Toy Models of Superposition Review

Sean Murphy[*]
Monash University
Melbourne, Australia
smur0055@student.monash.edu

Leo Qiao[*]
Monash University
Melbourne, Australia
lqia0021@student.monash.edu

1. Review

The *Toy Models of Superposition* paper [1] published in 2022 by researchers from Anthropic and Harvard University, investigates "superposition", a phenomenon where neural networks represent more features than they have available dimensions. They define a *feature* as an interpretable property of the input that neurons respond to. The paper focuses on how neural networks overcome their limited dimensions through the use of polysemantic neurons, which represent multiple features at once.

1.1. Summary

The authors use a toy model to demonstrate superposition by training it to project five features onto only two dimensions, then attempting to recover the original features. One key finding is that the sparsity of a feature—how rarely it occurs—is a primary factor in determining whether it is represented in superposition. For instance, when features are not sparse, only the two most important ones are represented in dedicated orthogonal dimensions. However, as sparsity increases, the model begins to encode all five features in superposition, arranging them geometrically in a pentagon-like structure within the two-dimensional space. This results in interference between the features, where one's presence also activates other unrelated neurons.

The middle sections of the paper provide a mathematical explanation for why superposition occurs, whilst also exploring the specific "phases" of superposition that can occur under different conditions. The paper concludes by discussing how "solving superposition" could lead to more interpretable and safer AI models. They suggest that a key step in achieving this would be the ability to identify and enumerate all features in a model—what they refer to as the fundamental units of a neural network. With such an understanding, one could potentially produce a universal quantifier over these features This framework could then be used to fully explain model behaviour. As a result, this would enable confidence that models would not simply avoid, but perhaps be entirely incapable of unethical behaviours like deception. The authors propose that this might be achieved either by designing models that avoid superposition entirely or by finding an overcomplete basis that disentangles the superimposed features, motivating future research directions.

1.2. Significance

The concept of superposition, as explored in this paper, holds significant implications for both our theoretical understanding of neural networks and practical advancements in mechanistic interpretability and AI safety. It provides an entirely new perspective on the fundamental units of neural networks by showing how polysemantic neurons can be broken down into multiple features. With over 200 citations in just 2 years, this paper has fueled a lot of modern day mechanistic interpretability research.

1.3. Relevance in the Deep Learning Era

In the current deep learning era, where advancements in model capabilities are positively correlated with model complexity, continuing to iterate this way only creates bigger, and bigger black boxes. The insights this paper presents regarding the superposition phenomenon suggest alternate directions that would allow us to better understand these black boxes by "solving superposition".

Additionally, this paper shows the importance of model architecture, illuminating the tradeoff between efficiency brought about by smaller model sizes, and the interference caused by the superpositional feature representations that it necessitates. As our understanding of model features matures, there is potential for model architecture hyperparams to be more analytically selected given information about the data, such as feature sparsity. If we are able to obtain universal quantifiers over features as the paper suggests, we may even see a shift in the deep learning era where engineers can build or modify parts of these networks manually, having the ability to dictate the inclusion of specific features, and change the semanticity of neurons at will.

[*] Sean and Leo contributed equally to Section 1. Sean wrote Section 2, and Leo wrote Section 3.

2. Superposition as a Cause of Hallucinations

A key question from the paper was whether interference from superposition causes hallucinations in large language models. To test this, we trained a classifier on high-sparsity data, inducing superposition, and evaluated it on low-sparsity data.

2.1. Dataset

We used a dataset with five features, represented by English affixes: *un-*, *re-*, *-able*, *-ful*, *-ness*, although any set of affixes could work. Two datasets were generated: one with words containing a single affix (pure words) and another with words containing multiple affixes (dual words, *e.g.*, *resourceful*). The pure dataset was split into training and test sets, while the dual dataset was reserved for evaluation.

2.2. Model

We trained a model with an embedding layer, a convolutional layer, a hidden layer, and ReLU activation. The hidden layer had only two neurons, forcing it to use superposition to encode all five features.

2.3. Results

Using the toy model framework, we visualized the five features in the hidden layer. Figure 1 shows the model encoded all features, though it didn't form the pentagon-like structure noted in the paper. Table 1 shows the results for the test set and the dual dataset. The model performed well on the test set, but accuracy dropped significantly on the dual dataset, even though each dual word had twice the correct labels. This suggests that when features are less sparse than in training, the additive interference from superposition may cause misinterpretations, contributing to hallucinations in large language models.

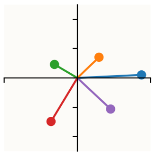

Figure 1. This figure can be reproduced with our Modelling Hallucinations Colab notebook.

Dataset	Accuracy
Test set	0.9708
Dual dataset	0.8651

Table 1. Model evaluation on the test set and the dual dataset.

3. Inhibiting Superpositions with Activations

The paper presents the activation function as a key in enabling the formation of superpositions due to their ability to filter out interference. As such, we demonstrate that modifying the activation can effectively control the model's usage of superpositions.

We propose ExReLU, a modified version of ReLU with a cut-off threshold of $t \in \mathbb{R}$, instead of 0. By setting $t < 0$, its filtering effect for negative interference is weakened, thus increasing the cost for the model to adopt superpositional representations.

$$\text{ExReLU}(x) = \begin{cases} x & \text{if } x \geq t \\ 0 & \text{if } x < t \end{cases}$$

3.1. Model

We trained two sets of models, each with hidden layers of two neurons on a dataset with five features whilst varying the feature sparsity. The first set used the normal ReLU activation, while the second used ExReLU with $t = -0.25$.

3.2. Results

We visualized the hidden layer's representation of the features using the previously mentioned toy model framework. Figure 2 shows that ExReLU has successfully suppressed the model's usage of superpositions in this instance. As such, it seems that the development of new activation functions akin to ExReLU may be key to "solving superposition" and controlling how models use their available dimensions as mentioned in the paper.

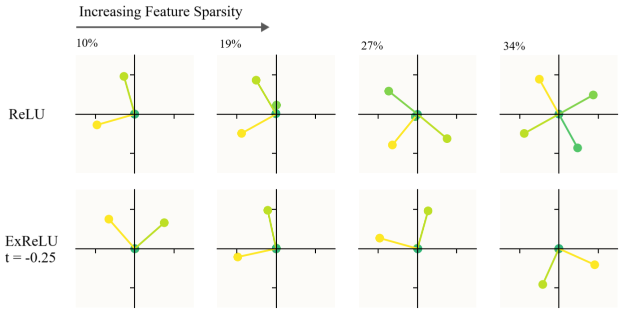

Figure 2. Comparison of feature embeddings by the two model sets, the code for which can be found in this notebook.

References

[1] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022. 1