REINFORCEMENT LEARNING

강화학습특론

-1주차-

Reinforcement Learning Overview

02

Reinforcement Learning: Characteristics, Formulation & Terminology

학습내용

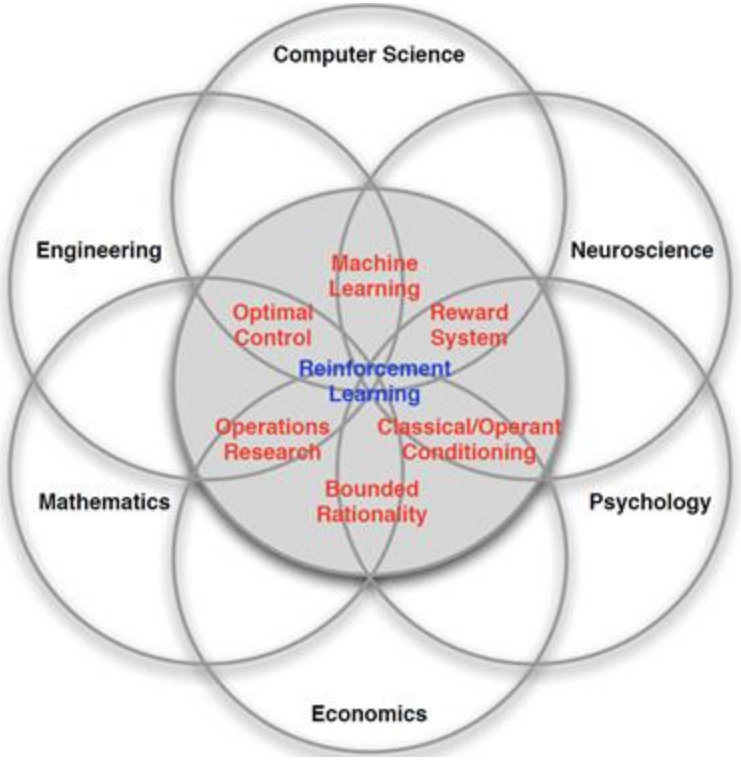
- Reinforcement Learning: Characteristics and Formulation
- Terminology: Agent & Environment, History & State, Rewards & Actions
- Markov State
- Fully Observable Environments
- Partially Observable Environments

Reinforcement Learning (RL):
Characteristics and Formulation

- The science and engineering of decision making and learning under uncertainty
- A type of machine learning that models learning from experience in a wide range of applications

Reinforcement Learning (RL):

Characteristics and Formulation



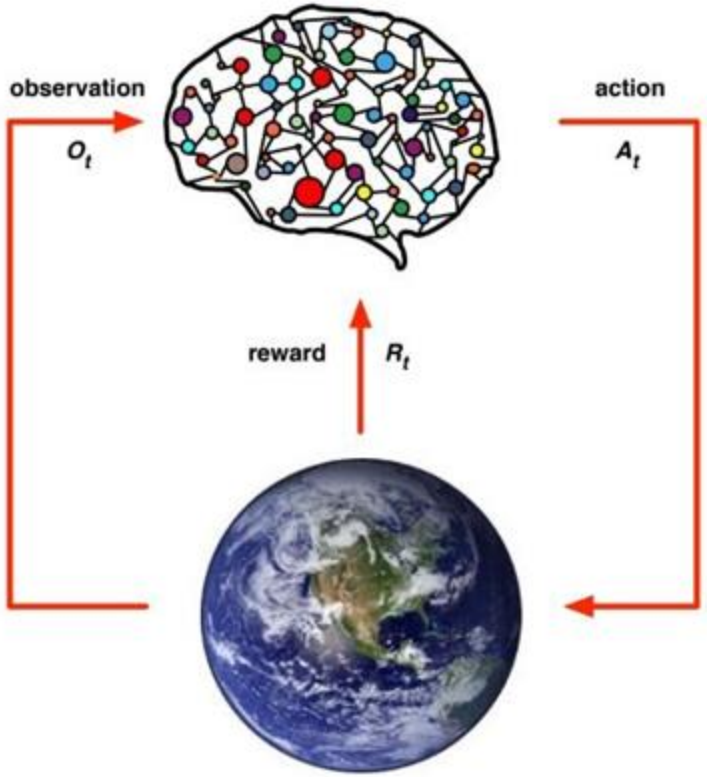
Relationship with Other Domains

Reinforcement Learning (RL):
Characteristics and Formulation

- What makes reinforcement learning different from other machine learning paradigms?
  - There is no supervisor, only a reward signal
  - Feedback is delayed, not instantaneous
  - Time really matters (sequential, non- independent identically distributed data)
  - Agent's actions affect the subsequent data it receives

Reinforcement Learning (RL):

Characteristics and Formulation



- At each time step t the agent:
  - Executed action A_t
  - Receives observation O_t
  - Receives scalar reward R_t
- The environment:
  - Receives action A_t
  - Emits observation O_{t+1}
  - Emits scalar reward R_{t+1}
- t increments at environment step

Reinforcement Learning (RL):

Characteristics and Formulation

Goal

Sequential Decision Making: select actions to maximize total future reward

- Actions may have long term consequences
- Reward may be delayed
- It may be better to sacrifice immediate reward to gain more long-term reward

Reinforcement Learning (RL):

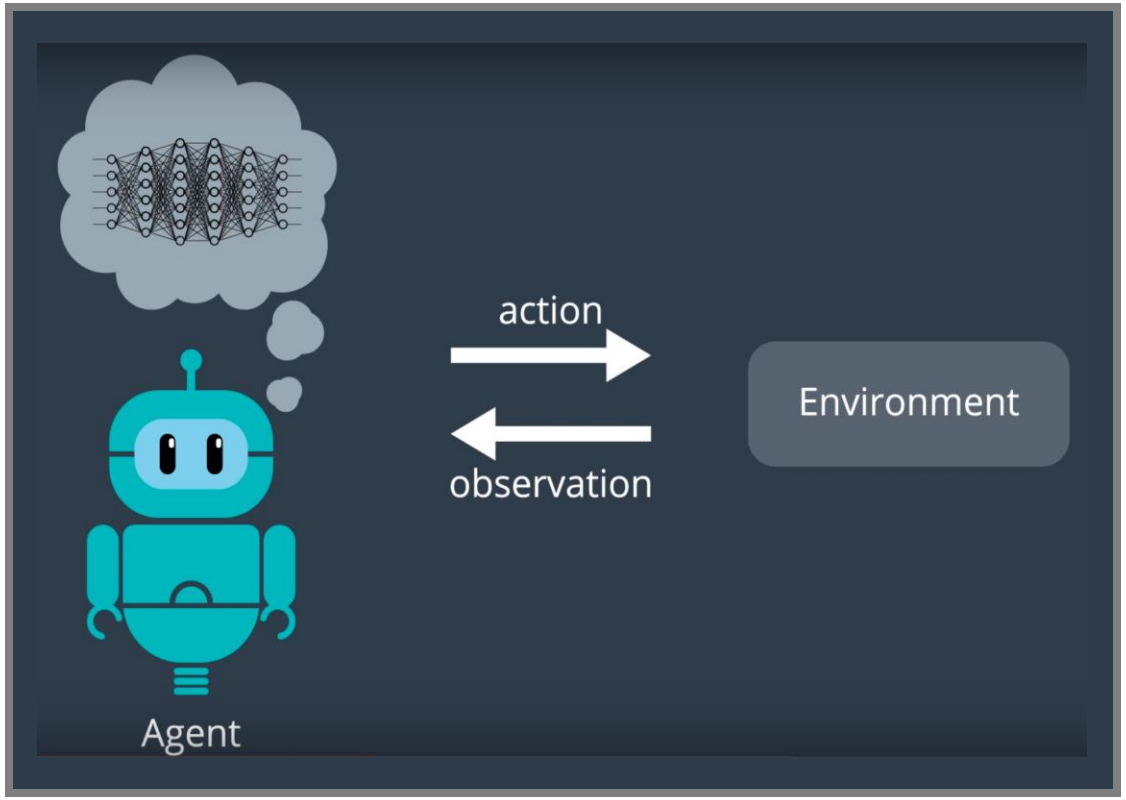
Characteristics and Formulation

Goal

Sequential Decision Making: select actions to maximize total future reward

- Examples
  - A financial investment (may take months to mature)
  - Re-fueling a helicopter (might prevent a crash in several hours)
  - Blocking opponent moves (might increase chances of winning after a certain number of moves)

