Journal of Korean Society of Industrial and Systems Engineering
https://doi.org/10.11627/jksie.2024.47.2.074

Vol. 47, No. 2 : 74-83, June 2024
ISSN : 2005-0461 / EISSN : 2287-7975

# Improving Dynamic Missile Defense Effectiveness Using Multi-Agent Deep Q-Network Model

Min Gook Kim* · Dong Wook Hong** · Bong Wan Choi* · Ji Hoon Kyung*†

*Department of Industrial Engineering, Hannam University
**Hanwha Systems

# 멀티에이전트 기반 Deep Q-Network 모델을 이용한 동적 미사일 방어효과 개선

김민국* · 홍동욱** · 최봉완* · 경지훈*†

*한남대학교 산업공학과
** 한화시스템(주)

The threat of North Korea's long-range firepower is recognized as a typical asymmetric threat, and South Korea is prioritizing the development of a Korean-style missile defense system to defend against it. To address this, previous research modeled North Korean long-range artillery attacks as a Markov Decision Process (MDP) and used Approximate Dynamic Programming as an algorithm for missile defense, but due to its limitations, there is an intention to apply deep reinforcement learning techniques that incorporate deep learning. In this paper, we aim to develop a missile defense system algorithm by applying a modified DQN with multi-agent-based deep reinforcement learning techniques. Through this, we have researched to ensure an efficient missile defense system can be implemented considering the style of attacks in recent wars, such as how effectively it can respond to enemy missile attacks, and have proven that the results learned through deep reinforcement learning show superior outcomes.

Keywords : Deep Reinforcement Learning, ADP, DQN, Multi-Agent, Missile Defense System

## 1. 서 론

2023년 10월 하마스가 이스라엘을 기습공격하였다. 언론엔 '하마스식 기습공격'이라 표현하며 북한이 대량의 장거리 화력으로 우리나라를 기습공격할 가능성이 있는지, 우리는 대비되어 있는지가 이슈가 되고 있다[5, 30]. 물론, 우리나라 역시 북한의 장거리 화력을 활용한 기습공격에 대비하고 있다. 단시간에 동시 다발적으로 우리나라의 주요시설을 공격할 수 있는 북한의 신형 방사포에 대한 위협을 인지하고, 국가의 중요 시설뿐만 아니라 군사시설 등을 보호하기 위해 북한의 장거리 화력에 대한 미사일방어체계 구축을 추진하고 있다[29].

미사일방어체계는 보호하기 위한 주요시설 주변에 위치하여 둠 형태의 방공망을 구축하고, 장거리 화력 공격에 대해 보유한 요격유도탄을 발사하여 장거리 화력의 탄두나 미사일을 요격하는 형태로 추진되고 있다[24]. 따라서 장거리 화력을 요격하기 위해 탄두나 미사일 표적에 요격유도탄인 무기를 할당하는 무기 표적 할당(WTA, Weapon Target Assignment) 문제가 중요할 수밖에 없고 이는 미사일방어체계의 교전 알고리즘과 연계된다.

최근 무기 표적 할당 문제에서는 최적해를 찾기 위해

Received 8 March 2024; Finally Revised 2 June 2024;
Accepted 3 June 2024
† Corresponding Author : kjh@hnu.kr

선형화[6], 유전자 알고리즘[9], 개미군집 알고리즘[12] 등 다양한 휴리스틱 알고리즘을 활용한 연구가 활발히 진행되고 있다. Naeem et al.[17]은 동적 무기 표적 할당 문제를 실시간 스케줄링 문제로 접근해 안정된 결혼 알고리즘을 변형하여 사용하였고, Cha et al.[2]는 동적계획법과 분기한정 알고리즘을 사용해 2단계 휴리스틱 알고리즘을 제안하였다. 하지만 이러한 접근 방식들은 불확실한 교전상황에 적합하지 않다는 한계가 있다. Jung et al.[7]은 휴리스틱 기반의 Rolling-Horizon 스케줄링 알고리즘을 제안해 기존의 SLS 전략보다 우수한 결과를 도출하였다.

Bertsekas et al.[1]은 강화학습 기반 알고리즘을 동적 무기 표적 할당 문제에 적용한 최초의 연구로, Neuro-Dynamic Programming을 통해 최적 함수를 근사화하였다. 이후 여러 연구들이 ADP 알고리즘을 활용해 미사일 방어 시나리오에서 우수한 성과를 보였으며, Davis et al.[3]과 Summers et al.[26]은 각각 현대전의 기술 발달을 반영한 실시간 자산 가치 평가와 다양한 시나리오 적용에서 ADP 알고리즘의 효율성을 입증하였다. Im et al.[4]은 탄도탄 궤적 예측을 기반으로 한 규칙기반 할당 방법론을, Shin et al.[25]은 멀티 에이전트 강화학습 모델을 제시하였으며, Lee et al.[10]은 장사정포 방어 연구에서 강화학습 알고리즘의 우수성을 확인하였다.

하지만 강화학습 알고리즘을 적용한 연구에서 ADP 알고리즘의 효율성을 입증하였지만, 강화학습의 한계점을 완전히 극복하지는 못하였다. 고려해야하는 변수가 늘어날수록 계산이 복잡해지는 것은 동일하며, 무엇보다도 강화학습으로 무기 표적 할당 문제를 해결하기 위해서는 환경에 대한 모든 값을 알아야 하는 한계점이 있다[27].

이를 해결하는 방법을 딥러닝과 강화학습을 활용한 심층강화학습에서 찾았다[14]. 심층강화학습은 기존의 강화학습과 달리 환경에 대한 모든 값을 알아야 할 필요가 없으며, 모든 변수를 계산하는 것이 아니기 때문에 대규모의 상태 공간에서도 효율적으로 작동할 수 있다. 이를 통해 복잡한 환경에서도 효과적인 정책을 학습할 수 있으며, 특히 실환경과 유사한 고차원 데이터에서도 뛰어난 성능을 발휘한다.

본 연구에서는 북한의 미사일 공격을 가정하고 이를 심층강화학습을 통해 방어할 수 있는 모델에 관한 연구를 진행하려 한다. 강화학습을 딥러닝으로 구현한 심층강화학습과 사전 연구된 요격통제 및 동시 교전의 알고리즘[8, 10]에 착안하여 심층강화학습 기법의 적용과 검증을 아래와 같이 3단계로 추진하겠다.

첫째, 심층강화학습 적용을 위해 북한의 미사일 위협을 MDP로 모형화하여 검증할 수 있는 모델을 구현하겠다. 둘째, MDP 강화학습 모델을 딥러닝 기법이 적용된 심층강화학습 기법에 적용할 수 있는 방안을 연구하여 멀티에이전트 기반의 수정된 DQN 심층강화학습 기법을 적용하였다. 셋째, 미사일 공격에 대한 교전상황을 반영한 시나리오를 통해 시뮬레이션을 수행하여 본 연구에서 적용한 심층강화학습 기반의 교전알고리즘의 우수함을 확인하였다.

## 2. 배경이론

북한의 미사일 위협을 순차적행동결정문제(MDP)로 모형화한 후 이 문제를 해결하기 위해 다양한 방법론을 적용할 수 있다. 전통적인 방법으로는 동적 프로그래밍(Dynamic Programing), 몬테카를로 방법(Monte Carlo Methods) 등이 있고 최근에는 강화학습을 활용하는 추세이다[18]. 본 연구에서는 딥러닝 기법이 적용된 심층강화학습 방법론을 적용하였다.

이는 심층강화학습 이전의 방법론 들은 대체로 계산의 효율성과 문제의 복잡성에 대한 한계가 있어 다양한 환경요소가 고려되어야 하는 미사일 방어체계 연구에도 한계가 있기 때문이다. 반면, 딥러닝을 활용한 심층강화학습을 통해 연구할 경우 상태공간이 크거나 연속적인 문제를 해결하기 용이하고 경험과 학습을 통해 문제를 해결하는 이점이 있다[10, 19]. 심층강화학습에 대해 더 알아보고 미사일 방어체계와 접목하는 방안을 알아본다.

### 2.1 심층강화학습 연구

심층강화학습(Deep Reinforcement Learning, DRL)은 강화학습(Reinforcement Learning)과 딥러닝(Deep-Learning)을 결합한 인공지능 기계학습의 한 분야로 최근 활발하게 연구되는 분야 중 하나이다. 심층강화학습이 최근 부각되는 이유는 현실 세계처럼 복잡한 환경에서 연구를 진행할 경우 기존 동적 프로그래밍이나 강화학습에 비해 많은 장점을 가지고 있기 때문이다.

MDP 모형을 이용한 대표적인 방법론인 동적 프로그래밍(Dynamic Programing) 등은 Full Knowledge 기반의 Model Based 기법이라 환경이 복잡함에 따라 계산 복잡도가 증가하고 계산량이 많아지는 차원의 저주라고 불리는 문제가 있다. 기본적으로 Full Knowledge 환경이라는 제한도 있다[20]. 이를 개선하기 위해 ADP를 적용하지만 환경이 복잡함에 따라 계산량이 증가하는 것은 동일하다[19].

반면에 심층강화학습은 이러한 문제가 해결된다. 샘플 기반의 학습으로 Full Knowledge 대신 경험-학습에 의존한다. 차원의 저주는 주요 특징만 학습하고 계산을 통해 근사치를 찾아감으로써 해소한다. 심층강화학습의 딥러닝 기법이 여기에 사용된다. 모든 상태값을 계산하여 테이블화하는 대신에 딥러닝 기법을 적용하여 신경망을 통해 함

수를 예측하고 함수식을 통해 근사치를 계산한다. 동적프로그래밍 등의 방법보다 계산량이 적고 더 복잡한 현실환경을 반영하여 시뮬레이션하기 용이한 이유가 여기에 있다. 우리가 연구할 미사일 방어체계에 관한 연구에 심층강화학습을 적용한 이유이다[13].

대표적인 심층강화학습 기법으로는 DQN(Deep Q-Network)이 있다. DQN은 Q-러닝을 딥러닝과 결합하여 고차원의 입력 데이터를 처리할 수 있는 강력한 기법이다[15]. Q-러닝은 각 상태에서 가능한 행동들의 가치를 추정하여 최적의 정책을 학습하는 심층강화학습 알고리즘이다. DQN은 컨볼루션 신경망을 사용하여, 각 상태에서 가능한 모든 행동의 예상 가치를 학습한다. 이 방법은 게임 환경과 같이 복잡한 상태 공간을 효과적으로 다룰 수 있도록 설계되었다. DQN은 기존 Q-러닝의 제한 사항들을 극복하기 위해 경험 재생(Experience Replay)과 타겟 네트워크(Target Network)와 같은 기술을 도입하였다[11]. 경험 재생은 에이전트가 얻은 경험을 메모리에 저장하고 무작위로 샘플링하여 학습함으로써 데이터의 상관성을 줄이고 학습 효율성을 높일 수 있다. 타겟 네트워크는 일정 간격으로만 업데이트되어 학습이 더 안정적이고 수렴 속도가 빨라진다.

이외에도 정책 기반의 심층강화학습 방법의 하나인 TRPO(Trust Region Policy Optimization)는 에이전트의 행동 정책을 직접적으로 최적화한다. PPO(Proximal Policy Optimization)와 같은 기법은 TRPO의 개념을 더욱 단순화하고 계산 효율성을 개선한 것으로 널리 사용되고 있다[22, 23]. 이러한 심층강화학습 기법들은 각각의 장단점을 가지고 있으며, 특정 문제에 가장 적합한 방법을 선택하는 것이 중요하다. 예를 들어, 높은 차원의 입력 데이터를 다루어야 하는 경우 DQN이 적합할 수 있으며, 연속적인 행동 공간을 가진 문제에는 TRPO나 PPO가 더 효과적일 수 있다. 이러한 기법들의 적절한 적용을 통해 무기체계 알고리즘 연구나 로봇 알고리즘 연구와 같은 복잡한 문제를 해결하는 데 큰 진전을 이룰 수 있다.

### 2.2 미사일 방어체계와 심층강화학습의 접목 이유

미사일 방어체계의 복잡성과 동적인 환경에 대응하기 위해 심층강화학습의 적용은 Exploration and Exploitation 접근법을 제공한다. 이 방법은 전통적인 방어 전략과 달리, 복잡한 상황에서 최적의 결정을 도출하는 데 도움을 줄 수 있다. 심층강화학습은 강화학습의 원리에 딥러닝의 강력한 데이터 처리 능력을 결합하여, 대규모의 데이터와 복잡한 패턴을 처리할 수 있다. 이러한 인공지능 기법의 적용은 미사일 방어체계를 효율적으로 운용하고, 다양한 공격 시나리오에 대응하는 최적의 전략을 개발하는 데 적합하다.

