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

Yun-Tao Ma[1], Tong Lu*[,1], Yi-Rui Wu[2]

[1]National Key Lab for Novel Software Technology, Nanjing University

[2]College of Computer and Information Hohai University

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 the 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 the multi-scale mechanism gives it the ability to model scaled relationships and makes it sensitive to numbers. We conduct experiments to show that our proposed architecture is effective and achieves competitive results VQA v2.

Proposed Method

As shown in Fig. 1, the proposed method consists of four modules: (a) feature extraction, (b) regional attention, (c) objects relation reasoning, (d) multimodal fusion.

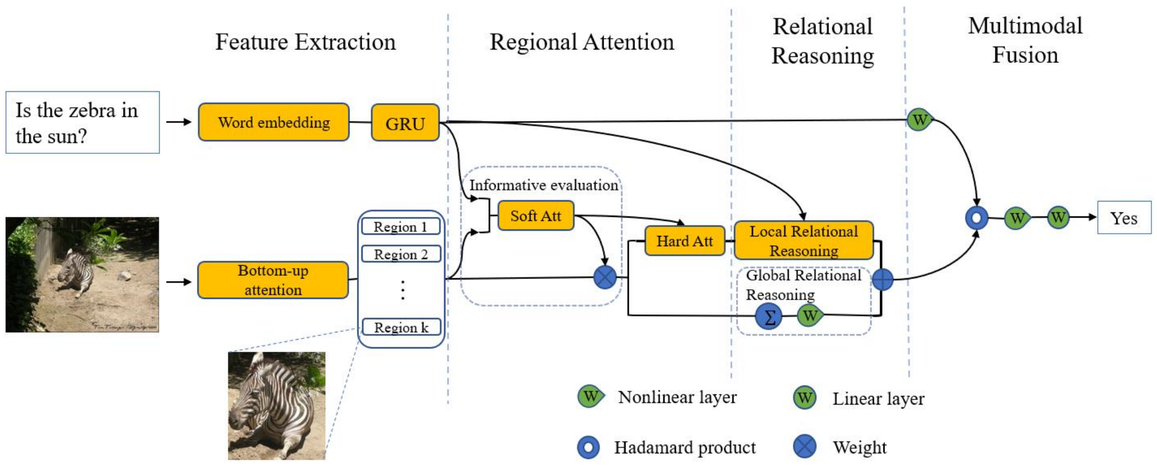

Fig.1 Overview of our proposed model for the example question-answering pair: "Is the zebra in the sun? Yes.". Although we call the last phrase "Multimodal Fusion", the fusion of information in different modalities actually runs through three phrases: regional attention, Relational Reasoning and Multimodal Fusion.

1) Feature Extraction: The input of VQA generally consists of two parts: an image and a text question. The image passed through a ResNet CNN, called bottom-up attention, to generate vector representations of K×2048. The questions are trimmed to a maximum of 14 words and each word is turned into 300-dimension vectors, which is initialized with pretrained GloVe word embeddings. The sequence of word embeddings is then passed through a Gated Recurrent Unit and then transformed into the final state q.

2) Regional Attention: As shown in Fig. 2, the regional attention is mainly composed of two parts, one is an informative evaluation implemented with soft attention, while the other is to pick up informative regions using hard attention. Each of the K areas that generated by feature extraction module, will be multiplied by a normalized informative evaluation based on textual and visual features. There are two parallel and different ways to utilize the weighted vectors. One is fusing the features of all locations to find overall relationship information of all regions. Another is taking informative areas, which picked up by hard attention, as important areas of the image for local relationship information.

