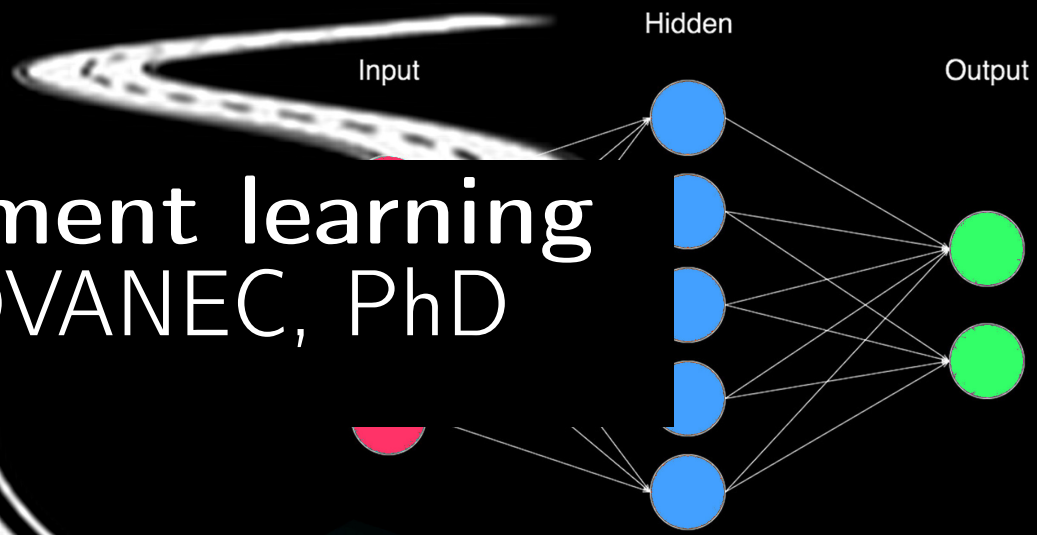
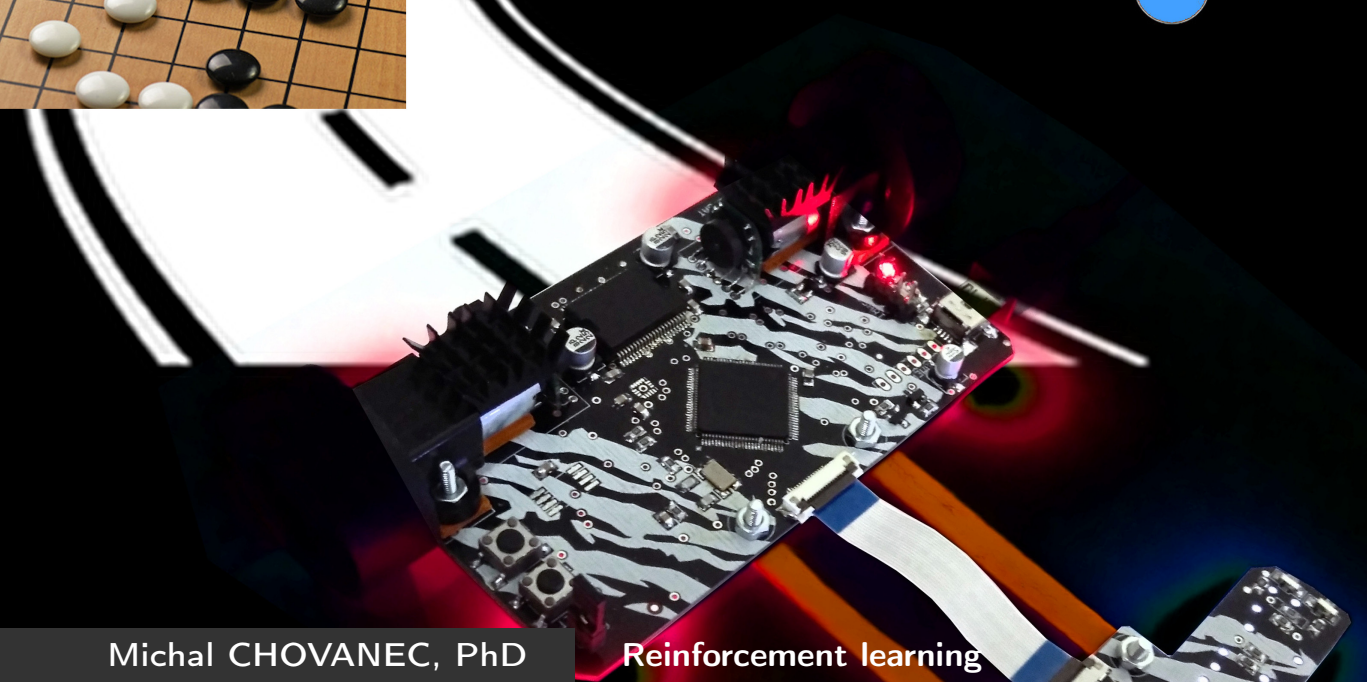
$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[r_{t+1} + \lambda \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right]$$

(The New Action Value = The Old Value) + The Learning Rate × (The New Information − the Old Information)



