Intelligent Automation & Soft Computing

Tech Science Press

DOI: 10.32604/iasc.2023.040521

ARTICLE

Improving VQA via Dual-Level Feature Embedding Network

Yaru Song*, Huahu Xu and Dikai Fang

School of Computer Engineering and Science, Shanghai University, Shanghai, 200444, China

*Corresponding Author: Yaru Song. Email: sonsgyr@shu.edu.cn

Received: 21 March 2023 Accepted: 08 June 2023 Published: 11 July 2024

ABSTRACT

Visual Question Answering (VQA) has sparked widespread interest as a crucial task in integrating vision and language. VQA primarily uses attention mechanisms to effectively answer questions to associate relevant visual regions with input questions. The detection-based features extracted by the object detection network aim to acquire the visual attention distribution on a predetermined detection frame and provide object-level insights to answer questions about foreground objects more effectively. However, it cannot answer the question about the background forms without detection boxes due to the lack of fine-grained details, which is the advantage of grid-based features. In this paper, we propose a Dual-Level Feature Embedding (DLFE) network, which effectively integrates grid-based and detection-based image features in a unified architecture to realize the complementary advantages of both features. Specifically, in DLFE, In DLFE, firstly, a novel Dual-Level Self-Attention (DLSA) modular is proposed to mine the intrinsic properties of the two features, where Positional Relation Attention (PRA) is designed to model the position information. Then, we propose a Feature Fusion Attention (FFA) to address the semantic noise caused by the fusion of two features and construct an alignment graph to enhance and align the grid and detection features. Finally, we use co-attention to learn the interactive features of the image and question and answer questions more accurately. Our method has significantly improved compared to the baseline, increasing accuracy from 66.01% to 70.63% on the test-std dataset of VQA 1.0 and from 66.24% to 70.91% for the test-std dataset of VQA 2.0.

KEYWORDS

Visual question answering; multi-modal feature processing; attention mechanisms; cross-model fusion

1 Introduction

The visual question answering (VQA) task [1] is a key challenge in artificial intelligence. VQA aims to deeply understand the semantic information of images and questions and integrate the information of these two different modes to answer questions related to images effectively. With the continuous progress of VQA tasks and related visual or voice tasks, various applications have emerged in the real world, such as intelligent navigators for visually impaired individuals, social robots to assist in treating autism patients [2], and intelligent education. After releasing many large-scale VQA datasets, including VQA v1 and VQA v2 [3], VQA received unprecedented support, resulting in hundreds of models.

At present, state-of-the-art VQA models use attention mechanisms to focus on image regions related to the question to answer the question accurately. There are two main categories of visual

attention mechanisms: detection-based methods [4] and grid-based methods [5]. In grid-based visual attention, the image is divided into several grids, and the attention is distributed among these grids. Although grid features can distribute attention to any region's size, it may focus on unrelated contexts or partial objects. For example, in Fig. 1a, "Which animal is on the grass?". The gird-based attention focuses on the part of the foreground object (dog), resulting in an incorrect answer (cat). It indicates that grid-based attention is intractable in answering questions about the precise positioning of foreground objects. Additionally, detection-based methods learn the attention distribution on the pre-specified detection box of the image, leading to improved accuracy in object localization. The most prominent regions in the image can be identified and represented by feature vectors, which can provide object-level information and reduce the difficulty of visual semantic embedding. The detection-based method can effectively answer questions about the foreground objects, significantly improving the performance of VQA. Despite the remarkable success, the detection-based methods lack fine-grained details and context information, which cannot effectively answer questions about the background form. For the question in Fig. 1b, "What color is the sky?" the detection box of the sky cannot be obtained from the image, resulting in incorrect answers. According to the above analysis, the critical challenge is to devise a robust attention allocation mechanism considering the background forms and the foreground object.

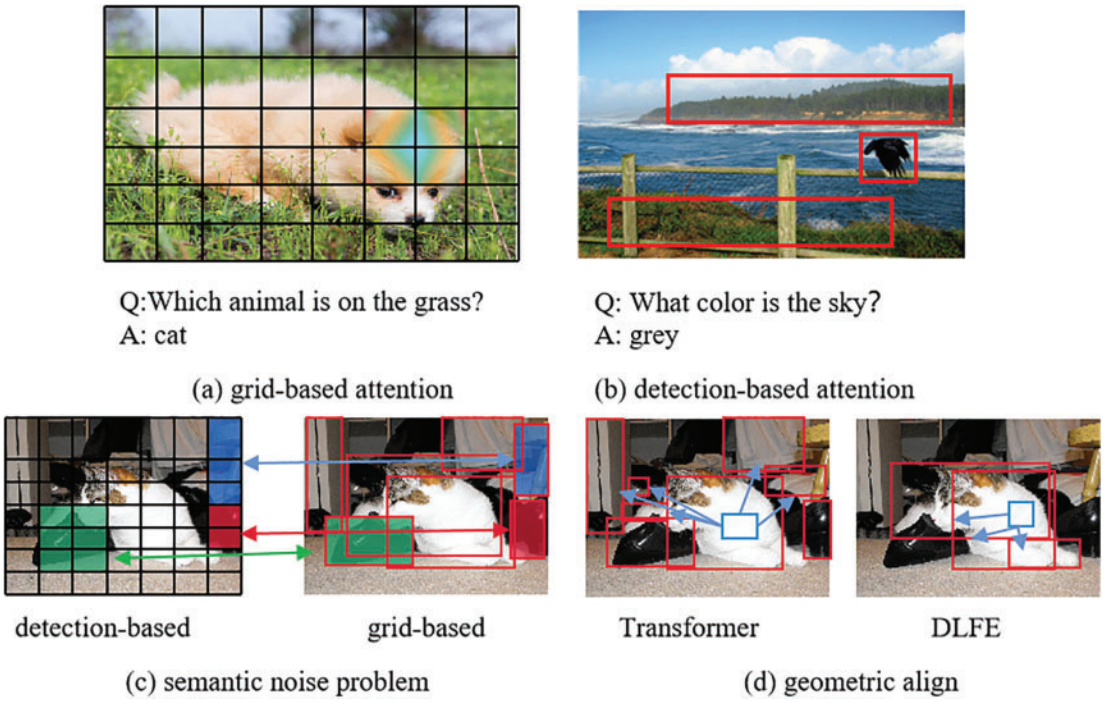

Figure 1: (a) Limitations of Grid-based attention. (b) Limitations of detection-based attention. (c) An example of semantic noise problem. (d) Geometric align

To solve this problem, Lu et al. [6] linearly combined detection-based and grid-based attention to infer the answer. However, it cannot effectively capture these two features' internal relationship and complementarity. In addition, directly utilizing the two feature sources during the attention process can lead to the creation of semantic noise quite easily. Grid features may interact with incorrect areas with similar appearances, such as the cat's abdomen and the white objects (Fig. 1c). In such a situation,

utilizing two features interferes with the complementarity of the two features and affects the overall performance. To address the above issues, we extend the idea in [4] to better utilize detection-based and grid-based features' complementary advantages, effectively answering questions about foreground objects and background forms.

In this article, we propose a novel Dual-Level Feature Embedding Network (DLFE). Specifically, firstly, the two feature sources are processed by a Dual-Level Self-Attention (DLSA) module to explore their intrinsic characteristics, where a Position Relation Attention (PRA) is introduced to embed the geometric information of the relative and absolute position relationships of the image. Then, a Feature Fusion Attention (FFA) module is further designed to address the semantic noise problem during the fusion process, where an alignment graph is constructed to align the two feature sources. As shown in Fig. 1d, in FFA, we replace the fully connected graph in the transformer with the proposed alignment graph, which can accurately fuse two feature sources. The fine-grained details and context information of grid features can be added to the detection features through FFA. The obtained image features can be used to answer questions about the background form and the foreground object.

