Multimedia Systems (2022) 28:445–456
https://doi.org/10.1007/s00530-020-00745-7

SPECIAL ISSUE PAPER

Question-relationship guided graph attention network for visual question answer

Rui Liu[1] · Liansheng Zhuang[1] · Zhou Yu[2] · Zhihao Jiang[3] · Tian Bai[1]

Received: 10 September 2020 / Accepted: 26 December 2020 / Published online: 15 March 2021

Abstract

A high-level of understanding about the surrounding context of an image is indispensable for VQA when faced with difficult questions. Previous studies address this issue by modeling object-level visual contents and transforming the internal relationships into a graph or tree. On one hand, however, this still leaves a gap between the modalities of language and vision. On the other hand, the abstract-level contents of the images and the meaning of the relationships between them are ignored. This paper proposes introducing a method of question-relationship guided graph attention network (QRGAT) to study a new representation of the visual features of an image through the guidance of a question and the explicit, internal relationships of objects. Specifically, to narrow the gap between different modalities, visual regions are represented as the combination of their attributes and visual features. Meanwhile, semantic relationships are transformed into the modality of language and used to form updated visual features. The three graph encoders with diverse relationships are considered to capture high-level features of images. Experimental results of the VQA 2.0 model show that our proposed QRGAT outperforms other interpretable visual context structures.

Keywords Visual question answer · Graph attention network · Visual relationship detection

1 Introduction

Visual question answering aims at answering a question related to the content of a given image. There are two available state-of-the-art approaches that can aid in performance improvement within VQA. One area is using the better representation provided by multimodality fusion strategies, such as bilinear fusion [35], DFAF [18], MCAN [34], and MUTAN [3]. The other is by gaining a better understanding of a given image [24]. Poor comprehension of visual regions leads to a strong dependency on language modality and enhances the effects of existing textual biases. However, these models pull surface information from images. Images do not typically contain explicit expressions of meaning and inter-relationality [10, 24]. Meanwhile, the process of updating visual features cannot refer various relationships and linguistic meanings of each image. Also, it is necessary to make questions participate in the encoding process of visual features early. To overcome the problem of an insufficient understanding of the visual regions and the deficiency of high-level information within images, this paper considers extracting attributes, visual features, and relational contents of objects to get the fine-grained features of images (Fig. 1). To capture an abundance of information from images, two kinds of relationships are contained: prior knowledge relations and implicit relations. Prior knowledge relations are divided into semantic relations and spatial relations. These relations, as the novel representations of an image, capture the location, action, state, and implicit information. Semantic

✉ Liansheng Zhuang
lszhuang@ustc.edu.cn

Rui Liu
ruilbwa@mail.ustc.edu.cn

Zhou Yu
yuz@hdu.edu.cn

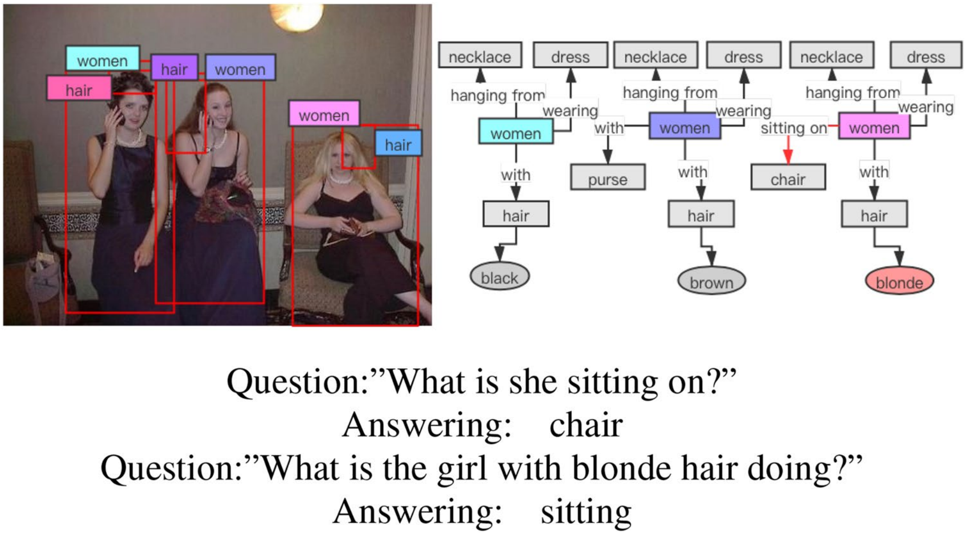
Zhihao Jiang
jzh1005@163.com

Tian Bai
baitian@ustc.edu.cn

[1] University of Science and Technology of China, Hefei, China

[2] Hangzhou Dianzi University, Hangzhou, China

[3] Information Engineering University of the People's Liberation Army, Zhengzhou, China



Question:"What is she sitting on?"
Answering: chair
Question:"What is the girl with blonde hair doing?"
Answering: sitting

Fig. 1 In the left figure, there are some objects recognized by the model. In the right figure, there is a part of relationships and attributes of the visual elements. QRGAT recognizes the semantic relation between woman and chair, and the attribute of the hair (blonde) and gives true answers

relations are transformed into vectors with the same encoders to narrow the gap between the different modalities. With a knowledge graph available for consultation, the graph is used to construct the connections among the image regions. Because the related and important objects and relationships should be paid attention to, a modified graph attention network is considered. Question participates in the interaction of relations and visual features in the early step. The high-level features of relations and visual regions are obtained in accordance with their neighbors. Compared with ReGAT, this paper uses individual attributes and classes as part of the representation of an object. Meanwhile, a better encoding method is applied to the semantic relations which guide the updating process of visual features. Questions participate in the encoding processes with different graph encoders. Also, there are interactions among nodes and relations. These methods help the model capture significant relations and objects within a specific question while considering the properties of objects. However, this method does have some challenges. First, the intended purpose of the process only requires an interpretable, efficient, and effective method, depending on the graph structure, for encoding relational information and visual features. To obtain an enormous capacity for justification and a deep comprehension of the image context, the question-guided updating layer and the modified graph attention network layer with edge information are designed. In these encoders, the updating of visual features and the relational features relies on a specific task represented by a question. And the self-attention layer is used for getting abstract features to enhance the ability to communicate robustness. Second, to narrow the gap between different modalities, attributes, relations, and questions are encoded by the same encoder. Third, different relations contain diverse information for predicting true answers. This paper is only interested in some relationships and pays more attention to the specific information formed from different relation-graphs. For assigning weight accordingly, an ultimate predictor is designed to make full use of the supporting evidence provided by different relation-graphs. In principle, the gap between language and vision is reduced. Also, more information is considered within this model than other VQA systems. It explicitly provides relational information, and this information permeates the updating of visual features to enhance the understanding of images and to enrich image representations. Experiments prove that this method, with its use of attributes and graph encoders, can improve the performance of the VQA model.

