Wen Sun

Title: The Role of Dataset Reset in Online Reinforcement Learning from Human Feedback

Abstract: Online Reinforcement Learning from Human Feedback (RLHF) is a paradigm for fine-tuning generative models such as Large Language Models (LLMs), which has produced by far the most powerful LLMs such as ChatGPTs and GPT4. In this work, leveraging the key property of text generation --- the ability to reset anywhere, we propose a new and specialized RL algorithm that can outperform standard RL algorithms such as Proximal Policy Optimization (PPO) when used in the RLHF pipeline. Our new algorithm Dataset Reset Policy Gradient (DR-PG), leverages the existing offline preference dataset during online policy training via dataset reset: it resets the policy gradient optimizer to the states in the offline dataset, instead of always starting from the initial state distribution. Offline preference dataset provides more informative states (i.e., more relevant to the underlying preference that we are optimizing) from which we can reset our RL optimizer and perform policy optimization. In theory, we show that when using DR-PG in the RLHF pipeline, DR-PG learns to perform at least as good as any policy that is covered by the offline dataset. In experiments, we demonstrate that in a standard RLHF benchmark, the generation from DR-PG is significantly better than the generation from PPO under the metric of GPT4 win-rate.

Bio: Wen Sun is an Assistant Professor in the Computer Science Department at Cornell. Before that, he was a postdoctoral researcher at Microsoft Research NYC and he completed his Ph.D in 2019 from the Robotics Institute at Carnegie Mellon University. He is generally interested in machine learning, especially Reinforcement Learning. Much of his current research is about designing algorithms for efficient sequential decision making, understanding exploration and exploitation, and how to leverage offline data to overcome exploration.