In our experiments, we set K = 50 as the value for top-K accuracy.

4.3 Implementation Detail

For the proposed UJSAH method, the parameters α, β, γ, and δ are used to balance the weights of different data items. In our experiments, when we set $\{\alpha = 0.08, \beta = 18, \gamma = 200, \delta = 0.12\}$, $\{\alpha = 1, \beta = 5, \gamma = 200, \delta = 1.2\}$ for

Table 1. The mAP results for all methods on two datasets.

Method	I → T								T → I							
	MIRFlickr-25K				**NUS-WIDE**				**MIRFlickr-25K**				**NUS-WIDE**			
	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
CVH	0.606	0.599	0.596	0.598	0.372	0.362	0.406	0.390	0.591	0.583	0.576	0.576	0.401	0.384	0.442	0.432
IMH	0.612	0.601	0.592	0.579	0.470	0.473	0.476	0.459	0.603	0.595	0.589	0.580	0.478	0.483	0.472	0.462
LCMH	0.559	0.569	0.585	0.593	0.354	0.361	0.389	0.383	0.561	0.569	0.582	0.582	0.376	0.387	0.408	0.419
CMFH	0.621	0.624	0.625	0.627	0.455	0.459	0.465	0.467	0.642	0.662	0.676	0.685	0.529	0.577	0.614	0.645
LSSH	0.584	0.599	0.602	0.614	0.481	0.489	0.507	0.507	0.637	0.659	0.659	0.672	0.577	0.617	0.642	0.663
DBRC	0.617	0.619	0.620	0.621	0.424	0.459	0.447	0.447	0.618	0.626	0.626	0.628	0.455	0.459	0.468	0.473
RFDH	0.632	0.636	0.641	0.652	0.488	0.492	0.494	0.508	0.681	0.693	0.698	0.702	0.612	0.641	0.658	0.680
UDCMH	0.689	0.698	0.714	0.717	0.511	0.519	0.524	0.558	0.692	0.704	0.718	0.733	0.637	0.653	0.695	0.716
DJSRH	0.810	0.843	0.862	0.876	0.724	0.773	0.798	0.817	0.786	0.822	0.835	0.847	0.712	0.744	0.771	0.789
AGCH	0.865	0.887	0.892	0.912	0.809	0.830	0.831	0.852	0.829	0.849	0.852	0.880	**0.769**	0.780	0.798	0.802
DUCH	0.850	0.863	0.873	0.893	0.753	0.775	0.814	0.827	0.826	0.855	0.864	0.877	0.726	0.758	0.781	0.795
UJSAH	**0.884**	**0.913**	**0.927**	**0.936**	**0.812**	**0.835**	**0.858**	**0.867**	**0.853**	**0.879**	**0.881**	**0.893**	0.765	**0.790**	**0.803**	**0.813**

MIRFlickr and NUS-WIDE datasets respectively. The network architecture is designed as follows: The image encoder ($d_v \to 4096 \to relu \to k \to tanh$), the text encoder ($d_t \to 2048 \to relu \to k \to tanh$), the dual prediction network ($k \to 1024 \to relu \to k \to tanh$), the image decoder ($k \to 4096 \to relu \to 4096 \to relu \to d_v \to relu$), the text decoder ($k \to 2048 \to relu \to 2048 \to relu \to d_t \to relu$). We conducted all experiments with the same experimental setting to ensure validity and accuracy.

