INTRODUCING ‘NEUROMORPHIC COMPUTING AND ENGINEERING’

Giacomo Indiveri

Institute of Neuroinformatics, University of Zurich and ETH Zurich, Zurich, Switzerland
* Author to whom any correspondence should be addressed.
E-mail: giacomo@ini.uzh.ch

Keywords: neuromorphic, intelligence, plasticity, interdisciplinary, spiking neural network

Abstract

The standard nature of computing is currently being challenged by a range of problems that start to hinder technological progress. One of the strategies being proposed to address some of these problems is to develop novel brain-inspired processing methods and technologies, and apply them to a wide range of application scenarios. This is an extremely challenging endeavor that requires researchers in multiple disciplines to combine their efforts and simultaneously co-design the processing methods, the supporting computing architectures, and their underlying technologies. The journal ‘Neuromorphic Computing and Engineering’ (NCE) has been launched to support this new community in this effort and provide a forum and repository for presenting and discussing its latest advances. Through close collaboration with our colleagues on the editorial team, the scope and characteristics of NCE have been designed to ensure it serves a growing transdisciplinary and dynamic community across academia and industry.

1. Introduction

As editor-in-chief I am pleased to announce the publication of the first content [1–6] in Neuromorphic Computing and Engineering (NCE). This editorial aims at motivating the need for creating such a journal by describing our view of some of the current challenges for information and communication technology (ICT) and the role that NCE can play to address them.

While Moore’s law has been fueling technological progress for decades, it is only in recent years that we could start to reap most of the benefits of this progress (see figure 1(a)): ICT is becoming pervasive and is affecting every facet of our daily lives. At a global scale, ICT is affecting almost all aspects of our society, ranging from global communication to education and health, from finance to automation, transportation, and climate change. With the digitalization of our society, computing technologies and electronic devices are producing increasing amounts of electronic data every year and people are getting access to larger and larger amounts of personalized information.

To cope with the demands that are emerging with this technological revolution, and to exploit the opportunities created with the availability of this data, novel data-science methods, machine-learning techniques, and artificial intelligence (AI) algorithms have emerged. AI algorithms typically employ neural networks and deep learning techniques to solve pattern recognition tasks and have been shown to be extremely successful in extracting information from large amounts of data [7, 8]. However, their training methods require massive amounts of data and computing resources, which in turn requires an amount of energy that is not sustainable with respect to both the global electricity supply and the carbon footprint of the computation. For example, it has been estimated that training a recent state-of-the-art AI neural network, such as GPT-3 [9], would take more than 27 years worth of processing time on a single computer, and that these computations would generate over 35000 kg of CO2 emissions in total, which is more than the average American adult produces in two years.¹ The reason for these exorbitant costs are due to the fact that current computing technologies, based on the classical von Neumann architectures, are not well matched to the parallel processing nature of neural networks [10]. On the other hand, biological brains clearly outperform AI systems in terms of the amount of power consumption requirements, the number of training data samples, and their ability to adapt to novel and

¹ From the Forbes article ‘deep learning’s carbon emissions problem’.
unexpected conditions. This is particularly true for those tasks that are still very difficult for computers and AI algorithms, but are done effortlessly by humans and animals, such as online learning, with small numbers of examples, interaction with the environment, or sensing and motor control.

A promising approach toward the development of novel computational paradigms and ICT systems that can interact intelligently with the environment, that could bridge the gap between natural intelligence and AI, and that could solve many of the open challenges facing the future of computing is the one pursued by the ‘neuromorphic computing and engineering’ field.

We are now at a very exciting time in which the convergence of the end or Moore’s law, with the renewed interest in neural networks, and the needs for low power and sustainable ‘green-AI’ are all pointing toward the huge potential that can come from research and development in neuromorphic computing and engineering.

