Deep Reinforcement Learning with Double Q-learning

AI 키워드	
날짜	@2024년 11월 18일
콘텐츠	논문
태그	유런

- 2015
- Google DeepMind
- DQN에서 target value가 overestimate되는 문제를 해결. 기존 DQN과 target value를 설정하는 부분에서만 차이를 가짐

Background

Q-learning과 DQN

$$Y_t^{\mathrm{Q}} \equiv R_{t+1} + \gamma \max_a Q(S_{t+1}, a; \boldsymbol{\theta}_t) .$$

$$Y_t^{\mathrm{DQN}} \equiv R_{t+1} + \gamma \max_a Q(S_{t+1}, a; \boldsymbol{\theta}_t^-) .$$

Double Q-learning

기존 알고리즘에서는 max 함수가 action을 선택/평가할 때 같은 인자를 사용

→ 더 overestimate한 값을 선택하게 함(=인자가 모두 틀린 값이더라도 더 큰 값만을 선택)
→ overoptimistic한 결과(=실제보다 더 유리하거나 높은 보상을 받을 것이라고 가정)

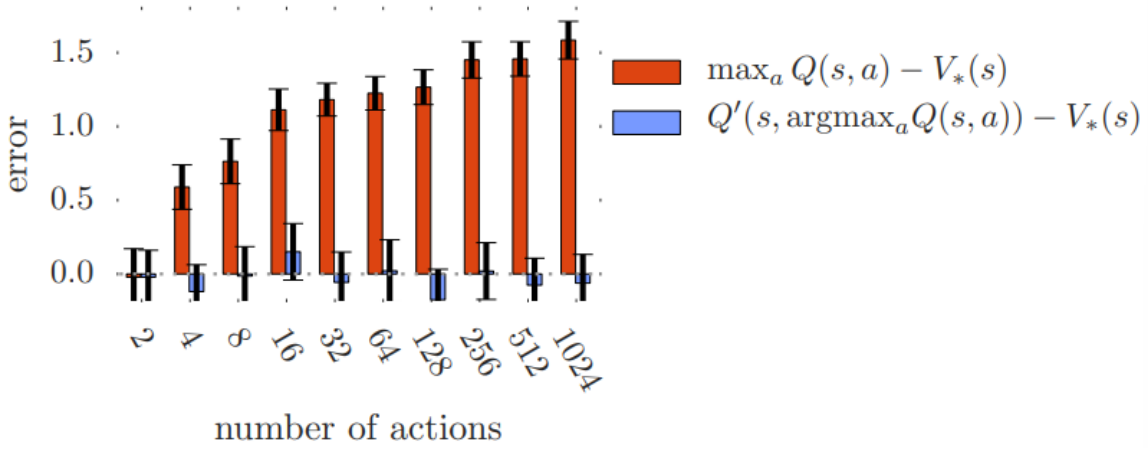
→ 선택($argmax()$)과 평가를 분리

$$Y_t^{\text{DoubleQ}} \equiv R_{t+1} + \gamma Q(S_{t+1}, \underset{a}{\text{argmax}}\, Q(S_{t+1}, a; \boldsymbol{\theta}_t); \boldsymbol{\theta}'_t)$$

Overoptimism due to estimation errors

타겟의 overestimation은 Q(s, a)의 에러 때문에 발생

에러는 양의 방향, 음의 방향 모두 존재하지만 DQN의 타겟은 항상 증가시키는 방향으로 overestimation이 일어남



Double DQN

greedy policy를 실시간 네트워크에 따라 평가. 단, 타겟 네트워크가 value를 추정할 수 있도록 함

기존 알고리즘과는 타켓 Y_T^{DQN}만이 다름

두번째 네트워크인 θ'_t가 θ_t^-로 바뀜 → 현재의 greedy policy 평가

$$Y_t^{\text{DoubleDQN}} \equiv R_{t+1} + \gamma Q(S_{t+1}, \underset{a}{\text{argmax}}\, Q(S_{t+1}, a; \boldsymbol{\theta}_t), \boldsymbol{\theta}_t^-) .$$

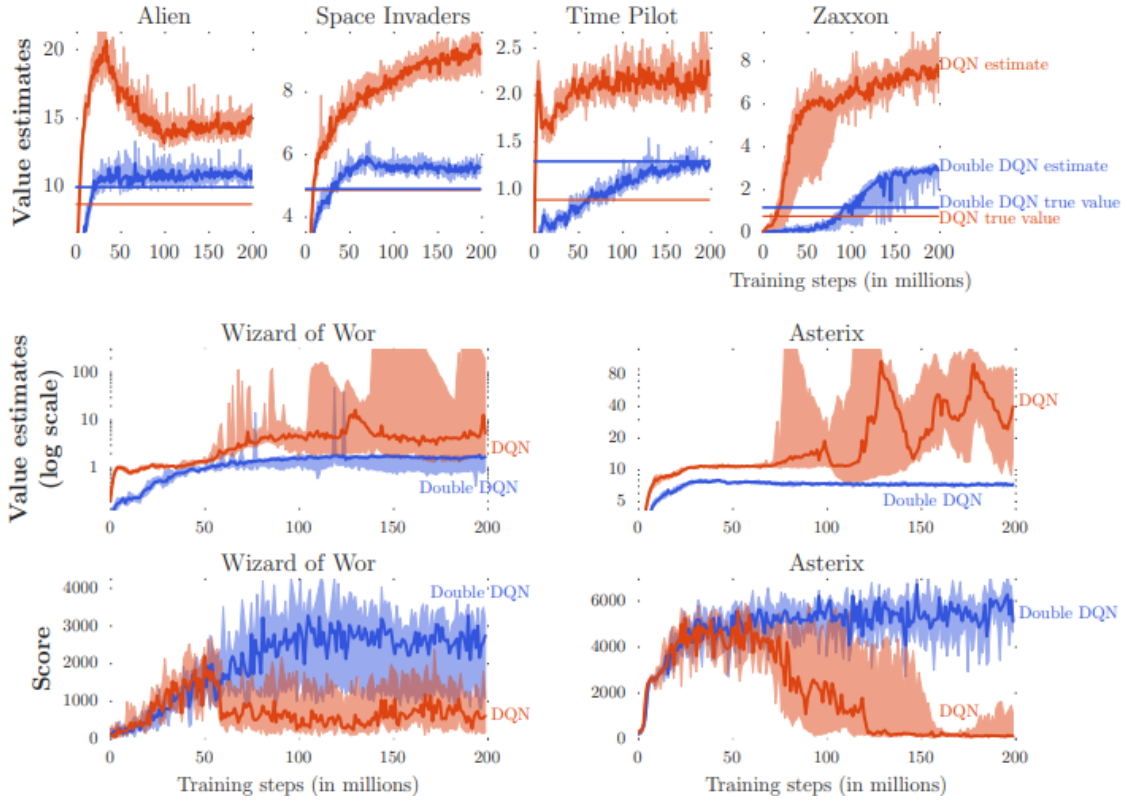
Empirical results

- Atari 2600 사용하여 테스트
- 3 CNN, 1 FC

DQN은 6개의 Atari 게임에서 overestimate

다만 overoptimism이 언제나 학습된 policy 질에 나쁜 영향을 끼치는 것은 아님. 그러나 학습 안정성 측면에서는 overestimation을 줄이는 것이 확실히 좋음

DDQN이 DQN에 비해 높은 score를 가지는 것으로 확인할 수 있음



Discussion

- 과대 추정 문제 해결
- 여러 환경에서 기존 DQN보다 우수한 성능
- 더 안정적인 학습