THEORIES OF NEURAL CODING AND REPRESENTATION

Our work explores representations in artificial "neurons". Neuroscientists study similar questions in biological neurons. There are a variety of theories for how information could be encoded by a group of neurons. At one extreme is a *local code*, in which every individual stimulus is represented by a separate neuron. At the other extreme is a *maximally-dense distributed code*, in which the information-theoretic capacity of the population is fully utilized, and every neuron in the population plays a necessary role in representing every input.

One challenge in comparing our work with the neuroscience literature is that a "distributed representation" seems to mean different things. Consider an overly-simplified example of a population of neurons, each taking a binary value of active or inactive, and a stimulus set of sixteen items: four shapes, with four colors (example borrowed from [4]). A "local code" would be one with a "red triangle" neuron, a "red square" neuron, and so on. In what sense could the representation be made more "distributed"? One sense is by representing *independent features* separately — e.g. four "shape" neurons and four "color" neurons. A second sense is by representing *more items than neurons* — i.e. using a binary code over four neurons to encode 2^4 = 16 stimuli. In our framework, these senses correspond to *decomposability* (representing stimuli as compositions of independent features) and *superposition* (representing more features than neurons, at cost of interference if features co-occur).

Decomposability doesn't necessarily mean each feature gets its own neuron. Instead, it could be that each feature corresponds to a "direction in activation-space"[25], given scalar "activations" (which in biological neurons would be firing rate). Then, only if there is a *privileged basis*, "feature neurons" are incentivized to develop. In biological neurons, metabolic considerations are often hypothesized to induce a privileged basis, and thus a "sparse code". This would be expected if the nervous system's energy expenditure increases linearly or sublinearly with firing rate.[26] Additionally, neurons are the units by which biological neural networks can implement non-linear transformations, so if a feature needs to be non-linearly transformed, a "feature neuron" is a good way to achieve that.

Any decomposable linear code that uses orthogonal feature vectors is functionally equivalent from the viewpoint of a linear readout. So, a code can both be "maximally distributed" — in the sense that every neuron participates in representing every input, making each neuron extremely polysemantic — and also have no more features than it has dimensions. In this conception, it's clear that a code can be fully "distributed" and also have no superposition.

A notable difference between our work, and the neuroscience literature we have encountered, is that we consider as a central concept the likelihood that features co-occur with some probability.[27] A "maximally-dense distributed code" makes the most sense in the case where items never co-occur; if the network only needs to represent one item at a time, it can tolerate a very extreme degree of superposition. By contrast, a network that could plausibly need to represent all the items at once can do so without interference between the items if it uses a code with no superposition. One example of high feature co-occurrence could be encoding spatial frequency in a receptive field; these visual neurons need to be able to represent white noise, which has energy at all frequencies. An example of limited co-occurrence could be a motor "reach" task to discrete targets, far enough apart that only one can be reached at a time

.

One hypothesis in neuroscience is that highly compressed representations might have an important use in long-range communication between brain areas[57]. Under this theory, sparse representations are used within a brain area to do computation, and then are compressed for transmission across a small number of axons. Our experiments with the absolute value toy model shows that networks can do useful computation even under a code with a moderate degree of superposition. This suggests that all neural codes, not just those used for efficient communication, could plausibly be "compressed" to some degree; the regional code might not necessarily need to be decompressed to a fully sparse one.

It's worth noting that the term "distributed representation" is also used in deep learning, and has the same ambiguities of meaning there. Our sense is that some influential early works (*e.g.* [5]) may have primarily meant the "independent features are represented independently" *decomposability* sense, but we believe that other work intends to suggest something similar to what we call superposition.

Comments & Replications

Inspired by the original Circuits Thread and Distill's Discussion Article experiment, the authors invited several external researchers who we had previously discussed our preliminary results with to comment on this work. Their comments are included below.

REPLICATION & FORTHCOMING PAPER

Kshitij Sachan *is a research intern at Redwood Research.*

Redwood Research has been working on toy models of polysemanticity, inspired by Anthropic's work. We plan to separately publish our results, and during our research we replicated many of the experiments in this paper. Specifically, we replicated all plots in the Demonstrating Superposition and Superposition as a Phase Change sections (visualizations of the relu models with different sparsities and the phase diagrams) as well as the plot in The Geometry of Superposition – Uniform Superposition. We found the phase diagrams look quite different depending on the activation function, suggesting that in this toy model some activation functions induce more polysemanticity than others.

Original Authors' Response: Redwood's further analysis of the superposition phase change significantly advanced our own understanding of the issue – we're very excited for their analysis to be shared with the world. We also appreciate the independent replication of our basic results.

REPLICATION & FURTHER RESULTS

Tom McGrath *is a research scientist at DeepMind.*

The results in this paper are an important contribution - they really further our theoretical understanding of a phenomenon that may be central to interpretability research and understanding network representations more generally. It's surprising that such simple settings can produce these rich phenomena. We've reproduced the experiments in the Demonstrating Superposition and Superposition as a Phase Change sections and have a minor additional result to contribute.

It is possible to exactly solve the expected loss for the $n = 2, m = 1$ case of the basic ReLU output toy model (ignoring bias terms). The derivation is mathematically simple but somewhat long-winded: the 'tricks' are to (1) represent the sparse portion of the input distribution with delta functions, and (2) replace the ReLu with a restriction of the domain of integration:

$$\int_D \mathrm{ReLU}(f(x))dx = \int_{D \cap f(x) > 0} f(x)dx$$

Making this substitution renders the integral analytically tractable, which allows us to plot the full loss surface and solve for the loss minima directly. We show some example loss surfaces below: