chapter 2

강화학습 연구 및 융합 기술 동향

김민석 ‖ 상명대학교 교수

I. 서론

강화학습(Reinforcement Learning)은 MDP(Markov Decision Process) 기반의 최적화 개념과 동물심리학 개념(trial-and-error)을 결합한 인공지능 기반 기계학습 알고리즘 중 하나이며, 시스템 최적화 문제를 풀기 위해 많은 연구 및 개발이 이루어지고 있다. 또한, 강화학습은 모든 시스템 환경 정보를 담당하고 관여하는 시뮬레이션 혹은 시스템 환경(Environment)을 중심으로 에이전트(Agent)가 환경에서 파생되는 데이터를 이용하여 보상함수(Reward Function)를 구성하고 이를 반복적으로 개선하여 최적의 목표를 달성하는 시스템 제어 방법이다. 이를 위해서 에이전트는 환경으로부터 파생되는 복수의 환경 상태(State) 변화, 에이전트의 행동(Action) 제어, 시스템 보상함수 설계, 정책(Policy) 개선 및 최적화(Optimization) 모델 도출이라는 유기적인 프로세스를 진행하여야 하며, 이에 따른 환경 상태 정의, 행동 결정, 보상함수 및 정책 설계 등의 학습 지표들이 잘 맞물려서 작동해야 좋은 학습 효과를 얻을 수 있다[1].

강화학습은 다른 기계학습 알고리즘과는 달리 에이전트의 행동에 따른 보상 학습을 기반으로 시스템 제어를 달성하는 최적 제어 솔루션이기 때문에, 보상함수를 설계하기 위한

* 본 내용은 김민석 교수(☎ 041-550-5113, minsuk.kim@smu.ac.kr)에게 문의하시기 바랍니다.
** 본 내용은 필자의 주관적인 의견이며 IITP의 공식적인 입장이 아님을 밝힙니다.

사용자(엔지니어)의 노력이 매우 중요하며, 시스템/시뮬레이션 환경에서 파생되는 데이터를 기반으로 학습 프로세스를 진행해야 하므로 시스템/시뮬레이션 환경이 반드시 존재해야 하는 제약 사항이 있어 다양한 분야에 적용하여 폭넓게 발전하기에는 여전히 한계가 있다. 하지만, 2013년 심층신경망(Deep Neural Network: DNN)을 결합한 심층 강화학습(Deep Reinforcement Learning) 방법이 도래하면서 학습을 위한 다양한 라이브러리(Library)들이 제공되어 수많은 학습 데이터를 처리하기 위한 리소스 문제점들이 해결되고 있으며, 실험 및 검증을 위한 많은 가상 시뮬레이션 환경들이 제공되어 강화학습 기반 최적 제어 문제를 풀기 위해 다양한 솔루션들과 함께 빠르게 발전해 나아가고 있다.

또한, 강화학습은 2015년 2월 네이처에 게재된 DQN(Deep Q_Network) 알고리즘으로 인해 학습 패러다임이 바뀌는 현상들이 본격적으로 나타나기 시작했다. 기존 강화학습은 학습 환경의 상태와 행동을 통해 출력되는 기댓값을 예측하는 방법으로 의사결정을 위한 Q 함수(Q-function)와 정책을 유기적으로 학습 개선하여 최적 제어를 달성하는 방법이다. 하지만 기존의 Q 함수를 학습하기 위해서는 많은 시간을 학습에 투자해야 하

[표 1] 최신 강화학습 알고리즘 종류 및 특징

연도	알고리즘	특 징
2015	Deep Q-Network(DQN)	- 신경망을 적용하여 Action 및 State에 대한 Q-function을 근사화 - Experience Replay Memory를 사용하여 Data Resource 절약
2015	Trust Region Policy Optimization(TRPO)	- Actor Critic(2000) 알고리즘 기반으로 반복되는 학습 정책을 정규화 - 규제(Penalty) 사용으로 불필요한 학습을 억제하여 빠르게 수렴
2016	Deep Deterministic Policy Gradient(DDPG)	- DQN의 Replay Memory를 사용하여 연속적인 Batch 업데이트가 가능 - 빠른 수렴을 위해 정책제어를 사용함
2016	Asynchronous Advantage Actor Critic(A3C)	- Actor Critic(2000) 알고리즘을 기반으로 A2C, A3C로 발전 - 멀티 A2C 환경을 동시에 학습하여 일관된 학습을 유도 - Discrete와 Continuous space에서 동시에 사용 가능
2017	Proximal Policy Optimization (PPO)	- TRPO와 유사한 알고리즘으로 대리 손실함수를 생성하여 더 빠른 수렴을 지원 - TRPO보다 구현이 간단하고 좋은 Sample Complexity를 사용할 수 있음
2018	Twin Delayed Deep Deterministic Policy Gradient (TD3)	- DDPG 기반 다중 네트워크를 동시에 학습하여 정책의 과대평가(Over-estimation)를 감소 - Target Action에 노이즈를 추가하여 Q-function의 오류를 방지
2018	Soft Actor Critic(SAC)	- DDPG와 유사한 방식으로 Q-function의 근사값으로 탐색을 제어 - Off-policy 알고리즘을 사용하여 Sample Inefficiency를 해결할 수 있음 - 로봇 시뮬레이션 모델링을 위해 많이 사용됨

〈자료〉 상명대학교 자체 작성

며, 환경에서 제공되는 과도한 환경 상태 정보에 따라 학습이 잘 이루어지지 않거나, 잘못된 목표로 학습이 진행되는 경우가 발생하였다. 이러한 문제점을 극복하기 위해 Q 함수를 심층신경망으로 구성하여 학습을 진행하는 연구 방법들이 연구되었고, 이 방법들을 통해 Q 함수의 학습효율 및 예측 정확도 성능을 높여 보다 효과적인 행동 제어 기반의 최적화 기계학습 알고리즘을 제공할 수 있게 되었다. 구글의 딥마인드는 DQN을 처음으로 소개하였고, 우리가 잘 알고 있는 Atari 게임부터 AlphaGo(2016)를 거쳐 현재 AlphaStar (2019)와 같은 고성능 게이밍 환경을 이용하여 강화학습의 성과를 이루어 내고 있다 [1],[2].

II. 비지도 강화학습

인공지능을 위한 기계학습(Machine Learning)은 일반적으로 지도학습, 비지도학습 및 강화학습으로 구분할 수 있으며, 최근 보다 효과적인 방법들을 찾기 위해 기존 방법들을 융합하여 연구 개발을 진행하는 현상들이 나타나고 있다. 예를 들어, 높은 학습 효과를 위해 비지도 학습 기반으로 데이터를 클러스터링한 후 지도학습 기반 분류법을 사용하는 방식으로 최소한의 리소스를 가지고 최대한의 학습 효과를 높이는 준지도(Semi-supervised) 기계학습 방법이 바로 대표적인 융합 기계학습 방법이다. 강화학습 또한 기존 Q-learning 기반 학습 방법들을 뛰어넘어 다양한 환경 변수를 이용한 비지도 강화학습을 고안하여 학습 오차를 줄이고 다양한 환경 요소들을 학습하여 학습 성능을 높이는 연구들이 개발되고 있다[3].

1. Intrinsic Reward Function

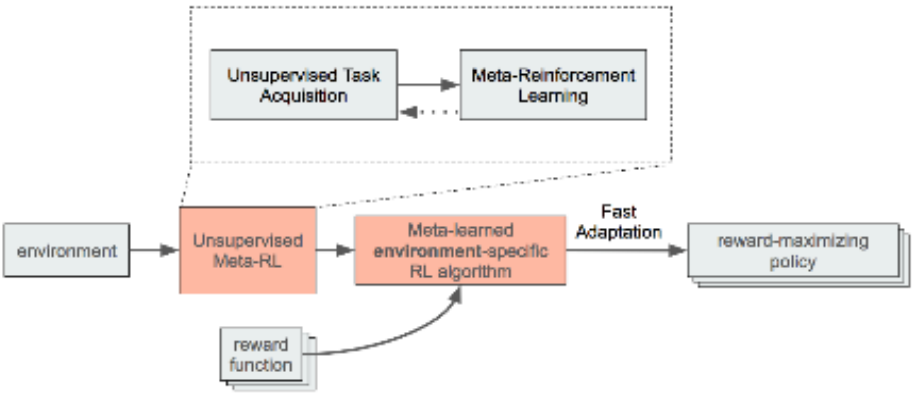
기존 강화학습 접근 방식은 시뮬레이션/시스템 환경으로부터 제공되는 상태 변화에 따른 외적 보상함수(Extrinsic Reward Function)를 이용하여 환경 변화에 대한 에이전트의 행동을 점차 개선하는 방법이다. 하지만, 이미 사용자로부터 정해진 외적 보상함수의 틀에 따른 환경 상태 변화 값을 학습하는 것은 에이전트가 환경의 복잡한 현상에 따른 행동을 효과적으로 제어하기에 무리가 따르는 경우가 발생할 수 있다. 따라서 이런 현상을

극복하는 방법으로 지정된 외적 보상함수 이외에 환경에 따른 내적 보상함수(Intrinsic Reward Function)를 사용하여 학습에 적용하는 강화학습 방법이 제안되고 있다[2],[3].

내적 보상함수는 시스템 혹은 시뮬레이션 환경에서 나타나는 외적 보상함수에서 설계된 학습 요구사항이 아닌 에이전트가 활동하면서 환경에 새롭게 내재되어 있는 현상들을 학습하여 개선하는 방법으로서 기존 강화학습에 비해 광범위한 환경에서 학습 효과를 올릴 수 있는 장점이 있다. 이러한 방법은 사람의 내적 동기 부여를 이용한 학습 방법으로부터 착안해서 생성된 방법이며[2], 에이전트의 행동 개선 및 동기 유발을 위해 시스템 환경 변화에서 나타나는 유기적 상호작용을 강조하고 있다[3]-[5]. 일반적으로 강화학습에서 에이전트는 한 에피소드가 끝난 이후에 초기 상태로 돌아가는 경우가 종종 발생한다. 이로 인해 에이전트는 학습에 대한 진로 개선이 전혀 되지 않은 초기 상태로 다시 학습을 재시작하여 성능을 감소시키는 경우를 방지하고, 매 스텝(행동)마다 미세한 보상(Fine Tuning)을 제공하여 학습 초기화 문제를 해결하는 방식으로 학습 성능을 높이는 효과를 얻을 수 있다. 또한, 매번 보상함수를 설계하거나 업데이트해야 하는 필요성을 감소시킬 수 있으므로 이와 같은 방법을 최근 비지도 강화학습의 개념으로 분류하고 있다. 물론, 학습 목표와 내적 보상 값의 동일성이나 보상의 비정상성(Non-stationary) 등에 대한 단점[2]들이 여전히 존재하지만, 예측 가능한 환경 변화 요소에 따른 학습 최적화 목표를 성취하기 위해서는 비지도 강화학습 방법이 좋은 효과를 얻을 수 있을 것이다.

2. Transferable Meta-skills(Learning)

앞서 언급한 것과 같이 강화학습은 주어진 환경에서 새로운 이슈를 빠르게 개선하여 문제를 해결하는 기계학습 방법이다. 일반적으로 인공지능 기반 지능학습인 심층신경망 학습 알고리즘은 성능 개선을 위해 최적화 파라미터 조정(Hyper Parameters Tuning) 방식을 채택하여 학습 개선을 시도하지만, 심층 강화학습은 내부에 있는 정책을 개선하는 SGD(Stochastic Gradient Descent) 또는 오류반송(Error back-propagation) 알고리즘을 통해 심층신경망을 학습하여 성능을 개선하는 방식이다[1]. 이와 같은 방법과 더불어 학습 성능을 향상시키는 방법으로는 다중 에이전트(Multi-agent)를 이용하여 학습을 분산하는 다중 에이전트 방식이 있다[6]. 또한, 데이터를 분산하여 저장하는 방법으로 딥러닝 분산 구조를 적용하여 환경 상태 정보로부터 얻은 경험을 신속하게 수집하고 이를



〈자료〉 Abhishek G, Benjamin E., Chelsea F, Sergey L, "Unsupervised meta-reinforcement learning: Given an environment, unsupervised meta-reinforcement learning produces an environment-specific learning algorithm that quickly acquire new policies that maximizes any task reward function", arXiv:1806.04640v3 Thu, 30 Apr. 2020.

