A conversation with Professor Jared Kaplan, January 23, 2020

Participants

- Prof. Jared Kaplan - Professor of Physics, Johns Hopkins 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 Prof. Kaplan.

Summary

Open Philanthropy spoke with Prof. Jared Kaplan of Johns Hopkins 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 application of Landauer's principle to the brain.

Landauer's principle

Landauer's principle states that erasing a bit of information requires a minimum energy expenditure -- specifically, $kT \ln 2$, where k is Boltzmann's constant, and T is the absolute temperature. This principle is grounded in the relationship between entropy and energy -- the same relationship that grounds the fact that heat doesn't flow from cold things to hot things, and the fact that you can't create a perpetual motion machine or an arbitrarily efficient engine.

For physicists, entropy is the logarithm of the number of accessible states. When a system changes, either this entropy stays the same, or it increases. Almost all fixed systems have more accessible states as the energy goes up. Temperature just is how the energy changes as the entropy changes (textbooks will often state this as: the reciprocal of the temperature is the derivative of the entropy with respect to the energy).

As an intuitive example: if your system (e.g., a set of gas molecules) has no energy at all, then all your molecules are just lying on the floor. As you add energy, they can bounce around, and there many more configurations they can be in.

The energy of a single moving particle is another example. It's kinetic energy is $\frac{1}{2} \times \text{mass} \times \text{velocity}^2$. The velocity is a vector, which in a three dimensional space will live on some

sphere. As you make the energy bigger, the surface area of this sphere increases. This corresponds to a larger number of accessible states (at the quantum mechanical level, these states are discrete, so you can literally count them).

Landauer and the brain

Mr. Carlsmith asked Prof. Kaplan's opinion of the following type of upper bound on the compute required to replicate the brain's task-performance. According to Landauer's principle, the brain, given its energy budget (~20 W) can be performing no more than ~1e22 bit-erasures per second. And if the brain is performing less than 1e22 bit-erasures per second, the number of FLOP/s required to replicate its task-performance is unlikely to exceed 1e22.

Prof. Kaplan thinks that this type of calculation provides a very reasonable loose upper bound on the computation performed by the brain, and that the actual amount of computation performed by the brain is almost certainly many orders of magnitude below this bound. Indeed, he thinks the true number is so obviously much lower than this that Landauer's principle does not initially seem particularly germane to questions about brain computation. One analogy might be attempting to upper bound the number of fraudulent votes in a US presidential election via the total population of the world.

However, he thinks that upper bounds based on Landauer's principle are a helpful counter to views on which "we really just don't know" how much computation the brain performs, or on which doing what the brain does requires the type of compute that would be implicated by very detailed biophysical simulations.

Hypotheses about brain computation

Prof. Kaplan places most of his probability mass on the hypothesis that most of the computation performed by the brain is visible as information transferred between synapses. However, in order to get close to the Landauer limit, you have to assume something very different from this.

It is theoretically possible that there is a large amount of additional computation taking place within neurons, but this seems very implausible, and Prof. Kaplan finds it difficult to evaluate arguments that condition on this possibility. One reason this seems implausible is that neurons aren't that different across species, and it does not seem plausible to Prof. Kaplan that in simple species with very few neurons, large amounts of computation are

taking place inside the neurons. One would need a story about when this complex internal computation developed in the evolutionary history of neurons.

Reversibility in biological systems

Prof. Kaplan expects that the operations performed in neurons and cells are not even close to being thermodynamically reversible. In general, Prof. Kaplan thinks it unlikely that big, warm things are performing thermodynamically reversible computations.

If you're in a regime where there is some signal to noise ratio, and you make your signal big to avoid noise, you can't be doing something thermodynamically reversible: the noise is creating waste heat, and you're extending your signal to get above that. Prof. Kaplan would have thought that basically all of the processes in the brain have this flavor.

For example, a lot of synapses, not too dissimilar from synapses in the brain, are used to send information to e.g. a muscle. Those synapses are using a lot of energy, and the brain is clearly going through a lot of effort to convey the relevant information confidently.

Processes that involve diffusion also cannot be thermodynamically reversible. Diffusion increases entropy. For example, if you take two substances and mix them together, you have increased the entropy of that system.

In general, it's extremely difficult to build reversible computers. For example, all of the quantum computers we have are very rudimentary (quantum computers are a type of reversible computer), and it's hard to keep them running for very long without destroying information.

In order to be performing thermodynamically reversible computations, each neuron would have to have some sort of very specialized component, operating in a specialized environment crafted in order to perform the computation in a thermodynamically reversible way. It would be hard to keep this running for very long, and Prof. Kaplan doesn't think this is happening.

RNA synthesis

Prof. Kaplan is surprised to hear that RNA synthesis is sometimes mentioned as an example of a close-to-reversible computational process in biological systems, and he would be curious to hear details about the type of reversibility in question. There is an important

difference between processes that are thermodynamically reversible, and those that are reversible in the same sense that e.g. zipping up a zipper is reversible -- e.g., you can undo it, but the process still involves creating a lot of waste heat.

Unitary matrices

Prof. Kaplan does not think that operations involving unitary matrices are a helpful example of a form of reversible computation that the brain could be performing.

FLOPs required per bit-erasure in the brain

Prof. Kaplan's intuition is that it is very unlikely that many FLOPs are required to do whatever the brain does per bit-erasure.

FLOPs in actual computers erase bits, and Prof. Kaplan expects that you generally have order one bit-erasures per operation in computational systems. That is, you don't do a lot of complicated things with a bit, and then erase it, and then do another set of very complicated things with another bit, and then erase it, etc. Prof. Kaplan's intuition in this respect comes from his understanding of certain basic operations you can do with small amounts of information.

In principle you can perform a very complicated set of transformations on a piece of information, like an image, without erasing bits. Prof. Kaplan can imagine some kind of order one factor increase in required compute from this type of thing.

At a thermodynamic level, though, implementing any process without erasing information is very difficult. And if you're doing a large number of complicated operations per bit-erasure, you're getting close to thermodynamic reversibility. Prof. Kaplan thinks that the brain is probably not getting anywhere close to being thermodynamically reversible, and therefore isn't doing a lot of complicated operations per bit-erasure.