Neural Computing and Applications (2023) 35:7529–7545
https://doi.org/10.1007/s00521-022-08048-w

ORIGINAL ARTICLE

Semantic-aware multi-branch interaction network for deep multimodal learning

Hao Pan[1,2] · Jun Huang[1,2]

Received: 14 June 2022 / Accepted: 7 November 2022 / Published online: 23 November 2022

Abstract

Deep multimodal learning has attracted increasing attention in artificial intelligence since it bridges vision and language. Most existing works only focus on specific multimodal tasks, which limits the ability to generalize to other tasks. Furthermore, these works only learn coarse-grained interactions at the object-level in images and the word-level in text, while ignoring to learn fine-grained interactions at relation-level and attribute-level. In this paper, to alleviate these issues, we propose a Semantic-aware Multi-Branch Interaction (SeMBI) network for various multimodal learning tasks. The SeMBI mainly consists of three modules, Multi-Branch Visual Semantics (MBVS) module, Multi-Branch Textual Semantics (MBTS) module and Multi-Branch Cross-modal Alignment (MBCA) module. The MBVS enhances the visual features and performs reasoning through three parallel branches, corresponding to the latent relationship branch, explicit relationship branch and attribute branch. The MBTS learns relation-level language context and attribute-level language context by textual relationship branch and textual attribute branch, respectively. The enhanced visual features then passed into MBCA to learn fine-grained cross-modal correspondence under the guidance of relation-level and attribute-level language context. We demonstrate the generalizability and effectiveness of the proposed SeMBI by applying it to three deep multimodal learning tasks, including Visual Question Answering (VQA), Referring Expression Comprehension (REC) and Cross-Modal Retrieval (CMR). Extensive experiments conducted on five common benchmark datasets indicate superior performance comparing with state-of-the-art works.

Keywords Semantic-aware · Multi-branch · Visual question answering · Referring expression comprehension · Cross-modal retrieval

1 Introduction

Our world is full of multimodal information, such as text, image, video and sound. Deep multimodal learning tasks, such as visual question answering [1, 7, 15, 23, 57], referring expression comprehension [27, 43, 51, 55] and cross-modal retrieval [20, 26, 36, 47], combining vision and language have attracted the attention of many researchers due to exposing many challenges to the artificial intelligence community. The key challenge in multimodal learning tasks lies in understanding a wide range of sophisticated semantics in an image, including attributes, spatial relationships, actions and intentions, and how all of these concepts are referred to and grounded in natural language.

In recent years, a large number of works have been proposed for multimodal learning tasks, which can be classified into two categories including learning object-word interactions [1, 15, 20, 57] and relational reasoning [3, 23, 26, 53, 55]. The methods of learning object-word interactions make use of object-level features from object detectors and discover alignments between salient regions and keywords. The image inputs are encoded into local object features by object detection methods such as Faster R-CNN [39], and the text inputs are encoded into textual

✉ Jun Huang
huangj@sari.ac.cn

Hao Pan
panhao2019@sari.ac.cn

[1] University of Chinese Academy of Sciences, Beijing 100049, China

[2] Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China

features by RNN based methods like LSTM [11] or GRU [6]. For the cross-modality interaction, most existing methods use co-attention [7, 20, 57] to connect two modalities. In this way, the problem of which objects to look at and what words to listen to can be solved. Although capturing object-word correspondences, these methods ignore the visual relationships in the image and the relational reasoning skill. As illustrated in Fig. 1, the model needs to recognize not only the objects ("surfer", "board", and "water") but also the visual relationships ("on" and "in") in the image.

To deal with the above issues, relational reasoning methods [3, 23, 26, 53, 55] have been proposed to capture the visual relationships between objects of the image. Unlike the object-word methods that learn coarse-grained cross-modal interactions, incorporating the relationships between image objects could benefit in learning fine-grained cross-modal interactions. Many methods have been proposed to achieve relational reasoning, which can be divided into two categories including latent relationships based methods [3, 12, 35, 53, 55] and explicit relationships based methods [23, 44]. Some latent relationships based methods generate relationships via combining pairwise regions and then mapping them through a MLP [3, 35, 53], others generate latent relationships by constructing the relative position and size of the bounding boxes [12, 55]. Although these methods have been proven to work well on relational reasoning, the relationships are learned in a weak-supervised manner, which lacks interpretability. The

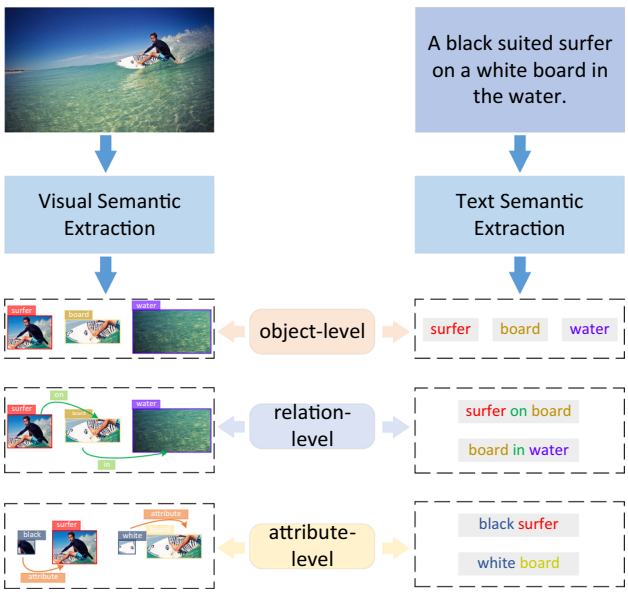

Fig. 1 Illustration of multi-level multimodal semantic extraction and fine-grained cross-modal alignment. Our SeMBI first models the visual semantic information in the image and models textual semantic information in the sentence, then it performs fine-grained cross-modal alignment in the built multimodal semantic space

explicit relationship-based methods [23, 44] utilize a pre-trained Scene Graph Generation (SGG) model [18] to predict the objects and relationships between them. A scene graph is a collection of explicit visual relationship triplets: <subject, predicate, object>, the subjects and objects are represented as nodes and the relationships or predicates between them are represented as edges. Although these methods successfully capture explicit visual relationships, the generated graph is very sparse and unbalanced due to the long tail of annotations, which limits its performance for downstream tasks. In addition, although the latent relationship-based methods and explicit relationship based methods are proposed, respectively, there is no method to comprehensively model the two relationships in an end-to-end framework.

Actually, in addition to focusing on the objects and relationships of the image, attributes are also important to better understand the image. As shown in Fig. 1, "black" and "white" are attribute semantics in the image. Therefore, it is vital to simultaneously model low-level semantic information (e.g. objects) and high-level semantic information (e.g. attributes and relationships). Moreover, it could be problematic to directly learn cross-modal interactions between various visual semantic components and full sentences. The reason is the lack of fine-grained cross-modal alignment at different levels, including nouns aligned with visual objects at object-level, adjectives aligned with visual attributes at attribute-level, and verbs or prepositions aligned with visual relationships at relation-level. Therefore, it is also necessary to capture the multi-level semantic information in the sentence and build fine-grained cross-modal alignment, which is ignored by other methods.

To overcome the aforementioned problems, we propose a novel Semantic-aware Multi-Branch Interaction (SeMBI) network for various multimodal learning tasks. The SeMBI performs intra- and inter-modality information interaction through multi-level semantic branches, which mainly consists of three modules: (1) the *Multi-Branch Visual Semantics* (MBVS) module jointly models low-level visual semantics (e.g. objects) and high-level visual semantics (e.g. attributes and relationships) in a parallel way. First, MBVS synchronously models low and high-level visual semantics through three parallel branches, which are latent relationship branch, explicit relationship branch (we prune the relationship of explicit relationship branch to avoid the problem of graph sparsity) and visual attribute branch. Then it updates the visual representation of each branch based on the cascaded self-attention model; (2) the *Multi-Branch Textual Semantics* (MBTS) module synchronously learns relation-level and attribute-level textual semantics through two parallel branches corresponding to textual relationship branch and textual attribute branch, respectively. In the textual relationship branch, we

first extract the relationship components (e.g. *surfer*-**on** -*board, board*-**in** -*water*) using the off-the-shelf Stanford Parser, then the relationship components are propagated to Graph Attention Network (GAT) to obtain the relation-level language context. In the textual attribute branch, we also use the Stanford Parser tool to extract the attribute components (e.g. **black** *surfer,* **white** *board*) , then the attribute components are propagated to Graph Convolution Network (GCN) to obtain the attribute-level language context; (3) the *Multi-Branch Cross-modal Alignment* (MBCA) module learns fine-grained cross-modal alignment and obtains the visual context under the guidance of both relation-level and attribute-level language context. On the top of the unified backbone, the visual and language contexts are used to task-specific heads for accomplishing each multimodal task.

Our contributions of this paper are summarized as follows.

- We propose a novel Semantic-aware Multi-Branch Interaction (SeMBI) network for various deep multimodal learning tasks by jointly modeling multi-level multimodal semantic information and learning fine-grained cross-modal alignment in a unified deep model.
- We propose an innovative Multi-Branch Visual Semantics (MBVS) module that comprehensively models multi-level visual semantic information by semantic-aware self-attention mechanism in an end-to-end framework.
- Multi-Branch Textual Semantics (MBTS) module is proposed to learn both relation-level language context via a relation-aware GAT and attribute-level language context through an attribute-aware GCN. And the Multi-Branch Cross-modal Alignment (MBCA) module is proposed to perform fine-grained cross-modal alignment under the guidance of both relation-level and attribute-level language context. As far as we know, this is the first work that performs cross-modal alignment at different semantic levels.
- We verify our proposed SeMBI on three common multimodal tasks: Visual Question Answering, Referring Expression Comprehension and Cross-Modal Retrieval. Extensive experiments conducted on five common benchmark datasets indicate superior performance compared with state-of-the-art works.

