https://doi.org/10.7840/kics.2024.49.1.1

오프라인 강화학습 기반 무인항공기를 통한 재난상황 네트워크 복구 방법 연구

어 제 연◆, 이 동 수*, 권 민 혜°

Offline Reinforcement Learning Based UAV Training for Disaster Network Recovery

Jeyeon Eo◆, Dongsu Lee*, Minhae Kwon°

요 약

전쟁, 산불 등의 재난상황에서 기존 유무선 네트워크가 손실되었을 때 무인항공기를 활용하여 신속히 네트워크 복구를 하는 방안에 대한 연구가 활발히 진행되고 있다. 기존 연구에 널리 활용되는 온라인 강화학습기반 무인항공기 학습 방식은 환경과 실시간 상호작용을 하며 행동정책을 학습한다. 이러한 온라인 학습 방식은 재난상황에서는 효율적이지 않다. 또한, 3차원 공간에서의 네트워크 복구 임무는 성공 확률이 다소 낮아 보상 신호를 획득하기 힘든 희소보상 환경이기에 기존의 강화학습 알고리즘을 적용하는데 어려움이 크다. 이러한 문제점들을 해결하기 위하여 본 연구는 오프라인 강화학습에 LSTM(Long Short-term Memory)을 적용한 무인항공기 복구 학습방법을 제안한다. 오프라인 강화학습을 기반으로 한 제안된 방식은 고정된 사전수집 데이터셋을 활용하여 정책을 학습한다. 따라서 실제 환경과 상호작용하지 않고 안전한 정책학습이 가능하다. 또한, 시계열 정보를 고려하는 LSTM을 활용하여 보상에 기여한 행동들에 간접적으로 보상을 할당하므로 희소 보상환경에서도 원활한 정책학습이 가능하다는 장점이 있다. 모의실험을 통하여 제안하는 방법을 통해 학습한 무인항공기가 네트워크 부분손실 환경 복구를 우수한 성능으로 수행하는 것을 확인하였다.

Key Words : Unmanned Aerial Vehicle, Reinforcement Learning, Offline Reinforcement Learning, Sparse Reward Environment

ABSTRACT

In disaster scenarios, ongoing research seeks to swiftly restore disrupted networks using unmanned aerial vehicles (UAVs) when conventional wired and wireless infrastructures falter. Current research predominantly relies on online reinforcement learning, wherein UAVs acquire behavioral policies through real-time interactions. However, network restoration in three-dimensional spaces presents formidable challenges due to the scarcity of reward signals in low-probability success scenarios, rendering traditional reinforcement learning approaches less effective. To address these challenges, this paper proposes an approach that integrates Long short-term memory

※ 본 연구는 정보(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(IITP-2021-0-00739)과 한국연구재단의 지원(RS-2023-00278812)을 받아 수행된 연구임
◆ First Author : Soongsil University Department of Intelligent Semiconductors, aircleaner@soongsil.ac.kr, 학생회원
° Corresponding Author : Soongsil university School of Electronic Engineering and Department of Intelligent Semiconductors, minhae@ssu.ac.kr, 종신회원
* Soongsil University Department of Intelligent Semiconductors, movementwater@soongsil.ac.kr, 학생회원
논문번호 : 202308-035-B-RN, Received August 4, 2023; Revised September 20, 2023; Accepted October 8, 2023

(LSTM) into offline reinforcement learning, utilizing a fixed pre-collected dataset to enable safe policy learning without direct real-world interaction. The LSTM's capability to assign rewards to action sequences contributing to success facilitates smoother policy development even within sparse reward environments. Empirical simulation experiments confirm the effectiveness of our method in enabling UAVs to efficiently recover partial network loss.

Ⅰ. 서 론

최근 드론 등 무인항공기 기술이 발달함에 따라, 무인항공기가 네트워크의 노드로서 네트워크를 구축 및 복구하는 기술이 주목받고 있다. 무인항공기는 공중 기동을 통해 사람 및 차량의 접근이 어려운 환경에도 접근가능하다. 이러한 장점으로, 기존 유무선망이 부분 손실된 산불 및 지진 등의 재난상황이나 군사상황에서 손실 지점으로 신속히 이동하여 기존 네트워크를 복구할 수 있다[1].

연속적인 의사결정이 가능한 강화학습을 통해 무인항공기의 이동목표를 결정하는 방법으로 중앙제어방식과 자율적분산형 의사결정방식이 있다. 중앙제어방식은 노드 수가 증가함에 따라 모든 노드의 정보를 실시간으로 수집 및 처리해야하기 때문에 효율성이 낮다[2]. 따라서 본 연구는 노드가 관측 가능한 정보를 바탕으로 스스로 의사결정을 내리는 자율적분산형 의사결정 방법을 고려한다. 자율적 분산형 의사결정을 토대로 하는 강화학습 시, 학습 주체인 개체는 환경의 모든 정보인 상태정보 중 관측 가능한 정보인 관측정보에 따라 행동을 결정한다. 개체는 해당 관측-행동 쌍에 따른 보상을 얻는 과정을 반복하여, 최대의 누적 보상을 얻는 행동을 취하는 것을 목표로 정책을 학습한다.

하지만 해당 문제 상황에서 강화학습을 통한 정책학습에는 두 가지 어려움이 있다. 첫 번째는 강화학습 방식으로 널리 쓰이는 온라인 강화학습(Online Reinforcement Learning) 방식이 재난상황에서의 적용이 불리하다는 점이다. 온라인 강화학습 방식에서는 개체가 환경에 직접 행동을 취하며 정책을 학습한다. 이때 불안정한 초기 정책학습과정에서 취하는 시행착오 과정은 사회적 및 경제적 손실을 야기할 위험이 있다. 이에 따른 해결방법으로 본 연구는 고정 데이터셋만을 통해 학습이 가능한 오프라인 강화학습(Offline Reinforcement Learning)[3] 방법을 고려한다. 오프라인 강화학습을 통해 정책을 사전 학습한 무인항공기는 재난환경에서의 시행착오로 인한 위험부담을 경감할 수 있다. 두 번째로, 무인항공기가 네트워크 유실 지점으로 이동하는 과정에서 점진적인 보상신호를 받는 것이 아닌, 네트워크를 복구하는 순간에만 보상을 획득하는 환경적 특징이다. 강화학습은 보상신호를 근거로 정책을 학습하기 때문에, 이와 같이 목표를 달성한 순간에만 보상을 받는 희소보상환경(Sparse Reward Environment)[4]에서는 정책학습이 어렵다. 이에 따른 해결방법으로는 본 연구는 직접적인 보상뿐만 아니라, 시계열 정보를 고려하여 간접적으로 보상을 할당한다. 구체적으로, 희소보상환경에서의 학습을 위해 시계열정보를 고려하는 TD3(Twin Delayed Deep Deterministic Policy Gradient Algorithm)-LSTM 방법을 제안한다. 제안 방법의 성능평가는 시뮬레이션을 통해 수행한다.

본 논문은 다음과 같이 구성되어 있다. Ⅱ장에서 본 연구와 관련된 선행 연구를 소개한다. Ⅲ장은 본 연구에서 정의하는 네트워크 손실 상황, 희소보상 환경으로서의 해당 상황, 그리고 그에 따른 해결방법으로 TD3-LSTM 오프라인 강화학습 방법을 설명한다. Ⅳ장에서는 시뮬레이션을 통해 제안한 알고리즘의 성능을 평가하고, Ⅴ장에서 결론을 맺는다.

Ⅱ. 선행 연구

2.1 심층 강화학습

강화학습은 학습과 행동의 주체인 개체(Agent)가 환경(Environment)과 상호작용하며 시행착오를 바탕으로 행동정책(Policy)을 학습하는 기계학습 방법이다. 강화학습은 심층신경망(Deep Neural Network)을 접목한 심층강화학습(Deep Reinforcement Learning)의 등장으로 복잡한 환경의 문제 학습에 높은 성능의 달성이 가능해졌다[5]. 이에 로보틱스(Robotics), 경제, 게임 등 많은 분야에서 활발하게 연구되고 있다[6,7,8]. 강화학습방법으로 Model-based 방법 혹은 Model-free 방법이 있으며, 현실 세계의 모델을 정확히 알기 어렵기 때문에 대부분의 연구에서 Model-free 방법을 사용한다. Model-free 방법은 환경에서 높은 누적보상을 받을 수 있는 행동정책을 학습한다. 구체적으로, Model-free 방법은 행동 및 관측정보의 가치를 근사

하는 가치함수를 통해 최적의 행동을 취하는 Value-based 방식과 주어진 관측정보에 따른 최적의 행동을 취하는 정책을 근사하는 Policy-based 방식이 있다. Value-based 방식으로 대표적으로 Q-learning[5]이 있으며, Policy-based 방식으로는 REINFORCE[9]가 있다. 두 가지 방법을 절충한 방식으로는 본 연구에서 다룰 TD3(Twin-delayed Deep Deterministic Policy Gradient)알고리즘[10]과 같은 액터-크리틱(Actor-critic)[11] 방법이 있다.

2.2 심층 강화학습 기반 네트워크 복구 UAV 학습

군사 상황이나 화재 및 산사태와 같은 재난 상황에서 기존 네트워크가 손실된 경우, 네트워크 노드로 동작할 수 있는 무인항공기가 신속히 이동하여 네트워크를 복구할 수 있다. 무인 항공기는 사람의 조종이 닿지 않는 곳에서도 다양한 기상상태 및 지형지물을 고려하여 연속적인 의사결정을 내려야한다. 이에 자율적이고 연속적인 의사결정이 가능한 강화학습 기반 무인항공기 학습 연구가 활발히 진행되고 있다[12]. 이때 UAV 기기의 통신 및 에너지 자원 등의 기기 제한사항을 고려하며 FANET(Flying Ad-hoc Network)을 구축하는 연구들이 주를 이루고 있다. [13]에서는 UAV의 통신범위와 에너지 자원이 제한되어 있다는 점을 고려하여 UAV의 에너지 소비, 즉 이동량을 최소화하며 FANET을 구축함을 보였다. [14]에서는 UAV가 유선통신에 비해 비교적 불안정하며 보안이 취약함에 주목하여, 주파수 공유 참여 여부를 고려하며 FANET을 구축한다. 앞서 언급한 선행 연구들을 통해 UAV의 기기 제한사항을 고려하여 네트워크를 복구하는 강화학습 학습 연구가 활발히 진행됨을 확인할 수 있다.

하지만 기존 선행연구들은 네트워크 복구 행동이 위기 상황에서 이루어질 수 있다는 점의 주목도가 낮다. 온라인 강화학습은 직접 환경에 시행착오를 취하며 정책을 학습한다. 이러한 재난상황에서의 시행착오 행동은 경제적 및 사회적 손해 또한 야기할 수 있다. 따라서 본 연구는 개체가 환경과 직접 시행착오 및 상호작용하지 않고 고정된 데이터셋을 기반으로 학습하는 오프라인 강화학습을 고려한다.

2.3 오프라인 강화학습

기존에 널리 쓰이던 온라인 강화학습에서, 개체는 환경에 행동을 취하고 반복적으로 경험을 수집하여 정책을 학습한다. 이때, 학습된 정책을 바탕으로 높은

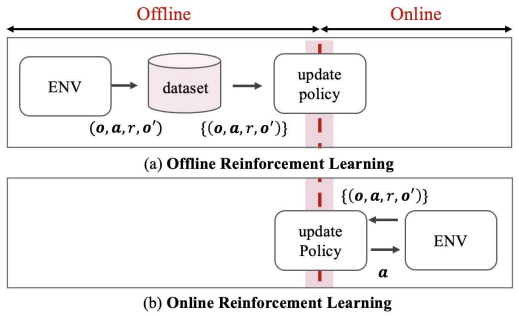

그림 1. 오프라인 강화학습과 온라인 강화학습의 비교
Fig. 1. Comparison between offline reinforcement learning and online reinforcement learning

누적보상이 예상되는 행동을 취하는 활용(Exploitation)과 새로운 행동을 취하는 탐색(Exploration)을 통해 정책을 학습한다. 탐색 시의 시행착오는 최적의 정책을 찾는 도움을 주지만, 의료 및 자율주행 등의 분야에서 치명적인 인적 및 물적 손실로 이어질 수 있다.

오프라인 강화학습은 이와 달리 환경과 직접 상호작용하며 탐색하지 않고, 고정된 사전수집 데이터셋만을 활용하여 정책을 학습한다(그림 1). 이러한 장점으로 최근 오프라인 강화학습에 대한 활발한 연구가 이어지고 있다[16,17]. 적용분야에 있어서 시행착오 행동으로 인한 에이전트 기기의 손괴 피해범위가 큰 로봇 분야에서 활발히 연구되고 있다[18,19]. 또한 시행착오 행동이 치명적인 인적 피해로 이어질 수 있는 자율주행 분야에서도 연구가 이어지고 있다[20,21]. 하지만 무인항공기를 활용한 신속한 네트워크 복구 상황 또한 안정적인 정책운용이 필요한 상황임에도 불구하고 온라인 강화학습을 고려하는 선행연구들이 주를 이루고 있다[12-14]. 재난상황에서의 안전한 네트워크 복구를 목표로 하는 본 연구는 오프라인 강화학습을 고려하여, 온라인 강화학습을 기반으로 하는 선행 연구들을 보완한다.

