버퍼 활용 무작위 빈 패킹 문제를 위한 강화학습-휴리스틱 결합 모델

김민지[1], 이강훈[2], 장병탁[3*]

[1,2,3]투모로 로보틱스

[1,3]서울대학교 컴퓨터공학부

[2,3]서울대학교 협동과정 인공지능전공

Reinforcement Learning with Heuristics for Buffer-Utilized Random Bin Packing Problem

Minji Kim[1], Ganghun Lee[2], Byoung-Tak Zhang[3*]

[1,2,3]Tommoro Robotics

[1,3]Department of Computer Science and Engineering, Seoul National University

[2,3]Interdisciplinary Program in Artificial Intelligence, Seoul National University

In this paper, we propose a reinforcement learning algorithm combined with heuristics to solve the bin packing problem (BPP) where the objects are randomly given and once placed objects cannot be moved, but the small load buffer can be utilized. This setting resembles the loading problem which has not been resolved despite of logistics automation. Since heuristics can be rapidly optimized by human intuition, and reinforcement learning is highly responsive to the environment, the combined method is highly available in the real-world industries. When the model learned through reinforcement learning determines the optimal object among the new object and objects in the buffer, then defined heuristics find optimal position and orientation for placement of the selected object. Through experiments, we verified the effectiveness by comparing the loading efficiency between the heuristic-only system and the combined system presented in this study.

Keywords: Buffer-utilized bin packing, Reinforcement learning, Heuristics, Value estimation, Uncertainty handling

논문접수일 : 2022.10.09. 논문수정일 : 2023.05.23. 게재확정일 : 2023.06.01.

1. 서울대학교 컴퓨터공학부 박사과정, 투모로 로보틱스 연구원

2. 서울대학교 협동과정 인공지능전공 박사과정, 투모로 로보틱스 연구원

3*. 서울대학교 컴퓨터공학부 교수, 서울대학교 AI 연구원 원장, 투모로 로보틱스 CEO, Corresponding Author: btzhang@bi.snu.ac.kr

1. 서론

1인 가구의 증가와 베이비붐 세대의 온라인 시장 진입은 상황에 따라 물건을 구매하는 소비 트렌드로 변화를 이끌었다. 이에 따른 다양한 이커머스(e-commerce)의 등장은 온라인에서 구매할 수 있는 상품군의 확대로까지 이어져 택배 물동량 증가의 주요한 원인이 되었다. 코로나19의 여파로 인한 봉쇄와 함께 택배 물류량이 정체될 것이라는 전망과 다르게, 물류 산업은 팬데믹 이전보다 성장한 실적을 보여주었다. 한국통합물류협회에서 발표한 통계에 의하면, 2021년 기준 총 택배 물량은 약 36억 2천만 개로 2020년에 비해 7.59% 성장했고, 코로나 이전인 2018년 기준으로는 25억 4천만 개에 비해서는 42.59%로 크게 증가하였다. 이러한 성장률은 코로나 팬데믹의 여파로 온라인 쇼핑과 모바일 쇼핑 시장의 확대에 기인한 것으로 볼 수 있다. 물건의 종류에 상관없이, 1~2일과 같은 짧은 기간 만에 소비자에게 배달되는 '도어 투 도어(door to door)' 배송 시스템을 기반으로 물류 산업은 다양한 비대면 문화의 기반이 되었다. 비대면 문화는 사람들의 생활 방식에 큰 변화를 불러일으켰다. 식료품 및 생필품을 애플리케이션을 이용해 주문하기 시작하면서 비대면 문화에서의 스마트 물류는 향상된 편리성을 넘어 삶을 유지할 수 있는 도구로 자리 잡았다. 이처럼 크게 향상된 주문량과 물동량 처리를 위해 물류 산업은 발전된 정보통신 기술(ICT)과 인공지능(AI)을 활용한 스마트 물류에 집중하였다.

다수의 물류 기업들은 스마트 물류를 위한 자동화 설비 인프라 구축을 위해 로봇과 인공지능을 도입하는 등 투자를 아끼지 않고 있다. 쿠팡의 물류 센터는 사람과 로봇의 협업에 집중했다. 개인에게 지급되는 PDA를 통해 가장 효율적인 동선을 제공해주고, 모바일 로봇을 통한 물건 운반을 통해 자동화를 이용한 생산성 향상을 이뤄냈다(쿠팡, 2022). CJ대한통운은 Technology, Engineering, System & Solution(TES) 기반의 물류 자동화 센터 구축을 통해 반복 작업을 AMR로 대체하여 인력을 효율적으로 분배할 수 있었고, 송장 인식 및 '오분류 관리 시스템' 등 높은 정확도를 가진 설비를 활용했다(CJ대한통운, 2021).

물류 전반의 과정에서 자동화가 이뤄졌음에도 불구하고, 물류 트럭에 상자를 올리거나 내리는 상하차는 다양한 물류 작업 중에서도 대표적으로 자동화가 더딘 부분이다. 이는 차량에 따른 자동화 구축 설비 구현이 힘들고, 변화하는 환경에서 다양한 규격의 물건을 사람처럼 빠르고 효율적으로 다루기 어렵기 때문이다. 그래도 하차의 경우는 보스턴 다이내믹스의 스트래치와 피클 로봇의 딜 등이 인공지능과 결합하여 적재함에 쌓인 상자들을 컨테이너 벨트 위에 빠르게 올려놓을 수 있도록 개발을 시도한 최근의 사례가 있다(김정은, 2022; 장길수, 2021). 반면 상차는 컨베이어 벨트를 타고 무작위로 접근하는 상자들에 순간적으로 대응해야 하므로 하차보다 더 난도가 높고 사람의 생산성을 따라가지 못하고 있다. 상차와 하차의 차이점은 물체 인식이 아닌, 적재 위치의 적절성에 대한 파악이다. 하차는 물체 파지 후 컨베이어 벨트에 올리는 것이라면, 상차는 동적으로 들어오는 물체들을 순간적으로 파악해서, 최대한의 적재 효율을 살리는 것이 중요하다. 컨베이어 벨트 위로 순차적으로 들어오는 임의의 물건들의 대기열을 보고, 가장 먼저 오는 물건을 두는 것이 아니라 무겁거나 큰 물건을 우선순위에 둔다거나, 빈 곳에 알맞은 크기의 물건을 따로 빼두는 등의 문제들은 자동화를 어렵게 한다. 단순 생산성을 떠나, 업무 난이도와 산재율이 높은 상차는 자동화를 이용해 해결해야 할 중요한 문제이다.

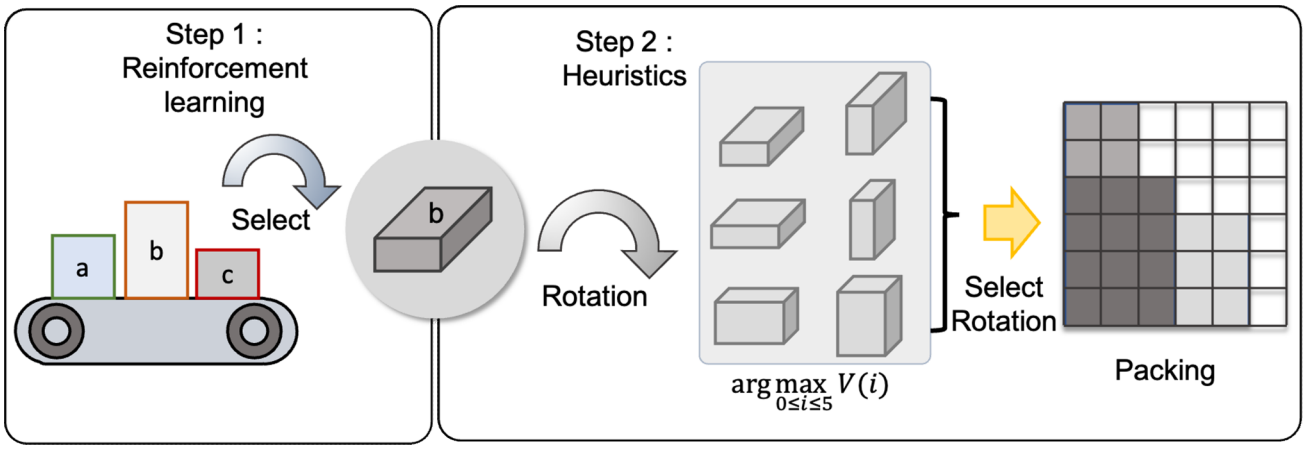
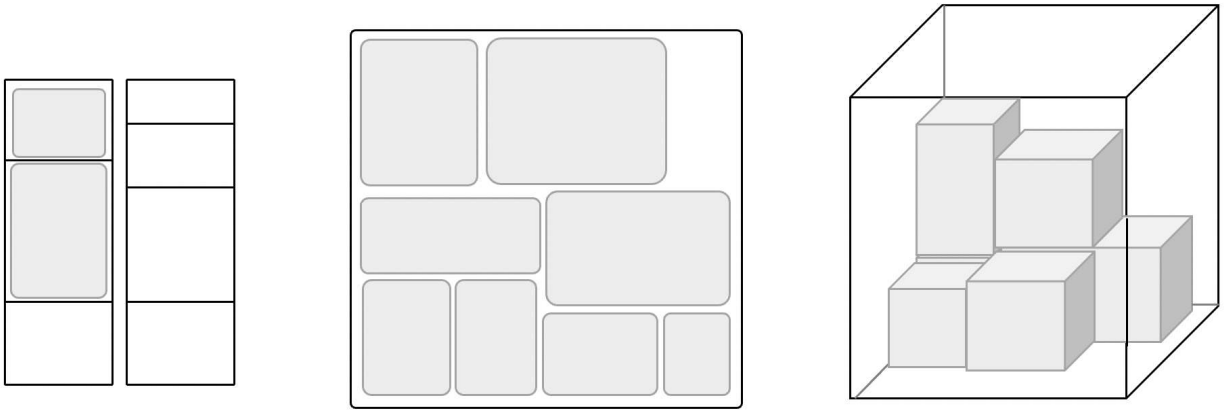


Figure 1. Reinforcement learning with heuristics
for buffer-utilized random bin packing problem model overview

상차 문제는 고정 컨테이너 안에 최대한 많은 물건을 효율적으로 적재해야 하므로 고전적인 최적화 문제인 BPP(Bin Packing Problem)로 볼 수 있다(Benkö et al., 2010). 적재할 영역의 크기, 적재할 물체들의 조건이 다양하므로, BPP는 최적화의 관점에서 여전히 어려운 과제이다. 직육면체의 물체를 다루는 3D-BPP는 여러 분야에서 다양한 휴리스틱(heuristics)을 이용한 조건부 설정으로 해답을 찾기도 하였으며, 학습을 이용하여 최적화를 시도하기도 하였다. 휴리스틱을 사용하면 특정 조건에서 빠른 해답을 찾을 수 있지만, 물체나 환경의 조건이 변함에 따라 전체적인 조건 수정이 필요하다. 인공지능을 이용한 학습의 경우 환경 조건에 변화에 유의미한 대처가 가능하지만, 학습에 따른 시간 및 계산적 비용이 기하급수적으로 커지는 위험이 있으며 학습된 규칙은 수정될 수 없어 전체 시스템이 학습에 기반한다면 모듈화와 사용자 맞춤화가 어렵다. 따라서 본 연구에서는 물류 환경과 같이 동적으로 들어오는 물건을 처리하는 3D-BPP 문제 해결을 위하여 휴리스틱과 강화학습을 결합한 해결책을 제시한다. 기존에 시도된 모든 경우를 시도 및 실패 경험으로 학습하는 강화학습과는 다르게 3D-BPP 학습과는 다르게 제안된 알고리즘은 휴리스틱과 함께 강화학습의 평가자의 가치 판단이라는 개념을 이용하여 문제를 해결하였다. Figure 1은 제안된 모델의 전체적인 모식도이다. 먼저 강화학습을 통해 가까운 미래에 조작할 물체 리스트 중 최적의 물체를 선택한다. 물체를 선택하고 나면, 물체의 가능한 6개의 방위 중 가장 높은 가치를 가지는 방위를 선택하여 물체를 적재한다.

2. 관련 연구

2.1 BPP(Bin Packing Problem)



a) 1D-BPP b) 2D-BPP c) 3D-BPP

Figure 2. Examples of multi-dimension bin packing problem

1차원 BPP는 수식 1과 같이 정리될 수 있다. 주어진 물체들의 세트 $L = \{1, \cdots, n\}$과 각 물체의 무게가 $w_i \in (0, 1)$, $i \in L$ 일 때, BPP에서는 L을 최소한의 적재 공간(Bin) $B_1, B_2, \ldots, B_m$에 넣을 수 있도록 한다.

$$\sum_{i \in B_j} w_i \leq 1, \ 1 \leq j \leq m \tag{1}$$

Figure 2는 차원에 따른 BPP의 예시를 보여준다. 차원이 높고, 영역이 넓을수록 BPP의 시간 복잡도는 크게 증가한다. BPP는 다항 시간 내에 완전한 해결책이 구할 수 없는 nondeterministic polynomial(NP)-hard 문제이다(Coffman et al., 1980). NP-hard 문제는 주로, 간단 추론 방식인 휴리스틱을 이용하여 빠른 최적해 근사를 이용한다. Johnson(1973)은 오프라인 휴리스틱을 이용한 BPP의 여러 방법론을 소개했다. 오프라인 휴리스틱은 적재 공간(bin) 안에 물체가 들어갈 공간이 있으면 배치하고, 들어가지 않으면 새로운 적재 공간을 열어 둔다는 기본적인 조건에 기반한다. 간단한 조건으로 보이지만, 적재 위치를 선택하는 기준에 따라 다양한 알고리즘들이 파생되었다. Best-fit(BF) 알고리즘은 적재 영역의 처음부터 끝까지 탐색하여 최적의 위치에 배치하고, first-fit(FF) 알고리즘은 탐색 중 첫 번째로 발견된 배치 가능 구역에 적재한다. 가장 대중적인 next-fit(NF) 알고리즘은 마지막 적재 위치로부터 적재 위치를 파악한다. 직관적으로는 BF가 가장 좋은 성능을 보일 것 같지만, BF의 높은 시간 복잡도 때문에 고차원의 문제에서는 선호되지 않는다. 반복적인 탐색 작업에서 나아가 Karmarkar and Karp(1982)에서는 오프라인으로 모든 물체를 적재 가능한 공간으로 정렬하여 배치하는 탐욕 정책들을 기반으로 최적해를 제시하였다.

1D-BPP와는 다르게 2D 및 3D-BPP에서는 고려해야 할 요소들이 추가된다. 다양한 규격의 물체들이 적재되면서 생기는 단차에 따른 물체의 기울어짐에 대해 생각해야 할 3D-BPP는 문제의 난이도가 크게 향상된다. 이 점에서 상차 시에 들어오는 물체들을 보고 가장 최적의 물체를 배치하는 직관이 필요하다. 3D BPP의 경우 마찬가지로 휴리스틱으로의 접근이 시도되었다(Gupta et al., 2017). Lodi at el.(2001)은 tabu search를 이용한 휴리스틱 해결책을 제시하였다. Edelkam et al.(2014)에서는 정확성과 랜덤 샘플링의 일반성을 결합한 최근 제안된 탐색 방법인 monte carlo tree search(MCTS)를 통해 기존 local search와 같은 반복적인 개선 알고리즘과 차별점을 가지는 동시에 시간 복잡도를 낮출 수 있었다(Browne et al., 2012). 앞선 선행 연구들은 BPP 최적화가 용이한 방식을 찾아 탐색했지만, 실제 물류 환경에서의 적용에서는 의도된

정도의 성능 기대하기 어렵다. 휴리스틱은 빠른 최적화에 이점을 가지는 건 분명하지만, 실제 물류 환경에 적용해야 하는 예측할 수 없는 미래의 대기열들까지 고려하긴 어렵기 때문이다. 더불어 물체 하나당 발생할 수 있는 모든 방위 상태의 고려 또한 계산 복잡도가 너무 높아져 학습 성능을 저해한다. 본 연구에서는 휴리스틱의 빠른 최적화에 대한 장점을 유지하면서 환경 대응성에 대한 한계를 극복하기 위해 강화학습과의 결합을 시도하였다.

