The Visual Computer (2022) 38:3097–3108
https://doi.org/10.1007/s00371-022-02524-z

ORIGINAL ARTICLE

SPCA-Net: a based on spatial position relationship co-attention network for visual question answering

Feng Yan[1] · Wushouer Silamu[2] · Yanbin Li[1] · Yachuang Chai[1]

Accepted: 10 May 2022 / Published online: 16 June 2022

Abstract

Recently, the latest method of VQA (visual question answering) mainly relies on the co-attention to link each visual object with the text object, which can achieve a rough interaction between multiple models. However, VQA models tend to focus on the association between visual and language features without considering the spatial relationship between image region features extracted by Faster R-CNN. This paper proposes an effective deep co-attention network to solve this problem. As a first step, BERT was introduced in order to better capture the relationship between words and make the extracted text feature more robust; secondly, a multimodal co-attention based on spatial location relationship was proposed in order to realize fine-grained interactions between question and image. It consists of three basic components: the text self-attention unit, the image self-attention unit, and the question-guided-attention unit. The self-attention mechanism of image visual features integrates information about the spatial position and width/height of the image area after obtaining attention so that each image area is aware of the relative location and size of other areas. Our experiment results indicate that our model is significantly better than other existing models.

Keywords BERT · Guided-attention · Self-attention · Faster R-CNN · Spatial position relationship

1 Introduction

VQA is a multimodal learning task that explores how to answer a question from a picture. It is more challenging and difficult to perform the VQA task than other multi-modal learning tasks [1–3], which require capturing nuances in the interactions between images and questions.

In recent years, work has focused mainly on improving the performance of the VQA model from three perspectives. Firstly, VQA can be improved by extracting the better images and questions features. The development and breakthroughs in neural network models such as ResNet [4], FishNet [5], and BUTD [6] have made it possible to extract better features from images. The second is multi-modal fusion, such as Bilinear Fusion [7], MCB [8], etc. Finally, it is the attention mechanism that predicts the answer by adapting which image regions and words to focus on. This paper focuses on the third aspect to improve the VQA task.

The early attention models learned coarse interactions, ignoring the interactions between every word in the question and every region of the image. Consequently, we cannot infer whether a question word correlates to the region of an image. This suggests that early co-attention models have many limitations.

DCN [9] and BAN [10] are examples of dense co-attention models, the two models can be stacked in depth. The results indicate that their deep models are much better than shallow models. However, these models do not consider the intramodal attention mechanism.

Therefore, two deep intra–inter attention models (DFAF [11] and MLIN [12]) and two co-attention models (MCAN [13] and MUAN [14]) have been introduced which are all inspired by the Transformer model [15]. They all use deep co-attention and achieve good results, but the model ignores the spatial position relationship and size of the regions within the images, use only vision features based on Faster-RCNN [16], and neglect to add the position information and size information of the image regions.

✉ Feng Yan
yanfeng@stu.xju.edu.cn

[1] School of Information Science and Engineering, Xinjiang University, Urumqi, China

[2] Laboratory of Multi-Lingual Information Technology of Xinjiang, Urumqi, China

Besides the image visual feature, the text feature of the question is also very important. For the question features, the existing models [17–19] rely on static word vectors. Different words have different meanings in different contexts in the natural language environment. The static word vectors cannot capture these differences effectively, so semantic and grammatical deviations may occur. When encoding text features, we try to use a multi-header mechanism and increase the depth of the text feature model. BERT [20] (Bidirectional Encoder Representations from Transformers) meets the above two assumptions and uses the self-supervised learning method on the basis of a massive corpus to provide a good feature representation for text. We use the dynamic word vector based on BERT to overcome these limitations. The contribution and innovation of the paper can be summed up as follows:

- We designed a deep cross-modality co-attention model, which includes three basic units: text self-attention unit, image self-attention unit, and question-guided-attention unit. Our design of the self-attention mechanism of image visual features integrates the spatial position, width, and height information of the image region in order to allow each image region to perceive the relative position and size information of other regions.
- In this paper, we encode question features using the dynamic word vector of BERT, which uses a multi-header mechanism to increase the depth of the text feature model, so that we can obtain more comprehensive question features.
- Our model achieved good performance on two VQA datasets. Extensive ablation experiments show that each module can play its part.

2 Related work

The VQA task involves both visual and text feature extraction as well as the fusion of multimodal features, which is a challenging task. Currently, the models for solving visual question answering can be categorized into four directions: the method based on attention mechanism, the method based on fusion, and the model based on high-level attributes and knowledge.

2.1 Method based on attention mechanism

As a result of the attention mechanism, humans are able to focus on important information and filter out irrelevant information. Researchers use a semantic expression as a query in question-guided visual attention to identify the region of the image that corresponds to the answer [17–19,21]. There are two methodologies for calculating soft attention scores in VQA: using concatenated question features and each region features of the image as inputs for a multi-layer perceptron, and using the dot product of two inputs.

2.2 Method based on fusion

An understanding of the image and question content and their relationship is necessary to successfully complete the VQA task. The multimodal features of VQA are merged using joint representation methods in an early stage of the process. A bilinear pooling-based feature fusion method [8] was applied to VQA to obtain high-level interaction functions between two modals. The MLB [22] method was later developed and shows comparable performance to the MCB [8], but with fewer weighted parameters.

2.3 Method based on pre-training model

Pre-training language models [23–26] have emerged as a new research hotspot to learn the joint representation of image and text through large-scale training [25] transforms the BERT structure into a model supporting two modal inputs and integrates visual and text features through an attention model. The author uses a conceptual caption dataset for pre-training and then applies the pre-training model to multiple visible text tasks. The pre-training model achieves better results in multiple tasks than the previous model for a single task. Pre-training tasks include text mask modeling, mask prediction, cross-modal alignment, and image question answering. The model has performed well with both VQA [27] and GQA [28] datasets after fine-tuning the pre-training model.

Models need to be pre-trained on large datasets, so they require more computing power and human resources. Such tasks are challenging for a small team.

2.4 High-level attributes and knowledge

References [29–32] relates to visual question answering based on information from an external knowledge base. The model can have 'common sense' like humans through the knowledge base. In order to answer the question 'what kind of flowers are in the picture?' the model needs to know what kinds of flowers there are. This method greatly improves the generalizability of the model.

3 Proposed model

In the section, we have improved on the basis of MCAN, integrated spatial location information on the visual features, and used BERT instead of glove to obtain better text features. An overview of the proposed model is shown in Fig. 1.

3.1 Question encoding

In the past, the text feature of the VQA model was based upon obtaining a static word vector from a corpus. This means that each word corresponds to a specific vector. The polysemy of English words is common. That is, the meaning of the same word varies depending on the context. Researchers introduced the dynamic word vector to solve polysemy, and BERT [20] is one of its representatives. BERT model uses a self-supervised learning technique on the basis of a massive corpus to provide a good feature representation for text, thus solving polysemy.

All the questions are cut to a maximum of 14 words, and the redundant words will be discarded. Because only 0.25% of the questions in the dataset have more than 14 words, in order to improve the computational efficiency. For a given question, text feature can be represented as follows:

$$X = \mathrm{BERT}(\mathrm{Ques}) \tag{1}$$

where $X \in \mathbb{R}^{d_x * M}$ is the question representations. d_x is the output dimension of the question representations.

3.2 Image feature embedding based on Faster R-CNN

Based on the Bottom-Up Attention [6], we pretrain Faster-RCNN on the large public dataset Visual Genome [33], and then use Faster-RCNN to extract the vision features of the image. The features of the image can be expressed as:

$$Y = \mathrm{FRCNN}(\mathrm{image}) \tag{2}$$

where $Y \in \mathbb{R}^{d_y * N}$ is the vision feature, $d_y = 2048$ is the dimension of each region feature and $N \in [10, 100]$, N is the number of objects detected in the image. Generally, a threshold will be set, and the objects exceeding this threshold will be selected.

When using Faster-RCNN to get the image region features, we obtain object's relative bounding box coordinates $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$ at the same time.

3.3 Co-attention mechanism based on spatial location relationship

3.3.1 Overview of the architecture

Our co-attention mechanism includes question self-attention mechanism (Q-SA), image vision self-attention based on spatial position relationship (V-SP-SA), and question guided-attention to image (GA). We use the multi-head attention mechanism of Transformer to realize the self-attention and the guided-attention, so we first introduce the multi-head attention mechanism. Then, we introduce the above three attention mechanisms units. Finally, we explain how to use the above three basic attention units for cascade combination. Figure 2 shows the overview of the three basic attention units.

3.3.2 Multi-head attention

Because VQA tasks require processing multi-modal data (images and text), they require more efficient calculation than tasks that only require processing unimodal data. At the same time, because the input image of the VQA task is highly correlated with the question text, the interaction between the data of the two modals helps to improve the accuracy of the results.

Suppose the query matrix $Q = \{q_1, q_2, \ldots, q_m\}$, where the query vector $q_i \in \mathbb{R}^{1*d}$, the key matrix $K = k_1, k_2, \ldots, k_m$, where the key vector $k_i \in \mathbb{R}^{1*d}$, the value matrix $V = \{v_1, v_2, \ldots, v_m\}$, where the query vector $v_i \in \mathbb{R}^{1*d}$. Then, the attention features are calculated as follows:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V \tag{3}$$

Next, we introduce the multi-head attention mechanism, which helps to further improve feature expression. Firstly, input Q, K, and V into h linear layers to obtain (Q_i, K_i, V_i), where $i = 1, 2, \ldots, h$ and then calculate the Attention(Q_i, K_i, V_i), where $i = 1, 2, \ldots, h$, and finally splice h attention features and obtain final output features by a linear layer. Multi-head attention is calculated as follows:

$$\begin{cases} \mathrm{head}_i = \mathrm{Attention}(Q_i, K_i, V_i) \\ \mathrm{MHA}(Q, K, V) = \mathrm{concat}(\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head}_h) W_o \end{cases} \tag{4}$$

where $W_o^{h*d_h*d}$ are the projection matrices, and d_h is the dimensionality of the output features from each head.

3.3.3 Self-attention and guided-attention unit

As shown in Fig. 2, we will introduce question self-attention (Q-SA), vision spatial position self-attention (V-SP-SA), and question guided-attention to images (GA).

$X = \{x_1; x_2; \ldots; x_M\} \in \mathbb{R}^{M*d_x}$ is the input features of Q-SA; the calculation formula of Q-SA is as follows:

$$Q\_\mathrm{SA}(X) = \mathrm{MHA}(X, X, X) \tag{5}$$

In the feedforward layer, it consists of two full connection layers. After the first full connection layer, we use the Relu activation function, then use dropout transformation, and finally use the second full connection layer. The output

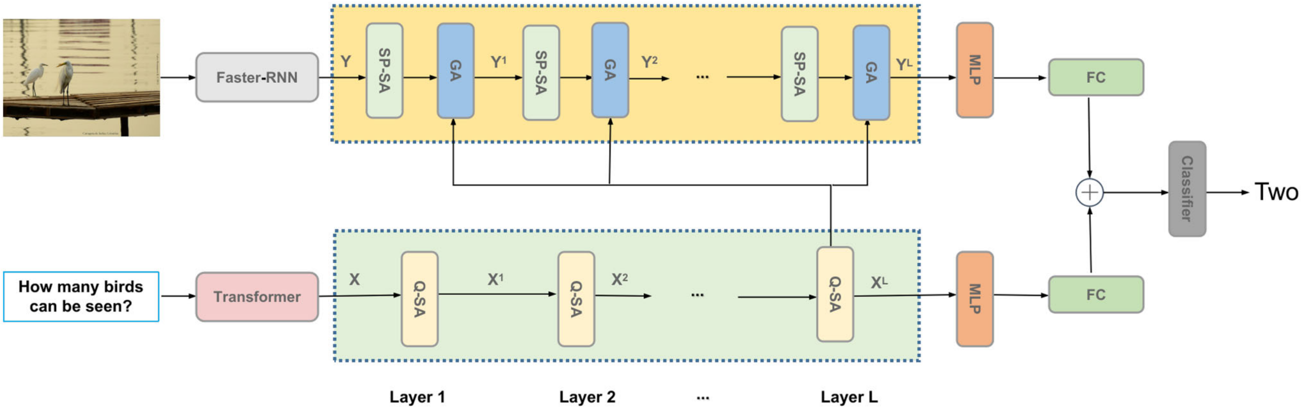

Fig. 1 Overview of the proposed model, which includes three basic units: text self-attention unit (Q-SA), image vision self-attention unit based on spatial position relationship (SP-SA), and question-guided-attention unit (GA)

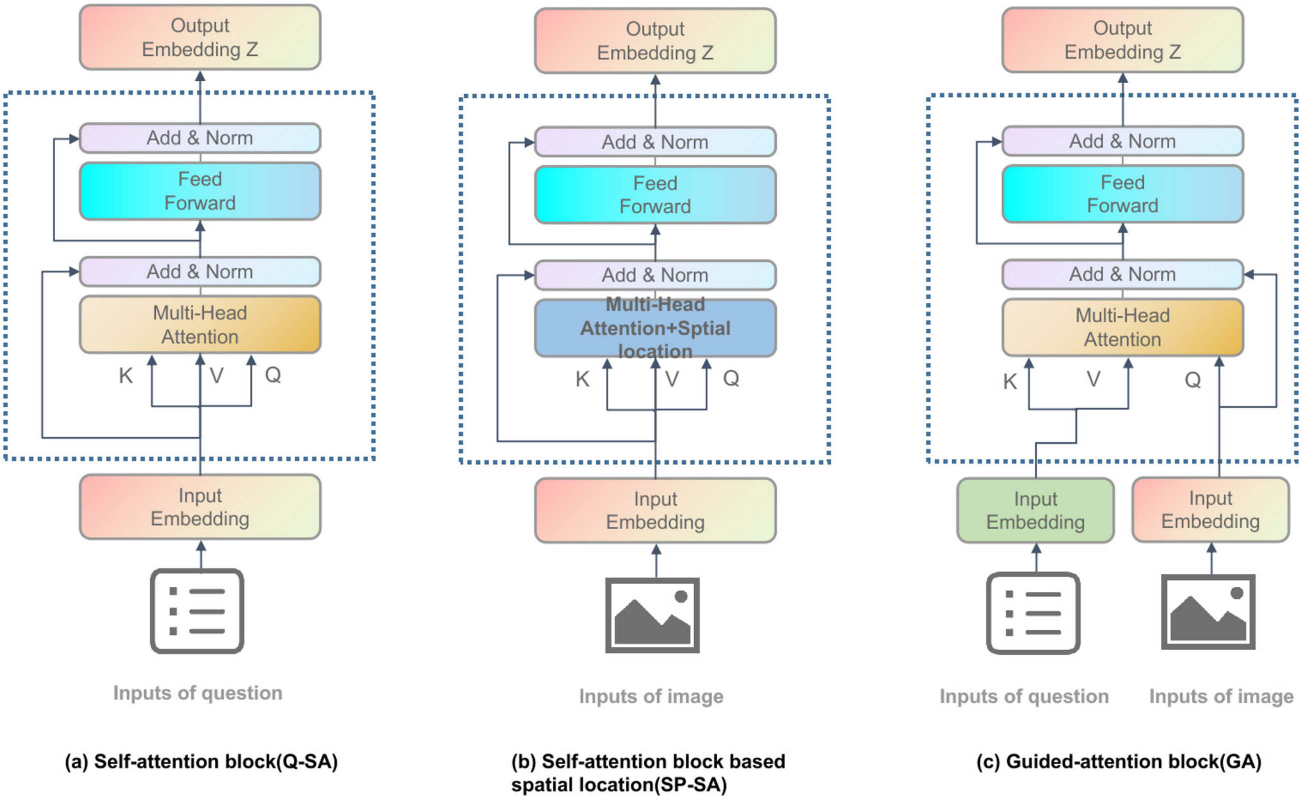

Fig. 2 Architecture of the three basic attention units

features are then processed by residual connection [4] and layer normalization. The formula is as follows:

$$\mathrm{FFN}(U) = \mathrm{FC}_d \circ \mathrm{Drop} \circ \mathrm{Relu} \circ \mathrm{FC}_{4d}(U) \tag{6}$$

where FC() is a fully connected layer, Drop() is a dropout layer. The symbol ∘ is a composition of two layers.

As shown in Fig. 2c, the GA unit has two input features $X = [x_1; x_2; \ldots; x_M] \in \mathbb{R}^{M*d_x}$ and $Y = [y_1; y_2; \ldots; y_N] \in \mathbb{R}^{N*d_y}$, where Y guides the attention learning for X. The calculation formula of the GA unit models is as follows:

$$\mathrm{GA}(X, Y) = \mathrm{MHA}(X, Y, Y) \tag{7}$$

3.3.4 Self-attention based on spatial position relationship

After we obtained object's relative bounding box coordinates $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$, we further process these information to obtain the center coordinates, width, and height of the image, as well as the size and position information of each area object relative to other area objects. As shown in Fig. 2b, when processing image features, we add spatial information about the objects in the image.

Firstly, calculate the center position coordinates, width, and height of each area object. Then, the center coordinates of every object are subtracted by each to get the relative distance between each object so that each object can perceive the relative position information. Then, we need to divide the abscissa and ordinate by the width and height of the area object, respectively, for normalization, so as to remove the common parts and highlight the individual differences. Finally, the logarithm of each data will not change the nature and relationship of the data, and the heteroscedasticity problem will be eliminated. The calculation formula is as follows:

$$\begin{cases} \bar{x} = (x_{\min} + x_{\max}) * 0.5 \\ \text{delta\_}x = \log\left(\dfrac{\bar{x} - \bar{x}^T}{w}\right) \\ \bar{y} = (y_{\min} + y_{\max}) * 0.5 \\ \text{delta\_}y = \log\left(\dfrac{\bar{y} - \bar{y}^T}{h}\right) \end{cases} \tag{8}$$

In the same way, we divide the width and height of each object to get the relative size information between each object, so that each object can perceive each other's relative size information. Finally, take the logarithm of data. The calculation formula is as follows:

$$\begin{cases} \text{delta\_}w = \log\left(\dfrac{w}{w^T}\right) \\ \text{delta\_}h = \log\left(\dfrac{h}{h^T}\right) \end{cases} \tag{9}$$

Then, the above four features are spliced together, and the formula is as follows:

$$R = \text{Concat}(\text{delta\_}x, \text{delta\_}y, \text{delta\_}w, \text{delta\_}h) \tag{10}$$

Finally, we use two fully connected layers for further processing. Each fully connected layer is followed by a Relu activation, and the formula is as follows:

$$r = \text{MLP}(R) = \text{Relu} \circ \text{FC}_{d_h}^{d_r} \circ \text{Relu} \circ \text{FC}_{d_r}^{4}(R) \tag{11}$$

where $\text{FC}_{d_r}^{4}()$, $\text{FC}_{d_h}^{d_r}$ are fully connected layers and $\text{FC}_{d_r}^{4}()$ indicate the dimension is converted from 4 to d_r, the sysbol $\circ$ is a composition of two layers.

After got the processed spatial information, we integrate it into the self-attention unit. $Y = \{y_1; y_2; \ldots; y_N\} \in \mathbb{R}^{N*d_y}$ is the input features of SPSA, the calculation formula of SPSA is as follows:

$$\begin{cases} \text{SP\_Attention}(Q, K, V) = \left(\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) + r\right) V \\ \text{SPSA} = \text{SP\_MHA}(Y, Y, Y) \end{cases} \tag{12}$$

where attention unit of SP_MHA is different from MHA, others remain the same.

3.3.5 Cascade of V-SP-SA, Q-SA and GA

MCAN paper proposes two cascading modes: encoder–decoder and stacking. Among them, the stacking method uses the output of the previous layer directly as the input of the next layer, while the encoder–decoder method uses the question self-attention feature of each layer as the query matrix.

Our model uses the cascade mode of encoder–decoder, assuming that $\text{SA}^1, \text{SA}^2, \ldots, \text{SA}^L$ represents the text self-attention of different layers, $\text{SPSA}^1, \text{SPSA}^2, \ldots, \text{SPSA}^L$ represents the image self-attention based on spatial position of different layers, $\text{GA}^1, \text{GA}^2, \ldots, \text{GA}^L$ represents the guided-attention of different layers, and X^k and Y^k represent the image features and text features output by layer k, respectively. Therefore, the formula of the attention module of the encoder–decoder cascade in layer k is:

$$\begin{cases} Y^k = \text{SA}^k\left(Y^{k-1}\right) \\ X^k = \text{GA}^k\left(Y^L, \text{SPSA}^k\left(X^{k-1}\right)\right) \end{cases} \tag{13}$$

3.4 Feature fusion

We need to fuse the image feature and the question text feature before sending them to predict the answer.

At first, the text feature of question $X = [x_1, x_2, \ldots, x_m]$ and the image feature $Y = [y_1, y_2, \ldots, y_n]$ was passed to a two-layer MLP to the weight of each component α_i, β_j.

Then, the model sums all component weights of visual features and text features.

$$\begin{cases} \text{MLP}(X) = \text{FC}_{2d}^{d} \circ \text{Relu} \circ \text{FC}_{d}^{d}(X) \\ \alpha = \text{softmax}\,(\text{MLP}(X)) \\ V = \sum_{i=1}^{m} \alpha_i x_i \\ \beta = \text{softmax}\,(\text{MLP}(Y)) \\ Q = \sum_{i=1}^{n} \beta_i y_i \end{cases} \tag{14}$$

where FC_{2d}^{d}, $\text{FC}_{d}^{d}()$ are fully connected layers and $\text{FC}_{2d}^{d}()$ indicate the dimension is converted from d to $2d$, the sysbol $\circ$ is a composition of two layers. Vector $\alpha =$

$[\alpha_1, \alpha_2, \ldots, \alpha_M] \in \mathbb{R}^M$ and vector $\beta = [\beta_1, \beta_2, \ldots, \beta_N] \in \mathbb{R}^N$.

Finally, we use a simple linear projection to obtain the fusion feature:

$$h = \mathrm{LayerNorm}(W_v^T V + W_q^T Q) \tag{15}$$

where vector W_v^T, $W_q^T \in \mathbb{R}^{d*d_z}$, $h \in \mathbb{R}^{d_z}$ is the fusion feature of the question and the image.

3.5 Answer prediction and loss function

VQA is mostly used as a classification task. Correct answers appearing more than eight times in the dataset will be classified as candidates. We pass the fusion features h through a linear classifier, followed by a Sigmoid function. As a result, a probability distribution for the candidate answers can be constructed. Choose the one with the greatest probability as the prediction answer.

$$\hat{y} = \mathrm{Sigmoid}(W_z^T h) \tag{16}$$

where vector $W_z^T \in \mathbb{R}^{d_z*A}$, A is the number of the candidate answers.

4 Experiment

4.1 Datasets

Many datasets are used in the VQA task, including Visual Genome [33], GQA [28], and VQA v2.0 [27]. VQA v2.0 dataset and GQA dataset are used for training and testing, which are described below:

VQA 2.0 is the most popular dataset. It is divided into training set, val set, and test set, which correspond to 248,349 questions, 121,512 questions, and 244,302 questions, respectively. Questions into three categories: yes/no, number, and other. It has 204k images from MS-COCO dataset. Each question has 10 answers, which are provided by 10 different people. A question may have multiple answers. As long as more than three people provide this answer, then this is the right answer:

$$\mathrm{acc(ans)} = \min\left\{\frac{\text{\#numb of the people provided ans}}{3}, 1\right\} \tag{17}$$

where acc is the answer accuracy.

The GQA [28] dataset consists of more than 110K images along with 22M questions. More questions in GQA dataset need multi-step reasoning and have more balanced answer distributions. About 94% of its questions need multi-step reasoning, and 51% query the relation between objects. Datasets are randomly divided into train, validation, and test splits, and test set includes test-dev, test-std, and test-challenge, totaling 420k questions. We train our model on the balanced training split and balanced validation split, then test results on the test split on the evaluation server.

4.2 Experimental setup

The dimension of the image d_y and the dimension of the question d_x are 2048 and 768, respectively. Each question is 14 words long. The dimensions of d and fusion feature d_z are 512 and 1024. The number of heads h is set to 8. The number of layers L is 6. The candidate's answer A is 3129. The batch size is set to 128.

The Adam optimizer [34] is used to optimize the model during training, and the parameters, β_1 and β_2, are, respectively, 0.9 and 0.98. We define the learning rate as $\min(2.5te^{-5}, e^{-4})$, where t is the number of iterations starting from 1. Every two iterations, the learning rate decreases by 0.2. Pytorch is used for all experiments.

The GQA dataset provided $A = 1878$ candidate answers. Tokenized words have a maximum length of 14 characters. The learning rate strategy of VQA v2.0 was also applied to the GQA dataset. To optimize our model on the GQA dataset, we used cross-entropy loss (CE).

