Osvaldo Simeone, Bipin Rajendran, André Grüning, Evangelos S. Eleftheriou,
Mike Davies, Sophie Deneve, and Guang-Bin Huang

Learning Algorithms and Signal Processing for Brain-Inspired Computing

The success of artificial neural networks (ANNs) in carrying out various specialized cognitive tasks has brought renewed efforts to apply machine learning (ML) tools for economic, commercial, and societal aims, while also raising expectations regarding the advent of an artificial "general intelligence" [1]–[3]. Recent highly publicized examples of ML breakthroughs include the ANN-based algorithm AlphaGo, which has proven capable of beating human champions at the complex strategic game of Go. The emergence of a new generation of ANN-based ML tools has built upon the unprecedented availability of computing power in data centers and cloud computing platforms. For example, the AlphaGo Zero version required training more than 64 GPU workers and 19 CPU parameter servers for weeks, with an estimated hardware cost of US$25 million [4]. OpenAI's video game–playing program needed training for an equivalent of 45,000 years of game play, costing millions of dollars in rent access for cloud computing services [2].

Recent studies have more generally quantified the requirements of ANN-based models in terms of energy, time, and memory consumption in both the training and inference (run-time) phases. An example is a recent work by researchers from the University of Massachusetts Amherst [5], which concluded that training a single ANN-based ML model can emit as much carbon as five cars during their lifetimes.

The massive resource requirements of ANN-based ML raise important questions regarding the accessibility of the technology to the general public and to smaller businesses. Furthermore, they pose an important impediment to deploying powerful ML algorithms on low-power mobile or embedded devices.

The importance of developing suitable methods to implement low-power artificial intelligence on mobile and embedded devices is attested by its central role in applications such as digital health, the tactile Internet, smart cities, and smart homes. In light of this, key industrial players, including Apple, Google, Huawei, and IBM, are investing in developing new chips optimized for streaming matrix arithmetic that promise to make ANN-based inference more energy efficient through complexity-reduction techniques such as quantization and pruning [6].

Neuromorphic, or brain-inspired, computing

In contrast to ANNs, the human brain is capable of performing more general and complex tasks at a miniscule fraction of the power, time, and space required by state-of-the-art supercomputers. An emerging line of work, often collectively called *neuromorphic computing*, aims at uncovering novel computational frameworks that mimic the operation of the brain, in a quest for orders-of-magnitude improvements in terms of energy efficiency and resource requirements.

The unmatched efficiency of the human brain as an adaptive learning and inference machine may be the result of a number of unique factors. Among these, none appears to be more fundamental, and more fundamentally different from the operation of digital computer, than the way in which neurons encode information: *with* time, rather than merely *over* time [7]. Biological neurons can be thought of as complex dynamic systems with internal analog dynamics that communicate through the timing of all-or-nothing—and hence digital—spikes. This is in stark contrast to the static analog operation of neurons in an ANN. Biological neurons are connected through networks characterized by large fan-out, feedback, and recurrent signaling paths, unlike the feedforward or chain-like recurrent architectures of ANNs. As studied in theoretical neuroscience, the sparse, dynamic, and event-driven operation of biological neurons makes it possible to implement complex online adaptation and learning mechanisms via local synaptic plasticity rules and minimal energy consumption.

Based on these observations, brain-inspired neuromorphic signal processing and learning algorithms and hardware platforms have recently emerged as low-power alternatives to energy-hungry ANNs. Unlike conventional neural networks, spiking neural networks (SNNs) are trainable dynamic systems that make

Digital Object Identifier 10.1109/MSP.2019.2935557
Date of current version: 29 October 2019

use of the temporal dimension, not just as a neutral substrate for computing but also as a way to encode and process information in the form of asynchronous spike trains. In SNNs, sparse spiking and hence time-encoded signals carry out interneuron communications and intraneuron computing.

