Computation with Sequences of Assemblies in a Model of the Brain

Max Dabagia MAXDABAGIA@GATECH.EDU
School of Computer Science, Georgia Tech

Christos H. Papadimitriou CHRISTOS@COLUMBIA.EDU
Department of Computer Science, Columbia University

Santosh S. Vempala VEMPALA@GATECH.EDU
School of Computer Science, Georgia Tech

Editors: Claire Vernade and Daniel Hsu

Abstract

Even as machine learning exceeds human-level performance on many applications, the generality, robustness, and rapidity of the brain's learning capabilities remain unmatched. How cognition arises from neural activity is *the* central open question in neuroscience, inextricable from the study of intelligence itself. A simple formal model of neural activity was proposed in Papadimitriou et al. (2020) and has been subsequently shown, through both mathematical proofs and simulations, to be capable of implementing certain simple cognitive operations via the creation and manipulation of assemblies of neurons. However, many intelligent behaviors rely on the ability to recognize, store, and manipulate temporal *sequences* of stimuli (planning, language, navigation, to list a few). Here we show that, in the same model, time can be captured naturally as precedence through synaptic weights and plasticity, and, as a result, a range of computations on *sequences* of assemblies can be carried out. In particular, repeated presentation of a sequence of stimuli leads to the memorization of the sequence through corresponding neural assemblies: upon future presentation of any stimulus in the sequence, the corresponding assembly and its subsequent ones will be activated, one after the other, until the end of the sequence. If the stimulus sequence is presented to two brain areas simultaneously, a scaffolded representation is created, resulting in more efficient memorization and recall, in agreement with cognitive experiments. Finally, we show that any finite state machine can be learned in a similar way, through the presentation of appropriate patterns of sequences. Through an extension of this mechanism, the model can be shown to be capable of universal computation. We support our analysis with a number of experiments to probe the limits of learning in this model in key ways. Taken together, these results provide a concrete hypothesis for the basis of the brain's remarkable abilities to compute and learn, with sequences playing a vital role.

Keywords: assemblies, neural network, neuroscience, plasticity, sequence learning, finite state machine

Overview

How does the activity of individual neurons and synapses lead to higher-level cognitive functionality? This is a central mystery in neuroscience which currently lacks an overarching theory. In Papadimitriou et al. (2020) a mathematical model of the brain was proposed in an attempt at such a theory. This neural model — which we call NEMO in this paper — entails brain areas, spiking neurons, random synapses, local inhibition, and plasticity. The dynamical system defined by NEMO

Extended abstract. Full version appears as (Dabagia et al., 2023).

has certain emergent attractors corresponding to *assemblies of neurons;* recall that an assembly is a stable set of highly intraconnected neurons in an area representing through their near simultaneous excitation a real-world object, episode, or idea (Hebb, 1949; Harris et al., 2003; Buzsáki, 2019). There is a growing consensus in neuroscience that assemblies of neurons play an important role in the way the brain works (Buzsáki, 2010; Huyck and Passmore, 2013; Yuste, 2015; Eichenbaum, 2018). It was established in Papadimitriou et al. (2020) and subsequent research, through both mathematics and simulation, that certain elementary behaviors of assemblies arise in NEMO: projection, association, merge, among others. Moreover, through NEMO one can implement certain reasonably complex cognitive phenomena, including learning to classify well-separated classes of stimuli (Dabagia et al., 2022), and parsing natural language sentences (Mitropolsky et al., 2021).

Many of the brain's remarkable capabilities rely on working with *sequences* (of stimuli, words, places, etc), with the capacity of the human brain for language being a particularly striking example. Arguably, it is through sequences of stimuli and their representations that brains deal with the all-important concept of *time*. The question arises: can NEMO capture this capability of the animal brain? In past work, NEMO did not have to deal explicitly with sequences or time. In the English parser implemented in Mitropolsky et al. (2021), the input sentence is presented sequentially, and the order of its words is not memorized by the device. In subsequent work on parsing (Mitropolsky et al., 2022), the need to memorize subsequences of the input language became apparent in connection to the *center embedding* of sentences; however, no mechanism for this memorization was proposed.

In this paper, we demonstrate the emergent formation of sequences of assemblies in NEMO. When a brain area is stimulated by the same sequence of stimuli a handful of times, assemblies are reliably created, and the entire sequence will subsequently be recalled when only the beginning of the stimulus sequence is presented. Importantly, the underlying mechanism involves the capture of time precedence between sequences through the establishment, via plasticity, of high synaptic weights between stimuli representations, in the direction of time. Moreover, we demonstrate that, by involving in this process additional brain areas during presentation (essentially forming a "scaffold" of interconnected assemblies) makes memorization faster and more robust; and more so if this new area already contains another memorized sequence. This provides theoretical support to experience: it is easier to memorize a sequence of stimuli when they are set to a tune, or when each stimulus is mentally associated by the subject with an element of an already memorized sequence — for example, the sequence of buildings next to the subject's home.

We use these ideas to further show that, in NEMO, assemblies can be configured to simulate an arbitrary *finite state machine* (FSM). Moreover, we show that this configuration can be learned quickly by presenting sequences of stimuli corresponding to state transitions; this captures the brain's ability to learn *algorithms* involving sequences. The implementation and learning of FSMs relies crucially one last feature of NEMO, namely *long-range interneurons (LRIs).* These are neural populations extrinsic to the brain areas of NEMO, which can be recruited by assemblies in adjacent areas, and whose function is to inhibit or disinhibit remote brain areas to achieve synchrony and control of the computation (Sik et al., 1995; Jinno et al., 2007; Zhang et al., 2014).

One interesting byproduct of the mechanism for implementing and learning FSMs is a simple demonstration that NEMO is *Turing complete*. In other words, NEMO with LRIs constitutes a hardware language capable of implementing any computation, within the constraints imposed by the parameters of the model. This is rather significant for a mathematical model that has the ambition to capture a large part of human cognition. The original exposition of NEMO in Papadimitriou et al.

(2020) did contain an argument of Turing-completeness as well; however, that proof relies on a biologically implausible computer-like program, with loops, conditional statements, and variables corresponding to assemblies. The Turing completeness argument in the present paper is carried out strictly within NEMO, and the required program is implemented with LRIs, yielding an entirely hardware-based and stimulus-driven general computer consistent with neurobiological principles.

The Neural Model

NEMO consists of the following ingredients: There are a finite number of *brain areas*, which are sets of n excitatory neurons, connected internally by directed edges (synapses) present independently with probability p. Some pairs of brain areas may be connected in one direction (or possibly both) by another, bipartite, random graph. All edges have nonnegative weights, initially 1. Input enters the model through designated *sensory* areas, which have only outgoing connections to other brain areas.

The dynamics of the system proceed in discrete time steps. At each step, each neuron determines its total synaptic input determined by summing up the current weights of incoming connections from neighbors which fired on the previous time step. For each brain area, *only the k neurons with the highest total input fire at each step* (with ties broken randomly). This operation, k-winners-take-all or the k-cap, is an important part of the model, capturing local inhibition and the area's inhibitory/excitatory balance. Synapse weights, both within and between areas, are nonnegative and governed by Hebbian plasticity with parameter $\beta > 0$, so that each time j fires immediately after i fires, the weight of the synapse from i to j increases by a multiplicative factor of $1 + \beta$.

Any brain area A can be inhibited — that is, no excitatory neuron in A can fire — by the activation of a designated population of inhibitory neurons, I_A, which are in turn inhibited by another population of interneurons, D_A. For our purposes, we simplify I_A and D_A to $0/1$-valued signals, such that area A is inhibited if and only if its inhibitory signal I_A is 1, and I_A is 1 precisely when D_A is 0. In full generality, the signal D_A may be a linear threshold function of the firing in other areas. These inhibition and disinhibition actions are assumed to be determined by activity on the round immediately prior.

In mathematical notation, with $\mathcal{A}$ as the set of areas, the dynamics of the model are fully captured by the following update equations:

$$
\begin{aligned}
I_A(t+1) &\leftarrow 1 - D_A(t+1) \\
x_A(t+1) &\leftarrow (1 - I_A(t+1)) \cdot k\text{-cap}\left(\sum_{B \in \mathcal{A}} W_{B,A} x_B(t)\right) \quad \forall A \in \mathcal{A} \\
W_{B,A}(t+1) &\leftarrow W_{B,A}(t) + \beta \cdot x_A(t+1) x_B(t)^\top \odot W_{B,A}(t) \quad \forall A, B \in \mathcal{A}
\end{aligned}
$$

where $x_A(t) \in \{0,1\}^n$ is the activity in area A (and so has either 0 or k nonzero elements) and $W_{B,A}(t) \in \mathbb{R}_+^{n \times n}$ is the weight matrix for the synapses from area B to area A. The function $k\text{-cap}\colon \mathbb{R}^n \to \{0,1\}^n$ maps a vector to the indicator vector of its k largest elements, while $\odot$ denotes element-wise multiplication.

Results

Below, we present a brief summary of various capabilities of NEMO related to computation on sequences. See Dabagia et al. (2023) for formal theorem statements and proofs of these results.

Sequence projection In sequence memorization (or projection), a sequence of stimuli or assemblies is copied into another area. The most natural way to project a sequence from one area to another is to simply activate, in the first area, the assemblies for each element of the sequence, one after the other in the given order. Assuming that there is a fiber connecting the first area to the second, and the second area is disinhibited, one would hope that this would result in the creation of a set of corresponding assemblies in the target area, so that activating any newly created assembly results in the sequential activation of the rest of the sequence of new assemblies. We show in theory and experiment that this occurs in NEMO, with quantitative bounds on plasticity and the other parameters.

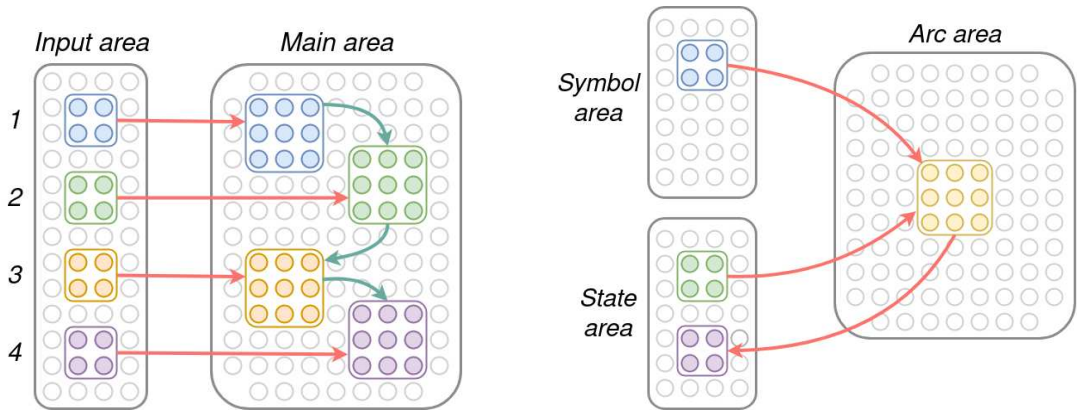

Figure 1: Left: Each stimulus in an input area evokes an associated assembly in a brain area. By repeated sequential presentation, directional connections between these assemblies are strengthened. . Right: A pair of symbol and state assemblies project to an arc assembly, which in turn projects to the next state assembly for the associated transition.

Scaffolded sequence projection A well-known phenomenon in cognitive science is that memorization is easier by creating associations such as mnemonics. For sequences, learning one sequence by associating with another sequence, element by element (e.g., learning the alphabet to a tune), helps with retention and recall. We consider a very simple form of this: when projecting a sequence, we create two copies of the sequence that "scaffold" each other. Remarkably, this leads to provably better recall of the sequence, with about half as many training rounds as simple sequence projection.

Finite state machine memorization We demonstrate that NEMO is powerful enough to simulate an arbitrary finite state machine (FSM). In fact, an FSM is learned — that is, memorized — simply by presenting all valid transitions between states in the FSM. Each each pair of state and symbol assemblies is linked with a unique arc assembly in a separate area. By repeated sequential presentation, this arc assembly becomes linked to the correct next state for that transition, so that in the future the firing of a state/symbol pair will trigger the associated arc assembly, and in turn the next state assembly, to fire.

Acknowledgments

MD and SV are supported in part by NSF awards CCF-1909756, CCF-2007443 and CCF-2134105, and a NSF Graduate Research Fellowship. CP is supported by NSF Awards CCF-1763970 and CCF-1910700, and a research contract with Softbank.

References

György Buzsáki. Neural syntax: cell assemblies, synapsembles, and readers. *Neuron*, 68(3):362–385, 2010.

György Buzsáki. *The Brain from Inside Out*. Oxford University Press, 2019.

Max Dabagia, Santosh S Vempala, and Christos Papadimitriou. Assemblies of neurons learn to classify well-separated distributions. In *Conference on Learning Theory*, pages 3685–3717. PMLR, 2022.

Max Dabagia, Christos H Papadimitriou, and Santosh S Vempala. Computation with sequences in a model of the brain. *arXiv preprint arXiv:2306.03812*, 2023.

Howard Eichenbaum. Barlow versus hebb: When is it time to abandon the notion of feature detectors and adopt the cell assembly as the unit of cognition? *Neuroscience letters*, 680:88–93, 2018.

Kenneth D Harris, Jozsef Csicsvari, Hajime Hirase, George Dragoi, and György Buzsáki. Organization of cell assemblies in the hippocampus. *Nature*, 424(6948):552–556, 2003.

Donald Olding Hebb. *The organization of behavior. A neuropsychological theory*. John Wiley, 1949.

Christian R Huyck and Peter J Passmore. A review of cell assemblies. *Biological cybernetics*, 107: 263–288, 2013.

Shozo Jinno, Thomas Klausberger, Laszlo F Marton, Yannis Dalezios, J David B Roberts, Pablo Fuentealba, Eric A Bushong, Darrell Henze, György Buzsáki, and Peter Somogyi. Neuronal diversity in gabaergic long-range projections from the hippocampus. *Journal of Neuroscience*, 27(33):8790–8804, 2007.

Daniel Mitropolsky, Michael J Collins, and Christos H Papadimitriou. A biologically plausible parser. *Transactions of the Association for Computational Linguistics*, 9:1374–1388, 2021.

Daniel Mitropolsky, Adiba Ejaz, Mirah Shi, Mihalis Yannakakis, and Christos H Papadimitriou. Center-embedding and constituency in the brain and a new characterization of context-free languages. *arXiv preprint arXiv:2206.13217*, 2022.

Christos H Papadimitriou, Santosh S Vempala, Daniel Mitropolsky, Michael Collins, and Wolfgang Maass. Brain computation by assemblies of neurons. *Proceedings of the National Academy of Sciences*, 117(25):14464–14472, 2020.

A Sik, M Penttonen, A Ylinen, and Gy Buzsáki. Hippocampal ca1 interneurons: an in vivo intracellular labeling study. *Journal of Neuroscience*, 15(10):6651–6665, 1995.

Rafael Yuste. From the neuron doctrine to neural networks. *Nature reviews neuroscience*, 16(8): 487–497, 2015.

Siyu Zhang, Min Xu, Tsukasa Kamigaki, Johnny Phong Hoang Do, Wei-Cheng Chang, Sean Jenvay, Kazunari Miyamichi, Liqun Luo, and Yang Dan. Long-range and local circuits for top-down modulation of visual cortex processing. *Science*, 345(6197):660–665, 2014.