Deep Q Networks 성능 향상을 위한 사인파형 Epsilon 감쇠에 관한 연구

유승찬, 이정우
서울대학교

scyu@cml.snu.ac.kr, junglee@snu.ac.kr

A Study on the Sinusoidal Epsilon Decay for Improving Deep Q Networks

Seung Chan Yu, Jung Woo Lee
Seoul National Univ.

요 약

본 논문에서는 Deep Q Networks(DQN)에서 Exploration 정도를 조절하는 상수인 Epsilon 의 감쇠(Decay) 방법을 새롭게 제안한다. 기존에 일반적으로 지수 감쇠로 설정하던 Epsilon 에 사인파를 더해 진동하면서 감쇠하도록 만듦으로써, DQN 의 과추정(Overestimation)을 방해하여 학습 성능을 향상시키는 것을 목적으로 한다.

Ⅰ. 서 론

Deep Q Network (DQN) 알고리즘의 등장으로, Atari 를 비롯한 다양한 환경에서 Q-value 추정(Estimation)을 이용한 다양한 학습이 가능하게 되었다.[1] 하지만 OpenAI Gym[2] 등의 새로 등장한 복잡하고 어려운 벤치마크 환경에서는 학습이 잘 되지 않는 문제점을 보였다. 가장 큰 원인으로는 DQN 의 과적합(Overestimation) 문제가 꼽힌다. 과적합이란 학습 중에 한번 특정 state-action 의 Q-value 가 크게 계산되면, 그 이후로는 해당 state-action 만 선택하게 되는 현상을 말한다. 특정 state-action 에만 고착되면 더 이상 탐험(Exploration)을 하지 못하고 학습에 실패하게 된다. 본 논문에서는 탐험 정도를 조절하는 상수인 ε(Epsilon)의 감쇠(Decay) 방법을 새롭게 설계하였다. 기존의 ε 감쇠는 지수 감쇠 (Exponential decay) 방법을 쓰는데, 여기에 사인파형을 더하여 진동하면서 감쇠하도록 설계하였다. 이를 통해 DQN 의 과추정을 방해하여 학습성능에 개선이 있는지, 기존의 DQN 과의 학습 성능을 비교해보았다.

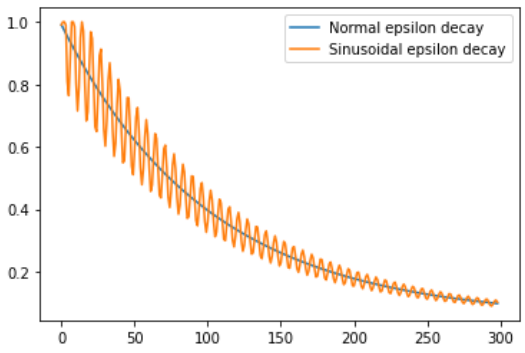
Ⅱ. 본론

제안 방법

DQN 은 특정 상태(State)에서 행동(Action)을 선택할 때, (1- ε)의 확률로 Q-value 가 가장 큰 행동을 선택하고, ε의 확률로 이외의 행동 중 무작위로 선택한다. 즉, ε값이 클수록 정답이라고 생각하는 행동 이외의 행동을 선택해보는 탐험을 많이 하게 된다. 물론 학습이 어느정도 진행된 후에는 ε 값이 0 에 가까운 것이 좋지만, 학습 초반에는 ε 값이 1 에 가까운 것이 학습에 더 도움이 된다. ε 이 1 일때는 균등분포(Uniform distribution)에서 행동을 선택하는 것과 수식적으로 같다. 때문에 학습초기에는 1 로 시작해서, 스텝(step) 이 지나면서 최종 ε 값(ε_{term}) 으로 지수함수적으로 점점 감소시키는 방법을 일반적으로 많이 사용한다. 그 식은 아래와 같다.

$$\varepsilon_{\exp\_decay} = \varepsilon_{term} + \left(1 - \varepsilon_{term}\right) * \exp\left(-\frac{current\_step}{decay\_weight}\right)$$

여기서 ε_{term} 는 학습 후반에도 유지되는 최소의 ε 값이고, decay_weight 는 ε의 감쇠 정도를 조절하는 하이퍼파라미터(Hyperparameter) 이다. ε 은 1 부터 시작하여 ε_{term} 까지 지수감쇠하고, decay_weight 이 커질수록 더 빠르게 감소한다. 아래 초반 300 스텝까지의 ε을 파란색 그래프로 나타냈다. ε_{term} 는 0.05 이고, decay_weight 은 100 으로, 실제 실험에서 사용한 값을 사용하여 그래프를 그렸다.