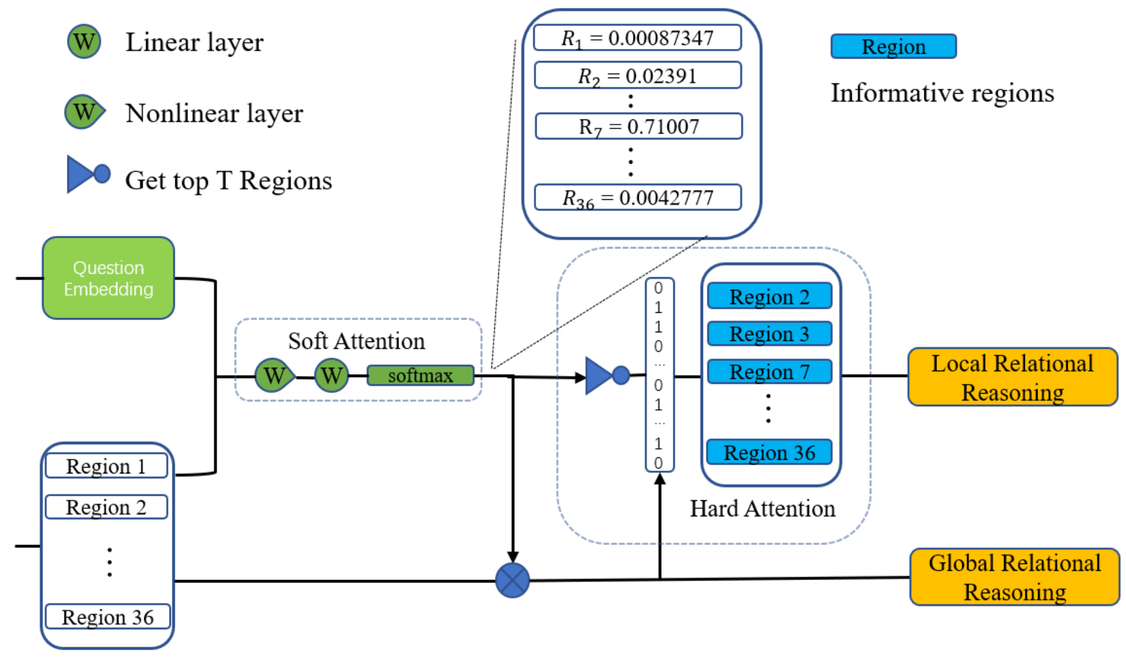

Fig.2 The process of Regional Attention consists of soft attention and hard attention. The output of soft attention module listed here is the intermediate result of the example in Fig. 1, while Region 7 has the biggest informative evaluation as 0.71007.

3) Relational Reasoning: relational reasoning module consists of two parts, Global Relational Reasoning fuses all the areas to extract global information, and Local Relational Reasoning only uses informative areas that picked by hard attention to extract local relationships among important regions.

4) Multimodal Fusion: As shown in Fig. 1, multimodal fusion runs through three phrases. Firstly, the informative evaluation is question-guided, which means that different questions for the same image will generate different informative estimation. Secondly, for local relational reasoning, each combination will concat with question embedding to extract different relationship information under different questions. Last but not least, the results of relational reasoning and question embedding will be combined with Hadamard product.

Experiments

Method	VQA v2 test-dev				VQA v2 test-std			
	All	Y/N	Num	Other	All	Y/N	Num	Other
VQA team-Prior	-	-	-	-	25.98	61.20	00.36	01.07
VQA team--Language only	-	-	-	-	44.26	67.01	31.55	27.37
VQA team-LSTM+CNN	-	-	-	-	54.22	73.46	35.18	41.83
MF-SIG+VG	64.73	81.29	42.99	55.55	-	-	-	-
Adelaide Model*	62.07	79.20	39.46	52.62	62.27	79.32	39.77	52.59
Adelaide Model+detector*	65.32	81.82	44.21	**57.10**	65.67	82.2	43.9	**56.26**
RUbi	64.75	-	-	-	-	-	-	-
Ours	**65.72**	**82.53**	**45.02**	56.08	**65.91**	**82.83**	**44.52**	56.09

Table.1. Results of the proposed method along with other published results on VQA v2 $ test-dev and test-standard splits in similar conditions (i.e., a single model; trained without external dataset). *: trained with external datasets.

As shown in Table 1, We compare the performance of our proposed model with state-of-the-art methods on VQA v2 test-dev and test-standard. Our method 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" shows that the relational reasoning of informative regions in multi scales gives the model ability of counting. 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, which needs more attention on question types. And we pay more attention to visual features when modeling, with the semantic information of questions mostly used to extract visual information.