Self-Supervised Multi-Modal Knowledge Graph Contrastive Hashing for Cross-Modal Search

Meiyu Liang, Junping Du*, Zhengyang Liang, Yongwang Xing, Wei Huang, Zhe Xue

Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing, China
{meiyu1210, junpingd, 1ce, 1416642324, xuezhe}@bupt.edu.cn

Abstract

Deep cross-modal hashing technology provides an effective and efficient cross-modal unified representation learning solution for cross-modal search. However, the existing methods neglect the implicit fine-grained multimodal knowledge relations between different modalities such as when the image contains information that is not directly described in the text. To tackle this problem, we propose a novel self-supervised multi-grained multi-modal knowledge graph contrastive hashing method for cross-modal search (CMGCH). Firstly, in order to capture implicit fine-grained cross-modal semantic associations, a multi-modal knowledge graph is constructed, which represents the implicit multimodal knowledge relations between the image and text as inter-modal and intra-modal semantic associations. Secondly, a cross-modal graph contrastive attention network is proposed to reason on the multi-modal knowledge graph to sufficiently learn the implicit fine-grained inter-modal and intra-modal knowledge relations. Thirdly, a cross-modal multi-granularity contrastive embedding learning mechanism is proposed, which fuses the global coarse-grained and local fine-grained embeddings by multihead attention mechanism for inter-modal and intra-modal contrastive learning, so as to enhance the cross-modal unified representations with stronger discriminativeness and semantic consistency preserving power. With the joint training of intra-modal and inter-modal contrast, the invariant and modal-specific information of different modalities can be maintained in the final cross-modal unified hash space. Extensive experiments on several cross-modal benchmark datasets demonstrate that the proposed CMGCH outperforms the state-of the-art methods.

Introduction

With the rapid development of the Internet and social networks, a large amount of muti-modal data such as text, image, and video have been generated. These massive amounts of multi-modal data contain very valuable information, and the descriptions of different modal data are complementary. Correspondingly, the demands for effective and efficient cross-modal search technologies are significantly increasing, which has attracted extensive attention in recent years. It aims to search one modal data from another different modal data that are most semantically relevant to the given query, such as searching images from texts, or searching texts from images. However, due to the problem of feature heterogeneity and semantic gap between different modal data, they are not directly comparable, which is an important challenge for achieving effective and efficient cross-modal search. To tackle this challenge, an effective solution is to learn a unified joint embedding space based on visual-semantic embedding, where the semantic similarities between the embeddings of different modal data are optimized to be maximum.

Owing to excellent nonlinear feature learning ability of deep learning, deep learning based cross-modal visual-semantic embedding learning methods have attracted broad attention which utilize the deep neural networks to extract global representations of both images and texts and then perform cross-modal alignment and fusion (Qian et al. 2021; Zhuo et al. 2020; Peng, Qi, and Yuan 2019; Wang et al. 2017; Wei et al. 2017). Due to low storage cost and fast searching speed, cross-modal deep hashing methods have been increasingly popular (Chen et al. 2021; Xu et al. 2019; Wang et al. 2019; Zhang, Peng, and Yuan 2018a; Wendel et al. 2019). However, the above methods do not fully utilize multi-modal knowledge for reasoning more cross-modal semantic knowledge relations hindered between image and text.

To address this problem, some works incorporate the commonsense knowledge for reasoning the high-level relations between image and text. The existing multimodal knowledge enhanced deep learning methods aim to incorporate multimodal knowledge into the networks, which have been utilized in visual question answering (Yang Ding and Wu 2022), and image-text retrieval (Fudong Nian and Xu 2017). Ding et al. (Yang Ding and Wu 2022) propose to represent multimodal knowledge based on triplets to correlate visual objects and answers. Nian et al. (Fudong Nian and Xu 2017) propose a multi-modal knowledge representation learning method that attempts to handle knowledge from both text and visual data. Since graph convolutional network (GCN) can make use of the semantic structure of data, some researchers have applied it to cross-modal hashing. Graph convolutional network hashing (GCH) (R. Xu and Liu 2019) introduces a GCN to bridge the modality

*Corresponding author.
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

gap and improve cross-modal retrieval. Aggregation-based graph convolutional hashing (AGCH) (P.-F. Zhang and Xu 2022) uses a GCN to mine the semantic structure of data and performs cross-modal fusion. Semi-supervised cross-modal graph convolutional network hashing (CMGCNH) (J. Duan and Huang 2020) uses an asymmetric GCN to learn modal-specific features and then generates a unified cross-modal hash representation. However, these methods neglect the implicit fine-grained multimodal knowledge relations between the image and text. When the image contains information that is not directly described in the text, the implicit multimodal knowledge relations can help to connect the image and text in the higher-level semantic space.

Recently, self-supervised learning, aiming to find supervised signals from the data itself, becomes a promising solution for the conditions without explicit labels. Contrastive learning, as one typical technique of self-supervised learning, has attracted wide attentions. Despite the wide use of contrastive learning in computer vision (K. He and Girshick 2020; T. Chen and Hinton 2020) and natural language processing (Z. Lan and Soricut 2018), little effort has been made on fine-grained cross-modal knowledge graph contrastive hash learning for cross-modal search.

In this paper, we study the problem of self-supervised learning on cross-modal hashing and propose a novel multi-grained multi-modal knowledge graph contrastive hashing for cross-modal search. In order to capture implicit fine-grained cross-modal semantic associations, a multi-modal knowledge graph is constructed, which represents the implicit multimodal knowledge relations between the image and text as inter-modal and intra-modal semantic associations. A cross-modal graph contrastive attention network is proposed to reason on the multi-modal knowledge graph to learn implicit fine-grained inter-modal and intra-modal knowledge relations. Moreover, in order to further promote more accurate cross-modal semantic alignment and fusion, in addition to local fine-grained multimodal graph contrastive learning, a multi-granularity cross-modal contrastive learning mechanism is proposed, which fuses global coarse-grained and local fine-grained embeddings by multi-head attention mechanism for inter-modal and intra-modal contrastive learning, aiming at enhancing the cross-modal unified representations with stronger discriminativeness and semantic consistency preserving power.

