무인기 경로 계획 최적화를 위한 훈련 데이터셋 정제 및 모방 기반 강화학습 연구

유현석 [1]*, 이세비 [1]
[1]Big & Deep Co., Ltd.
{[1]*hsyoo, [1]leesebi}@bigndeep.co.kr

A Study on Path Planning Optimization of UAV using Refined Dataset and Imitation-based Reinforcement Learning

Yoo Hyun Suk[1]*, Lee Se Bi[1]
[1]Big & Deep Co., Ltd.

요 약

최근 국방 산업에서는 공중 감시 정찰 임무를 수행하는 무인기(Unmanned Aerial Vehicle)로 하여금 최적 정찰 경로를 생성하도록 하는 경로 계획(Path Planning) 연구가 확대되고 있다.[1] 본 논문은 무인기가 자율적으로 최단 경로를 계획하도록 훈련시키기 위한 학습 데이터 정제 방법을 제안하며, 학습의 안정성을 향상시키고자 모방학습(Imitation Learning) 기반 심층 강화학습(Deep Reinforcement Learning) 방안을 제시한다. 학습에 사용되는 데이터는 방문 지점(Way Point)이 무작위로 주어진 산점도로 정의되며, 이때 학습 개체(Agent)가 다양한 유형의 환경에서 학습할 수 있도록 산점도 내 방문 지점들이 산포된 정도에 따라 학습 환경을 k개의 유형으로 분류하는 데이터셋 정제를 수행하였다. 이후 정제된 데이터에 OR-Tools 를 적용하여 만든 경로 그래프를 모방 기반 강화학습의 훈련 데이터로 사용하였으며, 강화학습 및 제안된 방식으로 학습된 무인기가 기존 방법에 비해 단축된 경로를 발견할 수 있음을 확인하였다.

Ⅰ. 서론

무인기 기술이 발전함에 따라 국방 영역의 감시 정찰 임무 등에서 무인기 운용의 중요성이 확대되고 있다. 감시 정찰 무인기가 제한된 배터리 용량 내에서 임무를 수행하기 위해서는 상황에 따라 자율적으로 최단 경로를 결정하는 경로 계획 기법을 확보해야 하며, 이를 위해 무인기가 실험 환경과 상호작용하며 학습하는 강화학습 연구가 활발히 진행되고 있다.[2] 본 논문에서는 심층 강화학습을 사용하여 무인기의 최단 경로를 도출하는 연구를 수행하였으며, 학습에 사용될 훈련 데이터셋 정제 방법 및 모방학습을 기반으로 한 심층 강화학습 방안을 제시한다. 무인기는 시뮬레이션 환경 내에 주어진 방문 지점을 각 한 번씩 방문하고 본래 위치로 복귀하는 TSP(Traveling Sales Person) 문제의 최적 해결 방법을 학습한다. 본 연구는 기존 TSP 최적화 도구인 OR-Tools 로 제작한 경로 대비 강화학습 무인기가 생성한 경로의 평균 누적 보상을 비교하며, 강화학습을 통한 경로 계획 기법이 새로운 최단 경로를 찾아낼 가능성을 보인다.

Ⅱ. 경로 계획 최적화를 위한 모방학습 기반 강화학습

무인기는 n개의 방문 지점을 잇는 경로를 이동하며 누적된 총이동 거리 $D = \Sigma_{i=1}^{n-1}\sqrt{(x_i - x_{i+1})^2 + (y_i - y_{i+1})^2}$ 를 최소화하는 TSP 경로를 찾는 것을 목적으로 학습한다.

2.1 강화학습 환경

본 연구는 경로 계획 문제를 강화학습 방식으로 해결하기 위해 무인기의 의사결정 형태를 마르코프 결정 과정 모델 $M =< i, S, A, T, R, \gamma >$으로 정의하였다.

시작점 i 는 학습을 시작할 때 개체의 초기 위치 인덱스를 나타내며, S, A, R은 각각 환경의 상태(State) 공간, 행동(Action) 공간, 보상함수를 의미한다. γ 는 현재 상태 이후에 누적되는 보상의 적용 비율인 감가율을, T 는 이전 상태에서 다음 상태로 전이될 확률인 상태전이확률을 나타낸다.

