Multimedia Tools and Applications (2024) 83:59577–59595
https://doi.org/10.1007/s11042-023-17945-8

Integrating multimodal features by a two-way co-attention mechanism for visual question answering

Himanshu Sharma[1] · **Swati Srivastava**[1]

Received: 22 May 2022 / Revised: 12 October 2023 / Accepted: 17 December 2023 /
Published online: 29 December 2023

Abstract
Existing VQA models predominantly rely on attention mechanisms that prioritize spatial dimensions, adjusting the importance of image regions or word token features based on spatial probabilities. However, these approaches often struggle with relational reasoning, treating objects independently, and failing to fuse their features effectively. This hampers the model's ability to understand complex visual contexts and provide accurate answers. To address these limitations, our innovation introduces a novel co-attention mechanism in the VQA model. This mechanism enhances Faster R-CNN's feature extraction by emphasizing image regions relevant to the posed question. This, in turn, improves the model's ability for visual relationship reasoning, making it more adept at analyzing complex visual contexts. Additionally, our model incorporates feature-wise multimodal two-way co-attentions, enabling seamless integration of image and question representations, resulting in more precise answer predictions. Our model achieves impressive results on VQA 1.0, surpassing the best existing model, Re-attention model by 1.14% on test-std. Moreover, on VQA 2.0, our model outperforms the best model, IAHOT model by a significant margin of 2.98% on test-std. These findings demonstrate that our approach not only outperforms earlier models but also establishes a new state-of-the-art performance level in Visual Question Answering.

Keywords VQA · Attention · Co-attention · Multimodal · Relational reasoning

1 Introduction

The remarkable advancement in deep learning has significantly advanced artificial intelligence (AI) [1] research, particularly in the fields of computer vision and natural language processing. Among the most recent and attractive areas in AI research today is Visual Question Answering (VQA) [2]. VQA involves generating natural language answers for questions related to given images, demanding a strong grasp of both visual

✉ Himanshu Sharma
himanshu.sharma@gla.ac.in

Swati Srivastava
swati.srivastava@gla.ac.in

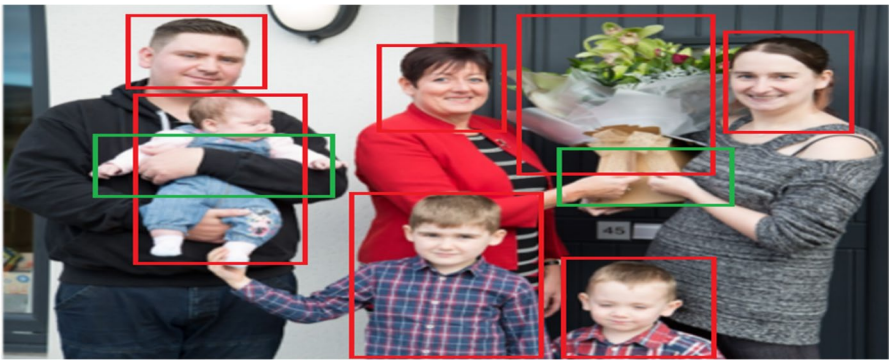
[1] Department of Computer Engineering and Applications, GLA University Mathura, Mathura, India

content and textual queries. VQA holds immense potential for diverse applications across domains [62]. It can aid visually impaired individuals in understanding their surroundings, facilitate seamless human-computer interaction, serve as a knowledgeable assistant for children and healthcare professionals, and even offer entertainment solutions. To accomplish VQA tasks, a sophisticated understanding of both images and questions is essential. Convolutional Neural Networks (CNNs) [3] or Faster R-CNN [4] are employed to extract discriminative features from image-question pairs. Correct answer generation relies on effectively extracting visual information relevant to the associated question from the corresponding image. Overall, VQA's concurrent comprehension of images and corresponding questions in natural language opens up exciting possibilities for innovative AI applications in various domains.

Mechanisms of attention have shown to be valuable tools to achieve this understanding. Especially attention-based models [5], which have been extensively studied, researchers have effectively utilized visual attention techniques in VQA tasks in recent years. These attention-based models produce spatial maps that employ visual cues to draw attention to pertinent image regions that are essential for resolving the inquiry. Traditional attention techniques including region-based, object-based, and semantic-based attention, however, have some drawbacks. They concentrate plenty of focus on specific visual aspects, such as regions, objects, or semantics, which may not adequately capture the complexity of feature representations needed for comprehensive VQA. The high-level semantic significance associated with both images and questions has presented difficulties in several past studies, including [6]. In addition, [7] has difficulty responding to inquiries requiring common sense, and its performance is hampered by problems like duplicates or missing detections. Innovative attention mechanisms that can better comprehend the complex relationships between visual and textual information, address the drawbacks of conventional approaches, and provide more precise and sophisticated answers to a variety of questions are required to advance VQA capabilities.

In this paper, we suggest an extensive approach to overcome the drawbacks of Visual Question Answering (VQA) roles that were previously emphasized. A question attention mechanism that concentrates on the most crucial terms in the phrase to enhance understanding of question semantics is introduced. We take the input query and extract word-level representations using the Gated Recurrent Unit (GRU) [8]. We integrate relational reasoning and the visual attention mechanism to generate enhanced image-question fusion. This combination enhances the model's understanding of the relationship between the representations of the image and the answer, resulting in enhanced VQA performance. To extract fine-grained features, we use a co-attention technique in which the image and the question are alternated. This iterative process allows the model to progressively emphasize appropriate visual information, enabling better collaboration between the image and question. We utilize Faster R-CNN, which performs better than earlier approaches like [7], for object detection to address the problem of duplicate or missed object detection. Additionally, we present a visual spatial attention module that highlights areas of the image that are highly relevant to the posed question. By focusing on the most instructive visual cues, the model may then precisely reply to questions.

In Fig. 1, accurately answering questions requires the model to not only count individuals but also focus on their age attributes for questions like "How many kids are there?" and "Are the kids of the same age?" Additionally, the model needs to employ relational reasoning to understand concepts like "holding" and "color of dresses," and fuse these relationships with visual features to answer questions like "What is the color of the dress of the youngest kid?" Combining visual relational reasoning and attention



Q. How many kids are there?
A. Three

Q. Are the kids of same age?
A. No

Q. What is the youngest kid's dress color?
A. Blue

Fig. 1 In this VQA example, the model detects distinct visual features (represented by red boxes) and utilizes visual reasoning to gain a better understanding of the image. This improved comprehension enhances the model's ability to predict answers accurately for questions about the attributes of objects in the scene, such as the number of kids, their ages, and the color of the dress worn by the youngest kid (related to green boxes representing semantics)

mechanisms empowers the model to obtain more fine-grained features, leading to enhanced VQA performance.