2 Related work

2.1 Low-level semantic modeling methods

Low-level semantic modeling methods aim to mine multimodal interactions between objects in an image and words in a sentence. VQA [1, 7, 15, 54, 57] aims to answer a question in natural language according to an image. BUTD [1] proposes a top-down attention on pre-detected salient regions, which is the first method to build object-word interaction, and subsequent methods are designed based on this model. Some other methods [7, 15, 54, 57] utilize attention mechanisms to explore object-word interaction. SANs [54] propose stacked attention networks to perform question-guided visual attention multiple times. Inspired by SANs, other co-attention models including BAN [15], MCAN [57] and DFAF [7] have been proposed. These models build sufficient interaction between each word in questions and each region in the image. REC [27, 43, 51, 55] aims to locate an object in an image referred to by a natural language expression. Most existing works [27] adopt co-attention mechanisms to build up the interactions between the expression and the objects in the image. CMR [20, 26, 47] is a task to get queries from one modality to retrieval information from another modality. Several works have been devoted to exploring the low-level semantic interactions for cross-modal retrieval [14, 20]. Karpathy et al. [14] first propose to extract image region features and text word features, and then align them in the embedding space. SCAN [20] follows SANs [54], using stacked cross-modal attention for similarity measures.

2.2 High-level semantic modeling methods

Due to the complex semantics of multimodal data, existing low-level semantic modeling methods do not well seize the high-level semantics, such as relationship and attribute. For the VQA task, methods like [3, 23, 53] take into account the high-level semantics of relationships. MuRel [3] is proposed to perform pairwise relationship modeling. TRRNet [53] and CRA-Net [35] propose a tiered relation reasoning method to improve the relational reasoning performance. ReGAT [23] is proposed to model multi-type visual relationships, such as spatial relationships, semantic relationships and implicit relationships. For the REC task, the methods in [51, 52, 55] decompose the model into components that deal with different semantics, and then fuse the matching scores of different components. MattNet [55] decomposes the model into subject, location and relationship modules, and computes a matching score for each module. CMRIN [51] highlights objects as well as multi-order relationships among them. DGA [52] network performs multi-step reasoning on top of the relationships among the objects. For the CMR task, only several works consider high-level semantic information [26, 49]. UniVSE [49] proposes an unified visual-semantic embedding approach that textual semantic is decomposed into different components such as relationships and attributes, which will

be aligned with the objects in the image. GSMN [26] learns fine-grained correspondence of relationships and attributes, which helps to improve the object correspondence. One limitation of these methods lies in the fact that they only consider the visual high-level semantics while ignoring the importance of textual high-level semantics. DC-GCN [13] proposes to simultaneously capture the visual relationships and the syntactic dependency relations between words in a question for the VQA task. MFM [33] comprehensively explores the multimodal matching relationships based on their multi-faceted representations. While these methods build both visual and textual relationships, they do not comprehensively build multi-level multimodal semantics and perform fine-grained cross-modal alignment at different semantic levels. Our work builds not only the high-level semantics of vision but also the multi-level semantics of text, as well as fine-grained cross-modal alignment.

2.3 Self-attention based methods

In this paper, we apply self-attention to capture multi-level visual semantics. Inspired by the self-attention in machine translation [41], lots of recent works [7, 47, 57] take use of the self-attention mechanism to implement multimodal learning tasks. MCAN [57] and DFAF [7] use self-attention to generate intra-modality attention maps and model the dense inter-modality interactions to improve VQA performance. CAMP [47] uses self-attention to obtain the aggregated message from other modalities for the CMR task. Different from the traditional self-attention mechanism, we add high-level visual semantic information to the traditional self-attention.

2.4 Graph networks

Graph neural networks have been very popular in recent years for aggregating information from neighbor nodes to target nodes in a graph. GCN is proposed to explore the relationships between objects in the image or words in the sentence, which has been applied to various multimodal tasks [13, 26]. For instance, DC-GCN [13] proposes a dual channel graph convolution network to explore the visual relationships and syntactic relationships. GSMN [26] utilizes GCN to update the matching vector of each node to address the CMR task. Our work applies GCN to capture the attribute semantics in the sentence. Recently, GAT [42] is proposed to overcome the disadvantages of GCN with attention mechanisms. Different attention weights represent the contribution of other nodes to the target nodes. ReGAT [23] introduces the GAT to do visual relational reasoning in VQA. LGRANs [43] proposes the language-guided GAT to dynamically learn object representations that better adapt to the referring expression. In this paper, we apply the GAT to obtain the relational semantics in the sentence. The difference between the previous graph attention mechanism and ours is their attention is obtained via the interaction between nodes, but our attention further adds relational semantics.

3 Method

In this section, we detail our proposed Semantic-aware Multi-Branch Interaction (SeMBI) network for deep multimodal learning. Figure 2 shows the framework of our proposed model. We will first describe the way of extracting features of image and text in sect. 3.1. Then, we will introduce the proposed MBVS module aiming to learn multi-level visual semantic information in Sect. 3.2 and introduce the proposed MBTS module for learning multi-level textual semantic information in Sect. 3.3. We will also present the MBCA module to establish the fine-grained cross-modal alignment in Sect. 3.4. Finally, the task-specific heads are discussed in Sect. 3.5.

3.1 Multimodal feature representation

Image representation. Almost all previous works [1, 7, 20, 26, 57] on multimodal learning use Faster R-CNN [39] to extract visual features. For a fair comparison, we extract the visual features by utilizing a Faster R-CNN model in conjunction with ResNet-101 [10], which is pre-trained on Visual Genome [19]. Specifically, given a image I, we extract an object feature vector f_i with 2048 dimensions and a box feature vector b_i with 4 dimensions for each image region r_i. The f_i is further transformed into a d dimensional vector v_i by a linear mapping. Then $V = \{v_1, ..., v_n\} \in \mathbb{R}^{n \times d}$ and $B = \{b_1, ..., b_n\} \in \mathbb{R}^{n \times 4}$ are used together to represent each image, where n is the number of detected regions in I.

Text representation. To extract the text features, given one-hot encoding w_i of each word in a sentence S, we first embed it into a 300-dimensional Glove vector [37] as $e_i = W_e w_i$. Then the word embeddings are input to a RNN based networks to produce the initial text representation. For VQA and REC tasks, we feed word embeddings into a LSTM [11]. For CMR task, we employ a bi-directional GRU [6] to enhance the text representation. Finally, we obtain the text representation $T = \{t_1, ..., t_k\} \in \mathbb{R}^{k \times d}$ for the sentence S, where k represents the number of words in one sentence. Note that text representation and image representation are in the same d-dimensional space.

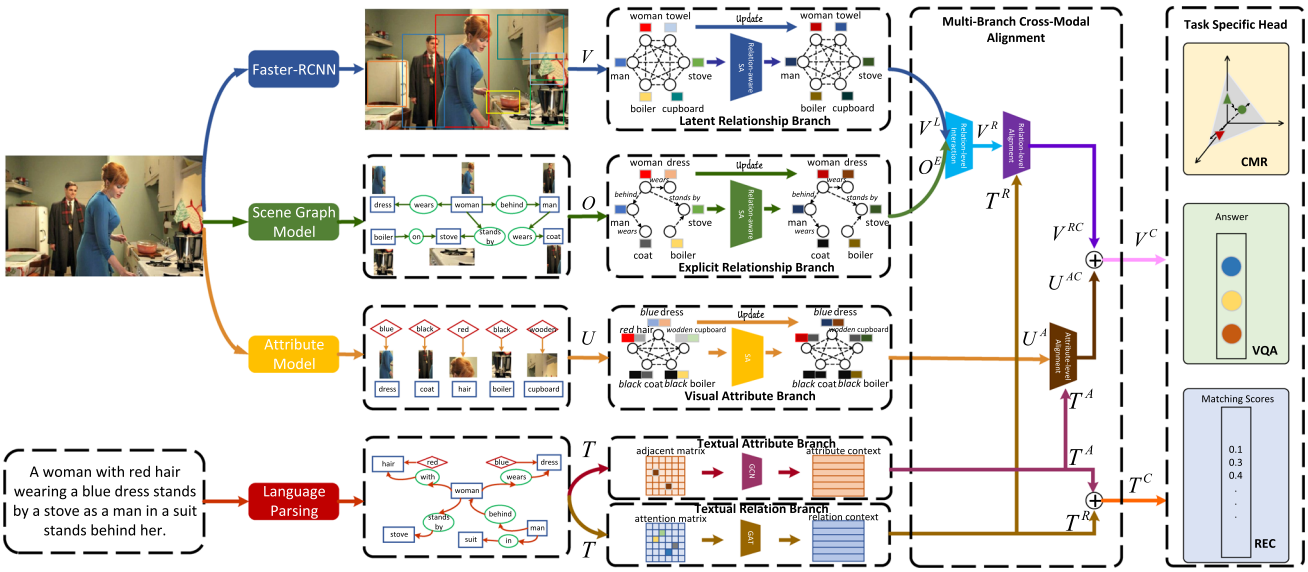

Fig. 2 An overview of our Semantic-aware Multi-Branch Interaction (SeMBI) network, which consists of three modules: (1) Multi-Branch Visual Semantic (MBVS) module: low and high-level visual semantics are captured through three parallel branches. (2) Multi-Branch Textual Semantic (MBTS) module: relation-level and attribute-level textual contexts are learned through two parallel branches. (3) Multi-Branch Cross-modal Alignment (MBCA) module: learning fine-grained cross-modal alignment and obtaining the visual context under the guidance of relation-level and attribute-level textual contexts

3.2 Multi-branch visual semantic

This part elaborates the method of visual semantic encoding. We first model the latent relationship and explicit relationship by latent relationship branch and explicit relationship branch, respectively. Then, we fuse the latent relation-aware visual feature and explicit relation-aware visual feature to get the unified visual representation. Finally, the attribute information in the image is modeled by the attribute branch. By the way, when we model the high-level visual semantics, the low-level visual semantics are also implicitly modeled. Due to the MBVS module is based on multi-head self-attention [41] mechanism, we first review a basic form of this mechanism.

3.2.1 Multi-head self-attention

In the multi-head self-attention, h parallel single-heads are concatenated to improve the expression capacity of the attended features. In the single-head attention, giving the set of feature representation $F \in \mathbb{R}^{m \times d_f}$, defining the query as $Q_F = FW_i^Q$, key as $K_F = FW_i^K$ and value as $V_F = FW_i^V$, where $W_i^Q \in \mathbb{R}^{d_f \times d_h}$, $W_i^K \in \mathbb{R}^{d_f \times d_h}$ and $W_i^V \in \mathbb{R}^{d_f \times d_h}$ are weights, the sub-script i donates for the i-th head and d_h is the dimensionality of the output features from each head. The single-head self-attention is computed as follows:

$$head_i = softmax\left(\frac{Q_F K_F^T}{\sqrt{d_h}}\right) V_F. \tag{1}$$

The multi-head attended output feature F' is given by:

$$F' = [head_1, head_2, ..., head_h] W_o. \tag{2}$$

where $W_o \in \mathbb{R}^{hd_h \times d_f}$.

3.2.2 Latent relationship branch

Inspired by [12], we model the latent relationship by constructing relative position of the bounding boxes. Then the latent relationship is added to the multi-head self-attention structure so that the model not only focuses on object-level semantic information, but also the relative geometric position between any pair of regions.

Specifically, the relative position $p_{ij} \in \mathbb{R}^4$ between two bounding boxes b_i and b_j is represented as

$$p_{ij} = \log\left(\frac{|x_i - x_j|}{w_i}\right), \log\left(\frac{|y_i - y_j|}{h_i}\right), \log\left(\frac{w_i}{w_j}\right), \log\left(\frac{h_i}{h_j}\right). \tag{3}$$

where (x, y) denotes the coordinate of the top-left point of the box, and h/w corresponds to the height/width of the box.

We then transform the p_{ij} to a d_l dimensional vector through a fully-connected layer to get the latent relationship:

$$R_{ij}^{L} = ReLU\left(p_{ij}W_1^l + b_1^l\right). \tag{4}$$

where $W_1^l \in \mathbb{R}^{4\times d_l}$ and $b_1^l \in \mathbb{R}^{1\times d_l}$ are trainable parameters, $R^L \in \mathbb{R}^{n\times n\times d_l}$.

Finally, V is fed into three independent fully-connected layers to get Q_V, K_V and V_V, corresponding to the query, key and value, respectively. We design relation-aware single-head self-attention as follows:

$$head_i = softmax\left(\frac{Q_V K_V^T}{\sqrt{d_h}} + ReLu\left(R^L W_2^l + b_2^l\right)\right)V_V. \tag{5}$$

where $W_2^l \in \mathbb{R}^{d_l\times 1}$ and $b_2^l \in \mathbb{R}^1$. The single-head attention is then substituted into Eq. 2 to obtain the latent relation-aware visual feature $V' \in \mathbb{R}^{n\times d}$. We borrow the residual skip-connection trick [10] to reserve the original visual information and stack this branch L times as follows :

$$V^{(l+1)} = LayerNorm\left(V^{(l)} + Dropout\left(V'^{(l)}\right)\right). \tag{6}$$

where *LayerNorm* is used here to stabilize training. For $V^{(1)}$, we set its inputs $V^{(0)} = V$ and $V'^{(0)} = V'$, respectively.

In this way, the updated features $V^L \in \mathbb{R}^{n\times d}$ from the relation-aware multi-head self-attention not only include the object-level semantic information, but also the semantic information with latent relationship.

3.2.3 Explicit relationship branch

To capture the explicit relationships from the image, we first use the off-the-shelf Scene Graph Generation (SGG) model to convert the input image into a scene graph [18], where nodes correspond to objects and the edges correspond to the relationships between objects. The SGG model is pre-trained on Visual Genome [19]. The original Visual Genome dataset consists of 150 most frequent objects and 50 relationship classes. However, the relationship classes are very imbalanced and show a long-tailed distribution, which creates a strong frequency bias and makes it particularly difficult for generalization. Therefore, we choose the 16 relationships (including no relationships) with the most frequency, which covers more than 80% of the instances in the Visual Genome.

Specifically, we represent a generated scene graph as $\mathcal{G}_v = (\mathcal{V}_v, \mathcal{E}_v)$, where $\mathcal{V}_v$ is the node-set, and $\mathcal{E}_v$ is the edge-set. Suppose there are the same n object nodes as original visual representation V. Each node will be encoded into a d-dimension visual feature vector by the SGG model. So the visual features can be represented as $O = \{o_1, ..., o_n\} \in \mathbb{R}^{n\times d}$. Each edge has a word label predicted by the SGG model. We utilize a word embedding layer to transform the label of each edge into a feature vector. Concretely, given the one-hot vector e_{ij}^o, which represents the relationship of o_i and o_j, it is passed through a word embedding layer to obtain the embedded feature $R_{ij}^E \in \mathbb{R}^{d_e}$. The overall feature is represented by $R^E \in \mathbb{R}^{n\times n\times d_e}$, where d_e is the dimension of word embedding.

Like the latent relationship branch, we transform the visual feature O into query Q_O, key K_O and value V_O, respectively, and take R^E as input to form relation-aware single-head self-attention as follows:

$$head_i = softmax\left(\frac{Q_O K_O^T}{\sqrt{d_h}} + ReLu\left(R^E W_1^e + b_1^e\right)\right)V_O. \tag{7}$$

where $W_1^e \in \mathbb{R}^{d_e\times 1}$ and $b_1^e \in \mathbb{R}^1$. The single-head attention is then substituted into Eq. 2 to obtain the explicit relation-aware visual feature $O' \in \mathbb{R}^{n\times d}$. Like Eq. 6, multiple stacks of this branch are also applied for O and O' to obtain the updated feature $O^E \in \mathbb{R}^{n\times d}$.

Thus, the updated feature O^E includes not only the object-level semantic information, but also the semantic information with explicit relationship. Some previous methods [23, 44] also use scene graph to extract explicit visual relationship, the difference between these methods and our Explicit Relationship Branch is our method designs a relation-aware multi-head self-attention mechanism to integrate relationships for updating visual features.

3.2.4 Relation-level interaction

With the relation-aware visual features V^L and O^E obtained by the two branches mentioned above, we fuse them by different fusion approaches to get the unified relation-aware visual representation as shown in Fig. 3. To the best of our knowledge, our method is the first work that comprehensively models latent relationships and explicit relationships in an end-to-end framework.

i) Linear Fusion. We simply fuse V^L and O^E via an element-wise summation.
ii) Gate Fusion. The gate fusion approach adopts a gate mechanism for relation-level fusion.

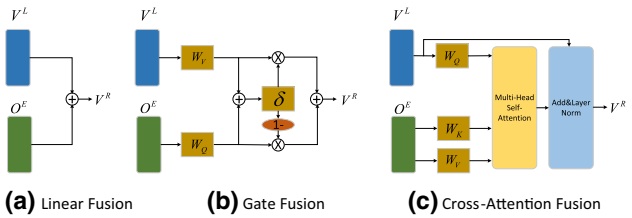

Fig. 3 Illustration of different fusion methods of latent relationship branch and explicit relationship branch

$$g = \sigma\left(W_V V^L + W_Q O^E\right).$$
$$V^R = g \cdot W_V V^L + (1-g) \cdot W_Q O^E. \tag{8}$$

where W_V and W_Q are to-be-learned parameters. g performs as a gate to select the most important information.

iii) Cross-Attention Fusion. The cross-attention fusion approach takes one feature as the primary feature and uses it to guide the attention learning for another feature. The attended feature are then integrated into the primary feature to get the output feature. Assuming that V^L corresponds to the primary feature, we project V^L as query Q_{V^L}, project O^E as key K_{O^E} and value V_{O^E}. Then, the projected Q_{V^L}, K_{O^E} and V_{O^E} are fed into Eq. 1 and Eq. 2 to get the attended feature $O^F \in \mathbb{R}^{n\times d}$. The obtained feature O^F has the same shape as V^L, and so O^F can be integrated with V^L via an element-wise summation just like Eq. 6. The feature after fusion is represented by $V^R \in \mathbb{R}^{n\times d}$.

3.2.5 Attribute branch

Inspired by [25], which focuses on multiple attributes to solve the graph clustering problem. Attributes are also important high-level semantics that can help understand images. In order to obtain visual attribute information, we use the attribute prediction model pre-trained on Visual Genome dataset (The dataset has 1,601 object classes and 401 attribute classes.).

Concretely, each object and corresponding attribute have a word label predicted by the attribute prediction model. Given the object one-hot vectors L^s and attribute one-hot vectors L^a, two label embedding layers are built to embed the L^s into $S = \{s_1, ..., s_n\} \in \mathbb{R}^{n\times d_e}$ and L^a into $A = \{a_1, ..., a_n\} \in \mathbb{R}^{n\times d_e}$, respectively. After obtaining the object feature and attribute feature, it is necessary to fuse them into a unified representation. Specifically, the object feature and attribute feature are concatenated and mapped through a nonlinear layer to obtain the fused feature $U \in \mathbb{R}^{n\times d}$.

Finally, we feed the unified representation U into Eqs. 1 and 2, then stack L layers as in Eq. 6 to get the updated feature $U^A \in \mathbb{R}^{n\times d}$. Therefore, the updated features obtained from this branch contain both the semantic information of the object-level and the attribute-level.

