Neuromorphic Cognition

Giacomo Indiveri* and Dr. Rodney Douglas
Institute of Neuroinformatics, University of Zurich and ETH Zurich, Zurich, Switzerland

Synonyms

Neuromorphic cognitive systems; Neuromorphic electronic systems; Neuromorphic real-time behaving systems

Definition

The hallmark of cognitive behavior is the ability of an agent to select an economically advantageous action based on immediate external stimuli as well as on their longer-term context. Neuromorphic cognition refers to the cognitive abilities of systems or agents implemented in neuromorphic electronic VLSI technology whose processing architecture is similar to the distributed, asynchronous one of biological brains. Neuromorphic agents are typically real-time behaving systems composed of multiple asynchronous event-based VLSI chips that integrate networks of silicon neurons and dynamic synapses together and that are interfaced to event-based neuromorphic sensors and robotic actuators. In order to express cognitive performance, these agents require a hardware infrastructure that supports local learning and decision making, for distributed communication and for the elaboration of state-dependent processing. We describe examples of such mechanisms and present a method for efficiently implementing neuromorphic cognition in these agents.

Detailed Description

Digital computers provide prodigious computational power and memory for the simulation of models of cognition. Nevertheless humans and many animals including insects still outperform the most powerful computers in real-world cognitive tasks. This disparity between the effectiveness of biological nervous systems and computers is primarily attributable to differences in their elementary processing devices and to the kinds of computational primitives they implement (Indiveri and Horiuchi 2011; Mead 1990). Rather than using Boolean logic, precise digital representations, and clocked operations, the nervous systems perform robust and reliable computation using hybrid analog/digital unreliable components; they employ distributed, event-driven, collective, and massively parallel mechanisms, and they implement adaptation, self-organizing, and learning strategies, on multiple spatial and temporal scales. Understanding these principles, and how they can lead to behaviors that exhibit cognitive qualities, remains a major challenge for science.

The goal of neuromorphic cognition is to understand the neural mechanisms used by the brain to achieve cognition, as well as to implement them using electronic neural systems whose physics of computation and architectural constraints are based on those of the biological nervous systems.

*Email: giacomo@ini.phys.ethz.ch
During the last decade the neuromorphic engineering community has made substantial progress in this endeavor by designing neuromorphic sensors and neuromorphic hardware and spiking neural networks and by developing the platform technology necessary for constructing distributed multi-chip systems of sensors and neuronal processors that operate asynchronously and communicate using action potential like signals (or spikes) (Chicca et al. 2007b; Merolla et al. 2007). In particular, it is now possible to build large multi-chip neuromorphic systems that combine multiple sensors with spiking neural network chips, robotic actuators, and end-effectors, as well as with conventional digital computing systems (e.g., for data logging, of further analysis). Usually, each chip specializes in a specific neural function (such as visual sensing, auditory sensing, filtering, spike-based learning).

The ability to construct distributed systems from collections of heterogeneous neuromorphic subsystems is largely due to the development of a very simple communication scheme, the “address-event representation” (AER), that is analogous to action potential communication between biological neurons. In AER, signal sources (such as silicon retina pixels or silicon neurons) each have a unique address. They signal their current activation by transmitting their address asynchronously on a digital bus. These AER events can then be routed to various target processors, in an analogous manner to axons routing somatic spikes to synapses.

Until recently, systems of this type have been mainly reactive, responding directly to stimuli with less regard for their embedding context in either the observed world or within the economic goals of the agent. This limitation has been due to their inability to autonomously extract, and reason over, more sophisticated models of world and self, abilities that are characteristic of cognitive systems. The step from reaction to cognition in neuromorphic systems is not an easy one, because the principles of cognition remain to be unraveled. Discovering these principles, and learning how to implement them in neuromorphic technology, is now an active domain of research. For example, considerable effort is devoted to developing electronic circuits that are considered to provide “necessary conditions” for cognition, such as the abilities to:

1. Interact with the environment efficiently in real time, with appropriate temporal dynamics
2. Store memories and to learn about the nature and statistics of their input signals
3. Make weighted decisions
4. Generate appropriate sequences of behavior

