오류 누적 완화를 위한 의사결정 트랜스포머-흐름 매칭 통합 구조

김민준, 지창훈, 최요한, 눌란벡 키즈 아셀, 한연희†
한국기술교육대학교 컴퓨터공학과 미래융합전공

{june573166, koir5660, yoweif, aselbaekki, yhhan}@koreatech.ac.kr

A Flow Matching-Integrated Decision Transformer Architecture for Mitigating Compounding Errors

Min-Jun Kim, Chang-Hun Ji, Yo-Han Choi, Nurlanbek kyzy Asel, Youn-Hee Han†
Future Convergence Engineering, Dept. of Computer Science and Engineering, KOREATECH

요 약

강화학습 연구 분야의 한 갈래인 오프라인 강화학습(Offline Reinforcement Learning)은 오래전부터 연구되어온 일반적인 온라인 강화학습(Online Reinforcement Learning) 방법론이 실제 세계에서 직면하는 여러 실용적인 문제들을 해결한다. 본 연구는 오프라인 강화학습 알고리즘 중 하나인 의사결정 트랜스포머(Decision Transformer)와 흐름 매칭(Flow Matching)을 접목한 구조를 제안한다. 제안한 구조는 기존 의사결정 트랜스포머의 오류 누적(Compounding Error) 문제를 완화한다. 이를 검증하기 위해 간단한 환경에서 실험하여 순수 의사결정 트랜스포머보다 더 효과적으로 목표에 도달함을 확인하였다. 본 연구에서는 오류 누적 문제를 가진 알고리즘들의 개선 가능성을 제시하고, 오프라인 강화학습이 더 다양한 문제를 해결할 것으로 기대한다.

Ⅰ. 서 론

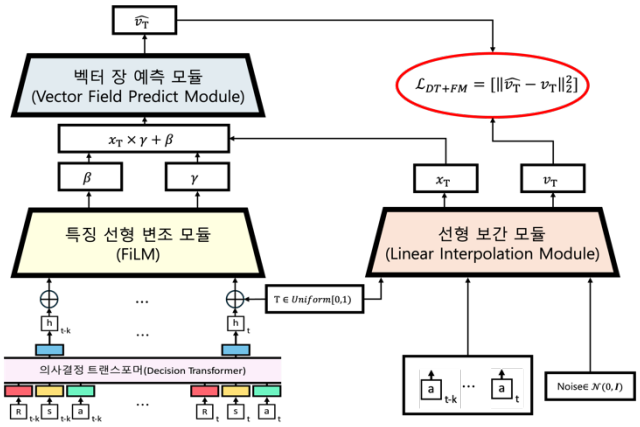
강화학습(Reinforcement Learning)은 에이전트가 환경과 상호작용하여 보상을 받으면서 누적 기대 보상을 최대화하도록 정책을 학습한다. 그런데 환경에 접근할 수 없거나 학습 데이터 수집 비용이 큰 경우, 오프라인 강화학습(Offline Reinforcement Learning)을 활용한다.

의사결정 트랜스포머(Decision Transformer, DT) [1]는 트랜스포머 구조를 기반으로 사전에 수집한 데이터를 활용해 학습하는 오프라인 강화학습 알고리즘이다. DT 모델은 시계열 데이터의 자기집중(Self-Attention) 기법을 통해 먼 과거의 행동과 현재 보상과의 관계를 이해할 수 있다. 하지만 DT 는 학습 과정에서 보지 못한 분포 밖(Out Of Distribution, OOD) 데이터 입력 이나 자기회귀적(Autoregressive) 추론 방식으로 인해 발생한 오류 누적(Compounding Error) 문제에 취약하다. 따라서 우리는 DT 의 오류 누적 문제를 완화할 수 있는 새로운 구조를 제안한다.

제안하는 구조는 DT에 흐름 매칭(Flow Matching, FM) [2]을 접목하여 초기 노이즈 데이터 분포에서 목표 데이터 분포로 향하는 벡터 장(Vector Field)를 학습한다. 오류가 발생하더라도, 학습된 벡터 장이 노이즈 분포를 목표 데이터 분포로 유도하기 때문에 안정적으로 수렴하며 결과적으로 오류 누적 문제가 완화된다. 이를 검증하기 위해 간단하고 제한적인 격자 (Grid) 환경에서 실험을 진행하였고, 오류 누적 문제를 완화함을 확인하였다.

