Paper summary, week 7

David McSharry

1 The Mythos of Model Interpretability

This paper attempts to better define what we mean we talk about interpratability. In the most general sense, what we want is a model that returns something besides simply a prediction, for example instead of simply diagnosing a disease we might want some details about why a model diagnosed that particular disease. It then introduces a few properties that we want to get from studying intepretability, for example ”trust”. The distinction is made between trusting a models accuracy, and trusting it won’t contain biases. Trust, like interpretability, is open to interpretation. The paper also mentions ”causality” as something we would like to get from studying intepretability. Deep networks can only really point out correlations by looking at their outputs. It also discusses some properties of interpretable networks, for example transparency. A transparent models processes can be understood, and they are therefore not just a ”black box”. For example if a model is simulatable then a human can run through the processes in their head and fully understand them (like a perceptron). Models could also be decomposed, so their component parts can be understood and so can how they fit together. Some other properties are also discussed, like visualization of how models make decisions, transparency of how a training algorithm works. The paper then then discusses some implications and key ideas to keep in mind when discussing intepretability, like that we should be careful to give up on the predictive power of black box models just because we don’t understand them even if they perform better than more intepretable methods.

One weakness of this paper is that when it motivates interpretability in a fairly narrow scope of AI applications, for example it basically maxes out at classifiers for use in medical settings and AI that can sentence convicts. These are two applications where interpretability would be nice, but most likely would not be disastrous in the worst case scenario without interpretability. It doesn’t discuss interpretability in the context of more powerful AI models.

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

This paper introduces ViM, a method for detecting OOD anomalies. Anomaly detection is important to avoid unexpected outcomes when deploying AI models, as coming across an OOD anomaly means the model has not been trained on any similar data. The purpose of ViM is to map an input to a score that is related to the likelihood that an input is OOD. If the score exceeds a threshold then the input is deemed OOD and the model can react accordingly. This method is based on the premise that when mapping from N neurons in a layer to C logits, information is lost, and that lost information is the key to detecting anomalies (assuming $N > C$). Consider the column space of the map from layer to logits (the feature space), since $N > C$ this cannot span all of $\mathcal{R}^N$, and so we can consider the orthogonal complement to this space. The projection of the layer onto the feature space becomes the logits and the projection of the layer onto the orthogonal complement (which we normally discard) is how we will determine an anomaly score. The norm of this projection onto complement space times a constant gives us the ViM. This results in better AUROC scores than other methods. (This is really cool)

A strength of this paper is that because it utilizes the information normally thrown away between layers of a neural network it offers insight into the inner workings of the network and therefore feels like a step toward intepretability. Perhaps further analysing information thrown away between layers in general can offer insight into an information theoretic approach to analysing the flow of an input through a neural network and what is thrown away vs kept. (This is quite speculative)

3 EXEMPLARY NATURAL IMAGES EXPLAIN CNN ACTIVATIONS BETTER THAN STATE-OF-THE-ART FEATURE VISUALIZATION

This paper explores a new way to visualize what features each layer in a neural network is maximally activated by. This helps us interpret what each layers job is in terms of building up higher order representations of different features. This was done by Olah et al by synthetically producing an image that maximally activated neurons to produce an abstract looking image that should help humans understand the feature being represented by those neurons. When shown to both untrained and trained humans (in terms of CNNs) these synthetic images helped them determine which of two natural images would most activate the neurons better than random chance. The innovation of this paper is that it shows that a better way to teach these representations to humans is to show them a natural image that highly activates the neurons, and then when asked which of two other natural images would maximally activate the functions the humans scored a higher prediction accuracy than both chance and than being shown synthetic maximally activating images. This is a step towards interpretability as it helps humans better visualize the feature being represented in neurons of a network.

A weakness of this paper is that it might lure humans into a false sense of security with neural networks. If the visualization of a feature representation is simply a dog, then an untrained human could be led to believe (subconsciously or consciously) that they understand the internal processes of the network and could lead to a false sense of intepretability, compared to if the feature rep visualization was a more abstract image.