Chatpor18

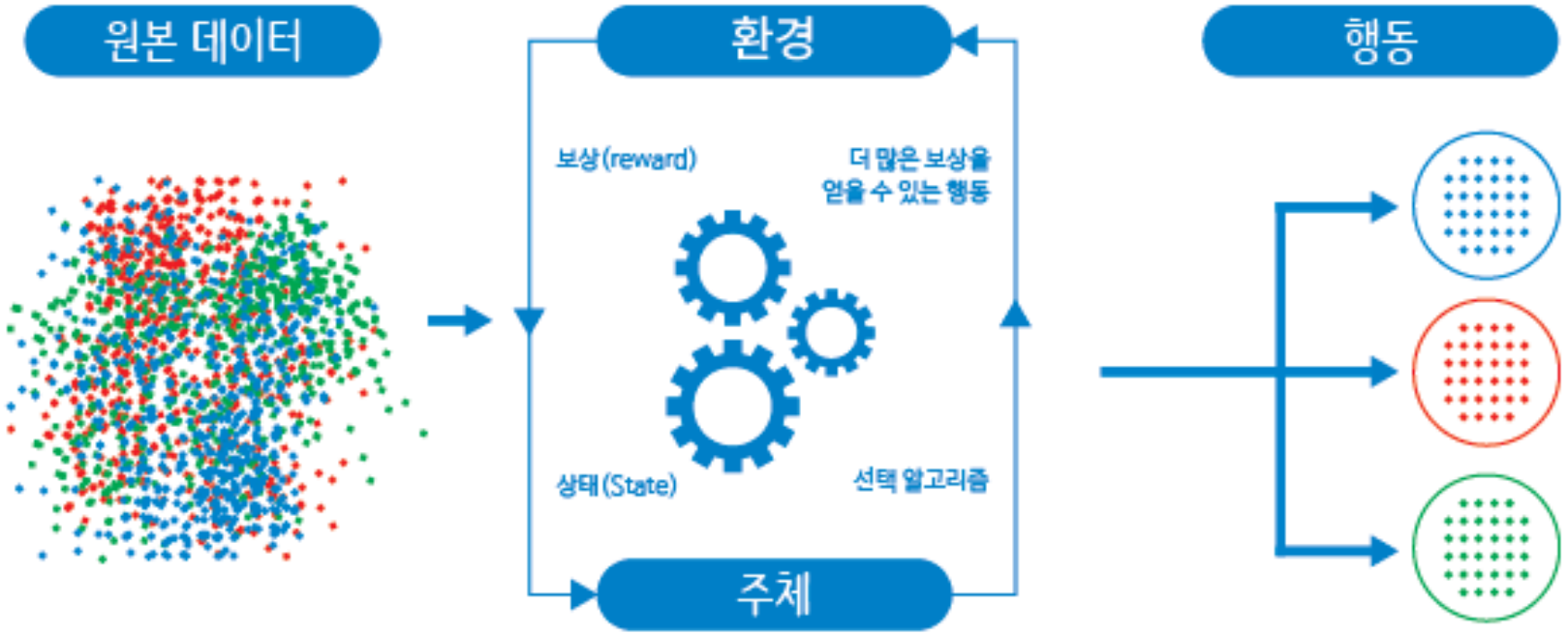
22.10.19

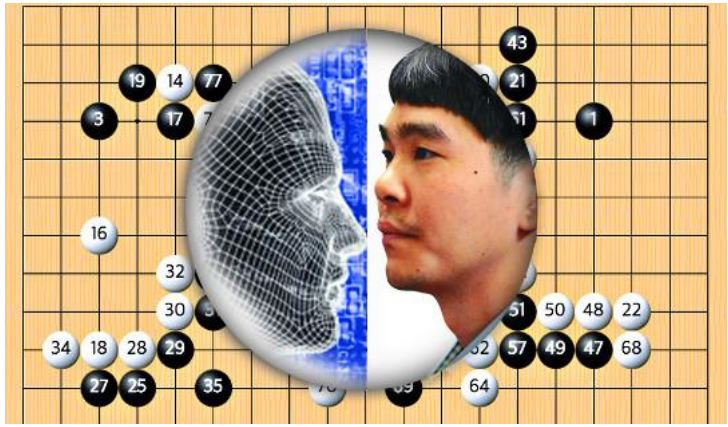
화학안전연구센터 최지원

목차

강화학습

- 정직인 환경에서 학습을 진행하는 지도/비지도학습과 달리, 어떤 환경 안에서 정의된 주체(agent)가 현재 상태(state)를 관찰하여 선택할 수 있는 행동(action)들 중 가장 최대 보상(reward)을 가져다 주는 행동을 학습하는 것
- 대표적으로 알파고가 있음

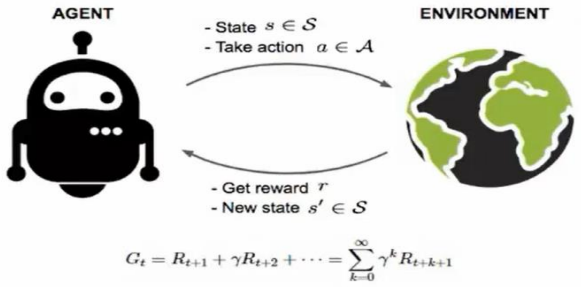




출처: http://tcpschool.com/

정책 탐색

- **1) 정책의 개념**
  - 주체가 행동을 결정하기 위해 사용하는 알고리즘
  - 주체가 어떤 상태에서 행동을 취하게 될 때 상태에 맞게 취할 수 있는 행동을 연결시켜주는 함수
- **2) 확률적 정책**
  - 어떤 상태에 대한 행동들의 확률분포를 반환
- **3) 유전 알고리즘**
  - 전자생존 이론을 기반으로 한 최적화 기법
  - 부모/자식 개념 -> 성능이 낮은 정책은 버리고 살아 있는 정책에서 자식 정책을 생산하게 함
  - 좋은 정책을 찾을 때까지 여러 세대를 걸쳐 반복
- **4) 정책 그레이디언트(Policy gradient, PG)**
  - 정책 파라미터에 대한 보상의 그래디언트를 평가한 후 높은 보상의 방향을 따르는 그래디언트로 파라미터를 수정하는 최적화 기법
  - 주체를 훈련하기 위한 최소한 환경 구축 -> OpenAI 짐* 이용



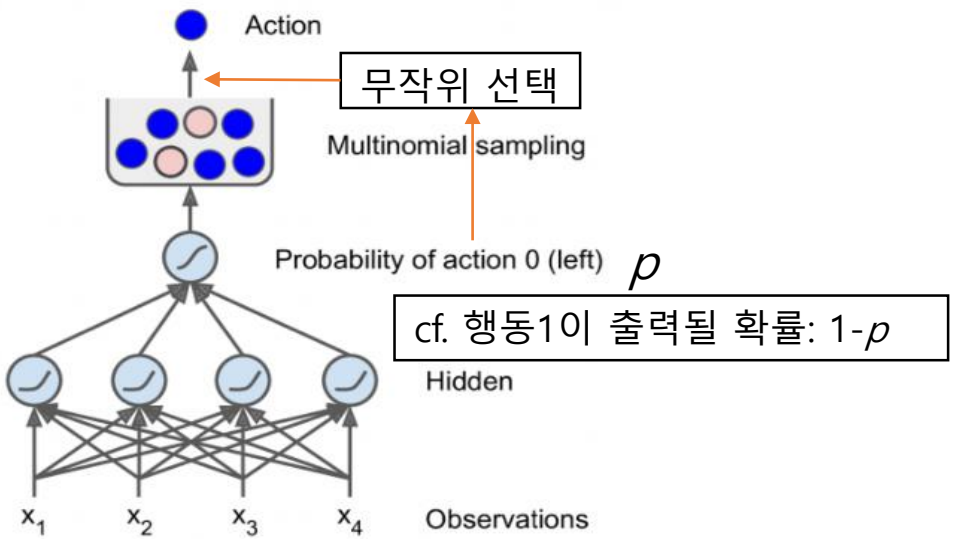
$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

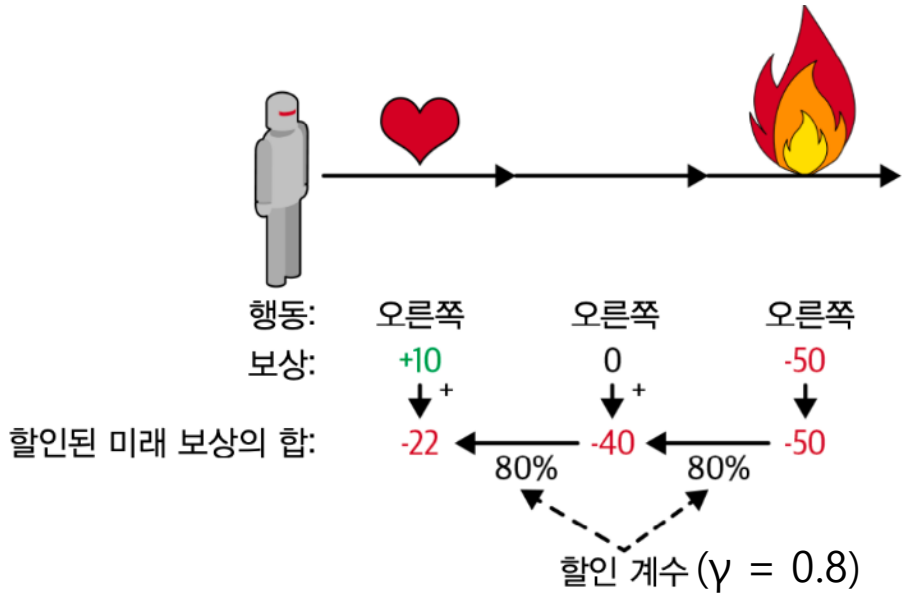
G_t (Return)를 Maximize 하는 최선의 정책 π를 찾는 것!