In summary, our contributions are as follows:

1) To our best knowledge, this is the first attempt to study the self-supervised multi-grained multi-modal knowledge graph contrastive hashing for cross-modal search. It fuses the global coarse-grained and local fine-grained embeddings by multihead attention for inter-modal and intra-modal co-contrastive learning, aiming at learning the high-level and implicit cross-modal semantic associations, and enabling it to be better applied to real world applications without label supervision.

2) A cross-modal graph contrastive attention network with co-contrastive mechanism is proposed to reason on the multi-modal knowledge graph to sufficiently learn the implicit fine-grained inter-modal and intra-modal semantic relations. With the joint training of intra-modal and inter-modal contrast, the invariant and modal-specific information of different modalities can be maintained, and the learnt cross-modal unified embeddings contain richer and more comprehensive information to boost cross-modal search.

3) We conduct diverse experiments on several public datasets and the proposed CMGCH outperforms the state-of-the arts methods, which demonstrates the effectiveness of CMGCH from various aspects.

Related Work

According to whether to use semantic labels as guidance information to learn cross-modal semantic association, it is mainly divided into supervised methods and unsupervised methods. The supervised cross-media hashing methods use semantic label information to guide the cross-modal association learning process to obtain a unified hash representation. Representative methods include adversary guided asymmetric hashing (AGAH) (Wendel et al. 2019), self-supervised cross-modal adversarial hashing (SSAH)(Chao et al. 2018), et al. The unsupervised cross-modal hashing methods maximize the semantic association between different modal data by learning cross-modal correlations. Representative methods include deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing (DGCPN) (Yu et al. 2021), unsupervised generative adversarial cross-modal hashing (UGACH) (Zhang, Peng, and Yuan 2018a), unsupervised coupled cycle generation adversarial hashing method (UCH) (C, C, and L 2019), et al.

In recent years, there are many impressive works which perform well on semantic feature learning by using contrastive learning. Hinton et al. (T. Chen and Hinton 2020)propose a simple framework for contrastive learning of visual representations (SimCLR), which requires larger batch sizes to achieve superior performance. He et al. (He et al. 2020) propose a momentum contrast method for unsupervised visual representation learning (MoCo), which replaces the memory bank just mentioned with queue. Dwibedi et al. (Dwibedi et al. 2021) propose a nearest-neighbor contrastive learning method of visual representations (NNCLR), which distinguishes the examples in the queue into positive examples and negative examples using the nearest neighbor method. A parametric contrastive learning (PaCo) method (Cui et al. 2021) is proposed to optimize the distinction between positive and negative samples in queue by using labels as a guide.

Inspired by the success of contrastive learning in intra-modal tasks, some cross-modal learning tasks based on contrastive learning have been gaining popularity. CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021) are the recent proposed multi-modal pre-training models based on contrastive learning, which form a task-agnostic model by predicting which text matches which image. In ALBEF (Li et al. 2021), it applies contrastive loss to align image and text features before modelling their joint representation. TCL (Yang et al. 2022) uses intra-modal and inter-modal contrastive learning, using knowledge distillation to guide the learning process. FACLCL (Bukchin et al. 2021) encodes structured event knowledge to enhance visual-linguistic pre-training with textual event graphs for contrastive learning.

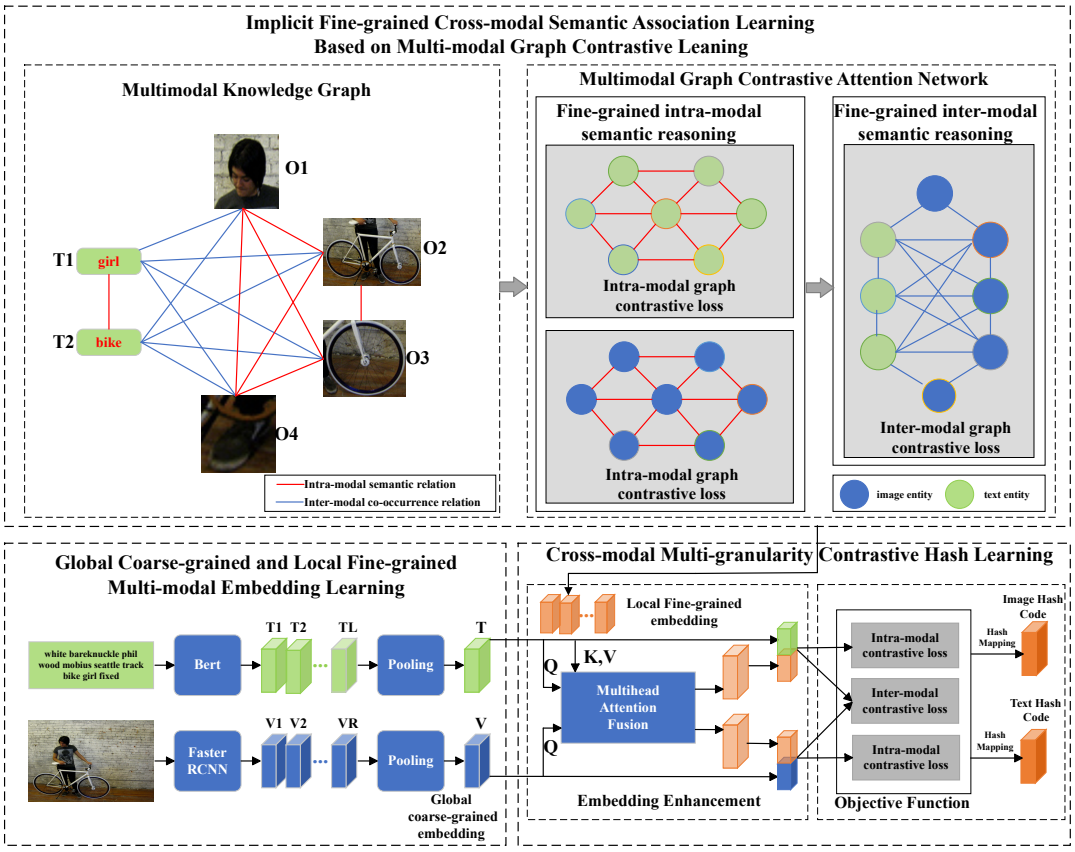

Figure 1: Framework of the proposed CMGCH method.

