Journal of Korea Robotics Society (2020) 15(2):197-204
https://doi.org/10.7746/jkros.2020.15.2.197 ISSN: 1975-6291 / eISSN: 2287-3961

현실 세계에서의 로봇 파지 작업을 위한 정책/가치 심층 강화학습 플랫폼 개발

Development of an Actor-Critic Deep Reinforcement Learning Platform for Robotic Grasping in Real World

김 태 원[1] · 박 예 성[1] · 김 종 복[2] · 박 영 빈[2] · 서 일 홍[†]
Taewon Kim[1], Yeseong Park[1], Jong Bok Kim[2], Youngbin Park[2], Il Hong Suh[†]

Abstract: In this paper, we present a learning platform for robotic grasping in real world, in which actor-critic deep reinforcement learning is employed to directly learn the grasping skill from raw image pixels and rarely observed rewards. This is a challenging task because existing algorithms based on deep reinforcement learning require an extensive number of training data or massive computational cost so that they cannot be affordable in real world settings. To address this problems, the proposed learning platform basically consists of two training phases; a learning phase in simulator and subsequent learning in real world. Here, main processing blocks in the platform are extraction of latent vector based on state representation learning and disentanglement of a raw image, generation of adapted synthetic image using generative adversarial networks, and object detection and arm segmentation for the disentanglement. We demonstrate the effectiveness of this approach in a real environment.

Keywords: Actor-Critic Deep Reinforcement Learning, Robotic Grasping

1. 서 론

최근의 심층 강화학습 기술의 발전으로 인해 로봇 파지 분야에서도 심층 강화학습에 기반한 파지기술들이 연구되고 있다[1-3]. 이중 Q-function Targets via Optimization (QT-Opt)[2]라는 구글에서 개발된 가치기반 심층 강화학습 기반한 연구는 강화학습 혹은 비강화학습 기반 연구를 포함하더라도 최고 수준(state-of-the-art)의 파지성공률을 나타내고 있다. 하지만 가치기반 심층 강화학습은 이산적인(Discrete) 행동 결정에서는 적합하나 연속적인(Continuous) 행동 결정에는 적합하지 않다. 그 이유는 해당 학습 방법이 정책망(Actor network)을 직접적으로 학습하지 않고 가치망(Critic network)만을 학습하고 이러한 가치망을 통해 정책을 추정하게 되는데, 연속행동 결정을 위해서는 이러한 정책 추정 과정을 위해서 추가적인 최적화(Optimization) 과정을 수행해야 하기 때문이다. 반면 정책/가치 심층 강화학습의 경우에는 정책망뿐만 아니라 가치망도 직접적으로 학습하기 때문에 연속행동 결정을 하기 위해 추가적인 최적화 과정이 요구되지 않는다.

한편, 최근까지 정책/가치 심층 강화학습에 기반한 경우에는 입력이 영상으로 주어지는 종단 간(End-to-end) 심층학습일 경우 다양한 물체를 파지하는데 효율적이지 못하였고[1], 이러한 문제를 해결하기 위해 상태표현 학습(State Representation Learning)을 이용하고 이때 학습 효율성을 향상하기 위하여 입력 영상을 독립된 요소로의 분리(Disentanglement)하는 방법이 제안되었다[3]. 하지만 이 방법은 시뮬레이터에서의 파지 성공률을 이전의 정책/가치 심층 강화학습 방법에 비해서 높게 향상시켰으나 학습을 위해 필요한 데이터의 양이 많아 현실 세계에는 직접적으로 적용하기 어렵다는 단점이 있다.

Received : Mar. 4. 2020; Revised : Apr. 8. 2020; Accepted : Apr. 21. 2020
※ This work was supported by the Technology Innovation Industrial Program funded by the Ministry of Trade, (MI, South Korea) [10073161, Technology Innovation Program], as well as by Institute for Information\&communications Technology Promotion (IITP) grant funded by MSIT (No. 2018-0-00622)
1. Principal Researcher, the Department of Electronics and Computer Engineering, Hanyang University, Seoul, Korea (incorl.ktw, ysp1026@incorl.hanyang.ac.kr)
2. Manager, the Department of Electronics and Computer Engineering, Hanyang University, Seoul, Korea (ggamir99, pa9301@hanyang.ac.kr)
† Associate Professor, Corresponding author: the Department of Electronics and Computer Engineering, Hanyang University, Seoul, Korea (ihsuh@hanyang.ac.kr)

요약하자면 현재의 심층 강화학습에 기반한 실세계에서의 로봇 파지 연구는 계산량이나 혹은 학습데이터의 양에 관한 문제로 인하여 실제적으로 사용할 수 있는 학습 방법은 존재하지 않는다고 볼 수 있다. 본 논문은 이러한 문제를 해결하고자 정책/가치 심층 강화학습에 기반하여 현실 세계에서의 로봇 파지를 위한 학습 플랫폼을 제안하고자 한다.

2. 관련연구

최근에는 심층학습에 기반하여 다양한 형태의 물체를 파지하기 위한 기술들이 많이 연구되고 있다[2,4,5]. 이중 지도학습[4], 자기 지도학습[5], 혹은 가치기반 심층 강화학습[2] 기반한 연구는 파지성공률이 70~80%에 이를 정도로, 심층신경망의 출력이 로봇의 파지 자세인 종단 간 심층 강화학습 연구 중에서는 최고 수준(state-of-the-art)의 파지성공률을 나타내고 있다. 하지만 대부분의 방법들이 많은 실세계 학습데이터를 필요로 한다.

