ISSN : 1226-7244 (Print)
ISSN : 2288-243X (Online)
논문번호 21-02-15
104

j.inst.Korean.electr.electron.eng.Vol.25,No.2,337~343,june 2021
https://doi.org/10.7471/ikeee.2021.25.2.337

심층 강화학습을 이용한 모바일 로봇의 맵 기반 장애물 회피 알고리즘

Map-Based Obstacle Avoidance Algorithm for Mobile Robot Using Deep Reinforcement Learning

선 우 영 민*, 이 원 창**★

Yung-Min Sunwoo*, Won-Chang Lee**★

Abstract

Deep reinforcement learning is an artificial intelligence algorithm that enables learners to select optimal behavior based on raw and, high-dimensional input data. A lot of research using this is being conducted to create an optimal movement path of a mobile robot in an environment in which obstacles exist. In this paper, we selected the Dueling Double DQN (D3QN) algorithm that uses the prioritized experience replay to create the moving path of mobile robot from the image of the complex surrounding environment. The virtual environment is implemented using Webots, a robot simulator, and through simulation, it is confirmed that the mobile robot grasped the position of the obstacle in real time and avoided it to reach the destination.

요 약

심층 강화학습은 학습자가 가공되지 않은 고차원의 입력 데이터를 기반으로 최적의 행동을 선택할 수 있게 하는 인공지능 알고리즘이며, 이를 이용하여 장애물들이 존재하는 환경에서 모바일 로봇의 최적 이동 경로를 생성하는 연구가 많이 진행되었다. 본 논문에서는 복잡한 주변 환경의 이미지로부터 모바일 로봇의 이동 경로를 생성하기 위하여 우선 순위 경험 재사용(Prioritized Experience Replay)을 사용하는 Dueling Double DQN(D3QN) 알고리즘을 선택하였다. 가상의 환경은 로봇 시뮬레이터인 Webots를 사용하여 구현하였고, 시뮬레이션을 통해 모바일 로봇이 실시간으로 장애물의 위치를 파악하고 회피하여 목표 지점에 도달하는 것을 확인하였다.

Key words : Mobile Robot, Path Planning, Deep Reinforcement Learning, Dueling Double DQN, Webots

* Dept. of Smart Robot Convergence and Application Engineering, Pukyong National University
** Dept. of Electronic Engineering, Pukyong National University
★ Corresponding author
E-mail : wlee@pknu.ac.kr, Tel : +82-51-629-6219
※ Acknowledgment
This work was supported by a Research Grant of Pukyong National University(2021)
Manuscript received May 14, 2021; revised Jun. 22, 2021; accepted Jun. 25, 2021.

Ⅰ. 서론

모바일 로봇이 모르는 환경에서 길을 찾아갈 수 있게 해주는 로봇 네비게이션 능력은 모바일 로봇을 사용하는 시스템에서 가장 중요하다. 네비게이션의 일반적인 목적은 시작점으로부터 목표지점까지 최적이나 혹은 준 최적경로를 2차원이나 3차원 환경을 따라 장애물을 회피하며 찾는 것이다[1]. 하지만 장애물 회피를 하면서 동시에 시간을 최소한으로 소비하는 경로를 찾는 것은 여전히 매우 어려운 과제이다[2].

최근 강화학습은 정교하고 사람이 설계하기 어려

운 작업을 로봇이 수행하도록 프레임워크와 다양한 도구들을 제공하기 위해 사용되고 있다[3]. 강화학습은 로봇이 환경과 상호작용하며 보상이라고 불리는 스칼라 피드백 신호를 통해 최적의 행동을 자율적으로 발견하게 해주고, 강화학습으로 인해 설계자는 문제의 정답을 명백하게 제시하는 것이 아니라 로봇의 매 단계의 수행을 평가하는 보상함수를 제시한다. 로봇과 관련된 다양한 문제를 해결하는데 강화학습이 사용되었으며 모바일 로봇 경로 생성 분야에서도 많은 중요한 업적들을 이루었다[4].

하지만 기존의 강화학습 알고리즘들을 모바일 로봇의 경로 생성 문제에 사용한다면 로봇에 설치된 다양한 센서들에서 입력되는 고차원의 데이터들을 그대로 상태로 사용할 수 없다는 한계가 있다[5]. 다행히 강화학습에 성공적으로 신경망을 적용한 심층 강화학습 알고리즘들이 등장했고 다양한 분야에서 큰 성과를 이루고 있다. 가장 대표적인 심층 강화학습 알고리즘으로 Deep Q-Network(DQN)이 있으며, 가공되지 않은 고차원의 이미지 데이터를 입력으로 받아 아타리 2600 게임에서 높은 점수를 얻는 정책을 성공적으로 학습하였다[6]. 특히 이미지처럼 고차원의 데이터를 그대로 입력받아 최적 정책을 발견할 수 있다는 장점으로 인해 많은 모바일 로봇의 경로 생성 문제에서도 심층 강화학습 알고리즘이 사용되어왔다[7-9].

이 논문에서는 모바일 로봇의 경로 생성 문제를 해결하기 위해 로봇이 이동할 환경의 이미지를 사용하는 방법을 제안한다. 위에서 바라보는 카메라를 통해 로봇의 현재위치, 목표 지점의 위치 그리고 장애물들의 위치를 파악하고 그 정보들을 바탕으로 그리드 맵을 만들어 심층 강화학습의 입력으로 사용한다. 다음으로 로봇의 현재위치에서 목표 지점까지 장애물과 충돌을 피하는 경로를 실시간으로 생성하고, 그 경로를 따라 로봇을 움직인다. 그리고 심층 강화학습 알고리즘으로는 우선순위 경험 재사용을 사용하는 D3QN을 사용한다. 이렇게 DQN을 개선시킨 심층 강화학습 알고리즘을 선택함으로써 DQN은 학습하지 못했던 복잡하고 어려운 환경에서도 모바일 로봇이 빠르고 안전하게 목표 지점에 도달하게 하는 경로를 생성할 수 있었다.

II. 관련기술

1. DQN

DQN은 핵심적인 심층 강화학습 알고리즘으로 강화학습 제어 알고리즘인 Q-learning과 심층 학습을 결합하여 최초로 이미지로부터 학습에 성공했다. DQN의 두 가지 성공요인으로 경험 재사용과 분리된 타겟 네트워크의 사용이 있다. 에이전트의 경험을 매 타임 스텝마다 경험 재사용 메모리에 저장하고 학습이 진행될 때 정해진 개수만큼 균일 랜덤 샘플링을 통해 미니배치를 구성하여 입력으로 넣어주게 된다. 그리고 타겟 네트워크란 학습이 진행되는 네트워크와 구조는 같지만 독립적인 파라미터를 가지는 네트워크이다. i번째 학습 네트워크 업데이트에 사용될 목표 값은 식 (1)과 같다.

$$y = r + \gamma \max_{a'} q(s', a'; \theta_i^-) \tag{1}$$

여기서 s'와 r은 각각 샘플링 된 경험에 존재하는 다음 상태와 보상이고 a'는 파라미터 θ^-를 가지는 타겟 네트워크에 의해 선택된 행동이다. 그리고 γ는 감쇠율로 미래 보상의 현재 가치를 결정한다.

2. Double DQN

Double DQN은 Q-learning의 행동 가치함수 과대평가 현상을 줄인 Double Q-learning을 DQN에 결합시킨 심층 강화학습 알고리즘으로 더 정확한 행동 가치함수 추정을 할 뿐만 아니라 몇몇 아타리 게임들에서 DQN보다 더 높은 점수를 얻었다[10]. Dobule DQN에서 사용될 목표 값은 식 (2)와 같다.

$$y = r + \gamma q(s', \arg max_{a'} q(s', a; \theta_i); \theta_i^-) \tag{2}$$

식 (2)에서 행동을 선택하는 과정과 그 행동을 평가하는 과정을 학습 네트워크와 타겟 네트워크에 역할을 분배함으로 행동 가치함수 과대평가 현상을 방지한다.

3. Dueling DQN

Dueling DQN은 dueling architecture를 사용하여 상태 가치함수와 이득함수로 표현되는 두 흐름의 결합으로 행동 가치함수를 추정한다[11]. Dueling DQN은 식 (3)과 같은 새로운 행동 가치함수를 추정한다.

$$q(s,a;\theta) = v(s;\theta) + \left(A(s,a;\theta) - \frac{1}{|A|}\sum_{a'} A(s,a';\theta)\right) \quad (3)$$

식 (3)에서 볼 수 있듯이 상태 가치함수와 이득함수의 조합으로 행동 가치함수가 만들어지며 이로 인해 행동 가치함수를 업데이트하는 과정이 더욱 안정해진다.

4. 우선순위 경험 재사용

우선순위 경험 재사용은 DQN의 경험 재사용 방법을 사용할 때 경험 재사용 메모리에서 학습에 사용될 데이터들을 균일 랜덤 샘플링이 아닌 시간차 오차의 크기에 따라 우선순위를 매겨 샘플링하는 방법으로 경험 재사용 방법을 더욱 효율적으로 만들어 준다[12]. 하지만 항상 우선순위가 큰 샘플만 샘플링하게 된다면 낮은 우선순위를 가지는 샘플들은 선택받지 못하게 되고 경험 재사용 메모리의 부분 집합만 사용하게 된다. 이를 해결하기 위해 확률적으로 샘플을 선택하게 되고 이때 메모리에 존재하는 i번째 샘플이 선택될 확률은 다음과 같다.

$$P(i) = \frac{p_i^{\alpha}}{\sum_k p_k^{\alpha}} \quad (4)$$

여기서 p_i는 샘플i의 우선순위이며 상수 α는 우선순위를 얼마나 사용할지 결정한다.

III. 시뮬레이션 환경

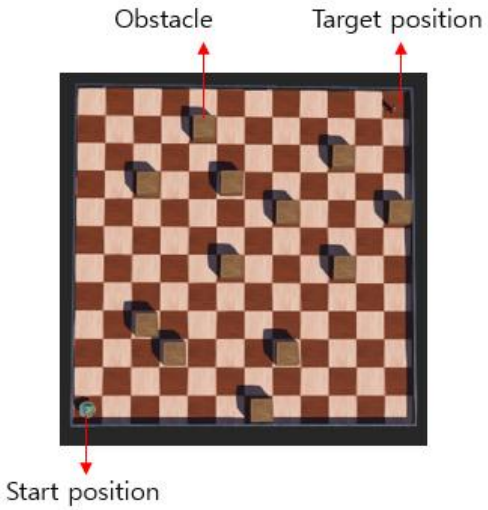

Fig. 1. Problem environment created using Webots.
그림 1. Webots를 이용하여 만든 문제 환경

본 논문에서 사용할 모바일 로봇과 로봇의 최적 경로 탐색을 수행하게 될 환경은 Cyberbotics에서 제공해주는 로봇 시뮬레이터인 Webots를 사용하여 구현하였으며, Webots는 사실적인 로봇 시뮬레이터로 현실 세계와 시뮬레이션 사이의 차이를 상당히 줄여준다는 장점이 있다[13-14]. Webots에서 생성한 환경은 그림 1과 같다.

시뮬레이션 환경은 12×12의 격자로 이루어져 있으며 4개의 벽으로 둘러싸여 있고 모바일 로봇의 출발 지점과 목표 지점은 고정되어 있다. 또한 장애물인 상자는 출발 지점과 목표 지점을 제외한 랜덤한 위치의 격자에 최대 20개까지 존재하게 된다. 매 에피소드마다 로봇은 항상 출발 지점에 위치하게 되며 장애물들은 랜덤하게 생성된다. 우리의 목적은 가능한 모든 에피소드에서 장애물 또는 벽과의 충돌을 피하며 가장 빠른 시간 내에 목표 지점

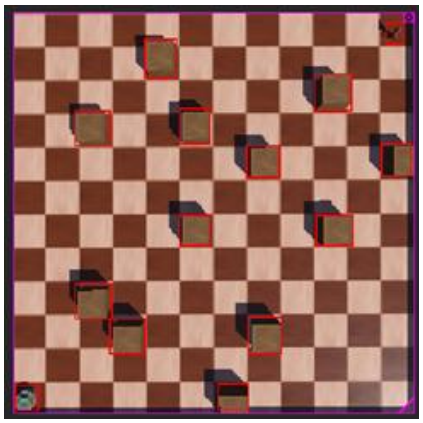

Fig. 2. Image of the environment obtained from the camera.
그림 2. 카메라로부터 얻은 환경의 이미지

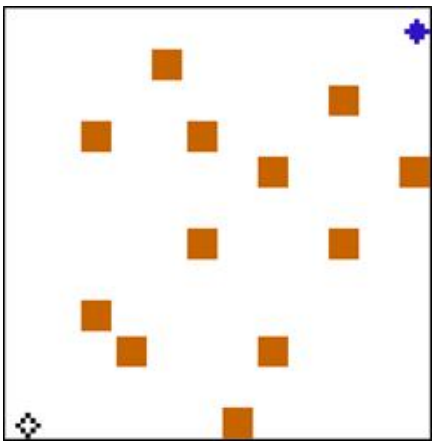

Fig. 3. Grid map created based on information obtained from images.
그림 3. 이미지에서 얻은 정보를 바탕으로 만들어진 그리드 맵

까지 도달하는 모바일 로봇의 경로를 생성하는 것이다. 카메라는 환경을 쳐다볼 수 있게 설치되어 있고 물체 인식 기능을 사용한다. 카메라를 통해 얻은 이미지에는 현재 모바일 로봇과 목표 지점 그리고 장애물들의 정보와 위치가 담겨있으며 그림 2와 같다. 그림 2에서 카메라에 인식된 모든 물체는 빨간 사각형으로 표현되어 있다. 카메라가 획득한 이미지로부터 얻은 물체의 정보와 위치를 바탕으로 만들어진 그리드 맵은 그림 3과 같다.

그리드 맵은 84×84의 크기를 갖는 이미지로 표현되고 우리가 설정했던 환경과 마찬가지로 12×12의 격자로 이루어진다. 장애물들은 주황색 사각형으로 표현되며 현재 로봇의 위치는 검은색 마름모 그리고 목표 지점은 파란색 마름모로 표현되어 있다. 이렇게 생성된 그리드 맵은 상태로써 심층 강화학습 알고리즘의 입력으로 들어가며 에이전트는 현재 그리드 맵에서 이득이 가장 크게 되는 행동을 하나 선택해야 한다. 모바일 로봇과 로봇이 수행할 수 있는 행동은 그림 4와 같고 로봇이 현재 위치한 격자에서 가능한 행동은 총 4가지로 한 번의 타임 스텝에 하나의 방향으로 1칸씩 이동할 수 있다.

Fig. 4. Mobile robot and admissible behaviors.
그림 4. 모바일 로봇과 허용되는 행동

로봇이 받게 될 스칼라 신호인 보상함수는 식 (4)과 같다.

$$r_{t+1} = \begin{cases} 0 & reach\,destination \\ -0.1 & otherwise \end{cases} \quad (4)$$

만약 로봇이 목표 지점에 도달한다면 0의 보상을 얻으며 장애물과 부딪히거나 목표 지점이 아닌 곳으로 이동한다면 -0.1의 음의 보상을 받게 된다. 에피소드마다 누적 보상이 제일 크게 되는 방법은 장애물과 부딪히지 않고 가장 빠른 타임 스텝 내에 목표지점에 도달하는 것이다. 시스템 구조는 그림 5와 같다.

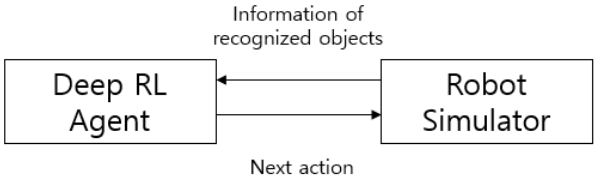

Fig. 5. System architecture.
그림 5. 시스템 구조

타임 스텝 t에서 카메라를 통해 인식된 모든 물체의 정보가 심층 강화학습 에이전트로 전송된다. 받은 정보를 바탕으로 상태 s_t인 그리드 맵이 만들어지고 에이전트의 입력으로 들어가며 t에서 모바일 로봇의 행동 a_t가 로봇 시뮬레이터로 전송되면 그에 따라 모바일 로봇을 움직이게 된다. 그리고 보상 r_{t+1}을 받고 다음 타임 스텝 $t+1$로 넘어가게 된다. 이 과정이 $t=0$에서부터 모바일 로봇이 목표 지점에 도달할 때 까지 반복된다.

Ⅳ. 심층 강화학습 훈련 결과

설정한 시뮬레이션 환경을 사용하여 두 가지의 심층 강화학습 알고리즘을 학습시켰다. 한 가지는 DQN이고 다른 하나는 우선순위 경험 재사용을 사용하는 D3QN 이다. 이때 보상함수는 식 (4)과 동일하며 심층 강화학습 에이전트의 입력은 그림 3과 같은 그리드 맵이다. 이때 학습 결과는 그림 6과 같다.

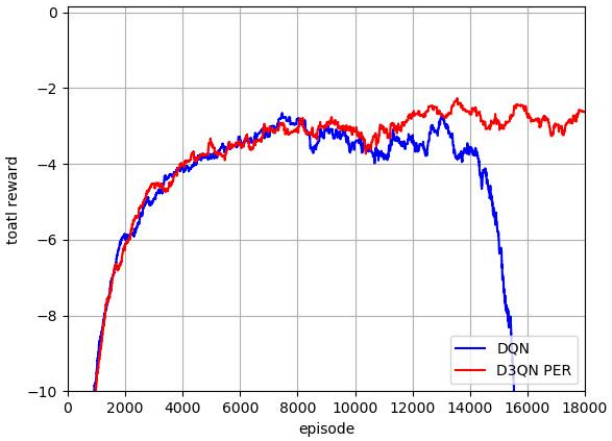

Fig. 6. Training results of deep reinforcement learning algorithms.
그림 6. 심층 강화학습 알고리즘들의 훈련 결과

세로축은 매 에피소드마다 가장 최근의 500개의 에피소드들의 누적 보상의 평균이고, 가로축은 에

피소드를 나타낸다. 모든 에피소드는 500번의 타임 스텝을 가지며 주어진 타임 스텝 내에 로봇이 목표 지점에 도달하지 못한다면 새로운 에피소드가 시작된다. 약 8,000번째 에피소드까지는 DQN과 우선순위 경험 재사용을 사용하는 D3QN 모두 로봇이 목표 지점에 장애물을 회피하며 도달하는 경로를 생성하는 최적 정책을 학습하는 것처럼 보인다. 하지만 DQN은 더 이상 좋은 정책을 찾지 못하고 오히려 평균 보상이 나빠지는 것을 볼 수 있다. 이것은 잘못된 행동-가치함수를 추정하고 있었기 때문이며 이로 인해 최적 정책이 아닌 정책으로 수렴하고 있었기 때문이라고 볼 수 있다. 따라서 DQN이 수렴된 정책을 사용하여 더 이상 최적경로를 생성할 수 없는 환경을 만났을 때 행동-가치함수들이 이상한 값을 가지기 시작했고 에피소드가 진행될수록 다른 모든 행동-가치함수가 잘못된 값을 가지게 되어 더 이상 보상이 크게 되는 행동을 선택하지 못하고 이미 성공적으로 경로를 찾았던 환경들에서도 더 이상 경로를 찾지 못하게 된다. 따라서 DQN의 그래프는 약 14,000번째 에피소드부터 무너지기 시작했다. 반면에 우선순위 경험 재사용을 사용하는 D3QN 알고리즘은 에피소드가 진행될수록 점점 평균 보상이 높아지는 것을 볼 수 있다. 이것은 올바른 행동-가치함수를 추정했고 최적 정책을 향해 계속 정책을 개선했기 때문이라고 볼 수 있다. 결국 우선순위 경험 재사용을 사용하는 D3QN 알고리즘은 주어진 시뮬레이션 환경에서 모바일 로봇의 경로를 생성하는 최적 정책을 발견했다고 할 수 있다. 학습된 우선순위 경험 재사용을 사용하는 D3QN 모델의 파라미터는 표 1과 같다. 이렇게 20,000개의 에피소드를 거쳐 학습을 진행한 우선순위 경험 재사용을 사용하는 D3QN 모델의 가중치를 저장하였고, 시뮬레이션을 통해 모바일 로봇이 경로를 탐색하는 데 사용하였다.

Table 1. Deep reinforcement learning model parameters.
표 1. 심층 강화학습 모델 파라미터

Hyperparameters	Value
Architecture	Conv(32-8×8-4)
	Conv(64-4×4-3)
	Conv(64-3×3-1)
	FullyConnected(512)
	FullyConnected(256)
	FullyConnected(64)
Batch size	128
Start ϵ	1.0
End ϵ	0.01
Annealing step	500000
Memory size	500000
Learning rate	0.0001
Discount rate	0.99
PER α	0.6
PER β	0.4
PER β increase	0.0000025

V. 심층 강화학습 훈련 결과

첫 번째로 실행한 테스트는 장애물이 없을 때 최적경로를 찾아가는지 확인하기 위함이다. 학습을 진행했던 환경에서는 최소 10개에서 최대 20개의 장애물이 존재하게 되는데 학습을 진행하지 않았던 장애물이 하나도 존재하지 않는 환경에서 최적경로를 찾아가는 모습을 그림 7을 통해 확인할 수 있다. 이때 검은색 마름모는 모바일 로봇의 현재 위치를 나타내며 파란색 마름모는 목표 지점을 나타낸다. 그리고 검은색으로 칠해진 사각형들은 모바일 로봇이 지나온 경로를 의미한다.