These are only necessary conditions. To build neuromorphic cognitive systems, the challenge is not only to develop the appropriate hardware circuits and systems but more generally to understand what kinds of computational models asynchronous spiking neurons can support and how to configure the hardware neurons to perform desired tasks using a particular computational approach. Recent projects have obtained promising results in this direction (Eliasmith et al. 2012; Giulioni et al. 2011; Indiveri et al. 2009; Nefci et al. 2013). For example, an important recent result offers a foundation for the design and construction of compact and effective cognitive neuromorphic systems (Rutishauser and Douglas 2009). These authors describe how state-dependent computation can be designed using coupled recurrent networks. In the next sections we will first present examples of basic neuromorphic building blocks that can be used for implementing the necessary conditions. Thereafter we describe a method that uses these building blocks and the principles of (Rutishauser and Douglas 2009) to “synthesize” a simple cognitive agent.
Temporal Dynamics in Neuromorphic Systems

Neuromorphic architectures and processing strategy are substantially different from those of conventional von Neumann computers. The von Neumann architecture calls for a small number of complex processors, which are each able to access a very large common memory storing code, data, and partial results. Memory access and the processing of that data are highly orchestrated, occurring synchronously at very high clock rates. The processors are time multiplexed, computing on partial results that must be transferred from and to the external memory banks.

By contrast, neuromorphic spiking neural network architectures are composed of massively parallel arrays of simple processing elements in which memory and computation are co-localized. This design does not allow for the virtualization of time and the rapid transfer of partial results back and forth between the computing units and memory banks situated outside the architecture core. Instead, the synapse and neuron circuits of neuromorphic architectures must process input spikes on demand, as they arrive, and must produce their output responses in real time. Consequently, the processing elements must operate with time constants that are well matched to those of the natural world signals they are designed to process. This constraint is not easy to satisfy using conventional analog VLSI circuits, whose transistors operate in the strong-inversion regime (Sze 1981), because the demand for long time constants can lead to area-expensive solutions. The area cost can be avoided by neglecting the requirement for matching and modeling the neural dynamics instead at accelerated, unrealistic time scales (Brüderle et al. 2011; Schemmel et al. 2007; Wijekoon and Dudek 2008). An alternative solution to this problem is to use current-mode design techniques (Tomazou et al. 1990) and log-domain circuits operating in the weak-inversion regime (Liu et al. 2002). The time constants of these circuits are proportional to the circuit capacitance and inversely proportional to a reference current. If the reference current is a subthreshold one, then this current can be as small as fractions of pico-amperes. Time constants of milliseconds would therefore require capacitors of only femtofarads, which can be implemented with very compact designs in standard VLSI technologies. Following this approach it is possible to build compact low-power neuromorphic circuits that can faithfully reproduce the dynamics of synaptic transmission observed in biological synapses (Destexhe et al. 1998) and that can implement sensory and signal processing systems for agents that need to interact with the environment in real time (Liu and Delbruck 2010).

An example of a compact subthreshold log-domain circuit that has often been used to reproduce biologically plausible temporal dynamics is the differential pair integrator (DPI) (Bartolozzi and Indiveri 2007). Using log-domain circuit analysis (Bartolozzi and Indiveri 2007; Drakakis et al. 1999; Frey 1993), and with reasonable assumptions of non-negligible input currents (Bartolozzi et al. 2006), it can be shown that the DPI circuit is equivalent to a first-order linear integrator which faithfully reproduces the dynamics of synaptic transmission observed in biological synapses (Destexhe et al. 1998).

Spike-Based Learning and Plasticity Mechanisms

One of the key properties of cognitive neural systems is their various forms of plasticity. In particular, “long-term” plasticity (Abbott and Nelson 2000) can produce sustained changes in the strength of individual synapses and so learn about the state of the environment. Implementations of long-term plasticity mechanisms in neuromorphic VLSI chips provide a method of self-calibration, by which synaptic weights can be set automatically, without the
overhead of dedicated wires and pins that would be required to load individual synapses from an external source.

