Applied Intelligence (2024) 54:5167–5182
https://doi.org/10.1007/s10489-024-05456-4

IMCN: Improved modular co-attention networks for visual question answering

Cheng Liu[1,2] · Chao Wang[1,2] · Yan Peng[1,2,3]

Accepted: 7 April 2024 / Published online: 16 April 2024

Abstract

Many existing Visual Question Answering (VQA) methods use traditional attention mechanisms to focus on each region of the input image and each word of the input question and achieve well performance. However, the most obvious limitation of traditional attention mechanisms is that the module always generates a weighted average based on a specific query. When all regions and words are unsatisfied with the query, the generated vectors, which are noisy information, may lead to incorrect predictions. In this paper, we propose an Improved Modular Co-attention Network (IMCN) by incorporating the Attention on Attention (AoA) module into the self-attention module and the co-attention module to solve this problem. AoA adds another attention process by using element-wise multiplication on the information vector and the attention gate, which are both generated from the attention result and the current context. With AoA, the attended information obtained by the model is more useful. We also introduce an Improved Multimodal Fusion Network (IMFN), which leverages various branches to achieve hierarchical fusion, to fuse visual features and textual features for further improvements. We conduct extensive experiments on the VQA-v2 dataset to verify the effectiveness of the proposed modules and experimental results demonstrate our model outperforms the existing methods.

Keywords Co-attention · Multimodal · Self-attention · Visual question answering

1 Introduction

As a type of multimodal learning, VQA [3] which needs both computer vision (CV) and natural language processing (NLP) [4] technologies, has attracted many researchers' interest. The VQA task is to give the correct linguistic answer based on the given image and the natural language question about the image [5]. It is one of the most challenging multimodal tasks since it requires a fine-grained semantic understanding of the input image and the corresponding question. As the attention mechanism continues to evolve, VQA has also made great progress. The attention mechanism, an integral part of most VQA models, has emerged since the continuous development of deep learning architecture. It is designed to enable the model to focus on the specific regions in the image and question and was first utilized in visual question answering by [6]. Nowadays, attention mechanisms can be seen in almost all the VQA architectures. And related research has proven that learning co-attention for the visual and textual modalities simultaneously can facilitate fine-grained representations of images and questions, thus leading to more accurate predictions [7, 8].

However, the deficiency of these co-attention models is that they learn only coarse interactions of multimodal instances. As for the correlation between every image region and every question word, the models don't concern. To solve this problem, bilinear attention networks (BAN) [9] and dense co-attention networks (DCN) [10] which consist of a stack of dense co-attention layers that can be cascaded in depth were proposed to model dense interactions between each image region and each question word. However, these models simultaneously lack modeling self-attention within

✉ Chao Wang
cwang@shu.edu.cn

Cheng Liu
563295220@shu.edu.cn

Yan Peng
pengyan@shu.edu.cn

[1] School of Future Technology, Shanghai University, Shanghai, China

[2] Institute of Artificial Intelligence, Shanghai University, Shanghai, China

[3] Shanghai Artificial Intelligence Laboratory, Shanghai, China

each modality, leading to poor performance along with an increase in model depth. Yu et al. [11] proposed modular co-attention networks (MCAN) to overcome it. MCAN is an improvement on the previous model by adding self-attention within each co-attention layer and gets better performance on VQA tasks. Specifically, the main part of the MCAN is made up of multiple Modular Co-Attention (MCA) layers. And the MCA layer is formed by combining two types of attention units: self-attention (SA) and guided attention (GA), which are used to capture intra-modal interactions (*e.g.*, region-to-region) and cross-modal interactions (*e.g.*, region-to-word) respectively.

Advanced attention mechanisms benefit VQA. Unfortunately, the traditional attention mechanism applied in the MCAN still has some limitations. One prominent of them is that the output is always a weighted combination of value pairs that the model is attending to [1]. It may be problematic when there is no closely related context for the model to attend to (*e.g.*, a word with no associated contextual word or image region). In this case, the attention mechanism may result in noisy or even distracting output vectors that can negatively impact the performance. Inspired by [12], in this paper, we propose an IMCN that utilizes the AoA module to refine the original model to solve this problem. AoA is an extension of the traditional attention mechanism, and its essence is to add a designed attention process to the traditional attention. The structure of AoA is shown in Fig. 1. We have labeled the second attention process with a blue dashed box. Initially, AoA generates an information vector (I) and an attention gate (G) through two linear transformations. Specifically, I is obtained from the present context (*i.e.*, the query) and the attention results through a linear transformation. It stores the information of the present context together with the newly acquired information that comes from the attention result. And G is obtained from the query and the attention result through another linear transformation followed by sigmoid activation. The value of each channel of G represents the relevance of the information on the corresponding channel in the information vector. Through element-wise multiplication to I and G, we ultimately acquire the attended information, which establishes a connection between multiple attention heads and preserves only the most pertinent while discarding all unrelated attention results. We use AoA to refine the SA and GA modules in MCAN and further construct three Improved Modular Co-attention (IMC) variants which are improvements on the three variants of MCA in MCAN. Finally, IMCN is acquired by cascading several IMC layers.

Moreover, the key point of method used in MCAN is leveraging attention networks to concentrate on key objects in images and keywords in questions. However, the distribution of attention in these prior attempts tends to localize similar regions, resulting in a lack of ability to derive important entities. From the perspective of the multimodal fusion module, the performance of MCAN is still limited by the fact that a significant amount of information may be lost during the multimodal interaction and fusion of two modalities. To solve this, we introduce a novel multimodal fusion module IMFN, which is capable of performing hierarchical multimodal fusion through multiple branches to capture fine-grained and intricate relationships across multiple levels: region, word, and their interaction. In addition, it is able to capture the distinct distribution of attention to the many different visual and textual components that are crucial for inferring answers.

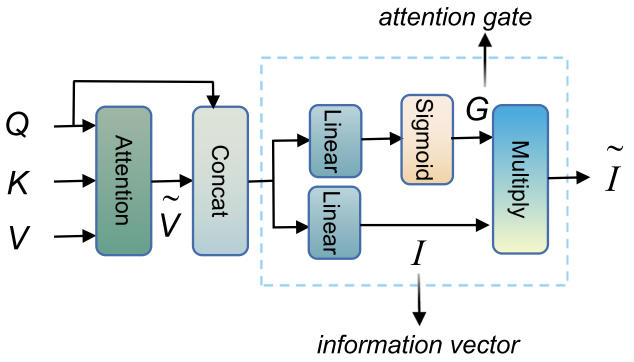


Fig. 1 Illustration of Attention on Attention. Q, K, V denote the query, key, and value respectively. $\widetilde{V}$ represents the first attention result. In addition, I, G denote the information vector and attention gate. And $\widetilde{I}$ represents the second attention result, which is the attended information

To summarize, the main contributions of this paper are:

- We propose the IMCN, which is capable of filtering the attention results of interactions between image regions and question words that are irrelevant to the prediction of the correct answer, and retaining only the useful attention results.
- We introduce an IMFN with different branches, which can hierarchically fuse visual and textual features through multiple stages and further enhance the model performance.
- We conduct extensive experiments on the VQA-v2 dataset. The experimental results outperform the baseline and prove the effectiveness of the proposed modules.

The rest of this paper is organized as follows: Section 2 introduces the related work of VQA in three parts. Section 3 introduces the overall framework of our model and the technical details used in them. Section 4 describes the dataset used, the specific experimental settings and the experimental results. Section 5 analyses the limitation of our method, and in Section 6, we conclude our work.

2 Related work

In this section, we will briefly introduce some research work related to this paper in two parts, visual question answering (Section 2.1) and attention mechanisms (Section 2.2).

2.1 Visual question answering

As a popular research direction, VQA has attracted more and more attention in recent years. It was first proposed by [5]. The approach used by most VQA models can be summarized as follows: firstly, extracting visual features and textual features from the given image and question; secondly, fusing the extracted features to predict the answer. To obtain visual and textual features, most VQA models leverage an attention mechanism to select useful information and reduce unrelated information. For example, to localize relevant image objects, Lee et al. [13] and Teney et al. [14] model visual attention mechanisms using different networks based on convolutional neural networks. Moreover, Nam et al. [15] and Fan et al. [16] proposed methods that both consider image-guided question attention and question-guided visual attention. Gurunlu et al. [17] proposed a novelty block-based image detection method. And Le et al. [18] applied explicit linguistic-visual grounding to guide the cross-attention.

In addition, in recent years, a number of challenging datasets that require reasoning were designed, such as OK-VQA [19], A-OKVQA [20], and WebQA [21]. Specifically, given an image-question pair, the model is not only required to localize the object in the image, but also needs to reason with external knowledge to give an accurate answer. And there has been a lot of work on commonsense VQA. Ravi et al. [22] proposed the Vision-Language-Commonsense BERT (VLC-BERT) which uses the commonsense transformer to incorporate contextualized knowledge. Garcia et al. [23] analysed the impact of efficient knowledge injection applied via E-BERT on the performance of vision-language models on relatively unexplored knowledge-based VQA tasks. Ding et al. [24] proposed the Multimodal Knowledge Extraction and Accumulation framework (MuKEA), which uses explicit triples that associate visual objects and factual answers with implicit relations to represent multimodal knowledge. Gao et al. [25] proposed a new paradigm for VQA task, which converts images into plain text, thus realizing knowledge passage retrieval and generative question-answering.

Apart from the above research, there has also been a boom in research on solving VQA tasks using large language models recently. For example, Yang et al. [26] proposed to use images to generate captions and introduce in-context examples to prompt GPT-3 to generate answers. Tiong et al. [27] proposed Plug-and-Play VQA (PNP-VQA) that leverages off-the-shelf pretrained models for VQA without additional training. And Guo et al. [28] further introduced question-answer pairs to construct a new prompt for large language models.

These methods have achieved superior performance on commonsense-based VQA datasets. Nevertheless, the performance of these methods in regular VQA datasets such as VQA-v2 is general. In addition, the computation overhead becomes large due to the large number of model parameters.

2.2 Attention mechanisms

The attention mechanism was proposed by [29]. It was designed to imitate the way humans see and has been used in many works on visual question answering. Early studies primarily adopted question-guided attention on image regions. Yang et al. [30] proposed a stacked attention network SAN that uses multiple layers of attention to query the image several times to predict the final answer progressively. Kim et al. [31] improved this by introducing residual learning to get better attention to information. Chen et al. [32] proposed a structured visual attention mechanism to capture the semantic structure of an image according to the question.

After that, image-guided attention to questions began to be adopted to achieve better performance. Lu et al. [7] proposed a refined model that uses a co-attention mechanism. With co-attention model can capture features from different regions of an image and different segments of a question. Yu et al. [33] proposed a multi-level attention network that can decrease the semantic gap with semantic attention and help fine-grained spatial inference with visual attention. To solve the problem of limited interaction between image regions and question words, Nguyen et al. [10] proposed DCN and Yu et al. [11] proposed MCAN which both consist of a stack of co-attention layers that can perform multiple image-question interactions.

With the use and success of pre-training in NLP, it has become a new trend to use pre-training to improve the performance of VQA models [34]. Moreover, some researchers think it vital to leverage visual and textual features in the VQA task. Therefore, these models focus on how to improve the existing attention mechanisms. For example, Zhou et al. [35] proposed TRAR which can select the corresponding attention dynamically according to the result of the previous inference step. Guo et al. [36] proposed SCAVQAN which can filter out useful features by setting thresholds for attention scores. Shen et al. [3] proposed LSAT which uses local windows of visual features to model intra-window and inter-window attention.

In addition, attention mechanisms were also applied in tasks other than VQA. For example, Cheng et al. [37] proposed a masked-attention mask transformer (Mask2Former) which uses masked attention to extract localized features in an image segmentation task. Liu et al. [38] proposed to use partial class activation attention to remove the intra-class inconsistency in the semantic segmentation task. Liang et al. [39] proposed a recurrent video restoration transformer (RVRT) to achieve long-range dependency modelling ability in video restoration. Song et al. [40] proposed a global-local attention module (GLAM) which consists of four kinds of attention to obtain local and global contextual information in image retrieval. To make the self-attention module concentrate on related regions, Xia et al. [41] chose to select the key and value pairs of self-attention in a data-dependent way. And Zhang et al. [42] proposed patch attention (PAT) which can capture the global shape context to solve the problem of high computational costs in point cloud learning.

We point out that the traditional attention mechanism used in the above work suffers from the flaw that we have mentioned. Therefore, in this paper, we experiment by replacing the traditional attention mechanism with a more advanced one (*i.e.*, AoA). Our results are improved compared to the baseline, thus validating the superiority of AoA over the traditional attention mechanism.

3 Method

In this section, we describe the IMCN and IMFN in detail. The overall framework of our method is shown in Fig. 2, which is stacked by multiple Improved Modular Co-attention (IMC) layers. In the following part, we introduce the method of extraction of image features and question features first and then explain the components of IMCN. Finally, after obtaining attended visual and textual features, the IMFN is introduced to fuse them and then we obtain the prediction answer eventually.

3.1 Notations

Before describing our method, we give the key mathematical notations and their descriptions of this paper, which are listed in Table 1.

3.2 Extraction of image features and question features

IMCN uses a group of regional visual features, which are extracted from Fast-RCNN that is pretrained on the Visual Genome dataset, to represent the input image. In order to ensure the quality of visual features, we set a confidence threshold and only select image regions with detection probability exceeding it. Thus, we obtain a dynamic number of regional image features which can be represented by a matrix $X \in \mathbb{R}^{m \times d_x}$, where $m \in [10, \ 100]$ is the number of selected image regions. And $x_i \in \mathbb{R}^{d_x}$ represents the i-th image region.

For the question feature representation, following previous works [9, 14], we first tokenize the input question into a set of words and limit the maximum number of words to 14. The extra words of questions whose number exceeds 14 are discarded to keep consistency with other questions. After that, the 300-dimensional GloVe word embeddings [29] which are pretrained on a large-scale corpus is used to further transform each word into a vector. Therefore, we can get the $n \times 300$ word embedding sequences, where $n \in [1, 14]$ denotes the number of words of the input question. Then we apply a one layer LSTM network with d_y hidden units to process the word embeddings and thus obtain a question features matrix $Y \in \mathbb{R}^{n \times d_y}$.When the number of image regions or the number of question words is less than the maximum size (i.e., m=100, n=14), we leverage zero-padding to fill their corresponding feature matrices to their maximum sizes.

3.3 Improved modular co-attention networks

IMCN is stacked by a set of IMC layers which consist of two basic attention units, i.e., the Improved Self-Attention (ISA) unit and the Improved Guided-Attention (IGA) unit. Through the combination of them in different ways, we further develop three variants of IMC. And we select the best one as the

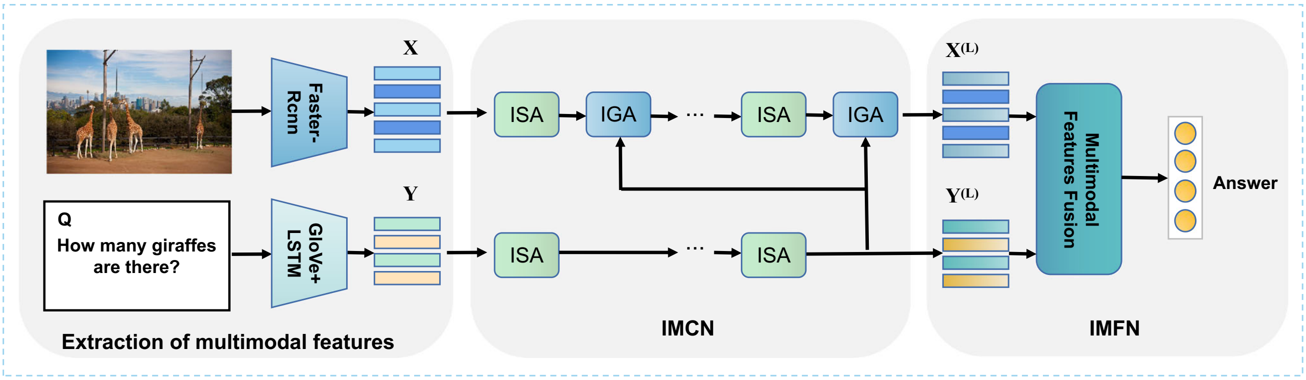

X
Faster-Rcnn
Q
How many giraffes are there?
GloVe+LSTM
Y
Extraction of multimodal features
ISA
IGA
...
ISA
IGA
ISA
...
ISA
IMCN
X(L)
Y(L)
Multimodal Features Fusion
Answer
IMFN

Fig. 2 Overall framework of the proposed model

Table 1 Key Notations and Descriptions

Notation	Description
$\mathbb{R}$	the set of real numbers
X	the image features
Y	the question features
m	the number of selected image regions
n	the number of words of the input question
I	the information vector
G	the attention gate
$\widetilde{I}$	the final attended information
l	the location of the IMC layer
L	the total number of the IMC layers
$X^{(l)}, Y^{(l)}$	the output of image features and question features of the l-th IMC layer
α, β	the attention weight matrices of image features and question features
$\mathbf{I}$	the identity matrix
$\|\|.\|\|_F$	the squared frobenius norm
$\hat{x}, \hat{y}$	the final image features and question features
s	the vector of final prediction answer

component of our model. The details are described in the following.

3.3.1 ISA and IGA units

ISA and IGA can be viewed as two extensions of self-attention and guided-attention proposed in [11] which use “scale dot-product attention” [29]. The scaled dot-product attention uses Q, K and V as inputs and outputs weighted combination of values. Specifically, it firstly computes the dot product of the Q and the K and divides by $\sqrt{d}$, where d is the dimension of K. The results is normalized with a soft-max function to acquire the attention weights of V. Then the attended features F can be obtained by applying weighted summation to attention weights and V:

$$F = att(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V. \tag{1}$$

To enhance the model’s representation capacity further, the *multi-head attention* mechanism, which consists of h parallel heads, is introduced in [29]. Then the attended features can be described as following formulas:

$$head_i = att(QW_i^Q, KW_i^K, VW_i^V), \tag{2}$$

and

$$F = mha(Q, K, V) = [head_1; head_2; ...; head_h]W^o, \tag{3}$$

where W_i^Q, W_i^K, W_i^V represent the projection matrices for each head, $head_i$ represents the computation of each head and “[;]” refers to the operation of concatenation. When nothing satisfies the given query, the multi-head attention module still generates a vector that is unrelated to the query, thus leading to generating a wrong answer for the VQA task. Therefore, we leverage the AoA proposed in [12] to improve the traditional multi-head attention module, which means that another attention function is used after obtaining the result of multi-head attention. The process of the second attention function can be described as follows:

$$I = W_I^Q Q + W_I^V \widetilde{V} + b_I, \tag{4}$$

$$G = \sigma(W_G^Q Q + W_G^V \widetilde{V} + b_G), \tag{5}$$

$$\widetilde{I} = G \odot I, \tag{6}$$

where $W_I^Q, W_I^V, W_G^Q, W_G^V \in \mathbb{R}^{d\times d}$, $b_I, b_G \in \mathbb{R}^d$, $\widetilde{V} = att(Q, K, V)$, $\odot$ represents the element-wise multiplication, and $I, G, \widetilde{I}$ represent the information vector, the attention gate and the final attended information respectively.

ISA also applies a feed-forward layer which uses two fully connected layers to perform ReLU activation and Dropout. Furthermore, to promote optimization, residual connection and layer normalization are used after the AoA and the feed-forward layer. The structure of the ISA unit is shown in Fig. 3. IGA is similar to ISA, the difference between them is that IGA takes visual features X and textual features Y as input where X is guided by Y, while ISA takes X or Y as input. And the structure of the IGA unit is shown in Fig. 4.

On the basis of ISA and IGA units, we develop three kinds of IMC layers which can be cascaded (see Fig. 5)

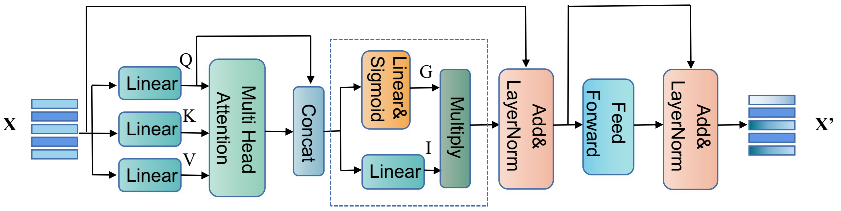


Fig. 3 Frame diagram of ISA unit

same as [11]. It means that the output from the last IMC layer is directly fed into the next IMC layer, and the number of features and the dimensionality of each feature remain unchanged.

- **IGA(X,Y)-ID(Y):** The IGA(X,Y)-ID(Y) (Fig. 5(a)) is the simplest unit that the input question features are straightly ouput. And the attended image features are obtained with an IGA unit by modeling the inter-modal interaction between each object $x_i \in X$ and each word $y_i \in Y$.
- **IGA(X,Y)-ISA(Y):** In comparison to IGA(X,Y)-ID(Y), the IGA(X,Y)-ISA(Y) (Fig. 5(b)) adds an ISA unit that models the intra-model interaction of all word pairs $(y_i, y_j) \in Y$.
- **ISGA(X,Y)-ISA(Y):** ISGA(X,Y)-ISA(Y) (Fig. 5(c)) adds an ISA unit to model the intra-model interaction of all objects $(x_i, x_j) \in X$ based on the IGA(X,Y)-ISA(Y).

It should be noted that the three IMC layers above do not cover all cases. We have also experimented with other IMC variants such as ISGA(X,Y)-ID(Y), IGA(X,Y)-IGA(Y,X) and ISGA(X,Y)-ISGA(Y,X). Nevertheless, we do not report these variants in the following since their performance is not comparative.

3.3.2 Cascade of IMC layers

By cascading, three aforementioned IMC variants are used to form the deep co-attention learning module. We denote the l-th IMC layer as IMC$^{(l)}$, where $l \in [1, L]$, L represents the total number of IMC layers. The output of visual features and textual features from the last IMC layer is directly inputted to the next IMC layer, which can be seen as a recursive process. Let $X^{(l-1)}, Y^{(l-1)}$ and $X^{(l)}, Y^{(l)}$ represent the input features and output features of the l-th IMC layer respectively. This process can be formalized as:

$$[X^{(l)}, Y^{(l)}] = IMC^{(l)}([X^{(l-1)}, Y^{(l-1)}]). \tag{7}$$

For IMC$^{(1)}$, the input features are set as: $X^{(0)} = X$ and $Y^{(0)} = Y$, where X, Y come from the Section 3.2.

We take the ISGA(X,Y)-ISA(Y) layer as an example to construct two kinds of co-attention models (Fig. 6) same as [11]. The stacking model (Fig. 6(a)) just simply stacks the L IMC layers together, while the encoder-decoder model (Fig. 6(b)) substitutes the input textual features $Y^{(l)}$ of the IGA unit of each IMC$^{(l)}$ to the textual features $Y^{(L)}$ which come from the final IMC layer. There is a special case that the two models are equivalent when $L = 1$.

3.4 Improved multimodal fusion networks

After obtaining the image features $X^{(L)} \in \mathbb{R}^{m\times d}$ and the question features $Y^{(L)} \in \mathbb{R}^{n\times d}$, an Improved Multimodal Fusion Network (IMFN) (Fig. 7) is introduced to fuse them [2].

Specifically, the image and question features are first passed through the multilayer perceptron (MLP) to calculate different attention weights over multiple heads. Let h_m represents the number of heads, and this process can be described as follows:

