UC SANTA BARBARA

THE Current

February 10, 2020
James Badham

Bring the Noise

Those who design deep neural networks for artificial intelligence often find inspiration in the human brain. One of the brain's more important characteristics is that it is a "noisy" system: not every neuron contains perfect information that gets carried across a synapse with perfect clarity. Sometimes partial or conflicting information is turned into action by the brain, and sometimes partial information is not acted upon until further information is accumulated over time.

"That is why, when you stimulate the brain with the same input at different times, you get different responses," explained Mohammad "Reza" Mahmoodi, a fifth-year Ph.D. candidate in the lab of UC Santa Barbara electrical and computer engineering professor Dmitri Strukov. "Noisy, unreliable molecular mechanisms are the reason for getting substantially different neural responses to repeated presentations of identical stimuli, which, in turn, allow for complex stochastic, or unpredictable, behavior."

The human brain is extremely good at filling in the blanks of missing information and sorting through noise to come up with an accurate result, so that "garbage in" does not necessarily yield "garbage out." In fact, Mahmoodi said, the brain seems to work best with noisy information. In stochastic computing, noise is used to train neural networks, "regularizing" them to improve their robustness and performance.

It is not clear on what theoretical basis neuronal responses involved in perceptual processes can be separated into a "noise" versus a "signal," Mahmoodi explained, but the noisy nature of computation in the brain has inspired the development of

stochastic neural networks. And those have now become the state-of-the-art approach for solving problems in machine learning, information theory and statistics.

“If you want a stochastic system, you have to generate some noise,” Mahmoodi and his co-authors, Strukov and Mirko Prezioso, write in a paper that describes an approach to creating such a noisy system. “Versatile stochastic dot product circuits based on nonvolatile memories for high performance neurocomputing and neurooptimization ” was published in a recent issue of the journal Nature Communications.

The most famous kind of network that operates based on stochastic computation is the so-called “Boltzmann” machine, which can solve difficult combinatorial optimization problems. Such problems are characterized by an essentially infinite number of possible reasonable solutions but no one absolutely best solution. The traveling salesman problem — that a salesman needs to pass through every state in the nation to sell products, but must do so by taking the shortest path possible — is a famous example.

No clear optimal, perfect solution exists because the space is so large and the possible combinations of routes within it are nearly limitless. Yet, Mahmoodi notes, “You can use neural networks and heuristic algorithms to find a kind of a semi-optimized solution. What matters is that you can generate a good response in a reasonable amount of time.”

This can be facilitated by applying an algorithm called “simulated annealing,” which is inspired by the crystallization process in physics.

“To obtain a crystal structure,” Mahmoodi said, “you heat up a solid to a very high temperature and then slowly cool it down. If you cool it slowly enough, all the molecules find their lowest-energy position, the most perfect location, and you get a beautiful, entirely uniform crystal.”

An analogous approach is used in simulated annealing. “Indeed,” Mahmoodi explains, “when we start solving the problem, we use too much noise — analogous to a too-high temperature in crystal formation. The result is that computations in the neural network are stochastic, or random. Then, we slowly reduce the amount of injected noise while moving toward deterministic, or entirely predictable computation, which, continuing the crystal-forming analogy, is referred to as ‘lowering the temperature.’ This procedure improves the network’s ability to explore

the search space and results in a much better final solution.”

The big question for the team is whether they can build a stochastic neural network that is fast and energy efficient, and can be operated with adjustable temperature (noise). Most artificial neural networks have two things in common: a huge number of weights, which are essentially the tunable parameters that networks learn during training; and a sprawling foundation of computational blocks, mostly performing multiplication and addition operations.

Building an energy-efficient, high-throughput neural network, therefore, requires devices that can store more information in a given area, and circuits that can perform the computation faster and with greater energy efficiency. While there have been many demonstrations of multiplication circuits and, separately, stochastic neurons, the efficient hardware implementation combining both functionalities is still missing.

In the Strukov lab, Mahmoodi and others are working on two mainstream technologies that are key to implementing neural networks: memristors and embedded flash.

“We are fortunate to be able to fabricate state-of-the-art analog memristor technology here at UCSB,” Mahmoodi said. “Each memristor or flash-cell device is small and can store more than five bits of data, as opposed to digital memories, like SRAM, which are much bulkier and can store only a single bit. Hence, we use these smaller, more efficient devices to design mixed-signal neural networks that have both analog and digital circuits and are therefore much faster and more efficient than pure digital systems.

“Indeed, in our paper, we report compact, fast, energy-efficient and scalable stochastic neural-network circuits based on either memristors or embedded flash,” he added. “The circuits’ high performance is due to mixed-signal (digital and analog) implementation, while the efficient stochastic operation is achieved by utilizing the circuit’s intrinsic noise. We show that our circuit can efficiently solve optimization problems orders of magnitude faster and with much greater energy efficiency than CPUs can.”

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.