Signal Processing: Image Communication 96 (2021) 116319

Contents lists available at ScienceDirect

Signal Processing: Image Communication

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

Multi-scale relation reasoning for multi-modal Visual Question Answering

Yirui Wu [a], Yuntao Ma [b], Shaohua Wan [c,*]

[a] College of Computer and Information, Hohai University, Fochengxi Road, Nanjing 210093, China
[b] National Key Lab for Novel Software Technology, Nanjing University, Xianling Road, Nanjing 210093, China
[c] School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan, China

ARTICLE INFO

Keywords:
Multi-modal data
Visual Question Answering
Multi-scale relation reasoning
Attention model

ABSTRACT

The goal of Visual Question Answering (VQA) is to answer questions about images. For the same picture, there are often completely different types of questions. Therefore, the main difficulty of VQA task lies in how to properly reason relationships among multiple visual objects according to types of input questions. To solve this difficulty, this paper proposes a deep neural network to perform multi-modal relation reasoning in multi-scales, which successfully constructs a regional attention scheme to focus on informative and question-related regions for better answering. Specifically, we firstly design regional attention scheme to select regions of interest based on informative evaluation computed by a question-guided soft attention module. Afterwards, features computed by regional attention scheme are fused in scaled combinations, thus generating more distinctive features with scalable information. Due to designs of regional attention and multi-scale property, the proposed method is capable to describe scaled relationships from multi-modal inputs to offer accurate question-guided answers. By conducting experiments on VQA v1 and VQA v2 datasets, we show that the proposed method has superior efficiencies than most of the existing methods.

1. Introduction

Visual understanding is one of the most challenging tasks in computer vision. Relying on the holistic scene understanding, the goal of Visual Question Answering (VQA) is to answer a question about an image by involving different levels of visual representation of images, and textual representation of questions. Due to its ability to automatically serve customers by answering different kinds of questions, VQA has been widely used in customer service [1], robotic answering [2,3] and so on [4], thus being one of the most hottest research topics in computer vision community.

The core idea for VQA problem is to learn the co-occurrence of a particular combination of features extracted from both images and questions. Following such idea, researchers try to construct space embedding for images and words via deep neural networks, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). To accurately build mappings from specific language elements to particular visual objects, attention schemes are adopted to improve performance of VQA on the basis of deep neural networks. Recently, there exists a new trend to solve VQA with the latest graph network by introducing external and structural knowledge, which has already achieved remarkable results.

In the stage of relational reasoning of VQA, most of current methods locally fuse features of different modes through joint feature embedding or attention mechanism. Meanwhile, VQA methods need to acquire information about position, size and etc among visual objects in global or semantical sense. Locally fusing or processing information of input images and questions cannot guarantee to achieve accurate question-related answers. In other words, most of current methods make use of local relationship rather than extracting global relationship among question-related objects. Considering local characteristics of input images, VQA task inherently requires to reason relationships among different regions or in global sense, thus discovering complex patterns among multiple modal inputs. To further explain such requirement, we show one example of public dataset, i.e., MSCOCO VQA v2 [5] in Fig. 1, where we can clearly view relations required for answering are "standing on" and "in the tree" respectively, even with the same visual objects, i.e., bears. Therefore, accurately answer questions about several visual objects can only be acquired by modeling relationship in a global sense.

Based on the former analysis, we further conclude two main difficulties in VQA, namely multi-modalities and relational reasoning. Firstly, different input modalities have different feature spaces, which brings challenges in recognizing the same object with different feature representations. We thus need to deal with feature processing by mapping different inputs into a united feature space. Secondly, the VQA task requires focusing on question-related context information in multiple scales, rather than only extracting features of objects themselves. In other words, input questions can be highly related to the quantity of

* Corresponding author.
E-mail address: shaohua.wan@ieee.org (S. Wan).

https://doi.org/10.1016/j.image.2021.116319
Received 23 August 2020; Received in revised form 5 May 2021; Accepted 6 May 2021
Available online 14 May 2021

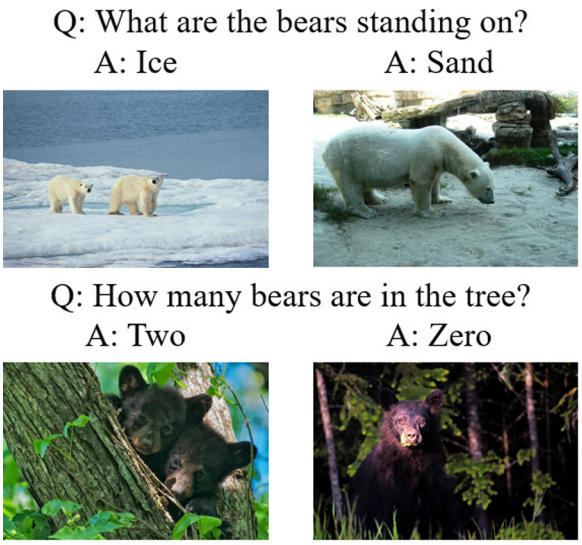

Fig. 1. Examples to show challenge of VQA task, i.e., relations required for answering are "standing on" and "in the tree" respectively, even with the same visual objects.

objects. Without prior knowledge of a number of involved objects, we are supposed to try combinations of multiple objects for answering, which is the core idea of performing VQA in multiple scales. How to appropriately fuse question-related information in different modalities and scales thus become the main challenge of VQA. Since the main task of multi-modality is to map feature spaces of different modalities to the same feature space, the information between different modalities can be transferred to each other, or they can be processed at the same time.

To handle these difficulties, we proposed to perform multi-modal relational reasoning in multi-scales, where a regional attention module is constructed to enhance question-related information and reduce computation complexity. The main contribution can be summarized as follows:

- We propose a novel scheme for the VQA task, which is capable to fuse multi-modal data in multiple scales for more accurate answers.
- We propose a regional attention module to compute informative weights for each region, which further picks up local regions to extract question-related context information for further processing.

The rest of the paper is organized as follows. Section 2 gives an overview of the related work on relative aspects. In Section 3, details of the proposed method is discussed, including Multimodal Feature Extraction, Regional Attention Module, Multiscale Relation Reasoning, and Multimodal Fusion. Section 4 shows our experimental results with several comparative methods. Finally, Section 5 concludes the paper.

2. Related work

As a challenging task, VQA has received extensive attention by recognizing as a visual Turing test. Considering relevance to our work, we mainly introduce two aspects, i.e., Introduction to VQA and attention scheme.

