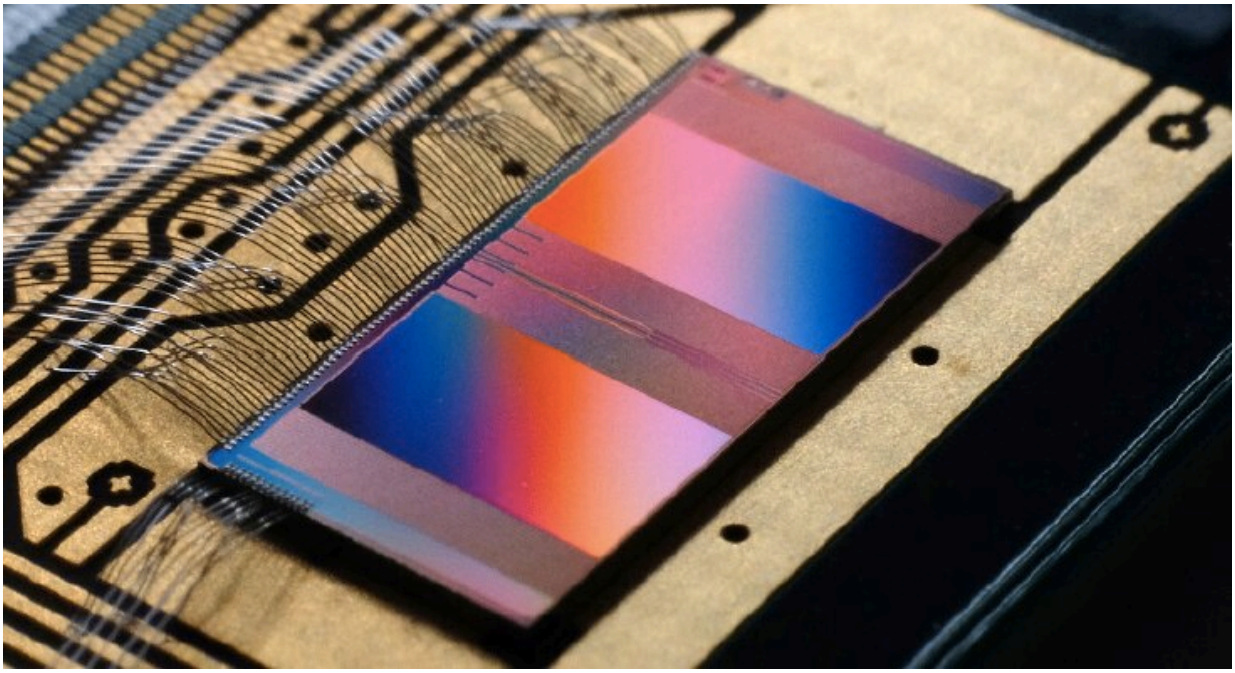
A framework to enhance deep learning using first-spike times

October 5 2021, by Ingrid Fadelli



Photograph of the BrainScaleS-2 chip used for the emulation. This mixed-signal neuromorphic research chip is used for various projects in Heidelberg and thanks to its analog accelerator the platform is characterized by speed and energy-efficiency. Credit: kip.uni-heidelberg.de/vision/

Researchers at Heidelberg University and University of Bern have recently devised a technique to achieve fast and energy-efficient computing using spiking neuromorphic substrates. This strategy,

introduced in a paper published in *Nature Machine Intelligence*, is a rigorous adaptation of a time-to-first-spike (TTFS) coding scheme, together with a corresponding learning rule implemented on certain networks of artificial neurons. TTFS is a time-coding approach, in which the activity of neurons is inversely proportional to their firing delay.

"A few years ago, I started my Master's thesis in the Electronic Vision(s) group in Heidelberg," Julian Goeltz, one of the leading researchers working on the study, told TechXplore. "The neuromorphic BrainScaleS system developed there promised to be an intriguing substrate for brain-like computation, given how its neuron and synapse circuits mimic the dynamics of neurons and synapses in the brain."

When Goeltz started studying in Heidelberg, deep-learning models for spiking networks were still relatively unexplored and existing approaches did not use spike-based communication between neurons very effectively. In 2017, Hesham Mostafa, a researcher at University of California—San Diego, introduced the idea that the timing of individual neuronal spikes could be used for information processing. However, the neuronal dynamics he outlined in his paper were still quite different from biological ones and thus were not applicable to brain-inspired neuromorphic hardware.

"We therefore needed to come up with a hardware-compatible variant of error backpropagation, the algorithm underlying the modern AI revolution, for single spike times," Goeltz explained. "The difficulty lay in the rather complicated relationship between synaptic inputs and outputs of spiking neurons."