[그림 1] 비지도 메타 강화학습 구조

기반으로 성능을 향상하는 분산 강화학습 방식 등이 강화학습의 활용 영역을 점차 넓히고 있다.

분산 강화학습 방법과 더불어 메타 기술(Meta Skill)을 이용한 강화학습은 시스템 환경에서 발생하는 환경 변화 요소들에 따라 나타나는 보상함수나 정책들을 에이전트에 의해 메타 테스트(Meta-test) 방식으로 분리하여 학습하는 방법으로 학습 성능을 효과적으로 최적화할 수 있는 비지도 강화학습 방법이다. [그림 1]의 구조와 같이 비지도 메타 강화학습 방법은 기존 강화학습의 보상함수 설계 시간을 단축하고 정책을 개선할 수 있으며 메타 기술을 이용한 별도의 개선 학습 과정을 추가하여 일반적인 학습 단계에서 경험하지 못하는 에이전트의 새로운 행동 제어, 환경 상태 정보, 정책 등을 일반화하여 학습에 효과적으로 반영하여 성능을 최대한 끌어 올릴 수 있는 장점들이 있다[7],[8]. 또한, 성능 개선을 위한 미세 조정(Fine Tuning) 없이 일반화된 강화학습 요소들을 학습에 적용하여 학습 효과를 높일 수 있으므로, 이를 비지도 강화학습으로 구분하여 정의할 수 있다.

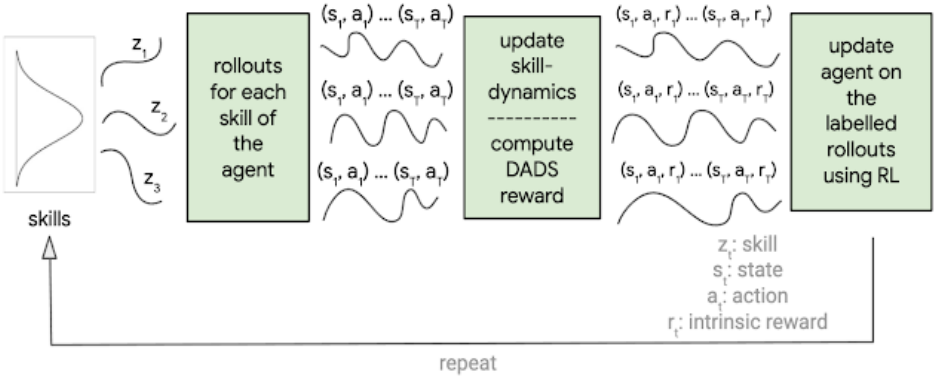
메타 기술 기반의 비지도 강화학습은 보상함수 및 정책과 연관된 상호작용을 분리하여 학습을 진행하기 위해 시간의 연속 관계를 이용한 학습 선순환 반복 기법이라고도 말할 수 있다. 메타 테스트/검증을 위한 시간은 보상함수와의 직접적인 연관성이 없지만, 학습 과정을 통해 환경 상태 정보와 동적 상호 작용이 존재한다고 가정할 수 있다[7],[8]. 따라서 이러한 방법은 환경 정보에 의존하지 않고 학습하는 비지도 학습 방법을 부분적으로

채택하고 있으며 해당 시간의 환경 요소를 사용하여 보상 기능으로부터 학습을 최적화할 수 있다는 점에서 매우 효과적인 학습 방법이라고 할 수 있다. 비록 에이전트가 경험하지 못한 부분을 학습하고 해결하는 과정에서는 어려움이 존재하지만, 학습에 시간적 의미를 적용하여 문제점을 해결하는 접근 방식으로 확장하면 사용자(엔지니어)가 학습에 개입하여 조정해야 하는 필수 조건을 배제하여 자율적으로 학습을 진행하는 비지도 강화학습의 완성형이 될 수 있다고 생각한다.