The significant contributions of this paper are:

- In the VQA model, we introduced a completely novel co-attention mechanism that enables Faster R-CNN to extract salient visual features and top-down visual attention to emphasize relevant regions corresponding to the question, strengthening visual relationship reasoning and reducing the impact of irrelevant features.
- We proposed Question-Adaptive Visual Attention Module (QA-VAM) and Question-Guided Region Attention Module (QG-RAM) to improve the precision of our answers and enhance question-answering accuracy. Both modules emphasize the image regions, which are significant to the words of a question.
- In our approach, image and question representations fuse through both feature-wise multimodal two-way co-attentions. By doing so, our model learns visual relations and attentions for specific image regions, enabling more accurate answer predictions.
- We conducted comprehensive evaluations using widely used VQA datasets: VQA 1.0 and VQA 2.0. The results exhibit that our approach performs exceptionally well at generating accurate answers.

2 Related works

Since the last decade, one of the most recent and fascinating topics in the area of computer vision is VQA [9–12]. By leveraging visual regions that are pertinent to the question, attention-based approaches [13–18] train the model to deliver the correct answers. Relational reasoning-based models [6, 19–21] mostly employ neural networks to model relationships among visual objects.

2.1 Visual question answering

The task of VQA gained significant interest from computer vision (CV) and natural language processing (NLP) domains. In recent years, researchers introduced various models to address the VQA task for normal and medical images. For instance, Zhang et al. [22] suggested a method based on a generative paradigm for addressing VQA on medical images by understanding the visual information using an encoder-decoder. Jiang et al. [23] found that grid features can work well for VQA which is faster than bottom-up region features that were computationally expensive where the semantic features play an important role in the effectiveness of the model. Chen et al. [24] suggested a model based on the synthesis of counterfactual samples that focus on visual objects and words for improved answering abilities by generating various counterfactual training samples and assigning ground-truth answers. To obtain suitable answers, Sharma and Jalal [25] developed a model that employs the knowledge gained from the image captions for the task of VQA. In this endeavor, the visual features from the image captioning task are integrated with the attended visual features.

2.2 Attention mechanism-based methods

By integrating the information from the question into the process of extracting deep visual features [63], the attention mechanism has enhanced the efficiency of VQA models. Consequently, the VQA approaches rely extensively on attention. In VQA tasks, most of the attentional approaches generate question-guided attention on visual regions. For instance, Yu et al. [26] introduced a co-attention network consisting of cascaded layers where both the self and guided attentions are present in each layer to model the interactions using the encoder-decoder approach for VQA. Li et al. [27] suggested a graph attention network [64] that encodes each image into a graph and models the object relations. Sharma and Jalal [21] proposed a model with two attention modules that exploit each other's knowledge for feature extraction to enhance the answering abilities. The relations between image regions and objects are employed by a graph neural network to generate captions, which are then used in the last layer of the hybrid architecture for answer prediction.

2.3 Visual relational reasoning

Relational reasoning plays a significant role in visual understanding which encompasses the relationship among visual objects. Only a visual understanding of each region separately cannot give sufficient information. To obtain reasonable relationship information, multiple regions need to be combined. Recent approaches use statistical learning on knowledge bases to perform relational reasoning. For example, Visual attention has been used to build an effective attention map on image regions. In an attempt to achieve better multi-modal feature fusion, Zhang et al. [20] suggested a module to reason complex relationships between visual objects by bringing together visual relationship and attention. Wu et al. [6] introduced a deep neural network to fuse multi-modal data where region-based attention focuses on question-related regions that generate distinctive features to offer accurate question-guided answers. For the study to generate acceptable answers, Cadene et al. [19]

suggested a model to represent end-to-end interactions between the input image, input regions, and the question.

2.4 Motivation

Our research aims to address the constraints of current VQA methods through a novel approach. Table 1 presents an overview of the limitations associated with existing methods. The integration of an attention mechanism enhances the integration of visual and linguistic features in visual question answering (VQA), enabling dense and bi-directional relations between the image and corresponding questions. The precision of answers predicted by VQA systems is significantly improved by employing the attention mechanism. The technique we propose introduces an entirely innovative co-attention approach that enhances the synthesis of visual and linguistic representations. This mechanism generates attention maps on both the image regions for each question word and the question words for individual image areas. By performing attended feature computation, multimodal representation concatenation, and transformation using a single-layered network with ReLU and remaining associations, our method enables comprehensive relations between all image regions and question words. We refer to this composite network as a dense co-attention layer, which can be arranged to create a hierarchical structure facilitating multi-step relations between the image and question in a fully symmetric manner.

3 Proposed method

An overview of our sophisticated Visual Question Answering (VQA) model and details on each component are presented in this section. Our method approaches the VQA task as a problem of classification, where the objective is to predict the most plausible answer $\widehat{a}$ (Eq. 1) from predefined responses based on image I and question q.

$$\widehat{a} = arg\max_{a} P(a/I, q) \tag{1}$$

where $a \in \{a_1, a_2, \ldots \ldots, a_M\}$ are the most common responses (answers) from the training data.

In Fig. 2, the proposed model executes visual relational reasoning and visual attention through utilizing the question, and bottom-up attention is used to generate the

Table 1 Limitations of existing methods

Methods	Limitations
MSRR [6]	They were unable to handle questions and images having high-level semantic meaning
PMC-VQA [22]	They used the ACC and Bleu scores as evaluation metrics, however since these metrics only consider string similarity irrespective of word order, they are incapable of capturing the fluency of the generated sentences
ReGAT [27]	They fail to fuse the semantic, spatial, and implicit relations and also in the utilization of each relation to answer specific questions
MCAN [26]	Their approach sometimes fails to distinguish the keywords in questions
UFSCAN [7]	Their approach fails to provide solutions to issues requiring some common sense

