심층 강화학습을 활용한 비동기식 Self-Play 사이버 공방 연구

김정현 [1], 정재혁 [1], 장준원 [2], 김민석 [2*]

[1] 상명대학원 전자정보시스템공학과
[2*]상명대학교 휴먼지능로봇공학과

[1]jungkim9898@gamil.com, [1]2023D1013@smu.ac.kr, [2*]minsuk.kim@smu.ac.kr

Asynchronous Self-Play Simulation of Cyber-attack/defense via Deep Reinforcement Learning

Jung Hyun Kim [1], Jae Hyeok Jeong [1], Jun Won Jang[2], Min-Suk Kim [2*]
[1]Sangmyung University, Dept. of Electronic Information System Engineering
[2*]Sangmyung University, Dept. of Human Intelligence & Robot Engineering

요 약

최근 매년 증가하는 사이버 위협으로 인해 경제적 손실과 국가적 피해가 심각하게 확대되고 있다. 본 논문에서는 사이버 공방 대응을 위한 비동기식 시뮬레이션 환경을 설계하고, 이를 활용하여 심층 강화학습 기반의 적대적 멀티 에이전트 학습 기법을 제안한다. 제안된 환경에서 공격자는 방어자가 보호하는 노드를 점거하는 것을 목표로 하며, 방어자는 해당 노드를 방어하는 것을 목표로 한다. 또한, 공격자의 정상적인 공격 시퀀스 생성을 확인하기 위해 DQN, PPO, SAC 기반 강화학습 모델을 적용하여 성능을 검증하였으며, 방어자는 이를 대상으로 한 방어 실험을 통해 성능을 평가하였다. 실험 결과, 공격자의 공격 성공률은 SAC 44.44%, DQN 2060%, PPO 300%로 나타났으며, 그 중 SAC 가 상대적으로 가장 안정적이고 우수한 성능을 보였다.

Ⅰ. 서 론

사이버 위협은 개인, 기업, 국가를 대상으로 매년 증가하고 있다. 한국인터넷진흥원(KISA)에서 파악한 국내 사이버 공격 시도는 23 년 1,277 건이며 24 년은 1887 건으로 510 건이 증가하여 약 15%증가함을 보였다. 2025 년 올해 사이버 공격 횟수는 SKT 침해사고 여파에 따라 상반기에만 1034 건을 달성하였고, 이는 2023 년 상반기에 발생한 664 건, 24 년 상반기에 발생한 899 건에 비해 각각 55.72%, 15.02%가 증가 했음을 확인할 수 있다.[1] 또한 World Economic Forum(WEF)에서 전세계의 사기 및 사이버범죄 피해는 최근 12 개월 동안 1 조 달러 이상이 발생했으며 일부 국가의 GDP 의 3%의 손실을 얻었음을 확인하였다.[2] 이러한 사이버 위협은 경제적으로 막대한 피해를 초래할 뿐만 아니라 국가 안보를 위협하는 수단으로 활용되고 있다. 이에 대응하기 위한 사이버 보안 기술은 현대 사회에서 핵심적인 안보 수단으로 자리 잡고 있으며, 최근에는 인공지능과의 융합을 통해 그 성능과 효율성이 지속적으로 발전하고 있다. 특히, 강화학습은 복잡한 사이버 공격 방식을 학습하고, 다양하게 생성된 공격 패턴에 대한 방어 기술을 개발하는 데 유용한 연구 방법[3]이다. 특히 공격과 방어를 독립적으로 학습하는 기존 방식과 달리, 멀티 에이전트 강화학습(MARL)을 활용하면 공격자와 방어자가 상호작용 속에서 동시에 학습할 수 있어 사이버 보안 연구에 효과적으로 적용되고 있다.

본 논문에서는 Unreal 엔진 기반의 사이버 공방 시뮬레이션 환경을 구축하고, 이를 활용하여 MARL 기반의 비동기식 사이버 공방 학습 환경을 설계하였다. 기존의 사이버 공방 시뮬레이션 환경은 동기식으로 제작되어 공격과 방어가 차례대로 학습하는 방식[4]을 일반적으로 사용한다. 이러한 구조는 공격과 방어 행위에 대한 정확성을 높일 수 있는 반면, 실시간성을 확보하기 어려움이 있다. 본 논문에서는 이러한 문제점을 해결하기 위해 Unreal 엔진을 활용하여 공격과 방어의 행위 과정을 시각화 함으로써, 실시간으로 학습 과정을 관찰할 수 있는 환경을 구축하였다. 또한, 실시간으로 변화하는 네트워크 환경을 대상으로, 공격자는 특정 노드를 점령하고 방어자는 이를 저지하는 것을 목적으로 시뮬레이션을 설계하여 실험을 진행하였다.

