물체 조작 정책의 효율적 습득을 위한 모방 학습과 강화 학습의 결합

정은진, 이상준, 김인철
경기대학교 컴퓨터과학과
e-mail: isk03276@kyonggi.ac.kr, dustashy@kyonggi.ac.kr, kic@kyonggi.ac.kr

Combining Imitation Learning with Reinforcement Learning for Efficient Manipulation Policy Acquisition

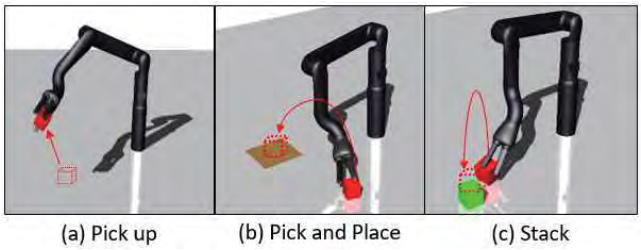
EunJin Jung, SangJoon Lee, Incheol Kim
Department of Computer Science, Kyonggi University

요 약

최근 들어 점차 지능형 서비스 로봇들이 인간의 실생활 속으로 들어옴에 따라, 로봇 스스로 다양한 물체들을 효과적으로 조작할 수 있는 지식을 습득하는 기계 학습 기술들이 매우 주목을 받고 있다. 전통적으로 로봇 행위 학습 분야에는 강화 학습 혹은 심층 강화 학습 기술들이 주로 많이 적용되어 왔으나, 이들은 대부분 물체 조작 작업과 같이 다차원 연속 상태 공간과 행동 공간에서 최적의 행동 정책을 학습하는데 여러 가지 한계점을 가지고 있다. 따라서 본 논문에서는 전문가의 데모 데이터를 활용해 보다 효율적으로 물체 조작 행위들을 학습할 수 있는 모방 학습과 강화 학습의 통합 프레임워크를 제안한다. 이 통합 프레임워크는 학습의 효율성을 향상시키기 위해, 기존의 GAIL 학습 체계를 토대로 PPO 기반 강화 학습 단계의 도입, 보상 함수의 확장, 상태 유사도 기반 데모 선택 전략의 채용 등을 새롭게 시도한 것이다. 다양한 성능 비교 실험들을 통해, 본 논문에서 제안한 통합 학습 프레임워크인 PGAIL의 우수성을 확인할 수 있었다.

1. 서론

전통적으로 로봇 행위 학습 분야에는 강화 학습(reinforcement learning) 기술들이 주로 많이 적용되어 왔다. 하지만 이들 대부분이 저차원(low-dimension)의 이산 상태-동작 공간(discrete state-action space)을 가정함으로써, 실제 로봇의 자율 행위를 학습하는데 한계가 있었다. 최근 들어서는 강력한 일반화 능력과 특징 학습 기능을 가진 심층 신경망(deep neural network)과 결합된 다양한 심층 강화 학습(deep reinforcement learning) 알고리즘들이 개발되어 영상이나 비디오와 같은 고차원(high-dimension)의 연속 입력 센서 데이터(continuous input sensory data)로부터 직접 행동 정책(policy)을 학습할 수 있는 수준으로까지 발전하였다. 하지만, 아직도 (그림 1)과 같이 9-자유도(DoF, Degree of Freedom) 다관절 로봇 팔과 손을 이용한 물체 조작 작업 학습에는 다음과 같은 여러 가지 어려운 난관들이 존재하고 있다[1]. 먼저, 다관절 로봇의 물체 조작 작업은 각 관절의 회전 모터(torque motor)의 회전력을 이용한 물리적인 행동 제어를 요구하기 때문에, 매우 높은 고차원의 연속 상태-행동 공간(continuous state-action space)을 갖는다. 또한 이러한 고차원의 연속 상태-행동 공간에서 최적의 행동 정책(optimal policy)을 학습하려면 대용량의 학습 데이터와 오랜 학습 시간을 필요로 한다. 하지만 로봇의 물리적인 특성상 시뮬레이션 환경이 아닌 실세계에서는 수많은 시행착오 경험을 통한 대용량의 학습 데이터의 확보는 사실상 불가능하다. 따라서 로봇 조작 학습을 위해서는 데이터 효율성(data efficiency)이 매우 높은 학습 알고리즘이 요구되며, 대부분 실세계가 아닌 시뮬레이션 환경에서 오랜 시행착오 경험을 통해 조작 지식을 학습한 후, 실세계 물리 로봇에 지식을 전이(knowledge transfer)하는 방식을 이용한다.



(그림 1) 다관절 로봇의 물체 조작 작업들

많은 물체 조작 행위 혹은 작업들은 로봇 손의 기계적 형태, 조작 대상 물체의 모양과 재질 등의 미세한 차이에도 큰 영향을 받는데 효율적인 학습을 위해 이러한 미세한 부분들까지 정확히 모델링하기는 어렵다. 또한 이러한 조작 작업들은 대부분 여러 단계의 복잡한 미세 제어를 거쳐 완성이 되는데, 이러한 작업들을 위한 보상 함수(reward function)를 설계하는 것도 쉽지 않은 일이다.

앞서 살펴본 바와 같이, 심층 강화 학습 기술은 로봇 스스로 시행착오 경험을 통해 필요한 조작 지식을 학습할 수 있다는 장점은 있으나, 대부분 데이터의 효율성이 낮아 의미 있는 조작 작업 지식을 습득하기까지 너무 오랜 학습 시간을 요구한다는 단점이 있다. 학습 시간을 단축해줄 수 있는 효과적인 방안 중의 하나가 사람 전문가의 작업 데모(expert demonstration)를 활용하는 모방 학습(imitation learning)이 있다[2]. 본 논문에서는 사람 전문

* 이 연구는 2018년도 산업통상자원부 및 산업기술평가관리원(KEIT) 연구비 지원에 의한 연구임('10077538')

가의 데모 데이터를 활용해 보다 효율적으로 물체 조작 작업들을 학습할 수 있는 모방 학습과 강화 학습의 통합 프레임워크를 제안한다. 이 통합 프레임워크는 학습의 효율성을 향상시키기 위해, 기존의 GAIL(Generative Adversarial Imitation Learning)[3] 학습 체계를 토대로 PPO [4] 기반 강화 학습 단계의 도입, 보상 함수의 확장, 상태 유사도 기반 데모 선택 전략의 채용 등을 새롭게 시도한 것이다. 본 논문에서 제안한 통합 학습 프레임워크의 우수성을 확인하기 위한 다양한 성능 비교 실험들을 수행하고, 그 결과를 소개한다.

2. 관련연구

본 논문에서 제안하는 강화 학습과 모방 학습을 결합한 통합 학습 프레임워크는 데모 데이터를 효과적으로 모방함으로써 강화 학습의 탐사 구간(exploration range)을 좁혀 데이터의 효율성을 높이는 방법이다. 대표적인 모방 학습 방법들로는 행위 복제(behavioral cloning)와 역 강화 학습(inverse reinforcement learning)이 있다. 행위 복제(behavior cloning)는 상태-행동 쌍들로 구성된 전문가의 데모 집합을 학습 데이터로 삼아 새로운 상태에 적합한 행동을 결정할 수 있는 일반화된 행동 정책을 감독 학습(supervised learning)하는 방식이다[2]. 반면에, 역 강화 학습(inverse RL)은 전문가의 데모 데이터로부터 최적의 비용 함수(cost function)를 역 추정해낸 뒤, 이 비용 함수를 이용하여 다시 강화 학습을 수행함으로써 학습자의 최적 행동 정책을 찾아낸다[5]. 하지만 이 방법들은 각각 고유한 문제점들을 갖고 있다. 행위 복제의 경우 연속 상태-행동 공간에서도 의미 있는 행동 정책을 학습해내기 위해서는 충분히 많은 양의 전문가 데모 데이터가 요구되는 반면, 역 강화 학습의 경우는 소량의 데모 데이터로부터 최적의 비용 함수를 역 추정해내기 위해서는 계산 비용이 너무 크다는 문제점이 있다. 이러한 문제점을 극복하고자 제안된 GAIL(Generative Adversary Imitation Learning)[3] 학습 체계는 대표적인 복합 심층 신경망 모델의 하나인 GAN(Generative Adversarial Network)과 유사하게, 서로 적대 관계를 이루는 생성자 네트워크(generator network)와 판별자 네트워크(discriminator network)를 포함한다. 생성자 네트워크, 즉 정책 네트워크는 보상을 최대로 높일 수 있도록 행동을 결정하는데 반해, 판별자 네트워크는 생성자가 결정한 행동이 전문가 데모와 얼마나 일치하는 지를 판별한다. 감독 학습을 통해 전문가의 데모 데이터를 그대로 학습자의 행동 정책으로 일반화하려는 행위 복제에 비해, GAIL은 소량의 전문가 데모 데이터를 토대로 학습자 스스로 일정한 정도의 시행착오 경험을 통해 자신의 고유한 행동 정책을 학습할 수 있는 특징이 있다. 따라서 GAIL은 행위 복제와는 달리 대용량의 데모 데이터 집합을 요구하지 않음으로써 비교적 높은 데이터 효율성을 보이며, 또한 역 강화 학습과는 달리 비용 함수의 역 추정을 위한 높은 계산 비용도 요구하지 않는다는 장점이 있다. 하지만 GAIL은 전문가의 데모 데이터를 무작위(random) 방식으로 선택하여 학습에 활용하고, 데모 데이터와의 일치도만을 고려하여 설계한 보상 함수를 이용함으로써 학습 성능과 확장성에 한계가 있다. 한편 [6]의 연구에서는 심층 강화 학습의 수렴성을 향상시키기 위해 리플레이 버퍼(replay buffer)에 저장된 과거 경험들 중에서 보상이 높은 우수한 데이터들만을 골라 정책 신경망을 업데이트하는 효과적인 경험 선택(experience selection) 전략을 제시하였다. 하지만 이 전략은 효과적인 모방 학습을 위한 데모 데이터의 선택 전략과는 거리가 있다.

3. 통합 학습 프레임워크

3.1 PPO 기반의 생성적 적대 모방 학습

앞서 소개한 GAIL[3]은 전문가의 데모 데이터를 효과적으로 활용함으로써, 강화 학습이 탐사해야할 넓은 상태-행동 공간을 줄이고 최적 행동 정책을 효율적으로 학습할 수 있는 일종의 혼합 학습 체계(hybrid learning system)이다. 이 학습 체계에서는 사람 전문가의 데모가 $D=(s_i,a_i)_{i=1,\dots,N}$와 같이 상태-행동 쌍들로 주어진다고 가정할 때, 이들을 효과적으로 활용하여 최적의 행동 정책 $\pi_{\theta^*}:S\rightarrow A$을 효율적으로 학습하는 것이 목표이다. GAIL 학습 체계에서는 대표적인 복합 심층 학습 모델의 하나인 GAN과 마찬가지로, 2개의 적대 관계 네트워크들을 포함한다. 하나는 행동을 결정하는 생성자 네트워크 $\pi_\theta:S\rightarrow A$이고, 다른 하나는 데모 데이터와의 일치도를 판별하는 판별자 네트워크 $D_\psi:S\,X\,A\rightarrow[0,1]$이다. 이 두 네트워크의 학습에는 (식 1)과 같은 최소-최대 목적 함수(min-max objective function)를 이용한다.

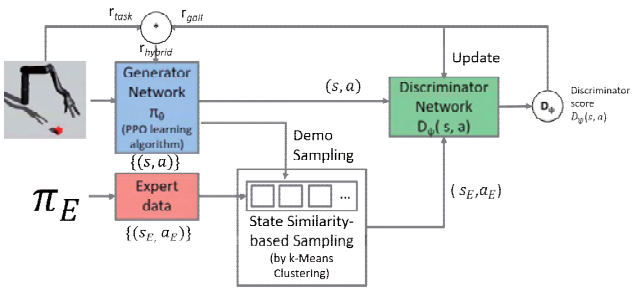
$$\min_\theta \max_\psi E_{\pi_E}[\log D_\psi(s,a)]+E_{\pi_\theta}[\log(1-D_\psi(s,a))] \quad \text{(식 1)}$$

(식 1)에서 π_E는 데모 경로들을 생성해놓은 전문가 정책(expert policy)을 나타낸다. GAIL의 학습 과정동안에는 매번의 에피소드마다 판별자 네트워크의 파라미터를 갱신하는 단계와 TRPO 알고리즘에 따라 생성자 네트워크의 파라미터를 갱신하는 단계를 교대로 수행한다.

한편, TRPO 강화 학습 알고리즘은 (식 2)와 같이 현재 정책과 새로운 정책 간의 확률 비(probability ratio) $\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$에 우세 함수(advantage function)의 추정값 $\hat{A}_t$을 곱하여 목적함수로 사용한다. 또 매개변수의 갱신량에 쿨백-레이블러 발산(Kullback-Leibler divergence, KL)값 $\beta KL\left[\pi_{\theta_{old}}(\cdot|s_t),\pi_\theta(\cdot|s_t)\right]$로 제약을 부과하여 기존 정책 기울기 알고리즘들의 매개변수 미세변화에 따른 높은 변량(high variance) 문제를 해결했다. 하지만, TRPO는 구현의 어려움, KL 분산값의 높은 계산량 등의 문제가 존재한다. 이를 해결하기 위해 근위 정책 최적화(Proximal Policy Optimization) 알고리즘인 PPO는 TRPO의 KL 분산값의 계산을 없애고 목적 함수를 클리핑시킴으로써, 목적 함수에 일정한 제약을 부과했다. 이를 통해, PPO는 TRPO의 높은 계산 복잡도 문제와 구현의 어려움을 해결하고, 높은 학습 성능을 보인다[4].

$$\hat{E}_t\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}\hat{A}_t-\beta KL\left[\pi_{\theta_{old}}(\cdot|s_t),\pi_\theta(\cdot|s_t)\right]\right] \quad \text{(식 2)}$$

따라서 본 논문에서 새롭게 제안하는 통합 학습 프레임워크에서는 GAIL에서 채용했던 TRPO 강화 학습 알고리즘 대신 새로운 PPO 알고리즘을 채용함으로써 학습의 효율성을 향상시키고자 한다. (그림 2)는 본 논문에서 제안하는 PPO 기반의 생산적 적대 모방 학습 프레임워크인 PGAIL의 구성을 보여준다. 그림에서 행동 정책을 나타내는 생성 네트워크의 파라미터들은 새로 채택한 PPO 알고리즘에 의해 매번 갱신된다. 또한 PPO 강화 학습 단계에 이용되는 보상 함수도 GAIL의 경와는 달리, 데모 데이터와의 일치도 뿐만 아니라 작업 수행에 따른 누적 보상도 반영될 수 있도록 확장된다. 그림의 하단부에 표시된 데모 데이터 선택부는 역시 GAIL의 임의 선택 전략에서 벗어나 현재 작업 상태와 유사한 데모 데이터를 골라 효과적으로 활용하는 상태 유사도 기반의 데모 선택 전략을 채택하고 있다.



(그림 2) PPO 기반의 생성적 적대 모방 학습 프레임워크

새롭게 확장된 보상 함수와 상태 유사도 기반의 데모 선택 전략은 3.2절과 3.3절에서 자세히 설명한다.

3.2 보상 함수 확장

보상 함수는 넓은 상태-행동 공간에서 탐사를 유도하는 강화 학습의 매우 중요한 요소이다. GAIL 학습 체계에서는 행동 정책을 나타내는 생산자 네트워크를 학습할 때 (식 3)에 정의된 보상 함수 r_{gail}를 이용하였다. 이 보상 함수는 판별자 네트워크의 출력인 데모 데이터와의 일치도 $D_\psi(s_t,a_t)$에만 의존하여 보상을 결정한다.

$$r_{gail} = -\log(1-D_\psi(s_t,a_t)) \quad \text{(식 3)}$$

본 논문에서 제안하는 통합 학습 프레임워크인 PGAIL에서는 기존의 GAIL에서 적용한 보상 함수를 확장하여 (식 4)와 같은 새로운 보상 함수 r_{hybrid}를 채택한다.

$$r_{hybrid}(s_t,a_t) = r_{gail}(s_t,a_t) * r_{task}(s_t,a_t) \quad \text{(식 4)}$$

