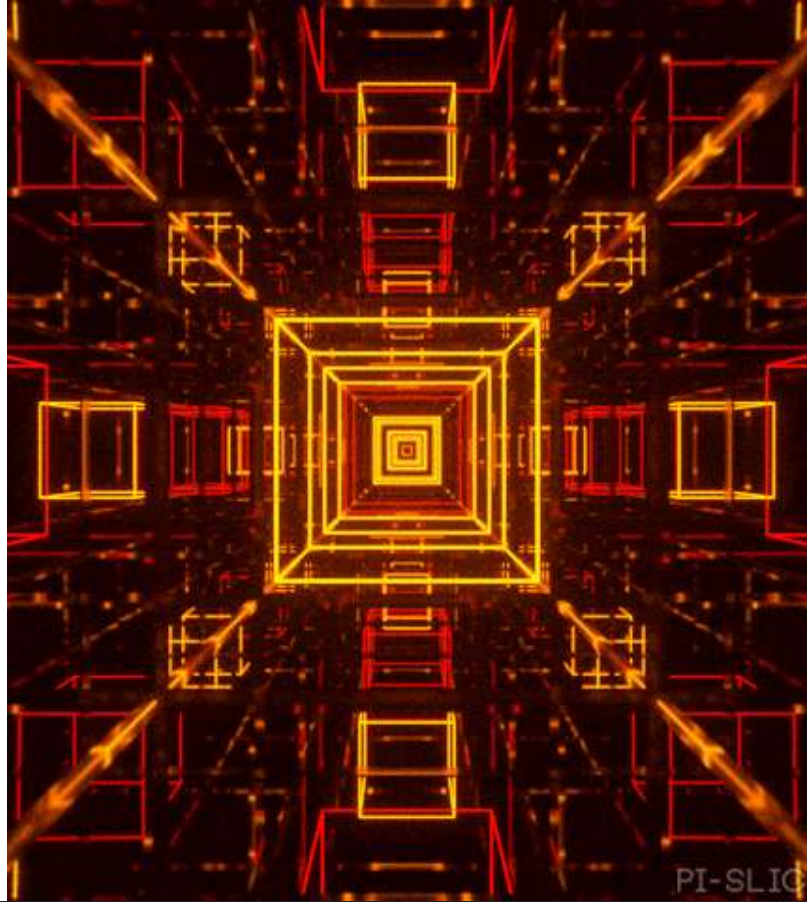
RCIF Applied AI Seminar Series: Explainable AI for Healthcare Applications

Introduction

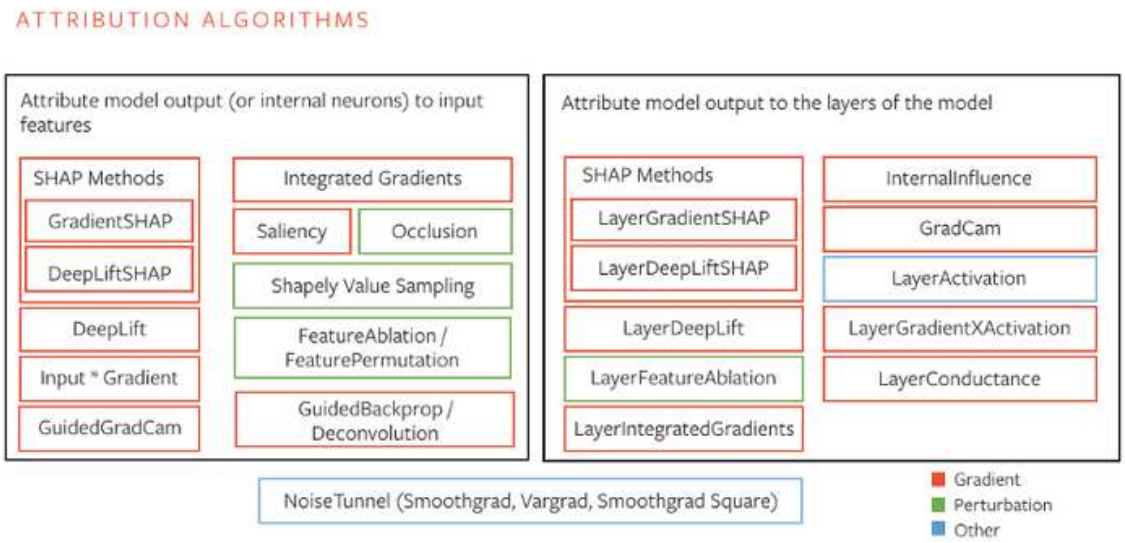
- Transparency: Explainable AI (xAI) clarifies AI decisions, building trust among clinicians and patients.
- Safety: Understanding AI processes helps identify and correct potential biases or errors, ensuring reliable interventions.
- Compliance: xAI meets regulatory standards requiring transparency and accountability, aiding approvals and integration.
- Collaboration: xAI fosters better clinician-AI interaction, enhancing the accuracy and effectiveness of patient care.
- Patient Trust: Clear explanations of AI decisions increase patient acceptance and adherence to treatments, improving outcomes.

