

Playing Atari with Deep Reinforcement Learning

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou,
Daan Wierstra, Martin Riedmiller

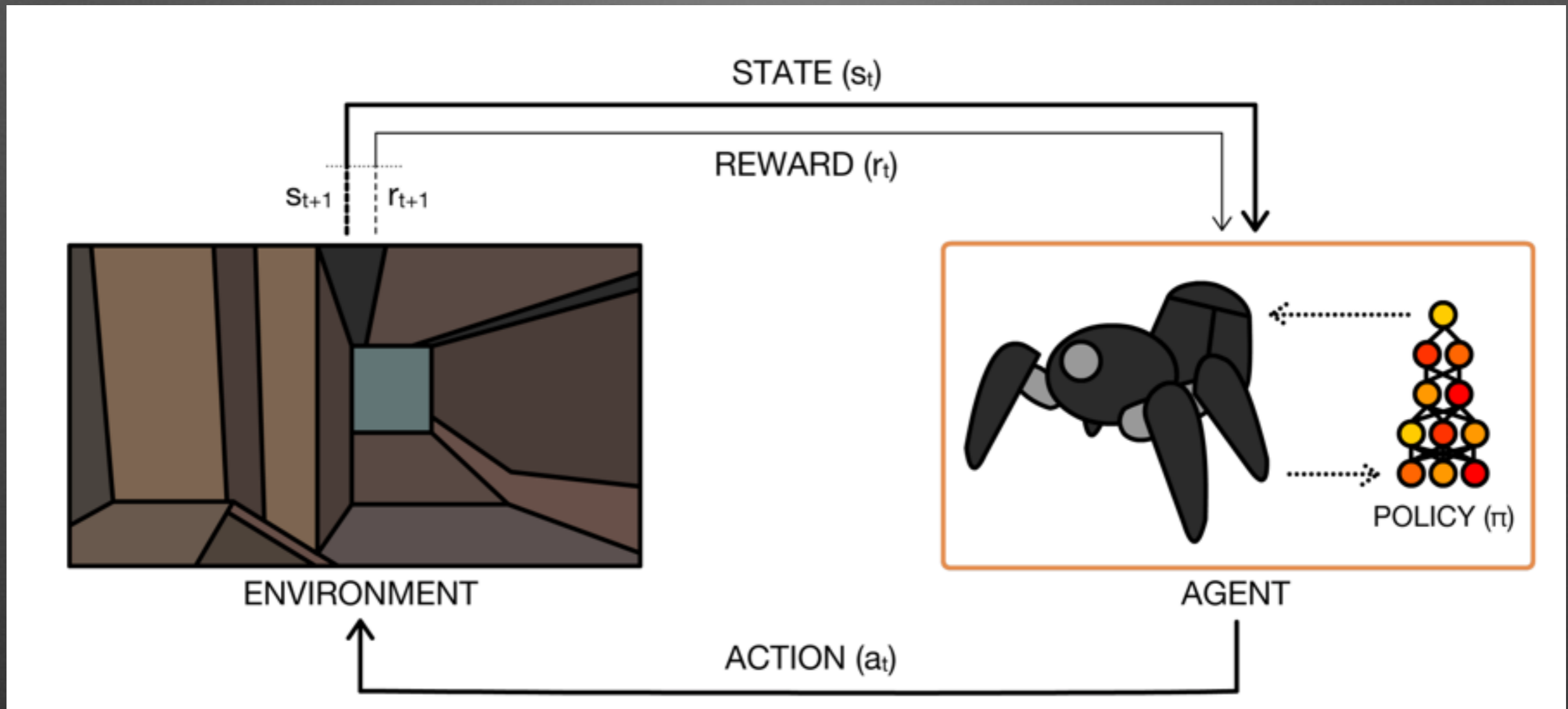
DeepMind (pre-Google)

Results

- end-to-end training
- Model free
- Better than humans for 3 out of 7 games!
- But fails badly for sparse rewards

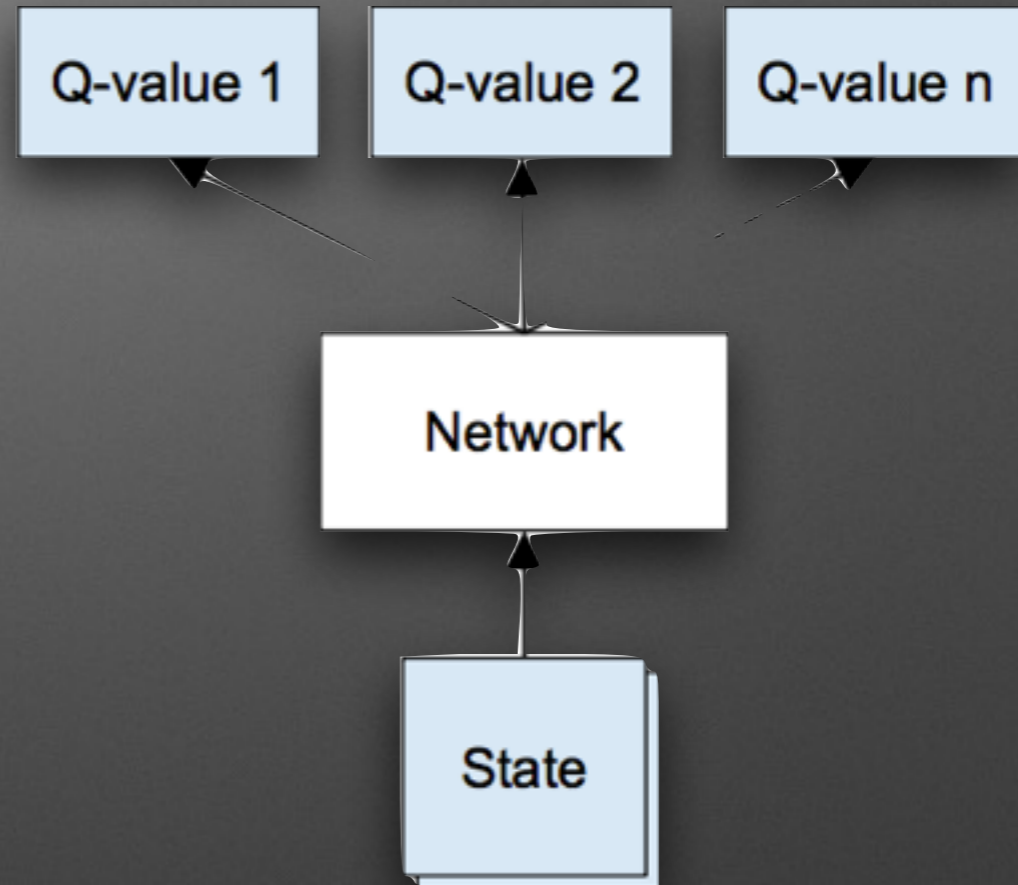


Reinforcement Learning



- Credit Assignment Problem
- Explore-Exploit dilemma

Q-Networks



- Reward function 'given' $R(t)$
- $Q(s, a) = \text{Max}(\text{all } R(t+1))$
- Action we take: $\text{argmax } a \text{ of } Q(s, a)$

Training the Q-network

ConV

ConV

ConV

FullyC

FullyC

Epsilon-greedy

- Simple strategy to pick best action.
- Slightly better: pick randomly with probability epsilon!



Policy Gradients

- End-to-End training
- Single 'Policy network' that we train directly.
- Loss function is modified supervised learning
 - with 'true' labels replaced with the actions we sampled
 - and a 'advantage' term of eventual score