Abstract Spatial-Temporal Reasoning via Probabilistic Abduction and Execution

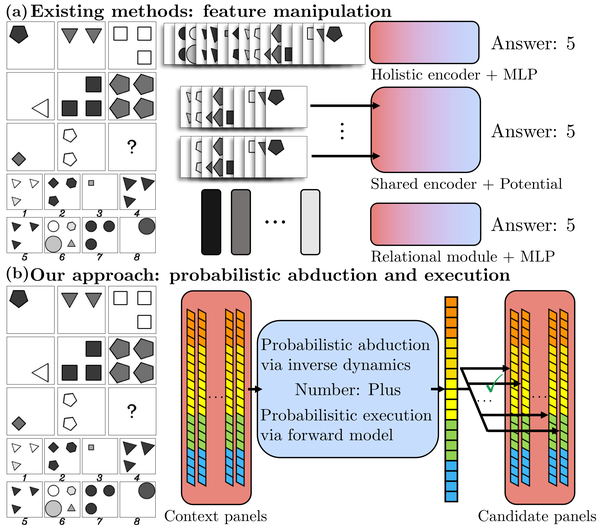
CVPR 2021
NASHVILLE, TENNESSEE

Chi Zhang★ Baoxiong Jia★ Song-Chun Zhu Yixin Zhu

UCLA Center for Vision, Cognition, Learning and Autonomy

{chi.zhang, baoxiongjia}@ucla.edu, sczhu@stat.ucla.edu, yixin.zhu@ucla.edu

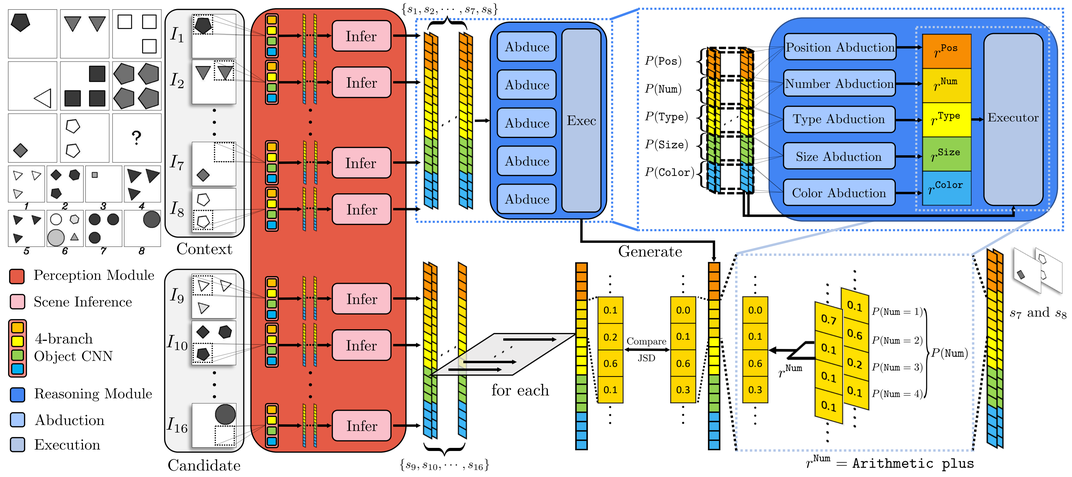
Motivation



What is ***not so right*** about previous methods:

- Existing methods only vary in how image features are manipulated, lacking in *transparency*, *interpretability*, *generalization*, and *the ability to incorporate knowledge*.
- Bottom-up only, without a top-down process, while humans arguably use top-down bottom-up reasoning: we could apply a generative process to abduce rules and execute them to synthesize a possible solution, and discriminatively select an answer.

PrAE



The PrAE Learner

- PrAE is a neuro-symbolic method that disentangles perception and reasoning.

Neural Visual Perception

- Perception composes of two modules: object CNN and scene inference engine.
- The object CNN slides across the spatial domain of each image and produces object attribute distributions for each image region, *i.e.*, *objectiveness*, *type*, *size*, and *color*.
- The scene inference engine aggregates each image region' object attribute distributions and marginalize them out for its probabilistic scene representation.

Symbolic Logical Reasoning

- *Abduction* works out the hidden rule: $P(r^a \mid I_1^a, \ldots, I_8^a) \propto \sum_{S^a \in \texttt{valid}(r^a)} \prod_{i=1}^{8} P(s_i^a = S_i^a)$
- Similar to computing inverse dynamics.
- *Execution* selects a rule from the distribution and executes it in a probabilistic planning manner to get the prediction: $P(s_3^a = S_3^a) \propto \sum_{\substack{(S_2^a, S_1^a) \in \texttt{pre}(r^a) \\ S_3^a = f(S_2^a, S_1^a; r^a)}} P(s_2^a = S_2^a) P(s_1^a = S_1^a)$
- Most similar candidate is selected.

Performance

PrAE significantly improves cross-configuration generalization

Method	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
WReN	9.86/14.87	8.65/14.25	29.60/20.50	9.75/15.70	4.40/13.75	5.00/13.50	5.70/14.15	5.90/12.25
LSTM	12.81/12.52	12.70/12.55	13.80/13.50	12.90/11.35	12.40/14.30	12.10/11.35	12.45/11.55	13.30/13.05
LEN	12.29/13.60	11.85/14.85	41.40/18.20	12.95/13.35	3.95/12.55	3.95/12.75	5.55/11.15	6.35/12.35
CNN	14.78/12.69	13.80/11.30	18.25/14.60	14.55/11.95	13.35/13.00	15.40/13.30	14.35/11.80	13.75/12.85
MXGNet	20.78/13.07	12.95/13.65	37.05/13.95	24.80/12.50	17.45/12.50	16.80/12.05	18.05/12.95	18.35/13.90
ResNet	24.79/13.19	24.30/14.50	25.05/14.30	25.80/12.95	23.80/12.35	27.40/13.55	25.05/13.40	22.15/11.30
ResNet+DRT	31.56/13.26	31.65/13.20	39.55/14.30	35.55/13.25	25.65/12.15	32.05/13.10	31.40/13.70	25.05/13.15
SRAN	15.56/29.06	18.35/37.55	38.80/38.30	17.40/29.30	9.45/29.55	11.35/28.65	5.50/21.15	8.05/18.95
CoPINet	52.96/22.84	49.45/24.50	61.55/31.10	**52.15**/25.35	68.10/20.60	65.40/19.85	39.55/19.00	34.55/19.45
PrAE Learner	**65.03/77.02**	**76.50/90.45**	**78.60/85.35**	28.55/**45.60**	**90.05/96.25**	**90.85/97.35**	**48.05/63.45**	**42.60/60.70**
Human	84.41	95.45	81.82	79.55	86.36	81.81	86.36	81.81

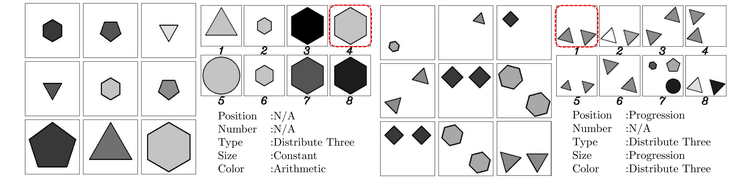
Jointly trained object CNN performance

Object Attribute	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
Objectiveness	93.81/95.41	96.13/96.07	99.79/99.99	99.71/97.98	99.56/95.00	99.86/94.84	71.73/88.05	82.07/95.97
Type	86.29/89.24	89.89/89.33	99.95/95.93	83.49/85.96	99.92/92.90	99.85/97.84	91.55/91.86	66.68/70.85
Size	64.72/66.63	68.45/69.11	71.26/73.20	71.42/62.02	73.00/85.08	73.41/73.45	53.54/62.63	44.36/40.95
Color	75.26/79.45	75.15/75.65	85.15/87.81	62.69/69.94	85.27/83.24	84.45/81.38	84.91/75.32	78.48/82.84

Accuracy of probabilistic abduction

Panel Attribute	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
Pos/Num	90.53/91.67	-	90.55/90.05	92.80/94.10	-	-	-	88.25/90.85
Type	94.17/92.15	100.00/95.00	99.75/95.30	63.95/68.40	100.00/99.90	100.00/100.00	100.00/100.00	86.08/77.60
Size	90.06/88.33	98.95/99.00	90.45/89.90	65.30/70.45	98.15/96.78	99.45/92.45	93.08/96.13	77.35/70.78
Color	87.38/87.25	97.60/93.75	88.10/85.35	37.45/45.65	98.90/92.38	99.40/98.43	92.90/97.23	73.75/79.48

Predicted results from a rendering engine



Future Work

- How to induce hidden rules from a few examples?
- How to crystalize such knowledge in the examples?
- How can visual features learned on other tasks be transferred to the reasoning problem?

VCLA
VCLA@UCLA