Massachusetts Institute of Technology

MIT News
ON CAMPUS AND AROUND THE WORLD

October 5, 2023

New tools are available to help reduce the energy that AI models devour

Amid the race to make AI bigger and better, Lincoln Laboratory is developing ways to reduce power, train efficiently, and make energy use transparent.

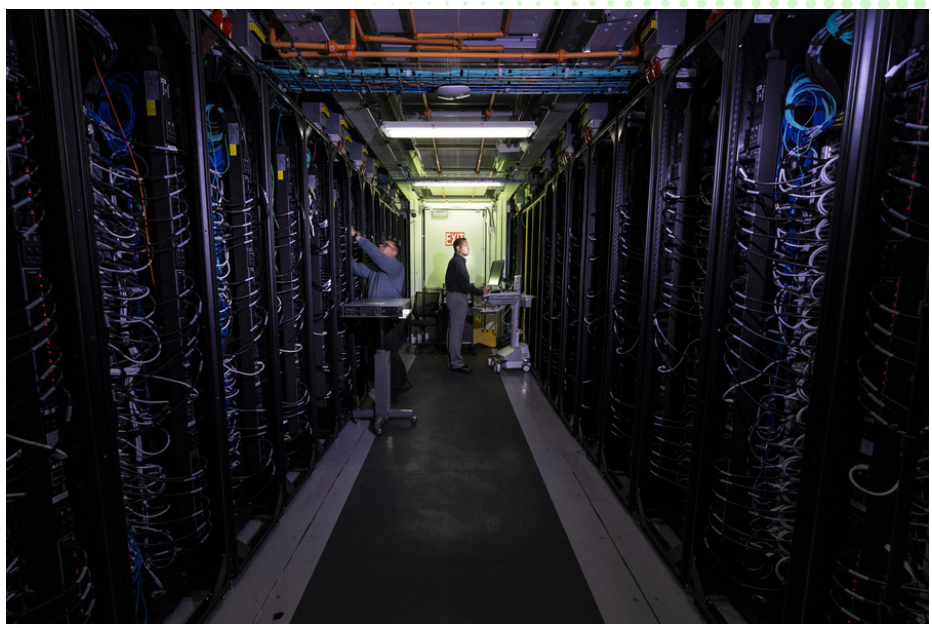
Kylie Foy | MIT Lincoln Laboratory

When searching for flights on Google, you may have noticed that each flight's carbon-emission estimate is now presented next to its cost. It's a way to inform customers about their environmental impact, and to let them factor this information into their decision-making.

A similar kind of transparency doesn't yet exist for the computing industry, despite its carbon emissions exceeding those of the entire airline industry. Escalating this energy demand are artificial intelligence models. Huge, popular models like ChatGPT signal a trend of large-scale artificial intelligence, boosting forecasts that predict data centers will draw up to 21 percent of the world's electricity supply by 2030.

The MIT Lincoln Laboratory Supercomputing Center (LLSC) is developing techniques to help data centers reel in energy use. Their techniques range from simple but effective changes, like power-capping hardware, to adopting novel tools that can stop AI training early on. Crucially, they have found that these techniques have a minimal impact on model performance.

In the wider picture, their work is mobilizing green-computing research and promoting a culture of transparency. "Energy-aware computing is not really a research area, because everyone's been holding on to their data," says Vijay Gadepally, senior staff in the LLSC who leads energy-aware research efforts. "Somebody has to start, and we're hoping others will follow."



At the Lincoln Laboratory Supercomputing Center, researchers are making changes to cut down on energy use. One of their techniques can reduce the energy of training AI models by 80 percent. Photo: Glen Cooper

Curbing power and cooling down

Like many data centers, the LLSC has seen a significant uptick in the number of AI jobs running on its hardware. Noticing an increase in energy usage, computer scientists at the LLSC were curious about ways to run jobs more efficiently. Green computing is a principle of the center, which is powered entirely by carbon-free energy.

Training an AI model — the process by which it learns patterns from huge datasets — requires using graphics processing units (GPUs), which are power-hungry hardware. As one example, the GPUs that trained GPT-3 (the precursor to ChatGPT) are estimated to have consumed 1,300 megawatt-hours of electricity, roughly equal to that used by 1,450 average U.S. households per month.

While most people seek out GPUs because of their computational power, manufacturers offer ways to limit the amount of power a GPU is allowed to draw. "We studied the effects of capping power and found that we could reduce energy consumption by about 12 percent to 15 percent, depending on the model," Siddharth Samsi, a researcher within the LLSC, says.

The trade-off for capping power is increasing task time — GPUs will take about 3 percent longer to complete a task, an increase Gadepally says is "barely noticeable" considering that models are often trained over days or even months. In one of their experiments in which they trained the popular BERT language model, limiting GPU power to 150 watts saw a two-hour increase in training time (from 80 to 82 hours) but saved the equivalent of a

U.S. household's week of energy.

The team then built software that plugs this power-capping capability into the widely used scheduler system, Slurm. The software lets data center owners set limits across their system or on a job-by-job basis.

"We can deploy this intervention today, and we've done so across all our systems," Gadepally says.

