ISSN : 1226-7244 (Print)
ISSN : 2288-243X (Online)
논문번호 21-04-09

j.inst.Korean.electr.electron.eng.Vol.25,No.4,644~649,December 2021
https://doi.org/10.7471/ikeee.2021.25.4.644

Self-Imitation Learning을 이용한 개선된 Deep Q-Network 알고리즘

Improved Deep Q-Network Algorithm Using Self-Imitation Learning

선 우 영 민*, 이 원 창**★

Yung-Min Sunwoo*, Won-Chang Lee**★

Abstract

Self-Imitation Learning is a simple off-policy actor-critic algorithm that makes an agent find an optimal policy by using past good experiences. In case that Self-Imitation Learning is combined with reinforcement learning algorithms that have actor-critic architecture, it shows performance improvement in various game environments. However, its applications are limited to reinforcement learning algorithms that have actor-critic architecture. In this paper, we propose a method of applying Self-Imitation Learning to Deep Q-Network which is a value-based deep reinforcement learning algorithm and train it in various game environments. We also show that Self-Imitation Learning can be applied to Deep Q-Network to improve the performance of Deep Q-Network by comparing the proposed algorithm and ordinary Deep Q-Network training results.

요 약

Self-Imitation Learning은 간단한 비활성 정책 actor-critic 알고리즘으로써 에이전트가 과거의 좋은 경험을 활용하여 최적의 정책을 찾을 수 있도록 해준다. 그리고 actor-critic 구조를 갖는 강화학습 알고리즘에 결합되어 다양한 환경들에서 알고리즘의 상당한 개선을 보여주었다. 하지만 Self-Imitation Learning이 강화학습에 큰 도움을 준다고 하더라도 그 적용 분야는 actor-critic architecture를 가지는 강화학습 알고리즘으로 제한되어 있다. 본 논문에서 Self-Imitation Learning의 알고리즘을 가치 기반 강화학습 알고리즘인 DQN에 적용하는 방법을 제안하고, Self-Imitation Learning이 적용된 DQN 알고리즘의 학습을 다양한 환경에서 진행한다. 아울러 그 결과를 기존의 결과와 비교함으로써 Self-Imitation Leaning이 DQN에도 적용될 수 있으며 DQN의 성능을 개선할 수 있음을 보인다.

Key words : Self-Imitation Learning, Actor-Critic Algorithm, Optimal Policy, Reinforcement Learning, Deep Q-Network

Ⅰ. 서론

강화학습의 에이전트(agent)가 알려지지 않은 환경에서 최적의 정책을 찾기 위해서는 탐험과 활용이라는 두 상반되는 개념이 필요하다. 활용이란 현재까지 가지고 있는 지식을 사용하여 단기적으로 보상이 가장 크게 되는 행동을 선택하는 것이다. 탐험이란 단기적으로는 보상이 적을지라도 더 좋은 행동을 찾는 것을 말한다. 탐험을 통해 더 좋은 행동을 찾아내고 그것을 활용하여 장기적으로는

* Dept. of Smart Robot Convergence and Application Engineering, Pukyong National University
** Dept. of Electronic Engineering, Pukyong National University
★ Corresponding author
E-mail : wlee@pknu.ac.kr, Tel : +82-51-629-6219
Manuscript received Dec. 1, 2021; revised Dec. 10, 2021; accepted Dec. 13, 2021.

더 큰 보상을 가질 수 있다. 하나의 행동을 선택할 때 활용과 탐험을 동시에 할 수 없으므로 이것은 종종 활용과 탐험의 갈등으로 인식된다[1].

강화학습에서 탐험과 활용 사이의 균형을 다루기 위해 다양한 알고리즘들이 사용되고 있지만, 가장 널리 사용되며 간단한 알고리즘으로 ϵ-greedy 알고리즘이 있다[2]. 이 알고리즘을 통해 현재 상태에서 탐욕적이지 않은 행동을 확률적으로 하나 선택하고 그 결과에 따른 보상을 얻을 수 있다. 하지만 그 주어진 경험을 충분히 활용하여 장기적으로 더 큰 보상을 얻는 정책을 찾는 것은 쉽지 않다. 따라서 에이전트가 행동-상태 공간이 큰 환경이나 매우 희소한 보상함수가 존재하는 환경에서 학습을 진행한다면 ϵ-greedy를 포함한 기존의 탐험과 활용 사이의 균형을 다루는 알고리즘만으로는 최적 정책을 발견하는 것에 한계가 있다. 그러나 Self-Imitation Learning(SIL)은 강화학습의 에이전트가 과거의 좋은 경험으로부터 학습할 수 있게 만들어 주는 알고리즘으로 과거의 좋은 경험을 재생산함으로 에이전트를 간접적으로 깊은 탐험으로 이끌어 더 좋은 정책을 발견할 수 있게 도와준다[3]. 그리고 SIL은 쉽게 구현되며 actor-critic 구조의 강화학습 알고리즘들에 간단히 결합될 수 있다. 또한 SIL을 사용한 몇몇 알고리즘들의 학습 결과를 통해 성능이 개선되었음이 증명되었고 에이전트가 모르는 환경의 새로운 영역을 탐험하는 것만큼이나 과거에 탐험으로 인해 알게 되었던 특별한 경험을 활용하는 것도 중요하다는 것이 밝혀졌다. 따라서 SIL을 actor-critic 구조가 아닌 강화학습 알고리즘에도 적용하여 더 좋은 정책을 찾게 해주자는 생각은 자연스럽다.

본 논문을 통해 Self-Imitation Learning(SIL)의 아이디어와 방법을 Deep Q-Network(DQN)[4]에도 적용시키는 방법을 제안한다. 그러나 SIL의 알고리즘은 actor-critic 구조를 가지지 않는 강화학습 알고리즘인 DQN에는 직접 적용될 수 없다. 따라서 본 논문에스는 SIL의 알고리즘을 DQN에 적용하기 위해 다음의 두 가지 방법을 사용했다. 첫 번째로 DQN의 가치 함수 업데이트 공식에 사용될 목표값을 수정하였다. 두 번째로 메모리에 저장되고 샘플링을 통해 학습에 사용되던 경험의 형태를 수정했다. 이 두 방법을 통해 SIL을 DQN에 적용할 때 생기는 문제들을 해결하였고 두 알고리즘을 결합할 수 있었다. 그리고 OpenAI Gym[5]에서 제공해주는 그림 1과 같은 대표적인 환경에서의 학습 결과를 통해 과거의 좋은 경험을 효율적으로 활용하는 SIL이 DQN을 개선할 수 있음을 보여준다.

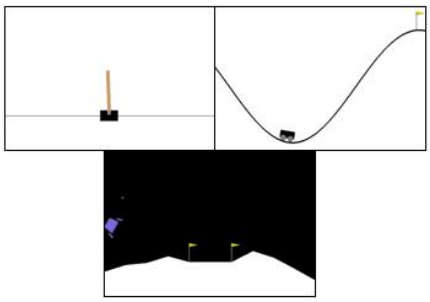

Fig. 1. Screen shots from OpenAI Gym Envrionments : CartPole, MountainCar, LunarLander.
그림 1. OpenAI Gym 환경들의 스크린 샷 : CartPole, MountainCar, LunarLander

II. 배경 지식

1. Deep Q-Network

Deep Q-Network(DQN)은 기존의 강화학습에서 표로 표현되던 행동 가치 함수를 신경망을 이용한 함수 근사를 통해 표현하여 상태공간의 규모가 큰 환경에서도 학습이 가능하도록 설계된 심층 강화학습 알고리즘이다. DQN은 비활성 정책 시간차 학습 제어 알고리즘인 Q-Learning[6]을 제어 알고리즘으로 사용하며 경험 재사용 방법을 통해 데이터 간의 상관관계를 줄였고 타겟 네트워크의 분리를 통해 학습의 안정성을 더하였다. DQN이 나온 뒤로 다양하게 개선된 DQN 알고리즘들이 제안되었다. 그리고 심층 강화학습 알고리즘의 성능을 끌어올리는 방법들이 DQN을 통해 실험적으로 확인되었다. 대표적인 개선된 DQN 알고리즘으로 Double DQN과 Dueling DQN 등이 있으며[7-8], DQN을 통해 실험적으로 그 성능을 확인한 알고리즘으로 Prioirtized Experience Replay, NoisyNet, Hindsight Experience Replay 등이 있다[9-11].

