



Cross-modal multi-relationship aware reasoning for image-text matching

Jin Zhang¹ · Xiaohai He¹ · Linbo Qing¹ · Luping Liu¹ · Xiaodong Luo¹

Received: 27 July 2020 / Revised: 14 December 2020 / Accepted: 29 December 2020 /

Published online: 27 January 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC part of Springer Nature 2021

Abstract

Cross-modal image-text matching has attracted considerable interest in both computer vision and natural language processing communities. The main issue of image-text matching is to learn the compact cross-modal representations and the correlation between image and text representations. However, the image-text matching task has two major challenges. First, the current image representation methods focus on the semantic information and disregard the spatial position relations between image regions. Second, most existing methods pay little attention to improving textual representation which plays a significant role in image-text matching. To address these issues, we designed a decipherable cross-modal multi-relationship aware reasoning network (CMRN) for image-text matching. In particular, a new method is proposed to extract multi-relationship and to learn the correlations between image regions, including two kinds of visual relations: the geometric position relation and semantic interaction. In addition, images are processed as graphs, and a novel spatial relation encoder is introduced to perform reasoning on the graphs by employing a graph convolutional network (GCN) with attention mechanism. Thereafter, a contextual text encoder based on Bidirectional Encoder Representations from Transformers is adopted to learn distinctive textual representations. To verify the effectiveness of the proposed model, extensive experiments were conducted on two public datasets, namely MSCOCO and Flickr30K. The experimental results show that CMRN achieved superior performance when compared with state-of-the-art methods. On Flickr30K, the proposed method outperforms state-of-the-art methods more than 7.4% in text retrieval from image query, and 5.0% relatively in image retrieval with text query (based on Recall@1). On MSCOCO, the performance reaches 73.9% for text retrieval and 60.4% for image retrieval (based on Recall@1).

Keywords Image-text matching · Visual multi-relationship · Graph neural network · Cross-modal retrieval

✉ Xiaohai He
hxx@scu.edu.cn

¹ College of Electronics and Information Engineering, Sichuan University, Chengdu, Sichuan, 610064, China

1 Introduction

Recently, the task of cross-modal image-text matching has garnered significant attention. Furthermore, image-text matching plays a critical part in cross-modal tasks such as image-text retrieval [5, 9, 33], multi-modal summarization [3], and image caption [53]; consequently, the task has a wide range of significant practical applications. In practice, image-text matching is an information transformation task between the two modalities that bridges the gap between computer vision and natural language processing (NLP). The challenge in image-text matching rests with rich cross-modal features in a common representational space, where the visual-semantic similarity between a text and an image is measured.

Several studies have been conducted to solve the image-text matching issue and have achieved significant progress. Most studies [7, 13, 20, 33] employed standard two-branch embedding architectures to process images and texts, and subsequently map them into one embedding space. Early methods [7, 13, 50, 54] attempted to utilize the convolutional neural network (CNN) to learn image representation, which captured only the local pixel-level features of the image. Later, to obtain region-level features, the image encoder commonly adopted Faster-RCNN [39] as a visual feature extractor [20, 23, 46, 47]. For text representation, recurrent networks were usually applied to capture features [7, 9, 13]. More recently, quite a few works solved the image-text matching problem with a relation-based approach [23, 30, 38], attempting to reason about the substructures of images and texts (i.e., regions and words, respectively).

The aforementioned studies have been proven to be reliable; however, many challenges for image-text matching still exist. Most existing image-text matching methods focus only on the visual features of an image, while ignoring the relative spatial position information between regions in an image. Both the semantic and spatial position information is crucial for image-text matching. For example, as shown in Fig. 1, “in”, “on” and “under” in the sentence description need the spatial position information in the corresponding image. So, it is not only necessary to detect the objects themselves but also to understand the spatial position information and even the abstract relations that link them together. In terms of text representations, most existing methods use either word2vec [34] or randomly initialized embedding to embed words. These methods use a fixed vector for each word. Therefore, due to the semantic richness and structural diversity of sentences, they are unable to solve the problem of polysemous words. For instance, given two sentences: “A woman is standing in the bank.” and “A woman takes pictures by the bank of the river.”, “bank” has different meanings. Therefore, extracting text features that contain contextual information is essential.

The above motivated the design of a multi-relationship aware reasoning method for image-text matching, as shown in Fig. 2, to address the challenges in image representation. The method models the relationships among static objects in an image on semantic and spatial levels and integrates the connections to produce relation-aware region-level representations. These visual relationship features can expose more fine-grained contents in images, which in return offers a whole scene interpretation that can be used to facilitate matching with complex textual semantic representations. To effectively reason about spatial relations, the image is processed as a scene graph with each node denoting an object and each directed edge representing the spatial label between two objects. The GCN with an attention mechanism is adopted to enrich object representation with its partial relation-essential neighbors, dynamically. Additionally, Bidirectional Encoder Representations from Transformers (BERT) [6] has made great achievements in a multitude of NLP tasks since

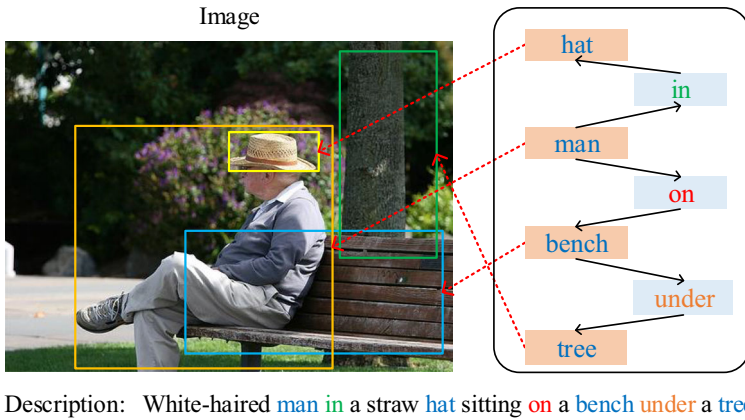


Fig. 1 An image with a description. The sentence description references to both objects (e.g., ‘hat’, ‘man’, ‘bench’ and ‘tree’) and relative position of objects (e.g., ‘in’, ‘on’ and ‘under’)

it was introduced. To address the issue of polysemous words in image-text matching, a text encoder based on BERT is introduced to learn plentiful textual representations, resulting in better image-text matching performance.

The contributions of our work can be summarized as follows.

- In this work, an effective approach is proposed to perform multi-relation aware reasoning among the detected objects in an image, which captures the spatial information between regions and semantic information for rich image representation.
- A spatial relation encoder for image-text matching is designed to capture the geometric position relations among image regions, taking advantage of the performance of GCN and the attention mechanism.
- A contextualized text encoder based on BERT is introduced, which exploits the semantic information on both sides of each word.

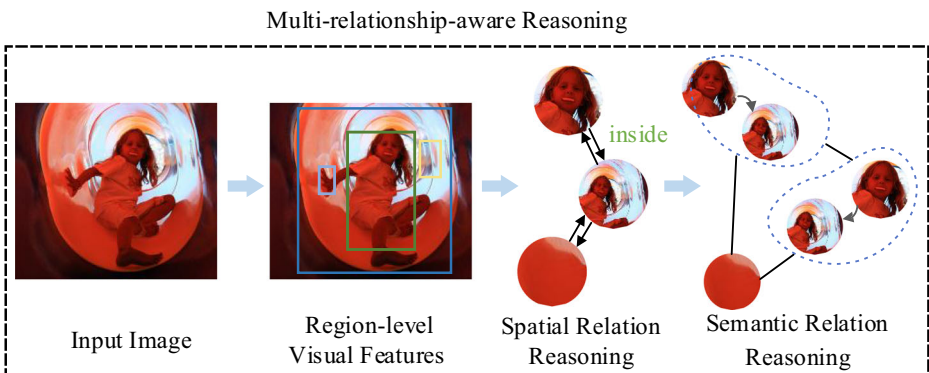


Fig. 2 Illustration of multi-relationship aware reasoning. Image representation is obtained by implementing reasoning on the region-level features. The representation includes key objects, spatial concepts, and semantic concepts of a scene consistent with the corresponding text

- Extensive experiments are conducted on two public datasets, and the experimental results demonstrate that the proposed model can achieve competitive performance than state-of-the-art methods.

