A Procedural Roadblock to Mechanistic Understanding of Neural Circuits

Venkat Ramaswamy (vramaswamy@ncbs.res.in)
Simons Centre for the Study of Living Machines,
National Centre for Biological Sciences,
Bangalore 560065, India.

Abstract

Neuroscience is witnessing impressive progress in techniques for observing and interrogating neural circuits. Advances include optical readout of neural circuit activity, capability to optogenetically stimulate/silence subsets of neurons in-vivo and ascertaining exact anatomical connectivity, for increasingly larger neural circuits. It is thought that progress in such technologies holds promise in ultimately enabling us to understand mechanistic computation in neural circuits leading to behavior. Here, using techniques from Theoretical Computer Science, we examine how many experiments are needed to establish an empirical understanding of mechanistic circuit computation, for a fixed behavior. It is proved, mathematically, that establishing the most extensive notions of understanding *needs* exponentially-many experiments in the number of neurons, in general, unless a widely-posited hypothesis about computation is false. To make matters worse, the feasible experimental regime is one where the number of experiments scales sub-linearly in the number of neurons. Together, this suggests that such a comprehensive understanding is de facto unknowable, in general. Determining which notions of understanding are algorithmically tractable, thus, becomes an important direction for investigation.

A similar roadblock may exist in our quest to comprehensively understand contemporary deep neural networks as well.

Keywords: Understanding; Circuit interrogation; Behavior.

"But on principle, it is quite wrong to try founding a theory on observable magnitudes alone. In reality, the very opposite happens. It is the theory which decides what we can observe."
– Albert Einstein

Neuroscience is making remarkable ongoing progress in experimental techniques, at the present time. At the neuronal and network level, advances include the ability to both image activity in as well as to activate/silence subsets of neurons, all-optically, in-vivo, in awake, behaving animals. Already, such techniques are being applied to study (nearly) whole-brain neural activity in zebrafish larva, C. elegans and hydra, albeit, currently, at low temporal resolution. Furthermore, connectomics is enabling us to determine the precise structure of these neural circuits. Presently, the connectome of the nematode C. elegans and the larval tadpole C. intestinalis have been fully reconstructed.

Indeed, ongoing national initiatives in the United States, Japan, China and Korea have made accelerating the development of these (and other) neurotechnologies, one of their central goals. The general premise has been that such technologies will ultimately enable us to reach the goal of understanding how networks of neurons mechanistically perform computations that lead to specific behaviors.

Empirically understanding mechanistic computation in these neural circuits will require efficient algorithms for large-scale (including whole-brain) neural circuit interrogation. The algorithms will seek to prescribe the smallest number of experiments necessary (that scale as a function of the number of neurons in the network) towards this goal. An experiment, for example, might entail perturbing activity of a subset of neurons, while imaging their activity and attempting to elicit behavior. The specifics of the next experiment prescribed by the algorithm could depend on the outcome of the current experiment. Theoretical Computer Science has known, from over half a century of work, that some problems have fast algorithms whereas certain others provably require intractably many steps of computation to solve, in general. It is as yet unclear in which class of problems those pertaining to neural circuit understanding fall in, although it has been suggested (Koch, 2012; Kumar, Vlachos, Aertsen, & Boucsein, 2013) that combinatorial explosion in number of interactions (among other considerations) might present a challenge to this end. Combinatorial explosion in the solution-space does not necessarily imply algorithmic intractability, and therefore the question has remained largely open.

