H2RL : Transferring Heuristic Knowledge to a Reinforcement Learning Agent for Solving the Online 3D Bin Packing Problem

Minji Kim[†]
Dept. of Computer Science and Engineering
Seoul National University
Seoul, Republic of Korea
mjkim@bi.snu.ac.kr

Ganghun Lee[†], Minsu Lee[*], Byoung-Tak Zhang[*]
Artificial Intelligence Institute (AIIS)
Seoul National University
Seoul, Republic of Korea
{khlee, mslee, btzhang}@bi.snu.ac.kr

***Abstract*—In this paper, we introduce an advanced technique for incorporating heuristic knowledge distillation into reinforcement learning agents. Our approach utilizes a two-stage learning process, where a teacher model, trained with heuristics, guides a student model in acquiring knowledge directly from the environment through reinforcement learning. To assess the performance of our method, we applied it to a logistics loading scenario, addressing an online 3D bin packing problem in both simulated and real-world robotic environments. Experimental results revealed that our method outperforms both the standalone heuristic and conventional RL models, highlighting the effectiveness of the heuristic knowledge distillation architecture in enhancing RL agent performance across diverse settings.**

***Index Terms*—Heuristic knowledge transfer, Reinforcement learning, Robotics, Bin packing problem, Real-world application**

I. Introduction

Reinforcement learning (RL) has emerged as a powerful paradigm for training intelligent agents to interact with complex environments and learn optimal behaviors [1]. However, in real-world applications, especially in robotics, RL-based approaches face several challenges, such as slow convergence, sample inefficiency, and generalization difficulties. Additionally, designing an appropriate reward function can be challenging, even with expert knowledge of the task. To address these challenges, researchers have explored various techniques for incorporating prior knowledge into RL methods [2]–[4].

One promising approach is to leverage heuristic knowledge, which refers to problem-specific strategies based on human knowledge that can guide the agent's learning process. By combining heuristic knowledge with RL, we can expect advantages such as increased sample efficiency, better generalization, and improved robustness. However, if the heuristic knowledge is not well-chosen or is too specific to the training environment, it can easily lead to overfitting and limited flexibility. Furthermore, striking a balance between heuristic knowledge and the RL learning process is crucial to ensure sufficient guidance for the RL model while preserving its capacity for learning and adaptation. Therefore, it is necessary to develop a methodology that actively utilizes heuristic knowledge in the early stages of RL learning to pursue sample efficiency and fast convergence, as well as to enable flexible learning and adaptation through interaction with the environment through RL.

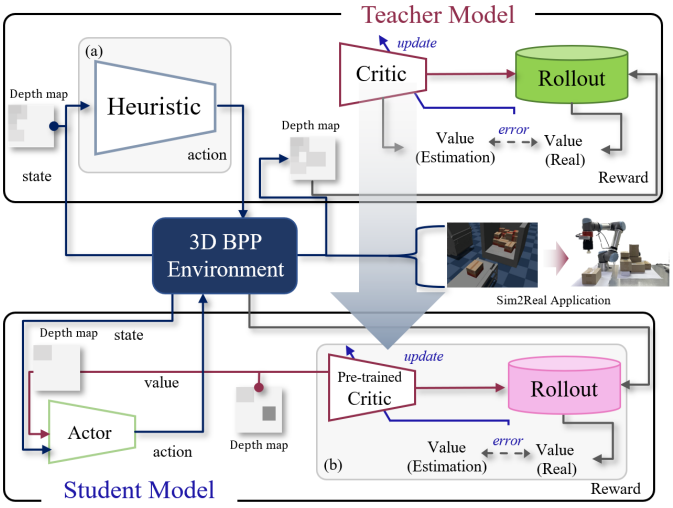

Fig. 1: The model overview of H2RL. (a) Heuristic replaces the actor in the teacher model to learn the *teacher critic*. (b) Using the pretrained teacher critic to train the *student agent*.

In this paper, we introduce a novel heuristic knowledge distillation framework, termed Heuristic to Reinforcement (H2RL), for the effective training of reinforcement learning (RL) agents. (Fig. 1). Specifically, the proposed architecture trains a teacher critic to predict the expected cumulative return from the observed state obtained by the heuristic under environmental and behavioral uncertainties. A student model transfers pretrained teacher critic's knowledge to the student agent while interacting with the environment.

To validate the effectiveness of the proposed approach, we applied it to the logistics loading problem. Solving the loading

[†]These two authors contribute equally to this work
[*]Corresponding authors

problem through learning is essential to increase production efficiency through collaboration with robotics [5]–[7]. We extended the loading task to the 3D Bin Packing Problem (3D-BPP) [8] and designed an online environment for loading 3D box objects into a bin. Results show that our method achieved higher space utilization (SU) than the heuristic and plain RL in both simulation and real-world settings. Moreover, our learned critic used as a buffer controller led to an increase in SU.

II. Related Work

A. Bin Packing Problem

The bin packing problem (BPP) is one of the classic optimization problems with diverse dimensions and conditions [9]–[11]. As an NP-hard problem, heuristics are typically preferred for BPP. Various offline heuristics have been proposed, including a greedy policy approach [12], [13]. 3D-BPP presents additional challenges, with heuristic approaches attempted [14]–[19]. However, unpredictable future constraints in real-world logistics environments limit the effectiveness of heuristics. To overcome these limitations, we combine RL with a heuristic to improve solution quality.

B. Reinforcement Learning with Heuristics

The combination of RL and heuristics has proven to be effective for tasks that can be defined in terms of case-based learning [20]–[22]. By incorporating heuristics, RL can leverage offline data and environmental information to improve learning efficiency, rather than starting from scratch [23], [24]. In contrast to previous research, we attempted to integrate heuristics within RL replacing role of teacher. This approach allows our model to easily adapt to optimal behavior and improve the efficiency of task performance.

C. Knowledge Distillation

Knowledge distillation is a technique for transferring knowledge from a large teacher model to a small student model [25]. It has been extensively utilized in various supervised scenarios [26]–[28]. Despite its relative lack of attention in RL, knowledge distillation can help student agents improve performance and learning efficiency [29]–[31].

III. Method

H2RL consists of two main steps. First, train the *teacher critic* to predict state-value (cumulative reward) following the action of the heuristic. Second, train the *student agent* interacting with the environment, using the pretrained teacher critic. In the second step, the knowledge of the teacher critic is distilled to the student agent in actor-critic RL style. In this way, we expect the student agent to learn heuristic behaviors while searching for more valuable actions from the heuristic.

A. Environment for Online 3D-BPP

We construct an RL environment of online 3D-BPP in the form of Markov Decision Process (MDP). The state has two elements: bin space and object size. We represent the bin space as the top-view 2D-depth map of size (W, H, D) (width, height, and depth, respectively). Each depth value is measured following uniform intervals to the bin. After placing the new object, depth map values on the region of placement decrease. The upcoming object size is represented as (w, h, d).

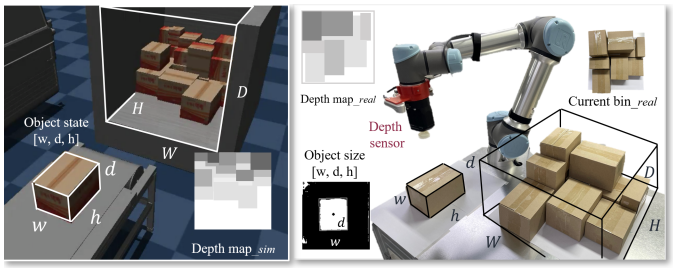

Fig. 2: An implementation of a simulation and real-world experimental setting.

Observing the state, the agent decides x and y $(0 \leq x < W, 0 \leq y < H)$, the placing location for the upcoming object as an action. The placing point is determined by choosing one point on the depth map grid. The agent can change the object orientation ($0°$ or $90°$ around the z-axis). The z-component of the placement is decided as the highest point at the support surface. A portion of actions in a certain state are physically infeasible due to the bin boundary violation and the unstable support surface. We assume the support surface is unstable if the ratio of the highest flat area is lower than 90%.

The reward is a volume of the object $w \times h \times d$ for each placement. The bigger object leads to a higher immediate reward, but space for future objects may decrease. Therefore, the agent should carefully determine the placement to maximize the cumulative rewards. Furthermore, the agent should consider the environmental constraints. As the episode terminates when an infeasible action is chosen, the agent will learn to avoid infeasible actions as they take away the chance of getting rewards afterward.

B. Training Teacher Critic from Heuristic

This paper aims to train the student model, an RL agent, by distilling heuristic knowledge, using a well-performing heuristic for online 3D-BPP as an example.It determines the placement using the sliding window K of size (w, h, d) equal to the upcoming object size. It finds the first feasible placement scanning the depth map with K. This process is conducted for each of the two object orientations, then determines the final placement as the one that causes least variance of the depth map values after placement.

To make a gradient-based teacher from the non-gradient based heuristic, we train the teacher critic network to evaluate the heuristic. Considering the proposed RL form, the teacher critic can learn to predict the expected cumulative return from the observed state obtained by the heuristic. We used temporal difference (TD) [32] learning to train the teacher critic network. The critic network learns rich state-value predictive knowledge from the heuristic by incorporating environmental and behavioral uncertainties. Notably, when training teacher

critics, a four-directional shifting noise is added to the heuristic decision (noisy heuristic) to improve the critic's generalization capacity by showing more diverse states.

C. Training Student Agent from Teacher Critic

After the teacher critic training phase, the heuristic knowledge can be distilled through the critic network to the student agent. Any RL algorithm with actor-critic structure [33] can be used to train the actor of the student agent with the pretrained teacher critic network. When the initialized student agent wanders the environment, the pretrained teacher critic routes the actor to the relatively higher value states. It allows the student agent to improve policies more effectively than with a plain critic. The teacher critic can be used frozen when the student learns, but we do not freeze it to mitigate potential distribution shifts between the heuristic and learning policy. To prevent the student from attempting infeasible actions, we train a feasibility network as an auxiliary task.This network predicts action feasibility and blocks unfeasible actions based on the prediction, as described in [34].

D. Buffer Control with Learned Critic

In some practical variations of online 3D BPP scenarios, the worker may utilize a small buffer to enhance load efficiency. Storing some upcoming objects in the buffer slots gives a chance to 'swap' the object soon to be placed. As the critic predicts future value based on the bin space and the object size, the optimal object in the buffer can be chosen by finding an object that maximizes critic value about the given bin space.

IV. Experiments

To investigate the proposed framework, we compared the space utilization (SU) of four methods: heuristic (zero noise), noisy heuristic (maximum noise is set to 3), plain RL (RL), and our H2RL, in two different environments. We compared our method with a heuristic using a UR5e robot arm and a RealSense D435i vision sensor in the real world.

A. Experimental Settings

We constructed two different environmental settings for the experiments, varying bin size and object size distribution. In *Env-1*, a $20 \times 20 \times 20$ bin and object sizes ranging from 6 to 14 with interval 2 for the all axes were prepared. In *Env-2*, a $30 \times 30 \times 30$ bin and object sizes ranged from 12 to 16 with interval 2 for the x-axis and y-axis, and from 5 to 10 for z-axis were prepared. *Env-2-r* is the real-world version of *Env-2*. For H2RL, the teacher critics were trained for 5M with the noisy heuristic, then student agents were trained for 10M with the teacher critic. Plain RL agents were trained from scratch for 10M. Proximal Policy Optimization (PPO) [35] is used for both H2RL and plain RL.

B. Results

The two upper rows in table I show the average SU for 100 episodes of each method in different simulated settings. RL attained higher SU than Heuristic in *Env-1*, but not in *Env-2*. However, H2RL found a more optimized solution than heuristic in both settings and also outperformed RL. Surprisingly, while the noisy heuristic performed worst, H2RL, which was influenced by the noisy heuristic, performed best. Moreover, H2RL demonstrated slightly faster convergence and enhanced sample efficiency compared to plain RL. This result indicates that H2RL benefited from the noisy heuristic to find the most valuable actions. Since the teacher critic had observed stable episodes made by the heuristic and had learned what states bring relatively high value, H2RL took these advantages from the student agent to learn the proper behaviors effectively. The row below table I demonstrates the averaged results for ten episodes in *Env-2-r*. The overall performance changed since the box distribution differed slightly from *Env-2*. Regardless of the real-world properties, H2RL held the highest SU.

