Journal of Engineering Science and Technology Review 14 (6) (2021) 116 - 123

JOURNAL OF
Engineering Science and Technology Review

www.jestr.org

Research Article

Co-attention Network for Visual Question Answering Based on Dual Attention

Feng Dong[1,2], Xiaofeng Wang[1,*], Ammar Oad[2] and Mir Sajjad Hussain Talpur[3]

[1]College of Information Engineering, Shanghai Maritime University, Shanghai 200120, China
[2]College of Information Engineering, Shaoyang University, Shaoyany 422000, China
[3]Information Technology Centre, Sindh Agriculture University Tandojam, Sindh 70060, Pakistan

Received 13 October 2021; Accepted 27 December 2021

Abstract

Attention mechanism is a modal feature processing method widely used in visual question answering (VQA) tasks. However, the attention bias may lead to the misalignment of key targets between modalities, which reduces the accuracy of VQA tasks. A co-attention network with dual attention mechanism was proposed to accurately align the key target between image and text modalities. First, the dual attention mechanism was used to accurately localize key targets within the modality. Then, the co-attention was employed for continuous fusion of image and text features. Finally, the key target alignment between modalities was achieved. A large number of experiments verified the validity of this model. Results demonstrate that the dual attention mechanism can accurately locate the target within the modality based on the existing attention. The modal fusion of image-guided text and text-guided image co-attention improves the alignment of key targets between modalities to some extent. Compared with the overall performance of several existing classic VQA models, that of the proposed model is improved by 0.14%–5.69%. This study provides some references for improving the performance of VQA tasks by target alignment between image and text modalities.

Keywords: Multimodal feature processing, Visual question answering, Co-attention network, Dual attention mechanism

1. Introduction

Visual question answering (referred to as VQA hereinafter) is a challenging task in artificial intelligence, and it covers many aspects, including target detection [1], target identification [2], fine-grained analysis [3], and natural language processing (NLP) [4]. Among existing methods, VQA is defined as an algorithm that utilizes deep learning network to take images and question text about the images as input and generate natural language answers as output. The key to the whole process lies in the extraction and fusion of critical image and text features. For this purpose, the attention mechanism has absolute advantages in the field of VQA. In mathematical form, the attention mechanism simply assigns weighting parameters to the inputs according to their importance. Considering that human beings are dealing with multimodal tasks, it is a process of using attention to continuously locate the key targets of each modality through the interaction between modalities [5]. By simulating the cognitive model of human brain, that is, focusing limited attention on critical parts of features according to actual needs, the understanding ability of neural networks is significantly enhanced.

With the continuous development of attention mechanism, various problems continue to emerge. Chief among them is that most of the methods use question-guided attention, that is, using the text of the question to focus on the image region unidirectionally. In other words, image-guided attention to text is ignored. The resulting linguistic bias further induces attention bias [6]. Attention bias refers to the fact that the process of computing attention does not really focus on the main features. According to the different modalities, it can be divided into image attention bias and text attention bias. Inside the modality, the attention focuses on the incorrect feature regions. Between modalities, the key regions of image is not aligned with the key regions of text. The abovementioned attention bias will affect the accuracy of the question answering of the VQA task.

On this basis, many studies have been conducted on how the VQA task utilizes attention to precisely locate the key regions of features and feature interactions between images and text [7-10]. However, the alignment of text features with image features' key regions is still not really solved. Therefore, how to use the interaction of features between modalities to accurately locate and align the key regions of the image and text modalities under the guidance of the attention mechanism has become an urgent problem to be solved.

To this end, this study performs precise localization of image and text features by implementing dual attention mechanism in the transformer structure with multi-head attention [11]. Image and text features guide each other to establish a co-attention model for aligning key feature regions and improving the accuracy of question answering.

2. State of the art

Many studies have been conducted on improving the performance of VQA tasks. At this stage, most of the innovation comes from study on feature fusion and model reasoning ability. Considering that the VQA task needs to semantically analyze and understand the modal features of images and texts, the fusion and alignment of inter-modal features is critical to solving the problem.

[*] E-mail address: xfwang58@163.com

ISSN: 1791-2377 © 2021 School of Science, IHU. All rights reserved.
doi:10.25103/jestr.146.13

