Applied Intelligence (2023) 53:586–600
https://doi.org/10.1007/s10489-022-03559-4

Sparse co-attention visual question answering networks based on thresholds

Zihan Guo[1] · Dezhi Han[1]

Accepted: 27 March 2022 / Published online: 20 April 2022

Abstract

Most existing visual question answering (VQA) models choose to model the dense interactions between each image region and each question word when learning the co-attention between the input images and the input questions. However, to correctly answer a natural language question related to the content of an image usually only requires understanding a few key words of the input question and capturing the visual information contained in a few regions of the input image. The noise information generated by the interactions between the image regions unrelated to the input questions and the question words unrelated to the prediction of the correct answers will distract VQA models and negatively affect the performance of the models. In this paper, to solve this problem, we propose a Sparse Co-Attention Visual Question Answering Network (SCAVQAN) based on thresholds. SCAVQAN concentrates the attention of the model by setting thresholds for attention scores to filter out the image features and the question features that are the most helpful for predicting the correct answers and finally improves the overall performance of the model. Experimental results, ablation studies and attention visualization results based on two benchmark VQA datasets demonstrate the effectiveness and interpretability of our models.

Keywords Visual question answering · Sparse co-attention · Attention score · Threshold

1 Introduction

Among artificial intelligence (AI) technologies, computer vision and natural language processing are the most popular and successful applications of deep learning, and these two fields also have the most direct impact on people's lifestyles. Computer vision aims to make machines have the same visual senses as humans to analyze and understand various kinds of visual information, and its most well-known applications are face recognition and autonomous driving. Natural language processing, on the other hand, involves various scenarios related to text and speech, mainly including text classification [1, 2], machine translation [3, 4] and speech recognition [5, 6]. As research continues to develop, researchers have gradually turned their attention to multi-modal learning tasks that combine visual information and textual information, such as image-text matching [7, 8], image captioning [9, 10] and visual question answering (VQA) [11–13]. While these multi-modal learning tasks are becoming more similar to the way the human brain processes information, they are also placing higher demands on the associated algorithms and models.

Given a text query or an image query, image-text matching models retrieve the visually relevant image or the semantic matching text by measuring the similarity between the input image and the input text. The task of image captioning is more complicated than the task of image-text matching, because the related models not only need to perceive various visual objects contained in the input images, but also need to capture the attributes of these visual objects and their interactions and express them with natural language. However, it has been shown that rough scene-level image understanding combined with n-gram statistics of words is sufficient to generate a reasonable general image caption. Given an image and a natural language question related to the content of the image, VQA task requires the model to give an accurate natural language answer. Compared with image captioning models, an effective VQA model must encode the textual information contained in

✉ Dezhi Han
dzhan@shmtu.edu.cn

Zihan Guo
guo_zihan11@163.com

[1] College of Information Engineering, Shanghai Maritime University, 1550 Haigang Avenue, Shanghai, 201306, China

the input questions to understand the task objective while encoding the input images. After obtaining the encoded features of the input images and the input questions, VQA models also need to capture the high-level interactions between the information of the two different modalities and combine multi-modal feature fusion methods to complete the inference and prediction of the answers. It can be seen that VQA task requires models to have the ability to recognize the specific details contained in the input images, such as the color of visual objects, human activities and the interactions between visual objects. In addition, some complex input questions may involve a certain degree of reasoning and even require information not contained in the input images, such as empirical knowledge existing in other instances and common knowledge that can be learned from external knowledge bases, to be answered correctly.

Since attention mechanism was introduced into VQA [14, 15], it has become an important part of advanced VQA models. At present, most of the mainstream VQA systems choose to use co-attention mechanisms combining visual attention and textual attention. Visual attention helps the model to filter out significant regional image features, while textual attention helps the model to focus on key words of the input questions. Most existing VQA models choose to model the dense interactions between each image region and each question word when learning the co-attention between the input images and the input questions. However, to correctly answer a natural language question related to the content of an image usually only requires understanding a few key words of the input question and capturing the visual information contained in a few regions of the input image. The noise information generated by the interactions between the image regions unrelated to the input questions and the question words unrelated to the prediction of the correct answers will distract VQA models and negatively affect the performance of the models. Although the traditional attention mechanism can empower the model with considerable noise-filtering ability, this effect is achieved by assigning less weight to irrelevant features and features with weak correlations, that is, the noise information is not completely eliminated. Therefore, MESAN [16] based on top-k selection has been proposed to concentrate the attention of the model by selecting a specified number of question key words. However, the drawback of this approach is that the model will ignore the important textual features when the number of the key words of the input question is greater than the preset value. Moreover, when the number of the key words contained in the input question is smaller than the preset value, the model will regard the words that are originally insignificant or even meaningless as key words, thus introducing unnecessary noise information. In this paper, to solve this problem, we propose a threshold-based Sparse Co-Attention Visual Question Answering Network (SCAVQAN) that is more suitable for VQA task and achieve better results. SCAVQAN concentrates the attention of the model by setting thresholds for attention scores to filter out the image features and the question features that are the most helpful for predicting the correct answers and finally improves the overall performance of the model. Experimental results, ablation studies and attention visualization results based on two benchmark VQA datasets demonstrate the effectiveness and interpretability of our models. Our threshold-based attention mechanism is a heuristic approach that can be applied to any model that infuses attention mechanism and can be tried in other tasks besides VQA. The code is available at https://github.com/644988396/SCAVQAN.

The rest of this paper is organized as follows: Section 2 introduces the research progress of VQA. Section 3 introduces the overall framework and specific technical details of SCAVQAN. Section 4 introduces the datasets used in the experiments, the settings of the experimental parameters, the ablation studies and the experimental results. In Section 5, we summarize our work and give the future research direction.

2 Related work

2.1 Visual question answering (VQA)

In addition to being a method to test the deep visual understanding ability of a model, VQA is regarded as the benchmark for general AI because it requires a large number of potential AI capabilities to be performed well. The deep visual understanding ability refers to the ability of algorithms to extract high-level visual information from the input images and make inferences based on it. The potential AI capabilities required for VQA task mainly include object detection (e.g., “How many bicycles are there in this picture”), behavior identify (e.g., “What is the person in this picture doing”), fine-grained recognition (e.g., “What vegetable is in the basket”), knowledge base reasoning (e.g., “Is this a vegetarian pizza or a meat pizza”) and common sense reasoning (e.g., “Is the person in the picture waiting for someone”). Thus, VQA requires the model to have a large amount of multi-modal knowledge beyond specific domains and it is regarded as a complete AI task. VQA systems have gone from being able to barely answer Yes/No questions to being able to correctly answer questions that require complex reasoning and additional common sense knowledge, and their achievements have far exceeded expectations. Moreover, VQA algorithms can be used in a wide range of scenarios, such as helping visually impaired users or intelligence analysts, providing medical assistance or automated customer service, and building surveillance video automated query systems.

2.2 Attention mechanisms in VQA

Attention mechanism, which is inspired by the analogy of human visual system, is one of the most effective improvements in multi-modal learning tasks. When viewing an image, humans often choose to focus on different regions of the image to quickly understand the visual information corresponding to these regions, rather than processing the whole image at once. Attention mechanisms based on deep neural networks have been widely used in the fields of machine translation, speech recognition and image captioning. Attention mechanisms help models to give different weights to the features of different degrees of importance to make the learning of neural networks more flexible. By infusing attention mechanisms, VQA models can focus attention to analyze and process various kinds of input information and reduce the negative impact of irrelevant information on their performance. Moreover, the attention weights calculated for the input images and the input questions can be used to realize attention visualization to enhance the interpretability of the relevant models. Among them, the attention weights can be understood as the importance of features. The higher the attention weight is, the more important the feature is, and vice versa. In VQA, visual attention helps the model to filter out salient regional image features, while textual attention helps the model to focus on key words of the input questions. Nowadays, most advanced VQA methods utilize co-attention mechanisms combining visual attention and textual attention to focus the models on both the image regions and the question key words that are the most relevant to correctly predicting the answers. Nam et al. [17] proposed a dual attention network (DAN) for multi-modal reasoning that enables visual attention and textual attention to guide each other in the process of collaborative reasoning. DAN focuses on specific words of questions and local regions of images through multiple steps and gathers essential information from the features of these two modalities. Nguyen et al. [18] proposed a stackable architecture that is completely symmetric between visual features and textual features to achieve multi-step interactions between image features and question features. In this architecture, each image region attends on question words and each question word attends on image regions to implement a dense bi-directional attention mechanism. Yu et al. [19] accomplished VQA task by modeling the intra-modal interactions (between image regions or question words) and the inter-modal interactions (between image regions and question words). They proposed a deep modular co-attention network (MCAN) consisting of modular co-attention layers cascaded in depth. Each modular co-attention layer models the self-attention of image features and question features, as well as the question-guided visual attention of image features through scaled dot-product attention. MESAN [16] based on explicit selection eliminates the noise information contained in the question words that are irrelevant to predicting the correct answers and ultimately helps VQA models to achieve higher accuracy.

3 Sparse co-attention visual question answering networks based on thresholds

The overall framework of SCAVQAN is shown in Fig. 1. SCAVQAN first uses an image feature extractor and a word embedding method to extract visual features and textual features required by subsequent experiments from the input images and the input questions. Next, a threshold-based sparse co-attention module takes the previously extracted image features and question features as inputs to learn the cross-modal interactions between the visual information and the textual information to make the model focus on significant image regions and key words of the input questions and output the attended image features and the attended question features. Finally, we use a

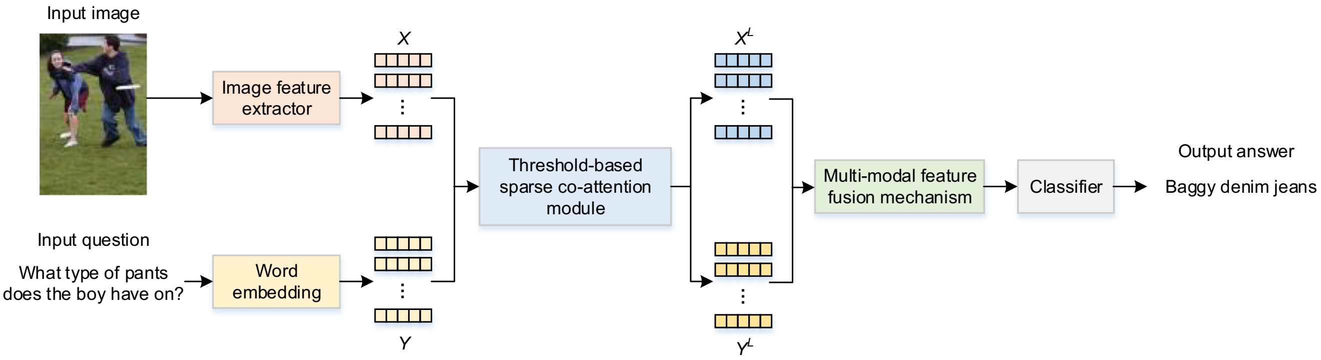

Input image
Image feature extractor
X
Input question
What type of pants does the boy have on?
Word embedding
Y
Threshold-based sparse co-attention module
X^L
Y^L
Multi-modal feature fusion mechanism
Classifier
Output answer
Baggy denim jeans

Fig. 1 The overall framework of Sparse Co-Attention Visual Question Answering Networks (SCAVQAN)

multi-modal feature fusion mechanism to fuse the attended image features and the attended question features belonging to different modalities, and input the fused features into a classifier to complete the prediction of the answers.

3.1 Extraction of image features

SCAVQAN uses Faster R-CNN [20] initialized by ResNet-101 and pre-trained on Visual Genome [21] as its image feature extractor. Faster R-CNN uses bottom-up mechanism to propose significant image regions and utilizes top-down mechanism to determine the weights of these regional image features. ResNet-101 reconstructs the layers as learning functions with reference to the layer inputs to reduce the training burden of the network and makes it easier to optimize. This network using the residual learning framework can achieve deeper depth with less computational cost and obtain higher accuracy with its greatly deepened depth. Visual Genome collects dense object annotations, attribute annotations and relationship annotations for visual images to aid visual models in object relationship modeling. Specifically, the dataset contains more than 100K images, each of which contains an average of 21 visual objects, 18 attributes and 18 pairwise relationships between the visual objects. Faster R-CNN is composed of a region proposal network (RPN) and a detection network, in which the region proposal network realizes nearly cost-free region proposals by sharing the full-image convolutional features with the detection network. Specifically, the region proposal network is a fully convolutional network, which is trained end-to-end to predict object boundaries and target scores for image regions. The region proposal network and the detection network jointly form the single Faster R-CNN model and guide the model which image regions should be concerned by sharing the full-image convolutional features.

In order to ensure the quality of the visual features, we only select the image regions whose detection probability exceeds the confidence threshold. After the images are fed into Faster R-CNN, we obtain a dynamic number of regional image features $X^0 \in \mathbb{R}^{m\times 2048}$, where $m \in [10, 100]$ is the number of the detected image regions of the input images. During the experiments, in order to facilitate calculation, we apply linear transformation to X^0 and obtain the regional image features $X \in \mathbb{R}^{m\times 512}$ to make the dimension of the image features and the dimension of the question features consistent. Among them, the feature of the i-th image region is $x_i \in \mathbb{R}^{512}$.

3.2 Extraction of question features

In order to facilitate calculation and improve the operation efficiency of the model, we limit the maximum length of the input questions to 14 for VQA 2.0 [22] and 29 for GQA [23] following previous works. The extra question words are discarded, but only a small percentage of the input questions are longer than the maximum length. First, SCAVQAN uses GloVe [24], a global log-bilinear regression model pre-trained on a large-scale corpus, to embed each word contained in the input questions into a 300-dimensional word vector. GloVe is a vector arithmetic method for learning vector space representations of words that can capture fine-grained semantics as well as grammatical rules. The GloVe model is a specific weighted least squares model based on global word-word co-occurrence counting matrix for efficient use of statistics.

Specifically, SCAVQAN tokenizes the input questions as individual words and inputs these words into a pre-trained GloVe word vector to obtain the $n \times 300$-dimensional word embedding sequences, where n is the number of the words contained in the input questions. Next, we input the word embedding sequences of the input questions into a single layer long short-term memory network (LSTM) with 512 hidden units to obtain the question features $Y \in \mathbb{R}^{n\times 512}$. Here, we use the output features of all the question words as the question features rather than just the output features of the last question word as the question features.

3.3 Threshold-based sparse co-attention module

SCAVQAN's sparse co-attention module consists of textual attention and visual attention. Among them, the textual attention is the self-attention of the input questions used to help the model focus on the question key words, and the visual attention contains the self-attention and the question-guided attention of the input images used to help the model focus on the image regions that are the most relevant to predicting the correct answers. In this section, we will introduce the details of threshold-based multi-head scaled dot-product attention and the threshold-based sparse co-attention module used by SCAVQAN.

3.3.1 Threshold-based multi-head scaled dot-product attention

Scaled dot-product attention [25] takes queries and key-value pairs as inputs and outputs attended features, i.e., the weighted sum of the values. The queries, keys, values, and attended features are all vectors. The weight of each value is computed by a compatibility function and its inputs are the query and the corresponding key. Specifically, scaled dot-product attention computes the dot product of queries Q and keys K, and divides the results by $\sqrt{k}$ to obtain the attention scores S. Next, it uses a softmax function to obtain the attention weights W of values V to calculate the attended features F. The specific calculation process of scaled dot-product attention can be described by the following formulas:

$$S = \frac{QK^{\mathrm{T}}}{\sqrt{k}} \tag{1}$$

$$W = \mathrm{softmax}(S) \tag{2}$$

$$F = WV \tag{3}$$

where k is the dimension of K. The purpose of dividing by $\sqrt{k}$ is to prevent the dot-products from being too large in magnitude and to ensure that this method can produce useful gradients. In practice, the attention functions are usually computed based on a set of queries, and the queries, keys and values are integrated into matrices for easy calculation.

