자율 이동 로봇 제어를 위한 강화학습-대규모 언어 모델 통합

채송화 [1], 성다훈 [2], 임유진 [3]
[1] 숙명여자대학교 IT 공학과 석사과정
[2] 숙명여자대학교 IT 공학과 박사과정
[3] 숙명여자대학교 인공지능공학부 교수

watermelon97@ sookmyung.ac.kr, ekgns324@sookmyung.ac.kr, yujin91@ sookmyung.ac.kr

Reinforcement Learning-Large Language Model Integration for Autonomous Mobile Robot Control

SongHwa Chae[1], Da-Hun Seong[2], Yujin Lim[3]
[1,2]Dept. of Information Technology Engineering, Sookmyung Women's University
[3]Div. of Artificial Intelligence Engineering, Sookmyung Women's University

요 약

다양한 분야에서 효율적인 물류 관리와 작업 자동화에 대한 수요가 증가함에 따라 자율 이동 로봇(AMR, Autonomous Mobile Robot)의 연구 및 활용이 활발히 이루어지고 있다. 자율 이동 로봇에 적용되는 강화학습 모델은 연속 제어 문제에서 안정적인 성능을 보이지만, 보상 함수가 고정적으로 설계되어 있어 환경 변화나 동적 상황에 유연하게 대응하기 어렵다는 한계가 존재한다. 이 한계를 보완하기 위해 본 연구에서는 대규모 언어 모델을 활용하여 학습 과정에서 주기적으로 성능 로그를 분석하고, 보상 스케일과 잠재 함수 가중치를 동적으로 조정하는 강화학습-대규모 언어 모델 통합 모델을 제안한다.

1. 서론

최근 효율적인 물류 관리와 작업 자동화에 대한 수요가 증가함에 따라 AMR(Autonomous Mobile Robot)에 대한 연구가 활발히 이루어지고 있다. AMR 은 제조 산업 뿐만 아니라 병원, 사무실, 식당과 같은 다양한 환경에서 배송, 청소, 안내와 같은 업무에 활용되며 이러한 동적인 환경에서는 불필요한 우회와 충돌을 최소화 하며 효율적인 주행을 하는 능력이 요구된다 [1].

하지만, PID 와 MPC 등의 전통적인 제어 기법은 환경 변화에 취약하거나, 빠르게 변화하는 환경에서 발생하는 높은 연산 비용 문제는 AMR 의 성능 저하로 이어질 수 있어 그 대안으로 강화학습이 제어 기법으로 제안되고 있다[2]. 그러나 강화학습 또한 충분한 환경 이해가 부족하거나, 강화학습이 취하는 행동에 대한 논리적 해석이 어렵고, 보상 함수 설계가 인간에 크게 의존하여 보상 해킹에 취약하다는 한계가 존재한다[3].

이러한 문제를 해결하기 위해 본 연구는 동적인 환경에서 관측 데이터 기반 맥락 추론 보상 설계를 자동화하는 강화학습-대규모 언어 모델 하이브리드 모델을 제안하고 독립 강화학습 모델과 비교 분석하여 AMR 제어에 대규모 언어 모델을 통합하는 방안을 탐구한다.

2. 관련 연구

AMR 분야에서는 연속 제어를 위한 강화학습 기법이 지속적인 주목을 받고있다. [4]는 AMR 제어 기법으로 PPO(Proximal Policy Optimization)를 적용하여 강화된 신경망 구조와 보상함수 설계를 통해 제어 기법의 성능을 개선하고, 복잡한 환경에서의 자율주행 실험을 통해 충돌 없이 효율적 경로 탐색을 수행하여 지정된 목표까지의 주행을 최적화하였다. [5]는 고정된 실내 환경에서 자율 주행 로봇의 경로 탐색 및 주행 목표를 달성하기 위해 환경 정보 없이 심층 강화학습을 통해 장애물을 회피하고 목표 지점에 도달하는 성능을 입증함과 동시에 보상 함수 설계의 변화가 에이전트의 행동과 성능에 큰 영향을 미친다는 것을 시사

하였다.