$$\alpha = softmax(MLP(X^{(L)})), \tag{8}$$

$$\beta = softmax(MLP(Y^{(L)})), \tag{9}$$

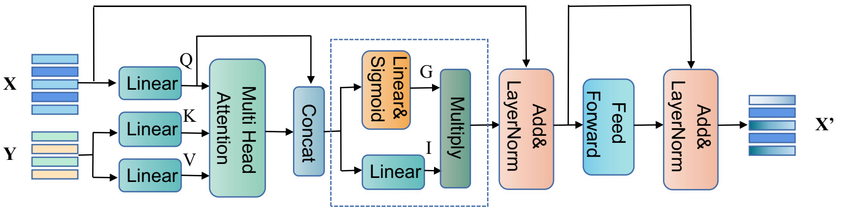
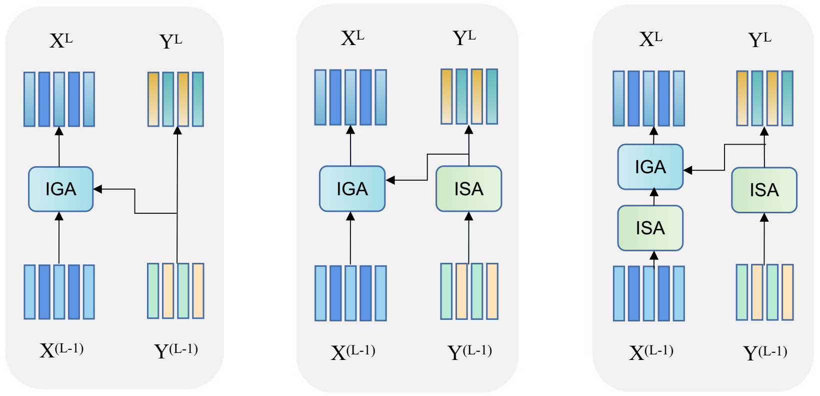
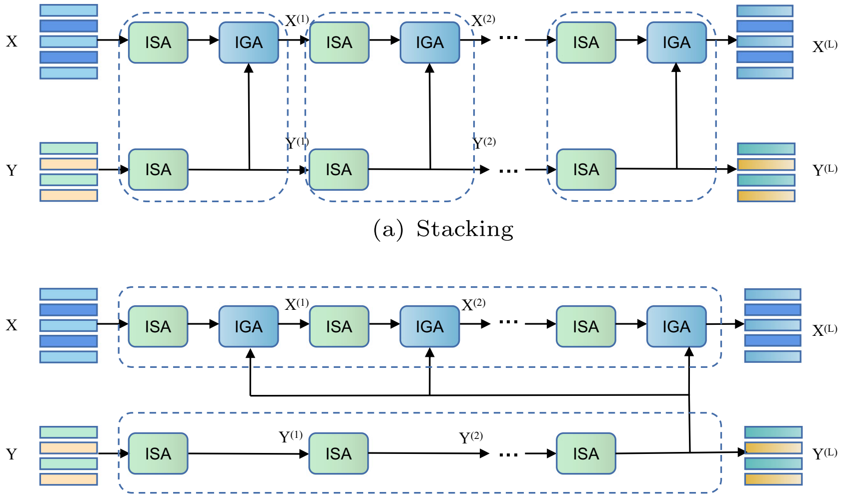


Fig. 4 Frame diagram of IGA unit



(a) IGA(X,Y)-ID(Y) (b) IGA(X,Y)-ISA(Y) (c) ISGA(X,Y)-ISA(Y)

Fig. 5 Three variants of IMC



(a) Stacking

(b) Encoder-Decoder

Fig. 6 Two cascading methods based on ISGA(X,Y)-ISA(Y)

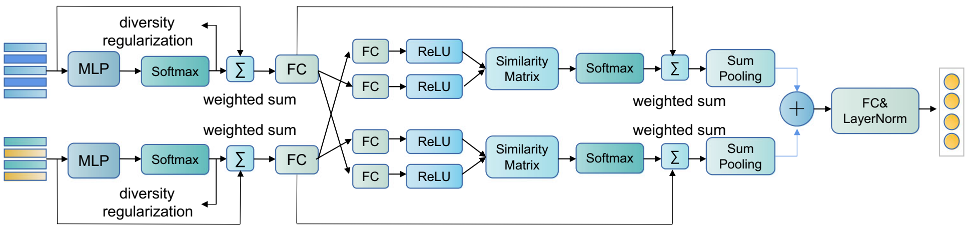


Fig. 7 Overall structure of Improved Multimodal Fusion Network

where $\alpha \in \mathbb{R}^{h_m \times m}$, $\beta \in \mathbb{R}^{h_m \times n}$. Then the attended image features $\widetilde{X}$ and the question features $\widetilde{Y}$ are acquired by applying a weighted summation:

$$\widetilde{X} = \sum_{i=1}^{m} \alpha_i x_i, \tag{10}$$

$$\widetilde{Y} = \sum_{i=1}^{n} \beta_i y_i, \tag{11}$$

where $\widetilde{X}, \widetilde{Y} \in \mathbb{R}^{h_m \times d}$, α_i is each column vector in α, and β_i is each column vector in β. Furthermore, to obtain complementary information between visual modality and textual modality, we use cross attention which firstly computes a similarity matrix of two modalities. After that, the attention weights are acquired with a softmax function. By leveraging weighted summation and sum pooling, we get two vectors $\hat{x}, \hat{y}$, which represent the final visual and textual features respectively. Then, they are combined together followed by a LayerNorm layer. Finally, we use a fully connected layer to predict the answer:

$$\widetilde{s} = sigmoid(Linear(LayerNorm(W_x^T \hat{x} + W_y^T \hat{y}))), \tag{12}$$

where $\widetilde{s} \in \mathbb{R}^N$, N is the number of the most common answers of the training set. We leverage binary cross-entropy same as [29] as loss function ($Loss_{BCE}$):

$$Loss_{BCE} = -\sum_{i}^{M}\sum_{j}^{N} s_{ij} \log(\widetilde{s}_{ij}) - (1 - s_{ij}) \log(1 - \widetilde{s}_{ij}), \tag{13}$$

where M, N represent the number of training questions and candidate answers, $s_{ij}, \widetilde{s}_{ij}$ represent the ground truth and prediction answer. To solve the redundancy problem that attention may focus on similar regions, motivated by [43], we add a diversity regularization loss ($Loss_{DR}$):

$$Loss_{DR} = ||(\alpha\alpha^T - \mathbf{I})||_F^2 + ||(\beta\beta^T - \mathbf{I})||_F^2, \tag{14}$$

where α, β representing the attention weight matrices are from (8) and (9). $\mathbf{I} \in \mathbb{R}^{h_m \times h_m}$ is the identity matrix. And "$||.||_F$" denotes the squared Frobenius norm. Finally, the total loss is described as follows:

$$Loss_{total} = Loss_{BCE} + \lambda Loss_{DR}, \tag{15}$$

where λ represents a coefficient.

4 Experiments

We conduct a series of experiments on VQA-v2 [44] and GQA [45] to validate the effectiveness of our proposed model. In this section, we first describe the dataset used in our experiments and the specific experimental setup. Then we perform some ablation studies to compare the performance of different variants of our model with MCAN. Finally, we provide a visualization of our model and compare the performance of our model with other existing VQA methods.

4.1 Dataset

VQA-v2, which consists of images and question-answer pairs annotated by humans, is one of the most commonly used VQA benchmark dataset [44]. The images used are from the MS-COCO dataset and can be classified into two parts: real images and abstract images. For each image, there are three questions with 10 answers to each question. The VQA-v2 dataset is divided into three parts: train set, val set, and test set. They include $80k, 40k, 80k$ images and $444k, 214k, 448k$ question answer pairs respectively. Moreover, the test set is further divided into the test-dev set and the test-std set. And the experiment results are classified into three types: *Yes/No*, *Number*, and *Other*, according to the types of questions. In this paper, for VQAv2 dataset, we use the standard accuracy and MRR to evaluate our method.

GQA, is a VQA dataset that aims to eliminate language priors. Compared to VQA-v2, the answer of the GQA [45] dataset tends to be acquired from the image itself. Therefore, the questions in GQA do not concern external knowledge. However, many GQA questions require multi-step inference and understanding of spatial, thus are more challenging than VQAv2. These questions are categorized into two kinds: *Binary* and *Open*. In addition, the dataset uses multiple evaluation metrics other than accuracy, such as consistency and validity. In this paper, for GQA dataset, we use accuracy, consistency, validity, and distribution to evaluate our method.

4.2 Experimental settings

Following the settings proposed in [11], the dimensions of input image features and input question features denoted by d_x, d_y are 2048, 512. The latent dimension d is 512, and the number of attention heads is 8.

We leverage the Adam solver to train our IMCN model, and set β_1, β_2 to 0.9 and 0.98. Moreover, we use a coefficient λ to adjust the proportion of two different loss functions and λ is 0.1. The models are trained for 13 epochs. From the first epoch, the learning rate is set to $2.5e^{-5}$ and during training of the first ten epochs, it changes as $\min(2.5te^{-5}, 1e^{-4})$, where t represents the number of iterations currently starting with 1. After 10 epochs the learning rate decays by 1/5 on the

Table 2 Comparative experimental results of the three variants of MCA and IMC with different layers

Model	L=1		L=2		L=4		L=6	
	MCA	IMC	MCA	IMC	MCA	IMC	MCA	IMC
IGA(X,Y)-ID(Y)	64.80%	64.80%	65.30%	65.50%	65.60%	65.80%	65.70%	66.00%
IGA(X,Y)-ISA(Y)	65.20%	65.30%	65.70%	66.30%	66.30%	66.90%	66.50%	67.20%
ISGA(X,Y)-ISA(Y)	65.40%	65.70%	66.30%	66.80%	66.90%	67.70%	67.30%	68.10%

Both MCA and IMC are cascaded in an encoder-decoder manner. The results are acquired on the val split

11-th and the last epochs. When experimenting to verify the validity of the IMCN module, we use (13) as a loss function. And when verifying the validity of the IMFN module, we use (15).

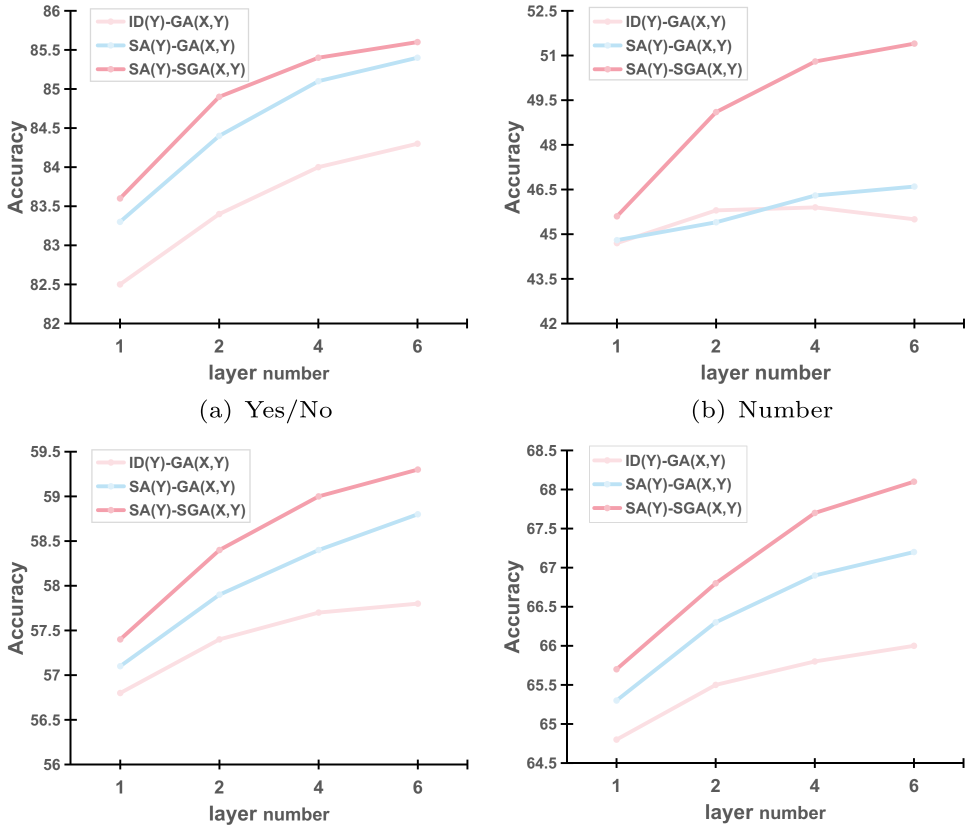
4.3 Ablation studies

We conduct a series of experiments to verify the effectiveness of the IMCN proposed by us. To verify which IMC variant works better, we conduct experiments by combining three IMC variants at different numbers of layers. To verify which stacking method is more effective, we compare the experimental results of the two methods. Moreover, we also conduct experiments with and without the IMFN module to verify its effectiveness of it.

Different variants of IMC

We perform experiments with different IMC variants at different number of layers to validate the effectiveness of introducing the AoA module. Our method develops three variants of IMC, i.e., IGA(X,Y)-ID(Y), IGA(X,Y)-ISA(Y), ISGA(X,Y)-ISA(Y), are improvements to the three variants of MCA proposed in MCAN [11], which corresponds to GA(X,Y)-ID(Y), GA(X,Y)-SA(Y), SGA(X,Y)-SA(Y) respectively (For convenience, we name each MCA variant and the corresponding IMC variant uniformly in Table 2).

From Table 2, we can see that almost all of our proposed variants outperform the corresponding variants in MCAN [11] whatever the number of layers, which verifies the effectiveness of introducing the AoA module.



(a) Yes/No

(b) Number

(c) Other

(d) Overall

Fig. 8 Each type and overall accuracy of the $\mathrm{IMCN}_{ed} - L$ model. The results are acquired on the val split

Table 3 Ablation experimental results of stacking and encoder-decoder with different layers

Model	L=2	L=4	L=6	L=8
IMCNsk	66.50%	67.30%	67.60%	67.50%
IMCNed	66.70%	67.70%	68.10%	67.90%

The results are acquired on the val split

Table 4 Ablation experimental results of IMCN and IMFN

Model	Yes/No	Number	Other	All	MRR
MCAN	86.82%	53.26%	60.72%	70.63%	64.83%
IMCN	87.02%	53.68%	60.89%	70.86%	64.91%
IMCN+IMFN	87.10%	53.74%	61.02%	70.95%	65.08%

The results are acquired on the test-dev split of VQA-v2

In addition, we also show the accuracy of different IMC variants on different types of questions in Fig. 8 to verify which IMC variant works better. From Fig. 8, it can be seen that IGA(X,Y)-ISA(Y) outperforms IGA(X,Y)-ID(Y) and ISGA(X,Y)-ISA(Y) outperforms IGA(X,Y)-ISA(Y) when the number of layers is equal. This conclusion is same as [11], revealing that modeling self-attention for visual features and textual features is useful. Thus, ISGA(X,Y)-ISA(Y) is selected as our default IMC in the experiments below.

Different stacking methods

To verify which stacking method works better, we experiment with two different stacking methods with different number of layers. From Table 3, we can see that the stacking model underperforms the encoder-decoder model at all layers. It can be explained that the self-attention to questions learned from the early ISA(Y) unit is not enough compared to the last ISA(Y) unit. Moreover, with the increase of L, we can see that both models perform better than before. And this phenomenon disappears when $L > 6$ due to the unstable gradients that cause optimization difficulty. The best results of the two models are obtained when $L = 6$, which is also the same as [11]. Therefore, we choose encoder-decoder as the default stacking method and $L = 6$ as the default number of layers in the following experiments.

Effectiveness of IMCN and IMFN

The improved part of our proposed model over MCAN mainly consists of two components: IMCN and IMFN. To verify the effectiveness of each component, we compare the results of IMCN and IMCN+IMFN with MCAN. From Tables 4 and 5, we can see that IMCN which consists of ISA units and IGA units outperforms the MCAN. In addition, we think it is necessary to adopt a more complex method to fuse visual and textual features to further improve performance. The results of the third row in Tables 4 and 5 verify the effectiveness of the IMFN module. For the results on VQAv2, it can be seen that IMCN achieves a 0.23% improvement in accuracy compared to MCAN, and IMCN+IMFN achieves a 0.09% improvement compared to IMCN. Moreover, for the results on GQA, IMCN achieves a 0.15% improvement in accuracy compared to MCAN, and IMCN+IMFN achieves a 0.16% improvement compared to IMCN.

4.4 Comparison with existing VQA methods

We evaluate our model against the state-of-the-art models for VQA on VQA-v2 dataset and GQA dataset. For VQA-v2, these methods can be classified to what based on fusion, attention and reasoning. For example, the models of **MCB** [46] and **ResNet-3000** [47] are based on fusion. We also compare our model with other models based on attention. **HAN** [48] leverages generated attention maps to supervise during training. **UpDn** [49] uses the object features which are extracted from Faster RCNN. Moreover, There are some methods based on reasoning such as **Dual-MFA** [50] and **Counting** [51]. In addition, **Img2LLM**$_{175B}$ [28] and **PNP-VQA**$_{11B}$ [27] are two methods that use large language models for VQA. For GQA, apart from MCAN [11], we also compare our model with some previous methods that have achieved the state-of-the-art, such as BUTD [14] and MAC [52]. The results of our model compared with others are shown in Tables 6 and 7.