The contributions of this paper are as follows: first, information flows through mutual effects of relations and among visual features based on graph networks to create abstract features. A question-guided layer for a graph-based relations encoder and a question-guided layer for a graph-based visual features encoder are proposed to focus on more meaningful and salient parts of relations and images. Second, to narrow the gap between language modality and visual modality, the attributes of visual objects are extracted as part of visual features, and the meanings of relations between visual objects are transformed into explicit features to make the model more interpretable and comprehensive. Third, this paper synthesizes different evidence provided by various graphs to predict a more integrated result answer.

2 Related work

2.1 Visual question answering

With the emergence of deep learning, the computer vision community and the NLP community have made great contributions to VQA studies. Meanwhile, many studies have been conducted on the challenges of VQA. One approach is to focus on better expression of the visual regions [8, 14, 16, 17, 22, 23, 33, 35]. To acquire richer visual information, the work of Noh et al. [16] developed a method for learning visual features based on a task-conditional classifier and unsupervised method without question information while attempting to transform this information into a VQA model. The works of Singh et al. [23] focus on the capability to read text within images and an attempt at mixing the OCR system and VQA model when answering questions. Another method endows the model with the capability to address and understand multisource modal information that also creates a better representation of a fusion feature [8, 14]. Peng et al. [18] and Yu et al. [34] think that the central goal of VQA is to learn the features of multimodal images and fuse them effectively. Their methods aim at dynamically fusing multimodal features between different modalities, which transmit both intra-modality and inter-modality information. It captures the high-level interactions between different domains. Nevertheless, these models tend to ignore the inherent semantic connections in images and the multiple steps of reasoning. Some works pay attention to a viewer's capability of reasoning [4, 10, 23, 27, 32]. Graphs and trees are used to construct connections among image regions [10, 24]. Despite existing models' attempts to enhance the capability of visual modality, these approaches lack the explicit application of semantic information and consideration of multisource information, which we address in this paper.

2.2 Attention mechanism in VQA

Attention mechanisms imitate human cognitive capabilities and try to mix internal experiments and external guided information to increase observational awareness in terms of regionality [15]. One common VQA [29] and image captioning task are to capture the most relevant regions in images. It should be noticed that the self-attention mechanism becomes a significant method in VQA for enhancing performance and is often applied in hidden states from a bidirectional LSTM of machine translation tasks. This results in an enhanced understanding of the relations in texts [12, 25]. To solve the task on the graph structure, graph attention networks emerged [26]. Using the local information of a graph, it breaks through the graph's structural limitations and gives a new ability to form generalizations and distinct processes. Nevertheless, the edge information in the graph is often ignored. For obtaining a better combination of visual features and relations information, this paper modifies the graph attention network to consider more factors within a specific task. Other researchers [11] have introduced a method to dynamically determine what media and what time to focus on in the sequential data to predict an answer. However, they have not explored the high-level information in images.

2.3 Visual relationship

Visual relationships are represented as triples ($\langle \text{object1} - \text{predicate} - \text{object2} \rangle$). Meanwhile, the tasks of visual relationship detection are explored to detect objects' positions and the relationships between objects which are used to obtain semantic relations and construct a graph over images. It is important to note that relationship is a combination of the object and the predicate. However, the set of possible relationships is larger than the set of possible objects and the set of predicates. Lu et al. [28] proposed a method that can learn the objects and the predicates separately by a visual appearance module. And the information of the objects and the predicates are fused to predict relationships in the last step. Meanwhile, a language module is used to create a vector space for relationships where similar relationships are close to each other. Some previous works [5, 6] provide methods to improve image segmentation by considering diverse relationships among objects, such as position relations. A two-stage pipeline is used by some papers to detect visual relationships between objects and to predict results for each object pair. Some [37] make great efforts toward solving entity-instance confusion, in which objects are related to many instances of the same class, and proximal-relationship ambiguity, in which subject–object pairs may have a similar connections. This paper also uses this structural analysis to extract the relationships of images. Spatial relations and implicit relations are additional sources of information utilized within this paper to predict an answer.

2.4 Graph network

Graph network plays as the product of blocks with "structured representation" to complete structured computation and to adapt to both Euclidean and non-Euclidean data [37]. It is always one of important research tasks of deep learning, including graph convolutional network, graph attention network, graph auto-encoder, graph generative network, and graph spatial–temporal network [13]. As for graph convolutional network, it can be divided into two categories: spectral-based method and spatial-based method. Graph attention network is an application of attention mechanism on graphs. Graph auto-encoder is a method of graph embedding which maps nodes into lower dimensional feature space and decodes graph information from that. Graph generative network aims to generative new graphs from the hidden representations of the given graphs. Time dimension is introduced into graph spatial–temporal network to capture the dynamicity of graphs. Graph network is widely used in reasoning. Some [2, 36] make great efforts toward combining end-to-end neural networks and inductive reasoning. Neural networks are one of the functions needed to complete relation reasoning within graph networks. However, it includes a good deal of complex neural networks of which structures are difficult to decide. It does not take into account the factors out of graph network in reasoning either. Some [7] propose a probabilistic graph that represents underlying semantics. It regarded as Neural State Machine to execute reasoning under the guidance of question. Nevertheless, the global structure of the graph is indispensable. The meaning of words within relations is not considered. Meanwhile, there are no interactions among nodes.

3 Question-relationship guided graph attention network

Visual question answering (VQA) is a complex task that involves the computer vision field and the natural language field. Given an image and a question, it requires neural networks to infer the correct answer through understanding and reasoning over visual regions and texts.

To capture high-level features, information flows through a mutual effect based on the graph network. A question joins the updating process early on. This section describes the approach to implement the question-relationship guided graph attention network in Fig. 2. To synthesize multisource information, two types of relations are considered to execute a specific task. Attributes and classes are extracted as part of the representations of the visual regions that help the attention mechanism capture the most relative visual objects of the question. Because of the same way in encoding to the questions, attributes, and classes, the attention mechanism becomes more explicable. Prior knowledge relations are divided into semantic relations and spatial relations. A semantic graph transforms relations into language modality to narrow the gap between different modalities. Three graph encoders learn individually fused features of different relations to obtain diverse information for predictions. This process is proven to be effective and available for improving performance, and reduces the difficulty to directly extract high-level and implicit information of a VQA model, which contains an image encoder, a language encoder, three graph encoders, and an answer predictor. This model takes a set of k visual features, the relations between objects, attributes, classes, and bounding boxes of objects as inputs. The details of these components are laid out below.