This paper proceeds as follows. Section 2 reviews relevant research related to image-text matching and graph neural networks (GNN). In Section 3, the proposed model is elaborated in detail. In Section 4, we present the experimental results and analysis thereof. Section 5 contains conclusions and some ideas for further work.

2 Related work

2.1 Image-text matching

The purpose of image-text matching is to infer a similarity score between an image and a sentence. The methods of image-text matching can be divided into similarity-based and classification-based methods. The former method [7, 12, 13, 23, 27, 28, 33, 46, 51] represents images and texts as feature vectors, and subsequently calculates the similarity in the same cross-modal embedding space to determine whether there is a match. For instance, Frome et al. [8] proposed a cross-modal feature extraction model to obtain text and image representations using Skip-Gram [34] and CNN, respectively. A hinge ranking loss is employed to ensure that the distance between a mismatched image-text pair is larger than between a matched image-text pair. Lee et al. [20] developed an attention mechanism for use on texts and images to learn better representations. Faghri et al. [7] improved image-text matching by incorporating hard negatives into the loss function. Zheng et al. [57] utilized instance loss to view each multi-modal data pair as a class to learn the inter-modal correlation between images and texts. Huang et al. [14] exploited bidirectional correlations between visual regions and textual words.

Classification-based methods [12, 22, 31, 45, 48, 55] classify the input image and text features by using the neural network fitting function, which outperforms cosine similarity. The inputs from two modals of the network are usually through a fusion process, which is common in visual question answering (VQA). Ma et al. [31] proposed a method to fuse image and text features early and consider text information of different granularity. Instead of exploring a shared embedded space for each image-text pair, Wang et al. [48] applied a specific similarity function between an image and a text with rank-based tensor fusion. Our work belongs to the similarity-based method, which is to capture spatial and semantic information through graph neural networks instead of just semantic information used in most algorithms, emphasizing improving the image representations.

2.2 Graph neural networks

The graph is an optimum representation of nonlinear-structured information that models a set of objects (nodes) and their pair-wise relationships (edges). A GNN is a learning-based method running on the graph domain. Owing to its compelling performance, GNN has been incorporated into various research fields. The concept of GNN was first proposed by Scarselli et al. [40], which arising from existing neural networks to process the data represented in the graph domain. In this model, neighborhood information is propagated through the graph. However, the information characteristics of the edges in the graph are ignored, so they cannot be modeled effectively. In [21], a classic spatial domain message-passing

model based on a gated recurrent unit (GRU) [4] was updated by receiving information from neighboring nodes and sending information to neighboring nodes, named the gated graph sequence network (GGNN). Kipf et al. [17] pioneered an extensible GCN to learn on the nonlinear-structured data through convolutional operations. Instead of employing fixed aggregation weights, wherein the above GCN methods used, Velickovic et al. [43] further proposed a graph attention network (GAT) model based on GCN. In the GAT model, features are extracted from graphs by a weighted summation of the features of neighboring nodes.

2.3 Visual relational reasoning

In recent years, exploiting visual relationships, which is the basis of image understanding, has received a great deal of attention. Several studies have explored graph-based architectures for visual relationship reasoning. Visual relation reasoning has been shown to be effective in image caption [53], VQA [36, 41, 52, 56], and image-text matching [23, 29]. For example, Li et al. [23] enhanced connections between the image regions and performed reasoning utilizing GCN [40] to generate features with semantic relationship information. Hou et al. [11] utilized prior knowledge in the form of a knowledge graph to infer relationships by making use of the semantic correlation and constraints between objects without any pre-trained detectors. Yang et al. [52] constructed a scene graph with the embedded representation of visual objects and relationships and proposed a GCN module with a two-stage reasoning process on the scene graph under the guidance of the question. A recent study [49] applied a sparse graph defined by inter-/intra-class edges. The associations between each target image and its neighborhood images were captured by a language-guided graph attention mechanism. Yao et al. [53] used a semantic relation classifier trained on the Visual Genome Dataset [19] and drew connections between objects in an image to enrich the region-level features. However, most of the work focused on semantic relations, which were less interpretable than spatial relations in image understanding.

3 Approach

In this section, the proposed cross-modal multi-relationship aware reasoning network (CMRN) model is presented. First, the main architecture is summarized in Section 3.1. The three key elements of image representation of our proposed model are detailed in Section 3.2. Afterward, Section 3.3 introduces the text encoder. Finally, the loss function is elaborated in Section 3.4.

3.1 Overall architecture

The detailed architecture of the proposed CMRN model is shown in Fig. 3. Our aim is to conclude the similarity between a whole sentence representation and a global image representation in a common embedding space. The network comprised three components: an image encoder (region-level image feature extractor, spatial relation encoder, semantic relation encoder, and global reasoning), a text encoder, and a loss function. First, a Faster-RCNN [39] was used to generate a set of region features in an image. A bidirectional spatial graph with label information was constructed based on the extracted convolution and bounding box features of each region. Then, GCN with an attention mechanism was employed to focus on the spatial relationships of the local surrounding regions. To further capture the

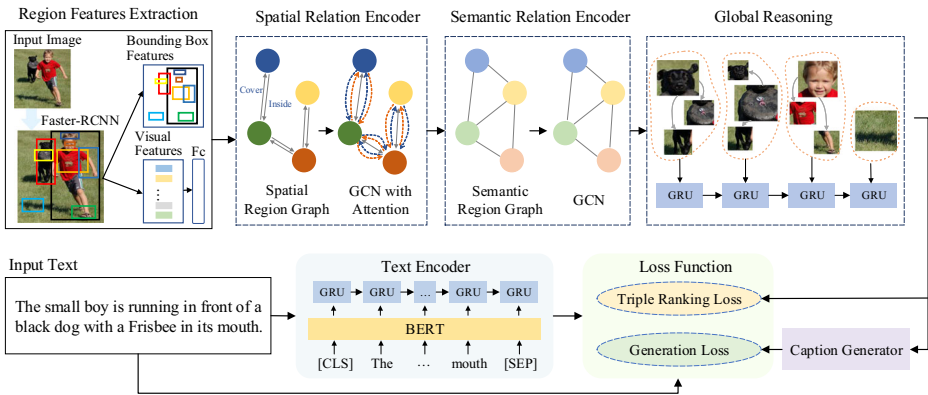


Fig. 3 Overview of the proposed CMRN model for image-text matching. Faster-RCNN was introduced to detect a set of target regions. These region-level features were fed into spatial relation encoder and semantic relation encoder to generate relation-aware features. After that, GRU was utilized to do global reasoning on the features containing spatial and semantic information, generating the representation of the whole image. Regarding text encoder, the global text representation was learned by BERT-GRU

semantic concept, we established connections between the region-level features with spatial relation perception by reasoning through GCN, generating the features with spatial and semantic relations. These visual features with relational perception were fed into the global reasoning module to obtain the final image representation. BERT-GRU was used to extract rich text features. Finally, based on image and text representations, the joint optimization of matching and caption generation was adopted to make the paired images and texts closer, and the unpaired ones more far away.

3.2 Image representation

3.2.1 Region-level visual features

Faster-RCNN [39] is an advanced object detector that has been applied in several tasks that require the detection of regions in images. Recently, many image-text matching algorithms [30, 33] started using Faster-RCNN to extract regional features. Inspired by these algorithms, we extracted region-level image features with a Faster-RCNN model in conjunction with ResNet-101 [10], which was pre-trained on the Visual Genomes dataset [19] by [1]. To denote feature representation containing plentiful semantic meaning and spatial position meaning, instance classes as well as spatial coordinates were predicted by the region-level feature encoder. Non-maximum suppression was applied to each class, with an intersection over union (IoU) threshold of 0.7. The confidence threshold was 0.3. We selected the top 36 regions of interests (ROIs) with the highest confidence scores. All of these thresholds and parameters in our experiment were set to be the same as in [1, 20, 23].

Given a region i detected in an image, features f_i with 2048 dimensions were selected after the average pooling layer. Then, f_i was converted into an h -dimensional embedding space by using a fully-connected layer. The transformation was formalized as follows:

$$v_i = W_v f_i + b_v \tag{1}$$

where W_v and b_v are a weight matrix and a bias, respectively.

