Applied Intelligence (2024) 54:5062–5080
https://doi.org/10.1007/s10489-024-05437-7

Relational reasoning and adaptive fusion for visual question answering

Xiang Shen[1] · Dezhi Han[1] · Liang Zong[2] · Zihan Guo[3] · Jie Hua[4]

Accepted: 30 March 2024 / Published online: 13 April 2024

Abstract

Visual relationship modeling plays an indispensable role in visual question answering (VQA). VQA models need to fully understand the visual scene and positional relationships within the image to answer complex reasoning questions involving visual object relationships. Accurate reasoning and an understanding of the relationships between different visual objects are particularly crucial. However, most reasoning models used in current VQA tasks only use simple attention mechanisms to model visual object relationships and ignore the potential for effective modeling using rich visual object features during the learning process. This work proposes an effective visual object Relationship Reasoning and Adaptive Fusion (RRAF) model to address the shortcomings of existing VQA model research. RRAF can simultaneously model visual objects' position, appearance, and semantic features and uses an adaptive fusion mechanism to achieve fine-grained multimodal reasoning and fusion. Specifically, we designed an effective image encoder to model and learn the relationship between the position and appearance features of visual objects. In addition, in the co-attention module, we employ semantic information from the question to focus on critical visual objects. Finally, we use an adaptive fusion mechanism to reassign weights and fuse different modalities of features to effectively predict the answer. Experimental results show that the RRAF model outperforms current state-of-the-art methods on the VQA 2.0 and GQA datasets, especially in visual object counting problems. We also conducted extensive ablation experiments to demonstrate the effectiveness of the RRAF model, achieving an overall accuracy of **71.33**% and **57.83**% on the VQA 2.0 and GQA datasets, respectively. Code is available at https://github.com/shenxiang-vqa/RRAF.

Keywords Visual question answering · Adaptive fusion · Visual relationship modeling · Attention mechanisms

1 Introduction

The Visual Question Answering (VQA) [1] task is an emerging research area in the field of artificial intelligence, which involves learning from both visual and textual modalities, thus bridging the gap between computer vision and natural language processing. In recent years, multimodal learning of language and vision has received extensive attention. Among various common multimodal learning tasks, including image captioning [2, 3], visual grounding [4], cross-modal information retrieval [5–7], and visual question answering [1]. VQA has emerged as a contemporary method for conducting the Turing test to evaluate visual intelligence. It learns to reason answers about real-world images when given natural language questions about the visual content. The VQA task requires models to provide answers based on given questions and images related to those questions, which places very high demands on the model to handle cross-modal information.

In recent years, advanced VQA methods [8–10] have mainly adopted attention mechanisms and multimodal joint representations of questions to achieve better performance. These methods aim to answer questions based on visual clues extracted from images and semantic information related to the questions, such as MCAN [11] and DFAF [12], which have significantly improved the performance of the VQA task. However, these models have not fully utilized spatial positional relationships and geometric image features in images, but only used partial positional information. For example, MCAN encodes some positional information in the image preprocessing stage through convolutional neural

Xiang Shen and Zihan Guo contributed equally to this work

✉ Dezhi Han
dzhan@shmtu.edu.cn

Extended author information available on the last page of the article

networks. In addition, there is another study that learns cross-modal alignment through pre-training on a large amount of unlabeled data. Although this method is effective, there is a natural deficiency in visual reasoning ability when answering questions that require a deep understanding of positional relationships between visual objects. Moreover, these large-scale pre-trained models directly connect positional features with semantic features before attention operations, which can often generate potential noise. Many methods [10, 13, 14]aim to extract information from rich images and questions, but do not consider possible noise information. Although image and question features can be extracted through deep convolutional and recurrent neural networks, they may not be effectively used for reasoning and predicting correct answers. Therefore, the VQA task places high demand on the cross-modal information processing capability of models.

As shown in Fig. 1, the model needs to understand the semantic, spatial, and appearance relationships of the input image. The image shown in Fig. 1 is from the public dataset MSCOCO VQA v2. The model needs to understand the semantic relationships between objects and learn the spatial relationships and appearance weights of visual objects with other objects. When answering the question "How many spoons are there beside the plate on the table?". The model first focuses on "spoons" based on the semantic information in the question. Then, it uses the critical information of the target object to learn the appearance weight of "spoons" and reasons based on the position information of "spoons" and other objects to predict the answer accurately. To obtain the semantic, spatial, and appearance information of these visual objects, we need to go beyond simple object detection and understand the interaction of fixed properties and relative position relationships between different objects in the image. The essence of the VQA task is to reason about the relationships between local and global objects in different regions of the image and establish complex multimodal feature relationships to answer questions.

However, deep mining of image feature information may also pose challenges for model prediction because it can cause a modal alignment problem between different features of images and questions. Deep modeling of visual object relationships may also include unnecessary information in the final obtained image features. In order to give the model better decision-making motivation, the current mainstream method is to use visual attention maps to tell the model the focus area in the image so that the model can focus on the area matching human visual focus when answering questions. However, recent research has shown that current VQA task models do not truly focus on the same areas as human attention [15]. Therefore, in order to answer the correct questions, it is crucial to search for key information in the image based on the semantic information of the questions. Most existing multimodal fusion models extract rich semantic features from images and questions, fuse multimodal features through multiplication and concatenation, and then map them to a common space for answer prediction. Although some linear models, such as MCB [16], and MFH [17], are considered effective fusion methods, these shallow multimodal fusion models lack fine-grained multimodal interactions. Attention mechanisms [18] are often one of the most commonly used methods for multimodal alignment and filtering of unnecessary information to allow multimodal modeling and improved interaction learning. The model can selectively focus on essential areas in the image

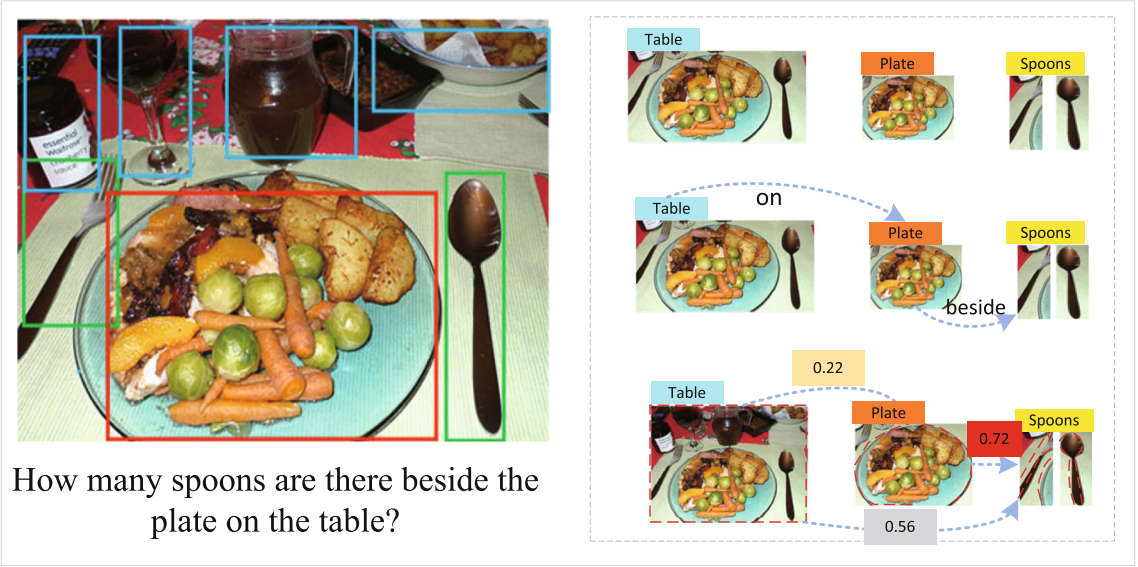

Fig. 1 Illustrating the need for a model to simultaneously understand both the spatial relationships and visual appearance of objects in order to comprehend complex reasoning tasks

through attention mechanisms. For example, BUTD [19] uses a bottom-up attention mechanism to align question information and each object feature. At the same time, DFAF [20] and MCAN [11] consider attention within each modality and interaction attention between different modalities. However, these attention mechanisms and fusion methods only focus on object information related to semantic questions when implementing modal alignment. Therefore, geometric position relationships and appearance weight information that are ignored near the focused objects will also affect the model's ability to answer complex reasoning questions.

Based on the above analysis and inspired by relevant work [21–23], we propose a relationship reasoning and adaptive fusion network for visual question answering tasks. In the process of image feature modeling, we designed an image feature encoder to achieve object position relationship reasoning and geometric appearance weight relationship modeling. At the same time, we use a deep co-attention network mechanism to obtain rich image semantic features. Given that irrelevant information may exist in image features, which may affect the alignment of image and question modalities and the prediction of correct answers, we use an adaptive fusion mechanism to reassign weights of different modalities before the fusion of multimodal features. By reassigning modality weights, we can filter out irrelevant information, and also align the question information with the relevant position in the image.

In order to address the challenges in VQA tasks, this paper proposes a novel visual object Relational Reasoning and Adaptive Fusion (RRAF) model. In visual object relation reasoning, an image encoder is used to model the spatial and geometric weight relations of visual objects. In contrast, the question encoder guides the image decoder to learn image semantic information. The network can obtain more re-fined image features, including position relationships, appearance weights, and semantic relationships, which support reasoning and answering of complex questions. In addition, we propose an effective multi-modal adaptive fusion mechanism, which can automatically adjust model features before fusion without designing complex attention mechanisms, increasing model parameters, or losing important feature information. Therefore, the main contributions of this work are as follows:

(1) This paper proposes a novel visual object relation reasoning network that employs an image encoder to simultaneously model object positions and appearance weight relations, while a question encoder guides the image decoder to learn crucial semantic information from the image. By using the object relation reasoning network, more fine-grained image features can be obtained simultaneously (including position relationships, appearance weights, and semantic relationships), which enable complex question reasoning and answering.
(2) An effective multi-modal adaptive fusion mechanism is proposed, which explores the effects of fusion using different weighting allocation strategies. The method is designed to be simple and effective, without the need for complex attention mechanisms, without increasing model parameters or losing important feature information.
(3) Experimental results on the benchmark VQA 2.0 and GQA datasets demonstrate that the RRAF model outperforms current state-of-the-art models. Through ablation experiments, we further analyze the impact of parameter settings on the RRAF model and reveal the model's interpretability.

The rest of this paper is organized as follows: Section 2 introduces the visual question answering models, visual relation reasoning models, and multi-modal fusion mechanisms related to this study. Section 3 describes the overall structure and specific design details of the RRAF model in detail. Section 4 presents the experimental results of comparisons on the VQA 2.0 and GQA benchmark datasets under different parameter settings, and the ablation experiment research, and visualizes the model through attention. Finally, the work of this paper is summarized, and future research directions are pointed out.

2 Related work

2.1 Visual question answering

Visual Question Answering (VQA) is a multimodal task that utilizes machine learning or deep learning techniques to analyze visual input (images and videos) and answer questions related to the visual content. In recent years, there has been an increasing focus on the understanding and reasoning of visual information in the VQA task. The VQA task aims to combine visual and textual information to correctly predict answers to given images and related questions by profoundly understanding and reasoning about the visual and textual information. This process requires the model to have strong reasoning abilities [1], especially for answering complex questions requiring even more complex reasoning abilities. Achieving answers to complex questions relies heavily on the sophisticated reasoning abilities of the model. A critical step in this process is the in-depth exploration and understanding of semantic cues within the questions by the model. Subsequently, the model engages in reasoning and judgment regarding image content through the analysis and modeling of semantic clues from the questions. For the analysis and reasoning of natural language, Sean Gerrish [24], in his

work "How Smart Machines Think," employs the principles of deep parse trees to analyze and comprehend the part-of-speech and language structures within sentence questions, thereby constructing an effective DeepQA system. However, implementing such intricate reasoning in visual question answering tasks is more challenging. This is because the model needs to analyze and reason about the semantic information in the questions and also utilize information from question clues for cross-modal understanding of image data that lacks the grammatical structure inherent in natural language. Therefore, researchers have constructed various VQA reasoning models.

VQA models traditionally map image and question features to the same high-dimensional space and fuse them through simple methods, which can easily ignore the introduction of noise and thus affect model performance. Currently, many VQA methods use attention mechanisms to improve the overall performance of the model. For example, many VQA methods [9, 10, 25–27] mainly study the relationship between visual regions and words in questions. Yang et al. [28] proposed a Stacked Attention Network (SAN) to locate image regions relevant to the questions. In [29], the authors proposed a compact trilinear interaction model that can simultaneously learn the relationships between images, questions, and answers, thereby achieving better model reasoning. Nguyen et al. [8] proposed a new reasoning framework to understand these features and effectively predicts in a coarse-to-fine manner. ReGAT [30] considers explicit and implicit relationships to enrich image representations. In addition to attention mechanisms, multimodal fusion strategies also critically impact the final performance of the VQA task [31]. Traditional multimodal fusion methods concatenate image and question features differently and then map them to a common space after alignment. At the same time, recent research [16, 32] explores more complex and effective multimodal fusion methods. Extracting meaningful features is also crucial for correctly answering questions in the VQA task. In image feature processing, grid features [33] and object region visual features [19, 34] are widely used, and these image features are more refined compared to region visual features. In term of question features, Glove [35] and BERT [36] commonly represent words and sentences. With the emergence of pre-training models, visual and language multi-tasking can effectively promote alignment between different modalities, thereby making the model perform well.

2.2 Visual relation reasoning

Visual relationship reasoning is crucial in visual-linguistic tasks and is the core of visual question answering. Visual relationship modeling aims to reason and understand the relationships between objects, such as spatial and logical relationships. These methods are based on symbolic reasoning [37] or graph-based simple reasoning. Modeling the position, appearance, shape, and semantics of visual objects is crucial for locating targets in contextual information. Recent research [38, 39] has shown that modeling position and semantic relationships between visual objects can help the model better understand the image and answer complex reasoning questions. Most existing research on visual object relationship reasoning focuses on implicit and explicit reasoning. Implicit relationships [30, 40, 41] are derived from the correlation of modal features (such as self-attention), while explicit relationships [30, 42, 43] represent geometric or semantic relationships between objects. However, more than explicit or implicit reasoning alone is needed to adequately achieve model reasoning. To better construct a reasoning model, graph neural network-based reasoning methods [30, 44] have received widespread attention. For example, Cadène et al. [45] proposed a multimodal relationship network for modeling spatial and semantic relationships between image regions. Li et al. [30] proposed a relationship-aware graph attention network, where image object regions are treated as nodes, and edges represent implicit relationships between objects. In addition, various attention mechanisms have been applied in existing methods to enhance the learning of implicit semantic relationships in the reasoning framework. These models assign an importance score to each region for a question and use them to weigh and summarize visual representations. Liu et al. [46] proposed the deep semantic-guided relationship attention network (DSGANet), which explicitly uses the formed 3D spatial relationships between objects to accurately align relationships inside/within objects. Chen et al. [47] proposed a cross-modal relational reasoning network (CRRN) to mask inconsistent attention graphs, highlight all potential alignments of corresponding word domains to align relationship consistency, and integrate the interpretability of VQA systems. Zhao et al. [48] leveraged enhanced semantic information within queries and employed graph neural networks to achieve entity reasoning. However, these models cannot model visual-spatial and semantic relationships simultaneously and lack interpretability to some extent. The RRAF model is capable of concurrently modeling the spatial, appearance, and semantic features of visual information. However, augmenting the number of reasoning modules to simultaneously capture diverse visual features will escalate the complexity and computational costs of the model.

2.3 Multimodal adaptive fusion

In VQA tasks, correctly selecting and fusing important feature information is crucial. Traditional modality fusion methods [16, 17] mainly focus on fusing shallow features. For example, Fukui et al. [16] proposed the Multimodal Compact Bilinear (MCB) method to map multimodal features to

high-dimensional space, which could be more computationally efficient due to its large memory requirement. To address the high-dimensional problem in MCB. Yu et al. proposed an enhanced version of MCB, named multimodal Factorized Bilinear (MFB) pooling [17], which uses the matrix factorization technique to compute the fused features, reducing the number of parameters and further enhancing the representation power of multimodal fused features. Kim et al. [49] further proposed MLB, which reduces the number of parameters by rewriting the weight matrix as the multiplication of two smaller matrices. These modality fusion methods have helped improve the performance of VQA models. Recently, some researchers have proposed adaptive fusion mechanisms to better select and allocate different modality information. For instance, Gu et al. [50] proposed an adaptive fusion network that utilizes complementary information from attention maps and adaptively fuses information based on word-level embeddings at different levels to represent image-question pairs appropriately. Chen et al. [51] proposed a unified Adaptive Rebalancing Network (ARN) for handling the classification of head and tail classes in the VQA dataset, which first learns a generic representation, and then gradually emphasizes tail data through an adaptively learned rebalancing branch. Zhang et al. [52] designed a practical Deep Multimodal Reasoning and Fusion network (DMRFNet) to achieve fine-grained multimodal reasoning and fusion. However, these methods did not consider the contribution weights for modality fusion at the end. Therefore, this paper designs a simple and effective adaptive fusion mechanism to reallocate the weights of different modalities, effectively improving the overall performance of the model. However, adaptive fusion may exhibit high sensitivity to subtle variations in input data, resulting in the model's instability when confronted with minor changes. This sensitivity could render the model highly responsive to noise or uncertainty. Consequently, we address the impact of employing distinct strategies for adaptive fusion mechanisms on the model separately.

3 Method

In this paper, the input of an image, question, and answer is represented by I, Q, and $A = \{a_i\}_{|\Omega|}$, respectively. Given an image I and a question Q, the goal is to predict an answer $a^* \in \Omega$ that best matches ground-truth answer a. The VQA task can be defined as a classification problem:

$$a^* \leftarrow arg \max_{a \in \Omega} p_\theta (a|I, Q) \tag{1}$$

This section presents a detailed description of the RRAF model, which is a visual object relationship reasoning method based on an adaptive fusion mechanism. The overall framework of the RRAF model is shown in Fig. 2, which mainly consists of feature extraction, visual object relationship reasoning, modality adaptive fusion, and answer prediction modules. Firstly, we describe the methods for extracting question and image features, and then describe the principles of the image and question encoders, as well as details about the question-guided image decoder. Finally, we introduce the design of the multi-modal adaptive fusion mechanism and answer prediction module. Through the introduction in this section, we can have a comprehensive understanding of the implementation details and working principles of the RRAF model.

3.1 Image and question representation

Image representation: In the section on image feature representation, inspired by the bottom-up attention mechanism, and following the approach proposed by [19], we employ a bottom-up approach to extract salient regions of the image as visual features. These features are based on intermediate features extracted from a Faster R-CNN [53] model pre-trained on the Visual Genomes [54] dataset, with ResNet-101 [55] as the backbone network. Specifically, a confidence threshold is set based on the probability of detected objects, dynamically determining the number of detected objects, denoted as $k \in [10, 100]$. Considering the varying number of salient regions in each image, we choose k=100 and fill the regions below this threshold with zeros to reach the maximum size. The feature representation $F_S = \left\{f_s^1, f_s^2, \cdots, f_s^k\right\} \in \mathbb{R}^{k \times d_v}$ (k=100) of the salient regions is obtained by convolution and mean pooling. The feature of the i-th visual object, denoted as $f_s^i \in \mathbb{R}^{d_v}$($d_v$=2048), can be represented. Simultaneously, we obtain the bounding box coordinates $F_G = \{b_i\}_{i=1}^k$ of the object, where the i-th visual object detection box is represented as $b_i^o = \left(x_{\min}^i, y_{\min}^i, x_{\max}^i, y_{\max}^i\right)$, with $x_{\min}^i$ and $y_{\min}^i$ indicating the coordinates of the top-left corner, and $x_{\max}^i$ and $y_{\max}^i$ indicating the coordinates of the bottom-right corner.

Question representation: In the question feature representation section, we follow the approach proposed by [11]. To ensure uniform length for all questions, each given question is initially tokenized into a sequence of words, which is then tokenized into words, and the number of these words is padded to the maximum value of 14; any exceeding portion is discarded. Subsequently, these words are represented in vector form, and using the 300-D GloVe word embedding model pre-trained on a large-scale corpus [35], they are transformed into a size of $t \times 300$ word sequence, where $t \in [1, 14]$ represents the number of words in the question. Finally, these word embeddings are passed through a single-layer LSTM network to obtain the question feature matrix represented as $Q_s = \{q_1, q_2, \cdots, q_t\} \in \mathbb{R}^{t \times d_q}$ (t=14), where $q_i \in \mathbb{R}^{d_q}$ (d_q=512).

