Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-18871-z

Advanced Visual and Textual Co-context Aware Attention Network with Dependent Multimodal Fusion Block for Visual Question Answering

Hesam Shokri Asri[1] · **Reza Safabakhsh**[1]

Received: 30 October 2023 / Revised: 10 February 2024 / Accepted: 4 March 2024

Abstract
Visual question answering (VQA) is a multimodal task requiring a simultaneous understanding of both visual and textual content. Therefore, image and question comprehension, finding a dense interaction among words and regions, and inference knowledge are the cores of VQA. In this paper, we propose the Advanced Visual and Textual Co-context Aware Attention Network with Dependent Multimodal Fusion Block for Visual Question Answering (ACOCAD), consisting of the image and the question representations and three proposed mechanisms: textual context-aware attention, a question-level & word-level visual attention, and a dependent multimodal fusion block. The textual context-aware attention mechanism marks the keywords of the question and captures rich features by modeling a context-aware unit beside the Universal Sentence Encoder model (USE) and a self-attention unit. The advanced visual attention approach is applied to attend on the regions with the aim of question-level and word-level visual attention. The dependent multimodal fusion block is employed to enhance associating keywords with key regions and generate more efficient vectors. Three sub-models are defined based on the three proposed mechanisms, and one ablation study is conducted on the benchmarks GQA and VQA-v2 datasets to evaluate the effectiveness of each mechanism of our ACOCAD model. Then, another ablation study for the overall accuracy of the ACOCAD model is carried out on one of the hyper-parameters to find its optimal value. Moreover, we explore how the Dependent Multimodal Fusion Block may relieve limitations of prior methods in answering questions including homograph words. In addition, to address the challenge regarding the length of question words, the potential efficiency of the USE model and the Visual Attention Mechanism are analyzed. For further review, a qualitative evaluation is done to visualize the effectiveness of the ACOCAD model using some samples. The results demonstrate that the ACOCAD model outperforms four out of seven criteria in the GQA dataset, and its overall accuracy criterion reaches 57.37%. Furthermore, our model achieves a significant enhancement compared to the previous state-of-the-art models and reaches 87.43%, 71.02%, and 71.18% accuracies in the Yes/No question type, overall test-dev dataset, and overall test-std of VQA-v2, respectively. Moreover, one of these sub-models attains the best accuracy of 60.95% among all models for the other question type.

Extended author information available on the last page of the article

Keywords Dependent multimodal fusion block · Question-level and word-level visual attention mechanisms · Textual context-aware attention · Universal sentences encoder

1 Introduction

Computer vision (CV) and natural language processing (NLP) are the most important and challenging fields in artificial intelligence. Computer vision consists of a set of methods to get meaningful data from visual input. The most popular tasks in this field, that receive significant success, are image classification [1], image segmentation [2], medical diagnosis [3, 4], and image processing [5, 6]. On the other hand, NLP aims to process text data and understand human language as it is spoken and written, mainly including Text classification [7], machine translation [8], and extracting data to analyze for health outcomes [9]. In the last few years, many researchers have become more enthusiastic to work on multimodal tasks which are a combination of visual and textual information. Image captioning problem [10, 11] is one of the multimodal tasks receiving an image as the input and producing a general and brief description as its output. Despite the success of image captioning, this method is not devoid of challenges, for example, a produced caption may not contain any details of the image, but the user usually intends to have access to the information of a specific image region. In order to address this issue, visual question answering (VQA) is introduced which allows the user to ask a question about each region of the image [12]. In fact, VQA takes an image and a question as inputs and generates an answer in the output presented in Fig. 1 Despite having more complexities, VQA possesses considerable advantages over image captioning. Evaluation metrics are easier to quantify in the VQA compared to image captioning because the answer to the VQA problem mostly includes one or at most three words [12, 13].

The VQA is an intersection of three fields, namely, computer vision, NLP, and inference knowledge [13]. Thus, it demands extensive knowledge of artificial intelligence to answer questions. In the following, we mention the main challenges that are considered in the VQA problem.

Image comprehension Visual semantic understanding is the ability of a machine to be concerned with the extraction of meaning from images, which is greatly addressed by the convolutional neural network (CNN) [14]. Most of the proposed methods in the VQA employ CNN or pretrained CNN in their models.

Question comprehension Human advantages over computer systems are question comprehension and sentence production. A word might convey different meanings in different contexts when placed beside other sets of words. Thus, it is another challenge for the VQA to comprehend the input question and understand what the question is about. Primary models used the bag of words method [15] to represent text data, then models were promoted and employed the Long Short Term Memory (LSTM) network [16] and word2vec model [24].

Inference knowledge Questions in the VQA are not usually simple such as asking about colors, objects, or the number of objects. For example, in a question such as "Does this man have a vision problem?", the model is required to detect the glasses on the person's eyes, and then, it should infer that the glasses on the person's eyes indicate a vision problem. The above example denotes that the model requires an ability to perform deep and complex inferences. Recent VQA methods use external

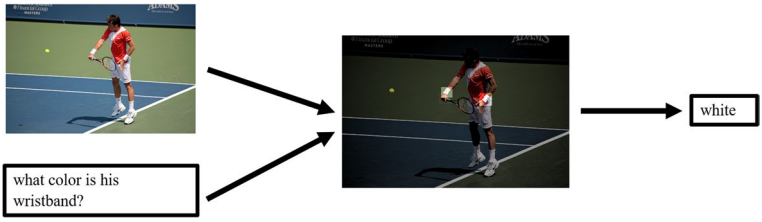

Fig. 1 An example of a VQA task

knowledge to achieve a more inferential model. To address this challenge, word2vec is used as an external knowledge to achieve a more inferential model [17].

Interaction of image and question Finding the relationship between the question and the image to understand which region of the image is related to the question is an important issue. Thus, the VQA problem needs a model to recognize a few regions of the image that are more relevant to the input question. In this respect, the visual attention mechanism is used to attend on the image based on the question and gives higher weights to the important regions [18, 19]. To improve the performance of the models, the co-attention mechanism employs visual and textual attention mechanisms simultaneously to identify keywords of the question [20, 21]. In another type of co-attention mechanism, a dense interaction is applied between each image region and question word in order to increase the relation between images and questions outperforming the previous mechanisms. The co-attention mechanism is modeled in many different ways which the novel models are inspired by the self-attention unit and multimodal encoder-decoder network [22, 23].

Contribution Based on the existing challenges, we propose a model named Advanced Visual and Textual Co-context Aware Attention Network with Dependent Multimodal Fusion Block (ACOCAD) which is employed in order to make some positive alterations in different sections of the previous VQA structures, such as adding the Universal Sentence Encoder (USE) [25] to both the question representation and the textual attention section, employing both word-level and question-level visual attention mechanisms, and the Dependent Multimodal Fusion (DMF) block.

In this respect, the main contributions of this paper can be summarized as follows:

1) We devise a textual context-aware attention to attend on the input question. Accordingly, besides the self-attention mechanism, which is implemented by default, a textual context-aware attention mechanism is applied in this module by adding the USE model in the form of Context-Aware Guide Attention unit (CAGA) to extract more meaningful features. This is because adding the USE model as another piece of knowledge to our model and employing it in the context-aware attention form leads to two attention stages on the input questions which results in more interpreted and discriminant features.
2) Despite all the advantages of the dense interaction mechanism, considering all relations between words of a question and all regions can transmit noisy information to our model because if the question is very long, it can impose over-interaction on our model which leads to distraction of the model from the correct direction, especially in simple questions. Therefore, we employ two levels of attention, word-level and question-level visual attention mechanisms using Question-level Guide Attention (QLGA) and Word-

level Guide Attention (WLGA) units to reduce the impact of the noisy information and generate highly constructive features.

3) Unlike the previous studies, which commonly employ an independent multimodal fusion mechanism, our model employs the DMF block to generate an efficient combination of the final vectors of the regions and words. Accordingly, the words of the question are revised based on the image regions.
4) Additionally, we independently examine the results of each sub-model to demonstrate the model's efficiency across various question types and criteria, showing that its performance surpasses the previous state-of-the-art models when evaluated with the two well-known datasets VQA-v2 [26] and GQA [27].

2 Related Work

VQA challenges and main mechanisms were briefly discussed in the introduction section. In this section, we consider in detail these mechanisms and review their important methods in each mechanism. All the methods can be divided into four main categories named non-attention mechanism, a visual attention mechanism, a co-attention mechanism, and a co-attention mechanism based on dense interaction which will be explained in the following.

2.1 Non-Attention Mechanism

Methods including this mechanism merely employ a joint embedding approach in their model which means, globally extracted features from the image and question are merged. For example, authors in [28] proposed the BOWING model in which the question and image features are concatenated together in a common layer and given to the prediction layer. This method is one of the simplest methods proposed in the VQA problem. To consider word order in the question, Ren et al. [29] propose an Image + LSTM Method. The architecture of this method is based on an LSTM network in which words and an image are entered, respectively. The limitation of the non-attention mechanism is that it employs a global feature to represent the input image. This might lead to the transfer of unnecessary information to the prediction layer. The attention mechanism approach is introduced to address this issue.

2.2 Visual Attention Mechanism

The human visual system focuses on prominent areas helping the person to understand visual input quickly and greatly. In this respect, the purpose of the visual attention mechanism is to use local features of the image [18]. Indeed, the image is firstly segmented into some regions, and then each region is represented in a feature vector and weighted based on its importance and relevance to the global feature of the question. Therefore, this mechanism is more efficient than the joint embedding approach because it extracts several features from the image and weights them to decrease the probability of transferring unnecessary information to the prediction layer.