Among the methods based on feature fusion, the most common are linear fusion (element-to-element concatenation, addition, or multiplication) and bilinear pooling. Lin [12] applied bilinear pooling to fine-grained visual recognition tasks. Specifically, two parallel CNNs extract the features of the target and the position of the target in the image, respectively, and then, the two features were fused through a bilinear pooling method. Finally, the fused vector was classified according to the answer. When two vectors represent features of different modalities, two modalities could exhibit fused interactions. However, bilinear pooling models required a large number of parameters. Therefore, subsequent models applying bilinear pool utilized parameter decomposition to optimize model size and computational efficiency. Fukui [13] proposed an MCB model that maps the outer product to a low-dimensional space, which could avoid calculating the outer product directly. Count Sketch was used to map feature vectors. At the same time, the fast Fourier transform replaced the convolution of the two feature vectors of image and text. In theory, the MCB model was indeed feasible. However, such a model might suffer from feature loss during the high-dimensional to low-dimensional mapping process. The BLOCK model proposed by Ben was based on the Tuchker decomposition of tensors, which transformed a three-dimensional matrix into three two-dimensional matrices and a core tensor by factorization [14]. This method achieved great results in the VQA task. However, the result of factorization might lead to the loss of important features of images and texts. With bilinear attention matrix, Kim [15] introduced bilinear fusion into matrix fusion to propose bilinear attention networks (BAN). On the one hand, BAN could better utilize full contextual and image features. On the other hand, this way leaded to a huge increase in computational workload. Do [16] reduced the computational cost by introducing a trilinear fusion model into the VQA task and using knowledge distillation, which was essentially the same as the bilinear model.

Feature fusion requires access to key feature information. In this regard, the attention mechanism works well. The attention mechanism, derived from the selective attention of human vision, was originally used in machine translation tasks and was later extended to the field of computer vision. In VQA, the question-guided image attention mechanism emerges first. In the SAN model proposed by Yang [17], image attention was represented as a layer of weight distribution generated by the SoftMax function after being guided by the question and fusing visual features. A vector representation of the image was obtained by weighted summation of the features of each region of the image. The image and question features through attention mechanism were fused and a SoftMax classifier was used to predict the answer distribution. Through multiple iterations of attention, the model was empowered to solve complex problems. However, the simple feature fusion process results in that key feature between modalities could not play a decisive role in the question answer. Komal [18] achieved the goal of correcting visual attention by explicitly training a model to learn the salient parts of the images available in the VQA-HAT dataset. Through attention, the key information of text and images was aligned to some extent. However, the interaction of the two modalities, text and image, was ignored. Modi [19] used an attention-based VQA method to deal with occluded object counts. In detail, the method generated answers by extracting image and text features and applying a multi-layer attention mechanism. Essentially, the precision for counting-related problems was improved. However, the answering accuracy for other types of questions in VQA still needed to be improved. Patro [20] implemented multiple images to provide a question answering based on the different attention regions of the image obtained by one or more supporting and dissenting samples. However, the obtained question answering was not general and generalizes poorly.

The abovementioned results are all question-guided attention mechanisms. However, this attention mechanism does not fully consider the influence of image features on text features. The co-attention mechanism is generated to improve the answering effect of question. Self- and guided-attention based on transformer framework can promote the development of co-attention mechanism. The MCAN model proposed by Yu [21] illustrated the basic idea of the co-attention mechanism, which was essentially a question-guided attention network. The guided-attention module only had the guidance of the question feature to the image feature, and no guidance process of the image feature is considered to the problem feature. Yang [22] proposed a new co-attention mechanism. For a given question, more important words should be assigned larger weights. Among them, the obtained weights guided the visual attention of the image. Gao [23] proposed a multi-path pyramid co-attention (MPC) structure to capture different feature information. Given that each attention branch of the original co-attention mechanism did not interact with other attention branches, the MPC mechanism was extended to a cascaded pyramid transformer co-attention (CPTC) module, which solved the interaction between modalities within co-attention problem. However, the MPC and CPTC modules had complex structures and a large number of parameters, which leaded to instability in the training process. Chen [24] applied a co-attention mechanism to combine structural and sequence models for further capture word-level interactions. Neural network-based affinity matrices were used to derive mutual attention weights between semantic and syntactic representations. Ultimately, this co-attention mechanism yielded fine-grained analysis results at the textual semantic level. However, the contribution of image fine-grained features to text semantics was still ignored.

Apart from co-attention mechanism, attention mechanism has also been widely explored. Messina [25] proposed a novel approach called Transformer Encoder Reasoning and Alignment Network (TEARAN). Based on multi-head attention, TEARAN focuses on scalable cross-modal information retrieval, aiming to maintain a good separation of image and text data pipelines. Notably, this approach ignores the interaction of the image and text data pipelines. Guo [26] associated images and questions by computing the similarity of each object–word pair in the feature space. The answer information to the question was used to learn the model to re-engage the corresponding visual objects in the image and reconstruct the initial attention map. Farazi [27] proposed a question-independent attention mechanism,which complemented existing question-dependent attention mechanisms. By modeling and parsing object instances, the mapping relationships in the "object graph" were used for visual features to generate question-agnostic attention features. Without question-specific training involved, the model could be embedded into existing VQA models as a plug-and-play module. However, VQA performance could only be greatly improved when used in conjunction with a question-related attention mechanism.

