Strategic Picture - The strategic picture articulated in this paper – What does superposition mean for interpretability and safety? What would a suitable solution be? How might one solve it? – developed in extensive conversations between authors, and in particular Chris Olah, Tristan Hume, Nelson Elhage, Dario Amodei, Jared Kaplan. Nelson Elhage recognized the potential importance of "enumerative safety", further articulated by Dario. Tristan brainstormed extensively about ways one might solve superposition and pushed Chris on this topic.

Writing - The paper was primarily drafted by Chris Olah, with some sections by Nelson Elhage, Tristan Hume, Martin Wattenberg, and Catherine Olsson. All authors contributed to editing, with particularly significant contributions from Zac Hatfield Dodds, Robert Lasenby, Kipply Chen, and Roger Grosse.

Illustration - The paper was primarily illustrated by Chris Olah, with help from Tristan Hume, Nelson Elhage, and Catherine Olsson.

Citation Information

Please cite as:

```
Elhage, et al., "Toy Models of Superposition", Transformer Circuits Thread, 2022.
```

BibTeX Citation:

```
@article{elhage2022superposition,
   title={Toy Models of Superposition},
   author={Elhage, Nelson and Hume, Tristan and Olsson, Catherine and Schiefer, Nicholas and Henighan,
Tom and Kravec, Shauna and Hatfield-Dodds, Zac and Lasenby, Robert and Drain, Dawn and Chen, Carol and
Grosse, Roger and McCandlish, Sam and Kaplan, Jared and Amodei, Dario and Wattenberg, Martin and Olah,
Christopher},
   year={2022},
   journal={Transformer Circuits Thread},
   note={https://transformer-circuits.pub/2022/toy_model/index.html}
}
```

Footnotes

1. Where “importance” is a scalar multiplier on mean squared error loss. [↩]

2. In the context of vision, these have ranged from low-level neurons like curve detectors [6] and high-low frequency detectors [18], to more complex neurons like oriented dog-head detectors or car detectors [1], to extremely abstract neurons corresponding to famous people, emotions, geographic regions, and more [19]. In language models, researchers have found word embedding directions such as a male-female or singular-plural direction [8], low-level neurons disambiguating words that occur in multiple languages, much more abstract neurons, and "action" output neurons that help produce certain words [2]. [↩]

3. This definition is trickier than it seems. Specifically, something is a feature if there *exists* a large enough model size such that it gets a dedicated neuron. This create a kind "epsilon-delta" like definition. Our present understanding – as we'll see in later sections – is that arbitrarily large models can still have a large fraction of their features be in superposition. However, for any given feature, assuming the feature importance curve isn't flat, it should eventually be given a dedicated neuron. This definition can be helpful in saying that something *is* a feature – curve detectors are a feature because you find them in across a range of models larger than some minimal size – but unhelpful for the much more common case of features we only hypothesize about or observe in superposition. [↩]

4. A famous book by Lakatos [23] illustrates the importance of uncertainty about definitions and how important rethinking definitions often is in the context of research. [↩]

5. This experiment setup could also be viewed as an autoencoder reconstructing x. [↩]

6. A vision model of sufficient generality might benefit from representing every species of plant and animal and every manufactured object which it might potentially see. A language model might benefit from representing each person who has ever been mentioned in writing. These are only scratching the surface of plausible features, but already there seem more than any model has neurons. In fact, large language models demonstrably do in fact know about people of very modest prominence – presumably more such people than they have neurons. This point is a common argument in discussion of the plausibility of "grandmother neurons'' in neuroscience, but seems even stronger for artificial neural networks. [↩]

7. For computational reasons, we won't focus on it in this article, but we often imagine an infinite number of features with importance asymptotically approaching zero. [↩]

8. The choice to have features distributed uniformly is arbitrary. An exponential or power law distribution would also be very natural. [↩]

9. Recall that $W^T = W^{-1}$ if W is orthonormal. Although W can't be literally orthonormal, our intuition from compressed sensing is that it will be "almost orthonormal" in the sense of Candes & Tao [25]. [↩]

