Week 10 Tuesday Paper 1: The Mythos of Model Interpretability

1. What aspects of Interpretability Research are discussed in the paper? Explain them in one sentence each.
Ans: Diverse Motivations for Interpretability: The paper identifies various motivations behind the quest for model interpretability, including trust in models, understanding causal relationships, ensuring fairness and ethical decision making, and other practical considerations like model debugging and regulatory compliance.

Transparency and post-hoc interpretability was also discussed in detail.

Feasibility and Desirability of Interpretability: The discussion touches upon the practical aspects of achieving interpretability, questioning the commonly held beliefs about the interpretability of linear models versus deep neural networks and suggesting that the situation is more nuanced.

2. What is the distinction between Transparency and Post-hoc Interpretability?
Ans: Transparency refers to the inherent understandability of a model's mechanism, where the model's operations can be comprehensively understood by humans. This can include the simplicity of the model, understandability of its parameters, and clarity of its decision making process.

Post-hoc Interpretability, on the other hand, involves explanations or insights provided after a model makes predictions. These explanations aim to make the model's decisions understandable, even if the model itself is complex and its internal workings are opaque.

3. Are models with simple architectures necessarily more interpretable than more complicated ones? What does the paper claim?
Ans: The paper challenges the conventional wisdom that simpler models are inherently more interpretable than complex ones. It argues that the interpretability of a model does not solely depend on its complexity or simplicity. Instead, it can vary based on the definition of interpretability being considered, such as transparency or post-hoc explanations, and sometimes more complex models can provide insights that simpler models cannot.

4. Do you agree with the paper's explanation of Interpretability? Are there components that you think it did not cover?

Ans: Personally, I find the paper's exploration of interpretability nuanced and insightful, especially its emphasis on the multifaceted nature of interpretability and the need to specify what kind of interpretability is being discussed. However, the paper also suggests that the conversation around interpretability in machine learning is still evolving, and there may be aspects not fully covered, such as the quantifiable metrics for interpretability, the role of user experience design in making models more interpretable, and the exploration of interpretability in unsupervised or reinforcement learning contexts.

Week 10 Tuesday Paper 2: Methods for Interpreting and Understanding Deep Neural Networks

Supplemental: Impossibility Theorems for Feature Attribution

1. What is an Expert in Activation Maximization, and why does it matter?
Ans: Activation Maximization seeks an input pattern that maximizes the response for a specific model output, often represented as a class probability. Incorporating an "expert" in AM, which is essentially a data density model p(x), leads to prototypes that not only produce strong class responses but also resemble the data more closely, making them more natural and representative. The choice of the expert significantly affects the resulting prototype, striking a balance between focusing on probable input regions and avoiding overfitting to specific data distributions.

2. What are some different ways to decompose the prediction of a neural network?
Ans: Decomposing neural network predictions can be approached through relevance propagation and other techniques like sensitivity analysis, simple Taylor decomposition, and specific methods like deconvolution and guided back-propagation. These techniques provide insights into how individual input features contribute to the network's prediction, emphasizing the importance of understanding the contributions of different layers and neurons within the network.

3. What is *locality* in the context of AM? Give a real-world example in which we might want to explore locality in a model.
Ans: Locality in AM refers to focusing on a specific region of the input space when analyzing complex, multimodal, or elongated probability functions. For instance, in biomedical data, analyzing data conditioned on a specific stage of a medical condition or relating to a particular subject or organ exemplifies exploring locality. This approach helps in creating prototypes that are representative of specific conditions or stages, rather than attempting to cover the entire conceptual space

4. Explain briefly how LRP works and its relationship with Taylor Decomposition.
Ans: LRP explains DNN decisions by redistributing the output prediction back through the network layers to the input layer, assigning relevance scores to individual inputs. The relationship between LRP and Taylor Decomposition, particularly for deep ReLU networks, is shown in the process where LRP's redistribution at each layer can be seen as performing a Taylor decomposition. This connection underlines the method's

capacity to provide detailed insights into how inputs influence network predictions by approximating the contribution of each input as part of a decomposed output function

5. What are the paper's recommendations about choosing the model to explain? What is good and bad about them?
Ans: The recommendations for choosing a model to explain with LRP suggest favoring models where LRP has been successfully applied, like convolutional neural networks with ReLU nonlinearities, and minimizing the number of fully connected layers to preserve explanation selectivity. These recommendations are grounded in practical observations and empirical validations, aiming to enhance the clarity and relevance of explanations. However, they might limit the applicability of LRP to a narrower range of models and potentially overlook the nuances in models that don't fit these criteria

6. What are explanation *continuity* and *selectivity*? What do they tell us about the quality of a model explanation?
Ans: Explanation continuity and selectivity are two important qualities of model explanations. Continuity ensures that small changes in the input lead to small changes in the explanation, which is crucial for understanding the stability and reliability of the explanations provided by a model. Selectivity measures how well the explanation identifies features that significantly impact the model's output. High selectivity means the model effectively pinpoints the most relevant input features for its predictions, indicating clearer and potentially more actionable insights into the model's behavior.