SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

July 2022, Vol. 65 172104:1–172104:13
https://doi.org/10.1007/s11432-021-3367-y

# Heterogeneous memory enhanced graph reasoning network for cross-modal retrieval

Zhong JI[1*], Kexin CHEN[1], Yuqing HE[1*], Yanwei PANG[1] & Xuelong LI[2]

[1]*School of Electrical and Information Engineering, Tianjin University, Tianjin* 300072*, China;*
[2]*Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi'an* 710129*, China*

Received 11 April 2021/Revised 16 July 2021/Accepted 15 October 2021/Published online 20 June 2022

**Abstract** Cross-modal retrieval (CMR) aims to retrieve the instances of a specific modality that are relevant to a given query from another modality, which has drawn much attention because of its importance in bridging vision with language. A key to the success of CMR is to learn more discriminative and robust representations for both visual and textual instances to further reduce the heterogeneous discrepancy existing in different modalities. In this paper, we address this challenging issue by proposing a heterogeneous memory enhanced graph reasoning network, named HMGR, to connect the semantic correlations between vision and language. On the one hand, we design a novel dual-path network architecture to generate relationship enhanced global representations by employing modality-specific graph reasoning on extracted local features for each instance. In this way, the topological interdependencies of both visual and textual intra-instance local fragments are fully mined to achieve a deeper semantic understanding of the relationships between them. On the other hand, we focus on utilizing inter-instance semantic correlated knowledge to enhance the discriminability of the final learned representations, which is achieved by introducing a joint heterogeneous memory network to iteratively restore both visual and textual instance-level information. Through interacting with long-term contextual multimodal knowledge, an encouraging shared latent feature space for mitigating the heterogeneous gap across different modalities can be learned. Extensive experiments under both image-text retrieval and video-text retrieval scenarios on three benchmark datasets demonstrate the effectiveness of our proposed method.

**Keywords** cross-modal retrieval, graph reasoning, memory network, visual semantic embedding, image-text retrieval, video-text retrieval

**Citation** Ji Z, Chen K X, He Y Q, et al. Heterogeneous memory enhanced graph reasoning network for cross-modal retrieval. Sci China Inf Sci, 2022, 65(7): 172104, https://doi.org/10.1007/s11432-021-3367-y

## 1 Introduction

With the explosive growth of multimedia data from social media and the Internet, cross-modal retrieval (CMR) [1–3] has become an interesting and fascinating research topic in recent years. Such a task aims at retrieving the instances of a specific modality that are relevant to a given query from another modality. Compared with traditional unimodal retrieval systems [4,5], the key challenge in cross-modal retrieval is to bridge the inherent heterogeneous gap resided in different modalities by fully exploiting the multimodal discriminative information.

In recent decades, a surge of deep learning based approaches [6–9] have been proposed to alleviate this problem and have made encouraging progress. Early studies [10–12] attempted to map the visual data (e.g., images or videos) and textual data (e.g., sentences) into a shared latent feature space so that the representations from different modalities can be directly compared with each other. For example, Kiros et al. [12] designed a convolution neural network (CNN) encoder and a recurrent neural network (RNN) encoder to extract the features for image and text respectively. However, such a one-to-one matching scheme neglects the fine-grained details (e.g., image regions and textual words) that reflect different semantic significance when depicting the whole visual or textual instance. Thus this coarse-grained representation embedding strategy cannot fully exploit the intra-modal and inter-modal correlations to learn discriminative representations, so as to limit the retrieval performance.

* Corresponding author (email: jizhong@tju.edu.cn, heyuqing@tju.edu.cn)

 

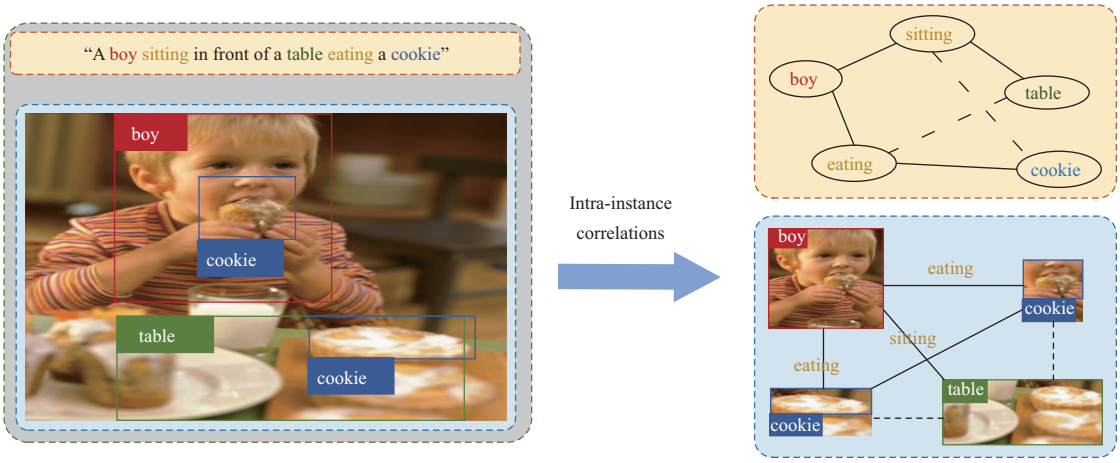


**Figure 1** (Color online) Illustration about the distinction of the semantic relevance among different words or visual objects, where the solid line represents a strong correlation between the two entities and the dotted line represents a weak correlation between them.

To mitigate this issue, a rich line of recent researches [8,13] focuses on learning local fragment features to achieve a many-to-many matching scheme. A straightforward approach is to obtain a cross-modal measurement by aggregating the similarities of all possible pairs between local fragments from both modalities. For example, Karpathy et al. [6] developed a unified deep model to infer the latent alignment between sentence segments and implicit local regions of an image that they depict. Lee et al. [8] proposed a stacked cross-modal attention mechanism to measure the image-text similarity by aligning fragments with all fragments from another modality. Such a many-to-many matching paradigm has made favorable improvement in cross-modal retrieval.

However, few of these methods pay enough attention to the robustness and discriminability of the learned features resided in individual modalities. On the one hand, the semantic relationship among intra-instance local fragments (e.g., object-object relationship in an image and word-word relationship in a sentence) always plays a key role in understanding the whole visual or textual instance. As illustrated in Figure 1, taking the caption "A boy sitting in front of a table eating a cookie" for an example, the semantic relevance of the pair "boy, eating" and the pair "table, eating" is apparently distinct. The same circumstance occurs in a given image, where the relationship among different visual semantic concepts is also different. However, few efforts have been devoted to mining such fine-grained intra-instance correspondences. On the other hand, most of the existing methods generally learn the representations of visual instance and textual instance relying on pair-wise multimodal data, and thereby the inter-instance information is always neglected during the learning process, which includes samples sharing similar semantic content from the same modality and semantic partly-correlated but labeled as unmatched entities from another modality. As a result, the learned features are not discriminative enough. Such a drawback leads to poor retrieval performance especially when encountering rarely appearing visual contents or textual contents for lacking sufficient context information.