가치기반 강화학습의 대표적인 방법인 QT-Opt[2]의 경우에는 연속행동 결정을 위한 최적화를 위해 Cross Entropy Method (CEM) 라는 방법을 사용하는데 이는 로봇 파지 자세의 샘플링을 통한 최적화 방법이다. QT-Opt에서는 CEM의 샘플링의 개수가 증가함에 따라 파지성공률은 향상되지만 계산량은 선형적으로 증가한다. 예를 들어 만약 1개의 파지 자세 샘플을 사용할 경우 정책/가치 심층 강화학습과 계산량이 유사하다. 하지만 일반적으로는 50~100개 정도의 파지 자세 샘플을 사용하는데 이 경우 정책/가치 심층 강화학습에 비해서 약 50~100배 정도 더 많은 연산을 필요로 하게 된다.

상태표현 학습은 심층 강화학습을 보조하기 위한 방법으로 최근 다양한 상태표현 학습 방법을 사용하여 다양한 작업에 사용되고 있다[6,7]. 심층 강화학습을 위한 심층신경망은 크게 영상정보를 처리하는 컨볼루션(Convolutional) 계층과 로봇 행동을 생성하는 완전연결(Fully-connected) 계층으로 이루어져 있다. 기존의 정책/가치 심층 강화학습이 입력이 영상으로 주어졌을 경우 학습 효율이 떨어지는 이유는 가치기반 심층 강화학습 방법에 비해서 학습해야 할 파라미터의 개수가 2배이기 때문이고 동시에 정책망과 가치망의 학습이 서로 상호의존적이기 때문이다[1]. 따라서 상태표현 학습을 통해서 컨볼루션 계층을 미리 학습하고 심층 강화학습 시에는 완전연결 계층만 학습함으로써 정책/가치 심층 강화학습의 효율성을 높일 수 있다. 하지만 현재까지 상태표현 학습 방법들의 공통적 한계점은 입력 영상이 비교적 단순해야 한다는 점이다. 즉 비교적 단순한 배경에 단순한 물체들일 경우에만 학습이 효율적으로 이루어진다.

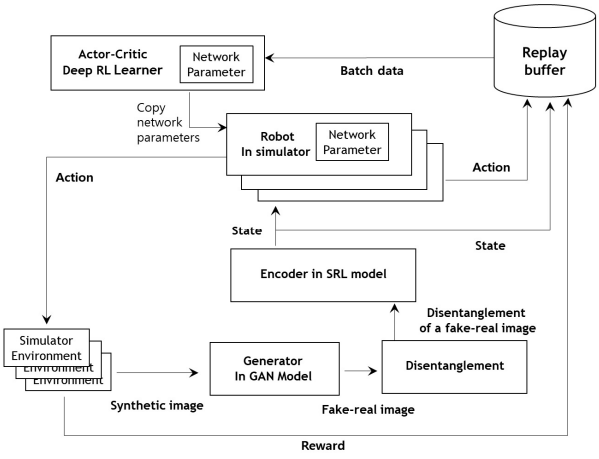
이러한 문제를 해결하기 위해 입력 영상을 독립적인 요소로 분리된 영상으로 분리(Disentanglement)하여 상태 표현 학습을 수행하는 방법이 제안되었다[3]. Disentanglement는 주로 상태학습에서 고차원의 데이터가 저차원의 벡터로 인코딩 될 때 인코딩 된 벡터의 각 차원이 서로 확률적 독립이어야 한다는 이론에서 사용되었던 개념이다. 마찬가지의 의미에서 해당 논문은 입력 영상의 3단계(L1 ~L3) Disentanglement를 제안했다. 즉 각 단계별로 원 영상을 점점 더 확률적으로 독립된 구성요소로 분리하였다. 이 방법을 통해 다양한 물체들을 복잡한 배경하에서 약 72%의 성공률로 파지하는 것이 가능하였다. 하지만 해당 방법은 많은 학습데이터를 요구하기 때문에 시뮬레이터에서만 적용이 가능한 기술이라는 한계점이 있다.

3. 제안하는 학습 플랫폼

[Fig. 1]은 시뮬레이터에서 로봇 파지 학습을 위한 시스템 구성도를 나타낸다. 본 연구에서는 사전의 상태표현학습에 기반한 정책/가치 심층 강화학습 방법을 사용하고 있으며 상태표현 신경망의 입력으로는 원 영상이 아니라 이를 독립된 요소로 분리하여 사용하였다. 이때 상태표현 모델로는 Beta Variational Autoencoder (β-VAE)[8]를 사용하였고 분리(Disentanglement) 수준으로는 L3[3]를 사용하였다.

상태표현 학습 시 β-VAE 는 인코더(Encoder) 망과 디코더(Decoder) 망으로 이루어져 있는데 정책망과 가치망의 영상정보를 처리하는 컨볼루션 계층은 상태표현학습 시 학습한 인코더의 파라미터를 그대로 복사하고 해당 계층의 파라미터는 고정된다. 강화학습 시에는 완전연결 계층의 학습만 수행된다.

시뮬레이터에서 학습 시 중요한 구성요소는 생성적 적대 신경망(GAN)을 이용한 시뮬레이터 영상으로부터 유사 현실(Fake-real) 영상으로의 영상변환이다. 영상변환을 위한 생성적 적대 신경망 모델로는 구글에서 제안한 GraspGAN[9]을 사용하였다. 본 시스템 구성도에서는 GraspGAN의 생성자

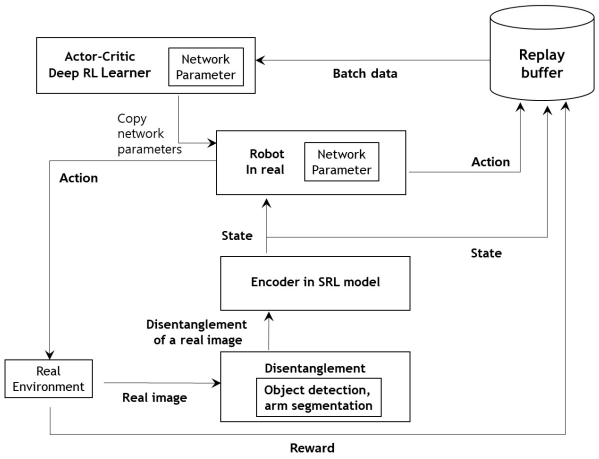


[Fig. 1] The block diagram for training in simulator

(Generator) 망과 판별자(Discrimi-nator) 망 중 생성자 망을 사용하여 영상변환을 수행한다. 시뮬레이터가 비록 현실 세계에서 보다 로봇 파지 학습을 위한 학습데이터를 빠르게 취득할 수는 있지만 더 효율적인 학습데이터 취득을 위해서 본 연구에서는 그림에서와 같이 복수의 시뮬레이터를 사용하여 학습데이터를 취득하였다. 이러한 학습데이터는 Replay buffer에 저장되고 정책/가치 심층 강화학습을 위한 learner가 Replay buffer로부터 학습데이터를 추출하여 학습을 수행한다. 다양한 정책/가치 심층 강화학습 방법 중에서 본 논문에서는 Distributed Distributional Deep Deterministic Policy Gradient (D4PG)[10]를 사용하였는데 이는 잘 알려진 Deep Deterministic Policy Gradient (DDPG)[11]를 확장한 알고리즘이다. Learner는 일정한 주기로 학습한 파라미터 중 정책망에 대한 파라미터를 시뮬레이터의 로봇에게 전달한다. 시뮬레이터의 로봇은 복사한 파라미터를 이용해 로봇의 정책을 결정하는데 사용한다.

[Fig. 2]는 현실 세계에서 로봇 파지 학습을 위한 시스템 구성도를 나타낸다. 학습할 정책망과 가치망의 영상 정보를 처리하는 컨볼루션 계층은 시뮬레이터에서의 학습 때와 마찬가지로 상태표현학습 시 학습한 파라미터를 그대로 복사하여 고정된다. 완전연결 계층 파라미터의 경우에는 시뮬레이터에서의 학습이 완료되고 나서의 파라미터를 복사해서 파라미터의 초기화 값으로 사용한다. 강화학습을 통해 해당 완전연결 계층의 파라미터는 수정된다. 시뮬레이터의 학습 때와 달리 복수의 로봇을 통해 학습데이터를 취득하는 것이 아니라 하나의 로봇을 이용해 학습데이터를 취득한다. 시뮬레이터에서 학습 시에 시뮬레이터 영상이 아니라 [Fig. 3]과 같이 현실 세계의 영상과 유사한 유사 현실 영상을 통해 학습했기 때문에 현실 세계에서 학습 시 실세계의 영상을 입력으로 사용할 수 있다. 시뮬레이터에서의 학습 때와 마찬가지로 입력 영상을 독립적인 요소로 분리된 영상으로 분리하여 상태표현 학습기의 인코더에 입력으로 사용하였다. L3 수준의 영상의 분리를 위해서는 파지하고자 하는 물체를 영상에서 검출해야 하고 로봇 암을 영상에서 분할해야 한다. 이는 시뮬레이터에서는 쉽게 추출 가능한 정보이지만 현실 세계에서의 학습에서는 그렇지 않다. 따라서 본 연구에서는 목표 물체를 영상에서 검출하기 위해서 기존의 물체 검출기(Object detector)와 의미적 객체 분할기(Semantic segmentator)를 사용하였다.

요약하자면 제안하는 학습 플랫폼은 시뮬레이터의 학습과 현실 세계의 학습으로 구성되어 있고, 주요 구성요소는 Disentanglement 기반한 상태표현 학습기, 영상변환을 위한 생성적 적대 신경망, 실세계에서 Disentanglement를 위한 물체 검출이기와 의미적 객체 분할기다.



[Fig. 2] The block diagram for training in real world

[Fig. 3] Left: a synthetic image, Center: a real image, Right: an adapted image that look similar to a real image, which is referred to “fake-real image” in this paper

4. 시뮬레이터에서의 학습

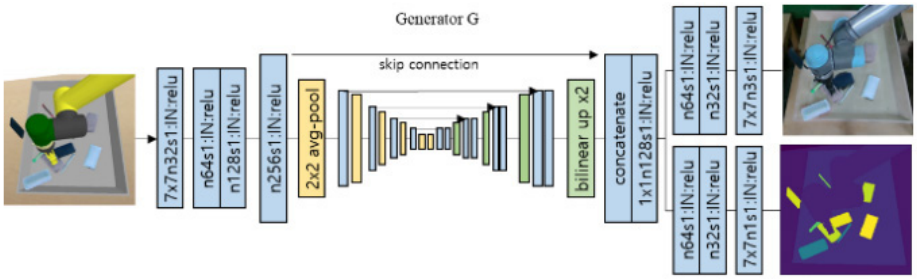
4.1 상태표현 학습

상태표현 학습은 특징학습(Feature learning)의 한 분야로서 효율적으로 정책을 학습할 수 있는 저차원의 벡터공간을 학습을 통해 찾아내는 것이 목적이다. 본 논문에서 사용하고 있는 β-VAE의 경우에는 Autoencoder (AE)[12]의 확장된 모델로서 기본적으로 인코더(Encoder) 망과 디코더(Decoder) 망으로 이루어져 있다. 본 논문에서 사용하고 있는 인코더의 신경망 구조는 Darknet-53[13]과 유사한 구조를 사용하고 있고 디코더의 신경망 구조는 인코더와 대칭적인 구조를 사용한다[3]. AE는 다음과 같은 복원 손실함수를 최소화한다.

$$L_{AE}(\theta,\phi) = \mathbb{E}_{p(x)}\left[\, \| f_\phi(f_\theta(x)) = x \|^2 \right], \quad (1)$$

위 식에서 f_θ과 f_ϕ는 각각 인코더와 디코더를 나타낸다. Variational AE[14]는 AE의 확장된 모델로서 아래와 같은 손실함수를 최소화한다.

$$L_{VAE}(\theta,\phi) = \mathbb{KL}(q_\theta(z|x) \,\|\, p(z)) - \mathbb{E}_{q\theta(z|x)}\left[log\, p_\phi(z|x) \right] \quad (2)$$



[Fig. 4] Architectures for Generator G. n64s1:IN:relu means 64 filters, stride 1, instance normalization IN and relu activation

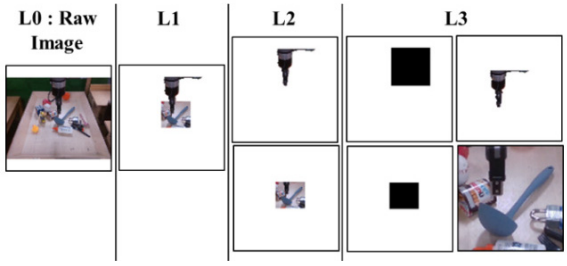
위 식에서 두 번째 항목은 복원 에러에 관련되어 있고 첫 번째 항목은 정규화(Regularization)와 관련되어 있다. 정규화 목표는 인코더의 잠재(latent) 변수 z에 대한 분포 $q_\theta(z|x)$와 $p(z)$의 KL-divergence가 최소화되도록 하는 것이다. β-VAE는 첫 번째 항목인 KL-divergence에 가중치를 주는 β 변수를 설정하는 것이다. 본 연구에서는 $\beta = 0.1$로 설정하였다.

4.2 생성적 적대 신경망 학습

GraspGAN과 일반적인 생성적 적대 신경망과의 학습 시 차이점은 GraspGAN은 시뮬레이터 영상을 현실 세계 영상으로 변환하고자 하는 목표함수 외에도 시뮬레이터 영상으로부터 객체 분할 영상을 추정해야 하는 목표함수를 동시에 사용한다. [Fig. 4]는 본 논문에서 GAN을 위해 사용하고 있는 생성자 망의 구조를 나타내고 있다. 좌측은 입력 영상인 시뮬레이터 영상, 우측 상단은 유사 현실 영상을 나타내고 있고 우측 하단은 객체 분할 영상을 나타내고 있다. 유사 현실 영상을 복원하는 망과 객체 분할 영상을 추정하는 망은 3개의 계층을 제외하고는 대부분의 파라미터를 공유하고 있다. 해당 모델에서 객체 분할 영상을 추가적으로 복원하는 이유는 GAN을 통한 영상변환과정에서 시뮬레이터에 있는 물체의 크기, 위치, 모양 등이 변하지 않고 영상변환과정을 통해서는 단지 색상이나 텍스처 등만이 변화하도록 하기 위함이다. 인코더에서 다운 샘플링 시에는 Average pooling을 사용하고 디코더에서 업 샘플링 시에는 bilinear upsampling을 사용한다. 인코더에서 합성곱 이후의 텐서(Tensor)를 저장하였다가 디코딩 과정에서 Concatenation 하는 skip connections을 사용한다.

5. 현실 세계에서의 학습

[Fig. 5]는 실세계의 영상을 물체 검출 알고리즘과 로봇 암 분할 알고리즘을 사용하여 독립적인 요소로 분리된 영상의 예제이다. 앞서 언급한 대로 시뮬레이터에서의 학습에 비해서 현실 세계에서의 학습을 위해 추가적으로 요구되는 과정은 원 영상을 독립적인 요소로 분리하기 위하여 물체 검출 모델과



[Fig. 5] Example of three levels of disentanglement. At the first level, a robot, a target object, and its periphery are contained in an image. This image is separated into the robot and the focused external world in the second level. There are four images on the third level of disentanglement. Upper Left: position of the robot, Upper Right: visual appearance of the robot, Bottom Left: position of the attended external world, Bottom Right: visual appearance of the attended external world

로봇 암 분할 모델을 학습하는 것이다. 본 절에서는 각각의 학습 방법과 학습에 사용된 데이터에 대해서 언급하고자 한다.

5.1 영상에서의 물체 검출

본 연구에서는 기존의 여러 물체 검출 알고리즘 중 물체 검출의 정확도가 높은 Faster R-CNN[15]을 사용하였고 컨볼루션 신경망(CNN) 구조로는 ResNet 101 구조를 사용하였다. 이때 물체 검출 성능에 가장 큰 영향을 미치는 것은 학습데이터를 어떻게 선정하느냐이다. 학습데이터는 크게 두 가지 종류로 나누어진다.

첫 번째 학습 데이터는 실제 환경으로부터 취득한 데이터를 증강(Augmentation)시켜 생성하였다. 구체적으로 말해서 학습 물체를 일정한 각도만큼 회전시켜 가면서 촬영하고, 객체 분할툴을 사용하여 수작업으로 물체만을 추출한 후, 선정한 배경 영상의 임의의 위치에 물체를 추가하는 방법으로 데이터를 증강시켰다. 둘째로는 3D 스캐너를 사용하여 물체의 3차원 모델을 만들고 이를 시뮬레이터에서 다양한 배경에 임의의 자세로 배치하여 학습에 필요한 데이터를 생성하였다. 첫 번째 방법에 의한 학습데이터는 약 40,000개, 두 번째 방법에 의한 학습데이터도 약 40,000개를 취득하였다.

각각의 방법은 장단점을 가지고 있다. 먼저 첫 번째 방법의 경우 실제 물체를 다양한 자세에서 촬영할 경우 수작업으로 객체 분할을 할 영상이 많아져서 데이터를 생성하는데 많은 시간과 비용을 필요로 한다. 대신 실제 물체를 촬영했기에 물체 검출의 성능을 향상시킬 수 있다. 반면 물체의 3차원 모델을 활용하면 손쉽게 물체의 다양한 자세에 대한 데이터를 생성할 수 있지만 실제 환경의 물체와 3차원 모델 간의 시각적 차이가 있다. 따라서 본 연구에서는 이 두 가지 방법을 혼용하여 데이터를 생성하였다.

강화학습으로 로봇 파지를 학습할 경우 하나의 로봇 파지 작업은 사실 여러 단계의 로봇 행동으로 분할되어 수행된다. 따라서 각 단계마다 물체 검출이 수행되어야 하는데, 이 과정에서 물체 검출이기만을 사용하면 중간에 목표물체의 검출을 실패하는 경우가 종종 발생한다. 특히 로봇이 물체를 잡는 순간에는 그리퍼에 의해서 물체가 일부 가려지는 경우가 많기 때문에 물체 검출이기를 통한 물체 검출의 성능이 저하될 경우가 많다. 그래서 본 연구에서는 물체를 파지하는 과정에서 목표 물체의 지속적인 검출을 위해 물체 검출이기와 함께 물체 추적기술(Object Tracking)을 적용하였다. 물체 추적기술은 영상에서 특정 관심 영역(Region of Interest)을 지정하면 이에 대해 프레임 간의 유사도를 추적해 대상을 추적하는 방법이다. 물체 추적기술로는 CSR-DCF[16] 모델을 사용하였다.

물체 파지 작업 시 최종적인 물체의 검출 과정은 다음과 같다. 물체 검출이기는 로봇의 각 행동단계마다 수행되는데 먼저 물체 검출이기를 이용해 목표 물체 영역을 추출하고 이 정보를 물체 추적기에 관심 영역으로 주면 추적기는 초당 약 30 프레임 정도의 속도로 계속 관심 물체를 추적한다. 로봇의 각 행동단계마다 만약 물체 검출기가 목표 물체의 검출에 성공하면 물체 추적기에 관심 영역을 추출된 물체의 영역 정보로 갱신한다. 만약 파지 단계 도중 물체 검출이기가 물체의 검출에 실패하면 물체 추적기가 지속적으로 추적하던 관심 영역을 물체의 영역 정보로써 사용한다. 이렇게 하면 물체 검출이기가 물체 검출에 실패하더라도 물체 추적기를 통해 지속적인 물체 검출이 가능해진다.

5.2 영상에서의 로봇 암 분할

로봇 암을 영상에서 분할하기 위한 의미적 객체 분할기로는 SegNet[17]을 이용하였다. 물체 검출과 마찬가지로 로봇 암 분할 성능에 가장 큰 영향을 미치는 것은 학습데이터를 어떻게 선정하느냐 이다. 학습데이터는 크게 두 가지 종류로 나누어진다.

첫 번째 학습 데이터는 실제 환경으로부터 로봇을 다양한 자세로 움직이며 취득한 데이터를 객체 분할툴을 사용하여 수작업으로 로봇 암만을 추출하였다. 둘째로는 시뮬레이터의 영상을 GAN을 이용한 영상변환 과정을 거쳐 유사 현실 영상으로 만들었다. 해당 영상에 대한 레이블 영상은 손쉽게 시뮬레이터에서 제공 받을 수 있다. 첫 번째 방법에 의한 학습 데이터는 약 800개, 두 번째 방법에 의한 학습데이터는 약 15,000개를 취득하였다.

