The Visual Computer (2023) 39:5783–5795
https://doi.org/10.1007/s00371-022-02695-9

ORIGINAL ARTICLE

Multi-modal co-attention relation networks for visual question answering

Zihan Guo[1] · Dezhi Han[1]

Accepted: 4 October 2022 / Published online: 29 October 2022

Abstract
The current mainstream visual question answering (VQA) models only model the object-level visual representations but ignore the relationships between visual objects. To solve this problem, we propose a Multi-Modal Co-Attention Relation Network (MCARN) that combines co-attention and visual object relation reasoning. MCARN can model visual representations at both object-level and relation-level, and stacking its visual relation reasoning module can further improve the accuracy of the model on Number questions. Inspired by MCARN, we propose two models, RGF-CA and Cos-Sin+CA, which combine co-attention with the relative geometry features of visual objects, and achieve excellent comprehensive performance and higher accuracy on Other questions respectively. Extensive experiments and ablation studies based on the benchmark dataset VQA 2.0 prove the effectiveness of our models, and also verify the synergy of co-attention and visual object relation reasoning in VQA task.

Keywords Computer vision · Visual question answering · Co-attention · Visual object relation reasoning

1 Introduction

With the development of natural language processing and computer vision, which are the core areas of general intelligent behavior, multi-modal learning that breaks the boundaries of language and vision and bridges them has received extensive attention in recent years. Multi-modal learning task involves image caption [1,2], visual question answering (VQA) [3–13], image-text matching [14,15], cross-modal retrieval (CMR) [16–18], etc. All these tasks require models to understand the visual information contained in images and the textual information contained in texts simultaneously. The difference is that VQA also requires models to have common sense knowledge and reasoning ability. Given a visual image and a natural language question related to the image, VQA requires the model to understand both the image and the question simultaneously, and use common sense knowledge and a certain degree of reasoning to predict the correct answer. Besides being the benchmark of general artificial intelligence, VQA also plays a significant role in various fields [19,28]. For example, VQA methods can be applied to medical imaging to enable automatic medical diagnosis; VQA systems can be used for aided navigation to help visually impaired users; VQA models contribute to the construction of surveillance video automatic query systems; other fields involving human–machine interaction, such as education, can be more intelligent by infusing VQA technology.

Introducing attention mechanisms into VQA task has become a widely used approach. Attention mechanism was first successfully applied to the field of natural language processing [29], which improved the performance of machine translation models. Given a group of elements, an attention module can effect an individual element through the aggregation weight automatically learned driven by task goal. In VQA task, visual attention networks [30–32] help models selectively focus on the visual information most relevant to answering the input questions. Similarly, textual attention, which can help models to focus on question key words and phrases, is also very important to the cross-modal learning task VQA. Now, most advanced VQA methods [33,34] use co-attention combining visual attention and textual attention to focus the models on the question key words and the image regions that are most relevant to predicting the correct answers. On the other hand, the computer vision commu-

✉ Zihan Guo
guo_zihan11@163.com

Dezhi Han
dzhan@shmtu.edu.cn

[1] College of Information Engineering, Shanghai Maritime University, 1550 Haigang Avenue, Shanghai 201306, China

nity has recognized that modeling the relationships between visual objects is crucial to improving the performance of VQA and object recognition models [35,36]. Therefore, some researchers have introduced visual relation reasoning [37–40], which is one of the latest advances in cross-modal learning of visual representations, into VQA models and achieved good results. A visual relation reasoning module is trained together with deep neural networks. It helps models to complete downstream tasks by modeling and reasoning the positional relationships, semantic relationships and implicit relationships between the visual objects of the input images to generate visual relation representations.

Inspired by previous studies [41,42], we believe that modeling the relationships between the visual objects of the input images is as important as focusing on salient image regions to help models provide semantic-rich and fine-grained visual features for downstream tasks. However, even advanced co-attention mechanisms can only learn object-level semantics, and they ignore the complex but semantically informative relationship clues contained in images. To solve this problem, we propose a Multi-Modal Co-Attention Relation Network (MCARN) to model visual representations at both object-level and relation-level. In the co-attention module, MCARN focuses the model on question key words and significant image regions by learning the self-attention of questions and images and the question-guided visual attention. The visual relation reasoning module in MCARN is also based on a basic attention module. In addition to the original feature-based attention weight, the visual relation reasoning module adds a new geometry weight to model the relative geometry relationships between the visual objects contained in the input images. Extensive experiments and ablation studies based on the benchmark dataset VQA 2.0 [8] prove the effectiveness of our models, and also verify the synergy of co-attention and visual object relation reasoning in VQA task. The major contributions of this paper are summarized as follows:

(1) We propose a Multi-Modal Co-Attention Relation Network (MCARN) to model visual representations at both object-level and relation-level.
(2) On the basis of MCARN, we stack its visual relation reasoning module to further improve the accuracy of the model on Number questions.
(3) Extensive experiments and ablation studies performed on a benchmark VQA dataset demonstrate the feasibility and effectiveness of our models.

The rest of this paper is organized as follows: We first review the research progress of VQA in Sect. 2. Section 3 introduces the overall framework and technical details of MCARN. The experimental settings, experimental results and further analysis are given in Sect. 4. Finally, we conclude our work and give the future research direction in Sect. 5.

2 Related works

2.1 Visual question answering (VQA)

The goal of VQA is to answer the input natural language questions according to the content of the input visual images. It is quite a challenging task, since it requires models to understand and reason over both textual and visual content, and may require external common sense knowledge. A general VQA model is mainly composed of a vision part, a question understanding part and an answer generation part. In the vision part, most VQA models use deep convolutional neural networks to extract the image features. Early VQA studies used VGG [43] or ResNet [44] to extract grid features from the input images, but such grid features could not accurately reflect the boundaries of visual objects. Now, most advanced VQA methods adopt Faster R-CNN [45], which combines region proposal network (RPN) and Fast R-CNN [46], to extract regional image features with more precise boundaries. In the question understanding part, VQA models use GloVe [47] or other word embedding methods to embed the input question words into word vectors, and then use recurrent neural networks such as Long Short-Term Memory (LSTM) [48] or Gated Recurrent Unit (GRU) [49] to encode the question word vectors at sentence-level to obtain the question features. In the answer generation part, VQA models fuse the image features and the question features by multi-modal feature fusion methods and feed the fused features to the answer decoder to generate the correct answers. Early VQA models used simpler methods such as concatenation, addition and element-wise multiplication to combine the image features with the question features. Some recent studies have proposed more complex multi-modal feature fusion methods, such as bilinear-pooling-based methods [50,51], which can reflect the relationship between features of two different modalities at an effective computational cost.

2.2 Attention mechanisms

Encoding semantic-rich and fine-grained representations from textual questions and visual images is important to improve the performance of VQA models. However, the methods described above are based on global features. Such global features are difficult to help VQA models to focus on question key words and significant image regions that are more important for the models to predict the correct answers, and may introduce noise. Therefore, many researches have introduced various attention mechanisms into VQA task to focus the models' attention on significant local features. Shih et al. [31] proposed a visual attention mechanism, which maps regional image features and textual features into a shared space, and uses inner product for relevance comparison. Zhang et al. [52] developed a hierarchical convolutional

self-attention encoder to capture the question-aware video context features. On the other hand, textual attention, which can help models to better understand textual questions by focusing on question key words, is also very important to VQA models. With the application and development of attention mechanisms, co-attention combining visual attention and textual attention has become the most popular attention method used in advanced VQA studies. Lu et al. [34] proposed a hierarchical co-attention model, which focuses attention on different segments of the input questions and different regions of the input images, and can model the questions at three levels to capture information of different granularities. After this, Nguyen et al. [33] designed a co-attention mechanism that could be stacked to form a hierarchical structure to achieve dense multi-step interactions between image-question pairs. Recently, Yu et al. [41] proposed a deep modular co-attention network consisting of a series of self-attention and guided-attention units to realize a stackable dense co-attention model with better performance.

