Self-Play 사이버 공방을 위한 심층 강화학습 시뮬레이션 연구

김정현 [1], 정재혁 [2], 김민석*

[1, 2] 상명대학원 전자정보시스템공학과
*상명대학교 휴먼지능로봇공학과

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

Study of Deep Reinforcement Learning Simulation for Self-Play in Cyber Attack and Defense

Kim Jung Hyun, Jeong Jae Hyeok, Kim Min Suk*
Sangmyung University, Dept. of Electronic Information System Engineering,
*Sangmyung University, Dept. of Human Intelligence & Robot Engineering,

요 약

사이버 위협이 증가함에 따라 인공지능을 활용한 다양한 사이버 대응 기술 개발이 활발하게 이뤄지고 있다. 본 논문에서는 Network Attack Simulator(Nasim)를 이용한 강화학습 기반 Self-Play 네트워크 공격 환경 개발을 위해 기존 공격 에이전트와 더불어 새로운 방어 에이전트를 추가하여 공방 대응 정책을 생성하였다. 해당 검증 정책을 검증하기 위해 DQN, PPO, SAC 기반 강화학습 모델을 대상으로 실험을 진행하였으며 이를 통해 다양한 공방 성능을 확인할 수 있다. 실험 결과, SAC 학습 모델이 다른 모델들에 비해 평균 60.67% 더 높은 승률을 나타내고 있다.

Ⅰ. 서 론

현재 정보화시대에 접어들며 사이버 기술이 급격히 발전하고 있지만, 이로 인해 사이버 위협도 증가하고 있다. 한국인터넷진흥원(KISA)의 사이버 위협 동향 보고서에 따르면, 2022 년 사이버 공격 시도는 1,142 회였고, 2023 년에는 1,277 회로 약 12% 증가했다.[1] 또한, 러시아-우크라이나 전쟁에서 사이버 무기체계의 확산으로 사이버 공격은 국가 안보를 위협하는 중요한 수단으로 부각되고 있다. 이러한 상황에서 인공지능을 활용한 사이버 공방 시뮬레이션, 예를 들어 Network Attack Simulator(Nasim) [2]이나 Microsoft 의 Cyber Battle Simulation(CBS)[3] 등이 제작되어 사이버 공격과 방어의 이해를 돕고 있다.

본 논문에서는 Nasim 환경에서 강화학습을 활용해 사이버 공방을 시도했다. Nasim 은 기본적으로 강화학습을 기반으로 네트워크를 장악하는 시뮬레이션이지만 공격 대응 기술이 존재하지 않는다. 따라서 공격에 대응이 가능한 방어 기술을 추가함으로써 방어자가 공격자에 대응하며 상호작용하는 환경을 구성하였다. 또한, Nasim 의 공격자는 네트워크 장악을 위해 현실의 공격 기법을 적용한 추상화된 공격 기술을 사용하고 있으며 이러한 사이버 공방 환경에서 적용된 강화학습 알고리즘의 특성에 따른 결과를 비교 및 검증하였다.

Ⅱ. 본론

2.1 Self-Play 사이버 공방 시뮬레이션 구성

본 논문에서는 Nasim 을 바탕으로 Self-Play 사이버 공방 시뮬레이션 환경을 구축하였다. 기존 Nasim 은 네트워크 공격을 추상화한 시뮬레이션 환경으로써 실제환경에서의 복잡한 공격 과정이 추상화되어 있기 때문에 네트워크 공격 방식을 보다 쉽게 이해할 수 있다. 하지만, Nasim 환경은 공격만 제공하기 때문에, 실제환경처럼 방어 기술이 존재하고 변화하는 네트워크 환경에 대한 공격자 성능 신뢰성도 낮다. 본 논문에서는 방어자와 방어 및 후속 조치 행동을 추가하여 더욱 정교한 사이버 공방 시뮬레이션을 구현하였다. 이때 방어 기술은 Node-Scan, Passwd Change, Reboot 등의 방어 도구를 적용하여 실험을 진행하였고 그림 1 은 방어자 에이전트 추가 이후의 전체적인 환경 구조도를 나타낸다.

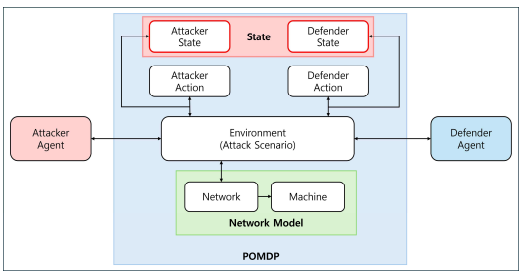

그림 1 Nasim Self-Play 공방 환경 구조

추가로 기존 제공되던 공격 기술인 Scan, Exploit 은 현실에 비해 너무 단순하게 설계가 되어 있었기 때문에, MITRE ATT&CK 의 실제 공격 기법을 응용하여 OS Credential Dumping, Privilege Escalation, Password Cracking 등을 활용해 실제 공격기법을 응용하였다.[4]

2.2 On/Off-Policy-based Reinforcement Learning

강화학습은 학습을 진행하는 에이전트가 주어진 환경에서 상호작용을 통해 보상을 최대화하는 행동 정책을 찾는 방법이다. 기본적으로 강화학습은 정책 학습 방식은 상태나 상태-행동 쌍의 가치를 계산하여 최적 정책을 학습하는 Value-based 와 정책을 직접적으로 학습하는 Policy-based 방법으로 구분할 수 있다. 이를 더 세부적으로 분류한다면 행동한 정책을 통해 목표 정책을 평가하는 방법인 On-Policy 와 이전에 경험했던 행동을 바탕으로 목표 정책을 평가하는 방법인 Off-Policy 로 구분할 수 있다. 본 논문에서는 정책 학습 방식에 따른 학습 결과를 비교 및 분석하며, 실험에서는 Deep Q-Network(DQN)[5], Proximal Policy

Optimization(PPO)[6], Soft Actor-Critic(SAC)[7]를 학습시키고 결과를 비교하였다.

Ⅲ. 실험

3.1 실험 환경 설정

본 논문에서는 사이버 공방 상황에서 방어자를 역할을 추가하였기 때문에, 방어자가 환경에서 작동할 수 있도록 환경 Process 를 일부 변환하여 공격자와 방어자가 동시에 작동 가능하도록 설계하여 실험을 진행하였다. 방어자와 공격자는 서로 다른 상태정보(State)를 사용하여 각각 필요한 State 를 보고 학습을 진행한다. 또한, 공격자와 방어자 사이에 State 정보가 달라지는 것을 방지하기 위해 일부 함께 사용하는 State 정보를 매 Step 마다 동기화하였다. 이를 통해 공격자와 방어자가 변화한 State 정보를 바로 공유하여, 학습 과정에서 두 에이전트가 상호작용함을 확인했다.

3.2 RL 기반 모델 성능 검증

공격자와 방어자를 학습하는 것에 있어 처음부터 학습이 되지 않은 공격자를 통해 방어자를 학습하는 것은 공격자의 성능과 방어자의 성능 둘 다 해칠 수 있기 때문에, 공격자를 먼저 학습하고 학습된 공격자를 기반으로 방어자를 학습하는 방식으로 성능 비교 실험을 진행하였다.

그림 2 는 공격자를 DQN, PPO, SAC 로 했을 때의 학습 결과를 비교한 결과 그래프이다.

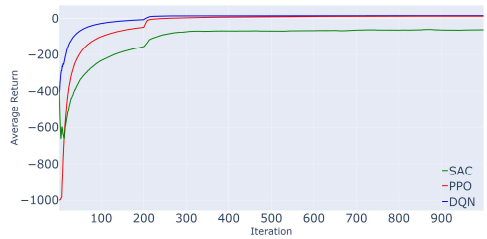

그림 2 Result for DQN, PPO, SAC Reward

수렴의 속도는 DQN, PPO, SAC 의 순서대로 빨랐으며, 결과 역시 Value-Based 인 DQN 이 좀더 안정적으로 높은 Reward 로 수렴하였다. 반면 Policy-Based 인 PPO 와 SAC 의 경우 DQN 에 비해 낮은 Reward 로 수렴하는 것을 확인할 수 있으며 이를 통해 외부의 방해 없는 상황에서는 DQN 과 같은 Value-Based 모델이 좀더 안정적인 모델을 생성함을 확인할 수 있었다. 이는 현재 state 에 대한 행동의 가치를 산출해서 학습되는 value-based 와 현재 state 에 대한 행동의 확률을 산출해서 학습되는 Policy-Based 의 특징의 차이가 나타남을 확인할 수 있다. 또한, 학습된 공격자를 방어자 학습을 위해 사용하여 방어자를 학습시킴과 동시에 공격자가 방어자의 행동에 맞춰 공격 전략을 바꾸는 것에 대해 실험을 진행하였고 학습을 진행한 총 5000 episode 에 대해 공격자의 초반부터 최종까지 승률은 Table 1 과 같이 나타난다.

Table 1 Attacker winning late.

Model	Episode	DQN Defender	PPO Defender	SAC Defender
DQN Attacker	1250	35.24%	24.72%	9.65%
	5000	55.65%	65.20%	39.58%
PPO Attacker	1250	98.22%	94.87%	79.55%
	5000	2.02%	0.48%	0.77%
SAC Attacker	1250	84.88%	100%	80.41%
	5000	87.02%	100%	76.84%

공격자와 방어자가 존재하는 상황에서는 DQN 방어자에 의해 State 가 지속적으로 바뀌기 때문에 성능이 많이 하락하는 것을 나타냈다. 또한 On-Policy 인 PPO 는 초반 방어자가 학습을 못한 상태에서는 높은 성능을 보이다 방어자가 학습이 진행된 이후에 업데이트가 진행된 정책이 유효하지 못하게 되는 상태로 변화하면서 성능이 급격하게 하락함을 보인 것을 확인하였다. SAC 의 경우 대부분의 상황에서 높은 승률을 도달함으로써 공격자와 방어자가 존재하는 상황에서 Policy-Based 와 Off-Policy 의 경우가 더 현재 환경에 적합함을 확인할 수 있다.

Ⅳ. 결론

본 논문에서는 기존 사이버 공격 시뮬레이션의 공격자의 신뢰도를 높이기 위해 방어자를 추가하여 학습 가능한 Self-Play 사이버 공방 환경을 구성하였다. 이를 통해 각 에이전트는 상대방의 행동이 서로에게 반영되는 상태로 최적의 정책을 생성할 수 있으며 Self-Play 환경에서의 강화학습 알고리즘별 모델 성능을 비교할 수 있었다. Self-Play 환경에서 Value-Based 인 DQN 은 단일 에이전트 환경보다 공격 성능이 하락함을 확인할 수 있었다. Policy-Based 인 PPO 나 SAC 가 일부 좋은 성능을 가짐을 알 수 있지만, On-Policy 로 학습되는 PPO 의 경우 학습이 지속될수록 방어자가 학습을 진행해 성능이 급격하게 하락하였다. 반면 off-Policy 로 학습되는 SAC 의 경우 대부분의 방어자를 상대로 우수한 성능을 가짐을 확인할 수 있었다.

ACKNOWLEDGMENT

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

참 고 문 헌

[1] 한국인터넷진흥원, “2023 년 하반기 사이버 위협 동향 보고서” 2024, (https://www.kisa.or.kr)

[2] Jjschwartz, “NetworkAttackSimulator,” 2019, (https://github.com/Jjschwartz/NetworkAttackSimulator,)

[3] Microsoft, “CyberBattleSim,” 2021, (https://github.com/microsoft/CyberBattleSim,)

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

[5] V. Mnih, K. Kavukcuoglu, D. Silver, et al., "Human-level control through deep reinforcement learning," Nature, pp. 529-533.

[6] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," arXiv preprint arXiv:1707.06347, 2017.

[7] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," Proceedings of the 35th International Conference on Machine Learning (ICML), pp. 1861-1870, July, 2018