Grid Path Planning with Deep Reinforcement Learning: Preliminary Results

Alelsandr Panov, Roman Suvorov and Konstantin Yakovlev

Federal Research Center "Computer Science and Control"
Russian Academy of Sciences (RAS)
National Research University Higher School of Economics
Moscow

August 2 – Fierces on BICA 2017

Sign based world model

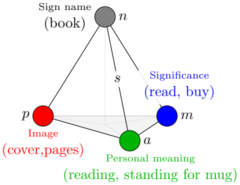
A component of knowledge representation is a sign:

- in sense of cultural-historical approach by L. Vygotsky,
- in sense of activity theory by A. Leontiev.

Sign based world model

A component of knowledge representation is a sign:

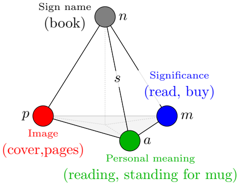
- in sense of cultural-historical approach by L. Vygotsky,
- in sense of activity theory by A. Leontiev.

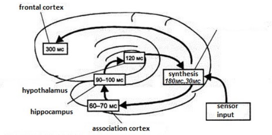


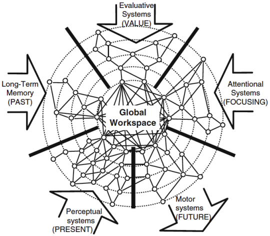
Sign based world model

A component of knowledge representation is a sign:

- in sense of cultural-historical approach by L. Vygotsky,
- in sense of activity theory by A. Leontiev.







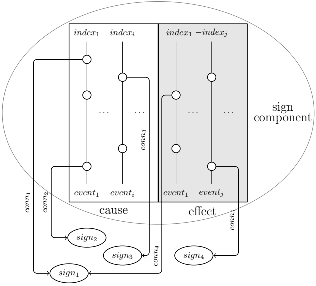
Supported ideas in psychology and biology:

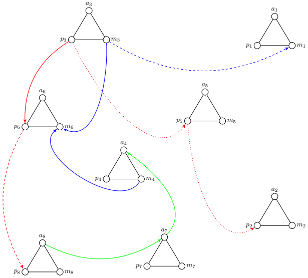
- neurophysiological data (Edelman, Ivanitsky, Mountcastle etc.),
- two and three levels psychological theories (Stanovich, Kahneman).

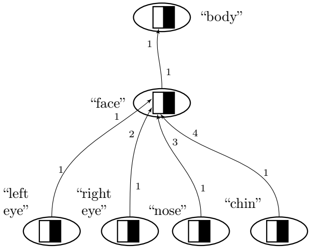
Osipov, G. S., A. I. Panov, and N. V. Chudova. "Behavior Control as a Function of Consciousness. II. Synthesis of a Behavior Plan". *Journal of Computer and Systems Sciences International*. 2015.

— ."Behavior control as a function of consciousness. I. World model and goal setting". *Journal of Computer and Systems Sciences International*. 2014.

Modelling of world model







Sign based world model (semiotic network):

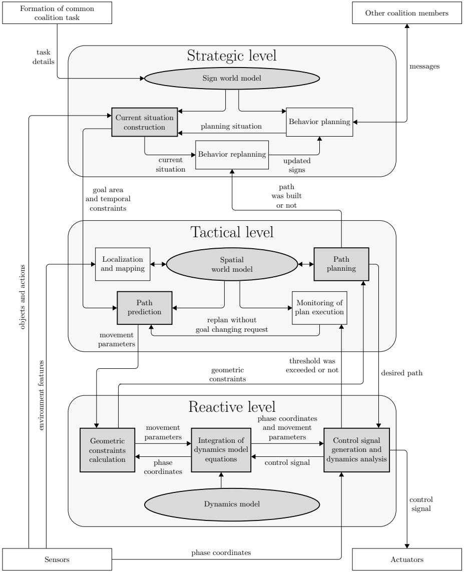
$$\Omega = \langle W_p, W_m, W_a, R_n, \Theta \rangle$$

Panov, Aleksandr I. “Behavior Planning of Intelligent Agent with Sign World Model”. *Biologically Inspired Cognitive Architectures*. 2017.

Osipov, Gennady S. “Signs-Based vs. Symbolic Models”. *Advances in Artificial Intelligence and Soft Computing*. 2015.

Applications

- Cognitive functions modeling and construction of models that explain psychological phenomena.
- Algorithm of synthesizing the plan of behavior (algorithms MAP, MultiMAP, GoalMAP).
- Solving symbol grounding and symbol anchoring problems.
- Reconstruction of sign based world model of the actor based on texts.
- Text generation based on specific world models (virtual assistants).
- Multi-level architectures of control (robotic systems).



Symbol anchoring in robotics

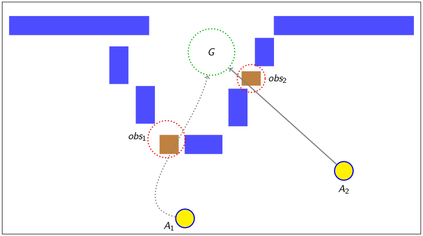
How to form symbols, concepts and signs on the basis of sensorimotor information:

- symbol grounding problem - Harnad, 1990; Barsalou, 1999, 2008; Sun, 2013;
- anchoring problem - Vernon, 2014; Karpov, 2016;
- semiotic schemas - Roy, 2005;
- stream model **DyKnow** - Heintz, 2010;
- **conceptors** - Jaeger, 2014;
- **SemLinks** system - Butz, 2016, 2017.

Learning rules of relocation

Features of the task:

- Using reinforcement learning to form components of the sign based world model.
- An agent uses "raw" sensory information as input data.
- The agent's task is to form a sign based world model as a result of learning: a certain conceptual description of the environment, including discrete rules of action in it.
- A broader formulation is the task of joint planning in a space with role distribution and communication in a coalition.



Reinforcement learning: general statement

General definitions:

- $a_t : s_t \to s_{t+1}$ - agent's actions in the environment,
- r_t - reward received by the agent from the environment,
- The agent's goal is to maximize the total reward $R = \sum_t \gamma^t r_t$, $0 < \gamma \leq 1$,
- $\pi : S \to A$ - agent's policy, taking into account previous experience and the need to study the environment: exploration and exploitation (ϵ-greedy method).

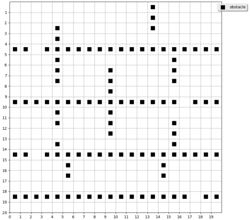
Solutions:

- If we know $T(s_t, a_t, s_{t+1})$ and $r(s_t, a_t)$, then this is a problem based on the model and we should solve a *Bellman equation*.
- Evaluation of value function $V(s) = \mathbf{E}[R|s, \pi]$ or action-value function $Q(s, a) = \mathbf{E}[R|s, a, \pi]$.

Sutton, Richard S. and Andrew G. Barto. *Reinforcement learning: An Introduction*. 2012.

Reinforcement learning: relocation rules

- $E = (M, G)$ - environment, where M - local map, $G(p_s, p_f)$ - an algorithm of reward generation,
- $a_t = p_t \to p_{t+1}$ - relocation actions of the agent,
- $s_t \in R^{(2d)^2}$ - agent's observation (sensory information).



Lets $Q^*(s_t, a_t) = \max_\pi \mathbf{E}[R|s_t, a_t, \pi]$ - optimal action-value function, then taking into account the definition of R we receive the Bellman equation:

$$Q^*(s, a) = \mathbf{E}_{s_t \sim E}\left[\boxed{r_t + \gamma \max_{a_t} Q^*(s_t, a_t)} \Big| s, a \right]$$

Reinforcement learning: approximation

To solve the Bellman equations by means the iterative methods it is used different approximations of the function $Q^*(s, a)$: $Q(s, a; \theta) \approx Q^*(s, a)$.

During the learning process parameters θ are adjusted as a result of minimization of the loss function $L(\theta)$:

$$L_i(\theta_i) = \mathbf{E}_{s,a\sim\rho(\cdot)} \left[\left(y_i - Q(s, a; \theta_i) \right)^2 \right],$$

$$y_i = \mathbf{E}_{s_t\sim E} \left[r_t + \gamma \max_{a_t} Q(s_t, a_t; \theta_{i-1}) | s, a \right]$$

$$\nabla_{\theta_i} L_i(\theta_i) = \mathbf{E}_{s,a\sim\rho(\cdot);s_t\sim E} \left[(r_t + \gamma \max_{a_t} Q(s_t, a_t; \theta_{i-1}) - \right.$$
$$\left. - Q(s, a; \theta_i)) \nabla_{\theta_i} Q(s, a; \theta_i) \right].$$

Reinforcement learning: replays

- An episode is a set of agent's actions and reactions of the environment to movements from the initial position to the final, or until the maximum number of actions N_a is reached ,
- $e_t = (s_t, a_t, r_t, s_{t+1})$ - a precedent saved into a memory D,
- Learning takes place according to some random sample e from the memory

$$D = \left\{ e_1, \boxed{e_2}, \ldots, \boxed{e_i}, e_{i+1}, \ldots e_j, \boxed{e_{j+1}}, \ldots \right\}$$

- One action can be used several times $\rightarrow$ expand the sample, eliminate the correlation of neighboring states.

Reward generation

The following algorithm was used to calculate the reward function:

$$G(s,g,t) = \begin{cases} \alpha_{opt} r_t^{opt} + \alpha_{rat} r_t^{rat} + \alpha_{euq} r_t^{euq}, & p_t \leftarrow 0, \\ r^{obs}, & p_t \leftarrow 1, \\ r^{tar}, & p_t = g, \end{cases}$$

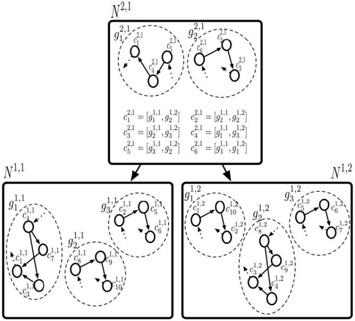
where

- $\sum \alpha_i = 1$ - normalization,
- $r_t^{opt} = l_t - l_{t-1}$ - changing optimal distance,
- $r_t^{rat} = e^{-l_t/l_0}$ - penalty for deviation from a goal,
- $r_t^{euq} = |p_t - g| - |p_{t-1} - g|$ - regularizer for straightening a path.

Heterarchical causal network

Biologically inspired learning model (formation of the sign component):

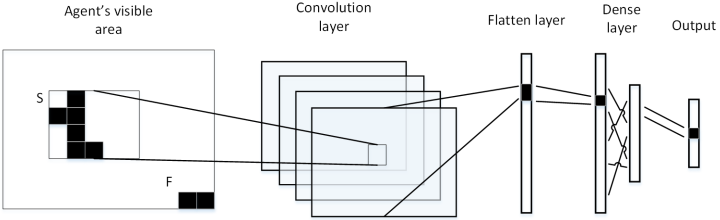
- scanning receptive field - pattern formation,
- spatial pooler (clasterization by online K-means),
- temporal pooler (agglomerative clasterization $\rightarrow$ Markovian chains).



Neuronal architectures

In the work we conducted experiments with various neural networks:

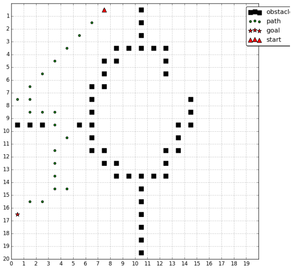
1. Ag_1 - a ¡¡shallow¿¿ fully-connected neural network,
2. Ag_2 - a convolution network of medium depth with a fully connected output layer,
3. Ag_3 - a deep network with Inception blocks.

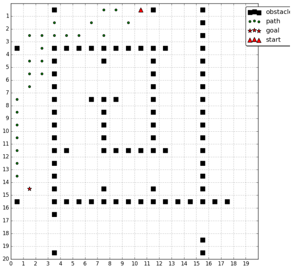


Maps, paths and parameters

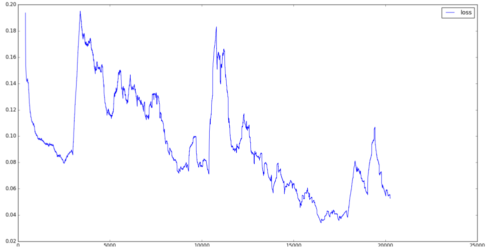
The most successful set of parameters:

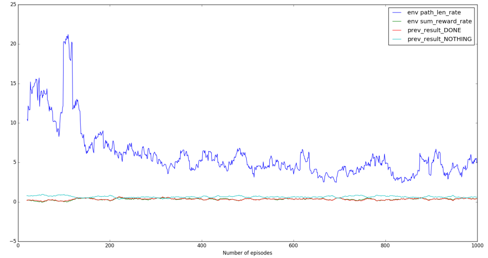
- $N_{ep} \sim 3000$ - number of episodes, $N_a = 100$ - maximum number of steps, $d = 20$ - observation radius of the agent,
- $\alpha_{opt} = 0.8, \alpha_{rat} = 0.1$ - reward coefficients,
- $r^{obs} = -4, r^{tar} = 10$ - reward parameters,
- $N_e = 10$ - size of the memory D,
- $\gamma = 4$ - discounting multiplier for reward function.





Convergence of the learning process





We used two quality metrics:

- M_p - the ratio of the path length found by the agent to the optimal,
- M_r - the ratio of total reward to its maximum value.

Thank you for your attention!

pan@isa.ru