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 Introduction

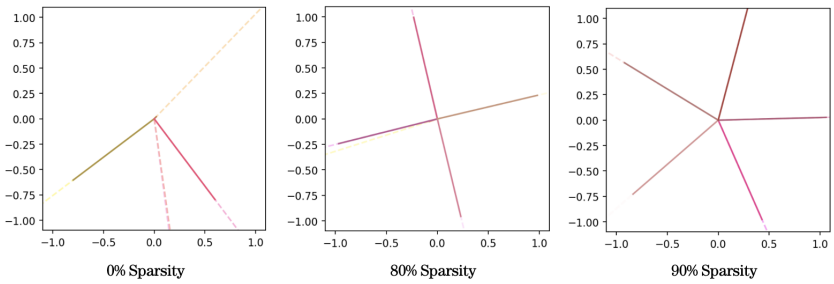

Figure 1: Replicated feature directions example.

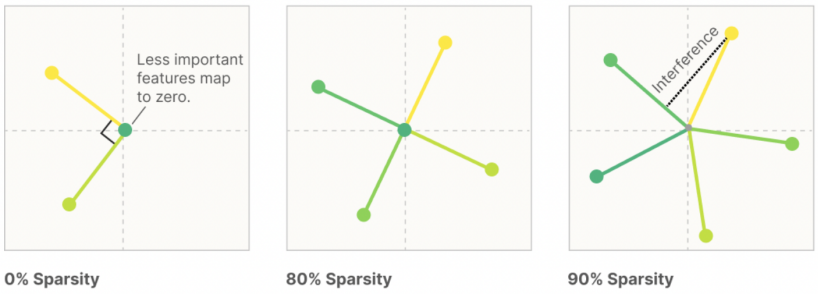

Figure 2: Graphic from Anthropic Paper

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.