Binarized neural networks show promise for fast, accurate machine learning

24 November 2020, by Allan Brettman

Removing bits and pieces along coding branches in machine learning algorithms can reduce complexity in decision trees and increase predictive performance. Credit: Nathan Johnson | Pacific Northwest National Laboratory

As anyone with a green thumb knows, pruning can promote thriving vegetation. A snip here, a snip there, and growth can be controlled and directed for a more vigorous plant.

The same principle can be applied to machine learning algorithms. Removing bits and pieces along coding branches in those algorithms can reduce complexity in decision trees and increase predictive performance.

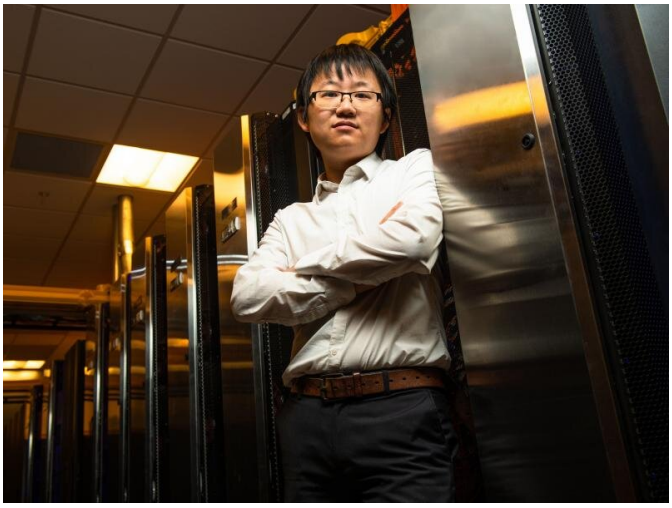
Researchers at the U.S. Department of Energy's Pacific Northwest National Laboratory (PNNL) have done just that. Exploring with binarized neural networks (BNNs), they used pruning principles to significantly reduce computation complexity and memory demands. BNNs are close cousins to deep neural networks, which require large amounts of computation. But BNNs differ in a significant way: they use single bits to encode each neuron and parameter, using much less energy and power for computation.

Pruning for faster growth

Researchers recognized the potential value of BNNs for machine learning starting in about 2016. If constructed—or pruned—just the right way, they consume less computing energy and are nearly as accurate as deep neural networks. That means BNNs have more potential to benefit resource-constrained environments, such as mobile phones, smart devices, and the entire Internet of Things ecosystem.

This is where pruning comes into play. As neural networks research has grown in recent years, pruning has gained more interest among computing researchers.

"Pruning is currently a hot topic in machine learning," said PNNL computer scientist Ang Li. "We can add software and architecture coding to push the trimming towards a direction that will have more benefits for the performance of computing devices. These benefits include lower energy needs and lower computing costs."



As neural networks research has grown in recent years, pruning has gained more interest among computing researchers, according to PNNL computer scientist Ang Li. Credit: Andrea Starr | Pacific Northwest National Laboratory

Pruning for precision

Li was among a group of PNNL researchers who recently published results in the Institute of Electrical and Electronics Engineers Transactions on Parallel and Distributed Systems showing the benefits of selective pruning. The research demonstrated that pruning redundant bits of the BNN architecture led to a custom-built out-of-order BNN, called O3BNN-R. Their work shows a highly-condensed BNN model—which already could display high-performing supercomputing qualities—can be shrunk signi?cantly further without loss of accuracy.

"Binarized neural networks have the potential of making the processing time of neural networks around microseconds," said Tong "Tony" Geng, a Boston University doctoral candidate who, as a PNNL intern, assisted Li on the O3BNN-R project.

"BNN research is headed in a promising direction to make neural networks really useful and be readily adopted in the real-world," said Geng, who will rejoin the PNNL staff in January as a postdoctoral research fellow. "Our finding is an important step to realize this potential."

Their research shows this out-of-order BNN can prune, on average, 30 percent of operations without any accuracy loss. With even more fine tuning—in a step called "regularization at training"—the performance can be improved an additional 15 percent.

Pruning for power

In addition to this out-of-order BNN's contributions to the Internet of Things, Li also pointed to potential benefits to the energy grid. Implementation of a modified BNN could also provide a boost to existing software that guards against cyberattacks when deployed in the power grid by helping existing sensors detect and respond to an attack, said Li.

"Basically," said Li, "we are accelerating the speed of processing in hardware."

More information: Tong Geng et al. O3BNN-R: An Out-of-Order Architecture for High-Performance and Regularized BNN Inference, *IEEE Transactions on Parallel and Distributed Systems* (2020). DOI: 10.1109/TPDS.2020.3013637

Provided by Pacific Northwest National Laboratory

APA citation: Binarized neural networks show promise for fast, accurate machine learning (2020, November 24) retrieved 20 August 2022 from https://techxplore.com/news/2020-11-binarized-neural-networks-fast-accurate.html