However, on the one hand, the above methods ignore the fine-grained cross-modal semantic associations, resulting in weak semantic discrimination of learned multimodal features; On the other hand, the above methods are all feature level semantic representation learning, without considering the semantic association and representation learning at the hash level. Thus they are not very efficient when facing large-scale cross-modal data search.

The Proposed CMGCH Method

Framework of CMGCH

In this paper, we propose a novel multi-grained multi-modal knowledge graph contrastive hashing method, aiming at obtaining a discriminative and consistent cross-modal hash representation for efficient cross-modal search. The framework is shown in Fig. 1, which includes three components: global coarse-grained and local fine-grained multi-modal embedding learning, implicit fine-grained cross-modal semantic association learning based on multi-modal graph contrastive leaning, cross-modal multi-granularity contrastive hash learning.

Problem Formulation

Given a cross-modal dataset $O = \{o_i\}_{i=1}^{N}$with N instances, where $o_i = (v_i, t_i)$ represents the image and text of the i-th instance in the dataset. The goal of the proposed CMGCH method is to jointly learn image hash mapping $f1 : H^v = f^v(F^v, \theta^v)$ and text hash mapping $f2 : H^t = f^t(F^t, \theta^t)$ by multi-grained multi-modal knowledge graph contrastive hashing, where F^v and F^t represent the feature representation of image and text respectively, θ^v and θ^t are the network parameters. Then by jointly learning two hash quantizers, $q1 : B^v = sign(H^v)$ and $q2 : B^t = sign(H^t)$, where $sign$ represent the symbolic function. We can obtain the K-bit unified binary hash representations $B^v \in \{-1, 1\}^K$ and $B^t \in \{-1, 1\}^K$ respectively, which can simultaneously preserve the inter-modal semantic similarity and the intra-modal semantic similarity.

Global Coarse-grained and Local Fine-grained Multi-modal Embedding Learning

A global coarse-grained and local fine-grained multi-modal embedding learning network is constructed, including two sub-networks of image embedding learning and text embedding learning.

In the image embedding learning network, we detect salient regions with the Bottom-Up and Top-Down attention model (Peter Anderson 2020), which selects the top R (R = 36) Regions of Interest (ROIs) with the highest class confidence scores. Then R region-level image features $V = [v_1, ..., v_R] \in \mathrm{R}^{R \times D_i}$, where D_i(D_i=2,048) is the dimension of the extracted region features. Afterwards, V is

projected into a D-dimensional space via a Fully Connected (FC) linear projection, denote the D-dimension by D.The obtained fine-grained visual region representation is denoted as $\widetilde{\mathrm{V}} = [\widetilde{v_1}, ..., \widetilde{v_R}] \in R^{R \times D}$. Moreover, we acquire the global embedding $\widetilde{v} \in R^D$ of the input image by adopting a Generalized Pooling Operator (GPO) on $\widetilde{\mathrm{V}}$.

In the text embedding learning network, Bert (Z. Lan and Soricut 2018) model is conducted to learn fine-grained word representation $T = [t_1, ..., t_L] \in R^{R_L \times D_t}$, where $t_j \in R^{D_t}$ denotes the representation of T's j-th word, L denotes the number of words, and D_t denotes the dimension of word embedding. Then T is projected into a D-dimensional space via an FC linear projection. The obtained textual word representation is denoted as $\widetilde{\mathrm{T}} = [\widetilde{t_1}, ..., \widetilde{t_L}] \in R^{R_L \times D}$. The global embedding $\widetilde{t} \in R^D$ of the input text is acquired by adopting the same pooling function GPO.

Implicit Fine-grained Cross-modal Semantic Association Learning Based on Multi-modal Graph Contrastive Leaning

Multi-modal Knowledge Graph Construction

Multimodal Knowledge Graph (MKG) is constructed by extracting visual object entities in images and word entities in texts, as well as semantic relationships between visual-visual, word-word, and visual-word to better mine the fine-grained and implicit semantic relationships within and between image and text modalities.

Visual object entities: Following MKVSE(Feng, He, and Peng 2023), we extract the visual object entities in the image as triple representation (V_i, T_i, O_i), where V_i denotes the original image, T_i denotes its corresponding text description, O_i denotes the set of visual object objects in each image, and the size of the triple tuples is N.

Text word entities: When extracting word entities, we focus on the meaning of words, phrases, and sentences, and eliminate meaningless words, such as "a", "the", etc. The most common text words $\{T_1, T_2, ...T_n\}$ corresponding to the set of image visual objects $\{O_1, O_2, ...O_n\}$ are selected as text word entities.

Inter-modal and intra-modal semantic relations: The inter-modal semantic relationships between modalities are represented by counting the number of co-occurrence relationships between visual object entities $\{O_1, O_2, ...O_n\}$ and text word entities $\{T_1, T_2, ...T_n\}$ in N triples (I_i, T_i, O_i). For the intra-modal semantic relations, the intra-modal semantic similarity is calculated by WordNet based path semantic similarity. The inter co-occurrence matrix is denoted as A^{inter}, which is calculated by the co-occurrence relationship between text words and visual objects. The intra-modal text similarity matrix is denoted as A^t with size n^t; the visual object similarity matrix is denoted as A^o with size n^o, and the feature space of the three is located as:

$$\begin{cases} A^{\text{inter}} \in R^{\left(n^t+n^o\right)\times\left(n^t+n^o\right)}, \\ A^t \in R^{n^t \times n^t}, \\ A^o \in R^{n^o \times n^o}. \end{cases} \tag{1}$$

Entity representation: following MKVSE (Feng, He, and Peng 2023), GloVe is used in MKG as an encoder for text entities and Bottom-Up and Top-Down(BUTD) attention model is used as an encoder for image entities. The most common text word nodes g_{n^i} and the most common visual object nodes b_{n^i} are selected for the construction of multimodal knowledge graph.

Cross-modal Graph Contrastive Attention Network

A cross-modal graph contrastive attention network (CGCAN) with co-contrastive learning is established to reason on the multi-modal knowledge graph to sufficiently learn the implicit fine-grained inter-modal and intra-modal relations. $G \in \mathbb{R}^{n^t \times D}$ indicates the embedding of text word entities, $B \in \mathbb{R}^{n^O \times D}$ represents the embedding of visual object entities, the whole cross-modal graph contrastive attention network can be represented as $M = CGCAN(G, B, A^o, A^t, A^{inter})$. CGCAN learns the embedding representations of entity nodes in MKG by two processes: intra-modal graph contrastive learning and inter-modal graph contrastive learning, the result obtained by CGCAN reasoning on MKG is denoted by M.

Intra-modal graph contrastive learning: based on Graph Attention Network (GAT), the visual object entity and the text word entity are reasoned separately.

The text word nodes z^t and the visual object nodes z^o are projected into a uniform contrastive space through an MLP and a hidden layer.

In N_1 text word nodes z^t, and N_2 visual object nodes z^o, the nodes with the highest similarity to themselves are selected as positive samples k^+ and the rest as negative samples k^-, τ is the temperature coefficient. The intra-modal graph contrastive learning is performed according to InfoNCE (Oord and Vinyals 2018), where the contrastive loss of text modality and visual modality in a single modal graph attention network are denoted as follows, respectiely:

$$\begin{cases} L^t_{\text{intra}} = -\frac{1}{N_1}\sum_i^{N_1} \log \frac{\exp\left(z_i^{tT}\cdot k^+/\tau\right)}{\sum_{j=1}^{N}\exp\left(z_i^{tT}\cdot k_j^-/\tau\right)}, \\ L^o_{\text{intra}} = -\frac{1}{N_2}\sum_i^{N_2} \log \frac{\exp\left(z_i^{oT}\cdot k^+/\tau\right)}{\sum_{j=1}^{N}\exp\left(z_i^{oT}\cdot k_j^-/\tau\right)}. \end{cases} \tag{2}$$

Inter-modal graph contrastive learning: inter-modal semantic inference is performed on the whole multimodal graph based on GAT. (O_i, T_i) denotes the semantic relationship between the visual object entity and the text word entity, and the one with the highest co-occurrence frequency with itself is selected as a positive sample according to the the co-occurrence relationship, and the rest are used as negative samples for inter-modal graph contrastive learning. According to InfoNCE, the inter-modal graph contrast loss function is defined as:

$$L_{inter} = -\frac{1}{N_1+N_2}\sum_i^{N_1+N_2} log\frac{exp(z_i^T\cdot k^+/\tau)}{\sum_{j=1}^{N_1+N_2} exp(z_i^T\cdot k_j^-/\tau)} \tag{3}$$

where k^+ represents a positive sample and k_j^- represents the j-th negative sample. Finally, the overall objective function L_{fgc} of implicit fine-grained cross-modal semantic association learning based on multi-modal graph contrastive leaning is as follows:

$$L_{fgc} = \beta L_{inter} + (1-\beta)(L^o_{intra} + L^t_{intra}) \tag{4}$$

where β is the hyperparameter that controls the weight of each loss function.

Cross-modal Multi-granularity Contrastive Hash Learning

We use multi-granularity feature fusion and self-supervised contrastive learning from different perspectives to guarantee the similarity of the generated hash codes in Hamming space. In the global coarse-grained feature embedding phase, we perform contrastive learning through different perspectives, specifically, our perspectives are divided into inter-modal and intra-modal.

To enhance the representational power of the model, inspired by MoCo, a dynamic dictionary with a queue and a moving-average encoder is built. The introduction of queue decouples the size of the dictionary from the size of the batch. As a result, the size of the dictionary can be much larger than the typical batch size. We construct momentum-update encoders for images and texts respectively.

We use $\mathbf{v}_g$ and $\mathbf{t}_g$ to denote the features obtained from the global embedding. At the end of the global embedding, the fine-grained representation learned based on multimodal knowledge graph is used to enhance the global embedding representation. Therefore, we fuse global coarse-grained features and local fine-grained features based on a multi-headed attention mechanism as follows.

$$\begin{cases} \mathbf{v}_a = \text{ FFN } (\text{ MultiHead } (\mathbf{v}_g, \mathbf{M})), \\ \mathbf{t}_a = \text{ FFN } (\text{ MultiHead } (\mathbf{t}_g, \mathbf{M})). \end{cases} \tag{5}$$

where FFN($\cdot$) denotes the feed forward network implemented by a multi-layer perceptron with the ReLU activation function in between. M is a fine-grained embedding representation obtained by reasoning on multimodal graph using the graph contrastive learning network. Thus, the obtained embedding representations of the image and text is as follows, λ_c is the hyperparameter:

$$\begin{cases} \mathbf{v}_f = \left[\sqrt{1-\lambda_c}\mathbf{v}_g, \sqrt{\lambda_c}\mathbf{v}_a\right], \\ \mathbf{t}_f = \left[\sqrt{1-\lambda_c}\mathbf{t}_g, \sqrt{\lambda_c}\mathbf{t}_a\right]. \end{cases} \tag{6}$$

$$\mathbf{v}_e = \text{FFN}\left(\mathbf{v}_f\right), \mathbf{t}_e = \text{FFN}\left(\mathbf{t}_f\right) \tag{7}$$

To achieve cross-modal semantic alignment and fusion, we use InfoNCE loss as the contrastive objective function, defined as follows:

$$\mathcal{L}_{(x,y)} = -\sum_i \log \frac{\exp\big((s(v_i^x, v_i^y)/\tau\big)}{\sum_{r=1}^{R} \exp\big((s(v_i^x, v_r^y)/\tau\big)} \tag{8}$$

where (x,y) represents different modal combinations, divided into two views with four different combinations, inter-modal combination (i,t), (t,i), intra-modal combination (i,i), (t,t), i and t denote image modality and text modality. $s\left(v_i^y, v_i^x\right)$ is the cosine similarity of v_i^x and v_i^y. The inter-modal and intra-modal contrastive objective functions are defined as follows:

$$\begin{aligned} \mathcal{L}_{\text{inter-loss}} &= \mathcal{L}_{(i,t)} + \mathcal{L}_{(t,i)} \\ \mathcal{L}_{\text{intra-loss}} &= \mathcal{L}_{(i,i)} + \mathcal{L}_{(t,t)} \end{aligned} \tag{9}$$

Finally, the overall cross-modal unified representation learning objective function L_{gcmh} is :

$$L_{gcmh} = \alpha\mathcal{L}_{inter-loss} + (1-\alpha)\mathcal{L}_{intra-loss} \tag{10}$$

where α is a hyperparameter to assign weights to inter-modal and intra-modal losses.

