Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

DMRFNet: Deep Multimodal Reasoning and Fusion for Visual Question Answering and explanation generation

Weifeng Zhang [a,*], Jing Yu [b], Wenhong Zhao [c], Chuan Ran [d]

[a] Jiaxing University, Zhejiang, China
[b] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[c] Nanhu College, Jiaxing University, Zhejiang, China
[d] IBM Corporation, NC, USA

ARTICLE INFO

Keywords:
Multimodal reasoning and fusion
Visual Question Answering
Explainable artificial intelligence

ABSTRACT

Visual Question Answering (VQA), which aims to answer questions in natural language according to the content of image, has attracted extensive attention from artificial intelligence community. Multimodal reasoning and fusion is a central component in recent VQA models. However, most existing VQA models are still insufficient to reason and fuse clues from multiple modalities. Furthermore, they are lack of interpretability since they disregard the explanations. We argue that reasoning and fusing multiple relations implied in multimodalities contributes to more accurate answers and explanations. In this paper, we design an effective multimodal reasoning and fusion model to achieve fine-grained multimodal reasoning and fusion. Specifically, we propose Multi-Graph Reasoning and Fusion (MGRF) layer, which adopts pre-trained semantic relation embeddings, to reason complex spatial and semantic relations between visual objects and fuse these two kinds of relations adaptively. The MGRF layers can be further stacked in depth to form Deep Multimodal Reasoning and Fusion Network (DMRFNet) to sufficiently reason and fuse multimodal relations. Furthermore, an explanation generation module is designed to justify the predicted answer. This justification reveals the motive of the model's decision and enhances the model's interpretability. Quantitative and qualitative experimental results on VQA 2.0, and VQA-E datasets show DMRFNet's effectiveness.

1. Introduction

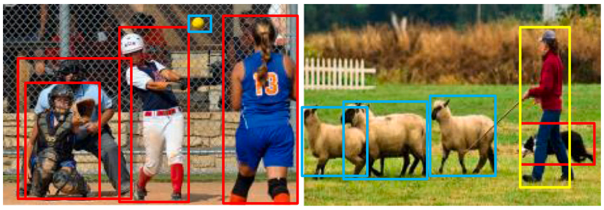
Visual Question Answering (VQA), which aims to jointly analyze multimodal content from images and natural language, is an attractive research direction. The VQA agent is expected to answer a question in natural language regarding an image. Thus, it is a more challenging task than traditional vision task such as classification [1,2] and detection [3], since it demands the agent to simultaneously understand the content of vision and language and fuse the information from both modalities to infer the correct answer.

Much effort has been made in literature and achieves great success in VQA task [4–9]. Most of these existing models devote themselves to fusing multimodal features extracted from image and question [10]. Element-wise multiplication and concatenation are most straight forward methods to fuse image and question features. Later on, more complex fusion models such as MLB [11], MCB [12], MFH [4], BLOCK [13], have been proposed, among which bilinear models are proven effective fusion approaches. However, these shallow multimodal fusion models are lack of fine-grained multimodal interactions. Then image region features obtained by pre-trained object detectors [3] and attention mechanism are widely adopted [14–16]. These shallow attention networks show that attention mechanism has the ability to highlight important and relevant visual features and textual features. Furthermore, researchers proposed deep co-attention models such as BAN [17], DCN [5], DFAF [18], MCAN [6], trying to model dense interactions between any image region and any question word. These studies updated the state-of-the-art for common benchmark datasets at the time of each publication. However these deep co-attention models are still insufficient to model complex relational features required for VQA task. Recent study shows that visual relation reasoning is crucial for enhance interactions between modalities [10] and it has also been widely studied and adopted by recent VQA models [19–23]. The relationships can be formally defined as triples $\langle subject, predicate, object\rangle$, i.e. $\langle woman, riding, horse\rangle$ or $\langle boy, kicking, ball\rangle$. Such visual relationships are working in conjunction with deep neural networks to generate relation-aware visual representation and significantly improve the accuracy of answer prediction.

Despite that promising performance has been reported, most existing VQA models are *black boxes* and there is still huge gap for humans to

[*] Corresponding author.
E-mail address: zhangweifeng@zjxu.edu.cn (W. Zhang).

https://doi.org/10.1016/j.inffus.2021.02.006
Received 25 August 2020; Received in revised form 25 December 2020; Accepted 7 February 2021
Available online 12 February 2021



Question: What is the woman in the middle doing?
Answer: hitting
Explanation: A woman wearing white hat is standing in the middle and hitting the baseball.

(a)

Question: What is the dog’s job?
Answer: herding
Explanation: A dog is helping people herd sheep on the grass.

(b)

Fig. 1. VQA examples show that both positional relations and semantic relations are vital for VQA models to predict accurate answer.

truly understand the model decisions without any explanation for them. To help people catch the motive of model’s decision, a popular way is to visualize attention maps to indicate the focused image regions. These regions attended by models are expected to match humans’ visual focus when people answer the same question. However, experimental researches show that current attention models in VQA do not seem to be looking at the same regions as humans [24]. For example, many cases show that the model attends to right regions but predicts wrong answers, or even predicts right answers when focuses on the regions irrelevant to the question. To solve this problem, several works [25,26] attempted to train attention model supervised by human-like attention annotations. Unfortunately, little performance gain is achieved and such human-like attention is expensive to obtain.

In this paper, we aim to address the above limitations of existing VQA models. Firstly, We observe that both positional relations and semantic relations are vital for VQA models to predict correct answer. For example, when the VQA model is asked *“What is the woman in the middle doing?”* about the image in Fig. 1(a), it not only needs to detect the *“woman in the middle”* by analyzing the positional relation between the players, but also reasons the semantic relations between the woman and the ball. Even for the image in Fig. 1(b), semantic relations between *“dog”*, *“people”*, and *“sheep”* are crucial for VQA model to predict correct answer, while the spatial relations play supplementary role. Secondly, humans naturally provide textual explanations for their decisions to communicate with others. Textual explanations for answers predicted by VQA models can be easier understood by users. And we believe that those relations are also the basis for generating rational explanations. In this work, we propose multimodal reasoning and fusion network to endow VQA model with relation reasoning ability and provide sufficient clues for predicting correct answer and generating textual explanation. Therefore, here we design a novel Multi-Graph Reasoning and Fusion (MGRF) layer which reasons positional and semantic relations between image regions by building two parallel graphs including positional graph and semantic graph which adopts pre-trained relation embeddings. And these two graphs can be adaptively fused guided by the question. The MGRF layers can be further stacked in depth to form Deep Multimodal Reasoning and Fusion Network (DMRFNet) to sufficiently reason and fuse multimodal relations for predicting correct answers. Furthermore, an explanation generation module is designed to generate textual explanation for the predicted answer. This textual explanation enhances the model’s interpretability, making the model decisions more understandable for humans. Quantitative and qualitative experimental results on two popular benchmarks including VQA 2.0 [27], and VQA-E [28] demonstrate that our DMRFNet is an effective multimodal reasoning and fusion approach for answering questions and generating explanations.

2. Related work

2.1. Visual question answering

VQA models need to predict answer for a question in natural language based on its understanding of a relevant image. Thus VQA models need the capability of reasoning and fusing visual and textual information from the image-question pair to infer the answer. The most typical methods for feature fusion are element-wise summation/multiplication or direct concatenation. Besides straightforward solutions, several works apply bilinear pooling [4,11,12,29] or more complex fusion methods [13,30]. Latter on, attention mechanism is adopted by VQA models [5,15,16,31] to highlight relevant visual and textual information. Furthermore, Anderson et al. first introduced object detector [3] to achieve object-level attention, rather than spatial grid. Most recently, impressive results on popular VQA benchmarks are obtained by deep co-attention models including BAN [17], DCN [5] and DFAF [18]. These models establish the complete interaction between each question word and each image region, resulting in better image-question representations.

Although promising performance has been reported, most existing VQA models are lack of interpretability. Recent works have turned to design explainable VQA models. The most straightforward way is to visualize attention maps [24–26]. Another approach to enhance the interpretability of VQA models is to offer text-based generation of post-hoc justifications [28,32]. In [28], a new dataset called VQA-E was introduced, in which textual explanation is also provided for each answer. Inspired by these works, we design and train our model to simultaneously predict answers and generate justifications. We find that training with the textual explanations not only yields better textual justification models, but also higher accuracy of answer prediction.

