Journal of Korea Robotics Society (2023) 18(3):352-357
https://doi.org/10.7746/jkros.2023.18.3.352 ISSN: 1975-6291 / eISSN: 2287-3961

강화학습의 신속한 학습을 위한 변이형 오토인코더 기반의 조립 특징 추출 네트워크

Variational Autoencoder-based Assembly Feature Extraction Network for Rapid Learning of Reinforcement Learning

윤 준 완[1] · 나 민 우[2] · 송 재 복†
Jun-Wan Yun[1], Minwoo Na[2], Jae-Bok Song†

Abstract: Since robotic assembly in an unstructured environment is very difficult with existing control methods, studies using artificial intelligence such as reinforcement learning have been conducted. However, since long-time operation of a robot for learning in the real environment adversely affects the robot, so a method to shorten the learning time is needed. To this end, a method based on a pre-trained neural network was proposed in this study. This method showed a learning speed about 3 times than the existing methods, and the stability of reward during learning was also increased. Furthermore, it can generate a more optimal policy than not using a pre-trained neural network. Using the proposed reinforcement learning-based assembly trajectory generator, 100 attempts were made to assemble the power connector within a random error of 4.53 mm in width and 3.13 mm in length, resulting in 100 successes.

Keywords: Robotic Assembly, Reinforcement Learning, Deep Learning

1. 서 론

최근 산업 현장에서 로봇을 이용한 조립 작업 자동화의 수요가 높아지고 있다. 일반적으로 로봇을 이용한 조립작업은 로봇과 조립 대상 부품의 자세 정보가 미리 정의된 정형화된 환경에서 수행된다. 하지만 부품이 무작위로 배치된 비정형 환경에서는 조립 작업을 자동화하기 어렵다. 이에 접촉력을 이용하여 부품 사이의 오차를 줄이기 위해 blind search[1]와 같은 방법들이 제안되었지만, 대상 부품에 따라 탐색 궤적을 미리 정의하고, 적절한 파라미터를 설정해야 한다. 또한 불확실한 환경 오차가 존재하는 비정형 환경에서는 성공률이 매우 낮아지는 단점이 존재한다.

한편, 조립 작업은 접촉이 수반되므로 부품 사이에 오차가 존재하면 과도한 힘이 발생하기에 적절한 힘제어가 필수적이다. 이때, 비정형 환경에서 안정적인 접촉을 유지하려면 적절한 접촉 모델이 필요하지만, 실제 접촉 거동을 정확히 모델링하는 것은 한계가 있다.

위와 같은 로봇 기반 조립의 한계를 극복하기 위해서 인공지능을 활용하는 연구들이 진행되고 있다[2,3]. 특히 복잡한 접촉 모델을 시행착오 방식을 통해 구현하는 강화학습이 로봇 조립작업에 적합한 방식으로 주목받고 있다[4-6]. 이들은 강화학습 기반의 로봇 궤적 생성기를 제안하였는데, [4]는 영상 정보를 기반으로, [5, 6]은 힘/토크 정보를 기반으로 두 부품 사이의 정렬을 위한 로봇 궤적을 출력하였다. 이 방법 모두 조립을 성공적으로 수행하였지만, 영상과 같은 고차원 정보를 네트워크 입력으로 곧바로 사용하여 실제 환경에서의 학습 시간이 길어지는 단점이 있다. 학습 시간의 증가는 로봇 동작 시간의 증가로 이어지며, 학습이 실패하는 경우가 늘어날수록 로봇과 조립물의 손상이 발생할 수 있으므로 실제 환경에서의 학습 시간 단축은 중요하다.