III. 비지도 강화학습 응용 분야

1. 로봇 시뮬레이션

구글 로봇팀(Google Brain team and the Robotics at Google team)에서 연구 개발하고 있는 Dynamics-Aware Unsupervised Discovery of Skills(DADS)를 통해 비지도 강화학습은 로봇 시뮬레이션 시나리오뿐만 아니라 실제 환경에서도 휴머노이드(Humanoid) 로봇과 같은 물체에 대한 행동 인식과 동작을 설계할 수 있다[3]. 이러한 방법들은 로봇의 복잡한 동작을 합성하여 민첩한 운동을 빠르게 학습하여 행동을 제어할 수 있는 비지도 강화학습 방법을 선택하고 있다. 이 방법을 통해 환경에 대한 예측 가능성



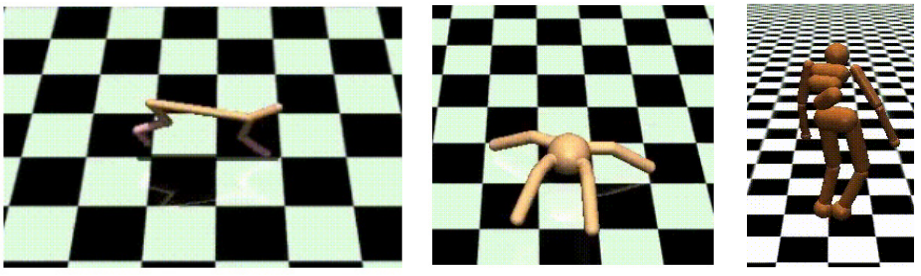
〈자료〉 Archit S, "DADS: Unsupervised Reinforcement Learning for Skill Discovery," posted by AI Resident, Google Research at the Google Brain team and the Robotics at Google team, Friday, May 29, 2020.

[그림 2] 비지도 강화학습 학습 프로세스

(Predictability)과 행동에 대한 다양성(Diverse)을 위한 해결 방법을 제시하고 있으며, 내적 보상함수 학습을 위한 기술적 방법을 함께 지원하고 있다. 또한, 내적 보상함수는 에이전트가 최대한 많은 잠재적 유용 동작을 학습할 수 있게 도와주며[3]-[5], 예측 가능한 다양한 변화를 학습 활동을 통해 환경에서 얻을 수 있다. 이러한 기술에 의해 정의된 본질적인 보상은 기존 학습 알고리즘과 마찬가지로 성능 최적화를 달성하는데 그 목표를 두고 있다.

[그림 2]와 같이 구글 로봇팀은 DADS를 통해 Skill-dynamics라고 불리는 신경망을 추가하여 학습을 진행하고 있다. 여기서 추가된 신경망은 환경이 제공하는 상태의 동적 변화를 감지하여 학습하는 보조의 심층신경망이며, 이 신경망을 통해 환경의 상태 변화(내적 보상)를 주기적으로 인지하여 학습하고 다음 행동을 예측할 수 있으므로 좋은 학습 결과를 기대할 수 있다[3]-[5].

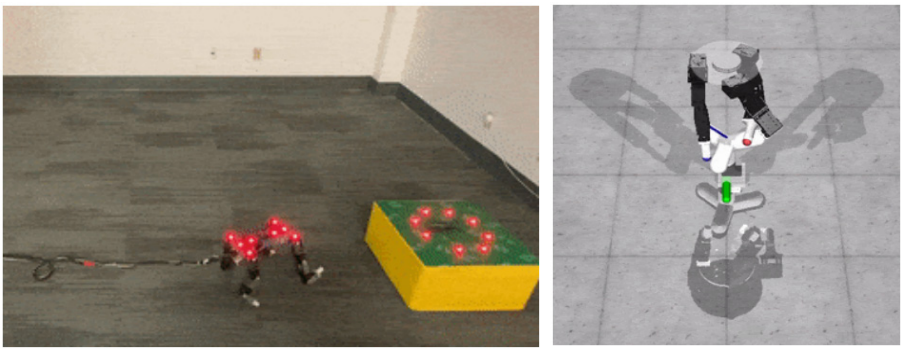
[그림 3, 4]는 최근 내적 보상함수를 이용한 비지도 강화학습 기반의 휴머노이드 로봇 시뮬레이션 예제들이다. 이와 같은 예제들은 로봇 시뮬레이션의 효과적인 학습을 실현하기 위해 예측 가능한 다양한 기술 요소들을 학습에 접목하고 있다. 내적 보상함수는 환경에 따라 제공되는 다양한 변화를 상태 정보로 전환할 수 있고, 이에 따라 제공되는 상태 정보를 이용하여 예측 가능한 환경적 변화로 학습하여 성능을 최적화할 수 있다.



〈자료〉 Simulation Example of 'Half Cheetah', 'Ant', Humanoid', Archit S, Shixiang G, Sergey L, Vikash K, Karol H "Dynamics-Aware Unsupervised Discovery of Skills," in Proceedings of International Conference on Learning Representations(ICLR), 2020.

[그림 3] 비지도 강화학습 기반 휴머노이드 로봇 시뮬레이션 예제 - 1

환경 변화에 따라 강화학습 기반 보상함수를 설계하는 이전 방식은 오히려 잠재적으로 예측 다양성을 억제할 수 있으므로 에이전트는 환경에 존재하는 잠재적 환경 변화를 포착하고 이에 따른 행동을 학습하여 내적 보상함수 요소를 개선하는 방식으로 기술적 확장성



〈자료〉 Simulation Example of 'Goal Navigation', 'Emergent Skills', Archit S, Michael A, Sergey L, Vikash K, Karol H, Shixiang G, "Emergent Real-World Robotic Skills via Unsupervised Reinforcement Learning," in Proceedings of Robotics: Science and Systems(RSS), 2020.

[그림 4] 비지도 강화학습 기반 휴머노이드 로봇 시뮬레이션 예제 - 2

을 높이고 있다.

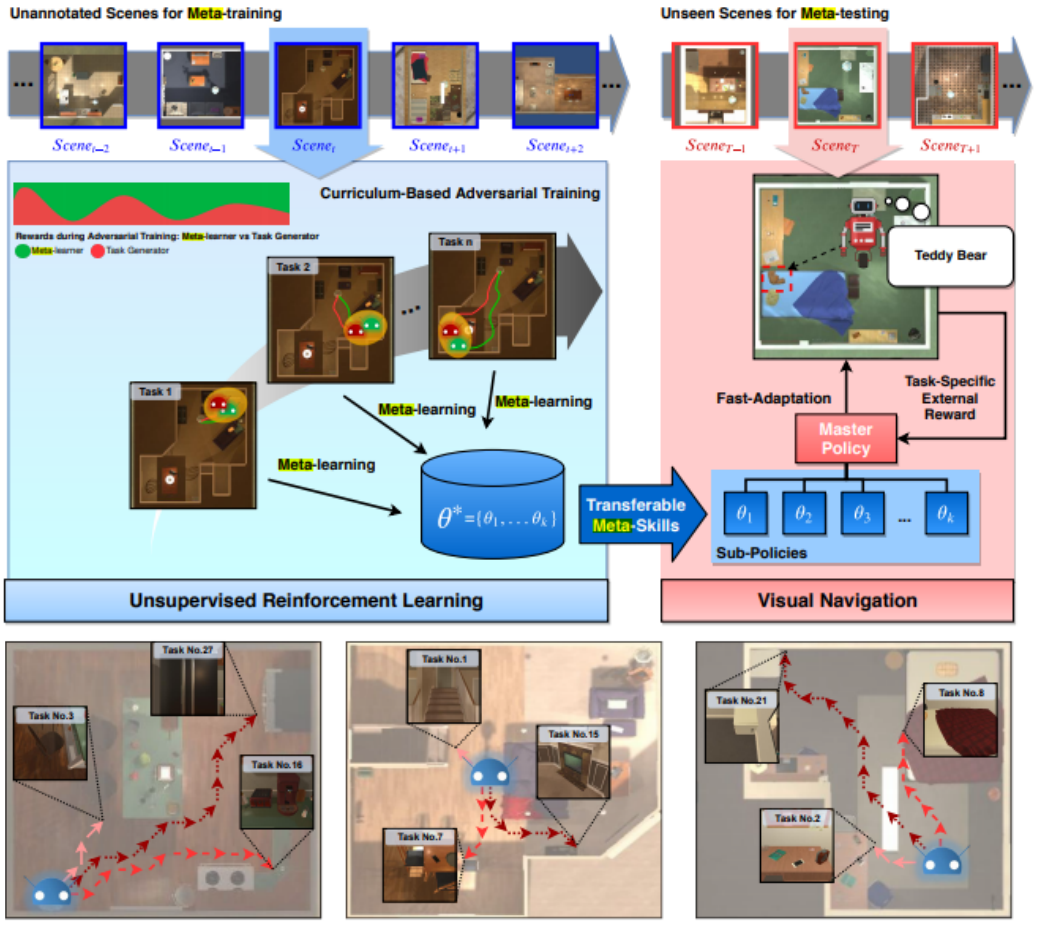
비지도 강화학습 알고리즘과 함께 다중 에이전트(Multi-agent) 방식을 채택하여 각각의 에이전트가 활동하면서 얻는 환경 정보와 학습 요소들을 공유하며 상호작용하는 방법으로 로봇 시뮬레이션 학습 속도를 가속화할 수 있다. 이 방법은 휴머노이드 로봇과 같은 고차원의 연속 제어 환경에서 매우 유용하게 사용될 수 있을 뿐만 아니라, 외적 보상함수의 환경적 제한으로 인한 영향을 직접적으로 받지 않기 때문에 학습을 통한 예측 제어의 다양성을 구현할 수 있는 장점들이 있다.

