A Basic Study on Indicator of Transfer Learning for Reinforcement Learning

Satoshi Sugikawa
Osaka Institute of Technology, 1-79-1 Kitayama, Hirakata City, Osaka, 573-0196 Japan
Email: satoshi.sugikawa@oit.ac.jp

Kenta Takeoka
Osaka Institute of Technology, 1-79-1 Kitayama, Hirakata City, Osaka, 573-0196 Japan

Naoki Kotani
Osaka Institute of Technology, 1-79-1 Kitayama, Hirakata City, Osaka, 573-0196 Japan

Abstract

Reinforcement learning requires a lot of learning time for the agent to learn. Transfer learning is a method to reduce this learning time, but it has the problem that the user does not know which knowledge is effective in which environment until it is learned. Therefore, it is necessary for the user to consider the relationship between the source and destination when transferring knowledge. Therefore, this study proposes Indicator of adaptation criteria that can determine this relationship in advance. In simulations, we demonstrate the usefulness of the proposed method by using some example problems.

Keywords: Reinforcement learning, Transfer learning, Maze problems

1. Introduction

In recent years, machine learning has achieved rapid growth with results in various fields such as natural language processing and image processing. Reinforcement learning [1], [2] which is capable of self-learning, is expected to further develop in all fields in the future. Reinforcement learning has the problem that it requires a lot of learning time because the agent learns from scratch in a new environment by trial and error.

Many studies have been conducted to reduce learning time, one of which is transfer learning. Transfer learning is a method of adapting and reusing previously learned material from similar tasks for a new task. Since there is no need to relearn, learning time can be reduced. However, the effectiveness of the transferred material is not known until it is transferred and learned. Therefore, when transfer learning is performed, the user needs to consider the relationship between the transferee and the transferee source, but even then, there is a possibility that the learning will not be successful and negative transfers will occur [3]. Therefore, this study proposes several indicators of adaptive criteria that can discriminate in advance the validity of knowledge.

2. Reinforcement learning

Reinforcement learning is a branch of machine learning in which an agent learns through interaction with its environment. Reinforcement learning is characterized by the fact that the agent chooses an action to achieve a certain goal and receives a reward signal for that action as it learns. Optimal behaviour rules are learnt by repeating steps 1 to 4 below.

1. Agent observes the state
2. Decides on a course of action based on cues from measures
3. Memorizes experience of which states, which actions and which rewards
4. Seeks measures based on experience

Reinforcement learning is also formulated as a Markov decision process, expressed as M=(S,A,P,R) as follows.

- S: Set of states
- A: Set of actions
- P: State transition $p(s_t = s' | s_t = s, a_t = a)$
- R: Behaviour rules $\pi(s, a) = R(s_t = s, a_t = a)$

Reinforcement learning aims to acquire behaviour rules that maximize the expected reward.

2.1. *Q-learning*

In this study, Q-learning [3] was used as a reinforcement learning algorithm learning updates the Q-values that the agent associates with a combination of states and actions, and finds an action rule that selects the action with the maximum Q-value in each state. The formula for updating the Q-values is as follows.

$$Q(s_t, a_t) = Q(s_t, a_t) + \eta * \left(R_{t+1} + \gamma \max Q\,(s_{t+1}, a) - Q(s_t, a_t)\right)$$

In this study, the ϵ-greedy method was also used as the agent's action selection method. In the ϵ -greedy method, the agent acts randomly with a probability of ϵ and selects an action with a probability of 1- ϵ with a maximum Q value.

3. Transfer learning

In transfer learning [4], [5], the goal is to reduce learning time by transferring knowledge learned at the transfer source as prior knowledge at the transfer destination with similar tasks. However, if there is no similarity between the source and destination, a negative transfer may occur. Therefore, the user must consider the similarity between the source and destination in advance. Therefore, it is necessary to determine the similarity between the source and destination to determine which knowledge is transferred to which destination.

In this study, we aim to formalize the adaptation criterion by determining the similarity between mazes using the maze problem.

3.1. *learning model*

The algorithm used for learning in reinforcement learning is Q-learning. There are five actions that the agent can take: up, down, left, right, left and right, and no action. The reward design is +1 if the agent reaches the goal, and -0.01 if it collides with a wall or a step elapse.

3.2. *transition method*

As the transfer method uses Q-learning, the transfer learning adopts the value function transfer type.
The Q-table obtained by the learning of the transfer source is transferred to the transfer destination agent.
The degree of re-use of the behavioural value function is adjusted using the transfer rate τ. The transfer rate τ is adjusted in the range of $0 < \tau < 1$. In this study, τ was set at 0.5.

$$Q^c(s,a) = Q^t(s,a) + \tau Q^s(s,a)$$

4. Similarity Indicators

The following five indices are used to compute the similarity.

4.1. *Similarity by pixel value*

In this calculation example, the similarity by pixel value is determined. As the mazes used in this study have the same size, the percentage of pixel values matched between the images is determined.

4.2. *Cosine similarity between maze sequences*

In this example calculation, the maze is transformed into arrays and the cosine similarity between the source and destination arrays is calculated. The transformed array is the same as the array used for training, with the maze walls set to 1 and the paths set to 0.

- s is a vectorization of the base maze
- t is a vectorization of the destination maze.

$$cos(s,t) = \frac{\sum_i s_i\, t_i}{\sqrt{\sum_i s_i^2}\sqrt{\sum_i t_i^2}}$$

4.3. *Similarity using maxQ as weights*

In this calculation example, maxQ is used to determine the similarity. The total of the differences between the mazes is calculated using maxQ of maze A, which is the source of the transition, as the weight. The sum of the differences between the mazes is calculated using maxQ of maze A as the weight, and divided by the total value of maxQ.

$$\mathrm{p} = 1 - \frac{\sum_{\mathrm{i}}(\mathrm{s_i} - \mathrm{t_i})\max_{\mathrm{a}}\mathrm{Q_{s_{i,a}}}}{\sum_{\mathrm{i}}\max_{\mathrm{a}}\mathrm{Q_{s_{i,a}}}}$$

4.4. *Cosine similarity using maxQ*

Finally, in this calculation example, the Q values obtained after learning are used to obtain the cosine similarity. The Q values obtained after normal learning in each maze $(12 \times 12 \times 5)$ are changed to an array of $12 \times 12$by finding the maximum value (maxQ) in each square. However, this is not known before application as it is an outcome. It is calculated as part of the evaluation for measuring similarity.

$$\cos(\mathrm{Max_aQ_s}, \mathrm{Max_aQ_t}) = \frac{\sum_{\mathrm{i}}\mathrm{Max_aQ_{s_{i,a}}Max_aQ_{t_{i,a}}}}{\sqrt{\sum_{\mathrm{i}}\mathrm{Max_aQ^2_{s_{i,a}}}}\sqrt{\sum_{\mathrm{i}}\mathrm{Max_aQ^2_{t_{i,a}}}}}$$

5. Simulation

Five mazes were prepared to test the usefulness of the index. Maze A is the original maze and was applied to mazes B, C, D, and E, respectively. Fig. 1, Fig. 2, Fig. 3, Fig. 4, and Fig. 5 show the mazes.

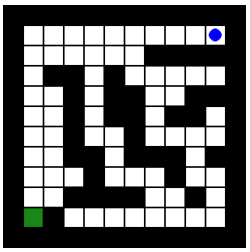

Fig.1 Maze A

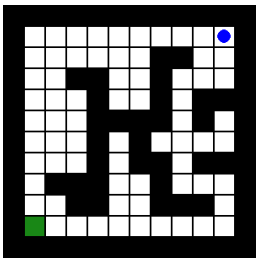

Fig.2 Maze B

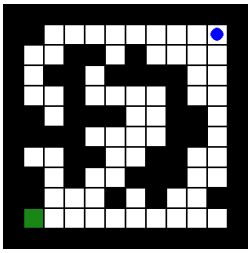

Fig.3. Maze C

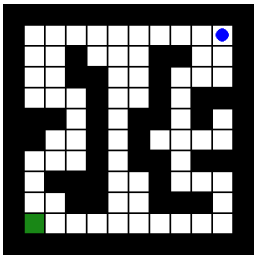

Fig.4 Maze D

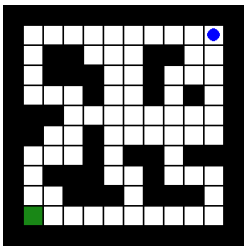

Fig.5 Maze E

The similarity of the mazes to each indicator is shown in Table 1.