2.2. Visual relation reasoning

Visual relation reasoning aims to represent and infer the relations among objects in an image, thus aggregating information from objects and their relations to enrich the image representation. It is a top priority for AI to achieve human intelligence [33]. How to model these relations is one of the major differences of the existing approaches. Early works [34] study the shallow positional relations based on spatial information (e.g. *left*, *above*) to improve visual segmentation. Later on, in [35], interactions (e.g. *wear*, *carry*) between paired objects are exploited, where visual relation reasoning is then formulated as a classification task. Afterwards, relationships are extended to richer definition [36], including positional, comparative, composition, action, *etc*. The most recent works propose to model visual relations by scene graphs based on prior human knowledge [37] and effectively improve the performance of image captioning [38].

Endowing VQA models with relational reasoning ability has also become a hot focus. [19] and [10] used simple fully connected networks or convolutional networks to implicitly reason relations between every pair of image regions, while [23] and [39] used Graph Neural Networks (GNN) to model relations between image regions. Wang et al. proposed VQA-Machine [21] which explicitly reasons the relation between objects. And visual facts as triplets such as *<man, play, football>* are needed to train it. All these models demonstrate that relation-aware information can significantly boost the performance of VQA models. However, these models fail to simultaneously model positional and semantic relations and they are also lack of interpretability. Our proposed MGRF layer not only reason multiple relations, but also adaptively fuse relations guided by the question, revealing an explainable information selection mode.

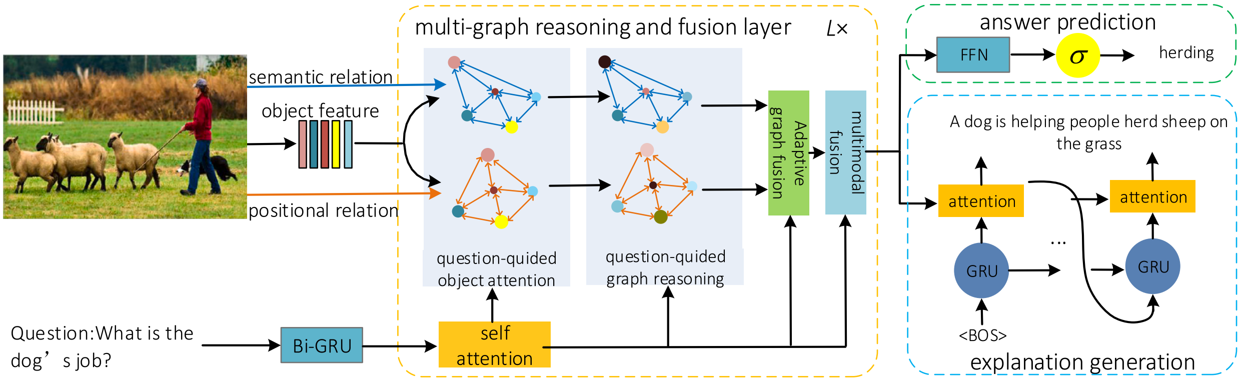


Fig. 2. The framework of our model, which is mainly composed of three components: (1) Deep Multimodal Reasoning and Fusion Network (in the yellow dashed box), stacked by L Multi-Graph Reasoning and Fusion layers which explicitly reason and fuse positional and semantic relations, plays a core role in this model. (2) Answer prediction module (in the green dashed box) composed of a feed forward network. (3) Explanation generation module (in the blue dashed box), which consists of recurrent neural networks and attention module, can generate justifications for predicted answers. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.3. Multimodal fusion

Multimodal fusion [40–42], which selects and fuses plentiful information from multiple modalities, is an essential module of VQA models. Most multimodal fusion methods in VQA models [4,11,12,29] fall into the feature-level fusion approach. Recent research shows that bilinear models are powerful in multimodal fusion, since they provide rich interaction between all elements of both vectors. However, they encounter inefficiency in VQA task due to its high dimensionality. [12] proposed MCB to speed up bilinear fusion by randomly projecting image and question features into a common space and use Fast Fourier Transformation to convolve the features. MFB [29], an enhanced version of MCB is proposed by Yu et al. and is cascaded to generate MFH [4] to further enhance the representation capacity of fused features. Kim et al. went step further and proposed MLB [11] to reduce parameters by rewriting the weight matrix into the multiplication of two small matrices. Recently, [13] proposed a new multimodal fusion named BLOCK based on the block-superdiagonal tensor decomposition. It optimizes the tradeoff between the expressiveness and complexity of the fusion model. However, these multimodal fusion mechanisms are shallow models, which are insufficient to fuse complex visual and textual information in VQA task. Most recently, some new approaches such as large-scale pretraining [43] and automatic neural architecture search frameworks [44] are introduced to learn multimodal representations and address vision-language tasks. In this paper, we proposed a novel Multi-Graph Reasoning and Fusion (MGRF) layer which explores positional and semantic relations and adaptively fuses them. Our MGRF layers can be cascaded to build Deep Multimodal Reasoning and Fusion Network (DMRFNet) to achieve sufficient reasoning and fusion.

3. Methodology

The typical VQA task can be described as follows: given an image I and a question q about the image, the task is to predict the answer according to the image. In this paper, we further generate textual explanations for the predicted answer to enhance the interpretability of our VQA model. As shown in Fig. 2, our VQA model is mainly composed of three components: (1) Deep Multimodal Reasoning and Fusion Network (in the yellow dashed box), that is stacked by L Multi-Graph Reasoning and Fusion layers which explicitly reason and fuse positional and semantic relations, plays a core role in this model. (2) Answer prediction module (in the green dashed box) composed of a Feed Forward Network (FFN) predicts the probability of each candidate answer based on the fused multimodal representation. (3) Explanation generation module (in the blue dashed box), which consists of recurrent neural networks and attention module, can generate textual justifications for the predicted answers.

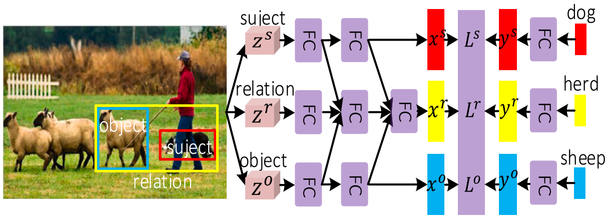


Fig. 3. The framework of semantic relation encoder.

3.1. Question representation

To make all the questions have the same length, we first trim each input question to a maximum of S words by simply discarding the extra words of the question longer than S words. Then we use word embedding to map the tokenized word into a 300 dimensional vector Thus the question is converted into a sequence of word embeddings $\{e_1, e_2, \dots, e_S\}$, which are then passed through a bi-directional GRU (Bi-GRU) to output the word representation as follows:

$$\overrightarrow{q_n} = Bi - GRU(\overrightarrow{q_{n-1}}, e_n) \tag{1}$$

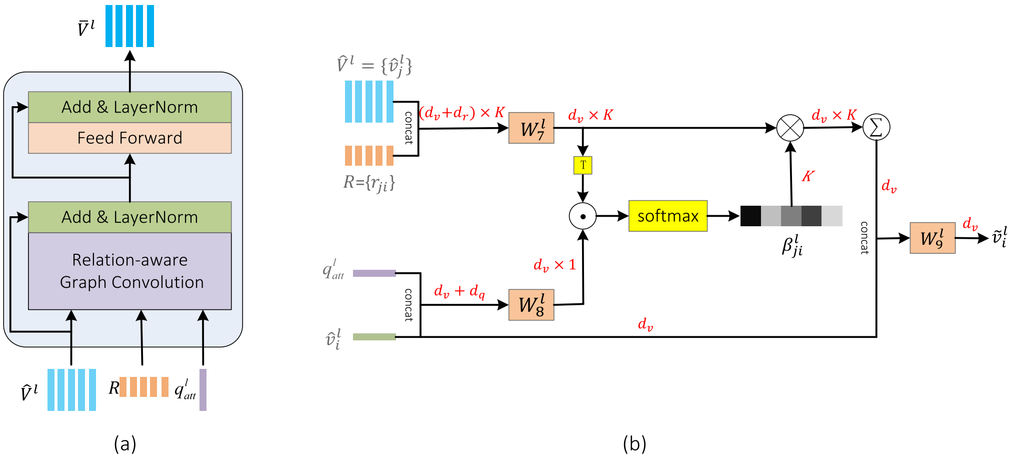
$$\overleftarrow{q_n} = Bi - GRU(\overleftarrow{q_{n+1}}, e_n) \tag{2}$$

Then each question can be represented as a matrix $Q = \{q_1, \dots, q_S\} \in \mathbb{R}^{d_q \times S}$, where $q_n = [\overrightarrow{q_n}, \overleftarrow{q_n}]$, and $[\cdot, \cdot]$ denotes concatenation.

