Deep Reinforcement Learning with Double Q-learning

Double DQN

IIP lab
정남규

Q-learning

$Q-value$: 상태($\boldsymbol{s_t}$)에서 행동($\boldsymbol{a_t}$)을 했을 때의 평균 $Return\ G_t = R_{t+1} + \gamma R_{t+2} + \cdots + R_{t+\infty}$

각각의 $Q-value$ 를 최대화하는 방향으로 학습하여 최종적인 optimal Q를 학습

Q-learning의 update

$$\boldsymbol{Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha(R_{t+1} + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}))}$$

Q-learning

Example)

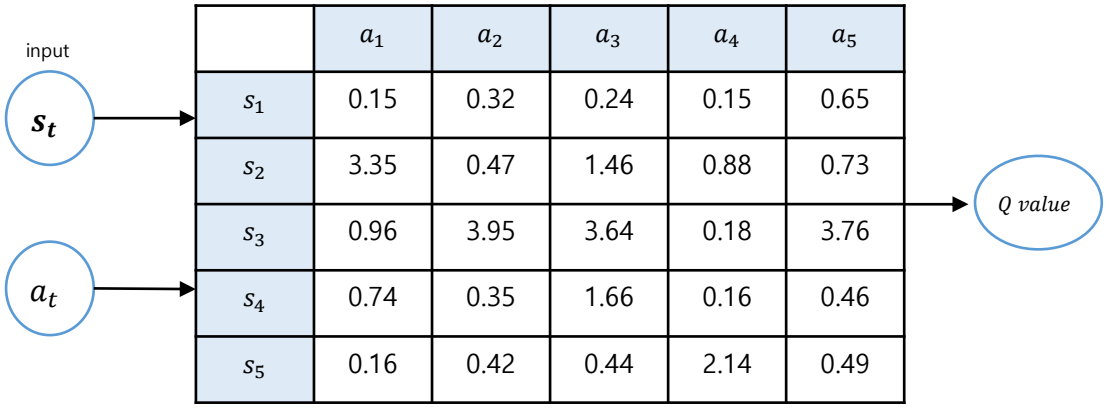
	a_1	a_2	a_3	a_4	a_5
s_1	0.15	0.32	0.24	0.15	0.65
s_2	3.35	0.47	1.46	0.88	0.73
s_3	0.96	3.95	3.64	0.18	3.76
s_4	0.74	0.35	1.66	0.16	0.46
s_5	0.16	0.42	0.44	2.14	0.49

단점 : n(s) 또는 n(a)가 많아질수록 많은 양의 공간을 필요로 한다.

DQN

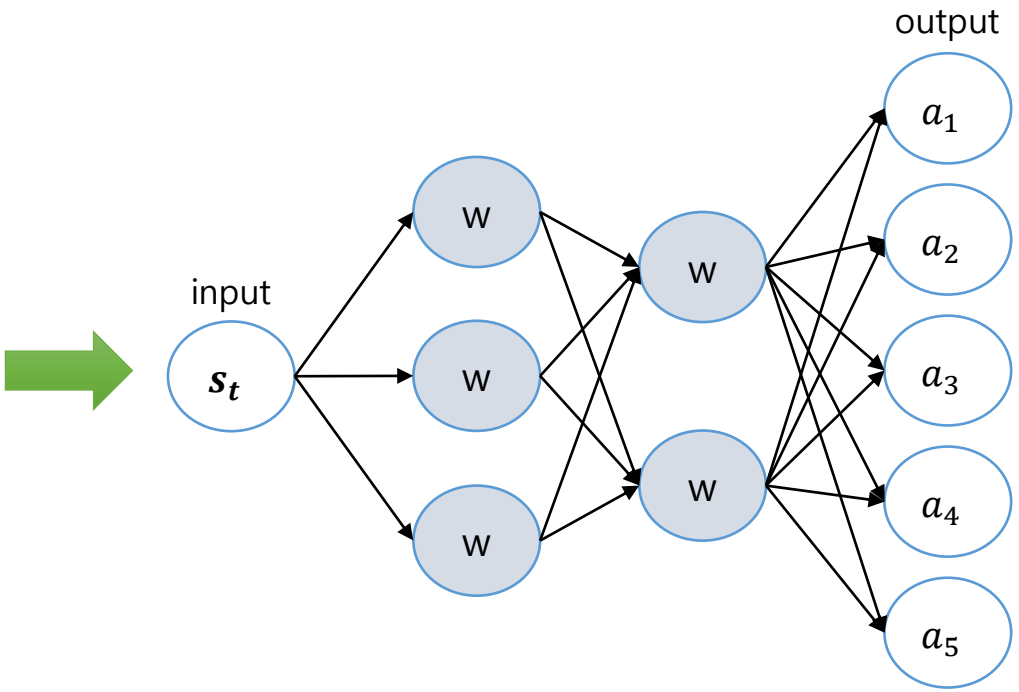
상태(s_t)에서의 모든 행동(a_t)에 대한 Q 값을 예측

Example)



	a_1	a_2	a_3	a_4	a_5
s_1	0.15	0.32	0.24	0.15	0.65
s_2	3.35	0.47	1.46	0.88	0.73
s_3	0.96	3.95	3.64	0.18	3.76
s_4	0.74	0.35	1.66	0.16	0.46
s_5	0.16	0.42	0.44	2.14	0.49

Q-learning



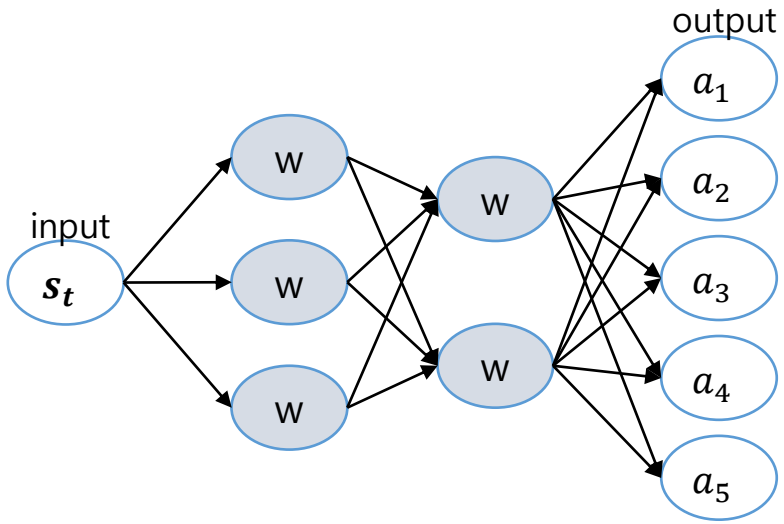
Deep Neural Network + Q-learning

DQN

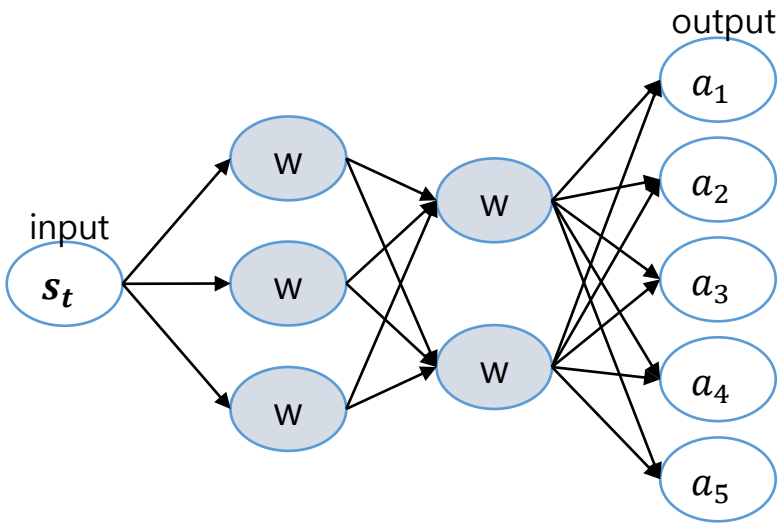
$$\boldsymbol{Q(s_t, a_t)} \leftarrow \boldsymbol{(1-\alpha)Q(s_t, a_t)} + \boldsymbol{\alpha(R_{t+1} + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}))}$$

Fixed Q_{target}

$$\boldsymbol{Q(s_t, a_t; w_{main})} \leftarrow \boldsymbol{(1-\alpha)Q(s_t, a_t; w_{main})} + \boldsymbol{\alpha(R_{t+1} + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; w_{target}))}$$



main

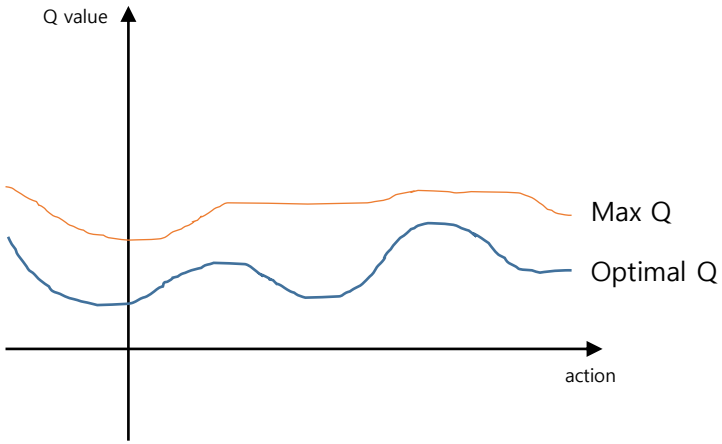


target

DQN

$$\boldsymbol{Q(s_t, a_t}; w_{main}) \leftarrow \boldsymbol{(1-\alpha)Q(s_t, a_t}; w_{main}) + \boldsymbol{\alpha(R_{t+1} + \gamma} \max_{a_{t+1}} \boldsymbol{Q\left(s_{t+1}, a_{t+1}}; w_{target}\right))$$

단점 : Q-learning의 overestimate action values



$$Max\ Q\ -Optimal\ Q\ \geq \sqrt{\frac{C}{m-1}} \qquad (C > 0, m: action\text{의 총 개수})$$

Let $\boldsymbol{\varepsilon_a = Q(s_t, a_t) - \ Q^*(s_t, a_t)}$

Goal

$$Max\ Q\ - Optimal\ Q\ \geq \sqrt{\frac{C}{m-1}} \qquad (C > 0, m: n(action))$$

Condition

1. s_t가 주어졌을 때 모든 action에 대해 $Q^*(s_t, a_t)$의 값이 같다.
2. $\sum_a \varepsilon_a = 0$
3. $\frac{1}{m}\sum_a {\varepsilon_a}^2 = C \qquad (C > 0, m: action$의 총 개수$)$

Condition

2. $\sum_a \varepsilon_a = 0$

3. $\frac{1}{m}\sum_a {\varepsilon_a}^2 = C$ $\quad (C > 0, m: action$의 총 개수$)$

Let $\quad n: {\varepsilon_a}^+(\varepsilon_a > 0)$의 $action$의 개수, $\quad m - n: {\varepsilon_a}^-(\varepsilon_a \leq 0)$의 $action$의 개수

Case 1) 모든 $\varepsilon_a = 0$이면 Condition 2는 성립하지만 Condition 3 불만족이라 성립이 되지 않는다.

Case 2) $\mathrm{m} - \mathrm{n} \geq 1,\ n \geq 1$이면 Condition 2가 성립하며 Condition 3가 성립할 수 있다.

Case 2) $\mathrm{m} - \mathrm{n} \geq 1,\ n \geq 1$

$$\therefore \sum_{a} {\varepsilon_a}^{+} \leq n \times \max {\varepsilon_a}^{+}$$

Assume : $\max \varepsilon_a < \sqrt{\frac{C}{m-1}}$

논리적 오류를 통한 증명

Assume : $\max \varepsilon_a < \sqrt{\frac{C}{m-1}}$ 이 모순임을 보여, Proof : $\max \varepsilon_a \geq \sqrt{\frac{C}{m-1}}$ 을 증명

Case 2) $\mathrm{m} - \mathrm{n} \geq 1,\ n \geq 1$

$$\therefore \sum_a {\varepsilon_a}^+ \leq n \times \max {\varepsilon_a}^+$$

Assume : $\max \varepsilon_a < \sqrt{\frac{C}{m-1}}$

1. $\sum_a |{\varepsilon_a}^-| \leq n \times \max |{\varepsilon_a}^-| < n \times \sqrt{\frac{C}{m-1}}$

2. $\sum_a {\varepsilon_a}^+ \leq n \times \max {\varepsilon_a}^+ < n \times \sqrt{\frac{C}{m-1}}$

1. $\sum_a |\varepsilon_a{}^-| \le n \times \max|\varepsilon_a{}^-| < n \times \sqrt{\frac{C}{m-1}}$

$$\Rightarrow \sum_a |\varepsilon_a{}^-| \times \sum_a |\varepsilon_a{}^-| \le \sum_a |\varepsilon_a{}^-| \times \max|\varepsilon_a{}^-| < \left(n \times \max|\varepsilon_a{}^-|\right)^2 < \left(n \times \sqrt{\frac{C}{m-1}}\right)^2$$

$$\Rightarrow \sum_a |\varepsilon_a{}^-|^2 \le \sum_a |\varepsilon_a{}^-| \times \max|\varepsilon_a{}^-| < \frac{n^2 \times C}{m-1}$$

$$\Rightarrow \sum_a |\varepsilon_a{}^-|^2 \le \sum_a |\varepsilon_a{}^-| \times \max|\varepsilon_a{}^-| < \frac{n^2 \times C}{m-1} < (m-1) \times C$$

$\because$ $m - n \ge 1 \rightarrow m - 1 \ge n$

2. $\sum_a {\varepsilon_a}^+ \le n \times \max {\varepsilon_a}^+ < n \times \sqrt{\frac{C}{m-1}}$

$$\Rightarrow \sum_a {\varepsilon_a}^+ \times \sum_a {\varepsilon_a}^+ \le \sum_a {\varepsilon_a}^+ \times \max {\varepsilon_a}^+ < n \times \left(\max {\varepsilon_a}^+\right)^2 < n \times \left(\sqrt{\frac{C}{m-1}}\right)^2$$

$$\Rightarrow \sum_a ({\varepsilon_a}^+)^2 \le n \times \frac{C}{m-1}$$

1. $\sum_a |\varepsilon_a^-|^2 \leq \sum_a |\varepsilon_a^-| \times \max|\varepsilon_a^-| < \frac{n^2 \times C}{m-1} < (m-1) \times C$

2. $\sum_a (\varepsilon_a^+)^2 \leq n \times \frac{C}{m-1}$

$\Rightarrow \sum_a |\varepsilon_a^-|^2 + \sum_a (\varepsilon_a^+)^2 \leq \sum_a |\varepsilon_a^-| \times \max|\varepsilon_a^-| + \sum_a (\varepsilon_a^+)^2 < \frac{n^2 \times C}{m-1} + \frac{n \times C}{m-1}$

$\Rightarrow \sum_a (\varepsilon_a)^2 \leq \sum_a |\varepsilon_a^-| \times \max|\varepsilon_a^-| + \sum_a (\varepsilon_a^+)^2 < \frac{n^2 \times C}{m-1} + \frac{n \times C}{m-1} < (m-1) \times C + C$ $\because$ $m - 1 \geq n$

$\Rightarrow \sum_a (\varepsilon_a)^2 < m \times C$

Assume : $\max \varepsilon_a < \sqrt{\frac{C}{m-1}}$

최종 식 : $\sum_a (\varepsilon_a)^2 \quad < \quad m \times C$ (Condition 3. $\frac{1}{m}\sum_a {\varepsilon_a}^2 = C$ 에 모순)

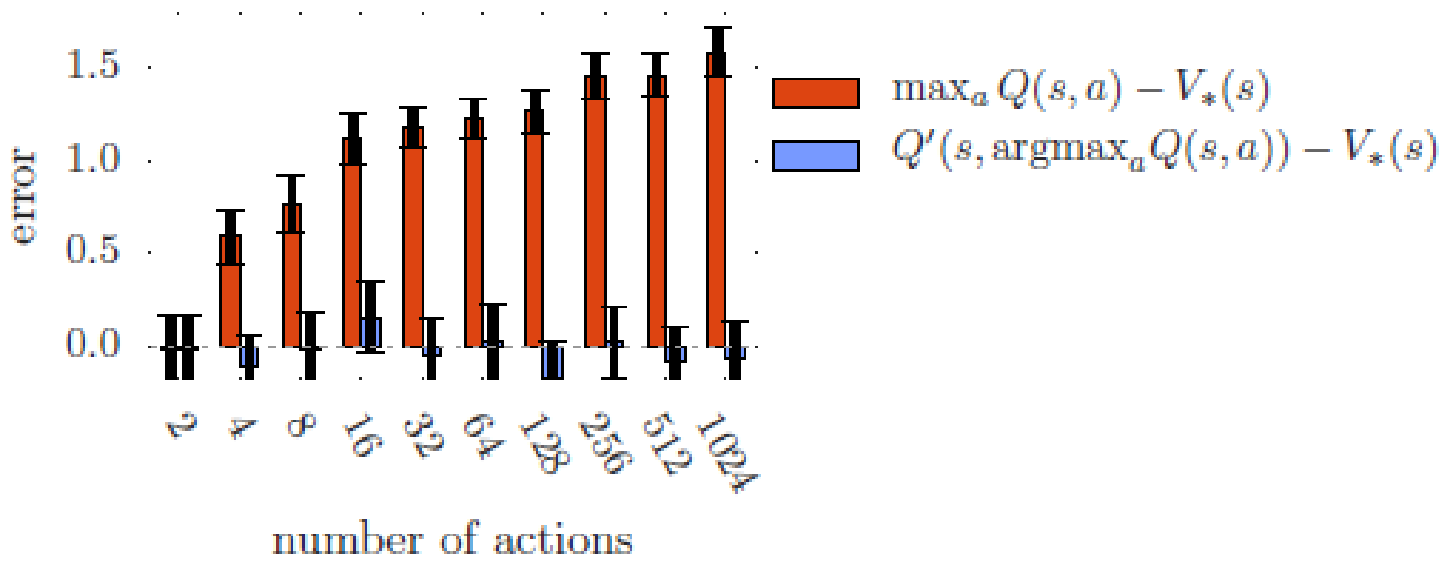
Conclusion

Assume : $\max \varepsilon_a < \sqrt{\frac{C}{m-1}}$ 의 역인 Proof : $\max \varepsilon_a \geq \sqrt{\frac{C}{m-1}}$ 가 참이다.

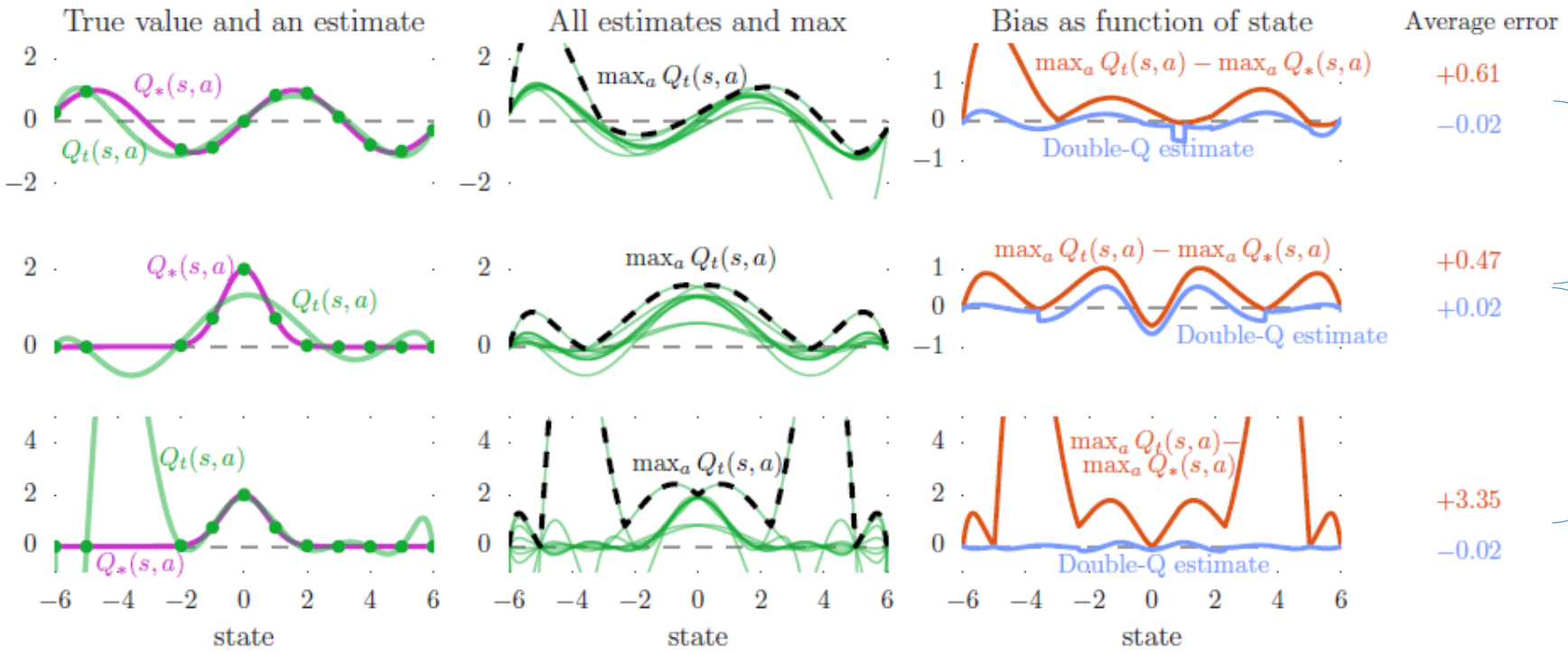
Double DQN

$$Q(s_t, a_t; w_{main}) \leftarrow (1-\alpha)Q(s_t, a_t; w_{main}) + \alpha(R_{t+1} + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; w_{target}))$$

$$Q(s_t, a_t; w_{main}) \leftarrow (1-\alpha)Q(s_t, a_t; w_{main}) + \alpha(R_{t+1} + \gamma Q(s_{t+1}, \underset{a_{t+1}}{\text{argmax}} Q(s_{t+1}, a_{t+1}; w_{main}); w_{target}))$$



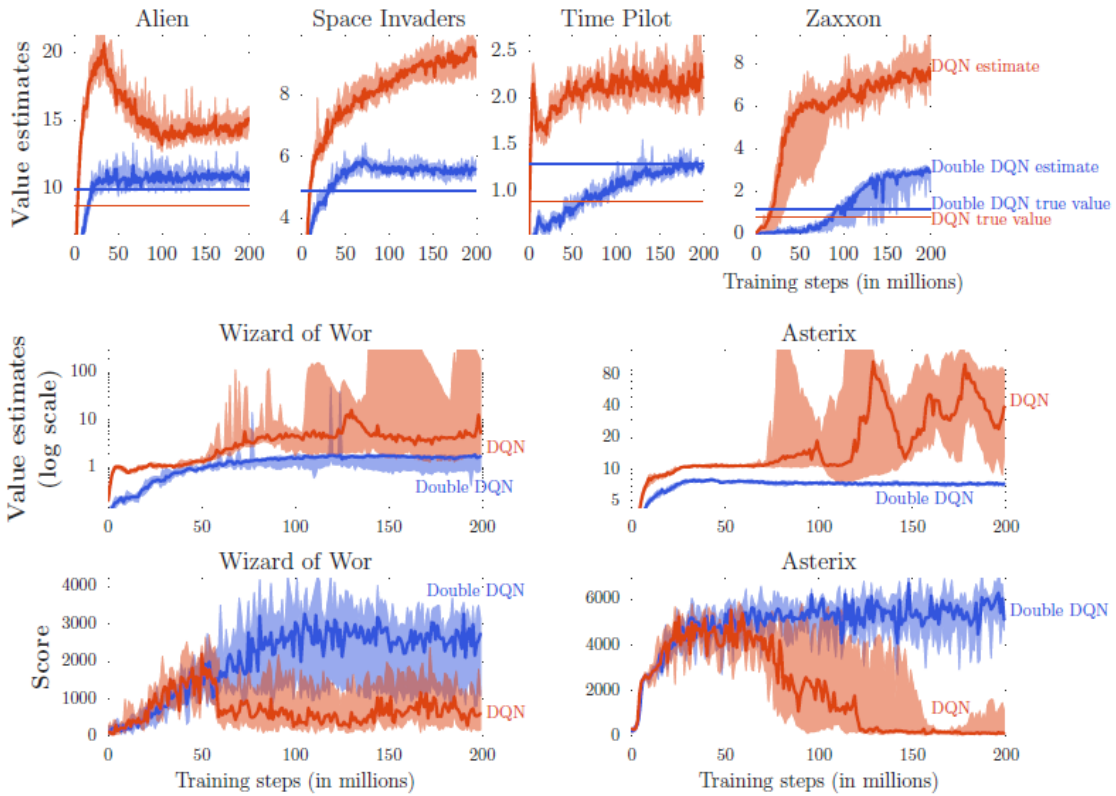
Double DQN



structure 비교

flexibility 비교

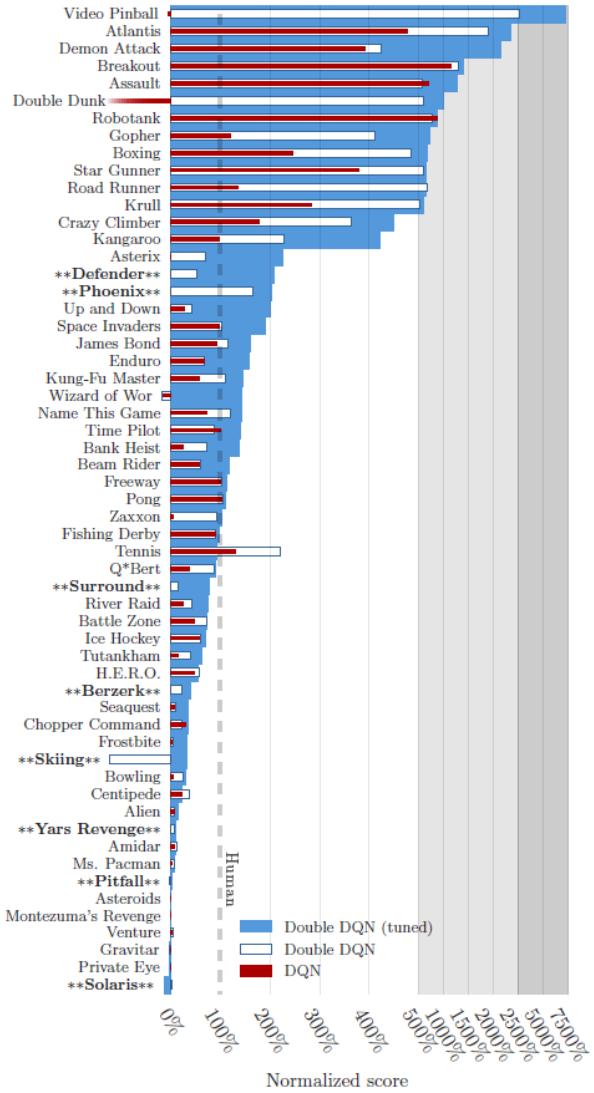
Double DQN



$$\frac{1}{T}\sum_{t=1}^{T}\max_{a} Q(s_t, a; w)\ , \qquad T = 125{,}000$$

Training step마다 $\max_{a} Q(s_t, a; w)$ 의 평균 값 비교

Double DQN



	DQN	Double DQN
Median	93.5%	114.7%
Mean	241.1%	330.3%

5분 학습 후 DQN(2015)와 성능 비교

	DQN	Double DQN	Double DQN (tuned)
Median	47.5%	88.4%	116.7%
Mean	122.0%	273.1%	475.2%

30분 학습 후 Human과 성능 비교

Double DQN

Five contributions

1. Q-learning의 내재인 추정 오류 때문에 overestimate가 될 수 있다는 것
2. Atari 2600 분석을 통해 overestimate가 실제로 더 흔하고 심각하다는 것
3. DQN에서 추가적인 네트워크나 매개 변수 없이 Double Q-learning을 적용
4. Double Q-learning을 적용함으로써 더 안정적인 학습이 가능하다는 것
5. Double DQN이 더 좋은 policy를 찾으며, Atari 2600에서 State-Of-The-Art를 보여주었다는 것

Thank you

Double DQN