https://doi.org/10.7840/kics.2025.50.6.884

군사적 지휘결심 지원을 위한 심층 강화학습 기반 무기-표적 할당 시스템 연구

이 재 휘•, 엄 찬 인*, 김 경 수**, 강 현 수**, 권 민 혜°

Deep Reinforcement Learning Based Weapon-Target Assignment to Support Military Decision-Making

Jaehwi Lee•, Chanin Eom*, Kyeongsoo Kim**, Hyunsu Kang**, Minhae Kwon°

요 약

최근 무기-표적 할당 (weapon-target assignment; WTA) 문제를 통해 지휘관의 지휘결심을 돕는 기술에 대한 연구가 활발히 진행되고 있다. 현대의 전장이 발전됨에 따라 WTA 문제 또한 현실적이고 복잡한 전장 환경을 고려한다. 이에 동적으로 최적의 결정을 내려야 하는 상황에서 심층 강화학습 방법이 활용될 수 있다. 따라서 본 논문은 심층 강화학습 기반의 WTA 문제에서 아군의 효율적인 의사결정을 위한 마르코프 의사결정 과정 (Markov decision process; MDP) 모델을 제안한다. 결과적으로 모의 실험을 통해 제안하는 MDP로 정의된 강화학습 모델이 heuristic 방법에 비해 요망 효과 달성률을 27.17%, 탄약 효율성을 38.61% 향상시켰으며, 무기 탄약 소모 비용은 11.98% 감소시킬 수 있음을 확인하였다.

키워드 : 무기-표적 할당, 심층 강화학습, 마르코프 의사결정 과정

Key Words : Weapon-target assignment, Deep reinforcement learning, Markov decision process (MDP), Deep deterministic policy gradient (DDPG), Twin delayed DDPG (TD3)

ABSTRACT

Recently, there has been significant research into technologies to assist commanders in military decision-making through the weapon-target assignment (WTA) problem. As the modern battlefield has evolved, WTA problems consider realistic and complex environments, where deep reinforcement learning methods can be utilized for dynamic decision-making. This paper proposes a Markov decision process (MDP) model for efficient decision-making in WTA problems, leveraging deep reinforcement learning to optimize the actions of friendly units. As a result, we confirmed that the reinforcement learning model based on the proposed MDP improved the commander's objective achievement by 27.17% and ammo efficiency by 38.61%, while reducing ammo usage cost by 11.98% compared to the heuristic approaches.

◆ First Author : Soongsil University Department of Intelligent Semiconductors, jaehwilee@soongsil.ac.kr, 학생회원
° 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
** Vision AI Research Institute, Konan Technology Inc., kyeongsoo.kim@konantech.com; hyunsu.kang@konantech.com
논문번호 : 202410-245-A-RN, Received October 17, 2024; Revised December 22, 2024; Accepted January 7, 2025

Ⅰ. 서 론

최근 군사 작전의 복잡성이 증가함에 따라 인공지능 기반의 지휘관 의사결정 지원 기술은 디지털 국방 체제 구축에 있어서 더욱 중요해지고 있다[1,2]. 이러한 상황에서 WTA는 지휘결심 AI 시스템, 지능형 전장 인식 및 판단, 자율적 전술 의사결정 등 다양한 최신 국방 기술들과 밀접하게 연관되어 있어 디지털 국방 체제 구축에 핵심적인 역할을 한다[3,4].

WTA는 군사적 의사결정의 핵심 요소로, 적군에 효과적인 피해를 입히는 동시에 아군의 무기 자원을 효율적으로 배분하는 것을 목표로 한다[5,6]. WTA는 전투 성과와 아군의 생존에 직접적인 영향을 미치며, 그 중요성은 현대 군사 작전에서 더욱 강조되고 있다. 특히, 지휘관의 전략적 목표를 달성하기 위해서는 단순한 피해 극대화가 아닌, 지휘관이 요망하는 목표 수준에 맞춘 자원 최적화와 함께 복잡하고 동적인 전장 환경에 대응하는 정교한 의사결정이 필수적이다.

기존 WTA 문제를 해결하기 위해 모든 가능한 해를 탐색하여 정확한 결과를 도출하는 exact 알고리즘이 사용되었지만, 계산 복잡도로 인한 한계가 존재하였다. 이에 빠른 계산 속도의 근사적인 방법인 heuristic 알고리즘 방법이 도입되었지만, 여전히 이러한 최적화 방법들은 현실적으로 복잡하고 전장 환경을 고려하는 의사결정 문제에서는 한계가 있었다.

심층 강화학습은 심층 신경망을 활용하여 동적인 환경에서도 유연한 의사결정이 가능한 특성으로 인해 기존의 WTA 해결방안의 한계를 완화할 수 있다[7]. 심층 강화학습 기반 의사결정 정책의 특성은 MDP 모델의 설계에 따라 달라질 수 있다. MDP는 강화학습 개체(agent)가 학습할 환경을 정의하며, 상태(state), 행동(action), 보상 함수(reward function) 등을 포함한다. 의사결정 개체는 MDP에서 정의된 상태에 따라 행동을 수행하고 설계된 보상 함수를 기반으로 보상을 받고 정책을 개선하기 때문에 해결하고자 하는 문제의 목표에 맞는 MDP의 설계가 매우 중요하다.

본 논문에서는 전장 환경이 동적으로 변화하는 상황에서 아군 부대의 적군 부대에 대한 효율적인 부대 및 무기 선택 의사결정 정책 구축을 목표로 한다. 기존 WTA 접근 방식의 한계를 보완하기 위해 심층 강화학습을 위한 MDP 모델을 제안하며, 정책 학습에는 심층 강화학습 알고리즘인 DDPG(Deep Deterministic Policy Gradient)[8]와 TD3(Twin Delayed DDPG)[9]를 사용한다.

본 논문의 주요 기여는 다음과 같다.

- 본 연구는 연속적이고 복잡한 상태-행동 공간을 갖는 군사적 의사결정 문제에 강화학습을 도입하여 기존 heuristic 방법에 비해 효과적으로 의사결정을 수행할 수 있음을 보였다.
- 강화학습을 적용하기 위한 MDP로 군사적 문제의 동적인 전장 환경 및 상호작용을 모델링하고, 군사적 특성을 고려한 상태, 행동, 보상 체계를 설계하였다.
- 제안한 MDP 기반의 강화학습 방법이 기존 heuristic 방법에 비해 다양한 군사적 성능 측면에서 우수함을 실험적으로 검증하였다.

본 논문의 구성은 다음과 같다. II장에서 본 연구의 선행연구에 대해 살펴보고, III장에서는 본 연구에서 해결하고자 하는 WTA의 시스템과 MDP 모델을 소개한다. IV장에서는 실험 설정과 제안한 모델을 기반으로 학습한 의사결정 정책을 분석하고 평가한 후 V장에서 결론을 맺는다. 본 논문에서 사용된 모든 기호와 표기법은 Appendix A1에서 확인할 수 있다.

