Explainable AI

Gentle Introduction to XAI methods.

Roberto Souza
Electrical and Software Engineering
(Slides courtesy of Mahsa Dibaji)

W2025

UNIVERSITY OF CALGARY

Outline

- Introduction – What is explainability about?
- Taxonomy – Different XAI techniques
  - SHAP
  - GradCAM
  - Intrinsic Gradients
- Evaluation of XAI

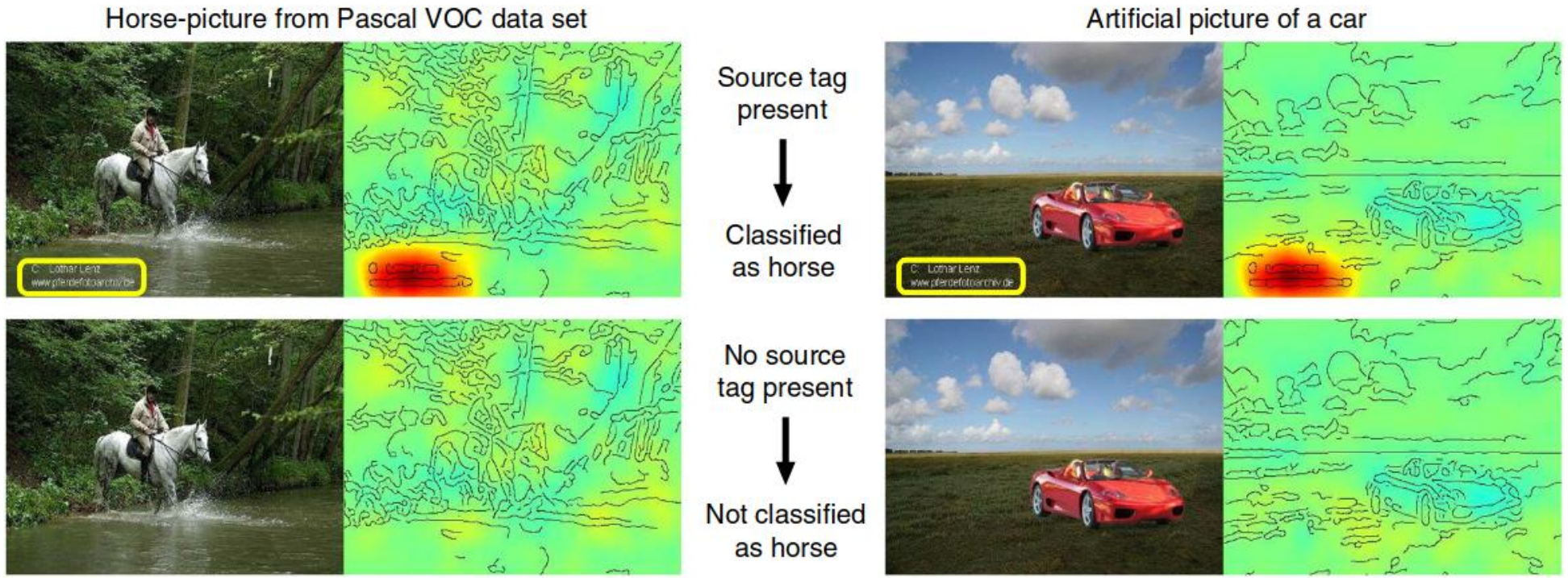
Learning Goals

- Get familiar with the purpose of XAI and the different categories model fall into.

- Get familiar with basics of some interpretability techniques.

Introduction

“Garbage in, Garbage out”



UNIVERSITY OF CALGARY

Introduction

Problem with “black box” model

Breast Cancer Detection Model Feng, 2020	
Accuracy rate:	0.85
Specificity:	0.857
Sensitivity:	0.847

