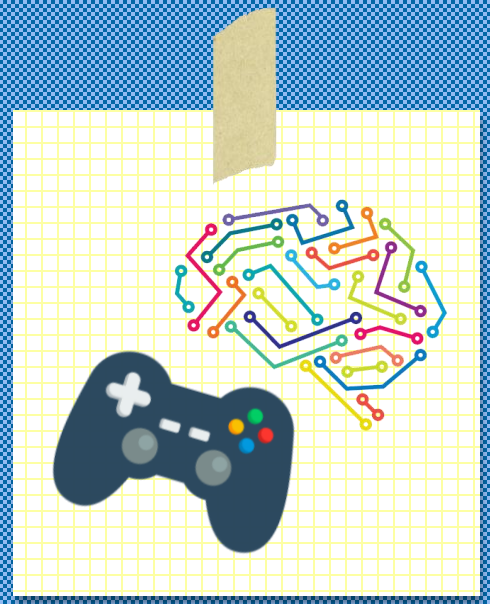
강화학습을 적용한 미로 탈출 프로젝트

Monte Carlo, Q-Learning, Deep SARSA



김웅재, 김현조, 김현정, 이규서

01

프로젝트 소개

1) 주제 선정 계기
2) 시행착오 과정

02

강화학습 개요

1) 강화학습이란?
2) MDP
3) 알고리즘

03

알고리즘 구현

1) 몬테카를로
2) 큐러닝
3) 딥살사

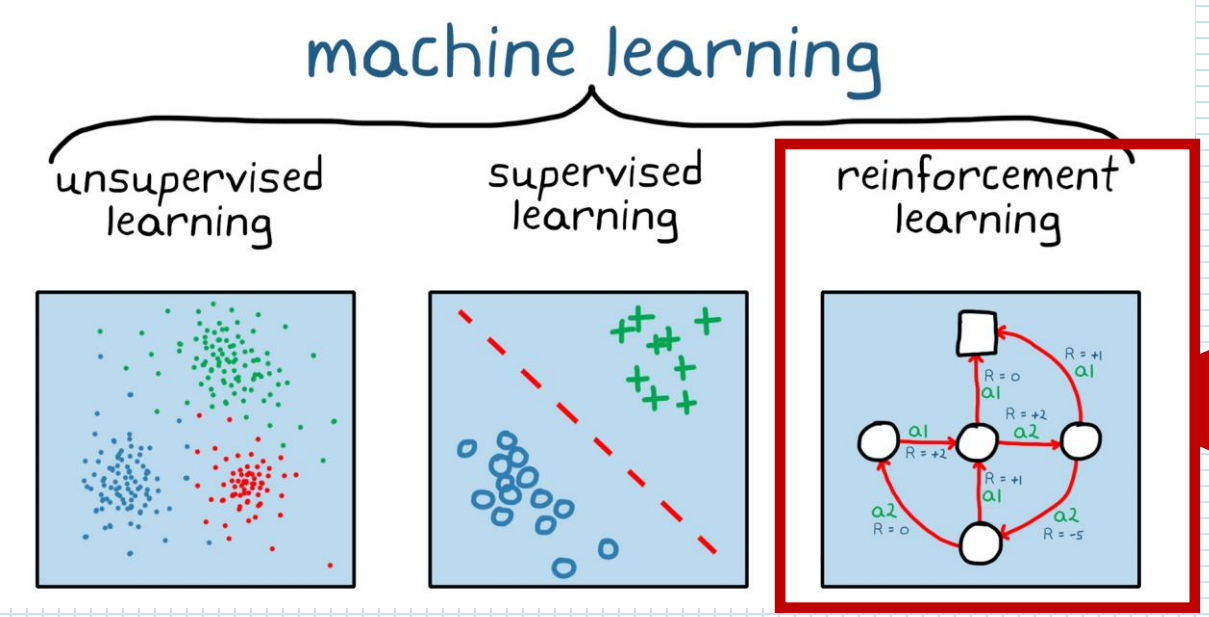
04

REVIEW

1) 느낀 점
2) 질의응답

1. 프로젝트 소개

1) 주제 선정 계기

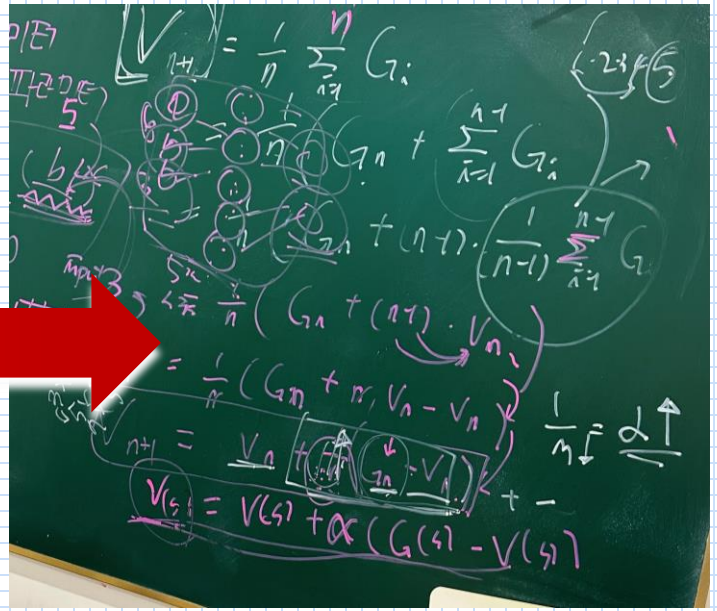


- ✓ 이번 교육 과정에서 자세히 다루지 않았던 기계학습 중 하나인 '강화학습 ' 에 대한 궁금증에서 출발
- ✓ 사용자 없이 경험을 통해 스스로 최적의 play를 한다는 것에 대한 흥미로움

1. 프로젝트 소개

2) 시행착오 과정

[직접 구현한 총알 피하기 게임]



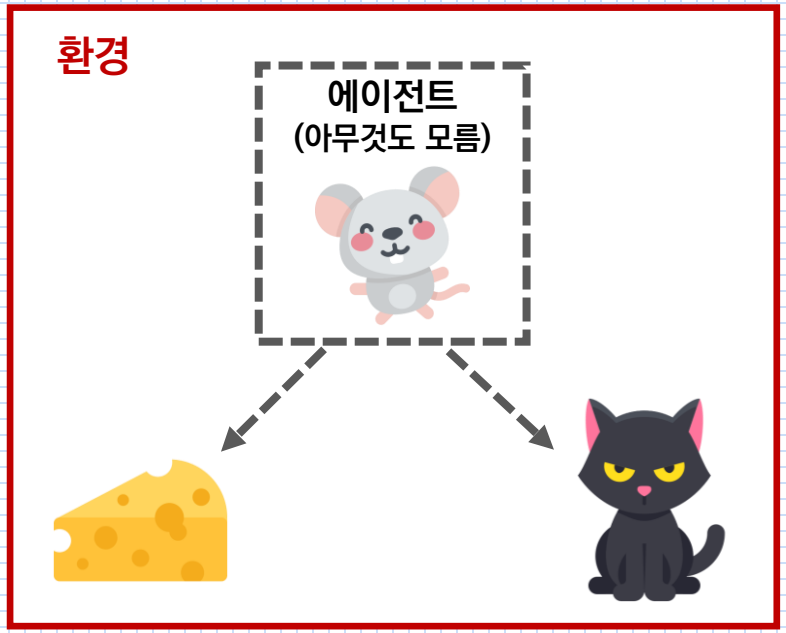
✓ 처음에는 단순히 강화학습을 적용하고 싶은 게임부터 만들기 시작

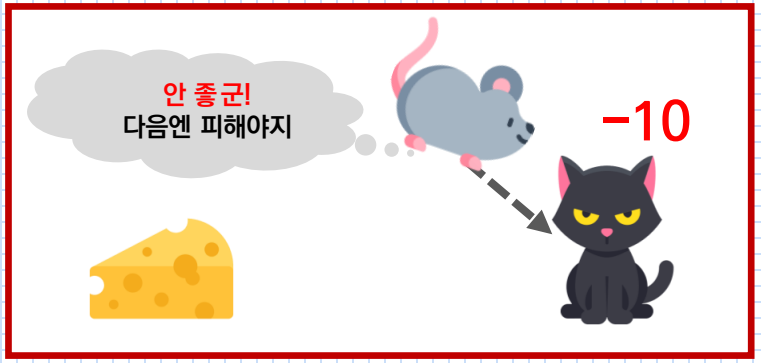
=〉 **강화학습에 대한 지식이 부족하다 보니 진행에 있어 한계를 느낌**

✓ 강화학습에 대한 기초 개념부터 다시 다지며 알고리즘을 통해 게임이 개선되어가는 과정을 보여주는 것으로 결정

2. 강화학습 개요

1) 강화학습이란 ?





- ✓ 시행착오를 통해 학습하는 방법 중 하나로, 실수와 보상을 통해 학습하여 목표를 찾아가는 알고리즘
- ✓ 라벨(정답)이 있는 데이터를 통해 가중치와 편향을 학습하듯이 강화학습은 보상(Reward)이라는 개념을 사용하여 가중치와 편향을 학습함

2. 강화학습 개요

2) MDP (Markov Decision Process)

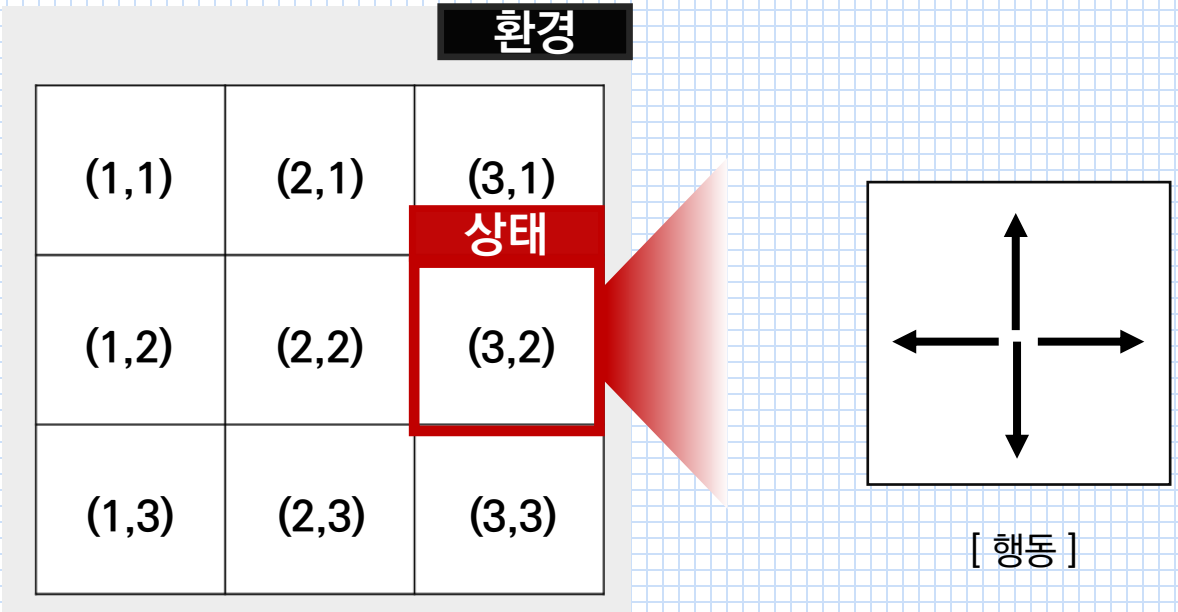
MDP의 구성요소

- 상태
- 행동
- 보상함수
- 상태 변환 확률

- ✓ 강화학습의 목적 : 에이전트가 환경을 탐색하면서 얻은 **보상의 합을 최대화 할 수 있는 최적의 행동 루트**를 찾는 것
- ✓ MDP : 순차적으로 행동을 결정해야 하는 문제를 다루는 강화학습을 컴퓨터가 알 수 있도록 수학적으로 표현한 방법

2. 강화학습 개요

2) MDP (Markov Decision Process) – 상태, 행동



[행동]

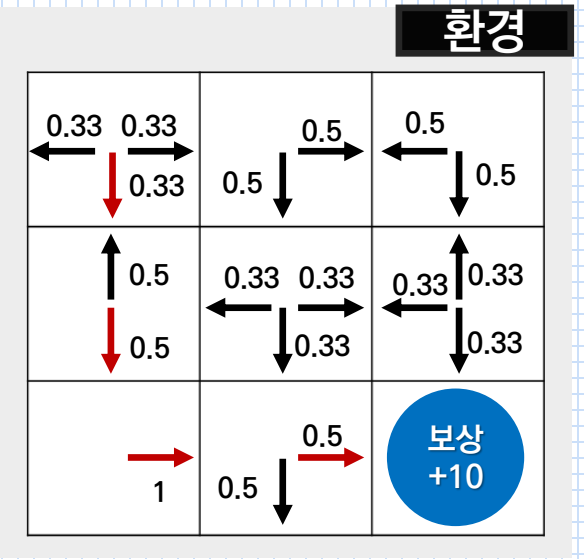
✓ 상태 : 자신의 상황에 대한 관찰로, 현재 에이전트의 위치를 파악할 수 있음

✓ 행동 : 다음 상태로 이동하기 위해 현재 상태에서 이동 가능한 방향

2. 강화학습 개요

2) MDP (Markov Decision Process) – 보상, 상태 변환 확률

[보상]



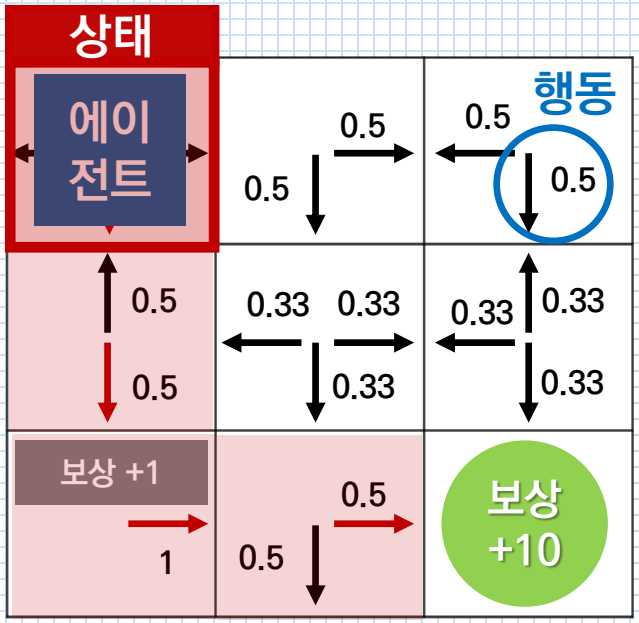
[상태 변환 확률]

- ✓ 보상 : 어떠한 상태에서 어떠한 행동을 했을 때 받을 수 있는 값
- ✓ 상태 변환 확률 : 에이전트가 어떠한 상태에서 행동을 통해 다른 상태로 가게 될 확률
- ✓ 보상과 상태 변환 확률은 환경의 모델로서 이에 따라 알고리즘이 개선됨