Ⅱ. 선행연구

2.1 Weapon-target Assignment

WTA는 군사 작전 및 방어 시스템에서 무기와 표적 간의 최적 할당을 다루는 문제로, 주어진 무기를 표적에 할당하여 손실을 최소화하거나 이익을 극대화하는 것을 목표로 한다. WTA는 조합-최적화 문제로 다뤄졌으며 이를 해결하기 위한 방법으로 exact 방법과 heuristic 방법이 고려되었다.

Exact 방법은 전통적인 최적화 문제를 다루는 방법 중 하나로, WTA에도 적용이 가능하다[10,11]. 비교적 단순한 WTA에서 모든 가능한 해를 고려하여 최적의 해결책을 보장할 수 있다는 장점이 있지만, 무기나 표적의 수가 증가하여 복잡해짐에 따라 계산 비용이 급격하게 증가한다는 한계가 존재한다.

WTA의 복잡도가 증가함에 따라 계산 효율성을 높이기 위해 빠르고 유용한 해를 제공할 수 있는 근사적인 방법인 heuristic 기반의 방법이 활용되었다. 또한 기존 WTA 문제는 아군의 이익 최대화 뿐만 아니라 효율적인 자원 활용 등과 같은 추가적인 목표도 함께 달성하기 위한 multi-objective WTA 문제로 확장되었다[12,13]. [14]에서는 다중 목표 달성을 위한 genetic 알고리즘을 통해 제약 조건이나 목표가 다양한 WTA에서 최적에 근사한 해를 찾는데 유용함을 보였고, [15]에서는 기존의 인공 벌집 알고리즘(Artificial Bee Colony; ABC) 방법을 변형하여 다중 목표를 위한 방공 WTA 문제에서 의사결정의 적시성과 정확성을 크게 향상시킬 수 있

음을 보였다. 하지만 이러한 heuristic 방법에도 불구하고 복잡하고 동적인 환경에서 의사결정에는 여전히 어려움이 존재하였고, 이에 심층 강화학습 기반의 방법이 고려되었다.

2.2 Deep Reinforcement Learning

심층 강화학습은 심층 학습(deep learning)과 강화학습을 결합한 방법으로 모델 기반의 방식(model-based)[16,17]과 모델이 없는 방식(model-free)으로 분류할 수 있다. 대부분의 공학 문제를 고려할 때, 환경 모델을 정확히 아는 것은 불가능하기 때문에 모델이 없는 방식이 적용될 수 있다. 이 방식에서 개체는 환경과의 직접적인 상호작용으로 정책을 학습한다. 모델이 없는 강화학습에는 Q 함수를 근사하는 가치 기반의 방식(value-based)과 개체의 정책을 근사하는 정책 기반의 방식(policy-based), 그리고 두 방법을 결합한 액터-크리틱 방식(actor-critic)이 존재한다.

대표적인 가치 기반의 방식에는 Q-learning이 존재한다. 이는 상태-행동 쌍에 대한 가치를 평가하는 Q 함수를 통해 정책을 학습하는 방식으로 각 상태에서 가장 높은 Q 값을 출력하는 행동을 채택한다. 최근에는 이러한 방식에 인공 신경망을 결합하여 DQN(Deep Q-Network)[18]과 DDQN(Double DQN)[19] 알고리즘으로 발전하였다.

정책 기반의 방식은 직접적으로 개체의 정책을 근사하는 방식으로 상태에 대한 행동을 출력하는 정책 네트워크를 학습한다. 구체적으로, 정책의 성능을 평가할 수 있는 목적 함수를 설정하고 신경망 가중치에 대해 최적화한다. 대표적인 정책 기반의 방식의 알고리즘에는 DPG (Deterministic Policy Gradient)[20]와 REINFORCE[21]가 존재한다.

액터-크리틱 방식은 가치 기반의 방식과 정책 기반의 방식을 결합한 방식으로 정책을 결정하여 개체의 행동을 선택하는 액터 네트워크와 액터가 선택한 행동을 평가하는 크리틱 네트워크를 통해 정책을 학습한다. 이 방식은 정책 업데이트와 정책에 대한 평가가 서로 다른 네트워크에서 이뤄지기 때문에 안정적이라는 장점이 존재한다. 대표적인 액터-크리틱 방식의 알고리즘에는 DDPG[8]와 TD3[9] 등이 있다.

2.3 강화학습 기반 Weapon-target Assignment

WTA 문제가 현실적인 군사 작전 환경을 반영하기 위한 방향으로 발전됨에 따라 복잡하게 변화하는 전장 상황에서 신속하고 정확한 의사결정이 요구되어왔다. 하지만 기존의 방법들은 고려하는 환경이 복잡해지거나, 환경이 동적으로 변화하는 등의 불확실성이 존재하는 경우 최적화하기 어렵다는 한계가 존재하였다[22,23]. 이러한 상황에서 강화학습 기반의 방법이 WTA에 도입되기 시작하였다. 강화학습은 환경과의 상호작용을 통해 정책을 학습하는 방법으로, 주어진 상태에서 최적의 행동을 선택하여 장기적인 보상을 최대화하는 방법이다. 이러한 강화학습 방법은 근사된 정책을 학습한 후, 동적인 전장 환경에서 학습된 정책을 활용하여 개체가 최적의 의사결정을 내릴 수 있다는 장점이 있다. [7]에서는 강화학습 기반의 방법이 기존의 heuristic 및 최적화 기반의 방법에 비해 의사결정에 걸리는 계산 시간을 대폭 감소시킬 수 있음을 확인하였고, [24]에서는 목표에 적합한 보상 함수의 설계를 통해 적군 피해 최대화, 아군 무기 사용 비용 최소화 등의 의사결정이 가능한 최적의 정책을 학습할 수 있음을 보였다.

이에 본 연구에서는 동적인 환경의 WTA 문제를 설정하고, MDP 설계를 통해 최적의 의사결정 정책 학습이 가능한 심층 강화학습을 기반으로 문제를 해결하고자 한다.

Ⅲ. 심층 강화학습 기반 Weapon-target Assignment 시스템

3.1 전장 환경

본 연구에서는 아군 부대의 적군 부대에 대한 효율적인 무기 선택 의사결정 정책 학습을 위해 동적으로 변화하는 전장 상황을 고려한다. 구체적으로, 적군 부대는 에피소드 내 매 시점마다 일정 범위 μ 내에서 무작위로

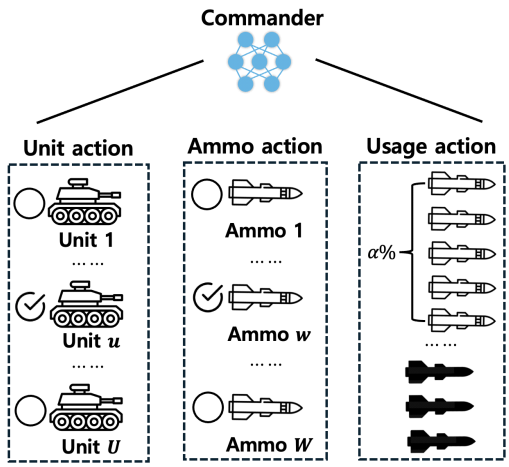


그림 1. 강화학습 기반 지휘결심 지원 시스템
Fig. 1. Reinforcement learning-based military decision support system

위치를 이동할 수 있다. 한 에피소드 내 위치가 고정되어 있는 의사결정 개체는 적군 부대의 위치나 종류, 상태에 따른 방어도 수준을 파악한 후 그림 1과 같이 U개의 아군 부대 $\boldsymbol{u} = \{1,2,\cdots,U\}$에서 W개의 무기 종류 $\boldsymbol{w} = \{1,2,\cdots,W\}$에 대해 적절한 탄약 발수를 선택하여 주어진 지휘관의 요망 효과를 달성하는 것을 목표로 한다.

3.2 Weapon-target Assignment 해결을 위한 Markov Decision Process 정의

대부분의 강화학습 문제는 MDP를 통해 모델링할 수 있다. MDP는 튜플 $\langle S,A,T,R,\gamma\rangle$로 정의되며, $s_{n,t}\in S$는 전장 환경의 상태 정보, $a_{n,t}\in A$는 개체의 의사결정 행동, $T(s_{n,t+1}|s_{n,t},a_{n,t})$는 상태 전이 확률(state transition probability), $R(s_{n,t},a_{n,t},s_{n,t+1})$는 보상 함수, $\gamma\in(0,1]$는 시간에 따른 감가율(discount factor)을 의미한다. 구체적으로, 학습 주체인 아군 부대는 개체로서 정의되며, 특정 전장 상태 $s_{n,t}$에서 무기 선택 행동 $a_{n,t}$를 수행하고 다음 상태 $s_{n,t+1}$에 도달하여 보상 $r_{n,t} = R(s_{n,t},a_{n,t},s_{n,t+1})$을 획득하게 된다. 이때, 개체는 누적되는 보상을 최대화하는 방식으로 의사결정 정책을 학습하게 된다.

3.2.1 상태 정보 (state)

전장 환경의 상태 정보 $s_{n,t}\in S$는 n번째 에피소드 내 t시점에서의 모든 정보를 의미하며, 다음과 같이 정의된다.

$$s_{n,t} = \left[m,k_n,b_n,\boldsymbol{c}^{\intercal},e_n,t,d_{n,t},h_{n,t},\boldsymbol{l_{n,t}}^{\intercal}\right]^{\top} \quad (1)$$

수식 (1)에서 m은 전장 환경의 크기를 나타내며, $k_n = \{k_{n,1},\cdots,k_{n,K}\}$는 n번째 에피소드에 대한 K개의 적군 부대 종류, $b_n = \{b_{n,1},\cdots,b_{n,B}\}$는 B개의 적군 상태에 따른 방어도 상수, $\boldsymbol{c} = [c_{1,1},\cdots,c_{1,W},\cdots,c_{U,1},\cdots,c_{U,W}]^{\top}$는 아군 부대의 종류와 무기 종류에 따라 일정 기간 가용 보급을 고려해 할당될 수 있는 탄약 보급률 벡터를 의미하며 $\mathbb{R}^{U\times W}$의 차원을 갖는다. e_n은 n번째 에피소드에 대한 지휘관의 요망 효과, t는 한 에피소드 내 현재의 시점을 의미하며, $d_{n,t}$는 n번째 에피소드의 t시점에서 아군 부대와 적군 부대와의 거리, $h_{n,t}$는 n번째 에피소드의 t시점 적군 부대의 체력, $\boldsymbol{l_{n,t}} = [l_{n,t,1,1},\cdots,l_{n,t,1,W},\cdots,l_{n,t,U,1},\cdots,l_{n,t,U,W}]^{\top}$는 n번째 에피소드의 t시점 아군 부대 종류와 무기 종류에 따른 잔여 탄수 벡터를 의미하고 $\mathbb{R}^{U\times W}$의 차원을 갖는다.

3.2.2 개체의 의사결정 행동 (action)

아군 부대의 n번째 에피소드 내 t시점 행동 $a_{n,t}\in A$는 다음과 같이 정의된다.

$$a_{n,t} = \left\{a_{n,t}^{unit},a_{n,t}^{ammo},a_{n,t}^{usage}\right\} \quad (2)$$

수식 (2)의 $a_{n,t}^{unit} = \left\{a_{n,t,1}^{unit},\cdots,a_{n,t,u}^{unit},\cdots,a_{n,t,U}^{unit}\right\}$은 부대 선택 행동으로 각각의 요소는 0 또는 1의 값을 갖는다. 구체적으로, $a_{n,t,u}^{unit} = 0$인 경우 아군 u번째 부대를 선택하지 않았음을 의미하고, $a_{n,t,u}^{unit} = 1$인 경우에는 선택했음을 의미한다. $a_{n,t}^{ammo} = \left\{a_{n,t,1}^{ammo},\cdots,a_{n,t,u}^{ammo},\cdots,a_{n,t,U}^{ammo}\right\}$은 무기 선택 행동으로 아군 u번째 부대에서 선택 가능한 W개의 무기 종류 중 하나를 선택하는 행동을 한다. 구체적으로, $a_{n,t,u}^{ammo}\in\{1,2,\cdots,W\}$의 값을 가지고 무기 종류는 모든 부대에서 동일하다고 가정한다. $a_{n,t}^{usage} = \left\{a_{n,t,1}^{usage},\cdots,a_{n,t,u}^{usage},\cdots,a_{n,t,U}^{usage}\right\}$는 선택한 부대와 무기에 대한 탄약 사용 비율을 결정하는 행동으로 $a_{n,t,u}^{usage}\in[1,100]$의 범위로 고려된다. 이때, 에피소드 n의 t시점 아군 u번째 부대의 w무기에 대한 실제 사용한 탄약 수는 탄약 보급률인 $c_{u,w}$를 고려하여 다음과 같이 정의된다.

$$a_{n,t,u}^{unit}\left(\alpha\times c_{u,w}\times a_{n,t,u}^{usage}\right) \quad (3)$$

수식 (3)에서 $\alpha\in(0,1]$는 에피소드 내 단위 시점당 사용 가능한 최대 탄약수를 조절하는 파라미터이다. 구체적으로, 개체는 t시점에서 최대 $\alpha\times c_{u,w}$ 만큼의 탄약을 사용할 수 있다. 또한 실제 사용한 탄약 수식을 기반으로 n번째 에피소드에서 $t+1$시점의 u부대, w무기에 대해 남은 탄약 수 $l_{n,t+1,u,w}$는 t시점 남은 탄약 수 $l_{n,t,u,w}$에서 사용한 탄약 수의 차이로 다음과 같이 정의된다.

$$l_{n,t+1,u,w} = l_{n,t,u,w} - a_{n,t,u}^{unit}\left(\alpha\times c_{u,w}\times a_{n,t,u}^{usage}\right) \quad (4)$$

수식 (4)에서 에피소드 종료시점 T_n까지 사용한 무기의 탄약 수의 총합이 해당 무기의 초기 잔존량인 $l_{n,0,u,w}$을 초과할 수 없으며, 특정 시점 실제 사용하고자 결정한 탄약 수가 해당 시점 잔존량보다 큰 경우 잔존량까지

만 선택이 가능하도록 제약한다. 따라서 최종 $a_{n,t,u}^{usage}$은 다음과 같이 정의된다.

$$a_{n,t,u}^{usage} = \min\left(\alpha \times c_{u,w} \times a_{n,t,u}^{usage}, l_{n,t,u,w}\right)$$
$$\alpha \times c_{u,w} \times \sum_{t=0}^{T_n} a_{n,t,u}^{usage} \le l_{n,0,u,w}$$

3.2.3 보상 함수 (reward)

아군 부대의 n번째 에피소드에서 t시점 보상 $r_{n,t}$는 현재 상태 $s_{n,t}$, 현재 행동 $a_{n,t}$, 다음 상태 $s_{n,t+1}$에 대한 함수 형태 $r_{n,t} = R(s_{n,t}, a_{n,t}, s_{n,t+1})$로 정의되며 보상항과 처벌항의 선형 결합으로 이뤄진다.

$$R(s_{n,t}, a_{n,t}, s_{n,t+1}) = \sum_{i=1}^{4} \eta_i R_{n,t,i} \quad (5)$$

수식 (5)에서 $\eta_{i\in\{1,2,3,4\}}$는 각 항에 대한 계수를 나타내고, $R_{n,t,i\in\{1,2,3,4\}}$는 보상항 또는 처벌항을 의미한다.

적군 부대 피해에 대한 보상항인 $R_{n,t,1}$은 현재 시점의 아군 부대의 적군 부대 공격에 대한 피해도가 클수록 높은 보상을 부여한다.

$$R_{n,t,1} = h_{n,t} - h_{n,t+1} \quad (6)$$

수식 (6)를 통해, 개체는 n번째 에피소드 t시점의 적군 부대의 체력 $h_{n,t}$과 아군 부대의 공격으로 인한 $t+1$시점 적군 부대의 체력 $h_{n,t+1}$의 차이가 클 수록 높은 보상을 획득한다. 이때, 아군 부대 공격에 의한 적군 부대의 체력은 다음과 같은 수식으로 정의된다.

$$h_{n,t+1} = h_{n,t} \prod_{u=1}^{U} \left(1 - p_{n,t,u}\right)^{a_{n,t,u}^{unit}\left(\alpha \times c_{u,w} \times a_{n,t,u}^{usage}\right)} \quad (7)$$

수식 (7)에서 $\prod_{u=1}^{U}\left(1-p_{n,t,u}\right)$는 적군 체력의 감소 비율로, 아군 각 부대 공격에 의한 적군 체력 잔존 비율의 누적 곱으로 고려된다. 이때, 피해량 $p_{n,t,u}$는 다음과 같다.

$$p_{n,t,u} = \mathcal{N}\left(\bar{p} \times b_n \times \delta\left(k_n, a_{n,t,u}^{ammo}\right) \times f(d_{n,t}), \sigma\right) \quad (8)$$

수식 (8)에서 $\bar{p}$는 피해도 가중치를 의미하고, $\delta\left(k_n, a_{n,t,u}^{ammo}\right)$는 적군 부대 종류와 아군 부대가 선택한 탄 종류에 따른 피해도 상수, σ는 피해량 정규분포에 대한 분산을 의미한다. 아군 부대는 적군 부대와의 거리에 대해 피해를 입힐 수 있는 사정거리가 존재하는데, 이에 $f(d_{n,t})$를 n번째 에피소드 내 t시점에서 아군 부대와 적군 부대와의 거리 $d_{n,t}$에 따른 피해도 감소 함수로 정의한다. 각각의 아군 부대는 부대별 최대 사정거리 $\tau_{u,max}$를 가지고 있으며, 이 범위를 초과하는 거리에 존재하는 적군에 대해서는 피해를 입힐 수 없다.

두 번째 항인 $R_{n,t,2}$는 무기의 탄약 사용 비용에 대한 항으로 무기의 사용한 탄약에 대해 페널티를 부여한다.

$$R_{n,t,2} = \sum_{u=1}^{U} a_{n,t,u}^{unit}\left(\alpha \times a_{n,t,u}^{usage}\right) \quad (9)$$

아군 부대는 수식 (9)을 통해 공격에 참여한 부대가 사용한 탄약에 대한 비용의 총 합을 페널티로 받게 된다.

군사 작전의 성공적인 완수를 위해서는 지휘관의 요망 효과에 맞춘 적절한 피해량을 달성하는 것이 필수적이다. 이에 본 연구에서는 $R_{n,t,3}$를 통해 적군 부대의 초기 체력 수준 $h_{n,0}$에서 요망 효과 e_n에 의한 요망하는 적군 체력 수준 $h_{e_n} = h_{n,0} - e_n$을 초과하는 경우 초과한 만큼 페널티를 부여하도록 설계하였다.

$$R_{n,t,3} = \begin{cases} 0, & h_{n,t+1} \ge h_{e_n} - \varepsilon \\ h_{n,t+1} - h_{e_n} + \varepsilon, & h_{n,t+1} < h_{e_n} - \varepsilon \end{cases} \quad (10)$$

수식 (10)에서 ε는 요망 체력 수준에 대해 허용되는 체력 마진으로써 적군의 체력이 요망 체력과의 차이가 ε 이내 인 경우 개체는 페널티를 받지 않는다. 아군 부대는 t시점 공격을 한 이후 $t+1$시점의 적군 부대의 체력 $h_{n,t+1}$이 요망 체력 h_{e_n}으로부터 ε이상을 초과하게 되면 그 차이만큼 페널티를 받게 된다.

지휘관의 요망 효과를 달성하지 못하는 경우는 작전 실패를 의미한다. 따라서 본 연구에서는 에피소드 보상인 $R_{n,t,4}$를 통해 에피소드 내 아군 부대가 적군 부대에게 가하는 총 피해량이 지휘관이 요망하는 체력 수준에 미치지 못하는 경우 페널티를 부여한다. 즉, 해당 항을 통해 요망 효과를 달성하지 못하고 에피소드의 최대 시점에 도달하여 종료되는 경우에 대해 페널티를 부여하게 되고, 수식은 다음과 같이 정의한다.

$$R_{n,t,4} = \begin{cases} -1, & h_{T_n} > h_{e_n} \\ 0, & otherwise \end{cases} \quad (11)$$

수식 (11)에서 h_{e_n}는 요망 체력을 의미하고, T_n은 에피소드의 총 길이로 정의되어 h_{T_n}은 에피소드 종료 시 적군 부대의 최종 체력을 의미한다.

Ⅳ. 실험 설정 및 분석

본 절에서는 심층 강화학습 기반으로 학습한 아군 부대 개체의 의사결정 정책을 분석하고 평가한다. 먼저 학습을 진행한 시뮬레이션 환경과 학습 알고리즘 설정에 대해 살펴본다. 이후 실험 결과에 대한 평가를 진행하고, exact 알고리즘과 heuristic 의사결정 방법과의 성능을 비교하고 분석한다.

아군 부대 개체의 의사결정 정책 학습은 군사 작전 환경을 구현한 시뮬레이터 상에서 환경과의 상호작용을 통해 획득하는 상태 정보, 의사결정 행동, 보상으로 이뤄진 데이터를 통해 이뤄진다. 시뮬레이션 상에서 개체는 매 에피소드마다 갱신되는 다양한 전장 상황을 경험하고 매 시점 환경과 상호작용하면서, 주어진 상태에서 최대의 보상을 받을 수 있는 행동을 선택하도록 정책을 점차적으로 개선한다. 학습은 총 3000번의 에피소드 동안 진행하며, 에피소드의 길이는 $T_n = 100t_s$로 설정된다. 이때, $1t_s = 1\text{second}$이다.

4.1 의사결정 정책 비교 알고리즘 및 평가 지표

본 연구에서는 제안한 MDP로 학습한 강화학습 기반의 의사결정 정책(RL-based)과 성능 비교를 위해 기존의 WTA 문제에서 다중 목표 최적화를 위한 heuristic 방법들과의 비교를 수행한다.

4.1.1 비교 알고리즘

1) RL-based: 본 연구에서는 심층 강화학습 기반의 의사결정 정책 학습에 다음과 같은 액터-크리틱 알고리즘을 고려한다.
   - DDPG[7]: 결정론적인 정책(deterministic policy)을 고려하는 알고리즘으로, 기존의 정책 기반 알고리즘인 DPG에 DQN 기반의 크리틱 네트워크를 적용함으로써, 연속적인 행동 공간(continuous action space)에서의 의사결정 정책을 학습한다.
   - TD3[8]: 두 개의 크리틱 네트워크 중 작은 Q값으로 액터 네트워크를 업데이트하는 double Q 기법을 활용함으로써 DDPG에서 발생하던 과대추정(overestimation) 문제를 완화한다. 또한 액터 네트워크의 업데이트를 크리틱 네트워크보다 지연시켜서 진행함 으로써 높은 학습 안정성을 제공할 수 있다.
2) Heuristic: 본 논문에서는 지휘관의 요망 효과 달성뿐만 아니라 자원의 효율적인 사용을 동시에 고려하는 것이 중요하기 때문에, 이러한 Multi-obective WTA를 해결할 수 있는 heuristic 알고리즘을 사용한다.
   - 다중 목표 유전 알고리즘(Multi-Objective Genetic Algorithm; MOGA)[14]: MOGA는 최적화 문제에서 다양한 해를 동시에 탐색할 수 있는 유전 알고리즘으로, 특히 복잡한 다중 목표 최적화 문제에서 유용하다. 이 알고리즘은 개체군 기반 탐색을 통해 전역 최적해를 찾을 수 있는 잠재력을 가지며, 다양한 해를 병렬로 평가하여 빠르게 수렴할 수 있다는 특징이 있다.
   - 다중 목표 인공 벌집 알고리즘(Multi-Objective Artificial Bee Colony; MOABC)[15]: MOABC는 기존 ABC 알고리즘의 확장으로, 다중 목표 최적화 문제를 처리하기 위해 변형된 최적화 알고리즘이다. 이 알고리즘은 탐색과 탐색 강화를 균형 있게 유지하며, 초기 해 공간을 광범위하게 탐색할 수 있다는 특징이 있다. 특히, 탐색 단계에서 다양한 초기 해를 고려하여 전역 최적해에 도달할 가능성을 높이고, 후속 탐색에서 탐색 강화 단계를 통해 해의 품질을 향상시킨다.

4.1.2 평가 지표

성능 비교를 위한 실험은 각 알고리즘별 5개의 랜덤 시드에 대해 각 시드별로 가능한 모든 적군 조합과 요망 효과 조합에 대해 10번씩 총 $N = 600$번의 에피소드를 진행하였다. 이때, 에피소드마다 환경과 상호작용하여 수집된 상태 정보, 의사결정 행동 정보, 보상 값을 포함하는 데이터를 통해 알고리즘별 성능 비교를 진행하였다. 다음과 같은 평가 지표들을 통해 성능을 측정하고 비교하였고, 각 평가 지표에서 ↑는 값이 클수록, ↓는 값이 작을수록 성능이 우수함을 의미한다.

- 달성률(↑): 적군의 최종 체력이 요망 체력의 허용마진 범위 내에 있을 때 해당 에피소드를 달성된 것으로 정의하며, 요망 체력에 미치지 못하거나 초과 피해를 입힌 경우 달성되지 않은 에피소드로 간주한다.

표 1. 알고리즘별 성능 비교
Table 1. Compare performance by algorithm

비교 알고리즘		달성률(%)	요망 체력 오차율(%)	무기 탄약 소모 비용	탄약 효율성
RL-based (Proposed)	DDPG [7]	91.27±6.71	3.72±0.86	**0.93±0.07**	0.69±0.03
	TD3 [8]	**91.77±3.82**	**3.67±0.73**	0.98±0.22	**0.71±0.21**
Heuristic	MOGA [14]	61.83±1.48	6.65±0.05	1.10±0.01	0.44±0.02
	MOABC [15]	66.97±0.46	5.74±0.02	1.07±0.03	0.57±0.01

$$f(n)=\begin{cases}1, & 0 \le h_{e_n} - h_{T_n} \le \varepsilon \\ 0, & otherwise\end{cases}$$

여기서, $f(n)$는 n번째 에피소드에 대한 달성 함수이며, 1은 요망 체력 달성 에피소드, 0은 미달성 에피소드를 의미한다. 이때 달성률은 전체 N개의 에피소드 중 달성된 에피소드의 비율로 다음과 같이 계산한다.

