ACTOR CRITIC METHODS IN SIMULATING SIMPLE NATURAL SELECTION ENVIRONMENT

Ümit Akköse
19190

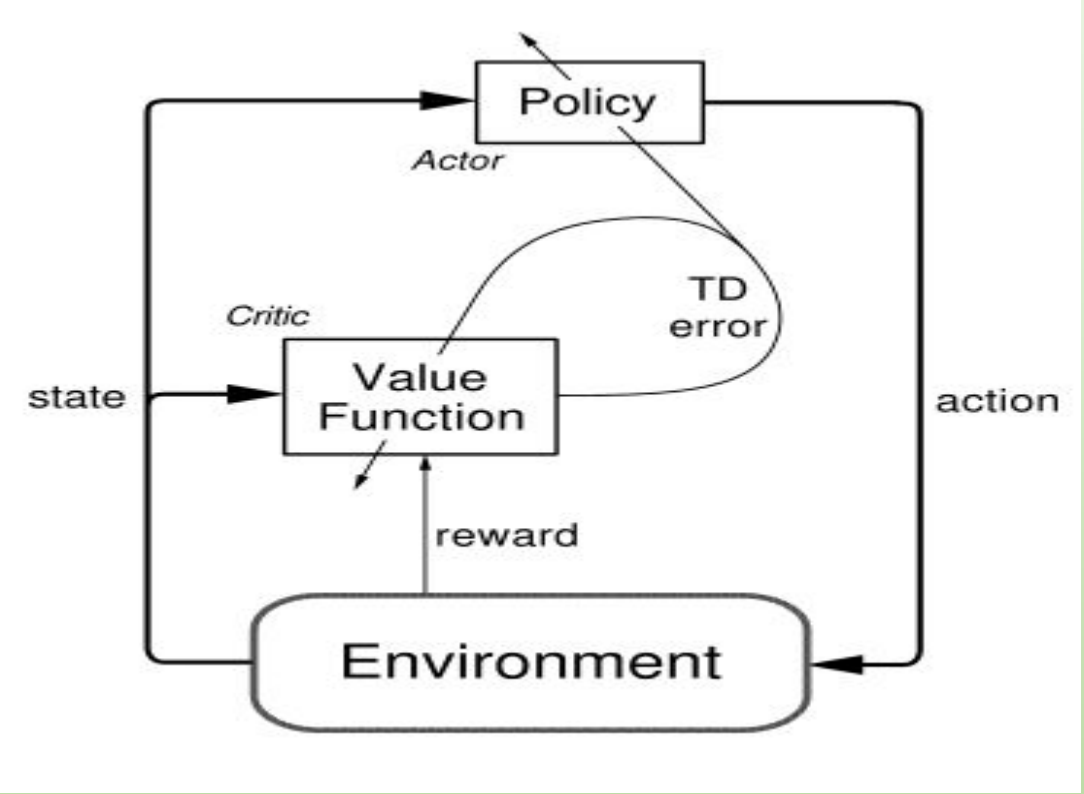
Barış Temel
19233

Veysel Oğulcan Kaya
17804

What is Deep Reinforcement Learning?

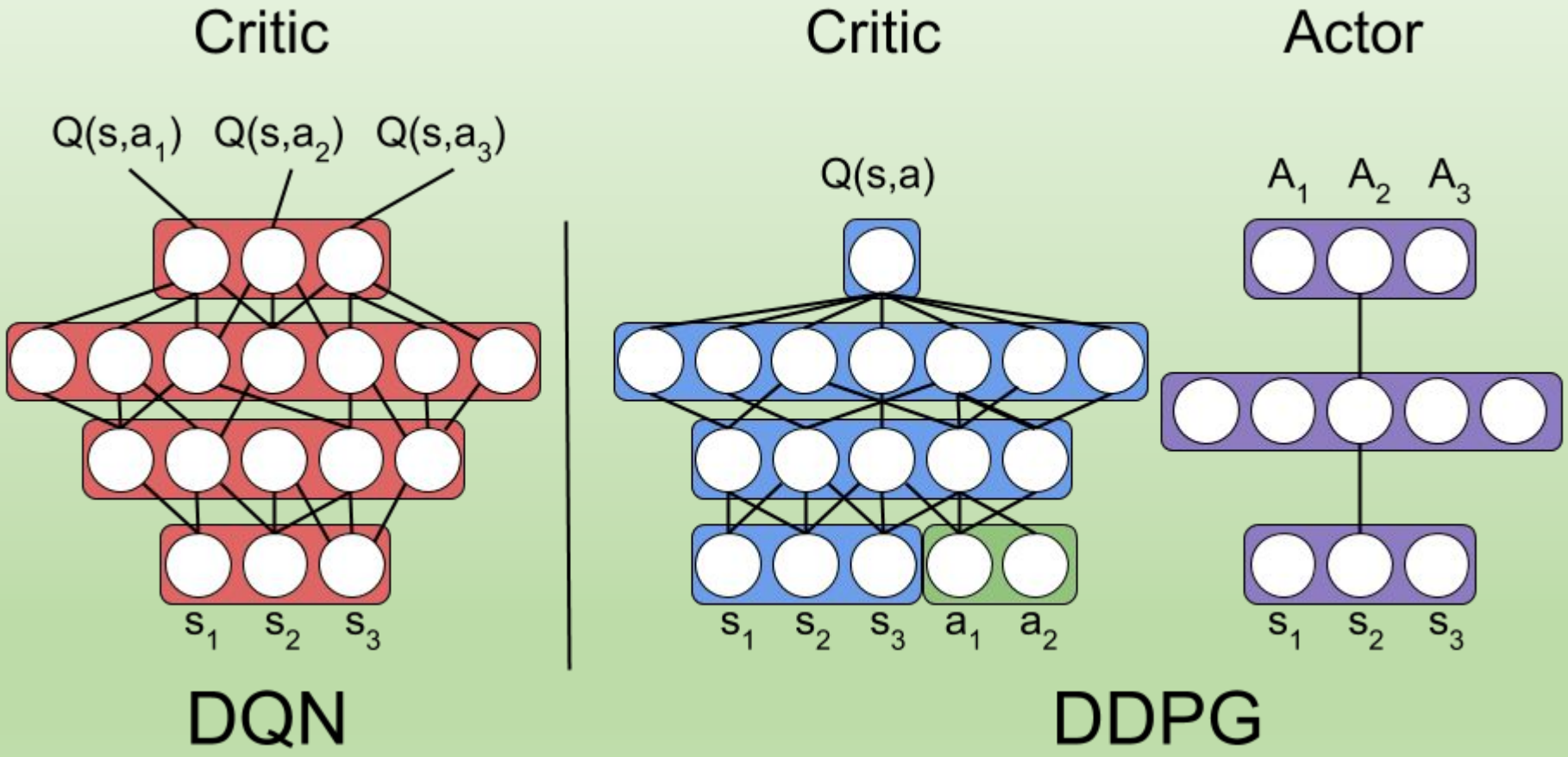
- Intelligent machines can learn from their actions similar to the way humans learn from experience.
- Inherent in this type of machine learning is that an agent is rewarded or penalised based on their actions.
- Actions that get them to the target outcome are rewarded (reinforced).

Actor-Critic Method



- The policy structure is known as the actor, because it is used to select actions, and the estimated value function is known as the critic.
- The critic is a state-value function.

DQN&DDPG STRUCTURE



DDPG Results

The pseudo-algorithm of DDPG

- Initialize actor network μ_θ and critic Q_ϕ with random weights.
- Create the target networks μ_θ 0 and $Q_\phi 0$.
- Initialize experience replay memory D of maximal size N.
- for episode $\in$ [1, M]:
  - Initialize random process ξ.
  - Observe the initial state s_0.
  - for t $\in$ [0, Tmax]:
    - Select the action at = $\mu\theta$(st) + ξ according to the current policy and the noise.
    - Perform the action at and observe the next state st+1 and the reward rt+1.
    - Store (st , at , rt+1, st+1) in the experience replay memory.
    - Sample a minibatch of N transitions randomly from D.
    - For each transition (sk, ak, rk, s0 k) in the minibatch: · Compute the target value using target networks yk = rk + γ Qϕ0(s 0 k , $\mu\theta$ 0(s 0 k)).
    - **Update** the critic by minimizing: L(ϕ) = 1 N X k (yk − Qϕ(sk, ak))2
    - **Update** the actor using the sampled policy gradient: $\nabla\theta J(\theta) = 1\ N\ X\ k\ \nabla\theta\mu\theta(sk) \times \nabla a Q\phi(sk, a)|a=\mu\theta(sk)$
    - **Update** the target networks: $\theta\ 0 \leftarrow \tau\theta + (1 - \tau)\ \theta\ 0\ \phi\ 0 \leftarrow \tau\phi + (1 - \tau)\ \phi$