

KNOWLEDGE REUSE DEGREE ASYMMETRY IN TRANSFER REINFORCEMENT LEARNING

Satoru Ikeda ¹, Kohei Esaki ², Hitoshi Kono ², Shota Chikushi ³,
Kaori Watanabe ⁴ and Hidekazu Suzuki ⁵

¹ Graduate School of Advanced Science and Technology, Tokyo Denki University,
Tokyo, Japan

² Department of Information and Communication Engineering, Tokyo Denki University,
Tokyo, Japan

³ Department of Robotics, Kindai University, Hiroshima, Japan

⁴ New Technology Foundation, Tokyo, Japan

⁵ Department of Engineering, Tokyo Polytechnic University, Kanagawa, Japan

ABSTRACT

In recent years, autonomous systems such as automated driving using machine learning have been developed and are being implemented in society. Therefore, reinforcement learning and transfer learning, which are considered useful for control methods of automated systems, are being studied. transfer learning is a method of reusing knowledge from the source task to the target task, and knowledge means that the policy, action-value function, model, and so on, in the machine learning domain. In the transferring situation, to avoid negative transfer such as over learning, the transfer rate can be adjusted transferring action value amplitude can be set. However, the transfer rates are usually determined by human experiences. The automatic transfer rate adjustment method proposed by Kono et al. has demonstrated improved environmental adaptability, and the function shape and a step size parameter determine the adjustment. However, the agent's environmental adaptation performance has not been achieved to the transfer rate set by humans. In this paper, separate step size parameters for when the transfer rate increases and decreases, a method is proposed that improves upon the automatic transfer rate adjustment method proposed by Kono et al. Experimental results have shown that reinforcement learning and transfer learning are conducted through simulations using a shortest path problem in two dimensions. Experimental results have verified that the transfer learning method improves adaptive performance compared to the method proposed by Kono et al.

KEYWORDS

Reinforcement learning, Transfer learning, Transfer rate, Agent simulation

1. INTRODUCTION

Autonomous driving technology is being researched and developed by many universities, research centers, and automobile manufacturers around the world. To promote technological development, the Defense Advanced Research Projects Agency (DARPA) has held three DARPA Challenges since 2004 [1]. Since then, many competitions and tests for autonomous driving have been conducted, and research has been accelerating simultaneously. In recent years, autonomous driving systems have been developed for automobiles and have begun to be implemented in society. In Japan, the Society of Automotive Engineers of Japan, Inc. (JSAE) has defined autonomous driving systems by levels based on the standards of the Society of Automotive Engineers (SAE) in the United States. Currently, autonomous driving systems are

classified as Level 4 according to the JSAE definition and are now capable of performing autonomous driving under certain conditions. However, even in the latest autonomous driving systems, accidents have occurred. Additionally, Level 5 autonomous driving systems, which are capable of autonomous all driving tasks, are not yet available. According to A. Swief and M. El-Habrouk, machine learning is used in the control of autonomous driving systems, and described that future issues include improving the accuracy of object recognition technologies, such as sensors, and improving the computational time and accuracy of learning algorithms [2]. Kiran et al. system is based on deep reinforcement learning, it is stated that transfer learning may also be effective, therefore discussion of transfer learning in reinforcement learning (hereinafter called transfer reinforcement learning) also become important [3]. At the same time, the discussion of transfer learning in machine learning is becoming more systematic [4]. This is not only due to transfer learning in machine learning, but also due to its connection to federated learning, which has been developing in recent years [5]. Reinforcement learning is a learning algorithm that allows an agent to learn optimal behavior through autonomous trial and error in a given environment [6]. In addition, transfer reinforcement learning reduces learning time and improves environmental adaptability by reusing policies previously acquired at the previous tasks (source task) at the current tasks (target task) [7][8][9]. Transfer reinforcement learning, an extension of reinforcement learning, may develop further as continual reinforcement learning [10][11].

