scientific reports

Check for updates

OPEN

Semantic embedding based online cross-modal hashing method

Meijia Zhang[1,3], Junzheng Li[2] & Xiyuan Zheng[1]✉

Hashing has been extensively utilized in cross-modal retrieval due to its high efficiency in handling large-scale, high-dimensional data. However, most existing cross-modal hashing methods operate as offline learning models, which learn hash codes in a batch-based manner and prove to be inefficient for streaming data. Recently, several online cross-modal hashing methods have been proposed to address the streaming data scenario. Nevertheless, these methods fail to fully leverage the semantic information and accurately optimize hashing in a discrete fashion. As a result, both the accuracy and efficiency of online cross-modal hashing methods are not ideal. To address these issues, this paper introduces the Semantic Embedding-based Online Cross-modal Hashing (SEOCH) method, which integrates semantic information exploitation and online learning into a unified framework. To exploit the semantic information, we map the semantic labels to a latent semantic space and construct a semantic similarity matrix to preserve the similarity between new data and existing data in the Hamming space. Moreover, we employ a discrete optimization strategy to enhance the efficiency of cross-modal retrieval for online hashing. Through extensive experiments on two publicly available multi-label datasets, we demonstrate the superiority of the SEOCH method.

Recently, with the exponential growth of Internet usage, there has been a surge in information data. Traditional single retrieval methods are no longer sufficient to meet the increasing retrieval needs of individuals. Cross-modal retrieval, as a more effective and in-demand search method, has garnered significant research attention in today's society. Commonly used cross-modal retrieval methods[1–4] employ real-valued vectors to represent multimodal data. However, these methods require extensive computation and suffer from low efficiency.

To enhance retrieval efficiency, hash-based cross-modal retrieval methods[5–24] have been proposed. For instance, Asymmetric Supervised Consistent and Specific Hashing (ASCSH)[5], Fast Discriminative Discrete Hashing (FDDH)[6], and A Nonlinear Supervised Discrete Hashing (NSDH)[7], among others. Cross-modal hashing methods can be categorized as unsupervised[11–15] or supervised[16–19]. In practical applications, supervised hashing methods have shown better performance than unsupervised ones. Despite the progress made in supervised cross-modal hashing research, several challenges remain, such as inadequate exploitation of semantic information, substantial quantization loss, and low retrieval efficiency.

The aforementioned methods all employ an offline learning model for batch-based training, which may fail to adapt to changing data and consequently reduce retrieval efficiency when faced with large volumes of streaming data. To address these limitations, several online hashing methods[25,26] have been proposed. Similar to offline hashing methods, online hashing methods can also be categorized as unsupervised or supervised. Unsupervised online hashing methods analyze the relationship between sample data, such as dimensionality reduction and the utilization of self-organizing mapping networks. Conversely, supervised online hashing methods often leverage label information to improve retrieval accuracy and mitigate the semantic gap problem.

Although numerous online cross-modal hashing methods have been proposed, existing approaches fail to fully exploit semantic information and accurately optimize hashing in a discrete manner.

To overcome these issues, we propose the Semantic Embedding-based Online Cross-modal Hashing (SEOCH) method, which integrates semantic information exploitation and online learning into a unified framework. To exploit semantic information, we map semantic labels to a latent semantic space and construct a semantic similarity matrix to preserve the similarity between new and existing data in the Hamming space. Moreover, we employ a discrete optimization strategy for online hashing. The main contributions of SEOCH are summarized as follows:

- To exploit semantic information, we map semantic labels to a latent semantic space. Instead of directly projecting semantic labels into binary hash codes B, we employ real-valued codes QB to leverage supervised information more effectively.

[1]School of Data Science and Computer Science, Shandong Women's University, Jinan 250300, China. [2]Network Information Management Center, Shandong Management University, Jinan 250357, China. [3]Shandong Provincial Key Laboratory of Network Based Intelligent Computing, Jinan 250022, China. ✉email: 306732399@qq.com

- Subsequently, we construct a semantic similarity matrix to preserve the similarity between new and existing data in the Hamming space, thus mitigating the information loss that occurs when learning hash codes solely based on new data.
- Additionally, we adopt a discrete optimization strategy for online hashing, which reduces quantization errors caused by relaxation-based optimization methods.

The remainder of this paper is organized as follows. We provides an overview of related work in cross-modal hashing methods in the first place. Then, our proposed method and training process are presented. Next, experimental results and corresponding analysis are presented. Finally, we summarize our work.

Related work

Numerous cross-modal hashing methods have emerged recently. Based on the utilization of semantic label information during the training process, these methods can be categorized into unsupervised and supervised approaches.

Unsupervised methods learn a shared Hamming space without incorporating semantic label information, such as the Inter-Media Hashing (IMH)[27] method, Collective Matrix Factorization Hashing (CMFH)[13] method, Fusion Similarity Hashing (FSH)[14] method, and Latent Semantic Sparse Hashing (LSSH)[12] method.

On the other hand, supervised hashing methods leverage semantic label information when learning hash codes. Examples include Semantic Correlation Maximization (SCM)[28], Semantics-Preserving Hashing (SePH)[29], Discriminant Cross-modal Hashing (DCH)[30], Subspace Relation Learning for Cross-modal Hashing (SRLCH)[31], and Semantic Topic Multimodal Hashing (STMH)[32] method. To take full advantage of heterogeneous correlation, many deep cross-modal retrieval methods have been proposed in recent years, such as references[33–35]. For instance, deep discrete cross-modal hashing with multiple supervision method[34] designs a semantic network to fully exploit the semantic information implicated in labels, which no longer focuses only on instance-pairwise and class-wise similarities, but also on instance-label level.

