2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Multi-scale Relational Reasoning with Regional Attention for Visual Question Answering

1st Yuntao Ma
National Key Lab for Novel Software Technology, Nanjing University
Nanjing, China
Email: mayuntao10@126.com

2rd Tong Lu *
National Key Lab for Novel Software Technology, Nanjing University
Nanjing, China
Email: lutong@nju.edu.cn

3nd Yirui Wu
College of Computer and Information Hohai University
Nanjing, China
Email: wuyirui@hhu.edu.cn

***Abstract*—One of the main challenges of visual question answering (VQA) lies in properly reasoning relations among visual regions involved in the question. In this paper, we propose a novel neural network to perform question-guided relational reasoning in multi-scales for visual question answering, in which each region of image is enhanced by regional attention.**

Specifically, we present regional attention module, which consists of a soft attention module and a hard attention module, to select informative regions of the image according to informative evaluations implemented by question-guided soft attention. Combinations of different informative regions are then concatenated with question embedding in different scales to capture relational information. Relational reasoning module can extract question-based relational information among regions, in which multi-scale mechanism gives it the ability to model scaled relationships with diversity making it sensitive to numbers. We conduct experiments to show that our proposed architecture is effective and achieves a new state-of-the-art on VQA v2.

***Index Terms*—Visual question learning, Attention, Multi-scale relational reasoning**

I. Introduction

Visual Question Answering (VQA) is to answer a natural language question about an image, which has been widely used in many fields, such as customer service, data management, robotics and so on. The main difficulty in VQA lies in the fact that VQA methods are requested to process information among multiple modes to generate correct answers, since the VQA task involves distinct types of input, such as visual and textual information.

Recently, CNN has achieved huge success in image classification [1] and other visual fields [2], which has also been introduced into VQA for extracting visual features. However, such powerful deep learning architecture is a visual features processor, but may not be well suited for relational reasoning, for that CNN focuses exclusively on processing local spatial structure. Differ from the usual tasks of image classification and detection, VQA task requires to not only extract the features of objects themselves, but also focuses on effective information that related to the question, which means that fusing visual and language information that in different modalities plays an important role in VQA.

Only recognizing informative regions of the image that related to the question is not enough for answering, we also need to reason relationships among different regions based on the question. As shown in Fig. I, which are examples of MSCOCO VQA v2 [3], we first need to understand questions and images respectively to find all the informative areas of the picture that related to the question. Besides, we also have to understand relationships among multiple objects based on the question, for that even if key areas are the same, the relations required for answering may not the same, such as the relationships ”standing on” and ”in the tree” in Fig. I.

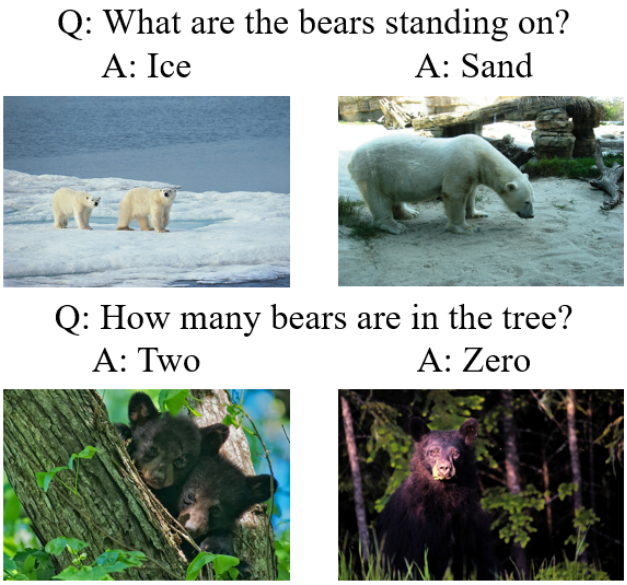

Fig. 1. We not only need to recognize objects, but also have to reason relationships among them. For example, the second picture in the second line has trees and a bear, however, the bear is not in the tree.

Thus, there are two main difficulties in VQA: multi-modalities and relational reasoning. Firstly, different input forms have different feature spaces, but features in different forms may represent the same thing. For example, the word ”bear” in the question means the region of bear in the pictures. And the question determines which parts of the image are important and which parts are redundant. Therefore, in the process of feature processing, two kinds of inputs need to be mapped into the same feature space. Secondly, some problems involve relationships among objects, which means

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9413140

that relationships need to be reasoned under the guidance of problems. It is also worth noticing that the number of regions involved in different relationships is different, which means that we need to analyze relationships in different scales.

