Benchmarking Deep Reinforcement Learning for Continuous Control

Reminders

Overview

- Benchmarking a number of different DRL algorithms to create a uniform comparison
- Benchmark tasks are classified into four groups:
  - Basic,
  - Locomotion
  - Partially observable (limited sensors, noisy observations or system identification),
  - Hierarchical tasks (where multiple things need to be learned separately - e.g. to walk and use walking to collect rewards)

Key ingredients

- Algorithms tested are:
  - REINFORCE,
  - TNPG (Truncated Natural Policy Gradient)
  - TRPO (Trust Region Policy Optimization)
  - RWR (Reward-Weighted Regression)
  - REPS (Relative Entropy Policy Search)
  - CEM and CMA-ES (Cross Entropy Method and Covariance Matrix Adaption Evolution Strategy)
  - DDPG (Deep Deterministic Policy Gradient)

REINFORCE

- Competitive with other state of the art
- Drawback that it sometimes suffers from local minima

TNPG and TRPO

- Listed as the same family, since their functionality is quite similar
- Both perform best on the following benchmarks, with TRPO being slightly better than TNPG

RWR

- Only one that does not require any parameter tuning
- Fails to solve challenging tasks such as locomotion

REPS

- Not a lot of said except that it does not perform great on average, due to local minima - highly dependent on initial policy
- Sometimes can achieve state-of-the-art results

Gradient-free methods (CEM and CMA-ES)

- Great performance on low dimensional tasks
- Suffer from curse of dimensionality - memory requirements increase substantially

DDPG

- Rewards rescaled by 0.1 to improve stability
- Found to be less stable than other algorithms
- Still performed well overall

Comments

- What I particularly liked about the paper is that the algorithms were briefly explained, but still well enough to have an intuition, and that a lot of great references with seminal work are listed
- What I dislike here is the lack of other important algorithms (such as GPS)
- Also, some algortihms were optimized by hyperparameter search, while the others were taken exactly as reported in the paper. I find this to be slightly biased, so I would not trust the results completely
- Nevertheless, DDPG, according to multiple other sources, seems to be less stable