심층 강화학습의 일반화를 위한 물리기반 3차원 미로게임 환경

박현○ , 한이삭, 백인창, 김경중*
광주과학기술원 융합기술원 융합기술학제학부
{andrew6303@gm., Issac7778@gm., inchang.baek@gm., kjkim@}gist.ac.kr

3D Physics-based Maze Game Environment for Generalization of Deep Reinforcement Learning

Hyun Park○, Isaac Han, In-Chang Baek, Kyung-Joong Kim*
School of Integrated Technology, Gwangju Institute of Science and Technology

요 약1)

3차원 물리 기반 제어 문제는 3차원 물리 환경에서 특정 물체를 조작하는 문제로, 인공지능 분야에서 연구되어 온 도전적인 문제이다. Ball in Maze(BiM) 게임은 대표적인 물리 기반 제어 문제로, 물리직관을 이해하는 능력이 요구된다. 본 논문에서는 심층 강화학습의 성능 평가를 위한 새로운 오픈 소스 Ball in Maze(BiM) 게임을 제안한다. 본 논문에서 제안하는 BiM 환경은 물리 직관을 요구하는 문제로, 다양한 구조를 통해 강화학습 알고리즘의 일반화 능력을 효과적으로 평가한다. 또한 기존 강화학습 알고리즘을 BiM 환경에 적용해 그 성능을 평가하고, 일반화 적용이 가능한지 테스트하여 가능성을 제시하였다.

1. 서 론

물리직관을 필요로 하는 제어 문제(control problem)는 로보틱스 및 인공지능 분야에서 도전적인 문제이다[1]. 최근 딥 러닝과 강화학습을 결합한 심층 강화학습은 많은 물리 직관 게임을 효과적으로 풀 수 있음이 연구되었다[2]. 대표적으로 로봇 조작 문제(Robotic manipulation)[3], 앵그리 버드(Angry bird)[4] 등이 있으며, 이들 중 카트폴 문제와 다양한 로봇 조작 문제는 최근 강화학습을 통해 인간 이상의 수준으로 플레이 할 수 있음이 밝혀졌다.

BiM(Ball in maze) 게임은 판을 기울여 공을 움직이는 게임으로, 미로를 통과하여 공을 목적지까지 도달하는 것이 목적이다. 이러한 BiM 게임을 풀기 위한 연구가 진행되었으며, 이를 위해 현실에서 직접 게임을 플레이 하는 방법이 제안되었다[5]. 하지만 실제 구현을 통해 물리 직관 연구를 진행하는 것은 시간과 노력이 많이 소요된다.

본 연구에서는 새로운 형태의 BiM 게임 시뮬레이션 환경을 제안한다. 본 연구에서 제안하는 BiM 환경은 다양한 구조를 가진 여러 개의 게임으로 구성되어, 강화학습 알고리즘의 일반화 성능을 효과적으로 측정한다. 또한 Unity Engine으로 구현되어 ML Agents등 Unity에서 지원하는 다양한 기능을 활용할 수 있으며, 오픈 소스2)로 공개되어 있어 누구나 개선하여 이용할 수 있다. 본 연구에서 제안하는 BiM환경에서 기존 강화학습 알고리즘의 일반화 성능을 평가한 결과, 일반화 성능이 매우 부족함을 알 수 있었다.

2. 물리 기반 3차원 게임 환경

Unity Engine에서는 게임을 위한 물리 시뮬레이션이 내장되어있다. 이를 이용하여, 본 연구에서는 게임엔진에서 제공되는 물리엔진을 통해 새로운 강화학습 환경을 개발하였다. 개발한 환경은 공이 미로를 통과하여 목적지에 도달하는 게임이고, 물리적인 직관을 필요로 하도록 게임을 구성하였다. 기존의 공을 에이전트로 한 게임 환경과 달리 보드(Board)를 회전시켜 중력을 통해 공을 이동시키는 방식으로 게임이 진행된다.

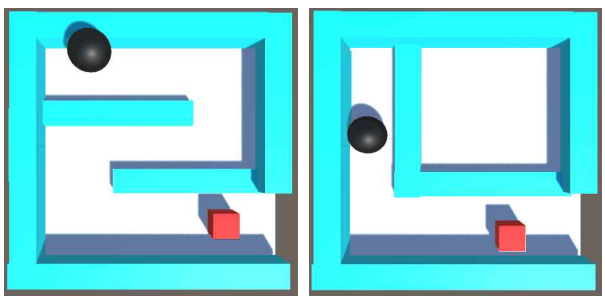

그림 1 물리 기반 3D 강화학습 시뮬레이션 환경

그림 1은 본 논문에서 제안한 시뮬레이션 환경이다. 검은색 공이 미로를 통과하여 빨간색 목적지로 가는 것을 성공으로 가정하였다. 검은색 공이 목적지로 가기 위해서는 보드(Board)를 회전시켜 중력을 통해 공을 이동시키는 방식으로 게임이 진행된다. Unity 3D ML-Agents SDK3) 기반으로 구현하였으며, 강화학습이 가능하도록 에

1) 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (2021R1A4A1030075).
2) https://github.com/hpark6303/Ball-in-Maze

피소드(Episode), 에이전트의 상태(State), 행동 (Action), 보상(Reward)을 정의하였다.

2.1 상태 (State)

물리직관 관련 연구들은 인공지능이 인간처럼 물리를 이해하여 문제를 해결하는 것을 목표로 한다. 따라서, 인공지능이 상황을 인지하는데 있어, 사람과 비슷한 방식으로 구성하였다. 본 게임에서 상태는 크게 두 가지로 나눌 수 있는데, 시각적인 정보에 해당하는 카메라(Camera)와 게임을 해결하기 위해 부가적인 정보를 제공하는 벡터 (Vector) 정보로 나눌 수 있다. 환경에서 사용한 상태 값은 표 1에 정리하였다.

센서 종류	입력 값
카메라 (Camera)	84 × 84 pixels
벡터 (Vector)	- 목적지의 좌표 - 공의 좌표 - 판의 기울기 - 중간지점의 좌표

표 1 상태 종류 및 입력 값

Unity 환경에서 제공한 카메라 화면으로 시각정보를 받아들이는 카메라 센서와 스크립트상에서 제공한 오브젝트들의 정보이다.

보드, 공, 벽, 목표 지점이 보이도록 'Camera Sensor'를 에이전트에 제공하였다. 에이전트는 A_{camera}만큼의 시야 각도를 가진다. 에이전트의 시야에서 분별을 용이하게 하기 위해 공과 벽, 보드의 색을 서로 다르게 설정하였다.

두 번째는 스크립트에서 공의 위치, 타겟의 위치, 보드의 회전 정도, 목표까지의 중간지점을 제공한다. 타겟의 위치와 중간 지점의 위치, 공의 위치는 x, y, z축에 대한 3차원 공간의 벡터 값으로 P_{Ball} = (x_b, y_b, z_b), P_{Target} = (x_t, y_t, z_t), P_{Help} = (x_h, y_h, z_h)와 같이 표현된다. 학습효과를 높이기 위해 중간지점인 P_{Help}를 구성하였다. 보드가 움직이기 때문에 공 이외의 다른 오브젝트들은 보드의 자식 오브젝트(Child Object)로 구성했는데, P_{Help}는 부모 오브젝트인 판을 기준으로 (-0.5, -0.26, -3.6)에 위치하여 그 근처에 도달하면 보상을 주도록 했다.

2.2 행동 (Action)

일정 크기의 각도까지 보드를 회전시키는 행동을 한다. 판의 회전정도는 R_{plane}으로, X축 회전을 R^x_{plane}, Z축 회전을 R^z_{plane}과 같이 표현한다. 판의 과한 이동이 학습을 저해시켜 이를 방지하기 위해 R^x_{plane}<15m, R^z_{plane}<10m으로 제한을 두었다.

3.3 보상 (Reward)

보상함수는 다음과 같다.

1. 목표한 지점과 충돌했을 때 reward = 1000
2. 한 스텝이 진행될 때마다 $S_{present}$ 변수를 늘리고, 그 값이 $S_{\max}$ 이상을 넘어가면 reward = -100
3. P_{Ball} 과 P_{Target}사이의 Z축 거리가 멀수록 보상이 음수가 되도록 reward = Target 위치의 z좌표 - 공 위치의 z좌표

$$\text{Reward} = P^z_{Target} - P^z_{Ball}$$

3. 심층 강화 학습 적용

본 연구에서는 물리기반 3차원 게임을 풀기 위해 Proximal Policy Optimization(PPO)[6] 알고리즘을 적용하였다. PPO 알고리즘은 연속적 제어(continuous control) 문제에 적합한 알고리즘으로, 신뢰 영역(trust region) 방법에 제한된 대체 목표(clipped surrogate objective) 손실함수를 적용해 안정적으로 가중치를 갱신한다. 이러한 PPO 알고리즘은 최근 많은 강화 학습 문제에서 좋은 성능을 보이는 것이 실험적으로 증명되었다.

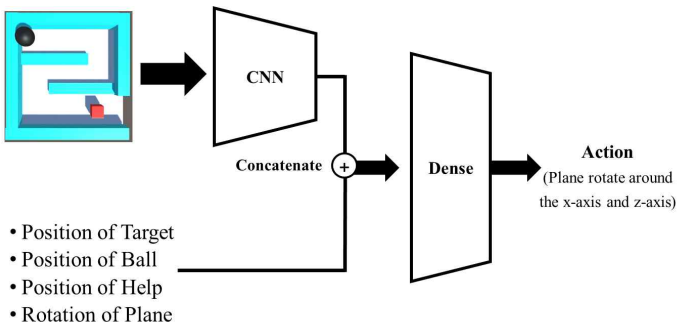

그림 2 인공신경망 구조

CNN층을 통해 이미지에서 특징을 추출하고, 추출된 특징과 게임상의 정보를 결합하여 다층 퍼셉트론에 입력된다. 최종적으로 보드의 X축 회전값, Z축 회전값으로 구성된 행동을 출력한다.

4. 실험결과

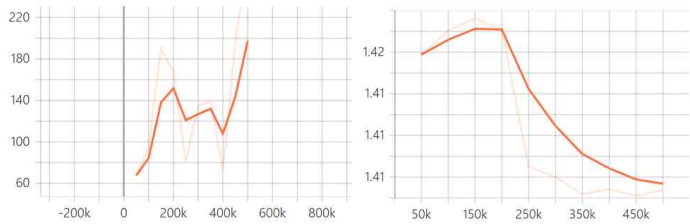

그림 3 학습과정에서 보상(좌)과 엔트로피(우)의 변화

학습을 진행하여 보상(Cumulative Reward)이 계속해서 증가하는 것을 확인할 수 있었다. 정책(Policy)의 엔트로피(Entropy) 또한 지속적으로 감소하는 모습을 볼 수 있다. 정책이 일관적인 방향으로 발전함을 알 수 있다.

3) https://github.com/Unity-Technologies/ml-agents

Batch size	1024	Epsilon	0.2
Buffer size	10240	Lambda	0.95
Learning rate	0.0003	Number of epoch	3
Beta	0.005	Gamma	0.99

표 2 강화학습 모델의 하이퍼 파라미터

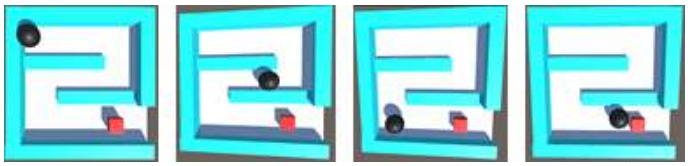

그림 4 학습된 에이전트와 공의 이동

4.2 일반화

학습한 인공지능이 일반화 가능한지 공의 위치(P_{Ball}), 미로의 형태, 타겟의 위치(P_{Target})를 다르게 하여 실험하였다. 여러 미로를 학습한다면 일반화가 더 수월해질 것이라고 생각하여, Maze1 미로에서 학습한 인공지능 (Agent 1)과 네 개의 미로 (Maze 1~Maze 4)를 모두 학습한 인공지능(Agent 2)으로 나누어 새로운 두개의 미로(Maze 5, Maze6)를 잘 해결하는지 평가하였다. 100번의 에피소드를 수행하여 성공률을 평가하였다.

	Maze1	Maze2	Maze3	Maze4	Maze5	Maze6
Agent1	2	70	13	12	10	60
Agent2	2	97	27	20	55	83

표 3 Agent1과 Agent2 일반화 성능 평가 결과
(각 미로에서의 성공률)

4.2.1 Maze1 미로를 사용하여 학습한 인공지능

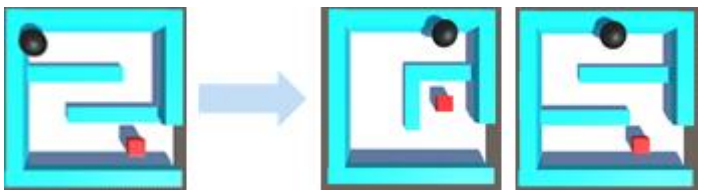

그림 5 에이전트가 학습한 미로(좌),
테스트한 미로 Maze5(중), Maze6(우)

Maze1 미로를 사용하여 학습시킨 에이전트는 처음 맞이한 환경에서도 P_{Target}이 비슷하면 어느 정도 문제를 해결하는 모습을 보였다. 하지만, P_{Target}이 바뀌면 탐색이 잘 되지 않는 문제를 확인할 수 있었다. 처음 학습한 환경의 P_{Target}에 과적합하여 Maze1 에서는 성공률 2%로 Maze2에서는 70%, Maze3 에서는 13%로 모든 수치에서 Agent2보다 낮게 나타나는 것을 볼 수 있다.

4.2.2 네개의 미로를 사용하여 학습한 인공지능

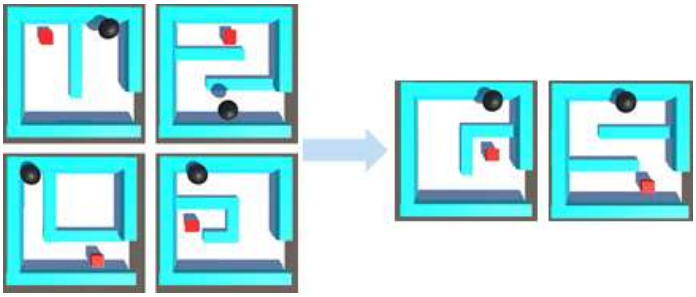

그림 6 에이전트가 학습한 미로 4개(Maze1, Maze2, Maze3, Maze4)(좌), 테스트한 미로 Maze5(중), Maze6(우)

다음과 같이 여러개의 미로를 학습한 인공지능이 새로운 미로를 해결할 때 비교적 잘 해결하는 것을 확인할 수 있었다. 하지만, Maze1에서 2%, Maze3에서 27%, Maze4에서 20%를 나타냄으로써 주어지는 환경에 따라 수행능력에 차이가 크고, 인간 수준 플레이를 하지 못하는 것으로 보인다.

5. 결 론

본 논문에서는 강화 학습의 일반화 성능을 테스트하기 위한 시뮬레이션 환경으로 물리 직관이 필요한 3D 게임 환경을 제시하였다. 본 논문에서 제안하는 환경은 에이전트가 스스로의 물리법칙을 경험하여 학습하던 환경과 달리 에이전트가 중력을 활용해 다른 오브젝트를 이동시키는 특징을 가졌다. 기존의 강화학습 알고리즘을 평가한 결과, 일반화 성능이 매우 약함을 알 수 있었다. 본 환경은 기존의 물리 기반 문제들과 달리 공이 튀거나, 밖으로 나가는 등의 현상이 일어나기 쉽기 때문에 다양한 상황에 일반화하거나, 최단 경로를 구하는 것이 힘든 문제이다.

참 고 문 헌

[1] Davis, Ernest. “Physical reasoning.“ Foundations of Artificial Intelligence 3 (2008): 597-620.

[2] Lillicrap, Timothy P., et al. “Continuous control with deep reinforcement learning.“ arXiv preprint arXiv:1509.02971 (2015).

[3] Todorov, Emanuel, Tom Erez, and Yuval Tassa. “Mujoco: A physics engine for model-based control.“ 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2012.

[4] Renz, Jochen. “AIBIRDS: The Angry Birds artificial intelligence competition.“ Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.

[5] Jha, Devesh K., et al. “Learning Tasks in a Complex Circular Maze Environment.“ (2018).

[6] Schulman, John, et al. “Proximal policy optimization algorithms.“ arXiv preprint arXiv:1707.06347 (2017).