For the study of question-guided attention and co-attention model, the important information of text and image features is fully considered, but the fusion process is an encoder and decoder structure dominated by the contextual information of the natural language text. Attention bias can easily lead to misalignment of key features in the absence of equal interactions between the two modalities. Therefore, dual attention is added to the transformer structure. Through the self- and guided-attention modules, the end-to-end multi-modal interaction co-attention model is built to solve the problem of key information alignment and attention bias in the process of modal fusion. The effectiveness of this model is proved through a large number of ablation experiments, which provides a theoretical basis for the practical application of subsequent VQA techniques.

The remainder of the study is organized as follows. Section 3 describes this study proposed VQA deep learning model, including feature representation, feature fusion, and question answering prediction. Among them, the emphasis is on the transformer-based dual attention module structure and the co-attention bolck. In Section 4, the number of attention heads and co-attention blocks that the model proposed in this study should have is obtained through extensive ablation experiments. The visualization of the results demonstrates the state of the art of the model. The last section summarizes the proposed method and provides an outlook for future study.

3. Methodology

The general VQA deep learning network model is divided into three modules: image and text feature representation module, feature fusion module, and answer prediction module. The presentation of the model structure proposed in this study will follow this basic structure.

3.1 Feature representation module

The primary task of the VQA is to represent the input image *I* and the question text *Q* for fusing and aligning the features of the two modalities. Consistent with the purpose of the target detection task, the target features in the input image must be obtained. On the basis of the Faster R-CNN network trained on the Visual Genome dataset, this study uses the bottom-up method to extract the target features in the image *I*. To facilitate network training, the number of targets per image is between [10, 100]. Among them, each target feature is obtained by mean-pooling of the convolutional layer. The image is represented as a feature matrix *X* as follows:

$$X = Fast - RCNN(I) \tag{1}$$

The representation of the question text *Q* is as follows. First, the input question sentence is divided into words, and each sentence has no more than 14 words. Thereafter, a 300-dimensional GloVe word embedding method is used to convert words into vectors. Next, the vectors are fed into a bidirectional LSTM network. Finally, the question feature matrix *Y* is obtained. By contrast, the bidirectional LSTM network captures the dependencies between each word better than the unidirectional LSTM. The formula is expressed as follows:

$$Y = Bi - LSTM(Q) \tag{2}$$

The network cannot adaptively handle matrices of different sizes. For the problem that the number of objects in image *I* does not match the number of words in question *Q*, 0 is used for padding. The number of words per question is padded to 14. The structure of the feature representation module is shown in Figure 4.

3.2 Feature fusion module

3.2.1 Dual attention mechanism

The dual attention mechanism is based on the transformer framework. The classic transformer structure was originally used in NLP, which has been widely used in image processing in recent years. Specifically, the image feature *X* and text feature *Y* obtained by the feature representation module generate three vectors through linear transformation, which are query vector *Q*, key vector *Q*, and value vector *V*, respectively. The calculation formula of scaled dot product attention is as follows:

$$Attention(Q,K,V) = soft\max(QK^T \big/ \sqrt{d_k})V \tag{3}$$

where d_k represents the dimension of the key vector. Multiple attention heads are generated through the parallel computation of multiple scaled dot product attention. The formula for concatenating the results of multi-head attention is as follows:

$$\begin{aligned} &F = MultiHead(Q,K,V) = \\ &Concat(head_1,...,head_h)W^O,\ i \in [1,h] \end{aligned} \tag{4}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{5}$$

where W_i^Q , W_i^K and W_i^V represent the mapping matrix of the i-th attention head; W^o represents the mapping matrix of the multi-head attention of the concatenation function; *h* represents the number of attention heads; *F* represents the result of the multi-head attention.

Dual attention is performed after the multi-head attention is completed. According to the residual structure, the feature *F* and the original query vectors *Q* are linearly transformed twice and their dimensions are unified. The first linear transformation formula is as follows:

$$S = W_q^S Q + W_v^S F + b^S \tag{6}$$

where W_q^s and W_v^s are the mapping matrices of *Q* and *F*, respectively; b^s is the bias variable; *S* is the result of linear transformation. Similarly, the second linear transformation formula is as follows:

$$G = \sigma(W_q^g Q + W_f^g F + b^g) \tag{7}$$

where W_q^g and W_q^g are the mapping matrices of *Q* and *F*, respectively; b^g is the bias variable; σ is the sigmoid function. The value range of the feature *G* generated by the sigmoid function in Formula (7) is 0 to 1, which is the attention gate value. This gate value is again multiplied element-wise by the result *S* of Formula (6) to produce the dual attention result F'. The formula is expressed as follows:

$$F' = G \otimes S \tag{8}$$

where ⊗ represents element-wise multiplication.