Table 1 Similarity Results

	Maze A to Maze B	Maze A to Maze C	Maze A to Maze D	Maze A to Maze E
Similarity by pixel value	82.95	77.08	80.02	79.44
Similarity between maze sequences	80.29	75.96	78.71	77.66
Similarity using Qmax	98.86	60.96	82.49	77.37
Cosine similarity using Qmax	72.97	5.89	41.17	28.91

S The Similarity results show that maze B, maze D, and maze C have the highest similarity in that order. Similarity by pixel value, similarity between maze sequences, and similarity using Qmax, in that order. The use of Q-values allows us to focus on differences only in important locations. Fig. 7, Fig. 8, and Fig. 9 show results of the transfer learning. Transition from maze A to maze C is not shown in the figure because the goal was not achieved. Reinforcement learning is represented by the red line and transfer learning by the blue line.

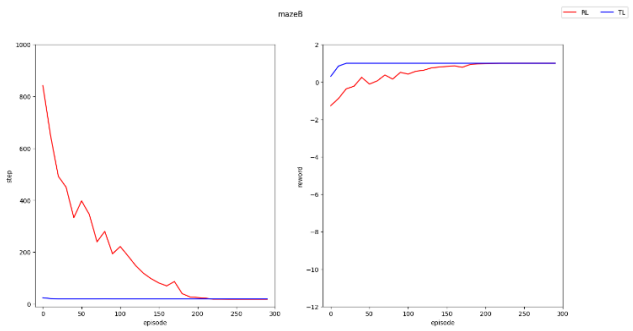

Fig.7. Transfer learning from maze A to maze B, left panel shows number of steps to goal and episodes, right panel shows total reward and number of episodes.

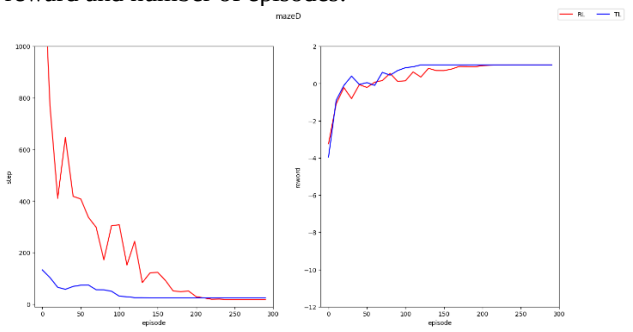

Fig.8. Transfer learning from maze A to maze D, left panel shows number of steps to goal and episodes, right panel shows total reward and number of episodes

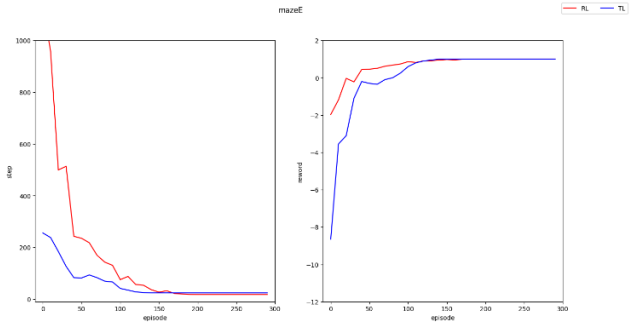

Fig.9. Transfer learning from maze A to maze E, left panel shows number of steps to goal and episodes, right panel shows total reward and number of episodes

From these figures, the goal is reached faster with transfer learning than with reinforcement learning. This indicates that transition learning is effective in these mazes. Based on the results in the figures and the cosine similarity using Qmax, it seems reasonable to focus on the Q value.