2. 네비게이션 모델

현재 사용되고 있는 지도 경로 탐색 네비게이션 모델에서도 비지도 강화학습을 적용할 수 있다. 심층 강화학습 기반의 네비게이션 모델을 구현하기 위해서는 막대한 데이터의 학습이 필요하므로, 네비게이션 모델 기반 지도 경로 탐색에서 가장 중요한 요소인 적은 양의 데이터를 이용하여 빠른 경로를 탐색(Path Planning)하는 방법으로는 비효율적이다. 또한, 시뮬레이션 혹은 실제 환경에 대한 개체 정보를 환경으로부터 제공받기 위해서는 수많은 데이터 어노테이션(Annotation)이 필요하며, 다른 환경에서 학습된 모델을 이전(Transferred Model)하여 사용하는 방법도 어려워서 심층 강화학습을 경로 탐색 모델에 적용하여 사용하는 것은 효과적이지 않을 수 있다.

이를 위한 솔루션으로 메타 기술(Transferable Meta-skills) 혹은 메타 학습을 통한 강화학습 방법들이 활발히 연구되고 있다. 이 방법은 다양한 환경에서 습득한 메타 정보를 필요한 환경에 이전 적용하여 사용할 수 있을 뿐만 아니라, 새로운 환경에서도 에이전트의 초기 학습 성능을 향상시킬 수 있으므로 빠른 경로 탐색이 필요한 탐색 네비게이션 모델로 적합하다.

[그림 5]는 유동적 메타 기술을 이용한 비지도 강화학습 시스템 구성도이다. 여기서 사용된 비지도 강화학습 방법은 메타 기술을 이용하여 경로 탐색을 유도하는 마스터 정책



〈자료〉 Simulation Example of 'Goal Navigation', 'Emergent Skills', Juncheng L, Xin W, Siliang T, Haizhou S, Fei W, Yueting Z, William Y,, "Unsupervised Reinforcement Learning of Transferable Meta-Skills for Embodied Navigation," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition(CVPR) 2020.

[그림 5] 유동적 메타 기술을 이용한 비지도 강화학습 시스템 구성도

(Master Policy)과 하위 정책(Sub-Policy)을 우선 분리한다. 분리된 하위 정책은 환경 상태 집합의 메타 학습을 미리 진행한 후 학습된 정보를 마스터 정책으로 이전하여 마스터 정책의 학습 성능을 신속하게 향상하는 방법을 시도하고 있다[7],[8]. 이 방법은 에이전트가 환경으로부터 개체 정보를 제공받지 않고, 오직 경로를 위한 하위 정책 프로세스와 일부 보상만으로 학습을 진행하기 때문에 이는 비지도 강화학습 방법 기반의 학습 방법이라고 할 수 있다. 또한, 기존 강화학습의 임의 탐색(Random Exploration)과 같은 방법을 통해 경로 학습을 무작위로 우선 진행하여 학습을 유도하는 방법이 아닌, 메타 학습을 이용한 하위 정책 학습 방법을 통한 학습 개선 방법으로 임의 탐색 기법보다 매우 빠른 학습 효과를 달성할 수 있다.

