A conversation with Professor Stephen Baccus, January 22, 2020

Participants

- Prof. Stephen Baccus - Professor of Neurobiology, Stanford University
- Joseph Carlsmith - Research Analyst, Open Philanthropy

Note: These notes were compiled by Open Philanthropy and give an overview of major points made by Prof. Baccus.

Summary

Open Philanthropy spoke with Prof. Stephen Baccus of Stanford University as part of its investigation of what we can learn from the brain about the computational power ("compute") sufficient to match human-level task performance. The conversation focused on the compute necessary to replicate the information-processing performed in the retina.

Recent work on CNN retinal models

Prof. Baccus and his collaborators have been developing convolutional neural network models (CNNs) that can be used to predict the firing patterns of retinal ganglion cells in response to naturalistic stimuli. These models are getting close to matching the variability in the response exhibited by the retina itself (the model's correlation with the retina's response to a stimulus is about 85-90% as high as the correlation between retinal responses across different trials with that stimulus, though this varies somewhat across cells). They have also observed similarities between the behavior of the internal units in the model, and that of interneurons in the retina -- similarities that emerge just from training on retinal ganglion cell spike data.

To gather this data, Prof. Baccus and collaborators record from a few dozen retinal ganglion cells per experiment, though the number varies somewhat across experiments. There is no special selection involved in choosing which cells to test, and Prof. Baccus would expect similar success with arbitrary sets of retinal ganglion cells, though one cannot account for every cell under every condition without testing it.

The CNN Prof. Baccus and his collaborators use has three layers. They've also been exploring models that attempt to capture the anatomy of the retina more accurately. The

model uses firing rate neurons, with responses averaged over 10 ms, but it would be easy to translate this into a spiking model, in which the firing rate drives a poisson process.

What these models don't yet do

Here are a few aspects of retinal computation missing from these models:

- These models focus on replicating the response of an individual retinal ganglion cell to a stimulus. However, it may also be necessary to replicate correlations between the responses of different cells in the retina, as these may carry important information. Some people think that replicating the firing patterns of individual cells is enough, but most people think that correlations are important. Prof. Baccus's lab has not yet assessed their model's accuracy with respect to these between-cell correlations, though it is on their agenda.
- The visual system works under a wide range of conditions -- for example, varying light levels and varying contrast levels. Experiments focused on a set of natural scenes only cover some subset of these conditions. For example, Prof. Baccus's lab has not really tested dim light, or rapid transitions between bright and dim light.

Standard of success

Prof. Baccus expects that there would be consensus in the field that if a model's correlation with an individual cell's response to a stimulus matches the correlation between that cell's responses across different trials with that stimulus, and the model also captures all of the higher-order correlations across different cells, this would suffice to capture everything that the retina is communicating to the brain. Indeed, it would do so almost by definition.

However, various correlation coefficient measures and information theory measures do not address the importance of the meaning of a given signal. For example, if your model misses a tiger hiding in the bushes, that's pretty important, even though the difference might account for only a very small fraction of the correlation coefficient between your model and the retina's response.

Prospects for success

Prof. Baccus thinks it likely that models of the type he has been working with would be adequate for replicating retinal computation, given arbitrary access to data about how different cells in the retina respond to different inputs. A fairly small network would likely be enough.

Prof. Baccus thinks that people familiar with recent work on the retina, and focused on capturing the retina's input-output properties, would generally appreciate that full models capable of replicating retinal responses to natural scenes are in sight, and that models of the same type as the fairly small models Prof. Baccus's lab uses, scaled up to the retina as a whole, would likely be adequate for this task.

Necessary compute to replicate the whole retina

We can think of the retina as receiving a 100 megapixel input and outputting a 1 megapixel output (though in bright light, it's more like 5 million inputs, because there are 5 million cones and 95 million rods). And there are something like 10 million other cells in the retina.

To build a functional computational model of the retina as a whole, you could use a linear filter and a threshold as a model unit, and you could have something like one model unit per cell in the retina. However, in some of Prof. Baccus's models, they have less than this. Whether you'd need e.g. one model unit for every interneuron, or one for every two or three interneurons, isn't clear, but it's around that order of magnitude.

Prof. Baccus does not think simulating more complex aspects of neuron biology, like dendrites, compartments and ion channels, would be necessary for replicating the retina's input-output relationship.

Dendritic processing

Prof. Bartlett Mel at the University of Southern California has popularized the idea that dendrites in the cortex act as a separate functional unit. On this view, a single neuron is the equivalent of a two-layer neural network, in which dendrites in the first layer perform their own separate integration and then pass the output to the soma. Prof. Mel has estimated the necessary number of dendritic units per cell. Prof. Baccus recalls that the estimate was something like five. This, too, would not require detailed simulation of dendrite biophysics.

There are dendritic spiking processes in retinal cells as well, so technically three layers of model units might not be enough. If you had two layers for the ganglion cells and two layers for the amacrine cells, you might need five layers instead of three. Prof. Bacccus's lab has not yet seen an improvement in performance from adding layers, but at times he has thought that maybe another layer is necessary.

Necessary biophysical detail

Neuroscientists study problems in different ways. One camp focuses on the meaning of neural signals. This camp mostly uses a class of models known as encoding models -- e.g., linear-non-linear models, deep networks, and other variations. Another camp uses statistical models to try to capture firing patterns, while remaining agnostic about both the meaning and the biophysics. And a third camp focuses on trying to understand and simulate biophysical mechanisms, while largely ignoring the meaning of signals.

There is a lot of dissociation between these camps. One possible argument in favor of detailed biophysical modeling is that it could reveal emergent phenomena that we otherwise would not know could emerge. There is also the argument that given that we don't know what all of the brain's biophysical mechanisms are for, we should include them in our models and study them in order to figure it out.