Based on the above observations, in this paper, we propose a heterogeneous memory enhanced graph reasoning network (HMGR) for cross-modal retrieval. Firstly, we design a novel dual-path network architecture to generate context-aware global representations by employing modality-specific graph reasoning on extracted local features. In this way, the intra-instance correlations are effectively mined. Secondly, inspired by the long-term memorability of memory networks, we introduce an external heterogeneous memory to integrate the inter-instance information during the representation learning process. Specifically, at every input time step, we first read from the memory items to obtain the memory enhanced visual and textual representations by leveraging the prior inter-instance information, and then we write new information to the memory contents based on the current input multimodal knowledge. With the reading and writing operations implemented on external memory contents, the inter-instance information is leveraged to learn more discriminative features. The theoretical difference of memory-enhanced cross-modal matching structure with traditional one-to-one matching and many-to-many matching systems is illustrated in Figure 2.

The framework of our proposed HMGR approach is illustrated in Figure 3, and the main contributions

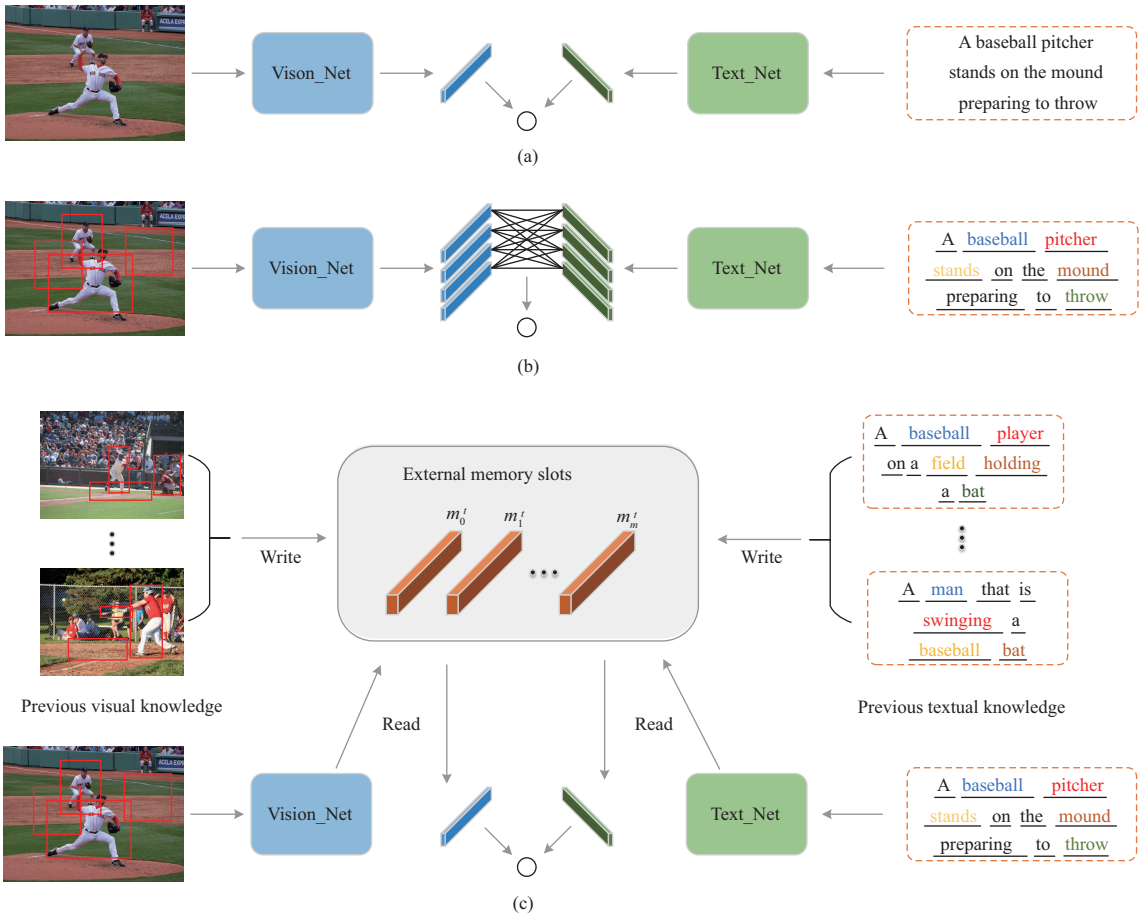


**Figure 2** (Color online) Illustration about the distinction between the memory-enhanced cross-modal matching scheme with traditional (a) one-to-one and (b) many-to-many matching methods. (c) The memory-enhanced matching model can utilize inter-instance information to enhance the discriminability of the learned representations.

of this paper are highlighted as follows.

(1) We propose a simple and interpretable dual-path graph reasoning network, which generates relationship enhanced visual and textual representations by exploiting the fine-grained semantic correlations among vision-vision elements and language-language elements.

(2) We integrate a joint heterogeneous memory network to a unified visual semantic embedding model. Through the reading and writing operations on external memory contents, inter-instance knowledge is utilized as side information to learn more discriminative features.

(3) Extensive experiments are conducted under two cross-modal retrieval scenarios including image-text retrieval and video-text retrieval. Our proposed method achieves state-of-the-art performance on three benchmark datasets, demonstrating the effectiveness of the proposed method.

## 2 Related work

Our work is related to several research topics including cross-modal retrieval, graph reasoning and memory networks. In this section, we briefly review the differences and connections between our work with some recent encouraging methods.

### 2.1 Cross-modal retrieval

According to the global or local perspective when depicting the visual and textual instance, the current mainstream approaches for cross-modal retrieval can be roughly divided into two categories: (1) one-to-one matching methods and (2) many-to-many matching methods. One-to-one matching methods attempt to directly map the image/video and sentence into a shared latent feature space for similarity

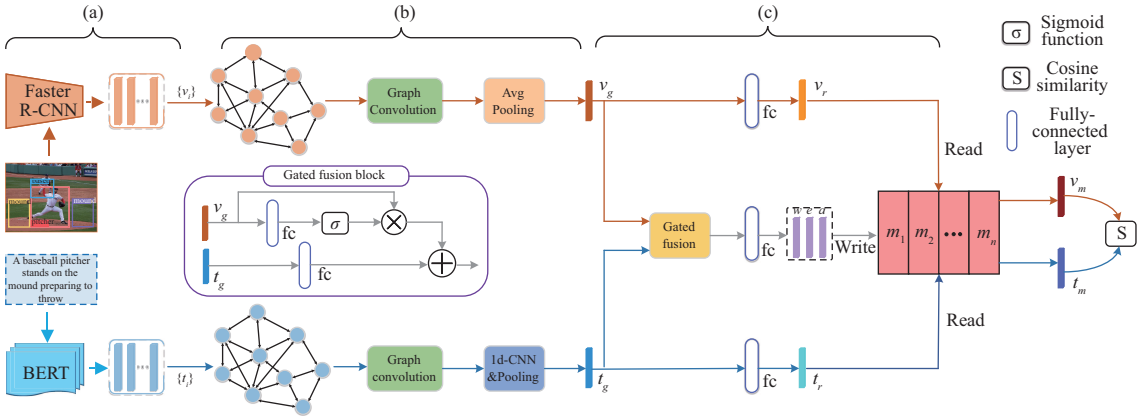


**Figure 3** (Color online) The overall framework of our heterogeneous memory enhanced graph reasoning network, which consists of three parts: (a) feature extraction to extract local features of each visual and textual instance, (b) dual-path graph reasoning module to generate global representations by capturing the fine-grained correlations between intra-instance fragments, and (c) memory-enhanced representation learning to obtain more discriminative features by exploiting the inter-instance information.

measurement. For example, Kiros et al. [12] extracted the global representations for images and captions with CNN encoder and RNN encoder respectively, and then applied hinge-based triplet ranking loss to align these heterogeneous data. Based on their study, Faghri et al. [7] refined the ranking loss function with hard negatives mining and boosted the retrieval performance significantly. Mithun et al. [14] constructed a joint representation embedding structure that encodes the video with multimodal cues such as image, motion and audio features. To measure the similarity of image/video and sentence more accurately, recently lots of many-to-many matching methods proposed to capture the fine-grained relationships between visual fragments and textual words. Karpathy et al. [6] simply aggregated all the local similarities of visual fragments and textual fragments to infer the final similarity measurement of the whole image and text. Further, Lee et al. [8] utilized faster R-CNN to extract object features of an image and proposed a stacked cross-attention network (SCAN) to align image objects with sentence words. To address the polysemous phenomenon existing in video-text retrieval scenario, Song et al. [15] learned multiple representations based on multi-head self attention mechanism for each visual and textual instance by incorporating local features with global features.