그래프의 주황색 선은, 본 논문에서 제시한 사인파형 epsilon 감쇠이다. 지수함수적으로 감쇠하는 기존 함수에 사인파형을 더하여 진동하면서 감쇠하도록 설계하였다. 수식은 아래와 같다.

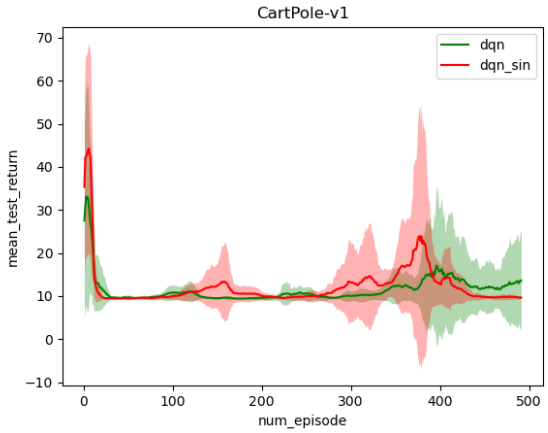
$$\varepsilon_{\text{sinusoidal\_decay}} = \varepsilon_{term} + \left(1 - \varepsilon_{term} + \frac{\sin(current_{step} - 1)}{\sin\_weight}\right) * \exp(-\frac{current\_step}{decay\_weight})$$

$$\varepsilon_{sinusoidal\_decay\_clip} = \min(1, \varepsilon_{sinusoidal\_decay})$$

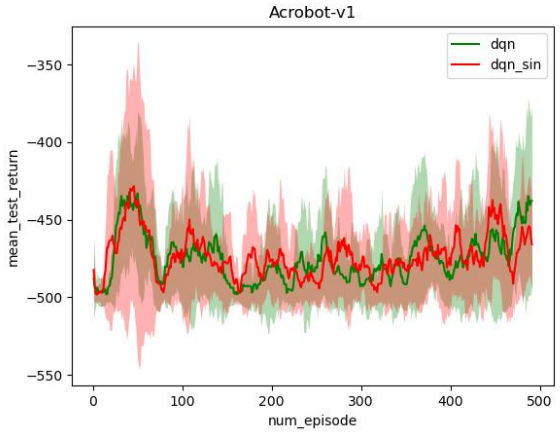
sin_weight 는 sin 함수가 더해지는 비중이고, 본 논문의 실험에서는 5 로 설정하였다. ε 이 1 보다 커지면, 1 로 선택하도록 min 함수를 취해주었다.

실험 결과

실험은 OpenAI gym 의 CartPole-v1 과 Acrobot-v1 에서 진행하였다. Cartpole-v1 에서의 결과는 아래와 같다.



녹색이 기존 방법, 붉은색이 본 논문에서 제안한 방법으로, 10 개의 무작위 시드(Random seed)로 학습을 진행하고, 평균값을 진한 선으로 표시하였다. Cartpole-v1 환경에서는 본 논문에서 제안한 방법, 즉 사인파형 Epsilon 감쇠가 더 좋은 성능을 보임을 확인할 수 있었다. 다음으로 Acrobot-v1 환경에서의 결과는 다음과 같다.



마찬가지로 녹색이 기존 지수감쇠 방법, 붉은색이 본 논문에서 제안한 사인파형 감쇠 방법이고, 같은 10 개의 무작위 시드로 실험을 진행하였다. 사인파형 감쇠 방법이 미세하게 좋은 성능을 보이긴 하였지만, 유의미한 차이라고 보기는 어려웠다.

Ⅲ. 결론

본논문에서는 DQN 의 고질적인 문제인 과추정 문제를 해결하기 위해, 기존의 지수함수를 이용한 ε 감쇠 방법에 사인파형을 더하여 진동시키며 실험을 진행해 보았다. 비교적 단순한 환경인 Cartpole-v1 환경에서는 성능의 개선을 볼 수 있었다. 하지만 상태가 훨씬 복잡한 Acrobot-v1 환경에서는 유의미한 차이를 보기 어려웠다. 복잡한 환경일수록 감쇠함수를 더 복잡하게 설계할 필요성을 확인하였다.

.

ACKNOWLEDGMENT

This work is in part supported by National Research Foundation of Korea (NRF, 2021R1A2C2014504(33%)), Institute of Information & communications Technology Planning & Evaluation (IITP, 2021-0-00106(33%), 2021-0-02068(33%)) grant funded by the Ministry of Science and ICT (MSIT), National R&D Program through the National Research Foundation of Korea(NRF) funded by Ministry of Science and ICT(2021M3F3A2A02037893), INMAC, and BK21-plus.

참 고 문 헌

[1] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2016). Openai gym. arXiv preprint arXiv:1606.01540.

[2] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.