Dueling Network Architectures for Deep Reinforcement Learning

By
Saptarashmi Bandyopadhyay
First Year PhD Student
Department of Computer Science and Engineering
Pennsylvania State University, University Park

Outline

- Motivation of the Paper
- Dueling Network Architecture
- Improvements Using the New Architecture
- Future Work

Introduction

- A conference paper 'Dueling Network Architectures for Deep Reinforcement Learning' at ICML 2016
- Authors:
  - Dr. Ziyu Wang, Dr. Tom Schaul, Dr. Matteo Hessel, Dr. Hado van Hasselt, Dr. Marc Lanctot, and Dr. Nando de Freitas, Research Scientists, Google DeepMind

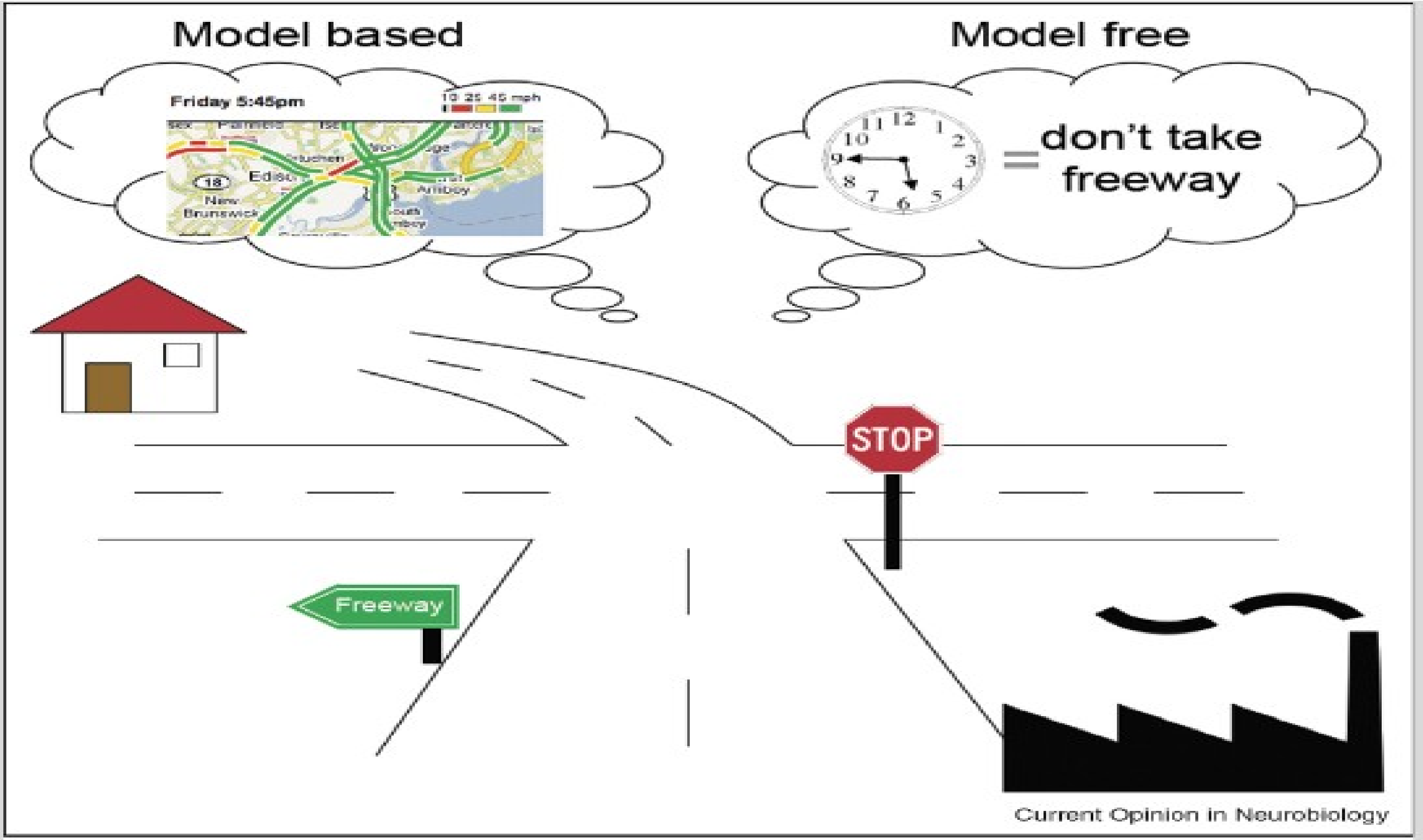
For reference

- Please read the following reference paper in addition to this paper for detailed information on convergence of advantage updating
- 'Residual Advantage Learning Applied to a Differential Game' in Neural Information Processing Systems, May 1995