$$\text{달성률} = \frac{1}{N}\sum_{n=1}^{N} f(n) \times 100$$

- 요망 체력 오차율(↓): 요망 효과에 의한 요망 체력과 적군의 최종 체력과의 오차율로 다음과 같이 정의된다.

$$\text{요망체력오차율} = \frac{1}{N}\sum_{n=1}^{N} \frac{\left| h_{T_n} - h_{e_n} \right|}{h_{e_n}} \times 100$$

- 무기 탄약 사용 비용(↓): 무기 탄약 사용 비용은 보상 함수의 $R_{n,t,2}$로 정의되며, 이는 아군 부대가 선택한 무기에서 사용한 탄약의 양에 따른 비용을 나타낸다.

$$\text{무기 탄약 사용 비용} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} R_{n,t,2}$$

- 탄약 효율성(↑): 에피소드 달성시 무기의 탄약을 얼마나 효율적으로 소모했는지를 나타내는 지표이다. 에피소드 달성률과 달성시의 평균 무기 탄약 소모 비용을 통해 다음과 같이 정의된다.

$$\text{탄약 효율성} = \frac{\sum_{n=1}^{N} f(n)}{\sum_{n=1}^{N}\sum_{t=1}^{T_n} f(n) R_{n,t,2}}$$

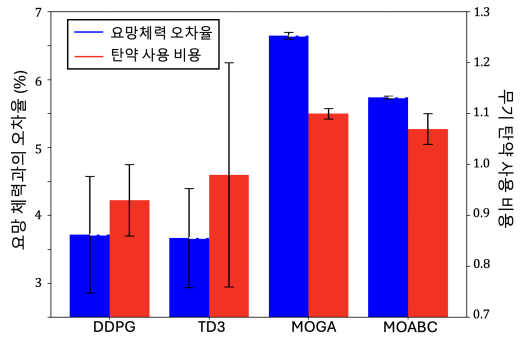


그림 2. 알고리즘별 요망 체력 오차율 및 탄약 사용 비용 평균
Fig. 2. Desired health point error rate and average cost of used ammo by algorithm

4.2 성능 분석 및 평가

본 절에서는 RL-based로 학습된 의사결정 정책과 heuristic 기반의 방법들과 비교하고 성능을 분석한다.

4.2.1 비교 알고리즘과의 성능 비교

표 1은 알고리즘별 요망 체력 오차율, 탄약 사용 비용, 달성률에 대해 에피소드 평균과 1차 표준편차를 나타낸 그래프이다. 해당 표를 통해 달성률 측면에서 RL-based 정책이 heuristic 의사결정 정책에 비해 27.12% 향상된 것을 확인할 수 있다. 또한 요망 체력 오차율을 2.5%, 무기 탄약 소모 비용은 11.98% 감소시킨 것을 확인할 수 있다. 특히, heuristic 방법은 낮은 요망 체력과의 오차율에 비해 달성률 또한 낮게 측정되었는데, 이는 적군의 최종 체력이 요망 체력에 근사하지만 지휘관이 요망하는 체력 마진 수준을 초과하거나 미치지 못하는 것을 의미한다. 반면 RL-based의 경우 낮은 오차율과 동시에 높은 달성률을 보였다. 또한 탄약 효율성[1] 측면에서도 RL-based 정책은 heuristic 정책에 비해 38.61% 높은 탄약 효율성을 가지는 것을 확인할 수 있다. 이를 통해 RL-based 정책이 heuristic 정책보다 탄약을 효율적으로 사용하여 요망 체력에 근사하게 피해를 입힘과 동시에 요구되는 체력 마진 정도에 맞춰

1) 해당 성능 수치는 0~1사이의 값으로 정규화하였다.

피해량을 조절할 수 있음을 보였다.

알고리즘별 요망 체력 오차율과 탄약 사용 비용에 대한 평균을 나타낸 그림 2을 통해 RL-based 와 heuristic의 차이를 한눈에 확인할 수 있다. 푸른색과 붉은색 박스는 모든 에피소드에 대한 평균 요망 체력 오차율과 평균 탄약 사용 비용을 나타내고, 박스 상단 수직선의 길이는 표준 편차를 의미한다. 해당 그림을 통해 RL-based 기반의 정책은 heuristic 기반의 정책에 비해 요망 체력과의 오차율과 무기 탄약 사용 비용 모두 더 낮게 측정된 것을 확인할 수 있다. 이러한 결과는 실험 환경이 다양한 전장 상황에서 연속적으로 변화하는 상태와 행동 공간의 높은 복잡도를 가지기 때문에 heuristic 방법이 최적화 과정에서 국소 최적해(local optimization)에 머물러 전역 최적화를 달성하기 어렵다는 한계를 보였기 때문이다[22,23]. 반면, RL-based 방법은 근사화된 정책을 통해 이러한 환경에서도 효과적으로 의사결정을 수행할 수 있음을 보인다.

4.2.2 강화학습 기반 의사결정 경향석 분석

본 절에서는 RL-based 의사결정 방법에 따른 무기선택 경향성을 비교한다. 그림 3는 DDPG, TD3 알고리즘별로 학습된 정책에 대해 실험을 진행하였을 때 적군 부대와의 거리에 따른 아군 부대 선택 히트맵이다. 그림에서 푸른색 세로줄은 각 부대별 최대 사정거리 $\tau_{u,max}$를 의미한다. 그림 3 (a)에서 DDPG로 학습된 정책은

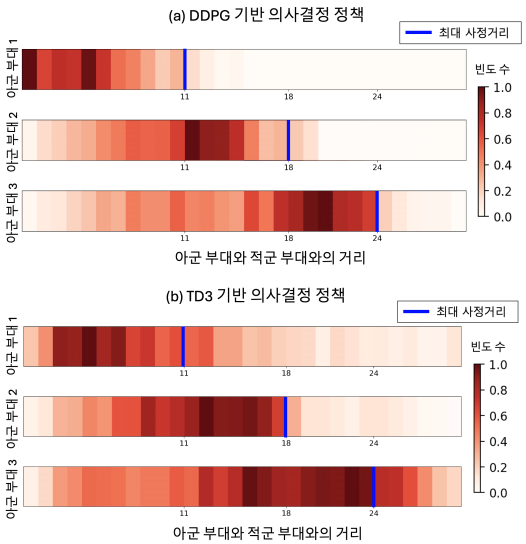


그림 3. 심층 강화학습 알고리즘별 적군 부대와의 거리에 따른 아군 부대 선택 행동 경향성
Fig. 3. Tendency of unit select actions according to distance from enemy units by deep reinforcement learning algorithm

