Brain-Inspired Learning for Intelligent Spectrum Sensing

Linda Katehi, Chaoyi He
Intelligent EM Sensor Lab, Texas A&M University, USA
Katehi@tamu.edu, Chaoyi_he@tamu.edu

***Abstract*—Today, the volume of data that needs to be communicated between wireless agents and the cloud has surpassed the ability of the available systems to transfer, manage and process it. Yet nature has told us that such an approach is ineffective and consumes all the resources we try to preserve with complex, expensive, and often inefficient solutions. In this paper, we will explore nature's design of intelligence and translate it to design guidelines for Intelligent Spectrum Sensors.**

***Keywords*—electromagnetics, sensors, intelligence, hybrid architectures, emergent memories, heterogeneous integration.**

I. Introduction

In the past three decades, advances in materials and electronic devices have extended Moore's law, increased device density and memory, and reduced computation and power consumption. They have driven the development of sensors able to collect data continuously and the design of software that can produce synthetic data at high bit data rates resulting in what we call today the Challenge of Data Deluge [1]. Following these advances, 6G intends to support autonomous driving and holographic communications, among other applications, implying the ability of systems to collect, store and process raw sensory data on the Cloud in real-time and with energy conservation. Contrary to these promises, experience with 4G and 5G technologies points in the opposite direction. As the number of sensors grows logarithmically, wireless data traffic and consumed energy are on an accelerated growth curve. The most significant growing component of electric energy consumption in the Information and Communication Technology (ICT) industry is communication to and from Data Centers. It is projected that by 2025 the amount of data stored will reach 187 Zettabytes (10^{21} Bytes), as shown in Figure 1 [2]. Due to the lack of information relative to the number and location of many worldwide Data Centers, it is currently difficult to estimate the electric energy required for data storage and the overall impact on CO_2 emissions. However, projections based on self-reported data [3] estimate that, by 2030, data storage will be the fastest-growing component of ICT, projected to reach 10-20% of global electric energy usage [4].

In parallel with the development of the personal computer and the growth in information processing and knowledge, our curiosity about the brain and its functions has led to an explosion in research in the past twenty years. Fundamental discoveries have shed light on the brain's physiology, the mechanisms of information gathering, unsupervised learning, and decision-making in an unmatched energy-efficient way. These discoveries drive a stark contrast between a laptop's brute force and energy-hungry computing power with the processes of information-based awareness, decision, and action of a brain that is inferior in mass and volume but superior in power use.

While there are still many unanswered questions relative to what drives human brain functions, including learning and decision-making, we have obtained a fundamental understanding of the human brain's architecture [5], which can guide engineering design principles as we try to answer two critical questions; What is intelligence and what drives it? How is it introduced into electronics? How do we reduce our dependence on data storage and contain energy usage?

In the following, we will discuss the concept of intelligence, the role of brain functions in intelligence, brain architecture for

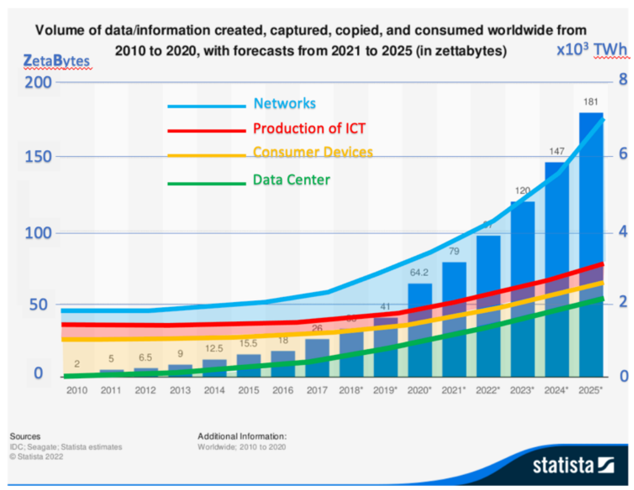

Fig. 1: Volume of data created/stored and expended energy per sector

intelligence, and its fundamental learning and decision-making processes. Then we will extend these concepts to engineering design principles and practices that lead to intelligence in edge devices. Finally, we will provide examples of applying these principles to designing intelligent electromagnetic sensors for spectrum sensing.