2.2 Deep Reinforcement Learning(DRL)

심층 강화학습은 최근 다양한 문제의 해결책으로 활발히 적용이 이뤄지고 있는 연구 분야이다. 기존 인공지능은 정답이 주어진 상태에서 학습하는 지도 학습(supervised Learning)과 정답이 주어지지 않은 상태에서 비슷한 입력의 군집을 형성하도록 하는 비지도 학습(unsupervised Learning)이 주를 이루었다. 강화학습은 이들과 다르게 시행착오를 거치며 환경(environment)과 에이전트(agent)의 상호작용을 통해 받는 보상의 최대화를 추구하는 학습법이다(Arulkumaran et al. 2017; François-Lavet et al.,2018). 수학적 계산을 위해 필요한 개념이 Markrov Decision Process(MDP)이다. MDP는 (S, A, P, R, γ)로 정의되며, S는 가능한 상태(state)들의 집합과 상태 이행 확률(probability)의 P로 이루어진 Markrov Process(MP)가 확장된 형태이다. 보상(Reward) R과 감가율(discount factor) γ를 통하여 행동(Action) $a \in A$ 에 대한 가치를 학습한다(Puterman, 1990). 수식 2는 감가율이 적용된 보상의 총합을 보여준다. 감가율이 없다면 가까운 시간 안에 받는 작은 보상에 치중할 수 있기 때문에, 먼 미래까지의 보상을 고려할 수 있도록 감가율을 적용해주는 것이다. 순차적 결정 문제(sequential decision problem)의 경우 기대되는 가치에 대한 계산이 가능하다. 벨만 기대 방정식(the bellman equation)은 현재 상태의 가치함수와 다음 상태의 가치함수 사이의 관계를 식으로 나타낸다(수식 3). 가치함수는 가능한 행동들의 영향을 받아 변하는 상태에 기반하기 때문에 상태 이행 확률에 따라 계산되는 것을 알 수 있다(Li, 2017).

$$R_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'} \tag{2}$$

$$v(s) = R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|s,a) v(s') \tag{3}$$

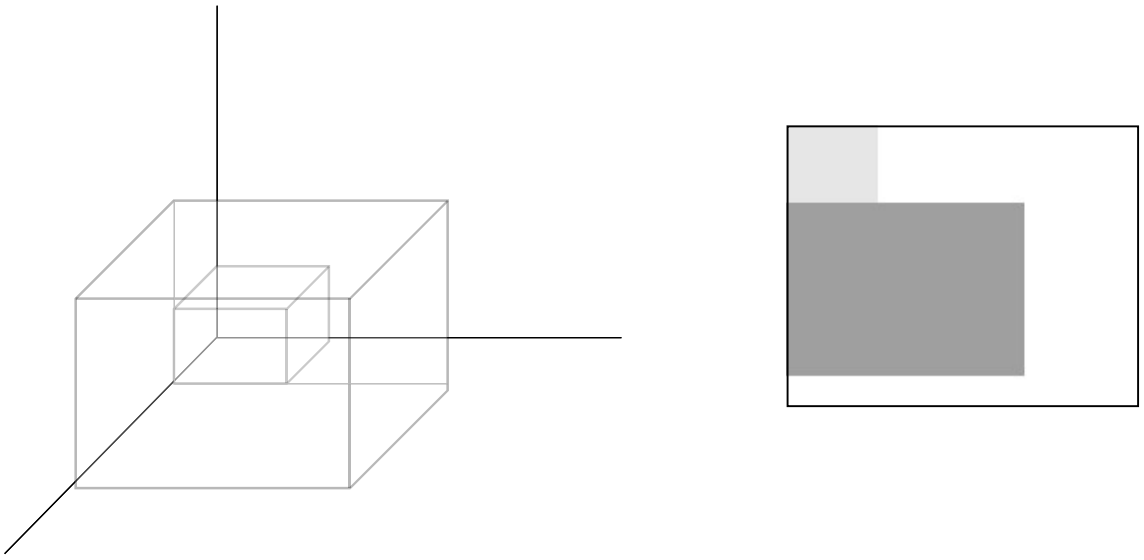
강화학습은 격자구조와 같은 간단한 환경에서부터, 복잡하고 연속적인 환경으로 확장되었다. 게임 환경을 넘어서, 실제 환경의 로봇 작업에 적용되고 있다(Johannink et al., 2019; Ibarz et al., 2021). 강화학습의 로봇 태스크 성공률이 증가함에 따라 물류 환경의 로봇 투입이 촉발되고 있다(Lobbezoo et al, 2021). Iriondo et al.(2019)에서는 물류 환경에서 모바일 로봇을 이용한 조작 태스크를 강화학습으로 해결하였으며, Yang et al.(2020)은 다중 로봇들의 경로 최적화 문제를 해결하였다. 하차에 대한 로봇 적용은 이뤄졌지만, 상차에 대한 로봇 적용이 미비한 이유는 크게 두 가지로 볼 수 있다. 움직이는 물체를 파지하는 것과 어떤 물건을 어디에 적재할지에 대한 유연한 사고를 기대할 수 없기 때문이다. 이러한 이유로 기존 휴리스틱에 강화학습을 결합한 연구들이 진행되고 있다(Cai et al., 2019). Hu et al.(2017)에서는 강화학습을 이용한 3D-BPP 연구가 이루어졌다. 해당 연구에서는 강화학습을 통해 적재할 위치를 학습하고 그에 따른 결과를 학습하였지만, 앞서 설정한 복잡한 환경의 BPP가 아닌 단순화된 형태의 BPP를 다룬다. Zhao et al.(2022) 또한 단순 3D-BPP를 문제를 다루었으나, feasibility 마스크를 통해 불가능한 행동을 고려하여 성능을 향상시켰다. 본 논문에서는 기존 3D-BPP 문제를 임의의 물체를 적재해야 하는 물류 환경에서의 상차 자동화를 위한 buffer-utilized BPP로 심화하고, 강화학습과 휴리스틱이 결합된 형태의 해결 모델을 제시한다.

3. 문제 정의 및 해결법

본 연구에서는 기존 BPP의 기본 전제와 물류 환경에 변동적 조건을 함께 결합한다. 기존 3D-BPP의 기본 전제와 같이 임의의 크기를 가진 직육면체 모양의 물체들이 제한된 영역에 최대한 많이 적재하는 것을 목표로 한다. 이때 물류 환경의 상차 환경은 컨테이너에서 적재할 물체들이 무작위의 크기로 연속적으로 배정되므로, 이를 반영하여 본 연구 환경에서도 물체에 대한 정보는 배정되기 이전에는 미리 알 수 없다. 또한, 상차 시에는 물체를 빠른 속도로 적재해야 하므로 한 번 배치된 물체를 다시 조작하기 어렵다. 이를 반영하여 본 연구 환경에서도 한 번 배치한 물체는 움직일 수 없다. 다만 현재 상태에서 적재가 어려운 물체가 배정되었을 때 숙련자가 잠시 그 짐을 치워두고 다음 물체를 우선 적재하는 것처럼, 작은 크기의 버퍼를 활용하여 무작위의 물체에 다소 대응할 수 있도록 하였다. 이때 해당 버퍼를 활용하여 버퍼의 물체와 새로 배정된 물체 중 적재할 물체를 선택하는 주체는 강화학습으로 학습할 모델이며, 선택된 모델은 휴리스틱에 의해 최적 방향과 위치를 고려하여 물체가 기울지 않도록 배치된다.

3.1 휴리스틱

본 논문에서 제안하는 방법에서 사용되는 휴리스틱 알고리즘은 직접 고안한 알고리즘으로, 사용자의 다양한 의도에 따른 휴리스틱 3D-BPP 중 하나의 종류로 고려할 수 있다. 본 논문에서는 depth map을 구할 수 있는 환경을 가정한다. Depth map은 Figure 3과 같이 적재 공간의 지면에 대하여 수직으로 발사된 광선에 의해 얻어진 톱뷰(Top-view) 형식의 이미지이다. 음영은 물체가 놓인 부분이며 음영이 진할수록 depth가 낮다(높이가 높다). 알고리즘 적재 공간의 가로, 세로, 높이를 각각 W, H, D, 적재 공간의 기본 위치를 0, 탐색 간격(window size)을 ω, depth map을 구하기 위한 광선의 간격과 발사되는 광선 시작점의 높이를 각각 r_g, r_h, depth map을 양자화할 단위를 z_g, 그리고 대상 지역의 평평함을 결정할 문턱 비율을 ρ라 하면 논문에서 제안하는 동적 휴리스틱 적재 알고리즘은 Algorithm 1과 같다(각 값과 관련한 도식은 Figure 4와 Figure 5를 참조한다).

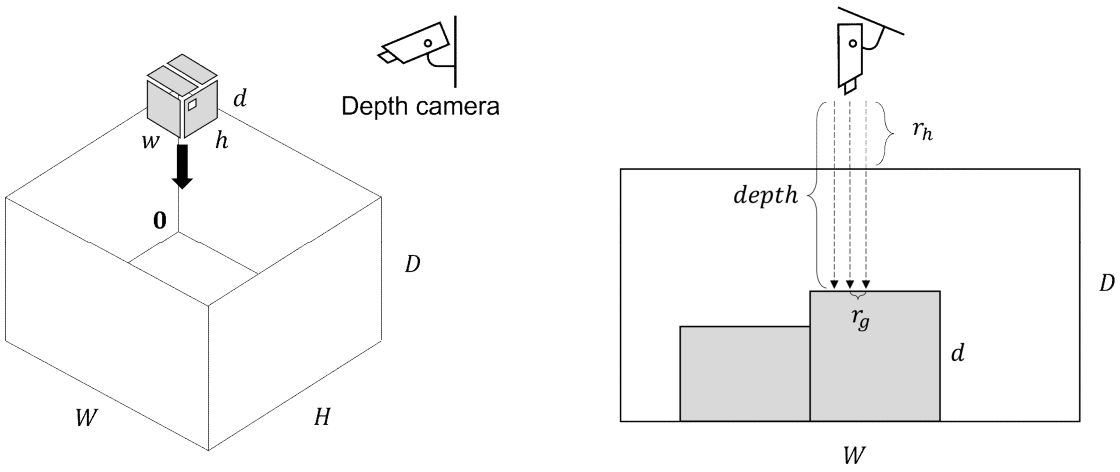


a) 3-D perspective view b) 2-D top-view of depth image

Figure 3. Two objects placed on the plane

초기 적재 공간은 비어있는 상태에서 적재를 시작하며, 매 순간 미리 알 수 없는 크기 w, h, d의 물체가 주어지면 6가지의 물체 방향에 따른 각각의 최적 적재 위치를 구한 후 그중 다음 depth map의 표준편차를 최소화할 물체 방향 및 최적 적재 위치에 따라 적재한다. 만약 어떤 물체 방향에서도 최적 위치가 구해지

지 않는 경우 적재 공간이 가득 찬 것으로 간주한다. 하나의 물체 방향에 대한 최적 위치를 구할 때는, Figure 5와 같이 주어진 탐색 간격에 따라 대상 위치를 움직이면서 모든 대상 위치에 대한 적재 가능 여부를 판단한다. 이때 적재 가능 여부는 적재 대상 높이 공간이 충분한가와 바닥 면이 평평한지에 따라 결정한다. 만약 적재할 물체가 차지할 대상 공간에 대한 depth map을 z_g를 단위로 양자화한 뒤 최빈값의 비율이 ρ 이상일 경우 평평하다고 간주한다. 단, 평평하더라도 최빈값이 최솟값(depth의 최솟값은 곧 높이의 최댓값을 의미한다)이 아니라면 오목한 부분이 평평한 것이므로 적재 불가능하다고 판단한다. 한 방향에 대하여 적재 가능하다고 판단된 모든 적재 위치가 모이면 그중 가장 기본 위치 0과 가까운 물체를 해당 방향에 대한 최적 위치로 간주한다. Figure 6은 고안된 휴리스틱을 사용하여 적재하는 예이다.



a) A container, object, and depth camera b) Side-view of the container and objects

Figure 4. The container and objects
(dotted lines denote the orthogonal rays to obtain a depth image)

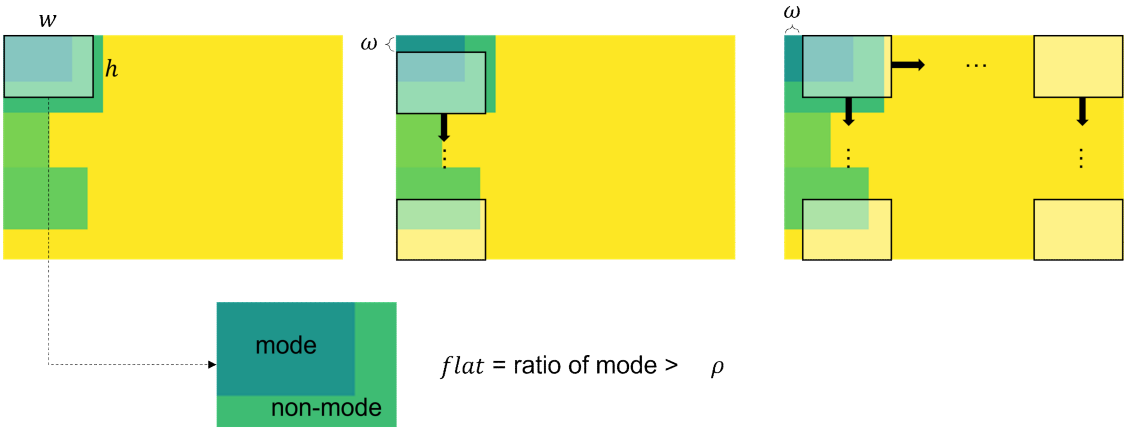


Figure 5. An example of sliding window
for the proposed heuristic algorithm for dynamic bin packing

Algorithm 1: Heuristic Algorithm for Dynamic Bin Packing

Input: $W, H, D, \mathbf{0}, \omega, r_g, r_h, z_g, \rho$
generate grid positions $\mathcal{G}$ with window size ω;
$t \leftarrow 0$;
while *not full* **do**
  capture depth map $\mathcal{M}_t(W, H, \mathbf{0})$;
  get next object of size w, h, d;
  define list of global position candidates $\mathcal{C}$;
  `// consider 6 orientations`
  for $obj \in \{(w,h,d),(w,d,h),(h,w,d),(h,d,w),(d,w,h),(d,h,w)\}$ **do**
    define list of local position candidates $\mathcal{C}_{obj}$;
    `// apply sliding window`
    for $\mathbf{p} \in \mathcal{G}$ **do**
      get depth map for target region $\mathcal{M}_t(w, h, \mathbf{p})$;
      `// find z limit violence of target region`
      $z_{max} \leftarrow D - (min(\mathcal{M}_t(w, h, \mathbf{p})) - r_h)$;
      $overz \leftarrow z_{max} + d > D$;
      `// find flatness of target region`
      quantize $\mathcal{M}_t(w, h, \mathbf{p})$ by z_g;
      $m \leftarrow mode(\mathcal{M}_t(w, h, \mathbf{p}))$;
      $flat \leftarrow ratio(m, \mathcal{M}_t(w, h, \mathbf{p})) \geq \rho$;
      `// find if the mode value expresses highest z`
      $highest \leftarrow m$ equal to $min(\mathcal{M}_t(w, h, \mathbf{p}))$;
      `// add acceptable position cadidate`
      $accept \leftarrow$ not $overz$ and $flat$ and $highest$;
      if *accept* **then**
        simulate next depth map $\mathcal{M}'_{t+1}(W, H, \mathbf{0})$;
        $s_{\mathbf{p}} \leftarrow std(\mathcal{M}'_{t+1}(W, H, \mathbf{0}))$;
        add $(\mathbf{p}, s_{\mathbf{p}})$ to $\mathcal{C}_{obj}$;
      end
    end
    $(\mathbf{p}^*_{obj}, s^*_{\mathbf{p}_{obj}}) \leftarrow (\mathbf{p}, s_{\mathbf{p}}) \in \mathcal{C}_{obj}$ which has minimum $distance(\mathbf{p}, \mathbf{0})$;
    add $(\mathbf{p}^*_{obj}, s^*_{\mathbf{p}_{obj}}, obj)$ to $\mathcal{C}$;
  end
  $(\mathbf{p}^*, obj^*) \leftarrow (\mathbf{p}, obj)$ in $(\mathbf{p}, s_p, obj) \in \mathcal{C}$ which has minimum $s_{\mathbf{p}}$;
  if $\mathcal{C}$ *is empty* **then**
    $full \leftarrow True$
  else
    place object at position $\mathbf{p}^*$ with orientation obj^*;
    $t \leftarrow t + 1$;
  end
end

Algorithm 1. Proposed heuristic algorithm for dynamic bin packing

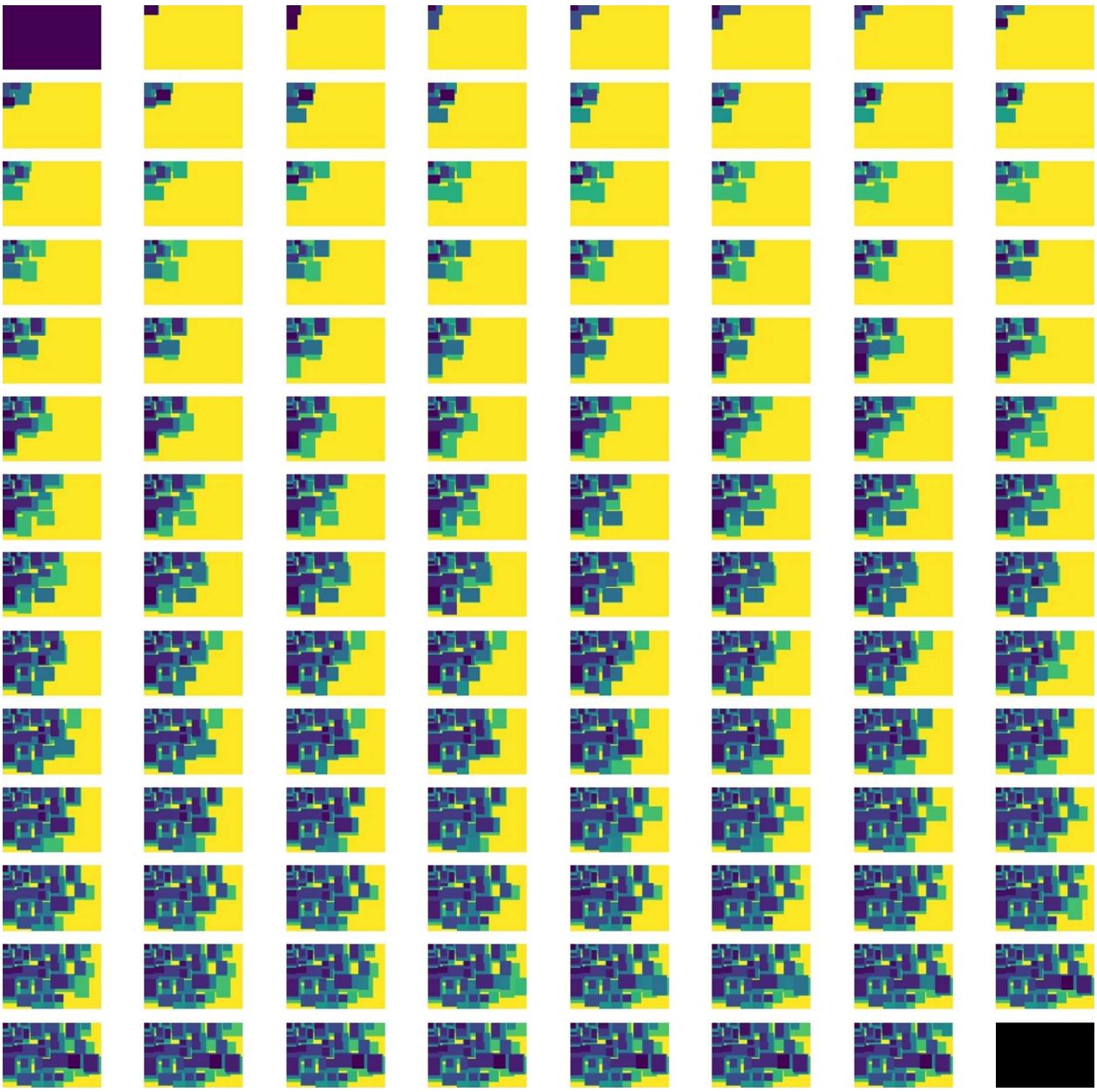


Figure 6. An example of packing sequence using proposed heuristic algorithm

3.2 강화학습과 휴리스틱을 융합한 버퍼 활용 동적 적재

실제로 동적 적재 작업을 수행할 때 상기한 휴리스틱 알고리즘을 직접 사용할 수 있지만, 적재 상태의 빈 곳과 배정된 물체의 크기 차이가 매우 큰 경우 배정된 물체를 처리하기 곤란할 수 있다. 이런 경우 배정된 물체를 하나씩 즉시 처리하는 것보다, 적은 수의 배정된 물체를 임시공간(버퍼)에 저장한 뒤 새로 배정받은 물체와 버퍼에 저장된 물체 중 하나를 선택하여 적재하는 것이 더 유연한 적재방식이다. 다만 주어진 적재 상태에서 최적의 물체를 선택하는 것은, 미래에 어떤 물체가 배정될지 알 수 없으므로 휴리스틱으로 설계하기 까다롭다. 이런 문제는 휴리스틱을 통한 설계 대신 경험적인 방법을 사용하여 접근할 수 있다. 2장 2절에서 서술했듯이 강화학습은 주어진 환경 내에서 여러 번의 경험을 통해 상태와 행동에 대한 가치를 일반화하여 최종 보상을 최대화하도록 행동을 학습하는 경험적 방법으로, 최적의 버퍼 활용을 위해 활용될 수 있다. 제안하는 방법은 강화학습의 개념을 통해 학습된 물체 선택자와 휴리스틱으로 설계된 물체 적재자를 활용하여 동적 적재 작업을 자동화하고 최적의 버퍼 활용을 통해 적재 효율을 증가시킨다.

Algorithm 2: Training Critic with Heuristic

Input: $V_\phi, heu, \mathcal{R}, \gamma, \eta$
while *training* **do**
  // gather heuristic experience
  while *not full* **do**
    capture depth map $\mathcal{M}$ and object size w, h, d;
    set current state $s = \{\mathcal{M}, w, h, d\}$;
    get heuristic action $a = heu(s)$;
    set reward $r = w \times h \times d$;
    capture next depth map $\mathcal{M}'$ and object size w', h', d';
    set next state $s' = \{\mathcal{M}', w', h', d'\}$;
    store transition $\{s, a, r, s'\}$ to rollout $\mathcal{R}$
  end
  // train critic V_ϕ from heuristic experience
  Compute target value $r + \gamma V_\phi(s')$ from $\mathcal{R}$;
  Evaluate advantage $A(s, a) = r + \gamma V_\phi(s') - V_\phi(s)$;
  $\phi \leftarrow \phi + \eta \nabla_\phi A(s, a)$;
end

Algorithm 2. Training critic with heuristic

제안하는 방법에서의 강화학습 방법은 전통적인 방법과 조금의 차이가 있다. 전통적인 강화학습에서는 일반적으로 상태를 관찰하고 행동을 결정하는 행위자(Actor), 그리고 상태를 관찰하고(또는 상태와 행동을 관찰하고) 그 가치를 결정하는 평가자(Critic)의 두 개체를 학습 대상으로 구성한다(Grondman et al., 2012). 그러나 Figure 7과 같이 제안하는 방법에서 학습하는 대상은 평가자뿐이며, 행위자는 없다. 평가자는 설계된 휴리스틱이 버퍼가 없는 적재 작업을 수행하는 동안 주어지는 상태와 보상을 관찰하고 상태에 따른 미래 가치를 추정하도록 학습한다. 이때 상태는 해당 순간의 depth map 및 해당 순간에서 배정된 물체의 크기 w, h, d이다. 보상은 배정된 물체의 부피 $w*h*d$이다. 단, 물류 환경에서는 안쪽부터 바깥쪽으로 쌓아가는 방식을 선호하므로 특별히 적재 공간 내부에 우선순위를 부여하였다. 현재 적재한 위치 우선순위가 이전에 적재한 위치 우선순위보다 높으면 부피 보상을 제공하고, 그렇지 않으면 보상을 제공하지 않는다. 에피소드는 더 이상 배정된 물체를 적재할 수 없을 때 종료된다. 따라서 에피소드가 종료될 때까지 받은 보상이 클수록 더 많은 부피의 물체를 적재한 것이다.

위의 시나리오를 통해 평가자가 충분히 학습되었다면, 평가자는 적재 상태를 나타내는 depth map과 그때 배정된 물체를 관찰하고 에피소드 종료까지 받을 보상에 대한 적절한 예측값을 출력할 수 있다. 만일 평가자에게 같은 depth map에 대해 다른 물체를 대입하여 예측값을 출력하도록 한다면 적재할 물체에 따른 미래의 가치를 비교할 수 있으며, 미래의 가치를 가장 크게 하는 물체를 선택하게 한다면 버퍼를 활용하는 적재 작업에서 최적의 물체를 선택하도록 응용할 수 있다. Figure 8은 학습된 평가자를 버퍼 활용 동적 적재 시나리오에 사용하는 예를 도식화하여 나타낸 것이다. 평가자가 버퍼 안에 들어있는 물체들 중 휴리스틱에 따라 배치할 경우의 가치를 예측하고, 잠재 가치가 가장 높은 물체를 적재하는 과정이 나타나 있다. Algorithm 2는 평가자를 휴리스틱을 이용하여 학습하는 방법에 대해 의사코드로 나타낸 것이다.

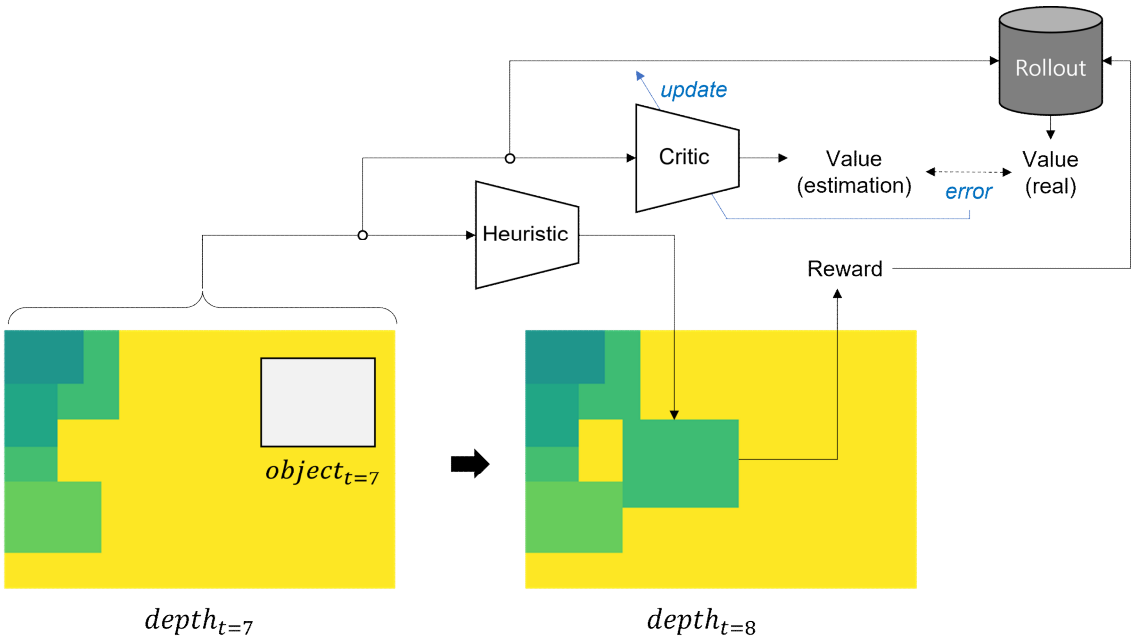


Figure 7. Training critic to estimate future value when depth and object size are given. The critic is trained by reinforcement learning algorithm

