Papers

Papers

By James Camacho

The Mythos of Model Interpretability

The paper criticizes how many researchers use the word ”interpretability” in a quasi-scientific way. It says that there are many papers and conferences where researchers work on interpretability without really defining it. Either it’s universally understood what it means to be interpretable, or it’s universally not understood. They show it’s the latter by how diverse the different metrics for diversity are.

They proceed with why interpretability is useful. In many fields (e.g. medical, financial, criminal), you want the model to pick up on casual relationships. Poor people are more likely to have heart disease, but that’s entirely correlation—more money in your wallet doesn’t change how your heart works. If you can’t interpret a model, you don’t know if it’s actually picking up on casual relationships or merely correlation. Other reasons you would want interpretability is to protect against adversarial attacks, deeper understanding of why a model suggests what it does, and trust. If you can’t understand a model’s hidden layers, it may have hidden motivations, be vulnerable to attacks you don’t know about, or give you information that sounds sketch without an explanation.

The paper explains a few avenues researchers are taking to interpret models: salient maps (what inputs most affect the outputs for a given class), training examples most similar to the output, or even separate neural networks to interpret the ”black box.”

I found the paper to not be very useful, more suited for a blog post than a publication. The premise was interpretability is a finicky thing, so they should not have talked about various interpretability strategies. Instead, they should have proved their claim, giving examples of the wildly different definitions other papers use for interpretability, or polls of data scientists. Instead, they made that claim without justification, and proceeded to try to define interpretability just as poorly as every other paper out there. The paper brings up many good points such as what interpretability should encompass, but it isn’t very rigorous, and doesn't actually provide any solutions. Overall, I feel it didn't add anything new to the AI/ML community.

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

Detecting out-of-distribution images is important to prevent disaster from adversarial attacks, changing environments, and black swans. Many methods aim to do so without changing a pretrained network. They look at the logits, or the features, or in the case of this paper, both. OOD images on average would have more variation in the logits and more information would be lost from the features to the logits. ViM capitalizes on this to detect OOD images.

The logits are essentially a projection from feature space to the classes. As with every projection, some information is lost—which is why only using the logits doesn't give a great OOD detector. ViM calculates this lost information with some fancy linear algebra, then normalizes it to match the largest logit in magnitude. It appends these values to the logits, and softmaxes. The smaller the softmax, the more likely to be OOD.

I think it's great that this paper is using more information to get better OOD detection, but I think it's going about it in the wrong way. It performed decently well, but certainly not state-of-the-art, and that's because it's not doing anything really novel. Using the feature space or the logit space had already been done, just perhaps not together. I think the method though is not using the data in the right way, it's like a poorly defined loss function. Maybe OOD examples will usually have a smaller softmax, but there's no reason for them to always. What if an OOD example has nearly identical feature space as an in-distribution example, even though it's clearly not in the same distribution?

Instead of coming up with more complicated ”loss functions” it would be better to really understand what it means for something to be out-of-distribution, to come up with some method of recognizing something new. I think some sort of reinforcement learning method could work, where it makes a prediction and when it's extremely wrong it changes it's OOD confidence accordingly.

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

They did a psychological (? social?) study, aiming to show feature visualization with synthetic images works well, but ended up finding their baseline of natural examples worked better than the synthetic images by a large margin. To perform the test, they gave the participants two images and asked which most activated a particular node. The participants were given examples of either synthetic images or other natural images that most activated the node to aid in the decision. Giving more images helped for both synthetic and natural examples, and both synthetic and natural examples helped in the prediction (the humans were right more than 50% of the time), but the participants were consistently more accurate, confident, and swift in their judgements with the natural examples.

This paper is not very technical, but I think it does more to aid interpretability than many technical papers. They find a very good test for interpretability—can a human guess what will activate the neural network?—and try it out. It makes sense intuitively, and gets significant results (natural examples are better than synthetic images). The only issue is with using humans, as humans are much slower and more costly than a computer algorithm. The whole point of machine learning is so humans don't have to do the work. So, human studies are not very scalable, but they do provide a good baseline for what it means to be interpretable.