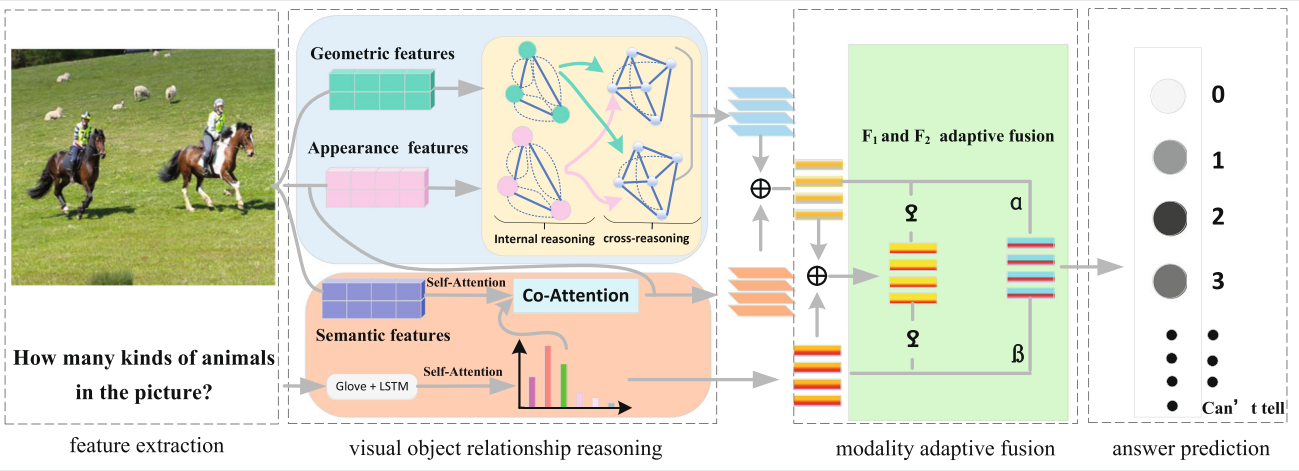

Fig. 2 Overall architecture of the proposed RRAF model

3.2 Visual relational reasoning

We adopted a structure consisting of two encoders and one decoder. Figure 3(a) illustrates the structure of the question encoder, which effectively learns the relationships among questions using the multi-head self-attention mechanism and guides self-attention to important regions of images related to the questions, thereby achieving the proposed image semantic relation modeling. Figure 3(b) shows the structure of the image encoder, which differs from the traditional Transformer framework structure as we constructed a novel image encoder that simultaneously models visual object appearance and geometric positional features. The traditional Transformer framework structure only employs unidirectional encoders and decoders and uses only self-attention and the question-guided attention of the decoder during the image modeling process. However, the RRAF model uses both image and text encoders and a decoder to achieve reasoning for complex problems.

3.2.1 Question encoder

As shown in Fig. 3 (a) represents the semantic encoder of the question. Using the question encoder, not only can the semantic features of the question be effectively learned through a self-attention mechanism, but the important regions of the image can also be attended based on the obtained relevant weights of the question semantic features. This is beneficial for the model to extract fine-grained image semantic features. In order to capture richer semantic features of the text, the question encoder is stacked with L self-attention (SA) units. Figure 3(a) represents a question encoder unit, which is mainly composed of multi-head self-attention layers and fully connected layers. Given the input question feature $Q_s = \{q_1, q_2, \cdots, q_t\} \in \mathbb{R}^{t \times d_q}$ as input, the multi-head attention layer can effectively learn the relationships between words $< q_i, q_j >$. For ease of writing, we will describe in detail the first layer of the question encoder, where the semantic relationship feature of the question is represented as F_{q_1}. The specific representation formula is as follows:

$$Q_q = Q_s W_{Q_q}, K_q = Q_s W_{K_q}, V_q = Q_s W_{V_q} \tag{2}$$

$$F_{q_1} = MA\left(Q_q, K_q, V_q\right) = concat\left(head_1, head_2, \cdots, head_h\right) W_l^{O_q} \tag{3}$$

$$head_i = soft\max\left(\frac{\left(Q_q W_i^{Q_q}\right) \cdot \left(K_q W_i^{K_q}\right)^T}{\sqrt{d}}\right)\left(V_q W_i^{V_q}\right) \tag{4}$$

where $W_i^{Q_q}, W_i^{K_q}, W_i^{V_q} \in \mathbb{R}^{d \times d_h}$($d$=512, d_h=64) are the projection matrices for the i-th head. $W_l^{O_q} \in \mathbb{R}^{h * d_h \times d}$($h$=8,$l = 1, \cdots, 6$) represents a linear mapping matrix. The notation $concat\,(\cdot)$ is used to represent the concatenation of all the heads. $MA\,(\cdot)$ is used to indicate multiple attention mechanisms. In order to avoid an increase in model size due to multiple concatenations, we usually have $d_h = d/h$. The output from the multi-head self-attention layer is then fed into a fully connected layer and subsequently into the next question encoder unit. The formula is defined as follows:

$$FFN\left(F_{q_1}\right) = \max\left(0, F_{q_1} W_1 + b_1\right) W_2 + b_2 \tag{5}$$

$$FFN\left(F_{q_1}\right) \rightarrow FFN\left(F_{q_L}\right) \tag{6}$$

where W_j and b_j (j=1,2) represent weight coefficients and biased variable respectively. $FFN\,(\cdot)$ represents a feed forward network.

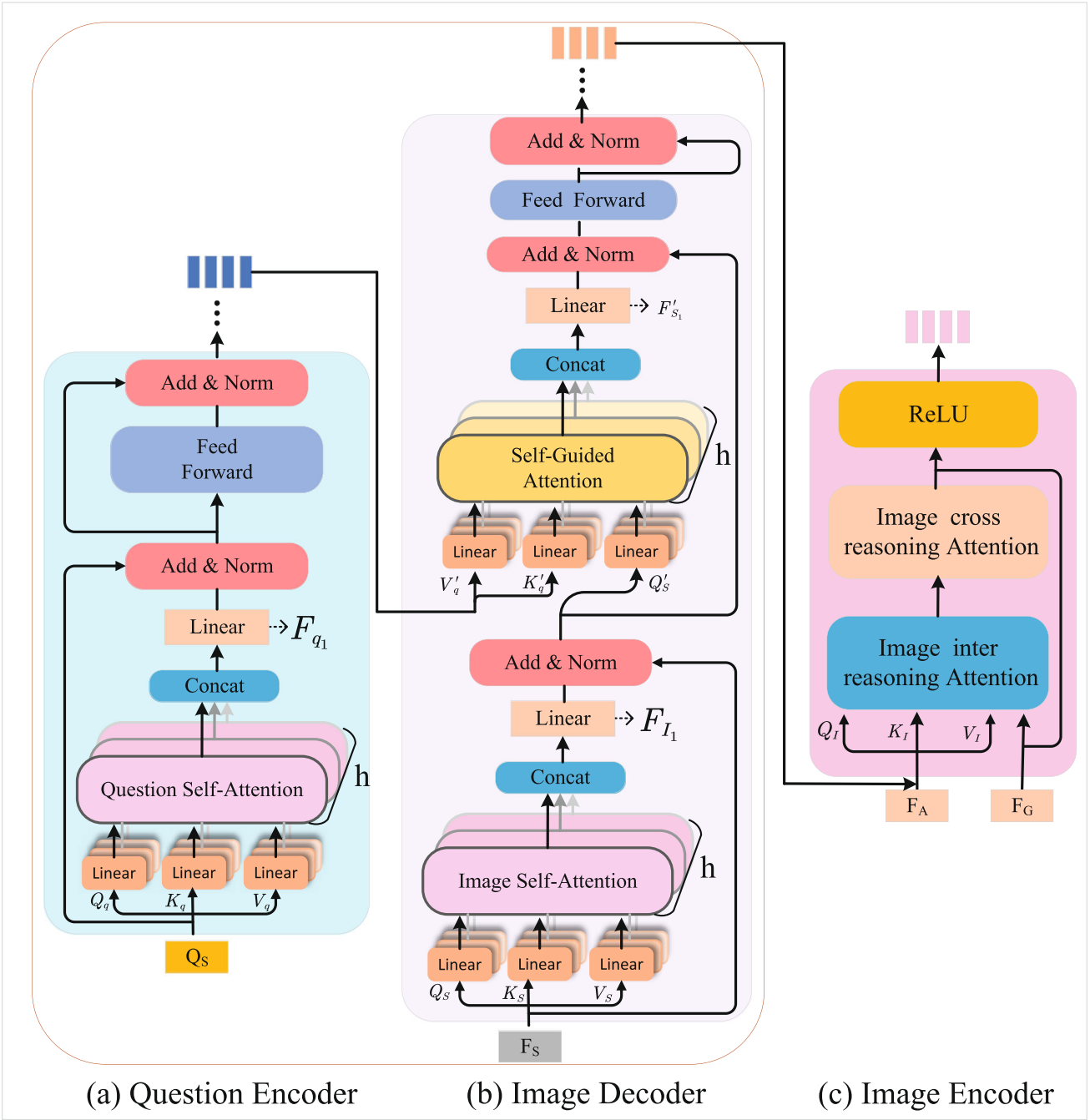

Fig. 3 Figures(a) and (b) represent the question encoder and image decoder in the co-attention mechanism, while figures(c) represents the image reasoning encoder

3.2.2 Image decoder

Figure 3(b) shows, an image decoder that employs a multi-head self-attention mechanism to process the semantic information of an image, similar to the self-attention mechanism for the question encoder described in Section 3.2.1. The difference is that in the image decoder, the semantic information of the question is used to guide attention to the semantic information of the image, thereby obtaining higher-level semantic information. The image encoder consists of multi-head self-attention layers for the image, multi-head question self-attention guided image attention layers, and fully connected layers. The semantic features of the question are used as input. The multi-head question self-attention guided image attention layer can effectively learn the relationship between the question words and a semantic image $< q_i, f_{s_i} >$. Firstly,

self-attention learning is performed based on the extracted image features. F_s and F_{I_1} represent the original low-level features of the input image and the semantic features of the image after being processed by multi-head self-attention, respectively. The equation is defined as follows:

$$\underbrace{F_S \xrightarrow{W^*} (K_S, V_S, Q_S)}_{LM} \tag{7}$$

$$F_{I_1} = MA(K_S, Q_S, V_S) = concat(H_1, H_2, \cdots, H_h) W_l^{O_S} \tag{8}$$

$$FFN\left(F_{I_1}\right) = \max\left(0, F_{I_1} W_3 + b_3\right) + b_4 \tag{9}$$