The aforementioned methods are all offline cross-modal retrieval models. However, in practical cross-modal retrieval applications, the input is typically in a streaming fashion. Consequently, several online methods have been proposed to cater to this scenario. In the online setting, as new data continuously arrives in a streaming manner, online methods solely utilize the newly arrived data to update the current model. This significantly reduces the computational complexity of the learning algorithm and the storage space requirements. Notable examples include Online Latent Semantic Hashing (OLSH)[25] and Online Collective Matrix Factorization Hashing (OCMFH)[26], which have garnered increasing attention. Nevertheless, these methods fail to fully exploit semantic information and accurately optimize hashing in a discrete manner.

The proposed method

Notation

In this paper, we consider a scenario where the number of image and text sample is equal. Let $\mathbf{X} = \{x_i\}_{i=1}^{n} \in \mathbb{R}^{d_x * n}$ represents the image samples and $\mathbf{Y} = \{y_i\}_{i=1}^{n} \in \mathbb{R}^{d_y * n}$ represents the text samples, where d_x and d_y denote the dimensions of the image and text modalities, respectively, and n is the number of samples. $\boldsymbol{L} = \{0,\ 1\} \in \mathbb{R}^{c * n}$ is the label matrix, where c is the number of classes. If $\{x_i,\ y_i\}$ belongs to the j-th class, $l_{ji} = 1$, otherwise $l_{ji} = 0$. $\boldsymbol{B} = \{0,\ 1\} \in \mathbb{R}^{k * n}$ is the hash code matrix, where k represents the number of bits in the hash codes.

Suppose the training data is received in a streaming manner. At the t-th round, a new data chunk $\overrightarrow{X}^{(t)} \in \mathbb{R}^{d_x * n_t}$ or $\overrightarrow{Y}^{(t)} \in \mathbb{R}^{d_y * n_t}$ with labels $\overrightarrow{L}^{(t)} \in \{0,1\}^{c * n_t}$ arrive, where n_t denotes the number of new data at t-th round. Correspondingly, $\tilde{X}^{(t)} \in \mathbb{R}^{d_x * N_{t-1}}$ or $\tilde{Y}^{(t)} \in \mathbb{R}^{d_y * N_{t-1}}$ with labels $\tilde{L}^{(t)} \in \{0,1\}^{c * N_{t-1}}$ is the existing data, where $N_{t-1} = \sum_{i=1}^{t-1} n_i$ is the number of the existing data before round t. The heterogeneous samples x_i and y_j are associated with similarity matrix S with its element s_{ij}, where $s_{ij} = 1$ means x_i and y_j share at least one common class label, and $s_{ij} = 0$ means x_i and y_j do not share common class label.

Hash-code learning

To facilitate the online cross modal hashing, the overall objective function (i.e. Loss function) can be written as:

$$\begin{aligned}\min_{\substack{Q^{(t)},\overrightarrow{B}^{(t)}\\ P^{(t)},U^{(t)},V^{(t)}}} \quad & \lambda_1\left\|Q^{(t)}\tilde{B}^{(t)} - P^{(t)}\tilde{L}^{(t)}\right\|_F^2 + \lambda_2\left\|\tilde{B}^{(t)} - U^{(t)}\tilde{X}^{(t)}\right\|_F^2 + \lambda_3\left\|\tilde{B}^{(t)} - V^{(t)}\tilde{Y}^{(t)}\right\|_F^2 + \lambda_1\left\|Q^{(t)}\vec{B}^{(t)} - P^{(t)}\vec{L}^{(t)}\right\|_F^2 \\ & + \lambda_2\left\|\vec{B}^{(t)} - U^{(t)}\vec{X}^{(t)}\right\|_F^2 + \lambda_3\left\|\vec{B}^{(t)} - V^{(t)}\vec{Y}^{(t)}\right\|_F^2 + \alpha\left(\left\|Q^{(t)}\right\|_F^2 + \left\|P^{(t)}\right\|_F^2 + \left\|U^{(t)}\right\|_F^2 + \left\|V^{(t)}\right\|_F^2\right) \\ & + \beta\left\|\vec{B}^{(t)T}\tilde{B}^{(t)} - k\overleftrightarrow{S}^{(t)}\right\|_F^2 \end{aligned} \tag{1}$$

where $\overleftrightarrow{S}^{(t)}$ is the similarity matrix at round t, $\tilde{B}^{(t)} \in \{0,1\}^{k * N_{t-1}}$ denotes the hash codes of existing data, $\overrightarrow{B}^{(t)} \in \{0,1\}^{k * n_t}$ denotes the hash codes of new data. $Q \in \mathbb{R}^{g * k}$, $P \in \mathbb{R}^{g * c}$, $U \in \mathbb{R}^{k * d_x}$ and $V \in \mathbb{R}^{k * d_y}$ are four mapping matrices, g is the dimension of latent semantic concept space. $\lambda_1, \lambda_2, \lambda_3, \alpha, \beta$ are five hyperparameters. The item $\beta\left\|\vec{B}^{(t)T}\tilde{B}^{(t)} - k\overleftrightarrow{S}^{(t)}\right\|_F^2$ preserves the similarity between the new data and the existing data in the hamming space, which can solve the problem of information loss caused by learning hash codes only with new data.

Training

The Semantic Embedding based Online Cross-modal Hashing (SEOCH) algorithm aims to optimize five variables. To address the objective in Eq. (1), an alternating learning strategy is employed, updating one variable at a time while keeping the others fixed. The entire training process is outlined below.

Update $\vec{B}^{(t)}$

By fixing all variables except $\vec{B}^{(t)}$, we can reformulate Eq. (1) as follows:

$$\min_{\vec{B}^{(t)}} \lambda_1 \left\|Q^{(t)}\vec{B}^{(t)} - P^{(t)}\vec{L}^{(t)}\right\|_F^2 + \lambda_2\left\|\vec{B}^{(t)} - U^{(t)}\vec{X}^{(t)}\right\|_F^2 + \lambda_3\left\|\vec{B}^{(t)} - V^{(t)}\vec{Y}^{(t)}\right\|_F^2 + \beta\left\|\vec{B}^{(t)T}\tilde{B}^{(t)} - k\overleftrightarrow{S}^{(t)}\right\|_F^2 \tag{2}$$

Differentiating Eq. (2) with respect to $\vec{B}^{(t)}$ and setting it to zero, we obtain:

$$\vec{B}^{(t)} = (\lambda_1 Q^{(t)T}Q^{(t)} + \lambda_2 I_1 + \lambda_3 I_1 + \beta\tilde{B}^{(t)}\tilde{B}^{(t)T})^{-1} * (\lambda_3 V^{(t)}\vec{Y}^{(t)} + \lambda_2 U^{(t)}\vec{X}^{(t)} + \lambda_1 Q^{(t)T}P^{(t)}\vec{L}^{(t)} + \beta\tilde{B}^{(t)}k\overleftrightarrow{S}^{(t)T}) \tag{3}$$

where $I_1 \in \mathbb{R}^{k*k}$ denotes an identity matrix. To compute $\tilde{B}^{(t)}$ in Eq. (3), we follow the steps below.

By fixing all variables except $\tilde{B}^{(t)}$, we can reformulate Eq. (1) as follows:

$$\min_{\tilde{B}^{(t)}} \lambda_1 \left\|Q^{(t)}\tilde{B}^{(t)} - P^{(t)}\tilde{L}^{(t)}\right\|_F^2 + \lambda_2\left\|\tilde{B}^{(t)} - U^{(t)}\tilde{X}^{(t)}\right\|_F^2 + \lambda_3\left\|\tilde{B}^{(t)} - V^{(t)}\tilde{Y}^{(t)}\right\|_F^2 + \beta\left\|\vec{B}^{(t)T}\tilde{B}^{(t)} - k\overleftrightarrow{S}^{(t)}\right\|_F^2 \tag{4}$$

Differentiating Eq. (4) with respect to $\tilde{B}^{(t)}$ and setting it to zero, we obtain:

$$\tilde{B}^{(t)} = \mathrm{sgn}(\beta\vec{B}^{(t)}\overleftrightarrow{S}^{(t)}) + (\lambda_1 Q^{(t)T}Q^{(t)} + \lambda_2 I_1 + \lambda_3 I_1)^{-1} * (\lambda_1 Q^{(t)T}P^{(t)}\tilde{L}^{(t)} + \lambda_2 U^{(t)}\tilde{X}^{(t)} + \lambda_3 V^{(t)}\tilde{Y}^{(t)}) \tag{5}$$

where $I_1 \in \mathbb{R}^{k*k}$ denotes an identity matrix.

Update $Q^{(t)}$

By differentiating Eq. (1) with respect to $Q^{(t)}$,

$$\frac{\partial Loss}{\partial Q^{(t)}} = 2\lambda_1(Q^{(t)}\tilde{B}^{(t)} - P^{(t)}\tilde{L}^{(t)})\tilde{B}^{(t)T} + 2\lambda_1(Q^{(t)}\vec{B}^{(t)} - P^{(t)}\vec{L}^{(t)})\vec{B}^{(t)T} + 2\alpha Q^{(t)} \tag{6}$$

By setting Eq. (6) to zero, we have

$$Q^{(t)} = C_2^{(t)}\left(C_1^{(t)} + \frac{\alpha}{\lambda_1}I_1\right)^{-1} \tag{7}$$

where

$$\begin{aligned} C_1^{(t-1)} &= \tilde{B}^{(t)}\tilde{B}^{(t)T} + \vec{B}^{(t)}\vec{B}^{(t)T} \\ C_2^{(t-1)} &= P^{(t)}\tilde{L}^{(t)}\tilde{B}^{(t)T} + P^{(t)}\vec{L}^{(t)}\vec{B}^{(t)T} \end{aligned} \tag{8}$$

$$\begin{aligned} C_1^{(t)} &= C_1^{(t-1)} + \vec{B}^{(t)}\vec{B}^{(t)T} \\ C_2^{(t)} &= C_2^{(t-1)} + P^{(t)}\vec{L}^{(t)}\vec{B}^{(t)T} \end{aligned} \tag{9}$$

Update $P^{(t)}$

By differentiating Eq. (1) with respect to $P^{(t)}$,

$$\frac{\partial Loss}{\partial P^{(t)}} = 2\lambda_1(Q^{(t)}\tilde{B}^{(t)} - P^{(t)}\tilde{L}^{(t)})(-\tilde{L}^{(t)T}) + 2\lambda_1(Q^{(t)}\vec{B}^{(t)} - P^{(t)}\vec{L}^{(t)})(-\vec{L}^{(t)T}) + 2\alpha P^{(t)} \tag{10}$$

By setting Eq. (10) to zero, we have

$$P^{(t)} = D_2^{(t)}(D_1^{(t)} + \frac{\alpha}{\lambda_1}I_2)^{-1} \tag{11}$$

where $I_2 \in \mathbb{R}^{c*c}$ is an identity matrix,

$$\begin{aligned} D_1^{(t)} &= \tilde{L}^{(t)}\tilde{L}^{(t)T} + \vec{L}^{(t)}\vec{L}^{(t)T} \\ D_2^{(t)} &= Q^{(t)}\tilde{B}^{(t)}\tilde{L}^{(t)T} + Q^{(t)}\vec{B}^{(t)}\vec{L}^{(t)T} \end{aligned} \tag{12}$$

$$\begin{aligned} D_1^{(t)} &= D_1^{(t-1)} + \vec{L}^{(t)}\vec{L}^{(t)T} \\ D_2^{(t)} &= D_2^{(t-1)} + Q^{(t)}\vec{B}^{(t)}\vec{L}^{(t)T} \end{aligned} \tag{13}$$

Update $U^{(t)}$

By differentiating Eq. (1) with respect to $U^{(t)}$,

$$\frac{\partial Loss}{\partial U^{(t)}} = 2\lambda_2(\tilde{B}^{(t)} - U^{(t)}\tilde{X}^{(t)})(-\tilde{X}^{(t)T}) + 2\lambda_2(\vec{B}^{(t)} - U^{(t)}\vec{X}^{(t)})(-\vec{X}^{(t)T}) + 2\alpha U^{(t)} \tag{14}$$

By setting Eq. (14) to zero, we have

$$U^{(t)} = E_2^{(t)} \cdot (E_1^{(t)} + \frac{\alpha}{\lambda_2} I_3)^{-1} \tag{15}$$

where $I_3 \in \mathbb{R}^{d_x * d_x}$ is an identity matrix,

$$\begin{aligned} E_1^{(t-1)} &= \tilde{X}^{(t)}\tilde{X}^{(t)T} + \vec{X}^{(t)}\vec{X}^{(t)T} \\ E_2^{(t-1)} &= \tilde{B}^{(t)}\tilde{X}^{(t)T} + \vec{B}^{(t)}\vec{X}^{(t)T} \end{aligned} \tag{16}$$

$$\begin{aligned} E_1^{(t)} &= E_1^{(t-1)} + \vec{X}^{(t)}\vec{X}^{(t)T} \\ E_2^{(t)} &= E_2^{(t-1)} + \vec{B}^{(t)}\vec{X}^{(t)T} \end{aligned} \tag{17}$$

Update $V^{(t)}$

By differentiating Eq. (1) with respect to $V^{(t)}$,

$$\frac{\partial Loss}{\partial V^{(t)}} = 2\lambda_3(\tilde{B}^{(t)} - V^{(t)}\tilde{Y}^{(t)})\ (-\tilde{Y}^{(t)T})\ + 2\lambda_3(\vec{B}^{(t)} - V^{(t)}\vec{Y}^{(t)})\ (-\vec{Y}^{(t)T}) + 2\alpha V^{(t)} \tag{18}$$

By setting Eq. (18) to zero, we have

$$V^{(t)} = F_2^{(t)} \cdot (F_1^{(t)} + \frac{\alpha}{\lambda_3} I_4)^{-1} \tag{19}$$

$$\begin{aligned} F_1^{(t-1)} &= \tilde{Y}^{(t)}\tilde{Y}^{(t)T} + \vec{Y}^{(t)}\vec{Y}^{(t)T} \\ F_2^{(t-1)} &= \tilde{B}^{(t)}\tilde{Y}^{(t)T} + \vec{B}^{(t)}\vec{Y}^{(t)T} \end{aligned} \tag{20}$$

$$\begin{aligned} F_1^{(t)} &= F_1^{(t-1)} + \vec{Y}^{(t)}\vec{Y}^{(t)T} \\ F_2^{(t)} &= F_2^{(t-1)} + \vec{B}^{(t)}\vec{Y}^{(t)T} \end{aligned} \tag{21}$$

where $I_4 \in \mathbb{R}^{d_y * d_y}$ is an identity matrix.

Out of sample

For a query that is not in the training set, we can generate the hash codes of a query point x_q or y_q as follows,

$$\mathrm{b}_q = \mathrm{sign}(x_q U^{(t)}), \quad \mathrm{b}_q = \mathrm{sign}(y_q V^{(t)}) \tag{22}$$

To obtain a comprehensive overview, the complete learning algorithm of our proposed SEOCH is presented in Algorithm 1.

Input: New data chunk $\overrightarrow{X}^{(t)} \in \mathbb{R}^{d_x * n_t}$ and $\overrightarrow{Y}^{(t)} \in \mathbb{R}^{d_y * n_t}$ with labels $\overrightarrow{L}^{(t)} \in \{0,1\}^{c * n_t}$, auxiliary variables $\vec{B}^{(t)}$, $Q^{(t)}$, $P^{(t)}$, $U^{(t)}$, $V^{(t)}$ and parameters λ_1, λ_2, λ_3, α, β.
Output: Hash code matrix $\vec{B}^{(t)}$ and hash functions.
1: **for** iter = 1, 2, $\cdots$, i do
2: Randomly initialize $\vec{B}^{(t)}$, $Q^{(t)}$, $P^{(t)}$, $U^{(t)}$, $V^{(t)}$;
3: Optimize $\vec{B}^{(t)}$ according to Equation (3);
4: Optimize $Q^{(t)}$ according to Equation (7);
5: Optimize $P^{(t)}$ according to Equation (11);
6: Optimize $U^{(t)}$ according to Equation (15);
7: Optimize $V^{(t)}$ according to Equation (19);
8: **end for**
Return Hash functions.

Algorithm 1. The comprehensive learning algorithm of our proposed SEOCH

Experiments

Datasets

In order to thoroughly assess the effectiveness of our approach, we conduct experiments on two publicly available multi-label datasets, namely the **MIRFLICKR-25K** dataset and the **NUS-WIDE** dataset. Detailed descriptions of these datasets are provided below.

The **MIRFLICKR-25K** dataset comprises 25,000 images with a total of 24 labels. Each image in this dataset is associated with one or more labels and connected to several textual tags. From this dataset, we randomly select 20,015 image-text pairs that possess at least 20 textual tags. Among these pairs, 2000 are chosen as queries and the remaining pairs form the training set. The image and text features used are 512-dimensional Scale-Invariant Feature Transform (SIFT) features and 1386-dimensional Bag of Words (BoW) features, respectively. To facilitate online cross-modal hashing, the training set is divided into 9 data chunks, with the first 8 chunks containing 2,000 instances each and the last chunk containing 2015 instances.