There are two types of image segmentations: (a) segmenting the input image into k same-sized regions [19], and (b) object-based segmentation type, in which the image is first given to a Fast R-CNN [30] for segmenting into meaningful regions. The Fast R-CNN is trained by the visual genome database [31] to focus on specific regions of the image where

objects are highly probable to be presented [32]. Figure 2 shows the difference between the two segmentation types. The first type of segmentation is used in [19] where each extracted region vector is independently concatenated to the global question vector, and then these vectors are weighted after passing through fully connected layers. In order to improve this method, Yang et al. [33] propose an iterative method named Stacked Attention Network (SAN) to update the global question feature in each iteration. Authors in [34, 35] devised a more complex combination with this assumption that point-wise addition and multiplication may not greatly denote the relationship among different features. Consequently, they use compact bilinear pooling [36] and Low-rank bilinear pooling using Hadamard product [37] to combine features, in order to attend on image regions. Later bottom-up and top-down attention (BUTD) [32, 38] method was proposed based on the object-based segmentation which outperforms the above studies.

2.3 Co-attention mechanism

Taking the question into account and weighing words can increase accuracy. This is due to the fact that some words of the question are usually more important such as the type of question or objects. Therefore, it is vital to consider the differences among the words in the question and weight them to generate an efficient global question vector. Consequently, visual and textual attention is applied to this mechanism simultaneously. In the study [21] authors added textual attention to their model and introduced a hierarchical architecture in which an attention mechanism was applied to the question and image sequentially. Zhou et al. [39] proposed the developed version of bilinear pooling as multimodal factorized bilinear pooling (MFH) to achieve effective fusion. Also, they employed a textual self-attention approach to generate an attended final question vector. Then, the visual attention approach was employed by this final vector of the question. Co-Attention Network with Question type (CAQT) [40] is another method containing two main contributions: (a) using question type in the last layer of the classifier layer; (b) employing a self-attention mechanism in the input question using bi-LSTM network. In this study, the question is given to the bi-LSTM network. Then each cell of its network represents the entered word in the form of a new vector. All of these vectors are weighed after passing through a linear layer and a Softmax layer. Each weight is multiplied by its corresponding word vector, and finally, all vectors are summed together, and the final vector of the question is generated. Despite the success of this mechanism, there is a weakness in the interaction of the image

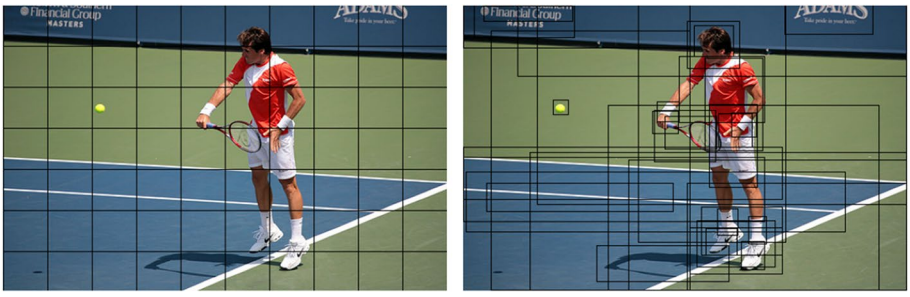

Fig. 2 (**a**) The k same-sized regions segmentation; (**b**) Object-based segmentation

and the question in which fine-grained correlation between each region and each word is not considered, therefore this issue leads to a lack of interaction.

2.4 Co-attention Mechanism based on dense interaction

To address the previous issue, dense co-attention models that calculate the complete interaction between each image region and each question word were proposed. Compared to the previous co-attention methods, which have coarse-grained interaction with the image, these methods benefit from fine-grained interaction. In this respect, the BAN [41] model was proposed, which multiplies each word vector in each region vector of the image to create an attention map. The method uses a deep cascaded model to comprehend complex interaction between the image and question such that vectors of regions and words are updated eight times by these attention maps. The modular co-attention networks (MCAN) [22] method was inspired by the transformers to improve question features, image features, and their interaction with each other. Accordingly, they proposed a novel architecture containing self-attention unit (SA). In the SA unit, words and regions attend to themselves. Moreover, the SA unit is used to apply dense interaction between words and regions. These units are implemented in the form of a cascaded and deep network. Chen et al. [23] proposed the Multimodal Encoder Decoder Attention Network method (MEDAN) employing an encoder module to attend on the question and the decoder module to attend on the image.

Authors in [42] only need a few keywords and regions to predict the correct answer because considering all the composition of the regions. Hence, they proposed the threshold-based Sparse Co-Attention Visual Question Answering Network (SCAVQAN) in which some of the regions and words are filtered based on the specified threshold before calculating the relation between regions and words. Therefore, only the important regions and words remain, and the interaction of these regions and words is calculated. It should be noted that the last three models achieve the highest accuracies.

3 Methodology

In this section, we propose our model named ACOCAD. Based on the above explanations, it can be concluded that employing a co-attention mechanism based on dense interaction is more efficient than the previous mechanisms. Hence, we construct our model architecture inspired by this improved model. However, this mechanism has a potential problem of transferring noisy information to the model due to its tendency for over-interaction. To decline this drawback, we incorporate two levels of visual attention mechanisms to increase overall efficiency. Furthermore, since external knowledge has the potential to elevate the efficacy of the method, the USE model, as an external pre-trained model, is added to our model to contribute to the refinement of the word attention mechanism, boost the model generalization, and embed sentences to convey external knowledge to other NLP tasks.

We first extract the question and image features and represent the corresponding vectors. Then the textual and visual context-aware attention mechanism is applied. Finally, we employ the DMF block to fuse region and word features. The fused vector is input to the classifier layer to predict the correct answer. The overall diagram of our model is shown in Fig. 3

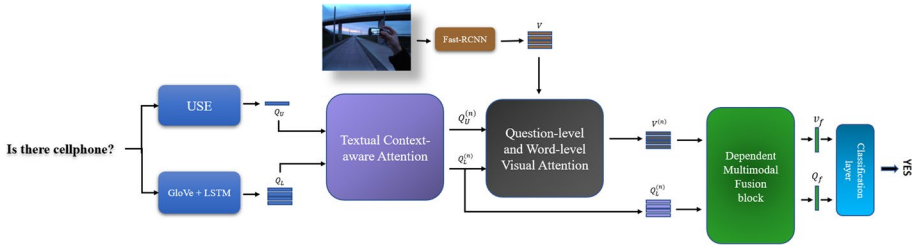

Fig. 3 The overall diagram of the Advanced Visual and Textual Co-context Aware Attention with Dependent Multimodal Fusion Block for Visual Question Answering (ACOCAD)

3.1 Question Representation

In the VQA task, questions have an unofficial format and informal writing. Therefore, we utilize common preprocessing techniques such as removing punctuation and stop words, as well as revising abbreviations and digits to rewrite them to a consistent structure. These steps were taken to prepare the data for subsequent stages. Then, the question Q is given to the word embedding layer which is pre-trained with an enormous corpus for converting each word to a vector. In our model, we use the GloVe model as an embedding layer to represent each word of the question with a 300–dimensional vector. Considering the fact that the embedding block is numerically limited, if the number of words in the question is more than the number of available blocks in the embedding layer, some words will be ignored. In contrast, if the number of words becomes less than the number of blocks, some blocks include zero padding Ref.[32, 42] has shown that the optimal numbers are 14 and 29 for the GQA and VQA-v2 datasets respectively. In the following, due to the prevention of misunderstanding and redundancy, we formulate our model based on the VQA-v2 dataset. Therefore, we have taken into account the output of the embedding layer $\boldsymbol{Q_e} \in R^{14*300}$. Then the output of the word embedding layer is passed through an LSTM network. Although the LSTM network causes the problem of giving a higher weight to the end block, the model addresses that by taking all outputs of the LSTM blocks and using them in the attention mechanism to weight them more rationally. Each LSTM block represents an input vector having 512 dimensions. Based on the length of the embedding layer, the output of the LSTM network is $\boldsymbol{Q_L} \in R^{14*512}$.

As the results of Ref. [44] show the close performance of the USE model to other word embedding methods such as GloVe [43] and word2vec in common NLP tasks, our proposed architecture incorporates the USE model as an additional network as depicted in Fig. 3. This inclusion aims to extract features from questions and transfer knowledge from other NLP sources to improve our model. Therefore, effective employing of this pre-trained sentence2vec along with other word2vec models can enhance the performance of the initial models to leverage their advantages. The USE model is trained using Wikipedia, web news, web question-answer pages, discussion forums, and Stanford natural language datasets which is available at https://tfhub.dev/google/universal-sentence-encoder/4. This pre-trained network takes a question as an input and represents it as a 512-dimensional vector, $\boldsymbol{Q_U} \in R^{1*512}$. $\boldsymbol{Q_e}$, $\boldsymbol{Q_L}$ and $\boldsymbol{Q_U}$ can be given as follow:

$$\boldsymbol{Q_e} = Glove(Q) \tag{1}$$

$$\boldsymbol{Q_L} = LSTM\left(\boldsymbol{Q_e}\right) \tag{2}$$

$$\boldsymbol{Q_U} = USE(Q) \tag{3}$$

Although the numbers 14 and 29 are determined as the optimal numbers for embedding layer, a potential limitation of some VQA methods, including our model, arises when the word length of the question exceeds these optimal numbers. Accordingly, certain words that can be vital in questions, may need to be ignored.

3.2 Image Representation

As discussed, it is concluded that object-based segmentation performs better than k same-sized type. As a result, we use a pre-trained object-based segmentation in this paper. Each image is represented by k 2048-dimensional vectors by setting k to 36 which is an optimal number [32]. Using a pre-trained Fast R-CNN rather than the common CNN reduces the number of network parameters, and increases the training speed and network response. To obtain better features, a normalizing layer is added to these vectors. As shown in Fig. 3, the pre-trained Fast R-CNN is used in our model to represent input image $\boldsymbol{I}$ as a feature $\boldsymbol{V} \in R^{k*2048}$:

$$\boldsymbol{V} = FRCNN(\boldsymbol{I}) \tag{4}$$

Even though leveraging Fast R-CNN offers benefits such as reducing the transmission of unnecessary information from input images to models and decreasing the number of learnable parameters, a potential downside is the possibility of ignoring some important regions in this process considered as one of the weaknesses of the model. Despite this limitation, the Fast R-CNN performs excellently on most VQA databases.