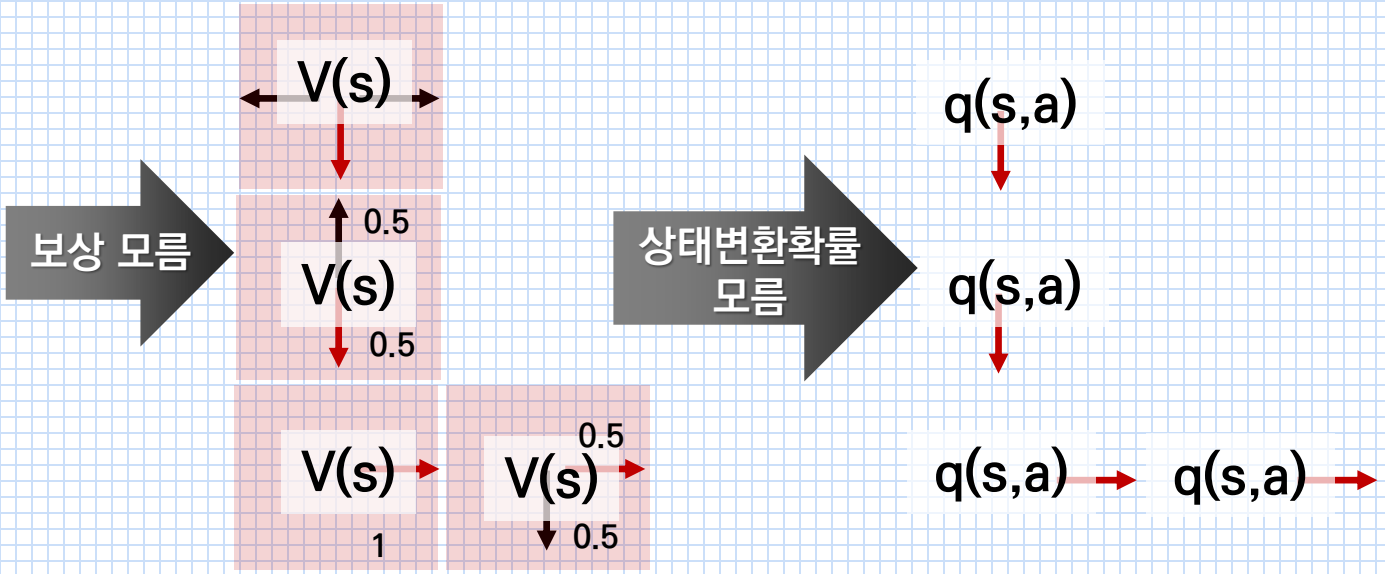
2. 강화학습 개요

3) 알고리즘

=〉 보상과 상태 변환 확률에 따라 가치함수 V(s)와 큐함수 q(s,a)을 통해 업데이트



[고전 알고리즘 – 보상, 상태변환확률 모두 아는 상태]

