Reinforcement learning concentrates on solving sequential problems. In each time step, an agent observes the state of an environment in which it operates and receives some reward. Based on these, it chooses an action, which is then executed. The environment passes to the next time step, and a new state is observed and so on, so forth. The goal is to find a policy, being a way of choosing actions, which generates the maximal total sum of rewards during an episode. Formally, a policy is a function from the states of an environment to actions; in deep reinforcement, learning is realized by a deep neural network. Archetypical examples of reinforcement learning problems are single-player games. However, its formalism is broad and can host a plethora of problems, e.g., autonomous driving, robot steering, power-grid optimization, etc. Some researches hope that reinforcement learning might be a way towards truly intelligent systems.

Based on these hopes, deep Reinforcement learning has undergone rapid growth, with thousands of papers in 2019 alone. There have been spectacular successes, most publicized ones in competitive multi-player games like Dota2 and Starcraft. There are multiple "buts", however. RL methods require a lot of experience and computational power (e.g., 2000 GPUs and 100k CPUs were used in Dota 2 training). This is not only a practical barrier for most real-world applications but also indications of deeper problems. There have been raising voices, including prominent researchers [3, 1, 2], calling for revisiting the field.

Such stories are common in science; after a period of exponential growth, there a need for a step back, reevaluate fundamental assumptions, and identify new research directions. This project aims to be a part of these attempts. We plan to contribute in two subfields *unsupervised reinforcement learning* and *multi-task reinforcement learning/meta learning.* Both of them are considered to be important. They intend to provide a more robust training signal and thus better efficiency and agile behaviors. Intuitively, an agent learning using solely the reward signal provided by the environment has little incentive to "understand better" the environment. One can illustratively, imagine the agent being a sluggish pupil making the minimal effort to solve the task. This is often detrimental in a broader perspective; the learned solutions are either mediocre or not being able to transfer to even slightly different settings.

References

[1] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI*, pages 3207–3214, 2018.

[2] Andrew Ilyas, Logan Engstrom, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Are deep policy gradient algorithms truly policy gradient algorithms? *CoRR*, abs/1811.02553, 2018.

[3] Marlos C. Machado, Marc G. Bellemare, Erik Talvitie, Joel Veness, Matthew J. Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents (extended abstract). In *IJCAI*, pages 5573–5577, 2018.