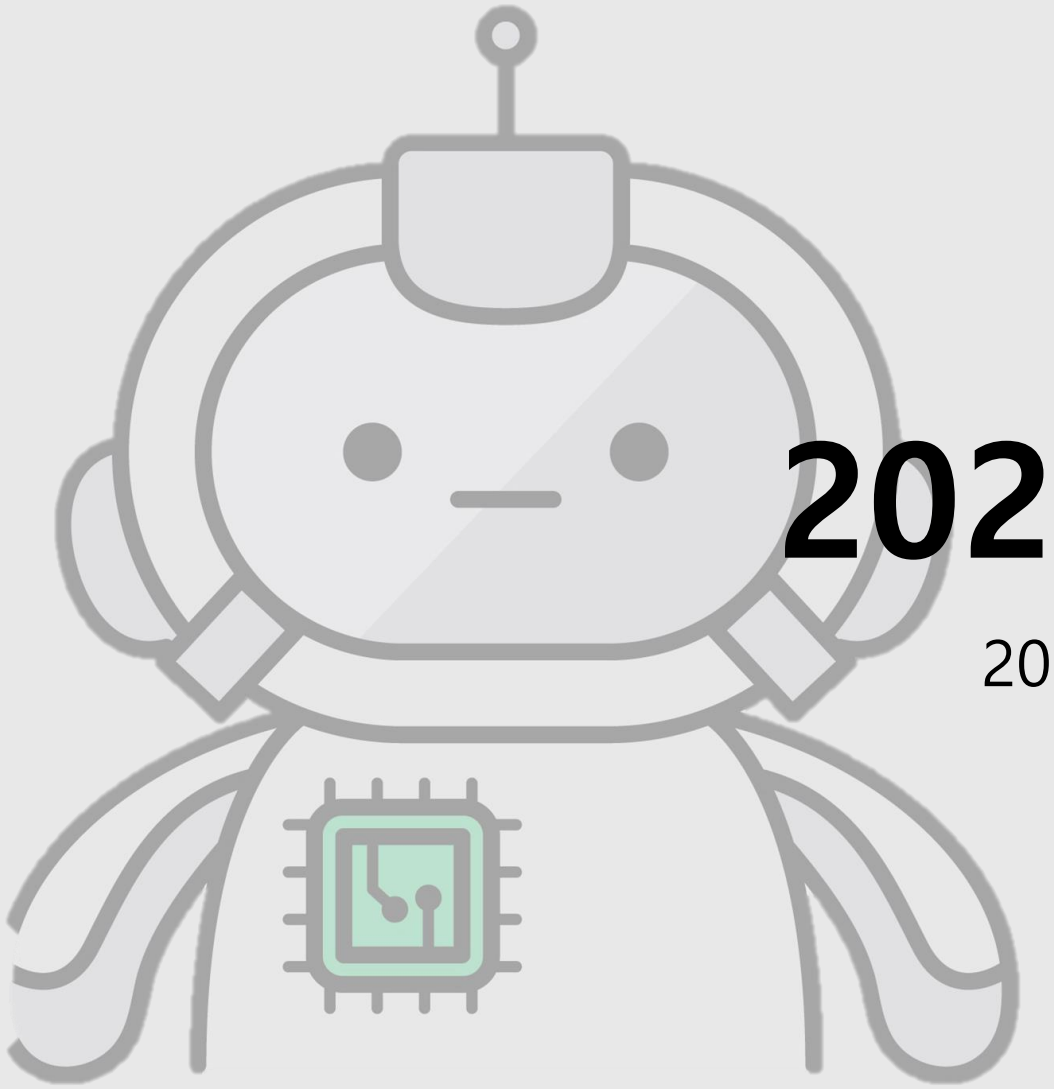
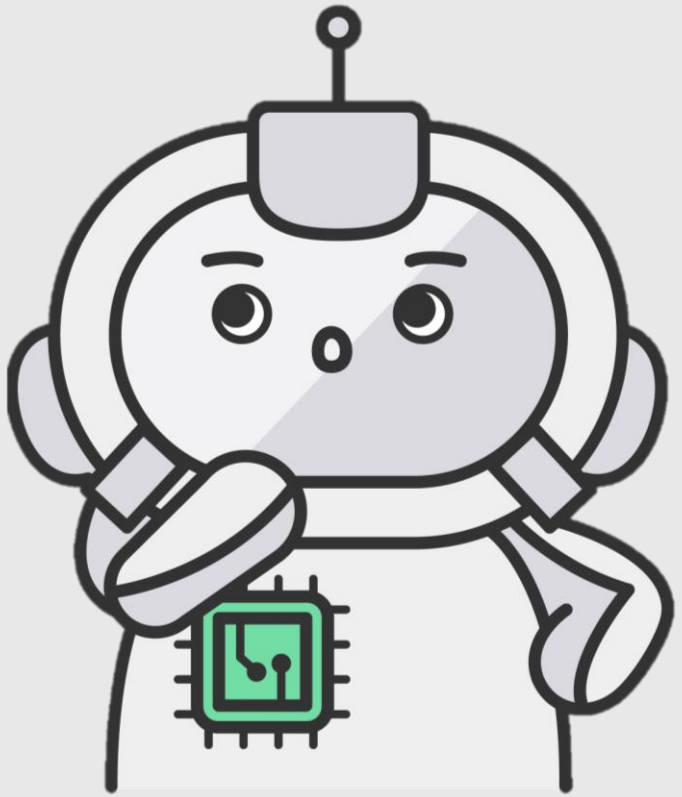


2025. 05. 16.

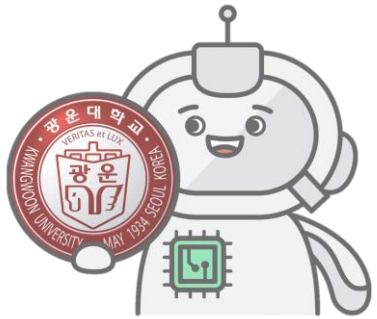
2025194110 박민규

자율 주행 드론 및 강화학습



1. 자율 주행 드론

- Unmanned Aerial Vehicle(UAV)
  - 무인 항공기, 조종사 없이 자동 혹은 원격으로 비행할 수 있는 항공 시스템
- Autonomous Driving
  - AI 기반 판단을 통해 스스로 주행
- Autonomous Drone Driving
  - UAV가 직접 주변 환경을 파악하고 판단하여 목적지 까지 이동
  - Amazon Prime Air의 드론 배송, 구조용 드론, 농업용 자율 비행기 등