Received : May. 31. 2023; Revised : Jul. 13. 2023; Accepted : Jul. 17. 2023
※ This research was funded by the MOTIE under the Industrial Foundation Technology Development Program supervised by the KEIT (No. 20008613)
1. Master Student, Mechanical Engineering, Korea University, Seoul, Korea (wns9677@korea.ac.kr)
2. Ph.D Student, Mechanical Engineering, Korea University, Seoul, Korea (jrmwl@korea.ac.kr)
† Professor, Corresponding author: Mechanical Engineering, Korea University, Seoul, Korea (jbsong@korea.ac.kr)

한편, 영상과 힘/토크 정보를 모두 사용하되, 학습시간을 단축하기 위해 영상정보에 대하여 선행 학습된 네트워크를 사용하는 방법도 제안되었다[7]. 하지만 힘/토크 및 속도와 같은 시계열 정보들은 선행 학습되어 있지 않았으며, 일부 알고리즘에서는 로봇의 자세를 사용하였으므로 비정형 환경에서는 적용이 불가능하다.

본 연구에서는 영상과 힘/토크 및 속도를 모두 포함한 고차원의 다양한 조립 정보들을 저차원으로 변환시켜주는 선행 학습된 신경망인 조립 특징 인코더(assembly feature encoder, AFE)를 사용하는 방법을 제안한다. 이를 통해 생성된 저차원 정보는 조립 정보들의 특징(feature)을 보유하고 있으므로 강화학습의 입력으로 적절하며, 실제 환경에서의 강화학습의 학습 시간을 단축시킬 수 있다.

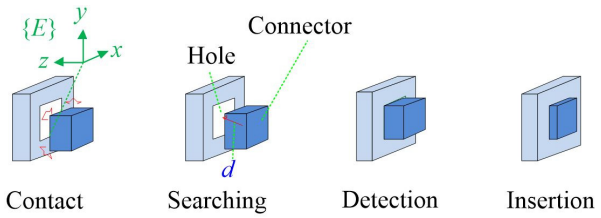
본 연구의 기여는 다음과 같다. 첫째, 고차원의 다양한 조립 정보들을 원본의 특징을 보유한 저차원 데이터로 변환한다. 둘째, 선행 학습된 네트워크인 AFE를 통하여 강화학습의 학습 시간을 단축한다. 셋째, 최적의 조립 궤적을 생성하는 강화학습 기반 로봇 조립 궤적 생성기를 구현한다.

본 논문의 구성은 다음과 같다. 2장에서는 강화학습 기반 로봇 조립 궤적 생성기를 설명한다. 3장에서는 AFE 구조, 학습에 사용한 데이터, 학습 결과, 그리고 AFE를 사용하는 강화학습 기반 로봇 조립 궤적 생성기를 설명한다. 4장에서는 AFE를 활용하는 강화학습 모델과 사용하지 않는 강화학습 모델의 학습 결과와 조립결과를 비교하며, 5장에서는 논문의 결론을 도출한다.

2. 강화학습 기반 로봇 조립

2.1 커넥터 조립전략

로봇을 이용한 커넥터 조립은 [Fig. 1]과 같이 접촉(contact), 탐색(searching), 탐지(detection), 삽입(insertion)으로 구성된다. 로봇에 의해 파지된 커넥터는 커넥터 구멍에 접촉한 후, 접촉을 유지한 채로 구멍을 탐색한다. 이때, 과도한 접촉력을 방지하기 위해 임피던스 제어[8]를 통해 적절한 접촉을 유지한다. 최종적으로 구멍이 탐지되면 커넥터를 구멍에 삽입하여 조립을 완료한다. 본 연구에서는 탐색 과정에서 커넥터와 커넥터 구멍 사이의 자세 오차 d를 줄이도록 영상과 접촉력을 이용한다. 커넥터와 구멍이 모두 포함되는 영상 정보를 통해 상대 자세 오차 d를 식별할 수 있다. d를 줄이도록 로봇이 구동되는 동안 접촉력을 사용하여 재밍과 같은 비정상적인 접촉 상황을 확인한다.



