휴먼형 로봇 손의 사물 조작 수행을 이용한 인간 행동 복제 강화학습 정책 최적화 방법 성능 평가

박나현 [1], 오지헌 [1], 류가현 [1], Edwin Valarezo Añazco[1,3], Patricio Rivera Lopez[1], 원다슬 [2], 정진균 [2], 장윤정 [2], 김태성 [1,2]
[1] 경희대학교 전자정보대학 전자정보융합공학과
[2] 경희대학교 전자정보대학 생체의공학과
[3]Faculty of Engineering in Electricity and Computation, FIEC., Escuela Superior Politécnica del Litoral, ESPOL. Guayaquil, Ecuador.
{nhpark, dhwlgjs3, yugacandy, edgivala, patoalejor, dsrina, ynjngchang, tskim}@khu.ac.kr
wjdwlsrbs77@naver.com

Evaluation of Human Demonstration Augmented Deep Reinforcement Learning Policy Optimization Methods Using Object Manipulation with an Anthropomorphic Robot Hand

Na Hyeon Park[1], Ji Heon Oh[1], Ga Hyun Ryu[1], Edwin Valarezo Añazco[1,3], Patricio Rivera Lopez[1], Da Seul Won[2], Jin Gyun Jeong[2], Yun Jung Chang[2], Tae-Seong Kim[1,2]
[1]Dept. of Electronics and Information Convergence Engineering,
[2]Dept. of Biomedical Engineering, College of Electronics and Information,
Kyung Hee University, Republic of Korea
[3]Faculty of Engineering in Electricity and Computation, FIEC., Escuela Superior Politécnica del Litoral, ESPOL. Guayaquil, Ecuador.

요 약

로봇이 사람과 같이 다양하고 복잡한 사물 조작을 하기 위해서 휴먼형 로봇손의 사물 파지 작업이 필수적이다. 자유도 (Degree of Freedom, DoF)가 높은 휴먼형(anthropomorphic) 로봇손을 학습시키기 위하여 사람 데모(human demonstration)가 결합된 강화학습 최적화 방법이 제안되었다. 본 연구에서는 강화학습 최적화 방법에 사람 데모가 결합된 Demonstration Augmented Natural Policy Gradient (DA-NPG)와 NPG 의 성능 비교를 통하여 행동 복제의 효율성을 확인하고, DA-NPG, DA-Trust Region Policy Optimization (DA-TRPO), DA-Proximal Policy Optimization (DA-PPO)의 최적화 방법의 성능 평가를 위하여 6 종의 물체에 대한 휴먼형 로봇손의 사물 조작 작업을 수행한다. 그 결과, DA-NPG 와 NPG 를 비교한 결과를 통해 휴먼형 로봇손의 사물 조작 강화학습에 행동 복제가 효율적임을 증명하였다. 또한, DA-NPG 는 DA-TRPO 와 유사한 성능을 보이면서 모든 물체에 대한 사물 파지에 성공하여 가장 안정적이었다. 반면, DA-TRPO 와 DA-PPO 는 사물 조작에 실패한 물체가 존재하여 불안정한 성능을 보였다. 본 연구에서 제안하는 방법은 향후 실제 휴먼형 로봇에 적용하여 휴먼형 로봇 손의 사물조작 지능 개발에 유용할 것으로 전망된다.

1. 서론

로봇이 사람 중심의 환경에서 물체를 재배치하거나 문 손잡이를 여는 등의 정교한 일을 수행하기 위해서는 휴먼형(anthropomorphic) 로봇의 파지 작업이 필수적이다. 휴먼형 로봇 손은 자유도(Degree of Freedom, DoF)가 높아 최근 로봇 사물 조작 강화학습에 Natural Policy Gradient (NPG) [1], Trust Region Policy Optimization (TRPO) [2], Proximal Policy Optimization (PPO) [3] 등 다양한 최적화 방법이 제시되었다.

그러나 로봇 손을 학습시키더라도 sample complexity 와 절대적인 학습 시간이 기하급수적으로 늘어나는 문제점이 존재하여, 이를 줄이기 위해 로봇 손 관절의 자유도를 줄이는 등의 제한이 존재하였다 [4]. 이러한 문제점을 해결하기 위해 강화학습에 사람 데모(human demonstration)를 결합시키는 행동 복제 (behavior cloning, BC)방법이 제안되었다 [5], [6]. 사람 데모는 사람의 물체 파지 정보를 측정한 것으로, 이

를 강화학습에 적용하면 문제점이었던 sample complexity 와 학습시간이 현저하게 줄어든다. 또한 사람 데모가 사람이 물체를 잡는 손 모양에 대한 정보를 제공하기 때문에 로봇 손은 물체를 사람처럼 자연스럽게 잡을 수 있도록 학습된다는 등의 장점이 있어, 최근 Demonstration Augmented Policy Gradient(DAPG) 등의 사람 데모와 결합한 강화학습의 다양한 정책 최적화 방법이 발표되었다 [5].

본 연구에서는 강화학습에 있어 사람 데모의 효용성 평가를 위하여 DA-NPG 와 NPG 의 성능을 비교하고, 강화학습의 NPG, TRPO, PPO 의 정책 최적화 방법에 사람 데모를 결합한 DA-NPG, DA-TRPO 및 DA-PPO 의 각 정책 최적화 성능을 휴먼형 로봇 손의 6 종 사물 조작(파지 및 재배치) 작업의 수행을 통하여 평가하였다.

2. 방법

높은 DoF 를 가지는 로봇 손의 사물 조작(파지 및 재배치)은 강화학습만으로는 사람과 유사하게 물체를 집을 수 있는 자연스러운 손동작을 수행하기가 어렵다. 따라서 본 연구에서는 행동 복제와 더불어 사람 데모를 정책 최적화 방법에 결합시켜 DA-NPG 와 NPG 의 성능을 비교하여 사람 데모의 효용성을 입증하고, NPG, TRPO 및 PPO 정책 최적화 방법에 사람 데모를 결합한 DA-NPG, DA-TRPO 및 DA-PPO 의 성능을 평가한다.

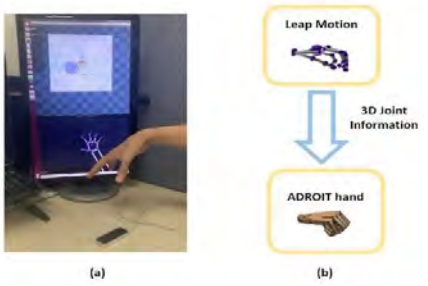
A. 사람 데모 측정 시스템

본 연구에서는 행동 복제를 시행하기에 앞서 각 물체마다 사람 데모를 측정했다. 사람 데모를 측정하기 위해 적외선 LED 센서를 이용하여 손과 각 관절의 위치 정보를 x, y, z (mm) 로 제공하는 Leap Motion 센서[7]와 로봇 시뮬레이션 패키지인 MuJoCo (Multi-Joint dynamics with Contact) 시뮬레이터 [8]에서 모델링한 ADROIT 로봇 손을 실시간으로 연동하였다 [9]. ADROIT 로봇 손은 손의 위치 정보와 각 관절의 각도를 입력 받아서 움직이는데 Leap Motion 센서의 손 모양과 ADROIT 로봇 손을 일치시키기 위해서 3D 관절 위치 정보를 각도로 변환한 후 각 관절의 조절 범위에 맞춰서 각도를 정규화 했다. 그림 1 에 행동 복제 환경을 도시하였다.

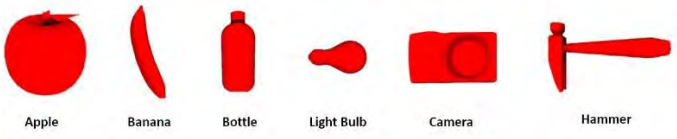
B. Human Demonstration Augmented DRL Policy Evaluation System

본 연구에서 시행한 로봇 손 사물 재배치 및 파지 작업에 사용한 물체들의 3D 모델은 그림 2 에, 시뮬레이션의 전체적인 과정은 그림 3 에 도시했다.

물체의 3D 모델을 이용하여 물체에 대한 사람 데모를 만들고 만든 사람 데모로 행동 복제를 시행하여 weight 값을 초기화해준다 [5]. 로봇 손은 주어진 환경과 상호작용하여 보상을 높이는 방향으로 정책을 최적화하는데, 이 과정에서 사람 데모를 입력하여 로봇 손의 학습을 돕는다. 여기서 정책을 최적화하는 방법은 DA-NPG, DA-TRPO 및 DA-PPO 세 가지로 구분되며, 학습이 끝나면 최적화 정책이 도출된다.



(그림 1) (a) 사람 데모 측정 시스템 (b)ADROIT 로봇 손과 사람의 손 모양을 동기화 하기 위해 Leap Motion 센서가 손 위치와 관절 정보를 실시간으로 ADROIT 로봇 손에 전달한다.



(그림 2) 로봇 손 사물 조작 (파지 및 재배치)에 사용한 6 종 물체의 3D 모델.

C. DA-NPG

NPG 는 기존의 standard gradient descent rule 에 기반한 Policy Gradient(PG) 방법에 Natural Gradient 기법을 접목시켜 steepest descent direction 으로 학습이 진행될 수 있도록 한 방법이다 [1]. NPG 의 gradient 식을 DA-NPG 로 확장하면 다음과 같다 [5].

$$\boldsymbol{g_{aug}} = \sum_{(s,a)\in\rho_\pi} \nabla_\theta ln\pi_\theta(a|s) A^\pi(s,a) + \sum_{(s,a)\in\rho_D} \nabla_\theta ln\pi_\theta(a|s) w(s,a) \quad (1)$$

$$w(s,a) = \lambda_0 \lambda_1^k \max_{(s',a')\in\rho_\pi} A^\pi(s',a') \;\; \forall (s,a) \in \rho_D \quad (2)$$

식 (1)은 정책 π의 데이터셋 ρ_π와 사람 데모의 데이터셋 ρ_D에 대한 부분으로 나눠진다. 데이터셋은 한 시점에 대한 상태(state, s)와 현 상태에서 로봇 손이 취하는 행동(action, a)의 상태-행동 쌍(s, a)을 포함하며, $A^\pi(s,a)$ 는 정책 π 에 대한 어드밴티지 함수이다. $w(s,a)$는 사람 데모 가중치 함수이며 식 (2)의 λ_0와 λ_1은 hyperparameter, k 는 iteration 을 의미한다. $\lambda_0 = 1.0$, $\lambda_1 = 0.95$ 로 설정함에 따라 iteration 이 커질수록 $w(s,a)$ 의 값이 작아져 사람 데모에 대한 가중치가 줄어든다.

(그림 3) 로봇손 사물 조작(파지 및 재배치) 학습을 위한 전체 강화학습 과정.

D. DA-TRPO

TRPO 는 NPG 방법에 Trust Region 이라는 constraint 를 추가하여 정책의 변화 정도를 제한한 기법이다 [2]. TRPO 의 surrogate objective function 은 다음과 같다.

$$\underset{\theta}{\text{maximize}}\ \hat{\mathbb{E}}_t\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}\hat{A}_t\right] \quad (3)$$

$$\text{subject to}\ \hat{\mathbb{E}}_t\left[KL\left[\pi_{\theta_{old}}(\cdot|s_t), \pi_\theta(\cdot|s_t)\right]\right] \le \delta$$

여기서 KL 은 KL divergence 로 업데이트 하기 전의 정책과 새로운 정책 간의 변화 정도를 측정하며, δ는 Trust Region 의 제한 범위를 나타낸다. DA-TRPO 는 식 (3)의 objective function 의 gradient 에 식 (1)을 적용시킨다.

E. DA-PPO

PPO 는 clipped surrogate objective 를 통해 TRPO 의 장점은 유지하면서 TRPO 의 단점인 복잡한 계산과 sample complexity 를 줄인 방법이다 [3]. PPO 의 surrogate objective function 은 다음과 같다.

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}\hat{A}_t \quad (4)$$

$$\underset{\theta}{\text{maximize}}\ \hat{\mathbb{E}}_t\left[min\left(r_t(\theta), clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right] \quad (5)$$

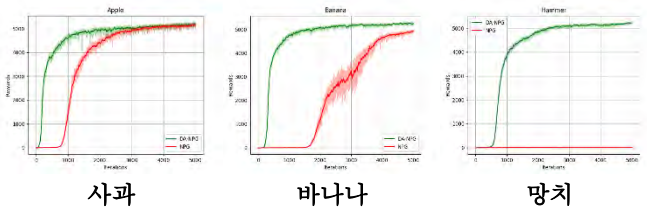
여기서 ϵ은 hyperparameter 로, Trust Region 의 제한 범위를 지정해주는 δ와 유사한 역할을 한다. 본 연구에서는 [3]에 따라 0.2 로 지정하였다. DA-PPO 는 식(4)의 probability ratio function 에 식 (1)을 적용시킨다.