3.2. Image representation

To fully understand the input image, we not only detect salient visual objects but also the relations between them. Following [14], pretrained Faster R-CNN [3] is implemented to extract visual features of K (typically $K = 36$) salient objects in the input image. Thus we obtain representation $V = \{v_1, v_2, \dots, v_K\}$ for each image.[1] Furthermore, the geometric features of the detected objects are also recoded, denoted as $G = \{g_1, g_2, \dots, g_K\}$, where $g_i = [x_i, y_i, w_i, h_i]$. ($x_i$, y_i), w_i, and h_i are the coordinates, width, and height of the detected region i respectively.

[1] To be specific, we obtain the visual feature $v_i \in \mathbb{R}^{2048}$ of object i ($i = 1, \dots, K$) from the $Res4b22$ feature map, which is the last convolutional layer in stage 4 of ResNet-101, through $RoI\ Pooling$ [3], since $Res4b22$ encodes rich category information of the detected bounding box and achieves best detection results in [45].



(a) (b)

Fig. 4. (1) The structure of our proposed Question-Guided Graph Reasoning module. (2) The internal structure of relation-aware graph convolution which is the core of Question-Guided Graph Reasoning. Red number on each line with arrow denotes the shape of the tensor. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Then positional relation between object i and object j can be simply embedded as follows:

$$r^p_{i,j} = \sigma(\mathbf{W}[\frac{x_j - x_i}{w_i}, \frac{y_j - y_i}{h_i}, \frac{w_j}{w_i}, \frac{h_j}{h_i}, \frac{w_j h_j}{w_i h_i}]) \tag{3}$$

We also try to model the semantic relations between visual objects based on semantic relation encoder which is proposed in [46], whose detailed structure is illustrated in Fig. 3. The encoder consists of two parts: (1) visual module which turns the features of subject z^s, relation z^r and object z^o into the embeddings of subject x^s, relation x^r and object x^o. (2) semantic module who takes the label of subject, relation and object encoded by a shared GRU as input, and outputs embeddings y^s, y^r and y^o. The visual and semantic embeddings are fed into a series of fully connected layers separately in order to align their representations in the semantic space, learned by the supervision of triplet losses L^s, L^r and L^o.[2] The learnt continuous vector x^r can preserve the discriminative capability and contextual awareness, and is denoted as the semantic relation $r^s_{i,j}$ between subject and object.

3.3. Multimodal reasoning and fusion

3.3.1. Multi-graph reasoning and fusion layer

Our multimodal reasoning and fusion is based on L Multi-Graph Reasoning and Fusion (MGRF) layers. Let question features $Q^l = \{q^l_1, \dots, q^l_S\}$, visual features $V^l = \{v^l_1, \dots, v^l_K\}$, semantic relations $R^s = \{r^s_{1,1}, \dots, r^s_{1,K}, \dots, r^s_{K,K}\}$ and positional relations $R^p = \{r^p_{1,1}, \dots, r^p_{1,K}, \dots, r^p_{K,K}\}$ denote the input, the MGRF outputs updated question features Q^{l+1} and visual features V^{l+1} by sufficiently reasoning and fusing multiple relations using the following six steps of operation. Note that question-guided object attention and question-guided graph reasoning for the semantic and positional graphs share the common operations but differ in their node and edge representations corresponding to the graph type. Thus we omit the superscript of node and edge representation when we introduce these two steps of operation.

(1) Graph Construction: For most VQA questions, both of semantic and positional relations play vital role for models to understand input image and predict correct answer. Hence, our model construct semantic graph and positional graph in parallel. These graph representations are served as the basis of the following multimodal reasoning and fusion. The semantic graph can be denoted as $\mathcal{G}^l_s = \{\mathcal{V}^l_s, \mathcal{E}^l_s\}$, where $\mathcal{V}^l_s = V^l$ is the node set (each node corresponds to a detected object) and $\mathcal{E}^l_s = R^s$ is the edge set (each edged denotes the semantic relation between objects). We also construct positional graph $\mathcal{G}^l_p = \{\mathcal{V}^l_p, \mathcal{E}^l_p\}$ in similar way, where $\mathcal{V}^l_p = V^l$ and $\mathcal{E}^l_p = R^p$. We assume that certain relationship exists between any pair of objects. Therefore, the above constructed graphs are fully-connected.[3]

(2) Question Self-Attention: To highlight the key words and draw dependencies between words of the question, question self-attention module designed following [47]. It contains a multi-head scaled dot-product attention sub-layer and a feed-forward sub-layer, where the feed-forward sub-layer is further composed of fully-connected layers. Then the question representation $Q^l \in \mathbb{R}^{d_q \times S}$ is updated by incorporating the correlations between words:

$$Q^{l+1} = LN(LN(\tilde{Q}^l + Q^l) + FF(LN(\tilde{Q}^l + Q^l))) \tag{4}$$

$$\tilde{Q}^l = \mathbf{W}^l_0[head_1, \dots, head_H] \tag{5}$$

$$head_h = (\mathbf{W}^l_{1,h} Q^l)\mathbf{a}, h = 1, \dots, H \tag{6}$$

where LN denotes layer normalization [48]. FF denotes the feed-forward layer. $\mathbf{a} \in \mathbb{R}^{S \times S}$ $(\sum_{j=1}^{S} \mathbf{a}_{i,j} = 1, i = 1, \dots, S)$ depicts dependence between each question word and can be calculated by using row-wise softmax on the dot-product of query $\mathbf{W}^l_{2,h} Q^l$ and key $\mathbf{W}^l_{3,h} Q^l$ as follows:

$$\mathbf{a} = softmax(\frac{(\mathbf{W}^l_{2,h} Q^l)^T \cdot (\mathbf{W}^l_{3,h} Q^l)}{\sqrt{d_q/H}}) \tag{7}$$

where H is the number of attention heads. $\mathbf{W}^l_{1,h}, \mathbf{W}^l_{2,h}, \mathbf{W}^l_{3,h} \in \mathbb{R}^{\frac{d_q}{H} \times d_q}$ are the projection matrices for the hth head. This multi-head attention can effectively improve the representation capacity of the updated features [47].

(3) Question-Guided Object Attention: The question-guided object attention examines all the objects to highlight the ones most relevant to the question and is commonly used by previous VQA models [5,15,16,31]. The attention weight of image region i can be calculated as follows:

$$\alpha^l_i = softmax[(\mathbf{W}^l_4 v^l_i)^T \cdot (\mathbf{W}^l_5 q^l_{att})] \tag{8}$$

Where $q^l_{att} \in \mathbb{R}^{d_q}$ is the summary semantics of the question, which can be calculated as follows,

$$q^l_{att} = (Q^{l+1})\mathbf{b}^l \tag{9}$$

[2] More details of these losses and training method can be found in [46]. We directly adopt the trained model as our semantic relation encoder.

[3] Following standard transformer encoder, we also try a variant model which only has the semantic graph whose nodes are the concatenation of visual features and spatial features. The performance can be found in Fig. 6(a)

where $\mathbf{b}^l \in \mathbb{R}^S$ is the attention map over question words of l_{th} layer, obtained by:

$$\mathbf{b}^l = softmax(\mathbf{W}_6^l Q^{l+1}) \tag{10}$$

Thus, we get the attentive visual features $\hat{V}^l = \{\hat{v}_i{}^l\}_{i=1}^K \in \mathbb{R}^{d_v \times K}$, where

$$\hat{v}_i{}^l = \alpha_i^l \cdot v_i^l \tag{11}$$

(4) Question-Guided Graph Reasoning: To correctly answer the question, the VQA model needs to reason and collect information from the constructed graphs guided by the question. Graph Attention Network (GAT) [49] is a straightforward method which aggregates the representations of neighborhood nodes. However, it fails to take semantic/positional relations into consideration, which contain rich information and is conducive to predicting correct answer. Hence, we extend the original GAT and design the Relation-aware Graph Convolution layer to aggregate context information including neighborhood representations and their relations to update representation of each node. As shown in Fig. 4(b), given the representations of each node $\hat{V}^l = \{\hat{v}_i{}^l\}_{i=1}^K$ (see Eq. (11)), we first calculate the importance of the neighborhood node j to node i according to their representations and relation guided by the question:

$$\beta_{ji}^l = softmax((\mathbf{W}_7^l[\hat{v}_j^l, r_{ji}])^T \cdot (\mathbf{W}_8^l[\hat{v}_i^l, q_{att}^l])) \tag{12}$$