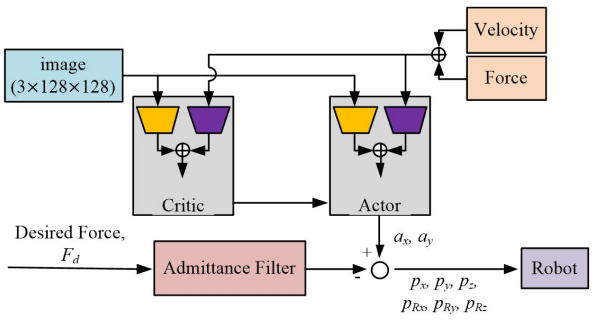
[Fig. 1] Connector assembly strategy

2.2 강화학습

강화학습은 에이전트(agent)가 현재 상태(current state, S_t)에서 취한 행동(action)으로부터 보상(reward)과 다음 상태(next state, S_{t+1})를 관찰하여 각 행동에 대해 획득할 수 있는 보상을 최대화하는 정책(policy)를 학습한다. 로봇 조립에서 에이전트는 로봇이며, 현재 상태에서 획득한 영상 및 접촉력을 상태 정보로 사용한다. 이때, 다음 상태에서 d를 최소화하도록 로봇의 궤적을 출력하면 높은 보상을 획득한다. 조립을 위한 로봇의 궤적은 연속 공간상에 존재하므로 본 연구에서는 연속 공간에서 강인한 성능을 보인다고 알려져 있는 Soft-Actor-Critic (SAC)[9]을 사용한다.

강화학습 기반 로봇 조립 궤적 생성기는 [Fig. 2]과 같다. 에이전트는 입력된 상태 정보로부터 행동(a_x, a_y)을 출력하며, 행동을 통해 로봇 조립의 탐색 과정을 수행한다. 상태 정보는 탐색 과정에서 로봇 행동의 결과로 획득한 영상, 이전 시점과 현재 시점에서 측정한 힘/토크 및 로봇의 말단 속도가 사용된다. 영상과 여러 시점에서 누적된 힘/토크 및 속도 정보는 고차원 정보이며, 이러한 정보로부터 오차 d를 줄이는 특징을 추출함으로써 궤적 생성기의 정책이 학습된다. 이 과정에서 고차원의 정보는 액터-크리틱 구조 내에서 저차원의 특징으로 변환되며, 이러한 저차원의 정보를 조립 특징으로 정의한다.

영상의 경우 인간의 시신경을 모방한 합성곱 신경망(convolutional neural network, CNN)을 통해 영상의 특징이 압축된 저차원 벡터인 특징맵(feature map)으로 변환된다. 힘/토크 및 속도와 같은 다시점 정보는 시계열 데이터 분석에 용이한 순



[Fig. 2] Structure of reinforcement learning-based assembly trajectory generator

환 신경망(recurrent neural network, RNN)을 사용한다. 이때, 적절한 특징을 획득하려면 다층의 신경망을 사용해야 하는데, 이는 학습 시간을 증가시키는 주요 요인 중 하나이다. 특히 강화학습은 에이전트인 로봇이 실제 환경에서 실시간으로 구동되어야 하므로, 학습 시간의 증가는 심할 경우에 로봇 및 커넥터를 손상시킬 수 있다. 만약, 액터-크리틱 내부의 특징 추출 구조를 오프라인 방식으로 선행 학습할 수 있다면 강화학습의 학습 시간을 단축시킬 수 있다.

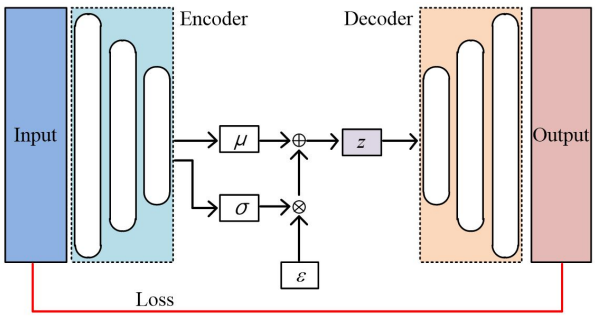
3. 오프라인 방식의 조립 특징 추출 모델

본 논문에서는 강화학습의 액터-크리틱 내부에 존재하는 조립 특징 추출 네트워크를 별도의 사전 학습된 조립 특징 인코더(assembly feature encoder, AFE)로 대체하여 특징 추출 과정을 오프라인으로 수행하는 조립 특징 추출 네트워크를 제안하였다. 특징 추출 네트워크는 상태 정보들의 특징을 추출하도록 Variational Auto Encoder (VAE)[10]를 사용하였다. VAE는 [Fig. 3]과 같이 고차원의 입력 정보를 압축하여 저차원 벡터인 잠재 벡터(Latent Vector, z)로 변환하는 전반부의 인코더와 잠재 벡터를 입력 데이터로 복원하는 디코더로 이루어진 비지도학습 신경망이다.

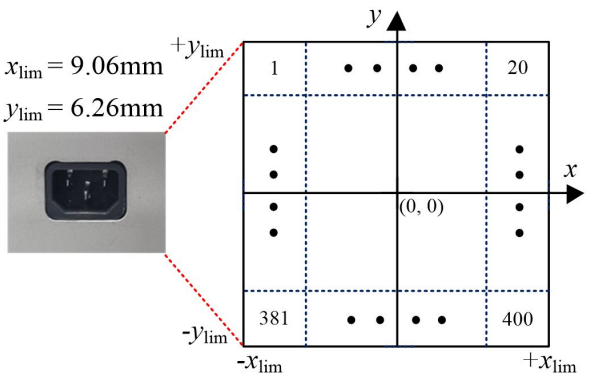
잠재벡터는 입력 데이터의 평균(μ), 표준편차(σ)와 무작위 노이즈(ε)을 통해 생성된다. 인코더를 통해서 입력 데이터의 분포는 정규분포에 가깝도록 변환되며, 변환된 데이터인 잠재벡터는 입력 데이터의 특징을 포함한다. 이때, 인코더의 성능이 좋을수록 표준편차가 1에 가까운 잠재 벡터가 생성되며, 노이즈에 강인한 잠재 벡터를 생성할 수 있다.

AFE를 통해 얻은 잠재벡터를 조립 특징으로 간주하므로, 미리 획득된 조립 정보를 오프라인 학습에 활용할 수 있다. 힘/토크 및 속도의 경우, 실제 조립 작업 중에 발생한 데이터를 사용한다. 앞서 설명한 조립전략을 통해 얻은 다품종 커넥터의 접촉 양상은 유사하므로, 다양한 커넥터로부터 미리 획득한 조립 정보를 활용하여 학습이 가능하다. 따라서 학습된 모델은 현재 대상 커넥터에 국한되지 않는다. 힘/토크 정보와 달리 영상 정보의 경우, 특정 상태에 대해 과학습(over-fitting)된 특징이 추출될 가능성이 높다. 이를 방지하려면, 다양한 상태를 포함하는 고른 분포를 가진 데이터가 필요하다. 본 연구에서는 [Fig. 4]와 같이 로봇의 궤적을 설정하여 영상 정보를 고르게 획득하도록 구성하였다. 궤적은 정해진 범위 내에서 400개의 구간으로 나뉘어져, 각 구간 내에서 로봇이 무작위로 움직인다. 커넥터 조립 시 로봇이 동작하는 범위가 작으므로, 이 궤적을 통해 조립 작업 중 발생 가능한 대부분의 상태에 대한 영상 정보들을 획득할 수 있다. 이처럼 데이터의 특성이 다르므로 AFE는 영상의 특징 추출을 위한 영상 특징 인코더(image feature encoder, IFE)와 힘/토크 및 속도를 포함하는 시계열 정보의 특징 추출을 위한 시계열 특징 인코더(time-series feature encoder, TFE)로 이루어져 있다. IFE의 구조는 7개의 층으로 구성된 CNN으로, TFE는 RNN을 경량화한 모델인 GRU[11]를 4개의 층으로 구성하였다.

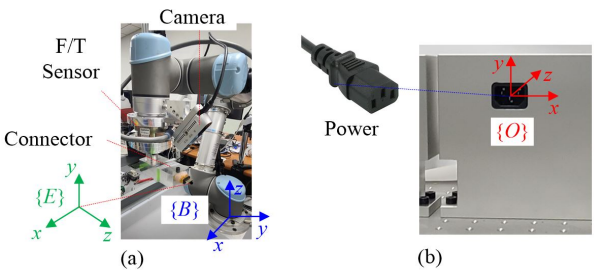
AFE의 학습 데이터 수집을 위한 환경 구성은 [Fig. 5(a)]와 같다. 로봇은 Universal Robot사의 UR5CB를 사용하였으며, 카메라는 Intel사의 Realsense D435, 힘/토크 센서는 ATI사의 Gamma를 사용하였다. 조립 대상 커넥터는[Fig. 5(b)]와 같은 파워 커넥터와 커넥터 구멍을 사용하였다. 좌표계 $\{B\}$는 로봇의 기저 좌표계, $\{E\}$는 로봇의 말단 좌표계, $\{O\}$는 커넥터 구멍의 좌표계이다.



[Fig. 3] Structure of VAE

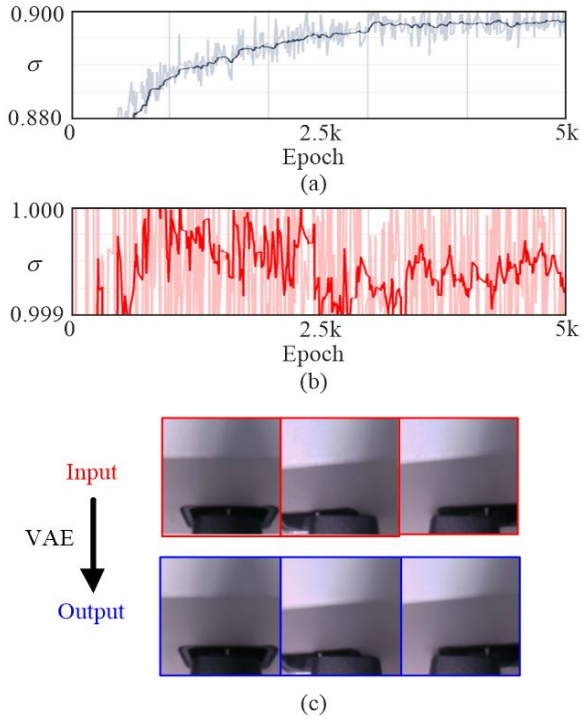


[Fig. 4] Divided regions for image acquisition

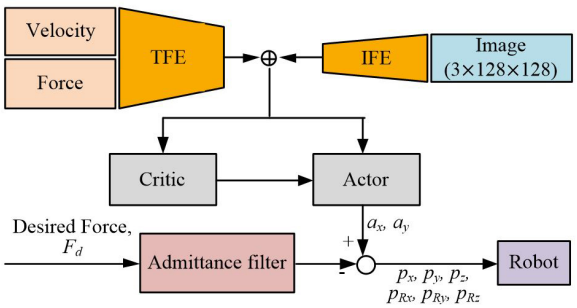


[Fig. 5] (a) Eenvironmental setup, and (b) assembly parts

AFE의 학습 결과는 [Fig. 6]과 같다. 영상 정보는 약 4만 개, 접촉 정보는 300회의 조립을 수행하여 얻은 약 10만 개의 데이터를 사용하였다. IFE를 학습하여 얻은 잠재벡터의 표준편차는 약 0.9, TFE의 경우는 1이 산출되었다. 또한 [Fig. 6(c)]로부터 IFE는 입력 이미지를 원본 이미지로 잘 복원함을 확인할 수 있다. [Fig. 7]은 AFE를 강화학습 네트워크에 적용한 구조를 보여준다. 기존의 강화학습 네트워크와 달리 이 구조는 액터와 크리틱의 외부에서 선행 학습된 특징 추출 네트워크를 사용하므로, 강화학습의 학습 파라미터의 수를 줄여 학습 시간을 단축시킬 수 있다. 또한 액터와 크리틱 네트워크의 구조가 변경되더라도 선행 학습된 AFE는 변경할 필요가 없다.



[Fig. 6] Results of AFE: (a) standard deviation of IFE, (b) standard deviation of TFE, and (c) input image and reconstruction image of IFE

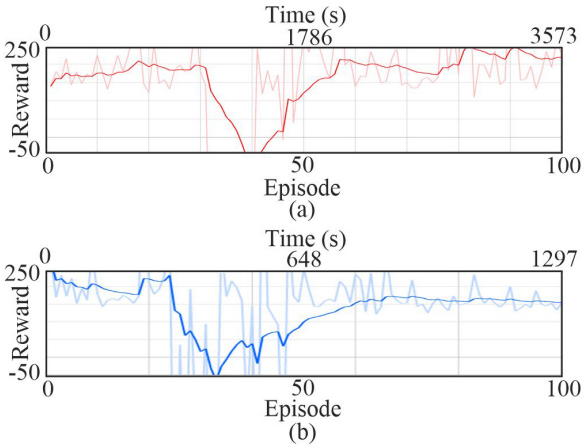


[Fig. 7] Structure of reinforcement learning-based assembly with AFE

4. 실험 결과

제안된 방법을 검증하기 위하여 가로 22.65 mm, 세로 15.65 mm의 형상을 가진 파워 커넥터의 탐색 작업에 대해 AFE를 사용한 경우와 사용하지 않은 강화학습 네트워크를 각각 학습하였다. AFE를 사용하지 않은 강화학습 모델의 액터는 AFE와 동일한 구조의 압축부와 ReLU를 사용한 4층의 전결합층으로 구성되었으며, 크리틱 또한 동일한 압축부와 ReLU를 사용한 5층의 전결합층으로 구성되었다. AFE를 사용할 경우도 동일한 구조의 액터 및 크리틱으로 구성되었다. 강화학습을 위한 실험 환경은 [Fig. 5]와 동일하다. 커넥터의 초기 위치 오차 범위는 커넥터 형상의 20%로 지정하였으며, 파워 커넥터의 경우, $\{O\}$의 x축 방향으로 ±4.53 mm, y축 방향으로 ±3.13 mm 이내의 무작위 값으로 지정하였다. 강화학습의 학습 결과는 [Fig. 8]과 같다. AFE를 사용하지 않은 경우는 약 63번째, AFE를 사용한 경우는 55번째의 에피소드에서 수렴하였다. 이때 학습 시간은 각각 31분 24초, 12분 32초가 소요되었다. 또한 [Table 1]과 같이 AFE를 사용하지 않으면 수렴 이후에도 보상의 변동이 심했지만, AFE를 사용한 경우 보상의 변동이 작았다. 이를 통해 AFE를 사용한 경우의 학습 속도가 약 3배 빠르며, 보상이 비교적 안정적으로 수렴함을 알 수 있다.

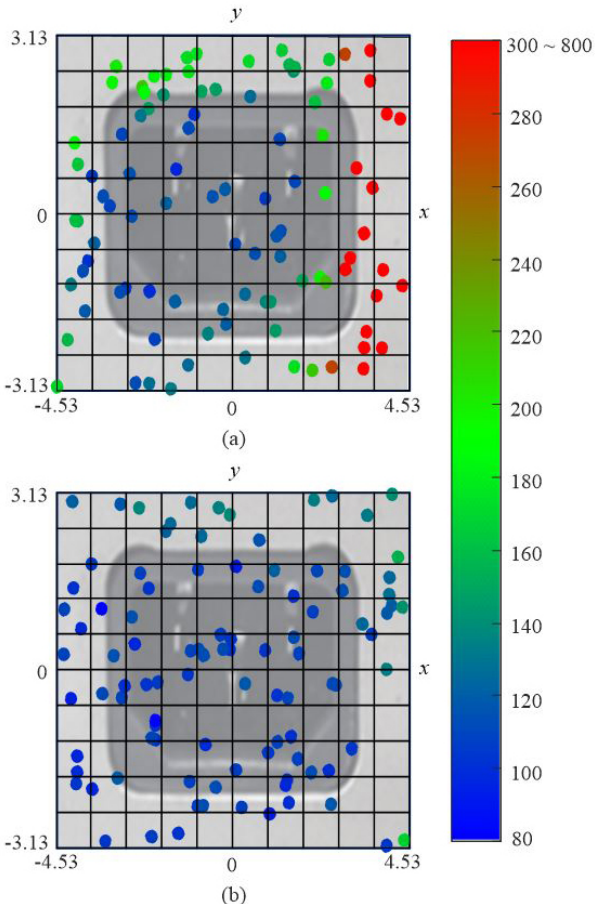
[Fig. 9]는 두 네트워크의 100번째 에피소드 모델을 사용하여 무작위 오차가 존재하는 경우에 대해 조립의 탐색 과정을 수행한 결과를 보여준다. 조립은 총 100회를 수행하였다. 각



