Faster Deep Reinforcement Learning with Slower Online Network

Kavosh Asadi, Rasool Fakoor, Omer Gottesman, Taesup Kim, Michael L. Littma, Alexander J. Smola

Proximal Bellman Operator

$(T_{c,f}^{\pi})^{n} = arg\min_{v'} \|v' - (T^{\pi})^{n}\|_2^2 + \frac{1}{c} D_f(v', v)$, где

$D_f(v', v) = f(v') - f(v) - \langle \nabla f(v), v' - v \rangle$

Если $D_f(v', v) = \|v' - v\|_2^2$, то

$(T_{c,f}^{\pi})^{n} = arg\min_{v'} \|v' - (T^{\pi})^{n}\|_2^2 + \frac{1}{c} \|v' - v\|_2^2$

Теорема. $T_{c,f}^{*}$ - сжимающий оператор с неподвижной точкой v^{*}

DQN PRO

$$\omega_{t+1} \leftarrow arg\min_{\omega} h(\omega_t, \omega) + \frac{1}{2\tilde{c}}\|\omega - \omega_t\|_2^2 \text{ , где}$$

$$h(\theta, \omega) = \hat{E}_{\langle s,a,r,s' \rangle}\left[\left(r + \gamma \max_{a'} \hat{Q}(s', a'; \theta) - \hat{Q}(s, a; \omega)\right)^2\right]$$

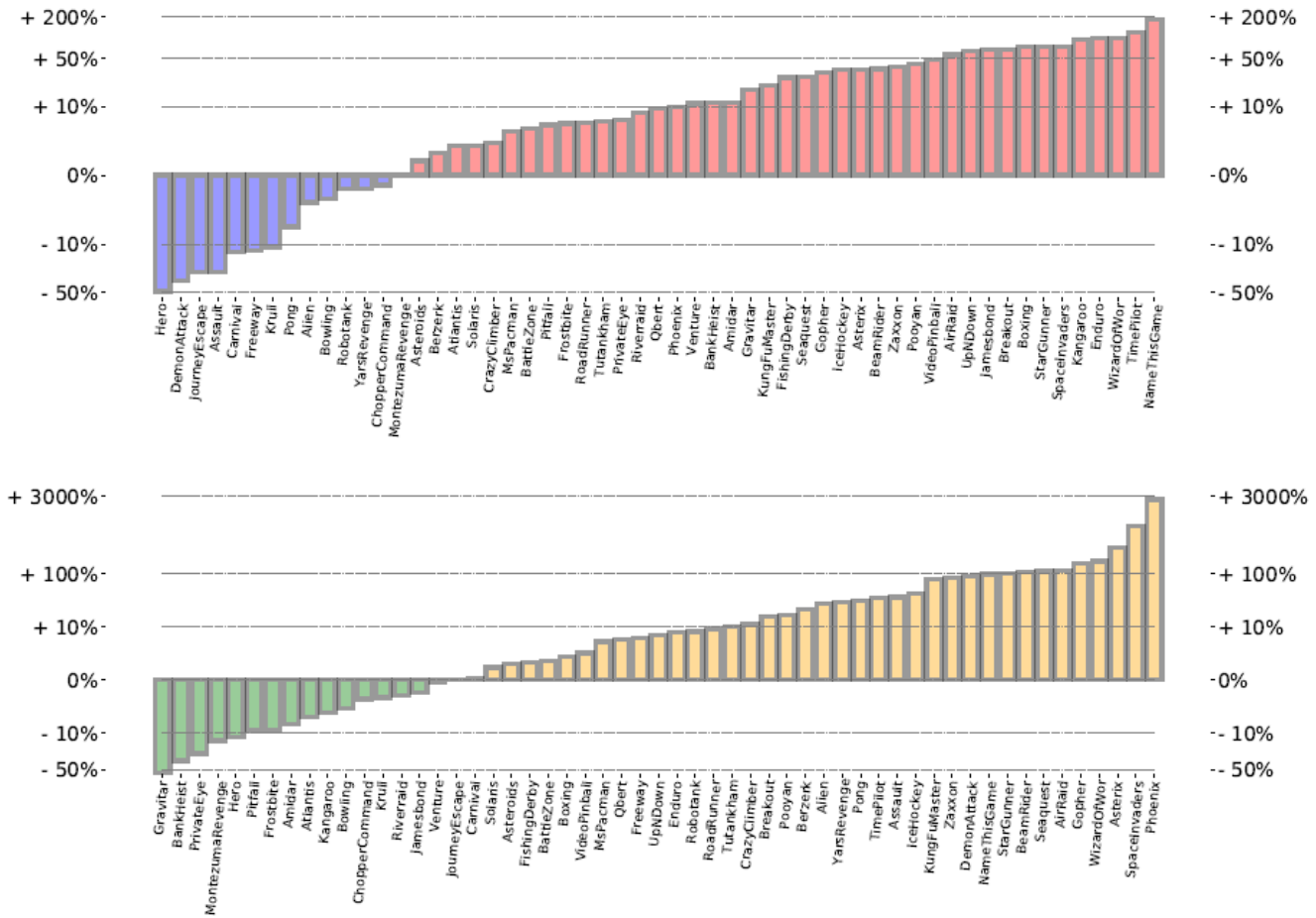
или

$$\omega_{t+1} \leftarrow \left(1 - \frac{\alpha}{\tilde{c}}\right)\omega + \frac{\alpha}{\tilde{c}}\omega_t - \alpha\nabla_2 h(\omega_t, \omega)$$

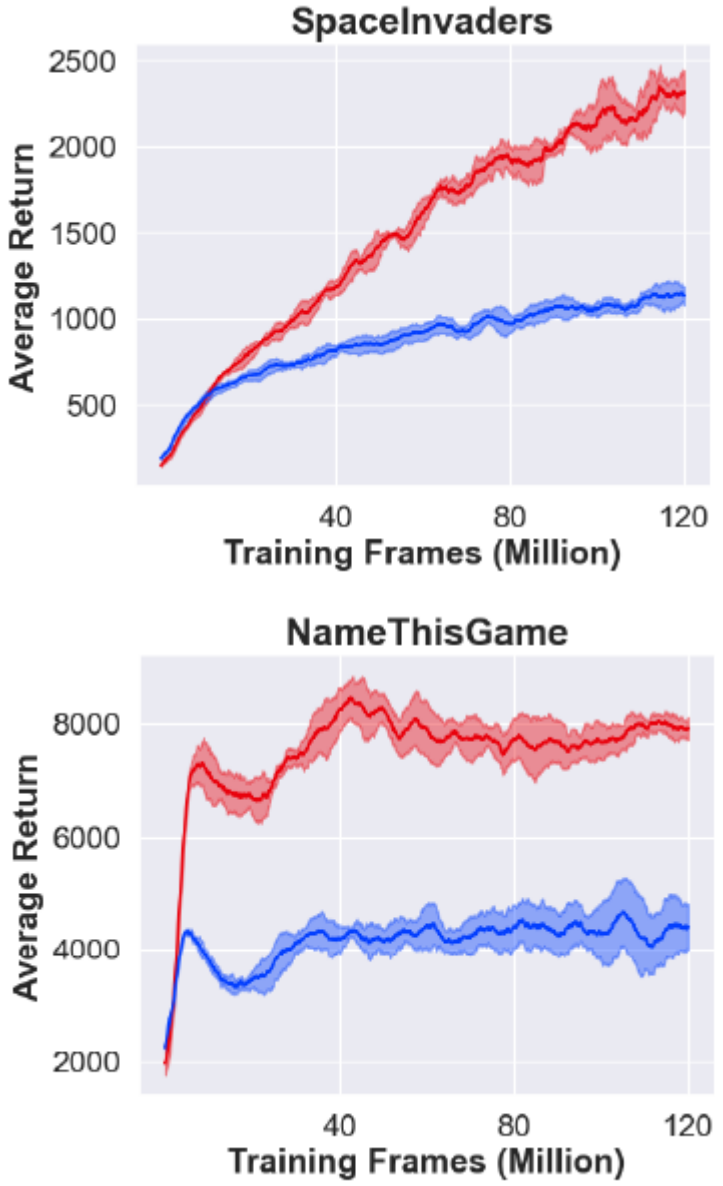
Algorithm 1 DQN with Proximal Iteration (DQN Pro)

Initialize θ, N, *period*, replay buffer $\mathcal{D}$, α, and $\tilde{c}$
$s \leftarrow$ env.reset(), $w \leftarrow \theta$, numUpdates $\leftarrow 0$
repeat
  $a \sim \epsilon$-greedy$\left(Q(s, \cdot; w)\right)$
  $s', r \leftarrow$ env.step(s, a)
  add $\langle s, a, r, s' \rangle$ to $\mathcal{D}$
  if s' is terminal **then**
    $s \leftarrow$ env.reset()
  end if
  for n in $\{1, \ldots, \text{N}\}$ **do**
    sample $\mathcal{B} = \{\langle s, a, r, s' \rangle\}$, compute $\nabla_w h(w)$
    $w \leftarrow \left(1 - (\alpha/\tilde{c})\right)w + (\alpha/\tilde{c})\theta - \alpha\nabla_w h(w)$
    numUpdates $\leftarrow$ numUpdates $+ 1$
    if numUpdates $\%$ *period* $= 0$ **then**
      $\theta \leftarrow w$
    end if
  end for
until convergence

