강화학습 실습

목차

❖ Introduction

❖ REINFORCE

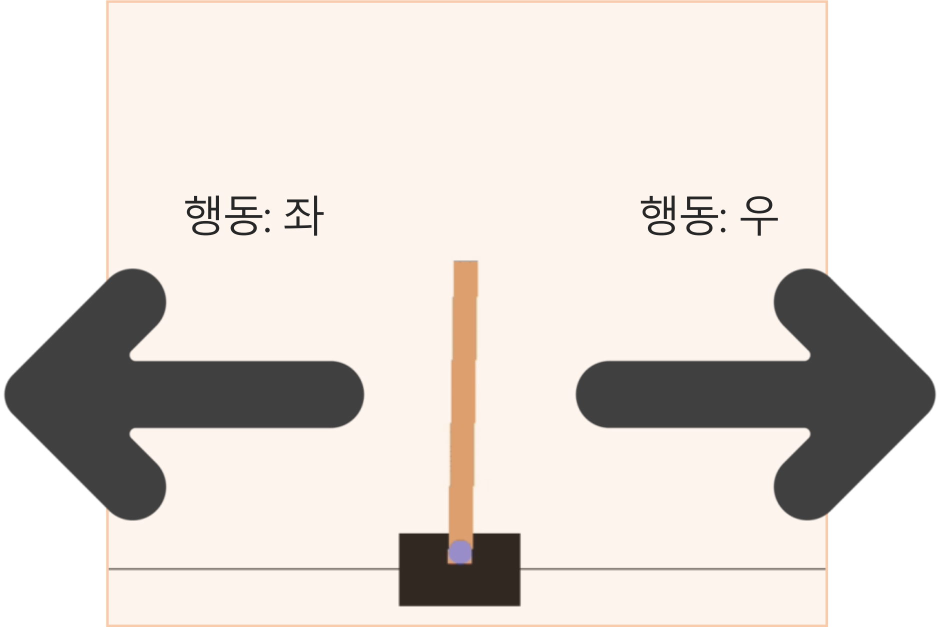
❖ Deep Q-Network

강화 학습 실습

Jupyter Notebook 실습

❖ Introduction

- CartPole
  a. 상태: cart의 위치, 속도 등
  b. 행동: 좌 & 우



- Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.

강화 학습 실습

Jupyter Notebook 실습

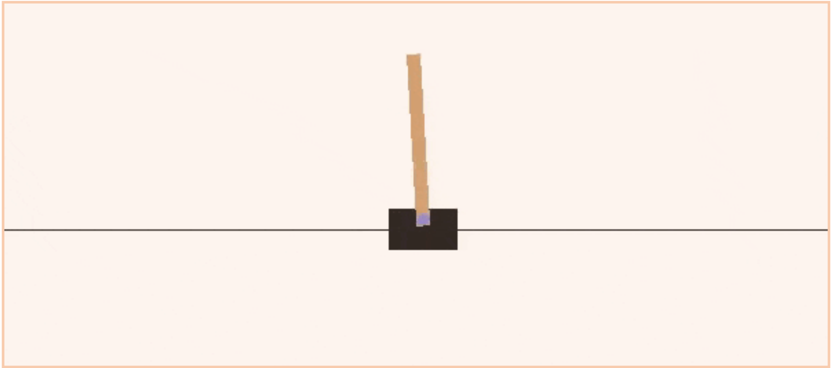
❖ Introduction

- CartPole

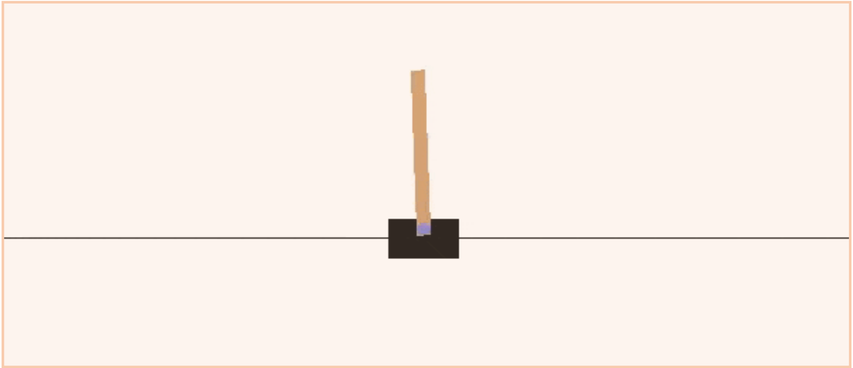
    a. CartPole 게임에 대한 상세 설명

    b. 관측 상태, 행동, 보상 등 상세 설명

게임의 특성



게임의 목표



- Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.