In the whole process of dual attention mechanism, the first attention is conducted by scaling dot product attention based on the global image and text features to capture the global important features. However, these global features are not further refined. Specifically, scaled dot product attention may find the approximate range of pivotal features. When multiple objects are included in the scope, only a few pivotal object features play a decisive role in the outcome of the VQA task. At this point, the dual attention operation plays a key role. The attention gate value produced by the sigmoid function is used to refine the first attention results for finding the more pivotal object features. Dual attention does not use the SoftMax function again because the gate value from 0 to 1 generated by the sigmoid function only reduces the proportion of some noncritical object features and does not completely remove them.

The entire dual attention process is represented as a dual attention block (DAB), as shown in Figure 1. The rest is the same as the transformer structure. Through the residual again, the query vector *Q* is added element-wise with F'. The result is subjected to LayerNorm operation, and the formula is as follows:

$$F'' = LayerNorm(F' \oplus Q) \tag{9}$$

where ⊕ represents the element-wise addition operation, and LayerNorm represents the level normalization operation, with the purpose of reasonably distributing features.

Pointwise feed forward is the basic structure for connecting LayerNorm, which consists of two linear transformations and a ReLU activation function. Two linear transformations flank the ReLU to ensure consistency of the input and output dimensions. ReLU can reduce the interdependence of parameters and avoid overfitting problems. The formula is as follows:

$$Z = PFF(F'') = Relu(F''W_1 + b_1)W_2 + b_2 \tag{10}$$

where W_i and b_i represent the weight coefficient and bias variable, respectively. *Z* is the output of pointwise feed forward.

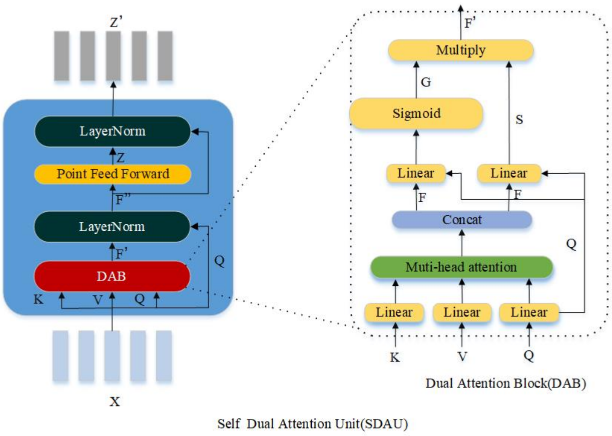
The final step repeats Formula (9), expressed as follows:

$$Z' = LayerNorm(F'' \oplus Z) \tag{11}$$

where Z' is the output of the entire transformer structure with dual attention mechanism.

Notably, when a single modal feature is input to the transformer structure with dual attention function, the structure is called a self-dual attention unit (SDAU). The structure is shown in Figure 1.

When the input is two modal features (image feature *X* and text feature *Y*), the structure is called a guided dual attention unit. The structure guided by image feature *X* is called image guided dual attention unit (IGDAU), as shown on the left of Figure 2. The structure guided by the question text feature *Y* is called a question guided dual attention unit (QGDAU), as shown on the right of Figure 2. At this point, image and text features are used as query vector *Q* for guided attention.



Self Dual Attention Unit(SDAU)

Fig. 1. Self-dual attention unit (SDAU)

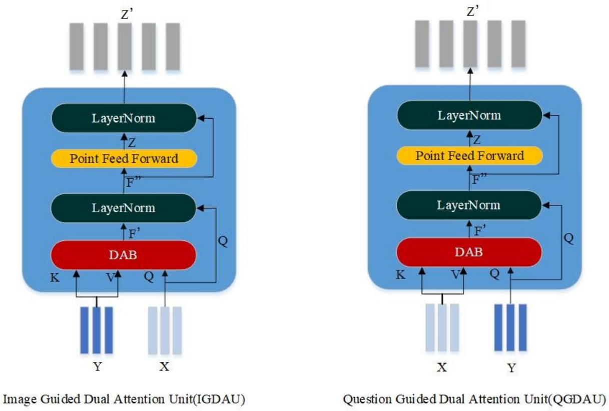


Image Guided Dual Attention Unit(IGDAU)
Question Guided Dual Attention Unit(QGDAU)

Fig. 2. Guided dual attention unit

3.2.2 Co-attention mechanism

The co-attention mechanism is constituted by the cascaded interaction of SDAU, QGDAU, and IGDAU. The specific process is as follows. The information generated after feature *Y* is input into SDAU is used as the input of IGDAU and QGDAU. The information generated after feature *X* is input to SDAU is used as the input of QGDAU. At the same time, the output of QGDAU is used as the input of IGDAU. The co-attention block (CAB) is thus formed, as shown in Figure 3.