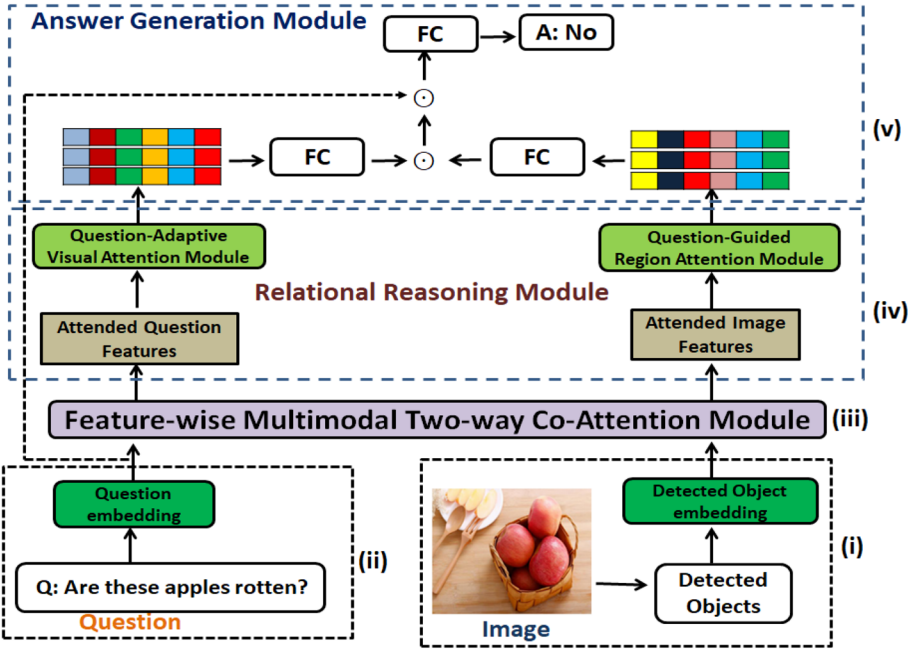


Fig. 2 The proposed VQA model framework. (i) Our model uses Faster R-CNN to capture visual features from K image regions. (ii) The question features are obtained using word embedding and GRU. (iii) We then employ FMulCoA to model the feature-wise using a two-way co-attention (MulIFA and MulQFA) module. (iv) These features are further used in the visual relational reasoning module which includes QA-VAM to obtain the fine-grained visual features and QG-RAM to generate spatial attention features related to the question. (v) Finally, the model predicts an answer from a set of possible answers using a multi-label classifier

boundary boxes as input. This enables the ability to generate answers that are precise and accurate. The model comprises five main components: (1) To extract visual features from K image regions, image modeling employs bottom-up attention based on ResNet within a Faster R-CNN architecture. (2) Question modeling, where the given question is minimized to a limit of 14 words and converted using word embeddings into vector representations. A Gated Recurrent Unit (GRU) processes these vectors, producing the final question representation. (3) The Feature-wise Multimodal Two-way Co-Attention Module (FMulCoA) module, explained in Section 3.3, includes Multimodal Image-Guided Feature-Wise Attention (MulIFA) module and Multimodal Question-Guided Feature-wise Attention (MulQFA) module, generating feature-wise attention features that enhance distinctiveness and fine-grained recognition capabilities. To identify relevant image regions, including spatial dimensions (object dimension), a Question-Adaptive Visual Attention Module (QA-VAM) is utilized as described in Section 3.4.1. (4) Detected image region proposals have been assigned weights by the Question-Guided Region Attention Module (QG-RAM) described in Section 3.4.2. which generates attended visual features based on the question's guidance. (5) To predict an accurate answer, a multi-label classifier enabled by deep NN is trained. It is essential to combine the visual relational reasoning module with the visual attention module simultaneously, enabling the fusion of fine-grained features.

3.1 Visual features

To extract visual information from the relevant regions of the input image, we utilize the Faster R-CNN framework with ResNet-101 pre-trained on the Visual Genome dataset [28]. Faster R-CNN is used to obtain object detection boxes [65], and non-maximum suppression is employed for selecting the top K (normally K = 36) detection boxes. Mean-pooled convolutional features v_i from these selected region proposals i are used to represent the input image as $V = [v_1, v_2, \ldots\ldots, v_K]$, where $V \in \mathcal{R}^{K \times d_v}$. This approach focuses on only a few salient image regions from a large number of probable configurations, serving as a "hard" attention mechanism. Additionally, We capture scaled geometric features of the identified as $B = [b_1, b_2, \ldots\ldots, b_K]^T$, where $b_i = [\frac{x_i}{w}, \frac{y_i}{h}, \frac{w_i}{w}, \frac{h_i}{h}]$, where $(x_i, y_i), w_i$ and h_i represents coordinates, width, and height information. The visual relational reasoning module will be fed with these features.

3.2 Question features

There is a restriction of a maximum 14 words for every input question q to ensure efficient computation. This selection is backed up by the observation from [29] that just 0.25 percent of the questions in the VQA dataset exceed fourteen words. Questions with fewer than 14 words are padded with zero vectors, whereas questions beyond 14 words have the additional words eliminated. The question is then tokenized, and using a word embedding layer initialized with pre-trained GloVe word embeddings, every word is transformed into a 300-dimensional vector [30]. A GRU (Gated Recurrent Unit) sequentially processes the subsequent order of word embeddings, with the hidden state size set to d_q dimensions. The GRU′s final hidden state $Q \in \mathcal{R}^{d_q}$ is considered as the embedding of the input question q.

3.3 Feature-wise multimodal two-way co-attention module

In our work, feature-wise learning modules are introduced that attend both the image and related question. Our contribution comprises of introducing a two-way co-attention mechanism that offers variations in the execution of image and question feature-wise attention methods. This mechanism exhibits distinct methods for conducting feature-wise attention on images and their associated questions. Our two-way co-attention method, referred to as alternating co-attention, involves sequentially alternating for conducting feature-wise attention between the image and the corresponding question. This method enables the model to emphasize relevant visual and textual information iteratively, enhancing the interaction between the image and question representations, as computed in (Eqs. 2) and (3).

$$V\prime = IMulFA(V, Q), Q\prime = QMulFA(V\prime, Q) \tag{2}$$

or

$$Q\prime = QMulFA(V, Q), V\prime = QMulFA(V, Q\prime) \tag{3}$$

Another two-way co-attention method we propose is called parallel two-way co-attention. Unlike the alternating two-way co-attention methods, parallel two-way

co-attention generates image and question attention concurrently, as defined in as computed in (Eqs. 4) and (5).

$$V\prime = IMulFA(V, Q) \tag{4}$$

$$Q\prime = QMulFA(V, Q) \tag{5}$$

3.4 Relational reasoning module

We highlight the significance of global and local relational reasoning in this subsection. The global scheme involves utilizing information from the entire image to implicitly answer the question, while the local scheme focuses on modeling relationships among multiple objects to generate answers. Together these schemes play a crucial role in analyzing visual information from various perspectives, forming the fundamental structure of the proposed relational reasoning framework. Researchers in the VQA domain have extensively explored methods for performing relation reasoning among objects. The prevailing approach involves constructing neural network-based functions to describe relationships as follows (Eq. 6):

$$f_2(O) = h(\sum_{i,j} g(o_i, o_j) \tag{6}$$

where $O = \{o_1, o_2, \ldots, o_K\}$ is a feature set corresponding to K different objects, $g()andh()$ are the functions representing fully-connected layer ofNN. The fundamental structure of the proposed relational reasoning framework comprises two main schemes: global and local relational reasoning. Within this architecture, the output feature is defined as follows (Eq. 7):

