무보상 피지컬 AI 환경에서 역강화학습 기반 안전 경로 의사결정 기법 연구

최진철, 박찬원

한국전자통신연구원

{spiders22v, cwp}@etri.re.kr

Inverse Reinforcement Learning - Based Safe Path Decision-Making in Reward-Free Physical AI Environments

Jinchul Choi, Chanwon Park

Electronics Telecommunications and Research Institute

요 약

피지컬 AI(Physical AI) 환경은 물리적 시스템과 지능형 에이전트가 상호작용하며 실제 환경에서의 예측 및 자율 제어를 수행하는 기술로, 최근 산업 제조, 스마트 물류, 공공 안전 분야를 중심으로 그 중요성이 확대되고 있다. 그러나 안전성과 운영 효율 등 상충하는 목표가 공존하는 피지컬 AI 기반 의사결정 환경에서는 위험 요소를 포함한 명시적인 보상 함수를 사전에 정의하는 데 어려움이 따른다. 본 연구에서는 이러한 문제를 해결하기 위해 전문가의 시연 데이터로부터 최적의 보상을 추정하는 역강화학습(Inverse Reinforcement Learning, IRL) 기반의 의사결정 기술을 제안한다. 피지컬 AI 환경을 Grid 기반 상태 공간으로 모델링한 후, 전문가의 안전 경로 데이터를 활용하여 Maximum Entropy IRL을 통해 상태별 보상 함수를 복원한다. 이후 복원된 보상을 기반으로 Soft Value Iteration을 수행함으로써 위험 회피 특성이 반영된 확률적 의사결정 정책을 도출한다. 또한 복원된 보상을 히트맵(Heatmap)으로 시각화함으로써 위험 및 안전 영역을 직관적으로 해석할 수 있도록 하였다.

Ⅰ. 서 론

피지컬 AI는 물리적 시스템과 지능형 에이전트가 실제 환경에서 상호작용하며 인지·판단·행동을 수행하는 기술로, 산업 제조, 스마트 물류, 공공 안전 등 다양한 분야에서 활용이 확대되고 있다. 특히 물리 환경을 반영한 가상 모델과 시뮬레이션을 통해 위험을 최소화하고 목표를 달성하는 자율적 의사결정 기능은 피지컬 AI의 핵심 요소로 주목받고 있다[1].

그러나 실제 피지컬 AI 환경에서는 안전성, 운영 비용, 작업 효율성과 같은 상충하는 요소를 동시에 고려해야 하며, 이를 강화학습의 보상 함수로 명시적으로 정의하는 데에는 구조적인 어려움이 따른다. 특히 보상 체계가 사전에 정의되지 않은 무보상 환경(Reward-free Environment)에서는 기존 강화학습 기반 접근법으로 안전한 의사결정 정책을 도출하는 데 명확한 한계가 존재한다[2].

본 연구는 이러한 문제의식을 바탕으로 보상 설계에 대한 의존도를 낮추고, 전문가의 의사결정을 직접 반영할 수 있는 역강화학습 기반 의사결정 기술을 제안한다. 전문가 시연 데이터를 활용한 보상 추정과 이를 기반으로 한 안전 의사결정 정책 도출을 통해, 피지컬 AI 환경에서의 안전 자율 제어 가능성을 제시한다.

Ⅱ. 본론

본 연구는 피지컬 AI 환경에서 에이전트가 목표 지점에 도달하는 동시에 위험 구역(Danger Zone)의 방문을 최소화하도록 의사결정하는 문제를 다룬다. 해당 피지컬 AI 환경은 상태 집합 S, 행동 집합 A, 상태 전이 함수 T, 보상 함수 R로 구성된 마르코프 결정 과정(Markov Decision Process, MDP)로 모델링될 수 있다. 그러나 실제 환경에서는 안전성, 위험도, 비용 등 복합적인 요소로 인해 보상 함수 R가 명시적으로 주어지지 않는 경우가 많다. 이에 따라 본 연구에서는 보상이 사전에 정의되지 않은 무보상(Reward-free) 환경에서 전문가 행동을 설명하는 잠재 보상을 추정하고, 이를 기반으로 안전 의사결정 정책을 도출하는 문제를 다룬다.