However, in VQA task, although the above traditional scaled dot-product attention can enhance the significant regional image features and the question key word features to a certain extent, other irrelevant regional image features and unimportant question word features will still participate in the prediction of the answers. The noise information contained in these unimportant features will prevent VQA models from predicting the correct answers and negatively affect the performance of the models. Therefore, SCAVQAN uses threshold-based scaled dot-product attention to filter out the regional image features and the question key word features that are the most helpful for predicting the correct answers by setting thresholds for the attention scores to focus the model's attention and finally improve the overall performance of the model. Specifically, we compare the values of the attention weight matrix W with threshold T. When the value of a weight is less than T, we set the value of the corresponding position in the attention scores S as $-\infty$. Then, we input the new attention scores S' into a softmax function to obtain the new attention weights W'. The specific calculation process can be described by the following formulas:

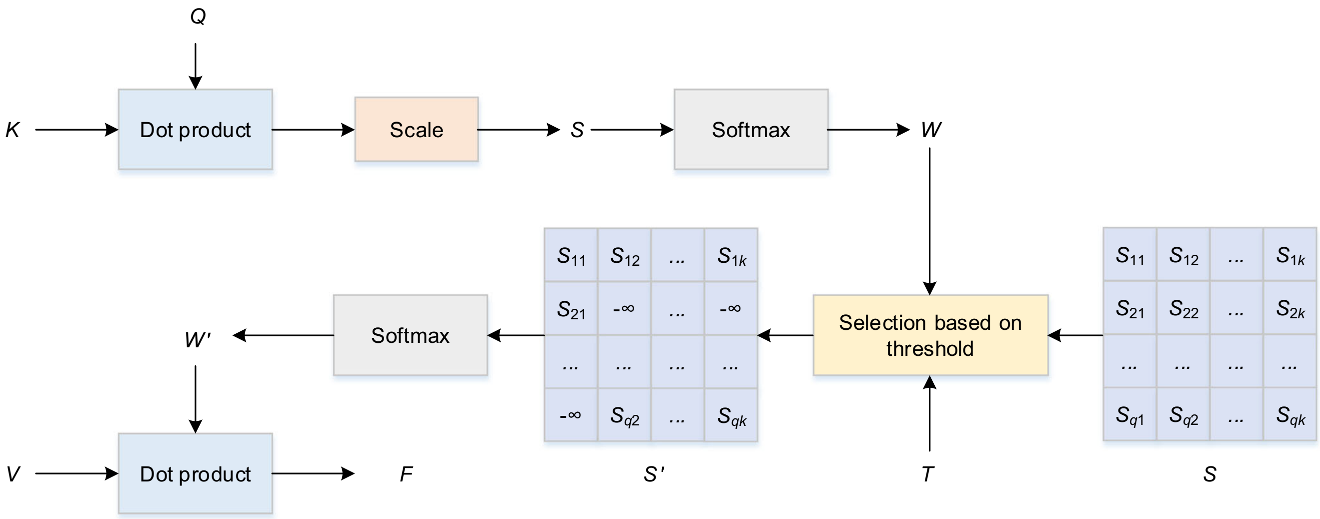
$$S'_{ij} = \begin{cases} S_{ij} & W_{ij} \geq T \\ -\infty & W_{ij} < T \end{cases} \tag{4}$$

$$W' = \mathrm{softmax}(S') \tag{5}$$

$$F = W'V \tag{6}$$

By doing so, the new attention weights of the features with lower attention scores will approach 0 infinitely, i.e., the irrelevant regional image features and the unimportant question word features will not participate in the calculation of the attended features F. Figure 2 shows the calculation process of our threshold-based scaled dot-product attention, where Q, K and V are the inputs, and q is the dimension of Q. S is the first calculated attention score, S' is the last calculated attention score, W is the first calculated attention weight, and W' is the last calculated attention weight. T is the threshold and F is the output attended feature.

Multi-head attention assigns features of different representation spaces to different attention heads so that these attention heads can independently learn the information from different locations. Specifically, multi-head attention uses linear projection to project queries and key-value pairs into H independent subspaces, where H is the number of the attention heads and the parameters of these attention heads are not shared. We execute threshold-based scaled dot-product attention in parallel in these subspaces and concatenate their outputs as the outputs of our threshold-based multi-head scaled dot-product attention. The specific calculation process of threshold-based multi-head scaled



S_{11}	S_{12}	...	S_{1k}
S_{21}	-∞	...	-∞
...	...	...	...
-∞	S_{q2}	...	S_{qk}

S_{11}	S_{12}	...	S_{1k}
S_{21}	S_{22}	...	S_{2k}
...	...	...	...
S_{q1}	S_{q2}	...	S_{qk}

Fig. 2 The calculation process of threshold-based scaled dot-product attention

dot-product attention can be expressed by the following formulas:

$$F_{\mathrm{head}_h} = \mathrm{threshold\_based\_attention}(QW_h^Q, KW_h^K, VW_h^V) \quad (7)$$

$$F_{\mathrm{multi\_head}} = \mathrm{concat}(F_{head_1}, F_{head_2}, ..., F_{head_H}) \quad (8)$$

where W_h^{Q}, W_h^{K} and W_h^{V} are linear projection parameter matrices, F_{head_h} is the output of the h-th threshold-based scaled dot-product attention head, and $F_{\mathrm{multi\_head}}$ is the output of threshold-based multi-head scaled dot-product attention. By infusing the multi-head attention technique, we further improve the representation ability of the attended features and effectively prevent overfitting.

3.3.2 Sparse co-attention module

In our sparse co-attention module, both the self-attention of the input questions and the self-attention of the input images are based on threshold-based multi-head scaled dot-product attention, while the question-guided visual attention is based on traditional multi-head scaled dot-product attention. The reason why we design the attention mechanisms in our sparse co-attention module in this way is because after learning the self-attention of the input questions and the self-attention of the input images, the model has already filtered out the textual features of question key words and the more significant regional image features. Next, the model only needs to learn the visual features that are more important for correctly answering the input questions through the question-guided visual attention. Figure 3 shows the self-attention layer of the input questions, the self-attention layer of the input images and the question-guided visual attention layer in SCAVQAN.

The self-attention layer of the input questions We denote the self-attention layer of the input questions as QSA layer. A one-layer QSA layer consists of a threshold-based multi-head scaled dot-product attention layer and a pointwise feed-forward layer. The pointwise feed-forward layer uses two fully connected layers to further transform its input and introduces ReLU function and Dropout to prevent overfitting. Its input is the output of the threshold-based multi-head scaled dot-product attention layer. The specific structure of the pointwise feed-forward layer is the same as previous studies, i.e., FC($4d$)-ReLU-Dropout(0.1)-FC(d), where d is the dimension of its input features. QSA layer takes the question features Y as inputs to learn the pairwise relationships between every two question words and outputs the attended textual features.

The self-attention layer of the input images We denote the self-attention layer of the input images as ISA layer. A one-layer ISA layer consists of a threshold-based multi-head scaled dot-product attention layer, which takes the regional image features X as inputs to learn the pairwise relationships between every two image regions and outputs the attended visual features.