Initially, Goeltz and his colleagues set out to develop a mathematical framework that could be used to approach the problem of achieving deep learning based on temporal coding in spiking neural networks. Their goal was to then transfer this approach and the results they

gathered onto the BrainScaleS system, a renowned neuromorphic computing system that emulates models of neurons, synapses, and brain plasticity.

"Assume that we have a layered network in which the input layer receives an image, and after several layers of processing the topmost layer needs to recognize the image as being a cat or a dog," Laura Kriener, the second lead researcher for the study, told TechXplore. "If the image was a cat, but the 'dog' neuron in the top layer became active, the network needs to learn that its answer was wrong. In other words, the network needs to change connections—i.e., synapses—between the neurons in such a way that the next time it sees the same picture, the 'dog' neuron stays silent and the 'cat' neuron is active."

The problem described by Kriener and addressed in the recent paper, known as the 'credit assignment problem," essentially entails understanding which synapses in a neural network are responsible for a network's output or prediction, and how much of the credit each synapse should take for a given prediction.

To identify what synapses were involved in a network's wrong prediction and fix the issue, researchers often use the so-called error backpropagation algorithm. This algorithm works by propagating an error in the topmost layer of a neural network back through the network, to inform synapses about their own contribution to this error and change each of them accordingly.

When neurons in a network communicate via spikes, each input spike 'bumps' the potential of a neuron up or down. The size of this 'bump' depends on the weight of a given synapse, known as 'synaptic weight."

"If enough upward bumps accumulate, the neuron 'fires'—it sends out a spike of its own to its partners," Kriener said. "Our framework

effectively tells a synapse exactly how to change its weight to achieve a particular output spike time, given the timing errors of the neurons in the layers above, similarly to the backpropagation algorithm, but for spiking neurons. This way, the entire spiking activity of a network can be shaped in the desired way—which, in the example above, would cause the 'cat' neuron to fire early and the 'dog' neuron to stay silent or fire later."

Due to its spike-based nature and to the hardware used to implement it, the framework developed by Goeltz, Kriener and their colleagues exhibits remarkable speed and efficiency. Moreover, the framework encourages neurons to spike as quickly as possible and only once. Therefore, the flow of information is both quick and sparse, as very little data needs to flow through a given neural network to enable it to complete a task.

"The BrainScaleS hardware further amplifies these features, as its neuron dynamics are extremely fast—1000 times faster than those in the brain—which translates to a correspondingly higher information processing speed," Kriener explained. "Furthermore, the silicon neurons and synapses are designed to consume very little power during their operation, which brings about the energy efficiency of our neuromorphic networks."

