sensors

MDPI

Article

Object-Level Visual-Text Correlation Graph Hashing for Unsupervised Cross-Modal Retrieval

Ge Shi, Feng Li, Lifang Wu * and Yukun Chen

Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China; shige@bjut.edu.cn (G.S.); lifeng9619@emails.bjut.edu.cn (F.L.); imbachenyukun@emails.bjut.edu.cn (Y.C.)
* Correspondence: lfwu@bjut.edu.cn

Abstract: The core of cross-modal hashing methods is to map high dimensional features into binary hash codes, which can then efficiently utilize the Hamming distance metric to enhance retrieval efficiency. Recent development emphasizes the advantages of the unsupervised cross-modal hashing technique, since it only relies on relevant information of the paired data, making it more applicable to real-world applications. However, two problems, that is intro-modality correlation and inter-modality correlation, still have not been fully considered. Intra-modality correlation describes the complex overall concept of a single modality and provides semantic relevance for retrieval tasks, while inter-modality correction refers to the relationship between different modalities. From our observation and hypothesis, the dependency relationship within the modality and between different modalities can be constructed at the object level, which can further improve cross-modal hashing retrieval accuracy. To this end, we propose a Visual-textful Correlation Graph Hashing (OVCGH) approach to mine the fine-grained object-level similarity in cross-modal data while suppressing noise interference. Specifically, a novel intra-modality correlation graph is designed to learn graph-level representations of different modalities, obtaining the dependency relationship of the image region to image region and the tag to tag in an unsupervised manner. Then, we design a visual-text dependency building module that can capture correlation semantic information between different modalities by modeling the dependency relationship between image object region and text tag. Extensive experiments on two widely used datasets verify the effectiveness of our proposed approach.

Keywords: cross-modal hash learning; deep model; hashing retrieval

Citation: Shi, G.; Li, F.; Wu, L.; Chen, Y. Object-Level Visual-Text Correlation Graph Hashing for Unsupervised Cross-Modal Retrieval. *Sensors* **2022**, *22*, 2921. https://doi.org/10.3390/s22082921

Received: 23 February 2022
Accepted: 7 April 2022
Published: 11 April 2022

1. Introduction

The rapid development of multimedia data, such as text, image, video, and audio, has greatly enriched multimedia data diversity. Recently, cross-modal retrieval [1–4] has attracted more and more attention. It aims to retrieve instances of another modality given a query instance of one modal. However, the rapid growth of multimedia data puts forward higher requirements for retrieval efficiency. Many hashing methods have been proposed to tackle these challenges. Hashing methods can map high-dimensional features into binary hash codes, transforming information retrieval into Hamming distance calculation and greatly improving retrieval efficiency.

Due to the existence of "heterogeneous gaps" between different modalities, the similarity of different modalities cannot be measured directly. Existing hashing methods solve this problem mainly by mapping data of different modalities into a common Hamming space. The supervised cross-modal hashing methods [5–7] use a large amount of manually annotated data to capture the correlation information between different modalities so as to achieve satisfactory performance. However, the data annotation is labor intensive and high cost, making it infeasible in practical scenarios. Unlike the supervised method, the unsupervised hashing method only relies on weakly correlated information between modalities. It does not require annotated data, making it easier to deploy to real-world applications.

Most existing unsupervised cross-modal hashing methods utilize global-level intra-modality and inter-modality similarities to learn the hash functions. UGACH [8] made full use of GAN's ability for unsupervised representation learning to exploit the underlying manifold structure of cross-modal data. UCH [9] combined generative adversarial networks and constructed two recurrent networks in an end-to-end framework to learn common feature representation and high-quality hash codes. DJSRH [10] improved the performance of retrieval by constructing the joint-semantic matrix and reconstructing with binary codes.

Driven by a large scale of image and text pairs with weakly semantic correlation, these methods have achieved great success. However, the text data consist of a series of tags provided by users, not annotated according to strict standards, and contains much noise. The direct use of the global level intra-modality and inter-modality similarity to learning hash functions will inevitably introduce a large amount of noise, leading to overfitting and degeneration models. In fact, both images and text have fine-grained granularities, i.e., objects in an image. Some object and tag pairs have slight similarities or even irrelevant at the fine-grained object level.

Figure 1 shows a set of image-text pairs in the MIRFlickr dataset. It can be seen that there is a corresponding relationship between the objects in the image and the tags in the text. Meanwhile, the unframed tags in the text are entirely unrelated to the image and are noise information. The existing methods based on global level mapping can not capture the similarities between objects and tags on fine-grained level, introducing much noise and resulting in non-optimal models.

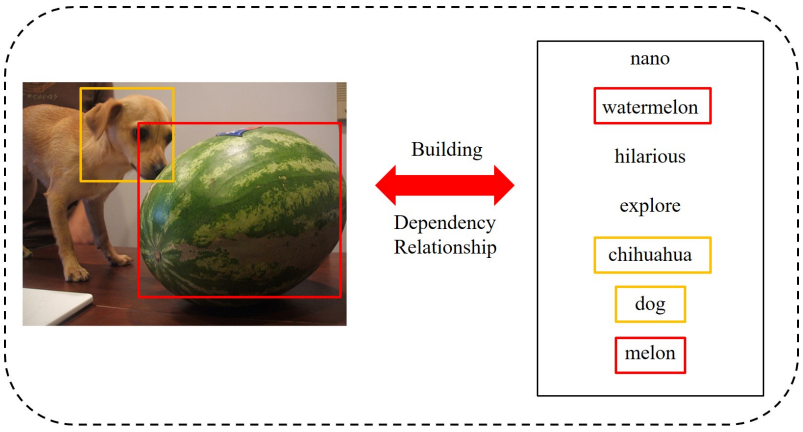

Figure 1. A pair of similar image text samples from the MIRFlickr dataset.

To solve the above problems, we propose a novel Object-level Visual-text Correlation Graph Hashing (OVCGH) to model the corresponding relationship between image object regions and text tags. Specifically, based on the idea of deep graph infomax [11], OVCGH can construct an object-level visual correlation graph and a text correlation graph based on the image object region and the text tag, which not only learn the correlation between the objects of the single model data, but also do not ignore the overall information lost due to the object-level correlation. OVCGH constructs a fine-grained image object area and text tag relationship pair based on the existing correlation graph, which ensures that similar object-tag pairs in cross-modal data can have higher scores and effectively avoid the impact of noise on the model. Finally, the accurate hash codes are learned under the joint supervision of object-level dependency and global-level mapping. The contributions of this paper are summarized as follows:

- We propose a novel object-level visual-text correlation graph hashing (OVCGH) to mine the fine-grained object-level similarity existing in cross-modal data and suppress noise interference. The results on two public datasets show that our method achieves the best experimental performance.
- We design an object-level correlation graph building module that can construct the object-level similarity in the modal. It can construct the graph structure for the image

modal and the text modal at the object-level in an unsupervised manner. Furthermore, the constructed graph structure contains the global information of its original semantic structure.

- We design a novel cross-modal dependency building module to build the object-level similarity between modalities while avoiding noise interference. It can model the dependency between the image object region and the text tag, ensuring that cross-modal data with similar objects can have a higher similarity score.

2. Related Work

The goal of cross-modal retrieval is to retrieve instances of another modality given a query instance of one modal. However, the amount of multimedia data is increasing rapidly, which sets higher requirements for improving the efficiency of retrieval. In order to solve this problem, many hashing methods [12–14] have been proposed to achieve fast and efficient cross-modal retrieval. Next, we divide the cross-modal hashing methods into two directions: unsupervised methods and supervised methods and introduce related work, respectively.

Supervised cross-modal hashing methods use labels to capture rich correlation information between different modalities to achieve cross-modal retrieval. Traditional supervised learning methods are mostly based on handcrafted features to learn semantic relevance in the common space. Bronstein et al. [14] proposed Cross-Modality Similarity Sensitive Hashing by using a boosted classification paradigm for hash learning. SCM [15] learned the hash function by constructing and maintaining the semantic similarity matrix. SePH [16] proposed a Semantics-Preserving Hashing method, which aimed to approximate the distribution of semantic labels and the distribution of hash codes on the Hamming space by minimizing the KL divergence.

Jiang et al. [17] proposed a Deep Cross-Modal Hashing method, which is one of the earliest methods to combine convolutional neural networks with cross-modal hashing. Li et al. [18] further improved this work and proposed Self-Supervised Adversarial Hashing, which was one of the first attempts to apply adversarial learning to the cross-modal hashing problem. Xu et al. [19] proposed a Graph Convolutional Hashing method, which used an association graph learning model to explore the inherent similarity structure among data. This method can help generate discriminants for the hash codes. Zhan et al. [7] proposed Supervised Hierarchical Deep Cross-modal Hashing, which implanted the similarity at each layer of the label hierarchy and the relatedness across different layers into the hash code learning. Although supervised methods achieved satisfactory performance, obtaining a large number of similar labels is usually expensive and very tricky, which makes supervised methods infeasible in real-world applications.

Unsupervised cross-modal hashing methods only rely on the relevant information in the paired data, making it easier to deploy to real-world applications. These methods usually learn hash codes by preserving inter-modality corrections. Traditional unsupervised cross-modal hashing methods mainly learn linear projection matrices by optimizing statistical analysis methods. Thompson et al. (CCA) [20] learned a subspace to maximize the correlation between two sets of heterogeneous data. IMH [21] proposed an Inter-Media Hashing method, which established a common Hamming space by maintaining the similarity between the inter-media and intra-media. Kumar et al. [22] extended Spectral Hash [23] to cross-modal retrieval and achieved better results.

Recently, deep learning techniques have also been developed to deal with cross-modal retrieval. UGACH [8] proposed an Unsupervised Generative Adversarial Cross-modal Hashing method, which made full use of GAN's ability for unsupervised representation learning to exploit the underlying manifold structure of cross-modal data. As an improvement, UCH [9] combined generative adversarial networks and constructed two recurrent networks in an end-to-end framework to learn common feature representation and high-quality hash codes. Su et al. [10] proposed Deep Joint-Semantic Reconstruction Hashing,

which further improved the performance of retrieval by constructing the joint-semantic matrix and reconstructing with binary codes.

Although unsupervised cross-modal hashing has obvious advantages in reducing the burden of data annotation, the accuracy is usually lower than a satisfactory level. Moreover, the existing methods mainly focus on mining the relationships between samples and consider them from the perspective of measuring the similarity of different samples without focusing on semantic information, so that the semantic similarity in the training data is not fully utilized, which motivates us to build the fine-grained dependence of the training data on the object level.

