Toy Models of Superposition

AUTHORS

Nelson Elhage*, Tristan Hume*, Catherine Olsson*, Nicholas Schiefer*, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg*, Christopher Olah‡

AFFILIATIONS

Anthropic, Harvard

PUBLISHED

Sept 14, 2022

* Core Research Contributor; ‡ Correspondence to colah@anthropic.com; Author contributions statement below.

Abstract: Neural networks often pack many unrelated concepts into a single neuron – a puzzling phenomenon known as 'polysemanticity' which makes interpretability much more challenging. This paper provides a toy model where polysemanticity can be fully understood, arising as a result of models storing additional sparse features in "superposition." We demonstrate the existence of a phase change, a surprising connection to the geometry of uniform polytopes, and evidence of a link to adversarial examples. We also discuss potential implications for mechanistic interpretability.

We recommend reading this paper as an HTML article.

It would be very convenient if the individual neurons of artificial neural networks corresponded to cleanly interpretable features of the input. For example, in an "ideal" ImageNet classifier, each neuron would fire only in the presence of a specific visual feature, such as the color red, a left-facing curve, or a dog snout. Empirically, in models we have studied, some of the neurons do cleanly map to features. But it isn't always the case that features correspond so cleanly to neurons, especially in large language models where it actually seems rare for neurons to correspond to clean features. This brings up many questions. Why is it that neurons sometimes align with features and sometimes don't? Why do some models and tasks have many of these clean neurons, while they're vanishingly rare in others?

In this paper, we use toy models — small ReLU networks trained on synthetic data with sparse input features — to investigate how and when models represent more features than they have dimensions. We call this phenomenon **superposition**. When features are sparse, superposition allows compression beyond what a linear model would do, at the cost of "interference" that requires nonlinear filtering.

Consider a toy model where we train an embedding of five features of varying importance[1] in two dimensions, add a ReLU afterwards for filtering, and vary the sparsity of the features. With dense features, the model learns to represent an orthogonal basis of the most important two features (similar to what Principal Component Analysis might give us), and the other three features are not represented. But if we make the features sparse, this changes: