심층 강화 학습을 통한 사람형 로봇손의 롱-호라이즌 주방 사물 조작

정진균, 정환석, 이진혁, 오지헌, Ismael Espinoza, Channabasava Chola, 김태성
경희대학교 전자정보대학 전자정보융합공학과

{wjdwlsrbs77, hwan136, qlqjs3647, dhwlgjs3, inespinoza24, channabasavac7, tskim}@khu.ac.kr

Solving Kitchen Long Horizon Tasks with an Anthropomorphic Robot Hand via Sub-Goal Network-based Deep Reinforcement Learning

Jin Gyun Jeong, Hwanseok Jung, Jin Hyuk Lee, Ji-Heon Oh, Ismael Espinoza, Channabasava Chola, Tae-Seong Kim
Dept. of Electronics and Information Convergence Engineering, College of Electronics and Information, Kyung Hee University.

요 약

본 연구에서는 단일 지능으로 다수의 이종 태스크를 순차로 수행하는 롱 호라이즌 작업 (Long Horizon Tasks) 학습 문제를 해결하기 위하여, 숏 호라이즌 (Short Horizon) 지능 기반의 데모 라이브러리와 서브 골 네트워크 (Sub-Goal Network)를 심층강화학습에 결합하여 해결하였다. 사람형 로봇 손을 부엌 가상 환경에서의 롱 호라이즌 작업을 강화학습을 통해 학습하고, 작업 수행 성능을 검증하였다. 학습한 롱 호라이즌 태스크 정책은 4 가지 태스크 조합으로 90.2%의 성공률을 달성하였다.

I. 서 론

로봇 손을 통해 한 번에 하나의 물체를 옮기고 조작하는 단순한 작업을 넘어, 단일 지능으로, 다수의 이종 작업을 순서대로 수행하는 롱 호라이즌(Long Horizon)작업은 최근 로봇 손 강화학습 연구의 주요 주제 중 하나이다. 최근 헬스케어 및 라이프 케어 서비스업 로봇의 수요 증가로 인해 하나의 로봇이 여러 작업을 연속으로 수행하는 지능이 요구되고 있다. 그러나, 기존의 한 번에 하나의 태스크(Task)만 수행하는 강화학습 방법인 숏 호라이즌 (Short Horizon) 태스크 강화학습은 여러가지 태스크를 연속적으로 수행하는 경우, 다양한 환경 변수의 변화에서 다음으로 수행해야 하는 작업을 찾지 못하기 때문에 연속적인 작업을 처리할 수 없다. 따라서 단일 로봇 지능이 변화하는 환경변수에 대응하여 작업이 완료된 후, 다음 작업 목표를 추정하여 여러 태스크를 연속적으로 수행할 수 있는 롱 호라이즌 태스크 강화학습이 필요하다.

본 논문에서는 사람형 로봇 손을 사용하여 롱 호라이즌 태스크 수행 지능을 개발한다. 그러나 기존의 Demo Augmented Policy Gradient(DAPG)[1] 방법만을 사용한 롱 호라이즌 태스크 정책의 학습은 마지막 태스크까지 수렴하지 못한다[2]. 따라서 이 DAPG 강화학습 알고리즘에 더해 숏 호라이즌 태스크 데모 레이블링과 서브 골 네트워크(Sub-Goal Network)를 적용하여 롱 호라이즌 태스크 학습을 수행한다. 우선, DAPG를 활용한 숏 호라이즌 태스크 수행 지능을 기반으로 학습된 숏 호라이즌 태스크 데모 라이브러리를 생성한다. 이 라이브러리를 기반으로 로봇 손의 행동 목표(Sub-Goal)을 추정하는 서브 골 네트워크를 학습시킨다. 이후, 이 데모 라이브러리와 서브 골 네트워크를 통해 롱 호라이즌 DAPG (Long-Horizon DAPG)를 구현하여 롱 호라이즌 태스크 강화학습을 수행한다.

II. 연구방법

본 연구는 MuJoCo[3] 로봇 물리 시뮬레이션 환경에서 Adroit[4] 로봇 손을 활용하여 숏 호라이즌 및 롱 호라이즌 태스크 정책의 강화학습을 수행한다.

A. Short-Horizon DAPG

숏 호라이즌 태스크를 수행하는 사람의 손 동작 데모 생성은 Leap Motion Controller[5]로부터 제공된 사람의 손 관절 위치 정보를 활용하였다. 생성된 사람 손 데모는 DAPG 알고리즘에 사용하여 4 가지 숏 호라이즌 태스크 정책을 학습하였다. 숏 호라이즌 태스크 정책은 부엌환경에서 4 가지 작업(캐비닛 열기, 경첩 문 열기, 바나나 옮기기, 주전자 옮기기)으로 구성되어 있다.

B. Long-Horizon DAPG

B.1 숏 호라이즌 태스크 데모 레이블링

학습한 숏 호라이즌 태스크 정책을 통해 숏 호라이즌 태스크 데모 라이브러리를 생성하고, 롱 호라이즌 태스크를 학습하기 위한 데모 데이터의 레이블링 과정을 진행한다. 학습될 정책이 효과적으로 구분할 수 있도록 서브 골에 따라 각 태스크 별로 환경변수를 다르게 레이블링(Labeling)하고 각 태스크 별로 총 3 개의 서브 골을 갖는다. 서브 골은 물체에 접근, 태스크 수행, 태스크 완료로 구성되어 있다.

B.2. 서브 골 네트워크(Sub-Goal Network)

서브 골 네트워크는 순서가 연결된 데모 데이터로부터 환경변수를 입력으로 받아 서브 골을 출력으로 하는 다층 퍼셉트론(Multi-Layer Perceptron) 네트워크이다. 학습 시 현재의 환경변수를 입력으로 받아 다음 수행해야 할 서브 골을 출력하는 역할을 한다.

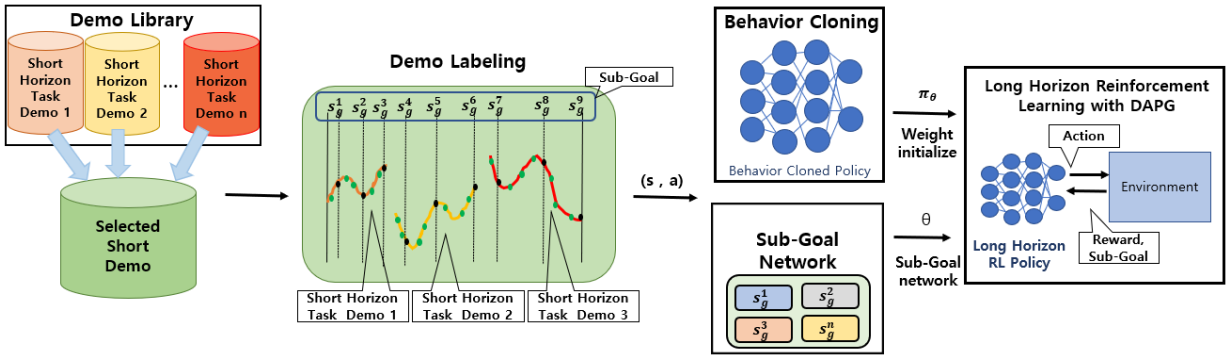

그림 1. 숏 호라이즌 태스크 데모와 서브 골 네트워크를 활용한 롱 호라이즌 태스크 DAPG 강화학습 진행 흐름도

B.3. 롱 호라이즌 태스크 학습 지능

숏 호라이즌 태스크 데모 라이브러리를 통해 롱 호라이즌 태스크를 학습하는 과정을 **그림 1** 에서 볼 수 있다. 우선, 데모 라이브러리에서 학습하려는 롱 호라이즌 태스크에 맞는 데모를 선택한다. 이후, 선택한 데모를 각 태스크 별로 레이블링하고 롱 호라이즌 태스크 시퀀스에 맞게 연결한다. 다음으로, 레이블링 된 데모는 행동 복제를(Behavior Cloning)에 사용되어 학습 정책의 파라미터를 초기화하고 동시에 서브 골 네트워크를 학습하기위해 사용된다. 마지막으로, 행동 복제로 초기화된 정책과 서브 골 네트워크를 통해 주어지는 롱 호라이즌 태스크 시퀀스를 바탕으로 DAPG 강화학습을 수행한다. 강화학습은 주어진 롱 호라이즌 태스크 시퀀스를 이상적으로 수행할 수 있도록 학습되며 이를 위해 태스크를 성공했을 때 보상함수가 극대화되는 방향으로 설계되었다.

III. 연구결과

본 논문에서 제안한 서브 골 네트워크와 DAPG 알고리즘으로 4 가지의 숏 호라이즌 태스크를 조합한 롱 호라이즌 태스크 정책을 학습시켰을 때, **그림 2** 과 같이 90.2% 의 성공률로 해당 태스크를 완료하였다. 태스크는 주전자 옮기기, 바나나 옮기기, 캐비닛 열기, 경첩 문 열기 순으로 학습되었다. **그림 3** 은 앞서 훈련한 **그림 2** 의 롱 호라이즌 태스크 정책의 수행 전과 후로 비교하였다.

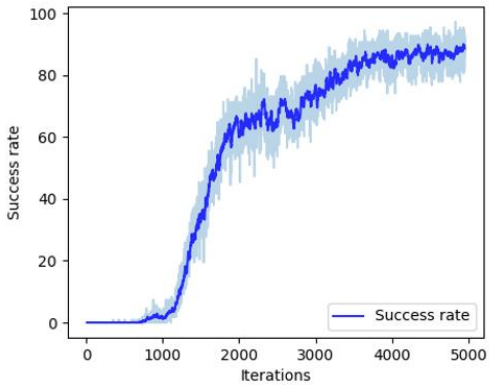

그림 2. DAPG 롱 호라이즌 태스크 강화학습으로 학습한 정책의 iteration 에 따른 성공률 그래프. ① 주전자 옮기기, ② 바나나 옮기기, ③ 캐비닛 열기, ④ 경첩 문 열기순으로 학습하였다.

IV. 고찰 및 결론

본 논문에서는 숏 호라이즌 태스크 데모와 서브 골 네트워크를 DAPG 강화학습 알고리즘에 적용하여 롱 호라이즌 태스크를 학습하고 수행할 수 있음을 확인하였다.

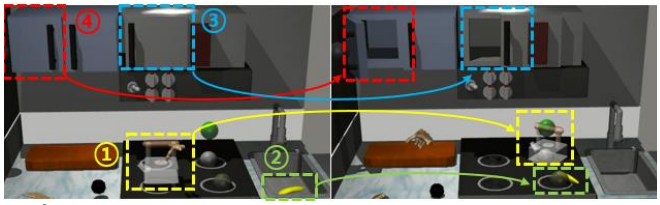

그림 3. 그림 2 에서 학습한 롱 호라이즌 태스크 정책의 수행 전(왼쪽)과 후(오른쪽)의 환경 변화. ① 주전자 옮기기, ② 바나나 옮기기, ③ 캐비닛 열기, ④ 경첩 문 열기순으로 학습하였다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 디지털 콘텐츠 원천기술개발사업의 연구결과로 수행되었음 (IITP-2017-0-00655). 이 논문은 2019 년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(2019R1A2C1003713)

참 고 문 헌

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

[2] A. Gupta, V. Kumar, C. Lynch, S. Levine, and K. Hausman, “Relay Policy Learning: Solving Long-Horizon Tasks via Imitation and Reinforcement Learning,” Oct. 2019, (http://arxiv.org/abs/1910.11956)

[3] 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, pp.5026-5033, 2012.

[4] 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, pp. 1512-1519, 2013.

[5] Leap Motion [Internet], https://www.ultraleap.com/