DMQA

강화 학습 실습

Jupyter Notebook 실습

❖ Introduction

- CartPole

  a. CartPole 게임에 대한 상세 설명

  b. 관측 상태, 행동, 보상 등 상세 설명

- **Observation :** [x, θ, dx/dt, dθ/dt]
  - x : track 상에서 cart의 위치
  - θ : pole과 normal line과의 각도
  - dx/dt : cart의 속도
  - dθ/dt : θ의 각속도

- **Ending condition(of episode)**
  1) θ가 15˚이상
  2) 원점(O: cetroid of track)으로부터의 거리가 2.4 units이상

- **Action :** cart의 가하는 힘의 방향 (0 or 1)
- **Reward :** episode가 유지되는 시간
- **Objective : Ending condition을 피하며 reward를 최대로(pole의 균형을 오랫동안 유지)**

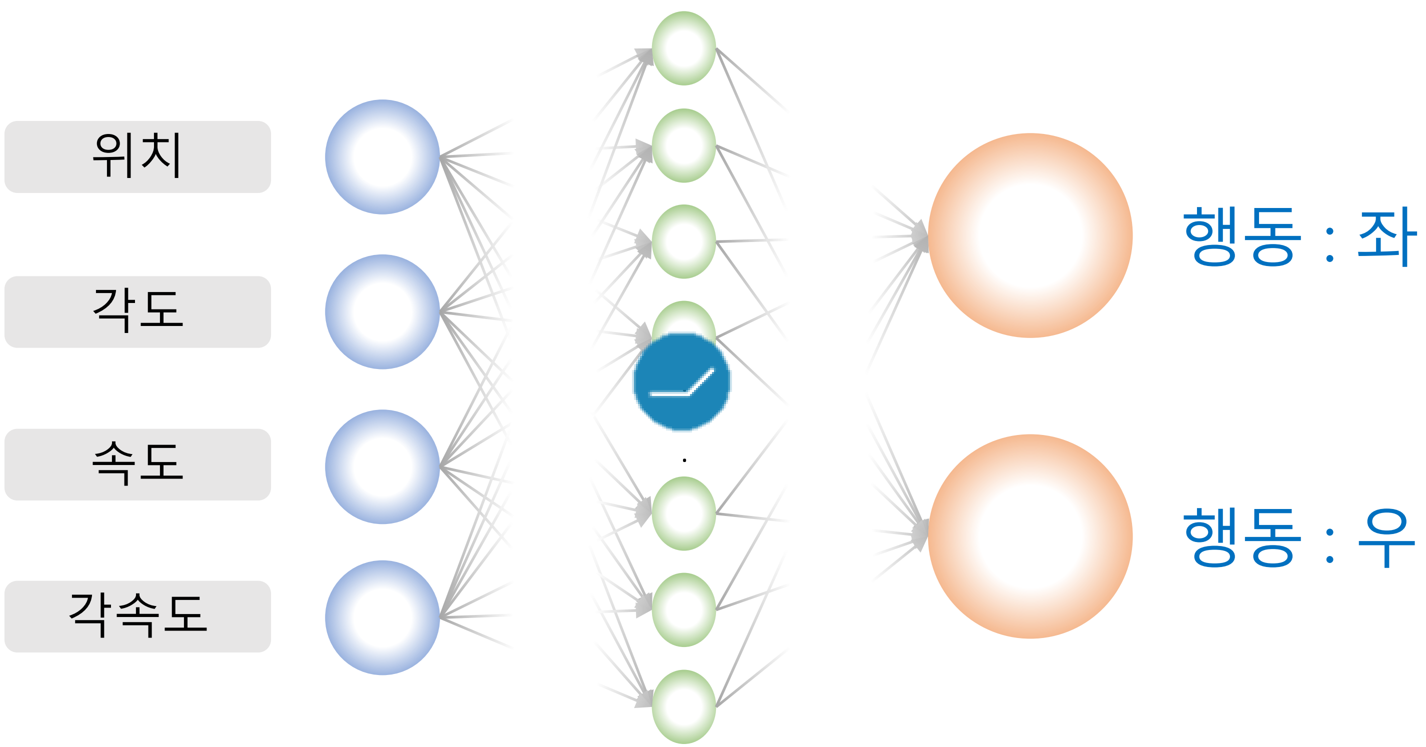
- Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.

강화 학습 실습

Jupyter Notebook 실습

❖ Introduction

- REINFORCE & Deep Q-Network 네트워크 구조



- Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.

