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

Motivation

- One of the main challenges of visual question answering (VQA) lies in properly reasoning relations among visual regions involved in the question.

Q: What are the bears standing on?

A: Ice

A: Sand

Q: How many bears are in the tree?

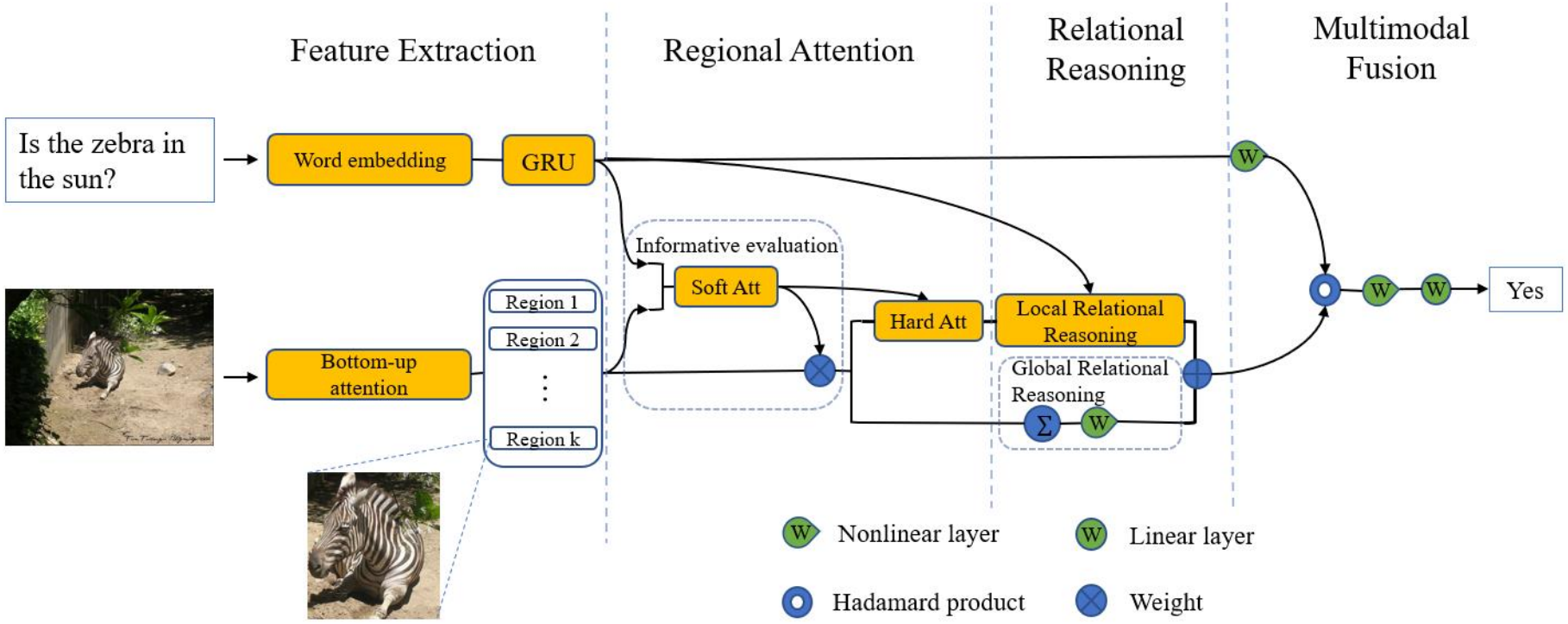
A: Two

A: Zero

- We not only need to recognize objects that related to the question, but also 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.