The goal of cross-modal hashing is to project different modalities into a common Hamming space. In the space, the unified codes of image and text are denoted as:$B^x = \{b_i^x\}_{i=1}^n$ for the image modality and $B^y = \{b_i^y\}_{i=1}^n$for the text modality, where $b_i^* \in \{+1, -1\}^L$, $* \in \{x, y\}$ and L is the length of hash codes. The Hamming distance is used to evaluate the similarity between image and text samples. In order to improve the retrieval efficiency of the model, we map the high-dimensional feature representation to Hamming space.

$$\mathbf{v}_q = \text{Sign}\left(\mathbf{v}_e\right), \mathbf{t}_q = \text{Sign}\left(\mathbf{t}_e\right) \tag{11}$$

where Sign($\cdot$) denotes symbolic function which is used to map feature representation to Hamming space.

Experimental Results and Analysis

Datasets and Evaluation Metrics

MSCOCO (Lin et al. 2014): This dataset totally contains 123,287 images. Each image is described with five annotated sentences with their annotations classified into 80 categories. We randomly select 5,000 image-text pairs as query set and the remaining ones are used as the retrieval set.

Flickr30k (Young et al. 2014): It contains 31,783 images from Flickr website, and each image is described by five different sentences. Following the settings in References(Tu et al. 2022), this dataset is split into 29,783 training images, 1,000 validation images, and 1,000 testing images.

Evaluation Metrics: The performance of the proposed CMGCH method is evaluated based on the Mean Average Precision (MAP) and R@K (defined as the percentage of ground truth being retrieved at top-K results).

Baselines

We use the cross-modal search task to evaluate the effectiveness of the proposed CMGCH method. We compare CMGCH with 16 state-of-the-art methods on MSCOCO dataset by MAP, including twelve unsupervised approaches (UCCH(Hu et al. 2022), DGCPN(Yu et al. 2021), LSSH(Wang et al. 2020), UKD-SS(Hu et al. 2020) and DSAH(Tu et al. 2022), CMFH(Lu et al. 2019), UCH(Li et al. 2019), FSH(Hong et al. 2017), JDSH(Zhang, Peng, and Yuan 2018b), DJSRH(Ding et al. 2016), UGACH(Zhou, Ding, and Guo 2014), CVH(Kumar and Udupa 2011)), and four supervised cross-modal hashing methods (FOMH(Tu et al. 2022), MTFH(Liu et al. 2021), DLFH(Jiang et al. 2019), DCH(Xing et al. 2017)). And also we compare with 4 state-of-the-art methods(UCCH(Hu et al. 2022), VSE++(Faghri et al. 2018), JDSH(Zhang, Peng, and Yuan 2018b), DJSRH(Ding et al. 2016)) on Flickr30k by R@K .

Comparison algorithms	*I→T (image query text)*				*T→I (text query image)*			
	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
CVH	0.503	0.504	0.471	0.425	0.506	0.508	0.476	0.429
LSSH	0.484	0.525	0.542	0.551	0.490	0.522	0.547	0.560
CMFH	0.366	0.369	0.370	0.365	0.346	0.346	0.346	0.345
FSH	0.539	0.549	0.576	0.587	0.537	0.524	0.564	0.573
DLFH	0.522	0.580	0.614	0.631	0.444	0.489	0.513	0.534
MTFH	0.399	0.293	0.295	0.395	0.335	0.374	0.300	0.334
FOMH	0.378	0.514	0.571	0.601	0.368	0.484	0.559	0.595
DCH	0.422	0.420	0.446	0.468	0.421	0.428	0.454	0.471
UGACH	0.553	0.599	0.598	0.615	0.581	0.605	0.629	0.635
DJSRH	0.501	0.563	0.595	0.615	0.494	0.569	0.604	0.622
JDSH	0.579	0.628	0.647	0.662	0.578	0.634	0.659	0.672
DGCPN	0.552	0.590	0.602	0.596	0.564	0.590	0.597	0.597
UCH	0.521	0.534	0.547	/	0.499	0.519	0.545	/
UKD-SS	0.549	0.572	0.604	/	0.549	0.576	0.625	/
DSAH	0.549	0.576	0.625	/	0.574	0.598	0.653	/
UCCH	0.605	0.645	0.655	0.665	0.610	0.655	0.666	0.677
CMGCH	**0.638**	**0.737**	**0.776**	**0.873**	**0.640**	**0.697**	**0.763**	**0.890**

Table 1: MAP Comparisons of cross-modal search with different lengths of hash codes on MSCOCO

Parameter Settings

The unified hash representation length is set to 16 bits, 32 bits, 64 bits, and 128 bits respectively. For each image, the Faster-RCNN detector provided by Bottom-Up and Top-Down (BUTD) attention model are taken to extract R (R = 36) region proposals and obtain a 2,048-dimensional feature for each region. And the BUTD model is pre-trained on ImageNet and Visual Genome datasets. For each input text, the basic version of the pre-trained Bert is leveraged to obtain the original word embeddings with dimension 768. The weight α is 0.9. The model is trained with batch size 256. The queue length of momentum encoder hyperparameter K is 8192 for Flickr30k and 65536 for MSCOCO, momentum encoder update hyperparameter m is 0.99, temperature coefficient τ is 0.07.

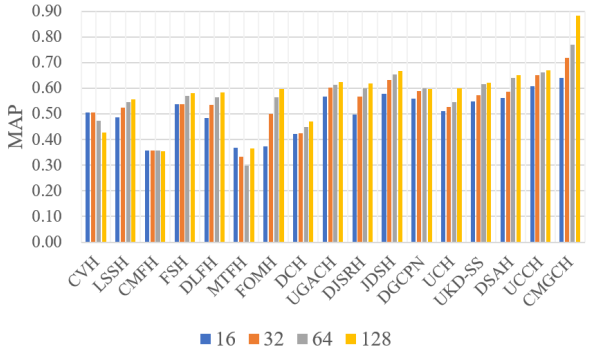

Figure 2: Comparison of average MAP for MSCOCO

Experimental Results and Analysis

Cross-modal Search Experimental Comparisons with Different Lengths of Hash Codes

In this experiment, we compare the proposed CMGCH with the state-of-the-art unsupervised and supervised cross-modal search methods in terms of MAP performance on two cross-modal hashing retrieval tasks of image query text (I→T) and text query image (T→I). We compared CMGCH with state-of-the-art algorithms with different lengths of hash codes B, including 16, 32, 64, and 128 bits. Table 1 lists the results of MAP comparison of different algorithms with different lengths of hash codes on MSCOCO. Table 2 lists the results of R@K comparison of different algorithms with different lengths of hash codes on Flickr30k. The average MAP and Recall curves of different algorithms for MSCOCO and Flickr30k with different lengths of hash codes are shown in Figure 2 and Figure 3, respectively. It can be seen that as the length of the hash codes increases, the accuracy of cross-modal search is improved. The reason lies in that longer hash codes can represent more information. However, the longer the hash code length is, the larger the search time costs are.

We can see that CMGCH obtains higher MAP performance and R@K performance on the MSCOCO and Flickr30k for the cross-modal hashing retrieval tasks respectively. Compared with previous best result, the proposed method achieves 16%, 17%, 19%, 31% increase in MAP for I→T and 10%, 16%, 16%, 32% increase in MAP for T→I on the MSCOCO separately in 16 bits, 32 bits, 64 bits and 128 bits. And it achieves an average of 159%, 191% and 178% increase in R@1 for I→T and an average of 165%, 182% and 186% increase in R@1 for T→I on the Flickr30k separately in 64 bits, 128 bits, 512 bits. The reason is that CMGCH completes inter-modal and intra-modal feature approximation by contrastive learning in the global embedding stage to extract coarse-grained feature. And fine-grained implicit features are also learned by contrastive learning on the multimodal knowledge graph, and fused with global coarse-grained features to ensure the similarity of feature vectors,

Bit	Method	$I \rightarrow T$			$T \rightarrow I$		
		R@1	R@5	R@10	R@1	R@5	R@10
64	VSE++	10.7	28.0	39.2	8.3	25.4	37.1
	DJSRH	3.6	14.4	22.1	3.4	11.6	18.5
	JDSH	10.0	28.6	39.3	8.0	23.6	34.5
	UCCH	14.5	37.6	50.8	10.9	32.3	44.0
	CMGCH	**36.7**	**67.0**	**77.1**	**28.9**	**56.2**	**67.0**
128	VSE++	11.3	31.14	42.6	9.2	27.7	40.4
	DJSRH	7.7	27.2	37.8	5.9	19.9	30.0
	JDSH	10.7	30.0	42.5	8.2	25.6	37.3
	UCCH	17.9	44.9	55.4	10.9	37.0	50.1
	CMGCH	**52.1**	**79.1**	**88.2**	**39.5**	**68.2**	**77.5**
512	VSE++	13.5	34.7	48.2	10.8	31.1	43.6
	DJSRH	17.9	43.5	56.3	13.3	36.3	48.9
	JDSH	13.6	35.6	49.4	9.8	29.1	42.6
	UCCH	22.8	48.1	61.0	16.9	41.8	54.9
	CMGCH	**63.6**	**86.7**	**92.5**	**48.4**	**77.1**	**85.3**

Table 2: Recall comparisons of cross-modal search with different lengths of hash codes on Flickr30K

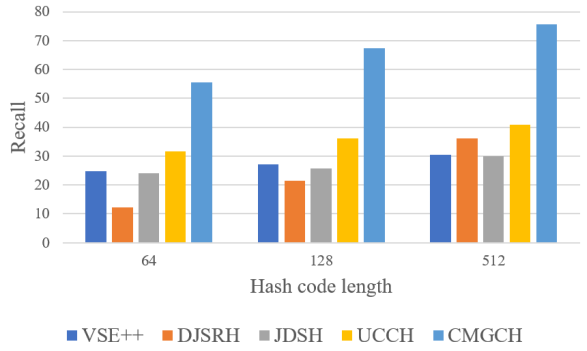

Figure 3: Comparison of average Recall for Flickr30k

thus promoting the improvement of cross-modal hashing retrieval performance.

Moreover, we compare the inference time of the proposed method with some other state-of-the-art unsupervised cross-modal hashing approaches. Comparison of inference time is shown in Table 3. The time efficiency of the proposed CMGCH is higher than that of UGACH and LSSH, and comparable to that of UCCH, while delivering superior MAP performance. Furthermore, the total parameters of the proposed CMGCH and the suboptimal UCCH are 141.2M and 257.8M, respectively. Thus we can see the proposed CMGCH achieves better overall performance.

Ablation Experiment of the Proposed CMGCH

In order to verify the effectiveness of different learning components, six types of CMGCH variants are compared and analyzed on MSCOCO. In the experiment, the length of hash code is 128 bits. CMGCH-mg: remove the multimodal knowledge graph and GAT model on the diagram, just use the global embedding; CMGCH-mgat: replacing GAT on the graph with GCN; CMGCH-oc: remove the global inter-modal contrastive learning; CMGCH-ic: remove the global intra-modal contrastive learning; CMGCH-goc: removes inter-modal contrastive learning on the graph; CMGCH-gic: remove intra-modal contrastive learning on graph.

Method	Inference Time	MAP(I→T)	MAP(T→I)
UGACH	0.254433s	0.615	0.635
LSSH	0.074445s	0.551	0.560
UCCH	**0.017874s**	0.665	0.677
CMGCH	0.018095s	**0.873**	**0.890**

Table 3: Inference time comparison on MSCOCO

CMGCH variants	MSCOCO	
	I → T	T → I
CMGCH-mg	0.823	0.837
CMGCH-mgat	0.853	0.844
CMGCH-oc	0.563	0.552
CMGCH-ic	0.714	0.792
CMGCH-goc	0.833	0.841
CMGCH-gic	0.793	0.705
CMGCH	**0.866**	**0.890**

Table 4: MAP performance of different CMGCH variants

Observing Table 4, it can be seen that, CMGCH has achieved better performance, which verifies that the integration of the six learning components can further enhance the cross-modal semantic association learning capabilities, and reduce the semantic gap between different modal.The multimodal knowledge graph enhances the ability of the model to acquire implicit relationships, and the implicit relationships on the graph are tighter through inter-modal and intra-modal contrastive learning on the graph, and the fine-grained features further enhance the model's ability to understand the implicit relationships, e.g., the implicit relationship between the image entity "apple" and the text entity "fruit" helps the model to perform cross-modal retrieval. The inter-modal and intra-modal contrastive learning in the global embedding stage maximizes the mutual information across and within modalities, and the fusion with local fine-grained features enhances the cross-modal retrieval capability of the model.