3.3 Multi-branch textual semantic

In this section, we first parse the sentence to get the relationship components and attribute components. Then, these two components are fed to textual relationship branch and textual attribute branch to get the relation-level language context and attribute-level language context, respectively.

3.3.1 Language parsing

Given a sentence S, we use an off-the-shelf Stanford Parser [40] to parser the sentence into relationship components and attribute components. For example, the sentence "*A brown fox chases a white rabbit*" is parsed to obtain the relationship components (e.g. fox **chases** rabbit) and attribute components (e.g. **brown** fox, **white** rabbit).

3.3.2 Textual relationship branch

After parsing, we find that there are thousands of relationship types in the whole dataset. We count the relationship types of VQA dataset, and there are 2,262 relationships in total. However, the relationship types are very imbalanced and show a long-tailed distribution. The same is true for other datasets. Based on this observation, we only select the N relationships (including no relationships) that occur most frequently.

Specifically, we construct a fully-connected graph $\mathcal{G}_r = (\mathcal{T}_r, \mathcal{E}_r)$ based on the extracted relationship components, where $\mathcal{T}_r = \{t_i\}_{i=1}^k$ is a set of nodes and t_i is text representation of each word. $\mathcal{E}_r$ is a set of edges and edge e_{ij}^t denotes the relationship between t_i and t_j in the relationship components. We use the word embedding layer to encode each edge e_{ij}^t into a relational representation[1] $R_{ij}^S \in \mathbb{R}^{d_e}$.

Following previous work [42], we employ the multiple stacked GAT layers to obtain the relation-level language context, so that it can capture key words and relational semantics of a sentence. More formally, given the node representations T and the edge representations R^S, the updated feature of the $l-$th layer is calculated as:

$$\alpha_{ij}^{(l)} = \frac{exp\left(t_i^{(l)} W_1^{(l)} \cdot \left(t_j^{(l)} W_2^{(l)}\right)^T + ReLU\left(W_3^{(l)} R_{ij}^S\right)\right)}{\sum_{t_j^{(l)} \in \mathcal{N}_{t_i^{(l)}}} exp\left(t_i^{(l)} W_1^{(l)} \cdot \left(t_j^{(l)} W_2^{(l)}\right)^T + ReLU\left(W_3^{(l)} R_{ij}^S\right)\right)}. \tag{9}$$

$$t_i^{(l+1)} = t_i^{(l)} + MLP\left(\sum_{t_j^{(l)} \in \mathcal{N}_{t_i^{(l)}}} \alpha_{ij}^{(l)} W_4^{(l)} t_j^{(l)}\right). \tag{10}$$

where $t_i^{(0)} = t_i$, $W_1^{(l)}, W_2^{(l)} \in \mathbb{R}^{d\times d}$, $W_3^{(l)} \in \mathbb{R}^{d_e\times 1}$ and $W_4^{(l)} \in \mathbb{R}^{d\times d}$ are trainable parameters, MLP denotes two fully-

[1] If there is no relationship between words t_i and t_j, the relational embedding between them is padded with 0.

connected layers with ReLU activation and dropout. Finally, we stack L-layer GAT and take output of the last layer from GAT to acquire the final relation-level language context $T^R = \{t_1^{(L)}, ...t_k^{(L)}\} \in \mathbb{R}^{k\times d}$. Compared with previous GAT methods that only obtain attention weights through interactions between nodes, our GAT method adds additional relational semantics to learn attention weights.

3.3.3 Textual attribute branch

After parsing the attribute components of a sentence, we construct a sparse graph $\mathcal{G}_a = (\mathcal{T}_a, \mathcal{E}_a)$. $\mathcal{T}_a = \{t_i\}_{i=1}^k$ is a set of nodes and t_i is text representation of each word. There exists graph edge between nodes if they are attribute components. We use matrix $\mathcal{M}$ to represent the adjacent matrix of each node where $\mathcal{M}_{ij} = 1$ if node i and node j form an attribute component, else $\mathcal{M}_{ij} = 0$.

Following previous work [17], we use multiple stacked GCN layers to obtain the attribute-level language context, so that it can capture key words and attribute semantics of a sentence. Concretely, given the node representations T and the adjacent matrix $\mathcal{M}$, the updated feature of the l-th layer is calculated as:

$$\beta_{ij}^{(l)} = \frac{exp(c_i^{(l)} W_5^{(l)} \cdot (c_j^{(l)} W_6^{(l)})^T)}{\sum_{c_j^{(l)} \in \mathcal{N}_{c_i^{(l)}}} exp(c_i^{(l)} W_5^{(l)} \cdot (c_j^{(l)} W_6^{(l)})^T)}. \tag{11}$$

$$c_i^{(l+1)} = c_i^{(l)} + \sum_{c_j^{(l)} \in \mathcal{N}_{c_i^{(l)}}} \mathcal{M}_{ij} \beta_{ij}^{(l)} W_7^{(l)} c_j^{(l)}. \tag{12}$$

where $c_i^{(0)} = t_i$, $W_5^{(l)}$, $W_6^{(l)}$ and $W_7^{(l)} \in \mathbb{R}^{d\times d}$ are trainable parameters. Following the stacked L-layer GCN, we obtain the attribute-level language context $T^A = \{c_1^{(L)}, ..., c_k^{(L)}\} \in \mathbb{R}^{k\times d}$ from the last layer. Due to the problem of vanishing gradients, the original GCN is difficult to perform multi-layer stacking, and our GCN solves this problem by using the residual skip-connection trick.

3.4 Multi-branch cross-modal alignment

To obtain the visual context related to language context, we align the relation-aware visual representation V^R by utilizing the relation-level language context T^R as the guided vector and align the attribute-aware visual representation U^A by utilizing the attribute-level language context T^A, respectively. As far as we know, this is the first work that performs cross-modal alignment at different semantic levels.

3.4.1 VQA and REC

VQA and REC tasks require joint reasoning over the text and the images. We first depict the cross-modal alignment on relationship branch in details, and then roughly describe that on attribute branch since this operation is same on two kinds of branch. Concretely, following Transformer [41] model, we first transform V^R into query feature and transform T^R into key and value features, where transformed features are denoted as Q_{V^R}, K_{T^R} and V_{T^R}. Then, these transformed features are fed into Eqs. 1 and 2 to get the multi-head attended feature. In order to adjust the representations, a feed-forward layer takes the multi-head attended features and further transforms them through two fully-connected layers as follows:

$$FFN(x) = FC(Dropout(ReLU(FC(x)))). \tag{13}$$

Finally, residual connection followed by *LayerNorm* is applied to the outputs of the two layers. The number of layers that this module stacks is set to L. The final output is represented as relation-aware visual context V^{RC}.

Similarly, the Transformer model is proceeded on U^A and T^A in the attribute branch, following the stacked L-layer Transformer, producing the attribute-aware visual context U^{AC}.

3.4.2 CMR

CMR between images and texts requires comparing the similarity across these two modalities on semantic level. Unlike VQA and REC tasks, which need to study fine-grained interactions and learn attention weight, CRM needs to map the image features and text features into a common semantic space for similarity measurement, so we compute the non-parametric attention weight between visual representations and language contexts using the cosine function. We first depict the cross-modal alignment on relationship branch in details, and then roughly describe that on attribute branch since this operation is same on two kinds of branch. Concretely, we first compute similarities between V^R and T^R, *i.e.*

$$sim_{ij} = \frac{v_i^{R^T} t_j^R}{\|v_i^R\| \cdot \|t_j^R\|}. \tag{14}$$

Attention is preformed on T^R with respect to each image region v_i^R:

$$v_i^{RC} = \sum_{j=1}^{k} \omega_{ij} t_j^R. \tag{15}$$

where

$$\omega_{ij} = \frac{exp(sim_{ij})}{\sum_{j=1}^{k} exp(sim_{ij})}. \tag{16}$$

We define V^{RC} as relation-aware visual context, in which each element is aligned by each v_i^R and the whole T^R.

Likewise, to attend on T^A with respect to each image region u_i^A, we define a weighted combination of language context. The final output is attribute-aware visual context U^{AC}.

3.5 Application to Specific Tasks

To apply our SeMBI model to deep multimodal learning tasks, we build task-specific output modules based on the relation-level language context T^R, attribute-level language context T^A, relation-aware visual context V^{RC} and attribute-aware visual context U^{AC}.

3.5.1 Architecture for VQA

The overall language context $T^C \in \mathbb{R}^{k\times d}$ is computed as the element-wise summation of language context at two level, and so does the visual context , we get the overall visual context $V^C \in \mathbb{R}^{n\times d}$. We use two independent attention model for V^C and T^C to obtain their attended feature $v_c \in \mathbb{R}^{1\times d}$ and $t_c \in \mathbb{R}^{1\times d}$, respectively. After that, the attended features are fused together as follows:

$$z = LayerNorm(W_v^T v_c + W_t^T t_c). \tag{17}$$

$$p = softmax(W_p z + b_p). \tag{18}$$

where $W_v \in \mathbb{R}^{d\times d_z}$, $W_t \in \mathbb{R}^{d\times d_z}$, $W_p \in \mathbb{R}^{d_z\times M}$ and $b_p \in \mathbb{R}^{1\times M}$ are learnable parameters, and p means the probability of the classified answers from the set of answer vocabulary which contains M candidate answers. During training, we use binary cross-entropy (BCE) loss function for answer classification.

