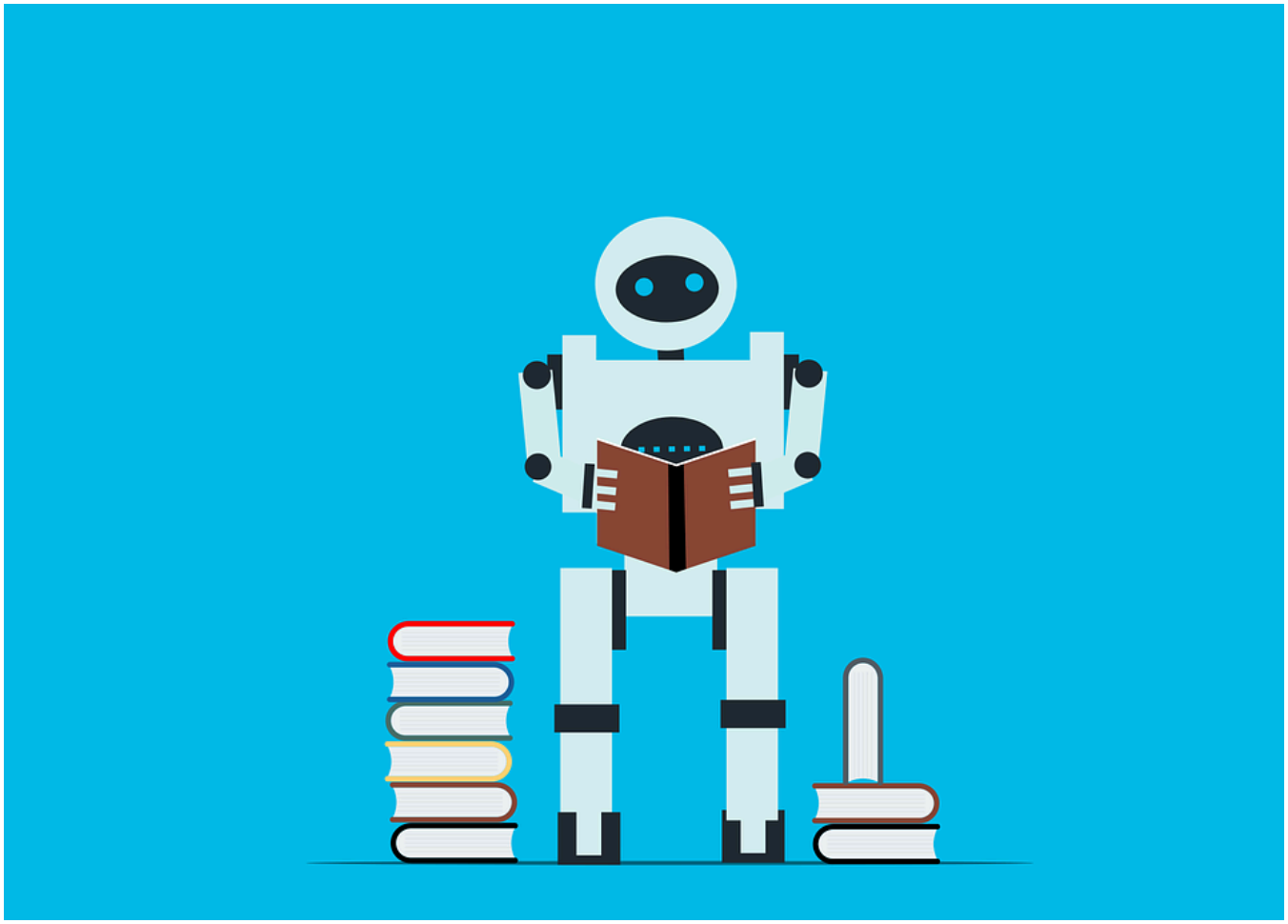


Introduction

1. 이 강의를 위한 사전지식

무작정 따라하는 강화학습 프로젝트

> - 만약, 코드 예제만 따라하고 싶다면, 아래 지식이 없어도 상관없다.