	Heuristic	Heuristic (Noisy)	RL	H2RL (Ours)
Env-1	48.60%	33.20%	50.67%	**54.46%**
Env-2	61.97%	33.89%	43.39%	**65.35%**
Env-2-r	58.93%	47.58%	27.75%	**59.10%**

TABLE I: Comparison of space utilization (SU) in different methods.

	Buffer-1	Buffer-2	Buffer-3	Buffer-4	Buffer-5
Env-1	57.58%	59.31%	61.77%	62.57%	**63.69%**
Env-2	68.73%	70.94%	72.55%	73.37%	**73.77%**

TABLE II: Comparison of space utilization (SU) by H2RL in different number of available buffer slots.

Table II demonstrates the SU of H2RL in buffer-utilized settings with a different number of available buffer slots. SU increased proportionally to the number of slots. This result reveals the essential ability of critics to estimate the impacts of various object sizes, enhancing SU through buffer control.

V. Conclusion

This paper introduces H2RL, a novel approach that addresses the limitations of RL by effectively leveraging human knowledge. The proposed method employs knowledge distillation to predict the critic's value for the state distribution using a heuristic. This leads to a substantial improvement in the performance of student agents, surpassing the heuristic with high SU in both simulated and real-world 3D BPP tasks. Although we evaluated the performance of our model on a single task, we plan to extend our approach to multiple tasks in future work to assess its robustness.

Acknowledgment

This work was partly supported by the IITP (2021-0-02068-AIHub/10%, 2021-0-01343-GSAI/20%, 2022-0-00951-LBA/20%, 2022-0-00953-PICA/20%) grants, and DAPA (No. KRIT-CT-23-003/20%) grant and the NRF of Korea (2021R1A2C1010970/10%) funded by the Korean government.

REFERENCES

[1] P. Dayan and Y. Niv, "Reinforcement learning: the good, the bad and the ugly," *Current opinion in neurobiology*, vol. 18, no. 2, pp. 185–196, 2008.

[2] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in neural information processing systems*, vol. 29, 2016.

[3] J. Hua, L. Zeng, G. Li, and Z. Ju, "Learning for a robot: Deep reinforcement learning, imitation learning, transfer learning," *Sensors*, vol. 21, no. 4, p. 1278, 2021.

[4] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, J. Peters, *et al.*, "An algorithmic perspective on imitation learning," *Foundations and Trends® in Robotics*, vol. 7, no. 1-2, pp. 1–179, 2018.

[5] F. M. Moura and M. F. Silva, "Application for automatic programming of palletizing robots," in *2018 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, pp. 48–53, IEEE, 2018.

[6] W. Echelmeyer, A. Kirchheim, and E. Wellbrock, "Robotics-logistics: Challenges for automation of logistic processes," in *2008 IEEE International Conference on Automation and Logistics*, pp. 2099–2103, IEEE, 2008.

[7] R. Krug, T. Stoyanov, V. Tincani, H. Andreasson, R. Mosberger, G. Fantoni, and A. J. Lilienthal, "The next step in robot commissioning: Autonomous picking and palletizing," *IEEE Robotics and Automation Letters*, vol. 1, no. 1, pp. 546–553, 2016.

[8] A. Benkö, G. Dósa, and Z. Tuza, "Bin packing/covering with delivery, solved with the evolution of algorithms," in *2010 IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA)*, pp. 298–302, IEEE, 2010.

[9] N. Karmarkar and R. M. Karp, "An efficient approximation scheme for the one-dimensional bin-packing problem," in *23rd Annual Symposium on Foundations of Computer Science (sfcs 1982)*, pp. 312–320, IEEE, 1982.

[10] J. Kang and S. Park, "Algorithms for the variable sized bin packing problem," *European Journal of Operational Research*, vol. 147, no. 2, pp. 365–372, 2003.

[11] A. E. F. Muritiba, M. Iori, E. Malaguti, and P. Toth, "Algorithms for the bin packing problem with conflicts," *Informs Journal on computing*, vol. 22, no. 3, pp. 401–415, 2010.

