회전형 2단 도립 진자에 대한 DDPG와 TD3 제어 성능 비교

지창훈[1], 임현교[2], 허주성[1], 한연희[1†]

한국기술교육대학교 미래융합공학전공[1], 한국기술교육대학교 창의융합공학협동과정[2]

{koir5660, glenn89, chil1207, yhhan}@koreatech.ac.kr

Comparison of DDPG and TD3 Control Performance for Rotary Double Inverted Pendulum

Chang-Hun Ji[1], Hyun-Kyo Lim[2], Joo-Seong Heo[1], Youn-Hee Han[1]

Future Convergence Engineering Major, Korea University of Technology and Education[1]

Department of Interdisciplinary Program in Creative Engineering, Korea University of Technology and Education[2]

요 약

강화 학습의 환경은 action space 를 기준으로 discrete action space 를 가진 환경과 continuous action space 를 가진 환경으로 나눌 수 있다. continuous action space 를 가진 환경을 학습시키는 강화 학습의 대표적인 알고리즘으로 Deep Deterministic Policy Gradient(DDPG)와 DDPG 의 단점을 보완한 Twin Delayed Deep Deterministic Policy(TD3)가 있다. 본 논문에서는 전통적인 제어 시스템 분야에서 controller 의 성능을 검증하는데 사용되는 Rotary Double Inverted Pendulum(RDIP)시뮬레이션을 활용하여 DDPG 와 TD3 의 실험을 진행한다. 그 후 DDPG 와 TD3 의 성능을 비교 분석하여 RDIP 시뮬레이션 환경 내에서 TD3 가 DDPG 보다 좋은 성능을 보여주고 있음을 확인한다.

Ⅰ. 서론

강화 학습은 행동을 결정하는 에이전트가 주어진 환경과 끊임없는 상호작용으로 에이전트가 얻는 보상을 최대화 시키는 것이 목적인 머신 러닝의 한 종류이다[1]. 강화 학습의 환경은 action space 를 기준으로 discrete action space 와 continuous action space 로 분류할 수 있다. 그 중 continuous action space 를 가진 강화 학습 환경에 적용할 수 있는 대표적인 강화 학습 알고리즘으로 Deep Deterministic Policy Gradient(DDPG)와 DDPG 의 단점을 보안한 Twin Delayed Deep Deterministic policy(TD3)가 있다. 본 논문에서는 기존의 전통적인 제어 시스템 분야에서 controller 의 성능을 검증하는데 활용되는 Rotary Double Inverted Pendulum(RDIP) 환경을 활용하여 DDPG 와 TD3 의 성능을 비교한다.

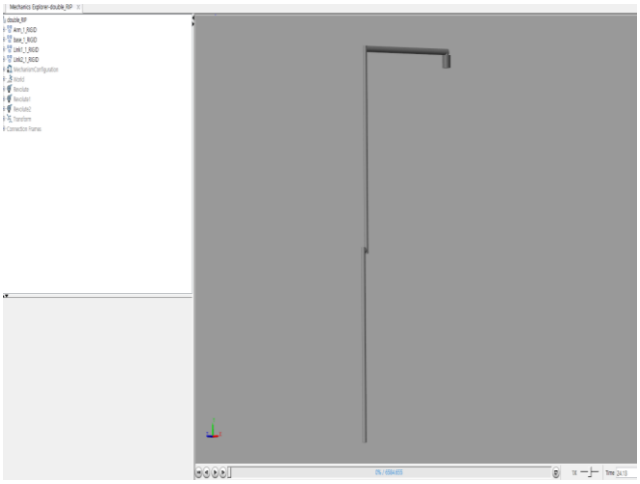
Ⅱ. 본론

1. Rotary Double Inverted Pendulum 환경

Rotary Inverted Pendulum(RIP)는 전통적인 제어 시스템 분야에서 controller 의 성능 검증을 위해 사용된 환경이다. pendulum 을 도립 시키는 것이 목적인 환경으로 주요 구성 요소로는 arm, motor, pendulum 이 있다[2]. 본 논문에서는 제어 난이도를 높여 알고리즘의 성능을 확인하기 위해 기존의 RIP 에서 pendulum 을 한 번 더 연결한 Rotary Double Inverted Pendulum(RDIP) 환경을 활용한다.

RDIP 는 기존의 RIP 와 동일한 구성요소에 pendulum 만 1 단 pendulum(pendulum1)과 2 단 pendulum(pendulum2)로 변경되었다. 목적은 motor 를 제어하여 pendulum1 과 pendulum2 를 동시에 도립 시키는 것이다. 실제 RDIP 를 활용하여 실험을 진행하기에는 현실적으로 많은 제약이 있으므로 본 논문에서는 MATLAB 과 SolidWorks 를 활용한 가상 RDIP 시뮬레이션으로 실험을 진행하였다.

강화 학습 에이전트는 MATLAB 시뮬레이션으로부터 RDIP 의 pendulum1 각도 θ_{p1}, pendulum1 속도 w_{p1}, pendulum2 각도 θ_{p2}, pendulum2 속도 w_{p2}와 arm 각도 θ_a, arm 속도 w_a를 얻는다. 초기 상태일 때 $\theta_{p1}, \theta_{p2}, \theta_a$는 모두 0 이다((그림 1) 참조).



(그림 1) RDIP MATLAB 시뮬레이션 초기 상태

† : 교신저자 한연희(한국기술교육대학교)

2. Deep Deterministic Policy Gradient

강화 학습 에이전트는 환경이나 목적에 따라 다양한 알고리즘을 사용한다. 로봇이나 기계 제어와 같은 환경은 continuous action space 를 가져야 한다. Continuous action space 를 가진 환경을 학습시키는 대표적인 알고리즘으로 Deep-Q-Network(DQN)의 기법들을 actor-critic 에 적용한 Deep Deterministic Policy Gradient(DDPG)가 있다[3].

DDPG 는 DQN 이 학습을 안정화 시키는 기법인 target-network 와 replay buffer 를 이용하여 기존의 actor-critic 과 neural network 를 성공적으로 접목시켰다. Actor 는 연속적인 행동 범위 내에서 가장 높은 행동가치를 지닌 행동을 뽑아내고 목적 함수를 최대화하는 경사 상승법을 수행한다. critic 은 actor 의 결과인 행동을 평가하고 DQN 과 유사하게 loss 를 최소화하는 방향으로 학습된다.

3. Twin Delayed Deep Deterministic Policy

Twin Delayed Deep Deterministic Policy(TD3)는 DDPG 의 단점인 critic 이 과대평가되는 overestimated bias 와 학습이 진행될수록 에러가 누적되는 것을 보완하여 DDPG 에 적용한 알고리즘이다[4].

TD3 는 overestimated bias 를 방지하기 위해 critic network 를 2 개로 설정한다. 그 후 2 개의 critic 결과값 중 낮은 값을 사용하는 것으로 overestimated bias 를 해결한다. 또한 학습이 진행되면서 에러가 쌓이는 것을 막기 위해 actor 와 critic 의 학습 주기를 다르게 하여 actor 의 학습을 지연시킨다.

4. 학습 환경, 상태, 행동, 보상

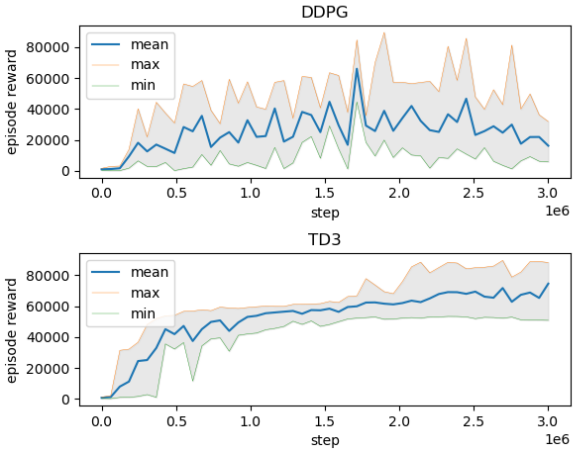
본 논문에서는 MATLAB 의 시뮬레이션 환경과 python 을 연동하여 MATLAB 의 시뮬레이션 결과들을 python 환경에게 실시간으로 전송한다. 그리고 python 에이전트는 MATLAB 시뮬레이션으로부터 받은 정보들을 토대로 행동을 결정해 MATLAB 환경에 전달한다. MATLAB 시뮬레이션이 에이전트에게 전달하는 정보는 pendulum1 각도 θ_{p1} 와 속도 w_{p1} , pendulum2 각도 θ_{p2} 와 속도 w_{p2} , motor 각도 θ_a 와 속도 w_a 로 총 6 개이다. Python 환경은 전달받은 정보들을 바탕으로 상태를 재구성하여 에이전트에게 전달한다(<표 1>참조).

<표 1> 에이전트의 상태, 보상

상태
$(\cos(\theta_{p1}), \sin(\theta_{p1}), w_{p1}, \cos(\theta_{p2}), \sin(\theta_{p2})\sin(\theta_{p2}), w_{p2}, w_a)$
보상
$\theta' + \theta'' - (\lvert w_{p1}\rvert + \lvert w_{p2}\rvert + 1.5 \times w_a)/150$ $\theta' = \begin{cases} 2\pi - \theta_{p1} & if\ \theta_{p1} > \pi \\ \theta_{p1} & else \end{cases}$ $\theta'' = \begin{cases} 2\pi - \theta_{p2} & if\ \theta_{p2} > \pi \\ \theta_{p2} & else \end{cases}$

보상 식에서 θ'와 θ''는 pendulum 이 도립 되어 있을수록 더 높은 값을 받을 수 있도록 θ_{p1}와 θ_{p2}를 이용하여 정의한 것이다. 또한 pendulum 이 움직이지 않고 제자리에 도립 시키기 위해 w_{p1}, w_{p2}, w_a가 작을수록 더 높은 보상 값을 얻을 수 있게 보상 식을 정의하였다. 행동은 actor network 에서 (-1,1) 사이의 값을 받은 후 action scale 을 곱해준다.

5. 실험 결과



(그림 2) 학습 결과 그래프

(그림 2)의 첫번째 그래프는 DDPG 로 실험한 결과이고, 두번째 그래프는 TD3 로 실험한 결과이다. 두 알고리즘 모두 5 번씩 실험을 진행하였다. 두 그래프 모두 파란색 선은 결과값들의 평균이고, 주황선은 결과값들의 최대, 초록선은 결과값들의 최소이다. 그래프를 보면 DDPG 보다 TD3 가 학습이 더 안정적으로 진행되고 진행 속도 또한 더 빠른 것을 확인할 수 있다.

Ⅲ. 결론

본 논문에서는 DDPG 와 TD3 의 성능 비교를 위해 MATLAB 시뮬레이션을 활용한 가상 RDIP 환경에서 실험을 진행하였다. 실험 결과 DDPG 의 단점인 overestimated bias 와 학습이 진행될수록 에러가 누적되는 것을 보완한 TD3 가 성능이 더 좋게 나왔음을 확인하였다. 앞으로 이 실험을 토대로 더 많은 환경에 DDPG 와 TD3 알고리즘 뿐만 아니라 다른 continuous action space 에 적용되는 다양한 알고리즘을 비교하고 분석할 예정이다.

ACKNOWLEGMENT

이 논문은 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행한 기초 연구사업임(No. NRF-2018R1A6A1A03025526).

참 고 문 헌

[1] Mnih et al. "Human-level control through deep reinforcement learning." Nature 518 , no. 7540 (2015): 529--533.

[2] Potsaid et al. "Optimal mechanical design of a rotary inverted pendulum.." Paper presented at the meeting of the IROS, 2002.

[3] Lillicrap et al. "Continuous control with deep reinforcement learning.." Paper presented at the meeting of the ICLR, 2016.

[4] Fujimoto et al. “Addressing Function Approximation Error in Actor-Critic Methods.” International Conference on Machine Learning. PMLR, 2018.