메타 기술을 이용한 비지도 강화학습은 리소스가 부족한 환경에서도 상태 정보를 제공받을 수 있으므로 보상함수의 설계가 잘 되어 있지 않은 시스템에서도 이전 가능한 메타 기술을 통해 단점을 보완할 수 있다. 또한, 마스터 정책을 통해 하위 정책의 시간적 학습 요소를 결정하여 학습된 메타 기술을 빠르게 습득할 수 있으며, 마스터/하위 정책들 간의 선순환 학습 과정을 유도하여 성능을 최적화할 수 있다.

Ⅳ. 결론

지금까지 살펴본 비지도 강화학습은 학습의 여러 가지 측면에서 예측 가능성, 보상함수의 확장성, 시간적 효율성, 학습 안정성 등의 새로운 도전과제들이 여전히 남아 있다[9],[10]. 비록 기술 확장 및 성능 최적화 측면에서 강화학습은 여전히 보상함수 설계, 환경 상태 정보의 제약, 시뮬레이션 환경과 실제 환경의 구조적 차이, 확장 이전성 문제, 적용 분야의 다양성 부재 등의 이슈들이 아직 해결해야 하는 문제로 남아 있지만, 다양한 측면에서 지속적인 연구 및 개발이 진행되고 있으므로 앞으로 더욱 기대가 높은 연구 분야이다. 또한, 여러 기술 간의 융합 연계성을 적용하여 학습 기반의 제어기술을 세분화한다면 단계별로 다양한 분야와 융합하여 사용 가능할 수 있을 것이며, 향후 인공지능 기술의 발전 가속화와 양적 인프라 향상에도 좋은 결과를 가져올 것으로 기대한다.

[참고문헌]

[1] 신승재, 조충래, 전홍석, 윤승현, 김태연, “심층 강화학습 라이브러리 기술 동향: A Survey on Deep Reinforcement Learning Libraries”, 한국전자통신연구원, 전자통신동향분석 34권 제6호 2019.

[2] 장수영. 윤현진, 박노삼, 윤재관, 손영성, “심층 강화학습 기술 동향: Research Trends on Deep Reinforcement Learning”, 한국전자통신연구원, 전자통신동향분석 34권 제4호 2019.

[3] Archit S, “DADS: Unsupervised Reinforcement Learning for Skill Discovery,” AI Resident, Google Research at the Google Brain team and the Robotics at Google team, 2020. 5. 29.

[4] Archit S, Shixiang G, Sergey L, Vikash K, Karol H “Dynamics-Aware Unsupervised Discovery of Skills,” International Conference on Learning Representations(ICLR), 2020. 4. 17.

[5] Archit S, Michael A, Sergey L, Vikash K, Karol H, Shixiang G, “Emergent Real-World Robotic Skills via Unsupervised Off-Policy Reinforcement Learning,” Robotics: Science and Systems(RSS), 2020.

[6] Megherbi, D. B., Kim, Minsuk., Madera, Manual., “A Study of Collaborative Distributed Multi-Goal and Multi-agent based Systems for Large Critical Key Infrastructures and Resources (CKIR) Dynamic Monitoring and Surveillance,” IEEE International Conference on Technologies for Homeland Security, 2013.

[7] Juncheng L, Xin W, Siliang T, etc. “Unsupervised Reinforcement Learning of Transferable Meta-Skills for Embodied Navigation,” IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2020.

[8] Abhishek G, Benjamin E, Chelsea F, Sergey L, “Unsupervised Meta-Learning for Reinforcement Learning,” International Conference on Learning Representations(ICLR), 2019.

[9] Volodymyr M, Adria P, Mehdi M, etc. “Asynchronous Methods for Deep Reinforcement Learning,” International Conference on Machine Learning, Volume 48, 2016, pp.1928-1937.

[10] Colin F, “Reinforcement Learning Reward Functions for Unsupervised Learning,” International Symposium on Neural Networks, Advances in Neural Networks-ISNN 2007, pp.397-402.