Human-level control through deep reinforcement learning

Volodymyr Mnih, Koray Kavukcuoglu, David Silver

(Google DeepMind)

Information about article

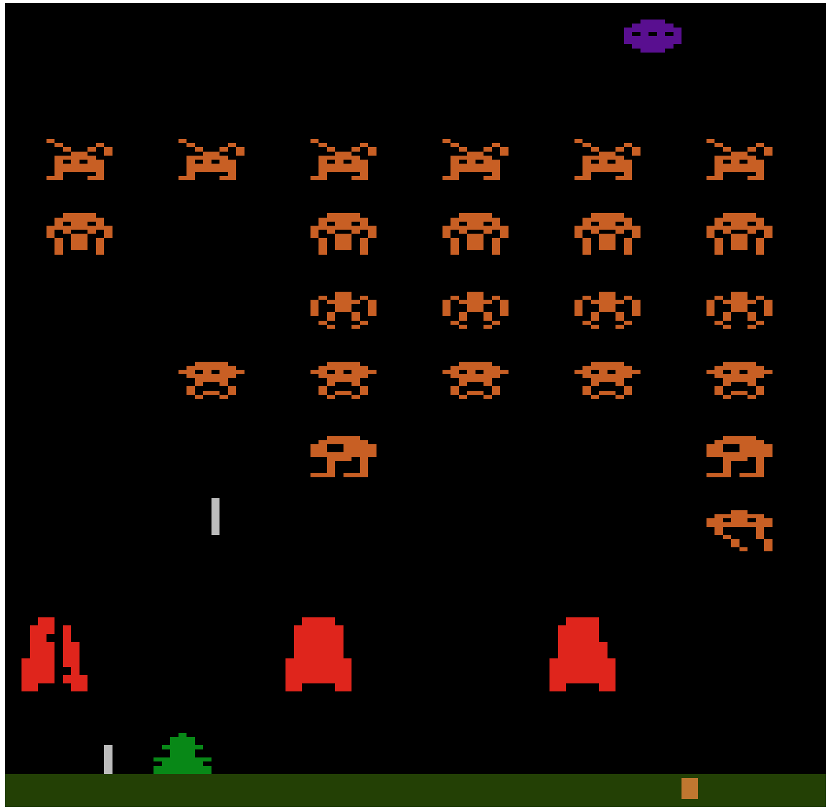
- Date published in “Nature” - 25 February 2015.
- Cited by – 687 times.
- Authors:
  - Volodymyr Mnih – 2154 citiations, h-index 16.
  - Koray Kavukcuoglu – 7214 citiations, h-index 24.
  - David Silver – 4575 citiations, h-index 29.

Problem statement

- Create a single algorithm that would be able to compete in wide range of challenging tasks with minimal prior knowledge.
- These challenging tasks – Atari 2600 games.

Game examples

Space Invaders



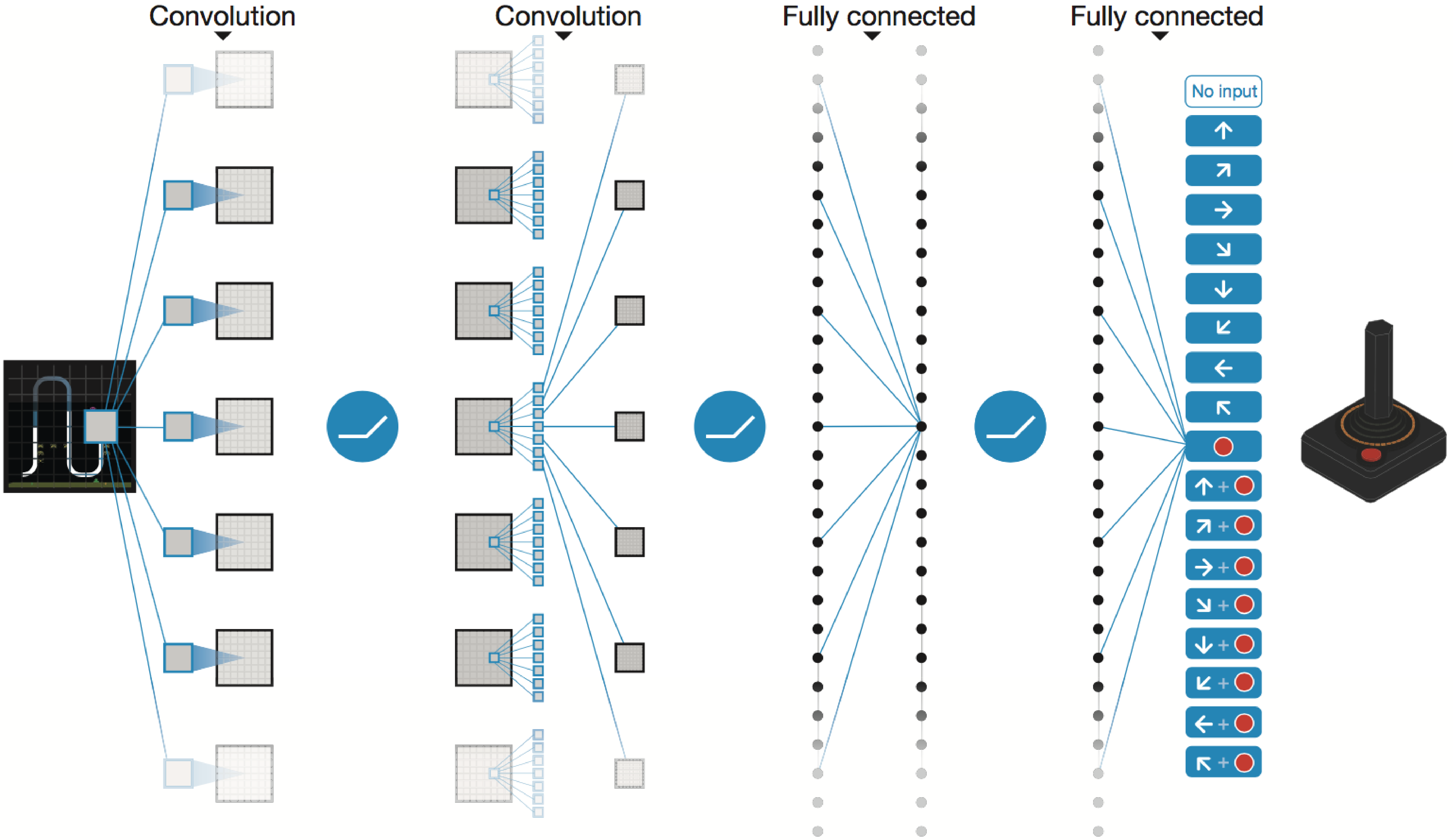
Breakout

Prior knowledge

- Visual images.
- The game-specific score.
- Number of actions, no specification though (agent doesn't have prior knowledge about what 'up' button does).
- The life count (if available).

Visual input

- Raw Atari 2600 frames – 210x160 pixel images, 128-colour palette.
- Artefacts of the Atari 2600 emulator:
  - Flickering.
  - Some objects appear only in even frames, some – in odd.
- Preprocessing:
  - Encoding a single frame – for each pixel take maximum colour value over current and the previous frames.
  - Rescaling to 84x84.



Convolution
Convolution
Fully connected
Fully connected
No input

Model architecture

- Input – 84x84x4 image produce by preprocessing.
- First hidden layer convolves 32 filters of 8x8 with stride 4.
- Second hidden layer – 64 filters of 4x4 with stride 2.
- Third – 64 filters of 3x3 with stride 1.
- Each hidden layer is followed by rectifier *max(0,x)*.
- Final hidden layer – fully-connected, 512 rectifier units.
- Output – fully-connected with a single output for each valid action.

Algorithm

- Goal – maximize rewards by selecting actions.
- Action-value function:

$$Q^*(s,a) = \max_{\pi} \mathbb{E}\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots \mid s_t = s,\ a_t = a,\ \pi\right]$$

- Loss function, optimized by stochastic gradient descent:

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

- Biologically inspired mechanism - experience replay.

Training details

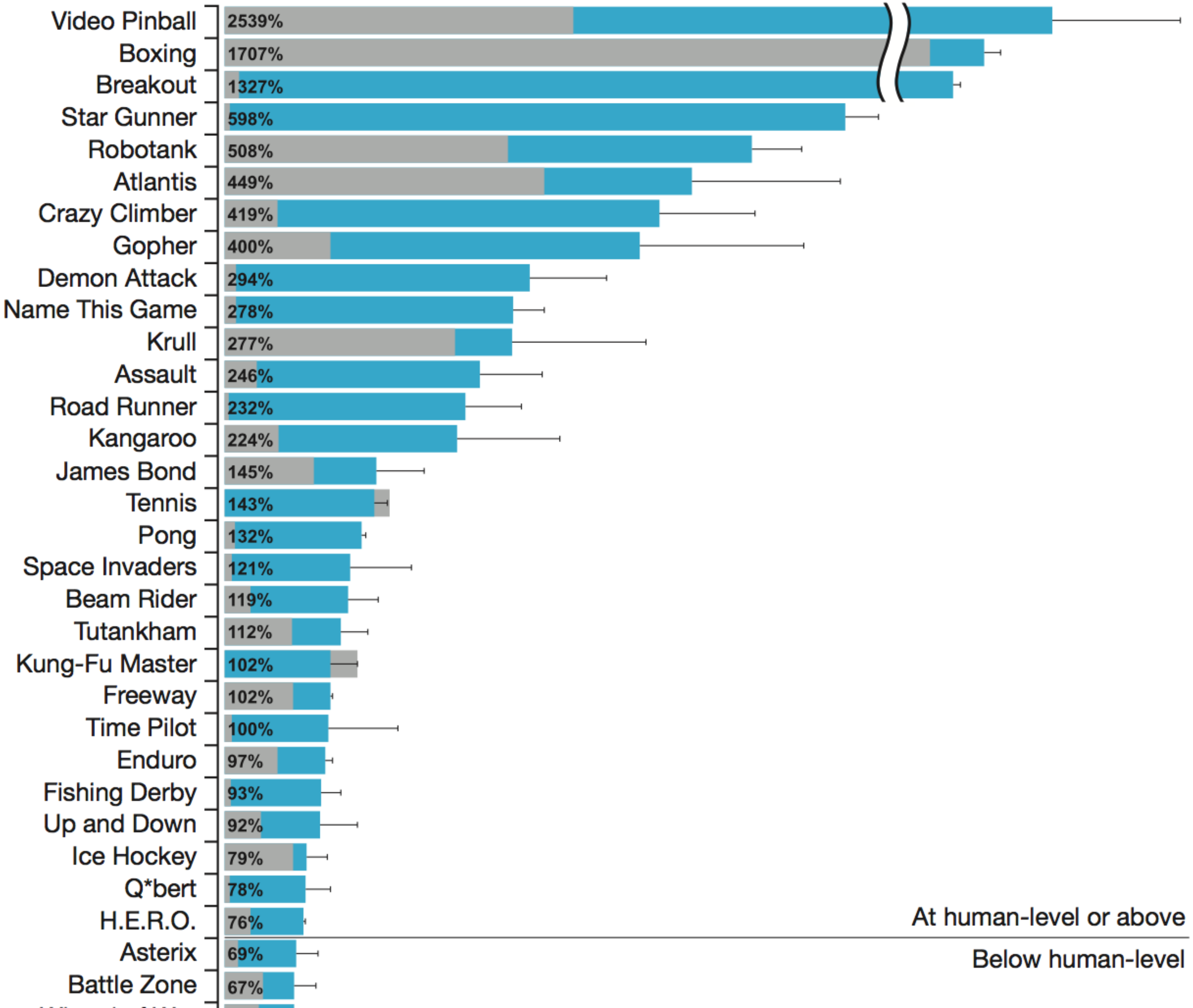
- 49 Atari 2600 games, different network for each game, but same architecture, learning rate and hyperparameters.
- Scores scaling
  - Positive reward – 1 point.
  - Negative reward – -1 point.
  - Reward unchained – 0 points.
- Frame-skipping technique – selecting action on each k-th frame. Used k = 4.
- The values of parameters were selected by informal search on 5 different games.

Evaluation details

- Trained agents were evaluated by playing each game 30 times from up to 5 min each time with different initial random conditions.
- Random agent – baseline comparison.
- The professional human tester:
  - Same emulator engine.
  - Emulator was run at 60 Hz.
  - Audio – disabled.
  - Performance – average reward from 20 episodes of each game lasting 5 min maximum.
  - 2 hours of practice for every game.

Results

- Outperforms the best existing reinforcement learning method in 43 out of 49 games, by the way other approaches use additional prior knowledge about games.
- Furthermore, performs at a level comparable to professional human games tester.
- Achieving more than 75% of the human score on more than half of the games.
- In certain games DQN was able to find long-term strategy.
- Challenging games – games which demand more extended strategy planning.


Video Pinball
Boxing
Breakout
Star Gunner
Robotank
Atlantis
Crazy Climber
Gopher
Demon Attack
Name This Game
Krull
Assault
Road Runner
Kangaroo
James Bond
Tennis
Pong
Space Invaders
Beam Rider
Tutankham
Kung-Fu Master
Freeway
Time Pilot
Enduro
Fishing Derby
Up and Down
Ice Hockey
Q*bert
H.E.R.O.
Asterix
Battle Zone
2539%
1707%
1327%
598%
508%
449%
419%
400%
294%
278%
277%
246%
232%
224%
145%
143%
132%
121%
119%
112%
102%
102%
100%
97%
93%
92%
79%
78%
76%
69%
67%
At human-level or above
Below human-level

Video time!