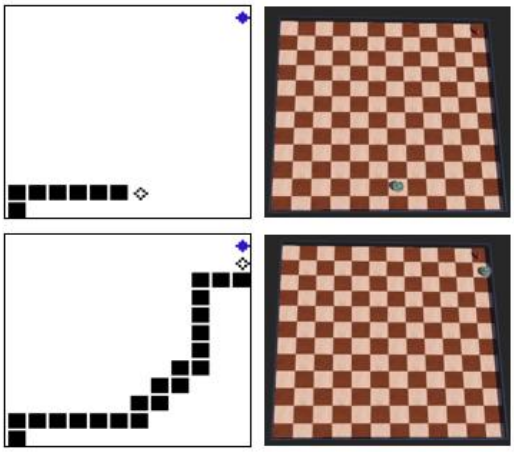

Fig. 7. Path generation when there are no obstacles.
그림 7. 장애물이 없을 때 경로 생성

두 번째로 실행한 테스트는 장애물이 10개 존재할 때 모바일 로봇이 최적경로를 찾아가는지 확인하는 것이다. 그 결과는 그림 8과 같으며, 10개의 장애물이 존재할 때도 장애물을 회피하며 최단 경로를 따라 모바일 로봇이 움직이는 것을 확인할 수 있다.

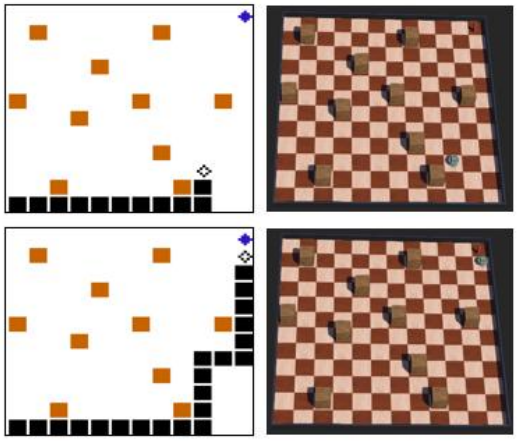

Fig. 8. Path generation when there are 10 obstacles.
그림 8. 10개의 장애물이 있을 때 경로 생성

마지막 테스트는 장애물이 20개 존재할 때 모바일 로봇이 최적경로를 찾아가는지 확인하는 것이다. 장애물이 20개 존재하는 환경은 본 논문에서 제안한 강화학습 알고리즘의 시뮬레이션 환경에서 나타날 수 있는 가장 어려운 환경이다. 테스트 결과는 그림 9와 같으며, 20개의 장애물이 존재할 때도 모바일 로봇이 출발 지점으로부터 목표 지점까지 충돌하지 않고 최단 경로로 도달하는 것을 확인할 수 있다.