W^* and LM represent the mapping parameter matrix and linear mapping, respectively, H_i defined in (4). $F_{S_1}^{'}$ denotes image feature maps that have been guided by the semantic features of the query through multi-head self-attention to attend to the target region features of the image. The formula is expressed as follows:

$$Q_{S_1}^{'} = LN\left(F_S + FFN\left(F_{I_1}\right)\right) \tag{10}$$

$$F_{S_1}^{'} = MA\left(Q_{S_1}^{'}, K_q^{'}, V_q^{'}\right) = concat\left(H_1^{'}, H_2^{'}, \cdots, H_h^{'}\right) W^{Q_{S_1}^{'}} \tag{11}$$

$$H_i^{'} = soft\max\left(\frac{\left(Q_{S_1}^{'} W_i^{Q_S^{'}}\right) \cdot \left(K_q^{'} W_i^{K_q^{'}}\right)}{\sqrt{d}}\right) \cdot \left(V_q^{'} W_i^{V_q^{'}}\right) \tag{12}$$

$$FFN\left(F_{S_1}^{'}\right) \xrightarrow{\cdots} FFN\left(F_{S_L}^{'}\right) = F_S^L \tag{13}$$

where $Q_{S_1}^{'}$ represents the query vector after self-attention learning on the image in the first layer, and $LN(\cdot)$ represents layer normalization. $W_i^{Q_S^{'}}, W_i^{K_q^{'}}, W_i^{V_q^{'}}$ and $W^{Q_{S_1}^{'}} \in \mathbb{R}^{h*d_h \times d}$ are learnable parameter matrices.

3.2.3 Image encoder

The visual object position and appearance modeling encoder is shown in Fig. 3(c). Section 3.2.2 introduced the image decoder, which obtains image features that are semantically related to the question but do not include information about the visual object's position and appearance shape. However, in the VQA task, it is vital to have a deep understanding and reasoning of the image while modeling the position and appearance shape of the visual object. As shown in Fig. 3(c), the image encoder first models the position relationship and appearance shape of the visual object separately. In contrast, appearance shape modeling uses self-attention to learn appearance weights between different visual objects. Visual object position relationship modeling constructs weight relationships based on object coordinate relationships. Then, the position and appearance weight relationships are combined with modeling of the visual object and ultimately obtains image features that contain visual object position and appearance weight relationships. These features help to answer complex VQA questions.

To model relationships between different objects and effectively represent the spatial positions of each visual object, we calculate the transformed position coordinates $b_i = [x_i, y_i, w_i, h_i]$ for the i-th detection box from its original coordinates $b_i^o = \left(x_{\min}^i, y_{\min}^i, x_{\max}^i, y_{\max}^i\right)$. Here, (x_i, y_i), w_i, and h_i represent the center coordinates, width, and height of the bounding box, respectively. The specific calculation is formulated as follows:

$$\begin{bmatrix} x_i = 0.5 * \left(x_{\min}^i + x_{\max}^i\right) \\ y_i = 0.5 * \left(y_{\min}^i + y_{\max}^i\right) \\ w_i = x_{\max}^i - x_{\min}^i + 1.0 \\ h_i = y_{\max}^i - y_{\min}^i + 1.0 \end{bmatrix} \tag{14}$$

The coordinates representing the bounding box position of the visual object are obtained through the calculation in (14) and denoted as $F_G = \{b_i\}_{i=1}^K$. In order to model the relationships between different objects, the image encoder first encodes the coordinates of visual objects, denoted by $F_G = \{b_i\}_{i=1}^K$. We first calculate the relative positional relationship between the m-th visual object at coordinate $b_m = [x_m, y_m, w_m, h_m]$ and the n-th coordinate $b_n = [x_n, y_n, w_n, h_n]$. The formula for calculating the object positional relationship is as follows:

$$pr_{mn} = \left[\log\left(\frac{|x_m - x_n|}{w_m}\right), \log\left(\frac{|y_m - y_n|}{h_m}\right), \log\left(\frac{w_n}{w_m}\right), \log\left(\frac{h_n}{h_m}\right)\right] \tag{15}$$

$$gr_{mn} = \mathrm{Re}LU\left(Emb\left(pr_{mn}\right) W_G\right) \tag{16}$$

where pr_{mn} denotes the computation of the positional relationship between two coordinates, while $Emb(\cdot)$ refers to the embedding of a 4-dimensional relative positional feature into a d_r-dimensional vector using sine and cosine functions of different wavelengths. Finally, the d_r-dimensional vector $W_G \in \mathbb{R}^{d_r}$ is transformed into a scalar gr_{mn}.

The modeling of the spatial relationship of visual objects also involves calculating the appearance weights of objects. The appearance features of visual objects are derived from the original semantic visual features F_S, which are decoded by an image decoder and combined with the semantic visual features F_S^L (L refers to the number of layers of encoder and

decoder) of the visual objects related to the question. Then, the appearance weight relationships of different objects are modeled using the information of the semantic visual features of the visual objects. By using the dot-product attention method on the input appearance features $F_A = \left\{v_i^*\right\}_{i=1}^K$, the appearance weights between different visual objects can be obtained, where ar_{mn} represents the similarity between the appearance of objects. The calculation is performed using the dot-product attention [11] as follows:

$$ar_{mn} = \sigma\left[\left(v_m^* Q_I\right) \cdot \left(v_n^* K_I\right) / \left(\sqrt{d_v / N}\right)\right] \tag{17}$$

By performing the aforementioned computations, we obtain the geometric positional relationship feature gr_{mn} and the appearance weight feature ar_{mn}. The integration of these two distinct visual features enables the learning of interactive reasoning among different object features. Given K sets of visual objects $\left\{f_a^n, f_g^n\right\}_{n=1}^K$, the relationship between the n-th visual object and the entire set of visual objects is represented by $F_R(n)$. The specific formula for this relationship is defined as follows:

$$\alpha_{nm} = \frac{gr_{mn} \cdot \exp\left(ar_{mn}\right)}{\sum_n^K gr_{mn} \cdot \exp\left(ar_{mn}\right)} \tag{18}$$

$$F_R(n) = \sigma\left(\sum_m \alpha_{mn} \cdot \left(W \cdot f_a^m\right)\right) \tag{19}$$

where W represents the projection matrix, $\sigma(\cdot)$ represents the activation function, and $F_R(n)$ represents the reasoning feature which contains geometric positional relationships and appearance weights. To further enhance visual object reasoning, we connect various object features to obtain the final visual reasoning feature F_{AG}^N.

$$F_{AG}^N = ||_{i=1}^N F_R^i(n) \oplus ar_{mn} \tag{20}$$

where $\oplus$ represents feature addition operation, $||$ represents a concatenation function, and N denotes the number of reasoning attention heads.

3.3 Modal adaptive fusion and answer prediction

By modeling the positional, appearance, and semantic features of visual objects simultaneously using image encoders and decoders, rich visual features can be obtained from the image. When answering questions, the model needs to retrieve information from visual regions or the relationships between them, or retrieve both simultaneously depending on the requirements of the question. However, these rich image features may contain information that is irrelevant to the question, making it difficult to align modal features from different domains during modal fusion. In addition, the abundance of image feature information can lead to a greater contribution of image features for classification and a lesser contribution of question features. Inspired by relevant research [56], we designed different types of multimodal adaptive fusion modules that dynamically control the contribution of image and question information to predict the answer through a modal adaptive fusion mechanism. As illustrated in Fig. 4, the utilization of adaptive fusion mechanisms enables effective incorporation of diverse features into the answer prediction process, taking into account their respective contributions.

By using a deep co-attention network, we output the semantic image feature $F_S^L = \left\{f_{s_1}^{(L)}, f_{s_2}^{(L)}, \cdots, f_{s_m}^{(L)}\right\} \in \mathbb{R}^{m \times d_s}$.Similarly, by encoding the image position and appearance features $F_{AG}^N \in \mathbb{R}^{m \times d_v}$ through visual reasoning, a fine-grained image feature is obtained that contains image semantics, position and appearance. The output question feature is represented by $Q_S^L = \left\{q_{s_1}^{(L)}, q_{s_2}^{(L)}, \cdots, q_{s_t}^{(L)}\right\} \in \mathbb{R}^{t \times d_s}$. We first use two MLP layers (FC(d_v)-ReLU-Dropout(0.1)-FC(1)) to compress the two different-dimensional features to the same dimension. Taking the image feature as an example, the final output image feature can be represented as $\tilde{F}_{sga}$:

$$F_{SAG} = F_{AG}^N \oplus F_S^L \tag{21}$$

$$\lambda = soft\max\left(MLP\left(F_{SAG}\right)\right) \tag{22}$$

$$\tilde{F}_{sag} = \sum_i^m \lambda_i f_{sag}^i \tag{23}$$

where $\Gamma = [\lambda_1, \lambda_2, \cdots, \lambda_m] \in \mathbb{R}^m$ is the learnable weight parameter. Feature $\tilde{q}_s$ of the question can be obtained using the same method. In order to improve the ability of the model to predict the correct answer, we use an MLP network to obtain two modal features with equal contribution. To avoid manually assigning weights to the modal features of the image and the question, we designed two different strategies for an adaptive fusion mechanism to reassign the weights of the output modal features. As shown in Fig. 4(b), this is a simple and efficient method. Specifically, we designed two different fusion methods. The first method is based on the overall features of the image and the question, and reassigns the weights of the new features. The specific equation is defined as follows:

$$C_1 = soft\max\left(W_{f_1}\left[\tilde{F}_{sag} \oplus \tilde{q}_s\right]\right) \tag{24}$$