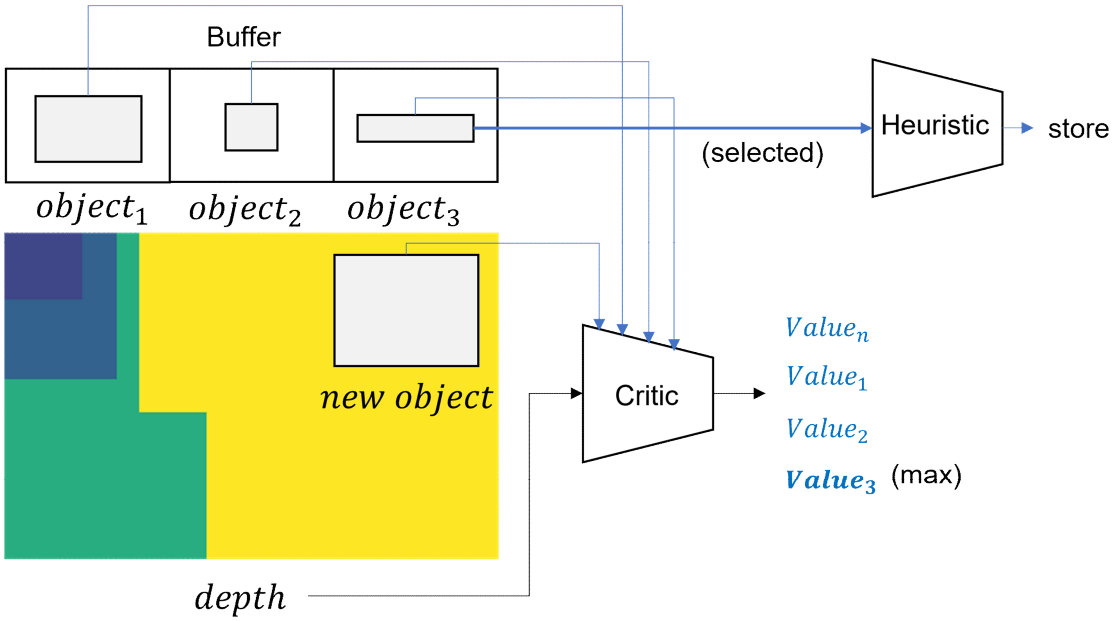


Figure 8. Using trained critic to select optimal object among objects in the buffer and newly given object. Heuristic will select the object which has maximum estimated value from critic, and place the object with predefined algorithm

4. 실험

제시한 방법의 효용을 검증하기 위해 적재 공간의 크기를 (W=0.4, H=0.7, D=0.5), (W=0.4, H=0.7, D=1.0), (W=0.6, H=0.9, D=0.6)의 3가지로 다양화하고, 각 적재 공간에서의 환경 및 휴리스틱 설정값은 탐색 간격 ω=0.02, 광선 간격 r_g=0.02, 광선 높이 r_h=2, depth map 양자화 단위 z_g=0.02, 평평함 문턱 비율 ρ=0.9로 설정하였다. 학습 시 매 순간 주어지는 물체의 크기는 w, h, d 각각 범위 [0.06, 0.14] 내의 0.02배수 크기로 무작위로 주어지도록 하였다. 강화학습 알고리즘으로는 proximal policy optimization(PPO) (Schulman et al., 2017)를 참조하였다. 모든 실험은 Python을 이용해 구축된 시뮬레이션 환경에서 수행되었다.

평가자의 학습에 따른 주요 실험 결과를 도출하기 위해 20번의 에피소드 학습마다 1번의 평가를 수행하였다. 학습 시에는 버퍼 없이 휴리스틱만으로 에피소드를 진행하지만, 평가 방식은 앞 장에서 서술한 바와 같이 매 평가 단계까지 학습된 평가자를 통해 버퍼 활용 적재를 수행한다. 이때의 버퍼 크기는 3으로 설정하였다. 따라서 평가 단계에서는 매 순간 버퍼에 담긴 물체 3개와 새로 배정된 물체 1개의 총 4개의 물체 중 평가자가 가장 고평가하는 물체를 적재할 물체로 선택한다. 1번의 평가에는 5번의 에피소드가 진행하고 그 결과의 평균으로 평가 결과를 도출하였다. 실험에서 주목해야 할 점은 같은 에피소드에 대하여 버퍼를 사용하지 않았을 때보다 학습된 평가자와 함께 버퍼를 사용하였을 때 에피소드 보상의 증분이다. 따라서 후자의 값에서 전자의 값을 뺀 값을 '개선된 에피소드 보상'으로 정의하고 해당 값을 평가 결과값으로 도출하였다.

Figure 9의 그래프는 각 적재 공간 크기에 따라 학습 중 평가된 개선된 에피소드 보상을 나타낸 것이다. 학습 초기에는 평가자가 올바르게 가치를 추정하지 못하여 버퍼를 사용하는 것이 사용하지 않는 것과 큰 차이를 보이지 않지만, 점차 학습이 진행됨에 따라 버퍼와 평가자를 사용할 때 개선된 누적 보상이 상승하는 경향을 보인다. 각 그래프 하단부의 그림은 해당 학습 부근의 최종 depth image 예시를 나타낸 것이다. 학습이 진행됨에 따라 점차 높은 밀도를 보이는 최종 depth image는 평가자의 학습이 진행될수록 버퍼를 더욱 잘 활용하여 적재 공간을 더 효율적으로 사용하고 있음을 나타낸다.

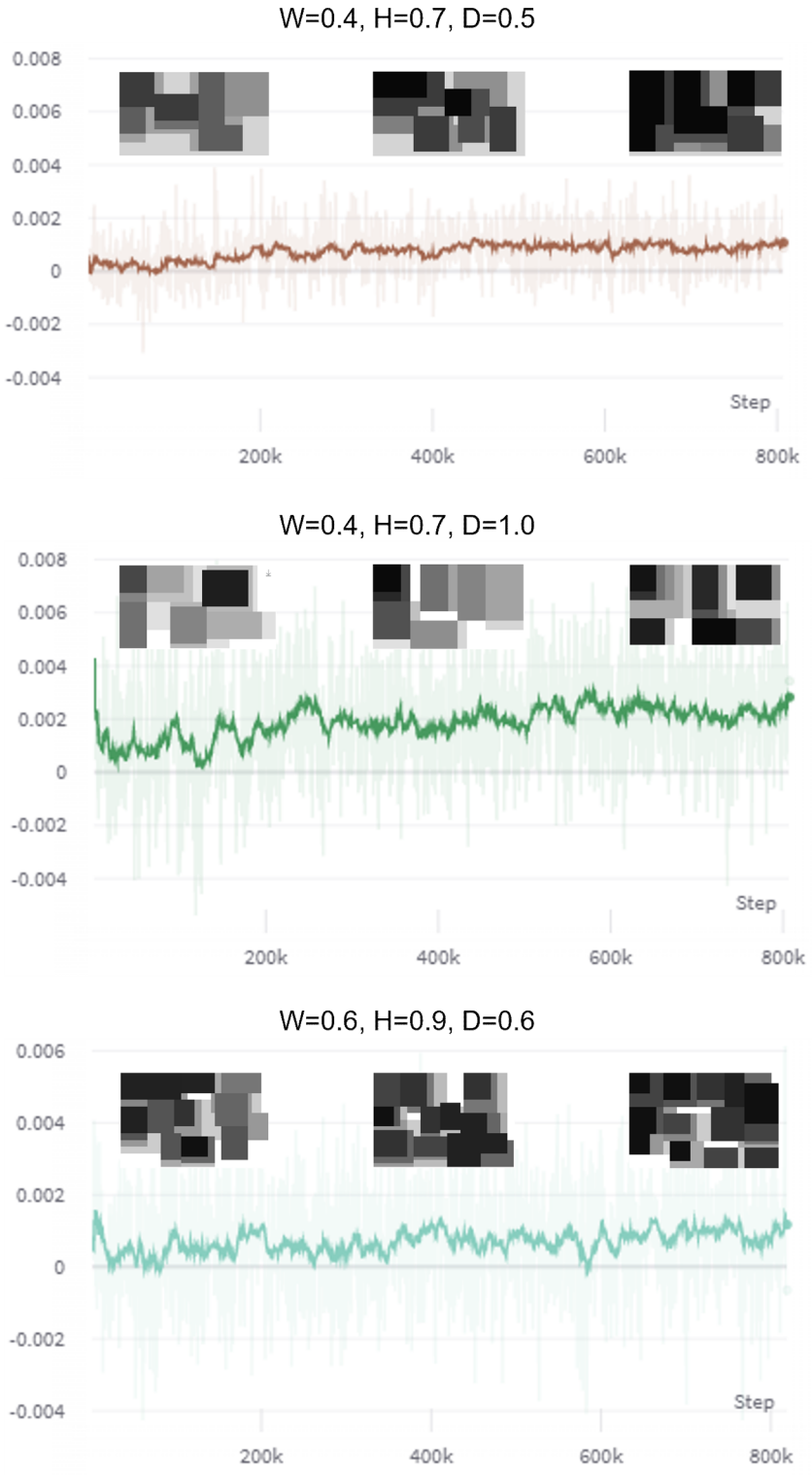


Figure 9. Improved episode reward and final depth images according to the training steps. As improved episode reward grows, the final depth images are getting more dense result