강화 학습 실습

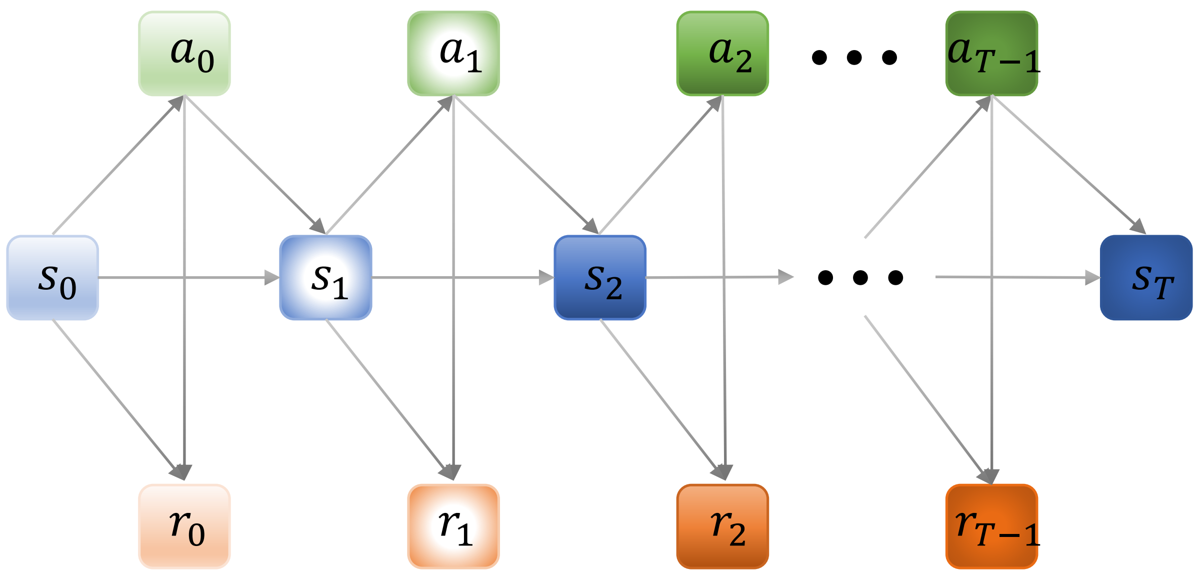
Jupyter Notebook 실습

❖ REINFORCE

- Weight update

a. 에피소드(τ) = $s_0, a_0, r_1, s_1, a_1, r_2, s_2, \dots, s_T$

b. $J(\theta) = E[\sum_{t=0}^{T-1} r_{t+1} | \pi_\theta] = E[r_1 + r_2 + r_3 + \cdots + r_\tau | \pi_\theta]$



- Sutton, R. S., McAllester, D. A., Singh, S. P., & Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In Advances in neural information processing systems (pp. 1057-1063).

강화 학습 실습

Jupyter Notebook 실습

❖ REINFORCE

- Weight update

a. $J(\theta) = E[\sum_{t=0}^{T-1} r_{t+1} | \pi_\theta] = E[r_1 + r_2 + r_3 + \cdots + r_\tau | \pi_\theta]$

b. $\theta' = \theta + \alpha \nabla_\theta J(\theta)$, $\nabla_\theta J(\theta)$ = Policy Gradient

$$\nabla_\theta E[\sum_{t=0}^{T-1} r_{t+1} | \pi_\theta] = E_\tau \nabla_\theta [\sum_{t=0}^{T-1} log\pi_\theta(a_t|s_t) r_{t+1}]$$

$$\approx E_\tau [\nabla_\theta \sum_{t=0}^{T-1} log\pi_\theta(a_t|s_t) G_t],$$

$$where\ G_t = \sum_{t=0}^{T-1} \gamma^t r_{t+1} = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots + \gamma^{T-1} r_T$$

Discounted G_t → 단순 보상의 합 발산 방지

- Sutton, R. S., McAllester, D. A., Singh, S. P., & Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In Advances in neural information processing systems (pp. 1057-1063).

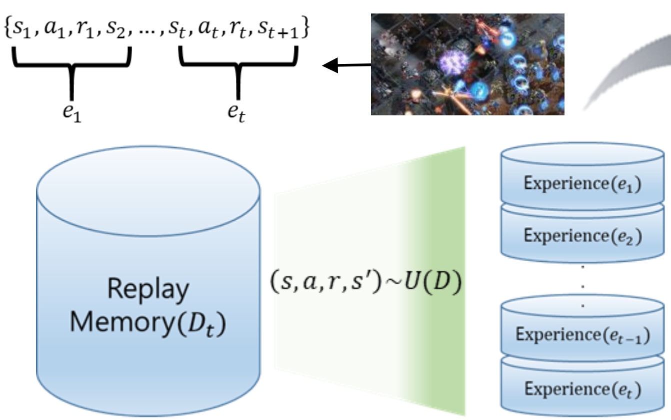
강화 학습 실습

Jupyter Notebook 실습

❖ Deep Q-Network

- Weight update
  a. Target network의 가중치는 매 C step마다 Train network의 가중치로 대체(C은 사용자 정의)
  b. C번의 iteration동안 Q-함수 업데이트 시 Target 움직임 방지

Q-함수 업데이트 $\nabla_{w_i}\mathcal{L}(w_i) = E[(r + \gamma \max_{a'} Q(s', a', w_{i-1}) - Q(s, a, w_i))\nabla_{w_i} Q(s, a, w_i)]$



$$\mathcal{L}_i(w_i) = E_{(s,a,r,s')\sim U(D)}\left[\left(r + \gamma \max_{a'} Q(s', a'; w_i) - Q(s, a; w_i)\right)^2\right]$$

C step 마다

$$\mathcal{L}_i(w_i) = E_{(s,a,r,s')\sim U(D)}\left[\left(r + \gamma \max_{a'} Q(s', a'; w_i^-) - Q(s, a; w_i)\right)^2\right]$$

Target y_i

- Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.

DMQA

감사합니다

DMQA