https://doi.org/10.7840/kics.2023.48.11.1405

실용적 강화학습 기술 동향: 모방학습부터 오프라인 강화학습까지

이 동 수•, 엄 찬 인*, 최 성 우**, 김 성 관**, 권 민 혜°

Survey on Practical Reinforcement Learning : from Imitation Learning to Offline Reinforcement Learning

Dongsu Lee•, Chanin Eom*, Sungwoo Choi**, Sungkwan Kim**, Minhae Kwon°

요 약

최근 강화학습의 패러다임은 온라인에서 오프라인으로 전환되고 있다. 이러한 변화는 시뮬레이션 기반 게임 과제에 국한된 온라인 강화 학습의 비실용성을 극복하기 위함이다. 본 논문에서는 사전 수집된 고정 데이터 세트를 기반으로 정책을 학습하는 실용적인 강화학습 기술을 소개하고자 한다. 이러한 시도는 모방학습부터 시작되었고, 최근 각광받고 있는 오프라인 강화학습(offline reinforcement learning) 방식으로 발전하였다. 오프라인 강화학습에서는 본질적인 문제인 분포 이동(distributional shift)을 해결하기 위해 제안된 오프라인 강화학습 알고리즘들에 대해 소개하고자 한다. 마지막으로 해당 분야에서 현재 해결되지 않은 문제들과 한계에 대해 논의한다.

키워드 : 심층 강화학습, 오프라인 강화학습, 모방학습
Key Words : Deep Reinforcement Learning, Offline Reinforcement Learning, Imitation Learning

ABSTRACT

The reinforcement learning paradigm has shifted from online to offline recently. Such a change is to overcome the impracticality of online reinforcement learning, which is limited to simulation-based game tasks (e.g., Go, Chess, Atari, and so on). This paper reviews an offline reinforcement learning approach that builds a policy by leveraging previously collected fixed datasets. To elaborate, we deal with the state-of-the-art offline reinforcement learning algorithms, which have been proposed to mitigate the distributional shift. Lastly, we discuss the open problems and limitations of current offline reinforcement learning.

Ⅰ. 서 론

강화학습(reinforcement learning)은 인공지능 개체(agent)의 자율적인 의사결정을 위한 강력한 패러다임으로, 기계학습의 방법론 중 하나이다. 행동심리학에서 영감을 받은 이 방법은 환경(environment)과의 상호작

※ 이 논문은 2022년 현대자동차그룹 미래기술연구과제의 지원을 받아 수행된 연구임.
◆ First Author : Soongsil University Department of Intelligent Semiconductors, movementwater@soongsil.ac.kr, 학생회원, 이동수 학생은 현대 정몽구재단의 장학금 지원을 받았음.
° Corresponding Author : Soongsil University Department of Intelligent Semiconductors and School of Electronic Engineering, minhae@ssu.ac.kr, 종신회원
* Soongsil University Department of Intelligent Semiconductors, eci0623@soongsil.ac.kr, 학생회원
** Hyundai Motor Company, {sungwoo.choi, sungkwan.kim}@hyundai.com
논문번호 : 202307-004-C-RN, Received July 5, 2023; Revised August 11, 2023; Accepted August 29, 2023

용 속에서 개체가 시행착오를 통해 최적의 의사결정 방법을 학습한다. 구체적으로, 개체는 환경에서 얻은 상태 정보(state)를 고려하여 행동(action)을 취하고 이에 대한 보상 신호(reward)를 최대화 하기 위한 행동 정책(policy)을 학습한다. 하지만 고전적인 강화학습 방법은 상태 및 행동 공간이 작고 이산적인 단순한 문제 해결에 사용되었다. 이러한 한계를 극복하며 강화학습이 주목을 받기 시작한 것은 딥러닝(deep learning)의 성공 이후, 심층 신경망(neural network)을 적용한 심층 강화학습(deep reinforcement learning) 방법이 대두된 이후이다.[1] 심층 강화학습 방법은 고용량의 함수 근사자(approximator)를 이용하여 연속적이며 비 선형적인 상태 및 행동 공간으로 구성된 복잡한 문제에서도 효과적인 성능을 보였다. 이와 같은 방법은 제어와 관련된 분야 뿐 아니라 각광받고 있는 자연어 처리 및 비전 분야에서도 유용하게 활용되고 있다.[2-5]

하지만, 온라인 강화학습(online reinforcement learning)의 경우 실제 세계 문제 해결에 적용하기에는 실용적 한계가 존재한다. 먼저, 개체의 정책이 불안정한 학습 초기 상태에서도 환경과의 상호작용을 통해 직접 데이터를 수집해야 하는데, 개체가 물리적 디바이스(자율주행차, 지능형 드론, 로봇)인 경우 사회적 및 금전적 위험을 초래할 수 있다. 또한, 개체는 능동적으로 환경과의 지속적인 상호작용을 통해 학습 데이터 세트를 수집해야 한다. 특히 on-policy 알고리즘을 기반으로 하는 경우 정책의 업데이트를 수행하는 시점마다 새로운 데이터 세트를 필요로 한다. 즉, 이전에 이용한 데이터는 폐기하기 때문에 샘플 효율성이 낮다. Off-policy 알고리즘의 경우 여전히 환경과의 능동적인 상호작용을 할 수 있지만, 이전에 수집된 데이터를 폐기하지 않고 replay buffer에 저장을 하며 샘플링을 통해 업데이트를 진행하기 때문에 샘플 효율성을 높일 수 있다. 온라인 강화학습의 실용적 문제 적용 한계를 해결하기 위한 대안으로 off-policy 알고리즘을 기반으로 하는 오프라인 강화학습(offline reinforcement learning; 또는 batch reinforcement learning) 방법이 논의되고 있다.[6]

오프라인 강화학습은 실시간으로 데이터 세트를 수집하지 않고 사전에 수집된 고정된 데이터 세트를 이용하여 학습을 진행하는 강화학습 방법이다. 이와 같은 방법은 온라인 강화학습이 효율적이지 않은 작업(로봇, 우주탐사, 자율주행, 의료 등)에 적용하는 경우 효율적으로 동작할 수 있다.[7-9] 하지만, 고정된 데이터 세트을 이용하는 방법은 앞서 말한 문제들을 해결하는 실용적인 해결책이 될 수 있는 반면에 온라인 상호작용을 추가로 고려할 수 있도록 기존에 고안되었던 온라인 강화학습의 알고리즘을 추가 적용하는 경우 종종 실패로 이어진다. 이러한 실패의 원인은 고정된 데이터 세트을 통해 경험한 분포 외부의 행동에 대한 낙관적 가치 평가 혹은 잘못된 일반화로 보고되고 있다.[10] 오프라인 방식에 온라인 강화학습 알고리즘의 단순한 적용이 실패로 이어진 현재, 문제를 완화할 수 있는 오프라인 강화학습 알고리즘의 개발 연구가 수행되고 있다.

대부분의 오프라인 강화학습 알고리즘의 경우 데이터 세트에서 경험한 궤적 분포와 실제 환경에서 경험할 분포의 차이를 최소화 하는 분포 이동(distributional shift) 문제를 최소화할 수 있는 학습 방법에 대한 제안을 포함하고 있다. 학습될 정책을 고정된 데이터 세트를 수집한 정책과 유사하게 제한하는 방법이 대표적이다. 구체적으로, 가치 함수에 대한 하한선(lower bound)을 설정하여 보수적으로 정책을 학습할 수 있도록 제약(regularizer 및 constraint) 및 페널티를 고려한다.[11,12] 제약으로는 명시적 밀도(explicit density)[12,13], 암시적 발산(implicit convergence)[14,15], 정책 목적 함수를 위한 지도학습(imitative learning)[16], Q-함수의 직접적 정규화[12] 등이 포함될 수 있다. 다수의 가치 평가 모델을 병렬적으로 사용하는 앙상블(ensemble) 방법 역시 보수적인 학습[17,18] 혹은 불확실성(uncertainty)의 추정을[16] 통해 보다 안정적인 정책 연구 학습을 위해 사용되고 있다. 최근 연구되고 있는 새로운 방법론은 이러한 제약을 이용하지 않고 중요도 샘플링(importance sampling)[19,20] 혹은 개체의 궤적 정보를 최적화하는 방법[21]을 고려하고 있다. 마지막으로 오프라인 강화학습을 통해서는 초기 정책을 학습하고 온라인 미세조정(fine-tuning)을 통해 정책을 강화 시키는 방법 역시 주목 받고 있다.[22,23]

본 논문에서는 강화학습의 실용적 문제 해결을 위한 오프라인 강화학습 패러다임의 전반적인 내용을 제공한다. 이를 위해 강화학습 문제를 공식화 하는 방법과 함께 최근 연구된 오프라인 강화학습 알고리즘들의 방법론에 대해 설명한다.

