위험 인지 기반 Look-ahead 관측을 활용한 A2AD 회피 강화학습 의사결정 모델 연구

김영근, 유연준, 윤소연, 이준[1]
육군사관학교

rladudrms82@gmail.com, c20681@kma.ac.kr, yeonyoonso@gmail.com,
jun.lee.mistra@gmail.com

A Risk-Aware Reinforcement Learning Decision Model for A2AD Penetration with Look-Ahead Observations

YoungKeun Kim, Yeonjun Yoo, So-Yeon Yoon, Jun Lee
Korea Military Academy

요 약

A2AD(Anti-Access/Area Denial) 환경에서 자율 에이전트가 목표 지점까지 안전하게 도달하기 위해서는 위험 지역을 효과적으로 회피하는 의사결정 전략이 요구된다. 본 연구에서는 위험 인지 기반 look-ahead 관측 기법을 활용하여, 에이전트가 이동 경로 상의 잠재적 위험을 사전에 고려할 수 있도록 하는 강화학습 기반 의사결정 모델을 제안한다. 제안한 모델은 현재 위치에서의 위험 정보뿐만 아니라, 현재 이동 방향을 유지한다고 가정했을 때 향후 K step 경로 상에서의 위험도를 관측에 포함하며, 부분 관측 환경에서의 시간적 정보 통합을 위해 LSTM 기반 정책망을 사용한다. 성능 평가 결과, 제안한 LSTM 기반 look-ahead 정책은 메모리 없는 MLP 기반 정책 대비 높은 임무 성공률과 낮은 충돌률을 보였다

Ⅰ. 서 론

현대 전장 환경에서 A2AD 전략은 상대 전력의 특정 지역으로의 접근이나 작전 수행을 제한함으로써 해당 지역에서의 영향력을 약화 또는 무력화하는 개념으로 정의된다[1]. 장거리 미사일, 지대공 방공망, 조기경보 레이더와 같은 위협 체계는 일정 범위의 위협 반경을 형성하며, 이른바 A2AD 버블(A2AD bubble) 내부로 진입하는 표적에 대해 탐지 및 요격을 수행한다. 이러한 환경에서 자율 무인기(UAV)나 무인 지상 차량(UGV)이 목표 지점까지 침투하거나 경로를 계획하는 것은, 위험 지역을 회피하면서도 임무 효율을 유지해야 하는 매우 도전적인 문제이다.

기존 경로 계획 기법으로 널리 알려진 인공 퍼텐셜 필드(Artificial Potential Field, APF)나 동적 윈도우 접근법(Dynamic Window Approach, DWA)은 장애물 회피에 대한 직관적인 해법을 제공하며 단순한 환경에서는 효과적인 성능을 보인다. 그러나 이러한 방법들은 사전에 정의된 규칙과 고정된 비용 함수에 기반하기 때문에, 복잡하거나 변화하는 환경에서는 적응성과 유연성이 부족하다는 한계를 가진다 [2]. 특히 A2AD 환경과 같이 위험 지역에 대한 접근이 완전히 금지되는 것이 아니라, 탐지되지 않는 범위 내에서는 근접 통과(near-miss)가 허용되는 상황에서는 기존의 경로 계획 기법이 지나치게 보수적이거나 반대로 안전 여유를 충분히 확보하지 못하는 문제가 발생할 수 있다 [3].

A2AD 회피 시나리오는 단순 장애물 회피와 달리, 위험에 대한 허용 범위와 목표 도달 효율 간의 균형이 핵심적인 요소로 작용한다. 기존 APF 기반 기법에서도 장애물과의 안전 거리를 고려하도록 개선한 연구들이 존재하며, 이는 위험장(risk field) 개념을 부분적으로 도입한 사례라 할 수 있다. 그러나 이러한 접근은 여전히 고정된 규칙에 의존하기 때문에 다양한 환경 조건이나 동적으로 변화하는 위협 상황에 대해 일관된 성능을 보장하기 어렵다. 이에 따라 환경과의 상호작용을 통해 전략을 학습할 수 있는 강화학습 기반 접근의 필요성이 제기된다[3].

본 연구에서는 A2AD 환경에서의 경로 계획 문제를 강화학습 기반으로 해결하고자 하며, 위험 인지 정보를 정책 학습에 직접 통합하는 새로운 접근을 제안한다. 특히 에이전트가 현재 위치에서의 위험 정보뿐만 아니라,

[1] 교신저자, 김영근, 유연준, 윤소연은 동일한 비율로 논문 작성에 기여 하였음

향후 K step 동안 이동할 경로 상의 위험을 사전에 인지할 수 있도록 하는 look-ahead 위험 관측 기법을 도입한다. 이는 인간이 전방 상황을 미리 인지하며 운전하는 방식과 유사한 직관을 강화학습 에이전트에 부여하는 것이다. 또한 환경이 부분 관측인 상황에서 이러한 위험 관측 정보를 효과적으로 활용하기 위해, LSTM 기반 순환 신경망 정책을 적용하여 시간적 위험 패턴을 통합하였다.

Ⅱ. 본론

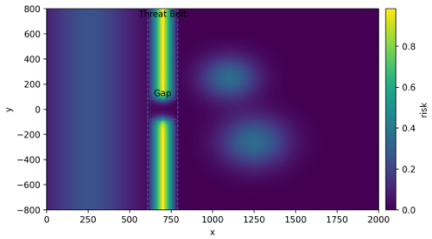

그림 1. A2AD 위험장 시각화

본 연구는 A2AD 환경에서 에이전트가 목표 지점에 도달하는 동시에 위협 구역을 회피하도록 학습하는 문제를 2 차원 연속 공간 시뮬레이션으로 정의한다. 에이전트는 일정 속도로 이동하며, 매 스텝마다 조향을 연속 행동으로 선택해 진행 방향을 조절한다. 기본 관측(state)은 현재 위치와 방향(yaw), 그리고 목표까지의 상대 위치(목표 방향/오프셋)로 구성되어 에이전트가 어디로 가야 하는지를 알 수 있다. 위험은 위험장(risk field)으로 모델링되며, 지도상 여러 위협(방공/레이더 등)이 각자 중심과 영향 반경을 가진다. 에이전트가 위협 중심에 가까울수록 위험도가 커지고 일정 임계치 이상에서는 사실상 격추(crash)에 해당한다. 복수 위협이 겹칠 때는 가장 치명적인 위협이 지배하도록 위험을 합성해, 에이전트가 실제로 피해야 할 최악의 위험이 반영되도록 한다.

시나리오는 난이도가 있는 A2AD 침투를 모사하기 위해, 그림 1 과 같이 특정 x 구간에 치명적 위협 벨트(threat belt)가 형성되어 있고 그 안에 좁은 안전 통로(narrow gap)가 존재하는 구조로 구성된다. 즉, 에이전트는 좁은 길을 찾아 통과해야 하는 탐색·회피 문제를 해결 하면서, 성공(SUCCESS), 격추(CRASH), 경계 이탈(OOB), 시간 초과(TIMEOUT) 등의 종료 조건을 통해 임무 결과를 구분한다. 보상은 에이전트가 목표를 향해서 전진하되, 위험을 억제 하도록 구성하였으며, 목표 접근 보상, 위험 페널티, 전방 위험 페널티, 종료 보상들을 고려한 종합 보상을 수행하였으며, 단순히 위험을 즉시 피하는 것이 아니라 근미래 위험까지 고려한 계획적 회피를 할 수 있도록 설계를 하였다.

$$r_t = r_t^{alive} + r_t^{prog} - w_{now} r_t^{now} - I[K>0] w_{look} r_t^{look} + r_t^{term}$$

본 논문에서 제안한 방법은 에이전트가 위험을 단순히 보상에서 벌점으로 주는 것을 넘어, 관측 자체에 위험 정보를 구조적으로 포함시키는 것이다. 이를 위해 기본 환경 위에 래퍼(RiskWrapper)를 적용하여 관측을 확장한다. 확장 관측은 크게 기본 상태, 현재 위치 위험도 및 전방 위험 요약으로 구성된다. 또한 노이즈와 전역 정보 부재로 인해 문제는 부분 관측(POMDP) 성격을 갖기 때문에, 과거 관측을 누적해 활용할 수 있도록 LSTM 기반 정책을 적용하고, 메모리가 없는 MLP 정책과 성능을 비교한다. 학습은 PPO 알고리즘을 사용하였다.

정책 학습은 Proximal Policy Optimization(PPO) 계열 알고리즘을 사용하였다. 기억 메커니즘의 효과를 비교하기 위해, 순환 구조가 없는 MLP 기반 정책과 LSTM 기반 정책을 각각 구성하였다. 총 학습 스텝 수는 200,000 step 으로 설정하였으며, rollout 길이는 2,048 step, 배치 크기는 256 으로 고정하였다. 할인 계수는 0.99, GAE 파라미터는 0.95 를 사용하였고, 학습률은 3×10^{-4} 로 설정하였다. PPO 의 클리핑 계수는 0.2 를 적용하였으며, 탐험에 따른 불확실성을 통제하기 위해 엔트로피 계수는 0.0 으로 설정하였다. 다음의 그림 1 은 성공률 결과를 look-ahead 길이에 따라 시각적으로 나타낸 것이다. LSTM 정책은 모든 K 에 대하여 MLP 대비 현저히 높은 성공률을 보였다.

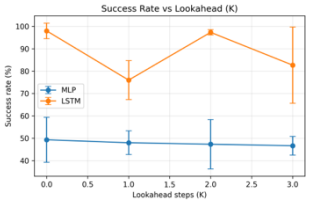

그림 1. Success/Crash/Timeout·OOB rate(%)

Ⅲ. 결론

본 연구에서는 위험 인지 기반 look-ahead 관측을 도입한 강화학습 정책을 통해 A2AD 회피 경로계획 문제를 다루었다. 정적인 위협 환경에서 에이전트가 근미래 K step 의 위험 정보를 관측하고, 이를 LSTM 기반 정책을 통해 시간적으로 통합함으로써 부분 관측 문제를 완화하도록 설계하였다. PPO 알고리즘을 이용한 학습 결과, 제안한 LSTM 기반 look-ahead 정책은 메모리 없는 정책 대비 높은 성공률과 낮은 충돌률을 보였으며, 경로 상의 최대 위험 노출 또한 감소하는 경향을 확인하였다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2024-00455788).

참 고 문 헌

[1] 김재학 , 김성현 and 홍건식, “중국의 반접근/지역거부(A2/AD) 전략과 딜레마: 삼각 억지와 강압을 중심으로”, 한국과 국제정치, 39(3), 37 - 74, 2023.

[2] Pan H., Luo M., Wang J., Huang T., Sun W. A safe motion planning and reliable control framework for autonomous vehicles. IEEE Trans. Intell. Veh. 2024;9:4780- 4793.

[3] Xiao X., Liu B., Warnell G., Stone P. Motion planning and control for mobile robot navigation using machine learning: A survey. Auton. Robot. 2022;46:569- 597. doi: 10.1007/s10514-022-10039-8.