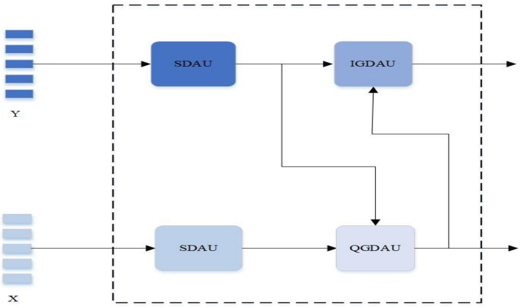


Fig. 3. Co-attention block (CAB)

Only question text features guide the attention of image features in the general co-attention mechanism, whereas text and image features guide the attention operation of each other in CAB. The two modal features are continuously fused through the cascade of multiple CAB structures, and finally, the key information between the modalities is aligned. At the same time, CAB fully simulates the process of human processing VQA task. Specifically, the interaction between the keywords of the text and the key targets in the image is used to make one-to-one correspondence. In this process, the key targets in the image are reacted to the text to explore the textual description of the information and state related to the image objects. This bidirection interaction process will play a decisive role in the results of subsequent question answering.

3.3 Answer prediction module

Through the cascade of multiple CABs, the processing result of the text branch is feature Y', and the processing result of the image branch is feature X'. In the question answering prediction stage, the two features are sequentially processed by multi-layer perceptron (MLP) and SoftMax function. Y' and X' are respectively subjected to residual operation to obtain features Y'' and X'', as shown in Equations (12) and (13).

$$Y'' = soft\,max(MLP(Y'))\otimes Y' \tag{12}$$

$$X'' = soft\,max(MLP(X'))\otimes X' \tag{13}$$

where $\otimes$ stands for element-wise multiplication.

On the basis of binary cross entropy function BCEloss, Y'' and X'' respectively perform full connection (FC) operation and element-by-element addition to obtain the final answer prediction, as shown in Formula (15).

$$P = BCEloss(FC(Y'')\oplus FC(X'')) \tag{14}$$

where P is the prediction result (probability value). The answer prediction module is shown in Figure 4.

3.4 Overall network structure

The feature representation, feature fusion, and answer prediction modules form a complete network structure, that is, co-attention with dual attention network (CADAN), as shown in Figure 4.

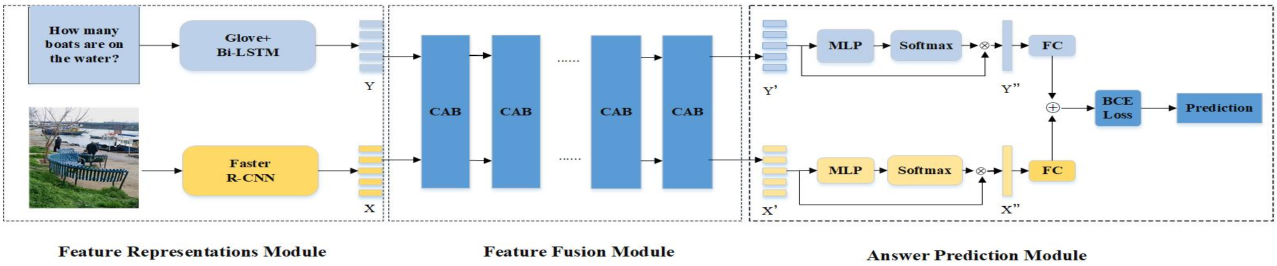


Fig 4. Co-attention with dual attention network(CADAN)

Considering the end-to-end structure of this deep learning network, the multiple cascades of CAB can achieve the alignment of multimodal key features. At the same time, the application of the dual attention mechanism can effectively focus on the key features inside the modality. In the subsequent results analysis and discussion sections, ablation experiments and visualization methods are used for validation.

4 Result Analysis and Discussion

Ablation and validation experiments based on the VQA-v2 dataset are conducted to verify the effectiveness of the CADAN. The specific number of CAB in feature fusion module is determined on the basis of the results of ablation experiments. All experimental procedures are performed on a server with RTX 2080Ti GPU installed.

4.1 VQA-v2 dataset

The VAQ-v2 dataset is used for training and validation. This dataset contains 204,721 images and 1,105,904 related manually annotated question–answer pairs. Three subsets are considered, namely, training set (82783 images and 443757 question–answer pairs), validation set (40504 images and 214354 question–answer pairs), and test set (81434 images and 447793 question–answer pairs). A total of 491809 processed Visual Genome question–answer pairs are added to facilitate the training process. Two online test sets, called test-dev and test-standard, are considered to evaluate the experimental results more accurately. All questions fall into three categories: yes/no, number, and other. Among them, the answer with the highest probability is considered to be the correct answer. The results are displayed as a percentage.

4.2 CAB ablation experiment

To determine the number N of CABs, N is set to 2, 4, 6, and 8 options under limited experimental conditions. The number of heads in multi-head attention is set to 2, 4, 8, and 16. A total of 16 ablation experiments are considered according to the number of CABs and the number of attention heads. The test-dev results of this process are shown in Table 1-4.

