순환신경망과 트랜스포머 모델을 활용한 POMDP 환경에서의 강화학습 성능 개선에 관한 연구

최요한[1], 지창훈[1], 석영준[1], 김평수[2], 한연희[1†]

[1]한국기술교육대학교 컴퓨터공학과 미래융합공학전공, [2]한국공학대학교 전자공학부

[1]{yoweif, koir5660, dsb04163, yhhan}@koreatech.ac.kr, [2]pskim@tukorea.ac.kr

A Study on Improving Reinforcement Learning Performance in POMDP Environments Using RNN and Transformer

Yo Han Choi[1], Chang-Hun Ji[1], Yeong-Jun Seok[1], Pyung Soo Kim[2], Youn-Hee Han[1]

[1]Future Convergence Engineering, Dept. of Computer Science Engineering, KOREATECH

[2]Dept. of Electronic Engineering, Tech University of Korea

요 약

본 연구는 부분 관찰 마르코프 결정 과정(POMDP) 환경에서 강화학습 성능을 개선하기 위해 순환신경망과 트랜스포머 모델을 비교 분석한다. 실험 결과, 트랜스포머 모델이 장기 의존성 처리에서 탁월한 성능을 보이며 POMDP 환경에서 가장 우수한 성능을 나타냈다. 반면, 다층 퍼셉트론 모델은 관찰 불완전성에 취약하여 성능이 저조하였다. 이 연구는 강화학습에서 트랜스포머 모델의 잠재력을 확인하며, POMDP 문제 해결에 있어 트랜스포머 모델이 특히 효과적임을 시사한다.

Ⅰ. 서 론

알파고의 성공은 딥러닝과 강화학습의 결합이 인공지능 분야에서 획기적인 성과를 이룰 수 있음을 입증하였다 [1]. 강화학습은 에이전트가 환경과 상호작용을 통해 최적의 행동을 학습하는 과정을 의미하며, 딥러닝은 복잡한 정책과 가치 함수를 근사하는 데 중요한 역할을 수행한다. 이러한 결합은 자율주행, 게임 플레이, 로봇 제어 등 다양한 복잡한 문제에 적용되어 인공지능의 성능을 비약적으로 향상시켜왔다.

강화학습은 이제 전통적인 기계 학습 문제를 넘어 자율주행, 재무 관리, 의료 등 다양한 산업 분야로 그 적용 범위를 빠르게 확장하고 있다. 특히, 강화학습은 복잡한 의사결정 문제를 해결하는 데 강점을 보이며, 현실 세계의 복잡한 시스템에서 최적화 문제를 해결하는 데 중요한 도구로 자리매김하고 있다. 이러한 확장은 강화학습의 효율성 향상, 새로운 알고리즘 개발, 그리고 더 강력한 모델에 대한 필요성을 촉진하고 있다.

그러나, 강화학습이 많이 훈련되는 시뮬레이션 환경과 달리 현실 세계의 많은 문제는 부분 관찰 마르코프 결정 과정(Partially Observable Markov Decision Process, POMDP)으로 모델링된다 [2]. POMDP 환경에서의 강화학습은 관찰 불완전성, 높은 상태 공간의 차원, 그리고 불확실한 행동 결과 등으로 인해 여러 가지 도전 과제에 직면한다. 에이전트는 제한된 정보로부터 환경의 현재 상태를 추정하고, 최적의 행동을 선택해야 한다. 이는 학습 과정에서 매우 불안정한 결과를 초래할 수 있으며, 학습의 수렴 속도가 느려지거나 경우에 따라 학습 자체가 불가능해질 위험이 있다. 따라서 POMDP 환경에서의 강화학습은 기존의 접근법들보다 더욱 정교한 신경망 설계와 학습 알고리즘이 요구된다. 본 연구에서는 이러한 문제를 해결하기 위한 방법으로써, POMDP 환경에서 순환 신경망(Recurrent Neural Network, RNN) [3]과 트랜스포머(Transformer) [4] 모델을 활용한 에이전트 네트워크의 강화학습 성능을 분석하고자 한다.

Ⅱ. 본 론

강화학습은 일반적으로 마르코프 결정 과정(Markov Decision Process, MDP) 환경을 해결하는 것을 목표로 하며, 이는 마르코프 속성(Markov Property)이 만족된다는 가정에 기반한다. 마르코프 속성이란, 현재 상태가 주어졌을 때 미래의 상태는 과거의 상태와는 독립적이라는 것을 의미하며, 이는 현재 상태가 모든 과거 정보를 함축하고 있다는 가정을 전제로 한다. 그러나, 실제 환경이 MDP가 아닌 POMDP일 경우, 이 속성은 더 이상 보장되지 않는다. POMDP 환경에서는 에이전트가 환경의 완전한 상태를 관찰할 수 없으므로, 과거의 관찰 시퀀스를 고려하지 않고는 최적의 결정을 내리기 어려워진다. 이러한 이유로, POMDP 문제를 해결하기 위해서는 전통적인 MDP 접근법과는 다른 방식으로 과거 시퀀스 데이터를 활용할 필요가 있다. 본 연구에서는 이러한 문제를 해결하기 위해 순환 신경망과 트랜스포머 모델을 활용하여, 이 모델들이 과거 시퀀스 데이터를 활용하여 POMDP 환경에서의 성능을 개선할 수 있는지 분석한다.

트랜스포머 모델은 자연어 처리(NLP) 분야에서 혁신적인 변화를 가져왔다. 멀티헤드 어텐션(Multi-Head Attention) 메커니즘을 통해 순차적 처리에 의존하지 않고도 병렬 연산이 가능하며, 문맥 정보를 효과적으로 처리한다. 또한 장기 의존성(Long-Term Dependency) 처리에 뛰어나며, 이를 통해 순환 신경망 기반 모델들이 직면했던

† 교신저자: 한연희 (yhhan@koreatech.ac.kr)

기울기 소실(Vanishing Gradient) 문제를 극복할 수 있었다. 이러한 특성은 강화학습에서도 잠재적인 혁신을 가져올 수 있다.

특히, 트랜스포머의 어텐션 메커니즘은 POMDP 환경에서 관찰 불완전성을 극복하고, 중요한 정보를 강조함으로써 더 나은 상태 추정을 가능하게 한다. 또한, 병렬 처리 능력은 대규모 데이터에서의 학습 속도를 크게 향상시킬 수 있으며, 이를 통해 강화학습의 효율성을 더욱 증진시킬 수 있다.

따라서 본 연구는 POMDP 환경에서 강화학습의 성능을 개선하기 위해, 순환신경망 및 트랜스포머 모델을 비교하고 검증한다.

Ⅲ. 실 험

본 연구에서 각 모델을 평가하기 위해 MuJoCo 시뮬레이터 [5]를 사용하여 물리적 환경을 모델링하였다. 추가로 POMDP 환경을 구현하기 위해 MuJoCo 시뮬레이터에서 에이전트에 주어지는 관찰 정보를 조정하였다. 구체적으로, 에이전트는 위치 정보만을 관찰할 수 있는 환경과 속도 정보만을 관찰할 수 있는 환경에서 학습을 수행하였다. 시뮬레이터 내 태스크로는 HalfCheetah, Hopper, Walker2D를 사용하였다.

본 연구에서는 세 가지 모델을 사용하여 POMDP 환경에서의 성능을 비교하였다. 첫 번째 모델은 다층 퍼셉트론(MLP)으로, 과거의 관찰 정보를 활용하지 않고 현재 관찰된 정보를 바탕으로 즉각적인 결정을 내리는 모델이다. 두 번째 모델은 순환 신경망의 일종인 GRU (Gated Recurrent Unit)로, 과거의 관찰 정보를 활용하여 현재의 상태를 추정하고 행동을 결정하는 능력을 가진다 [6]. 세 번째 모델은 트랜스포머 아키텍처를 기반으로 한 GPT (Generative Pre-trained Transformer) 모델로, 장기 의존성을 처리하는 데 강점을 가지며, 과거 시퀀스의 중요한 정보를 효과적으로 통합하여 최적의 결정을 내릴 수 있다 [7].

모든 모델의 학습에는 Proximal Policy Optimization (PPO) 알고리즘을 사용하였다. PPO는 정책의 급격한 변경을 방지하여 업데이트를 안정적으로 수행하며, 높은 수렴 속도와 성능을 제공하는 강화학습 알고리즘이다.

이러한 구성으로 수행된 실험 결과는 표 1에서 볼 수 있다. 위치 정보만 주어지는 환경은 P, 속도 정보만 주어지는 환경은 V와 함께 표기하였다. 표 1에 제시된 실험 결과에 따르면, 각 모델은 POMDP 환경에서 상이한 성능을 보였다. MLP 모델은 과거의 관찰 정보를 활용하지 않는 구조적 한계로 인해, 다른 두 모델에 비해 일관되게 낮은 성능을 기록하였다. 이는 POMDP 환경에서 과거 정보를 적절히 활용하는 것이 성능 향상에 필수적임을 시사한다. GRU와 GPT 모델 간의 성능 비교에서, GPT 모델이 대부분의 경우 더 우수한 성능을 나타내었다. 특히, 복잡한 POMDP 환경에서 GPT 모델의 성능이 두드러졌으며, 이는 GPT가 장기 의존성을 보다 효과적으로 처리할 수 있음을 보여준다. 한편, Hopper 환경에서는 상대적으로 문제의 난이도가 낮아, 모델 간의 성능 차이가 크게 드러나지 않았다.

표 1. 실험 결과

	MLP	GRU	GPT
HalfCheeta-P	1922±587	3047±682	3899±406
HalfCheeta-V	1612±204	2305±695	3320±497
Hopper-P	885±171	953±99	945±83
Hopper-V	872±184	939±86	943±77
Walker2D-P	647±150	926±263	1363±211
Walker2D-V	567±290	801±277	1287±364

Ⅳ. 결론

본 연구는 MLP, GRU, GPT 모델을 POMDP 환경에 적용하여 각 모델의 성능을 비교하였다. 실험 결과, GPT 모델이 POMDP 환경에서 관찰 불완전성을 극복하는 데 있어 가장 효과적임을 확인하였다. GRU 모델은 시간적 종속성 처리에서 우수한 성능을 보였으나, GPT 모델에 비해 제한적인 성능을 나타냈다. MLP 모델은 가장 단순한 구조로, 복잡한 POMDP 문제에서 성능의 한계를 드러냈다. 연구 결과는 트랜스포머 모델이 강화학습에서 특히 POMDP 환경에 적합한 모델임을 시사하며, 향후 연구에서 이 모델의 활용 가능성을 더욱 탐구할 필요가 있음을 보여준다.

ACKNOWLEDGMENT

이 논문은 2023년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(No. NRF-2023R1A2C1003143)이며, 또한 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-학·석사연계ICT핵심인재양성의 지원(IITP-2024-RS-2022-00156326, 50)을 받아 수행된 연구임

참 고 문 헌

[1] Silver, D., et al. "Mastering the game of Go with deep neural networks and tree search," Nature, vol. 529, no. 7587, pp. 484-489, Jan. 2016.

[2] Cassandra, A. R., Kaelbling, L. P., Littman, M. L. "Acting optimally in partially observable stochastic domains," Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1023-1028, Aug. 1994.

[3] Hochreiter, S., Schmidhuber, J. "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[4] Vaswani, A., et al. "Attention is all you need," Advances in Neural Information Processing Systems, pp. 5998-6008, Dec. 2017.

[5] Todorov, E., et al. "MuJoCo: A physics engine for model-based control," 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 5026-5033, Oct. 2012.

[6] Cho, K., et al. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724-1734, Oct. 2014.

[7] Radford, A., et al. "Language Models are Unsupervised Multitask Learners," OpenAI preprint, 2019, (https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf).