ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230607

Boosting Visual Question Answering Through Geometric Perception and Region Features

Hong Yu[a], Zhiyue Wang[a], Yuanqiu Liu[a] and Han Liu[a;*]

[a]School of Software, Dalian University of Technology

Abstract. Visual question answering (VQA) is a crucial yet challenging task in multimodal understanding. To correctly answer questions about an image, VQA models are required to comprehend the fine-grained semantics of both the image and the question. Recent advances have shown that both grid and region features contribute to improving the VQA performance, while grid features surprisingly outperform region features. However, grid features will inevitably induce visual semantic noise due to fine granularity. Besides, the ignorance of geometric relationship makes VQA models difficult to understand the object relative positions in the image and answer questions accurately. In this paper, we propose a visual enhancement network for VQA that leverages region features and position information to enhance grid features, thus generating richer visual grid semantics. First, the grid enhancement multi-head guided-attention module utilizes regions around the grid to provide visual context, forming rich visual grid semantics and effectively compensating for the fine granularity of the grid. Second, a novel geometric perception multi-head self-attention is introduced to process two types of features, incorporating geometric relations such as relative direction between objects while exploring internal semantic interactions. Extensive experiments demonstrate that the proposed method can obtain competitive results over other strong baselines.

1 Introduction

Visual question answering (VQA) is a challenging task of multimodal learning to bridge vision and language. With the advancement of deep learning technology, the accuracy of VQA has improved dramatically, which makes it popular in many realistic scenarios, such as blind assistance, autonomous driving, and other fields.

VQA generally answers questions based on visual clues extracted from images according to the semantic information of the corresponding questions. Existing approaches commonly leverage deep learning models to extract image and language features, and perform multimodal feature fusion to correlate the question with critical image regions using attention mechanisms to answer the question accurately. Recently, the attention mechanism is widely applied in VQA tasks leading to more accurate answers [2, 5, 36, 38]. As an advanced representative, Transformer [32] relies entirely on attention mechanisms to draw global dependencies between input and output. Many VQA paradigms with Transformer are proposed and achieve excellent performance.

Most existing VQA models emphasize the following challenges. The first challenge is how to extract precise visual information from a given image, which has prompted researchers to explore more robust semantic representations of images. Pre-trained Convolutional Neural Networks (CNNs) are used to extract image features [3] previously, which only capture general information about images and limit the performance of VQA. In [2], bottom-up attention is proposed to provide image features based on Faster R-CNN [31], which helps to identify crucial regions in an image. This approach has been widely used in subsequent works. Moreover, well pre-trained grid features show strong descriptive capabilities in VQA and image captioning [13].

To get more expressive visual information, the integration of multiple visual sources are utilized in recent studies. A multimodal multiplicative feature embedding scheme [24] is proposed to merge free-form image regions, detection boxes, and question representations. While it does not consider the relationship between visual features comprehensively. Some research [14, 27, 34] incorporate segmentation features, OCR features or object attributes and relationship to provide a comprehensive visual-linguistic view of the input image. However, these models mostly focus on the region features. The combination of grid features, which perform excellently in VQA is less studied especially the problem of the semantic inconsistency [22] and spatial misalignment.

The second challenge is how to explore and leverage the spatial relationship between objects effectively. Spatial information, such as object position, relative size, and relative direction, is critical for a comprehensive understanding of visual content. Many VQA studies [38, 43] mainly consider the semantic correlation between visual objects which brings low visual reasoning capabilities, particularly for questions related to spatial relations. To deal with the problem, some work [9, 20, 33] encodes bounding box position, size, and other information into image features and has achieved improvement. But it perceives relative direction or absolute position information alone which results in a significant limitation.

Grid features and region features are independently explored to deal with the aforementioned challenges. Despite the success, there are still limitations. Although grid features perform well in VQA, they introduce visual semantic noise after global interaction and fail to focus on critical visual information due to fragmentation. While widely applied region features concentrate on objects, whose semantic information is intensive, they cannot cover the whole image or describe the global scene comprehensively. The two types of visual features complement each other's information, and they should be effectively integrated into a unified framework. In this paper, we propose a visual enhancement network for VQA that accurately locates and enriches target objects by leveraging region features and geomet-

* Corresponding Author. Email: liu.han.dut@gmail.com.

rical visual relations to augment grid features during the fusion of vision and language features. Our proposed network consists of four modules: a self-attention module (SA), a grid enhancement module (GE), a geometric perception module (GP), and a guided-attention module (GA). The SA module captures the dependencies of words in the question and regions in the image respectively. The GE module augments the grid features with region features to exploit visual information at different granularities. The GP module extends the SA module to model the geometric relationship between input objects explicitly and efficiently. Finally, the GA module uses the question to guide the image to focus on areas relevant to it.

The contributions of this paper can be summarized as follows:

- We propose a grid enhancement multi-head guided-attention (GE-MGA), which utilizes regions around the grid to provide visual context for it. The coarse-grained and fine-grained information complement each other to provide rich visual grid semantics for answer inference.
- A novel geometric perception multi-head self-attention (GP-MSA) is introduced to improve image comprehension ability by exploiting geometric relationship such as absolute position and relative direction.
- By combining the two modules and applying them to the VQA network, the visual enhancement network (VE) is proposed. Extensive experiments on the VQA-v2 dataset demonstrate that our network can obtain competitive results compared with other strong baselines.

2 Related Work

2.1 Visual Question Answering

Visual question answering (VQA) aims to answer questions based on images and corresponding questions. It is commonly treated as a classification task with fixed categories [3, 33]. With the rapid development of VQA, various benchmark datasets [1, 3, 12, 15, 19, 30] and methods [6, 10, 36, 37, 38, 41, 43] have been proposed successively. Previous studies have thoroughly explored feature fusion methods [41], moving beyond simple mechanisms like concatenation, element-wise multiplication, and element-wise addition. Multimodal compact bilinear pooling model [6], multimodal low-rank bilinear pooling model [17], multimodal factorized bilinear pooling model [39] and multimodal factorized high-rank pooling model [40] have been proposed, which not only capture the complex interaction information between multimodalities but also reduce the number of model parameters. Attention mechanisms have been proven to be valid in various tasks, such as machine translation [4] and image captioning [9]. Several studies have explored attention mechanisms for VQA, including the early attention method [24, 36] by stacking multi-step visual attention and using questions to identify question-relevant regions in images, co-attention method [23, 38] for simultaneous visual and text feature attention learning, and relation attention method [5, 20] modeling explicit and implicit relations between regions through graph attention mechanism. With the popularity of Transformer networks [32], recent advances in VQA benefit from stacking multiple attention layers, such as bilinear attention layers [16], self-attention layers [37, 43], to capture the relationship both within and across modalities.

