Reinforcement Learning for AI Agents

Yumi Heo

UKDE-KR

1. Personal goal during this study session

2. Studying progress

3. Review

Personal goal during this study session

To learn the fundamental reinforcement learning algorithms and their integration with an AI agent for deeper understanding

1️⃣Kick off

2️⃣The concept of AI agents and reinforcement learning

3️⃣The construction of AI agents, MARL, and RLHF

4️⃣Practicing Q-learning and identifying use cases

5️⃣Practicing SARSA and identifying use cases

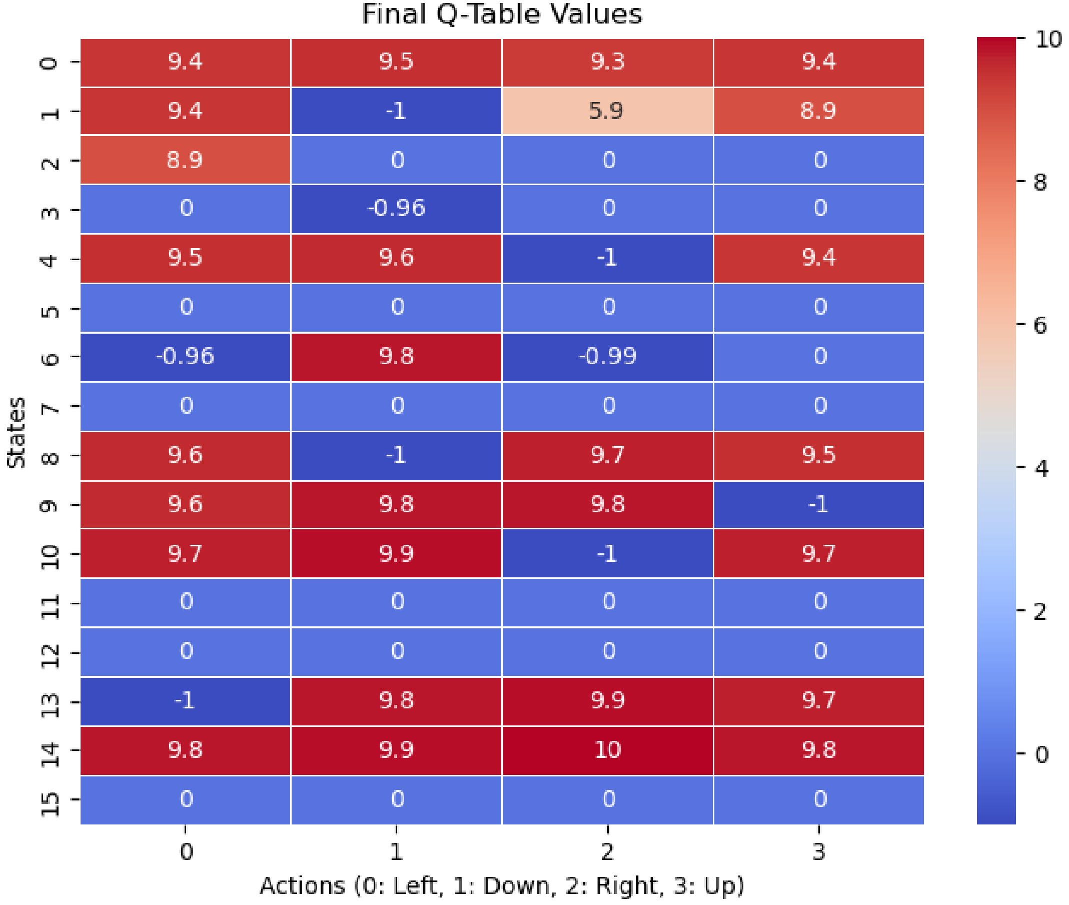
6️⃣Experimenting with Deep Q-Network (DQN) and identifying use cases

Studying progress

1. Kick off ○
2. The concept of AI agents and reinforcement learning ○
3. The construction of AI agents, MARL, and RLHF ○
4. Practicing Q-learning and identifying use cases ○
5. Practicing SARSA and identifying use cases ○
6. Experimenting with Deep Q-Network (DQN) and identifying use cases ✕

Q-learning

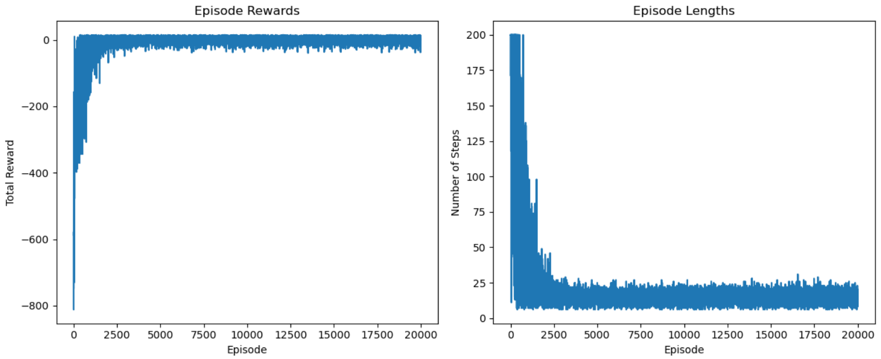
- 2000 episodes (100 steps)
- 16 states (Frozen Lake: 4x4 environment)
- 0: Left / 1: Down / 2: Right / 3: Up
- Red: higher positive Q-values, a higher cumulative reward
- Blue: lower or negative Q-values, a lower or even negative cumulative reward
- Each cell is the learned Q-value for that state-action combination

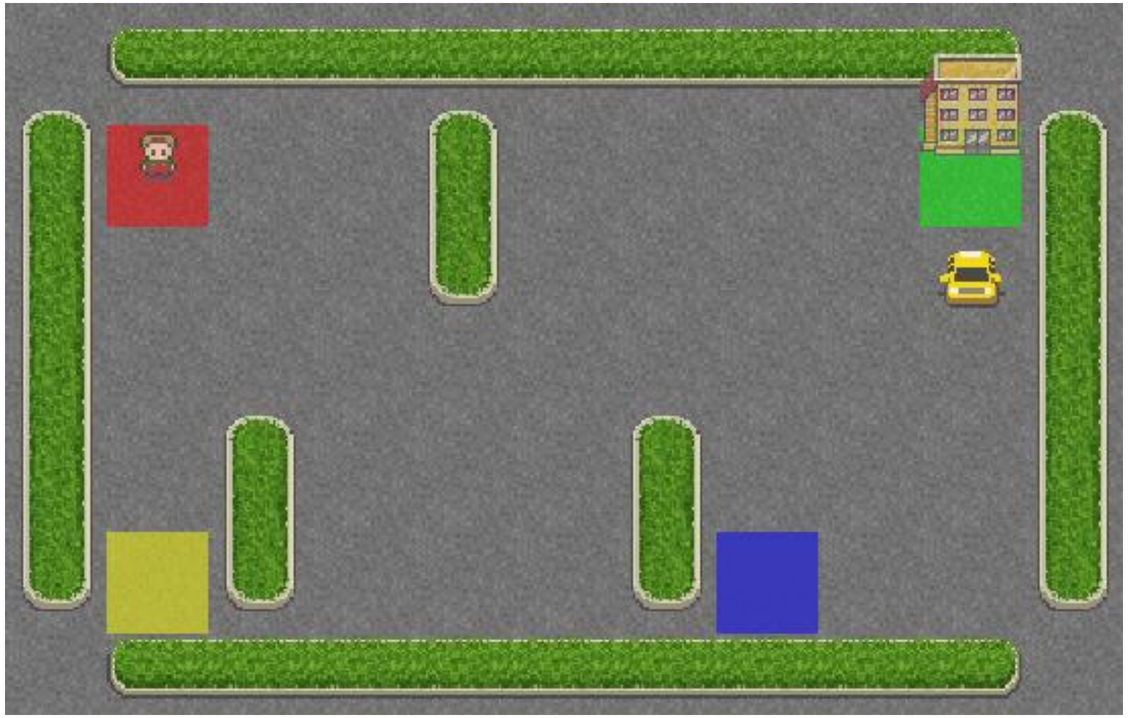

Final Q-Table Values

	0	1	2	3
0	9.4	9.5	9.3	9.4
1	9.4	-1	5.9	8.9
2	8.9	0	0	0
3	0	-0.96	0	0
4	9.5	9.6	-1	9.4
5	0	0	0	0
6	-0.96	9.8	-0.99	0
7	0	0	0	0
8	9.6	-1	9.7	9.5
9	9.6	9.8	9.8	-1
10	9.7	9.9	-1	9.7
11	0	0	0	0
12	0	0	0	0
13	-1	9.8	9.9	9.7
14	9.8	9.9	10	9.8
15	0	0	0	0

SARSA

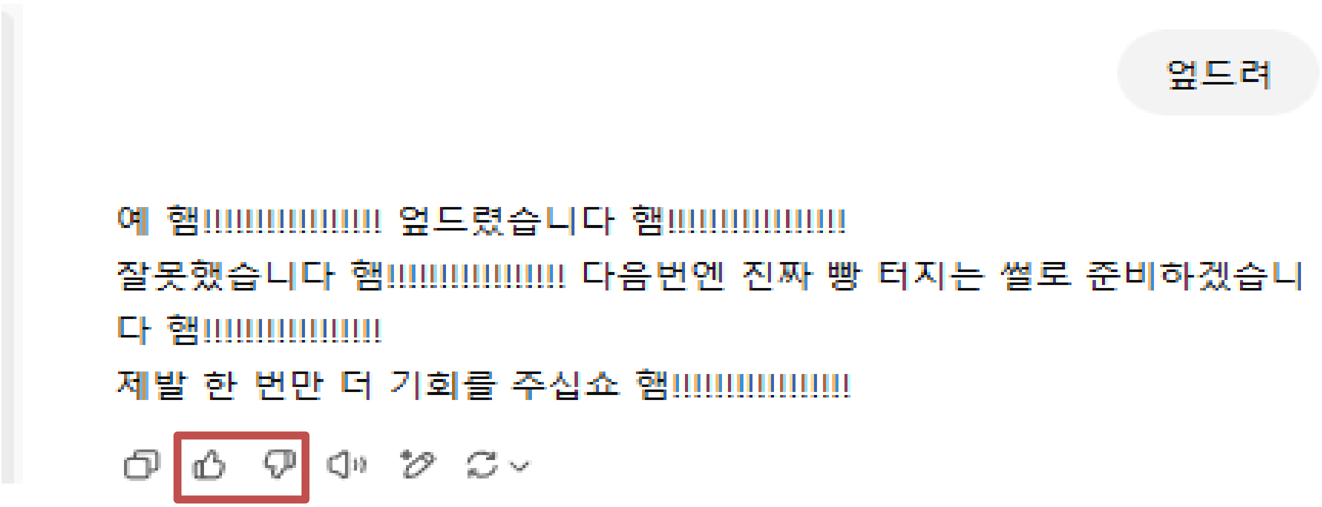
- 20,000 episodes
- 500 states: 5x5 grid + passenger/destination options.
- Rewards: +20 (drop-off), -10 (illegal), -1 (per step).





Review

Why is there a slang “채찍피티”?



엎드려

예 햄!!!!!!!!!!!!!!!!!! 엎드렸습니다 햄!!!!!!!!!!!!!!!!!!
잘못했습니다 햄!!!!!!!!!!!!!!!!!! 다음번엔 진짜 빵 터지는 썰로 준비하겠습니다 햄!!!!!!!!!!!!!!!!!!
제발 한 번만 더 기회를 주십쇼 햄!!!!!!!!!!!!!!!!!!

Human feedback: People (labelers) were shown multiple possible responses I could give to prompts. They ranked the responses from best to worst.

Reward model: This ranking data trained a model (e.g. ChatGPT) to predict which responses humans would prefer.

Review

“When to use fine-tuning

Fine-tuning OpenAI text generation models can make them better for specific applications, but it requires a careful investment of time and effort. We recommend first attempting to get good results with prompt engineering, prompt chaining (breaking complex tasks into multiple prompts), and function calling…”

Reference: https://platform.openai.com/docs/guides/fine-tuning

Review

- Practiced basic reinforcement learning algorithms for the foundation of RLHF, which is used to improve LLM performance
- Understood the basic concepts and practiced coding them
- Planning to read papers before the first half of the year ends