О термодинамических ограничениях искусственного интеллекта

М.В.Алтайский [1]

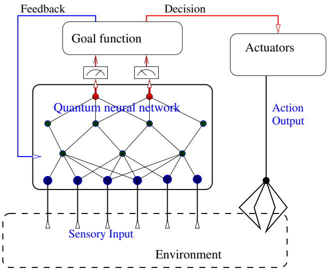
[1]ИКИ РАН

Oct 8, 2021

according to:
M.V.Altaisky and N.E.Kaputkina,
Thermodynamic restrictions on artificial intelligence based on quantum systems, *pp.14 -17 in Procceedings of V Scientific School "Dynamics of Complex Networks and their Applications", Kaliningrad, Sep 13-15, 2021, IEEE. [BF-NAICS 2021]*

Abstract

We discuss the possibility of emerging conscious behaviour in open quantum systems based on the definition of consciousness as the capability of a system to model the behaviour of its environment and act accordingly to the model. According to the laws of thermodynamics the entropy of the learning system and its environment cannot decrease, unless a quantum measurement is performed on it by external consciousness. This prohibits the emergence of conscious behaviour in artificial systems. The consideration is illustrated by numerical examples.



Subjects

- Strong AI - 'интеллект человеческого уровня'
- Intelligent agents (IA) - автономные агенты
- Thermodynamics of classical stochastic learning
- Quantum machine learning
- Thermodynamics of quantum machine learning
- Thermodynamic restrictions on quantum AI

Energy budget of AI systems

Turing test

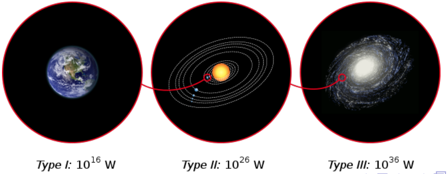
AI is *strong* if it cannot be distinguished from a human by means of interrogation using a computer keyboard.

... no system has passed the Turing test as yet.

Energy budget of civilizations

Kardashev scale:

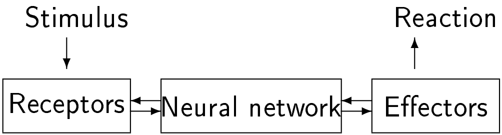
1. Planetary civilzation
2. Stellar civilization
3. Galactic civilization



[Picture from Wikipedia]

Intelligent agents (IA) functionality

- Recording facility – input devices enabling the perception of information from the environment.
- Discriminating facility – a neural network or an algorithmic system, which classifies the data perceived, builds a model of environment, and controls the actuators.
- Control facility – it stores the perceptions into the memory as *individual knowledge*, and directs the learning of discriminating facility by rewards or by other means.
- Actuators – devices, acting upon environment according to signals from discriminating facility.



Let $\{w\}$ be set of weights of a neural network. The weight-update rule for supervised learning can be written in differential form:

$$\dot{w}(t) = -w(t) + f(w(t), \xi, \sigma_T, t) + \zeta(t),$$

where $f(\cdot)$ represents a learning algorithm, ξ are vectors from the training set, and σ_T are the true labels for these vectors, $\zeta(t)$ is Gaussian noise. Considering a single neuron ought to classify N-dimensional vectors into two classes, labelled by $\sigma_T = \pm 1$, as a linear adder, the recognition of any new vector $\vec{\xi}$ is performed by a stochastic process

$$\sigma = F(A[w,\xi], \cdot) = \pm 1, \quad A[w,\xi] = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \xi_k w_k,$$

The algorithm $f(\cdot)$ ought to be chosen to maximize the probability of $F(A[w,\xi^{(\mu)}], \cdot) = \sigma_T^{(\mu)}$, where μ labels the data in the training set.

Thermodynamics of classical stochastic learning

Having the learning completed, we need to know how much the generated label σ reduces our uncertainty about the true label σ_T for the shown vector ξ:

$$I(\sigma:\sigma_T) \equiv S(\sigma_T) - S(\sigma_T|\sigma), \quad S(X|Y) := -\sum p(x,y)\log\frac{p(x,y)}{p(y)}$$

The efficiency of learning is given by the ratio of the mutual information $I(\sigma:\sigma_T)$ to the total entropy production S. Goldt и U. Seifert. "Stochastic Thermodynamics of Learning". в: *Phys. Rev. Lett.* 118 (2017), с. 010601:

$$\eta = \frac{I(\sigma:\sigma_T)}{\Delta S(w) + \Delta Q}, \quad \Delta S(w) = S(w(0),0) - S(w(t),t),$$

where ΔQ is the heat dissipated by the weight tuning process, and $S(w,t)$ is the Shannon entropy of the marginalized distribution $p(w,t) = \sum_{\sigma_T,\sigma} p(\sigma_T,w,\sigma,t)$.

Open quantum systems

Closed quantum system obeys the Schrödinger equation

$$\imath\hbar\frac{\partial}{\partial t}|x\rangle = \hat{H}|x\rangle$$

For an open system

$$|\psi\rangle = \sum_{x,y} c_{xy}|x\rangle|y\rangle$$

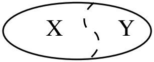
For an *observable* A measured on a system X:

$$\langle\psi|\hat{A}|\psi\rangle = \sum c^*_{x'y'}c_{xy}\langle x'|\hat{A}|x\rangle\langle y'|y\rangle$$
$$= A_{xx'}\rho_{xx'} = \mathrm{Tr}\hat{A}\hat{\rho}, \quad \hat{\rho} = cc^\dagger$$

in view of orthogonality $\langle y|y'\rangle = \delta_{yy'}$ of states in the unobserved space Y

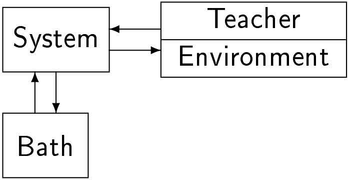
Density matrix ρ obeys the von Neumann equation:

$$\imath\hbar\frac{\partial\hat{\rho}}{\partial t} = [\hat{H},\hat{\rho}].$$



Thermodynamically our approach is similar to

I. A. Luchnikov и др. “Machine Learning Non-Markovian Quantum Dynamics”. в: *Phys. Rev. Lett.* 124 (14 2020), с. 140502



Quantum machine learning

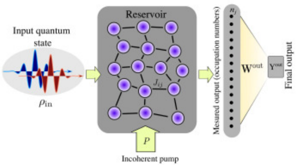
Neural networks (e^N)

- Nonlinear activation function
- Massive parallelism of synaptic connections
- Open dissipative system working at room temperature

Quantum computers (N^P)

- Linear operators acting on quantum states
- Quantum parallel processing of superposed states
- Unitary evolution in a closed system preserves coherence

Data:
Classical
or
Quantum



Picture from S. Ghosh и a.o. "Quantum reservoir processing". в: *npj Quantum Inf.* 5 (2019), c. 35

Output:
Classical
or
Quantum

Quantum spin (S) learns the direction of 'classical' spin (T)

Classical spin $\rho_T = \begin{pmatrix} p_\uparrow & 0 \\ 0 & p_\downarrow \end{pmatrix}$ is equivalent to a constant magnetic field. Its interaction with a spin qubit (S) is given by the Hamiltonian

$$H_{ST} = -\frac{w}{2}\hat{\sigma_z}, \quad \text{where } w = W\langle \sigma_T \rangle.$$

The dynamic of a single Ising spin can be then described by a Lindblad- Gorini-Kossakowski-Sudarshan- type master equation see e.g., H. Breuer и F. Petruccione. *The theory of open quantum systems*. Oxford University Press, 2002:

$$\frac{d\rho}{dt} = \frac{\imath\Omega}{2}[\sigma_x, \rho(t)] + \frac{\imath w}{2}[\sigma_z, \rho(t)] + \gamma_0(n+1)\Big(\sigma_-\rho\sigma_+ - \\ - \frac{1}{2}\{\sigma_+\sigma_-, \rho(t)\}\Big) + \gamma_0 n\Big(\sigma_+\rho\sigma_- - \frac{1}{2}\{\sigma_-\sigma_+, \rho(t)\}\Big),$$

where n is the number of quanta in the reservoir B, and γ_0 is the reservoir coupling constant.

Substituting

$$\rho(t) = \begin{pmatrix} a(t) & \xi(t) + \imath\eta(t) \\ \xi(t) - \imath\eta(t) & 1 - a(t) \end{pmatrix}$$

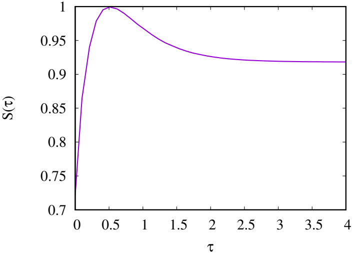
into the master equation we get

$$\frac{d\xi}{d\tau} = -\xi - w'\eta,$$
$$\frac{d\eta}{d\tau} = -\eta + w'\xi - \Omega' a + \frac{\Omega'}{2},$$
$$\frac{da}{d\tau} = -2a + \Omega'\eta + r$$

The entropy $S = -\mathrm{Tr}\rho\log_2\rho$

$$S = -x_+ \log_2 x_+ - x_- \log_2 x_-$$

$$x_\pm = \frac{1}{2} \pm \sqrt{\frac{1}{4} + a^2 - a + \xi^2 + \eta^2}$$



Entropy $S = S(\tau)$, calculated for arbitrary values of parameters $\Omega' = 1/3$, $r = 2/3$, $w' = 1/5$, and initial conditions $\xi(0) = \eta(0) = 0$, $a(0) = 0.8$

where $\tau = t\gamma_0\left(n + \frac{1}{2}\right)$ is dimensionless time, $w' = \frac{w}{\gamma_0\left(n+\frac{1}{2}\right)}$ is renormalised 'magnetic field', $\Omega' = \frac{\Omega}{\gamma_0\left(n+\frac{1}{2}\right)}$ is renormalized Rabi frequency, $r = \frac{n}{n+\frac{1}{2}}$

Learning on general classical data ($\xi_i^{(\mu)} = \pm 1, \sigma_T^{(\mu)} = \pm 1$)

Let the training set consist of P classical data vectors $\xi^{(\mu)} = (\xi_1^{(\mu)}, \dots, \xi_N^{(\mu)}; \sigma_T^{(\mu)})$, not affected by any interaction. Its density matrix is diagonal $\hat{\rho}_T = \sum_{\xi=0}^{2^{N+1}-1} P(\xi)|\xi\rangle\langle\xi|$. The initial density matrix of the whole system is the direct product $\hat{\rho}(0) = \hat{\rho}_S \otimes \hat{\rho}_T \otimes \hat{\rho}_B$. The evolution of the learning system is given by the von Neumann equation traced over T and B:

$$\imath\hbar\dot{\rho}_S(t) = \mathrm{Tr}_{T,B}[H, \hat{\rho}(t)].$$

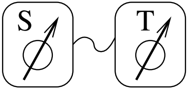
The trace over bath degrees of freedom is reduced to the weighted sum with the probabilities P_μ of each vector from the training set:

$$\imath\hbar\dot{\rho}_S(t) = \sum_{\mu=1}^{P} P_\mu \mathrm{Tr}_B[H_\mu, \rho_{SB}(t)],$$

$$H_\mu = -\lambda\sigma_T^\mu \hat{\sigma}_z^0 - \sum_{i=1}^{N} \xi_i^\mu \hat{\sigma}_z^i + \sum_k \sum_{i=0}^{N} \lambda_k^i (b_k^\dagger + b_k)\hat{\sigma}_z^i$$

$$-\frac{1}{2}\sum_{i=0}^{N} \epsilon_i \hat{\sigma}_z^i - \frac{1}{2}\sum_{i\neq j} J_{ij}\hat{\sigma}_z^i\hat{\sigma}_z^j - \frac{1}{2}\sum_{i=0}^{N} \Omega_i \hat{\sigma}_x^i + \sum_k \omega_k b_k^\dagger b_k$$

Quantum system learns on quantum data?



$$\hat{H} = -\frac{\Omega}{2}\hat{\sigma}_x^S - \frac{w}{2}\hat{\sigma}_z^S - \frac{J}{2}\hat{\sigma}_z^S\hat{\sigma}_z^T - \frac{\Delta}{2}\sigma_z^T.$$

The notion of "learning on quantum data" is not well defined. An attempt to learn the quantum state of the teacher results in quantum evolution of both S and T. The information S gains from T is given by the *mutual information*

$$I[\rho_S : \rho_T] := S[\rho_S(t)] + S[\rho_T(t)] - S[\rho(t)],$$

$$\rho_S(t) = \mathrm{Tr}_T \rho(t), \qquad \rho_T(t) = \mathrm{Tr}_S \rho(t)$$

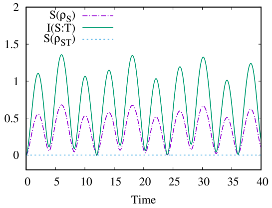
are the partial density matrices,

$S[\rho] = -\mathrm{Tr}\rho \log_2 \rho$.

$S[\rho_S]$ is entanglement entropy

At the absence of the heat bath the entropy of the combined system is conserved, but S becomes entangled with T:

$\rho(t) = e^{-\imath \hat{H} t}\rho(0)e^{\imath \hat{H} t}$.



The entropy oscillates synchronously with the mutual information between S and T.

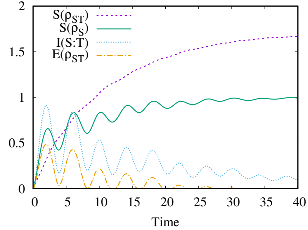
The system S starts from the state $\frac{|0\rangle+|1\rangle}{\sqrt{2}}$, the teacher T starts from the state $\frac{|0\rangle-|1\rangle}{\sqrt{2}}$. The parameters of the Hamiltonian are

Quantum learning at the presence of heat bath

At the presence of external environment, the heat bath B, connected to the learning system, the entropy of the total system (S+T) grows faster than the entropy of S. The evolution of entropies, entanglement and mutual information were calculated according to the master equation:

$$
\begin{aligned}
\frac{d\rho}{dt} = \quad & \imath[\frac{\Omega}{2}\hat{\sigma}^S_x + \frac{w}{2}\hat{\sigma}^S_z + \frac{J}{2}\hat{\sigma}^S_z\hat{\sigma}^T_z + \frac{\Delta}{2}\sigma^T_z, \rho(t)] \\
+ \quad & \gamma_0(n+1)\left(\sigma^S_-\rho\sigma^S_+ - \frac{1}{2}\{\sigma^S_+\sigma^S_-, \rho(t)\}\right) \\
+ \quad & \gamma_0 n\left(\sigma^S_+\rho\sigma^S_- - \frac{1}{2}\{\sigma^S_-\sigma^S_+, \rho(t)\}\right),
\end{aligned}
$$

$E[\rho_{ST}]$ C.H.Bennet et al. PRA 54(1996)3824



Evolution of the entropy $S[\rho_{ST}]$, mutual information, and entanglement of formation $E[\rho_{ST}]$, calculated for the density matrix $\rho_{ST}(t)$. Initially the system is in the state $\frac{|0\rangle+|1\rangle}{\sqrt{2}}$ state, the teacher starts its evolution from the state $\frac{|0\rangle-|1\rangle}{\sqrt{2}}$. The parameters of the Hamiltonian are: $\Omega = 1.0, w = 0.2, J = 1.0, \Delta = 0.$

Thermodynamic restrictions on learning

At the assumption of an infinite heat bath B at equilibrium at a temperature T, the entropy production by the combined system $S+T$ is non-negative:

$$\sigma = \frac{dS_{ST}}{dt} + J \geq 0, \quad J = -\frac{1}{T}\frac{d}{dt}\mathrm{Tr}(H\rho_{ST}),$$

where J is the entropy flux from the system $S+T$ to the heat bath B.

The non-negativity of the total entropy production in the system ST takes the form

$$\dot{S}_S + \dot{S}_T - \frac{1}{T}\frac{dE_{ST}}{dt} \geq \dot{I}(\sigma:\sigma_T), \quad E_{ST} \equiv \mathrm{Tr}(H\rho_{ST}),$$

Integrating the latter equation we get

Inequality

$$\Delta S_S + \Delta S_T + \Delta Q/T \geq \Delta I(\sigma:\sigma_T)$$

Impossibility of perpetual learning

1. $\Delta Q > 0$ – The system $S+T$ is heating the environment. The excess of energy of $S+T$ system is dissipated to the bath Bin the form of heat ΔQ:

$$\Delta E_{ST} = \Delta Q, \quad \Delta S_S + \Delta S_T + \Delta Q/T \geq \Delta I(\sigma : \sigma_T).$$

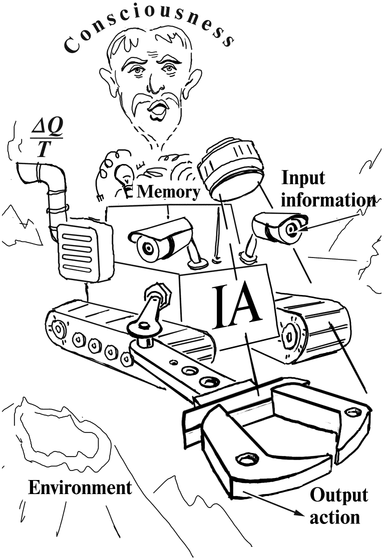
We can't extract more information ΔI than the energy ΔE spent.

2. $\Delta Q < 0$ – The system $S+T$ is embedded in a hot environment. The heat flux goes from B to $S+T$:

$$-\Delta Q/T = \Delta S_S + \Delta S_T - \Delta I(\sigma : \sigma_T) \geq 0.$$

Since the l.h.s. is positive, the gained ΔI is less than the sum of entropy increase in S and T. If the system ST relaxes to equilibrium with B, the heat flux should vanish, and so will do the mutual information $I(\sigma : \sigma_T)$.

Thank You for your attention!



Future space mission: Drawing by K.Zabusik