https://transformer-circuits.pub/2022/toy_model/index.html

Introduction

Ideally (from an explainability point of view) models would allocate one neuron per input feature. IRL this doesn't always happen.

Why does this happen in some models and not others?

This paper uses small ReLU models to explore how and when models can represent features in superposition - i.e more features than dimensions.

- Models can represent more features than dimensions as non-orthogonal embeddings with a some amount of "interference" between the features
- In certain cases models can perform computations while in superposition, the authors say this is like a smaller network noisily simulating a larger highly sparse network
- Neurons can be "polysemantic" i.e represent multiple features, or "monosemantic"
  - This is governed by a phase change
- Superposition shows geometric structure.

Features, Directions, and Superposition

The authors make the following claims:

- Decomposability: Network representations can be described in terms of independently understandable features.
- Linearity: Features are represented by direction.

They claim there are two oppositional forces that can explain why models sometimes have features correspond to neurons or not:

- Privileged Basis(?): Only some representations have a privileged basis which encourages features to align with basis directions (i.e. to correspond to neurons).
- Superposition: Linear representations can represent more features than dimensions, using a strategy we call superposition. This can be seen as neural networks simulating larger networks. This pushes features away from corresponding to neurons.

"...We tend to think of neural network representations as being composed of features which are represented as directions. We'll unpack this idea in the following

sections."

What are Features

Discussion of what actually constitutes a feature:

- Features as functions of the input:
  - "Doesn't quite fit", the features are "fundamental" abstractions for reasoning about data, with the same features occuring across models
- Interpretable properties
  - "human understandable concept", but should also be able to refer to a newly discovered "feature" (by the model) that we may not understand
- Neurons in sufficiently large models
  - Ideally in large enough network it would devote a single neuron to the property we are calling a feature.

Features as Directions

- we tend to think of neural network representations as being composed of features which are represented as directions. We'll unpack this idea in the following sections.
  - But isn't there an activation function between them?

They claim linear representations should be favoured for three reasons:

- Linear representations are the natural outputs of obvious algorithms a layer might implement. If one sets up a neuron to pattern match a particular weight template, it will fire more as a stimulus matches the template better and less as it matches it less well.
- Linear representations make features "linearly accessible." A typical neural network layer is a linear function followed by a non-linearity. If a feature in the previous layer is represented linearly, a neuron in the next layer can "select it" and have it consistently excite or inhibit that neuron. If a feature were represented non-linearly, the model would not be able to do this in a single step.
- Statistical Efficiency. Representing features as different directions may allow non-local generalization in models with linear transformations (such as the weights of neural nets), increasing their statistical efficiency relative to models which can only locally generalize.
  - Here non-local generalisation means being able to generaliset to data "far" from that seen during training.

Privileged vs Non-privileged Bases

- Word embeddings have no priviledged basis, you can transform the embedding and inversely transform the weights of the model and you get the same model with different basis dimensions.
  - In such a case we have to identify "interesting directions", like the one between man and woman.
- If a coordinate-wise, non-linear activation function is used, this "breaks the symmetry" and makes the basis of the activations the one that features are incentivised to align with.

The Superposition Hypothesis

There is some vaguely mathematical arguments made that NNs should be able to represent more features than they have neurons.

- Superposition vs Non-Superposition: A linear representation exhibits superposition if W^TW is not invertible. If W^TW is invertible, it does not exhibit superposition. (Fancy way of saying if W has orthogonal directions it is not in superposition)

Demonstrating Superposition

The authors set up small two layer models (Linear and ReLU) to demonstrate this superposition.

To go with this model they create some synthetic data with the following properties:

- Sparse features
- More features than neurons
- Features that vary in importance

They enforce sparsity controlled by a parameter S and each feature (dimension) has an importance I_i.

They project the features down to a latent space, then try to recover the feature vectors by applying the transposed projected (with a bias). They claim the projection matrix should be "close" to orthonormal and the bias allows the model to set features it doesn't represent to "their expected value". It also alwos the model to discard small amounts of noise, important for superposition.

Without an activation function, superposition doesn't occur.

Loss is defined as:

$$L = \sum_x \sum_i I_i (x_i - x'_i)^2$$

In their experiments, as sparsity increases, the ReLU model starts representing more features in superposition, initially placing the features in antipodal pairs, and eventually representing them with interference.

Mathematical understanding

- In the linear case the model essentially performs PCA, so there can be no superposition
- Setting this understanding aside, in the linear case (as shown by Saxe et. al. https://arxiv.org/abs/1312.6120) NN weights can be "can be thought of as optimizing a simple closed-form solution". The loss function above can be rewritten

$$L \sim \sum_i I_i(1 - ||W_i||^2)^2 + \sum_{i \neq j} I_j (W_j \cdot W_i)^2$$

Revealing a feature benefit part (first term) and an interference part (second term).

- This shows there is a tradeoff between representing a given feature and potential interference with other features from representing that feature.
- There is a similar result in the ReLU case, but due to the "zeroing out" aspect of ReLU, features which cause "negative" interference are free.
- The authors note the interference term is similar to a Thomson problem (i.e packing points on an n-sphere according to some energy function), which goes some way to explaining the geometrical results later in the paper

Superpositions as a Phase Change

Phase Change:

Discontinuous change in a system as some variable is changed.

- By using a "toy model of the toy model" which can be solved analytically they compare the toy model as they change the relative feature importance and feature density $(1 - S)$
- This reveals phase diagrams with discrete regions referring to situations where its advantageous to either not represent, represent with a dedicated dimension, or represent in superposition.
- Why don't they consider situations where all three features are represented in superposition in the last example?