Decision Transformer 의 성능과 하이퍼 파라미터의 변화

주석훈, 이정우*
서울대학교

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

Performance of Decision Transformer and hyperparameter variation

Ju Seok Hun, Lee Jung Woo*
Seoul National Univ.

요 약

최근의 많은 연구에서 Transformer 을 사용한 모델이 시퀀스 모델링 문제에서 좋은 성능을 보여주고 있다. 이러한 영향으로, 강화학습 연구에서도 기존의 벨만 방정식에 기반한 알고리즘으로 접근하는 방법 뿐만 아니라 강화학습 문제를 시퀀스 모델링 문제로 정의하고, Transformer 를 이용하여 정책을 직접 학습하는 모델이 제안되었다. 본 연구에서는 제안된 Transformer 기반 강화학습 모델 중 Decision Transformer 에서 모델 크기와 입력 시퀀스의 길이 등 하이퍼 파라미터 변화에 따른 성능을 분석하고 하이퍼 파라미터 변화에 대한 강인함을 확인한다.

Ⅰ. 서 론

시퀀스 모델링 문제에서 Transformer[1]기반 모델은 좋은 성능을 보여주고 있다. Transformer 모델을 기반으로 거대 언어 모델이 개발되고, GPT 와 같은 모델은 최근 자연어 처리 연구에서 널리 사용되고 있다. 강화학습에서 다루는 순차적 의사 결정 문제 또한 시퀀스 모델링 문제로 접근하고자 하는 시도가 이루어 졌다. 대표적으로 Decision Transformer[2] 모델과 Trajectory Transformer[3] 모델이 제안되었다. 본 연구에서는 제안된 Transformer 기반 강화학습 모델에서 모델 구성과 관련된 주요 하이퍼 파라미터의 변화에 따른 성능의 변화를 확인하고, 하이퍼 파라미터 변화에 대한 강인함을 확인한다.

Ⅱ. 본론

Transformer 기반 강화학습 모델로는 대표적으로 Decision Transformer 와 Trajectory Transformer 를 들 수 있다. 두 모델 중에서 본 연구에서는 Decision Transformer 를 기반으로 연구를 진행하고, 분석한다. Decision Transformer 의 구조는 아래 그림 1 과 같다.

Decision Transformer 는 입력으로 Return-to-go 와 상태, 행동의 시퀀스를 받아 각 타임스텝의 행동을 출력하도록 작동한다. 학습 단계에서는 실제 각 타임스텝의 행동과 모델에서 출력한 각 타임스텝의 행동 사이의 차이를 손실함수로 하여 학습을 진행하고, 추론 단계 에서는 초기 상태와 목표로 하는 Return-to-go 를 입력하여 행동을 추론하고, 추론한 행동을 환경에서 수행하여 얻게되는 상태와 보상을 다시 입력해 다음 타임스텝의 행동을 추론한다.

기존의 벨만 방정식에 기반한 강화학습 알고리즘과 이를 구현한 딥러닝 모델의 경우, 학습 단계에서 상태, 행동, 보상, 다음 상태로 구성된 시퀀스 길이가 1 인 데이터를 이용해서 모델을 학습시킨다. 또한 구현에 사용되는 딥러닝 모델의 경우 주로 간단한 MLP 구조를 주로 사용하여 구현한다.

반면, Transformer 를 기반으로 한 모델은 시퀀스 길이가 긴 입력을 학습에 사용하고, 여러 레이어의 Attention 모듈을 사용하여 파라미터수가 MLP 를 기반으로 한 모델보다 많으므로 이 차이점과 관련된 하이퍼 파라미터를 변화시키면서 실험하여 성능의 변화를 확인하고, 파라미터 변화에 대해 얼마나 강인한 성질을 보이는지 분석한다.

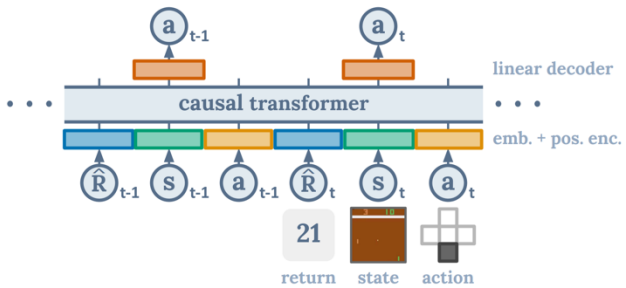

그림 1 Decision Transformer 개요도 [2]

Ⅲ. 실험 및 결론

실험은 Atari 게임 환경에서 진행하였다. Atari 환경의 여러 게임 중에서 Decision Transformer 의 실험 결과로 보고 된 환경 중 Breakout, Seaquest 환경에서 실험을 진행하였다.

하이퍼 파라미터의 경우 모델의 파라미터 수에 직접적인 영향을 미치는 Attention 레이어의 수와 시퀀스 길이를 변화시키면서 실험하고, 나머지 하이퍼 파라미터의 경우 Decision Transformer 논문에서 사용한 수치를 그대로 사용하였다. 입력 시퀀스 길이의 경우 기본값은 30 으로 하고 길이 10, 50 인 경우에 대해 실험하고, Attention 레이어 수의 경우 기본 값은 6 으로 설정하고 4, 8 인 경우에 대해 실험하여 성능을 비교하였다.

실험 결과는 아래와 같다. 두 게임의 실험 결과에서 레이어 수의 경우 공통적인 경향성이 보이지 않는 반면, 시퀀스 길이의 경우 기본 값인 30 이나 길이가 긴 50 에 비해 길이가 10 인 짧은 시퀀스에서 길이가 긴 경우들과 비교해서 비슷하거나 더 좋은 성능이 나오는 것을 확인할 수 있다.

강화학습 환경에서 Transformer 구조를 이용하면, 시퀀스 길이가 긴 입력에 대해 타임스텝 사이에서 Attention 을 통해 맥락을 파악하고, 이 정보를 기반으로 행동을 추론한다. 따라서 입력의 길이가 길수록 많은 정보를 활용할 수 있어 좋은 성능을 보일 것으로 기대할 수 있는데, 환경의 특성에 따라 짧은 길이의 입력이 더 좋은 성능을 보일 수 있음을 확인할 수 있다.

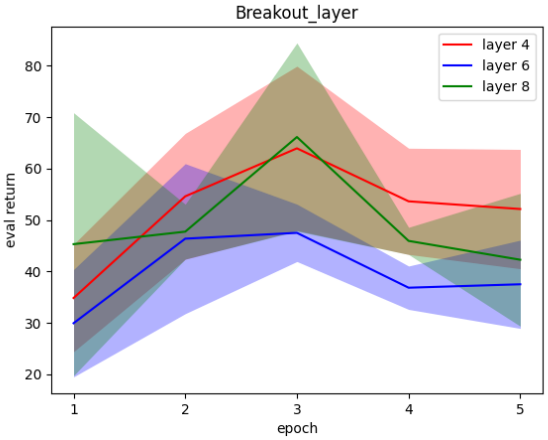

그림 2 Breakout 레이어 수 변화

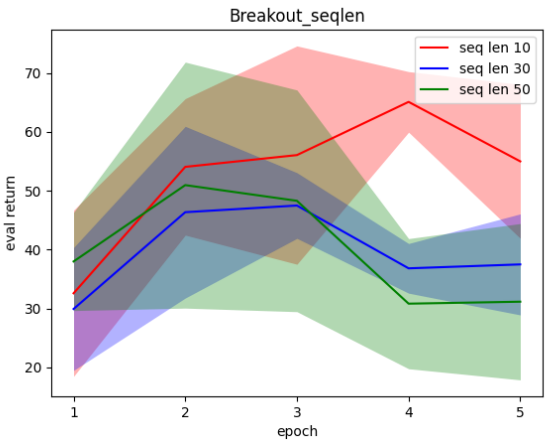

그림 3 Breakout 시퀀스 길이 변화

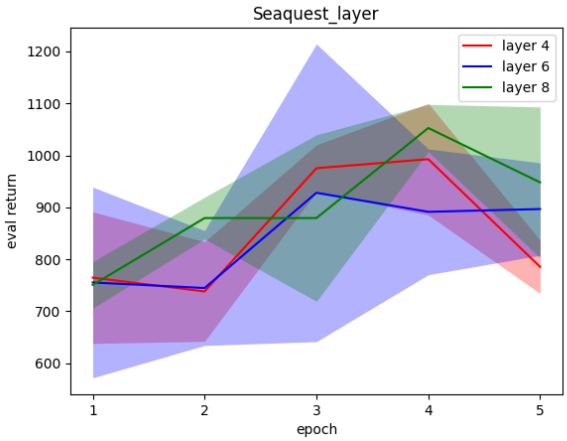

그림 4 Seaquest 레이어 수 변화

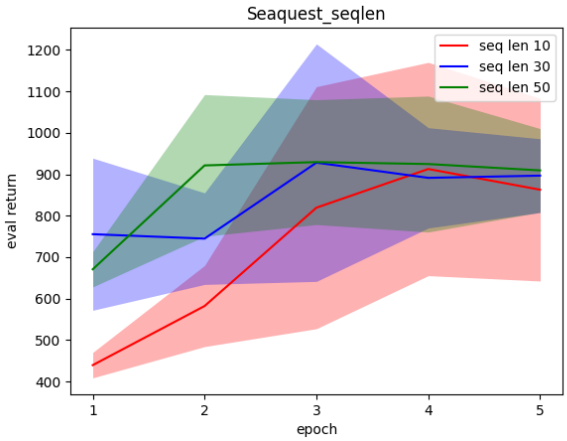

그림 5 Seaquest 시퀀스 길이 변화

ACKNOWLEDGMENT

This work is in part supported by National Research Foundation of Korea (NRF, 2021R1A2C2014504(34)), Institute of Information & communications Technology Planning & Evaluation (IITP- 2021-0-00106(33),IITP-2021-0-01059(33)) grant funded by the Ministry of Science and ICT (MSIT), INMAC, and BK21 FOUR program.

참 고 문 헌

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł ., and Polosukhin, I. Attention is all you need. In Advances in neural information processing systems, pp. 5998- 6008, 2017.

[2] 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 Neural Information Processing Systems (NeurIPS), 2021.

[3] Janner, M., Li, Q., and Levine, S. Offline reinforcement learning as one big sequence modeling problem. In Advances in Neural Information Processing Systems, 2021.