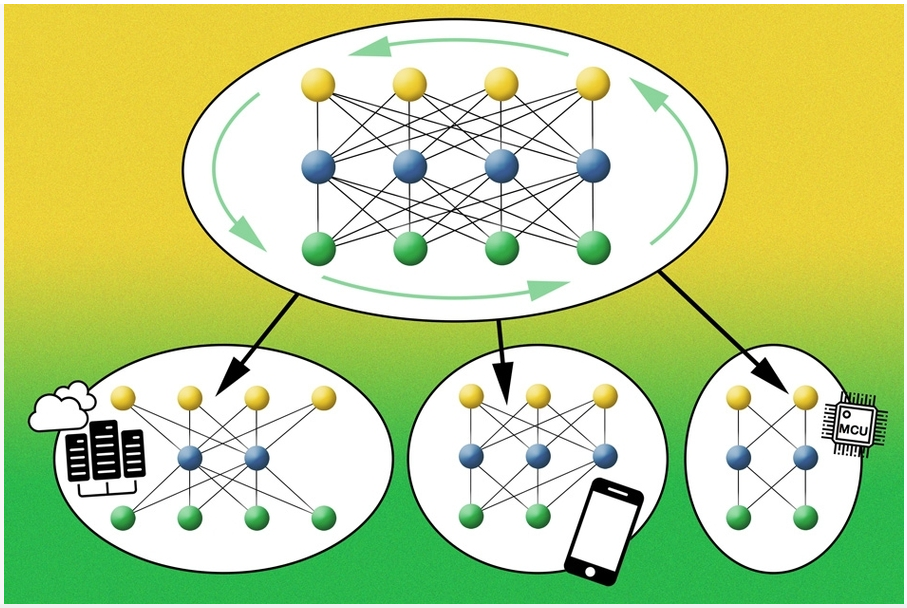
MIT News
ON CAMPUS AND AROUND THE WORLD

Reducing the carbon footprint of artificial intelligence

MIT system cuts the energy required for training and running neural networks.

Rob Matheson | MIT News Office
April 23, 2020



MIT researchers have developed a new automated AI system with improved computational efficiency and a much smaller carbon footprint. The researchers' system trains one large neural network comprising many pretrained subnetworks of different sizes that can be tailored to diverse hardware platforms without retraining.

Image: MIT News, based on figures courtesy of the researchers

Artificial intelligence has become a focus of certain ethical concerns, but it also has some major sustainability issues.

Last June, researchers at the University of Massachusetts at Amherst released a startling report estimating that the amount of power required for training and searching a certain neural network architecture involves the emissions of roughly 626,000 pounds of carbon dioxide. That's equivalent to nearly five times the lifetime emissions of the average U.S. car, including its manufacturing.

This issue gets even more severe in the model deployment phase, where deep neural networks need to be deployed on diverse hardware platforms, each with different properties and computational resources.

MIT researchers have developed a new automated AI system for training and running certain neural networks. Results indicate that, by improving the computational efficiency of the system in some key ways, the system can cut down the pounds of carbon emissions involved — in some cases, down to low triple digits.

The researchers' system, which they call a once-for-all network, trains one large neural network comprising many pretrained subnetworks of different sizes that can be tailored to diverse hardware platforms without retraining. This dramatically reduces the energy usually required to train each specialized neural network for new platforms — which can include billions of internet of things (IoT) devices. Using the system to train a computer-vision model, they estimated that the process required roughly 1/1,300 the carbon emissions compared to today's state-of-the-art neural architecture search approaches, while reducing the inference time by 1.5-2.6 times.

"The aim is smaller, greener neural networks," says Song Han, an assistant professor in the Department of Electrical Engineering and Computer Science.

“Searching efficient neural network architectures has until now had a huge carbon footprint. But we reduced that footprint by orders of magnitude with these new methods.”

The work was carried out on Satori, an efficient computing cluster donated to MIT by IBM that is capable of performing 2 quadrillion calculations per second. The paper is being presented next week at the International Conference on Learning Representations. Joining Han on the paper are four undergraduate and graduate students from EECS, MIT-IBM Watson AI Lab, and Shanghai Jiao Tong University.

Creating a “once-for-all” network

The researchers built the system on a recent AI advance called AutoML (for automatic machine learning), which eliminates manual network design. Neural networks automatically search massive design spaces for network architectures tailored, for instance, to specific hardware platforms. But there’s still a training efficiency issue: Each model has to be selected then trained from scratch for its platform architecture.

“How do we train all those networks efficiently for such a broad spectrum of devices — from a $10 IoT device to a $600 smartphone? Given the diversity of IoT devices, the computation cost of neural architecture search will explode,” Han says.

The researchers invented an AutoML system that trains only a single, large “once-for-all” (OFA) network that serves as a “mother” network, nesting an extremely high number of subnetworks that are sparsely activated from the mother network. OFA shares all its learned weights with all subnetworks — meaning they come essentially pretrained. Thus, each subnetwork can operate independently at inference time without retraining.

The team trained an OFA convolutional neural network (CNN) — commonly used for image-processing tasks — with versatile architectural configurations, including different numbers of layers and "neurons," diverse filter sizes, and diverse input image resolutions. Given a specific platform, the system uses the OFA as the search space to find the best subnetwork based on the accuracy and latency tradeoffs that correlate to the platform's power and speed limits. For an IoT device, for instance, the system will find a smaller subnetwork. For smartphones, it will select larger subnetworks, but with different structures depending on individual battery lifetimes and computation resources. OFA decouples model training and architecture search, and spreads the one-time training cost across many inference hardware platforms and resource constraints.

This relies on a "progressive shrinking" algorithm that efficiently trains the OFA network to support all of the subnetworks simultaneously. It starts with training the full network with the maximum size, then progressively shrinks the sizes of the network to include smaller subnetworks. Smaller subnetworks are trained with the help of large subnetworks to grow together. In the end, all of the subnetworks with different sizes are supported, allowing fast specialization based on the platform's power and speed limits. It supports many hardware devices with zero training cost when adding a new device.

In total, one OFA, the researchers found, can comprise more than 10 quintillion — that's a 1 followed by 19 zeroes — architectural settings, covering probably all platforms ever needed. But training the OFA and searching it ends up being far more efficient than spending hours training each neural network per platform. Moreover, OFA does not compromise accuracy or inference efficiency. Instead, it provides state-of-the-art ImageNet accuracy on mobile devices. And, compared with state-of-the-art industry-leading CNN models , the researchers say OFA provides 1.5-2.6 times speedup, with superior accuracy.

“That’s a breakthrough technology,” Han says. “If we want to run powerful AI on consumer devices, we have to figure out how to shrink AI down to size.”

“The model is really compact. I am very excited to see OFA can keep pushing the boundary of efficient deep learning on edge devices,” says Chuang Gan, a researcher at the MIT-IBM Watson AI Lab and co-author of the paper.

“If rapid progress in AI is to continue, we need to reduce its environmental impact,” says John Cohn, an IBM fellow and member of the MIT-IBM Watson AI Lab. “The upside of developing methods to make AI models smaller and more efficient is that the models may also perform better.”