강화학습의 생성적 적대 모방 학습 모델의 Discriminator에 분포 기반 강화학습을 적용시킨 효과에 관한 연구

김형진, 이정우*
서울대학교, *서울대학교
hjkim@cml.snu.ac.kr, *junglee@snu.ac.kr

A Study on the Application of Distributional RL to Discriminator of Generative Adversarial Imitation Learning model

Kim Hyung Jin, Lee Jung Woo*
Seoul National Univ., *Seoul National Univ.

요 약

본 논문은 강화학습의 생성적 적대 모방 학습 모델인 Generative Adversarial Imitation Learning 모델의 Discriminator의 output으로 sigmoid function을 대신하여 분포 기반 강화학습인 Distributional Reinforcement Learning의 방식을 적용하여 하나의 scalar 값이 아닌 distribution으로 예측하는 알고리즘을 제안하였다. 강화학습 실험환경인 Mujoco의 3 가지 환경에서 실험을 진행하여 기존 Generative Adversarial Imitation Learning과 비교하여 다소의 초기학습 속도 향상 또는 학습 최고점 향상을 확인하였다.

Ⅰ. 서 론

강화학습의 생성적 적대 모방 학습 모델인 Generative Adversarial Imitation Learning 모델은 Imitation Learning 분야와 Inverse RL 분야에서 획기적인 모델이다.[1] Generative Adversarial Imitation Learning 모델의 Discriminator의 output을 reward로써 사용하는데 이 output function에 sigmoid function 대신 Distributional Reinforcement Learning 방식을 채택하여 좀 더 다양한 상황에 대처할 수 있는 알고리즘을 고안하였다.

Ⅱ. 본론

Generative Adversarial Imitation Learning은 적대적 생성 방법인 GAN의 방식을 Imitation Learning에 적용시킨 알고리즘이다. 새로운 state action pair를 만들어내는 Generator와 이를 expert의 state action pair인지 아닌지를 판단하는 Discriminator의 적대적 생성 학습을 통해 학습시킨다.

$$\mathbb{E}_{\pi}[log(D(s,a))] - \mathbb{E}_{\pi_{\mathbb{E}}}[log(1-D(s,a))] - \lambda H(\pi)$$

수식 1 GAN 방식을 토대로 만든 Generative Adversarial Imitation Learning의 Discriminator의 loss function

$D(s,a)$는 Discriminator로 output이 1에 가까우면 Generator, 0에 가까우면 expert의 satate action pair라고 구별하도록 학습시킨다. 이를 통해 Generative Adversarial Imitation Learning은 expert의 state action pair만을 통해서 Generator로 expert에 가까운 policy를 만들어냄과 동시에 Discriminator에서의 output에 log를 취한 값인 cost function 또한 얻을 수 있는 Imitation Learning과 Inverse RL을 동시에 행하는 알고리즘이다.

Distributional Reinforcement Learning은 value function Q의 scalar 값

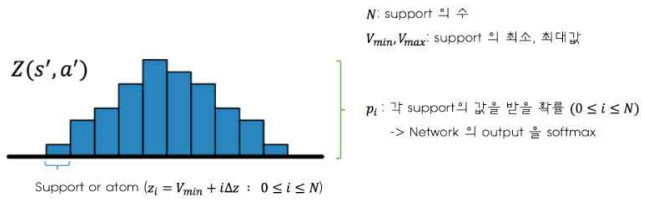
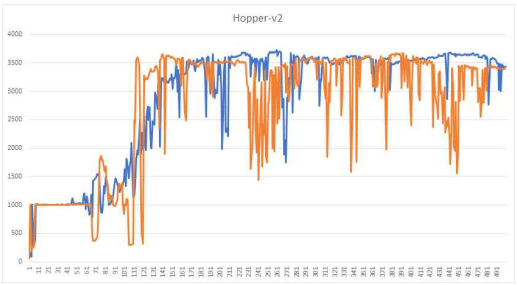


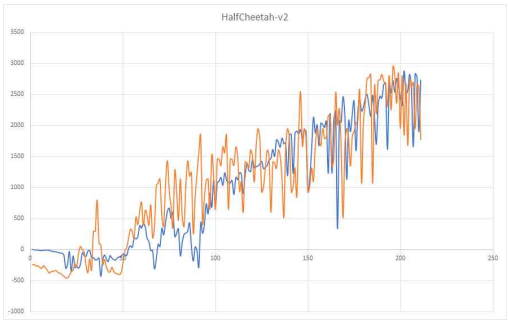
그림 1 Distribution RL은 distribution function Z로 scalar 값이 아닌 distribution을 output하여 사용한다

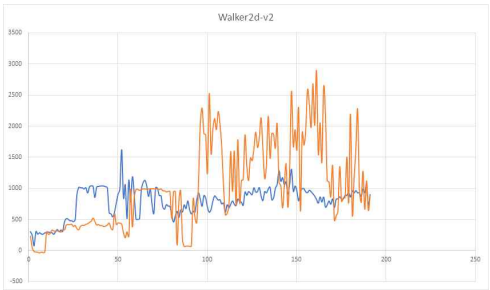
대신 distribution function Z를 사용하여 환경이 좀 더 랜덤성이 있거나 어려울 경우에도 대처할 수 있는 강화학습이다.

우리는 이 Generative Adversarial Imitation Learning 알고리즘의 Discriminator의 output이 sigmoid fucntion을 통해 0부터 1로 output하는 대신에 Distributional RL의 방식을 채택하여 결과적으로 Generative Adversarial Imitation Learning의 cost function에 영향을 끼쳐 환경이 좀 더 랜덤성이 있거나 어려울 경우에도 대처할 수 있게끔 하는 목적으로 이 알고리즘을 고안하였다.

우리는 Distributional RL 분야에서의 c51 논문 알고리즘을 채택했고 이를 Generative Adversarial Imitation Learning의 Discriminator의 output에 적용시켰다.[2] Output이 0에서 1사이인 sigmoid function을 대체하는 것이므로 min, mix를 0과 1로 잡고 support의 수를 8로 잡고 각각의 support가 나올 확률 p_i를 인공신경망 softmax function을 통해 output하였다. 실제 output 되는 $D(s,a)$의 값은 $\sum_i z_i p_i(s_t,a_t)$로 정하고 그 이후로는 Generative Adversarial Imitation Learning 알고리즘과 동일하게 진행된다. 실험은 강화학습 실험환경인 Mujoco의 Hopper, HalfCheetah, Walker2d에서 진행되었고 뒤로 갈수록 보통 학습 난이도가 어려운 실험환경이다.







위 3가지 그래프는 각각 Hopper-v2, HalfCheetah-v2, Walker2d-v2의 환경에서 실험한 Generative Adversarial Imitation Learning과 본 논문의 알고리즘을 비교하는 학습그래프이다. x축은 total iteract수를 y축은 학습도중 evaluation에서의 Return값을 나타낸다. 각각 파란선이 Generative Adversarial Imitation Learning, 주황선이 본 논문의 알고리즘의 학습그래프를 보여준다. Hopper-v2와 HalfCheetah-v2 환경의 경우에는 Generative Adversarial Imitation Learning와 비교하여 본 논문의 알고리즘의 학습 최고점은 동일하지만 초기 학습 속도가 좀더 빠른 것을 확인할 수 있었고, Walker2d-v2 환경의 경우에는 초기 학습속도는 비슷하나 학습 최고점에서 차이가 분명히 드러나는 것을 확인 할 수 있었다.

Ⅲ. 결론

본 논문에서는 Generative Adversarial Imitation Learning 모델의 Discriminator의 output으로 Distributional Reinforcement Learning의 방식 중 하나인 c51 방식으로 바꾸는 알고리즘을 제시하였다. 본 논문에서 제시한 알고리즘으로 Mujoco의 3가지 환경에서 기존 Generative Adversarial Imitation Learning 알고리즘과 성능 비교를 했을 때 초기 학습 속도 또는 학습 최고점에 있어서 다소 향상된 성능을 확인할 수 있었다.

ACKNOWLEDGMENT

This work is in part supported by National Research Foundation of Korea (NRF, 2021R1A4A1030898(25%)), Institute of Information & communications Technology Planning & Evaluation (IITP, 2021-0-00106(25%), 2021-0-02068(25%)) grant funded by the Ministry of Science and ICT (MSIT), Center for Applied Research in Artificial Intelligence(CARAI, UD190031RD(25%)) grant Funded by Defense Acquisition Program Administration(DAPA), Agency for Defense Development(ADD), INMAC, and BK21-plus.

참 고 문 헌

[1] Ho, Jonathan, and Stefano Ermon. "Generative adversarial imitation learning." Advances in neural information processing systems 29 (2016).

[2] SBellemare, Marc G., Will Dabney, and Rémi Munos. "A distributional perspective on reinforcement learning." International Conference on Machine Learning. PMLR, 2017.