Model Agnostic Meta-Learning for Efficient Federated Deep Reinforcement

Chae-Rim Han* Sun-Jin Lee and Il-Gu Lee

Sungshin Women's University, Seoul, Korea
{20200969, 220214013, iglee}@sungshin.ac.kr

Keyword : Data optimization, Deep neural networks, Deep reinforcement learning, Federated learning

1 Introduction

Reinforcement learning is a learning method that rewards through trial and error and finds rules inherent in the data[1]. Agents determine behaviors based on the state values and policies of their surroundings, receive rewards accordingly, and learn policies through backward shaping to maximize the rewards. However, because this method is optimized for a simulation environment, there is a simulation of a real-world(sim2real) problem that cannot be learned properly in a real environment[2]. This is because only some agents are provided with a reward for the state action, and the rest learn with an observed state without a reward. Recently, research has been conducted on single-agent reinforcement learning that adds noise to the DNN reward; however, if the estimated value is greater than the maximum value, a maximum bias problem arises[3]. In other words, conventional studies do not solve the memory overload problem caused by an increase in the size of the learning data and queue matrix and cannot specify the exact q-value.

2 Proposed Model

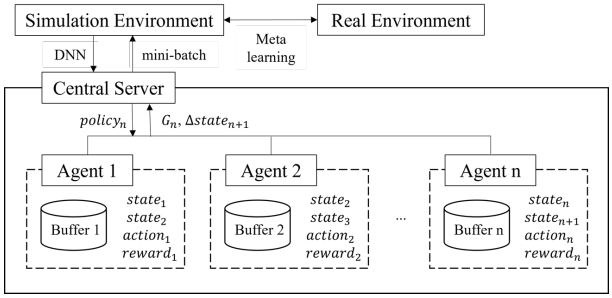

Figure 1: Architecture of Model Agnostic Meta-Learning for Efficient Federated Deep Reinforcement

The federated deep RL framework based on agnostic learning proposed in this study is illustrated in Figure 1. To improve the accuracy of the model, we used a meta-learning method for distributed learning by creating a mini-batch composed of states with random noise. For fine-tuning, the proposed model determines and learns an initial weight that minimizes the sum of the losses calculated in each minibatch(MAML).

Even if the q value is estimated with a value that deviates from the distribution, overfitting can be prevented by removing the loss value with a large compensation variance. To minimize memory usage, the proposed model performs learning such that the n^{th} agent selects the nth $+1^{th}$ Q value when determining the action.

The 7th International Conference on Mobile Internet Security (MobiSec'23), Dec. 19-21, 2023, Okinawa, Japan, Article No.P-7

* Corresponding author: Sungshin Women's University, Seoul, 02844, South Korea, Email: 20200969@sungshin.ac.kr

3 Results and Conclusions

In this study, performance was evaluated by comparing the accuracy, memory usage, and latency of the proposed framework with those of the TD method(TD). The TD is a combination of reward shaping and policy transfer used for the target task using the policy obtained from the source task, which selects the action with the largest q value among the possible actions in the next state. In this case, the policy refers to a series of rules set by the agent to determine the optimal action. In Q-learning, when α was defined as the rate of learning and δ as the discount factor, the experimental parameters were set to α = 0.05 and δ = 1.0, and the number of nodes was increased from 1 to 25. The accuracy was measured as the proportion of correctly classified data among all data, and the latency was defined as the time when the n^{th} policy was updated. Figure 2 shows the results of evaluating the performance of the proposed model of agnostic learning for efficient federated deep RL and the conventional model of the Q-learning method according to the increase in the number of nodes. The accuracies of the TD and proposed models were 92.3% and 97.8%, respectively, and the memory usage was 546.8 MB and 243.4 MB. Latency is measured by increasing the number of iterations to 4000, and TD is inversely proportional to the number of iterations; however, the proposed model shows a relatively stable pattern, deriving an average latency of 28.9% less than that of TD.

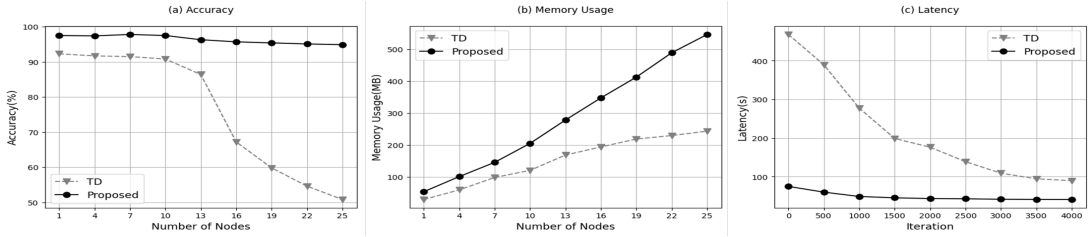

Figure 2: Comparisons models in terms of performance metrics

In this study, federated learning and agnostic meta-learning were considered together to increase the accuracy and reduce memory usage and latency of model learning in a real environment. Future research will focus on preventing the overestimation of q values while maintaining consistency[4].

Acknowledgement :This work was partly supported by grants of the Korea Institute for Advancement of Technology (KIAT) funded by the Korean Government (MOTIE) (P0008703, The Competency Development Program for Industry Specialist) and the MSIT under the ICAN (ICT Challenge and Advanced Network of HRD) program (No. IITP-2022-RS-2022-00156310) supervised by the Institute of Information & Communication Technology Planning & Evaluation (IITP). This study was also supported by a Korea Foundation for Women In Science, Engineering and Technology (WISET) grant funded by the Ministry of Science ICT (MSIT) under the team research program for female engineering students (WISET-2023-141).

References

[1] Badnava, B. Esmaeili, M. Mozayani, N. Zarkesh-Ha P. A new potential-based reward shaping for reinforcement learning agent. https://ieeexplore.ieee.org/document/10099211, 2023.
[2] Yang, Y. Cao, W. Guo, L. Gan, C. Wu M. Reinforcement learning with reward shaping and hybrid exploration in sparse reward scenes. https://ieeexplore.ieee.org/document/10128012, 2023.
[3] Lu, Q. Giannakis G. B. (2021). Gaussian process temporal-difference learning with scalability and worst case performance guarantees. https://ieeexplore.ieee.org/document/9414667, 2021.
[4] X. Xu J. Li, S. Huang and G. Zuo. Generative adversarial imitation learning from human behavior with reward shaping. https://ieeexplore.ieee.org/document/10034038, 2022