강화학습 기반 달 탐사선 랜딩 제어기 설계

김규선, 백한결, 박수현, 김중헌

고려대학교

kingdom0545@korea.ac.kr, 67back@korea.ac.kr, soohyun828@korea.ac.kr, joongheon@korea.ac.kr

Reinforcement Learning-based Lunar Module Landing Controller Design

Gyu Seon Kim, Hankyul Baek, Soohyun Park, Joongheon Kim

Korea Univ

요 약

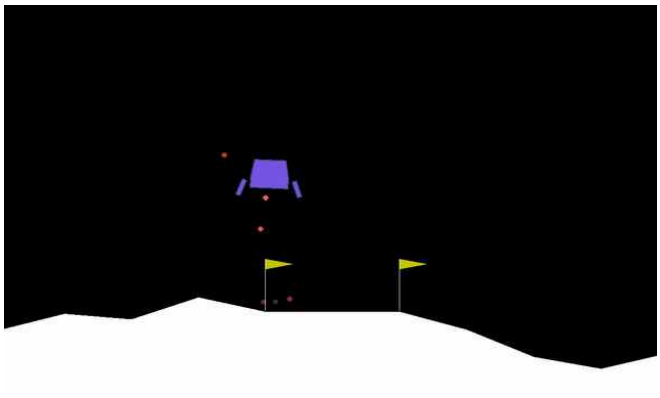
과거 냉전시대부터 이어져 온 우주개발의 활발한 진전과 함께 과학적 탐구, 기술적 혁신, 경제적 잠재력 등의 이유로, 최근 달 탐사 프로젝트에 대한 관심이 급증하고 있다. 달은 지구의 유일한 천연 위성으로써, 특히 지구 초기 역사와 태양계 형성 과정에 대한 중요한 단서를 제공해줄 뿐만 아니라, 달 내부 구조에 대한 연구는 행성 과학에 대한 인류의 지식을 확장시킬 잠재력을 가지고 있다. 이러한 배경 속에서, 달 탐사선을 안정적으로 달에 착륙시키기 위해서는 고급 착륙 제어 기술이 필수적이다. 이에 본 논문에서는 기존의 고전적인 Proportional-Integral-Differential (PID) 제어기 대신 강화학습 (Reinforcement Learning, RL)을 이용하여 달 착륙선의 제어 시스템을 설계하는 새로운 접근법을 제시한다. 제어 시스템을 설계하는 데 있어, 달 탐사선의 역학과 root locus를 분석하지 않고 fully connected layer를 활용한 인공신경망 기반 제어기 설계를 목표로 한다. 특히, value based method와 policy gradient based method의 대표적 알고리즘인 Deep Q-Network (DQN)과 Actor-Critic (AC) 그리고 Reinforce 알고리즘을 활용하여 착륙 제어 시스템의 현실적용 가능성을 실험적으로 탐색한다. 달의 물리적 환경을 묘사한 실험을 통해 해당 제어기 알고리즘의 우수성을 평가함과 동시에 우주개발과 달 탐사 분야에서 인공지능(Artificial Intelligence, AI) 기술의 적용 가능성을 시사한다.

Ⅰ. 서 론

현재의 우주개발은 지구 밖의 탐험 범위 확장 및 새로운 과학적, 기술적 도전을 실현시키고 있다. 이러한 맥락에서, 달 탐사는 여러 이유로 우주 탐사의 핵심 요소로 자리 잡고 있다. 지구의 유일한 천연 위성인 달은 지구 및 태양계의 형성과 발전에 대한 중대한 정보를 제공하며, 그 표면과 내부 구조에 대한 연구는 지구와 다른 행성들의 탄생에 대한 인류의 이해를 심화시킬 수 있다. 또한, 태양계 내외의 다른 천체를 탐험하는 데 있어 달은 전략적인 중간 기지 및 출발점으로서의 역할을 할 수 있다. 뿐만 아니라, 대기가 없는 달의 환경은 대기 간섭 없이 우주를 관측할 수 있는 독특한 천체 관측 플랫폼도 제공한다. 이러한 달의 이점을 활용하기 위해선, 달 탐사선의 정밀한 제어가 필수적이다. 역사적으로, 달 탐사선은 비례-적분-미분(Proportional-Integral-Differential, PID) 컨트롤러를 활용하여 제어되었다. 이 PID 컨트롤러는 시스템의 현재 상태와 목표 상태 간의 오차를 최소화하기 위해 출력을 조절하는데, 비례항, 적분항, 미분항을 사용한다. 매개변수 조정이 상대적으로 단순하며, 다양한 산업 분야에서 효과적인 제어 알고리즘으로 입증된 PID 컨트롤러이지만, 복잡하거나 비선형적인 시스템에서 최적의 성능을 발휘하는 데 한계를 보인다 [1]. 특히, 역학 시스템의 변화에 대응하는 데 유연성이 부족하여, 달 탐사선의 구조나 무게 등이 변화할 경우, 새로운 제어기를 재설계해야 하는 문제점이 있다. 반면, 강화학습(Reinforcement Learning, RL) 기반의 제어기는 역학 시스템의 변화에도 불구하고, 하이퍼 파라미터를 재조정함으로써 제어기의 재설계를 용이하게 한다. 강화학습에서 환경과의 지속적인 상호작용은 다양한 시나리오와 환경 변화에 대응할 수 있는 높은 유연성을 제공한다. 이는 복잡하고 예측하기 어려운 환경에서도 효과적으로 작동하여 비선형 시스템에서 강건한 제어기 설계를 가능하게 한다.

Ⅱ. 1. 달 착륙 시나리오와 제어시스템

달 탐사선의 착륙 시나리오는 달 탐사선이 달 표면에 안정적으로 착륙하기까지의 모든 과정을 포함한다. 즉, 달 탐사선의 착륙 시나리오는 단순히 달 탐사선이 달에 착륙하는 순간뿐만 아니라, 달 탐사선이 달의 중력권에 들어오고 달 표면에 안전하게 착륙하는 데까지의 모든 과정을 포함한다. 이 과정은 다음과 같은 단계로 구분된다: ⅰ) 접근 단계 : 여기서 달 탐사선은 달 궤도에 진입하고 착륙 지점 접근을 시작한다; ⅱ) 하강 단계 : 이 때 탐사선은 달의 중력에 의해 가속되며, 속도 조절과 정확한 착륙 지점 타겟팅이 중요해진다; ⅲ) 착륙 단계 : 최종 단계에서는 달 표면에 접근하며 속도를 줄이고 부드럽게 착륙한다.; 이러한 단계들을 안전하고 효율적으로 수행하기 위해서, 달 탐사선의 제어 시스템은 정밀하게 설계되어야 한다. 이 시스템에는 유도 및 항법 시스템(Guidance, Navigation and Control, GNC), 추진 시스템, 센서 및 계측 시스템, 연산 컴퓨터 등이 포함되는데, 이들은 충격으로부터 보호되어야 하기 때문이다. 특히, 달 탐사선의 경우, 지구와 달 사이의 거리로 인한 통신 지연 때문에, 상황에 따라 자율적으로 의사결정을 내릴 수 있는 능력이 요구된다. 또한 달의 예측 불가능하고 복잡한 지형에 대응하기 위해서는 정밀한 착륙 알고리즘이 필수적이다. 제어 시스템의 설계와 구현에는 정밀한 센서와 강인한 센서 융합 기술이 중요하지만, 가장 중요한 것은 정밀한 제어 알고리즘의 설계이다. 달 탐사선 환경의 불확실성과 동적 특성으로 인해, 강화학습은 달 탐사선의 착륙 문제를 해결하기 위한 유망한 방법으로 부상하고 있다. 강화학습은 동적이고 불확실한 환경에 쉽게 적응하며, 최적의 정책을 학습하는 데 효과적이다 [2]. 이러한 강화학습의 장점은 달 탐사선 착륙 시나리오의 동적이고 불확실한 환경을 쉽게 처리할 수 있어, 강화학습이 제어기 설계에 적합하다고 볼 수 있다. 본 논문에서는 기계학습의 한 분야인 강화학습을 활용하여, 달 탐사선의 착륙 제어 알고리즘을 설계하는 방법을 탐구한다.



[그림 1] 달 탐사선의 착륙 환경

Ⅱ. 2. 강화학습 - DQN, AC, Reinforce

강화학습은 환경과의 상호작용을 통해 에이전트가 누적 보상을 최대화하는 방향으로 학습을 진행하는 기계 학습의 한 방법론이다. 이 과정에서 에이전트는 다양한 상태에 따른 행동의 가치를 학습하고, 이를 극대화하기 위해 지속적으로 행동을 조정한다. 강화학습에서 에이전트와 환경 간의 반복적인 상호작용은 강화학습만의 가장 큰 특징으로, 연속적인 결정을 내려야 하는 달 탐사선의 착륙 문제와 같은, 순차적 의사결정 문제에 적합하다. 강화학습 알고리즘은 가치 기반 방법(value based method)과 정책 그래디언트 기반 방법(policy gradient based method)으로 크게 나뉜다. 본 논문에서는 가치 기반 방법의 대표적인 알고리즘인, Q-value를 신경망을 통해 학습하는 Deep Q-Learning(DQN)과 정책 그래디언트 기반 방법의 주요 알고리즘인 Actor-Critic(AC), Reinforce를 달 탐사선의 제어 시스템 설계에 적용한다. DQN은 최적의 행동 정책을 결정하기 위해 Q-value를 신경망을 통해 학습하는 반면, Actor-Critic과 Reinforce 방식은 행동 정책 자체를 직접 학습하는 접근을 취한다. 이러한 강화학습 알고리즘에서 달 탐사선은 다양한 환경 조건과 상황에 적응하며, 최적의 의사결정을 내리는 능력을 학습해 나간다. 에이전트는 알고리즘마다 정의된 손실함수를 최소화하고 목적함수를 최대화하는 방향으로 학습하며, 이는 최고의 누적 보상을 야기하는 행동 선택으로 이어진다. 본 논문의 실험에서 사용된 강화학습 알고리즘인 DQN의 손실함수, AC의 목적함수, 그리고 Reinforce의 목적함수는 각각 식 (1), 식 (2), 식 (3)으로 정의된다.

$$L(\theta)' = E[(R+\gamma \max Q_\theta(s',a') - Q_\theta(s,a))^2] \quad (1)$$

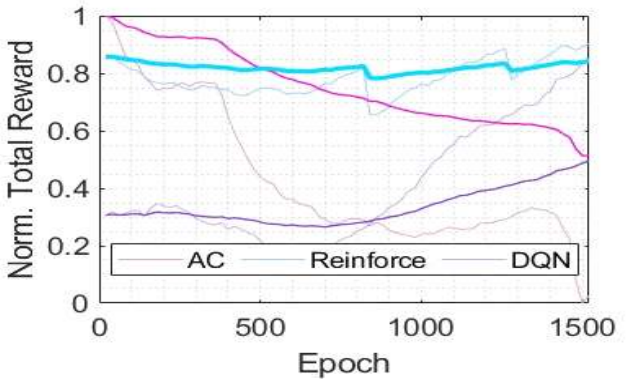
$$\nabla_\theta J(\theta) = E[\nabla_\theta \log \pi_\theta(a|s) Q_w(s,a)] \quad (2)$$

$$\nabla_\theta J(\theta) = E[\nabla_\theta \log \pi_\theta(a|s) G_t] \quad (3)$$

식 (1)에서 $L(\theta)'$, R, γ, $Q_\theta(s',a')$, 그리고 $Q_\theta(s,a)$는 각각 손실함수, 보상 함수, 감쇄비, 타겟 네트워크에 의해 추정된 Q값, 그리고 에이전트가 얻은 Q값이다. 식(2)와 (3)에서 $J(\theta)$, $Q_w(s,a)$, 그리고 G_t는 각각 목적함수, 상태-행동 가치함수, 그리고 리턴이다.

Ⅱ. 3. 달 탐사선 착륙 환경

[그림 1]은 달 착륙선의 실험 환경을 보여준다. 실험 환경은 open AI gym에서 제공해주는 'LunarLander'이다. LunarLander에서 에이전트가 관측할 수 있는 state와 action의 차원은 각각 8, 4이다. 에이전트가 관측하는 state는 착륙선의 x좌표, y좌표, 착륙선의 x축 선속도, y축 선속도, 착륙선의 각도, 각속도, 그리고 착륙선의 다리가 지면에 접촉했는지를 나타내는 2개의 bool값으로 이루어진다. 에이전트가 취할 수 있는 행동 차원 4가지로 아무것도 하지 않음, 왼쪽 엔진 분사, 메인 엔진 분사, 오른쪽 엔진 분사로 구분된다. 보상함수의 경우는, 달 탐사선이 착륙 패드로부터 멀어지거나 충돌하면 보상을 잃는다. 이와 반대로 착륙선이 지면에 잘 착지하거나 다리가 지면과 접촉하면 양의 보상을 받는다. 또한 엔진을 발사할 때 마다 소량의 음의 보상을 줘서 최소한의 엔진 분사 즉, 최소한의 에너지로 안전하게 착륙할 수 있도록 보상함수가 설계되었다.



[그림 2] 훈련에서의 보상값 추이

Ⅱ. 4. 성능평가

[그림 2]는 각 알고리즘별 epoch에 따른 정규화된 보상함수 값의 추이를 보여준다. [그림 2]를 통해 알 수 있듯, 달 탐사선의 착륙 시나리오 있어, 정책 그래디언트 기반 방법인 AC, Reinforce가 가치 기반 방법인 DQN보다 좋은 성능을 보인다. DQN은 이산적인 행동 공간에 더 적합하며, 연속적인 행동 공간에서는 성능이 떨어지는 것을 알 수 있다. 또한 DQN의 특징 중 하나인 Q-target network는 학습의 안정성을 돕지만, 빠르게 변화하는 환경에 적응하는 데는 지연을 야기할 수 있다.

Ⅲ. 결론

본 논문에서는 달 탐사선이 달에 착륙할 때 발생하는 환경의 고유한 특성으로 인해 겪는 제어시스템 문제를 강화학습을 통해 해결하였다. 달 탐사선의 착륙 시나리오에서의 동적 변화와 불확실성을 고려하여 기존의 고전적인 PID 제어기 대신 강화학습을 활용하여 달 착륙선의 제어시스템을 설계하였다. 본 논문은 달 탐사선 착륙 제어에 강화학습을 접목시키므로써, 강화학습이 전통적인 제어 시스템에 비해 가지는 유연성과 적응성을 강조하고, 이러한 접근 방식이 우주 탐사 및 개발 분야에서 중요한 발전을 이룰 수 있음을 시사한다.

ACKNOWLEDGMENT

본 연구는 정부(과학기술정부통신부)의 재원으로 정보통신기획평가원의 지원(No. 2022-0-00907, (2 세부) AI Bots 협업 플랫폼 및 자기조직 인공지능 기술 개발), 그리고 한국연구재단 기초연구실지원사업(2021R1A4A1030775)의 연구비 지원을 받아 수행된 연구임.

참 고 문 헌

[1] H. Sira-Ramírez, E. W. Zurita-Bustamante and C. Huang, "Equivalence Among Flat Filters, Dirty Derivative-Based PID Controllers, ADRC, and Integral Reconstructor-Based Sliding Mode Control," *IEEE Transactions on Control Systems Technology*, vol. 28, no. 5, pp. 1696-1710, September 2020.

[2] Y. Hu, J. Fu and G. Wen, "Safe Reinforcement Learning for Model-Reference Trajectory Tracking of Uncertain Autonomous Vehicles With Model-Based Acceleration," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 3, pp. 2332-2344, March 2023.