Fortunately, plasticity mechanisms based on the timing of the spikes map very effectively onto silicon neuromorphic devices (Fusi et al. 2000; Häfliger et al. 1997; Indiveri et al. 2006; Mitra and Indiveri 2009; Bofill-i Petit and Murray 2004). A popular class of spike-driven learning rules is based on the spike-timing-dependent plasticity (STDP) (Abbott and Nelson 2000; Markram et al. 1997). In STDP the relative timing of pre- and postsynaptic spikes determines how the efficacy of a synapse is updated. There are several implementations of STDP learning chips (Arthur and Boahen 2006; Indiveri et al. 2006; Mitra and Indiveri 2009; Bofill-i Petit and Murray 2004), and a wide range of theoretical models have been proposed. Both the theoretical models and the VLSI implementations indicate that STDP can be effective in learning to classify spatiotemporal spike patterns (Arthur and Boahen 2006; Gütig and Sompolinsky 2006), although, in its simplest form, the STDP algorithm is unsuitable for learning different patterns of mean firing rates (Abbott and Nelson 2000).

An interesting issue that arises for physical implementations of plasticity mechanisms, be they biological or electronic, is that the synaptic weights they update are both bounded (they can neither increase indefinitely nor assume negative values) and have limited resolution. This restriction on the precision and range of the weights imposes strong constraints on the network’s capacity to store and retain synaptic memories. For example, in attractor networks with bounded synapses (Giulioni et al. 2011), if the long-term changes in synaptic weight cannot be arbitrarily small, the network’s memory trace decays exponentially with the number of stored patterns.

In addition to the problem of overwriting the stored synaptic values with new ones, neuromorphic VLSI implementations must also address the problem of possible decay of the stored synaptic weight voltages with time (e.g., due to leakage). One effective strategy for protecting stored memories against decay is to permit only two stable synaptic efficacy states per synapse and allow only a very low average number of transitions from one stable state to the other (Fusi 2002). By modifying only a random subset of the synapses with a small probability, memory lifetimes increase by a factor inversely proportional to the probability of synaptic modification (Fusi 2002).

Given a stochastic mechanism for the weight update, and provided neurons are stimulated via a large number of redundant synaptic circuits, hardware implementations of spike-based learning networks can be shown to have computational properties analogous to the ones of ideal networks, with unbounded synapses (Fusi 2002). Where neuromorphic sensors interface with the environment, the input signals are inherently noisy. The noise of these input spike patterns is sufficient to drive the stochastic update mechanism required to increase memory lifetimes. In addition, by permitting only two stable synaptic states to store on long time scales the values of the weights, it is sufficient to use a bistable circuit that restores the synaptic state to either its high rail or its low one, depending if the weight is above or below a set threshold. In this way memory is preserved even in the absence of stimuli or when the presynaptic activity is very low.

Brader and colleagues (Brader et al. 2007) have proposed a spike-based learning rule using bistable synapses that is able to classify patterns of mean firing rates. This model is also able to account for classical STDP behavior, as well as the additional rich phenomenology observed in neurophysiological experiments on synaptic plasticity. This rule has been implemented using neuromorphic analog/digital circuits (Chicca et al. 2013) and has been applied to real-time spike-based learning and classification (Giulioni et al. 2009, 2011; Mitra et al. 2008).
Soft Winner-Take-All Circuits

Winner-take-all (WTA) networks are well suited for implementing decision making and context-dependent action selection operators and are easily implemented using neuromorphic circuits. These types of networks typically consist of a group of interacting neurons that compete with one another in response to an input stimulus. The neurons with the strongest responses suppress competing neurons to win the competition. Competition is achieved through a recurrent pattern of connectivity involving both excitatory and inhibitory connections. Cooperation between neurons with similar response properties (e.g., close receptive field or stimulus preference) is mediated by excitatory connections. Competition and cooperation make the output of individual neuron depend on the activity of all neurons in the network and not just on its own input. According to the relative strengths of recurrent excitation versus inhibition, the network may select only one winner (hard winner-take-all) or a small population of likely winners (soft winner-take-all) (Amari and Arbib 1977; Dayan and Abbott 2001; Hansel and Sompolinsky 1998; Wilimzig et al. 2006). WTAs are useful computational primitives because they perform linear operation such as amplification and locus invariance, as well as nonlinear operations such as selective amplification (or selective suppression) and multi-stability (Chicca et al. 2007a; Douglas and Martin 2007; Hahnloser et al. 2000).

