

크리틱의 불확실성을 낮추는 액터 목적함수를 통해 강화학습 알고리즘 DDPG 성능 향상에 관한 연구

김형진, 이정우*
서울대학교

kevin909@snu.ac.kr, junglee@snu.ac.kr*

A Study on the Improvement of DDPG Performance by Lowering the Uncertainty of Critic through Objective Function

Kim Hyung Jin, Lee Jung Woo*
Seoul National Univ.

요 약

본 논문은 강화학습의 크리틱의 불확실성을 낮춤으로써 모델 성능 향상을 이끌어내기 위한 목적함수를 설정하였다. 이를 강화학습 알고리즘 DDPG에 적용한 알고리즘을 Mujoco 환경에 실험해본 결과 유의미한 성능 향상을 확인할 수 있었다.

I. 서 론

본 논문에서는 액터-크리틱류의 강화학습 알고리즘에서 크리틱의 불확실성을 낮추는 것이 학습을 향상시킨다는 것을 실험적으로 보였다. 크리틱의 불확실성에 대한 패널티를 액터의 목적함수에 적용하여 크리틱의 불확실성을 줄이는 식의 업데이트 방식을 사용했다. 본 논문에서는 DDPG 알고리즘에 이 아이디어를 적용하여 OpenAI gym Mujoco 의 여러 환경에서 실험을 하였고 이를 통해 강화학습 알고리즘 학습 정도를 향상시킨 것을 확인할 수 있었다.

II. 본론

본 논문에서는 크리틱의 불확실성을 줄이기 위해 액터의 목적함수에 패널티로 추가하였다. 액터-크리틱류의 강화학습 알고리즘에서 보편적으로 적용할 수 있게 기존 알고리즘의 목적함수에 더하는 방식을 채택하였다.[1] 본 논문에서는 액터-크리틱 방식 중 하나인 DDPG에 적용하였다.[2] 본 논문에서 DDPG에 적용한 액터의 목적함수는 다음과 같다.

$$J(\theta) = E[Q_{\omega}(s_t, \mu_{\theta}(s_t)) - \alpha * \sigma]$$

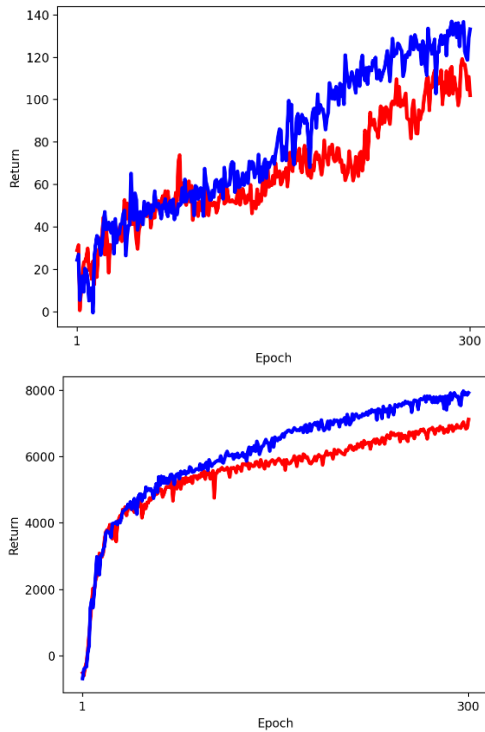
θ 는 액터의 매개변수, s_t 는 시간 t 에서의 상태, ω 는 크리틱의 매개변수, μ_{θ} 는 θ 가 파라미터인 액터, α 는 크리틱의 불확실성을 반영하는 정도를 표현하는 초매개변수, σ 는 크리틱의 불확실성을 뜻한다. σ 는 dropout layer가 포함된 크리틱의 신경망에서 10번 출력한 Q function 값의 표준편차로 구한다. Q function 값의 표준편차가 클수록 크리틱의 불확실성이 크다는 뜻이다. 기존의 DDPG 알고리즘에서 $-\alpha * \sigma$

부분만 더해주는 방식을 택했기 때문에 DDPG가 아닌 다른 액터-크리틱 방식의 강화학습 알고리즘에서도 적용가능한 장점이 있다.

실험은 기존의 DDPG 알고리즘과 크리틱의 불확실성을 적용한 본논문의 알고리즘의 학습 정도를 OpenAI gym Mujoco에서 직접 비교하였다. DDPG와 본논문의 알고리즘의 초매개변수는 모두 동일하게 진행하였다.

Mujoco 환경	기존 DDPG	본논문 알고리즘
Walker2d-v2	2116	2171
Swimmer-v2	98	135
Ant-v2	1414	2000
Hopper-v2	2614	2837
HalfCheetah-v2	7541	8015

실험결과는 위의 표와 같이 나왔다. OpenAI gym Mujoco의 5가지 환경에 대해서 각각 5번씩 독립적으로 실험한 후 테스트했을 때 나온 Return 값의 평균값을 적어놓았다. 5가지 환경에 대해서 모두 기존 DDPG 알고리즘에 비해 성능이 향상되는 것을 확인할 수 있었다. 초매개변수 α 값은 학습이 진행되는 동안 고정하였다. 표에서 실행한 각 환경의 α 값은 1.0, 1.0, 0.1, 0.6, 0.06으로 설정하였다.



각 그래프는 Swimmer-v2 와 HalfCheetah-v2 환경에서 학습되는 정도를 Test Return 으로 표현한 것이다. 빨간 그래프가 기존 DDPG 이고 파란 그래프가 본논문의 알고리즘을 적용한 그래프이다. 이 그래프도 우연히 더 좋은 성능을 보인 것이 아니라는 것을 보여주기 위해서 5 번의 독립된 실험 결과를 평균을 냈다.

III. 결론

본논문에서는 크리티크의 불확실성을 줄이는 것이 액터-크리티크류의 강화학습 알고리즘의 성능을 향상시킬 수 있다는 것을 보이기위해 기존 DDPG 에 적용하여 OpenAI gym Mujoco 의 5 가지 환경에서 실험을 통해 보여주었다. 각 환경에 대해서 5 번의 독립된 실험을 평균낸 결과값을 통해 본논문의 알고리즘의 효과가 있는 것을 보여주었다. DDPG 가 아닌 액터-크리티크류의 강화학습 알고리즘에 모두 적용할 수 있을 것으로 보이므로 활용도가 높을 것으로 예상된다.

ACKNOWLEDGMENT

This work is in part supported by Bio-Mimetic Robot Research Center Funded by Defense Acquisition Program Administration, and by Agency for Defense Development (UD190018ID), MSIT-IITP grant (No.2019-0-01367, BabyMind), Grant(UD190031RD) from Defense Acquisition Program Administration(DAPA) and Agency for Defense Development(ADD), INMAC, and BK21-plus.

참 고 문 헌

- [1] V. R. Konda, *Actor-Critic Algorithms*. PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 2002.
- [2] Lillicrap, Timothy P, Hunt, Jonathan J, Pritzel, Alexander, Heess, Nicolas, Erez, Tom, Tassa, Yuval, Silver, David, and Wierstra, Daan. Continuous control