The main task of multi-modality [4], [5] is to map feature spaces of different modalities to the same feature space. So that the information between different modalities can be transferred to each other, or they can be processed at the same time. Zhou et al. [6] proposed "iBOWIMG" to concatenate word features from the question and CNN features from the image to predict the answer. For the sake of better extracting semantic information, CNN+LSTM architectures [7], [8] are then proposed. However, such methods simply fused results of different modalities, which will lead to the remaining of redundant visual information that irrelevant to the question.

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.

In this work, we present a model with multimodal relational reasoning in multi-scales, in which regions are enhanced by regional attention, and achieve state-of-art results. The main contribution of this work can be summarized as follows:

1. We propose regional attention, which uses soft attention to evaluate the importance of regions in the image and uses hard attention to pick up informative regions for further relational reasoning.

2. We propose multi-scale relational reasoning, which combines the question information with visual information to carry out relational reasoning in different scales.

II. Related work

VQA is a challenging task receiving extensive attention while Malinowski et al. [9] proposed VQA as a visual turing testing. The main difficulties of VQA lie in multi-modality and relational reasoning and the previous work also mainly focused on these two aspects.

A. Attention

Attention mechanisms [10], [11] have been a breakthrough for multimodal systems, and are commonly used in VQA tasks to bring question information into visual feature extraction or processing. Xu et al. [12] proposed a memory network with spatial attention and use the question to choose relevant regions for answering. Yang et al. [13] presented a multi-layer stacked attention network(SAN) to infer the answer progressively. Such methods are top-down visual attention, which achieves question-guided attention on every region in the image. Anderson et al. [14] used bottom-up(Faster R-CNN based) to detect visual objects and weighted the detection boxes through question-guided top-down attention, highlighting key objects in the picture. Lu et al. [15] presented a co-attention model that joint reasons image and question attention.

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. In the example mentioned in I, ignoring the relationship between the bear and the tree will lead to wrong answers.

B. Relational reasoning

Relational reasoning is one of the most important tasks of visual understanding and the central component of general artificial intelligence. To answer the question about an image, relational reasoning is a very important ability. Johnson et al. [16] proposed a model consists of a program generator that constructs an explicit representation of the reasoning process, and an execution engine that executes the resulting program to produce an answer.

Such an explicit reasoning framework, however, needs strong prior to train it. Santoro et al. [17] proposed a simple neural network module that reasons over all the possible pairs of objects in the picture and proved the efficiency of implicitly visual reasoning without strong prior. Perez et al. [18] introduced a general-purpose conditioning method called FiLM: Feature-wise Linear Modulation, which influences neural network computation via a transformation based on the question. Hudson et al. [19] proposed a model that approaches problems by decomposing them into a series of attention-based reasoning steps, each performed by a novel Memory, Attention and Composition(MAC). However, such methods didn't take the scales of relational reasoning into consideration and didn't pay enough attention to important regions of the image.

Inspired by these previous works, we propose question-guided regional attention to evaluate the importance of regions, which only let informative regions passing through for local relational reasoning. Meanwhile, we design a novel multi-modal relational reasoning module built without strong priors for regions. It is composed of global relational reasoning and local relational reasoning, one for processing global relational reasoning and the other for local relational reasoning. Global relational reasoning uses all the regions detected from the image, while local relational reasoning only uses few of them but in different scales.

III. Proposed model

This section presents details of objects' relational reasoning model enhanced by multimodal fusion, which is based on informative estimation and relational reasoning. The network design of the proposed mothod, as is shown in Fig. 2, consists of four modules: (a) feature extraction, (b) regional attention, (c) objects relation reasoning, (d) multimodal fusion. For the sake of transparency, we describe the model with the specific structure and hyperparameters values of best performance. In section IV-B we will discuss the influence of structure