2.3 Visual relation reasoning

Recently, visual relation reasoning has been introduced into VQA task to help models better answer questions and images that require logical understanding ability and achieved impressive results. Santoro et al. [38] designed a dedicated module to calculate the relationship between entities to help deep learning architectures deal with tasks requiring rich relation reasoning. Perez et al. [53] proposed a general-purpose conditioning method based on conditional information for visual reasoning requiring multi-step and high-level processes. Yu et al. [54] designed a visual relation reasoning module to reason the pair-wise and inner-group visual relationship between visual objects to enhance visual representations at relation-level.

3 Multi-modal co-attention relation networks

MCARN combines deep modular co-attention with visual relation reasoning to guide the model to achieve visual relation reasoning and correctly answer the textual questions related to the input images. The overall framework of MCARN is shown in Fig. 1. We first introduce how to extract image features and question features from the input visual images and the input textual questions. Then the co-attention module and the visual relation reasoning module will be introduced. Finally, we employ a simple feature fusion method to fuse the extracted multi-modal features and feed the fused features into a classifier to complete the prediction of the answers.

3.1 Image and question representations

We use Faster R-CNN [45] (based on ResNet-101) pre-trained on Visual Genome [55] to extract image features from the input visual images the same as MCAN [41]. Faster R-CNN adopts bottom-up mechanism to represent the input images as regional image features. By mean-pooling the convolutional feature from its detected region, we represent the i-th visual object as a feature $\mathbf{x}_i \in \mathbb{R}^{2048}$. Based on the confidence threshold we set to the probabilities of the detected regions, the extracted regional image features are represented as a feature matrix $\mathbf{X} \in \mathbb{R}^{i \times 2048}$, where $i \in [10, 100]$ is the number of the visual objects.

For the input textual question, we first tokenize it into words and limit its maximum length to 14. Then, we use the 300-D GloVe word vectors [47], which have been pre-trained on a large-scale corpus, to embed each question word into a word vector. By doing so, we obtain a word embedding sequence of size $w \times 300$ of the input textual question, where $w \in [1, 14]$ is the number of the question words. Finally, we use a single layer LSTM with 512 hidden units to encode the word embedding sequence and obtain the question features $\mathbf{Y} \in \mathbb{R}^{w \times 512}$.

3.2 Co-attention module

The co-attention module in MCARN is based on scaled dot-product attention [56]. Its inputs include queries and keys of dimension k, and values of dimension v. For ease of calculation, we set k and v to the same value d. Scaled dot-product attention first calculates the dot products of the queries with the keys, divides each by $\sqrt{d}$ and then feeds the results into a softmax function to obtain the weights on the values. In practice, the queries, keys and values are packed together into matrices $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$. The result calculated by $\mathbf{Q}$ and $\mathbf{K}$ represents the attention. The attended feature is obtained by weighted summation over $\mathbf{V}$ with respect to the attention:

$$\text{attention}\,(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d}}\right)\mathbf{V} \tag{1}$$

Similar to reference [56], we adopt multi-head attention to perform the attention function in parallel. We concatenate the attended features and project them to obtain the final values:

$$\text{multi\_head}\,(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{concat}\,(\text{head}_1, \text{head}_2, ..., \text{head}_H)\mathbf{W}^{\mathrm{MH}} \tag{2}$$

$$\text{head}_p = \text{attention}\,(\mathbf{Q}\mathbf{W}_p^{\mathrm{Q}}, \mathbf{K}\mathbf{W}_p^{\mathrm{K}}, \mathbf{V}\mathbf{W}_p^{\mathrm{V}}) \tag{3}$$

where H is the number of the attention heads, $p \in [1, H]$, $\mathbf{W}_p^{\mathrm{Q}}$, $\mathbf{W}_p^{\mathrm{K}}$, $\mathbf{W}_p^{\mathrm{V}}$ and $\mathbf{W}^{\mathrm{MH}}$ are projection parameter matrices, and $head_p$ is the output of the p-th scaled dot-product

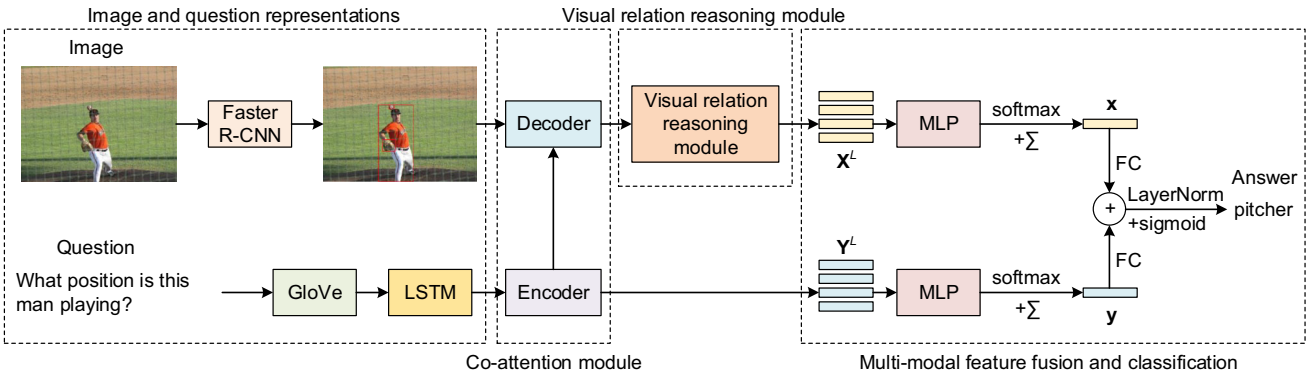
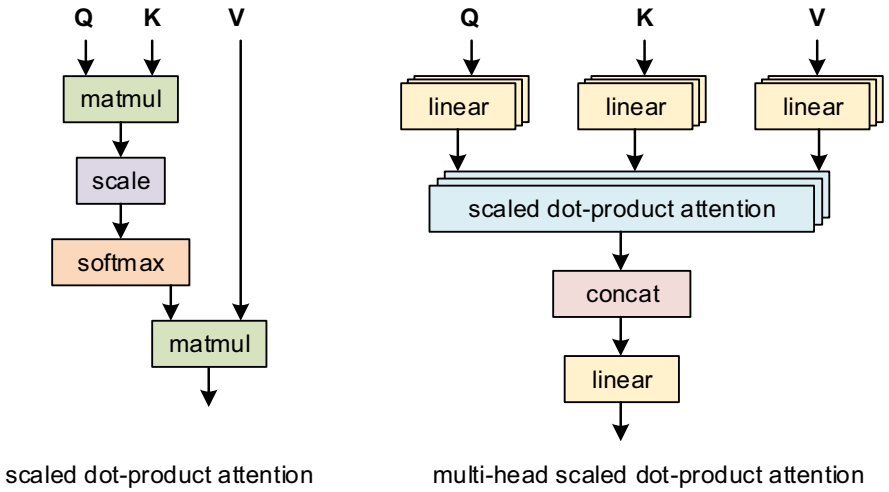


Fig. 1 The overall framework of Multi-Modal Co-Attention Relation Networks (MCARN)

Fig. 2 The calculation steps of scaled dot-product attention and multi-head scaled dot-product attention



attention head. By using multi-head attention, MCARN can focus on information from different representation subspaces at different positions. Figure 2 shows the calculation steps of scaled dot-product attention and multi-head scaled dot-product attention.

Yu et al. [41] proved that the encoder-decoder model is steadily superior to the stacking model. Therefore, we also adopt the encoder-decoder structure to construct the co-attention module in MCARN. The co-attention module takes the regional image features $\mathbf{X}$ and question features $\mathbf{Y}$ as inputs to learn the self-attention of the input images and the input questions as well as the question-guided visual attention. We denote the L-layer co-attention module as Encoder-Decoder^L. As shown in Fig. 3, it learns the self-attended question features $\mathbf{Y}^L$ and the attended image features $\mathbf{X}^L$ through L layers Encoder-Decoder.

The L-th Encoder-Decoder is shown in Fig. 4. Specifically, an Encoder consists of a multi-head scaled dot-product attention layer and a pointwise feed-forward layer. The multi-head scaled dot-product attention layer takes the question features as input to learn the self-attention of the input questions, and the pointwise feed-forward layer takes the output of the

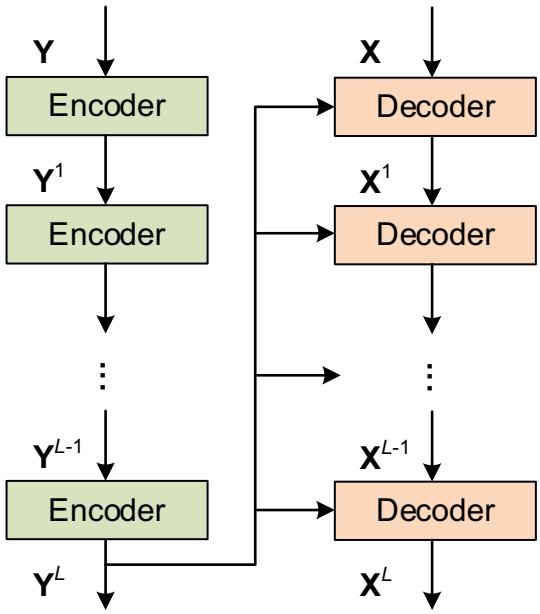


Fig. 3 The co-attention module in MCARN

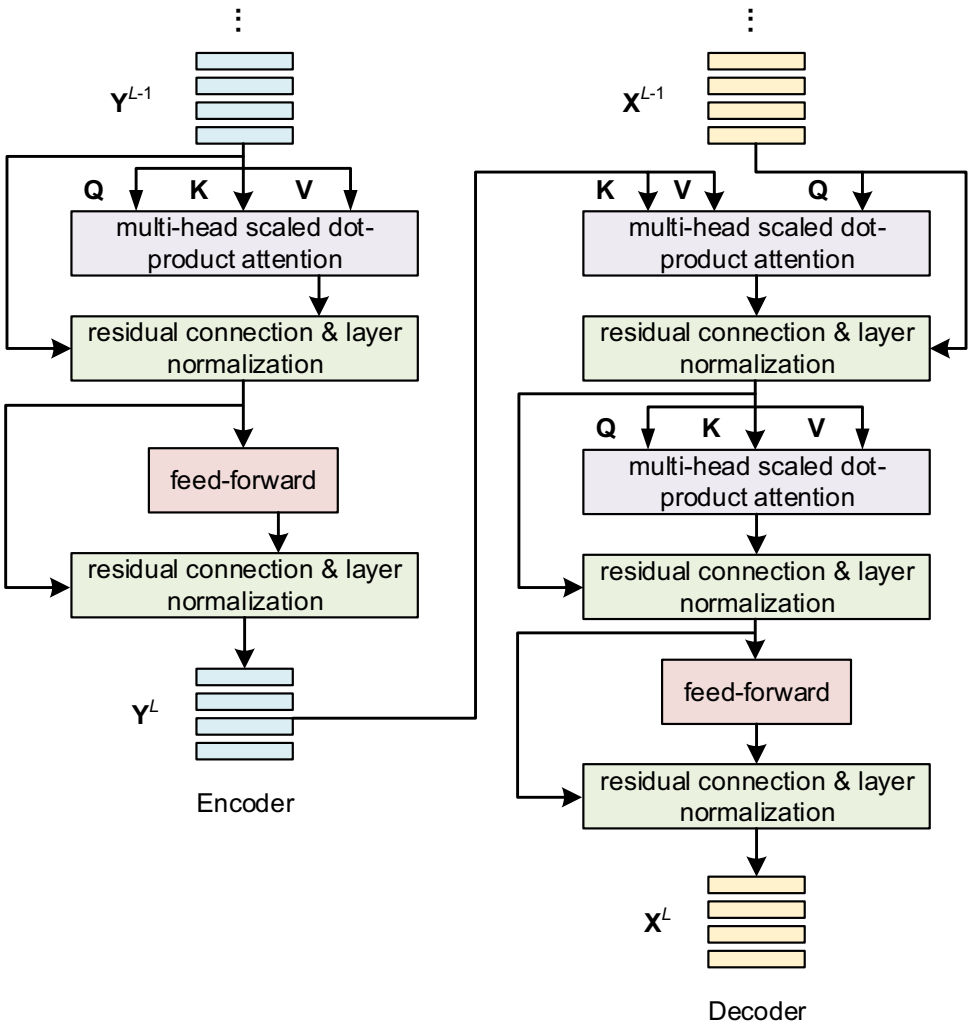


Fig. 4 The L-th Encoder-Decoder

multi-head scaled dot-product attention layer as input and adopts two fully connected layers to further transform the attended features. The purpose of Encoder is to reconstruct the question features according to the normalized similarity between each sample and other samples. A Decoder consists of two consecutive multi-head scaled dot-product attention layers and a pointwise feed-forward layer. The first multi-head scaled dot-product attention layer takes the output of Encoder and the regional image features as inputs to learn the question-guided visual attention, and the second multi-head scaled dot-product attention layer takes the output of the first multi-head scaled dot-product attention layer as input to learn the self-attention of the input images. Similar to Encoder, Decoder aims to reconstruct the image features according to the normalized cross-modal similarity between two groups of samples. In addition, we apply residual connection [44] and layer normalization [57] after each multi-head scaled dot-product attention layer and each pointwise feed-forward layer to stabilize training. In Encoder-Decoder^L, the input of the first Encoder is $\mathbf{Y}$, the input of each other Encoder is the output of the previous Encoder, and the output of the last Encoder is $\mathbf{Y}^L$. The input of the first Decoder is $\mathbf{X}$ and $\mathbf{Y}^L$, the inputs of each other Decoder are the output of the previous Decoder and $\mathbf{Y}^L$, and the output of the last Decoder is $\mathbf{X}^L$.