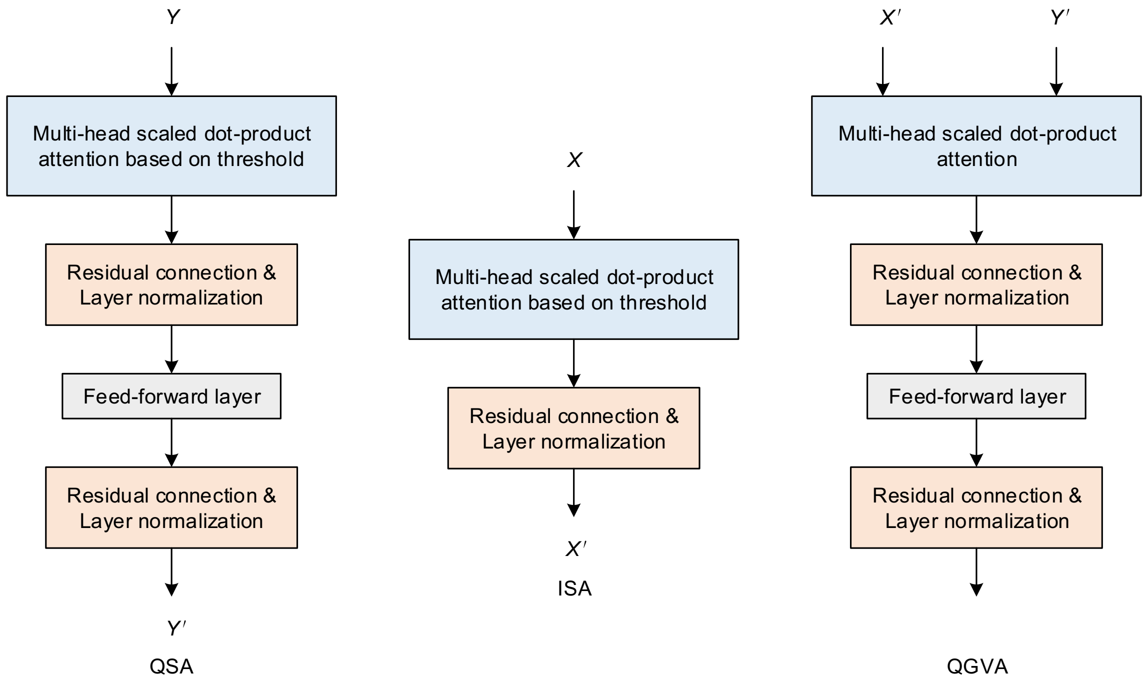

Y
Multi-head scaled dot-product attention based on threshold
Residual connection & Layer normalization
Feed-forward layer
Residual connection & Layer normalization
Y′
QSA
X
Multi-head scaled dot-product attention based on threshold
Residual connection & Layer normalization
X′
ISA
X′
Y′
Multi-head scaled dot-product attention
Residual connection & Layer normalization
Feed-forward layer
Residual connection & Layer normalization
QGVA

Fig. 3 The self-attention layer of the input questions (QSA), the self-attention layer of the input images (ISA), and the question-guided visual attention layer (QGVA) in SCAVQAN

The question-guided visual attention layer We denote the question-guided visual attention layer as QGVA layer. A one-layer QGVA layer is composed of a traditional multi-head scaled dot-product attention layer and a pointwise feed-forward layer. The structure and functions of the pointwise feed-forward layer are the same as that of the QSA layer. QGVA layer takes the outputs of QSA layer and ISA layer, i.e., the attended textual features and the attended visual features, as inputs to learn the question-guided visual attention and outputs the new attended visual features.

All the attention layers in our sparse co-attention module can be understood as a process of reconstructing the input features through the normalized similarity between the input samples. In addition, we also apply residual connection [26] and layer normalization to the outputs of all the multi-head scaled dot-product attention layers and the pointwise feed-forward layers to stabilize training. Considering that the inputs and outputs of all the attention layers have the same dimension, we stack the attention layers to form a hierarchy to help the model achieve more complex reasoning. An L-layer sparse co-attention module consists of an L-layer QSA layer and an L-layer ISA-QGVA layer. Specifically, the L-layer QSA layer takes the question features Y as inputs and outputs the attended textual features Y^L, and the inputs of the middle layers and the last layer are the outputs of their previous layer. The L-layer ISA-QGVA layer takes the regional image features X and the attended textual features Y^L as inputs and outputs the attended visual features X^L, and the inputs of the middle layers and the last layer are the outputs of their previous layer and the attended textual features Y^L. Figure 4 shows the specific structure of an L-layer sparse co-attention module.

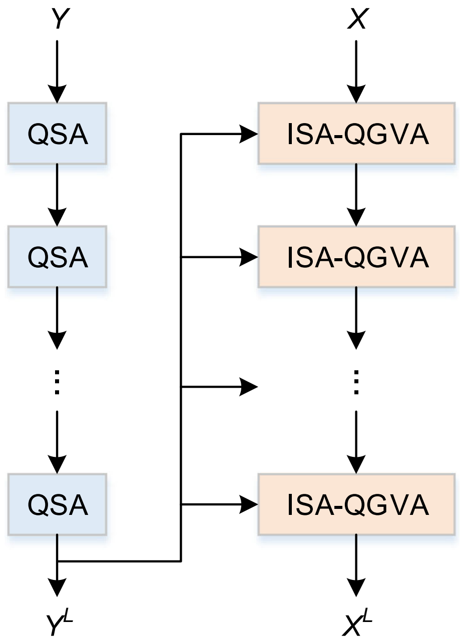


Fig. 4 The specific structure of an L-layer sparse co-attention module

3.3.3 Multi-modal feature fusion mechanism and classifier

Now, the attended visual features $X^L \in \mathbb{R}^{m\times 512}$ and the attended textual features $Y^L \in \mathbb{R}^{n\times 512}$ output by our sparse co-attention module contain rich information about the attention weights of the image regions and the question key words that are more important for correctly answering the input questions. First, we use a two-layer multi-layer perceptron (MLP) to compress X^L and Y^L to the same size, i.e., $\mathbb{R}^{512}$. We use a softmax function to calculate the attention weights of the image features and the question features and multiply them with the corresponding features to obtain the final image features $\widetilde{X}$ and the final question features $\widetilde{Y}$. Taking the final image features $\widetilde{X}$ as an example, the specific calculation process can be expressed by the following formulas:

$$\alpha = \text{softmax}(MLP(X^L)) \tag{9}$$

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

where $\alpha \in \mathbb{R}^m$ is the attention weight. Similarly, we can calculate the final question features $\widetilde{Y}$ by the above method. Then, we project $\widetilde{X}$ and $\widetilde{Y}$ to the same dimension z and the results are combined to achieve linear multi-modal feature fusion:

$$Z = \text{LayerNorm}(\widetilde{X} W_{\text{X}} + \widetilde{Y} W_{\text{Y}}) \tag{11}$$

where W_{X} and W_{Y} are linear projection parameter matrices, layer normalization is used to promote optimization, and $Z \in \mathbb{R}^z$ is the fused multi-modal feature. Finally, we project the fused multi-modal feature Z to the same dimension as the number of the most common answers of the training set and fed it into a sigmoid function to complete the prediction of the answers:

$$A = \text{sigmoid}(Linear(Z)) \tag{12}$$

where Linear$(\cdot)$ represents a linear layer, $A \in \mathbb{R}^a$ is the prediction result of the model, and a is the number of the most common answers of the training set.

4 Results and discussion

We conduct experiments on two benchmark VQA datasets to verify the effectiveness of our proposed models. In this section, we will describe the datasets used in the experiments and the specific experimental settings. We conduct ablation studies to test the performance of different variants of our models and the effects of different thresholds on their performance. Finally, we compare our models with advanced VQA methods based on the experimental results and attention visualization methods.

4.1 Datasets

VQA 2.0 Data bias in the real world often leads to VQA models ignoring the visual information contained in the input images and instead utilize the language priors implied in the input questions to predict the correct answers. For example, in VQA 1.0 [27], for Yes/No questions, VQA models can achieve more than 50% accuracy by only taking "yes" as the prediction answer without providing the input images. To solve this problem, Goyal et al. [22] created a more balanced and larger dataset VQA 2.0 based on VQA 1.0. Each question in VQA 2.0 corresponds to a pair of images that are similar but have different answers. Experimental results show that the performance of some existing VQA models on VQA 2.0 is significantly worse than their performance on VQA 1.0. It means that these models do make use of the language priors contained in VQA 1.0 and proves that VQA 2.0 effectively eliminates these language priors to some extent. The VQA 2.0 dataset composed of such image-question pairs makes the visual information contained in the input images more important for the relevant models to answer the input questions correctly and can more effectively track the progress of research related to VQA. Specifically, the complete VQA 2.0 dataset is divided into the train set, the val set, and the test set, which contains more than 443K, 214K, and 453K image-question pairs, respectively. The test set is further divided into the test-dev set, the test-std set, the test-challenge set and the test-reserve set. According to the types of the questions and the answers, the accuracy results of VQA models are divided into four types: Yes/No, Number, Other and Overall.

GQA Like the VQA 2.0 dataset, the GQA dataset [23] also seeks to eliminate the strong and prevalent real-world priors existing in previous VQA datasets. The difference is, GQA balances not only binary questions, but also open questions, by applying a tunable smoothing technique that makes the answer distribution for each question group more uniform. In addition, GQA designs a strong and robust question engine that leverages Visual Genome scene graph structures [21] to generate diverse reasoning questions, which all come with functional programs that represent their semantics, and are visually grounded in the image scene graphs, to retain the semantic and visual richness of real-world images for visual reasoning and compositional question answering. The GQA dataset consists of 113,018 images and 22,669,678 questions of assorted types and varying compositionality degrees, measuring performance on an array of reasoning skills. The GQA dataset is split into 70% train, 10% validation, 10% test and 10% challenge.