3.1 Visual and language feature extraction

To represent the visual features, this paper takes the method of bottom–up and top–down attention model [1]. The visual region features are obtained from Faster RCNN [20] model (Resnet101 as the backbone). For each image, 36 region proposals and their related bounding boxes are extracted. The obtained visual features are denoted as $\mathbf{V} \in \mathbb{R}^{k \times 2048}$, where a visual feature is denoted as $\mathbf{v_i}$, and there are k regions in total ($\mathbf{V} = \{\mathbf{v_1}, \mathbf{v_2} \ldots \mathbf{v_k}\}$). Bounding box $\mathbf{b} = [x_1, y_1, x_2, y_2]$ represents a 4-dimensional spatial coordinate, where (x_1, y_1) denotes the coordinate of the top-left point and (x_2, y_2) denotes the coordinate of the bottom-right point. All bounding boxes are normalized.

A bidirectional RNN with Gated Recurrent Unit (GRU) is adopted for GLove word embedding of question-word features. A question is denoted as $\mathbf{q} \in \mathbb{R}^{d_q}$. All questions are padded or truncated to the length 14.

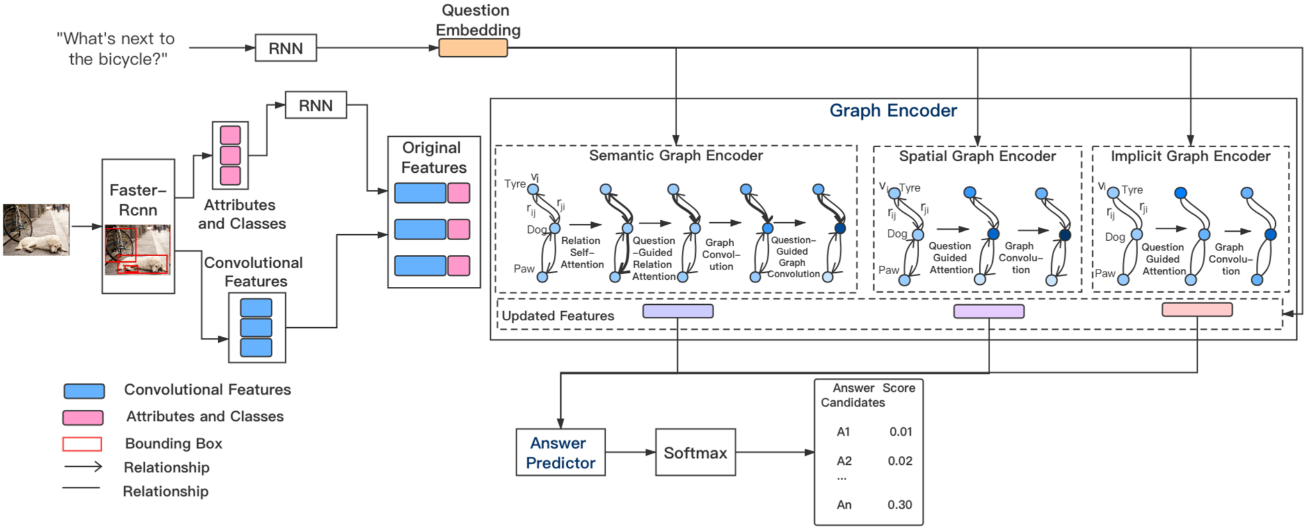


Fig. 2 An overview of GRGCAT model. It contains three kinds of graph encoder. Visual features, attributes, bounding boxes, and classes are extracted from Faster RCNN. Relations are detected for updating visual features. Putting them into graph encoder to learn question-relation guided visual features, which will be fused with linguistic features and then will be put into answering predictor

3.2 Visual relationship and attribute extraction

To obtain semantic relationship features, this paper uses the method in [37] to extract relations in images. There are 51 kinds of relations (e.g., walking in, watching, wearing, and sitting on), including a no-relation class retained for entities pairs. Each $\mathbf{r_{ij}}$ denotes the relationship between subject i and object j, where i and j both denote the entities detected by the architecture. Glove word embedding $\mathbf{e_r}$ of each relation word is adopted as the input of the RNN (in this paper, GRU is used) to encode relationships, which is the same as the question encoder. $\mathbf{R} \in \mathbb{R}^{k\times k\times \mathrm{dim}}$ denotes the matrix of relations, where $\mathbf{r_{ij}}$ denotes the relation. This shows that visual region i has the relationship $\mathbf{r_{ij}}$ with visual region j. All relation words also are padded or truncated to the same length 2. This paper adopts a bottom–up and top–down attention model [1] for getting attributes. Attributes (including classes) are obtained from a classifier that takes the visual features from Faster RCNN as inputs. Because each object may have more than one attribute, and each attribute may have more than one word, an RNN is considered. The treatment of attributes (including classes) is the same as the semantic relationships.

3.3 Graph construction

Each object $\mathbf{v_i'}$ is regarded as a vertex. A visual feature $\mathbf{v_i}$ and a attribute feature $\mathbf{a_i}$ are concatenated as a new vertex feature $\mathbf{v_i'}$.

3.3.1 Implicit graph

$\mathbf{G_{img}} = \{\mathbf{V}', \mathbf{E}\}$ is a complete graph of all visual regions. Each vertex has an undirected edge with any other vertex. The edge represents the relation between two objects that cannot be explicitly expressed. This graph demonstrates an abstract feature for every vertex through graph convolution as shown in the following section and is regarded as a supplement for other relationship graphs.