$\mathbf{W}_7^l \in \mathbb{R}^{d_v \times (d_v + d_r)}$ and $\mathbf{W}_8 \in \mathbb{R}^{d_v \times (d_v + d_q)}$ are all learnable matrices. d_v, d_q, d_r are dimensions of visual feature, relation embedding, and question embedding respectively. Then the context information containing neighborhood representations and their relations can be expressed as follows:

$$c_i^l = \sum_{j \in \mathcal{N}_i} \beta_{ji}^l \mathbf{W}_7^l [\hat{v}_j^l, r_{ji}] \tag{13}$$

Finally we fuse the gathered context information to update the representation of node i as follows,

$$\tilde{v}_i^l = ReLU(\mathbf{W}_9^l [c_i^l, \hat{v}_i^l]) \tag{14}$$

where $\mathbf{W}_9^l \in \mathbb{R}^{d_v \times (d_v + d_v)}$ is learnable.

Furthermore, as shown in Fig. 4(a), a Feed-Forward (FF) layer is used to nonlinearly transform the outputs of the graph convolution layer. Residual connection [2] and Layer Normalization (LN) [48] is also applied to facilitate optimization. Thus the representation is finally updated as:

$$\dot{v}_i^l = LN[FF(LN(\hat{v}_i^l + \tilde{v}_i^l)) + LN(\hat{v}_i^l + \tilde{v}_i^l)] \tag{15}$$

(5) Question-Guided Adaptive Graph Fusion: As shown in Fig. 1, accurately fusing the positional relations and semantic relations according to the question is vital for VQA models to predict correct answer. Inspired by this idea, we design a question guided graph fusion approach which estimates the weights of positional relation and semantic relation using a single-layer fully connected neural network,

$$[w_V^l, w_G^l] = softmax(\mathbf{W}_{10}^l q_{att}^l) \tag{16}$$

Thus the representation of each image region can be updated by adaptively fusing nodes from two graphs,

$$\bar{v}_i^l = w_V^l \dot{v}_i^{V,l} + w_G^l \dot{v}_i^{G,l} \tag{17}$$

where $\dot{v}_i^{V,l}$ and $\dot{v}_i^{G,l}$ are the nodes of semantic and positional graphs respectively.

(6) Multimodal Fusion: Finally, multimodal fusion is used to fuse attentive question representation q_{att}^l into each image region representation $\overline{v}_i^l$. In practice, we linearly project the concatenation of the visual and textual features to obtain the fused multimodal representation:

$$v_i^{l+1} = \mathbf{W}_{11}^l [\overline{v}_i^l, q_{att}^l] \tag{18}$$

Note that, as we have reviewed in Section 2.3, there are number of bilinear fusion approaches can be used, such as MLB [11], BLOCK [13]. We will compare the performance of our linear projection with those bilinear fusion methods in Section 4.4.

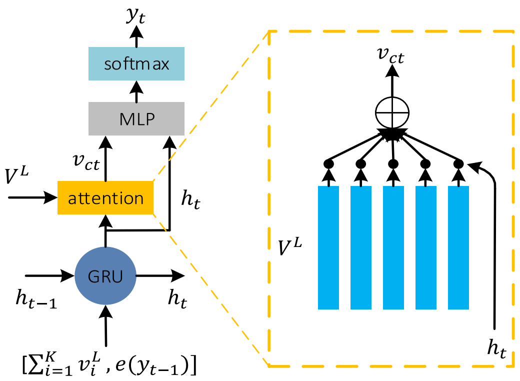


Fig. 5. The structure of explanation generator, which uses GRU and attention mechanism to produce textual justification for predicted answer based on the fused multimodal representation.

3.3.2. Deep multimodal reasoning and fusion network

Our Deep Multimodal Reasoning and Fusion Network (DMRFNet) is composed of L Multi-Graph Reasoning Fusion Layers (denoted by MGRF^1,MGRF^2, …, MGRF^L). Denoting the input for MGRF^l as V^{l-1}, Q^{l-1}, R^p and R^s, the output of MGRF^l is V^l and Q^l, which are further fed to the MGRF^{l+1} as its inputs in a recursive manner.

$$[V^l, Q^l] = MGRF^l([V^{l-1}, Q^{l-1}, R^p, R^s]) \tag{19}$$

Particularly, for MGRF^0, we set $V^0 = V$ and $Q^0 = Q$. Note that the input R^p and R^s for all MGRFs are the same. By stacking these MGRFs, fine-grained interactions are conducted and rich relations are fused.

3.4. Answer prediction

After passing through DMRFNet with L MGRFs, the output $V^L = \{v_1^L, \ldots, v_K^L\}$ already contains rich information of the input image, question and their relations. Following existing work [6,10], we design a Feed Forward Neural Network composed of two fully connected layer with ReLU in its between as a classifier to predict the probabilities of all candidate answers (top-M frequent answers in the training set),

$$\hat{\mathbf{p}} = softmax(FF(\sum_{i=1}^K v_i^L)) \tag{20}$$

$\hat{p}_i \in \hat{\mathbf{p}}$ is the predicted probability for the ith answer, while p_i is its ground-truth label. Following [6,10], the cross-entropy loss is used to train this classifier, formally defined as,

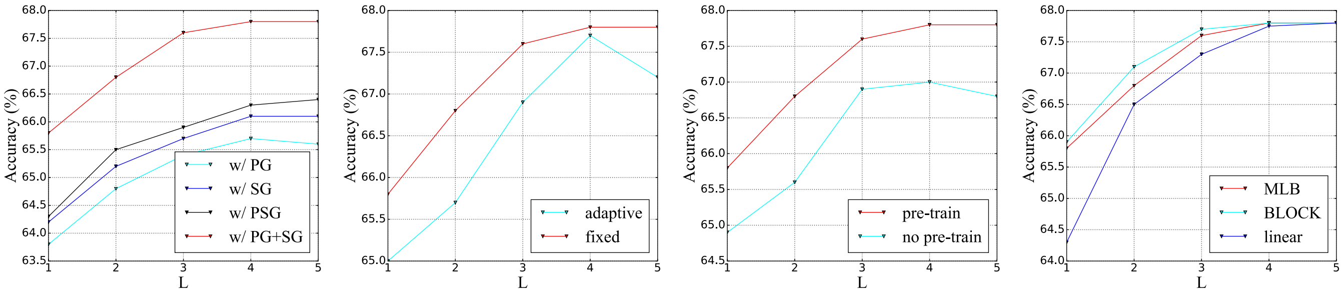
$$\mathcal{L}_{AP} = -\sum_{j=1}^B \sum_{i=1}^M (p_i^{(j)} log(\hat{p}_i^{(j)})) + (1 - p_i^{(j)}) log(1 - \hat{p}_i^{(j)}) \tag{21}$$

where $\hat{p}_i \in \hat{\mathbf{p}}$ is the predicted probability for the ith answer, while p_i is its ground-truth label. B denotes the batch-size, j subscripts the index of the training sample in the batch, and M is the size of the candidate answer set.