Ⅱ. 배경 지식

본 섹션에서는 강화학습 문제를 정의하는 방법과 함께 강화학습 문제를 해결하는 방법 및 표기법(notation)에 대해 소개한다. 먼저, 강화학습에서 개체와 환경 사이의 의사결정을 다룬 마르코프 의사결정(Markov decision process; MDP) 모델에 대해 확인한 후, 강화학습 문제를 해결하는 방법 및 정의에 대해 살펴본다. 이어서, 오프라인 강화학습 알고리즘의 이해를 위한 배

경 지식에 대해 논의한다.

2.1 마르코프 의사결정

강화학습은 역동적 환경(dynamical environment)에서의 의사결정 모델의 학습 문제를 해결할 수 있는 방법이다. 개체는 상태(state) s_t에서 행동(action) a_t를 결정하고, 이에 대한 대한 보상 신호(reward) r_t를 획득하는 의사결정 모델을 갖는다. 이와 같은 의사결정 모델은 마르코프 의사결정 모델로 정의할 수 있다. 이때, 마르코프 의사결정은 이상적인 환경 상태를 고려하는 강화학습 문제를 표현할 수 있는 수학적 공식화 방법이다.[24]

마르코프 의사결정 모델은 튜플 $M = <S, A, T, \rho_0, R, \gamma>$로 나타낼 수 있다. 여기서 S는 상태 공간(state space), A는 행동 공간(action space), $T(S_{t+1} \mid s_t, a_t)$는 상태전이 확률(state transition probabilities), ρ_0는 초기 상태 분포(initial state distribution), R은 보상함수(reward function), 그리고 $\gamma \in (0,1]$는 감가율(temporal discounted factor)을 나타낸다.

하지만 상태 정보를 모두 관측할 수 있다는 이상적인 가정을 포함하는 마르코프 의사결정 모델은 실용적인 문제에 적용하기에는 한계가 있다. 따라서, 대부분의 실용적 문제 해결을 위해서는 부분적 관측 마르코프 의사결정(partially observable Markov decision process; POMDP)를 이용한다. POMDP는 완전한 상태 정보가 아닌 일부 관측한 불완전한 정보인 o_t를 이용한다. 해당 모델은 $M = <S, A, O, T, \Omega, \rho_0, R, \gamma>$과 같이 정의할 수 있다. O는 관측 공간을 나타내며, $\Omega(o_t \mid s_t)$는 관측 전이 확률(observation transition probabilities)을 나타낸다.

강화학습에서 개체의 궁극적인 목표는 궤적에서 얻어지는 누적 보상을 최대화 하는 정책을 학습하는 것이며, 이는 수학적으로 다음과 같이 표현할 수 있다.

$$J(\pi) = \mathbb{E}_{s_0 \sim \rho_0, a_t \sim \pi(s_t), s_{t+1} \sim T(s_t, a_t)}[\sum_t \gamma^t R(s_t, a_t, s_{t+1})] \quad (1)$$

여기서 π는 개체의 행동 정책을 의미하며, 만약 궤적의 전체 길이가 ∞로 고려되는 경우, $\pi(a_t \mid s_t) T(s_{t+1} \mid s_t, a_t)$에 의해 정의되는 (s_t, a_t)에 대한 Markov chain을 에르고딕(ergodic)하다고 가정하여 문제를 해결할 수 있다.[24]

2.2 강화학습

강화학습에서 궁극적인 목표는 모든 궤적에서 예상되는 누적 보상을 최대화 할 수 있는 최적의 정책을 학습하는 것이다. 강화학습 문제를 해결하는 방법에는 크게 세가지 방법 있다.

가치 기반 강화학습(Value-based RL): 가치 기반의 강화학습 방법은 상태 가치함수 $V(S_t)$ 혹은 상태-행동 가치 함수인 $Q(s_t, a_t)$를 최적화함으로써 최적의 정책을 간접적으로 학습한다. 여기서, $V(S_t) = \mathbb{E}_{a_t \sim \pi(s_t)}[Q(s_t, a_t)]$로 나타낼 수 있으며, 상태 가치 함수는 상태-행동 가치 함수 $Q(s_t, a_t)$의 모든 행동에 대한 기댓값을 의미한다. 따라서, $Q(s_t, a_t) - V(S_t)$는 특정 행동 a_t가 어느 정도의 이점을 가지는지 평가하는 지표로 이득 함수 (advantage function) $A(s_t, a_t)$로 정의할 수 있으며, 이는 상태 가치 함수에 대한 낮은 분산 정도를 갖기 때문에 대안으로 고려될 수 있다. 최적의 정책을 간접적으로 학습한다는 특징은 벨만 최적 연산자 (Bellman optimality operator) B^π에서 기인하며, 유일한 바나흐 고정점(Banach fixed point)인 $Q^{\pi*}(or\ V^{\pi*})$에서 벨만 연산자의 γ-축소(contraction)를 다음과 같이 만족해야만 한다.[24]

$$B^\pi Q(s_t, a_t) = \mathbb{E}_{s_{t+1} \sim T(s_t, a_t)}\big[R(s_t, a_t, s_{t+1}) + \gamma Q\big(s_{t+1}, \pi(s_{t+1})\big)\big]$$

$$B^\pi V(s_t) = \mathbb{E}_{a_t \sim \pi(s_t), s_{t+1} \sim T(s_t, a_t)}[R(s_t, a_t, s_{t+1}) + \gamma V(s_{t+1})]$$

여기서 $R(s_t, a_t, s_{t+1}) + \gamma Q(s_{t+1}, \pi(s_{t+1}))$은 Bellman target y로 정의할 수 있다. 위의 가치 함수를 계산하기 위한 가장 기초적인 방법으로는 우리는 동적 프로그래밍 (dynamic programming) 및 반복적인 추론과 평가를 거치는 방법을 고려할 수 있다. 하지만 이러한 방법은 상태 및 행동 공간이 연속적이며 높은 차원을 갖게 되는 경우 계산 복잡도에 의해 적용하기가 어렵다는 단점이 있다.

이와 같은 문제는 2013년 심층 신경망을 함수 근사자로 고려하여 행동 가치 함수 $Q(s_t, a_t)$를 근사한 Deep-Q network(DQN)을 시작으로 빠르게 해결되었다.[25] DQN은 벨만 오류 (Bellman error) $Q_\theta(s_t, a_t) - B^\pi Q_{\theta'}(s_{t+1}, \pi(s_{t+1}))$를 최적화 하는 것을 신경망의 목적함수로 고려하였다. 여기서 θ와 θ'은 Q-network 및 목표 Q-network(target Q-network)의 파라미터를 의미한다.

정책 경사(policy gradient): 식 (1)을 최적화 하는 가장 직접적인 방법은 정책의 기울기를 직접 추정하는 것이다. 이 경우 가치 기반 방법과 달리 정책 자체가

매개변수화 될 수 있다. 여기서 매개변수에 대한 목적함수의 경사$\nabla_\phi J(\pi_\phi)$는 다음과 같다.

$$\mathbb{E}_{\tau\sim\pi_\phi}\left[\sum_t \nabla_\phi \ln \pi_\phi(a_t|s_t)\left(G_t - b(s_t)\right)\right]$$

여기서 G_T는 t 시점부터 궤적의 마지막 시점 까지의 누적 보상(return)을 의미하며, $b(s_t)$는 baseline으로 샘플링된 궤적에 대한 평균 보상을 의미한다. $b(s_t)$는 V(s_t)로 대체될 수 있다. 정책 경사 방법은 가치함수에 대한 추정을 위해 일반적으로 Monte Carlo 방식으로 해결된다.

Actor-critic: Actor-critic 알고리즘은 가치 함수를 매개변수화 하는 가치 기반의 방식과 정책을 매개변수화 하는 정책 경사 방식을 통합한 방식이다. Actor 네트워크의 경우 정책에 대한 근사를, critic의 경우 가치함수에 대한 근사를 수행한다. 일반적으로 SAC(soft actor critic) 기반의[26] 확률론적 알고리즘과 DPG (deterministic policy gradient) 기반의 결정론적 알고리즘으로 구분할 수 있다.[27,28] 먼저 확률론적 알고리즘의 경우 정책은 행동의 분포를 출력하며, 액터 네트워크 파라미터 ϕ 및 크리틱 네트워크 파라미터 θ 목적함수는 다음과 같다.

$$J(\theta) = \frac{1}{|B|}\sum_{(s_t,a_t,r_t,s_{t+1})=1}\Big(Q_\theta(s_t,a_t) - r_t + \gamma\left(Q_{\theta'}\left(s_{t+1},\pi_\phi(s_{t+1})\right) - \beta\log\pi_\phi(s_{t+1})\right)\Big)$$

$$J(\phi) = \frac{1}{|B|}\sum_{s_t\in B} Q_\theta\left(s_t,\pi_\phi(s_t)\right) - \alpha\log\pi_\phi(s_t)$$

여기서 B는 업데이트에 사용되는 batch의 크기를 의미하며, β는 temperature 파라미터를 의미한다.

결정론적 알고리즘의 경우 정책은 직접적인 행동 값을 출력하며, 각 네트워크의 목적함수는 다음과 같다.

$$J(\theta) = \frac{1}{|B|}\sum_{(s_t,a_t,r_t,s_{t+1})=1}\Big(Q_\theta(s_t,a_t) - r_t + \gamma Q_{\theta'}\left(s_{t+1},\pi_\phi(s_{t+1})\right)\Big)$$

$$J(\phi) = \frac{1}{|B|}\sum_{s_t\in B} Q_\theta(s_t,\pi_\phi(s_t))$$

이어서, 이와 같은 알고리즘의 안정성을 위해 사용되는 대표적인 기술 두 가지를 소개한다.

1) Soft target update: Soft update[28]는 신경망의 매개 변수를 업데이트 하는데 사용되는 기술이다. 타겟 네트워크의 가중치를 점진적으로 조금씩 업데이트 하는 방식으로 다음과 같이 작동한다. 이와 같은 점근 방식은 분산을 줄이고 네트워크가 보다 느리게 학습하게 유도하여 학습 과정의 안정화를 돕는데 사용된다.

$$\theta' = \tau\theta + (1-\tau)\theta'$$
$$\phi' = \tau\phi + (1-\tau)\phi'$$

2) Clipped double Q[29]: 해당 방법은 심층망을 사용하는 경우 하나의 Q 네트워크가 아닌 두개의 Q 네트워크를 이용하는 방식이다. 이는 Q 네트워크에서 자주 발생할 수 있는 과대추정 문제를 보수적인 방법으로 해결해주는 방법으로, Bellman target을 고려할 때 다음과 같이 이를 적용할 수 있다.

$$y = R(s_t,a_t,s_{t+1}) + \gamma\min_{j=1,2} Q_{\theta_j}\left(s_{t+1},\pi(s_{t+1})\right)$$

2.3 오프라인 강화학습

기존의 강화학습은 다른 기계학습 패러다임과는 다르게 대부분 비 실용적이며 비교적 좁은 도메인에서의 성공[30,31]을 거두었다. 이는 강화학습의 능동적인 학습 방법에 기인한다. 능동적인 강화학습 방식의 문제는 정책을 업데이트할 때마다 데이터 세트를 재수집해야 하는데, 이는 위험성을 동반하는 동시에 데이터 수집에 대한 비용이 많이 드는 작업이다.[32] 이와 같은 문제를 완화하기 위해 현대 기계 학습의 성공을 장려한 크고 다양한 데이터 세트를 고려하게 된다. 오프라인 강화학습 문제의 경우 앞서 정의하였던 강화학습 문제를 고정된 데이터 세트를 기반으로 공식화하여 정의할 수 있다. 즉, 개체는 학습을 위해 추가적인 상호작용 없이 사전 수집된 데이터 세트만을 이용하여 최적의 정책을 학습해야 한다.

데이터 세트 $\mathfrak{D} = \{(s_t,a_t,s_{t+1},r_t)_i\}$는 특정 행동 정책을 통해 수집될 수 있으며, 이는 하나의 정책이 아닌 다수의 정책 혼합 π_β으로 고려될 수 있다. 데이터

세트의 상태는 행동 정책에 의해 유도된 상태 분포에서 샘플링 $s_t \sim d^{\pi_\beta}(\cdot)$되며, 행동은 이와 같은 혼합정책에서 샘플링 $a_t \sim \pi_\beta(\cdot|s_t)$된다. 다음 상태는 상태 전이 확률에 의해 샘플링 $s_{t+1} \sim T(\cdot|s_t, a_t)$되며, 보상은 보상 함수에 의해 계산 $r_t = R(s_t, a_t, s_{t+1})$된다. 즉, 데이터 세트에 포함된 상태 및 동작 샘플의 분포는 행동 정책을 표현할 수 있다.

Ⅲ. 오프라인 알고리즘 리뷰

이번 장에서는 현재 활발히 사용되고 있는 대표적인 오프라인 알고리즘에 대한 리뷰를 수행한다. 사전에 수집된 고정된 데이터 세트 기반으로 정책을 학습하는 오프라인 패러다임의 학습 방법인 모방 학습(imitation learning), imitative learning, 오프라인 강화학습 알고리즘을 포함한다.

3.1 모방학습

대표적인 모방학습 알고리즘으로는 행동복제 알고리즘인 BC(behavioral cloning)이 있다.[33] BC 알고리즘은 오프라인 패러다임의 초기에 고려된 알고리즘으로 데이터 세트 내의 상태 및 행동 페어를 모방하기 위해 가장 기본적인 형태의 지도학습을 따른다. 알고리즘의 행동 정책 학습을 위한 손실 함수는 다음과 같다.

$$L(\phi) = \mathbb{E}_{s_t, a_t \sim \mathfrak{D}} \left(a_t - \pi_\phi(s_t) \right)^2$$

이와 같은 손실함수는 네트워크의 생성물인 예측된 행동 $\tilde{a}_t = \pi_\phi(s_t)$과 실제 데이터 세트에서의 행동 a의 오차를 최소화 하는 것을 목표로 한다. 모방학습의 경우 전문가가 생성한 최적의 데이터 세트에서는 좋은 성능을 보이지만, 차선 데이터 세트에서는 좋은 성능을 보이지 못한다.