Benefits of the New Architecture

- Better generalization can be achieved in the learning process across actions by modifying the conventional neural architecture (e.g. LSTMs or auto-encoders) without changing the algorithms
- New neural architecture for model-free reinforcement learning (RL)
- Better policy evaluation in the event of many similar-valued actions
- Better performance while playing the Atari 2600 games

Model-free RL vs Model-based RL



Current Opinion in Neurobiology

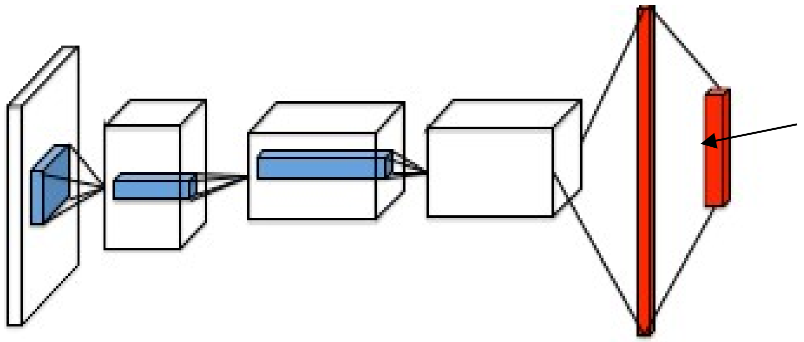
Figure 1: Two ways to choose which route to take when traveling home from work on friday evening.

[2]

Overview of the Dueling Network Architecture

- Two separate estimators
  - One for state value function
  - Another for state-dependent action advantage function
- Benefit that the new Network Architecture can be applied to existing algorithms in RL and can be extended to future algorithms

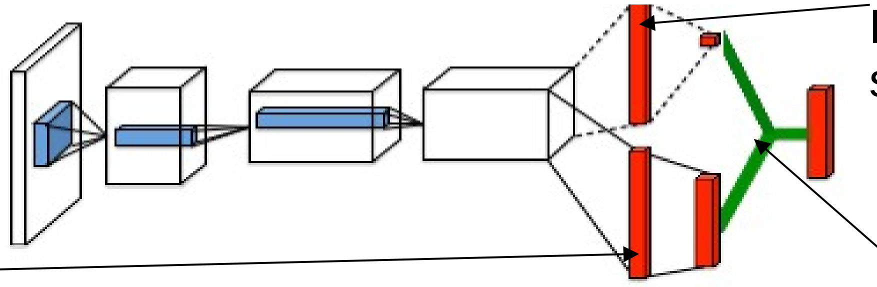
Diagrammatic Representation



Q function

[1]

Single Stream Q Network (Conventional Architecture)



Estimator for scalar state-value (V)

[1]

Estimator of advantages for each action (A)

Q function combining the V and A function

Dueling Q Network with 2 streams (New Architecture)