3. 결과 및 분석

A. DA-NPG vs. NPG 의 성능 비교

그림 4 의 그래프는 DA-NPG 와 NPG 의 사물 조작 강화학습 보상 그래프를 보여준다. 그래프를 보면 DA-NPG 와 NPG 의 성능은 사과 에서는 유사하고 바나나는 DA-NPG 가 우세한 경향을 보이며, 망치는 DA-NPG 만이 사물 파지 및 재배치 작업에 성공한다. 전반적으로 DA-NPG 가 NPG 에 비해 성능이 뛰어나며, 학습에 필요한 시간도 적게 걸린다.

DA-NPG 와 NPG 는 사물 조작 작업을 수행할 때 로봇 손이 물체를 잡는 모양에서도 차이가 난다. 그림 5 는 3 종 사물 조작 작업에 대한 학습이 완료된 로봇 손이 사물을 잡는 손 모양을 보여준다. NPG 로 학습시킨 로봇 손이 사과와 바나나를 잡는 모습은 사람이 잡는 모습과 달리 관절이 부자연스럽지만 DA-NPG 로 학습시킨 로봇 손은 사람과 같이 안정적으로 물체를 잡는다. 사람 데모는 최적화 정책의 최종 보상을 향상시키고 학습시간을 줄이며 로봇 손이 사람과 같은 자연스러운 손동작이 가능하게 한다는 것을 알 수 있다.



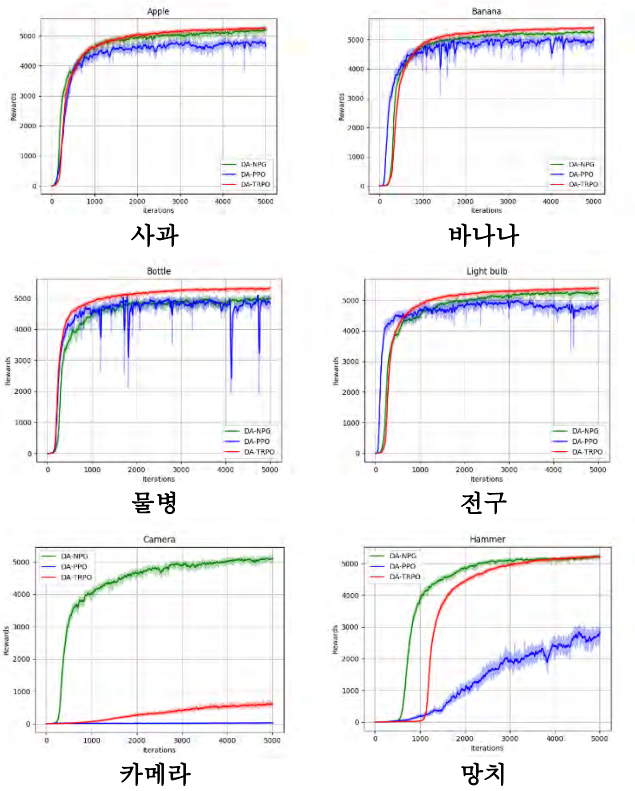
(그림 4) DA-NPG(초록색 선), NPG(빨간색 선)의 3 종 물체에 대한 사물 조작 강화학습 보상 그래프

