THE 37TH AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE
FEBRUARY 7-14, 2023 • WASHINGTON, DC, USA
WALTER E. WASHINGTON CONVENTION CENTER

Hierarchical ConViT with Attention-based Relational Reasoner for Visual Analogical Reasoning

Wentao He[1], Jialu Zhang[1], Jianfeng Ren[1], Ruibin Bai[1], Xudong Jiang[2]
[1] University of Nottingham Ningbo China
[2] Nanyang Technological University

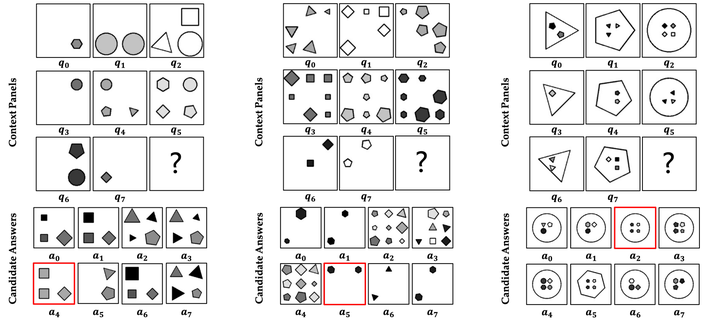
University of Nottingham UK | CHINA | MALAYSIA
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Summary

This poster presents a novel Hierarchical Convolutional Vision Transformer with Attention-based Relational Reasoner (HCV-ARR) for visual abstract reasoning on RPMs, and it achieves the best performance compared with state-of-the-art models on three popular benchmark datasets.

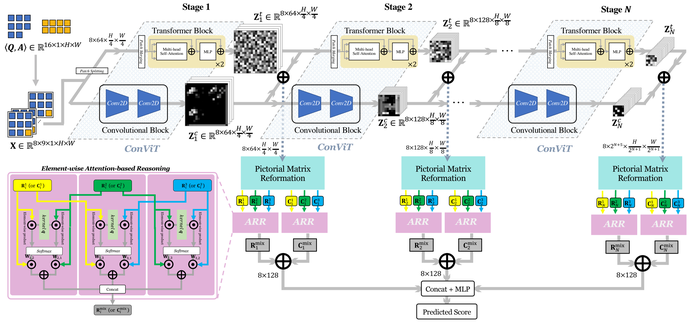
Background/Motivation

The RPM tasks are originally designed as a non-verbal assessment for human intelligence. By using pictorial matrices containing visually simple patterns, the intention is to minimize the impact of language barrier and culture bias.



Method

- A convolutional block is designed to capture the low-level visual attributes, and a transformer block is designed to capture the high-level image semantics.
- Furthermore, we propose to hierarchically recognize the RPM panel in different levels of receptive fields. The hierarchically designed network structure can capture different aspects of the RPM panels at different scales, from overall global insights of attribute knowledge (e.g., *Number* and *Position*) to specific local understandings of image details (e.g., *Type* and *Size*).



- The core intuition of RPMs is that the same reasoning rules are applied across 3 rows/columns, but different attributes may rely on different rules. To conduct robust analogical reasoning based on the extracted visual features, instead of simply detecting the common recurrent patterns among rows/columns, we design an Attention-based Relational Reasoner (ARR) that dynamically learns the combination of rules applied to attributes across rows/columns. The designed element-wise attention mechanism better models the non-linear relations in each attribute among images. The proposed ARR can uncover a combination of a wide range of relational rules in inductive reasoning.

Results

■ *Ablation studies on different model components*

Methods	Accuracy (%) in Different Configurations							
	Avg.	C	2×2G	3×3G	L-R	U-D	O-IC	O-IG
MRNet (CVPR'21)	86.8	97.0	72.7	69.5	98.7	98.9	97.6	73.3
TSCNN (TIFS'19)	87.9	96.8	73.5	74.7	94.4	91.8	94.5	88.9
HCV	92.7	99.9	85.7	78.4	99.9	99.8	99.8	85.4
Conv + ARR	93.4	99.9	86.3	79.8	99.8	99.7	99.6	88.7
ViT + ARR	44.9	52.0	39.4	41.0	35.8	35.8	57.8	50.4
Proposed HCV-ARR	95.4	99.8	92.9	87.9	99.8	99.6	99.7	88.5