2.4 희소보상환경

희소보상환경은 강화학습에서 에이전트가 올바른 행동을 취했을 때 받는 보상이 드물게 주어지는 환경이다. 이는 보상을 기반으로 정책을 학습하는 개체가 목표에 도달할 수 있는 행동을 학습하기 어렵게 만들어, 정책 학습의 어려움을 야기한다[22]. 본 논문의 문제상황 또한 3차원 공간에서 네트워크를 복구하는 정확한 공간으로 이동해야만 보상을 받는 희소보상환경에 해당한다[23].

이러한 희소보상환경에서 정책 학습을 위해 기존 보상을 보완하여 내적 보상신호를 개체에게 전달하는 Intrinsic Reward 방법이 제안되었다. 그 중 활발히 연구되고 있는 Curiosity 기반 Intrinsic Reward 부여 방법[24]은 개체가 환경을 탐색하고 새로운 정보를 얻는 정도를 측정하여 보상으로 사용하는 방법이다. 개체는 자신이 얼마나 예측을 잘하고 있는지 평가하면서 환경을 탐험한다. 구체적으로, 개체는 다음 상태를 예측하고 실제 상태와 비교하며, 이 예측 오차를 Curiosity로 사용하여 에이전트에게 Intrinsic Reward를 부여한다. 이는 개체가 미래의 상태를 얼마나 정확하게 예측하는지에 대한 측정 지표가 된다. 온라인 강화학습을 고려하는 많은 선행연구들은 해당 방법을 활용하지만, 본 논문에서 고려하는 오프라인 강화학습의 경우 탐색과정이 존재하지 않기 때문에 효율적인 탐색을 전제로 하는 Intrinsic Reward 방법을 적용할 수 없다. 따라서 오프라인 강화학습에서도 희소보상환경에서 정책을 원활히 학습할 수 있도록, 본 논문은 시계열 정보를 고려해 간접적으로 보상을 할당하는 알고리즘을 제안한다.

Ⅲ. 재난상황 FANET 복원을 위한 TD3-LSTM 기반 UAV 오프라인 강화학습 방법

본 장에서는 문제상황을 정의하고 그 해결방법으로 TD3-LSTM 기반 오프라인 강화학습 방법을 제안한다.

3.1 문제상황 설정

본 연구는 $N\times N\times N$ 크기의 3차원 공간에서 일부 노드가 통신 불능 상태가 된 재난상황 부분손실 네트워크를 고려한다. 해당 상황에서 문제 노드의 정확한 위치 파악이 불가하고 문제 지역으로 사람의 직접적인 접근이 어렵다. 따라서, 무인 항공기는 자율적인 의사결정을 바탕으로 문제 노드의 위치로 이동 및 대체함으로써 네트워크 부분 복구를 진행하고자 한다. 고려하는 문제 상황에 대한 예시 이미지는 그림 2와 같다.

네트워크를 구성하는 노드 e_i의 종류는 $i\in\{\mathrm{r,a,s,d}\}$이다. 각각 정보를 수집하는 소스 노드 e_s, 데이터 센터로 수집한 정보를 전달하는 목적 노드 e_d , 그리고 소스 노드와 목적 노드 사이의 연결을 담당하는 A개의 보조 노드 $e_{a1},\cdots,e_{aA}$, 그리고 릴레이 노드 e_r을 의미한다. 따라서, 모든 노드의 집합은

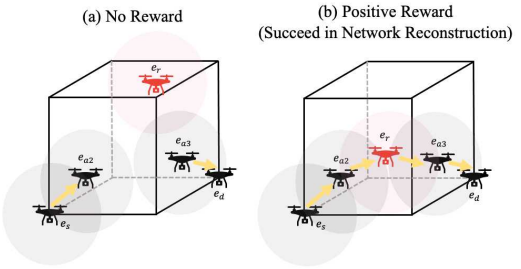

그림 2. 고려하는 부분손실 네트워크 복구 문제 상황
Fig. 2. Considered network recovery scenario for partial network loss

$\boldsymbol{E}=\{e_s,e_d,e_{a1},e_{a2},\cdots,e_{aA},e_r\}$로 정의된다. 모든 노드는 이중통신(Full Duplex)을 가정하며, 모든 노드의 통신 가능 거리이자 릴레이 노드의 관측 가능 거리는 δ로 고정되어 있다. 이때, 고정된 위치의 소스 노드와 목적 노드가 떨어져 있어 직접 통신이 불가능한 네트워크를 고려한다. 네트워크 연결은 보조 노드와 릴레이 노드를 통하여 소스 노드부터 목적 노드까지 통신가능 거리로 끊임없이 연결되었을 때로 정의한다. 보조 노드는 각 위치에 고정되어 움직이지 않으며, 이 중 랜덤하게 1개의 보조 노드가 동작 불능 상태가 되어 네트워크 연결이 부분 손실된 상황을 고려한다. 릴레이 노드는 변화하는 네트워크 부분 손실 상황에서 손실을 복구하는 위치로 이동하는 것을 목표로 한다.

3.2 마르코프 결정 과정(Partially Observable Markov Decision Process) 모델 제안

현실적인 강화학습 문제는 개체가 제한적인 관측정보를 통해 의사 결정을 수행하는 POMDP(Partially Observable Markov Decision Process)로 표현할 수 있다. POMDP는 튜플 $<S, A, O, R, \gamma>$ 로 정의한다. 개체는 학습과 행동의 주체로서, 상태 $\boldsymbol{s}\in S$를 관측하여 관측정보 $\boldsymbol{o}\in O$를 얻는다. 관측정보를 바탕으로 행동 $\boldsymbol{a}\in A$를 결정하며, 해당 관측-행동 쌍 $(\boldsymbol{o},\boldsymbol{a})$에 대한 보상 r을 획득한다. 개체는 시간에 따른 중요도인 감가율 γ가 적용된 누적 보상 $\sum_t \gamma^t r_t$을 최대화하는 행동 정책을 학습하는 것을 목표로 한다. 이때, t는 타임스텝을 의미한다.