3. The Proposed Method

The overall framework of our proposed approach is shown in Figure 2, which consists of two stages: (1) the graph-level representations learning stage; (2) the graph-guided hashing learning stage. Based on the similarity between image object regions and tags, the graph-level representations learning stage constructs two relation graph structures named visual feature graph and textual feature graph. The graph uses the strategy of maximizing the mutual information of local and global features to learn the graph-level representation of the image object regions and the text tags, respectively. In this training method, the nodes of the relation graph learn high-level representations that contain global information. Furthermore, the dependency between the object region and the tag can be modeled according to the existing relation graph structure. In this way, different modality data can be combined at a fine-grained level rather than as a whole. Through the feature interaction between the object-tag related pairs, the similarity score of the image and text at the object level S^O can be obtained, which will be more conducive to judging the similarity between the image and the text. In addition, the joint representations that aggregate all nodes of the graph are used to calculate the global-level score S^L between the image and the text. Finally, according to S^O and S^L, the graph-guided hashing learning stage obtains the object-level pair-wise loss and joint semantic reconstruction loss, respectively, and the two loss functions are used to jointly optimize the learning of accurate hash codes.

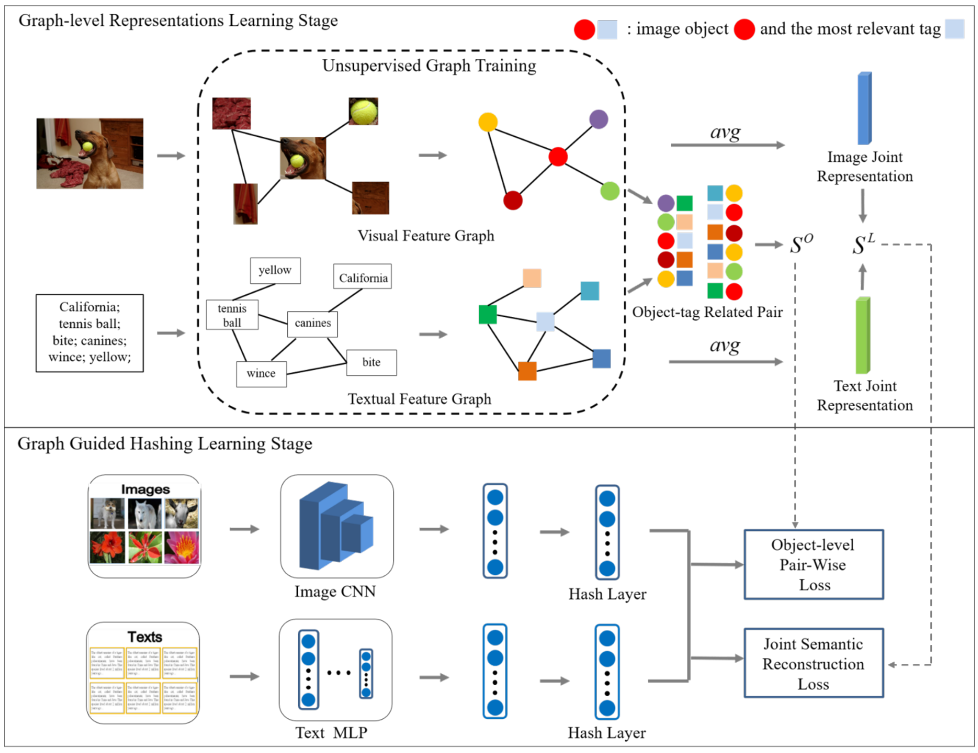

Figure 2. The overall framework of our proposed visual-textual correlation graph hashing (VCGH) approach.

3.1. Problem Definition

Next, we introduce some definitions used in this paper. For the cross-modal retrieval task, given the p instances of image-text pair input: $X = \{(x_i^I, x_i^T)\}_{i=1}^{p}$, where x_i^I denotes the ith image sample and x_i^T denotes the ith text sample. The goal of cross-modal hashing is to learn a hash function to generate hash codes of length l for different modalities, which maps data of different modalities into a common Hamming space for direct comparison.

In our method, a pre-trained visual feature extractor is used to encode the image object regions into feature vectors, and the visual feature extractor uses a pre-trained CNN network and an object detector (Faster-RCNN [24]). For each image x_i^I, the visual feature extractor encodes each image object region as a d_1 dimension visual feature vector $O_i = \{o_{i,j}\}_{j=1}^{n_i}$, where n_i is the number of object regions contained in the ith image, and $o_{i,j}$ is the jth visual feature vector of the image x_i^I. Each text x_i^T is composed of multiple tags, so it is defined as $x_i^T = \{t_{i,j}\}_{j=1}^{m_i}$, where $t_{i,j}$ is the jth tag and m_i is the number of tags contained in the ith text.

3.2. Graph-Level Representations Learning

For each image and text, VCGH constructs visual undirected graph $G_i^I = (V_i^I, E_i^I)$ and text undirected graph $G_i^T = (V_i^T, E_i^T)$, respectively. For visual undirected graphs, V^I is the set of all image object features related to x_i^I, and E^I is the set of edges. The node set in G^I is defined as follows:

$$V_i^I = \{o_{i,j}\}_{j=1}^{n_i} V^I \in R^{n_i \times d_1} \tag{1}$$

where $o_{i,j}$ defines each node in the node set, and two nodes are defined as neighbors if the cosine similarity between them is larger than zero. For G^T, each node defines the word embedding of the corresponding text tag, so the node set of the text undirected graph is defined as follows:

$$V_i^T = \{F(t_{i,j})\}_{j=1}^{m_i} V^T \in R^{m_i \times d_2} \tag{2}$$

where $v_{i,j}$ is a d_2 dimension word embedding representation of each text tag, which is mapped by the FastText model [25]. The FastText model is a widely used word-embedding extraction tool. Since the original text tags contain many uncommon words, the FastText model can be used to achieve better results. For each node in the text undirected graph, we use the same neighbor definition method as the visual undirected graph.