Here, using techniques from Computational Complexity Theory, we ask what is the smallest number of experiments *necessary*, in general, in order to arrive at a comprehensive empirical understanding of neural circuit computations that lead to a fixed behavior, in a hypothetical experimental setting. We find that no general algorithms exist to solve this class of problems that always use sub-exponential number of experiments in the number of neurons, unless the complexity class $\mathcal{P} = \mathcal{NP}$ [1]. If, remarkably, $\mathcal{P} = \mathcal{NP}$ were true, it would mean that hundreds of problems – many of them commercially important and extensively studied for decades – would have sub-exponential algorithms, where none have been found to date. Performing exponentially-many experiments in the number of neurons would lead one to require more experiments than the estimated number of atoms in the observable universe, even

[1] The hypothesis mentioned in the abstract (that is widely thought to be true) is that $\mathcal{P} \neq \mathcal{NP}$.

for modest-sized nervous systems – rendering it an impracticable undertaking. The exposition here will focus on the main ideas and intuition behind the methods; we skip a number of essential details for want of space.

Experimental scenario

The experimental scenario we have in mind is one where we have an individual animal that has been trained to do a certain behavioral task and we wish to obtain a comprehensive mechanistic understanding of how its neural circuits (acutely) cause this behavior to be manifested at the present point in time. This understanding must be a causal account and also encompass the degeneracy that has been associated with neural computations (Edelman & Gally, 2001). For simplicity, we consider behaviors with a binary behavioral readout[2]. For instance, a possible task could be discriminating two odors A and B by a trained animal, wherein our goal is to obtain an understanding of how mechanistic computation in neural circuits currently causes the animal to recognize odor A and perform the correct behavior to indicate the same. For the sake of analysis, we will assume that we have access to its entire nervous system, where we have the ability to, for example, image activity in and stimulate/silence subsets of neurons and we know its connectome, although the main result is largely independent of the specific experimental capabilities/technology at hand.

Understanding mechanistic computation

What does it mean to understand mechanistic computation in neural circuits? There is as yet no standard definition of understanding in this context and conceivably, there exist multiple concomitant descriptions constituting notions of understanding that might, for example, include details spanning different spatial/temporal scales. We wish to have our theory be applicable to a wide variety of such notions. A characteristic of understanding is the ability to answer questions about the subject of understanding, in short order. A notion of understanding might be considered more *extensive* than another if a description corresponding to the former can so answer a larger repertoire of questions, than is the case with the latter. Accordingly, the most extensive notions of understanding – which we will call *comprehensive notions of understanding* – ought to enable us to answer certain central questions about mechanistic computation in the circuit leading up to behavior. We will posit one such central question that involves determining a certain subset of neurons that causally participate in computations leading to the said behavioral readout. This general, if seemingly unusual, approach is a key step in proving the result, as the reader will soon see. Some definitions follow, to make this precise.

Definition 1 (Degenerate Circuit)**.** *A subset N of neurons is said to constitute a* d*egenerate circuit for a behavior B, if B can be successfully elicited (with respect to the behavioral readout) by the silencing of all neurons, except those in N.*

Now, it is possible that whether a subset of neurons forms a degenerate circuit for a certain behavior or not might be a function of the current state of the network, which we could mean to include a variety of phenomena including the dynamical state of the network, stochastic variability, plasticity, neuromodulation or even neuroimmunological and other considerations that might contribute to trial-to-trial variability in the neural circuits. We will assume the availability of an *oracle* that will guarantee that the neural circuit is in the exact same state before the beginning of each trial, since we wish to understand mechanistic computation in the circuit at the present point in time, embodying a single state. This is a fairly routine construct in Theoretical Computer Science, where one often shows that even with such strong capabilities, certain things are hard to do. Going forward, we will assume that the state is so fixed in each instance. Note that this does not compel the algorithm to use the oracle. We now define another notion.

Definition 2 (Minimal Degenerate Circuit (MDC))**.** *A subset N of neurons is said to constitute a* m*inimal degenerate circuit for a behavior B, if N constitutes a degenerate circuit for B and furthermore no proper subset of N is a degenerate circuit for B.*

There exists at least one MDC with respect to each behavior, although there are likely many. It is interesting to determine the neurons that are present in every MDC; we call this set of neurons the *vital set* of neurons for that behavior (with respect to said behavioral readout).

Definition 3 (Vital Set)**.** *A subset N of neurons is said to constitute the* v*ital set of neurons for a behavior B, if N is the intersection of every MDC for B.*

This *vital set* is especially of interest; it follows that it is exactly the set of neurons, with the property that silencing any of them will extinguish the said behavior[3]. For example, silencing the T4/T5 cells in Drosophila using temperature-sensitive shibire has been shown (Bahl, Ammer, Schilling, & Borst, 2013) to completely abolish the optomotor response. We therefore posit that any notion of understanding that claims to be comprehensive ought to allow us to determine this set fairly quickly.

Definition 4 (Vital Set problem)**.** *Given a whole-brain and a behavior B, determine the vital set for B.*

Thus, more precisely, we define a notion of understanding to be *comprehensive*, if from its description one can quickly (i.e. in number of steps that scale as a polynomial in the number of neurons) determine its Vital Set for the said behavior. The definition utilizes the notion of a *reduction*, which is fundamental in Theoretical Computer Science. Informally, a reduction is a recipe to quickly convert any instance of one problem (Problem A) into an instance of another problem (Problem B),

[2] Strictly speaking, we only require a way of partitioning possible behaviors into two classes.

[3] The vital set could be an empty set. Even if this is the case, it would be useful to know that to be the case.

such that a solution to Problem B can be quickly mapped back to a solution of Problem A as well. Therefore, the existence of an efficient algorithm for Problem B immediately implies the existence of an efficient algorithm for Problem A, via the reduction. More profoundly, if there exists no efficient algorithm for Problem A, the reduction implies that Problem B cannot have an efficient algorithm either; otherwise there would be a logical contradiction.

The definition of comprehensive understanding above, thus, asserts that there exists a reduction from the Vital Set problem to the stated problem of determining (any notion of) comprehensive understanding. We will end up showing that the Vital Set problem is hard to solve using number of experiments that scale as a polynomial in the number of neurons, which implies that the comprehensive understanding problem is likewise. Towards this end, we define another problem.

Definition 5 (k-Vital Set problem)**.** *Given a whole-brain, a behavior B and a positive integer k, is the Vital Set for B of size k?*

There is a straightforward reduction from the k-Vital Set problem to the Vital Set problem. The reduction involves solving the corresponding instance of the Vital Set problem, counting the number of neurons in the obtained Vital Set, and answering if it is of size k.

Next, we establish a reduction from k-CLIQUE – a known $\mathcal{NP}$-complete problem – to the k-Vital Set problem, which shows that the k-Vital Set problem is in fact an $\mathcal{NP}$-hard problem. This implies that no general sub-exponential algorithms exist for any of the problems participating in the aforementioned reductions, unless $\mathcal{P} = \mathcal{NP}$.

The *clique* of an undirected graph is a subgraph of it, such that every pair of vertices in that subgraph has an edge between them. The k-CLIQUE problem seeks to determine if a given graph has a clique of size k. Our reduction, in effect, will provide a quick recipe to construct a neural circuit from a given undirected graph, with the guarantee that the graph will have a clique of size k, if and only if the said neural circuit has a Vital Set of size $k+2$. The simplest version of the construction has a single "sensory" and a single "motor" neuron. The sensory neuron signals the arrival of the pertinent stimulus by firing a single spike and the motor neuron, likewise, signals execution of the said behavior by firing a single spike (Figure 1).

The "interneuron" circuit connecting the sensory and motor neuron is constructed from the undirected graph by having a neuron in place of each graph vertex and bi-directional[4] connections whenever there is an undirected edge between vertices. Additionally, there is a coincidence-generator made up of two neurons that also connects the interneuron circuit to the motor neuron. The synaptic responses are set up in order to result in three properties: (a) The entire circuit forms a degenerate circuit for the behavior. (b) The undirected graph has a k-clique, if and only if the circuit formed by the corresponding interneurons plus the coincidence-generator neurons, sensory and motor neuron form a minimal degenerate circuit of size $(k+4)$ for the behavior. (c) The circuit, by design, does not have a degenerate circuit of size less than $(k+2)$. We skip the rest of the details of the construction, for want of space. What the reduction from k-CLIQUE to k-Vital Set establishes is that if there were a general algorithm that always solved k-Vital Set with a sub-exponential number of experiments, that could be used to construct an algorithm using sub-exponential steps for k-CLIQUE as well and in turn for every $\mathcal{NP}$-complete problem, implying that $\mathcal{P} = \mathcal{NP}$. We have proofs establishing that a few other related problems are $\mathcal{NP}$-hard as well (details skipped). This implies that notions of understanding that these problems reduce to are algorithmically hard to establish as well; the k-Vital Set Problem isn't the only such algorithmically hard question. Another interesting consequence of this reduction is that, even if one had an exact simulation of the entire neural circuit (or the whole-brain), extracting a comprehensive understanding needs exponentially-many steps, in general, unless $\mathcal{P} = \mathcal{NP}$.

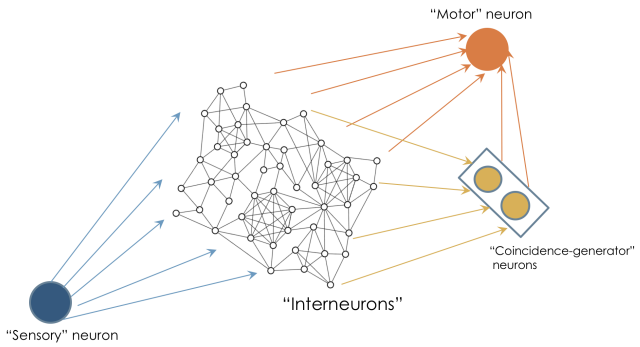

Figure 1: A schematic of the neural circuit that the reduction constructs, given any undirected graph. The interneuron circuit is what differs in the neural circuit for each given undirected graph.

An important potential caveat is that the analysis above – as is typical in Computational Complexity Theory – is of the worst-case. That is, the theory implies that there is a sub-class of neural circuits that provably require exponentially-many experiments for the said problems, unless $\mathcal{P} = \mathcal{NP}$. Worryingly though, this sub-class of circuits is a rather simple-looking one that maps stimulus to response, while potentially being mediated by a smaller degenerate circuit.

The above result raises the question of what notions of understanding – even if they might not be comprehensive – are experimentally establishable, in practice. This is a foundational question that will likely require extensive study. But first, we need to consider the question of how the number

[4] Bi-directional connections are not central to this construction. One could have a more complicated construction, where the number of bidirectional connections is arbitrarily small (details skipped).

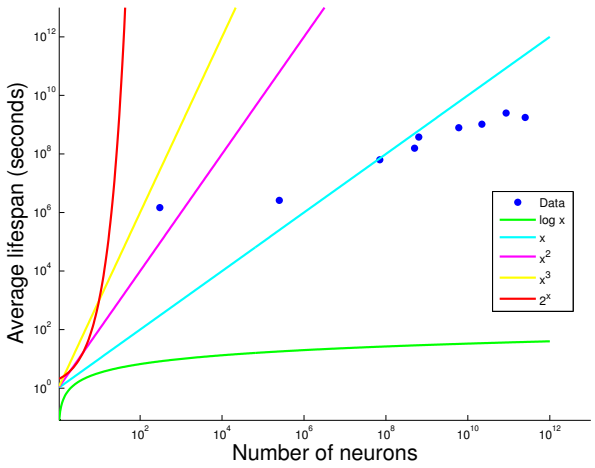

Figure 2: A log-log plot of estimates of the average lifespan as a function of (estimated) number of neurons in the nervous system for C. elegans, fruit fly, mouse, octopus, marmoset, Human, Chimpanzee, Rhesus monkey and Elephant. The plot suggests that running even a linear number of experiments in the number of neurons for most organisms would need time that exceeds their expected lifetime.

of behavioral-timescale experiments ought to scale with the number of neurons, for it to be realizable, in practice. We plotted (Figure 2) the average lifetime of a number of organisms as a function of the estimated number of neurons in the central nervous system. Superimposed on this plot is the amount of time it would take to finish an experimental protocol that scales as certain functions in the number of neurons, assuming each experiment takes a second (corresponding to the behavioral timescale) and there is no gap period between successive experiments – both conservative assumptions. From this plot, it is immediate that, for most organisms, an experimental protocol that uses even a linear number of experiments requires time roughly equal to the average lifetime of the animal – rendering such a protocol infeasible, in practice. Thus, not only do the most extensive notions of empirical understanding seem to need exponentially-many experimental steps, the regime for realizable algorithms is one that uses a sub-linear number of experiments – a significantly more stringent prospect.

Discussion

The results here suggest that in addition to current technological barriers, there exists another fundamental roadblock – a *procedural* one – to neural circuit understanding. Specifically, the most extensive notions of understanding might elude us, in general. This has intriguing philosophical implications in our quest to understand the brain.

To be sure, we haven't already encountered this roadblock. The current neurotechnological renaissance will likely lead to a significantly enhanced understanding of the brain. That said, it is arguably only a matter of time before the ambition of our questions approaches this barrier. In the meanwhile, demarcating the boundary between algorithmically tractable and intractable notions of neural circuit understanding will be an important new foundational direction for investigation in Neuroscience.

The question of mechanistically understanding contemporary deep neural networks shares many parallels with the question of so understanding biological neural circuits. We do not yet understand either to any reasonably comprehensive degree – in spite of the remarkable fact that we can manipulate deep networks at will. Like in Systems Neuroscience, there has been a genre of work in deep neural networks (e.g. (Erhan, Bengio, Courville, & Vincent, 2009; Li, Chen, Hovy, & Jurafsky, 2015)) focused on attempting to understand these networks via a study of selective neurons/units. On the other hand, more recent work (Morcos, Barrett, Rabinowitz, & Botvinick, 2018), has suggested that in networks that generalize well, such selective neurons are no more important for network performance than non-selective neurons, arguably bringing us back to (conceptual) square one. While our results here do not directly map to contemporary deep networks, we close by speculating that an analogous roadblock may exist with deep networks as well. This direction, therefore, merits further investigation.

Acknowledgments

The work was supported by the Simons Foundation. The author wishes to thank Kambadur Ananthamurthy, Arunava Banerjee, Upi Bhalla, Albert Cardona, Vishaka Datta, Vivek Jayaraman, Konrad Kording, Arvind Kumar, Sahil Moza, Dinesh Natesan, Christos Papadimitriou, Sanjay Sane and Shuchita Soman for useful discussions.

References

Bahl, A., Ammer, G., Schilling, T., & Borst, A. (2013). Object tracking in motion-blind flies. *Nature neuroscience*, *16*(6), 730.

Edelman, G. M., & Gally, J. A. (2001). Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences*, *98*(24), 13763–13768.

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). *Visualizing higher-layer features of a deep network* (Tech. Rep.). Universite de Montreal.

Koch, C. (2012). Modular biological complexity. *Science*, *337*(6094), 531–532.

Kumar, A., Vlachos, I., Aertsen, A., & Boucsein, C. (2013). Challenges of understanding brain function by selective modulation of neuronal subpopulations. *Trends in neurosciences*, *36*(10), 579–586.

Li, J., Chen, X., Hovy, E., & Jurafsky, D. (2015). Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066*.

Morcos, A. S., Barrett, D. G., Rabinowitz, N. C., & Botvinick, M. (2018). On the importance of single directions for generalization. *arXiv preprint arXiv:1803.06959*.