Additionally, the bounding-box feature g_i was denoted as $[x_i, y_i, w_i, h_i]$, representing a 4-dimensional spatial coordinate, where (x_i, y_i) is the coordinate of the upper-left spot in the box and (w_i, h_i) is the width and height of the bounding box.

Ultimately, each image was represented as a set of objects $O = \{o_1, o_2, \dots, o_n\}$, where n represented the number of regions in the image; in our experiments $n = 36$. Each object o_k is associated with a visual feature vector $v_i \in R^{d_v}$ and a bounding-box feature vector $g_i \in R^{d_g}$ ($d_v = 2048$ and $d_g = 4$ in our experiments). The whole image can be written as $O = [v_1 \parallel g_1, v_2 \parallel g_2, \dots, v_n \parallel g_n]$.

3.2.2 Multi-relationship visual features

Based on region-level visual features, the multi-relationship visual features among static objects in an image on semantic and spatial levels were obtained by GNN.

First, a spatial relation encoder was introduced to generate valuable spatial correlations between the regions in the image. In particular, spatial information has been proven to be an essential factor in visual understanding on the VQA [24] task, which reflects the spatial form of a single object and the relationship between objects by connecting isolated regions. A bidirectional spatial graph was constructed to fully explore the spatial relationship between every two regions in the image. The spatial relationship is a triplet $\varepsilon_{i,j} = \langle \text{object}_i, \text{relation}, \text{object}_j \rangle$, which denotes the relative geometric position, given two object regions. The edges and corresponding labels were determined by the IoU, relative distance, and angle. There are 11 types of positional relationships [53], some examples are illustrated in Fig. 4. In addition, when two regions are far apart, or there are few overlapping regions, the spatial correlation between them tends to be weak, classified as a no-relation class. The spatial relation graph $\mathcal{G}_{spa} = (v, \varepsilon_{spa})$ is directional, in which the edges $\varepsilon_{i,j}$ and $\varepsilon_{j,i}$ are symmetrical.

A GCN with an attention mechanism was applied to perform reasoning on the spatial graph. The original GCN used on an undirected graph to aggregate information from neighbor nodes can be described as follows:

$$v'_i = \sigma \left(\sum_{v_j \in N(v_i)} W v_j + b \right) \quad (2)$$

where $W \in R^n \times R^n$ is a weight matrix. b is the bias term and $\sigma(\cdot)$ denotes an activation function, such as $\text{ReLU}(\cdot) = \max(0, \cdot)$. $N(v_i)$ represents the set of neighbors of v_i , including the object v_i itself.

The original GCN could not incorporate the information of directed edges and labels. Therefore, since the spatial graph is directional and includes label information, different directions and the labels of edges were transformed by transformation matrices and biases, respectively. It made the spatial encoder sensitive to both directionality and label. Additionally, instead of collecting information uniformly from all connected nodes, an attention mechanism was incorporated into GCN to focus on important edges automatically. Therefore, each node v_i was encoded by a linear weighted combination of neighbor features (after potentially applying a nonlinearity) as follows:

$$v'_i = \sigma \left(\sum_{v_j \in N(v_i)} W_{dir(i,j)} \cdot (\alpha_{ij} v_j) \right) \quad (3)$$

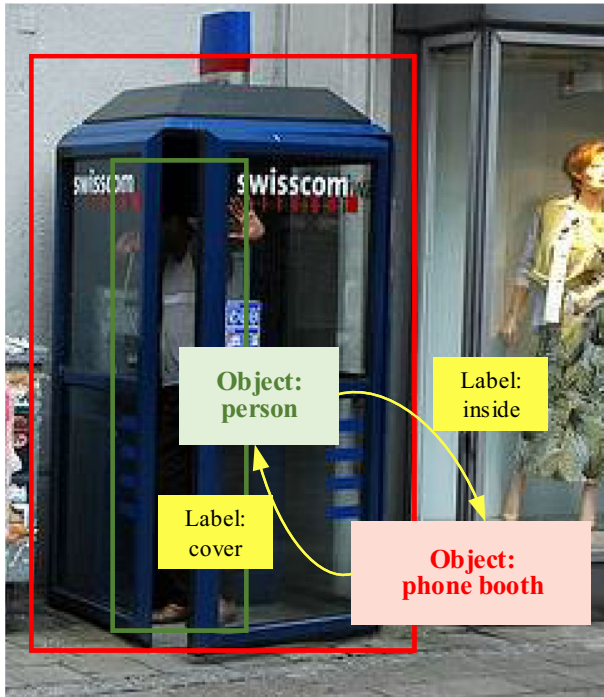


Fig. 4 An example of spatial relations. The yellow arrows and the labels in the yellow boxes indicate the orientation and class of the spatial relations, respectively

where $W_{\{i,j\}}$ is a matrix. α_{ij} is the attention coefficient and can be calculated as follows:

$$\alpha_{ij} = \frac{\exp\left(\left(W_{\varphi}v'_i\right)^T \cdot V_{dir(i,j)}v_j + b_{lab(i,j)}\right)}{\sum_{v_j \in N(v_i)} \exp\left(\left(W_{\varphi}v'_i\right)^T \cdot V_{dir(i,j)}v_j + b_{lab(i,j)}\right)} \tag{4}$$

where W_{φ} denotes the transformation matrix. $dir(i, j)$ selects the transformation matrix, which is sensitive to the directionality of each edge $(v_i - v_i, v_i - v_j, v_j - v_i)$, and $lab(i, j), lab(j, i)$ represents the label between $v_i - v_j, v_j - v_i$.

To stabilize the training process and enrich the model’s ability, multi-head attention [42] was also adopted to enhance the above GCN. The output features from each attention head were concatenated to obtain the region-level image representation with spatial relation perception as follows:

$$v'_i = \parallel_{m=1}^M \sigma \left(\sum_{v_j \in N(v_i)} W_{dir(i,j)}^m \cdot \left(\alpha_{ij}^m v_j\right) \right) \tag{5}$$

where M represents the number of independent attention mechanism.

Consequently, after encoding regions through the attention-based spatial relationship encoder as in (5), the refined region-level features were endowed with the inherent spatial relationships between objects.

After obtaining the region-level features with spatial relation perception, a GCN-based reasoning model [23] was used to derive the semantic relationship between the regions

with spatial relation perception to enrich the image representation. A fully connected graph $\mathcal{G}_{sem} = (v, \varepsilon_{sem})$ was constructed by treating each region as a node, where the edge set ε_{sem} is $n \times (n - 1)$ region pairs grouped by the detected n image regions. ε_{sem} was described as the affinity of each region pair as follows:

$$\varepsilon_{sem}(v'_i, v'_j) = (Uv'_i)^T \cdot (Vv'_j) \quad (6)$$

where U, V are embedding matrices.

A GCN with residual connection was applied in the fully connected graph. The response of each node was updated as follows:

$$v_i^* = W_r (\varepsilon_{sem} v'_i W_g) + v'_i \quad (7)$$

where W_r represents the weight of the residual connection. W_g is the weight matrix of the GCN layer.

Afterwards, the final output $V^* = \{v_1^*, v_2^*, \dots, v_k^*\}$, $v_i^* \in R^h$ was the relation-aware visual representation.

3.2.3 Image-level visual features

Rest on the region-level features with spatial and semantic relation perceptions, we fed V^* into GRU [4] in sequence for global reasoning, which selectively filtered out features to obtain the final image representation I following [23]. The update gate can be formulated as follows:

$$z_t = sigmoid(W^z v_t^* + U^z m_{t-1} + b_z) \quad (8)$$

where W^z, U^z and b_z are weights and bias. v_t^* is the current input region feature and m_{t-1} is the whole image feature at last step. *Sigmoid* is an activation function that maps any real value to the range (0, 1).

The image representation was updated as follows:

$$\tilde{m}_t = tanh(W^m v_t^* + U^m (r_t \circ m_{t-1}) + b_m) \quad (9)$$

where W^m, U^m and b_m are weights and bias. \circ represents an element-wise multiplication. r_t denotes the reset gate that computes similarly to the update gate. *Tanh* is an activation function that maps any real value to the range (-1, 1).

