Internship: Optimization and fine tuning of LLM (Large Langage Models)

Supervisor: Prof. E-G. Talbi
INRIA & University of Lille
Contact: el-ghazali.talbi@univ-lille.fr

This internship will be carried out in the framework of the PEPR (Programme et Equipement Prioritaire de Recherche Numpex (Exama project).

Context

Many scientific and industrial disciplines are concerned by big optimization problems (BOPs). The goal of this work is to come up with breakthrough in optimization algorithms on LLMs (Large Language Models) composed of trillions of parameters. The convergence between optimization algorithms and generative AI is an important in AI and High Performance Computing (HPC).

Inference with Large Language Models is costly and often dominates the life cycle cost of LLM-based services. Naural Architecture Search (NAS) can automatically find architectures optimizing the tradeoffs between accuracy and inference cost.

Roadmap

NAS for LLMs architectures is computationally prohibitive.

In this work, we will investigate the use of efficient optimization algorithms (example: parallel fractal optimization) to reduce the latency of real-world commercial web-scale text prediction system. The goal of this work is to solve the NAS problem to find an architecture that when trained with data D and training algorithm A, produces a model that has similar accuracy but significantly reduced latency.

The tasks composting this work are summarized below:

- Modeling and analysis of the NAS problem
- Solving of the problem using original and high-performance optimization algorithms
- Application to well known LLM such as GPT

Location: INRIA Lille

References

[1] T. Firmin, E-G. Talbi, «A framework for fractal based optimization", 2024.
[2] Raiaan, M. A. K., et al. (2024). A review on large Language Models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*.
[3] Javaheripi, M., et al. (2022). Litetransformersearch: Training-free neural architecture search for efficient language models. *Advances in Neural Information Processing Systems*, *35*, 24254-24267.
[4] Javaheripi, Mojan, et al. "Litetransformersearch: Training-free neural architecture search for efficient language models." *Advances in Neural Information Processing Systems* 35 (2022): 24254-24267.
[5] E-G. Talbi, "Metaheuristics: from design to implementation", *Wiley*, 2009.