According to the experimental results in Tables 6 and 7, we get the following meaningful conclusions:

- For VQAv2, compared with attention-based models such as SCAVQAN [36], while maintaining the same level of performance on test-std, IMCN+IMFN achieves a 0.13% improvement on Test-dev. Moreover, compared to MCAN [11] used as our improved original model, our model achieves a 0.32% enhancement on Test-dev and a 0.28% enhancement on Test-std.
- Especially for more complex questions such as “Number”, our model has a significant 0.48% improvement compared to MCAN, which reveals that our model performs better on complex questions. In addition, we also compare our method with Img2LLM$_{175B}$ and PNP-VQA$_{11B}$. The results in Table 6 show that our method still has a large advantage in terms of accuracy.
- For GQA, the results in Table 7 show that our method maintains a similar consistency to MCAN while gaining improvements in all other metrics. The experimental results above demonstrate the effectiveness and the generalizability of our model.

Table 5 Ablation experimental results of IMCN and IMFN

Methods	Binary	Open	Accuracy↑	Consistency↑	Validity↑	Distribution↓
MCAN	75.56%	40.38%	56.84%	87.19%	96.85%	1.31
IMCN	75.65%	40.59%	56.99%	87.07%	96.83%	1.22
IMCN+IMFN	75.68%	40.86%	57.15%	87.18%	96.88%	1.18

The results are acquired on the test-dev split of GQA

Table 6 Results of our model compared with the state-of-the-art models on VQA-v2

Model	Test-dev				Test-std			
	Yes/No	Number	Other	All	Yes/No	Number	Other	All
MCB [46]	-	-	-	-	78.82%	38.28%	53.36%	62.27%
HAN [48]	78.54%	37.94%	53.38%	61.99%	-	-	-	-
ReNet-3000 [47]	-	-	-	-	79.20%	39.50%	52.60%	62.10%
UpDn [49]	81.82%	44.21%	56.05%	65.32%	-	-	-	65.67%
Dual-MFA [50]	83.59%	40.18%	56.84%	66.01%	83.37%	40.39%	56.89%	66.09%
DCN [10]	84.48%	41.66%	57.44%	66.83%	84.61%	41.27%	56.83%	66.66%
Counting [51]	83.14%	51.62%	58.97%	68.09%	83.56%	51.39%	59.11%	68.41%
CoR-3 [53]	85.22%	47.95%	59.15%	68.62%	85.76%	48.40%	59.43%	69.14%
MFH [8]	84.27%	49.56%	59.89%	68.76%	-	-	-	-
BAN [9]	85.31%	50.93%	60.26%	69.52%	-	-	-	-
BAN+Counter [9]	85.42%	54.04%	60.52%	70.04%	-	-	-	70.35%
MCAN [11]	86.82%	53.26%	60.72%	70.63%	-	-	-	70.90%
SCAVQAN [36]	86.96%	53.49%	60.95%	70.82%	-	-	-	71.14%
Img2LLM$_{175B}$ [28]	-	-	-	61.90%	-	-	-	-
PNP-VQA$_{11B}$ [27]	-	-	-	64.80%	-	-	-	-
IMCN+IMFN(Ours)	87.10%	53.74%	61.02%	70.95%	87.32%	53.48%	61.29%	71.18%

Table 7 Results of our model compared with the state-of-the-art models on GQA

Model	Binary	Open	Accuracy↑	Consistency↑	Validity↑	Distribution↓
CNN+LSTM	63.26%	31.80%	46.55%	74.57%	96.02%	7.46
BUTD [14]	66.64%	34.83%	49.74%	78.71%	96.18%	5.98
MAC [52]	71.23%	38.91%	54.06%	81.59%	96.16%	5.34
MCAN [11]	75.56%	40.38%	56.84%	87.19%	96.85%	1.31
SCAVQAN [36]	74.97%	41.18%	56.99%	87.16%	96.67%	1.25
IMCN+IMFN(Ours)	75.68%	40.86%	57.15%	87.18%	96.94%	1.16