The **NUS-WIDE** dataset consists of approximately 270,000 images annotated with a total of 81 labels. For our experiments, we select 186,577 image-text pairs that are associated with at least one of the 10 most frequent concepts. Within the NUS-WIDE dataset, we randomly choose 1,867 pairs as queries, while the remaining pairs serve as the database. The image and text features in the database are represented by 500-dimensional Bag-of-Visual Words (BoVW) features and 1000-dimensional BoW features, respectively. Similar to the previous dataset, the training set is divided into 18 data chunks, with the first 17 chunks containing 10,000 instances each and the last chunk containing 14,710 instances to facilitate online cross-modal hashing.

Baselines and evaluated metrics

The proposed method is evaluated against six state-of-the-art cross-modal hashing methods, which can be categorized as follows: (1) offline methods: SCM-seq[28], DCH[30], SRLCH[31], JIMFH[36]; (2) online methods: OLSH[25], OCMFH[26]. The source codes of these baselines are publicly available online, and the parameters are set based on the recommendations provided in the corresponding papers. In JIMFH, the mAP value is calculated with the number of query data set to 100. To ensure a fair comparison, we set the number of query data to 2,000 and 1,867 for the MIRFLICKR-25K and NUS-WIDE datasets, respectively.

Consistent with previous studies, we employ mean Average Precision (mAP) and Precision-Recall curves to evaluate the retrieval accuracy for two retrieval tasks: Image Retrieval Text (I2T) and Text Retrieval Image (T2I).

In the experiments, the parameters are set empirically. For the MIRFLICKR-25K dataset, we set λ_1 =1e3, λ_2 =0.1, λ_3=1, α=0.8, β=1, and g=50. For the NUS-WIDE dataset, we set λ_1 =1e4, λ_2 =0.1, λ_3 =0.1, α=0.1, β =0.1, and g=100.

Experimental results and analysis

The mean Average Precision (mAP) scores of SEOCH and the comparison methods in the final round on the MIRFLICKR-25K and NUS-WIDE datasets are presented in Tables 1 and 2, respectively. Moreover, Figs. 1 and 2 display the mAP scores for each round of different methods in the two datasets, using 8-bit and 32-bit hash codes.

From the above results, it can be observed that: (1) The proposed SEOCH significantly outperforms all the offline baselines in almost all tasks, demonstrating its efficiency for streaming data scenarios. (2) The SEOCH outperforms the online baselines in most of the retrieval tasks, indicating the superiority of the semantic embedding-based learning method. (3) The discrete methods, namely JIMFH, significantly outperform the relaxation-based methods, i.e., SCM-seq, validating that the discrete hashing methods are more effective for semantic similarity preservation. (4) With the increase of the code length, the performance of all methods is improved, which is

Task	Methods	code length				
		8 bits	16 bits	32bits	64 bits	128 bits
Image to Text	SCM-Seq	0.601	0.616	0.626	0.632	**0.634**
	DCH	0.586	0.59	0.589	0.609	0.623
	SRLCH	0.597	0.604	0.629	0.614	0.627
	JIMFH	0.632	0.636	0.612	0.637	0.626
	OLSH	0.578	0.577	0.60	0.595	0.593
	OCMFH	0.558	0.556	0.555	0.555	0.554
	SEOCH	**0.68**	**0.645**	**0.663**	**0.64**	0.62
Text to Image	SCM-Seq	0.605	0.611	0.617	0.62	0.622
	DCH	0.587	0.592	0.591	0.604	0.618
	SRLCH	0.596	0.605	0.635	0.619	0.633
	JIMFH	0.655	0.659	0.679	0.679	**0.69**
	OLSH	0.579	0.582	0.601	0.602	0.602
	OCMFH	0.558	0.556	0.555	0.555	0.554
	SEOCH	**0.693**	**0.695**	**0.696**	**0.68**	0.66

Table 1. The mAP scores of SEOCH and the comparison methods in the final round on the **MIRFLICKR-25K** dataset (Numbers in boldface indicate the highest scores).

Task	Methods	code length				
		8 bits	16 bits	32bits	64 bits	128 bits
Image to Text	SCM-Seq	0.476	0.479	0.487	0.486	0.491
	DCH	0.432	0.434	0.453	0.493	0.504
	SRLCH	0.367	0.377	0.347	0.352	0.365
	JIMFH	0.512	0.512	0.515	0.518	0.52
	OLSH	0.487	0.516	0.524	0.535	0.526
	OCMFH	0.425	0.452	0.484	0.465	0.45
	SEOCH	**0.525**	**0.526**	**0.535**	**0.536**	**0.532**
Text to Image	SCM-Seq	0.456	0.459	0.467	0.468	0.471
	DCH	0.421	0.424	0.443	0.479	0.487
	SRLCH	0.355	0.356	0.362	0.365	0.364
	JIMFH	0.622	0.624	0.628	0.623	0.62
	OLSH	0.516	0.559	0.564	0.592	0.58
	OCMFH	0.405	0.438	0.472	0.455	0.441
	SEOCH	**0.635**	**0.636**	**0.641**	**0.642**	**0.638**

Table 2. The mAP scores of SEOCH and the comparison methods in the final round on the **NUS-WIDE** dataset (Numbers in boldface indicate the highest scores).