새로운 보상 함수 r_{hybrid}에는 데모 데이터와의 일치도를 반영하는 모방 보상 r_{gail}외에 작업 보상 r_{task}이 새로 추가되었다. 이 작업 보상은 데모 데이터와의 일치도와는 무관하게, 수행하고자 하는 작업의 완성도에 따라 결정되는 보상을 의미한다. 이러한 작업 보상은 일반적으로 순수 강화 학습에서 주로 채용하는 보상들과 같은 것으로 해석할 수 있다. 예컨대, 로봇 손을 물체에 접근하기 위한 reach 작업에서는 로봇 손이 물체에 근접한 정도에 따라 경정되는 보상이 작업 보상의 한 예가 될 수 있다. 모방 보상 외에 이러한 작업 보상 요소를 포함한 보상 함수 r_{hybrid}는 학습을 전문가의 데모에 근접하도록 유도할 뿐만 아니라, 스스로 작업 성과를 높일 수 있는 방향으로도 유도할 수 있다. 이러한 두 보상 요소의 보완적 기능은 전문가 데모에 오류가 있거나 품질이 낮은 경우에 이러한 데모를 단순히 모방하기 보다는, 학습자인 로봇 스스로 전문가 데모를 뛰어넘는 양질의 행동 정책을 학습할 수 있는 기회를 제공한다는 면에서 큰 의미를 가진다.

3.3 상태 유사도 기반의 데모 선택 전략

모방 학습에서는 현재 작업 상태에 도움을 줄 수 있는 전문가의 데모 데이터를 어떤 방식으로 선택하는 지가 학습 성능에 큰 영향을 미칠 수 있다. 하지만 기존의 GAIL 학습 체계에서는 현재 자업 상태와는 무관하게 전문가의 데모 데이터를 무작위(random) 방식으로 선택하여 학습에 활용하였다. 이러한 무작위 데모 선택 전략으로 인해 GAIL은 학습 성능과 확장성에 한계가 있었다. 본 논문에서 제안하는 통합 학습 프레임워크인 PGAIL에서는 k-평균 군집화(k-means clustering)을 이용하여, 현재 상태 상태에 가장 유사한 데모 데이터들만을 선택해 활용하는 상태 유사도 기반의 데모 선택 전략(State Similarity-based Demo Sampling Strategy)을 구현하였다.

논 논문에서 다루는 3 가지 조작 작업들의 경우, 각 데모 데이터의 상태는 로봇의 9개 관절의 각도, 속도와 물체의 위치 좌표로 이루어진 크기 21의 실수 벡터로 정의할 수 있다. 로봇의 현재 작업 상태와 유사한 상태를 갖는 데모 데이터들을 선택하기 위해 여러 방법들이 존재한다. 가장 직관적인 방법은 상태 벡터의 각 21개의 값들을 모두 비교하여 차가 가장 적은 데모 데이터를 선택하는 방법이다. 하지만, 이 방법은 매번 데모 데이터를 선택할 때마다 많은 계산량을 요구하기 때문에 적절하지 않다. 이를 해결하기 위해 본 논문에서는 k-평균 군집화를 통해 유사한 상태의 데모 데이터를 선택한다. 즉, 각 데모 데이터의 상태를 기준으로 미리 데모 데이터들을 사전에 군집화 시켜 놓은 후, 로봇의 현재 상태와 유사한 군집에서 데모 데이터를 추출하여 학습에 이용한다. K-평균 군집화 알고리즘을 통해 현재 작업 상태에 유사한 데모 데이터를 선택할 경우, 각 군집의 중심값과의 거리만을 비교하기 때문에 계산량을 대폭 감소시킬 수 있다.

