PhD offer: New Efficient Algorithms for Large-Scale Training of Transformers

Job Offer: Internship available immediately, leading to a fully-funded 3-year PhD.
Advisor: Dr. Edouard Oyallon (HDR, CNRS, Sorbonne University), `edouard.oyallon@cnrs.fr`
Main Location: ISIR, Sorbonne University: Jussieu campus, centrally located in Paris.
Note: Possible collaboration with MILA under Pr. Eugene Belilovsky.
Application: Send CV, grade transcript, and optionally 1 or 2 referees.

Context: Current training algorithms struggle to reach peak Model FLOPS Utilization (MFU), typically around 70% [1], due to communication bottlenecks, synchronization overhead, memory constraints, and inefficient parallelism. Despite advances like Tensor Parallelism (TP), Distributed Data Parallelism (DDP), Fully Sharded Data Parallelism (FSDP), and Pipeline Parallelism, many challenges remain.This research seeks to maximize MFU and develop highly parallelizable algorithms for exascale (and decentralized) multi-GPU training. Key challenges include:

- **Accelerating communication** [2]: Optimizing topology-aware workload, minimizing synchronization, and compressing data to reduce transfer latency.
- **Overlapping communication and computation** [3]: New algorithms to maximize GPU throughput with minimal memory overhead.
- **Optimizing training** [4]: Enhancing parallelism, memory efficiency, and adaptive optimizers.
- **Improving model parallelism** [5]: Advancing dynamic layer partitioning and higher-parallelism beyond standard pipelining.
- **Leveraging asynchrony** [6]: Developing robust training procedure with stale gradient and communication tolerance.

This research is highly applied but welcomes theoretical contributions with tangible impacts. (e.g., how [6] influenced [2]).

References

[1] PyTorch Team. Maximizing training throughput, 2024. Blog post.

[2] A. Nabli, E. Belilovsky, and E. Oyallon. A2cid2: Accelerating asynchronous communication in decentralized deep learning. In *NeurIPS*, 2023.

[3] A. Nabli, L. Fournier, P. Erbacher, L. Serrano, E. Belilovsky, and E. Oyallon. Acco: Accumulate while you communicate, hiding communications in distributed llm training, 2024. Preprint.

[4] B. Thérien, C.-É. Joseph, B. Knyazev, E. Oyallon, I. Rish, and E. Belilovsky. Lo: Compute-efficient meta-generalization of learned optimizers, 2024. Preprint.

[5] S. Rivaud, L. Fournier, T. Pumir, E. Belilovsky, M. Eickenberg, and E. Oyallon. Petra: Parallel end-to-end training with reversible architectures, 2025. ICLR.

[6] A. Nabli and E. Oyallon. Dadao: Decoupled accelerated decentralized asynchronous optimization. In *ICML*, 2023.