This has motivated the development of prototype neuromorphic hardware platforms that are able to process time-encoded data. These platforms include IBM's TrueNorth, SpiNNaker, developed within the "Human Brain Project"; Intel's Loihi; and proof-of-concept prototypes based on nanoscale memristive devices. These systems are typically based on hybrid digital–analog circuitry and in-memory computing, and they have already provided convincing proof-of-concept evidence of the remarkable energy savings that can be achieved with respect to conventional neural networks. Furthermore, SNNs have the unique advantage of being able to natively process spiking data as it is produced by emerging audio and video sensors inspired by biology, such as silicon cochleas or dynamic vision sensor cameras.

The role of signal processing in neuromorphic computing

Work on neuromorphic computing has been carried out by researchers in ML, computational neuroscience, and hardware design, often in parallel. While the problems under study—regression, classification, control, and learning—are central to signal processing, the signal processing community, by and large, has not been involved in the definition of this emerging field. Nevertheless, with the increasing availability of neuromorphic chips and platforms, the guest editors believe that progress in the field of neuromorphic computing calls for an interdisciplinary effort by researchers in signal processing in concert with researchers in ML, hardware design, system design, and computational neuroscience.

From a signal processing perspective, the specific features and constraints of neuromorphic computing platforms open interesting new problems concerning regression, classification, control, and learning. In particular, SNNs consist of asynchronous distributed architectures that process sparse binary time series by means of local spike-driven computations, local or global feedback, and online learning. Ideally, they are characterized by a graceful degradation in performance as the network's number of spikes, and hence, the energy usage, increases. For example, recent work has shown that SNNs can obtain satisfactory solutions of the sparse regression problem much more quickly than conventional iterative algorithms [8]. Solutions leverage tools that are well known to signal processing researchers, such as variational inference, nonlinear systems, and stochastic gradient descent.

In this issue

The field's scope encompasses neuroscience, hardware design, and ML, which makes it difficult for a nonexpert to find a suitable entry point in the literature. This special issue brings together key researchers in this area to provide readers of *IEEE Signal Processing Magazine* with up-to-date and survey-style articles on algorithmic, hardware, and neuroscience perspectives on the state-of-the-art aspects of this emerging field.

Overview

The first special issue article, "The Importance of Space and Time for Signal Processing in Neuromorphic Agents," by Indiveri and Sandamirskaya, introduces the role of time-encoded information and parallel neuromorphic computing architectures in enabling more efficient learning agents as compared to state-of-the-art ANNs.

Sensing and time-encoded representations

Neuromorphic computing architectures take as inputs time-encoded signals that are either produced by neuromorphic sensors or converted from natural signals such as images, video, or audio. The next two articles in this special issue describe these two scenarios. In "Event-Driven Sensing for Efficient Perception" by Liu et al., the authors discuss the main properties of the data produced by neuromorphic sensors and show how these features enable energy-efficient, low-latency, and real-time computing on neuromorphic platforms. In their article, "Signal Processing Foundations for Time-Based Signal Representations," Sevüktekin et al. discuss signal processing foundations for time-based signal representations of exogenous signals and for the reconstruction of these signals from their time-encoded versions.

Learning and signal processing applications

Neuromorphic platforms can be trained to carry out a variety of inference and control tasks. The next set of articles review training algorithms and applications. In "Surrogate Gradient Learning in Spiking Neural Networks," Neftci, Mostafa, and Zenke review training algorithms for standard deterministic models of SNNs via surrogate gradient methods, which aim at overcoming the nondifferentiability of the relevant loss functions. The next article, "An Introduction to Probabilistic Spiking Neural Networks," by Jang et al., discusses an alternative solution that is based on the use of probabilistic models by reviewing the resulting learning rules and applications. To further reduce the complexity of training, reservoir-computing techniques have been proposed, which are based on adapting only a subset of weights while others are randomly selected. Soures and Kudithipudi next present an overview of this topic in "Spiking Reservoir Networks." Finally, in "Neuroscience-Inspired Online Unsupervised Learning Algorithms," Pehlevan and Chklovskii focus on the special class of unsupervised learning algorithms, for which they provide a principled derivation of similarity-based local learning rules that are applied to problems such as linear dimensionality reduction, sparse or nonnegative feature extraction, and blind nonnegative source separation.

Hardware platforms

Standard computing systems based on the von Neumann architecture are not well suited to harness the efficiency of computing in SNNs. In "Low-Power Neuromorphic Hardware for Signal

Processing Applications," Rajendran et al. review architectural and system-level design aspects that underlie the operation of neuromorphic computing platforms for efficient implementation of SNNs.

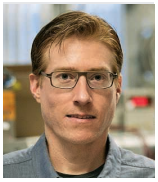
Guest Editors

Osvaldo Simeone (osvaldo.simeone@kcl.ac.uk) received his M.Sc. degree (with honors) and Ph.D. degree in information engineering from the Politecnico di Milano, Italy, in 2001 and 2005, respectively. He is a professor of information engineering with the Centre for Telecommunications Research, Department of Engineering, King's College London. From 2006 to 2017, he was a faculty member at the New Jersey Institute of Technology. He is a corecipient of the 2019 IEEE Communication Society Best Tutorial Paper Award, the 2018 IEEE Signal Processing Best Paper Award, the 2017 Best Paper by the *Journal of Communication and Networks,* the 2015 IEEE Communication Society Best Tutorial Paper Award, and the IEEE International Workshop on Signal Processing Advances in Wireless Communications 2007 and IEEE Conference on Wireless Rural and Emergency Communications 2007 Best Paper Awards. He was awarded a Consolidator Grant by the European Research Council in 2016. He is a Fellow of the IEEE and of the Institution of Engineering and Technology.

Bipin Rajendran (bipin@njit.edu) received his B.Tech. degree in instrumentation from the Indian Institute of Technology Kharagpur in 2000 and his M.S. and Ph.D. degrees in electrical engineering from Stanford University, California, in 2003 and 2006, respectively. He is an associate professor of electrical and computer engineering at the New Jersey Institute of Technology. He was a master inventor and research staff member at the IBM T.J. Watson Research Center, New York, from 2006 to 2012 and a faculty member in the Electrical Engineering Department, the Indian Institute of Technology Bombay, from 2012 to 2015. His research focuses on building scalable architectures and systems for neuromorphic computing using nanoscale devices. He has authored more than 75 papers in peer-reviewed journals and conferences and received 59 U.S. patents, four of which were awarded IBM's high-value patent award. His research was supported by the U.S. National Science Foundation, Semiconductor Research Corporation, and companies such as Intel and IBM. He is a Senior Member of the IEEE.

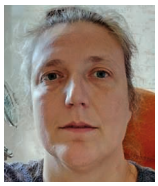
André Grüning (a.gruning@surrey.ac.uk) received his Ph.D. degree in computer science from the Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany, working with the Complex Systems Group. He is a senior lecturer in computational intelligence in the Department of Computer Science, University of Surrey, United Kingdom, as well a visiting researcher at the Institute of Physiology, University of Bern, Switzerland. He is a task leader in the H2020 E.U. flagship project "The Human Brain Project," where he and his team work on implementing biologicals for spiking neural networks on neuromorphic hardware. His research focuses on computational and cognitive neuroscience; he has more than 10 years of experience in designing and analyzing bioinspired learning algorithms for rate-based and spiking neural networks. He is a member of the European Institute for Theoretical Neuroscience, Paris, where he recently organized the "From Neuroscience to Machine Learning" international workshop.

Evangelos S. Eleftheriou (ele@zurich.ibm.com) received his diploma of engineering from the University of Patras, Greece, in 1979 and his M.Eng and Ph.D. degrees in electrical engineering from Carleton University, Ottawa, Canada, in 1985. He is currently responsible for the neuromorphic computing activities of IBM Research–Zürich. He has authored or coauthored more than 200 publications. He was a corecipient of the 2003 IEEE Communications Society Leonard G. Abraham Paper Award. He was also a corecipient of the Eduard Rhein Foundation 2005 Technology Award. In 2005, he was appointed an IBM fellow and was also inducted into the IBM Academy of Technology. In 2009, he was a corecipient of the IEEE Control Systems Society Control Systems Technology Award and the IEEE Transactions on Control Systems Technology Outstanding Paper Award. In 2016, he received an honoris causa professorship from the University of Patras, Greece. In 2018, he was inducted into the U.S. National Academy of Engineering as foreign member. He is a Fellow of the IEEE.



