On AGI architecture

December 18, 2019 mid-term report

YKY

Independent researcher, Hong Kong

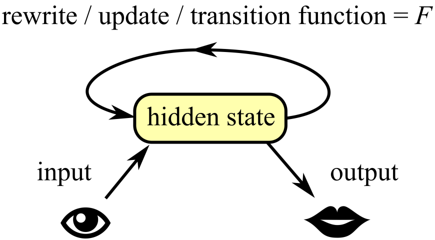
generic.intelligence@gmail.com

Table of contents

Hello friends ☺️
I have made some intermediate progress recently, which I share below.
I am also looking for collaborators.

The simplest AGI architecture

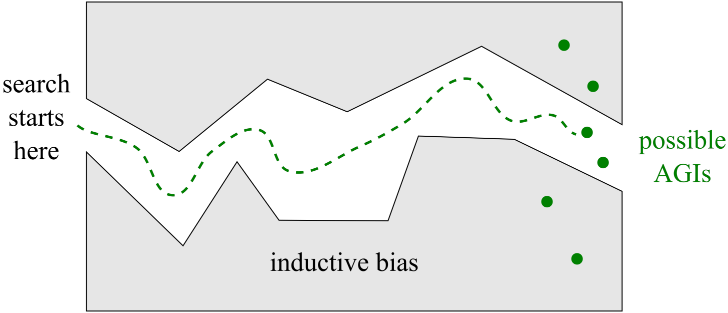
- The simplest AGI architecture consists of a single recurrent loop:

 (1)

- It operates under reinforcement learning, maximizing rewards by the Bellman optimality condition
- The transition function F can be implemented by a neural network
- According to No Free Lunch theorem, problem with this architecture is lack of inductive bias, learning is too slow

Some musings on No Free Lunch (NFL)

- According to NFL, there is no such things as “good” or “bad” inductive bias
- As long as it accelerates learning, and still accomodates AGI, it is good bias

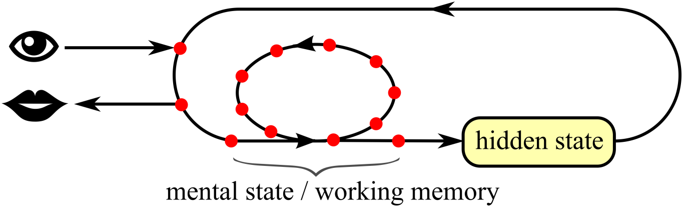


(2)

- For example, the neural network F can be made sparse while preserving deepness
- Yet, I proposed earlier to use logic as bias. Is that redundant? 😖

“Double loop” architecture

- Assume working memory consists of disparate propositions (•), residing in an inner loop. As this loop is iterated, the propositions are condensed into the hidden state. Hence the “double-loop” architecture

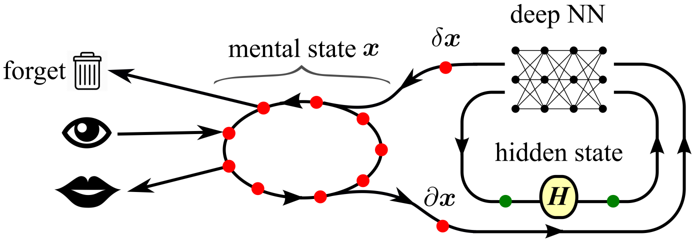


(3)

- This same architecture may be shared by the human brain, as its structure is simple and thus could have been evolved
- As far as I know, BERT also contains an implicit recurrence, where words in a sentence are condensed into a hidden state, from which target words are generated one by one
- So it seems that if we modify BERT to be a “double loop”, we can get an AGI:

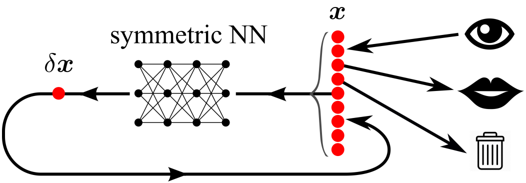
$$\square \xrightarrow{\text{loop 1}} \square \xrightarrow{\text{loop 2}} \boxed{\text{mental state}} \tag{4}$$

- Diagram (3) is a bit inaccurate, here is a more detailed diagram:



(5)

- In contrast, one can use a symmetric NN to make the architecture even simpler:



(6)

- But it's not easy to decide whether (5) or (6) is faster 😖

Connection between reinforcement learning & quantum mechanics

- **Bellman**

$$\boxed{\text{Bellman}} \quad S_t(x) = \max_u \{L(x,u) + \gamma S_{t+1}(x)\} \tag{7}$$

- Bellman **Hamilton-Jacobi** 1970s Kalman Pontryagin Hamilton-Jacobi-Bellman (HJB)

$$\boxed{\text{Hamilton-Jacobi}} \quad \frac{\partial S(x,t)}{\partial t} = -H \tag{8}$$

- Lagrangian L Hamiltonian H

$$L = \text{K.E.} - \text{P.E.} \quad , \quad H = \text{K.E.} + \text{P.E.} \tag{9}$$

K.E. = kenetic energy, P.E. = potential energy.

-

- Hamilton-Jacobi Schrödinger exact $\Psi = e^{-i\hbar S}$

$$\boxed{\text{Hamilton-Jacobi}} \qquad \frac{\partial S}{\partial t} = -H \qquad \Longrightarrow \qquad i\hbar\frac{\partial \Psi}{\partial t} = H\Psi \quad \boxed{\text{Schrödinger}} \tag{10}$$

- (quantization) [Field 2010]
- AI Hilbert Schrödinger
- Schrodinger (imaginary time) (diffusion)

$$\boxed{\text{wave eqn.}} \qquad \frac{\partial \Psi}{\partial t} + i\Delta\Psi = 0 \qquad \leftrightsquigarrow \qquad \frac{\partial u}{\partial t} + \Delta u = 0 \quad \boxed{\text{heat eqn.}} \tag{11}$$

- AI discrete Laplacian Δ discrete Schrödinger operators graph graph
- 😖

Topos theory

Topos Lawvere 1950s

topos (objects)

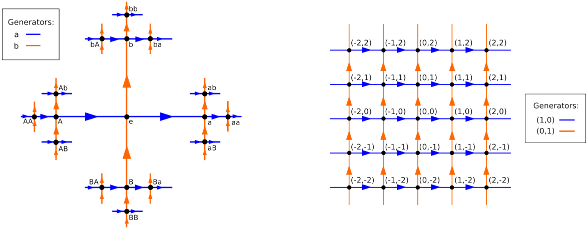
- Cartesian product $A \times B$ $A \wedge B$
- exponentiation $A \longrightarrow B = B^A$ $A \Rightarrow B$
- subobject classifier $A \hookrightarrow B$ $A \subseteq B$

Topos relation graphsalgebras

F $U \xrightarrow{F} V$ exponentiation V^U U, V $A \times B$ embed $A \times B = B \times A$ Abel

Permutation invariance

- Word2Vec Poincaré disc / hyperbolic space [Nickel and Kiela 2017]
- vectors hyperbolic space
- F_2 Cayley   Cayley



(12)

- F_n Cayley hyperbolic disc F_n Abelianization $F_n^{\mathrm{Ab}} \cong \mathbb{Z}^n = \mathbb{Z} \times \ldots \mathbb{Z}$ n- grid
- (representation theory) Abel   1- F_n^{Ab} n 1-

- $\mathbb{Z}^n$ fractal fractals
- weights-sharing permutation invariant (= Symmetric NN)
- activation function = polynomial
- 1-2 = 2
- sparse bias
-

Field, JH (2010). “Derivation of the Schrödinger equation from the Hamilton-Jacobi equation in Feynman’s path integral formulation of quantum mechanics”. In: *European Journal of Physics*.

Nickel, Maximillian and Douwe Kiela (2017). “Poincaré embeddings for learning hierarchical representations”. In: *Advances in neural information processing systems*, pp. 6338–6347.

☺️