3.2 Imitative Learning

Imitative learning은 모방학습과 강화학습 방식을 결합한 중간 단계의 학습 방법이다.[16] Imitative learning의 경우 기존의 강화학습에서 정책을 학습시키기 위한 목적 함수에 모방학습의 손실함수 $\mathbb{E}_{s_t, a_t \sim \mathfrak{D}} \left(a_t - \pi_\phi(s_t) \right)^2$를 auxiliary 항으로 고려한 학습 방법을 의미한다. 가치 기반 네트워크의 학습 방법은 유지한 채, 다음과 같은 정책 손실함수를 고려한다.

$$L_{BC}(\phi) = \mathbb{E}_{s_t, a_t \sim \mathfrak{D}} \Big[\big(Q_\theta(s_t, \pi_\phi(s_t) - \alpha \ln \pi_\phi(s_t) \big) - \alpha \left(a_t - \pi_\phi(s_t) \right)^2 \Big]$$

3.3 오프라인 강화학습

3.3.1 Policy Constraints

정책 규제 알고리즘은 명시적인 규제 방식을 이용하는 explicit policy constraint 방법과 암묵적으로 규제를 가하는 implicit policy constraint 방식으로 구분할 수 있다. 정책 규제기반의 오프라인 강화학습 알고리즘은 오프라인 패러다임으로의 전환과 함께 초기부터 연구되었으며 현재 가장 많은 연구가 진행되었다. 정책 규제 알고리즘의 기본적인 목적함수 형태는 다음과 같다.

$$L_{PC}(\phi) = \mathbb{E}_{s_t \sim \mathfrak{D}, a_t \sim \pi(s_t)} [Q_\theta(s_t, a_t)]$$
$$\text{subject to } D\left(\pi_\phi(s_t), \pi_\beta(s_t) \right) \leq \epsilon$$

여기서 $D(\cdot,\cdot)$은 분포 간 거리(divergence) 측정 방법을 의미하며, ϵ은 divergence threshold를 의미한다. $\pi_\beta(s_t)$의 경우 데이터 세트를 만든 행동 정책을 의미하기 때문에 데이터 세트 $\mathfrak{D}$에서 샘플링될 수 있는 행동의 분포를 의미할 수 있다.

1) Explicit Policy Constraint

명시적인 정책 규제는 학습중인 정책 π_ϕ를 규제 방식을 이용하여 π_β 분포에 가깝게 유도하는 방식이다. 명시적인 정책 규제를 가해주는 알고리즘에는 BCQ(batch-constrained Q), BEAR(bootstrapping error accumulation reduction), BRAC(behavior-regularized actor-critic), Fisher BRC(Fisher behavior-regularized critic) 등이 있다.

BCQ[11]: BCQ 알고리즘의 경우 다음과 같은 정책을 고려한다.

$$\pi_\phi(a_t|s_t) = \arg \max_{a_i + \zeta_\psi(s_t, a_i)} Q_\theta \left(a_i + \zeta_\psi(s_t, a_i) \right)$$
$$\text{for } a_i \sim \pi_\beta(\cdot|s_t), \quad i = 1, \cdots, N$$

여기서 $\zeta_\psi(s_t, a_t)$는 사전에 정의된 범위 내에서 클리핑되는 noise를 출력하는 perturbation 네트워크를 의미하며, N은 샘플의 수를 의미한다. BCQ는 행동의 완전한 모방이 아닌 유사한 행동을 수행할 수 있도록 정책

을 규제한다. 이때 완전한 규제를 원한다면 perturbation 모델 ζ_ψ를 고려하지 않거나, 하나의 샘플 N= 1만을 고려할 수 있다. 이와 같이 N개의 sample을 사용하거나, noise가 추가된 모델은 제한된 데이터 세트 내에서 행동의 다양성을 일정 부분 보장하며 분포 이동 문제를 완화할 수 있다.

BRAC[34]: 강화학습에서 가장 대표적으로 사용되는 actor-critic 알고리즘을 기반하여 행동을 규제하는 방법으로, 정책 개선 및 정책 평가 단계 중 하나의 목적함수 항에 분포의 거리 개념을 처벌로써 고려해줄 수 있다. 본 논문에서는 다양한 거리(KL divergence, Wasserstein distance, MMD 등)를 측정할 수 있는 방법의 경우 성능 면에서 큰 차이를 보이지 않는 다는 것을 보였으며, 정책을 정규화하는 작업보다는 가치함수에 처벌항을 고려해주는 경우가 보다 긍정적인 효과를 가져온다고 보고하였다. 처벌항을 고려한 가치 함수는 다음과 같이 정의할 수 있다.

$$V_D(s_t) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{s_{t\sim\mathfrak{D}}} \Big[R^{\pi_\phi}(s_t) - \alpha D\left(\pi_\phi(\cdot\,|s_t), \pi_\beta(\cdot\,|s_t)\right)\Big]$$

위 식에서 $R^{\pi_\phi}(s_t) = \mathbb{E}_{a_t\sim\pi_\phi(\cdot|s_t)}[\mathrm{R}(\mathrm{s_t}, \mathrm{a_t})]$를 의미하며, α는 처벌항의 가중치를 조절해주는 상수항을 의미한다.

Fisher-BRC[35]: Fisher BRC는 BRAC에서 보고한 actor-critic 알고리즘의 규제 관련 항에 Fisher divergence를 이용하여 엔트로피 정규화 기법을 고려한 알고리즘이다. 고려되는 목적함수는 기존의 TD-error와 더해 볼츠만(Boltzmann) 분포를 고려해준 다음의 정규화 항을 고려한다.

$$J_{Fisher}(\theta) = J(\theta) + \lambda \mathbb{E}_{s_{t\sim\mathfrak{D}}}\left[F\left(\frac{\exp\big(Q_\theta(s_t,\cdot)\big)}{\sum_a \exp\big(Q_\theta(s_t,a)\big)}, \pi_\beta(\cdot|s_t)\right)\right]$$

여기서 $F(\cdot,\cdot)$은 두 정책 분포의 Fisher divergence를 의미하며 $\frac{\exp(Q_\theta(s_t,\cdot))}{\sum_a \exp(Q_\theta(s_t,a))}$는 볼츠만 정책으로 정의할 수 있다.

2) Implicit Policy Constraints

BEAR[13]: BEAR 알고리즘은 f-divergence 로서 Gaussian kernel 과 함께 최대 평균 불일치(maximum mean discrepancy; MMD) 거리 개념을 고려하여 π_ϕ를 규제하는 방법으로 다음과 같이 정책을 규제할 수 있다.

$$\pi_\phi(a_t|s_t) = \max_\pi \mathbb{E}_{s_t\sim\mathfrak{D}, a_t\sim\pi(s_t)}\left[\min_{j=1,\cdots,K} Q_j(s_t,a_t)\right]$$
$$\text{subject to } \mathbb{E}_{s_t\sim\mathfrak{D}}[MMD(\mathfrak{D}(s_t), \pi(\cdot\,|s_t))] \le \epsilon$$

이와 같은 방법은 경험적으로 샘플의 수가 적어지는 경우 π_ϕ와 π_β 사이의 MMD가 유사함을 보장하며, 차선의 행동을 포함하는 경우에도 좋은 정책을 찾을 수 있음을 보였다.

CRR[15]: CRR 알고리즘은 분포 이동 문제로 인해 발생하는 데이터 세트에 존재하지 않는 행동에 대한 $Q(s_t, a_t)$ 값의 직접적인 평가를 피하기 위해 필터함수를 고려하여 다음과 같은 정책을 최적화 한다.

$$\pi_\phi(a_t|s_t) = \arg\max_\pi \mathbb{E}_{s_t,a_t\sim\mathfrak{D}}[f(Q_\theta,\pi,s_t,a_t)\log\pi(a_t|s_t)]$$

여기서 $f(\cdot)$은 양에서만 존재하는 단조 증가 함수(monotonically increasing function)로 정의되는 모든 함수가 고려될 수 있다. 만약 $f(\cdot)$ = 1이 성립하는 경우, 정책 학습 방식은 모방학습(BC)와 유사하게 여겨질 수 있으며, advantage function $A(s_t, a_t)$를 기반으로 하여 $f(\cdot) := \mathbf{1}[A(s_t,a_t) > 0]$ 혹은 $f(\cdot) := \exp(A(s_t,a_t)/\beta)$와 같이 정의할 수 있다. 여기서 $\mathbf{1}[\cdot]$은 지시함수(indicator function)를 의미하며 β는 항의 가중치로 고려되는 temperature 파라미터를 의미한다.

AWR[36]: 행동 정책을 규제하는 것은 오프라인 강화학습 기간동안 기존의 off-policy 알고리즘과 비교하여 안정적인 정책을 형성할 수 있었다. 하지만, 행동 정책의 추정에 대한 bootstrapping 에러는 피할 수 없었으며, 이를 해결하기 위해 저자는 AWR 알고리즘을 제안하였다. AWR 알고리즘은 데이터 세트에 포함된 궤적 전체를 고려하는 Monte Carlo 방식을 고려하여 다음과 같은 advantage function $A(s_t,a_t) = G_t - V(s_t)$을 이용하여 다음과 같이 정책과 가치함수를 학습한다.

$$V_\rho(s_t) = \arg\min_V \mathbb{E}_{s_t,a_t\sim\mathcal{D}}\left[\left|\left|G_t - V_\rho(s_t)\right|\right|^2\right]$$

$$\pi_\phi(a_t|s_t) = \arg\max_\pi \mathbb{E}_{s_t,a_t\sim\mathcal{D}}\left[\log\pi(a_t|s_t)\exp\left(\frac{1}{\beta}\left(G_t - V_\rho(s_t)\right)\right)\right]$$

AWAC[14]**:** AWR을 기반으로 하는 AWAC는 advantage function $A(s_t, a_t)$를 추정하기 위해 상태 행동 가치 함수 TD-error를 최소화 할 수 있는 $Q(s_t, a_t)$를 고려한다. 즉, AWAC에서 advantage function은 다음과 같이 정의할 수 있다.

$$A(s_t, a_t) = Q_\theta(s_t, a_t) - \mathbb{E}_{a\sim\pi_\phi}[Q_\theta(s_t, a)]$$

3.3.2 Regularization

Regularization은 정책 및 가치 함수의 학습에 고려될 수 있으며 value regularization 및 policy regularization으로 구분할 수 있다.

1) Value Regularization

CQL[12]**:** Value regularization 기법이 고려된 가장 대표적인 오프라인 강화학습 알고리즘으로는 CQL이 있다. CQL은 분포 이동 문제를 해결하기 위해 데이터 세트에 포함된 상태-행동 쌍의 가치는 최대화 하며 포함되지 않은 상태-행동 쌍의 가치는 최소화하여 정책 학습에 있어 데이터 세트에 포함된 정책 분포와 유사하도록 규제를 가하는 효과를 준다. 이에 대한 구체적인 regularizer는 다음과 같다.

$$\mathcal{R}(\theta) = \max_{\pi_\phi}\Big[\mathbb{E}_{s_t\sim\mathcal{D},a\sim\pi_\phi(\cdot|s_t)}[Q_\theta(s_t, a)] - \mathbb{E}_{s_t,a_t\sim\mathcal{D}}[Q_\theta(s_t, a_t)] + \mathcal{R}(\pi_\phi)\Big]$$

여기서 $\mathcal{R}(\pi_\phi)$는 정책의 regularizer를 의미한다.

N-SAC[18]**:** N-SAC는 앙상블(ensemble) 네트워크와 함께 clipped Q-learning 기법을 고려한 알고리즘으로 N개의 critic network를 고려하며 주어진 상태 및 행동 쌍에 대한 가장 비관적인 (가장 낮은) Q 함수를 고려하여 actor 및 critic network를 업데이트 한다.

$$Q_{\theta_i} = \min_{\theta_i}\mathbb{E}_{s_t,a_t,s_{t+1}\sim\mathcal{D}}\Big[\Big(Q_{\theta_i}(s_t, a_t) - \Big(R(s_t, a_t, s_{t+1}) + \gamma\mathbb{E}_{a'\sim\pi_\phi(s_{t+1})}\Big[\min_{j=1,\cdots,N} Q_{\theta'_j}(s_{t+1}, a') - \beta\log\pi_\phi(a'|s_{t+1})\Big]\Big)\Big)^2\Big]$$

$$\pi_\phi = \min_\phi \mathbb{E}_{s_t\sim\mathcal{D},a\sim\pi(s_t)}\Big[\min_{j=1,\cdots,N} Q_{\theta_j}(s_{t+1}, a) - \beta\log\pi_\phi(a|s_t)\Big]$$

EDAC[18]**:** 비관적인 Q 함수를 고려하기 위해 앙상블 네트워크를 이용하는 N-SAC의 경우 특정 업무에서 굉장히 많은 양의 앙상블 네트워크를 요구할 수 있다. 이는 계산 및 시간 비용을 크게 증가시킬 수 있으며, 이를 해결하기 위한 알고리즘으로 cosine similarity를 이용한 알고리즘인 EDAC이 제안되었다. EDAC은 학습된 Q 함수 사이의 유사도를 regularizer로 고려하여 학습과정에서 Q 함수 다양성을 보장하며 다수의 네트워크를 학습할 수 있다. 즉, 유사도가 높은 네트워크의 비율을 줄여 N-SAC의 문제를 해결할 수 있다. 해당 알고리즘의 Q 함수는 다음과 같이 regularizer를 고려한다.

$$\mathcal{R}_{EDAC}(\theta) = \mathbb{E}_{s_t,a_t\sim\mathcal{D}}\left[\frac{1}{N-1}\right]\sum_{1\le i\ne j\le N}\left\langle\nabla_{a_t}Q_{\theta_i}(s_t, a_t), \nabla_{a_t}Q_{\theta_j}(s_t, a_t)\right\rangle$$

여기서 $\langle\cdot,\cdot\rangle$은 두 함수간의 cosine similarity를 의미한다.

2) Policy Regularization

SAC[26]**:** 대표적인 policy regularization 방법은 2장에서 살펴본 SAC 알고리즘이 있다. 해당 알고리즘은 entropy $\log\pi_\phi(s_t)$을 regularizer로 고려하며 정책 학습을 수행한다. SAC 알고리즘은 최적 정책의 확률적인 제어를 가능하게 하며, entropy 항을 통해 정책의 조기 수렴을 방지하여 훈련과정에서 exploration-exploitation 문제를 자동적으로 고려함으로서, 견고성과 안정성을 개선하였다.

3.3.3 In-sample Learning

IQL[37]**:** 앞서 확인한 모든 방법은 데이터 셋에 존재하지 않는 상태-행동 쌍에 대한 가치 평가가 수행된다. 따라서, 기존의 알고리즘은 이러한 상태-행동 쌍에 대한 가치 평가에 대한 정확한 평가가 어렵기 때문에 이를 규제하는 항을 고려하게 된다. In-sample learning의 경우 데이터 세트 내의 정보만을 고려하여 데이터 세트에 존재하지 않는 상태-행동 쌍에 대한 query를 전적으로 피하는 방법을 채택한다. 대표적인 알고리즘은 IQL이 있다. 해당 알고리즘은 AWR 기반으로 정책을 학습하며 on-policy 알고리즘인 SARSA 방식을 채택하여 데이터 세트 내 샘플만을 이용하여 문제를 해결한다. 또한, 기존의 대칭적인 손실 함수 대신 아닌 비대칭적인 손실 함수를 이용하여 가치 평가 모델의 학습을 수행한다. 총 3개의 네트워크(상태 가치 함수, 상태-행동 가치 함수, 정책)를 고려하며, 각각의 손실함수는 다음과 같다.

$$L(\rho) = \mathbb{E}_{s_t,a_t\sim\mathcal{D}}\left[L_2^\tau\left(Q_{\theta'}(s_t,a_t) - V_\rho(s_t)\right)\right]$$

$$L(\theta) = \mathbb{E}_{s_t,a_t,s_{t+1}\sim\mathcal{D}}\left[\left(R(s_t,a_t,s_{t+1}) + \gamma V_\rho(s_{t+1}) - Q_\theta(s_t,a_t)\right)^2\right]$$

$$L(\phi) = \mathbb{E}_{s_t,a_t\sim\mathcal{D}}\left[\exp\left(\beta\left(Q_{\theta'}(s_t,a_t) - V_\rho(s_t)\right)\right)\log\pi_\phi(a_t|s_t)\right]$$

여기서 $L_2^\tau(x) = |\tau - \mathbf{1}(x<0)|\ x^2$은 비대칭적 손실 함수인 expectile 손실 함수를 의미하며, τ는 비대칭도를 결정하는 계수를 의미한다.