Introduction

Importance of Explainability

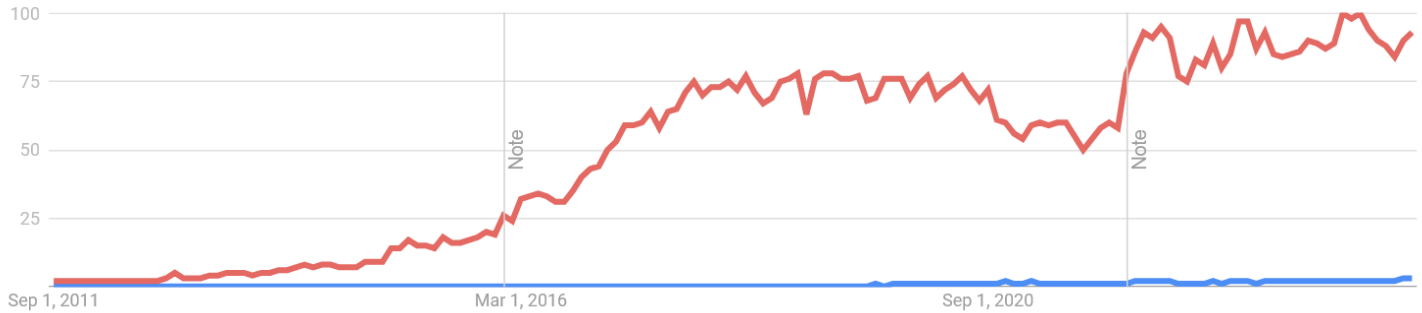
1. Trust and Reliability
   - Building Trust between user and AI system, Enabling User Feedback
   - Risk management
2. Compliance and Ethical Considerations
   - Making decision-process of AI systems more transparent
3. Model Debugging and Improvement
   - Debug AI models
   - Identify and mitigate biases

Introduction

- Explainability
  - Understanding the cause-and-effect relationships between inputs and outputs of a model. Focuses on the intuition behind model outputs.
- Interpretability
  - Understanding the internal logic and mechanics of a model. Focuses on the inner workings and processes of the model.

Explainable AI (Google trends)

Rising Interest in “Explainable AI”: A Google Trends Analysis Over Time (2011-2024)

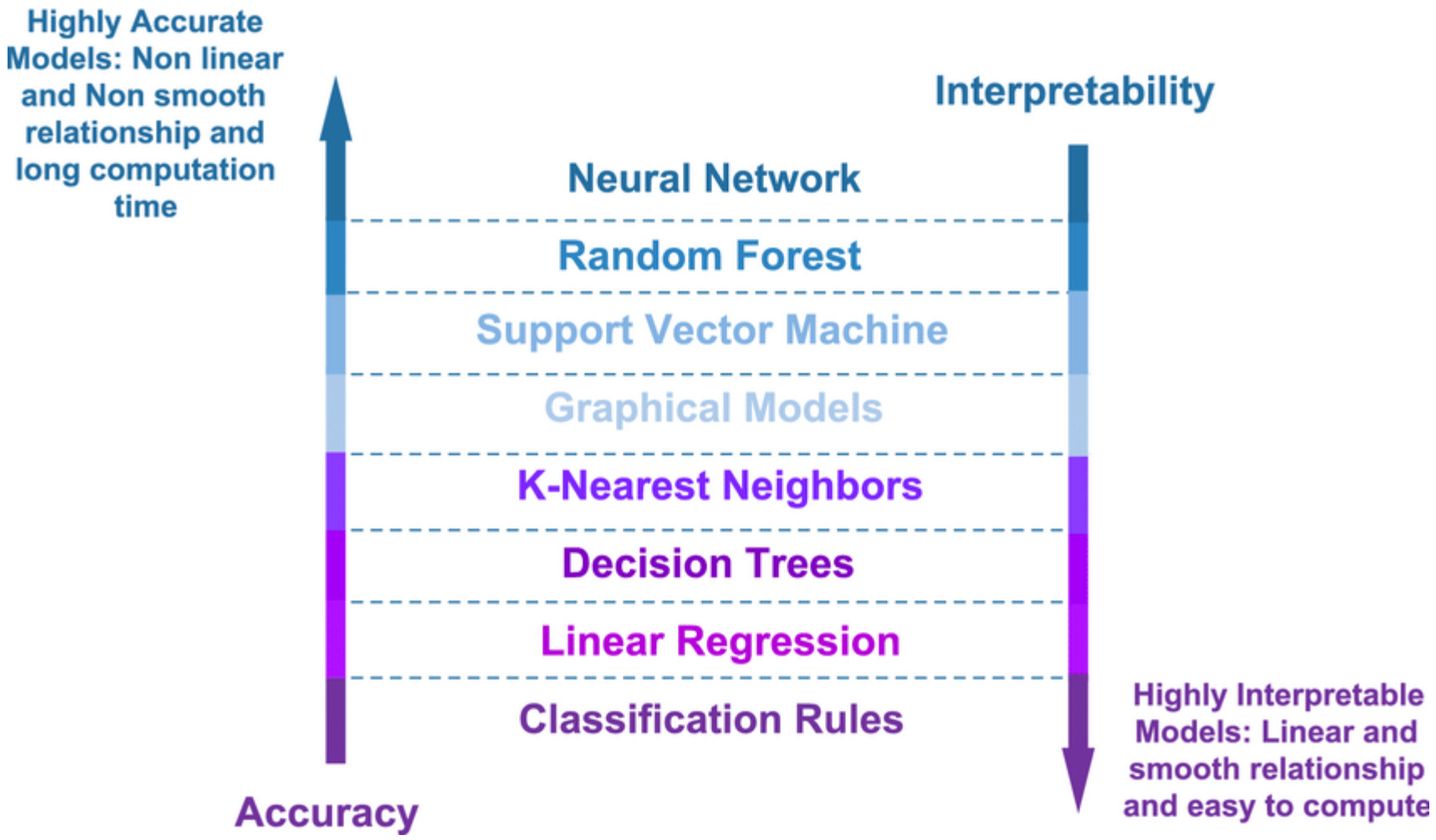


Deep Learning
Explainable AI

But still low compared to “deep learning”: A Google Trends Analysis Over Time (2011-2024)