그러나 대부분의 선행 연구는 정적 환경만을 고려하여 동적 환경 시나리오의 성능 전이 및 환경 변화 적응에 제한적이다. 이를 보완하기 위해 동적 환경을 고려하는 연구가 주목받고 있다. [6]은 경로 계획 성능을 발휘하기 어려운 좁은 통로나 동적 요소가 혼재된 환경에서 기존 알고리즘의 환경 변화 대응 속도와 계산 효율성이 떨어짐에 대해 지적하며 이를 해결하기 위해 GAP SAC(Gated Attention Prioritized Experience Reply Soft Actor-Critic) 알고리즘을 제안하였다. 해당 알고리즘은 상태 표현을 확장하고 동적인 휴리스틱 보상을 제공하며 학습 효율과 환경 인식 능력을 높여 변동이 심한 실제 환경에서도 주행 가능함을 입증한다. 또한 [7]은 동적인 환경에서 이동 로봇의 로컬 경로 계획 문제를 해결하기 위해 심층 강화학습과 전통 경로 계획 알고리즘을 결합한 새로운 방식을 제안하여 로봇의 실시간 장애물 회피와 진행 방향, 속도 등을 포함한 보상 함수 설계를 통해 로봇의 안정성과 효율성을 높여 동적 환경에서도 안정적이고 매끄러운 주행 경로 생성 가능성을 보였다.

뿐만 아니라 대규모 언어 모델 연구는 기존 강화학습이나 전통 알고리즘만으로는 한계가 있던 영역에서 새로운 가능성을 제공하고 있다. [8]은 대규모 언어 모델을 활용하여 자연어 명령을 동적으로 해석하고 웨이 포인트로 변환하는 경로 계획 프레임워크를 제안하며 복잡한 환경에 적용할 수 있는 가능성을 제시한다. [9]는 복잡한 동적 환경 속에서 모바일 로봇이 사회적인 맥락을 고려한 안전 경로를 실시간으로 계획할 수 있도록 대규모 언어 모델 기반의 멀티 모달 센서 융합 프레임워크를 제공하며 LiDAR 와 RGB 카메라 데이터를 FPGA 에서 실시간으로 융합하고 경로 우선순위를 동적으로 조정해 보행자 예측 오차를 줄이는 등 인간 친화적인 경로 계획이 가능함을 시사한다.

이와 같이 선행 연구들은 강화학습을 통한 자율 이동 로봇 제어의 가능성을 보여주었으나, 고정된 보상 함수로 인해 동적인 환경에 유연하게 대응하지 못한다는 점은 중요한 한계로 지적된다. 본 연구에서는 대규모 언어 모델이 가진 맥락 이해와 적응적 추론 능력을 활용하여 보상 함수를 조정함으로써 동적 환경에서의 이동 안정성과 효율성을 향상시키는 방안을 탐구한다.

3. 방법론

본 연구는 자율 이동 로봇의 경로 계획 문제를 MDP(Markov Decision Process)로 정의하고, 기본 제어 기법으로 PPO 을 적용하였다.

$$\mathcal{L}^{clip}(\theta) = E_t\left[min\left(r_t(\theta)\widehat{A_t}, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\widehat{A_t}\right)\right]$$

PPO 는 정책 경사 기반 강화학습 알고리즘으로, 정책 업데이트 시 클리핑 기법을 사용하여 새로운 정책과 기존 정책 간의 변화 폭을 제한하여 급격한 성능 저하나 정책 붕괴를 방지하고, 연속 제어 과제에서 안정적인 학습을 가능하게 한다. 이러한 특성은 환경 변화가 잦고 불확실성이 높은 문제에서 특히 유용하게 활용된다[10].

3.1 PPO 단독 모델

PPO 단독 모델은 로봇의 현재 위치, 목표 좌표, 목표까지의 상대 벡터, 5 개 방향 거리 센서 값, 충돌 여부, 로컬 점유 그리드, 그리고 로봇의 방향 각도까지 포함한 122 차원 상태 벡터를 입력으로 사용한다. CNN 기반 네트워크는 점유 그리드에서 공간적인 특징을 추출하고, MLP(Multi-Layer Perceptron)은 수치형 상태 벡터를 처리하여 두 특징을 통합한다. 이를 통해 에이전트는 주변 환경을 정밀하게 인식하고, 최적의 행동을 선택하도록 학습된다. 최종적으로 Actor-Critic 구조를 통해 다섯 가지 행동(전진, 후진, 정지, 좌회전, 우회전)과 상태 가치를 산출한다.

