A Study of Parallel Computing for Machine Learning: Which Platform for which Application?

Benoit Corda1, Clement Farabet1,2, Yann LeCun1
1. Courant Institute, New York University, 715 Broadway, New York
2. Electrical Engineering Department, Yale University, New Haven

Instances of Machine Learning Models share a set of basic operations

In particular for convolutional networks, models are dominated by convolutions and dot-product operations. For general purpose CPUs, several machine learning libraries have been developed, which all provide efficient multi-dimensional arrays and modular description frameworks, this includes LuaL [1], Torch [2] and Theano [3]. Essentially, all these libraries, and more generally all scientific packages (e.g. NumPy, MatLab, ...) rely on compiled code to handle the most expensive operations, and provide a high-level, dynamically-typed scripting language to easily explore and develop new algorithms. Although the theoretical peak performance of modern CPUs is about 10 gigaflops per core, the typical performance obtained on a real convolutional network is typically on the order of 1 gigaflops, mostly due to memory bandwidth limitations.


Problem: the cost of the feature extraction slow down the overall system.

References