2.1. Introduction to VQA

There are two main difficulties in the task of VQA: one is the fusion of different modes between image and natural languages, the other is the relationship reasoning between the subjects involved in the problem. In fact, VQA, image retrieval [6], image classification [7,8], image ranking [9] and other tasks all need relational reasoning. According to the different ideas of works, we roughly divide current methods into three categories, i.e., joint embedding, attention mechanism, and relational reasoning.

Joint Embedding based Methods. Methods based on joint embedding mainly focuses on the fusion of the learned features of image and question embedding. After obtaining the semantic information after fusing, methods can successfully infer the answer of the question based on the fused information.

Early, Zhou et al. [10] propose "ibowimg", which firstly extracts semantic features by bag of words model and visual features by GoogleNet, and then combines both features to directly predict the answer of the question. Simple splicing of the extracted features of different modal inputs may lead to the lack of problem guidance in the extraction of visual features, resulting in the low correlation between the extracted information and the problem. To better extract the semantic information related to the question, Malinowski et al. [11] propose Neural-Image-QA, which firstly extracts the image features through the pre trained CNN, and then uses LSTM to process the problem information. Meanwhile, their proposed method takes the image features together as the input to participate in each step of the problem feature extraction. However, their proposed method simply fuses results of different modalities, which leads to the remaining of redundant visual information that irrelevant to the question.

Later, bilinear model is adopted to complete task of multi-modal information fusion, which helps to learn the high-level relationship between the semantic information of the question and the visual information of the picture. Based on the idea of bilinear model, Ben-younes et al. [12] propose Mutan, which is a tucker decomposition based on multi-modal tensor and parameterizes the bilinear interaction between visual and text representation. In addition to the Tucker framework, the core tensor is constrained by the sparsity of the structure to further reduce the parameters of the model, which is added to the training process as a regularization method to prevent overfitting.

Attention Mechanism based Methods. The introduction of attention mechanism to implement in VQA has brought a breakthrough for multi-modal system, which is usually used to integrate the information of question into the process of visual feature extraction or processing or extract the common features between two different modes through attention mechanism. We category current attention based methods into two types, i.e., spatial attention and co-attention mechanism.

Methods based on spatial attention mechanism generally guide extraction of image spatial information by problem features or joint features of problem and image. Early, Xu et al. [13] propose a memory network based on a specially designed problem-oriented spatial attention mechanism. Their proposed network firstly captures the fine-grained mapping relationship between the problem and the image space by embedding each word, and then selects the region related to the problem. Later, Shih et al. [14] firstly extracts the regions of input image by using pre-training network for edge detection, and then removes the duplicate region by using the non-maximum suppression. Afterwards, set of language items and image regions are embedded into a hidden space, in which a correlation weight is generated for each region by inner product. Finally, visual features of regions and text feature are combined to generate weighted and fused feature, thus computing final answers. Inspired by the framework of faster R-CNN, Anderson et al. [15] propose a bottom-up attention mechanism to extract the regions containing objects from the images. Meanwhile, the problem-oriented attention mechanism is used to assign weights to the regions containing subjects from top to bottom. The higher the weight, the greater the correlation between the region and the problem, thus successfully reducing the interference of redundant information.

Spatial attention mechanism focuses on the visual information processing of images, and extracts the most important local information. However, VQA requires to fully understand the problem semantics rather than only understanding the visual content. Some researchers have proposed methods based on co-attention mechanism [16], which

can learn the text attention of the problem and the visual attention of the image at the same time. Later, Yu et al. [17] set up two basic attention units, where one is self attention unit to interact within modes and extracts important information, and the other one is guided attention unit to interact between modes. Finally, MCAN (Modular Co-Attention Network) is formed for better answering by connecting two different attention units via collaborative attention module.

Relational Reasoning based Methods. The reasoning of the problem may involve the relationship among quantity of visual objects. Therefore, relational reasoning plays an important role in visual understanding and text understanding, especially when the position and size of objects in the picture are compared. Only understanding semantical meanings of each region cannot get such comparing based relationship. It is necessary to combine multiple regions to obtain such relationship information.

Early, Johnson et al. [18] propose a program generator and an execution engine, where the former unit constructs an explicit representation of the reasoning process, and the other unit executes the resulting program to produce an answer. However, such an explicit framework needs strong prior knowledge to train. Later, Perez et al. [19] introduce a general-purpose conditioning method called FiLM (Feature-wise Linear Modulation), which tries to involve neural network computation via a question-related transformation. Following the trial, Hudson et al. [20] propose to decompose the question into a series of attention-based reasoning steps, where each step is answered by a novel scheme named as Memory, Attention and Composition (MAC). Afterwards, Cadene et al. [21] propose structure of MuRel cell and incorporate it to develop a full MuRel network. Their proposed MuRel cell not only represent interactions between question and image regions by a rich vectorial representation, but also model region relations with pairwise combinations. Inspired by the fact that brain function or cognition can be described as a global and dynamic integration of local neuronal connectivity, Wu et al. [22] propose a connective cognition network (CCN) to dynamically reorganize the visual neuron connectivity that is contextualized by the meaning of questions and answers. They first develop visual neuron connectivity to fully model correlations of visual content, and then fuse the sentence representation with that of visual neurons. Finally, they propose directional connectivity to infer answers or rationales.

There exists a new trend to solve VQA with the latest graph network, which has achieved remarkable results so far. For example, Hu et al. [1] propose Language-Conditioned Graph Networks (LCGN), in which each node represents an object, and each object is described by a context-aware representation. Their proposed network is simple in design and implementation, but achieves significatively effective results across multiple tasks and datasets. Li et al. [23] propose a Relation-aware Graph Attention Network (ReGAT), which learns question-adaptive relation representations by encoding each image into a graph and modeling multi-type inter-object relations via a graph attention mechanism. To automatically answer questions about movies, Han et al. [24] firstly introduce a new dataset called PlotGraphs as external knowledge, which contains massive graph-based information of movies. Then, they propose a novel VQA model containing two main parts, i.e., a layered memory network (LMN) and a plot graph representation network (PGRN). LMN is capable to represent frame-level and clip-level movie content by their designed word memory module and the adaptive subtitle memory module respectively, meanwhile PGRN can represent the semantic information and the relationships of the entire graph.

2.2. Attention scheme

Attention scheme is commonly used in visual tasks to bring feature enhancement, thus adopted for question-related information extraction in VQA [25]. Attention mechanism was introduced into multimodal systems to solve this problem. Question-guided attention can effectively reduce redundant information and highlight informative regions of the image at the same time. Meanwhile, when analyzing relationships among different regions, the number of subjects involved in a specific relationship is often uncertain in advance, so the reasoning of relationships needs to be considered from multiple different scales, which is ignored in previous works.

Early, Yang et al. [26] present a multi-layer stacked attention network to infer the answer progressively, which achieves question-guided attention on every region in the image. Later, Woo et al. [27] build Convolutional Block Attention Module (CBAM) as a lightweight attention scheme, which sequentially computes attention values along spatial and channel dimensions. Afterwards, they fuse attention and image feature map for automatical feature enhancement. Zhao et al. [28] propose end-to-end Recurrent Attention (RA) model for pedestrian recognition, which highlight the spatial property of generated feature map by involving strength of Recurrent Learning and Attention scheme. Experiments show that they successfully extract context relationships among attribute categories of pedestrians to achieve more accurate recognition results. Further emphasizing on the importance of receptive fields, Li et al. [29] design SKNet (Selective Kernel Network) to assign weights for both channel related information and size of convolution kernel. The methods above focused on extracting key regions of the image according to the question. They model "where to look" or "what words to listen to", ignoring relationships among these regions, which are also important in some questions.

3. Proposed model

In this section, we present the overall structure of our proposed scheme in Fig. 2, which consists of four modules: (a) Multimodal Feature Extraction, (b) Regional Attention Module, (c) Multiscale Relation Reasoning, and (d) Multimodal Fusion.

3.1. Multimodal feature extraction

We adopt different approaches to extract features from multimodal inputs, i.e., an image and a question. We construct a bottom-up attention module to extract feature set with size of $k \times 2048$, where each feature is represented as v_i and k refers to the number of detected regions. Essentially, the core of bottom-up attention module is Faster R-CNN to generate regions of interest and ResNet to extract distinguish feature description.

The input question is trimmed to a maximum of 14 words for computational efficiency. With pretrained $GloVe$ word embeddings, each word is initialized as a $300-$dimension feature vector. Afterwards, the sequence of word features is processed by a Recurrent Gated Unit, where the final state of output is considered as feature representation of input question q.

3.2. Regional attention module

As shown in Fig. 3, the proposed regional attention module is composed of two parts, where the region informative evaluation is implemented by the soft attention part, and the hard attention part is responsible to pick up informative regions based on informative weights computed by soft attention part. Therefore, the proposed soft attention part is designed to not only help reduce redundant regions performed by the later hard attention part, but also assign informative weights to features of different regions based on input question.

We extract question-related context information by a combination of soft and hard attention parts. The reason for such design lies in the fact that only adopting soft attention for feature enhancement still brings abundant information for further processing. Essentially, the core idea of our proposed method is to perform multiscale relation reasoning. In other words, further processing is supposed to try combinations of multiple regions in different scales. Adopting too many regions by weighting with only soft attention greatly increase the number of region candidates, thus offering a large amount of computation

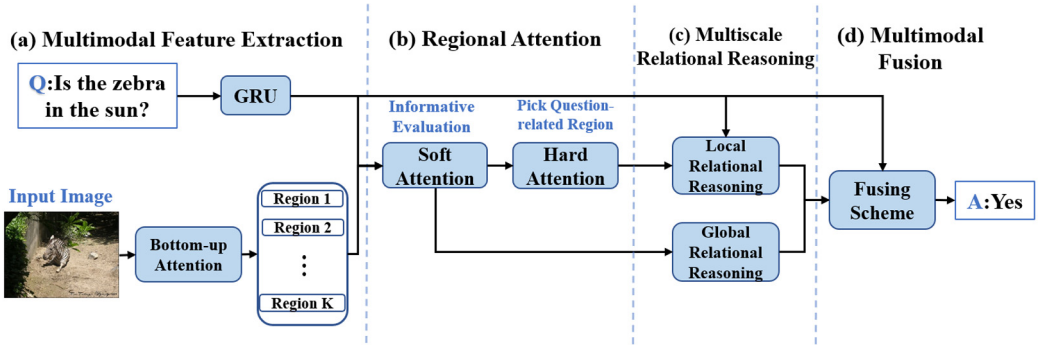

Fig. 2. Workflow of our proposed model to answer a question under guidance of an input image, i.e., "Is the zebra in the sun? Yes".

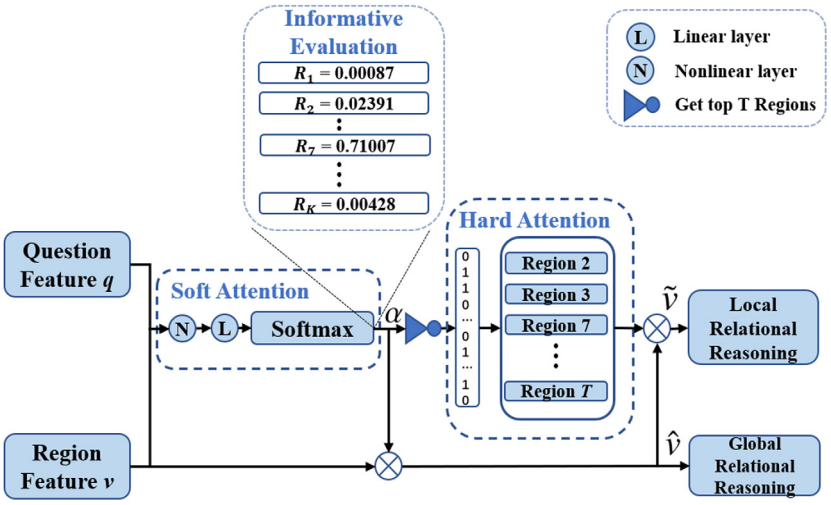

Fig. 3. Architecture design of the proposed Regional Attention module.

burden and even introducing noise information for accurate question-guided answering. Moreover, soft attention part offers an intuitive descriptor on the relevance between the question and each region. Meanwhile, hard attention part picks up informative regions to pursue high similarity between the input image and question. Both attention parts implicitly build a convinced alignment between the input image and question to help achieve question-guided answering.

Our proposed regional attention module is composed of soft attention and hard attention. The soft attention is dedicated to the evaluation of regional information, and the hard attention is to select information regions according to the evaluation information of the soft attention.

Specifically, we firstly construct region feature set $v = \{v_i, i = 1, \ldots, K\}$ for the extracted regions. Regarding question feature q and region feature v as input, we perform informative evaluation to obtain weight vector α with the proposed soft attention part:

$$\alpha = softmax(W_2 * (Relu(W_1 * ([v, q]) + b))) \tag{1}$$

where [] refers to concatenate operation, matrix W_1 and W_2 represent weight parameters in linear and nonlinear layer respectively, b is the bias vector, $Relu()$ and $softmax()$ represent ReLu and softmax functions respectively. After determining weight vector α, we compute weighted region feature $\hat{v} = \alpha \cdot v$, where $\cdot$ refers to element-wise multiplication. It is noted that $\hat{v}$ would be input for further global relational reasoning.

Meanwhile, we form question-related region feature set $\tilde{v}$ with T highest weight values in $\hat{v}$ with the proposed hard attention part:

$$\tilde{v} = f_s(\hat{v}, T) \tag{2}$$

where T is a hyper-parameter in experiments, function $f_s(\hat{v}, T)$ refers to operations of sorting $\hat{v}$ in descending order and choosing T feature vectors with highest value to construct output feature set. It is noted $\tilde{v}$ will be utilized for local relational reasoning in different scales, while the other feature vectors are discarded in later processing.

3.3. Multi-scale relational reasoning

In this subsection, we first emphasize the importance of global and local relation reasoning, where the global scheme tries to answer question implicitly by utilizing information of the whole image, and the local scheme models relationship among multiple objects for answers. Essentially, we believe both schemes contribute to answering questions by analyzing visual information from different aspects, which form the basic structure of the proposed relation reasoning architecture.

VQA task is complicated to solve especially when question is related with multiple objects, due to its dynamic embedding of objects corresponds to different questions [30]. For example, the questions of "What are the bears standing on" and "How many bears are on the ice" for the first image in Fig. 1 is and is not related to object "ice", even though the answering image is the same. To deal with flexibility brought by different questions, VQA inherently requires to perform multi-scale relation reasoning, thus owning comprehensively enough ability to answer questions. In other words, dynamic variation of question asks local relation reasoning scheme to prepare relationship descriptions of multiple objects ahead for low running time and fast feedback.

How to perform relation reasoning among objects have been widely discussed by researchers in VQA domain, where the general idea is to construct function by form of neural network to describe relationship [31,32]. We firstly define relation in the mth scale as relation by grouping m objects as one. In other words, we are supposed to construct

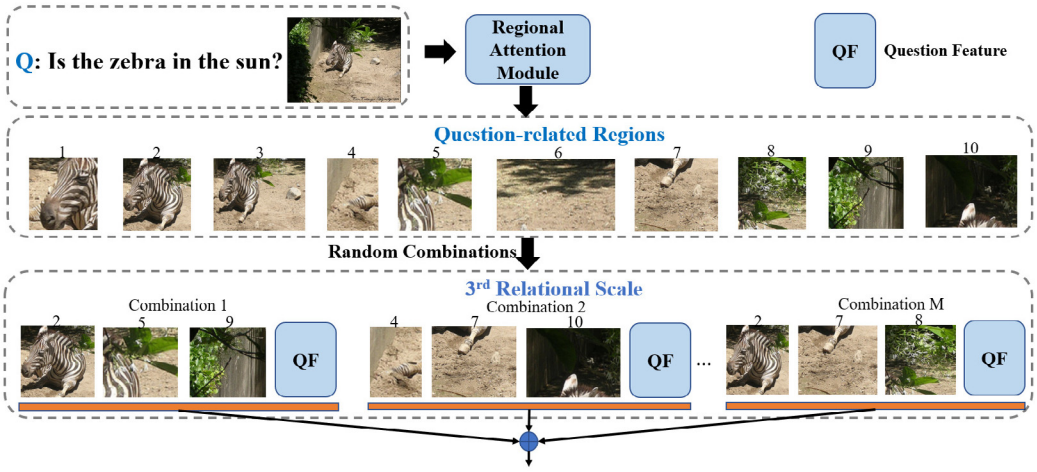

Fig. 4. An example of local relational reasoning scheme at the 3rd scale with M different combinations.

relationship descriptors for every m objects, thus performing relation reasoning in the mth scale. Then, we take $m = 2$ as an example and write function to perform relational reasoning with form of neural network as follows:

$$f_2(O) = h(\sum_{i,j} g(o_i, o_j)) \tag{3}$$

where the input O is a feature set corresponding to multiple objects represented as $O = \{o_1, o_2, \ldots, o_K\}$, K is the total number of objects, and function $g(), h()$ represent fully-connected layer of neural network.

Even though Eq. (3) is powerful in building relationship at the 2nd scale, expanding it to describe multi-scale relationship is still not easy, due to dynamic property of number of question-related objects. Instead of dynamically modeling relationship, we propose to construct as many as relationship descriptors in multi-scales, thus solving the complexity brought by unsettling objects and implicitly improving ability to reason relations between different object.

Recall the basic structure of the proposed relational reasoning architecture are global and local relational reasoning schemes, we define its output feature as

$$V = f_g(\hat{v}) + f_l(\tilde{v}, q) \tag{4}$$

where q is the constructed feature for input question, $\tilde{v}$ and $\hat{v}$ refer to question-related and weighted region feature set respectively, and functions $f_g()$ and $f_l()$ represent global and local relational reasoning schemes respectively.

Specifically, global relational reasoning scheme firstly sums all the weighted regional features, and then computes feature representation via a non-linear layer, which could be represented as

$$f_g(\hat{v}) = Relu(W_g * (\sum_{i=0}^{K} \hat{v}_i) + b_g) \tag{5}$$

where W_g and b_g refer to parameter matrix and bias vector.

To better explain how local relation reasoning scheme works, we show an example of relation reasoning at the 3rd scale in Fig. 4, where we firstly extract question-related regions by regional attention module, and then define the index of scale as the number of question-related regions in a combination. It is noted that three regions and question feature form a combination. Besides, we can find M combinations of regions are listed for computation, where M is a hyper-parameter and generally defined smaller than all possible combinations, thus saving huge computation cost in running time.

Defining number of scales S as a hyperparameter to be settled by user, the proposed local relation reasoning scheme with total S scales can be represented as

$$f_l(\tilde{v}, q) = R_1(\tilde{v}, q) + R_2(\tilde{v}, q) + \cdots + R_S(\tilde{v}, q) \tag{6}$$

where function $R_s()$ represents the relational reasoning at sth scale.

Essentially, the relational reasoning at sth scale could form $N = C_K^s$ combinations, where C represents combination function. To build up these combinations, we adopt random selection to group question-related regions with M combinations, where M is smaller than N. The final output for relational reasoning at sth scale thus could be represented as

$$R_s(\tilde{v}, q) = r(c_2, q) + r(c_3, q) + \cdots + r(c_M, q) \tag{7}$$

where c_m is the ith possible combination at sth scale, each relation term represented as function $r(c_m, q)$ captures relationships between m ordered regions and can be constructed by a nonlinear layer following with a linear layer to extract relationships of regions:

$$r(c_m, q) = W_{c,2} * (Relu(W_{c,1} * ([c_m, q]) + b_{c,1})) + b_{c_2} \tag{8}$$

where $W_{c,2}$ and $W_{c,1}$ refer to parameter matrix, $b_{c,1}$ and $b_{c,2}$ refer to bias vectors.

3.4. Multimodal fusion

As shown in Fig. 2, the process of multimodal fusion has indeed happened across the whole workflow, which can be divided into three phases. In region attention module, we design process of informative evaluation with question-guided idea, where different questions for the same image could generate different informative estimations. In other words, we involve textual information of question to help extract question-related context part of visual image information. During local relational reasoning scheme, each possible combination in different scales is grouped with question feature, which introduces textual information to handle scalable feature modeling.

Last but not least, a fusing scheme is designed at the end of the workflow, which fuse feature output by relational reasoning scheme V and question feature q by Hadamard product to output final feature descriptor. After fusing, the result feature is processed by a nonlinear layer following with a linear layer to generate answering for the question A, which can be represented as:

$$A = softmax(Relu(V \circ q)) \tag{9}$$

where $\circ$ refers to Hadamard product.

4. Experiments

In this section, we firstly introduce dataset and measurement. Then, we conduct multiple sets of parameter setting and ablation experiments to evaluate sensitivity to parameters and impact of different structure designs, respectively. Finally, we compare the proposed method with the existing methods to demonstrate efficiency.

Table 1
Comparisons on performance with different parameter settings tested on VQA v2 dataset.

Parameter settings	All	Yes/no	Numbers	Other
T: 5 regions	63.92	81.6	44.08	55.73
T: 7 regions	63.99	81.35	44.69	55.89
T: 10 regions	**64.11**	**82.18**	**44.73**	**55.91**
T: 15 regions	63.94	81.43	44.41	55.81
T: 20 regions	63.99	81.38	44.62	55.89
S: 3 scales	63.86	81.55	43.23	55.88
S: 4 scales	64.06	81.67	44.53	55.74
S: 5 scales	**64.11**	**82.18**	**44.73**	**55.91**
S: 6 scales	64.02	81.23	44.69	55.53
S: 7 scales	63.93	81.93	44.41	55.42
S: 8 scales	63.94	81.99	44.37	55.39
S: 9 scales	63.95	81.51	44.08	55.86
M: 1 combination	63.89	81.52	44.64	55.56
M: 2 combinations	64.03	82.10	44.70	55.41
M: 3 combinations	**64.11**	**82.18**	**44.73**	**55.91**
M: 4 combinations	64.02	82.05	44.54	55.67
M: 5 combinations	64.04	82.23	44.50	55.37

4.1. Dataset and measurement

We adopt MSCOCO VQA V1 and VQA V2 as dataset, which both consist of a training, a validation and an online testing set. Essentially, VQA v2 contains $440k$ question–answer pairs for training and $214k$ pairs for validation, which is carefully designed in number of training and testing samples to decrease the possibility of being overfitting. We evaluate performance by VQA accuracy:

$$Acc(ans) = min(1, \frac{\#humans\ that\ said\ ans}{3}) \tag{10}$$

where the accuracy is 1 only if the predicted answer appears no less than 3 times in human labeled answer list.

4.2. Parameter setting experiments

As shown in Table 1, we perform multiple experiments to determine values for hyper-parameters, where T, S and M refer to the number of chosen regions in hard attention module, total scale number, and total combination number in multiscale reasoning, respectively. It' noted that we fix two parameters with best choice at first, and then testing performance with varying another parameter.

Specifically, small value of T leads to insufficient information passing through hard attention module, thus failing to generate correct answers. Meanwhile, larger value of T helps better describe global relationships with more inputs. However, too large setting of T could largely weaken local relationships, thus resulting in lower overall and single item performance. In other words, most VQA questions are related with and could be answered by modeling of local relationship, rather than description of global relationship. Moreover, larger T brings much more computation burden for the proposed method. Therefore, we define $T = 10$ after quantity of experiments.

We can also observer that more scales does not contribute to better performance in Table 1. The reason lies in the inherent property of VQA tasks, where VQA generally involves a limited number of objects for answering. Therefore, larger S is not necessary for better performance. Meanwhile, few scales with smaller S fails to meet the information needs of different VQA tasks, where the corresponding questions should be answered under guidance of local relationship modeling in multiscale. Based on experiments, we define $S = 5$ for best performance. For each scale, inferring all possible combinations of informative regions will bring huge computing burden. We thus define $M = 3$ to keep balance between performance and computation cost.

Table 2
Comparisons on performance with different structure designs tested on VQA v2 dataset, where MF, HA, SA and MR represent designs of Multimodal Fusion, Hard Attention, Soft Attention and Multiscale Reasoning, respectively.

Structure designs	Overall	Yes/no	Numbers	Other
Proposed Method	**64.11**	**82.18**	**44.73**	**55.91**
Without MF	63.92	81.96	44.48	55.60
Without HA and MR	63.15	80.07	42.87	55.81
Without SA	63.35	80.84	43.01	55.27
Without HA and SA	62.78	79.98	42.16	54.79
Without HA				
Random 3 regions	63.60	81.34	43.29	55.48
Random 5 regions	63.67	81.61	42.90	55.56
Without MR				
With the 2nd scale	63.81	81.35	43.47	55.85
With the 3rd scale	63.84	81.37	43.64	55.85
With the 4th scale	63.94	81.4	44.46	55.81
With the 5th scale	63.95	81.41	44.47	55.83

4.3. Ablation experiments

To show efficiency of the proposed designs, we perform four sets of ablation experiments as shown in Table 2. Specifically, we replace hard attention module by randomly selecting several regions for comparisons. Meanwhile, we utilize operation of pre-defined single scale to compare with multiscale reasoning design. It is noted that all testings are carried out with parameter sets chosen in Table 1 if applicable.