2. NCE thematic areas

The term neuromorphic was originally coined by Carver Mead in the late 1980s at CALTECH, to describe very large scale integration (VLSI) systems containing electronic circuits that mimic neuro-biological architectures present in the nervous system [11, 12]. However, throughout the years, the term has ‘morphed’ its original meaning and started to be used to describe a broader set of concepts, and approaches. In addition to the original ‘neuromorphic engineering’ meaning, the term also started to be used to describe ‘neuromorphic computing’ systems that contain pure digital circuits or conventional processors, for simulating spiking neural networks and neural models of computation. In parallel, the same term started to be used to refer to systems comprising nano-scale ‘memristive’ devices, developed within the field of emerging memory technologies. Today, the term is also being used to refer to algorithmic and machine learning approaches that simulate biologically plausible and hardware friendly spiking neural networks and learning mechanisms.

2.1. Neural circuits and systems integrated in CMOS technology

The main goal of the original neuromorphic engineering approach was to directly emulate the physics of computation of biological neural networks, via the use of transistors operated in their ‘weak-inversion’ or ‘subthreshold’ regime [13–15]. This approach was aimed at building artificial neurons, synapses, networks, and sensory systems using the same organizing principles used by the nervous system in animal brains. This effort had the dual objective of both understanding neural computation by building physical emulations of real neural circuits, and developing compact and low power sensory processing and computing architectures radically different from the standard digital computers of the time. Given the high-risk and basic research aspect of this approach, there is only a small number of academic groups that are still pursuing it today. This community is mostly focused on the development of small-scale prototype chips to explore different aspects of neural computation, ranging from sensory systems [16–20] to reconfigurable networks with biologically plausible neural dynamics [21–24], to spike-based learning and plasticity circuits [25–29].
In more recent times the term neuromorphic was also adopted to describe mixed-signal and pure digital VLSI systems which implement computing platforms that could be used to simulate models of spiking neural networks. This effort was mainly driven by the possibility to exploit progress in computing and integrated circuit technology for building large-scale dedicated brain-inspired computing systems. For example, the EU Human Brain Project funded the development of wafer-scale integrated systems designed to faithfully reproduce simulations of neuroscience modeling studies comprising large numbers of neurons at accelerated speeds [30]. Similarly, the SpiNNaker system, also developed with the support of the Human Brain Project, is a multi-core computer designed with the goal of simulating very large numbers of spiking neurons in real time [31]. In its current state of development the SpiNNaker machine, built by stacking together 600 printed circuit boards each carrying 48 SpiNNaker processors, supports the simulation of hundreds of millions of neurons. An alternative strategy to scale up the size of simulated spiking neural networks is the one proposed by IBM, which in 2014 presented the ‘TrueNorth’ neuromorphic systems that integrated on the same chip 4096 cores, each comprising pure digital asynchronous circuits able to simulate 256 neurons with 256 × 256 synaptic connections [32]. This was a significant breakthrough in the field, as it demonstrated how advanced technology nodes, such as the one used of the Samsung 28 nm bulk CMOS process, could support the integration of very large numbers of silicon neurons, while keeping the overall power consumption extremely low (e.g., with an average consumption of 70 mW total power while running a typical recurrent network in biological real-time, four orders of magnitude lower than a conventional computer running the same network). The more recent Loihi chip built by Intel [33] has some common design choices with the IBM TrueNorth chip, in that it uses pure digital asynchronous circuits to simulate neurons and synapses. Taking advantage of the progress in integration technology, this chip was fabricated using the Intel 14 nm FinFET process. Rather than focusing on the large-scale, the Loihi designers chose to focus on more complex neural and synaptic features, including spike-based learning mechanisms. So the Loihi chip integrates ‘only’ 128 cores, and supports the simulation of networks containing up to 130 000 spiking neurons. Similarly to the SpiNNaker and the TrueNorth chips, Loihi is being used as a research platform to support the development of spike-based processing architectures that can be applied to the solution of practical problems. Indeed, researchers and students are encouraged to develop novel spike-based computing solutions using Loihi, via the support of the Intel Neuromorphic Research Community (INRC) program, which has been very successful in promoting the growth of the community and producing a range of promising results [34].