1. 자율 주행 드론

그림 1. 자율 주행 드론 예시

1. 자율 주행 드론

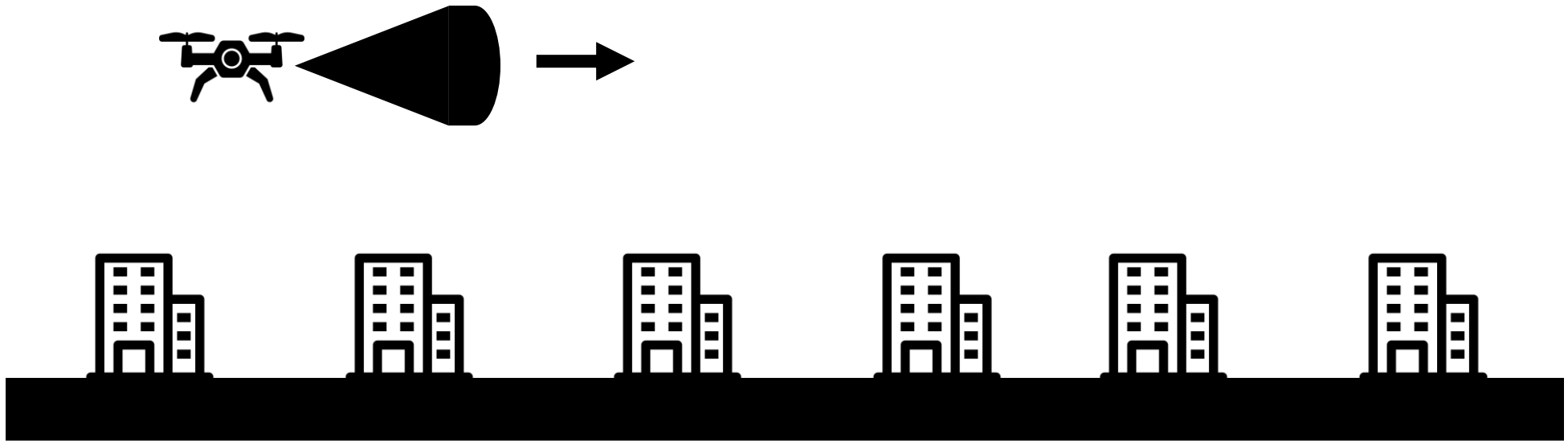


그림 1. 자율 주행 드론 예시

1. 자율 주행 드론

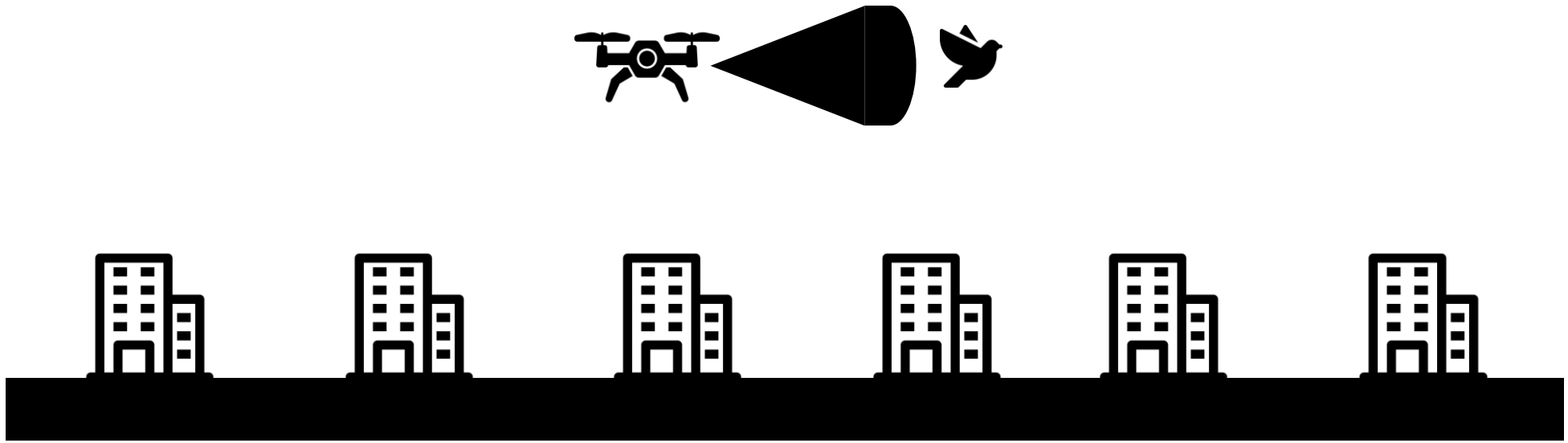


그림 1. 자율 주행 드론 예시

1. 자율 주행 드론

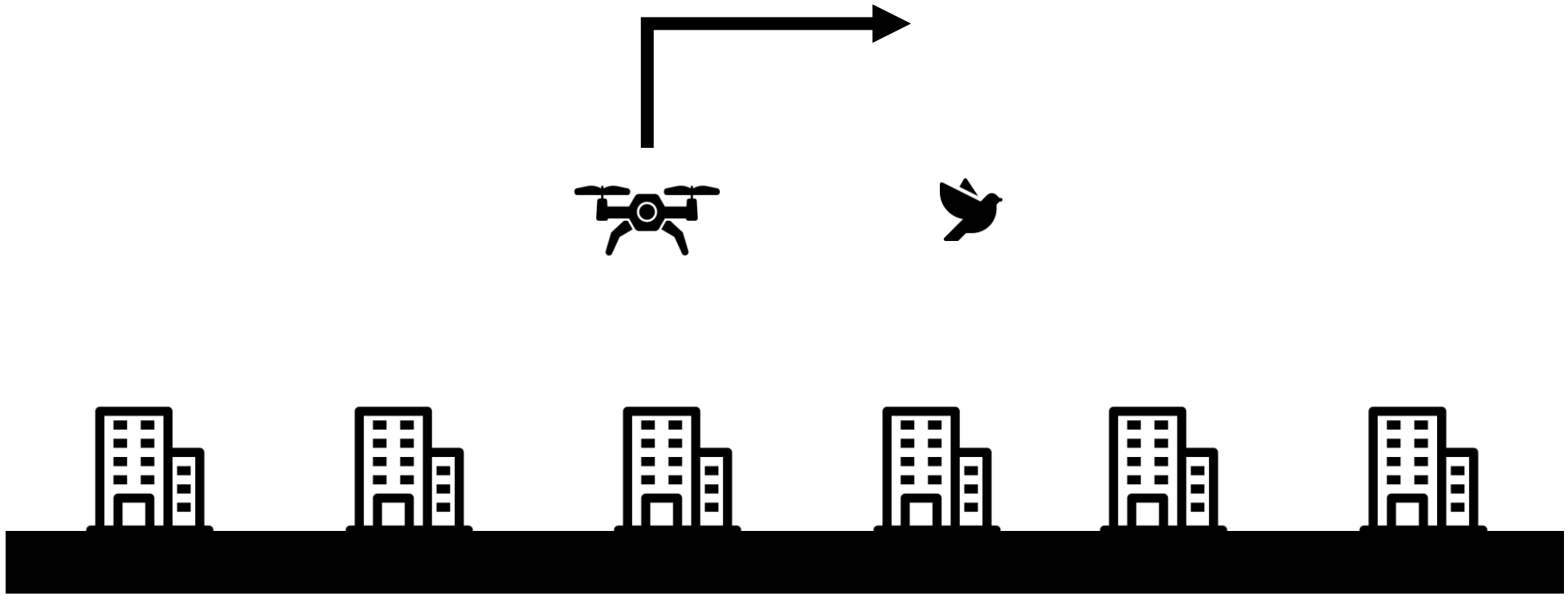


그림 1. 자율 주행 드론 예시

1. 자율 주행 드론

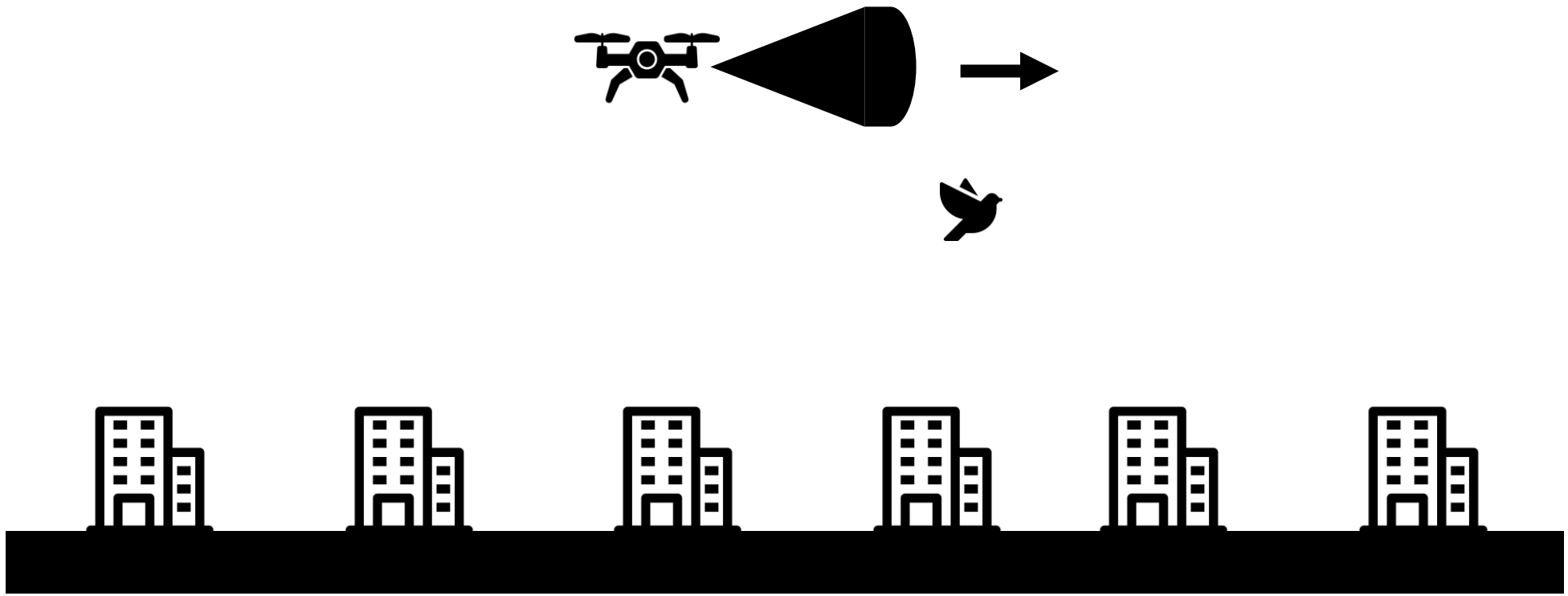


그림 1. 자율 주행 드론 예시

1. 자율 주행 드론

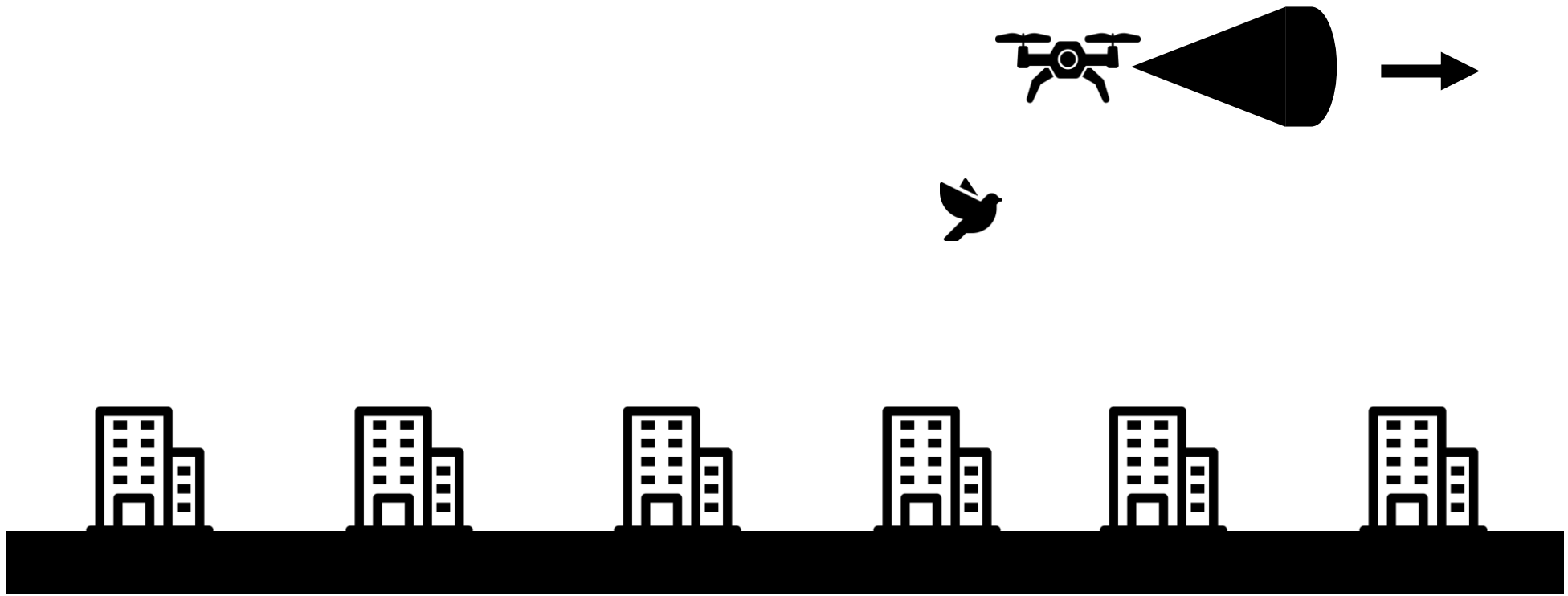
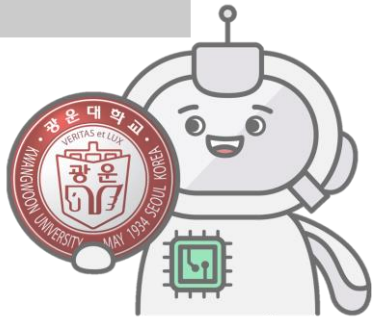


그림 1. 자율 주행 드론 예시

1. 자율 주행 드론

표 1. 자율 주행 UAV의 구성 단계 및 적용 기술

동작	의의	활용 기술
탐지	경로 상 장애물을 인식해 충돌 방지	컴퓨터 비전(YOLO. Faster R-CNN 등), LiDAR, SLAM 등
경로 탐색	장애물 회피를 위한 최적 경로 재계산	경로 계획 알고리즘(A*, D* 등), 강화학습 등
제어	경로에 따라 드론을 이동	PID 제어, 딥러닝 기반 제어 등



1. 자율 주행 드론

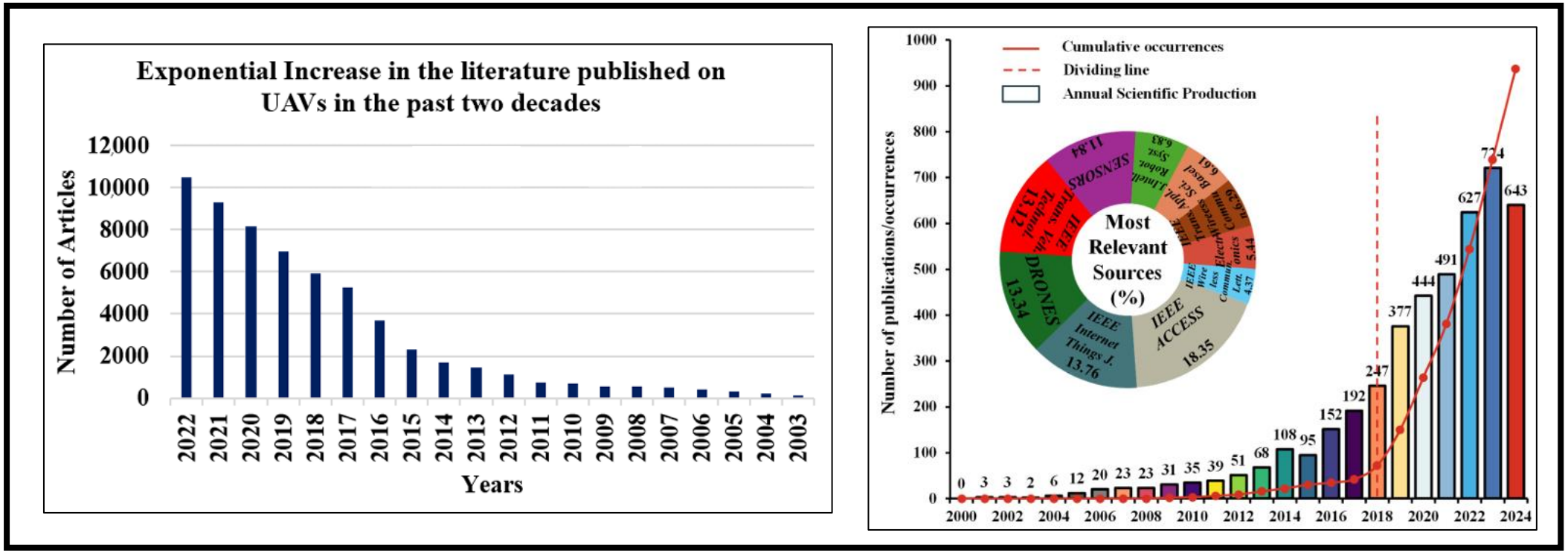


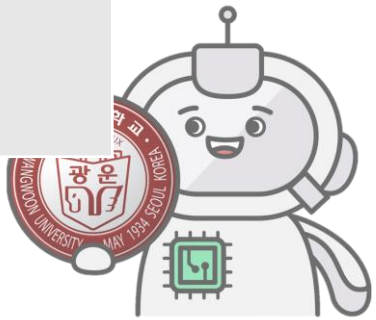
그림 2. 최근 UAV 경로 탐색 관련 연구의 출판 증가 추이 및 주요 학술 저널 분포[1][2]

[1] A Systematic Literature Review (SLR) on Autonomous Path Planning of Unmanned Aerial Vehicles (Husnain et al. in 2023)
[2] UAV Path Planning Trends from 2000 to 2024: A Bibliometric Analysis and Visualization (Wu et al. in 2025)

1. 자율 주행 드론

표 2. 경로 탐색과 관련 활용 기술

기술 분류	설명
A*	휴리스틱 기반의 대표적인 경로 탐색 알고리즘 장애물 회피와 효율적인 경로 계산 가능 정적 환경에 최적화되어 있기 때문에 동적 장애물 대응이 어려움
Model Predictive Control (MPC)	일정 시간 동안의 예측을 통해 최적 경로와 제어 명령을 동시에 계산 동적 환경 대응에 유리 연산량이 많아 실시간 계산에 부담이 크며 고속 비행시 한계가 존재
Grid-based Search	격자 기반의 탐색 방식으로, 지도 기반 환경에서는 단순 해상도가 낮으면 정밀도가 낮아지고, 높으면 계산 복잡도가 증가
강화학습 (Reinforcement Learning)	환경과 상호작용하며 보상을 통해 스스로 경로를 학습. 비정형 환경에 강함. 학습시간이 길고, 일반화가 어려우며 안정성이 낮을 수 있음



2. 강화학습(Reinforcement Learning)

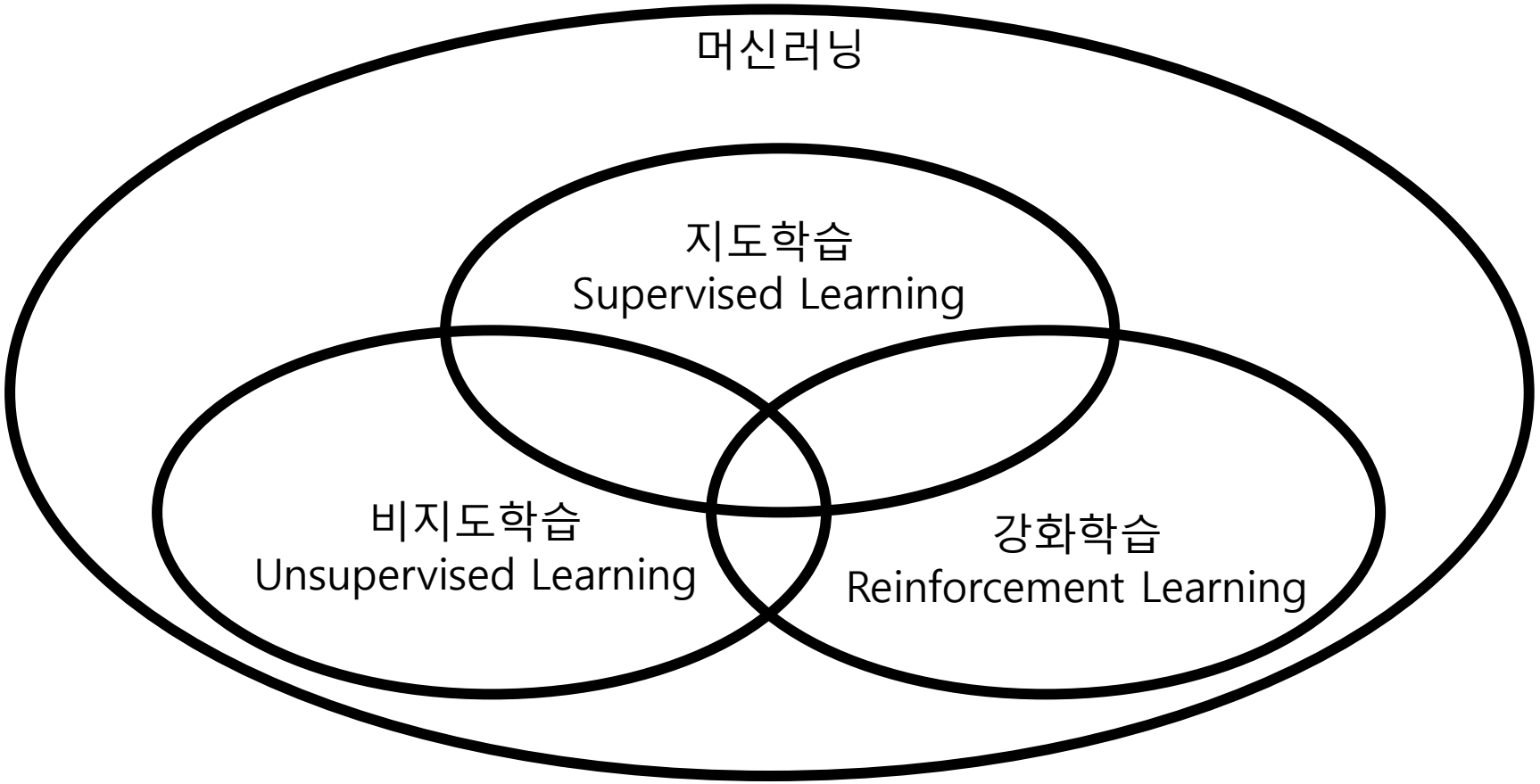
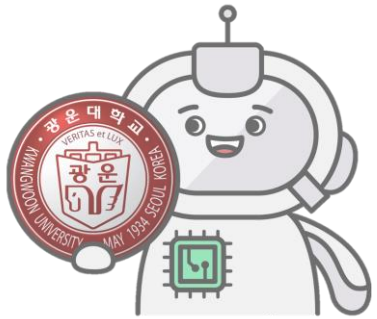


그림 4. 머신러닝의 대표적인 세 가지 학습 방식



2. 강화학습(Reinforcement Learning)

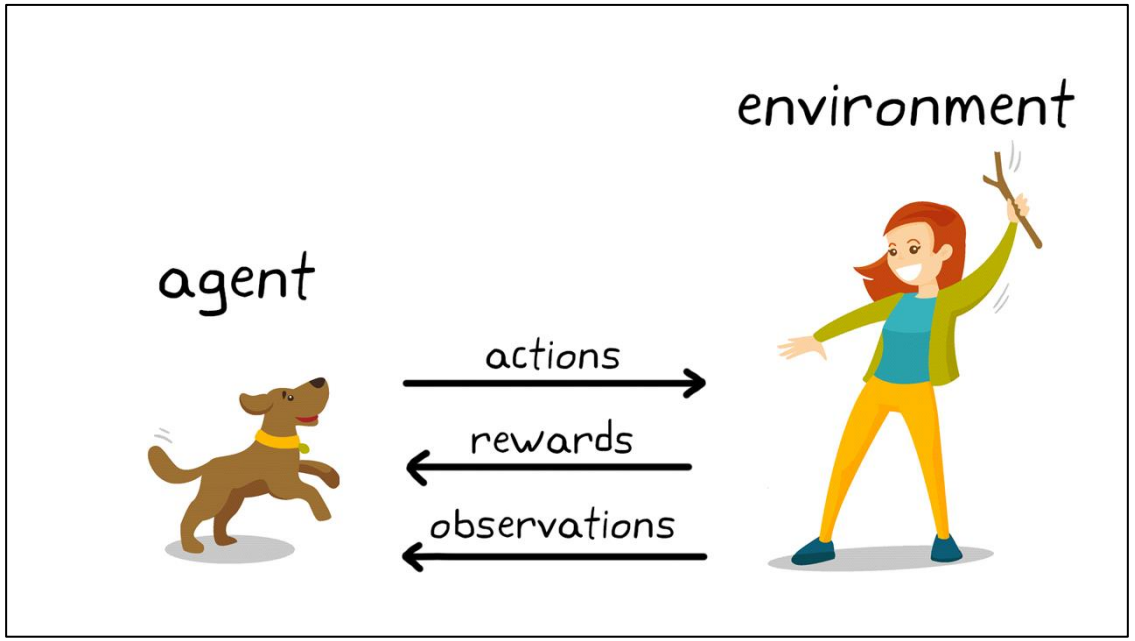


그림 5. 강화학습 예시

2. 강화학습(Reinforcement Learning)