Ⅱ. 의사결정 트랜스포머

DT 는 사전에 수집된 정적 데이터셋으로부터 정책을 학습한다. 데이터셋은 길이 k의 시계열 샘플로, 각 시점 t에서의 목표 누적 보상(Return-To-Go, RTG) R_t, 상태 s_t, 행동 a_t로 구성된다. 모델은 이러한 시계열 데이터셋



[그림 1] 의사결정 트랜스포머 + 흐름 매칭의 학습 구조

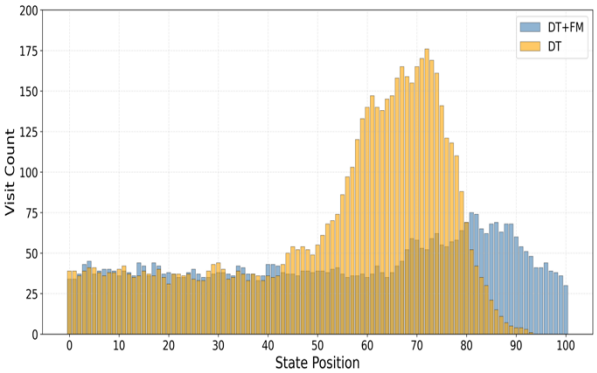
에 대해 주어진 RTG 를 만족하는 행동을 예측하도록 학습된다. 구체적으로, 모델이 예측한 행동 $\hat{a}_t$ 와 실제 데이터셋에 기록된 행동 a_t간의 차이를 최소화해 모델이 주어진 RTG 를 만족하는 행동을 정확히 예측하도록 학습한다.

그러나 DT 는 모방 학습(Imitation Learning)의 학습 방식을 따라, 학습 과정에서 보지 못한 OOD 데이터 입력에 대해 오류가 발생하며, 더 나아가 자기회귀적 구조의 특성상 초기에 발생한 이러한 오류가 다시 입력으로 반영되면서 오류가 누적된다.

Ⅲ. 흐름 매칭

FM 은 데이터 분포의 변화를 확률 흐름(Probability Flow)으로 표현하고, 확률 흐름을 흐름 매칭 시간에 대해 미분하여 그 변화를 기술하는 벡터 장을 직접 회귀 방식으로 학습한다. 구체적으로, 흐름 매칭 시간 T에 대

† 한연희(Youn-Hee Han, yhhan@koreatech.ac.kr): 교신저자



[그림 2] 격자(Grid) 환경 실험 방문 수 비교 그래프

해 노이즈 분포 x_0 에서 목표 분포 x_1 로 향하게 하는 연속적인 확률 흐름 x_T 가 정의된다. 이때 목표 벡터 장 v_T 는 이 확률 흐름을 T 에 대해 미분하여 얻어지며, 학습 과정에서는 v_T 와 모델이 예측한 벡터 장 $\hat{v}_T$ 간의 차이를 최소화함으로써 노이즈 분포에서 목표 분포로 수렴하도록 유도한다. 따라서 학습 과정에서 보지 못한 OOD 데이터가 입력되어 오류가 발생하더라도, 학습된 벡터 장이 이를 목표 분포로 유도함으로써 오류가 보정되고 결과적으로 오류 누적 문제가 완화된다.

IV. 제안한 구조

그림 1 은 제안한 DT 와 FM 을 결합한 학습 구조이다. 먼저 DT 는 입력으로 윈도우 샘플 k만큼의 RTG, 상태, 대응하는 행동을 받아들인다. 이들을 임베딩하여 인과적 마스킹(Causal Masking)을 적용한 DT 를 통해 k 개의 특징(Hidden State)들을 출력한다. 이 특징들에 FM 을 적용하기 위해 임베딩한 흐름 매칭 시간 T 를 각각 더한다. 이어서 흐름 매칭 시간 T 가 더해진 특징들을 특징 선형 변조 모듈(FiLM) [3]에 통과시켜 β 와 γ 를 구한다.

한편, 선형 보간 모듈(Linear Interpolation Module)의 입력으로 초기 분포인 가우시안 노이즈, 목표 분포인 목표 행동, 흐름 매칭 시간 T 를 주어 확률 흐름 x_T 와 목표 벡터 장 v_T 를 출력한다. 마지막으로 앞서 구한, β 와 γ 를 사용해 확률 흐름 x_T 를 선형 변환한 뒤, 벡터 장 예측 모듈(Vector Field Predict Module)의 입력으로 넣어 예측 벡터 장 $\hat{v}_T$ 를 출력한다. 모든 모델의 학습은 아래의 손실 함수 $\mathcal{L}$ 을 최소화하는 방향으로 종단간(End-to-End) 진행된다.

$$\mathcal{L}_{DT+FM} = [\|\hat{v}_T - v_T\|_2^2] \quad (1)$$

학습 후 추론 시에는 목표 RTG R_0 와 초기 상태 s_0 를 DT 의 입력으로 주어 특징 h 를 출력한다. 이를 윈도우 샘플 k 만큼 생성한 $h_{0\sim k-1}$ 중 마지막 특징 h_{k-1} 와 노이즈를 디노이징 단계를 N번 거쳐 노이즈 없는 새로운 행동 a_{k-1}을 생성한다.

V. 실험

실험은 0 에서 시작해 100 에 도달하는 1 차원 격자 환경에서 진행되었으며, 에이전트는 매 스텝마다 왼쪽(-1) 또는 오른쪽(+1)로 이동할 수 있다. 에이전트가 계속 오른쪽 행동을 선택하여 100 에 방문하면 성공 처리하고 보상 1 을 받는다. 이외의 상태에는 보상이 주어지지 않는다. 100에 끝내 방문하지 못하고 200스텝 이상 지나면 타임 아웃하여 실패 처리한다. 학습에 사용한 정적 데이터셋은 에이전트가 항상 오른쪽으로 움직였을 때의 상태, 행동, RTG 이다.

<표 1> 격자(Grid) 환경 실험 결과 비교

Model	Success rate (%)	Most visited states (number of visits to that state)
DT	0.0	72 (176)
DT+FM	100.0	81 (75)

추론 시에는 의도적으로 OOD 데이터 입력에 의한 오류 누적을 발생시키기 위해 새로운 환경을 설계하였다. 구체적으로 새로운 환경은 10% 확률로 추출한 행동과 무관하게 에이전트는 왼쪽(-1)으로 이동한다. 실험은 총 30 에피소드를 실행하였다. 표 1 은 이들의 성공률과 가장 많이 방문한 상태 및 해당 상태의 방문 횟수를 비교한 결과이다.

표 1 의 결과는 DT+FM 이 100%로 성공한 반면, DT 는 한 번도 성공하지 못한 것을 보여준다. 이는 DT 가 가진 OOD 데이터에 대한 오류 누적 문제를 보여주고, DT+FM 은 이를 효과적으로 해결했음을 나타낸다. 또한 가장 많이 방문한 상태를 비교하면, DT 는 72 상태를 176 번으로 가장 많이 방문한 반면, DT+FM 은 81 상태에서 75 번 방문하였다. 이는 DT 의 최대 방문 횟수가 DT+FM 의 2.3 배이고, 그 당시의 상태 또한, DT 가 더 빨리 나타나 오류 누적에 취약한 것을 보여준다.

그림 2 의 결과는 표 1 의 결과를 도식화한 것으로, 각 그래프의 100 상태와 피크 값을 통해 표 1 의 결과를 확인할 수 있다. 특히 주목할 점은 그래프가 상승하기 시작하는 지점으로, DT 는 43 상태부터 상승하기 시작해 오류 누적 현상이 나타나는 반면, DT+FM 은 64 상태부터 상승하여 오류 발생이 지연됨을 보여준다.

VI. 결론

본 연구에서는 DT 의 오류 누적 문제를 완화할 수 있는 FM 결합 구조를 제안하고, 격자 환경 실험 결과를 통해 제안한 구조가 오류 누적 문제를 완화함을 보였다. 더 나아가 본 연구는 오류 누적 문제를 가진 기존 알고리즘들의 개선 가능성을 제시하며, 오프라인 강화학습이 보다 다양한 실제 문제 상황에 적용될 수 있는 잠재력을 보여준다.

ACKNOWLEDGMENT

이 논문의 연구는 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행되었음

(No.NRF-2023R1A2C1003143&No.NRF-2018R1A6A1A03025526).

참 고 문 헌

[1] Chen, Lili, et al. "Decision transformer: Reinforcement learning via sequence modeling." Advances in neural information processing systems 34 (2021): 15084-15097.

[2] Lipman, Yaron, et al. "Flow matching for generative modeling." arXiv preprint arXiv:2210.02747 (2022).

[3] Perez, Ethan, et al. "Film: Visual reasoning with a general conditioning layer." Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1. 2018.