로봇 협업 행동을 제어하기 위한 순차적 작업 설계 및 학습에 관한 연구

김성현, 노삼열, 장인국
한국전자통신연구원

{kim-sh, samuel, ingook}@etri.re.kr

A Study on the Sequential Task Design and Learning to Control Robot Collaborative Behavior

Seonghyun Kim, Samyeul Noh and Ingook Jang
Electronics and Telecommunications Research Institute

요 약

최근, 강화학습은 단일 에이전트 환경의 연구 발전을 토대로 다중 에이전트 환경에 대한 연구가 지속적으로 진행되고 있다. 다중 에이전트 강화 학습에는 에이전트 사이의 인터랙션으로 인한 Non-stationary 문제가 필수적으로 다뤄지고 있다. 본 논문은 실제 환경과 유사한 다중 로봇 조작 문제를 다룬다. 로봇의 협업에 중점을 두기 위해 순차 조작 작업을 핸드오버 작업으로 설계하고, 순차적 협업에서 발생하는 도전적인 이슈에 대하여 논의한다.

Ⅰ. 서 론

최근 게임, 자율주행, 로봇공학 등 많은 분야에서 딥러닝을 활용한 강화학습 기술이 연구되고 있다. 이러한 강화학습 기술의 발전은 주로 단일 에이전트 환경에서 연구되어온 반면, 다중 에이전트 강화 학습은 단일 에이전트 강화 학습에 비해 상대적으로 많이 연구되지 않았다. 본 논문에서는 실제 문제와 상당히 유사한 다중 에이전트 환경을 고려하기 위해 다중 로봇 조작 문제를 다룬다. 로봇 제어는 실제 환경의 문제 중에서도 난이도가 높은 복잡한 문제에 해당한다 [1]-[2]. 로봇의 협업에 중점을 두기 위해 순차적 조작 작업을 설계하고 순차적 협업에서 발생하는 도전적인 이슈에 대하여 논의한다.

Ⅱ. 본론

단일 에이전트 강화 학습에서는 에이전트가 스스로 문제를 효과적으로 해결하는 것에 초점을 두고 있고, 다중 에이전트 강화 학습에서는 에이전트가 다른 에이전트와 협력하거나 경쟁하면서 문제를 이해하고 해결하는 것을 목표로 한다. 단일 에이전트 강화 학습과 다중 에이전트 강화 학습의 주요 차이점은 Non-stationary 문제의 존재이다. 즉, 다른 에이전트의 행동이 에이전트의 행동에 영향을 미치기 때문에 다른 에이전트에 의해 상태와 행동의 전이 확률이 변경된다 [3].

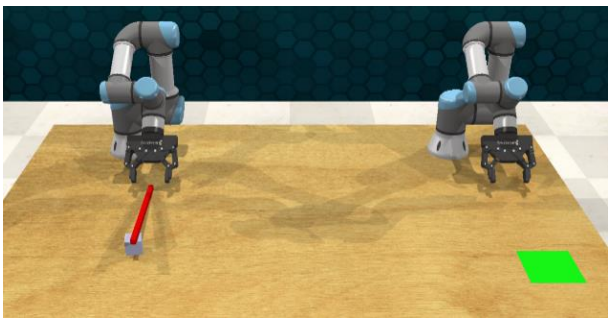

그림 1. 핸드오버 작업 환경

1. 핸드오버 작업

그림 1 은 두 개의 로봇, 대상 물체(빨간색), 목표 위치(녹색), 더미 물체(회색)로 구성된 핸드오버 작업을 위한 환경을 나타낸다. 각 로봇의 팔 관절 자유도는 6 그리고 그리퍼 자유도는 1 이다. 핸드오버 작업에서 성공 조건은 '대상 물체가 목표 위치에 도달한다'로 정의된다. 이러한 환경에서 성공 조건을 만족시키려면 목표 물체와 목표 위치 사이의 먼 거리로 인해 두 로봇이 반드시 협업해야 한다. 즉 왼쪽의 로봇이 대상 물체를 잡은 후, 오른쪽의 로봇에게 넘겨준 후, 오른쪽 로봇이 목표 위치로 대상 물체를 이동시켜야 한다.

2.보상함수 설계

멀티 에이전트 강화 학습을 통해 핸드오버 과제를 해결하기 위해서는 다양한 보상 함수가 필요하고, 이에 대한 정의는 다음과 같다.

- Distance reward for robots:

$$r_{d,o}^{i} = -\sqrt{\left\|\mathbf{b}_{p}^{i} - \mathbf{o}_{p}\right\|^{2}}$$

- Grasping reward:

$$r_{g}^{i} = \begin{cases} c_g, & \text{for gripper touches target object} \\ 0, & \text{for otherwise} \end{cases}$$

- Distance reward for success condition:

$$r_{d,o}^{g} = -\sqrt{\left\|\mathbf{g}_{p} - \mathbf{o}_{p}\right\|^{2}}$$

- Handover reward:

$$r_{h} = \begin{cases} c_h, & \text{for grasping reward pair } (0, c_g) \\ 0, & \text{for otherwise} \end{cases}$$

여기서 $\mathbf{b}_{p}^{i}$ 는 로봇 i 의 그리퍼 포지션, $\mathbf{o}_{p}$ 는 대상 물체의 포지션, $\mathbf{g}_{p}$ 는 목표 위치, and c_g 와 c_h 는 양의 값을 갖는 상수이다. 이와 같은 보상함수들을 이용하여, 전체 보상함수는 다음과 같이 정의된다.

$$r_t = \sum_{i} \left(r_{d,o}^{i} + r_{g}^{i}\right) + r_{d,o}^{g} + r_{h} \ .$$

3. MA-DDPG

다중 에이전트 환경에서 Non-stationary 문제를 해결하기 위해, MA-DDPG 라고 하는 중앙집중 훈련 및 분산 실행 접근 방식에 기반한 방법이 제안되었다 [4]. Actor-Critic 방식을 채택한 MA-DDPG 에서는 여러 에이전트의 정책이 분산 방식으로 실행되고 가치 함수가 중앙집중 방식으로 업데이트된다. 가치 함수는 업데이트 단계에서 여러 에이전트의 모든 행동 정보가 주어진 중앙집중 방식에 기반하기 때문에 Non-stationary 문제는 업데이트 단계에 영향을 미치지 않는다. 시스템 적용의 관점에서, 에이전트의 학습된 정책은 다른 에이전트의 작업을 고려하지 않으므로, 여러 에이전트는 분산 방식으로 동작한다.

Ⅲ. 실험결과

핸드오버 작업을 위한 두 로봇을 훈련하기 위해 MA-DDPG 를 학습 알고리즘으로 사용하고. 네트워크 구조 및 매개변수는 기존 MA-DDPG 연구[4]의 것과 동일하다. 훈련의 경우 에피소드당 최대 길이는 50 회, 최대 에피소드는 50000 회이다.

그림 2 는 에피소드에 따른 전체 보상을 나타낸다. 그림 2 에서 확인할 수 있듯이, 학습 초기에는 성능이 증가하지만 일정 수준에서 한계에 도달한다. 더 자세히 설명하면, 첫 번째 로봇은 대상 물체를 잡을 수 있지만 두 번째 로봇은 대상 물체를 잡을 수 없는 상태, 즉 핸드오버가 이루어지지 않는다. 핸드오버는 매우 섬세한 작업이기 때문에 MA-DDPG 의 단순한 적용으로는 작업을 해결할 수 없다. 첫 번째 로봇의 관점에서 핸드오버는 잡는 보상을 포기하는 것이기 때문에 핸드오버 작업이 잘 수행되지 않게 된다.

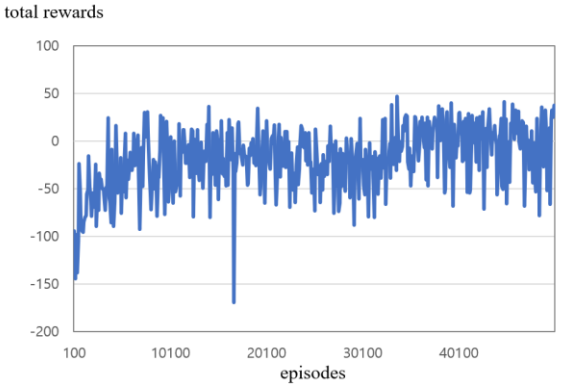

그림 2. 에피소드에 따른 전체 보상

순차적 협업 과제를 해결하기 위해서는 다음과 도전적인 이슈를 고려해야 한다.

- 협업을 위한 정보를 설계하는 방법,
- 협업의 수준을 평가하는 방법,
- 개인보상을 희생하면서 협력보상을 추구하는 방법,
- 개인기술에서 협업기술로 학습을 전환하는 방법

Ⅳ. 결론

본 논문에서는 핸드오버 작업으로 정의된 다중 로봇 조작 문제를 다루었다. 시뮬레이션 결과를 통해 MA-DDPG 의 단순한 적용은 학습에 한계가 있는 것을 확인할 수 있다. 향후 작업에서는 순차적인 협업 과제를 해결하기 위해 도전적인 이슈를 해결하는 방법에 대하여 연구를 진행하고자 한다.

ACKNOWLEDGMENT

본 연구 논문은 한국전자통신연구원 연구운영지원사업의 일환으로 수행되었음. [22ZR1100, 자율적으로 연결· 제어· 진화하는 초연결 지능화 기술 연구].

참 고 문 헌

[1] M. Zhang, et al, “Work chain-based inverse kinematics of robot to imitate human motion with Kinect,” ETRI Journal, vol. 40, no. 4, pp. 511-521, Aug., 2018.

[2] M. Ilbeygi and M. R. Kangavari, “Comprehensive architecture for intelligent adaptive interface in the field of single-human multiple-robot interaction,” ETRI Journal, vol. 40, no. 4, pp. 483-498, Aug., 2018.

[3] H. Kim et al, “Avoiding collaborative paradox in multi-agent reinforcement learning,” ETRI Journal pp. 1004-1012, 43(6), 2021.

[4] R. Lowe et al, “Multiagent actor-critic for mixed cooperative-competitive environments,” in Proc. NIPS, 2017, pp. 2094-2100.