

Автоматическое формирование правил перемещения с использованием обучения с подкреплением

Александр Панов и Роман Суворов

Федеральный исследовательский центр «Информатика и управление»
Российской академии наук

16 июня – САИТ 2017
г. Светлогорск, Россия



Картина мира субъекта деятельности

Картина мира субъекта деятельности - это представления субъекта о внешней среде, о своих собственных характеристиках, целях, мотивах, о других субъектах и операции (произвольные и произвольные), осуществляемые на основе этих представлений.

Картина мира субъекта деятельности

Картина мира субъекта деятельности - это представления субъекта о внешней среде, о своих собственных характеристиках, целях, мотивах, о других субъектах и операции (произвольные и произвольные), осуществляемые на основе этих представлений.

Элементом картины мира является знак:

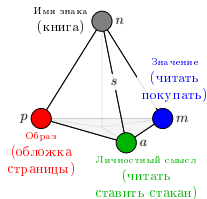
- в смысле культурно-исторического подхода Выготского-Лурии,
- выполняющий функции в соответствии с теорией деятельности Леонтьева.

Картина мира субъекта деятельности

Картина мира субъекта деятельности - это представления субъекта о внешней среде, о своих собственных характеристиках, целях, мотивах, о других субъектах и операции (произвольные и произвольные), осуществляемые на основе этих представлений.

Элементом картины мира является знак:

- в смысле культурно-исторического подхода Выготского-Лурии,
- выполняющий функции в соответствии с теорией деятельности Леонтьева.

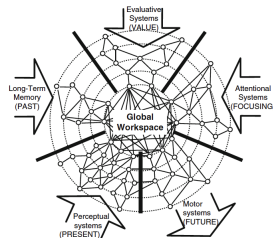
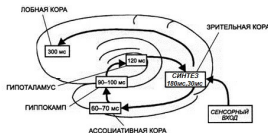
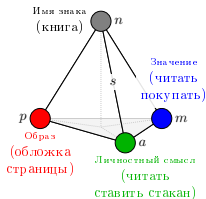


Картина мира субъекта деятельности

Картина мира субъекта деятельности - это представления субъекта о внешней среде, о своих собственных характеристиках, целях, мотивах, о других субъектах и операции (произвольные и произвольные), осуществляемые на основе этих представлений.

Элементом картины мира является знак:

- в смысле культурно-исторического подхода Выготского-Лурии,
- выполняющий функции в соответствии с теорией деятельности Леонтьева.



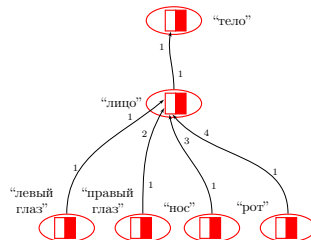
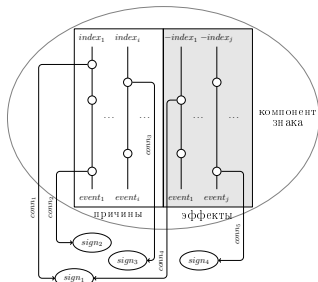
В пользу существования такой структуры свидетельствуют:

- нейрофизиологические данные (Эдельман, Иваницкий, Маунткастл и др.),
- другие психологические теории (например, трехкомпонентная модель Станович).

Осипов, Г. С., А. И. Панов и Н. В. Чудова. «Управление поведением как функция сознания. II. Синтез плана поведения». *Известия Российской академии наук. Теория и системы управления*. 2015.

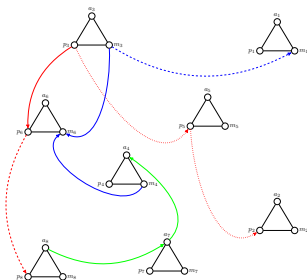
— «Управление поведением как функция сознания. I. Картина мира и целеполагание». *Известия Российской академии наук. Теория и системы управления*. 2014.

Моделирование картины мира



Модель картины мира
(семиотическая сеть):

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



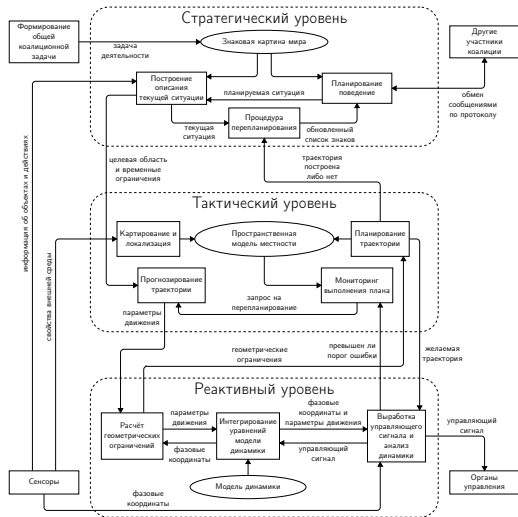
Panov, Aleksandr I. "Behavior Planning of Intelligent Agent with Sign World Model". *Biologically Inspired Cognitive Architectures*. 2017.



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

Применение модели

- Моделирование когнитивных функций и построение моделей, объясняющих психологические феномены.
- Алгоритмы синтеза плана поведения (алгоритмы MAP, MultiMAP, GoalMAP).
- Решение проблемы символизации.
- Построение картины мира субъекта на основе текстов.
- Генерация сообщений на основе картин мира определенного типа (виртуальные ассистенты).
- Построение многоуровневых архитектур управления.



Проблема символизации в робототехнике

Как на основе сенсомоторной информации сформировать символы, концепты, понятия, знаки:



- symbol grounding problem - Harnad, 1990; Barsalou, 1999, 2008; Sun, 2013;
- anchoring problem - Vernon, 2014; Карпов, 2016;
- семиотические схемы - Roy, 2005;
- потоковая модель DyKnow - Heintz, 2010;
- концепторы - Jaeger, 2014;
- система SemLinks - Butz, 2016, 2017.

Обучение правилам перемещения

Особенность задачи:

- Использование обучения с подкреплением для формирования компонент знаковой картины мира.
- В качестве знаний о среде агент использует «сырую» сенсорную информацию.
- Задача агента - в результате обучения сформировать картину мира: некоторое понятийное описание среды, включающее дискретные правила действия в нем.
- Более широкая постановка - задача совместного планирования в пространстве с распределением ролей и коммуникацией.

Обучение с подкреплением: общая постановка

Основные понятия:

- $a_t : s_t \rightarrow s_{t+1}$ - действия агента в среде,
- r_t - вознаграждение, получаемое агентом от среды,
- цель агента - максимизация суммарного вознаграждения
$$R = \sum_t \gamma^t r_t, 0 < \gamma \leq 1,$$
- $\pi : S \rightarrow A$ - стратегия агента, учитывающая предыдущий опыт и необходимость исследования среды (ϵ -жадный метод).

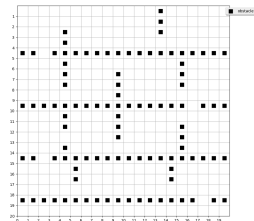
Способы решения:

- Если известны $T(s_t, a_t, s_{t+1})$ и $r(s_t, a_t)$, то это задача, основанная на модели, решение *уравнения Беллмана*.
- Оценка функции полезности $V(s) = \mathbf{E}[R|s, \pi]$ или функции полезности действия $Q(s, a) = \mathbf{E}[R|s, a, \pi]$.

Саттон, Р.С. и Э. Г. Барто. *Обучение с подкреплением*. 2011.

Обучение с подкреплением: правила перемещения

- $E = (M, G)$ - среда, где M - карта местности, $G(p_s, p_f)$ - алгоритм генерации вознаграждения,
- $a_t = p_t \rightarrow p_{t+1}$ - действия агента по перемещению,
- $s_t \in R^{(2d)^2}$ - наблюдения агента (сенсорная информация).



Пусть $Q^*(s_t, a_t) = \max_{\pi} \mathbf{E}[R | s_t, a_t, \pi]$ - оптимальная функция полезности, тогда с учетом определения R получаем следующее уравнение Беллмана:

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

Обучение с подкреплением: аппроксимация

Для решения итерационными методами уравнения Беллмана используют различные аппроксимации функции $Q^*(s, a)$:

$$Q(s, a; \theta) \approx Q^*(s, a).$$

В процессе обучения происходит настройка параметров θ в результате минимизации функции потерь $L(\theta)$:

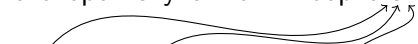
$$L_i(\theta_i) = \mathbf{E}_{s, a \sim \rho(\cdot)} \left[\left(y_i - Q(s, a; \theta_i) \right)^2 \right],$$

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

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

Обучение с подкреплением: переигровки

- Эпизод - это набор действий агента и реакций среды на перемещения от начального положения до конечно, либо до достижения максимального количества действий N_a ,
- $e_t = (s_t, a_t, r_t, s_{t+1})$ - прецедент сохраняется в память агента D ,
- обучение идет по некоторой случайной выборке e из памяти

$$D = \{e_1, e_2, \dots, e_i, e_{i+1}, \dots, e_j, e_{j+1}, \dots\}$$


- одно действие можно использовать несколько раз \rightarrow расширяем выборку, устраняем корреляции соседних состояний.

Генерация вознаграждения

Для расчета функции вознаграждения использовали следующий алгоритм:

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

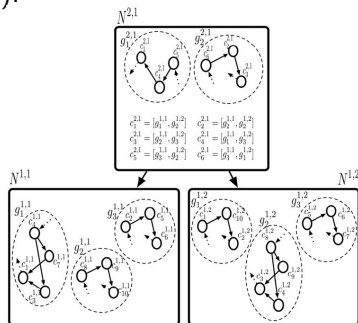
где

- $\sum \alpha_i = 1$ - нормировка,
- $r_t^{opt} = l_t - l_{t-1}$ - изменение оптимального расстояния,
- $r_t^{rat} = e^{-l_t/l_0}$ - штраф за отклонение от цели,
- $r_t^{euq} = |p_t - g| - |p_{t-1} - g|$ - регуляризатор для спрямления пути.

Гетерархическая каузальная сеть

Биологически правдоподобная модель обучения (формирования компонента знака) включает:

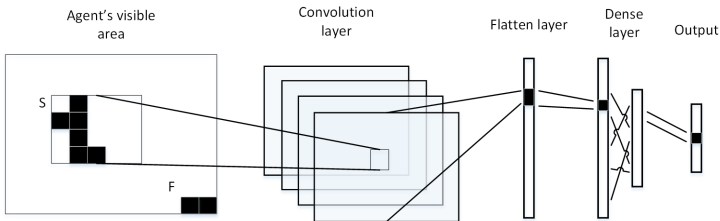
- сканирующее рецептивное поле - формирование паттерна,
- пространственный группировщик (кластеризация паттернов online K-means),
- временной группировщик (агломеративная кластеризация → марковские цепи).



Нейросетевые архитектуры

В работы мы проводили эксперименты с различными нейронными сетями:

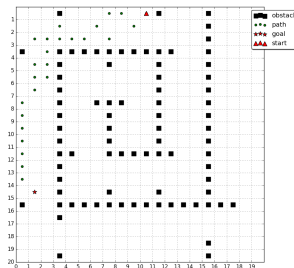
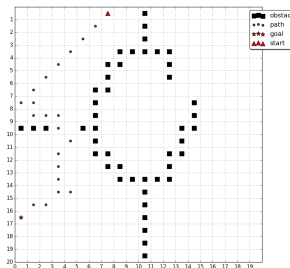
- 1 Ag_1 - «мелкая» полносвязная нейронная сеть,
- 2 Ag_2 - сверточная сеть средней глубины с полносвязными выходным слоем,
- 3 Ag_3 - глубокая сеть, состоящая из блоков Inception.



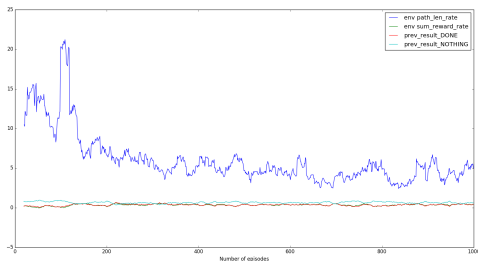
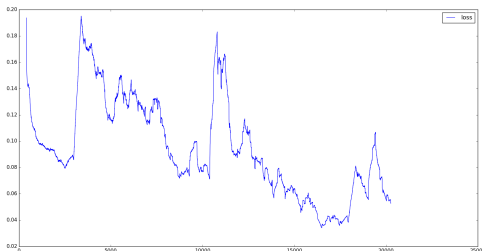
Карты, пути, параметры

Самый успешный наборов параметров:

- $N_{ep} \sim 3000$ - количество эпизодов, $N_a = 100$ - ограничение по шагам, $d = 20$ - радиус видимости агента,
- $\alpha_{opt} = 0.8, \alpha_{rat} = 0.1$ - значения коэффициентов вознаграждения,
- $r^{obs} = -4, r^{tar} = 10$ - значения параметров вознаграждения,
- $N_e = 10$ - размер памяти D ,
- $\gamma = 4$ - дисконтирующий множитель при расчете вознаграждения.



Сходимость процесса обучения



Мы использовали две метрики качества:

- M_p - отношение длины пути, построенного агентом, к оптимальному,
- M_r - отношение суммарного вознаграждения к максимальному.

Спасибо за внимание!

pan@isa.ru

ФИЦ ИУ РАН, лаб. 0-2