The computational abilities of sWTA networks have been used for solving feature extraction and pattern classification problems in simulation (Ben-Yishai et al. 1995; Bennett 1990; Pfeiffer et al. 2010) and in neuromorphic VLSI (Chicca et al. 2007b; Hahnloser et al. 2000). Interestingly, the pattern of connection of the WTAs is consistent with the pattern of connections observed in the superficial layers of the neocortex (Binzegger et al. 2004; Douglas and Martin 2004), suggesting that the neurons of these layers may make use of this computational primitive (Douglas et al. 1995; Somers et al. 1995).

The examples of WTAs listed above make use of continuous valued outputs that can be interpreted as discharge rates. However, in neuromorphic implementations, where communication is usually by spikelike address events, it is more convenient (and more biologically correct) to implement these primitives using spiking neurons. Several examples of VLSI sWTA networks of spiking neurons can be found in the literature (Abrahamsen et al. 2004; Chicca et al. 2004; DeYong et al. 1992; Hylander et al. 1993; Indiveri et al. 2001; Oster and Liu 2004). Figure 1 shows experimental data measured from one of such neuromorphic WTA networks (Chicca 2006; Chicca et al. 2004). This example shows how the VLSI spiking neural network can perform nonlinear selection. An input stimulus (see Fig. 1a) consisting of Poisson trains of spikes, with a mean frequency profile showing two Gaussian-shaped bumps with different amplitude, is applied to the input synapses of each neuron in the soft WTA network. The chip output response is a series of spike trains produced by the output silicon neurons (see Fig. 1b). The mean frequencies measured from each spike raster in Fig. 1b show how the soft WTA network selects and amplifies the Gaussian bump with higher activity while suppressing the other one, with respect to the baseline condition.

Synthesizing Cognition in Neuromorphic Architectures

The previous sections described examples of circuits and circuit design techniques that can be used to implement the necessary components for endowing neuromorphic systems with cognitive abilities. Neftci and colleagues have recently described a method that combines many of these elements into a neuromorphic agent able to express simple real-time cognitive behaviors (Neftci et al. 2013).

Their method recognizes three conceptual layers of processing: firstly, a high-level behavioral model, in which the target cognitive task is cast as a state machine that captures the state-dependent
organization of the required behavior (Fig. 2b); secondly, an intermediate abstract computational layer, comprising a field of neurons configured as many modular sWTA networks; and finally, an implementation layer composed of distributed neuromorphic hardware such as CMOS VLSI neurons and synapses and their event communication infrastructure.

The method begins by creating the intermediate general-purpose computational layer. The model neurons of this theoretical and therefore precise layer are supported by the relatively imprecise electronic neurons of the implementation layer. These two layers are related by mapping the necessary model neuron parameters onto the circuit bias voltages of the electronic neurons via an automatic procedure in which the circuit biases are calibrated against the model parameters by a series of population activity measurements.

Finally, the high-level behavioral model is embedded in the computational layer by the introduction of sparse connections between sWTA modules so that the overall network gives rise to attractors that encode the behavioral states of the cognitive model. Additional connections are introduced to provide the necessary transitions between the attractor states (Rutishauser and Douglas 2009). The transitions are conditional on external input. These coupled sWTA networks form a “soft-state machines” which, unlike conventional finite-state machines, combine analog and digital processing in the same circuit.

