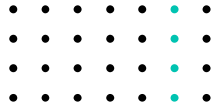
06

DQN 알고리즘

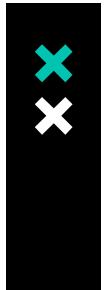
1. 기본개념

DQN 기본개념 가치기반과 정책기반



DQN 알고리즘

- 가치기반 강화학습 : 인공신경망의 학습 대상이 가치, 가치함수를 기반으로 정책 결정(DQN)
- 정책기반 강화학습 : 인공신경망의 학습 대상이 정책, 정책을 직접 학습한다.(REINFORCE, A2C, PPO)



DQN 기본개념

DQN 알고리즘

- 행동가치함수를 인공신경망에 근사해서 학습하는 강화학습 알고리즘

Goal(MSE)

$$J(w) = E_{\pi}\left[(q_{\pi}(S, A) - \hat{q}(S, A, w))^2\right] \quad ①$$

TD

$$\Delta w = \propto \left(\underbrace{R_{t+1} + \gamma \hat{q}(S_{t+1}, A_{t+1}, w)}_{②\text{-}1} - \hat{q}(S_t, A_t, w) \right) \nabla_w \hat{q}(S_t, A_t, w) \quad ②$$

Prediction Error

$$ERR = \underbrace{R_{t+1} + \gamma \hat{q}(S_{t+1}, A_{t+1}, w)}_{③\text{-}1} - \underbrace{\hat{q}(S_t, A_t, w)}_{③\text{-}2} \quad ③$$

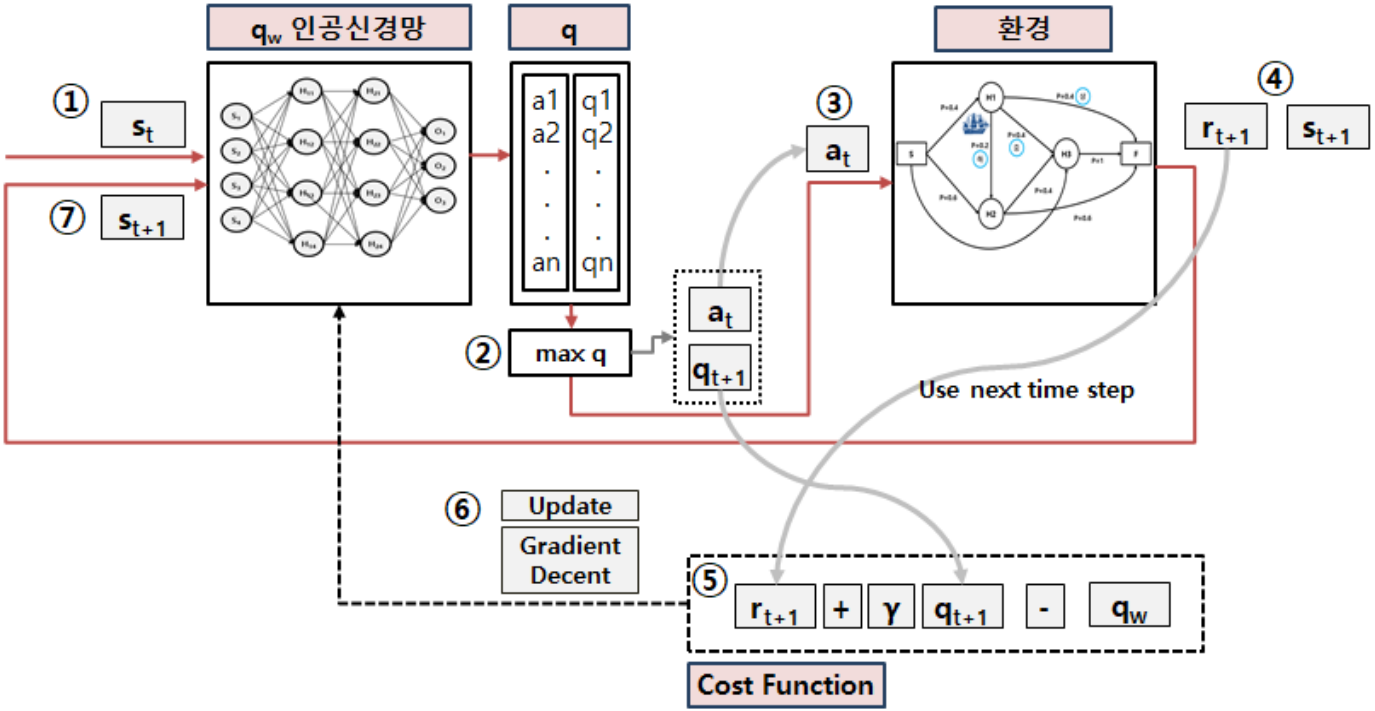
③-1 에이전트를 실행해서 얻은 행동가치함수

③-2 인공신경망에서 예측한 행동가치함수

DQN 기본개념

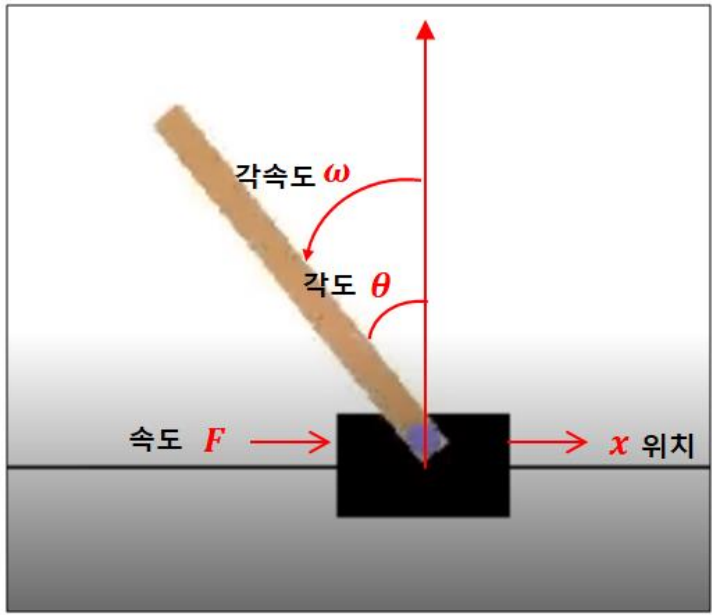
기본 로직

[에이전트 실행]



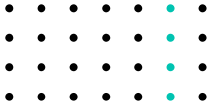
DQN 기본개념 카트폴

카트폴



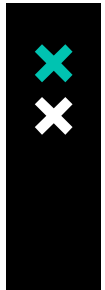
- State :수레의 위치(x), 수레의 속도(F), 막대의 각도(θ), 막대의 각속도(ω)
- Action : 좌/우

DQN 기본개념 탐험과 탐욕

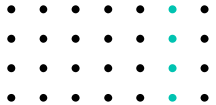


탐험(Exploration)과 탐욕(Exploitation)

- 탐험(Exploration): 에이전트가 다양한 경험을 하는 것
- 탐욕(Exploitation): 에이전트가 누적 보상이 최대가 방향으로 행동하는 것
- 탐험과 탐욕의 문제: 정책이 불완전한 학습 초기에 탐욕 정책에 따라 행동하면 다른 좋은 상태를 경험할 수 없는 문제 발생



DQN 기본개념 탐험과 탐욕

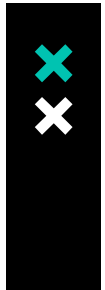


입실론 탐욕 정책

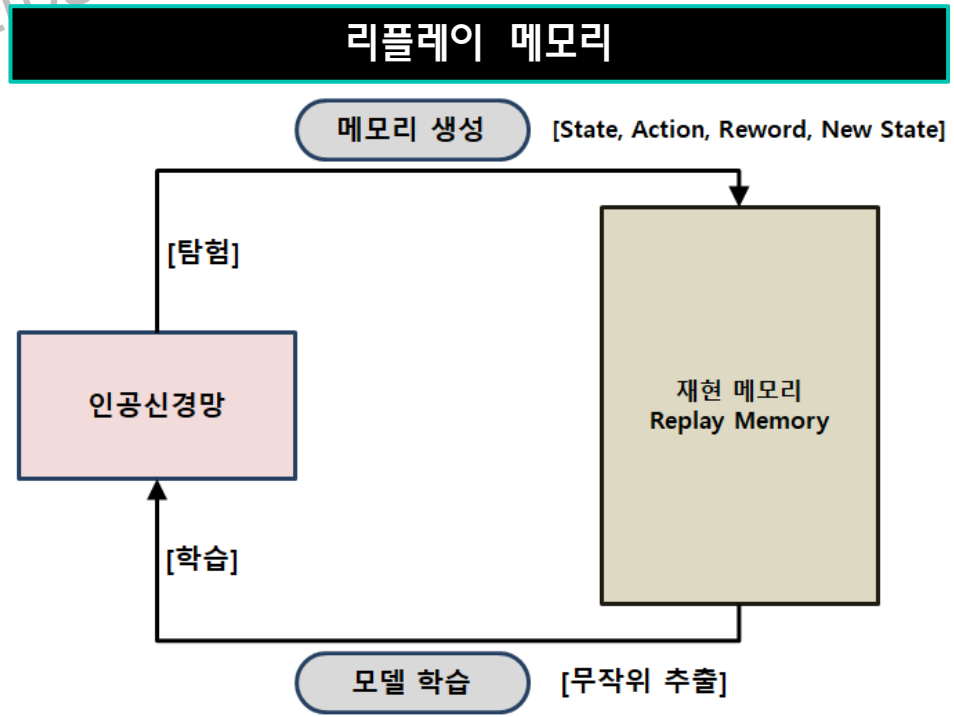
- 에이전트가 탐욕 정책을 사용하면서 다양한 상태를 탐험할 수 있도록 하는 기법
- 입실론 값은 시간이 지남에 따라 줄어들어야 한다.

입실론
탐욕 정책

```
if random value > ε
    argmax(a)
else
    random(a)
```



DQN 기본개념 탐험과 탐욕



- 에이전트가 탐험을 통해 리플레이 메모리 생성
- 리플레이 메모리 생성 과정이 끝나면 리플레이 메모리에서 데이터를 무작위로 추출해 모델을 학습
- 학습 데이터가 시간적인 상관관계를 없애주는 역할