Leonardo et al. [26] pointed out that existing graph convolutional networks either use global nodes, which assume that each node is connected to other nodes, or use local nodes, which are connected to only a part of the nodes. The former ignores the structure of the graph, while the latter lacks global information. Inspired by [11], through the strategy of maximizing local and global mutual information, the relationship graph can capture not only the local structure information of the graph, but also model the overall knowledge of the graph. Therefore, Deep Graph Infomax [11] is used to learn the relationship between image object regions and text tags. Specifically, taking the image feature graph as an example, our training goal is to maximize the mutual information between each image object region and the entire image, so we can learn a mapping from the undirected visual graph to a high-level representation of each node:

$$\varepsilon_{\theta_I} : \left(V_i^I, E_i^I\right) \rightarrow Z_i^I \tag{3}$$

where $Z_i^I = \{z_{i,1}, z_{i,2}, \cdots, z_{i,j}, \cdots, z_{i,n_i}\}$ is the high-level representation set of the nodes in the image feature graph, and the parameter θ_I is the learnable parameter of the image feature graph. Through this strategy, the image object region representation $o_{i,j}$ can be mapped to a d dimensional representation $z_{i,j}$ that contains the global information of the image. The text undirected graph can then be obtained by analogy as follows:

$$\varepsilon_{\theta_T} : \left(V_i^T, E_i^T\right) \rightarrow Z_i^T \tag{4}$$

where $Z_i^I = \{z_{i,1}, z_{i,2}, \cdots, z_{i,j}, \cdots, z_{i,m_i}\}$ is the high-level representation set of the nodes in the textual feature graph, the parameter θ_T is the learnable parameter of the textual feature graph, and $z'_{i,j}$ is a d dimension high-level representation containing the global information of the text.

When matching images and texts, the main judgment is based on the similarity between objects in the image and the text. Inspired by this, it is necessary to construct the similarity score between the image and the text at the object level. Thus, we construct a dependency between the image object region and the text tag based on the above graph structure. Specifically, based on the trained visual feature graph and textual feature graph, we can transform each image and each text into a d-dimensional vector set. Each vector in the set represents a high-level representation of the image object region or text tag, respectively. Furthermore, if an image object region is associated with a text tag, the object region and the tag will have a higher similarity score. The inner product method is used to measure the similarity score between the image region vector and the text tag vector. Therefore, the similarity score of the text-to-image is defined as follows:

$$S = \frac{1}{m_i} \sum_{k=1}^{m_i} \max_{j \in [1,n_i]} z_{i,j} \top z'_{i,k} \tag{5}$$

where $\max_{j \in [1,n_i]} z_{i,j}^T z'_{i,k}$ represents the similarity score between the kth tag in the text x_i^T and the most relevant object region in the image x_i^I. Furthermore, the object-level similarity score of the text-to-image is obtained by the mean value method. Using the same method, we can obtain the similarity score of the image-to-text. The final object-level score is defined as follows:

$$\begin{aligned} S^O = \frac{1}{2}\Big(&\frac{1}{m_i} \sum_{k=1}^{m_i} \max_{j \in [1,n_i]} z_{i,j} \top z'_{i,k} \\ &+ \frac{1}{n_i} \sum_{k=1}^{n_i} \max_{j \in [1,m_i]} z'_{i,j} \top z_{i,k}\Big) \end{aligned} \tag{6}$$

In addition to the object-level score, the global-level score is also considered to obtain the similarity between the image and the text. Specifically, after training the visual feature graph and the text feature graph, the global-level representation r_i and r'_i of the image and text is calculated by the mean method:

$$\begin{aligned} r_i &= \frac{1}{m_i} \sum_{k=1}^{m_i} z_{i,k} \\ r_i' &= \frac{1}{n_i} \sum_{k=1}^{n_i} z'_{i,k} \end{aligned} \tag{7}$$

Finally, we calculate the cosine similarity of r_i and r'_i to obtain the global-level score $S^G = \cos(r_i, r'_i)$ between the image x_i^I and the text x_i^T.

3.3. Graph Guided Hashing Learning

The goal of VCGH is to learn accurate hash codes for different modalities through the supervision of visual feature graphs and text feature graphs so that different modal data with similar semantic information are close in the public Hamming space. Specifically, the VCGH network contains two branch structures. Each branch contains an original space layer and a hash layer. The original space layer maps the original features of different modalities to the same common space for direct measurement, which can be expressed as:

$$\phi_c(f) = \tanh(w_c f + b_c) \tag{8}$$

where $f = \{f^I, f^T\}$ is the original feature of different modalities, i.e., f^I for an image and f^T for a text, w_c denotes the weights, and b_c is the bias parameter.