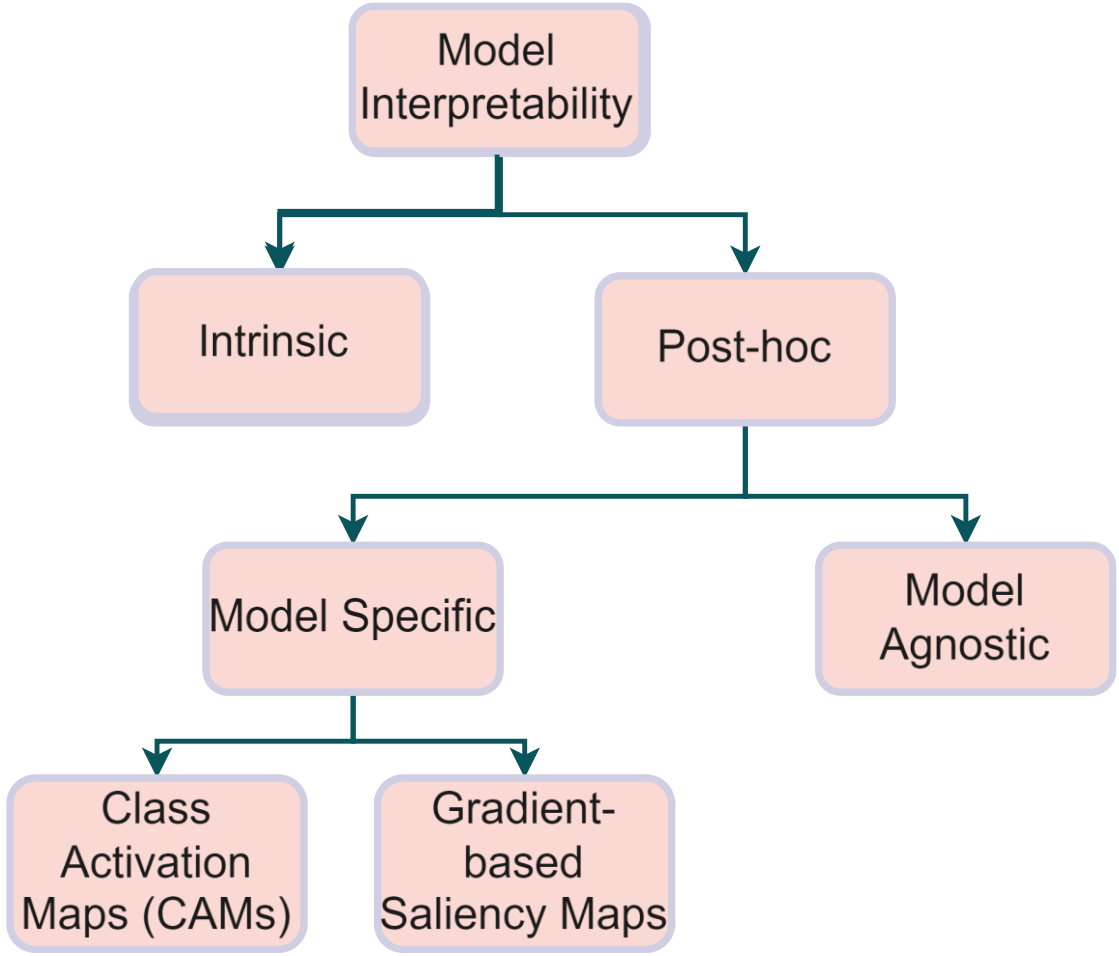
UNIVERSITY OF CALGARY

Introduction

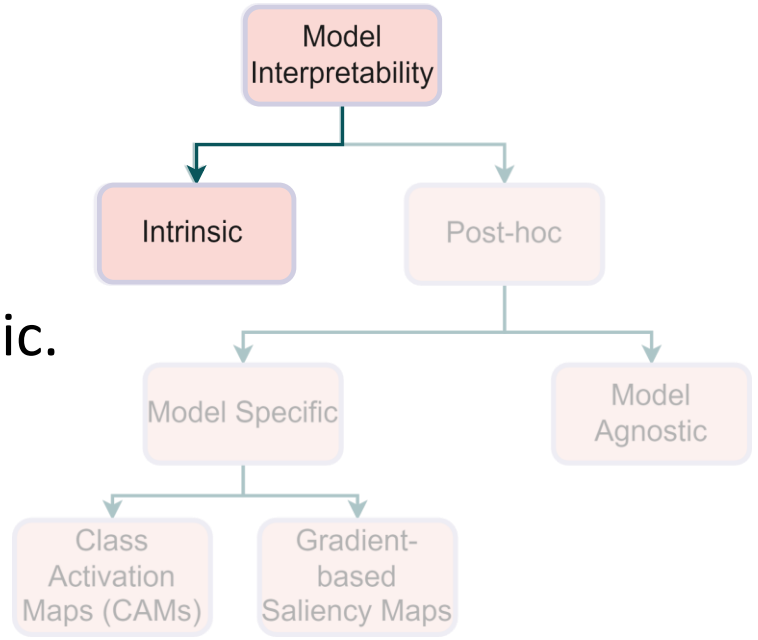


Demystifying ML Model Interpretability, Manish Kumar, Medium

Interpretability Taxonomy



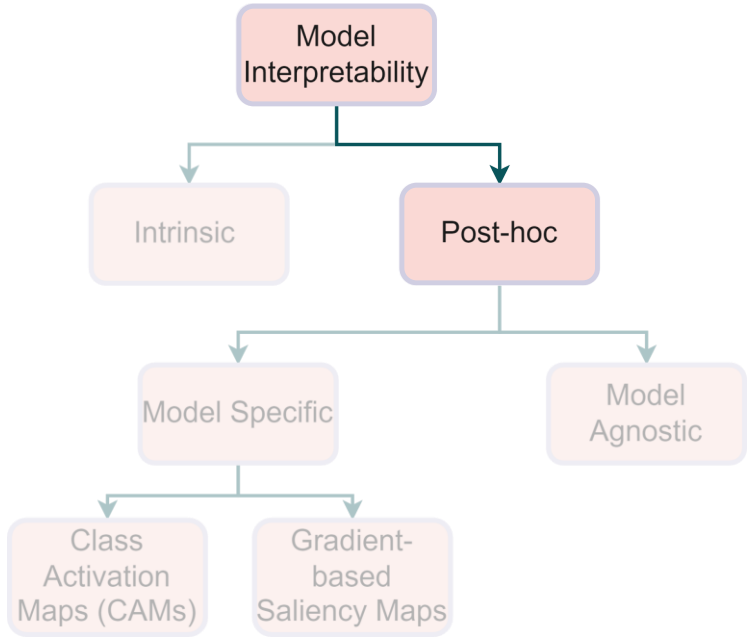
Intrinsic Interpretability



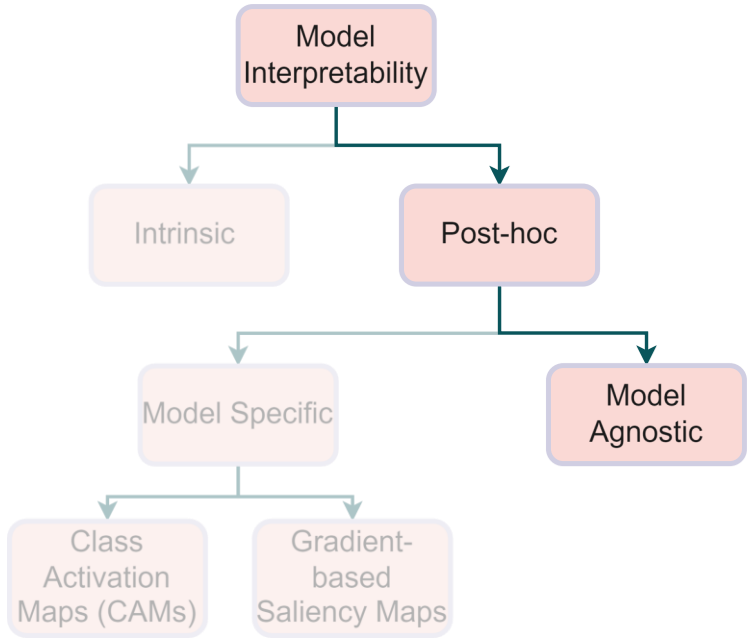
- The interpretability of ‘white-box’ models is defined as intrinsic.
- These models are interpretable by design.
- Examples
  - Rule-based systems such as decision trees.
  - A shallow neural network
  - Logistic regression
- Good choice for high-stakes decisions!

