효율적 학습을 위한 State Representation 기반 Decision Transformer 알고리즘

조예령*, 민준서*, 박수현† , 김중헌*

{joyena0909*, joongheon*}@korea.ac.kr, soohyun.park@sookmyung.ac.kr† , minjs4562@gmail.com*

State Representation-Based Decision Transformer Algorithm for Efficient Learning

Yeryeong Cho*, Junseo Min*, Soohyun Park† , Joongheon Kim*

Korea University*, Sookmyung Women's University†

요 약

본 논문은 연속 제어 환경 내 학습 효율을 향상시키기 위해 state representation 기반 decision transformer (DT)알고리즘을 제안하고, 그 효과를 Gymnasium HalfCheetah 환경에서의 실험을 통해 검증한다. 기존 DT는 궤적(trajectory)을 시퀀스로 모델링하여 Return-to-Go 조건하에서 행동을 예측하나, 원시 상태를 그대로 입력할 경우 불필요한 정보와 분포 변동성이 학습 안정성과 샘플 효율을 저해할 수 있다. 이에 따라 본 연구는 상태를 저차원 잠재 표현으로 정규화 및 압축하는 학습 모듈을 도입하고, 해당 표현을 기반으로 시퀀스 예측을 수행함으로써 정책 학습의 효율적 수렴을 유도한다. 실험 결과와 함께 관련 연구 동향 또한 소개하면서, DT 기반 연속 제어 학습에 대해 구체적으로 논의하여 실질적인 연구 방향에 대해 시사한다.

Ⅰ. 서 론

강화학습(reinforcement learning, RL)은 에이전트가 환경과의 상호작용을 통해 누적 보상을 최대화하는 정책을 학습하는 순차적 의사결정(sequential decision-making) 방법론이다 [1]. 특히 로봇 제어, 자율주행, 게임 플레이 등과 같이 연속 제어(continuous control)가 요구되는 문제에서 RL은 복잡한 동역학과 비선형 제약을 내재적으로 다룰 수 있다는 장점이 있다. 그러나 연속 제어 환경은 행동(action) 및 상태(state)가 고차원이며, 보상(reward)이 희소하거나 지연되는 경우가 많아 학습이 불안정해지기 쉽다. 또한 시뮬레이터 구동 비용, 탐험(exploration)에 따른 시행착오 비용 등으로 인해, 충분히 많은 상호작용 데이터를 수집한 뒤 학습 가능한 전통적 RL 가정이 현실 적용에서 한계가 존재한다 [2].

한편, transformer 모델이 시퀀스 모델링에서 주목받으면서, RL에서도 이를 결합하여 상태, 행동, 보상을 시퀀스로 재구성하고, 효율적으로 정책을 학습하려는 시도가 증가하고 있다 [3]. 이때 실제 성능을 좌우하는 요소 중 하나는 상태를 어떤 형태로 입력에 제공하는가이다. 원시 상태를 그대로 사용하면 불필요한 정보가 포함되거나 관측 분포의 변동성이 커져 학습 안정성과 샘플 효율을 저해할 수 있기 때문이다. 따라서 상태를 저차원 잠재 표현(latent representation)으로 정규화 및 압축하여 정책 학습에 필요한 정보만 선별적으로 전달하는 방식은, 연속 제어 환경에서의 효율적 수렴과 일반화에 유리할 수 있으며, 본 논문은 이에 대한 기초 연구를 통해 가능성을 검토하고자 한다.

Ⅱ. Decision Transformer의 개념 및 동향

Decision Transformer (DT)는 RL의 보상, 상태, 행동을 시퀀스 데이터로 모델링한 후 Transformer를 활용해 행동을 생성하는 모델이다 [4]. DT는 오프라인 데이터셋을 정답(ground truth)으로 활용하는 지도학습 방식을 따른다. 특히, DT는 목표 보상을 입력받는 Return-to-Go (RTG) 방식으로 학습을 진행한다. 목표 보상에 도달하는 최적값을 학습하는 방식을 통해, 단순한 모방학습(behavior cloning)이 데이터의 평균적 행동을 모방하는 것에 그치는 한계를 극복한다 [5, 6]. 이러한 지도학습 방식은 bootstrapping과 discounting을 배제하여 장기적 의사결정에서 전통적

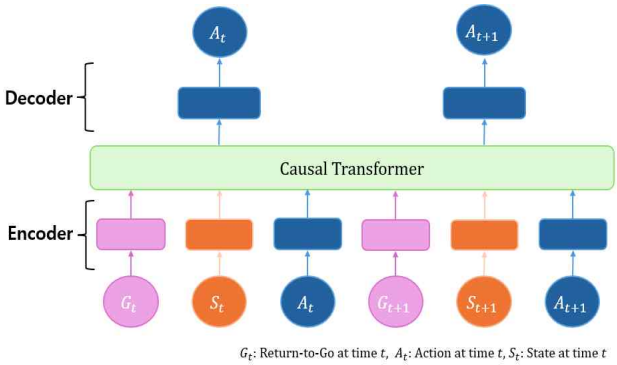

그림 1. Decision Transformer(DT)의 구조

RL보다 안정적인 성능을 보인다. 특히, transformer 구조의 self-attention을 활용해 행동과 보상을 직접 연결하는 학습 방식을 통해 기존 Bellman backup방식이 가진 기여도 할당 문제(credit assignment)와 지연전달(slow propagation)문제를 완화하여 빠르고, 정확한 학습과 추론을 가능하게 한다 [7]. 이렇게 DT가 제안된 이후 멀티 에이전트 학습 환경으로의 확장, 한계점 개선 및 구조 확장에 대한 연구가 활발하게 진행되었다. Multi-Agent Decision Transformer (MADT)는 기존 DT를 Multi-Agent System (MAS)으로 확장한 모델이다 [8]. MADT는 각 에이전트의 행동을 시퀀스로 모델링하며, 오프라인 Actor-Critic 방식의 학습을 도입하였다. 이를 통해 개별 에이전트 정보와 전체 집단 정보의 공유를 통해 MAS에서 우수한 협력 성능을 확보하였다.

하지만 기존 DT는 최적 경로 결합 문제(stitching ability)와 보상설계 문제(reward design)에서 한계가 존재한다. 이러한 한계를 개선하기 위해 Q-learning Decision Transformer (QDT)는 Q-learning과의 결합을 통해 가치를 평가하여 가치가 높은 상태를 연결하여 최적 경로 결합 문제를 완화하였다 [9, 10]. 이에 대한 예시로, StARformer는 상태, 행동, 보상 구조를 통해 인과관계를 학습하고, 보상을 명시적으로 포함하여 복잡한 보상설계 문제를 완화하였다. 하지만, 확률적 환경에서는 동일한 행동이

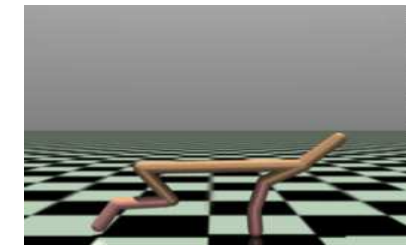

그림 2. Gymnasium의 HalfCheetah

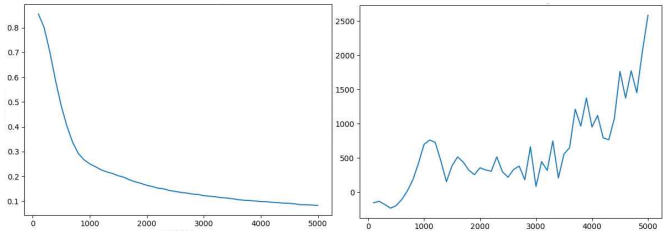

그림 3. HalfCheetah 환경 내 학습 Loss(좌)와 Reward(우)

서로 다른 결과를 도출할 수 있기 때문에 DT와 같은 지도학습기반 방식은 실패하거나 성능이 저하될 수도 있다는 점을 유의해야한다[11].

Ⅲ. 실험 환경 구성 및 결과

본 논문은 연속 제어 환경에서 DT의 학습 효율을 향상시키기 위해 state representation 기반 DT 알고리즘을 제안하고, 그 효과를 Gymnasium HalfCheetah 환경에서의 실험을 통해 검증한다. 그림 2에서 볼 수 있듯이, HalfCheetah는 관절 구동 기반의 2족 로봇이 전진 속도를 최대화하도록 제어하는 과제로, 고차원 연속 상태와 연속 행동 공간을 포함하므로 시퀀스 모델 기반 정책 학습의 성능과 안정성을 관찰하기에 적절하다. 입력 시퀀스는 시간 순서에 따른 토큰화 및 마스킹을 통해 미래 정보 누출을 방지하도록 구성하였으며, 학습 과정에서는 손실 함수 기반의 행동 예측 정확도와 에피소드 단위 누적 보상(episode return)을 중심 지표로 사용하였다. 실험 결과, DT는 HalfCheetah 환경에서 RTG 조건에 따라 행동을 생성하며 정책을 학습하는 경향을 보였고, 학습이 진행됨에 따라 에피소드 누적 보상이 점진적으로 개선되는 양상이 관찰되었다. 다만 연속 제어 환경 특성상 데이터 품질 및 분포에 민감하게 반응하는 경향이 존재하며, 학습 안정성은 시퀀스 길이, 배치 구성, 그리고 RTG 스케일링 등 전처리 설정에 의해 유의미하게 영향을 받는 것으로 확인되었다. 결론적으로, DT는 HalfCheetah에서 시퀀스 모델 기반 정책 학습의 가능성을 보여주었으며, 동일 환경에서의 성능 향상을 위해서는 데이터 구성 방식 및 설정에 대한 체계적 검토가 필요함을 시사한다.

Ⅳ. 결론

본 논문은 연속 제어 환경에서 시퀀스 모델 기반 정책 학습의 적용 가능성을 확인하기 위해 Gymnasium HalfCheetah 환경에 DT를 적용하고, 행동을 예측하는 학습 절차를 구성하였다. 실험 결과, DT는 RTG에 조건화된 행동 생성 양상을 보였으며, 학습 진행에 따라 에피소드 누적 보상이 점진적으로 개선되는 경향이 관찰되었다. 한편, 연속 제어 환경의 특성상 성능은 시퀀스 길이, 배치 구성, RTG 스케일링과 같은 전처리 및 학습 설정이 안정성에 유의미한 영향을 미치는 것으로 확인되었다. 이러한 분석을 통해, 본 논문은 DT 기반 학습의 재현성과 성능 향상을 위해서는상태 입력의 정규화 전략 구성, 하이퍼파라미터에 대한 체계적 분석 등이 필요함을 시사한다. 향후 연구는 상태를 저차원 잠재 표현으로 압축 및 정규화하는 표현 학습 모듈을 결합하고, 연속 제어 기준 기법 및 주요 오프라인 RL 기법과의 비교를 통해 학습 효율 및 일반화 성능을 정량적으로 검증하는 단계가 필수적이다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원을 받아 수행된 연구임 (IITP-2026-RS-2024-00436887). 본 논문의 교신 저자는 김중헌임.

참 고 문 헌

[1] K. Arulkumaran, M. P. Deisenroth, M. Brundage and A. A. Bharath, "Deep Reinforcement Learning: A Brief Survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26-38, Nov. 2017.

[2] A. Okutan Kara, M. Kara and A. Boyaci, "A Comparative Analysis of Machine Learning and Deep Reinforcement Learning Approaches for Adaptive Intrusion Detection," *IEEE Access*, vol. 13, pp. 189833-189849, Oct. 2025.

[3] H. Gao et al., "A Spatial - Temporal Predictive Transformer Network for Level-3 Autonomous Vehicle Decision-Making," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 7, pp. 12228-12242, Jul. 2025.

[4] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, "Decision Transformer: Reinforcement Learning via Sequence Modeling," in *Advances in Neural Information Processing Systems (NeurIPS)*, Virtual, Dec. 2021, pp. 15084-15097.

[5] I. Kpereobong Friday, S. Prasanna Pati and D. Mishra, "A Multi-Modal Approach Using a Hybrid Vision Transformer and Temporal Fusion Transformer Model for Stock Price Movement Classification," *IEEE Access*, vol. 13, pp. 127221-127239, Jul. 2025.

[6] S. Fujimoto, D. Meger and D. Precup, "Off-Policy Deep Reinforcement Learning without Exploration," in *Proc. International Conference on Machine Learnin (ICML),* Long Beach, California, Jun. 2019, pp. 2052-2062.

[7] A. Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NIPS)*, Long Beach, California, Dec. 2017, pp. 5998-6008.

[8] V. Nurmanova, Y. Akhmetov, M. Bagheri, A. Zollanvari, B. T. Phung and G. B. Gharehpetian, "Confidence Level Estimation for Advanced Decision-Making in Transformer Short-circuit Fault Diagnosis," *IEEE Transactions on Industry Applications*, vol. 58, no. 1, pp. 233-241, Jan. 2022.

[9] T. Yamagata, A. Khalil, and R. Santos-Rodriguez, "Q-Learning Decision Transformer," in *Proc. International Conference on Machine Learnin (ICML)*, Honolulu, pp. 39076-39093, Jul. 2023.

[10] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3, pp. 279-292, May 1992.

[11] J. Shang, X. Kahatapitiya, K. Li, and M. S. Ryoo, "StARformer: Transformer with State-Action-Reward Representations for Visual Reinforcement Learning," in *Proc. European Conference on Computer Vision (ECCV)*, Israel, Oct. 2022, pp. 462-479.