Program of the CNRS AISSAI Workshop on

Energetics of computation in artificial and natural networks

Ecole Normale Supérieure, Paris

December 15 and 16, 2022

Thursday - Salle Dussane, Ecole Normale Supérieure, 45 rue d'Ulm, Paris 5th

09:00-09:15   *Welcome and introductory remarks*

09:15-10:15   **Nathalie Rochefort (Univ. of Edimburgh)**
Paying the brain's energy bill

10:15-10:45   *Coffee Break*

10:45-11:45   **Damien Querlioz (Center for Nanoscience and Nanotechnology, Paris-Saclay)**
Toward Energy-Efficient Artificial Intelligence with Brain Inspiration

12:00-13:00   *Colloquium Data Science*
Wolfram Pernice (Univ. of Münster)
Computing beyond Moore's law with photonic hardware

13:00-14:30   *Lunch (on site)*

14:30-15:30   **Pierre-Yves Plaçais (ESPCI, PSL)**
Energy availability at the cellular and systemic levels controls memory dynamics in the Drosophila brain

15:30-16:30   **Eva Garcia-Martin (Ekkono Solutions)**
Edge machine learning: theory vs. application

16:30-17:00   *Coffee Break*

17:00-18:00   **Arvind Murugan (Univ. of Chicago)**
Energy dissipation cost of learning and computation in physical systems

Friday - Site Ulm, Collège de France, 3 rue d'Ulm, Paris 5th

09:00-10:00 **A. Mizrahi (CNRS-Thales, Paris-Saclay)**
Neural networks with radiofrequency spintronic nanodevices

10:00-10:30 *Coffee Break*

10:30-11:30 **Julian Götzl (Univ. of Heidelberg and Bern)**
From biology to silicon substrates: neural computation with physics

11:30-12:30 **Shahar Kvatinsky (Technion)**
On-Device Machine Learning with Memristors in the Neuromorphic Era

12:30-14:00 *Lunch (on your own)*

14:00-15:00 **Renaud Jolivet (Univ. of Genève)**
Energy-efficient information transfer in brain circuits

15:00-16:00 **Sylvain Gigan (ENS, PSL)**
Exploiting light scattering for optical computing

16:00-16:30 *Coffee Break*

16:30-18:00 **Round table animated by R. Zecchina (Univ. Bocconi)**

Abstracts

Paying the brain's energy bill

How have animals managed to maintain metabolically expensive brains given the volatile and fleeting availability of calories in the natural world? I will first review studies in support of three strategies that involve: 1) a reallocation of energy from peripheral tissues and functions to cover the costs of the brain, 2) an implementation of energy-efficient neural coding, enabling the brain to operate at reduced energy costs, and 3) efficient use of costly neural resources during food scarcity. I will then present a recent study from the lab showing metabolic state-dependent mechanisms by which the mammalian cortex regulates coding precision to preserve energy in times of food scarcity.

Altogether, these results reveal energy-saving mechanisms that make energy-costly brains fit for survival.

Toward Energy-Efficient Artificial Intelligence with Brain Inspiration

When performing artificial intelligence (AI) tasks, computers consume considerably more energy for moving data between logic and memory units than for doing actual arithmetic. This inefficiency leads to the unsustainable energy cost of AI: training modern AI models requires gigawatt-hours of electricity. Brains, by contrast, achieve superior energy efficiency by fusing logic and memory entirely, performing a form of "in-memory" computing. Until now such an integration between logic and memory was impossible at a large scale using conventional CMOS technology. However, companies such as Intel, Samsung, ST Microelectronics, or TSMC, have recently reached production status on new memory devices such as (mem)resistive, phase change, and magnetic memories, which give us an opportunity to achieve an extremely tight integration between logic and memory. Unfortunately, these new devices also come with important challenges due to their unreliable nature. In this talk, we will look at neuroscience inspiration to extract lessons on the design of in-memory computing systems with unreliable devices. We will first study the reliance of brains on approximate memory strategies, which can be reproduced for machine learning. We will give the example of a memristor-based Bayesian machine. Based on measurements on a hybrid CMOS/memristor integrated system, we will see that such a system can recognize human gestures using thousands times less energy than a competing microcontroller unit. Then, we will present a second approach where the probabilistic nature of emerging memories, instead of being mitigated, can be fully exploited to implement a type of probabilistic learning. We train experimentally an array of 16,384 memristors to recognize images of cancerous tissues using this technique. Finally, we will present prospects concerning the implementation of different learning algorithms with emerging memories.

Computing beyond Moore's law with photonic hardware

Conventional computers are organized around a centralized processing architecture, which is well suited to running sequential, procedure-based programs. Such an architecture is inefficient for computational models that are distributed, massively parallel and adaptive,

most notably those used for neural networks in artificial intelligence. In these application domains demand for high throughput, low latency and low energy consumption is driving the development of not only new architectures, but also new platforms for information processing.

Photonic circuits are emerging as one promising candidate platform and allow for realizing the underlying computing architectures, which process optical signals in analogy to electronic integrated circuits. Therein electrical connections are replaced with photonic waveguides, which guide light to desired locations on chip. Through heterogeneous integration, photonic circuits, which are normally passive in their response, are able to display active functionality and thus provide the means to build neuromorphic systems capable of learning and adaptation. In reconfigurable photonic architectures in-memory computing allows for overcoming separation between memory and central processing unit as a route for designing artificial neural networks, which operate entirely in the optical domain.

Energy availability at the cellular and systemic levels controls memory dynamics in the Drosophila brain

Contrary to modern, man-made machines, brains – and more generally nervous systems- were constrained to evolve in a context of limiting energy availability. An intriguing question is thus to investigate how this translates into brain function, even in well-fed animals and regular physiological state.

By studying the dynamics of olfactory memory formation in the fruitfly *Drosophila melanogaster*, we uncovered several principles that rule the interplay between energy metabolism and memory, and described their underlying molecular mechanisms:

 1) competition between metabolic pathways: glial cells, the non-neuronal cells of the brain, provide metabolites to neurons that fuel memory formation. I will describe how hormonal inhibition of glial metabolic pathways that would compete with neuronal fueling is triggered by learning to allow for proper memory encoding.

2) upregulation of energy fluxes as a gating process: formation of long-term memory involves a high metabolic cost. Whether the fly's memory brain center will engage into long-term memory formation following a learning experience is governed by a dopamine signaling that activates mitochondrial energy flux in neurons. This metabolic regulation, which occurs very early after learning, is the limiting step that sorts the learning experiences to be consolidated into long-term memory.

3) fuel flexibility: in regular conditions, the fueling of brain activity essentially relies on glucose metabolism. Upon food shortage, adaptive mechanisms are implemented to prevent long-term memory formation, which is beneficial for survival. However, less persistent memories can be formed in replacement, involving a distinct source of energy for neurons, namely ketone bodies derived from glial lipid stores.

Altogether, these findings reveal simple design principles that allow adaptive memory encoding in an energy-efficient way.

Edge Machine Learning: Theory vs Applications

Edge Machine learning has started to gain track in industrial and research applications. In this talk, we explore the key differences between research in edge machine learning and real world industrial IoT applications, with a specific focus on energy efficiency, memory constraints, embedded devices, and data availability.

Energy dissipation cost of learning and computation in physical systems

Since Landauer, we have known in the abstract that information processing, and in particular the erasure of memory, costs energy. In recent decades, this principle has provided insight into how specific biomolecular processes spend energy to process information. We review work on some of these foundational processes that range from correcting errors, measuring with higher precision or computing derivatives. We also relate such Landauer energy costs to the energy cost of contrastive learning in a physical system.

Neural networks with radiofrequency spintronic nanodevices

We use spintronic nanodevices to implement neurons and synapses and build artificial neural networks directly in hardware. These devices are CMOS-compatible, small, fast and low energy consumption. Furthermore, they can emit and receive radiofrequency signals, that we can use to achieve dense connectivity. Finally, their rich non-linear dynamics could be leveraged to implement on-chip learning.

From biology to silicon substrates: neural computation with physics

Whether biological or artificial, intelligence ultimately boils down to the ability of physical substrates to perform (complex) computations efficiently. Understanding intelligence thus requires overcoming a set of interrelated - and interdisciplinary - challenges. Starting from a subset of biological dynamics that computational neuroscience identifies as relevant for computation, we need to come up with compatible models for coding, computation and learning. Once these building blocks have been identified, we would like to find efficient in-silico implementations, be that in classical devices or custom-engineered novel forms of hardware.
In my talk, I will discuss some of our answers to certain subsets of these grand challenges. In particular, I will address the following questions:

- How can cortical hierarchies perform credit assignment?
- How can deep networks learn precise spike timing?
- How can we emulate such neuro-synaptic dynamics in silico?

On-Device Machine Learning with Memristors in the Neuromorphic Era

Memristive technologies are attractive candidates to replace conventional memory technologies and can also be used to combine data storage and computing to enable novel

non-von Neumann architecture. One such non-von Neumann architecture is neuromorphic computing, where brain-inspired circuits are built for massive parallelism and in-place computing.

This talk focuses on neuromorphic computing with memristors at the edge. I will show how we can get inspiration from the brain to build electronic circuits that are energy efficient and perform both inference and training extremely fast and efficiently. We will see that this approach can be used not only to accelerate machine learning applications but also for novel mixed-signal circuits and for near-sensor processing.

Energy-efficient information transfer in brain circuits

The nervous system consumes a disproportionate fraction of the resting body's energy production. In humans, the brain represents 2% of the body's mass, yet it accounts for ~20% of the total oxygen consumption. Expansion in the size of the brain relative to the body and an increase in the number of connections between neurons during evolution underpin our cognitive powers and are responsible for our brains' high metabolic rate. Despite the significance of energy consumption in the nervous system, how energy constrains and shapes brain function is often under-appreciated. I will illustrate the importance of brain energetics and metabolism, and discuss how the brain trades information for energy savings in the visual pathway. Indeed, a significant fraction of the information those neurons could transmit in theory is not passed on to the next step in the visual processing hierarchy. I will discuss how this can be explained by considerations of energetic optimality. Finally, I will briefly discuss how energetic constraints might impact coding strategies in neural networks and how this provides an elegant approach for a more holistic view of brain circuits.

Exploiting light scattering for optical computing

Optics, be it in integrated photonics circuits or with free space transformation, is a promising platform for neuromorphic computing. Scattering of light in strongly disordered materials is an extremely complex yet linear and deterministic process, equivalent to a large scale random matrix multiplication. Our group has been exploiting this approach for optical imaging in complex media for more than a decade, but we realized in the last few years that it was also an operation that could be exploited for several computing tasks: I will present some of our recent work ranging from image classification, to time series prediction, to solving large scale Ising Hamiltonians.