Dueling Network Architectures for Deep Reinforcement Learning

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado van Hasselt, Marc Lanctot, and Nando de Freitas

Reviewed by Zhao Song

January 27, 2017
The deep Q-network (DQN) [Mnih et al. 2015]:

- Architecture: Convolutional neural network + Fully connected linear layer;
- Target function:
  \[
y_{i}^{DQN} = r + \gamma \max_{a'} Q(s', a'; \theta^-)
\]
- Lost function at iteration \(i\):
  \[
  L_i(\theta_i) = E_{s, a, r, s'} \left[ y_{i}^{DQN} - Q(s, a; \theta_i) \right]^2
  \]
The Proposed Dueling Architecture

Figure 1. A popular single stream $Q$-network (top) and the dueling $Q$-network (bottom). The dueling network has two streams to separately estimate (scalar) state-value and the advantages for each action; the green output module implements equation (9) to combine them. Both networks output $Q$-values for each action.
• $Q$—function:

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

• Value function:

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

• Advantage function:

$$A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$$
Figure 2. See, attend and drive: Value and advantage saliency maps (red-tinted overlay) on the Atari game Enduro, for a trained dueling architecture. 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.
The Dueling Network

Key insight: Unnecessary to estimate the value of each action, for many states.

- With the goal of estimating $Q-$function
- Two separated streams with fully connected layers to estimate the value and advantage functions
  - Value of only one action updated in the single-stream architecture
  - Value function $V$ updated in the dueling architecture
- Streams then combined to estimate the $Q-$function
How to combine the two streams, $V(s; \theta, \beta)$ and $A(s, a; \theta, \alpha)$?

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

- Zero advantage at the chosen action:

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

- Another choice:

$$Q(s, a; \theta, \alpha, \beta) = V(s; \theta, \beta) + \left[ A(s, a; \theta, \alpha) - \frac{1}{|A|} \sum_{a'} A(s, a'; \theta, \alpha) \right]$$
The double DQN (DDQN) [Van Hasselt et al. 2016]:

- Proposed to mitigate the problem of overestimation in Q-learning
- Target function:

\[ y_{i}^{DDQN} = r + Q(s', \arg \max_{a'} Q(s', a' | \theta_i); \theta^-) \]

- Trained with back-propagation
Policy Evaluation

The corridor environment

- $\epsilon$-greedy policy
- Actions: Go up, down, left, right, and additional no-op(s)
- Three MLP layers followed by two MLP layer in the dueling part
- Performance measured by square error

$$\sum_{s,a} [Q(s, a; \theta) - Q^\pi(s, a)]^2$$
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.
Figure 4. Improvements of dueling architecture over the baseline Single network of van Hasselt et al. (2015), using the metric described in Equation (10). Bars to the right indicate by how much the dueling network outperforms the single-stream network.
A few tricks

- Rescale the gradients entering the last convolutional layer
- Clip the gradients to constraint their norm
- Prioritized experience replay [Schaul et al. 2016]
### Table 1. Mean and median scores across all 57 Atari games, measured in percentages of human performance.

<table>
<thead>
<tr>
<th></th>
<th>30 no-ops</th>
<th>Human Starts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Median</td>
</tr>
<tr>
<td>Prior. Duel Clip</td>
<td>591.9%</td>
<td>172.1%</td>
</tr>
<tr>
<td>Prior. Single</td>
<td>434.6%</td>
<td>123.7%</td>
</tr>
<tr>
<td>Duel Clip</td>
<td>373.1%</td>
<td>151.5%</td>
</tr>
<tr>
<td>Single Clip</td>
<td>341.2%</td>
<td>132.6%</td>
</tr>
<tr>
<td>Single</td>
<td>307.3%</td>
<td>117.8%</td>
</tr>
<tr>
<td>Nature DQN</td>
<td>227.9%</td>
<td>79.1%</td>
</tr>
</tbody>
</table>