Post-hoc Interpretability

- Explaining a prediction of the model without explaining the exact internal mechanism of that model.
- Allows for the interpretation of complex models such as CNNs that are not inherently interpretable.
- Model-agnostic vs Model specific



Model-agnostic Techniques



- Techniques that can be applied to any model, regardless of its internal structure.
- To provide a way to interpret the decisions of complex models without needing access to their internal workings.
- Key Characteristics:
  - Independence from model type.
  - Focus on input-output relationships.
- Challenges: May not capture the intricacies of specific models, leading to less precise explanations.
- Examples:
  - LIME
  - SHAP
  - Perturbation-based methods

SHAP - SHapley Additive exPlanations

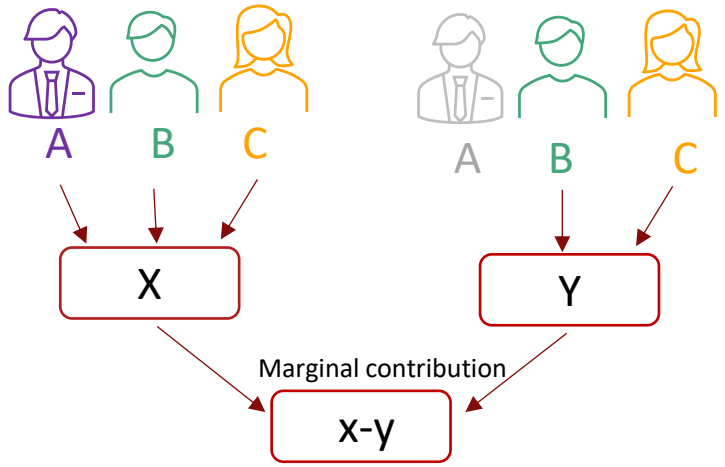
- **Shapley Values:** concept from cooperative game theory used to fairly distribute the total payoff (predictions) among players (features) based on their individual contributions.
- It quantifies the importance of each feature in model's decision-making process.

$$ShapV_i = \frac{1}{\#\ Coalitions} \sum_{All\ Coalitions} (Prediction\ with\ F_i - Prediction\ without\ F_i)$$

Marginal contributions

Coalitions: A subset of features working together to influence the prediction.

SHAP - SHapley Additive exPlanations

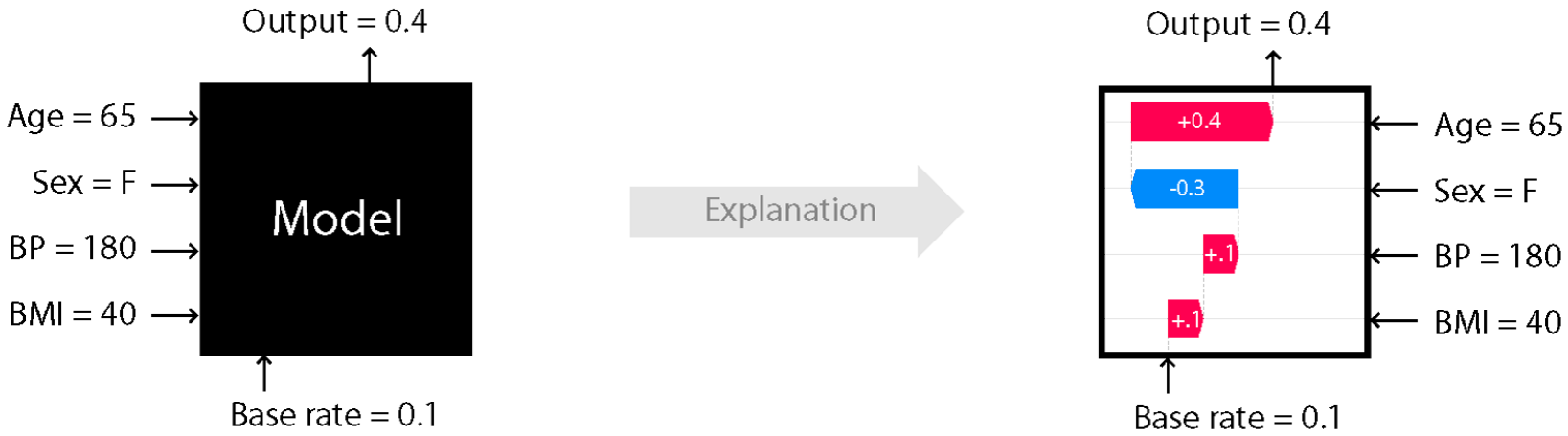


	Cost
A	$50
B	$60
C	$80
AB	$70
BC	$90
AC	$100
ABC	$120

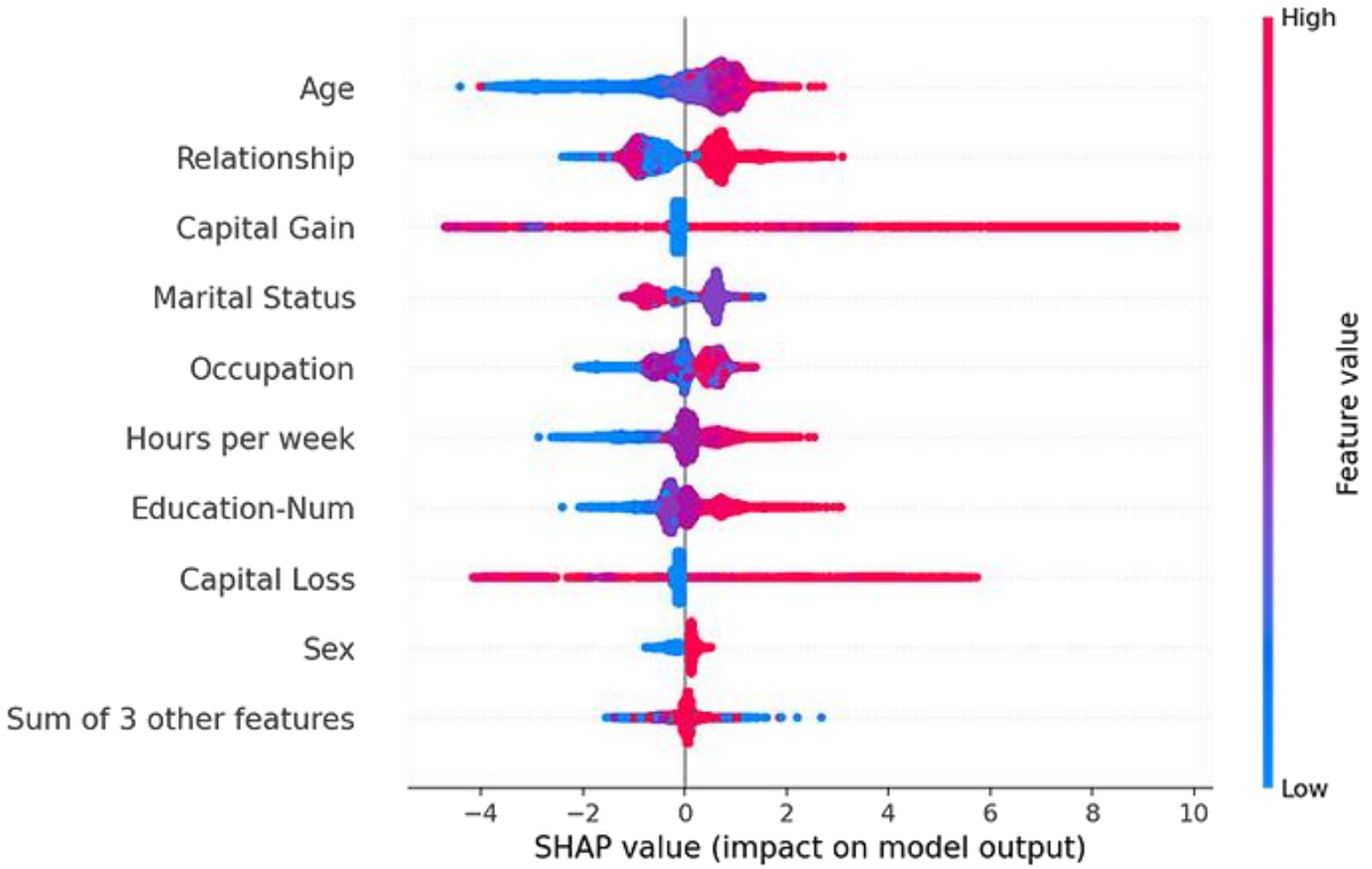
Constellation	A	B	C
ABC	$50	$20	$50
BAC	$10	$60	$50
ACB	$50	$20	$50
BCA	$30	$60	$30
CAB	$20	$20	$80
CBA	$30	$10	$80
Average	**31.7**	**31.7**	**56.7**

1. **Efficiency:** the final reward must be shared among all
2. **Symmetry:** players who made the same contribution will receive the same amount of reward.
3. **Dummy:** players who did not contribute will receive no reward.
4. **Additivity:** Shapley values for different predictions can be added together

SHAP - SHapley Additive exPlanations



Example on Tabular Data (Adult Census Income)



- **Magnitude:** A larger absolute SHAP value a larger impact on the model's output.
- **Sign:** whether the effect of that feature is to increase or decrease the prediction.
- **Color:** Value of the feautre

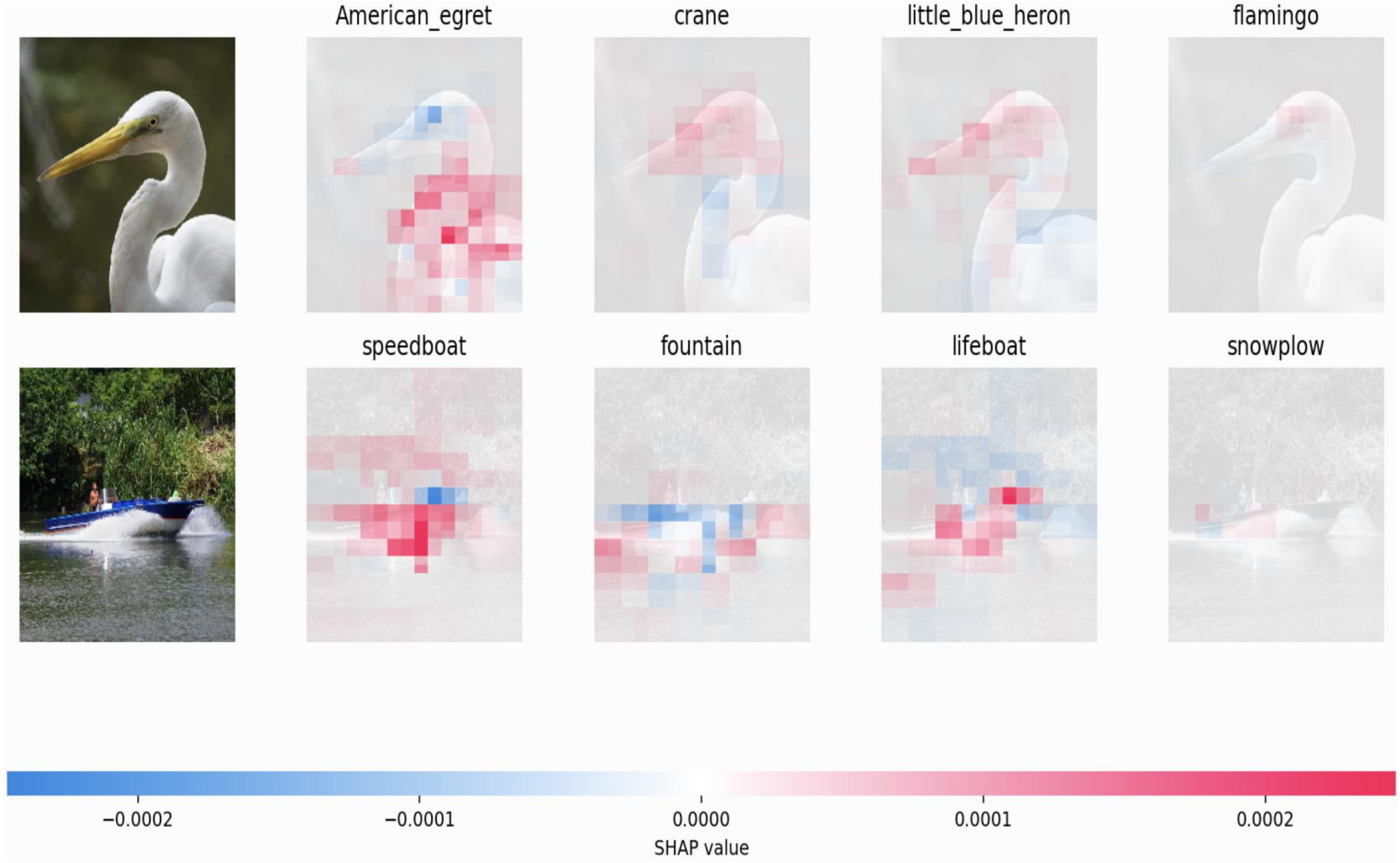
How SHAP works?