3.2.1 상태정보(State)

상태 정보 $\boldsymbol{s}\in S$는 네트워크 내 모든 정보를 의미하며, 아래와 같이 정의한다.

$$\boldsymbol{s}=[\boldsymbol{P}_{\mathrm{r}},\boldsymbol{P}_{\mathrm{a}},\boldsymbol{P}_{\mathrm{s}},\boldsymbol{P}_{\mathrm{d}}]$$

$\boldsymbol{P}_{\mathrm{i}}=[\mathrm{x_i},\mathrm{y_i},\mathrm{z_i}]_{\mathrm{i}\in\{\mathrm{r,s,a,d}\}}$는 노드 e_i의 상태 정보인 x,y,z 축별 위치를 담고 있는 벡터를 의미한다. 이때, 릴레이 노드 e_r를 제외한 모든 노드 $e_{i\in\{\mathrm{a,s,d}\}}$의 위치는 단일 에피소드 내에서 시간과 관계없이 동일한 위치를 유지한다. 본 연구는 $N\times N\times N$ 크기의 3차원 공간의 FANET을 고려하므로, 노드 $e_{i\in\{\mathrm{a,s,d}\}}$의 축별 위치는 아래와 같이 $[0,N]$에서 정의된다.

$$0\le[x_i,y_i,z_i]_{i\in\mathrm{a,s,d}}\le\mathrm{N}$$

릴레이 노드 e_r은 UAV의 공중 주행 특징을 고려하여 높이 H이하로 주행하지 않는다. 따라서 e_r의 축별 위치는 아래와 같이 정의된다.

$$0\le[x_r,y_r]\le N$$
$$H\le z_r\le N$$

3.2.2 관측정보(Observation)

관측 정보 $\boldsymbol{o}\in O$는 릴레이 노드가 관측 가능한 정보로, 상태 정보 $\boldsymbol{s}$의 부분집합이다. 이는 아래와 같이 정의한다.

$$\boldsymbol{o}=[\boldsymbol{P}_{\mathrm{r}},\boldsymbol{l}_{\mathrm{x}},\boldsymbol{l}_{\mathrm{y}},\boldsymbol{l}_{\mathrm{z}}]$$

$\boldsymbol{P}_{\mathrm{r}}$은 릴레이 노드 e_r의 위치정보 벡터이다. $\boldsymbol{l}_{\mathrm{i}}=[\mathrm{l_{i,1}},\mathrm{l_{i,2}},\cdots,\mathrm{l_{i,J}}]_{\mathrm{i}\in\mathrm{x,y,z}}$은 릴레이 노드가 관측 가능한 J개의 인접 노드 $j_{i,}\cdots,j_J$ 까지의 축별 직선거리 벡터이다. 이때, 노드 i와 j간 유클리디안 거리(Euclidean Distance)가 δ 이하일 때, 노드 i와 j는 인접하며, 관측가능 및 통신 가능하다고 정의한다. 이때 무인항공기는 한정된 메모리 크기를 가짐에 따라 릴레이 노드가 관측할 수 있는 노드의 수는 최대 J개로 제한한다.

3.2.3 행동(Action)

행동 $\boldsymbol{a}\in A$는 릴레이 노드의 축별 이동거리 벡터 값이며, 아래와 같이 정의한다.

$$\boldsymbol{a}=[\Delta x_{r,}\Delta y_{r,}\Delta z_r]$$

이때, 각 원소는 아래와 같이 이동 가능거리 범위 $[-A_{\max},A_{\max}]$에서 정의된다.

$$-A_{\max}\le\Delta x_r,\Delta y_r,\Delta z_r\le A_{\max}$$

3.2.4 보상(Reward)

보상 R은 다음과 같이 정의한다.

$$R=\eta_1 v-\eta_2 c$$

$$v=\begin{cases}1, & \text{if network is recovered}\\ 0, & \text{otherwise}\end{cases}$$

$$c=\sqrt{(\Delta x_r)^2+(\Delta y_r)^2+(\Delta z_r)^2}$$

첫 번째 항 v는 네트워크 복구 여부에 따른 보상항이다. 네트워크가 복구된 경우 1, 복구 되지 않는 경우 0으로 정의된다. 두 번째 항 c는 네트워크가 복구된 후 릴레이 노드의 무의미한 움직임을 감소시키는 처벌항이다. 이를 통하여 개체인 릴레이 노드의 배터리 소모량을 최소화할 수 있도록 한다. η_1과 η_2는 전체 보상에 대한 보상항과 처벌항의 반영 정도를 조정하

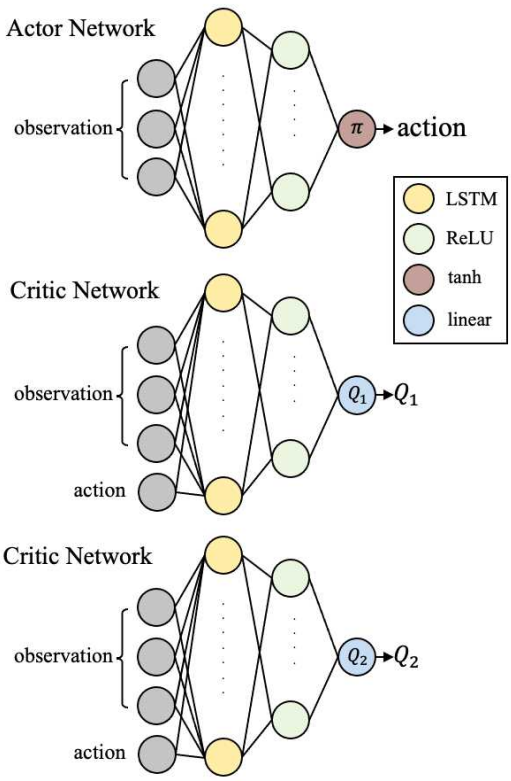

그림 3. TD3-LSTM 네트워크 구조
Fig. 3. TD3-LSTM network structure

는 계수이다.

3.3 TD3-LSTM 기반 오프라인 강화학습 알고리즘

본 절에서는 재난상황 FANET 복원을 위한 TD3-LSTM 알고리즘을 제안한다. TD3 알고리즘 설명 후, 본 연구의 문제상황의 특징인 희소보상환경을 고려한 TD3-LSTM의 알고리즘을 설명한다.

3.3.1 TD3 알고리즘

TD3 알고리즘은 액터-크리틱 알고리즘의 일종으로, 연속적인 관측 및 행동공간에서 정책을 학습할 수 있다. 수행할 행동을 결정하는 액터 네트워크 π_ϕ와 행동을 평가한 가치값을 액터 네트워크에게 피드백으로 제공하는 크리틱 네트워크 Q_{θ_1}, Q_{θ_2}로 구성된다. 액터 및 크리틱 네트워크는 반복적으로 업데이트되어 누적 보상을 최대화하는 정책을 학습한다. TD3 알고리즘은 크리틱 네트워크에서 두 개의 네트워크 Q_{θ_1}, Q_{θ_2}를 활용하여, 이 중 낮은 값을 학습에 사용한다. 이러한 Clipped Double Q-learning 기법을 통해 기존 크리틱 네트워크의 문제였던 과대 추정 문제를 완화한다. 또한 TD3 알고리즘은 원본 네트워크 $Q_{\theta_1}, Q_{\theta_2}, \pi_\phi$와 목표 네트워크 $Q'_{\theta_1}, Q'_{\theta_2}, \pi'_\phi$를 분리하는 특징이 있다. 목표 네트워크의 학습을 진행하다, 갱신 주기 d에 원본 네트워크와 목표 네트워크를 갱신한다. 이처럼 원본 네트워크의 갱신이 온건히 이루어지도록 하여 학습의 안정성을 향상한다. 더불어 행동을 결정하는 액터 네트워크가 가치를 판단하는 크리틱 네트워크에 비해 낮은 빈도로 학습을 진행하도록 하여 학습의 안정성을 향상한다.

TD3의 크리틱 네트워크 목적함수는 다음과 같다.

$$y = r + \gamma \min_{i=1,2} Q_{\theta'_i}(\boldsymbol{o}', \tilde{\boldsymbol{a}} + \epsilon)$$
$$\text{where } \tilde{\mathbf{a}} \sim \pi_{\phi'}(\mathbf{o}') + \epsilon, \epsilon \sim \text{clip}(\mathbb{N}(0, \sigma^2), -\mu, \mu) \quad (1)$$

$$\theta_i = \arg\min_{\theta_i} (y - Q_{\theta_i}(\boldsymbol{o}, \boldsymbol{a}))^2 \quad (2)$$

Q_θ는 $(\boldsymbol{o}, \boldsymbol{a})$의 가치를 평가하는 크리틱 네트워크로서, 식 1과 같이 TD-target인 y를 계산하여 식 2와 같이 업데이트된다. 수식 (1)에서 $\tilde{\boldsymbol{a}}$ 는 액터 목표 네트워크에서 계산된 행동을 의미하며 $\boldsymbol{o}'$는 $\boldsymbol{o}$에서 행동 $\boldsymbol{a}$을 취한 후 다음 타임스텝에서의 관측 정보를 의미한다. ϵ는 정책 평활화(smoothing)를 위한 노이즈이다. $\mathbb{N}(0, \sigma^2)$는 평균이 0이고 분산이 σ^2인 정규분포이며, ϵ는 해당 정규분포에서 범위 $[-\mu, \mu]$로 제한되어 추출된다.

액터 네트워크의 목적함수는 다음과 같다.

$$\phi = \arg\max_{\phi} Q_{\theta_1}(\boldsymbol{o}, \pi_\phi(\mathbf{o})) \quad (3)$$

액터 네트워크는 행동정책을 근사하는 네트워크로서, 식 (3)을 통해 Q값을 극대화하는 행동을 취하도록 업데이트됨을 알 수 있다. 앞서 기술한 네트워크의 갱신은 아래와 같이 이루어진다.

$$\theta_i' \leftarrow \tau\theta_i + (1-\tau)\theta_i' \quad (4)$$

$$\phi' \leftarrow \tau\phi + (1-\tau)\phi' \quad (5)$$

τ는 네트워크의 갱신 반영 정도를 조절하는 소프트 업데이트 파라미터이다. 즉, $\tau \in [0,1]$는 1에 가까울수록 원본 네트워크를 크게 반영하여 목표 네트워크를 갱신하고, 0에 가까울수록 적게 반영하여 갱신한다.

3.3.2 희소보상 환경을 고려한 TD3-LSTM 알고리즘

본 연구에서 고려하는 네트워크 부분손실 상황은 네트워크를 복구한 순간에 한정되어 보상을 받는다. 따라서, 3차원 공간에서 다양한 행동을 취하여도 네트워크를 복구하는 위치로 정확히 이동하여 보상을 받을 확률이 희박하다. 이러한 환경은 보상신호를 기반으로 하는 정책 학습이 어려운 희소보상환경으로 정의된다. 이러한 희소보상환경에서의 정책 학습 시, 시계열 정보를 고려하는 순환신경망을 활용하여 보상에 이르는 단계들에게 간접적으로 보상을 할당할 수 있다. 구체적으로, DQN 알고리즘에 관측 정보의 시계열을 고려하는 LSTM(Long Short-term Memory)[25]를 결합한 DRQN(Deep Recurrent Q-network)[26]은 기존 강화학습 DQN(Deep Q-network) 알고리즘에 비해 뛰어난 성능을 보여주었다. 이어서 ADRQN(Action-specific Deep Recurrent Q-network)[27]은 관측정보에 행동정보 또한 LSTM을 통해 고려하여 개선된 성능을 보였다. 하지만, 해당 연구들은 연속공간이 아닌 이산공간을 기반으로 한다는 한계가 존재했다. 이에 연속된 공간을 고려할 수 있는 DPG(Deterministic Policy Gradient)[28]알고리즘에 LSTM을 적용한 RDPG(Recurrent DPG)[29]가 연속된 공간에서 의 정책학습을 보였다. 본 연구는 DPG 알고리즘과 같이 연속 공간을 고려하되, 과대 추정 문

Algorithm 1. TD3-LSTM offline learning algorithm

Require: dataset $\mathcal{D}$, initial critic network parameters θ_1, θ_2 and an actor network parameter ϕ, initial target critic network parameters θ'_1, θ'_2 and target actor network parameter ϕ', target update frequency u, soft update ratio τ, train iteration step t, total train iteration T, LSTM hidden cell $\rho_Q, \rho_{Q'}, \rho_\pi, \rho_{\pi'}$, LSTM sequence length l

Load $\mathcal{D}$
for step $t = 0, 1, \cdots, T$
  Sample sequence tuples $< \boldsymbol{o}, \boldsymbol{a}, \boldsymbol{o}', r >_1, \ldots$
  $, < \boldsymbol{o}, \boldsymbol{a}, \boldsymbol{o}', r >_l$ from $\mathcal{D}$
  Initialize hidden cells $\rho_Q, \rho_{Q'}, \rho_\pi, \rho_{\pi'}$
  $\tilde{\boldsymbol{a}}_{seq} \leftarrow \pi_{\phi'}(\boldsymbol{o}'_{seq}, \rho_{\pi'})$
  $y \leftarrow r_{seq} + \gamma \min_{i=1,2} Q_{\theta'_i}(\boldsymbol{o}'_{seq}, \tilde{\boldsymbol{a}}_{seq}, \rho_{Q'})$
  Update critic networks:
  $\theta_i \leftarrow \arg\min_{\theta_i} (y - Q_{\theta_i}(\boldsymbol{o}_{seq}, \boldsymbol{a}_{seq}, \rho_Q))^2$
  if t mod u then
    Initialize hidden cell ρ_Q
    Update actor network:
    $\phi \leftarrow \arg\max_\phi Q_{\theta_1}(\boldsymbol{o}, \pi_\phi(\boldsymbol{o}_{seq}, \rho_\pi), \rho_Q)$
    Update target networks:
    $\theta'_i \leftarrow \tau\theta_i + (1-\tau)\theta'_i$
    $\phi' \leftarrow \tau\phi + (1-\tau)\phi'$
  end if
end for