Parameter Sensitivity Analysis

(1) Performance Evaluation with Different Parameter α. From Figure 4(a), we can see that the performance of cross-modal search has an improvement as α rises, but decreases when α approaches 1. The possible reason is that MSCOCO and Flickr30k datasets we used are feature vectors obtained by Faster RCNN using the BUTD mechanism, rather than the original dataset, so data augmentation of images may degrade intra-modal contrastive learning effect. So we set α to 0.9.

(2) Performance Evaluation with Different Length K of momentum encoder queue. We set different queue lengths for the momentum encoder shown in Figure 4(b), and we find that the longer the queue, the better the model works, which indicates that the long queue contains historical feature vectors and the model is able to learn a richer feature representation. The feature vectors in queue are usually

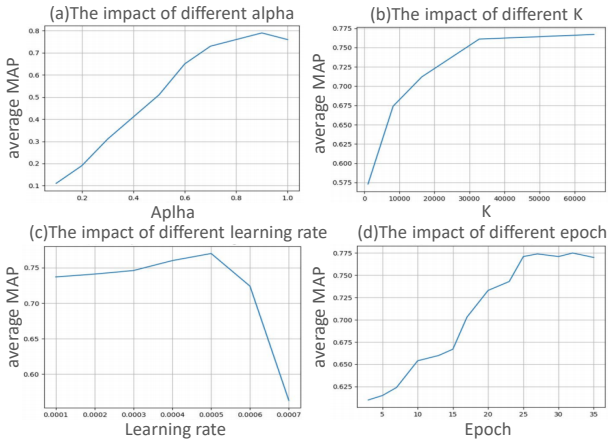

Figure 4: Performance evaluation with different parameters

from different data samples, and the relationship between these samples is more difficult to capture, so the model training yields a representation with stronger generalization ability.

(3) Performance Evaluation with Different Learning Rate. In this experiment, the impact of learning rate lr on the performance of CMGCH is analyzed. Figure 4(c) shows the impact of different lr on MAP value of cross-modal search performance when the code length is 64 bits. It can be seen that when lr=0.0005, the MAP performance achieved in the retrieval task is the highest. Therefore, we set lr=0.0005 in the experiment.

(4) Performance Evaluation with Different Epoch. In this experiment, the performance of cross-modal search under different epoch sizes is analyzed. Experiments are carried out on the MSCOCO dataset, and the experimental results are shown in Figure 4(d). It can be seen that as the epoch increases, the algorithm gradually achieves convergence. When epoch is 25 approximately, the convergence of cross-modal search algorithm tends to stabilize.

Conclusion

The proposed work is the first attempt to study the self-supervised multi-grained multi-modal knowledge graph contrastive hashing for cross-modal search, aiming at learning the high-level and implicit cross-modal semantic associations, enabling it to be better applied to real world applications without label supervision. In order to mine implicit fine-grained cross-modal semantic associations, a multi-modal knowledge graph is constructed, and a cross-modal graph contrastive attention network is proposed to reason on the multi-modal knowledge graph to sufficiently learn the implicit fine-grained inter-modal and intra-modal knowledge relations. In addition, for further promoting more accurate cross-modal semantic alignment and fusion, a multi-granularity contrastive learning mechanism is proposed, which fuses the global coarse-grained and local fine-grained embeddings by multihead attention mechanism for inter-modal and intra-modal contrastive learning.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62192784, U22B2038, 62172056, 62272058), and CAAI-Huawei MindSpore Open Fund (No.CAAIXSJLJJ-2021-007B).

References

Bukchin, G.; Schwartz, E.; Saenko, K.; Shahar, O.; Feris, R.; and Giryes, e. a., R. 2021. Fine-grained angular contrastive learning with coarse labels. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 8730–8740.

C, L.; C, D.; and L, W. 2019. Coupled cycleGAN: unsupervised hashing network for cross-modal retrieval. In *AAAI*, 176–183.

Chao, L.; Cheng, D.; Ning, L.; Wei, L.; Xinbo, G.; and Dacheng, T. 2018. Self-supervised adversarial hashing networks for cross-modal retrieval. In *CVPR*, 4242–4251.

Chen, Y.; Wang, S.; Lu, J.; Chen, Z.; Zhang, Z.; and Huang, Z. 2021. Local graph convolutional networks for cross-modal hashing. In *ACM MM*, 1921–1928.

Cui, J.; Zhong, Z.; Liu, S.; Yu, B.; and Jia, J. 2021. Parametric contrastive learning. In *ICCV*, 695–704.

Ding, G.; Guo, Y.; Zhou, J.; and Yue, G. 2016. Large-Scale Cross-Modality Search via Collective Matrix Factorization Hashing. *IEEE Transactions on Image Processing*, 25(11): 5427–5440.

Dwibedi, D.; Aytar, Y.; Tompson, J.; Sermanet, P.; and Zisserman, A. 2021. With a little help from my friends: nearest-neighbor contrastive learning of visual representations. In *ICCV*, 9568–9577.

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. *2018 British Machine Vision Conference (BMVC)*.

Feng, D.; He, X.; and Peng, Y. 2023. MKVSE: Multimodal Knowledge Enhanced Visual-Semantic Embedding for Image-Text Retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Fudong Nian, T. L., Bing-Kun Bao; and Xu, C. 2017. Multi-modal knowledge representation learning via webly-supervised relationships mining. In *Proc.the 25th ACM International Conference on Multimedia*, 411–419.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 9729–9738.

Hong, L.; Ji, R.; Wu, Y.; Huang, F.; and Zhang, B. 2017. Cross-Modality Binary Code Learning via Fusion Similarity Hashing. In *Computer Vision Pattern Recognition*.

Hu, H.; Xie, L.; Hong, R.; and Tian, Q. 2020. Creating Something From Nothing: Unsupervised Knowledge Distillation for Cross-Modal Hashing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hu, P.; Zhu, H.; Lin, J.; Peng, D.; Zhao, Y.-P.; and Peng, X. 2022. Unsupervised contrastive cross-modal hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

J. Duan, Z. W., Y. Luo; and Huang, Z. 2020. Semi-supervised cross-modal hashing with graph convolutional networks. In *Databases Theory and Applications*, 93–104.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y. T.; Parekh, Z.; and Pham, e. a., H. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proc. Int. Conf. Machine Learning*, 4904–4916.

