Computer Science and Information Systems 22(1):199–217 https://doi.org/10.2298/CSIS240321003L

ASAM: Asynchronous Self-Attention Model for Visual Question Answering

Han Liu[1], Dezhi Han[1], Shukai Zhang[1], Jingya Shi[1], Huafeng Wu[2,⋆], Yachao Zhou[3], and Kuan-Ching Li[4,⋆]

[1] College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China
liuhanshmtu@163.com
dzhan@shmtu.edu.cn
zhang_shukai11@163.com
jingyashi00@163.com
[2] Merchant Marine College, Shanghai Maritime University, Shanghai 201306, China
hfwu@shmtu.edu.cn
[3] Shanghai Anheng Times Information Technology Co., Ltd.
Shanghai 200131, China
anna.zhou@dbappsecurity.com.cn
[4] Dept of Computer Science and Information Engr (CSIE), Providence University, Taiwan
kuancli@pu.edu.tw

Abstract. Visual Question Answering (VQA) is an emerging field of deep learning that combines image and question features and generates collaborative feature representations for classification by uniquely fusing the components. To enhance the effectiveness of models, it is crucial to fully utilize the semantic information from both text and vision. Some researchers have improved the accuracy of the model's training by either adding new features or enhancing the model's ability to extract more detailed information. However, these methods have made experimentation more challenging and expensive. We propose a model called asynchronous self-attention model (ASAM) that makes use of an asynchronous self-attention component and a controller, integrating the asynchronous self-attention mechanism and collaborative attention mechanism effectively to leverage the rich semantic information of the underlying visuals. It realizes an end-to-end training framework that can extract and exploit the rich representational information of the underlying visual images while performing coordinated attention with text features, as it does not over-emphasize fine-grained but finds a balance within it, thus allowing the model to learn more valuable information. Extensive ablation experiments were conducted on the proposed ASAM using the VQA v2 dataset to verify its effectiveness. The results of the experiments demonstrate that the proposed model outperforms other state-of-the-art models, without increasing the model complexity and the number of parameters.

Keywords: Visual Question Answering, Asynchronous Self-Attention, Deep Collaborative, Controller.

⋆ Corresponding authors

1. Introduction

With the comprehensive development of deep learning, single-modal tasks involving computer vision (CV) and natural language processing (NLP) no longer meet the demands of technological advancement. Consequently, there has been a growing emphasis on multimodal task learning involving the collaboration of image and text modalities. Among these multi-modal tasks [28,43,25], a pivotal research direction is visual question answering (VQA)[37,23].

The VQA task aims to accurately answer the natural language questions about images. This requires the model to not only comprehend the content of both the images and the questions but also to grasp the intricate relationships between them. Initially, training models of convolutional neural network developed rapidly and involved in many fields [8,5]. The research on VQA also focused on the training model using convolutional neural networks [9,29,35] and subsequently evolved to incorporate attention mechanisms which are now widely used in scientific research [6]. Several studies [27,31,44] have demonstrated that methods leveraging conventional self-attention mechanisms have provided significant impetus to advancing visual question answering (VQA) tasks. However, their performance remains limited as traditional attention mechanisms fail to effectively model the intricate relationships between the two modalities.

With the advancement of computing and communication technologies, a disruptive new architecture has emerged-the Transformer [40]. Initially devised to address problems in NLP, the Transformer has progressively found success in applications to computer vision and multimodal tasks. This success is attributed to its attention mechanism, which excels in comprehensively modeling the relationships between modalities and has been applied in various fields[5,26] Yu et al. [45] were the first to employ the transformer model in the VQA, leading to the model securing the championship in the 2019 VQA Challenge. Due to the simplicity and effectiveness of the transformer's encoder-decoder structure and attention mechanism and its ability to capture long-distance dependencies, there have been many attempts in VQA so far[3,42,48,4,12,20]. However, it has been observed that fine granularity can be used in all fields[24].If most of the work emphasizes and over-attention to it, it may lead to the loss of some effective information [24,21]. For this reason, we consider whether there can be a solution to the issue above.

Regarding fusing features from different layers, ResNet [17] introduces an identity shortcut connection structure that directly skips one or more layers and fuses features between different layers to solve the problem of gradient disappearance. Qin et al. [34] propose a Residual Weight-Sharing Attention Network (RWSAN), wherein within each attention unit of the RWSAN layer, residual learning is performed using learnable connectivity patterns and shared parameters. Drawing upon this conceptual framework, we propose an asynchronous self-attention mechanism combined with collaborative attention, resulting in the design of the Asynchronous Self-Attention Model (ASAM). ASAM can perform bottom-up connections to the attention map from the previous layer's output, thereby balancing the coarseness of the granularity in image representation. This ensures that the model focuses on crucial regions of the image without introducing additional complexity, thereby allowing the effective extraction of rich semantic information from the image. Extensive ablation experiments based on the VQA v2 benchmark dataset prove the effectiveness of our proposed models. The main contributions of this paper are as follows:

(1) An asynchronous self-attention mechanism is proposed to optimize the balance between coarse-grained and fine-grained image representations. Simultaneously, a controller is designed to optimize the features computing attention scores during the self-attention modeling process.

(2) By integrating collaborative attention with the designed self-attention mechanism, we propose an Asynchronous Self-Attention Model (ASAM). This model is capable of coordinating relationships between objects of different granularities and collaboratively attending to image features in conjunction with text.

(3) We conducted extensive experiments on the benchmark dataset VQA v2, and the results indicate that the proposed ASAM achieved favorable performance without increasing model complexity or the numbers of parameters.

The remainder of this work is organized as follows. We introduce the work related to Visual Question Answering research in section 2. Then section 3 describes the asynchronous self-attention mechanism in detail. Next, section 4 verifies the validity of the model through extensive experiments, and finally, the concluding remarks and a prospect for future directions are given in section 5.

2. Related Work

2.1. Visual Question Answering

The essence of the visual question answering task lies in the simultaneous comprehension of the input question and image, coupled with a capacity for reasoning to accurately respond to natural language inquiries about the image. Over the past few years, an increasing number of researchers have devoted themselves to investigating VQA tasks, leading to a diverse array of methods that contribute to enhancing task performance. Models based on the transformer architecture have gained more widespread application [3,16,30,15]. Mao et al. [30]proposed an approach guided by positional attention, significantly enhancing the model's performance by incorporating three distinct positional attention modules into a single transformer model. Chen et al. [2] proposed for the first time to introduce contextual information with different combinations of representations into VQA, and proposed a context-aware attention network (CAAN) to solve the problem of existential comprehension bias, marking a novel breakthrough built upon the foundation of MCAN [45]. Furthermore, visual question answering requires models to possess extensive multimodal knowledge beyond specific domains to enable models to answer more abundant questions. Consequently, some researchers leverage large-scale knowledge bases for information extraction, allowing the models to infer image content and answer questions requiring common-sense knowledge not explicitly covered in the image [47,7,41]. Building upon optimizing features, some researchers have proposed methods such as feature filtering, gating mechanisms, and stepwise refinement of features from coarse to fine. Nguyen et al. [32] extract predicates simultaneously with features, enabling dual learning of coarse-grained and fine-grained information and achieving robust reasoning. Guo et al. [13] utilized top-k filtering, explicitly selecting the most crucial information from both the image and the question to concentrate attention, proposing a novel multi-modal explicit sparse attention network. Diverging from other methods that focus on refining features, we leverage the output of the model's preceding layer to influence the input of the next layer, thereby balancing the coarseness in granularity between feature representations. We optimize the model by integrating heterogeneous self-attention and modular co-attention networks.

2.2. Attention Model

Referring to how humans process information when seeing images, the researchers consider that the model should be able to recognize what and where the object in the images is when faced with a question. The location of the model's gaze should be the object's position in the image most relevant to the question. Due to its capability to dynamically modulate attention towards critical regions or words across multimodal modeling processes, and to allocate weights based on feature importance, the attention mechanism confers significant advantages in addressing VQA tasks. Consequently, it has been extensively adopted in this task to enhance modeling processes across diverse modalities effectively. VQA models can concentrate attention on relatively important information by incorporating attention mechanisms, thereby reducing interference from irrelevant information. Yang et al. [44] devised a multi-layer attention network to address noise induced by global features, which is the first attempt of attention mechanisms in VQA tasks, yielding promising performance. Anderson et al. [1] introduced a Bottom-up and Top-down Attention (BUTD) network to identify prominent image region features within the model. However, aside from image features, learning textual features is equally crucial. In general, visual attention assists the model in focusing on critical image regions, while textual attention attends to essential words. Consequently, dense co-attention over both images and questions is currently prevalent. Lu et al. [27] devised a hierarchical collaborative attention model, wherein the architecture constructs a collaborative attention graph at three levels: word level, phrase-level, and question level. Emphasizing mutual guidance between text and modalities, the model, however, is limited to learning coarse interactions between modalities. Nam et al. [31] introduced a dual attention network (DAN) for multimodal reasoning, employing multi-step reasoning to mutually guide visual and textual attention. Yu et al. [45] proposed a multimodal dense co-attention network by modeling dense interactions within and between modalities, representing a significant breakthrough in attention mechanisms. Chen et al. [3] designed a textual global-context module and a compact attention mechanism, introducing a multimodal vision-language paradigm that enhances the modeling dependencies capability of image tokens and the model's reasoning ability.

3. Method

3.1. Model Components

Before presenting the complete model framework, this section first introduces the essential components of the model. The collaborative asynchronous self-attention attentive layer consists of three basic units: Self-Attention unit (SA), Asynchronous Self-Attention unit (ASA), and Guided-Attention unit (GA).

As shown in Fig. 1(a), only one input, denoted as X, represent either text or visual features. In Fig. 1(b), Y_i and $Y_{(i-1)'}$, are image features, and $Y_{(i-1)'}$ is the reserved feature of the previous layer. In Fig. 1(c), X and Y denote text features and image features respectively. And Z represents output features.

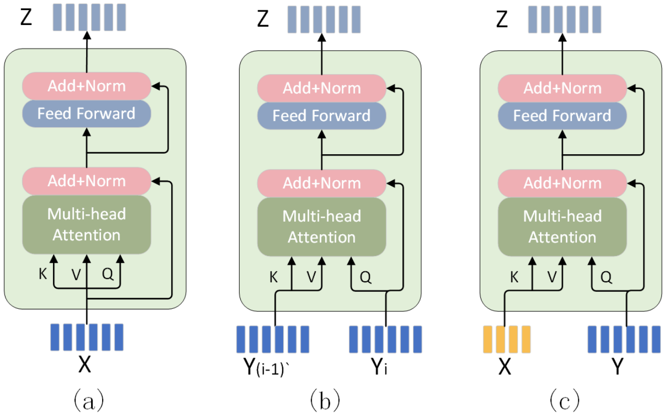

Fig. 1. Three base components in the proposed SASM model

Self-Attention, Asynchronous Self-Attention and Guided-Attention Units. The attention mechanism used in this paper is drawn from [40]. In feature processing, question and image features are transformed into queries, keys, and values feature. d_k and d_v are the dimensions of the keys and queries that make up the scaled dot-product attention's input in attention mechanisms, respectively. We calculate the dot-product of queries with all keys and divide the result by $\sqrt{d}$. Finally, we use the softmax function to obtain the attention weights on the values. In practice, to compute the attention weights on a set of queries simultaneously, we pack queries into matrix $Q\epsilon\mathbb{R}^{1\times d}$, and pack the keys and values into matrices $K\epsilon\mathbb{R}^{n\times d}$ and $V\epsilon\mathbb{R}^{n\times d}$. The calculation process of self-attention specific is shown as follows:

$$f = Attention\left(Q, K, V\right) = softmax\left(\frac{QK^T}{\sqrt{D_k}}\right)V \tag{1}$$

To improve the representational capability of the features, the attention mechanism uses a multi-head attention mechanism to enrich the feature information by jointly focusing on the representation subspace at different locations. Taking the example of h separate heads, the output feature f can be represented as follows:

$$f = MHA\left(Q, K, V\right) = Concat\left(head_1, head_2..., head_h\right)W^O \tag{2}$$

$$head_i \ = \ Attention\left(Q^i, K^i, V^i\right) \tag{3}$$

where $W_i^Q, W_i^K, W_i^V \epsilon\mathbb{R}^{d\times d_h}$ are the projection matrices of Q, K ,V in the i-th head, respectively. $W^O\epsilon\mathbb{R}^{h\times d_h\times d}$ is the projection parameter matrix, d_h is the dimension of each header output feature and is generally set to $d_h = d/h$. This setting aims to prevent multi-head attention models from becoming too large and consuming too many computing resources.

Based on the above description and inspiration from [45], we have independently designed an asynchronous self-attention module ASA (see Fig. 1(b)). The input feature can be flexibly represented as SA's text or image features. After the feature obtains the attention weight in the multi-head attention layer, it guides the attention of features X, connects the output result with the residual of the original feature, and then normalizes it with the LayerNorm function to facilitate optimization. After performing the above operations, we feed the processed features into a feed-forward layer and then perform the residual and the normalization operations again, finally outputting the attention features $Z\epsilon\mathbb{R}^{m\times d}$. In GA, $X\epsilon\mathbb{R}^{m\times d_x}$ represents the text features and $Y\epsilon\mathbb{R}^{n\times d_y}$ represents the image features. Different from SA, GA uses text features to guide the attention learning of image features.

The Asynchronous Self-Attention (ASA) we designed is different from SA and GA. $Y, Y_{(i-1)}\epsilon\mathbb{R}^{n\times d_y}$ in ASA represent the same type of features, i.e., text or image features. Different from the synchronous SA ($Q,\ K,\ V$ all come from the same component output), $Y_{(i-1)}\epsilon\mathbb{R}^{n\times d_y}$ in ASA are derived from the output of the previous component, and the $Y_{(i-1)}$, which are closer to the original features than the attended features, retain the relatively rich semantic information of the original features. Taking image features as an example, high-level image features have richer semantic information compared to the underlying features. Applying the above characteristics to multi-head attention can be represented by the following equation:

$$f_Y = Attention\left(Y_i, Y_{(i-1)}, Y_{(i-1)}\right) = softmax\left(\frac{Y_i Y_{(i-1)}^T}{\sqrt{d_y}}\right) Y_{(i-1)} \qquad (4)$$

The more detailed model structure of the combined components will be described in the following sections.

Component Combination. As shown in Fig. 2, we can obtain different model structures by combining the three components in Sect.3.1, in which the text and image features are consistent with those described in Fig. 1. The figure presents the image Asynchronous Self-Attention Model ASAM-I (using asynchronous self-attention components on image features), the text Asynchronous Self-Attention Model ASAM-Q (using asynchronous self-attention components on question features), and the common Asynchronous Self-Attention Model ASAM-QI (applying asynchronous self-attention components on image and question features simultaneously). These three models are all cascade structures, and we provide a detailed description of the multi-modal feature transfer process. Image Asynchronous Self-Attention Model in Fig. 2(a) is our baseline model. The text features $X_{(i)}$ are passed to the next layer after intensive interaction with themselves in SA. In the image processing part, the image features $Y_{(i)}$ are first asynchronous self-attention with the image features $Y_{(i-1)}$ from the previous layer through the ASA component, the result features are guided by the text features X in the GA. The modelling of the image features is completed in the GA to obtain more detailed image features. In contrast to the baseline model described above, text asynchronous self-attention (Fig. 2(b)), as a contrasting model, swaps the component SA with the component ASA in the baseline and uses an asynchronous self-attention approach in the modeling of the text features to simulate the interaction process of the text features in unimodality. The model (Fig. 2(c)) also used as a comparison model, has a structure that combines the two structures described previously, using asynchronous self-attention for both image and text processing, which is guided attention for multi-modal features and interaction between unimodal features at different moments in time.

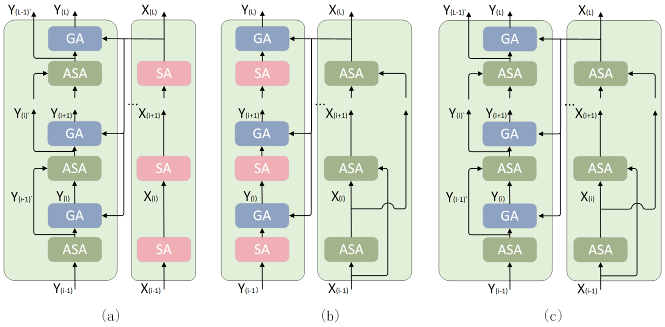

Fig. 2. Three different structures of the model. (a) denotes the image asynchronous self-attention model. (b) is the text asynchronous self-attention model. (c) is the text-image asynchronous self-attention model

3.2. Asynchronous Self-attention Model

The overall model structure of ASAM is shown in Fig. 3. ASAM contains three parts: features representation, model modeling, and features fusion and answer classifier. We improve the model in the interaction modeling part, which will be described in the subsequent sections. The above section had provided the introduction to the basic structure of the model. In this section, we will describe the entire model structure in detail. The entire framework can be specifically represented as three parts. The first is to address the way in which the image and text features that are represented as input, and the second is the separate cascade of the three models in the model section mentioned that models the features of the two modalities in a detailed interactive manner. Finally, for the output image and question features, we use a multi-modal fusion model to fuse the features and feed them to a multi-label classifier for predicting the answer.

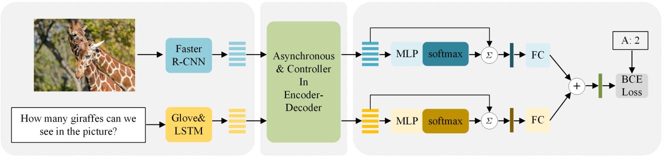
Faster R-CNN
How many giraffes can we see in the picture?
Glove& LSTM
Asynchronous & Controller In Encoder-Decoder
MLP
softmax
Σ
FC
+
A: 2
BCE Loss

Fig. 3. The overall structure of the Asynchronous Self-Attentive Model (ASAM)

Image and Question Representations. The input question is processed first. For a question, we take the sentence into words and trim it to a maximum of 14 words. Next, we use the previous words as input, transform them into a vector using the 300-D GloVe word embedding [38] pre-trained on a large-scale corpus to obtain a words sequence of size $m \times 300$, where the maximum is 14 and the minimum is 1. Finally, we input the word embedding sequence to a single layer 512-dimensional long short-term memory (LSTM) network to obtain the question features $X\epsilon\mathbb{R}^{m\times 512}$.

For image features, we use a Faster R-CNN [36] model with a ResNet-101 as its backbone and pre-trained on the Visual Genome dataset [19] to extract them, we can obtain objects features $Y\epsilon\mathbb{R}^{n\times 2048}$ with a dynamic number of objects,$n\epsilon\left[10,100\right]$.

With the above description, the process of the question and image features extraction can be expressed by the following equations:

$$Y = Faster\_RCNN\left(image\right) \tag{5}$$

$$X = LSTM\left(Glove\left(question\right)\right) \tag{6}$$

The above describes that the images have different number of object regions and the questions also have variable number of words. To facilitate the calculation, we use the zero-padding method to fill the number of image objects and question words to the maximum, which fills the image object regions n to 100 and the number of question words m to 14. In practice, we use a linear transformation to unify the image and question dimensions, transforming the image features to the exact 512 dimensions as the question features.

Collaborative Asynchronous Self-Attention with Controller. In Sect.3.1, we have introduced our models, including the baseline model ASAM-I (Fig. 2(a)) and two comparison models ASAM-Q and ASAM-QI (Fig. 2(b)(c)). This section will focus on describing our baseline model. The asynchronous self-attention model consists of a deep cascade of modules shown in Fig. 2, where the question features X and the image features Y can be described in the deep model as $X_{(L)}$ and $Y_{(L)}$ respectively, where $Y_{(L)}$ represents the intermediate features in a hierarchy that the ASA component has processed. In all models, the first layer will uniformly use the SA component instead of ASA component, because there are no incoming features from the previous layer for the first layer of the model. In addition to the baseline model described above, we have additionally designed a controller as shown in Fig. 4 (the left half of the figure shows the processing of the features within the controller, while the right half adds the controller on the basis as indicated in Fig. 2(a)), which can provide a self-learning parameter for the features and can effectively further improve the model performance. The process can be summarized in the following equations:

$$k_1 = Linear\left(AAP\left(k\right)\right) \tag{7}$$

$$k_2 = Linear\left(ReLU\left(k\right)\right) + Linear\left(ReLU\left(k_1\right)\right) \tag{8}$$

$$k_{parm} = Linear\left(k_2\right) \tag{9}$$

$$k_{fin} = k_{parm}\cdot k + q \tag{10}$$

where AAP() is the AdaptiveAvgPool2d() adaptive average pooling function. We feed the processed K into a linear layer, and the result is activated simultaneously with the unprocessed K subsequently by the $ReLU()$ function. Then we pass them to another linear layer separately and finally add them together. We obtain a parameter by summing the results of the features through a linear layer, multiplying this parameter with the original K and finally, adding it to Q. The input V is treated in the same way as K. With the above processing, we can obtain more detailed features.

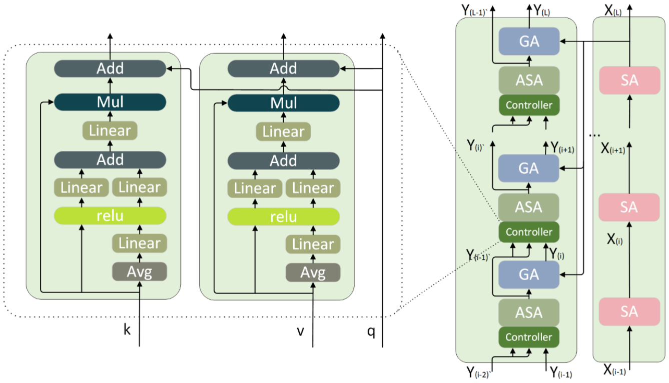

Fig. 4. Asynchronous Self-Attention with Controller

Feature Fusion and answer Classifier. With a deep collaborative asynchronous self-attention model, we obtain question features $X_{(L)}$ and image features $Y_{(L)}$ that contain rich semantic information. To fuse the features of these two modalities, we use an MLP layer consisting of two fully connected layers to transform the multi-modal features. Then the softmax function calculates the attention weight α^X (or α^Y) for the question word features X_i^m (or the image region features Y_j^n). Finally, by multiplying the attention weights with the multi-modal features separately, we can obtain the final question features $\bar{X}$ and image features $\bar{Y}$. Using the question features as an example, the above process can be summarized in the following equations:

$$\alpha^X = softmax\left(MLP\left(X_{(L)}\right)\right) \tag{11}$$

$$\bar{X} = \sum_{i}^{m} \alpha_i^X X_{(L)i} \tag{12}$$

Similarly, we can obtain image features $\bar{Y}$ by using the same method.

Next, we embed features $\bar{X}$ and $\bar{Y}$ into the same dimension and project the multi-modal features onto a vector $F \epsilon \mathbb{R}^c$ by using a linear layer, where c represents the number of features classified in the training set. Finally, the sigmoid activate function is used to obtain the final classification. The formula can be expressed as:

$$f = LN\left(W_X^T \bar{X} + W_Y^T \bar{Y}\right) \tag{13}$$

$$F = Linear\left(f\right) \tag{14}$$

$$A = sigmoid\left(F\right) \tag{15}$$

where W_X^T and W_Y^T denote two linear projection matrices. In the final training process, as similar to the paper [39], we use binary cross-entropy (BCE) as the loss function to train the model.

4. Experimental Results

In this section, we conduct a series of experiments using the collaborative asynchronous self-attention model. Experiments will be conducted on the benchmark VQA v2 dataset to validate the model's performance. We provide a brief description of the parameter settings and conduct extensive ablation experiments on the number of layers of the depth model stack and multiple variants of the model under selected parameters in this section. Then, we show the model validity using attention visualization. Finally, we compare the performance with some of the previous state-of-the-art.

4.1. Implementation Details

Following [45], we set the number of heads h of the multi-head attention mechanism to 8, such a setting makes the originally 512-dimensional intermediate dimension d evenly distributed to the 8 heads, and the dimension of each head is $d_h = 64$. In terms of the number of layers, we set the number of model layers to $L\epsilon \{2, 4, 6\}$ and performed adequate experiments. During training, we set the batch size to 64 and applied the Adam optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.98$. For all model training epochs and learning rate settings, we use 13 training epochs and set the base learning rate to $1e^{-4}$. The model starts training with an initial learning rate of one quarter of the base learning rate, which is gradually increased in subsequent training epochs until the learning rate becomes $1e^{-4}$, and remains constant in the tenth training epoch. Finally, the learning rate decreases in the following training epochs at a rate of 0.2 times every two epochs.

4.2. Datasets

All experiments in this paper are based on the most commonly VQA v2 dataset [11], which uses the MS-COCO dataset as well as question and answer pairs annotated by humans. Compared to the initial version of the VQA dataset, VQA v2 minimizes linguistic bias. In the VQA v2 dataset, each image corresponds to three questions and each question will have ten answers to be answered, the answer chosen most frequently will be considered the correct answer to the question. The entire dataset is divided into three parts: training set (train), validation set (val) and test set (test). Our training process will use both the training and validation sets described above and an additional vg dataset (the additional VQA samples from Visual Genome). After completing the training, the results will be uploaded online for evaluation. The test-dev and test-standard subsets, which divided by test set, will be used for online evaluation, and all training results will be evaluated online with more excellent stability and accuracy than local testing. All test results will be divided into four sections: Yes/No (Y/N), Number, Other and Qverall (All) accuracy.

4.3. Ablation Studies

Model Layers. We first proceeded with several ablation experiments on the layers of the model, which results are depicted in Fig. 5 and Table 1. The model's validity is discussed in detail in the light of the results.

As shown in Fig. 5, with the number of layer settings increasing, the performance of all three different models steadily increases in the other three test criteria (Yes/No, Number and Overall accuracy), except for the ASAM-Q model which decreases after 4 layers in terms of other. This verifies that the use of asynchronous self-attention components is effective for modeling self-attention as the depth of the model increases. It is also not hard to notice that when setting the number of layers to 2, the effect of using asynchronous self-attention on the question self-attention modeling (ASAM-Q) are higher than the other two models. The analysis shows that the word sequences in the question features are only 14-dimensional space, compared to the 100 object dimensions in the image features, and the question features can be focused on the correct words more quickly with the asynchronous self-attention component. However, as the depth of the model increases and the attention on the image continues to be refined, the accuracy of both the ASAM-QI and ASAM-I models gradually approaches or even surpasses that of the ASAM-Q model after 4 layers. In the case of the transition from 2 to 4 layers, the model performance improves rapidly and slows down when the transition from 4 to 6 layers, which also indicates a gradual saturation of the model performance as the depth increases. Therefore, we set our baseline model layer parameter to 6 layers, which saves the time cost of model training and also controls the number of parameters in the model.

Based on 6 layers, the accuracy of each model is presented in Table 1 for comparison with the MCAN model. The experimental results indicate that the ASAM outperforms the MCAN on all other problems except in the Number type. Specifically, it improves the model performance and does not increase the number of model parameters after using the asynchronous self-attention component.

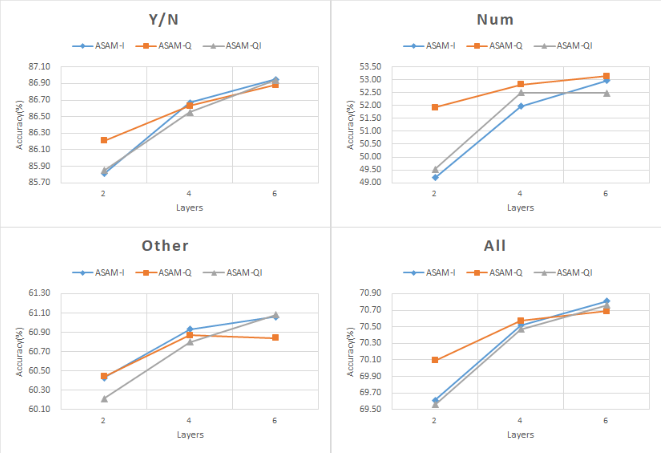

Fig. 5. Results of ablation on the ASAM-I, ASAM-QI, and ASAM-Q model layers

Model Variation. Table 2 and Fig. 6 present the results of the proposed ASAM model with a controller. As depicted in Table 2, incorporating a controller results in a further

Table 1. Accuracy of ASAM-Q, ASAM-QI and ASAM-I with six model layers on the test-dev of VQA v2

Model	Y/N	Number	Other	All
MCAN	86.82	53.26	60.72	70.63
ASAM-Q	86.88	53.15	60.84	70.69
ASAM-QI	86.94	52.48	61.08	70.76
ASAM-I	86.95	52.97	61.06	70.81

augmentation of the model's performance across various metrics, substantiating the effectiveness of the controller design. This observation highlights that the collaborative interaction between the asynchronous self-attention model and the controller can unlock more significant potential.

The performance of the basic model is compared with the performance of the model using the controller as shown in Fig. 6 (The blue part is the original model and the orange part is the model with the addition of the controller). Experimental results show that adding a controller is superior to the model without the controller in some of the test types as well as in overall accuracy, which further demonstrates that the controller we have designed optimizes the features to some extent and allows the model to locate the target location more accurately when converging on the attention. In addition, the figure also depicts the trend lines for the original model and the model with the added controller under different test contents. The results of the original model show an upward trend under all test contents except for a slight downward trend in the trend line for the original model under Y/N. This is one reason why we ended up using the model with the asynchronous self-attention component alone on the image as our baseline model.

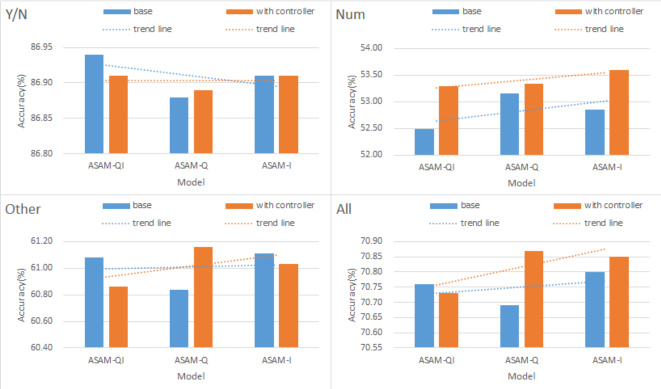

Fig. 6. Comparison between the original model and the model with the Controller

Table 2. Accuracy of ASAM-Q, ASAM-QI and ASAM-I model with controller on the test-dev and test-std set of VQA v2

	Test-dev Test-dev				**Test-std**
Model	**Y/N**	**Number**	**Other**	**All**	**All**
ASAM-QI(c)	86.91	53.29	60.86	70.73	70.96
ASAM-Q(c)	86.89	53.34	61.16	70.87	71.30
ASAM-I(c)	86.91	52.59	61.03	70.85	71.20

4.4. Comparison with State-of-the-Art

Table 3 lists the experimental results of the ASAM and the other state-of-the-art model on the benchmark dataset VQA v2. BUTD [1] used a bottom-up visual feature of attention and was the winning model in the 2017 VQA Challenge. In addition to focus on where to look, HieCoAttVQA [27] also focuses on what words to listen for, and the model was able to make joint inferences about attention to images and questions. Compared to other bilinear pooling methods, MFH [46] used generalized higher-order models to capture the more complex interactions between multi-modal features. BAN [18] effectively extends a single attention network using a bilinear attention mapping, and also keeps the computational cost constant while taking into account each pair of multimode input channels. Peng et al. [33] devised a self-guided word relation attention scheme and two problem-adaptive visual relation attention modules to explore the semantic latent relationships between words and extract precise binary relationships between objects. Qin et al. [34] propose a Residual Weight-Sharing Attention Network (RWSAN). By using a method that dynamically fuses multi-modal features with intra-modal and inter-modal information flow, DFAF [10] achieved high-level interaction between visual and language modalities. To understand the visual scene in an image, ReGAT [22] encoded the image as a graph and used the graph attention mechanism to model the objects in the graph with multiple types of relationships. Re-atten [12] reconstructs attention based on answer re-attention, which allows the model to re-learn visual objects in the image. Guo et al. [14] set thresholds for attention score to filter out the text or image features, to choose the most relevant information for predicting the correct answer and avoid the distract of unrelated question or image areas. MCAN [45] proposes a deep-modular co-attention network to address the issue of insufficient deep interactions in models, winning the 2019 VQA Challenge. MESAN [13] proposes top-k-based filter method of attention scores. From the experimental results, it is evident that the proposed method outperforms current state-of-the-art models across most metrics, except for a slight decrement in performance for Yes/No type questions. Compared to MCAN, ASAM also applied an encoder-decoder architecture and used an asynchronous self-attention (with controller) model on the images, with an overall accuracy 0.22 points higher than MCAN on test-dev and 0.3 points higher on test-std, with significant accuracy improvements on other validation metrics. This validates the superiority of our model for Visual Question Answering with leading performance.

Table 3. Accuracies of the model proposed in this paper on the Visual Question Answering Dataset VQA v2 to compare with the state-of-the-art methods

	Test-dev				Test-std
Model	**Y/N**	**Number**	**Other**	**All**	**All**
Bottom-up[1]	81.82	44.21	56.05	65.32	65.67
HieCoAttVQA[27]	79.70	40.00	59.80	65.80	66.10
MFH[46]	84.27	49.56	59.89	68.76	-
BAN[18]	85.31	50.93	60.26	69.52	-
MRA-Net[33]	85.58	48.92	59.46	69.02	69.46
RWSAN[34]	86.45	52.18	60.38	70.19	-
DFAF[10]	86.09	53.32	60.49	70.22	70.34
ReGAT[22]	86.08	54.42	60.33	70.27	70.58
Re-attn[12]	87.00	53.06	60.19	70.43	70.72
MCAN[45]	86.82	53.26	60.72	70.63	70.90
SCAVQA-I[14]	87.00	53.31	60.83	70.76	71.09
MESAN[13]	87.05	53.21	60.72	70.71	71.08
ASAM-QI	86.94	52.48	61.08	70.76	-
ASAM-Q	86.88	53.15	60.84	70.69	-
ASAM-I	86.95	52.97	61.06	70.81	-
ASAM-QI(c)	86.91	53.29	60.86	70.73	70.96
ASAM-Q(c)	86.89	53.34	61.16	70.87	71.30
ASAM-I(c)	86.91	52.59	61.03	70.85	71.20

4.5. Attention Visualization

Fig. 7 compares the visualization results of the MCAN with ASAM (The darker-colored parts are the model's attentional focus). In the Visual Question Answering task, the asynchronous self-attention mechanism proposed in this paper no longer singularly pursues fine-grained but takes a broader view of the image and retains as much information as possible. Each of the three examples in Fig. 7 combines the real situation, the situation predicted by the MCAN model, and the situation predicted by our model. The more highlighted areas in the image regions indicate that the model is focusing more attention on them. Based on the images, it is apparent that the model proposed in this paper focuses on objects from multiple perspectives, pinpointing the key objects and distracting some attention from other relevant objects. In the first instance, the man is playing a sport, and the model focuses not only on the man himself and the surfboard beneath his feet, but also the waves and ultimately get a valid answer. For the counting question in the example in the third column, our model covers all objects more comprehensively than the fine-granularity of the comparison model, and thus correctly answers the given question. It is not difficult to find that our model can always focus on more information, but this also implies another problem, which is information interference. When faced with problems requiring precise targeting of attention, MCAN reduces the interference of redundant information, which has more advantages.

5. Concluding Rewarks

This paper proposed an Asynchronous Self-Attention Model (ASAM) and a controller, balancing coarse-grained and fine-grained attention within the attention model for VQA. Its component can coordinate the relationship between the upper and lower attention layers and the VQA model using this component. The controller can retain other related object information as much as possible without reducing the attention on key objects. Experimental results and ablation studies on the benchmark dataset VQA v2 validate the effectiveness of the proposed model and demonstrate the effectiveness of coordinating the upper and lower attention layers to improve model performance. However, the choice of feature matching fusion method still needs to be continuously explored, and this is a direction for our future research. In future research, we will aim to explore more effective models and apply them to the field of Visual Question Answering to assist the models in better understanding visual and question features and to advance further the research work related to Visual Question Answering.

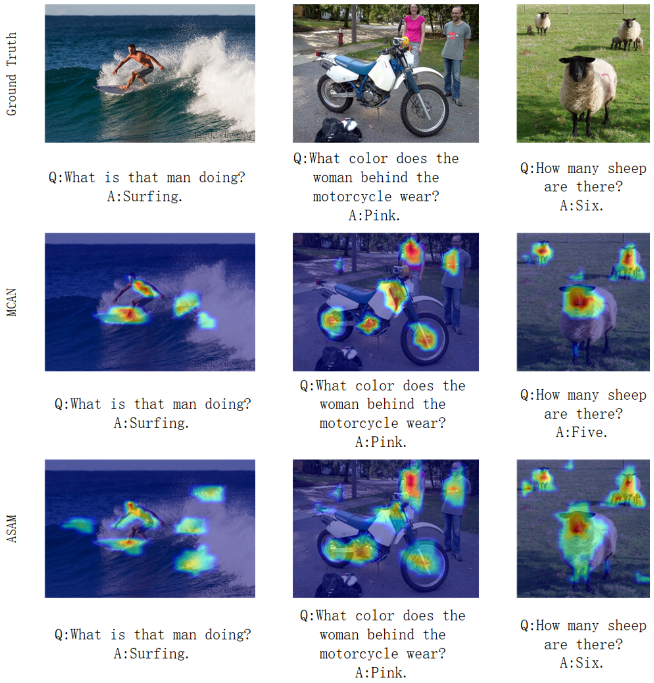
Ground Truth
Q:What is that man doing?
A:Surfing.
Q:What color does the woman behind the motorcycle wear?
A:Pink.
Q:How many sheep are there?
A:Six.
MCAN
Q:What is that man doing?
A:Surfing.
Q:What color does the woman behind the motorcycle wear?
A:Pink.
Q:How many sheep are there?
A:Five.
ASAM
Q:What is that man doing?
A:Surfing.
Q:What color does the woman behind the motorcycle wear?
A:Pink.
Q:How many sheep are there?
A:Six.

Fig. 7. Attention visualization result. The first row is the input images, questions and ground truth answers. The second is the baseline model MCAN. The third row is the proposed ASAM model

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (Grant No. 52331012) and the Natural Science Foundation of Shanghai under Grant 21ZR1426500.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)
2. Chen, C., Han, D., Chang, C.C.: Caan: Context-aware attention network for visual question answering. Pattern Recognition 132, 108980 (2022)
3. Chen, C., Han, D., Chang, C.: MPCCT: multimodal vision-language learning paradigm with context-based compact transformer. Pattern Recognit. 147, 110084 (2024)
4. Chen, C., Han, D., Shen, X.: CLVIN: complete language-vision interaction network for visual question answering. Knowl. Based Syst. 275, 110706 (2023)
5. Diao, C., Zhang, D., Liang, W., Li, K.C., Hong, Y., Gaudiot, J.L.: A novel spatial-temporal multi-scale alignment graph neural network security model for vehicles prediction. IEEE Transactions on Intelligent Transportation Systems (2022)
6. Diao, C., Zhang, D., Liang, W., Li, K., Hong, Y., Gaudiot, J.: A novel spatial-temporal multi-scale alignment graph neural network security model for vehicles prediction. IEEE Trans. Intell. Transp. Syst. 24(1), 904–914 (2023)
7. Ding, Y., Yu, J., Liu, B., Hu, Y., Cui, M., Wu, Q.: 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 (2022)
8. Fan, Y., Xu, B., Zhang, L., Song, J., Zomaya, A.Y., Li, K.: Validating the integrity of convolutional neural network predictions based on zero-knowledge proof. Inf. Sci. 625, 125–140 (2023)
9. Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., Xu, W.: Are you talking to a machine? dataset and methods for multilingual image question. Advances in neural information processing systems 28 (2015)
10. Gao, P., Jiang, Z., You, H., Lu, P., Hoi, S.C., Wang, X., Li, H.: Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6639–6648 (2019)
11. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6904–6913 (2017)
12. Guo, W., Zhang, Y., Yang, J., Yuan, X.: Re-attention for visual question answering. IEEE Trans. Image Process. 30, 6730–6743 (2021)
13. Guo, Z., Han, D.: Multi-modal explicit sparse attention networks for visual question answering. Sensors 20(23), 6758 (2020)
14. Guo, Z., Han, D.: Sparse co-attention visual question answering networks based on thresholds. Applied Intelligence 53(1), 586–600 (2023)
15. Guo, Z., Han, D., Massetto, F.I., Li, K.C.: Double-layer affective visual question answering network. Computer Science and Information Systems 18(1), 155–168 (2021)
16. Han, D., Zhou, S., Li, K.C., de Mello, R.F.: Cross-modality co-attention networks for visual question answering. Soft Computing 25, 5411–5421 (2021)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 770–778. IEEE Computer Society (2016)

18. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear attention networks. Advances in neural information processing systems 31 (2018)
19. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision 123, 32–73 (2017)
20. Li, H., Han, D., Chen, C., Chang, C., Li, K., Li, D.: A visual question answering network merging high- and low-level semantic information. IEICE Trans. Inf. Syst. 106(5), 581–589 (2023)
21. Li, J., Han, D., Wu, Z., Wang, J., Li, K., Castiglione, A.: A novel system for medical equipment supply chain traceability based on alliance chain and attribute and role access control. Future Gener. Comput. Syst. 142, 195–211 (2023)
22. Li, L., Gan, Z., Cheng, Y., Liu, J.: Relation-aware graph attention network for visual question answering. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10313–10322 (2019)
23. Li, S., Gong, C., Zhu, Y., Luo, C., Hong, Y., Lv, X.: Context-aware multi-level question embedding fusion for visual question answering. Inf. Fusion 102, 102000 (2024)
24. Liang, W., Yang, Y., Yang, C., Hu, Y., Xie, S., Li, K., Cao, J.: Pdpchain: A consortium blockchain-based privacy protection scheme for personal data. IEEE Trans. Reliab. 72(2), 586–598 (2023)
25. Lin, W., Chen, J., Mei, J., Coca, A., Byrne, B.: Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023)
26. Long, J., Liang, W., Li, K.C., Wei, Y., Marino, M.D.: A regularized cross-layer ladder network for intrusion detection in industrial internet of things. IEEE Transactions on Industrial Informatics 19(2), 1747–1755 (2022)
27. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. Advances in neural information processing systems 29 (2016)
28. Ma, F., Zhou, Y., Rao, F., Zhang, Y., Sun, X.: Image captioning with multi-context synthetic data. In: Wooldridge, M.J., Dy, J.G., Natarajan, S. (eds.) Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver,Canada. pp. 4089–4097. AAAI Press (2024)
29. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: A neural-based approach to answering questions about images. In: Proceedings of the IEEE international conference on computer vision. pp. 1–9 (2015)
30. Mao, A., Yang, Z., Lin, K., Xuan, J., Liu, Y.J.: Positional attention guided transformer-like architecture for visual question answering. IEEE Transactions on Multimedia (2022)
31. Nam, H., Ha, J.W., Kim, J.: Dual attention networks for multimodal reasoning and matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 299–307 (2017)
32. Nguyen, B.X., Do, T., Tran, H., Tjiputra, E., Tran, Q.D., Nguyen, A.: Coarse-to-fine reasoning for visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4558–4566 (2022)
33. Peng, L., Yang, Y., Wang, Z., Huang, Z., Shen, H.T.: Mra-net: Improving vqa via multi-modal relation attention network. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(1), 318–329 (2020)
34. Qin, B., Hu, H., Zhuang, Y.: Deep residual weight-sharing attention network with low-rank attention for visual question answering. IEEE Transactions on Multimedia (2022)

35. Ren, M., Kiros, R., Zemel, R.: Exploring models and data for image question answering. Advances in neural information processing systems 28 (2015)
36. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28 (2015)
37. Shen, X., Han, D., Zong, L., Guo, Z., Hua, J.: Relational reasoning and adaptive fusion for visual question answering. Appl. Intell. 54(6), 5062–5080 (2024)
38. Sturman, D.J., Zeltzer, D.: A survey of glove-based input. IEEE Computer graphics and Applications 14(1), 30–39 (1994)
39. Teney, D., Anderson, P., He, X., Van Den Hengel, A.: 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 (2018)
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 5998–6008 (2017)
41. Wang, Y., Yasunaga, M., Ren, H., Wada, S., Leskovec, J.: Vqa-gnn: Reasoning with multimodal knowledge via graph neural networks for visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21582–21592 (2023)
42. Xia, H., Lan, R., Li, H., Song, S.: ST-VQA: shrinkage transformer with accurate alignment for visual question answering. Appl. Intell. 53(18), 20967–20978 (2023)
43. Yan, S., andWeifeng Chen, M.B., Zhou, X., Huang, Q., Li, L.E.: Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling. CoRR abs/2402.06118 (2024)
44. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 21–29 (2016)
45. Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 6281–6290. Computer Vision Foundation / IEEE (2019)
46. Yu, Z., Yu, J., Xiang, C., Fan, J., Tao, D.: Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE transactions on neural networks and learning systems 29(12), 5947–5959 (2018)
47. Zheng, W., Yin, L., Chen, X., Ma, Z., Liu, S., Yang, B.: Knowledge base graph embedding module design for visual question answering model. Pattern recognition 120, 108153 (2021)
48. Zhou, Y., Ren, T., Zhu, C., Sun, X., Liu, J., Ding, X., Xu, M., Ji, R.: TRAR: routing the attention spans in transformer for visual question answering. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 2054–2064. IEEE (2021)

Han Liu received a B.S. degree in computer networking from the Shenyang Normal University, Shenyang, China, in 2013, and the M.S. and Ph.D. degrees in computer software & theory and information management & information systems from the Shanghai Maritime University, Shanghai, China, in 2016 and 2022, respectively. He is a Postdoctoral Fellow with the Department of Logistics, Shanghai Maritime University, Pudong, China, in 2022. His main research interests include blockchain, IoT, cloud security, and machine learning.

Dezhi Han received a B.S. degree in applied physics from the Hefei University of Technology, Hefei, China, in 1990, and the M.S. and Ph.D. degrees in computing science from the Huazhong University of Science and Technology, Wuhan, China, in 2001 and 2005, respectively. He is a Professor with the Department of Computer, Shanghai Maritime University, Pudong, China, in 2006. His research interests include cloud and outsourcing security, wireless communication security, network and information security, and visual question answering.

Shukai Zhang received an M.S. from the School of Information Engineering at Shanghai Maritime University, China. His research interest is visual question answering.

Jingya Shi is pursuing an M.S. at the School of Information Engineering at Shanghai Maritime University, China. Her current research interest is visual question answering.

Huafeng Wu received a Ph.D. degree in computer science from Fudan University in 2009, a master's degree in traffic information engineering and control from Dalian Maritime University in 2004, conducted Postdoctoral Research at Carnegie Mellon University from 2008 to 2009, and also a Visiting Scholar with Shanghai Jiao Tong University from 2012 to 2013. He is currently a Professor and a Ph.D. Supervisor with the Merchant Marine College, Shanghai Maritime University. His research interests include the Internet of Shipping, the Internet of Things, and wireless sensor networks. He also serves as an Editorial Board Member for the Computer Communications journal.

Yachao Zhou received a B.S. degree from Beijing University of Posts and Telecommunications, an M.S. from Tsinghua University, Beijing, China, and a Ph.D. from Dublin City University, Dublin, Ireland. She is the chief scientist of Technology Co., Ltd. Her specific interests include big data and cloud computing, cybersecurity, and deep packet inspection.

Kuan-Ching Li is a Life Distinguished Professor at Providence University, Taiwan. He is a recipient of distinguished and chair professorships from universities in several countries and awards and funding support from a number of agencies and high-tech companies. Besides publishing numerous journal articles, book chapters, and refereed conference papers, he is co-author/co-editor of more than 50 technical professional books published by CRC Press/Taylor & Francis, Springer, and McGraw-Hill. His research interests include parallel and distributed computing, Big Data, and emerging technologies. He is a senior member of the IEEE and a fellow of the IET.

Received: March 21, 2024; Accepted: September 17, 2024.