$$V = f_g(\widehat{v}) + f_l(\widetilde{v}, q) \tag{7}$$

where q is the feature produced for the given question, $\widetilde{v}$ refers to the question-related feature set, and $\widehat{v}$ denotes the weighted region, $f_g()$ and $f_l()$ signifies the global and local relational reasoning, respectively. In global relational reasoning, the process starts by summing all the weighted regional features. Subsequently, a non-linear layer is employed to compute the feature representation. This computation can be represented as follows (Eq. 8):

$$f_g(\widehat{v}) = Relu(W_g * \left(\sum_{i=0}^{K} \widehat{v}_i\right) + b_g) \tag{8}$$

where W_g is the parameter matrix and b_g refers to the bias vector. The local relation reasoning scheme involves extracting question-guided regions through a regional attention module and defining the index of scale based on the number of question-dependent regions in a combination. To optimize memory usage, we perform nonlinear projection on the image region features and question representation, reducing them to a lower-dimensional subspace. Subsequently, we efficiently integrate the question embeddings into the image region embeddings (Eq. 9).

$$f_l(\widetilde{v}, q) = Relu(W_V \bullet V + b_V) + Relu(W_Q \bullet Q + b_Q) \tag{9}$$

where W_V and W_Q are the learning weights, b_V and b_Q indicate the biases.

3.4.1 Question-adaptive visual attention module

In VQA, to accurately answer a question, it is crucial to emphasize image regions that are pertinent to the related question. To address this, we introduce a Question-Adaptive Visual Attention Module (referred to as QA-VAM). The QA-VAM module incorporates multiple spatial attention heads, also known as glimpses, to filter out irrelevant information and emphasize the regions that are strongly related to the question. For each glimpse, we first combine the visual feature $V\prime \in \mathcal{R}^{K\times M}$ with the question feature $Q\prime \in \mathcal{R}^{1\times N}$ obtained from the bilinear model. These fused features are then passed through a softmax function to generate attention distributions over the regions of the image. This process helps to identify the regions that require focused attention when answering the given question. The specific formulation (Eqs. 10–13) and details of the attention distributions will be elaborated in the subsequent discussion.

$$h_i = BM\left(V_i\prime, Q^{'T}\right) \tag{10}$$

$$h = \left[h_0, h_1, \ldots\ldots, h_{K-1}\right] \tag{11}$$

$$p = softmax\left(W_h^v h\right) \tag{12}$$

$$v_j = \sum_{i=1}^{K} p_{j.i} V_i^{'}, j \in \{1, 2, \ldots\ldots, g\} \tag{13}$$

where V_i represents the *i-th* object feature,$h_i \in \mathbb{R}^C$ denotes the *i-th* fusion feature, [.] represents the stacking operation between vectors. $h \in \mathbb{R}^{C\times K}$ and $W_h^v \in \mathbb{R}^{g\times C}$ are a parameter matrix. g represents the number of glimpses. $v_j \in \mathbb{R}^M$ represents the *j-th* spatial attention visual feature.

3.4.2 Question-guided region attention module

Attention has emerged as a crucial element in VQA models [31–34], particularly in the context of top-down visual attention. In this work, top-down visual attention is employed to selectively emphasize the image regions that are utmost appropriate to the given question, effectively dropping the impact of inappropriate visual elements. A review of the low-rank bilinear pooling method, which forms the ground of the Question-Guided Region Attention Module (QG-RAM), is presented.

The most basic multimodal bilinear model combines the visual features of an image region, denoted as $v \in \mathcal{R}^{d_v}$, with the features of a question, denoted as $Q \in \mathcal{R}^{d_q}$. This model incorporates a bilinear interaction between the two feature sets. Mathematically, it can be represented as (Eq. 14):

$$z_i = v^T W_i Q \tag{14}$$

where $W_i \in \mathcal{R}^{d_v\times d_q}$. Bilinear models are known for their ability to capture pairwise interactions between feature dimensions effectively. However, they're plagued by two significant challenges: an excessive number of parameters, that result in high computational

costs, and the possibility of overfitting. Pirsiavash et al. [35] offered a low-rank bilinear model to address these issues, which minimizes the number of parameters by substituting the original parameter matrix W_i with two smaller matrices, $H_i G_i^T$, where $H_i \in \mathcal{R}^{d_v \times d}$ and $G_i \in \mathcal{R}^{d_q \times d}$.

$$z_i = v^T W_i Q = v^T H_i G_i^T Q = 1^T \left(H_i^T v \circ G_i^T Q\right) \tag{15}$$

where $1 \in \mathcal{R}^d$ signifies a vector of ones and $\circ$ indicates element-wise multiplication.

The following formulation can be used to generate an attention map equivalent to Eq. (16) and figure out the attended weight ω_i for image region i.

$$\omega_i = \frac{\exp(z_i)}{\sum_{k=1}^{K} \exp(z_k)} \tag{16}$$

To reduce parameters and promote parameter sharing across image regions, as mentioned in references [36, 37], the similar projection matrices $H_i \in \mathcal{R}^{d_v \times d}$ and $G_i \in \mathcal{R}^{d_q \times d}$ are used for all image regions. Therefore, in Eq. (17), the variable z_i can be defined as follows:

$$z_i = P^T \left(H_i^T v \circ G_i^T Q\right) \tag{17}$$

where $P \in \mathcal{R}^d$ represents a learnable vector. To obtain the attended feature representation $V_{att} \in \mathcal{R}^{d_v}$ for all regions in an image, we can calculate it as the weighted sum of the region visual features. The formulation for V_{att} can be expressed as follows:

$$V_{att} = A^T . V \tag{18}$$

where $A = [\omega_1, \omega_2, \ldots\ldots, \omega_K]^T$ represents the attention map.

3.5 Answer generation

The visual features are integrated with the question representation using either of the following formulations Eq. 19 or Eq. 20 after obtaining the relational visual feature representation V_{vr} and the attended feature representation V_{att}.

$$f = \left(W_r V_{vr} \circ W_a V_{att}\right) \circ W_q Q \tag{19}$$

or

$$f = \left(W_r V_{vr} \circ W_a V_{att}\right) + W_q Q \tag{20}$$

where $W_r \in \mathcal{R}^{d_f \times d_v}$, $W_a \in \mathcal{R}^{d_f \times d_v}$, and $W_r \in \mathcal{R}^{d_f \times d_q}$ represent learning weight matrices. The symbol $\circ$ denotes element-wise multiplication, and Q represents the question representation. The resulting fused vector f has a dimension of d_f. The bias terms are omitted in these equations for simplicity.

Next, to compute the probability of answer a_i given the image and question, a simple two-layer MLP (Multi-Layer Perceptron) with ReLU nonlinearity in its hidden layer is used (Eq. 21):

$$P(a_i/I, q) = \sigma(MLP(f))_i \tag{21}$$

Here, MLP represents the Multi-Layer Perceptron with ReLU activation, and σ denotes the sigmoid activation function.

