UC SANTA BARBARA

THE Current

August 5, 2024
James Badham

Proving the value of an old algorithm in training deep-learning models for AI

For about a decade, computer engineer Kerem Çamsari employed a novel approach known as probabilistic computing. Based on probabilistic bits (p-bits), it's used to solve an array of complex combinatorial optimization problems. In one of the best known of these, "the traveling salesperson problem," a salesperson must find the shortest route to visit a given number of cities, none more than once.

But with "everything moving to AI," said Çamsari, an associate professor in UC Santa Barbara's Department of Electrical and Computer Engineering, he began applying his optimization algorithms to the new task of training a deep generative artificial intelligence (AI) model.

Recently, Shaila Niazi, a third-year doctoral student in Çamsari's lab, achieved a significant breakthrough in that effort, becoming the first to use probabilistic hardware to train a deep generative model on a large scale to address a real-life problem, such as recognizing handwritten digits or images of real objects like birds, dogs and automobiles. Niazi used those novel tools to generate an image that was not in the training dataset, a basic task for a generative AI model.

"After training our network," she added, "we can tell it to dream up a new image, and it can do that." The work appears in the "Training deep Boltzmann networks

with sparse Ising machines," published in the journal Nature Electronics.

As far as we know, Çamsari noted, "This may be the first paper describing the use of Ising machines — a recently developed physics-based probabilistic computer designed to perform optimization problems — to train a large-scale machine-learning (ML) model without any simplifications of the dataset. That's something new, made possible only by recent advances in probabilistic computers." Previously, to simplify the recognition task, people might have lowered the image quality of a 28x28-pixel image, by converting it to, say, 6x6 pixels.

Traditional computing is based on deterministic bits, which must have one of two values — 0 or 1 — at any given time and change only according to any specific computation. A probabilistic bit differs in that it is never a definite 0 or 1, but fluctuates constantly, as rapidly as every nanosecond. The p-bit is a physical hardware building block that can generate that string of 0s and 1s, providing built-in randomness that is often useful in algorithms.

Niazi's accomplishment relied on the tremendous computing ability in a machine made more powerful by a piece of adaptive hardware designed in Çamsari's lab. There, a type of nanodevice used in magnetic memory technology is modified to make it highly "memory-less," such that it naturally fluctuates in the presence of thermal noise at room temperature. Çamsari's team also uses an algorithm that has been out of favor with the AI community for more than a decade. "We're not following the current paradigm," Çamsari said.

That approach allowed Niazi to create a very "deep" three-layer neural network. "Each layer in a neural network consists of a set of neurons that process information received from the previous layer, transform it in some way and pass it on to the next layer," Niazi explained. "These layers are like steps in a process, with each step dealing with increasingly complex aspects of the information it receives."

Çamsari pointed out that while all neurons in the human brain are the same, that is not the case in the algorithm. "Suppose you show the model an image of a cat," he said. "The first layer might recognize triangular shapes, like the ears, that make it possible to recognize a cat. The second layer captures higher-level features, maybe some finer detail inside the ear.

"Usually, in what is called energy-based ML, two layers of neurons are connected to each other," he continued. "In the past, the limitations of hardware meant that

having more than two layers was difficult, even though it was well-established that increasing the depth of a network by adding layers would be tremendously useful. The phrase deep learning refers to that network depth, the hierarchical structure of the neural network on which today's whole deep-learning revolution has been built."

In recent years, the ML field has been dominated by what is called the backpropagation algorithm, or backprop*,* which, Çamsari said, "is basically driving everything right now, but in my lab, we use what's called a contrastive algorithm, a physics-based model used to solve optimization problems, but that we are now repurposing to train a neural network for an AI model."

For some time, backprop and the contrastive algorithm were about equal in terms of their ability to power AI applications. But around 2010, graphical processing units (GPUs), which Nvidia introduced in 1999, began to be used for AI. "Backprop was more amenable to that hardware, so people stopped using contrasted algorithms, which were too hard to train with hardware that was not optimized for them," Çamsari said. "But they fit our physics-based Ising machines and probabilistic computers really well."

Niazi got her results by training a deeper model. "When Shaila added layers, she was immediately rewarded," Çamsari said. "Not only could she generate new images, but she did so using only thirty thousand parameters (the number of unique bits of information a machine can hold), compared to the three million parameters used by shallower models that failed to generate images. That convinced the journal reviewers that there might be something here."

Making an Image

Çamsari explains how p-bits work to create an image, perhaps a simple black-and-white image of a numeral in a square having 28 pixels on each side, where each 1x1-pixel square therein is a p-bit, which is synonymous with a neuron and is the basic building block of a probabilistic computer. To draw an image of a numeral — a one, a three, and so on — the correct p-bits in the square have to be on, and the correct ones have to be off at the right time.

In principle, finding the correlations to train the network is easy, but in practice, it can be very difficult. According to Niazi, it would take several months of training on a classical computer to get results equivalent to what she gets in less than a day using

the fast Field Programmable Gate Array (FPGA)-based p-computer, which makes roughly sixty billion decisions per second.

“Initially, we used an easy dataset — called an MNIST dataset — to draw the digits zero, one, two, three as a way to verify our algorithms and the hardware's strength,” she said. “Later, the journal reviewers asked us to train some more-difficult datasets that contain more-complex images of airplanes or automobiles. They suggested that even a simulation would be fine if we could not train the model on our hardware. Based on various considerations, we ended up successfully training those more-difficult datasets on our hardware, and it worked, even with our limited p-bit resources.”

Others who are using the non-backprop algorithm are not solving the same problem as Niazi is, but rather, an easier version of it, said Çamsari. “Shaila’s p-bits, or neurons, in this context, have 15 or 20 neighbors,” he noted. “The GPU problems that other people have solved for similar Ising problems have only three or four neighbors, and those other researchers were not trying to train datasets or generate new images; they were just trying to solve a simple probabilistic problem as a way to perform a speed test on their machines. Solving our real-world problem requires many more than three or four neighbors.”

Niazi has been able to generate grayscale images, but color images will require further advances. “We’re at the limits of our machine’s capacity and hungry for more computing power, so we need to scale up, but doing that with silicon is going to be tough. We are at the end of something here,” Çamsari explained. “Shaila can currently fit only 5,000 p-bits using the world’s best computer chips, but for color, we would need around 15,000 to 20,000 p-bits. We are working on alternative implementations with nanodevices to make that happen.”

Tags
Artificial Intelligence

Media Contact
James Badham

About UC Santa Barbara

The University of California, Santa Barbara is a leading research institution that also provides a comprehensive liberal arts learning experience. Our academic community of faculty, students, and staff is characterized by a culture of interdisciplinary collaboration that is responsive to the needs of our multicultural and global society. All of this takes place within a living and learning environment like no other, as we draw inspiration from the beauty and resources of our extraordinary location at the edge of the Pacific Ocean.