II. Edge Intelligence

We can argue that living species are governed by only one policy: survival under energy constraints. Their brain has been an essential body component to ensure they follow this policy successfully. However, survival has different implications depending on the size and lifetime of living organisms. A brain is a biological mass that requires energy to operate effectively. As a result, it does not exist if its added value does not effectively increase the probability of survival. This leads us to the realization that not all living species have a brain. Only multicellular organisms, with a size of mm or more and with a lifetime longer than a few days, have a brain, such as *C'elegans* [5], which does have a brain with only 302 neurons. For these living organisms, where there is a need for coordination between cells, their brain may be responsible for a large percentage of their size, indicating that their presence substantially increases their ability for survival.

As the size of an organism grows from *C'elegans* to *Humans*, the brain increases substantially in complexity and number of neurons (the human brain has 100 billion neurons between the cerebellum and the cerebral) to ensure coordination in function and action. Evolution through constant redesign has resulted in the human brain coordinating sensing, processing, and control executed in real-time and with the least possible energy. This is achieved by following policies and optimizing functions based on acceptable probability measures of success. Using policies in a brain results in a layered optimization of process and decision-making. In the most intelligent species, a slew of derivative policies leads to suboptimal solutions that are very effective most of the time. Depending on the hierarchy of a policy, the closer it is to the elemental policy of survival, the more challenging it becomes to alter; most derivative policies, however, are influenced by external conditions creating a continuously changing web of individual behaviors. Based on the above observations, we can conclude that edge intelligence represents the ability to collect information about the environment, process it as fast as possible, correlate it effectively, and use it to make decisions and take actions that increase the probability of achieving the intended outcomes, under severe energy constraints.

III. ARCHITECTURE FOR EDGE INTELLIGENCE

Edge intelligence has been spoken about more than demonstrated, with a recent exception in voice and text processing [6]. Most edge devices have already converted to platforms of sensors, i.e., cell phones with audio, video, camera, comms transceiver, and a GPS; automotive vehicles with everything included in the phone plus multiple radars; biosensors (bio-watch, bio-ring, glucose sensor) which record vital body signs and body chemistry and communicate the data directly to the cloud. While most of the collected information remains in storage unutilized, it is the most significant contributor to energy usage for storage and retrieval. In nature, however, not a single bit of data remains unprocessed.

Using a distributed architecture, a nervous system senses the environment, communicates, computes, and actuates movement with speed, accuracy, sparsity, noise, and energy trade-offs. Bodies are richly endowed with sensors having a wide range of dynamic time scales, from pain sensors with a time scale of seconds to acoustic sensors in echolocating bats that can time sound pressure changes at the microsecond level to detect source locations and distances. Many other sensors, such as olfactory sensors, vestibular accelerometers, dynamic vision sensors, and many pressure, temperature, and touch sensors, fill in a million-fold dynamic range. How does the brain combine and coordinate all these data streams? It does so with multiple layers of sensorimotor integration.

Sensory and motor control streams are not isolated from each other but interact at every layer of the hierarchy. Fast, low-resolution decisions are made at the bottom of the order (i.e., our eyes follow a flying bird). Each successive layer improves information quality from the summaries it receives about the state of lower controllers. The highest layers only know what needs to be accomplished (i.e., walking on a given trail). Nevertheless, brains achieve remarkably fast, accurate, and robust decision and control performance due to a highly effective layered processing, decision, and control architecture [5] that is designed to adhere to a few fundamental rules;

- Process data as close to the sensor as possible
- Send only what is essential,
- Send information at the lowest rate possible,
- Act when necessary.

