Chain of Reasoning for Visual Question Answering

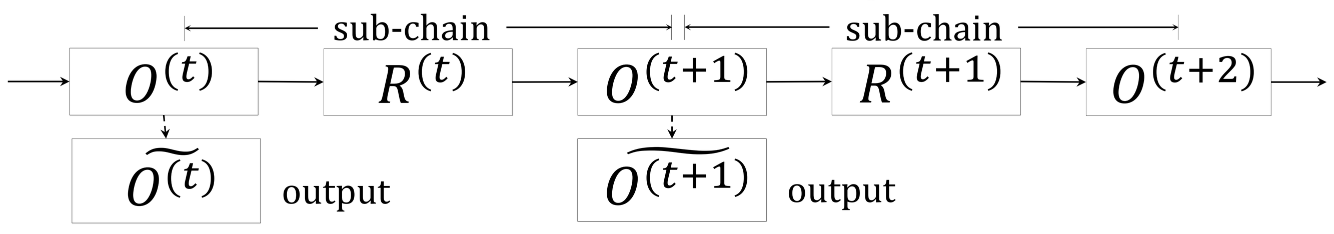
Chenfei Wu, Jinlai Liu, Xiaojie Wang, Xuan Dong
Center for Intelligence Science and Technology
Beijing University of Posts and Telecommunications

insight

"The technical issues of acquiring knowledge, representing it, and using it appropriately to construct and explain **lines-of-reasoning**, are important problems in the design of knowledge-based systems, which illuminates the art of Artificial Intelligence"

-- Edward A. Feigenbaum, "father of expert systems"

How to construct "lines-of-reasoning" ?



