Bulletin of Networking, Computing, Systems, and Software – www.bncss.org, ISSN 2186-5140
Volume 9, Number 1, pages 53–57, January 2020

Online optimization of AGV transport systems using deep reinforcement learning

Kei Takahashi
Engineering department,
The University of Electro-communications
Tokyo, Japan
t1833082@edu.cc.uec.ac.jp

Sogabe Tomah
i-PERC & Engineering department
The University of Electro-communications
Tokyo, Japan
sogabe@uec.ac.jp

Abstract—**In recent years, the distribution industry and the manufacturing industry have faced many challenges such as labor shortages and product diversification. For this reason, there is an attempt to automate the distribution and production process by using an automated guided vehicle (AGV) that can automatically carry a package to a predetermined place. However, in order to automate, there is a problem of how to optimize the movement path and transfer of AGV, and various studies have been conducted to solve it. In this paper, we propose a method to control multiple AGVs using deep reinforcement learning. To evaluate the deep reinforcement learning methodology, simulation experiments are performed to first train one model and then use the learned network to optimize another model. And simulation results show that the proposed method learns optimal or near-optimal solutions from past experience and provides superior performance in new environments.**

Index Terms—**Reinforcement Learning, Online Optimization, DQN, Transport systems, AGV**

I. Introduction

In recent years, automated guided vehicles (AGVs) that can automatically transport packages to specified locations are widely used in transport systems in factories and automated warehouses. In the construction of transport systems, automated guided vehicles have a great degree of freedom and play an important role in flexibly transporting materials and products [1]. In order for the system to increase the work efficiency, AGV decision making according to many situations is important. Such decision-making includes many items such as determination of transported goods, selection of destination, determination of route from current location to destination, and can be considered as a combinatorial optimization problem. However, when the production system becomes larger and more complex, such combinatorial optimization problems generally belong to NP complete problems and are almost difficult to solve strictly [2]. For this reason, optimal algorithms have not been established, and most of them are controlled by a centralized production system that operates according to a schedule determined from the experience of workers working in factories. Moreover, it is necessary to implement and verify AGV control logic every time the transport system is changed.

There are two approaches to such large-scale optimization problems: batch optimization and online optimization. Batch optimization is a technique for obtaining an optimal solution using all data obtained at the time of calculation. Examples of batch optimization algorithms include mathematical optimization such as Newton's method and linear programming. On the other hand, online optimization is a technique in which future information is unknown and every time new data is acquired, optimization is performed based on that information [3]. While batch optimization is highly accurate, every time new data is added, it is necessary to learn again using all the data, and there is a disadvantage that there is no expansion capability for a new environment. Online optimization using machine learning has the disadvantage that learning is unstable and does not converge well unless parameters are set well. However, the machine learning method may be able to optimize the unknown environment by discovering the model law based on the training data [4]. For this reason, online optimization has better generalization performance than batch optimization, which is optimized for only one environment.

In this paper, we propose a method to improve the efficiency of AGV transport system by online optimization using Deep Q Network [6] [7], which is one of the reinforcement learning methods. Train with one model as the first approach. Start with a random strategy and iteratively optimize the parameters of the neural network using the states obtained during training. Next, using the learned network, we try to test the optimization in the modified model. At the time of testing, inference in an unknown state is executed, and an action based on the inference is selected. These results provide insights on how to use deep reinforcement learning for combinatorial optimization problems, especially those where it is difficult to design heuristics.

II. Related works

Several methods using reinforcement learning have been proposed to obtain efficient control of AGV in the transport system. Reinforcement learning can learn AGV behavior by repeating trial and error. By using Q-learning as information such as AGV operation status, position, and distance to the destination, multiple AGVs can be controlled efficiently [8] [9]. However, when Q-learning is used to optimize a large-scale production system, the state and behavior increase, making it difficult to learn the optimal strategy [10]. To solve the problem, Kamoshida et al. [11] showed that by using Deep Q Learning, one of the basic methods of deep reinforcement learning, it is possible to learn appropriate control even when high-dimensional model information is used as input. In general, AGV mostly operates in cooperation with other machines, and AGV behavior affects each other. It is ideal to minimize transport time while preventing AGV traffic in the transport system, so ignoring this effect can cause inefficiencies. Therefore, Dong et al.

