To the Rainbow 제 3 편

Double DQN

정원석

순서

1. Abstract
2. 배경지식
3. Double DQN
4. 비교

1. Abstract

Q-learning

Q-learning

Overstimate 발생

▷ Unrealistic한 high value 학습 ▷ Performance 저하

DQN

Q-learning

+

DeepLearning

▷ DQN

DQN은 몇 게임에서 Human-level performance를 보였다

DQN

Overestimate 보안

Double Q-learning

+ Deeplearning ▷ Double DQN

2.배경지식

True value of an action

선택된 action ▷ Policy를 따라 action 선택

Expected sum of future reward

$$Q_\pi(s,a) \equiv \mathbb{E}\left[R_1 + \gamma R_2 + \ldots \,\middle|\, S_0 = s, A_0 = a, \pi\right]$$

Optimal action value

가장 높은 value를 받
게하는 Policy

$$Q_*(s,a) = \max_\pi Q_\pi(s,a)$$

Parameterized value function

모든 state에서의 action의 value
를 구하는것이 매우매우 어렵다.

$$Q(s, a; \boldsymbol{\theta}_t)$$

Parameterized value function을 배우자

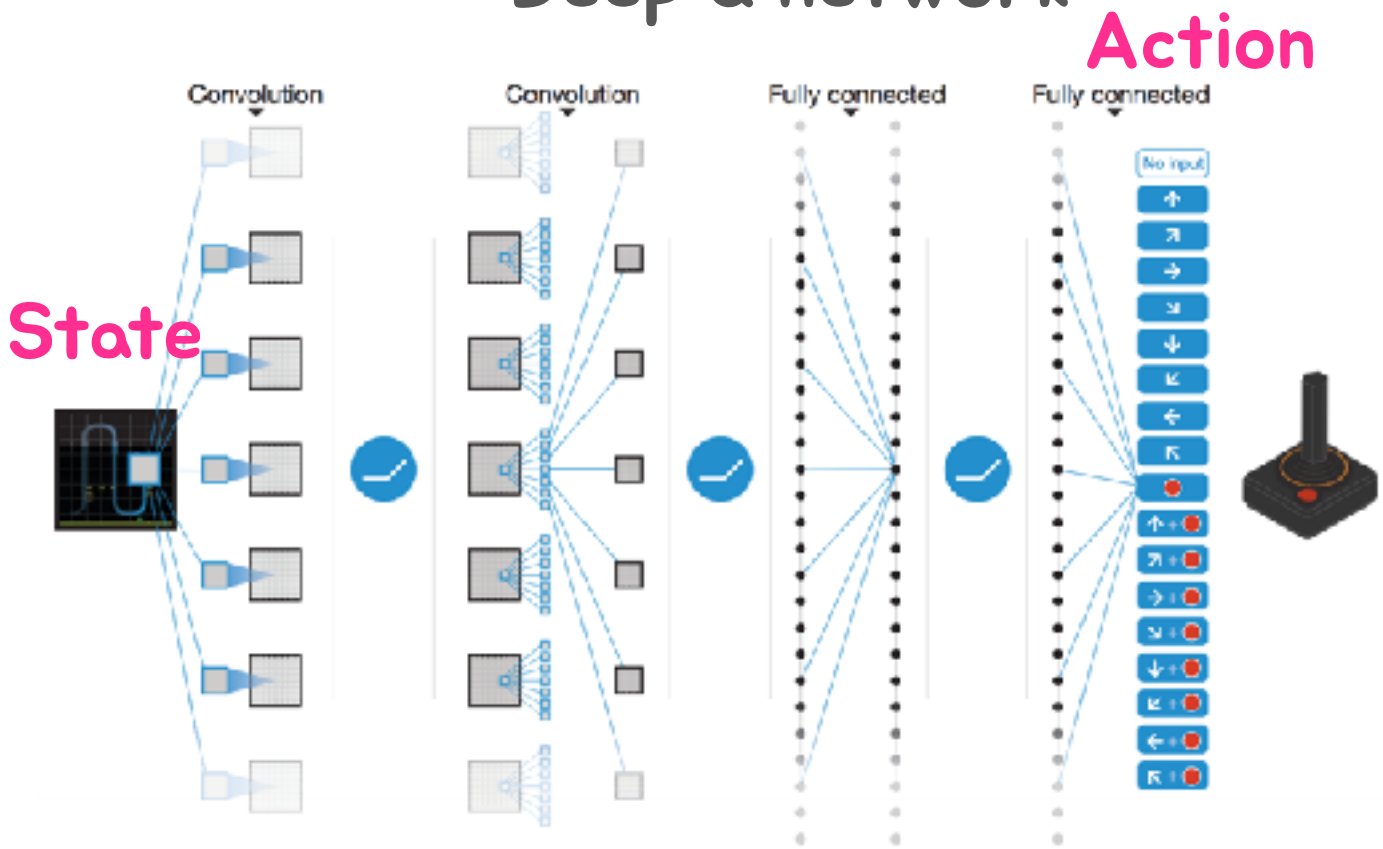
Q-learning UpdateRule

Target

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha(Y_t^{\mathrm{Q}} - Q(S_t, A_t; \boldsymbol{\theta}_t))\nabla_{\boldsymbol{\theta}_t} Q(S_t, A_t; \boldsymbol{\theta}_t)$$

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

Deep Q network



https://goo.gl/images/iQfbNG

Deep Q network -Target

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

Double Q-learning

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

Q-learning Target

$$Y_t^{\mathrm{Q}} = R_{t+1} + \gamma Q(S_{t+1}, \operatorname*{argmax}_a Q(S_{t+1}, a; \boldsymbol{\theta}_t); \boldsymbol{\theta}_t)$$

Double Q-learning Target

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

Symmetrically switching and update

Double DQN

분리!

Target의 action

$$Y_t^{\mathrm{Q}} = R_{t+1} + \gamma Q(S_{t+1}, \underset{a}{\operatorname{argmax}}\, Q(S_{t+1}, a; \boldsymbol{\theta}_t); \boldsymbol{\theta}_t)$$

Evaluation의 action

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

Double DQN

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

Target update 비교

Q-learning Target

$$Y_t^{\mathrm{Q}} = R_{t+1} + \gamma Q(S_{t+1}, \operatorname*{argmax}_a Q(S_{t+1}, a; \boldsymbol{\theta}_t); \boldsymbol{\theta}_t)$$

Double Q-learning Target

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

Symmetrically switching and update

DQN

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

Double DQN

$$Y_t^{\mathrm{DoubleDQN}} \equiv R_{t+1} + \gamma Q(S_{t+1}, \operatorname*{argmax}_a Q(S_{t+1}, a; \boldsymbol{\theta}_t), \boldsymbol{\theta}_t^-)$$