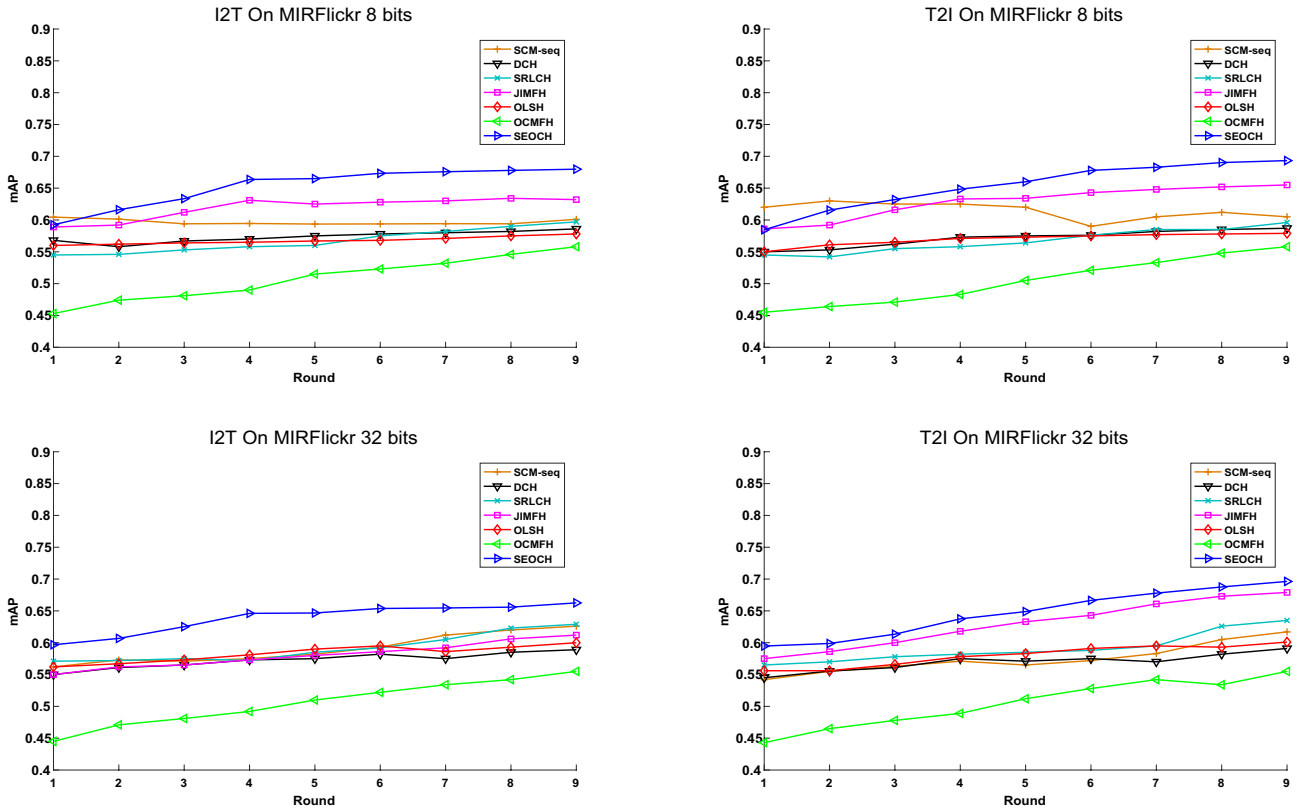

Figure 1. The mAP scores at each round for two retrieval tasks on the **MIRFLICKR-25K** dataset.

consistent with the general observation in hashing research. (5) Compared with the 8-bit and 16-bit hash codes, the performance improvement of SEOCH is more significant when using longer codes (e.g., 32 bits or above), indicating its ability to exploit the semantic structure of the data in high-dimensional Hamming space.

Further analysis

Ablation study

Moreover, three variations of SEOCH have been devised to assess the performance of the proposed method, as presented in Table 3. SEOCH-I sets λ_1 to 0. SEOCH-II sets λ_2 and λ_3 to 0. SEOCH -III eliminates the similarity matrix. From Table 3, it can be observed that for 8 bits, SEOCH-III achieves the lowest result; for 16, 32, and 64

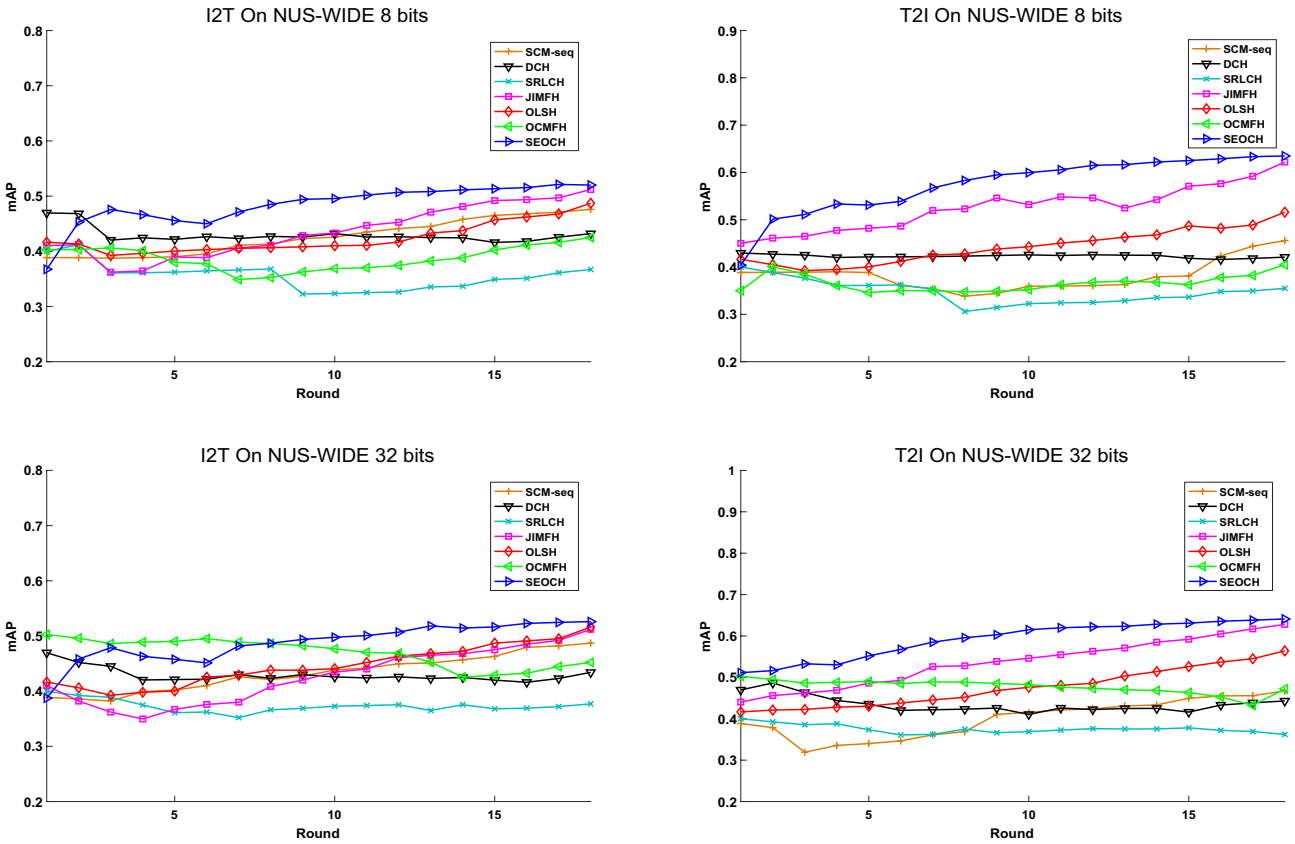

Figure 2. The mAP scores at each round for two retrieval tasks on the **NUS-WIDE** dataset.

Task	Methods	code length				
		8 bits	16 bits	32bits	64 bits	128 bits
Image to Text	SEOCH-I	0.609	0.607	0.587	0.571	0.561
	SEOCH-II	0.64	0.554	0.579	0.549	0.568
	SEOCH-III	0.577	0.605	0.612	0.635	0.612
	SEOCH	**0.68**	**0.645**	**0.663**	**0.64**	**0.62**
Text to Image	SEOCH-I	0.623	0.619	0.602	0.581	0.566
	SEOCH-II	0.667	0.578	0.598	0.59	0.577
	SEOCH-III	0.575	0.617	0.632	0.649	0.658
	SEOCH	**0.693**	**0.695**	**0.696**	**0.68**	**0.66**

Table 3. Ablation study on the **MIRFLICKR-25K** dataset (The numbers in bold indicate the best performance).

bits, SEOCH-II exhibits the lowest result; for 128 bits, SEOCH-I demonstrates the lowest result. Hence, it can be concluded that each component in our proposed SEOCH plays a significant role in the retrieval outcomes.

Time cost analysis

To validate the efficiency of the proposed SEOCH, we conducted additional experiments on the MIRFLICKR-25K dataset to compare the training times of the baseline methods and SEOCH. In these experiments, we configured the hash code length to be 8 bits and 32 bits respectively. The training times of the two online methods under the same configurations are presented in Table 4.

From Table 4, it is evident that the proposed SEOCH not only achieves superior retrieval performance but also exhibits the shortest training time. Hence, the retrieval efficiency has been significantly enhanced.

Conclusion

This paper is focused on harnessing the semantic correlation between different modalities and enhancing the efficiency of cross-modal retrieval in online scenarios. In this paper, we propose an innovative approach called Semantic Embedding based Online Cross-modal Hashing (SEOCH). SEOCH integrates the exploitation of

Bits	Methods	Chunk1	Chunk2	Chunk3	Chunk4	Chunk5
8	OLSH	4.6	1.34	1.32	1.32	1.33
	OCMFH	13.35	1.52	1.36	1.43	1.49
	SEOCH	2.01	0.64	1.75	1.03	1.19
32	OLSH	5.4	1.41	1.42	1.52	1.53
	OCMFH	15.74	1.43	1.34	1.93	1.79
	SEOCH	2.42	0.82	1.93	1.37	1.9

Table 4. Training time (in seconds) on the **MIRFLICKR-25K** dataset.

semantic information and online learning into a unified framework. To leverage semantic information, we map semantic labels to a latent semantic space and construct a semantic similarity matrix to preserve the similarity between new and existing data in the Hamming space. Moreover, we employ a discrete optimization strategy for online hashing. Extensive experiments on two publicly available multi-label datasets validate the superiority of SEOCH.

Data availability

The datasets generated and/or analysed during the current study can be accessed as follows: Download the [NUSWIDE.mat] dataset from https://pan.baidu.com/s/1WEAezxn6mbEbqekPjBnRQw, password: 8888. Download the [MIRFLICKR.mat] dataset from https://pan.baidu.com/s/1GT-mrUutslGhp3lP2i_rYQ, password: 8888. The source code of Semantic embedding based online cross-modal hashing method are available from the corresponding author on reasonable request.

Received: 4 July 2023; Accepted: 17 December 2023
Published online: 06 January 2024

References

1. Rasiwasia, N. Pereira, J. C., & Coviello, E. A new approach to cross-modal multimedia retrieval. in *International Conference on Multimedia*, pp. 251–260 (2010).
2. Lew, M. S., Sebe, N., Djeraba, C. & Jain, R. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl* **2**(1), 1–19 (2006).
3. Yu, E., Sun, J., Wang, L., Wan, W. & Zhang, H. Coupled feature selection based semisupervised modality-dependent cross-modal retrieval. *Multimed. Tools Appl.* **78**, 28931–28951 (2018).
4. Wang, L. *et al.* Joint feature selection and graph regularization for modality-dependent cross-modal retrieval. *J. Vis. Commun. Image Represent.* **54**, 213–222 (2018).
5. Meng, M. *et al.* Asymmetric supervised consistent and specific hashing for cross-modal retrieval. *IEEE Trans. Image Process.* **30**, 986–1000 (2021).
6. Liu, X., Wang, X. & Cheung, Y. M. FDDH: Fast discriminative discrete hashing for large-scale cross-modal retrieval. *IEEE Trans. Neural Netw. Learn. Syst.* **33**(11), 6306–6320 (2022).
7. Yang, Z. *et al.* NSDH: A nonlinear supervised discrete hashing framework for large-scale cross-modal retrieval. *Knowl. Based Syst.* **217**(3), 106818 (2021).
8. Hu, M. *et al.* Collective reconstructive embeddings for cross-modal hashing. *IEEE Trans. Image Process.* **28**(6), 2770–2784 (2019).
9. Wu, L., Wang, Y. & Shao, L. Cycle-consistent deep generative hashing for cross-modal retrieval. *IEEE Trans. Image Process.* **28**(4), 1602–1612 (2019).
10. Zhuang, Y., Wang, Y., Wu, F., Zhang, Y., & Lu, W. Supervised coupled dictionary learning with group structures for multi-modal retrieval. in *Proc. 27th AAAI Conf. Artif. Intell.*, pp. 1070–1076 (2013).
11. Wu, G., *et al.* Unsupervised deep hashing via binary latent factor models for large-scale cross-modal retrieval. in *Proc. 27th Int. Joint Conf. Artif. Intell.*, pp. 2854–2860 (2018).
12. Zhou, J., Ding, G., & Guo, Y. Latent semantic sparse hashing for cross-modal similarity search. in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, pp. 415–424 (2014).
13. Ding, G., Guo, Y., & Zhou, J. Collective matrix factorization hashing for multimodal data. in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2083–2090 (2014).
14. Liu, H., Ji, R., Wu, Y., Huang, F., Zhang, B. Cross-modality binary code learning via fusion similarity hashing. in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (CVPR), pp. 6345–6353 (2017).
15. Long, M., Cao, Y., Wang, J., & Yu, P. S. Composite correlation quantization for efficient multimodal retrieval. in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, pp. 579–588 (2016).
16. Li, C., *et al.* Selfsupervised adversarial hashing networks for cross-modal retrieval. in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 4242–4251 (2018).
17. Tang, J., Wang, K. & Shao, L. Supervised matrix factorization hashing for cross-modal retrieval. *IEEE Trans. Image Process.* **25**(7), 3157–3166 (2016).
18. Cao, Y., Long, M., Wang, J., & Zhu, H. Correlation autoencoder hashing for supervised cross-modal search. in *Proc. ACM Int. Conf. Multimedia Retr. (ICMR)*, pp. 197–204 (2016).
19. Huang, H. J., *et al.* Supervised cross-modal hashing without relaxation. in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, pp. 1159–1164 (2017).
20. Luo, X., Zhang, P., Huang, Z., Nie, L. & Xin Shun, Xu. Discrete hashing with multiple supervision. *IEEE Trans. Image Process.* **28**(6), 2962–2975 (2019).
21. Luo, X., Wu, Y., & Xu, X. Scalable supervised discrete hashing for large-scale search. in *Proc. World Wide Web Conf*, pp. 1603–1612 (2018).
22. Luo, X. SDMCH: Supervised discrete manifold-embedded cross-modal hashing. in *Proc. Int. Joint Conf. Artif. Intell*, pp. 2518–2524 (2018).

23. Li, C., *et al.* SCRATCH: A scalable discrete matrix factorization hashing for cross-modal retrieval. in *Proceedings of the ACM International Conference on Multimedia*, pp. 1–9 (2018).
24. Zhu, L. *et al.* Discrete multimodal hashing with canonical views for robust mobile landmark search. *IEEE Trans. Multimed.* **19**(9), 2066–2079 (2017).
25. Yao, Tao *et al.* Online latent semantic hashing for cross-media retrieval. *Pattern Recogn.* **89**, 1–11 (2019).
26. Wang, D., Wang, Q., An, Y., Gao, X., Tian, Y. Online collective matrix factorization hashing for large-scale cross-media retrieval. in *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 1409–1418 (2020).
27. Song, J., Yang, Y., Yang, Y., *et al.* Inter-media hashing for large-Scale retrieval from heterogeneous data sources. in *Proceedings of ACM SIGMOD*, pp. 785–796 (2013).
28. Zhang, D., & Li, W. Large-scale supervised multimodal hashing with semantic correlation maximization. in *Twenty-eighth Aaai Conference on Artificial Intelligence*, pp. 2177–2183 (2014).
29. Lin, Z. *et al.* Cross-view retrieval via probability-based semantics-preserving hashing. *IEEE Trans. Cybern.* **47**(12), 4342–4355 (2017).
30. Xu, X. *et al.* Learning discriminative binary codes for largescale cross-modal retrieval. *IEEE Trans. Image Process.* **26**(5), 2494–2507 (2017).
31. Liu, L., Yang, Y., Hu, M., *et al.* Index and retrieve multimedia data: Cross-modal hashing by learning subspace relation. in *International Conference on Database Systems for Advanced Applications,* pp. 606–621 (2018).
32. Wang, D., Gao, X., Wang, X., *et al.* Semantic topic multimodal hashing for cross-media retrieval. in *Proceedings of the International Conference on Artificial Intelligence*, pp. 3890–3896 (2015).
33. Liu, X. Q. *et al.* Deep cross-modal hashing based on semantic consistent ranking. *IEEE Trans. Multimed.*https://doi.org/10.1109/TMM.2023.3254199 *(2023).*
34. Yu, E. *et al.* Deep discrete cross-modal hashing with multiple supervision. *Neurocomputing* **14**, 486 (2022).
35. Zhang, L. *et al.* Deep top-k ranking for image-sentence matching. *IEEE Trans. Multimed* **22**(3), 775–785 (2019).
36. Wang, D. *et al.* Joint and individual matrix factorization hashing for large-scale cross-modal retrieval. *Pattern Recogn.* **107**, 107479 (2020).

Acknowledgements

This work was supported by the following Grants: Talent Project of Shandong Women's University under Grant 2020RCYJ21, 2018RC34061; Opening Fund of Shandong Provincial Key Laboratory of Network Based Intelligent Computing; Cultivation Fund of Shandong Women's University High-level Scientific Research Project (2022GSPSJ02).

Author contributions

M.Z. designed the experiments and wrote the main manuscript text. J.L. conducted computational work, and prepared all figures and/or tables. X.Z. provided the some idea about model. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.