3.3 Visual relation reasoning module

The visual relation reasoning module in MCARN is also based on scaled dot-product attention, and Fig. 5 shows the specific calculation steps. The features of visual objects consist of their appearance features and geometry features. In this paper, the appearance features of visual objects refer to the attended image features $\mathbf{X}^L$ output by the co-attention module, and the geometry features are 4-dimensional visual object bounding boxes denoted by $\mathbf{G}$. Given the input set $\{(\mathbf{x}_n^L, \mathbf{G}_n)\}_{n=1}^i$ of i visual objects, the visual relation reasoning module calculates the relationship between each visual object and other visual objects to obtain the relation features

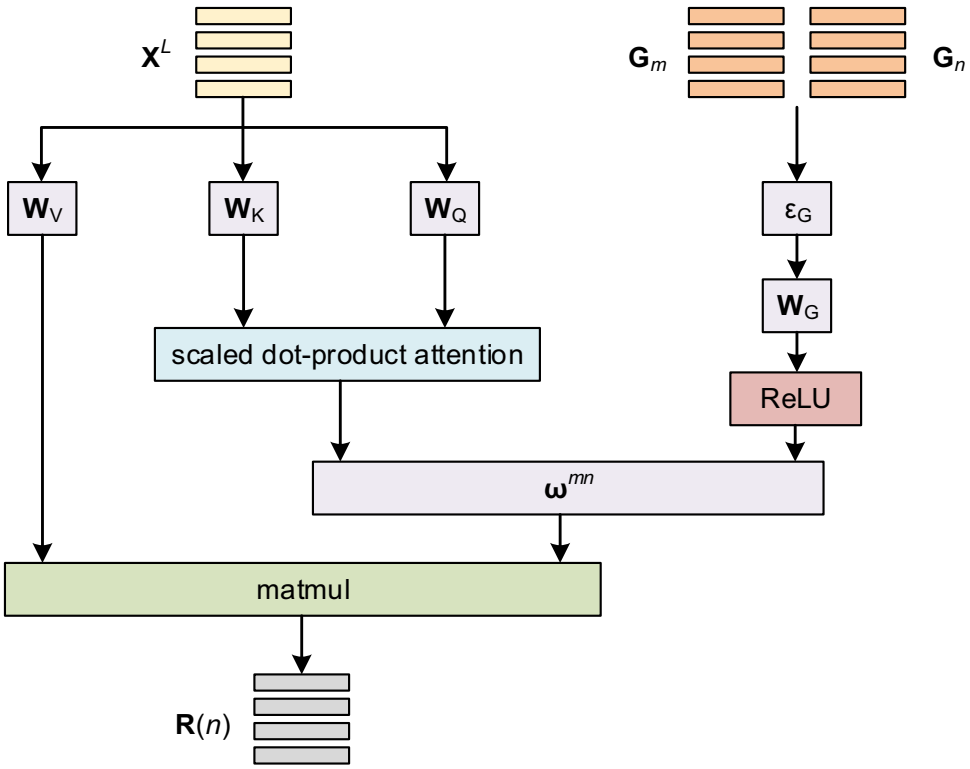


Fig. 5 The calculation steps of the visual relation reasoning module

$\mathbf{R}(n)$:

$$\mathbf{R}(n) = \sum_m \boldsymbol{\omega}^{\mathrm{mn}} \cdot \left(\mathbf{W}_{\mathrm{V}} \cdot \mathbf{x}_m^L\right) \tag{4}$$

where $\mathbf{W}_{\mathrm{V}}$ corresponds to values $\mathbf{V}$ in Eq. 1. The relation weight $\boldsymbol{\omega}^{\mathrm{mn}}$ represents the influence of other visual objects on the object, and its calculation method is as follows:

$$\boldsymbol{\omega}^{\mathrm{mn}} = \frac{\boldsymbol{\omega}_{\mathrm{G}}^{\mathrm{mn}} \cdot \exp(\boldsymbol{\omega}_{\mathrm{X}}^{\mathrm{mn}})}{\sum_k \boldsymbol{\omega}_{\mathrm{G}}^{kn} \cdot \exp(\boldsymbol{\omega}_{\mathrm{X}}^{kn})} \tag{5}$$

The appearance weight $\boldsymbol{\omega}_{\mathrm{X}}^{\mathrm{mn}}$ is calculated as a dot product according to Eq. 1:

$$\boldsymbol{\omega}_{\mathrm{X}}^{\mathrm{mn}} = \frac{\mathbf{W}_K \mathrm{x}_m^L \left(\mathbf{W}_Q \mathrm{x}_n^L\right)^T}{\sqrt{d}} \tag{6}$$

where $\mathbf{W}_{\mathrm{K}}$ and $\mathbf{W}_{\mathrm{Q}}$ correspond to $\mathbf{K}$ and $\mathbf{Q}$ in Eq. 1, which project the appearance features $\mathbf{x}_m^L$ and $\mathbf{x}_n^L$ into the subspace to calculate their matching degree. The dimension of the projected features is d, and the geometry weight $\boldsymbol{\omega}_{\mathrm{G}}^{\mathrm{mn}}$ is given by the following formula:

$$\boldsymbol{\omega}_G^{\mathrm{mn}} = \max\{0, \mathbf{W}_G \cdot \varepsilon_G(\mathbf{G}_m, \mathbf{G}_n)\} \tag{7}$$

where ε_G represents the method [56] used to embed the geometry features of the two visual objects into high-dimensional representations and the dimension of the embedded features is 64. MCARN adopts 4-dimensional relative geometry features $(\log(\frac{|u_m - u_n|}{g_m}), \log(\frac{|b_m - b_n|}{h_m}), \log(\frac{g_n}{g_m}), \log(\frac{h_n}{h_m}))^{\mathrm{T}}$ as the geometry features of visual objects, where u, b, g and h are, respectively, the abscissa and ordinate of the center point of visual object bounding boxes as well as the width and height of visual object bounding boxes. $\mathbf{W}_{\mathrm{G}}$ is used to transform the embedded features into scalar weights and we trim the scalar weights at 0 to restrict the relationship between visual objects to a certain geometry relationship. The visual relation reasoning module simultaneously models N_{r} relationships, and these multiple relation features are added with the appearance features after being concatenated to achieve feature enhancement:

$$\mathbf{x}_n^L = \mathbf{x}_n^L + \mathrm{concat}\left[\mathbf{R}^1(n), ..., \mathbf{R}^{N_{\mathrm{r}}}(n)\right] \tag{8}$$

3.4 Multi-modal feature fusion and classification

By modeling the co-attention and learning the relationship between visual objects, we obtain the image features $\mathbf{X}^L$ and the question features $\mathbf{Y}^L$. The image features contain rich object-level and relation-level information, and the question key word features are also given greater weight. Now, we use a two-layer MLP consisting of two fully connected layers to calculate the attended features $\mathbf{x}$ and $\mathbf{y}$. Taking $\mathbf{x}$ as an example, the calculation method is as follows:

$$\boldsymbol{\alpha}^{\mathrm{X}} = \mathrm{softmax}\left(\mathrm{MLP}(\mathbf{X}^L)\right) \tag{9}$$

$$\mathbf{x} = \sum_{j=1}^{i} \boldsymbol{\alpha}_j^{\mathrm{X}} \mathbf{x}_j^{L} \tag{10}$$

where $\boldsymbol{\alpha}^{\mathrm{X}}$ is the attention weight of image features. The attended image features $\mathbf{x}$ can be obtained by multiplying image features by their corresponding attention weight. Similarly, we can obtain the attended question features $\mathbf{y}$ through the above method.

We design a simple linear multi-modal feature fusion function to calculate the fused feature $\mathbf{z}$:

$$\mathbf{z} = \text{layernorm}\left(\mathbf{W}_x^T \mathbf{x} + \mathbf{W}_y^T \mathbf{y}\right) \tag{11}$$

where $\mathbf{W}_{\mathrm{x}}^{\mathrm{T}}$ and $\mathbf{W}_{\mathrm{y}}^{\mathrm{T}}$ are linear projection matrixes, and layer normalization is used to facilitate optimization. Finally, we project the fused feature $\mathbf{z}$ into a vector $\mathbf{s}$ and feed $\mathbf{s}$ into a sigmoid function to predict the correct answer, where $\mathbf{s} \in \mathbb{R}^A$ and A is the number of the most common answers of the training set.

4 Experimental studies

All the experiments are conducted on the benchmark dataset VQA 2.0 [8]. VQA 2.0 is based on Microsoft COCO image data [58] and attempts to minimize the effectiveness of learning data bias by balancing the answers to each question. VQA 2.0 is divided into the train set (82,783 images and 443,757 question-answer pairs), the validation set (40,504 images and 214,354 question-answer pairs) and the test set (81,434 images and 447,793 question-answer pairs). The test set is further divided into the test-dev set and test-standard set to evaluate VQA models online. Following previous work, we train our models on the train set and the validation set, and we also add a subset of VQA samples from Visual Genome [55] to facilitate training. We report the experimental results on the test-dev set and the test-standard set of the VQA evaluation server. The accuracies of the models are classified into four categories based on the type of the questions and the answers: Yes/No, Number, Other and Overall.