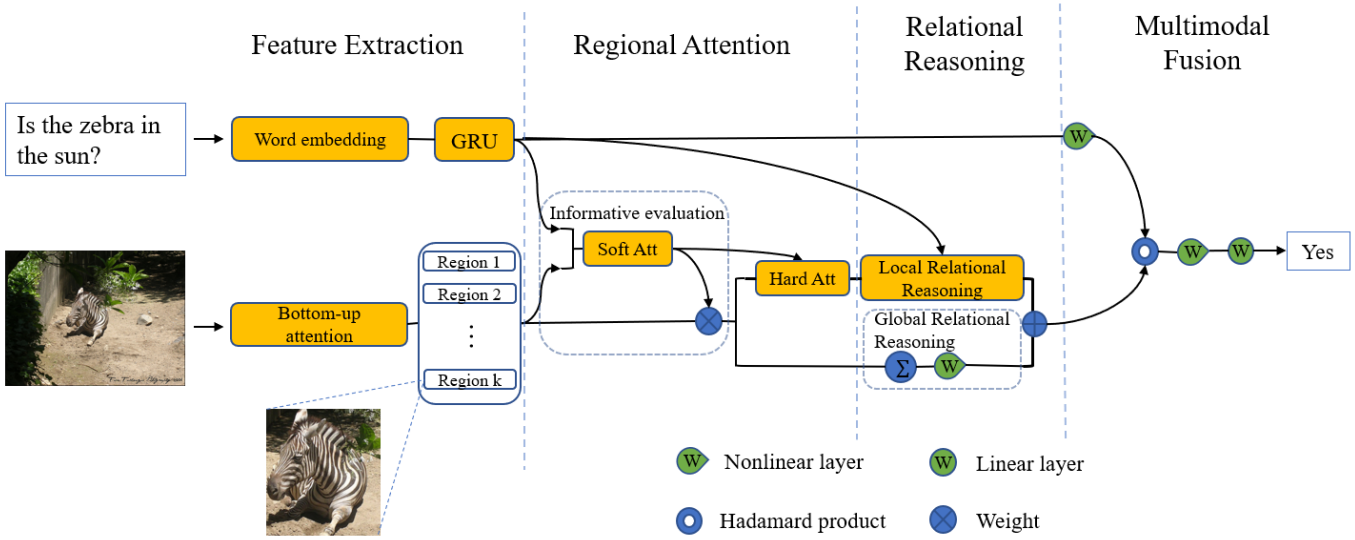

Fig. 2. Overview of our proposed model for the example question-answering pair: "Is the zebra in the sun? Yes.". Although we call the last phrase "Multimodal Fusion", the fusion of information in different modalities actually runs through three phrases: regional attention, Relational Reasoning and Multimodal Fusion. The process of Regional Attention and Local relational reasoning are illustrated in Fig. 3 and Fig. 5

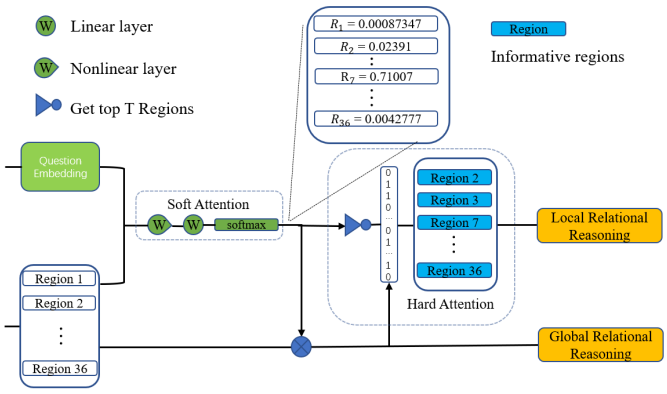

Fig. 3. The process of Regional Attention, which consists of soft attention and hard attention. The result of soft attention is obtained form the inference of the picture in Fig. 2 and its corresponding question, while Region 7 has the biggest informative evaluation as 0.71007.

and hyperparameters on performance and demonstrate the efficiency of our model.

A. Feature extraction

The input of VQA generally consists of two parts: an image and a text question, having essentially different forms. For the image, it is passed through a Resnet CNN within a Faster R-CNN framework, called bottom-up attention [14], to generate vector representations of $K \times 2048$, and K is a hyperparameter representing the number of regions extracted from an image. Each region of the image is encoded into a 2048-dimensional vector. All features here are pretrained and provided by [14]. We take the vector of a specific region as v_i and the stacking of vectors as v.

The questions are trimmed to a maximum of 14 words for computational efficiency with the extra words discarded. Each word is turned into $300-$dimension vectors and initialized with pretrained $GloVe$ word embeddings [20]. The sequence of word embeddings is then passed through a Recurrent Gated Unit [21]. We take the final state after processing the word embeddings as q.

B. Regional attention

The regional attention is mainly composed of two parts, one is the informative evaluation implemented by soft attention, while the other is hard attention for picking up informative regions. Informative evaluation (similar to top-down attention in [14]) is based on textual information and visual features. For K areas of the image, each area will have a normalized informative weight, and corresponding area features will be multiplied by the weight. Meanwhile, there are two parallel and different ways to utilize evaluation results. Firstly, the features of all locations will be fused to find overall relationship information of all regions. Secondly, informative areas, which can pass through the regional attention, will be taken as important areas of the image for local relationship information. The process of extracting global and local relationship information will be detailed in Section III-C.