In this context, it is becoming increasingly important for machine learning to reuse knowledge and to continuously learn, rather than simply learning once. For this reason, it is important to have technology that does not forget knowledge that is to be reused, and that allows transfer learning while adjusting the degree of knowledge reuse. However, if transfer learning is performed in an environment different from the one in the which it was learned, it may not be possible to complete the task, and when reusing the learned policy, there is a problem known as over learning, in which there is excessive adaptation to the environment or task in which it was learned. A transfer efficiency that adjusts the proportion of policy that are reused has been proposed to suppress over-learning [12]. Furthermore, by using the transfer surface proposed by Osgood, it is possible to visually grasp the trend of the transfer learning effect [13]. However, manual adjustment of the transfer rate requires appropriate verification, but it is a necessary trial-and-error process. Kono et al. successfully achieved transfer learning without manual adjustment using an automatic transfer rate adjustment method based on a sigmoid function [14]. However, while the automatic adjustment of the transfer rate has succeeded in reducing the time required for trial and error in the early stages of learning, the learning progression remains similar to that of reinforcement learning, and convergence has not been achieved. Therefore, the aim of this study is to improve the parameter update method for automatic adjustment of the transfer rate based on the automatic adjustment method of Kono et al., in order to reduce the computation time of the learning algorithm required for the automotive system and to enhance the environmental adaptability during transfer learning.

The remainder of this paper is organized as follows. Section II discusses previous and related theories and researches. Section III proposed the method which modified Kono et al. method to improve environmental adaptivity. Section IV presents computer simulation experiments using a two-dimensional grid world. Section V concludes the paper.

2. PREVIOUS RESEARCH

2.1. Reinforcement Learning

Reinforcement learning (RL) is a method which kind of the machine learning. RL agent aims to maximize the reward by the environment through trial-and-error actions, acquiring knowledge of

the optimal solution regarding the set rewards for the given task. Therefore, it does not require the preparation of training data, allowing for implementation with less workload compared to supervised learning. This paper focuses on Q-learning, a method in reinforcement learning that is still actively researched today [15]. Next, the update formula for the action-value function in Q-learning will be presented.

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left\{ r + \gamma \max_{a' \in A} Q(s', a') - Q(s, a) \right\}. \quad (1)$$

Here, s is the current state, and s' is the state after the transition by an action a . a' is the action taken in the state after the transition. α ($0 < \alpha \leq 1$) is the learning rate, γ ($0 < \gamma \leq 1$) is the discount rate, and r is the reward obtained after the transition. Various methods exist for action selection from the policy; however, the soft-max method based on the Boltzmann distribution is used in this paper, allowing for probabilistic action selection. Equation for the action selection method will be presented following.

$$P(s|a) = \frac{\exp\{Q(s, a)/T\}}{\sum_b \exp\{Q(s, b)/T\}}. \quad (2)$$

T is a parameter known as the temperature constant, which is used to adjust the ease of selecting actions with high action values.

2.2. Transfer Learning

Transfer learning is a method for reusing knowledge such as policy, action-value function and so on, obtained from reinforcement learning in new environments, and transferring learning from reinforcement learning is referred to as transfer reinforcement learning. Reinforcement learning tends to take time in the early stages due to random actions, making it difficult to execute multiple times. By using transfer reinforcement learning, it is possible to shorten the learning time. It is believed that knowledge reuse reduces the initial random actions in new environments, enabling more efficient learning. Additionally, Inter-task Mapping has been proposed as a method to describe the correspondence between agents with different structures, environments, and actions [16]. Equation for transfer reinforcement learning is presented based on the [14].

$$Q_c(s, a) = Q_t(s, a) + Q_s\{\chi(s), \chi(a)\}. \quad (3)$$

where $Q_c(s, a)$ is the integrated policy used by the agent for action selection in the target task, and $Q_t(s, a)$ is the policy for recording newly acquired action values. $Q_s\{\chi(s), \chi(a)\}$ is a policy that has been learned in advance from source task. $\chi(\cdot)$ is a function called inter-task mapping, and $\chi(s)$ and $\chi(a)$ indicate the mapping states and actions respectively [16]. For example, inter-task mapping is defined as $\chi : S_t \rightarrow S_s$, it means that the elemental s_t of the target task's set S_t is mapped to elemental S_s of the set S_s in source task. In some studies, $Q_c(s, a)$ is used as the initial value for the target task.

2.3. Transfer Rate

Kono et al. proposed a transfer rate adjustment method based on Takano et al. approach [9][12]. Transferring method with transfer rate can be defined as

$$Q_c(s, a) \leftarrow Q_t(s, a) + \tau Q_s\{\chi(s), \chi(a)\}. \quad (4)$$

The transfer rate $\tau(0 \leq \tau \leq 1)$ discounts the behavioral value of the reutilization measure, allowing us to adjust the effect of the transfer [14]. Kono et al. proposed a variable transfer rate automatic adjustment method using a sigmoid function.

$$\tau = \frac{1}{1 + e^{-\sigma g}}, \quad (5)$$

$$\sigma \leftarrow \sigma + \Delta d, \quad (6)$$

$$\Delta d = \begin{cases} +d & \text{if } s_t \neq s_{t+1}, \\ -d & \text{if } s_t = s_{t+1}. \end{cases} \quad (7)$$

Here, $\sigma(0 \leq \sigma \leq 1)$ is the input value of the sigmoid function, and the gain g is determined the shape of gradient with an arbitrary value that is adjusted by a human experiences. When the determine the transfer rate τ , σ is adjusted by learning agent's behavior using step size d . The transfer method is reformulated using the sigmoid function as defined by

$$Q_c(s, a) \leftarrow Q_t(s, a) + \frac{1}{1 + e^{-\sigma g}} Q_s\{\chi(s), \chi(a)\}. \quad (8)$$

In transfer reinforcement learning, existing methods required humans to manually adjust the transfer rate, and transfer reinforcement learning was executed with a fixed transfer rate. This led to situations where it became like over learning, or the transfer rate was underestimated too low, making it impossible to verify the three major effects (learning speed improvement, asymptotic improvement, jumpstart improvement), which are the evaluation indicators of transfer reinforcement learning [8]. As a result, humans had to readjust the transfer rate and re-execute transfer reinforcement learning. Kono et al. automated the transfer rate adjustment process in transfer reinforcement learning by determining whether there were environmental changes, Δd , at each step. They used a constant d to decrease the transfer rate if there was no change state by an action and to increase it if there was a change. However, while the method showed the effects of some transfer reinforcement learning case, the convergence speed was slow, and the results indicated that it did not converge in the later stages of additional learning.

3. PROPOSED METHOD

In Kono et al. method, the transfer rate were controlled using the step size d which was implemented in a sigmoid function. In this case, the fluctuation of transfer rate τ depends on only the shape on the function in both the increase and decrease. The main idea of the proposed method is that when adaptation to the environment is required, the transfer rate value is rapidly decreased, and when learning continues from there or when reuse of the action value function is effective, the transfer rate is slowly increased. Therefore, the following equation will be presented for controlling transfer rate τ .

$$\Delta d = \begin{cases} +d_1 & \text{if } s_t \neq s_{t+1}, \\ -d_2 & \text{if } s_t = s_{t+1}. \end{cases} \quad (9)$$

Proposed method is prepared two constants, d_1 and d_2 , using different parameters $d_1 \neq d_2$ for the increase and decrease. By appropriately setting the constants d_1 and d_2 , improvements in learning efficiency and accuracy are expected.