Mike Davies (mike.davies@intel.com) received his B.S. and M.S. degrees from the California Institute of Technology, Pasadena, in computer engineering, mathematics, and electrical engineering. Since joining Intel Labs in 2014, he has researched neuromorphic prototype architectures and is responsible for Intel's recently announced Loihi research chip. He leads Intel's Neuromorphic Computing Lab. His group's research interests span asynchronous circuits, fine-grain parallel computing architectures, neuromorphic algorithms, and software frameworks for event-driven distributed systems. He began his career in 2000 as a founding employee of Fulcrum Microsystems. As Fulcrum's director of silicon engineering, he pioneered high-performance asynchronous design methodologies as applied to several generations of industry-leading Ethernet switch products.

Sophie Deneve (sophie.deneve@ens.fr) received two B.S. degrees, in



mathematics and biology; an M.S. degree in cognitive science; and a Ph.D. degree from the University of Rochester. She is a director of research and group leader in the Department of Cognitive Science, Ecole Normale Supérieure (ENS), Paris. Her research interests focus on how biological neural circuits learn and solve probabilistic problems, as well as probabilistic approaches to human perception and psychiatry. Her recent results include networks learning to compute as efficiently and robustly as possible with spikes, entirely derived from an objective function combining errors in estimation (both in reproducing trained example and generalizing to new ones) and cost in number of spikes. She was awarded a Dorothy Hudgkin Fellowship from Royal Society London in 2004, a Marie Curie Team of Excellence fellowship in 2006, a MacDonnell Foundation Award in 2012, and a Brain and Human Cognition grant and a European Council Consolidator grant (from 2012 to 2017).

Guang-Bin Huang (egbhuang@ntu.edu.sg) received his B.Sc. and M.Eng. degrees from Northeastern University, Shenyang, China, and his Ph.D. degree from Nanyang Technological University, Singapore. He is a full professor in the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He is a principal investigator with the BMW-NTU Joint Future Mobility Lab on Human–Machine Interface and Assisted Driving; a principal investigator of data and video analytics with the Delta–NTU Joint Lab; a principal investigator of scene understanding with ST Engineering–NTU Corporate Lab; and a principal investigator for marine data analysis and prediction for autonomous vessels with the Rolls Royce–NTU Corporate Lab. He serves as an associate editor of *Neurocomputing*, *Cognitive Computation*, *Neural Networks*, and *IEEE Transactions on Cybernetics*. He has led or implemented several key industrial projects. He is a member of Elsevier's Research Data Management Advisory Board and of China's International Robotic Expert Committee.

References

[1] J. Brockman, *Possible Minds*. East Rutherford, NJ: Penguin Press, 2019.

[2] C. Metz, "With $1 billion from Microsoft, an A.I. lab wants to mimic the brain," *New York Times*, 2019. [Online]. Available: https://www.nytimes.com/2019/07/22/technology/open-ai-microsoft.html

[3] J. Lovelock, *Novacene: The Coming Age of Hyperintelligence*. East Rutherford, NJ: Penguin Press, 2019.

[4] AlphaGo Zero. [Online]. Available: https://en.wikipedia.org/wiki/AlphaGo

[5] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proc. 57th Annu. Meeting the Association Computational Linguistics*, 2019. [Online]. Available: http://arxiv.org/abs/1906.02243

[6] S. S. Sarwar, G. Srinivasan, B. Han, P. Wijesinghe, A. Jaiswal, P. Panda, A. Raghunathan, and K. Roy, "Energy efficient neural computing: A study of cross-layer approximations," *IEEE Trans. Emerg. Sel. Topics Circuits Syst.*, vol. 8, no. 4, pp. 796–809, Dec. 2018. doi: 10.1109/JETCAS.2018.2835809.

[7] F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek, *Spikes: Exploring the Neural Code*. (Computational Neurosciences series). Cambridge, MA: MIT Press, 1997.

[8] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, et al., "Loihi: A neuromorphic manycore processor with on-chip learning," *IEEE Microw.*, vol. 38, no. 1, pp. 82–99, 2018. doi: 10.1109/MM.2018.112130359.

SP