# **Related**

* [What are polysemantic neurons?](https://docs.google.com/document/d/1Xn_fyjjbUZAb7hghg0H-HbpmRAom4Wp_M9ALb9-A9fs/edit)
* [How much can we learn about AI with interpretability tools?](https://docs.google.com/document/d/14e9YGIel-4pah_vgFGmVgwXm-RNqkAcQZkFa8I4LeVo/edit)
* [What is interpretability and what approaches are there?](https://docs.google.com/document/d/1hHAx92e89YQfBXT96C7BLiuMipcl6LOszGlG2L7rrZo/edit)
* [What are circuits in interpretability?](https://docs.google.com/document/d/1h8CYsKzh769Hp7ikniUieU5bo86I3_nO-5qdXI0qqws/edit)

(possibly):

* [What is mechanistic interpretability?](https://docs.google.com/document/d/1an0GeIelvAIdvFK9UpNDHCn_WGtiI2DKnQSaHfR8EeU/edit)
* [What are latent spaces?](https://docs.google.com/document/d/1WtGffKsnT9XEHfPEGUwArmSt8lRNaDooBtqqIlHiicQ/edit)

# **Sources**

* <https://distill.pub/2020/circuits/zoom-in/>
* <https://distill.pub/2017/feature-visualization/>
* <https://www.transformer-circuits.pub/2022/mech-interp-essay/index.html>
* toy models
* https://transformer-circuits.pub/2023/superposition-composition/index.html
* Superposition, Memorization, and Double Descent [save for later]
* softmax linear units [save for later]
* Neurons in a haystack
* https://arena-ch1-transformers.streamlit.app/[1.7]\_Toy\_Models\_of\_Superposition

[place the below text above Related section later]

NOTE: The following sections below until “Scratchpad” are the writings I made during the hackathon. Each numbered section may be considered a new article.

1. What is superposition?

[Superposition](https://transformer-circuits.pub/2022/toy_model/index.html) refers to the phenomenon where neural networks “represent more features (learned concepts) than they have dimensions or neurons”.[[1]](#footnote-2) A model exhibiting superposition cannot represent a feature with just one neuron, and thus features must be represented by sharing neurons through approaches such as “distributed representations”, as shown in Figure 1. In this figure, a model is trying to tell apart three shapes based on how they activate its neurons. It cannot pair each of the three shape features with its own neuron “N”, as there are only two neurons. So to assign each feature with a uniquely identifiable label using neurons, the model represents them each as a “[distributed representation](https://arxiv.org/abs/2305.01610)”, or a weighted sum, of two neurons. In this manner, the model is able to distinguish the three features from one another based on the activations of only two neurons. A circle would have a high N1, and a low N2, while a triangle would have a low N1 and a high N2, and a square would never occur when N1 activates. Even though each feature cannot be provided with its own neuron, it is able to share neurons with others, which is an efficient representation.

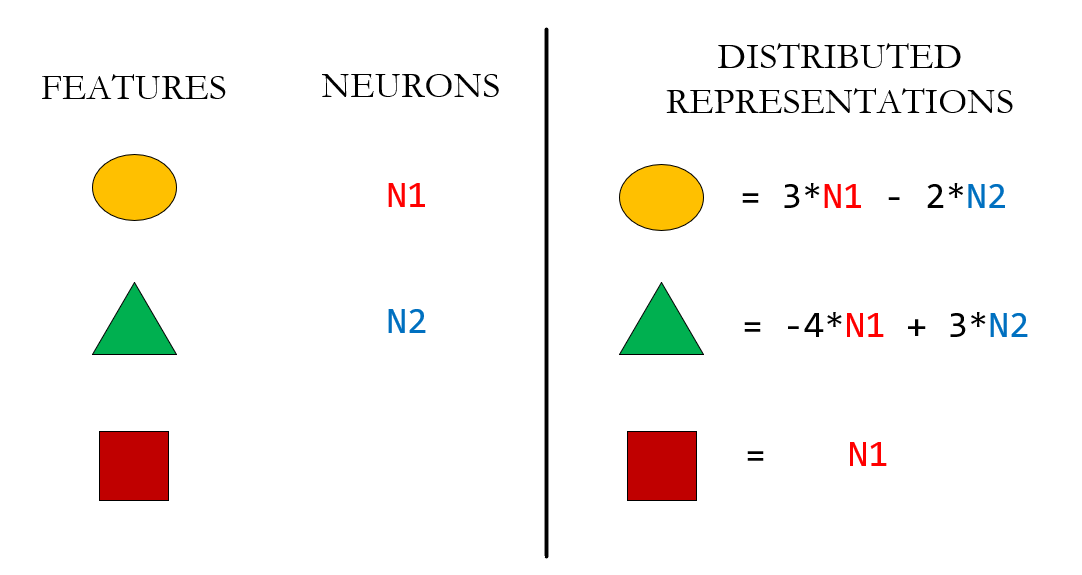


Figure 1: A case of superposition with distributed representations

In this configuration, a circle and a square would both have a “high N1”, meaning they are associated in this dimension. This representation was decided to be beneficial based on its performance for a given dataset. However, what does this dimension even “mean” when extrapolating it to tasks beyond the objective of correctly labeling that dataset? Subsequently, what does this “association” of the circle and square mean along their shared dimensions? When the AI must make decisions that take into account human ethical concerns, is this “association” of the circle and square aligned with human beliefs, or was it merely a good “trick” for the limited dataset it was trained on? AI interpretability research that studies superposition aims to answer these questions by identifying the manner of superposition a model is in, based on the features it must represent. This is done by analyzing the distributed representations that features are in with one another for a model.

What are the dangers of superposition from the view of AI Existential Safety?

We often hear that someone has an “irrational fear”. This could be someone who avoids eating anything said to “contain a chemical”, as their reasoning is that chemicals like bleach are harmful to consume. Though this is a decent guess on what is “bad to eat”, it is wrong to use this reasoning to claim that “all chemicals are bad because bleach is bad”, since consumable products, such as drinks that contain H2O, contain chemicals. Many human beings already have strange beliefs and associations that lead to decisions which are not rationally beneficial for their survival. As society develops superintelligent AI, it is important to understand what features the AI decides are correlated or associated with one another in order to understand its reasonings and beliefs, and thus ensure we are developing the AI in the right directions.

On the surface, we might see an AI making decisions to “eradicate a bad compound X, which is of type Y”, and we might overly trust it. But this AI may be deceiving us, as its internal reasoning is actually that “all these chemicals of type Y are bad”, even though not all of them are bad, and some are even required for our survival. It is thus crucial to catch this reasoning mistake before the AI makes decisions to eradicate all compounds of type Y (this scenario may also be regarded as a form of overfitting). Moreso, it is even more difficult to identify that the AI is making predictions which align with morals and ethics that are beneficial to humans, as these may be harder to define and correctly convey to an AI. When a model displays superposition, ensuring that representing hard-to-define, general concepts such as “good” are not just based on “association Y”, but are based on more nuanced forms of reasoning, would mean that the model must re-arrange how it represents its features so that its distributed representations better align with values beneficial to humans. Solutions to problems posed by superposition include creating models without superposition, and better interpretability of features organized due to superposition, which would reduce the unknowns of how AI piece features together into larger representations.

1. How are features represented as directions in latent space?

Though there are multiple ways to define a feature, to describe superposition, a good definition to use is to view features as directions in a “latent space”, which is a coordinate system that represents each dimension with a neuron in a layer, measuring its activation strength. Each layer has its own latent space, so the input data is transformed to a new latent space in each layer. Since each dimension is where a basis vector lies, that means each vector is described as a linear combination, or a weighted sum, of basis vectors. Some of these vectors may correspond to a feature. Figure 2 illustrates a simple example of this using height and tail as basis vectors to describe a cat feature and a bear feature, measuring how much an item is “like a cat” or “like a bear” based on how long its tail and height are. An item that is shorter with a longer tail compared to other items seen in training data is “more like a cat” than “like a bear”.

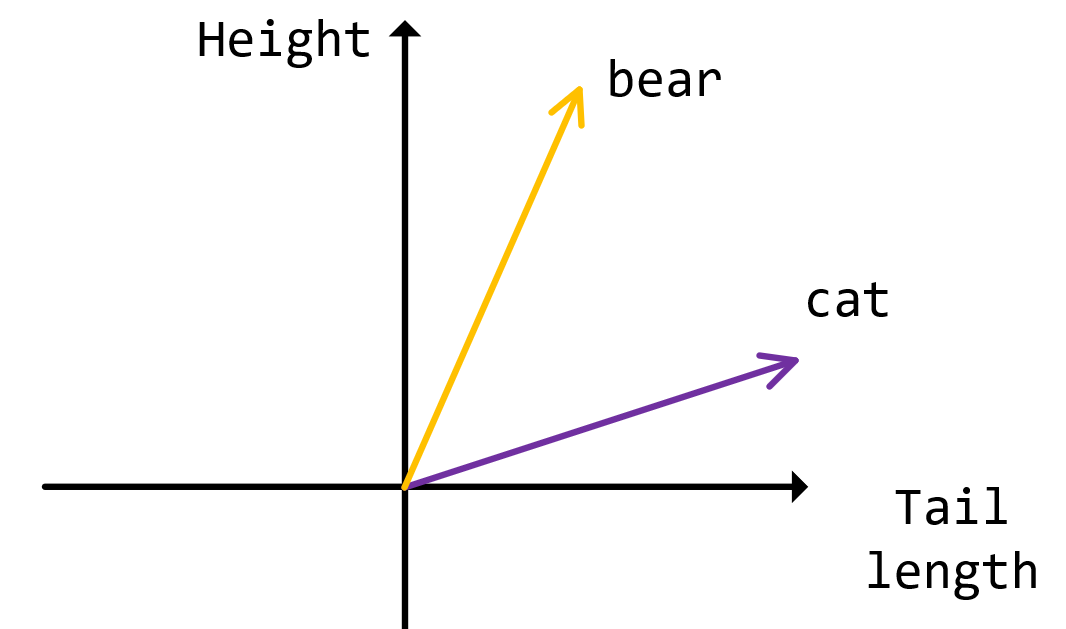


Figure 2: Features as directions in latent space

Note that features are not explicitly defined during training; rather, they are patterns in data which are used to achieve predictions that minimize a loss function. One of the goals of AI interpretability is to discover if there are any internal activation patterns which can be described as human-understandable features, such as “like a cat”. Multiple interpretability studies have discovered evidence for this so far for a variety of models, ranging from [image recognition models](https://distill.pub/2020/circuits/zoom-in/) to [large language models](https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html).

Moreover, some activation patterns do not necessarily correspond to features as human-understandable “traits”. For instance, studies have provided evidence for the existence of circuits, in which groups of neurons act as modules that each perform a task, which then piece together into larger groups of neurons to perform compound tasks in an algorithmic fashion. One such discovered circuit type is the [Indirect Object Identification circuit](https://arxiv.org/abs/2211.00593) in GPT-2 Small, which involves inputs such as “John and Mary went to the store, John gave a book to” that should predict “Mary” as the next subject token. This circuit type first uses a module to identify subjects in an input token sequence, and then uses another module to identify duplicate subjects, and finally inhibits the duplicate subjects so that the model’s prediction value for the non-duplicate subject, which would be “Mary”, would be higher.