- 상호작용을 통해 목표를 달성하는 방법을 학습하는 알고리즘

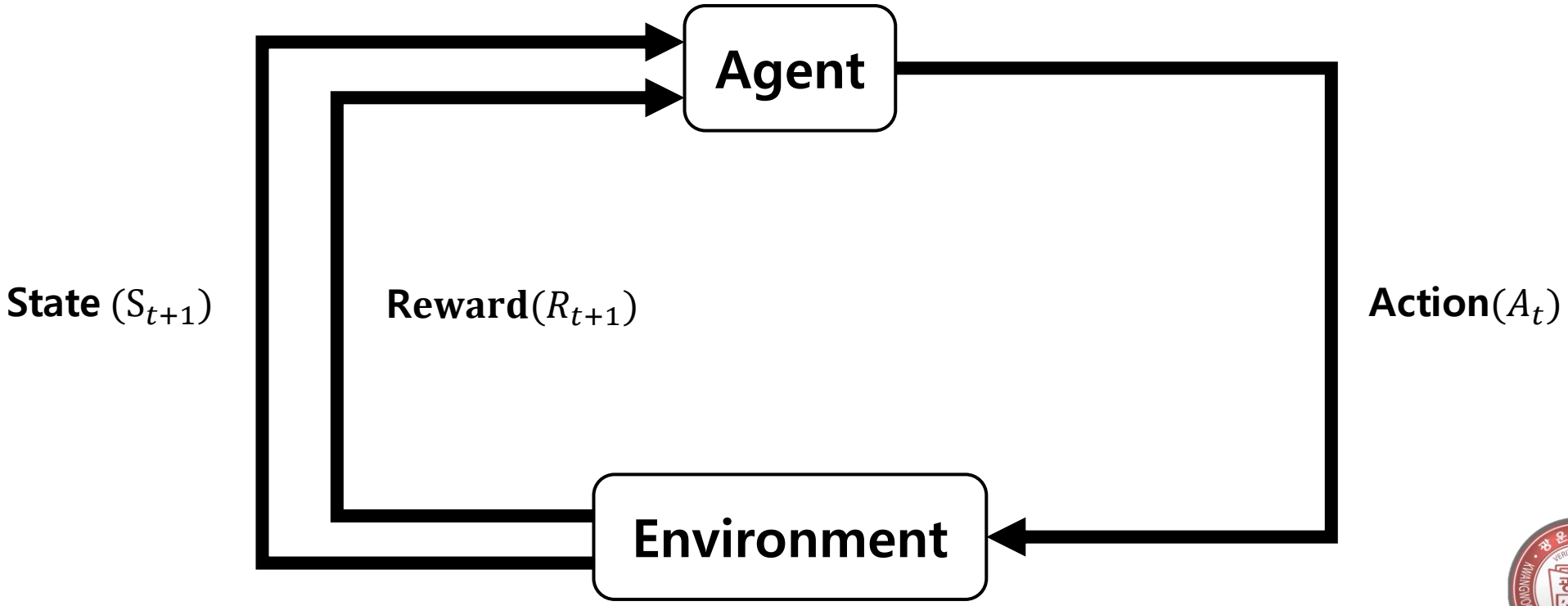
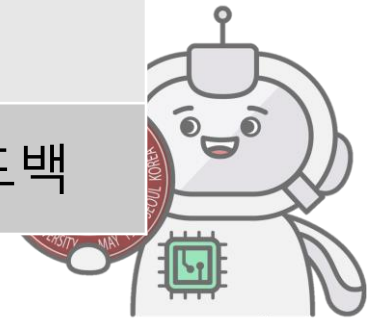


그림 6. 강화학습 순서도

2. 강화학습(Reinforcement Learning)

표 3. 강화학습의 구성 요소

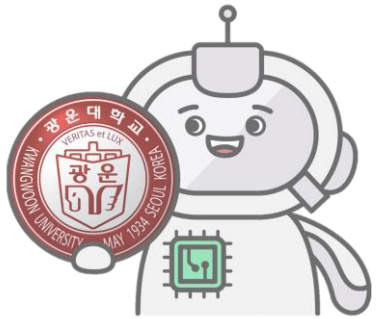
구성 요소	의미
에이전트(Agent)	학습하고 행동하는 주체
환경(Environment)	에이전트가 상호작용하며 행동의 결과를 받는 배경
상태(State)	현재 환경의 정보를 나타내는 관측 값
행동(Action)	에이전트가 상태에서 취할 수 있는 선택
보상(Reward)	에이전트의 행동에 따라 환경이 주는 수치적 피드백



2. 강화학습(Reinforcement Learning)

- 강화 학습 예제: 그리드 월드(Grid World)
  - S_0부터 S_7까지의 8개의 상태
  - 액션은 상하좌우 이동

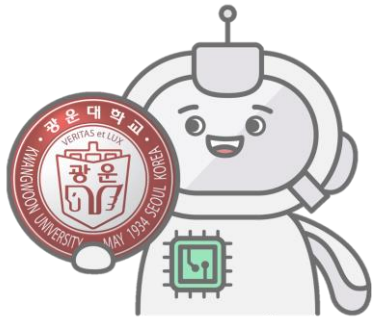
S_0	S_1	S_2	S_3
S_4	S_5	S_6	S_7



2. 강화학습(Reinforcement Learning)

- 강화 학습 예제: 그리드 월드(Grid World)
  - ➢ 골 지점의 보상 값을 1, 이외의 지점을 0

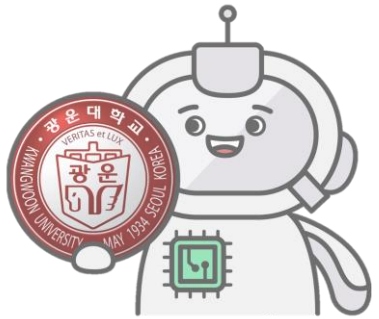
0	0	0	1
0	0	0	0



2. 강화학습(Reinforcement Learning)

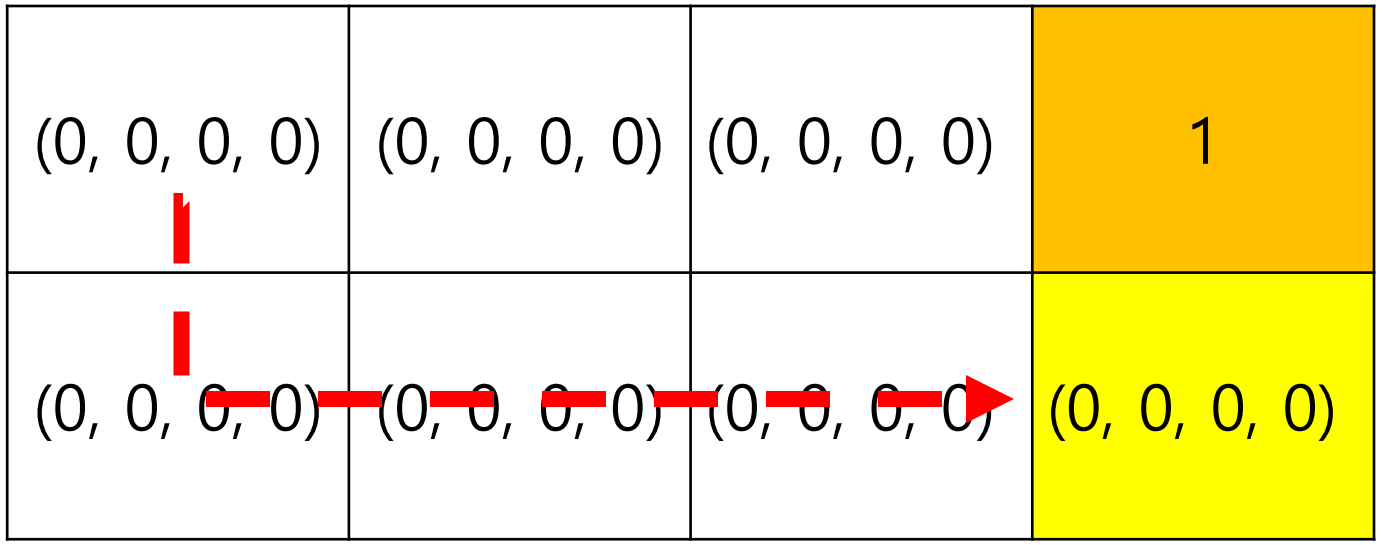
- 강화 학습 예제: 그리드 월드(Grid World)
  - 각 상태에서 상하좌우로 움직였을 때의 보상 값을 기록

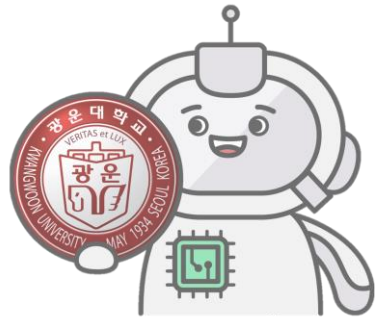
(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	1
(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)



2. 강화학습(Reinforcement Learning)

- 강화 학습 예제: 그리드 월드(Grid World)
  - ➢ 우연히 빨간 선과 같은 경로로 이동한다 가정

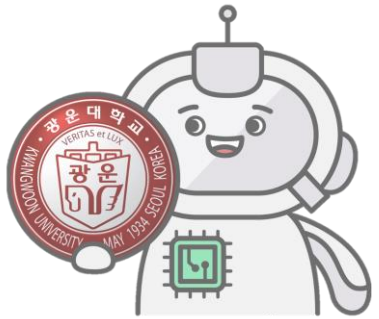




2. 강화학습(Reinforcement Learning)

- 강화 학습 예제: 그리드 월드(Grid World)
  - 골 지점에 도착한 경우, 이전 상태에서 취한 액션의 보상 값을 업데이트
  - $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} - Q(s_t, a_t)]$

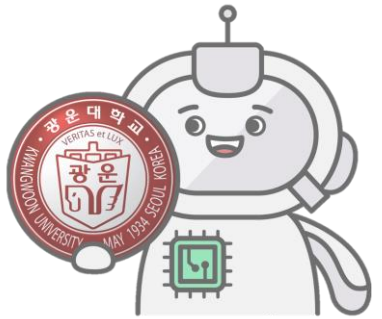
(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	1
(0, 0, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	(1, 0, 0, 0)



2. 강화학습(Reinforcement Learning)

- 강화 학습 예제: 그리드 월드(Grid World)
  - 각 상태에서 이동 가능한 상태들 중, 보상 값을 가장 많이 얻을 수 있는 상태로 이동하기 위한 액션-보상 업데이트
  - $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \cdot \max_{a`} Q(s_{t+1}, a`) - Q(s_t, a_t)]$

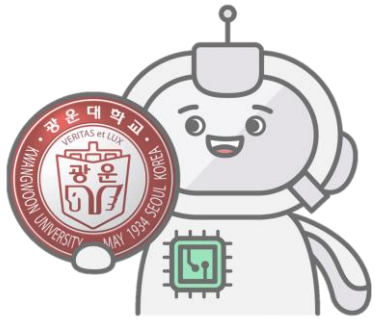
(0, 1, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	1
(0, 0, 0, 1)	(0, 0, 0, 1)	(0, 0, 0, 1)	(1, 0, 0, 0)



2. 강화학습(Reinforcement Learning)

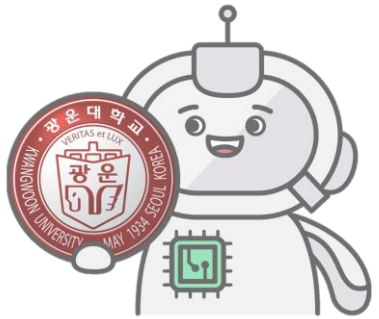
- 강화 학습 예제: 그리드 월드(Grid World)
  - ➢ 다음 에피소드에서 Q-Table을 기반으로 액션을 선택

(0, 1, 0, 0)	(0, 0, 0, 0)	(0, 0, 0, 0)	1
(0, 0, 0, 1)	(0, 0, 0, 1)	(0, 0, 0, 1)	(1, 0, 0, 0)



2. 강화학습(Reinforcement Learning)

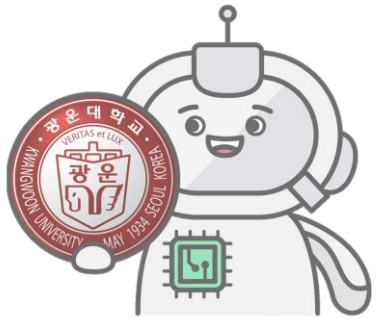
- $\epsilon - Gredy$
  - 탐험(exploration)과 이용(exploitation)의 균형을 조절하는 행동 선택 전략
  - 확률적으로 최적의 행동을 선택하지 않고, 가끔은 랜덤하게 행동함
  - $a` = \begin{cases} \arg\max_a Q(s,a) & with\ probability\ 1-\epsilon \\ random\ action & with\ probability\ \epsilon \end{cases}$
- **사용 목적**
  - 탐험 보장 : 최적이 아닐 수도 있는 Q값에 의존하지 않음
  - 수렴 보조 : 다양한 상태-액션을 통해 Q값의 정확도를 높임
  - 균형 조절 : ϵ을 줄이면서 점점 exploitation 중심 학습으로 전환