3.3.2 Semantic graph

$\mathbf{G_{sem}} = \{\mathbf{V}', \mathbf{E}\}$ is constructed referring to relationships R extracted from VQA v2 dataset by [37] which is pretrained on a visual relationship dataset (Visual Genome). If a relation between two objects is attainable, according to $\mathbf{R}$, there will be a corresponding edge between them. Different from previous works, for each pair of objects i and j, if $\langle i - r - j \rangle$ (e.g., girl eating cake, man sitting on the chair) is a valid relation, an explicit feature of relation encoded by the aforementioned way is given for this edge, which will then take effect with the updating of $\mathbf{v_i'}$. It should be noticed that object i has a relation $\mathbf{r}$ with object j, but object j does not necessarily have a relation with object i. It also indicates that the graph is directed. It proves that the relation features would help to increase the understanding of images. Examples of semantic relations are shown in Fig. 3

3.3.3 Spatial graph

$\mathbf{G_{spa}} = \{\mathbf{V}', \mathbf{E}\}$, the method of construction is the same as the semantic graph. The spatial relation $\langle i - r - j \rangle$ denotes the position connections between objects. The key to constructing a spatial graph is building a position matrix of all detected objects. 11 different classes for different layouts like the way proposed by [31] are defined. A no-relation class is set in the situations where the objects are too far away from each other during classification. Of course, spatial relation $\langle i - r - j \rangle$ and spatial relation $\langle j - r - i \rangle$ are not identical, but the two relations must be simultaneously valid. Both in semantic graph relations and spatial graph relations, it is interchangeable for subjects and objects, illustrating that the edges of the graph are not symmetric.

3.4 Graph encoder

For the sake of utilizing the visual contents, three kinds of relations are considered for capturing the locations, the actions, and the states. The semantic graph encoder is used for extracting semantic relations, such as the states and the actions of instances. The spatial graph encoder is used for extracting the position relations between objects in the layout of the image. The implicit graph encoder is a supplement with implicit relations.

For each graph as follows, each visual feature is transformed into a new feature by the linear operation. Every vertex is denoted as a new feature by concatenating the visual feature with the corresponding attribute, and represented as:

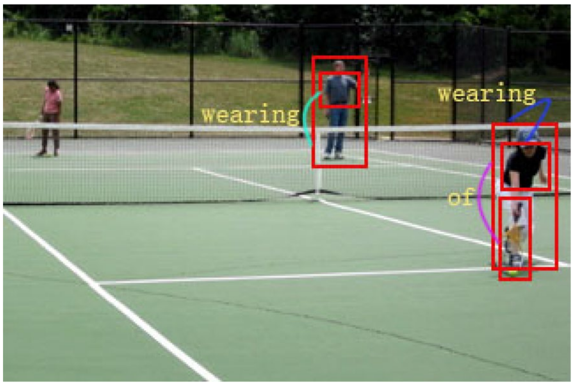


Fig. 3 There are the examples of the semantic relations detected by the model. There are two relations: man wearing skirt and racket of man

$$\mathbf{v}_{\mathbf{i}}' = [\mathbf{v}_{\mathbf{i}}||\mathbf{a}_{\mathbf{i}}] \quad \text{for} \quad i = 1, \dots, k, \tag{1}$$

where $\mathbf{v}_{\mathbf{i}}$ is the ith node in $\mathbf{V} \in \mathbb{R}^{k\times\dim}$ which denotes visual features. $\mathbf{a}_{\mathbf{i}} \in \mathbb{R}^{\dim}$ is the ith node in $\mathbf{A} \in \mathbb{R}^{k\times\dim}$ which represents attribute features.

3.4.1 Notation

This paper denotes scalars, vectors, and tensors using lower case, bold lower case, and bold upper case letters. The details are shown in Table 1.

3.4.2 Semantic graph encoder

Figure 2 gives a detailed illustration of semantic graph encoder. The semantic graph encoder learns to capture the important and relative relation between each pair of visual regions by the guidance of the question. This paper proposes four layers to encode the semantic graph: a relation self-attention layer, a question-guided relation attention layer, a vision self-attention layer, and a question-guided graph attention layer. It makes a deep excavation of the information of relations and transforms this information into visual features. The information on relations and questions would guide the model learn weights of importance and aggregate features to update each visual feature. The relation self-attention layer learns to capture the influence between different connections. In terms of a relationship $\mathbf{r}_{\mathbf{ij}}$, this paper considers all relations including subject $\mathbf{v}_{\mathbf{i}}'$ (considering the edges which come out from subject $\mathbf{v}_{\mathbf{i}}$) and explores their effect on this relationship $\mathbf{r}_{\mathbf{ij}}$. Each relation feature is transformed into query, key, and value feature as follows [21, 30]:

$$\alpha^{ij}_{im} = (\mathbf{U}^{\mathbf{sem}}_{\mathbf{1}}\mathbf{r}_{\mathbf{ij}})^{\mathrm{T}}\mathbf{V}^{\mathbf{sem}}_{\mathbf{1}}\mathbf{r}_{\mathbf{im}}, \tag{2}$$

$$\alpha^{ij}_{im} = \frac{\alpha^{ij}_{im}}{\sqrt{d_r}}, \tag{3}$$

$$(\alpha^{ij}_{im})' = \frac{\exp(\alpha^{ij}_{im})}{\sum_{n\in N_i}\exp(\alpha^{ij}_{in})}, \tag{4}$$

$$\mathbf{r}^{\mathbf{s}}_{\mathbf{ij}} = \sum_{m\in N_i}(\alpha^{ij}_{im})'\mathbf{W}^{\mathbf{sem}}_{\mathbf{1}}\mathbf{r}_{\mathbf{im}}, \tag{5}$$

where $\mathbf{U}^{\mathbf{sem}}_{\mathbf{1}}, \mathbf{V}^{\mathbf{sem}}_{\mathbf{1}} \in \mathbb{R}^{\dim\times d_r}$are projection matrices. Self-attention is then performed on the relations. α^{ij}_{im} is computed by scaled dot product. This way is adopted in the following sections. The equation will be omitted in the following content. This paper uses this information flows to update relation features as $\mathbf{r}^{\mathbf{s}}_{\mathbf{ij}}$. Specifically, α^{ij}_{im} represents the similarity between relation features and N_i are the neighbors of node i. m represents the subscript of the neighborhood of i. For the question-guided relation attention layer, it learns to capture the most relative and salient relations to extract new features. Given a question $\mathbf{q}$ and relation features as inputs, $\mathbf{r}^{\mathbf{q}}_{\mathbf{ij}}$ is a new relation feature:

$$\beta_{ij} = (\mathbf{U}^{\mathbf{sem}}_{\mathbf{2}}\mathbf{q})^{\mathrm{T}}\mathbf{V}^{\mathbf{sem}}_{\mathbf{2}}\mathbf{r}^{\mathbf{s}}_{\mathbf{ij}}, \tag{6}$$

$$\beta_{ij}' = \frac{\exp(\beta_{ij})}{\sum_{0<m,n<k}\exp(\beta_{mn})}, \tag{7}$$

$$\mathbf{r}^{\mathbf{q}}_{\mathbf{ij}} = \beta_{ij}'\mathbf{W}^{\mathbf{sem}}_{\mathbf{2}}\mathbf{r}^{\mathbf{s}}_{\mathbf{ij}}, \tag{8}$$

where β_{ij}' represents the correlation between the relation feature $\mathbf{r}^{\mathbf{s}}_{\mathbf{ij}}$ and the question $\mathbf{q}$. The vision self-attention layer calculates the association weights between every pair of visual features and is endowed with the ability to extract abstract information. The process is similar to the relation self-attention layer, so this paper only provides equations of vision self-attention layer:

$$\gamma_{ij} = (\mathbf{U}^{\mathbf{sem}}_{\mathbf{3}}\mathbf{v}_{\mathbf{i}}')^{\mathrm{T}}\mathbf{V}^{\mathbf{sem}}_{\mathbf{3}}\mathbf{v}_{\mathbf{j}}', \tag{9}$$

Table 1 The descriptions of notations

Name	Description
$\mathbf{U}$	A tensor as a projection matrix
$\mathbf{V}$	A tensor as a projection matrix
$\mathbf{W}$	A tensor as a projection matrix
$\mathbf{Y}$	A tensor as a projection matrix
$\mathbf{X}$	A tensor as a projection matrix
α^{ij}_{im}	The similarity between the $\mathbf{r}_{\mathbf{ij}}$ and the $\mathbf{r}_{\mathbf{im}}$
α_i	The correlation between the visual feature $\mathbf{v}_{\mathbf{i}}$ and the question $\mathbf{q}$
α^{v}_{ij}	The similarity between the $\mathbf{v}_{\mathbf{i}}$ and the $\mathbf{v}_{\mathbf{j}}$
α_{ij}	A scalar of weights
β_{ij}	A scalar of weights
γ_{ij}	A scalar of weights
δ_{ij}	A scalar of weights
k	The number of visual regions
$\mathbf{r}_{\mathbf{ij}}$	The relationship between the region i and the region j
$\mathbf{v}_{\mathbf{i}}$	The visual region i
$\mathbf{q}$	The question
$\mathbf{h}$	The fused feature
Ua	The query
Va	The key
Wa	The value

$$\gamma'_{ij} = \frac{\exp(\gamma_{ij})}{\sum_{m\in N_i}\exp(\gamma_{im})}, \tag{10}$$

$$\mathbf{v_i^s} = \sum_{j\in N_i}\gamma'_{ij}\mathbf{W_3^{sem}}\mathbf{v'_j}, \tag{11}$$

This paper argues that relationships and question information are complementary to the process of updating of visual features, and should be taken into account for the model. For example, "Are the bird's legs touching the water", the model should understand the relation "touching", and locate relative regions. Thus, a question-guided graph attention layer is proposed. It compromises the visual contents and relation contents, and guides the updating of visual features. This layer is modified on the conventional attention mechanism:

$$\delta_{ij} = (\mathbf{U_4^{sem}q})^{\mathrm{T}}\mathbf{V_4^{sem}}[\mathbf{v_i^s}, \mathbf{r_{ij}^q}], \tag{12}$$

where $\mathbf{U_4^{sem}} \in \mathbb{R}^{\dim\times d_q}$, $\mathbf{V_4^{sem}} \in \mathbb{R}^{\dim\times(d_r+d_v)}$ are projection matrices. The rest equations are similar to the vision self-attention layer. This paper final encodes visual feature as $\mathbf{v_i^q} \in \mathbb{R}^{\dim}$. In contrast to the conventional graph attention network, this approach of encoding visual features considers the influence of relation and question on visual features, which also effectively takes advantage of the relation contents. It learns different weights of importance to nodes according to the relevance of a specific question. The outputs would then be fed into a multimodal fusion module.

3.4.3 Spatial graph encoder

This encoder adopts a simpler structure which includes two layers: a question-guided graph attention layer and a graph attention convolution layer. It encodes visual features which guided by a question. The question-guided graph attention layer only considers question $\mathbf{q}$ and visual features as inputs. The operation mode of this layer is similar to the question-guided relation attention layer of the semantic graph encoder:

$$\alpha_i = (\mathbf{U^{spa}q})^{\mathrm{T}}\mathbf{V^{spa}v'_i}, \tag{13}$$

$$\alpha'_i = \frac{\exp(\alpha_i)}{\sum_{0<j<k}\exp(\alpha_j)}, \tag{14}$$

$$\mathbf{v_i^q} = \alpha_i^s\mathbf{W^{spa}v'_i}, \tag{15}$$

where α'_i represents the correlation between the visual feature $\mathbf{v_i}$ and question $\mathbf{q}$. It is easy to explicate that a bigger weight is assigned to the visual feature which is more relevant to a specific task. The graph attention convolution layer would consider the spatial relation information. Every edge in the spatial graph is encoded by one-hot. Each edge is transformed into another vector space. This layer is endowed with the sensitive capacity both for different layout between subject and object and for capturing the important parts of visual features. This paper omits the equation of β_{ij}; specifically:

$$\beta'_{ij} = \frac{\exp(\beta_{ij} + \mathbf{Xr_{ij}^{spa}})}{\sum_{m\in N_i}\exp(\beta_{im} + \mathbf{Xr_{im}^{spa}})}, \tag{16}$$

$$\mathbf{v_i^s} = \sum_{j\in N_i}\beta'_{ij}(\mathbf{Wv_j^q} + \mathbf{Yr_{ij}^{spa}}), \tag{17}$$

where $\mathbf{U}, \mathbf{V}, \mathbf{W}, \mathbf{X}, \mathbf{Y}$ are projection matrices, $\mathbf{r_{ij}^{spa}}$ and $\mathbf{r_{ij}^{spa}} \in \mathbb{R}^{d_{\text{label}}}$. $\mathbf{r_{ij}^{spa}}$ represents the spatial information of each edge. The outputs contain the prior spatial relations information between each pair of objects through the spatial graph encoder.

