Institut de Recherche
en Informatique de Toulouse

Equipe APO
IRIT-CNRS
ENSEEIHT
2 rue Charles Camichel
31000 TOULOUSE CEDEX

Re-computation and Memory Consumption in Parallelism Models for training Deep Neural Networks

Background

The training phase of Deep Neural Networks (DNNs) has become an important source of computing resource usage that involves memory-intensive operations. Performing the gradient descent via backpropagation produces extensive data (activations) during the execution. The number of layer in the DDN and the batch size on which the DNN is trained is thus often limited by the memory size of the computing architecture. To cope with this problem [1, 2, 3], evicting strategies exist to decide which data to store during the forward phase of the backpropagation and which will be re-computed later in the execution. Many of these evicting strategies remain rudimentary in the implementations (such as periodically saving the activations) but theoretical studies have been conducted to decide which activation to store in order to minimize the amont of re-computations in the sequential model [4, 5].

At the same time distributed deep learning systems [6] have gained significant attention due to their ability to leverage distributed computational resources. Parallelization techniques for DNNs can be separated mainly into two categories :

— Data Parallelism in which the DNN is replicated on several workers. Each workers run the model on a subset of the input batch and have to synchronize between each others to update the model weights. This method suffers from a linear increase in memory as the number of workers increases, while communications costs increase either logarithmically or linearly.

— Model Parallelism in which the layers of the DNN are distributed on the computation ressources. This technique induce high communication cost but allow the input batch to be propagated to the next worker as soon as it has completed its forward pass, allowing workers to compute in parallel via pipelining.

Both these techniques increase the memory consumption when training DNNs making it even more crucial to develop efficient evicting strategies.

Objective of the internship

The primary objective of this internship is to investigate how various parallelism models for training Deep Neural Networks (DNNs) impact the theoretical trade-off between re-computation and memory requirements. The aim is to formalize distributed backpropagation as a memory-aware scheduling problem, which will aid in developing efficient strategies for data eviction.

Key Responsibilities

During the internship, the intern will :

1. Conduct a comprehensive literature review on existing parallelism models for deep neural networks, such as pipelining [1], and explore delayed gradient techniques.
2. Develop task-aware scheduling algorithms aimed at optimizing resource allocation and communication bandwidth in decentralized deep learning systems.
3. Implement and test these algorithms across various decentralized deep learning scenarios, evaluating their effectiveness in improving system efficiency and performance.
4. Document research findings and contribute significantly to the writing of research papers or reports.

Location

The internship will be based in Toulouse and will involve collaboration with Sorbonne University and the French National Centre for Scientific Research (CNRS). This opportunity may potentially lead to a PhD position between both locations.
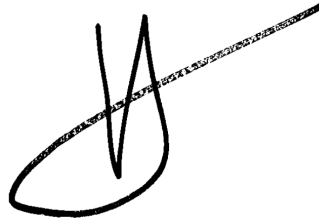
An ideal candidate should :

— Be at the Master 2 level (or equivalent)
— Have a strong background in applied mathematics
— Have strong programming skills in Python and experience with Pytorch
— Be familiar with distributed computing and optimization techniques

Organization

Assignment	:	IRIT-APO team
Supervision	:	Edouard Oyallon (CNRS, Sorbonne University, Flatiron Institute)
		Julien Herrmann (CNRS, IRIT)
Duration	:	5 to 6 months
Stipend	:	approximately €600 per month
Expected Start Date	:	February / March 2024
Location	:	ENSEEIHT-IRIT, 2 rue Claude Camichel, 31000 Toulouse
Contact	:	Julien Herrmann < julien.herrmann@irit.fr >

Toulouse, le 29 décembre 2023,
Julien Herrmann

Références

[1] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, et al. Gpipe : Efficient training of giant neural networks using pipeline parallelism. *Advances in neural information processing systems*, 32, 2019.

[2] Huiping Zhuang, Zhenyu Weng, Fulin Luo, Toh Kar-Ann, Haizhou Li, and Zhiping Lin. Accumulated decoupled learning with gradient staleness mitigation for convolutional neural networks. In *International Conference on Machine Learning*, pages 12935–12944. PMLR, 2021.

[3] Eugene Belilovsky, Michael Eickenberg, and Edouard Oyallon. Decoupled greedy learning of cnns. In *International Conference on Machine Learning*, pages 736–745. PMLR, 2020.

[4] Julien Herrmann and Guillaume Pallez (Aupy). H-revolve : a framework for adjoint computation on synchronous hierarchical platforms. *ACM Transactions on Mathematical Software (TOMS)*, 46(2) :1–25, 2020.

[5] Olivier Beaumont, Julien Herrmann, Guillaume Pallez, and Alena Shilova. Optimal memory-aware backpropagation of deep join networks. *Philosophical Transactions of the Royal Society A*, 378(2166) :20190049, 2020.

[6] Anastasia Koloskova, Tao Lin, Sebastian U Stich, and Martin Jaggi. Decentralized deep learning with arbitrary communication compression. *arXiv preprint arXiv :1907.09356*, 2019.

julien.herrmann@irit.fr