Toy Models of Superposition Replication and Findings

Abstract

Toy Models of Superpostion[1] is a groundbreaking paper published by researchers affilated with Anthropic and Harvard University in 2022. The paper demonstrates that neural networks can represent more features than they have demensions buy training small models with under 100 neurons. Additionally, they use these so called "toy models" to understand the relationship between how neural networks are trained and how they represent the data internally. This paper was able to the finding from this paper and make new observations about "toy models" and how they behave under different training circumstances.

1 Background and Motivation

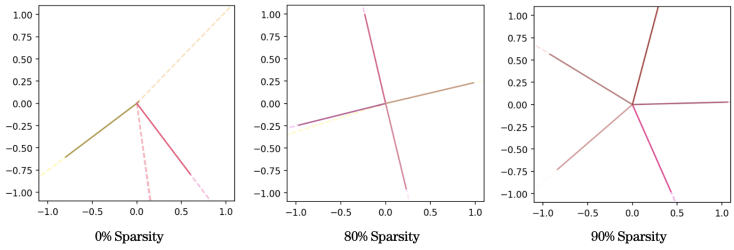

Figure 1: Replicated feature directions example.

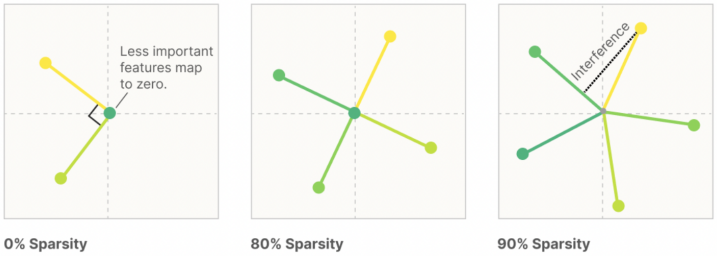

Figure 2: Graphic from Anthropic Paper

2 Demonstrating Superposition

In this section of Toy Models of Superpostion[1], the researchers provide context and define terms. This provides a good overveiw for people unfamiliar with some of the concepts discussed in this replication.

In this paper, I make a few additional comments about some key terms to provide context in this paper.

(1) Defining Features: The original paper Toy Models of Superpostion defines features broadly as "properties of the input which a sufficiently large neural network will reliably dedicate a neuron to representing." The authors do however describe this definition as "slightly circular" and note that they are not "overly attached to it." I find the definition especially problematic because a network that is small or has unconventional archetecture may represent a feature that a larger network or a network with a more typical archetecture may represent. These representations are clearly still features, but are not treated as so under the original definition.

As a result, I propose an alternative definition: features are aspects of the input that a neural network represents accuratly with a higher propability than a randomly initialized network. In other words, features are parts of the input that a model determines to be important enough to represent internally.

(2) Role of linear functions in Neural Networks: The original authors of Toy Models of Superpostion strongly emphasize the role of linear functions in neural networks. They claim that "Linear representations are the natural format for neural networks to represent information in!" Although this is clearly mostly true, I dont think there is sufficient evidence to justify the confidence of the authors in the original paper. Nonlinearities are also clearly doing lots of useful computation and it seems just as important to study their role in how a neural network represents features from layer to layer. The nonlinearity at each layer of a neural network may be just as important as the linear transformation in understanding how a model makes decisions.

(3) Defining Superposition: The original paper has a fantastic and simple defination for Superposition: "Roughly, the idea of superposition is that neural networks 'want to represent more features than they have neurons', so they exploit a property of high-dimensional spaces to simulate a model with many more neurons." This is the definition I will used throughout this paper.

References

[1] 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, and Christopher Olah. Toy models of superposition, 2022.