In this manner, circuit behavior can be described as analogous to the parable of the “[Blind Men and an Elephant](https://americanliterature.com/author/james-baldwin/short-story/the-blind-men-and-the-elephant)”, where several blind men each touch an elephant but don’t know what it is as a whole. One touches a tail and thinks it’s a snake, while another touches a leg and thinks it’s a tree trunk. But if they each tell their own observations to a blind judge, then the judge can use all this information to deduce that this object is actually an elephant. In this analogy, each blind man is like a module, and their message passing to a judge would constitute a circuit. Therefore, activation patterns may not only serve to identify “traits”, but to pass information along to computational units that process these traits.

1. What are the differences between superposition, polysemanticity, and distributed representations?

Now that definitions for “feature” and “circuit” have been established (link to section 2, previous article), other terms that build upon them can be introduced. The following three words are commonly confused terms that are often used interchangeably with one another:

* Superposition is a phenomenon for a model when there are more features than neurons
* Polysemanticity is a property of a neuron when one neuron corresponds to multiple features
* Distributed Representation is a property of a feature where the feature is represented as a linear combination of neurons

Superposition implies the existence of Polysemantic Neurons (and often, Distributed Representations). However, the converse is not true.

A sparse feature and a sparse distributed representation may also be confused with one another:

* A sparse feature is not “frequently encountered” in a data distribution compared to other features
* A sparse distributed representation uses a few neurons in its summation compared to the total number of neurons in a layer

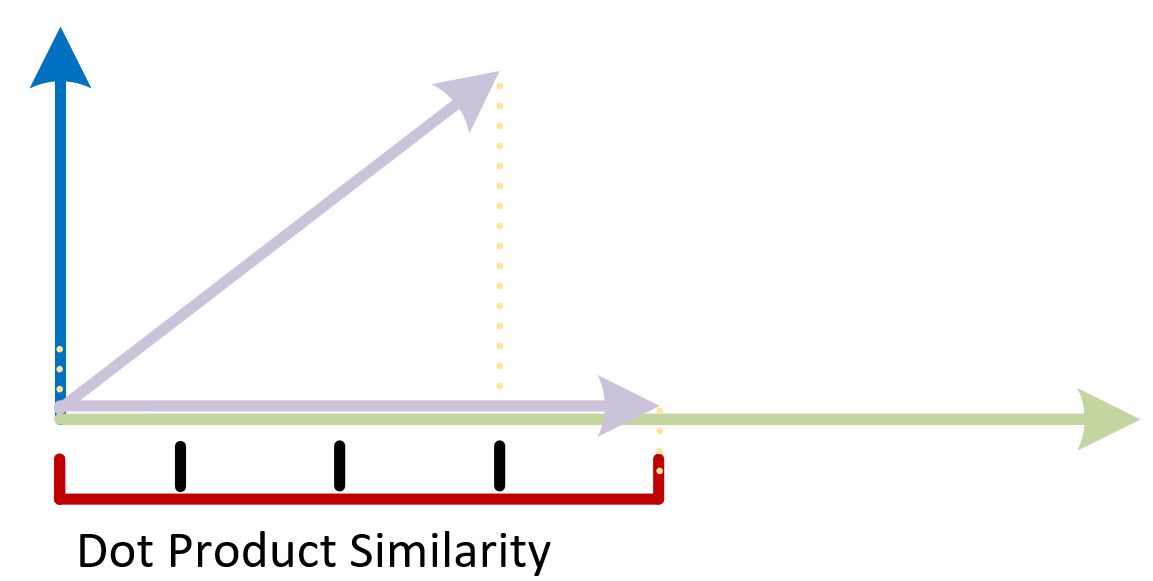
1. How do we experimentally study superposition?

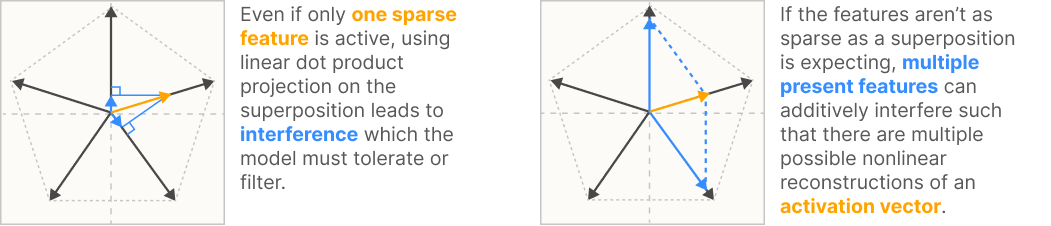
[In previous work](https://transformer-circuits.pub/2022/toy_model/index.html), was observed that the “degree” of superposition a model has, and thus the way a model represents features, is affected primarily by two properties:

1. Importance: how important is a feature for correctly performing a certain task?
2. Sparsity: how frequent is that feature encountered by the model?

Therefore, superposition can be studied by comparing how models represent features differently as feature importance and sparsity are varied.

Another topic of interest to study is the relation of correlated features with interference. When superposition occurs, the distributed representations that features are organized in are affected by feature correlation and interference. The dot product can be thought of as a similarity measure between two vectors, as seen in the figure below. Vectors that are closer in direction are more similar, and orthogonal vectors, at 90 degrees with one another, have 0 similarity. Consequently, the concept of interference measures the dot product between two vectors- vectors that have a higher dot product have higher interference. The cosine similarity measure is a normalization of this dot product similarity.





Correlated features were found (by the study in the link above) to be represented in orthogonal directions. One hypothesis for why this occurs is because when two features often co-occur in an input, the model requires a way to tell them apart. If fur and tail, which are often used together to describe animals, were activated in the same way by the same neurons, then when an input contains both fur and tail, a model would have trouble telling apart which is fur and which is tail. Thus, it aims to represent them using features which have little similarity with one another, at least in some directions.[[2]](#footnote-3)

In contrast, anti-correlated features are not correlated with one another. Most features are anticorrelated; for instance, the features “calculus” and “furries” are often not used in the same context. Thus, the model may decide for them to share the same neuron activation patterns, as it often doesn’t need to distinguish which is which to make predictions, since they are often not used together to make a decision involving both calculus and furries; additionally, the model can tell them apart by other features in their surrounding context. Anti-correlated features were often found to be represented in opposite- or antipodal- directions. Other, more complex geometric arrangements of features were discovered by varying properties such as sparsity, correlation, number of dimensions, and importance.

# Scratchpad

[avoid going into more as will make the article too long and technical (save for other posts)]

How do we measure superposition?

W^T W: measure correlations

[figure showing in dot in]

To do:

• Explain how to measure superposition using W^T W with Basic Results diagram: measure the sum of a row of W^T W to find the direction of a feature in relation to other features. This is 0 (black, orthogonal to all) if no interference, and 1 (yellow). The more interference, the more superposition?

◦ W^T W is a central concept that many measurements use

• Explain, in simple terms, features by features matrix W^T W (correlations; show input dot input matrix diagram?)

• L ~= benefit + interference : (trade-off)

• Froebenius norm measures number of features a model has learned

• Dimensionality D equation(also uses W dot W)

• Antipodal to more geometric structures

• Linear vs nonlinear case

• For now, explain up to and including sec 4 (geometry)

• list of open questions

# [the work below was done by the previous author, not me]

# **Scratchpad**

**[redacted due to not knowing if author gives permission to upload]**

——————————————————— EOF ——————————————————

1. Features and dimensions will be defined in more detail in the Section (or Article) “How are features represented as directions in latent space?” [↑](#footnote-ref-2)
2. It might be confusing how correlated features can be orthogonal while still being conceptually similar. It can be speculated that many correlated concepts would be similar in some directions- that is, they have high cosine similarity with “like a cat” features- but there would be other directions in which they are not in order to tell them apart. However, this is pure speculation with no evidence so far. [↑](#footnote-ref-3)