Efficient Large-scale Machine Learning

Under the classical learning theory setting, if we have sufficient training samples, computational resources, along with a powerful first-order method to optimize an empirical risk properly, then the output of the first-order method is expected to achieve a small test error. For high-dimensional and non-convex settings with deep neural networks (DNNs), minimizing the empirical risk is a challenging optimization task due to non-convexity and lack of guarantees in terms of global optimality. Beyond empirical risk minimization, formulating the problems of training generative adversarial networks (GANs) and more general and possibly non-zero-sum game-theoretic settings require more complicated mathematical frameworks.

In terms of implementation in a synchronous system with K nodes, first-order solvers for empirical risk minimization and VI-solvers are scaled by distributing computation among nodes, e.g., by partitioning the entire dataset in a cloud data center, followed by aggregation of local computations. Nodes can be, e.g., hospitals and cell phones that train a global model or personalized models collaboratively in a federated learning setting.

Learning under Resource Constraints.

Energy: GPT-3 has been trained on 10,000 V100 GPUs in a Microsoft cloud data center with unprecedented communication costs and Carbon footprint. The total emissions of training Llama 2 family of models and GPT-3 are estimated to be 539 and 552 tons of CO2 equivalent, respectively, which is not sustainable. Quantization, sparisification, and pruning of the underlying neural networks and data movement minimization are proposed to make ML more efficient as data movements between global and shared memory dominates the energy consumption of hardware platforms and the total energy consumption is a monotonically increasing function of the model size.

We have significantly accelerated ML in multi-node systems, i.e., handled scalability issues by developing highly communication-efficient and large-scale training in supervised learning, and training deep generative adversarial networks without sacrificing accuracy [1–3].

To make large-scale and data-parallel ML more efficient, we need to handle computation (in chip) and communication costs. In this project, our goal is to develop efficient in-chip elementary operations in lower precisions jointly with communication-efficient model training and inference to make AI training and inference energy-efficient.

This project aligns with the theme of “Energy systems” in particular, energy efficiency and smart energy and flexible energy systems.

Supervisor: Associate Professor Ali Ramezani-Kebrya

Preferred background: computer science, statistics, machine learning, data science, mathematics, physics, engineering, or other relevant field

Number of Available Projects: 2

Project period: flexible

Expected outcome: developing novel and efficient large-scale ML training and inference schemes

References:

[1] Fartash Faghri, Iman Tabrizian, Ilia Markov, Dan Alistarh, Daniel M. Roy, and Ali Ramezani-Kebrya. Adaptive gradient quantization for data-parallel SGD. In Advances in Neural Information Processing Systems (NeurIPS), 2020.

[2] Ali Ramezani-Kebrya, Fartash Faghri, Ilya Markov, Vitalii Aksenov, Dan Alistarh, and Daniel M. Roy. NUQSGD: Provably communication-efficient data-parallel SGD via nonuniform quantization. Journal of Machine Learning Research (JMLR), 22(114):1–43, 2021.

[3] Ali Ramezani-Kebrya, Kimon Antonakopoulos, Igor Krawczuk, Justin Deschenaux, and Volkan Cevher. Distributed extra-gradient with optimal complexity and communication guarantees. In International Conference on Learning Representations (ICLR), 2023.