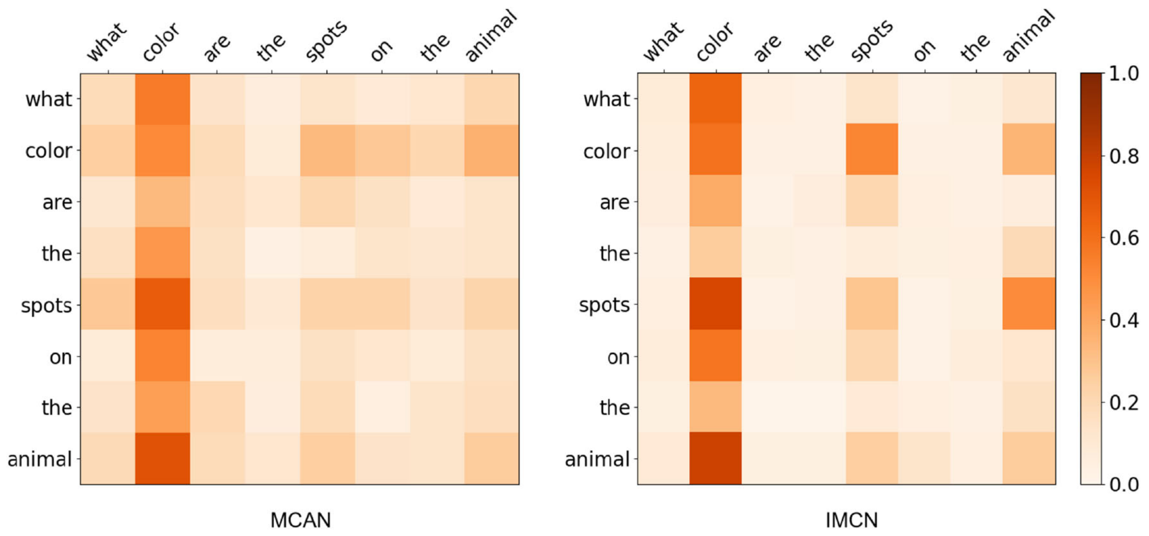


Fig. 9 Textual self-attention maps of MCAN and IMCN

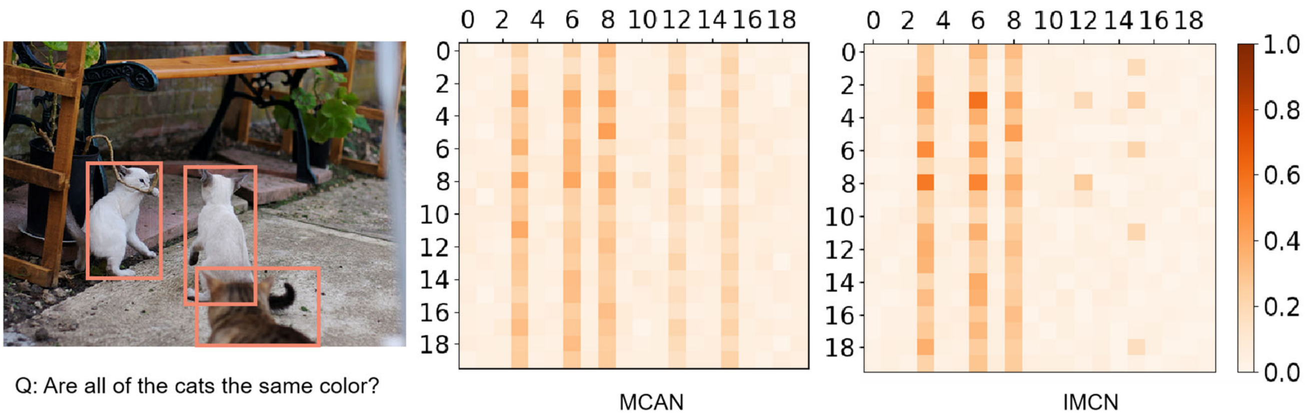


Fig. 10 Visual self-attention maps of MCAN and IMCN

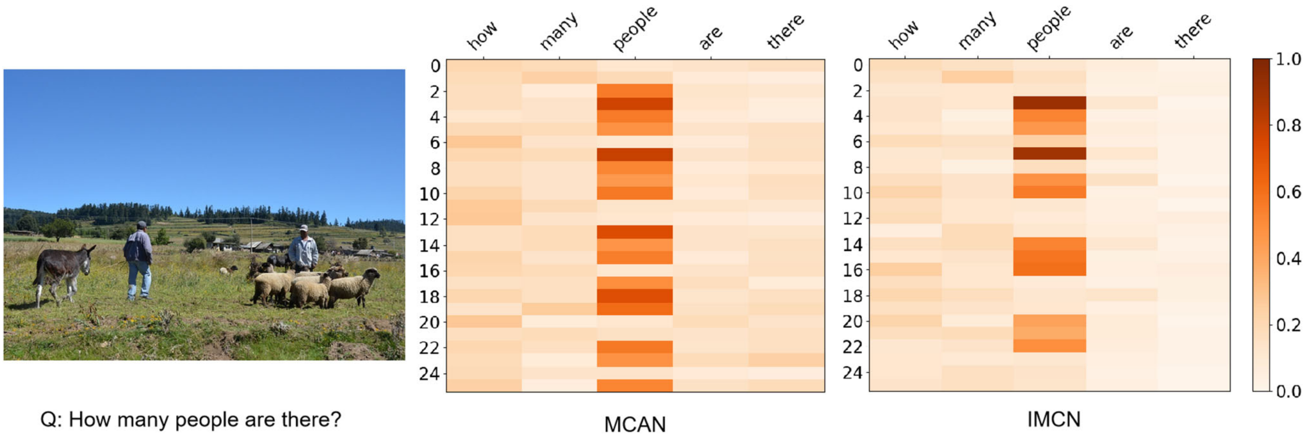


Fig. 11 Visual question-guided attention maps of MCAN and IMCN

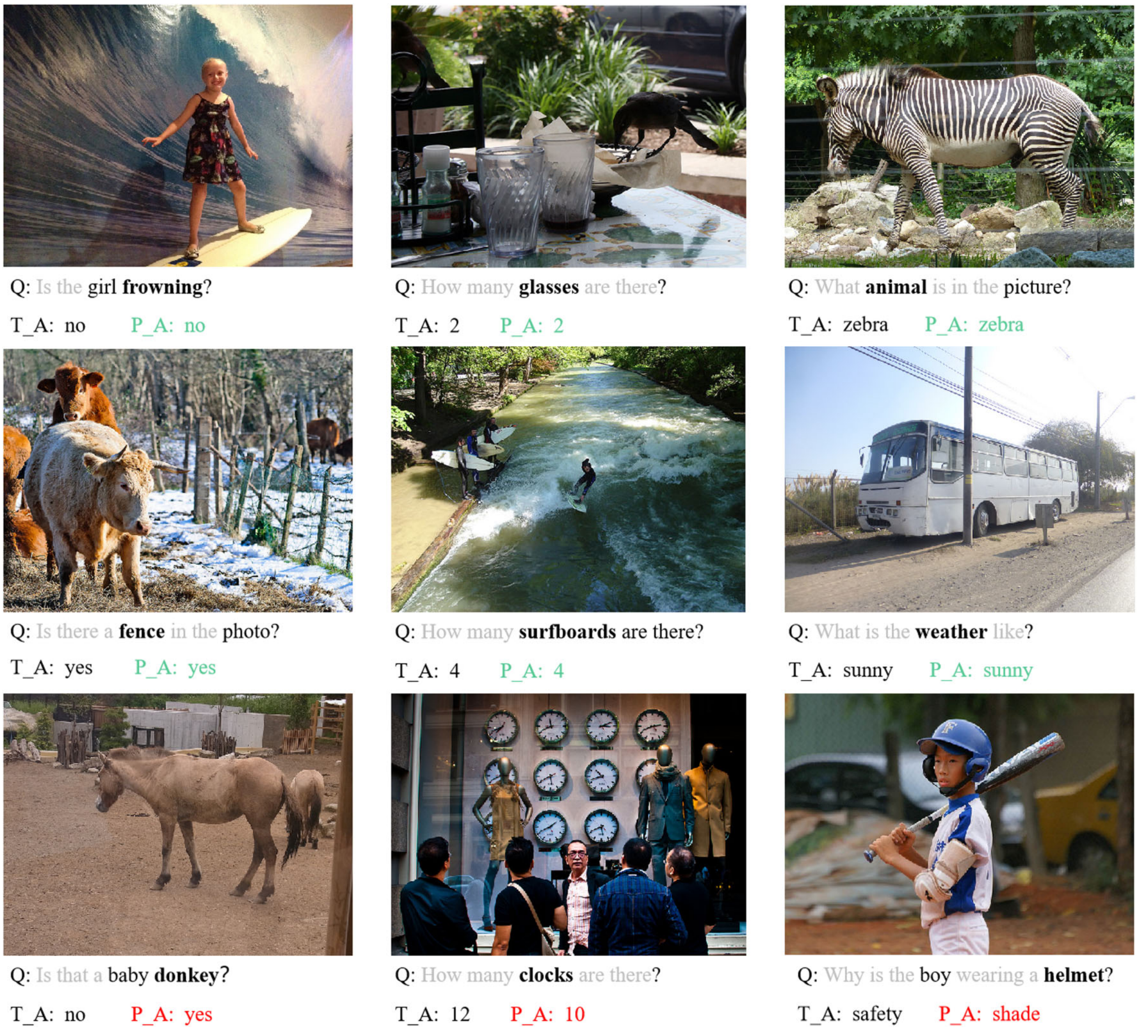
4.5 Attention visualization

To verify the interpretability of our model, the visualization of the learned attention results of our model and MCAN are shown in Figs. 9, 10 and 11.

- Figure 9 shows the learned self-attention of textual features. For the input question: *What color are the spots on the animal?*, we can see that compared to the MCAN, IMCN focuses more on the keywords “color”, “spots” and “animal”, and weakens the focus on other irrelevant words.
- Figure 10 shows the learned self-attention of image features. For the question “Are all of the cats the same color”, both MCAN and IMCN obtain high values on the key objects 3, 6, and 8 (three cats). However, IMCN is able to filter objects 12 and 15 that are not related to predicting the final answer.
- Figure 11 shows the learned question-guided visual attention results. For question: “*How many people are there ?*”, it can be seen that both MCAN and IMCN can find the keyword “people”. However, for the input image, there are multiple image objects related to the question keyword “people”, and IMCN pays more attention to these image objects that facilitate answering the question correctly.

4.6 Qualitative analysis

We also show some qualitative results of our model in Fig. 12. The brightness of each word represents the level of attention of it, which means the word in bold is the most meaningful



Q: Is the girl **frowning**?
T_A: no P_A: no

Q: How many **glasses** are there?
T_A: 2 P_A: 2

Q: What **animal** is in the picture?
T_A: zebra P_A: zebra

Q: Is there a **fence** in the photo?
T_A: yes P_A: yes

Q: How many **surfboards** are there?
T_A: 4 P_A: 4

Q: What is the **weather** like?
T_A: sunny P_A: sunny

Q: Is that a baby **donkey**?
T_A: no P_A: yes

Q: How many **clocks** are there?
T_A: 12 P_A: 10

Q: Why is the boy wearing a **helmet**?
T_A: safety P_A: shade

Fig. 12 Illustration of some results of our method on test-dev split

word. The first row and the second row show instances of correctly answered questions, and the third row shows instances of incorrectly answered questions. For successful instances, the keywords of the question and the key objects corresponding to the question are accurately attended. As for instances of failures, it can be concluded to two reasons. The first is that the model is unable to distinguish the keywords in question or key objects in the image. As shown in the third row of Fig. 12, the first instance fails to distinguish the keyword “baby” and the second instance fails to distinguish all the key objects “clocks”. And the other reason is that the model still can’t predict the true answer despite the extracted visual and textual features already most related, which is common in attention-based models. Taking the third picture as an example, though the keyword “helmet” is attended, the model still gives a wrong answer due to the lack of commonsense of what helmet is used for.

5 Limitation and discussion

One limitation of our proposed approach, which we have shown the example above, is the inability to correctly answer the question requiring commonsense. It is a common problem with models based on the attention mechanisms, since these methods only acquire information through input images and questions. Apart from this limitation, another one is the incapability of mitigating the bias caused by the inconsistent distribution of the datasets. For example, for the question“what colour is the apple?”, when the number of answers in the training dataset that are red is much more than the number of answers that are green, then the model still tends to answer the “red” even though the apple in the picture is green when testing.

6 Conclusion

In this paper, we propose an Improved Modular Co-attention Network which consists of a cascade of Improved Modular Co-attention layers to solve the problem of visual question answering. To further improve the performance, we also introduce an Improved Multimodal Fusion Network. Our proposed model outperforms the existing methods on the VQA benchmark dataset. Extensive experiments which include ablation studies, attention visualization and qualitative analysis verify that our model is capable of filtering out irrelevant visual and textual features. Moreover, we also demonstrate that using AoA to replace the traditional attention mechanism in VQA task effectively improves the model performance. Compared to MCAN, our method requires only about 30% additional computing overhead. AoA is a more advanced attention module that can be used to improve other models that employ the traditional attention mechanism for VQA apart from MCAN. We hypothesize that adopting AoA on other tasks will also yield improved results since we have not validated it on other tasks. However, our model still has some limitations, which have been mentioned in Section 5. For commonsense questions, it may be solved by combining the model with the knowledge base or using the large language model. To debias, using causal attention and data augmentation are two approaches that deserve trying. And these are what we will further explore during future research.

