Research Evaluation Report

Saptam Bakshi

January 16, 2022

Courses attended

Semester	Subject	Marks
I	AIML	86
	Cryptology-I	74
	Automata & Formal Languages	67
	Discrete Mathematics	82
II	Advanced Machine Learning	84
	Design & Analysis of Algorithms	76
	Quantum Information & Cryptology	76
	Research Methodology	91

Research work carried out

We present a new off-policy Reinforcement Learning algorithm called Opportunistic Actor-Critic (OPAC) that combines both TD3 and SAC retaining their core features. Both TD3 and SAC, two recent Reinforcement Learning algorithms, have been immensely successful in tackling robotic control tasks in the past. Unlike TD3 or SAC though, it uses three critics for a more optimistic decision making. We develop the theory of Clipped Triple Q-learning and establish its proof of convergence. We present detailed empirical result which confirm OPAC's superior performance in terms of cumulative rewards compared to SAC and TD3 in various robotic control tasks.

More details about our work can be found in the accompanying pre-recorded video presentation.

Future plan

We will try to efficiently solve either of the following two problems using Reinforcement Learning. Both the problems can be thought of as a type of hyperparameter optimisation task where we are to find out the *optimal hyperparameter* of a model for which some pre-defined performance metric is optimised.

- **Neural architecture search (NAS):** It is a technique for automating the design of artificial neural networks, a widely used model in the field of machine learning. NAS has been used to design networks that are on par or outperform hand-designed architectures.

  NAS can be characterised as a system with three major components namely: i) search space, ii) search algorithm, and iii) evaluation strategy. The search algorithm samples a population of network architecture candidates. It receives the child model performance metrics as rewards (e.g. high accuracy, low latency) and optimises to generate high-performance architecture candidates.

- **Parameter search in evolutionary computation:** In evolutionary computation, differential evolution (DE) is a method that optimises a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. Such methods are commonly known as metaheuristics as they make few or no assumptions about the problem being optimised and can search very large spaces of candidate solutions.

  The choice of DE parameters can have a large impact on optimisation performance. Selecting the DE parameters that yield good performance has therefore been the subject of much research.

Papers read

[CVLH19] Kamil Ciosek, Quan Ho Vuong, Robert Tyler Loftin, and Katja Hofmann, *Better Exploration with Optimistic Actor-Critic*, NeurIPS, 2019.

[DS10] Swagatam Das and Ponnuthurai Nagaratnam Suganthan, *Differential Evolution: A Survey of the State-of-the-Art*, IEEE transactions on evolutionary computation **15** (2010), no. 1, 4–31.

[FHM18] Scott Fujimoto, Herke Hoof, and David Meger, *Addressing Function Approximation Error in Actor-Critic Methods*, International Conference on Machine Learning, PMLR, 2018, pp. 1587–1596.

[Has10] Hado Hasselt, *Double Q-learning*, Advances in neural information processing systems **23** (2010), 2613–2621.

[HZAL18] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine, *Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor*, International conference on machine learning, PMLR, 2018, pp. 1861–1870.

[HZH$^+$18] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine, *Soft Actor-Critic Algorithms and Applications*, arXiv e-prints (2018), arXiv:1812.05905.

[SJLS00] Satinder Singh, Tommi Jaakkola, Michael L. Littman, and Csaba Szepesvári, *Convergence Results for Single-Step On-Policy Reinforcement-Learning Algorithms*, Machine Learning **38** (2000), no. 3, 287–308.

[SLH$^+$14] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller, *Deterministic Policy Gradient Algorithms*, International conference on machine learning, PMLR, 2014, pp. 387–395.