■ *Ablation studies on different depths of network*

ID	Index of Stages 1st	2nd	3rd	Accuracy (%) in Different Configurations Avg.	Center	2×2G	3×3G	L-R	U-D	O-IC	O-IG
1.	✔	✗	✗	74.9	83.2	53.1	58.1	90.6	90.3	87.4	61.7
2.	✗	✔	✗	85.8	98.9	66.9	64.5	99.1	99.5	98.5	73.6
3.	✗	✗	✔	93.7	99.8	86.7	84.1	99.6	99.6	99.7	86.9
4.	✔	✔	✗	87.8	98.9	68.1	68.7	99.3	99.2	99.1	78.1
5.	✔	✔	✔	95.4	99.8	92.9	87.9	99.8	99.6	99.7	88.5

■ *Results on the RAVEN-FAIR*

Methods	Accuracy (%) in Different Configurations Avg.	C	2×2G	3×3G	L-R	U-D	O-IC	O-IG
CoPINet (NIPS'19)	36.5	–	–	–	–	–	–	–
LEN (NIPS'19)	50.9	–	–	–	–	–	–	–
DCNet (ICLR'21)	57.0	57.2	48.4	58.2	57.5	59.4	62.0	56.2
SRAN (AAAI'21)	76.7	87.4	60.4	62.8	86.5	86.7	77.5	75.9
MRNet (CVPR'21)	86.8	97.0	72.7	69.5	98.7	98.9	97.6	73.3
Proposed HCV-ARR	**95.4**	**99.8**	**92.9**	**87.9**	**99.8**	**99.6**	**99.7**	**88.5**

■ *Results on the I-RAVEN*

Methods	Accuracy (%) in Different Configurations Avg.	C	2×2G	3×3G	L-R	U-D	O-IC	O-IG
CoPINet (NIPS'19)	46.1	54.4	36.8	31.9	51.9	52.5	52.2	42.8
LEN (NIPS'19)	41.4	56.4	31.7	29.7	44.2	44.2	52.1	31.7
DCNet (ICLR'21)	46.6	56.2	32.7	32.9	54.7	53.9	55.9	39.8
SRAN (AAAI'21)	60.8	78.2	50.1	42.4	70.1	70.3	68.2	46.3
MRNet (CVPR'21)	81.0	99.6	63.4	59.2	98.7	98.3	95.7	51.9
Proposed HCV-ARR	**93.9**	**99.9**	**96.2**	**75.5**	**99.4**	**99.6**	**99.5**	**87.3**

■ *Results on the original RAVEN*

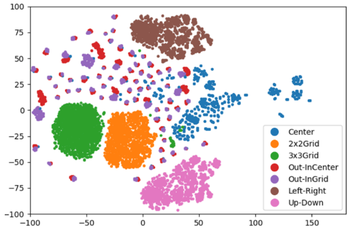
Methods	Accuracy (%) in Different Configurations Avg.	C	2×2G	3×3G	L-R	U-D	O-IC	O-IG
CoPINet (NIPS'19)	18.4	–	–	–	–	–	–	–
SRAN (AAAI'21)	46.2	49.0	45.4	52.8	42.4	46.0	49.1	48.8
MRNet (CVPR'21)	84.0	–	–	–	–	–	–	–
Proposed HCV-ARR	**87.3**	**99.8**	**71.4**	**65.9**	**99.9**	**99.8**	**98.0**	**76.2**

(a) Without contrasting on candidate answers

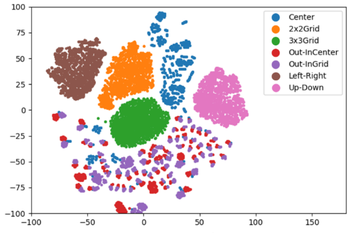
Methods	Accuracy (%) in Different Configurations Avg.	C	2×2G	3×3G	L-R	U-D	O-IC	O-IG
CoPINet (NIPS'19)	91.4	95.1	77.5	78.9	99.1	99.7	98.5	91.4
LEN (NIPS'19)	72.9	80.2	57.5	62.1	73.5	81.2	84.4	71.5
Rel-AIR (ICCV'20)	94.1	99.0	**92.4**	87.1	98.7	97.9	98.0	85.3
DCNet (ICLR'21)	93.6	97.8	81.7	86.7	99.8	99.8	99.0	91.5
MRNet (CVPR'21)	**96.6**	–	–	–	–	–	–	–
Proposed HCV-ARR	96.0	**99.4**	86.9	**89.1**	**99.9**	**99.9**	**99.8**	**96.8**

(b) With contrasting on candidate answers

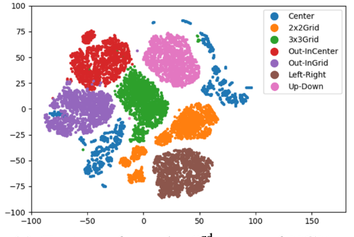
■ *Visualization of the perception module*



(a) Features from the 1st stage of HCV.



(b) Features from the 2nd stage of HCV.



(c) Features from the 3rd stage of HCV.