From Policy Gradient to Actor-Critic methods

Proximal Policy Optimization (PPO)

Olivier Sigaud

Sorbonne Université
http://people.isir.upmc.fr/sigaud
There are two PPO algorithms
They are well covered on youtube videos
So only a quick overview here
Easy implementation, a lot used
Key question: is it Actor-Critic?
Proximal Policy Optimization (Algorithm 1)

- The conjugate gradient method of TRPO is not available in tensor libraries
- Same idea as TRPO, but uses a soft constraint on trust region rather than a hard one
- Instead of:
  \[
  \max_{\theta} \mathbb{E}_t \left[ \pi_{\theta}(a_t|s_t) \frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} A_{\pi_{\theta_{old}}}(s_t, a_t) \right] \\
  \text{subject to } \mathbb{E}_t \left[ KL(\pi_{\theta_{old}}(\cdot|s)||\pi_{\theta}(a_t|s_t)) \right] \leq \delta
  \]
- Rather use:
  \[
  \max_{\theta} \mathbb{E}_{s \sim \rho, a \sim \pi} \left[ \pi_{\theta}(a_t|s_t) \frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} A_{\pi_{\theta_{old}}}(s_t, a_t) \right] - \beta \mathbb{E}_{s \sim \rho} \left[ KL(\pi_{\theta_{old}}(\cdot|s)||\pi_{\theta}(a_t|s_t)) \right]
  \]
- Makes it possible to use SGD instead of conjugate gradient


Proximal Policy Optimization (Algorithm 2)

\[
\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)} \text{ may get huge if } \pi_{\theta_{old}} \text{ is very small}
\]

\[
\text{Clipped importance sampling loss (clipping the surrogate objective)}
\]

\[
r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}
\]

\[
L^{CLIP}(\theta) = \mathbb{E}_t[min(r_t(\theta) \hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon) \hat{A}_t)]
\]

\[
\text{Back-propagate } L^{CLIP}(\theta) \text{ through a policy network}
\]
Is PPO actor-critic?

- Improvement over TRPO, thus REINFORCE-like policy update
- But:
  - Algorithm: “PPO, actor-critic style”
  - In the Dota-2 paper: “PPO, a variant of advantage actor-critic, …”
- What matters is the critic (or baseline) update method
- Uses N-step Generalized Advantage Estimate instead of Monte Carlo
- Thus somewhere between MC and TD (same for ACKTR)
- Other properties:
  - Simpler implementation, better performance than TRPO
  - Does not use a replay buffer → more stable, less sample efficient
  - Still on-policy, πθ and πθold cannot differ much

PPO applications

- Massive parallel versions of PPO, with dedicated architectures
- Very few teams can afford such engineering and computing effort


Massive parallel updates

- Several workers in parallel: more i.i.d and faster exploration
- The acceleration is better than linear in the number of workers
- No need for a replay buffer (as in A3C), but loss of sample efficiency


From Policy Gradient to Actor-Critic methods

OpenIA five

- The LSTM deals with non-Markov data
- The vision layers are problem specific

Any question?

Send mail to: Olivier.Sigaud@upmc.fr
Igor Adamski, Robert Adamski, Tomasz Grel, Adam Jedrych, Kamil Kaczmarek, and Henryk Michalewski.  
Distributed deep reinforcement learning: Learn how to play atari games in 21 minutes.  

Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al.  
Solving rubik's cube with a robot hand.  

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al.  
Dota 2 with large scale deep reinforcement learning.  

IMPALA: Scalable distributed deep-rl with importance weighted actor-learner architectures.  

Emergence of locomotion behaviours in rich environments.  

Learning dexterous in-hand manipulation.  

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov.  
Proximal policy optimization algorithms.  