4.3 Ablation analysis

In this section, ablation experiments are performed on VQA 2.0 to explore how effective the components are. We train the models on the training set and test them on the validation set.

Table 1 shows the ablation results.

- MCAN + LSTM [13]: denotes deep modular co-attention networks and static word vector based on LSTM.
- MCAN + BERT: represents deep modular co-attention networks and dynamic word vector based on BERT.
- MCAN + BERT + SPSA: deep modular co-attention networks, image vision self-attention based on spatial position relationship, and dynamic word vector based on BERT.
- MCAN + BERT + SPSA (large): deep modular co-attention networks, image vision self-attention based on spatial position relationship, and dynamic word vector based on BERT.

For the first row, we used deep modular co-attention networks and static word vectors based on LSTM.

With the dynamic word vector based on BERT, we get a 0.15% improvement in the second row, proving that the

Table 1 Evaluation of the ablation model using the VQA v2.0 validation set

Ablation model	Accuracy (%)
MCAN + LSTM	67.20
MCAN + BERT	67.35
MCAN + BERT + SPSA	67.59
MCAN + BERT + SPSA (large)	**67.74**

Results in bold are those with the highest scores

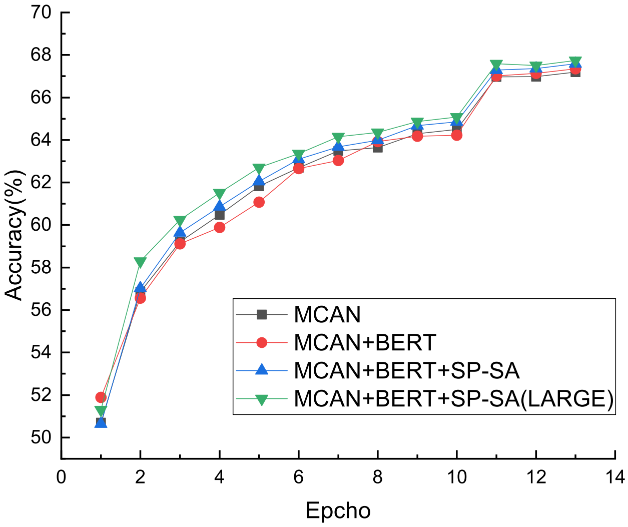

Fig. 3 Accuracy curve of ablation model

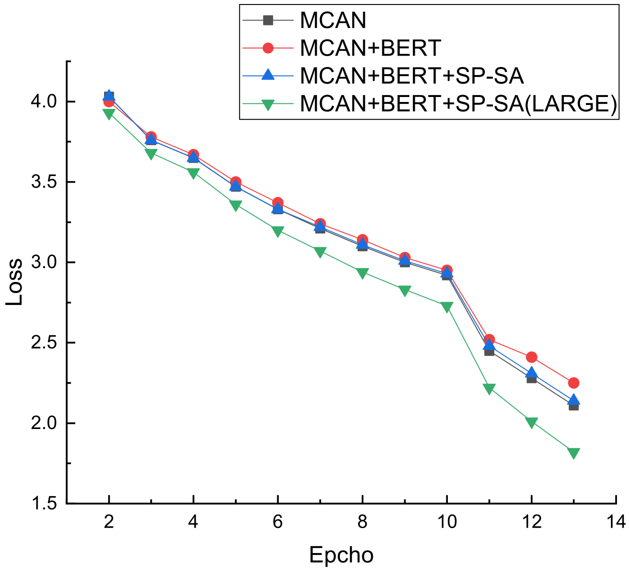

Fig. 4 Loss curve of ablation model

Table 2 The results of different loss functions on VQA v2.0 val set

Fusion method	Y/N	Num	Other	Overall
Mul	84.93	51.79	58.01	67.31
Add	**85.01**	**51.94**	**58.48**	**67.59**

Results in bold are those with the highest scores. Questions have three categories: yes/no, number, and other. Overall represents the overall accuracy

dynamic word vector significantly improves prediction performance.

On the third row, we evaluated image vision self-attention based on spatial position relationship, which was improved by 0.39%, proving that this method is effective.

Based on the third row, we built a larger model for the fourth row, and we improved it by 0.46%.

During the training, the accuracy curve and loss curve are shown in Figs. 3 and 4.

Table 2 shows the results of different operations in the fusion of image and question features. Add and Mul represent element-wise addition and multiplication, respectively. Add provides the better accuracy. It shows that element-wise addition is more effective in fusing the two features.

4.4 Qualitative analysis

In Fig. 5, we select the three largest weights from the regions in the image for visualization. We mark them with boxes and then display the corresponding weights next to the boxes. The boxes in the picture correspond to the question framing the two birds, as we'd expect.

Figure 5 displays the attention maps of Q-SA, V-SP-SA, and GA. By increasing the number of layers of Q-SA, it becomes more intuitive which words have a larger weight. 'How' and 'birds' get more attention. We have found keywords in the question, which explains our findings. For the V-SP-SA, the areas with high brightness in the picture are several important image areas found by the model. For GA, the last layer shows the word 'bird' is most closely related to some areas of the image. From this image and question, our model associates the important text of the question and image.

4.5 Comparison with the state-of-the-arts

We test our model against the current state-of-the-art on VQA v2.0. Table 3 shows the evaluation results.

The BUTD [6] was proposed and won the VQA 2017 challenge, our model exceeds BUTD by 6.35 percentage points. MFB [35] and MFH [36] mainly focus on fusion text features and visual features, our model is improved by 5.07%, and 2.59%, respectively. MuRel [37] is a multi-modal relational network that can reason over images using end-to-end learning. Our model is 3.26% higher than MuRel. To improve performance and interpretability, the MRA-Net [38] model combines textual and visual relations. Even without exploring pairwise relationships, our model achieves