To extract key information among different regions, we utilize soft attention to model information contained in the region that related to the question, which can be further used to not only to reduce redundant information of regions, but also evaluate the importance of regions for the corresponding

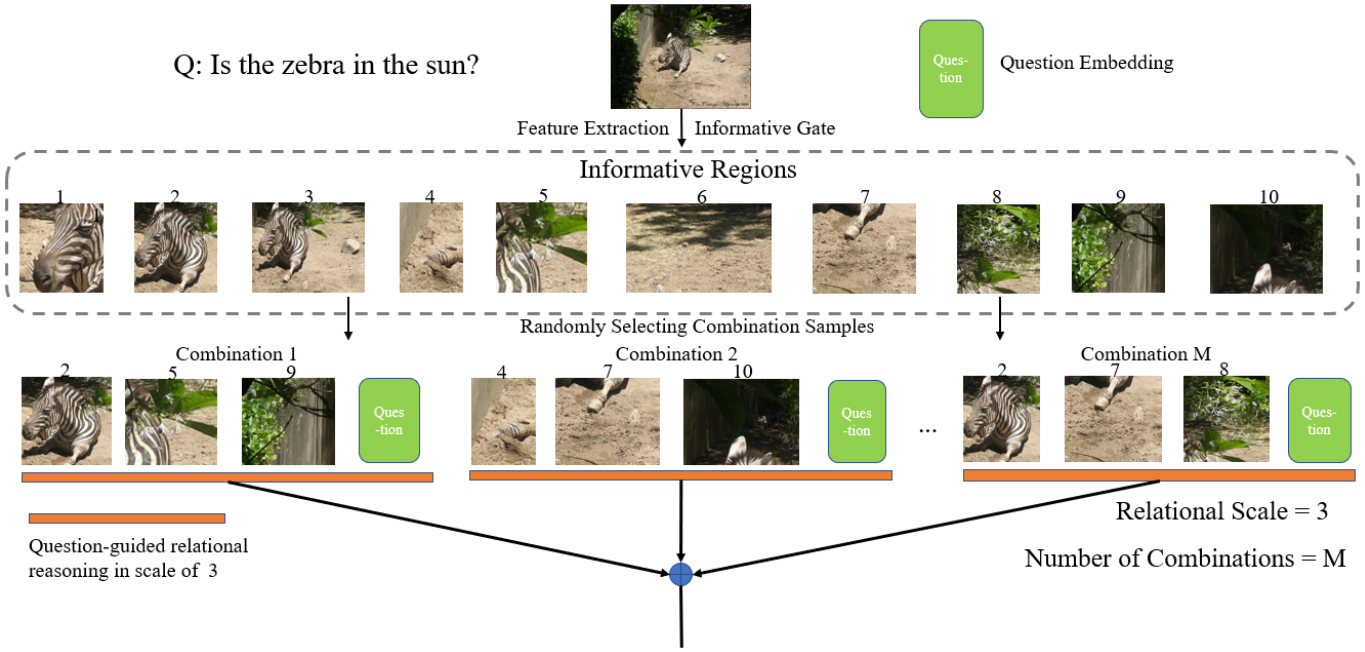

Fig. 4. An example of local relational reasoning module for an image and the corresponding question "Is the zebra in the sun?" in specific scale of 3 with M different combinations. The picture is passed through the feature extraction module and regional attention, generating T informative regions. The num of informative regions "T" here is 10. The different combinations are chosen randomly from all the combinations of informative regions and results of each combination are added up as the relational reasoning result in scale of 3. The results of different scales will be sumed up as local relational reasoning result.

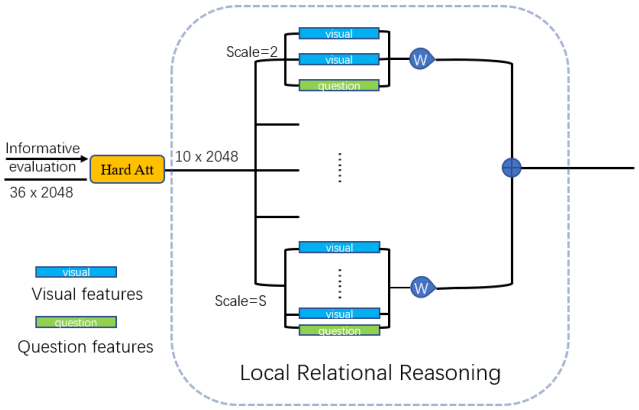

Fig. 5. The process of local multimodal relational reasoning, which has different scales. After passing through the regional attention, 36 regions of the image only left T, which we called informative regions. The local relationships will be reasoned in different scales, and each scale has different combinations of regions. The process in a specific scale can be seen in Fig. 4

question. In other words, soft attention scheme offers an intuitive descriptor on the relevance between the question and each region. Hard attention scheme picks up informative regions according to the result of the informative evaluation to pursue high similarity between the image and the question, which implicitly builds an alignment between them.