As the final prediction, the answer with the highest probability among all the candidates is picked. For training, the prediction is often penalized using the binary cross-entropy loss function.

4 Experiment

4.1 Datasets and evaluation metric

Our primary evaluation for VQA models is conducted on two widely used datasets: VQA 1.0 [38] and VQA2.0 [39].

VQA 1.0 The VQA 1.0 has been developed employing the Microsoft COCO image dataset [40]. The set, typically referred to as the "test-standard" set, consists of 123,287 images that were used to generate a total of 248,349 training questions, 121,512 validation questions, and 244,302 test questions. A subset of the "test-standard" set called "test-dev," which comprises 25% of the test questions, is also provided. There are three different categories of questions in these datasets: yes/no questions, questions involving numbers, and other questions. Each question is related to ten free-response answers generated by unlike individuals.

VQA 2.0 An updated and enhanced version of the VQA 1.0 dataset is the VQA 2.0. It aims to overcome linguistic bias and evaluate increasingly complex recognition models for VQA. VQA 2.0 provides larger dimensions with more than 204,000 images extracted from the MS COCO dataset, more than 1 million questions, and more than 11 million answers. The dataset consists of 214,354 validation pairs, 447,793 testing pairs, and 443,757 training pairs (images, question, answer). Assuring consistency in the evaluation process, the evaluation metric employed in VQA 2.0 is the same as that used in VQA 1.0.

For the VQA 1.0 dataset and the VQA 2.0 dataset, our evaluation results are presented based on the challenging Open-Ended task. The review procedure becomes more complicated and diverse by considering that roughly 50% of the questions in both VQA 1.0 and VQA 2.0 fall into the "other" category.

Evaluation metric The performance of VQA models is evaluated using VQA accuracy as follows (Eq. 22):

$$Acc(ans) = \min(\frac{\#humans\ that\ said\ ans}{3}, 1) \tag{22}$$

The accuracy of a VQA model is considered 1 only if the predicted answer appears at least 3 times in the human-labeled answer list.

4.2 Implementation details

The PyTorch library is employed in the building process of our model. We apply the Adamax solver with $\beta 1 = 0.9$ and $\beta 2 = 0.9992$ for the VQA 1.0 and VQA 2.0 datasets. The learning rate is set for the initial three epochs at 0.001,0.002, and 0.0030. It remains constant until the tenth epoch; after that, it decreases every two epochs. We apply gradient clipping and use a batch size of 512. Dropouts are utilized after each fully connected layer to prevent overfitting. Question encoding involves embedding each word into a

300-dimensional vector, and the GRU's hidden state size is set to 1024. For the CLEVR dataset [41], we follow a method outlined in reference [42]. We train our VQA model end-to-end rather than relying upon previously learned ImageNet features. There is also batch normalization with a small CNN with 4 convolutional layers, *ReLU* activations, 128 kernels of size 3-*by*-3, and strides of 2. As a result, each image is denoted as an $8 \times 8 \times 128$ t ensor. Words are embedded into 64-dimensional vectors and fed into a single-layer *GRU* with a hidden state size of 128. Other settings align with the reference [42].

4.3 Ablation study

We employ an array of modules that have various hyperparameters in our comprehensive VQA model. To evaluate how each module impacts the accuracy of the overall prediction, we perform ablation. Table 2 demonstrates the results of the ablation test corresponding to each module employed by our full model and corresponding model size. We apply the VQA2.0 dataset to train multiple versions of our VQA model, and then we evaluate how well they perform on the val split. Several versions of our VQA model are as follows:

(1) **Baseline model:** In the baseline model, element-wise multiplication is employed to combine the visual representation and question representation, which is then subjected to non-linear projection through a fully connected layer.
(2) **Baseline model + multimodal feature-wise attention:** By employing relations between visual and textual features (Row 2 and Row 3), multimodal feature-wise attention builds attention weights to highlight significant features and overpower less useful ones. The performance of the model is improved by 1.65% over the baseline model.
(3) **Baseline model + Relational Reasoning:** By employing visual relational reasoning component, the model captures visual interactions among distinct regions within an image. We can observe that it boosts the performance of our VQA model by 4.72% over the baseline model.
(4) **Baseline model + feature-wise multimodal two-way co-attention:** We utilize a two-way co-attention mechanism, which implements image and question feature-wise attention that varies to integrate feature-wise attention learning modules. The first two co-attention modules, also known as alternating co-attention involve alternatively for conducting feature-wise attention between the image and the corresponding question. The gain in the performance is 3.70% over the Baseline model, which indicates that our feature-wise multimodal co-attention module can concurrently capture significant features of image and question encodings, and overlook less relevant features.

Table 2 Conducting Ablation Tests on Individual Modules and Model Size on VQA 2.0

Model	Accuracy	Model size
Baseline	66.64	15.6 M
Question multimodal feature-wise attention	67.13	16.2 M
Image multimodal feature-wise attention	68.29	21.6 M
Relational Reasoning	71.36	30.7 M
Multimodal feature-wise co-attention	72.19	31.6 M
Full model	**75.89**	33.8 M

Bold values specifies the highest accuracy

(5) **Full Model:** It uses a feature-wise multimodal two-way co-attention component together with the relational reasoning module. The overall gain in the performance of our proposed method over the Baseline model is 9.25%. The observed enhancement in performance can be attributed to the cumulative effect of the three modules, underscoring their mutual compatibility.

4.4 Quantitative results

In Tables 3 and 4, we have compared the proposed method with the existing methods.

In Table 3, we demonstrate the superiority of our model over existing methods by comparing the performance of our model on the VQA 1.0 dataset that surpasses the best-published results. We achieve notable improvements on all question types, Num by 0.62%, Y/N by 0.31% and other by 1.04% for the test-dev set and Num by 0.83%, Y/N by 1.12% and other by 1.06% for the test-standard set compared to the Re-attention [11] model. Moreover, when compared to the method MRA-Net [43] on the VQA 1.0 dataset, our model showcases improved values on all question types, Num by 3.33%, Y/N by 1.49%, other by

Table 3 Comparison of performance of our method with existing methods on VQA 1.0

Method	Test-dev				test-std			
	Num	Y/N	Other	All	Num	Y/N	Other	All
VQA team [38]	36.77	80.5	43.08	57.75	36.53	80.57	43.73	58.16
SAN [13]	37.32	80.87	43.12	58.7	37.53	80.8	43.48	58.24
FDA [14]	36.16	79.3	45.77	59.24	-	-	-	58.9
HieCoAtt [15]	38.7	79.7	51.7	61	-	-	-	62.1
DAN [45]	39.1	83	53.9	64.3	38.1	82.8	54	64.2
SAAA [46]	39.1	82.2	55.2	64.5	39.1	82	55.2	64.6
MLAN [47]	40.2	83.8	53.7	64.6	40.9	83.7	53.7	64.8
VQA-Machine [48]	38.4	81.5	53	63.1	38.2	81.4	53.2	63.3
MFH [36]	39.7	85	57.4	66.8	39.5	85	57.4	66.9
DCN [49]	41.66	84.48	57.44	66.83	41.27	84.61	56.83	66.66
MSMLAN [50]	41.2	83.8	56.7	66.1	41.3	83.7	56.6	66.2
QLOB [17]	39.51	82.26	52.17	63.13	37.95	82.24	52.25	63.12
VRR [20]	41.43	84.36	58.71	67.37	41.33	84.18	58.58	67.33
ALSA [51]	42.94	87.12	59.06	69.52	43.84	86.94	58.21	69.32
IASSM [52]	42.86	87.04	58.94	69.35	43.68	86.85	58.17	69.05
MRA-Net [43]	**43.89**	**86.79**	**59.62**	**69.06**	**44.16**	**86.37**	**60**	**69.22**
CAM [53]	42.78	86.77	60.27	69.25	42.36	86.55	60.38	69.29
AT-CMG [54]	43.05	86.72	60.7	69.47	42.63	86.87	60.77	69.64
Re-attention [11]	**46.6**	**87.9**	**61.3**	**70.5**	**46.8**	**87.2**	**61.5**	**70.8**
GNN-CAA [55]	42.2	84.12	58.81	68.21	41.43	84.16	59.12	68.35
ICIVQA [56]	-	-	-	-	41.63	84.76	60.12	70.35
UFSCAN [7]	-	-	-	70.19	-	-	-	70.24
Ours	47.22	88.21	62.34	71.23	47.63	88.32	62.56	71.94

Bold values represent the models with which we have compared the proposed model

Table 4 Comparison of performance of our method with existing methods on VQA 2.0

Method	Test-dev				Test-std			
	Num	Y/N	Other	All	Num	Y/N	Other	All
MCB [39]	-	-	-	-	38.28	78.82	53.36	62.27
DCN [49]	46.6	83.5	56.72	66.6	46.93	83.89	56.9	67
Tips & Tricks [29]	44.21	81.82	56.05	65.32	43.9	82.2	56.26	65.67
MUREL [19]	49.84	84.77	57.85	68.03				68.41
MSMLAN [50]	45.5	82.6	56.8	66.3	-	-	-	-
ODA-GCN [16]	47.02	83.73	56.57	66.67	83.77	47.28	56.96	66.87
QLOB [17]	-	-	-	-	39.25	76.46	49.62	59.5
VRR [20]	45.51	83.31	58.41	67.2	44.96	83.39	58.49	67.34
ALSA [51]	48.98	85.73	59.17	69.21	-	-	-	-
MRA-Net [43]	48.92	85.58	59.46	69.02	49.22	85.83	59.86	69.46
CAM [53]	47.35	85.18	59.76	68.82	46.98	85.22	59.91	68.99
AT-CMG [54]	47.91	85.58	59.92	69.13	47.72	85.85	60.35	69.55
Re-attention [11]	54.39	87.51	61.51	71.6	53.38	88.95	61.54	71.94
JE-MHA [57]	50.8	86.3	59.9	69.7	-	-	-	70.79
BGNs [58]	54.09	88.6	62.46	72.28	-	-	-	72.41
GNN-CAA [55]	46.12	84.12	58.13	67.96	45.21	83.34	58.87	67.98
ICIVQA [56]	-	-	-	-	48.23	86.63	59.98	70.67
IAHOT [21]	**56.33**	**88.23**	**67.74**	**75.4**	**-**	**-**	**-**	**73.34**
TRAR [44]	**55.33**	**88.11**	**63.31**	**72.62**	**-**	**-**	**-**	**72.93**
APN [59]	52.68	87.44	61.18	71.14	-	-	-	71.83
PHOC-FV [60]	54.43	86.98	61.34	72.28	-	-	-	72.89
Visual + STF [61]	-	-	-	74.97	-	-	-	-
Ours	57.21	89.23	67.83	75.89	49.23	87.21	62.05	76.32

Bold values represent the models with which we have compared the proposed model

2.72% for the test-dev set and Num by 3.47%, Y/N by 1.95% and other by 2.56%, for the test-standard set. Our model achieves a state-of-the-art performance on VQA 1.0, with an overall best accuracy of 71.23% on the test-dev set and 71.94% on the test-standard set.

In Table 4, our co-attention-based model demonstrates significant improvements over the state-of-the-art methods IAHOT [21] and TRAR [44] on the VQA 2.0 dataset. Our model gains improvement on all question types, Num by 0.88%, Y/N by 1.0%, and other by 0.09%, when compared to the IAHOT method and Num by 1.98%, Y/N by 1.12% and other by 4.52%, when compared to TRAR method, on the test-dev set. Our model achieves state-of-the-art performance on VQA 2.0, with an overall best accuracy of 75.89% on the test-dev set and 76.32% on the test-standard set.

Overall, our results highlight the effectiveness of our two-way co-attention-based model, outperforming the state-of-the-art methods, as indicated by the boldface values in both Tables 3 and 4.

4.5 Qualitative results

Figures 3 and 4 illustrate the qualitative results of our model on the VQA 1.0 and VQA 2.0 datasets respectively, where the examples are randomly picked from the dataset. In Fig. 3,

Question: What is the color of grapes?
Re-attention [11]: orange
MRA-Net [43]: green
Ours: green

Question: How many women are there?
Re-attention [11]: two
MRA-Net [43]: two
Ours: one

Question: Which is the green vegetable in the plate?
Re-attention [11]: spinach
MRA-Net [43]: broccolini
Ours: Broccoli

Fig. 3 Qualitative results obtained by our model on the VQA 1.0 dataset. Re-attention [11] and MRA-Net [43] are comparable state-of-the-art methods

Question: What is near to the clock?
IAHOT [21]: cat
TRAR [44]: cat
Ours: rabbit

Question: What are animals shown in the image?
IAHOT [21]: deer
TRAR [44]: giraffe
Ours: giraffes

Question: Who runs ahead between both persons?
IAHOT [21]: Man
TRAR [44]: Man
Ours: Woman

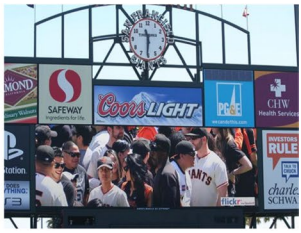
Fig. 4 Qualitative results obtained by our model on the VQA 2.0 dataset. IAHOT [21] and TRAR [44] are comparable state-of-the-art methods

it is shown that for the VQA 1.0 dataset, our model gives more accurate predictions as compared to MRA-Net [43] and Re-attention [11] on the same test images. Similarly, in Fig. 4, improved predictions for VQA 2.0 are shown when comparing our model with published methods IAHOT [21] and TRAR [44]. The effectiveness of our proposed model is evident in its ability to fuse relational reasoning with visual features in the image. By incorporating the two-way co-attention mechanism with Faster RCNN, our model successfully detects objects and their relationships in the image, enabling a more comprehensive image description. It efficiently focuses on relevant objects and their corresponding features while understanding the relations between them, leading to accurate answer generation through the integration of relational and visual reasoning. The inclusion of visual spatial attention further enhances the model's performance in the VQA task by concentrating on objects relevant to the given question. Through the fusion of detected visual features and their relationships, our model achieves higher confidence in predicting the correct answers.

In Fig. 3 (leftmost), our co-attention mechanism effectively attends to the object's "grapes" and "orange" along with the attribute "color" using relational reasoning. This allows our model to correctly predict the answer "green" to the question "What is the color

Question: What are the bears standing on?
A: sand
GT: ice

Question: What a person is holding?
A. stick
GT: ski pole



Question: The logo of which company is at the rightmost side?
A. purple
GT: CHW health services

Fig. 5 Example of failure cases. Sometimes our model fails to use common sense reasoning like in the left and middle images. Also, the scene text is inaccessible by the proposed model, as is observed in the rightmost image. The two letters GT and A represent the answer for the ground truth and the predicted answer, respectively

of grapes?" Similarly, in Fig. 4 (rightmost), the attended objects are "man" and "woman". Our model demonstrates its capability to understand the relative positions of both persons, leading to an accurate prediction of "woman" in response to the question "Who runs ahead?" These qualitative results highlight the efficacy of our co-attention mechanism, which seamlessly integrates visual features and relational reasoning, enabling accurate answers prediction.

4.6 Failure cases

The negative examples reveal several fail cases of our method. Specifically, our model struggles to answer the questions where commonsense knowledge is required. For example, in Fig. 5 (left), the correct answer to the question "What are the bears standing on?" should be "ice," but our model erroneously predicts "sand," indicating a lack of contextual understanding. Similarly, in Fig. 5 (middle), the question "What is a person holding?" should be answered as "ski pole," but our model predicts "stick," showing the challenge of handling complex reasoning without access to external knowledge bases. Furthermore, in Fig. 5 (right), our model faces difficulty in comprehending scene text as it is not designed for text reading. Consequently, it predicts an irrelevant answer, "purple," instead of correctly identifying "CHW health services" as the logo of the company on the rightmost side. These limitations emphasize the need for further improvement in our method, particularly in integrating external knowledge and addressing complex reasoning tasks effectively.

5 Conclusion

The present research presents a feature-wise attention approach that enhances the extraction of distinguishing features in image and question representations by paying attention to valuable features while suppressing insignificant ones. Novel modules have been developed to simulate question-guided image feature-wise attention and image-guided question feature-wise attention. These modules are fused with visual spatial attention to integrate the feature-wise and spatial co-attention network. The dense two-way co-attention layer,

which integrates dense symmetric interactions between the input image and question in an attempt to enhance the fusion of visual and linguistic representations, is the primary element of the network. Our findings from experiments support the claim that our method exhibits state-of-the-art performance on two significant real-world datasets.

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. Zhang C, Lu Y (2021) Study on artificial intelligence: the state of the art and future prospects. J Ind Inf Integr 23:100224
2. Sharma H, Srivastava S (2021) Visual question-answering model based on the fusion of multimodal features by a two-way co-attention mechanism. Imaging Sci J 69(1–4):177–189
3. Bhatt D, Patel C, Talsania H, Patel J, Vaghela R, Pandya S, ..., Ghayvat H (2021) CNN variants for computer vision: history, architecture, application, challenges, and future scope. Electronics 10(20):2470
4. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28, Montreal Convention Center, Montreal, Canada, December 7–10
5. Schwartz I, Schwing A, Hazan T (2017) High-order attention models for visual question answering. Advances in Neural Information Processing Systems, 30. Long Beach, California, USA, December 4–9, 3667–3677
6. Wu Y, Ma Y, Wan S (2021) Multi-scale relation reasoning for multi-modal visual question answering. Signal Process: Image Commun 96:116319
7. Zhang S, Chen M, Chen J, Zou F, Li YF, Lu P (2021) Multimodal feature-wise co-attention method for visual question answering. Information Fusion 73:1–10
8. Dey R, Salem FM (2017) Gate-variants of gated recurrent unit (GRU) neural networks. In: 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS). IEEE, pp 1597–1600
9. Zhan H, Xiong P, Wang X, Xin WANG, Yang L (2022) Visual question answering by pattern matching and reasoning. Neurocomputing 467:323–336
10. Zheng W, Yin L, Chen X, Ma Z, Liu S, Yang B (2021) Knowledge base graph embedding module design for visual question answering model. Pattern Recogn 120:108153
11. Guo W, Zhang Y, Yang J, Yuan X (2021) Re-attention for visual question answering. IEEE Trans Image Process 30:6730–6743
12. Zheng X, Wang B, Du X, Lu X (2021) Mutual attention inception network for remote sensing visual question answering. IEEE Trans Geosci Remote Sens 60:1–14
13. 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, Las Vegas, Nevada, June 26-July 1, 21–29
14. Ilievski I, Yan S, Feng J (2016) A focused dynamic attention model for visual question answering. arXiv preprint arXiv:1604.01485
15. Lu J, Yang J, Batra D, Parikh D (2016) Hierarchical question-image co-attention for visual question answering. Advances in neural information processing systems, 29, Barcelona, Spain, December 5–10, 289–297
16. Zhu X, Mao Z, Chen Z, Li Y, Wang Z, Wang B (2021) Object-difference drived graph convolutional networks for visual question answering. Multimed Tools Appl 80(11):16247–16265
17. Gao L, Cao L, Xu X, Shao J, Song J (2020) Question-led object attention for visual question answering. Neurocomputing 391:227–233
18. Singh A, Natarajan V, Shah M, Jiang Y, Chen X, Batra D, ..., Rohrbach M (2019) Towards vqa models that can read. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, June 16–20, 8317–8326