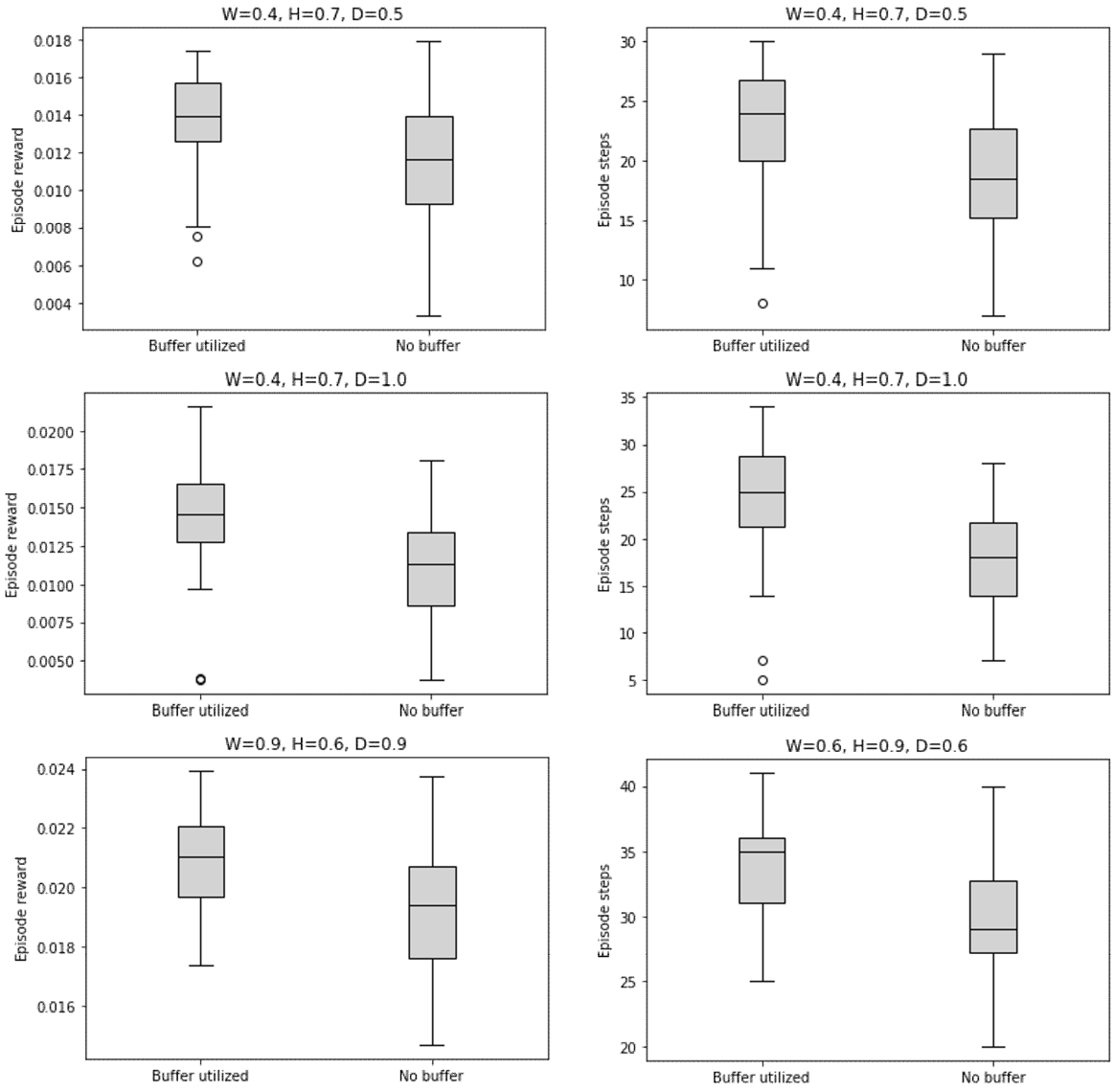


Figure 10. Box plots of episode rewards and episode steps(placed objects + 1) for different container sizes. All results indicate that buffer-utilized with trained critic performed bin packing more efficiently than without buffer

Figure 10은 학습된 평가자와 버퍼를 활용하여 50 에피소드 수행한 결과와 버퍼 없이 휴리스틱만으로 50 에피소드 수행한 결과를 에피소드 보상 및 에피소드 스텝 수로 구분(열)하여 각 적재 공간 크기에 따라(행) 상자 수염 그래프로 비교한 것이다. 에피소드 보상은 최종 적재 부피를 나타내며, 에피소드 스텝 수는(적재한 물체 수 + 1)을 나타낸다. 그래프에서 볼 수 있듯이 학습된 평가자와 버퍼를 활용하였을 때의 최종 적재 부피 및 적재한 물체 수가 버퍼를 사용하지 않았을 때보다 유의미하게 상승하였다. 평가자와 버퍼를 활용한 50 에피소드 수행 결과와 버퍼를 사용하지 않은 50 에피소드 수행 결과는 같은 에피소드에서 각 경우에 대해 2회 수행되었으므로, 두 집단 간 차이의 통계적 유의미성을 확인하기 위해 대응표본 t-검정을 수행할 수 있다. 검정 결과 W=0.4, H=0.7, D=0.5에서의 에피소드 보상 및 에피소드 스텝 수에 대한 p-value는 각각 0.0007, 0.0002, W=0.4, H=0.7, D=1.0에서의 p-value는 최소 단위 이하로 0에 근사한 서로 다른 값 0.0000, 0.0000, 그리고 W=0.6, H=0.9, D=0.6에서의 p-value는 각각 0.0001, 0.0000(근사치)으로 세 적재 공간 크기의 경우 모두에 대해서 평가자와 버퍼를 활용한 집단과 버퍼를 활용하지 않은 집단의 에피소드 보상 및 에피소드 스텝 수는 통계적으로 매우 유의미한 차이를 보였다. 따라서 제안된 방법을 통해 학습한 critic은 다양한 적재 공간에 대하여 휴리스틱과 함께 버퍼를 효율적으로 활용하여 BPP 수행 능력의 품질을 개선한다고 볼 수 있다.

5. 결론

실험을 통해 본 논문에서 제안한 방법이 실제로 버퍼 활용을 통해 적재 효율을 개선할 수 있음을 보였다. 근래 딥러닝 기술의 발달로 물류 자동화가 가속화되고 있지만, 실적용 시에는 학습한 규칙을 수정하기 어려운 딥러닝의 특성 때문에 딥러닝만을 이용하여 물류 자동화를 수행하는 데에는 한계가 있다. 사람의 직관이 반영된 휴리스틱은 실적용 시 수정이 쉽더라도, 정해진 규칙 외의 상태에는 유의미한 선택을 하지 못한다는 한계점을 가진다. 본 논문에서는 딥러닝의 특성을 가지는 강화학습과 사용자 맞춤화가 쉬운 휴리스틱 방법을 적절히 융합한 방법으로, 물류 자동화를 위한 하나의 현실적인 방안을 제시하였다.

해당 모델은 버퍼 내의 적재할 물건 및 새로 배정된 물체 중 어떤 물건을 적재해야 휴리스틱이 더 높은 효율을 가져올 수 있을지 추론한다. 휴리스틱에서는 기존의 연구들과 다르게 적재할 물건의 방위를 고려한다는 점에서 차별점을 가진다. 실제 물류 환경 적용에 의의를 두어, 물체의 기울어짐에 대한 역치를 설정하여 현실적인 적재 환경을 고려할 수 있도록 하였다. 휴리스틱은 기존 환경에 최적화된 적재 결과를 유도할 수 있도록 설계되었기 때문에 성능 향상의 여지가 적다는 예측에도 불구하고, 다양한 실험 조건에서 향상된 생산성을 보여주었다. 그리고 구역화를 통해 적재 공간 안에 우선순위를 두어 적재함으로써 실제 작업 현장에 맞는 적재 형태를 유도할 수 있었다. 또한, 분할 구역 안에 작은 물체를 활용하여 더욱 많은 물건을 배치할 수 있었다는 긍정적인 효과 또한 실험을 통해 확인할 수 있었다.

상차 자동화 문제를 해결하기 위해 해결되어야 할 점이 아직 많지만, 본 논문에서는 기존의 BPP를 상차 환경에 대응하도록 고도화하였다는 의의가 있다. 또한, 고도화된 환경에서도 학습에 필요한 정보는 depth 정보뿐으로, 다수의 특수한 장비가 필요하지 않아 확장성이 좋다. 적재할 용량 설정과 적재 공간 내부의 depth 측정이 가능하다면, 파지 위치를 추정하는 인공지능 로봇과 같은 설비와 함께 자동화가 가능할 것으로 기대된다. 또한, 로봇과의 협업 등을 통해 인력난과 상해 및 업무 강도 문제를 해결할 수 있을 것으로 기대된다. 늘어나는 물동량의 처리를 위한 효율적인 인력 배분이 될 수 있을 것으로 기대된다. 현재로서는 향후 연구로는 단순 평가자를 통한 가치 판단을 넘어, 학습된 평가자를 통한 사람의 지식 전이를 가능하게 하여, 강화학습만으로 BPP를 학습하였을 때의 성능을 향상하고자 한다. 또한, 실제 환경에서의 로봇을 이용한 실험 환경을 구축하고 강화학습의 역할을 확대하여 휴리스틱만 사용한 알고리즘과 학습된 평가자를 사용했을 경우의 비교 실험을 진행하여, 제안한 알고리즘의 유효성을 보일 예정이다.

참고문헌

Arulkumaran, K., Deisenroth, M. P., Brundage, M., Bharath, A. A.(2017), A brief survey of deep reinforcement learning, arXiv preprint arXiv:1708.05866.

Benkö, Attila, György Dósa, and Zsolt Tuza.(2010), Bin Packing/Covering with Delivery, solved with the evolution of algorithms, *2010 IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA).*

Browne, C. B., Powley, E., Whitehouse, D., Lucas, S. M., Cowling, P. I., Rohlfshagen, P., and Colton, S.(2012), A survey of monte carlo tree search methods, *IEEE Transactions on Computational Intelligence and AI in games,* 4(1), 1-43.

Cai, Q., Hang, W., Mirhoseini, A., Tucker, G., Wang, J. and Wei, W.(2019), Reinforcement learning driven heuristic optimization. *arXiv preprint* arXiv:1906.06639.

Edelkamp, Stefan, Max Gath, and Moritz Rohde.(2014), Monte-Carlo tree search for 3D packing with object orientation, *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz).* Springer, Cham.

Edward G Coffman, Jr, Michael R Garey, David S Johnson, and Robert Endre Tarjan.(1980), Performance bounds for level-oriented two-dimensional packing algorithms, *SIAM Journal on Computing,* 9(4):808–826,1980.

François-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., and Pineau, J.(2018), An introduction to deep reinforcement learning, *Foundations and Trends® in Machine Learning,* 11(3-4), 219-354.

Garey, M. R, Johnson, D. S. and Victor Klee(ed.).(1979), Computers and Intractability: A Guide to the Theory of NP-Completeness. *A Series of Books in the Mathematical Sciences.* San Francisco, Calif.: W. H. Freeman and Co.

Grondman, I., Busoniu, L., Lopes, G. A., and Babuska, R.(2012), A survey of actor-critic reinforcement learning: Standard and natural policy gradients, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* 42(6), 1291-1307.

Gupta, A., Guruganesh, G., Kumar, A., and Wajc, D.(2017), Fully-dynamic bin packing with limited repacking, *arXiv preprint arXiv*:1711.02078.

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

Ibarz, J., Tan, J., Finn, C., Kalakrishnan, M., Pastor, P. and Levine, S.(2021), How to train your robot with deep reinforcement learning: lessons we have learned. *The International Journal of Robotics Research,* 40(4-5), 698-721.

Iriondo, A., Lazkano, E., Susperregi, L., Urain, J., Fernandez, A., and Molina, J.(2019), Pick and place operations in logistics using a mobile manipulator controlled with deep reinforcement learning, *Applied Sciences,* 9(2), 348.

Johannink, T., Bahl, S., Nair, A., Luo, J., Kumar, A., Loskyll, M., ... and Levine, S.(2019), Residual reinforcement learning for robot control. In 2019 International Conference on Robotics and Automation (ICRA), 6023-6029 pp.

Johnson, David S.(1973), Near-optimal bin packing algorithms. *Diss. Massachusetts Institute of Technology*.

Karmarkar, N and Karp, R. M.(1982), An efficient approx-imation scheme for the one-dimensional bin-packing prob-lem. *In 23rd Annual Symposium on Foundations of Computer Science (sfcs 1982)*, 312-320. IEEE.

Li, Yuxi.(2017), Deep reinforcement learning: An overview, *arXiv preprint* arXiv:1701.07274

Lobbezoo, Andrew, Yanjun Qian, and Hyock-Ju Kwon.(2021), Reinforcement Learning for Pick and Place Operations in Robotics: A Survey, *Robotics 10.3* : 105.

Lodi, Andrea, Silvano Martello, and Daniele Vigo.(2002), Heuristic algorithms for the three-dimensional bin packing problem. *European Journal of Operational Research 141.2*, 410-420.

Martello, Silvano, David Pisinger, and Daniele Vigo.(2000), The three-dimensional bin packing problem. *Operations research 48.2*, 256-267.

Puterman, Martin L.(1990), Markov decision processes, *Handbooks in operations research and management science 2*, 331-434.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O.(2017), Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

Yang, Yang, Li Juntao, and Peng Lingling(2020), Multi-robot path planning based on a deep reinforcement learning DQN algorithm, *CAAI Transactions on Intelligence Technology 5.3* : 177-183.

Zhao, H., Zhu, C., Xu, X., Huang, H., and Xu, K.(2022), Learning practically feasible policies for online 3D bin packing." *Science China Information Sciences 65.1*, 1-17.

김정은,「보스턴 다이내믹스, "물류 로봇 '스트레치' 올해 생산량 이미 매진"」, THE DAILYPOST, 2022.04.04, 1쪽. https://www.thedailypost.kr/news/articleView.html?idxno=86670

장길수,「피클 로봇, 고속 박스 하적 로봇 '딜(Dill)' 개발」, 로봇신문, 2021.04.20, 1쪽. http://www.irobotnews.com/news/articleView.html?idxno=24627

CJ대한통운,「'AI 혁신 기술'이 이끄는 CJ대한통운의 스마트 물류 혁명」, CJ대한통운, 2021.07.28, 1쪽. https://www.cjlogistics.com/ko/newsroom/latest/LT_00000238

쿠팡,「로봇과 인공지능: 쿠팡 물류의 최신 기술을 소개합니다」, 쿠팡뉴스룸, 2022.09.06, 1쪽. https://news.coupang.com/archives/19485/

요약문

본 논문에서는 강화학습과 휴리스틱을 결합하여 매 순간 무작위로 물체가 주어지지만 작은 크기의 적재 버퍼를 활용할 수 있는 bin packing problem (BPP) 문제를 해결하는 방법을 제안한다. 이때 한번 놓인 물체의 위치는 변경될 수 없다. 이러한 환경은 전반적인 물류 프로세스 자동화의 물결에도 아직 원활히 해결되지 않은 상차 환경과 유사하다. 휴리스틱은 인간의 직관에 따라 신속한 최적화가 가능하고, 강화학습은 환경에 대한 반응성이 뛰어나므로 두 방법을 결합한 방법은 실제 물류 환경으로의 적용이 용이하다. 강화학습을 통해 학습한 모델이 신규 물체 및 버퍼 내 일부 물체들에 대한 각각의 가치를 판별하고 그에 따른 최적의 물체를 선택하면, 선택된 물체는 정의된 휴리스틱 알고리즘에 따라 물체의 최적 방위와 위치를 고려하여 배치된다. 이런 방식으로 결합된 자동화 방법은 버퍼를 효과적으로 활용하여 적재 효율을 증대시킬 수 있다. 실험을 통해 실제로 본 연구에서 제시한 방법을 통해 버퍼를 함께 고려했을 때 휴리스틱 모델만을 사용했을 때보다 적재 효율이 증가함을 확인하였다.

키워드: 버퍼 활용 빈 패킹, 강화학습, 휴리스틱, 가치 측정, 불확실성 제어