Adversarial Reinforcement Learning 기반 탐사 로봇 안정성 증강에 관한 연구

김종원[01]
[1]하나고등학교
kkcasl21331@gmail.com

Research on Enhancing the Stability of Exploration Robots Based on Adversarial Reinforcement Learning

Jongwon Kim[01]
[1]Hana Academy Seoul

요 약

본 연구는 적대적 강화학습(Adversarial Reinforcement Learning) 방식을 적용하여, 가혹한 환경에서 작동하는 탐사 로봇의 안정성을 증가시키는 방안을 제안한다. 작은 노이즈 또는 방해 요소에도 결정되는 정책의 방향성이 크게 달라질 수도 있는 변수가 존재하던 기존 방식의 강화학습과 달리, 본 연구에서는 적대적 강화학습을 이용하여 학습 과정에 적대적 요소와의 상호작용을 포함하여 학습시키고, 이를 통해 강건한 정책을 도출하게 함으로써 기존 방식의 문제점을 완화할 수 있을 것으로 예상한다.

1. 서 론

탐사 로봇은 예측 불가능한 환경에서 주로 작동하므로, 로봇의 안정성과 신뢰성은 매우 중요하다. 기존의 강화학습(RL) 방식은 로봇이 복잡한 환경에서 보상을 최대화 시키는 방향의 정책을 학습하여 경로 탐색, 장애물 회피, 작업 수행 등의 능력을 갖추도록 하는 데 있어 큰 가능성을 보여주었다. 그러나 안정되지 않고, 변수가 많은 환경에서는 작은 방해나 노이즈에도 정책의 방향이 크게 변화할 수 있어 이러한 시스템의 신뢰성이 저하될 가능성이 존재한다. 따라서, 본 연구는 적대적 강화학습(Adversarial Reinforcement Learning)을 적용하여 탐사 로봇의 안정성을 강화할 수 있는 방안을 제안한다. 적대적 강화학습은 기존 강화학습의 성능을 개선하기 위해 환경에 적대적 요소를 추가하여 학습하는 기법이다. 일반적인 강화학습에서는 에이전트가 정해진 환경 안에서 보상을 최대화하는 방향으로 학습하지만, 적대적 강화학습에서는 적대적 에이전트(Adversarial Agent)가 등장하여 학습 환경에 방해 요소를 추가하거나 환경을 예측이 어렵게 변화시킨다. 또한, 일반적인 강화학습에서는 주어진 환경에서 보상을 최대화하는 방향으로 정책을 학습하지만, 적대적 강화학습에서는 주 에이전트와 적대적 에이전트 간의 상호작용을 통해 최적화 과정이 이루어진다. 이는 게임 이론의 관점에서 극소화-극대화(min-max) 문제로 볼 수 있으며, 주 에이전트는 방해를 최소화하면서 보상을 최대화하려 하고, 적대적 에이전트는 그 반대로 방해 효과를 극대화하려 하는 방식으로 작동한다. 이를 통해 에이전트는 다양한 변동과 방해에 점차 적응하며, 강건한(robust) 정책을 학습하게 된다. 이를 통해 기존 방식에서 발생했던 문제를 완화할 수 있을 것으로 기대된다.

2. 강화학습 환경 설계

본 연구를 위한 학습 상황의 시나리오는, 미개척 지역을 탐사하는 탐사 로봇으로 가정한다. 이때, 탐사 로봇은 센서들을 기반으로 환경에 대해 정보를 수집한다. 환경을 구성하는 요소는, 목적지와 장애물이 존재한다. 또한, 에이전트는 목적지에 도달할 경우 보상(Reward)을 받으며, 탐사 도중 장애물에 충돌하거나, 제한시간 내 목적지에 도달하지 못할 경우에는 처벌(Punishment)을 받는 식으로 에피소드가 진행된다. 적대적 에이전트는 장애물의 위치 또는 크기를 바꾸는 방식으로 주 에이전트의 목적 달성을 방해한다. 이처럼 공격-수비(Attack-Defense)의 형태로, 주 에이전트가 목적지에 도달하는 것을 방해하며, 성공적으로 방해할 시에 보상을 받고, 반대로 주 에이전트가 목적지에 도달하는 경우에는 처벌을 받게 된다.

