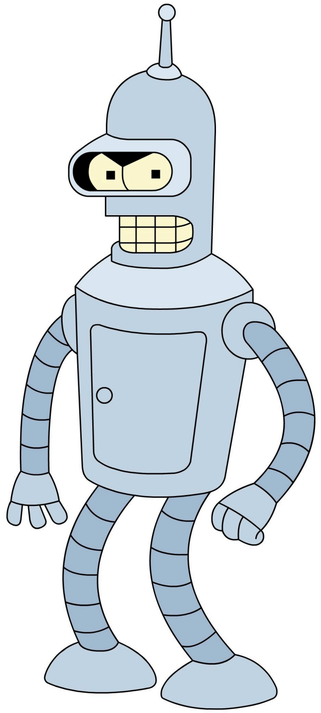
Reinforcement Learning

HSE, autumn - winter 2022

Lecture 4: Deep RL



Sergei Laktionov
slaktionov@hse.ru
LinkedIn

Background

1. Practical RL course by YSDA, week 4
2. Past iteration course, lecture 4
3. RL for Finance Book, Chapter 13

Recap: Q-Learning

Setup: $|\mathcal{A}| < +\infty, |\mathcal{S}| < +\infty$

Recall the Bellman optimality equation for Q^* and apply it as an update rule:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_t + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)]$$

TD-error

Bellman target

In this case, the learned action-value function, Q, directly approximates Q^*, the optimal action-value function, independent of the policy being followed.

Recap: Q-Learning

Setup: $|\mathcal{A}| < +\infty, |\mathcal{S}| < +\infty$

Recall the Bellman optimality equation for Q^* and apply it as an update rule:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_t + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)]$$

TD-error

Bellman target

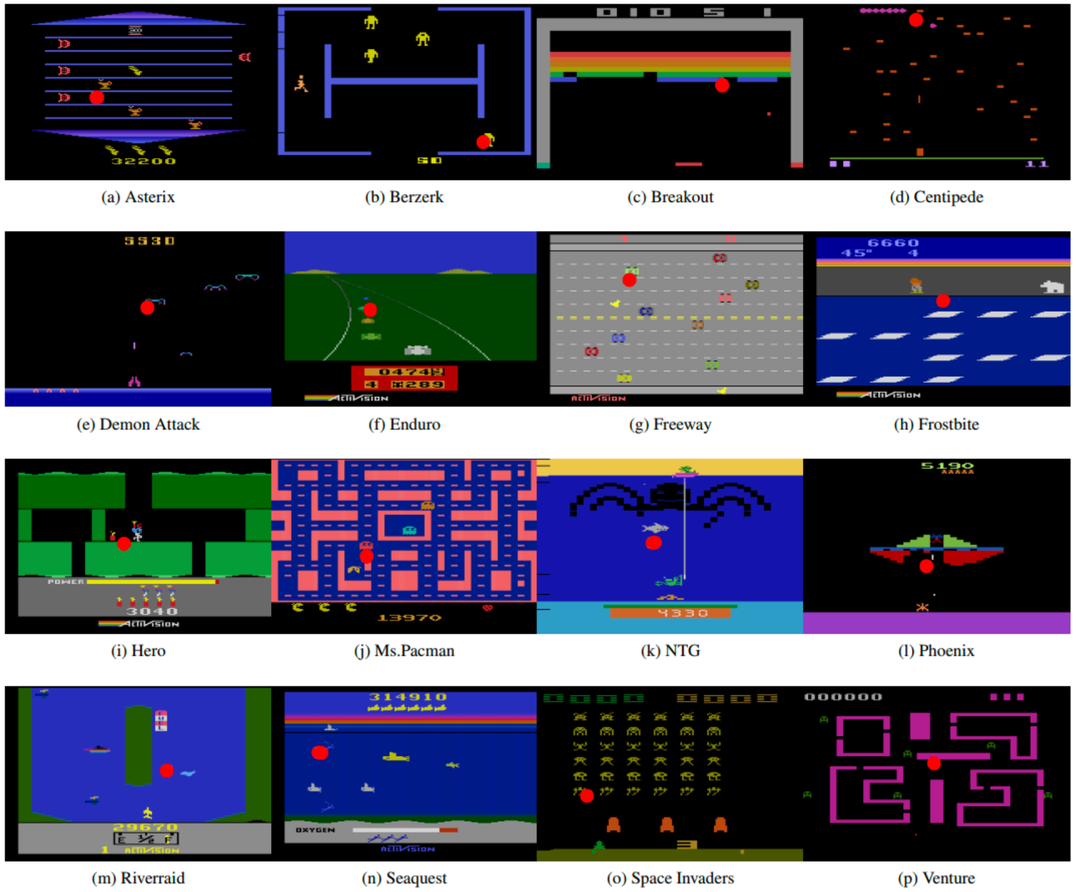
In this case, the learned action-value function, Q, directly approximates Q^*, the optimal action-value function, independent of the policy being followed.

Recall the trajectories are generated following the ε-greedy (behaviour) policy while the Q-function's update corresponds to the greedy (target) strategy. That is the reason why Q-learning is an off-policy method.

Atari

Setup: $|\mathcal{A}| < +\infty, |\mathcal{S}| < +\infty$

1. More than 57 different games
2. Only frames are available
3. State space is actually finite but too large to apply tabular methods
4. Few actions are available



(a) Asterix (b) Berzerk (c) Breakout (d) Centipede

(e) Demon Attack (f) Enduro (g) Freeway (h) Frostbite

(i) Hero (j) Ms.Pacman (k) NTG (l) Phoenix

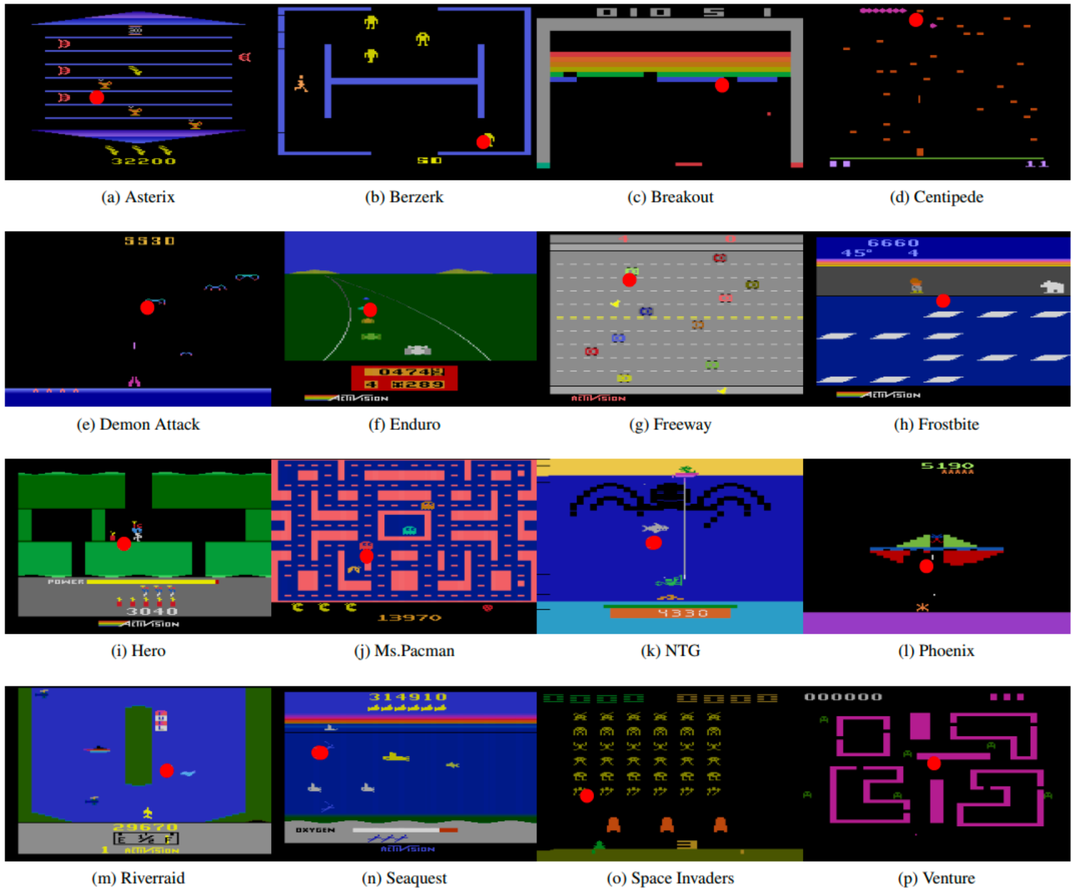
(m) Riverraid (n) Seaquest (o) Space Invaders (p) Venture

Source

Atari

Setup: $|\mathcal{A}| < +\infty, |\mathcal{S}| < +\infty$

1. More than 57 different games
2. Only frames are available
3. State space is actually finite but too large to apply tabular methods
4. Few actions are available