[12] J. de Castro Silva, N. Soma, and N. Maculan, "A greedy search for the three-dimensional bin packing problem: the packing static stability case," *International Transactions in Operational Research*, vol. 10, no. 2, pp. 141–153, 2003.

[13] R. Verma, A. Singhal, H. Khadilkar, A. Basumatary, S. Nayak, H. V. Singh, S. Kumar, and R. Sinha, "A generalized reinforcement learning algorithm for online 3d bin-packing," *arXiv preprint arXiv:2007.00463*, 2020.

[14] F. Alvelos, T. M. Chan, P. Vilaça, T. Gomes, E. Silva, and J. Valério de Carvalho, "Sequence based heuristics for two-dimensional bin packing problems," *Engineering Optimization*, vol. 41, no. 8, pp. 773–791, 2009.

[15] D. Mack and A. Bortfeldt, "A heuristic for solving large bin packing problems in two and three dimensions," *Central European Journal of Operations Research*, vol. 20, pp. 337–354, 2012.

[16] W. F. Maarouf, A. M. Barbar, and M. J. Owayjan, "A new heuristic algorithm for the 3d bin packing problem," in *Innovations and Advanced Techniques in Systems, Computing Sciences and Software Engineering*, pp. 342–345, Springer, 2008.

[17] Y. Wu, W. Li, M. Goh, and R. De Souza, "Three-dimensional bin packing problem with variable bin height," *European journal of operational research*, vol. 202, no. 2, pp. 347–355, 2010.

[18] H. Zhao, C. Zhu, X. Xu, H. Huang, and K. Xu, "Learning practically feasible policies for online 3d bin packing," *Science China Information Sciences*, vol. 65, no. 1, p. 112105, 2022.

[19] H. Hu, X. Zhang, X. Yan, L. Wang, and Y. Xu, "Solving a new 3d bin packing problem with deep reinforcement learning method," *arXiv preprint arXiv:1708.05930*, 2017.

[20] R. A. Bianchi, R. Ros, and R. Lopez de Mantaras, "Improving reinforcement learning by using case based heuristics," in *International Conference on Case-Based Reasoning*, pp. 75–89, Springer, 2009.

[21] N. Mazyavkina, S. Sviridov, S. Ivanov, and E. Burnaev, "Reinforcement learning for combinatorial optimization: A survey," *Computers & Operations Research*, vol. 134, p. 105400, 2021.

[22] M. Fang, Y. Li, and T. Cohn, "Learning how to active learn: A deep reinforcement learning approach," *arXiv preprint arXiv:1708.02383*, 2017.

[23] C.-A. Cheng, A. Kolobov, and A. Swaminathan, "Heuristic-guided reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 13550–13563, 2021.

[24] M. Waltz and K. Fu, "A heuristic approach to reinforcement learning control systems," *IEEE Transactions on Automatic Control*, vol. 10, no. 4, pp. 390–398, 1965.

[25] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[26] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, pp. 1789–1819, 2021.

[27] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2604–2613, 2019.

[28] P. Yun, Y. Liu, and M. Liu, "In defense of knowledge distillation for task incremental learning and its application in 3d object detection," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2012–2019, 2021.

[29] S. Green, C. M. Vineyard, and C. K. Koç, "Distillation strategies for proximal policy optimization," *arXiv preprint arXiv:1901.08128*, 2019.

[30] J. H. Cho and B. Hariharan, "On the efficacy of knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4794–4802, 2019.

[31] W. Son, J. Na, J. Choi, and W. Hwang, "Densely guided knowledge distillation using multiple teacher assistants," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9395–9404, 2021.

[32] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine learning*, vol. 3, pp. 9–44, 1988.

[33] I. Grondman, L. Busoniu, G. A. Lopes, and R. Babuska, "A survey of actor-critic reinforcement learning: Standard and natural policy gradients," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 1291–1307, 2012.

[34] H. Zhao, Q. She, C. Zhu, Y. Yang, and K. Xu, "Online 3d bin packing with constrained deep reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 741–749, 2021.

[35] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.