Massachusetts Institute of Technology

MIT News

ON CAMPUS AND AROUND THE WORLD

February 11, 2020

Brainstorming energy-saving hacks on Satori, MIT's new supercomputer

Three-day hackathon explores methods for making artificial intelligence faster and more sustainable.

Kim Martineau | MIT Quest for Intelligence

Mohammad Haft-Javaherian planned to spend an hour at the Green AI Hackathon — just long enough to get acquainted with MIT's new supercomputer, Satori. Three days later, he walked away with $1,000 for his winning strategy to shrink the carbon footprint of artificial intelligence models trained to detect heart disease.

"I never thought about the kilowatt-hours I was using," he says. "But this hackathon gave me a chance to look at my carbon footprint and find ways to trade a small amount of model accuracy for big energy savings."

Haft-Javaherian was among six teams to earn prizes at a hackathon co-sponsored by the MIT Research Computing Project and MIT-IBM Watson AI Lab Jan. 28-30. The event was meant to familiarize students with Satori, the computing cluster IBM donated to MIT last year, and to inspire new techniques for building energy-efficient AI models that put less planet-warming carbon dioxide into the air.

The event was also a celebration of Satori's green-computing credentials. With an architecture designed to minimize the transfer of data, among other energy-saving features, Satori recently earned fourth place on the Green500 list of supercomputers. Its location gives it additional credibility: It sits on a remediated brownfield site in Holyoke, Massachusetts, now the Massachusetts Green High Performance Computing Center, which runs largely on low-carbon hydro, wind and nuclear power.

Several dozen students participated in the Green AI Hackathon, co-sponsored by the MIT Research Computing Project and MIT-IBM Watson AI Lab.

A postdoc at MIT and Harvard Medical School, Haft-Javaherian came to the hackathon to learn more about Satori. He stayed for the challenge of trying to cut the energy intensity of his own work, focused on developing AI methods to screen the coronary arteries for disease. A new imaging method, optical coherence tomography, has given cardiologists a new tool for visualizing defects in the artery walls that can slow the flow of oxygenated blood to the heart. But even the experts can miss subtle patterns that computers excel at detecting.

MIT LINCOLN LABORATORY SUPERCOMPUTING CENTER

Brainstorming energy-saving hacks on Satori, MIT's new supercomputer (continued)

At the hackathon, Haft-Javaherian ran a test on his model and saw that he could cut its energy use eight-fold by reducing the time Satori's graphics processors sat idle. He also experimented with adjusting the model's number of layers and features, trading varying degrees of accuracy for lower energy use.

A second team, Alex Andonian and Camilo Fosco, also won $1,000 by showing they could train a classification model nearly 10 times faster by optimizing their code and losing a small bit of accuracy. Graduate students in the Department of Electrical Engineering and Computer Science (EECS), Andonian and Fosco are currently training a classifier to tell legitimate videos from AI-manipulated fakes, to compete in Facebook's Deepfake Detection Challenge. Facebook launched the contest last fall to crowdsource ideas for stopping the spread of misinformation on its platform ahead of the 2020 presidential election.

If a technical solution to deepfakes is found, it will need to run on millions of machines at once, says Andonian. That makes energy efficiency key. "Every optimization we can find to train and run more efficient models will make a huge difference," he says.

To speed up the training process, they tried streamlining their code and lowering the resolution of their 100,000-video training set by eliminating some frames. They didn't expect a solution in three days, but Satori's size worked in their favor. "We were able to run 10 to 20 experiments at a time, which let us iterate on potential ideas and get results quickly," says Andonian.

As AI continues to improve at tasks like reading medical scans and interpreting video, models have grown bigger and more calculation-intensive, and thus, energy intensive. By one estimate, training a large language-processing model produces nearly as much carbon dioxide as the cradle-to-grave emissions from five American cars. The footprint of the typical model is modest by comparison, but as AI applications proliferate its environmental impact is growing.

One way to green AI, and tame the exponential growth in demand for training AI, is to build smaller models. That's the approach that a third hackathon competitor, EECS graduate student Jonathan Frankle, took. Frankle is looking for signals early in the training process that point to subnetworks within the larger, fully-trained network that can do the same job. The idea builds on his award-winning Lottery Ticket Hypothesis paper from last year that found a neural network could perform with 90 percent fewer connections if the right subnetwork was found early in training.

The hackathon competitors were judged by John Cohn, chief scientist at the MIT-IBM Watson AI Lab, Christopher Hill, director of MIT's Research Computing Project, and Lauren Milechin, a research software engineer at MIT.

The judges recognized four other teams: Department of Earth, Atmospheric and Planetary Sciences (EAPS) graduate students Ali Ramadhan, Suyash Bire, and James Schloss, for adapting the programming language Julia for Satori; MIT Lincoln Laboratory postdoc Andrew Kirby, for adapting code he wrote as a graduate student to Satori using a library designed for easy programming of computing architectures; and Department of Brain and Cognitive Sciences graduate students Jenelle Feather and Kelsey Allen, for applying a technique that drastically simplifies models by cutting their number of parameters.

"We pushed the system — in a good way," says Cohn. "In the end, we improved the machine, the documentation, and the tools around it."

IBM developers were on hand to answer questions and gather feedback. "We pushed the system — in a good way," says Cohn. "In the end, we improved the machine, the documentation, and the tools around it."

Going forward, Satori will be joined in Holyoke by TX-Gaia, Lincoln Laboratory's new supercomputer. Together, they will provide feedback on the energy use of their workloads. "We want to raise awareness and encourage users to find innovative ways to green-up all of their computing," says Hill.

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the United States Air Force under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.