Then m_t was defined as the whole image representation I , where t is the length of V^* .

$$m_t = (1 - z_t) \circ m_{t-1} + z_t \circ \tilde{m}_t \quad (10)$$

where \circ is an element-wise multiplication.

3.3 Text representation

Two versions of text encoder were used, depending on the embedding methods of each word. In the first method, words were embedded into a d -dimensional vector ($d = 300$) by an embedding matrix with random initialization as follows:

$$C_i = W_i \times w_i, i \in [1, n] \quad (11)$$

where w_i is a one-hot vector in a sentence with n words $[w_1, w_2, \dots, w_n]$.

The other text encoder exploited pre-trained BERT [6] to initialize word embeddings. After the input text was pretreated, by adding two special tokens: [CLS] and [SEP], it was fed into BERT for embedding. The BERT model has 12 bidirectional transformer blocks [42] with a hidden state size of 768. The structure of the transformer combined with the

attention mechanism can fully reflect the context information of texts. Therefore, the BERT-based text encoder can achieve more accurate alignment when facing the complex image-text matching problem.

$$S = \{[\text{CLS}], w_1, w_2, \dots, w_n, [\text{SEP}]\} \quad (12)$$

$$C_i = \text{BERT}(S_i), i \in [1, n] \quad (13)$$

Subsequently, the GRU [4] was adopted to generate text representation, which was mapped into the h -dimensional embedding space.

$$T_i = \text{GRU}(C_i), i \in [1, n] \quad (14)$$

The hidden state of the last moment was the textual semantic representation of the whole sentence. Therefore, the extracted text representation expressed highly concentrated semantic information to achieve a proper alignment with the visual features of an image.

3.4 Loss function

The loss function is the summation of the cross-modal matching loss and the generation loss to learn the alignment between T and I , which is expressed as follows:

$$L = L_{\text{match}} + L_{\text{gen}} \quad (15)$$

For the generation loss portion [23], the goal was to optimize the learning of visual features. The image caption model [44] was used to encourage the learned visual representation to generate captions that were as close as possible to the ground-truth texts. The generation loss is defined as:

$$L_{\text{gen}} = - \sum_{k=1}^m \log p(e_k | e_{k-1}, V; \theta) \quad (16)$$

where m denotes the length of the generated description sequence $E = (e_1, e_2, \dots, e_m)$, and θ is the parameter from the image caption model.

For image-text matching loss portion, a triplet ranking loss based on hinge [7, 15, 20] with emphasis on hard negatives [7] was utilized to constrain the similarity score of the matched image-text pairs, which were larger than the similarity score of the unmatched ones. The matching loss is as follows:

$$L_{\text{match}} = [\alpha - S(I, T) + S(I, \hat{T})]_+ + [\alpha - S(I, T) + S(\hat{I}, T)]_+ \quad (17)$$

where α is a margin parameter and $[x]_+ \equiv \max(x, 0)$. $S(\cdot)$ represents the similarity function denoting the matching score, by calculating the inner product from the distance of the visual features I and textual features T in the public embedding space. For a mismatched image-text pair (I, T) , $\hat{I} = \text{argmax}_{m \neq I} S(m, T)$ and $\hat{T} = \text{argmax}_{d \neq T} S(I, d)$ are the hardest negatives. To improve computing efficiency, we only calculated in small batches, rather than acquire the hardest negatives throughout the entire training set.

4 Experiments and discussion

4.1 Datasets and evaluation metrics

To verify the effectiveness of the proposed method, extensive experiments were conducted on two publicly available datasets, Flickr30K [54] and MSCOCO [26].

Flickr30K: This dataset contains 31,783 images, collected from the Flickr website. Each image has five manually annotated descriptions. The data were split up as in [15] with a 28,000-image training set, 1,000-image validation set, and 1,000-image test set.

MSCOCO: We utilized the splits reported in [7], as most image-text matching algorithms do. This large-scale benchmark comprised 123,287, 1,000, and 5,000 images were employed for training, validation, and testing, respectively. Each image is associated with five descriptive sentences.

Evaluation Metrics: In this paper, $R@K$ ($K = 1, 5, 10$), short for *Recall@K*, was adopted as performance evaluation indicators of the image-text matching. It was interpreted as the percentage of correct matched pairs in the top- k retrieved results. Higher $R@K$ means better performance.

$$R@K = \frac{\text{queries retrieved the ground truth at top-}k}{\text{all queries}} \quad (18)$$

4.2 Implementation details

For CMRN, the Adam optimizer [16] was used to optimize the parameters of the model with a learning rate of 0.0002, which decays 0.1 every 10 epochs on Flickr30K, and decays 0.1 every 15 epochs on MSCOCO. The mini-batch size was 128 for the experiments. We adopted 300-dimensional word embeddings and one-layer GRU for text representations in 2048 dimensions. For BERT-based CMRN, BERT-base-cased¹ was used with a learning rate of 1e-5. The change process of the learning rate was the same as that of CMRN. The training lasted 30 epochs with a batch size of 32 on MSCOCO and Flickr30K. BERT and a layer of GRU were utilized for 2048-dimensional text representations. Furthermore, cross-modal features were aligned in the joint embedded 2048 dimension space. The α parameter of the hinge-based triplet ranking loss was empirically set to 0.2. During the training, we validated the model at each epoch on the validation set, and selected parameters with the highest total recall to solve the overfitting problem. The model was implemented based on PyTorch [37] with a GeForce GTX 1080 Ti GPU.

4.3 Comparisons with state-of-the-art methods

4.3.1 Baselines

The proposed model is compared with several SOTA models to verify its effectiveness. The following compared methods can be divided into two groups: similarity-based methods and classification-based methods.

- Similarity-based methods contain DVSA [15], Fisher Vector [18], m-CNN [31], VQA [25], VSE++ [7], DAN [35], SCO [13], GXN [9], SCAN [20], BSSAN [14], PFAN [46], MDM [32], VSRN [23], TBNN [47], CycleMatch [27], CASC [51] and TERN [33]. These baselines projected the global or local features of two modals into a common embedding space and measured their similarity.
- Classification-based methods include DSPE [45] and sm-LSTM [12]. These methods regarded image-text alignment as a classification problem, in which the answer directly represented whether the image and the text match each other.

¹Available at https://storage.googleapis.com/bert_models/2018.10.18/cased_L-12_H-768_A-12.zip

4.3.2 Results on Flickr30K

Table 1 represents the experimental results on Flickr30K for cross-modal retrieval compared with state-of-the-art (SOTA) methods. The backbones used for image and text feature extractions are listed, such as VGG, Resnet, Faster-RCNN for image, GRU, CNN for text. For VSRN [23], the ensemble results of two independently trained models were given in the original paper. However, ensemble learning can only improve the accuracy to a certain extent and require a large amount of training time. For this reason, the accuracy of a single model was considered. We used the original code and the pre-trained model provided

Table 1 Evaluation results of image-text matching on Flickr30K test set in terms of $R@K$

