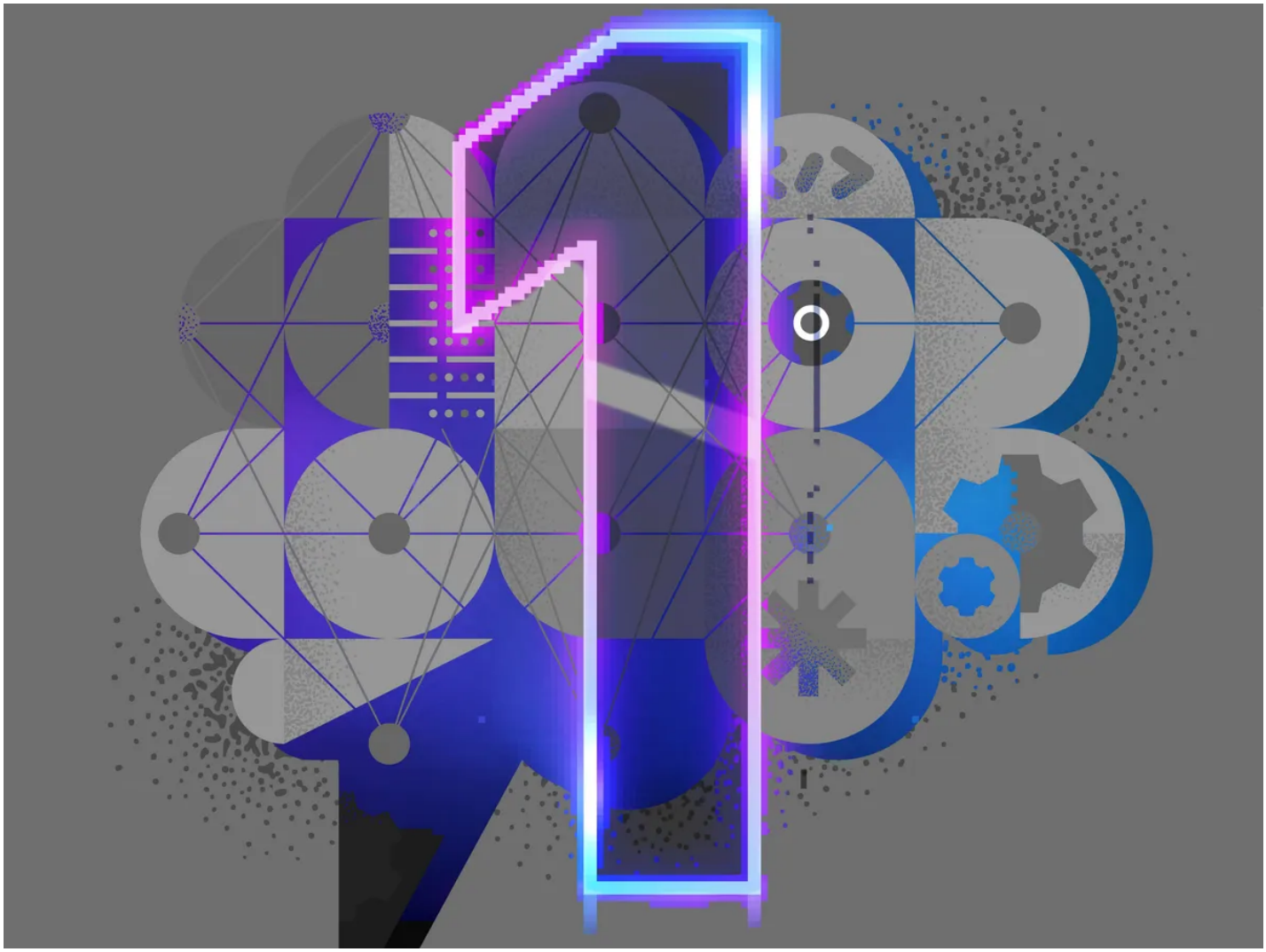
NEWS ARTIFICIAL INTELLIGENCE

1-bit LLMs Could Solve AI's Energy Demands

"Imprecise" language models are smaller, speedier—and nearly as accurate

BY MATTHEW HUTSON

15 HOURS AGO



GETTY IMAGES

■ arge language models, the AI systems that power

Large language models, the AI systems that power chatbots like ChatGPT, are getting better and better —but they're also getting bigger and bigger, demanding more energy and computational power. For LLMs that are cheap, fast, and environmentally friendly, they'll need to shrink, ideally small enough to run directly on devices like cellphones. Researchers are finding ways to do just that by drastically rounding off the many high-precision numbers that store their memories to equal just 1 or -1.