피지컬 AI 환경을 2D Grid로 표현하면 상태 공간은 다음과 같다.

$$S = (x, y) | 0 \le x < W,\ 0 \le y < H,\ |S| = W \times H. \quad (1)$$

여기서 x와 y는 각각 grid 상의 가로 및 세로 좌표를 나타내며, W와 H는 환경의 폭과 높이를 의미한다. 각 상태는 grid의 하나의 셀(cell)에 해당한다. 행동 공간 A는 상, 하, 좌, 우의 네 가지 이동으로 정의한다. 상태 전이는 결정적 전이(deterministic transition)를 가정하며, 장애물 또는 맵의 경계로 이동을 시도할 때는 상태가 유지된다. 전문가 시연 데이터는 상태 시퀀스로 구성된 궤적(trajectory) 집합으로 정의한다.

$$\tau_E = \{s_0, s_1, \ldots, s_T\} \quad (2)$$

본 연구의 목적은 전문가의 행동을 설명할 수 있도록 에이전트가 목표 지점에 도달하도록 유도하는 잠재 보상 함수 $R_\theta(s)$를 추정하고, 추정된 보상에 기반하여 상태 s에서 행동 a를 선택하는 의사결정 정책 $\pi(a \mid s)$를 도출하는 것이다.

그림 1은 제안 방법의 전체 흐름을 나타낸다. 전문가 궤적에서 상태 방문 분포를 계산하고(MaxEnt IRL)[3], 추정된 보상으로 Soft Value Iteration[4]을 수행해 확률적 정책을 산출한다. 이때 학습된 보상은 히트맵으로 시각화하여 위험 및 안전 영역을 직관적으로 해석할 수 있도록 한다.

MaxEnt IRL은 전문가 궤적이 높은 확률로 생성되도록 하는 보상 파

그림 1. 제안한 IRL 기반 무보상 안전 의사결정 절차

라미터 θ를 추정한다. 전문가 궤적 확률은 다음과 같이 정의된다.

$$P(\tau|\theta) \propto \exp\left(\sum_t R_\theta(s_t)\right) \quad (3)$$

학습은 전문가 상태 방문 분포 μ_E와 현재 정책이 유도하는 상태 방문 분포 μ_θ의 차이를 줄이는 방향으로 이루어진다.

$$\nabla_\theta \propto (\mu_E - \mu_\theta) \quad (4)$$

즉, 전문가가 자주 방문한 상태는 상대적으로 높은 보상으로, 회피한 위험 구역은 낮은 보상으로 형성되도록 θ를 반복 갱신한다.

추정된 R_θ가 주어졌을 때, 본 연구는 불확실성이 존재하는 피지컬 AI 환경에 적합한 확률적 정책을 위해 Soft Value Iteration을 사용한다. Soft Bellman backup은 다음과 같다.

$$V(s) = \log \sum_a \exp(Q(s,a)), \quad Q(s,a) = R_\theta(s) + \gamma \sum_{s'} T(s,a,s')V(s') \quad (5)$$

정책은 softmax 형태로 도출된다.

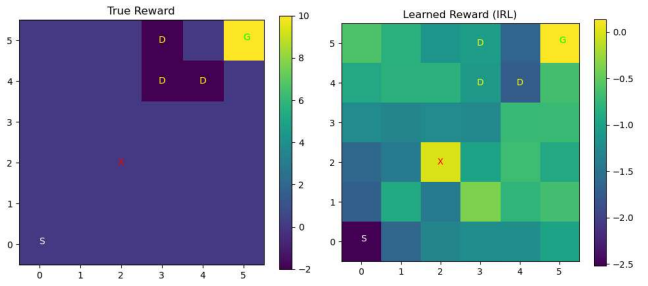
$$\pi(a \mid s) = \frac{\exp(Q(s,a))}{\sum_{a'} \exp(Q(s,a'))} \quad (6)$$

이와 같은 정책 구조는 위험 구역 회피를 유도하면서도, 환경 변화에 대해 과도하게 결정론적이지 않은 피지컬 AI 특성에 적합한 안전 의사결정 정책을 제공한다.

학습된 보상 R_θ는 상태별 위험도를 내재적으로 표현하므로, 2D Grid 환경에서 히트맵으로 시각화할 수 있다. 이를 통해 위험 영역 주변의 음의 보상 분포와 목표 지점 인근의 양의 보상 집중 현상을 관찰할 수 있으며, 이러한 보상 분포는 피지컬 AI 환경에서 에이전트의 의사결정 근거를 설명 가능하게 만드는 중요한 단서를 제공한다.

Ⅲ. 실험 결과 및 분석

실험 환경은 피지컬 AI 환경을 단순화한 6×6 GridWorld 기반 물리 공간 모델로 구성되며, 시작점(Start), 목표점(Goal), 장애물(Obstacle), 위험 구역(Danger Zone)을 포함한다. 전문가 정책은 평가 기준으로 사용되는 참 보상(True Reward)에 기반하여 안전 경로를 생성하며, 역강화학습(IRL)은 해당 보상 정보를 사전에 알지 못한 상태에서 전문가 시연 궤적만을 입력으로 학습을 수행한다. 학습을 통해 추정된 보상 함수는 Soft Value Iteration을 적용하여 확률적 의사결정 정책으로 변환된다.



(a) 설계자가 정의한 참 보상(True Reward) 히트맵 (b) IRL을 통해 학습된 보상(Learned Reward) 히트맵

그림 2. 무보상 환경에서 전문가 데모로부터 복원된 보상 구조 비교

그림 2는 설계자가 정의한 참 보상과 IRL을 통해 학습된 보상(Learned Reward)을 히트맵 형태로 비교한 결과를 보여준다. 참 보상은 목표 지점에 양의 보상과 위험 구역에 음의 보상을 명시적으로 부여한 구조로, 본 연구에서는 성능 평가 기준으로만 사용된다. 반면 학습된 보상은 전문가 시연 데이터만을 기반으로 추정된 결과로, 목표 지점 주변에는 높은 보상

그림 3. 무보상 환경에서 전문가 데모로부터 복원된 보상 구조 비교

이 형성되고 위험 구역 인근에는 낮은 보상이 분포하는 특성을 보인다. 이는 전문가의 목표 지향적이면서도 위험 회피적인 행동 패턴이 보상 공간에 자연스럽게 반영되었음을 의미한다. 또한 학습된 보상은 연속적인 분포 형태로 나타나며, 이는 상태 방문 빈도와 확률적 정책 학습의 영향에 따른 결과이다.

보상 복원 결과가 실제 의사결정 성능으로 이어지는지를 검증하기 위해, 그림 3에서는 전문가(Expert) 정책과 IRL 정책의 성능을 정량적으로 비교하였다. 성능 평가는 누적 보상(Return), 목표 도달률(Success Rate), 위험 구역 방문 비율(Danger Rate)을 기준으로 수행하였다. 실험 결과, IRL 정책은 전문가 정책과 거의 동일한 누적 보상을 달성하였으며, 목표 도달률 또한 두 정책 모두 동일하게 나타났다. 이는 IRL이 전문가 수준의 목표 지향적 의사결정 정책을 성공적으로 학습했음을 의미한다.

특히 주목할 점은 위험 구역 방문 비율(Danger Rate)이다. IRL 정책은 전문가 정책에 비해 위험 구역 방문 비율이 더 낮게 나타났으며, 이는 IRL이 위험 상태에 대해 보다 보수적인 의사결정 전략을 학습했음을 시사한다. 이러한 결과는 Soft-RL 기반 확률적 정책의 특성과 IRL 학습 과정에서 형성된 보상 기울기(Reward Gradient)의 영향으로, 위험 상태에 대한 회피 성향이 강화된 것으로 해석할 수 있다.

이러한 결과를 통해 IRL은 무보상 피지컬 AI 환경에서도 전문가의 안전 지향적 의사결정 특성을 효과적으로 복원할 수 있으며, 복원된 보상을 기반으로 목표 달성과 위험 회피를 동시에 고려하는 안정적인 안전 경로 의사결정이 가능함을 확인하였다.

Ⅳ. 결론

본 논문에서는 무보상 피지컬 AI 환경에서 전문가 시연을 활용한 역강화학습 기반 안전 의사결정 기법을 제안하고, 그 유효성을 검증하였다. 실험 결과, 제안한 방법은 피지컬 AI 에이전트가 따르는 전문가의 위험 회피 및 목표 지향적 판단 기준을 잠재 보상 함수 형태로 효과적으로 복원할 수 있음을 확인하였다. 또한 복원된 보상을 기반으로 학습된 정책은 전문가 정책과 유사한 수준의 누적 보상과 안정적인 목표 지향성을 보였다. 특히 위험 구역 방문률이 전문가 정책보다 낮게 나타나, 명시적인 보상 설계 없이도 피지컬 AI 환경에서 안전 중심의 의사결정이 가능함을 확인하였다.

ACKNOWLEDGMENT

본 연구는 2026년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 2022-0-00438, (총괄1세부)지능형 디지털 트윈 연합 운용 및 예측 핵심기술 개발)

참 고 문 헌

[1] D. Liu et al, Generative Physical AI in Vision: A Survey, arXiv:2501.10928v2 [cs.CV], Apr. 2025. doi:10.48550/arXiv.2501.10928.

[2] C. Jin et al. (2020). Reward-free exploration for reinforcement learning. Proceedings of the 37th International Conference on Machine Learning (ICML 2020), PMLR 119:4870-4879.

[3] B. Ziebart et al. (2008). Maximum entropy inverse reinforcement learning. In AAAI Conference on Artificial Intelligence, vol. 8, pp. 1433-1438.

[4] T. Haarnoja et al (2017) Reinforcement Learning with Deep Energy-Based Policies. International Conference on Machine Learning (ICML), 1352-1361.