4. 구현 및 실험

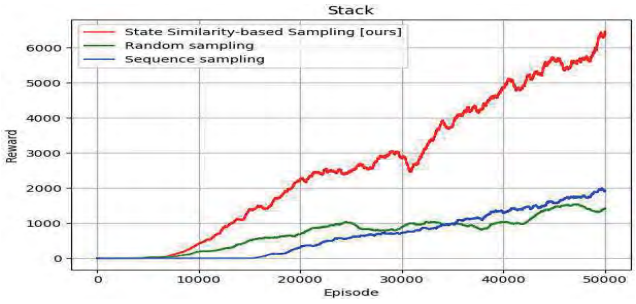
4.1 구현

본 논문에서 제안하는 PPO 기반의 생산적 적대 모방 학습 프레임워크(PGAIL)은 2개의 신경망 서브 네트워크들로 구성된다. 정책 신경망인 생성자 신경망과 판별자 신경망 모두 3층의 완전 연결 층(fully connected layer)으로 구현했다. 한편 MuJoCo 시뮬레이션 환경의 Jaco 로봇 팔과 손을 사람이 직접 제어하여 데모 데이터를 확보하기 쉽지 않아, 본 연구에서는 대신 PPO 강화 학습 알고리즘으로 정책 신경망을 장시간 학습시켜 일정 수준의 성능을 보이는 행동 정책을 얻은 후 이것을 이용하였다. 즉 로봇이 학습된 행동 정책에 따라 스스로 조작 작업을 수행할 수 있게 되었을 때, 관절들의 이동 경로(trajectory)를 기록하여 해당 작업의 모방 학습을 위한 초기 전문가 데모 데이터로 활용하였다. 구현과 실험은 Geforce GTX 1080 Ti GPU가 탑재된 컴퓨터에서 수행되었다. 구체적인 작업 환경과 실험 환경은 MuJoCo 시뮬레이터와 Jaco 로봇 손 모델, mujoco-py 라이브러리 등을 이용해 Python으로 구현하였다. PPO 기반의 생산적 적대 모방 학습 프레임워크(PGAIL)의 효과적인 구현과 학습을 위해 Tensorflow 심층 신경망 라이브러리를 이용하였다.

4.2 실험

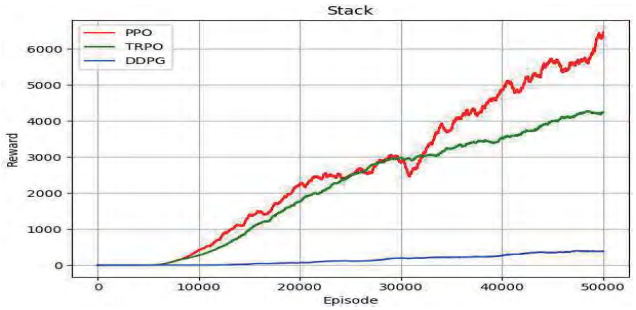
본 논문에서 제안하는 PPO 기반의 생산적 적대 모방 학습 프레임워크(PGAIL)의 성능을 분석하기 위한 다양한 실험들을 수행하였다. 첫 번째 실험은 PGAIL 통합 학습 프레임워크에서 채택한 상태 유사도 기반 데모 선택(State Similarity-based Sampling, SSS) 전략의 효과를 분석하기 위한 실험이다. 이 실험에서는 Stack 조작 작업에서 본 논문의 상태 유사도 기반 선택(SSS) 전략을 무작위 선택(random sampling), 순차 선택(sequential sampling) 등과 같은 베이스라인 데모 선택 전략들과 성능 비교를 수행하였으며, 실험 결과는 (그림 3)과 같다. 본 논문에서 제안한 상태 유사도 기반 선택(SSS) 전략이 가장 우수한 성능을 보였으며, 다른 베이스라인 전략들에 비해 뚜렷한 성능 향상이 있었음을 확인할 수 있다.

두 번째 실험은 PGAIL 통합 학습 프레임워크에서 채택한 PPO 강화 학습 알고리즘의 효과를 분석하는 실험이다. 이 실험에서는 GAIL에서 채용했던 알고리즘인 TRPO, 연속 제어 정책 학습을 위한 또 다른 강화 학습 알고리즘인 DDPG(Deep Deterministic Policy Gradient)들을 본 논문에서 채택한 PPO 알고리즘과 성능 비교를 수행하였으며, 실험 결과는 (그림 4)와 같다.



(그림 3) 데모 선택 전략에 따른 성능 비교

전체적으로 본 논문에서 채택한 PPO 알고리즘이 DDPG와 TRPO 알고리즘들에 비해 학습 향상 속도가 빠르다는 것을 확인할 수 있다. DDPG에 비교해서는 학습 초기부터 성능 차이가 뚜렷한데 반해, 유사한 특성을 가진 TRPO 알고리즘에 비교해서는 학습 초기에 거의 비슷한 학습 성능을 보였다. 하지만 학습이 약 3만 에피소드까지 충분히 진행된 이후부터는 TRPO에 비해 뚜렷한 성능 향상을 보인 것을 알 수 있다.