Jiang, Q. Y.; Li, W. J.; Member; and IEEE. 2019. Discrete Latent Factor Model for Cross-Modal Hashing. *IEEE Transactions on Image Processing*, 28(7): 3490–3501.

K. He, Y. W. S. X., H. Fan; and Girshick, R. B. 2020. Momentum contrast for unsupervised visual representation learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 9726–9735.

Kumar, S.; and Udupa, R. 2011. Learning Hash Functions for Cross-View Similarity Search. In *International Joint Conference on Artificial Intelligence*.

Li, C.; Deng, C.; Wang, L.; Xie, D.; and Liu, X. 2019. Coupled CycleGAN: Unsupervised Hashing Network for Cross-Modal Retrieval. In *2019 AAAI Conference on Artificial Intelligence (AAAI)*, 176–183.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Proc. Int. Conf. Neural Inf. Process. Syst.*, 9694–9705.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *2014 European Conference on Computer Vision (ECCV)*, 740–755. Springer.

Liu, X.; Hu, Z.; Ling, H.; and Cheung, Y. M. 2021. MTFH: A Matrix Tri-Factorization Hashing Framework for Efficient Cross-Modal Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3): 964–981.

Lu, X.; Zhu, L.; Cheng, Z.; Li, J.; and Zhang, H. 2019. Flexible Online Multi-modal Hashing for Large-scale Multimedia Retrieval. In *the 27th ACM International Conference*.

Oord, Y. L., Aaron van den; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

P.-F. Zhang, Z. H., Y. Li; and Xu, X.-S. 2022. Aggregation-based graph convolutional hashing for unsupervised cross-modal retrieval. *IEEE Trans. Multimedia*, 24: 466–479.

Peng, Y.; Qi, J.; and Yuan, Y. 2019. CM-GANs: cross-modal generative adversarial networks for common representation learning. *TOMM*, 15(1): Article No.22,1–24.

Peter Anderson, C. B. D. T. M. J. S. G. L. Z., Xiaodong He. 2020. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4116–4126.

Qian, S.; Xue, D.; Zhang, H.; Fang, Q.; and Xu, C. 2021. Dual adversarial graph neural networks for multi-label cross-modal retrieval. In *AAAI*, 2440–2448.

R. Xu, J. Y. C. D., C. Li; and Liu, X. 2019. Graph convolutional network hashing for cross-modal retrieval. In *Proc.IJCAI Int. Joint Conf. Artif. Intell.*, 982–988.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; and Agarwal, e. a., S. 2021. Learning transferable visual models from natural language supervision. In *Proc. Int. Conf. Machine Learning*, 8748–8763.

T. Chen, M. N., S. Kornblith; and Hinton, G. E. 2020. A simple framework for contrastive learning of visual representations. In *Proc. ICML*, 1597–1607.

Tu, R. C.; Mao, X. L.; Ma, B.; Hu, Y.; Yan, T.; Wei, W.; and Huang, H. 2022. Deep Cross-Modal Hashing with Hashing Functions and Unified Hash Codes Jointly Learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(2): 560–572.

Wang, B.; Yang, Y.; Xu, X.; Hanjalic, A.; and Shen, H. T. 2017. Adversarial cross-Modal retrieval. In *ACM MM*, 154–162.

Wang, D.; Gao, X. .; Wang, X.; and He, L. 2019. Label consistent matrix factorization hashing for large-scale cross-modal similarity search. *TPAMI*, 41(10): 2466–2479.

Wang, X.; Zou, X.; Bakker, E. M.; and Wu, S. 2020. Self-Constraining and Attention-based Hashing Network for Bit-Scalable Cross-Modal Retrieval. *Neurocomputing*, 400.

Wei, Y.; Zhao, Y.; Lu, C.; Wei, S.; Liu, L.; Zhu, Z.; and Yan, S. 2017. Cross-modal retrieval with CNN visual features: a new baseline. *IEEE Transactions on Cybernetics*, 47(2): 449–460.

Wendel, G.; Xiaoyan, G.; Gu, J.; Bo, L.; Zhi, X.; and Weiping, W. 2019. Adversary guided asymmetric hashing for cross-modal retrieval. In *ICMR*, 159–167.

Xing; Shen; Fumin; Yang; Tao, H.; and Xuelong. 2017. Learning Discriminative Binary Codes for Large-scale Cross-modal Retrieval. *IEEE Transactions on Image Processing*.

Xu, R.; Li, C.; Yan, J.; Deng, C.; and Liu, X. 2019. Graph convolutional network hashing for cross-modal retrieval. In *IJCAI*, 982–988.

Yang, J.; Duan, J.; Tran, S.; Xu, Y.; Chanda, S.; and Chen, e. a., L. 2022. Vision-Language Pre-Training with Triple Contrastive Learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 15671–15680.

Yang Ding, B. L. Y. H. M. C., Jing Yu; and Wu, Q. 2022. MuKEA: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In *Proc.IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5089–5098.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Nlp.cs.illinois.edu*.

Yu, J.; Zhou, H.; Zhan, Y.; and Tao, D. 2021. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In *AAAI*, 4626–4634.

Z. Lan, S. G. K. G. P. S., M. Chen; and Soricut, R. 2018. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proc. ICLR*, 6077–6086.

Zhang, J.; Peng, Y.; and Yuan, M. 2018a. Unsupervised generative adversarial cross-modal hashing. In *AAAI*, 539–546.

Zhang, J.; Peng, Y.; and Yuan, M. 2018b. Unsupervised Generative Adversarial Cross-modal Hashing. In *2018 AAAI Conference on Artificial Intelligence (AAAI)*, 539–546.

Zhou, J.; Ding, G.; and Guo, Y. 2014. Latent semantic sparse hashing for cross-modal similarity search. In *International Acm Sigir Conference on Research Development in Information Retrieval*.

Zhuo, C.; Hao, D.; Yufei, W.; Tong, X.; and Enhong, C. 2020. Cross-modal video clip retrieval based on visual-text relationship alignment. *SCI CHINA INFORM SCI*, 50(6): 862–876.