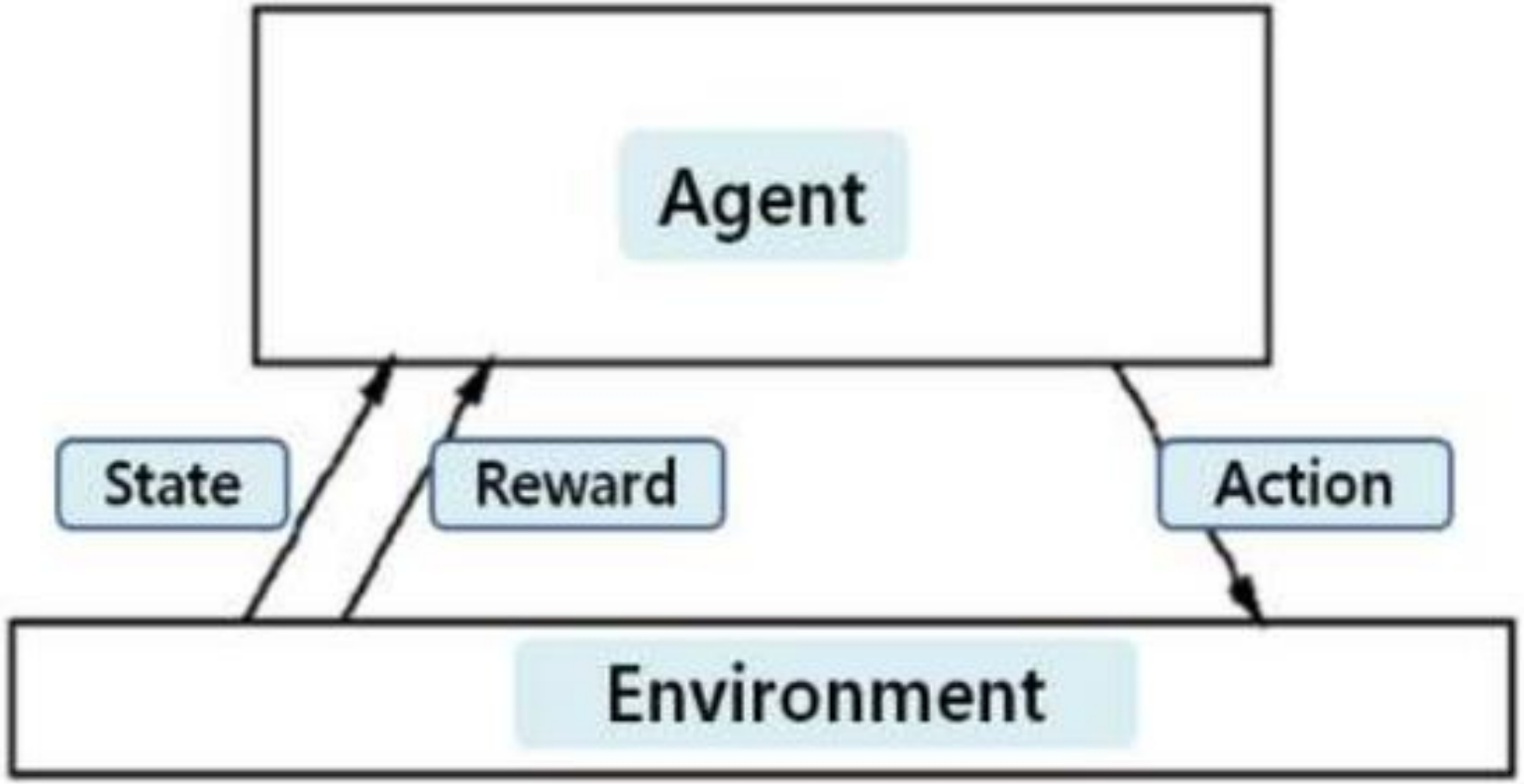
Dueling Network Architectures for Deep Reinforcement Learning

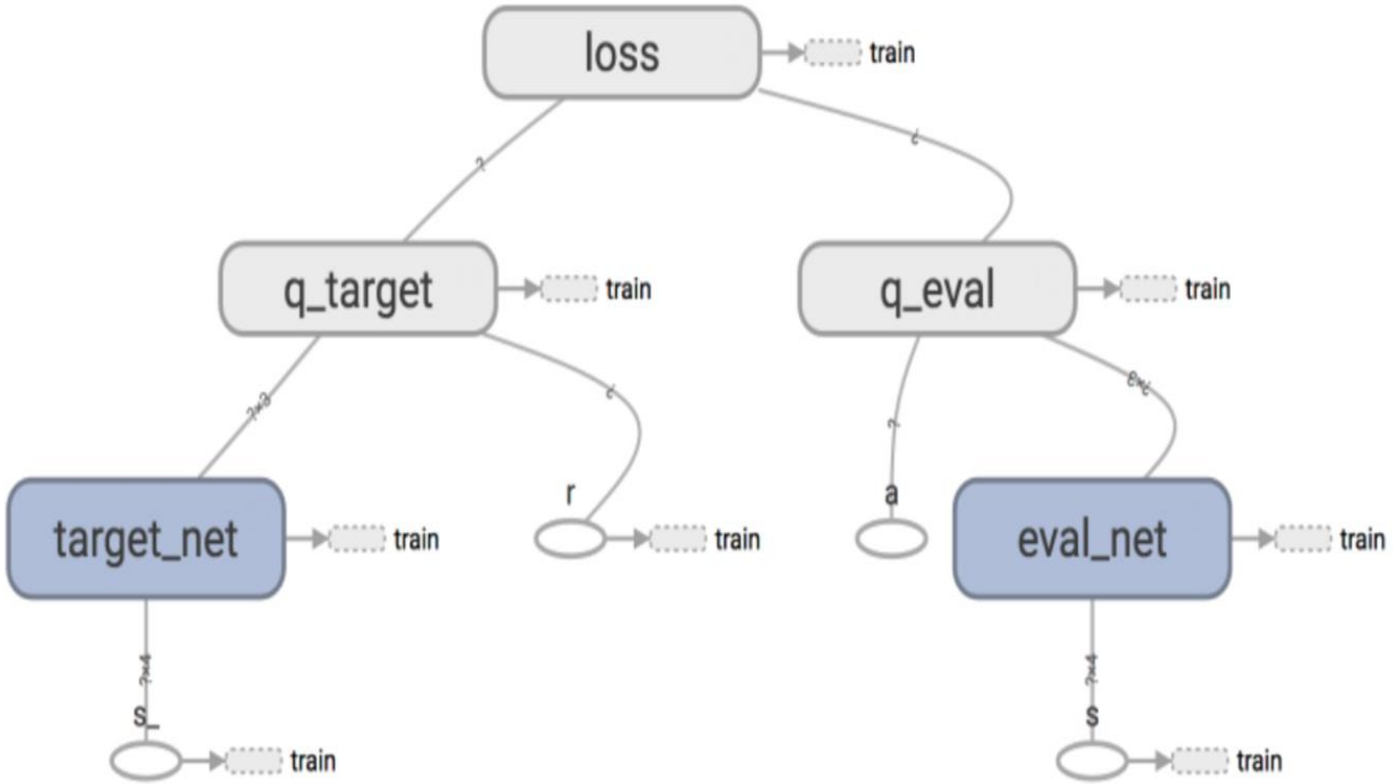
목차

- 강화학습의 기본
- Q-learning -> Deep Q learning
- Abstract, Introduction
- ATARI 게임
- 함수 수식
- Experiments
- Conclusion

강화학습의 기본

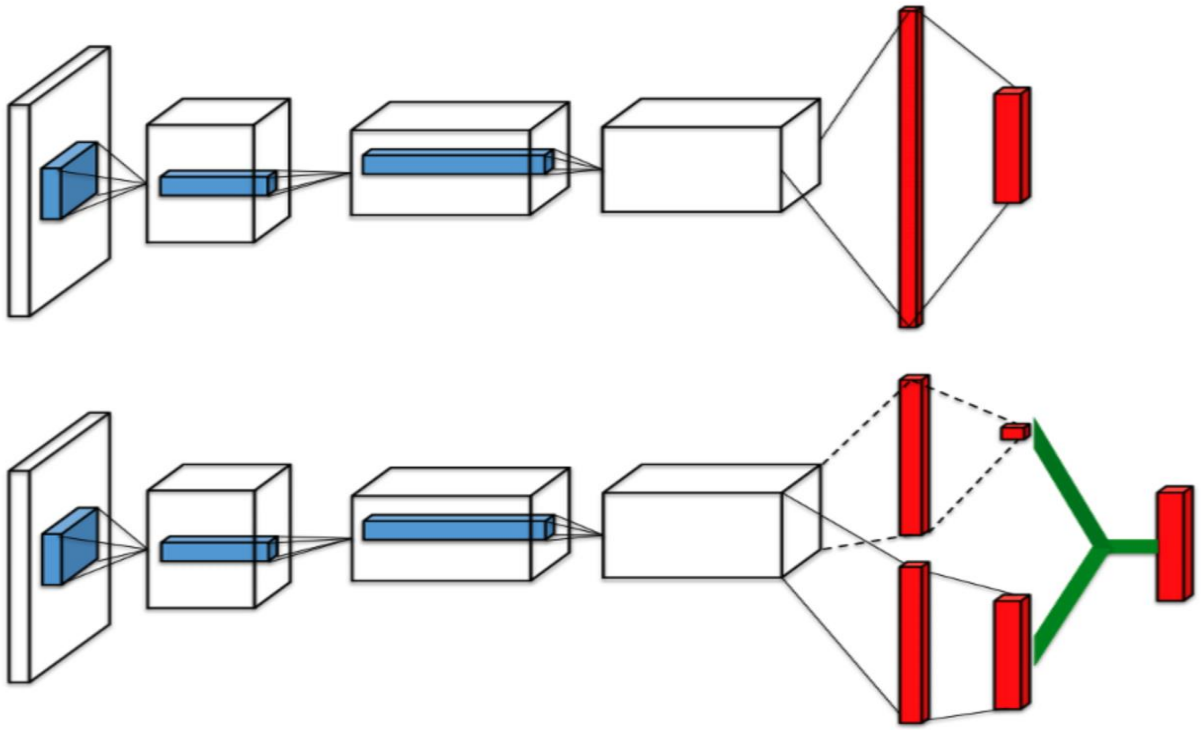


Q-learning -> Deep Q learning



- 다층 합성곱 신경망(CNN)
- 네트워크가 기억 속 사전 지식을 통해 스스로 학습할 수 있도록 경험 리플레이를 사용
- 갱신 중 목표 Q-값을 계산하기 위해 타겟 네트워크를 사용합니다.

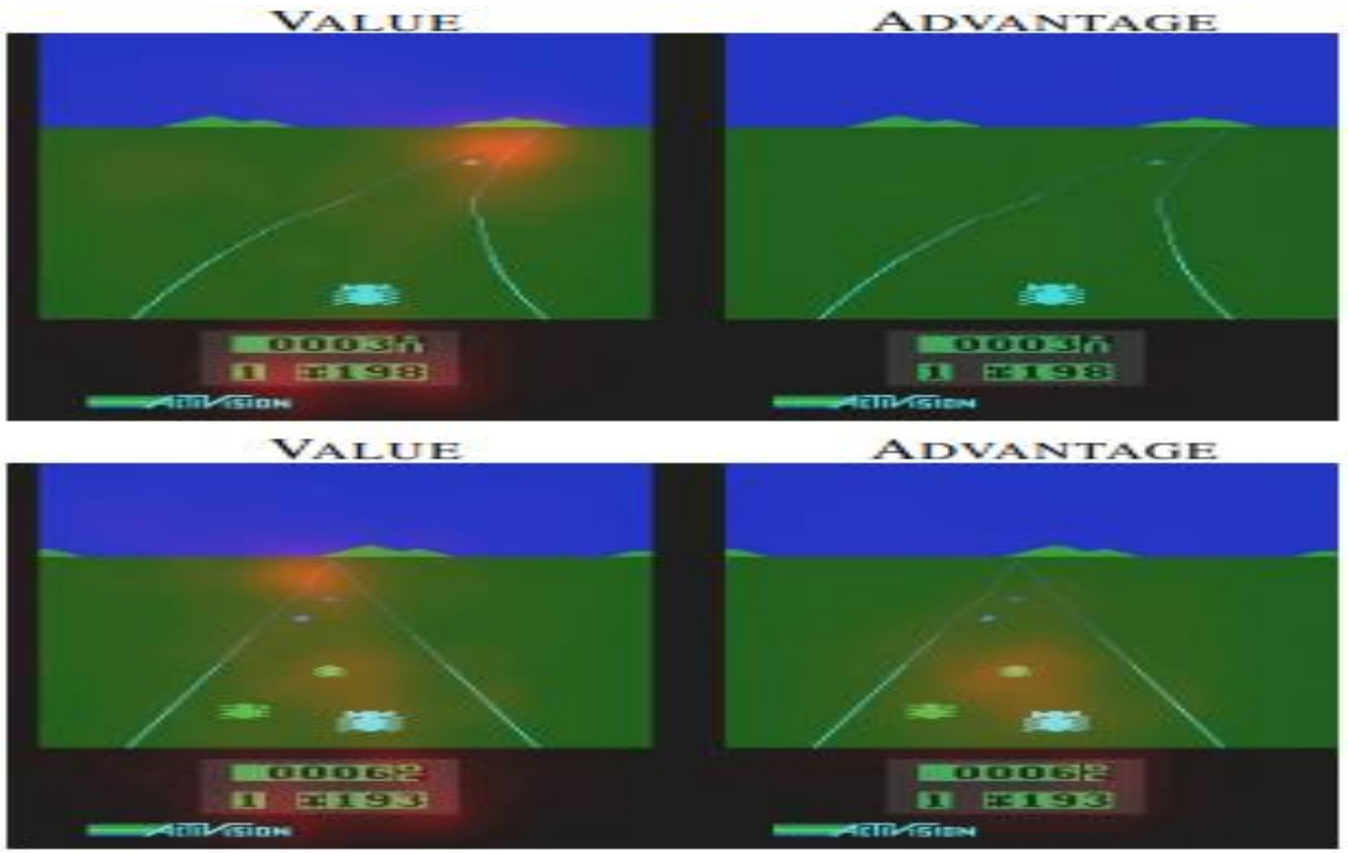
Abstract, Introduction



$$Q(s,a) = A(s,a) + V(s)$$

- 위 그림은 DQN,
  아래 그림은 DDQN 입니다.
- 가치함수 V(s) -> 특정 상태에
  있는 것이 얼마나 좋은지
  나타내는 값입니다.
- 어드밴티지 함수 A(s, a) :
  어드밴티지 함수는 특정 행동이
  다른 행동에 비해 얼마나 좋은 지
  알려줍니다.

ATARI 게임



논문의 최종 수식

$$Q(s,a;\theta,\alpha,\beta) = V(s;\theta,\beta) + \left(A(s,a;\theta,\alpha) - \underbrace{\frac{1}{\mathcal{A}}\sum_{a'} A(s,a';\theta,\alpha)}_{\text{Average advantage}}\right)$$

Common network parameters

Value stream parameters

Advantage stream parameters

Hence, all the experiments reported in this paper use the module of equation (논문 인용구)

정책 평가(Policy evaluation) - Experiments

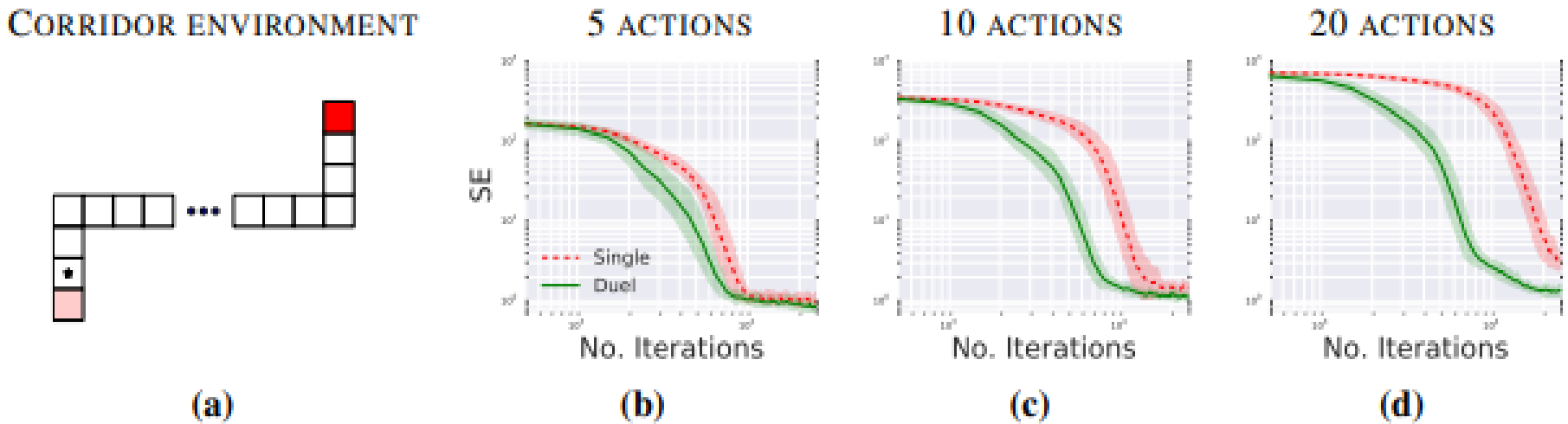
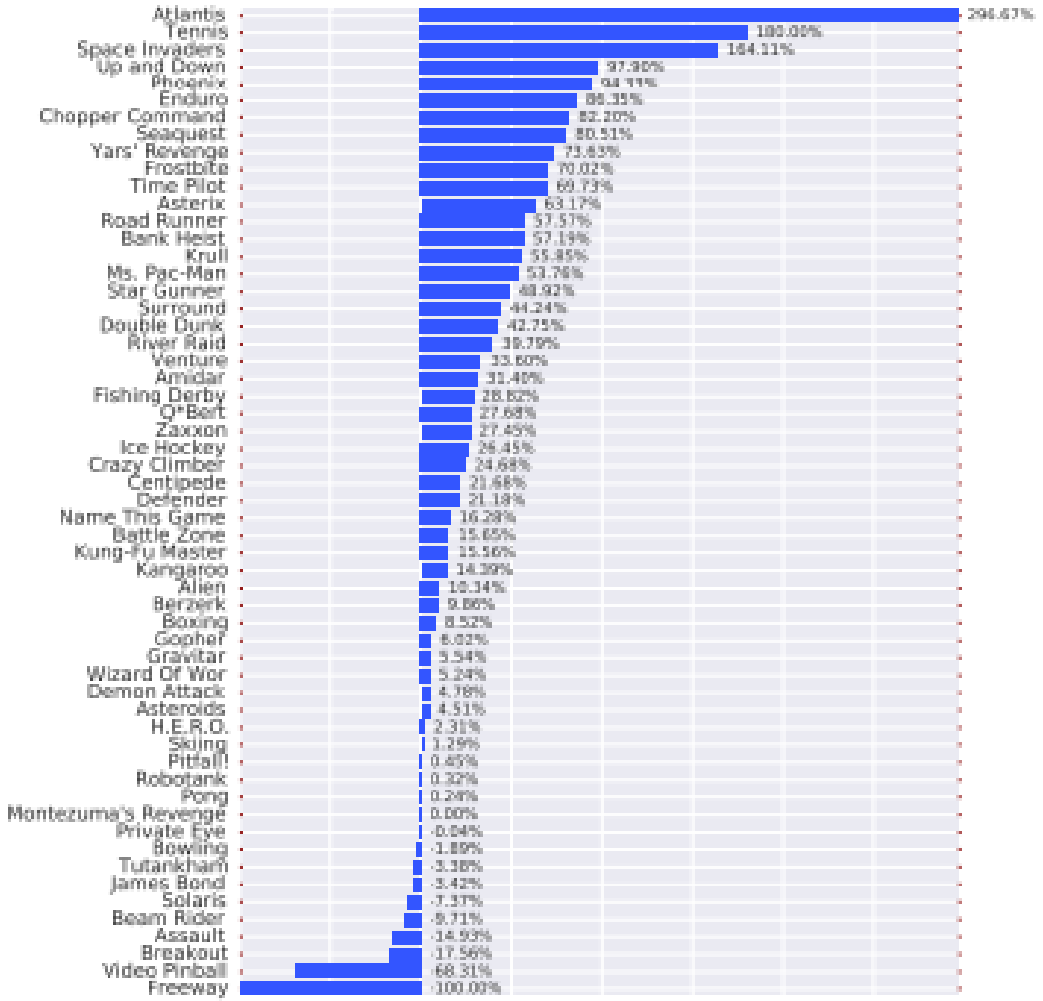


Figure 3. **(a)** The corridor environment. The star marks the starting state. The redness of a state signifies the reward the agent receives upon arrival. The game terminates upon reaching either reward state. The agent's actions are going up, down, left, right and no action. Plots **(b)**, **(c)** and **(d)** shows squared error for policy evaluation with 5, 10, and 20 actions on a log-log scale. The dueling network (Duel) consistently outperforms a conventional single-stream network (Single), with the performance gap increasing with the number of actions.

아타리 게임 플레이 결과



$$\frac{\text{Score}_{\text{Agent}} - \text{Score}_{\text{Baseline}}}{\max\{\text{Score}_{\text{Human}}, \text{Score}_{\text{Baseline}}\} - \text{Score}_{\text{Random}}}. \qquad (10)$$

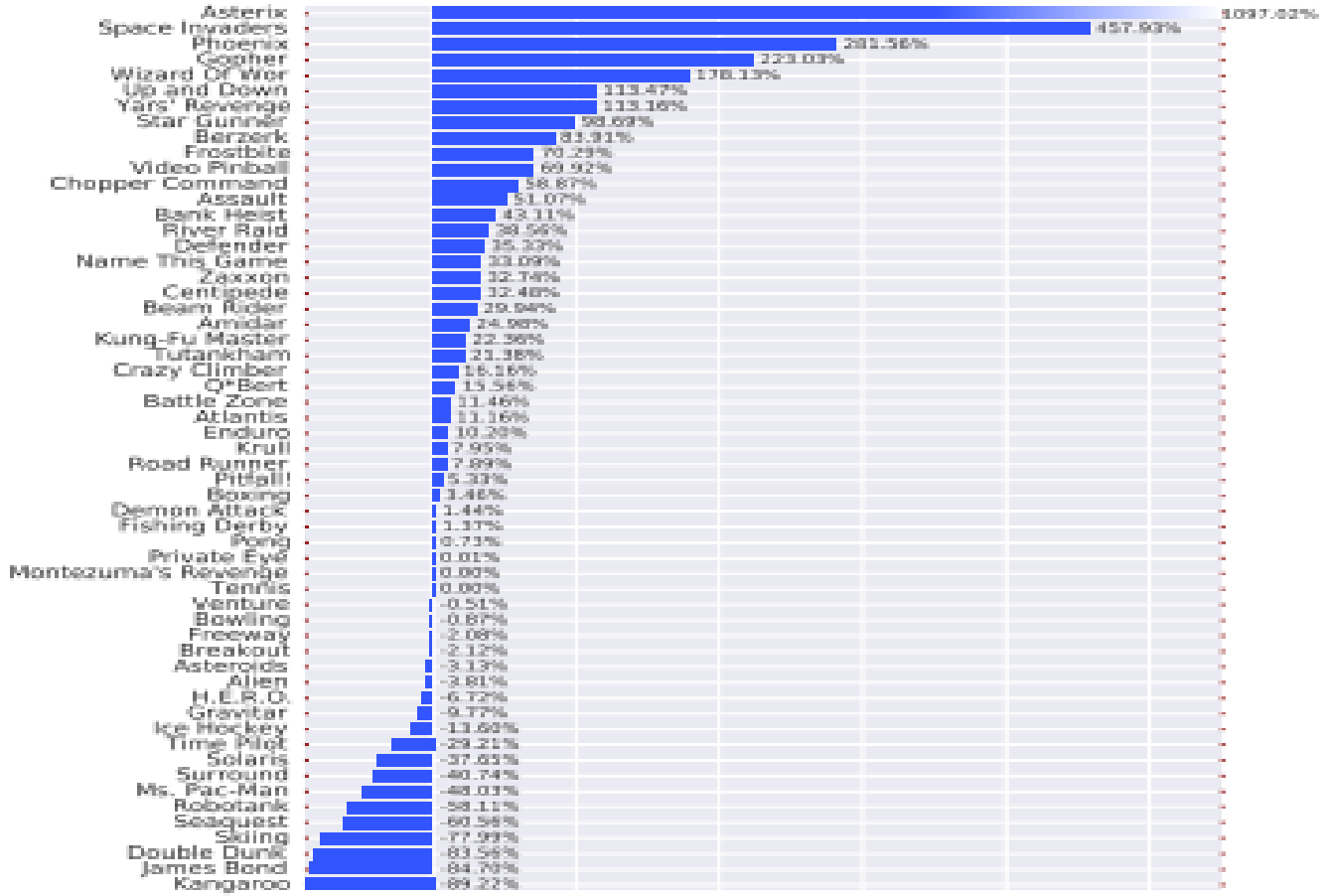
Table 1. Mean and median scores across all 57 Atari games, measured in percentages of human performance.

	30 no-ops		Human Starts	
	Mean	**Median**	**Mean**	**Median**
Prior. Duel Clip	**591.9%**	**172.1%**	**567.0%**	**115.3%**
Prior. Single	434.6%	123.7%	386.7%	112.9%
Duel Clip	**373.1%**	**151.5%**	**343.8%**	**117.1%**
Single Clip	341.2%	132.6%	302.8%	114.1%
Single	307.3%	117.8%	332.9%	110.9%
Nature DQN	227.9%	79.1%	219.6%	68.5%

Prioritization, dueling and gradient clipping 상호작용

Note that, although orthogonal in their objectives, these extensions (prioritization, dueling and gradient clipping) interact in subtle ways. For example, prioritization interacts with gradient clipping, as sampling transitions with high absolute TD-errors more often leads to gradients with higher norms. To avoid adverse interactions, we roughly re-tuned the learning rate and the gradient clipping norm on a subset of 9 games. As a result of rough tuning, we settled on 6.25×10^{-5} for the learning rate and 10 for the gradient clipping norm (the same as in the previous section).

아타리 게임 플레이 결과


Asterix
Space Invaders
Phoenix
Gopher
Wizard Of Wor
Up and Down
Yars' Revenge
Star Gunner
Berzerk
Frostbite
Video Pinball
Chopper Command
Assault
Bank Heist
River Raid
Defender
Name This Game
Zaxxon
Centipede
Beam Rider
Amidar
Kung-Fu Master
Tutankham
Crazy Climber
Q*Bert
Battle Zone
Atlantis
Enduro
Krull
Road Runner
Pitfall!
Boxing
Demon Attack
Fishing Derby
Pong
Private Eye
Montezuma's Revenge
Tennis
Venture
Bowling
Freeway
Breakout
Asteroids
Alien
H.E.R.O.
Gravitar
Ice Hockey
Time Pilot
Solaris
Surround
Ms. Pac-Man
Robotank
Seaquest
Skiing
Double Dunk
James Bond
Kangaroo

감사합니다.