Terminology: Agent and Environment



- At each step the agent:
  - Receives observation
  - Receives scalar reward
  - Executes action
- The environment:
  - Receives action
  - Emits observation
  - Emits scalar reward
- increments at environment step

Terminology: History and State

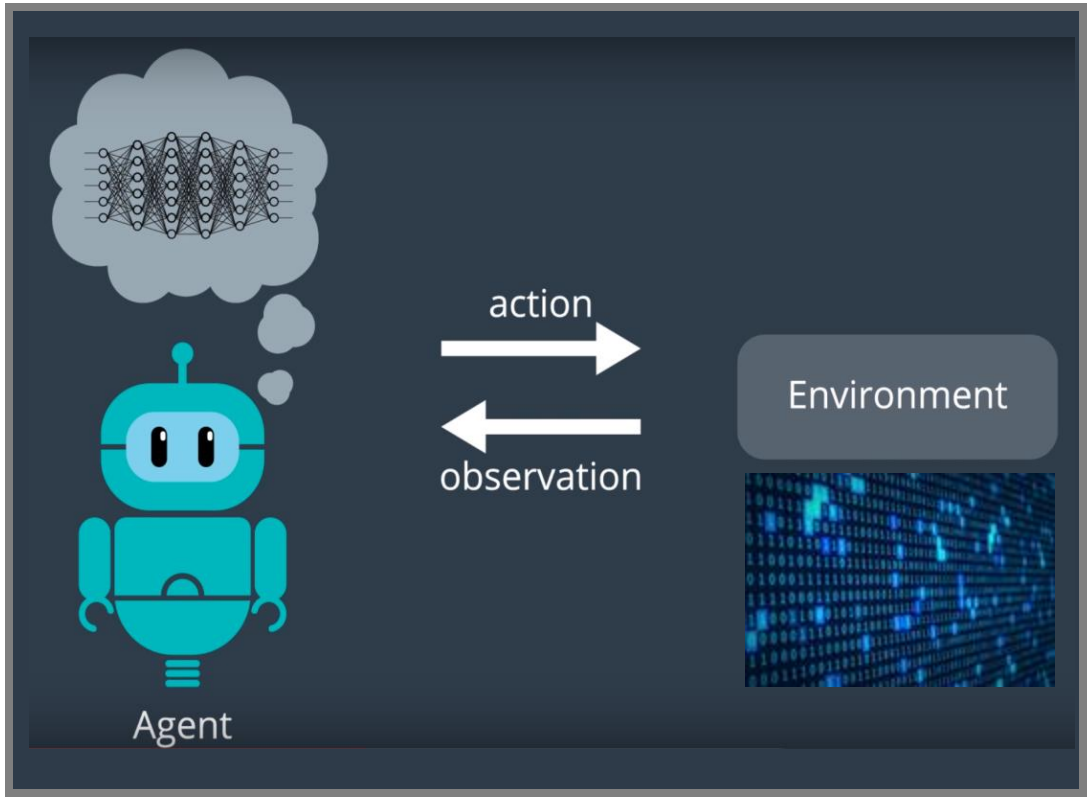
- The history is the sequence of observable variables (observations, actions, rewards) up to the current time step

$$H_t = O_1, R_1, A_1, \ldots, A_{t-1}, O_t, R_t$$

- State is a function of history and contains information related to what happens to the agent in the next time step

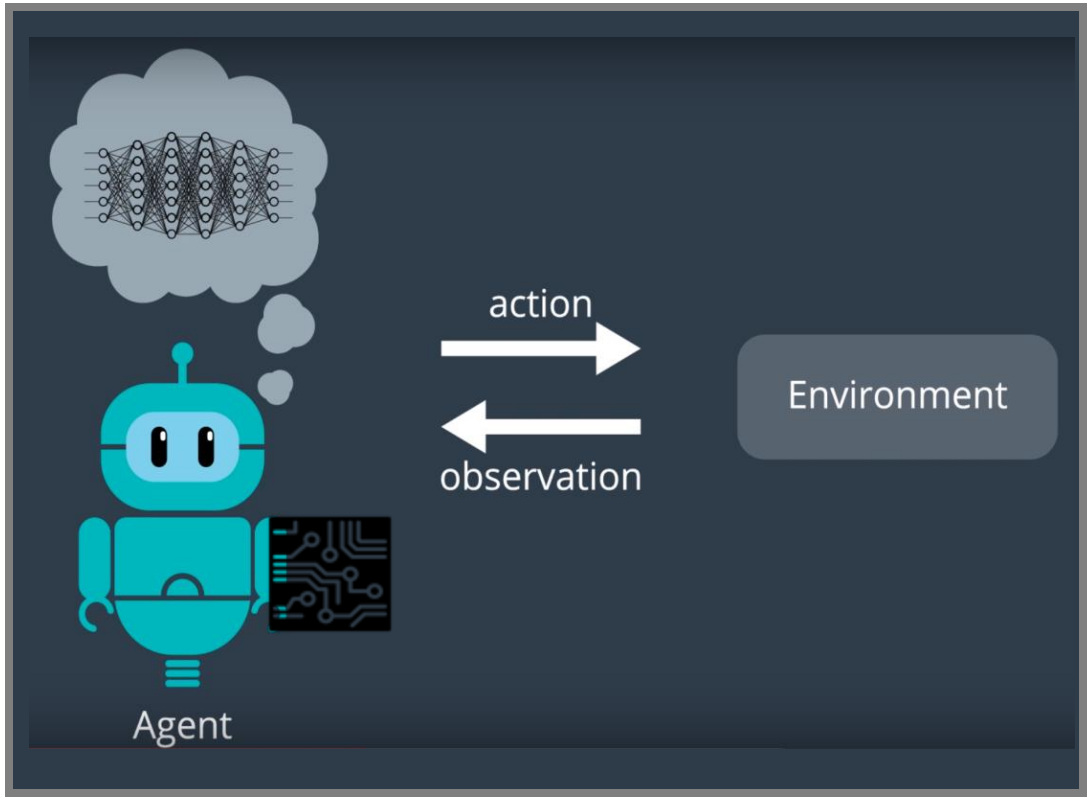
$$S_t = f(H_t)$$

Terminology: Environment State (S_t^e)



- The environment state S_t^e is the environment's private representation
- S_t^e is not usually visible to the agent
- Even if S_t^e is visible to the agent, it may contain irrelevant information

Terminology: Agent State (S_t^a)



- S_t^a is the agent's internal representation
- Information the agent uses to pick the next action
- It is the information used by reinforcement learning algorithms

Terminology: Rewards

- A reward is a scalar feedback signal
- Indicates how well agent is doing at each step
- The agent's job is to maximize cumulative reward over time

Terminology: Examples of Rewards

- Defeat the world champion at Backgammon /Chess/Go
  - reward for winning/losing a game
- Play many different Atari games
  - reward for increasing/decreasing score
- Manage an investment portfolio
  - reward for each $ in bank

Terminology: Examples of Rewards

- Control a power station
  - reward for producing power (+ve)
  - reward for exceeding safety thresholds (-ve)

- Make a humanoid robot walk
  - reward for forward motion (+ve)
  - reward for falling over (-ve)

Terminology: Actions

- Actions are a set of operations that an agent can perform which allows it to interact and change its environment.
- This allows the agent to transfer between different states.
- Every action taken by the agent yields a reward from the environment.
- The decision of which action to choose is decided based on the policy.

Markov State

- Each Markov state contains all useful information from the history
- A state is Markov if and only if

$$\mathbb{P}[S_{t+1} \mid S_t] = \mathbb{P}[S_{t+1} \mid S_1, \dots, S_t]$$

- The future (S_{t+1}) is independent of the past given the present (S_t).
- Once the current state is known, the history may be thrown away.
- The state is a sufficient to develop a prediction for the future.

Markov State: Rat Example

- What if agent state = last 3 items in sequence?
- What if agent state = counts for lights, bells and levers?
- What if agent state = complete sequence?

출처 https://www.gettyimagesbank.com/s/?lv=si&st=union&mi=2&q=1388704407%2C+1413726124%2C+1068565928%2C+1298716789%2C+1455391555&ssi=go

Fully Observable Environments

- Full observability: agent directly observes environment state

$$O_t = S_t^a = S_t^e$$

- Environment state == Agent state == Markov state
- Formally, this is a Markov Decision Process (MDP)

Partially Observable Environments

- Partial observability: agent indirectly observes environment:
  - A robot with camera vision isn't told its absolute location
  - A trading agent only observes current prices
  - A poker playing agent only observes public cards
- Now agent state is NOT equal to environment state
- Formally this is a Partially Observable MDP (POMDP)
- Agent must construct its own state representation