실시간 다중 로봇 제어를 위한 Mamba 기반 아키텍처의 확장성 분석

윤정란, 박수현
숙명여자대학교
yojr3001@sookmyung.ac.kr, soohyun.park@sookmyung.ac.kr

A Scalability Analysis of Mamba-Based Architecture for Real-Time Multi-Robot Task Allocation

Jungran Yoon, Soohyun Park
Sookmyung Women's Univ.

요 약

본 논문은 Mamba 기반의 아키텍처를 통해 Multi-Robot Task Allocation (MRTA) 문제에서 발생하는 계산 복잡도와 실시간 의사결정의 병목 현상을 해결하고자 하였다. 기존 Transformer 기반 접근법은 이차적 시간 복잡도로 인해 로봇 규모 확장에 한계를 보였다. 이에 입력 시퀀스 길이에 대해 선형적인 시간 복잡도를 갖는 선택적 State Space Model (SSM)을 MRTA에 도입하여 한계를 구조적으로 개선하였다. 확장성 검증을 위한 실험에서 제안하는 모델이 로봇의 수 증가에 따라 완만한 추론 속도 저하를 보이며 대규모 MRTA에서 제안 모델이 안정적인 실시간 할당을 수행할 수 있음을 확인하였다.

Ⅰ. 서 론

다중 로봇 시스템의 효율성은 주어진 임무를 로봇에게 최적으로 배분하는 Multi-Robot Task Allocation (MRTA) 기술에 달려 있다. MRTA는 대표적인 NP-hard 조합 최적화 문제로, 로봇 수가 증가함에 따라 탐색 공간이 기하급수적으로 확장되는 특성을 갖는다[1]. 최근 Transformer 기반 강화학습 모델이 뛰어난 성능을 보였으나[2], Self-Attention 메커니즘의 이차적 계산 복잡도는 로봇 규모 확장에 따른 실시간 의사결정의 심각한 병목 현상을 초래한다. 이에 본 논문은 선택적 State Space Model (SSM)에 기반하여 입력 길이에 대한 선형 복잡도를 보장하는 Mamba 아키텍처를 MRTA 문제에 도입한다. Mamba는 순전파 속도와 병렬 학습 능력을 동시에 확보하여 에이전트 수 증가에 따른 연산 부하를 줄일 수 있다[3]. 실험을 통해 로봇 규모를 변화시키며 확장성 측면에서 Mamba와 Transformer의 성능을 비교하고, Mamba가 대규모 다중 로봇 환경에 효율적임을 확인하였다.

Ⅱ. 본론

2.1 MRTA에서 Transformer 기반 접근법과 한계

MRTA에서 심층 강화학습을 활용하여 데이터로부터 최적의 할당 정책을 학습하는 연구가 활발하다. 특히 Transformer 아키텍처는 MRTA 문제를 일련의 노드 시퀀스를 생성하는 과정으로 해석하여 탁월한 성능을 보였다[2]. Transformer의 핵심인 Self-Attention 메커니즘은 모든 로봇 쌍 간의 관계를 계산함으로써 임무 간의 전역적인 상호 의존성을 효과적으로 포착할 수 있다. 그러나 Self-Attention은 입력 시퀀스 길이 N에 대해 $O(N^2)$의 시간 복잡도와 메모리 공간을 요구한다[4]. 이는 대규모 로봇 환경에서 연산량이 기하급수적으로 폭증함을 의미하며, 실시간성이 보장되어야 하는 MRTA 시스템에서 병목 현상을 초래한다. 결과적으로 Transformer 기반 모델은 제한된 컴퓨팅 자원을 가진 엣지 디바이스나

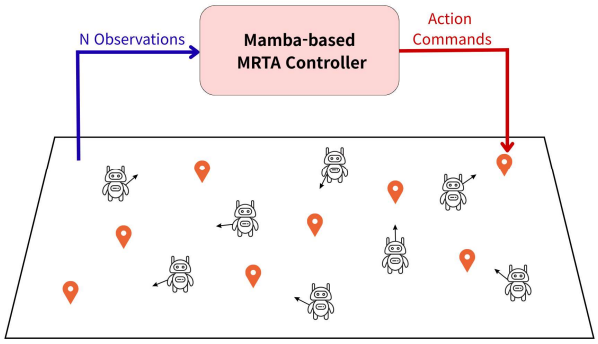

그림 1. MRTA 시스템

대규모 군집 제어 환경에서 선형적인 성능 확장을 기대하기 어려우며, 이를 대체할 효율적인 아키텍처가 필요하다.

2.2 Mamba 아키텍처와 선택적 상태 공간 모델

Transformer의 이차적 연산 비용 문제를 해결하기 위해 경량화 기법이 제안되었으나[5], 이는 종종 모델의 표현력 저하를 동반한다. 이에 대한 대안으로 최근 선택적 상태 공간 기반의 Mamba 아키텍처가 주목받고 있다. Mamba는 연속적인 시스템의 이산화를 통해 Recurrent Neural Network (RNN)과 유사하게 시계열 데이터를 순차적으로 처리하면서도, 훈련 시에는 Convolution Neural Network (CNN)과 같은 병렬 처리가 가능하다는 장점이 있다[6]. 그리고 입력 데이터에 따라 매개변수를 동적으로 조절하여 임무 할당에 유의미한 정보만을 상태 변수에 압축하여 전달한다. 이는 입력 길이 N에 대해 선형적인 시간 복잡도 $O(N)$를 가능하게 해[3], 로봇 규모가 확장되어도 Transformer 대비 낮은 메모리 사용량과 빠른 추론 속도를 보장한다. 이러한 Mamba의 특성이 대규모 MRTA 문제에 적합하다는 점에 착안하여, 그림 2와 같이 Mamba 기반의 인코더와 어텐션 기반의 디코더를 결합한 하이브리드 아키텍처를 제안한다.

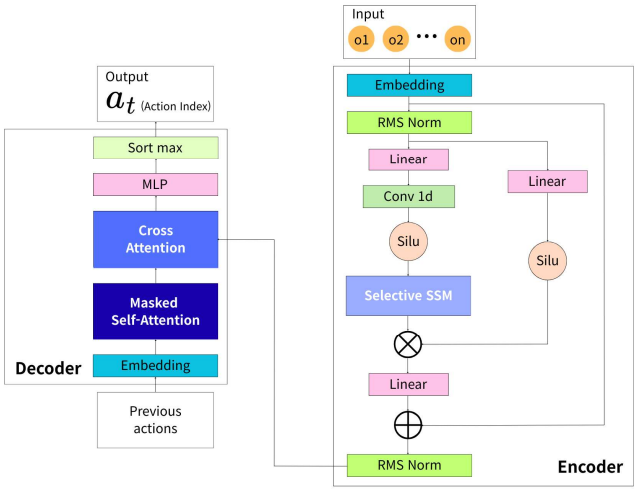

그림 2. Mamba 기반 모델 아키텍처

모델의 인코더는 다중 로봇의 관측값 시퀀스를 입력으로 받아 Mamba 블록을 통해 각 로봇의 전역적 문맥 정보를 추출한다. 이 과정에서 인코더는 선택적 SSM을 활용하여 로봇 간의 복잡한 상호작용 정보를 효과적으로 압축하고 임베딩한다. 이어지는 디코더는 자기 회귀적 방식으로 각 로봇의 행동을 순차적으로 결정한다. 구체적으로 Masked Self-Attention을 통해 이전 로봇들의 할당 결정을 참조하고, Cross-Attention을 통해 Mamba 인코더가 생성한 전역 상태 임베딩을 조회함으로써 현재 순서의 로봇에게 가장 적합한 임무를 할당한다.

2.3 실험

2.3.1 실험 환경 및 설정

실험은 NVIDIA A100 GPU 환경에서 수행되었다. 모델 학습을 위해 Adam Optimizer를 사용하였으며, 학습률은 0.0001로 설정하였다. 실험은 Multi-Agent Particle Environments (MPE)의 'Simple Spread' 시나리오에서 진행되었으며, 에이전트와 랜드마크가 동일한 수로 배치되는 환경을 구축하였다. 아키텍처별 확장성을 검증하기 위해 로봇의 수를 15대와 50대로 구분하여 실험했고, 각 실험은 총 200,000 스텝 동안 진행되었다.

2.3.2 추론 속도 및 확장성 분석

본 연구에서는 제안하는 아키텍처의 성능을 검증하기 위한 기준 모델로 Multi-Agent Transformer(MAT)를 선정하였다. MAT은 다중 에이전트의 상태 관측값을 시퀀스 데이터로 간주하고 이를 인코더-디코더 구조를 통해 처리하여 에이전트 간의 최적 협력 정책을 도출한다. 대규모 환경에서의 모델 확장성을 평가하기 위해 로봇의 수를 15대에서 50대로 점진적으로 증가시키며 실험을 수행하였다. 각 시나리오에서 Mamba 기반 모델과 Transformer 기반 모델의 Frames Per Second (FPS) 변화 추이를 측정하고, 모델의 확장성을 정량적으로 평가하기 위해 속도 감소폭을 지표로 사용하였다. 로봇의 수가 N_{base}에서 N_{target}으로 확장될 때의 비율을 $\alpha = N_{target}/N_{base}$라 하면, 시간 복잡도에 따른 이론적 예측 감소폭 D_{pred}는 다음과 같이 정의된다[3, 4].

$$D_{pred} = \begin{cases} \alpha^2 & \text{for } Transformer(O(N^2)) \\ \alpha & \text{for } Mamba(O(N)) \end{cases} \quad (1)$$

로봇 수를 15대에서 50대로 확장하여 $\alpha \approx 3.33$의 확장 비율을 적용한 결과는 표 1과 같다. Transformer 기반 모델 MAT은 로봇 수가 증가함에 따라 FPS가 약 9.8배 급감하는 양상을 보였다. 이는 Self-Attention의 이차적 복잡도로 인해 로봇 간의 모든 상호작용을 계산하는 과정에서 연산 부하가 로봇 수의 제곱에 비례하여 증가했음을 의미한다. 반면 Mamba 기반 모델은 동일한 환경에서 FPS 저하 폭이 약 4.75배에 그쳤다. 이는 Mamba가 로봇 수 증가에 대해 선형적인 복잡도를 유지한다는 이론적 가설을 뒷받침한다. 로봇 수 증가에 따른 Mamba 기반 모델의 완만한 성능 저하가 대규모 다중 로봇 임무 할당 문제에서 안정적인 추론 속도를 보장할 수 있음을 확인할 수 있다.

표 1. 로봇 수 증가에 따른 모델별 추론 속도 비교

모델	연산 복잡도	FPS(N=15)	FPS(N=50)	속도 감소폭
MAT	$O(N^2)$	49	5	약 9.8배
Mamba 기반 모델	$O(N)$	19	4	약 4.75배

Ⅲ. 결론

본 논문에서는 MRTA 문제에서 발생하는 계산 복잡도와 실시간 의사결정의 병목 현상을 해결하기 위해 Mamba 기반의 아키텍처를 제안하였다. 기존 Transformer 기반 접근법은 Self-Attention 메커니즘의 이차적 시간 복잡도로 인해 로봇 규모 확장에 한계를 보였다. 이에 본 연구는 입력 시퀀스 길이에 대해 선형적인 시간 복잡도를 갖는 선택적 SSM을 MRTA에 도입하여 한계를 구조적으로 개선하였다. 확장성 검증을 위한 실험 결과, 제안하는 모델은 로봇의 수가 15대에서 50대로 증가할 때 추론 속도가 약 4.75배 감소에 그쳐 MAT 대비 완만한 성능 저하를 보였다. 이는 대규모 로봇이 운용되는 복잡한 물류 및 산업 현장에서도 제안 모델이 안정적인 실시간 할당을 수행할 수 있음을 시사한다. 향후 연구에서는 동적 환경이나 이종 로봇 군집 등 현실적인 제약 조건이 포함된 시나리오에 모델을 적용하고자 한다.

ACKNOWLEDGMENT

본 연구는 산업통상자원부 및 산업기술평가관리원(KEIT)의 지원(RS-2024-00442168) 및 과학기술정보통신부 및 정보통신기획평과원의 대학ICT연구센터사업(IITP-2026-RS-2024-00436887)의 연구결과로 수행되었음. 본 논문의 교신저자는 박수현임.

참 고 문 헌

[1] Gerkey, Brian P., and Maja J. Matarić. "A formal analysis and taxonomy of task allocation in multi-robot systems." The International journal of robotics research 23.9 (2004): 939-954.

[2] Wen, Muning, et al. "Multi-agent reinforcement learning is a sequence modeling problem." Advances in Neural Information Processing Systems 35 (2022): 16509-16521.

[3] Gu, Albert, and Tri Dao. "Mamba: Linear-time sequence modeling with selective state spaces." First conference on language modeling. 2024.

[4] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[5] Tay, Yi et al. "Efficient Transformers: A Survey." ACM Computing Surveys 55 (2020): 1 - 28.

[6] Gu, Albert, Karan Goel, and Christopher Ré. "Efficiently modeling long sequences with structured state spaces." arXiv preprint arXiv:2111.00396 (2021).