LLMs, like all neural networks, are trained by altering the strengths of connections between their artificial neurons. These strengths are stored as mathematical parameters. Researchers have long compressed networks by reducing the precision of these parameters—a process called quantization —so that instead of taking up 16 bits each, they might take up 8 or 4. Now researchers are pushing the envelope to a single bit.

How to Make a 1-bit LLM

There are two general approaches. One approach, called post-training quantization (PTQ) is to quantize the parameters of a full-precision network. The other approach, quantization-

aware training (QAT), is to train a network from scratch to have low-precision parameters. So far, PTQ has been more popular with researchers.

In February, a team including Haotong Qin at ETH Zurich, Xianglong Liu at Beihang University, and Wei Huang at the University of Hong Kong introduced a PTQ method called BiLLM. It approximates most parameters in a network using 1 bit, but represents a few salient weights—those most influential to performance—using 2 bits. In one test, the team binarized a version of Meta's LLaMa LLM that has 13 billion parameters.

»

"One-bit LLMs open new doors for designing custom hardware and systems specifically optimized for 1-bit LLMs."

—FURU WEI, MICROSOFT RESEARCH ASIA

To score performance, the researchers used a metric calledperplexity, which is basically a measure of how surprised the trained model was by each ensuing piece of text. For one dataset, the original model had a perplexity of around

5, and the BiLLM version scored around 15, much better than the closest binarization competitor, which scored around 37 (for perplexity, lower numbers are better). That said, the BiLLM model required about a tenth of the memory capacity as the original.

PTQ has several advantages over QAT, says Wanxiang Che, a computer scientist at Harbin Institute of Technology, in China. It doesn't require collecting training data, it doesn't require training a model from scratch, and the training process is more stable. QAT, on the other hand, has the potential to make models more accurate, since quantization is built into the model from the beginning.

1-bit LLMs Find Success Against Their Larger Cousins

Last year, a team led by Furu Wei and Shuming Ma, at Microsoft Research Asia, in Beijing, created BitNet, the first 1-bit QAT method for LLMs. After fiddling with the rate at which the network adjusts its parameters, in order to stabilize training, they created LLMs that performed better than those created using PTQ methods. They were still not as good as full-precision networks, but roughly 10 times as energy

full-precision networks, but roughly 10 times as energy efficient.

In February, Wei's team announced BitNet 1.58b, in which parameters can equal -1, 0, or 1, which means they take up roughly 1.58 bits of memory per parameter. A BitNet model with 3 billion parameters performed just as well on various language tasks as a full-precision LLaMA model with the same number of parameters and amount of training—Wei called this an "aha moment"—but it was 2.71 times as fast, used 72 percent less GPU memory, and used 94 percent less GPU energy. Further, the researchers found that as they trained larger models, efficiency advantages improved.

A BitNet model with 3 billion parameters performed just as well on various language tasks as a full-precision LLaMA model.

This year, a team led by Che, of Harbin Institute of Technology, released a preprint on another LLM binarization method, called OneBit. OneBit combines elements of both PTQ and QAT. It uses a full-precision pretrained LLM to generate data for training a quantized version. The team's 13

generate data for training a quantized version. The team's 13-billion-parameter model achieved a perplexity score of around 9 on one dataset, versus 5 for a LLaMA model with 13 billion parameters. Meanwhile, OneBit occupied only 10 percent as much memory. On customized chips, it could presumably run much faster.

Wei, of Microsoft, says quantized models have multiple advantages. They can fit on smaller chips, they require less data transfer between memory and processors, and they allow for faster processing. Current hardware can't take full advantage of these models, though. LLMs often run on GPUs like those made by Nvidia, which represent weights using higher precision and spend most of their energy multiplying them. New hardware could natively represent each parameter as a -1 or 1 (or 0), and then simply add and subtract values and avoid multiplication. "One-bit LLMs open new doors for designing custom hardware and systems specifically optimized for 1-bit LLMs," Wei says.

"They should grow up together," Huang, of the University of Hong Kong, says of 1-bit models and processors. "But it's a long way to develop new hardware."