Table 1. Ablation experiment based on N=2

	overall	other	yes/no	number
2head	68.79	58.93	84.92	51.90
4head	69.14	59.06	85.04	52.02
8head	**69.93**	**59.67**	**85.32**	**53.10**
16head	69.19	58.53	85.18	52.97

Table 2. Ablation experiment based on N=4

	overall	other	yes/no	number
2head	68.98	58.96	84.96	52.07
4head	69.74	59.50	85.89	52.21
8head	**70.22**	**60.09**	**86.38**	**53.29**
16head	70.01	60.02	85.98	53.10

Table 3. Ablation experiment based on N=6

	overall	other	yes/no	number
2head	69.71	59.48	85.86	52.19
4head	70.25	60.51	86.93	53.44
8head	**71.01**	**61.33**	**87.07**	**53.97**
16head	70.29	60.67	86.98	53.63

Table 4. Ablation experiment based on N=8

	overall	other	yes/no	number
2head	69.52	59.22	85.26	51.98
4head	70.13	60.71	86.57	52.89
8head	**71.28**	**61.01**	**86.88**	**53.12**
16head	69.89	59.52	85.90	52.34

Among the four sets of experiments, the highest performing values are presented in bold. When the number of attention heads is 8, the accuracy rates of all categories of questions are the highest. Therefore, the number of attention heads is set to 8. At the same time, as shown in the four tables, with the change of the number N of CABs, under the condition of the same number of attention heads, the accuracy of various categories of questions increases as N increases. The performance index peaks when N is 6. But Table 4 shows that performance starts to degrade when N is 8. So the number N of CABs is set to 6. Figure 5 shows a column chart of the performance change for various categories of questions caused by the number of CABs when the attention heads is 8.

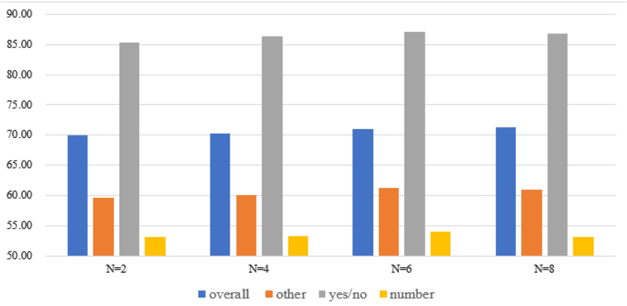


Fig 5. Performance of various types of questions with the changes of CAB

Figure 5 shows that the accuracy rates of other, yes/no, and number of the three categories of questions, as well as the overall indicators, reach the highest when N is 6. Through ablation experiments, the number of CABs is determined to be 6, and the number of attention heads is determined to be 8.

Experiments show that dual attention in the feature fusion module is focused on intra-modal and inter-modal features. The interaction of modal feature information is a critical process. However, the performance is not better when the number of CABs is higher. If the key features of the image and text are aligned, then continuing the CAB operation will lead to deviations in the aligned features. As a result, the performance of the model decreases.

4.3 Confirmatory experiment

In this study, the main purpose of the dual attention mechanism is to align key targets in images and texts. This way verifies the effectiveness of the dual attention mechanism, keeps the overall structure of CADAN unchanged, and removes the dual attention mechanism in the transformer framwork. This replacement model is named co-attention with primary attention network (CAPAN). The test-dev results of the CADAN and CAPAN comparative experiments are shown in Table 5.

Table 5. Comparative experiment of CADAN and CAPAN

	overall	other	yes/no	number
CADAN	71.01	61.33	87.07	53.97
CAPAN	70.05	60.49	85.73	52.98

According to Table 5, CAPAN has decreased in all metrics compared with CADAN. Therefore, validation experiments demonstrate the effectiveness of the proposed dual attention mechanism. The superiority of CADAN is further highlighted through the visualization of the results. The visualization results are shown in Figure 6. Notably, only the objects that are decisive for question answering are annotated in the visualization results.

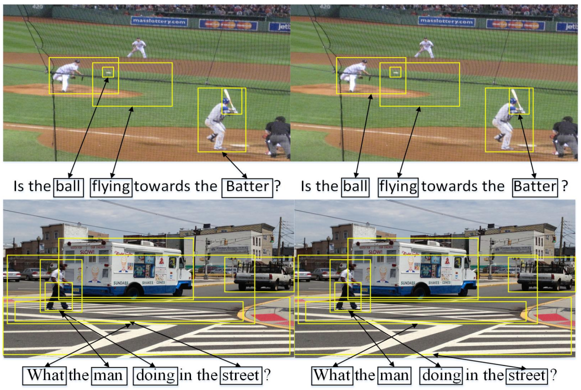

Is the ball flying towards the Batter ?
Is the ball flying towards the Batter ?
What the man doing in the street ?
What the man doing in the street ?