[12] showed that predicting future tasks of the AGV system and sending AGV to the predicted task start position would lead to improved transport efficiency. Thus, it has been shown that deep learning, which has been progressing rapidly in recent years, is effective in various prediction problems. Therefore, in this study, we consider that using the prediction performance for inference of unknown states in reinforcement learning leads to improvement of generalization performance in optimization problems.

III. EXPERIMENT ENVIRONMENT AND METHOD

A. *Learing Process*

The virtual transport system created by the production simulation software WITNESS (made by Lanner) is used for learning and testing the proposed method. Figure 1 shows the learning process.

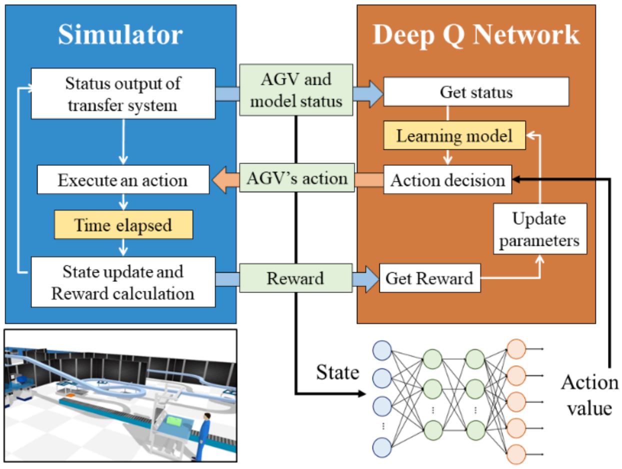

Fig.1. Overview of the learning process

The left side of the figure represents the simulator, and the right side represents the learning algorithm. First, the AGV and model states are obtained from the simulator, and the information is input to the network to determine the action. After that, the action is executed by the simulator, the obtained reward is input again to the algorithm, and the parameter is updated.

B. *AGV transport systems*

In this study, we assumed the AGV transport system shown in Fig. 2 and created a model using the production simulation software WITNESS (made by Lanner). The transport system will ship three types of packages sequentially. Set up a buffer that can store up to 10 packages at the loading point. A signal is sent when there is a cargo in the buffer, and the received idle AGV comes to load the load. The cargo is then transported to the unloading point by AGV.

It takes 6 seconds for the vehicle to load and unload the cargo. When another vehicle approaches a stopped vehicle for unloading, the approaching vehicle needs to be temporarily stopped to prevent a collision. This stop time is a factor that reduces the efficiency of the transport system. In this model, it is possible to select whether the vehicle goes to route A or route B in advance, and it is ideal to predict the congestion status of each path and select a route that can pass without stopping.

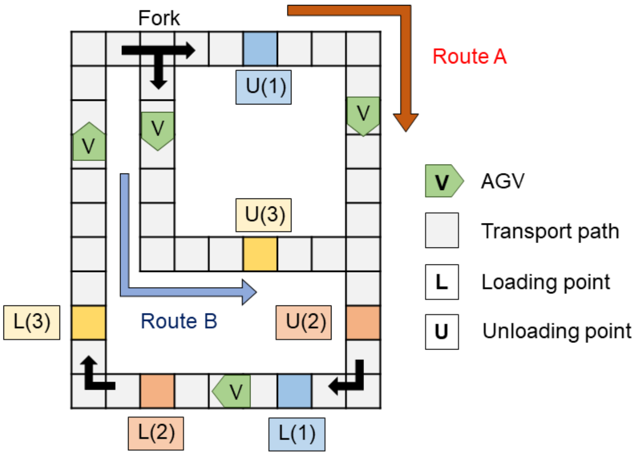

Fig.2. A Transport model with driving lanes for AGVs

C. *State variables*

Four AGVs circulate along the lane and carry the cargo. AGVs can only travel in one direction and cannot overlap on the lane. The upper limit speed of AVG is $5\,m/s$, the acceleration / deceleration is $1\,m/s^2$. Table 1 shows the data items that can be acquired from the simulation.

Table 1. Experimental data item

	Data Item
(a)	Vehicle state
(b)	Vehicle position
(c)	Vehicle destination
(d)	Cargo presence

(a) Vehicle state takes one of the following six states. (1: Idle, 2: Transfer request acceptance, 3: Blocking, 4: Transfer 5: Loading, 6: Unloading).

(b) Vehicle position divides the conveyance path into 23 sections, and indicates 1 if there is a vehicle in that section and 0 if there is no vehicle.

(c) Vehicle destination represents the current destination of each AGV and takes one of seven states: 3 loading points, 3 unloading points, or no destination.

(d) Cargo presence is 1 if there is a cargo at each loading point and 0 if there is no cargo.

D. *Action*

The action in this model is the allocation of cargo transport to AGV. Table 2 shows the possible actions of AGV and their definitions. When heading from LP2 to UP2, it is possible to select route A or route B. By avoiding a congested route, transport efficiency can be improved.

Table 2. List of action definition

Actions	Definition
1	Carry from L(1) to U(1)
2	Carry from L(2) to U(2) (taking route A)
3	Carry from L(2) to U(2) (taking route B)
4	Carry from L(3) to U(3)
0	Do not carry

E. *Reward*

The purpose of this experiment is to maximize the amount of luggage transported per hour through action selection. The cause of the decrease in the conveyance efficiency is a temporary suspension of AGV due to a mistake in the conveyance order or congestion of paths. Therefore, it can be considered that the smaller the number of AGV blockings, the better the conveyance efficiency. Therefore, in each step, the reward is -1 if the state of each AGV is the blocking state, and the reward is 0 if it is not the blocking state. Also, if the transport amount at time t is larger than the previous episode, the reward is set to 1.

F. *Algorithm*

The ε-greedy algorithm is adopted as a method for balancing the trade-off between accumulation of experience obtained from action results and utilization. DNN is used to evaluate the state and action.

```
Algorithm : optimization opretation
procedure Deep Q Learrning
    for Episode = 1 : n do
        reset simulator
        for t = 1 : T do
            get state from simulator
            $a_t = \begin{cases} \text{take random action} & \text{prob.}\varepsilon \\ argmin_a Q(s_t, a, \theta) & \text{otherwisw} \end{cases}$
            send and execute action
            observe rewaard $r_t$ and state $s_{t+1}$
            minibatch ($s_j$ , $a_j$ , $r_j$ , $s_{t+1}$ )
            $y_j = \begin{cases} r_j & \text{if episode terminates} \\ r_j + \min(Q(s, a, \theta)) & \text{otherwise} \end{cases}$
            loss function $(y_j - Q(s_j, a_j, \theta))^2$
        end for
    end for
end procedure
```

IV. EXPERIMENTAL RESULT

In this experiment, we verified the series of episodes from the time when the model shown in Fig. 2 started at T = 0 to the time when it stopped at T = t.

A. *Random Action*

First, table 3 shows the transport amount when an action is selected at random with a uniform probability when the model is operated up to T = 1000. In this model, even if an action is selected at random, **76.6** cargo can be transported on average.

Table 3. Transport amount in random action

Seed value	1	2	3	4	5	**Ave**
Trans Amount	79	77	73	79	75	**76.6**

B. *Domain knowledge*

Based on the researcher's domain knowledge, we examined the state of vehicles under the rules adjusted by conditional branching so that the vehicle does not jam. The figure shows how much each vehicle was actually suspended due to crowding when the model was run up to T = 1000 based on domain knowledge.