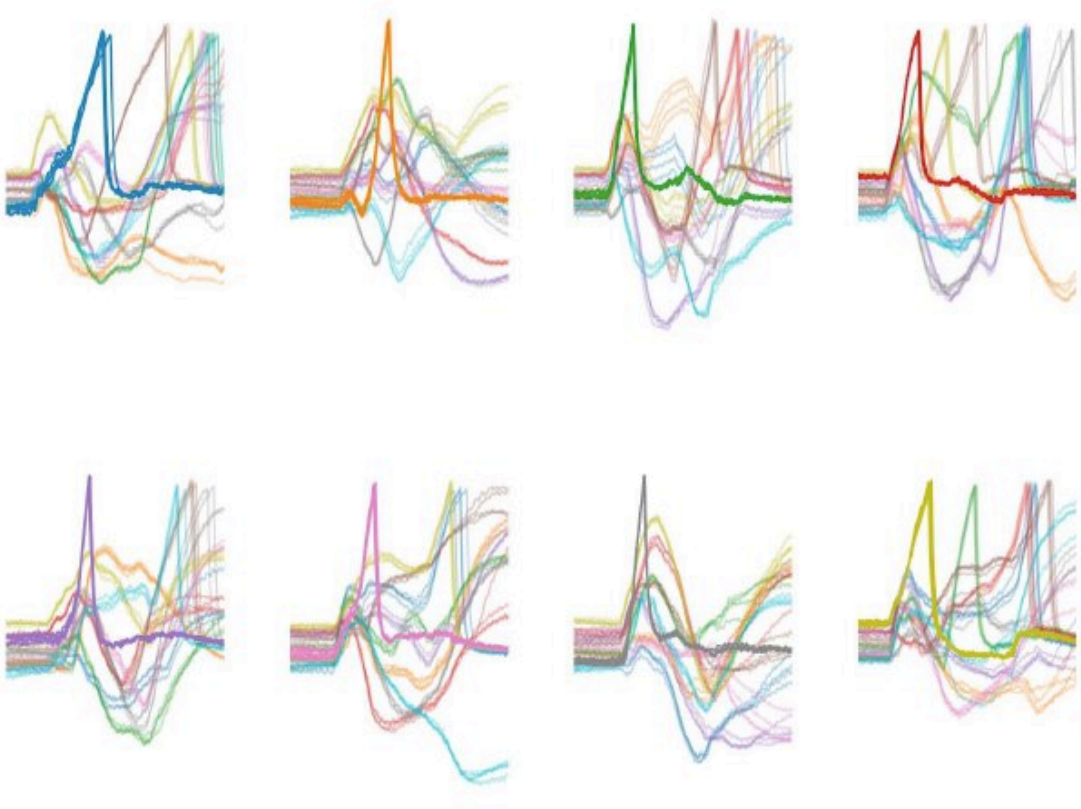


Illustration of the on-chip classification process. The traces in the eight panels show the membrane voltages of the classifying neurons. The sharp peak is when the neuron spikes. Our algorithm aims to have the 'correct' label neuron spike first while delaying the spikes of the other label neurons. Multiple recordings for each trace show the variation due to the analog nature of the circuitry, but nevertheless the algorithm succeeds in training. Credit: Goltz et al.

The findings could have important implications for both research and development. In addition to informing further studies, they could, in fact, pave the way toward the development of faster and more efficient neuromorphic computing tools.

"With respect to information processing in the brain, one longstanding

question is: Why do neurons in our brains communicate with spikes? Or in other words, why has evolution favored this form of communication?" M. A. Petrovici, the senior researcher for the study, told TechXplore. "In principle, this might simply be a contingency of cellular biochemistry, but we suggest that a sparse and fast spike-based information processing scheme such as ours provides an argument for the functional superiority of spikes."

The researchers also evaluated their framework in a series of systematic robustness tests. Remarkably, they found that their model is well-suited for imperfect and diverse neural substrates, which would resemble those in the human cortex, where no two neurons are identical, as well as hardware with variations in its components.

"Our demonstrated combination of high speed and low power comes, we believe, at an opportune time, considering recent developments in chip design," Petrovici explained. "While on modern processors the number of transistors still increases roughly exponentially (Moore's law), the raw processing speed as measured by the clock frequency has stagnated in the mid-2000s, mainly due to the high power dissipation and the high operating temperatures that ariseas a consequence. Furthermore, modern processors still essentially rely on a von-Neumann architecture, with a central processing unit and a separate memory, between which information needs to flow for each processing step in an algorithm."

In neural networks, memories or data are stored within the processing units themselves; that is, within neurons and synapses. This can significantly increase the efficiency of a system's information flow.

As a consequence of this greater efficiency in information storage and processing, the framework developed by this team of researchers consumes comparatively little power. Therefore, it could prove particularly valuable for edge computing applications such as

nanosatellites or wearable devices, where the available power budget is not sufficient to support the operations and requirements of modern microprocessors.

So far, Goeltz, Kriener, Petrovici and their colleagues ran their framework using a platform for basic neuromorphic research, which thus prioritizes model flexibility over efficiency. In the future, they would like to implement their framework on custom-designed neuromorphic chips, as this could allow them to further improve its performance.

"Apart from the possibility of building specialized hardware using our design strategy, we plan to pursue two further research questions," Goeltz said. "First, we would like to extend our neuromorphic implementation to online and embedded learning."

For the purpose of this recent study, the network developed by the researchers was trained offline, on a pre-recorded dataset. However, the team would like to also test it in real-world scenarios where a computer is expected to learn how to complete a task on the fly by analyzing online data collected by a device, robot or satellite.

"To achieve this, we aim to harness the plasticity mechanisms embedded on-chip," Goeltz explained. "Instead of having a host computer calculate the synaptic changes during learning, we want to enable each synapse to compute and enact these changes on its own, using only locally available information. In our paper, we describe some early ideas towards achieving this goal."

In their future work, Goeltz, Kriener, Petrovici and their colleagues would also like to extend their framework so that it can process spatiotemporal data. To do this, they would need to also train it on time-varying data, such as audio or video recordings.

"While our model is, in principle, suited to shape the spiking activity in a network in arbitrary ways, the specific implementation of spike-based error propagation during temporal sequence learning remains an open research question," Kriener added.

More information: J. Göltz et al, Fast and energy-efficient neuromorphic deep learning with first-spike times, *Nature Machine Intelligence* (2021). DOI: 10.1038/s42256-021-00388-x

Steve K. Esser et al, Backpropagation for energy-efficient neuromorphic computing. *Advances in neural information processing systems*(2015). papers.nips.cc/paper/2015/hash … d4ac0e-Abstract.html

Sebastian Schmitt et al, Neuromorphic hardware in the loop: Training a deep spiking network on the brainscales wafer-scale system. *2017 international joint conference on neural networks (IJCNN)*(2017). DOI: 10.1109/IJCNN.2017.7966125

Citation: A framework to enhance deep learning using first-spike times (2021, October 5) retrieved 27 April 2024 from https://techxplore.com/news/2021-10-framework-deep-first-spike.html