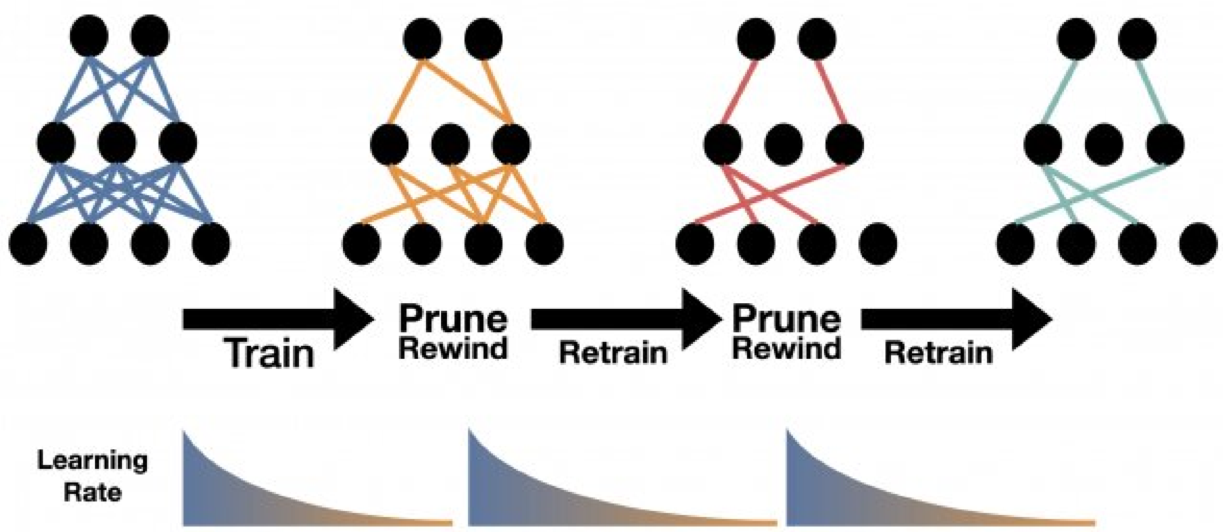
TechXplore

Researchers unveil a pruning algorithm to make artificial intelligence applications run faster

April 30 2020, by Kim Martineau



MIT researchers have proposed a technique for shrinking deep learning models that they say is simpler and produces more accurate results than state-of-the-art methods. It works by retraining the smaller, pruned model at its faster, initial learning rate. Credit: Alex Renda

As more artificial intelligence applications move to smartphones, deep learning models are getting smaller to allow apps to run faster and save battery power. Now, MIT researchers have a new and better way to compress models.

It's so simple that they unveiled it in a tweet last month: Train the model, prune its weakest connections, retrain the model at its fast, early training rate, and repeat, until the model is as tiny as you want.

"That's it," says Alex Renda, a Ph.D. student at MIT. "The standard things people do to prune their models are crazy complicated."

Renda discussed the technique when the International Conference of Learning Representations (ICLR) convened remotely this month. Renda is a co-author of the work with Jonathan Frankle, a fellow Ph.D. student in MIT's Department of Electrical Engineering and Computer Science (EECS), and Michael Carbin, an assistant professor of electrical engineering and computer science—all members of the Computer Science and Artificial Science Laboratory.

The search for a better compression technique grew out of Frankle and Carbin's award-winning Lottery Ticket Hypothesis paper at ICLR last year. They showed that a deep neural network could perform with only one-tenth the number of connections if the right subnetwork was found early in training. Their revelation came as demand for computing power and energy to train ever larger deep learning models was increasing exponentially, a trend that continues to this day. Costs of that growth include a rise in planet-warming carbon emissions and a potential drop in innovation as researchers not affiliated with big tech companies compete for scarce computing resources. Everyday users are affected, too. Big AI models eat up mobile-phone bandwidth and battery power.

The Lottery Ticket Hypothesis triggered a series of mostly theoretical follow-on papers. But at a colleague's suggestion, Frankle decided to see what lessons it might hold for pruning, in which a search algorithm trims the number of nodes evaluated in a search tree. The field had been around for decades, but saw a resurgence after the breakout success of neural networks at classifying images in the ImageNet competition. As models got bigger, with researchers adding on layers of artificial neurons to boost performance, others proposed techniques for whittling them down.

Song Han, now an assistant professor at MIT, was one pioneer. Building on a series of influential papers, Han unveiled a pruning algorithm he called AMC, or AutoML for model compression, that's still the industry standard. Under Han's technique, redundant neurons and connections are automatically removed, and the model is retrained to restore its initial accuracy.

In response to Han's work, Frankle recently suggested in an unpublished paper that results could be further improved by rewinding the smaller, pruned model to its initial parameters, or weights, and retraining the smaller model at its faster, initial rate.

In the current ICLR study, the researchers realized that the model could simply be rewound to its early training rate without fiddling with any parameters. In any pruning regimen, the tinier a model gets, the less accurate it becomes. But when the researchers compared this new method to Han's AMC or Frankle's weight-rewinding methods, it performed better no matter how much the model shrank.

It's unclear why the pruning technique works as well as it does. The researchers say they will leave that question for others to answer. As for those who wish to try it, the algorithm is as easy to implement as other pruning methods, without time-consuming tuning, the researchers say.

"It's the pruning algorithm from the 'Book,'" says Frankle. "It's clear, generic, and drop-dead simple."

Han, for his part, has now partly shifted focus from compression AI models to channeling AI to design small, efficient models from the start. His newest method, Once for All, also debuts at ICLR. Of the new learning rate method, he says: "I'm happy to see new pruning and retraining techniques evolve, giving more people access to high-performing AI applications."

More information: Comparing Rewinding and Fine-tuning in Neural Network Pruning, arXiv:2003.02389 [cs.LG] arxiv.org/abs/2003.02389

Provided by Massachusetts Institute of Technology

Citation: Researchers unveil a pruning algorithm to make artificial intelligence applications run faster (2020, April 30) retrieved 25 April 2024 from https://techxplore.com/news/2020-04-unveil-pruning-algorithm-artificial-intelligence.html