19. 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, Long Beach, CA, June 16–20, 1989–1998
20. Zhang W, Yu J, Hu H, Hu H, Qin Z (2020) Multimodal feature fusion by relational reasoning and attention for visual question answering. Information Fusion 55:116–126
21. Sharma H, Jalal AS (2022) An improved attention and hybrid optimization technique for visual question answering. Neural Process Lett 54(1):709–730
22. Zhang X, Wu C, Zhao Z, Lin W, Zhang Y, Wang Y, Xie W (2023) PMC-VQA: visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415
23. 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
24. Chen L, Yan X, Xiao J, Zhang H, Pu S, Zhuang Y (2020) Counterfactual samples synthesizing for robust visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 10800–10809
25. Sharma H, Jalal AS (2022) Image captioning improved visual question answering. Multimedia tools and applications 81(24):34775–34796
26. 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
27. 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
28. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, ..., Fei-Fei L (2017) Visual genome: connecting language and vision using crowdsourced dense image annotations. Int J Comput Vis 123(1):32–73
29. Teney D, Anderson P, He X, Van Den Hengel A (2018) Tips and tricks for visual question answering: learnings from the 2017 challenge. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, Utah, June 18–22, 4223–4232
30. 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), association for computational linguistics, Doha, Qatar. 1532–1543
31. Yang Z, He X, Gao J, Deng L, Smola A (2016) Stacked attention networks for image question answering. In: CVPR, pp 21–29
32. Lu J, Yang J, Batra D, Parikh D (2016) Hierarchical question-image co-attention for visual question answering. In: NIPS. pp 289–297
33. Kazemi V, Elqursh A (2017) Show, ask, attend, and answer: a strong baseline for visual question answering, arXiv:1704.03162v2
34. Nguyen D, Okatani T (2018) Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In: CVPR. pp 6087–6096
35. Ramanan D, Pirsiavash H, Fowlkes C (2009) Bilinear classifiers for visual recognition. In: NIPS, pp 1482–1490
36. 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
37. O.K. Kim J., W. Lim, Hadamard product for low-rank bilinear pooling, ICLR, 2017.
38. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D (2015) VQA:visual question answering. In: Proc. IEEE Int. Conf. Computer Vision (ICCV). pp 2425–2433, https://doi.org/10.1109/ICCV.2015.279
39. 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: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). pp 6325–6334. https://doi.org/10.1109/CVPR.2017.670
40. Lin T, Maire M, Belongie SJ, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft COCO: common objects in context. In: Fleet DJ, Pajdla T, Schiele B, Tuytelaars T (eds) Computer vision - ECCV 2014 - 13th European conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V, in: Lecture Notes in Computer Science. Springer, vol. 8693, pp 740–755. https://doi.org/10.1007/978-3-319-10602-1_48
41. Hariharan B, Johnson J, Maaten L, Li F-F (2017) Clevr: a diagnostic dataset for compositional language and elementary visual reasoning. In: CVPR, pp 1988–1997
42. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980
43. Peng L, Yang Y, Wang Z, Huang Z, Shen HT (2020) Mra-net: improving vqa via multi-modal relation attention network. IEEE Trans Pattern Anal Mach Intell 44(1):318–329

44. Zhou Y, Ren T, Zhu C, Sun X, Liu J, Ding X, ..., 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, June 19–25, 2074–2084
45. Nam H, Ha JW, Kim J (2017) Dual attention networks for multimodal reasoning and matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, Hawaii, July 21–26, 299–307
46. Kazemi V, Elqursh A (2017) Show, ask, attend, and answer: a strong baseline for visual question answering. arXiv preprint arXiv:1704.03162
47. Yu D, Fu J, Mei T, Rui Y (2017) Multi-level attention networks for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, Hawaii, July 21–26, 4709–4717
48. Wang P, Wu Q, Shen C, van den Hengel A (2017) The vqa-machine: learning how to use existing vision algorithms to answer new questions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, Hawaii, July 21–26, 1173–1182
49. Nguyen DK, Okatani T (2018) Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, Utah, June 18–22, 6087–6096
50. Yu D, Fu J, Tian X, Mei T (2019) Multi-source multi-level attention networks for visual question answering. ACM Trans Multimed Comput Commun Appl (TOMM) 15(2s):1–20
51. Liu Y, Zhang X, Zhao Z, Zhang B, Cheng L, Li Z (2020) ALSA: adversarial learning of supervised attentions for visual question answering. IEEE Trans Cybern 52(6):4520–4533
52. Liu Y, Zhang X, Huang F, Zhou Z, Zhao Z, Li Z (2020) Visual question answering via combining inferential attention and semantic space mapping. Knowl-Based Syst 207:106339
53. Peng L, Yang Y, Zhang X, Ji Y, Lu H, Shen HT (2020) Answer again: improving VQA with cascaded-answering model. IEEE Trans Knowl Data Eng 34(4):1644–1655
54. Li W, Sun J, Liu G, Zhao L, Fang X (2020) Visual question answering with attention transfer and a cross-modal gating mechanism. Pattern Recogn Lett 133:334–340
55. Sharma H, Jalal AS (2021) Visual question answering model based on graph neural network and contextual attention. Image Vis Comput 110:104165
56. Sharma H, Jalal AS (2022) Image captioning improved visual question answering. Multimed Tools Appl 81(24):34775–34796
57. Kim JJ, Lee DG, Wu J, Jung HG, Lee SW (2021) Visual question answering based on local-scene-aware referring expression generation. Neural Netw 139:158–167
58. Guo D, Xu C, Tao D (2021) Bilinear graph networks for visual question answering. IEEE Trans Neural Netw Learn Syst
59. Yang X, Gao C, Zhang H, Cai J (2021) Auto-parsing network for image captioning and visual question answering. In: Proceedings of the IEEE/CVF international conference on computer vision, June 19–25 (pp. 2197–2207)
60. Sharma H, Jalal AS (2022) A framework for visual question answering with the integration of scene-text using PHOCs and fisher vectors. Expert Syst Appl 190:116159
61. Sharma H, Jalal AS (2022) Improving visual question answering by combining scene-text information. Multimed Tools Appl 81(9):12177–12208
62. Barra S, Bisogni C, De Marsico M, Ricciardi S (2021) Visual question answering: which investigated applications? Pattern Recogn Lett 151:325–331
63. Gao H, Xu K, Cao M, Xiao J, Xu Q, Yin Y (2021) The deep features and attention mechanism-based method to dish healthcare under social iot systems: an empirical study with a hand-deep local–global net. IEEE Trans Comput Soc Syst 9(1):336–347
64. Gao H, Xiao J, Yin Y, Liu T, Shi J (2022) A mutually supervised graph attention network for few-shot segmentation: the perspective of fully utilizing limited samples. IEEE Trans Neural Netw Learn Syst
65. Xiao J, Xu H, Gao H, Bian M, Li Y (2021) A weakly supervised semantic segmentation network by aggregating seed cues: the multi-object proposal generation perspective. ACM Trans Multimed Comput Commun Appl 17(1s):1–19