Internship Topic:

“Efficient Training of Neural Networks”

Problem: modern neural networks might be too big to fit one GPU during training (besides model parameters, we need to store optimizer state, gradients, and activations needed to perform backpropagation).

Solution: To reduce memory footprint per GPU during training iteration we can apply different approaches.

- One of them is *activation checkpointing*. The idea is that instead of storing all intermediate activations required for gradient computation during backward, we save for backward only part of activations and recompute the rest during backward. Such a technique allows reducing memory footprint for one GPU due to some computational overhead.
- Another is *model parallelism,* which proposes storing different parts of a model on different devices and requires additional time for communication during training (to send/receive activations and gradients).
- *Pipeline parallelism* tries to optimize multi-GPU training by additionally splitting a batch of training data into micro batches and proposing an efficient schedule to compute forward and backward passes through micro batches.

Task:

The goal of the current project is to study and implement different techniques for efficient multi-GPU training. Firstly, you’ll implement baselines using different activation checkpointing techniques from Rockmate (Topal library with advanced activation checkpointing techniques) and pipelining techniques from PyTorch and DeepSpeed (Microsoft library for training optimization DeepSpeed). Then you’ll develop and implement new approaches for combining pipelining and checkpointing. You will compare them in terms of memory footprint / computational time trade-off, and contribute to the software package. You'll perform experiments with modern neural networks, including GPT-like models and Neural Operators. You'll analyze the performance of models using NVIDIA and PyTorch profiling tools.

Outcome: You'll familiarize yourself with

- efficient neural network training techniques (e.g., activation checkpointing, pipelining)
- modern deep learning architectures (e.g., transformer models, neural operators)
- neural network profiling tools
- finding optimal activation checkpointing schedule for NN training under memory constraints
- optimizing multi-GPU training
- developing software for automatic memory-efficient training (based on Rockmate package of Topal team)
- contributing to scientific paper

Team:

During the internship, you'll be supervised by members of the Topal team at INRIA Bordeaux, namely, Julia Gusak, Lionel Eyraud-Dubois, and Olivier Beaumont.

For the last several years members of the Topal team have been working on optimizing the training of neural networks by applying techniques from high performance computing, linear and tensor algebra (please, see papers from ICML'23, ICML'23, IJCAI'22, NeurIPS'21).

Also, you'll have an opportunity to work with Caltech and apply efficient training techniques to different data science applications within the existing collaboration between Topal and Anima AI + Science Laboratory.

For a deeper dive into the efficient deep learning field, do consider attending WANT@NeurIPS 2023, a workshop organized by the Topal team and its collaborators at the upcoming NeurIPS 2023.

Contact: yulia.gusak@inria.fr (Julia Gusak)

Internship Format:

Hybrid (offline & online)