08

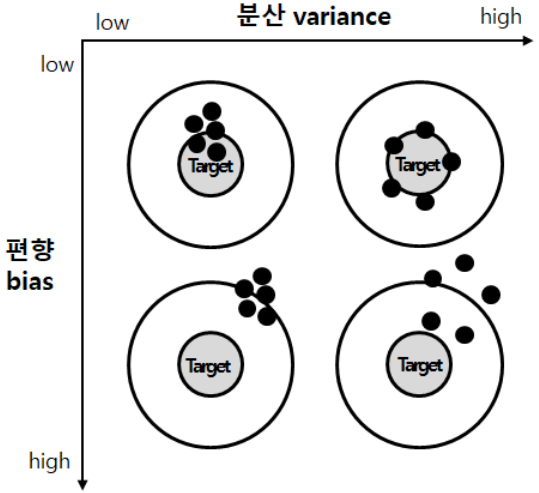
A2C 알고리즘

1. 기본개념

A2C 기본개념

분산과 편향

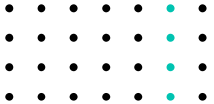
분산과 편향



- 분산(variance)
  - 데이터가 얼마나 넓게 분포하는지를 의미
- 편향(bias)
  - 데이터가 목표 지점에서 얼마나 떨어져 있는지를 의미

A2C 기본개념

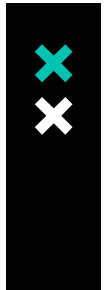
분산과 편향



알고리즘의 분산과 편향

- REINFORCE 알고리즘
  - 하나의 정책으로 에피소드가 끝날 때까지 계속 행동해서 데이터를 수집
  - 데이터의 편향(bias)이 작다.
  - 학습에 사용하는 데이터는 에피소드가 끝날 때까지 수집한 보상의 누적 합
  - 각각의 보상에 들어있는 분산 값도 누적된다.
  - 분산(variance)이 크다.

- DQN 알고리즘
  - 하나의 행동을 해서 수집된 데이터를 바탕으로 전체 데이터를 추정
  - 데이터의 편향(bias)이 크다.
  - 학습 데이터는 하나의 행동에 대한 값이다.
  - 분산(variance)이 작다.



A2C 기본개념

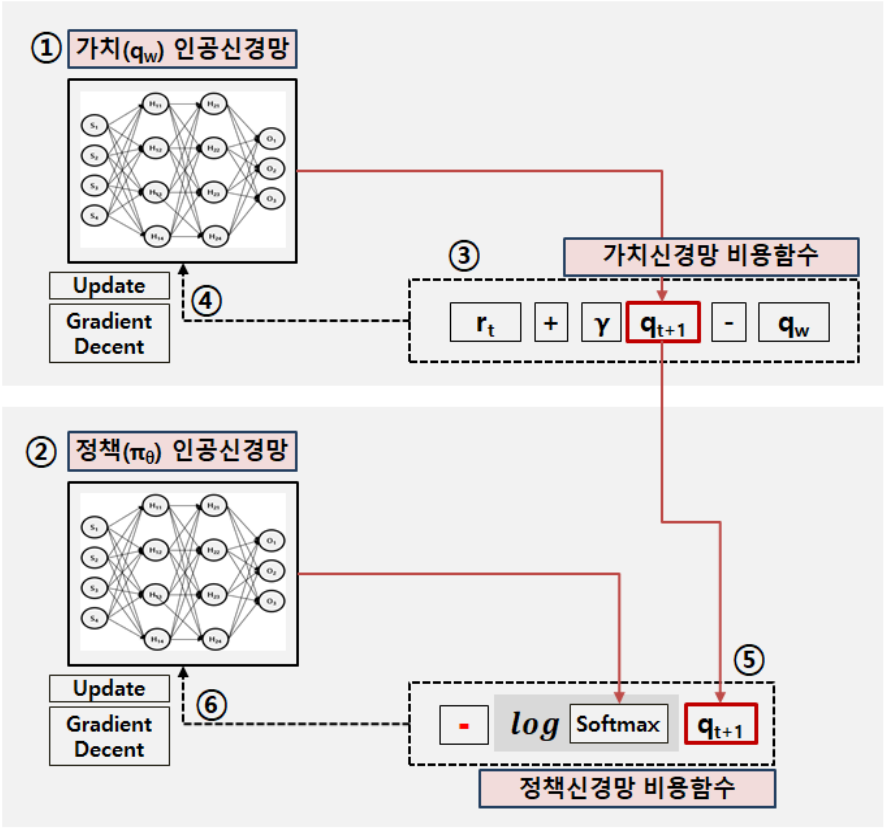
AC 알고리즘

AC(Actor Critic) 알고리즘 개념

- 편향이 작은 REINFORCE 알고리즘의 장점과 분산이 작은 DQN의 장점을 결합
- 정책 신경망과 가치 신경망을 별도로 분리
- 가치 신경망을 사용해서 정책을 통해서 얻을 수 있는 가치를 계산해서 정책을 평가
- 정책 신경망을 사용해서 에이전트의 행동을 결정하는 정책을 계산

A2C 기본개념

AC 알고리즘



AC(Actor Critic) 알고리즘 개념

원스탭 MDP(하나의 타임스텝만 고려하는 MDP) 환경에서는 반환 값이 행동가치함수의 편향되지 않은(unbiased) 샘플

A2C 기본개념

A2C 알고리즘

A2C(Advantage Actor Critic) 알고리즘 개념

- AC 정책신경망에서 REINFORCE 알고리즘은 여전히 분산(variance)이 큰 단점 존재
- 변화를 줄여 주기 위해 베이스라인(Baseline)을 지정해주면 데이터의 분산을 어느 정도 제어가능
- 가치함수를 베이스라인으로 많이 사용
- 행동가치함수에서 가치함수(베이스라인)를 빼서 REINFORCE 알고리즘에서 사용
- 이것을 Advantage라 한다.

A2C 기본개념

A2C 알고리즘

어드밴티지(Advantage) 개념

- 어드벤티지(행동가치함수-가치함수)를 사용하면 기대값에 변화 없이 분산을 줄일 수 있다.

$$A^{\pi\theta}(s, a) = Q^{\pi\theta}(s, a) - V^{\pi\theta}(s) \quad ①$$

$$V^{\pi\theta}(s) \fallingdotseq V_v(s) \quad ②$$

$$Q^{\pi\theta}(s, a) \fallingdotseq Q_w(s, a) \quad ③$$

$$A(s, a) = Q_w(s, a) - V_v(s) \quad ④$$

A2C 기본개념

A2C 알고리즘

어드밴티지(Advantage) 계산

TD

$$V^{\pi\theta}(s) \leftarrow V^{\pi\theta}(s) + \propto(r + \gamma V^{\pi\theta}(s') - V^{\pi\theta}(s)) \quad ①$$

Cost Function

$$\delta = r + \gamma V^{\pi\theta}(s') - V^{\pi\theta}(s) \quad ②$$

기댓값

$$E[\delta^{\pi\theta}|s,a] = \underbrace{E[r + \gamma V^{\pi\theta}(s')|s,a]}_{③\text{-}1} - \underbrace{E[V^{\pi\theta}(s)|s,a]}_{③\text{-}2} \quad ③$$

$$= Q^{\pi\theta}(s, a) - V^{\pi\theta}(s) \quad ④$$

$$= A^{\pi\theta}(s, a) \quad ⑤$$

- 비용함수에 대한 기대값을 구하면 어드벤티지와 동일하다

가치함수와 행동가치함수의 정의

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) q_\pi(s,a)$$

$$q_\pi(s, a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s')$$

A2C 기본개념

A2C 알고리즘