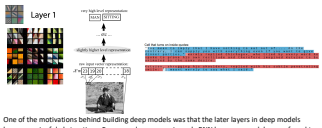
Week 7 - Concept-Based Explanations

Wednesday, March 5, 2025 9:50 PM

Activation Analysis

1. Deep learning models are made from simple layers. When an input is passed through the network, each layer transforms the output of the previous layer. The intermediate outputs at a given layer are often called that layer's "activations." Each activation is association with one neuron.

2. There is a long, if informal, practice of relating model activations to properties of the data. For example, one of the earliest neural network explainability papers found that many neurons early on in the network activate when presented edges with specific orientations.

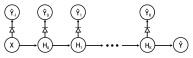


3. One of the motivations behind building deep models was that the later layers in deep models learn meaningful abstractions. For example, neurons in early RNN language models were found to activate only on text that appear within a block of quotes. These explanations are arguably at a higher, and more useful, level of abstraction than those that refer to individual inputs or samples.

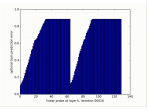
4. The recent literature on concept-based explanations formalizes these early explorations. The appeal of the approach is that it helps bridge the vocabulary we already use to reason about a topic with the computational representations living within networks -- it is an exercise in creating a shared language. If we let ourselves indulge in a little scifi, if we expect that one day we'll have machines capable of solving extremely complex problems, then the hope is that the methods descending from this literature will let us discover the underlying problem solving techniques.

Linear Probes

1. A "linear probe" is just a linear model that predicts a feature of interest from intermediate model activations. If the model has high accuracy, then we can say this feature has been learned by the model in that layer.



2. This can be useful for scrutinizing claims about *how* models achieve their performance. In the toy example in their paper, they show that a skip connection across layers 1 - 64 prevents the model from learning anything useful within those layers. The implication is that while skip connections may be popular in many types of architectures, they might also might lead to a lot of wasted computation and inadvertent shallowness if not attended to carefully (Click the image below to see how probe prediction accuracy only improves for the deep layers).



Network Dissection

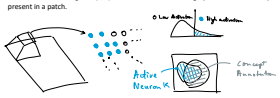
1. The network dissection paper advances over linear probes in two ways: (i) The authors realize that richly annotated datasets are a great source for probing what a model knows and (ii) they attempt to understand models at the level of individual neurons, not just layers.

2. For (i), the authors introduce the Broden ("**Bro**ad and **Den**sely Labeled") dataset. For each image in this dataset, each pixel is annotated with many complementary labels, some at several levels of resolution. There are 5 types of labels, which are easier to understand through an example.

They refer to the possible label values as concepts. E.g., "cottage", "sky", ... "lined" are concepts.

3. The authors want to see whether a given neuron k is related to a concept c. To this end:

   - For every input image x, compute the activations for the current neuron across all spatial patches of the image.
   - Make a histogram of those activations across all inputs and patches. For each input, create a binary mask that selects patches whose activations are above a given quantile.
   - Compute the intersection-over-union (IoU) of these binary masks with the concept's annotation mask. Large IoU(k, c) means that neuron k is highly active when concept c is present in a patch.



4. The overall process makes it possible to create images like the one below. E.g., it seems that neuron 1410 in the resSc model always activates when it sees a specific type of conical house.

5. A skeptical reader would argue that these associations could be purely due to chance. We have searched so many neuron-concept pairs that some are bound to seem associated. If this is true, then the picture above would be overinterpreting something that has no inherent meaning (only large subsets of neurons, not individual neurons, should be interpreted). To this end, they run an experiment where they apply a random rotation to all the activation vectors and rerun the dissection algorithm. They find that the IoUs in this case are much lower, which suggests that isolated neurons really may be associated with these concepts.

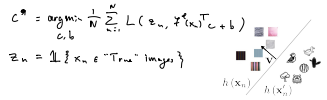
6. One of the wildest applications of this idea is to support interactive image generation. Network dissection can isolate concept-related neurons in generative image models, and by cleverly controlling the activations of those neurons, the user can interactively paint new objects in an existing generative image output. Click the link to see the demo.

Concept Activation Vectors

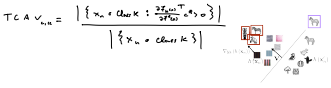
1. Concept Activation Vectors have a similar motivation as Network Dissection -- using human-curated annotation to interpret model components -- but focus on interpreting layers. Their approach applies more generally (not just images) and requires less dense annotation. To see whether layer l uses a concept to improve model performance, we follow four steps.

2. *Step 1*: Define a collection of images x[n] that represents a concept. Also construct a control pool of random images x'[n] as a reference point. In this example, the concept is "stripes."

3. Step 2: Extract layer l's activations f[l](x) for both the positive and control groups. Train a linear classifier to distinguish the groups. This is like linear probes, except the task is derived from Step 1. (My figures need to be updated with v -> c and h -> f[l]).



4. *Step 3*: Use the estimated concepts c to compute a concept score, which can be interpreted as "the fraction of training samples [in class k] whose model classifier scores increase when the input is infinitesimally moved in the direction of the concept."



5. *Step 4*: Test the significance of T[k] by estimating concepts trained to distinguish randomly chosen example images. The scores derived from these concepts define a null distribution that can be used to compute a p-value for the observed T[k]. Though honestly, I'm not sure what hypothesis this is testing.

Developments

1. We can think about multiple concepts simultaneously using Interpretable Concept Basis. Instead of considering inner products with concept directions, we just learn a regression onto them. More precisely, first run a regression of the layer l activations onto the pre-classification logits.

2. Then run a second regression to see how those coefficients can be expanded in terms of the known concepts. The new weights are constrained to be nonnegative, so we don't have to worry about weights cancelling one another out. A large coefficient in this regression means that the associated concept is important for this layer's contribution to the class k prediction.

3. How can we know whether we've found all the relevant concepts for a layer? Yeh et al (2020) relate concepts to the statistical principle of sufficiency. They define a summary statistic based on the concepts alone. If the model performance after this reduction is comparable to the original model, then the collected set of concepts are considered to be sufficient.

4. Formally, for a collection of concepts c[1], ..., c[m], compute the summaries T(x).

5. Let h denote the part of the model that maps layer l's activations to the final classification logits. Find a function g that optimizes:

This is the best possible model using the concept-based summary. The authors propose a particular ratio between this performance and the original model's, and this can be interpreted as the extent to which the current m concepts are sufficient.