CPU algorithm trains deep neural nets up to 15 times faster than top GPU trainers

April 7 2021, by Jade Boyd

Anshumali Shrivastava is an assistant professor of computer science at Rice University. Credit: Jeff Fitlow/Rice University

Rice University computer scientists have demonstrated artificial intelligence (AI) software that runs on commodity processors and trains

deep neural networks 15 times faster than platforms based on graphics processors.

"The cost of training is the actual bottleneck in AI," said Anshumali Shrivastava, an assistant professor of computer science at Rice's Brown School of Engineering. "Companies are spending millions of dollars a week just to train and fine-tune their AI workloads."

Shrivastava and collaborators from Rice and Intel will present research that addresses that bottleneck April 8 at the machine learning systems conference [MLSys](#).

Deep [neural networks](#) (DNN) are a powerful form of [artificial intelligence](#) that can outperform humans at some tasks. DNN training is typically a series of [matrix multiplication](#) operations, an ideal workload for graphics processing units (GPUs), which cost about three times more than general purpose central processing units (CPUs).

"The whole industry is fixated on one kind of improvement—faster matrix multiplications," Shrivastava said. "Everyone is looking at specialized hardware and architectures to push matrix multiplication. People are now even talking about having specialized hardware-software stacks for specific kinds of deep learning. Instead of taking an expensive algorithm and throwing the whole world of system optimization at it, I'm saying, 'Let's revisit the algorithm.'"

Shrivastava's lab did that in 2019, recasting DNN training as a search problem that could be solved with hash tables. Their "sub-linear deep learning engine" (SLIDE) is specifically designed to run on commodity CPUs, and Shrivastava and collaborators from Intel showed it could outperform GPU-based training when they [unveiled it at MLSys 2020](#).

The study [they'll present this week at MLSys 2021](#) explored whether

SLIDE's performance could be improved with vectorization and memory optimization accelerators in modern CPUs.

"Hash table-based acceleration already outperforms GPU, but CPUs are also evolving," said study co-author Shabnam Daghaghi, a Rice graduate student. "We leveraged those innovations to take SLIDE even further, showing that if you aren't fixated on matrix multiplications, you can leverage the power in modern CPUs and train AI models four to 15 times faster than the best specialized hardware alternative."

Study co-author Nicholas Meisburger, a Rice undergraduate, said "CPUs are still the most prevalent hardware in computing. The benefits of making them more appealing for AI workloads cannot be understated."

Provided by Rice University