3.5. Explanation generation

To generate an explanation, we adopt an GRU-based language model that takes the fused multimodal representation V^L as input. The detailed structure of our explanation generator G is shown in Fig. 5. At time step t, the GRU takes the concatenation of the mean-pooled multimodal representation, its previous output, and the previously generated word embedding $e(y_{t-1})$ as input, and generates the output as follows:

$$h_t = GRU(h_{t-1}, [\sum_{i=1}^K v_i^L, e(y_{t-1})]) \tag{22}$$



(a) (b) (c) (d)

Fig. 6. The overall accuracy of our variants: (a) impacts of positional graph and semantic graph. (b) impact of graph fusion. (c) impact of semantic relation encoder pre-training. (d) comparison with bilinear fusion. The number of layers $L = \{1, 2, 3, 4, 5\}$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To make the generator focus on the important fused information, the attention mechanism is used to compute the attention weight $\omega_{t,i}$ over $v_i^L \in V^L$:

$$\omega_{t,i} = \frac{exp((\mathbf{W}_{12}h_t)^T(\mathbf{W}_{13}v_i^L))}{\sum_{j=1}^{K} exp((\mathbf{W}_{12}h_t)^T(\mathbf{W}_{13}v_j^L))} \tag{23}$$

Finally, we use a simple fully-connected layer to obtain the probability of the next word y_t based on the attentive multimodal feature $v_{ct} = \sum_{i=1}^{K} \omega_{t,i} v_i^L$ and the output of GRU h_t:

$$p_G(y_t|\mathbf{y}_{1:t-1}, V^L) = softmax(\mathbf{W}_{14}(v_{ct} + h_t)) \tag{24}$$

Given the ground-truth explanation $\mathbf{y}^*_{1:T}$, we train our explanation generator G by minimize the following loss:

$$\mathcal{L}_{EG} = -\sum_{t=1}^{T} log(p_G(y_t^*|\mathbf{y}^*_{1:t-1}, V^L)) \tag{25}$$

4. Experiments

4.1. Datasets

VQA 2.0 [27]: It is one of the largest and most popular VQA datasets, which contains 443,757 train questions-answer pairs, 214,354 validation questions-answer pairs and 447,793 test questions-answer pairs (named as $test - standard$), relating to 123,287 MSCOCO images [50]. Additionally, a 25% subset of the $test - standard$ referred to as $test - dev$ is also provided to facilitate model evaluation. There are three kinds of questions including *yes/no*, *number* and *other*. Different voters annotate 10 answers for each question. The accuracy of answer prediction is reported using the tools provided by [51].

VQA-E [28]: This dataset is automatically derived from VQA 2.0 dataset by synthesizing an explanation for each question–answer pair. For each question–answer pair, the most relevant MSCOCO caption is retrieved by measuring the similarity between the question–answer pair and captions. Then the explanation is generated by fusing the question, answer, and relevant caption. Finally, 269 786 question–answer–explanation triplets based on 10 325 images are generated, including 18 298 training samples and 88 488 validation samples. We use the training set to train our model and report the predicted answers and generated explanations of the validation set. We evaluate the generated textual explanations using BLEU-4 [52], METEOR [53], ROUGE [54], CIDEr [55].

4.2. Training details

Our model is implemented with the PyTorch library on a machine with 4 Nvidia Geforce 2080Ti GPUs. We use the Adamax solver with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The batch size is set to be 64. Dropouts ($p = 0.2$) are used after each fully connected layers to prevent overfitting. For encoding questions, word embedding is adopted to map each word into a 300 dimensional vector. The hidden state of Bi-GRU is set to be 512, thus $d_q = 1024$. The number of attention heads in question self-attention is set to 8 ($H = 8$). For VQA 2.0 dataset, we choose the top-3000 frequent answers in the training set to form the set of candidate answers ($M = 3000$). The questions are truncated or padded, so that $S = 14$. We follow the same settings as VQA 2.0 for the VQA-E dataset and use the same candidate answers.

As shown in Fig. 2, our model has three parts: Deep Multimodal Reasoning and Fusion Network (DMRFNet), Answer Predictor (AP), and Explanation Generator (EG). DMRFNet and AP make up the answering model, while the explanation model is composed of DMRFNet and EG. We train our model as follows: Following previous works [6,17,27,56], we use the training samples from VQA 2.0 to train the answering model using the loss function $\mathcal{L}_{AP}$. Since these samples have no explanation annotations, we freeze the explanation generator when training the answering model. We set the learning rate to min($3te-5$, $1e-4$), where t is the current epoch number. After 10 epochs, the learning rate is decayed by 1/4 every 2 epochs. The total number of epoch is set to 15. Then we report the test results on VQA 2.0. Next, we unfreeze the explanation generator, and use the image–question–answer–explanation samples from VQA-E train split to train the full model including DMRFNet, AP, and EG together under the supervision of $\mathcal{L}_{AP} + \mathcal{L}_{EG}$. specifically, we initialize the answering model with the weights from the model trained on VQA 2.0. And we set a small initial learning rates $5e-5$ to fine tune the answering model, while the initial learning rate of the explanation generator is set to be 0.001. Both of these learning rates are decayed by 1/2 every 3 epochs.

4.3. Comparison with state-of-the-arts

Table 1 compares our models with state-of-the-art discriminative models on VQA 2.0. All the results are obtained by single model. Simple multimodal fusion models, including MCB [12], MLB [11], MFB [4], MFH [4], and BLOCK,BLOCK, are shown in the first block of the table. Deep fusion models based on stacking attention layers, such as DCN [5], BAN [17], DFAF [18], and MCAN [6], are shown in the second block. These models adopt pretrained Faster R-CNN to detect salient object and they also use GloVe [60] to initialize word embedding. We also show some recent models based on visual relational reasoning in the third block. And our models are in the last block: DMRFNet+AP denotes our answering model while DMRFNet+AP+EG is our full model. The experimental results demonstrate that: (1) Our models significantly outperform the simple multimodal fusion models in the first block. (2) Using the same strong Faster R-CNN features and GloVe, our full model consistently outperform the deep fusion approaches in the second block and relational reasoning models in the third block on most metrics. DCN [5], BAN [17], DFAF [18], and MCAN [6] achieved fine-grained multimodal fusion by stacking multiple co-attention layers. MuRel [22], ReGAT [57], and our previous work VRR [10] build deep reasoning

Table 1
Comparison of single model performance on VQA 2.0.

Model	Test-dev				Test-standard
	Overall	Other	Number	Yes/No	Overall
MCB reported in [4]	65.4	57.4	37.2	82.3	–
MLB reported in [4]	65.8	56.8	37.9	83.9	–
MFB [4]	66.9	58.4	39.1	84.9	66.6
MFH [4]	67.7	59.2	40.2	84.9	67.5
BLOCK [13]	67.5	58.5	47.33	83.6	67.9
DCN [5]	66.60	56.72	46.60	83.50	67.00
BAN [17]	69.66	60.50	50.66	85.46	–
DFAF [18]	70.22	60.49	53.32	86.09	70.34
MCAN [6]	70.63	60.72	53.26	**86.82**	70.90
VRR [10]	67.20	58.41	45.51	83.31	67.34
MuRel [22]	68.03	57.85	49.84	84.77	68.41
ReGAT [57]	70.27	60.33	**54.42**	86.08	70.58
ViLBERT [58]	70.55	–	–	–	70.34
VisualBERT [59]	70.80	–	–	–	71.00
DMRFNet+AP	71.18	61.45	53.58	86.80	71.10
DMRFNet+AP+EG	**71.24**	**61.97**	53.84	86.81	**71.27**

Table 2
Comparison of single model performance on VQA-E.

Model	Accuracy	BLEU-4	METEOR	ROUGE	CIDEr
Bottom-Up [56]	65.32	–	–	–	–
BAN [17]	70.65	–	–	–	–
VRR [10]	68.36	–	–	–	–
MCAN [6]	71.09	–	–	–	–
QI-E [28]	–	8.60	16.57	34.92	84.07
QI-AE [28]	67.35	9.40	17.37	36.33	93.08
DMRFNet+AP	71.26	–	–	–	–
DMRFNet+EG	–	12.30	20.45	39.96	96.76
DMRFNet+AP+EG	**71.74**	**14.50**	**21.85**	**41.18**	**97.08**

Table 3
Ablation experiments for DMFNet on the *val* split of VQA 2.0.

(a) **MGRF variants:** Accuracies of the DMRFNet models with different MGRF variants. All variants use **one** layer of MGRF.

Model	All	Other	Num	Y/N
-w/o QSA	65.3	54.5	45.5	83.2
-w/o OA	64.8	54.3	44.5	82.5
-w/o GR	64.6	54.0	44.6	82.4
Full MGRF	65.8	55.6	45.5	83.3

(b) **Stacking layer:** Accuracies of the DMRFNet model with different layers of MGRF.

