Biologically-plausible hierarchical chunk learning on mixed-signal neuromorphic hardware

Atilla Schreiber[1], Shuchen Wu[2], Chenxi Wu[1], Eric Schulz[2], Giacomo Indiveri[1]
1. Institute of Neuroinformatics, UNI–ETH Zurich, Switzerland 2. Max Planck Institute for Biological Cybernetics Tübingen, Germany

Summary. Humans seamlessly and effortlessly group patterns in perceptual sequences into chunks, which are memorized as separate entities. Chunking is a central computational principle essential for memory compression, structural decomposition, and predictive processing. On an algorithmic level, cognitive models such as the Hierarchical Chunking Model (HCM) propose grouping proximal observational units as chunks, which has been demonstrated to resemble human chunk learning behaviorally. However, on the circuitry level, the question remains: Which mechanisms of a neural population enable chunk learning, and how chunking empowers neural circuit computation. We propose a biologically plausible implementation of the HCM: the neuromorphic HCM (nHCM) in recurrent spiking neural networks (SNN). When parsing through perceptual sequences, nHCM uses sparsely connected neurons to construct hierarchical chunk representations in an event-driven way. We first applied nHCM's working principles on a CPU, which already excels in speed, power consumption, and memory usage compared to its original counterpart. Then, we developed the neural circuitry for recurrent SNN and applied this algorithm to the mixed-signal dynamic asynchronous neuromorphic processor (DYNAP) − a chip that emulates neural networks with silicon neurons and synapses. The hardware uses analog electronic circuits to mimic biological dynamics and emits digital spikes asynchronously, which is many times more energy-efficient than modern computers on biological timescales. By mapping onto DYNAP, we showed that the algorithm is inherently robust and overcomes the high failure rate and inaccuracy of bio- and electronic-analog components. The nHCM demonstrates the ability of neuro-adapted algorithms to meet the increasing demand for energy-efficient computing and proposes an alternative programming paradigm. This project represents a building block toward understanding how a recurrent SNN learns perceptual and action execution patterns as chunks and provides a basis for event-driven time-domain bio-signal processing.

Human chunking presents us with a learning model that leads to more structured and hierarchical results than traditional deep learning. The HCM [1] is a cognitive chunking model where repeated proximal units of perception are grouped to construct a hierarchical representation. The nHCM adapts the algorithm's implementation to a neural population. As illustrated in Fig. 1A, nHCM parses a temporally encoded perceptual sequence input such as strings of letters in units of chunks and learns new chunks online. We have implemented parsing on the neuromorphic chip [2] and with a neuron-emulating data structure on a CPU.

After every input, the algorithm checks if the previous most recent representation and current input represent a known chunk. If so, their respective neuron populations are inhibited and the population representing their combination is activated. This process is repeated with the new chunk and the second most recent and uninhibited chunk until no further concatenations are possible. When a chunk is activated, a signal travels through a designated chain of neurons, where each neuron encodes a specific time after the initial activation to keep track of the temporal information.

If a neuron population has gone for a sufficient time without inhibition it reaches the last neuron in its population chain and is used as output as it is assumed that it has reached the maximum concatenation possible in the current sequence and is sampled probabilistically in combination with the previous maximally concatenated chunk which was stored in a short-term memory and replaced by the newer chunk afterward. If the same pair of chunks gets sampled k-times in a row, it will be combined into a new chunk and transition to long-term memory.

When a transition between two chunks is learned, a new population of neurons is designated for that combination. The population is constitutively inhibited and needs disinhibition through the first child chunk to be primed for activation. The activatory signal stems from an excitatory synapse of the second child chunk. The disinhibition and activation only sync up to produce a signal when the temporal information matches the initial encoding (Fig. 1B and Fig. 1E). Disinhibition surpasses other coincidence detection methods in accuracy and robustness. When a parent chunk activates, an inhibitory signal propagates back to the populations that activated it, and suppresses their signal propagation before they produce an output. Individual components are thus inhibited in favor of the complete representation.

We used the Kullback–Leibler (KL) divergence to measure the deviation of distribution across models. The neuronal data structure and probabilistic online learning increase the processing speed by almost two magnitudes while reducing the memory needs to a minimum on traditional hardware (Fig. 1C). The

nHCM on neuromorphic hardware achieves comparable results in the experiments (Fig. 1D). Similar to the brain's computation with unreliable and stochastic components, the nHCM learns robust representations and self-corrects the propagation of errors despite the variability and malfunction of neural populations on the chip. For now the time-interval of the sequence is fixed and the input restrained to characters but could be expanded to work with any sequential inputs on biological timescales such as speech, or even fMRI data. In conclusion, steps towards neuro-adapted algorithms on CPU and neuromorphic hardware, as demonstrated with the nHCM could help with the increasingly power-demanding computational needs and profit from nature's ability to achieve viable results in a resource-scarce and competitive environment.