## 2.2 Graph reasoning

Graph reasoning has gradually been shown to be an effective way of relational reasoning in many computer vision tasks. Recently, various graph convolution network (GCN) based refined methods has been proposed, such as gated graph neural networks (GGNN) [16], dynamic graph neural networks (DGNN) [17] and graph attention networks (GAT) [18]. The most relevant technique to our work is GAT [18], which was proposed to refine nodes' representations by dynamically attending over neighborhoods' features. Graph reasoning has also been widely employed in many multi-modal networks. For example, Li et al. [19] applied it to generate representations that capture key objects and semantic concepts of a scene. In this study, we design a dual-path graph reasoning network to exploit the intra-instance semantic relationship among vision-vision elements and language-language elements.

## 2.3 Memory networks

Memory networks have achieved great progress and successful application in several domains of artificial intelligence. As one of the pioneering approaches, Graves et al. [20] proposed the neural Turing machines to mimic the computer memory paradigm. Sukhbaatar et al. [21] integrated the external memory with a recurrent attention model and the network is trained end-to-end. In view of multi-modal networks, Xiong et al. [22] developed a dynamic memory network for visual and textual question answering. To address video question answering, Fan et al. [23] designed a heterogeneous memory to learn global context information from appearance and motion features of videos. Ref. [24] was the most closely related research to ours, which also applied memory network in cross-modal retrieval. They focused on preserving the fine-grained multimodal clews (e.g., visual objects and textual words) directly in their memory network and updated the memory contents alternatively by them. Differently, in this paper we address to leverage

memory network to preserve the instance-level information and update our memory contents by applying a gated fusion operation on multimodal representations, which is more efficient.

## 3 Approach

Figure 3 illustrates the overall framework of our heterogeneous memory enhanced graph reasoning network (HMGR). It has three sub-modules: features extraction, dual-path graph reasoning and memory-enhanced representation learning. In the following, we will elaborate on each part in detail and introduce our deployed objective function.

### 3.1 Features extraction

**Image feature extraction.** Given an image $I$, following [8,25], we aim to detect $k$ salient objects $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_k\}$. Specifically, we utilize faster R-CNN model with ResNet-101 as backbone pretrained on Visual Genorme dataset, then feed these regional features into a fully-connected layer to obtain the final D-dimensional feature vectors $\boldsymbol{V} = \{\boldsymbol{v}_i\}_{i=1}^{k} \in \mathbb{R}^{k\times D}$:

$$\boldsymbol{v}_i = \boldsymbol{W}_o \boldsymbol{o}_i + b_o. \tag{1}$$

**Video feature extraction.** As for video, we follow [15] to regard the averagely sampled $T$ frames $\boldsymbol{F} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_T\}$ of a video as its local fragments, and then employ Resnet-152 pretrained on Imagenet as the backbone to encode these local features. We take the final output 2048-dim vector of the mean pooling layer as the frame representations. Similar to image encoder, a fully-connected layer is adopted to transform them into a D-dimensional space:

$$\boldsymbol{v}_i = \boldsymbol{W}_f \boldsymbol{f}_i + b_f. \tag{2}$$

**Sentence feature extraction.** BERT (bidirectional encoder representations from transformers) [26] is a popular language understanding model that has achieved state-of-the-art results on plenty of downstream NLP tasks. Inspired by the huge success of BERT, in this study, we utilize a BERT-base model pretrained on a large text corpus (Wikipedia) to extract 768-dimensional word-level representations for each sentence. We then feed them into one fully-connected layer to get the final D-dimensional vectors: $\boldsymbol{T} = \{\boldsymbol{t}_i\}_{i=1}^{n} \in \mathbb{R}^{n\times D}$, where $n$ is the number of words in sentence.

### 3.2 Dual-path graph reasoning module

After obtaining the local fragment representations of visual data and textual data, we aim to leverage the intra-modal semantic correlations contained in them to improve the modality-specific representations. Next, we take the visual features $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k\}$ as examples to describe our deployed graph reasoning module. We omit the same operation carried out on textual word features for simplicity.

**Graph construction.** By regarding all the visual local fragments $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k\}$ as vertexes, we construct an undirected fully-connected graph $G = (\boldsymbol{V}, E)$, as shown in Figure 3(b). And we define the adjacency matrix of the dynamic graph according to the semantic correlations of different nodes: $\tilde{A}_{i,j} = \boldsymbol{v}_i \cdot \boldsymbol{v}_j$. Considering that each element in matrix $\boldsymbol{A}$ should be non-negative, we then perform normalization along each row of matrix $\boldsymbol{A}$ as follows:

$$\boldsymbol{A}_{i,j} = \frac{\tilde{A}_{i,j}^2}{\sum_{j=1}^{k} \tilde{A}_{i,j}^2}. \tag{3}$$

Afterwards, an identity matrix $\boldsymbol{I} \in \mathbb{R}^{k\times k}$ is added to address the self-loop relationships of nodes, and then the final adjacency matrix is defined as $\hat{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}$.

**Graph reasoning network.** In graph reasoning network, the self-attention operation is firstly performed on the nodes. We compute attention coefficients that indicate the significance of node $i$ on node $j$ as

$$\hat{e}_{i,j} = \boldsymbol{W}_\varphi \boldsymbol{v}_i \cdot \boldsymbol{W}_\phi \boldsymbol{v}_j. \tag{4}$$

In order to make coefficients of different nodes can be compared directly, a softmax layer is utilized to normalize $e_{i,j}$ across all candidates of $j$:

$$e_{i,j} = \frac{\exp(\hat{e}_{i,j})}{\sum_{j \in N_i} \exp(\hat{e}_{i,j})}, \tag{5}$$

where $N_i$ is the neighborhood nodes number of node $i$, and then node $i$ is represented as the weighted sum of all other corresponding nodes' feature representation based on the normalized attention coefficients in (5):

$$\boldsymbol{v}_i^* = \sum_{j \in N_i} e_{i,j} \boldsymbol{v}_j. \tag{6}$$

Finally, we apply a multi-layer graph convolution on the newly transformed graph $G^* = (\boldsymbol{V}^*, E^*)$ to further learn reasonable embeddings with residual connection. The network's response at node $i$ takes every neighborhood nodes defined by graph correlations into consideration:

$$\boldsymbol{H}^{(l+1)} = \boldsymbol{W}_r(\sigma(\boldsymbol{D}^{-1/2}\hat{\boldsymbol{A}}\boldsymbol{D}^{-1/2}\boldsymbol{H}^{(l)}\boldsymbol{W}_l)) + \boldsymbol{H}^{(l)}, \tag{7}$$

where $\boldsymbol{H}^{(l)}$ denotes the $l$-th layer output of GCN, $\boldsymbol{H}^{(0)} = \boldsymbol{V}^*$, $\boldsymbol{D}$ is the diagonal degree matrix and $\boldsymbol{D}_{i,i} = \sum_j \hat{A}_{i,j}$, $\boldsymbol{W}_l$ is the learnable weight matrix of GCN with dimension of $D \times D$, $\boldsymbol{W}_r$ is the weight parameter for residual connection, $\sigma$ represents activation function, e.g., Relu function.

**Modality-specific global representations.** We take the last layer output of GCN $\boldsymbol{H}_v = \{\boldsymbol{h}_v^i\}_{i=1}^k \in \mathbb{R}^{k \times D}$ as the final relationship enhanced visual local fragment features. Similarly, we can get the relationship enhanced sentence words representations $\boldsymbol{H}_t = \{\boldsymbol{h}_t^i\}_{i=1}^n \in \mathbb{R}^{n \times D}$. Next, a simple average pooling layer and L2-normalization are added to aggregate these local fragments into a final global representation:

$$\boldsymbol{v}_g = \left\| \frac{1}{k} \sum_{i=1}^{k} \boldsymbol{h}_v^i \right\|_2. \tag{8}$$

As for sentence, we apply a convolutional neural network to aggregate local context information. Concretely, three different sizes of kernels ($1 \times D$, $2 \times D$, $3 \times D$) are carried out separately to capture phrase level information, and then Max_ pooling is conducted across all the words to reduce redundancy. Finally the three output vectors are concatenated and fed into a linear mapping layer to acquire the final sentence representation. This process is expressed as

$$p_{s,i} = \text{relu}(\boldsymbol{W}_s \boldsymbol{h}_t^{i:i+s-1} + b_s), \quad s \in \{1, 2, 3\}, \tag{9}$$

$$q_s = \max\{p_{s,1}, \ldots, p_{s,n}\}, \tag{10}$$

$$\boldsymbol{t}_g = \|\boldsymbol{W}_e \text{concat}(q_1, q_2, q_3) + b_e\|_2, \tag{11}$$

where $\boldsymbol{W}_e \in \mathbb{R}^{D \times 3D}$ and $b_e \in \mathbb{R}^D$ are learnable weight matrix and bias of the linear layer, $\boldsymbol{t}_g$ is the final sentence representation.

### 3.3 Memory-enhanced representation learning

Although the above graph reasoning module effectively learns the instance global representation by exploiting the intra-modal semantic correlations, it still only pays attention to the information contained in pairwise instances. In this subsection, we introduce how to take advantage of the external common knowledge to enhance the modality-specific representations, which is achieved by employing the memory network to restore both visual and textual instance-level information via multi-step iterative reasoning. Then through the content-based addressing mechanism, memory-enhanced visual and textual representations are obtained by reading from heterogeneous memory slots. The implementation details are as follows.

**Reading.** At every input time step, we review the previous learned instance-level semantic knowledge preserved in heterogeneous memory to obtain a more discriminative representation. We first feed $\boldsymbol{v}_g$ and $\boldsymbol{t}_g$ into a fully-connected layer to generate two modality-specific read heads $\boldsymbol{v}_r$ and $\boldsymbol{t}_r$. To determine how

and which memory cell to read from, firstly, we apply a content-based addressing mechanism to calculate the soft attention weights for each individual row in memory slots based on the read head:

$$w_{r,v}^i = \frac{\exp(\boldsymbol{v}_r \cdot M_{t-1}^i)}{\sum_{i=1}^N \exp(\boldsymbol{v}_r \cdot M_{t-1}^i)}, \tag{12}$$

$$w_{r,t}^i = \frac{\exp(\boldsymbol{t}_r \cdot M_{t-1}^i)}{\sum_{i=1}^N \exp(\boldsymbol{t}_r \cdot M_{t-1}^i)}. \tag{13}$$

Then the memory-enhanced representation is defined as the convex combination of every memory slot $\boldsymbol{v}_m = \sum_{i=1}^N w_{r,v}^i M_{t-1}^i$. Similarly, we get textual memory-enhanced representation $\boldsymbol{t}_m = \sum_{i=1}^N w_{r,t}^i M_{t-1}^i$.

**Writing.** Since the memory's volume (memory size) is limited, at the $t$-th input time, it is required to selectively delete the previous information and write new information to the memory. Specifically, we first implement gated fusion strategy on $\boldsymbol{v}_g$ and $\boldsymbol{t}_g$ to combine the currently input visual and textual knowledge into one vector $\boldsymbol{f}_g$, and then feed $\boldsymbol{f}_g$ into a fully-connected layer to generate 3 M-dimensional vectors $\boldsymbol{w}, \boldsymbol{e}, \boldsymbol{a}$ as the write head, erase vector and add vector, respectively. The diagrammatic view of reading and writing operation on memory network is illustrated in Figure 4.

$$\begin{aligned} &\boldsymbol{f}_g = g * \boldsymbol{v}_g + (1-g) * \boldsymbol{t}_g, \\ &g = \text{sigmoid}\left(W_g \text{cat}\left(\boldsymbol{v}_g, \boldsymbol{t}_g\right) + b_g\right), \end{aligned} \tag{14}$$

where $*$ is an element-wise product operation.

Giving an erase vector $\boldsymbol{e}$ and an add vector $\boldsymbol{a}$ emitted by controller network, we first regularize $\boldsymbol{e}$ by $\boldsymbol{e} = \text{sigmoid}(\boldsymbol{e})$ to [0, 1], whose $k$-th element means the extent of the $k$-th dimension in $M_t$ would be erased, and then the $M$-length vector $\boldsymbol{a}$ is added to each memory location modified by writing weights as follows:

$$M_t^i = M_t^{i-1}(\mathbf{1} - w_w^i \boldsymbol{e}) + w_w^i \boldsymbol{a}. \tag{15}$$

The writing weights for every memory unit are calculated as

$$w_w^i = \frac{\exp(\boldsymbol{w} \cdot M_{t-1}^i)}{\sum_{i=1}^N \exp(\boldsymbol{w} \cdot M_{t-1}^i)}. \tag{16}$$

It is worthy to claim that the memory reading and writing operation carried out on video-text retrieval task is similar with the operation performed on image-text retrieval task. The difference between the two tasks lies on the different extracted local features for images and videos. For a given image, we extract salient objects as local fragment features, while for videos, we extract key frames to capture the temporal-level information. Considering some recent studies for CMR are also built upon graph reasoning and memory networks, here we make some comparisons with these existing methods for a clear understanding of this domain. Most of the existing GCN-based methods either apply graph reasoning to capture the connections of features within individual modalities [19], or formulate the visual-semantic matching task as a scene graph matching problem relying on additional parsing toolkit [27]. Compared with them, we employ graph reasoning in conjunction with attention mechanisms on both visual and textual modality to acquire the instance-level global representations. In this way, the correlations of intra-instance fragment features are effectively mined to better align the information from two modalities. To address the long-tail distribution of multimodal data, Refs. [24, 28] also resorted to memory network for cross-modal retrieval, while they directly utilize the fragment-level features (image regions and textual words) as interaction knowledge, which lacks the global understanding of such intra-instance features and leads to inefficient cross-modal retrieval.

### 3.4 Optimization

As illustrated in Figure 3, our model outputs two pairs of embeddings for each instance pair, i.e., $(\boldsymbol{v}_g, \boldsymbol{t}_g)$ and $(\boldsymbol{v}_m, \boldsymbol{t}_m)$. It is optimized by deploying the following objective function:

$$\mathcal{L} = \max[0, \Delta - s(V,T) + s(V,\widehat{T})] + \max[0, \Delta - s(V,T) + s(\widehat{V},T)], \tag{17}$$

where $\mathcal{L}$ denotes the triplet ranking loss with hard negative mining, which is widely used in many cross-modal retrieval systems to encourage the similarity scores of positive pairs larger than negative

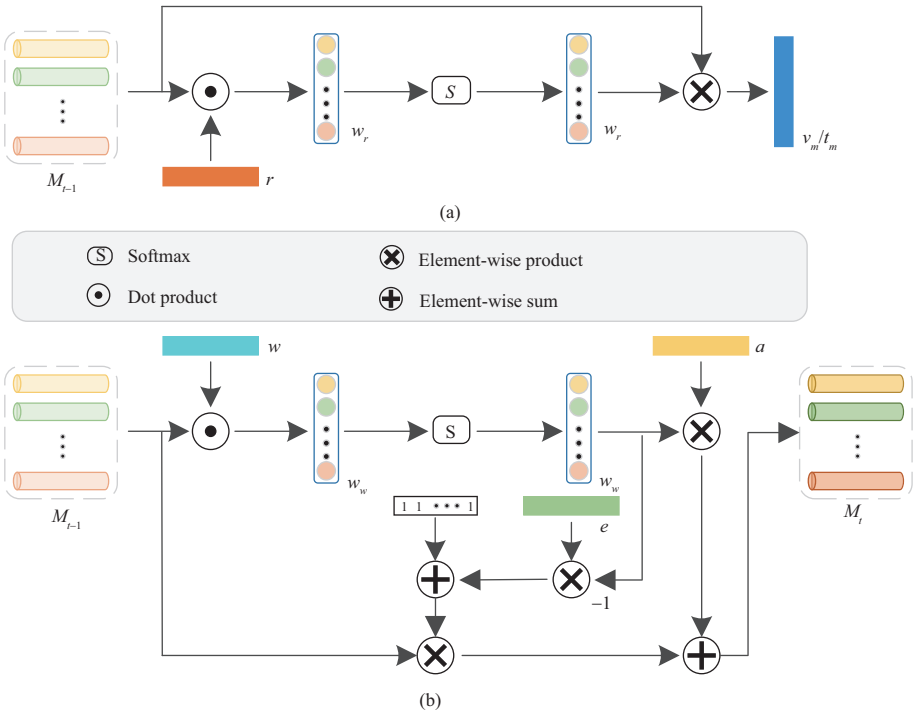


**Figure 4** (Color online) Illustration about the (a) reading and (b) writing mechanism operated on external heterogeneous memory network.

ones'. $(\widehat{V}, \widehat{T})$ denotes hard negative instances in a mini-batch. $s$ is the similarity score for visual data $V$ and sentence $T$, which is calculated by $s_g = [\cos(\boldsymbol{v}_g, \boldsymbol{t}_g) + \cos(\boldsymbol{v}_m, \boldsymbol{t}_m)]/2$, where cos denotes the cosine similarity measurement, and $\Delta$ is the margin for triplet ranking loss.

## 4 Experiments

We conduct two groups of experiments on two cross-modal retrieval tasks, which include image-text retrieval and video-text retrieval to demonstrate the effectiveness of our proposed HMGR. For image-text retrieval, two benchmark datasets, i.e., Flickr30K and MS-COCO datasets, are used to testify our model. For video-text retrieval, we use the TGIF dataset to compare our performance with state-of-the-art methods.

### 4.1 Datasets and implementation details

**Datasets.** MSCOCO contains 123287 images, and each image is annotated with 5 sentences. Following [8], we split the dataset 113287 images for training, 5000 images for validation and 5000 images for testing. We report the evaluation results by averaging over 5-folds of 1K testing images and the whole 5K testing images. Flickr30K consists of 31783 images and each image is associated with 5 text descriptions. We adopt the same split in [6], where both 1000 images for validation and testing, and the rest 29000 images for training. TGIF dataset is used to evaluate retrieval performance for video-text matching, and following [15], the dataset is split into 80K training videos, 10708 validation videos and 34101 testing videos, all the videos annotated with one caption.

**Implementation details.** We implement all our experiments in PyTorch toolkit with a single NVIDIA GeForce RTX 2080ti GPU. For image feature extraction, we select 36 salient objects for each image as local fragment features, whose dimensionality is 2048-dim. For videos, we averagely subsample 8 frames spread across each video as local features. As for the textual data, we use the BERT model pretrained by [26] to represent each word token into a 768-dim vector. Note that the weights of all feature extraction models are fixed during training procedure, and we update the network's rest trainable parameters by adopting Adam optimizer with batch size 128 for image-text retrieval and 32 for video-text retrieval. We train our model on all three datasets for 30 epochs, where the learning rate is set 0.0001

**Table 1** Image-text retrieval results on Flickr30K testing set in terms of Recall@$K$ (R@$K$)

| Method | Image-backbone | Text-backbone | Text retrieval | | | Image retrieval | | | mR |
|---|---|---|---|---|---|---|---|---|---|
| | | | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| m-CNN [11] | VGG-19 | CNN | 33.6 | 64.1 | 74.9 | 26.2 | 56.3 | 69.6 | 54.1 |
| VSE++ [7] | ResNet-152 | GRU | 52.9 | 79.1 | 87.2 | 39.6 | 69.6 | 79.5 | 68.0 |
| TIMAM [29] | ResNet-152 | Bert | 53.1 | 78.8 | 87.6 | 42.6 | 71.6 | 81.9 | 69.3 |
| SCO [30] | ResNet-152 | LSTM | 55.5 | 82.0 | 89.3 | 41.1 | 70.5 | 80.1 | 69.8 |
| SCAN [8] | Faster R-CNN | Bi-GRU | 67.4 | 90.3 | 95.8 | 48.6 | 77.7 | 85.2 | 77.5 |
| CAMP [31] | Faster R-CNN | Bi-GRU | 68.1 | 89.7 | 95.2 | 51.5 | 77.1 | 85.3 | 77.8 |
| SAEM [32] | Faster R-CNN | Bert | 69.1 | 91.0 | 95.1 | 52.4 | 81.1 | 88.1 | 79.5 |
| ACMM [24] | Faster R-CNN | Bi-GRU | **80.0** | **95.5** | **98.2** | 50.2 | 76.8 | 84.7 | 80.9 |
| MMCA [13] | Faster R-CNN | Bert | 74.2 | 92.8 | 96.4 | 54.8 | 81.4 | 87.8 | 81.2 |
| GSMN [27] | Faster R-CNN | Bi-GRU | 76.4 | 94.3 | 97.3 | 57.4 | 82.3 | 89.0 | 82.8 |
| HMGR (ours) | Faster R-CNN | Bert | 78.4 | 94.2 | 97.9 | **60.0** | **86.2** | **91.7** | **84.7** |

**Table 2** Image-text retrieval results on MSCOCO testing 1K set in terms of Recall@$K$ (R@$K$)

| Method | Image-backbone | Text-backbone | Text retrieval | | | Image retrieval | | | mR |
|---|---|---|---|---|---|---|---|---|---|
| | | | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| m-CNN [11] | VGG-19 | CNN | 42.8 | 73.1 | 84.1 | 32.6 | 68.6 | 82.8 | 64.0 |
| VSE++ [7] | ResNet-152 | GRU | 64.7 | – | 95.9 | 52.0 | – | 92.0 | – |
| SCO [30] | ResNet-152 | LSTM | 69.9 | 92.9 | 97.5 | 56.7 | 87.5 | 94.8 | 83.2 |
| SCAN [8] | Faster R-CNN | Bi-GRU | 72.7 | 94.8 | 98.4 | 58.8 | 88.4 | 94.8 | 83.6 |
| CAMP [31] | Faster R-CNN | Bi-GRU | 72.3 | 94.8 | 98.3 | 58.5 | 87.9 | 95.0 | 84.5 |
| SAEM [32] | Faster R-CNN | Bert | 71.2 | 94.1 | 97.7 | 57.8 | 88.6 | 94.9 | 84.5 |
| MMCA [13] | Faster R-CNN | Bert | 74.8 | 95.6 | 97.7 | 61.6 | 89.8 | 95.2 | 85.8 |
| ACMM [24] | Faster R-CNN | Bi-GRU | **81.9** | **98.0** | **99.3** | 58.2 | 87.3 | 93.9 | 86.4 |
| GSMN [27] | Faster R-CNN | Bi-GRU | 78.4 | 96.4 | 98.6 | 63.3 | 90.1 | 95.7 | 87.1 |
| HMGR (ours) | Faster R-CNN | Bert | 81.8 | 96.5 | 99.0 | **67.5** | **92.2** | **96.3** | **88.9** |

for the first 15 epochs and 0.00001 for another 15 epochs. The margin delta in triplet ranking loss is set to 0.2. The dimensionality of the final transformed joint embedding space $D$ is set to 1024, and the sizes of memory matrices $M$ and $N$ are set to 256 and 1024, respectively.

### 4.2 Evaluation metrics

Following previous study, we employ the recall at $K$ (R@$K$, $K$ = 1, 5, 10) as the evaluation metric, which describes the fraction of ground truth instance being retrieved at top 1, 5, 10 results. Additionally, the average recall rate "mR" is also used to testify the overall retrieval performance. The performance results of the selected approaches are all from the original papers.

### 4.3 The performance of image-text retrieval

Tables 1–3 [29–34] depict the quantitative retrieval results on Flickr30K, MS-COCO 1K and 5K testing sets, respectively. We observe that our model achieves promising performance with respect to all the evaluation metrics on the two datasets. Concretely, it brings about 1.9%, 1.8% and 2.9% gains for "mR" on Flickr30K, MS-COCO 1K and 5K testing set respectively compared with the previously leading model, i.e., GSMN, which learns the correspondence of object, relation and attributes through graph structured matching network. In addition, HMGR has an impressive advantage in image retrieval tasks compared with its competitions. For example, it outperforms the second best approach GSMN in 2.6%, 3.9%, and 2.7% absolute points on Flickr30K on R@1, R@5, and R@10, respectively. As for the text retrieval task, ACMM is the best, and this is probably because the fine-grained textual information preserved in their memory network could provide more exhaustive clews to alleviate the semantic bias for understanding sentences. Our HMGR also achieves comparable performance.

**Table 3** Image-text retrieval results on MS-COCO testing 5K set in terms of Recall@$K$ (R@$K$)

| Method | Text retrieval | | | Image retrieval | | | mR |
|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| VSE++ [7] | 41.3 | – | 81.2 | 30.3 | – | 72.4 | – |
| DPC [33] | 41.2 | 70.5 | 81.1 | 25.3 | 53.4 | 66.4 | 56.3 |
| SCAN [8] | 50.4 | 82.2 | 90.0 | 38.6 | 69.3 | 80.4 | 68.5 |
| CAMP [31] | 50.1 | 82.1 | 89.7 | 39.0 | 68.9 | 80.2 | 68.3 |
| MMCA [13] | 54.0 | 82.5 | 90.7 | 38.7 | 69.7 | 80.8 | 69.4 |
| IMRAM [34] | 53.7 | 83.2 | 91.0 | 39.7 | 69.1 | 79.8 | 69.4 |
| ACMM [24] | **63.5** | **88.0** | **93.6** | 36.7 | 65.1 | 76.7 | 70.6 |
| HMGR (ours) | 62.9 | 87.6 | 92.2 | **43.1** | **72.5** | **82.9** | **73.5** |

**Table 4** Video-text retrieval results on TGIF dataset in terms of Recall@$K$ (R@$K$)

| Method | Text retrieval | | | Video retrieval | | | mR |
|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| DeViSE [10] | 0.84 | 3.53 | 6.02 | 0.83 | 3.38 | 5.99 | 3.43 |
| VSE++ [7] | 0.42 | 1.63 | 3.60 | 0.55 | 1.89 | 3.77 | 1.98 |
| Order [35] | 0.51 | 2.09 | 3.80 | 0.48 | 2.13 | 3.86 | 2.15 |
| Corr-AE [36] | 0.89 | 3.41 | 5.61 | 0.90 | 3.48 | 5.97 | 3.38 |
| PVSE [15] | 2.32 | 7.49 | 11.94 | 2.17 | 7.76 | 12.25 | 7.32 |
| HMGR (ours) | **3.57** | **10.58** | **15.84** | **3.53** | **10.65** | **15.08** | **9.88** |

### 4.4 The performance of video-text retrieval

Table 4 [35, 36] presents the video-text retrieval results on TGIF dataset. For comparison, we cite the performance of competitive algorithms reported in PVSE [15]. From Table 4, it can be observed that our proposed method brings about 1%–3% improvement for R@1, R@5 and R@10 in both videos to text retrieval and text to video retrieval compared with PVSE [15]. Additionally, it also achieves the best performance on "mR", which validates the effectiveness of our proposed method for video-text retrieval.

### 4.5 Ablation studies and analysis

In this subsection, we perform ablation studies to evaluate the effect of each component in our proposed HMGR model. Specifically, we define the compared different model settings as follows:

**baseline.** It indicates we simply take the average over all extracted local fragments features as the final global representation and without memory-enhanced representation learning module.

**baseline+VR.** It indicates we add visual graph reasoning module only to obtain global image representation.

**baseline+TR.** It indicates we add textual graph reasoning module only to obtain global sentence representation.

**baseline+VR+TR.** It consists of both visual and textual graph reasoning modules to get global representation.

**baseline+mem.** This model introduces our memory-enhanced module and just uses the averaged local features to interact with external memory slots.

**HMGR (separate).** In this experiment, we devise two separate modality-specific external memory blocks to preserve the previous input visual and textual knowledge, respectively, which means the final image memory-enhanced representation learning only utilizes visual inter-instance information and the final sentence memory-enhanced representation learning only utilizes textual inter-instance information.

**Analysis of each component.** Table 5 shows the experimental results of different ablation settings on Flickr30K and TGIF datasets. We observe that the retrieval performance of baseline model degenerates significantly on both text retrieval and image/video retrieval because we removed our reasoning module and memory-enhancement module. The retrieval accuracy increases apparently when baseline equipped with visual or textual or both modality reasoning, which demonstrates the effectiveness of our dual-path graph reasoning module. Note that baseline+TR performs better than baseline+VR, which is probably because it is more difficult to capture the correlations between object-object elements than word-word elements. Besides, we can see the memory-enhanced representation learning is beneficial to the overall

**Table 5** Analysis of each component of our HMGR on Flickr30K and TGIF datasets

| Method | Flickr30K dataset | | | | | | | TGIF dataset | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Text retrieval | | | Image retrieval | | | mR | Text retrieval | | | Video retrieval | | | mR |
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| baseline | 45.7 | 75.2 | 84.9 | 32.0 | 62.8 | 73.9 | 62.4 | 0.63 | 3.44 | 4.76 | 0.59 | 2.93 | 5.64 | 3.00 |
| baseline+VR | 50.1 | 80.1 | 87.7 | 36.6 | 68.5 | 79.0 | 67.0 | 0.89 | 5.64 | 9.37 | 0.71 | 4.34 | 8.92 | 4.98 |
| baseline+TR | 62.8 | 87.4 | 93.0 | 45.1 | 74.9 | 82.8 | 74.3 | 1.24 | 6.72 | 9.89 | 1.16 | 5.63 | 10.46 | 5.85 |
| baseline+VR+TR | 67.9 | 89.7 | 94.9 | 51.7 | 80.2 | 87.7 | 78.7 | 1.96 | 7.55 | 11.27 | 1.94 | 6.59 | 12.37 | 6.95 |
| baseline+mem | 61.4 | 85.4 | 91.7 | 42.6 | 74.0 | 82.6 | 72.9 | 1.67 | 7.02 | 10.96 | 1.73 | 6.08 | 11.84 | 6.55 |
| HMGR (separate) | 72.3 | 91.4 | 95.4 | 53.1 | 81.4 | 88.1 | 80.3 | 2.46 | 8.31 | 12.18 | 2.32 | 7.81 | 13.06 | 7.69 |
| HMGR | **78.4** | **94.2** | **97.9** | **60.0** | **86.2** | **91.7** | **84.7** | **3.57** | **10.58** | **15.84** | **3.53** | **10.65** | **15.08** | **9.88** |

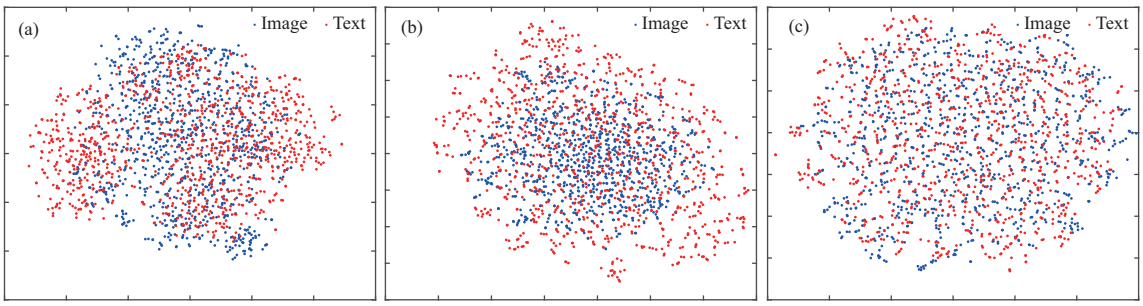


**Figure 5** (Color online) t-SNE visualization of distributions of image and sentence at different output steps on MS-COCO testing sets. (a) Original features; (b) reasoning module output; (c) memory reading output.

retrieval accuracy. Furthermore, the performance improvements achieved by HMGR (separate) model compared with baseline+VR+TR model demonstrate that it is also effective when we learn the final memory-enhanced representations through interacting with two modality-specific memories, while the better results of HMGR verify that more discriminative features can be learned by making full use of the information of the two modalities together.

### 4.6 Qualitative results

To give a more comprehensive understanding of the learning process of our heterogeneous memory enhanced graph reasoning network, we visualize the distributions of visual feature and textual feature at every output step in our model with the technique of t-SNE visualization on MS-COCO testing sets. Specifically, we convert the original features, output features from graph reasoning module and output features from memory enhanced module to 2-dimensional vectors with PCA, respectively, then plot these points with different colors in Figure 5 according to which modality they come from. As illustrated in Figure 5, we can see that after graph reasoning module, the distributions of visual data and textual data become more consistent. The comparisons between Figure 5(c) with (a) and (b) reveal that after reading from memory network, visual and textual instances achieve better alignment, which demonstrates that our memory network fully exploits the external joint information to improve the discriminability of the embedding representations, so as to alleviate the modality heterogeneous gap. Since the graph reasoning module aims at generating modality-specific global representations that include key semantic concepts in the scene. To validate this, we visualize the attention assignment of the key regions in a given image in Figure 6. Specifically, we compute the inner-product similarity between each regional features $\boldsymbol{V}^* = \{\boldsymbol{v}_1^*, \boldsymbol{v}_2^*, \ldots, \boldsymbol{v}_k^*\}$ and image global representation $\boldsymbol{v}_g$ as the attention weights and highlight the regions whose assigned weights are high. From Figure 6, we can observe that the key semantic concepts are well addressed by our model after graph reasoning, e.g., for Figure 6(a), the model attends on the horses which are reasonable to the ground-truth textual query. Figure 7 illustrates the qualitative retrieving results from image-text and video-text bidirectional retrieval on MS-COCO and TGIF testing sets, respectively. For each query item, we list the top-4 ranking instances from another modality. We can observe that our model performs well under both retrieval tasks, and even the false retrieved results have partial relation to the queries, which proves that our method can generate general and reasonable representations over similar local fragments.

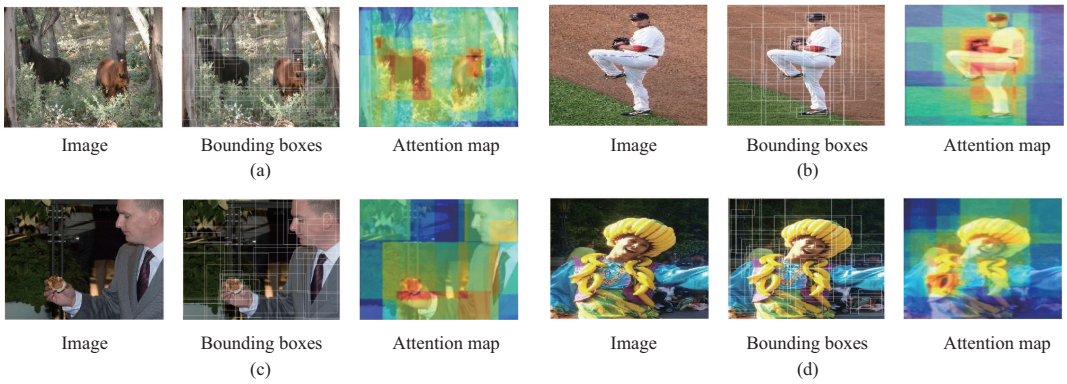


**Figure 6** (Color online) Visualization of attention weights on image activation maps. Specifically, we compute the inner-product similarity between regional features $\boldsymbol{V}^* = \{\boldsymbol{v}_1^*, \boldsymbol{v}_2^*, \ldots, \boldsymbol{v}_k^*\}$ and image global representation $\boldsymbol{v}_g$ as the attention weights. Then based on the generated attention weights, we visualize the corresponding attended visual information produced by graph reasoning module. (a) Two horses are looking towards the camera while standing in the woods; (b) a baseball player with one leg kicking up preparing to throw a ball; (c) a man in a tie is eating a hot dog; (d) a man wearing a hat and necklace made of bananas.