Inspired by successful biosystems, we suggest a layered approach for edge-intelligence architecture, which can distribute the burden of sensing and decision-making from the centralized cloud to the end nodes. Figure 2 provides a

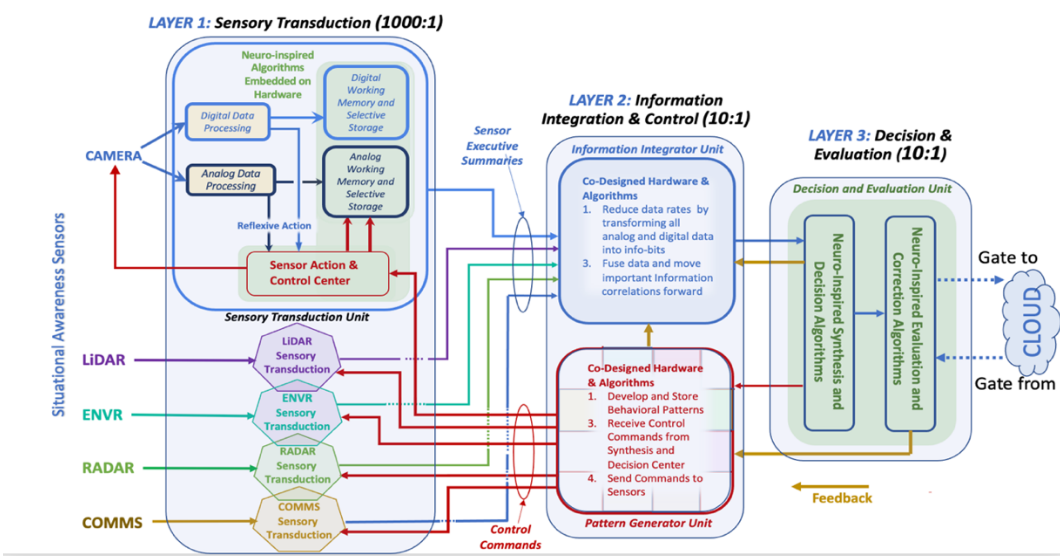

Fig. 2: Layered architecture for an intelligent sensor platform. Expected data-to-information reduction 100,000:1

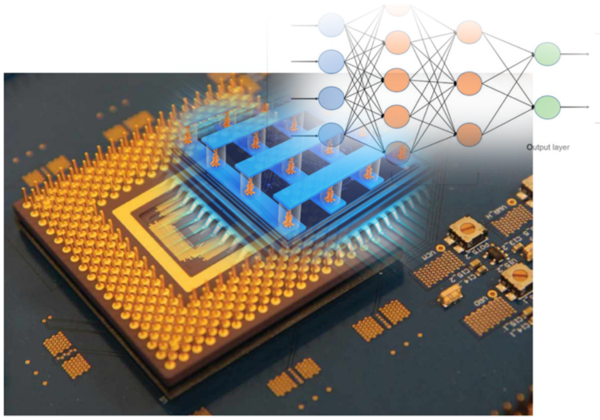

Fig. 3: Memristor-based crossbar configuration with an embedded MLP algorithm

schematic of this layered hierarchical architecture, specifically developed for a platform of sensors and includes three processing layers. At each layer, information is combined and distilled to allow decisions to be shaped by the task at hand. Different combinations of sensory information address each situational context, which is controlled by top-down projections. Information flow toward decision and action can be rapidly and flexibly controlled to accomplish action goals set at the highest layers. Consequently, only a subset of the sensory information must be transmitted to the next layer. In addition, state-of-the-art online sensor-action systems greatly ignore such possibilities in applications that store data for later (or never) usage. Siri or Alexa, for example, sends voice command data to the cloud to be processed, sometimes by deep neural networks (DNNs), and then sends back a voice response from the cloud to the car or mobile phone. This excessive data traffic to and back from the cloud can be avoided if the processing is done as close to the sensor as possible. In the proposed distributed architecture, there are three primary layers where significant data reduction is achieved.

A. Layer 1 - Sensory Transduction (1000:1 data reduction)

At this level, raw data sent by each sensor are processed to extract information that may either lead to a reflexive action or be reduced to an executive summary of events sent to the next hierarchical level (see Figure 2). The Sensory Transduction unit includes analog and digital data processing circuits co-located with the sensor and designed to translate the raw data outputs into information. To address energy-efficient processing of the multi-modal sensor data, an algorithm-hardware co-design approach is employed to re-use the model across different modalities of data, improve working memory efficiency and reduce energy overheads. Most present-day sensors generate analog outputs, instilling the need to rethink analog hardware design that can embed and sustain deep neural networks (DNNs) at the edge. Analog computation in memory (ACiM) is an attractive paradigm that can accelerate local NNs. However, ACiM hardware suffers from a significant overhead (both in latency and energy) posed by Analog to Digital Converters (ADCs). To overcome this barrier, training the algorithm with hardware-in-the-loop could achieve ADC-less ACiM hardware accelerators for embedded algorithms.