To evaluate the proposed DLFE model, we conducted extensive experiments on the VQA v1 and VQA v2 datasets. We achieved superior performance, with the "All" score of 70.63% on the VQA v1 test-dev and 70.91% on the VQA v2 test-dev. Our contributions are concluded as follows:

(1) A new DLFE is proposed to achieve the complementarity of detection and grid features, and we have conducted many experiments on the VQA dataset to demonstrate its performance compared to other models. (2) We designed a Dual-Level Self-Attention mechanism with position relationships, which models input features' complex visual and positional relationships by introducing relative and absolute positional information of image regions. (3) We introduce FFA to fuse the two feature sources. The alignment graph constructed in FFA can reduce the semantic noise of fusion, accurately fuse two feature sources, and obtain more effective multi-view features.

This paper is divided into five chapters, and the contents of each chapter are arranged as follows: Section 2 describes the work related to visual question answers and the co-attention mechanism. Section 3 gives the proposed model in detail. Section 4 discusses the experimental results. The conclusion will be described in Section 5.

2 Literature Review

With in-depth research in artificial intelligence, more complex deep neural network models have been proposed. Research on artificial intelligence based on deep learning has entered a golden age and can be applied to various fields, such as education and medicine [7]. Most state-of-the-art VQA methods are based on deep neural networks.

2.1 Attention Mechanism for VQA

Recently, much work has focused on integrating attention mechanisms to solve VQA. The attention mechanism can make the model pay more attention to the key words in the problem and the parts related to the problem in the image.

The first mechanism is grid-based attention. Reference [8] proposes a VQA architecture with a circular attention mechanism. The recurrent layer helps guide visual and text attention because the network can infer the relationship between the image and several parts of the problem. Reference [9] introduces a novel VQA model that integrates inferential attention and semantic space mapping.

Reference [10] introduces an Object-Difference Attention model to calculate the attention weights by computing the difference vector between each pair of objects in the image.

The second mechanism is detection-based attention. The paper [4] proposes aligning the problem with the object suggestion box generated by Faster R-CNN and using top-down and bottom-up attention to learn critical objects in the image according to the given problem. Since then, the visual features extracted by Faster R-CNN have been widely used. Reference [11] improves this model [4] by using multiple techniques, such as using sigmoid output instead of traditional softmax output in the classifier, which can ensure that a question may have multiple correct answers. A gated relationship awareness mechanism is proposed in [12] to capture useful relationship features to predict correct answers. The model uses a gate mechanism to determine the required information and uses context information to expect answers. A new parse tree-guided inference network is introduced by [13], which consists of an attention module, a gated residual synthesis module, and a parse tree-guided propagation module. Reference [14] proposes a multi-step attention framework that gradually adjusts the essential visual areas through the reasoning representation of the problem and filters irrelevant information based on a multimodal correlation cross-modal gating strategy.

However, most attention-based methods focus on only one image feature (grid-based or detection-based), therefore, cannot simultaneously capture the global relationship and local details of the image. They need to improve in solving different types of problems. In contrast to these methods, our proposed method effectively integrates the two attention mechanisms in a unified framework. It can simultaneously obtain the global relationship and local details of the image.

2.2 Co-Attention Mechanism

The co-attention method considers not only the use of textual features to obtain visual features but also the importance of specific words in the question, i.e., which words are more crucial for answering the question. Reference [15] proposes a deep modular collaborative attention network connected by multiple modular collaborative attention modules in parallel. Each joint attention module focuses on problems and image self-attention through two simple attention units. Reference [16] introduces dual self-attention and co-attention networks. Self-attention mechanisms extract essential information from both the image and the text. The co-attention module is designed to capture the cross-modal correlation between the text sentences and image content. Reference [17] proposes an efficient dense co-attention network consisting of a Bi-LSTM network for encoding questions and answers to improve the accuracy of extracted semantics and better capture word relationships. Yan et al. [18] have introduced a co-attention mechanism for VQA, which includes three distinct components: visual self-attention mechanism with spatial position, text self-attention mechanism, and question-guided attention mechanism. Combining and stacking these three units can increase the depth of the model to extract more detailed visual and text features.

3 Method

The structure of our model is shown in Fig. 2. It takes the question, grid features, and detection-based features as input and learns to simultaneously associate questions with grid features and detection features to infer answers. The proposed visual representation module consists of Dual-Level Self-Attention (DLSA) and Feature Fusion Attention (FFA). The proposed DLSA based on positional relationships is used to explore the inherent characteristics of the two image features. Then, FFA is designed to address the semantic noise caused by the fusion of two types of image features, where an alignment graph is constructed to guide the semantic alignment between two feature sources.

For the question, we first use GRU and LSTM to obtain a vector representation of the question and then use a self-attention mechanism to receive dependency relationships between sentences. Finally, co-attention is used to learn the interactive features of the image and question more accurately and use the cross-entropy loss function to train the fused features.

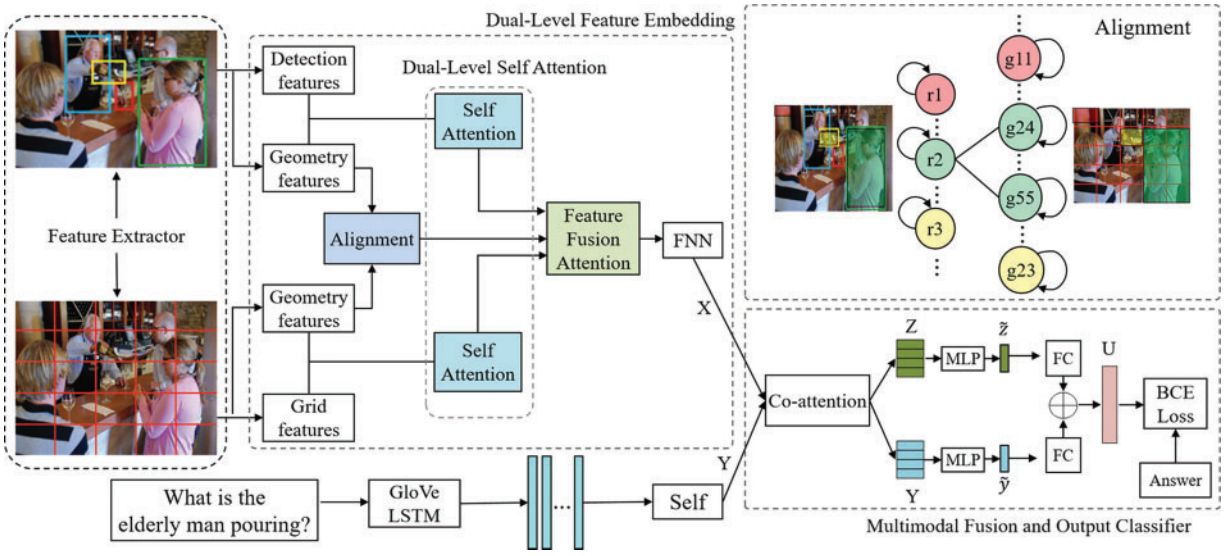

Figure 2: Structure of the proposed model

3.1 Input Feature Encoding

Grid features encoding: ImageNet pre-trained Resnet-152 is used to learn the visual features of images. The size of an image in ResNet is adjusted to 448 × 448. The last convolution layer output feature map is used to encode the whole-image visual feature $V_G = \{v_g\}^{N_G}$, where N_G represents the number of relevant features.

Detection features encoding: Faster R-CNN framework is used to obtain the detection boxes in the input image. The non-maximum suppression with an intersection over the union of 0.5 was applied to all the object's suggestions, and associated detection scores produced by Faster R-CNN and chose the top-ranked 10 detection boxes as the detection features. The visual characteristic of each detection box is encoded using the 4096-dimensional visual features of the fc7 layer of Faster R-CNN coupled with the box detection scores. Let $V_D = \{v_d\}^{N_D}$ represent the visual feature, where N_D is the number of detection boxes.

Encoding question features: For question sentences, recurrent neural networks (RNN) with sequential architecture are frequently employed to model phrases [19]. However, RNN fails to accurately capture the intrinsic connections between the words in various places. Additionally, it has a high time cost, especially for long sentences. The attention mechanism [20] captures the internal dependencies between words and enables more efficient computation. Therefore, we input the question features obtained through LSTM into the self-attention mechanism to capture the dependence within the sentence. We first use spaces and punctuation to split sentences into words. After the work of [21], we first pruned the extra words for questions with more than 14 words, and for questions with fewer than 14 words, we filled the question with zero vectors. Using pre-trained 300-d GloVe word embeddings to embed word vectors, each word is converted into a 300-dimensional vector. The size of the resulting

word embedding sequence is 14 × 300. Then, the words are embedded into a one-layer LSTM network with a hidden dimension [22], and a problem feature matrix is output. The feature matrix is input into the self-attention mechanism, and the problem features $Y = R^{n \times d_y}$ are finally obtained.

3.2 Dual-Level Feature Embedding

Our DLFE includes dual-level self-attention (DLSA) and feature fusion attention (FFA). DLSA consists of two improved self-attention mechanisms to explore each image feature's internal relationship. FFA is proposed to capture the complex interaction between detection and grid features to achieve inter-layer fusion.

3.2.1 Dual-Level Self Attention

The transformer can enrich visual features and model the relationship between inputs through self-attention. Therefore, to better model the intra-layer relationship between the two features, we designed a dual-level self-attention (DLSA) composed of two independent self-attention modules. As shown in Fig. 3a, the traditional self-attention mechanism excels at directly capturing the dependency relationship between inputs by calculating the correlation between Q and K. However, it overlooks the spatial relationship between image features, and traditional models without positional information often lead to comprehension biases due to the similarity of regional surfaces, resulting in incorrect answers. Introducing location information can help the self-attention mechanism overcome understanding biases caused by regional appearance similarity. Therefore, a self-attention mechanism based on position relation attention was constructed to reconstruct the complex visual and spatial relationships between input image features. The network structure is shown in Fig. 3b.

Absolute position provides crucial information to the model by indicating the feature's location. Consider two identically shaped objects, one on the edge and the other in the center. In this instance, absolute position helps the model differentiate them precisely. We account for grids and detection as two visual cues for Absolute position. To determine the grid absolute position (GAP), we concatenate two 1-d sines and cosine embeddings.

$$GAP(r, c) = [AP_r; AP_c] \tag{1}$$

where r and c are the row and column indexes of the grid, respectively. $AP_r, AP_c \in R^{(d_m/2)}$ are denoted as:

$$AP(pos, 2k) = sin(pos/10000^{2k/(d_m/2)}) \tag{2}$$

$$AP(pos, 2k+1) = cos(pos/10000^{2k/(d_m/2)})$$

where k denotes the dimension and pos denotes the position. For detection, we incorporated 4-d $B_i = (\mathrm{x_{left}}, \mathrm{y_{left}}, \mathrm{x_{right}}, \mathrm{y_{right}})$ bounding box into detection absolute position (DAP):

$$DAP(i) = B_i W_{emb} \tag{3}$$

where i represents the number of the boxes and (x_{left}, y_{left}), (x_{right}, y_{right}) indicate the box's top-left and bottom-right corners, respectively. $W_{emb} \in R^{dm \times 4}$ is a parameter matrix.

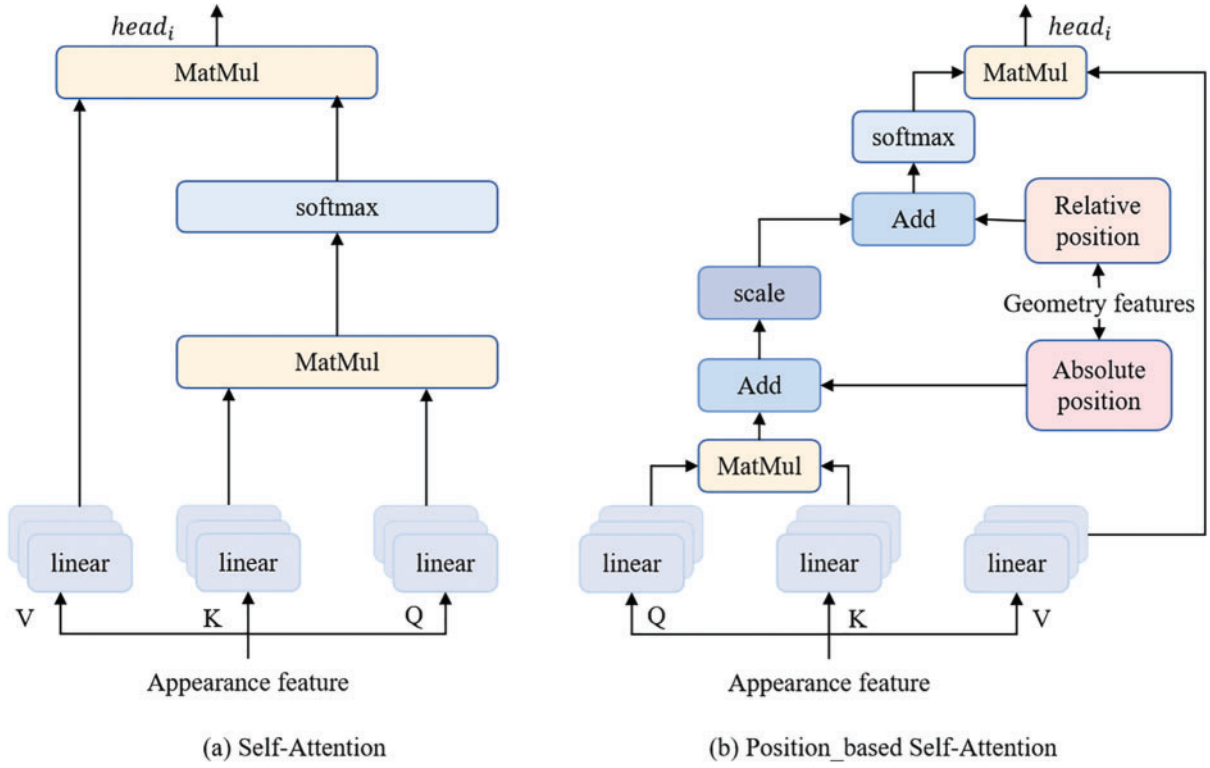

Figure 3: Self-attention to integrate position information

Relative positional represents the geometric features of the bounding box and is added to compensate for the loss of relative position information during self-attention. A target's bounding box can be represented as (W, H, x, y), where W, H, and x, y indicate the bounding box's height, width, and center coordinates. The grid is also considered a bounding box. Therefore, according to each field, the grids are also expressed as (W, H, x, y). Thus, for box_i and box_j, we can use a four-dimensional vector to describe their geometric relationship:

$$R(i,j) = \left(log\left(\frac{|x_i - x_j|}{w_i}\right), log\left(\frac{|y_i - y_j|}{h_i}\right), log\left(\frac{w_i}{w_j}\right), log\left(\frac{h_i}{h_j}\right)\right)^T \tag{4}$$