제를 해결하며 더 빠르고 안정적인 학습 성능을 보이는 TD3 알고리즘에 LSTM 레이어를 적용한다. 해당 TD3-LSTM 알고리즘은 알고리즘1에서 확인할 수 있다. 구체적으로, LSTM[은 hidden cell ρ를 통해 이전 시간 단계의 정보를 장기적으로 기억한다. 이때, 시계열 학습단위인 시퀀스의 길이는 l로 정의하며, 길이 l

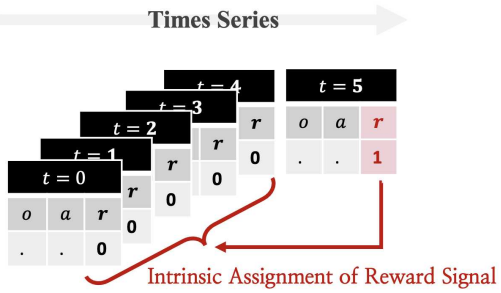

그림 4. 시계열을 고려한 궤적에의 보상신호 간접할당
Fig. 4. Intrinsic assignment of reward signal to trajectory considering time series

의 시퀀스 $\boldsymbol{o}_{1,\cdots,l}, \boldsymbol{a}_{1,\cdots,l}, \boldsymbol{o}'_{1,\cdots,l}, r_{1,\cdots,l}$는 각각 $\boldsymbol{o}_{seq}, \boldsymbol{a}_{seq}, \boldsymbol{o}'_{seq}, r_{seq}$로 정의한다. 본 연구는 LSTM 적용을 통해 최종 목표를 이루어 직접 보상을 받는 관측-행동쌍 뿐만 아니라, 해당 목표에 기여하였으나 보상을 받지 못했던 궤적에 간접적으로 보상을 할당한다(그림 4). 이를 통해 희소보상환경에서 원활히 정책을 학습하고자 한다.

Ⅳ. 시뮬레이션을 통한 알고리즘 성능 분석

4.1 시뮬레이션 및 학습환경 설정

제안하는 방법의 네트워크 복구 성능을 검증하기 위해 다음과 같은 설정 아래 시뮬레이션이 수행되었다. 모든 노드가 존재하는 3차원 환경의 크기는 $10 \times 10 \times 10$으로 설정한다. 소스 노드 e_s, 목적 노드 e_d 의 개수는 각각 1개로 고정되며, 이동과 학습의 주체인 릴레이 노드 e_r의 개수는 1개로 고정한다. 릴레이 노드는 각 에피소드마다 랜덤하게 주어지는 총 3가지의 네트워크 부분 손실 상황 시나리오에서 네트워크를 복구한다(표 1). 모든 시나리오에서 소스 노드와 목적 노드는 고정된 위치를 가지며, 각 시나리오에서 2개의 보조노드의 위치는 시나리오별로 고정된다. 각 에피소드는 40 타임스텝으로 이루어져있다. 즉, 릴레이 노드 e_r은 $10 \times 10 \times 10$ 3차원 환경 내의 랜덤한

표 1. 네트워크 부분손실 상황 시나리오 별 노드의 위치
Table 1. Node locations of partially network loss scenarios

		e_d	e_{a_1}	e_{a_2}
Scenario 1	[0,0,0]	[10, 10, 0]	[3.0, 2.0 ,2.7]	[5.95, 5.05, 3.3]
Scenario 2			[5.95, 5.05, 3.3]	[8.9, 8.1, 3.9]
Scenario 3			[8.9, 8.1, 3.9]	[3.0, 2.0 ,2.7]

위치에서 시작하여 40 타임스텝 안에 네트워크를 복구하는 위치로 이동해야 한다. 이때, 한 타임스텝에서 움직일 수 있는 릴레이 노드의 최대 이동거리인 $A_{\max}$는 0.3으로, 릴레이 노드의 이동가능거리는 $[-0.3,0.3]$에서 정의된다. 또한 릴레이 노드의 주행제한 높이 H는 2로 고정된다. 릴레이 노드의 관측 가능 및 모든 노드의 통신 가능 거리 δ는 5.5로 고정된다. 이때, 목적노드는 소스노드부터 시작하여 각 노드를 거쳐 차례로 전달되는 패킷을 수신하면, 네트워크 복구 성공 신호를 모든 노드에게 브로드캐스팅하는 네트워크를 고려한다. 이를 통해 각 타임스텝마다 릴레이 노드는 네트워크 복구 성공 여부를 알 수 있다고 가정한다.

4.2 학습에 사용된 데이터셋

본 연구에서 오프라인 강화학습에 필요한 데이터셋은 시뮬레이션 환경과 동일한 환경에서 수e_s집한다. 데이터셋 $\mathcal{D}$는 단위 경험 데이터 $<\boldsymbol{o},\boldsymbol{a},\boldsymbol{o}',r>$로 이루어진 길이 T_{epi}의 에피소드 $<\boldsymbol{o}_1,\boldsymbol{a}_1,\boldsymbol{o}_1',r_1,\cdots,\boldsymbol{o}_{T_{epi}},\boldsymbol{a}_{T_{epi}},\boldsymbol{o}_{T_{epi}}',r_{T_{epi}}>$ d개로 구성된다. 데이터셋은 Curriculum Learning[30]을 통해 학습한 전문가 정책에서 수집된 Expert Dataset와 무작위 정책에서 수집된 Random Dataset을 절반씩 혼합하여 사용한다[31,32].

4.3 비교에 사용된 알고리즘 및 설정

본 논문에서는 제안하는 방법의 검증을 위해 두 가지 알고리즘과의 비교를 진행하였다. 본 절은 비교에 사용한

알고리즘을 간략히 설명한다.

4.3.1 Behavior Cloning

BC(Behavior Cloning)는 데이터셋과 가장 유사한 행동을 취하도록 하는 지도학습법이다. 데이터셋을 기반으로 정책을 학습한다는 점에서 오프라인 강화학습과 유사하다. 하지만 오프라인 강화학습에서 관측-행동쌍의 가치를 판단하여 가장 높은 가치를 가지는 것으로 판단되는 행동을 취하는 것과 다르게, BC는 관측정보가 주어지면 데이터셋 내의 관측-행동쌍과 가장 비슷한 행동을 취하는 것이 목표이다. W_ψ가 $\boldsymbol{o}$가 입력으로

주어지면 $\boldsymbol{a}$를 출력하는 네트워크일 때, BC의 정책 학습을 위한 목적함수는 아래와 같다.

$$\psi = \arg\min_\psi (W_\psi(\boldsymbol{o}) - \boldsymbol{a})^2 \quad (6)$$

4.3.2 Imitative Learning

IL(Imitative Learning)[21]은 TD3 알고리즘의 액터 네트워크에 규제항으로 BC를 추가한 방법이다. IL은 BC 규제항을 추가하여 학습의 안정성을 향상한다. BC 규제항 이외는 TD3 알고리즘과 동일하며, IL의 액터 네트워크의 목적함수는 다음과 같다.

$$\phi = \arg\max_\phi (\lambda Q_{\theta_1}(\boldsymbol{o},\pi_\phi(\boldsymbol{o})) - (\pi_\phi(\boldsymbol{o}) - \boldsymbol{a})^2) \quad (7)$$

이때 λ는 규제항의 반영 정도를 조절하는 파라미터로서, 본 시뮬레이션에서는 0.5로 설정한다.

