Less chat leads to more work for machine learning

October 15 2019

The team is using deep analysis of parallel computations to accelerate machine learning at scale. Credit: Onur Oymak / Alamy

By deconstructing and analyzing the tried-and-tested methods used in massively parallel computations, a KAUST-led collaboration has developed a groundbreaking framework for efficient parallel computations at scale. The framework has particular relevance for the types of processing needed for optimization in machine learning.

"Parallelizing" an optimization or data processing task allows the task to

be distributed among many computational nodes. Ideally, this would divide the time needed for computation by the number of nodes recruited to the task. However, with parallelization comes the need to pass increasing amounts of information between the nodes, which means the ideal degree of acceleration is never achieved in practice.

"In distributed optimization, a common issue is the communication bottleneck," explains Konstantin Mishchenko from the Visual Computing Center. "Imagine that you had a computer with four cores, and you want to run your parallelized program on a new computer with 16 cores. Naturally, you would expect the new computer to be about four times faster. But, even though the new computer has four times the total computing power, much of it is taken up by synchronizing the cores at each model update. This communication bottleneck reduces the positive effect of increasing the number of cores and becomes severe when we scale the number of cores to hundreds or thousands."

Recent research by Peter Richtárik's group has addressed this problem in two ways—by improving the compression of information passed at each synchronization and by generalizing the learning algorithm so that it can be used with any compression scheme.

"The hardest thing to understand was why existing ideas always work," says Mishchenko. "Commonly, researchers first guess what trick needs to be used, and only later do we start understanding why it works. This is exactly what we did: by using simple counterexamples, we reanalyzed two well-known tricks and came to the realization that there is a better way to use them."

Those techniques, called quantization and random sparsification, are compression methods that are typically used in isolation. By combining both, and crucially, only compressing the difference between new information and the previous update, the team proved mathematically

that a more efficient compression scheme is possible with less loss of information.

"The most important point is that this new technique, where we compress the difference between current and previous information—and not just the new information itself—ensures that less information is lost when we perform a compression," says Mishchenko. "And we have proved and observed in experiments that scaling using our method is closer to the ideal."

The other finding generalizes the learning algorithm for a range of different optimization tasks in a way that allows it to be used with any compression scheme.

"Our motivation was to create a general theory that does not rely on any specific compression scheme in order to understand the effects of compression on distributed training," says Samuel Horvath from the research team.

Using this theory makes it possible to construct algorithms for distributed computation without the problems of incomplete optimization and dependence on specific compression schemes faced by existing methods.

"This work helps us to better understand the effects of different compression methods and helps us choose the right compression scheme for the given problem," says Horvath.

More information: Horvath, S., Kovalev, D., Mishchenko, K., Richtárik, P. & Stich, S. U. Stochastic distributed learning with gradient quantization and variance reduction, arXiv:1904.05115 (2019). arxiv.org/abs/1904.05115

Mishchenko, K., Hanzely, F. & Richtárik, P. 99% of parallel optimization is inevitably a waste of time, arxiv.org/abs/1901.09437 (2019). arxiv.org/abs/1901.09437

Mishchenko, K., Gorbunov, E., Takac, M. & Richtárik, P. Distributed learning with compressed gradient differences, arxiv.org/abs/1901.09269 (2019). arxiv.org/abs/1901.09269