Max-Mean N-스텝 시간차 학습

황규영*, 김주봉*, 허주성*, 한연희*†
*한국기술교육대학교 미래융합공학전공

{to6289, rlawnqhd, chil1207, yhhan}@koreatech.ac.kr

Max-Mean N-step Temporal Difference Learning

Gyu-Young Hwang*, Ju-Bong Kim*, Joo-Seong Heo*, Youn-Hee Han*
*Future Convergence Engineering, Korea University of Technology and Education.

요 약

적절한 n을 선택할 경우 n-스텝 시간차 학습은 몬테카를로 방법과 1-스텝 시간차 학습보다 성능이 좋은 알고리즘으로 알려져 있지만 최적의 n값은 파라미터에 민감하며 편향-분산 트레이드오프 문제가 있기 때문에 n을 선택하는 것에 어려움이 있다. 기존 n-스텝 시간차 학습에서 n값 선택의 어려움을 해소하기 위해, 본 논문에서는 $1 \le k \le n$에 대한 모든 k-스텝 누적 보상의 최댓값과 평균으로 구성된 새로운 학습 타겟인 $\Omega - return$을 제안한다. 마지막으로 OpenAI Gym 의 Atari 게임 환경에서 기존 n-스텝 시간차 학습과의 성능 비교 평가를 진행하여 본 논문에서 제안하는 알고리즘이 기존 n-스텝 시간차 학습 알고리즘보다 성능이 우수하다는 것을 입증한다.

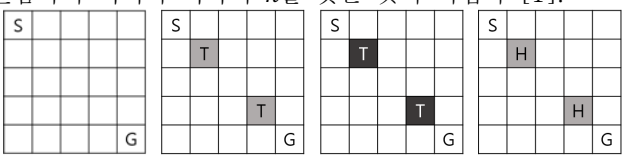
Ⅰ. 서 론

몬테카를로 방법은 완전 누적 보상을, 1-스텝 시간차 학습은 1-스텝 누적 보상을 학습의 업데이트 타겟으로 사용한다. n-스텝 시간차 학습은 몬테카를로 방법과 1-스텝 시간차 학습을 결합한 것으로, n-스텝까지 관찰된 누적 보상과 n 번째 스텝에서 기대되는 행동 가치가 결합된 값을 학습의 업데이트 타겟으로 사용한다. n 의 값을 적절하게 선택할 경우 n-스텝 시간차 학습은 강화 학습의 성능을 높일 수 있어 1-스텝 시간차 학습보다 좋을 수 있다. 하지만 파라미터 n은 강화 학습 환경의 변화에 대하여 민감하여 최적의 n을 찾는 것이 어렵고, n의 값이 커지면 n-스텝 누적 보상의 분산이 커지고 n 의 값이 작아지면 편향이 발생하는 편향-분산 트레이드오프 문제가 있다. 본 논문에서는 n -스텝 시간차 학습에서 최적의 n 을 선택해야 하는 문제를 해결하기 위해, $\Omega - return$ 이라는 새로운 n -스텝 업데이트 타겟을 제안한다.

Ⅱ. 배경

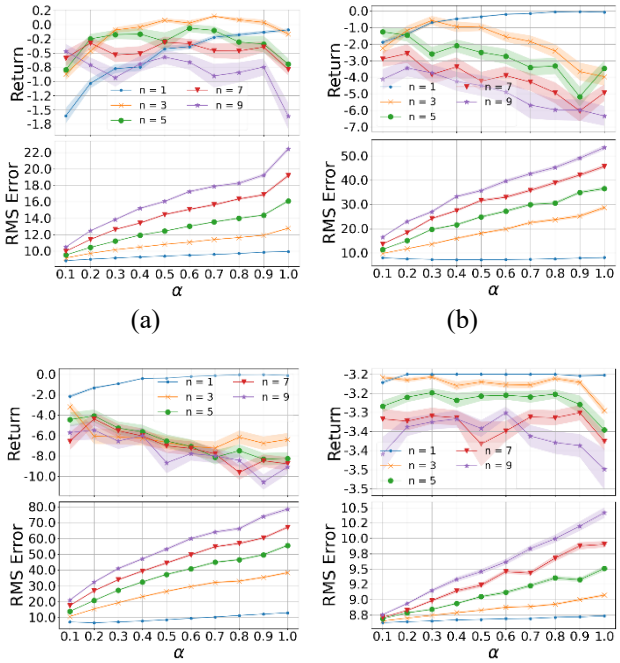
1. n-스텝 시간차 학습의 파라미터 민감성

(그림 1)은 네 가지의 5×5 그리드월드 환경이다. 매 스텝마다 -0.1 의 보상을 받고, 'S'에서 출발해 'G'에 도달하면 1.0 의 보상을 받으며 에피소드가 끝난다. (b)의 'T'에 도달하면 -3.0, (c)의 'T'에 도달하면 -6.0, (d)의 'H'에 도달하면 -3.0 의 보상을 받으며 에피소드가 종료된다. (그림 2)는 (그림 1) 환경에서의 실험 결과이다. 실험 결과와 같이 n은 실험 환경과 학습률 α의 변화에 민감하여 최적과 최악의 n을 찾는 것이 어렵다 [1].



(a) (b) (c) (d)

(그림 1) 네 가지의 5×5 그리드월드



(a) (b) (c) (d)

(그림 2) 네 가지의 5×5 그리드월드에서의 실험 결과

2. 편향-분산 트레이드오프

Q-learning n -스텝 누적 보상은 다음의 2 가지로 분리하여 볼 수 있다.

$$\sum_{k=1}^{n} \gamma^{k-1} R_{t+k} \quad (1)$$

$$\gamma^{n} \max_{a} Q_{t+n-1}(S_{t+n}, a) \quad (2)$$

수식(1)은 에이전트가 환경과 상호작용하며 받은 보상을 의미하며, 수식(2)는 행동 가치 함수 Q 를 사용하여 누적 보상을 추정한 값이다. n의 값이 커지면 수식(1)에 의해 확률적으로 행동하며 받은 보상의 비중이

† 교신 저자 한연희

커지므로 n-스텝 누적 보상의 분산이 커지게 된다. 반대로 n의 값이 작아지면, Q는 학습 중인 에이전트가 받을 누적 보상을 추정한 값이기 때문에 실제 값과는 차이가 있고 이로 인해 편향이 발생하게 된다 [2].

Ⅲ. 본론

본 논문은 $1 \le k \le n$에 대한 모든 k-스텝 누적 보상의 최댓값과 평균으로 구성된 $\Omega - return$이라는 새로운 n-스텝 업데이트 타겟을 제안한다.

제안하는 방법은 n-스텝 누적 보상에 기초한 가장 좋은 업데이트 타겟을 결정하는 것에서 시작된다.

Q-learning 알고리즘의 학습 타겟인 1-스텝 누적 보상을 보면 $\max_a Q_t(S_{t+1}, a)$가 사용된다. Q의 max를 취하는 것은 실제 값보다 과대평가되어 편향이 발생해 문제를 야기할 수 있지만, 만약 높은 보상을 받을 수 있는 상태-행동에 대한 Q 값이 과대평가되어 있다면 에이전트가 해당 영역을 탐험하도록 장려하여 학습에 이점을 줄 수 있다 [3]. 따라서 $1 \le k \le n$에 대한 모든 $G_{t:t+k}$ 값들의 최댓값을 G_{max}라 명명하고 다음과 같이 정의한다.

$$G_{max} = \max_k G_{t:t+k}, \quad 1 \le k \le n \tag{3}$$

만약 모든 상태에 대한 최적의 행동 가치 Q^*를 구했다면, $1 \le k \le n$에 대한 모든 $G_{t:t+k}$는 같은 값을 가질 것이다. 상태 S_t에서 행동 A_t가 최적 정책을 따른다면 다음과 같이 표기할 수 있다.

$$\frac{1}{n}\left[\sum_{k=1}^{n} G_{t:t+k}\right] - Q(S_t, A_t) \approx 0 \tag{4}$$

수식(4)로부터 $1 \le k \le n$에 대한 모든 $G_{t:t+k}$ 값들의 평균이 업데이트 타겟으로 좋다는 것을 알 수 있다. 따라서 이 평균을 G_{mean}이라 명명하고 다음과 같이 정의한다.

$$G_{mean} = \frac{1}{n}\left[\sum_{k=1}^{n} G_{t:t+k}\right] \tag{5}$$

G_{max}는 초기 단계에 학습을 가속화 시킬 수 있지만 Q값을 과대평가하는 경향이 있어 학습하는 동안 문제가 될 수 있다 [4]. 반면, G_{mean}은 학습 후반 업데이트의 타겟으로써 좋다. 그러므로 G_{max}와 G_{mean}을 결합한 업데이트 타겟인 $\Omega - return$을 제안한다.

$$\Omega - return = \beta G_{max} + (1-\beta) G_{mean}, \quad 0 \le \beta \le 1 \tag{9}$$

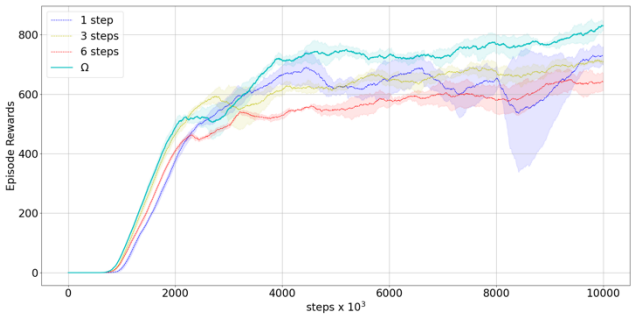
β는 G_{max}와 G_{mean}의 비율을 결정하는 파라미터이다. 본 논문은 파라미터 β를 다음의 수식으로 대체하였다.

$$\beta = \frac{\max_k |G_{t:t+k}| - \min_k |G_{t:t+k}|}{\max_k |G_{t:t+k}|} \tag{10}$$

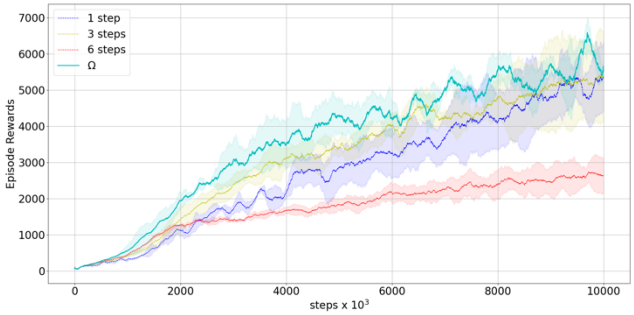
학습 초기에는 β 값이 높아 G_{max}의 비율이 높으며, 학습이 진행되면서 β 값은 줄어들어 G_{mean}의 비율이 높아진다.

Ⅳ. 실험

제안하는 $\Omega - return$의 성능을 평가하기 위해 Nature DQN, Double DQN, Dueling DQN, Prioritized replay buffer 알고리즘들을 통합한 기법 위에서 기존 n-스텝 시간차 학습과의 비교 실험을 진행한다. 실험 환경은 OpenAI Gym Atari 게임 중 Enduro 와 Seaquest 이다. 실험 결과는 (그림 3), (그림 4)와 같으며 3 번 측정한 것의 평균과 표준편차를 그래프에 나타내었다. 실험 결과를 통해 기존 n-스텝 시간차 학습 알고리즘의 성능이 1-스텝 시간차 학습보다 떨어지는 경우에도 본 논문에서 제안하는 알고리즘의 성능이 좋다는 것을 알 수 있다.



(그림 3) OpenAI Gym Enduro 환경에서의 실험 결과



(그림 4) OpenAI Gym Seaquest 환경에서의 실험 결과

Ⅴ. 결론

n-스텝 시간차 학습은 하이퍼-파라미터 n을 선택하기 어려운 문제가 있다. 본 논문에서는 $1 \le k \le n$에 대한 모든 n-스텝 누적 보상의 최댓값과 평균으로 구성된 $\Omega - return$이라는 새로운 n-스텝 업데이트 타겟을 제안한다. 실험은 DQN 알고리즘의 기본이 되는 Nature DQN 과 DQN 의 확장 알고리즘인 Double DQN, Dueling DQN, Prioritized replay buffer 알고리즘들을 통합한 기법 위에서 기존 n-스텝 시간차 학습과 제안하는 알고리즘의 성능 비교 평가를 진행하였다. OpenAI Gym Atari 게임 환경에서의 실험 결과 본 논문에서 제안하는 알고리즘이 기존 n-스텝 시간차 학습 알고리즘보다 성능이 우수하다는 것을 입증하였다. 따라서 $\Omega - return$이 기존 n-스텝 시간차 학습 알고리즘에서 적절한 n을 선택해야 하는 문제를 해결하는 새로운 학습 타겟으로 적절하다고 생각된다.

ACKNOWLEDGMENT

이 논문은 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2018R1A6A1A03025526 및 No. NRF-2020R1I1A3065610).

참 고 문 헌

[1] Hessel, M., J. Modayil, H. van Hasselt, et al. Rainbow: Combining Improvements in Deep Reinforcement Learning. In *AAAI*, pages 3215-3222. AAAI Press, 2018.

[2] Hernandez-Garcia, J., Sutton, R. Understanding Multi-Step Deep Reinforcement Learning: A Systematic Study of the DQN Target. 2019. Cite arXiv:1901.07510 Comment: NIPS Deep Learning Workshop 2018.

[3] Lan, Q., Pan, Y., Fyshe, A., White, M. Maxmin Q-learning: Controlling the Estimation Bias of Q-learning. In *ICLR* 2020.

[4] Thrun, S., A. Schwartz. Issues in using function approximation for reinforcement learning. In *Proceedings of the Fourth Connectionist Models Summer School*. Erlbaum, 1993.