By utilizing multimodal fusion scheme, the proposed method involves question embedding information to infer relationship among regions, which results in not only question-related feature, but also higher overall and single item performance. Without both hard attention and multiscale reasoning, the proposed method degenerates to utilizing soft attention for informativeness evaluation only. However, simple weighting scheme could bring too complex information to further process, thus largely decreasing performance. In fact, we believe structural information with regions better fits with problem of VQA, since questions of VQA are mostly related with several informative regions rather than the whole image.

As shown in Table 2, randomly selecting regions largely decreases performance of the proposed method, which reflects the importance of regional attention scheme to choose proper and informative regions. Furthermore, irrelative regions could even bring noise for further processing. Due to the multiscale nature of VQA questions, modeling local relationship with single scale could not guarantee the question can be answered, especially for "Numbers" category. Therefore, we can observe large decrease in "Numbers" category with 2nd and 3rd scale. With more objects are grouped in 4th and 5th scale, question in "Numbers" category answered with small value could be answered. However, several complicated questions involved multiple objects can still not be answered, thus resulting with a small decrease in performance.

In addition, we conducted ablation experiments on different attention operations. As shown in Table 2, when HA or SA is removed, the performance of the proposed method decreases to varying degrees. It means that all different attention operations play a positive role in our proposed method.

4.4. Performance comparison experiments

We compare our method with existing methods on VQA V1 and VQA V2 datasets. For fairness, all methods are trained with MSCOCO VQA train+val, and tested with both test-dev and test-standard, where these two different testing sets are designed to prevent overfitting performance in VQA challenge 2020.

Table 3 shows that the proposed method on VQA v2 dataset achieves best performance in overall and each single testing category, except for "Other" category in test-std. All these facts prove the efficiency

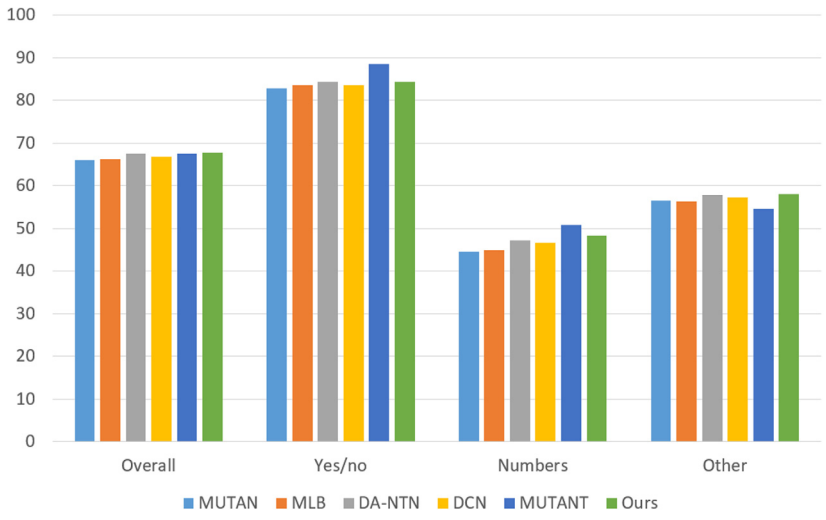

Fig. 5. Performance comparison between the proposed method and existing methods on VQA v2 dataset.

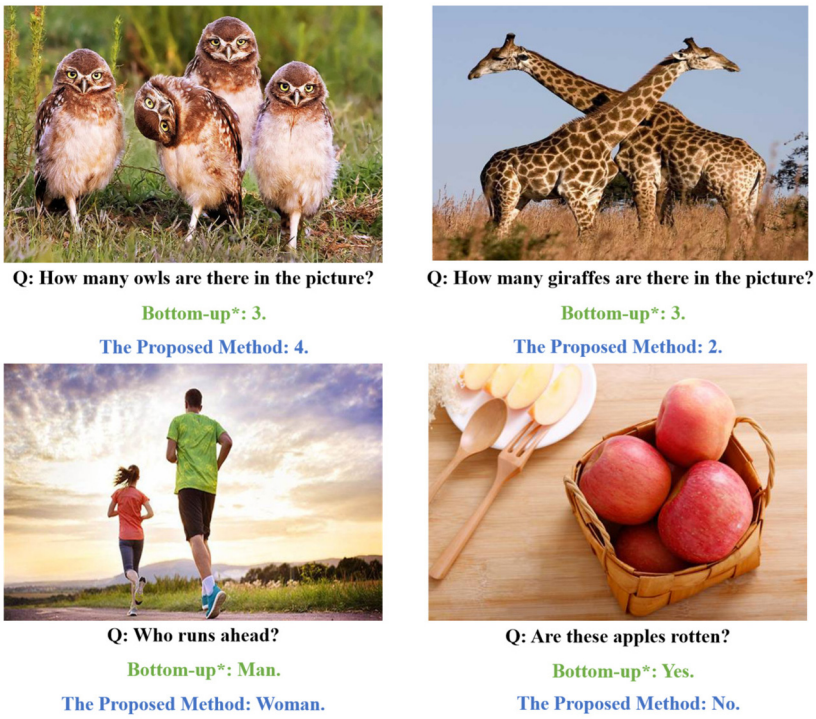

Fig. 6. Several VQA examples in MSCOCO VQA v2 dataset achieved by the proposed method and Adelaide Model+detector* [33].

Table 3
Performance comparison between the proposed method and the existing methods on VQA v2 dataset. It is noted that * represents that the corresponding method is trained with external datasets. † have been trained by [34] and [33] presented the details of [15] on VQA task.

