Toy Models of Superposition

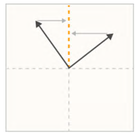
UNIVERSITY OF CAMBRIDGE

Guangyu Yang, Xueyan Li, Yawen Duan

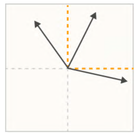
Background

Neural networks often contain "**polysemantic neurons**" that respond to multiple unrelated inputs.

e.g. 4e:55 in InceptionV1 responses to both cat faces and fronts of cars.

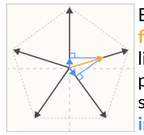


Polysemanticity is what we'd expect to observe if features were not aligned with a neuron.

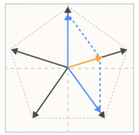


In the **superposition hypothesis**, features can't align with the basis because the **model embeds more features than there are neurons**. Polysemanticity is inevitable if this happens.

Superposition is when models represent more features than they have dimensions.



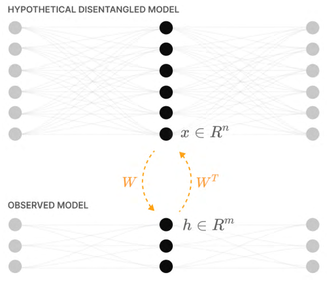
Even if only one sparse feature is active, using linear dot product projection on the superposition leads to interference which the model must tolerate or filter.



If the features aren't as sparse as a superposition is expecting, multiple present features can additively interfere such that there are multiple possible nonlinear reconstructions of an activation vector.

Goal: demonstrating superposition

Toy Model and Experiment Settings



- **Observations**:
  - Real-world feature is sparse -- S_i
  - More features than neurons -- $n > m$
  - Features vary in importance -- I_i
- **Parameters:**

Feature Vector: $x \in \mathbb{R}^n$
Feature Sparsity: S_i
Feature Importance: I_i

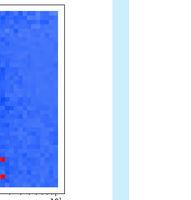
$$\text{Feature: } x_i = \begin{cases} 0 & \text{with probability } S_i, \\ \text{Uniform}(0,1) & \text{otherwise.} \end{cases}$$

$$\text{Model: } x' = \text{ReLU}\left(W^T W x + b\right), W \in \mathbb{R}^{n \times m}$$

$$\text{Loss: } L = \sum_x \sum_i I_i \left(x_i - x'_i\right)^2$$

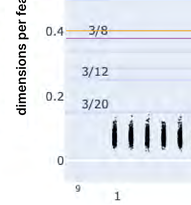
As sparsity increases, models use "Superposition" to represent more features than dimensions

Sparsity: 0%



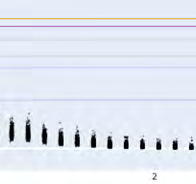
Dedicated orthogonal dimensions for the two most important features. Less important features are **not embedded.**

Sparsity: 80%



The four most important features are represented as **antipodal pairs**. The least important features are **not embedded.**

Sparsity: 90%



All five features are embedded **as a pentagon**, but there is now "positive interference".

Feature Importance
- Most
- Medium
- Least

Parameters
$n = 5$
$m = 2$
$I_i = 0.9^i$

Demonstrating Superposition

- **Conclusion 2: Both monosemantic and polysemantic neurons can form.**

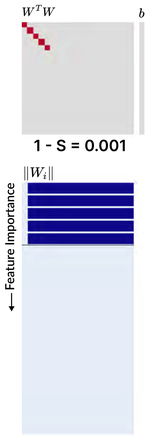
Weights / Bias Element Values

Superposition $\sum_j (\hat{x}_i \cdot x_j)^2$

Parameters
$n = 20$
$m = 5$
$I_i = 0.7^i$

Linear Model

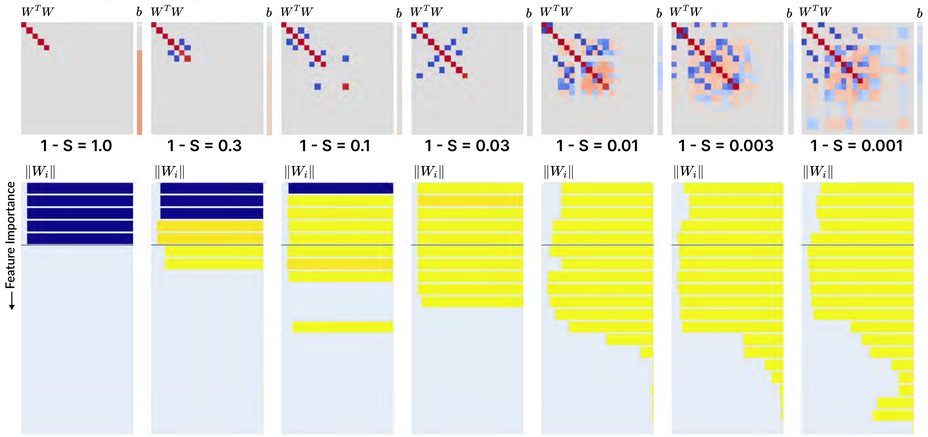
$x' = W^T W x + b$



Linear models learn the top *m* features.

ReLU Output Model

$x' = \text{ReLU}\left(W^T W x + b\right)$



In the **dense** regime, ReLU models also

As **sparsity increases,** superposition allows models to represent more features. The most important features are initially untouched. This early superposition is organized in **antipodal pairs**.

In the **high sparsity** regime, models put all features in superposition, and continue packing more. Note that at this point we begin to see ***positive interference*** and ***negative biases***.

Superposition as a Phase Change

- **Conclusion 3: Whether features are stored in superposition is governed by a phase change.**
- We investigate into a simple 2-layer model with ***n*** input (x) and output features (y) and ***m=n-1*** hidden neurons is generated with different **importance *r*** and **sparsity *s***.

$$y = \text{ReLU}(W^t W x + b)$$

$$\text{loss} = (x_1 - y_1)^2 + \ldots + (x_{n-1} - y_{n-1})^2 + r(x_n - y_n)^2$$

Theoretical loss for **feature of interest**

Not represented	$r(\frac{s}{3} - \frac{s^2}{4})$
Superposition	$\frac{(1+r)s^2}{6}$
Dedicated dimension (Others in superposition)	$\frac{s}{3} - \frac{s^2}{4}$
Dedicated dimension (Others not represented)	$\frac{s^2}{3}$

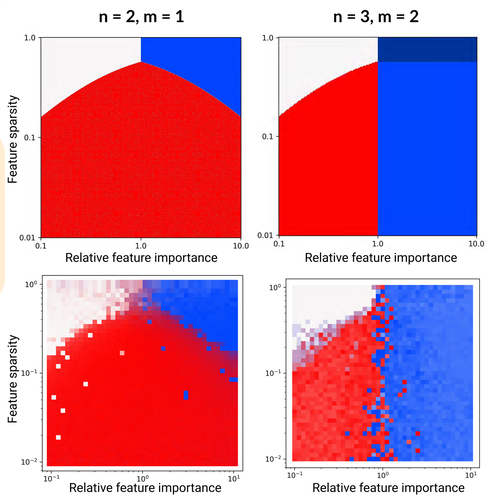
$W = [1 \quad 0]$
$W = [0 \quad 1]$
$W = [1 \quad -1]$

Each configuration is colored by the norm and superposition of the extra feature

Superposition of the extra feature $\sum_j (\hat{x}_i \cdot x_j)^2$ (0 to 1)

$\|W_i\|$ (0)

$W = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$
$W = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \end{bmatrix}$
$W = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
$W = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$



The Geometry of Uniform Superposition

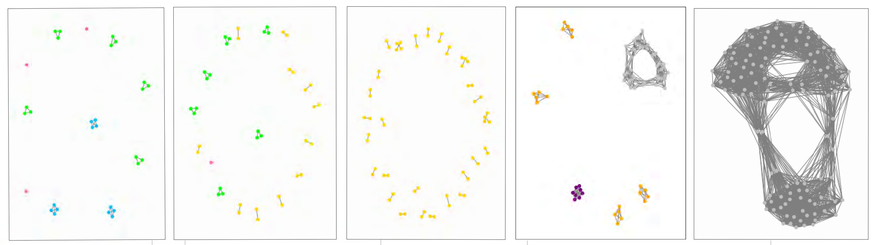
- **Conclusion 4: Superposition organizes features into geometric structures such as digons, triangles, pentagons, and tetrahedrons.**
- Uniform superposition is where all features are identical: **independent, *equally important* and *equally sparse*.**

Number of Hidden Dimensions per Embedded Feature

$$D^* = \frac{m}{||W||_F^2}$$

Feature Dimensionality

$$D_i = \frac{||W_i||^2}{\sum_j (\hat{W}_i \cdot W_j)^2}$$



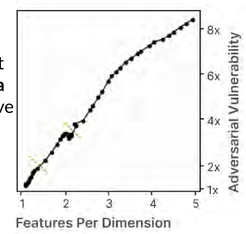
- The number of dimensions per feature (blue curve) is "**sticky**" at 1 and 1/2.
- The feature dimensionalities (black dots) cluster at certain fractions labeled with colored lines.
- These fractions correspond to certain geometric structures.
- Features seem to go through different phases *within* superposition.

Relationship to Adversarial Robustness*

Non-Superposition: $(W^T W)_0 = (1, 0, 0, 0, \ldots)$

With-Superposition: $(W^T W)_0 = (1, \epsilon, -\epsilon, \epsilon, \ldots)$

Idea: attack the most important feature **via interference,** effective even in the **infinite data limit.**



Conclusions

1. **Superposition is a real, observed phenomenon.**
2. **Both monosemantic and polysemantic neurons can form.**
3. Whether features are stored in superposition is governed by a **phase change.**
4. **Superposition organizes features into geometric structures** such as digons, triangles, pentagons, and tetrahedrons.

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