$$\alpha = T_1 C_1, \beta = T_2 C_1 \tag{25}$$

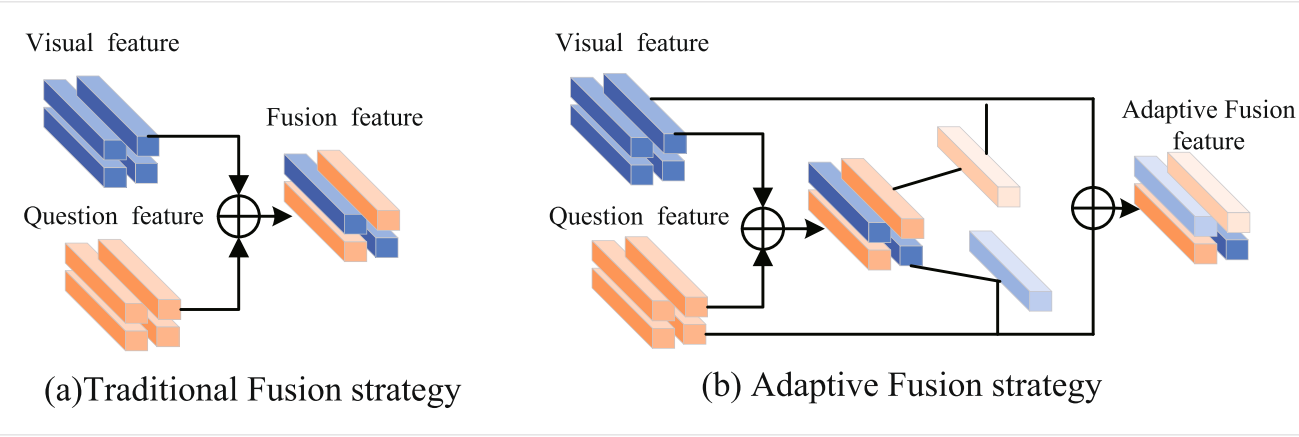

Fig. 4 Comparison of traditional fusion strategy and adaptive fusion strategy

We also developed an additional weight distribution method, which can be expressed by the following equation:

$$C_2 = sigmoid\left(W_{f_2}\left[concat\left(\tilde{F}_{sag}, \tilde{q}_s\right)\right]\right) \tag{26}$$

$$\varepsilon = C_2 \otimes \tilde{q}_s, \eta = 1 - \varepsilon \tag{27}$$

where W_{f_1} and W_{f_2} are projection matrices, T_1 and T_2 represent operations for extracting and allocating weights. We express it using a multi-modal linear fusion function as follows:

$$Z_1 = LN\left(W_\alpha\left(\alpha \otimes \tilde{q}_s\right)\right) + W_\beta\left(\beta \otimes \tilde{F}_{sag}\right) \tag{28}$$

$$Z_1 = LN\left(W_\varepsilon\left(\varepsilon \otimes \tilde{q}_s\right)\right) + W_\eta\left(\eta \otimes \tilde{F}_{sag}\right) \tag{29}$$

where Z_1 and Z_2 represent the features obtained by using different fusion strategies. W_α and W_β are the trainable parameter matrix. Since a question may have multiple correct answers, binary cross-entropy loss (BCE) is used as the model optimization objective during the training process.

$$loss = -\sum_{i=1}^{|A|}\left(y_i \log\left(\hat{y}_i\right) + \left(1 - y_i\right)\log\left(1 - \hat{y}_i\right)\right) \tag{30}$$

$|A|$ represents the size of the candidate answer set, $\hat{y}_i$ is the score predicted by the model for each candidate answer, and y_i is the soft score of the answer provided in the dataset.

4 Experiments and discussion

All experiments in this paper were conducted on a Linux Ubuntu 18.04 system with 4 NVIDIA Geforce GTX 3090TI graphics cards, each with 24GB of memory. The deep learning framework was PyTorch, and the CUDA version was 10.0. In Sections 4.1 and 4.2, we first introduce the two datasets used in this paper, namely VQA 2.0 [56] and GQA [56], and provide detailed information on the experimental settings. Section 4.3 describes the experimental results on these two benchmark datasets and compares them with state-of-the-arts models, while Section 4.4 describes the comparison of ablation experiments. Finally, in Section 4.5, we demonstrate the effectiveness of our method through visual examples.

4.1 DataSet

VQA 2.0: The VQA 2.0 dataset [56] is a fully manually annotated open-domain benchmark dataset for visual question answering. It consists of images from the MS-COCO dataset along with associated question-answer (QA) pairs. Each image is associated with a minimum of three questions, and each question has ten corresponding answers. Due to the manual annotation of questions, the dataset inevitably carries certain language biases. Researchers have attempted to address the effectiveness of minimizing (Table 1) model learning bias by balancing answers for each question. The VQA 2.0 dataset is partitioned into training, validation, and test sets, with the sample distribution outlined as follows:

Table 1 VQA 2.0 dataset sample distribution

Split	Images	Questions	Answers
Train	82,783	443,757	4,437,570
Val	40,504	214,354	2,143,540
Test	81,434	447,739	-
All	204,721	1,105,904	-

The test set contains two subsets: *test-dev* and *test-std*. The questions in the VQA 2.0 dataset can be divided into three types based on the category of their answers:"Yes/No", "Number", and "Other". For open-ended tasks, following the work of Antol et al. [57], a voting mechanism is used to score the accuracy of the predicted answers, as shown below:

$$Accuracy\,(a) = \min\left\{\frac{count\,(a)}{3}, 1\right\} \tag{31}$$

where $count(a)$ is the number of votes for answer a by 10 different annotators.

GQA: The GQA [58] dataset comprises 113K images, encompassing 22M questions. In contrast to VQA 2.0, the GQA dataset emphasizes the inference and compositional language understanding capabilities of questions, while also prioritizing the objectivity of inquiries, underscoring that answers to questions can only be derived from the images themselves. Relative to the VQA 2.0 dataset, the GQA dataset mitigates language biases by establishing connections between keywords in questions and corresponding regions within images, facilitating research to more accurately pinpoint the reasons for model errors. Furthermore, GQA introduces a more extensive set of evaluation metrics, including consistency, validity, plausibility, and distribution, to comprehensively assess the performance of models in complex scene-based question reasoning. The sample distribution of the GQA dataset is shown in Table 2.

4.2 Implementation details

In the experiment, we first used a pre-trained Faster-RCNN model with ResNet-101 as the backbone to extract object features from the images. The dimensions of the object features and the question word features are 2048 and 512, respectively. In the multi-head attention mechanism, we set the number of attention heads h to 8, and the dimension of the features outputted by each head is $d/h = 64$. Following the suggestion in [11], the number of layers L in both the image decoder and question encoder is set to 6, with a feedforward layer structure of FC(512,2048)-ReLU-dropout(0.1)-FC(2048,512). The reasoning module in the image is set to $N \in [4, 8, 16, 32, 64]$. During the training process, we optimized the RRAF model using AdamW [55] with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The batch size was set to 64, and binary cross-entropy (BCE) was used as the loss function. The learning rate was set to min $\left(2.5te^{-5}, 1e^{-4}\right)$, where t is the current epoch number starting from 1. After 10 epochs of training, the learning rate was reduced to 1/5 of the current learning rate every two epochs. The VQA 2.0 training set includes the train subset, val subset, and additional *vg* subset, *i.e.*, *train+val+vg*. The vg subset consists of QA pairs from the Visual Genome dataset. The GQA training set consists of the train and val subsets, *i.e.*, *train+val*.

Table 2 GQA dataset sample distribution

Split	Images	Questions	Vocab
Train	72,140	94,300	-
Val	10,234	132,062	-
Test-dev	398	12,578	-
Test	2,987	95,336	-
All	85,759	1,182,976	3,097

4.3 Comparison with state-of-the-art

Table 3 compares the RRAF model and current advanced visual reasoning models on the VQA 2.0 dataset. The comparative experimental results in Table 3 show that using the proposed relation reasoning module and adaptive fusion mechanism can improve accuracy in predicting answers. BAN [59] is a bilinear attention network that considers a bilinear interaction between multimodal inputs to fully utilize information from a question and image features. BAN-Counter [59] combines BAN with Counter, a neural network module that allows robust counting between object proposals and further improves the model's accuracy on counting metrics. The DFAF [12] and MCAN [11] models use deep cooperative attention networks to capture the interaction between modalities effectively. The DFAF model designed an information flow interaction within and between modalities. The MCAN model uses self-attention and question-guided attention units to establish the relationship between modalities. MuRel [45] and ReGAT [30] use graph neural networks to construct deep reasoning networks. MuRel adopts residual characteristic learning for end-to-end reasoning, and ReGAT combines explicit and implicit relationships to achieve complex reasoning. TRRNet [60] adopts the method of reasoning relation features to enhance the model's understanding of the image. ViBERT [61] uses pre-training to effectively improve model performance, with significantly increased model parameter volume compared to end-to-end models. MDFNet [62] proposes a multi-graph reasoning and modality fusion method to achieve fine-grained multimodal reasoning and fusion. Compared with the proposed reasoning module and adaptive fusion method in this paper, the RRAF model presented in this paper performs significantly better than the MDFNet model. The GMA [44] model builds a graph for the question from syntactic and embedded information to implement reasoning between images and questions. The LSAT [9] model enhances the self-attention mechanism in the image decoder, achieving fine-grained reasoning on images through local window interactions. UFSCAN + counter [63] constructs an effective joint feature and spatial co-attention network (UFSCAN) to improve model perfor-

Table 3 Performance comparison on VQA 2.0 with SOTAs

Model	Test-dev				Test-std
	Yes/No	Number	Other	All	All
BAN [59]	85.42	50.93	60.26	69.52	-
BAN-Counter [59]	85.42	54.04	60.52	70.04	70.35
DFAF [12]	86.09	53.32	60.49	70.22	70.34
MuRel [45]	84.77	49.84	57.85	68.03	68.41
ReGAT [30]	86.08	54.42	60.33	70.27	70.58
MCAN [11]	86.82	53.26	60.72	70.63	70.90
ViBERT [61]	-	-	-	70.80	71.10
TRRNet [60]	87.27	51.89	61.02	70.80	71.20
MDFNet [62]	86.85	53.73	**61.78**	**71.19**	71.32
GMA [44]	85.62	51.74	60.51	69.86	70.16
LSAT-R [9]	87.06	53.32	61.04	70.88	71.13
UFSCAN-counter [63]	85.52	54.99	61.08	70.46	70.73
MCAoA [64]	87.05	53.81	60.97	70.90	71.14
CLVIN [65]	87.09	53.65	60.89	70.86	71.16
CAAN [10]	87.09	53.37	61.13	70.94	71.31
RRAF(ours)	**87.03**	**55.39**	60.95	71.06	**71.33**

mance. MCAoA [64] adds attention modules in the encoder and decoder to judge the relationship between attention results and queries. CLVIN [65] and CAAN [10] are both based on the Transformer architecture, aiming to enhance the representation capability of questions and improve the mode of interaction, thereby boosting the model's inference abilities.

