Measuring abstract reasoning in neural networks

David G.T. Barrett, Felix Hill, Adam Santoro, Ari S. Morcos, Timothy Lillicrap

ICML, 2018

Presenter: Xinran(Nicole) Han

April 6, 2020

Penn Engineering

Problem & Motivation

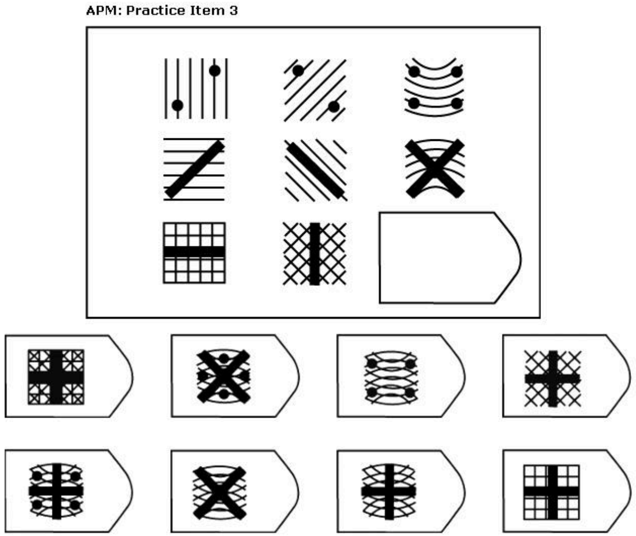
- Can neural networks learn abstract reasoning?
  - Or do they merely rely on superficial statistics?
  - Can they solve visual reasoning problems?

Major contributions:

- Dataset to measure abstract reasoning and evaluate generalizability
- Wild Relation Network (WReN)
- Propose a way to improve generalizability through auxiliary training

Problem & Motivation

- How do we measure human intelligence?
- One popular IQ test: Raven's Progressive Matrices (RPMs)
  - Developed in the 1930s to examine general intelligence
  - Consists of multiple choice visual analogy problems
  - Strongly diagnostic of abstract verbal, spatial and mathematical reasoning ability, discriminating even among populations of highly educated subjects
  - Potential pitfall: can be invalidated if subjects prepare too much



Dataset Generation

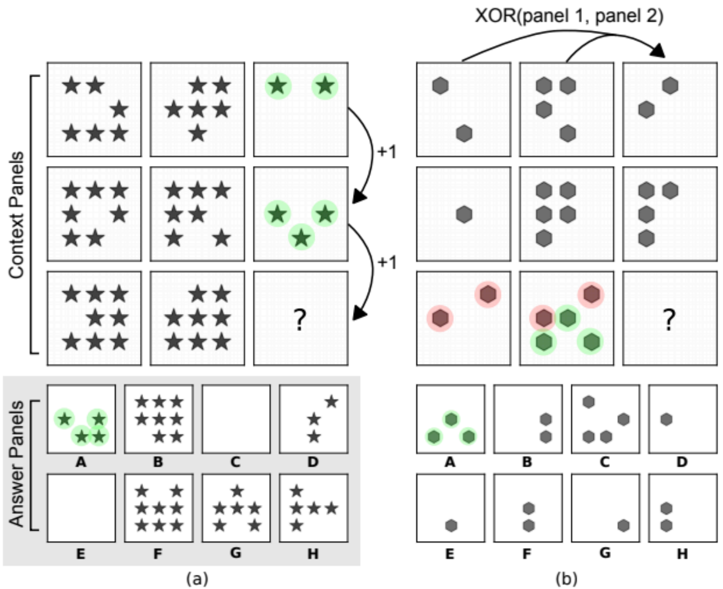
- Human intelligence test: RPM
- The right answer tends to be the one that can be explained with simplest justification using basic relations

- Procedurally Generated Matrices(PGM) Dataset
- Abstract Structure $S = \{[r, o, a]: r \in R, o \in O, a \in A\}$:
  - Relation types R : progression, XOR, OR, AND, consistent union
  - Object types O : shape, lines
  - Attribute types A : size, type, color, position, number
- Example: [progression, shape, color] changes in color intensity of shapes

Dataset Generation

- Procedurally Generated Matrices(PGM) Dataset
- S_a: set of attributes among the triples in S
- Generation process:

1. Sample 1-4 triples
2. Sample values $v \in V$ for each $a \in S_a$
3. Sample values $v \in V$ for each $a \notin S_a$
4. Render symbolic forms to pixels



Generalization Regime

- How to evaluate reasoning and generalizability?
- Have different patterns/rules for train/test sets

1. *Neutral*
2. *Interpolation*
3. *Extrapolation*
4. *Held-out Attribute shape-color*
5. *Held-out attribute line-type*
6. *Held-out Triples: randomly choose 7 out of 29 possible unique triples for test set*
7. *Held-out Pairs of Triples: 360/400 triple pairs for training, 40 for testing*
8. *Held-out Attribute Pairs*

Baseline Models

- Models are trained to predict the label of the correct missing panel
  - **CNN-MLP**: four-layer convolution + 2 layer fully connected layer, with ReLU and dropout layer
  - **ResNet-50**: from He et al. (2016)
  - **LSTM**: each panel is passed sequentially and independently through a 4-layer CNN, tagged with position, then passed to a standard LSTM module
  - **Context-blind ResNet**: train ResNet-50 model with only the right multiple-choice panels as input.
  - Random guessing should yield around 12.5% accuracy. Strong models can exploit statistical regularities among multiple choice inputs alone.

Proposed Model

- **Wild Relation Network (WReN)**
- Applied a Relation Network Module (Santoro et al. 2017) multiple times to infer inter-panel relationships
- Each candidate choice panel is assigned a score using a Relation Network (RN)

$$
\begin{aligned}
s_k &= \mathrm{RN}(\mathcal{X}_k) \\
&= f_\phi\Big(\sum_{y,z \in \mathcal{X}_k} g_\theta(y,z) \Big),
\end{aligned}
$$

- Similarly,
- Proposed **Wild-ResNet**: one multiple choice candidate + eight context panels are provided as input for a score

Model	WReN	Wild-ResNet	ResNet-50	LSTM	CNN+MLP	Blind ResNet
Test Acc(%)	62.6	48.0	42.0	35.8	33.0	22.4

