High-Performance Communication for Deep Learning

Torsten Hoefler

Computer Science Department, ETH Zürich, Switzerland

One of the main drivers behind the rapid recent advances in machine learning has been the availability of efficient system support. Despite existing progress, scaling compute-intensive machine learning workloads to a large number of compute nodes is still a challenging task. In this talk, we provide an overview of communication aspects in deep learning. We address the communication challenge, by proposing SparCML, a general, scalable communication layer for machine learning applications. SparCML is built on the observation that many distributed machine learning algorithms either have naturally sparse communication patterns, or have updates which can be sparsified in a structured way for improved performance, without loss of convergence or accuracy. To exploit this insight, we analyze, design, and implement a set of communication-efficient protocols for sparse input data, in conjunction with efficient machine learning algorithms which can leverage these primitives. Our communication protocols generalize standard collective operations, by allowing processes to contribute sparse input data vectors, of heterogeneous sizes. Our generic communication layer is enriched with additional features, such as support for non-blocking (asynchronous) operations and support for low-precision data representations. We validate our algorithmic results experimentally on a range of large-scale machine learning applications and target architectures, showing that we can leverage sparsity for order-of-magnitude runtime savings, compared to existing methods and frameworks.