

Занятие 10

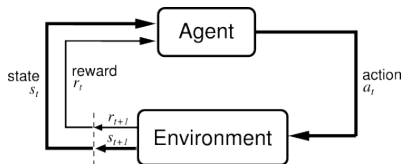
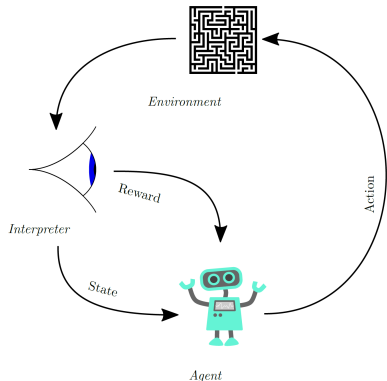
Обучение с подкреплением и иерархическая временная
память

МФТИ
ФИЦ ИУ РАН

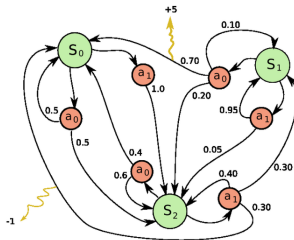
10 мая, 2018

- 1 Обучение с подкреплением
- 2 Иерархическая временная память

Обучение с подкреплением: постановка задачи



- t - дискретные моменты времени,
- $a_t \in A$ - действие агента в момент времени t ,
- E - среда, $s_t \rightarrow s_{t+1}$ - состояния среды,
- r_t - вознаграждение, поступающее от среды,
- $R = \sum_t \gamma^t r_t$ - суммарное вознаграждение,
 $0 < \gamma \leq 1$ - дисконтирующий множитель.



Марковский процесс (Markov decision process (MDP)) - кортеж $\langle S, A, P, R \rangle$:

- S - конечное число состояний,
- A - конечное число действий
- $P = \{P_a(s, s') = P(s_{t+1} = s' | s_t = s, a_t = a)\}$ - вероятности переходов,
- $R = \{r_a(s, s')\}$ - вознаграждение.

Цель агента - обучиться стратегии π выбора действия в наблюдаемых состояниях среды $\pi : S \rightarrow A$, в результате применения которой он получит максимальное суммарное вознаграждение R :

$$\sum_t \gamma^t r_t \rightarrow \max_{\pi}$$

Баланс между исследованием среды и учетом предыдущего опыта (exploration vs exploitation) - ε -жадный метод:

- с вероятностью $1 - \varepsilon$ выбирается действие на основе предыдущих прецедентов,
- с вероятностью ε - случайное действие из доступных на данный момент

Параметр ε уменьшают с течением времени.

- Если известны $P(s_t, a_t, s_{t+1})$ и $R(s_t, a_t)$, то это задача, основанная на модели, решение *уравнения Беллмана*:

$$V(s) = \max_a Q(s, a),$$

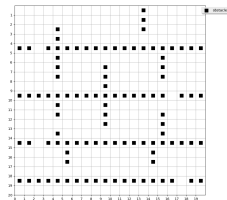
$$Q(s, a) = \sum_{s_{t+1}} P(s_t, a_t, s_{t+1}) (R(s_t, a_t) + \gamma V(s_{t+1})),$$

где $V(s) = \mathbf{E}[R|s, \pi]$ - функция полезности, а
 $Q(s, a) = \mathbf{E}[R|s, a, \pi]$ - функция полезности действия.

- Оценка функций полезности $V(s)$ или $Q(s, a)$.

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 $L(\theta)$:

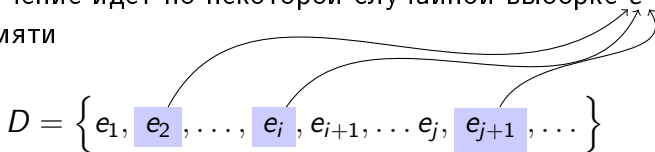
$$L_i(\theta_i) = \mathbf{E}_{s, a \sim \rho(\cdot)} \left[(y_i - Q(s, a; \theta_i))^2 \right],$$

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

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

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



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

Для расчета функции вознаграждения использовали следующий алгоритм:

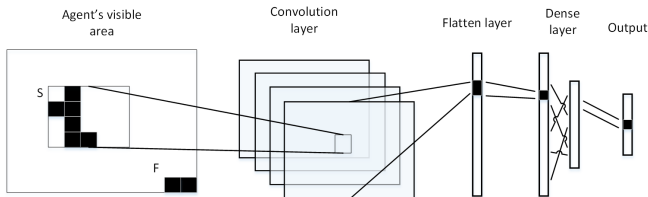
$$G(s, g, t) = \begin{cases} \alpha_{opt} r_t^{opt} + \alpha_{rat} r_t^{rat} + \alpha_{euq} r_t^{euq}, & p_t \leftarrow 0, \\ r_t^{obs}, & p_t \leftarrow 1, \\ r_t^{tar}, & p_t = g, \end{cases}$$

где

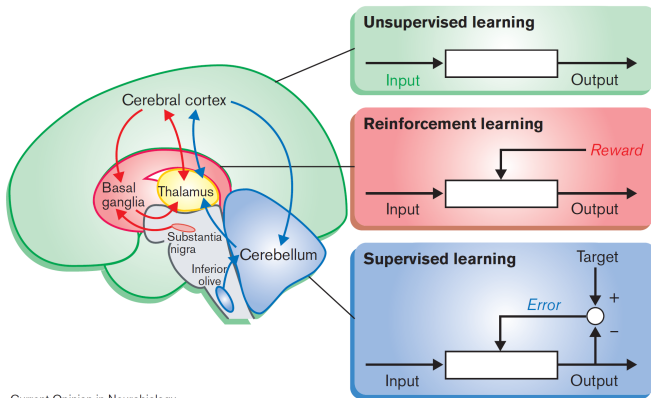
- $\sum \alpha_i = 1$ - нормировка,
- $r_t^{opt} = l_t - l_{t-1}$ - изменение оптимального расстояния,
- $r_t^{rat} = e^{-l_t/l_0}$ - штраф за отклонение от цели,
- $r_t^{euq} = |p_t - g| - |p_{t-1} - g|$ - регуляризатор для спрямления пути.

Можно использовать разные варианты сетей:

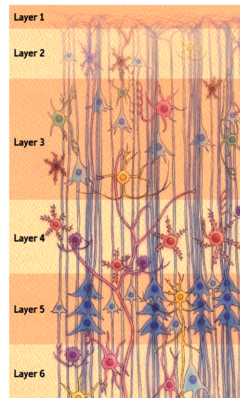
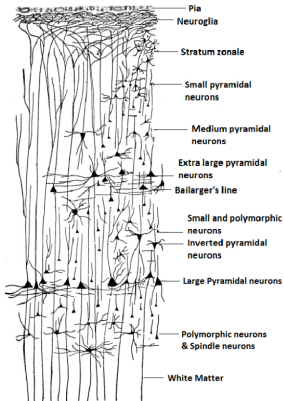
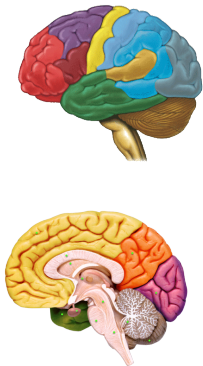
- 1 Ag_1 - «мелкая» полносвязная нейронная сеть,
- 2 Ag_2 - сверточная сеть средней глубины с полносвязными выходным слоем,
- 3 Ag_3 - глубокая сеть, состоящая из блоков Inception.



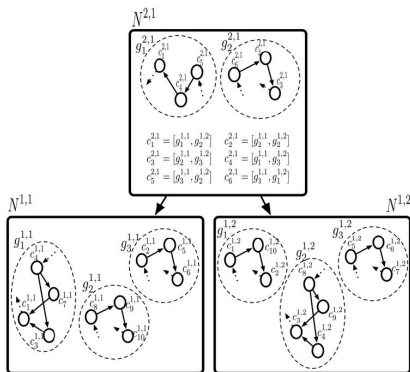
Модели обучения в мозге



Histological Structure of the Cerebral Cortex

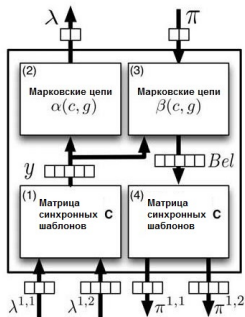
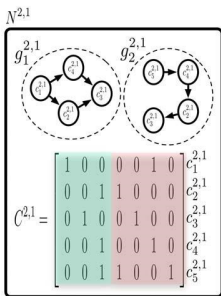


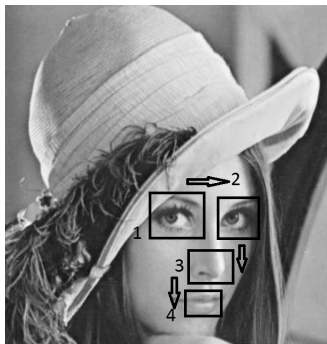
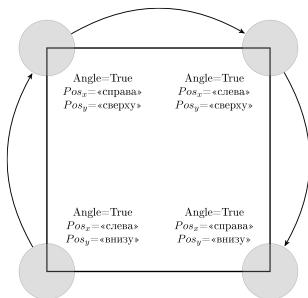
Иерархическая временная память



- $N^{i,j}$ - узлы сети,
- $g^{i,j}$ - временные группы,
- $c_k^{i,j}$ - паттерн (синхронные шаблоны).

Иерархическая временная память



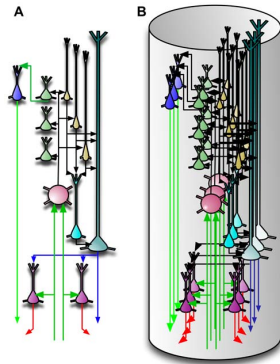
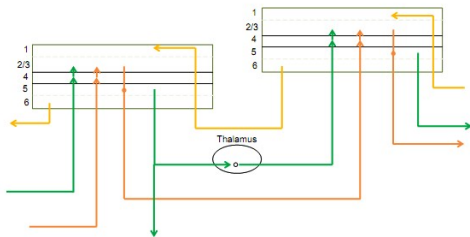


$$Z = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

К основным принципам работы механизма обучения относятся:

- использование иерархии вычислительных узлов с восходящими и нисходящими связями,
- использование Хэббовских правил обучения,
- разделение пространственного и временного группировщиков,
- подавление второстепенной активации для формирования разреженного представления.

Нейронная организация



Темы проектов, курсовых и дипломных работ

- Принцип гарантированного управления и его применение в нелинейных задачах.
- Приближенное решение задачи нелинейного управления на конечном интервале регулирования и его применение.
- Фильтрация в нелинейных системах.
- Сравнение эффективности различных нелинейных законов управления.
- Методы и алгоритмы эвристического поиска.
- Планирование траектории в виртуальных мирах как задача эвристического поиска пути на графе особой структуры.
- Планирование траектории беспилотных транспортных средств: методы и алгоритмы.
- Планирование траектории перемещения в пространстве положений для сложных робототехнических систем (манипуляторов и др.).
- Задачи картирования и локализации для беспилотных транспортных средств.
- Картирование и локализация для беспилотных летательных аппаратов по данным инерциальной навигационной системы и видеопотоку.
- Методы и алгоритмы выделения особенностей на растровых изображениях (SURF, SIFT и др.).
- Методы и алгоритмы обработки изображений для робототехники.
- Обучение с подкреплением по модели Actor-Critic.
- Иерархическое обучение с подкреплением.
- Глубокое обучение с подкреплением.
- Обучение с подкреплением для манипулятора по видеопотоку.
- Нейросимвольные способы представления и приобретения знаний.
- Планирование поведения в коалиции агентов с распределением ролей.
- Нейробайесовские методы обучения.

Спасибо за внимание!