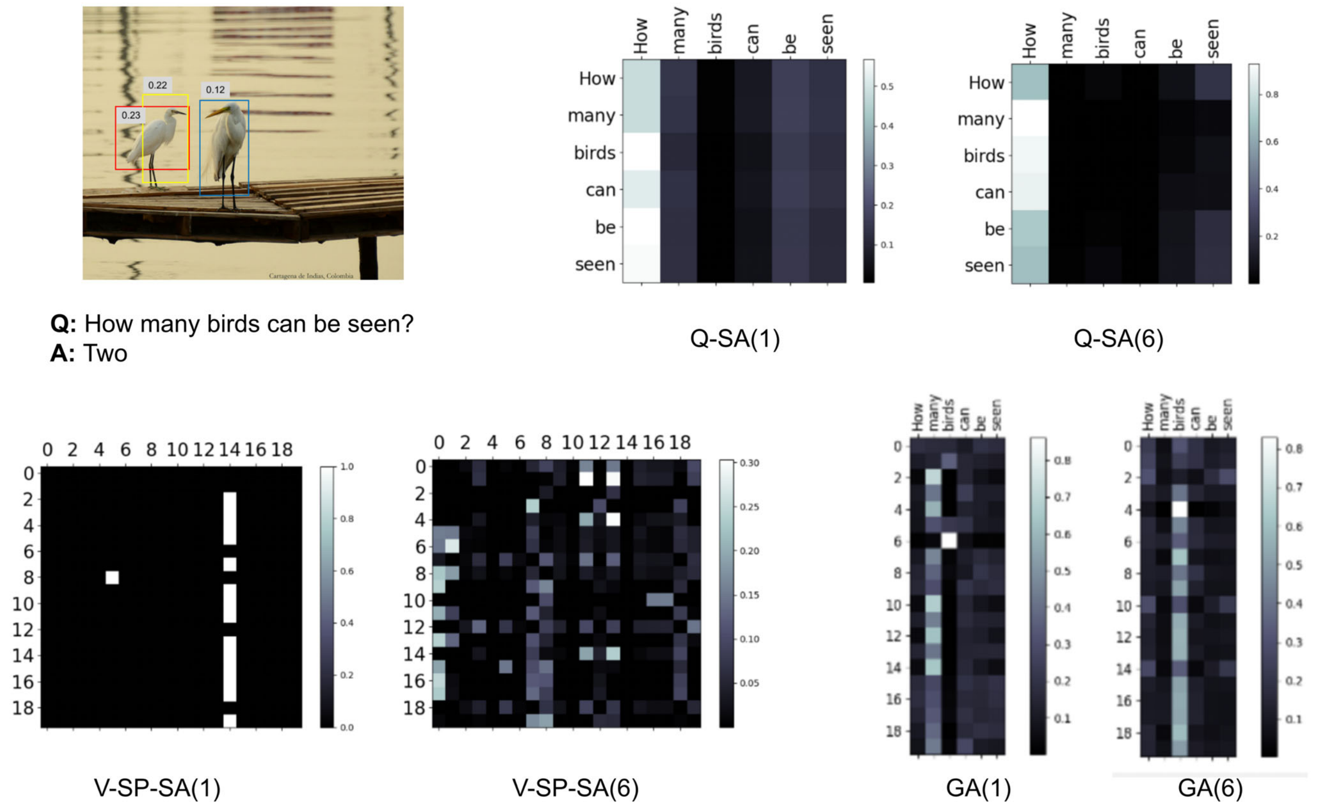

Fig. 5 Six layers were set for the model. In the first and last layers, attention maps of the Q-SA, V-SP-SA, and GA are displayed. GA and Q-SA, V-SP-SA are guided-attention networks, question self-attention networks, and vision self-attention networks based on spatial position, respectively

Table 3 Accuracy of single model on VQA v2.0 test-dev and test-standard dataset

Model	Test-dev (%)				Test-std (%)			
	Y/N	Num	Other	Overall	Y/N	Num	Other	Overall
Bottom-Up [6]	81.82	44.21	56.05	65.32	82.20	43.90	56.26	65.67
MFB [35]	84.1	39.1	58.4	66.9	84.2	38.1	57.8	66.6
MuRel [37]	84.77	49.84	57.85	68.03	–	–	–	68.41
MFH [36]	84.27	49.56	59.89	68.76	–	–	–	–
MRA-NET [38]	85.58	48.92	59.46	69.02	85.83	49.22	59.86	69.46
MCAN [13]	86.82	53.26	60.72	70.63	–	–	–	70.9
DFAF [11]	86.73	52.92	61.04	70.59	–	–	–	70.81
DMBA-NET (train + val) [39]	87.55	51.15	60.72	70.69	87.81	50.26	60.79	70.85
MDFNet [43]	86.85	53.73	61.78	71.19	–	–	–	71.32
Our	87.18	54.98	61.52	71.35	87.50	54.83	61.73	71.67

2.23% higher than MRA-Net. Based on modular co-attention (MCA) layers cascaded in depth, MCAN [13] proposes a deep modular co-attention network. The modular composition of two basic attention units is used in each MCA layer to model the attention to images and questions, as well as the attention to questions when viewing images. Our model is 0.86% higher than MCAN. DMBA-NET uses a bilinear attention network instead of multi-head attention to calculate intramodal and intermodal attention. Without the VG dataset, DMBA-NET achieves an equivalent performance to MCAN, and our model is improved by 0.81%. MDFNet [43] design an effective multimodal deep fusion network to achieve fine-grained multimodal fusion, our model is 0.35% higher than MDFNet.

In Table 4, we compare the results of our model and the SOTA models on the GQA dataset. BUTD [6] is the winner model of the 2017 VQA challenge. The overall accuracy of our model is 3.67% higher than BUTD. BAN [10] (bilinear attention network) takes into account the interaction of words in the question and the objects in the image to build a bilinear co-attention map utilizing each pair of multi-modal channels. Our model achieves 0.86% higher than BAN. Both

Table 4 Accuracy of single model on GQA test-dev and test-standard dataset

Model	Test-dev						Test-std
	Accu (%)	Binary (%)	Open (%)	Valid (%)	Plausib (%)	Consist (%)	Accu (%)
BUTD [6]	53.38	67.78	40.72	96.62	84.81	77.62	–
BAN [10]	56.19	73.31	41.13	96.77	85.58	84.64	–
LCGN [40]	55.8	–	–	–	–	–	56.1
OCCAM [41]	56.2	–	–	–	–	–	56.3
MCAN [13]	56.00	75.61	38.76	96.69	85.35	87.03	–
MCLN [42]	56.80	–	–	–	–	–	57.00
Our	57.05	76.23	40.20	96.44	85.23	87.78	57.34

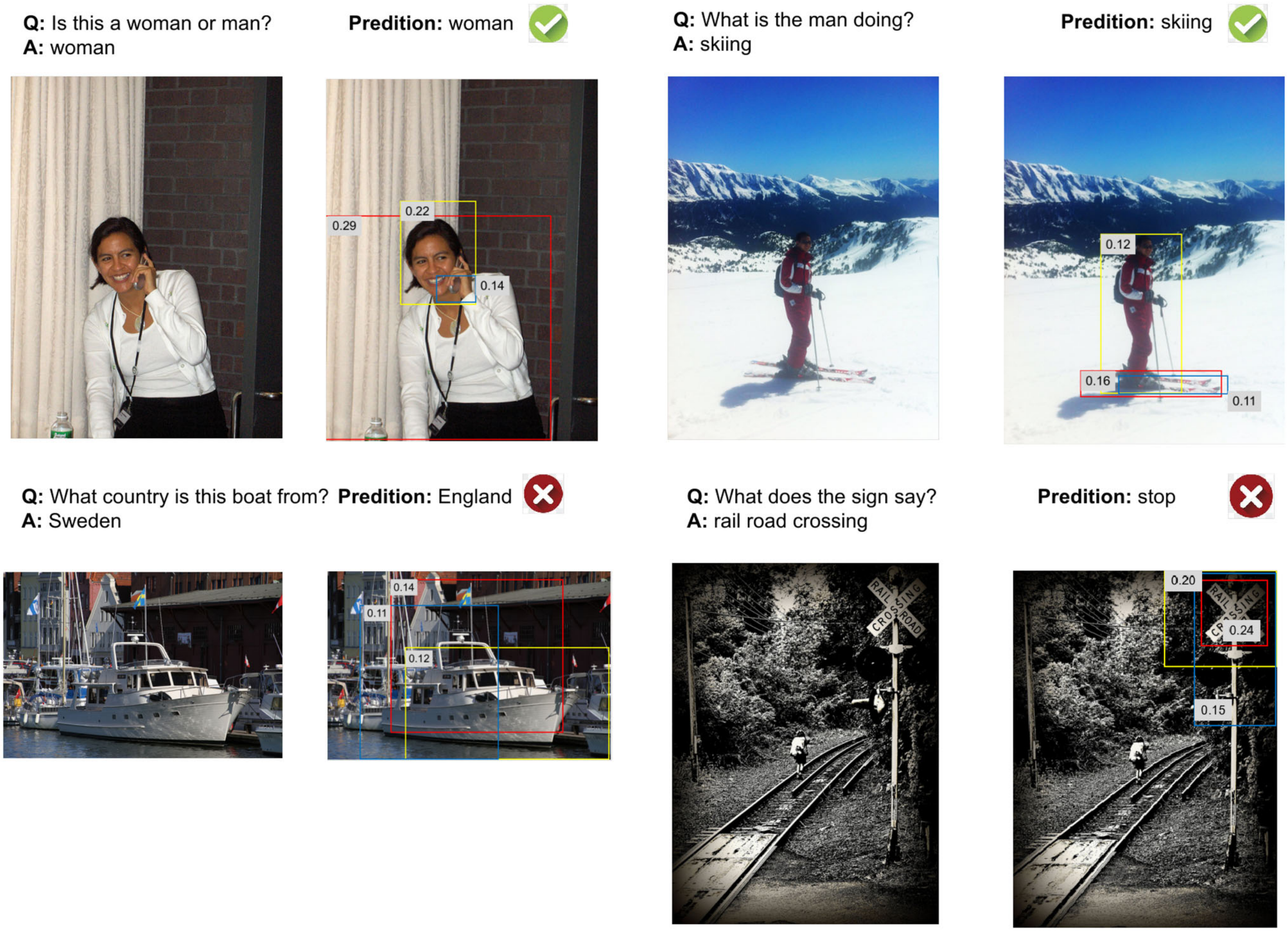

Fig. 6 Some typical examples of our model prediction. The first two examples are correct predictions and the last two are wrong predictions. Each example only shows the top-3 regions, while the object regions are highlighted by boxes. In the images, the number next to the box indicates how much attention each region receives

LCGN [18] and OCCAM [30] are deep context learning models. However, both of them only consider the context of the image modality. Our model exceeds them by 1.25% and 0.85%, respectively. MCLN [42] design corresponding context learning modules and compose them to create a multiple context learning layer. Our model achieves 0.25% higher than MCLN.

5 Conclusion

In order to demonstrate the effects of attention, we randomly selected some questions and visualized the image attentions in Fig. 6. We can see that the first three regions in the box with the highest score are all related to questions. The image attention can focus on the keyword of the questions. Therefore, our model works well. Through the incorrect examples, we found the weakness of our model: (1) the model cannot recognize the text in the picture(e.g., text on the sign in the 4th picture); (2) the model lacks certain common sense (e.g., the model does not know what the national flag is. Even if the national flag in the picture can be detected, it does not know which country it is from, in the 3th example). The shortcomings described here provide a basis for further improvements in VQA.

We propose a co-attention mechanism for VQA in this paper. The model includes three units: a text self-attention mechanism, a visual self-attention mechanism integrated into spatial position, and a question-guided-attention mechanism. To obtain more fine-grained visual and text features, these three units can be stacked to increase the depth of the model.

Based on our experimental results, we found that our model performed better than other co-attention mechanisms (MCAN), and we were able to demonstrate that spatial location and size information aid in improving the accuracy of our model. Moreover, visualizing the regions boxes with high attention shows that our model can identify the image regions related to the question.

In the future, we intend to focus on using more efficient models to extract the features of image instead of Faster R-CNN. In addition, through some wrong answer information predicted by the model, we find that there are some obstacles in the character recognition in the image. The model combined with understanding the character information in the image should be a research hotspot in the field of VQA.

Acknowledgements This work was supported in part by National Natural Science Foundation of China under Grant U1911401 and Key Project of Science and Technology Innovation 2030 supported by the Ministry of Science and Technology of China under Grant ZDI135-96.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Donahue, J., Anne Hendricks, L., Gua-Darrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 677–691 (2017)
2. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of ICML, Lille, FR, pp. 2048–2057 (2015)
3. Nam, H., Ha, J.-W., Kim, J.: Dual attention networks for multimodal reasoning and matching. In: Proceedings of CVPR, Honolulu, HI, UT, pp. 2156–2164 (2017)
4. He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
5. Sun, S.Y., Pang, J.M., Shi, J.P., Yi, S., Ouyang, W.L.: Fishnet: a versatile backbone for image, region, and pixel-level prediction. In: Advances in Neural Information Processing Systems, pp. 760–770 (2018)
6. 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: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE (2018)
7. Gao, Y., Beijbom, O., Zhang, N., Darrell, T.: Compact bilinear pooling. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 317–326 (2016)
8. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2016)
9. Nguyen, D.-K., Okatani, T.: Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In: Proceedings of CVPR, Salt Lake, UT, pp. 6087–6096 (2018)
10. Kim, J.-H., Jun, J., Zhang, B.-T.: Bilinear attention networks. In: Advances in Neural Information Processing Systems (NIPS) (2018)
11. Gao, P., Li, H., You, H., Jiang, Z., Lu, P., Hoi, S.C.H., Wang, X.: Dynamic fusion with intra- and inter-modality attention flow for visual question answering. In: Proceedings of CVPR, Long Beach, CA, UT, pp. 6639–6648 (2019)
12. Gao, P., You, H., Zhang, Z., Wang, X., Li, H.: Multi-modality latent interaction network for visual question answering. Available: https://arxiv.org/abs/1908.04289 (2019)
13. Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: 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 (2019)
14. Yu, Z., Cui, Y., Yu, J., Tao, D., Tian, Q.: Multimodal unified attention networks for vision-and-language interactions. Available: https://arxiv.org/abs/1908.04107 (2019)
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 6000–6010 (2017)
16. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 1137–1149, 2017-01-01 (2017)
17. Xu, H., Saenko, K.: Ask, attend and answer: exploring question-guided spatial attention for visual question answering. In: Computer Vision—ECCV, vol. 2016, pp. 451–466. Springer International Publishing (2016)
18. Sun, Q., Fu, Y.: Stacked self-attention networks for visual question answering. In: Proceedings of the 2019 on International Conference on Multimedia Retrieval. ACM (2019)
19. Chowdhury, M.I.H., Nguyen, K., Sridharan, S., Fookes, C.: Hierarchical relational attention for video question answering. In: 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE (2018)
20. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv e-prints arXiv:1810.04805 (2018)
21. Yu, D., Fu, J., Mei, T., Rui, Y.: Multi-level attention networks for visual question answering. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2017)
22. Kim, J., On, K., Kim, J., Ha, J., Zhang, B.: Hadamard product for low-rank bilinear pooling. arXiv preprint arXiv:1609.02907, 2016-10-14 (2016)
23. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., Gao, J.: Unified vision-language pre-training for image captioning and VQA. In:

Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 13041–13049, 2020-01-01 (2020)
24. Tan, H., Bansal, M.: LXMERT: learning cross-modality encoder representations from transformers. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics (2019)
25. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: NeurIPS (2019)
26. Li, L.H., Yatskar, M., Yin, D., Hsieh, C., Chang, K.: VisualBERT: a simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
27. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: visual question answering. In: International Conference on Computer Vision (ICCV) (2015)
28. Hudson, D.A., Manning, C.D.: GQA: a new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, June, pp. 6700–6709 (2019)
29. Wu, Q., Wang, P., Shen, C., Dick, A., Van Den Hengel, A.: Ask me anything: free-form visual question answering based on knowledge from external sources. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2016)
30. Wang, P., Wu, Q., Shen, C., Dick, A., van den Hengel, A.: Explicit knowledge-based reasoning for visual question answering. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization (2017)
31. Wang, P., Wu, Q., Shen, C., Dick, A., van den Hengel, A.: FVQA: fact-based visual question answering. IEEE Trans. Pattern Anal. Mach. Intell. **40**, 2413–2427, 2018-01-01 (2018)
32. Wu, Q., Shen, C., Wang, P., Dick, A., van den Hengel, A.: Image captioning and visual question answering based on attributes and external knowledge. IEEE Trans. Pattern Anal. Mach. Intell. **40**, 1367–1381, 2018-01–01 (2018)
33. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual genome: connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. **123**, 32–73, 2017-01–01 (2017)
34. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
35. Yu, Z., Yu, J., Fan, J., Tao, D.: Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE (2017)
36. Yu, Z., Yu, J., Xiang, C., Fan, J., Tao, D.: Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans. Neural Netw. Learn. Syst. **29**(12), 5947–5959 (2018)
37. Cadene, R., Ben-Younes, H., Cord, M., Thome, N.: Murel: multimodal relational reasoning for visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1989–1998 (2019)
38. Peng, L., Yang, Y., Wang, Z., Huang, Z., Shen, H.T.: MRA-NET: improving VQA via multi-modal relation attention network. IEEE Trans. Pattern Anal. Mach. Intell. **44**, 318–329 (2020)
39. Yan, F., Silamu, W., Li, Y.: Deep modular bilinear attention network for visual question answering. Sensors **22**(3), 1045 (2022). https://doi.org/10.3390/s22031045
40. Hu, R., Rohrbach, A., Darrell, T., Saenko, K.: Language-conditioned graph networks for relational reasoning. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, June, pp. 10294–10303 (2019)
41. Wang, Z., Wang, K., Yu, M., et al.: Interpretable visual reasoning via induced symbolic space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, Seattle, WA, USA, June, pp. 1878–1887 (2020)
42. Zhang, P., Lan, H., Khan, M.A.: Multiple context learning networks for visual question answering. Sci. Program. **2022**, 1–11 (2022)
43. Zhang, W., Yu, J., Wang, Y., et al.: Multimodal deep fusion for image question answering. Knowl. Based Syst. **212**, 106639 (2021)

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

Feng Yan is currently a Ph.D. candidate with School of Information Science and Engineering, Xinjiang University. His current research interests include computer vision, natural language processing, information retrieval, data mining, and social computing.

Wushouer Silamu is academician of Chinese Academy of Engineering, Executive Director of the Chinese Association for Artificial Intelligence. Professor of Xinjiang University and doctoral supervisor. He has published more than 200 papers and presided over 65 key projects including 7 national 863 projects and 1 national 973 project.

Yanbin Li is Xinjiang University lecturer. The research interest is natural language processing and computer vision.

Yachuang Chai is currently pursuing a Ph.D. degree at Xinjiang University. The research interest is computer vision and intelligent transportation.