2.2 Visual Feature

The exploitation of visual features of an image is crucial to the accuracy of modern VQA models. There are typically three types of visual features, namely: global features, region features, and grid features. Global features are obtained through a conventional CNN network to encode the images [26, 28]. They fail to capture details and local information in the image. After the innovative proposal of the combined bottom-up and top-down attention mechanism [2], region features are widely used in VQA, which precisely locate objects and provide rich visual representations, such as categories and attributes. In [13], researchers revisit grid features as an alternative to the widely used bottom-up region features for vision and language tasks, skipping all the region-related steps in the existing VQA pipeline and using C5 output of the adapted ResNet backbone as grid convolution features. Their experiments show that grid features can perform better than region features in less inference time.

2.3 Visual Relationship Modeling

Some recent works focus on the detection and simulation of visual relationship. Visual relations represent interactions between objects, which are essential for locating targets by contextual information. Existing VQA approaches implement visual relationship modeling through implicit relationship and explicit relationship. Implicit relationship [5, 38] is captured by the attention module, or higher-order interactions on fully connected graphs. Explicit relationship refers to geometric positions and semantic interactions between objects. Previous methods [9, 25, 37] measure geometric relationship, considering distance and scale. However, they ignore relative direction, which is essential for object localization and relationship understanding. Other works [20] model spatial and semantic relations using bounding boxes and object features. But they ignore the absolute position, which is helpful to distinguish two objects with the same appearance at different locations in an image. In contrast to these works, we consider both absolute positions and relative directions to model complex visual and position relationship between input features comprehensively and accurately.

3 Method

As shown in Figure 1, our visual enhancement (VE) visual question answering model consists of three main components: visual and textual representation (Section 3.1), VE network (Section 3.2), and answer reasoning (Section 3.3). In the following, we elaborate on these three components sequentially.

3.1 Visual and Textual Representation

By considering each visual object v_i in the image and word w_i in a question as a node, we can construct a fully-connected undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{E}$ is the set of $K \times (K-1)$ edges. Each edge represents an implicit semantic and explicit geometric relationship, which consists of the weights assigned to each edge by graph attention.

Node $\mathcal{V}$ is composed of visual features (containing regions $\boldsymbol{V}_r$ and grids $\boldsymbol{V}_g$) and textual features $\boldsymbol{Q}$. Following [2, 13], we use pre-trained Faster R-CNN [31] model with ResNeXt backbone to identify a set of objects $\boldsymbol{V}_r = \{\boldsymbol{v}_i^r\}_{i=1}^M$, where $\boldsymbol{v}_i^r \in \mathbb{R}^{d_v}$ represents the visual feature vector of i-th object. Meanwhile, each visual object corresponds to a position information vector $\boldsymbol{b}_i^r =$

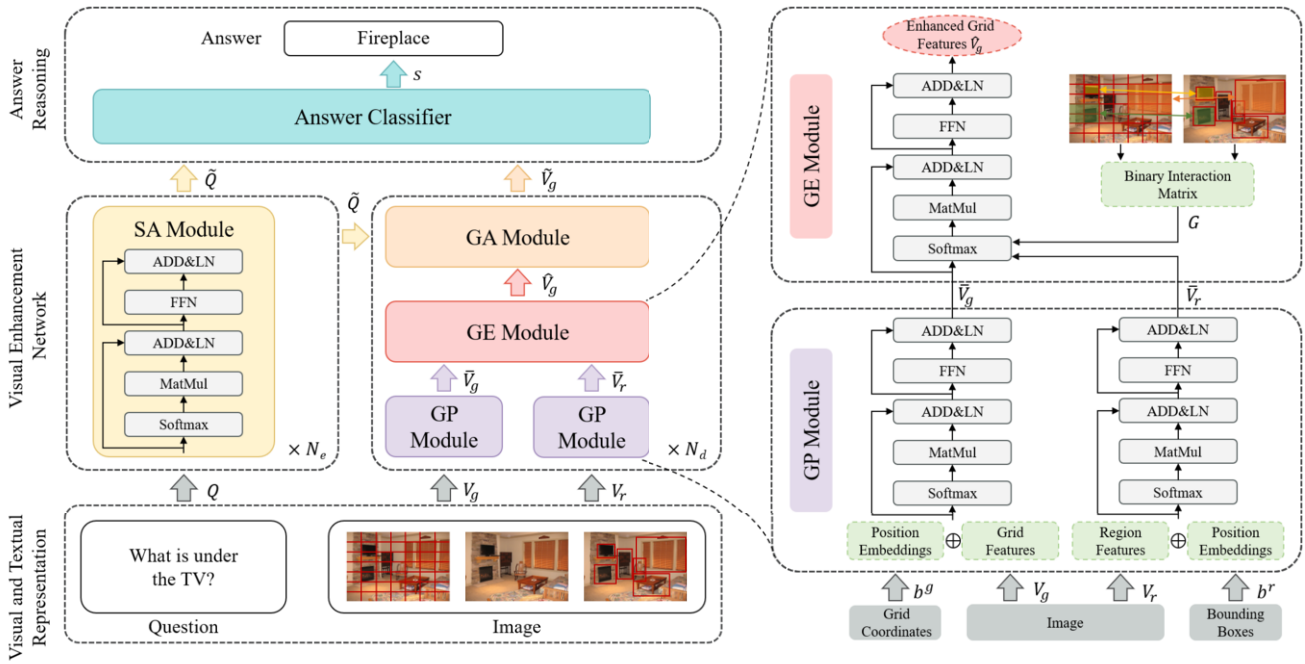

Figure 1. The overall architecture of our proposed visual enhancement network. The input image and question are first represented as a series of region features, grid features, and word embeddings in the visual and textual representation module. Our proposed visual enhancement network consists of N_e encoder layers and N_d decoder layers. The encoder layer encodes question textual features through the self-attention module (SA). The decoder layer obtains enhanced grid features through the geometric perception module (GP) and grid enhancement module (GE), and uses the guided-attention module (GA) to guide to focus on visual information related to the question. Finally, the answer classifier is used to fuse two high-level semantic vectors into one vector for answer prediction.

$(x_{min}, y_{min}, x_{max}, y_{max}) \in \mathbb{R}^4$, where x_{min} and y_{min} correspond to top left coordinates, as well x_{max} and y_{max} denote bottom right coordinates respectively.

Grid features $\boldsymbol{V}_g = \{\, \boldsymbol{v}_i^g \,\}_{i=1}^N$ are extracted from ResNeXt [13], where $\boldsymbol{v}_i^g \in \mathbb{R}^{d_f}$ is the representation of the i-th grid, and N is the number of grid features. We view the grid as a special kind of bounding box, so each grid feature can also calculate the corresponding position vector $\boldsymbol{b}_i^g = (x_{min}, y_{min}, x_{max}, y_{max}) \in \mathbb{R}^4$.

For the textual representation of the question, we use LSTM [11] to generate context-aware embedding for each word, denoted as $\boldsymbol{Q} = \{\, \boldsymbol{w}_i \,\}_{i=1}^L$, where i-th word is formulated as $\boldsymbol{w}_i \in \mathbb{R}^{d_q}$. L is the length of the question.

3.2 Visual Enhancement Network

Our proposed VE network adopts an encoder-decoder structure by stacking N_e encoder layers for modeling interactions between textual features and N_d encoder layers for processing visual features and cross-modal alignment. The encoder layer consists of the self-attention module (SA), and the decoder layer consists of the grid enhancement module (GE), the geometric perception module (GP), and the guided-attention module (GA).

3.2.1 Self-Attention Module

We compute the implicit relationship between text pairs and visual object pairs separately based on multi-head self-attention in [38]. It consists of h identical heads, and each attention head i takes query $\boldsymbol{Q}_i = \boldsymbol{Q}\boldsymbol{W}_i^Q$, key $\boldsymbol{K}_i = \boldsymbol{K}\boldsymbol{W}_i^K$ and value $\boldsymbol{V}_i = \boldsymbol{V}\boldsymbol{W}_i^V$ as input. $\boldsymbol{W}_i^Q$, $\boldsymbol{W}_i^K$ and $\boldsymbol{W}_i^V$ are the projection matrices for the i-th head. We adopt a scaled dot product function to calculate the correlation score $\boldsymbol{E}_i$, and apply a softmax function to obtain the attention weights on the values. This process is visualized as Figure 2(a). The output of each head is computed as follows:

$$\boldsymbol{E}_i = \frac{\boldsymbol{Q}_i \boldsymbol{K}_i^T}{\sqrt{d_k}}, \tag{1}$$

$$head_i = \mathrm{Att}(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i) = \mathrm{Softmax}(\boldsymbol{E}_i)\boldsymbol{V}_i. \tag{2}$$

The output of all heads is concatenated and multiplied with a learned projection matrix $\boldsymbol{W}^O$:

$$\mathrm{MSA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = [head_1, head_2, ..., head_h]\boldsymbol{W}^O. \tag{3}$$

When we set the inputs $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$ to the same $\boldsymbol{Q} = \{\, \boldsymbol{w}_i \,\}_{i=1}^L$, it forms the SA module used in our network. In addition, residual connection and layer normalization are applied to facilitate optimization. (They are also used in the GE, GP, and GA modules. For simplification, we provide a detailed elaboration only in this section and omit the explanation in the other three modules.) The output question features $\tilde{\boldsymbol{Q}}^l$ of the l-th layer can be calculated by:

$$\begin{aligned} \boldsymbol{Q}^l = &\mathrm{LayerNorm}\big(\tilde{\boldsymbol{Q}}^{l-1} + \mathrm{MSA}(\tilde{\boldsymbol{Q}}^{l-1}\boldsymbol{W}_{l-1}^Q, \\ &\tilde{\boldsymbol{Q}}^{l-1}\boldsymbol{W}_{l-1}^K, \tilde{\boldsymbol{Q}}^{l-1}\boldsymbol{W}_{l-1}^V)\big), \end{aligned} \tag{4}$$

$$\tilde{\boldsymbol{Q}}^l = \mathrm{LayerNorm}\big(\boldsymbol{Q}^l + \mathrm{FeedForward}(\boldsymbol{Q}^l)\big), \tag{5}$$

$$\mathrm{FeedForward}(\boldsymbol{Q}^l) = \mathrm{ReLU}(\boldsymbol{Q}^l\boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2, \tag{6}$$

where FeedForward$(\cdot)$ consists of two fully-connected layers with ReLU activation. $\boldsymbol{W}_{l-1}^Q, \boldsymbol{W}_{l-1}^K, \boldsymbol{W}_{l-1}^V \in \mathbb{R}^{d_q \times d_q}$, $\boldsymbol{W}_1 \in \mathbb{R}^{d_q \times (4d_q)}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{(4d_q) \times d_q}$ are learnt parameter matrices.

3.2.2 Grid Enhancement Module

In VQA, both grid features and region features contain information about objects in an image, but they have different characteristics of their own. Grid features may introduce visual semantic noise after global interaction due to fine granularity and fragmentation. Region features locate objects accurately and provide rich visual semantics. Therefore, combining grid features with region features will obtain comprehensive and accurate object information, thus improving performance.

We use improved visual attention called grid enhancement multi-head guided-attention (GE-MGA) to achieve enhanced grid features. The main idea of it is to find, for each grid feature, the regions with which it has positional overlap. The input is a set of region features and grid features, and the output is a set of enhanced grid features. The module consists of the following steps.

As shown in the middle part of Figure 2(b), we construct an interaction matrix $\boldsymbol{G}$, indicating the existence of implicit relationship between grids and regions, which is calculated based on the relative geometric position. $\boldsymbol{G}$ consists of binary numbers. $\boldsymbol{G}_{mn}$ is set to 1 when the m-th grid and n-th region coincide, and 0 otherwise.

A segmentation function is used to calculate the corresponding attention weight $\boldsymbol{E}_{mn}$ between $< \boldsymbol{v}_m^g, \boldsymbol{v}_n^r >$ when the $\boldsymbol{G}_{mn}$ value is 1, which means that the m-th grid and the n-th region have an intersection. Otherwise, we set the attention weight to negative infinity. GE-MGA can be obtained by:

$$\boldsymbol{E}_{mn} = \begin{cases} \frac{\boldsymbol{Q}_m \boldsymbol{K}_n{}^T}{\sqrt{d_k}}, & \boldsymbol{G}_{mn} = 1 \\ -\infty, & \boldsymbol{G}_{mn} = 0, \end{cases} \tag{7}$$

$$head_i = \text{GE-Att}(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i) = \text{Softmax}(\boldsymbol{E}_i)\boldsymbol{V}_i, \tag{8}$$

$$\text{GE-MGA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = [head_1, head_2, ..., head_h]\boldsymbol{W}^O. \tag{9}$$

We can compute the enhanced grid features $\hat{\boldsymbol{V}}_g$, which are the output of the GE module in Figure 1, as follows:

$$\hat{\boldsymbol{V}}_g = \text{GE-MGA}(\bar{\boldsymbol{V}}_g, \bar{\boldsymbol{V}}_r, \bar{\boldsymbol{V}}_r), \tag{10}$$

where $\bar{\boldsymbol{V}}_g$ and $\bar{\boldsymbol{V}}_r$ are the grid features and region features produced by the GP module mentioned in Section 3.2.3.