4.1 Experimental settings

During the experiments, we employ zero-padding to fill $\mathbf{X}$ and $\mathbf{Y}$ to their maximum sizes to set the number of the visual objects and the number of the question words as invariants, i.e. i=100 and w=14. To solve the underflow problem during training, we use $-\infty$ before every softmax layer to mask the padding logits. The dimension of the scaled dot-product attention layer is 512, i.e. d=512. The number of the attention heads H is set to 8 and the dimension of the output of each head is d/H=64. The layer L of Encoder-Decoder^L is set to 6 and the structure of its feed-forward layers is FC($4d$)-ReLU-dropout(0.1)-FC(d). The structure of the MLP used to calculate the attended features is FC(d)-ReLU-dropout(0.1)-FC(1). ReLU [57] is the activation function and dropout [59] is used to prevent overfitting. The number of the relation features modeled by the visual relation reasoning module $N_{\mathrm{r}} \in \{4, 8, 16, 32\}$. The dimension of the fused feature $\mathbf{z}$ is 1024, and the number of the most common answers A is 3129. We use Adam solver [60] (β_1=0.9, β_2=0.98) to train our models, set the batch size to 64 and adopt binary cross-entropy (BCE) as the loss function. The warm-up learning rate is set to $\min(2.5te^{-5}, 1e^{-4})$, where t is the current epoch number starting from 1.

Table 1 Experimental results of MCARN on the test-dev set of VQA 2.0

N_{r}	Yes/No (%)	Number (%)	Other (%)	Overall (%)
4	86.69	53.83	60.74	70.65
8	86.84	54.28	**60.78**	70.78
16	**87.07**	54.40	60.66	**70.83**
32	86.95	**54.43**	60.70	70.80

4.2 Ablation studies and analysis

This section mainly discusses the effects of the number of the relation features N_{r} and different variants of visual object relation modeling with the co-attention module on the performance of VQA models.

4.2.1 The number of the relation features N_{r}

As shown in Table 1 and Fig. 6, we explore the effects of the number of relation features $N_{\mathrm{r}} \in \{4, 8, 16, 32\}$ on the performance of MCARN. The maximum value of each column is shown in bold. The experimental results show that with the increase in N_{r}, the accuracy of the model roughly increases first and then decreases. When N_{r}=16, MCARN achieves the highest overall accuracy of 70.83%. It is worth noting that with the increase in N_{r}, the accuracy of MCARN on Number questions keeps improving and has not reached the peak, which proves that modeling the relation features of visual objects helps the model to correctly answer Number questions. SceneGCN [61] and v-VRANet [54], which combine visual reasoning methods, are far less accurate than MCAN [41], which introduces co-attention. Therefore, in Fig. 6, we only compare MCARN with the advanced co-attention model MCAN. It can be seen from Fig. 6 that MCARN outperforms MCAN on all types of questions when the appropriate parameter N_{r} is selected. The above experimental results demonstrate the effectiveness of combining co-attention with visual relation reasoning in VQA task.

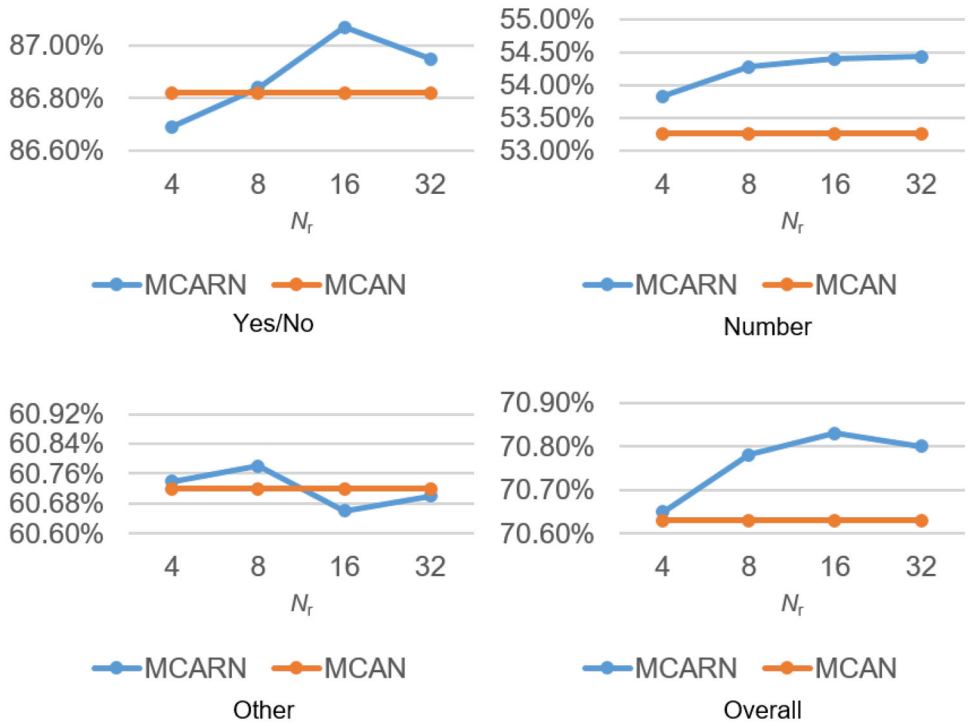


Fig. 6 The Yes/No, Number, Other and Overall accuracies of MCARN and MCAN on the test-dev set of VQA 2.0

Table 2 Experimental results of MCARN-1 on the test-dev set of VQA 2.0

N_r	Yes/No (%)	Number (%)	Other (%)	Overall (%)
4	86.91	53.58	60.79	70.73
8	86.75	53.49	60.74	70.63
16	86.91	53.21	60.57	70.58
32	86.73	53.30	60.65	70.56

To verify the effectiveness of the learnable geometry weight ω_G^{mn}, we conduct experiments on a model named MCARN-1 whose geometry weight is set as 1 and the results are shown in Table 2. By comparing with Table 1, it can be seen that when N_r values are the same, MCARN performs better than MCARN-1 on all types of questions except when N_r=4, which indicates that the learnable geometry weight can improve the performance of VQA models. When N_r=4, we believe that the advantage of MCARN-1 over MCARN is that compared with the learnable geometry weight with fewer relation features, the fixed geometry weight is more conductive to the training and optimization of the model. In addition, the accuracy of MCARN is always higher than that of MCARN-1 on Number questions regardless of N_r, which further proves the effectiveness of the learnable geometry weight in improving the counting ability of VQA models.

4.2.2 Stacked visual relation reasoning modules

Considering that the inputs and outputs of the visual relation reasoning module have the same dimension, we design a model named MCARNS that stacks two consecutive visual relation reasoning modules. In MCARNS, the input of the latter visual relation reasoning module is the output of the former visual relation reasoning module, and other parts and parameter settings of the model are the same as MCARN. We test the effects of the number of relation features N_r on the performance of MCARNS and the results are shown in Table 3 and Fig. 7. The maximum value of each column is shown in bold. To make a fair comparison, we must use the same number of relation features N_r in experiments. It can be seen that although the accuracy of the model decreases after stacking the visual relation reasoning modules, when N_r=8, MCARNS still has advantages over MCAN which only infuses co-attention. By comparing with Table 1 and Fig. 6, it can be found that stacking the visual relation reasoning modules on the basis of MCARN can further improve the accuracy of the model on Number questions, but it will reduce the accuracy of the model on the other two types of question and the overall performance of the model. This proves that the stacked visual relation reasoning modules can help MCARN to achieve more complex reasoning. In future studies, we will explore how to better combine multiple stacked visual relation reasoning modules with co-attention mechanisms.