Reinforcement learning

Michal CHOVANEC, PhD

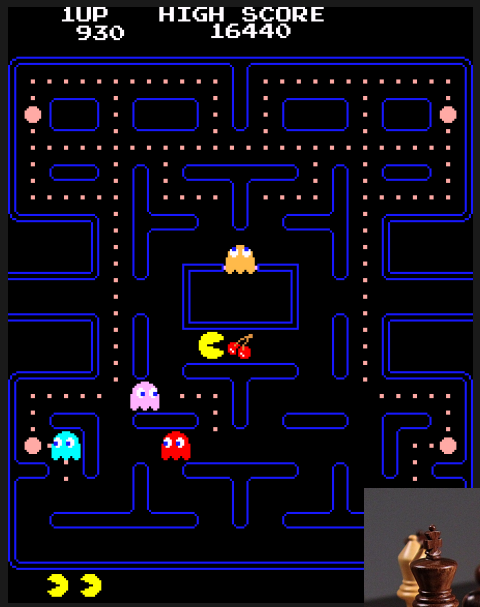


Reinforcement learning

"Master Kihara will listen to noise of fire"

Reinforcement learning

- learn from punishment and rewards
- learn to play a game with unknow rules



Reinforcement learning

1. supervised learning based on human player

$$action = F(observation)$$

$$loss = \left(required\_action - F(observation)\right)^2$$

2. RL method

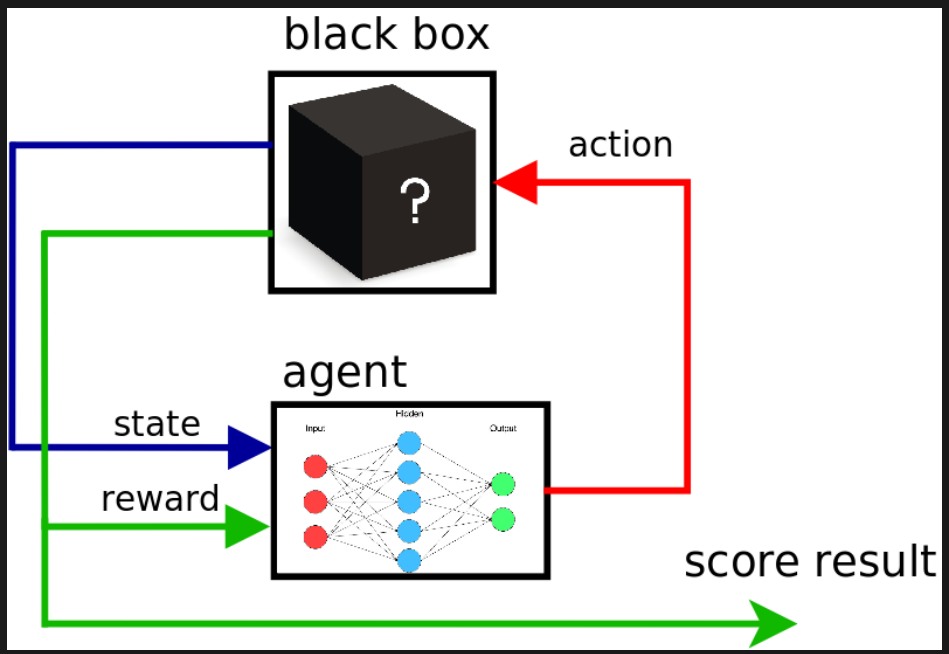
$$action = F(observation)$$

$$loss = \left(required\_action - F(observation)\right)^2$$

required_action = ?

Reinforcement learning

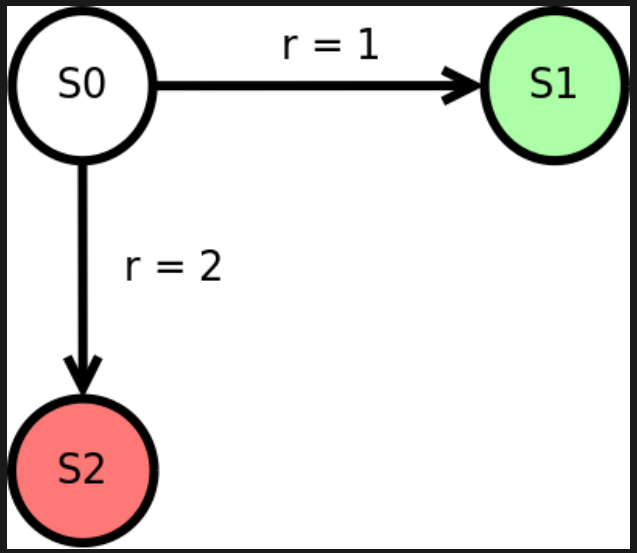
- obtain **state**
- choose **action**
- **execute** action
- obtain **reward**
- learn from **experiences**



Making decisions

two possible strategies

- strategy 1 : S0->S1, score = 1.0
- strategy 2 : S0->S2, score = 2.0



$$Q(s, a) = R(s, a)$$

where

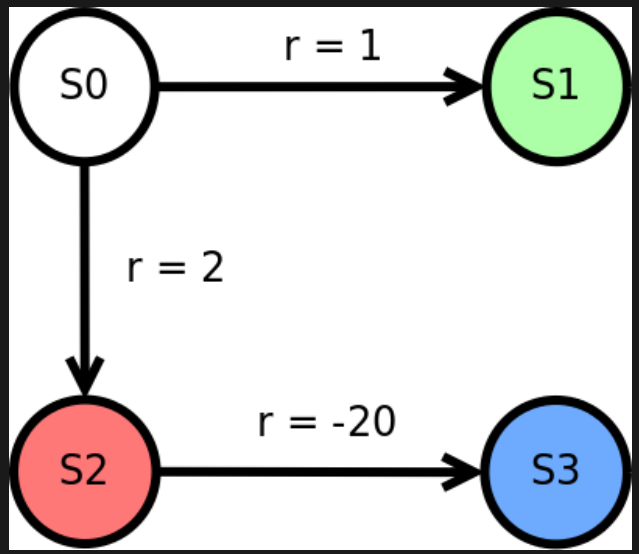
s is state

a is action

Making decisions

two possible strategies, greedy = trap

- strategy 1 : S0->S1, score = 1.0
- strategy 2 : S0->S2->S3, score = 2.0 + (-20.0) = -18

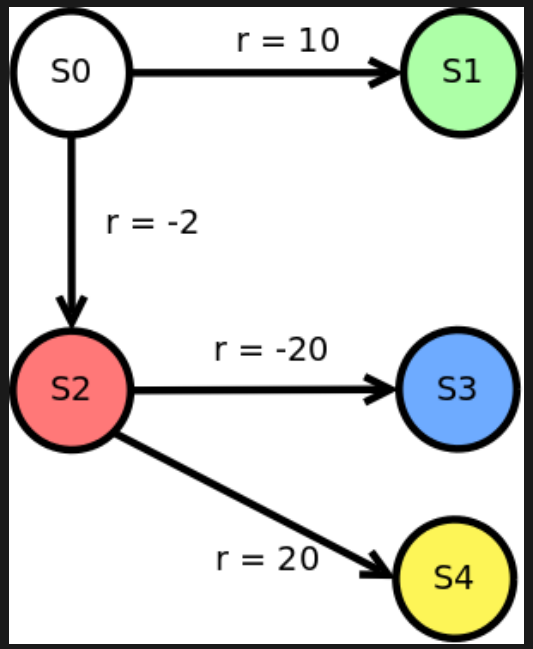


$$Q(s, a) = R(s, a) + FutureReward$$

Making decisions

three possible strategies

- strategy 1 : S0->S1, score = 10.0
- strategy 2 : S0->S2->S3, score = -2.0 + (-20.0) = -22
- strategy 3 : S0->S2->S4, score = -2.0 + (20.0) = 18



$$Q(s, a) = R(s, a) + max(FutureRewards)$$

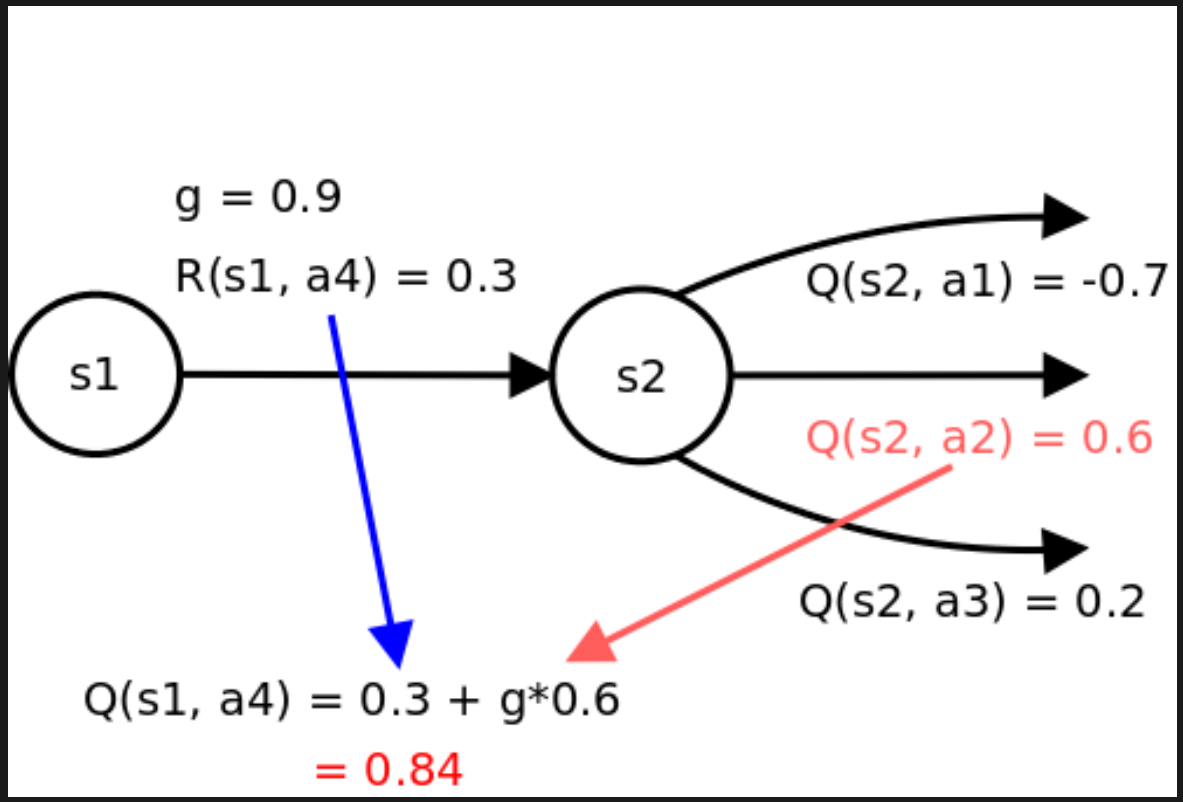
Q learning

$$Q(s,a) = R(s,a) + \gamma \max_{a'} Q'(s',a')$$

where
s is state
a is action
s' is next state
a' is best action in next state
$R(s,a)$ is reward
$\gamma \in \langle 0,1 \rangle$ is discount factor

Q learning

$$Q(s,a) = R(s,a) + \gamma \max_{a'} Q'(s',a')$$



Q learning - stochastic

$$Q(s,a) = R(s,a) + \gamma \max_{a'} Q'(s',a')$$

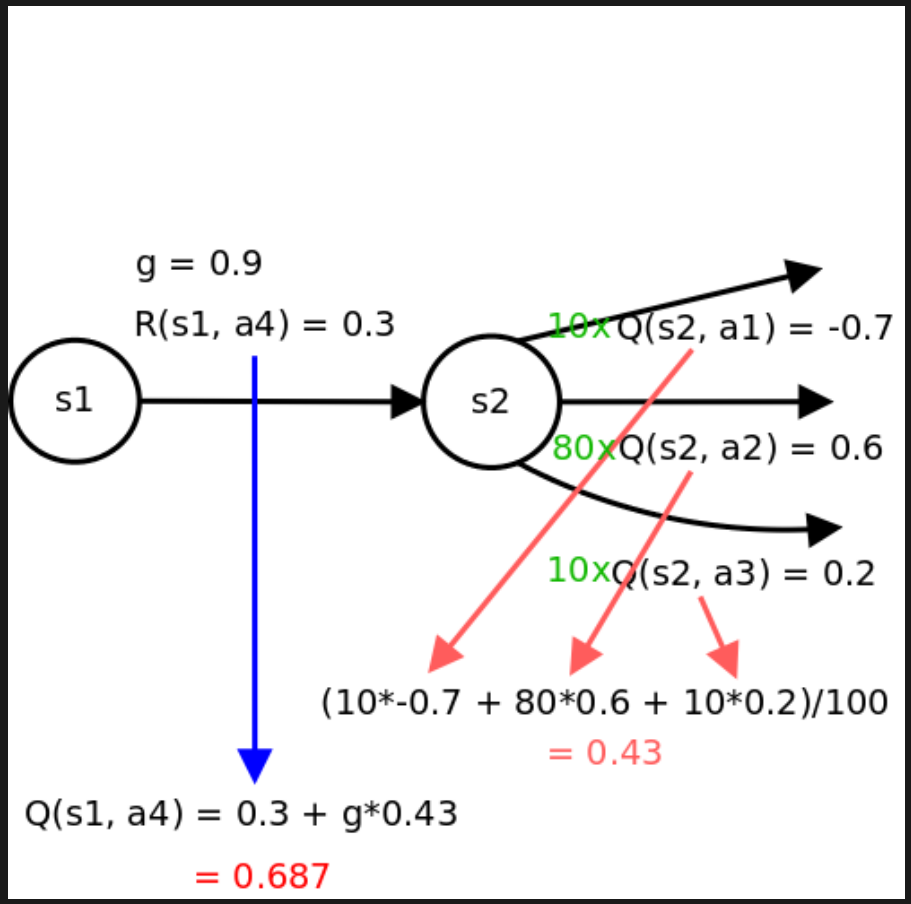
$$Q(s,a) = (1-\alpha)Q'(s,a) + \alpha\Big(R(s,a) + \gamma \max_{a'} Q'(s',a')\Big)$$

Q learning

- **obtain reward**
  reward = env.get_reward();
- **update state**
  state_old = state;
  state = env.get_observation();
- **select action**
  action_old = action;
  action = select_action(Q(state));
- **process learning**
  Q(state_old, action_old)+= α(reward + γ $\max_{a'}$Q(state, a') - Q(state_old, action_old));
- **execute action**
  env.action(action);

SARSA learning

$$Q(s,a) = (1-\alpha)Q'(s,a) + \alpha\Big(R(s,a) + \gamma Q'(s',a')\Big)$$



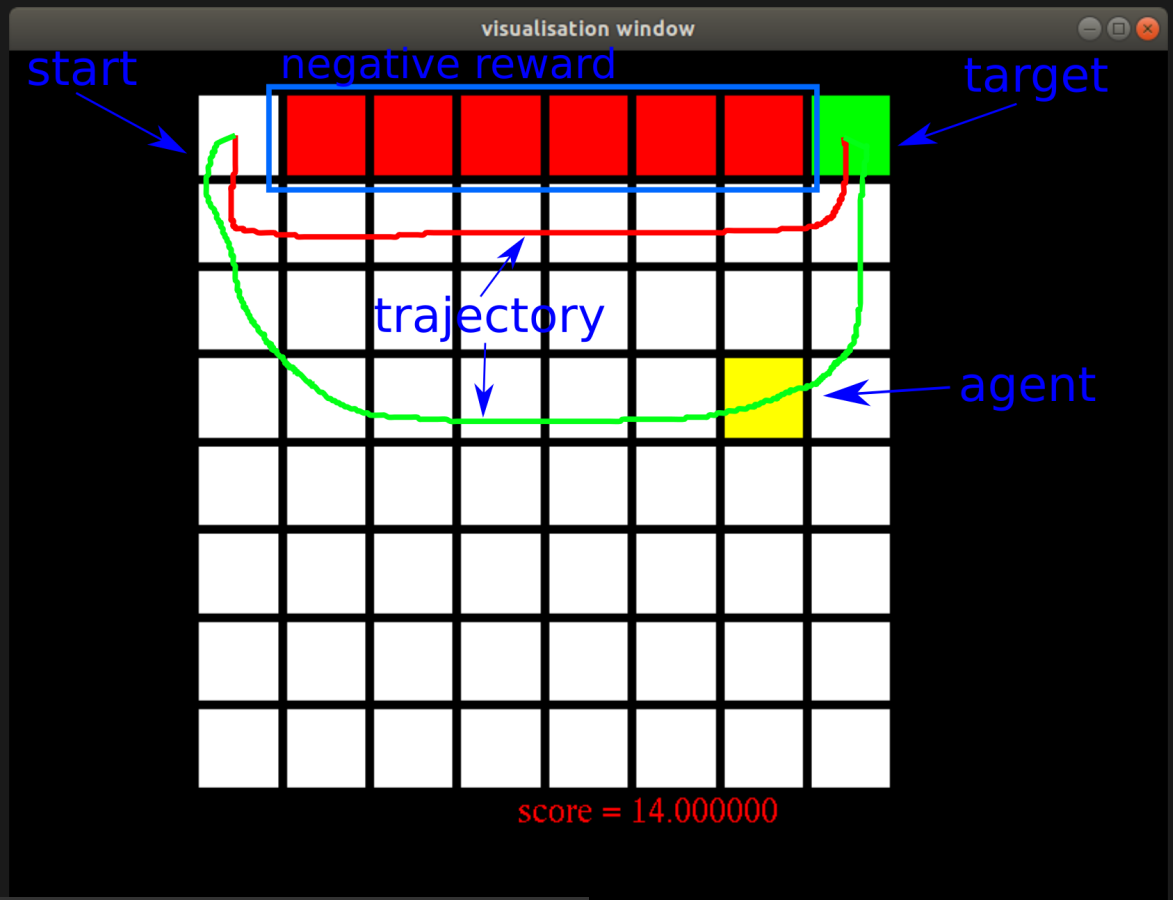
SARSA learning

- **obtain reward**
  reward = env.get_reward();
- **update state**
  state_old = state;
  state = env.get_observation();
- **select action**
  action_old = action;
  action = select_action(Q(state));
- **process learning**
  Q(state_old, action_old)+= α(reward + γQ(state, action) - Q(state_old, action_old));
- **execute action**
  env.action(action);

Q-learning vs SARSA - cliff example

$$Q(s,a) = (1-\alpha)Q'(s,a) + \alpha\Big(R(s,a) + \gamma \max_{a'} Q'(s',a')\Big)$$

$$Q(s,a) = (1-\alpha)Q'(s,a) + \alpha\Big(R(s,a) + \gamma Q'(s',a')\Big)$$



Deep reinforcement learning

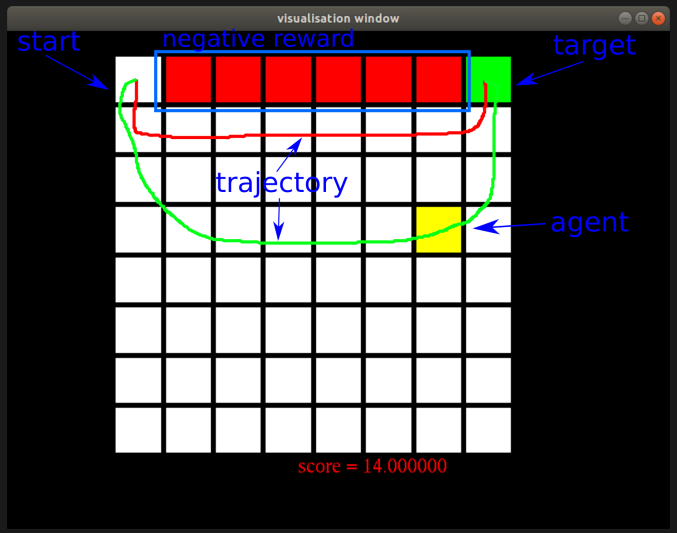
Possible states in some games

- **tic-tac-toe** 26830 states
- **2048** 44096709674720289 states
- **atoms in observable universe** 10^{82}
- **chess** 10^{120} states
- **GO** 10^{170} states

storing Q values

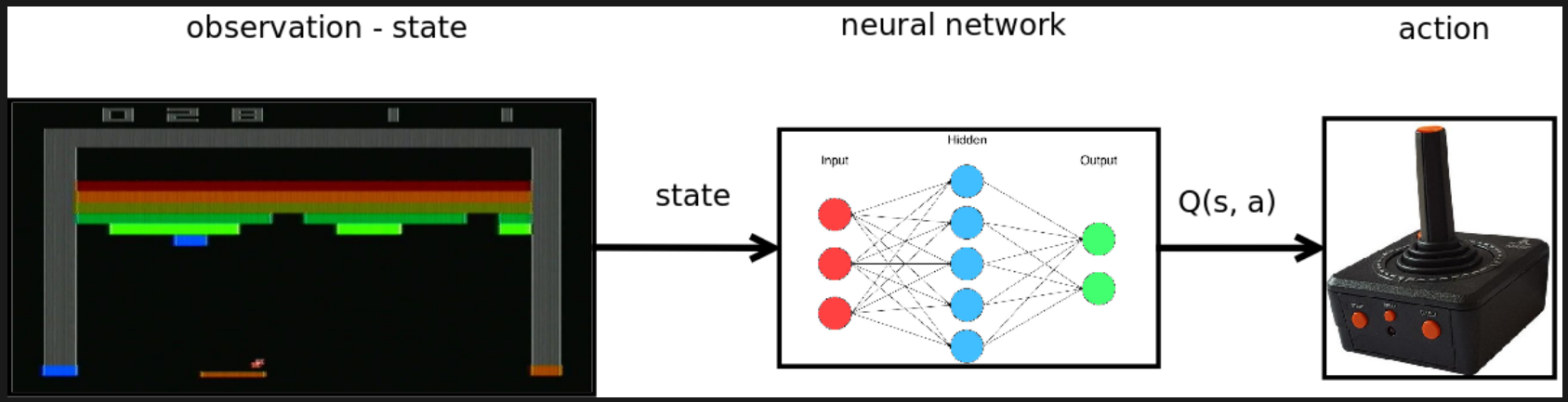
- table
- linear combination of features, $Q(s,a) = \sum_{i=1}^{N} w_i F_i(s,a)$
- neural network

Q table example

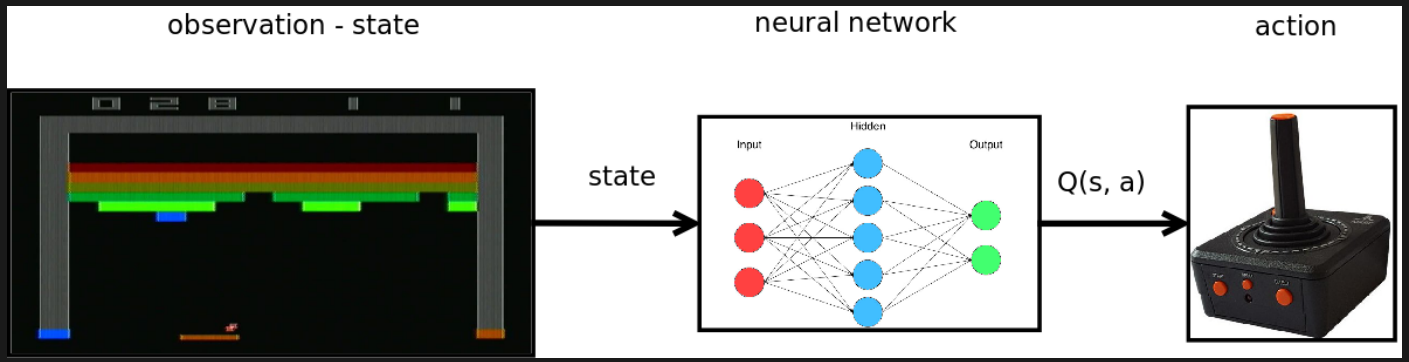


State	A0 UP	A1 LEFT	A2 DOWN	A3 RIGHT
0			0.43	-1
1, 2, 3, 4, 5, 6	T	T	T	T
7	T	T	T	T
8				0.49
9	-1			0.53
10	-1			0.59
11	-1			0.65
12	-1			0.72
13	-1			0.81
14	-1			0.9
15	1			

Neural network for $Q(s, a)$ approximation - Q networks



Neural network for $Q(s, a)$ approximation - Q networks



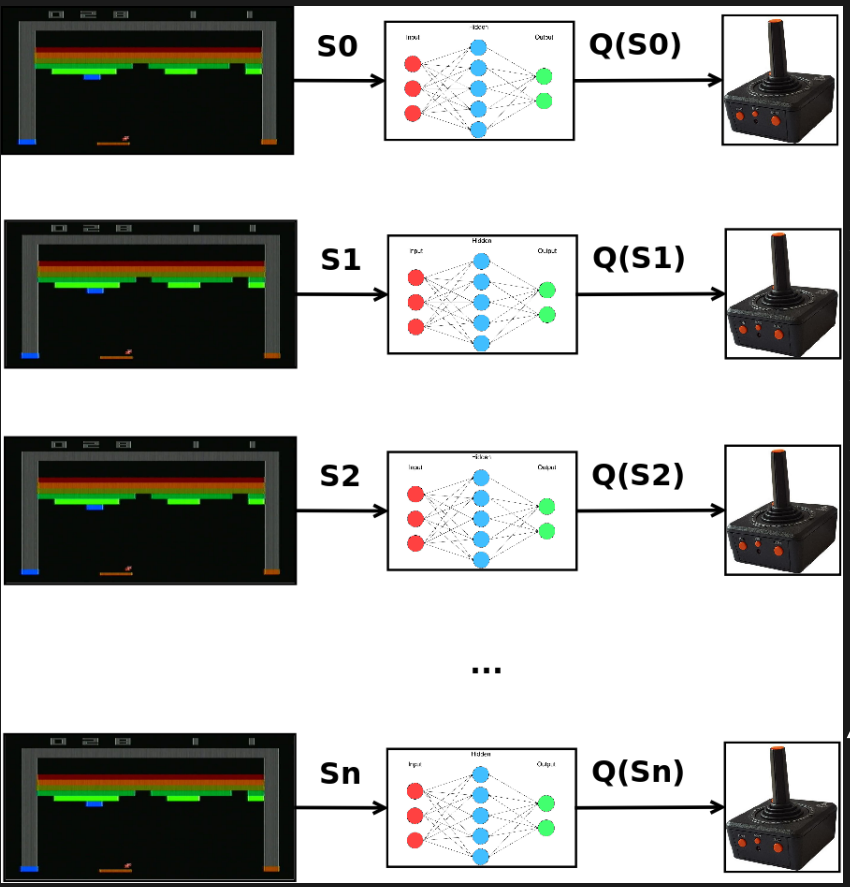
$$SARSA : Q(s,a) = (1-\alpha)Q'(s,a) + \alpha\Big(R(s,a) + \gamma Q'(s',a')\Big)$$

$$loss = \Big(required\_value - Q(state, action)\Big)^2$$

$$loss = \Big(Q'(s,a) - Q(s,a)\Big)^2$$

$$loss = \Big(R(s,a) + \gamma Q'(s',a') - Q(s,a)\Big)^2$$

Training network



$$\Delta Q'(s_0, a_0) = \alpha\Big(R(s_0, a_0) + \gamma Q'(s_1, a_1)\Big)$$

$$\Delta Q'(s_1, a_1) = \alpha\Big(R(s_1, a_1) + \gamma Q'(s_2, a_2)\Big)$$

$$\Delta Q'(s_2, a_2) = \alpha\Big(R(s_2, a_2) + \gamma Q'(s_3, a_3)\Big)$$

$$\dots$$

$$\Delta Q'(s_n, a_n) = \alpha\Big(R(s_n, a_n) + \gamma Q'(s_{n+1}, a_{n+1})\Big)$$

Training network

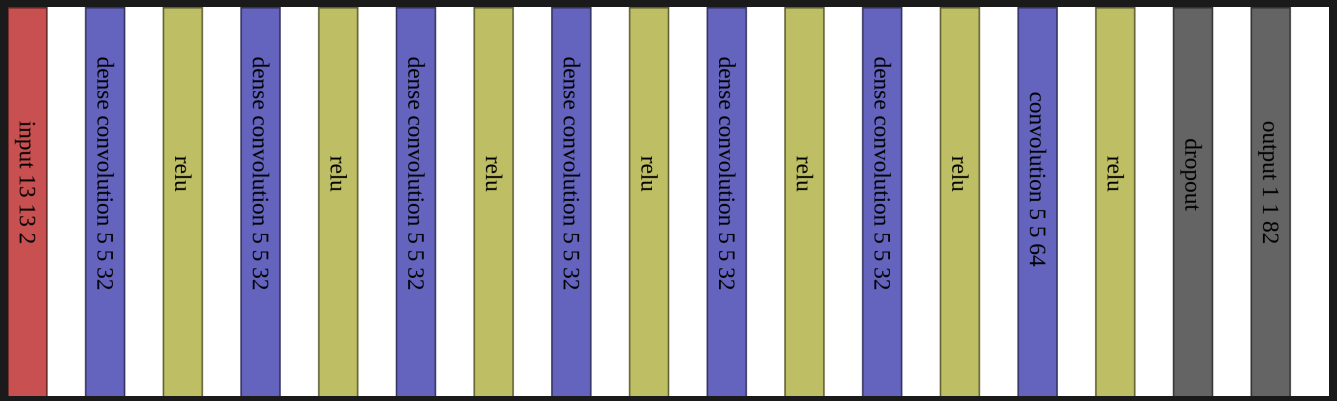
1. computer temporary Q_t batch

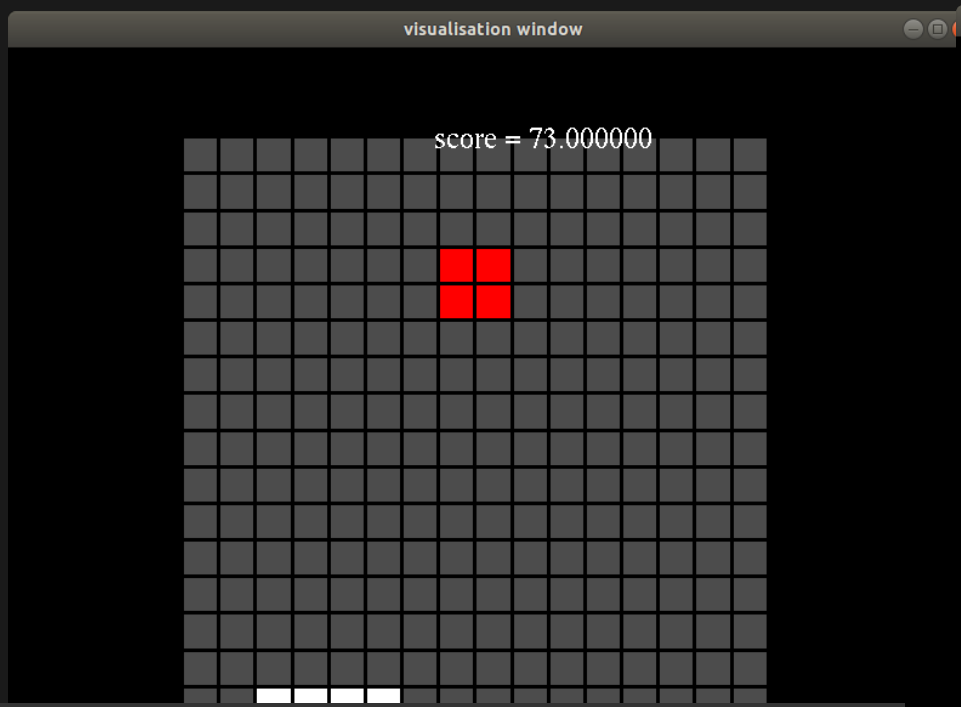
$$
\begin{aligned}
Q_t(s_0, a_0) &= R(s_0, a_0) + \gamma Q_t(s_1, a_1) \\
Q_t(s_1, a_1) &= R(s_1, a_1) + \gamma Q_t(s_2, a_2) \\
Q_t(s_2, a_2) &= R(s_2, a_2) + \gamma Q_t(s_3, a_3) \\
&\dots \\
Q_t(s_n, a_n) &= R(s_{n+1}, a_{n+1})
\end{aligned}
$$

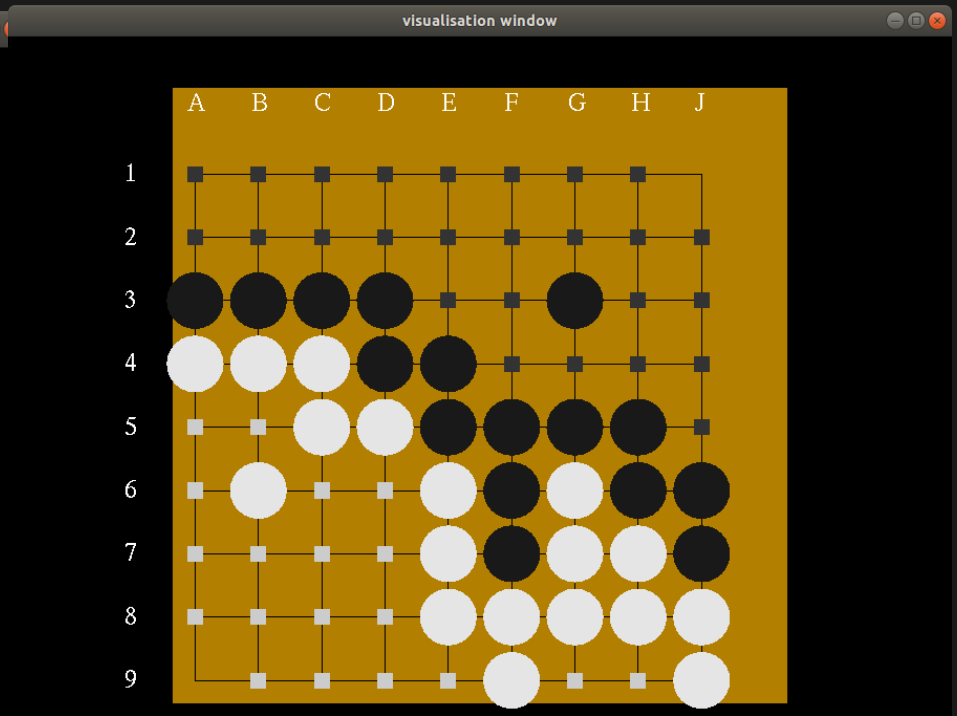
2. train neural network using

$$Q_{new}(s_i, a_i) = (1 - \alpha) Q_{old}(s_i, a_i) + \alpha Q_t(s_i, a_i)$$

Examples


input 13 13 2
dense convolution 5 5 32
relu
dense convolution 5 5 32
relu
dense convolution 5 5 32
relu
dense convolution 5 5 32
relu
dense convolution 5 5 32
relu
dense convolution 5 5 32
relu
convolution 5 5 64
relu
dropout
output 1 1 82


visualisation window
score = 73.000000


visualisation window
A B C D E F G H J
1
2
3
4
5
6
7
8
9

Usefull links

Richard S. Sutton : Reinforcement Learning: An Introduction
https:
//www.amazon.com/Reinforcement-Learning-Introduction-Adaptive-Computation/dp/0262193981

Maxim Lapan : Deep reinforcement learning
https://www.amazon.com/Practical-Reinforcement-Learning-Maxim-Lapan/dp/1788834240

Andrej Karpathy : Pong from pixels
http://karpathy.github.io/2016/05/31/rl/

CHRISTOPHER J.C.H. WATKINS : Q-learning
http://www.gatsby.ucl.ac.uk/~dayan/papers/cjch.pdf

Mohit Sewak : Practical Convolutional Neural
Networks
https://www.amazon.com/Practical-Convolutional-Neural-Networks-Implement/dp/1788392302

Densely Connected Convolutional Networks
https://arxiv.org/pdf/1608.06993.pdf

Q&A

michal chovanec (michal.nand@gmail.com)

www.youtube.com/channel/UCzVvP2ou8v3afNiVrPAHQGg

github https://github.com/michalnand