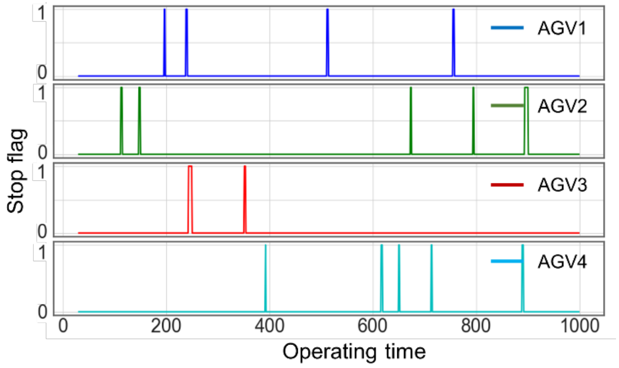

Fig.3. Each vehicle stopped due to traffic jam

In fig.3, the horizontal axis is the operation time, and the vertical axis is a vehicle stop flag indicating that the smooth conveyance state is when y = 0, and that the vehicle is temporarily stopped due to traffic congestion when y = 1. Figure 3 shows that the vehicle has been suspended more than 20 times due to traffic jams, and it can be seen that the control based on domain knowledge cannot always select the optimal action. Therefore, in order to confirm that the transport efficiency is improved by continuing to select paths that are not crowded, we created a reference model with a virtual path without an unloading point as shown in Fig.4.

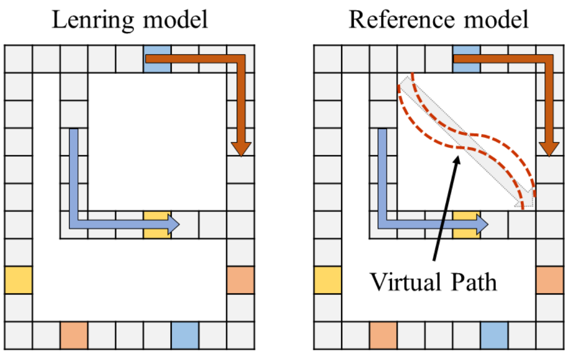

Fig.4. Learning model and reference model with virtual path

Table 4 shows the transport amount of the learning model and the reference model under control based on domain knowledge. From table 4, it was found that even if the same control rule is used, a final transport amount increases if a path without congestion is always passed.

Table 4. Transport amount based on domain knowledge

	Operating time	**Total amount**
Learning Model	1000	123
Reference Model	1000	125
Learning Model	2000	248
Reference Model	2000	254

C. *Deep Q Learning*

Using deep reinforcement learning which is a proposed method, we optimized T = 0 to T = 1000 as a series of episodes. The 103 state variables obtained from the simulator were input to the 4-layer DNN for learning. Fig. 5 shows the transport amount for each episode of DQN, and Fig. 6 shows the overall learning process.

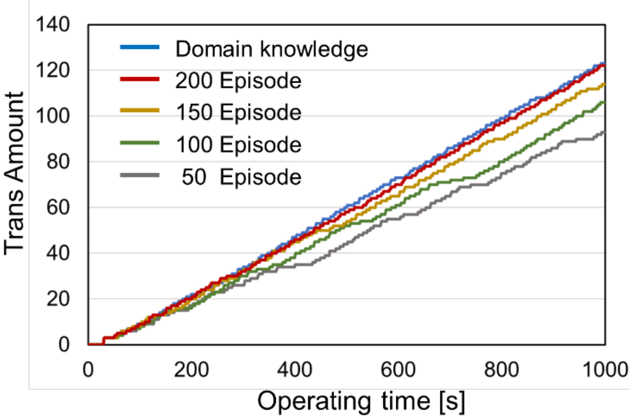

Fig.5. The transport amount for each episode

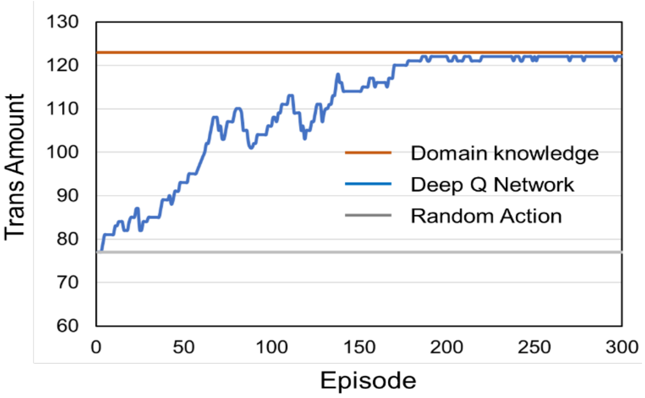

Fig.6. Learning process with DQN

Fig 5. shows the transport amount of domain knowledge and the transport amount for every 50 episodes of the DQN learning process. Each time the episode is repeated, the carrying amount increases, and it can be seen that learning of appropriate behavior is progressing. In addition, Fig.6 also shows that learning with DQN has converged in about 200 episodes. The convergence value of the conveyance amount by optimization using DQN is 122, which is slightly less than 123 when domain knowledge is used. However, since a sufficient transport amount has been achieved, it can be seen that the congestion situation on the branch path was predicted and an appropriate action was selected to some extent. From these results, it was shown that AGV control optimization by deep reinforcement learning is sufficiently possible.

D. *Reuse learned networks*

To confirm the generalization performance to the unknown environment by deep RL, we tried to control the model with modified Fig3 using the trained DNN. First, the trained DNN was applied to a model in which the number of AGVs was increased or decreased, that is, the number of agents was changed. The results are shown in Table 5.

Table 5. Transport amount according to the number of AGV

Number of AGV	AGV:3	AGV:4	AGV:5
Transport Amount	101	123	127

In the case of three AGVs, the number of AGVs was insufficient for the number of transportation demands, so the transportation amount was considered to have decreased. In addition, in the case of five AGVs, in addition to appropriate actions, it is considered that the upper limit of transportation demand was achieved by increasing the number of AGVs. It can be seen that even when the number of agents increases or decreases, deep reinforcement learning can flexibly select appropriate actions.

Next, the position of the transfer point is moved, and the generalization performance of the deep reinforcement learning for the model whose environmental information has changed is examined. Figure 7 shows the modified model.

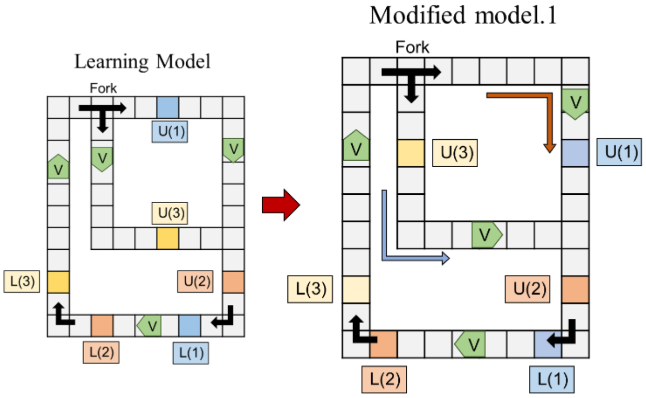

Fig.7. Learning model and modified model.1

In modified model.1, only the coordinates of Loading point (1), (2) and Unloading point (1), (3) were changed. Table 6 shows the results of applying trained DNN to this model.

Table 6. Transport amount in the modified model.1

	Operating time	Total amount
Learning Model	1000	123
Modified Model.1	1000	120

Compared to the learned model, the transport amount in the modified model decreased slightly. This is because L (3) and L (2), and L (1) and U (2) are close to each other, so the transfer of each other has an effect, and new blocking factors have increased. However, the results show that deep reinforcement learning can flexibly select appropriate actions even when the environment changes slightly.

Finally, a model with a significant change in the environment is created, and to what extent the deep reinforcement learning has generalization performance is examined. Fig. 8 shows the modified model.

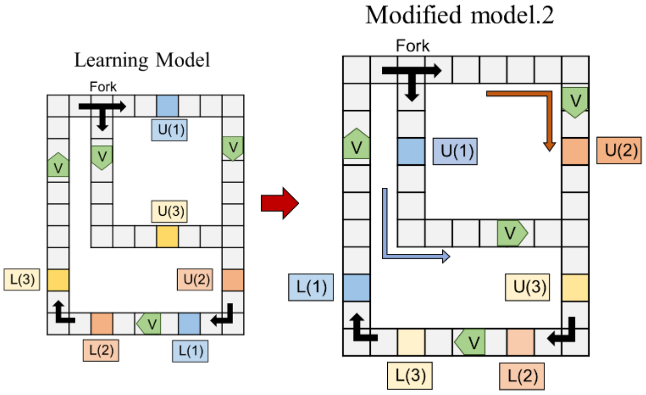

Fig.8. Learning model and modified model.2

In Modified model.2, the coordinates of all the transfer points were changed, and the order of arrangement was also changed, so that changes were also made to the order of cargo transport and the destination of the cart. Table 7 shows the results of applying trained DNN to this model.

Table 7. Transport amount in the modified model.2

	Operating time	Total amount
Learning Model	1000	123
Modified Model.2	1000	92

Compared to the learned model, the transport amount in the modified model.2 was significantly reduced. In the two models, the obtained state variables are the same, but since the priority order of transport is different from the learning model, it seems that even if the action was selected according to the learned policy, it did not work effectively.

As a result, it was confirmed that deep reinforcement learning has a generalization performance that can select an action by making a prediction even in an unknown state if learning is completed. However, if the environment of the model changes significantly and the learned strategy is not effective for the model, it is necessary to learn again.

A. Conclusion

In this paper, we proposed an optimization method by deep reinforcement learning for the control optimization problem of AGV transport system, and confirmed the effectiveness of the method by experiments using a transport simulator. In addition, it was confirmed that the approach has a generalization performance that can select an optimal action flexibly to some extent even when the environment changes by using the network once learned. The first problem in the future is that the transport amount by optimization using deep reinforcement learning is lower than the value by domain knowledge. It is desirable that it is at least equal to or greater than the value based on domain knowledge. To that end, reward settings can be reviewed and approaches other than deep q network can be considered. The second problem is that we have not been able to relatively evaluate how good the solution obtained by optimization by deep reinforcement learning. For this purpose, it is necessary to calculate the theoretical value of the transport amount by a mathematical optimization method and perform comparison.

REFERENCES

[1] L. Sabattini, V. Digani, C. Secchi, G. Cotena, D. Ronzoni, M. Foppoli, and F. Oleari, “Technological roadmap to boost the introduction of agvs in industrial applications,” in 2013 IEEE 9th International Conference on Intelligent Computer Communication and Processing (ICCP). IEEE, 2013, pp. 203–208.

[2] Murao H., Kitamura S., "Online scheduling of a multi-robot system by using genetic algorithms". IEEE International Symposium., vol.2, pp.709713,1998.

[3] Xiao, L., “Dual averaging methods for regularized stochastic learning and online optimization”, The journal of machine learning research volume 11 (2010), pp. 2543-2596

[4] Irwan Bello∗, Hieu Pham∗, Quoc V. Le, Mohammad Norouzi, Samy Bengio , “NEURAL COMBINATORIAL OPTIMIZATION WITH REINFORCEMENT LEARNING”, arXiv preprint arXiv:1611.09940, 2017

[5] R. S. Sutton and A. G. Barto, “Reinforcement Learning,” A Bradford Book, MIT Press, 1998.

[6] Mnih,V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. and Riedmiller, M., Playing Atari with deep reinforcement learning, arXiv preprint arXiv:1312.5602, 2013.

[7] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare,M. G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G.et al., Human-level control through deep reinforcement learning, Nature, 518[7540], 529-533, 2015

[8] Michiko Watanabe, Masashi Furukawa, Masahiro Kinoshita, Yukinori Kakazu, “Acquisition of Efficient Transportation Knowledge by Q-Learning for Multiple Autonomous AGVs and Their Transportation Simulation”, Journal of the Japan Society for Precision Engineering, Vol. 7, pp. 1609-1614, 2001

[9] Asahikawa National College of Technology Kenta MURATA, Michiko WATANABE and Masashi FURUKAWA, Autonomous Driving of Multiple AGVs by Reinforcement Learning, Journal of the Japan Society for Precision Engineering Spring competition, 2004, pp. 593-594

[10] A. Gosavi, “Reinforcement learning: A tutorial survey and recent advances,” INFORMS Journal on Computing, vol. 21, no. 2, pp. 178–192, 2009

[11] Ryota Kamoshida, Yoriko Kazama, “Acquisition of Automated Guided Vehicle Route Planning Policy Using Deep Reinforcement Learning”, IEEE International Conference on Advanced Logistics and Transport , 2017

[12] Dong Li, Bo Ouyang, Duanpo Wu, Yaonan Wang, “Artificial intelligence empowered multi-AGVs in manufacturing systems”, arXiv:1909.03373vl