4. EXPERIMENT

This experiment aims to verify the proposed method and conduct tests under different conditions to determine appropriate parameter settings. Q-learning is adopted as reinforcement learning to acquire the action value function, and the obtained action value function will be applied to execute transfer reinforcement learning, along with existing studies and the proposed method, and the results will be compared for evaluation. The learning environment will be a 10×10 grid world focused on the shortest path problem. For the reward conditions, in Q-learning, a positive reward r_p is given upon reaching the goal, while a negative reward r_n is given for actions that result in a collision with wall or obstacles. In transfer learning, a positive reward is similarly awarded when reaching the goal. Regarding the negative reward, it will be provided in the case of a fixed transfer rate, but since learning can occur without it in existing studies and the proposed method, it will not be given. When the agent moves in the grid world, one move or one action resulting in a collision with a wall will be considered one step. The episode will be considered complete when the agent either reaches the goal or takes 10,000 steps, and a total of 300 episodes will be conducted. Three types of tasks will be prepared, and by evaluating them, the effectiveness of the proposed method will be verified to be independent of the environment. Each of the three types of tasks involves blocking the shortest path of the Source Task in the Target Task. The evaluation will be conducted using the number of steps per episode and the area of the learning curve representing the total number of steps. Learning curve consists number of steps until reaching the goal of the agent in each episodes. Number of steps in each episodes is described e_i , total number of steps is defined $\sum_i e_i$. The set of states \mathbf{S} that the agent observes and the set of actions \mathbf{A} that can be executed during learning are described in Equation (10) to Equation (11). The three types of tasks are illustrated in the Figure 1 through Figure 3. Additionally, the parameter settings used in the experiment are shown in Table 1.

$$\mathbf{S} = \{s_1, s_2, s_3, \dots, s_i\} \quad (10)$$

$$s_i = (x, y) \quad (11)$$

$$\mathbf{A} = \{forward, backward, right, left\} \quad (12)$$

Table 1. Hyperparameters used in the experiments for source task and target task

Parameter	Value
Learning rate α	0.1
Discount rate γ	0.99
Boltzmann constant T	0.05
Positive reward r_p	1.0
Negative reward r_n	-0.5
Total number of episodes	300
Trial number of trials	100
Gain value g of sigmoid function	5

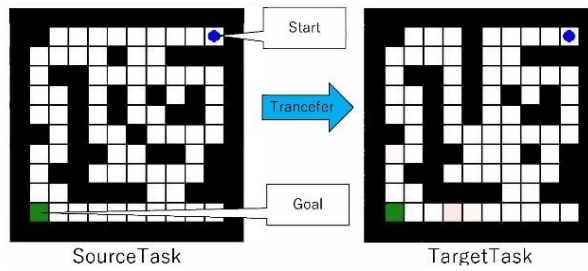


Figure 1. Grid world set up of shortest path problem as named TASK1

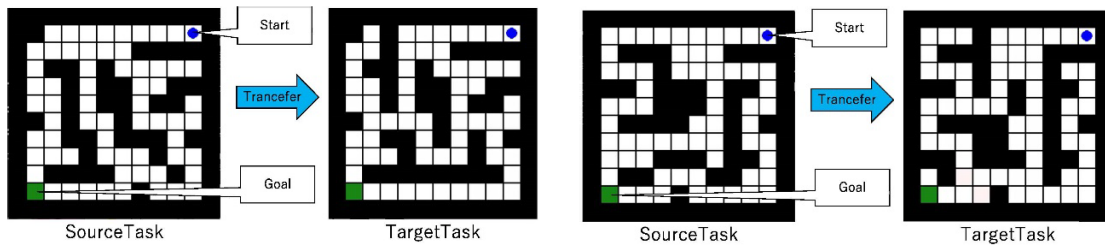


Figure 2. Grid world set up of shortest path problem TASK2 Figure 3. Grid world set up of shortest path problem TASK3

4.1. Determining Constant d Value for Previous Method

4.1.1. Conditions

The existing method of Kono *et al.*, which uses a sigmoid function to automatically adjust the transfer rate, is used to change the value of the step size parameter d to increase/decrease in order to decide a more effective transfer rate. In this experiment, the sigmoid function in the existing method is evaluated within the following range:

$$d = \{\pm 0.005, \pm 0.01, \pm 0.02, \pm 0.03, \pm 0.04, \pm 0.05, \pm 0.06, \pm 0.07, \pm 0.08, \pm 0.09, \pm 0.1, \pm 0.2, \pm 0.3, \pm 0.4\}.$$

Additionally, transfer reinforcement learning is conducted with a fixed transfer rate of $\tau = 0.1$ for comparison purposes. This constant transfer rate is determined by human experience and trial-and-error in the simulation. Moreover, reinforcement learning without transfer is also conducted for comparison. In this condition, the agent haven't reusing action value function. In other words, this case is reinforcement learning in the target task.

4.1.2. Results

The comparison results are shown in Figure 4 to Figure 6. Average and standard deviation is calculated with 10 trials data of end of obtained learning curve. The results showed that when $d = 0.04$ and $d = 0.03$ for Task 2 and Task 3, respectively, the results were close to the RL result. In the case of TASK1, RL result is high performance compared with all of d conditions. Therefore, when the positive transfer can be emerged, there was a possibility that Kono *et al.* method using a sigmoid function was superior to reinforcement learning, the average total number of steps was lower when $d = 0.05$ to 0.1 were used than when reinforcement learning was used. However, the Kono *et al.* method does not reach the optimal fixed transfer rate which is hand tuned, when a function is used to auto adjustment. Figure 7 and Figure 8 below shows the transition of the number of steps and transfer rate during the adjustment of TASK1 at the

beginning and at the end of the learning process. In Figure 7 and Figure 8, the transfer rate trend is reproduced as in the literature of Kono et al. It can be seen that the transfer rate is not maintained at 1 all the time, but increases and decreases even at the end of the learning period.

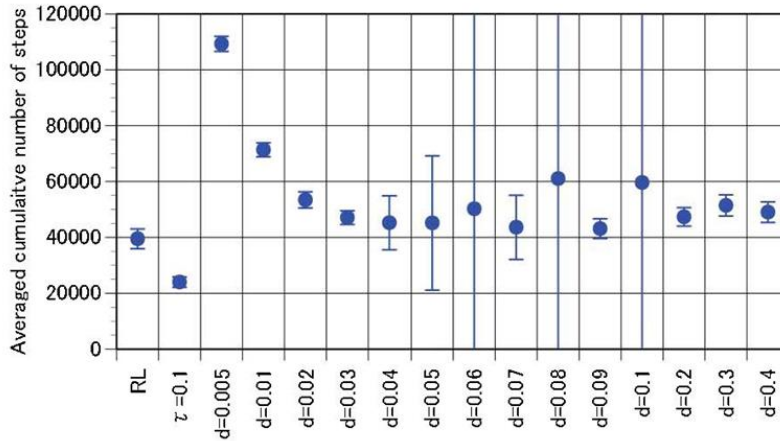


Figure 4. Number of steps in end of the learning curve on TASK1 condition

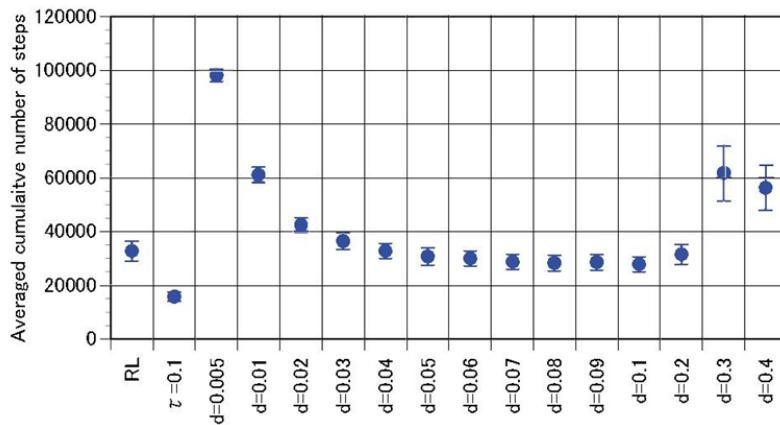


