Compas'2023 : Parallélisme / Architecture/ Système
LISTIC/USMB - Annecy France, 4-7 juillet 2023

A New Parallelization Approach in Deep Learning Using CPU/GPU Architectures for Memory Optimization

Luis A. Torres N.

Universidad Industrial de Santander,
CAGE - SC3UIS
680002 Carrera 17 Calle 9 - Colombie
luis.torres@correo.uis.edu.co

Résumé
Deep Learning algorithms are highly relevant in fields such as computer vision and natural language processing. Similarly, the increasing availability of data has improved the models based on Deep Learning and therefore increased their size and efficiency. The combination of these factors has incurred the need to have increased increasingly efficient computational resources, among which are greater processing capacity and memory available for the execution of the training of these models.
As for memory consumption, the training algorithm of Deep Learning models requires a significant amount that could exceed the capacities of the GPU and, in some cases, the CPU memory. New methods have been created to train the model to resolve this issue, such as Model Parallelism, Data Parallelism, and Pipeline Parallelism. However, these methods do not necessarily reduce the memory footprint but distribute memory requirements among devices such as servers, GPUs, and TPUs.
The high processing capacities provided by GPUs and TPUs are limited by the low memory available on these devices. In search of a solution, new techniques have been developed, such as using sparse arrays, pruning the activations of the network layers, and offloading the activations to the CPU memory to reduce the memory footprint of the training process. Despite these new techniques and methods, memory remains the most significant bottleneck present in the training of Deep Learning models, which is why it has become one of the most active research fields. Many researchers have directed their studies to use the CPU as a more active actor in the training process. The new CPU architectures show some important optimizations directly aimed at training neural networks, particularly with its AVX instruction set.
It is proposed a method of parallelization of the Deep Learning training algorithm by distributing the workload between the CPU and the GPU/TPU to reduce the memory footprint present in these devices without loss of accuracy and performance. To achieve this objective, it is proposed to design an asynchronous method to execute the training distributed between the CPU and the GPU/TPU, reducing waiting times during the forward pass and backward pass balancing workloads.
