TechXplore

Deep learning rethink overcomes major obstacle in AI industry

March 2 2020

Rice University's Anshumali Shrivastava led a team that demonstrated how to implement deep learning technology without specialized acceleration hardware like graphics processing units. Credit: Jeff Fitlow/Rice University

Rice University computer scientists have overcome a major obstacle in the burgeoning artificial intelligence industry by showing it is possible to speed up deep learning technology without specialized acceleration hardware like graphics processing units (GPUs).

Computer scientists from Rice, supported by collaborators from Intel, will present their results today at the Austin Convention Center as a part of the machine learning systems conference MLSys.

Many companies are investing heavily in GPUs and other specialized hardware to implement deep learning, a powerful form of artificial intelligence that's behind digital assistants like Alexa and Siri, facial recognition, product recommendation systems and other technologies. For example, Nvidia, the maker of the industry's gold-standard Tesla V100 Tensor Core GPUs, recently reported a 41% increase in its fourth quarter revenues compared with the previous year.

Rice researchers created a cost-saving alternative to GPU, an algorithm called "sub-linear deep learning engine" (SLIDE) that uses general purpose central processing units (CPUs) without specialized acceleration hardware.

"Our tests show that SLIDE is the first smart algorithmic implementation of deep learning on CPU that can outperform GPU hardware acceleration on industry-scale recommendation datasets with large fully connected architectures," said Anshumali Shrivastava, an assistant professor in Rice's Brown School of Engineering who invented SLIDE with graduate students Beidi Chen and Tharun Medini.

SLIDE doesn't need GPUs because it takes a fundamentally different approach to deep learning. The standard "back-propagation" training technique for deep neural networks requires matrix multiplication, an ideal workload for GPUs. With SLIDE, Shrivastava, Chen and Medini turned neural network training into a search problem that could instead be solved with hash tables.

This radically reduces the computational overhead for SLIDE compared to back-propagation training. For example, a top-of-the-line GPU platform like the ones Amazon, Google and others offer for cloud-based deep learning services has eight Tesla V100s and costs about $100,000, Shrivastava said.

We have one in the lab, and in our test case we took a workload that's perfect for V100, one with more than 100 million parameters in large, fully connected networks that fit in GPU memory," he said. "We trained it with the best (software) package out there, Google's TensorFlow, and it took 3 1/2 hours to train.

"We then showed that our new algorithm can do the training in one hour, not on GPUs but on a 44-core Xeon-class CPU," Shrivastava said.

Deep learning networks were inspired by biology, and their central feature, artificial neurons, are small pieces of computer code that can learn to perform a specific task. A deep learning network can contain millions or even billions of artificial neurons, and working together they can learn to make human-level, expert decisions simply by studying large amounts of data. For example, if a deep neural network is trained to identify objects in photos, it will employ different neurons to recognize a photo of a cat than it will to recognize a school bus.

"You don't need to train all the neurons on every case," Medini said. "We thought, 'If we only want to pick the neurons that are relevant, then it's a search problem.' So, algorithmically, the idea was to use locality-sensitive hashing to get away from matrix multiplication."

Hashing is a data-indexing method invented for internet search in the 1990s. It uses numerical methods to encode large amounts of information, like entire webpages or chapters of a book, as a string of digits called a hash. Hash tables are lists of hashes that can be searched very quickly.

"It would have made no sense to implement our algorithm on TensorFlow or PyTorch because the first thing they want to do is convert whatever you're doing into a matrix multiplication problem," Chen said. "That is precisely what we wanted to get away from. So we wrote our own C++ code from scratch."

Shrivastava said SLIDE's biggest advantage over back-propagation is that it is data parallel.

"By data parallel I mean that if I have two data instances that I want to train on, let's say one is an image of a cat and the other of a bus, they will likely activate different neurons, and SLIDE can update, or train on these two independently," he said. "This is much a better utilization of parallelism for CPUs.

"The flipside, compared to GPU, is that we require a big memory," he said. "There is a cache hierarchy in main memory, and if you're not careful with it you can run into a problem called cache thrashing, where you get a lot of cache misses."

Shrivastava said his group's first experiments with SLIDE produced significant cache thrashing, but their training times were still comparable to or faster than GPU training times. So he, Chen and Medini published the initial results on *arXiv* in March 2019 and uploaded their code to GitHub. A few weeks later, they were contacted by Intel.

"Our collaborators from Intel recognized the caching problem," he said. "They told us they could work with us to make it train even faster, and they were right. Our results improved by about 50% with their help."

Shrivastava said SLIDE hasn't yet come close to reaching its potential.

"We've just scratched the surface," he said. "There's a lot we can still do to optimize. We have not used vectorization, for example, or built-in accelerators in the CPU, like Intel Deep Learning Boost. There are a lot of other tricks we could still use to make this even faster."

Shrivastava said SLIDE is important because it shows there are other ways to implement deep learning.

"The whole message is, 'Let's not be bottlenecked by multiplication matrix and GPU memory,'" Chen said. "Ours may be the first algorithmic approach to beat GPU, but I hope it's not the last. The field needs new ideas, and that is a big part of what MLSys is about."

More information: SLIDE : In Defense of Smart Algorithms over Hardware Acceleration for Large-Scale Deep Learning Systems, arXiv:1903.03129 [cs.DC] arxiv.org/abs/1903.03129

Provided by Rice University

Citation: Deep learning rethink overcomes major obstacle in AI industry (2020, March 2) retrieved 25 April 2024 from https://techxplore.com/news/2020-03-deep-rethink-major-obstacle-ai.html