본 논문에서 명시된 오프라인 강화학습 알고리즘의 분류는 표 1을 통해 확인할 수 있다.

Ⅳ. 토 의

4.1 해결되지 않은 문제 및 한계

오프라인 강화학습 연구 분야에서는 해결 및 발전이 필요한 다양한 요소들이 존재한다. 본 절에서는 오프라인 강화학습 분야에서의 open problem 및 앞으로의 방향성에 대해 토의한다.

오프라인 강화학습에서 해결해야 할 가장 큰 어려움은 분포 이동 문제이다. 해당 문제는 데이터 세트 내 존재하지 않는 uncertainty set에 대한 과대 추정(over estimation)으로 인해 발생한다. 해당 문제를 해결하기 위해 대부분의 오프라인 강화학습 연구들은 데이터 세트 내에 존재하지 않는 행동을 보수적으로 평가하거나, policy 자체에 constraint를 부여하는 방법을 사용한다. 이러한 규제 기반의 접근은 데이터 세트 내에 존재하지 않는 행동 선택 확률을 낮춤으로써, 실제 환경에서의 분포 이동 문제를 완화한다. 하지만 제약에 기반한 방식은 고정된 데이터 세트에 대해 높은 의존성을 갖기 때문에 데이터 세트 내 최적의 행동이 포함되지 않는 경우 개체는 최적의 행동을 학습하기 어렵다. 또한, 양질의 데이터 세트가 제공되지 않을 경우 과적합(over fitting)과 같은 문제가 쉽게 발생할 수 있다. 이러한 문제를 해결하기 위해서는 uncertainty set에 대한 정확한 평가가 선행되어야 한다.

오프라인 강화학습 기반 정책의 평가 방법 또한 꾸준히 연구돼야 할 요소이다. 현재 많은 오프라인 강화학습 연구에서는 정책 평가를 위해 학습 기간 중 주기적으로 온라인 환경에서 평가를 수행한다. 이러한 평가 방식에서는 불완전한 정책을 지닌 개체가 환경과 상호작용을 수행한다. 따라서, 상호작용의 위험성이 큰 작업에서는 적합한 평가 방법으로 고려되기 어렵다. 이에, 이전 경험을 통해서만 정책을 평가하는 OPE(off-policy

표 1. 오프라인 강화학습 알고리즘 분류
Table 1. Offline reinforcement learning algorithms classification

	TD3 +BC	BCQ	BEAR	BRAC	Fisher BRC	CRR	AWR	AWAC	CQL	N-SAC	EDAC	IQL
Value Regularize				O	O	O			O	O	O	O
Policy Regularize		O		O					O			
Explicit Constraint	O				O							
Implicit Constraint			O				O	O				O
Clipped Q	O			O				O		O	O	

evaluation) 기반의 평가가 고려되고 있으나 현재의 OPE 방식은 종종 부정확한 평가를 제공한다고 보고되고 있다[6]. 정책 평가 방법은 최적의 정책 선택 및 파라미터 조정 등 알고리즘의 개선을 위한 기준으로 사용되기 때문에 신뢰성 높은 평가 방법은 반드시 확립되어야 한다. OPE 기반 평가 방법 연구의 발전은 정확한 정책 성능 제공을 통해 오프라인 강화학습 알고리즘 개선에 크게 기여할 것이다.

마지막으로, 오프라인 강화학습을 위한 데이터 세트 구축 및 활용 관점에서도 다양한 연구가 시도될 수 있다. 특히, 실제 세계를 고려하는 오프라인 강화학습에서는 보상 신호를 labelling 하는 연구의 필요성이 강조될 수 있다. 실제 세계 데이터에는 기본적으로 보상 신호가 존재하지 않기 때문에, 설계자가 직접 보상 label을 부여해야 한다. 이는 매우 높은 비용을 요구하며, 사전 설계된 보상 함수에 제약을 받는다는 문제점이 존재한다. 따라서, 고품질의 데이터 세트를 구성하기 위한 보상 신호 부여와 관련한 연구는 높은 중요성을 갖는다. 동일한 관점에서, 비지도 학습 기반의 오프라인 강화학습 방법의 연구도 유망한 분야로 고려될 수 있다. 데이터의 label을 요구하지 않는 학습 방식은 실제 세계의 unlabeled 데이터 세트의 활용도를 크게 높일 것이다.

4.2 오프라인 알고리즘 별 상충점

오프라인 강화학습 알고리즘의 경우 본문 및 4.1에서 논의된 학습 및 적용 과정에서의 분포 이동 문제를 해결하기 위해 개발 되었다. 본 논문에서 조사된 오프라인 강화학습의 경우 1) Policy에 대한 규제 및 제약, 2) Value function에 대한 규제 및 제약, 그리고 3) In-sample learning을 통해 직접적인 규제를 최소화 한 방법과 같이 3가지로 구분 가능하다. 본 섹션에서는 각각의 알고리즘이 갖는 상충에 대해 논의한다.

4.2.1 Value function 규제 및 제약

Value function에 대한 직접적인 규제 및 제약 방법의 주된 목적은 데이터 세트 내에 포함되지 않은 데이터 샘플의 평가를 비관적으로 조정하는 것에 있다. 데이터 세트 내에 포함되지 않은 데이터 샘플의 평가는 과소 및 과대 평가 될 수 있는데, 이때 많은 오프라인 강화학습 알고리즘은 과대평가를 방지하는 것에 초점을 맞춘다. 즉, 경험해보지 못한 행동의 경우 보수적/비관적으로 낮은 가치를 갖는다고 평가를 수행한다. 이와 같이 비관적인 가치 평가를 수행하는 이유는 강화학습에서 exploitation 과정에서 행동을 선택하는 원리를 통해 이해할 수 있다. 강화학습에서 exploitation 과정에서 선택되는 행동은 높은 가치를 갖고 있다는 것을 의미한다. 이는 선택 가능한 행동 중 최적의 보상을 얻을 수 있는 행동이라는 의미를 내포한다. 즉, 경험해보지 못한 행동의 실제 가치를 판단하는 것은 어려울 수 있으며, 해당 행동이 실제로는 부정적인 결과를 가져올 수 있기 때문에 이에 대한 평가 자체를 보수적으로 수행하는 데 의의가 있다.

4.2.2 Policy 규제 및 제약

Policy를 규제 및 제약하는 방법의 목적은 데이터 세트 내에 존재하는 행동과 유사한 행동을 선택하게 하는데 있다. 알고리즘의 특성에 따라서 value function의 제약과 동시에 수행될 수 있다. Policy를 규제하는 가장 기본적인 방법은 실제 데이터 세트 내에 존재하는 행동과 정책을 통해 생성된 행동을 유사하게 만드는 것이다. 이를 수행하기 위해 기존의 BC와 같은 제약을 도입할 수 있으며, 행동 정책과 학습된 정책 간의 distance를 줄이는 방법 등이 존재한다.