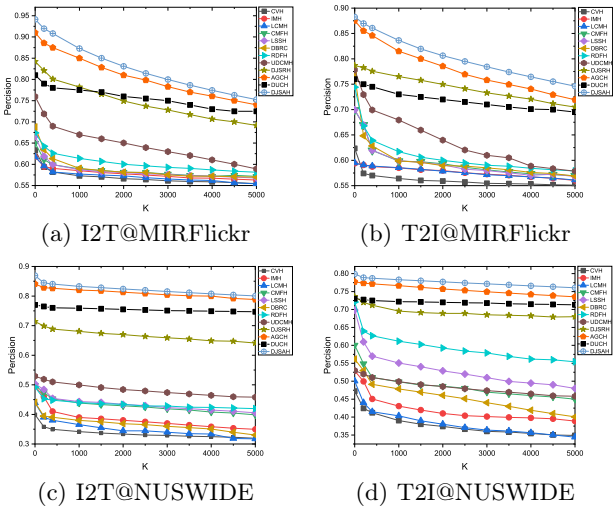
4.4 Retrieval Accuracy Comparison

In this subsection, Table 1 manifests the mAP scores of our UJSAH compared to baselines in the "Image-to-Text (I→T)" and "Text-to-Image (T→I)" retrieval studies, with hash code lengths ranging from 16 to 128 bits. By analyzing Table 1, we can obtain the following conclusions:

1) Our UJSAH method reaches a satisfactory result compared to baselines with various hash code lengths and verifies the validity of the method. In particular, on the I→T task, the mean mAP scores of the proposed UJSAH are 2.9% and 1.5% higher compared to the AGCH in the MIRFlickr and NUS-WIDE datasets, respectively. On the T→I task, the mean mAP scores of the proposed UJSAH are 2.3% and 1.1% higher compared to the second highest baseline in the MIRFlickr and NUS-WIDE datasets. We propose that UJSAH outperforms baseline methods in all datasets in the cross-modal retrieval.

2) Data analysis indicates that the performance of all baseline methods shows significant improvement as the hash code length increases. Longer hash codes have the potential to capture and represent richer semantic information. However, it is essential to note that some baseline methods may experience a degradation in retrieval performance as the hash code length increases. This can be attributed to adding redundant information and introducing potential noise in more extended hash codes.

The top-K precision curves for a hash code length of 128 bits on the two datasets are depicted in Fig.2. Experimental results show that the UJSAH method outperforms the baseline hashing method at various return numbers. The experimental analysis further demonstrates the effectiveness and excellence



(a) I2T@MIRFlickr (b) T2I@MIRFlickr

(c) I2T@NUSWIDE (d) T2I@NUSWIDE

Fig. 2. The top-K precision cures of UJSAH with 128-bit on two datasets.

of UJSAH method in the field of unsupervised large-scale multimedia information retrieval.

4.5 Parameter Sensitivity Analysis:

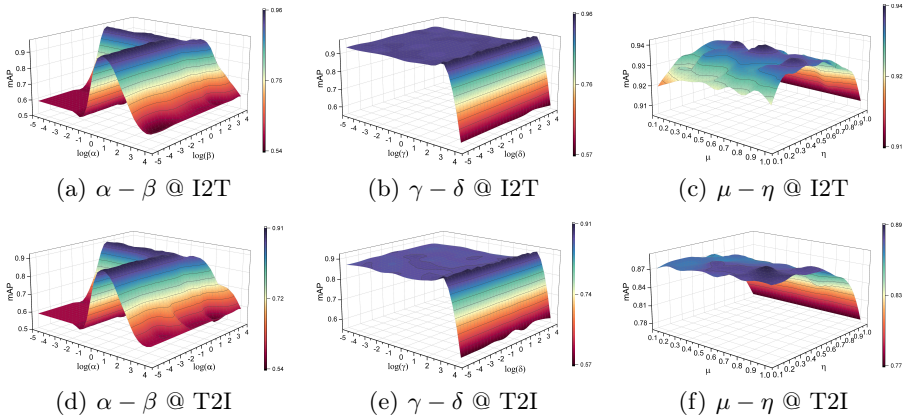
Fig.3 shows the sensitivity analysis of parameters α, β, γ, and δ set from $1e^{-5}$ to $1e^{4}$, and parameters μ and η set from 0.1 to 1.0. 1) From Fig.3 (a) and (b), it can be seen that parameter a is set from 0.01 to 100, and β can achieve good results regardless of the value of the model. 2) From the analysis of Fig.3 (c) and (d), we can have the conclusion that the parameter γ is less than $1e^{4}$, and parameter δ is set at any value, the model effect is very stable, and the retrieval results are very satisfactory. 3) From the analysis of Fig.3 (e) and (f), it can be concluded that the I2T of the model is greater than 0.9 and T2I is greater than 0.75 for any value of parameters μ, η. The comprehensive performance of the UJSAH method is stable and not sensitive to the changes of parameters μ, η.

4.6 Ablation Experiments

To assess the effectiveness of each component and confirm their usefulness, we conduct ablation experiments by designing various variants for each component.

UJSAH -1: We modify the similarity matrix $\mathbf{S}$ as $\mathbf{S} = \mu\mathbf{S}^{v} + (1-\mu)\mathbf{S}^{t}$, using the traditional similarity matrix fusion as semantic constraint information, and have verified the validity of the joint-semantics similarity matrix.

UJSAH -2: To verify the effectiveness of our designed dual prediction autoencoder, we modify the input ($\bar{\mathbf{Z}}^{m}$) of the decoding module in UJSAH framework



(a) $\alpha - \beta$ @ I2T (b) $\gamma - \delta$ @ I2T (c) $\mu - \eta$ @ I2T

(d) $\alpha - \beta$ @ T2I (e) $\gamma - \delta$ @ T2I (f) $\mu - \eta$ @ T2I

Fig. 3. The effects of the parameters with 128-bit on MIRFlickr-25K.

Table 2. The ablation experiments for different variants of UJSAH on MIRFlickr-25K.

Method	I → T				T → I			
	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
UJSAH	**0.884**	**0.913**	**0.927**	**0.936**	**0.853**	**0.879**	**0.881**	**0.893**
UJSAH-1	0.763	0.880	0.909	0.920	0.785	0.842	0.865	0.881
UJSAH-2	0.695	0.844	0.922	0.934	0.690	0.798	0.875	0.886

to input ($\mathbf{Z}^m$) in the traditional way to experiment the importance of dual prediction network.

As the analysis in Table 2 can be concluded, the joint-semantics similarity matrix and dual prediction autoencoder in the proposed UJSAH model can effectively improve the retrieval accuracy.

5 Conclusion

In this paper, we present a Unsupervised Joint-Semantics Autoencoder Hashingmethod for multimedia retrieval. The generated hash codes are guaranteed to retain more information about the similarity of the raw data by autoencoder. The design of the joint-semantics similarity matrix achieves efficient mining of multimedia data similarity matrix by using a mixture of similarity matrix within each modality and the cross-modal similarity matrix construction proposed in this paper. An autoencoder established on a dual prediction network is proposed to realize the association of semantic information of cross-modal data by converting hash codes of different modal data to each other. Finally, we execute extensive experiments on the widely used datasets to prove the significance and sophistication of the UJSAH method. In future research, we plan to study fine-grained correlations, capture more complex relationships between different

modalities in multimodal data, and extend the proposed architecture and similarity relation mining to shallow hash models.

Acknowledgements This work is supported in part by the National Natural Science Foundation of China under the Grant No.62202501 and No.U2003208, in part by the National Key R&D Program of China under Grant No.2021YFB3900 902 and in part by the Science and Technology Plan of Hunan Province under Grant No.2022JJ40638.