상태 $s \in S$는 $s = [[x_1, \cdots x_n]^T, [y_1, \cdots y_n]^T]$ 로 정의한다. n 개의 방문 지점이 존재할 때 개체는 $(x_{(i\,|i\in[1,n])}, y_{(i\,|i\in[1,n])})$에서 학습을 시작하며, 방문 지점의 지상 좌표 $(x_1 \cdots x_n,\ y_1 \cdots y_n)$를 파악한 상태에서 경로를 계획한다.

개체는 주어진 상태공간 내에서 행동 $a \in A$를 취한다. j번째 노드에서 k번째 노드로 이동할 때 행동 a는 $a = [x_k - x_j,\ y_k - y_j]$로 정의된다.

개체는 다음과 같은 보상함수에 따라 학습된다.

$$R_{done} = \begin{cases} -d_{f,i} + reward, & Loop\ is\ made \\ -reward, & otherwise \end{cases}$$

$$R_{not\_done} = -\frac{\sqrt{(x_k - x_j)^2 + (y_k - y_j)^2}}{D}$$

개체는 학습 에피소드가 종료된 경우 R_{done} , 종료되지 않은 경우에는 R_{not_done} 을 보상으로 받는다. 개체는 이동할 때마다 이전 노드 위치와 현재 노드 위치의 유클리드 거리를 경로 전체 길이 D로 나눈 음의 보상을 받으며, 한 번 방문한 지점에 다시 방문할 경우 임의로 지정된 상수 페널티 $-reward$ 를 얻는다. 개체가 모든 지점을 한 번씩 방문하고 시작 지점으로 돌아올 경우, 마지막 방문 지점 $(x_f,\ y_f)$에서 시작 지점까지의 거리를 음수로 표현한 $-d_{f,i} = -\sqrt{(x_f - x_i)^2 + (y_f - y_i)^2}$ 에 $reward$를 합한 값을 보상으로 얻는다.

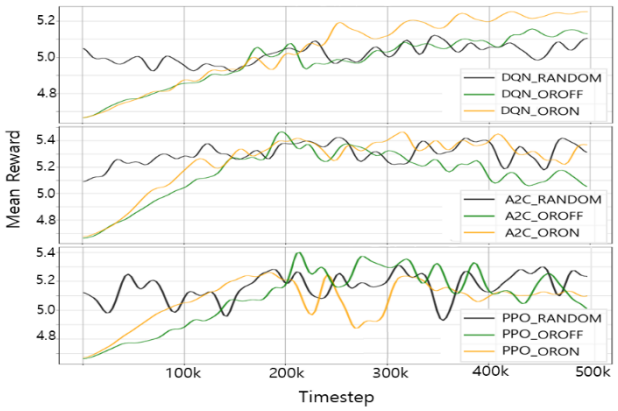

그림 1. 학습 유형별 평균 누적 보상 비교

2.2 훈련 데이터셋 정제 및 모방학습 데이터셋 구축

본 연구에서는 강화학습 개체가 다양한 유형의 경로에서 학습하도록 하기 위하여 방문 지점이 분포하는 군집 밀도 $\rho = \frac{n}{m^2}$ 에 따라 k 개의 유형으로 산점도를 분류하는 데이터셋 정제를 수행하였다. 제작된 데이터셋들의 ρ 는 유클리드 평균 거리에 따라 k 개 범위로 구분되며, 이후 같은 범위 내에 해당되는 데이터셋들이 동일한 $type \in [1, k]$을 부여받는다.

본 논문에서는 개체를 안정적으로 학습시키기 위해 모방학습 기법을 강화학습에 결합하였다. 모방학습에 사용된 훈련 데이터셋은 k 개의 유형으로 분류된 학습 데이터셋에 OR-Tools 를 적용하여 제작한 경로 그래프이며, 거리 합이 최단에 가까운 경로를 보장하는 전문가 데이터셋(Expert Dataset)이다.

Ⅲ. 모의실험

본 연구는 강화학습을 활용한 경로 계획 최적화 시뮬레이션을 진행하였으며, PPO[3], DQN[4], A2C[5] 알고리즘을 사용하여 훈련 데이터 정제 및 모방학습 적용에 따른 실험 결과 비교를 수행하였다. 또한 제안된 방식의 성능을 확인하기 위해 현재 Google 사에서 오픈소스로 제공되는 최적화 도구인 OR-Tools 를 통해 생성된 경로를 기준선(Base Line)으로 가정한다.

3.1 모의실험 설정

실험은 $100 \times 100m^2$ 환경에서 진행되었으며, 방문 지점 n은 감시 정찰 임무 상황에서 무인기의 비행 가능 시간을 고려하여 20 개로 설정하였다. 방문 지점 산점도의 유형 분류 개수인 k는 10^2, $reward$는 10^3으로 지정하였으며, 5×10^3 번의 타임스텝(timestep)동안 실험을 수행하였다.

3.2 경로 자율 계획 최적화 평가

본 논문은 제안된 방식으로 학습된 무인기의 성능을 기준선 $(y = 0)$과 비교하기 위해 타임스텝 경과에 따른 학습 유형별 평균 누적 보상 비교를 수행하였다. 그림 1 의 RANDOM 은 방문 지점들이 무작위로 선정된 위치에 분포하는 환경에서 학습된 개체를, OROFF 는 100 가지 유형으로 분류된 훈련 환경($k = 100$)에서 학습한 개체를 나타내며, ORON 은 $k = 100$ 인 전문가 데이터셋으로 모방 기반 강화학습을 수행한 개체를 의미

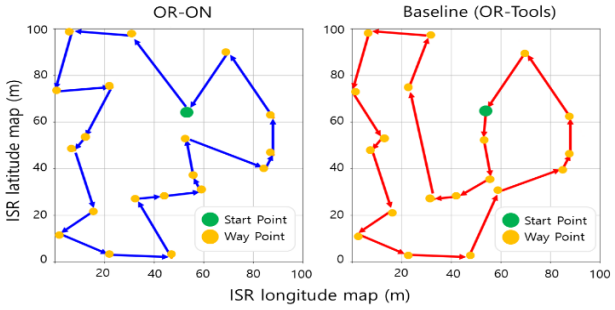

그림 2. 기준선 대비 DQN ORON 경로 비교

한다. 실험 결과 정책 기반 알고리즘인 A2C, PPO 의 경우 정제된 훈련 데이터셋을 사용하는 개체에 비해 RANDOM 유형과 모방 기반 강화학습에서 우수한 성능을 보임을 확인하였다. 반면 가치 기반 알고리즘인 DQN 의 경우 정제된 데이터셋을 사용하는 유형에 대해 안정적으로 우상향하는 모습을 보였으며, RANDOM 유형에 대해 일정한 성능을 보임을 확인하였다. 그림 2 에서는 본 과제에 대해 가장 강건한 성능을 보이는 DQN 알고리즘의 ORON 실험과 기준선 알고리즘을 통해 도출된 경로 계획을 비교하였다. DQN ORON 과 기준선의 총이동 거리는 각각 425.36m, 446.07m 로, DQN ORON 이 기존 알고리즘에 비해 단축된 거리를 발굴함을 보임으로써 제안된 방식의 유효성을 입증하였다.

Ⅳ. 결론

본 논문에서는 정찰 임무를 수행하는 무인기의 최단 경로 계획 능력을 향상시키기 위한 강화학습 연구를 진행하였다. 개체가 다양한 유형의 방문 지점 산점도에서 학습할 수 있도록 방문 지점의 밀도에 따라 훈련 환경을 분류하였으며, 모방학습 기법을 도입하여 임무 수행 개체의 안정적인 학습을 유도하고 그에 따른 비교 실험을 진행하였다. 실험 결과, 제안 방식으로 학습된 개체가 OR-Tools 에 비해 단축된 경로를 생성해 내는 양상을 확인하였으며, 이를 통해 경로 결정 과제에서 정제된 데이터셋 및 모방 기반 강화학습을 활용하여 기존 방법에 비해 개선된 경로를 도출할 가능성을 제시하였다.

ACKNOWLEDGMENT

본 연구는 대한민국 정부(방위사업청) 재원으로 국방기술진흥연구소에서 수행하는 방산혁신기업 100 전용 R&D 지원사업의 연구비 지원으로 수행된 연구임 (과제관리번호: R230106).

참 고 문 헌

[1] H. T. Kang, "Direction of operation of ROKA RPA for future warfare," Defense Policy Research, vol. 35, no. 1, pp. 7-33, 2019.
[2] Azar, Taher, et al., "Drone deep reinforcement learning: A review," Electronics, 10.9, pp. 999, 2021
[3] Schulman, John, et al., "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.
[4] Mnih, Volodymyr, et al., "Human-level control through deep reinforcement learning," Nature, 518.7540, pp. 529-533, 2015.
[5] Mnih, Volodymyr, et al., "Asynchronous methods for deep reinforcement learning," International Conference on Machine Learning, PMLR, pp. 1928-1937, 2016.