Designing for Neural Information Processing at Scale
Some lessons from Self-Organization in Spiking Neural Networks

Felix Wang[1,2], Fred Rothganger[2], Stephen E. Levinson[1] [1]University of Illinois at Urbana-Champaign, [2]Sandia National Laboratories

ECE ILLINOIS
Department of Electrical and Computer Engineering

Sandia National Laboratories

Neuro-inspired Design

- Neural information processing rests on emergent structures, where the complexity of problems is managed only through the complexity of solutions
- Regulation within a critical operating regime optimizes for network stability, information transmission, information storage, and computational power
- A balance between signal propagation and attenuation is required to scale local interaction to global communication

Spiking Network Architecture

Distributed computation and communication is necessitated by the substantial size of a spiking neural network:

- When plated on multielectrode arrays and grown to confluency, there are on the order of 10^3 cortical neurons/mm^2 with up to 10^6 neurons total, and 10^3 outgoing connections per neuron
- The human brain is on the order of 10^{11} neurons

Neural computation yields a sequence of spiking events:

- Spikes are qualitatively distinctive impulses in contrast to the otherwise quiescent nature of neurons
- For a given neuron, neurocomputational behavior is determined from the relationship between applied stimuli and the resulting spiking activity
- Spikes may be treated as the primary information bearing signal of the network

Communication of spikes performs implicit computation, specified through synaptic connectivity:

- Excitatory mechanisms facilitate the generation of spikes at the post-synaptic neuron, whereas
- Inhibitory mechanisms depress the generation of spikes
- Synaptic weights determines efficacy of transmission
- Plasticity rules change synaptic weights with respect to spiking activity, and
- Reward mechanisms modulate plasticity rules with respect to external events

Analogs to Parallel Computing

Both systems may be formulated as graphs $G(V, E)$:

- Vertices: spiking neurons $\leftrightarrow$ compute nodes
- Edges: synaptic connections $\leftrightarrow$ communication links

A key difference is found in their component interaction:

- *Local data movement* - synaptic connectivity patterns are spatially dependent, with connection lengths limited to within a small neighborhood, vs.
- *Global data movement* - parallel interconnects route messages potentially across the entire system

Emergent Structure

Self-organization of spiking networks yields persistently reproducible spatio-temporal patterns of spiking activity:

- *Polychronous* groups of neurons are characterized by both a spatial component (*which* constituent neurons spiked) and a temporal component (*when*)
- In contrast to single neurons, these groups are sensitive to spatio-temporal patterns of stimuli
- Group activation may be treated as higher-order information bearing signals of the network

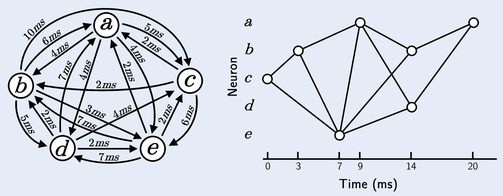

Figure 1: Toy network (left) illustrating the graphical representation of a polychronous group (right)

Formation of groups follows from selection on variety:

- *Primary repertoire* - composed of potential groups that are supported structurally by the network
- *Secondary repertoire* - acquired through activity dependent experiential selection from primary repertoire

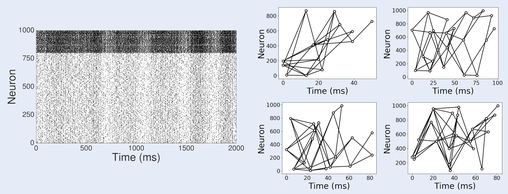

Figure 2: Spiking activity (left) yields multiple polychronous groups (right) through application of synaptic plasticity rules

Distributed Information

The spiking activity of polychronous groups of neurons is distributed both spatially and temporally:

- Constituent neurons may have no direct connection
- Group activation duration may extend beyond the transmission delay between any given pair of neurons

The self-organization of spiking networks may be considered as a distributed method of learning:

- Strengthening or weakening of synaptic weights occurs only through local plasticity rules, yet
- Acquisition of polychronous groups requires mutual facilitation among remote neurons
- The transition from asynchronous to polychronous spiking provides a primitive associative mechanism

Regulation of Information Processing

Operation near criticality offers a number of advantages with respect to neural information processing

Network stability - balance between excitatory and inhibitory mechanisms drives network toward stable dynamics where spiking activity is sustained

Information transmission - balance between signal propagation and attenuation enables potentially long range communication between neurons in the network

Information storage - balance between strong and weak connections allow for increased number of independently stable patterns of spiking activity

Computational power - balance between fixed and variable synaptic weights maximizes the diversity of reliable mappings between inputs and outputs of the network

Scaling Communication

The ratio of signal propagation to attenuation is characterized by the branching parameter: $\sigma = \frac{1}{n}\Sigma_{i=1}^{n}\Sigma_{j=1}^{n} p_{ij}$

where p_{ij} is the probability that activity in component i will induce activity in components j

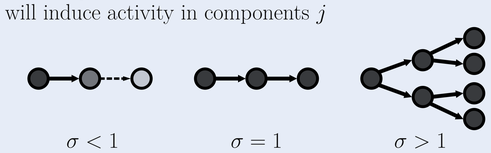

Figure 3: Depending on the value of the branching parameter, network activity over time either dies out (left), is approximately sustained (center), or blows up (right)

Scaling local interactions to global communication requires balancing the branching parameter near unity:

- Increased excitatory mechanisms should be met with appropriately increased inhibitory mechanisms
- Increased variability (e.g. of computation) should be counterbalanced by increased regulatory mechanisms

Algorithmic Trade-offs

Neural information processing at scale, in general, trades architectural complexity for algorithmic complexity:

- Forgoing more straightforward global synchronization routines in favor of local coordination methods
- Decreased average communication latency, but interaction beyond local neighborhood needs to pass through intermediate computation
- Increased data locality per computational unit, but application relevant data structures are distributed

What is Criticality?

Criticality refers to the behavior of a system operating near a critical point or phase transition (e.g. the transition between order and disorder)

Case study: the Ising model examines a simple system of magnetic spins subject to nearest neighbor interactions and thermal fluctuations on a 2D lattice

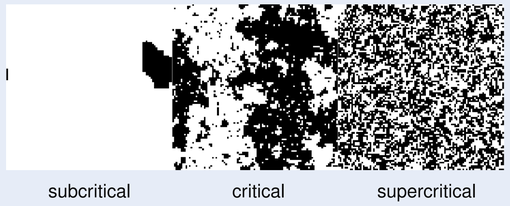

Figure 4: ferromagnetic (left), phase transition (center), and paramagnetic (right) behaviors of the Ising model

A proxy measure of communication is the dynamic correlation of magnetic spin between lattice points:

$$C_{ij} = \langle (s_i - \langle s_i \rangle)(s_j - \langle s_j \rangle) \rangle$$

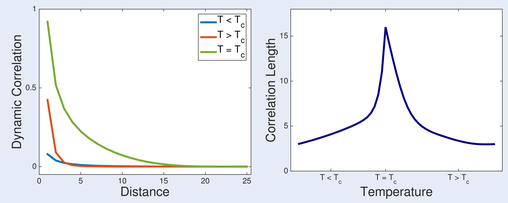

Figure 5: Average dynamic correlation with respect to separation distance (left); maximum average dynamic correlation length with respect to temperature (right)

References

[1] W. R. Ashby, "Requisite variety and its implications for the control of complex systems," *Cybernetica*, vol. 1, no. 2, pp. 83–99, 1958.

[2] J. M. Beggs, "The criticality hypothesis: how local cortical networks might optimize information processing," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 366, no. 1864, pp. 329–343, 2008.

[3] E. M. Izhikevich, "Polychronization: Computation with spikes," *Neural Computation*, vol. 18, pp. 245–282, 2006.

Acknowledgements

The author would like to thank the Language Acquisition and Robotics Laboratory at the University of Illinois at Urbana-Champaign and the Data Driven and Neural Computing group at Sandia National Laboratories for their invaluable guidance. This work was supported by Sandia National Laboratories Laboratory Directed Research and Development (LDRD) program under the Hardware Acceleration of Adaptive Neural Networks (HAANA) Grand Challenge project and the LDRD Academic Alliance program. Sandia National Laboratories is a multi-mission laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

SAND2017-4201 C