Policy만을 규제하는 것과 value function을 규제하는 것의 가장 큰 차이는 데이터 세트에 존재하지 않는 샘플에 대한 가치 평가를 임의로 조정하지 않는 것에 있다. 즉, value function에서의 단점인 추가적인 데이터 샘플 및 온라인 미세조정 과정에서 발생할 수 있는 문제가 완화될 수 있다는 장점이 있다. 또한, 샘플이 적절한 경우 평균적인 성능을 보장할 수 있다. 하지만 샘플 내에 행동의 비율이 부적절한 행동들에 편중되어 있는 경우 잘못된 정책 규제가 수행될 수 있기 때문에 기존의 모방학습에서의 문제를 동일하게 겪을 수 있다.

4.2.3 In-sample Learning

In-sample learning의 골자는 학습 과정에서 데이터 세트 내에 존재하지 않는 샘플에 대한 평가 자체를 수행하지 않는 것에 있다. 일반적으로 강화학습에서 TD-target을 구하는 과정에서 next action에 대한 inference 과정이 요구된다. 이때 next state와 next action 페어는 데이터 세트 내에 존재하지 않는 경우가 존재할 수 있다. 즉, TD-target의 추정 값에 bootstrapping 에러가 발생하여 학습 과정에 문제를 발생시킬 수 있다. In-sample learning은 이를 방지하고자 TD-target을 state-action value function이 아닌 state value function으로 설정한다. 그리고, state value function은 기존과 같이 state의 평균적인 가치가 아닌 최적의 행동을 결정했을 때의 가치에 가깝게 평가한다.

이와 같은 방식은 정책 및 value function에 대한 어떠한 규제도 존재하지 않기 때문에 value function 규제

방식의 단점이었던 추가적인 데이터 샘플 및 온라인 미세조정 과정에서 우려되는 문제를 방지할 수 있다. 또한, policy 규제 방식의 단점이었던 샘플 내에 행동의 비율이 부적절한 행동들에 편중된 경우에도 특정 state에서 경험해본 행동 중 최선의 행동을 선택하게 될 수 있다. 하지만, 데이터 세트 내에 우수한 데이터가 존재하지 않는 경우에는 해당 방식 역시 좋은 성능의 정책을 형성하는 것을 기대하기에는 어렵다.

이와 같이 모든 방식들은 generalization을 수행하여 분포 이동 문제를 직접적으로 하는 것은 아니며, 분포 이동 문제를 무시한 채 정책이 보수적으로 최선의 행동을 결정하는 것을 목적으로 한다.

4.3 오프라인 강화학습을 위한 데이터 세트

오프라인 강화학습에서 개체는 고정된 데이터 내에서 학습을 수행한다. 이에 높은 데이터 품질을 구성하기 위한 연구가 꾸준히 수행되고 있다. 오프라인 강화학습에서 대표적으로 고려되는 데이터 세트는 synthetic 및 실제 세계 데이터로 구분될 수 있다.

4.3.1 Synthetic 데이터의 활용

Synthetic 데이터는 사전 학습된 정책을 통해 수집된 정보로, 학습된 정책을 지닌 에이전트가 환경과 직접 상호작용하여 수집된 데이터를 의미한다. 이때, synthetic 데이터의 분포는 정보 수집을 위한 에이전트의 정책 완성도에 따라 서로 다르게 형성된다. 이에 대부분의 연구에서는 각각 상이하게 학습된 정책을 활용하여 synthetic 데이터를 수집하며, 무작위의 행동을 수행하는 정책을 활용하기도 한다. 또한, 여러 개의 정책을 활용하여 하나의 데이터 세트를 구성하는 방법 또한 고려되고 있다. 이는 주어진 데이터 세트 내 경로 정보의 가치 평가를 통해 최선의 행동을 학습하는 오프라인 강화학습의 특성에 기반한 것으로 데이터의 다양성을 높이기 위한 방법으로 고려될 수 있다. 이처럼 synthetic 데이터를 이용한 강화학습은 활발히 수행되고 있다. 하지만, 미완성 정책을 통한 synthetic 데이터 수집 시 상호작용 단계에서 높은 위험성을 수반할 수 있다는 문제점이 존재한다.

4.3.2 실제 데이터의 활용

Synthetic data의 문제점 완화 및 현실과의 차이 극복을 위한 방법으로는 실제 세계의 데이터 세트를 활용하는 방식이 고려될 수 있다. 이를 위해서는 실제 데이터를 MDP에 기반한 경로(trajectory) 정보가 포함된 데이터 세트를 수집해야 한다. 하지만, 대부분의 실제 세계 데이터는 경로 정보가 존재하지 않기 때문에 사전 설계된 MDP에 기반하여 데이터 가공을 수행하는 과정이 요구된다. 이러한 데이터 가공 과정은 높은 비용을 요구하며, 신뢰성 높은 MDP를 사전에 설계해야 한다는 어려움이 존재한다. 이에 경로 정보가 일부 누락된 데이터 세트를 활용하여 오프라인 강화학습을 수행하는 연구가 활발히 진행되고 있다[38]. 실제 데이터를 활용하기 위한 또 다른 시도로는 준 지도학습(semi-supervised learning) 기반의 방법이 고려될 수 있다. 해당 방법에서는 label이 부여된 소수의 데이터를 활용하여, 누락된 특정 경로 정보를 할당한다[39]. 해당 방법을 통해 경로 정보가 누락된 데이터 또한 오프라인 학습이 가능한 데이터의 형태로 수집될 수 있다.

V. 결 론

본 논문에서는 고정된 데이터 세트를 기반으로 정책을 학습하는 오프라인 강화학습에 대해 소개한다. 기존의 실시간 상호작용을 필요로 하는 온라인 강화학습 패러다임에서 벗어난 오프라인 강화학습 방식은 다양한 측면에서 실용성을 제공한다. 하지만 단순한 시뮬레이터 환경이 아닌 실제 세계의 문제에 적용을 위해서는 많은 문제들이 남아있다. 강화학습 패러다임에서 사용 가능한 데이터 세트 가공 문제, 데이터 세트를 기반으로 학습된 정책을 신뢰하고 적용하기 위한 정책 평가 문제, 분포 이동에 따른 정책 동작 오류 문제 등 아직까지 실제 적용을 위해서는 부족한 실정이다. 구체적으로, 데이터 세트의 품질 향상 및 검증 방법, 행동 및 보상 정보와 같은 손실 등에 따른 정책 학습 방법, 분포 이동 문제를 해결하기 위한 온라인 및 오프라인 방식의 결합 등에 대한 연구는 실용성 향상에 크게 기여할 수 있는 새로운 연구 주제로 고려될 수 있다

References

[1] K. Arulkumaran, et al., "Deep reinforcement learning: A brief survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26-38, Nov. 2017. (https://doi.org/10.1109/MSP.2017.2743240)

[2] C. Eom, D. Lee, and M. Kwon, "Autonomous driving strategy for bottleneck traffic with prioritized experience replay," *J. KICS*, vol. 48, no. 6, pp. 690-703, Jun. 2023. (https://doi.org/10.7840/kics.2023.48.6.690)

[3] J. Degrave, et al., “Magnetic control of tokamak plasmas through deep reinforcement learning,” *Nature*, vol. 602, pp. 414-419, Feb. 2022. (https://doi.org/10.1038/s41586-021-04301-9)

[4] N. E. H. Ouamane and H. Belhadef, “Deep reinforcement learning applied to NLP: A brief survey,” *NTIC*, pp. 1-5, Mila, Algeria, Dec. 2022. (https://doi.org/10.1109/NTIC55069.2022.10100477)

[5] N. Le, et al., “Deep reinforcement learning in computer vision: A comprehensive survey,” *Artificial Intell. Rev.*, vol. 55, pp. 1-87, Apr. 2022. (https://doi.org/10.1007/s10462-021-10061-9)

[6] R. F. Prudencio, et al., “A survey on offline reinforcement learning: Taxonomy, review, and open problems,” *IEEE Trans. Neural Netw. and Learn. Syst.*, Mar. 2023. (https://doi.org/10.1109/TNNLS.2023.3250269)

[7] N. Gürtler, et al., “Benchmarking offline reinforcement learning on real-robot hardware,” *ICLR*, Kigali, Rwanda, May 2023.