Then, using the embedding technique in [11] $R(i,j)$ is embedded in a high-dimensional embedding. $R(i,j)$ passes the geometric relationship between the two boxes, which is mapped to a scalar:

$$R(i,j) = ReLU(Emb(R(i,j))W_\Omega) \tag{5}$$

where W_Ω is a learned parameter matrix, we can combine the absolute and relative position information using positional relation attention (PRA). We changed the attention layer's query and key for the absolute position:

$$W = \frac{\left(pos_q + Q\right)\left(pos_k + K\right)^T}{\sqrt{d_k}} \tag{6}$$

where pos_q and pos_k are queries and keys of absolute position, respectively. Then we use the relative position relationship to adjust the attention weight by:

$$W_{ij}^{\cdot} = W_{ij} + log(R(i,j)) \tag{7}$$

Then, we use softmax to calculate the output of positional relation attention (PRA) and normalize the weights. Our proposed definition of multi-head PRA (MHPRA) is as follows:

$$\mathrm{PRA} = softmax\left(\frac{\left(pos_q + Q\right)\left(pos_k + K\right)^T}{\sqrt{d_k}} + log(R)\right)V \tag{8}$$

$$head_i = PRA(QW_i^Q, KW_i^K, VW_i^V, pos_q, pos_k, R) \tag{9}$$

$$\mathrm{MHPRA(Q, K, V)} = \mathrm{Concat(head_1, \cdots, head_h)W} \tag{10}$$

To learn relation-aware representation, the detection and grid hidden states $H_d^{(l)}$ and $H_g^{(l)}$ are fed into the next layer of DLSA:

$$C_d^{(l)} = MHPRA(H_d^{(l)}, H_d^{(l)}, H_d^{(l)}, DAP, DAP, R_{dd}) \tag{11}$$

$$C_g^{(l)} = MHPRA(H_g^{(l)}, H_g^{(l)}, H_g^{(l)}, GAP, GAP, R_{gg}) \tag{12}$$

where $H_d^{(0)} = V_D$ and $H_g^{(0)} = V_G$. R_{dd} and R_{gg} represent the relative position matrix of detection and the grid, respectively, and then, for two different categories of visual features, we use two separate position level feed-forward networks (FFN):

$$C_d'^{(l)} = FFN_d(C_d^{(l)}) \tag{13}$$

$$C_g'^{(l)} = FFN_g(C_g^{(l)}) \tag{14}$$

After that, the relationship-aware representation will be input into the next module.

3.2.2 Feature Fusion Attention

We propose feature fusion attention (FFA) to model the complex interactions between detection and grid features to achieve inter-layer fusion. Fig. 4 displays its architecture. An alignment graph G = (V, E) comprising two features was first built to prevent the introduction of semantic noise. We create a visual node set V by representing all detection and grid features as separate nodes. The grid nodes for the edge set E are linked to the detection feature node only when their bounding boxes intersect. Based on the above guidelines, we can build an alignment graph. According to the map of geometric alignment, we applied FFA to identify the attention of two different types of visual features.

FFA aims to enhance visual features by fusing two kinds of visual attention. We integrate the absolute position information and the relative position information through Eqs. (1) and (2) to obtain the weight matrix and normalize it:

$$\alpha_{ij} = \frac{e^{W'_{ij}}}{\sum_{j \in N(v_i)} e^{W'_{ij}}} \tag{15}$$

where v_i and $N(v_i)$ represent the visualization node and the neighbor node of the visualization node, respectively. The weighted sum is expressed as:

$$M_i = \sum_{j \in N(v_i)} \alpha_{ij}^{(l)} V_j \tag{16}$$

where V_j is the value of the j-th visualization node. In general, multi-head FFA (MHFFA) is defined as:

$$DFFA(K, Q, V, pos_q, pos_k, R, G) = \underset{G}{graph - softmax}\left(\frac{(pos_q + Q)(pos_k + K)^T}{\sqrt{d_k}} + log(R)\right)V \tag{17}$$

$$head_i = DFFA(QW_i^Q, QW_i^Q, QW_i^Q, pos_q, pos_k, R, G) \tag{18}$$

$$MHDFFA(K, Q, V) = Concat(head_1, \cdots, head_h)W^O \tag{19}$$

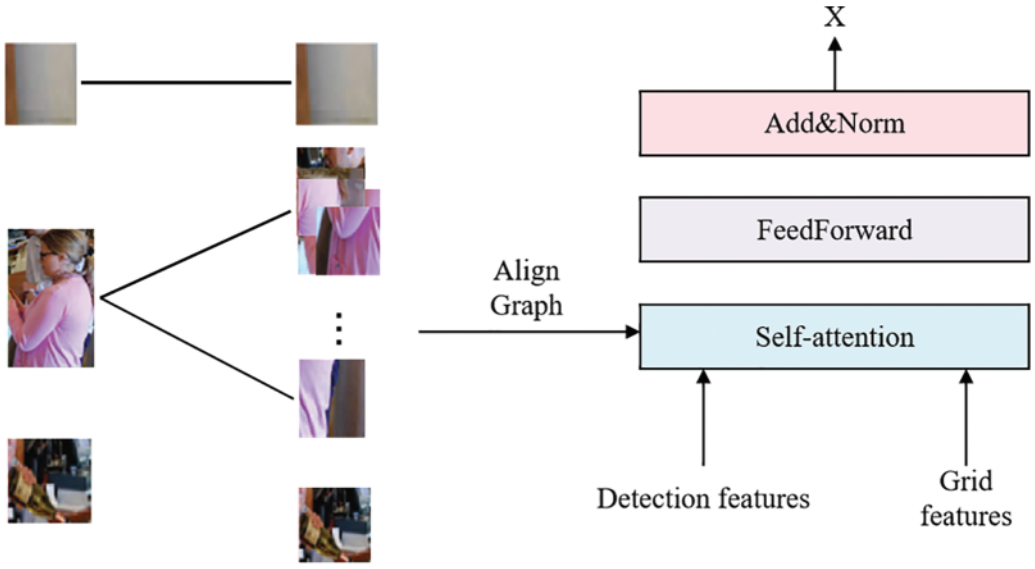

Figure 4: Architecture of the proposed feature fusion attention

After this process, the grid features and detection features are fused together, and the final output of the visual features is:

$$X = MHDFFA(C_d^{\prime(l)}, C_g^{\prime(l)}, GAP, DAP, R, G) \tag{20}$$

where R is the relative position matrix between the detection and grid. After multi-level encoding, grid-based and detection-based features are concatenated and fed into the next module. The fused features absorb the advantages of the detection and grid and can better answer questions related to the foreground object and the background form.

3.3 Co-Attention and Answer Prediction

Following the existing VQA methods [3,6,15], we modeled VQA as a multi-category classification problem. In this paper, we introduced a co-attention mechanism combining image and question features to obtain cross-modal feature representations. Specifically, for a given question feature Y and fused image feature X, we input question feature Y as k and q and image feature X as v into the co-attention mechanism. In this process, key targets in the image will respond to the text to explore features related to the text. The question guide image feature $Z = [Z_1; \cdots; Z_m] \in \mathrm{R}^{\mathrm{m\times d}}$ can be obtained through the co-attention module. In the answer prediction stage, MLP and softmax functions are used

to process the two features in sequence. We designed an attention reduction model for $Z(Y)$ and two-layer MLP to obtain the target feature $\tilde{z}$ ($\tilde{y}$):

$$\alpha = softmax(MLP(Z)) \tag{21}$$

$$\tilde{z} = \sum_{i=1}^{m} \alpha_i z_i \tag{22}$$

where $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_m] \in \mathrm{R}^m$ represents the learnable attention weight. Similarly, we can obtain $\tilde{y}$. Finally, they are input into linear multimodal fusion:

$$U = LayerNorm(W_x^T \tilde{z} + W_y^T \tilde{y}) \tag{23}$$

where $W_z, W_y \in R^{d \times d_u}$ are two linear projection matrices. d_u is the dimension of fused features. LayerNorm is used for stable training [23]. We apply a sigmoid function to map the fused feature U to a vector. Like [4], we utilize binary cross-entropy as the loss function and train an n-way classifier on the fused feature U.

4 Experiments

4.1 Datasets and Evaluation Metric

In this section, we will evaluate the performance of the DLFE model on the benchmark datasets listed below.