Ⅱ. 본문

2.1 사이버 공방 기반 시뮬레이션 환경

본 논문에서는 기존의 동기식 사이버 공방 구조가 아닌 비동기식 시뮬레이션 환경을 구현하기 위해 Unreal 기반의 사이버 공방 시뮬레이션을 구축하였다. 전체 시뮬레이션 구조는 (1) 환경 상태를 관리하는 클라이언트(Unreal), (2) 환경과 에이전트 간의 통신을 담당하는 서버, (3) 환경 내 행동을 수행하는 클라이언트(Agent)로 구성된다.

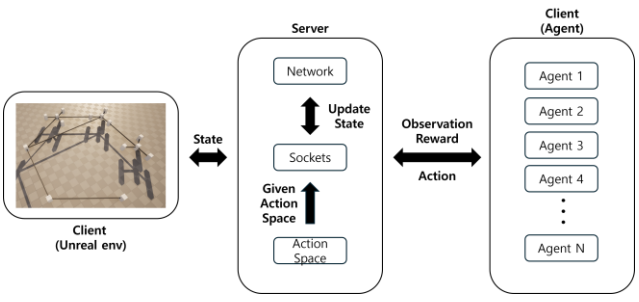

그림 1 Unreal 엔진 기반 사이버 공방 환경 구조

그림 1 은 사이버 공방 환경 시뮬레이션의 전체 구조를 보여주며, 학습 에이전트는 공격팀과 방어팀으로 구분되어 학습이 진행된다. 이때 공격팀은 공격자 단말을 통해 점진적으로 정보를 획득하여 학습을 수행하며, 방어팀은 개별 에이전트가 각 노드를 독립적으로 담당하도록 설계되었다. 공격도구와 방어도구는 단순화 작업을 최소화한 상태에서 MITRE ATT&CK 프레임워크에 기반하여 구현하였으며, 환경 조건에 따른 성공 유무를 판단하여 보상함수를 설계하고 이를 기반으로 학습을 진행하였다.[4]

2.2 Method: Decentralized Training with Decentralized Execution

MARL 학습은 크게 두 가지 접근 방식으로 구분된다. 첫째, 중앙에서 에이전트의 학습을 통합적으로 관리하는 방식인 CTDE(Centralized Training with Decentralized Execution)이며, 다른 하나는 학습과 실행 모두 에이전트가 독립적으로 수행하는 DTDE(Decentralized Training with Decentralized Execution)이다. 본 논문에서는 공격 에이전트가 점진적으로 정보를 취득하여 학습하는 POMDP(Partially observable Markov decision Process) 구조와 방어 에이전트가 자신이 관리하는 노드의 정보를 기반으로 학습하는 MDP(Markov Decision Process) 구조를 채택하기 때문에 두방식을 동시에 중앙에서 관리하는 것이 어렵다. 이에 각자의 역할에 최적화된 학습을 수행할 수 있도록 DTDE 방식을 기반으로 학습을 진행하였다.

Ⅲ. 실험

3.1 실험 환경 구성

본 논문에서는 하나의 공격 에이전트와 세 개의 방어 에이전트로 학습 환경을 구성하였으며, 최종 학습 목표는 방어 중인 세 개의 노드를 탈취하는 것으로 설정하였다. 반대로, 방어 에이전트는 각자 담당하는 노드의 점거를 방지하는 것을 목표로 학습을 수행하며, 이를 통해 공격자의 침투 시도를 저지하도록 구성하였다. 공격 순서는 실제 해킹 순서와 동일하게 미리 점거한 단말기와 연결된 노드들 스캔하고 공격하고자 하는 노드를 기준으로 취약점을 찾는 행위를 통해 정보를 취득하여 최종 Exploit 을 통해 노드를 탈취하는 방식이다. 방어는 공격의 스캔 행위가 실제 환경에서 대응하기 어렵거나 불가능하기 때문에, 공격 에이전트가 생성한 취약점 정보를 패치(Patch)하거나 공격 수행에 필요한 상태를 사전에 변경하는 방식으로 학습을 진행하였다.

3.2 MARL 기반 강화학습 모델 성능 검증

본 논문의 실험 환경에서는 학습에 앞서, 공격 에이전트가 실제 해킹 절차와 동일한 순서로 공격 과정을 생성하는지 확인하였다. 또한, 해당 에이전트가 주어진 공격 도구를 사용하여 상태공간을 확인하고 상황에 맞는 공격 학습이 진행됨을 확인하였다. 그림 2 는 강화학습 모델의 성능 검증을 위해 DQN(Deep Q-learning Network), PPO(Proximal Policy Optimization), SAC(Soft Actor-Critic)을 학습시킨 Reward-Episode 학습 성능 결과 그래프이다.

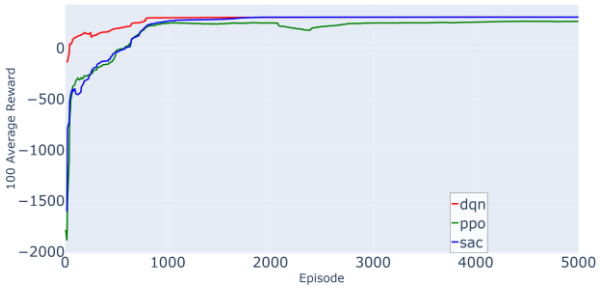

그림 2 Single Attacker Result for DQN, PPO, SAC Reward

DQN 의 공격 시퀀스는 최종 10 개, SAC 는 9 개, PPO 는 27 개의 공격 시퀀스를 생성함을 확인하였다. 공격 순서 역시 현재 설계된 시나리오에서 가정한 공격 순서와 유사하게 생성했기 때문에 현재 제작된 사이버공방 환경에서 의도한 결과를 도출했음을 확인하였다. 또한, 방어 에이전트를 도입하여 에이전트간 Self-Play 학습을 진행하였다. 그림 3 은 방어 에이전트를 적용했을 때 나타난 DQN Reward-Episode 결과 그래프이며, 결과적으로 기존 단일 공격자에 비해 도달한 보상함수의 최대 값이 감소함을 확인하였다.

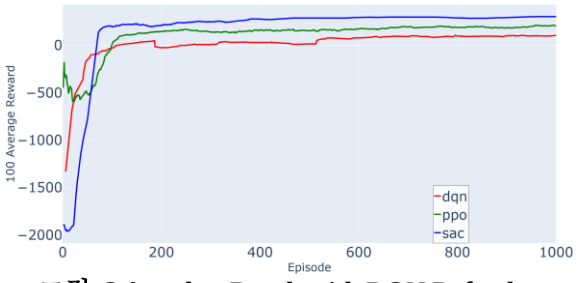

그림 3 Attacker Result with DQN Defender

최종 1000 에피소드를 기준으로 SAC 는 13 의 시퀀스를 생성했으며, PPO 는 108, DQN 은 216 의 시퀀스를 생성함을 확인하였고, 이는 방어자가 수행한 행동이 공격자의 유효한 행위 실행을 성공적으로 저지했음을 나타내고 있다.

Ⅲ. 결론

본 논문은 비동기식 사이버 공방 시뮬레이션 환경을 설계하고, 공격/방어 에이전트 간의 상호 적대적 학습을 통해 해당 환경의 신뢰성을 검증하였다. 비동기식 Self-play 학습방식은 동기식과는 다르게 통신 지연 및 작동 순서에 영향을 받는다. 따라서 단일 공격 행위만을 학습한 경우, 설계된 환경에서 선택할 수 있는 최선의 수인 9 개의 시퀀스가 아닌, 평균 15 개의 시퀀스를 생성하므로 일부 비효율적으로 작동함을 확인하였다. 또한, 공격 에이전트의 학습을 확인한 뒤 비동기식 Self-Play 환경에서 방어 에이전트를 적용하여 학습 진행하였을 때, 공격 에이전트의 공격 시퀀스 개수가 기존 대비 평균 약 646.67% 증가함을 확인할 수 있다.

ACKNOWLEDGMENT

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea Government (MSIT) (No. RS-2022-II220961)

참 고 문 헌

[1] 한국인터넷진흥원, “2025 년 상반기 사이버 위협 동향 보고서” 2025, (https://www.kisa.or.kr)

[2] R. Muggah and M. Margolis. Global Cybersecurity Outlook 2025. 2025, (https://www.weforum.org/publications/global-cybersecurity-outlook-2025/)

[3] Landolt, Christoph R., et al. "Multi-Agent Reinforcement Learning in Cybersecurity: From Fundamentals to Applications." arXiv preprint arXiv:2505.19837 (2025).

[4] MITRE, “MITRE ATT&CK,” 2024 (https://attack.mitre.org)