Proposed Model

- From original paper

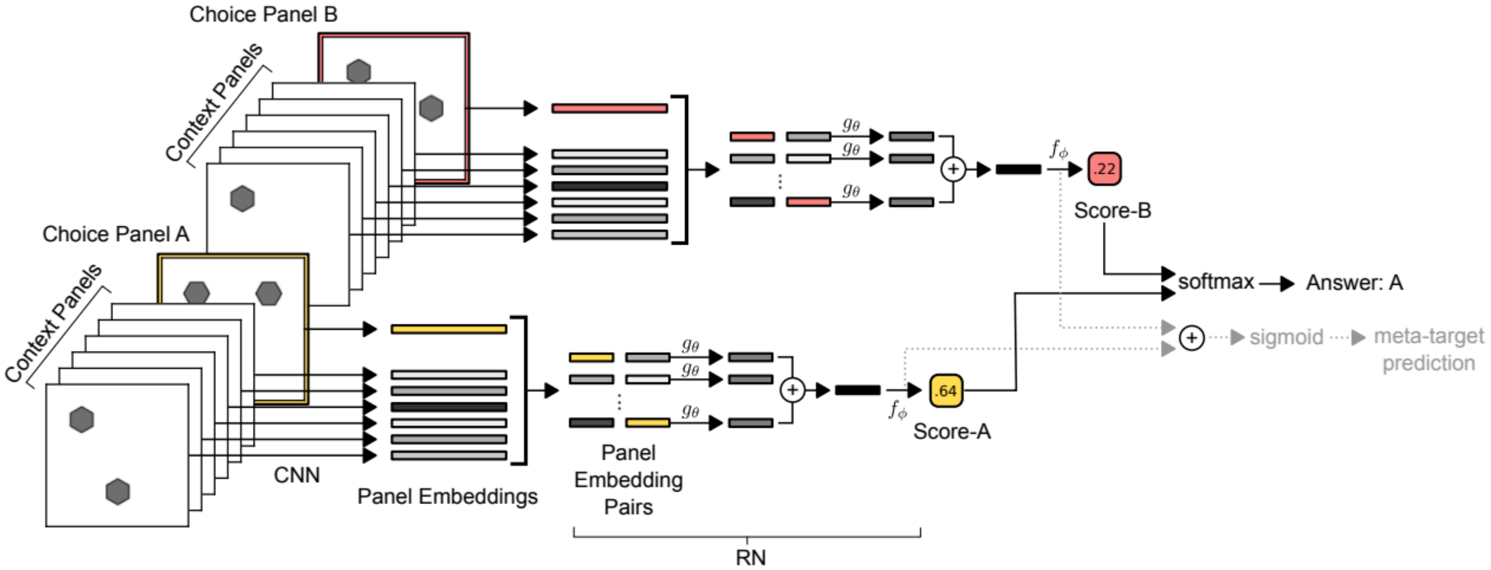


Figure 3. **WReN model** A CNN processes each context panel and an individual answer choice panel independently to produce 9 vector embeddings. This set of embeddings is then passed to an RN, whose output is a single sigmoid unit encoding the "score" for the associated answer choice panel. 8 such passes are made through this network (here we only depict 2 for clarity), one for each answer choice, and the scores are put through a softmax function to determine the model's predicted answer.

Auxiliary Training

- Which shapes, attributes and relations does the model think are present in the PGM?
- Construct 'meta-targets': (shape, line, color, number, position, size, type, progression, XOR, OR, AND, consistent union). Each entry is binary, based on whether the shape/attribute/relation exists
- $L_{total} = L_{target} + \beta L_{meta-target}$
- In the Neutral regime, leads to 13.9% improvement in test accuracy

Experiment Results

- From original paper

Regime	$\beta = 0$ Val. (%)	Test (%)	Diff.	$\beta = 10$ Val. (%)	Test (%)	Diff.
Neutral	63.0	62.6	-0.6	77.2	76.9	-0.3
Interpolation	79.0	64.4	-14.6	92.3	67.4	-24.9
H.O. Attribute Pairs	46.7	27.2	-19.5	73.4	51.7	-21.7
H.O. Triple Pairs	63.9	41.9	-22.0	74.5	56.3	-18.2
H.O. Triples	63.4	19.0	-44.4	80.0	20.1	-59.9
H.O. `line-type`	59.5	14.4	-45.1	78.1	16.4	-61.7
H.O. `shape-colour`	59.1	12.5	-46.6	85.2	13.0	-72.2
Extrapolation	69.3	17.2	-52.1	93.6	15.5	-78.1

Experiment Results Analysis

- Performance vary for different relation types:
  - For single relation triples: OR(64.7%) XOR(53.2%)
  - For triples involving lines(78.3%), involving shapes(46.2%)
  - Shape-Number(80.1%), Shape-size(26.4%)

  - For training with meta-target:

Accuracy(%)	Shape	Attribute	Relation	All
Correct Target	78.2	79.5	86.8	87.4
Wrong Target	62.2	49.0	32.1	34.8

Discussion

- Contributions
  - Proposed a dataset with means to measure different generalization abilities
  - Architecture of model made a critical difference on the reasoning ability
  - Better performance if model is required to decode representations into symbols
- Limitations
  - Model's world is highly constrained, does not resemble human knowledge acquirement
  - Poor generalization in many settings, such as Extrapolation.
  - Limited rules, relations and geometry
  - Other possible definitions for abstract reasoning
- Possible Future Work
  - Transfer knowledge from other datasets that also contains similar relations, such as counting, OR, etc. (VQA)

Follow-up Works

- **RAVEN: A Dataset for Relational and Analogical Visual REasoNing,** CVPR 2019

  Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, Song-Chun Zhu

- **Learning to Make Analogies by Contrasting Abstract Relational Structure,** ICLR 2019

  Felix Hill, Adam Santoro, David G.T. Barrett, Ari Morcos & Tim Lillicrap