10. We have the model be $x' = W^T W x$, but leave x Gaussianaly distributed as in Saxe. [↩]

11. As a brief aside, it's interesting to contrast the linear model interference, $\sum_{i \neq j} |W_i \cdot W_J|^2$, to the notion of coherence in compressed sensing, $\max_{i \neq j} |W_i \cdot W_J|$. We can see them as the L^2 and L^∞ norms of the same vector. [↩]

12. To prove that superposition is never optimal in a linear model, solve for the gradient of the loss being zero or consult Saxe et al. [↩]

13. Here, we use “phase change” in the generalized sense of “discontinuous change”, rather than in the more technical sense of a discontinuity arising in the limit of infinite system size. [↩]

14. Scaling the importance of all features by the same amount simply scales the loss, and does not change the optimal solutions. [↩]

15. Note that there's a degree of freedom for the model in learning W_1: We can rescale any hidden unit by scaling its row of W_1 by α, and its column of W_2 by α^{-1}, and arrive at the same model. For consistency in the visualization, we rescale each hidden unit before visualizing so that the largest-magnitude weight to that neuron from W_1 has magnitude 1. [↩]

16. These specific values were chosen to illustrate the phenomenon we're interested in: the absolute value model learns more easily when there are more neurons, but we wanted to keep the numbers small enough that it could be easily visualized. [↩]

17. One question you might ask is whether we can quantify the ability of superposition to enable extra computation by examining the loss. Unfortunately, we can't easily do this. Superposition occurs when we change the task, making it sparser. As a result, the losses of models with different amounts of superposition are not comparable – they're measuring the loss on different tasks! [↩]

18. Ultimately we want to say that a model doesn't implement some class of behaviors. Enumerating over all features makes it easy to say a feature doesn't exist (e.g. "there is no 'deceptive behavior' feature") but that isn't quite what we want. We expect models that need to represent the world to represent unsavory behaviors. But it may be possible to build more subtle claims such as "all 'deceptive behavior' features do not participate in circuits X, Y and Z." [↩]

19. Superposition also makes it harder to find interpretable directions in a model without a privileged basis. Without superposition, one could try to do something like the Gram–Schmidt process, progressively identifying interpretable

directions and then removing them to make future features easier to identify. But with superposition, one can't simply remove a direction even if one knows that it is a feature direction. [↩]

20. More formally, given a matrix $H \sim [d,m] ~=~ [h_0, h_1, \ldots]$ of hidden layer activations $h \sim [m]$ sampled over d stimuli, if we believe there are n underlying features, we can try to find matrices $A \sim [d,n]$ and $B \sim [n,m]$ such that A is sparse. [↩]

21. In particular, it seems like we should expect to be able to reduce superposition at least a little bit with essentially no effect on performance, just by doing something like L1 regularization without any architectural changes. Note that models should have a level of superposition where the derivative of loss with respect to the amount of superposition is zero – otherwise, they'd use more or less superposition. As a result, there should be at least some margin within which we can reduce the amount of superposition without affecting model performance. [↩]

22. A more subtle issue is that GANs and VAEs often assume that their latent space is Gaussianly distributed. Sparse latent variables are very non-Gaussian, but central limit theorem means that the superposition of many such variables will gradually look more Gaussian. So the latent spaces of some generative models may in fact force models to use superposition! [↩]

23. Note that this has a nice information-theoretic interpretation: $\log(1-S)$ is the surprisal of a given dimension being non-zero, and is multiplied by the expected number of non-zeros. [↩]

24. Note that in the compressed sensing case, the phase transition is in the limit as the number of dimensions becomes large - for finite-dimensional spaces, the transition is fast but not discontinuous. [↩]

25. We haven't encountered a specific term in the distributed coding literature that corresponds to this hypothesis specifically, although the idea of a "direction in activation-space" is common in the literature, which may be due to ignorance on our part. We call this hypothesis *linearity* [↩]

26. Experimental evidence seems to support this [55] [↩]

27. A related, but different, concept in the neuroscience literature is the "binding problem" [56] in which e.g. a red triangle is a co-occurrence of exactly one shape and exactly one color, which is not a representational challenge, but a binding problem arises if a decomposed code needs to represent simultaneously also a blue square — which shape feature goes with which color feature? Our work does not engage with the binding question, merely treating this as a co-occurrence of "blue", "red", "triangle", and "square". [↩]