Method	Image backbone	Text backbone	Image-to-text retrieval			Text-to-image retrieval		
			R@1 ↑	R@5 ↑	R@10 ↑	R@1 ↑	R@5 ↑	R@10 ↑
DVSA [15] (2015)	RCNN	BRNN	22.2	48.2	61.4	15.2	37.7	50.5
Fisher Vector [18] (2015)	VGG-19	GMM+ HGLMM	35.0	62.0	73.8	25.0	52.7	66.0
m-CNN [31] (2015)	VGG-19	CNN	33.6	64.1	74.9	26.2	56.3	69.6
VQA [25] (2016)	VGG-19	GRU	33.9	62.5	74.5	24.9	52.6	64.8
DSPE [45] (2016)	VGG-19	HGLMM	40.3	68.9	79.9	29.7	60.1	72.1
sm-LSTM [12] (2017)	VGG-19	Bi- LSTM	42.5	71.9	81.5	30.2	60.4	72.3
DAN [35] (2017)	ResNet-152	LSTM	55.0	81.8	89.0	39.4	69.	79.1
VSE++ [7] (2018)	ResNet-152	GRU	52.9	79.1	87.2	39.6	69.6	79.5
SCO [13] (2018)	ResNet-152	LSTM	55.5	82.0	89.3	41.1	70.5	80.1
GXN [9] (2018)	ResNet-152	GRU	56.8	–	89.6	41.5	–	80.1
SCAN(T2I) [20] (2018)	Faster-RCNN	Bi-GRU	61.8	87.5	93.7	45.8	74.4	83.0
SCAN(I2T) [20] (2018)	Faster-RCNN	Bi-GRU	67.9	89.0	94.4	43.9	74.2	82.8
PFAN t-i [46] (2019)	Faster-RCNN	Bi-GRU	66.0	89.6	94.3	49.6	77.0	84.2
PFAN i-t [46] (2019)	Faster-RCNN	Bi-GRU	67.6	90.0	93.8	45.7	74.7	83.6
TBNN [47] (2019)	VGG-19	HGLMM	43.2	71.6	79.8	31.7	61.3	72.4
MDM [32] (2019)	VGG	CNN	44.9	75.4	84.4	34.4	67.0	77.7
BSSAN [14] (2019)	VGG/Faster-RCNN	LSTMs	44.6	74.9	84.3	33.2	62.6	72.9
CycleMatch [27] (2019)	ResNet-152	RNN	58.6	83.6	91.6	43.6	75.3	84.2
VSRN* [23] (2019)	Faster-RCNN	GRU	66.8	90.5	95.2	51.5	78.7	86.1
CASC-SCAN(T2I) [51] (2020)	Faster-RCNN	Bi-GRU	62.2	87.8	93.9	46.2	75.1	82.7
CASC-SCAN(I2T) [51] (2020)	Faster-RCNN	Bi-GRU	68.4	89.6	94.9	44.0	74.6	83.4
[4pt] CMRN(ours)	Faster-RCNN	GRU	71.8	90.8	95.1	54.1	80.1	87.0
CMRN-BERT(ours)	Faster-RCNN	BERT-GRU	70.8	91.5	95.4	55.2	81.8	88.1

The best results appear in bold

by the author to evaluate VSRN [23]. From Table 1, we can notice that the proposed model outperforms the SOTA methods on Flickr30K.

The performance of the proposed models is compared with VSRN [23], which also utilized the image-level global reasoning and the same feature extraction backbone. It is noteworthy that CMRN obtains 54.1% in R@1, 80.1% in R@5, and 87.0% in R@10 for text-to-image retrieval task, which outperforms VSRN [23] more than 5.0%, 1.7%, and 1.0%, respectively. Moreover, CMRN achieves 71.8% in R@1, 90.8% in R@5, and 95.1% in R@10 for the image-to-text retrieval task. VSRN [23] tried to enrich region-level image representation by focusing on the semantic associations between image regions. Our approach captures both semantic and spatial relationships among different regions and employs an attention mechanism to deal with the weight of the graph edge, which is the spatial label between regions. These results demonstrate that the visual representation of CMRN can be better matched with the corresponding text representation. This improvement indicates that multi-relationship aware reasoning can significantly improve the image-text matching performance. We also adopt the pre-trained BERT in the text encoder, and the results are recorded in the last line of Table 1. With the help of the pre-trained model, CMRN-BERT enjoys a further performance gain with $R@K$ ($K = 1, 5, 10$) in text-to-image retrieval, which is 55.2%, 81.8%, and 88.1%, respectively.

4.3.3 Results on MSCOCO

The evaluation results on the MSCOCO 1,000-image test set are listed in Table 2, which are obtained by averaging over 5 folds of 1,000 test images. The proposed model CMRN achieves 73.9% and 60.4% in R@1 for text retrieval and image retrieval tasks with a single trained model, which outperforms the previous best model, i.e. SCAN(T2I) [20], with more than 3.0% and 4.0%, respectively. We can find that the spatial information between regions and the semantic information is effective in improving the performance of image-text matching. Based on the results, presented in Table 2, a difference in the performance of R@1 between CMRN and CMRN-BERT is apparent. This reflects that there is an imbalance between the complexity of image representation and that of text representation. By doing an analysis, a potential reason is the characteristics of the dataset. First, the average number of objects in per image is larger on MSCOCO. Second, most objects in the image of MSCOCO are not centrally distributed, and small objects account for a large proportion.

4.4 Qualitative results

To give an instinctive sense of the experimental results of the proposed framework, the analysis of qualitative results is revealed by visualizing cases. The results compared to VSRN [23], which achieves the first-best performance among the baselines, are demonstrated in Figs. 5 and 6. In particular, two texts and two images are randomly selected for text-to-image and image-to-text retrieval, respectively, to observe the results. Figure 5 shows the qualitative results of text-to-image retrieval on Flickr30K. Given a text, the top-3 retrieved images are shown, as derived by VSRN [23], CMRN, and CMRN-BERT according to the similarity scores. The mismatched images are outlined in red boxes, and the matched images are outlined in green boxes. Figure 6 illustrates the qualitative results of the image-to-text retrieval. The top-3 ranked texts produced by VSRN [23], CMRN, and CMRN-BERT are listed from left to right, respectively. The correct results are marked in black.

The following conclusions can be drawn from the results displayed in Figs. 5 and 6.

Table 2 Evaluation results of image-text matching on MSCOCO 1000-image test set in terms of $R@K$

Method	Image	Text	Image-to-text retrieval			Text-to-image retrieval		
	backbone	backbone	R@1 ↑	R@5 ↑	R@10 ↑	R@1 ↑	R@5 ↑	R@10 ↑
DVSA [15] (2015)	RCNN	BRNN	38.4	69.9	80.5	27.4	60.2	74.8
Fisher Vector [18] (2015)	VGG-19	GMM+ HGLMM	39.4	67.9	80.9	25.1	59.8	76.6
m-CNN [31] (2016)	VGG-19	CNN	42.8	73.1	84.1	32.6	68.6	82.8
DSPE [45] (2016)	VGG-19	HGLMM	50.1	79.7	89.2	39.6	75.2	86.9
sm-LSTM [12] (2017)	VGG-19	Bi-LSTM	53.2	83.1	91.5	40.7	75.8	87.4
VSE++ [7] (2018)	ResNet-152	GRU	64.7	–	95.9	52.0	–	92.0
GXN [9] (2018)	ResNet-152	GRU	68.5	–	97.9	56.6	–	94.5
SCO [13](2018)	ResNet-152	LSTM	69.9	92.9	97.5	56.7	87.5	94.8
SCAN(T2I) [20] (2018)	Faster-RCNN	Bi-GRU	70.9	94.5	97.8	56.4	87.0	93.9
SCAN(I2T) [20] (2018)	Faster-RCNN	Bi-GRU	69.2	93.2	97.5	54.4	86.0	93.6
MDM [32] (2019)	VGG	CNN	54.7	84.1	91.9	44.6	79.6	90.5
CycleMatch [27] (2019)	ResNet-152	RNN	61.1	86.8	94.2	47.9	80.9	90.9
VSRN* [23] (2019)	Faster-RCNN	GRU	73.6	94.0	97.6	60.7	88.4	94.0
TERN [33] (2020)	Faster-RCNN	BERT	–	–	–	51.9	85.6	93.6
CMRN(ours)	Faster-RCNN	GRU	73.9	93.9	97.9	60.4	88.5	94.0
CMRN-BERT(ours)	Faster-RCNN	BERT-GRU	68.5	92.5	97.0	57.3	88.4	94.8

The best results appear in bold

- From the number and ranking order of correct retrieval shown in Figs. 5 and 6, it can be concluded that the proposed modal is superior to the SOTA methods in cross-modal bidirectional retrieval, which verifies the competitiveness of our method.
- Images retrieved by CMRN and CMRN-BERT contain more distinctive information than those retrieved by VSRN [23], as shown in Fig. 5. The above qualitative results demonstrate the necessity and rationality of performing multi-relation reasoning between image regions.
- As can be seen from Fig. 6, enhancing BERT-based text representation containing contextual information is beneficial for cross-modal retrieval, especially image retrieval.
- Through error analysis, it can be found that some incorrect retrievals from CMRN and CMRN-BERT involve relevant search information. For example, given the text query B in Fig. 5, water, sunlight, man, etc., are visible in all of images retrieved by CMRN-BERT, which is not true of the images retrieved by VSRN [23]. This indicates the effectiveness of our proposed method.