Encoding models have had the most success in replicating functionally-relevant input-output relationships, and these don't focus on the biophysics very much. Prof. Baccus does not think there is a widespread camp of people who study encoding models, but who think that you need very complex, biophysical models in order to replicate neural computation.

Indeed, in many cases, in the retina, detailed biophysical dynamics combine to produce a fairly linear system. For example, photoreceptors involve a complex biochemical cascade, but at any one mean light level, they're pretty linear, and you can implement them using a linear filter. And as the light level changes, they adjust gain and timing based on the intensity of the light -- another complex biophysical process, but one that can be replicated in software fairly easily.

You can also borrow ideas from biophysics and then simplify them. For example, Prof. Baccus's lab has used a biophysical model used for capturing biophysical details like synaptic vesicles and ion channels, and abstracted the elements of that model to create a simplified model of retinal adaptation that does a good job of capturing adaptation across a wide range of contexts.

Indeed, Prof. Baccus expects that you could probably make his CNN retinal models simpler if you added in some of the non-linear mechanisms present in the biophysics, though he would not advocate throwing in every biophysical mechanism.

Generalizing from the retina to the brain

One lesson from the last few decades is that a lot of things people thought were happening in the cortex are first happening in the retina (and maybe only in the retina). The cortex is probably performing many computations similar to what the retina performs -- for example, creating efficient and sparse representations, adjusting to the changing statistics of the inputs, and in some cases removing correlations from the input.

There is also recent evidence from the Allen Institute of Brain Sciences pointing to parallels between the architecture of the retina and the architecture of the cortex. Specifically, in both the retina and cortex, you see a small number of cell types functioning as inputs, and a larger number functioning as outputs.

Prof. Dan Yamins and Prof. James DiCarlo's labs have compared activations in neural network models trained to recognize images with the activity patterns in different layers of the cortex, and they see a lot of parallels (though not a one-to-one correspondence). This might suggest that modeling methods that work in the retina will also work in different levels of the cortex.

Prof. Baccus thinks the answer is "maybe" to the question of whether the compute necessary to model neurons in the retina will be similar to the compute necessary to model neurons in the cortex. You might expect a volume by volume comparison to work as a method of scaling up from the retina to the cortex.

Differences between the cortex and the retina

There are, though, a number of differences between the cortex and the retina. For example:

- There is higher connectivity in the cortex than in the retina. A neuron in the cortex can have tens of thousands of inputs. In a model, it's easy to implement another connection between model units, but whether one can optimize such models in the same way isn't clear.
- Recurrence might be the trickiest difference. The retina can be largely approximated as a feedforward structure (there is some feedback, but a feedforward model does pretty well), but in the cortex there is a lot of feedback between different brain regions. This might introduce oscillations and feedback signals that make precise details about spike timings (e.g., at a 1 ms level of precision) more important, and therefore make firing rate models, which blur over 10 ms, inadequate. That said, even if this were true, Prof. Baccus expects that there would be a way of

implementing these dynamics that does not involve modeling detailed biophysical mechanisms. For example, you could find a way to keep track of timing in the model.

Learning

Computational descriptions of learning in the brain are pretty simple. For example, Hebbian plasticity, anti-Hebbian plasticity, timing-dependent plasticity, and neuromodulation can all be implemented in a pretty simple way, without using full biophysical models.

Neuromodulation

Prof. Eve Marder's lab has shown that there are dozens of neuromodulators at work in something as simple as a crab stomach, and these change how a neural circuit works. In biology, this functionality requires synthesizing the relevant signaling molecules, releasing them, sensing them via a receptor that keeps them separate from other signals, and so forth. In software, though, these processes can be made very simple. For example, if you have 20 neuromodulators, you can use a vector with 20 members.

CNN FLOP/s

Prof. Baccus and his colleagues have calculated that their CNN requires ~20 billion floating point operations to predict the output of one ganglion cell over one second (these numbers treat multiply and addition as separate operations - if we instead counted multiply-add operations (MACCs), the numbers would drop by a factor of roughly 2).

The input size is 50x50 (pixels) x 40 time points (10 ms bins). Layer 1 has 8 channels and 36x36 units with 15x15 filters each. Layer 2 has 8 channels and 26x26 units with 11x11 filters each. Layer 3 (to the ganglion cell) is a dense layer with a 8x26x26 filter from layer 2.

This leads to the following calculation for one ganglion cell:

- Layer 1: (40x15x15x2 + 1 (for the ReLU)) x 36 x 36 units x 8 channels = 1.87e8
- Layer 2: (8x11x11x2 + 1) x 26 x 26 units x 8 channels = 1.05e7
- Layer 3: 8x26x26x2 = 10,816.

Total: 1.97e8 FLOP per 10 ms bin. Multiplied by 100, this equals 1.97e10 FLOP/s.

Simulating more ganglion cells simultaneously only alters the last layer of the network, and so results in only a relatively small increase in computation. A typical experiment involves around 5-15 cells, but Prof. Baccus can easily imagine scaling up to 676 cells (26x26 — the

size of the last layer), or to 2500 (50x50 — the size of the input). 676 cells would require 20.4 billion FLOPs per second. 2500 would require 22.4 billion.

The largest amount of computation takes place in the first layer of the network. If the input size was larger, these numbers would scale up.

Other people to talk to

- Prof. Michael Berry - Associate Professor of Neuroscience, Princeton University
- Prof. Kwabena Boahen - Professor of Bioengineering and Electrical Engineering, Stanford University
- Prof. Henry Markram - Professor at the École Polytechnique Fédérale de Lausanne and Director of the Blue Brain Project