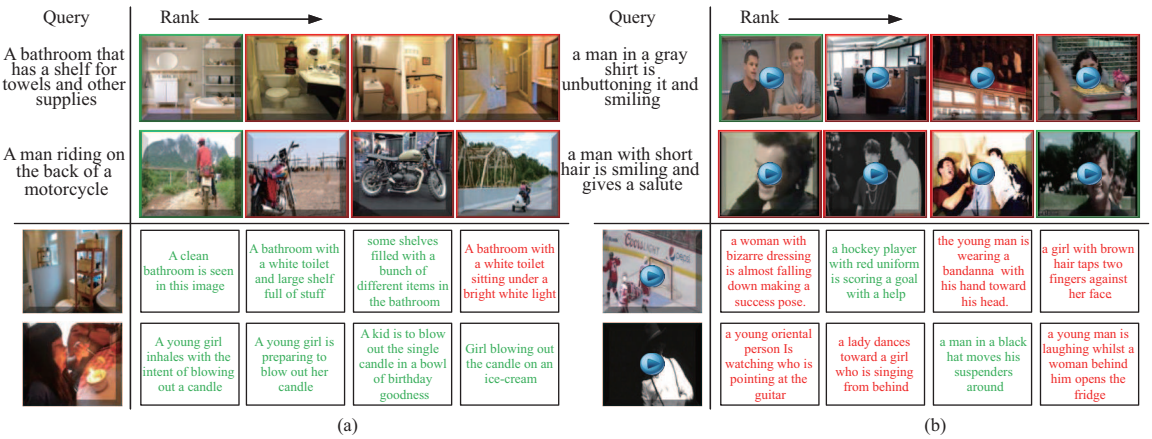


**Figure 7** (Color online) Visualization of (a) image-text and (b) video-text bidirectional retrieval on MS-COCO and TGIF datasets. For each query sample, we show top 4 ranked instances from another modality, where mismatched ones are with red boxes and matched ones are with green boxes.

## 5 Conclusion

In this paper, we have proposed a novel heterogeneous memory enhanced graph reasoning network for cross-modal retrieval. A dual-path graph reasoning module is designed to model the semantic relevance of the intra-instance fragments to learn modality-specific instance-level representations. We also introduce a heterogeneous memory network to enhance the discriminability of multimodal features by effectively exploiting the inter-instance information. Extensive experiments show that our method achieves competing results both on image-text retrieval and video-text retrieval tasks.

**Acknowledgements** This work was supported by Natural Science Foundation of Tianjin (Grant No. 19JCYBJC16000) and National Natural Science Foundation of China (Grant No. 61771329).

**References**

1 Chen Y, Huang R, Chang H, et al. Cross-modal knowledge adaptation for language-based person search. IEEE Trans Image Process, 2021, 30: 4057–4069
2 Zhang L, Ma B, Li G, et al. Generalized semi-supervised and structured subspace learning for cross-modal retrieval. IEEE Trans Multimedia, 2018, 20: 128–141
3 Ji Z, Wang H, Han J, et al. SMAN: stacked multimodal attention network for cross-modal image-text retrieval. IEEE Trans Cybern, 2022, 52: 1086–1097
4 Ji Z, Yan J T, Wang Q, et al. Triple discriminator generative adversarial network for zero-shot image classification. Sci China Inf Sci, 2021, 64: 120101
5 Wang Z H, Liu X, Lin J W, et al. Multi-attention based cross-domain beauty product image retrieval. Sci China Inf Sci, 2020, 63: 120112
6 Karpathy A, Li F-F. Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 3128–3137

7 Faghri F, Fleet D J, Kiros J R, et al. VSE++: improving visual-semantic embeddings with hard negatives. 2017. ArXiv:1707.05612
8 Lee K H, Chen X, Hua G, et al. Stacked cross attention for image-text matching. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018. 201–216
9 Hu Z, Luo Y, Lin J, et al. Multi-level visual-semantic alignments with relation-wise dual attention network for image and text matching. In: Proceedings of International Joint Conference on Artificial Intelligence, 2019. 789–795
10 Frome A, Corrado G, Shlens J, et al. DeViSE: a deep visual-semantic embedding model. In: Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013
11 Ma L, Lu Z, Shang L, et al. Multimodal convolutional neural networks for matching image and sentence. In: Proceedings of the IEEE International Conference on Computer Vision, 2015. 2623–2631
12 Kiros R, Salakhutdinov R, Zemel R S. Unifying visual-semantic embeddings with multimodal neural language models. 2014. ArXiv:1411.2539
13 Wei X, Zhang T, Li Y, et al. Multi-modality cross attention network for image and sentence matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 10941–10950
14 Mithun N C, Li J, Metze F, et al. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, 2018. 19–27
15 Song Y, Soleymani M. Polysemous visual-semantic embedding for cross-modal retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 1979–1988
16 Li Y, Tarlow D, Brockschmidt M, et al. Gated graph sequence neural networks. 2015. ArXiv:1511.05493
17 Jiang J, Wei Y, Feng Y, et al. Dynamic hypergraph neural networks. In: Proceedings of International Joint Conference on Artificial Intelligence, 2019. 2635–2641
18 Veličković P, Cucurull G, Casanova A, et al. Graph attention networks. 2017. ArXiv:1710.10903
19 Li K, Zhang Y, Li K, et al. Visual semantic reasoning for image-text matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 4654–4662
20 Graves A, Wayne G, Danihelka I. Neural turing machines. 2014. ArXiv:1410.5401
21 Sukhbaatar S, Szlam A, Weston J, et al. End-to-end memory networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, 2015. 2: 2440–2448
22 Xiong C, Merity S, Socher R. Dynamic memory networks for visual and textual question answering. In: Proceedings of International Conference on Machine Learning, 2016. 2397–2406
23 Fan C, Zhang X, Zhang S, et al. Heterogeneous memory enhanced multimodal attention model for video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 1999–2007
24 Huang Y, Wang L. ACMM: aligned cross-modal memory for few-shot image and sentence matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 5774–5783
25 Anderson P, He X, Buehler C, et al. 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, 2018. 6077–6086
26 Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. 2018. ArXiv:1810.04805
27 Liu C, Mao Z, Zhang T, et al. Graph structured network for image-text matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 10921–10930
28 Song G, Wang D, Tan X. Deep memory network for cross-modal retrieval. IEEE Trans Multimedia, 2019, 21: 1261–1275
29 Sarafianos N, Xu X, Kakadiaris I A. Adversarial representation learning for text-to-image matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 5814–5824
30 Huang Y, Wu Q, Song C, et al. Learning semantic concepts and order for image and sentence matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 6163–6171
31 Wang Z, Liu X, Li H, et al. CAMP: cross-modal adaptive message passing for text-image retrieval. In: Proceedings of Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 5764–5773
32 Wu Y, Wang S, Song G, et al. Learning fragment self-attention embeddings for image-text matching. In: Proceedings of the 27th ACM International Conference on Multimedia, 2019. 2088–2096
33 Zheng Z, Zheng L, Garrett M, et al. Dual-path convolutional image-text embeddings with instance loss. ACM Trans Multimedia Comput Commun Appl, 2020, 16: 1–23
34 Chen H, Ding G, Liu X, et al. IMRAM: iterative matching with recurrent attention memory for cross-modal image-text retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 12655–12663
35 Vendrov I, Kiros R, Fidler S, et al. Order-embeddings of images and language. 2015. ArXiv:1511.06361
36 Feng F, Wang X, Li R. Cross-modal retrieval with correspondence autoencoder. In: Proceedings of the 22nd ACM International Conference on Multimedia, 2014. 7–16