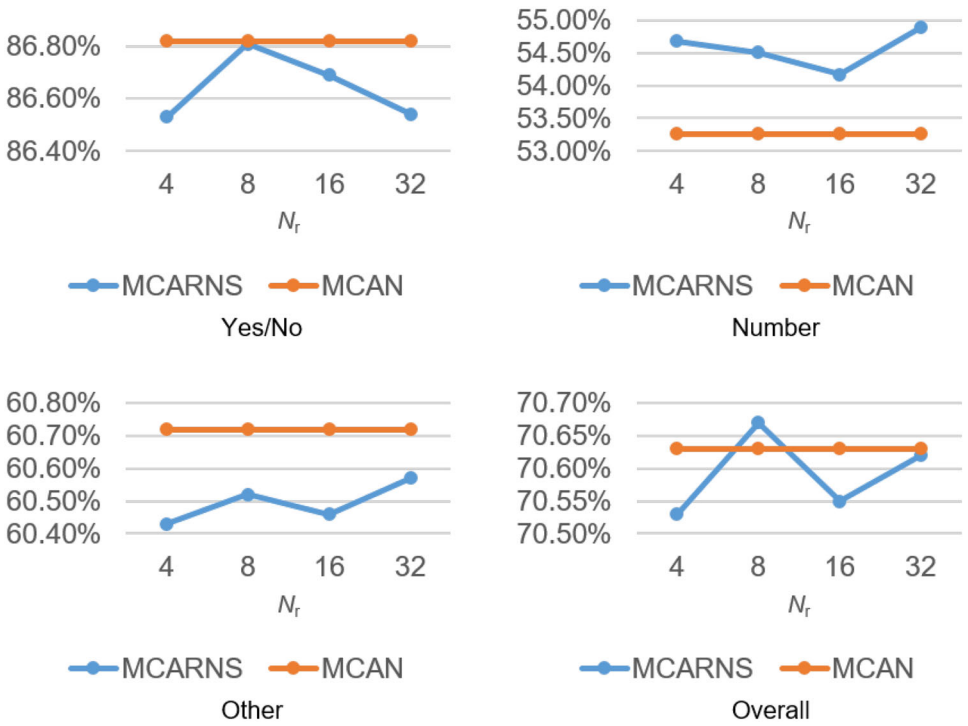


Fig. 7 The Yes/No, Number, Other and Overall accuracies of MCARNS and MCAN on the test-dev set of VQA 2.0

Table 3 Experimental results of MCARNS on the test-dev set of VQA 2.0

N_r	Yes/No (%)	Number (%)	Other (%)	Overall (%)
4	86.53	54.68	60.43	70.53
8	**86.81**	54.51	60.52	**70.67**
16	86.69	54.17	60.46	70.55
32	86.54	**54.89**	**60.57**	70.62

4.2.3 Relative geometry features of visual objects

We design two co-attention models, RGF-CA and Cos-Sin+CA, which utilize the relative geometry features of visual objects to further explore the effects of the relative position information between the visual objects contained in the input images on the performance of VQA models. Specifically, we concatenate the 2048-dimensional regional image features with the 4-dimensional relative geometry features of the 100 visual objects as the 2448-dimensional input image features of RGF-CA, and other parts and parameter settings of the model are the same as MCAN. We also try to project the 400-dimensional relative geometry features into 512 dimensions, then add them to the input image features of Decoder and take them as the new input image features. However, the effect of this method is not ideal, and we will not list it in the experimental results. Cos-Sin+CA takes the high-dimensional relative geometry features of visual objects embedded into 64 dimensions in Eq. 7 as the input of the visual part, and other parts and parameter settings of the model are the same as MCAN. Specifically, we project the 6400-dimensional relative geometry features of the 100 visual objects into 512 dimensions, then add them to the input image features of Decoder and take them as the new input image features.

The above two models only use the co-attention mechanism of encoder-decoder structure without introducing the visual relation reasoning module, and the experimental results are shown in Table 4. The maximum value of each column is shown in bold. It can be seen that the relative geometry features of visual objects can also improve the performance of VQA models. Among them, RGF-CA with simpler relative geometry features has better comprehensive performance and Cos-Sin+CA with embedded high-dimensional relative geometry features achieves the best results on Other questions.

4.3 Comparison with advanced VQA models

Table 5 shows the experimental results of our models and the advanced VQA single models on the test-dev set and the test-standard set of VQA 2.0. The maximum value of each column is shown in bold. Among them, BUTD [62] is a model combining regional image features with question-guided attention, which considers the natural basis of attention. BAN [63] is a bilinear attention network, which takes into account the bilinear interactions between multi-modal inputs to make full use of the given textual information and visual information. BAN+Counter combines BAN with Counter [64],

Table 4 Experimental results of MCAN, RGF-CA, Cos-Sin+CA, MCARN and MCARNS on the test-dev set of VQA 2.0

Model	Yes/No (%)	Number (%)	Other (%)	Overall (%)
MCAN	86.82	53.26	60.72	70.63
RGF-CA	86.94	53.60	60.90	70.80
Cos-Sin+CA	86.97	53.07	**60.98**	70.79
MCARN(N_r=16)	**87.07**	54.40	60.66	**70.83**
MCARNS(N_r=8)	86.81	**54.51**	60.52	70.67

Table 5 Experimental results of our models and the advanced VQA single models on the test-dev set and the test-standard set of VQA 2.0

Model	test-dev Yes/No (%)	Number (%)	Other (%)	Overall (%)	test-standard Overall (%)
BUTD [62]	81.82	44.21	56.05	65.32	65.67
BAN [63]+Counter [64]	85.42	54.04	60.52	70.04	70.35
MCAN [41]	86.82	53.26	60.72	70.63	70.90
MUAN [65]	86.77	54.40	60.89	70.82	71.10
SceneGCN [61]	82.72	46.85	57.77	66.81	67.14
v-VRANet [54]	83.31	45.51	58.41	67.20	67.34
MCARN(N_r=16)	**87.07**	54.40	60.66	**70.83**	71.16
MCARNS(N_r=8)	86.81	**54.51**	60.52	70.67	70.91
RGF-CA	86.94	53.60	60.90	70.80	**71.17**
Cos-Sin+CA	86.97	53.07	**60.98**	70.79	**71.17**

a neural network component that allows robust counting in object proposals to further improve the accuracy of the model on Number questions. MCAN [41] and MUAN [65] are both VQA models that introduce co-attention mechanisms to model the dense inter-modal interactions and intra-modal interactions. SceneGCN [61] proposes a Scene Graph Convolution Network, which performs VQA task by jointly inferring the semantic relationships and attributes of visual objects. v-VRANet [54] proposes a visual relation reasoning module to infer the visual relationships between visual objects under the guidance of the input questions. It can be seen from Table 5 that our models achieve the best results on all types of questions on the test-dev set and the highest overall accuracy on the test-standard set. Specifically, MCARN using only one visual relation reasoning module achieves the highest accuracy on Yes/No questions. For Number questions, MCARNS stacking two visual relation reasoning modules achieves the highest accuracy. For Other questions, it is a better choice to combine the high-dimensional relative geometry features of visual objects with the attended image features. All the above experimental results demonstrate the effectiveness of our models and the importance of modeling visual representations at both object-level and relation-level in VQA task.

