A conversation with Dr. Kate Storrs, June 11, 2020

Participants

- Dr. Kate Storrs - Alexander von Humboldt Research Fellow, Justus Liebig University
- Joseph Carlsmith - Research Analyst, Open Philanthropy

Note: These notes were compiled by Open Philanthropy and give an overview of the major points made by Dr. Storrs.

Summary

Open Philanthropy spoke with Dr. Kate Storrs of Justus Liebig 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 deep neural networks as models of human cognition.

Comparing brains and artificial neural networks

There is a lot of debate in neuroscience about what you need to include in a model in order to call it a model of the brain. The units in artificial neural network models are highly abstracted, and they leave out a lot of biophysical complexity. However, Dr. Storrs thinks it is at least interesting to pursue this level of abstraction, as you can't know in advance which biological idiosyncrasies are strictly computationally necessary, and which are accidents of implementation.

If you pursue this level of abstraction, then one way you might compare your network's capacity to the brain's is in terms of number of units or connections between units, which have been scaling up exponentially in recent years. The new GPT-3 model, for example, has 175 billion trainable parameters. This is still far off the brain, which probably has 100 trillion or maybe even a 1000 trillion synapses (if we treat synapses as relevantly analogous to connections between units, see below). However, it's not *that* wildly off.

Ten years ago the concept of having models only three orders of magnitude smaller than the brain was unfathomable. Today such models still require enormous computational resources, but we're much closer to being within the brain's range.

Synapses as trainable parameters

Dr. Storrs' sense is that, in the parts of the field she engages with most closely (e.g., systems level modeling, visual/cognitive/perceptual modeling, human behavior), and maybe more broadly, a large majority of people treat synaptic weights as the core learned parameters in the brain.

That said, she is not a neurophysiologist, and so isn't the right person to ask about what sort of biophysical complexities could imply larger numbers of parameters. She is peripherally aware of papers suggesting that glia help store knowledge, and there are additional ideas as well. The truth probably involves mechanisms other than synaptic weights, but she believes that the consensus is that such weights hold most of the knowledge.

Comparisons between deep neural networks and the visual system

It's hard to say how much of the task that the brain's visual system is performing is captured by a network trained on e.g. image recognition. Depending on the size of the image, such a network may be receiving far less information as input (e.g., a 128-pixel image, vs. the size of the whole retina) than the brain. It's also performing a more limited task using that information.

Returning the name of the main object in an image is a tiny portion of what the visual system can do. Core vision involves understanding the visual world as a navigable 3D space of objects, equipped with orientations, materials, depth, properties, and behavioral affordances. Dr. Storrs would guess that object-recognition only occurs on top of that kind of description of the world. Models analogous to the visual system would need to perform a wider range of the tasks that the visual system performs, which suggests that they would need to be more powerful.

Deep neural network vision models are also fragile, as a recent paper by Mr. Robert Geirhos and collaborators on "shortcut learning" nicely summarizes. This fragility isn't very surprising. If you train a network only on the task of naming objects, this objective won't drive the kind of robust visual representation that the brain employs.

On the other hand, a lot of our impression of the richness of human vision is illusory. For example, we don't see crisply, or in color, in the periphery of our visual field. So perhaps biological vision uses its own shortcuts.

Models of primary visual cortex (V1)

In order to model V1, plausibly you'd need lateral recurrent connections, because a lot of the activity of V1 neurons depends on the activity of neighboring neurons sensitive to nearby spatial regions. However, it seems possible to estimate the compute required to simulate the functionally necessary aspects of V1's computation to within a couple of orders of magnitude (though simulating the rest of the biology is a very different story, and it can be hard to tell which processes are which).

In Dr. Storrs' area of neuroscience, there can be a narrative to the effect that: "the early visual system is basically done. We understand the canonical computations: e.g., edge, orientation and color selection. You link them up with local exhibition and inhibition, and you have feedback that probably has some kind of predictive function (e.g., you get less and less response from V1 neurons to a predictable stimulus, suggesting that feedback is creating some kind of short-term memory). Once you've got all of this, you can explain most of V1 activity." (This is not necessarily Dr. Storrs' view; it's just a summary of a common narrative.)

However, if you spoke to a V1 electrophysiologist, they might emphasise aspects of V1 neural activity which we cannot yet explain, or the fact that some V1 neurons (20-30%) don't appear to be visually responsive.

Later regions

Some recent work from the labs of Prof. Jim DiCarlo and Prof. Dan Yamins has focused on using task-trained deep neural networks to predict activity in later visual regions, e.g., inferior temporal cortex (IT). The sense in which these task-trained networks are models of the visual system is a point of contention in the visual neuroscience field at the moment.

On the one hand, these models can explain a lot more variance in IT than baseline models/techniques (e.g. banks of Gabor filters, earlier computer vision models, or use of pixel-wise similarity). And these models also make available new applications, such as possible ways of decoding the images that people are seeing, or ways of improving brain-computer interfaces.

On the other hand, these models do not provide the same kind of intuitive and interpretable explanation that some previous models of vision provided. And you can argue that while deep neural networks give us a feature space that we can combine to predict a neuron's activity, they don't really yet tell us what the computational role of that neuron is, and we already knew that cells high-up in the visual system were responsive to complex features: e.g., there are face-selective cells, and cells selective for particular conjunctions of shapes (though the fact that you see differential predictive success using different layers of the network does point to the visual complexity that different biological neurons are tuned to).

Some people think that the fact that you get much better predictions of neural activity by recombing and reweighting the features in a deep neural network layer means that these networks aren't very good models of the visual system. On this view, a good model ought to learn the specific features that neurons in the visual system detect, or the specific distribution of features.

Alternatively, though, one might think that the specific feature that any neuron is tuned to is fairly arbitrary. E.g., if you tried to predict the activity of a neuron in one person's brain, on the basis of the activity of a neuron in another person's brain, you won't get a 1-1 correspondence, so we shouldn't expect that a model will arrive at exactly the tunings and prevalences in brains. Rather, it's more likely that there are sets of features that serve as a good basis set for then recombining to predict neuron behavior.

Most papers in this area will have a noise ceiling, indicating, for example, how well you can predict one monkey's brain activity using another monkey's brain activity, or one human's fMRI data using data from a collection of people. These are important upper bounds: any model of "the brain" can only ever be a model of the average brain (though depending on your experiment, the relevant noise ceiling may be across trials for an individual rather than across many individuals).

Image categories humans can recognize

The question of how many categories humans can recognize is sort of impossible, because the concept of a category is fairly fuzzy, and it isn't rich enough to capture what human visual recognition involves. For example, you've probably seen tens of thousands of chairs over the course of your life. You were able to immediately recognize them as chairs, but you were also able to immediately see a large number of individuating properties. Indeed, one of the great powers of the visual system is that it arrives at a description that is flexible enough that you can then carve it up in whatever ways are behaviorally relevant.

Looking at common nouns, and budgeting a certain number of instances of each (maybe 100 or 1000) as individually recognizable, might be one way to put a very rough number on the categories that humans can recognize.

All Open Philanthropy conversations are available at
http://www.openphilanthropy.org/research/conversations