3.3 Textual Context-aware Attention Mechanism

The textual context-aware attention mechanism is a method of weighing text words according to their importance which is depicted in Fig. 3. The network learns to assign more weights to the question keywords. Therefore, they have a greater effect on the final vector of the question. Inspired by the self-attention unit, and based on the Multi-Head Attention [45], we employ a combination module attending to the question words forrepresenting them in the rich and discriminative features. This module contains two main units and a feedforward layer at the end. Unlike the previous methods merely employed one stage of attention, SA unit, our model introduces an additional unit of context-aware attention named CAGA. This unit aims to generate more meaningful vectors.

At first, the CAGA is applied to the question, where all words are attended based on the general concept of the question – the USE model in this case. The words are then assigned weights and represented according to their importance in questions. The CAGA unit helps to enhance inference and the general understanding of our model by adding external knowledge. In Fig. 4 , the CAGA unit which is composed of the Multi-Head Attention module, skip connection layer, and normalization layer, is shown. As the Multi-Head Attention module is illustrated in Fig. 5, this module generally takes three inputs named key (**k**), query (**q**) and value (**v**) which can be initialized with the same or different values. Each head attends to the input words based on its learnable weight, and the outputs of all heads are eventually concatenated

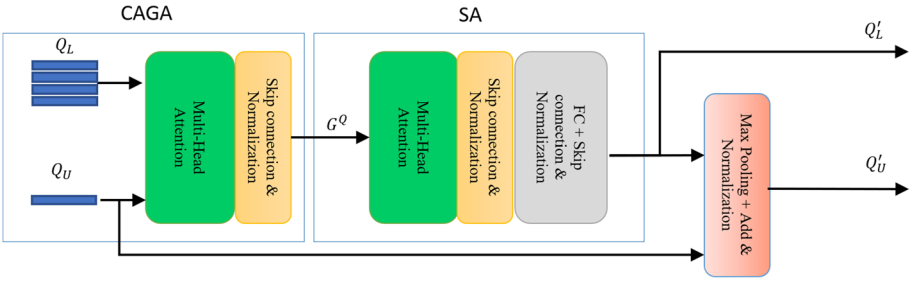

Fig. 4 Textual Context-aware Attention module using USE model, including context-aware guide attention unit (CAGA), self-attention unit (SA), and Maxpooling layer

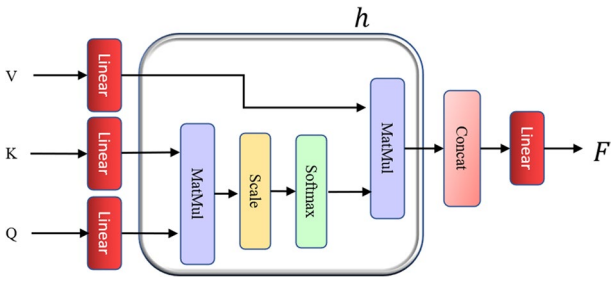

Fig. 5 The architecture of Multi-Head Attention module

with each other and passed through a linear projection layer. Therefore, the final vector of this module is formed.

In the CAGA unit, the vector F is the output of the mentioned Multi-Head Attention module which $q = v$:

$$\boldsymbol{F} = MultiHeadAtt(\boldsymbol{Q_U}, \boldsymbol{Q_L}, \boldsymbol{Q_U}) \tag{5}$$

where $\boldsymbol{F} \in R^{14\times512}$. To sum up, $\boldsymbol{F}$ represents the attended vectors of the question words using the USE model. Then, the skip connection layer is applied to the $\boldsymbol{F}$ vector to address vanishing as well as inflating gradients, and finally the normalization layer is appended to stabilize the training process. According to Fig. 4, our textual attention network attends on the question in two steps. In the first step, as explained above, we design textual context-aware attention using the CAGA unit imposed on the word vectors. Then, as a common procedure, self-attention using the SA unit is employed in the output of the CAGA to be reattended. In this work, we benefit from both approaches of self-attention and context-aware attention mechanism targeting different aspects of input questions. Additionally, to generate a general vector of the question and use it in the question-level visual attention mechanism on the regions, the attended word vectors $\boldsymbol{Q_L'} = [w_1, w_2, \ldots, w_{14}]$ are passed through a max pooling layer Then, it is added to $\boldsymbol{Q_U}$ to maintain its generalization. $\boldsymbol{Q_L'}$ and $\boldsymbol{Q_U'}$ are the outputs of the textual attention network.

$$\boldsymbol{G^Q} = CAGA\left(\boldsymbol{Q_U}, \boldsymbol{Q_L}\right) \tag{6}$$

$$\boldsymbol{Q_L'} = SA(\boldsymbol{G^Q}) \tag{7}$$

$$Q'_U = Normalize(MaxPooling\left(Q'_L\right) + Q_U) \tag{8}$$

where G^Q and $Q'_L \in R^{14\times512}$, $Q'_U \in R^{1\times512}$. Q'_U includes the general meaning and concept of questions and Q'_L contains more detailed information and keywords of questions.

3.4 Question-level & Word-level Visual Attention Mechanism

In this section, we explain the advanced visual attention mechanism based on both question-level and word-level visual attention approaches shown in the overall diagram. In this respect, regions of the image that are more relevant to the question are specified and higher weights are assigned to them.

Accordingly, two attention approaches are used in this study with two different purposes named: (a) the question-level visual attention on the regions; and (b) the word-level visual attention on the regions. Each region is attended by the general concept of the question in the question-level visual attention mechanism, therefore, the regions are weighted based on their importance in the question. This helps the model to extract coarse-grained features from the regions which is carried out by the QLGA unit using the main vector of the question which was made by the max pooling layer and USE vector formulated in Eq (8). On the other side, each region is attended by all the words in the word-

level visual attention mechanism, and the weight of each region is computed by the relation with each word. Hence, the fine- grained features are extracted from the regions using the WLGA unit. In most cases, the model in simple questions only needs to find the relation among some keywords and key regions, and it is relatively enough to predict the correct answer, however, this dense interaction sometimes transfers noisy information to the model. In contrast, finding that relation is not necessarily enough to get the correct answer in complicated questions. This is why adding the general concept of the question as an extra inference is useful to help the model to predict the correct answer and reduce the effect of the noisy information by detecting the important words in questions. Hence, a question-level visual attention mechanism is applied to the model in addition to the word-level one to the regions. To sum up, we employ the two stages of visual attention so that the first QLGA unit attends and filters some of the regions, and the second WLGA unit reattends the filtered regions and weighs them. As shown in Fig. 6, our question-level and word-level visual attention mechanisms consist of three units designed based on the Multi-Head Attention module. In the first step, the question-level visual attention mechanism is employed on the image regions by the QLGA unit to enhance the generalization of the model.

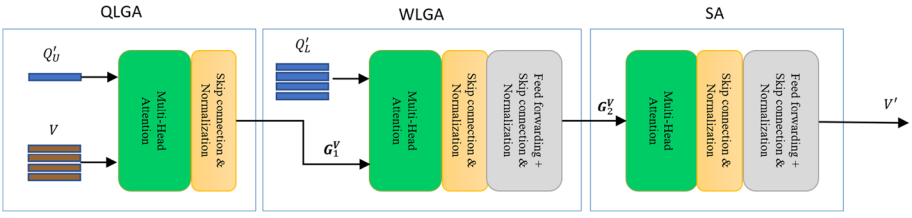

Fig. 6 Question-level and Word-level Visual Attention module including question-level guide attention (QLGA), word-level guide attention (WLGA), and self-attention (SA) units

$$\boldsymbol{G}_1^V = QLGA(\boldsymbol{Q}_U^{'}, \boldsymbol{V}) \tag{9}$$

where $G_1^V \in R^{36\times512}$. In the next step, the word-level visual attention mechanism is employed on the image regions by modeling the second WLGA unit as follows to extract fine-grained features:

$$\boldsymbol{G}_2^V = WLGA(\boldsymbol{Q}_L^{'}, \boldsymbol{G}_1^V) \tag{10}$$

where $\boldsymbol{G}_2^V \in R^{36\times512}$. Then, the self-attention mechanism is applied to the image regions to improve the extracted features of the two former units. Finally, the skip connection and layer normalization are employed to represent the final visual attention vector $\boldsymbol{V}^{'}$ as follows:

$$\boldsymbol{V}^{'} = SA(\boldsymbol{G}_2^V) \tag{11}$$

where $\boldsymbol{V}^{'} \in R^{36\times512}$. We can implement deep cascaded form of the visual and textual attention modules in which the input of the further textual attention module is the output of the previous one and so on for the image attention module. $\boldsymbol{Q}_L^{(n-1)} \rightarrow \boldsymbol{Q}_L^{(n)}$ and $\boldsymbol{V}^{(n-1)} \rightarrow \boldsymbol{V}^{(n)}$, where n is the depth of these cascades.

3.5 Dependent Multimodal Fusion Block and Output Classifier

In the preceding sections, we identified salient regions within images and highlighted keywords by input questions. As we approach the final stage, our purpose is to provide answers to questions. The process of answering relies on both words and regions. This is because responding to questions without considering input images is inherently flawed. Similarly, relying solely on salient regions poses significant risks because different questions with diverse concepts and types can refer to the same regions. Hence, the fusion of both keywords and regions is crucial for accurately answering questions. In this respect, we need to use a proper multimodal fusion block which is an integral part of the VQA problem.

One of the common modules for fusing multimodal problems is the independent multimodal fusion (IMF) block which is used by Ref. [22, 23]. In this module, before fusing regions and words, the question words and the regions are attended independently, and the words of the question are weighted just based on the question. However, the dependency aspect between the regions and words is neglected in the IMF block. For example, some words such as “bat” and “match” have different meanings in different sentences, homograph words, which make the question ambiguous for the model. For instance, the question “What color is the bat?” the word “bat” is vague to be answered because the target of this question can be an animal or the stuff which is used to play cricket. Therefore, attending to the question based on its corresponding region helps to understand the purpose of the question. Although the USE model reduces this ambiguity, it can happen. Thus, simultaneously attending to the question and the image not only clarifies this ambiguity more, but also improves the accuracy of weighting the words. This is because each word of the question is weighted based on not only the inner attention and general concept of the question, but also based on its relevance to the image. In this respect, to deal with this challenge which is not considered in IMF blocks [22, 23, 42], we propose the DMF block which attends on the words and regions concurrently to revise the weights and the word vectors based on image regions. Also, we do the same method for the regions to weight them based on the words. To sum up, one of our motivations

for applying the DMF block is to attend on words based on regions. Moreover, another reason is to fuse two different types of vectors in our multimodal problem which are the image and text vectors.

In this regard, the extracted vectors $\boldsymbol{V}^{(n)}$ and $\boldsymbol{Q}_L^{(n)}$ enter the DMF block as shown in Fig. 7. Then, the regions and words are passed through a linear layer to compress the same size, and finally they are multiplied to each other, and the attention map is calculated.

$$\boldsymbol{A} = (\boldsymbol{W}_r \boldsymbol{V}^{(n)})^T (\boldsymbol{W}_w \boldsymbol{Q}_L^{(n)}) \tag{12}$$

where $A \in R^{36\times14}$ determines the similarity of each region and the word. The calculated attention map is summed once along the row side and once along the column side

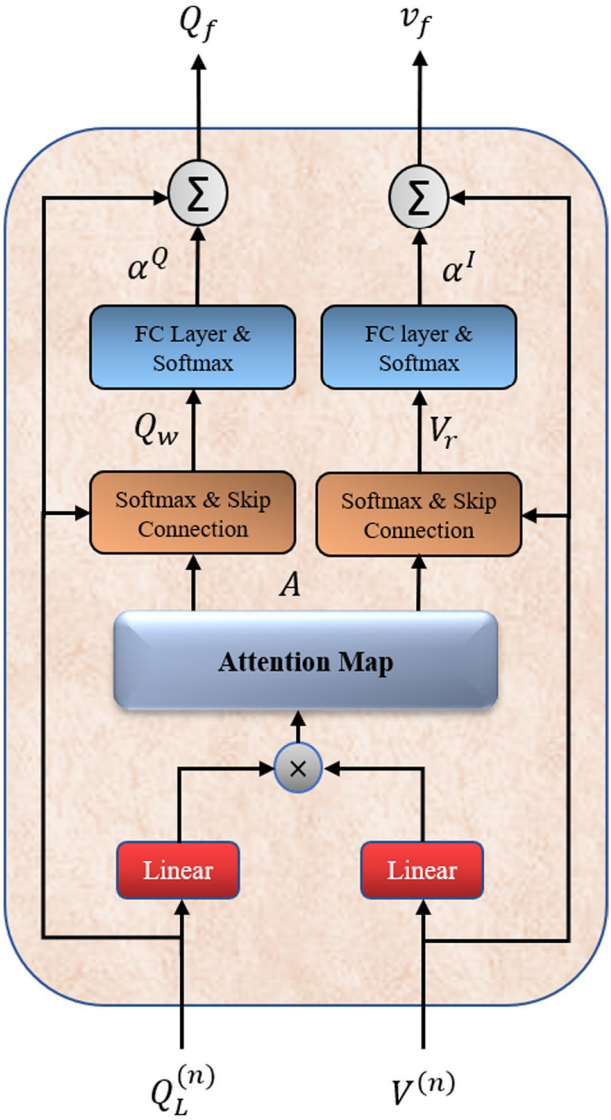

Fig. 7 Dependent Multimodal Fusion (DMF) block

to compute cumulative attention maps for each region and word. These maps are passed through softmax and add layers to represent the vectors of the words and regions.

$$\boldsymbol{V_r} = softmax\big(sum_{col}(\boldsymbol{A})\big).\boldsymbol{V}^{(n)} + \boldsymbol{V}^{(n)} \tag{13}$$

$$\boldsymbol{Q_w} = softmax\big(sum_{row}(\boldsymbol{A})\big).\boldsymbol{Q}_{\boldsymbol{L}}^{(n)} + \boldsymbol{Q}_{\boldsymbol{L}}^{(n)} \tag{14}$$

To calculate the importance of the words and regions,$\boldsymbol{V_r}$ and $\boldsymbol{Q_w}$ enter an FC and a Softmax layer, respectively. $\boldsymbol{\alpha^Q}$ and $\boldsymbol{\alpha^I}$ determine final weights of each region and word:

$$\boldsymbol{\alpha^Q} = Softmax(MLP\big(\boldsymbol{V_r}\big)) \tag{15}$$

$$\boldsymbol{\alpha^I} = Softmax(MLP\big(\boldsymbol{Q_w}\big)) \tag{16}$$

where $\boldsymbol{\alpha^Q} \in R^{1\times14}$ and $\boldsymbol{\alpha^I} \in R^{1\times36}$. Figure 11 shows some samples of these obtained weights. These weights are multiplied in corresponding vectors and summed together. The final vectors of image $\boldsymbol{v_f}$ and question $\boldsymbol{Q_f}$ are formed as:

$$\boldsymbol{v_f} = \sum_{i=1}^{14} \alpha_i^Q.\boldsymbol{V}_i^{(n)} \tag{17}$$

$$\boldsymbol{Q_f} = \sum_{i=1}^{36} \alpha_i^I.\boldsymbol{Q}_{\boldsymbol{L}(i)}^{(n)} \tag{18}$$

where $\boldsymbol{v_f}$ and $\boldsymbol{Q_f} \in R^{1\times512}$**.** There are three common methods to fuse these final vectors: element-wise multiplication, concatenation, and summation. Previous studies [22, 23, 41] have demonstrated that the summation method is more efficient than other techniques in the VQA problem. Consequently, we employ the summation method to fuse the final question, $\boldsymbol{Q_f}$, and image vectors, $\boldsymbol{v_f}$ by feeding them into a fully connected layer, FC, for dimension alignment and subsequent summation.

$$Fu = FC\big(\boldsymbol{v_f}\big) + FC(\boldsymbol{Q_f}) \tag{19}$$

where $Fu \in R^{1\times512}$. As illustrated in Fig. 3, in the classification section, the fused vector undergoes modification in the linear layer before being passed through the sigmoid layer to predict the correct answer as follows:

$$Answer = sigmoid(linear(Fu)) \tag{20}$$

4 Experiment

In this section, we conduct experiments to evaluate the performance of the ACOCAD model on the benchmark GQA and VQA-v2 datasets. First, our benchmark dataset and the evaluation metric are introduced. In the next step, the settings of our experiment such as the number of parallel heads, and the loss function are presented. Then, we perform ablation studies to examine four sub-models of our model and determine the optimal dimensional learnable weights. In the subsequent section, the performance of the model is visualized, and finally, the results of our model are compared with the current state-of-the-art models.

4.1 Datasets

The GQA dataset has a part named object-based features which is compatible with our model, therefore, we assess the ACOCAD model on this dataset including 22,669,678 questions and 113k images, and measuring diverse criteria of the model which will be mentioned in section 4.2. Besides, we use the VQA-v2 benchmark dataset to evaluate our model. The VQA-v2 dataset is split into three parts: a training set (443757 question-answer pairs and 80k images); a validation set (214354 question-answer pairs and 40k images); and a testing set (447793 question-answer pairs and 80k images). All images are obtained from the MSCOCO dataset [46]. There are ten candidate answers per question which are proposed by different annotators. Therefore, the answers may not be unique and there might be different answers for a question. Additionally, the VQA-v2 dataset consists of three types of questions: Yes/No, Number, and Other Questions with a share of %41, %10, and %49, respectively.

4.2 Evaluation Metric

The following metric is considered for evaluations of the VQA-v2 dataset in the experiment:

$$\mathrm{acc}(\mathrm{x_i}) = \min\left(\frac{\text{number of annotators that answer}}{3}, 1\right) \tag{21}$$

where y_i is the predicted answer of the model to the question x_i. Using this method, if more than two annotators propose the exact answer that the model predicts, y_i, then the accuracy of the model for the question x_i is considered equal to one. On the other hand, the GQA has some other criteria like Consistency measuring inference of the model by bringing up a concept in different formats, Validity measuring evaluates whether the predicted answer is in the scope of the valid data or not. In other words, the model merely assesses the answered type. Plausibility is another measure checking if the predicted answer could be reasonable in the real world or not, for instance, we do not have a blue apple. Furthermore, there is another measure of Distribution comparing true answer distribution with that of the predicted answers. As much as this measure is close to one, it shows the model performance is better.

4.3 Experimental Settings

The proposed model consists of some hyper-parameters that should be specified. The dimension of all learnable weights, the output of the multi-head attention, the output of the LSTM network, and all the feedforward layers are all equal to a parameter named (d). As shown in Table 1, the number of parallel heads (h) in the multi-head attention and sequential encoder-decoder (n) are other hyper-parameters that are evaluated in detail in the ablation studies section of Ref. [22, 42]. The optimal values for hyper-parameters h and n are 8 and 6, respectively [45]. Accordingly, we take these hyper-parameters into account in our model and do not analyze them again. Also, the length of the questions is limited to the 14 and 29 first words and the number of the answers in the classification section is set to 3129 and 1843 for the GQA and the VQA-v2 datasets respectively, which are the highest frequency answers [32, 42]. The model is trained in 60 epochs and the batch size increases gradually from 256 to 512. The Cross-Entropy loss function and Adam optimizer are used.

Table 1 Implemented settings of our model, ACOCAD

Basic Hyperparameters	value
Number of parallel heads (h)	8
Sequential encoder-decoder (n)	6
Feedforwarding and LSTM layer	512
Dropout rate	0.2
Loss Function	Cross-Entropy
Batch size	512-256
Epoch	60

The structure of the feedforwarding layer in our model is FC(4d)-ReLU/LeakyReLU-Dropout(0.2)- FC(d). The model is implemented based on the TensorFlow framework available at https://github.com/hshokriAI/ACOCAD. the training and validation sets, besides a subset of the visual genome dataset, are also used to evaluate the accuracy of the testing set. Because the ACOCAD model benefits from Fast R-CNN, a huge reduction in the number of learnable parameters is achieved, as well as more concentration to be given to attention approaches. The ACOCAD model is trained by a system equipped with Nvidia GTX 1080 Ti GPU, 40G of RAM. The model is designed based on parallel processing, therefore, each epoch lasts 10 minutes on average and the entire training process is completed within approximately 6 hours. To the best of my knowledge, the VQA task is immensely sensitive to hyperparameters, and evaluating on diverse VQA datasets brings substantial modifications in both architecture and preprocessing techniques.

4.4 Ablation Studies

In this section, we initially assign each proposed contribution to a distinct sub-model and then evaluate the improvement in accuracy for each sub-model in comparison with the baseline model. Secondly, we assess the hyper-parameter of d, representing the number of feedforwarding and fully connected layers in our model.

4.4.1 Ablation Study in Sub-Models

As mentioned, the ACOCAD model has three contributions. One sub-model for each of the three contributions of this paper is defined by individually adding to the baseline model, which is designed based on the simple combinations of the SA and GA units, and their efficiencies are independently investigated. Our model and these sub-models are as follows:

- baseline + CAGA unit: We merely add the textual context-aware attention mechanism using the CAGA unit in the first sub-model.
- baseline + QLGA + WLGA units: In this sub-model, QLGA and WLGA units are added to the baseline model for the purposes of the question-level and word-level visual attention approaches, respectively.
- baseline + DMF block: The DMF block is replaced with the independent fusion block, which is the last block of the baseline model, to improve the features of images and questions, and generate superior fusion vectors as well.

- ACOCAD: All above contributions are added together to the baseline which form our final model.

Figure 8 compares the accuracies of the three proposed sub-models and our overall model with the baseline one. In the Yes/no question type including more inferential questions than other types, the results demonstrate that all contributions cause better results compared to the baseline model. This is due to the positive impact achieved by incorporating additional information into the basic model (the USE model in this case), mitigating noisy effect through the inclusion of two levels of attention mechanisms, and employing DMF block to identify ambiguity in questions.

Moreover, the ACOCAD model achieves the highest accuracy because adding these contributions leads to an improvement in the inference of the model. In the Number type, all sub-models perform weakly except for the 'baseline + DMF block' sub-model showing its plus point in counting objects. The GQA and VQA-v2 datasets are basically unbalanced, particularly in Number Type, with classes numbers one and two being significantly more frequent than others. In addition, the distribution of training and validation datasets are disparate. Hence, this imbalance contributes to increased bias towards the more frequent classes in the model, potentially justifying the observed decrease in accuracy. In the Other type which includes simpler questions than others, all sub-models outperform the baseline model because QLGA and WLGA units, which detect in the visual attention section the correct regions and diminish noisy effects, perform greatly.

In the both overall GQA and VQA-v2 datasets, the ACOCAD model obtains the best accuracy. It is needed to note that a minor enhancement is regarded as a breakthrough in the VQA problem due to its complexity and difficulty.

4.4.2 Ablation Study in Internal Layer *d*

There is a direct correlation between resource/time consumption and number of the parameters in the same structure. As the number of the parameters exceeds their optimal number, the model is prone to overfitting and the accuracy probably decreases. In this regard, we investigate the impacts of the number of parameters in our model to find its optimal value having both acceptable accuracy and less resource/time consumption. The dimension of all multi-head attention is considered 512 based on the recent Ref. [22, 23] to find the optimal

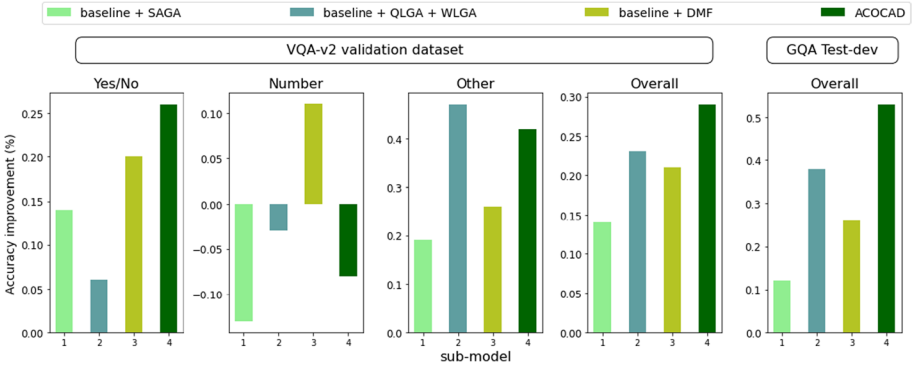

Fig. 8 Accuracy improvement caused by each sub-model compared to the baseline

number for another hyper-parameter d for feedforwarding and FC layers. We evaluate the model for three different values of d : 256, 512, and 1024. According to the success of the ACOCAD model shown in section 4.4, an ablation study on the parameter d is carried out just for this model. For the values of 256, 512, and 1024, the total number of parameters is approximately 30M, 38M, and 54M, respectively. As shown in Fig. 9, overall accuracies of the ACOCAD model for $d = 512$ and $d = 1024$ are higher compared to $d = 256$. Moreover, the model with $d = 1024$ needs more time/resource consumption, and its accuracy is slightly lower, therefore, the optimal value for the parameter d is 512.

4.5 Visualization and Effectiveness of Dependent Multimodal Fusion Block

In this section, we compare the IMF block with the DMF block to assess the effectiveness of the DMF block. Figure 10 illustrates two samples to visualize the efficiency of the DMF block. The salient regions of each image were detected by Fast R-CNN, and the rectangles were drawn around these areas. Moreover, we label only three or four regions in each image and accentuated them with a red bounding box to prevent overcrowding the images and causing confusion for readers. We chose these labeled areas based on their highest relevance to the input question selected from the datasets. In this visualization, we analyze the correlation between labeled areas and words in questions within the IMF and DMF blocks. The color of each cell indicates the similarity between the corresponding words and regions so that the darker color demonstrates higher similarity.

In the first sample, which is 'Is there bat in man hand?', the ambiguous word is 'bat'. Typically, the word 'bat' is primarily defined as an animal; however, in this sample the word 'bat' refers to a stuff for playing. In the model constructed with the IMF block, questions lack a sense of relevance to their corresponding images, resulting in confusion. Although the USE model attempts to slightly mitigate this misunderstanding in the

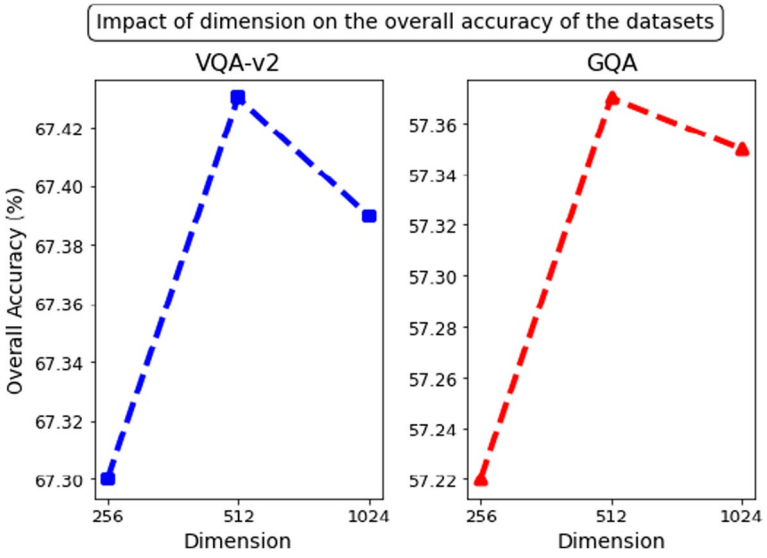

Fig. 9 The overall accuracy of the ACOCAD model for different dimension values d ∈{256, 512, 1024}as the hyper-parameter of the model.

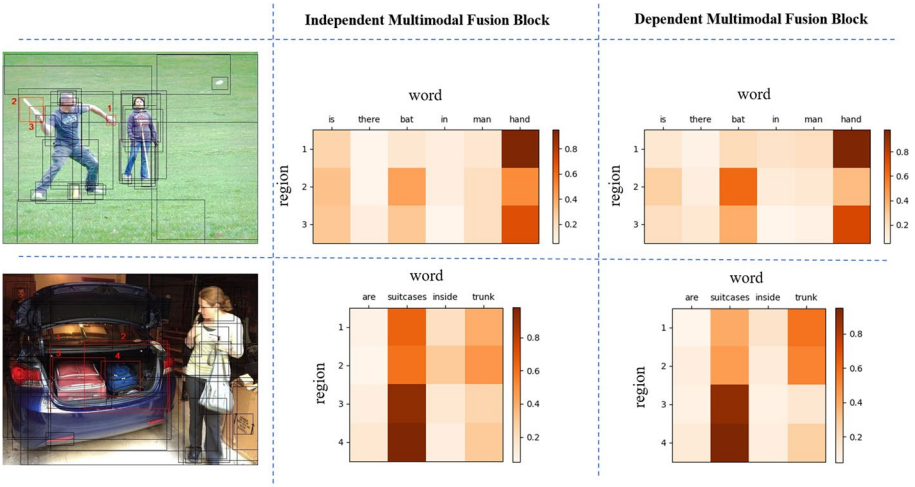

Fig. 10 Comparative Analysis: Independent Multimodal Fusion (IMF) vs. Dependent Multimodal Fusion (DMF) Blocks in Performance Visualization

ACOCAD model, its effectiveness is still insufficient. Hence, the DMF block is used to assist the model in reducing this misunderstanding by focusing on the image regions. In the DMF block, as depicted, the correlation between the region number 2 and the word 'bat' is higher compared to the IMF block. This is because the DMF block identifies the misunderstanding in the word 'bat' and adjusts its vector. Moreover, the DMF block has an additional positive impact, such as enhancing the correlation between the region number 3 and the word 'bat', reducing the correlation between the region number 2 and the word 'hand', and so on.