Method	VQA v2 *test-dev*				VQA v2 *test-std*			
	Overall	Yes/no	Numbers	Other	Overall	Yes/no	Numbers	Other
VQA team-Prior [5]	–	–	–	–	25.98	61.20	00.36	01.07
VQA team-Language only [5]	–	–	–	–	44.26	67.01	31.55	27.37
MUTAN† [35]	66.01	82.88	44.54	56.5	66.38	83.06	44.28	56.91
MLB† [36]	66.27	83.58	44.92	56.34	66.62	83.96	44.77	56.52
Bottom-up* [33]	65.32	81.82	44.21	57.10	65.67	82.20	43.90	56.26
Graph Att. [37]	–	–	–	–	65.77	82.39	45.77	56.14
DA-NTN [34]	67.56	84.29	47.14	57.92	67.94	84.60	47.13	**58.20**
DCN [38]	66.87	83.51	46.61	57.26	67.04	83.85	47.19	56.95
RUbi [39]	64.75	–	–	–	–	–	–	–
MUTANT [40]	67.63	**88.56**	**50.76**	54.56	–	–	–	–
Improved Performance	0.16	–	–	0.17	0.11	0.07	0.95	–
The Proposed Method	**67.79**	84.33	48.28	**58.09**	**68.05**	**84.67**	**48.14**	58.15

Table 4
Performance comparison between the proposed method and the existing methods on VQA v1 dataset.

Method	VQA v1 *test-dev*				VQA v1 *test-std*			
	Overall	Yes/no	Numbers	Other	Overall	Yes/no	Numbers	Other
MUTAN [35]	65.7	83.3	39.7	56.6	65.8	83.2	40.3	56.4
MLB [36]	–	–	–	–	65.07	84.02	37.09	54.77
DA-NTN [34]	67.9	85.8	41.9	58.6	68.1	85.8	42.5	58.5
DCN [38]	66.89	84.61	42.35	57.31	67.02	85.04	42.34	56.98
The Proposed Method	67.95	85.82	43.41	58.79	68.47	85.89	43.41	58.18

of the proposed method. With the help of regional attention module and multiscale reasoning design, the proposed method is capable to achieve more promising performance to deal with questions in category of "Numbers", compared with the improved performance on category of "Yes/no" and "Other". The reason lies in the idea of multiscale modeling is specially designed to describe relationship among multiple objects, where questions in "Numbers" is inherently difficulty with nature of local relational reasoning in multiple scales. Therefore, we believe the higher improved performance in "Numbers" category proves the successful description on multiscale nature of VQA task achieved by the proposed method.

Moreover, category of "Yes/no" is high in performance achieved by the latest methods, where we believe extracting high-level semantic meaning can contribute to large improvement in performance, rather than relationship modeling. Since category of "Other" is complicated in representation of questions, understanding questions in deeper thought could help improve the corresponding performance. Therefore, we believe we can perform better in "Other" category with similar thoughts borrowing from [34]. Above all, the proposed method is specially designed to describe multiscale nature of VQA, where experiments have shown multiscale modeling ability of the proposed method. From Fig. 5, we can compare the performance of these methods more intuitively.

Table 4 shows the comparison results between our method and the existing methods on VQA v1. Our method can keep the best performance both on test-dev and test-standard. To qualitatively evaluate the proposed method, we show several VQA examples in Fig. 6, where these examples correspond to category of "Number", "Number", "Other" and "Yes/No", respectively. We could observe the proposed method achieves correct results in different categories of questions, due to nature of regional attention to generate context feature and multiscale design to describe local relationship among multiple objects.

4.5. Implementation details

All experiments are carried out on a single NVIDIA 1080Ti card with Intel E5-2620 CPU (2.60 GHz) and 250 GB RAM. To be fair, we adopt the same extracted feature for all experiments, where the number of detected regions k is defined as 36. The dimension of final question feature is set as 2048 for the convenience of fusion between question and image features. We use Adamax optimizer for training, where we set the initial learning rate as $2e^{-3}$ and batch size as 512. We use dropout and early stopping to avoid overfitting.

5. Conclusion

In this paper, we propose to utilize multi-scales relation reasoning for multimodal Visual Question Answering, where we specially design a novel regional attention scheme to help extract informative and question-related regions of interest. We validate the proposed method on VQA v2 dataset and show its efficiency by comparing with several latest works. With the development of edge or cloud system [41,42], we think the proposed method is promising to be applied in real-life with fast user feedback. Our future work is to further develop the proposed method to involve high-level semantic meaning from both images and questions, thus better solving complicated VQA tasks.

CRediT authorship contribution statement

Yirui Wu: Supervision, Investigation, Visualization, Writing, Editing. **Yuntao Ma:** Software, Algorithms, Investigation. **Shaohua Wan:** Visualization, Writing, Editing, Revision.

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.

Acknowledgments

This work was supported by National Key R&D Program of China under Grant 2018YFC0407901, the Fundamental Research Funds for the Central Universities, China under Grant B200202177, the Natural Science Foundation of China under Grant 61702160, the Natural Science Foundation of Jiangsu Province, China under Grant BK20170892.

References

[1] R. Hu, A. Rohrbach, T. Darrell, K. Saenko, Language-conditioned graph networks for relational reasoning, in: Proceedings of International Conference on Computer Vision, 2019, pp. 10293–10302.
[2] Y. Xi, Y. Zhang, S. Ding, S. Wan, Visual question answering model based on visual relationship detection, Signal Process., Image Commun. 80 (2020).
[3] L. Xu, H. Huang, J. Liu, SUTD-TrafficQA: A question answering benchmark and an efficient network for video reasoning over traffic events, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021.
[4] Z. Gao, Y. Li, S. Wan, Exploring deep learning for view-based 3D model retrieval, ACM Trans. Multim. Comput. Commun. Appl. 16 (1) (2020) 18:1–18:21.
[5] 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: Proceedings of Computer Vision and Pattern Recognition, 2017, pp. 6325–6334.
[6] J. Yu, Y. Rui, D. Tao, Click prediction for web image reranking using multimodal sparse coding, IEEE Trans. Image Process. 23 (5) (2014) 2019–2032.
[7] J. Yu, Y. Rui, Y.Y. Tang, D. Tao, High-order distance-based multiview stochastic learning in image classification, IEEE Trans. Cybern. 44 (12) (2014) 2431–2442.
[8] Y. Zhao, H. Li, S. Wan, A. Sekuboyina, X. Hu, G. Tetteh, M. Piraud, B. Menze, Knowledge-aided convolutional neural network for small organ segmentation, IEEE J. Biomed. Health Inform. 23 (4) (2019) 1363–1373.
[9] J. Yu, X. Yang, F. Gao, D. Tao, Deep multimodal distance metric learning using click constraints for image ranking, IEEE Trans. Cybern. 47 (12) (2017) 4014–4024.
[10] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, R. Fergus, Simple baseline for visual question answering, 2015, CoRR arXiv:abs/1512.02167.
[11] M. Malinowski, M. Rohrbach, M. Fritz, Ask your neurons: A neural-based approach to answering questions about images, in: Proceedings of IEEE International Conference on Computer Vision, 2015, pp. 1–9.
[12] H. Ben-younes, R. Cadène, M. Cord, N. Thome, MUTAN: multimodal tucker fusion for visual question answering, in: Proceedings of IEEE International Conference on Computer Vision, 2017, pp. 2631–2639.
[13] H. Xu, K. Saenko, Ask, attend and answer: Exploring question-guided spatial attention for visual question answering, in: Proceedings of European Conference on Computer Vision, vol. 9911, 2016, pp. 451–466.
[14] K.J. Shih, S. Singh, D. Hoiem, Where to look: Focus regions for visual question answering, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4613–4621.
[15] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6077–6086.

[16] J. Kim, J. Jun, B. Zhang, Bilinear attention networks, in: Proceedings of Neural Information Processing Systems, 2018, pp. 1571–1581.
[17] Z. Yu, J. Yu, Y. Cui, D. Tao, Q. Tian, Deep modular co-attention networks for visual question answering, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6281–6290.
[18] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C.L. Zitnick, R.B. Girshick, Inferring and executing programs for visual reasoning, in: Proceedings of International Conference on Computer Vision, 2017, pp. 3008–3017.
[19] E. Perez, F. Strub, H. de Vries, V. Dumoulin, A.C. Courville, FiLM: Visual reasoning with a general conditioning layer, in: Proceedings of AAAI Conference on Artificial Intelligence, 2018, pp. 3942–3951.
[20] D.A. Hudson, C.D. Manning, Compositional attention networks for machine reasoning, in: Proceedings of International Conference on Learning Representations, 2018.
[21] R. Cadène, H. Ben-younes, M. Cord, N. Thome, MUREL: Multimodal relational reasoning for visual question answering, in: Proceedings of Computer Vision and Pattern Recognition, 2019, pp. 1989–1998.
[22] A. Wu, L. Zhu, Y. Han, Y. Yang, Connective cognition network for directional visual commonsense reasoning, in: Proceedings of Advances in Neural Information Processing Systems, 2019, pp. 5670–5680.
[23] L. Li, Z. Gan, Y. Cheng, J. Liu, Relation-aware graph attention network for visual question answering, in: Proceedings of International Conference on Computer Vision, 2019, pp. 10312–10321.
[24] Y. Han, B. Wang, R. Hong, F. Wu, Movie question answering via textual memory and plot graph, IEEE Trans. Circuits Syst. Video Technol. 30 (3) (2020) 875–887.
[25] S. Wan, Y. Xia, L. Qi, Y. Yang, M. Atiquzzaman, Automated colorization of a grayscale image with seed points propagation, IEEE Trans. Multimedia 22 (7) (2020) 1756–1768.
[26] Z. Yang, X. He, J. Gao, L. Deng, A.J. Smola, Stacked attention networks for image question answering, in: Proceedings of Computer Vision and Pattern Recognition, 2016, pp. 21–29.
[27] S. Woo, J. Park, J.-Y. Lee, I. So Kweon, Cbam: Convolutional block attention module, in; Proceedings of European Conference on Computer Vision, 2018, pp. 3–19.
[28] X. Zhao, L. Sang, G. Ding, J. Han, N. Di, C. Yan, Recurrent attention model for pedestrian attribute recognition, in: Proceedings of AAAI Conference on Artificial Intelligence, 2019, pp. 9275–9282.
[29] X. Li, W. Wang, X. Hu, J. Yang, Selective kernel networks, in: Proceedings of Computer Vision and Pattern Recognition, 2019, pp. 510–519.
[30] S. Ding, S. Qu, Y. Xi, S. Wan, Stimulus-driven and concept-driven analysis for image caption generation, Neurocomputing 398 (2020) 520–530.
[31] B. Zhou, A. Andonian, A. Oliva, A. Torralba, Temporal relational reasoning in videos, in: Proceedings of European Conference on Computer Vision, 2018, pp. 831–846.
[32] L. Li, F. Zhu, H. Sun, Y. Hu, Y. Yang, D. Jin, Multi-source information fusion and deep-learning-based characteristics measurement for exploring the effects of peer engagement on stock price synchronicity, Inf. Fusion 69 (2021) 1–21.
[33] D. Teney, P. Anderson, X. He, A. van den Hengel, Tips and tricks for visual question answering: Learnings from the 2017 challenge, in: Proceedings of Computer Vision and Pattern Recognition, 2018, pp. 4223–4232.
[34] Y. Bai, J. Fu, T. Zhao, T. Mei, Deep attention neural tensor network for visual question answering, in: Proceedings of European Conference on Computer Vision, vol. 11216, 2018, pp. 21–37.
[35] H. Ben-younes, R. Cadène, M. Cord, N. Thome, MUTAN: Multimodal tucker fusion for visual question answering, in: Proceedings of International Conference on Computer Vision, 2017, pp. 2631–2639.
[36] J. Kim, K.W. On, W. Lim, J. Kim, J. Ha, B. Zhang, Hadamard product for low-rank bilinear pooling, in: Proceedings of the International Conference on Learning Representations, 2017.
[37] W. Norcliffe-Brown, S. Vafeias, S. Parisot, Learning conditioned graph structures for interpretable visual question answering, in: Proceeding of Neural Information Processing Systems, 2018, pp. 8344–8353.
[38] D. Nguyen, T. Okatani, Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering, in: Proceedings of Computer Vision and Pattern Recognition, 2018, pp. 6087–6096.
[39] R. Cadène, C. Dancette, H. Ben-younes, M. Cord, D. Parikh, RUBi: Reducing unimodal biases for visual question answering, in: Proceedings of Neural Information Processing Systems, 2019, pp. 839–850.
[40] T. Gokhale, P. Banerjee, C. Baral, Y. Yang, MUTANT: A training paradigm for out-of-distribution generalization in visual question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 878–892.
[41] X. Xu, B. Shen, X. Yin, M.R. Khosravi, H. Wu, L. Qi, S. Wan, Edge server quantification and placement for offloading social media services in industrial cognitive iov, IEEE Trans. Ind. Informatics 17 (4) (2021) 2910–2918.
[42] X. Xu, Q. Wu, L. Qi, W. Dou, S. Tsai, M.Z.A. Bhuiyan, Trust-aware service offloading for video surveillance in edge computing enabled internet of vehicles, IEEE Trans. Intell. Transp. Syst. 22 (3) (2021) 1787–1796.