VQA v1 [1] is generated using the MS-COCO dataset, and crowdsourcing employees marked it up. This dataset includes three types of questions: (1) number; (2) yes/no; and (3) other. Each question and the two subtasks included in this dataset were marked as having 10 free answers. The top 1000 most frequent responses, or 82.7% of all responses, were chosen as potential outputs.

VQA v2 [3] is a large-scale public dataset of commonly used VQA tasks, and the images are also from MSCOCO. VQA v2.0 includes a train set, value set, and test set. The test set is divided into test-dev and test-std for online debugging and validation experiments. For each image, generate an average of three questions. The questions are divided into four categories: (1) overall, (2) yes/no, (3) number and (4) other. Compared to VQA v1.0, VQA v2.0 has a larger scale, and the questions in the dataset are more challenging and diverse while increasing the diversity of answers and limiting the overlap between question answers.

VQA is usually defined as a multi-category classification question, and its category label comes from a predefined set of candidate answers [24]. The creator of the datasets set up an open evaluation server to perform blind tests on the candidate answers on the test set and used a voting mechanism to calculate the accuracy rate. We use the precision metric proposed by [3], which is considered reliable in capturing individual variations in answers:

$$Accuracy(\alpha) = min\left(\frac{count\,(\alpha)}{3}, 1\right) \tag{24}$$

where $count(\alpha)$ represents the total number of answers selected by the various annotators.

4.2 Implementation Details

This section describes the model's hyperparameters that were used in the experiments. The input question features' dimension d_y, the input image feature's dimension d_x and the fused feature's dimension d_u are 512, 1024 and 1024, respectively. According to [6], the potential dimension d_m of multihead attention is 512, the number of heads h = 8, and we can calculate that the potential dimension of each head is $DH = 512/8 = 64$.

We utilized the Adam solver [25] to train the model. The model is trained for 18 epochs using a batch size of 64. Only the training set was employed for training to obtain results on the test-val split. To make training simplify it on test-dev or test-std, both the train and value sets and the VQA sample subset of the visual genome [26] were utilized for training.

4.3 Comparison with State-of-the-Art Methods

In this section, we present the performance of our model on three well-known VQA datasets and compare it with the state-of-the-art models.

Table 1 presents our results and the results of other currently popular methods on the test-dev and test-std of VQA v1. Our model outperforms other methods on most metrics, indicating significant performance improvements. More specifically, like our model, Dual-MFA [6] and ALSA [27] learn the relationship between the question and the image from different perspectives (detection boxes and free-form regions), and ALSA implements a confrontation learning mechanism between the two supervised attention modules. However, our DLFE considers both intra- and inter-modal interactions of the two image features, taking full advantage of the complementary nature of the two features, which leads to a higher performance improvement. GRA [12] uses bilinear attention for the initial interaction between visual features and question features and filters visual features irrelevant to the question, identifying a particular object within an image and obtaining the best results on the 'number' types. DLFE increases the 'All' accuracy of the test-dev from 70.41% to 70.63% by effectively coordinating detection-based features and context features. At the same time, we can see that our model performance is 0.22% higher than that of the best model MEDAN [28] for 'All' types of questions on the test-dev dataset. Our DLFE model demonstrates similar improvements in the other question categories, and these advantages are also reflected in the test-std results. These findings further confirm the strong performance of the DLFE model on VQA tasks.

Table 1: Comparison between DLFE and other models on the test-dev and test-standard sets of VQA 1.0

Methods	VQA v1.0 test-dev				VQA v1.0 test-std			
	All	Yes/No	Number	Other	All	Yes/No	Number	Other
Bottom-up [4]	65.53	81.10	44.34	56.32	65.42	82.45	43.81	55.93
CAM [29]	69.25	86.77	42.78	60.27	69.29	86.55	42.36	60.38
MSA [14]	69.47	86.72	43.05	60.70	69.64	86.87	42.63	60.77
Dual-MFA [6]	66.01	83.59	40.18	56.84	66.09	83.37	40.39	56.89
GRA [12]	70.41	86.85	53.90	60.13	70.30	86.15	54.41	60.25
MRA-Net [30]	69.06	86.79	43.89	59.62	69.22	86.37	44.16	60.00
DSACA [16]	69.65	85.63	44.32	59.73	69.53	85.77	43.77	59.57
ALSA [27]	69.52	86.12	42.94	59.06	69.32	86.94	43.84	58.21

(Continued)

Table 1 (continued)

Methods	VQA v1.0 test-dev				VQA v1.0 test-std			
	All	Yes/No	Number	Other	All	Yes/No	Number	Other
MEDAN [28]	70.41	86.85	52.58	60.77	70.23	86.63	52.45	60.73
DLFE (ours)	70.63	86.97	52.65	60.90	70.41	87.04	52.66	60.82

Table 2 shows our results and the results of other currently popular methods on test-dev of VQA 2.0. Compared with other methods, our model has more advantages for all indicators. Bottom-up [4] is the 2017 VQA challenge winner, which first incorporates an object detection network Faster R-CNN to enhance the visual features. Our model also uses the same image features and outperforms the UpDn model on all sub-tasks on the test-dev. The ALSA model has some similarities with our model. They increase the accuracy of answer prediction by fusing the two image features through confrontation learning. However, this approach lacks the ability to capture the internal relationships among image features, which can produce semantic noise when the two features are fused. Therefore, our model improved by 1.70% in 'All' type accuracy on the test-dev set. The MRA-Net [30] model explores visual and textual relations to improve performance. Our model is 1.89% higher than MRA-Net in 'All' type accuracy. SPRA [18] achieves fine-grained interaction between questions and images through a deep cascade of co-attention models, thus outperforming our model for the "Number" and "other" question types. However, DLFE learns visual information from a different perspective to obtain comprehensive image features and achieves higher accuracy on other types of questions using only co-attention to interact with question features. The highest accuracy obtained on the 'All' type proves that DLFE enables two visual features to complement each other.

Table 2: Comparison between DLFE and other models on the test-dev set of VQA 2.0

Methods	VQA v2.0 test-dev			
	All	Yes/No	Number	Other
Bottom-up [4]	65.32	81.82	44.21	56.05
CAM [29]	70.39	85.18	53.86	60.57
Dual-MFA [6]	66.24	84.59	41.08	57.46
MSA [11]	69.13	85.58	47.91	59.92
MRA-Net [30]	69.02	85.58	48.92	59.46
GRA [12]	70.52	88.68	53.94	62.97
ODA-GCN [13]	66.67	84.28	47.02	56.57
DSACA [16]	69.42	85.95	48.01	59.87
ALSA [27]	69.21	85.73	48.98	59.17
CMCN [15]	68.03	86.27	49.84	57.85
MEDGA [31]	70.11	85.97	51.56	60.09
SPRA [18]	70.35	86.98	54.98	61.52
QD-GFN [32]	70.51	86.45	52.41	60.52
DMBA-NET [33]	70.69	86.94	51.15	60.72
MEDAN [28]	70.76	87.02	52.69	60.77

(Continued)

Table 2 (continued)

Methods	VQA v2.0 test-dev			
	All	Yes/No	Number	Other
DLFE (ours)	70.91	87.09	52.74	61.07

The model can achieve excellent results because DLFE fully exploits visual information and can obtain more effective visual features. The reason for the superiority of the GRA model in "Yes/No" and "Other" type questions is that it has a relation-aware module to extract semantic relations between image objects, which allows the model to answer questions requiring relational reasoning effectively. In addition, GRA introduces bilinear attention for the initial interaction between visual features and question features and filters visual features irrelevant to the question for better counting. The GRA model only uses GRU to classify the question, ignoring the underlying semantic relationships between the key words. Our model can explore the implicit relationships between words and outperforms GRA by 0.39% on the "All" type. Although our model does not model the semantic relationship between the objects in the image, the attention visualization in Fig. 5 shows that DLFE can learn the correct visual features and question features. The DLSA and FFA modules designed in DLFE can be built on any type of attention model, including GRA, to mutually reinforce them from different viewpoints. Namely, DLFE can be applied to other excellent VQA models for better performance.

