Deep Learning at the Speed of Light

Lightmatter bets that optical computing can solve AI's efficiency problem

By DAVID SCHNEIDER

In 2011, **Marc Andreessen,** general partner of venture capital firm Andreessen Horowitz, wrote an influential article in *The Wall Street Journal* titled, "Why Software Is Eating the World." A decade later now, it's *deep learning* that's eating the world.

Deep learning, which is to say artificial neural networks with many hidden layers, is regularly stunning us with solutions to real-world problems. And it is doing that in more and more realms, including natural-language processing, fraud detection, image recognition, and autonomous driving. Indeed, these neural networks are getting better by the day.

But these advances come at an enormous price in the computing resources and energy they consume. So it's no wonder that engineers and computer scientists are making huge efforts to figure out ways to train and run deep neural networks more efficiently.

An ambitious new strategy that's coming to the fore this year is to perform many of the required mathematical calculations using photons rather than electrons. In particular, one company, Lightmatter, will begin marketing late this year a neural-network accelerator chip that calculates with light. It will be a refinement of the prototype Mars chip that the company showed off at the virtual Hot Chips conference last August.

While the development of a commercial optical accelerator for deep learning is a remarkable accomplishment, the general idea of computing with light is not new. Engineers regularly resorted to this tactic in the 1960s and '70s, when electronic digital computers were too feeble to perform the complex calculations needed to process synthetic-aperture radar data. So they processed the data in the analog domain, using light.

Because of the subsequent Moore's Law gains in what could be done with digital electronics, optical computing never really

caught on, despite the ascendancy of light as a vehicle for data communications. But all that may be about to change: Moore's Law may be nearing an end, just as the computing demands of deep learning are exploding.

There aren't many ways to deal with this problem. Deep-learning researchers may develop more efficient algorithms, sure, but it's hard to imagine those gains will be sufficient. "I challenge you to lock a bunch of theorists in a room and have them come up with a better algorithm every 18 months," says Nicholas Harris, CEO of Lightmatter. That's why he and his colleagues are bent on "developing a new compute technology that doesn't rely on the transistor."

So what then does it rely on?

The fundamental component in Lightmatter's chip is a Mach-Zehnder interferometer. This optical device was jointly invented by Ludwig Mach and Ludwig Zehnder in the 1890s. But only recently have such optical devices been miniaturized to the point where large numbers of them can be integrated onto a chip and used to perform the matrix multiplications involved in neural-network calculations.

Keren Bergman, a professor of electrical engineering and the director of the Lightwave Research Laboratory at Columbia University, in New York City, explains that these feats have become possible only in the last few years because of the maturing of the manufacturing ecosystem for integrated photonics, needed to make photonic chips for communications. "What you would do on a bench 30 years ago, now they can put it all on a chip," she says.

Processing analog signals carried by light slashes energy costs and boosts the speed of calculations, but the precision can't match what's possible in the digital domain. "We have an 8-bit-equivalent system," say Harris. This limits his company's chip to neural-network inference calculations–the ones that are carried out after the network has been trained. Harris and his colleagues hope their technology might one day be applied to training neural networks, too, but training demands more precision than their optical processor can now provide.

Lightmatter is not alone in the quest to harness light for neural-network calculations. Other startups working along

► LIGHT PIPE: Lightmatter uses an integrated Mach-Zehnder interferometer, part of which is simulated here.
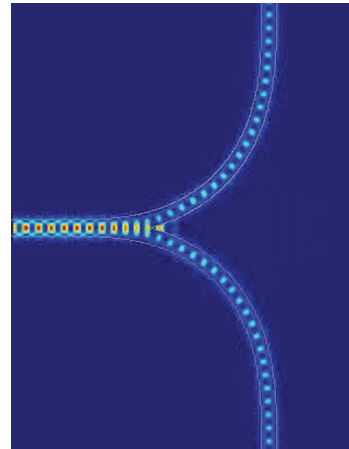
these lines include Fathom Computing, LightIntelligence, LightOn, Luminous, and Optalysis. One of these, Luminous, hopes to apply optical computing to spiking neural networks, which take advantage of the way the neurons of the brain process information–perhaps accounting for why the human brain can do the remarkable things it does using just a dozen or so watts.

Luminous expects to develop practical systems sometime between 2022 and 2025. So we'll have to wait a few years yet to see what pans out with its approach. But many are excited about the prospects, including Bill Gates, one of the company's high-profile investors.

It's clear, though, that the computing resources being dedicated to artificial-intelligence systems can't keep growing at the current rate, doubling every three to four months. Engineers are now keen to harness integrated photonics to address this challenge with a new class of computing machines that are dramatically different from conventional electronic chips yet are now practical to manufacture. Bergman boasts: "We have the ability to make devices that in the past could only be imagined." ∎

◄ PLUG AND PLAY: Lightmatter's prototype board uses a normal PCI bus.

POST YOUR COMMENTS AT
spectrum.ieee.org/lightmatter-jan2021