Результаты исследователей

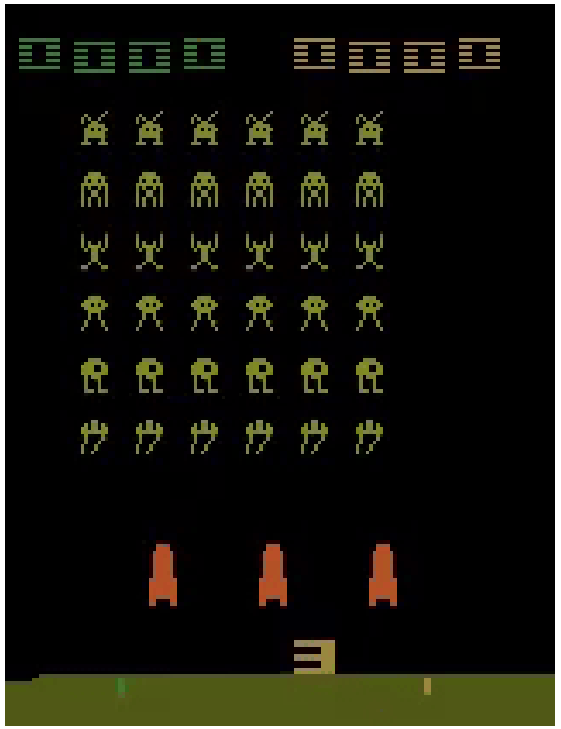


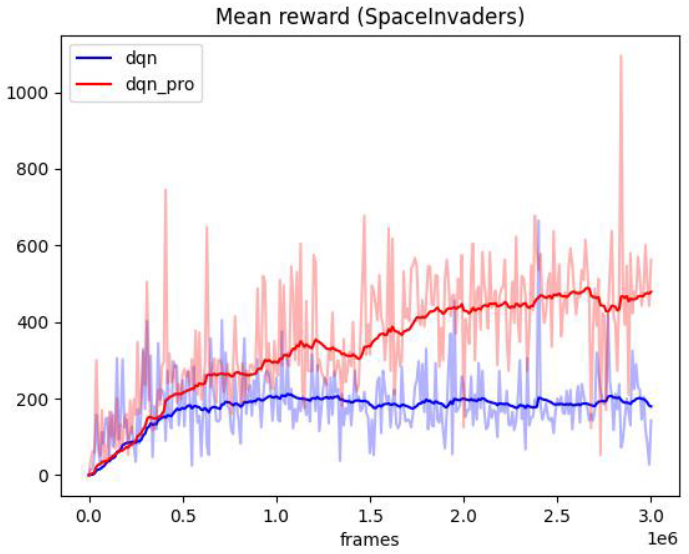
Финальное преимущество DQN Pro над DQN (вверху) и Rainbow Pro над Rainbow (внизу)

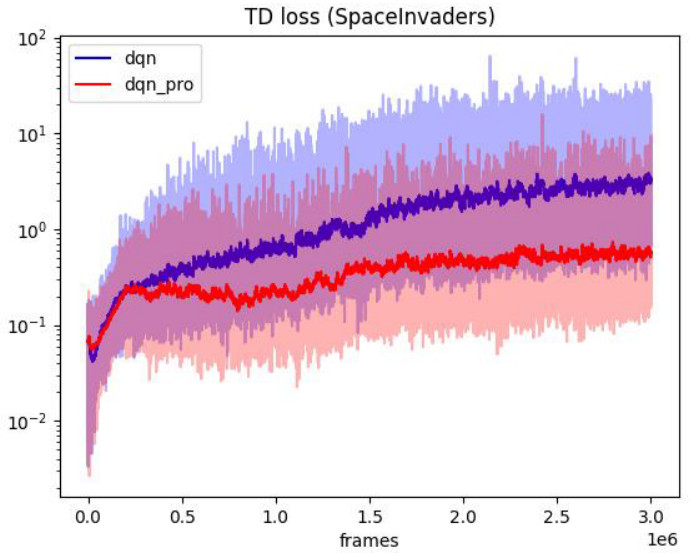


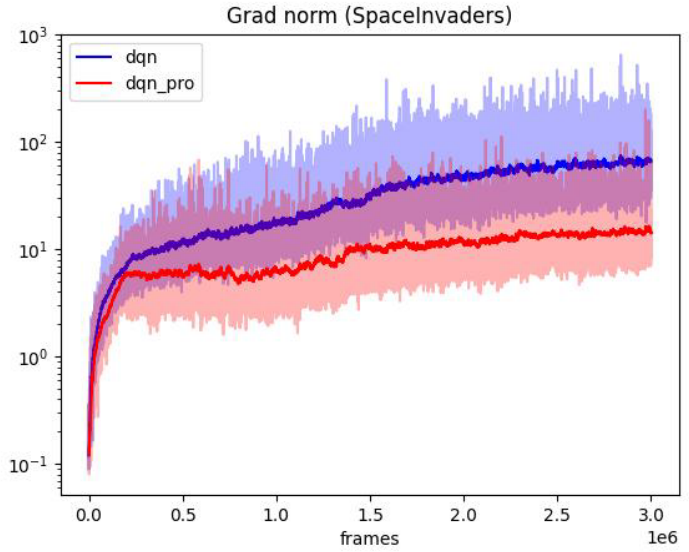
DQN Pro – красный, DQN - синий

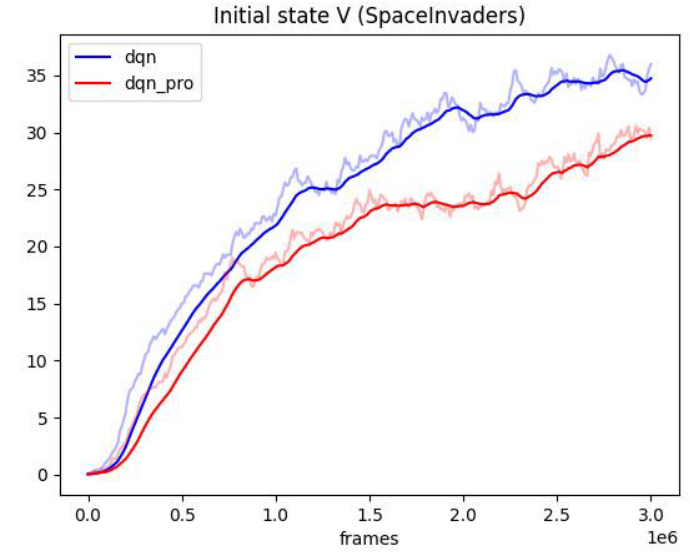
SpaceInvaders



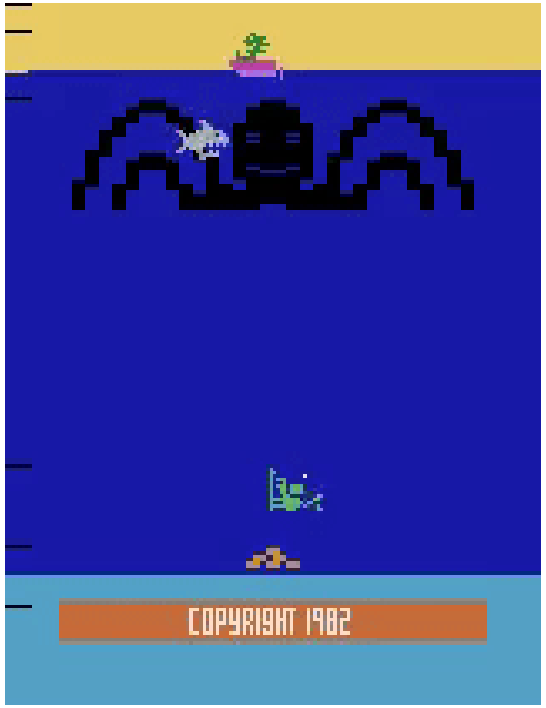


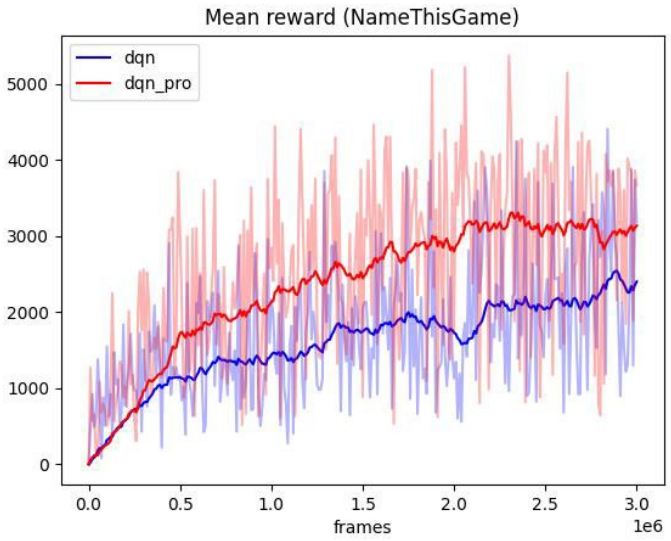


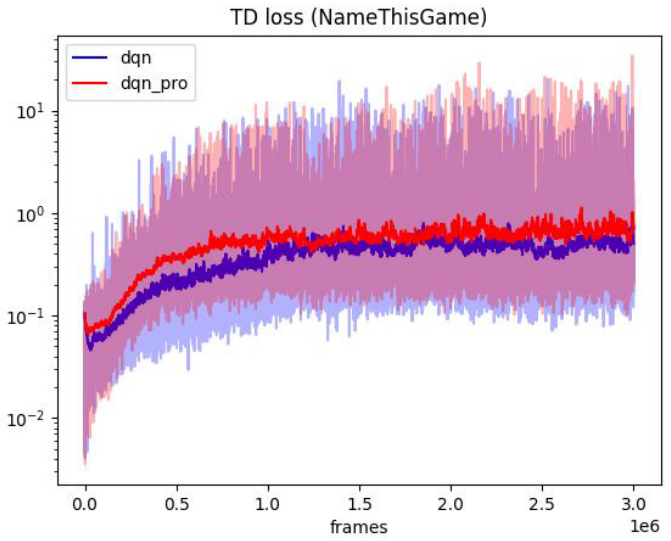