Fig. 10 shows the results of applying reinforcement learning to the mazes. Fig. 11, Fig. 12, Fig. 13 and Fig. 14 show the results of applying transition learning to each maze. Red indicates locations with high rewards. The reward of maze A was applied to each maze. Therefore, the maze was routed from the top towards the goal. This demonstrates the validity of this simulation.

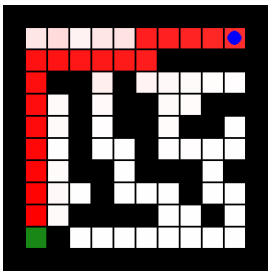

Fig.10 Result of Reinforcement Learning on maze A

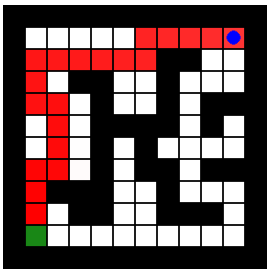

Fig.11 Result of Transfer learning on maze B

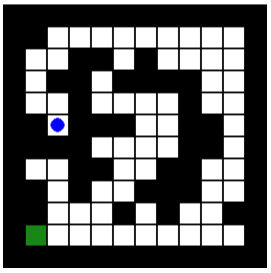

Fig.12 Result of Transfer learning on maze C

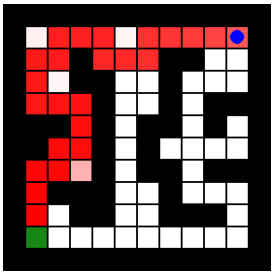

Fig.13 Result of Transfer learning on maze D

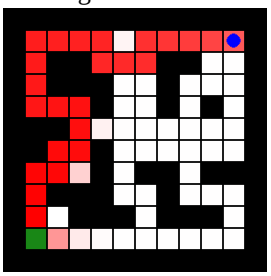

Fig.14 Result of Transfer learning on maze E

6. Conclusion

In this study, we conducted basic research on measures focusing on environmental similarity in reinforcement learning. Simulation results confirmed the validity of focusing on maxQ. The validity of the indicator was also confirmed to some extent. In the future, the creation of more accurate indicators is required.

References

1. Sutton, R, S., Barto, A, G: Reinforcement Learning: An Introduction. The MIT Pres（2007）
2. Leslie Pack Kaelbling. Michael L. Littman and Andrew W. Moore: Reinforcement Learning-A Survey. Journal of Artificial Intelligence Research.vol.4.237/285（1996）
3. C. J. C. H. Watkins: Learning from Delayed Rewards, PhD thesis, King's College, Cambridge, UK（1989）
4. Haitham B. Ammar, Karl Tuyls, Matthew E. Taylor, Kurt Driessens, Gerhard Weiss: Reinforcement Learning Transfer via Sparse Coding, Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems（2012）
5. Matthew E. Taylor, Peter Stone: Transfer Learning for Reinforcement Learning Domains, A Survey ,Journal of Machine Learning Research, vol.10,1633/1685（2009）

Authors Introduction

Dr. Satoshi Sugikawa

He received his Dr. Eng. degrees from Kobe University, Japan, in 2011. In 2013, he joined Osaka Institute of Technology, where he is currently an Assistant Professor. His research interests include Mathematical Optimization. He is a member of IEEJ

Mr. Kenta Takeoka

He received his Bachelor's degree in Engineering in 2022 from the Faculty of Information Science and Technology, Osaka Institute of technology in Japan. He is currently a master student in Osaka Institute of Technology, Japan

Dr. Naoki Kotani

He received a Ph.D. degree from Osaka University, Osaka, Japan, in 2011. He is an assistant professor of Information Science and Technology at Osaka Institute of Technology. His research interests machine learning and robotics. Kotani is a member of the ISCIE, the SICE, the RSJ, the JSAI and hte IEEE.