2. Self-Imitation Learning

Self-Imitation Learning(SIL)은 미지의 영역을 탐험하는 다른 탐험 전략들과는 달리 에이전트가 겪었던 사건을 다시 학습에 효율적으로 활용하는 비활성 정책 actor-critic 알고리즘이다. SIL은 기

존의 활성 정책 actor-critic 알고리즘에 쉽게 결합되어 상당한 성능 개선을 보여준다. 실제로 SIL은 Advantage-Actor critic(A2C) 및 Proximal Policy Optimization(PPO)에 적용되었고[12-13], 다양한 환경을 통해 실험적으로 과거의 좋은 경험을 재사용하는 것이 탐험에 큰 도움이 된다는 것이 증명되었다. SIL은 비활성 정책 actor-critic 알고리즘이므로 기존의 활성 정책 actor-critic 알고리즘들과는 달리 경험 재사용 메모리 $D=\{(s_t,a_t,G_t)\}$를 사용한다. 여기서 s_t, a_t, G_t는 각각 시간 t에서의 상태, 행동 그리고 이득을 의미한다. 그리고 과거의 좋은 경험만을 경험 재사용에 사용하기 위해 다음과 같은 비활성 정책 actor-critic 손실함수를 사용한다.

$$L^{sil}=E_{s,a,G\in D}[L^{sil}_{policy}+\beta^{sil}L^{sil}_{value}] \tag{1}$$

$$L^{sil}_{policy}=-\log\pi_\theta(a|s)(G-V_\theta(s))_+ \tag{2}$$

$$L^{sil}_{value}=\frac{1}{2}\|(G-V_\theta(s))_+\|^2 \tag{3}$$

여기서 $(\bullet)_+$연산은 $\max(\bullet,0)$를 의미하며 π_θ와 $V_\theta(s)$는 각각 모델의 파라미터 θ를 통해 표현된 정책과 상태 가치 함수를 의미한다. β^{sil}은 SIL의 하이퍼-파라미터이다. 이를 통해 과거의 좋은 경험만 학습에 사용하며 과거의 좋은 경험을 모방한다는 의미를 알 수 있다.

III. 제안하는 알고리즘

SIL은 과거의 좋은 경험을 활용할 수 있게 도와주는 알고리즘이지만 DQN에 적용할 때 고려할 점이 두 가지 있다. 첫 번째로 DQN은 actor-critic 구조를 가지는 강화학습 알고리즘이 아니다. DQN은 행동 가치 함수를 근사하는 하나의 네트워크를 가지며 그 가치를 기반으로 더 나은 정책을 선택하는 가치 기반 강화학습이다. 하지만 actor-critic 구조를 가지는 강화학습 알고리즘은 정책을 근사하는 actor와 그 정책을 평가하는 가치함수를 근사하는 critic, 이렇게 두 개의 네트워크를 유지하며 학습에 사용한다. 따라서 식 (1)-(3)의 SIL의 손실함수를 그대로 DQN에 사용할 수 없음은 명백하다. 두 번째로 SIL은 비활성 정책 방법으로써 스스로 경험 재사용 메모리를 유지하면서 활성 정책 방법 강화학습 알고리즘의 뒤에 단순히 결합되어 사용되었다. 하지만 DQN은 이미 경험 재사용 메모리를 가지고 있으며 에이전트가 겪었던 수많은 경험들을 저장하고 학습에 사용한다. 따라서 경험 재사용 메모리를 두 개를 유지하는 것은 굉장히 비효율적이며 하나의 경험 재사용 메모리를 사용하여 DQN과 SIL을 결합할 방법이 필요하다.

우선 우리는 에피소드의 매 시간마다 생기는 경험 : $m_t=(s_t,a_t,R_{t+1},s_{t+1},p_t)$을 에피소드 메모리 M에 저장한다. 이때 s_t와 a_t는 시간 t에서의 상태와 행동이다. R_{t+1}과 s_{t+1}은 시간 $t+1$에서 주어지는 보상과 상태이다. p_t는 시간 t에서 목표 정책과 행동 정책의 비로 식 (4)와 같다.

$$p_t=\frac{\pi(s_t|a_t)}{\mu(s_t|a_t)} \tag{4}$$

여기서 π는 목표 정책이고 μ는 행동 정책이다. 에피소드가 시간 $t=0$에서 시작하여 $t=T$에서 끝난다면 에피소드 메모리는 $M=\{m_0,m_1,\cdots,m_{T-1}\}$과 같이 된다. 하나의 에피소드가 끝이 난다면 M이 가지고 있던 모든 경험에 대해 다음과 같은 경험 재사용 메모리 $D=\{e_1,\cdots,e_N\}$에 저장될 경험 $e_t=(s_t,a_t,R_{t+1},s_{t+1},G_t,w_t)$를 생성한다. 여기서 N은 경험 재사용 메모리의 크기를 나타내며 G_t는 이득으로 식 (5)와 같고 w_t는 중요도 추출법 가중치로 식 (6)과 같다. 식 (5)에서 γ는 감쇠율로 미래 보상의 현재에서의 가치를 결정하게 된다.

$$G_t=\sum_{k=0}^{\infty}\gamma^k R_{t+k+1} \tag{5}$$

$$w_t=\prod_{k=t}^{T-1}p_k=\prod_{k=t}^{T-1}\frac{\pi(s_t|a_t)}{\mu(s_t|a_t)} \tag{6}$$

이렇게 만들어진 e_t가 저장된 경험 재사용 메모리에서 정해진 개수만큼 랜덤 샘플링을 통해 경험을 추출하여 미니 배치를 형성하고 학습에 사용된다.

다음으로 우리는 제어 알고리즘의 가치 함수 업데이트 공식에 사용될 목표 값을 수정하였다. DQN의 경험 재사용 메모리에서 샘플링한 경험 e_t로부터 학습에 사용되는 목표 값 y_{DQN}은 식 (7)과 같다.

$$y_{DQN}=R_{t+1}+\gamma\max q(s_{t+1},a_{t+1};\theta^-) \tag{7}$$

여기서 θ^-는 DQN의 타겟 네트워크 모델의 파라미터이고 a_{t+1}는 s_{t+1}에서 선택할 수 있는 행동이

다. 그리고 $q(s_{t+1},a_{t+1};\theta^-)$는 s_{t+1}에서 a_{t+1}의 행동 가치 함수를 파라미터 θ^-를 가지는 모델이 추정한 값이다. 우리가 제시하는 새로운 목표 값 $y_{DQN+sil}$은 식 (8)과 같다.

$$y_{DQN+sil} = w_t G_t K_t \lambda + y_{DQN}(1-K_t\lambda) \tag{8}$$

여기서 λ는 중요도 추출법 가중치가 곱해진 몬테카를로 이득 $w_t G_t$를 학습에 얼마나 사용할 것인지를 결정하는 하이퍼-파라미터로 과거의 좋은 경험을 학습에 얼마나 사용할 것인지를 결정한다. K_t는 $w_t G_t$를 목표 값에 사용할 것인지 아닌지 결정하는 파라미터로 식 (9)와 같다.

$$K_t = \begin{cases} 1 & w_t G_t > q(s_t,a_t;\theta) \text{and } w_t > 0 \\ 0 & otherwise \end{cases} \tag{9}$$

θ는 현재 학습 네트워크 모델의 파라미터이고 식 (9)가 의미하는 것은 다음과 같다. 만약 과거에 s_t에서 a_t를 선택한 뒤 실제로 관측했던 $w_t G_t$가 학습 네트워크가 추측한 가치보다 크다면 학습에 사용한다. 그리고 시간 t로부터 에피소드가 끝날 때까지 행동 정책에 의해 발생했던 상태-행동 궤적이 목표 정책에 의해서도 발생할 수 있어야 학습에 사용할 수 있다. 그렇지 않다면 $w_t G_t$를 사용하지 않는다. 만약 식 (8)에서 $K_t = 1$이라면 y_{DQN}과 $w_t G_t$의 가중평균이 되며, $K_t = 0$이라면 y_{DQN}과 정확히 같아진다. 따라서 $y_{DQN+sil}$를 사용한다면 과거의 좋은 경험을 적극적으로 활용할 수 있으며, 좋은 경험이 아니거나 목표 정책에 의해 일어날 수 없는 에피소드에 의해 생성된 경험에 대해서는 y_{DQN}을 통해 계속 가치함수를 업데이트할 수 있다.