An example of a task similar to those used in laboratory settings to probe cognition in primates is shown in Fig. 2a. The cognitive neuromorphic agent is challenged with this same task. The agent is implemented using both electronic spiking neural networks and a neuromorphic silicon retina sensor that responds to visual patterns displayed on a screen (Fig. 2a) with trains of spikes. The visual scene is composed of independently moving horizontal and vertical bars. Depending on the current context, the neuromorphic agent is required to respond differently to these bars. The context in force is determined by a cue that is transiently shown at the beginning of the experiment and that must be remembered. In the first context the agent must produce response A when it observes a horizontal bar entering the right half of the screen; and in the second context it must produce response B when it observes a vertical bar entering the left half of the screen. Under all other conditions, the agent should remain silent. The relevance of this task is twofold: It requires processing of real-time dynamic visual stimuli in a state-dependent fashion, and it is comparable...
in complexity to the tasks used to probe cognition in behavioral studies of humans and animals. The neuromorphic agent was able to perform this task successfully and reliably and produce the correct output by activating the appropriate place-encoded neurons.

The important contribution of this work is that it provides a systematic way to install reliable processing on the underlying unreliable neuromorphic hardware, in a way that simplifies the programming of the desired high-level behavior. This approach is analogous to the software one used to program and compile computational processes in general-purpose digital computers, except that the underlying neuromorphic hardware is radically different from digital ones in both system concept and electronic implementation. Because our networks are implemented using spiking neurons and synapses with biologically plausible dynamics, they cast light on how both brains and future neuromorphic technologies could implement cognitive behaviors.

**Discussion**

So far neuromorphic engineering has succeeded in providing the physical infrastructure for constructing visual and auditory sensors, spiking neural networks, and event-driven effectors that are similar in organization if not in size to the nervous systems of biology. Until now the tasks that these neuromorphic systems have been able to perform are quite simple, so that even the most sophisticated VLSI systems created were reactive in quality, mapping rather directly sensory...
percepts to simple actions. Of course, intelligent systems are not simply reactive. Instead, given some knowledge of its environment and its internal state, and some approximate behavioral objectives, an intelligent agent comes to reason that certain combinations of actions are more likely to achieve an objective than others. While we may recognize these properties in the behavior of animals, it has been extraordinarily difficult to evoke them in artificial systems, be they either symbolic or connectionist in design.

Neuromorphic cognition faces the challenge of implementing such properties in real-time behaving sensory-motor systems. This not only requires to identify the organization principles of the brain but also to develop the technology and method for instantiating cognition on physical implementations of neurons and synapses. Although the neuromorphic neuron and synapse circuits are fabricated using standard CMOS VLSI foundries, they are unlike conventional industrial digital or analog circuits. They are a hybrid of analog and digital asynchronous circuits, and their analog transistors usually run in the subthreshold regime in order to emulate very compactly the crucial functional characteristics of neural systems (Indiveri et al. 2011; Liu et al. 2002). Consequently, neuromorphic circuits are subject to substantial fabrication variability (device mismatch) and operating noise. Unlike digital circuits, analog circuits require dedicated signal restoration mechanisms to avoid signal corruption through successive stages of processing (Sarpeshkar 1998). A similar signal restoration problem holds for biological neurons, as was quickly recognized by von Neumann when he considered alternative architectures for early computers (von Neumann 1958). The sWTA networks described in the previous sections and the neuromorphic circuits that emulate neural dynamics can be used to form signal restoration mechanisms that support reliable computation, even when implemented with imprecise hardware components. Furthermore, given the progress made in spike-based learning and plasticity methods for neuromorphic architectures (Mitra et al. 2009; Pfeiffer et al. 2010; Serrano-Gotarredona et al. 2013) and given the recent developments in methods for synthesizing cognition in neuromorphic systems (Eliasmith et al. 2012; Neftci et al. 2013), the challenge of building bioinspired intelligent agents that can interact with the environment in real time and express cognitive abilities is much more accessible.

Acknowledgments

This work was supported by the EU ERC Grant “neuroP” (257219). Part of this work, including the one on soft WTA networks, was done in collaboration with Elisabetta Chicca. We acknowledge also the CapoCaccia and Telluride workshops on neuromorphic engineering (The CapoCaccia Cognitive Neuromorphic Engineering Workshop (http://capocaccia.ethz.ch/) and the Telluride neuromorphic cognition engineering workshop (http://www.ine-web.org)) for fruitful discussions on neuromorphic cognition.

References