[Fig. 8] Learning results: (a) without AFE, and (b) with AFE

[Table 1] Results of RL learning reward

	Without AFE	With AFE
Median	179.3	153.5
Average	214.8	166.3
Standard deviation	102.0	35.1



[Fig. 9] Assembly results and the number of action steps per episode: (a) without AFE, and (b) with AFE

[Table 2] Result of action step

	Without AFE	With AFE
Median	152	109
Average	232	111.39
Standard deviation	221.40	13.95

축은 각각 $\{O\}$ 기준의 x, y축을 나타내며, 각 점은 무작위로 분포된 탐색 초기 위치이다. 이때 탐색 과정을 완료할 때까지 출력된 행동 횟수를 청색과 적색 사이의 보간 색상표로 표시하였다. 커넥터의 초기 위치 오차 범위는 학습 시의 초기 오차와 동일하다. AFE를 사용한 경우에는 100회 모두 조립에 성공하였고, AFE를 사용하지 않은 경우에는 99회 성공하였다.

정책이 이상적으로 학습된 경우, 각 상태에 대해 적절한 Q가 형성되었으므로, [Fig. 9]의 원점으로부터 멀어질수록 행동의 횟수가 선형적으로 증가한다. [Fig. 9(a)]로부터 AFE를 적용하지 않은 경우에 출력된 행동의 수가 우측이 높다는 것을 알 수 있다. 반면에, [Fig. 9(b)]에서 보듯이 AFE를 사용할 경우에는 상대적으로 출력된 행동이 고르게 분포되었다. 또한 [Table 2]는 출력된 행동 횟수에 대한 중간값, 평균, 표준편차를 산출한 결과를 보여주는데, AFE를 사용할 경우 조립 성공까지 출력된 행동의 수와 변동이 더 작았다. 이로부터 AFE를 사용할 경우에 강화학습의 정책이 다양한 상태에 대해 더욱 최적의 행동을 출력함을 알 수 있다.

5. 결 론

본 연구에서는 조립 상태의 특징 추출을 위해 선행 학습된 인코더인 AFE를 활용하여 강화학습의 학습 시간을 단축시키고, 학습 시 보상의 안정성을 증가시키는 전략을 제안하였다. 이를 위하여 적절한 조립 데이터와 고른 분포를 가진 RGB 영상 데이터를 생성하여 AFE를 학습시켰다. 학습된 AFE를 이용하여 강화학습을 구성하였으며, AFE를 사용하지 않을 경우와 비교하였을 때, 신속하고 안정적인 학습이 가능하였다. 학습된 AFE를 사용하는 강화학습 모델을 이용하여 100회의 파워 커넥터 탐색 작업을 진행하였으며, 모두 성공하였다. 또한 AFE를 사용하지 않을 경우에 비해 최적의 정책을 학습하였다는 점을 확인할 수 있었다. 향후 연구에서는 파워 커넥터 외의 다른 커넥터로 대상을 확대하여 연구를 수행할 예정이다.