Fig 6. Rendering of key target alignment between modalities

The left part in Figure 6 is the CADAN visualization with dual attention mechanism, and the right part is the CAPAN visualization. In the first row, the key words of the question text on the left are aligned with the key targets of the image, but the word "batter" in the right part is misaligned with bat in the image. In the second row, the critical to question answering is the zebra crossing on the street where pedestrians are located. The word "street" on the left pinpoints the target box of the zebra crossing, while the image on the right locates the sidewalk with traffic signs. Therefore, attention bias occurs in the right picture in the CAPAN model, while no attention bias occurs in the left column, which fully illustrates the role of the proposed dual attention mechanism. In the ablation and verification experiments, the accuracy of the "number" category is between 51% and 54% when many objects are present in the picture and multiple objects cover each other. There is still plenty of room for improvement.

4.4 Comparison with existing models

Existing models have achieved good results, and CADAN also has decent performance advantages compared with these classic models. Bottom-up is the first model to employ target detection features. MFH consists of a cascade of multiple MFN modules and represents a generalized multimodal decomposition high-order pooling. BAN makes full use of image and textual information through structural bilinear interaction. The purpose of MuRel is to utilize the MuRel unit for automatically inferring the source language to express the information interaction between the question and image regions. DCN improves the fusion of vision and language through dense interactions between different modalities. DFAF obtains high-level information between visual and linguistic domains through the fusion of intra-modal and inter-modal information flow. Table 6 shows that CADAN is 5.64, 0.96, 2.92, and 0.97 greater than Bottom-up, BAN+counter, MuRel, and DFAF on Test-Std, respectively. MCAN and MCAoAN are all both based on a multimodal fusion co-attention mechanism with transformer framwork. The difference is that MCAoAN has multiple reinforcement attention. In the transformer structure-based model, CADAN improves by 0.41 and 0.15 compared with MCAN and MCAoAN, respectively. The results show that the dual attention mechanism in the self-attention unit and guided attention unit in CADAN and the interaction of two modal features in CAB can achieve precise alignment of key targets between modalities. However, CADAN is less accurate than BAN+counter on the number problem. The reason is that the BAN+counter model focuses on the "number" category qusetion.

Table 6. Comparison of CADAN with other advanced models

Model	**Test-Dev**				**Test-Std**
	overall	other	yes/no	**number**	**overall**
Bottom-up	65.32	56.05	81.82	44.21	65.67
MFH	68.76	59.89	84.27	49.56	—
BAN	69.52	60.26	85.31	50.93	—
BAN+counter	70.04	60.52	85.42	54.04	70.35
MuRel	68.03	57.85	84.77	49.84	68.41
DFAF	70.22	57.26	86.09	53.32	70.34
MCAN	70.63	60.72	86.82	53.26	70.90
MCAOAN	70.84	61.01	86.96	53.45	71.16
CADAN	71.01	61.33	87.07	53.97	71.31

5. Conclusions

In order to improve the accuracy of the VQA task and reveal the role of the attention mechanism in image and text modalities, this study used a combination of deep learning network technology and experimental research to discuss and analyzed object alignment methods between modalities. The following conclusions could be drawn:

(1) The application of DAB can more accurately locate the object related to the answer of the VQA task in the modality, and it can solve the problem of attention bias to a certain extent to make sufficient preparations for the subsequent feature fusion.

(2) The mutual guiding attention of the image and text features of the CAB module in the feature fusion module plays an important role in the alignment of key targets between modalities. In other words, the cascade process of CAB is the process of gradually aligning the objects.

(3) The alignment of key targets between modalities actually simulates the process by which humans solve the VQA task.The answers of the VQA task are more accurate when the alignment accuracy is higher.

The features between modalities are fused with each other by combining theoretical and simulated experimental study, and finally, the effect of key target alignment based on the dual attention mechanism is achieved. The CADAN model established in this study can simplify and approximate the way humans process VQA tasks, and it has certain reference value for the subsequent application of VQA tasks to real-scene environments. A long process of improvement is still required to achieve intelligent question answering for full visual scenes due to the variety of questions in the VQA task. In addition to simulating the human attention mechanism, future work should conduct in-depth discussions on visual logical reasoning.

Acknowledgements
This work was supported by National Science Foundation of China (Grant Nos. 61872231 and 61701297).

References

1. Bosquet, B., Mucientes, M., Brea, V. M., "STDnet: A Conv Net for Small Target Detection". In: *29th British Machine Vision Conference*, Newcastle, UK: CMT, 2018, pp.253.
2. Kortylewski, A., Liu, Q., Wang, A., Sun, Y., Yuille, A., "Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion". *International Journal of Computer Vision*, 129(3), 2021, pp.736-760.
3. Rodriguez, P., Velazquez, D., Cucurull, G., Gonfaus, J. M., Roca, F. X., Gonzalez, J., "Pay attention to the activations: a modular attention mechanism for fine-grained image recognition". *IEEE Transactions on Multimedia*, 22(2), 2019, pp.502-514.
4. Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R., "FLAIR: An easy-to-use framework for state-of-the-art NLP". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Minneapolis, USA: ACL, 2019, pp.54-59.
5. Zhang, L., Lin, W., "*Selective visual attention: computational models and applications*". Singapore: John Wiley & Sons, Singapore, 2013, pp.39-46.
6. Hovy, D., Prabhumoye, S., "Five sources of bias in natural language processing". *Language and Linguistics Compass*, 15(8), 2021, pp.12432.
7. Malinowski, M., Doersch, C., Santoro, A., Battaglia, P., "Learning visual question answering by bootstrapping hard attention". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany: Springer, 2018, pp.3-20.
8. Gupta, D., Suman, S., Ekbal, A., "Hierarchical deep multi-modal network for medical visual question answering". *Expert Systems with Applications*, 164, 2021, pp.113993.
9. Das, A., Agrawal, H., Zitnick, L., Parikh, D., Batra, D., "Human attention in visual question answering: Do humans and deep networks look at the same regions?". *Computer Vision and Image Understanding*, 163, 2017, pp. 90-100.
10. Burt, R., Cudic, M., Principe, J. C., "Fusing attention with visual question answering". In: *2017 International Joint Conference on Neural Networks (IJCNN)*, Anchorage, USA: IEEE, 2017, pp. 949-953.
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, l., Polosukhin, I., "Attention is all you need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems,* Guangzhou, China, 2017, pp. 6000-6010.
12. Lin, T. Y., RoyChowdhury, A., Maji, S., "Bilinear cnn models for fine-grained visual recognition". In: *Proceedings of the IEEE International Conference on Computer Vision*, Boston, USA: IEEE, 2015, pp.1449-1457.
13. 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*, Austin, USA: ACL, 2016, pp.457-468.
14. Ben-Younes, H., Cadene, R., Cord, M., Thome, N., "MUTAN: Multimodal Tucker Fusion for Visual Question Answering". In: *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy: IEEE, 2017, pp.2631-2639.
15. Kim, J. H., Jun, J., Zhang, B. T., "Bilinear Attention Networks". In: *Thirty-second Conference on Neural Information Processing Systems*, Montreal, Canada: NIPS, 2018, pp.1571-1581.

16. Do, T., Do, T. T., Tran, H., Tjiputra, E., Tran, Q. D., “Compact trilinear interaction for visual question answering”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Long Beach, USA: IEEE, 2019, pp.392-401.

17. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.*, "*Stacked Attention Networks for Image Question Answering*".* In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Las Vegas, USA: IEEE, 2016, pp.21-29.

18. Sharan, K., Ganesan, A., Oates, T., “Improving visual reasoning with attention alignment”. In: *International Symposium on Visual Computing*, Lake Tahoe, USA: Springer, 2019, pp.219-2.

19. Modi, S., Pandya, D., “A-VQA: Attention Based Visual Question Answering Technique for Handling Improper Count of Occluded Object”. In: *The International Conference on Recent Innovations in Computing*, Singapore, Singapore: Springer, 2019, pp.331-343.

20. Modi, S., Pandya, D., “Differential attention for visual question answering”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA: IEEE, 2018, pp.7680-7688.

21. 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 International Conference on Computer Vision*, Long Beach, USA: IEEE, 2019, pp.6281-6290.

22. Yang, C., Jiang, M., Jiang, B., Zhou, W., Li, K., “Co-attention network with question type for visual question answering”. *IEEE Access*, 7, 2019, pp.40771-40781.

23. Gao, L., Chen, T., Li, X., Zeng, P., Zhao, L., Li, Y. F., “Generalized pyramid co-attention with learnable aggregation net for video question answering”. *Pattern Recognition*, 120, 2021, pp.108145.

24. Chen, Y., Wu, C., Huang, Y., “Enhancing structure modeling for relation extraction with fine-grained gating and co-attention”. *Neurocomputing*, 467, 2022, pp.282-291.

25. Messina, N., Amato, G., Esuli, A., Falchi, F., Gennaro, C., Marchand-Maillet, S., “Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders”. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(4), 2021, pp.1-23.

26. Guo, W., Zhang, Y., Wu, X., Yang, J., Cai, X., Yuan, X., “Re-attention for visual question answering”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, New York, USA: AAAI, 34(01), 2020, pp.91-98.

27. Farazi, M., Khan, S., Barnes, N., “Question-agnostic attention for visual question answering” In: *2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy: IEEE, 2020, pp.3542-3549.