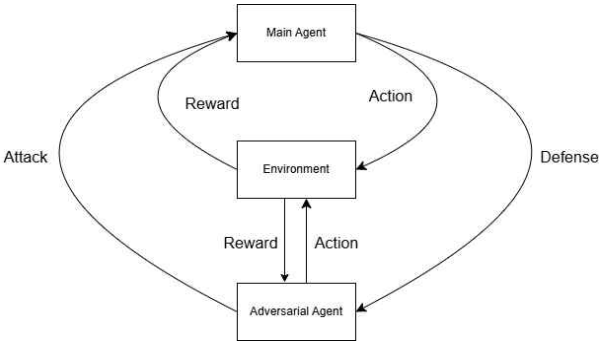

그림 1 적대적 강화학습 에피소드 구성도

그림 1과 같은 흐름으로 에피소드가 진행되며, 각 에이전트에 대한 보상함수를 설계하였다.

$$R_{main}(s,a)=\begin{cases}5-(1/T)\,(\text{if } S=true)\\ -2 \qquad\quad (\text{if } S=false \text{ or } C=true)\end{cases}$$

수식 1 주 에이전트 보상함수

$$R_{adversarial}(s,a)=\begin{cases}1\,(\text{if } S=false \text{ and } W=true)\\ 4\,(\text{if } S=false \text{ and } C=true)\\ -5\,(\text{if } S=true)\end{cases}$$

수식 2 적대적 에이전트 보상함수

두 수식에서, $R_{main}(s,a)$과 $R_{adversarial}(s,a)$는 각각 상태 s와 행동 a에 의존하는 에이전트의 보상을 뜻하며, S는 주 에이전트가 목표에 도달했는지에 대한 참/거짓, C는 주 에이전트가 장애물에 충돌했는지에 대한 참/거짓, W는 주 에이전트가 벽에 충돌했는지에 대한 참/거짓, 그리고 $T\,(T>0)$는 주 에이전트가 목표물에 도달한 후 남은 시간을 의미한다.

3. 강화학습 시뮬레이션 구현

앞서 언급한 내용들을 바탕으로, Unity ML-Agent Toolkit을 이용하여 강화학습 시뮬레이션을 구현하였다.

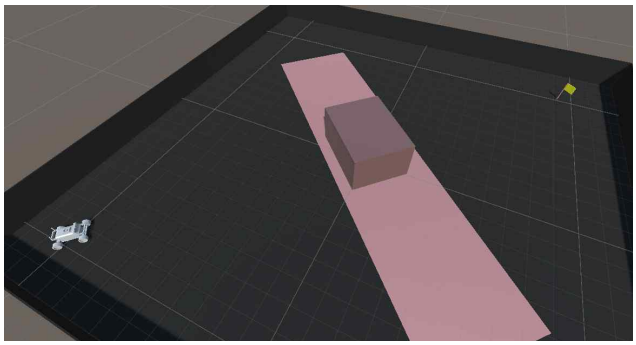

그림 2 Unity 내 학습 환경 구현

그림 2와 같이 Unity 상에 학습을 위한 환경을 구현하였다.
자동차가 주 에이전트이며, 영역 중간의 박스가 적대적 에이전트, 그리고 깃발이 목적지이다. 앞서 언급한 구조도 및 보상함수를 기반으로 모델을 학습시켰다.

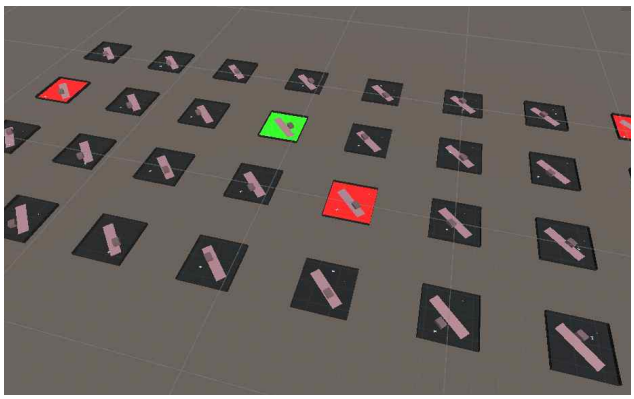

그림 3 모델 학습

학습시간 단축을 위해 그림 3과 같이 학습 영역을 복사하여 병렬적으로 배치해주었고, 하이퍼 파라미터를 조정해가며 학습을 완료하였다.

4. 결과

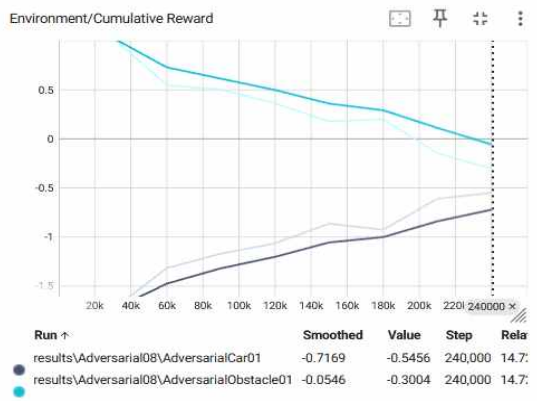

그림 4 보상 그래프

두 에이전트가 상호작용하며 학습한 결과 보상에 대한 그래프가 그림 4와 같이 도출되었다. x축은 진행한 에피소드 수, y축은 에이전트의 평균 보상을 나타낸다. 그림에서 볼 수 있듯, 두 그래프가 시간이 지남에 따라 한 지점으로 수렴하는 모습을 보여준다. 이는 두 모델이 적절한 균형으로 학습되었다는 것을 의미한다.

이후 생성된 onnx 모델을 주 에이전트와 적대적 에이전트에 적용하여 모델의 성능을 테스트하였다.

모델명 (.onnx)	Normal-06	Adversarial-08
에피소드 (회)	537	392
목표달성 (회)	92	150
성공률 (%)	17.13	38.27

표 1 모델 성능 평가

모델의 성능을 테스트하기 위해 테스트 맵을 구현하고, 학습이 완료된 onnx 모델을 이식하였다. 전체 에피소드의 횟수와 목표달성 횟수를 각각 기록하였고, 이를 바탕으로 성공률을 도출하였다. 적대적 강화학습이 포함되지 않은 기존 방식의 강화학습은 갑작스럽게 나타난 장애물에 대해 잘 대처하지 못하는 경향성을 보여주었고, 그에 반해 적대적 강화학습은 더 높은 성공률을 보여주며, 기존 강화학습보다 2배가량 상승하였다.

5. 결론 및 향후 연구

본 연구에서는 적대적 강화학습(Adversarial Reinforcement Learning) 기법을 탐사 로봇에 적용하여, 복잡하고 예측 불가능한 환경에서의 안정성을 강화하는 방안을 제안하였다. 적대적 에이전트가 제공하는 방해 요소를 통해 주 에이전트는 다양한 환경 변동에 적응하고, 극한 상황에서도 강건한 정책을 학습하게 되었다. 실험 결과, 기존의 강화학습 방식보다 높은 안정성과 강건성을 확보할 수 있었다.

향후 연구에서는 적대적 에이전트의 방해 전략을 더욱 다양화하여 실제 환경에서의 복잡성을 증가시키고, 주 에이전트가 더욱 다양한 상황에 적응할 수 있도록 하는 방안을 모색할 것이다. 에이전트 환경을 다중 에이전트(Multi-agent) 환경으로 확장하여 협력적 또는 경쟁적 상황에서의 적대적 강화학습 효과를 평가할 계획이다. 또한, 실제 센서가 장착된 하드웨어를 통해 본 시뮬레이션을 재구현하고, Sim-to-Real 기법을 이용해 실제 환경에 에이전트들을 적응시킨 뒤, 안정성을 확인하는 방법으로 연구 주제를 확장할 수 있다.

참 고 문 헌

[1] Pinto, Lerrel, James Davidson, Rahul Sukthankar, and Abhinav Gupta. "Robust Adversarial Reinforcement Learning." arXiv preprint arXiv:1703.02702 (2017).

[2] Ji, Xiang, Sanjeev Kulkarni, Mengdi Wang, and Tengyang Xie. "Self-Play with Adversarial Critic: Provable and Scalable Offline Alignment for Language Models." arXiv preprint arXiv:2406.04274 (2024).

[3] Chu, S. "Leverage of Generative Adversarial Model for Boosting Exploration in Deep Reinforcement Learning." 2024 6th International Conference on Internet of Things, Automation and Artificial Intelligence (IoTAAI), Guangzhou, China, 2024, pp. 315-318. doi: 10.1109/IoTAAI62601.2024.10692624.

[4] Rahman, Md Masudur, and Yexiang Xue. "Adversarial Policy Optimization in Deep Reinforcement Learning." arXiv preprint arXiv:2304.14533 (2023).

[5] Sun, Z., and Y. Zhu. "A Counting Method Based on Deep Reinforcement Learning Combined with Generative Adversarial Network." 2022 International Conference on Machine Learning, Cloud Computing and Intelligent Mining (MLCCIM), Xiamen, China, 2022, pp. 431-434. doi: 10.1109/MLCCIM55934.2022.00079.

[6] 권기덕, 김인철, "적대적 멀티 에이전트 환경에서 효율적인 강화 학습을 위한 정책 모델링", 정보과학회논문지 : 소프트웨어 및 응용, 제35권, 제3호, pp. 179-188, 2008.

[7] 최용찬, 박성수, "강화학습 기반 적대적 위협 환경 하에서의 정찰드론 경로 계획", 차세대융합기술학회논문지, 제6권, 제4호, pp. 624-631, 2022.