Neuromorphic Computing and Engineering

EDITORIAL • OPEN ACCESS

Editorial: Focus issue on energy-efficient neuromorphic devices, systems and algorithms

To cite this article: Adnan Mehonic *et al* 2023 *Neuromorph. Comput. Eng.* **3** 040201

NEUROMORPHIC
Computing and Engineering

OPEN ACCESS

RECEIVED
11 October 2023

ACCEPTED FOR PUBLICATION
25 October 2023

PUBLISHED
31 October 2023

EDITORIAL

Editorial: Focus issue on energy-efficient neuromorphic devices, systems and algorithms

Adnan Mehonic[1,*], Charlotte Frenkel[2] and Eleni Vasilaki[3]

1 Department of Electronic and Electrical Engineering, UCL, London, United Kingdom
2 Microelectronics Department, Delft University of Technology, Delft, The Netherlands
3 Department of Computer Science, The University of Sheffield, Sheffield, United Kingdom
* Author to whom any correspondence should be addressed.

E-mail: adnan.mehonic.09@ucl.ac.uk

Artificial intelligence (AI) is undergoing exponential growth and is being ubiquitously deployed across a myriad of sectors, including but not limited to, self-driving automotive technologies, robotics, automation, Internet of Things, healthcare solutions, security systems, and the entertainment industry. This widespread adoption is substantially contributing to the ever-increasing demand for computing power. Current estimates indicate that computational requirements are doubling approximately every two to three months [1].

The vast energy consumption of AI technologies is a focal point of concern. For instance, the energy consumed by AlphaGo during a single game of Go astonishingly exceeds the total energy required to manufacture and fuel a car over its entire lifespan [2]. This voracious appetite for power is accentuating cost barriers in AI development. Lambdalabs estimated in 2020 that training the GPT-3 model on a singular GPU would take a staggering 355 years and cost around $4.6 million [3]. Training its successor, GPT-4, is projected to cost upwards of $100 million [4].

Conversely, there is a parallel trend toward the proliferation of edge computing devices that incorporate localized intelligence. Unlike traditional models that are highly dependent on cloud computing, these emerging devices necessitate that AI algorithms operate with significantly lower energy consumption, in the sub-watt range, and even down to tens of microwatts at extreme edge locations. One of the key research domains focusing on this challenge encompasses the development of energy-efficient nanoelectronic devices, novel neuromorphic architectures, and innovative algorithmic approaches.

The term 'neuromorphic' is interpreted differently across various academic and industrial communities [5, 6]. However, the common denominator among these perspectives is the aspiration to emulate brain-like functionalities in computing hardware. This includes in-hardware learning, spike-based processing, temporal and spatial locality, resilience to environmental noise, adaptability, asynchronous communication, and analogue processing. These features are largely absent in today's mainstream computing technologies. The unique blend of these characteristics in neuromorphic systems paves the way for exploiting intricate and multifaceted physics. A key enabler for this exploitation is foundational research at the materials and device levels, which allows unlocking new dimensions in the capabilities and applications of computing technologies.

In this special issue, we are pleased to feature three cutting-edge articles that delve into the realm of post-CMOS technologies, with a particular focus on various types of memristors and their applications in energy-efficient computing. These articles collectively illuminate the rapidly evolving landscape of next-generation computing architectures, showcasing the tremendous potential of memristors as viable alternatives or supplements to traditional CMOS technologies.

The first study [7], from the IBM Zurich group, investigates the challenges and opportunities of using in-memory computing for efficient computational tasks. While in-memory computing has shown promise, it suffers from low computational accuracy due to device variability. The paper explores bit slicing as a technique to improve this accuracy. Using models of phase-change memory arrays (type of memristor technology), the study finds that bit slicing in this context should focus on minimizing error through averaging within a specific dynamic range, rather than extending the dynamic range as in digital processors. The findings are validated through experiments using a prototype chip, and the impact on neural network inference accuracy is assessed using CIFAR-10 and ImageNet benchmarks. The second study [8] comes from research groups in Aachen and Julich and explores the same topic of in-memory computation but uses a different type of memristive technology—resistance switches based on valence change mechanisms (VCMs). When employed for vector-matrix multiplications, VCM devices exhibit non-idealities that impact the performance of vector-matrix multiplications and, consequently, the overall accuracy of the application. These non-idealities can be either time-independent, such as programming variability, or time-dependent, like read noise. The paper conducts both empirical and theoretical analyses of these issues, concluding that variability in the low resistive state plays a key role. It also recommends reading in the RESET direction as preferable to the SET direction. The third paper in this series [9], authored by groups from Peking University, the Chinese Institute for Brain Research, and the Beijing Academy of Artificial Intelligence, focuses on using memristive devices for the processing of unlabeled data. This is a critical area in AI due to the scarcity of well-structured labeled data. The paper reviews the potential of memristive devices to efficiently implement various algorithms used for unlabeled data analysis, such as *k*-means clustering and restricted Boltzmann machines. It provides an overview of design principles and applications of these devices in neuromorphic computing and other AI tasks related to processing unlabeled data. The final paper [10] in the issue, coming from another two groups in Julich and Aachen, utilizes a high-end AMD CPU to explore full-scale simulations of neuronal network models of the brain. The paper reports that these simulations can run in less time than the biological timespan they aim to model. This finding is particularly significant for applications in robotics and closed-loop systems, where real-time performance is crucial. The study is valuable not only for understanding learning and development processes in the brain, which can extend over longer biological periods, but also for furthering the development of neuromorphic technologies.

While many questions remain unanswered in the field of neuromorphic technologies, it is clear that diverse research communities are converging to explore this intriguing domain from various angles. This includes work on materials, devices, circuits, systems, and algorithms—all of which are crucial and call for a comprehensive, multidisciplinary approach. We hope this special issue serves as an inspiration to both readers and researchers interested in this field.

Data availability statement

No new data were created or analysed in this study.

ORCID iDs

Adnan Mehonic https://orcid.org/0000-0002-2476-5038
Eleni Vasilaki https://orcid.org/0000-0003-3705-7070

References

[1] Mehonic A and Kenyon A J 2022 Brain-inspired computing needs a master plan *Nature* **604** 255–60
[2] van Wynsberghe A 2021 Sustainable AI: AI for sustainability and the sustainability of AI *AI Ethics* **1** 213–8
[3] Li C 2020 (available at: https://lambdalabs.com/blog/demystifying-gpt-3) (accessed 25 August 2023)
[4] WILL KNIGHT 2023 (available at: www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/) (accessed 25 August 2023)
[5] Mehonic A and Eshraghian J 2023 Brains and bytes: trends in neuromorphic technology *APL Mach. Learn.* **1** 020401
[6] Frenkel C, Bol D and Indiveri G 2023 Bottom-up and top-down approaches for the design of neuromorphic processing systems: tradeoffs and synergies between natural and artificial intelligence *Proc. IEEE* **111** 6
[7] Le Gallo M, Nandakumar S R, Ciric L, Boybat I, Khaddam-Aljameh R, Mackin C and Sebastian A 2022 Precision of bit slicing with in-memory computing based on analog phase-change memory crossbars *Neuromorph. Comput. Eng.* **2** 014009
[8] Bengel C, Mohr J, Wiefels S, Singh A, Gebregiorgis A, Bishnoi R, Hamdioui S, Waser R, Wouters D and Menzel S 2022 Reliability aspects of binary vector-matrix-multiplications using ReRAM devices *Neuromorph. Comput. Eng.* **2** 034001
[9] Xiao Z, Yan B, Zhang T, Huang R and Yang Y 2022 Memristive devices based hardware for unlabeled data processing *Neuromorph. Comput. Eng.* **2** 022003
[10] Kurth A C, Senk J, Terhorst D, Finnerty J and Diesmann M 2022 Sub-realtime simulation of a neuronal network of natural density *Neuromorph. Comput. Eng.* **2** 021001