NEURAL NETWORKS REINFORCEMENT LEARNING

Кетков С.

Кондратьев Н.

Макарова О.

Прибыткина Д.

Семенов С.

Part 2

Q-Learning

Q-learning (Q-обучение) — метод, применяемый в искусственном интеллекте при агентном подходе. На основе получаемого от среды вознаграждения агент формирует функцию полезности **Q**, что впоследствии дает ему возможность уже не случайно выбирать стратегию поведения, а учитывать опыт предыдущего взаимодействия со средой. Применяется для ситуаций, которые можно представить в виде марковского процесса принятия решений.
Одно из преимуществ Q-обучения — то, что оно в состоянии сравнить ожидаемую полезность доступных действий, не формируя модели окружающей среды.

Algorithm

Пусть агент находится в какой-то среде, у которой есть текущее состояние ***s***, и в каждый момент можно выполнить действие ***a*** из дискретного набора. Это действие переводит систему в новое состояние (стохастически, т.е. в системе есть всякое случайное) и может выдать **reward** (вознаграждение) или закончить игру.

Algorithm

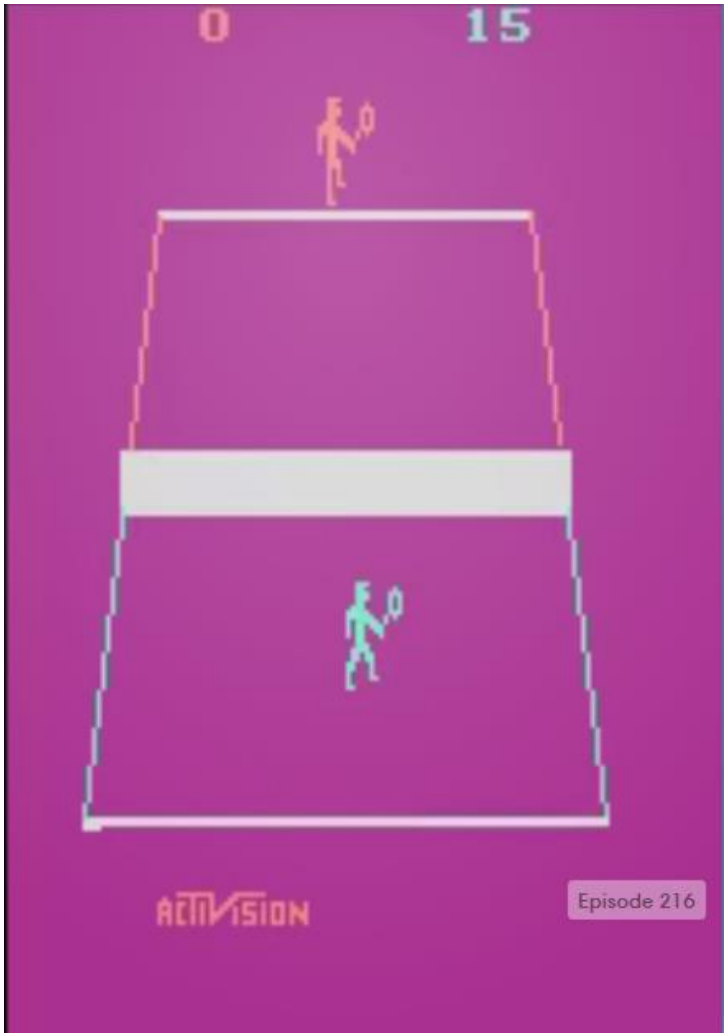
Введем понятие **cumulative return over time** — общее количество вознаграждений, которые можно получить от текущего момента до конца игры, причем будущие вознаграждения уменьшаются на ***y*** (коэффициент дисконтирования) за каждый период времени, т.е. вознаграждение в следующий момент времени — это ***y*r***, еще через один — y^2*r.

Оптимальное поведение можно описать некой функцией ***Q*(s,a)***.

Task definition

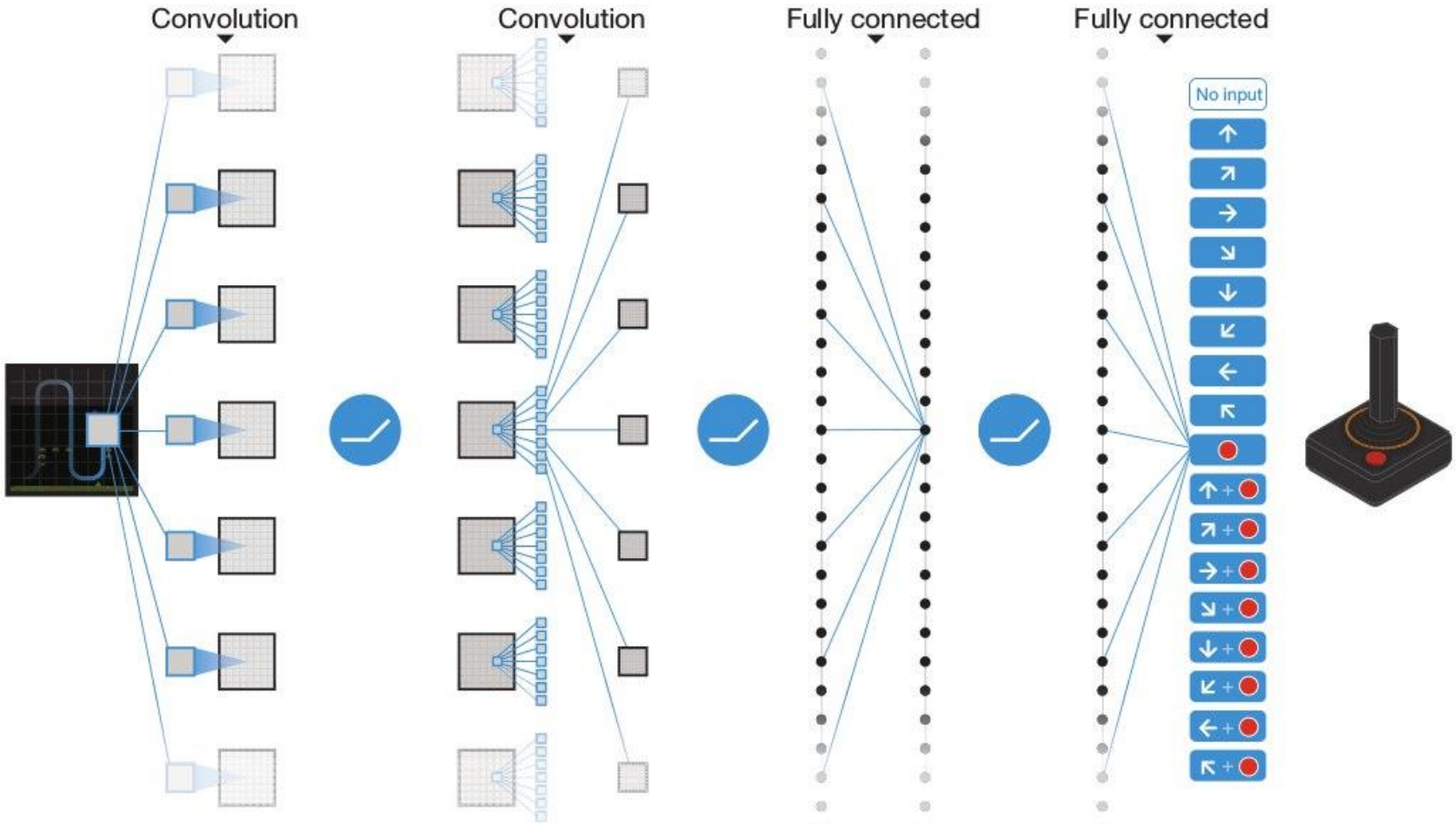
Tennis

На каждом шаге игры среда даёт RGB изображение размером 160 x 210 пикселей (ширина x высота). Агент может выполнить 18 действий.



Architecture

Schematic illustration of the convolutional neural network*



** Mnih, Volodymyr, et al. "Human-level control through deep reinforcement learning." Nature 518.7540 (2015): 529-533.*

Loss function

$$L_i(\theta_i) = \mathbb{E}_{(s,a,r,s') \sim \mathrm{U}(D)} \left[\left(r + \gamma \max_{a'} Q(s', a'; \theta_i^-) - Q(s, a; \theta_i) \right)^2 \right]$$