Figure 5. Number of steps in end of the learning curve on TASK2 condition

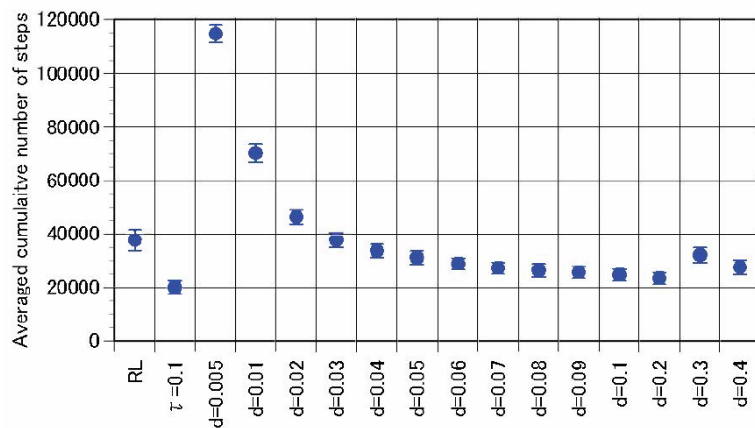


Figure 6. Number of steps in end of the learning curve on TASK3 condition

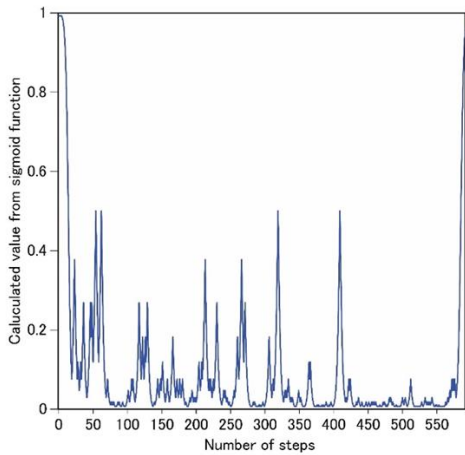


Figure 7. Transition of the value output from the sigmoid function as TASK1(episode no. 1)

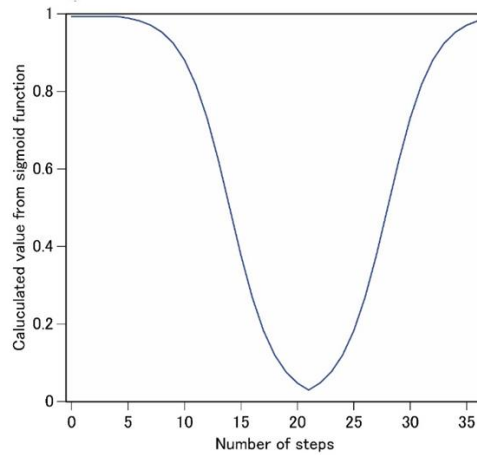


Figure 8. Transition of the value output from the sigmoid function as TASK1(episode no. 300)

4.2. Determining d_1 and d_2 Values in Proposed Method

4.2.1. Conditions

This experimental condition is conducted to investigate how the average total number of steps changes when two values $d_1 \neq d_2$ for the increase/decrease are prepared in the proposed method. It is possible to obtain better performance from previous method by changing the rate of increase or decrease of the transfer rate. Based on the results of above mentioned experiment, the value d_1 and d_2 will be varied in the range of $0.05 \leq d_i \leq 0.3$. The action value function of the source task is used for the transfer was the same as in above experiment.

4.2.2. Results

Figures 9 show the total number of steps in the results. The results show that setting a relatively large value for d_2 relative to d_1 tends to reduce the total number of steps. Therefore, it can be considered that a setting of $d_1 < d_2$ is sufficient to lower the total number of steps.