3.2 대규모 언어 모델-강화학습 결합 모델

본 연구에서 제안하는 결합 모델은 PPO 단독 모델의 구조를 기반으로 대규모 언어 모델 기반 보상 함수 조정 모듈을 추가한 것으로, 학습 과정에서 에피소드 단위로 수집되는 성공률, 충돌 횟수, 목표 도달 시간, 평균 보상 등을 바탕으로 일정 간격마다 대규모 언어 모델(GPT-4o-mini)에 질의한다. 대규모 언어 모델은 수집된 로그를 바탕으로 보상 스케일과 세 가지 잠재 함수 가중치를 조정한다. 만약, 대규모 언어 모델이 제안한 값이 안전 기준을 벗어나면 20% 범위 내로 클리핑 되며, EMA(Exponential Moving Average) 기반 스무딩을 통해 안정성을 보장한다. 이러한 절차를 통해 모델은 환경 변화나 학습 정체 상황에서도 보상 함수를 적응적으로 최적화 할 수 있다.

3.3 보상함수 설계

PPO 단독 모델에서의 보상 함수(R_{base})는 이동 보상, 충돌 페널티, 목표 도달 보상, 시간 효율성 보상으로 구성된다.

$$R_{total} = R_{base} + \beta\big(\Phi(s') - \Phi(s)\big)$$

PPO 단독 모델의 보상 함수 R_{base}를 기반으로 대규모 언어 모델이 설계한 보상의 조정 항 β를 목표와

의 거리, 목표 방향과의 정렬 정도, 장애물과의 간격을 반영하는 잠재 함수 Φ로 구성된다. 대규모 언어 모델은 로그를 분석하여 가중치를 조정하여 목표 다양성을 우선시하면서 충돌 감소, 시간 단축 순으로 학습 목표를 세분화 하여 고정된 보상 함수가 가지는 한계를 극복하고 환경 변화가 잦은 환경에서도 안정적인 정책 학습을 가능하게 한다.

4. 실험 및 결과

본 연구에서 제안하는 강화학습과 대규모 언어 모델 통합 모델은 기존 PPO 단독 모델 구조에 적응적인 보상 조정 모듈을 추가한 것으로, 학습 과정에서 성능 저하가 발생할 경우 대규모 언어 모델이 개입하도록 설계되었다.

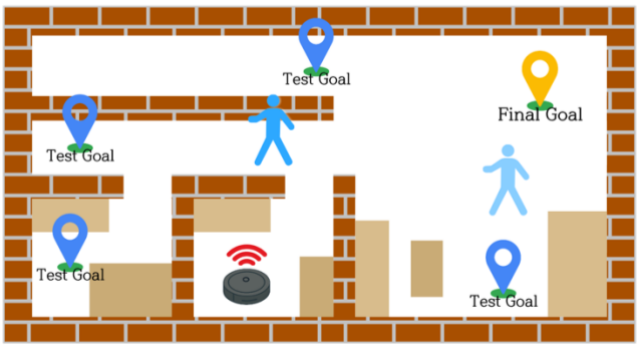
<표 1> 대규모 언어 모델 연동 보상 조정 개입

Algorithm 1 Conditional LLM Intervention

Input: Episode number e, Performance statistics S, Current parameters θ
Output: Updated parameters θ'

```
if e mod 100 ≠ 0 then return θ
if S.success_rate ≥ 0.6 AND S.avg_collisions ≤ 15 then return θ
if S.success_rate < 0.5 OR S.avg_collisions > 25 then
    context ← BuildContext(S, θ, goal_statistics)
    try
        llm_response ← QueryGPT4oMini(context, temperature=0.2, max_tokens=500)
    catch
        llm_response ← HeuristicFallback(context, θ)
    θ_proposed ← ExtractParameters(llm_response)
    θ_safe ← Clip(θ_proposed, 0.8θ, 1.2θ)
    θ' ← 0.3θ_safe + 0.7θ
    return θ'
return θ
```

시스템은 100 에피소드 단위로 성공률, 충돌 횟수, 평균 시간, 평균 보상과 같은 성능 지표를 점검하고 해당 지표가 일정 기준 이하일 경우 GPT-4o-mini 에 질의를 수행한다. 이 때 반환되는 보상 조정 값 β는 20% 이내로 제한되며, EMA 스무딩 기법을 거쳐 안정성을 확보한다. 이를 통해 대규모 언어 모델이 불필요하게 자주 호출되는 것을 줄이고 필요한 시점에만 조언을 반영하여 학습 효율을 높인다.



(그림 1) 학습 및 평가 실험 환경

실험은 15m × 9m 크기의 실내 맵에 정적 장애물과 동적 장애물이 혼재하도록 Gym 환경을 구축하여 수행되었다. AMR 에이전트는 네 개의 고정된 목표를 순차적으로 학습한 후, 학습되지 않은 무작위 목표에서 일반화 성능을 검증하였다.

<표 2> 에피소드 별 성능 분석

Episode	Method	Success Rate (%)	Avg. Collisions	Avg. Time (s)	Avg. Reward
1-20	LLM-PPO	100.0 ± 0.0	9.2 ± 9.0	216.6 ± 173.9	14,650 ± 3,265
	PPO	95.0 ± 4.9	10.1 ± 20.6	210.0 ± 215.6	14,403 ± 4,630
	Diff	+5.3%	-8.9%	+3.1%	+1.7%
21-40	LLM-PPO	100.0 ± 0.0	8.0 ± 6.5	193.2 ± 91.0	14,473 ± 1,726
	PPO	100.0 ± 0.0	11.4 ± 7.9	251.5 ± 161.0	13,590 ± 2,846
	Diff	0.0%	-29.8%	-23.2%	+6.5%
41-60	LLM-PPO	90.0 ± 6.7	7.2 ± 5.7	208.2 ± 105.9	14,394 ± 1,998
	PPO	100.0 ± 0.0	11.4 ± 8.1	261.7 ± 168.0	13,611 ± 2,953
	Diff	-10.0%	-36.8%	-20.4%	+5.8%
61-80	LLM-PPO	95.0 ± 4.9	6.5 ± 4.7	195.9 ± 81.2	14,540 ± 1,552
	PPO	95.0 ± 4.9	11.3 ± 7.8	253.8 ± 160.3	13,657 ± 2,817
	Diff	0.0%	-42.5%	-22.8%	+6.5%
81-100	LLM-PPO	90.0 ± 6.7	7.9 ± 6.3	212.3 ± 93.6	14,365 ± 1,735
	PPO	90.0 ± 6.7	17.7 ± 24.3	351.8 ± 242.1	12,120 ± 4,465
	Diff	0.0%	-55.4%	-39.6%	+18.5%

실험 결과는 에피소드를 거듭함에 따라 PPO 단독 모델보다 강화학습-대규모 언어 모델 통합 모델의 성능이 점진적으로 개선됨을 확인할 수 있다. 초기 구간(1-20 에피소드)에서는 두 방법 간 차이가 미미했으나, 중반 구간(21-80 에피소드)에서는 충돌 횟수가 최대 43%까지 감소하고, 목적지 도달 시간은 최대 23% 단축되는 등 유의미한 개선을 보였다. 후반부에서 이 격차는 더욱 커져 충돌이 약 55%까지 감소하고, 목적지 도달 시간이 약 40% 빨라졌으며, 평균 보상도 약 19% 향상됨을 확인할 수 있다.

전체 에피소드에 대한 성능을 종합적으로 살펴보면, 모든 지표에서 일관적이고 예측 가능한 성능을 달성했음을 확인할 수 있다. 제안된 강화학습-대규모 언어 모델 통합 모델의 성능 향상은 단순한 보상 조정을 넘어 대규모 언어 모델이 제공하는 맥락 이해와 적응적 추론이 강화학습 과정에 반영된 결과로 추측된다. 이를 통해 환경 변화에 따른 대규모 언어 모델 기반 동적 보상 설계가 강화학습의 효율성을 높일 수 있음을 시사한다.

5. 결론 및 향후 연구

본 연구는 AMR 제어를 위한 대규모 언어 모델과 강화학습을 통합한 하이브리드 모델을 제안한다. 환경 및 에이전트의 성능 맥락에 따른 동적 보상 조정을 통해 대규모 언어 모델-강화학습 결합 모델은 기존의 PPO 단독 모델 대비 평균 충돌 횟수를 38% 감소시키고, 목적지 도달 시간을 21% 단축시켰으며, 평균 보상을 8% 향상시켰다.

하지만 단일 환경에서만 수행된 실험이고, 다양한 환경과 다양한 장애물이 존재하는 공간에서의 강건성은 검증되지 않았다. 뿐만 아니라, 대규모 언어 모델의 반복 질의에 따른 출력 결과가 일관성이 있는지 측정되지 않았다.

따라서 향후 연구에서는 다양한 공간적 토폴로지를 포함한 다중 환경 검증을 통해 제안 기법의 일반화 가능성을 확인할 필요가 있다. 또한 대규모 언어 모델 출력의 일관성을 정량적으로 평가하고, 실제 실험을 고려한 센서 노이즈와 같은 현실적 제약 조건을 실험 환경으로 조성하여 현실 적용 가능성을 탐색하고 검증해야 한다. 더 나아가 다른 강화학습 알고리즘과의 호환성 검토 및 멀티 모달 융합 확장을 통한 제안 모델의 실용성과 범용성을 강화할 수 있을 것으로 기대된다.

사사문구

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원 학·석사연계 ICT 핵심인재양성 지원을 받아 수행된 연구임(IITP-2025-RS-2022-00156299)

참고문헌

[1] T. Lackner, J. Hermann, C. Kuhn, and D. Palm, "Review of autonomous mobile robots in intralogistics: state-of-the-art, limitations and research gaps," *Procedia CIRP*, vol. 130, pp. 930-935, 2024.

[2] S. Akki and T. Chen, "Benchmarking model predictive control and reinforcement learning based control for legged robot locomotion in MuJoCo simulation," *IEEE Access*, 2025.

[3] J. Ni, F. Li, X. Jin, X. Peng, and Y. Tang, "Reinforcement learning based constrained optimal control: an interpretable reward design," *arXiv preprint arXiv:2502.10187*, 2025.

[4] Taheri, S. R. Hosseini, and M. A. Nekoui, "Deep reinforcement learning with enhanced PPO for safe mobile robot navigation," *arXiv preprint arXiv:2405.16266*, 2024.

[5] Quinones-Ramirez, J. Rios-Martinez, and V. Uc-Cetina, "Robot path planning using deep reinforcement learning," *arXiv preprint arXiv:2302.09120*, 2023.

[6] Z. Zhang, H. Fu, J. Yang, and Y. Lin, "Deep reinforcement learning for path planning of autonomous mobile robots in complicated environments," *Complex & Intelligent Systems*, vol. 11, no. 6, article 277, 2025.

[7] B. Tao and J. H. Kim, "Deep reinforcement learning-based local path planning in dynamic environments for mobile robot," *Journal of King Saud University-Computer and Information Sciences*, vol. 36, no. 10, article 102254, 2024.

[8] M. T. Tariq, Y. Hussain, and C. Wang, "Robust mobile robot path planning via LLM-based dynamic waypoint generation," *Expert Systems with Applications*, vol. 282, article 127600, 2025.

[9] X. Liu, A. Farid, R. Ukyoh, T. Amano, H. Rizk, and H. Yamaguchi, "LLM-Driven Adaptive Autonomous Robot Navigation via Multimodal Fusion for Diverse Environments," in *2025 IEEE Intelligent Vehicles Symposium (IV)*, June 2025, pp. 2361–2368.

[10] W. Zhang, L. Shan, L. Chang, W. Wang, and Y. Dai, "HDCPO: A PPO-based path following and obstacle avoidance method for USV considering environmental disturbances," *Advanced Engineering Informatics*, vol. 68, p. 103735, 2025.