[8] X. Fang, et al., “Offline reinforcement learning for autonomous driving with real world driving data,” *ITSC*, Macau, China, Oct. 2022. (https://doi.org/10.1109/ITSC55140.2022.9922100)

[9] H. Emerson, et al., “Offline reinforcement learning for safer blood glucose control in people with type 1 diabetes,” *J. Biomedical Informatics*, vol. 142, no. 6, Jun. 2023.

[10] C. Zhang, et al., “BRAC+: Improved behavior regularized actor critic for offline reinforcement learning,” *ACML*, Nov. 2021.

[11] S. Fujimoto, et al., “Off-policy deep reinforcement learning without exploration,” *ICML*, California, USA, Jun. 2019.

[12] A. Kumar, et al., “Conservative Q-learning for offline reinforcement learning,” *NeurIPS*, Vancouver, Canada, Dec. 2020.

[13] A. Kumar, et al., “Stabilizing off-policy Q-learning via bootstrapping error reduction,” *NeurIPS*, Vancouver, Canada, Dec. 2019.

[14] A. Nair, et al., “AWAC: Accelerating online reinforcement learning with offline datasets,” *arXiv preprint arXiv:2006.09359*, 2020. (https://doi.org/10.48550/arXiv.2006.09359)

[15] Z. Wang, et al., “Critic regularized regression,” *NeurIPS*, Vancouver, Canada, Dec. 2020.

[16] S. Fujimoto and S. S. Gu, “A minimalist approach to offline reinforcement learning,” *NeurIPS*, Dec. 2021.

[17] R. Agarwal, et al., “An optimistic perspective on offline reinforcement learning,” *ICML*, Vienna, Austria, Jul. 2020.

[18] G. An, et al., “Uncertainty-based offline reinforcement learning with diversified Q-ensemble,” *NeurIPS*, Dec. 2021.

[19] O. Nachum, el al., “DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections,” *NeurIPS*, Vancouver, Canada, Dec. 2019.

[20] O. Nachum, et al., “AlgaeDICE: Policy gradient from arbitrary experience,” *arXiv preprint arXiv:1912.02074*, 2019. (https://doi.org/10.48550/arXiv.1912.02074)

[21] L. Chen, et al., “Decision transformer: Reinforcement learning via sequence modeling,” *NeurIPS*, Dec. 2021.

[22] S. Lee, et al., “Offline-to-online reinforcement learning via balanced replay and pessimistic Q-ensemble,” *ICML*, Maryland, USA, Jul. 2022.

[23] K. Fang and P. Yin, “Planning to practice: Efficient online fine-tuning by composing goals in latent space,” *IROS*, Kyoto, Japan, Oct. 2022.

[24] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, MIT Press, 2018.

[25] V. Mnih, et al., “Human-level control through deep reinforcement learning,” *nature*, vol. 518, pp. 529-533, Feb. 2015.

[26] T. Haarnoja, et al., “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,” *ICML*, Stockholm, Sweden, Jul. 2018.

[27] D. Silver, et al., "Deterministic policy gradient algorithms," *ICML*, Beijing, China, Jun. 2014.
[28] T. Lillicrap, et al., "Continuous control with deep reinforcement learning," *ICLR*, San Juan, Puerto Rico, May 2016.
[29] H. V. Hasselt, et al., "Deep reinforcement learning with double Q-learning," *AAAI*, Arizona, USA, Feb. 2016.
[30] M. G. Bellemare, et al., "The arcade learning environment: An evaluation platform for general agents," *J. Artificial Intell. Res.*, vol. 47, no. 1, pp. 253-279, May 2013. (https://doi.org/10.1613/jair.3912)
[31] E. Todorov, et al., "Mujoco: A physics engine for model-based control," *IROS*, Algarve, Portugal, Mar. 2012.
[32] S. Levine, et al., "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," *arXiv preprint arXiv:2005.01643*, 2020. (https://doi.org/10.48550/arXiv.2005.01643)
[33] F. Torabi, et al., "Behavioral cloning from observation," *IJICAI*, Stockholm, Sweden, Jul. 2018.
[34] Y. Wu, et al., "Behavior regularized offline reinforcement learning," *arXiv preprint arXiv:1911.11361*, 2019.
[35] I. Kostrikov, R. Fergus, et al., "Offline reinforcement learning with Fisher divergence critic regularization," *ICML*, 2021.
[36] X. B. Peng, et al., "Advantage-weighted regression: Simple and scalable off-policy reinforcement learning," *arXiv preprint arXiv:1910.00177*, 2019.
[37] I. Kostrikov, et al., "Offline reinforcement learning with implicit Q-learning," *ICLR*, 2022.
[38] A. Li, et al., "MAHALO: Unifying Offline Reinforcement Learning and Imitation Learning from Observations," *ICML*, Hawaii, USA, Jul. 2023.
[39] Q. Zheng, et al., "Semi-supervised offline reinforcement learning with action-free trajectories," *ICML*, Hawaii, USA, Jul. 2023.

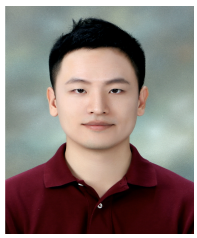
이 동 수 (Dongsu Lee)

2022년 2월 : 숭실대학교 의생명시스템학부 빅데이터컴퓨팅융합전공 학사
2022년 3월~현재 : 숭실대학교 지능형반도체학과 석박사통합과정
<관심분야> 강화학습, 계산신경과학, 자율주행
[ORCID:0000-0002-9238-4106]

엄 찬 인 (Chanin Eom)

2022년 8월 : 숭실대학교 전자정보공학부 IT융합전공 학사
2022년 9월~현재 : 숭실대학교 지능형반도체학과 석사과정
<관심분야> 강화학습, 인공지능, 자율주행
[ORCID:0009-0005-6340-6635]

최 성 우 (Sungwoo Choi)

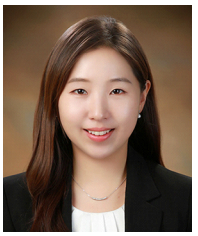


2005년 2월 : 카이스트 전기및전자공학부 학사
2007년 5월 : Mines ParisTech 로봇공학과 석사
2010년 9월 : Mines ParisTech 로봇공학과 박사
2011년 1월~2014년 6월 : 르노닛산 얼라이언스 책임연구원
2014년 6월~현재 : 현대자동차 자율주행사업부 책임연구원
<관심분야> 강화학습, 자율주행, 경로계획, MPC
[ORCID:0009-0001-3950-6856]

김 성 관 (Sungkwan Kim)

2019년 2월 : 한국산업기술대학교 메카트로닉스공학과 학사
2021년 2월 : 카이스트 조천식녹색교통대학원 석사
2021년 10월~현재 : 현대자동차 자율주행사업부 연구원
<관심분야> 자율주행, 강화학습, 경로계획
[ORCID:0000-0002-7035-4109]

권 민 혜 (Minhae Kwon)



2011년 8월 : 이화여자대학교 전자정보통신공학과 학사
2013년 8월 : 이화여자대학교 전자공학과 석사
2017년 8월 : 이화여자대학교 전자전기공학과 박사
2017년 9월~2018년 8월 : 이화여자대학교 전자전기공학과 박사 후 연구원
2018년 9월~2020년 2월 : 미국 Rice University, Electrical and Computer Engineering, Postdoctoral Researcher
2020년 3월~현재 : 숭실대학교 전자정보공학부 및 지능형반도체학과 조교수
<관심분야> 강화학습, 자율주행, 모바일네트워크, 연합학습, 계산신경과학
[ORCID:0000-0002-8807-3719]