# A Study on Application of Curriculum Learning in Deep Reinforcement Learning

Author: Ikumi KODAKA [s20s3011ea@s.chibakoudai.jp]
Fumiaki SAITOH [fumiaki.saitoh@p.chibakoudai.jp]

**Published: IEEE *Xplore*** 30 November 2021

# Outline

- Introduction
- Deep Reinforcement Learning
- Curriculum Learning
- Experimental Environment
- Experiment
- Result
- Discussion

# Introduction

- Deep reinforcement learning can be applied to tasks with many states.
- However, as the number of task states increases, the number of trials increases significantly.
- Therefore, we aim to improve the efficiency of the number of trials by using curriculum learning, in which the difficulty of a task is changed step by step during learning.

# Deep Reinforcement Learning

- Deep reinforcement learning receives a state (st) every time (t) and selects an action (at) based on a strategy (π). A reward (rt+1) is given for each of these actions.
- The goal is to maximize the reward by trial and error.
- Each action for each state has a Q value.
- Each action for each state has a Q value.

# Curriculum Learning

- Generate tasks that serve as the basis for the target task, called a "curriculum
- Training agents for each task to learn the target task
- Gradually increasing the difficulty level improves learning efficiency in reinforcement learning.

# Experimental environment

- An overview of the convolutional neural network used in the DQN for the experiments is shown in figure

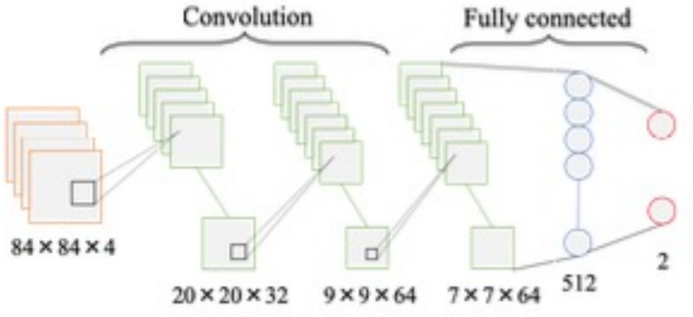

Fig. 4. Convolutional Neural Network

# Experimental environment

- This time, Boltzmann selection was used as the agent's action selection method.

$$\pi(s,a) = \frac{exp\left(\frac{Q(s,a)}{\tau}\right)}{\sum_{b \in A} \exp\left(\frac{Q(s,b)}{\tau}\right)}$$

# Experimental environment

- The game to be studied is a vertical-scrolling shooter game.
- There is a variable that counts time called "step".
- The reward is +1 when the player defeats the enemy, and -1 when the player is defeated by the enemy.
- An episode ends when the player is defeated by an enemy.

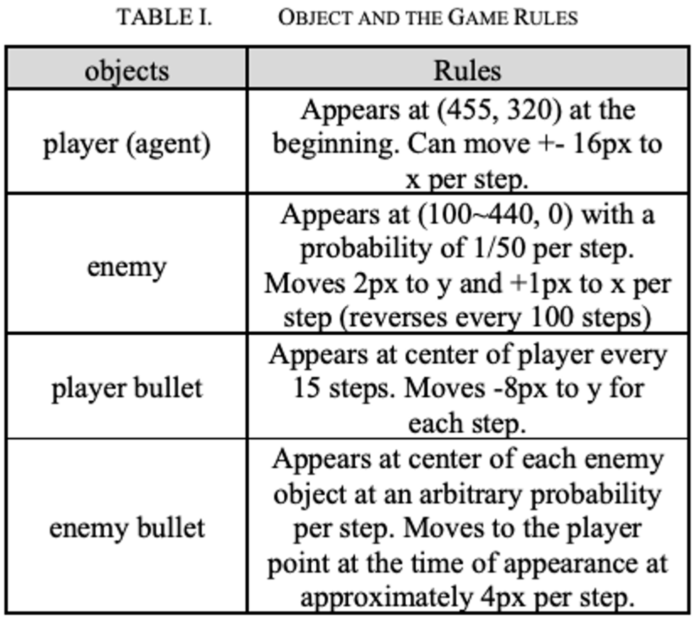

TABLE I. OBJECT AND THE GAME RULES

| objects | Rules |
|---|---|
| player (agent) | Appears at (455, 320) at the beginning. Can move +- 16px to x per step. |
| enemy | Appears at (100~440, 0) with a probability of 1/50 per step. Moves 2px to y and +1px to x per step (reverses every 100 steps) |
| player bullet | Appears at center of player every 15 steps. Moves -8px to y for each step. |
| enemy bullet | Appears at center of each enemy object at an arbitrary probability per step. Moves to the player point at the time of appearance at approximately 4px per step. |

# Experiment

・This time, we will experiment with normal learning (DQN) and curriculum learning (CL) and compare the two.

・The difficulty level is changed by the probability of the enemy firing bullets (others are not changed).

・The highest difficulty level has a probability of 1/30 (bullets/step).

・The number of trials is 5 times with 10000 episodes as one trial.

# Experiment

- DQN without changing it from a probability of 1/30. Do 10000 episodes.

- The following table is used for CL.

  - There are CL_A,C,E that change the difficulty every 1000 episodes and CL_B,D,F that change the difficulty every 2000 episodes.

- After 6000 episodes of CL, the probability does not change from 1/30.

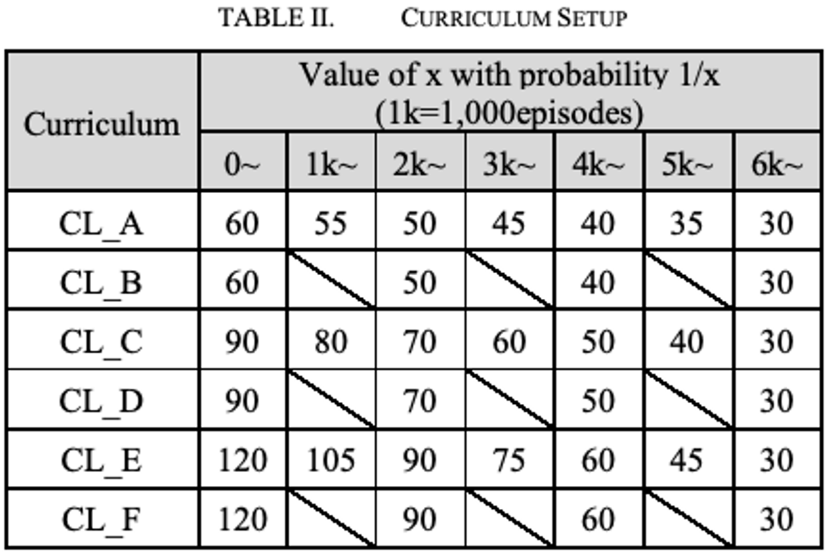

TABLE II. CURRICULUM SETUP

| Curriculum | Value of x with probability 1/x (1k=1,000episodes) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0~ | 1k~ | 2k~ | 3k~ | 4k~ | 5k~ | 6k~ |
| CL_A | 60 | 55 | 50 | 45 | 40 | 35 | 30 |
| CL_B | 60 | | 50 | | 40 | | 30 |
| CL_C | 90 | 80 | 70 | 60 | 50 | 40 | 30 |
| CL_D | 90 | | 70 | | 50 | | 30 |
| CL_E | 120 | 105 | 90 | 75 | 60 | 45 | 30 |
| CL_F | 120 | | 90 | | 60 | | 30 |

# Result

- The following chart compares the average reward obtained for each episode.

- The following chart shows the average reward for each episode.

- The following chart shows the average reward for each episode.

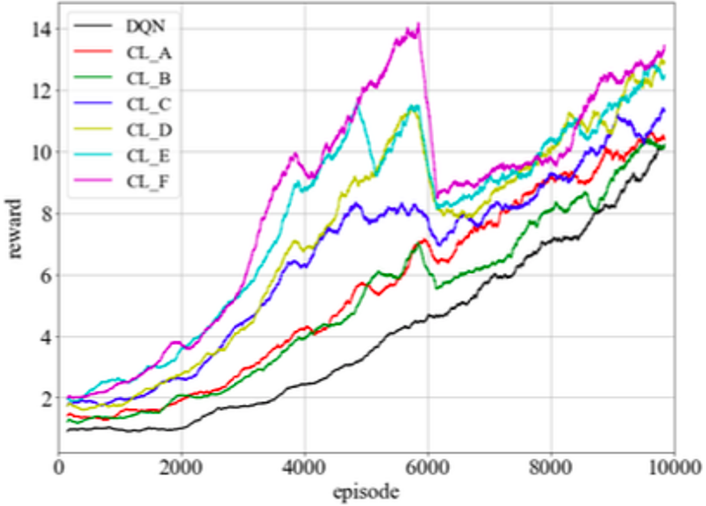

Fig. 7. The relationship between episodes and rewards

# Result

- This figure compares the rewards obtained after 6000 episodes.

- In all CLs, the rewards immediately after 6000 episodes were lower.

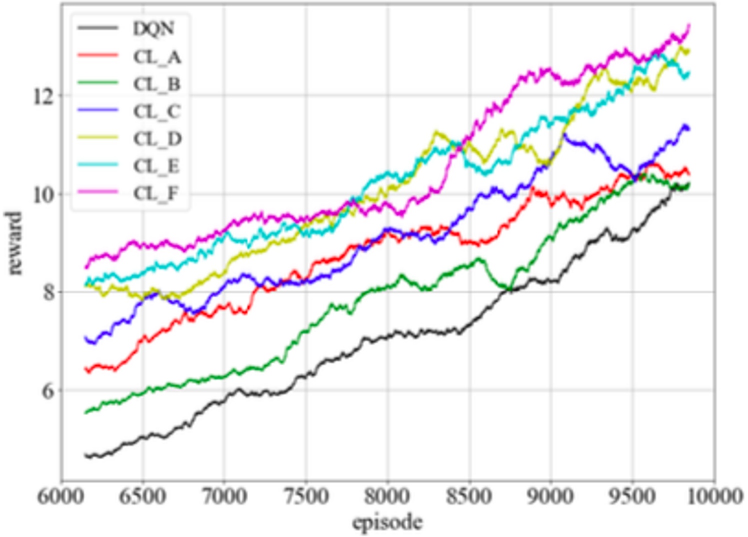

Fig. 8. The relationship between episodes and rewards (After 6,000 episodes)

# Result

・This figure shows the average of the Q values obtained for each episode.

・The increase in Q-values was higher for CL than for DQN.

・The probability of choosing an action depends on the Q-value of Boltzmann's choice, so the knowledge gained from the action is easily reflected.

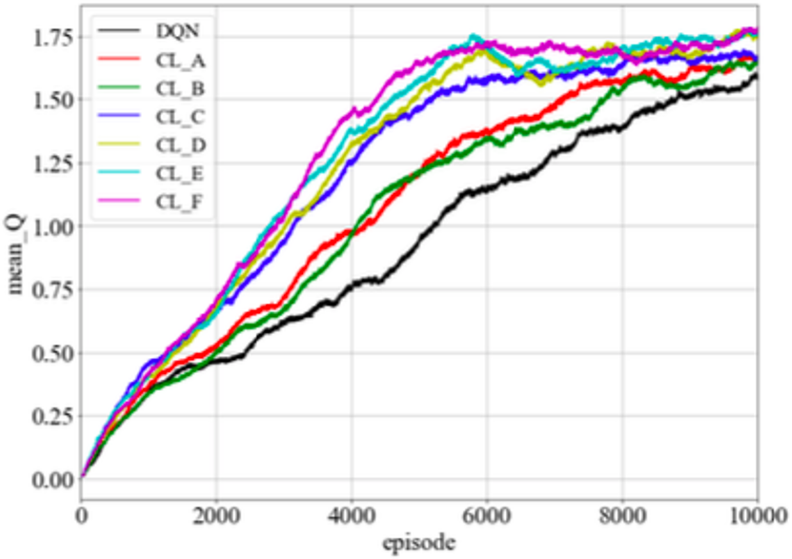

Fig. 9. The relationship between episodes and Q-value

# Discussion

- The results showed that the number of trials decreased when curriculum learning was used.
- The results also showed that the same number of trials produced a difference in learning performance when the difficulty level was adjusted.
- However, they could not confirm the effectiveness of the simplification.

# Thank you for listenning