Trajectory의 마지막 State를 활용한 Decision Tranformer의 성능 향상에 관한 연구

김형진, 이정우*

서울대학교, *서울대학교

hjkim@cml.snu.ac.kr, *junglee@snu.ac.kr

A Study on the Performance Improvement of Decision Tranformer Using the Last State of Trajectory

Kim Hyung Jin, Lee Jung Woo*

Seoul National Univ., *Seoul National Univ.

요 약

본 논문은 최근 강화학습의 주류인 Decision Transformer 모델에서 입력 trajectory의 마지막 state, action, return-to-go 세트를 training 하기 전에 미리 prompt로 입력해준 결과 학습속도가 향상된 것을 확인할 수 있었다. 환경은 Mujoco 의 Hopper와 Ant에서 진행하였고 expert dataset을 이용하였다. medium dataset을 이용한 Hopper 환경에서도 학습속도가 향상된 것을 확인할 수 있었다.

Ⅰ. 서 론

본 논문에서는 최근 강화학습의 주류인 Decision Transformer[1] 모델의 training 성능을 향상시켜 보기 위하여 실험을 진행하였다. Decision Transformer의 경우 Offline Reinforcement Learning에서 주로 사용되어 실시간으로 환경과 소통하여 training하지 않고 미리 진행한 trajectory의 dataset을 이용하여 training을 진행한다. 이 과정에서 trajectory의 dataset의 마지막 state, action, return-to-go 세트를 이용하여 성능향상을 노려보았다.

Ⅱ. 본론

기존 Decision Transformer의 경우 Offline Reinforcement Learning으로 미리 진행한 trajectory dataset만을 사용하여 training을 진행한다. dataset의 경우 전문가(expert)가 진행한 trajectory, 학습이 덜 된 자(medium)가 진행한 trajectory 등 여러 가지가 있다. 기존 Decision Transformer에서 training은 dataset의 trajectory에서 랜덤하게 일정한 길이의 sequence를 잘라와서 이를 Decision Transformer에 순서대로 넣어서 진행한다. 그렇게 함으로써 Decision Transformer가 기존 dataset의 trajectory를 따라가는 법을 배우는 것이다.

본 논문에서는 dataset의 trajectory에서 랜덤하게 일정한 길이의 sequence를 잘라와서 입력하기 전에 그 trajectory의 마지막 state, action, return-to-go 세트를 먼저 입력을 해준 후에 기존과 동일한 training 루트를 따라가게 하였다. expert trajectory의 경우 대부분 task를 성공하므로 마지막 state, action, return-to-go가 마치 현재 agent가 가야될 goal로서의 역할을 담당할 수 있다고 보았다. 실제로 goal-oriented Reinforcement Learning에서의 실험을 보면 training할 때 state와 goal을 함께 입력하면 성능이 향상되는 것을 볼 수 있다.

실험은 OpenAI의 gym의 Mujoco에서 Hopper와 Ant환경에서 진행하였다. 그림 1을 보면 두 가지 환경에서 모두 최종 성능은 같았지만 training

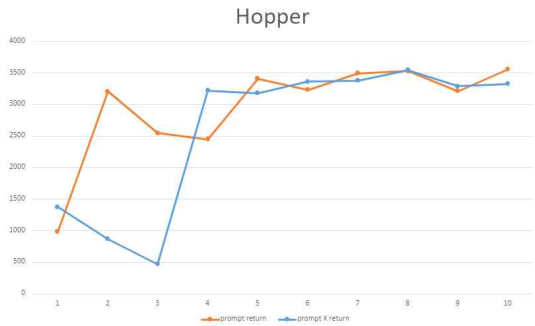

그림 1 Hopper에서의 Return값 비교

그림 2 Hopper에서의 Failure rate 비교

속도가 훨씬 향상된 것을 확인할 수 있었다. 그림에서 주황색 그래프가 본 논문 모델의 성능이고 파란색 그래프가 기존 Decision Transformer의 성능이다. 그림 2 같은 경우는 Hopper가 달리다가 쓰러졌을 때를 Failure이라 보고 Failure를 할 확률을 실험하여 나타낸 결과이다. 본 논문의 주황색 그래프가 Failure rate가 더 낮아서 성능이 더 좋은 것을 확인할 수 있다. 마찬가지로 Ant 환경에서도 똑같이 실험을 진행하였고 실험결과는 그림 3, 4와 같다.

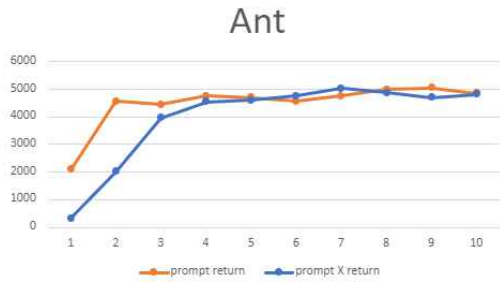

그림 3 Ant에서의 Return값 비교

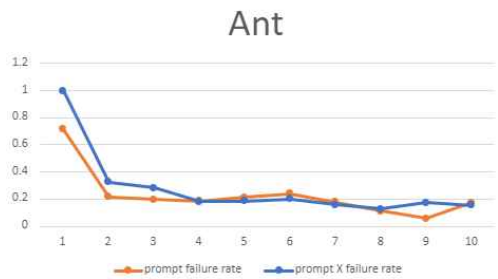

그림 4 Ant에서의 Failure rate 비교

그리고 expert dataset이 아닌 medium dataset에서도 똑같이 실험을 진행해 보았다. expert dataset을 사용한 실험 만큼은 아니지만 training 속도가 빨라진 것을 확인할 수 있었다. medium dataset의 경우 data 속 trajectory가 항상 성공한 trajectory가 아닐 수도 있기 때문에 expert dataset을 사용한 경우보다 효과가 덜 한 것 같다고 결론지었다.

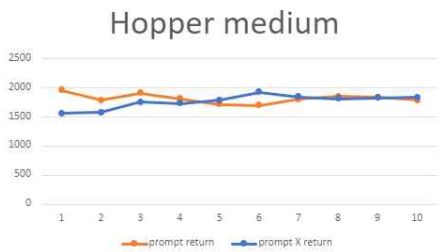

그림 5 Hopper에서 medium dataset을 사용했을 때 Return값 비교

Ⅲ. 결론

본 논문에서는 최근 강화학습에서 주류인 Decision Transformer 모델의 성능 향상을 위해 Offline Reinforcement Learning에서 사용되는 expert dataset에서의 trajectory의 마지막 state, action, return-to-go 세트를 마치 goal로서 활용하였다. OpenAI의 Mujoco의 Hopper와 Ant 환경에서 실험을 진행하였고 training 속도가 향상된 것을 실험적으로 확인 할 수 있었다.

ACKNOWLEDGMENT

This work is in part supported by National Research Foundation of Korea (NRF, 2021R1A2C2014504(50%)). National R&D Program through the National Research Foundation of Korea(NRF) funded by Ministry of Science and ICT(2021M3F3A2A02037893(50%)), INMAC, and BK21 FOUR program.

참 고 문 헌

[1] Lili Chen. Decision Transformer: Reinforcement Learning via Sequence Modeling, NeurIPS 2021 Poster.