It is worth noting that the RRAF model performs well in object counting, as shown in the"Number" column of Table 3. Our model outperforms previous state-of-the-art models by at least 0.4% in this category. In addition, in the counting models specifically designed, BAN-Counter and UFSCAN + counter, although the "Number" category has shown significant improvement, there is no substantial improvement in the accuracy of the "ALL" category. Our model's advantage is in achieving the best performance in the "Number" category and improving accuracy in other categories. The reason is that the RRAF model can simultaneously model visual object semantics, position, and appearance shape and use adaptive fusion mechanisms to effectively help the model improve performance. The MCAN model is the wining model in the 2019 VQA Challenge. The comparison between the RRAF model and the MCAN model shows good accuracy, with an increase of 0.21% in Yes/No, 2.13% in Number, 0.23% in Other, and 0.43% in "All" on *test-dev* and *test-std*, respectively. In addition, the accuracy of the "Number" category compared to advanced counting models UFSCAN + counter and ReGAT increased by 0.4% and 0.97%, respectively, which indicates that our proposed model has good performance.

Table 4 Performance comparison on GQA

Methods	Accuracy	Open	Binary	Validity	Plausibility	Consistency
Human [58]	89.3	87.4	91.2	98.9	97.2	98.4
CNN+LSTM [58]	46.55	31.80	63.26	96.02	84.25	74.57
Bottom-up [19]	49.74	34.83	66.64	96.18	84.57	78.71
MAC [58]	54.06	38.91	71.23	96.16	84.48	81.59
BAN [59]	56.19	41.13	73.31	96.77	85.58	84.64
DMFNet [52]	57.05	41.86	73.98	**97.62**	84.87	86.98
LGCN [40]	56.10	-	-	-	-	-
MCAN_Base [66]	56.00	38.76	75.61	96.69	85.35	87.03
SPCA-Net [66]	57.05	40.20	76.23	96.44	85.23	87.78
GMA [44]	57.26	42.30	73.20	96.41	85.05	83.95
MSRAN [67]	57.56	41.99	75.20	96.05	84.62	**88.61**
RRAF (ours)	**57.83**	**42.32**	**76.50**	96.80	**85.61**	88.01

Table 5 Experimental results of the RRAF model using region features and grid features under different adaptive fusion strategies

Model	VR(s/g/a)	FC_1	FC_2	Yes/No	Number	Other	All
RRAF-R	s	×	×	86.82	53.26	60.72	70.63
RRAF-G	s	×	×	87.43	53.80	61.81	71.45
RRAF-R (32)	s/g/a	×	×	86.79	54.14	60.77	70.74
RRAF-G	s	√	-	87.55	54.53	61.78	71.57
RRAF-G	s	-	√	87.35	54.14	61.71	71.46
RRAF-R	s	√	-	86.95	53.44	60.93	70.81
RRAF-R	s	-	√	87.17	53.27	60.93	70.87

Table 4 presents the comparison results of RRAF with state-of-the-art models on the GQA dataset, where the first row shows human performance, which can be considered as the upper limit of the current VQA task. CNN+LSTM [58] is a method that predicts the answer by linearly combining image and question features. Other models use Faster-RCNN to extract image features. MAC is a milestone model on the GQA reasoning dataset, decomposing a task into continuous reasoning steps. LGCN [40] model visual object regions using neural graph networks and complete VQA tasks by jointly inferring semantic relationships among visual objects and relationships among attributes. DMFNet [52] uses multiple graph reasoning and fusion layers with pre-trained semantic relationship embedding to explain complex spatial and semantic relationships between visual objects. SPCA-Net [66] and MSRAN [67] employ encoder-decoder structures to infer images using spatial coordinates of visual objects. According to Table 4, the RRAF model has higher accuracy compared to the current state-of-the-art reasoning models. Compared with the latest GMA [44] model, RRAF has improved in "Accuracy", "Open", "Binary", and "Consistency" by 0.57%, 0.02%, 2.66%, and 3.59%, respectively, but "Validity" did not reach the current best level. We speculate that incorporating the visual object positional relationships for reasoning and adaptive fusion mechanism is effective. However, the original baseline model has limitations in checking "Validity" and the rationality of the answer within the scope of the question.

4.4 Ablation study

This section primarily discusses the test results of the RRAF model on the VQA 2.0 and GQA benchmark datasets *test-dev*. In order to analyze the role of each module in the model and demonstrate the superiority of the proposed method, ablation experiments were performed on the complete model on the two benchmark datasets to explore the function of each module. The RRAF model variants we propose are discussed as follows.

4.4.1 Ablation analysis on VQA 2.0 and GQA dataset

Table 5 shows the performance of the RRAF model using region and grid visual features under two adaptive fusion strategies. "RRAF-R" and "RRAF-G" respectively represent the use of region visual feature $m \in [10, 100]$ and the grid visual feature, where the grid visual feature has a resolution of 8×8. VR(s/g/a) represents the image reasoning module, where s, g, and a reference to semantic, positional, and appearance reasoning for visual objects, respectively. FC_1 and FC_2 represent two different adaptive fusion strategies. The experimental results in Table 5 show that the adaptive fusion mechanism performs better than the baseline model. To better verify the effectiveness of the proposed adaptive fusion mechanism, we conducted a validation experiment using grid visual features as the baseline model. In addition, due to the absence of detection boxes for grid visual features, it was impossible to model the positional relationship. Therefore we should have included a position reasoning module in the ablation experiment. Based on the comparative experimental results, the adaptive fusion method can substantially enhance the overall accuracy of the model. However, this improvement is contingent on a simplistic analysis of image and question features for adaptive fusion. Consequently, the model cannot attain superior reasoning abilities and achieve optimal outcomes. More complex feature analysis techniques

Table 6 Ablation experiments on different reasoning modules under the exploration of region visual features and adaptive fusion strategy FC_1

Model	VR(s/g/a)	FC_1	FC_2	Yes/No	Number	Other	All
RRAF-R (4)	s/g/a	√	-	**87.05**	54.34	60.64	70.80
RRAF-R (8)	s/g/a	√	-	86.88	54.50	60.89	70.87
RRAF-R (16)	s/g/a	√	-	86.88	54.65	60.66	70.78
RRAF-R (32)	s/g/a	√	-	86.88	55.05	**61.03**	71.00
RRAF-R (64)	s/g/a	√	-	87.03	**55.39**	60.95	**71.06**

Table 7 Exploring ablation experiments on varying numbers of reasoning modules based on region visual features and the adaptive fusion strategy FC_2

Model	VR(s/g/a)	FC_1	FC_2	Yes/No	Number	Other	All
RRAF-R (4)	s/g/a	-	√	86.84	54.69	60.87	70.87
RRAF-R (8)	s/g/a	-	√	86.88	54.78	60.91	70.91
RRAF-R (16)	s/g/a	-	√	86.95	54.53	60.84	70.88
RRAF-R (32)	s/g/a	-	√	86.99	54.79	60.93	70.97
RRAF-R (64)	s/g/a	-	√	**86.99**	**54.80**	**60.95**	**70.98**

may need to enhance the model's reasoning capacity further and achieve outstanding results.

Table 6 explores the impact of a different number of reasoning modules on the model's performance when using the first adaptive fusion mechanism. As shown in the table, using the relation reasoning module can effectively improve the accuracy of the "Number" category. As the number of attention heads in the relation reasoning module increases, the accuracy of the model's "Number" and "All" categories also improves continuously. Due to limited computing resources, we only verified the attention head number of object relation reasoning to 64. Among them, the RRAF-R (64) is the model that we finally adopted.

Table 7 analyzes the effect of using a different numbers of reasoning module heads on the performance of the model when using the second adaptive fusion mechanism. The table shows that although the second adaptive fusion mechanism can significantly improve the performance of the model, its effect could be better than that of the first adaptive fusion strategy. This may be due to the imbalance in modal allocation between the question and image features, since the weight contribution of the automatically allocated question features is ε (see (27)). In contrast, that of the image features is $1-\varepsilon$; this also indicates the necessity of designing a reasonable adaptive fusion module to correctly predict answers in VQA models.

Table 8 shows the performance of the variant models of RRAF on the benchmark dataset GQA. The table compares the differences between the models that do not use adaptive fusion strategies and visual object relation reasoning modules and our proposed model. All variant models on the GQA dataset are trained using region visual features and detection boxes (frcn+bbox) for training and validation. Table 6 shows that employing our proposed visual object relation reasoning module and adaptive fusion mechanism can improve the model's overall performance. Moreover, the first adaptive fusion mechanism is more reasonable than the second. Since we did not use question information to guide the construction of the image encoder, some irrelevant information may have been incorporated into the final image features. Accordingly, how can one filter out these irrelevant features again? Undoubtedly, using the adaptive fusion mechanism is a simple and effective method. It does not lose essential image features and also improves overall model performance.

As shown in Table 9, the RRAF model effectively improves model performance and does not excessively increase the number of parameters. In contrast, the MCAN+PA [38] model introduces position information in intra-modal interaction learning, which can adaptively adjust the inter-modal interaction according to different inputs. Although it achieved good results, it increases many parameters and performs worse than the RRAF model in the "Number" category.