4.4 Parameter Sensitivity

To analyze the parametrization and how it affects the performance of DLFE, we present the experimental results with different parameter settings on the VQA datasets.

Under limited experimental conditions, to determine the number of FFA layers, we set N as 2, 4, 6, and 8 options. We also test different numbers of heads in the multi-head attention module by setting it to 2, 4, 8 and 16. According to the number of layers and attention heads, 16 experiments were considered. Table 3 shows the "All" prediction accuracy on the VQA 2.0 test-dev dataset. After conducting experiments with different numbers of FFA layers and attention heads in multi-head attention, we observed that the highest accuracy rate for the "All" type was achieved when the number of attention heads was set to 8. Therefore, we set the number of attention heads to 8 for our final model. With the increase in the number of layers N of FFA and the same number of attention heads, the accuracy of "All" increases. When N is 6, the performance index reaches its peak. However, when N is 8, the performance starts to decline. Therefore, the number of layers N of FFA is set to 6.

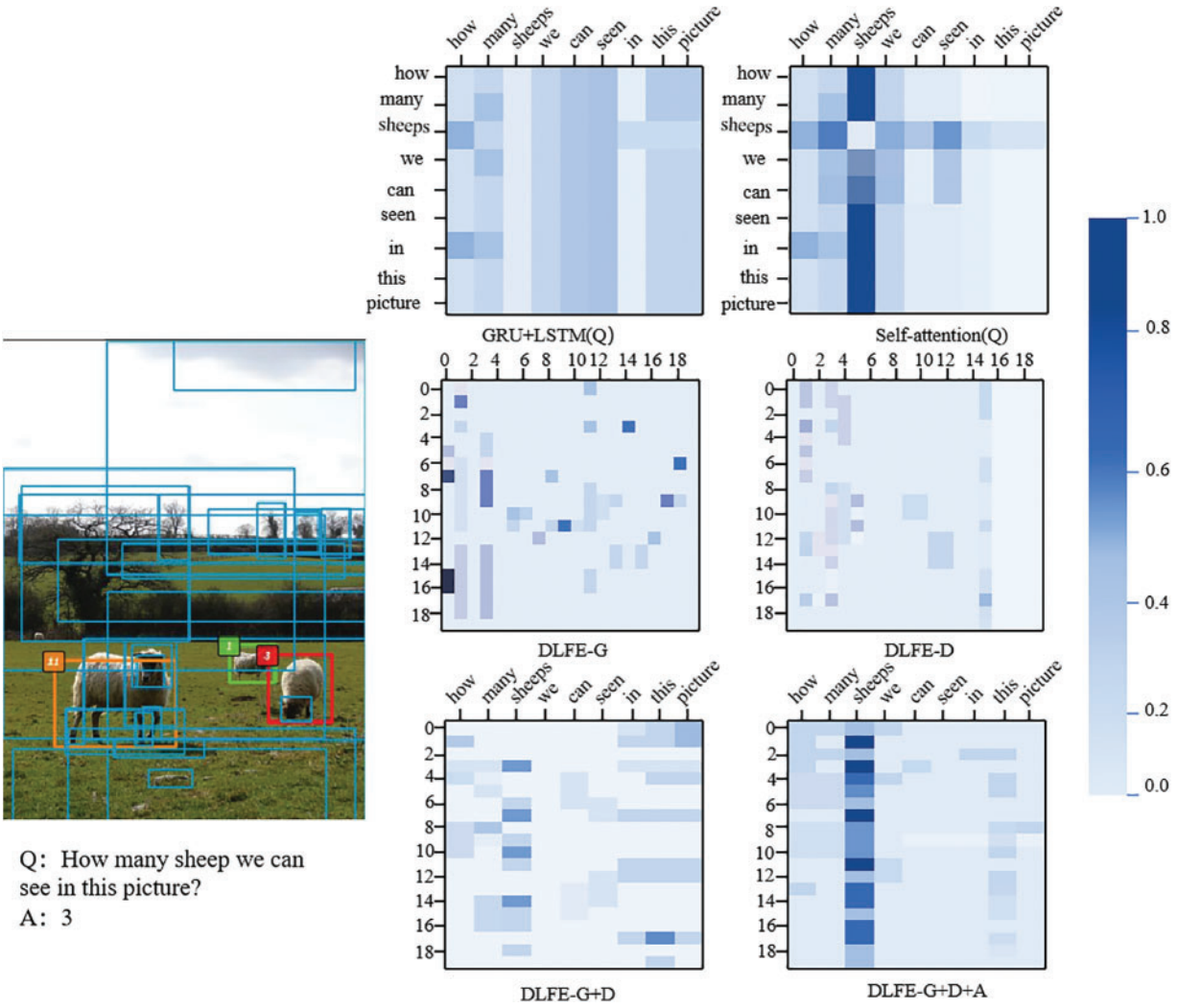

Figure 5: Visualization of learned attention map

Table 3: The results of our model on test-dev set of VQA 2.0 with different parameter setting

	N = 2	N = 4	N = 6	N = 8
2 head	66.50	69.47	69.56	69.86
4 head	68.61	69.60	70.01	70.18
8 head	69.69	69.73	70.43	70.25
16 head	69.81	69.95	70.03	69.91

Experiments show that the FFA module mainly focuses on intra-modal features and inter-modal features. The interaction of modal feature information is a crucial process. However, increasing the number of FFA layers may result in better performance. If the key features of the grid-based and

the detection-based features are aligned, continuing the FFA operation will result in deviation of the aligned features, which can lead to a degradation in the model's performance.

4.5 Ablation Study

Our DLFE is composed of several components. To measure the impact of each component on the overall prediction accuracy, we performed ablation studies on the VQA 2.0 value set. The following modifications to our suggested model:

DLFE-G: This model retains only grid-based features in the vision module.

DLFE-D: This model retains only the features based on the detection frame in the vision module.

DLFE-G+D: This model combines two image features.

DLFE-G+D+A: In the model feature fusion, the alignment graph is introduced to solve the semantic noise of the fusion.

ADD: Directly fuse text features and image features.

Co-attention: Co-attention is given to text features and image features.

The first four rows of Table 4 evaluate the effect of each significant element of the visual feature embedded network. Here, we use co-attention to fuse image and text features. DLFE-G considers the grid to enhance local feature representation while ignoring the relationship between the objects related to the problem. DLFE-D considers the detection frame feature to improve the global context information and neglects the local details. Compared with that of DLFE-G+D+A, their performance is worse. DLFE-G+D only directly fuses the two features, neglecting the relationship between the two features. It is challenging to optimize fusion, and semantic noise is introduced. Although performance is improved compared with that of DLFE-G and DLFE-D, it still needs improvement as that of DLFE-G+D+A. DLFE-G+D+A combines the two features and introduces an alignment module to reduce semantic noise and parameters in the fusion process. The accuracy is better than the above models.

Table 4: Results of ablation experiment on VQA 2.0

Methods	VQA v2.0 test-dev			
	All	Yes/No	Number	Other
DLFE-G	68.50	86.70	50.94	59.29
DLFE-D	69.84	86.88	51.56	59.95
DLFE-G+D	70.02	86.95	52.03	60.37
DLFE-G+D+A	70.59	87.04	52.47	60.58
ADD	70.55	86.90	52.06	60.35
CO-attention	70.76	87.15	52.63	60.82

The last two rows of Table 4 evaluate the impact of text and image feature fusion methods on model results. DLFE-G+D+A is used to get visual features. From the experimental results, Co-attention contributes to the performance because it can capture the cross-modal correlation between visual content and the question sentences.