4.5 Ablation analysis

In this subsection, we discuss the extensive ablation analysis that was conducted on the Flickr30K test set to verify the effectiveness of each pivotal component by starting with a baseline model.



Fig. 5 Qualitative text-to-image retrieval examples on Flickr30K. For each text query, the top-3 retrieved images are shown. The correctly matched images are in green boxes, and the mistakenly matched ones are in red boxes

In Table 3, we report different ablated versions of the models, described as follows.

- The baseline model considered only the semantic relation reasoning to obtain image representation. The text encoder utilized random initialization embedding.



Fig. 6 Visual examples of image-to-text retrieval on Flickr30K. For the image query, the top-3 retrieved descriptions are shown. Properly matched descriptions are in black, and the mismatched ones are marked as red

Table 3 Performance on Flickr30K test set for ablation analysis

Method	Image-to-text retrieval			Text-to-image retrieval		
	R@1 ↑	R@5 ↑	R@10 ↑	R@1 ↑	R@5 ↑	R@10 ↑
baseline	66.8	90.5	95.2	51.5	78.7	86.1
CMRN	71.8	90.8	95.1	54.1	80.1	87.0
CMRN-BERT*	69.4	90.7	95.4	54.6	81.3	88.6
CMRN-BERT	70.8	91.5	95.4	55.2	81.8	88.1

The best results appear in bold

- An attention-based spatial relation encoder and semantic relation encoder were combined to obtain the CMRN model. CMRN outperformed the baseline model in R@1 by more than 7.5% and 5.0% for image-to-text and text-to-image retrieval, respectively. The above results denoted that the utilization of multi-relation aware reasoning can really learn a better image representation.
- CMRN-BERT*: Random initialization embedding was replaced with BERT as the word embedding in the text encoder while the rest of CMRN remained unchanged. The comparison showed that the pre-trained BERT model was conducive to achieving better performance, especially for text-to-image retrieval.
- CMRN-BERT: There were two image semantic extraction layers in CMRN-BERT. It outperformed CMRN and CMRN-BERT* by a relative margin for most recall metrics. The performance of image retrieval was improved from 54.6%/81.3% to 55.2%/81.8% at R@1 and R@5 when compared with CMRN-BERT*. This is attributed to the fact that the image features extracted from multiple layers are rich enough to be aligned with the text features based on BERT.

4.6 The convergence of model

In this subsection, the convergence of the proposed model is discussed. To save space, we only elaborated on the convergence and hyperparameter selection results of both image-to-text and text-to-image retrieval on the Flickr30K dataset.

The convergence of the loss function is shown in Fig. 7. It is apparent that the training loss decreases with the increasing number of train iterations and gradually converges to a stabilized value, which proves the convergence of the model.

We also recorded the values of R@1, R@5, and R@10 along with each iteration using the optimal hyperparameter settings. Figure 8 illustrates the experimental results of image-to-text and text-to-image retrieval with an increasing number of train iterations. It can be observed that the convergence process was divided into two stages.

- With more iterations, the values of R@1, R@5, and R@10 increased and the image-text matching performance improved. The model converged quickly in the first 3.0k iterations, and reached its optimal status around the 8.0k iteration. This showed the rationality of our hyperparameter settings.
- After the 8.0k iteration, the performances of R@1, R@5, and R@10 were relatively stable, even for the 10.0k and 16.0k iterations. According to [2], these phenomena also indicate the convergence of the proposed model.

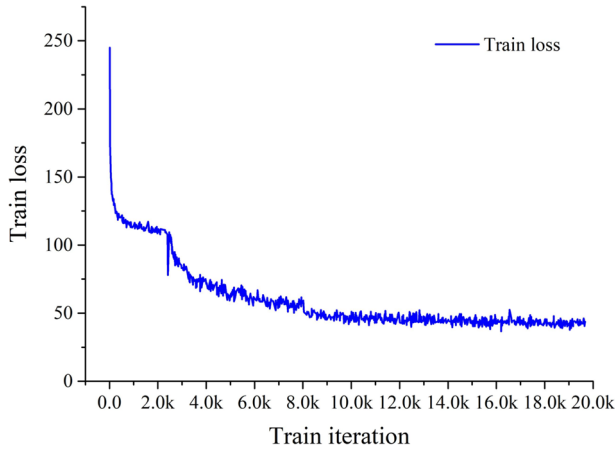


Fig. 7 The convergence of loss function on Flickr30K train data. The abscissa is the number of train iteration, and the ordinate is the train loss

4.7 Impact of neighbor number

To verify the impact of the neighbor number, extensive experiments were conducted on Flickr30K. In this study, when constructing the spatial relation encoder, the number of neighbor nodes in the spatial graph (N) was a sensitive parameter, as it determined the degree of information aggregation. With a small number of N , the network filtered out a wide range of nodes, leaving only a few highly relevant nodes. Inversely, a large number introduced some space-independent nodes. Hence, an appropriate parameter was important in the proposed network. Here, the image-text matching performance was investigated by setting $N \in [12, 32]$, with a step of 4. Since N mainly determined the image representation, it had the same effect on CMRN and CMRN-BERT. CMRN was taken as the representative to conduct the experiments. The results are reported in Fig. 9. Besides, to show the

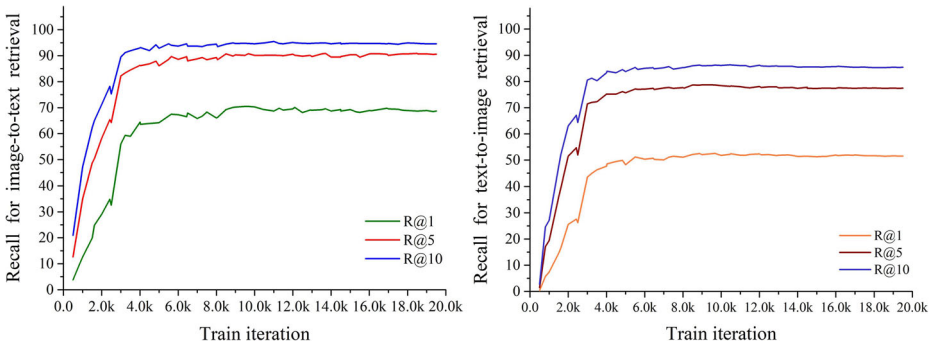


Fig. 8 The valid accuracies on Flickr30K valid data. From left to right, the R@1, R@5, and R@10 curves for image-to-text and text-to-image retrieval are presented. The abscissa is the number of train iteration, and the ordinate is the matching accuracy

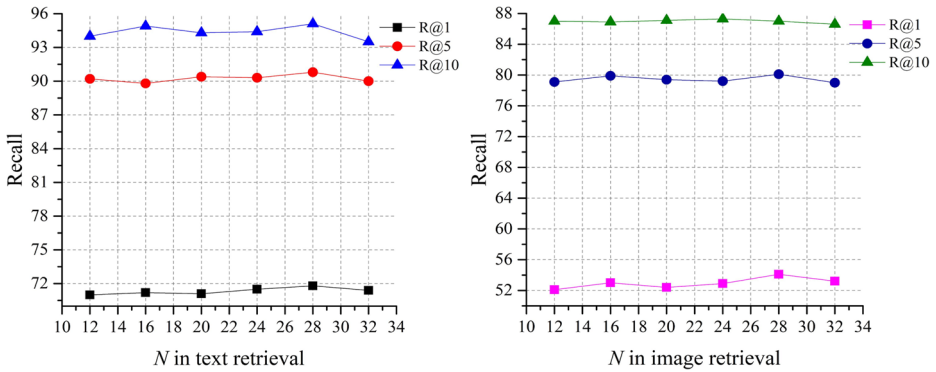


Fig. 9 Results of parameter exploration. $R@K$ ($K = 1, 5, 10$) is reported for a variety of number of neighborhood nodes (N). The left-hand plot shows quantitative results of text retrieval; the right-hand plot shows quantitative results of image retrieval

overall matching performance, the sum of Recall ($rsum$) was also computed for image and text retrieval, shown in Fig. 10.