아군 부대에 대해 적군 부대의 위치가 최대 사거리를 초과하는 경우 거의 선택하지 않도록 학습된 것을 확인할 수 있는 반면, 그림 3 (b)에서 TD3로 학습된 정책은 적군 부대의 위치가 최대 사거리를 초과하는 경우에도 종종 선택을 하는 경향을 보이는 것을 확인할 수 있다. 이는 RL-based 방법 중 DDPG 알고리즘이 TD3 알고리즘에 비해 심층 신경망이 아군 부대의 최대 사거리에 따른 차이를 더 잘 학습할 수 있음을 의미한다. 추가적으로 TD3의 최대 사거리 외 비효율적인 부대 선택 행동으로 인해 표 1의 무기 탄약 소모 비용도 DDPG에 비해 TD3가 소폭 높게 측정된 것을 확인할 수 있다. 이러한 행동 경향성 학습 결과는 TD3 알고리즘의 double Q 기법에 있다. 구체적으로, 2개의 critic 네트워크 중 작은 Q값을 target 값으로 설정하는 double Q 기법은 target 값이 불안정해지는 문제를 유발할 수 있다. 본 연구에서 고려하는 시나리오 환경에서 TD3는 이러한 문제로 인해 불안정한 critic 학습을 보이며, 결과적으로 critic 네트워크의 평가 결과를 통해 학습된 actor는 거리별 부대 선택 경향성을 성공적으로 학습하지 못하는 결과로 이어졌다. 반면, DDPG의 경우 단일 critic 네트워크를 사용함으로써, 비교적 안정된 critic 학습을 보였으며 거리에 따른 적절한 부대 선택 경향성을 학습하였다. DDPG와 TD3 알고리즘의 critic 학습 그래프는 Appendix A2에서 확인할 수 있다.

V. 결 론

본 연구에서는 심층 강화학습을 통해 동적으로 변화하는 전장 환경에서 효율적으로 아군 부대 및 무기 선택을 할 수 있는 의사결정 정책 학습을 위한 MDP를 제안하였다. DDPG, TD3 알고리즘으로 개체를 학습시켜 분석하고, 기존 WTA 문제 접근 방식인 heuristic 알고리즘과의 성능도 비교하고 분석하였다. 제안하는 MDP를 통해 학습된 RL-based 정책은 적군 위치에 대해 최대 사거리가 각각 다른 아군 부대를 선택하는 행동을 성공적으로 학습할 수 있음을 확인하였다. 또한 heuristic 알고리즘 기반 정책에 비해 요망 효과 달성률을 27.12%, 탄약 효율성을 38.61% 향상시키고, 요망 체력 오차율을 2.5%, 무기 사용 비용을 11.98% 절감할 수 있음을 확인하였다.

References

[1] J. Han, et al., “Conceptual design of infrastructure and framework for a futuristic

surveillance imagery fusion system," *J. KICS*, vol. 46, no. 9, pp. 1426-1439, 2021. (https://doi.org/10.7840/kics.2021.46.9.1426)

[2] S. Jin, et al., "A study on multiple reasoning technology for intelligent battlefield situational awareness," *J. KICS*, vol. 45, no. 6, pp. 1046-1055, 2020. (https://doi.org/10.7840/kics.2020.45.6.1046)

[3] IITP, *ICT R&D Technology Roadmap 2025, 2020*, Retrieved Oct. 8, 2024, from https://www.iitp.kr/kr/1/knowledge/openReference/view.it?ArticleIdx=5239&count=true.

[4] O. Altinoz, "Evolving model for synchronous weapon target assignment problem," *Int. Conf. INISTA*, 2021. (https://doi.org/10.1109/INISTA52262.2021.9548606)

[5] J. Kim, et al., "A study on the weapon–target assignment problem considering heading error," *Int J. Aeronautical and Space Sci.*, vol. 25, pp. 1105-1120, 2024. (https://doi.org/10.1007/s42405-024-00717-5)

[6] R. Gao, et al., "The weapon target assignment in adversarial environments," *Commun. in Comput. and Inf. Sci.*, vol. 2029, pp. 247-257, 2024. (https://doi.org/10.1007/978-981-97-0885-7_21)

[7] H. Na, et al., "Weapon-target assignment by reinforcement learning with pointer network," *J. Aerospace Inf. Syst.*, vol. 20, no. 1, pp. 53-59, 2023. (https://doi.org/10.2514/1.I011150)

[8] T. Lillicrap, et al., "Continuous control with deep reinforcement learning," *ICLR*, 2016.

[9] S. Fujimoto, et al., "Addressing function approximation error in actor-critic methods," *ICML*, 2018.

[10] D. Ahner and C. Parson, "Optimal multi-stage allocation of weapons to targets using adaptive dynamic programming," *Optimization Lett.*, vol. 9, pp. 1689-1701, 2015. (https://doi.org/10.1007/s11590-014-0823-x)

[11] Y. Lu and D. Chen, "A new exact algorithm for the weapon-target assignment problem," *Omega*, vol. 98, no. 102138, 2021. (https://doi.org/10.1016/j.omega.2019.102138)

[12] W. Li, et al., "Knowledge-guided evolutionary optimization for large-scale air defense resource allocation," *IEEE TAI*, 2024. (https://doi.org/10.1109/TAI.2024.3375263)

[13] X. Chang, et al., "Bi-objective multi-stage weapon target assignment problem with limited ammunition," *IEEE BigDIA*, 2023. (https://doi.org/10.1109/BigDIA60676.2023.10429060)

[14] C. Wang, et al., "Multi-objective optimization of weapon target assignment based on genetic algorithm," *Int. Conf. Computer, CITCE*, 2021. (https://doi.org/10.1109/CITCE54390.2021.00013)

[15] H. Xing and Q. Xing, "An air defense weapon target assignment method based on multi-objective artificial bee colony algorithm," *Computers, Materials & Continua*, vol. 76, no. 3, pp. 2685-2705, 2023. (https://doi.org/10.32604/cmc.2023.036223)

[16] C. Atkeson, et al., "A comparison of direct and model-based reinforcement learning," *IEEE ICRA*, 1997. (https://doi.org/10.1109/ROBOT.1997.606886)

[17] A. Nagabandi, et al., "Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning," *IEEE ICRA*, 2018. (https://doi.org/10.1109/ICRA.2018.8463189)

[18] V. Mnih, et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529-533, 2015. (https://doi.org/10.1038/nature14236)

[19] H. Hasselt, et al., "Deep reinforcement learning with double Q-learning," *AAAI*, 2016. (https://doi.org/10.1609/aaai.v30i1.10295)

[20] D. Silver, et al., "Deterministic policy gradient algorithms," *ICML*, 2014.

[21] R. Sutton, et al., "Reinforcement learning: An introduction," *MIT Press*, pp. 1-22, 2018.

[22] Y. Zhao, et al., "Multi-weapon multi-target assignment based on hybrid genetic algorithm in uncertain environment," *Int. J. Advanced Robotic Syst.*, vol. 17, no. 2, 2020.

(https://doi.org/10.1177/1729881420905922)
[23] X. Change, et al., “Adaptive large neighborhood search algorithm for multi-stage weapon target assignment problem,” *Computers & Industrial Eng.*, vol. 181, no. 109303, 2023. (https://doi.org/10.1016/j.cie.2023.109303)
[24] T. Li, et al., “An intelligent algorithm for solving weapon-target assignment problem: DDPG-DNPE algorithm,” *Computers, Materials & Continua,* vol. 76, no. 3, 2023. (https://doi.org/10.32604/cmc.2023.041253)

이 재 휘 (Jaehwi Lee)

2024년 2월 : 숭실대학교 전자정보공학부 IT융합전공 학사
2024년 3월~현재 : 숭실대학교 지능형반도체학과 석사과정
<관심분야> 인공지능, 강화학습, 지능형지휘결심, 자율주행
[ORCID:0009-0001-8014-6493]

엄 찬 인 (Chanin Eom)

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

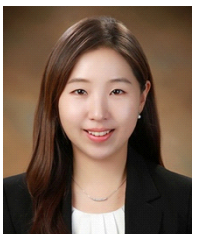
김 경 수 (Kyeongsoo Kim)

2022년 2월 : 한남대학교 정보통신공학과 학사
2023년 8월 : 한남대학교 정보통신공학과 석사
2023년 8월~현재 : 코난테크놀로지 비전AI연구소 연구원
<관심분야> 강화학습, 인공지능, 지능형지휘결심, 비전인식
[ORCID:0009-0005-6453-0240]

강 현 수 (Hyunsu Kang)

1996년 8월 : 전북대학교 전자계산학과 학사
1999년 2월 : 전북대학교 전산통계학과 석사
2000년 1월~2000년 12월 : 창신소프트 개발팀장
2001년 1월~2013년 9월 : 코난테크놀로지 책임연구원
2013년 10월~2016년 8월 : 시스트란 수석연구원
2016년 9월~현재 : 코난테크놀로지 비전AI연구소 이사
<관심분야> 강화학습, 지능형지휘결심, 인공지능, 객체인식
[ORCID:0009-0001-5184-0259]

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~2025년 2월 : 숭실대학교 전자정보공학부 IT융합 전공 조교수
2025년 3월~현재 : 숭실대학교 전자정보공학부 IT융합 전공 부교수
<관심분야> 강화학습, 지능형지휘결심, 자율주행, 모바일네트워크, 연합학습, 계산신경과학
[ORCID:0000-0002-8807-3719]

Appendix

표 A1. 표기법 정의
Table A1. Notation declaration

표기법	정의	표기법	정의
S	상태 공간	$s_{n,t}$	상태
A	행동 공간	$a_{n,t}$	행동
$T(s_{n,t+1}\|s_{n,t},a_{n,t})$	상태 전이 확률	$R(s_{n,t},a_{n,t},s_{n,t+1})$	보상 함수
γ	감가율	t	에피소드 내 시점
U	아군 부대의 수	$\boldsymbol{u}$	아군 부대 집합
W	아군 무기의 수	$\boldsymbol{w}$	아군 무기 집합
$a_{n,t}^{unit}$	t시점 아군 부대 선택 행동	$a_{n,t}^{ammo}$	t시점 아군 무기 선택 행동
$a_{n,t}^{usage}$	t시점 선택한 부대와 무기에 대한 탄약 사용 비율 결정 행동	m	전장 환경의 크기
K	적군 부대 종류의 수	k_n	n번째 에피소드의 적군 부대 종류
B	적군 상태의 수	b_n	n번째 에피소드의 적군 상태에 따른 방어도 상수
e_n	n번째 에피소드의 요망 효과	$\boldsymbol{c}$	아군 부대의 종류와 무기 종류에 따른 탄약 보급률 벡터
$d_{n,t}$	n번째 에피소드의 t시점에서 아군 부대와 적군 부대와의 거리	$\boldsymbol{l_{n,t}}$	n번째 에피소드 t시점에서 아군 부대와 무기 종류에 따른 잔여 탄수 벡터
$h_{n,t}$	n번째 에피소드의 t시점에서 적군 부대의 체력	h_{e_n}	n번째 에피소드에서 요망하는 적군 부대의 최종 체력
η	보상 함수의 각 항에 대한 계수	α	한 시점당 사용 가능한 최대 탄약 비율
$p_{n,t,u}$	n번째 에피소드의 t시점 아군 부대 u가 가한 피해량	T_n	n번째 에피소드의 에피소드 총 길이
h_{T_n}	n번째 에피소드 종료 시점 적군 부대 체력	$\bar{p}$	피해도 가중치
$f(d_{n,t})$	$d_{n,t}$에 따른 피해도 감소 함수	$\tau_{u,max}$	부대별 공격 가능 최대 사정거리
$\delta(k,a_{n,t,u}^{ammo})$	적군 부대 종류와 선택한 무기 종류에 따른 피해도 상수	ε	요망 체력 마진
$f(n)$	n번째 에피소드에 대한 달성 함수	N	전체 에피소드 수

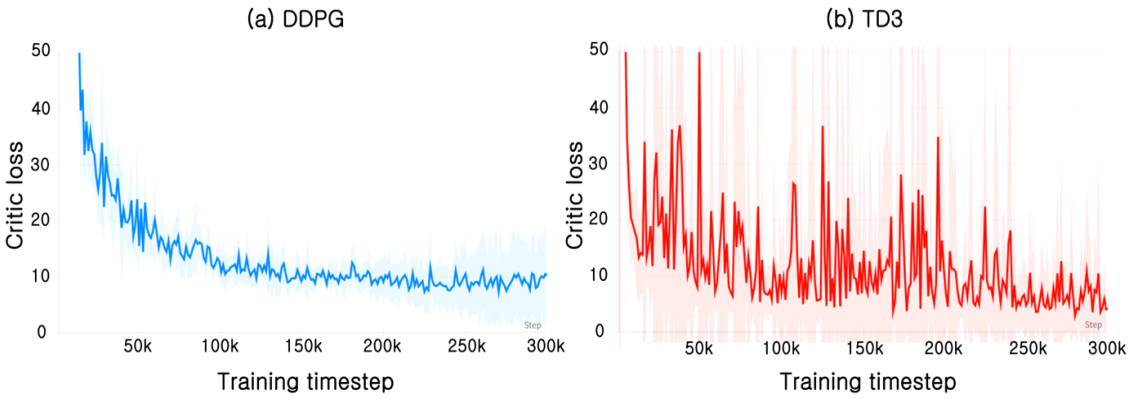


그림 A2. 강화학습 알고리즘별 critic 학습 그래프 (a) DDPG, (B) TD3
Fig. A2. Critic loss curve by RL algorithms