Table 5 shows ablation experiments on different question representation methods. Rand_{ft} indicates the word embedding is initially randomized and subsequently fine-tuned, while PE the use of positional

encoding [34]. GLoVe$_{pt}$ indicates the word embeddings [35] are pre-trained with GloVe, GLoVe$_{pt+ft}$ is further fine-tuned We can observe that GloVe word embeddings is significantly outperforms than random initialization. Other methods, such as fine-tuning the GloVe embedding or using LSTM network to replace the position encoding to model the temporal information, can slightly improve the performance.

Table 5: Accuracies of our model with different question representations

Model	All	Yes/No	Number	Other
Rand$_{ft}$+PE	66.4	85.1	47.88	59.51
GLoVe$_{pt}$+PE	68.3	86.3	49.32	60.64
GLoVe$_{pt}$+LSTM	69.7	86.7	49.32	60.88
GLoVe$_{pt+ft}$+LSTM	69.8	86.7	49.28	60.91

In Fig. 5, we will visualize the learned attention. We only show one example and visualize six attention maps from different attention units.

Question: The attention map of GRU+LSTM forms vertical stripes, and words such as "how" and "see" gain greater attention weight. This unit serves as a question-type classifier. In addition, the large value of SA (Q) is in the "sheep" column. This shows that all concerned features tend to use "sheep" features for reconstruction. In other words, the keyword "sheep" is correctly identified.

Image: The values in the attention map of DLFE-G are uniformly distributed, indicating that sheep's key objects are unclear. The large values in the attention map of DLFE-D appear in columns 1, 3, and 15, which correspond to the three sheep in the image. This explains why detection-based attention can significantly improve object counting performance. The attention map of DLFE-G+D cannot focus correctly on the current object in the image. The attention map of DLFE-G+D+A tends to focus on all values in the "sheep" column. This can be explained by the fact that the input feature is reconstructed from the "sheep" feature in feature fusion attention. In addition, the DLFE-G+D unit contains more noise than the DLFE-G+D+A.

4.6 Qualitative Evaluation

We have visualized the attention map of DLFE to study the impact of the attention mechanism after fusion. At the same time, the attention map based on the region and the attention model based on the detection frame is visualized for comparison. In Fig. 6, we offer five instances from the VQA test-dev. From Figs. 6a and 6b, we can infer that both the region-based and check frame-based models can obtain the correct answers, and the fused model can consistently get the right answers. In some cases, just one attention mechanism can focus on the appropriate image areas to obtain the correct answer. For example, in Fig. 6c, the correct answer is generated based on the detected attention having a higher weight in all three horses, while the grid attention map pays attention to the incorrect area. In Fig. 6d, the attention map based on the grid can focus on the sky and ground, while the attention map based on the detection fails because these image areas are not covered by a fixed detection frame. Another instance is shown in Fig. 6e. The model cannot provide the right answer because although the attention is on the correct area, the symbol cannot be obtained from the attention area and contains the meaning "cannot turn right".

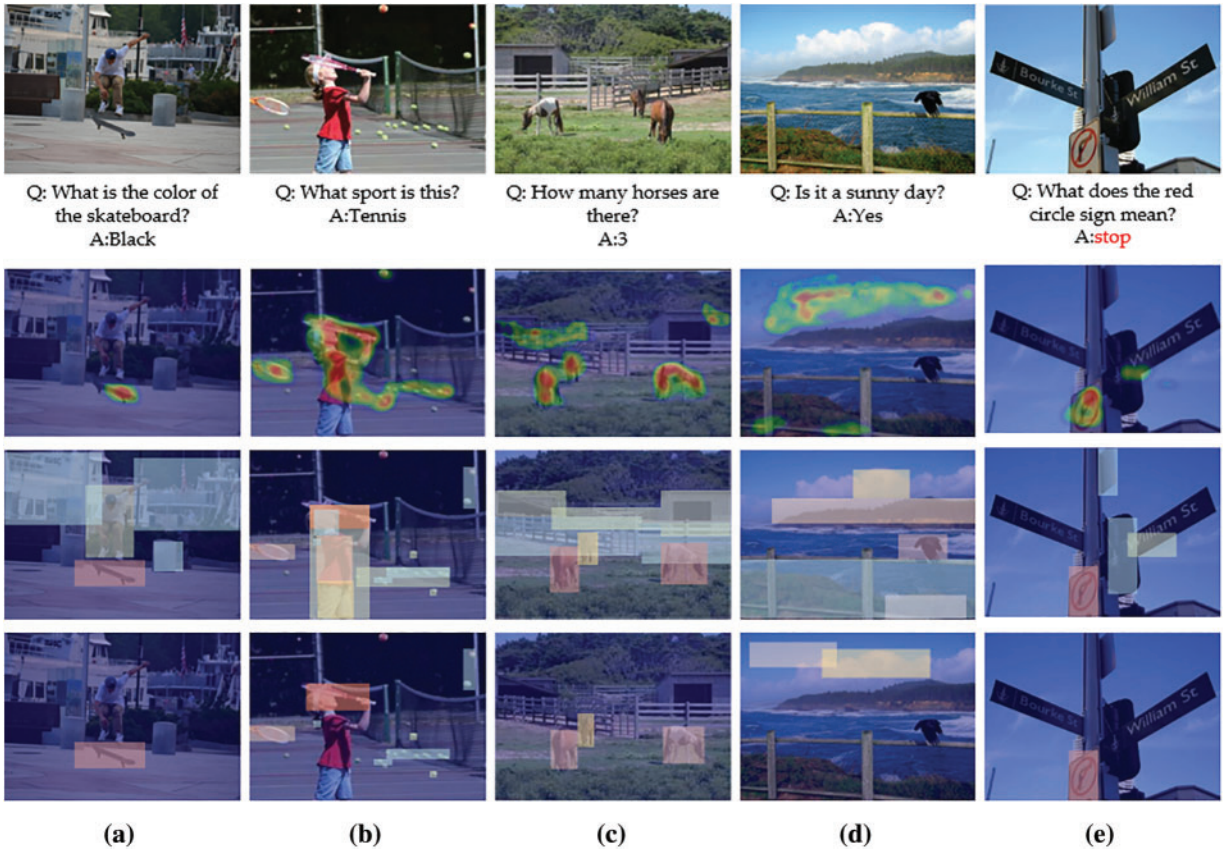

Figure 6: Five examples for attention visualization. The first row in each instance shows the original image along with the corresponding question-answer pair. The second row displays the grid-based attention maps, while the third row displays the detection-based attention maps. The last row shows the fused feature maps

5 Conclusions

In this paper, we explore the visual attention of VQA tasks from different perspectives and propose a DLFE model. The method achieves complementarity between grid and detection features and can comprehensively mine image information to answer questions about foreground objects and background forms. Our method applies a Dual-Level Self-Attention module to explore intrinsic characteristics of detection and grid features, where a Position Relation Attention is introduced to embed the geometric information of the image's relative and absolute position relationships. We also propose a Feature Fusion Attention module with an alignment graph to effectively enhance the two features and solve the problem of random noise caused by the direct fusion of the two feature sources. Finally, the co-attention module captures the relationship between question features and image features to predict answers. Extensive experiments on two public VQA datasets confirm the effectiveness and superiority of DLFE over baseline. In addition, we also demonstrate the superiority of the proposed model components through ablation experiments.

Although our model achieves a high accuracy rate, some limitations still need to be addressed. Fig. 6e shows that although the model focuses on the correct region, it cannot get the meaning of

"cannot turn right" from the captured symbols. The model needs external knowledge to support questions that cannot be answered directly from the image. Therefore, we need to consider further integrating an external knowledge base into the model to embed the knowledge related to the question and the image objects to provide rich information to improve the model's accuracy. Furthermore, our perception of VQA as a categorical task does not align with people's intuitive understanding. In subsequent research, we can add a natural language generation module to the answer prediction component to enable the model to combine questions with classification to generate more flexible responses.