4.4 학습성능 평가 기준

본 논문은 제안하는 무인항공기의 부분손실 네트워크 복구방법의 성능 기준으로 Hit Ratio를 고려하며 Hit Ratio는 다음과 같이 계산한다.

$$\text{Hit Ratio} = \frac{1}{T_{epi}} \sum_{t=1}^{T_{epi}} r_t \quad (8)$$

즉, Hit Ratio가 1에 가까울수록 높은 네트워크 복구 및 유지성능을 보임을 알 수 있다.

4.5 시뮬레이션 결과 분석

그림 5와 6은 제안하는 TD3-LSTM 오프라인 강화학습과 타 알고리즘과의 성능을 비교하는 그래프이다. 그래프의 x축은 학습 Iteration이며, 그래프의 y축은 Hit Ratio, 즉 학습 성능을 나타낸다. 실선과 음영은 각각 10번의 실험 결과에 대한 평균과 95% 신뢰구간을 나타낸다.

그림 5는 제안하는 방법의 LSTM 적용 유무에 따른 학습 성능을 나타낸다. LSTM을 적용하지 않은 오프라인 강화학습방법의 성능을 나타내는 초록색 실선은 0.2 Hit Ratio를 유지함을 알 수 있다. 이를 통해 제안하는 방법이 LSTM을 적용하여 시계열 정보를 고려함으로써 희소보상환경에서 우수한 학습성능을 보임을 알 수 있다.

그림 6은 TD3-LSTM 오프라인 강화학습 알고리즘과 IL, BC 알고리즘과의 비교를 진행한다. 정확한 비교를 위하여 IL과 BC 알고리즘 또한 LSTM 레이어를 적용하였다. IL은 학습 Iteration 1만회 이전까지 제안하는 알고리즘과 비슷한 학습 성능을 보이다 약 0.4 Hit Ratio를 유지한다. 반면, 제안하는 알고리즘은 학

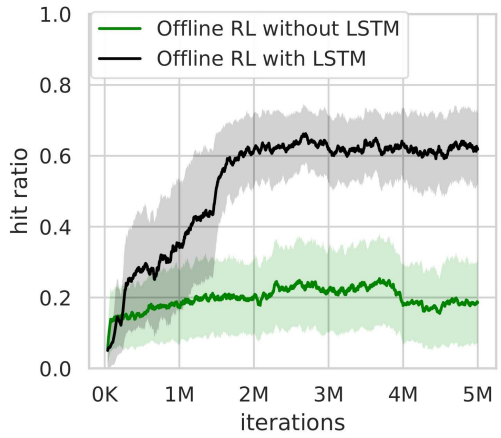

그림 5. LSTM 적용 여부별 학습 진행에 따른 학습 성능 비교
Fig. 5. Comparison of performance per training iteration among algorithms

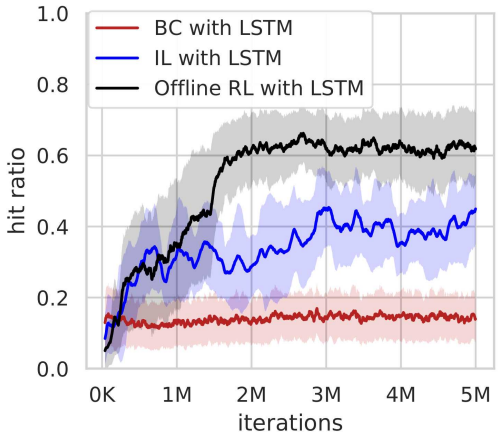

그림 6. 알고리즘별 학습 진행에 따른 학습 성능 비교
Fig. 6. Comparison of performance per training iteration among algorithms

습 진행에 따라 성능이 증가하여 약 0.6 Hit Ratio를 유지한다. 즉, 40 타임스텝중 약 16 타입스텝 동안 이동하여 약 24 타임스텝동안 네트워크를 복구하여 연결 상태를 유지함을 알 수 있다. BC 방법은 학습 진행과 무관하게 0.1 Hit Ratio를 유지한다. 해당 결과를 통해 오프라인 강화학습 기반 TD3-LSTM방법이 다른 알고리즘에 비해 우수한 네트워크 복구 성능을 가지고 있음을 알 수 있다.

V. 결 론

본 연구는 무인항공기의 네트워크 복원을 위한 오프라인 강화학습 기반 방법을 제안한다. 해당 TD3-LSTM 방법은 강화학습 알고리즘인 TD3의 신경망에 시계열 정보를 고려할 수 있는 LSTM 레이어를 적용하였다. 이를 통하여 희소보상 환경인 네트워크 복원상황에서 다른 알고리즘에 비해 높은 학습성능을 보이는 것을 확인하였다. 제안한 방법으로 학습한 UAV는 우수한 네트워크 복원 성능을 보여줄 뿐만 아니라, 오프라인 강화학습을 고려하기 때문에 재난상황에서의 활용을 기대할 수 있다. 후속 연구로는 더욱 복잡한 문제 상황에 대한 효과적인 대응을 할 수 있도록 여러 대의 복구 무인항공기를 활용하는 멀티에이전트 강화학습 연구를 진행하고자 한다.

Appendix

표 A1. 연구 환경 시스템 사양
Table A1. System specification of the research

CPU	ADM Ryzen 9 5900X 12-Core
GPU	NVIDIA GeForce GTX 1660 SUPER
RAM	128G
SSD	500GB

표 A2. 알고리즘 별 신경망 노드와 계층 정보
Table A2. Neural network node and layer of algorithm

	Input node	LSTM layer number	LSTM layer node
TD3-critic	13	1	[13, 32]
TD3-actor	10	1	[10, 32]
Imitative Learning	10	1	[10, 32]
Behavior Cloning	10	1	[10, 32]

	LSTM sequence length	Hidden layer number	Hidden layer node	Output node
TD3-critic	15	1	[32, 16]	1
TD3-actor	15	1	[32, 16]	3
Imitative Learning	15	1	[32, 16]	3
Behavior Cloning	15	1	[32, 16]	3

References

[1] J. Eo, D. Lee, and M. Kwon, “Offline reinforcement learning based UAV training for

emergency network recovery," *JCCI*, pp. 445-446, Yeosu, Korea, Apr. 2023.

[2] N. Kim, M. Kwon, and H. Park, "Q-learning based ad-hoc network formation strategy for wireless nodes with random mobility models," *J. KICS*, vol. 46, no. 11, pp. 1834-1845, 2021.

[3] S. Lange, et al., "Batch reinforcement learning," *Springer*, pp. 45-73, 2012. (https://doi.org/10.1007/978-3-642-27645-3_2)

[4] M. Riedmiller, et al., "Learning by playing solving sparse reward tasks from scratch," *ICML*, pp. 4344-4353, Vienna, Austria, 2018.

[5] V. Mnih, et al., "Playing atari with deep reinforcement learning," *NIPS Deep Learning Workshop*, Nevada, USA, 2013. (https://doi.org/10.48550/arXiv.1312.5620)

[6] P. Kormushev, et al., "Reinforcement learning in robotics: Applications and real-world challenges," *Robotics*, vol. 2, no. 3, pp. 122-148, Jul. 2013. (https://doi.org/10.3390/robotics2030122)

[7] P. Abbeel, et al., "An application of reinforcement learning to aerobatic helicopter flight," *Advances in NIPS*, vol. 19, Dec. 2006.

[8] A. Mosavi, et al., "Comprehensive review of deep reinforcement learning methods and applications in economics," *Mathematics*, vol. 8, no. 10, 2020. (https://doi.org/10.3390/math8101640)

[9] R. J. Williams, "Simple statistical gradient following algorithms for connectionist reinforcement learning," *Mach. Learn.*, vol. 8, no. 3-4, pp. 229-256, May 1992. (https://doi.org/10.1007/BF00992696)

[10] S. Fujimoto, et al., "Addressing function approximation error in actor-critic methods," *ICML*, pp. 1587-1596, Stockholm, Sweden, Jul. 2018.

[11] V. Konda, et al., "Actor-critic algorithms," *Advances in NIPS*, vol. 12, pp. 1008-1014, 1999.

[12] X. Liu, et al., "Reinforcement learning in multiple-UAV networks: Deployment and movement design," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8036-8049, 2019. (https://doi.org/10.1109/TVT.2019.2922849)

[13] C. H. Liu, et al., "Energy-efficient UAV control for effective and fair communication coverage: A deep reinforcement learning approach," *IEEE J. Sel. Areas in Commun.*, vol. 36, no. 9, pp. 2059-2070, 2018. (https://doi.org/10.1109/JSAC.2018.2864373)

[14] A. Shamsoshoara, et al., "Distributed cooperative spectrum sharing in UAV networks using multi-agent reinforcement learning," *CNCC*, pp. 1-6, Las Vegas, USA, 2019. (https://doi.org/10.1109/CCNC.2019.8651796)

[15] A. Kumar, et al., "Conservative Q-learning for offline reinforcement learning," *NeurIPS*, vol. 33, pp. 1179-1191, 2020.

[16] J. Eo, D. Lee, and M. Kwon, "The impact of dataset on offline reinforcement learning of UAV for emergency network recovery," *J. KICS*, Jeju, Korea, 2023.

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

[18] G. Zhou, et al., "Real world offline reinforcement learning with realistic data source," *ICRA*, pp. 7176-7183, 2023. (https://doi.org/10.1109/ICRA48891.2023.10161474)

[19] A. X. Lee, et al., "How to spend your robot time: Bridging kickstarting and offline reinforcement learning for vision-based robotic manipulation," *IROS*, pp. 2468-2475, Kyoto, Japan, 2022. (https://doi.org/10.1109/IROS47612.2022.9981126)

[20] X. Fang, et al., "Offline reinforcement learning for autonomous driving with real world driving data," *ITSC*, pp. 3417-3422, Macau, China, 2022. (https://doi.org/10.1109/ITSC55140.2022.9922100)

[21] B. Osiński, et al., "Simulation-based

reinforcement learning for real-world autonomous driving," *ICRA*, pp. 6411-6418, Paris, France, 2020. (https://doi.org/10.1109/ICRA40945.2020.9196730)

[22] J. Eschmann, et al., "Reward function design in reinforcement learning," *Reinforcement Learning Algorithms: Analysis and Applications*, Springer, pp. 25-33, 2021. (https://doi.org/10.1007/978-3-030-41188-6_3)

[23] C. Wang, et al., "Deep-reinforcement -learning-based autonomous UAV navigation with sparse rewards," *IEEE Internet of Things J.*, vol. 7, no. 7, pp. 6180-6190, 2020. (https://doi.org/10.1109/JIOT.2020.2973193)

[24] D. Pathak, et al., "Curiosity-driven exploration by self-supervised prediction," *ICML*, pp. 2778-2787, Sydney, Australia, 2017.

[25] S. Hochreiter, et al., "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997. (https://doi.org/10.1007/978-3-642-24797-2_4)

[26] M. Hausknecht, et al., "Deep recurrent Q-learning for partially observable MDPs," *AAAI*, Texas, USA, 2015.

[27] P. Zhu, et al., "On improving deep reinforcement learning for POMDPs," *Computing Research Repository*, 2017. (https://doi.org/10.48550/arXiv.1704.07978)

[28] D. Silver, et al., "Deterministic policy gradient algorithms," *ICML*, pp. 387-395, Beijing, China, 2014.

[29] N. Heess, et al., "Memory-based control with recurrent neural networks," *NIPS Deep Reinforcement Learning Workshop*, 2015. (https://doi.org/10.48550/arXiv.1512.04455)

[30] Y. Bengio, et al., "Curriculum learning," *ICML*, pp. 41-48, Montreal, Canada, 2009. (https://doi.org/10.1145/1553374.1553380)

[31] K. Schweighofer, et al., "Understanding the effects of dataset characteristics on offline reinforcement learning," *NeurIPS Deep RL Workshop*, Sydney, Australia, 2021.

[32] J. Eo, D. Lee, and M. Kwon, "Can expert dataset guarantee offline performance in sparse reward environment?" *ICML DMLR Workshop*, Hawaii, USA, 2023.

[31] S. Fujimoto, et al., "A minimalist approach to offline reinforcement learning," *NeurIPS*, vol. 34, pp. 20132-20145, 2021.

어 제 연 (Jeyeon Eo)
한국통신학회논문지 vol 48, no 11 참조

이 동 수 (Dongsu Lee)
한국통신학회논문지 vol 48, no 11 참조
[ORCID:0000-0002-9238-4106]

권 민 혜 (Minhae Kwon)
한국통신학회논문지 vol 48, no 11 참조
[ORCID:0000-0002-8807-3719]