For analog-in-memory processing, there are two possible hardware-algorithm co-designed architectures. The first one uses a crossbar configuration of emerging memory devices that results in a massive reduction of computations involving vector-matrix multiplications (VMMs) commonly utilized in filtering, compression, data synchronization, and fusion signal processing methods. Analog crossbar arrays efficiently implement processing in one time-step instead of a digital implementation resulting in O(NlogN) time-steps or neuromorphic implementation O(N) where NxN is the matrix size in VMM. The second approach incorporates a memristive cellular neural network to process the analog signal pixel-parallelly (see Figure 3). The neural network allows for orders of a magnitude speed advantage over digital hardware by performing analog in-memory computing. In the case of image processing, as each cell processes one pixel of the image in parallel with all others, the one-to-one connectivity between cells leads to high throughput and low latency (<10 μs).

Compared with a purely CMOS implementation, a memristive MLP offers more than a 50% reduction in the transistor count, non-volatility, better power efficiency, and substantially faster processing than a convolutional neural network (CNN). A CNN's processing speed depends on various factors, including kernel size and input image size. On the contrary, CeNN's computation times are independent of network size if the image to be processed has dimensions within the bounds of the network [8],[9]. This data processing approach reduces high data rates to as low levels as possible while preserving the quality of information. Similar observations have been made using fully connected and recurrent neural networks embedded in crossbar configurations. Event sensors, whenever available, achieve part of the 1000:1 reduction goal. For these types of sensors, we employ a class of bio-inspired neural networks, namely, spiking neural networks (SNNs), to process visual information with discrete spikes or events over multiple time steps.

B. Layer 2 - Information Integration and Control (10:1 data reduction)

Layer 2 consists of two components – the *Information Integrator Unit* and the *Pattern Generator Unit*. At the Information Integrator Unit, heterogeneous information from Layer 1 is converted to the same event-information unit (i.e.,

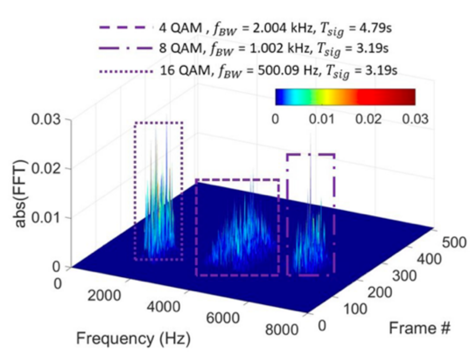

Fig. 4: Identification of three QAM signals using an MLP NN on a Crossbar Configuration. The signal has been Fourier transformed by a Memristive Cross-Bar before its identification, and results have been confirmed via YOLO.

spikes) and fused by aligning the individual sensor's temporal, spatial, and spectral frames. For most high-data-rate sensors, onboard processing can achieve some data reduction and compression. Still, the need for co-processing all sensor data for multi-modality detection and estimation limits how much that can be achieved just at the sensor front. State-of-the-art techniques in synthetic data fusion incorporate the use of social media and rely on deep learning computing in a combination of CPUs, GPUs, or FPGAs. In that context, deep learning methods can enhance the end-to-end capability for fusing and efficiently processing multi-modal sensor data [12]. The Information Integrator Unit transports only critical information to the next Layer in the hierarchy. This information is defined by the difference between the actual maximum observed entropy minus what information is already known (stored) about the observed system. The critical information sent forward to L3 includes information grouped in temporal and spatial frames determined by the integrator and stamped by its clock.

The Pattern Generation Unit in L2 stores a set of behavioral patterns designed for the platform or learned and optimized through the platform's operation. These patterns are replicated in every control center and every layer of the platform's architecture. Each behavioral pattern in this unit does not store individual actuator commands. It only stores information about which sensors or mobility control centers need to be activated as part of this pattern.