2.2. Memristive devices and emerging memory technologies

The material science and device physics community has been carrying out research on new materials and technologies for memory and long-term storage applications for many years. Recently however, this community started using the term ‘neuromorphic’ to refer to new devices and systems that, on one hand exhibit different types of behaviors that can be linked to those of biological synapses, and on the other represent important building blocks for the development of large-scale AI computing systems [35]. Indeed, shortly after the proposal of using these new nano-scale devices as ‘memristors’ [36], this community quickly embraced the idea that these devices could be used to implement synapses in artificial neural networks, and to store locally their synaptic weights [37]. These devices and technologies hold a great promise of enabling ‘in-memory computing’ in neural networks, and of supporting, through their physics, complex non-linear features that could be exploited to emulate many interesting properties of biological synapses. Within this context, a wide range of research efforts are being pursued, including the development of different types of non-volatile and volatile memristive devices, and the design of spike- or pulse-based control schemes for inducing biologically plausible learning behaviors in memristive cross-bar arrays [38, 39]. As there is not yet a single solution to the problem of finding the optimal artificial synapse, a broad range of materials, devices, and techniques are still active areas of investigation [40–46].

2.3. Algorithms and computational models of spike-based learning and inference

An important aspect of the neuromorphic computing and engineering domain is the one of hardware–software co-design. It is no surprise then that the term neuromorphic is also being used to describe research in computational models and algorithms that can be readily mapped onto memristive, CMOS, or hybrid memristive–CMOS neuromorphic architectures [47]. Most efforts in this domain focus either on exploring spike-based learning methods that approximate the backpropagation algorithm [48, 49], or on identifying local stochastic and complex non-linear plasticity mechanisms that can be reproduced by memristive devices or CMOS learning circuits [47, 50–53]. Very promising results are also being obtained by combining the latest advances in AI and machine learning algorithms with brain-inspired and computational neuroscience modeling efforts [54–57]. These investigations provide useful specifications for the design of new non-volatile and volatile memristive devices and for the design of spiking neural network chips that can use the principles
of computation derived from the brain to carry out low-power and robust computation using local learning rules and unreliable and low-precision components, such as the ones present in animal brains.

3. Open challenges

The most important challenge that NCE faces is that of supporting a diverse and interdisciplinary community of researchers working on different aspects of neuromorphic computing and engineering. This journal will be instrumental in aligning the goals and objectives of this community, and in promoting its growth (see figure 2). The success of NCE strongly depends on how much the experts in the different thematic areas will be willing to broaden their horizons, learn the ‘language’ of the experts of other thematic areas, and create synergies with them. This is indeed already happening to some degree: material science and device experts are collaborating with circuit designers to both extend current CMOS technologies and integrate nano-scale devices in newly developed ones; computer science and machine learning experts are collaborating with neuroscientists to develop brain-inspired processing theories that are compatible with the neuromorphic technologies being developed; electronic engineers, physicists and memory device experts are working closely with computational neuroscientists to implement devices and circuits that can emulate the biophysics of synapses and neurons in the electronic circuits and systems being developed.

By defining a common vision and by bridging the traditional boundaries between standard subjects, the NCE journal will enable the development and dissemination of breakthroughs in neuromorphic computing and engineering that will have disruptive effects, and which could revolutionize the nature of computation (see ‘thinking outside the box’ figure 2).

4. Conclusions and outlook

Neuromorphic computing and engineering is becoming an extremely important and timely research area. Basic research in brain-inspired neuromorphic electronic circuits has been carried out for a significant number of years. In parallel tremendous progress has been made in computational neuroscience, nanotechnologies, and machine learning. This is the best time to combine the know-how gained so far in these disciplines to obtain breakthroughs that can potentially solve many of the problems that ICT is starting to face. The NCE journal has been established to provide a new open access medium explicitly designed to support this convergence of efforts, and to bring together all researchers engaged in these areas in a way that unites the community and defines the field of ‘neuromorphic computing and engineering’ for years to come.

Data availability statement

No new data were created or analysed in this study.
References


[2] Leroux N et al 2021 Hardware realization of the multiply and accumulate operation on radio-frequency signals with magnetic tunnel junctions *Neuromorph. Comput. Eng.* 1 011001


[57] Lillicrap T P and Santoro A 2019 Backpropagation through time and the brain Curr. Opin. Neurobiol. 55 82–9 Machine Learning, Big Data, and Neuroscience