Acknowledgements We thank the anonymous reviewers for their valuable comments. This work was supported by the Program of Natural Science Foundation of Shanghai (No. 23ZR1422800).

Author Contributions Cheng Liu: Writing - Original Draft, Writing - Editing, Software, Data curation.
Chao Wang: Writing - Editing, Investigation.
Yan Peng: Writing - Review, Editing.

Data Availability Data will be available upon request.

Code Availability The code for reproducing the results provided in the manuscript will be made public upon acceptance.

Declarations

Competing interests The authors have no relevant financial or nonfinancial interests to disclose other than the aforementioned funding.

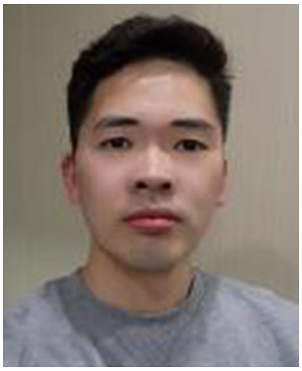
References

1. 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
2. Zhang H, Li R, Liu L (2022) Multi-head attention fusion network for visual question answering. 2022 IEEE International Conference on Multimedia and Expo (ICME), pp 1–6
3. 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
4. Khurana D, Koli A, Khatter K, Singh S (2023) Natural language processing: State of the art, current trends and challenges. Multimed Tools Appl 82(3):3713–3744
5. 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
6. Shih KJ, Singh S, Hoiem D (2016) Where to look: Focus regions for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4613–4621
7. Lu J, Yang J, Batra D, Parikh D (2016) Hierarchical question-image co-attention for visual question answering. Adv Neural Inf Process Syst 29
8. 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

9. Kim J-H, Jun J, Zhang B-T (2018) Bilinear attention networks. Adv Neural Inf Process Syst 31
10. Nguyen D-K, 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, pp 6087–6096
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. Huang L, Wang W, Chen J, Wei X-Y (2019) Attention on attention for image captioning. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 4634–4643
13. Lee K-H, Chen X, Hua G, Hu H, He X (2018) Stacked cross attention for image-text matching. In: Proceedings of the European conference on computer vision (ECCV), pp 201–216
14. 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, pp 4223–4232
15. Nam H, Ha J-W, Kim J (2017) Dual attention networks for multimodal reasoning and matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 299–307
16. Fan H, Zhou J (2018) Stacked latent attention for multimodal reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1072–1080
17. Gurunlu B, Ozturk S (2022) Efficient approach for block-based copy-move forgery detection. In: Smart trends in computing and communications: proceedings of SmartCom 2021, pp 167–174. Springer
18. Le TM, Le V, Gupta S, Venkatesh S, Tran T (2023) Guiding visual question answering with attention priors. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 4381–4390
19. Marino K, Rastegari M, Farhadi A, Mottaghi R (2019) Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, pp 3195–3204
20. Schwenk D, Khandelwal A, Clark C, Marino K, Mottaghi R (2022) A-okvqa: A benchmark for visual question answering using world knowledge. In: European conference on computer vision, pp 146–162. Springer
21. Chang Y, Narang M, Suzuki H, Cao G, Gao J, Bisk Y (2022) Webqa: Multihop and multimodal qa. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 16495–16504
22. Ravi S, Chinchure A, Sigal L, Liao R, Shwartz V (2023) Vlc-bert: Visual question answering with contextualized commonsense knowledge. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1155–1165
23. Garcia-Olano D, Onoe Y, Ghosh J (2022) Improving and diagnosing knowledge-based visual question answering via entity enhanced knowledge injection. Companion Proceedings of the Web Conference 2022:705–715
24. Ding Y, Yu J, Liu B, Hu Y, Cui M, Wu Q (2022) Mukea: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5089–5098
25. Gao F, Ping Q, Thattai G, Reganti A, Wu YN, Natarajan P (2022) Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5067–5077
26. Yang Z, Gan Z, Wang J, Hu X, Lu Y, Liu Z, Wang L (2022) An empirical study of gpt-3 for few-shot knowledge-based vqa. Proceedings of the AAAI conference on artificial intelligence 36:3081–3089
27. Tiong AMH, Li J, Li B, Savarese S, Hoi SC (2022) Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training. Findings of the Association for Computational Linguistics: EMNLP 2022:951–967
28. Guo J, Li J, Li D, Tiong AMH, Li B, Tao D, Hoi S (2023) From images to textual prompts: Zero-shot visual question answering with frozen large language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10867–10877
29. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30
30. 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
31. Kim J-H, Lee S-W, Kwak D, Heo M-O, Kim J, Ha J-W, Zhang B-T (2016) Multimodal residual learning for visual qa. Adv Neural Inf Process Syst 29
32. Zhu C, Zhao Y, Huang S, Tu K, Ma Y (2017) Structured attentions for visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp 1291–1300
33. 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, pp 4709–4717
34. Zhou L, Palangi H, Zhang L, Hu H, Corso J, Gao J (2020) Unified vision-language pre-training for image captioning and vqa. Proceedings of the AAAI conference on artificial intelligence 34:13041–13049
35. 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
36. Guo Z, Han D (2023) Sparse co-attention visual question answering networks based on thresholds. Appl Intell 53(1):586–600
37. Cheng B, Misra I, Schwing AG, Kirillov A, Girdhar R (2022) Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1290–1299
38. Liu S-A, Xie H, Xu H, Zhang Y, Tian Q (2022) Partial class activation attention for semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 16836–16845
39. Liang J, Fan Y, Xiang X, Ranjan R, Ilg E, Green S, Cao J, Zhang K, Timofte R, Gool LV (2022) Recurrent video restoration transformer with guided deformable attention. Adv Neural Inf Process Syst 35:378–393
40. Song CH, Han HJ, Avrithis Y (2022) All the attention you need: Global-local, spatial-channel attention for image retrieval. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 2754–2763
41. Xia Z, Pan X, Song S, Li LE, Huang G (2022) Vision transformer with deformable attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4794–4803
42. Zhang C, Wan H, Shen X, Wu Z (2022) Patchformer: An efficient point transformer with patch attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11799–11808
43. Park G, Han C, Yoon W, Kim D (2020) Mhsan: multi-head self-attention network for visual semantic embedding. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1518–1526

44. 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
45. 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
46. Fukui A, Park DH, Yang D, Rohrbach A, Darrell T, Rohrbach M (2016) Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Proceedings of the 2016 conference on empirical methods in natural language processing. Assoc Comput Linguist
47. Ma C, Shen C, Dick A, Wu Q, Wang P, van den Hengel A, Reid I (2018) Visual question answering with memory-augmented networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6975–6984
48. Qiao T, Dong J, Xu D (2018) Exploring human-like attention supervision in visual question answering. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
49. 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
50. Lu P, Li H, Zhang W, Wang J, Wang X (2018) Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
51. Zhang Y, Hare J, Prügel-Bennett A (2018) Learning to count objects in natural images for visual question answering. In: International conference on learning representations
52. Hudson DA, Manning CD (2018) Compositional attention networks for machine reasoning. In: International conference on learning representations
53. Wu C, Liu J, Wang X, Dong X (2018) Chain of reasoning for visual question answering. Adv Neural Inf Process Syst 31

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



Cheng Liu received the B.S. degree from the School of Architecture and Planning, Anhui Jianzhu University, China, in 2019. He is currently pursuing the Master degree at the School of Future Technology, Shanghai University, China. His research interests include visual question answering and large language models.

Chao Wang received his Ph.D. degree from the School of Computer Science, Fudan University in 2022. He is a lecturer at the School of Future Technology, Shanghai University. His research interests include natural language processing, knowledge bases, and causal inference.

Yan Peng received her Ph.D. degree in pattern recognition and intelligent systems from the Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, China, in 2009. She is currently the Dean of the School of Future Technology at Shanghai University, Shanghai, China. Her current research interests include multi-modal machine learning, modeling and control of energy harvesting, unmanned surface vehicles, field robotics, and locomotion systems.