The input of informative evaluation is question-oriented. For each region $i = 1...K$ of the image, it is concatenated with the question embedding and sent into the informative evaluation. The informative evaluation passes the input through a nonlinear layer and then a linear layer to generate informative evaluations of each location. The values of evaluation are normalized through softmax to generate the final informative weight. The process of informative evaluation could be represented as

$$\tilde{y_i} = Relu(Wv_i + b) \tag{1}$$

$$g = Relu(W^{'}q + b^{'}) \tag{2}$$

$$y_i = \tilde{y_i} \circ g \tag{3}$$

$$Imp_i = W_i y_i \tag{4}$$

$$\alpha = softmax(Imp) \tag{5}$$

$$\hat{v} = \alpha v \tag{6}$$

where v_i, $q$are regional features and question embedding respectively, and $\circ$ is Hadamard product.

All the region features are weighted by the normalized informative values. Then features of all areas will be passed through the hard attention, which only lets the top T areas with high informative values passing. These areas will be utilized for local relational reasoning in different scales. The process of hard attention can be represented as

$$\alpha_{idx} = argmax[T](\alpha) \tag{7}$$

$$\beta_i = \begin{cases} 0, & \alpha_i \text{ not in } \alpha_{idx}, \\ 1, & \alpha_i \text{ in } \alpha_{idx}. \end{cases} \tag{8}$$

$$\hat{v}_{imp} = \beta \hat{v} \tag{9}$$

where $argmax[T](\alpha)$ means the T biggest informative evaluations of α, and only the top T regions will be passed through for local relational reasoning, while the others are discarded directly.

C. Multi-scale relational reasoning

The design philosophy of relational reasoning, as is also illustrated in the works of Santoro et al. [17] and Zhou et al. [22], is to constrain the functional form of the neural network. For example, the functional form relational reasoning in the scale of 2 can be represented as

$$f_2(O) = h_\phi(\sum_{i,j} g_\theta(o_i, o_j)) \tag{10}$$

where the input is a set of "objects" $o_1, o_2, ..., o_K$, and $r_1, r_2, ..., r_{s_k}$ means the combinations of objects in scale of s_k.

The capacity of relational reasoning is built-in to such functional form, for that such architecture takes mutiple objects into consideration, and the connection between different objects creates a relationship. Meanwhile, this form has to take different combinations into consideration, which implies that such form is not privy to the actual meaning of any particular relation, for that the relationships between different objects are different.

The multi-scale relation reasoning forces the model to consider their relations from different scales, which implicitly improves the model's ability to reason the relations between various objects, especially quantity related problems.

In our work, relational reasoning module consists of two parallel parts, one is used to extract global information, and the other is used to extract local relationships among important regions picked by hard attention. The final results of these two parts will be added, and the overall process can be represented as

$$V = f_{local}(\hat{v}_{imp}, q) + f_{global}(\hat{v}) \tag{11}$$

where $\hat{v}_{imp}$ is features of top T regions, and $\hat{v}$ is features of all regions. Both of them are weighted by the normalized informative values.

For global information, our model takes all regional features as sum, and then the sum is passed through a non-linear layer. The whole process is as follows.

$$f_{global}(\hat{v}) = Relu(W(\sum_{i=0}^{K} \alpha v_i) + b) \tag{12}$$

Local relational reasoning, as is shown in Fig. 5, is much more complicated than global relation reasoning. The scale of regions for relational reasoning required in different textual contexts is different. Some questions involve location information between multiple objects, while others may only be related to a small number of areas. Therefore, the relationship extraction of local features has many different scales. Besides, it is unnecessary to utilize all possible combinations of the T regions for relation reasoning which needs huge computing overhead. Therefore, for each scale, our method only samples different M combinations from all combinations to perform local relationship inference.

The number of scales is a hyperparameter, and the model with S scales can be represented as

$$f_{local}(\hat{v}_{imp}, q) = R_1(\hat{v}_{imp}, q) + R_2(\hat{v}_{imp}, q)... + R_S(\hat{v}_{imp}, q) \tag{13}$$

where R_i represents the relational reasoning of specific scale and $1, 2...S$ is just the number of scales instead of the number of regions for relation reasoning. The choices of the number of regions can be different even if the number scales is same. For example, $[10, 7, 5, 3]$ vers $[10, 5, 3, 2]$. They both have four scales, but their scales are different.

The relationship reasoning for each different scale is question-oriented, and the question embedding will concat into each relationship. Therefore, the relational reasoning of scale i actually consists of $i+1$ vectors. In addition, each relational inference of scale i randomly samples several combinations of i regions, and then performs relational reasoning with question embedding. The results of different combinations will be added. The process can be represented as

$$R_i(\hat{v}_{imp}, q) = r(C_1, q) + r(C_2, q)... + r(C_M, q) \tag{14}$$

where C_i is different combinations of the T vectors for the specific scale of R_i, and q is the question embedding. M is a hyperparameter that decides the number of combinations that sampled from vectors. $r(C_i, q)$ is a nonlinear layer following with a linear layer to extract relationships of regions in the combination, and q is concatenated with C_i, the process can be presentated as

$$r(C_i, q) = W'(Relu(W([C_i, q]) + b)) + b' \tag{15}$$

D. Multimodal fusion

Multimodal fusion, as is shown in Fig. 2, has three phrases. Firstly, the informative evaluation is question-guided. Different questions for the same image will generate different informative estimation, thus resulting in different regions passing through hard attention. Secondly, for local relational reasoning, each combination will concat with question embedding. The same combination of regions will have various relations under different questions. Last but not least, the results of relational reasoning and question embedding will be combined with Hadamard product. The combining process can be represented as

$$F = V \circ Q \tag{16}$$

where V is the sum of global and local relations and Q is the result of question embedding passing through a nonlinear layer. $\circ$ is Hadamard product. After combination, the result will pass through a nonlinear layer following with a linear layer as illustrated before.

TABLE I

DIFFERENT MODELS COMPARISON ON VALIDATION SPLIT OF VQA V2. RESULTS OF THE ABLATIVE EXPERIMENTS EVALUATED ON THE VQA V2 VALIDATION SET. EVERY ROW PRESENTS THE RESULTS OF THREE EXPERIMENTS WITH DIFFERENT RANDOM SEEDS, WHICH CHANGES ONLY ONE VARIABLE COMPARING TO THE PROPOSED MODEL AND THE FIRST ROW IS OUR PROPOSED MODEL. EACH BLOCK IS EXPERIMENTS FOR DIFFERENT VALUES OF THE SAME PARAMETERS.

	VQA v2 validation			
	All	Yes/no	Numbers	Other
Proposed model	64.07 ± 0.04	82.01	44.9	55.57
Without hard attention and local relational reasoning	63.15 ± 0.08	80.07	42.87	55.81
Hard attention				
Randomly select regions(1 combinations)	63.46 ± 0.05	81.43	42.35	55.42
Randomly select regions(2 combinations)	63.55 ± 0.03	81.43	43.13	55.32
Randomly select regions(3 combinations)	63.60 ± 0.09	81.34	43.29	55.48
Randomly select regions(4 combinations)	63.62 ± 0.02	81.37	43.61	55.42
Randomly select regions(5 combinations)	63.67 ± 0.05	81.61	42.9	55.56
Elasticity of Hard Attention				
5 regions (With scales of $[5, 4, 3, 2]$)	63.96 ± 0.07	81.67	44.59	55.66
7 regions (With scales of $[7, 5, 4, 3, 2]$)	63.97 ± 0.02	81.92	44.72	55.41
15 regions (With scales of $[15, 7, 5, 3, 2]$)	64.05 ± 0.04	82.03	44.73	55.48
20 regions (With scales of $[20, 7, 5, 3, 2]$)	64.03 ± 0.02	82.13	44.68	55.41
Multimodal fusion				
Without question embedding in relational reasoning	63.92 ± 0.08	81.96	44.48	55.6
Local relational reasoning scales				
With scales of $[10, 9, 8, 7, 6]$	63.92 ± 0.03	82.22	44.3	55.25
With scales of $[10, 5, 4, 3, 2]$	63.99 ± 0.02	81.92	44.66	55.45
With scales of $[10, 7, 4]$	63.98 ± 0.09	81.64	44.18	55.69
With scales of $[10, 7, 5, 3]$	63.93 ± 0.13	82.03	44.65	55.39
Local relational reasoning with M combinations				
With 1 combination	63.98 ± 0.05	81.51	45.19	55.65
With 2 combinations	63.99 ± 0.08	82.11	44.99	55.43
With 4 combinations	64.05 ± 0.06	82.16	44.65	55.67
With 5 combinations	64.08 ± 0.07	82.08	44.57	55.54

IV. EXPERIMENTS

We conducted multiple sets of experiments with different model structures and parameters to evaluate the impact of different structures for performance and the sensitivity of the model to the parameters. All results are shown in Table III-C.

A. Experimental setup

In this section, we present some experiments with alternative architecture and hyperparameters to compare with the proposed model and demonstrate effective and efficient of it. And then compare our model with other competing methods. Each experiment is the result of a single model trained independently. All models are trained on the MSCOCO VQA v2 training set.

For ablative experiments, each choice is trained three times with different random seeds. The results are reported on MSCOCO VQA v2 validation set at the best epoch and the performance is metriced by the standard VQA accuracy. The first row is our proposed model. In the feature extraction stage, Teney et al. [23]'s work indicates that fixed $K = 36$ will lead to better performance, and they demonstrate the effectiveness of bottom-up feature extraction for images. Thus we use bottom-up and corresponding parameters as feature extraction. At the same time, we do experiments on our innovative module: regional attention, relational reasoning, the elasticity of hard attention, the scales of local relational reasoning, and the sampling of each scale. Our proposed model samples three combinations for each scale of $[10, 7, 5, 3, 2]$ and let 10 regions pass through hard attention according to the question-guided informative evaluation.

TABLE II

THE COMPARISON BETWEEN STRATEGIES OF RANDAMLY SELECTING REGIONS AND HARD ATTENTION.

Number of Combinations	Hard Attention	Randomly select
With 1 combination	63.98±0.05	63.46±0.05
With 2 combinations	63.99±0.08	63.55±0.03
With 3 combinations	64.07±0.04	63.60±0.09
With 4 combinations	64.05±0.06	63.62±0.02
With 5 combinations	64.08±0.07	63.67±0.05

When comparing with other competing methods, we report the results of MSCOCO VQA v2 test-dev and MSCOCO VQA v2 test-standard, which are returned from the official VQA challenge 2020.

B. Ablative experiments

1) Regional Attention: The ultimate goal of regional attention is to evaluate the importance of regions and extract T important regions from K regions. We mainly do experiments on hard attention for that the effectiveness of soft attention has been improved by [14]. To control variables, we still evaluate the informative of regions, and features of regions also have been weighted by the informative evaluation.

Hard Attention. The contrast experiment selects regions randomly to pass through hard attention instead of referring to the informative evaluations. As can be seen in Table III-C, the strategy of randomly selecting regions makes the performance of the model decline sharply. As is shown in Table

TABLE III

PERFORMANCE COMPARISON BETWEEN THE PROPOSED METHOD AND THE EXISTING METHODS ON VQA V2 DATASET. RESULTS OF THE PROPOSED METHOD ALONG WITH OTHER PUBLISHED RESULTS ON VQA V2 *test-dev* AND *test-standard* SPLITS IN SIMILIAR CONDITIONS (I.E., A SINGLE MODEL; TRAINED WITHOUT EXTERNAL DATASET). *: TRAINED WITH EXTERNAL DATASETS.

	VQA v2 *test-dev*				VQA v2 *test-std*			
Method	All	Yes/no	Numbers	Other	All	Yes/no	Numbers	Other
VQA team-Prior [3]	-	-	-	-	25.98	61.20	00.36	01.07
VQA team-Language only [3]	-	-	-	-	44.26	67.01	31.55	27.37
VQA team-LSTM+CNN [3]	-	-	-	-	54.22	73.46	35.18	41.83
MF-SIG+VG [11]	64.73	81.29	42.99	55.55	-	-	-	-
Adelaide Model* [23]	62.07	79.20	39.46	52.62	62.27	79.32	39.77	52.59
Adelaide Model+detector*(Bottom-up) [23]	65.32	81.82	44.21	**57.10**	65.67	82.2	43.9	**56.26**
RUbi [24]	64.75	-	-	-	-	-	-	-
Ours	**65.72**	**82.53**	**45.02**	56.08	**65.91**	**82.83**	**44.52**	56.09

IV-A, with the increase of combinations, the performances of both randomly selecting and regional attention are better, however, the model enhanced by hard attention with only one combination sample outperforms randomly select strategy with 5 combinations. Too much redundant information hinders the reasoning of local relationships, which makes the performance of the model significantly reduced.

Elasticity of Hard Attention. The hard attention needs to determine the number of regions to be passed through, called the elasticity of the regional attention. We have experimented with different numbers of $5, 7, 10, 15, 20$ and our proposal model is 10. According to the results, when there are only a few areas passing through the hard attention, the model can not capture enough information for reasoning, which will have a bad impact on the performance of the model. When the number of informative regions is more than 10, more information does not bring significant changes in model performance, but it will lead to the increasing of computing overhead.

2) Multimodal fusion: The multimodal fusion here mainly refers to the fusion in the local relational reasoning phrase, which combines the question embedding and the regional features to infer the relationship between regions. The contrast experiment makes inference only with regional features, ignoring the question context. The result shows that the question information plays an important role in relational reasoning.

3) Local relational reasoning: Local relation reasoning mainly includes two hyperparameters: scales of local relational reasoning and the number of regional combinations for each scale.

Local relational reasoning scales. Scale refers to how many areas are taken as a basis for relationship analysis. More areas can better extract global relationships, but it will weaken local strong relationships, while fewer areas can pay more attention to local relationship information. For example, the positional relationships between two objects. In experiments, we compared multiple different strategies with the proposed model. Namely, scales that favor more regions, scales that favors fewer regions, and scales that varied but fewer. We can see that no matter which kind of strategy, results are not as good as the proposed method. This may be due to the diversity of the questions themselves. Biased strategies of scales or too few scales can not meet the variable needs of different problems, where the relationships need to be extracted in the question have different scales.