To sum up, the enhanced grid features incorporate the information of the region features. We calculate the attention weights of the region features corresponding to each grid feature with trade-offs, so the method can remove the redundant portion of the vision effectively and add truly helpful object information to the grid.

3.2.3 Geometric Perception Module

We construct the geometric perception multi-head self-attention (GP-MSA) incorporating position information based on the SA module. It introduces absolute position information and relative position relationship. We use a hybrid encoding approach to combine relative position encoding and absolute position encoding. This module preserves both relative position information and absolute position information between visual elements. Specifically, the module is divided into two stages: geometric transformation and feature aggregation.

In the geometric transformation stage, we first process the absolute position of the original visual features. For the i-th region feature, we perform a simple position encoding of its bounding box $\boldsymbol{b}_i^r = (x_{min}, y_{min}, x_{max}, y_{max})$:

$$\boldsymbol{pos}_r^i = \boldsymbol{b}_i^r \boldsymbol{W}_{emb}, \tag{11}$$

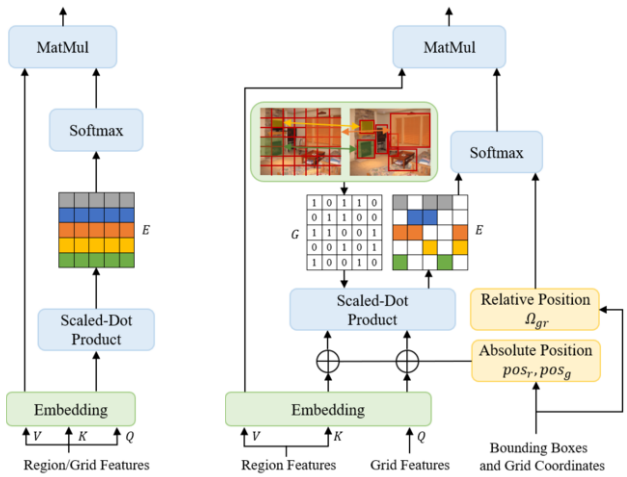

Figure 2. Illustration of the different attention modules. (a) Initial scaled dot-product attention in multi-head self-attention (MSA). (b) Scaled dot-product attention with the interaction matrix $\boldsymbol{G}$ in the grid enhancement module (GE) and position encodings in the geometry perception module (GP).

where $\boldsymbol{pos}_r^i \in \mathbb{R}^{d_m}$ is the encoded absolute position feature, and $\boldsymbol{W}_{emb} \in \mathbb{R}^{4 \times d_m}$ is a learnable parameter.

For each grid, we encode its absolute position in the whole image as a vector, using the same sine and cosine embedding methods for positional encoding as in [32], as follows:

$$\begin{cases} PE(pos, 2t) = sin(pos/10000^{2t/(d_m/2)}) \\ PE(pos, 2t+1) = cos(pos/10000^{2t/(d_m/2)}), \end{cases} \tag{12}$$

$$\boldsymbol{pos}_g^i = [\boldsymbol{PE}_m, \boldsymbol{PE}_n], \tag{13}$$

where pos is the position and t is the dimension. m and n represent the row and column index of the i-th grid respectively. $\boldsymbol{pos}_g^i \in \mathbb{R}^{d_m}$ is a concatenation of $\boldsymbol{PE}_m \in \mathbb{R}^{d_m/2}$ and $\boldsymbol{PE}_n \in \mathbb{R}^{d_m/2}$, which represents the absolute position encoding of grid features.

In order to better integrate the relative position encoding of visual features, our method calculates it based on the geometric structure of the bounding boxes. According to $\boldsymbol{b}_i = (x_{min}, y_{min}, x_{max}, y_{max})$, the width w_i and height h_i can be calculated. For i-th object and j-th object, their relative geometric space relationship is expressed as a five-dimensional vector as:

$$\begin{aligned} \boldsymbol{\Omega}(i,j) = \big(&log(\frac{|x_i - x_j|}{w_i}), log(\frac{|y_i - y_j|}{h_i}), \\ &log(\frac{w_i}{w_j}), log(\frac{h_i}{h_j}), log(r_{ij})\big), \end{aligned} \tag{14}$$

where the fifth of these terms represents relative direction encoding to enhance the orientation perception of the image. We will elaborate on it in the following exposition.

$4k$ unit direction vectors are predefined in the two-dimensional space, where each vector represents a direction category. We inscribe direction vectors in the rectangular coordinate system when k adopts various values, as shown in Figure 3. The i-th unit direction vector is denoted as:

$$\boldsymbol{\alpha}_i = (cos(\frac{i\pi}{2k}), sin(\frac{i\pi}{2k})), i \in \{0, 1, \ldots, 4k-1\}. \tag{15}$$

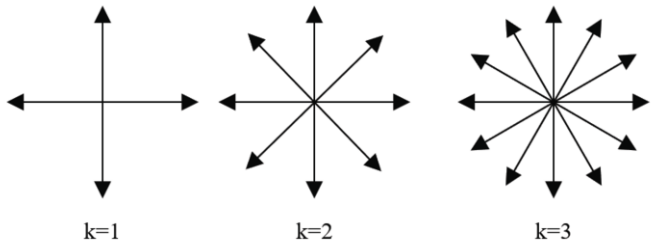

Figure 3. Examples of unit direction vectors for $k = 1, 2, 3$.