C. Layer-3 (L3): Decision and Evaluation (10:1 data reduction)

This layer consists of two units, the *Synthesis and Decision Center* and the *Evaluation and Error Correction Center*. Information communicated from L2 is correlated with information from previously stored time frames. Emerging events are identified and assessed as dangers, urgent needs, or opportunities. Balancing risk against need and opportunity leads to selecting a behavioral pattern. For the design of this layer, an SNN-crafted normalization or input encoding technique decouples the learnable parameters across different time steps and data streams to yield low-energy and high-performance training.

Recent advances in AI have vastly improved the analysis of high-dimensional signals sensed from dynamic environments. This has been accomplished primarily through learning from examples using deep learning, a simplified model of the cerebral cortex, and goal searching based on reinforcement learning, a simplified model of the basal ganglia. Unlike engineered systems, brain control is not centralized but is widely distributed and organized in a layered fashion. State of the Art (SOTA) NN algorithms use neural learning to achieve a level of perception that can lead to prediction and decision. Neural Networks that Temporally Change (NNTC) can be employed in the proposed Layer 3 designs. Instead of being parametrized by scalar weights, these networks have their learned weights as temporal functions. These NNTCs immediately search for temporal explanations leading to divergence in information (that is, congestion is likely on Friday and not likely on Sunday). They transform uncertainty and likelihood into temporal representations and accurate predictions. The evaluation part of Layer 3 controls two gates: One for transferring information to the cloud and the other for receiving information from it. Cloud information projects directly to Layer 3, which would take it down as necessary. In a real-time system where every fraction of a second may be significant, the Platform allows for direct projections from Layer 3 to 1, just as done in sensory cortices inside the brain.

IV. Conclusions

Designing for edge processing and intelligence faces expectations for real-time effective response and limitations in available space that lead to requirements for ultra-high-density hardware that operates on limited energy budgets and with high levels of security and reliability. Nature's frugality in tying capacity to the need for survivability and continuous learning provides lessons we can and should utilize in designing intelligent edge sensory platforms. A successful approach to such a design requires learning from nature how to trade off perception with action and survivability. In this case, the goal is to identify mechanisms from neuroscience [5], [6] and apply them to design an intelligent, autonomous mobile sensor platform that senses, perceives, acts, adapts, and learns in real-time and within a limited energy budget.

Acknowledgments

The State of Texas Excellence Fund and the National Science Foundation funded this research.

References

[1] G. Bell, T. Hey, and A. Szalay, "Beyond the Data Deluge," *Science (1979)*, vol. 323, no. 5919, pp. 1296–1297, Mar. 2009.
[2] Statista Inc. Research Department, "Amount-of-data-created-consumed-and-stored-2010-2020-with-forecasts-to-2025," *Statista Inc.*, 2021.
[3] D. D'Ambrosio and P. Gonzalez, "Electricity Sector Analysis - IEA," Paris, France, Sep. 2022.
[4] C. Garcia, "The Real Amount of Energy a Data Center Uses." Available: https://akcp.com
[5] Peter Sterling and Simon Laughlin, *Principles of Neural Design*, The MIT Press Cambridge, Massachusetts, London, England, 2017.
[6] N. A. Sulieman, L. R. Celsi, W. Li, A. Zomaya, and M. Villari, "Edge-Oriented Computing: A Survey on Research and Use Cases," Energies (Basel), vol. 15, no. 2, Jan. 2022.
[7] B. A. Anderson et al., "The past, present, and future of election history," Neurosci Biobehav Rev, vol. 130, pp. 326–350, Nov. 2021.
[8] Chua & Yang, "Cellular Neural Networks: Theory and Applications," IEEE Trans. Circuits & Systems, Vol. 35, 1988.
[9] C. Li et al. "3D Memristor Circuits as Complex Neural Networks," Nature Electronics, 2020.
[10] Chaoyi He et al. "A Memristor Crossbar Chip for Analog Signal Processing," Proceedings of the 2021 COMCAS October 2021.
[11] A Grimaldi et al. "A Homeostatic gain control mechanism to improve event-driven object recognition," International CBMI Conference 2021.
[12] Pedram Ghamisi et al. "Multisource and Multitemporal Data Fusion in Remote Sensing," IEEE Geoscience and Remote Sensing Magazine March 2019.