- 파이썬
- 뉴럴 네트워크
- 통계학

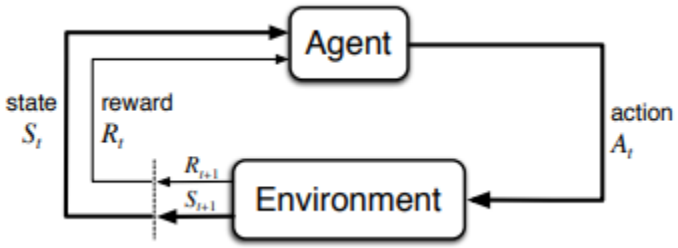
- 미적분학

2. 강화학습(Reinforcement Learning) 이란?

강화학습은 다음과 같이 정의된 환경(**Environment**) 과 에이전트(**Agent**) 의 상호작용이다.

- **환경** : 에이전트와 상호작용하는 에이전트 밖의 모든 것 이다.
    - **보상** : 강화학습 문제의 목표를 정의합니다. 그리고 에이전트의 행동으로 인해 환경으로부터 받는 (좋은)신호이다.
    - **상태** : 정책과 가치함수의 입력값으로써 환경이 보내주는 신호이다.
- **에이전트** : 학습자이면서 정책 결정자이다.
    - **정책** : 에이전트의 행동을 결정하는 함수이다.
    - **가치함수** : 장기적인 관점으로 어떤 행동이 좋은지를 결정한다.

> 사람을 생각해보면 사람의 감각은 몸에서 느껴지므로 에이전트라고 생각할 수 있으나 이는 환경에서 주는 상태이므로 이는 환경에 포함된다. (이 또한 강화학습 문제를 어떻게 정의하느냐에 따라서 달라질 수 있다.)



종류에 따라서 다르지만 일반적인 강화학습의 목적은 장기적인 관점에서 얻을 수 있는 보상을 최대화 하는 정책함수 찾기이다.

3. Deep Q-learning이란?

알파고와 같은 인공지능으로 분야를 선두하는 **구글의 딥마인드**(**deep mind**)가 2013년에 발표한 논문 *Playing Atari with Deep Reinforcement Learning*에서 뉴럴 네트워크와 강화학습을 결합한 **Deep Q-learning** 알고리즘이 소개되었다.

하지만 강화학습을 딥러닝과 결합하는 과정에서 다음과 같은 문제들이 발생한다.

- 지도학습에서의 데이터와 달리, 강화학습은 에이전트가 *행동*(*Action*)을 하고 *보상*(*Reward*)를 받아야 하므로 모델의 가중치 업데이트와 샘플이 만들어지는 사이에 시간차가 발생한다. (논문에서는 강화학습의 데이터가 sparse, noisy and delayed 하다고 표현했다.)
- 딥러닝 지도학습은 샘플 사이의 독립성을 가정하는데, 강화학습은 *에이전트*(*Agent*)의 후 *상태*(*State*)가 전 상태에 영향을 받으므로 독립성을 보장할 수 없다.
- 딥러닝은 고정된 데이터 분포에서 학습을 진행하지만, 강화학습은 학습이 진행됨에 따라 데이터의 분포가 변화한다.

위와 같은 문제를 완화하기 위해서, ***Deep Q-learning***에서는 다음과 같은 방법을 사용한다.

1. Experience Replay Mechanism

Episode[1]를 진행하면서 얻는 정보들을 **Replay Memory**[2]로 저장한다. 그리고 이 *Replay Memory*에 있는 데이터로 학습을 진행함으로, 데이터 분포의 일관성이 유지된다.

2. 데이터를 무작위로 샘플링한다.

에이전트의 진행 방향에 따라 학습을 진행할 경우 데이터의 독립성이 보장되지 않는다. 이러한 문제를 보완하기 위해 *Replay Memory*에서 데이터를 무작위로 추출해 데이터의 독립성을 보완한다.

이러한 방법을 포함해, 뉴럴 네트워크 그리고 **Stochastic Gradient Descent** 방법을 사용하여 ***Policy***를 학습시키는 알고리즘을 ***Deep Q-learning***이라 한다.

다음 챕터에서는 이러한 *Deep Q-learning*을 이해하기 위해, 바탕이 되는 *강화학습*의 기본 개념을 코드와 함께 공부하고자 한다.

---

1) Episode : 체스에서 킹이 죽으면 게임이 끝나듯이, 에이전트(체스 플레이어)와 환경(체스판)의 상호작용이 자연스럽게 끝나면, 그 전까지 얻어지는 상태, 보상 그리고 행동의 정보들을 Episode라 한다.

2) Replay Momory : $(State_t, Action_t, Reward_t, State_{t+1})$ 과 같은 튜플의 배열