*Open AI 짐: 다양한 종류의 시뮬레이션환경을 제공하는 툴킷

신경망 정책

- **신경망 정책**
  - 관측을 입력으로 받고 실행할 행동에 대한 확률을 추정하여 출력하는 신경망
  - 신경망이 만든 확률을 기반으로 랜덤하게 행동을 선택
- **행동평가(신용 할당 문제)**
  - 주체가 보상을 받았을 때 어떤 행동 때문에 받는건지 알 수 없음
  - 행동이 일어난 후 각 단계마다 할인 계수 γ를 적용한 보상을 모두 합한 결과로 행동 평가
  - 할인된 보상의 합 = 행동의 대가



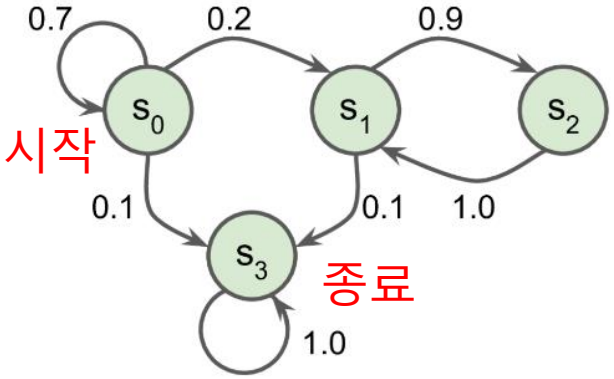


정책 그레이디언트

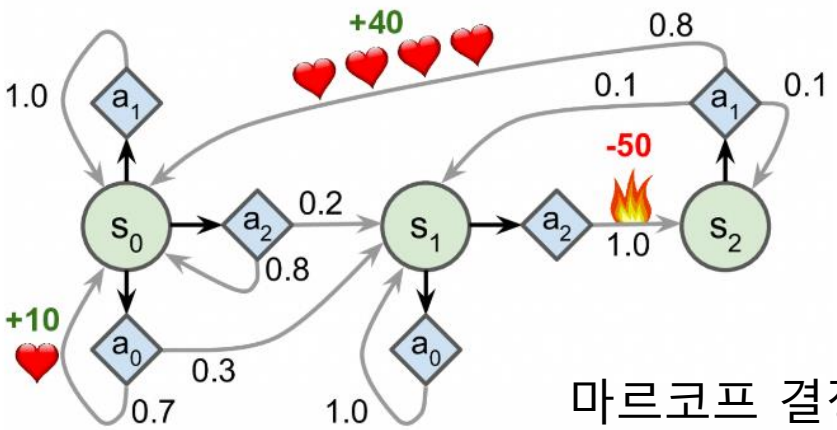
- **높은 보상을 얻는 방향의 그레이디언트를 따르도록 정책의 파라미터를 최적화하는 알고리즘**
  - 1) 신경망 정책이 여러 번에 걸쳐 게임을 플레이하고 매 스텝마다 선택된 행동이 더 높은 가능성을 가지도록 그레이디언트 계산
  - 2) 행동의 점수를 할인율을 적용하여 계산하고, 여러 번의 게임에 걸친 많은 행동들에 대한 점수 정규화
  - 3) 각 파라미터에 대해 [그레이디언트*점수]를 평균내어 그레이디언트를 계산하고, 경사 하강법 스텝을 수행
  - 나쁜 행동은 점수가 음수가 되므로 해당 행동을 행하지 않는 방향으로 경사 하강법 스텝 수행
- **보상을 증가시키기 위해 정책을 직접적으로 최적화 함**

마르코프 결정 과정(Markov decision process, MDP)

- **1) 마르코프 연쇄**
  - 정해진 개수의 상태를 가지고 있음
  - 각 단계마다 시스템은 상태를 유지하거나 시스템의 상태가 바뀌게 되는데 상태의 변화를 전이라고 함
  - 종료상태: 다른 상태로의 전이가 더 이상 일어나지 않음
- **2) 마르코프 결정 과정**
  - 마르코프 연쇄와 비슷하지만 각 스텝에서 주체는 여러 가능한 행동 중 하나를 선택할 수 있고, 행동에 따라 전이 확률이 달라짐
  - 어떤 전이는 보상을 반환하며 주체는 보상을 최대화하는 정책을 찾음



