Neuromorphic Artificial Intelligence

Prof. Jean Anne Incorvia, The University of Texas at Austin, USA

Information processing machines, driven by microelectronics, have enabled the global computing and communication revolutions and the information age. As we look to the future of computing, we see challenges in continuing to do high-performance computing, such as increasing the density of our computing chips, improving their energy efficiency, and efficiently removing heat from the computing system. At the same time, when looking to the future, it is clear that computing is rapidly branching out from the traditional laptop, cell phone, or data center, and that it is the cornerstone of advanced artificial intelligence (AI).

Computers for AI have vastly different requirements from computers in other areas such as data centers or supercomputers. AI computers require high energy efficiency to operate in *edge* devices, since they need to operate in a small footprint with often limited battery life. These edge devices could be anything from a robot, to a smart car, to satellites in space collecting astronomy data. As such, AI computers must be able to be immersed in harsh conditions, such as different weather, high radiation, or changes in temperature. The computers need to be adaptive to sensory inputs, so the AI can respond in real-time to the environment, a requirement called *online learning*. The computers therefore need to efficiently process sensory inputs, known as *in-sensor computing*. AI computers also need to be able to do local computation, to avoid the energy cost of excessive back-and-forth between the edge device and the computing cloud (called the *analog-to-digital bottleneck*).

It is obvious that such AI computers have vastly different needs from traditional computing. One obvious but grand example of a computer that satisfies many of these requirements is the human or mammalian brain. Thus, in looking to the future, we can take inspiration from the brain and how it functions to design brain-inspired computers for AI.

2022 celebrates 75 years since the invention of the transistor. We can consider the development of the traditional computer as akin to walking inside a tunnel. In the example of the compute function, this tunnel was refined over these years through tight co-design between the material (silicon), device (field-effect transistor, FET), and architecture (von Neumann) [1]. With a new application of AI ahead of us, silicon transistors and how they are arranged in circuits may no longer be the best candidates. AI will require its own tight co-design, and we can imagine in 75 more years there may be new device(s), using new materials, connected in new ways into circuits for AI. Thus, AI is a wide open field for exploring how new physics, new materials, new devices, and new architectures can be used.

Neuromorphic artificial intelligence combines these two desires of a) mimicking the brain and b) using application-specific materials and devices. This is a large task, since the brain is still not well understood, and which specific brain-like features that should be integrated into an artificial computer is an open question. Today, at minimum, we aim to design artificial neurons that can mimic the basic features of neurons in the brain: they can act as an activation function that takes in an input and produces a non-linear output. Similarly, we can design artificial synapses that can mimic the basic features of synapses in the brain: they can act as weights that determine the connectivity between the neurons. The synapses and neurons can then be arranged in a circuit, e.g. as a layer of a neural network.

Going forward, there are many open research questions and opportunities to explore this space between computer engineering, electrical engineering, neuroscience, materials science, and physics. For example, what higher-order features of neurons and synapses are important for learning, and how can these be applied to AI? As an example answer, my research group has worked on implementing asymmetric synapse behavior, called metaplasticity, by leveraging the properties of magnetic and low-dimensional materials [2-3]. We showed this brain-inspired, device-level behavior can have system-level benefits by allowing an AI to learn new data without forgetting previous data it has been trained to classify. Another example question is, how important are temporal effects in the brain, and how could those be leveraged for AI? Similarly, we know the brain using imperfect devices (neurons) to perform complex functions. How can such stochasticity and randomness be leveraged for AI [4-5]? I will touch on some directions to answer each of these questions.

References:

[1] S. Guha, H. -S. P. Wong, J. A. C. Incorvia, and S. Chowdhury. Future Directions Workshop: Materials, Processes, and R&D Challenges in Microelectronics. US Department of Defense Basic Research Office (2022). https://basicresearch.defense.gov/Programs/Future-Directions-Workshops/
[2] T. Leonard, S. Liu, M. Alamdar, H. Jin, C. Cui, O. G. Akinola, L. Xue, T. P. Xiao, J. S. Friedman, M. J. Marinella, C. H. Bennett, and J. A. C. Incorvia. "Shape-Dependent Multi-Weight Magnetic Artificial Synapses for Neuromorphic Computing." *Advanced Electronic Materials,* 2200563 (2022). <DOI 10.1002/aelm.202200563>.
[3] D. Kireev*, S. Liu*, H. Jin, T. P. Xiao, C. H. Bennett, D. Akinwande, and J. A. C. Incorvia. "Metaplastic and Energy-Efficient Biocompatible Graphene Artificial Synaptic Transistors for Enhanced Accuracy Neuromorphic Computing." *Co-first authors. *Nature Communications* 13, 4386 (2022). <DOI 10.1038/s41467-022-32078-6>.
[4] S. Misra, L. C. Bland, S. G. Cardwell, J. A. C. Incorvia, C. D. James, A. D. Kent, C. D. Schuman, J. D. Smith, and J. B. Aimone. "Probabilistic Neural Computing with Stochastic Devices." *Advanced Materials* 2204569 (2022).
[5] S. Liu, T. P. Xiao, J. Kwon, B. J. Debusschere, S. Agarwal, J. A. C. Incorvia, and C. H. Bennett. "Bayesian Neural Networks Using Magnetic Tunnel Junction-Based Probabilistic In-Memory Computing." *Frontiers in Nanotechnology* 4: 1021943 (2022). <DOI 10.3389/fnano.2022.1021943>.