$$rsum = \underbrace{R@1 + R@5 + R@10}_{\text{image retrieval}} + \underbrace{R@1 + R@5 + R@10}_{\text{text retrieval}} \tag{19}$$

As observed, the performance gradually increased from $N = 12$ to the highest point $N = 28$. When N exceeded 28, the performance of the proposed model decreased dramatically because of the redundant information. The results suggested that $N = 28$ was the optimal parameter for text retrieval with 71.8% in $R@1$, 90.8% in $R@5$, and 95.1% in $R@10$ and image retrieval with 54.1% in $R@1$, 80.1% in $R@5$, and 87.0% in $R@10$. It is

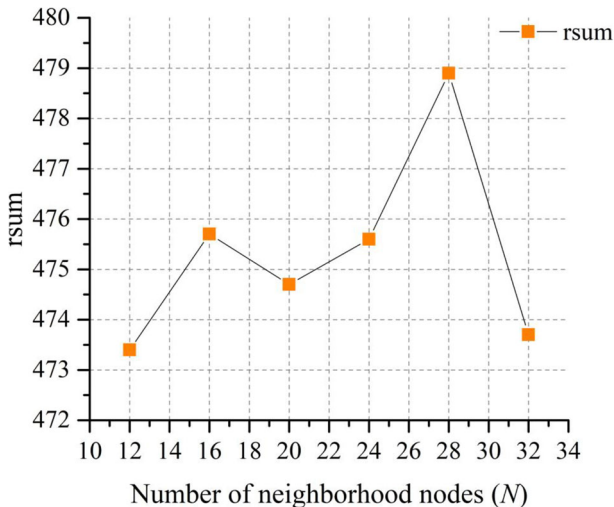


Fig. 10 Performance comparison on $rsum$ with the number of neighbor nodes (N)

Table 4 Performance comparison on Flickr30K test set for ensemble

Method	Image-to-text retrieval			Text-to-image retrieval		
	R@1 ↑	R@5 ↑	R@10 ↑	R@1 ↑	R@5 ↑	R@10 ↑
VSRN [23]	71.3	90.6	96.0	54.7	81.8	88.2
CMRN-ensemble($M = 2$)	73.5	92.1	96.6	55.4	81.3	88.3
CMRN-ensemble($M = 4$)	73.0	91.5	96.1	56.0	82.3	89.2

The best results appear in bold

worth mentioning that $rsum$ was up to 478.9. Therefore, $N = 28$ was set as our default value in all experiments.

4.8 Ensemble

To further improve cross-modal retrieval performance, the common practice of ensembling was applied to the image-text matching task [20, 23]. It is a typical post-processing operation that fuses the outputs of multiple baseline models in a certain way to reduce variance and improve results. In this work, the predicted similarity scores of two trained models were averaged and compared with VSRN [23] (two models were also used). The results are manifested in Table 4.

Meanwhile, the number of integrated models was validated on the Flickr30K test set. The most basic form of ensembling was utilized, that is, using multiple entities of the same trained model (same network architecture, same parameters, and same dataset). M sets of models were selected for the ensemble by averaging the relevance score from each model. As can be seen from Figs. 11 and 12, with the gaining number of models, the accuracy is also gradually increasing. When the number of models was greater than 4, the matching performance showed no obvious upward trend with the increase of M . When $M = 4$, the performance of the models was at its highest, and stable at 73.0%, 91.5%, and 96.1%, for R@1, R@5, and R@10, respectively, for text retrieval, and 56.0%, 82.3%, and 89.2% for R@1, R@5, and R@10, for image retrieval, respectively. The $rsum$ reached 488.1.

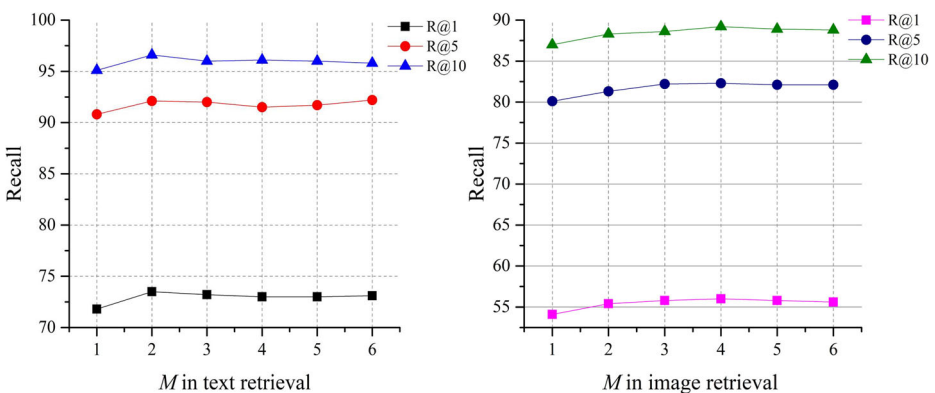


Fig. 11 Performance of the number of integrated models (M) on $R@K$ ($K = 1, 5, 10$). The left-hand plot shows the experimental results of text retrieval; the right-hand plot shows the results of image retrieval

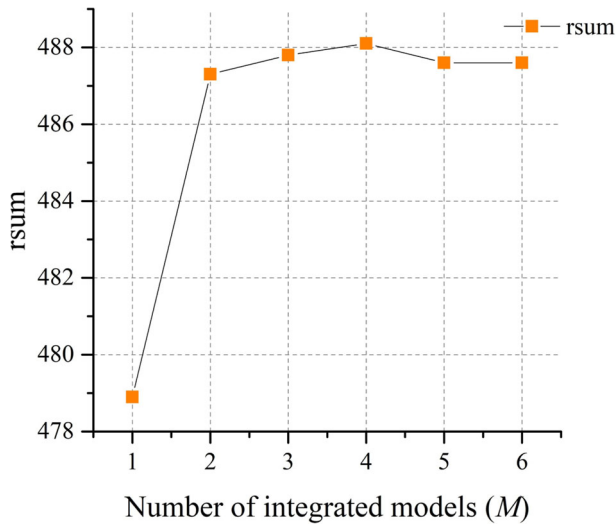


Fig. 12 Performance comparison on $rsum$ with the number of integrated models (M)

5 Conclusion and future work

In this study, a new framework named CMRN was proposed to address the cross-modal image-text matching task by extracting more abundant image and text representations. A multi-relationship aware reasoning method was designed to perform reasoning among static objects in an image on semantic and spatial levels. The relation-aware image representation in turn result in better image-text matching performance. Furthermore, a BERT-based text encoder was adopted to learn distinctive textual representations containing contextual information. Extensive experiments were conducted on two public datasets in cross-modal retrieval. The experimental results demonstrated that the proposed model was superior in handling image-text matching problems.

In the future, we plan to extend our work in the following two directions. First, we will develop some reconstruction constraints for better text feature extraction, such as generating image regions close to the ground-truth from the learned text representation. Second, we will consider designing a region counting module for image feature extraction to align the numbers in the corresponding text.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grant 61871278, the Industrial Cluster Collaborative Innovation Project of Chengdu (no. 2016-XT00-00015-GX), the Sichuan Science and Technology Program (no. 2018HH0143).

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) Bottom-up and top-down attention for image captioning and visual question answering, pp 6077–6086, 06

2. Arqub OA, Abo-Hammour ZS (2014) Numerical solution of systems of second-order boundary value problems using continuous genetic algorithm. *Inf Sci* 279:396–415
3. Chen J, Zhuge H (2019) Extractive summarization of documents with images based on multi-modal rnn. *Future Gener Comput Syst* 99:04
4. Chung J, Gülçehre Ç, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555
5. Cornia M, Baraldi L, Tavakoli HR, Cucchiara R (2020) A unified cycle-consistent neural model for text and image retrieval. *Multimed Tools Appl* 1–25, 07
6. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, vol 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, pp 4171–4186
7. Faghri F, Fleet DJ, Kiros JR, Fidler S (2018) Vse++: improving visual-semantic embeddings with hard negatives. In: *BMVC*
8. Frome A, Corrado GS, Shlens J, Bengio S, Dean J, Ranzato M, Mikolov T (2013) Devise: a deep visual-semantic embedding model. In: *NIPS*, pp 2121–2129
9. Gu J, Cai J, Joty S, Niu L, Wang G (2018) Look, imagine and match: improving textual-visual cross-modal retrieval with generative models. In: *2018 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp 7181–7189
10. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *2016 IEEE conference on computer vision and pattern recognition (CVPR)*, pp 770–778
11. Hou J, Wu X, Qi Y, Zhao W, Luo J, Jia Y (2019) Relational reasoning using prior knowledge for visual captioning. *Computer Vision and Pattern Recognition*. arXiv:1906.01290
12. Huang Y, Wang W, Wang L (2017) Instance-aware image and sentence matching with selective multi-modal lstm. In: *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp 7254–7262
13. Huang Y, Wu Q, Song C, Wang L (2018) Learning semantic concepts and order for image and sentence matching, pp 6163–6171, 06
14. Huang F, Zhang X, Zhao Z, Li Z (2019) Bi-directional spatial-semantic attention networks for image-text matching. *IEEE Trans Image Process* 28:2008–2020
15. Karpathy A, Feifei L (2015) Deep visual-semantic alignments for generating image descriptions. In: *2015 IEEE conference on computer vision and pattern recognition (CVPR)*, pp 3128–3137
16. Kingma D, Adam JBA (2014) A method for stochastic optimization. In: *International conference on learning representations*, p 12
17. Kipf T, Welling M (2017) Semi-supervised classification with graph convolutional networks. arXiv:1609.02907
18. Klein B, Lev G, Sadeh G, Wolf L (2015) Associating neural word embeddings with deep image representations using fisher vectors. In: *IEEE conference on computer vision and pattern recognition (CVPR)*, pp 4437–4446, 06
19. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li L, Shamma DA et al (2017) Visual genome: connecting language and vision using crowdsourced dense image annotations. *Int J Comput Vis* 123(1):32–73
20. Lee K, Chen X, Hua G, Hu H, He X (2018) Stacked cross attention for image-text matching. In: *ECCV*. Springer, Cham, pp 212–228
21. Li Y, Tarlow D, Brockschmidt M, Zemel RS (2015) Gated graph sequence neural networks. *CoRR*, arXiv:1511.05493
22. Li S, Xiao T, Li H, Yang W, Wang X (2017) Identity-aware textual-visual matching with latent co-attention. In: *2017 IEEE international conference on computer vision (ICCV)*, pp 1908–1917
23. Li K, Zhang Y, Li K, Li Y, Fu Y (2019) Visual semantic reasoning for image-text matching. In: *2019 IEEE/CVF International conference on computer vision (ICCV)*, pp 4653–4661
24. Li L, Gan Z, Cheng Y, Liu J (2019) Relation-aware graph attention network for visual question answering. In: *2019 IEEE/CVF international conference on computer vision (ICCV)*, pp 10313–10322
25. Lin X, Parikh D (2016) Leveraging visual question answering for image-caption ranking 9906:261–277, 10
26. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick C (2014) Microsoft coco: common objects in context, 8693, 04
27. Liu Y, Guo Y, Liu L, Bakker EM, Lew MS (2019) Cyclematch: a cycle-consistent embedding network for image-text matching. *Pattern Recognit* 93:365–379, 05

28. Liu C, Mao Z, Liu A-A, Zhang T, Wang B, Zhang Y (2019) Focus your attention: a bidirectional focal attention network for image-text matching. In: Proceedings of the 27th ACM international conference on multimedia, MM '19. Association for Computing Machinery, New York, pp 3–11
29. Liu C, Mao Z, Zhang T, Xie H, Wang B, Zhang Y (2020) Graph structured network for image-text matching. arXiv:2004.00277
30. Lu J, Batra D, Parikh D, Lee S (2019) Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in neural information processing systems, pp 13–23
31. Ma L, Lu Z, Shang L, Li H (2015) Multimodal convolutional neural networks for matching image and sentence. In: 2015 IEEE international conference on computer vision (ICCV), pp 2623–2631
32. Ma L, Jiang W, Jie Z, Wang X (2019) Bidirectional image-sentence retrieval by local and global deep matching. *Neurocomputing* 345:36–44, 02
33. Messina N, Falchi F, Esuli A, Amato G (2020) Transformer reasoning network for image-text matching and retrieval. arXiv:2004.09144
34. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv:1301.3781
35. Nam H, Ha J, Kim J (2017) Dual attention networks for multimodal reasoning and matching. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 2156–2164
36. Norcliffebrown W, Vafeias S, Parisot S (2018) Learning conditioned graph structures for interpretable visual question answering. In: *NeurIPS*, pp 8334–8343
37. Paszke A, Gross S, Chintala S, Chanan G, Yang E, Devito Z, Lin Z, Desmaison A, Antiga L, Lerer A (2017) Automatic differentiation in pytorch
38. Qi D, Su L, Song J, Cui E, Bharti T, Sacheti A (2020) Imagebert: cross-modal pre-training with large-scale weak-supervised image-text data. *Comput Vis Pattern Recognit*. arXiv:2001.07966
39. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell* 39:1137–1149, 06
40. Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G (2009) The graph neural network model. *IEEE Trans Neural Netw* 20(1):61–80
41. Trott A, Xiong C, Socher R (2018) Interpretable counting for visual question answering. arXiv:1712.08697
42. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: *NIPS*
43. Velickovic P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y (2018) Graph attention networks. arXiv:1710.10903
44. Venugopalan S, Rohrbach M, Donahue J, Mooney RJ, Darrell T, Saenko K (2015) Sequence to sequence—video to text. In: 2015 IEEE international conference on computer vision (ICCV), pp 4534–4542
45. Wang L, Li Y, Lazebnik S (2016) Learning deep structure-preserving image-text embeddings. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp 5005–5013, 06
46. Wang Y, Yang H, Qian X, Ma L, Lu J, Li B, Fan X (2019) Position focused attention network for image-text matching. *Computation and Language*. arXiv:1907.09748
47. Wang L, Li Y, Huang J, Lazebnik S (2019) Learning two-branch neural networks for image-text matching tasks. *IEEE Trans Pattern Anal Mach Intell* 41(2):394–407
48. Wang T, Xu X, Yang Y, Hanjalic A, Shen H, Song J (2019) Matching images and text with multi-modal tensor fusion and re-ranking. In: Proceedings of the 27th ACM international conference on multimedia, pp 12–20, 10
49. Wang P, Wu Q, Cao J, Shen C, Gao L, Van Den Hengel A (2019) Neighbourhood watch: referring expression comprehension via language-guided graph attention networks. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 1960–1968
50. Wei Y, Zhao Y, Lu C, Wei S, Liu L, Zhu Z, Yan S (2017) Cross-modal retrieval with cnn visual features: a new baseline. *IEEE Trans Syst Man Cybern* 47(2):449–460
51. Xu X, Wang T, Yang Y, Zuo L, Shen F, Shen HT (2020) Cross-modal attention with semantic consistence for image-text matching. *IEEE Trans Neural Netw Learn Syst* 02:1–14
52. Yang Z, Qin Z, Yu J, Hu Y (2018) Scene graph reasoning with prior visual relationship for visual question answering. arXiv: Multimedia
53. Yao T, Pan Y, Li Y, Mei T (2018) Exploring visual relationship for image captioning. In: European conference on computer vision, pp 711–727
54. Young P, Lai A, Hodosh M, Hockenmaier J (2014) From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. *Trans Assoc Comput Linguist* 2(1):67–78

55. Zhang Y, Lu H (2018) Deep cross-modal projection learning for image-text matching. In: The European conference on computer vision (ECCV), pp 707–723
56. Zhang Y, Hare JS, Prugelbennett A (2018) Learning to count objects in natural images for visual question answering. arXiv:[1802.05766](https://arxiv.org/abs/1802.05766)
57. Zheng Z, Zheng L, Garrett M, Yang Y, Xu M, Shen Y-D (2020) Dual-path convolutional image-text embeddings with instance loss. *ACM Trans Multimed Comput Commun Appl* 16(2): 1–23

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