(그림 4) 강화 학습 알고리즘에 따른 성능 비교

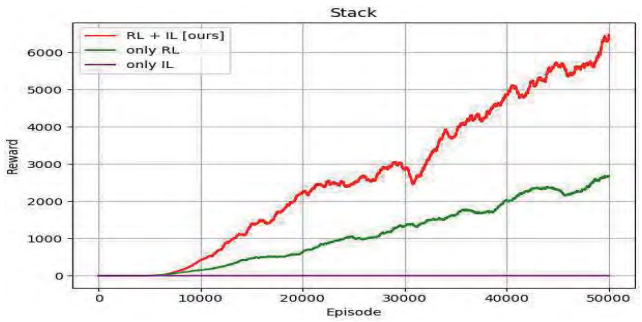
세 번째 실험은 모방 학습과 강화 학습을 모두 포함하는 본 논문의 PGAIL 통합 학습 프레임워크(RL + IL)의 성능을 PPO 알고리즘을 이용한 순수 강화 학습(only RL)과 행위 복제 방식의 순수 모방 학습(only IL)과 비교함으로써, PGAIL 통합 학습 프레임워크(RL + IL)의 우수성을 입증하기 위한 실험이다. 이 실험은 Pickup, Pick and Place, Stack 등 총 3 가지 조작 작업들 모두에서 수행되었으며, 실험 결과는 (그림 5), (그림 6), (그림 7)과 같다. 서로 다른 3 가지 조작 작업 모두에서 본 논문에서 제안한 PGAIL 통합 학습 프레임워크(RL + IL)이 다른 순수 강화 학습(only RL)이나 순수 모방 학습(only IL)에 비해 매우 뚜렷한 성능 향상을 보인 것을 확인할 수 있다. 또한 학습이 더 진행될수록 이들 간의 성능 격차는 더 심화되는 것을 확인할 수 있다. 이러한 실험결과들을 종합해볼 때, 본 논문에서 제안한 PGAIL 통합 학습 프레임워크의 우수성을 충분히 확인할 수 있다.

5. 결론

본 논문에서는 전문가의 데모 데이터를 활용해 보다 효율적으로 물체 조작 행위들을 학습할 수 있는 모방 학습과 강화 학습의 통합 프레임워크를 제안한다. 이 통합 프레임워크는 학습의 효율성을 향상시키기 위해, 기존의 GAIL 학습 체계를 토대로 PPO 기반 강화 학습 단계의 도입, 보상 함수의 확장, 상태 유사도 기반 데모 선택 전략의 채용 등을 새롭게 시도한 것이다. 다양한 실험들을 통해, 제안한 통합 학습 프레임워크의 우수성을 확인할 수 있었다.

(그림 5) 학습 전략 간의 성능 비교: Pickup 작업

(그림 6) 학습 전략 간의 성능 비교: Pick and Place 작업



(그림 7) 학습 전략 간의 성능 비교: Stack 작업

참고문헌

[1] Y. Zhu, Z. Wang, and J. Merel, et al., "Reinforcement and Imitation Learning for Diverse Visuomotor Skills," *Proceedings of the International Conference on Robotics Science and Systems*, 2018

[2] S. Ross, G. J. Gordon, and D. Bagnell "A Reduction of Imitation Learning and Structured Prediction to No-regret Online Learning," *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 627 - 635, 2011.

[3] J. Ho, and S. Ermon "Generative Adversarial Imitation Learning," *Advances in Neural Information Processing System*, pp. 4565 - 4573, 2016.

[4] J. Schulman, F. Wolski, and P. Dhariwal, et al., "Proximal Policy Optimization Algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[5] C. Finn, S. Levine, and P. Abbeel "Guided Cost Learning: Deep Inverse Optimal Control via Policy Optimization," *Proceeding of the International Conference on Machine Learning*, pp. 49 - 58, 2016.

[6] T. de Bruin, J. Kober, and K. Tuyls, et al., "Experience Selection in Deep Reinforcement Learning for Control," *Journal of Machine Learning Research*, vol. 19, no. 9, 2018.