trlX: A Framework for Large Scale Open Source RLHF

Louis Castricato
EleutherAI

Abstract

Reinforcement learning from human feedback (RLHF) utilizes human feedback to better align large language models with human preferences via online optimization against a learned reward model. Current RLHF paradigms rely on Proximal Policy Optimization (PPO), which quickly becomes a challenge to implement and scale up to large architectures. To address this difficulty we created the trlX library (Havrilla et al., 2023) as a feature-complete open-source framework for RLHF fine-tuning of models up to and exceeding 70 billion parameters. We implemented support for multiple types of distributed training including distributed data parallel, model sharded, as well as tensor, sequential, and pipeline parallelism.

Biography

Louis Castricato is a research scientist at EleutherAI, working on RLHF infrastructure and engineering. Previously, Louis was head of LLMs at Stability AI and team lead at CarperAI, the largest open source RLHF group, as well as a PhD student at Brown University.

References

Alexander Havrilla, Duy Van Phung, Maksym Zhuravinskyi, Aman Tiwari, Jonathan Tow, Shivanshu Purohit, Stella Biderman, Quentin Anthony, Ethan Kim, and Louis Castricato. 2023. trlx: A framework for large scale reinforcement learning from human feedback. In *Conference on Empirical Methods in Natural Language Processing*.

Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023), pages 245–245