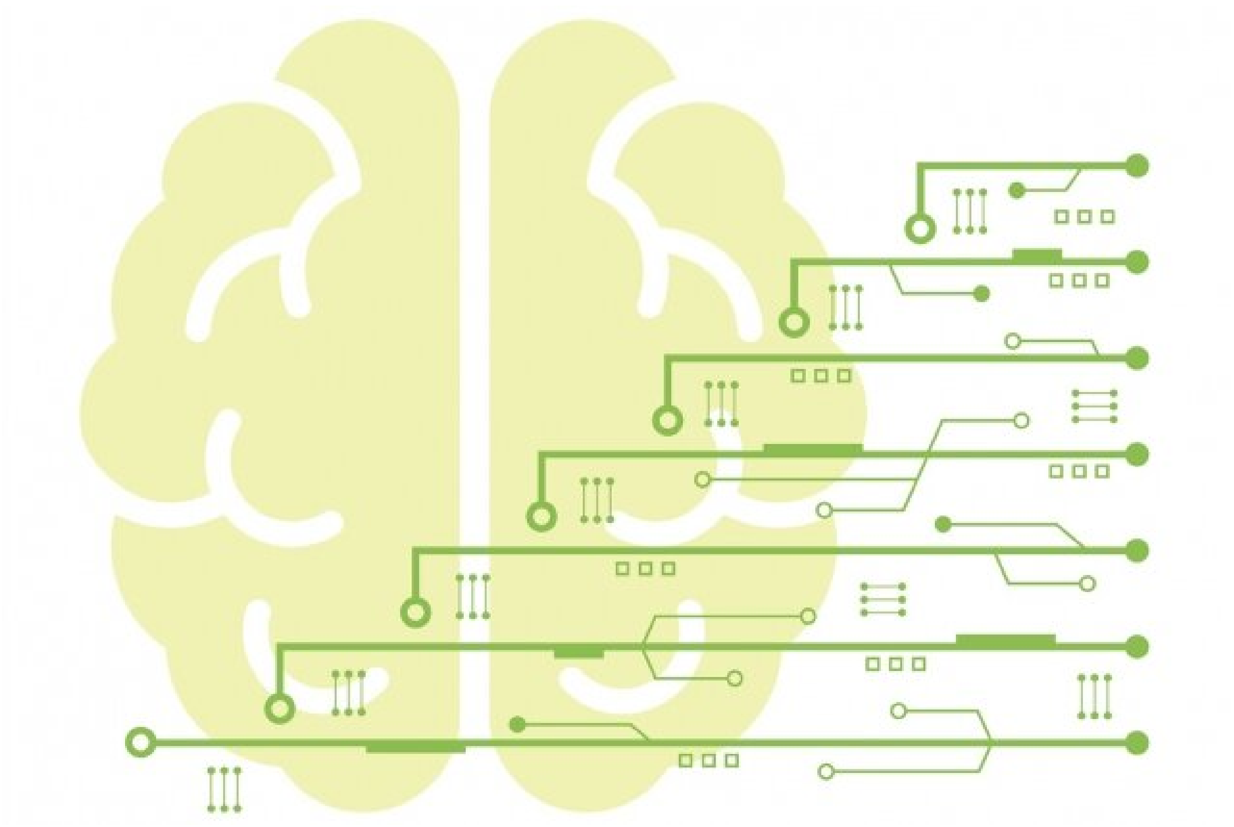
Shrinking deep learning's carbon footprint

August 10 2020, by Kim Martineau



Deep learning has driven much of the recent progress in artificial intelligence, but as demand for computation and energy to train ever-larger models increases, many are raising concerns about the financial and environmental costs. To address the problem, researchers at MIT and the MIT-IBM Watson AI Lab are experimenting with ways to make software and hardware more energy efficient, and in some cases, more like the human brain. Credit: Niki Hinkle/MIT Spectrum

In June, OpenAI unveiled the largest language model in the world, a text-generating tool called GPT-3 that can write creative fiction, translate legalese into plain English, and answer obscure trivia questions. It's the latest feat of intelligence achieved by deep learning, a machine learning method patterned after the way neurons in the brain process and store information.

But it came at a hefty price: at least $4.6 million and 355 years in computing time, assuming the model was trained on a standard neural network chip, or GPU. The model's colossal size—1,000 times larger than a typical language model—is the main factor in its high cost.

"You have to throw a lot more computation at something to get a little improvement in performance," says Neil Thompson, an MIT researcher who has tracked deep learning's unquenchable thirst for computing. "It's unsustainable. We have to find more efficient ways to scale deep learning or develop other technologies."

Some of the excitement over AI's recent progress has shifted to alarm. In a study last year, researchers at the University of Massachusetts at Amherst estimated that training a large deep-learning model produces 626,000 pounds of planet-warming carbon dioxide, equal to the lifetime emissions of five cars. As models grow bigger, their demand for computing is outpacing improvements in hardware efficiency. Chips specialized for neural-network processing, like GPUs (graphics processing units) and TPUs (tensor processing units), have offset the demand for more computing, but not by enough.

"We need to rethink the entire stack—from software to hardware," says Aude Oliva, MIT director of the MIT-IBM Watson AI Lab and co-director of the MIT Quest for Intelligence. "Deep learning has made the recent AI revolution possible, but its growing cost in energy and carbon emissions is untenable."

Computational limits have dogged neural networks from their earliest incarnation—the perceptron—in the 1950s. As computing power exploded, and the internet unleashed a tsunami of data, they evolved into powerful engines for pattern recognition and prediction. But each new milestone brought an explosion in cost, as data-hungry models demanded increased computation. GPT-3, for example, trained on half a trillion words and ballooned to 175 billion parameters—the mathematical operations, or weights, that tie the model together—making it 100 times bigger than its predecessor, itself just a year old.

In work posted on the pre-print server arXiv, Thompson and his colleagues show that the ability of deep learning models to surpass key benchmarks tracks their nearly exponential rise in computing power use. (Like others seeking to track AI's carbon footprint, the team had to guess at many models' energy consumption due to a lack of reporting requirements). At this rate, the researchers argue, deep nets will survive only if they, and the hardware they run on, become radically more efficient.

Toward leaner, greener algorithms

The human perceptual system is extremely efficient at using data. Researchers have borrowed this idea for recognizing actions in video and in real life to make models more compact. In a paper at the European Conference on Computer Vision (ECCV) in August, researchers at the MIT-IBM Watson AI Lab describe a method for unpacking a scene from a few glances, as humans do, by cherry-picking the most relevant data.

Take a video clip of someone making a sandwich. Under the method outlined in the paper, a policy network strategically picks frames of the knife slicing through roast beef, and meat being stacked on a slice of bread, to represent at high resolution. Less-relevant frames are skipped over or represented at lower resolution. A second model then uses the abbreviated CliffsNotes version of the movie to label it "making a sandwich." The approach leads to faster video classification at half the computational cost as the next-best model, the researchers say.

"Humans don't pay attention to every last detail—why should our models?" says the study's senior author, Rogerio Feris, research manager at the MIT-IBM Watson AI Lab. "We can use machine learning to adaptively select the right data, at the right level of detail, to make deep learning models more efficient."

In a complementary approach, researchers are using deep learning itself to design more economical models through an automated process known as neural architecture search. Song Han, an assistant professor at MIT, has used automated search to design models with fewer weights, for language understanding and scene recognition, where picking out looming obstacles quickly is acutely important in driving applications.

In a paper at ECCV, Han and his colleagues propose a model architecture for three-dimensional scene recognition that can spot safety-critical details like road signs, pedestrians, and cyclists with relatively less computation. They used an evolutionary-search algorithm to evaluate 1,000 architectures before settling on a model they say is three times faster and uses eight times less computation than the next-best method.

In another recent paper, they use evolutionary search within an augmented designed space to find the most efficient architectures for machine translation on a specific device, be it a GPU, smartphone, or tiny Raspberry Pi. Separating the search and training process leads to huge reductions in computation, they say.

In a third approach, researchers are probing the essence of deep nets to see if it might be possible to train a small part of even hyper-efficient networks like those above. Under their proposed lottery ticket hypothesis, Ph.D. student Jonathan Frankle and MIT Professor Michael Carbin proposed that within each model lies a tiny subnetwork that could have been trained in isolation with as few as one-tenth as many weights—what they call a "winning ticket."

They showed that an algorithm could retroactively find these winning subnetworks in small image-classification models. Now, in a paper at the International Conference on Machine Learning (ICML), they show that the algorithm finds winning tickets in large models, too; the models just need to be rewound to an early, critical point in training when the order of the training data no longer influences the training outcome.

