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: