

강화 학습

18-8: 시간차 학습

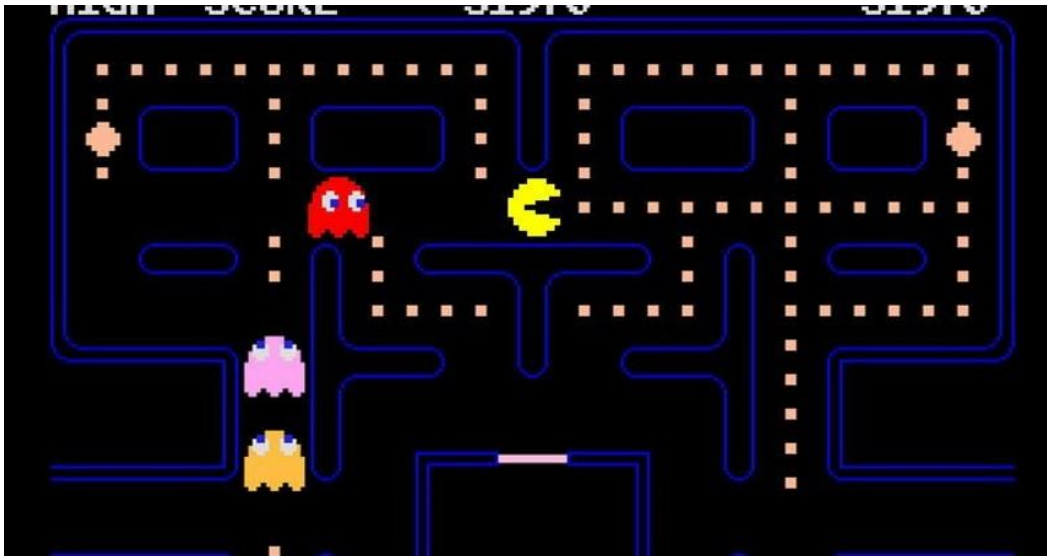
Presentation: Jeiyoon Park

6th Generation, TAVE

18.8 시간차 학습

- 시간차 학습 (Temporal Difference Learning)

- (1) 가정: 에이전트는 초기에 가능한 상태와 행동만 알고 다른건 모름
- (2) 즉, 초기의 에이전트는 전이 확률, $T(s, a, s')$, 과 보상, $R(s, a, s')$, 에 대해 알지 못함




Q-Learning

- 시간차 학습 (Temporal Difference Learning)

(3) 시간차 학습이란 매 타임스텝마다 가치함수를 업데이트 하는 방법

(4) 시간차 예측에서는 다음 스텝의 보상과 가치함수를 샘플링 하여 현재 상태의 가치함수를 업데이트 한다.


$$R + \gamma V(s_{t+1}) - V(s_t)$$

Q-Learning

- 시간차 학습 (Temporal Difference Learning)

$$V(s) \leftarrow V(s) + \alpha(G(s) - V(s))$$



$$v(s) = E[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots | S_t = s]$$

$$= E[R_{t+1} + \gamma(R_{t+2} + \gamma^1 R_{t+3} + \dots) | S_t = s]$$

$$= E[R_{t+1} + \gamma(G_{t+1}) | S_t = s]$$

$$= E[R_{t+1} + \gamma v(S_{t+1}) | S_t = s]$$

반환값이긴 하지만 사실 에이전트가 실제로 받은 보상이 아직은 아님.
따라서 가치함수 형태로 나타낼 수 있음



$$V(S_t) \leftarrow V(S_t) + \alpha(R + \gamma V(S_{t+1}) - V(S_t))$$

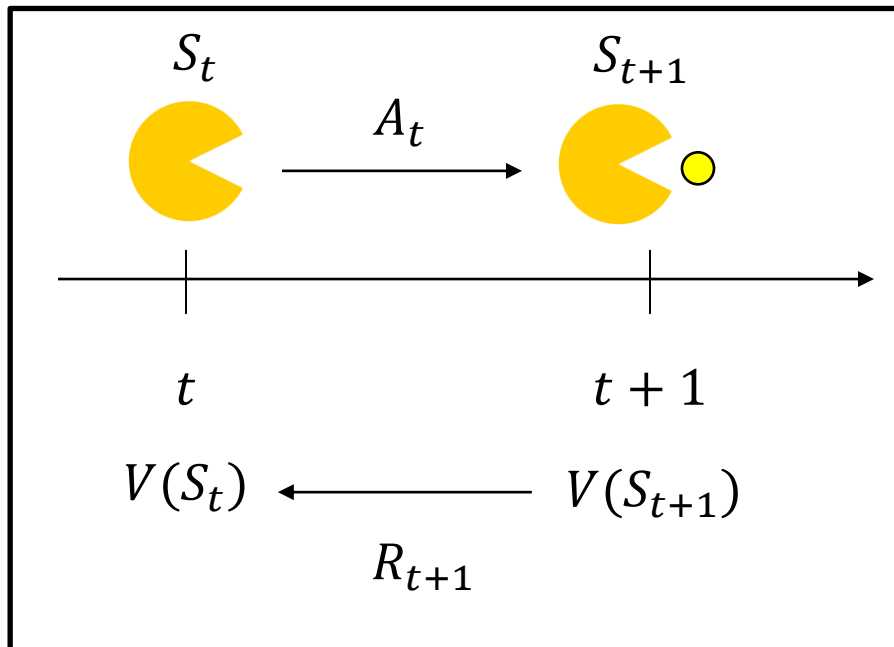
여기서 $R + \gamma V(S_{t+1})$ 를 시간차 에러(Temporal-difference error)라고 함

Q-Learning

- 시간차 학습 (Temporal Difference Learning)

- 따라서 시간차 예측은 어떤 상태에서 행동을 하면 보상을 받고 다음 상태를 알게되고 다음 상태의 가치함수와 알게된 보상을 더해 그 값을 업데이트의 목표로 삼는다는 것. 이 과정을 반복

$$V(S_t) \leftarrow V(S_t) + \alpha(R + \gamma V(S_{t+1}) - V(s))$$

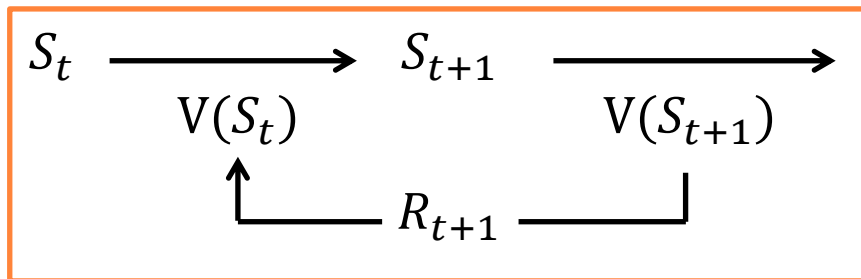


Q-Learning

- 시간차 학습 (Temporal Difference Learning)

(5) 시간차 은 매 타임스텝마다 현재 상태에서 하나의 행동을 하고 환경으로 부터 보상을 받고 다음 상태를 알게 됨

(6) 다음 상태의 예측값을 통해 현재의 가치함수를 업데이트 하는 방식을 강화학습에서는 **부트스트랩(Bootstrap)**이라고 함. 즉, 목표가 정확하지 않은 상태에서 현재의 가치함수를 업데이트 함



Thank you



<https://jeiyoon.github.io/>