Through the hash layer, the common representations of different modalities are mapped to binary hash codes. which can be expressed as:

$$h(f) = sigmoid(W_h \phi_c(f) + v)) \tag{9}$$

where W_h is the weight matrix and v is the bias parameter. Then a threshold function is used to get a final binary hash codes of length l:

$$b(f) = sgn(h_k(f) - 0.5), k = 1, 2, \ldots, l \tag{10}$$

To preserve the original semantic information of the obtained hash codes as much as possible, VCGH uses the object-level score and the global-level score obtained from the graph as a joint guide to supervising network learning. Specifically, we first construct a similarity structure H by using the average inner product of different model hash codes:

$$H = \frac{1}{l} b(f^I) \top b(f^T) \tag{11}$$

where $b(f^I)$ and $b(f^T)$ are the hash codes of the image and text, respectively, and the average result is obtained by dividing the code length l. We minimize the L_2 loss of similar structure between the hash codes and the object level for different modal data, which is defined as the object-level pair-wise loss:

$$L_o = \min (H - S^o)^2 \tag{12}$$

For our proposed joint semantic reconstruction loss, we first calculate cosine similarity matrices $S^I = \cos(f^I, f^I)$ and $S^T = \cos(f^T, f^T)$ to represent the original semantic similarity structure of the image modality and text modality, respectively, and a weighted fusion strategy is used to jointly express the global-level similarity of different semantic information structures:

$$S^F = \alpha S^I + \beta S^T + \gamma S^G \tag{13}$$

where α, β, and γ are weight parameters and S^F is the global-level similarity matrix. Furthermore, we construct the joint semantic reconstruction loss to keep the hash codes discriminative from the intra-modal and inter-modal perspectives:

$$\begin{aligned} L_G = \min \parallel \lambda S^F - \cos(b(f^I), b(f^T)) \parallel_F^2 \\ + \gamma \parallel \lambda S^F - \cos(b(f^I), b(f^I)) \parallel_F^2 \\ + \eta \parallel \lambda S^F - \cos(b(f^T), b(f^T)) \parallel_F^2 \end{aligned} \tag{14}$$

where $\| \cdot \|_F^2$ represents the *Frobenius* norm, λ is a hyper-parameter used to adjust the proportion of the global-level similarity structure, and λ and η are used to adjust the proportion of the intra-modal similarity. Finally, the loss function of VCGH is defined as follows:

$$L = L_o + L_G \tag{15}$$

4. Experiments

In this section, we conduct some experiments on our proposed method. We first introduce the datasets, evaluation methods, and implementation details. Then, we compare and analyze the experimental results of VCGH with eight state-of-the-art methods. Finally, we conduct an ablation study to verify the importance of each part of VCGH.

4.1. Datasets and Compared Methods

The proposed VCGH and compared methods are evaluated on two widely used datasets: MIRFlickr [27] and NUS-WIDE [28].

MIRFlickr [27] contains 25,000 images and is annotated with 24 labels. Each image is associated with text tags and annotated with at least one label. Following [29], we use 20,015 image-text pairs in our experiments, where 2000 are preserved as the query set and the rest are used for retrieval database. A total of 5000 image-text pairs are randomly selected from the retrieval database as the training set.

NUS-WIDE [28] contains 269,648 images and is annotated with 81 labels. Each image is also associated with text tags and labeled with one or more of the labels. However, there are overlaps among the labels. Following [29], we selected the top 10 most frequent labels and the corresponding 186,577 image-text pairs as the database set. We randomly selected 2000 image-text pairs from the dataset as the query set and the others as the retrieval database, while 5000 image-text pairs are randomly selected from the retrieval database as the training set.

The proposed VCGH is compared with eight state-of-the-art methods. They are six unsupervised methods, i.e., CVH [22], PDH [30], CMFH [31], CCQ [32], DJSRH [10], MGAH [33], and two supervised methods, i.e., CMSSH [14] and SCM [15].

4.2. Evaluation Methods

In the experiment, we use two cross-modal retrieval tasks: using image queries to retrieve text in the database (image→text) and using text queries to retrieve images in the database (text→image). Specifically, we use the proposed VCGH method and all compared methods to obtain the hash codes of all query and retrieval database samples. We calculate the Hamming distance between the hash codes of the query and retrieval database samples and get a ranking list. Finally, we use two evaluation methods to evaluate the ranking list to verify the effectiveness of the retrieval: Mean Average Precision (MAP) and precision-recall curve (PRcurve). They are defined as follows:

MAP is a commonly used evaluation tool for information retrieval. It is computed as the mean of average precision (AP) of the retrieval results. The AP is computed as:

$$AP = \frac{1}{R}\sum_{k=1}^{n}\frac{k}{R_k} \times rel_k \tag{16}$$

where n is the size of database, R is the number of relevant images or text in the database, R_k is the number of related images or text in the ranking list of length k, and rel_k means if the kth sample and the query sample have the same label then $rel_k = 1$, otherwise $rel_k = 0$.

The PR curve is obtained according to the precision and recall rate corresponding to each retrieval position.