Ⅳ. 시뮬레이션 결과

우리는 OpenAI Gym에서 제공해주는 대표적인 환경들인 CartPole, MountainCar, LunarLander에 대한 학습 결과를 통해 SIL을 결합한 DQN과 그렇지 않은 DQN을 비교한다. 그리고 본 시뮬레이션에서는 양쪽 알고리즘 모두 학습을 시작하기에 앞서 랜덤 정책을 사용해 경험 재사용 메모리를 가득 채우고 ϵ-greedy 에서 ϵ을 0으로 설정하고 학습을 진행하였다. 즉 행동 정책과 목표 정책 모두 탐욕적인 정책인 채로 학습을 진행했다. 이렇게 함으로 SIL의 도움만으로 DQN의 에이전트가 실제로 과거의 좋은 경험을 잘 활용하고 그로 인해 탐험에 도움이 되었는지 확인하고자 했다.

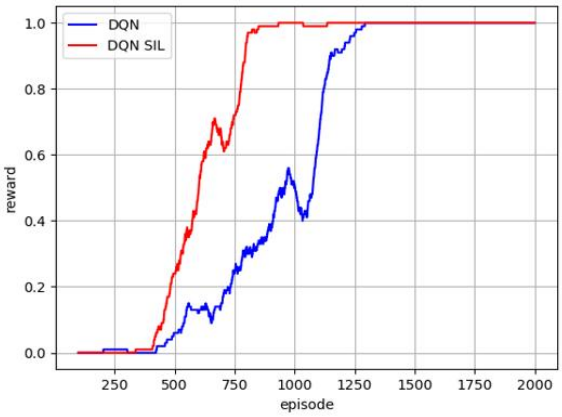

Fig. 2. Training results in MountainCar with Sparse reward.
그림 2. 희소한 보상이 존재하는 MountainCar에서의 학습 결과

가로축은 진행된 에피소드이며 세로축은 가장 최근 100개의 에피소드에서 받은 각각의 총 보상을 평균 낸 값이다. DQN은 약 200번째 에피소드에서 오른쪽 산으로 올라가는 경험을 발견하였지만, 그것을 활용하여 100개의 에피소드의 평균 보상이 +1이 되기까지 약 1150개의 에피소드를 소비했다. 반면에 SIL을 적용한 DQN은 약 300번째 에피소드에서 오른쪽 산으로 올라가는 경험을 발견하였고 평균 보상이 +1이 되기까지 약 600개의 에피소드를 소비하였다. 이 단순한 결과로부터 SIL을 적용한 DQN이 좋은 경험을 발견한 순간부터 그 경험을 적극적으로 활용하고 결국에는 탐험을 더 잘 한다는 것을 확인할 수 있다.

두 번째로 테스트한 환경들에는 앞의 희소한 보상이 존재하는 MountainCar 예제와는 달리 부드러운 보상 함수들이 존재한다. CartPole에서의 목적은 막대가 연결된 카트가 막대를 떨어트리지 않고 최대한 버티는 것이다. 막대를 떨어트리지 않는 매 타입 스텝마다 +1의 보상을 얻고 떨어트린다면 에피소드는 끝나게 되며 +0의 보상을 받는다. 또한 이번에 사용한 MountainCar에서의 목적은 최대한 빨리 오른쪽 산 위로 도달하는 것이며, 오른쪽 산 위로 도달하지 못하는 매 타입 스텝마다 -1의 보상을 얻고 도달하게 되면 +0의 보상을 받고 에피소드는 끝나게 된다. 마지막으로 LunarLander는 달 착