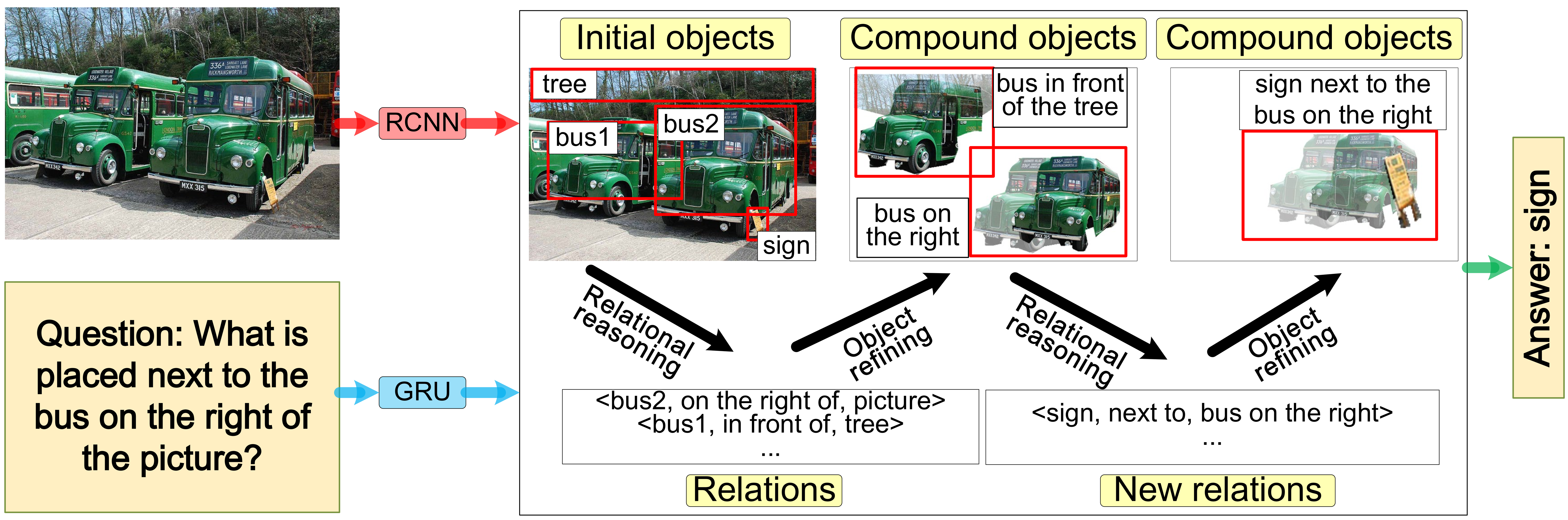
Related Works

Models	How they view reasoning	Deficiencies
Relation-based methods	View reasoning procedure as **one-step** relational reasoning	Not enough to answer **complex** questions
Attention-based methods	View reasoning procedure as to update the attention distribution on **original objects**.	Cannot generate **new objects.**
Module-based methods	View reasoning procedure as a layout generated from **manually pre-defined modules**.	Cannot form **new relations.**

Our work

The reasoning procedure is viewed as the **alternate updating** of objects and relations

Chain of Reasoning Model for VQA



Comparison with state-of-the-arts

We achieve state-of-the-art results on four public datasets: the VQA 1.0 dataset, the VQA 2.0 dataset, the COCO-QA dataset and the TDIUC dataset.

Table 1: Comparision with the state-of-the-arts on the VQA 1.0 dataset.

		VQA 1.0 Test-dev					VQA 1.0 Test-std				
		Open-Ended				MC	Open-Ended				MC
	Method	All	Y/N	Num.	Other	All	All	Y/N	Num.	Other	All
Single image feature	HighOrderAtt[12]	-	-	-	-	69.4	-	-	-	-	69.3
	MLB(7)[14]	66.77	84.54	39.21	57.81	-	66.89	84.61	39.07	57.79	-
	Mutan(5)[16]	67.42	85.14	39.81	58.52	-	67.36	84.91	39.79	58.35	-
Multi image feature	DualMFA[21]	66.01	83.59	40.18	56.84	70.04	66.09	83.37	40.39	56.89	69.97
	ReasonNet[22]	-	-	-	-	-	67.9	84.0	38.7	**60.4**	-
Single image feauture	CoR-2(36boxes) (ours)	68.16	85.57	43.76	58.80	72.60	68.19	85.61	43.10	58.75	72.61
	CoR-3(36boxes) (ours)	**68.37**	**85.69**	**44.06**	**59.08**	**72.84**	**68.54**	**85.83**	**43.93**	59.11	**72.93**

Table 2: Comparision with the state-of-the-arts on the VQA 2.0 dataset.

	VQA 2.0 Test-dev				VQA 2.0 Test-std			
Method	All	Y/N	Num.	Other	All	Y/N	Num.	Other
MF-SIG-VG[23]	64.73	81.29	42.99	55.55	-	-	-	-
Up-Down(36 boxes)[24]	65.32	81.82	44.21	56.05	65.67	82.20	43.90	56.26
LC_Baseline(100 boxes)[25]	67.50	82.98	46.88	58.99	67.78	83.21	46.60	59.20
LC_Counting(100 boxes)[25]	68.09	83.14	**51.62**	58.97	68.41	83.56	**51.39**	59.11
CoR-2(36 boxes) (ours)	67.96	84.7	47.1	58.42	68.15	84.82	46.8	58.52
CoR-3(36 boxes) (ours)	68.19	84.98	47.19	58.64	68.59	85.16	47.19	59.07
CoR-3(100 boxes) (ours)	**68.62**	**85.22**	47.95	**59.15**	**69.14**	**85.76**	48.4	**59.43**

Table 3: Comparision with the state-of-the-arts on the COCO-QA dataset.

Method	All	Obj.	Num.	Color	Loc.	WUPS0.9	WUPS0.0
QRU [26]	62.50	65.06	46.90	60.50	56.99	72.58	91.62
HieCoAtt [11]	65.4	68.0	51.0	62.9	58.8	75.1	92.0
Dual-MFA [21]	66.49	68.86	51.32	65.89	58.92	76.15	92.29
CoR-2(36 boxes) (ours)	68.67	69.76	55.14	73.36	59.52	77.47	92.68
CoR-3(36 boxes) (ours)	**69.38**	**70.42**	**55.83**	**74.13**	**60.57**	**78.10**	**92.86**

Table 4: Comparision with the state-of-the-arts on the TDIUC dataset.

Question Type	MCB-A[13]	RAU[27]	CATL-QTA^W [28]	CoR-2 (ours)	CoR-3 (ours)
Sceen Recognition	93.06	93.96	93.80	94.48	**94.68**
Sport Recognition	92.77	93.47	95.55	**95.94**	95.90
Color Attributes	68.54	66.86	60.16	73.59	**74.47**
Other Attributes	56.72	56.49	54.36	59.59	**60.02**
Activity Recognition	52.35	51.60	60.10	60.29	**62.19**
Positional Reasoning	35.40	35.26	34.71	39.34	**40.92**
Sub. Object Recognition	85.54	86.11	86.98	88.38	**88.83**
Absurd	84.82	96.08	**100.00**	95.17	94.70
Utility and Affordances	35.09	31.58	31.48	**40.35**	37.43
Object Presence	93.64	94.38	94.55	95.40	**95.75**
Counting	51.01	48.43	53.25	57.72	**58.83**
Sentiment Understanding	66.25	60.09	64.38	66.72	**67.19**
Overall (Arithmetric MPT)	67.90	67.81	69.11	72.25	**72.58**
Overall (Harmonic MPT)	60.47	59.00	60.08	65.65	**65.77**
Overall Accuracy	81.86	84.26	85.03	86.58	**86.91**

Ablation study

Table 5: Effectiveness of the chain structure on the VQA 2.0 validation.

Method	MLB[14]	MLB-Stack-2	MLB-Stack-3	MLB-Parallel-2	MLB-Parallel-3	CoR-2 with MLB	CoR-3 with MLB
Val	62.91	63.28	63.55	63.20	63.28	64.90	**64.96**
Method	Mutan[16]	Mutan-Stack-2	Mutan-Stack-3	Mutan-Parallel-2	Mutan-Parallel-3	CoR-2	CoR-3
Val	63.61	63.78	63.90	63.66	63.80	64.96	**65.14**

Tab. 5 shows that our proposed **chain structure** is superior to **stack structure** or **parallel structure**.

Table 6: Effectiveness of relational reasoning operation on the VQA 2.0 validation.

Method	Val
CoR-2 with $[O_i^{(t)}; O_j^{(1)}; G]W_1$	62.46
CoR-2 with $(O_i^{(t)} + O_j^{(1)}) \odot G$	64.73
CoR-2 with $(O_i^{(t)} \odot G_l) \oplus (O_j^{(t)} \odot G_r)$	64.24
CoR-2	**64.96**

Tab. 6 shows that our **relational reasoning** operation $R_{ij}^{(t)} = \left(O_i^{(t)} \odot G_l\right) \oplus \left(O_j^{(1)} \odot G_r\right)$ is superior than others.

Table 7: Effectiveness of object refining operation on the VQA 2.0 validation.

Method	Val
CoR-2 with $\sum_{i=1}^{m} \alpha_i^{(t)} R_{ji}^{(t)}$	64.42
CoR-2	**64.96**

Tab. 7 shows that our **object refining** operation $O_j^{(t+1)} = \sum_{i=1}^{m} \alpha_i^{(t)} R_{ij}^{(t)}$ is superior than the others.

Table 8: Effectiveness of the model on different question types on the CLEVR dataset.

Method	Overall	Count	Exist	Compare Numbers	Query Attribute	Compare Attribute
MLB	85.0	90.0	76.7	78.8	91.1	82.7
Mutan	86.3	92.5	80.2	81.7	91.2	84.5
RN	96.4	-	-	-	-	-
CoR-2	**98.7**	**98.8**	**97.7**	**92.3**	**99.9**	**99.7**

Tab. 8 shows that **the whole model** is superior than MLB, Mutan and RN at the same setup.