4.3. Implementation Details

In this section, we present the implementation details of our VCGH in the experiments. For the graph-level representation learning stage, we use the pre-trained object detection model FasterRCNN [24] to detect the object regions in the image and use the 19-layer VGGNet [34] pre-trained on the ImageNet dataset [35] to get the 4096-dimensional features of the object region for each image. Then, the Deep Graph Infomax model is trained for 30 epochs to map the image object region into a 512-dimensional graph-level representation. For text, we use the FastText model [25] to extract the 300-dimensions text feature of each tag and then use the same training method to obtain a 512-dimensional text graph-level representation. For the graph-guided hashing learning stage, we use 19-layer VGGNet and Multilayer Perceptron as the backbone network of our image and text modalities, respectively, and then, a fully connected layer of length is used to obtain the hash codes. We train the network using the mini-batch method and set the batch size to 32. The learning rates are set to 0.001 for the image model and 0.01 for the text model. Moreover, the network is trained for 80 epochs and decreased by a factor of 10 at 60 epochs.

In addition, for a fair comparison, we apply the same implementation of compared methods provided by the authors and follow their best settings to perform the experiments.

We use the same image and text features for all compared methods and use a 19-layer VGGNet pre-trained on the ImageNet dataset to extract image deep features. We finally apply the bag of words model to extract Bow text features for texts.

4.4. Experiment Results

Tables 1 and 2 show the MAP scores from 16 to 128 hash bits on the MIRFlickr dataset and NUS-WIDE dataset. From the results, we can clearly observe that VCGH significantly outperforms the eight state-of-the-art methods on the two datasets. Specifically, on the MIRFlickr dataset, our method achieves the best average MAP score of 0.712 on the image→text task and 0.719 on the text→image task. Compared to the best unsupervised method MGAH, our method achieves a performance improvement from 0.696 to 0.712 on the image→text task and from 0.681 to 0.719 on the text→image task. On the NUS-WIDE dataset, our method also achieves the best average MAP score of 0.628 on the image→text task and 0.632 on the text→image task. These improvements demonstrate the effectiveness of modeling the dependency relationship between the image object region and the text tag by constructing visual feature graphs and text feature graphs.

Table 1. The MAP scores of two retrieval tasks on the MIRFlickr dataset with different lengths of hash codes. The best results are highlighted in bold.

Method	Image→Text				Text→Image			
	16	**32**	**64**	**128**	**16**	**32**	**64**	**128**
CVH [22]	0.602	0.587	0.578	0.572	0.607	0.591	0.581	0.574
PDH [30]	0.623	0.624	0.621	0.626	0.627	0.628	0.628	0.629
CMFH [31]	0.659	0.660	0.663	0.653	0.611	0.606	0.575	0.563
CCQ [32]	0.637	0.639	0.639	0.638	0.628	0.628	0.622	0.618
CMSSH [14]	0.611	0.602	0.599	0.591	0.612	0.604	0.592	0.585
SCM [15]	0.636	0.64	0.641	0.643	0.661	0.664	0.668	0.670
DJSRH [10]	0.659	0.661	0.675	0.684	0.655	0.671	0.673	0.685
MGAH [33]	0.685	0.693	0.704	0.702	0.673	0.676	0.686	0.690
VCGH	**0.701**	**0.711**	**0.714**	**0.723**	**0.705**	**0.713**	**0.725**	**0.733**

Table 2. The MAP scores of two retrieval tasks on the NUS-WIDE dataset with different lengths of hash codes. The best results are highlighted in bold.

Method	Image→Text				Text→Image			
	16	**32**	**64**	**128**	**16**	**32**	**64**	**128**
CVH [22]	0.458	0.432	0.410	0.392	0.474	0.445	0.419	0.398
PDH [30]	0.475	0.484	0.480	0.490	0.489	0.512	0.507	0.517
CMFH [31]	0.517	0.550	0.547	0.520	0.439	0.416	0.377	0.349
CCQ [32]	0.504	0.505	0.506	0.505	0.499	0.496	0.492	0.488
CMSSH [14]	0.512	0.470	0.479	0.466	0.519	0.498	0.456	0.488
SCM [15]	0.517	0.514	0.518	0.518	0.518	0.510	0.517	0.518
DJSRH [10]	0.503	0.517	0.528	0.554	0.526	0.541	0.539	0.570
MGAH [33]	0.613	0.623	0.628	0.631	0.603	0.614	0.640	0.641
VCGH	**0.615**	**0.628**	**0.631**	**0.639**	**0.617**	**0.628**	**0.635**	**0.648**

Figure 3 shows the precision-recall curves of 16-bit hash codes on two datasets. We can observe that our proposed VCGH approach also maintains the best accuracy on image→text tasks and text→image tasks among all compared methods, which further demonstrates the effectiveness of our proposed method.

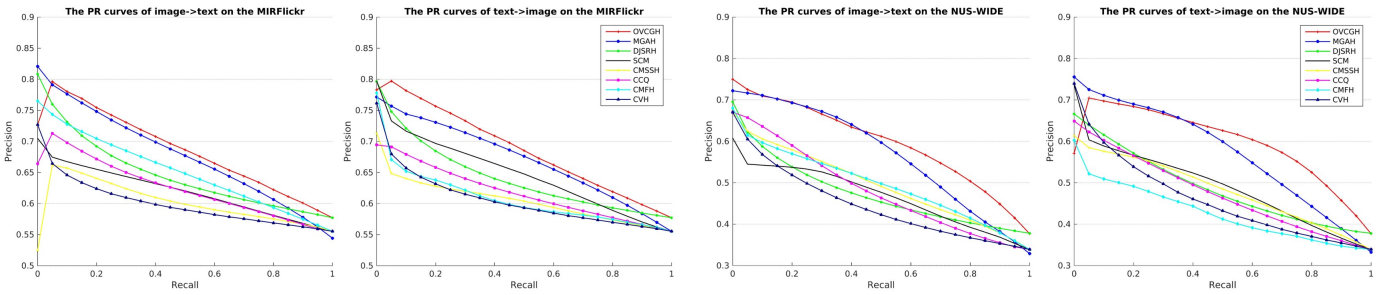

Figure 3. The PR curves on the MIRFlickr and NUS-WIDE datasets with 16-bit hash codes.