In less than two years, the lottery ticket idea has been cited more than more than 400 times, including by Facebook researcher Ari Morcos, who has shown that winning tickets can be transferred from one vision task to another, and that winning tickets exist in language and reinforcement learning models, too.

"The standard explanation for why we need such large networks is that overparameterization aids the learning process," says Morcos. "The lottery ticket hypothesis disproves that—it's all about finding an appropriate starting point. The big downside, of course, is that, currently, finding these 'winning' starting points requires training the full overparameterized network anyway."

Frankle says he's hopeful that an efficient way to find winning tickets will be found. In the meantime, recycling those winning tickets, as Morcos suggests, could lead to big savings.

Hardware designed for efficient deep net algorithms

As deep nets push classical computers to the limit, researchers are pursuing alternatives, from optical computers that transmit and store data with photons instead of electrons, to quantum computers, which have the potential to increase computing power exponentially by representing data in multiple states at once.

Until a new paradigm emerges, researchers have focused on adapting the modern chip to the demands of deep learning. The trend began with the discovery that video-game graphical chips, or GPUs, could turbocharge deep-net training with their ability to perform massively parallelized matrix computations. GPUs are now one of the workhorses of modern AI, and have spawned new ideas for boosting deep net efficiency through specialized hardware.

Much of this work hinges on finding ways to store and reuse data locally, across the chip's processing cores, rather than waste time and energy shuttling data to and from a designated memory site. Processing data locally not only speeds up model training but improves inference, allowing deep learning applications to run more smoothly on smartphones and other mobile devices.

Vivienne Sze, a professor at MIT, has literally written the book on efficient deep nets. In collaboration with book co-author Joel Emer, an MIT professor and researcher at NVIDIA, Sze has designed a chip that's flexible enough to process the widely-varying shapes of both large and small deep learning models. Called Eyeriss 2, the chip uses 10 times less energy than a mobile GPU.

Its versatility lies in its on-chip network, called a hierarchical mesh, that adaptively reuses data and adjusts to the bandwidth requirements of different deep learning models. After reading from memory, it reuses the data across as many processing elements as possible to minimize data transportation costs and maintain high throughput.

"The goal is to translate small and sparse networks into energy savings and fast inference," says Sze. "But the hardware should be flexible enough to also efficiently support large and dense deep neural networks."

Other hardware innovators are focused on reproducing the brain's energy efficiency. Former Go world champion Lee Sedol may have lost his title to a computer, but his performance was fueled by a mere 20 watts of power. AlphaGo, by contrast, burned an estimated megawatt of energy, or 500,000 times more.

Inspired by the brain's frugality, researchers are experimenting with replacing the binary, on-off switch of classical transistors with analog devices that mimic the way that synapses in the brain grow stronger and weaker during learning and forgetting.

An electrochemical device, developed at MIT and recently published in *Nature Communications*, is modeled after the way resistance between two neurons grows or subsides as calcium, magnesium or potassium ions flow across the synaptic membrane dividing them. The device uses the flow of protons—the smallest and fastest ion in solid state—into and out of a crystalline lattice of tungsten trioxide to tune its resistance along a continuum, in an analog fashion.

"Even though is not yet optimized, it gets to the order of energy consumption per unit area per unit change in conductance that's close to that in the brain," says the study's senior author, Bilge Yildiz, a professor at MIT.

Energy-efficient algorithms and hardware can shrink AI's environmental impact. But there are other reasons to innovate, says Sze, listing them off: Efficiency will allow computing to move from data centers to edge devices like smartphones, making AI accessible to more people around the world; shifting computation from the cloud to personal devices reduces the flow, and potential leakage, of sensitive data; and processing data on the edge eliminates transmission costs, leading to faster inference with a shorter reaction time, which is key for interactive driving and augmented/virtual reality applications.

"For all of these reasons, we need to embrace efficient AI," she says.

More information: The Computational Limits of Deep Learning. arXiv:2007.05558v1 , arxiv.org/pdf/2007.05558.pdf

Xiahui Yao et al. Protonic solid-state electrochemical synapse for physical neural networks, *Nature Communications* (2020). DOI: 10.1038/s41467-020-16866-6

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology

Citation: Shrinking deep learning's carbon footprint (2020, August 10) retrieved 13 March 2024 from https://techxplore.com/news/2020-08-deep-carbon-footprint.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.