2. 강화학습(Reinforcement Learning)

- 강화 학습 예제: 그리드 월드(Grid World)
  - 최적의 경로 학습 혹은 더 높은 보상 값을 위한 $\epsilon - Gredy$방법을 통한 탐험(Exploration)과 활용(Exploitation)을 조율

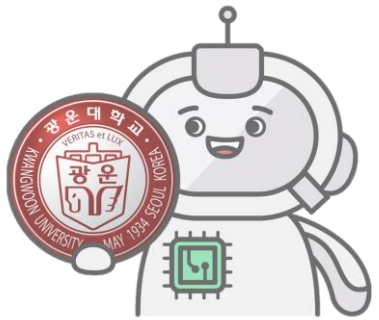
S_0	S_1	S_2	S_3
S_4	S_5	S_6	S_7
S_8	S_9	S_{10}	S_{11}



2. 강화학습(Reinforcement Learning)

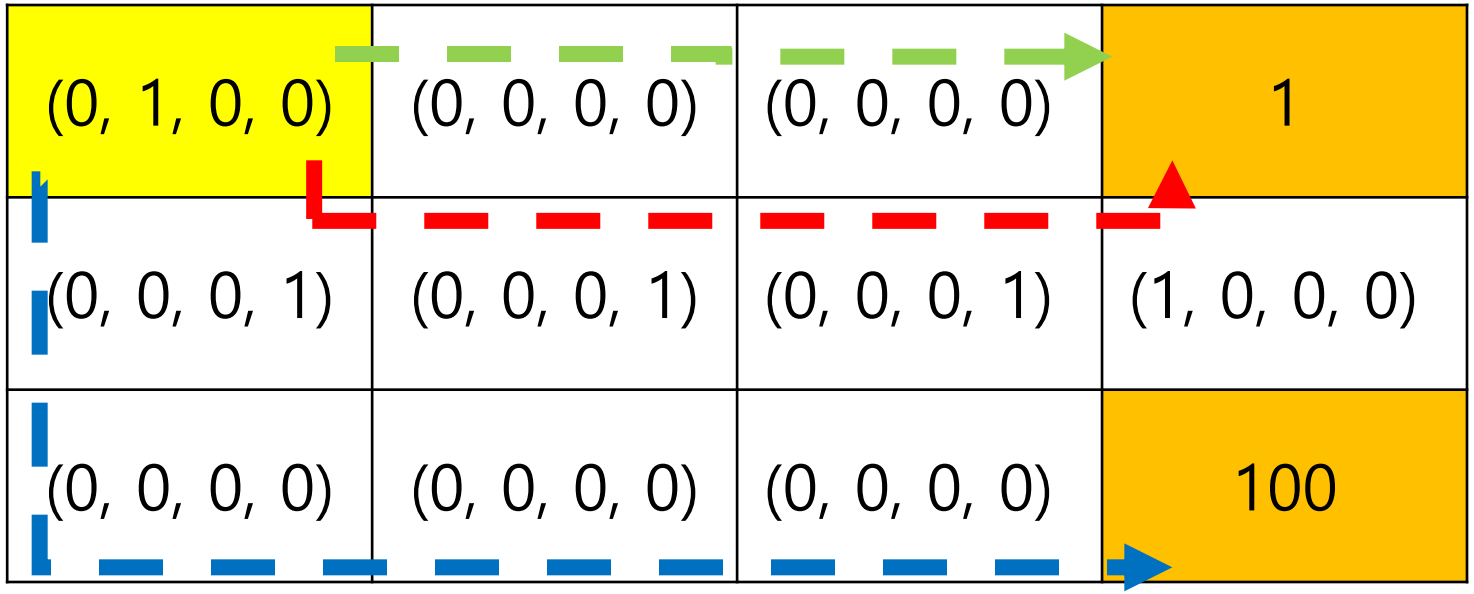
- 강화 학습 예제: 그리드 월드(Grid World)
  - 최적의 경로 학습 혹은 더 높은 보상 값을 위한 $\epsilon - Gredy$방법을 통한 탐험(Exploration)과 활용(Exploitation)을 조율

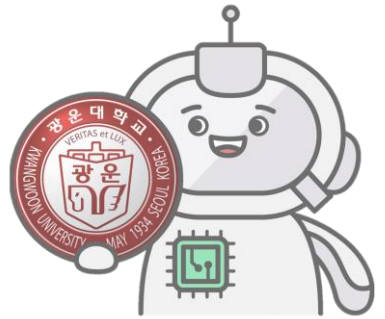
0	0	0	1
0	0	0	0
0	0	0	100



2. 강화학습(Reinforcement Learning)

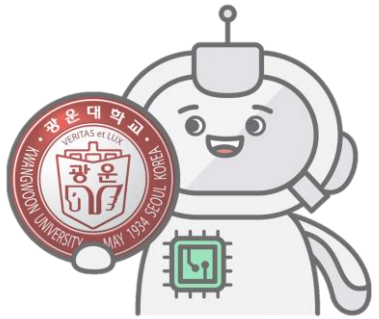
- 강화 학습 예제: 그리드 월드(Grid World)
  - ➢ 최적의 경로 학습 혹은 더 높은 보상 값을 위한 $\epsilon - Gredy$방법을 통한 탐험(Exploration)과 활용(Exploitation)을 조율





2. 강화학습(Reinforcement Learning)

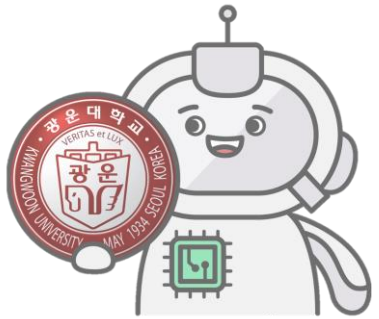
- 할인율(Discount factor, γ)
  - 에이전트가 **미래 보상**을 얼마나 중요하게 생각하는지를 결정하는 계수
  - 보통 $0 \leq \gamma \leq 1$ 사이의 값을 가짐
  - $V(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$
  - 값이 1에 가까울 수록 이후 에피소드에서 얻는 보상 값을 크게 반영
  - 값이 0에 가까울 수록 미래 보상은 무시하고 현재 보상을 크게 반영



2. 강화학습(Reinforcement Learning)

- 강화 학습 예제: 그리드 월드(Grid World)
  - 할인율(Discount factor, γ)
  - 좀 더 효율적인 action이 되도록 계산

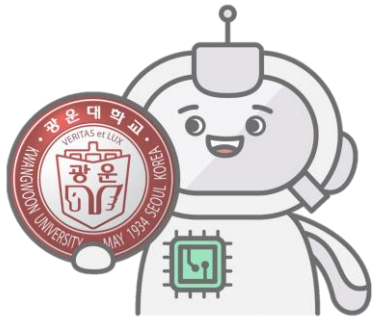
(0, 1, 0, 1)	(0, 0, 0, 1)	(0, 0, 0, 1)	1
(0, 0, 0, 1)	(0, 0, 0, 1)	(0, 0, 0, 1)	(1, 0, 0, 0)



2. 강화학습(Reinforcement Learning)

- 강화 학습 예제: 그리드 월드(Grid World)
  - 할인율(Discount factor, γ)
  - $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \cdot \max_{a`} Q(s_{t+1}, a`) - Q(s_t, a_t)]$

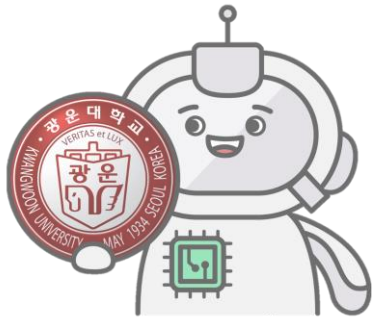
$(0, \gamma^5, 0, \gamma^3)$	$(0, 0, 0, \gamma^2)$	$(0, 0, 0, \gamma)$	1
$(0, 0, 0, \gamma^4)$	$(0, 0, 0, \gamma^3)$	$(0, 0, 0, \gamma^2)$	$(\gamma, 0, 0, 0)$



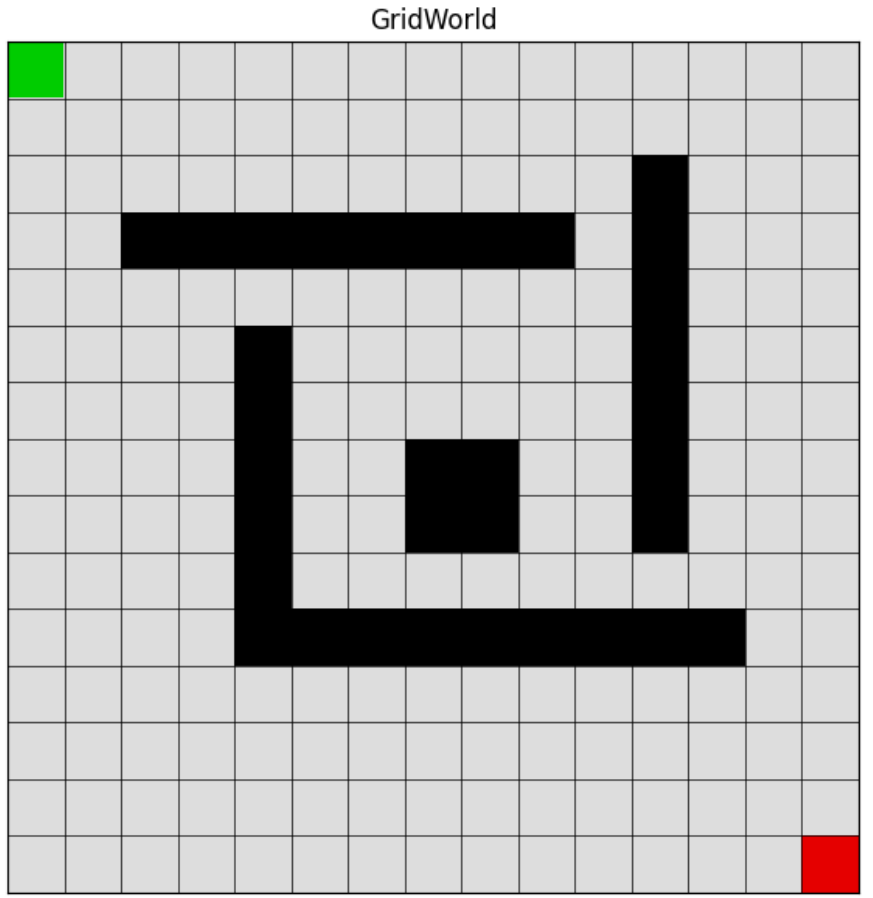
2. 강화학습(Reinforcement Learning)

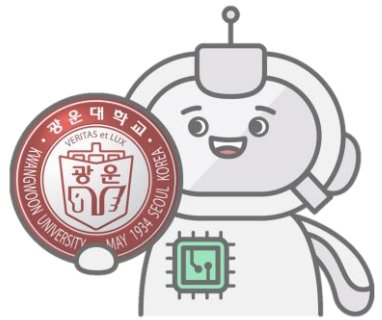
표 4. Q-learning 기반 경로 탐색 실험을 위한 환경 설정 및 하이퍼파라미터

가정	설명
시작 위치	(0, 0)
목표 위치	(14, 14)
액션	상(0), 하(1), 좌(2), 우(3)
보상	일반(-1), 장애물(-100), 도착(100)
학습률	$\alpha = 0.1$
할인율	$\gamma = 0.9$
ϵ	1.0, 매 에피소드 마다 0.005씩 감소
Max_step	200
Episode	5000

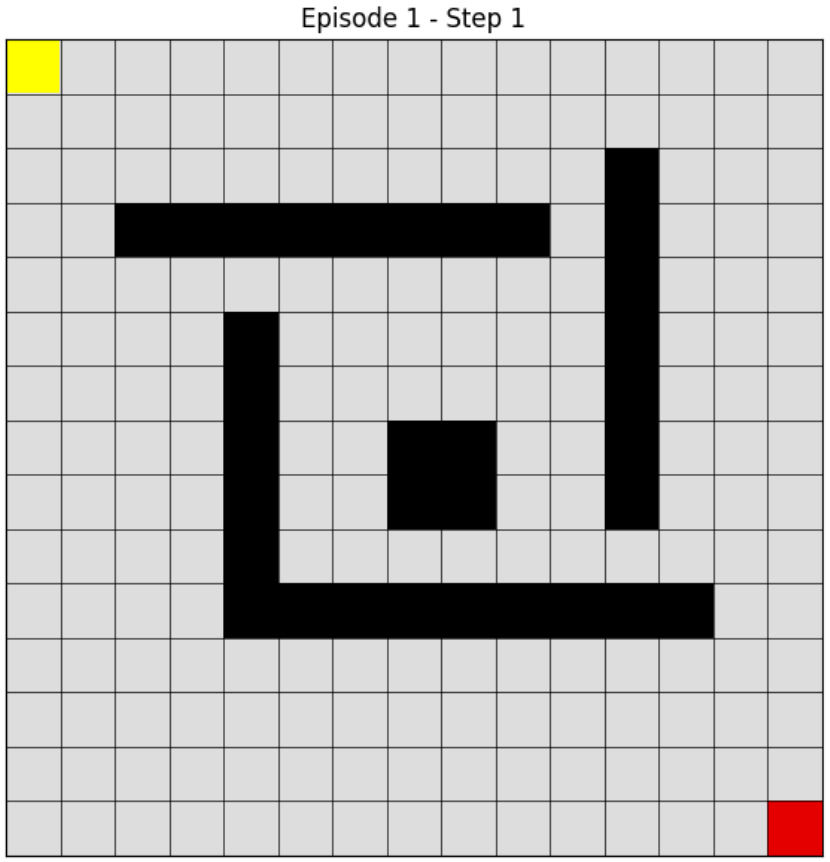


2. 강화학습(Reinforcement Learning)



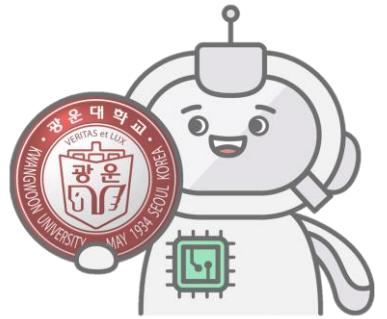


2. 강화학습(Reinforcement Learning)

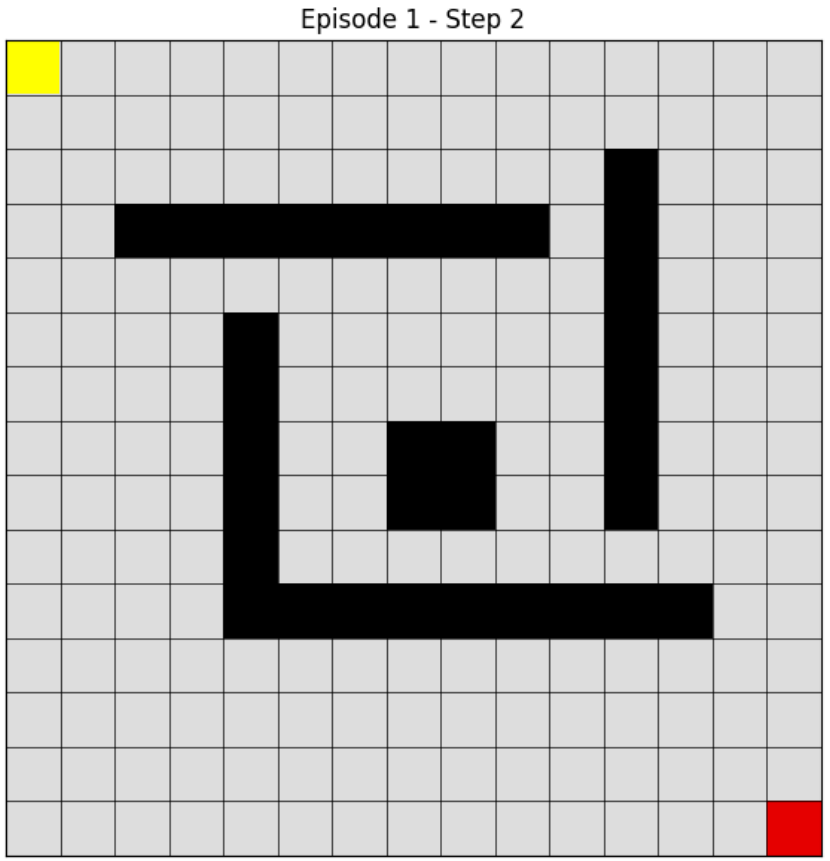


- Step 1
  - Action: 0 → (-1, 0)
  - New State: (0, 0)
  - $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \cdot \max_{a`} Q(s_{t+1}, a`) - Q(s_t, a_t)]$

    $Q(0,0) \leftarrow 0 + 0.1 \cdot (-1 + 0.9 \cdot 0 - 0) = -0.1$
  - Q(0, 0): [-0.1, 0.0, 0.0, 0.0]

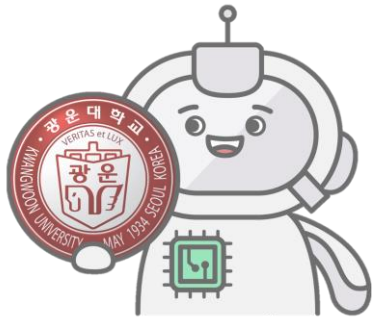


2. 강화학습(Reinforcement Learning)

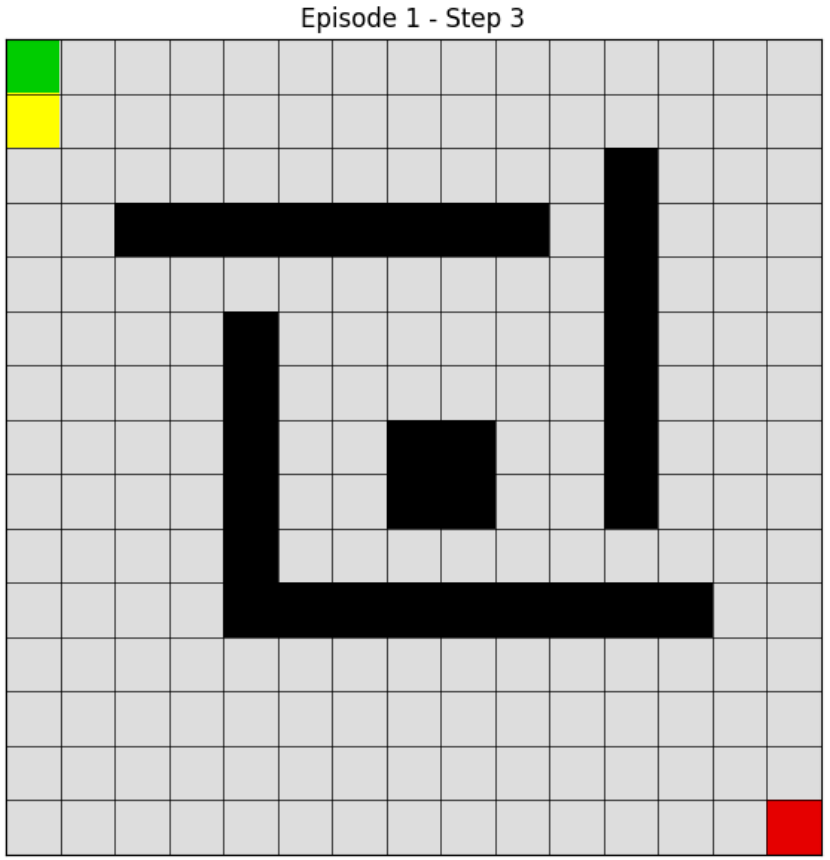


- Step 2
  - Action: 2 → (0, -1)
  - New State: (0, 0)
  - $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \cdot \max_{a`} Q(s_{t+1}, a`) - Q(s_t, a_t)]$

    $Q(0,2) \leftarrow 0 + 0.1 \cdot (-1 + 0.9 \cdot 0 - 0) = -0.1$
  - Q(0, 0): [-0.1, 0.0, -0.1, 0.0]

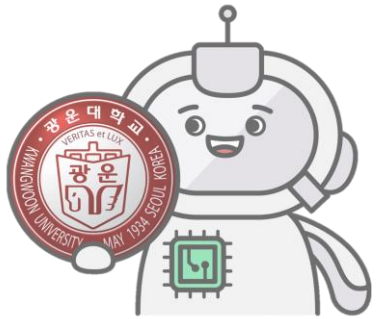


2. 강화학습(Reinforcement Learning)

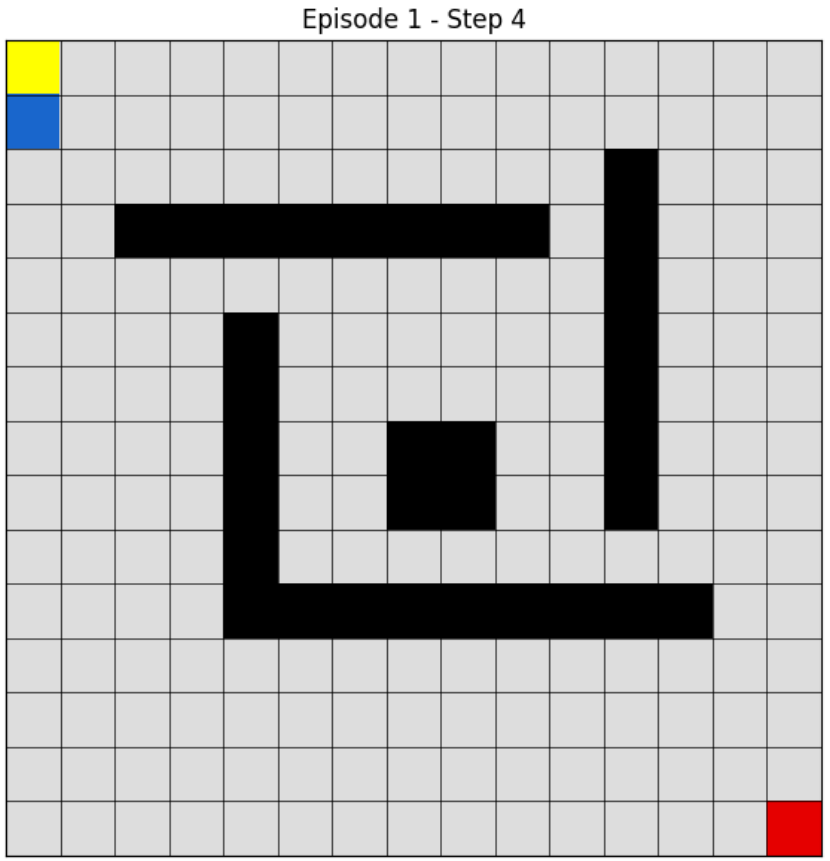


- Step 3
  - Action: 1 → (1, 0
  - New State: (1, 0)
  - $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \cdot \max_{a`} Q(s_{t+1}, a`) - Q(s_t, a_t)]$

    $Q(0,1) \leftarrow 0 + 0.1 \cdot (-1 + 0.9 \cdot 0 - 0) = -0.1$
  - Q(0, 0): [-0.1, -0.1, -0.1, 0.0]
  - Q(1, 0): [0.0, 0.0, 0.0, 0.0]

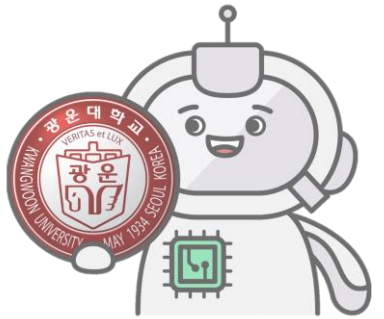


2. 강화학습(Reinforcement Learning)

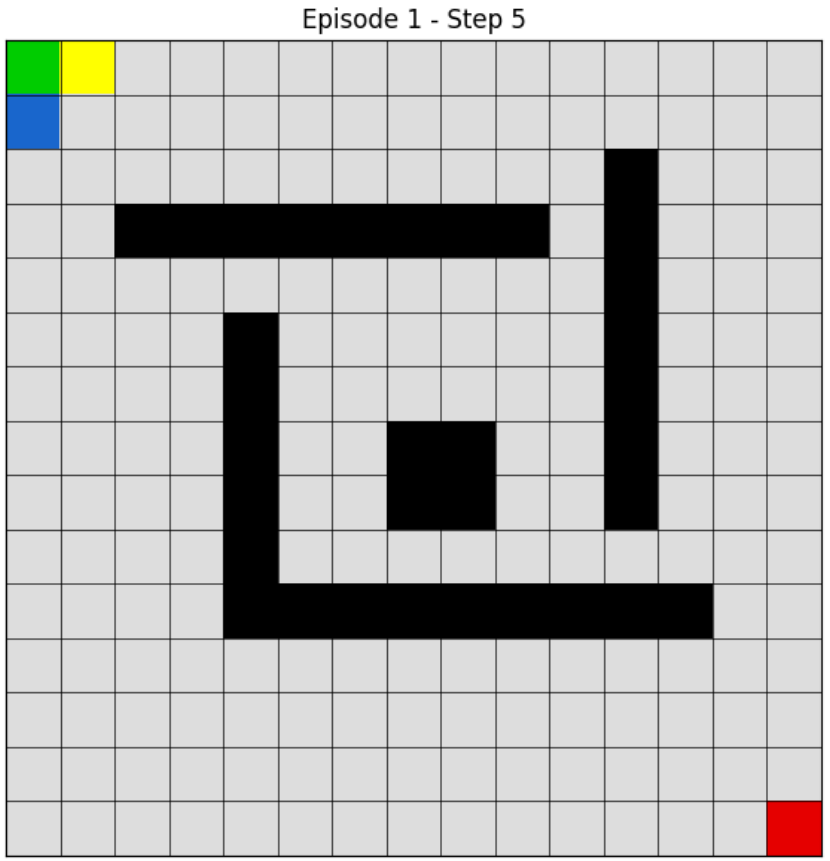


- Step 4
  - Action: 0 → (-1, 0)
  - New State: (0, 0)
  - $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \cdot \max_{a`} Q(s_{t+1}, a`) - Q(s_t, a_t)]$

    $Q(1,0) \leftarrow 0 + 0.1 \cdot (-1 + 0.9 \cdot 0 - 0) = -0.1$
  - Q(0, 0): [-0.1, -0.1, -0.1, 0.0]
  - Q(1, 0): [-0.1, 0.0, 0.0, 0.0]

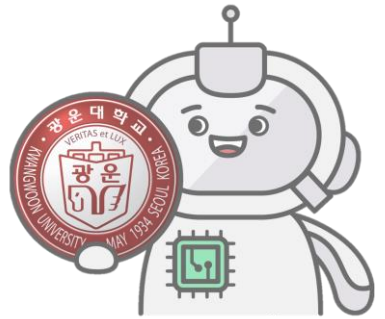


2. 강화학습(Reinforcement Learning)

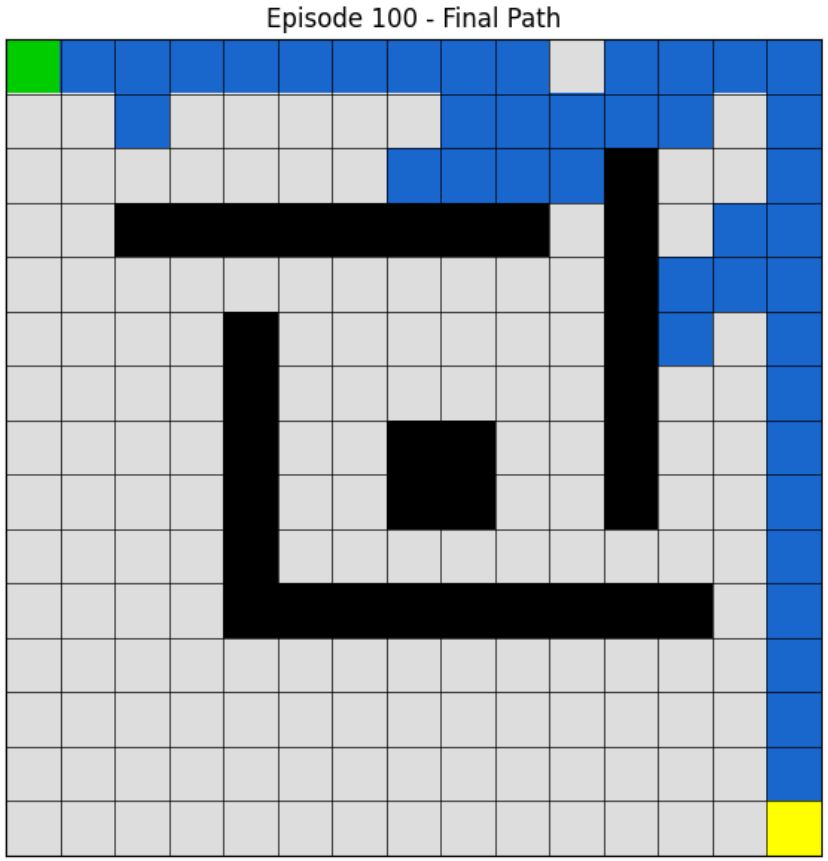


- Step 5
  - Action: 3 → (0, 1)
  - New State: (0, 1)
  - $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \cdot \max_{a`} Q(s_{t+1}, a`) - Q(s_t, a_t)]$

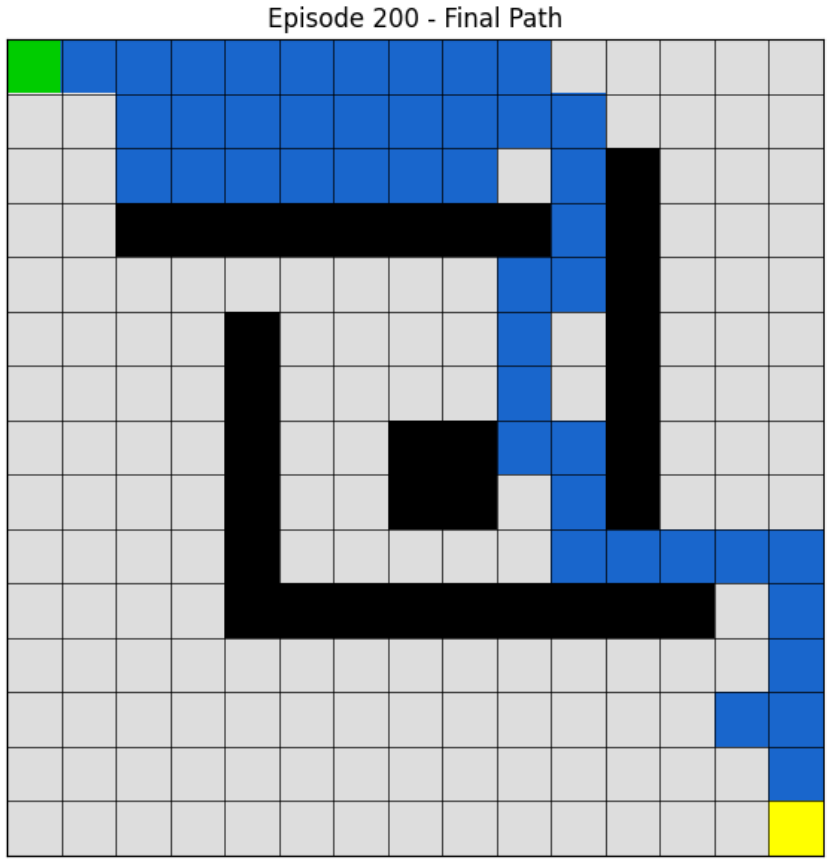
    $Q(0,2) \leftarrow 0 + 0.1 \cdot (-1 + 0.9 \cdot 0 - 0) = -0.1$
  - Q(0, 0): [-0.1, -0.1, -0.1, -0.1]
  - Q(1, 0): [-0.1, 0.0, 0.0, 0.0]
  - Q(0, 1): [0.0, 0.0, 0.0, 0.0]

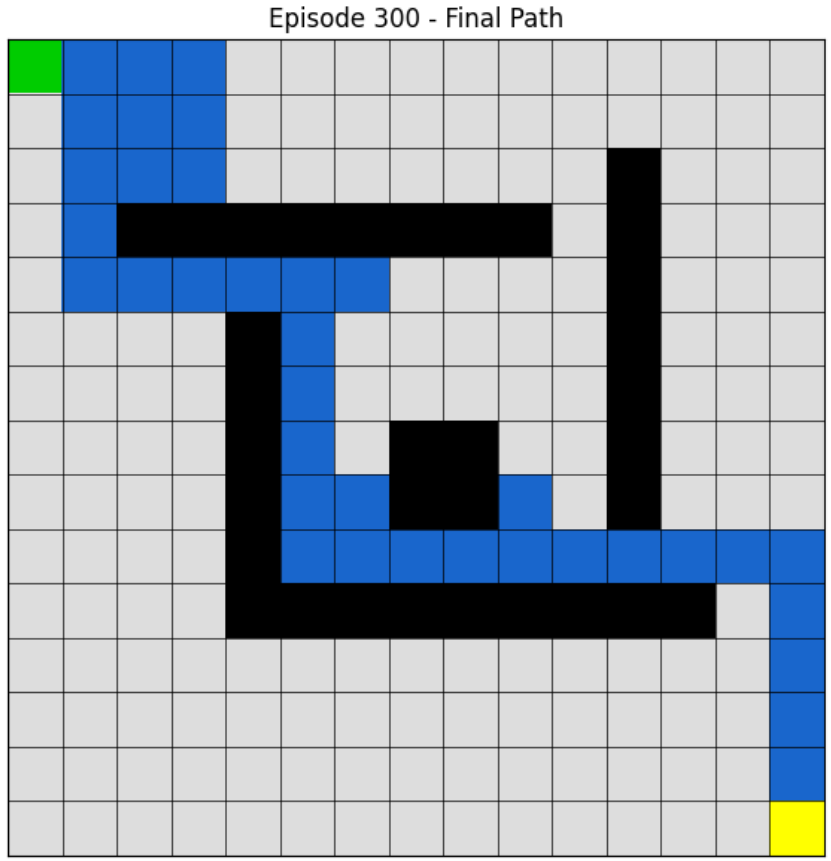


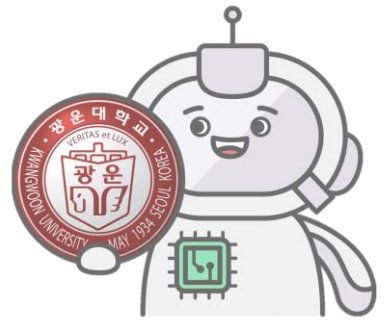
2. 강화학습(Reinforcement Learning)



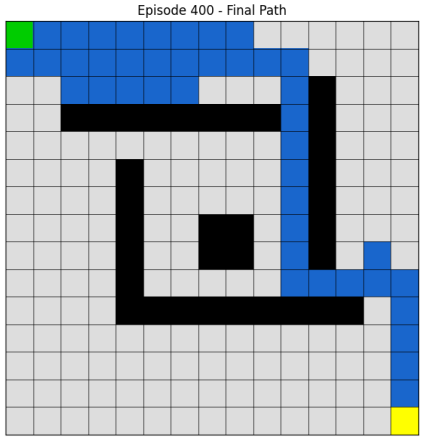
2. 강화학습(Reinforcement Learning)

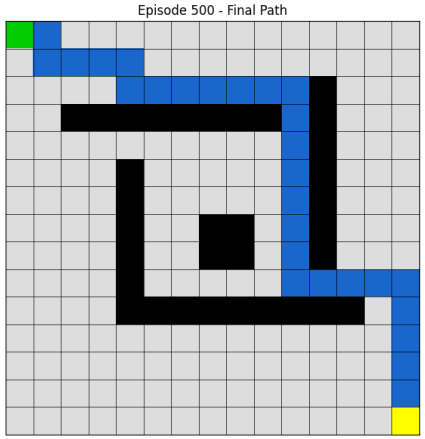


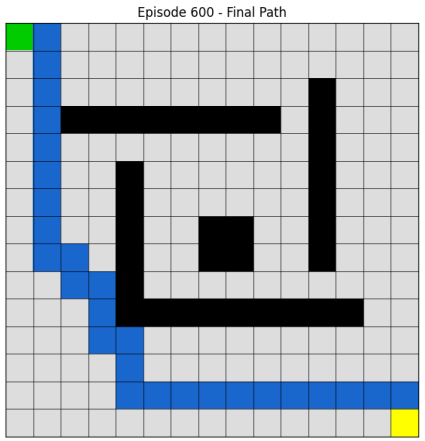


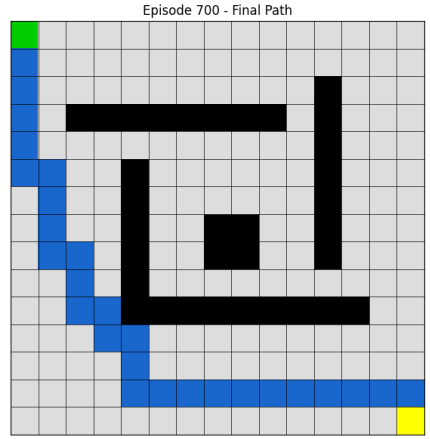


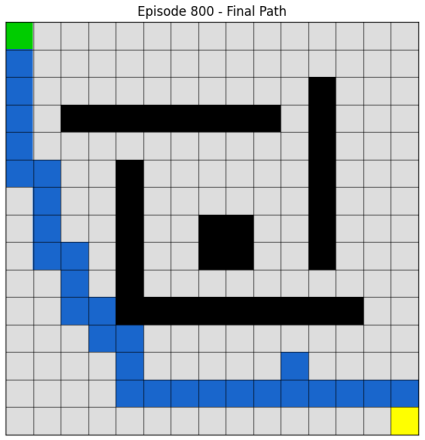
2. 강화학습(Reinforcement Learning)

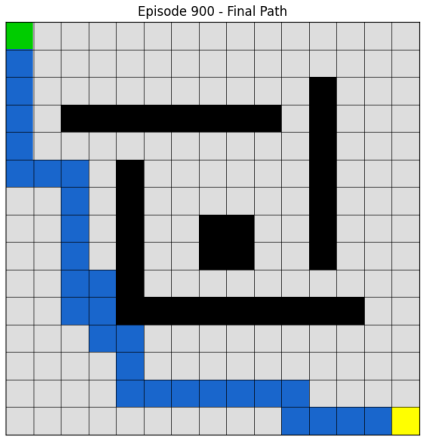


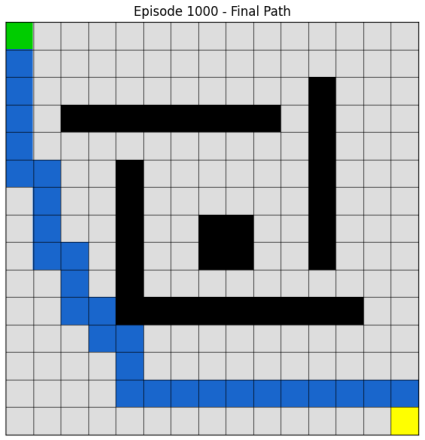




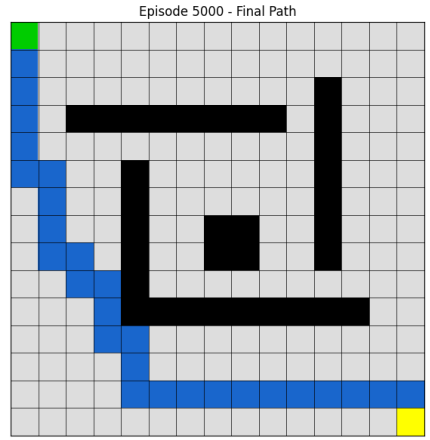


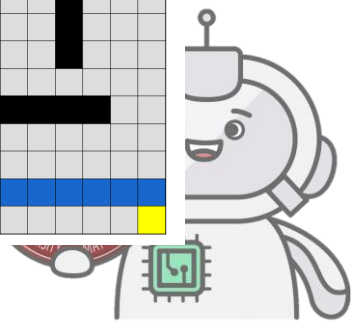






...





2. 강화학습(Reinforcement Learning)

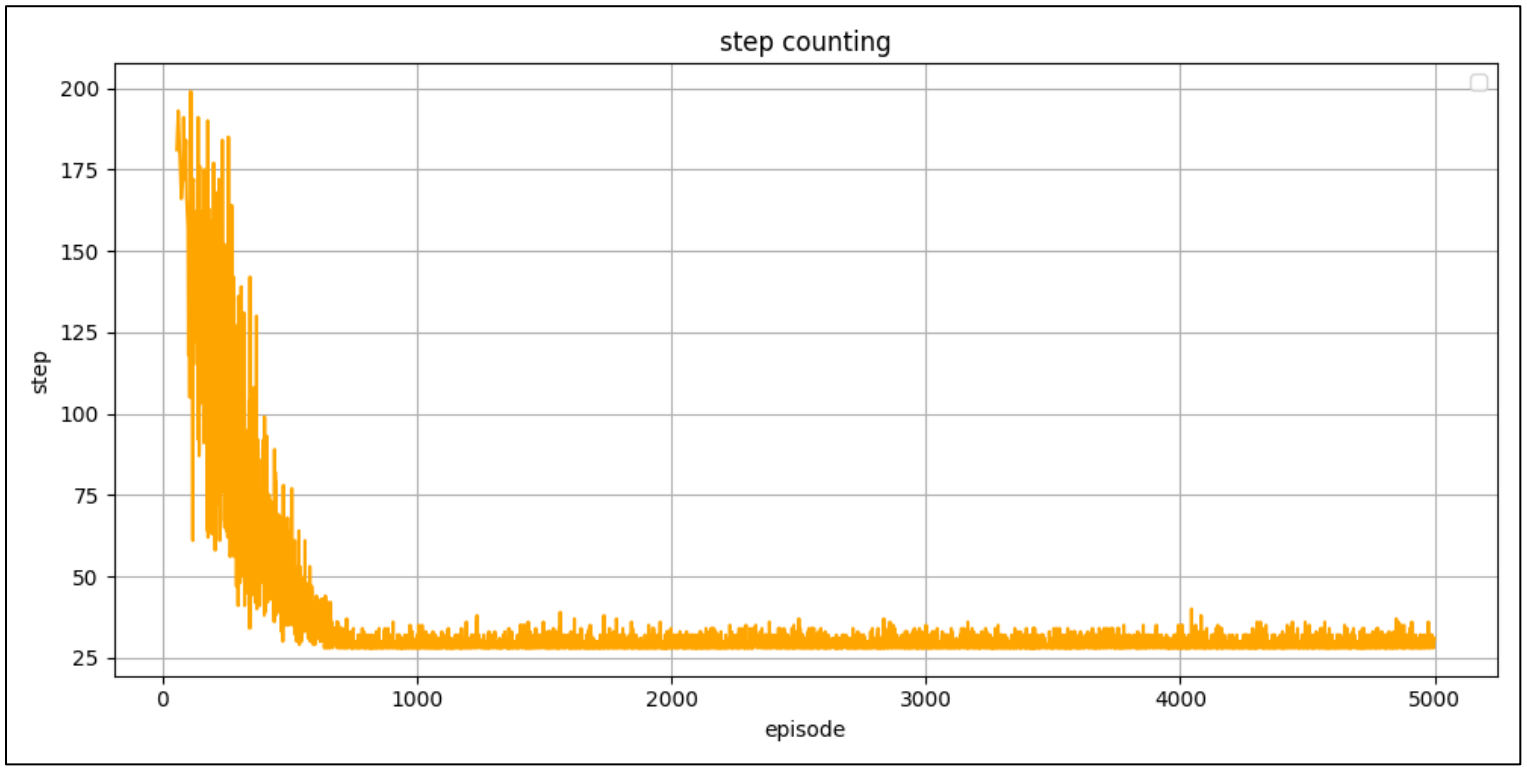


그림 7. Q-learning을 이용한 2차원 그리드 환경에서 에이전트가 골 지점에 도달하는 데 소요된 스텝 수의 변화

3. Future Work

- Markov Decision Process(MDP)
- 상태 가치함수 V & 행동 가치 함수 Q
- 벨만 방정식
- Monte Carlo(MC)
- Temporal difference(TD) & SARSA