Layer	All	Other	Num	Y/N
1	65.8	55.6	45.5	83.3
2	66.8	56.8	47.8	84.3
3	67.6	57.9	48.9	84.9
4	67.8	58.9	49.7	85.5
5	67.8	58.8	49.8	85.5

network which implicitly reasons relation between objects, obtaining impressive results. Our models in this paper obtain better result by designing deep network stacked of Multi-Graph Reasoning and Fusion layers and explicitly building two graphs to respectively exploit positional and semantic relations. Especially, our full model significantly promotes the accuracy of *"Other"* question which need the models conduct complex reasoning from 60.72 to 61.97. (3) Furthermore, our full model achieves similar results with recent large-scale pre-trained models such as ViLBERT [58] and VisualBERT [59] which is shown in the fourth block. (4) There is another interesting phenomenon that our full model DMRFNet+AP+EG consistently outperforms our answering model DMRFNet+AP on all question types. This means that forcing the model to explain and training the answering model and explanation model together can significantly help predict correct answers. The reason for this phenomenon may be that the textual explanations give detailed description of the image as shown in Fig. 1, and our DMRFNet are forced to generate multimodal representation containing more clear sense of the image content under the supervision of $\mathcal{L}_{EG}$.

We further compare our model with state-of-the-arts on the VQA-E dataset in Table 2. We test Bottom-Up [14], BAN [17], VRR [10], and MCAN [6] using the codes released by the authors. These models can only predict answers based on the fused image-question representation and there is no explanation generated. Deep co-attention models including BAN and MCAN have obvious advantage on the accuracy of predicted answers. The results are shown in the first block of Table 2. The second block shows two models from [28]. These models are similar to Bottom-up, which use Faster R-CNN features and attention mechanism to obtain discriminative image features. QI-E is an explanation model which can only generate textual explanations based on fused features, while QI-AE can simultaneously predict answers and generate explanations. The results of our models are shown in the bottom block. We can see that our answering model (DMRFNet+AP) and full model (DMRFNet+AP+EG) consistently outperform state-of-the-arts on answer prediction. To be specific, our full model achieves highest accuracy 71.74%, surpassing the current best model MCAN. We also find that both of our explanation model (DMRFNet+EG) and full model generate higher-quality textual explanations than the state-of-the-art models. This comparison demonstrates that the output of our DMRFNet contains rich clues for answer prediction and explanation generation.

4.4. Ablation study

In this section, we design variants to investigate the contribution of each component of our model. We train them on the *train* split and report the performance on the *val* split. The results are shown in Table 3 and discussed in detail below.

MGRF variants: To evaluate the importance of components of MGRF, as shown in Table 3, we design the following variants: (1) MGRF without question self-attention, denoted as "-w/o QSA", which causes the overall accuracy to drop from 65.8 to 65.3; (2) MGRF without question-guided object attention, denoted as "-w/o OA", which only achieves 64.8 overall accuracy; (3) MGRF without question-guided graph reasoning, denoted as "-w/o GR", which causes a sharp drop of performance. All of these variants have only one MGRF layer. We can see that all of question self-attention module, question-guided object attention module, and graph reasoning module are important for predicting correct answer. Our full MGRF combines all the above modules together and obtains best performance, achieving 65.8 overall accuracy.

Stacking layer: Our DMRFNet achieves fine-grained deep reasoning and fusion by stacking MGRF layers. Hence, we try DMRFNet with different MGRF layers. We set $L = \{1, 2, 3, 4, 5\}$ and compare the results in Table 3. As expected, the performance of DMRFNet steadily improves with increasing L, and finally saturates at $L = 4$. The saturation can be the result of unstable gradients during training when $L > 4$, which makes the optimization difficult. It is also observed in [61].

Impacts of multi-graph: To evaluate the importance of semantic graph and positional graph in our MGRF layer, we design four MGRF variants: (1) MGRF with only positional graph, denoted as "w/ PG". (2)MGRF with only semantic graph, denoted as "w/ SG". (3) MGRF with only one graph whose input is the concatenation of visual features and positional features, denoted as "w/ PSG". This method is similar to the Transformer Encoder [47]. (4) full MGRF with both positional and semantic graphs, denoted as "w/ PG+SG". We set the number of MGRF layers $L = \{1, 2, 3, 4, 5\}$ and report the overall accuracies in Fig. 6(a). We can see that removing any one of these two graphs causes sharp drop of accuracy, proving that both positional relation and semantic relation are vital for VQA models to predict correct

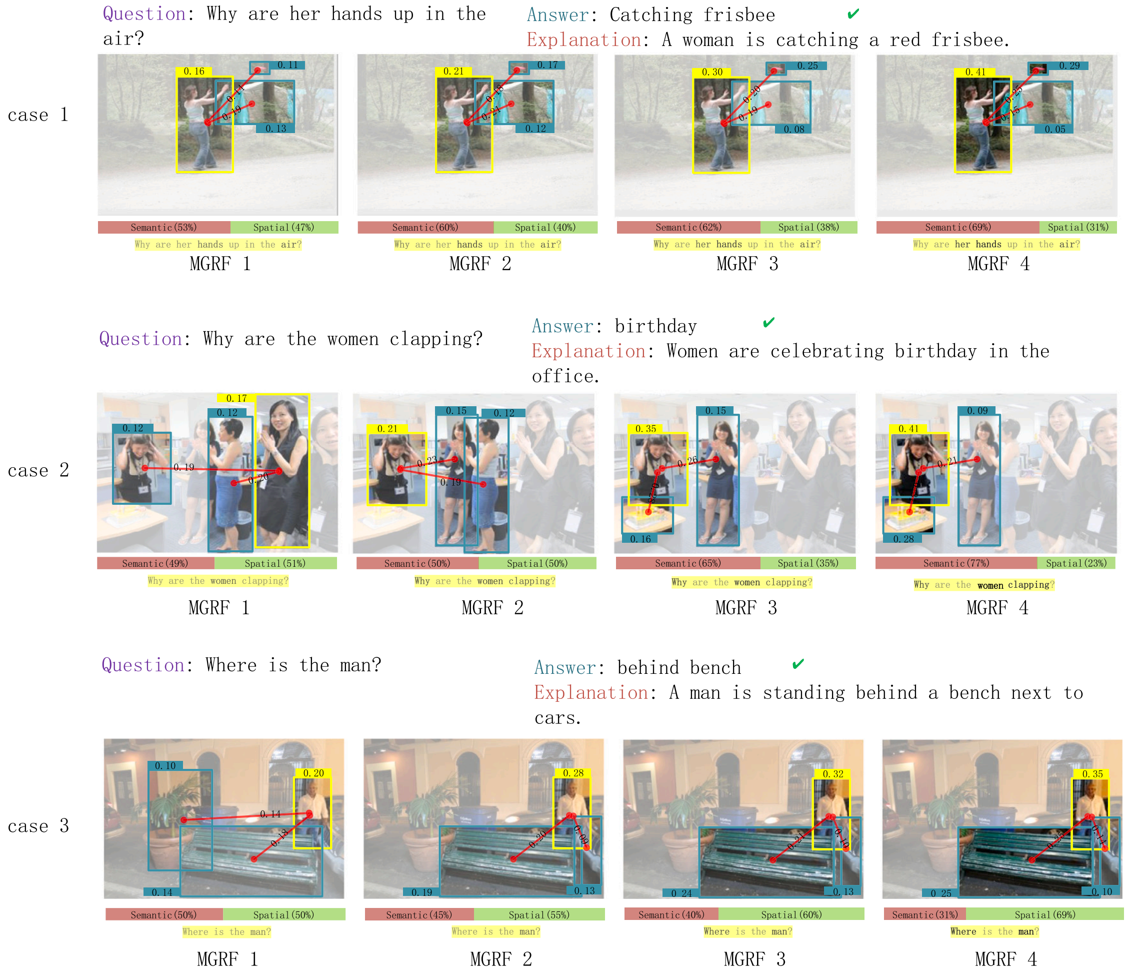

Question: Why are her hands up in the air?
Answer: Catching frisbee ✔
Explanation: A woman is catching a red frisbee.
case 1
Semantic(53%) Spatial(47%)
Semantic(60%) Spatial(40%)
Semantic(62%) Spatial(38%)
Semantic(69%) Spatial(31%)
Why are her hands up in the air?
MGRF 1
MGRF 2
MGRF 3
MGRF 4
Question: Why are the women clapping?
Answer: birthday ✔
Explanation: Women are celebrating birthday in the office.
case 2
Semantic(49%) Spatial(51%)
Semantic(50%) Spatial(50%)
Semantic(65%) Spatial(35%)
Semantic(77%) Spatial(23%)
Why are the women clapping?
Question: Where is the man?
Answer: behind bench ✔
Explanation: A man is standing behind a bench next to cars.
case 3
Semantic(50%) Spatial(50%)
Semantic(45%) Spatial(55%)
Semantic(40%) Spatial(60%)
Semantic(31%) Spatial(69%)
Where is the man?

Fig. 7. Qualitative examples from VQA-E. Four columns correspond to 4 layers of MGRFs respectively. The most selected objects by visual attention (α^l in Eq. (8)) are highlighted in yellow box. And its top-2 attended neighbors (β^l in Eq. (12)) are shown in green boxes. Their semantic relations are also shown using red lines with weights on them. The bar below image shows the contributions of spatial and semantic graphs ($[w_V^l, w_G^l]$ in Eq. (16)). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

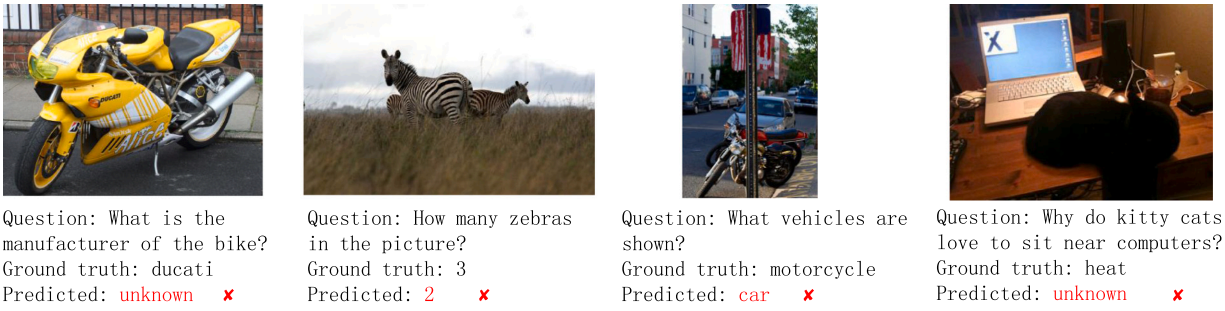

Question: What is the manufacturer of the bike?
Ground truth: ducati
Predicted: unknown ✘
Question: How many zebras in the picture?
Ground truth: 3
Predicted: 2 ✘
Question: What vehicles are shown?
Ground truth: motorcycle
Predicted: car ✘
Question: Why do kitty cats love to sit near computers?
Ground truth: heat
Predicted: unknown ✘

Fig. 8. Typical failure cases of our model.

answer. In addition, the “w/ PSG” variant outperforms “w/ PG” and “w/ SG”, meaning that concatenating visual and positional features can improve performance. However, “w/ PSG” obtains inferior result to “w/ PG+SG”, demonstrating that explicitly and adaptively reasoning and fusing semantic and positional relations is a better approach to exploit useful clues for predicting correct answers.

To effectively fusing the information from two graphs according to the current question, our MGRF is equipped with question-guided adaptive graph fusion module. We also try a baseline to fuse graphs which directly fusing semantic and positional graphs by setting $\omega_V^l = \omega_G^l = 0.5$ (see Eq. (17)). From the results in Fig. 6(b), we can see that our proposed question guided adaptive graph fusion approach (the red curve) achieves better accuracy using any number of MGRF layers than directly summation (the green curve).

Impacts of semantic relation encoder pre-training: Our model uses pre-trained semantic relation encoder (see Fig. 3) to obtain semantic relation embeddings between visual objects. And our semantic graph is constructed based on these embeddings. To prove the effectiveness of semantic relation encoder pre-training, we design a variant which do not use pre-training but plugs the relation encoder network into our model and train it in an end-to-end manner with the VQA samples. The comparison is shown in Fig. 6(c). It is demonstrated that pre-training of relation encoder provides significant performance improvement.

Comparison with bilinear fusion: In our final model, we simply concatenate the visual and textual features and linearly project the concatenation to update the visual representation of each image region using Eq. (18). As we have discussed in Section 3.3.1, any kinds of bilinear fusion module can be plugged into our model. Therefore, we try two popular bilinear fusion methods including MLB [11] and BLOCK [13] and compare their performance with our simple linear method. The results in Fig. 6(d) show that BLOCK and MLB achieve better results, especially when we use fewer than 3 MGRF layers. However, as the number of MGRF layers increases, our linear method achieves similar performance. Hence, we choose this linear method in our final model since it has less parameters.

4.5. Visualization

In this section, we analyze the behavioral character and interpretability of our model by visualizing several typical examples. From case study in Figs. 7 and 8, we sum up the following three insights:

DMRFNet is capable to reason and fuse information from vision and language modalities. In Fig. 7, we visualize the behavior of the DMRFNet with 4 layers of MGRFs. The four columns correspond to the 4 layers. The first column shows the question self-attention map over words (b^l in Eq. (10)), visual attention map over objects (α^l in Eq. (8)), semantic relations between the most attended object and its top-2 relevant objects (β^l in Eq. (12)). For a clear display, the positional graph, which is similar to the semantic graph, is not given. The rest columns can be deduced by analogy. We can see that iterations through the MGRFs tend to gradually discard regions, keeping only the most relevant ones, under the guidance of input question. This makes the fused multimodal feature more discriminative, that is beneficial to the downstream task. We also can see that the relations captured by our model match human intuition.

DMRFNet is capable to reveal the information selection mode from different graphs. We also quantitatively analyze the contributions of spatial and semantic graph in our MGRF layer, by showing $\{\omega_V^l, \omega_G^l\}, l = 1, 2, 3, 4$ in the bar below every image. By visualizing these contribution values, we can reveal the information selection mode of our model. In the first example (first row in Fig. 7), the interaction between *"woman"* and *"frisbee"* is the key to answer the question *"Why are her hands up in the air?"*. We can see that semantic graph contributes more than positional graph in each MGRF layer in this case. This information selection mode matches human intuition and makes our model more interpretable. This phenomenon also can be observed in the other examples, which demonstrates that our proposed approach has the ability to mine complementary information that is vital for model predicting correct answer.

DMRFNet fails mostly in four conditions: scene text relevant answers, inadequate visual evidence, ambiguous question and limited external knowledge. These failure examples are shown in Fig. 8. (1) Our model fails when the question needs the model recognize optical characters. This case is called *TextVQA* which is a recently proposed task [62]. (2) Some cases fail when there is inadequate visual evidence, such as the second example in which the third zebra is sheltered. (3) Some cases fail because of the ambiguity of the question. For instance, both *"motorcycle"* and *"car"* appear in the image and can be the answer to the question. (4) Our model also fails to answer questions requiring external knowledge and common sense. We can see that the VQA model needs prior knowledge that computers can generate heat that makes cat feel comfortable, when answering *"Why do kitty cats love to sit near computers?"*. Unfortunately, our model has no such knowledge.

5. Conclusion

In this paper, we proposed a novel Multi-Graph Reasoning and Fusion layer (MGRF), which can reason positional and semantic relations simultaneously, and adaptively fuse these relations according to the question. By stacking multiple MGRFs, we build Deep Multimodal Reasoning and Fusion Network (DMRFNet) to achieve fine-grained reasoning and fusion. This network is used to address VQA task. Furthermore, we also proposed an explanation generator to generate textual justification for predicted answer, which reveals the motivation of the model's decision. Quantitative and qualitative experimental results on both VQA 2.0 and VQA-E demonstrate the effectiveness and interpretability of our model.

CRediT authorship contribution statement

Weifeng Zhang: Conceptualization, Methodology, Software, Writing - original draft. **Jing Yu:** Visualization, Formal analysis, Writing - review & editing. **Wenhong Zhao:** Data curation, Investigation. **Chuan Ran:** Software.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by the Natural Science Foundation of Zhejiang Province, China (Grant No. LQ21F020022).

References

[1] K. Simonyan, A. Zisserman, Very deep convolutional networks for large scale image recognition, in: ICLR, 2015.

[2] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778.

[3] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: NIPS, 2015, pp. 91–99.

[4] Z. Yu, J. Yu, J. Fan, D. Tao, Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering, IEEE Trans. Neural Netw. Learn. Syst. 99 (2017) 1–13.

[5] D. Nguyen, T. Okatani, Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering, in: CVPR, 2018, pp. 6087–6096.

[6] Z. Yu, J. Yu, Y. Cui, D. Tao, Q. Tian, Deep modular co-attention networks for visual question answering, in: CVPR, 2019, pp. 6281–6290.

[7] Z. Zhao, Z. Zhang, S. Xiao, Z. Yu, Open-ended long-form video question answering via adaptive hierarchical reinforced networks, in: IJCAI, 2018, pp. 4383–4389.

[8] Y. Jang, Y. Song, Y. Yu, G. Kim, Y. Kim, Tgi-qa: Toward spatio-tempporal reasoning in visual question answering, in: CVPR, 2017, pp. 1359–1367.

[9] H. Xue, Z. Zhao, D. Cai, Unifying the video and question attentions for open-ended video question answering, IEEE Trans. Image Process. (2017) 5656–5666.

[10] W. Zhang, J. Yu, H. Hu, H. Hu, Z. Qin, Multimodal feature fusion by relational reasoning and attention for visual question answering, Inf. Fusion 55 (2020) 116–126.

[11] J. Kim, K. On, W. Lim, Hadamard product for low-rank bilinear pooling, in: ICLR, 2017.

[12] A. Fukui, D.H. Park, D. Yang, A. Rohrbach, Multimodal compact bilinear pooling for visual question answering and visual grounding, in: EMNLP, 2016, pp. 457–468.

[13] H. Ben-younes, R. Cadene, N. Thome, M. Cord, Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection, in: AAAI, 2019, pp. 8102–8109.

[14] P. Anderson, X. He, C. Buehler, D. Teney, Bottom-up and top-down attention for image captioning and visual question answering, in: CVPR, 2018, pp. 6077–6086.

[15] Z. Yang, X. He, J. Gao, L. Deng, A. Smola, Stacked attention networks for image question answering, in: CVPR, 2016, pp. 21–29.

[16] J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, in: NIPS, 2016, pp. 289–297.

[17] J.-H. Kim, J. Jun, B.-T. Zhang, Bilinear attention networks, in: NIPS, 2018, pp. 1571–1581.

[18] P. Gao, Z. Jiang, H. You, et al., Dynamic fusion with intra- and inter-modality attention flow for visual question answering, in: CVPR, 2019, pp. 6639–6648.

[19] D. Raposo, A. Santoro, D. Barrett, M. Malinowski, A simple neural network module for relational reasoning, in: NIPS, 2017, pp. 4967–4976.

[20] E. Perez, F. Strub, H. Vries, V. Dumoulin, A. Courville, Film: visual reasoning with a general conditioning layer, in: AAAI, 2018, pp. 3942–3951.

[21] P. Wang, Q. Wu, C. Shen, The vqa-machine: Learning how to use existing vision algorithms to answer new questions, in: CVPR, 2017, pp. 3909–3918.

[22] R. Cadene, H. Ben, M. Cord, N. Thome, Murel: Multimodal relational reasoning for visual question answering, in: CVPR, 2019, pp. 1989–1998.

[23] R. Hu, A. Rohrbach, T. Darrell, K. Saenko, Language-conditioned graph networks for relational reasoning, in: ICCV, 2019.

[24] A. Das, H. Agrawal, L. Zitnick, D. Parikh, D. Batra, Human attention in visual question answering: Do humans and deep networks look at the same regions?, in: EMNLP, 2016, pp. 932–937.

[25] C. Gan, Y. Li, H. Li, C. Sun, B. Gong, Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation, in: ICCV, 2017, pp. 1829–1838.

[26] T. Qiao, J. Dong, D. Xu, Exploring human-like attention supervision in visual question answering, in: AAAI, 2018, pp. 7300–7307.

[27] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh, Making the v in vqa matter: elevating the role of image understanding in visual question answering, in: CVPR, 2017, pp. 6325–6334.

[28] Q. Li, Q. Tao, S.R. Joty, J. Cai, J. Luo, Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions, in: ECCV, 2018, pp. 570–586.

[29] Z. Yu, J. Yu, J. Fan, D. Tao, Multi-modal factorized bilinear pooling with co-attention learning for visual question answering, in: ICCV, 2017, pp. 1839–1848.

[30] M. Cord, N. Thome, H. Ben-younes, R. Cadene, Mutan: Multimodal tucker fusion for visual question answering, in: ICCV, 2017, pp. 2631–2639.

[31] V. Kazemi, A. Elqursh, Show, ask, attend, and answer: A strong baseline for visual question answering, 2017, in: arXiv:1704.03162v2.

[32] D.H. Park, L.A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, M. Rohrbach, Multimodal explanations: Justifying decisions and pointing to the evidence, in: CVPR, 2018, pp. 8779–8788.

[33] P.W. Battaglia, J.B. Hamrick, V.e.a. Bapst, Relational inductive biases, deep learning, and graph networks, 2018, in:.

[34] S. Gould, J. Rodgers, D. Cohen, G. Elidan, D. Koller, Multi-class segmentation with relative location prior, Int. J. Comput. Vis. 80 (3) (2008) 300–316.

[35] S.K. Divvala, A. Farhadi, C. Guestrin, Learning everything about anything: Webly-supervised visual concept learning, in: CVPR, 2014, pp. 3270–3277.

[36] L. Cewu, K. Ranjay, B. Michael, F. Li, Visual relationship detection with language priors, in: ECCV, 2016, pp. 852–869.

[37] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. Shamma, M. Bernstein, L. Fei-Fei, Visual genome: Connecting language and vision using crowdsourced dense image annotations, IJCV 123 (1) (2017) 32–73.

[38] A. Peter, F. Basura, J. Mark, G. Stephen, Spice: Semantic propositional image caption evaluation, in: ECCV, 2016, pp. 382–398.

[39] W. Norcliffe-Brown, E. Vafeias, S. Parisot, Learning conditioned graph structures for interpretable visual question answering, in: NIPS, 2018, pp. 8344–8353.

[40] S. D'mello, J. Kory, A review and meta-analysis of multimodal affect detection system, ACM Comput. Surv. 47 (3) (2015) 43–79.

[41] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Inf. Fusion 37 (2017) 98–125.

[42] L. Piras, G. Giacinto, Information fusion in content based image retrieval: A comprehensive overview, Inf. Fusion 37 (2017) 50–60.

[43] Y.-C. Chen, L. Li, L. Yu, A.E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, J. Liu, Uniter: Universal image-text representation learning, in: ECCV, 2020, pp. 104–120.

[44] Z. Yu, Y. Cui, J. Yu, M. Wang, D. Tao, Q. Tian, Deep multimodal neural architecture search, in: ACM MM, 2020, pp. 3743–3752.

[45] S. Ren, K. He, R. Girshick, X. Zhang, J. Sun, Object detection networks on convolutional feature maps, in: CVPR, 2017, pp. 1476–1481.

[46] J. Zhang, Y. Kalantidis, M. Rohrbach, M. Paluri, A. Elgammal, M. Elhoseiny, Large-scale visual relationship understanding, in: AAAI, 2019, pp. 9185–9194.

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, et al., Attention is all you need, in: NIPS, 2017, pp. 5998–6008.

[48] J.L. Ba, J.R. Kiros, G. Hinton, Layer normalization, 2016, in: arXiv preprint arXiv:1607.06450.

[49] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, et al., Graph attention networks, in: ICLR, 2018.

[50] T. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, Microsoft coco: common objects in context, in: ECCV, 2014, pp. 740–755.

[51] S. Antol, A. Agrawal, J. Lu, M. Mitchell, Vqa: visual question answering, in: ICCV, 2015, pp. 2425–2433.

[52] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: ACL, 2002, pp. 311–318.

[53] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: ACL, 2005, pp. 65–72.

[54] C. Lin, Rouge: a package for automatic evaluation of summaries, in: ACL, 2004, pp. 168–175.

[55] R. Vedantam, L. Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: CVPR, 2015, pp. 4566–4575.

[56] D. Teney, P. Anderson, X. He, A. Hengel, Tips and tricks for visual question answering: learnings from the 2017 challenge, 2017, in: arXiv:1708.02711v1.

[57] L. Li, Z. Gan, Y. Cheng, J. Liu, Relation-aware graph attention network for visual question answering, in: ICCV, 2019.

[58] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: KDD, 2019.

[59] D. Yin, C.-J. Hsieh, K.-W. Chang, L.H. Li, M. Yatskar, Visualbert: A simple and performant baseline for vision and language, 2019, in: arXiv preprint arXiv:1908.03557.

[60] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: EMNLP, 2014, pp. 1532–1543.

[61] B. Ankur, M. Chen, O. Firat, Y. Cao, Y. Wu, Training deeper neural machine translation models with transparent attention, 2018, in: arXiv preprint arXiv:1808.07561.

[62] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, M. Rohrbach, Towards vqa models that can read, in: CVPR, 2019, pp. 8317–8326.