강화학습을 적용한 그네 타기 로봇의 시뮬레이션

최호승[1], 김대원[2], 박희재[1*]

Simulation of swing robot using reinforcement learning

H.S. Choi[1], D.W. Kim[2], H.J. Park[1*]

서울과학기술대학교 기계시스템디자인공학과[1],
서울과학기술대학교 기계설계로봇공학과[2]

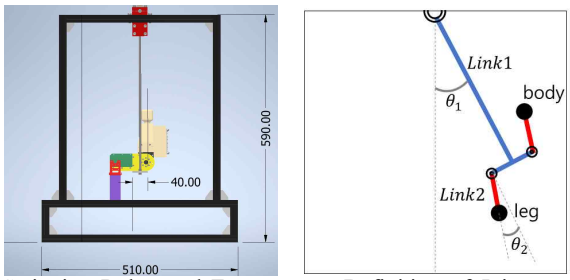
Key Words : Reinforcement learning, Actor-critic algorithm, Dynamics simulation

1. 서 론

본 논문에서는 그네 타기 로봇에 강화학습을 적용하고 그 효용성을 검증하고자 한다. 단순화된 휴머노이드 형태로 2 자유도를 갖는 로봇을 제작하였고 강화학습의 일종인 A2C 알고리즘을 이용하는 방법을 제안한다. 반복 훈련을 통해 로봇이 그네 타기 운동을 하는 방법을 스스로 터득할 수 있음을 시뮬레이션을 통하여 확인하였고 로봇에 적용하는 강화학습의 유효성을 확인하고자 한다.

2. 시뮬레이션 및 훈련

실험에 앞서 **Fig. 1**과 같이 실제 제작이 가능한 단순화된 휴머노이드 형태의 그네 타기 로봇의 3D 모델을 설계한 후 이 로봇에 대한 시뮬레이션을 위한 동역학적 모델을 구하였다.



Swinging Robot and Frames　　Definition of Joints

Fig. 1 Swinging robot

동역학적 시뮬레이션은 OpenAi에서 제공하는 Gym 라이브러리의 Acrobot-v1 환경을 기반으로 진행하였다. **Fig. 1**과 같이 로봇의 다리(Leg)와 상체(Upper body)가 각 자유도를 가지고 있는데 상체를 고정한 경우를 이중 진자 모델로 근사화하여 시뮬레이션을 진행하였다.

강화학습 모델의 상태 공간은 $\cos(\theta_1), \sin(\theta_1), \cos(\theta_2), \sin(\theta_2)$와 정규화된 각속도 $\dot{\theta}_1$, $\dot{\theta}_2$를 사용하고, 행동 공간의 경우 로봇이 다리를 뻗고 있는 자세를 행동 0, 다리를 굽히고 있는 자세를 행동 1로 정의하였다. 연속적인 행동을 하는 것처럼 만들기 위해 Link2의 이전 스텝 각도 θ_2에 따라 상이한 토크를 줄 수 있도록 토크 프로파일을 만들어 적용하였다.

학습을 위하여 사용된 A2C 알고리즘에서는 액터 신경망과 크리틱 신경망으로 분리되는데 액터 신경망의 입력층은 ReLU 활성화 함수를 가지는 은닉층을 거쳐 Softmax 활성화 함수로 정책을 결정한다.

크리틱 신경망은 ReLU 활성화 함수를 적용하여 상태에 대한 가치를 판단하여 훈련이 진행된다. 보상의 척도는 $-|\cos(\theta_1)|$로 정의하여 Link1 각도 θ_1가 수평에서 멀어질수록 더 큰 음의 보상을 받도록 구성하고, 에피소드당 1,000 스텝을 진행하도록 하였다.

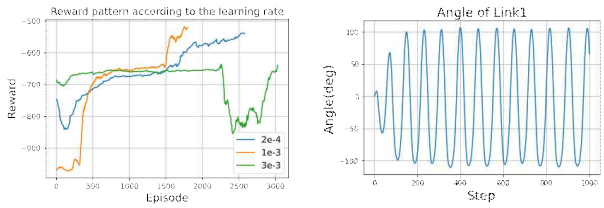
학습의 종료 조건은 한 에피소드 동안 θ_1을 기록하고 에피소드가 끝나면 기록한 각도 데이터를 이용하여 FFT를 수행하고 지배적인 주파수가 0.3Hz 이상 0.5Hz 이하이며 주파수들의 평균 크기보다 100배 이상이 되는 주파수가 검출되면 학습이 종료되었다고 판단하였다.

3. 결 론

훈련 결과 **Fig. 3 (a)**와 같이 중반까지 정체기를 거친 후에 특정 에피소드 부근에서 보상이 급격히 증가하는 양상을 보였다. 훈련이 완료된 모델로 테스트를 수행한 결과 **Fig. 3. (b)**와 같이 이상적인 스윙 동작을 만들어내는 것을 확인함으로써, 신경망을 이용한 강화학습을 통해 로봇이 그네 타기 운동을 하도록 제어할 수 있음을 확인하였다.

다양한 학습률로 학습을 진행한 결과, 0.0002에서 0.003 사이의 값에서 학습이 수렴하고, 그 외의 학습률에서는 수렴을 보장할 수 없는 것을 확인하였다. **Table 1**에서 확인할 수 있듯이 같은 종료 조건에서는 0.001의 학습률을 가질 때 가장 좋은 보상을 얻을 수 있음을 확인하였다.

후속 연구에서는 실제 환경에서 강화학습의 효용성을 확인할 것이다. 그리고 로봇 상체의 자유도를 사용하는 경우, 즉 2 자유도를 갖는 로봇에도 적용하여 비교 연구를 진행할 것이다. 궁극적으로는 보행 및 자세 유지를 위한 로봇제어에 강화학습 방법을 적용하는 연구를 계속할 것이다.



(a) Rewards of episodes　　**(b) Angle of well trained model**

Fig. 3 Learning result

Table 1 Comparison of reward

learning rate	0.0002	0.001	0.003
reward	-550	-487	-599
final episode	2578	1800	3031

참 고 문 헌

(1) Daisuke Uragami, yu Kohno, Tatsuji Takahashi, 2016, *Robotic action acquisition with cognitive biases in coarse-grained state space*, Biosystems. 145, 41-52

(2) Shalabh Bhatnagar et. al. 2009, Natural Actor-Critic Algorithms, automatica, 07.008