륙선을 빠르고 안전하게 착지시키는게 목적으로 위의 두 환경처럼 보상 함수가 단순하지 않다. 달 착륙선이 땅에 부딪히거나 착륙하면 에피소드는 끝나게 된다. 이번 테스트의 목적은 SIL을 적용한 DQN이 과거의 좋은 경험을 사용하는 데 있어 다

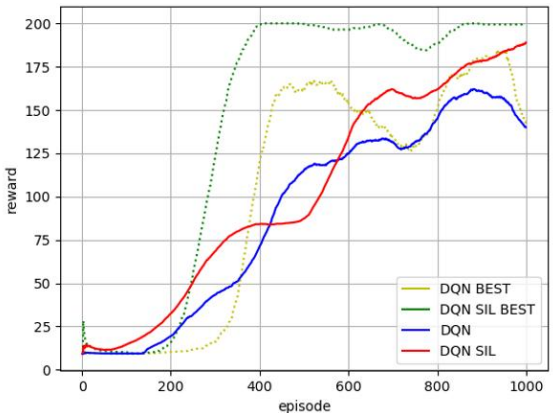

Fig. 3. Training results in CartPole.
그림 3. CartPole에서의 학습 결과

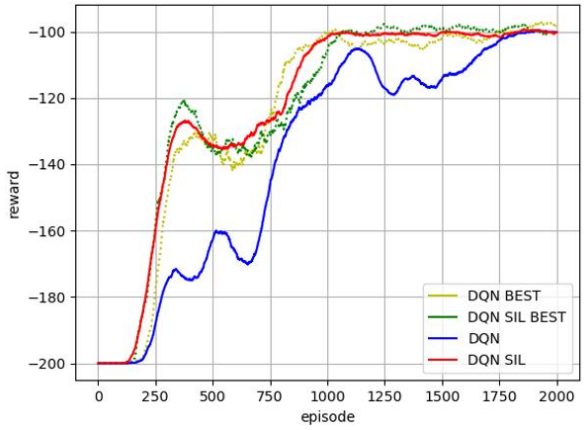

Fig. 4. Training results in MountainCar.
그림 4. MountainCar에서의 학습 결과

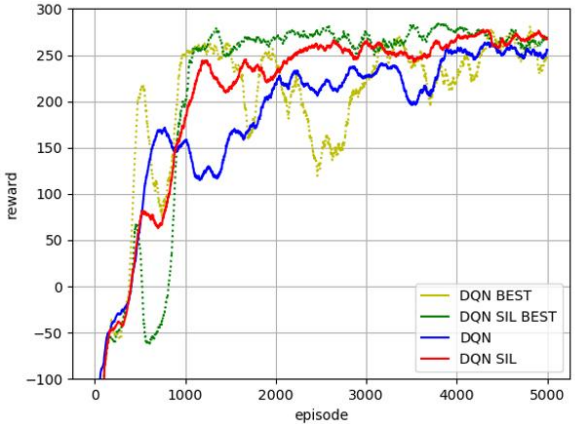

Fig. 5. Training results in LunarLander.
그림 5. LunarLander에서의 학습 결과

양한 목적과 보상을 가진 환경에서 일반적으로 DQN보다 좋은 성능을 보여주는지 확인하기 위함이다. 그 결과는 그림 3, 그림 4, 그림 5와 같다.

이번 테스트 결과에서도 마찬가지로 가로축은 진행된 에피소드이며 세로축은 가장 최근 100개의 에피소드에서 받은 각각의 총 보상을 평균 낸 값이다. 첫 번째 테스트 결과 그래프와의 차이점은 두 알고리즘을 사용하여 각각 다섯 번의 시뮬레이션을 진행한 뒤 그것들을 평균 낸 결과를 그래프에 실선으로 출력하였으며 최고로 높은 평균 보상을 얻었던 결과 하나를 추가로 점선으로 출력하였다. 이 결과에서 알 수 있듯이 SIL을 적용한 DQN이 그렇지 않은 DQN보다 다양한 환경에서 일반적으로 더 높은 평균 보상을 얻는 결과를 관찰할 수 있다. 우리가 이번 시뮬레이션에 사용했던 심층 강화학습 알고리즘의 하이퍼-파라미터는 표 1과 같다.

Table 1. Deep reinforcement learning model parameters.
표 1. 심층 강화학습 모델 파라미터

Hyper-parameters	Value
Architecture	FullyConnected(512)
	FullyConnected(256)
	FullyConnected(64)
Batch size	32
Start ϵ	0.0
End ϵ	0.0
Annealing step	0
Memory size	100000
Learning rate	0.0001
Discount rate	0.99
SIL λ	0.1

V. 결론

본 논문에서는 가치 기반 심층 강화학습 알고리즘인 DQN에 비활성 정책 actor-critic 알고리즘인 Self-Imitation Learning(SIL)을 적용하는 방법을 제안했고 다양한 환경에서의 학습 결과를 통해 과거의 좋은 경험을 사용하는 것이 DQN의 학습에도 큰 도움을 줄 수 있다는 것을 확인하였다. 경험 재사용 메모리에 저장될 경험의 수정과 실제로 강화학습 에이전트가 관찰했던 이득을 사용하는 새로

운 목표 값의 사용으로 인해 기존의 DQN의 알고리즘에 매우 간단히 SIL의 알고리즘과 아이디어를 접목할 수 있었다. 아울러 본 논문에서는 DQN에 SIL을 적용했지만, 제안된 방법을 활용하면 DQN을 개선한 심층 강화학습 알고리즘들뿐만 아니라 다양한 가치 기반 강화학습 알고리즘들에도 쉽게 SIL을 적용할 수 있고 에이전트를 깊은 탐험으로 이끌어 더 좋은 정책을 찾는 데 도움을 줄 수 있을 것이다.

References

[1] Sutton, Richard S., and Andrew G. Barto. “Reinforcement learning: An introduction,” MIT press, 2018.

[2] Kuleshov, Volodymyr, and Doina Precup. “Algorithms for multi-armed bandit problems,” arXiv preprint arXiv:1402.6028, 2014.

[3] Oh, Junhyuk, et al. “Self-imitation learning,” *International Conference on Machine Learning. PMLR,* 2018.

[4] Mnih, V., Kavukcuoglu, K., Silver, D. et al. “Human-level control through deep reinforcement learning,” *Nature,* Vol.0518, pp.529-533, 2015.

[5] https://gym.openai.com/docs/

[6] Watkins, Christopher JCH, and Peter Dayan. “Q-learning.” *Machine learning,* Vol.8, No.3-4, pp.279-292, 1992.

[7] Van Hasselt, Hado, Arthur Guez, and David Silver. “Deep reinforcement learning with double q-learning.” *Proceedings of the AAAI conference on artificial intelligence*. Vol.30. No.1. 2016.

[8] Wang, Ziyu, et al. “Dueling network architectures for deep reinforcement learning.” *International conference on machine learning. PMLR*, 2016.

[9] Schaul, Tom, et al. “Prioritized experience replay,” arXiv preprint arXiv:1511.05952, 2015.

[10] Fortunato, Meire, et al. “Noisy networks for exploration.” arXiv preprint arXiv:1706.10295, 2017.

[11] Andrychowicz, Marcin, et al. “Hindsight experience replay.” arXiv preprint arXiv:1707.01495, 2017.

[12] Mnih, Volodymyr, et al. “Asynchronous methods for deep reinforcement learning.” *International conference on machine learning.* PMLR, 2016.

[13] Schulman, John, et al. “Proximal policy optimization algorithms.” arXiv preprint arXiv: 1707.06347, 2017.

BIOGRAPHY

Yung-Min Sunwoo (Member)

2020 : BS degree in Electronic Engineering, Pukyong National University.
2020~present : MS student in Smart Robot Convergence and Application Engineering, Pukyong National University.

Won-Chang Lee (Member)

1983 : BS degree in Instrumentation and Control Engineering, Seoul National University.
1985 : MS degree in Electrical and Electronic Engineering, KAIST.
1992 : PhD degree in Electronic and Electrical Engineering, POSTECH.
1993~present: Professor, Pukyong National University