Side benefits have arisen, too. Since putting power constraints in place, the GPUs on LLSC supercomputers have been running about 30 degrees Fahrenheit cooler and at a more consistent temperature, reducing stress on the cooling system. Running the hardware cooler can potentially also increase reliability and service lifetime. They can now consider delaying the purchase of new hardware — reducing the center's "embodied carbon," or the emissions created through the manufacturing of equipment — until the efficiencies gained by using new hardware offset this aspect of the carbon footprint. They're also finding ways to cut down on cooling needs by strategically scheduling jobs to run at night and during the winter months.

"Data centers can use these easy-to-implement approaches today to increase efficiencies, without requiring modifications to code or infrastructure," Gadepally says.

Taking this holistic look at a data center's operations to find opportunities to cut down can be time-intensive. To make this process easier for others, the team — in collaboration with Professor Devesh Tiwari and Baolin Li at Northeastern University — recently developed and published a comprehensive framework for analyzing the carbon footprint of high-performance computing systems. System practitioners can use this analysis framework to gain a better understanding of how sustainable their current system is and consider changes for next-generation systems.

Adjusting how models are trained and used

On top of making adjustments to data center operations, the team is devising ways to make AI-model development more efficient.

When training models, AI developers often focus on improving accuracy, and they build upon previous models as a starting point. To achieve the desired output, they have to figure out what parameters to use, and getting it right can take testing thousands of configurations. This process, called hyperparameter optimization, is one area LLSC researchers have found ripe for cutting down energy waste.

"We've developed a model that basically looks at the rate at which a given configuration is learning," Gadepally says. Given that rate, their model predicts the likely performance. Underperforming models are stopped early. "We can give you a very accurate estimate early on that the best model will be in this top 10 of 100 models running," he says.

In their studies, this early stopping led to dramatic savings: an 80 percent reduction in the energy used for model training. They've applied this technique to models developed for computer vision, natural language processing, and material design applications.

"In my opinion, this technique has the biggest potential for advancing the way AI models are trained," Gadepally says.

Training is just one part of an AI model's emissions. The largest contributor to emissions over time is model inference, or the process of running the model live, like when a user chats with ChatGPT. To respond quickly, these models use redundant hardware, running all the time, waiting for a user to ask a question.

One way to improve inference efficiency is to use the most appropriate hardware. Also with Northeastern University, the team created an optimizer that matches a model with the most carbon-efficient mix of hardware, such as high-power GPUs for the computationally intense parts of inference and low-power central processing units (CPUs) for the less-demanding aspects. This work recently won the best paper award at the International ACM Symposium on High-Performance Parallel and Distributed Computing.

Using this optimizer can decrease energy use by 10-20 percent while still meeting the same "quality-of-service target" (how quickly the model can respond).

This tool is especially helpful for cloud customers, who lease systems from data centers and must select hardware from among thousands of options. "Most customers overestimate what they need; they choose over-capable hardware just because they don't know any better," Gadepally says.

Growing green-computing awareness

The energy saved by implementing these interventions also reduces the associated costs of developing AI, often by a one-to-one ratio. In fact, cost is usually used as a proxy for energy consumption. Given these savings, why aren't more data centers investing in green techniques?

"I think it's a bit of an incentive-misalignment problem," Samsi says. "There's been such a race to build bigger and better models that almost every secondary consideration has been put aside."

They point out that while some data centers buy renewable-energy credits, these renewables aren't enough to cover the growing energy demands. The majority of electricity powering data centers comes from fossil fuels, and water used for cooling is contributing to stressed watersheds.

Hesitancy may also exist because systematic studies on energy-saving techniques haven't been conducted. That's why the team has been pushing their research in peer-reviewed venues in addition to open-source repositories. Some big industry players, like Google DeepMind, have applied machine learning to increase data center efficiency but have not made their work available for others to deploy or replicate.

Top AI conferences are now pushing for ethics statements that consider how AI could be misused. The team sees the climate aspect as an AI ethics topic that has not yet been given much attention, but this also appears to be slowly changing. Some researchers are now disclosing the carbon footprint of training the latest models, and industry is showing a shift in energy transparency too, as in this recent report from Meta AI.

They also acknowledge that transparency is difficult without tools that can show AI developers their consumption. Reporting is on the LLSC roadmap for this year. They want to be able to show every LLSC user, for every job, how much energy they consume and how this amount compares to others, similar to home energy reports.

Part of this effort requires working more closely with hardware manufacturers to make getting these data off hardware easier and more accurate. If manufacturers can standardize the way the data are read out, then energy-saving and reporting tools can be applied across different hardware platforms. A collaboration is underway between the LLSC researchers and Intel to work on this very problem.

Even for AI developers who are aware of the intense energy needs of AI, they can't do much on their own to curb this energy use. The LLSC team wants to help other data centers apply these interventions and provide users with energy-aware options. Their first partnership is with the U.S. Air Force, a sponsor of this research, which operates thousands of data centers. Applying these techniques can make a significant dent in their energy consumption and cost.

"We're putting control into the hands of AI developers who want to lessen their footprint," Gadepally says. "Do I really need to gratuitously train unpromising models? Am I willing to run my GPUs slower to save energy? To our knowledge, no other supercomputing center is letting you consider these options. Using our tools, today, you get to decide."