(그림 5) DA-NPG 와 NPG 학습 후 최종 사물 조작 작업 결과

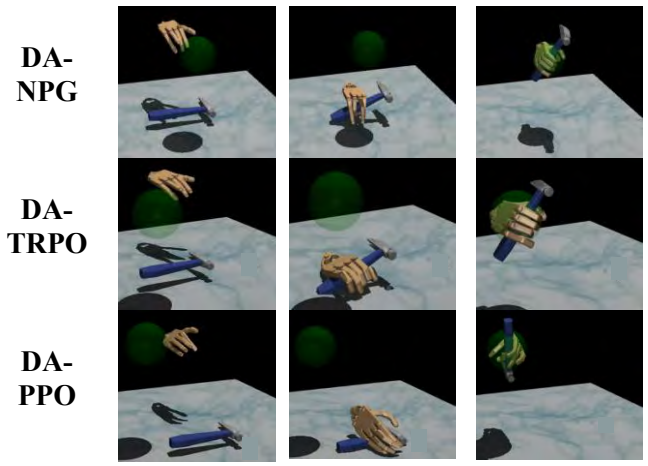
B. 정책 최적화 방법에 따른 성능 비교

그림 6 은 정책 최적화 방법에 따른 6 종 사물 조작 작업의 보상 그래프를, 그림 7 은 망치 조작 작업의 시뮬레이션 결과를 도시했다. 그래프를 보면 전반적인 물체들에 대해 DA-TRPO 의 성능이 가장 높고 안정적이며 그림 7 과 같이 사물을 잡는 손 모양도 자연스럽지만 카메라 조작 작업에 실패하였고, DA-NPG 는 DA-TRPO 와 비슷한 성능을 보이며 모든 사물 조작에 대하여 평균 99.33%의 성공률을 보였다. DA-PPO 는 바나나, 병, 전구 등의 작업에 대해서 비교적 빠르게 학습되는 경향을 보였으나 카메라 작업에 실패하였고,

DA-NPG 와 DA-PPO 에 비하여 정책이 업데이트 되는 과정에서 받는 보상의 그래프 경향이 가장 불안정한 것을 확인할 수 있다. 따라서, 정책 최적화 알고리즘 중 성능이 가장 안정적인 것은 DA-NPG 이었다.



(그림 6) DA-NPG (초록색 선), DA-TRPO(빨간색 선), DA-PPO(파란색 선)의 6 종 물체에 대한 사물 조작 강화학습 보상 그래프.



(그림 7) DA-NPG, DA-TRPO, DA-PPO 의 학습 후 망치 사물 조작 결과. 왼쪽에서부터 시간 순서대로 로봇 손이 망치를 잡는 과정을 보여준다.

4. 결론

본 연구에서는 DA-NPG 와 NPG 의 학습 성능을 비교하여 사람 데모의 효용성을 입증하고, 행동 복제 방법과 DA-NPG, DA-TRPO, DA-PPO 를 이용한 강화학습으로 학습시킨 로봇 손이 수행하는 6 종 물체에 대한 사물 조작 작업의 학습 성능을 평가하였다. DA-NPG 가 NPG 보다 성능 수치와 더불어 물체를 잡는 손 모양 또한 자연스러웠으며, DA-NPG 는 DA-TRPO 와 유사한 성능을 보이면서 일부 사물에 대해 작업에 실패한 DA-TRPO 와 DA-PPO 와 달리 모든 물체에 대한 작업에 성공하였기 때문에 그 성능이 가장 안정적이었다.

ACKNOWLEDGEMENTS

이 논문은 2019 년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (2019R1A2C1003713). 이 논문은 2020 년도 정부(미래창조과학부)의 재원으로 한국연구재단 –현장맞춤형 이공계 인재양성 지원사업의 지원을 받아 수행된 연구임(No. 2017H1D8A1031522). 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 디지털콘텐츠원천기술개발사업의 연구결과로 수행되었음 (IITP-2017-0-00655).

참고문헌

[1] S. Kakade. “A Natural Policy Gradient”. Nerual Information Processing Systems (NIPS), 2001, 14:1531-1538.

[2] J. Schulman, S. Levine, P. Abbeel, M. Jordan, P. Moritz. “Trust Region Policy Optimization”. Proceedings of the 32nd International Conference on Machine Learning, PMLR 2015, 37:1889-1897.

[3] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov. “Proximal Policy Optimization Algorithms”. arXiv:1707.06347v2 [cs.LG]

[4] S. Gu, E. Holly, T. Lillicrap, S. Levine, "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates," 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 2017, pp. 3389-3396.

[5] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, S. Levine. “Learning Complex Dexterous Manipulation with Deep Reinforcement Learning and Demonstrations”. arXiv:1709.10087v2 [cs.LG]. 2018.

[6] A. Gupta, C. Eppner, S. Levine, P. Abbeel, "Learning dexterous manipulation for a soft robotic hand from human demonstrations," 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, 2016, pp. 3786-3793.

[7] https://www.ultraleap.com/

[8] E. Todorov, T. Erez and Y. Tassa, "MuJoCo: A physics engine for model-based control" 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura, 2012, pp. 5026-5033.

[9] V. Kumar, Z. Xu and E. Todorov, "Fast, strong and compliant pneumatic actuation for dexterous tendon-driven hands," 2013 IEEE International Conference on Robotics and Automation, Karlsruhe, 2013, pp. 1512-1519.