강화학습 기반 최적 경로 탐색 UAV 정찰 시나리오

이성준, 박수현*, 김중헌
고려대학교, *숙명여자대학교

{ssungjoon, joongheon}@korea.ac.kr, *soohyun.park@sookmyung.ac.kr

Reinforcement learning-based optimal path search UAV reconnaissance scenario

Sungjoon Lee, Soohyun Park*, Joongheon Kim
*Korea Univ., **Sookmyung Women's Univ.

요 약

강화학습(Reinforcement Learning, RL)은 에이전트가 환경과 상호작용하며 보상을 최대화하는 최적의 행동을 학습하는 기법이다. 본 연구에서는 5x5 격자 환경에서 UAV 에이전트가 세 개의 목표 지점을 탐색하는 정찰 시나리오를 구현하고, 이를 위해 보상 함수를 설계하였다. 설계된 보상 함수는 목표 지점 탐색 성공 시 보상, 에너지 소비 패널티, 충돌 및 중복 탐색 패널티를 포함한다. 이를 통해 UAV 가 효율적이고 신속하게 목표 지점을 탐색하도록 유도하며, UAV 의 자율적 탐색 능력을 향상시킨다.

Ⅰ. 서 론

무인 항공기(Unmanned Aerial Vehicle, UAV)는 실제 조종사가 탑승하지 않고 지상에서 무선으로 조정되거나 프로그램적으로 설정한 경로를 이동하는 자율 이동체이다. 여러 법적 규제가 완화됨에 따라 점차 상업적인 응용 분야에서도 널리 사용되고 있다[1]. 지상 이동체와 달리 비행을 통해 이동의 제약을 극복한 UAV 는 앞으로의 시대를 새롭게 변화시킬 차세대 핵심 기술로 평가받고 있다. 그럼에도 불구하고, 현재 전기차를 비롯한 지상 이동체들의 자율 주행 기술이 주목받고 있는 시점에서, UAV 를 포함한 여러 자율 이동체들의 주행 기술은 아직 초기 단계에 머물러 있다. UAV 를 이용한 정찰의 이점은 사람이 직접 적진에 침투하여 중요 핵심 시설을 파악하는 감시 작업은 정보적 우위를 확보하여 전세를 확보하는 것에 큰 기여를 하지만, 이를 인공지능 기술을 적용한 자율 이동 UAV 가 대체하기에는 아직 기술적 한계가 있다[2]. 본 논문에서는 인공지능 기술의 하나인 강화학습을 UAV 에 적용하여 그리드 환경에서의 정찰 시나리오를 설계하고 성능평가를 통해 해당 알고리즘의 우수성을 평가한다[3].

Ⅱ. 본론

2-1 강화학습

RL 은 기계학습의 한 분야로, 에이전트가 환경과 상호작용하면서 주어진 목표를 달성하기 위한 최적의 행동을 학습하는 방법론이다. 이를 통해 에이전트는 장기적으로 최대의 보상을 얻을 수 있는 정책(policy)을 학습하게 된다. 본 논문에서는 강화학습의 기본 개념과 핵심 요소를 다루고, 이들이 에이전트의 학습 과정에서 어떻게 작용하는지를 설명하고자 한다.

2.1.1 상태(State)

상태는 환경의 현재 상황을 나타내는 정보로, 에이전트가 다음 행동을 결정하는 데 중요한 역할을 한다. 상태는 환경의 특정 시점에서의 모든 관련 정보를 포함하며, 이를 통해 에이전트는 최적의 행동을 선택할 수 있다. 예를 들어, 바둑 게임에서 상태는 현재 바둑판에 놓인 모든 돌의 위치와 색상 정보를 포함한다.

2.1.2 행동(Action)

행동은 에이전트가 취할 수 있는 모든 가능한 조치를 의미한다. 예를 들어, 자율주행 차량의 경우 행동은 속도 조절, 방향 변경, 정지 등의 조치로 구성될 수 있다.

2.1.3 보상(Reward)

보상은 에이전트가 특정 행동을 취한 후 환경으로부터 받는 피드백으로, 긍정적일 수도 있고 부정적일 수도 있다. 예를 들어, 바둑 게임에서 승리하는 행동은 높은 보상을 받게 되며, 패배하는 행동은 부정적인 보상을 받게 된다.

2.1.4 정책(Policy)

정책은 에이전트가 특정 상태에서 어떤 행동을 취할지를 결정하는 전략이다. 정책은 상태와 행동 간의 매핑으로 표현되며, 최적의 정책은 주어진 환경에서 장기적으로 최대의 보상을 얻을 수 있는 행동을 선택하는 것이다. 정책은 신경망과 같은 함수 근사기를 통해 표현될 수 있으며, 이를 통해 복잡한 환경에서도 효과적인 학습이 가능하다.

2.1.5 정책 기반 방법(Policy-based Methods)

정책 기반 방법(Policy-based methods)은 에이전트가 정책을 직접 학습하는 접근 방식이다. 정책이란 주어진 상태에서 에이전트가 취할 행동을 결정하는 전략을 의미하며, 일반적으로 확률 분포의 형태로 표현된다. 신경망은 상태를 입력받아 각 가능한 행동에 대한 확률을 출력하며, 이를 통해 에이전트가 취할 행동을 샘플링한다. 이 접근 방식은 상태 공간이 크거나 연속적인 경우에도 유연하게 적용될 수 있다는 장점이 있다.

2-2 Proximal Policy Optimization (PPO)

PPO 는 정책 경사 방법의 한계를 극복하기 위해 제안된 알고리즘으로, 정책 업데이트 시 급격한 변화를 클리핑(Clipping) 기법을 통해 이전 정책과 현재 정책이 너무 큰 차이가 나지 않도록 제한하여 안정성을 높인다. PPO 는 중요도 샘플링(Important Sampling)을 사용하여, 새로운 정책의 행동이 얼마나 이전 정책과 일치하는 지 평가한다. 또한, 여러 번의 미니 배치 업데이트를 통해 정책을 근사하며, 실험적으로 높은 성능과 안정성을 보여준다. 이 알고리즘은 우수한 성능을 입증하고 있다. PPO 를 비롯한 여러 정책 기반 방법은 환경의 모델이 불명확하거나, 상태-행동 공간이 매우 큰 경우에도 효과적으로 적용될 수 있는 장점이 있다. 또한, 연속적인 행동 공간을 다룰 수 있어 로봇 제어, 자율 주행 등 다양한 실세계 문제에 유용하게 활용될 수 있다.

2-3 PPO 기반 UAV 정찰 시나리오 성능평가

본 연구에서는 5x5 격자 환경에서 UAV 에이전트가 강화학습을 통해 세 개의 목표 지점을 탐색하는 정찰 시나리오를 PPO 알고리즘을 적용하여 구현하였다. 이 시나리오에서 UAV 의 행동을 최적화하기 위해 강화학습의 보상 함수 R을 설계하였다.

$$R = \begin{cases} +10 \; if \; s \in T & (\textit{목표 지점 탐색}) \\ -0.1 & (\textit{탐색 시간 패널티}) \\ -0.1 & (\textit{이동 시 에너지 소비 패널티}) \\ -5 \quad if \; collision & (\textit{충돌 시 패널티}) \\ -2 \quad if \; s \in V & (\textit{목표 지점 중복 탐색 패널티}) \\ +50 \; if \; T = \emptyset & (\textit{모든 목표 지점 탐색 완료 보상}) \end{cases}$$

s 는 에이전트의 현재 상태(위치)이다. a 는 에이전트의 행동이다. T 는 남은 목표 지점의 집합이다. V는 이미 방문한 위치의 집합이다. $steps$ 는 현재까지의 스텝 수이다. $collision$은 충돌 여부를 나타내는 논리 값이다. 설계된 보상 함수는 UAV 가 목표 지점을 성공적으로 탐색하면 높은 보상을 부여하고, 각 스텝마다 에너지 소비를 고려한 패널티를 추가하였다. 또한, 장애물과 충돌 시 큰 패널티를 부여하고, 이미 방문한 위치를 중복 탐색할 경우에도 패널티를 적용하였다. 이러한 보상 구조는 UAV 가 효율적이고 신속하게 목표 지점을 탐색하면서도 에너지 소비를 최소화하고 충돌을 회피하도록 유도한다. 그림 1 은 실제 랜덤으로 경로를 탐색하는 것보다 PPO 를 적용하였을 때 우수한 보상 값을 찾는 것을 확인할 수 있다.

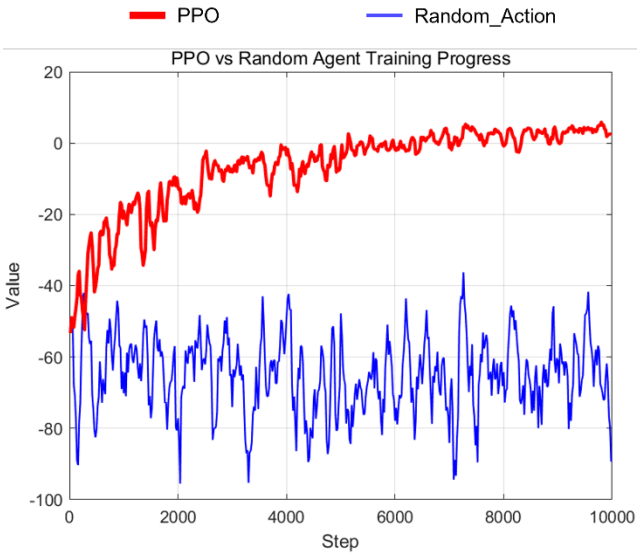

그림 1 평균 보상 그래프

Ⅲ. 결론

본 연구의 목적은 강화학습 기법을 활용하여 UAV 가 주어진 환경에서 최적의 경로를 학습하고, 정찰 임무를 효과적으로 수행할 수 있는 가능성을 탐구하는 것이다. 이를 통해 UAV 의 자율적 탐색 능력을 향상시키고, 다양한 실세계 응용 분야에서의 활용 가능성을 제시하고자 한다. 또한, 차후 연구에서는 UAV 이동 환경을 실제와 항공역학을 적용하여 더 유사한 행동과 장애 조건을 적용하여 환경을 설계할 것이고, 양자강화학습을 이용하여 이 시나리오를 적용하여 성능을 평가할 것이다.

ACKNOWLEDGMENT

본 연구는 2022 년 한국연구재단의 지원을 받아 수행됨 (NRF 2022R1A2C2004869). 본 논문의 교신저자는 김중헌임.

참 고 문 헌

[1] C. Park, G. S. Kim, S. Park, S. Jung, and J. Kim, “Multi-Agent Reinforcement Learning for Cooperative Air Transportation Services in City-Wide Autonomous Urban Air Mobility,” *IEEE Transactions on Intelligent Vehicles*, vol. 8 no. 8, pp. 4016-4030, June 2023.

[2] S. Jung, W. J. Yun, M. Shin, J. Kim, and J.-H. Kim, “Orchestrated Scheduling and Multi-Agent Deep Reinforcement Learning for Cloud-Assisted Multi-UAV Charging Systems,” *IEEE Transactions on Vehicular Technology*, vol. 70, no. 6, pp. 5362-5377, June 2021.

[3] W. J. Yun, S. Park, J. Kim, M. Shin, S. Jung, D. A. Mohaisen, and J.-H. Kim, “Cooperative Multiagent Deep Reinforcement Learning for Reliable Surveillance via Autonomous Multi-UAV Control,” IEEE Transactions on Industrial Informatics, vol. 18, no. 10, pp. 7086-7096, October 2022.

[4] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, “Policy Gradient Methods for Reinforcement Learning with Function Approximation,” in *Proc. Advances in Neural Information Processing Systems (NIPS)*, vol. 12, Denver, CO, USA, November 1999.