(a) Asterix (b) Berzerk (c) Breakout (d) Centipede

(e) Demon Attack (f) Enduro (g) Freeway (h) Frostbite

(i) Hero (j) Ms.Pacman (k) NTG (l) Phoenix

(m) Riverraid (n) Seaquest (o) Space Invaders (p) Venture

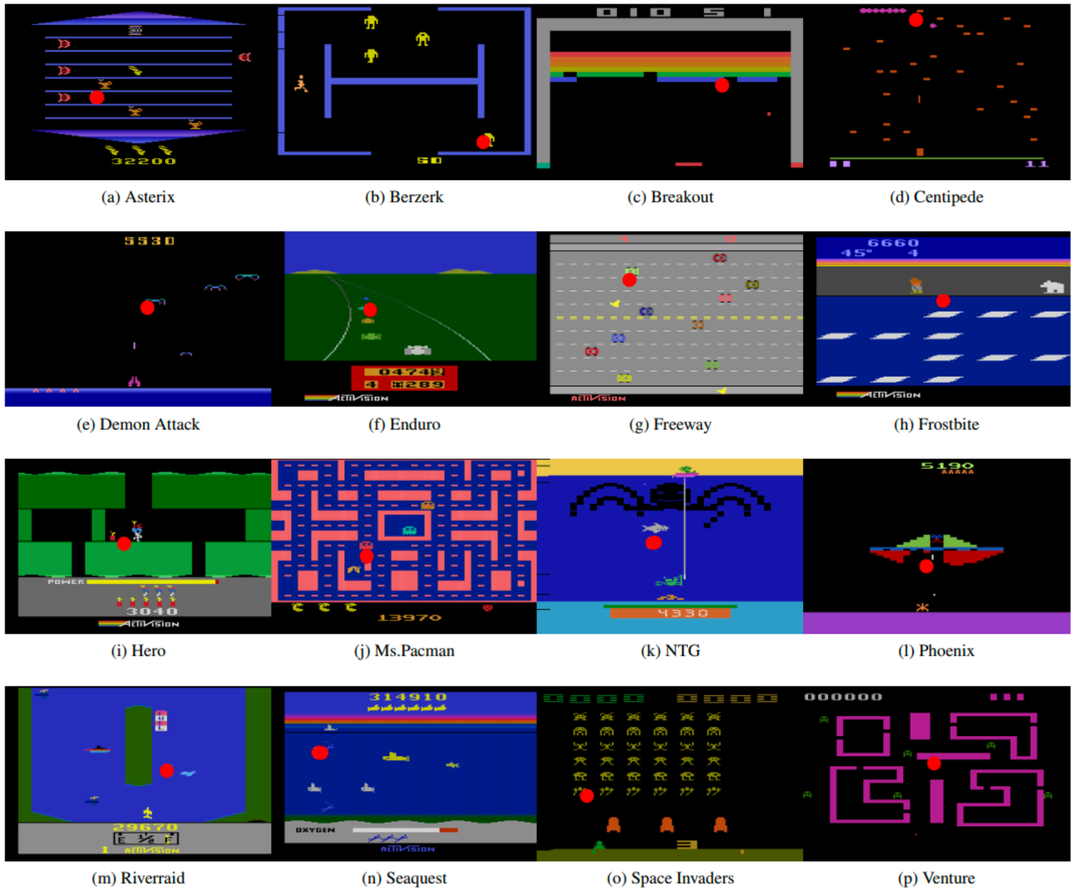
Source

The ultimate goal is to build a universal algorithm which can be applied to all of these environments without modifications and specific hyperparameters.

Atari

Setup: $|\mathcal{A}| < +\infty, |\mathcal{S}| < +\infty$

1. More than 57 different games
2. Only frames are available
3. State space is actually finite but too large to apply tabular methods
4. Few actions are available



(a) Asterix (b) Berzerk (c) Breakout (d) Centipede

(e) Demon Attack (f) Enduro (g) Freeway (h) Frostbite

(i) Hero (j) Ms.Pacman (k) NTG (l) Phoenix

(m) Riverraid (n) Seaquest (o) Space Invaders (p) Venture

Source

Let's approximate action-value function with a neural network: $Q^*(s,a) \approx Q(s,a;\theta)$

Deep Q-Networks (DQN, 2013)

Playing Atari with Deep Reinforcement Learning

Volodymyr Mnih **Koray Kavukcuoglu** **David Silver** **Alex Graves** **Ioannis Antonoglou**

Daan Wierstra **Martin Riedmiller**

DeepMind Technologies

{vlad,koray,david,alex.graves,ioannis,daan,martin.riedmiller} @ deepmind.com

Abstract

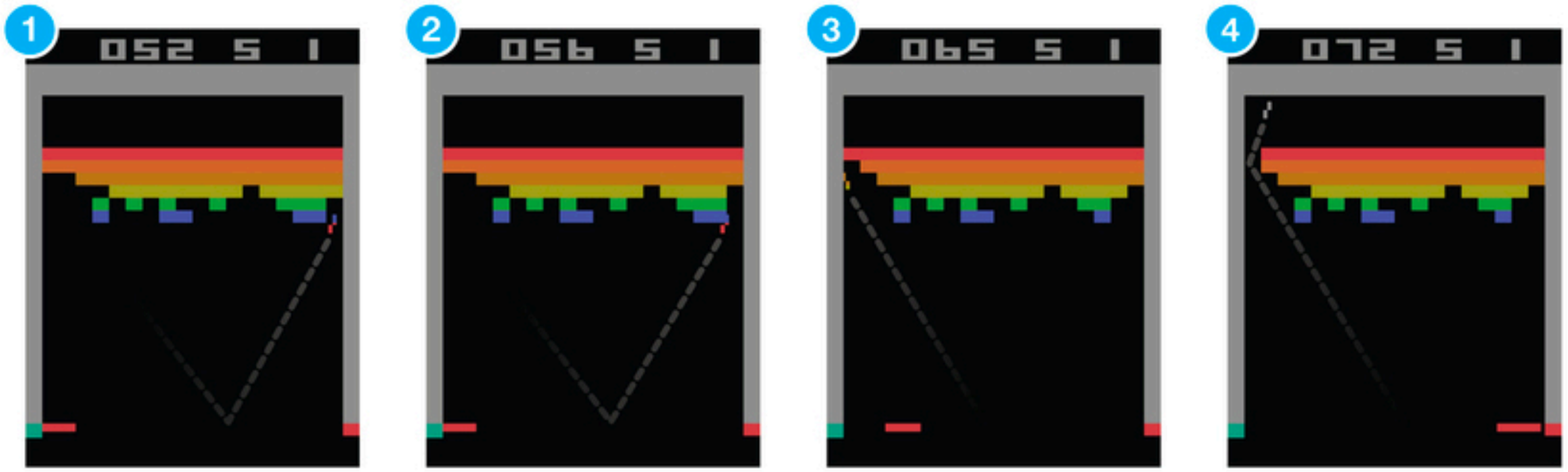
We present the first deep learning model to successfully learn control policies directly from high-dimensional sensory input using reinforcement learning. The model is a convolutional neural network, trained with a variant of Q-learning, whose input is raw pixels and whose output is a value function estimating future rewards. We apply our method to seven Atari 2600 games from the Arcade Learning Environment, with no adjustment of the architecture or learning algorithm. We find that it outperforms all previous approaches on six of the games and surpasses a human expert on three of them.

Original paper

Is Atari Environment MDP?

Source

MDP from Frames



Source

Preprocessing

States:

- Crop image
- Grayscale
- Downsampling
- Stack 4 consecutive frames

Actions' frequency:

- MaxAndSkip
- Sticky actions (we won't use)

Environment:

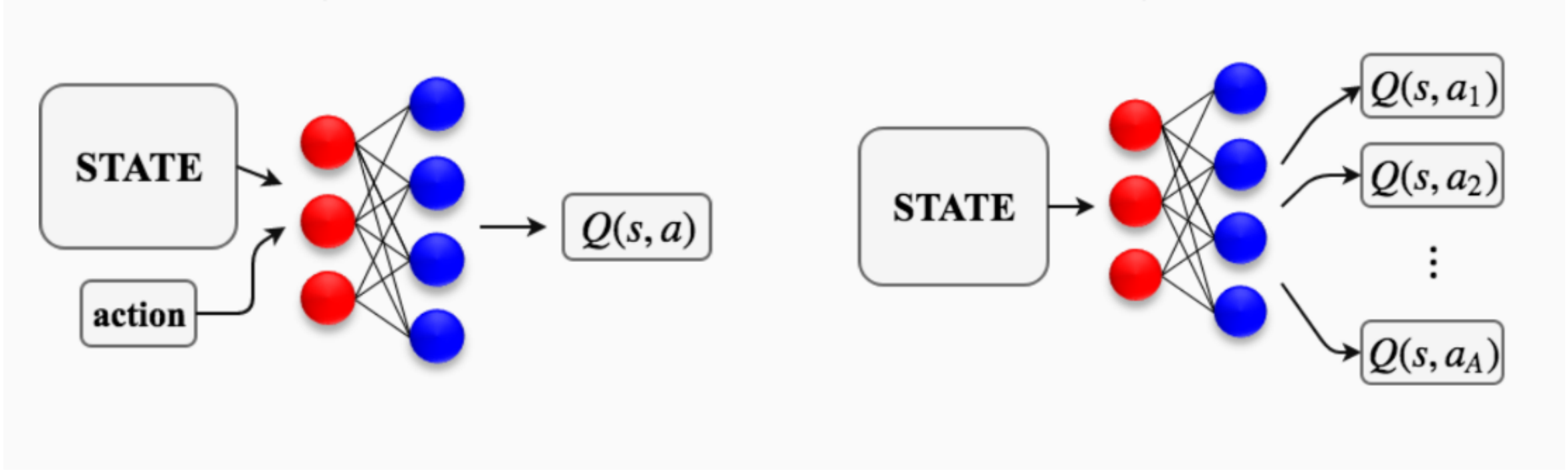
- EpisodicLife
- FireReset

Reward:

- ClipReward ($\{-1,0,1\}$)

Deep Q-network

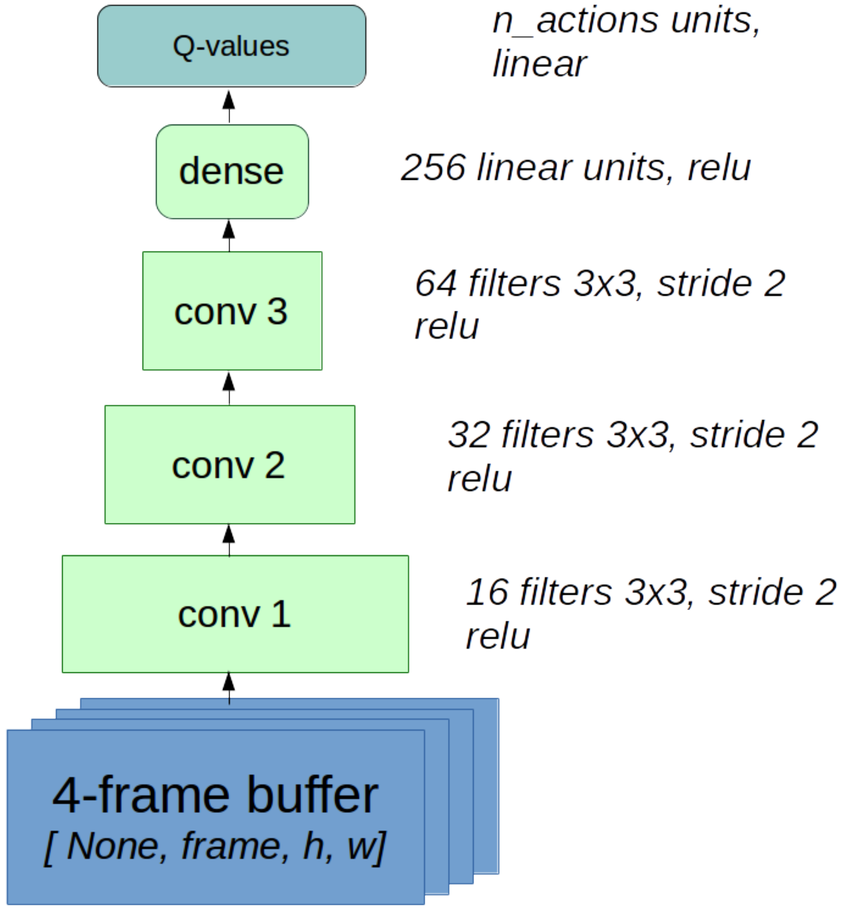
Choose your fighter:



Source

Deep Q-network

- No MaxPool
- No Dropout
- No BatchNorm



Source

Target

$Q^*(s, a) = \mathbb{E}[R_t + \gamma \max_a Q^*(S_{t+1}, a) \,|\, S_t = s, A_t = a]$ is Bellman optimally equation.

At each iteration i:

$y_i = \mathbb{E}_{s' \sim \mathscr{E}}[r + \gamma \max_{a'} Q(s', a', \theta_{i-1}) \,|\, s, a]$ is Bellman target, $\mathscr{E}$ is an environment

$L_i(\theta_i) = \mathbb{E}_{s,a \sim \rho(.)}[(y_i - Q(s, a; \theta_i))^2]$ - MSE loss, $\rho(s, a)$ is a probability distribution over sequences s and actions a, which authors refer to as the behaviour distribution.

Gradients

At each iteration i:

$y_i = \mathbb{E}_{s' \sim \mathscr{E}}[r + \gamma \max_{a'} Q(s', a', \theta_{i-1}) \,|\, s, a]$ is Bellman target, $\mathscr{E}$ is an environment

$L_i(\theta_i) = \mathbb{E}_{s,a \sim \rho(.)}[(y_i - Q(s, a; \theta_i))^2]$ - MSE loss, $\rho(s, a)$ is a probability distribution over sequences s and actions a, which authors refer to as the behaviour distribution.

$$\nabla_{\theta_i} L_i(\theta_i) = \mathbb{E}_{s,a \sim \rho(.); s' \sim \mathscr{E}}\left[\left(r + \gamma \max_{a'} Q(s', a'; \theta_{i-1}) - Q(s, a; \theta_i)\right) \nabla_{\theta_i} Q(s, a; \theta_i)\right]$$

Gradients

At each iteration i:

$y_i = \mathbb{E}_{s' \sim \mathscr{E}}[r + \gamma \max_{a'} Q(s', a', \theta_{i-1}) \,|\, s, a]$ is Bellman target, $\mathscr{E}$ is an environment

$L_i(\theta_i) = \mathbb{E}_{s,a \sim \rho(.)}[(y_i - Q(s, a; \theta_i))^2]$ - MSE loss, $\rho(s, a)$ is a probability distribution over sequences s and actions a, which authors refer to as the behaviour distribution.

TD-error

Step size

$$\nabla_{\theta_i} L_i(\theta_i) = \mathbb{E}_{s,a \sim \rho(.); s' \sim \mathscr{E}}\Big[\Big(r + \gamma \max_{a'} Q(s', a'; \theta_{i-1}) - Q(s, a; \theta_i)\Big) \nabla_{\theta_i} Q(s, a; \theta_i)\Big]$$

Bellman target

Gradients

At each iteration i:

$y_i = \mathbb{E}_{s' \sim \mathscr{E}}[r + \gamma \max_{a'} Q(s', a', \theta_{i-1}) \,|\, s, a]$ is Bellman target, $\mathscr{E}$ is an environment

$L_i(\theta_i) = \mathbb{E}_{s,a \sim \rho(.)}[(y_i - Q(s, a; \theta_i))^2]$ - MSE loss, $\rho(s, a)$ is a probability distribution over sequences s and actions a, which authors refer to as the behaviour distribution.

TD-error

Step size

$$\nabla_{\theta_i} L_i(\theta_i) = \mathbb{E}_{s,a \sim \rho(.); s' \sim \mathscr{E}}\Big[\Big(r + \gamma \max_{a'} Q(s', a'; \theta_{i-1}) - Q(s, a; \theta_i)\Big) \nabla_{\theta_i} Q(s, a; \theta_i)\Big]$$

Bellman target

Let's apply SGD!

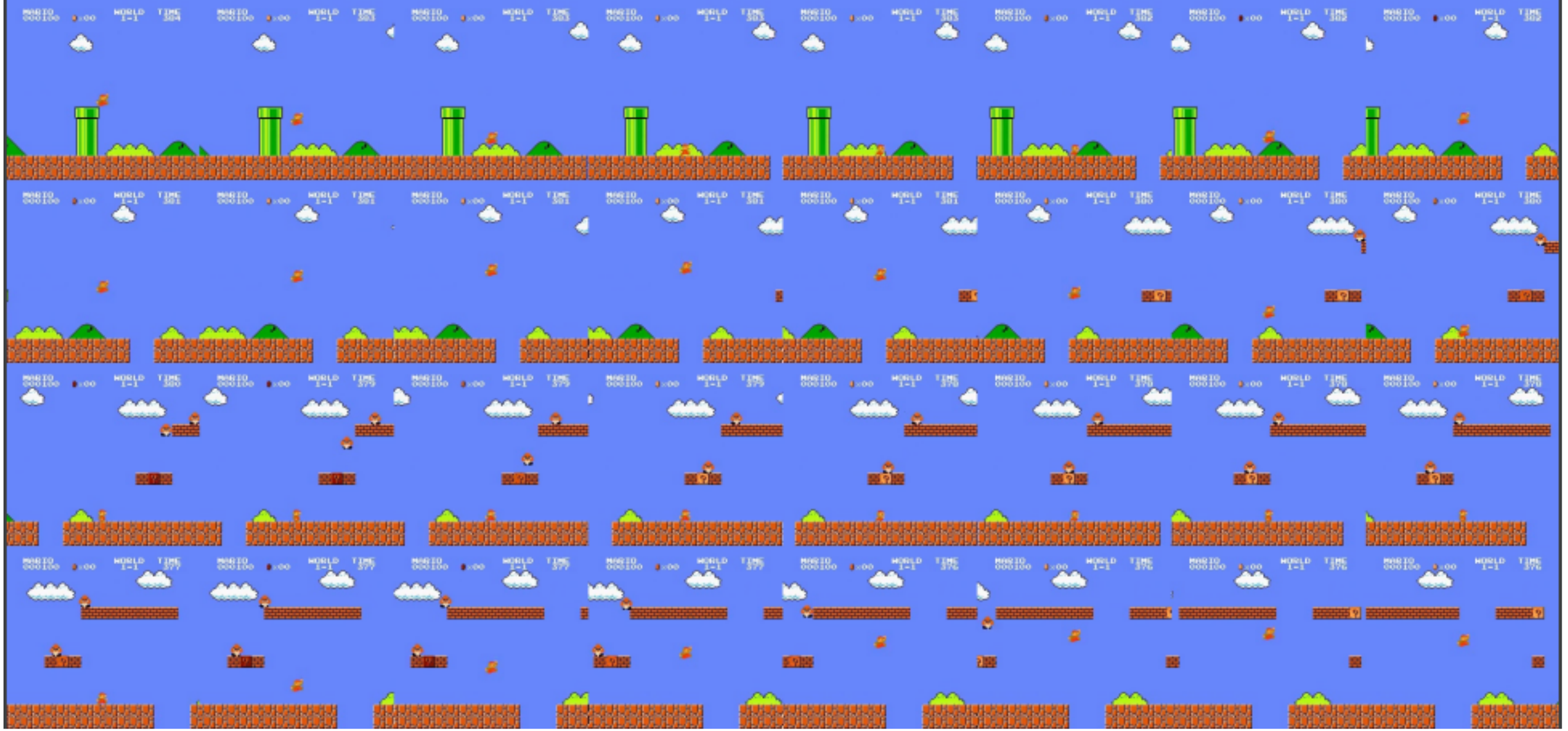
Target Network

If we use weights from the previous step targets are highly non-stationary.

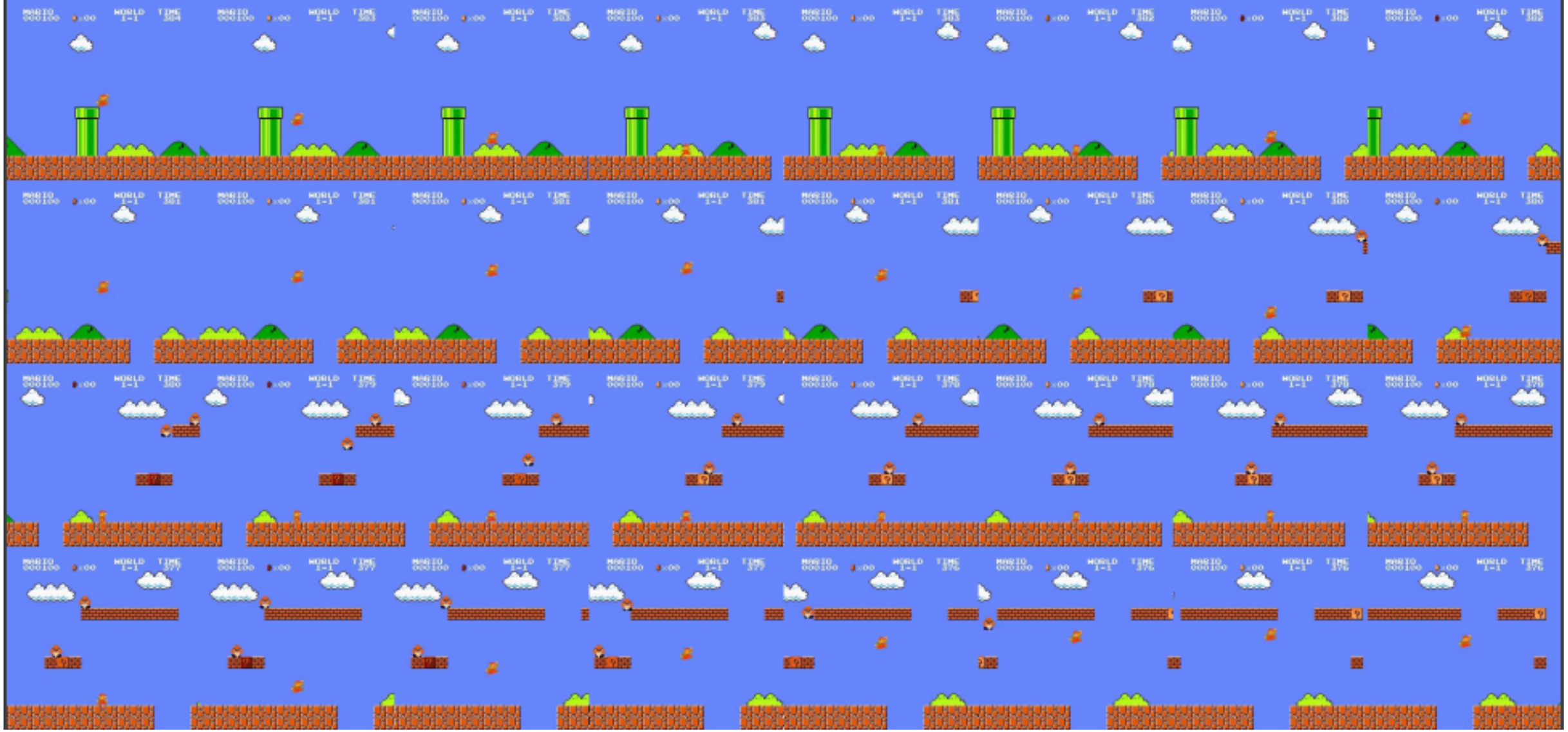
Let's update weight of the network for target generation (target network) every K steps or apply soft updates (Polyak update) with parameter β:

A. $\theta^- \leftarrow \theta$ every K SGD iterations

B. $\theta^- \leftarrow (1-\beta)\theta^- + \beta\theta$ on each iteration

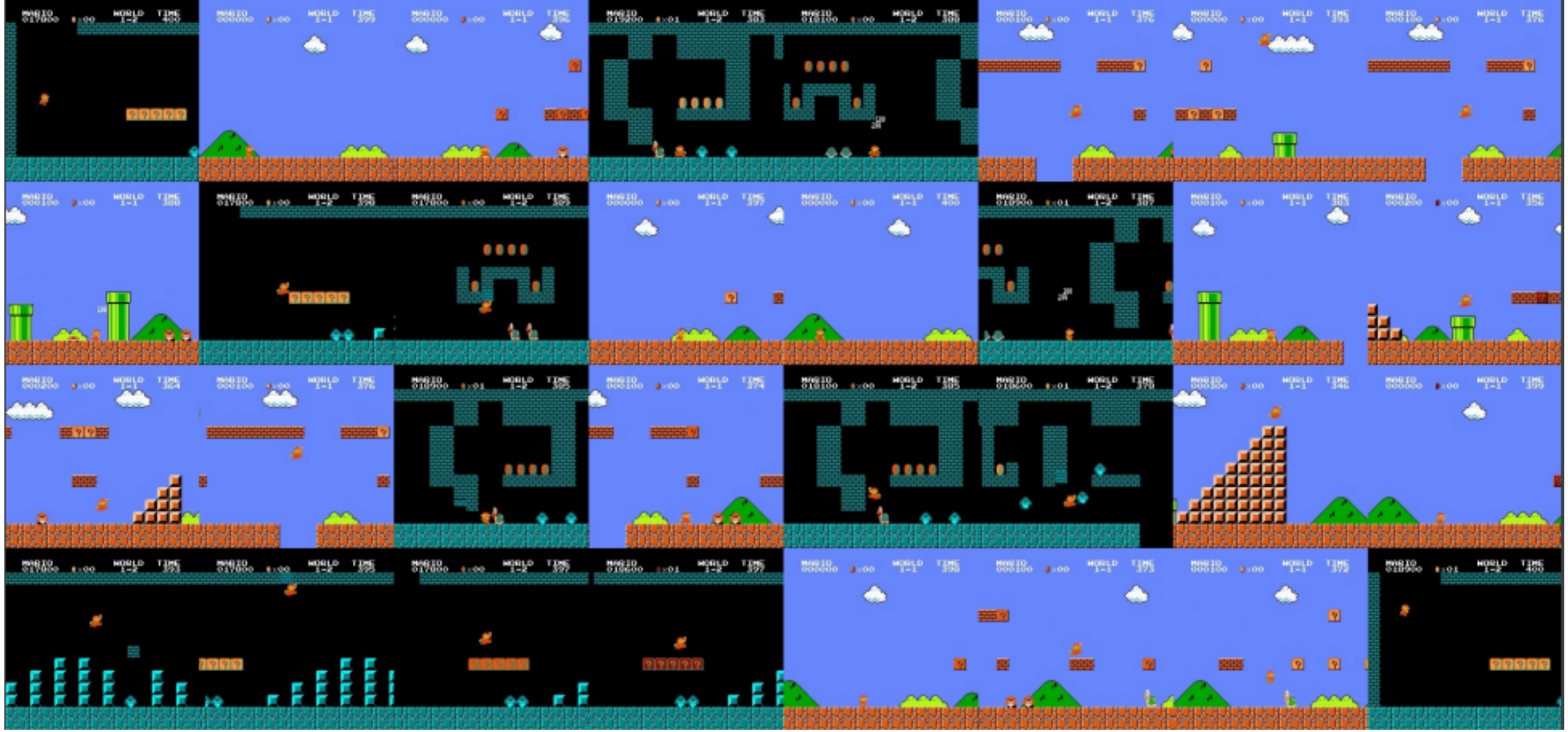


Source



Source

Consecutive samples are extremely correlated so batch’s elements are not i.i.d.

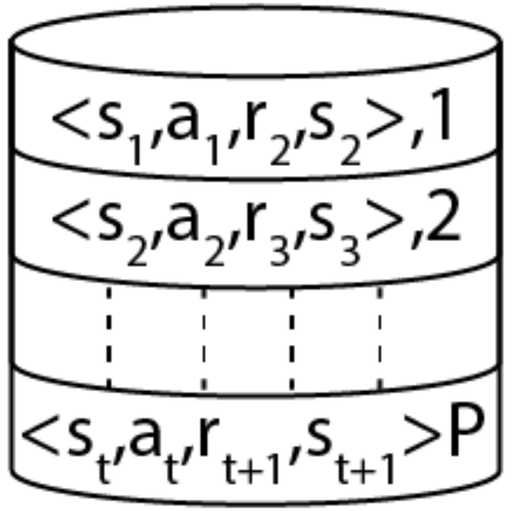


Source

Let’s sample randomly from the Experience Replay Buffer which stores played transitions

Experience Replay Buffer

- On each step store $< s, a, r, s' >$ in the buffer
- Sample n random transitions from the buffer
- Train on them



Source

Advantages:

- No need to revisit same states many times
- Make the estimators consistent with the current policy, update estimators
- Decorrelate update samples to maintain i.i.d. assumption

Disadvantages:

- Not applicable for the on-policy learning

Exploration

Recall so-called ε-greedy policy:

$$\pi = \begin{cases} \text{select random action with probabily } \varepsilon \\ \text{select greedy action with probabily } 1 - \varepsilon \end{cases}$$

We use ε-greedy strategies to avoid being stuck in local optima

DQN Algorithm

Initialise $Q(s, a; \theta)$; $\theta^- = \theta$; replay buffer D is empty;

Observe s_0;

$t = 0,1,\ldots$

- Take action a sampled from the ε-greedy policy w.r.t. $Q(s, a; \theta_t)$;
- Observe new (r, s', done), store $(s, a, r, s', \text{done})$ in D;
- Sample batch of transitions $(s_j, a_j, r_j, s_{j+1}, \text{done}_{j+1})_{j=1}^{B}$ from D
- $y_j = r_j + \gamma(1 - \text{done}_{j+1}) \max_{a'} Q(s_{j+1}, a'; \theta^-)$;
- Perform a gradient descent step: $\theta_{t+1} = \theta_t - \frac{\alpha}{B} \sum_{j=1}^{B} \nabla_{\theta_t} (y_j - Q(a_j, s_j; \theta_t))^2$
- If t mod $K = 0$ update target network: $\theta^- \leftarrow \theta_t$

Rainbow (2017)

Rainbow: Combining Improvements in Deep Reinforcement Learning

Matteo Hessel DeepMind
Joseph Modayil DeepMind
Hado van Hasselt DeepMind
Tom Schaul DeepMind
Georg Ostrovski DeepMind

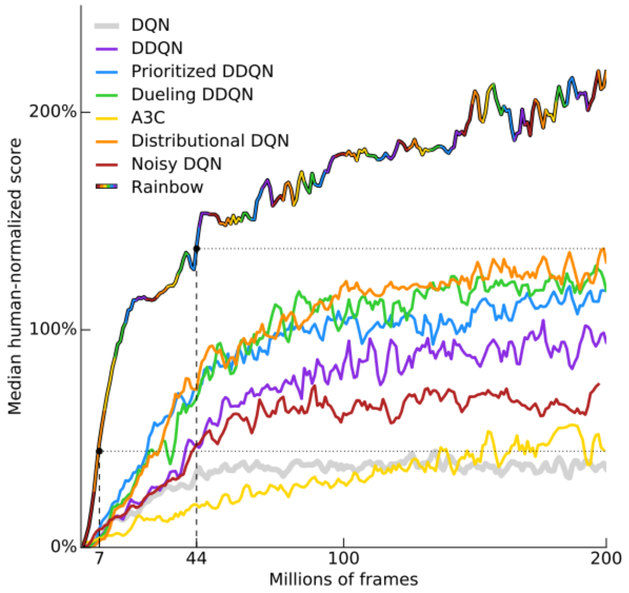
Will Dabney DeepMind
Dan Horgan DeepMind
Bilal Piot DeepMind
Mohammad Azar DeepMind
David Silver DeepMind

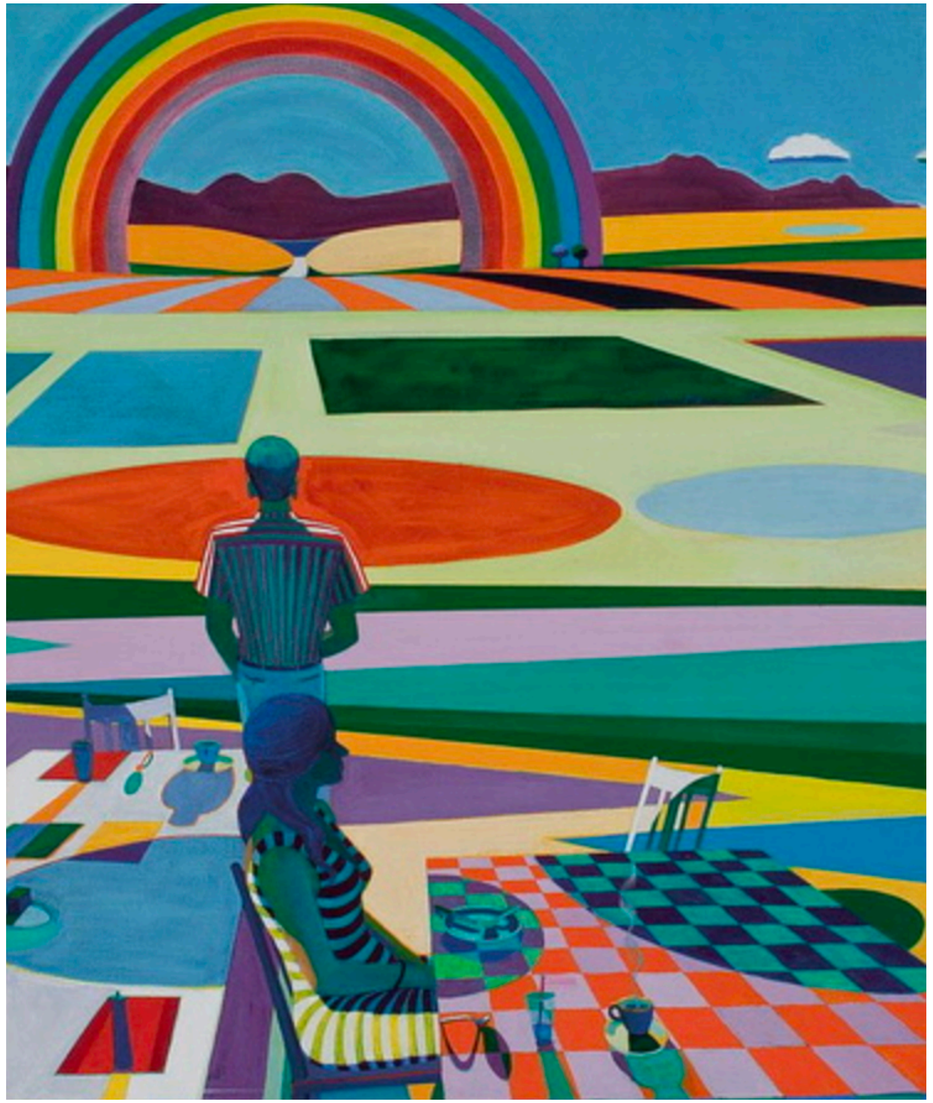
Abstract

The deep reinforcement learning community has made several independent improvements to the DQN algorithm. However, it is unclear which of these extensions are complementary and can be fruitfully combined. This paper examines six extensions to the DQN algorithm and empirically studies their combination. Our experiments show that the combination provides state-of-the-art performance on the Atari 2600 benchmark, both in terms of data efficiency and final performance. We also provide results from a detailed ablation study that shows the contribution of each component to overall performance.

Introduction

The many recent successes in scaling reinforcement learning (RL) to complex sequential decision-making problems were kick-started by the Deep Q-Networks algorithm (DQN; Mnih et al. 2013, 2015). Its combination of Q-learning with convolutional neural networks and experience replay enabled it to learn, from raw pixels, how to play many Atari games at human-level performance. Since then, many exten





Original paper

Source

Overestimation Bias

The overestimation of the Q-function by the algorithms which can be caused by several reasons:

1. Maximum over estimated values is used implicitly as an estimate of the maximum value, which can lead to a significant positive bias due to approximation error.

2. Due to using the same samples (plays) both to determine the maximizing action and to estimate its value.

Action selection

$$\max_{a'} Q(s', a') = Q(s', \text{argmax} Q(s', a'))$$

Action evaluation

Double DQN

Let's decouple action selection and action evaluation. Like in the tabular setting we can use two weakly correlated networks with independent buffers D_i,:

$$y_1 = r + \gamma Q(s', \text{argmax}_{a'} Q(s, a', \theta_1); \theta_2)$$

$$y_2 = r + \gamma Q(s', \text{argmax}_{a'} Q(s, a', \theta_2); \theta_1)$$

but it can be too expensive…

Double DQN

Let's decouple action selection and action evaluation. Like in the tabular setting we can use two weakly correlated networks with independent buffers D_i,:

$$y_1 = r + \gamma Q(s', \text{argmax}_{a'} Q(s, a', \theta_1); \theta_2)$$

$$y_2 = r + \gamma Q(s', \text{argmax}_{a'} Q(s, a', \theta_2); \theta_1)$$

but it can be too expensive…

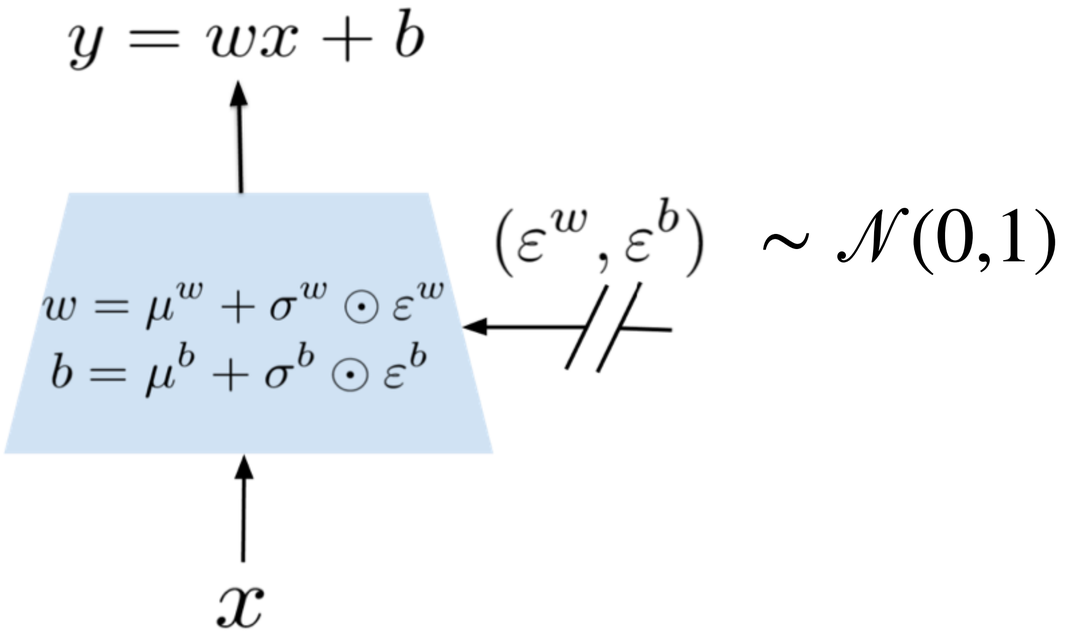
We can use the target network as the second network:

$$y = r + \gamma Q(s', \text{argmax}_{a'} Q(s, a', \theta_t); \theta_t^-)$$

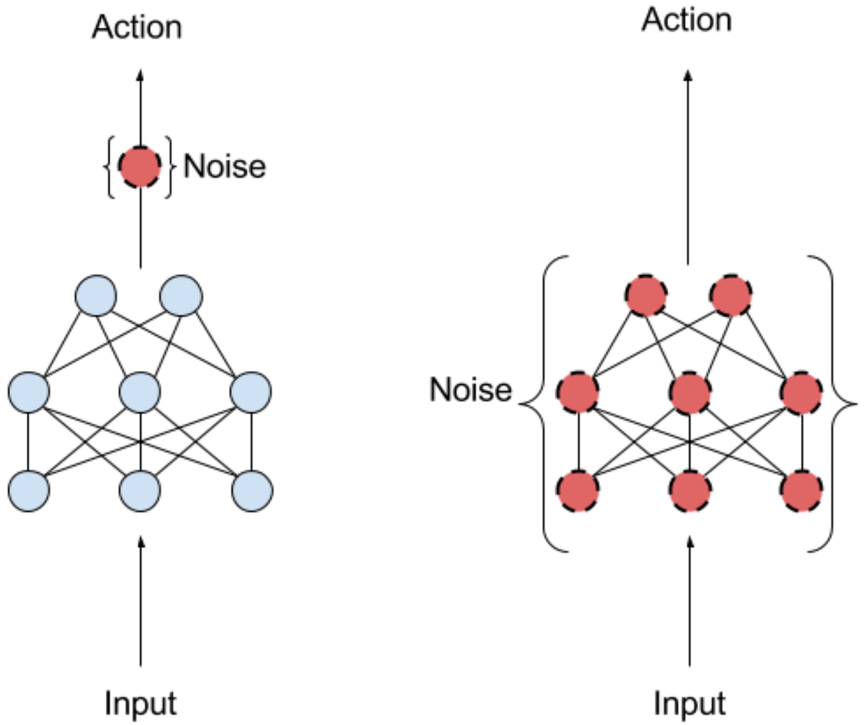
Noisy Networks

Issues:

- State-independent exploration
- ε-greedy is naive



Source

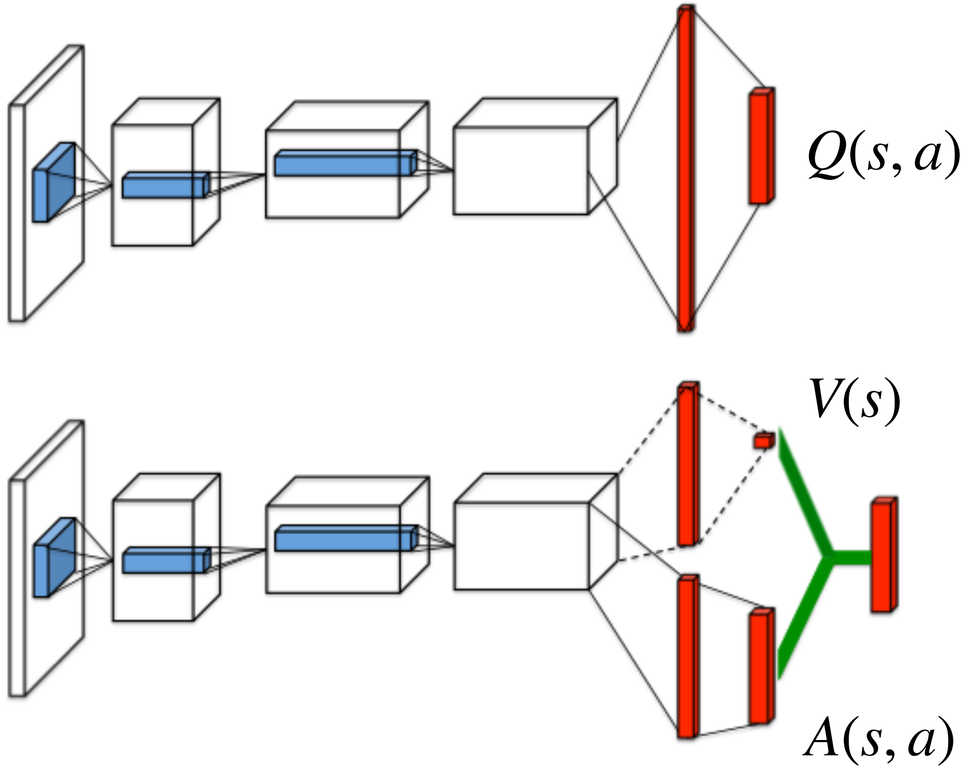


Source

Dueling DQN

$A(s,a) = Q(s,a) - V(s)$ is a relative measure of importance of each action, advantage.

$$Q(s,a) = V(s) + A(s,a)$$



Source

Multi-step DQN

$$y = r + \gamma r' + \gamma r'' + \ldots + \gamma^n \max_{a^{(n)}} Q(s^{(n)}, a^{(n)})$$

We assume that all transitions are generated with greedy policy but it's not the case.

No theoretical guaranties in case of off-policy algorithm!

https://arxiv.org/abs/1901.07510

Other Improvements

1. Prioritized Experience Replay
2. Distributional RL

Prioritized Experience Replay

The uniform sampling from the experience replay make an agent replay transitions at the same frequency that were originally experienced, regardless of their significance.

Prioritized Experience Replay

The uniform sampling from the experience replay make an agent replay transitions at the same frequency that were originally experienced, regardless of their significance.

$\delta_i = y_i - Q(s, a; \theta)$ - TD-error

$p_i = |\delta_i| + \epsilon, \epsilon > 0$

$P(i) = \frac{p_i^\alpha}{\sum_k p_k^\alpha}$ is a the probability of sampling transition i

However we induce a bias in a Q-function approximation.

Prioritized Experience Replay

The uniform sampling from the experience replay make an agent replay transitions at the same frequency that were originally experienced, regardless of their significance.

$\delta_i = y_i - Q(s, a; \theta)$ - TD-error

$p_i = |\delta_i| + \epsilon, \epsilon > 0$

$P(i) = \frac{p_i^\alpha}{\sum_k p_k^\alpha}$ is a the probability of sampling transition i

However we induce a bias in a Q-function approximation.

We can correct this bias by using Importance-Sampling (IS) weights

$w_i = \left(\frac{1}{N}\frac{1}{P(i)}\right)^\beta$ with beta annealing from 0 to 1 and $w_i\delta_i$ instead of δ_i.

Ablation Study

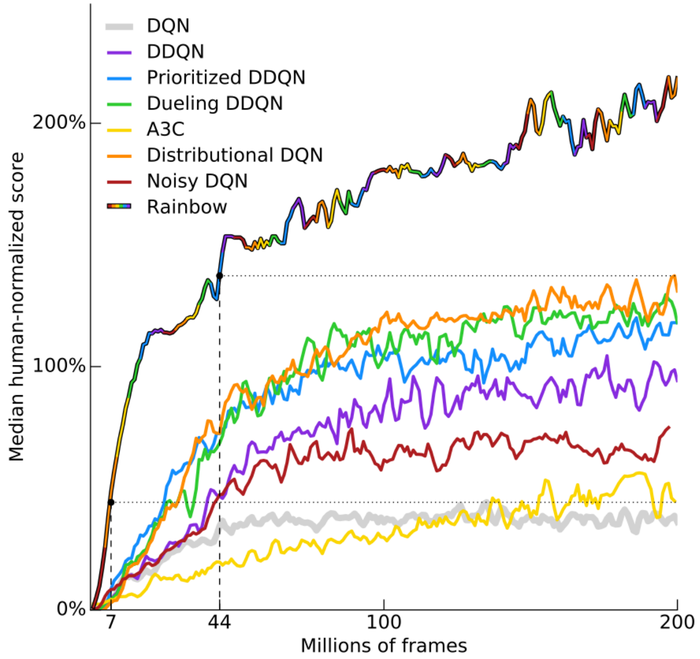


Figure 1: **Median human-normalized performance** across 57 Atari games. We compare our integrated agent (rainbow-colored) to DQN (grey) and six published baselines. Note that we match DQN's best performance after 7M frames, surpass any baseline within 44M frames, and reach substantially improved final performance. Curves are smoothed with a moving average over 5 points.

Source

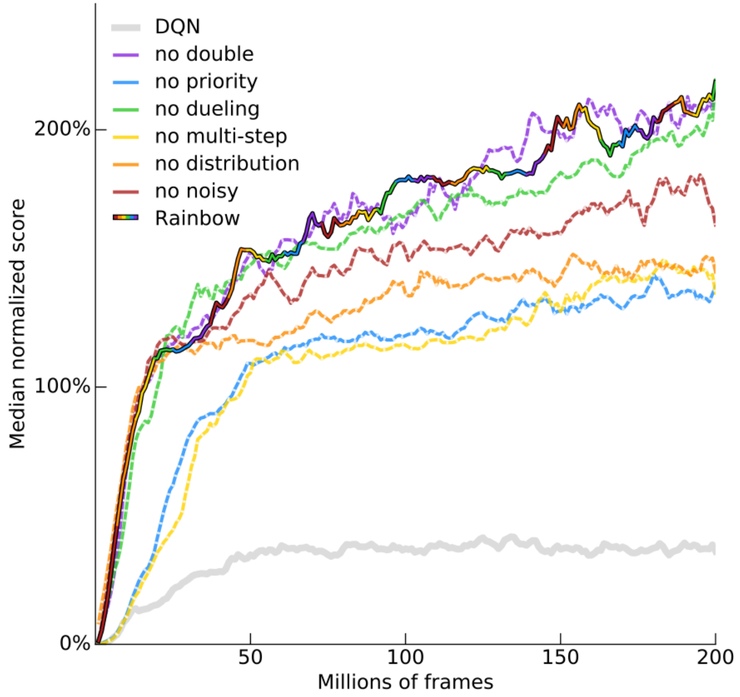


Figure 3: **Median human-normalized performance** across 57 Atari games, as a function of time. We compare our integrated agent (rainbow-colored) to DQN (gray) and to six different ablations (dashed lines). Curves are smoothed with a moving average over 5 points.

Source

Recap: MDP

MDP is a 4-tuple $(\mathcal{S}, \mathcal{A}, p, r)$:

1. $\mathcal{A}$ is an action space
2. $\mathcal{S}$ is a state space
3. $p(s' \mid s, a) = \mathbb{P}(S_{t+1} = s' \mid S_t = s, A_t = a)$ is a state-transition function
4. $r : \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is a reward function giving an expected reward: $r(s, a) = \mathbb{E}[R \mid s, a]$

POMDP

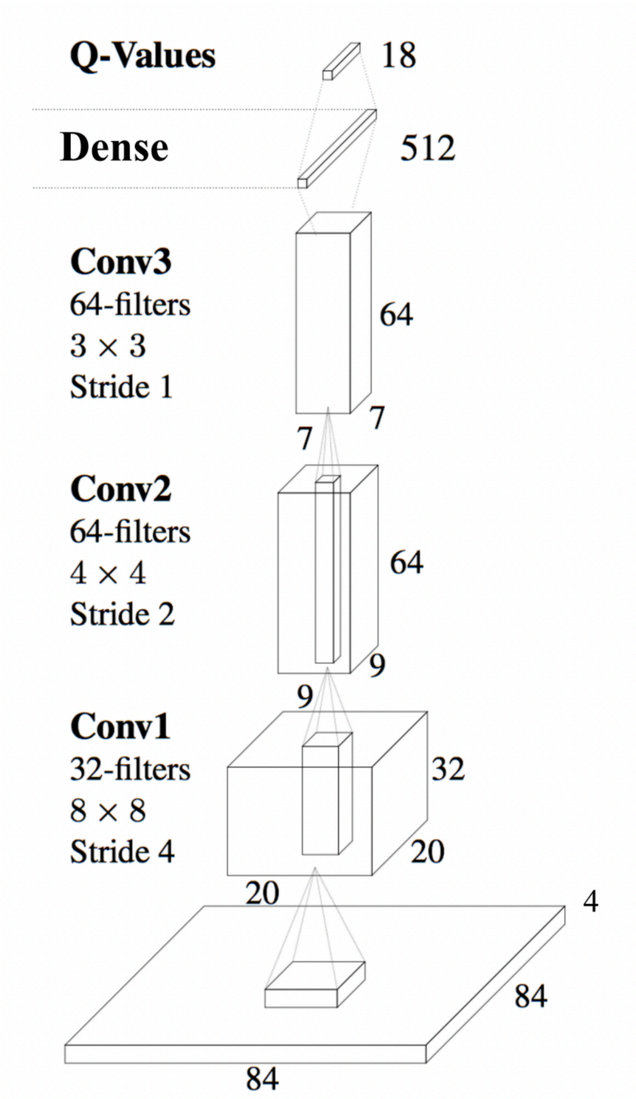
1. $(\mathcal{S}, \mathcal{A}, r)$ are the same as in MDP
2. $\mathcal{O}$ is a set of possible observations
3. $p(s' \mid s, a) = \mathbb{P}(S_{t+1} = s' \mid S_t = s, A_t = a)$ is a state-transition function
4. $p(o \mid s', a) = \mathbb{P}(O_t = o \mid S_{t+1} = s', A_t = a)$ is a observation-transition function

DRQN

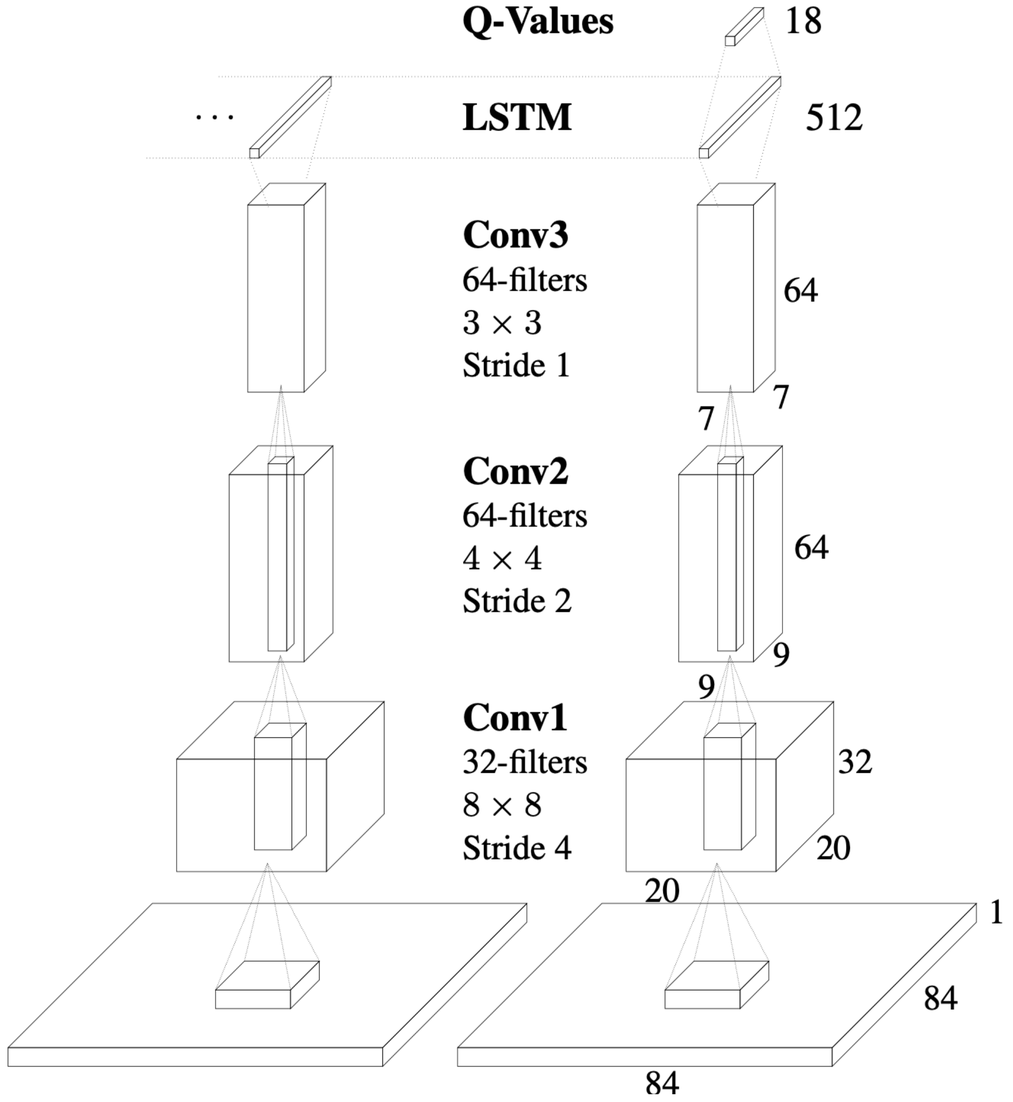
Vanilla Deep Q-Learning has no explicit mechanisms for deciphering the underlying state of the POMDP and is only effective if the observations are reflective of underlying system states. In the general case, estimating a Q-value from an observation can be arbitrarily bad since $Q(o, a; \theta) \neq Q(s, a; \theta)$.

- Let's equip agent with memory h_t
- $Q(s_t, a_t) \approx Q(o_t, h_{t-1}, a_t)$
- $h_t = LSTM(o_t, h_{t-1})$

DQN vs DRQN



Original paper

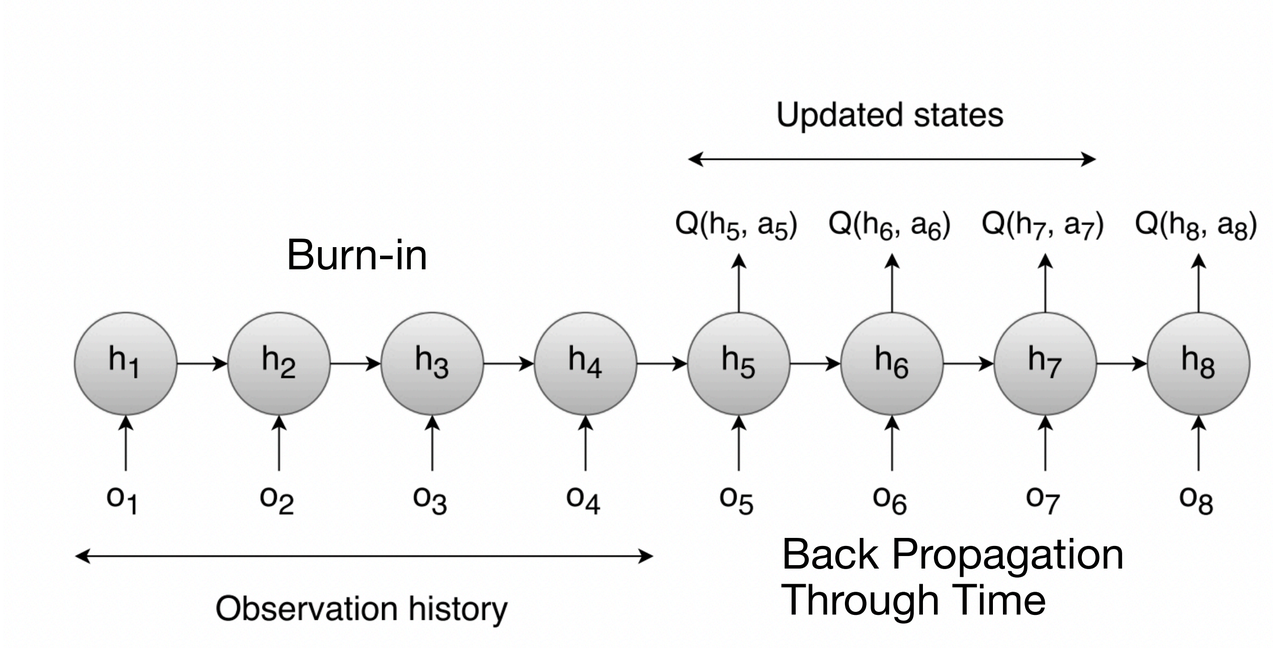


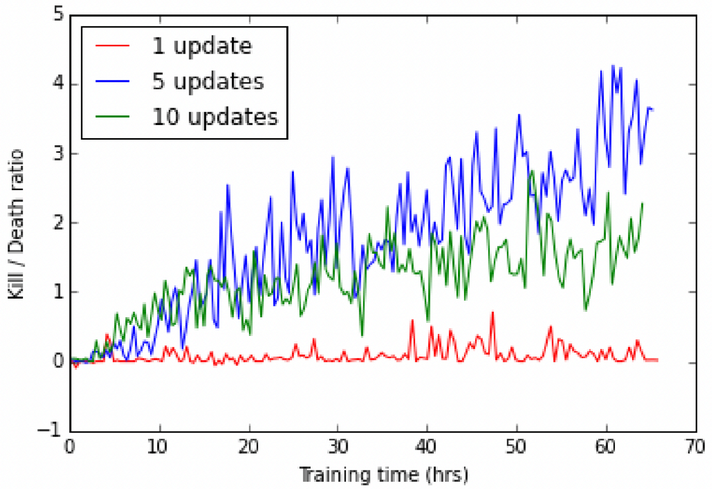
Original paper

DRQN Experience Replay

Original paper

- Sample random time step
- Consider N consecutive transitions
- Update only last M





Thank you for your attention!