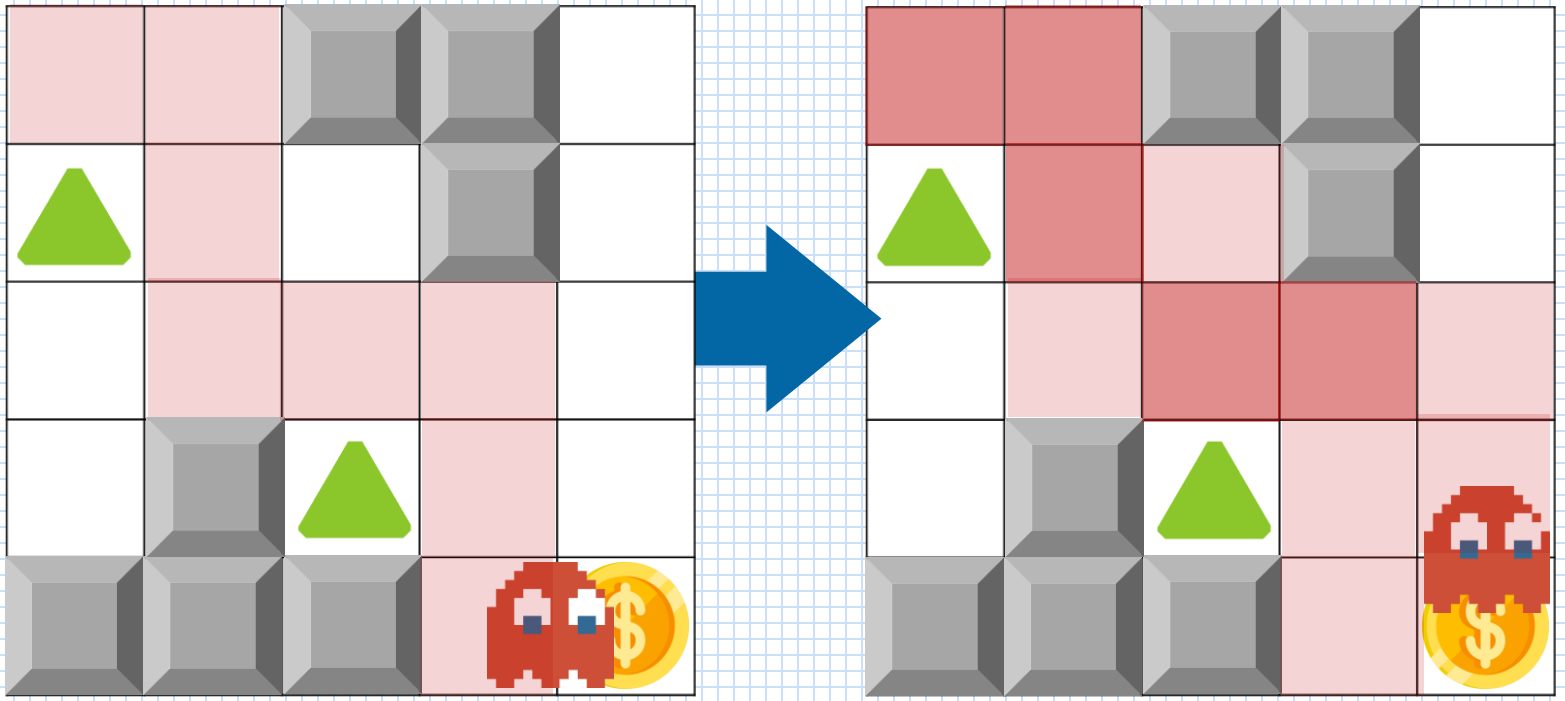


✓ 가치함수 : 각각의 칸에 대한 가치를 업데이트하며 최적의 루트를 찾음

✓ 큐함수 : 어떠한 행동에 대한 가치를 업데이트하며 최적의 루트를 찾음

2. 강화학습 개요

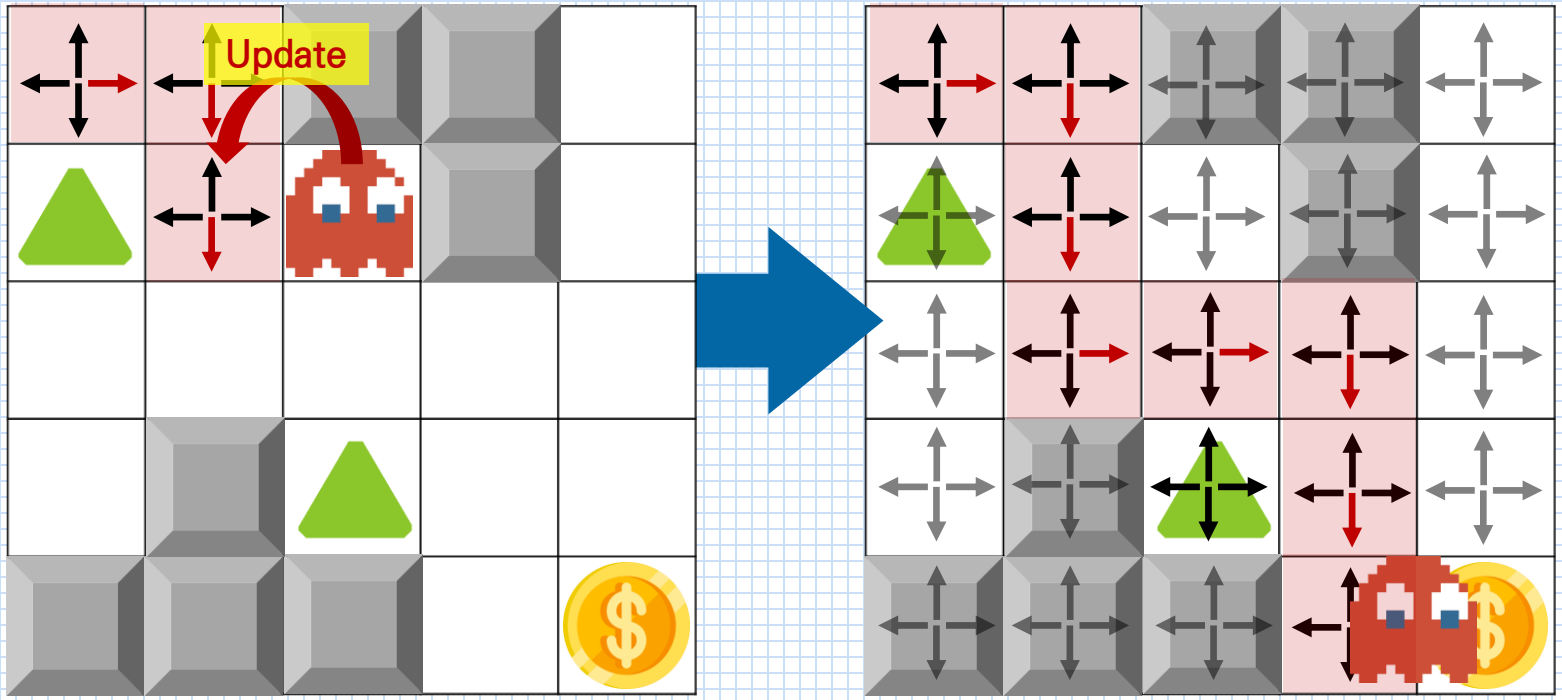
3) 알고리즘 – 몬테카를로 (가치함수 업데이트)



- ✓ 에피소드를 1번 완료한 후 지나온 상태에 대하여 가치함수 업데이트
- ✓ 에피소드가 끝나야 업데이트 적용이 가능하기 때문에 시간이 오래 걸림

2. 강화학습 개요

3) 알고리즘 – 큐러닝 (큐함수 업데이트)

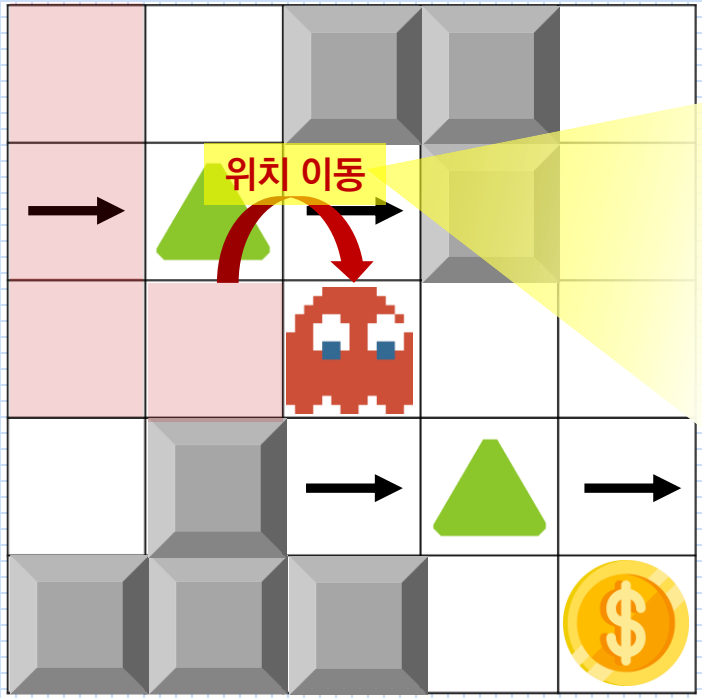


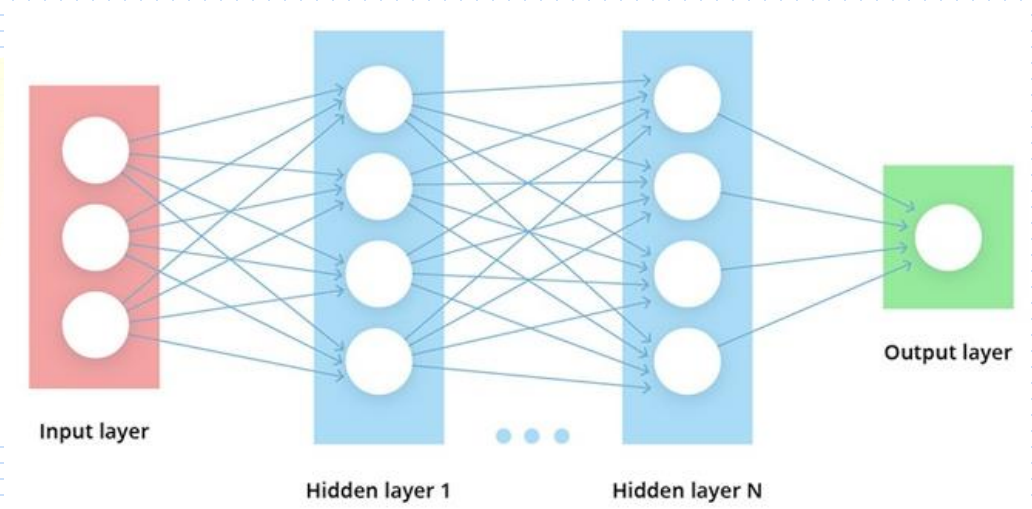
✓ 큐함수를 통해 실시간으로 업데이트 진행

✓ 모든 상태를 경험함으로써 업데이트되며 그 중에서 최적의 가치를 가지는 루트로 가게 됨

2. 강화학습 개요

3) 알고리즘 – 딥살사 (딥러닝을 통한 학습)





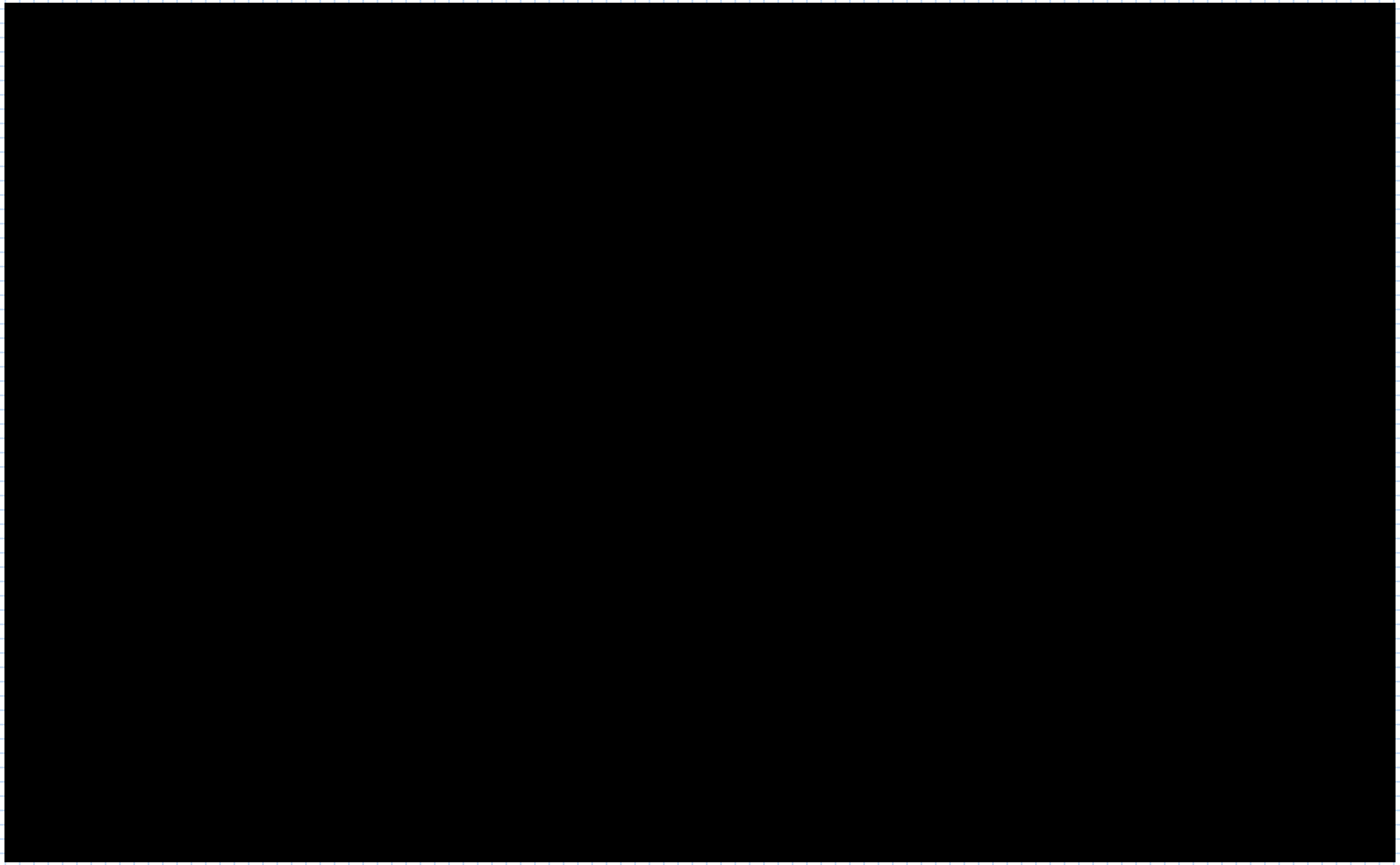
- ✓ 움직이는 장애물이 생기며 경우의 수가 더욱 많아짐 –〉 딥러닝을 통해 학습 진행
- ✓ 이동 1번 = 학습 1번이며, 이동할 때마다 가중치가 역전파를 통해 계속해서 업데이트됨

3. 알고리즘 구현

1) 몬테카를로 (Monte Carlo)

3. 알고리즘 구현

2) 큐러닝 (Q-Learning)



3. 알고리즘 구현

3) 딥살사 (Deep SARSA)

시연 진행

4. REVIEW

1) 느낀 점

✓ 지도학습, 비지도학습에 대한 개념과는 완전히 다르다 보니 강화학습에 대한 확실한 기초 개념 필요

✓ 데이터 분석을 하지 못한 것에 대한 아쉬움이 있음 (강화학습은 데이터가 없이 학습되기 때문)

✓ 고전 학문임에도 불구하고 찾아볼 수 있는 자료가 부족하여 어려움이 많았음

✓ 강화학습의 알고리즘이 어떻게 개선되어 왔는지를 완벽하게 이해할 수 있어서 좋았음

✓ 더 많은 알고리즘 공부를 통해 우리가 만들었던 총알 피하기 게임에도 꼭 적용해야겠다는 목표가 생김

4. REVIEW

질의응답

감사합니다.