역시 각각의 방법은 장단점을 가지고 있다. 먼저 첫 번째 방법의 경우 실제 로봇 암을 다양한 자세에서 촬영할 경우 레이블 데이터를 얻기 위해 수작업으로 로봇 암 분할을 할 영상이 많아져서 데이터를 생성하는데 많은 시간과 비용을 필요로 한다. 대신 실제 로봇 암을 촬영했기에 로봇 암 분할 성능을 향상시킬 수 있다. 반면 시뮬레이터에서 변환한 유사 현실 영상을 활용하면 로봇 암의 다양한 자세에 대한 레이블 데이터를 손쉽게 생성할 수 있지만 실제 환경의 로봇 암과 유사 현실 영상의 로봇 암 간의 시각적 차이가 다소 존재한다. 따라서 본 연구에서는 이 두 가지 방법을 혼용하여 데이터를 생성하였다.

6. 실 험

6.1 로봇 파지실험을 위한 설정

시뮬레이터에서 로봇 파지를 학습하기 위해서 본 연구에서는 Bullet 시뮬레이터를 사용하였다[3]. 학습 시에는 한 대의 PC에 다섯개의 시뮬레이터를 동시에 구동시켜 학습을 진행하였다. 실세계와 시뮬레이터 모두 [Fig. 6]의 좌측과 같은 UR5 로봇과 Robotiq 그리퍼를 사용하였고 한 대의 RGB 카메라가 고정된 위치에서 로봇과 물체가 놓여있는 선반을 관측하였다. 영상은 자르고 크기를 변환하여서 360x360 의 해상도로 변환된 후에 Disentanglement 과정 후에는 각 분리된 영상의 해상도는 128x128이 된다. 파지에 사용된 물체는 [Fig. 6]의 우측과 같이 17개 이다. 상태표현 학습을 위해서 앞서 언급했듯이 β-VAE (β=0.1) 를 사용하였고 분리 수준으로는 L3 를 사용하였다.

심층 강화학습을 위해서는 D4PG를 사용하였다. 강화학습의 정책망과 가치망의 입력은 사전에 학습한 상태표현 학습기

[Fig. 6] Left: the experimental setup. Right: 17 objects for the grasping task

의 인코더의 출력 벡터를 사용하였고 이는 256차원의 벡터이다. 정책망의 출력은 로봇 파지 자세로 5차원의 벡터이다. 구체적으로 로봇 엔드 이펙터의 위치 (x, y, z)와 마지막 조인트의 각도(θ), 그러퍼의 상태(0.5이상: 열림, 0.5미만: 닫힘)이다.

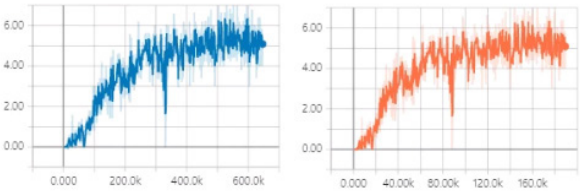
로봇 파지 학습을 위한 데이터를 취득하기 위해서 로봇은 정책망의 로봇 파지 자세에 가우시안 노이즈를 추가하여 행동한다. 하지만 이와 같은 정책 만으로는 학습 초기에 파지에 성공한 데이터를 획득하기는 쉽지 않다. 따라서 학습 초기에도 파지 성공을 10~20%내외로 달성할 수 있는 휴리스틱 정책을 동시에 사용하게 되는데 이를 대본(scripted) 정책이라고 한다[3]. 학습 초기에는 이러한 대본 정책의 비중이 1이고 정책망을 이용한 정책이 0으로 시작하지만 학습이 진행됨에 따라 정책망을 이용한 정책은 선형적으로 증가하며 95%까지 증가한다. 강화학습 관련 파라미터 설정은 정책망과 가치망을 위한 학습률은 각각 10^{-4}와 10^{-3}이고 Adam optimizer[3]를 사용했다. Discount factor $\gamma = 0.99$는 soft target update $\tau = 0.01$을 사용했다. 학습 시의 배치 데이터의 수는 32이다.

실세계에서 강화학습을 수행할 경우에는 다른 학습 관련 파라미터는 시뮬레이터에서의 학습 시와 동일하지만 크게 다른 점은 우선 학습 속도와 데이터 취득 속도의 비율을 조정하기 위해 새로운 파라미터가 필요하다는 점이다. 시뮬레이터에서의 학습 시에는 로봇이 5개가 구동되는 것과 마찬가지이고 시뮬레이터에서의 각 로봇의 움직임은 실세계에서의 로봇보다 빠르다. 따라서 학습데이터의 취득 속도는 실세계에서의 취득 속도보다 10~20배 정도 빠르다. 따라서 실세계에서 학습 시 학습 속도를 시뮬레이터와 동일하게 했을 경우 적은 데이터를 기반으로 최적화를 시도해서 학습이 국부적 최소점에 빠져 실패하게 된다. 따라서 실세계에서 학습 시 학습 속도를 시뮬레이터에서의 학습 속도에 비해 20배 느리게 해서 학습하였다.

