Dueling Network Architectures for Deep Reinforcement Learning

Bio-Medical Computing Laboratory

TAEHEUM CHO

17th May, 2018

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado van Hasselt, Marc Lanctot, Nando de, Freitas, Google DeepMind, London, UK. arXiv: 5 Apr 2016.

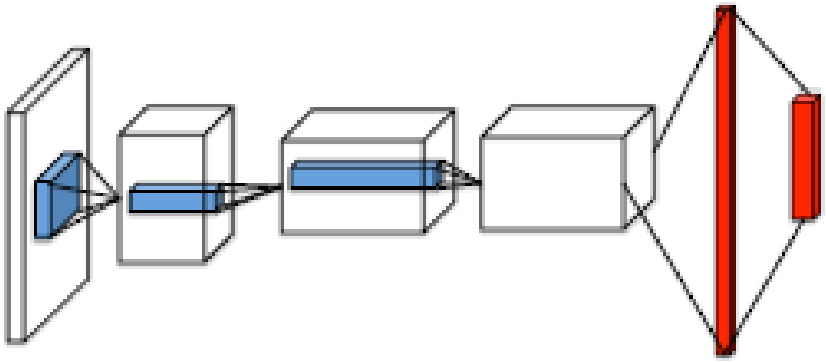
Introduction

Dueling architecture…

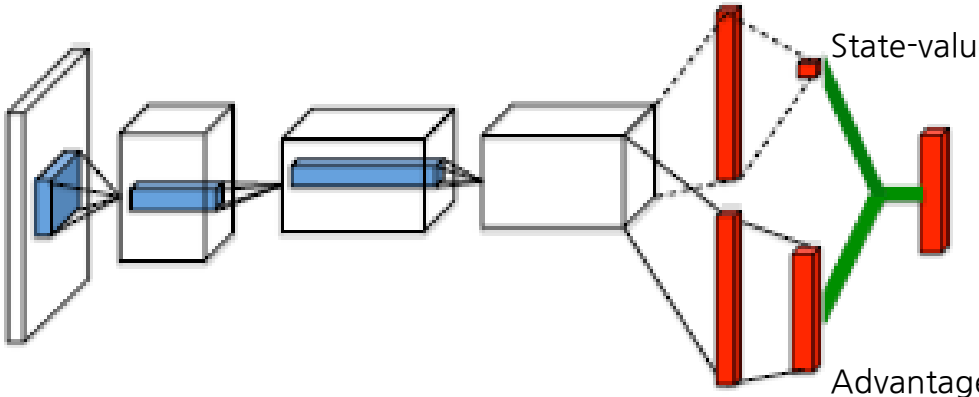
1. useful in states where its actions do not affect the environment in any relevant way.
2. takes an alternative but complementary approach that is better suited for model-free RL.
3. explicitly separates the representation of state values and (state-dependent) action advantages.
4. The two streams are combined via a special aggregating layer to produce an estimate of the state-action value function Q
5. can learn which states are (or are not) valuable, without having to learn the effect of each action for each state.

Introduction

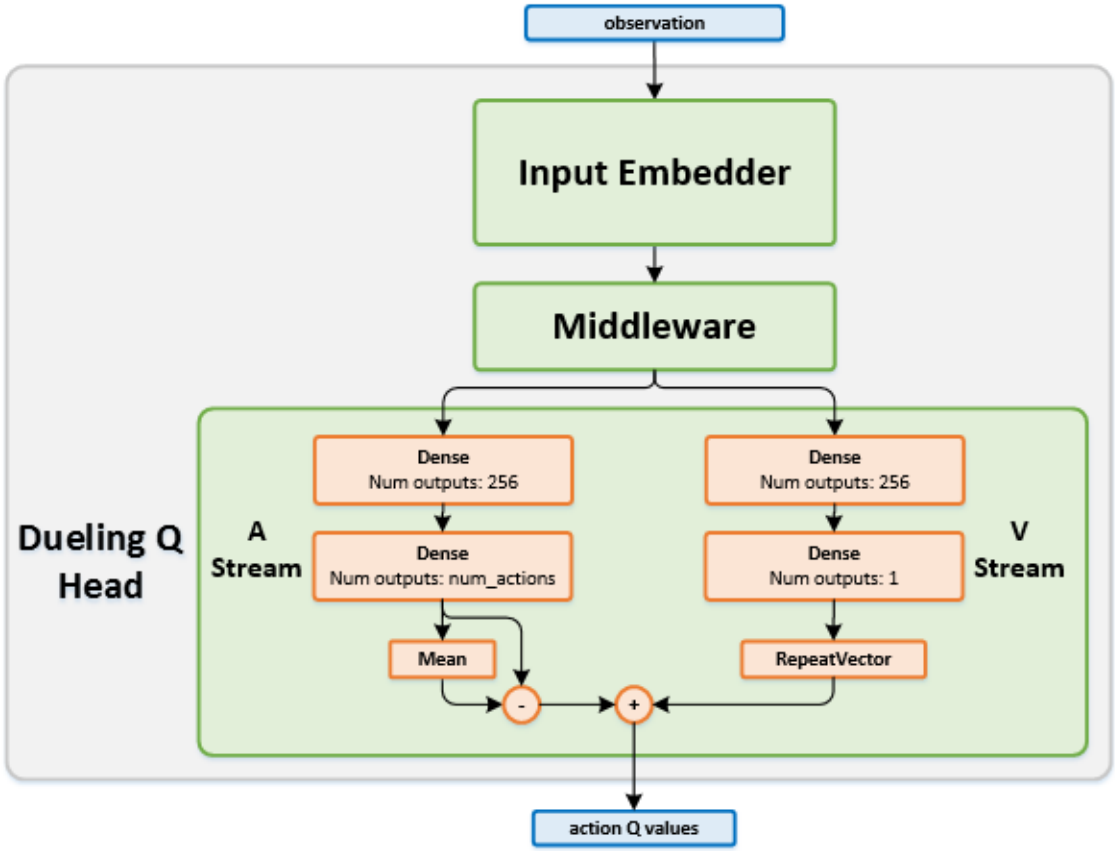
Single stream
Q-network



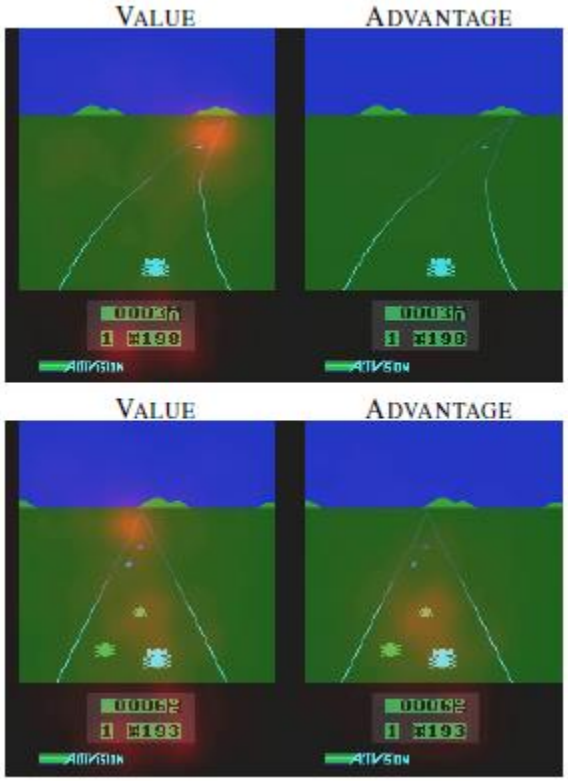
Dueling
Q-network



Introduction



Introduction



- The value stream learns to pay attention to the road.
- The advantage stream learns to pay attention only when there are cars immediately in front, so as to avoid collisions.

Background

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

Q-function using Bellman expectation equation

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

Q-function using Bellman Optimality equation

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

We define …

Background

Loss function of deep Q-network:

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

with

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

Loss function of double deep Q-network:

$$L_i(\theta_i) = \mathbb{E}_{s,a,r,s'}\left[\left(y_i^{DDQN} - Q(s,a;\theta_i)\right)^2\right],$$

with

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

Background

Prioritized Replay:

1. Their key idea was to increase the replay probability of experience tuples that have a high expected learning progress
2. This led to both faster learning and to better final policy quality across most games
3. we show that dueling architecture improves performance for both the uniform and the prioritized replay baselines

Schaul et al., 2016

The Dueling Network Architecture

Properties of dueling architecture:

1. It is unnecessary to estimate the value of each action choice
2. we instead use two sequences (or streams) of fully connected layers
3. Finally, the two streams are combined to produce a single output Q function.
4. Since the output of the dueling network is a Q function, it can be trained with the many existing algorithms, such as DDQN and SARSA.

$$Q(s,a;\theta,\alpha,\beta) = V(s;\theta,\beta) + A(s,a;\theta,\alpha)$$

θ: parameters of the convolutional layers

α: parameter of the state-value stream of F.C. layers

β: parameter of the advantage stream of F.C. layers

The Dueling Network Architecture

But we cannot recover V and A uniquely from this equation

$$Q(s,a;\theta,\alpha,\beta) = V(s;\theta,\beta) + A(s,a;\theta,\alpha)$$

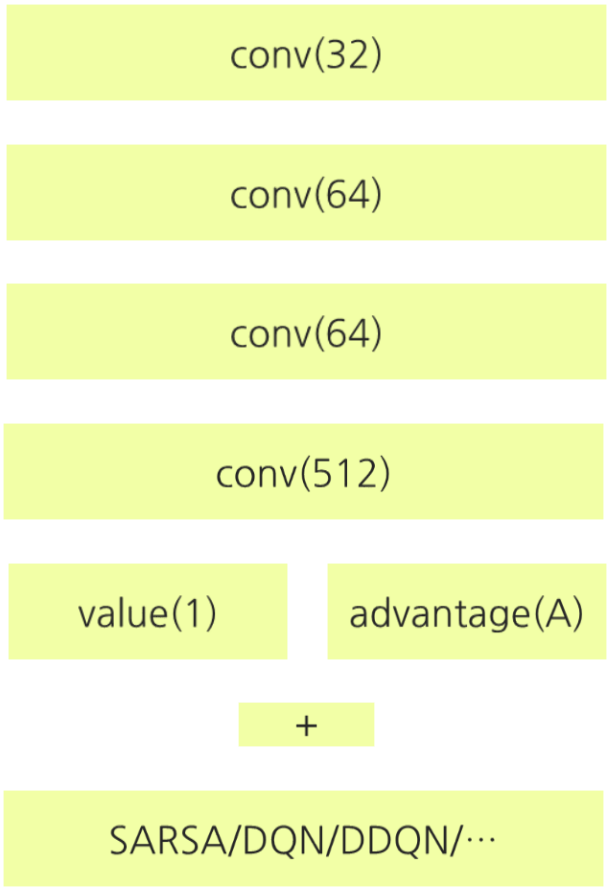
To address this issue of identifiability, we can force the advantage function estimator to have zero advantage at the chosen action

$$Q(s,a;\theta,\alpha,\beta) = V(s;\theta,\beta) + \left(A(s,a;\theta,\alpha) - \max_{a' \in |\mathcal{A}|} A(s,a';\theta,\alpha) \right)$$

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

time-efficiency
stability
same result

Experiment



1. We use 1024 hidden units for the first fully-connected layer so that dueling and single architectures have roughly the same number of parameters.
2. we start the game with up to 30 no-op actions to provide random starting positions for the agent.
3. we measure improvement in percentage in score over the better of human and baseline agent scores:

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

4. We took the maximum over human and baseline agent scores as it prevents insignificant changes to appear as large improvements when neither the agent in question nor the baseline are doing well.
5. To isolate the contributions of the dueling architecture, we re-train DDQN with a single stream network using exactly the same procedure as described above.

Experiment

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%

1. Prior. = Prioritized experience replay
2. Clip = Re-trained DDQN
3. Single = Original DDQN
4. Duel = Dueling architecture
5. 30 no-ops = random starting points
6. Human Starts = starting points sampled from a human expert's trajectory.

Experiment

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%

1. Duel Clip does better than Single Clip on 75.4% of the games (43 out of 57)
2. It also achieves higher scores compared to the Single baseline on 80.7% (46 out of 57) of the games
3. under the Human Starts metric, Duel Clip once again outperforms the single stream variants
4. In particular, our agent does better than the Single baseline on 70.2% (40 out of 57) games

Experiment

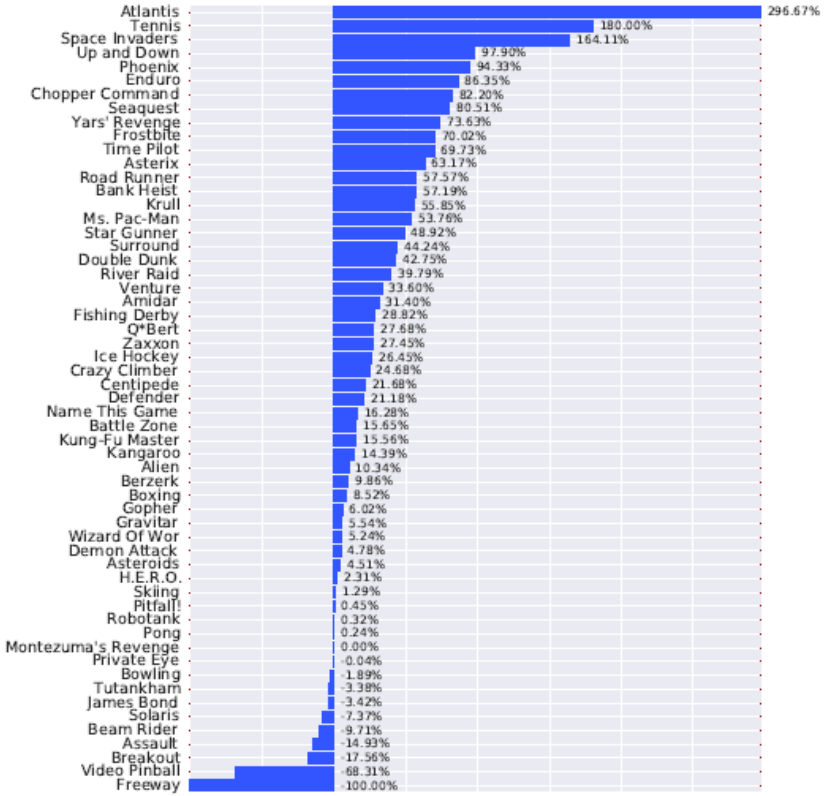


Figure 4. Improvements of dueling architecture over the baseline Single network of van Hasselt et al. (2015), Bars to the right indicate by how much the dueling network outperforms the single-stream network.

Experiment

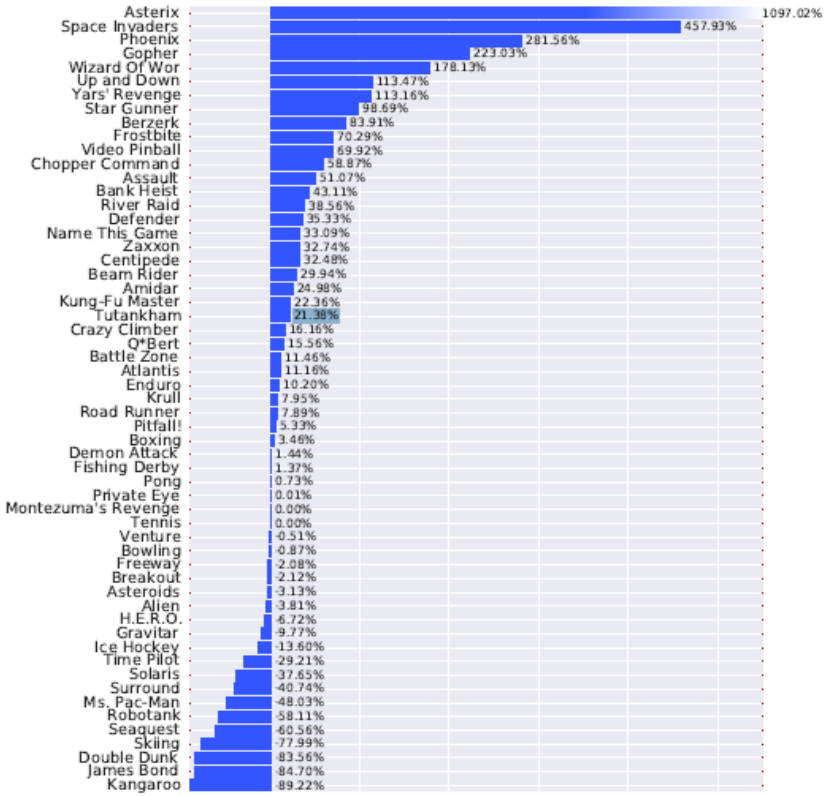


Figure 5. Improvements of dueling architecture over Prioritized DDQN baseline, Again, the dueling architecture leads to significant improvements over the single-stream baseline on the majority of games.

Discussion & Conclusions

1. The advantage of the dueling architecture lies partly in its ability to learn the state-value function efficiently.
2. With every update of the Q values in the dueling architecture, the value stream V is updated
3. This more frequent updating of the value stream in our approach allocates more resources to V
4. thus allows for better approximation of the state values, which in turn need to be accurate for temporal difference-based methods like Q-learning to work (Sutton & Barto, 1998).
5. The new dueling architecture leads to dramatic improvements over existing approaches for deep RL in the challenging Atari domain.