Likewise, in the second sample, which is 'Are suitcases inside trunk?', the problematic word is 'trunk' which can refer to a part of a tree or the space in the back of a car. The result indicates superior performance of the DMF block than the IMF one because the DMF block in our model identifies a stronger correspondence between region number 1 and the word 'trunk', and also, between region number 2 and this word. Furthermore, the DMF block reduces the correlation between each of these regions and the word 'suitcase' showcasing another advantage of this block. These two visualized samples highlight the beneficial impact of the DMF block. However, there are shortcomings in this block. While it can carefully detect misunderstandings, it may not perfectly revise word vectors. After analyzing the above samples, we found out the presence of homograph words in the questions as a detrimental factor in the VQA problem. Consequently, we conducted a test to assess the effectiveness of the DMF block in the ACOCAD model on VQA datasets including homograph words in order to credit to our claim. In the first test, we collected a 300-word list of the most common homograph words. Subsequently, we selected questions containing at least one of these homograph words and compared the results between the baseline model including the IMF block and the ACOCAD model including the DMF block on the validation datasets. According to Table 2, the baseline model achieves an accuracy of 67.45%, and our ACOCAD model demonstrates an improvement, reaching an accuracy of 67.92%. This reflects a 0.47% enhancement in all questions of the VQA-v2 validation dataset. While The accuracy for questions including homograph words increases from 64.66% to 65.70%, demonstrating a 1.04% improvement. Similarly, in the GQA

Table 2 Homograph word analysis: A comparative study between IMF block in baseline model vs. DMF block in ACOCAD

Dataset	VQA-v2 validation dataset		GQA Test-dev	
	Questions including homograph words	All Questions	Questions including homograph words	All Questions
Baseline	64.66%	67.45%	53.17%	56.64%
ACOCAD	65.70%	67.92%	54.75%	57.37%

dataset, there is a 0.73% improvement across all questions, while the enhancement in questions involving homograph words is 1.58%.

To sum up, our experiment demonstrates that the improvement in the questions including homograph words is greater than all questions of the VQA-v2 validation and GQA all questions. This is because the ACOCAD model benefits from the DMF block and USE model as an external knowledge which are plus points of the proposed model compared to the baseline model. Although these two approaches serve as solutions for homograph words, the DMF block outperforms the USE model in the VQA task due to involving images. This is because, unlike the USE model, which defines the context of questions to partially determine the intended meaning of words, the DMF block examines images to rectify misunderstood words.

4.6 Effectiveness of the USE Model and Visual Attention Mechanism

The length of questions poses another limiting factor in the VQA task. The examinations conducted on the VQA datasets reveal that as the word length of questions increases the model performance declines. This is because long questions raise likelihood of model error in identifying keywords and introduce greater complexity and noisy information on the model, hence, Table 3 is illustrated to assess the positive influence of the ACOCAD model on questions with varying lengths across three ranges: questions with the length of 1 to 7, 7 to 14 and more than 14 words. This experiment demonstrates the improvement of the ACOCAD model compared to the baseline one across various question lengths. Note that these percentages of this table present the accuracy differences between these two models.

The results indicate a positive trend, where the accuracy increases from 0.38% to 0.64% on the VQA-v2 dataset and from 0.63% to 082% on the GQA dataset as the question length increases from the range of 1-7 words to the range of 7-14. This enhancement trend continues for the GQA dataset so that it reaches 1.15% for the questions including more than 14 words. This is due to the USE model, which offers a comprehensive understanding of the question content. It efficiently assigns weights to words by giving less importance to

Table 3 ACOCAD's better performance in different question lengths compared to the baseline model

Question length	Enhancement on VQA-v2 validation dataset	Enhancement on GQA Test-dev
1 to 7	0.38%	0.63%
7 to 14	0.64%	0.82%
More than 14	0.58%	1.15%

redundant words. Importantly, the dual levels of Visual attention mechanisms, QLGA and WLGA, minimize the effect of over-interaction in long questions.

4.7 Qualitative Evaluation

The attention weights images, α^I and questions, α^Q are given by Eq (15, 16) showing the importance of regions and words. In Fig. 11, we illustrate 8 samples (6 correctly predicted and 2 incorrectly predicted) of the dataset in which visual attention located the regions that are more relevant to the correct answer, and textual attention determines the keywords. Two questions are brought up for each original image, and the output of the visual attention and histogram of the textual attention are illustrated in this figure. The bars show the weights of words, and the bright regions locate important predicted regions. Using these samples, we can figure out the effectiveness of the question and its keywords in identifying the attention regions in the image. In the first two images, the model has nearly performed well, and it was able to detect the target regions and words. In the first question of the third image, the keywords 'plane', 'taking off' and also the important regions are specified, but it is not adequate to predict the correct answer. Thus, the model requires more inference. In the second question of the third image, the model perfectly detects the 'tail' region in the image and predicts its color correctly. Although the model counts the signs truly in the fourth image, it could not generate the correct answer in the second question of this image because it is a complicated question, and the model was unable to recognize the keywords and relevant regions. If we visually analyze the model with the aid of the above samples, we conclude that the performance of the model implemented based on the word-level visual attention is satisfactory in simple questions including only one or two keywords. This is mainly because these models implicitly match keywords with salient regions easily.

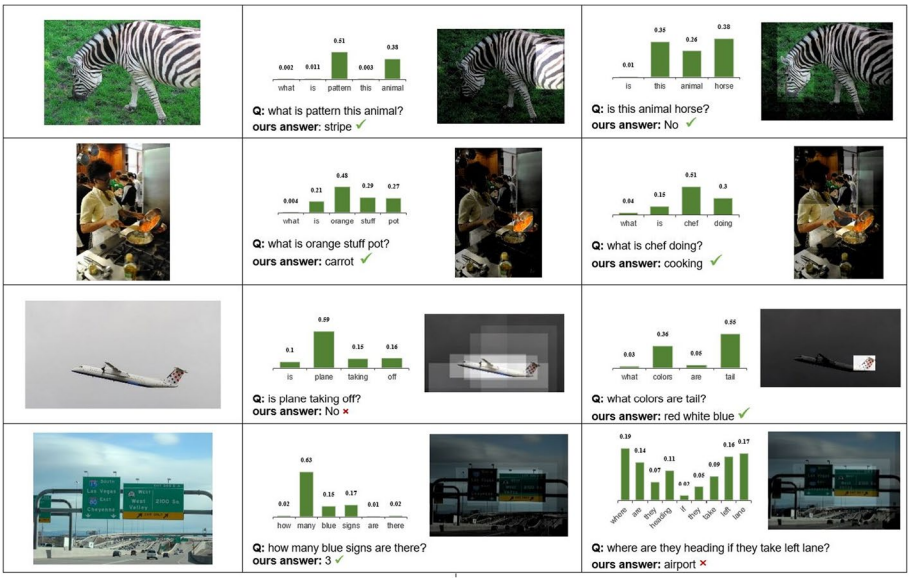

Fig. 11 Visualization of some samples. Original image (left), the output of the visual attention and the histogram of the textual attention weight (middle, right) in the two different questions

However, in the complicated or sometimes lengthy questions, models may be confused about implicit matching keywords with meaningful regions which leads to adding noisy information to the model, or maybe these matchings are insufficient to detect the target regions. Thus, the model needs inference to weigh regions correctly so that the question-level visual attention which refers to the concept of the question can be a solution to this type of question. For instance, it is evident that detecting target regions based on the single words of the question is difficult and baffling in the example of 'Where are they heading if they take the left lane?'. Hence, using the question-level visual attention mechanism besides the word-level visual attention is effective to assign better weights and diminish the effect of the noisy information. Moreover, in the question 'What colors are tail?', the word 'tail' refers to many things like the back of the animal body, the back part of the aircraft, or the bottom part of the shirt. Thus, the DMF block assist the question comprehension to meet challenges by taking a look at the image regions. Thus, the DMF block assist the question comprehension to meet challenges by taking a look at the image regions.

4.8 Comparison with the State-of-the-Art

The ACOCAD model is designed by the optimal parameter found in the ablation studies section. In this section, the ACOCAD model is compared to the current state-of-the-art models implemented based on the object-based segmentation approach, as depicted in Figs. 12 and 13. Models employing this approach, including ACOCAD, have shorter training time as they used a pre-trained Fast R-CNN model, leading reduction in the number of the epochs and parameters significantly.