또 한 가지 고려할 점은 정책망과 가치망을 위한 학습률 설정이다. Adam optimizer의 특성상 시뮬레이터에서 학습이 진행될수록 학습 목표를 만족하는 최소점에 거의 도달했기 때문에 해당 학습률은 적어지도록 되어 있다. 실세계에서의 학습은 시뮬레이터의 학습과 달리 학습 시작단계부터 학습 목표를 만족하는 최소점에 상당히 근접해 있다고 볼 수 있다. 따라서 학습률은 시뮬레이터 보다 상당히 작게 설정하지 않으면 역시 학습이 실패하게 된다. 실세계에서 학습 시 정책망과 가치망을 위한 학습률은 각각 30^{-6}과 30^{-5}으로 설정하였다.

6.2 로봇 파지 실험 결과

[Fig. 7]은 시뮬레이터에서의 학습 결과를 나타내고 있다. 평가를 하기 위해 학습 중에 일정한 주기로 최대 17개의 물체를 선반에 배치하고 이 중 10개의 임의로 선택된 물체를 파지



[Fig. 7] Performance curve during training in simulation. Y-axis is the number of successful grasp. Left : x-axis is the training iteration. Right: x-axis is the number of grasp

하는 테스트를 수행하였다. 이 경우 물체들의 위치는 모두 랜덤이고 따라서 파지 대상의 물체 주변에 물체들이 존재하여 올바른 파지 자세 임에도 불구하고 그리퍼와 주변 물체가 충돌이 되는 경우가 빈번히 발생되었다. 두 그래프의 y 축은 파지에 성공한 횟수이다. 왼쪽 그래프의 x축은 시뮬레이터에서의 학습 횟수(학습 1회는 하나의 배치데이터를 이용해 모델 파라미터를 한 번 변경한 경우임)이고 최대 약 600 k이다. 오른쪽 그래프의 x축은 학습데이터(학습데이터 1개는 사전 상태, 로봇 동작, 사후 상태, 보상으로 이루어짐)의 양이고 최대 약 180 k이다. 최대 성공률은 약 53%이다.

실세계에서 학습의 경우 학습 횟수는 5 k, 학습데이터의 양은 3 k이다. 학습 시간의 경우 약 60시간이 소요되었다. 실세계에서 파지 실험은 [Fig. 6]의 우측 그림의 물체 중 5~7개의 물체를 임의로 선택하여 파지를 수행하였다. 물체는 [Fig. 6]의 좌측 그림과 같이 일반적으로 파지 시 주변 물체에 충돌하지 않도록 어느 정도 떨어진 위치에 배치하였다. 각 물체에 대해서 한 번씩 파지를 수행하고 다시 5~7개의 물체를 임의로 선택하여 파지를 수행하였다. 이런 식으로 반복하여 수행한 총 파지 시도 횟수는 103번이고 성공 횟수는 73이다. 따라서 파지 성공률은 73/103 * 100 = 68.8%이다. 이는 구글의 연구들[2,5]에서 약 70~80%의 파지 성공률을 달성하기 위해 취득한 실세계에서의 학습데이터의 양이 약 500~700 k인 것에 비하면 1/100 보다 적은 실세계 데이터를 사용한 것이다.

7. 결 론

구글에서 연구한 가치 기반 심층 강화학습 혹은 최신의 심층 강화학습 기반 연구 방법들은 계산량 혹은 취득해야 할 실세계에서의 데이터의 양이 너무 많아서 실제로 사용하는 데에 제약이 있다. 본 논문에서는 정책/가치 심층 강화학습에 기반하여 현실 세계에서의 로봇 파지를 위한 학습 플랫폼을 제안하였다.

구체적으로 말해서 제안하는 학습 플랫폼은 시뮬레이터의 학습과 현실 세계의 학습으로 구성되어 있고 주요 구성요소는 Disentanglement 기반한 상태표현 학습기, 영상변환을 위한 생성적 적대 신경망, 실세계에서 Disentanglement을 위한 물체

검출이기와 의미적 객체 분할기다. 제안하는 방법은 계산량과 실제데이터의 양 두 가지 측면을 고려했을 때 현실적으로 로봇 파지를 위해 사용 가능한 방법이고 또한 파지 성공률도 약 70%에 이를 정도이다.

한 가지 실험과정에서 추가적으로 확인할 수 있었던 점은 시뮬레이터에서 학습한 파라미터를 초기 파라미터로 사용하고 실세계에서는 전혀 학습하지 않은 상태에서도 정책망의 추정한 로봇 파지 자세 중 비록 z 값이 상당히 차이가 있어서 실제로 파지에 성공하지는 못했지만 (x, y, θ)는 거의 정확했다는 점이다. 실세계의 데이터 숫자가 기존의 다른 연구들에 비해서 약 1/100 정도만을 사용해서 높은 파지 성공률을 달성했다는 점뿐만 아니라 이와 같은 관찰을 통해서도 시뮬레이터의 학습이 효율적으로 실세계의 학습으로 전이됐다는 것을 확인할 수 있었다.

본 논문에서 제안한 학습 플랫폼은 플랫폼이라는 정의에 따라 특정 로봇이나 물체에 의존적이지 않다. 따라서 본 연구에서는 UR5 로봇과 17개의 물체를 사용했지만 제안하는 학습 플랫폼 하에서 시뮬레이터에서의 로봇 모델의 변화나 물체 검출기 혹은 로봇 암 분할 학습의 학습 대상 물체를 변경한다면 다양한 로봇 혹은 다양한 물체를 대상으로 강화학습 기반의 파지를 수행할 수 있다.