M combinations. For each scale, simply inferring all possible combinations of important regions will bring huge computing overhead. Therefore, we have experimented with different sampling numbers: $1, 2, 3, 4, 5$. The result shows that the sampling number can bring better performance, but the effect is very small, and it will increase computing overhead, So we choose 3 as the final number of samples.

C. Comparison with existing methods

We compare the performance of our proposed model with state-of-the-art methods. To prevent overfitting, the VQA challenge 2020 uses two different test sets to test the model: test-dev and test-standard. We display the results of both test sets in Table IV-A. For fairness, all the scores correspond to models trained on VQA v2 $train + val$ split and tested on VQA v2 test-dev and test-standard. Our model surpasses all the models in questions of "Yes/no" and "Numbers", which emphasizes more on reasoning the relationships of regions in the images instead of the form of output. Interestingly, the result of questions of "Numbers", which requires a strong ability of counting, shows that the relational reasoning of informative regions in multi scales gives the model such ability.

Our model has achieved success in extracting informative regions and reasoning regional relations related to the corresponding question, but it is also valuable to keep an eye on the failure cases. Our model doesn't surpass other models in questions of "Others", mainly because the answers to such questions are diverse, and the type and the range of them are determined by the question. But we pay more attention to visual features when modeling, and the semantic information of questions is mostly used to extract visual information. We hope that our work and such critical outlook will eoncourage more breakthroughs in the future.

V. CONCLUSION

In this paper, we proposed regional attention and multi-scales relation reasoning for Visual Question Answering, one for extracting important regions according to the question and the other for reasoning the relationships among them.

We exhibited various ablation studies, demonstrating the efficiency of regional attention and the robustness of joint Multimodal relation reasoning. We validate our approach on VQA v2 and attribute the success of our model to regional attention and multimodal relational reasoning. Our final network is very competitive and outerforms state-of-the-art results on VQA v2.

VI. Acknowledge

This work is supported by Scientific Foundation of State Grid Corporation of China (Research on Icewind Disaster Feature Recognition and Prediction by Fewshot Machine Learning in Transmission Lines).

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition (2016)," in *Proceedings of Conference on Computer Vision and Pattern Recognition*, pp. 770–778.

[2] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks (2015)," in *Proceedings of Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems*, pp. 91–99.

[3] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering (2017)," in *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6325–6334.

[4] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence (2015)," in *Proceedings of 2015 IEEE International Conference on Computer Vision*, pp. 2623–2631.

[5] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text (2015)," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3441–3450.

[6] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering (2015)," *CoRR*, vol. abs/1512.02167.

[7] M. Ren, R. Kiros, and R. S. Zemel, "Image question answering: A visual semantic embedding model and a new dataset (2015)," *CoRR*, vol. abs/1505.02074.

[8] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images (2015)," in *Proceedings of 2015 IEEE International Conference on Computer Vision*, pp. 1–9.

[9] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input (2014)," in *Proceedings of Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems*, pp. 1682–1690.

[10] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention (2015)," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. JMLR Workshop and Conference Proceedings, vol. 37, pp. 2048–2057.

[11] C. Zhu, Y. Zhao, S. Huang, K. Tu, and Y. Ma, "Structured attentions for visual question answering (2017)," in *Proceedings of IEEE International Conference on Computer Vision*, pp. 1300–1309.

[12] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering (2016)," in *Proceedings of Computer Vision European Conference*, ser. Lecture Notes in Computer Science, vol. 9911, pp. 451–466.

[13] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, "Stacked attention networks for image question answering (2016)," in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 21–29.

[14] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering (2018)," in *Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086.

[15] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering (2016)," in *Proceedings of Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems*, pp. 289–297.

[16] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. L. Zitnick, and R. B. Girshick, "Inferring and executing programs for visual reasoning (2017)," in *Proceedings of IEEE International Conference on Computer Vision*, pp. 3008–3017.

[17] A. Santoro, D. Raposo, D. G. T. Barrett, M. Malinowski, R. Pascanu, P. W. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning (2017)," in *Proceedings of NIPS*, pp. 4967–4976.

[18] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, "Film: Visual reasoning with a general conditioning layer (2018)," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)*, pp. 3942–3951.

[19] D. A. Hudson and C. D. Manning, "Compositional attention networks for machine reasoning (2018)," in *Proceedings of 6th International Conference on Learning Representations*.

[20] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation (2014)," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543.

[21] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation (2014)," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734.

[22] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos (2018)," in *Proceedings of ECCV*, pp. 831–846.

[23] D. Teney, P. Anderson, X. He, and A. van den Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge (2018)," in *Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4223–4232.

[24] R. Cadène, C. Dancette, H. Ben-younes, M. Cord, and D. Parikh, "Rubi: Reducing unimodal biases for visual question answering," in *Proceedings of Neural Information Processing Systems*, 2019, pp. 839–850.