Metric For VQA 2.0, we use the standard accuracy as the evaluation metric. For GQA, we use four additional metrics to get further insight into the models, i.e., Validity, Plausibility, Consistency, and Distribution (lower is better). The validity score checks whether a given answer is in the question scope. The plausibility metric measures whether an answer is reasonable or makes sense. The consistency score measures responses consistency across different questions. The distribution metric measures the overall match between the predicted answer distribution and the true answer distribution.

4.2 Experimental settings

In order to fix the number of the image regions detected in the input images and the number of the words contained in the input questions, we use zero padding to fill X and Y to their maximum sizes (m=100 and n=14 or 29), i.e., the number of the input image regions is 100 and the number of the input question words is 14 or 29. During training, we use $-\infty$ before the softmax function to mask the padded values of the attention scores to make their probabilities approach 0 infinitely to ensure that the previously padded regional image features and padded question word features will not negatively affect the performance of the models. The dimensions of the threshold-based multi-head scaled dot-product attention layer and the traditional multi-head scaled dot-product attention layer are both 512, and the number of the attention heads is 8, i.e., H=8. The number of the layers of our sparse co-attention module is set to 6. The dimension of the fused multi-model feature Z is 1024, i.e., z=1024. The number of the most common answers of the training set is 3129 in VQA 2.0 and 1843 in GQA.

We use the Adam optimizer to train our models, where $\beta_1 = 0.9$ and $\beta_2 = 0.98$. Binary Cross-Entropy Loss is adopted as the loss function and the batchsize is set to 64. The structure of the multi-layer perceptron used in the fusion of the multi-modal features is FC(512)-ReLU-Dropout-FC(1). Our VQA 2.0 models use the train set and the val set of VQA 2.0 and a subset of Visual Genome [21] as the training set, and adopt the test-dev set and the test-std set of VQA 2.0 as the test set. Our GQA models use the train set and the val set of GQA as the training set, and adopt the test-dev set of GQA as the test set. Our VQA 2.0 models are trained up to 13 epochs and GQA models are trained up to 11 epochs. The Dropout is set to 0.1. The above experimental parameter settings and the structure of the multi-layer perceptron are consistent with

Table 1 Ablation experimental results of SCAVQAN-Q on the test-dev set of VQA 2.0

T	Yes/No	Number	Other	Overall
0.05	**87.08%**	52.98%	60.77%	70.72%
0.06	87.01%	53.33%	60.88%	70.79%
0.07	86.93%	53.16%	60.94%	70.77%
0.08	86.93%	**53.50%**	60.88%	70.78%
0.09	87.07%	52.68%	**60.96%**	70.78%
0.10	86.96%	53.49%	60.95%	**70.82%**
0.11	87.03%	53.08%	60.78%	70.72%
0.12	86.84%	52.94%	60.86%	70.66%

those in existing stacked co-attention VQA models to fairly and clearly compare the performance of the models.

4.3 Ablation studies

We have fixed the dimensions and the number of the attention heads of the threshold-based multi-head scaled dot-product attention layer and the traditional multi-head scaled dot-product attention layer, and the number of the layers of our sparse co-attention module according to the experience of other advanced co-attention VQA models. Therefore, we only need to test the performance of different variants of our models and the effects of different thresholds on their performance through ablation experiments. We report our experimental results on the test-dev set and the test-std set of VQA 2.0, and the test-dev set of GQA.

SCAVQAN based on the thresholds of the input question words We denote SCAVQAN based on the thresholds of the input question words as SCAVQAN-Q. SCAVQAN-Q uses threshold-based multi-head scaled dot-product attention when modeling the self-attention of the input questions, while it uses traditional multi-head scaled dot-product attention when learning the self-attention of the input images and the question-guided visual attention. We conduct ablation experiments for the threshold $T \in \{0.05, 0.06, 0.07, 0.08, 0.09, 0.10, 0.11, 0.12\}$ of the input question words on VQA 2.0, and the experimental results are shown in Table 1 and Fig. 5. The best value of each column is shown in bold. It can be seen that with the increase of the threshold T of the input question words, the accuracies of SCAVQAN-Q on Number questions, Other questions and all types of the input questions generally increase first and then decrease. When T=0.10, the model achieves its highest overall accuracy, i.e., 70.82%.

SCAVQAN based on the thresholds of the input image regions We denote SCAVQAN based on the thresholds of the input image regions as SCAVQAN-I. SCAVQAN-I uses threshold-based multi-head scaled dot-product attention when modeling the self-attention of the input images, while it uses traditional multi-head scaled dot-product attention when learning the self-attention of the input questions and the question-guided visual attention. We

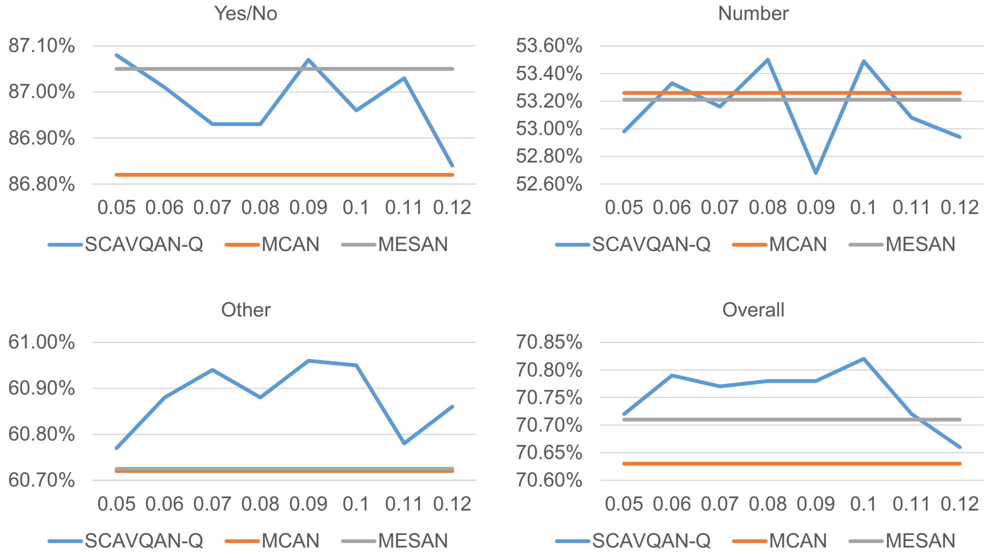


Fig. 5 Experimental results of MCAN, MESAN and SCAVQAN-Q with $T \in \{0.05, 0.06, 0.07, 0.08, 0.09, 0.10, 0.11, 0.12\}$ on the test-dev set of VQA 2.0

Table 2 Ablation experimental results of SCAVQAN-I on the test-dev set of VQA 2.0

T	Yes/No	Number	Other	Overall
0.06	86.84%	52.60%	60.89%	70.68%
0.07	86.75%	52.61%	**60.99%**	70.65%
0.08	**87.09%**	52.61%	60.84%	70.72%
0.09	86.94%	53.04%	60.78%	70.68%
0.10	86.87%	52.66%	60.92%	70.67%
0.11	87.00%	**53.31%**	60.83%	**70.76%**
0.12	86.85%	52.75%	60.92%	70.67%
0.13	86.68%	53.10%	60.75%	70.56%

conduct ablation experiments for the threshold $T \in \{0.06, 0.07, 0.08, 0.09, 0.10, 0.11, 0.12, 0.13\}$ of the input image regions on VQA 2.0, and the experimental results are shown in Table 2 and Fig. 6. The best value of each column is shown in bold. It can be seen that more accurate and significant image regions can help VQA models to better answer Other questions, but a small number of image regions can not provide enough visual information when the models face Number questions, which leads to a decrease in the corresponding accuracies. When an appropriate threshold T of the input image regions is selected, our models still achieve better overall performance. For example, when T=0.11, SCAVQAN-I achieves its highest overall accuracy, i.e., 70.76%.