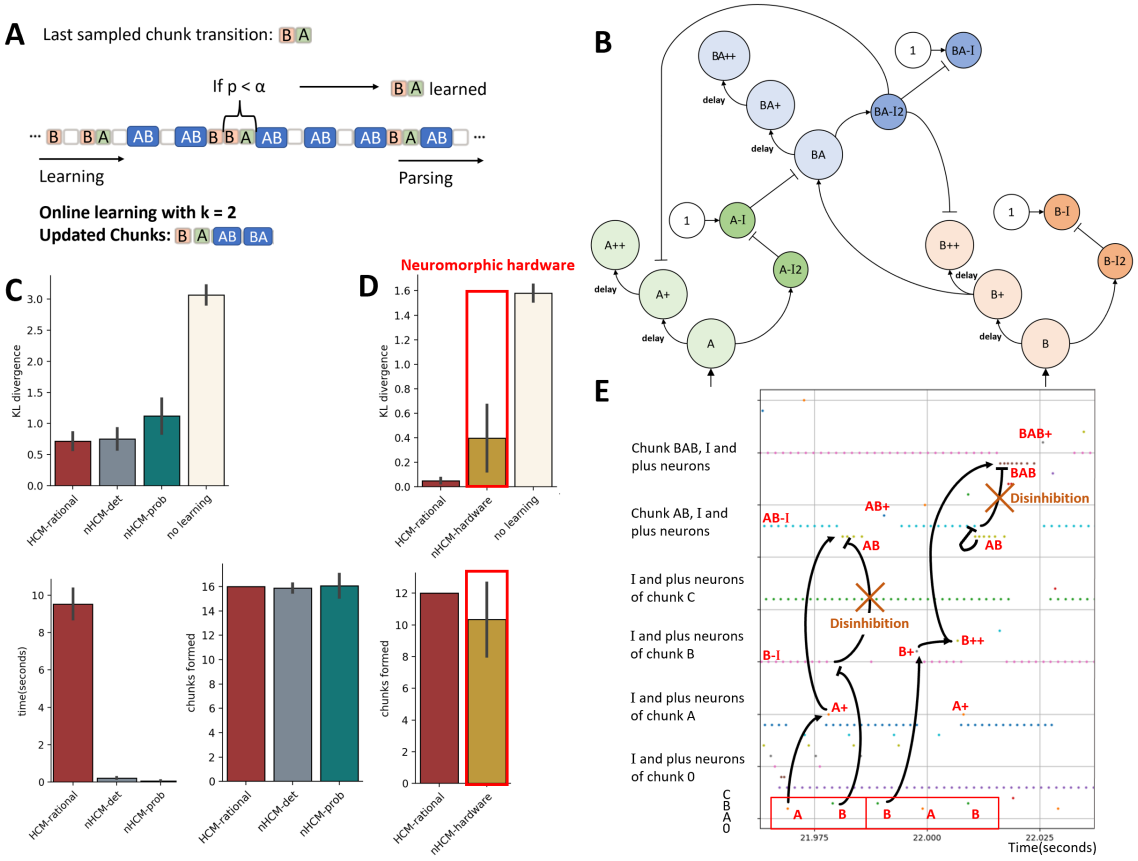

Figure 1: **nHCM implementation and results** (A) Online learning while parsing the sequence with $k = 2$. (B) Neuronal disinhibition circuit of chunks A, B, and BA. 1 indicates constitutive activation of inhibitory interneurons. Subsequent activation of population B and then A will lead to disinhibition and activation of AB neurons. Feedback from AB neurons suppresses downstream neurons of chunk population A and B. (C) Experiments with rational HCM, deterministic version of nHCM with postprocessing and probabilistic version with online learning on CPU (Chunks 16, Sequence length 5000, n = 30). (Top left) KL divergence when compared to no learning (Bottom left) Speed comparison. (Bottom middle) Number of chunks formed. (D) Experiments with rational HCM on CPU and nHCM on neuromorphic hardware (Chunks 12, Sequence length 500, n = 3). (Top) KL divergence compared to no learning. (Bottom) Number of chunks formed. (E) Spike train output of the nHCM on the neuromorphic hardware showing chunking of inputs AB and BAB. Primary activation of atomic elements at the bottom. Chunks are ordered from bottom to top as primary activation, constitutive inhibition, and downstream delayed neurons with temporal information. I2 interneurons not shown.

References

[1] Shuchen Wu et al. "Learning Structure from the Ground up—Hierarchical Representation Learning by Chunking" Thirty-Sixth Conference on Neural Information Processing Systems, 2022

[2] Saber Moradi et al. "A scalable multi-core architecture with heterogeneous memory structures for Dynamic Neuromorphic Asynchronous Processors (DYNAPs)", CoRR, abs/1708.04198, 2017,