References

[1] D. Quillen, E. Jang, O. Nachum, C. Finn, J. Ibarz, and S. Levine, "Deep reinforcement learning for vision-based robotic grasping: A simulated comparative evaluation of off-policy methods," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, QLD, Australia, 2018, DOI: 10.1109/icra.2018.8461039.

[2] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, and S. Levine, "Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation," *arXiv:1806.10293,* 2018, [Online], https://arxiv.org/abs/1806.10293.

[3] T. Kim, Y. Park, Y. Park, and I. H. Suh, "Acceleration of Actor-Critic Deep Reinforcement Learning for Visual Grasping in Clutter by State Representation Learning Based on Disentanglement of a Raw Input Image," *arXiv:2002.11903*, 2020, [Online], https://arxiv.org/abs/2002.11903.

[4] J. Mahler, M. Matl, X. Liu, A. Li, D. Gealy, and K. Goldberg, "Dex-Net 3.0: Computing robust vacuum suction grasp targets in point clouds using a new analytic model and deep learning," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, QLD, Australia, 2018, DOI: 10.1109/icra.2018.8460887.

[5] S. Levine, P. Pastor, A. Krizhevsky, and D. Quillen, "Learning hand-eye coordination for robotic grasping with large-scale data collection," *International Symposium on Experimental Robotics*, pp. 173-184, 2016, DOI: 10.1007/978-3-319-50115-4_16.

[6] C. Finn, X. Y. Tan, Y. Duan, T. Darrell, S. Levine, and P. Abbeel, "Deep spatial autoencoders for visuomotor learning," *2016 IEEE International Conference on Robotics and Automation (ICRA)*, Stockholm, Sweden, 2016, DOI: 10.1109/icra.2016.7487173.

[7] H. van Hoof, N. Chen, M. Karl, P. van der Smagt, and J. Peters, "Stable reinforcement learning with autoencoders for tactile and visual data," *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Daejeon, South Korea, 2016, DOI: 10.1109/iros.2016.7759578.

[8] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework," *ICLR*, 2017, [Online], https://www.semanticscholar.org/paper/beta-VAE%3A-Learning-Basic-Visual-Concepts-with-a-Higgins-Matthey/a90226c41b79f8b06007609f39f82757073641e2.

[9] K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. Pastor, K. Konolige, S. Levine, and V. Vanhoucke, "Using simulation and domain adaptation to improve efficiency of deep robotic grasping," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, QLD, Australia, 2018, DOI: 10.1109/ICRA.2018.8460875.

[10] G. Barth-Maron, M. W. Hoffman, D. Budden, W. Dabney, D. Horgan, A. Muldal, N. Heess, and T. Lillicrap, "Distributed distributional deterministic policy gradients," *arXiv:1804.08617*, 2018, [Online], https://arxiv.org/abs/1804.08617.

[11] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv:1509.02971*, 2016, [Online], https://arxiv.org/abs/1509.02971.

[12] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, 2010, [Online], https://www.semanticscholar.org/paper/Stacked-Denoising-Autoencoders%3A-Learning-Useful-in-Vincent-Larochelle/e2b7f37cd97a7907b1b8a41138721ed06a0b76cd.

[13] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv:1804.02767*, 2018, [Online], https://arxiv.org/abs/1804.02767.

[14] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv:1312.6114*, 2013, [Online], https://arxiv.org/abs/1312.6114.

[15] S. Ren, K. He, R. Grisshick, and J. Sun, "Faster R-CNN: Toward Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, DOI: 10.1109/tpami.2016.2577031.

[16] A. Lukezic, T. Vojir, L. C. Zaic, J. Matas, and M. Krstan, "Discriminative Correlation Filter Tracker with Channel and Spatial Reliability," *International Journal of Computer Vision*, 2019, DOI: 10.1007/s11263-017-1061-3.

[17] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481-2495, 2017, DOI: 10.1109/tpami.2016.2644615.

김 태 원

2017 목원대학교 지능로봇공학(공학사)

2018~현재 한양대학교 지능형로봇학과 (석사과정)

관심분야: Deep Learning, Reinforcement Learning, Robotic Manipulation, Computer Vision

박 예 성

2011 홍익대학교 서울캠퍼스 기계시스템디자인공학과(학사)

2018~현재 한양대학교대학원 서울캠퍼스 지능형로봇학과(석사)

관심분야: Robotic Manipulation, Machine Learning, Visual Recognition

김 종 복

2001 조선대학교 제어계측공학과(학사)

2003 한양대학교대학원 전자전기제어계측공학과(석사)

2003~현재 한양대학교대학원 전자컴퓨터통신공학과(박사 수료)

관심분야: 기계 학습, 로봇 제어

박 영 빈

2001 한양대학교 사회학과(학사)

2007 한양대학교 전자정보통신공학과(석사)

2013 한양대학교 전자정보통신공학과(박사)

2014~현재 한양대학교 네트워크기반지능형로봇교육센터 연구교수

관심분야: 기계학습, 로봇제어, 컴퓨터비전

서 일 홍

1977 서울대학교 전자공학과(공학사)

1979 한국과학기술원 전자공학과(공학석사)

1982 한국과학기술원 전자공학과(공학박사)

1985~현재 한양대학교 공과대학 교수

관심분야: 기계학습, 로봇제어, 실내 및 실외 주행