Table 8 The experimental results of RRAF using different modules on the GQA benchmark dataset

Model	VR(s/g/a)	FC_1	FC_2	Accuracy	Open	Binary	Validity	Plausibility	Consistency
RRAF-R	s	×	×	56.00	75.61	38.76	96.69	85.35	87.03
RRAF-R	s	√	-	56.98	74.86	41.25	96.90	85.40	87.61
RRAF-R	s	-	√	56.84	74.91	40.95	96.86	85.26	87.43
RRAF-R (32)	s/g/a	×	×	57.18	75.68	41.23	96.72	85.45	87.12
RRAF-R (32)	s/g/a	√	-	57.48	75.74	41.73	97.02	85.75	87.12
RRAF-R (32)	s/g/a	-	√	57.46	74.84	41.37	97.12	**85.78**	87.02
RRAF-R (64)	s/g/a	×	×	57.23	74.98	41.51	**96.98**	85.62	86.34
RRAF-R (64)	s/g/a	√	-	**57.83**	**76.50**	**42.32**	96.80	85.61	**88.01**
RRAF-R (64)	s/g/a	-	√	57.79	75.55	42.27	96.85	85.63	86.99

Table 9 Comparison of parameter size and performance of RRAF model on VQA 2.0 and GQA datasets

Model	Dataset	Params	Test-dev	Test-std
MCAN [11]	VQA 2.0	57.80M	70.63	70.90
MCAN+PA [38]	VQA 2.0	68.50M	71.05	71.52
MCAN	GQA	53.06M	56.00	-
RRAF-R (32)	VQA 2.0	58.19M	71.00	71.23
RRAF-R (32)	GQA	54.80M	57.48	-
RRAF-R (64)	VQA 2.0	61.34M	71.06	71.33
RRAF-R (64)	GQA	55.60M	57.83	-

In this study, we followed the above experimental settings and trained the VQA 2.0 dataset using the *train+val+vg* training method and the GQA dataset using the *train+val* training method. Since the adaptive fusion mechanism does not change the number of parameters of the model, we used the results of the first adaptive fusion mechanism when comparing the RRAF model in the table (see Table 4). According to the experimental results, increasing the number of visual objects for reasoning relation attention from 32 to 64 only increased the parameters by 1.61M but significantly improved performance. Notably, we proposed a plug-and-play visual object relation reasoning method in end-to-end VQA tasks, which can be applied to different VQA model tasks and effectively improve the model performance.

4.5 Qualitative analysis

In order to better demonstrate the effectiveness of the proposed visual object relation module and adaptive fusion mechanism, we conducted ablation experiments and compared the image reasoning process using different attention modules. As shown in the visual examples in the first row of Fig. 5 (a) represents the case where object position and appearance are not modeled. The model only focuses on the "computers" on the table based on semantic information and cannot effectively utilize information from surrounding computer objects for reasoning. At the same time, it only focuses on the local area (screen color) of the middle computer based on semantic information, leading the model to believe that "black" is the correct answer. (b) When object position, semantics, and appearance (VR-s/a/g) are modeled, reasoning more accurately based on semantic information and calculating the relationship weights between different objects is possible. However, irrelevant information unrelated to objects is easily introduced during modeling. When using visual object relation reasoning to learn visual relation reasoning and the adaptive fusion mechanism (VR-s/a/g+FCn), the model can better combine the semantic information of the problem and effectively filter out irrelevant visual information. This enables the model to answer complex reasoning questions accurately.

In the second row of Fig. 5, we show a counting question for complex objects in visual objects. To accurately answer

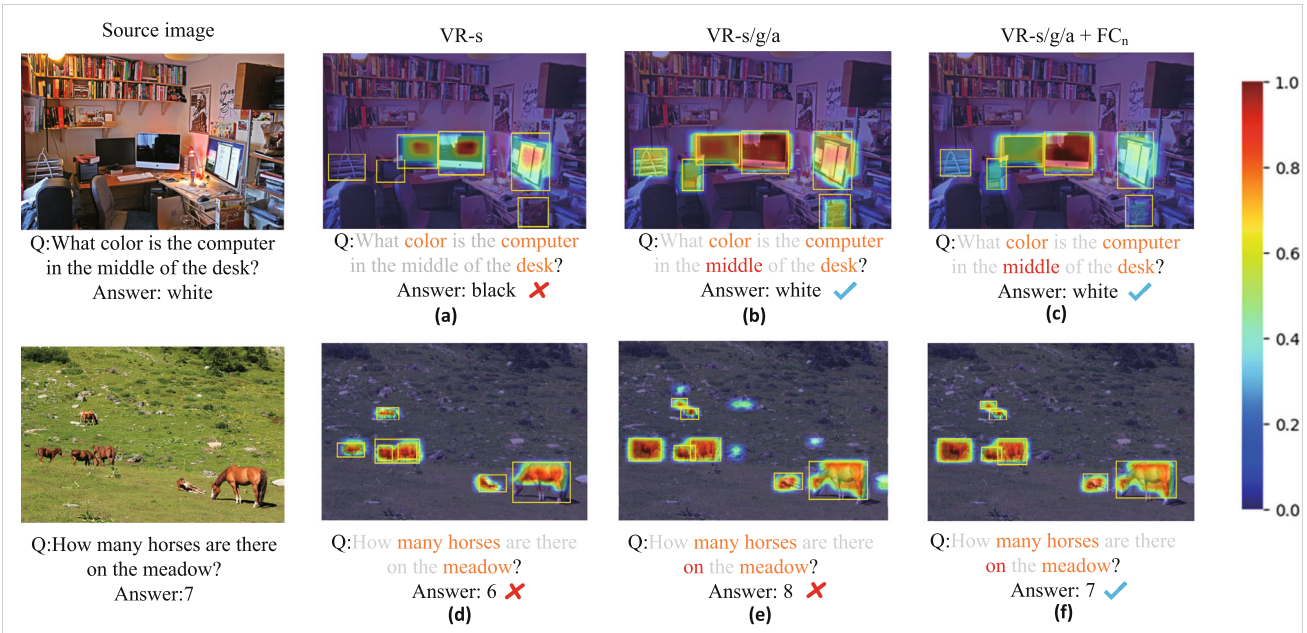

Fig. 5 Visual object relation reasoning attention visualization example based on adaptive fusion mechanism

the count of complex targets (small or overlapping targets) in visual objects, it is necessary to understand and calculate the image accurately. (d) represents a method that only focuses on visual objects based on semantic information, which does not help the model to better understand the positional relationship between objects and is prone to ignoring counting of complex objects. (e) depicts a modeling approach that integrates visual object semantics, position, and appearance information, enabling comprehension and reasoning among objects. However, this approach also considers irrelevant information surrounding visual objects, leading to the inclusion of information unrelated to objects and potential errors in model calculations. (f) shows the attention result using the adaptive fusion mechanism and visual object relation reasoning. The model understands the appearance and positional relationship between visual objects and also uses an adaptive fusion mechanism to avoid irrelevant information from participating in the features of the predicted answer. Therefore, the model can effectively perform reasoning and calculation for complex problems.

5 Conclusion and future work

We proposed a visual object relation reasoning method based on an adaptive fusion mechanism to efficiently reason complex relationships among visual objects thru a visual relationship reasoning and adaptive fusion (RRAF) model. The model adopt an efficient and applicable image encoder that can simultaneously learn the interaction between visual objects' position and appearance features, achieving spatial positional relationship reasoning and appearance shape reasoning for complex visual objects. In a deep co-attention network, the question semantic-guided attention mechanism can achieve more accurate semantic alignment between image regions and problem text information. Meanwhile, the adaptive fusion mechanism can reassign the contribution of modalities based on the question and semantic features, effectively filtering out irrelevant information features and ensuring consistency between the relationships among visual objects and the position descriptions in the problem. This paper conducted ablation experiments on the model to verify its performance, and the results show that each component in the model plays an important role. Performance analysis on the VQA 2.0 and GQA benchmark datasets shows that the RRAF model effectively improved the accuracy of the model's counting category and also improves overall model performance. This indicates that the RRAF can better understand visual content and reason multiple visual relationships (position, semantics, and appearance) among visual regions.

This work explored the inherent property relationships of visual objects. However, there are many other interaction relationships among visual objects, such as spatial positional and action interaction behavior relationships. In future research work, more visual reasoning relationships will be explored, and novel natural language reasoning methods will also continue to be explored to help machines think and understand problems more intelligently. Additionally, RRAF is an easily transferable network model that can be used in related visual language tasks to improve model performance.

Acknowledgements This research received partial funding from the National Natural Science Foundation of China (Grant No. 52331012) and the Natural Science Foundation of Shanghai (Grant No. 21ZR1426500). Additionally, support was provided by the Hunan Provincial Natural Science Foundation (Grant No. 2022JJ50245) and the Scientific Research Fund of Hunan Provincial Education Department (Grant No. 22B0753). This work also supported by the Shanghai Maritime University's Top Innovative Talent Training Program (Grant No. 2022YBR014) for Graduate Students in 2022.

Author Contributions Methodology, material preparation, data collection, and analysis were performed by Xiang Shen and Zihan Guo. Xiang Shen wrote the first draft of the manuscript, Liang Zong and Zihan Guo commented on previous versions of the manuscript. Dezhi Han did the supervision, reviewing, and editing. All authors read and approved the final manuscript.

Declarations

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D (2015) Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp 2425–2433
2. Wang Y, Xu N, Liu A-A, Li W, Zhang Y (2021) High-order interaction learning for image captioning. IEEE Trans Circuits Syst Video Technol 32(7):4417–4430
3. Stefanini M, Cornia M, Baraldi L, Cascianelli S, Fiameni G, Cucchiara R (2022) From show to tell: A survey on deep learning-based image captioning. IEEE Trans Pattern Anal Mach Intell 45(1):539–559
4. Deng J, Yang Z, Liu D, Chen T, Zhou W, Zhang Y, Li H, Ouyang W (2023) Transvg++: End-to-end visual grounding with language conditioned vision transformer. IEEE Trans Pattern Anal Mach Intell
5. Hu P, Peng D, Wang X, Xiang Y (2019) Multimodal adversarial network for cross-modal retrieval. Knowl-Based Syst 180:38–50
6. Xu X, Lin K, Yang Y, Hanjalic A, Shen HT (2020) Joint feature synthesis and embedding: Adversarial cross-modal retrieval revisited. IEEE Trans Pattern Anal Mach Intell 44(6):3030–3047
7. Esposito M, Damiano E, Minutolo A, De Pietro G, Fujita H (2020) Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering. Inform Sci 514:88–105
8. Nguyen BX, Do T, Tran H, Tjiputra E, Tran QD, Nguyen A (2022) Coarse-to-fine reasoning for visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4558–4566

9. Shen X, Han D, Guo Z, Chen C, Hua J, Luo G (2022) Local self-attention in transformer for visual question answering. Appl Intell 1–18
10. Chen C, Han D, Chang C-C (2022) Caan: Context-aware attention network for visual question answering. Pattern Recognition 132:108980
11. Yu Z, Yu J, Cui Y, Tao D, Tian Q (2019) Deep modular co-attention networks for visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6281–6290
12. Gao P, Jiang Z, You H, Lu P, Hoi SC, Wang X, Li H (2019) Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6639–6648
13. Zhang H, Zeng P, Hu Y, Qian J, Song J, Gao L (2023) Learning visual question answering on controlled semantic noisy labels. Pattern Recognition 138:109339
14. Yanagimoto H, Nakatani R, Hashimoto K (2022) Visual question answering focusing on object positional relation with capsule network. In: 2022 12th International congress on advanced applied informatics (IIAI-AAI), IEEE, pp 89–94
15. Das A, Agrawal H, Zitnick L, Parikh D, Batra D (2017) Human attention in visual question answering: Do humans and deep networks look at the same regions? Comput Vision Image Understand 163:90–100
16. Fukui A, Park DH, Yang D, Rohrbach A, Darrell T, Rohrbach M (2016) Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv:1606.01847
17. Yu Z, Yu J, Xiang C, Fan J, Tao D (2018) Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans Neural Netw Learn Syst 29(12):5947–5959
18. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems 30
19. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6077–6086
20. Gao P, Jiang Z, You H, Lu P, Hoi SC, Wang X, Li H (2019) Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6639–6648
21. Wang C, Shen Y, Ji L (2022) Geometry attention transformer with position-aware lstms for image captioning. Expert Syst Appl 201:117174
22. Hu H, Gu J, Zhang Z, Dai J, Wei Y (2018) Relation networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3588–3597
23. Wei J, Li Z, Zhu J, Ma H (2022) Enhance understanding and reasoning ability for image captioning. Appl Intell 1–17
24. Gerrish S (2018) How Smart Machines Think. The MIT Press, London
25. Guo Z, Han D (2023) Sparse co-attention visual question answering networks based on thresholds. Appl Intell 53(1):586–600
26. Zhou Y, Ren T, Zhu C, Sun X, Liu J, Ding X, Xu M, Ji R (2021) Trar: Routing the attention spans in transformer for visual question answering. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2074–2084
27. Shen X, Han D, Chang C-C, Zong L (2022) Dual self-guided attention with sparse question networks for visual question answering. IEICE Trans Inform Syst 105(4):785–796
28. Yang Z, He X, Gao J, Deng L, Smola A (2016) Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 21–29
29. Do T, Do T-T, Tran H, Tjiputra E, Tran QD (2019) Compact trilinear interaction for visual question answering. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 392–401
30. Li L, Gan Z, Cheng Y, Liu J (2019) Relation-aware graph attention network for visual question answering. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10313–10322
31. Zhang D, Cao R, Wu S (2019) Information fusion in visual question answering: A survey. Inform Fusion 52:268–280
32. Ben-Younes H, Cadene R, Cord M, Thome N (2017) Mutan: Multimodal tucker fusion for visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp 2612–2620
33. Jiang H, Misra I, Rohrbach M, Learned-Miller E, Chen X (2020) In defense of grid features for visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10267–10276
34. Nguyen A, Tran QD, Do T-T, Reid I, Caldwell DG, Tsagarakis NG (2019) Object captioning and retrieval with natural language. In: Proceedings of the IEEE/CVF international conference on computer vision workshops, pp 0–0
35. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
36. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805
37. Zhao Z, Samel K, Chen B et al (2021) Proto: Program-guided transformer for program-guided tasks. Advances in neural information processing systems 34:17021–17036
38. Mao A, Yang Z, Lin K, Xuan J, Liu Y-J (2022) Positional attention guided transformer-like architecture for visual question answering. IEEE Trans Multimed
39. Li W, Sun J, Liu G, Zhao L, Fang X (2020) Visual question answering with attention transfer and a cross-modal gating mechanism. Pattern Recognition Lett 133:334–340
40. Hu R, Rohrbach A, Darrell T, Saenko K (2019) Language-conditioned graph networks for relational reasoning. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10294–10303
41. Yu J, Zhang W, Lu Y, Qin Z, Hu Y, Tan J, Wu Q (2020) Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval. IEEE Trans Multimed 22(12):3196–3209
42. Huang Q, Wei J, Cai Y, Zheng C, Chen J, Leung H-f, Li Q (2020) Aligned dual channel graph convolutional network for visual question answering. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 7166–7176
43. Yang Z, Qin Z, Yu J, Wan T (2020) Prior visual relationship reasoning for visual question answering. In: 2020 IEEE International conference on image processing (ICIP), IEEE, pp 1411–1415
44. Cao J, Qin X, Zhao S, Shen J (2022) Bilateral cross-modality graph matching attention for feature fusion in visual question answering. IEEE Trans Neural Netw Learn Syst
45. Cadene R, Ben-Younes H, Cord M, Thome N (2019) Murel: Multimodal relational reasoning for visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1989–1998
46. Liu Y, Wei W, Peng D, Mao X-L, He Z, Zhou P (2022) Depth-aware and semantic guided relational attention network for visual question answering. IEEE Trans Multimed
47. Chen H, Liu R, Peng B (2021) Cross-modal relational reasoning network for visual question answering. In: Proceedings of the

IEEE/CVF international conference on computer vision, pp 3956–3965
48. Zhang J, Huang B, Fujita H, Zeng G, Liu J (2023) Feqa: Fusion and enhancement of multi-source knowledge on question answering. Expert Syst Appl 227:120286
49. Kim J-H, On K-W, Lim W, Kim J, Ha J-W, Zhang B-T (2016) Hadamard product for low-rank bilinear pooling. arXiv:1610.04325
50. Gu G, Kim ST, Ro YM (2017) Adaptive attention fusion network for visual question answering. In: 2017 IEEE International conference on multimedia and expo (ICME), IEEE, pp 997–1002
51. Chen H, Liu R, Fang H, Zhang X (2021) Adaptive re-balancing network with gate mechanism for long-tailed visual question answering. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 3605–3609
52. Zhang W, Yu J, Zhao W, Ran C (2021) Dmrfnet: deep multimodal reasoning and fusion for visual question answering and explanation generation. Inform Fusion 72:70–79
53. Ren S, He K, Girshick RB, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149
54. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li L-J, Shamma DA et al (2017) Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int J Comput Vision 123:32–73
55. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
56. Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D (2017) Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6904–6913
57. Pishchulin L, Insafutdinov E, Tang S, Andres B, Andriluka M, Gehler PV, Schiele B (2016) Deepcut: Joint subset partition and labeling for multi person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4929–4937
58. Hudson DA, Manning CD (2019) Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6700–6709
59. Kim J-H, Jun J, Zhang B-T (2018) Bilinear attention networks. In: Advances in neural information processing systems 31
60. Yang X, Lin G, Lv F, Liu F (2020) Trrnet: Tiered relation reasoning for compositional visual question answering. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16, Springer, pp 414–430
61. Lu J, Batra D, Parikh D, Lee S (2019) Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in neural information processing systems 32
62. Zhang W, Yu J, Wang Y, Wang W (2021) Multimodal deep fusion for image question answering. Knowl-Based Syst 212:106639
63. Zhang S, Chen M, Chen J, Zou F, Li Y-F, Lu P (2021) Multimodal feature-wise co-attention method for visual question answering. Inform Fusion 73:1–10
64. Rahman T, Chou S-H, Sigal L, Carenini G (2021) An improved attention for visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1653–1662
65. Chen C, Han D, Shen X (2023) Clvin: Complete language-vision interaction network for visual question answering. Knowl-Based Syst 110706
66. Yan F, Silamu W, Li Y, Chai Y (2022) Spca-net: a based on spatial position relationship co-attention network for visual question answering. Visual Comput 38(9–10):3097–3108
67. Yao H, Wang L, Cai C, Sun Y, Zhang Z, Luo Y (2023) Multi-modal spatial relational attention networks for visual question answering. Image Vision Comput 140:104840

Xiang Shen is currently pursuing the Ph.D. degree with Shanghai Maritime Unversity. His research interests include visual question answering and maching learning.

Dezhi Han received the B.S. degree in applied physics from the Hefei University of Technology, Hefei, China, in 1990, and the 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 2006. His current research interests include cloud and outsourcing security, wireless communication security, network and information security, Visual Question Answering. He is currently a Member of the IEEE.

Liang Zong received B.S. degree from Sichuan University of Science and Engineering in 2006, Zigong, China, M.S. degree from Ningbo University in 2010, Ningbo, China, and Ph.D. degree from Hainan University in 2016, Haikou, China. From September 2017 to December 2017, he was a visiting scholar in the Faculty of Science and Engineering, Anglia Ruskin University, UK. He is the director of the Internet of things lab with the College of Information Engineering, Shaoyang University, China. He is a member of Hunan Provincial Key Laboratory of Information Service in Rural Southwest Hunan, and also a member of Hunan Provincial Academician Expert Workstation Team. His current research interest is satellite networks, machine learning, and transmission control for air-space ground network.

Zihan Guo received his B.E. degree from Tianjin Polytechnic University, China, in 2017, and his Ph.D. degree 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.

Jie Hua received his B.S. degree in software engineering from Wuhan University, Wuhan, China, in 1999 and a PhD degree in software engineering from University of Technology Sydney (UTS), Sydney, Australia, in 2014. He is currently working as Professor at Shaoyang University and a researcher at UTS. His research interests include graph drawing, visualisation, big data, and deep learning.

Authors and Affiliations

Xiang Shen[1] · Dezhi Han[1] · Liang Zong[2] · Zihan Guo[3] · Jie Hua[4]

Xiang Shen
shenxiang1107@163.com

Liang Zong
zongliang@hnsyu.edu.cn

Zihan Guo
guo_zihan11@163.com

Jie Hua
jie.hua@uts.edu.au

1 College of Information Engineering, Shanghai Maritime University, 1550 Haigang Avenue, Shanghai 201306, People's Republic of China

2 College of Information Engineering, Shaoyang University, Shaoyang 422099, People's Republic of China

3 College of Information Engineering, Changzhi University, Changzhi, Shanxi 046011, People's Republic of China

4 TD School, University of Technology, Sydney, Sydney, Ultimo, NSW 2007, Australia