3.5.2 Architecture for REC

We reuse the language context T^C and visual context V^C in the VQA head. Then, we feed the language context T^C into the attention model to obtain the attended feature $t_c \in \mathbb{R}^{1\times d}$. After that, t_c is broadcasted and integrated with visual context V^C as follows:

$$Z = LayerNorm(W_v^T V^C + W_t^T t_c). \tag{19}$$

Then, we apply two fully-connected layers to project each attended feature $z \in Z$ into a score $s \in \mathbb{R}^1$ and a 4-D bounding box coordinate $c \in \mathbb{R}^4$. We utilize KL-divergence as a ranking loss $\mathcal{L}_{rank}$ and $l1$ loss as a regression loss $\mathcal{L}_{reg}$ to optimize the model. Formulated as,

$$\begin{aligned}\mathcal{L}_{REC} &= \mathcal{L}_{rank} + \lambda\mathcal{L}_{reg} \\ &= \frac{1}{n}\sum_{i=1}^{n} s_i^* log(\frac{s_i^*}{s_i}) + \lambda\frac{1}{n}\sum_{i=1}^{n} L_1(c_i^*, c_i).\end{aligned} \tag{20}$$

where λ is a hyper-parameter to balance the two losses. s_i^* and c_i^* correspond to the ground-truth score and the ground-truth bounding box of *i-th* proposal, respectively.

3.5.3 Architecture for CMR

CMR aims to learn a matching score to measure the cross-modal similarity between the image-text pair. We can derive a matching score between image I and text S:

$$G(I,S) = \frac{1}{n}\sum_{i=1}^{n} sim(v_i^R, v_i^{RC}) + \frac{1}{n}\sum_{i=1}^{n} sim(u_i^A, u_i^{AC}). \tag{21}$$

where $sim()$ is the cosine function that measures the similarity between two input features. $v_i^R \in V^R$ and $u_i^A \in U^A$ correspond to the relation-aware region feature and attribute-aware region feature, respectively. $v_i^{RC} \in V^{RC}$ and $u_i^{AC} \in U^{AC}$ are the context features.

Following previous works [20], we employ the triplet loss as the loss function to enforce positive image-text pairs to be clustered and negative ones to be separated in the embedding spaces. We use the hard negative mining strategy, which makes the negative closest to the anchor, that is:

$$\begin{aligned}\mathcal{L}_{CMR} = &\sum_{(I,S)} [m - G(I,S) + G(I,S^-)]_+ \\ &+ [m - G(I,S) + G(I^-,S)]_+.\end{aligned} \tag{22}$$

where I^-, S^- are hard negatives, $[x]_+ = max(x, 0)$. m is a margin value.

4 Experiments

To evaluate the effectiveness of the proposed Semantic-aware Multi-Branch Interaction Network (SeMBI), we perform experiments in terms of Visual Question Answer (VQA), Referring Expression Comprehension (REC) and Cross-Modal Retrieval (CMR) on five publicly available datasets. Ablation studies are conducted to validate each module of our model. We also compare with recent state-of-the-art methods on the three tasks.

4.1 Datasets and protocols

We train and evaluate our model on VQA dataset, namely VQA-2.0 [8], three REC datasets, namely RefCOCO [56], RefCOCO+ [56] and RefCOCOg [34] and CMR dataset, namely MS-COCO [24]. The details are as follows:

VQA-2.0: VQA-2.0 [8] is a widely-used dataset for VQA task which has about 1,105,904 image-question pairs. The dataset contains 443,757 train questions, 214,354 validation questions, 447,793 test questions. The test set is further split into test-dev and test-std sets. The results are classified into three categories: *yes/no, number and other*. For evaluation metric, we use the tools provided by [2] to evaluate the accuracy of the predicted answer a:

$$Acc(a) = min\left(1, \frac{n_a}{3}\right). \tag{23}$$

where n_a is the total number of people that give the same answer as a.

RefCOCO, RefCOCO+ and RefCOCOg: The three datasets are collected on MS-COCO [24]. RefCOCO [56] contains 142,210 referring expressions for 19,994 images. RefCOCO+ [56] contains 141,564 referring expressions for 19,992 images. The two dataset are split into train, validation, testA, and testB. RefCOCOg [34] contains 95,010 referring expressions for 25,799 images. This dataset is split into train, validation and test. RefCOCO and RefCOCO+ datasets include short expressions while RefCOCOg has longer complex expressions. To compare our model with state-of-the-arts, we use the same evaluation metrics as those in [55]. The *Precision@1* metric is used for performance evaluation.

MS-COCO: MS-COCO [24] is an image captioning dataset containing about 123,287 images. Each image in MS-COCO [24] has 5 captions. As previous works [20, 47], we use 113,287 images to train all models, 5,000 images for validation and another 5,000 images for testing. To show the retrieval performance, we use $R@K$ (K=1,5,10) for cross-modal retrieval. $R@K$, defined as the percentage of queries in which the ground-truth matchings are contained in the first K retrieved results. The higher $R@K$ represents better performance. We also report *rSum*, which is the sum of $R@K$ for both image retrieval and sentence retrieval.

4.2 Implementation details

In our experiment, we use the following hyper-parameters as our default setting. We set the number of heads in multi-head self-attention as $h = 8$. The dimension of word-embedding d_e is set to 300. We train the proposed model on 2 NVIDIA 3080 GPUs. For each dataset, the specific parameter settings are as follows:

VQA Setup: For VQA-2.0, we use the Adam optimizer [16] with parameters $\alpha = 0.0001, \beta_1 = 0.9, \beta_2 = 0.99$. The size of candidate answers $M = 3,129$. The model is trained up to 13 epochs with batch size 64. For adjustment of learning rate, we use a warm-up strategy. Specifically, we begin with a learning rate of 0.00025, linearly increasing it till 0.0001 at epoch 4. After 10 epochs, the learning rate is decayed by 1/5 every 2 epochs. The hidden size of d is set to 512 and the dimensionality of fused feature d_z is set to 1,024. Each image has $n \in [0, 100]$ object features. The sequence of embedded words is fed into LSTM [11] for each time step. All the questions are padded and truncated to the same length 14.

REC Setup: For the three REC datasets, the length of textual queries is set to 15 and encode text feature through a LSTM [11]. The loss weight λ is set to 1. We use two visual features pre-trained on COCO [24] dataset and Visual Genome [19] dataset, respectively. During pre-training, we exclude the images in the training, validation and testing sets of RefCOCO [56], RefCOCO+ [56] and RefCOCOg [34]. We set the other hyper-parameters to be same as VQA-2.0.

CMR Setup: The model is trained up to 20 epochs with batch size 32. The initial learning rate is set as 0.0002 with decaying 1/10 every 10 epochs. For the texts, we use a bi-directional GRU [6] with one layer. For the images, we use Faster R-CNN [39] pre-trained on Visual Genome [19] to

Table 1 Accuracy of single model on VQA-2.0 test-dev and test-std dataset, it is trained on training, validation splits and Visual Genome dataset. (‡) denotes the most important evaluation metric

Model	test-dev				test-std
	All‡	Y/N	Num	Other	All‡
BUTD [1]	65.32	81.82	44.21	56.05	65.67
BAN [15]	69.52	85.31	50.93	60.26	–
SSCN [48]	69.78	–	–	–	70.09
DFAF [7]	70.22	86.09	53.32	60.49	70.34
DenIII [28]	70.5	86.3	50.9	61.5	70.8
MLVQA [32]	70.59	86.71	53.36	60.66	70.91
MCAN [57]	70.63	86.82	53.26	60.72	70.90
Murel [3]	68.03	84.77	49.84	57.85	68.41
CRA-Net [35]	68.61	84.87	49.46	59.08	68.92
TRRNet [53]	70.80	87.27	51.89	61.02	71.20
BGNs [9]	70.97	87.03	53.56	61.18	–
AGAN [61]	71.16	86.87	54.29	**61.56**	71.50
ReGAT [23]	70.27	86.08	54.42	60.33	70.58
VC RCNN [45]	71.21	**87.41**	53.28	61.44	71.49
DC-GCN [13]	71.21	87.32	53.75	61.45	71.54
SeMBI (Our)	**71.32**	87.16	**55.27**	61.30	**71.58**

extract initial 36 region features for each image. The hidden size of image feature and word feature is set as 1,024. The number of stacked layer L is experimentally set as 4. And the margin m is empirically set as 0.2.

4.3 Comparison with state-of-the-arts

We compare the proposed SeMBI model against existing state-of-the-art models on five publicly available datasets. As shown in Tables 1, 2 and 3, SeMBI achieves superior performance on all tasks, which verifies the effectiveness of our model. For a fair comparison, we do not compare SeMBI to the multimodal-BERT approaches which using large-scale data to pre-train a generalize model.

4.3.1 Results on VQA-2.0

In Table 1, we compare our results on the VQA-2.0 dataset to the state-of-the-art model. For a fair comparison, all the results are obtained by single model trained on the training, validation splits and Visual Genome dataset. The table is splitted into three blocks. The first block shows results of learning only object-word interactions. BUTD [1] is first model to use features based on Faster R-CNN [39] instead of grid features. BAN [15], DFAF [7] and MCAN [57] use complicated attention mechanism such as self-attention and co-attention. In the middle block, all the models are designed with semantic awareness, but contain only latent visual relationships. The third block shows results that combine several semantic information. ReGAT [23] also simultaneously captures latent visual relationships and explicit visual relationships, but it combines the two relationships through late fusion, which is inefficient. VC RCNN [45] uses extra visual common sense. While DC-GCN [13] captures both visual and syntactic relationships simultaneously, it lacks focus on attribute information and fine-grained cross-modal alignment, which are equally important. Our SeMBI belongs to the third category, but its cross-modal alignment is more fine-grained than others.

It is obvious that SeMBI outperforms all the state-of-the-art models, which proves the effectiveness of our proposed multi-branch model. Comparing with the results in the first block, our model surpasses them by a large margin. It demonstrated that our semantic-aware model achieves better performance compared to the ones that do not consider any kind of semantic. Comparing the results in the second block, our model achieves superior performance due to our model considers multimodal multi-level semantic information, including object, relationship and attribute. Despite complicated semantic information are considered in the third block, our model achieves better performance. Since other models only build intra-modality semantic information, our model considers fine-grained cross-modal alignment. It is obvious that SeMBI increases the overall accuracy of ReGAT by 1.05% on the test-dev set. However, SeMBI surpasses VC RCNN and DC-GCN by only 0.11% on the test-dev set, the reason is that the sentences in VQA-2.0 dataset are very short and lack complex reasoning, which leads to limited improvement of the model by the learned relational semantics and attribute semantics. Although *Y/N* and *Other* do not reach the best, our SeMBI outperforms them in other categories (e.g., *All*

Table 2 Comparison with state-of-the-art referring expression comprehension approaches on region proposals from detection model. For RefCOCO and RefCOCO+, testA is for grounding persons, and testB is for grounding objects

Model	Detector	Pre-trained Dataset	RefCOCO			RefCOCO+			RefCOCOg	
			val	testA	testB	val	testA	testB	val	test
LGRANs [43]	frcnn-vgg16	COCO	–	76.60	66.40	–	64.00	53.40	–	–
NMTree [27]	frcnn-vgg16	COCO	71.65	74.81	67.34	58.00	61.09	53.45	61.01	61.46
DGA [52]	frcnn-resnet101	COCO	–	78.42	65.53	–	69.07	51.99	–	63.28
MattNet [55]	frcnn-resnet101	COCO	76.40	80.43	69.28	64.93	70.26	56.00	66.67	67.01
MattNet [55]	mrcnn-resnet101	COCO	76.65	81.14	69.99	65.33	71.62	56.02	66.58	67.27
DDPN [58]	frcnn-resnet101	Genome	76.8	80.1	72.4	64.8	70.5	54.1	66.7	67.0
CM-Att-Erase [29]	frcnn-resnet101	COCO	78.35	83.14	71.32	68.09	73.65	58.03	67.99	68.67
MCN [31]	darknet53	COCO	80.08	82.29	74.98	67.16	72.86	57.31	66.46	66.01
CMRIN [51]	frcn-resnet101	COCO	–	82.53	68.58	–	75.76	57.27	–	67.38
HFR [38]	frcn-resnet101	COCO	79.76	83.12	75.51	66.80	72.53	57.09	69.71	69.08
CM-A-E+Ref-NMS [5]	mrcnn-resnet101	COCO	80.70	84.00	76.04	68.25	73.68	59.42	70.55	70.62
SeMBI (Our)	frcnn-resnet101	COCO	81.13	82.68	**79.65**	70.34	73.11	64.56	72.30	72.97
SeMBI (Our)	frcnn-resnet101	Genome	**83.13**	**86.67**	77.32	**74.21**	**79.79**	**65.15**	**74.57**	**74.79**

Table 3 Comparison of performance of our model with the state-of-the-art methods on MS-COCO dataset

Model	Image Query			Text Query			rSum
	R@1	R@5	R@10	R@1	R@5	R@10	
hiMoCS [62]	27.4	60.5	73.1	20.5	50.9	66.4	298.8
MFM [33]	58.9	86.3	92.4	47.7	81.0	90.9	457.2
UniVSE [49]	64.3	89.2	94.8	48.3	81.7	91.2	469.5
SAEM [50]	71.2	94.1	97.7	57.8	88.6	94.9	504.3
CAMP [47]	72.3	94.8	98.3	58.5	87.9	95.0	506.8
SCAN [20]	72.7	94.8	98.4	58.8	88.4	94.8	507.9
SGM [44]	73.4	93.8	97.8	57.5	87.3	94.3	504.1
CRGN [60]	73.8	95.6	98.5	60.1	88.9	94.5	511.4
VSRN [22]	74.0	94.3	97.8	60.8	88.4	94.1	509.4
BCAN [30]	74.2	95.6	98.4	58.6	87.3	93.9	508.0
MEMBER [21]	75.2	**96.1**	97.8	60.7	89.2	94.8	513.8
PFAN++ [46]	75.4	95.5	98.2	60.9	88.9	94.7	513.6
CCAN [59]	75.5	95.4	98.5	61.3	**89.7**	95.2	515.6
IMRAM [4]	76.1	95.3	98.2	61.0	88.6	94.5	513.7
GSMN [26]	76.1	95.6	98.3	60.4	88.7	95.0	514.0
SeMBI (Our)	**76.3**	95.5	**98.5**	**61.5**	89.3	**95.3**	**516.4**

of test-dev, *Num* and *All* of test-std). Since our model builds relative location and multi-level semantics, it is capable of counting and relational reasoning. However, the VQA-2.0 dataset lacks complex reasoning questions, so it cannot reflect the relational reasoning ability of our model, resulting in sub-optimal performance on *Y/N* and *Other*.

4.3.2 Results on the REC datasets

In Table 2, we show evaluation results on RefCOCO, RefCOCO+ and RefCOCOg, respectively. Our proposed SeMBI consistently outperforms existing methods across all the datasets. When using visual features pre-trained on COCO, the SeMBI improves the average accuracy over all the three datasets. Specially, it also surpasses all the existing models on the RefCOCOg dataset which has relatively longer expression requires relational reasoning capabilities. When using visual features pre-trained on Visual Genome, SeMBI improves the average accuracy over the testing sets achieved by the DDPN [58] method by 4.76%, 10.17% and 7.79%, respectively, on the RefCOCO, RefCOCO+ and RefCOCOg datasets. The results show that SeMBI outperforms existing state-of-the-arts methods regardless of the pre-trained features.

4.3.3 Results on MS-COCO

Table 3 shows the comparison between our model and state-of-the-art approaches on the MS-COCO dataset. We can see that our proposed SeMBI outperforms all the existing models, which further demonstrates the advantages of our model. Note that we only report the results of our method based on single model, it can be easily applied to ensemble model. While our model shows slightly lower scores than others under some metrics, it yields clearly superior performance against other competitors under the more crucial metric R@1 for retrieval task. Compared with the previous best model SCAN, our model achieves 2.6% and 1.8% improvement on R@1 for two directions. And compared with GSMN, our single model can achieve competitive results on MS-COCO. Although several recent works exploit multi-level semantic information between the two modalities and utilize cross-modal attention or graph neural network to build cross-modal alignment, which achieve superior improvement on CMR tasks. We believe that our SeMBI jointly modeling multi-level multimodal semantic information and fine-grained cross-modal alignment has greater advantages.

4.4 Ablation study

Module Analysis: To demonstrate the influence of different modules, we conduct ablation studies incrementally. To be more specific, first, we only add the latent relationship branch, and then add other modules incrementally. Note that in order to speed up the ablation study, we use 512 dimensional hidden size in all modules of the Cross-Modal Retrieval task.

As reported in Table 4, when combining latent relationship branch and explicit relationship branch, the results are better than just using latent relationship branch (in line2). This demonstrates the vital importance of jointly

Table 4 The ablation study on VQA-2.0 validation set, RefCOCO+ (pre-trained on Visual Genome) validation set and MS-COCO validation set to investigate the effect of different modules. The best results are highlighted in bold. We take the *rSum* as the evaluation metric for MS-COCO dataset. The "LRB" represents the Latent Relationship Branch in Sect. 3.2.2. "ERB" stands for Explicit Relationship Branch in Sect. 3.2.3. "AB" represents Attribute Branch in Sect. 3.2.5. "TRB" means Textual Relationship Branch in Sect. 3.3.2. "TAB" means Textual Attribute Branch in Sect. 3.3.3

Line	LRB	ERB	AB	TRB	TAB	VQA-2.0	RefCOCO+	MS-COCO
1	✔	✗	✗	✗	✗	67.48	72.57	510.5
2	✔	✔	✗	✗	✗	67.64	72.79	511.1
3	✔	✔	✔	✗	✗	67.72	72.83	511.5
4	✔	✔	✔	✔	✗	67.76	73.12	512.0
5	✔	✔	✔	✔	✔	**67.81**	**74.21**	**513.0**

investigating latent relationships and explicit relationships in an unified framework. Besides, when combining the three branches to model visual semantics, the model is further improved (line 3), revealing that multi-branch visual semantic module can model multi-level visual semantics and boost the model performance. Moreover, the performance improvement of adding textual relationship branch can be observed (line 4), indicating that it is important to model relational semantics in text and build fine-grained cross-modal alignment of relation-level. Finally, we add the textual attribute branch to form complete SeMBI (line 5). The results reach the highest on the three datasets. Compared to the RefCOCO+ and MS-COCO datasets, our model has a lower improvement on the VQA-2.0. The reason is that the VQA-2.0 dataset doesn't contain complex relational and attribute semantics. In general, our proposed model exceeds all other modules, verifying the effectiveness and complementarity of different modules.

Hyper-parameter: In Table 5, we compare the effects of different dimensions of relation features in latent relationship branch. For VQA task, we observe that $d_l = 64$ is better than $d_l = 128$ or $d_l = 256$. As for the REC task, the best result is when $d_l = 32$.

Moreover, we gradually increase the stacking layers L from 1 to 6 to train and evaluate them on the benchmark datasets. As shown in Fig. 4a, the performance of SeMBI steadily improves with increasing L on VQA-2.0, and $L = 6$ performs better. This observation well demonstrates that the stacking scheme effectively improves model performance. As shown in Fig. 4b, we can find that increasing the number of layers in an appropriate range (i.e., from 1 to 2) can improve the performance of referring expression comprehension task. This demonstrates that more layers offer more information interaction. However, when L is greater than 2, the performance begins to drop. The reason may be that too many layers bring more parameters, limiting the model optimization.