With the position coordinates of the m-th bounding box and the n-th bounding box, we can calculate the relative direction vector $\boldsymbol{v}_{mn}$. The relative relationship category r_{mn} is calculated by the cosine similarity algorithm between $\boldsymbol{v}_{mn}$ and $\boldsymbol{\alpha}_i$:

$$\boldsymbol{v}_{mn} = (\frac{x_m - x_n}{\sqrt{(x_m - x_n)^2 + (y_m - y_n)^2}}, \frac{y_m - y_n}{\sqrt{(x_m - x_n)^2 + (y_m - y_n)^2}}), \tag{16}$$

$$r_{mn} = \arg\max_i \frac{\boldsymbol{v}_{mn} \cdot \boldsymbol{\alpha}_i}{\|\boldsymbol{v}_{mn}\| \cdot \|\boldsymbol{\alpha}_i\|}. \tag{17}$$

We map $\boldsymbol{\Omega}(i, j)$ into the higher dimensional space by the embedding method in [32], and feed it into a linear layer to transform it into a scalar, which describes the geometric relationship between the two visual areas:

$$\boldsymbol{\Omega}(i, j) = \mathrm{ReLU}(\mathrm{Emb}(\boldsymbol{\Omega}(i, j)\boldsymbol{W}_G). \tag{18}$$

In the feature aggregation stage, we revisit SA for visual features by integrating the absolute position embedding and relative geometric information obtained above respectively. We add the corresponding absolute position information $\boldsymbol{pos}_q$ and $\boldsymbol{pos}_k$ to query $\boldsymbol{Q}$ and key $\boldsymbol{K}$, and integrate the relative geometric relations $\boldsymbol{\Omega}$ after multiplying them together. GP-MSA is calculated as:

$$\boldsymbol{E} = \frac{(\boldsymbol{Q} + \boldsymbol{pos}_q)(\boldsymbol{K} + \boldsymbol{pos}_k)^T}{\sqrt{d_k}} + log(\boldsymbol{\Omega}), \tag{19}$$

$$\begin{aligned} head_i =& \text{GP-Att}(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i, \boldsymbol{pos}_{q_i}, \boldsymbol{pos}_{k_i}, \boldsymbol{\Omega}_i) \\ =& \text{Softmax}(\boldsymbol{E}_i)\boldsymbol{V}_i, \end{aligned} \tag{20}$$

$$\begin{aligned} &\text{GP-MSA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}, \boldsymbol{pos}_q, \boldsymbol{pos}_k, \boldsymbol{\Omega}) \\ &= [head_1, head_2, ..., head_h]\boldsymbol{W}^O. \end{aligned} \tag{21}$$

The grid features $\bar{\boldsymbol{V}}_g$ incorporating geometric information, which are the output of the GP module in Figure 1, can be obtained by:

$$\bar{\boldsymbol{V}}_g = \text{GP-MSA}(\boldsymbol{V}_g, \boldsymbol{V}_g, \boldsymbol{V}_g, \boldsymbol{pos}_g, \boldsymbol{pos}_g, \boldsymbol{\Omega}_g). \tag{22}$$

The region features $\bar{\boldsymbol{V}}_r$ are calculated in the same way as $\bar{\boldsymbol{V}}_g$. In addition, the GE module can complement the geometric information between grids and regions in a similar way to perceive the spatial relationship between them, with absolute position encodings $\boldsymbol{pos}_r$ and $\boldsymbol{pos}_g$ and relative position encoding $\boldsymbol{\Omega}_{gr}$, as shown in the yellow part of Figure 2(b). Specifically, Eq. 7 is modified to:

$$\boldsymbol{E}_{mn} = \begin{cases} \frac{\boldsymbol{Q}'_m \boldsymbol{K}'^T_n}{\sqrt{d_k}} + log(\boldsymbol{\Omega}_{mn}), & \boldsymbol{G}_{mn} = 1 \\ -\infty, & \boldsymbol{G}_{mn} = 0, \end{cases} \tag{23}$$

$$\boldsymbol{Q}'_m = \boldsymbol{Q}_m + \boldsymbol{pos}_q^m, \tag{24}$$

$$\boldsymbol{K}'_n = \boldsymbol{K}_n + \boldsymbol{pos}_k^n. \tag{25}$$

In summary, the GP module computes absolute position embeddings and relative position relationship of bounding boxes and grid coordinates through geometric transformation and incorporates them into GP-MSA through feature aggregation.

3.2.4 Guided-Attention Module

The guided-attention module establishes a connection between the two modalities, text and vision, guiding the image to focus on the regions which are relevant to the question. A set of visual features $\hat{\boldsymbol{V}}_g$ and text features $\tilde{\boldsymbol{Q}}$ are inputs. We denote the output attended visual features as $\tilde{\boldsymbol{V}}_g$, which are the output of the GA module in Figure 1. GA module computes the pairwise relationship of samples $< \hat{\boldsymbol{v}}_i^g, \tilde{\boldsymbol{w}}_i >$ by multi-head guided-attention (MGA):

$$\tilde{\boldsymbol{V}}_g = \text{MGA}(\hat{\boldsymbol{V}}_g, \tilde{\boldsymbol{Q}}, \tilde{\boldsymbol{Q}}) = \text{MSA}(\hat{\boldsymbol{V}}_g, \tilde{\boldsymbol{Q}}, \tilde{\boldsymbol{Q}}). \tag{26}$$

3.3 Answer Reasoning

According to the VE network, we obtain visual features $\tilde{\boldsymbol{V}}_g$ and text features $\tilde{\boldsymbol{Q}}$ with rich information after the image object and sentence word attention interactions. Motivated by the multimodal fusion module in [38], we use the soft attention mechanism to calculate the aggregated features $\tilde{\boldsymbol{v}}$ as:

$$\boldsymbol{\alpha} = \text{Softmax}(\text{MLP}(\tilde{\boldsymbol{V}}_g)), \tag{27}$$

$$\tilde{\boldsymbol{v}} = \sum_{i=1}^{N} \alpha_i \boldsymbol{v}_i^g. \tag{28}$$

$\tilde{\boldsymbol{q}}$ is computed in the same way as $\tilde{\boldsymbol{v}}$. $\tilde{\boldsymbol{v}}$ and $\tilde{\boldsymbol{q}}$ are fused into a unified representation $\boldsymbol{z}$ by multimodal fusion function as follows:

$$\boldsymbol{z} = \text{LayerNorm}(\boldsymbol{W}_v^T \tilde{\boldsymbol{v}} + \boldsymbol{W}_q^T \tilde{\boldsymbol{q}}), \tag{29}$$

where $\boldsymbol{W}_v$ and $\boldsymbol{W}_q$ represents two trainable projection matrices. The role of LayerNorm is to stabilize the training.

The fused feature $\boldsymbol{z} \in \mathbb{R}^{d_z}$ is fed to a fully connected layer and sigmoid activation function to generate the predicted answer vector $\boldsymbol{s} \in \mathbb{R}^N$, where N is the number of answers. We use binary cross-entropy loss (BCE) to train the model.

4 Experiments

In this section, we perform experiments to demonstrate the performance of the VE network proposed in this paper on the largest VQA benchmark dataset, VQA-v2 [8], and compare it with state-of-the-art methods to validate its effectiveness in VQA.

4.1 Experimental Settings

4.1.1 Datasets

The commonly available large benchmark dataset VQA-v2 is constructed based on MSCOCO [21] images. Each image corresponds to three questions and each question corresponds to ten answers. The dataset has about 1105904 VQA examples, of which 443757, 214354, and 447793 examples are used for training, validation, and testing respectively. In addition, two test subsets, including test-dev and test-std, are provided for online testing to evaluate performance. To measure the overall accuracy, three types of answers are considered: Number, Yes/No, and Other. Compared with VQA-v1, the distribution of the dataset is more balanced.

Table 1. Performance comparisons on the val, test-dev, and test-std splits of the VQA-v2 dataset. We compare the VE network with state-of-the-arts, including attention-based methods, graph-based methods, fusion-based methods, methods that introduce position information, etc. The top two best results are highlighted in bold.

	Val				Test-dev				Test-std
Method	All	Y/N	Num	Other	All	Y/N	Num	Other	All
UpDn [2]	63.15	80.07	42.87	55.81	65.32	81.82	44.21	56.05	65.67
Dual-MFA[24]	59.82	-	-	-	66.01	83.59	40.18	56.84	66.09
BAN [16]	66.04	-	-	-	70.04	85.42	54.04	60.52	70.35
DFAF [7]	66.66	-	-	-	70.22	86.09	53.32	60.49	70.34
ReGAT [20]	65.30	-	-	-	70.27	86.08	54.42	60.33	70.58
MCAN [38]	67.20	84.80	49.30	58.60	70.63	86.82	53.26	60.72	70.90
AGAN [42]	67.38	-	-	-	71.16	86.87	54.29	61.56	71.50
MCAN(Grid) [13]	-	-	-	-	72.59	**88.46**	55.68	62.85	-
MMNAS[37]	67.80	85.10	**52.10**	**58.90**	71.24	87.27	**55.68**	61.05	71.46
APN[35]	67.38	84.99	49.71	58.66	71.14	87.44	52.68	61.18	71.33
LENA [10]	66.59	84.35	48.71	57.79	70.31	86.63	54.26	60.22	70.48
TRAR [43]	67.70	**85.20**	49.60	-	**72.62**	88.11	55.33	**63.31**	**72.93**
MHAFN [41]	-	-	-	-	71.54	87.31	53.65	62.13	71.65
CFR [27]	**69.70**	-	-	-	72.50	-	-	-	-
Ours(VE)	**69.58**	**86.42**	**53.10**	**61.14**	**73.55**	**88.80**	**57.56**	**64.14**	**73.68**

4.1.2 Implementation Details

To extract grid features and region features, we use the improved Faster R-CNN model with ResNeXt-152 backbone [2, 13] pre-trained on the Visual Genome dataset [19]. For region features, we use zero-padding to fill the number of detected objects M to 100 if there are less than 100 candidate objects. For grid features, we average-pool them to 7×7 grid size, setting N to 49. For text features, we trim the input question to a maximum of 14 words and generate 300-dimensional word embeddings from pre-trained GloVe [29]. The word embeddings are fed into the LSTM network [11] to extract the final question features.

The framework is implemented by PyTorch. The dimensions of the input region feature d_v, grid feature d_f, question feature d_q, and fused multimodal feature d_z are 2048, 2048, 512, and 1024 respectively. The number of unit direction vectors is 8, which means k is 2. The number of heads in multi-head attention is 8. Similar to the strategy in [38], the size of the answer vocabulary is 3129. The number of encoder layers N_e and decoder layers N_d are set to 6 and 3. We train our VE network up to 15 epochs with batch size 64. For model training, we use Adam optimizer [18] with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The base learning rate is set to $1.25e^{-5}$ and increases to $5.0e^{-5}$ linearly during the first epoch and the fourth epoch and maintains $5.0e^{-5}$. After ten epochs, it decays to 1/5 of its original value every two epochs.

4.2 Experimental Results

To demonstrate the effectiveness of our VE network for VQA, we compare it with some state-of-the-art methods. In Table 1, our proposed method achieves highly competitive performance, with 69.58%, 73.55%, and 73.68% overall accuracy on the val, test-dev, and test-std set of the VQA-v2 dataset.

Table 1 shows the results of the val (left column) and test-dev and test-std splits (middle and right columns) in the VQA-v2 dataset. Specifically, Dual-MFA [24] uses a common attention mechanism to fuse the free-form image regions and detection boxes associated with the input question, lacking a comprehensive consideration of the relationship between visual features. By augmenting the grid features with region features, the VE network outperforms it by 7.54% and 7.59% on the test-dev and test-std sets, respectively. LENA [10] proposes a focusing mechanism to eliminate visual-semantic redundancy and performs vision-semantic compositionality modeling of multiple visual features. MCAN [38] and MCAN(Grid) [13] are Transfomer-based co-attention networks, which utilize different types of image features, namely regions, and grids, respectively. They stack SA and GA layers and achieve dense interaction of visual and linguistic modalities. VE network outperforms them by 3.24%, 2.92%, and 0.96% on the test-dev set, respectively, indicating that our enhanced visual features have more adequate visual semantics. ReGAT [20] and MMNAS [37] explore position relationship between object pairs to enhance feature learning, but do not take absolute position information and relative orientation into account. On the test-dev set, the VE network outperforms them by 3.28% and 2.31%, respectively. TRAR [43], CFR [27] and MHAFN [41] are recent VQA models, and we outperform them, demonstrating the strengths of our VE network.

Table 2. Performance comparisons of ablation experiments on the val and test-dev splits of VQA-v2 dataset. "R" denotes only region features are used, "G" denotes only grid features are used, and "G+R" means both are used. "GE" and "GP" represent our proposed grid enhancement module and geometric perception module, respectively.