Qualitative evaluation

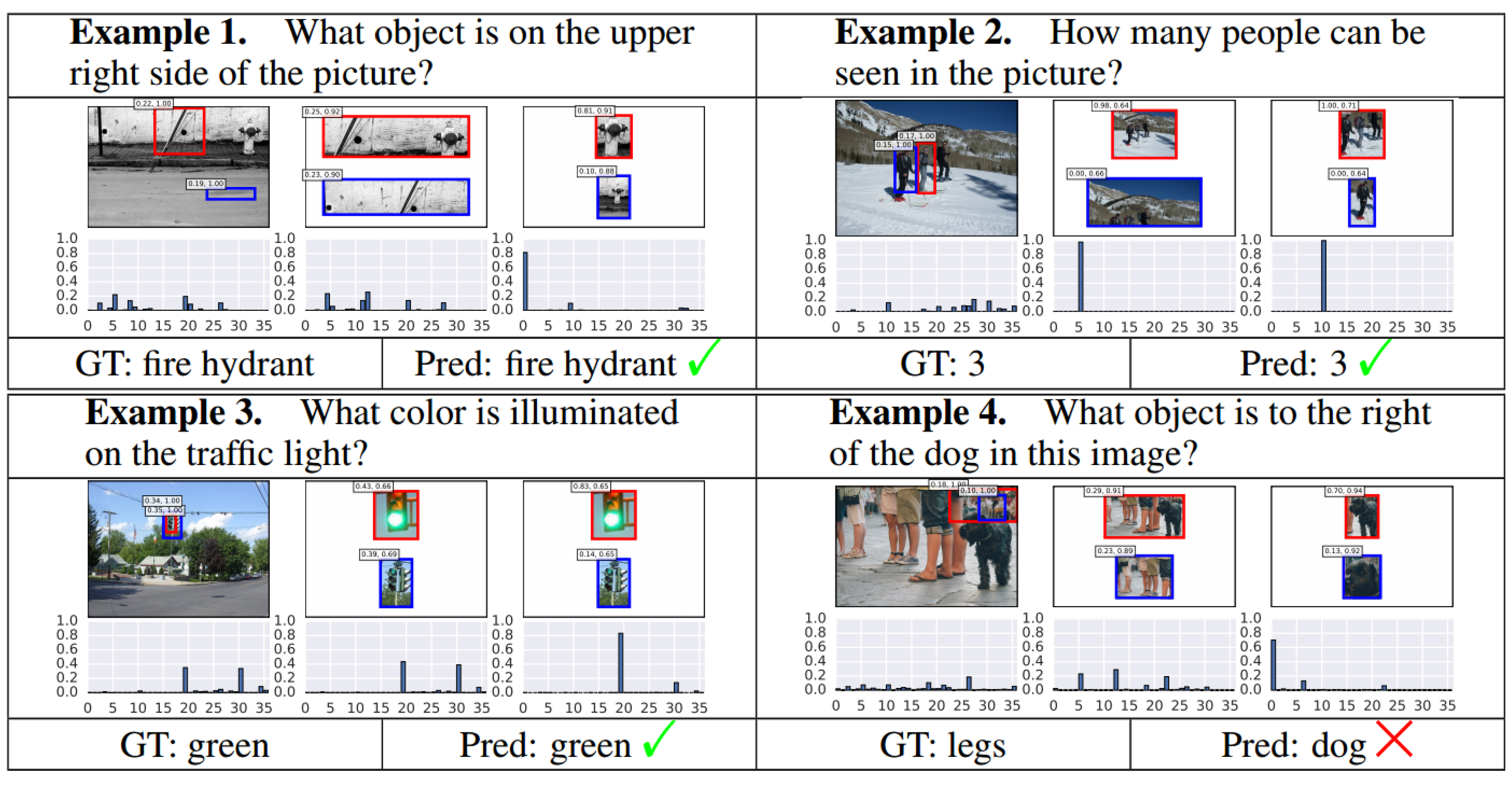


Figure 4: Visualization of the reasoning procedure of CoR-3.

We visualize the compound objects generated by CoR-3 and their attention weights. The upper part of each example shows the top-2 compound objects and the lower part shows the attention distributions. Interestingly, the attention distribution changes from **dispersion** to **concentration**. Statistics show that **96.76%** of the sucess cases in CoR-3 satisfy the phenomenon.