Explainable AI

Group Members:

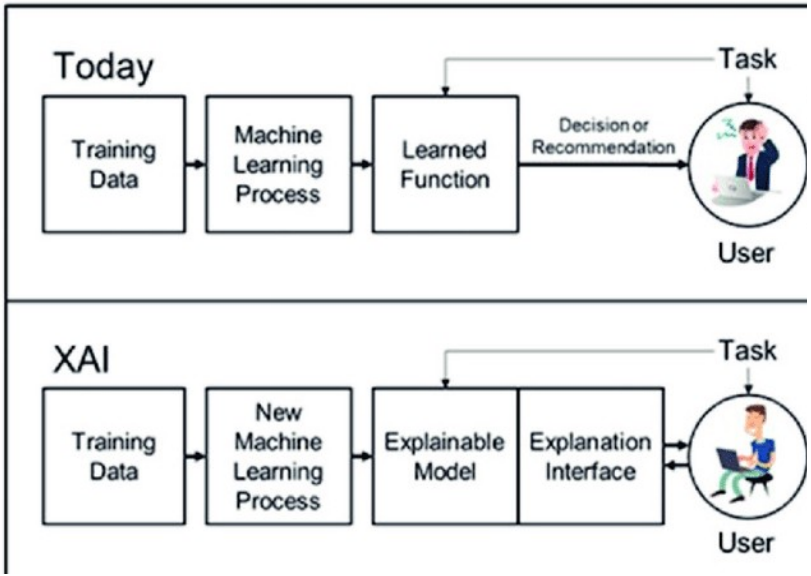
Akshay Daberao (IIT2020182)

Saurabh Patel (IIT2020185)

Tanmay Solanki (IIT2020188)

Introduction :

- Explainable AI addresses the **"black box"** nature of many machine learning models.
- Techniques used for X-AI
  - Lime, SHAP, LRP.
- **Why XAI ?**
  - Provide clear understanding **how solution is reached?**
  - It's a key element of responsible AI, embedding ethics, trust, and accountability into AI systems.



Four Properties of Good Explanation Techniques :

- **Conservation:** if we find explainable evidence in the output, it must show up somewhere in the input features (no loss of evidence)

- **Positivity:** either a feature is relevant (positive) or irrelevant (zero)

- **Continuity:** if two inputs are almost the same, and the prediction is almost the same, then the explanation should be almost the same

- **Selectivity:** models must agree with explanation. Removing evidence from input should reduce confidence in the output

Local
Interpretable
Model-agnostic
Explanations

LIME(Local Interpretable Model-agnostic Explanations) :

- **Local** : Focuses on providing explanations at the local level instead of explaining the global behavior of the entire model.
- **Interpretable**: The goal of LIME is to generate interpretable explanations for machine learning models.
- **Model-agnostic**: It doesn't rely on the specifics of the internal workings of a given model, making it versatile and applicable to different types of models.
- **Explanations**: LIME generates explanations to shed light on why a model makes specific predictions.

- LIME often provides visualizations, such as feature importance plots or heatmaps, to make the explanation more understandable.

How LIME works?

- **Perturbed Samples**: Generate perturbed samples by making small changes into it.
- **Model Predictions**: Both the original CNN model and its predictions on the perturbed samples are used.
- **Feature Weights**: LIME assigns weights to the perturbed samples based on the proximity of each sample to the original instance
- **Build Surrogate Model**: LIME uses the perturbed samples and their corresponding model predictions to train a locally interpretable surrogate model.
- **Feature Importance**: The coefficients of the surrogate model provide insights into the importance of different features.
- **Interpretability**: The surrogate model, being simpler, is more interpretable than the original CNN model. It allows users to understand how changes in specific features influence the model's predictions in the local region around the selected instance.
- **Explanation Visualization**: LIME often provides visualizations, such as feature importance plots or heatmaps, to present the explanation in an easily understandable format.

SHAP

SHAP(Shapely Additive Explanation):

- The foundational concept in SHAP is drawn from cooperative game theory, specifically Shapley values. In cooperative game theory, the Shapley value assigns a value to each player (feature in our context) indicating its importance in the overall outcome achieved by collaborating players (features).

- SHAP values assign an importance value to each feature in a model. Features with positive SHAP values positively impact the prediction, while those with negative values have a negative impact. The magnitude is a measure of how strong the effect is.

How SHAP works?:

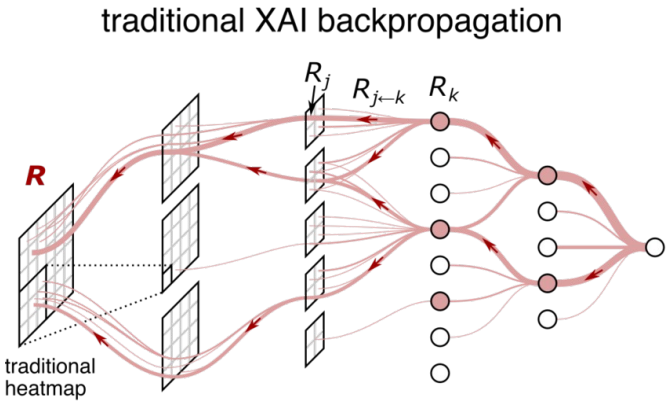
- **Baseline Model Prediction**: SHAP begins by defining a baseline or reference prediction, typically the average output of the model based on a background dataset.
- **Feature Attribution**: For each instance, SHAP explores all possible combinations of features and computes the contribution of each feature in predicting the output compared to the baseline prediction. This is essentially an exhaustive process to calculate the contribution of each feature to the difference between the model's prediction for an individual instance and the baseline prediction.
- **Shapley Values Calculation**: SHAP uses a weighted average of these contributions to compute the Shapley values for each feature. These values indicate the average contribution of a feature to the prediction across all possible combinations.
- **Interpreting Results**: The Shapley values provide an intuitive explanation of a model's output by attributing contributions to each feature. Positive Shapley values indicate features pushing the prediction higher, while negative values suggest features pushing the prediction lower.

Types of SHAP Explainers :

- **KernelExplainer:** Approximates feature importance in predictions by using a weighted linear regression model with a specialized function (kernel).
- **TreeExplainer:** Tailored for tree-based models, Tree SHAP efficiently computes exact Shapley values by leveraging the structure of decision trees, random forests, or gradient boosting machines.
- **DeepExplainer:** Tailored for complex neural networks, it reveals the significance of individual features in making predictions.
- **GradientExplainer:** Gradient SHAP estimates Shapley values by computing expected gradients from the model's input to output, suitable for models capable of computing gradients efficiently.

LRP (Layer wise relevance propagation) :

- LRP stands for Layer-wise Relevance Propagation, The goal of LRP is to attribute the model's output to its input features, providing insights into why a particular prediction was made.
- Propagates prediction backwards based on some set of rules.



$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$$

How does LRP works? :

1. **Forward Pass:** During the forward pass, the neural network processes input data through its layers to make a prediction. Each neuron's output is computed based on its inputs, weights, and activation function.

2. **Initialize Relevance:** Initialize the relevance score for the output layer (final prediction) with the prediction score obtained from the forward pass.

3. **Backward Pass :** Iterate backward through the layers of the network, propagating the relevance score from the output layer to the input layer. For each layer, compute the layer-wise relevance.

Advantages of Lime

- Model Agnostic: Applied to wide Range of models
- Local Interpretability: LIME provides local explanations, making it suitable for understanding the behavior of a model for specific instances, which can be valuable for individual predictions.
- Human-Interpretable Explanations: The surrogate models created by LIME are typically simple and interpretable, such as linear models, making the explanations more understandable for humans.
- Perturbation-Based Approach: The perturbation-based approach of LIME, where it generates perturbed samples around a specific instance, is a practical and intuitive method for approximating model behavior.

Disadvantages of Lime

- Sample Dependency: LIME's explanations can be sensitive to the choice of perturbed samples, and different samples may lead to different local surrogate models.
- Local Scope: LIME is designed for local interpretability, providing insights into individual predictions.
- Surrogate Model Simplification: Can miss complex interactions present in the original model.

Advantage Of SHAP

Unified Framework: SHAP provides a unified framework for interpreting a wide range of models

Theoretical Foundation: SHAP values are based on game theory, providing a solid theoretical foundation for understanding feature contributions.

Global and Local Interpretability: SHAP allows for both global and local interpretability

Handling Feature Interactions: SHAP values handle complex feature interactions, providing a nuanced understanding of how different features contribute to model predictions.

Disadvantage Of SHAP

Computational Complexity: SHAP may become computationally expensive, especially for high-dimensional models or large datasets, impacting the scalability of the method.

Advantage of LRP

Consistency with Model Output: LRP strives to ensure that the sum of relevance scores across all input features equals the model's output, maintaining consistency and interpretability.

Applicability to Deep Neural Networks: LRP can be applied to deep neural networks, including convolutional and recurrent architectures, making it versatile for interpreting a wide range of models.

Handling Non-Linearity: LRP addresses non-linear activation functions in neural networks, allowing for a more accurate representation of feature importance in complex models.

Disadvantage of LRP

Sensitivity to Network Architecture: LRP's performance may vary based on the specific architecture of the neural network.

Global Context Limitations: LRP primarily focuses on local explanations and may have limitations in capturing global contextual information, especially when dealing with long-range dependencies in sequential data.

Complexity in Implementation: Implementing LRP correctly and efficiently can be complex, Misimplementation may lead to incorrect interpretations.

Thank You

:)