Part 3

A Perspective View and Survey of Meta-Learning

By Ricardo Vilalta and Youssef Drissi

Presented by Rasheed Ali R.

Solutions to Base-Learning Limits

- ★ Direct application of Meta-Learning
  - Stacked Generalization
- Unbounded adaptive bias learners
  - Evolution Based Learners
  - Analogy Based Learners

Evolution Theory

1. Natural Selection:

   An entity exists unless it parishes (Duh!)

   $$C: E \rightarrow E + \emptyset$$

2. Reproduction:

   Existing increases odds of replication

   $$R: E \rightarrow \{E, E\}$$

3. Mutation:

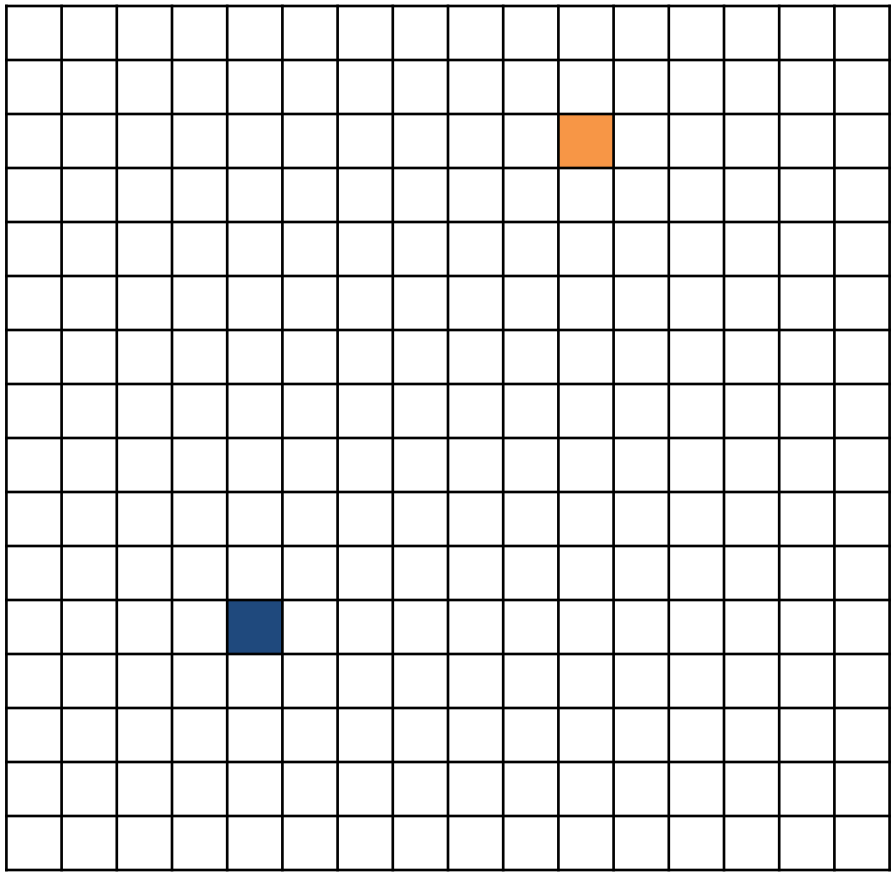
   An entity may transform into another entity

   $$T(E) \rightarrow E'$$

Intelligence Agents & Evolution

- The pool of human thoughts:
  - C: A thought is forgotten unless it is minimally:
    - amusing
    - simple
    - useful…
  - R: A thought gets replicated:
    - communication…
  - T: A thought is modified
    - Accident: Misinterpretation/misrepresentation
    - Deliberate: Scientific or artistic process…

Simple Evolutionary Learning



Agent Makeup

- General memory: 2040 Int64 cells
- Observation memory: 8 Int64 cells
  - Cell 0: remaining energy
  - Cell 1,2: x,y-coordinate of self
  - Cell 3,4: x,y-direction of self
  - Cell 5,6: x,y-coordinate of target

Agent Makeup

- Instruction Set:
  - Move: 100 energy
  - Turn: 5 energy
  - Copy: 1 energy
  - Think: 1 energy

Agent Makeup

- Think instruction representation:
  - Instruction ID
  - Address of next instruction
  - Address of input start
  - Address of output start

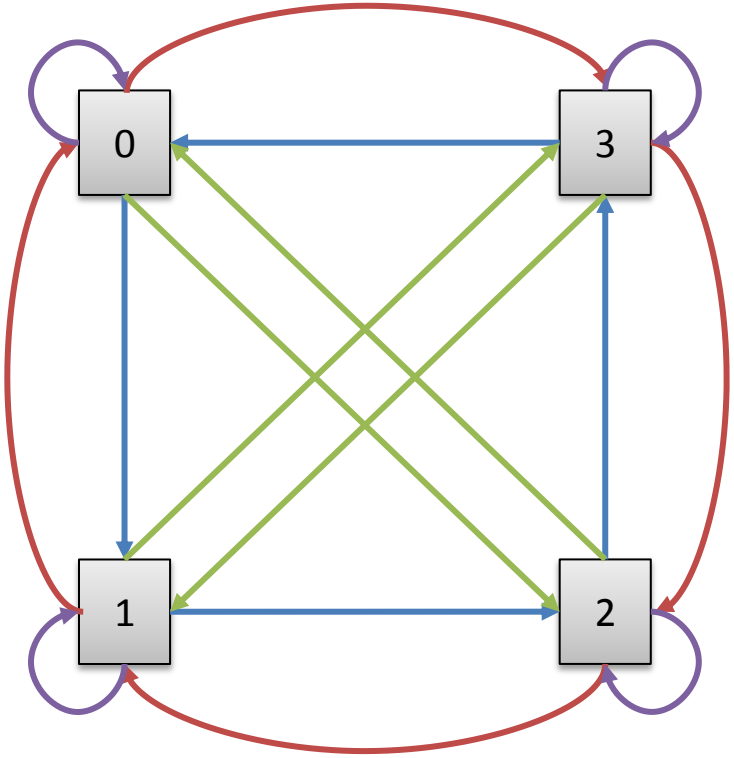
Agent Thinking

- Agents are Turing complete
- Thinking operation:
  - $Think(n, m) \rightarrow \pi_n(m)$ where $\pi_n \in S_k(\mathbb{N}, \circ)$
- Thinking simulates any TM DFA δ
- Memory simulates TM tape/DFA states
- Copy action simulate all TM tape actions

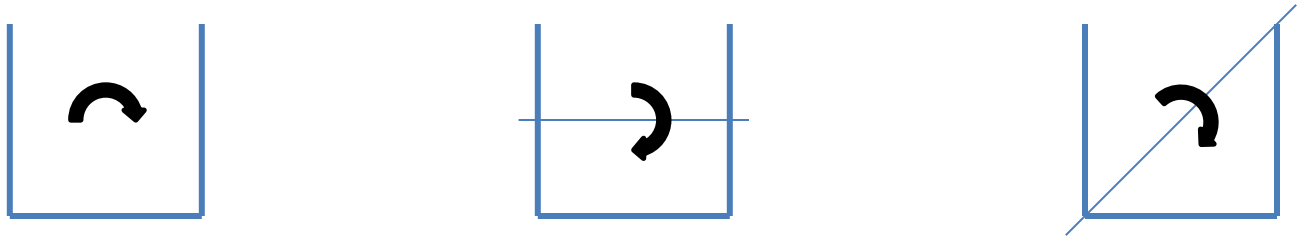
Thinking Addition

- Addition in $\mathbb{Z}/_4 = (\{0,1,2,3\}, +)$
- $0 + 0 = 0$
- $0 + 1 = 1$
- $0 + 2 = 2$
- $0 + 3 = 3$

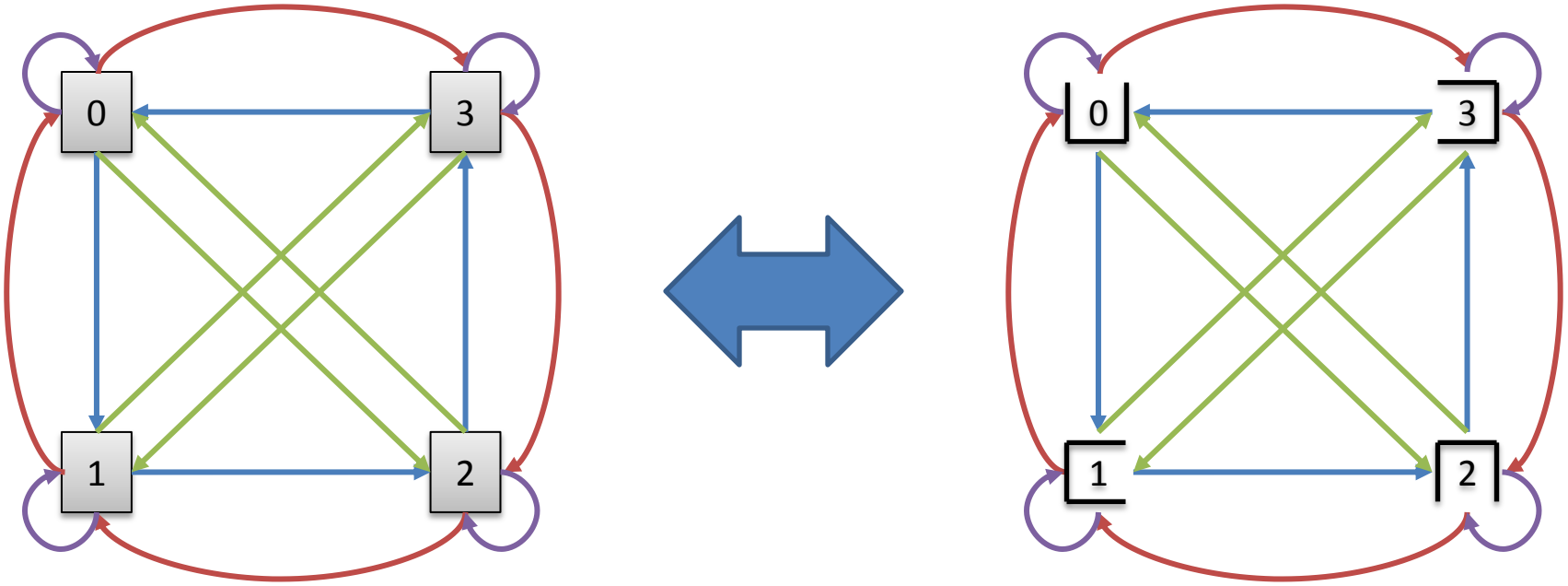
Thinking Addition



Rotating a Surface

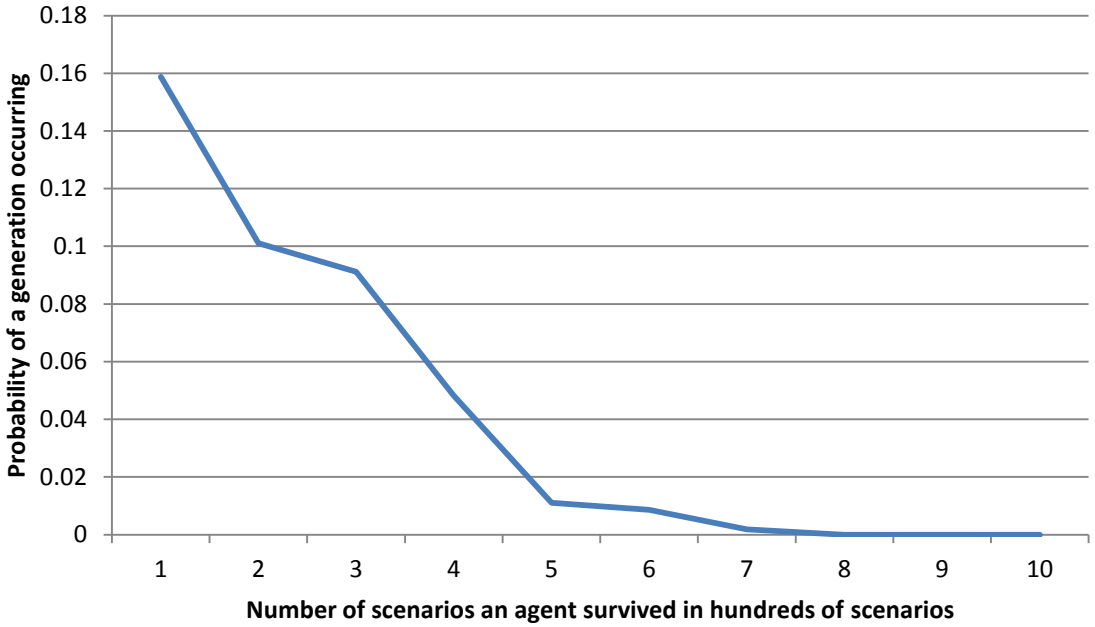


Analogy

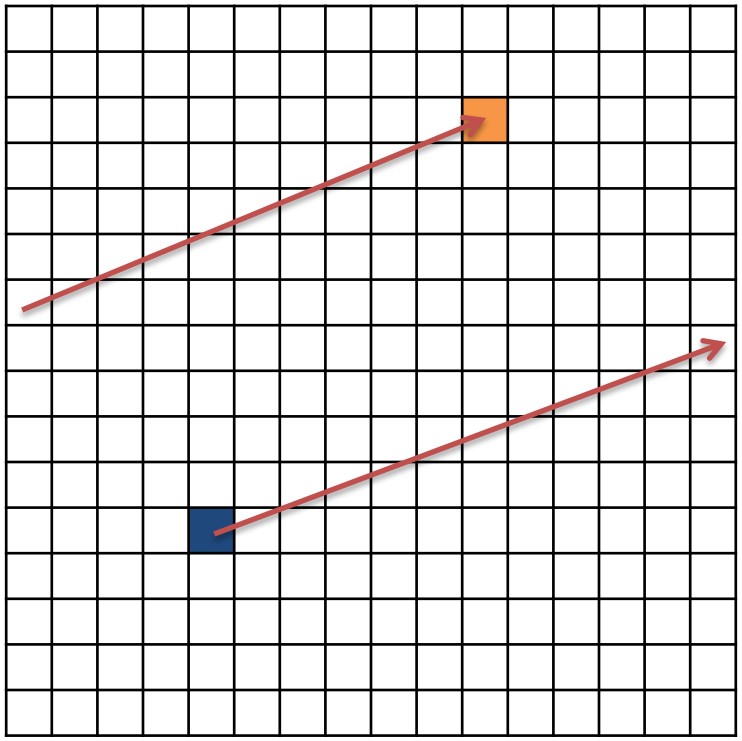


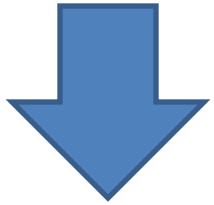
Initial Results

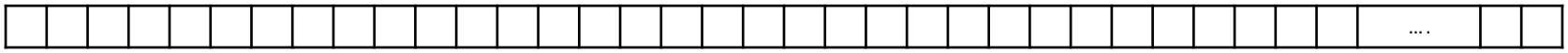
- 100,000 generations
- 10,000 copies per generation
- p_a: survival probability of agent $a \in A$
- $\max_{a \in \mathrm{A}} p_a = 0.0742$



Best Static Strategy







Exploitation

- $\frac{\#\ of\ possible\ steps}{total\ leniarized\ space} = \frac{20}{256} = 0.078125$
- $p_a > 0.078125 \rightarrow a$ interprets observation
- Modify environment to better target agents capable of interpretation

Improved Results

- $O(10^8)$ generations: $\max_{a \in A} p_a = 0.097$
- They must be interpreting the target coordinates

Occam's Razor

The simplest explanation that fits the facts.

- Robert Heinlein

Bayes Theory

- The most probable hypothesis h given observations d

$$p(h|d) = \frac{p(d|h)p(h)}{p(d)}$$

Information Theory

- Information identifies x out of possible $\mathcal{X}$
- Data encodes information

The more probable x the less information required

Information Theory

- Data-information relation:

$$H(\mathcal{X}) = -\sum_{x \in \mathcal{X}} p(x) \log_2\big(p(x)\big)$$

- Example:
  - $\mathcal{C}$: $\{Male, Female\}$
  - $\mathcal{X}$: $\{Height, ColorPreference\}$

Descriptive Complexity Theory

- $K(x)$: minimum resources required to compute x.
- Description: data + interpretation
- Example:
  - Compression: Information ← WinRar → Data
  - Description: Information ← Any program → Data

Fractal

Illustrate

Analogy Learner Theory

"The best analogy between two cases is the one which minimizes the amount of description necessary for the derivation of the source from the target."

Minimum Message Length

- Find more probable hypotheses by finding shorter message
  - Bayes:

    most probable explanation given data
  - Shannon:

    more probable explanations require less data
  - Kolmogorov:

    min description is related to computation and data

Minimal Message Length

- First Result of description complexity:
  - description complexity cannot be described
- $length(h \wedge d) = -\log\big(p(h \wedge d)\big) = -\log\big(p(h)\big) - \log\big(p(d|h)\big)$
- $-\log\big(p(h)\big)$: encodes the hypothesis
- $-\log\big(p(d|h)\big)$: encodes information

Conclusion

- Statistically invariant
- Scale invariant
- Account for measurement precision
- Hypotheses independent
- Nearly optimal converge to true hypotheses

Pros

- Very logical, everything has a reason for being in the paper
- Prepares the reader to understand complicated papers

Cons

- Missing performance or technical comparison
- Missing generalization and categorization
- No recent work is considered