4.5. Ablation Study

In this section, we compare three baseline methods to verify the importance of each part in VCGH.

Baseline: we design a baseline method without using visual feature graphs and textual feature graphs named Baseline. Specifically, Baseline trains the network using the joint semantic reconstruction loss without containing the global-level score of the relational graph, $S = \alpha S^I + \beta S^T$, so the training objective function is:

$$\begin{aligned} L = \min \parallel \lambda S - \cos(b(f^I), b(f^T)) \parallel_F^2 \\ +\gamma \parallel \lambda S - \cos(b(f^I), b(f^I)) \parallel_F^2 \\ +\eta \parallel \lambda S - \cos(b(f^T), b(f^T)) \parallel_F^2 \end{aligned} \tag{17}$$

Baseline-global: We introduce the global-level score obtained from the relationship graph on the basis of Baseline, so the objective training function of Baseline-global is defined in Formula (14).

Baseline-object: We add an object-level paired-wise loss based on the Baseline, which is defined in Equation (12).

Table 3 shows the MAP scores by using different variants of the proposed VCGH on MIRFlickr and NUS-WIDE. We can clearly observe that Baseline-global achieves an average performance improvement of 0.014 and 0.01 compared with Baseline for the image→text task and text→image task on the MIRFlickr dataset, which demonstrates the effectiveness of the proposed visual feature graph and text feature graph. We can also see an average performance improvement of 0.048 and 0.029 for the two tasks on the NUS-WIDE dataset, which further demonstrates the above conclusion. By comparing Baseline-object with Baseline, we can see that the average performance of the two tasks on the MIRFlickr dataset increased by 0.016 and 0.017, and the NUS-WIDE dataset also increased by 0.01 and 0.018, which further demonstrates the effectiveness of building the dependency relationship between the image object region and the text tag. By comparing our proposed VCGH method with Baseline-object, We can see that the average performance of the two tasks on the MIRFlickr dataset increased by 0.015 and 0.016, and the NUS-WIDE dataset also increased by 0.016 and 0.016, which demonstrates that the object-level score and the global-level score complement each other and further improve the accuracy of cross-modal retrieval.

Table 3. The MAP scores by using different variants of the proposed VCGH on MIRFlickr and NUS-WIDE. The best results are highlighted in bold.

Task	Method	MIRFlickr				NUS-WIDE			
		16	**32**	**64**	**128**	**16**	**32**	**64**	**128**
image→text	Baseline	0.656	0.682	0.689	0.698	0.507	0.559	0.569	0.581
	Baseline-global	0.681	0.688	0.692	0.703	0.589	0.593	0.611	0.615
	Baseline-object	0.686	0.691	0.699	0.710	0.594	0.611	0.619	0.623
	VCGH	**0.701**	**0.711**	**0.714**	**0.723**	**0.615**	**0.628**	**0.631**	**0.639**
text→image	Baseline	0.664	0.690	0.691	0.700	0.498	0.583	0.576	0.619
	Baseline-global	0.680	0.683	0.687	0.694	0.581	0.588	0.610	0.614
	Baseline-object	0.691	0.696	0.710	0.718	0.597	0.616	0.621	0.628
	VCGH	**0.705**	**0.713**	**0.725**	**0.733**	**0.617**	**0.628**	**0.635**	**0.648**

5. Conclusions

This paper proposes a novel visual-textful correlation graph hashing (VCGH) approach for unsupervised cross-modal retrieval. VCGH uses an unsupervised method to construct visual feature graphs and text feature graphs for different modalities. The proposed features graphs model the information association in the single modality to obtain more semantic information graph-level representations. Furthermore, based on the existing graph structure, VCGH can model the dependency relationship between the image object region and the text tag and improves the accuracy of cross-modal retrieval from the perspective of global-level scores and object-level scores. Experimental results on two widely used datasets demonstrate the superiority of the proposed method, which illustrates the positive effect of mining semantic information to explore cross-modal retrieval tasks.

With the widespread popularity and application of multi-modal data, in the future, we will further extend our approach to more modals for cross-modal hash retrieval, and we wonder whether the semantic information would have great impact on multi-modal scenarios performances.

Author Contributions: Conceptualization, G.S. and F.L.; methodology, F.L.; software, Y.C.; validation, F.L. and Y.C.; formal analysis, G.S. and F.L.; investigation, F.L.; resources, G.S. and L.W.; data curation, F.L. and Y.C.; writing—original draft preparation, F.L.; writing—review and editing, G.S. and L.W.; visualization, G.S.; supervision, L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 62106010, 61976010, 62176011, and 62106011.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing is not applicable to this paper as no new data were created or analyzed in this study.

Acknowledgments: We thank the anonymous reviewers for their valuable comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Deng, C.; Chen, Z.; Liu, X.; Gao, X.; Tao, D. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Trans. Image Process.* **2018**, *27*, 3893–3903. [CrossRef] [PubMed]
2. Kiros, R.; Salakhutdinov, R.; Zemel, R.S. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv* **2014**, arXiv:1411.2539.
3. Wu, Y.; Wang, S.; Huang, Q. Online asymmetric similarity learning for cross-modal retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4269–4278.
4. Zhang, T.; Wang, J. Collaborative quantization for cross-modal similarity search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2036–2045.

5. Liu, W.; Wang, J.; Ji, R.; Jiang, Y.G.; Chang, S.F. Supervised hashing with kernels. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2074–2081.
6. Wang, J.; Liu, W.; Sun, A.X.; Jiang, Y.G. Learning hash codes with listwise supervision. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3032–3039.
7. Zhan, Y.W.; Luo, X.; Wang, Y.; Xu, X.S. Supervised Hierarchical Deep Hashing for Cross-Modal Retrieval. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 20–24 October 2020; pp. 3386–3394.
8. Zhang, J.; Peng, Y.; Yuan, M. Unsupervised generative adversarial cross-modal hashing. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
9. Li, C.; Deng, C.; Wang, L.; Xie, D.; Liu, X. Coupled cyclegan: Unsupervised hashing network for cross-modal retrieval. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January –1 February 2019; Volume 33, pp. 176–183.
10. Su, S.; Zhong, Z.; Zhang, C. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 3027–3035.
11. Velickovic, P.; Fedus, W.; Hamilton, W.L.; Liò, P.; Bengio, Y.; Hjelm, R.D. Deep graph infomax. *Stat* **2018**, *1050*, 21.
12. Liu, H.; Ji, R.; Wu, Y.; Huang, F.; Zhang, B. Cross-modality binary code learning via fusion similarity hashing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7380–7388.
13. Zhang, X.; Lai, H.; Feng, J. Attention-aware deep adversarial hashing for cross-modal retrieval. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 591–606.
14. Bronstein, M.M.; Bronstein, A.M.; Michel, F.; Paragios, N. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 13–18 June 2010; pp. 3594–3601.
15. Zhang, D.; Li, W.J. Large-scale supervised multimodal hashing with semantic correlation maximization. In Proceedings of the AAAI Conference on Artificial Intelligence, Quebec City, QC, Canada, 27–31 July 2014; Volume 28.
16. Lin, Z.; Ding, G.; Hu, M.; Wang, J. Semantics-preserving hashing for cross-view retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3864–3872.
17. Jiang, Q.Y.; Li, W.J. Deep cross-modal hashing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3232–3240.
18. Li, C.; Deng, C.; Li, N.; Liu, W.; Gao, X.; Tao, D. Self-supervised adversarial hashing networks for cross-modal retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4242–4251.
19. Xu, R.; Li, C.; Yan, J.; Deng, C.; Liu, X. Graph Convolutional Network Hashing for Cross-Modal Retrieval; In Proceedings of the International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; pp. 982–988.
20. Thompson, B. Canonical correlation analysis. In *Encyclopedia of Statistics in Behavioral Science*; Wiley: Hoboken, NJ, USA, 2005.
21. Song, J.; Yang, Y.; Yang, Y.; Huang, Z.; Shen, H.T. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, New York, NY, USA, 22–27 June 2013; pp. 785–796.
22. Kumar, S.; Udupa, R. Learning hash functions for cross-view similarity search. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, Barcelona, Spain, 16–22 July 2011.
23. Weiss, Y.; Torralba, A.; Fergus, R. Spectral hashing. *Nips. Citeseer* **2008**, *1*, 4.
24. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv* **2015**, arXiv:1506.01497.
25. Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T. Bag of tricks for efficient text classification. *arXiv* **2016**, arXiv:1607.01759.
26. Ribeiro, L.F.; Zhang, Y.; Gardent, C.; Gurevych, I. Modeling global and local node contexts for text generation from knowledge graphs. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 589–604. [CrossRef]
27. Huiskes, M.J.; Lew, M.S. The mir flickr retrieval evaluation. In Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, Vancouver, BC, Canada, 30–31 October 2008; pp. 39–43.
28. Chua, T.S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; Zheng, Y. Nus-wide: A real-world web image database from national university of singapore. In Proceedings of the ACM International Conference on Image and Video Retrieval, Santorini Island, Greece, 8–10 July 2009; pp. 1–9.
29. Ding, G.; Guo, Y.; Zhou, J. Collective matrix factorization hashing for multimodal data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, Ohio, USA, 23–28 June 2014; pp. 2075–2082.
30. Rastegari, M.; Choi, J.; Fakhraei, S.; Hal, D.; Davis, L. Predictable dual-view hashing. In Proceedings of the International Conference on Machine Learning PMLR, Atlanta, GA, USA, 16–21 June 2013; pp. 1328–1336.
31. Ding, G.; Guo, Y.; Zhou, J.; Gao, Y. Large-scale cross-modality search via collective matrix factorization hashing. *IEEE Trans. Image Process.* **2016**, *25*, 5427–5440. [CrossRef] [PubMed]
32. Long, M.; Cao, Y.; Wang, J.; Yu, P.S. Composite correlation quantization for efficient multimodal retrieval. In Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, Italy, 17–21 July 2016; pp. 579–588.

33. Zhang, J.; Peng, Y. Multi-pathway generative adversarial hashing for unsupervised cross-modal retrieval. *IEEE Trans. Multimed.* **2019**, *22*, 174–187. [CrossRef]
34. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
35. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]