4.4 Attention visualization

As shown in Fig. 8, in this section, we compare the attention visualization results of our model with MCAN for specific image instances to enhance the interpretability of our model. The three input questions in the figure belong to the Yes/No type, the Number type and the Other type from top to bottom. For the first input instance, although the MCAN model correctly focuses on the black clad snowboarder, the model also divides its attention to other image regions that do not contain the black clad snowboarder, while our MCARN model successfully focuses only on the image regions related to answering the input question. For the second input instance, the MCAN model not only fails to focus on all the frames, but also diverts attention to wrong image regions, such as the image region containing the chair, while our MCARN model correctly focuses on all the image regions containing frames. For the last input instance, both MCAN model and MCARN model obtain correct attention visualization results, but the attention map of MCARN model is more compact, which can help the model to predict the correct answer more accurately.

5 Conclusion

The current mainstream VQA models only model the object-level visual representations but ignore the relationships between visual objects. To solve this problem and effectively use the position information of visual objects and their relative geometry relationships in VQA task, we propose a Multi-Modal Co-Attention Relation Network (MCARN) that combines co-attention and visual object relation reasoning. MCARN uses the co-attention module to learn the textual

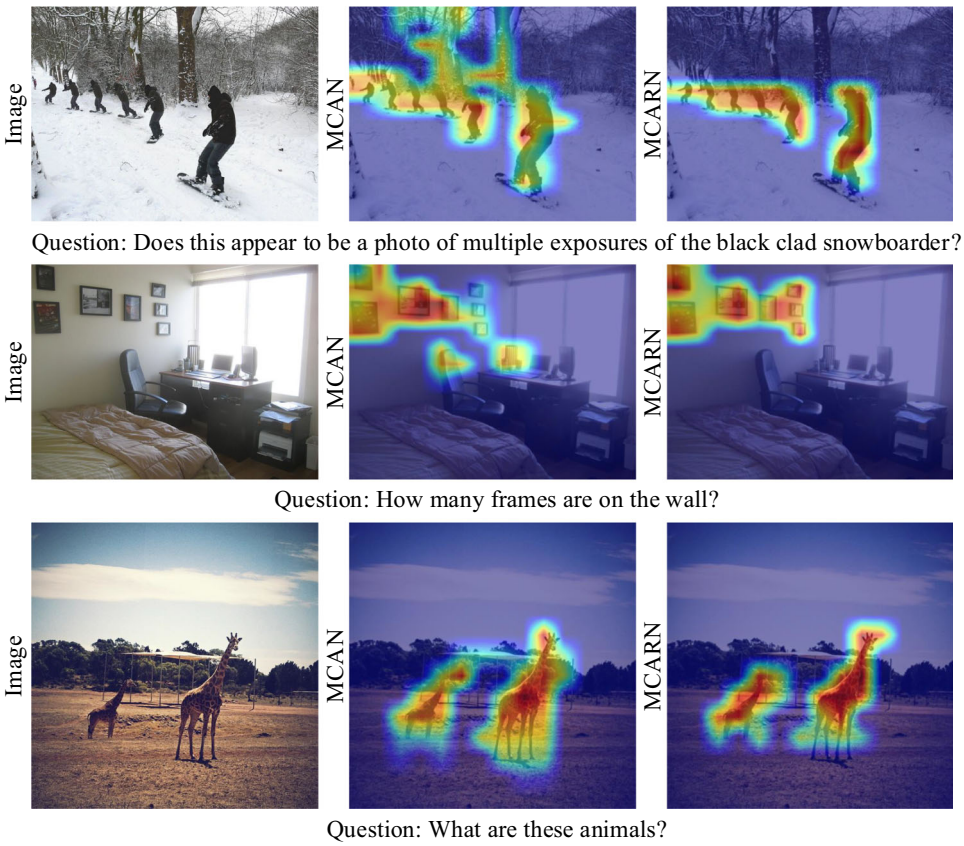


Fig. 8 The attention visualization results of MCAN and MCARN

features and object-level visual representations that are more critical for predicting the correct answers, and further utilizes the visual relation reasoning module to model the visual representations at relation-level. On the basis of MCARN, we stack its visual relation reasoning module to further improve the accuracy of the model on Number questions. Inspired by MCARN, we propose two models, RGF-CA and Cos-Sin+CA, which combine co-attention with the relative geometry features of visual objects, and achieve excellent comprehensive performance and higher accuracy on Other questions respectively. Extensive experiments and ablation studies based on the benchmark dataset VQA 2.0 prove the effectiveness of our models, and also verify the synergy of co-attention and visual object relation reasoning in VQA task. In future studies, we will explore more effective geometry feature representations of visual objects and how to better combine multiple stacked visual relation reasoning modules with co-attention mechanisms.

Acknowledgements This research is supported by the National Natural Science Foundation of China under Grant 61873160 and Grant 61672338, and the Natural Science Foundation of Shanghai under Grant 21ZR1426500. We thank all the reviewers for their constructive comments and helpful suggestions.

Author Contributions Methodology, material preparation, data collection, and analysis were performed by ZG. ZG wrote the first draft of the manuscript, and DH and ZG commented on previous versions of the manuscript. DH did the supervision, reviewing, and editing. All authors read and approved the final manuscript.

Data availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

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

References

1. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: neural image caption generation with visual attention. ICML. **37**, 2048–2057 (2015)
2. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. CVPR. **1**, 3156–3164 (2015)
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: visual question answering. ICCV. **1**, 2425–2433 (2015)
4. Noh, H., Seo, P.H., Han, B.: Image question answering using convolutional neural network with dynamic parameter prediction. CVPR. **1**, 30–38 (2016)
5. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. NIPS. **1**, 1682–1690 (2014)

6. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: overcoming priors for visual question answering. CVPR. **1**, 4971–4980 (2018)
7. Lao, M., Guo, Y., Wang, H.: Zhang, Xin: Cross-modal multistep fusion network with co-attention for visual question answering. IEEE Access. **6**, 31516–31524 (2018)
8. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: elevating the role of image understanding in visual question answering. CVPR. **1**, 6325–6334 (2017)
9. Han, D., Pan, N., Li, K.-C.: A traceable and revocable ciphertext-policy attribute-based encryption scheme based on privacy protection. IEEE Trans. Dependable Secur. Comput. **19**(1), 316–327 (2022)
10. Cui, M., Han, D., Wang, J.: An efficient and safe road condition monitoring authentication scheme based on fog computing. IEEE Internet Things J. **6**(5), 9076–9084 (2019)
11. Cui, M., Han, D., Wang, J., Li, K.-C., Chang, C.-C.: ARFV: an efficient shared data auditing scheme supporting revocation for fog-assisted vehicular ad-hoc networks. IEEE Trans. Veh. Technol. **69**(12), 15815–15827 (2020)
12. Han, D., Zhu, Y., Li, D., Liang, W., Souri, Alireza, Li, Kuan-Ching.: A blockchain-based auditable access control system for private data in service-centric IoT environments. IEEE Trans. Ind. Inf. **18**(5), 3530–3540 (2022)
13. Li, H., Han, D.: Tang, Mingdong: A privacy-preserving storage scheme for logistics data with assistance of blockchain. IEEE Internet Things J. **9**(6), 4704–4720 (2022)
14. Lee, K.-H., Chen, X., Hua, G., Houdong, H., He, X.: Stacked cross attention for image-text matching. ECCV. **4**, 212–228 (2018)
15. Wang, L., Li, Y., Huang, J., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. IEEE Trans. Pattern Anal. Mach. Intell. **41**(2), 394–407 (2019)
16. Ranjan, V., Rasiwasia, N., Jawahar, C.V.: Multi-label cross-modal retrieval. ICCV. **1**, 4094–4102 (2015)
17. Song, G., Wang, D.: Tan, X.: Deep memory network for cross-modal retrieval. IEEE Trans. Multim. **21**(5), 1261–1275 (2019)
18. He, Y., Xiang, S., Kang, C., Wang, J., Pan, C.: Cross-modal retrieval via deep and bidirectional representation learning. IEEE Trans. Multim. **18**(7), 1363–1377 (2016)
19. Lu, P., Ji, L., Zhang, W., Duan, N., Zhou, M., Wang, J.: R-VQA: learning visual relation facts with semantic attention for visual question answering. KDD. **1**, 1880–1889 (2018)
20. Ren, F.: Zhou, Y.: CGMVQA: a new classification and generative model for medical visual question answering. IEEE Access. **8**, 50626–50636 (2020)
21. Sayedshayan, H.H., Mehran, S., Abdolreza, M.: Multiple answers to a question: a new approach for visual question answering. Vis. Comput. **37**(1), 119–131 (2021)
22. Yan, F., Silamu, W., Li, Y.: Chai, Yachuang: SPCA-Net: a based on spatial position relationship co-attention network for visual question answering. Vis. Comput. (2022). https://doi.org/10.1007/s00371-022-02524-z
23. Rahman, T., Chou, S.-H., Sigal, L., Carenini, G.: An improved attention for visual question answering. CVPR Workshops **2021**, 1653–1662 (2021)
24. Yang, C., Wu, W., Wang, Y., Zhou, H.: Multi-modality global fusion attention network for visual question answering. Electronics **9**(11), 1882 (2020)
25. Guo, Z., Han, D.: Multi-modal explicit sparse attention networks for visual question answering. Sensors **20**(23), 6758 (2020)
26. Liu, H., Gong, S., Ji, Y., Yang, J., Xing, T., Liu, C.: Multimodal cross-guided attention networks for visual question answering. CMSA 2018. (2018)
27. Han, D., Zhou, S., Li, K.-C.: Rodrigo Fernandes de Mello: cross-modality co-attention networks for visual question answering. Soft Comput. **25**(7), 5411–5421 (2021)
28. He, S., Han, D.: An effective dense co-attention networks for visual question answering. Sensors **20**(17), 4897 (2020)
29. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. ICLR (2015)
30. Chen, K., Wang, J., Chen, L.-C., Gao, H., Xu, W., Nevatia, R.: ABC-CNN: an attention based convolutional neural network for visual question answering. CVPR. (2015). arXiv:1511.05960
31. Shih, K.J., Singh, S., Hoiem, D.: Where to look: focus regions for visual question answering. CVPR. 4613-4621 (2016)
32. Liu, Y., Zhang, X., Huang, F., Tang, X., Li, Z.: Visual question answering via attention-based syntactic structure tree-LSTM. Appl. Soft Comput. **82**, 1055484 (2019)
33. Nguyen, D.-K., Okatani, T.: Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. CVPR. 6087–6096 (2018)
34. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. NIPS. 289-297 (2016)
35. Lao, M., Guo, Y., Wang, H.: Zhang, Xin: Cross-modal multistep fusion network with co-attention for visual question answering. IEEE Access. **6**, 31516–31524 (2018)
36. Mottaghi, R., Chen, X., Liu, X., Cho, N.-G., Lee, S.-W., Fidler, S., Urtasun, R., Yuille, A.L.: The role of context for object detection and semantic segmentation in the wild. CVPR. **1**, 891–898 (2014)
37. Chen, X., Gupta, A.: Spatial memory for context reasoning in object detection. ICCV. **1**, 4106–4116 (2017)
38. Santoro, A., Raposo, D., Barrett, D.G.T., Malinowski, M., Pascanu, R., Battaglia, P.W., Lillicrap, T.: A simple neural network module for relational reasoning. NIPS. 4967–4976 (2017)
39. Peng, L., Yang, Y., Wang, Z., Wu, X., Huang, Z.: CRA-Net: composed relation attention network for visual question answering. ACM Multimedia. 1202–1210 (2019)
40. Wang, P., Wu, Q., Shen, C., van den Hengel, A.: The VQA-machine: learning how to use existing vision algorithms to answer new questions. CVPR. **1**, 3909–3918 (2017)
41. Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering. CVPR. **1**, 6281–6290 (2019)
42. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. CVPR. 3588–3597 (2018)
43. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. ICLR. **1** (2015)
44. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CVPR. **1**, 770–778 (2016)
45. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. NIPS. **1**, 91–99 (2015)
46. Girshick, R.B: Fast R-CNN. ICCV. **1**, 1440–1448 (2015)
47. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. EMNLP. 1532–1543 (2014)
48. Hochreiter, Sepp, Schmidhuber, Jürgen.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
49. Chung, J., Gülçehre, Ç., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. NIPS. **1** (2014), arXiv:1412.3555
50. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. EMNLP. **1**, 457–468 (2016)
51. Yu, Z., Yu, J., Fan, J., Tao, D.: Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. ICCV. **1**, 1839–1848 (2017)
52. Zhang, Z., Zhao, Z., Lin, Z., Song, J., He, X.: Open-ended long-form video question answering via hierarchical convolutional self-attention networks. IJCAI. **1**, 4383–4389 (2019)

53. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.C.: FiLM: visual reasoning with a general conditioning layer. AAAI. **1**, 3942–3951 (2018)
54. Yu, J., Zhang, W., Yuhang, L., Qin, Z., Hu, Y., Tan, J., Wu, Q.: Reasoning on the relation: enhancing visual representation for visual question answering and cross-modal retrieval. IEEE Trans. Multim. **22**(12), 3196–3209 (2020)
55. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D.A., Bernstein, M.S., Fei-Fei, Li.: Visual genome: connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. **123**(1), 32–73 (2017)
56. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. NIPS. **1**, 5998–6008 (2017)
57. Ba, L.J., Kiros, J.R., Hinton, G.E.: Layer normalization. Mach. Learn. **1**, 2016. arXiv:1607.06450
58. Microsoft, C.O.C.O., Tsung-Yi, L., Michael, M., Serge, J.B., James, H., Pietro, P., Deva, R., Piotr, D.C., Lawrence, Z.: Common objects in context. ECCV **5**, 740–755 (2014)
59. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
60. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. ICLR.**1**, (2015)
61. Yang, Z., Qin, Z., Yu, J., Wan, T.: Prior visual relationship reasoning for visual question answering. ICIP. **1**, 1411–1415 (2020)
62. 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. CVPR. **1**, 6077–6086 (2018)
63. Kim, J.-H., Jun, J., Zhang, B.-T.: Bilinear attention networks. NeurIPS. **1**, 1571–1581 (2018)
64. Zhang, Y., Hare, J.S., Prügel-Bennett, A.: Learning to count objects in natural images for visual question answering. ICLR. **1** (2018)
65. Yu, Z., Cui, Y., Yu, J., Tao, D., Tian, Q.: Multimodal unified attention networks for vision-and-language interactions. CVPR. **1**, (2019). arXiv:1908.04107

Zihan Guo Zihan Guo received his B.E. degree from Tianjin Polytechnic University, China, in 2017. He is currently a Ph.D. student at the College of Information Engineering, Shanghai Maritime University, China. His research interests include computer vision, natural language processing and visual question answering based on deep learning.

Dezhi Han Dezhi Han received his B.S. degree in applied physics from the Hefei University of Technology, Hefei, China, in 1990, and his M.S. and Ph.D. degrees in computing science from the Huazhong University of Science and Technology, Wuhan, China, in 2001 and 2005, respectively. He is currently a Professor with the Department of Computer, Shanghai Maritime University, Pudong, China, in 2010. His research interests include visual question answering, cloud and outsourcing security, wireless communication security, network, and information security. He is currently a Member of the IEEE.