References

1. Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009.
2. Guiguang Ding, Yuchen Guo, and Jile Zhou. Collective matrix factorization hashing for multimodal data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2075–2082, 2014.
3. Lixin Fan, KamWoh Ng, Ce Ju, Tianyu Zhang, and Chee Seng Chan. Deep polarized network for supervised learning of accurate binary hashing codes. In *IJCAI*, pages 825–831, 2020.
4. Wentao Fan, Chao Zhang, Huaxiong Li, Xiuyi Jia, and Guoyin Wang. Three-stage semisupervised cross-modal hashing with pairwise relations exploitation. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
5. Jiaen Guo and Xin Guan. Deep adversarial cascaded hashing for cross-modal vessel image retrieval. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:2205–2220, 2023.
6. Di Hu, Feiping Nie, and Xuelong Li. Deep binary reconstruction for cross-modal hashing. *IEEE Transactions on Multimedia*, 21(4):973–985, 2018.
7. Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 39–43, 2008.
8. Shaishav Kumar and Raghavendra Udupa. Learning hash functions for cross-view similarity search. In *Twenty-second international joint conference on artificial intelligence*, 2011.
9. Qi Li, Zhenan Sun, Ran He, and Tieniu Tan. A general framework for deep supervised discrete hashing. *International Journal of Computer Vision*, 128(8):2204–2222, 2020.
10. Yewen Li, Mingyuan Ge, Mingyong Li, Tiansong Li, and Sen Xiang. Clip-based adaptive graph attention network for large-scale unsupervised multi-modal hashing retrieval. *Sensors*, 23(7):3439, 2023.
11. Haomiao Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Deep supervised hashing for fast image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2064–2072, 2016.
12. Xiaoqing Liu, Huanqiang Zeng, Yifan Shi, Jianqing Zhu, Chih-Hsien Hsia, and Kai-Kuang Ma. Deep cross-modal hashing based on semantic consistent ranking. *IEEE Transactions on Multimedia*, 2023.

13. Georgii Mikriukov, Mahdyar Ravanbakhsh, and Begüm Demir. Deep unsupervised contrastive hashing for large-scale cross-modal text-image retrieval in remote sensing. *arXiv preprint arXiv:2201.08125 , year=2022*.
14. Yang Shi, Xiushan Nie, Xingbo Liu, Li Zou, and Yilong Yin. Supervised adaptive similarity matrix hashing. *IEEE Transactions on Image Processing*, 31:2755–2766, 2022.
15. Yufeng Shi, Yue Zhao, Xin Liu, Feng Zheng, Weihua Ou, Xinge You, and Qinmu Peng. Deep adaptively-enhanced hashing with discriminative similarity guidance for unsupervised cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):7255–7268, 2022.
16. Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Heng Tao Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *Proceedings of the 2013 ACM SIGMOD international conference on management of data*, pages 785–796, 2013.
17. Hai Su, Meiyin Han, Junle Liang, Jun Liang, and Songsen Yu. Deep supervised hashing with hard example pairs optimization for image retrieval. *The Visual Computer*, pages 1–16, 2022.
18. Shupeng Su, Zhisheng Zhong, and Chao Zhang. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3027–3035, 2019.
19. Rong-Cheng Tu, Jie Jiang, Qinghong Lin, Chengfei Cai, Shangxuan Tian, Hongfa Wang, and Wei Liu. Unsupervised cross-modal hashing with modality-interaction. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
20. Di Wang, Quan Wang, and Xinbo Gao. Robust and flexible discrete hashing for cross-modal similarity search. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2703–2715, 2017.
21. XianHua Zeng, Ke Xu, and YiCai Xie. Pseudo-label driven deep hashing for unsupervised cross-modal retrieval. *International Journal of Machine Learning and Cybernetics*, pages 1–20, 2023.
22. Yu-Wei Zhan, Xin Luo, Yongxin Wang, and Xin-Shun Xu. Supervised hierarchical deep hashing for cross-modal retrieval. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3386–3394, 2020.
23. Peng-Fei Zhang, Guangdong Bai, Hongzhi Yin, and Zi Huang. Proactive privacy-preserving learning for cross-modal retrieval. *ACM Transactions on Information Systems*, 41(2):1–23, 2023.
24. Peng-Fei Zhang, Yang Li, Zi Huang, and Xin-Shun Xu. Aggregation-based graph convolutional hashing for unsupervised cross-modal retrieval. *IEEE Transactions on Multimedia*, 24:466–479, 2021.
25. Qi Zhang, Liang Hu, Longbing Cao, Chongyang Shi, Shoujin Wang, and Dora D Liu. A probabilistic code balance constraint with compactness and informativeness enhancement for deep supervised hashing. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 2022.
26. Jile Zhou, Guiguang Ding, and Yuchen Guo. Latent semantic sparse hashing for cross-modal similarity search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 415–424, 2014.
27. Xiaofeng Zhu, Zi Huang, Heng Tao Shen, and Xin Zhao. Linear cross-modal hashing for efficient multimedia search. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 143–152, 2013.