Peeking into the Black Box

- Integrated Gradients & Feature Ablation: Identify feature importance across modalities.
- Layer Activation & Layer Attribution: Understand how specific layers contribute to model decisions, crucial for deep imaging models.
- Saliency Maps: Visualize which parts of an input image influence the model's predictions.
- Cross-Model Comparisons: Facilitate understanding of how different model architectures process and prioritize information.
- Customizable & Extensible: Tailor explanations to specific multimodal interactions and imaging modalities.



xAI with Captum

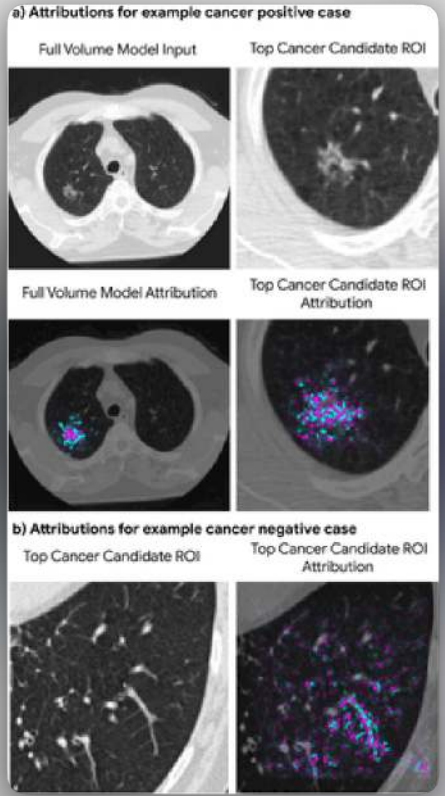


Methods and Use Cases

Algorithm	Use Case for CNNs	Pros	Cons
SHAP Methods	Understanding feature attributions across different image regions.	Accurate, interpretable, model-agnostic.	Computationally expensive, complex to implement.
GradientSHAP	Adding robustness to gradient-based explanations in CNNs.	Improves robustness, combines gradients with SHAP.	Still computationally intensive, can be sensitive to noise.
GuidedGradCam	Highlighting important regions in images processed by CNNs.	Combines benefits of Guided Backpropagation and GradCam, highly interpretable.	Requires careful implementation, may not capture global features.
Integrated Gradients	Providing comprehensive feature attributions for CNN outputs.	Theoretically sound, accurate, scalable.	Requires baseline selection, computationally intensive.
Saliency	Quick approximation of important regions in images processed by CNNs.	Simple, fast, easy to implement.	Less precise, can be noisy, lacks robustness.
Occlusion	Systematically masking parts of the input image to understand impact on CNN output.	Intuitive, easy to understand, model-agnostic.	Computationally expensive, may miss interactions between features.
GuidedBackprop/Deconvolution	Visualizing which features in input images contribute most to CNN's predictions.	High-resolution visualizations, interpretable.	May produce misleading results, only highlights positive gradients.
GradCam	Highlighting the most relevant regions in images for CNN decisions.	Intuitive, effective for localization tasks, visually interpretable.	Limited to convolutional layers, lower resolution.
LayerActivation	Visualizing activations of CNN intermediate layers.	Provides insight into layer behaviors, interpretable.	Does not directly attribute features to outputs.
NoiseTunnel (Smoothgrad, Vargrad, Smoothgrad Square)	Adding noise to gradient-based methods for more robust attributions in CNNs.	Enhances robustness, reduces noise in attributions.	Adds computational overhead, may still require parameter tuning.

Intro to Integrated Gradients

- Definition: Integrated Gradients (IG) is an attribution method that helps explain the predictions of deep neural networks.
- Purpose: It assigns importance scores to input features based on their contribution to the model's output.
- Transparency: Provides insight into how features influence predictions, enhancing model interpretability.
- Gradient-Based: Utilizes gradients to measure changes in predictions relative to input features.
- Path Integration: Integrates gradients along a path from a baseline to the actual input.



How does it work?

- Baseline Selection: Choose a baseline input (e.g., a black image) for comparison.
- Gradient Calculation: Compute gradients of the model's output with respect to the input features.
- Path Integration: Integrate these gradients along the path from the baseline to the input.
- Attribution Mapping: Visualize the importance scores for each input feature.

Occlusion

- Definition: Occlusion is a method for interpreting neural network predictions by systematically masking parts of the input and observing the change in the output.
- Purpose: Helps determine the importance of different regions in an input image for the model’s decision-making.
- Sliding Window: Occlude parts of the input image using a sliding window.
- Replace with Baseline: Replace the occluded region with a baseline value (e.g., black).
- Attribute Calculation: Measure the impact on the model's output for each occluded region.

How does it work?

Let's Open a Black Box...

Open

Schedule a AI/HPC Consult