심층강화학습의 핵심은 에이전트가 환경과 상호작용을 통해 학습하고, 보상을 최대화하는 방향으로 행동을 조정한다는 점이다. 이러한 접근 방식은 미사일 방어체계에서 적의 공격 패턴을 인식하고, 최적의 요격 전략을 학습하는 데 효과적이다. 심층강화학습의 에이전트는 구현된 환경에서 다양한 시나리오와 위협 수준에 대한 데이터를 경험과 학습을 통해 이를 끌어낼 수 있다.

특히 심층강화학습은 다양한 유형의 공격에 대응하는 능력을 개선한다. 전통적인 방법론들, 예를 들어 동적 프로그래밍이나 몬테카를로 방법은 계산을 통해 전체 상황을 인식하는 특징 때문에 특정 상황에 맞춘 솔루션을 제공하는 데 초점을 맞추었다. 반면, 심층강화학습은 이러한 제한을 극복하고, 동적으로 변화하는 다양한 공격 유형과 환경에 대응할 수 있는 유연한 전략을 개발한다.

이러한 심층강화학습의 적용은 미사일 방어체계를 더 지능적이고 적응력 있는 시스템으로 변모시킬 수 있다. 연속적인 학습과 자가 적응을 통해, 시스템은 지속해서 발전하며 새로운 위협에 신속하게 대응할 수 있다. 이는 미사일 방어체계의 자산을 보다 효율적으로 운용하면서도, 높은 수준의 보호를 보장할 수 있도록 한다.

본 연구는 이러한 심층강화학습 기법을 미사일 방어체계에 접목하여, 그 가능성과 효과를 탐구하고자 한다. 이를 통해 더욱 발전된 방어 전략과 시스템을 개발하고, 더욱 안전한 환경을 구축하는 데 기여할 것으로 기대된다.

## 3. 방법론

### 3.1 문제정의

북한 미사일 방어체계에 대한 논의를 진행하기 위해 유사한 연구[8, 10]와 ADP[20] 및 심층강화학습[27] 관련 서적에서 사용한 기호를 참고하여 아래와 같이 문제를 정의한다.

미사일방어체계의 문제는 순차적으로 행동(적 미사일 요격)을 결정해야 하는 문제이므로 순차적 행동에 대한 정의가 필요하다. 순차적 행동에 대해서는 T = {1, 2, ⋯ ,$T$}, $T \rightarrow \infty$에 걸쳐서 미사일을 요격하는 연속적인 행동들을 결정해야 한다. T는 의사결정 시점(epoch)의 집합이고, $t$를 임의의 의사결정 시점이라 한다.

의사결정 시점은 적 미사일 공격으로 정해지며, 적 미사일 공격은 외생(exogenous)적 요인으로 강요된다. 적 미사일 공격은 M으로 표현하며, $t$ 시점에서 공격목표 $i$를 향해 발사된 공격 벡터는 $M_{ti} = (M_{ti})_{i \in A} = (M_{t1}, M_{t2}, ..., M_{t|A|})$로 나타난다.

미사일방어체계가 보호해야 하는 자산을 A라고 하며 $t$

시점에서 자산 $i$의 상태는 $A_t=(A_{ti})_{i\in A}=(A_{t1}, A_{t2}, ..., A_{t|A|})$로 표현된다. 자산 $A_{ti}=1$이면 피해가 없는 상태이며 $A_{ti}=0$ 이면 자신이 피해를 입어 완파된 상태를 나타낸다. 여기서 자산 $i$는 적 미사일 공격 목표 $i$가 된다.

미사일방어체계는 자산을 보호하기 위한 방어포대를 보유하고 있으며, 방어포대에서 보유한 요격유도탄의 재고를 R이라 하고 $t$시점에서 재고를 $R_t=(R_{ti})_{i\in A}=(R_{t1}, R_{t2}, ..., R_{t|A|})$로 나타낸다.

따라서 본 논문의 문제는 적 미사일 공격 M에 대해 자산 A를 방어하기 위해서 미사일방어체계가 R만큼의 요격유도탄을 활용하여 방어하는 문제로 정의할 수 있다. 이는 적 미사일 공격에 요격유도탄을 할당하는 동적 무기 표적 할당 문제(DWTA, Dynamic Weapon Target Assignment)로 정의할 수 있으며, 이는 적의 다량의 미사일 공격에 대해 의사결정 시점마다 순차적으로 요격에 대한 행동을 결정하는 문제로 순차적행동결정 모형인 MDP로 정의할 수 있다.

### 3.2 MDP 모형

MDP(Markov Decision Process) 모형은 5가지 요소로 정의되며, 각 의사결정 시점에서의 상태(state), 행동(action), 전이확률(transition probability), 보상(reward), 할인인자(discount factor)로 정의된다.

본 논문에서 상태는 앞서 문제를 정의한 대로 $S_t=(A_t, R_t, M_t)\in S$로 나타내며, 여기에서 S는 적 미사일 공격과 공격에 따른 자산, 및 요격유도탄의 모든 가능한 상태의 전체 집합이다.

행동은 $t$ 시점에서 요격유도탄 자산을 방어하는 방어포대 $i$에서 적 미사일 $j$에 대해 발사된 요격유도탄의 수이며 이를 $x_{tij}$라고 표현한다. 행동은 요격유도탄의 재고수에 제한을 받고 동시에 발사하는 수량도 제한이 있다.

전이확률은 일반적으로 $P(S_{t+1}|S_t, x_t)$로 표현되며 상태 $S_t$에서 행동 $x_t$를 취한 후 $S_{t+1}$로 도달할 확률을 의미한다. 우리는 전이확률을 상태 전이함수로 표현하면 $S_{t+1}=M(S_t, x_t, W_{t+1})$로 정의하며 여기에서 $W_{t+1}=(A_{t+1}, M_{t+1})$이며 이는 $t+1$시점에서 확률변수로 알게 되는 자산의 가치와 적 공격에 대한 정보를 나타낸다. $A_{t+1}$은 $t$시점에서 적 미사일 공격 $M_t$와 요격유도탄의 행동 xt에 의해 결정되는 확률변수이며 $M_{t+1}$은 $t+1$시점에서 알게 되는 확률변수이다.

보상에 대해 표현하기 전에 먼저 자산 가치의 손실을 표현하는 식에 대해 먼저 알아보도록 하겠다. $t$시점에서 요격유도탄의 행동이 결정되어도 요격유도탄이 미사일을 피격했는지는 확정되지 않고, 확률적으로 표현된다.

요격유도탄이 미사일을 100% 요격한다고 할 수 없고 당시의 환경조건, 적 미사일의 상태, 요격유도탄의 상태 등 다양한 조건에 따라 요격유도탄이 적 미사일을 요격했는지 알 수 있기 때문에 본 논문에서는 이를 확률적으로 표현한다. 따라서 확률적으로 표현되는 자산의 가치 손실(Cost) 식은 식 (1)과 같이 표현된다.

$$C(S_t, x_t, A_{t+1,i})=\sum_{i\in A} v_i(A_{ti}-A_{t+1,i}) \quad (1)$$

여기에서 $v_i$는 자산 $i$의 가치다. 이 수식은 상태 $S_t$에서 행동 $x_t$를 취했을 때, 다음 시점 $t+1$에서 자산가치 $A_{t+1}$로 변환에 따른 비용을 나타낸다. 여기서 각 $A_{ti}$는 시점 $t$에서의 자산 $i$의 가치를 나타내고, $v_i$는 해당 자산 가치 변화에 대한 가중치이다. 이를 기댓값을 활용하여 조건부 기대 비용함수로 상태 $S_t$와 행동 $x_t$가 주어졌을 때, 모든 가능한 자산 $A_{t+1}$에 대한 기대비용을 계산한다.

$$C(S_t, x_t)=E\{\sum_{i\in A} v_i(A_{ti}-A_{t+1,i}) \mid S_t, x_t\} \quad (20$$

MDP 모형에서 보상함수는 비용함수를 최소화할 수 있는 행동을 보상함수로 한다. 상태별로 결정되는 행동을 정책이라 부르고 $X_t^{\pi}(St)$를 요격유도탄을 결정하는 함수(즉, 정책)라 하면 목적함수는 식 (3)과 같이 표현할 수 있고 이를 통해 보상을 계산할 수 있다.

$$\min_{\pi\in\Pi} E^{\pi}\{E^{T}\{\sum_{t=1}^{T} C(S_t, X_t^{\pi}(S_t))\}\} \quad (3)$$

할인인자는 미래의 보상을 현재 가치로 할인하여 계산하는 데 사용된다. 이 인자는 미래의 보상이 현재 시점에 미치는 가치를 결정하는 데 중요한 역할을 한다. 할인인자는 0과 1 사이의 값으로 설정되며, 이 값에 따라 에이전트의 의사결정이 단기적 보상에 중점을 둘지, 아니면 장기적 보상을 고려할지 결정된다. 본 논문에서는 통상적인 임의의 수 0.8를 선택했다.

### 3.3 심층강화학습을 통한 시뮬레이션

앞서 설명한 MDP 모형을 벨만 방정식을 이용하여 동적 프로그래밍(Dynamic Programing)을 통해 정확한 최적해를 구할 수 있지만 실제 문제를 MDP로 모형화하여 최적해를 구하기엔 제약사항이 따른다.

첫째 문제는 MDP로 모형화하면 상태의 크기가 너무 커서 계산하기 어렵다는 문제이다. 두 번째 문제는 MDP 모형에서 어떤 상태가치를 계산하기 위해서는 그 상태로부

터 이후 모든 시간에 걸쳐 직간접적으로 전이될 수 있는 모든 상태의 가치를 알고 있어야 하는 Model Based 모형이라는 점이다[8, 10, 20, 27] 물론 이를 해결하기 위해서 최적해의 근사치를 계산하는 근사적 동적 프로그래밍(Approximate Dynamic Programming, ADP)을 적용할 수 있다. 하지만, ADP에도 한계는 있다. 근사치의 오류가 있을 수 있고 근삿값을 계산하기 위한 함수 선택의 어려움이 있다. 또한 여전히 대규모 문제의 경우 계산량이 많고 복잡할 수 있는 한계가 있다[21].

반면 심층강화학습을 이용할 경우 경험에 따른 학습을 통해 새로운 전략을 시도하고 최적의 해를 탐색할 수 있다. 환경과의 상호작용을 통해 학습하므로 Model Based처럼 모든 상태값을 알아야 할 필요가 없다. 더욱이 최근에는 딥러닝과 같은 최신기술을 통해 복잡한 문제에도 효과적인 성능을 보이는 심층강화학습을 적용할 수 있다. 본 논문에서는 심층강화학습을 적용하여 미사일방어체계의 DWTA 문제를 해결하려 한다. 심층강화학습의 다양한 방법 중 DQN 기법을 적용하여 미사일방어체계의 요격효과 개선을 입증하려 한다.

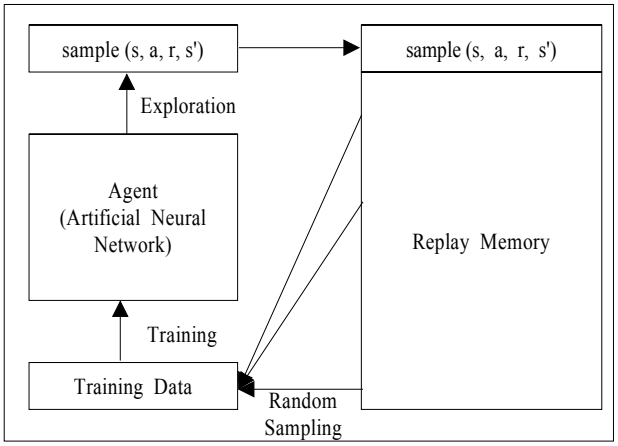
MDP 모형을 이용하여 상태(S), 보상(R), 할인인자($r$)는 동일하게 적용하며 행동(A)도 동일하나 식 (4)와 같이 정의한다.

$$x_{S_t}=\{x_t : \sum_{j\in M_{ti}^A} x_{tij} \le \min(R_{ti}, x_i^{\max}),\ \forall i \in A\} \quad (4)$$

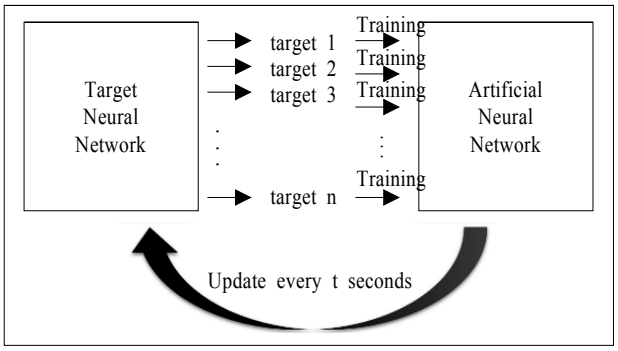
전이확률(P)은 심층강화학습에서 학습을 통해 근사치로 수렴하게 된다.

DQN을 개선해서 사용하기 위해 DQN의 대표적인 방법 2가지를 적용한다. 하나는 경험 리플레이(Experience Reply)이다[16]. 경험 리플레이는 에이전트가 환경에서 탐험하며 얻는 샘플 상태를 메모리에 저장한다는 것이다. 샘플을 저장하는 메모리를 리플레이 메모리(Replay Memory)라고 한다[11]. 에이전트가 학습할 때 리플레이 메모리에서 여러 개의 샘플을 무작위로 뽑아서 뽑은 샘플에 대해 인공신경망을 업데이트한다. 이 과정을 매 타임 스텝마다 반복한다. 이 과정은 <Figure 1>과 같다.

다른 하나는 타깃신경망(Target network)이다. DQN은 심층강화학습의 한 형태로, 행동가치함수(Q함수)를 근사화하기 위해 신경망을 사용한다. 이때, 예측값을 이용하여 또 다른 값을 예측하다 보니 업데이트의 목표가 되는 정답이 계속 변하는 문제가 있다. 따라서 일정 시간 동안 정답을 유지하기 위해 사용하는 것이 타깃신경망이다. 타깃신경망을 따로 만들어서 타깃신경망에서 정답에 해당하는 값을 구하고 이 정답을 통해 다른 인공신경망을 학습시키다가 일정한 시간 간격마다 타깃신경망을 인공신경망을 업데이트해 준다. 이 과정은 <Figure 2>와 같다.

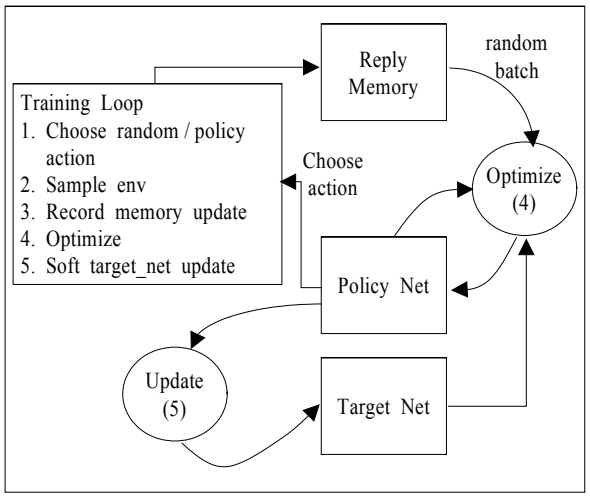



<Figure 1> Training Artificial Neural Networks with Replay Memory [11]




<Figure 2> Learning through Target Neural Networks and Artificial Neural Networks [16]

DQN을 적용하기 위한 전체흐름도는 <Figure 3>과 같고 이를 적용하기 위한 일반적인 DQN 수도코드는 <Figure 4>와 같다.




<Figure 3> Overall Flowchart of DQN [28]

**Algorithm 1** Deep Q-learning with Experience Replay

Initialize replay memory D to capacity N
Initialize action-value function Q with random weights
for episode = 1, M do
 Initialize sequence $s_1 = \{x_1\}$ and preprocessed sequenced $\phi_1 = \phi(s_1)$
 for t = 1, T do
 With probability $\varepsilon$ select a random action $a_t$
 otherwise select $a_t = \max_a Q^*(\phi(s_t), a; \Theta)$
 Execute action $a_t$ in emulator and observe reward $r_t$
 and image $x_{t+1}$
 Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$
 Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in D
 Sample random minibatch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$
 from D
 Set $y_j = \{$
 $r_j$ for terminal $\phi_{j+1}$
 $r_j + \gamma \max_{a'} Q(\phi_{j+1}, a'; \Theta)$ for non-terminal $\phi_{j+1}$
 $\}$
 Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \Theta))^2$
 end for
end for

<Figure 4> DQN Algorithm for Deep Reinforcement Learning Application [15]

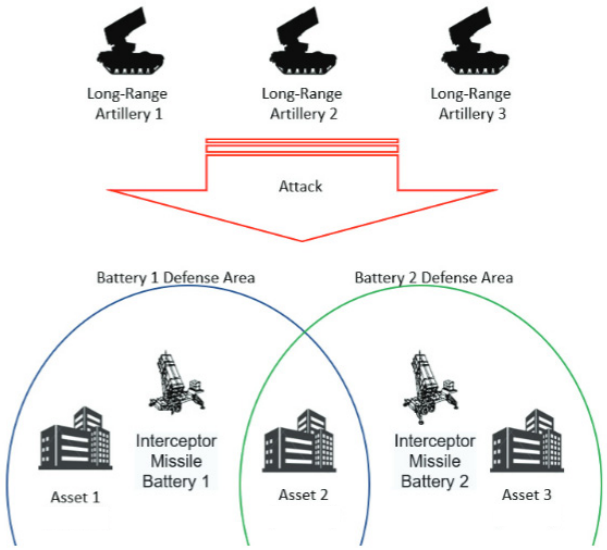
## 4. 시뮬레이션 및 결과 분석

### 4.1 시뮬레이션 시나리오

시뮬레이션을 위한 적 미사일 포대와 아 방어자산 및 방어포대는 <Figure 5>에 요약되어 있다. 적은 3개의 미사일 포대가 있으며 아군의 방어자산 3곳을 목표로 공격한다. 3곳을 각각 공격할 수도 있고 한곳에 집중하여 공격할 수도 있다. 공격벡터(M)은 무작위로 선정된다.

요격유도탄 포대는 2개가 있으며 1개의 포대가 2개의 방어자산을 방어할 수 있고 1개의 자산은 공유하여 방어한다. 적 미사일 공격을 요격 시에는 확률에 의해 요격하며, 요격에 실패하면 적 미사일 1발마다 5%씩 가치에서 차감하는 것으로 가정한다. 즉, 1발 피격시 0.95. 2발 피격시 약 0.9, 3발 피격시 약 0.86 순으로 낮아진다. 요격유도탄의 요격 성공률은 90%로 가정했다.

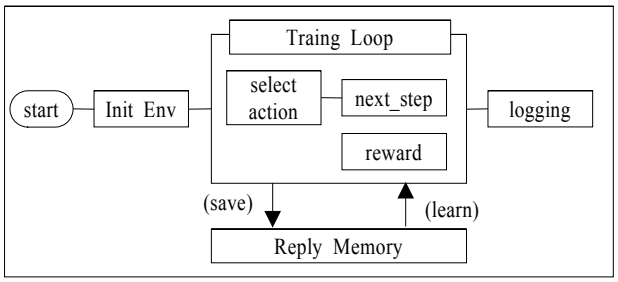
심층강화학습에서 사용하는 할인인자 r은 0.8로 고정하고, 적의 공격은 40발을 랜덤으로 적 미사일 포대에서 할당하여 공격하고 3개의 포대에서 공격하여 총 120발을 공격한다. 방어포대의 요격유도탄 재고는 각 포대마다 60발씩 할당한다. 총 300회 반복하여 학습한다.




<Figure 5> Scenario Summary [8, 10]

DQN이 적용된 심층강화학습은 <Figure 6>의 flowchart와 같이 진행된다. 환경을 초기화하면서 적 미사일의 공격 시나리오가 결정되고, 딥러닝 모델을 통해 학습하여 상태에 따른 행동을 선택한다. 환경으로부터 다음 상태와 보상을 얻고, 샘플을 리플레이 메모리에 저장하고 리플레이 메모리에서 무작위 추출한 샘플로 학습하여 모델을 업데이트한다. 공격 시나리오가 끝날 때까지 반복하고 기록한다.

자산별 방어 활동을 결정하는 멀티에이전트로 학습하며, 적 공격 시나리오의 공격에 방어자산별로 미사일 방어 여부를 판단한다. 2번 자산의 경우에는 요격이 결정될 시 남은 요격포탄이 많은 방어포대가 적 미사일을 요격한다.




<Figure 6> Deep Reinforcement Learning Model Flowchart

시뮬레이션 실험을 위해 Python과 Tensorflow를 활용하였고, 실험 결과 기록을 위해 Weights and Biases(wandb)를 이용하였다.

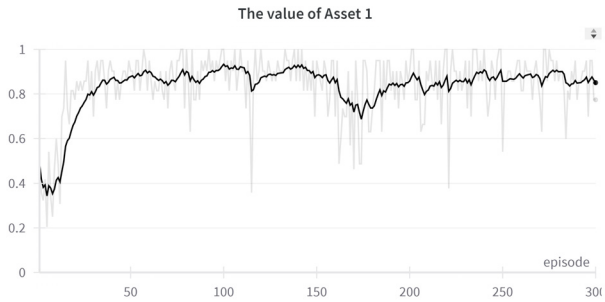
### 4.2 시뮬레이션 결과

설명한 시나리오대로 심층강화학습을 통한 시뮬레이션

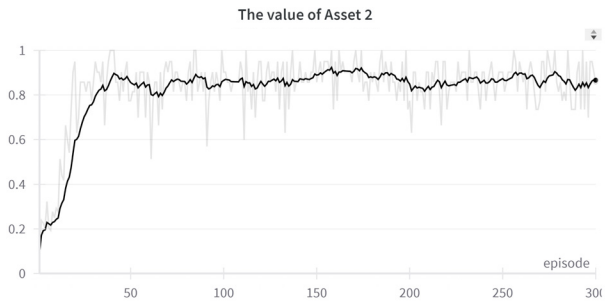
을 시행하였다. 시뮬레이션의 결과는 <Table 1>과 같다. 또한, 각 자산의 학습에 따른 자산의 가치 변화는 <Figure 7>~<Figure 9>처럼 그래프로 나타낼 수 있다.

<Table 1> Simulation Results

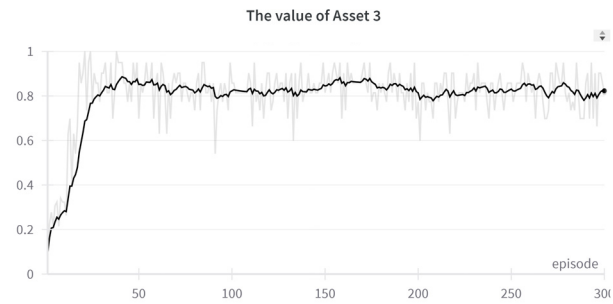
| | Asset1 | Asset2 | Asset3 |
|---|---|---|---|
| Average | 0.8628 | 0.8642 | 0.8276 |
| Standard Deviation | 0.1180 | 0.0855 | 0.0781 |
| Minimum Value | 0.4584 | 0.5134 | 0.5404 |
| 25 percentile | 0.8299 | 0.8265 | 0.7631 |
| Median | 0.9025 | 0.8574 | 0.8574 |
| 75 percentile | 0.9602 | 0.9263 | 0.8897 |
| Maximum Value | 1.0000 | 1.0000 | 1.0000 |
| 95% Confidence Interval | [0.8413, 0.8843] | [0.8491, 0.8793] | [0.8152, 0.8399] |




<Figure 7> The Value Graph of Asset 1




<Figure 8> The Value Graph of Asset 2

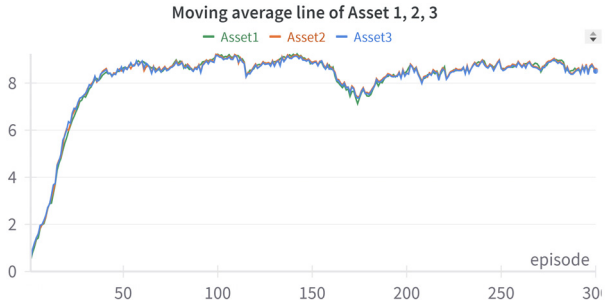



<Figure 9> The Value Graph of Asset 3

결과 그래프에서 보듯이 학습에 따라 Asset 1, 2, 3의 가치가 최소 0.4584에서 최대 1까지 미사일 요격을 통해 기지의 가치를 방어하는 모습을 볼 수 있다. 학습에 따라 적 미사일을 요격하여 기지의 가치가 1에 수렴하는 모습을 보여주지만, 확률적으로 요격에 실패할 수 있기 때문에 변동성이 있는 그래프를 그리고 있다.

기지 가치의 이동평균선을 보면 심층강화학습을 통한 학습 패턴이 보인다. 초기 급등을 거쳐 안정화를 거치며, 점진적인 성능개선과 고점에 도달한 모습을 보여주고 있고 이는 각 기지가 동일한 학습 패턴을 보여주고 있다. 기지의 최대 가치는 1로 한계가 있어서 학습에 따른 이동평균선 역시 고점에 도달하여 수평으로 안정적인 학습을 이루어 가는 모습을 보여준다.

<Figure 10>은 각 기지의 이동평균선의 그래프이다.




<Figure 10> Moving average line of Asset 1, 2, 3

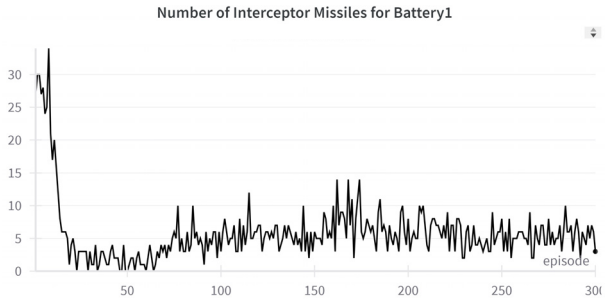
미사일 요격 확률이 90% 이므로 적 미사일에 요격유도탄을 1:1로 할당했을 때의 각 기지가 확률적으로 피격받을 수 있는 적 미사일 수는 식 (5)에 의해 40발당 4발이며, 이에 따라 예상되는 기지 가치는 0.8145가 된다.

$$E(N_{\_hits}) = N \times (1-p) \tag{5}$$

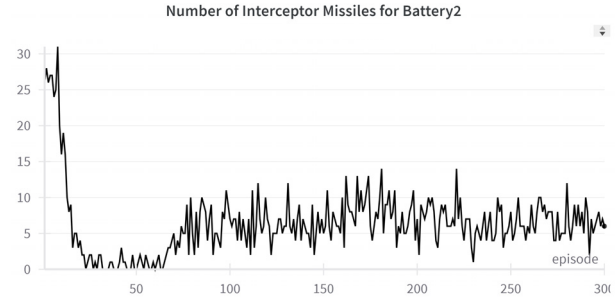
따라서 시뮬레이션을 통한 결과값이 유의미한 결과를 가질 수 있는지를 확인하면. <Table 1>의 95% 신뢰구간이 0.8145를 포함하고 있지 않으므로 유의미하게 높은 평균값을 가지고 있음을 알 수 있고 이는 표준 편차를 고려해도 동일하게 적용되므로 DQN을 통한 학습의 결과가 개선된 미사일 요격을 할 수 있음을 나타낸다. 또한 25 percentile 값과 비교했을 때 약 5%, Maximum Value와 비교했을 때는 약 23%의 개선된 기지의 가치이다.

<Figure 11>, <Figure 12>는 기지를 보호하는 요격포대가 방어에 사용한 요격미사일의 숫자를 보여주고 있다. 요격을 많이 할수록 방어의 확률이 높아지고, 방어를 많이 할수록 Asset의 가치가 줄어들지 않고, 높은 보상을 받을 수 있기 때문에 포탄을 많이 사용하는 방향으로 학습하는

것을 볼 수 있다. 또한 심층강화학습 모델의 적 미사일 공격의 끝을 알 수 없으므로 일정 포탄수를 확보하며 학습하는 모습을 볼 수 있다.




<Figure 11> Number of Interceptor Missiles for Battery 1




<Figure 12> Number of Interceptor Missiles for Battery 1

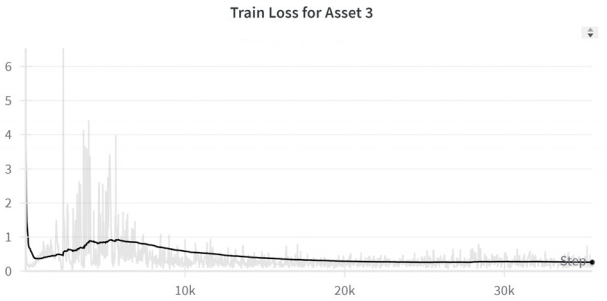
<Figure 13>은 심층강화학습을 통한 각 에피소드별 보상을 시각화해 주고 있다. 기지가치의 이동평균선과 동일한 패턴을 보인다. 요격의 성공 여부가 이산확률에 기반을 두고 있고, 각각의 에피소드에서 다른 경험을 함으로써 변동성이 있는 것처럼 보이지만 시행착오를 겪으며 점진적인 성능이 향상되고 있음을 보여준다.


<Figure 13> Reward Graph

<Figure 14>~<Figure 16>은 딥러닝 모델을 적용한 loss 값에 대한 그래프로 학습에 따라 최적화되어 가는 모습을 보여준다.


<Figure 14> Traing Loss Graph for Asset 1


<Figure 15> Traing Loss Graph for Asset 2




<Figure 16> Traing Loss Graph for Asset 3

## 5. 결 론

적 미사일 공격에 대한 방어체계에 적용하기 위한 심층강화학습 기반의 인공지능 방법론을 연구하였다. 선행연구를 바탕으로 WTA 문제를 MDP로 모형화하였고, 기존 연구에서 수행한 ADP 방법론과의 차이점과 심층강화학습을 통한 이점을 제시함으로써 방어체계 알고리즘을 심층강화학습을 통해 발전시킬 수 있음을 보였다.

연구결과 무작위적 포대의 공격에 대해 적절한 대응을 심층강화학습을 통한 시도와 경험을 통해 학습하여 대응하는 모습을 볼 수 있었으며, 이를 딥러닝을 통한 심층강화학습 모델을 통하여 학습하여 대응할 수 있음을 입증하였다. 이로써 본 논문의 학문적 공헌은 대표적 심층강화학

습의 알고리즘인 DQN을 개선하여, 장거리 화력 위협 도메인에 적용을 통하여 심층강화학습의 적용 범위를 확대하여도 좋은 방법론임을 검증하였고, 실제 사격과 유사하게 적의 공격을 사전에 알 수 없고 패턴화 시키지 않음으로서 더 실제사례와 유사한 시뮬레이션을 보여주었다.

향후에는 심층강화학습을 위해 변형된 DQN 기반의 멀티에이전트를 이용하였듯이 DQN 외에 최신의 다양한 기법을 적용하여 연구하는 것도 또 다른 연구영역의 한 분야라고 생각된다.

더욱이 지금까지 연구가 단순하게 무장할당에 기반한 WTA 문제를 MDP로 모형화하여 학습시켰다면 디지털 트윈처럼 실제 미사일의 기동과 요격미사일의 요격까지 시뮬레이션하여 학습시키는 고차원의 연구도 필요하다. 이를 통해 실데이터를 가상화하여 얻을 수 있기 때문에 한국형 미사일방어체계를 구현하는데 꼭 필요한 연구가 될 것으로 생각한다.

## Acknowledgement

This study has been partially supported by industry-academic research of Hannam University and Hanwha System.

## References

[1] Bertsekas, D., Homer, M., Logan, D., Patek, S., and Sandell, N., Missile Defense and Interceptor Allocation by Neuro-Dynamic Programming, *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 2000, Vol. 30, No. 1, pp. 42-51.

[2] Cha, Y.H. and Jeong, B., Exact Algorithm for the Weapon Target Assignment and Fire Scheduling Problem, *Journal of the Society of Korea Industrial and Systems Engineering*, 2019, Vol. 42, No. 1, pp. 143-150.

[3] Davis, M.T., Robbins, M.J., and Lunday, B.J., Approximate Dynamic Programming for Missile Defense Interceptor Fire Control, *European Journal of Operational Research*, 2017, Vol. 259, pp. 873-886.

[4] Im, J.S., Yoo, B.C., Kim, J.H., and Choi, B.W., A Study of Multi-to-Majority Response on Threat Assessment and Weapon Assignment Algorithm: by Adjusting Ballistic Missiles and Long-Range Artillery Threat, *Journal of Korean Society of Industrial and Systems Engineering*, 2021, Vol. 44, No. 4, pp. 43-52.

[5] Jang, B.C. and Kwon, H.J.., Consideration on Our Asymmetric Response through the Israel-Hamas Surprise Attack, *Defense & Technology*, 2023, Vol. 538, pp. 116-125.

[6] Jang, J.G., Kim, K., Choi, B.W., and Suh, J.J., A Linear Approximation Model for an Asset-based Weapon Target Assignment Problem, *Journal Society of Korea Industrial and System Engineering*, 2015, Vol. 38, No. 3, pp. 108-116.

[7] Jung, J.K., Uhm, H.S., and Lee, Y.H., Rolling - Horizon Scheduling Algorithm for Dynamic Weapon - Target Assignment in Air Defense Engagement, *Journal of the Korean Institute of Industrial Engineering*, 2020, Vol. 46, No. 1, pp. 11-24.

[8] Kim, H.H., Kim, J.H., Kong, J.H., and Gyeong, J.H., Reinforcement Learning-based Dynamic Weapon Allocation for Multiple Long-range Artillery Attacks, *Journal of the Korean Institute of Industrial Management Systems*, 2022, Vol. 45, No. 4, pp. 42-52.

[9] Kim, J.H., Kim, K., Choi, B.W., and Suh, J.J., An Application of Quantum-inspired Genetic Algorithm for Weapon Target Assignment Problem, *Journal Society of Korea Industrial and System Engineering*, 2017, Vol. 40, No. 4, pp. 260-267.

[10] Lee, C.S., Kim, J.H., Choi, B.W., and Kim, K.T., Approximate Dynamic Programming Based Interceptor Fire Control and Effectiveness Analysis for M-To-M Engagement, *Journal of the Korean Society for Aeronautical & Space Sciences*, 2022, Vol. 50, No. 4, pp. 287-295.

[11] Lee, W.W., Yang, H.R., Kim, G.W., Lee, Y.M., and Lee, E.R., *Reinforcement Learning with Python and Keras* (Revised Edition), Published April 7, 2020, pp. 227-247.

[12] Lee, Z.J., Lee, C.Y., and Su, S.F., An Immunity Based Ant Colony Optimization Algorithm for Solving Weapon-Target Assignment Problem, *Applied Soft Computing*, 2002, Vol. 2, No. 1, pp. 39-47.

[13] Li, S.E., Deep Reinforcement Learning, *Reinforcement Learning for Sequential Decision and Optimal Control*, Singapore: Springer Nature Singapore, 2023, pp. 365-402.

[14] Li, Y., Deep reinforcement learning: An overview. arXiv preprint arXiv:1701.07274, 2017, pp. 5-28.

[15] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M., Playing Atari with Deep Reinforcement Learning, arXiv preprint arXiv:1312.5602 , 2013, pp. 2-5.

[16] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A.,

Veness, J., Bellemare, M.G., and Hassabis, D., Human-level Control Through Deep Reinforcement Learning, *Nature*, 2015, Vol. 518, No. 7540, pp. 529-533.

[17] Naeem, H. and Masood, A., An Optimal Dynamic Threat Evaluation and Weapon Scheduling Technique, *Know-ledge-Based Systems*, 2010, Vol. 23, No. 4, pp. 337-342.

[18] Park, Y.W., and Jung, J.W., Formulation of a Defense Artificial Intelligence Development Plan, *Korean Society for Defense Technology*, Dec. 2020, pp. 3-8.

[19] Powell, W.B., *Approximate Dynamic Programming: Solving the Curse of Dimensionality*, 2011, Second Edition, John Wiley & Sons, Hoboken, NJ., pp. 315-346.

[20] Powell, W.B., *Approximate Dynamic Programming: Solving the Curse of Dimensionality*, Second Edition, 2011, John Wiley & Sons, Hoboken, NJ., pp. 235-276.

[21] Powell, W.B., Perspectives of Approximate Dynamic Programming, *Annals of Operations Research*, 2012, Vol. 13, No. 2, pp. 1-38.

[22] Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P., Trust Region Policy Optimization, In *Proceedings of The 32nd International Conference on Machine Learning*, 2015, pp. 1-9.

[23] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O., Proximal Policy Optimization Algorithms, arXiv preprint arXiv:1707.06347, 2017, pp. 365-402.

[24] segye news, https://www.segye.com/newsView/20231102526999 (accessed 2023/2/7).

[25] Shin, M.K., Park, S.-S., Lee, D., and Choi, H.-L., Mean Field Game based Reinforcement Learning for Weapon-Target Assignment, *Journal of the Korea Institute of Military Science and Technology*, 2020, Vol. 23, No. 4, pp. 337-345.

[26] Summers, D.S., Robbins, M.J., and Lunday, B.J., An Approximate Dynamic Programming for Comparing Firing Policies in a Networked Air Defense Environment, *Computers & Operations Research*, 2020, Vol. 117, pp. 1-29.

[27] Sutton, R.S. and Barto, A.G., *Reinforcement learning: An introduction*, 2nd ed., 2018, pp. 30-39.

[28] Tutorials for Reinforcement Learning, https://tutorials.pytorch.kr/intermediate/reinforcement_q_learning.html (accessed 2024/1/5).

[29] Yonhapnews, https://www.yna.co.kr/view/AKR20220410019151504 (accessed 2023/2/7).

[30] Yonhapnews, https://www.yna.co.kr/view/MYH20231012022600641 (accessed 2024/2/7).

**ORCID**

Min Gook Kim | https://orcid.org/0009-0005-1121-7592
Dong Wook Hong | https://orcid.org/0009-0002-4801-6098
Bong Wan Choi | https://orcid.org/0000-0002-9609-1714
Ji Hoon Kyung | http://orcid.org/0000-0002-0359-5594