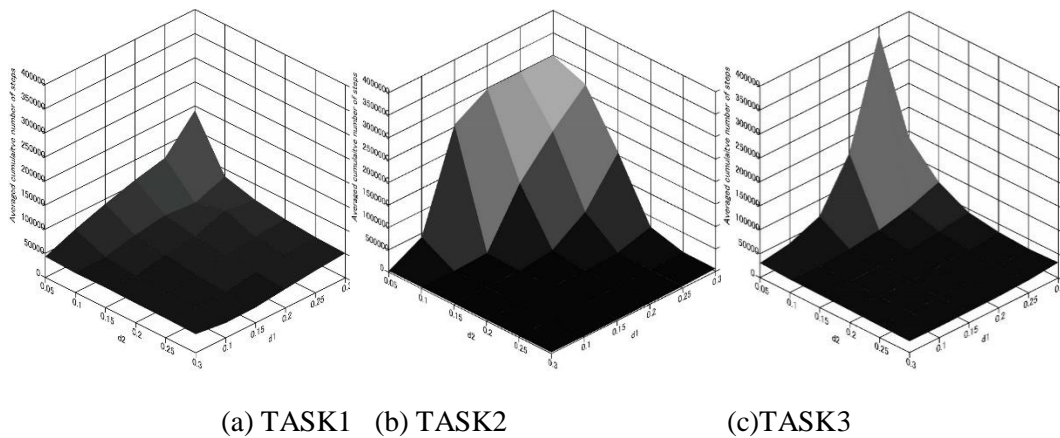


Figure 9. Change in the number of steps required for convergence due to changes in d_1 and d_2

4.2.3. Discussions

Based on the results obtained from Experiment 2, the parameters of the proposed method were set to $d_1 = 0.1$ and $d_2 = 0.3$ and compared with Kono et al. existing method and transfer reinforcement learning with a fixed transfer rate. The results are shown in Figures 10. From left to right, the plots show RL, the optimal fixed transfer rate, the existing method of Kono et al. and the proposed method, in that order. Since the lower the average total number of steps, the shorter the learning time, the graphs in Figures. Figure 10 show that the proposed method with $d_1 \neq d_2$ has a higher learning time reduction and better performance in adapting to the environment when transfer learning than the existing method. However, compared to the case of a fixed transfer rate, the proposed method still yielded results that were approximately 5000 steps worse. The objective of the method proposed in this paper was to improve environmental adaptation performance by preparing two parameters, d_1 and d_2 , for transfer rate adjustment, which was originally a single parameter d , in order to suppress fluctuations in the transfer rate during the later stages of learning. While this objective was achieved, the process of automatic adjustment of the transfer rate takes time, and there is potential for reducing the adjustment time by making d_1 and d_2 variable constants.

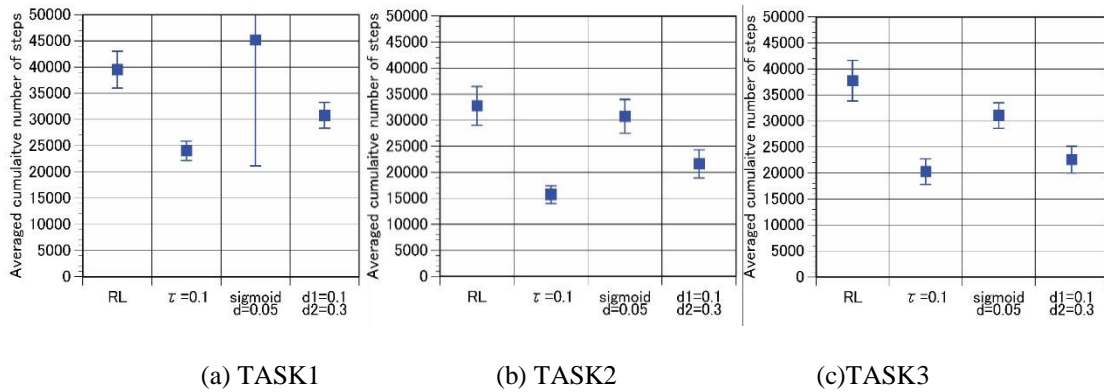


Figure 10. Comparison of the number of steps at the end of the learning curve with conventional methods

5. CONCLUSION

This paper is focused on transfer reinforcement learning, which is regarded as useful for autonomous systems such as autonomous driving systems, and is proposed to enhance environmental adaptation performance and reduce learning time compared to existing methods by automatically adjusting the transfer rate with different step sizes. Basic experiments were conducted to compare the effect of transfer learning using a grid world for the comparison of different conditions, demonstrating that the proposed method may be more significant than existing methods. However, experiments have not been conducted in real environments, and the effectiveness of the proposed method has only been verified in a grid world. When conducting experiments in real environments, it is also necessary to verify the method's performance under environmental changes, such as variations in the scale and configuration of the agent. Furthermore, the sigmoid function used in both the existing research and the proposed method is not always optimal, making it essential to conduct verification using other functions as well.

REFERENCES

- [1] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixão, F. Mutz, L. de P. Veronese, T. Oliveira-Santos, A. F. De Souza “Self-driving cars: A survey.”, *Expert systems with applications* vol.165, 2021.
- [2] A. Swief and M. El-Habrouk, “A survey of Automotive Driving Assistance Systems technologies,” In *Proceedings of the 2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, pp.1-12, 2018.
- [3] B. R. Kiran et al., “Deep Reinforcement Learning for Autonomous Driving: A Survey,” in *IEEE Transactions on Intelligent Transportation Systems*, vol.23, no.6, pp.4909-4926, June 2022, doi: 10.1109/TITS.2021.3054625.
- [4] J. Wang and Y.Chen, “Introduction to Transfer Learning Algorithms and Practice”, *Machine Learning: Foundations, Methodologies, and Applications*, Springer Nature Singapore, 2023.
- [5] R. Razavi-Far, B. Wang, M. E. Taylor, and Q. Yang, “Federated and Transfer Learnig”, *Springer Nature Sqtzlerl and*, 2023.
- [6] R. S. Sutton and A. G. Barto (1998). *Reinforcement learning: An introduction*. MIT press.
- [7] M. E. Taylor and P. Stone (2009). “Transfer learning for reinforcement learning domains: A survey.” *Journal of Machine Learning Research* vol.10 (Jul): 1633-1685.
- [8] A. Lazaric, “Transfer in Reinforcement Learning: a Framework and a Survey,” *Reinforcement Learning - State of the art*, vol.12, Springer, pp.143-173, 2012.
- [9] H. Kono, A. Kamimura, K. Tomita, Y. Murata, and T. Suzuki (2014)“Transfer Learning Method Using Ontology for Heterogeneous Multiagent Reinforcement Learning.” *International Journal of Advanced Computer Science and Application* vol.5, no.10, pp.156-164.
- [10] K. Khetarpal, M. Riemer, I Rish, and D. Precup, “Towards Continual Reinforcement Learning: A Review and Perspectives”, *Journal of Artificial Intelligence Research*, vol.75, no.2022, pp.1401-1476, 2022.
- [11] D. Abel, A. Barreto, B. V. Roy, D. Precup, H. van Hasselt, and S. Singh, “A definition of continual reinforcement learning”, in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp.50377-50407, 2023.
- [12] T. Takano, H. Takase, H. Kawanaka, H. Kita, T. Hayashi and S.Tsuruoka (2011). “Transfer Learning based on Forbidden Rule Set in Actor-Critic Method.” *International Journal of Innovative Computing, Information and Control* vol.7, no.5(B).
- [13] C. E. Osgood. “The similarity paradox in human learning: A resolution”, *Psychological Review*, vol.56, no.3, pp.132-143. 1949.
- [14] H. Kono, Y. Sakamoto, Y. Ji, and H. Fujii, “Automatic Transfer Rate Adjustment for Transfer Reinforcement Learning”, *International Journal of Artificial Intelligence & Applications*, vol.11, no.5/6, pp.47-54, 2020.
- [15] C. J. C. H. Watkins and P. Dayan.“Q-learning,” *Machine Learning* 8, pp. 279-292, 1992.
- [16] M. E. Taylor, P. Stone, and Y. Liu. “Transfer learning via inter task mappings for temporal difference learning,” *Journal of Machine Learning Research*, vol.8, no.1, pp.2125-2167, 2007.