Aspects of the Dueling Architecture

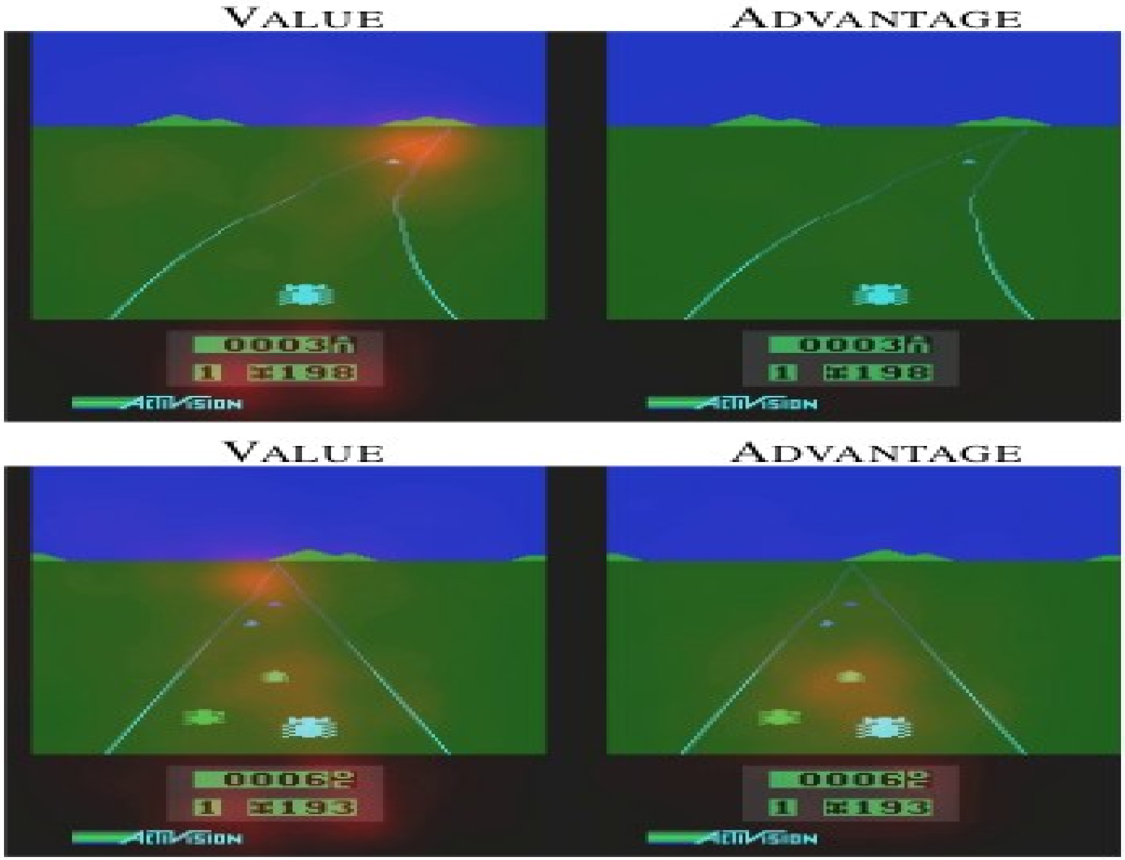
- The representation of the state-values and the action-advantages are separated in the two streams
- The separate value and advantage functions share a common convolutional feature learning module
  - Separate estimates are provided without additional supervision
- They are combined by a special aggregating layer to produce the Q function

Intuition

- Valuable states can be learned with this architecture without learning the effect of each action for each stage
- Very pertinent when the actions do not impact the environment signficantly
- Experiments have illustrated quick identification of correct actions during the policy evaluation

Illustration of the intuition on saliency maps in Atari 2600

Value function pays attention on the road



Advantage function pays attention to stop collision

The Atari Game-Enduro

Environment of the Experiments

- Sequential decision making set-up
- Interaction with the environment over discrete time steps
- Goal is to maximize the discounted reward R_t
- $R_t = \Sigma_{i=t}{}^{\infty} \gamma^{(i-t)} r_i$
- $\gamma \in [0, 1]$ is the discount factor between immediate and future reward

Q and V Functions for a Stochastic Policy π

$$Q^{\pi}(s,a) \quad = \quad \mathbb{E}\left[R_t \middle| s_t = s, a_t = a, \pi\right]$$

$$Q^{\pi}(s,a) = \mathbb{E}_{s'}\left[r + \gamma \mathbb{E}_{a' \sim \pi(s')}\left[Q^{\pi}(s',a')\right] \mid s, a, \pi\right]$$

The Q function measures the value of choosing a particular action a at a certain state s

$$V^{\pi}(s) \quad = \quad \mathbb{E}_{a \sim \pi(s)}\left[Q^{\pi}(s,a)\right].$$

The Value Function determines the goodness of the state s

Advantage Function

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s)$$

The advantage function gives a relative measure of the importance of each action at a certain state.

$$\mathbb{E}_{a \sim \pi(s)}\left[A^{\pi}(s,a)\right] = 0.$$

Review of Deep Q Networks and Experience Replay

$$L_i(\theta_i) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{U}(\mathcal{D})}\left[\left(y_i^{DQN} - Q(s,a;\theta_i)\right)^2\right].$$

$$y_i^{DQN} = r + \gamma \max_{a'} Q(s', a'; \theta^-)$$

- Experience replay improves the data efficiency through re-use of the samples with multiple updates
- Variance is reduced due to uniform sampling of the minibatches of experience with the replay buffer. Correlation decreases among the samples in the update.

Double Deep Q Network

$$y_i^{DDQN} = r + \gamma Q(s', \arg\max_{a'} Q(s', a'; \theta_i); \theta^-).$$

y_i^{DQN} is replaced by y_i^{DDQN}.

It is useful to address over-optimism in the value estimates.

Prioritized Replay

- For those experience tuples that have a high expected learning progress, their replay probability is increased.
- Faster learning
- Better final policy quality

Q and V Functions for a Deterministic Policy

- $a* = \arg\max_{a' \in A} Q(s, a')$
- $Q(s, a*) = V(s)$
- $A(s, a*) = 0$

Output of the Two Streams

- For the 2 streams in the dueling architecture
  - One of them generates V (s; θ, β) (scalar)
  - Another stream generates A(s, a; θ, α), an |A|-dimensional vector
  - Θ refers to the parameters of the convolutional layers
  - α and β refers to the parameters of the 2 streams of fully connected layers
- Aggregation of the V and A functions to generate the Q function by addition is problematic due to the non-unique recovery of V and A functions.

Solution to the Aggregation Problem

- 0 advantage is generated by the advantage function estimator at the chosen action.
- Forward mapping is implemented in the last module of the network.
- $Q(s, a; \theta, \alpha, \beta) = V(s; \theta, \beta) + A(s, a; \theta, \alpha) - \max_{a' \in |A|} A(s, a'; \theta, \alpha)$

Stable Optimization

- $Q(s, a; \theta, \alpha, \beta) = V(s; \theta, \beta) + A(s, a; \theta, \alpha) - (1/|A|) \Sigma_{a'} A(s, a'; \theta, \alpha)$
- The advantage can now change as fast a the mean instead of having to pace up with changes in the optimal action's advantage
- No free-lunch: Q function is off-target by a constant. But stability criteria is over-powering the results as observed in experiments

Policy Evaluation

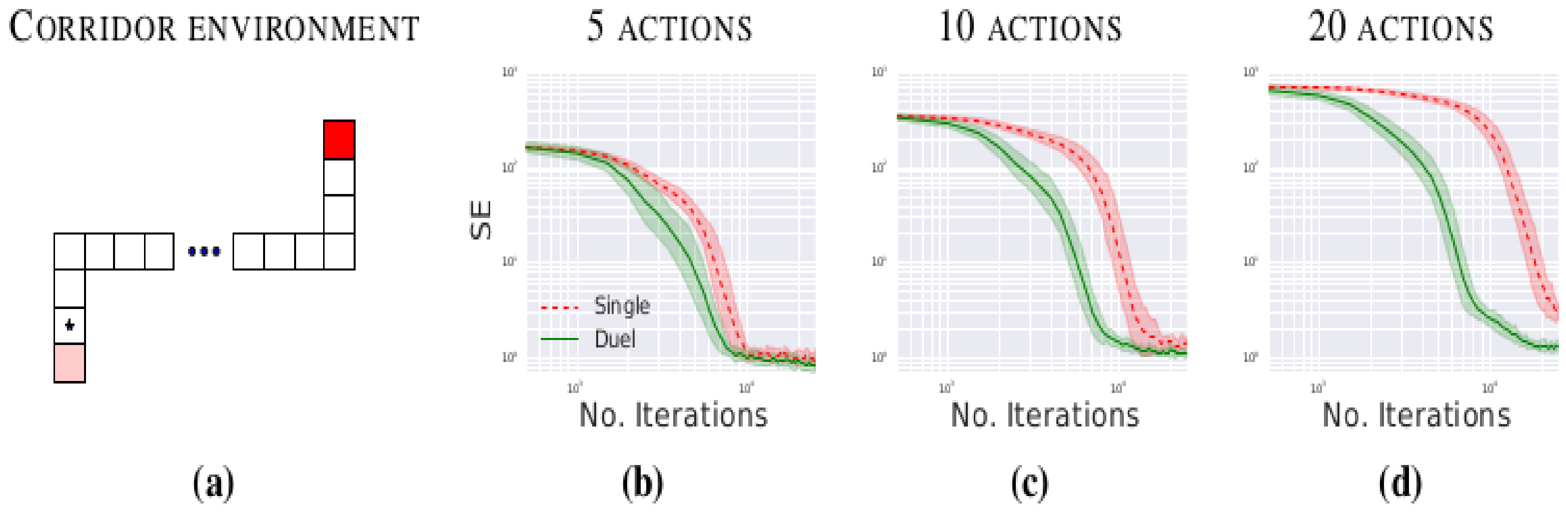
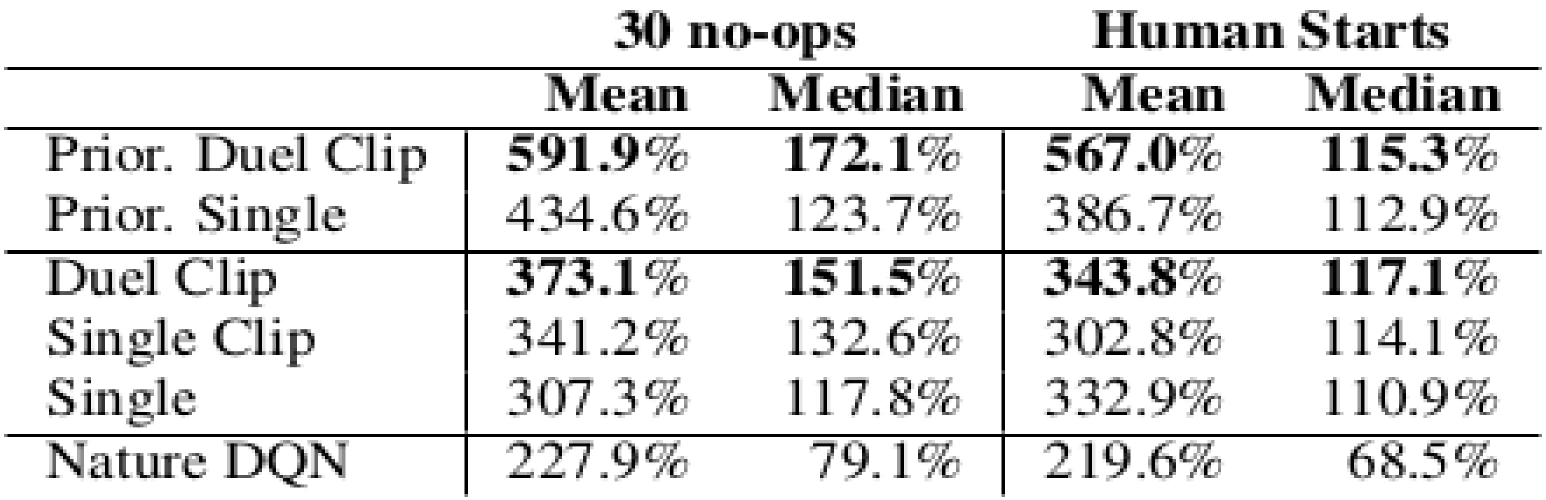


Figure 3. **(a)** The corridor environment. The star marks the starting state. The redness of a state signifies the reward the agent receives upon arrival. The game terminates upon reaching either reward state. The agent's actions are going up, down, left, right and no action. Plots **(b)**, **(c)** 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.

Improvement in the Performance of Atari Games



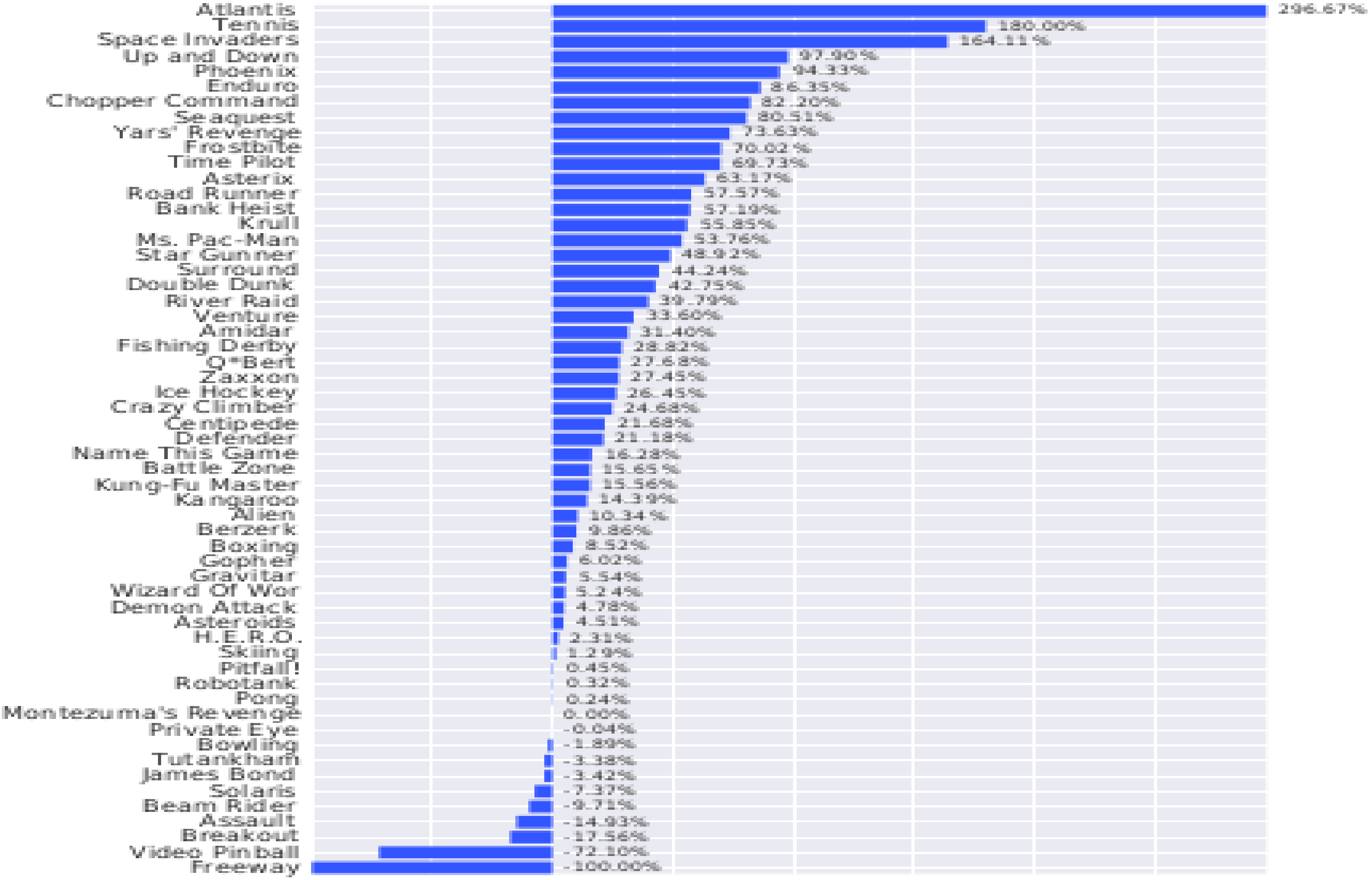
	30 no-ops		Human Starts	
	Mean	Median	Mean	Median
Prior. Duel Clip	**591.9%**	**172.1%**	**567.0%**	**115.3%**
Prior. Single	434.6%	123.7%	386.7%	112.9%
Duel Clip	**373.1%**	**151.5%**	**343.8%**	**117.1%**
Single Clip	341.2%	132.6%	302.8%	114.1%
Single	307.3%	117.8%	332.9%	110.9%
Nature DQN	227.9%	79.1%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

Why is % used for expressing mean and median?

$$\frac{\text{Score}_{\text{Agent}} - \text{Score}_{\text{Baseline}}}{\max\{\text{Score}_{\text{Human}}, \text{Score}_{\text{Baseline}}\} - \text{Score}_{\text{Random}}},$$

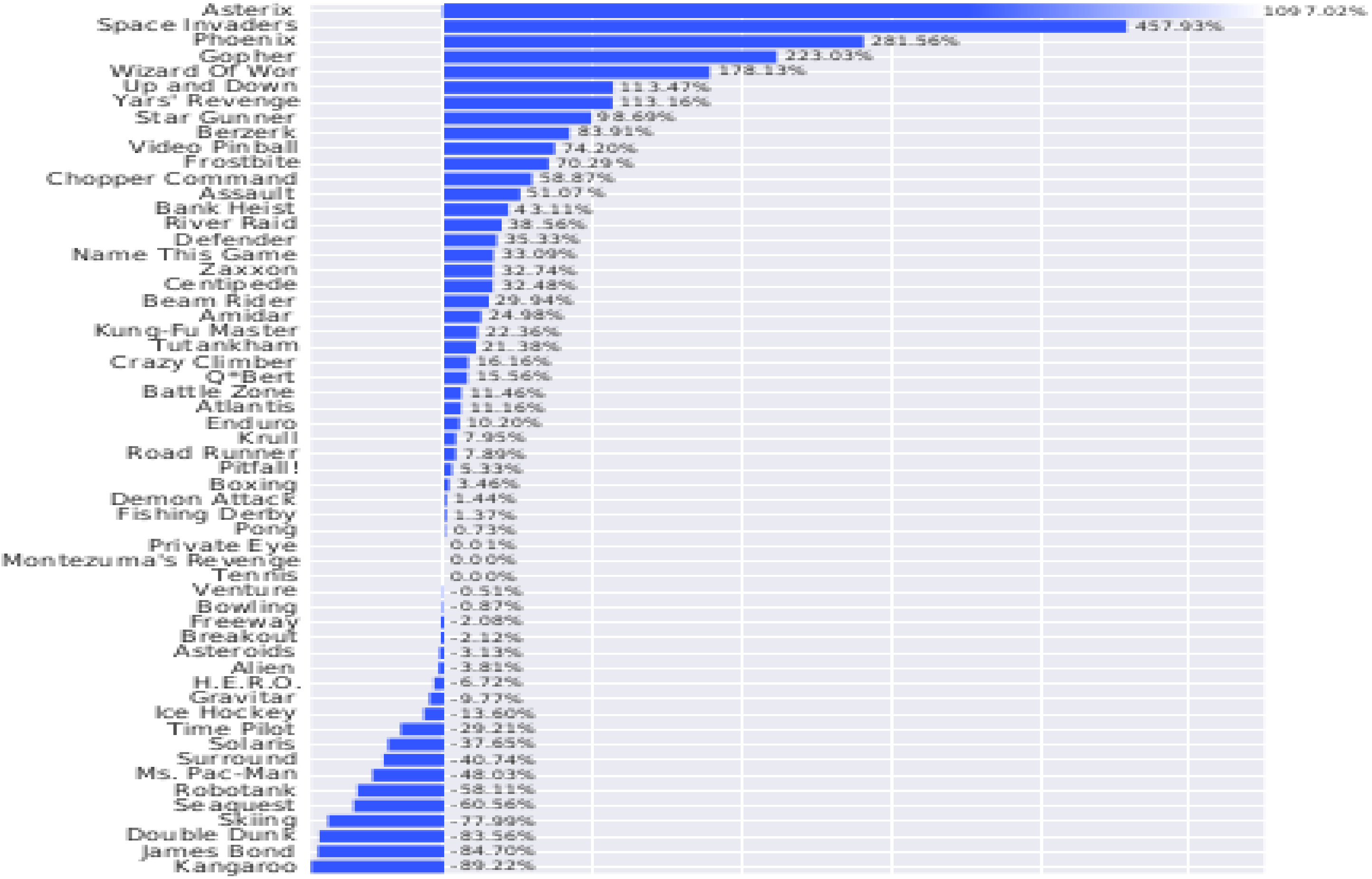
Improvement of Dueling Network Architecture

Baseline is single Q Network



Improvement of Dueling Network Architecture with Prioritized Performance Replay

Baseline is prioritized DDQN



Advantages of Dueling Network Architecture

- V of all actions are updated for every update of the Q values which leads to better approximation of the state values
- The advantage is prominent for large number of actions in the experiments
- The architecture is robust to noise which could have otherwise reordered actions due to the large scale difference between state values and action values.

References

1. Dueling Network Architectures for Deep Reinforcement Learning by Ziyu Wang, Tom Schaul, Matteo Hessel, Hado van Hasselt, Marc Lanctot, and Nando de Freitas, Proceedings of the 33rd International Conference on Machine Learning New York, NY, USA, 2016. JMLR: W&CP volume 48
2. Reinforcement learning: The Good, The Bad and The Ugly by Peter Dayan and Yael Niv in Current Opinion in Neurobiology Journal, 2008, volume 18, pp.1-12, DOI: 10.1016/j.conb.2008.08.003

Thank You