SCAVQAN based on the multi-modal thresholds We denote SCAVQAN based on the multi-modal thresholds as SCAVQAN-M. SCAVQAN-M uses threshold-based multi-head scaled dot-product attention when modeling the self-attention of the input questions and the self-attention of the input images, while it uses traditional multi-head scaled dot-product attention when learning the question-guided visual attention. We conduct ablation experiments for the threshold $T_Q \in \{0.09, 0.10, 0.11\}$ of the input question words and the threshold $T_I \in \{0.08, 0.09, 0.10, 0.11\}$ of the input image regions on VQA 2.0, and the experimental results are shown in Table 3 and Fig. 7. The best value of each column is shown in bold. Since the accuracy levels of the models have similar variation curves under different thresholds, we only take SCAVQAN-M($T_Q = 0.10$) as an example to discuss the performance of the models. As can be seen from Fig. 7, with the increase of the threshold T_I of the input image regions, the accuracies of SCAVQAN-M on Other questions and all types of questions generally increase first and then decrease. When T_I=0.10, the model achieves its highest overall accuracy, i.e., 70.76%.

Based on the above experimental results, we conduct ablation studies on GQA for the threshold $T \in \{0.22, 0.23, 0.24, 0.25, 0.26, 0.27, 0.28\}$ of the input question words on our best model SCAVQAN-Q to further verify the effectiveness of our threshold-based method, and the experimental results are shown in Table 4. The best value of each column is shown in bold. It can be seen that SCAVQAN-Q performs better than MCAN when the

Fig. 6 Experimental results of MCAN, MESAN and SCAVQAN-I with $T \in \{0.06, 0.07, 0.08, 0.09, 0.10, 0.11, 0.12, 0.13\}$ on the test-dev set of VQA 2.0

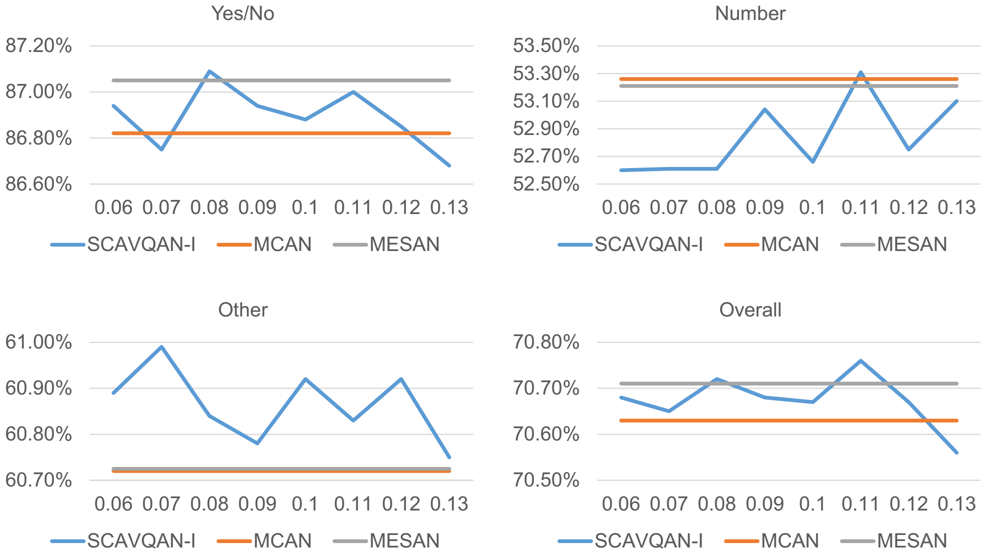


Table 3 Ablation experimental results of SCAVQAN-M on the test-dev set of VQA 2.0

T_Q	T_I	Yes/No	Number	Other	Overall
0.09	0.08	86.88%	52.55%	60.78%	70.59%
0.09	0.09	86.91%	52.98%	60.88%	70.71%
0.09	0.10	86.85%	52.95%	60.89%	70.68%
0.09	0.11	86.75%	52.73%	60.77%	70.56%
0.10	0.08	86.83%	52.68%	60.92%	70.66%
0.10	0.09	86.75%	53.10%	**61.03%**	70.72%
0.10	0.10	**87.02%**	52.82%	60.94%	**70.76%**
0.10	0.11	86.87%	52.63%	60.90%	70.66%
0.11	0.08	86.89%	52.42%	60.84%	70.61%
0.11	0.09	86.83%	**53.25%**	60.88%	70.70%
0.11	0.10	86.87%	53.09%	60.78%	70.65%
0.11	0.11	86.76%	52.76%	60.88%	70.62%

threshold value is appropriate. For example, when T=0.26, the overall accuracy of SCAVQAN-Q is 57.13%, which is 0.29% higher than that of MCAN.

As can be seen from the results of the ablation studies above, compared with the advanced co-attention network MCAN and MESAN based on top-k selection, the overall performance of our proposed models is better mainly because they achieve higher accuracies on Other questions. We think this is because when modeling the self-attention of the input questions and the self-attention of the input images, the models based on our threshold-based multi-head scaled dot-product attention can more accurately filter out the question key words and the more significant image regions to better understand the input questions and the input images, so as to more accurately answer the relatively complex Other questions.

4.4 Comparison with advanced VQA models

The accuracy results of various models on VQA 2.0 and GQA are shown in Tables 5 and 6 respectively. The best value of each column is shown in bold. In order to make a fair comparison, all the models in Table 5 are single models rather than integrated models, and all of these models are trained with the regional image features extracted by Faster R-CNN. BUTD [28] is based on bottom-up attention and won the 2017 VQA Challenge. MCAN [19] models the dense interactions between internal elements of the input questions and the input images through a stacked encoder-decoder structure. When modeling visual attention, MEDAN [29] chooses to learn the question-guided visual attention first and uses a new optimizer named AdamW to train the model. MUAN proposes a

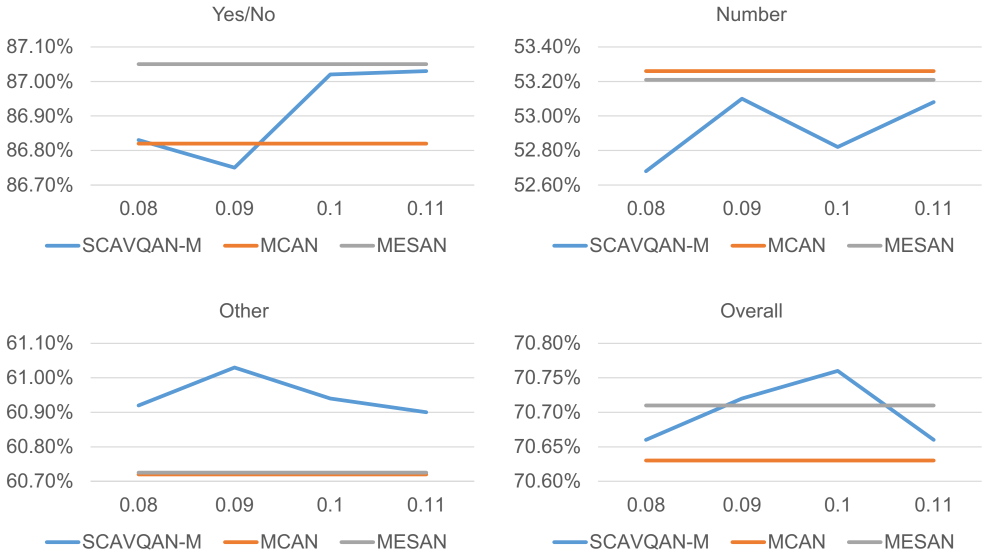


Fig. 7 Experimental results of MCAN, MESAN and SCAVQAN-M with $T_Q = 0.1$ and $T_I \in \{0.08, 0.09, 0.10, 0.11\}$ on the test-dev set of VQA 2.0

Table 4 Ablation experimental results of SCAVQAN-Q on the test-dev set of GQA

T	Accuracy	Binary	Open	Validity	Plausibility	Consistency	Distribution
0.22	56.79%	74.86%	40.90%	96.54%	84.92%	87.49%	1.20
0.23	56.36%	74.17%	40.70%	96.54%	85.09%	87.96%	1.31
0.24	57.12%	75.48%	40.97%	96.79%	85.24%	87.22%	1.13
0.25	56.99%	74.97%	**41.18%**	96.67%	**85.43%**	87.16%	1.25
0.26	**57.13%**	**76.00%**	40.54%	96.82%	84.96%	87.99%	**1.09**
0.27	56.90%	74.79%	**41.18%**	96.56%	84.92%	**88.13%**	1.25
0.28	56.27%	75.05%	39.77%	96.78%	85.28%	87.01%	1.23
MCAN [19]	56.84%	75.56%	40.38%	**96.85%**	85.32%	87.19%	1.31

Table 5 Results of our models and the state-of-the-art models on VQA 2.0

	Test-dev				Test-std
Model	Yes/No	Number	Other	Overall	Overall
BUTD [28]	81.82%	44.21%	56.05%	65.32%	65.67%
MCAN [19]	86.82%	53.26%	60.72%	70.63%	70.90%
MEDAN(Adam) [29]	**87.10%**	52.69%	60.56%	70.60%	71.01%
MUAN	86.77%	**54.40%**	60.89%	**70.82%**	71.10%
MESAN [16]	87.05%	53.21%	60.72%	70.71%	71.08%
SCAVQAN-Q(T=0.1)	86.96%	53.49%	**60.95%**	**70.82%**	**71.14%**
SCAVQAN-I(T=0.11)	87.00%	53.31%	60.83%	70.76%	71.09%
SCAVQAN-M($T_{\mathrm{Q}} = T_{\mathrm{I}} = 0.1$)	87.02%	52.82%	60.94%	70.76%	71.08%

Table 6 Results of SCAVQAN-Q and the state-of-the-art models on GQA

Model	Accuracy	Binary	Open	Validity	Plausibility	Consistency	Distribution
CNN+LSTM	46.55%	63.26%	31.80%	96.02%	84.25%	74.57%	7.46
BUTD [28]	49.74%	66.64%	34.83%	96.18%	84.57%	78.71%	5.98
MAC [30]	54.06%	71.23%	38.91%	96.16%	84.48%	81.59%	5.34
MCAN [19]	56.84%	75.56%	40.38%	**96.85%**	85.32%	87.19%	1.31
SCAVQAN-Q(T=0.25)	56.99%	74.97%	**41.18%**	96.67%	**85.43%**	87.16%	1.25
SCAVQAN-Q(T=0.26)	**57.13%**	**76.00%**	40.54%	96.82%	84.96%	87.99%	**1.09**
SCAVQAN-Q(T=0.27)	56.90%	74.79%	**41.18%**	96.56%	84.92%	**88.13%**	1.25

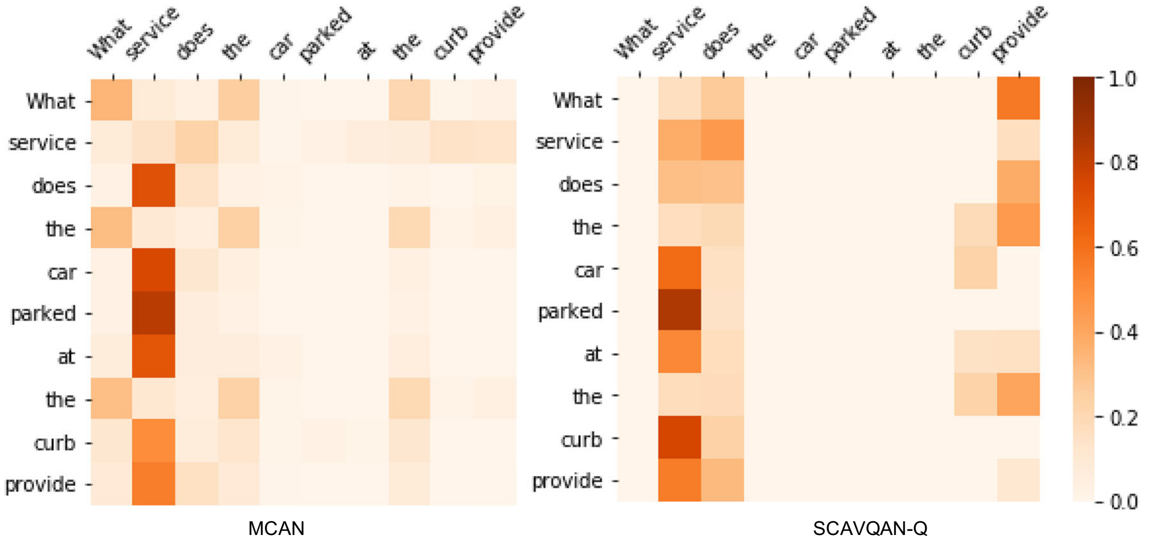


Fig. 8 Textual self-attention maps of MCAN and SCAVQAN-Q

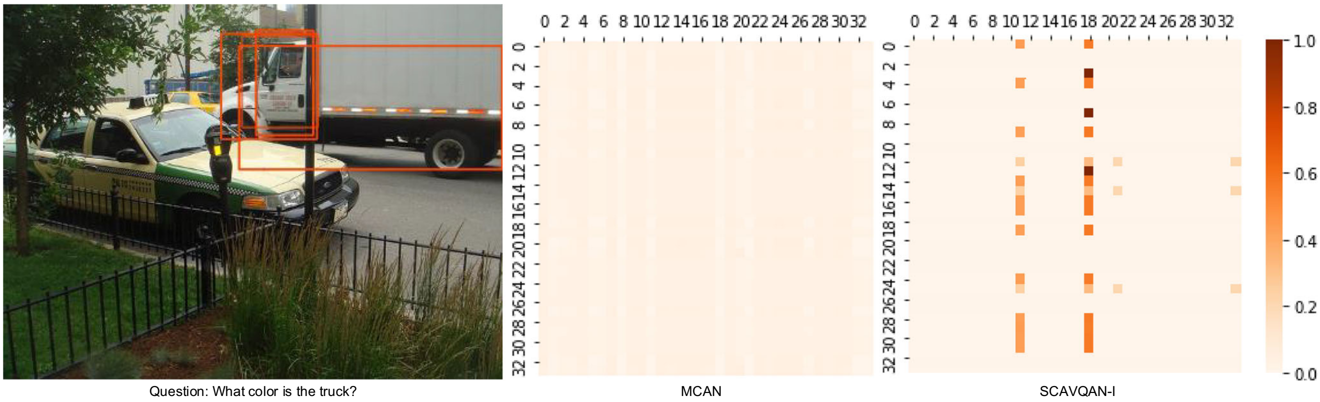


Fig. 9 Visual self-attention maps of MCAN and SCAVQAN-I

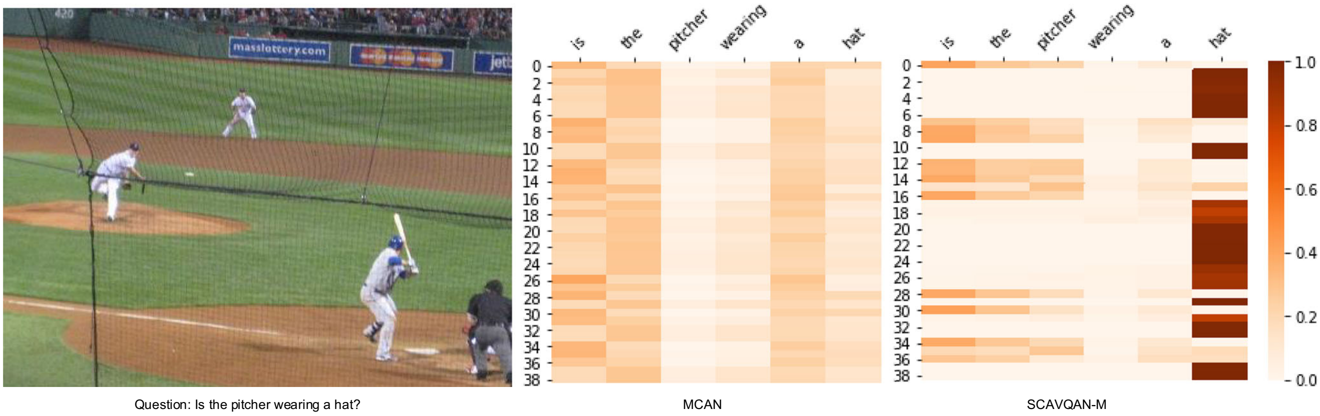


Fig. 10 Visual question-guided attention maps of MCAN and SCAVQAN-M

general unified attention module to simultaneously model the relationships between elements within and between modalities and to perform VQA task by deeply stacking this module. In order to better focus the attention of VQA models, MESAN [16] based on top-k selection explicitly filters out the input features that are the most relevant to predicting the correct answers. As can be seen from Table 5, while maintaining the accuracy levels similar to that of the advanced VQA models on Yes/No questions and Number questions, our models achieve higher accuracies on the more complex Other questions and better overall performance. In Table 6, MAC [30] approaches questions by decomposing them into attention-based reasoning steps and imposes structural constraints to perform iterative reasoning processes that are directly inferred from the data in an end-to-end approach. As can be seen from Table 6, while maintaining the validity scores similar to that of MCAN, our models perform better on all other metrics. All the above experimental results prove the effectiveness of our proposed models.

4.5 Attention visualization

In this section, we compare the attention visualization results of our models with MCAN for specific question instances and specific image instances to prove the effectiveness and interpretability of our models. Figure 8 shows the textual self-attention maps of MCAN and SCAVQAN-Q for the input question "What service does the car parked at the curb provide". It can be seen that the attention map captured by SCAVQAN-Q is more sparse. This relatively sparse attention map means that our model is more focused, and thus can infer and predict the correct answers more accurately. Figure 9 shows the visual self-attention maps of MCAN and SCAVQAN-I for the input question "What color is the truck". It can be seen that SCAVQAN-I is able to accurately focus on the image regions containing "truck", i.e., the 11th, 18th, 21st, and 33rd image boxes, while MCAN, on the other hand, diverts more attention to other image regions that are not relevant to the input question. Figure 10 shows the question-guided visual attention maps of MCAN and SCAVQAN-M for the input question "Is the pitcher wearing a hat". It can be seen that SCAVQAN-M pays more attention to the question key words "pitcher" and "hat". Moreover, since there are multiple image regions associated with the question key word "hat" in the input image, our model also focuses more attention on these image regions that are more important for correctly answering the input question.

5 Conclusion

In this paper, we propose a threshold-based Sparse Co-Attention Visual Question Answering Network (SCAVQAN) to solve the problem of distraction caused by modeling the interactions between all the input image regions and all the input question words in some advanced VQA models. SCAVQAN focuses the models' attention and achieves advanced performance by setting thresholds for attention scores to filter out the visual features and the textual features that are the most relevant to answering the input questions correctly. Experimental results, ablation studies and attention visualization results based on two benchmark VQA datasets demonstrate that our models can filter out the question key words of the input questions and the significant image regions of the input images more accurately to better understand the input questions and the input images, so as to answer the relatively complex Other questions more accurately. Our threshold-based attention mechanism can be applied to any model that infuses attention mechanism and can be tried in other tasks besides VQA. In future research, we will explore more effective attention mechanisms to help VQA models better respond to the most difficult Number questions at the present stage and further promote the progress of VQA related research.

Acknowledgements This research is supported by the National Natural Science Foundation of China under Grant 61873160 and Grant 61672338, and the Natural Science Foundation of Shanghai under Grant 21ZR1426500. We thank all the reviewers for their constructive comments and helpful suggestions.

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

Declarations

Competing interests The authors declare that they have no conflict of interest.

References

1. Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J (2021) Deep learning-based text classification: a comprehensive Review. ACM Comput Surv 54(3):62:1–62:40
2. Kim J, Jang S, Park EL, Choi S (2020) Text classification using capsules. Neurocomputing 376:214–221

3. Chen K, Wang R, Utiyama M, Sumita E, Zhao T, Yang M, Zhao H (2020) Towards more diverse input representation for neural machine translation. IEEE ACM Trans Audio Speech Lang Process 28:1586–1597
4. Sun H, Wang R, Chen K, Utiyama M, Sumita E, Zhao T (2020) Unsupervised neural machine translation with cross-lingual language representation agreement. IEEE ACM Trans Audio Speech Lang Process 28:1170–1182
5. Fan C, Yi J, Tao J, Tian Z, Liu B, Wen Z (2021) Gated recurrent fusion with joint training framework for robust end-to-end speech recognition. IEEE ACM Trans Audio Speech Lang Process 29:198–209
6. Miao H, Cheng G, Gao C, Zhang P, Yan Y (2020) Transformer-based Online CTC/attention end-to-end speech recognition architecture. ICASSP:6084–6088
7. Lee K-H, Xi C, Hua G, Hu H, He X (2018) Stacked cross attention for Image-Text matching. ECCV (4):212–228
8. Wang L, Li Y, Huang J, Lazebnik S (2019) Learning Two-Branch neural networks for Image-Text matching tasks. IEEE Trans Pattern Anal Mach Intell 41(2):394–407
9. Yu N, Hu X, Song B, Yang J, Zhang J (2019) Topic-Oriented Image captioning based on Order-Embedding. IEEE Trans Image Process 28(6):2743–2754
10. Yang M, Liu J, Shen Y, Zhao Z, Chen X, Wu Q, Li C (2020) An ensemble of generation- and Retrieval-Based image captioning with dual generator generative adversarial network. IEEE Trans Image Process 29:9627–9640
11. Qiao Y, Yu Z, Liu J (2020) VC-VQA: Visual Calibration mechanism for visual question answering. ICIP:1481–1485
12. Yang C, Jiang M, Jiang B, Zhou W, Li K (2019) Co-Attention Network with question type for visual question answering. IEEE Access 7:40771–40781
13. Agrawal A, Batra D, Parikh D, Kembhavi A (2018) Don't just assume; look and answer: Overcoming priors for visual question answering. CVPR:4971–4980
14. Zhang L, Liu S, Liu D, Zeng P, Li X, Song J, Gao L (2021) Rich visual Knowledge-Based augmentation network for visual question answering. IEEE Trans Neural Netw Learn Syst 32(10):4362–4373
15. Yu J, Zhang W, Lu Y, Qin Z, Hu Y, Tan J, Wu Q (2020) Reasoning on the relation: Enhancing visual representation for visual question answering and Cross-Modal retrieval. IEEE Trans Multim 22(12):3196–3209
16. Guo Z, Han D (2020) Multi-Modal Explicit sparse attention networks for visual question answering. Sensors 20(23):6758
17. Nam H, Ha J-W, Kim J (2017) Dual attention networks for multimodal reasoning and matching. CVPR:2156–2164
18. Nguyen D-K, Okatani T (2018) Improved fusion of visual and language representations by dense symmetric Co-Attention for visual question answering. CVPR:6087–6096
19. Yu Z, Yu J, Cui Y, Tao D, Qi T (2019) Deep modular Co-Attention networks for visual question answering. CVPR:6281–6290
20. Ren S, He K, Girshick RB, Sun J (2017) Faster r-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149
21. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li L-J, Shamma DA, Bernstein MS, Li F-F (2017) Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int J Comput Vis 123(1):32–73
22. Goyal Y, Khot T, Agrawal A, Summers-Stay D, Batra D, Parikh D (2019) Making the v in VQA matter: Elevating the role of image understanding in visual question answering. Int J Comput Vis 127(4):398–414
23. Hudson DA, Manning CD (2019) GQA: A new dataset for Real-World visual reasoning and compositional question answering. CVPR:6700–6709
24. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. EMNLP:1532–1543
25. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is All you Need. NIPS:5998–6008
26. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. CVPR:770–778
27. Agrawal A, Lu J, Antol S, Mitchell M, Lawrence Zitnick C, Parikh D, Batra D (2017) VQA: Visual Question answering. Int J Comput Vis 123(1):4–31
28. Teney D, Anderson P, He X, van den Hengel A (2018) Tips and tricks for visual question answering: Learnings from the 2017 challenge. CVPR:4223–4232
29. Chen C, Han D, Wang J (2020) Multimodal Encoder-Decoder attention networks for visual question answering. IEEE Access 8:35662–35671
30. Drew A, Hudson CD (2018) Manning. Compositional Attention Networks for Machine Reasoning. ICLR (Poster)

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

Zihan Guo received his B.E. degree from Tianjin Polytechnic University, China, in 2017. He is currently a Ph.D. student at the College of Information Engineering, Shanghai Maritime University, China. His research interests include computer vision, natural language processing and visual question answering based on deep learning.

Dezhi Han received his B.S. degree in applied physics from the Hefei University of Technology, Hefei, China, in 1990, and his M.S. and Ph.D. degrees in computing science from the Huazhong University of Science and Technology, Wuhan, China, in 2001 and 2005, respectively. He is currently a Professor with the Department of Computer, Shanghai Maritime University, Pudong, China, in 2010. His research interests include visual question answering, cloud and outsourcing security, wireless communication security, network, and information security. He is currently a Member of the IEEE.