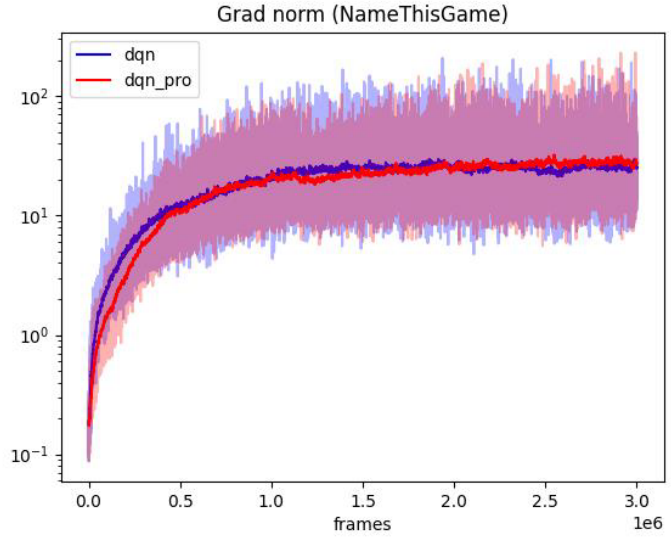


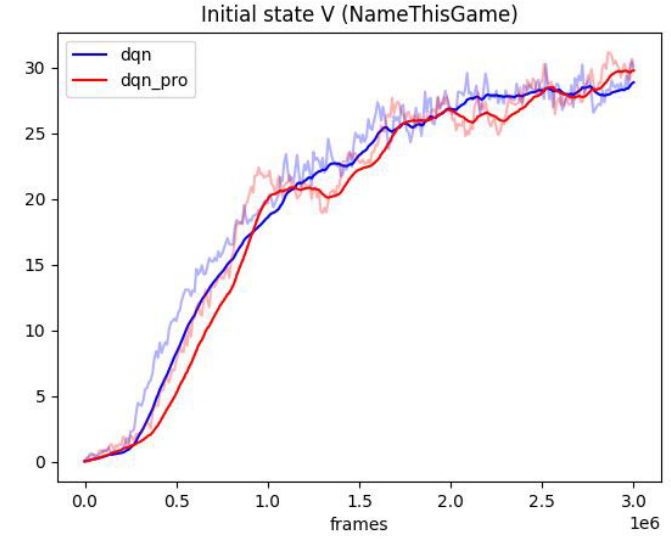
NameThisGame











Выводы

- DQN Pro позволяет добиться заметно большей награды по сравнению с DQN (но для этого по-прежнему нужно долго его обучать)
- Меньше склонен к переоценке значения V-функции

Спасибо за внимание!