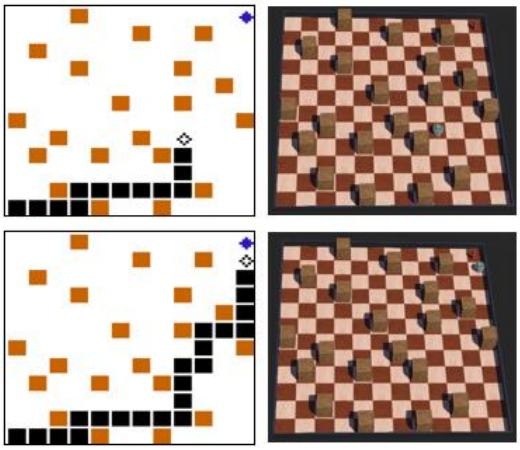

Fig. 9. Path generation when there are 20 obstacles.
그림 9. 20개의 장애물이 있을 때 경로 생성

Ⅵ. 결론

본 논문에서는 장애물이 존재하는 미지의 환경에서 모바일 로봇이 장애물과 충돌하지 않고 주행하도록 하는 최적경로를 생성하기 위한 개선된 강화학습 알고리즘을 제안하였다. 제안된 방법을 사용한다면 스마트 팩토리에서 사용하는 자율주행 로봇이나 실내 서비스 자율주행 로봇 등 실내에서 동작하는 모바일 로봇이 장애물의 위치나 개수 등이 동적으로 변하는 환경에서도 주위 환경의 이미지만 주어진다면 충돌을 회피하며 최단 경로로 출발지점에서 목표 지점까지 이동할 수 있을 것이다. 또한 본 논문에서는 심층 강화학습 훈련 결과를 통해 개선된 DQN 알고리즘의 필요성을 확인하였다. 우선순위 경험 재사용을 사용하는 D3QN보다 보다 더 개선된 학습 능력을 보여주는 심층 강화학습 알고리즘이 구현된다면 더욱 어렵고 복잡한 환경에서도 성공적으로 모바일 로봇의 최적경로를 생성할 수 있을 것이다.

References

[1] K. Zhu and T. Zhang, “Deep reinforcement learning based mobile robot navigation: A review,” in *Tsinghua Science and Technology*, vol.26, no.5, pp.674-691, 2021.
DOI: 10.26599/TST.2021.9010012.

[2] X. Xue, Z. Li, D. Zhang and Y. Yan, “A Deep Reinforcement Learning Method for Mobile Robot Collision Avoidance based on Double DQN,” *2019 IEEE 28th International Symposium on Industrial Electronics (ISIE)*, pp.2131-2136, 2019.
DOI: 10.1109/ISIE.2019.8781522.

[3] Kober, Jens, J. Andrew Bagnell, and Jan Peters. “Reinforcement learning in robotics: A survey,” *The International Journal of Robotics Research,* Vol.32, No.11, pp.1238-1274, 2013.
DOI: 10.1177/0278364913495721

[4] Otte, Michael W. “A survey of machine learning approaches to robotic path-planning,” *University of Colorado at Boulder, Boulder* 2015.
DOI: 10.1109/RISE.2017.8378228

[5] S. Zhou, X. Liu, Y. Xu and J. Guo, “A Deep Q-network (DQN) Based Path Planning Method for Mobile Robots,” *2018 IEEE International Conference on Information and Automation (ICIA)*, pp.366-371, 2018. DOI: 10.1109/ICInfA.2018.8812452.

[6] Mnih, Volodymyr, et al. “Playing atari with deep reinforcement learning,” *arXiv preprint*

arXiv:1312.5602, 2013.
[7] J. Xin, H. Zhao, D. Liu and M. Li, "Application of deep reinforcement learning in mobile robot path planning," *2017 Chinese Automation Congress (CAC)*, pp.7112-7116, 2017.
DOI: 10.1109/CAC.2017.8244061.
[8] X. Ruan, D. Ren, X. Zhu and J. Huang, "Mobile Robot Navigation based on Deep Reinforcement Learning," *2019 Chinese Control And Decision Conference (CCDC)*, pp.6174-6178, 2019.
DOI: 10.1109/CCDC.2019.8832393.
[9] H. Sasaki, T. Horiuchi and S. Kato, "A study on vision-based mobile robot learning by deep Q-network," *2017 56th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, pp.799-804, 2017.
DOI: 10.23919/SICE.2017.8105597.
[10] Van Hasselt, Hado, Arthur Guez, and David Silver. "Deep reinforcement learning with double q-learning," *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol.30, No.1, 2016.
[11] Wang, Ziyu, et al. "Dueling network architectures for deep reinforcement learning." *International conference on machine learning*. PMLR, 2016.
[12] Schaul, Tom, et al. "Prioritized experience replay." *arXiv preprint arXiv:1511.05952,* 2015.
[13] "Webots: robot simulator" https://cyberbotics.com/#features
[14] Kirtas, M., et al. "Deepbots: A Webots-Based Deep Reinforcement Learning Framework for Robotics." *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, Cham, 2020.

BIOGRAPHY

Yung-Min Sunwoo (Member)

2020 : BS degree in Electronic Engineering, Pukyong National University.
2020~present : MS student in Smart Robot Convergence and Application Engineering, Pukyong National University.

Won-Chang Lee (Member)

1983 : BS degree in Instrumentation and Control Engineering, Seoul National University.
1985 : MS degree in Electrical and Electronic Engineering, KAIST.
1992 : PhD degree in Electronic and Electrical Engineering, POSTECH.
1993~present : Professor, Pukyong National University