Week 7 Paper Summaries

Joseph Camacho

August 8, 2022

The Mythos of Model Interpretability

Interpretability seems to be a loose conglomeration of ideals. Fundamentally, humans want to understand why models give the decisions they do, and that is what they are looking for when they say a model is "interpretable". Some kinds of interpretability are *transparancy*, which focuses on understanding the components of a model (i.e. making it less of a black box), and *post hoc interpretations*, which focus on explaining a decision after it has been made (but not revealing the inner working of the model--this is similar to how brains are black boxes but we can still get explanations from humans).

There are many goals of interpretability. One of the biggest is being able to trust models. If we can understand them, we are better able to know when they might fail. Being interpretable can also help us see where causal relationships might make models fail, and where they would be able to transfer well to other datasets. Interpretability is further useful in providing information beyond a simple number.

What does an interpretable model look like?

- Simulatable -- A human could reasonably run through the calculations needed. Basically only sparse linear models and short decision trees are interpretable under this criterion.
- Decomposable -- The model can be broken down into intuitive/understandable parts.
- Algorithmic transparency -- The model's algorithm can be proven to have certain properties such as convergence or nice decision boundaries.
  What are some post-hoc methods?
- Text explanation -- The model gives a written explanation of why it made its choices.
- Visualization -- Use something like t-SNE to visualize high dimensional embeddings. You can also look at what inputs cause certain nodes in a neural network to be highly activated.
- Local explanations -- E.g., saliency maps
- Explain by example -- Give an example of what the model thinks is the closest image to the given one, along with a class prediction.

Linear models are not always more interpretable than deep neural networks. Although the algorithm is simpler, they don't have the same kind of clear features that the nodes of neural networks have. Also, transparency might not be what we need; humanity could be better off with less transparent models that are much more accurate or robust. Finally, be wary of post-hoc interpretations. They often look really cool, but aren't actually as meaningful as they look.

Judge

The authors seem to be trying to consolidate the various interpretations of "interpretability", but don't ever give a single definition of interpretability that others could use. Instead, they just list a bunch of ways interpretability has been used in the past. I think this paper would have been more useful to outside researchers if they gave more specific guidelines on how the word "interpretability" should be used.

ViM: Out-Of-Distribution with Virtual-logit Matching

Out-of-distribution (OOD) detection is important for a variety of reasons (e.g. detecting when the model is being deployed in situations it shouldn't be). This paper describes a method to do OOD detection by creating a virtual logit that represents how out of distribution the input is.

They review other methods and point out how they are lacking (e.g. methods using logits/probabilities, features that depend on the nullspace of the weight matrix are not considered).

They next build up to what their Virtual-logit Matching is by reviewing some of the math regarding OOD scores based on the null space and principal space (where "principal space" is defined as a small subspace spanned by the largest eigenvectors of $X^T X$).

Here's how the Virtual-logit Matching works. First, construct the "principal subspace P" by taking the space spanned by the largest D eigenvectors of $X^T X$, (where X are the values taken right before the logits, and D is some constant). Then, create a virtual logit for the OOD by taking the norm of the component of x orthogonal to P, and scaling this norm so that it is approximately equal to the maximum of the original logits. The probability of being OOD is calculated using softmax, just like the probabilities for the other classes.

ViM is similar to adding energy and some residual information.

The authors also created their own dataset, called OpenImage-O, to benchmark various OOD methods. Their results significantly outperformed almost all other methods (using the AUROC metric) across a large number of OOD benchmarks, including OpenImage-O, Texture, iNaturalist, and ImageNet-O. Mahalanobis, though, had similar performance, although it is much more computationally costly.

Hyperparameters can also be tuned. ViM is fairly robust to changes in dimension (D), but figuring out a good value of α can be hard to do.

Judge

The paper fails to evaluate how ViM performs against adversarial images. (I.e., images generated specifically to fool it.) I suspect that ViM would be rather weak against adversarial attacks.

Exemplary Natural Images Explain CNN Activations Better than State-of-the-Art Feature Visualization

Subjects were given groups of images that highly activate and minimally activate certain neurons in a neural network. They were asked to choose which one of two more presented images would be most likely to highly activate the neuron, and their confidence for it. Response time was also tracked. Overall, natural images vastly outperformed (>2 standard deviations) images generated via feature visualization in accuracy, and also outperformed in confidence and reaction time. In fact, natural images outperformed even hand-picked feature visualizations from feature visualization experts.

More sample images led to higher accuracy in both synthetic and natural images.

Judgement

The whole "confidence level" part of their study is too subjective, and doesn't really add much because (1) considering that both test groups had, on average, more than 75% accuracy, a well calibrated human should be hitting "3" (most confident) almost every time, (2) most subjects are poorly calibrated, so this doesn't actually give much meaningful data, and (3) what does this actually tell us that the accuracies don't? That humans feel more confident, regardless of whether they actually are more accurate? That seems like a rather *dangerous* thing to have in an interpretability tool, not something to be praised.