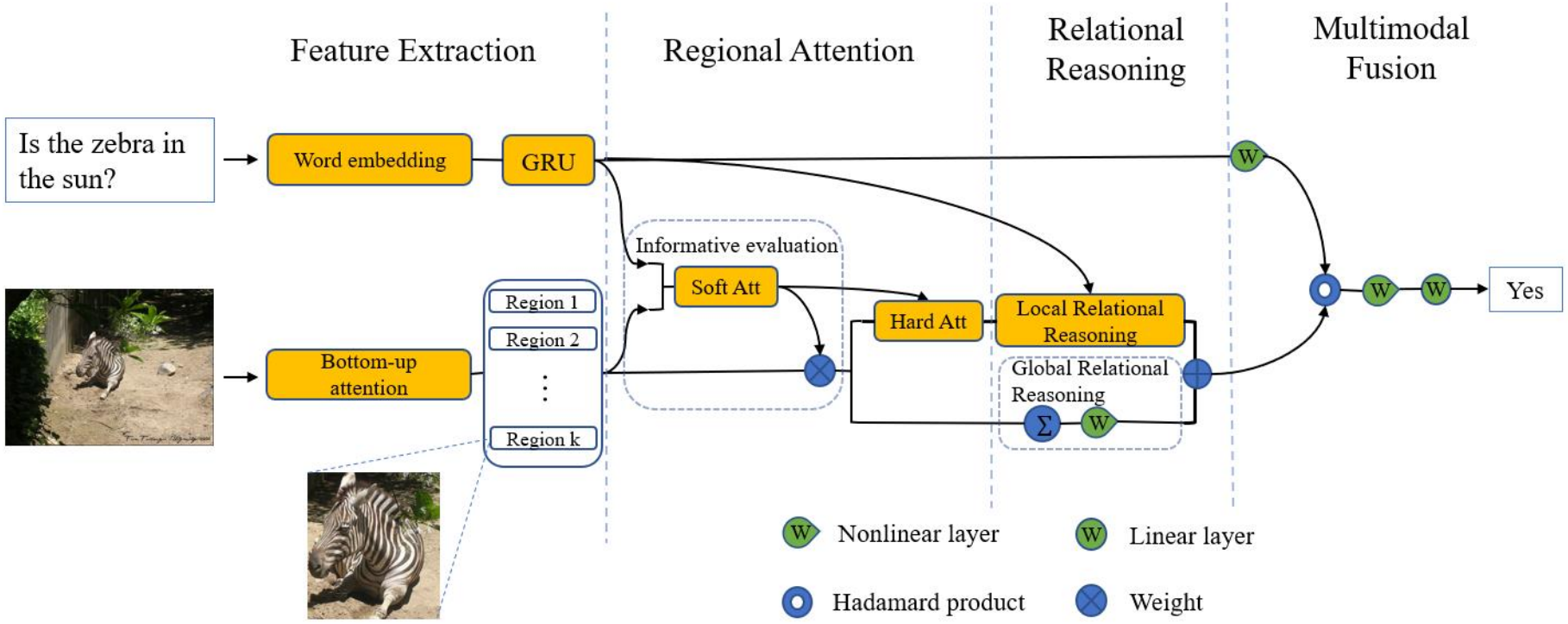
Proposed Method

- The network design of the proposed method consists of four modules: (a) Feature extraction, (b) Regional attention, (c) Relational reasoning, (d) Multimodal fusion.