3.4.4 Implicit graph encoder

It is similar to the spatial graph encoder, including two layers. Comparing with the spatial graph encoder, this paper only changes the equation to calculate the weights of different visual features like [10]:

$$\alpha_{ij}^v = (\mathbf{U^{imp}v_i^q})^{\mathrm{T}}\mathbf{V^{imp}v_j^q}, \tag{18}$$

$$\alpha_{ij}^b = \max\{0, \mathbf{w}\cdot f_b(\mathbf{b_i}, \mathbf{b_j})\}, \tag{19}$$

$$\alpha'_{ij} = \frac{\alpha_{ij}^b\exp(\alpha_{ij}^v)}{\sum_{m\in N_i}\alpha_{im}^b\exp(\alpha_{im}^v)}, \tag{20}$$

where f_b computes a 4-dimensional relative geometry feature, and computes cosine and sine functions of different wavelengths. Then, this approach transforms them into a new feature $\mathbf{d_h}$. And $\mathbf{w} \in \mathbb{R}^{d_h}$ projects the new feature into a scalar weight. In all graph encoders, multihead attention is adopted to improve the architecture performance. It means that there are M independent attention mechanisms simultaneously work. Then, the results of every attention mechanism are concatenated.

$$\mathbf{v_i} = ||_{m=1}^{M}\sigma\Big(\sum_{j\in N_i}\alpha_{ij}\mathbf{Wv_j}\Big), \tag{21}$$

3.5 Multimodal fusion and answer prediction layer

After each graph encoder process, the outputs are considered to contain information about different graphs and relations between each pair of objects. Then, the multimodal fusion approach is applied to fuse various features, including linguistic features and visual features. The same operation is executed on every graph. Through the multimodal fusion strategy, a fused feature **h** is learned by a neural network:

$$\mathbf{h} = f(\mathbf{v}^{*}, \mathbf{q}; \theta), \tag{22}$$

where f denotes a fusion approach, and $\mathbf{v}^{*}$ represents the final visual feature after a graph encoder. So far, diverse relation information from different graphs is obtained. The final features of different graphs are fed into a 2-layer multilayer perceptron and then are fed into a softmax function. The loss function in the model is cross-entropy. To achieve selecting different source information, a 2-layer multilayer perceptron (MLP) is employed to learn the weights of information from different sources obtained by three graphs:

$$\mathbf{p}^{'} = \mathrm{MLP}(\mathbf{p}^{\mathbf{sem}}, \mathbf{p}^{\mathbf{spa}}, \mathbf{p}^{\mathbf{imp}}; \theta), \tag{23}$$

where $\mathbf{p}^{'} \in \mathbb{R}^{\mathrm{dim}}$ represents the result of the whole model. $\mathbf{p} \in \mathbb{R}^{\mathrm{dim}}$ denotes the result of a single graph.

4 Experiment

4.1 Setting

The proposed VQA model is evaluated on VQA 2.0 and VQA-CP v2. VQA 2.0 contains numerous varied real images from MSCOCO images. Human-annotated questions and answers based on images are provided by VQA 2.0. There are three questions provided in an image and ten answers provided per question. The ground-truth accuracy of a candidate answer is the average of min $(\frac{\#\text{Humans} \quad \text{votes}}{3}, 1)$ over all 10 select 9 sets. Three types of questions are contained in VQA 2.0 (yes/no, number, and other). VQA-CP v2 dataset is a derivation of the VQA 2.0 dataset. The distributions of answers in training and test splits are different.

The hyper-parameters of this paper in the experiments are as follows: visual features that are extracted from Faster RCNN with the backbone of Resnet101 have 2048 dimensions, while each word of the question, relationship, and attribute is embedded by 600 dimensions. The sentence of a question is fixed length at 14 by padding or truncating. The key, query, and value vectors are transformed into 1024 dimensions for the implicit graph and the spatial graph (512 dimensions for the semantic graph on the VQA-CP dataset). All graph encoders adopt 4 multihead attention with 256 (128) dimensions for each head. Three fusion strategies are adopted to fuse visual features and linguistic features: bottom–up top–down (BUTD) [1], multimodal tucker fusion (MUTAN) [3], and bilinear attention network (BAN) [8]. Adam solver [9] with beta1 = 0.9 and beta2 = 0.999 is used to train the model. The base learning rate is set to 0.001. This paper takes a gradual warm-up learning rate: 0.5*base learning rate, 1.0*base learning rate, 1.5*base learning rate, and 2.0*base learning rate at the first 4 epochs. 2.0*base learning rate is kept in the next epochs, and the learning rate decays by 0.25 every 2 epochs after 16 epochs. Full connected layers have a dropout rate of 0.2. The environment of experiment is Pytorch platform, and the variables are initialized to default value.

4.2 Ablation studies

All ablation studies are conducted on the validation split of VQA 2.0. These ablation studies explore the influence of different layers and different modules for three graph encoders. The results are shown in Table 2.

4.2.1 Implicit graph encoder and spatial graph encoder

To investigate the effect on the question-guided graph attention layer (GGAL) and the graph attention convolution layer (GACL) on the implicit graph encoder and the spatial graph encoder, comparative experiments are executed with the different layers. The fusion strategy is BUTD [1]. From the results, the model with both of two layers outperforms other structures on validation. It proves that the GGAL improves the performance of the VQA model and means that question information is a significant complement for implicit graph encoder when there is no additional information. Meanwhile, the positive influence of the modified graph attention network is certified. When the GACL is built into the structure, the performance of the VQA model simultaneously improves.

Table 2 Results of ablation studies

Graph Encorder	Component	Accuracy (%)
Implicit(BUTD)	W/o question-guided layer	63.38
Implicit(BUTD)	W/ question-guided layer	64.02
Spatial(BUTD)	W/o question-guided layer	63.84
Spatial(BUTD)	W/ question-guided layer	64.05
Semantic(BUTD)	W/o relation self-attention	63.67
Semantic(BUTD)	W/o relation-question attention	63.53
Semantic(BUTD)	Only w visual self-attention	63.36
Semantic(BUTD)	Complete network	63.76
Semantic(ReGAT)	MLP as pretrained network	65.59
Semantic(BAN)	MLP as pretrained network	65.75
Semantic(BAN)	ContrastiveLosses4VRD as pretrained network	65.92
Semantic(BAN)	W/o attributes	65.26
Semantic(BAN)	W/ attributes	65.92

4.2.2 Semantic graph encoder

This paper investigates the influence of 4 layers in the semantic graph encoder. The default model contains all four layers to predict an answer. The first two layers are tried to abandon separately in the model. The fusion strategy is BUTD [1]. Using all the four layers, performance can improve by 0.4%. It notices that when the module of the layers about relations is abandoned, the performance of the VQA model decreases. There is a reasonable conjecture that the semantic graph encoder has captured the information of semantic relations in images and then passes such information into predictor. The attention mechanism also is proved to have an immense effect on semantic graph encoder. With the same pretrained network (MLP), the semantic graph encoder gets better performance than the ReGAT [10].

4.2.3 Attributes

This paper compares the results between the approach with attributes (including classes) and the approach without attributes (including classes). The results are in Table 2. A semantic graph encoder with a fusion strategy of BAN [8] is adopted. The number of objects is fixed at 36. By adding the attribute features, performance improves by 0.66%. There is a reasonable conjecture that the introduction of attributes reduces the gap between visual modality and linguistic modality.

4.3 Baselines

By evaluating the performance of this model, this paper compares the results with some baselines. The results are in Table 3. Bottom–up top-down (BUTD) [1], multimodal tucker fusion (MUTAN) [3], dynamic tree structures (VCTREE-HL) [24], bilinear attention networks (BAN) [8], relation-aware graph attention (ReGAT) [10], DFAF [18], and multimodal relational reasoning (MuRel) [19] methods are considered as baselines. They proposed different schemes to address the task of VQA. BUTD [1] thinks the salient parts of the image should be paid more attention to. DFAF [18], MUTAN [3], and BAN [8] aim to find better multimodality fusion approaches. VCTREE-HL [24] and ReGAT [10] focus on structuring the spatial relationship and explicit relationship in images to enhance the understanding of pictures. For modeling complex reasoning features for high-level tasks, MuRel [19] introduces the MuRel cell [19] to reason the interactions between question and image regions by a rich vectorial representation. The results are on the VQA 2.0 validation. Imp/Spa/Sem means a single type of implicit, semantic, or spatial relation. When compared with the baselines of BUTD [1] and MUTAN [3], it reports the single-type graph encoder with BUTD [1] and MUTAN [3] delivers the best performance. The total result of graph encoders has the best performance.

This paper conducts experiments on the VQA-CP v2 dataset. MuRel [19], BUTD [1], BAN [8], and ReGAT [10] methods which have the stronger reasoning ability are considered as baselines. Table 4 shows the results on the test split. The model in this paper surpasses the baselines by a large margin. With only a single graph encoder, the model achieved the best performance on baselines (41.24 vs 41.17).

Table 3 Model accuracy on the VQA 2.0 validation

Model				Accuracy (%)
BUTD [1]				63.15
MUTAN [3]				58.16
MUTAN+MLB [3]				58.76
BAN [8]				65.36
VCTREE-HL [24]				65.1
ReGAT(BUTD fixed) [10]				64.98
ReGAT(BAN fixed) [10]				66.62
ReGAT(BAN ada) [10]				67.18
DFAF [18]				66.66
MCAN [34]				67.2
	Semantic (%)	Implicit (%)	Spatial (%)	All
Ours (MUTAN fixed)	62.34	62.27	62.14	n/a
Ours (BUTD fixed)	63.76	64.05	64.02	65.23 %
Ours (BAN fixed)	65.92	66.39	65.97	**67.30** %

The bold letters denote the best results of the experiments

Table 4 Model accuracy on the VQA-CP v2 benchmark

Model				Accuracy (%)
Bottom-up [1]				38.01
BAN [8]				39.31
MuRel [19]				39.54
ReGAT [10]				40.42
UpDn+ Q-Adv+DoE [38]				41.17
	Semantic (%)	Implicit (%)	Spatial (%)	All
Ours (BAN fixed)	39.88	41.24	40.76	**42.50** %

The bold letters denote the best results of the experiments

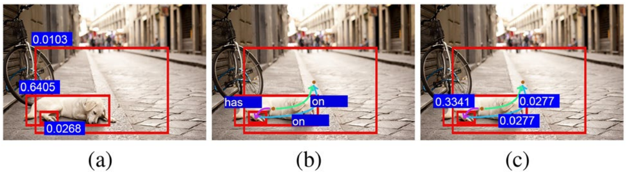


Fig. 4 **a** Visual attention weights, **b** relations between pairs, **c** correlation of relations with question. Question: Where is the dog laying? The answer is sidewalk , and the ground truth is sidewalk

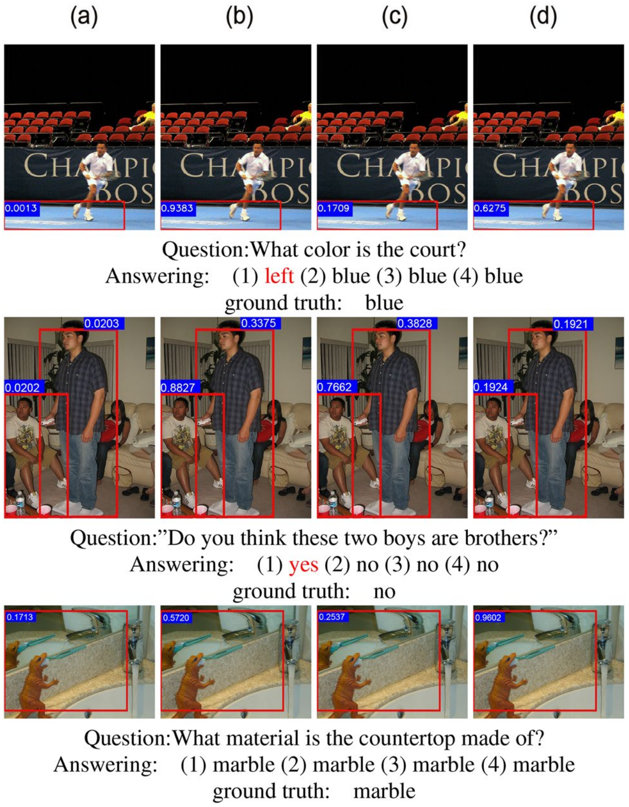

(a)
(b)
(c)
(d)
0.0013
0.9383
0.1709
0.6275
Question:What color is the court?
Answering: (1) left (2) blue (3) blue (4) blue
ground truth: blue
0.0203
0.0202
0.3375
0.8827
0.3828
0.7662
0.1921
0.1924
Question:"Do you think these two boys are brothers?"
Answering: (1) yes (2) no (3) no (4) no
ground truth: no
0.1713
0.5720
0.2537
0.9602
Question:What material is the countertop made of?
Answering: (1) marble (2) marble (3) marble (4) marble
ground truth: marble

Fig. 5 **a** Graph attention, **b** semantic graph encoder, **c** implicit graph encoder, and **d** spatial graph encoder. There are different weights of objects in four methods

4.4 Visualization analysis

In Fig. 4, this paper visualizes the relations, which are detected and used by this model, to illustrate the effectiveness of the semantic graph encoder. It shows how detected relations help to improve the performance of the model. When relations are valid for visual regions, the model is more inclined to give more weight on the relations related to question.

Figure 5 provides the weights at the runtime and illustrates the effect of the different graph encoders. Comparing different methods, it shows that the graph encoders in this paper help model to locate the important objects, capture the interactions between regions, and assist with providing a better alignment between different regions. These examples give new evidence proving that the question-guided attention layers play a considerable role in graph encoders.

5 Conclusion

In this paper, question-relationship guided graph attention neural Network (QRGAT) is proposed for VQA. More detailed divisions are carried on to the image contents, while graph encoders, as the keywords, are contained in this model to capture the actions, states, layout, and implicit relations in the images. One relation influences the other relation in reasoning and encoding process. In addition, the question is offered as well as a guider role in the updating process of visual features. It reduces the gap between language modality and relation modality. Meanwhile, it tries to obtain a new abstract-level feature over a relation graph. In the future, we intend to study an end-to-end network, including different relations, and find a more effective and efficient way to precisely capture and combine diverse information of the images.

Acknowledgements This work was supported in part to Dr. Liansheng Zhuang by NSFC under contract No. U20B2070 and No. 61976199, and in part to Dr. Houqiang Li by NSFC under contract o. 61836011, Next Generation AI Project of China no. 2018AAA0100602.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6077–6086 (2018)
2. Battaglia, P.W., Hamrick, J.B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al.: Relational inductive biases, deep learning, and graph networks. arXiv:1806.01261 (2018)
3. Ben-Younes, H., Cadene, R., Cord, M., Thome, N.: Mutan: Multimodal tucker fusion for visual question answering. 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2631–2639 (2017)
4. Cadene, R., Ben-Younes, H., Cord, M., Thome, N.: Murel: Multimodal relational reasoning for visual question answering. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1989–1998 (2019)
5. Dai, B., Zhang, Y., Lin, D.: Detecting visual relationships with deep relational networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3298–3308 (2017)
6. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8359–8367,18-23 (2018)
7. Hudson, D.A., Manning, C.D.: Learning by abstraction: the neural state machine. Conference on Neural Information Processing Systems, pp. 5871–5884 (2019)
8. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear attention networks. 32nd Conference on Neural Information Processing Systems (NeurlPS), pp. 1564–1574 (2018)
9. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
10. Li, L., Gan, Z., Cheng, Y., Liu, J.: Relation-aware graph attention network for visual question answering. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1031–10321 (2019)
11. Liang, J., Jiang, I., Cao, L., Kalantidis, Y., Li, L.J., Hauptmann, A.G.: Focal visual-text attention for memex question answering. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 1 (2019)
12. Lin, Z., Feng, M., Santos, C.N.d., Yu, M., Xiang, B., Zhou, B., Bengio, Y.: A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130 (2017)
13. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. ECCV 2016:14th European Conference on Computer Vision, pp. 852–869 (2016)
14. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. 30th Annual Conference on Neural Information Processing Systems (NeurlPS), pp. 289–297(2016)
15. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
16. Noh, H., Kim, T., Mun, J., Han, B.: Transfer learning via unsupervised task discovery for visual question answering. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8377–8386 (2019)
17. Norcliffe-Brown, W., Vafeias, S., Parisot, S.: Learning conditioned graph structures for interpretable visual question answering. 32nd Conference on Neural Information Processing Systems (NeurlPS), pp. 8834–9343(2018)
18. Peng, G., Jiang, Z., You, H., Lu, P., Hoi, S., Wang, X., Li, H.: Dynamic fusion with intra-and inter-modality attention flow for visual question answering. arXiv:1812.05252 (2018)
19. Remi, C., Hedi, B.-Y., Cord, M., Thome, N.: Murel: multimodal relational reasoning for visual question answering. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1989–1998 (2019)
20. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149 (2017)
21. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. arXiv preprint arXiv:1803.02155 (2018)
22. Shih, K.J., Singh, S., Hoiem, D.: Where to look: Focus regions for visual question answering. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4613–4621 (2016)
23. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8309–8318 (2019)
24. Tang, K., Zhang, H., Wu, B., Luo, W., Liu, W.: Learning to compose dynamic tree structures for visual contexts. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6612–6621 (2019)
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. 31st International Conference on Neural Information Processing Systems, vol. 30, pp. 5998–6008 (2017)
26. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. 2018 International Conference on Learning Representations (2018)
27. Wu, C., Liu, J., Wang, X., Dong, X.: Chain of reasoning for visual question answering. 32nd International Conference on Neural Information Processing Systems, vol. 31, pp. 275–285 (2018)
28. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P.S.: A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 1, pp. 4–24 (2020)
29. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: neural image caption generation with visual attention. 32nd International conference on machine learning (ICML), volume 3 of 3, pp. 477–499 (2015)
30. Xu, K., Wang, Z., Shi, J., Li, H., Zhang, Q.C.: A2-net: molecular structure estimation from cryo-em density volumes. Proc. AAAI Conf. Artif. Intell. **33**, 1230–1237 (2019)
31. Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring visual relationship for image captioning. ECCV 2018: 15th European Conference on Computer Vision, pp. 711–727 (2018)
32. Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., Tenenbaum, J.: Neural-symbolic vqa: disentangling reasoning from vision and language understanding. Advances in Neural Information Processing Systems, vol. 31, 2018, pp. 1031–1042 (2017)
33. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: a dataset for understanding complex web videos via question answering. Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 1, pp. 9127–9134 (2019)
34. Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6274–6283 (2019)
35. Yu, Z., Yu, J., Xiang, C., Fan, J., Tao, D.: Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans. Neural Netw. Learn. Syst. **29**(12), 5947–5959 (2018)

36. Zhang, C., Chao, W.L., Xuan, D.: An empirical study on leveraging scene graphs for visual question answering. 2018 British Machine Vision Conference (BMVC), p. 288 (2018)

37. Zhang, J., Shih, K.J., Elgammal, A., Tao, A., Catanzaro, B.: Graphical contrastive losses for scene graph parsing. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11527–11535 (2019)

38. Yang, Z., He, X., J.L.A.: Stacked attention networks for image question answering. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21–29 (2016)

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