Acknowledgement: This work does not require additional acknowledgment.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm the contributions to the paper as follows: study conception and design: S., X.; data collection: S., F.; analysis and interpretation of results: S.; draft manuscript preparation: S. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data supporting the findings of this study are openly available in VQA Dataset at https://visualqa.org/.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. Antol *et al.*, "VQA: Visual question answering," in *Proc. Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 7–13.

[2] A. Singh, K. Raj, T. Kumar, S. Verma, and A. M. Roy, "Deep learning-based cost-effective and responsive robot for autism treatment," *Drones*, vol. 7, no. 2, pp. 81–98, 2023. doi: 10.3390/drones7020081.

[3] Y. Goyal and T. Khot, "Making the v in VQA matter: Elevating the role of image understanding in visual question answering," *Int. J. Comput. Vis.*, vol. 127, no. 1, pp. 398–414, 2017. doi: 10.1007/s11263-018-1116-0.

[4] P. Anderson and X. He, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 6077–6086.

[5] Y. Ming, N. N. Hu, C. X. Fan, and F. Feng, "Visuals to text: A comprehensive review on automatic image captioning," *IEEE/CAA J. Autom. Sin.*, vol. 9, no. 8, pp. 1339–1365, 2022. doi: 10.1109/JAS.2022.105734.

[6] P. Lu, H. Li, W. Zhang, J. Wang, and X. Wang, "Co-attending free-form regions and detections with multimodal multiplicative feature embedding for visual question answering," in *Proc. AAAI Conf. Artif. Intell.*, New Orleans, Louisiana, USA, 2018, vol. 32, no. 1.

[7] S. Jamil and A. M. Roy, "An efficient and robust phonocardiography (PCG)-based valvular heart diseases (VHD) detection framework using vision transformer (VIT)," *Comput. Biol. Med.*, vol. 158, no. 25, pp. 106734, 2023. doi: 10.1016/j.compbiomed.2023.106734.

[8] A. Osman and W. Samek, "DRAU: Dual recurrent attention units for visual question answering," *Comput. Vis. Image Underst.*, vol. 185, no. 8, pp. 24–30, 2019. doi: 10.1016/j.cviu.2019.05.001.

[9] Y. Liu, X. M. Zhang, F. R. Huang, and Z. H. Zhao, "Visual question answering via combining inferential attention and semantic space mapping," *Knowl. Based Syst.*, vol. 207, no. 9, pp. 106339, 2020. doi: 10.1016/j.knosys.2020.106339.

[10] C. Wu, J. Liu, X. Wang, and X. Dong, "Object-difference attention: A simple relational attention for visual question answering," in *Proc. 26th ACM Int. Conf. Multimedia*, Seoul, South Korea, 2018, pp. 519–527.

[11] D. Teney, P. Anderson, X. He, and A. Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 4223–4232.
[12] X. J. Shao, Z. G. Xiang, and Y. X. Li, "Visual question answering with gated relation-aware auxiliary," *IET Image Process.*, vol. 16, no. 5, pp. 1424–1432, 2022. doi: 10.1049/ipr2.12421.
[13] X. Zhu, Z. Mao, Z. Chen, Y. Li, Z. Wang and B. Wang, "Object-difference drived graph convolutional networks for visual question answering," *Multimed. Tools Appl.*, vol. 80, no. 11, pp. 16247–16265, 2021. doi: 10.1007/s11042-020-08790-0.
[14] W. Li, J. H. Sun, G. Liu, L. L. Zhao, and X. Z. Fang, "Visual question answering with attention transfer and a cross-modal gating mechanism," *Pattern Recognit. Lett.*, vol. 133, no. 5, pp. 334–340, 2020. doi: 10.1016/j.patrec.2020.02.031.
[15] D. Z. Han, S. L. Zhou, and K. C. Li, "Cross-modality co-attention networks for visual question answering," *Soft Comput.*, vol. 25, no. 4, pp. 5411–5421, 2021. doi: 10.1007/s00500-020-05539-7.
[16] Y. Liu, X. M. Zhang, Q. Y. Zhang, C. Z. Li, and F. R. Huang, "Dual self-attention with co-attention networks for visual question answering," *Pattern Recognit.*, vol. 117, no. 9, pp. 107956, 2021. doi: 10.1016/j.patcog.2021.107956.
[17] S. He and D. Z. Han, "An effective dense co-attention networks for visual question answering," *Sensors*, vol. 20, no. 8, pp. 4897, 2020. doi: 10.3390/s20174897.
[18] Y. Feng, W. Silamu, Y. B. Li, and Y. C. Chai, "A based on spatial position relationship co-attention network for visual question answering," *Vis. Comput.*, vol. 38, no. 6, pp. 3097–3108, 2022. doi: 10.1007/s00371-022-02524-z.
[19] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A deep learning approach to visual question answering," *Int. J. Comput. Vis.*, vol. 125, no. 8, pp. 110–135, 2017. doi: 10.1007/s11263-017-1038-2.
[20] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An attentive survey of attention models," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 5, pp. 1–32, 2021. doi: 10.1145/3465055.
[21] Y. Liu, X. Zhang, F. Huang, L. Cheng, and Z. Li, "Adversarial learning with multi-modal attention for visual question answering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 3894–3908, 2021. doi: 10.1109/TNNLS.2020.3016083.
[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.
[23] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," arXiv:1607.06450, 2016.
[24] Q. Wu, D. Teney, P. Wang, C. H. Shen, and A. Dick, "Visual question answering: A survey of methods and datasets," *Comput. Vis. Image Underst.*, vol. 163, no. 10, pp. 21–40, 2017. doi: 10.1016/j.cviu.2017.05.001.
[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980, 2014.
[26] K. Ranjay *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 2, pp. 32–73, 2017. doi: 10.1007/s11263-016-0981-7.
[27] Y. Liu, X. Zhang, Z. Zhao, B. Zhang, L. Cheng, Z. Li, "ALSA: Adversarial learning of supervised attentions for visual question answering," *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 4520–4533, 2020. doi: 10.1109/TCYB.2020.3029423.
[28] C. Chen, D. Han, and J. Wang, "Multimodal encoder-decoder attention networks for visual question answering," *IEEE Access*, vol. 8, pp. 35662–35671, 2020. doi: 10.1109/ACCESS.2020.2975093.
[29] L. Peng, Y. Yang, Z. Wang, Z. Huang, and H. T. Shen, "Answer again: Improving VQA with cascaded-answering model," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 4, pp. 1644–1655, 2020. doi: 10.1109/TKDE.2020.2998805.
[30] L. Peng, Y. Yang, Z. Wang, Z. Huang, and H. T. Shen, "MRA-Net: Improving VQA via multi-modal relation attention network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 318–329, 2020.
[31] H. Li and D. Han, "Multimodal encoders and decoders with gate attention for visual question answering," *Comput. Sci. Inf. Syst.*, vol. 18, no. 3, pp. 1023–1040, 2021. doi: 10.2298/CSIS201120032L.
[32] Y. Qian, Y. Hu, R. Wang, F. Feng, and X. Wang, "Question-driven graph fusion network for visual question answering," in *2022 IEEE Int. Conf. Multimedia and Expo*, Taipei City, Taiwan, China, 2022, pp. 1–6.

[33] F. Yan, W. Silamu, and Y. Li, "Deep modular bilinear attention network for visual question answering," *Sensors*, vol. 22, no. 3, pp. 1045, 2022. doi: 10.3390/s22031045.
[34] W. Zhang, J. Yu, H. Hu, and Z. C. Qin, "Multimodal feature fusion by relational reasoning and attention for visual question answering," *Inf. Fusion*, vol. 55, no. 3, pp. 116–126, 2020. doi: 10.1016/j.inffus.2019.08.009.
[35] W. Guo, Y. Zhang, J. Yang, and X. Yuan, "Re-attention for visual question answering," *IEEE Trans. Image Process.*, vol. 30, pp. 6730–6743, 2021. doi: 10.1109/TIP.2021.3097180.