마르코프 연쇄



마르코프 결정 과정

S: 상태
a: 행동

마르코프 결정 과정(Markov decision process, MDP)

- **1) 벨먼 최적 방정식**
  - 어떤 상태 s 의 최적의 상태 가치 V*(s)를 추정하는 방법
  - 주체가 상태 s에 도달한 후 최적으로 행동한다고 가정하고 평균적으로 기대할 수 있는 할인된 미래 보상의 합
  - 가능한 모든 상태에 대한 최적의 상태 가치를 정확하게 추정할 수 있도록 도와주는 방정식
  - 모든 상태 가치를 0으로 초기화 -> 가치반복 알고리즘을 사용하여 반복적으로 업데이트

$$V^*(s) = \max_a \sum_{s'} T(s,a,s')\left[R(s,a,s') + \gamma \cdot V^*(s')\right] \quad \text{for all } s$$

벨먼 최적 방정식

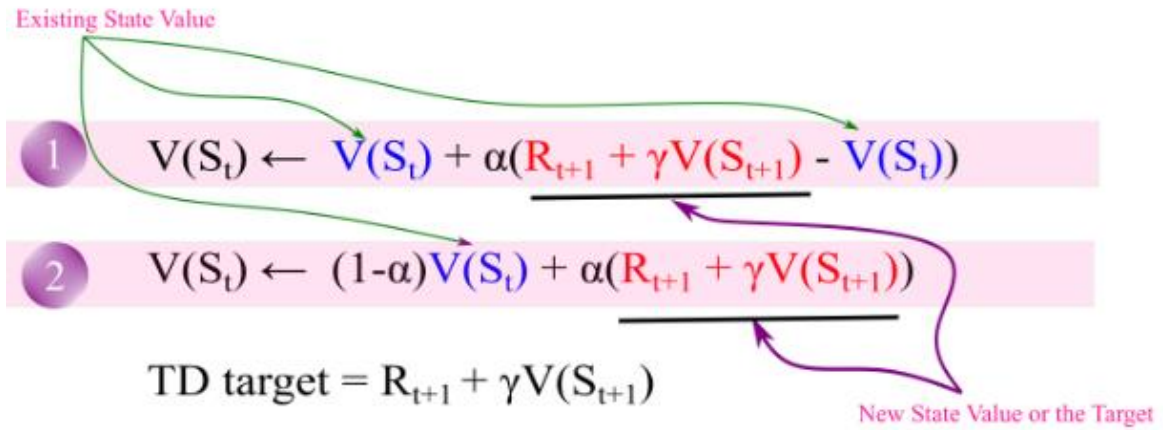
$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s')\left[R(s,a,s') + \gamma \cdot V_k(s')\right] \quad \text{for all } s$$

가치 반복 알고리즘

- T(s,a,s′): 주체가 행동 a를 선택했을 때 상태 s에서 상태 s′로 전이될 확률
- R(s,a,s′): 주체가 행동 a를 선택해서 상태 s에서 상태 s′로 전이했을 때 받을 수 있는 보상
- γ: 할인 계수

시간차 학습(temporal difference learning, TD 학습)

- **가치반복 알고리즘과 비슷하나 주체가 MDP에 대해 일부 정보만 알고 있을 때 다룰 수 있도록 변형**
  - 독립적인 행동으로 이루어진 강화 학습 문제는 보통 마르코프 결정 과정으로 모델링 될 수 있지만 초기 에이전트는 전이 확률과 보상이 얼마나 되는지 알지 못함
  - 보상에 대해 알기 위해서는 적어도 한번은 각 상태와 전이를 경험해봐야 함
  - 전이 확률에 대해 신뢰할 많나 추정을 얻으려면 여러 번 경험해야 함
- **주체가 탐험 정책을 사용해 MDP를 탐험하여 탐험이 진행될수록 TD 학습 알고리즘이 실제 관측된 전이와 보상에 근거하여 상태 가치 추정값 업데이트**



Q-러닝

- 전이 확률과 보상을 모르는 초기 상황에서 Q-가치 반복 알고리즘 적용
- 주체의 플레이를 보고 점진적으로 Q-가치 추정을 향상하는 방식으로 작동
- 정확한 Q-가치 추정을 얻으면 최적의 정책은 가장 높은 Q-가치를 가진 행동을 선택하는 것(그리디 정책)

https://colab.research.google.com/drive/19ayCIgntfMt4m3D3gvXZbQqdJWZpYXun#scrollTo=HcQk17uPpLZP