References

[1] S. R. Chhatpar and M. S. Branicky, “Search strategies for peg-in-hole assemblies with position uncertainty,” *IEEE International Workshop on Intelligent Robots and Systems (IROS)*, Maui, USA, pp. 1465-1470, 2001, DOI: 10.1109/IROS.2001.977187.

[2] L. Xie, H. Yu, Y. Zhao, H. Zhang, Z. Zhou, M. Wang, and R. Xiong, “Learning to Fill the Seam by Vision: Sub-millimeter Peg-in-hole on Unseen Shapes in Real World,” *IEEE International Conference on Robotics and Automation (ICRA)*, Philadelphia, USA, pp. 2982-2988, 2022, DOI: 10.1109/ICRA46639.2022.9812429.

[3] J. Jiang, L. Yao, Z. Huang, G. Yu, L. Wang, and Z. Bi, “The state of the art of search strategies in robotic assembly,” *Journal of Industrial Information Integration*, vol 26, Mar., 2022, DOI: 10.1016/j.jii.2021.100259.

[4] G. Schoettler, A. Nair, J. Luo, S. Bahl, J. A. Ojea, E. Solowjow, and S. Levine, “Deep Reinforcement Learning for Industrial Insertion Tasks with Visual Inputs and Natural Rewards,” *IEEE International Workshop on Intelligent Robots and Systems (IROS)*, Las Vegas, USA, pp. 5548-5555, 2020, DOI: 10.1109/IROS45743.2020.9341714.

[5] X. Zhao, H. Zhao, P. Chen, and H. Ding, “Model accelerated reinforcement learning for high precision robotic assembly,” *International Journal of Intelligent Robotics and Applications*, vol. 4, pp. 202-216, Jun., 2020, DOI: 10.1007/s41315-020-00138-z.

[6] Y.-G. Kim, M. Na, and J.-B. Song, "Reinforcement Learning-based Sim-to-Real Impedance Parameter Tuning for Robotic Assembly," *International Conference on Control, Automation and Systems (ICCAS)*, Jeju, Republic of Korea, pp. 833-836, 2021, DOI: 10.23919/ICCAS52745.2021.9649923.

[7] J. Luo, O. Sushkov, R. Pevceviciute, W. Lian, C. Su, M. Vecerik, N. Ye, S. Schaal, and J. Scholz, "Robust Multi-Modal Policies for Industrial Assembly via Reinforcement Learning and Demonstrations: A Large-Scale Study," *Robotics: Science and Systems*, 2021, DOI: 10.15607/RSS.2021.XVII.088.

[8] N. Hogan, "Impedance control: An approach to manipulation: Part II–Implementation," *ASME Journal of Dynamic Systems Measurement and Control*, vol. 107, no. 1, pp. 8-16, Mar., 1985, DOI: 10.1115/1.3140713.

[9] T. Haarnoja, A. Zhou, P. Abbeal, and S. Levine, "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," *International Conference on Machine Learning (ICML)*, Stockholm, Sweden, 2018, DOI: 10.48550/arXiv.1801.01290.

[10] D. P. Kingma and M. Weling, "Auto-Encoding Variational Bayes," *International Conference on Learning Representations (ICLR)*, Banff, Canada, 2014, DOI: 10.48550/arXiv.1312.6114.

[11] K. Cho, B. V. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, DOI: 10.48550/arXiv.1406.1078.

윤 준 완

2021 울산과학기술원 기계항공공학과(학사)
2021~현재 고려대학교 기계공학과(석사)

관심분야: 로봇 제어, 로봇 기반 조립, 강화학습

나 민 우

2017 고려대학교 기계공학과(학사)
2019 고려대학교 스마트융합학과(석사)
2019~현재 고려대학교 기계공학과(박사)

관심분야: 로봇 파지, 로봇 기반 조립, 3차원 측정

송 재 복

1983 서울대학교 기계공학과(공학사)
1985 서울대학교 기계공학과(공학석사)
1992 MIT 기계공학과(공학박사)
1993~현재 고려대학교 기계공학부 교수

관심분야: 로봇의 설계 및 제어, AI 기반 로봇 매니퓰레이션