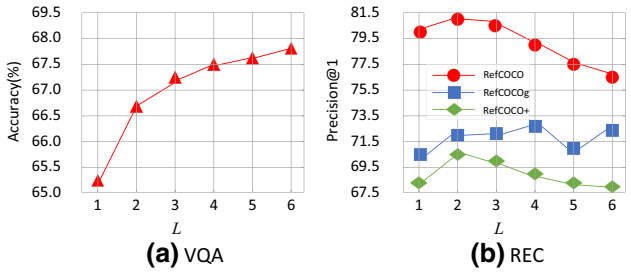

Fig. 4 Results comparison on VQA and REC (pre-trained on COCO) regarding different number of stacking layers L related to all branches

Relation-level fusion approaches: In the second column of Table 5, we compare various relation-level fusion approaches introduced in Sect. 3.2.4. We can observe that the cross-attention fusion approach achieves substantially better performance than other fusion approaches.

Table 5 Ablation experiments for hyper-parameters on VQA-2.0 and RefCOCO+. We train on the train split and report the results on the val split

Component	Setting	VQA-2.0	RefCOCO+
Latent Relationship Branch	$d_l = 32$	67.74	**74.21**
	$d_l = 64$	**67.81**	73.49
	$d_l = 128$	67.76	73.72
Visual Relation-Level Fusion	Linear Fusion	67.25	72.98
	Gate Fusion	67.39	73.46
	Cross-Attention Fusion	**67.81**	**74.21**

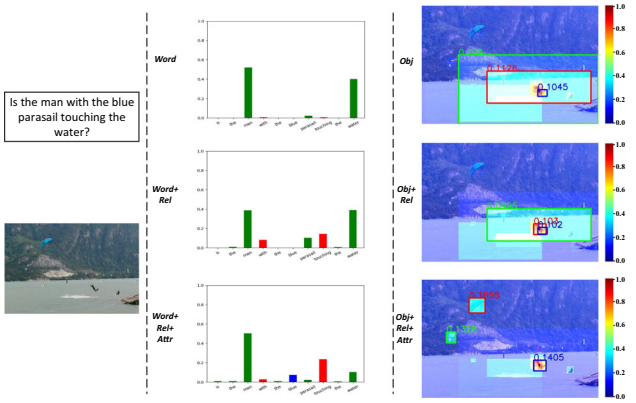

Fig. 5 Visualization of the learned textual attention maps and visual attention maps by three ablated models. The **Word** and **Obj** represent learning only the object-word semantics. The **Word+Rel** and **Obj+Rel** represent adding relation-level semantics. The **Word+Rel+Attr** and **Obj+Rel+Attr** represent adding both relation-level semantics and attribute-level semantics. Words representing relational semantics are marked in red, words representing attribute semantics are marked in blue, and other words are marked in green. The three bounding boxes shown in each image are the top-3 attended regions. The most interested object regions are marked by boxes in red. The numbers are attention weights

4.5 Visualization and analysis

To better illustrate the effectiveness of adding relation-level semantics and attribute-level semantics, we compare the attention maps learned by three ablated models. As shown in Fig. 5, comparing row 1 and row 2, we can see that relation-level semantics helps to capture the verbs or prepositions in the sentence which represent relational semantics, such as "with" and "touching". In addition, it can be seen from the visual attention map in the row 2 that with the relation-level semantics, the model pays more attention to the regions where man touches the water. Row 2 and row 3 show that adding the attribute-level semantics helps to focus on more adjectives in the sentence which represent attribute semantics, such as "blue". And from the visual attention map in the row 3 that with the attribute-level semantics, the model focuses more on the "blue parasail". These visualization results are consistent with the results reported in Table 4.

In Fig. 6, we visualize the learned attentions from relation-level cross-modal alignment branch and attribute-level cross-modal alignment branch, respectively. The effective alignment of vision and language is the key to multimodal learning. We can see that the relation-level cross-modal alignment enables our model to capture the relational words, such as "next to" in the first example and "standing behind" in the second example. The attribute-level cross-modal alignment enables our model to capture the attribute words, such as "circular, white" in the first example and "blue" in the second example. In addition, under the guidance of these key words, both models pay more attention to the relevant visual regions. These demonstrate that our model can accurately build fine-grained cross-modal alignment.

Figure 7 further shows four examples comparing the cross-modal retrieval results between the modal with and without semantics. Our semantic-aware approach is able to learn correct information of relationship and attribute. For example, in Fig. 7a, our semantic-aware model correctly comprehends the visual relationship "spoons laid out across table" in candidate images, so that it successfully ranks the image that contains the relationship at the top. When without semantics, the image contains "spoons in basket" will be ranked at the top instead. Another example in Fig. 7b, by correctly identifying the attribute of "snow", our semantic-aware model successfully ranks the correct image at the top. However, when without semantics, the

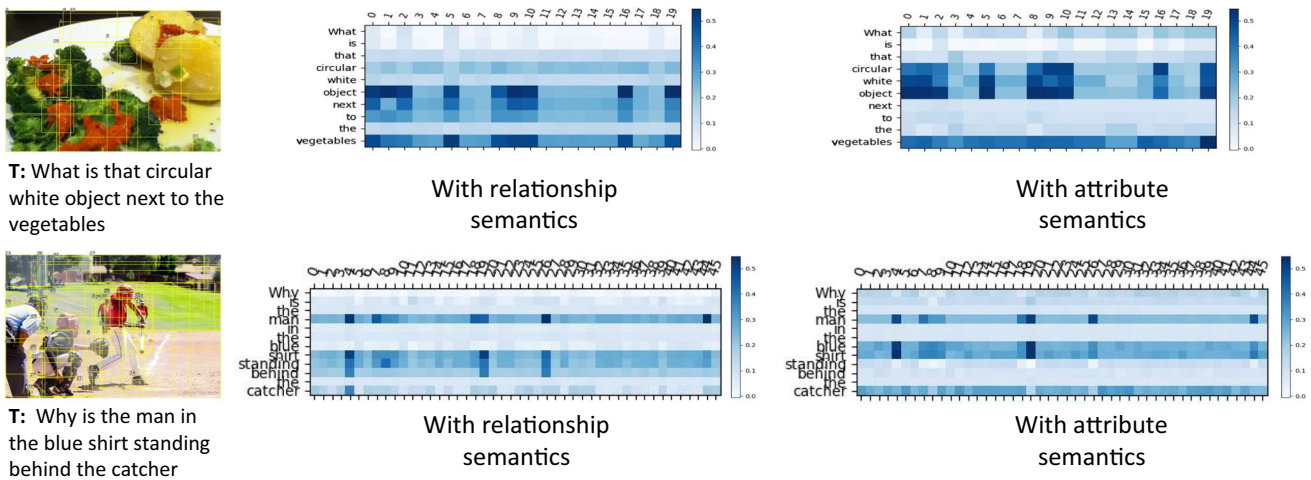

Fig. 6 Visualizations of the learned cross-attention maps. The index within [0-19] or [0-45] on the axes of the attention maps corresponds to the object in the image

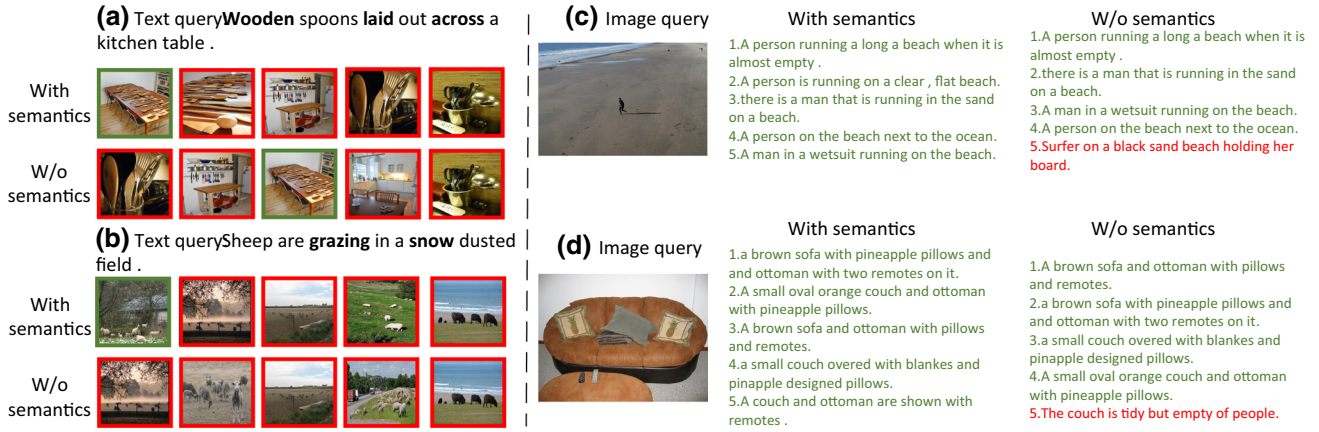

Fig. 7 Results of cross-modal retrieval. **a** and **b** are examples of text-to-image retrieval, we show top 5 ranked images from left to right. The mismatched images are with red boxes and matched images are with green boxes. **c** and **d** are examples of image-to-text retrieval, we show top 5 ranked texts. The mismatched texts are marked as red and matched texts are marked as green

image with "snow" attribute cannot be retrieved correctly. Figure 7c and d shows examples for image-to-text retrieval. Our semantic-aware model can accurately capture the visual relationships ("run", "next to" and "over") and visual attributes ("clear" and "flat"), so that it can correctly retrieve all the matching text. However, the model without semantics fails to retrieval all matching text.

5 Conclusion

In this paper, we present an effective semantic-aware model SeMBI to capture both multi-level visual semantics and multi-level textual semantics. Moreover, we conduct cross-modal alignment of the corresponding semantic branch, so that the image can better align with the corresponding text. To evaluate the effectiveness and generalizability of our model, we inject SeMBI into the models for Visual Question Answering, Referring Expression Comprehension and Cross-Modal Retrieval tasks. Extensive experiments demonstrate significant improvement compared with state-of-the-art approaches. Further, we do ablation studies proving the effectiveness of different modules. In the future, we will go further to apply this model to more multimodal learning tasks such as Image Captioning and Natural Language for Visual Reasoning.

Acknowledgements This paper was supported by National Key R &D Program of China (2019YFC1521204).

Data availability The datasets analysed during the current study are available from the corresponding author on reasonable request.

Declarations

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. In Proceedings of the IEEE conference on computer vision and pattern recognition 6077–6086
2. Antol S, Agrawal A, Jiasen L, Mitchell M, Dhruv BC, Zitnick L, Parikh D (2015) Vqa: visual question answering. In Proceedings of the IEEE international conference on computer vision 2425–2433
3. Cadene R, Ben-Younes H, Cord M, Thome N (2019) Murel: multimodal relational reasoning for visual question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 1989–1998
4. Chen H Ding G, Liu X, Lin Z, Liu J, Han J (2020) Imram: iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp 12652–12660
5. Chen L, Ma W, Xiao J, Zhang H, Chang SF (2021) Ref-nms: breaking proposal bottlenecks in two-stage referring expression grounding. Proceed AAAI Conf Artif Intell 35:1036–1044
6. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using rnn encoder-decoder for statistical machine translation. Proceedings of the 2014 conference on empirical methods in natural language processing
7. Gao P, Jiang Z, You H, Pan Lu, Hoi Steven CH, Wang X, Li H (2019) Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 6639–6648
8. Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D (2017) Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, 6904–6913
9. Guo D, Chang X, Tao D (2021) Bilinear graph networks for visual question answering. IEEE Transactions on neural networks and learning systems 1–12
10. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition 770–778

11. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
12. Han H, Jiayuan G, Zhang Z, Dai J, Wei Y (2018) Relation networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition 3588–3597
13. Huang Q, Wei J, Cai Y, Zheng C, Chen J, Leung HF, Li Q (2020). Aligned dual channel graph convolutional network for visual question answering. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 7166–7176,
14. Karpathy A, Joulin A, Fei-Fei L (2014) Deep fragment embeddings for bidirectional image sentence mapping. Advances in neural information processing systems 27: Annual conference on neural information processing systems
15. Kim JH, Jun J, Zhang BT (2018) Bilinear attention networks. Advances in Neural information processing systems 31: annual conference on neural information processing systems
16. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. International conference on learning representations
17. Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. International conference on learning representations
18. Knyazev B, de Vries H, Cangea C, Taylor GW, Courville A, Belilovsky E (2020) Graph density-aware losses for novel compositions in scene graph generation. British machine vision conference
19. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li LJ, Shamma David A et al (2017) Visual genome: connecting language and vision using crowdsourced dense image annotations. Int J Comp Vis 123(1):32–73
20. Lee KH, Chen X, Hua G, Houdong H, He X (2018) Stacked cross attention for image-text matching. In Proceedings of the European conference on computer vision (ECCV) 201–216
21. Li J, Liu L, Niu L, Zhang L (2021) Memorize, associate and match: embedding enhancement via fine-grained alignment for image-text retrieval. IEEE Trans Image Process 30:9193–9207
22. Li K, Zhang Y, Li K, Li Y, Yun F (2019) Visual semantic reasoning for image-text matching. In Proceedings of the IEEE/CVF international conference on computer vision. pp 4654–4662
23. Li L, Gan Z, Cheng Y, Liu J (2019) Relation-aware graph attention network for visual question answering. In Proceedings of the IEEE/CVF international conference on computer vision. pp 10313–10322
24. Tsung-Yi L, Michael M, Serge B, James H, Pietro P, Deva R, Piotr D, Lawrence ZC (2014). Common objects in context Microsoft coco. In: European conference on computer vision 10: 740–755. Springer
25. Lin Z, Kang Z, Zhang L, Tian L (2021) Multi-view attributed graph clustering. IEEE Transactions on knowledge and data engineering
26. Liu C, Mao Z, Zhang T, Xie H, Wang B, Zhang Y (2020) Graph structured network for image-text matching. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 10921–10930
27. Liu D, Zhang H, Feng W, Zha ZJ (2019) Learning to assemble neural module tree networks for visual grounding. In Proceedings of the IEEE/CVF international conference on computer vision 4673–4682
28. Liu F, Liu J, Fang Z, Hong R, Hanqing L (2021) Visual question answering with dense inter- and intra-modality interactions. IEEE Trans Multim 23:3518–3529
29. Liu X, Wang Z, Shao J, Wang X, Li H (2019) Improving referring expression grounding with cross-modal attention-guided erasing. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 1950–1959
30. Liu Y, Wang H, Meng F, Liu M, Liu H (2021) Attend, correct and focus: a bidirectional correct attention network for image-text matching. In: 2021 IEEE International conference on image processing (ICIP), pages 2673–2677
31. Luo G, Zhou Y, Sun X, Cao L, Chenglin W, Deng C, Ji R (2020) Multi-task collaborative network for joint referring expression comprehension and segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 10034–10043
32. Ma J, Liu J, Lin Q, Bei W, Wang Y, You Y (2021) Multitask learning for visual question answering. IEEE Transactions on neural networks and learning systems 1–15
33. Ma L, Jiang W, Jie Z, Jiang YG, Liu W (2020) Matching image and sentence with multi-faceted representations. IEEE Trans Circ Sys Video Technol 30(7):2250–2261
34. Mao J, Huang J, Toshev A, Camburu O, Yuille AL, Murphy K (2016) Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition 11–20
35. Peng L, Yang Y, Wang Z, Wu X, Huang Z (2019) Cra-net: composed relation attention network for visual question answering. In: Proceedings of the 27th ACM international conference on multimedia, pp 1202–1210
36. Peng Y, Huang X, Zhao Y (2018) An overview of cross-media retrieval: concepts, methodologies, benchmarks and challenges. IEEE Trans Circ Sys Video Technol 28(9):2372–2385
37. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
38. Qiu H, Li H, Wu Q, Meng F, Shi H, Zhao T, Ngan KN (2020) Language-aware fine-grained object representation for referring expression comprehension. In Proceedings of the 28th ACM international conference on multimedia. pp 4171–4180
39. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. Adv Neural Infor Process Sys 28:91–99
40. Schuster S, Krishna R, Chang A, Fei-Fei L, Manning CD (2015) Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In: Proceedings of the fourth workshop on vision and language. pp 70–80
41. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
42. Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y (2017) Graph attention networks. arXiv preprint arXiv:1710.10903
43. Wang P, Wu Q, Cao J, Shen C, Gao L, van den HA (2019) Neighbourhood watch: referring expression comprehension via language-guided graph attention networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1960–1968
44. Wang S, Wang R, Yao Z, Shan S, Chen X (2020) Cross-modal scene graph matching for relationship-aware image-text retrieval. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1508–1517
45. Wang T, Huang J, Zhang H, Sun Q (2020) Visual commonsense r-cnn. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10760–10770
46. Wang Y, Yang H, Bai X, Qian X, Ma Lin, Jing Lu, Li Biao, Fan Xin (2021) Pfan++: bi-directional image-text retrieval with position focused attention network. IEEE Trans Multim 23:3362–3376
47. Wang Z, Liu X, Hongsheng LL, Sheng JY, Wang X, Shao J (2019) Camp: cross-modal adaptive message passing for text-

image retrieval. In Proceedings of the IEEE/CVF International conference on computer vision pp. 5764–5773
48. Whitehead S, Wu H, Ji H, Feris R, Saenko K (2021) Separating skills and concepts for novel visual question answering. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 5628–5637
49. Wu H, Mao J, Zhang Y, Jiang Y, Li L, Sun W, Ma WY (2019) Univse: robust visual semantic embeddings via structured semantic representations. arXiv preprint arXiv:1904.05521
50. Wu Y, Wang S, Song G, Huang Q (2019) Learning fragment self-attention embeddings for image-text matching. In: Proceedings of the 27th ACM international conference on multimedia, pp 2088–2096
51. Yang S, Li G, Yizhou Y (2019) Cross-modal relationship inference for grounding referring expressions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 4145–4154
52. Yang S, Li G, Yizhou Y (2019) Dynamic graph attention for referring expression comprehension. In Proceedings of the IEEE/CVF international conference on computer vision, pp 4644–4653
53. Yang X, Lin G, Lv F, Liu F (2020) Trrnet: tiered relation reasoning for compositional visual question answering. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI16
54. Yang Z, He X, Gao J, Deng L, Smola A (2016) Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 21–29
55. Licheng Y, Lin Z, Shen X, Yang J, Xin L, Bansal M, Mattnet BTL (2018) Modular attention network for referring expression comprehension. In: Proceedings of the IEEE conference on computer vision and pattern recognition 10:1307–1315
56. Yu L, Poirson P, Yang S, Berg AC, Berg TL (2016) Modeling context in referring expressions. In: European conference on computer vision, pp 69–85. Springer
57. Zhou Y, Jun Y, Cui Y, Tao D, Tian Q (2019) Deep modular co-attention networks for visual question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 6281–6290
58. Yu Z, Yu J, Xiang C, Zhao Z, Tian Q, Tao D (2018) Rethinking diversified and discriminative proposal generation for visual grounding. In: Proceedings of the international joint conference on artificial intelligence
59. Zhang Q, Lei Z, Zhang Z, Li SZ (2020) Context-aware attention network for image-text retrieval. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 3533–3542
60. Zhang Y, Zhou W, Wang M, Tian Q, Li H (2021) Deep relation embedding for cross-modal retrieval. IEEE Trans Image Process 30:617–627
61. Zhou Y, Ji R, Sun X, Luo G, Hong X, Su J, Ding X, Shao L (2020) K-armed bandit based multi-modal network architecture search for visual question answering. In: Proceedings of the 28th ACM international conference on multimedia, pp 1245–1254
62. Zhuang Y, Song J, Fei W, Li X, Zhang Z, Rui Yong (2018) Multimodal deep embedding via hierarchical grounded compositional semantics. IEEE Trans Circ Sys Video Technol 28(1):76–89