In the GQA dataset as illustrated in Fig. 12, the ACOCAD model attains the best accuracy in four criteria: Binary, Consistency, Validity, and Overall Accuracy which are defined in the Section 4.2. About Consistency, which nearly refers to the inference ability, the model improves the accuracy by 0.37%. In addition, the results demonstrate that our model generates more valid data than other models because the USE model as an external knowledge has a vital role in this criterion. The overall Accuracy as the most paramount criterion is enhanced by 0.24% and reaches 57.37%. In the VQA-v2 dataset shown in Fig. 13, the ACOCAD model achieves the highest accuracy of 87.43% in the Yes/No question type, and its accuracy is 0.33% and 0.47% higher than MEDAN and SCAVQAN models,

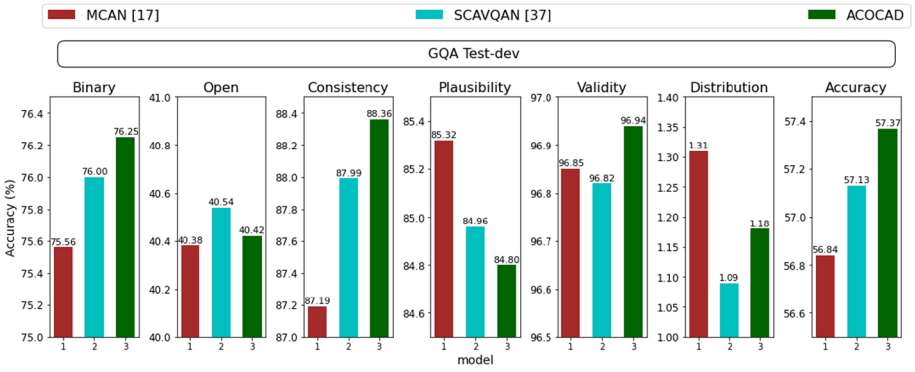

Fig. 12 Comparing the ACOCAD model with other object-based state-of-the-art models evaluated on the GQA dataset based on seven criteria

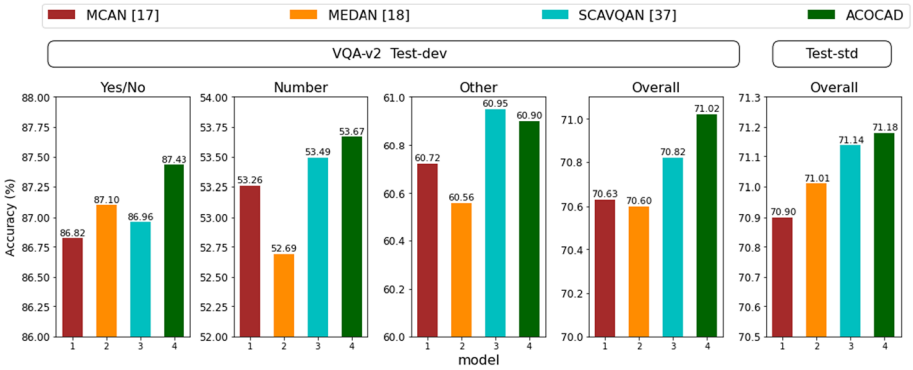

Fig. 13 Result of the ACOCAD model compared to other object-based state-of-the-art models on the VQA-v2 dataset

respectively. As mentioned, the Yes/no question type is more inferential than the Other question type, and we improve this ability using our contributions. Unlike the VQA-v2 validation dataset, the model performance was successful in the Number type of the test-dev dataset. Besides, the overall test-dev accuracy of the model is 0.2% and 0.41% higher than the SCAVQAN and MCAN models which are in the subsequentranks. This amount of enhancement in the VQA problem is considered a significant improvement. Finally, the overall test-std accuracy of our model 71.18% obtains the first rank among all state-of-the-art models.

Unlike Figs. 12 and 13 which compare only object-based and dense interaction models, Table 4 compares the ACOCAD model and its sub-models with current state-of-the-art models. The MUAN model [47] proposes a general unified attention network to realize relationships among elements of questions and images. The MESAN [48] model concentrates on some certain sections of images while filtering out others that may be irrelevant.

Table 4 Comparing the accuracies of our sub-models, the ACOCAD model, and state-of-the-art models on GQA and VQA-v2 datasets

Model	VQA-v2 Test-dev				VQA-v2 Test-std	GQA Test-dev
	Yes/No	Number	Other	Accuracy	Accuracy	Accuracy
BUTD [32]	81.82	44.21	56.05	65.32	65.67	49.74
ICIV [49]	86.63	48.23	59.98	70.46	70.67	-
MCAN [22]	86.82	53.26	60.72	70.63	70.90	56.84
MUAN [47]	86.77	**54.40**	60.89	70.82	71.10	49.74
MEDAN [23]	87.10	52.69	60.56	70.60	71.01	-
MESAN [48]	87.05	53.21	60.72	70.71	71.08	56.37
SCAVQAN [42]	86.96	53.49	**60.95**	70.82	71.14	57.13
Baseline	86.78	53.15	60.48	70.48	70.73	56.25
Baseline + SAGA	86.97	53.20	60.56	70.60	70.85	56.74
Baseline + QLGA + WLGA	86.94	53.62	**60.95**	70.81	70.98	56.86
Baseline + DMF	87.22	53.70	60.64	70.78	71.03	57.10
ACOCAD	**87.43**	53.67	60.90	**71.02**	**71.18**	**57.37**

Himanshu et al. [49] suggest a model named "Image Captioning Improved Visual Question Answering". This model generates a caption from images and employs it in the attention mechanism to generate more semantic visual features.

The ACOCAD model improves the accuracy by 0.65% compared to the baseline model and achieves the top rank by reaching 87.43% in the Yes/No question type of the VQA Test-dev dataset. Additionally, the baseline model + DMF block attains the second rank among the state-of-the-art models. This improvement is achieved because the Yes/No question type is both more inferential and more frequent compared to other types, and the proposed innovations such as adding external knowledge, introducing two levels of visual attention, and implementing the DMF block, contribute to generating more constructive features.

In the Other question type, the ACOCAD model achieves 60.90%, reflecting a 0.42% improvement over the baseline model. Furthermore, both the SCAVQAN and Baseline + QLGA + WLGA models achieve the highest accuracy of 60.95%. Owing to that these two models benefit of the unique visual attention mechanisms. The SCAVQAN defines a threshold for image weights and eliminates ones that are smaller than the threshold. Similarly, the QLGA + WLGA model performs the same manner but with a different approach such that it assigns less weight to irrelative regions by two levels of attention.

In terms of overall accuracy for the test-dev and test-std measures, the ACOCAD model surpasses all state-of-the-art models, achieving percentages of 71.02% and 71.18%, respectively. Additionally, it improves upon the baseline model by 0.5% and 0.55% in the test-dev and test-std VQA-v2 datasets. While the SCAVQAN model achieves the highest accuracy among all state-of-the-art models on the GQA dataset at 57.13%, the ACOCAD model surpasses it with an accuracy of 57.37%. Furthermore, the ACOCAD model demonstrates significant improvement over the baseline model, achieving an accuracy increase of 1.12% on the GQA dataset.

It is worth mentioning that the VQA task is one of the most challenging tasks in the world. Thus, a slight growth in each measure is considered a major breakthrough in the VQA task.

5 Conclusion

In this paper, our proposed ACOCAD model for the VQA task comprises four key sections. The initial section involves image and question representations, with the innovative use of the USE model as external knowledge for the question representation section which has been never used before to the best of our knowledge. The second section introduces the textual context-aware attention mechanism, utilizing the CAGA unit to enhance the question comprehension. The third section employs the question-level and word-level visual attention mechanism, implemented by the QLGA and WLGA units, to focus on regions with diverse objectives. The final section incorporates the DMF block to enhance the interaction between word and region vectors by considering the impact of regions on word vectors.

To evaluate the performance of the model, two well-known datasets namely GQA and VQA-v2 are used. We conduct an ablation study on each proposed sub-model to investigate the impact of each introduced mechanism. Another ablation study is carried out on one hyper-parameter, aiming to identify its optimal value. Then, we assess the potential of the DMF block in smoothing the limitations observed in the previous methods, particularly in

handling questions with homograph words. Furthermore, to tackle the complexity associated with the length of question words, we examine the effectiveness of the USE model and the Visual Attention Mechanism. To provide a comprehensive visualization, we perform a qualitative evaluation that offers a visual representation of the effectiveness of the ACOCAD model through some samples.

The experiment results on the GQA dataset indicate that the ACOCAD model excels in four out of seven criteria, with the highest overall accuracy among all state-of-the-art models. The ACOCAD model excels in answering Yes/No questions on the VQA Test-dev dataset, securing the top rank. Additionally, one of its sub-models attains the highest accuracy in the Other question type. Furthermore, the overall accuracy of the ACOCAD model is higher than all state-of-the-art models on both the test-dev and test-std measures.

6 Limitations and Future Works

In this section, we propose limitations of the ACOCAD model and suggest corresponding solutions for future works.

Classification section Although the second-best answer in the prediction layer may sometimes outperform the final answer, it is not chosen as the final answer just due to a minor probability difference. Parametric softmax classifier has several limitations such as lack of simplicity and explainability and paying less attention to modeling the latent data structure. Therefore, improving the prediction layer and its settings is very enlightening for future research. According to the study [50], we can utilize the Deep Nearest Centroid (DNC) algorithm to classify final fusion vectors, formulated as follows:

$$Fu_{DNC} = linear(FC(\boldsymbol{v}_f) + FC(\boldsymbol{Q}_f)) \tag{22}$$

Where Fu_{DNC} is the final represented feature in the VQA problem which is composed of fusing final image vector, $\boldsymbol{v}_f$, and question vector, $\boldsymbol{Q}_f$. Accordingly, the DNC algorithm is applied to Fu_{DNC}. The DNC algorithm defines a centroid for each class in training datasets and classifies based on the distance of test datasets from these centroids in the feature space. Integrating the DNC algorithm with the softmax layer can improve simplicity, transparency, discovering underlying data structure, and representation learning.

Fusion block limitation While the ACOCAD model effectively identifies keywords and regions, it faces challenges in fusing these regions with questions and accurately detecting answers. This limitation primarily arises from the combination of two distinct content types – vision and text – leading to potential discrepancies. To address this, captions of the regions, inspired by the approach in study [49] but with innovative structures, can be utilized to enhance the efficiency of the fusion block

Monocular Depth Estimation (MDF) MDF is a technique for estimating depth from 2D images [51]. VQA datasets include questions related to object locations and their positions relative to each other, showcasing the model's ability to recognize depth in images. For instance, the VQA-v2 dataset comprises questions starting with the word 'Where' indicating a focus on identifying depth in images. However, it exhibits weaknesses in accurately determining the exact depth of objects, a limitation that can be addressed through adversarial training in self-supervised monocular depth estimation across various aspects.

Vulnerability to attack The VQA task faces challenges from both users and physical agents, particularly in the form of complex, lengthy questions designed to deceive the system. Additionally, there is a threat from image attacks using optimal adversarial patches [52]. Furthermore, the vulnerability of the common fusion block in advanced deep neural networks is demonstrated in study [53]. To address these challenges, advanced self-supervised approaches for extracting more meaningful features can be employed, coupled with reinforcement learning to enhance model performance through user feedback.

Authors Contribution All authors, listed on the title page, had active roles in the preparation of the manuscript, however, the first draft of the manuscript was written by 'Hesam Shokri Asri'. All authors commented on previous versions of the manuscript and approved the final one. Furthermore, they attest to the validity of data and results.

Data availability The validity and source code of this study are available in https://github.com/hshokriAI/ACOCAD

Declarations

All authors agree on submitting the paper to the Multimedia Tools and Applications Journal. No plagiarism has occurred in this manuscript and all references have been reviewed and cited. The authors did not receive support from any organization for the submitted work. In addition, this study, and the datasets used, did not involve human or animal participation, and therefore it did not require consent.

Competing Interests Authors have no relevant financial or non-financial interests to disclose, and no funds or grants were received for the preparation of this study.

References

1. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inform Proc Syst (NIPS). https://doi.org/10.1145/3065386
2. Pham DL, Xu C, Prince JL (2020) A survey of current methods in medical image segmentation. Annual Rev Biomed Eng 2(3), 315-337. https://doi.org/10.1146/annurev.bioeng.2.1.315
3. Selim, Md, Jie Zhang, Faraneh Fathi, Michael A. Brooks, Ge Wang, Guoqiang Yu, and Jin Chen (2023). Latent Diffusion Model for Medical Image Standardization and Enhancement. *arXiv preprint* arXiv:2310.05237. *https://arxiv.org/2310.05237*
4. Bolhassani M, Oksuz I (2021) Semi-Supervised Segmentation of Multi-vendor and Multi-center Cardiac MRI. In 2021 29th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE. https://ieeexplore.ieee.org/abstract/9477818
5. Mohtasebi M, Huang C, Zhao M, Mazdeyasna S, Liu X, Haratbar SR, ..., & Yu G. (2023) A Wearable Fluorescence Imaging Device for Intraoperative Identification of Human Brain Tumors. IEEE J Trans Eng Health Med. https://ieeexplore.ieee.org/abstract/document/10339301
6. Irwin D, Mazdeyasna S, Huang C, Mohtasebi M, Lui X, Chen L, Yu G (2022) Near-infrared Speckle Contrast Diffuse Correlation Tomography for Noncontact Imaging of Tissue Blood Flow Distribution. CRC Press
7. Kowsari K, JafariMeimandi K, Heidarysafa M, Mendu S, Barnes L, Brown D (2019) Text classification algorithms: A survey. Information 10(4):150
8. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. Advances in neural information processing systems, 27. https://doi.org/10.1145/3439726
9. Nawar N, Omar E-G, Loknath SS, Giridhar RB (2022) Social media for exploring adverse drug events associated with multiple sclerosis. https://hawaii.edu/10125/79851
10. Mao J, Huang J, Toshev A, Camburu O, Yuille AL, Murphy K (2016) Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition. 11-20

11. Mao J, Xu W, Yang Y, Wang J, Huang Z, Yuille A (2014) Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632. https://doi.org/10.48550/1412.6632
12. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D (2015) Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision. 2425-2433
13. Gupta AK (2017) Survey of visual question answering: Datasets and techniques. arXiv preprint arXiv:1705.03865. https://doi.org/10.48550/1705.03865
14. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks, In: Proceedings of the European Conference on Computer Vision (ECCV). 818–83. https://doi.org/10.1007/978-3-319-10590-1_53
15. Sethy A, Ramabhadran B (2008) Bag-of-word normalized n grammodels. In Ninth Annual Conference of the International Speech Communication Association
16. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural computation 9(8):1735–1780. https://doi.org/10.1162/6795963
17. Wu Q, Teney D, Wang P, Shen C, Dick A, Van Den Hengel, A (2017) Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding, 163, 21-40. https://doi.org/10.1016/j.cviu.2017.05.001
18. Chen K, Wang J, Chen LC, Gao H, Xu W, Nevatia R (2015) Abc-cnn: An attention based convolutional neural network for visual question answering. arXiv preprint. https://doi.org/10.48550/1511.05960
19. Shih KJ, Singh S, Hoiem D (2016) Where to look: Focus regions for visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4613-4621)
20. Yu Z, Yu J, Fan J, Tao D (2017) Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In Proceedings of the IEEE international conference on computer vision (pp. 1821-1830)
21. Lu J, Yang J, Batra D, Parikh D (2016) Hierarchical question-image co-attention for visual question answering. Advances in neural information processing systems, 29
22. Yu Z, Yu J, Cui Y, Tao D, Tian Q (2019) Deep modular co-attention networks for visual question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 6281-6290)
23. Chen C, Han D, Wang J (2020) Multimodal encoder-decoder attention networks for visual question answering. IEEE Access, 8, 35662-35671. https://doi.org/10.1109/ACCESS.2020.2975093
24. Rong X (2014) word2vec parameter learning explained. arXiv preprint. https://doi.org/10.48550/1411.2738
25. Cer D, Yang Y, Kong SY, Hua N, Limtiaco N, John RS, Kurzweil R (2018) Universal sentence encoder. arXiv preprint. https://doi.org/10.48550/1803.11175
26. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6904-6913) (2017)
27. Hudson DA, Manning CD (2019) Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 6700-6709
28. Zhou B, Tian Y, Sukhbaatar S, Szlam A, Fergus R (2015) Simple baseline for visual question answering. arXiv preprint. https://doi.org/10.48550/1512.02167
29. Ren M, Kiros R, Zemel R (2015) Exploring models and data for image question answering. In: Advances in Neural Information Processing Systems (NIPS)
30. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NIPS)
31. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Fei-Fei L (2017) Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int J Comput Vision 123(1):32–73. https://doi.org/10.1007/S11263-016-0981-7
32. Anderson P, He X, Buehler C (2018) Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In: CVPR
33. Yang Z, He X, Gao J, Deng L, Smola A (2016) Stacked attention networks for image question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 21-29)
34. Fukui A, Park DH, Yang D, Rohrbach A, Darrell T, Rohrbach M (2016) Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint. 1048550/1606.01847
35. Ben-Younes H, Cadene R, Cord M, Thome N (2017) Mutan: Multimodal tucker fusion for visual question answering. In: Proceedings of the IEEE international conference on computer vision. 2612-2620
36. Gao Y, Beijbom O, Zhang N, Darrell T (2016) Compact bilinear pooling. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 317-326)

37. Kim JH, On KW, Lim W, Kim J, Ha, JW, Zhang BT (2016) Hadamard product for low-rank bilinear pooling. arXiv preprint. https://doi.org/10.48550/1610.04325
38. Teney D, Anderson P, He X, Van Den Hengel, A (2018) Tips and tricks for visual question answering: Learnings from the 2017 challenge. In: Proceedings of the IEEE conference on computer vision and pattern recognition.4223-4232
39. Yu Z, Yu J, Xiang C, Fan J, Tao D (2018) Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE transactions on neural networks and learning systems, 29(12): 5947-5959. https://doi.org/10.1109/TNNLS.2018.2817340
40. Yang C Jiang M, Jiang B, Zhou W, Li K (2019) Co-attention network with question type for visual question answering. IEEE Access, 7, 40771-40781. https://doi.org/10.1109/ACCESS.2019.2908035
41. Kim JH, Jun J, Zhang BT (2018) Bilinear attentionnetworks. Advances in neural information processing systems, 31
42. Guo Z, Han D (2022) Sparse co-attention visual question answering networks based on thresholds. Applied Intelligence, 1-15. https://doi.org/10.1007/s10489-022-04355-w
43. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).1532-1543
44. Cer D, Yang Y, Kong SY, Hua N, Limtiaco N, John RS, Kurzweil R (2018): Universal sentence encoder for English. In: Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations. 169-174
45. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Proc. NIPS, Long Beach, CA, USA. 5998–6008
46. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Zitnick CL (2014) Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755) Springer, Cham. https://doi.org/10.1007/978-3-319-10602-1_48
47. Yu Z, Cui Y, Yu J, Tao D, Tian Q (2019) Multimodal unified attention networks for vision-and-language interactions. arXiv preprint arXiv:1908.04107. https://arxiv.org/1908.04107
48. Guo Z, Han D (2020) Multi-modal explicit sparse attention networks for visual question answering. Sensors 20(23):6758
49. Sharma H, Jalal AS (2022) Image captioning improved visual question answering. Multimed Tools Appli 81(24):34775–34796. https://doi.org/10.1007/s11042-021-11276-2
50. Wang W, Han C, Zhou T, Liu D (2022) Visual recognition with deep nearest centroids. *arXiv preprint arXiv:2209.07383*. https://arxiv.org/2209.07383
51. Cheng Z, Liang J, Tao G, Liu D, Zhang X (2023) Adversarial training of self-supervised monocular depth estimation against physical-world attacks. arXiv preprint arXiv:2301.13487. https://arxiv.org/2301.13487
52. Cheng Z, Liang J, Choi H, Tao G, Cao Z, Liu D, Zhang, X (2022) Physical attack on monocular depth estimation with optimal adversarial patches. In European Conference on Computer Vision (pp. 514-532). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19839-7_30
53. Cheng Z, Choi H, Liang J, Feng S, Tao G, Liu D., ... & Zhang X (2023) Fusion is Not Enough: Single-Modal Attacks to Compromise Fusion Models in Autonomous Driving. arXiv preprint-arXiv:2304.14614. https://arxiv.org/2304.14614

Authors and Affiliations

Hesam Shokri Asri[1] · **Reza Safabakhsh**[1]

✉ Hesam Shokri Asri
h.shokri@aut.ac.ir

Reza Safabakhsh
safa@aut.ac.ir

[1] Computer Engineering Department, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran