

DDQN

Deep Reinforcement Learning with Double Q-learning

Google DeepMind

2015

Summary


1. Q-learning은 학습에서 estimation error 때문에 large-scale problem에서 overestimation한다.
2. Atari game에서 value estimation을 통해 overestimation이 생각보다 심하고 흔히 일어남을 확인함.
3. Double Q-learning으로 overestimation을 확실하게 줄일 수 있음.
4. 추가적인 네트워크없이 기존 DQN에서 DoubleDQN으로 발전시킴.
5. 실험 결과, DDQN이 Atari 2600에서 더 좋은 결과를 얻으면서 DQN보다 더 나은 policy를 찾는다는 것을 확인함.

Motivation

Main Problem : Q-learning의 action value overestimation

우리가 알고 있는 standard Q-learning(1989)

$$\theta_{t+1} = \theta_t + \alpha(Y_t^Q - Q(S_t, A_t; \theta_t)) \nabla_{\theta_t} Q(S_t, A_t; \theta_t).$$

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


RL에서 function approximation을 할 때 환경의 noise 영향이 있음.
이 noise가 max operator때문에 action value을 overestimation한다.
Overestimation으로 인해 optimal policy 학습에 지장을 받는다.

Motivation

Solution : Double Q-Learning(2010)

Selection / Evaluation 파라미터 분리시키자.


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

Next state에서의 action value
maximize하는 action

Next state에서 Maximize하는 action의
value

Motivation

기존 Q-learning과 마찬가지로 DQN(2015.2)에서도 overestimation 문제가 남아있음

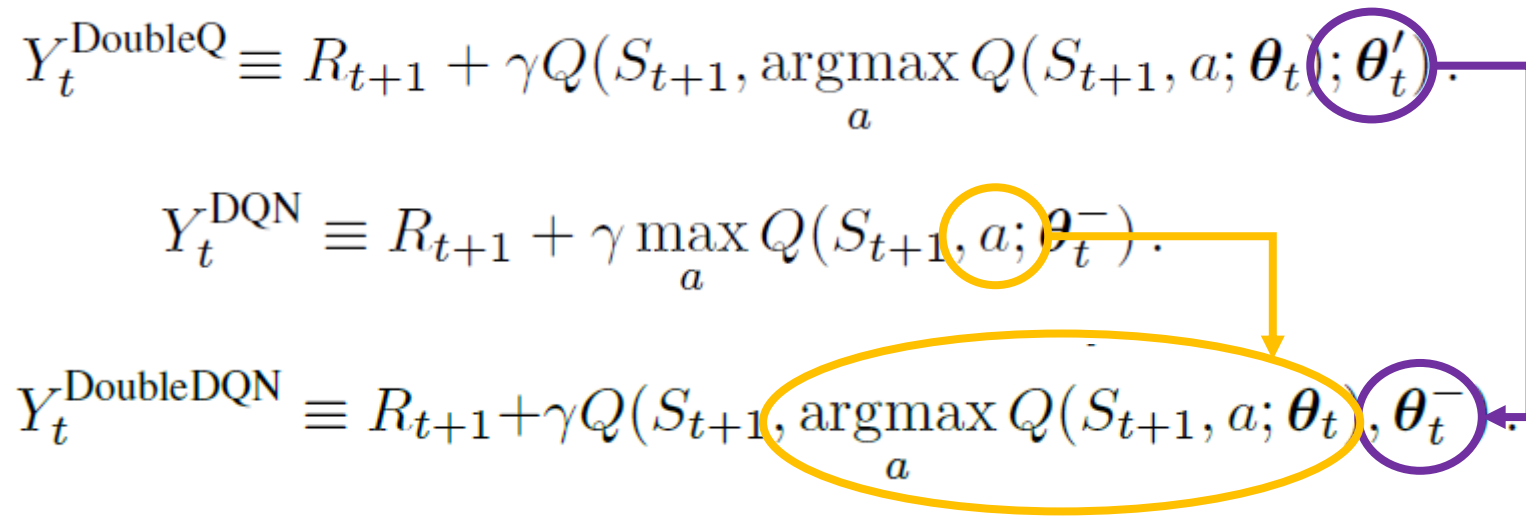
$$Y_t^{\text{DQN}} \equiv R_{t+1} + \gamma \max_a Q(S_{t+1}, a; \theta_t^-).$$


Target network에서 next state에서 action a를 selection 할 때와 Action a의 value를 evaluation할 때 같은 파라미터를 사용함.

Noise가 있는 상태에서 max를 취하니 action-value가 overestimation 될 수 밖에..

Method

DQN도 Double로!

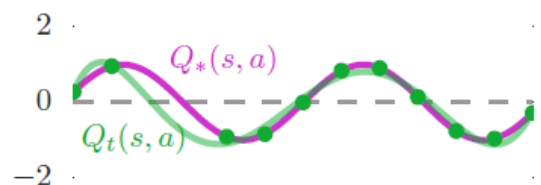
$$Y_t^{\text{DoubleQ}} \equiv R_{t+1} + \gamma Q(S_{t+1}, \underset{a}{\operatorname{argmax}} Q(S_{t+1}, a; \theta_t); \theta'_t).$$
$$Y_t^{\text{DQN}} \equiv R_{t+1} + \gamma \underset{a}{\operatorname{max}} Q(S_{t+1}, a; \theta_t^-).$$
$$Y_t^{\text{DoubleDQN}} \equiv R_{t+1} + \gamma Q(S_{t+1}, \underset{a}{\operatorname{argmax}} Q(S_{t+1}, a; \theta_t), \theta_t^-).$$


Result

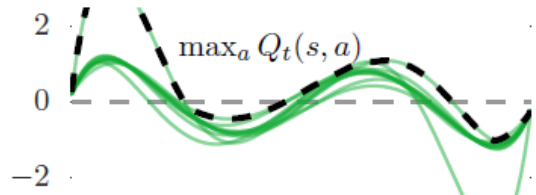
Overestimation 확인

Double-Q unbiased함

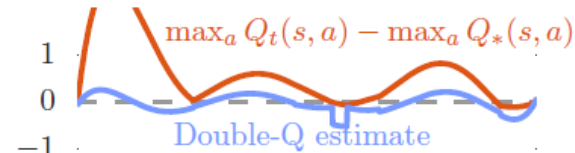
True value and an estimate



All estimates and max



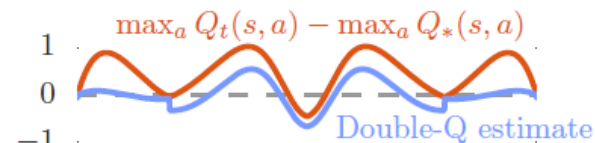
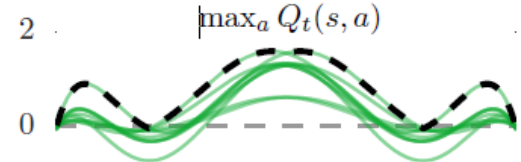
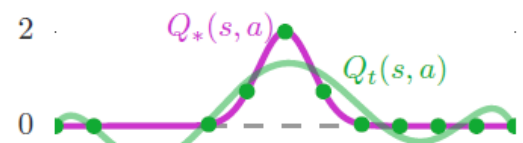
Bias as function of state



Average error

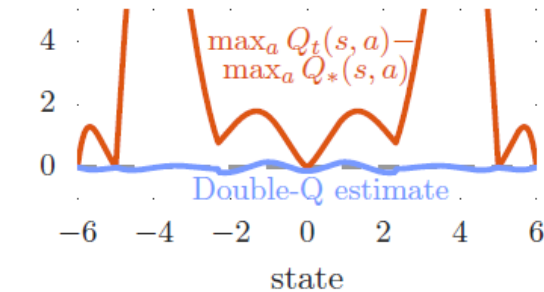
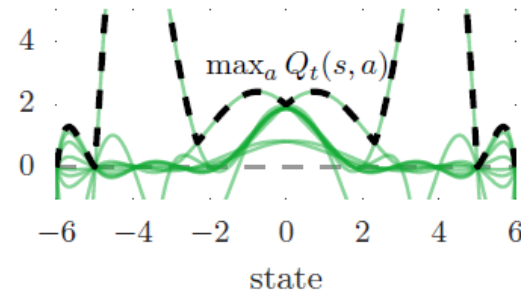
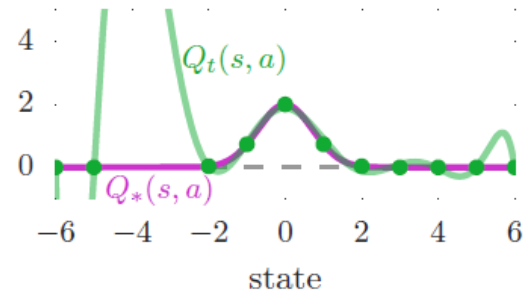
+0.61

-0.02



+0.47

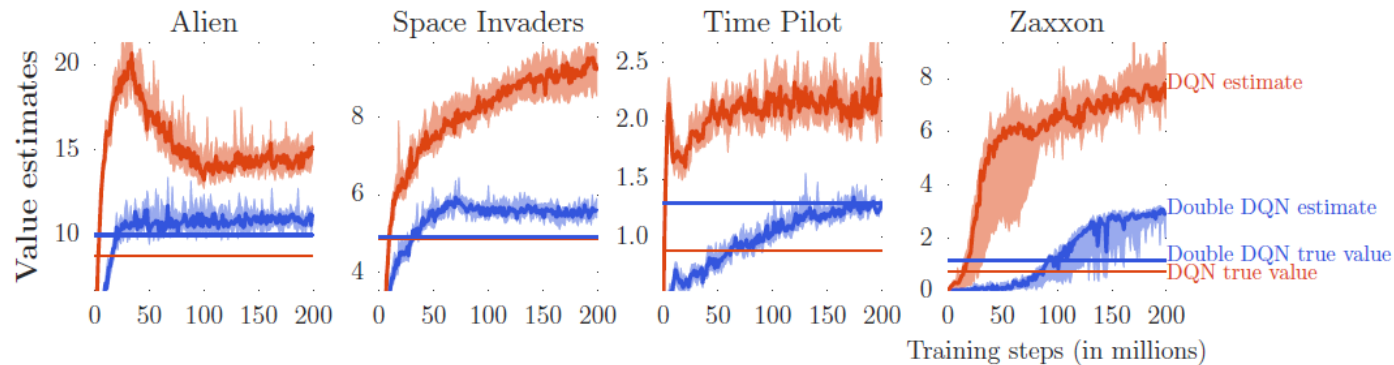
+0.02



+3.35

-0.02

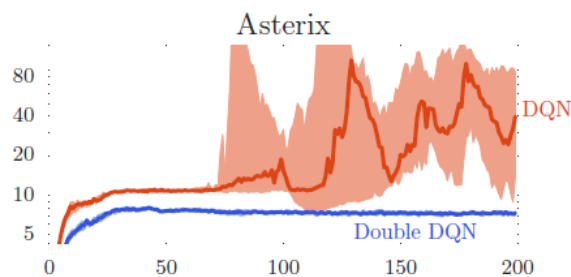
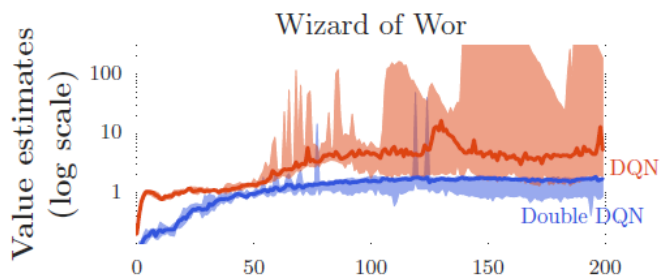
Result



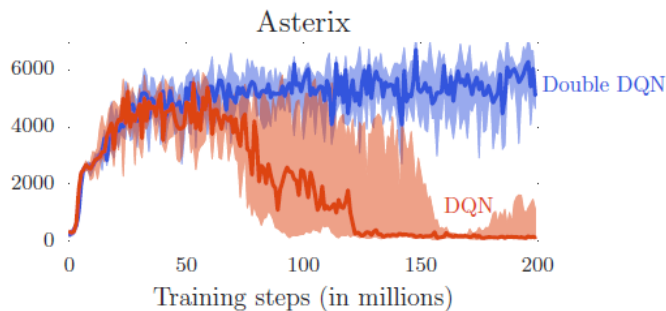
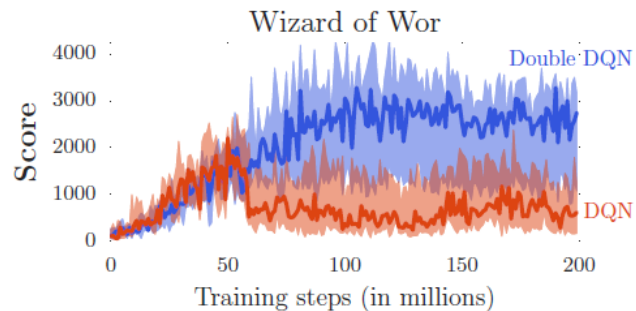
Orange - DQN / Blue - DDQN

Value estimates: DQN이 굉장히 overestimate하는 것을 볼 수 있다.

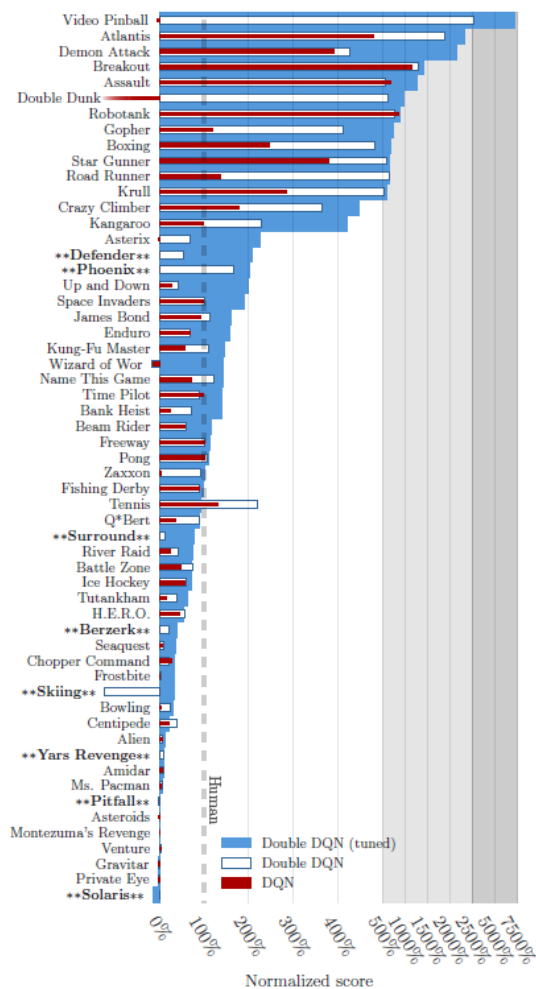
Score : Overestimation으로 시작 시 점수가 떨어지는 것을 확인할 수 있다.



True value와 가까운 estimation을 하는 DDQN이 더 정확하고 더 나은 policy를 찾는다는 것을 확인할 수 있다.



Result



결과적으로 DDQN은 deterministic한
순서보다는 일반화된 해결책을 찾는 방향으로
동작해서 더 robust함을 알 수 있었다.