Searching for Model Concepts by their Conditional Independencies

David Reber

Abstract

Your abstract here

So far Victor Veitch and Kaarel Hänni have helped brainstorm and soundboard! ChatGPT probably helped a lot with formatting, rephrasing, and catching spelling mistakes. All remaining mistakes are my own.

1 Rough Sketch of the Search Process

Assume we have three binary concepts X, Y, and Z which satisfy the conditional independence of $(X \perp\!\!\!\perp Y|Z)$, and we want to find them in a model. We can do this by searching for a set of three positions in the model which satisfy the conditional independence.

1.1 What are we searching over?

Let's assume concpets are represented linearly as subspaces of the embedding space, and the i^{th} layer is simply reasoning over these concepts (but not changing the embedding space itself). If this were true, our search is a search over subspaces in the embedding space, and we would need a way to determine which subspaces correspond to which concepts.

(An alternative possibility is there is a different embedding space at each layer; that is, that each layer is a transformation from the previous layer's embedding space to a new embedding space. If this is the case, the subspace corresponding to a concept in one layer may not correspond to the same subspace in the next layer. Additionally, if the transformations are not full-rank then we would expect to see concept collapse, where some concepts are not represented in later embedding spaces at all.)

Here, *position* refers to We can do this by training a probe to predict X, Y, and Z from the positions in the model, and then searching for a set of three positions which satisfy the conditional independence.

2 April 25 Victor summary

- Assume you have infinite compute and infinite data
- It seems like even then you may not be able to do this search easily, because there are infinite rotations you would need to check (there's no privileged basis).
- The counterpoint is that LLMs seem to be able to also have an understanding of which concepts are conditionally independent: if you assume that knowledge is also represented somewhere in the model, does that buy you the extra constraints you need to narrow down the search space to something that ever terminates?

3 Rough Notes from April 25 discussion with Kaarel

- Transformer model
- Causal graph has 3 nodes (actually a CI)
- Try all possible variations of 3 places in the residual stream to look at
  - Both token position and layer: all possible combinations of places
- Look for a feature in each position, testing whether each feature has the independence you want
- Should be possible to do using a differentiable loss function
  - If binary concepts are sigmoids of affine transformations of features
  - Then loss depends on whether the CI holds in that batch
  - KL divergence between product distribution and joint distribution you observe
  - Need an empirical variant of the mutual information.
- Gradient descent is on how to choose 3 positions in the model.
- If there are any other concepts which have the same causal structure, this won't be able to differentiate between them
  - So we could add a supervised term which matches prompt = "man"
- If the 3 variables have deterministic value from a given input, then we could train a supervised probe to match the labels I know to the positions in the model.