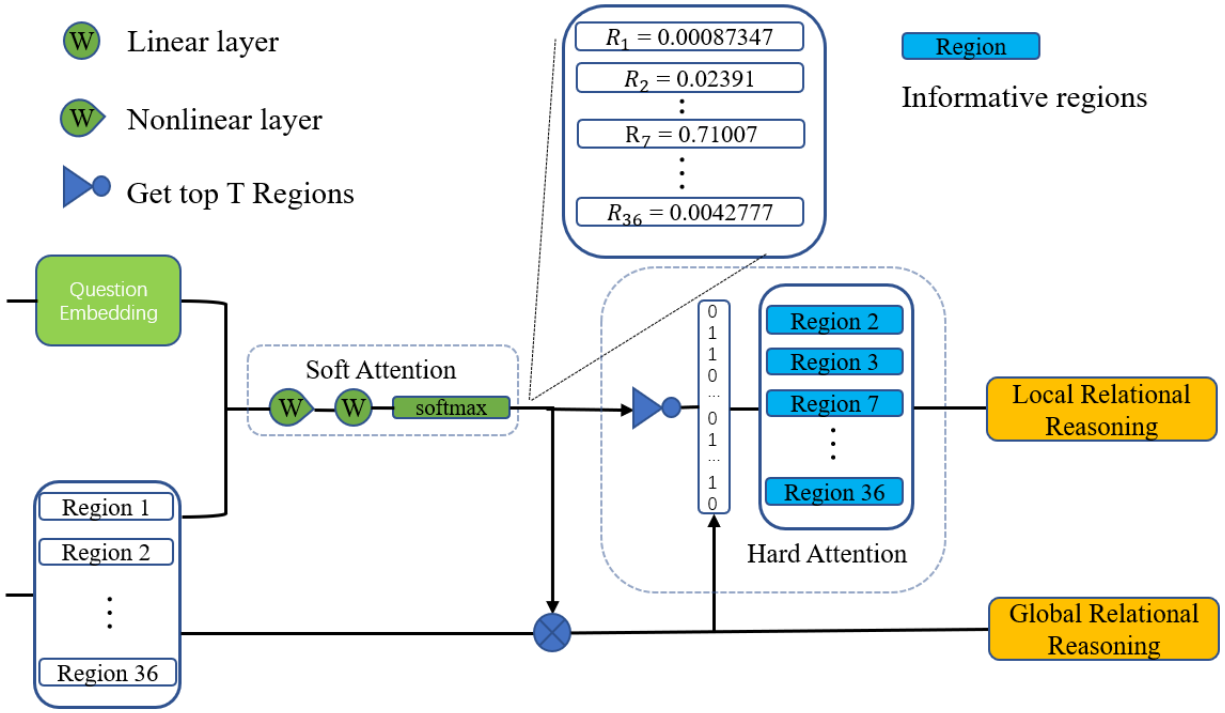


Feature Extraction

- The input of VQA task consists of two parts: an image and a text question.
- Image -> K×2048-vectors Question -> q (2048-vector)

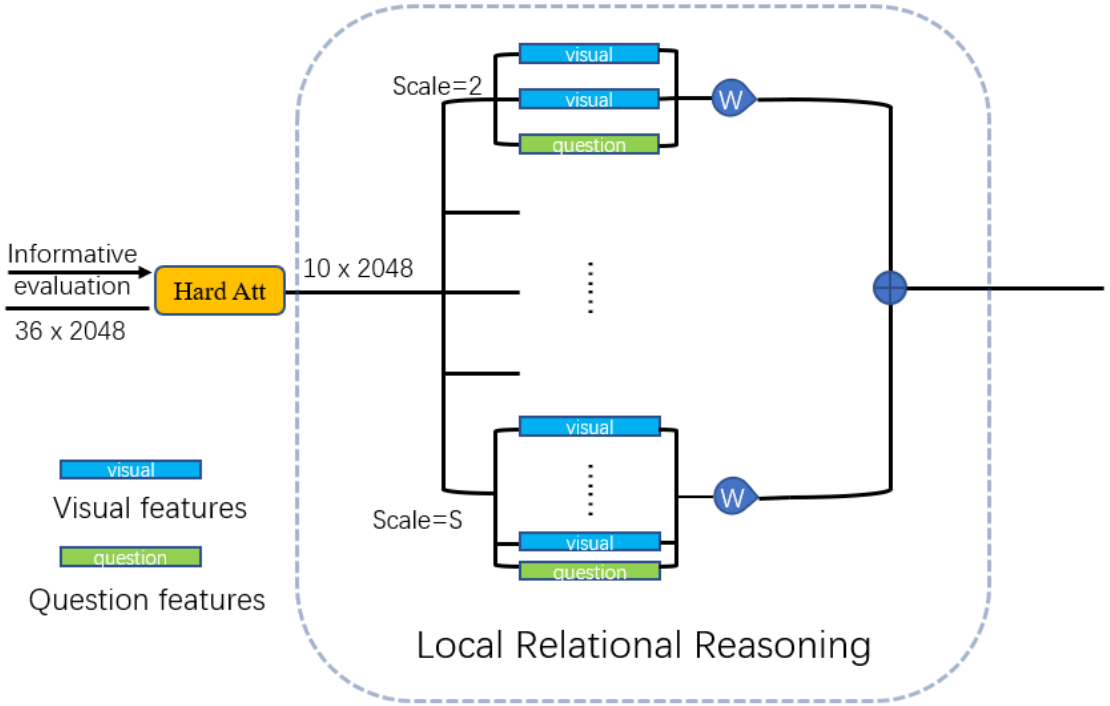


Regional Attention



- Consists of a soft attention module and a hard attention module.
- Select informative regions of the image according to informative evaluations implemented by question-guided soft attention.

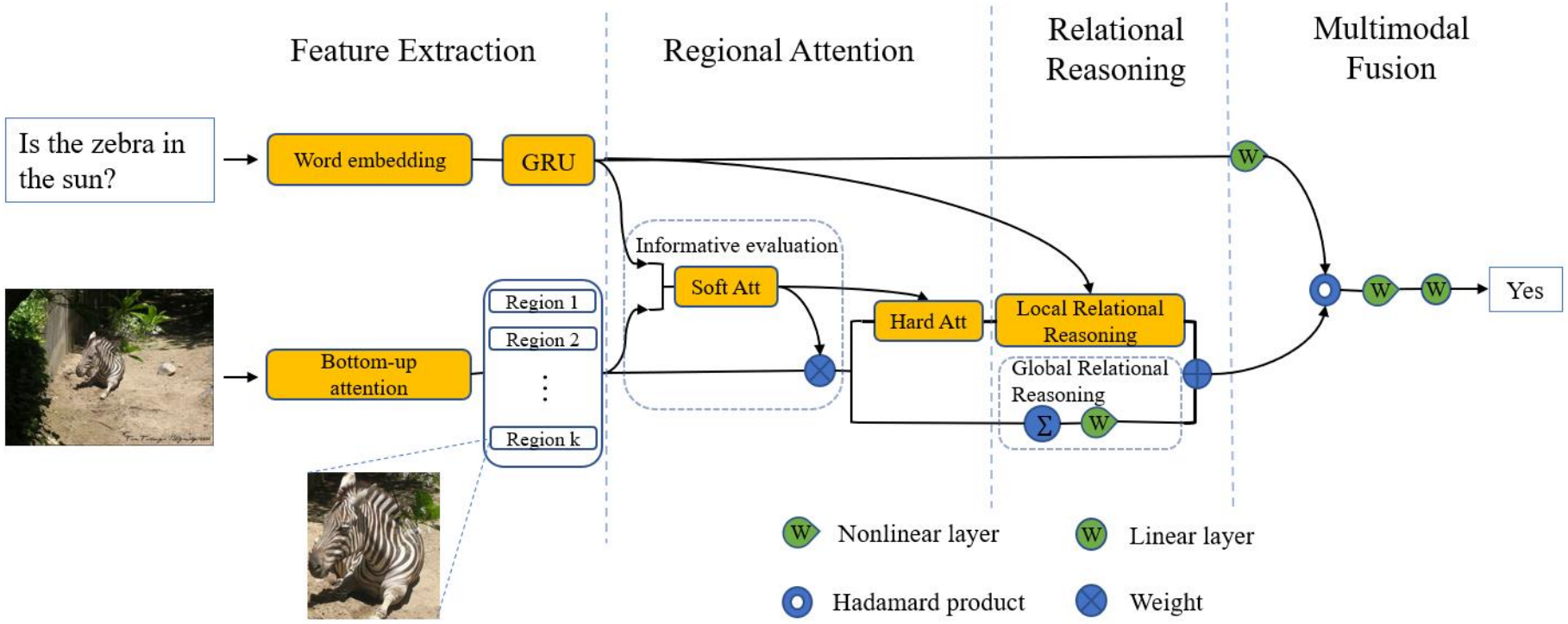
Relational reasoning



- Extract question based relational information among regions.
- In different scales.
- Multi-scale mechanism makes it sensitive to numbers.

Multimodal fusion

- Multimodal fusion runs through three phrases.
- Question-guided.



Experiment Results

- Our proposed architecture is effective and achieves competitive result on VQA v2.

Method	VQA v2 *test-dev*				VQA v2 *test-std*			
	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

Thanks