	Val				Test-dev			
Method	All	Y/N	Num	Other	All	Y/N	Num	Other
Base(R)	68.81	85.95	50.40	60.65	72.39	88.13	54.88	62.91
Base(G)	68.38	85.72	50.24	60.00	72.16	88.14	54.08	62.61
Base(G+R)+GE	69.29	86.23	51.71	61.06	73.20	88.81	56.19	63.73
Base(G)+GP	68.52	85.68	51.61	59.94	72.51	88.27	55.05	63.00
Base(G+R)+GE+GP	69.58	86.42	53.10	61.14	73.55	88.80	57.56	64.14

4.3 Ablation Study

We perform an ablation study to verify the effectiveness of each proposed component of the whole model. The results are shown in Table 2. "Base(R)" indicates the base model with region features, which consists of the self-attention module for the question and image and the guided-attention module for the image. "Base(G)" input grid features on the base model. "Base(G+R)+GE" adds the grid enhancement module proposed in section 3.2.2 in addition

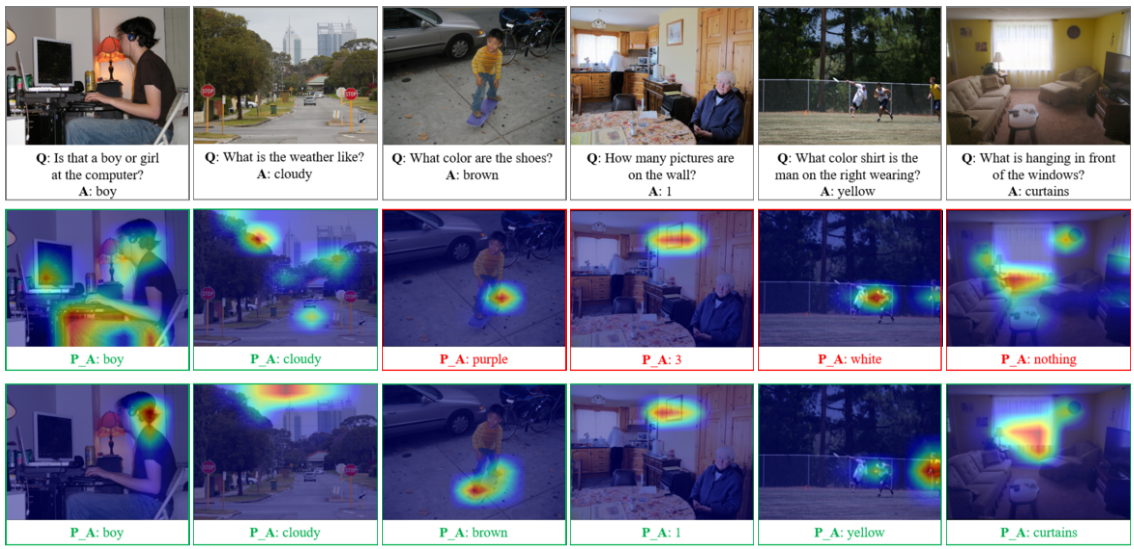

Figure 4. Visualization of some typical examples from the val set of VQA-v2 dataset. For each example, we show the original image, the given question (Q), and the ground-truth answer (A). We also display the learned visual attention map, where the highlighted object receives a higher attention weight and the predicted answer (P_A) below the attention map. The second row is the result of the "Base(G)" model described in Section 4.3. The third row is the result of our VE network.

to "Base(G)". "Base(G)+GP" incorporates the geometric perception module proposed in section 3.2.3 on the basis of "Base(G)". "Base(G+R)+GE+GP" represents the complete VQA network, which combines the grid enhancement module and geometric perception module.

"Base(G+R)+GE" outperforms "Base(R)" and "Base(G)" significantly, demonstrating the effectiveness of the grid enhancement module, which augments grid features with region features and take advantage of different granularities of visual information. "Base(G)+GP" performs better than "Base(G)", because the introduction of position information in the grid features can help focus on the objects and answer position-related questions. "Base(G+R)+GE+GP" outperforms "Base(G+R)+GE" by combining absolute position encoding and relative position information for region and grid features. It models the geometric relationship of input objects and grids explicitly and efficiently. We replace GP with other geometric relation calculations in [20] and obtain the result of 69.11% on the val split, confirming the superiority of our GP module. The final complete model "Base(G+R)+GE+GP" achieves the best performance, indicating that the two modules can provide consistent improvements, using enhanced grid features and geometric information to provide adequate evidence for VQA.

4.4 Qualitative Analysis

We visualize some attention maps generated by our model in Figure 4 and present six examples from the val set. The first row represents the original image, the corresponding question, and the ground-truth answer. The second row represents the attention map and predicted answer obtained by the "Base(G)" model described in Section 4.3, and the third row is obtained by our proposed VE network. The green box represents the correct predicted answer, and the red represents the wrong. Figure 4(a) and 4(b) show the cases where the correct answer is generated with both models, but our model focuses on the more appropriate regions that help to answer the question. In the samples of Figure 4(c) and 4(d), our VE network produces the correct answers "brown" and "1". The grid features enhanced with regions contain the surrounding information, which not only sufficiently exploits the characteristic of grid features to pay attention to details, but also incorporates the advantage that region features can accurately identify targets in images. It generates a rich visual context. Taking Figure 4(c) as an example, it can be observed that this enhancement mechanism causes the model to notice vital cues about "shoes". The questions in Figure 4(e) and 4(f) are related to relative orientation, for instance, "on the right" and "in front of", our GP module captures the position relationship and identifies objects precisely, generating answers "yellow" and "curtains". When answering such questions, our network takes the spatial relationship between objects into account, while other methods fail.

5 Conclusion

In this paper, we propose a visual enhancement network to boost VQA which includes two main modules, grid enhancement multi-head guided-attention (GE-MGA) and geometric perception multi-head self-attention (GP-MSA). GE-MGA augments the grid with its surrounding regions and generates accurate visual grid semantics with visual context, which effectively solves the semantic noise issue caused by fragmented grid features. Meanwhile, GP-MSA integrates geometric position information to help the model understand spatial relationship between objects. Extensive experiments prove that both the GE-MGA and GP-MSA deliver competitive performance improvements.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 62106035, 62206038). We also would like to thank Dalian Ascend AI Computing Center and Dalian Ascend AI Ecosystem Innovation Center for providing inclusive computing power and technical support.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi, 'Don't just assume; look and answer: Overcoming priors for visual question answering', in *CVPR*, pp. 4971–4980, (2018).

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang, 'Bottom-up and top-down attention for image captioning and visual question answering', in *CVPR*, pp. 6077–6086, (2018).

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh, 'VQA: visual question answering', in *ICCV*, pp. 2425–2433, (2015).

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, 'Neural machine translation by jointly learning to align and translate', in *ICLR*, (2015).

[5] Rémi Cadène, Hédi Ben-Younes, Matthieu Cord, and Nicolas Thome, 'MUREL: multimodal relational reasoning for visual question answering', in *CVPR*, pp. 1989–1998, (2019).

[6] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach, 'Multimodal compact bilinear pooling for visual question answering and visual grounding', in *EMNLP*, pp. 457–468, (2016).

[7] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven C. H. Hoi, Xiaogang Wang, and Hongsheng Li, 'Dynamic fusion with intra- and inter-modality attention flow for visual question answering', in *CVPR*, pp. 6639–6648, (2019).

[8] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh, 'Making the V in VQA matter: Elevating the role of image understanding in visual question answering', in *CVPR*, pp. 6325–6334, (2017).

[9] Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu, 'Normalized and geometry-aware self-attention network for image captioning', in *CVPR*, pp. 10324–10333, (2020).

[10] Yudong Han, Yangyang Guo, Jianhua Yin, Meng Liu, Yupeng Hu, and Liqiang Nie, 'Focal and composed vision-semantic modeling for visual question answering', in *ACM Multimedia*, pp. 4528–4536, (2021).

[11] Sepp Hochreiter and Jürgen Schmidhuber, 'Long short-term memory', *Neural Computation*, 1735–1780, (1997).

[12] Drew A. Hudson and Christopher D. Manning, 'GQA: A new dataset for real-world visual reasoning and compositional question answering', in *CVPR*, pp. 6700–6709, (2019).

[13] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik G. Learned-Miller, and Xinlei Chen, 'In defense of grid features for visual question answering', in *CVPR*, pp. 10264–10273, (2020).

[14] Luoqian Jiang, Yifan He, and Jian Chen, 'Text-aware dual routing network for visual question answering', *CoRR*, **abs/2211.14450**, (2022).

[15] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick, 'CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning', in *CVPR*, pp. 1988–1997, (2017).

[16] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang, 'Bilinear attention networks', in *NeurIPS*, pp. 1571–1581, (2018).

[17] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang, 'Hadamard product for low-rank bilinear pooling', in *ICLR*, (2017).

[18] Diederik P. Kingma and Jimmy Ba, 'Adam: A method for stochastic optimization', in *ICLR*, (2015).

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei, 'Visual genome: Connecting language and vision using crowdsourced dense image annotations', *International Journal of Computer Vision*, 32–73, (2017).

[20] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu, 'Relation-aware graph attention network for visual question answering', in *ICCV*, pp. 10312–10321, (2019).

[21] Yuetan Lin, Zhangyang Pang, Donghui Wang, and Yueting Zhuang, 'Feature enhancement in attention for visual question answering', in *IJCAI*, pp. 4216–4222, (2018).

[22] Fenglin Liu, Yuanxin Liu, Xuancheng Ren, Xiaodong He, and Xu Sun, 'Aligning visual regions and textual concepts for semantic-grounded image representations', in *NeurIPS*, pp. 6847–6857, (2019).

[23] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh, 'Hierarchical question-image co-attention for visual question answering', in *NIPS*, pp. 289–297, (2016).

[24] Pan Lu, Hongsheng Li, Wei Zhang, Jianyong Wang, and Xiaogang Wang, 'Co-attending free-form regions and detections with multimodal multiplicative feature embedding for visual question answering', in *AAAI*, pp. 7218–7225, (2018).

[25] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji, 'Dual-level collaborative transformer for image captioning', in *AAAI*, pp. 2286–2293, (2021).

[26] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz, 'Ask your neurons: A neural-based approach to answering questions about images', in *ICCV*, pp. 1–9, (2015).

[27] Binh X. Nguyen, Tuong Do, Huy Tran, Erman Tjiputra, Quang D. Tran, and Anh Nguyen, 'Coarse-to-fine reasoning for visual question answering', in *CVPR Workshops*, pp. 4557–4565, (2022).

[28] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han, 'Image question answering using convolutional neural network with dynamic parameter prediction', in *CVPR*, pp. 30–38, (2016).

[29] Jeffrey Pennington, Richard Socher, and Christopher D. Manning, 'Glove: Global vectors for word representation', in *EMNLP*, pp. 1532–1543, (2014).

[30] Mengye Ren, Ryan Kiros, and Richard S. Zemel, 'Exploring models and data for image question answering', in *NIPS*, pp. 2953–2961, (2015).

[31] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun, 'Faster R-CNN: towards real-time object detection with region proposal networks', in *NIPS*, pp. 91–99, (2015).

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, 'Attention is all you need', in *NIPS*, pp. 6000–6010, (2017).

[33] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel, 'Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks', in *CVPR*, pp. 1960–1968, (2019).

[34] Mingrui Wu, Xuying Zhang, Xiaoshuai Sun, Yiyi Zhou, Chao Chen, Jiaxin Gu, Xing Sun, and Rongrong Ji, 'Difnet: Boosting visual information flow for image captioning', in *CVPR*, pp. 17999–18008, (2022).

[35] Xu Yang, Chongyang Gao, Hanwang Zhang, and Jianfei Cai, 'Auto-parsing network for image captioning and visual question answering', in *ICCV*, pp. 2177–2187, (2021).

[36] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola, 'Stacked attention networks for image question answering', in *CVPR*, pp. 21–29, (2016).

[37] Zhou Yu, Yuhao Cui, Jun Yu, Meng Wang, Dacheng Tao, and Qi Tian, 'Deep multimodal neural architecture search', in *ACM Multimedia*, pp. 3743–3752, (2020).

[38] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian, 'Deep modular co-attention networks for visual question answering', in *CVPR*, pp. 6281–6290, (2019).

[39] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao, 'Multi-modal factorized bilinear pooling with co-attention learning for visual question answering', in *ICCV*, pp. 1839–1848, (2017).

[40] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao, 'Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering', *IEEE Transactions on Neural Networks and Learning Systems*, 5947–5959, (2018).

[41] Haiyang Zhang, Ruoyu Li, and Liang Liu, 'Multi-head attention fusion network for visual question answering', in *ICME*, pp. 1–6, (2022).

[42] Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Gen Luo, Xiaopeng Hong, Jinsong Su, Xinghao Ding, and Ling Shao, 'K-armed bandit based multi-modal network architecture search for visual question answering', in *ACM Multimedia*, pp. 1245–1254, (2020).

[43] Yiyi Zhou, Tianhe Ren, Chaoyang Zhu, Xiaoshuai Sun, Jianzhuang Liu, Xinghao Ding, Mingliang Xu, and Rongrong Ji, 'TRAR: routing the attention spans in transformer for visual question answering', in *ICCV*, pp. 2054–2064, (2021).