1. **Select a prediction to explain:** Choose an instance from your dataset that has been run through the model to get a prediction.
2. **Create a baseline output:** The baseline is typically the average prediction over the dataset, representing the prediction that would be made without any information about the current instance.
3. **Compute SHAP Values:**
   - For each feature, consider all possible sets of features (coalitions) that could include that feature. Compute the change in prediction with and without the feature across these sets.
   - The average of these differences across all coalitions is the SHAP value for that feature.
   - The SHAP value represents the feature's average marginal contribution to the prediction compared to the baseline.

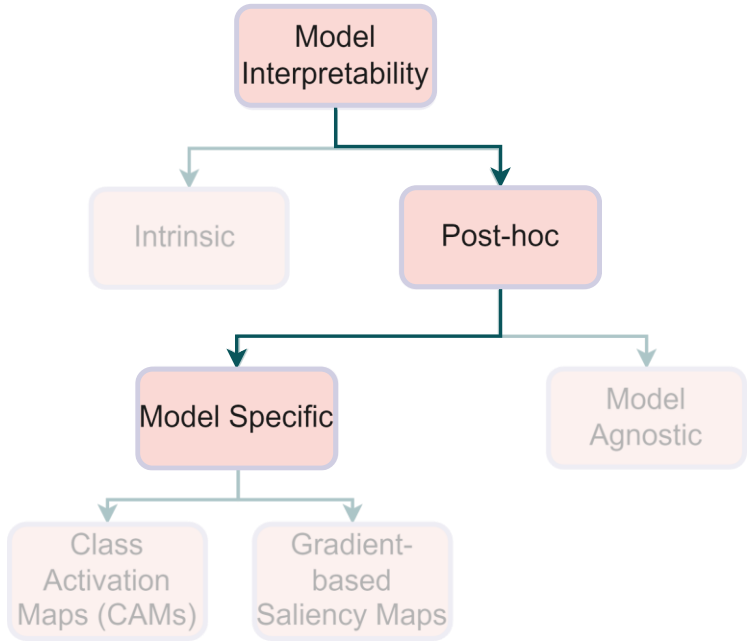
➢ Computation of Shapley values is complex for models with large number of features.

➢ Different SHAP explainers optimize computation of shapley values for different types of ML/DL.

Example on Image Data (ImageNet) – Deep SHAP



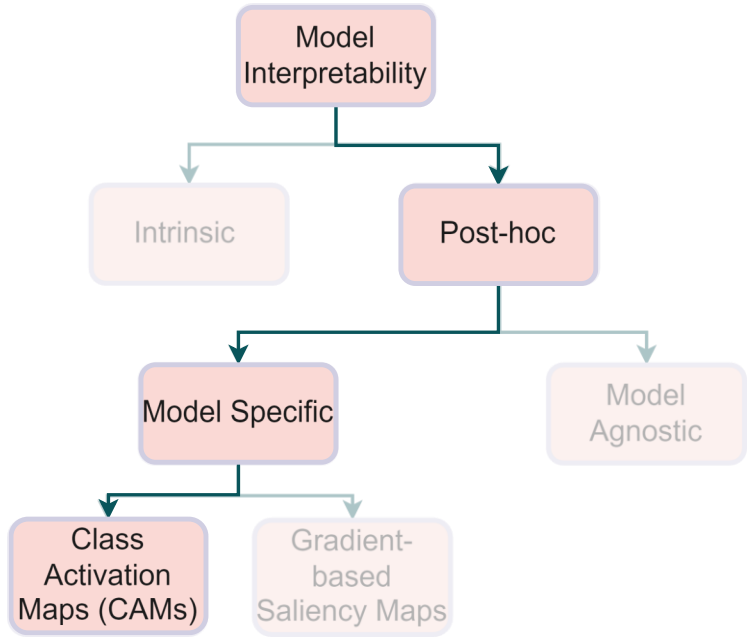
Model-specific Techniques



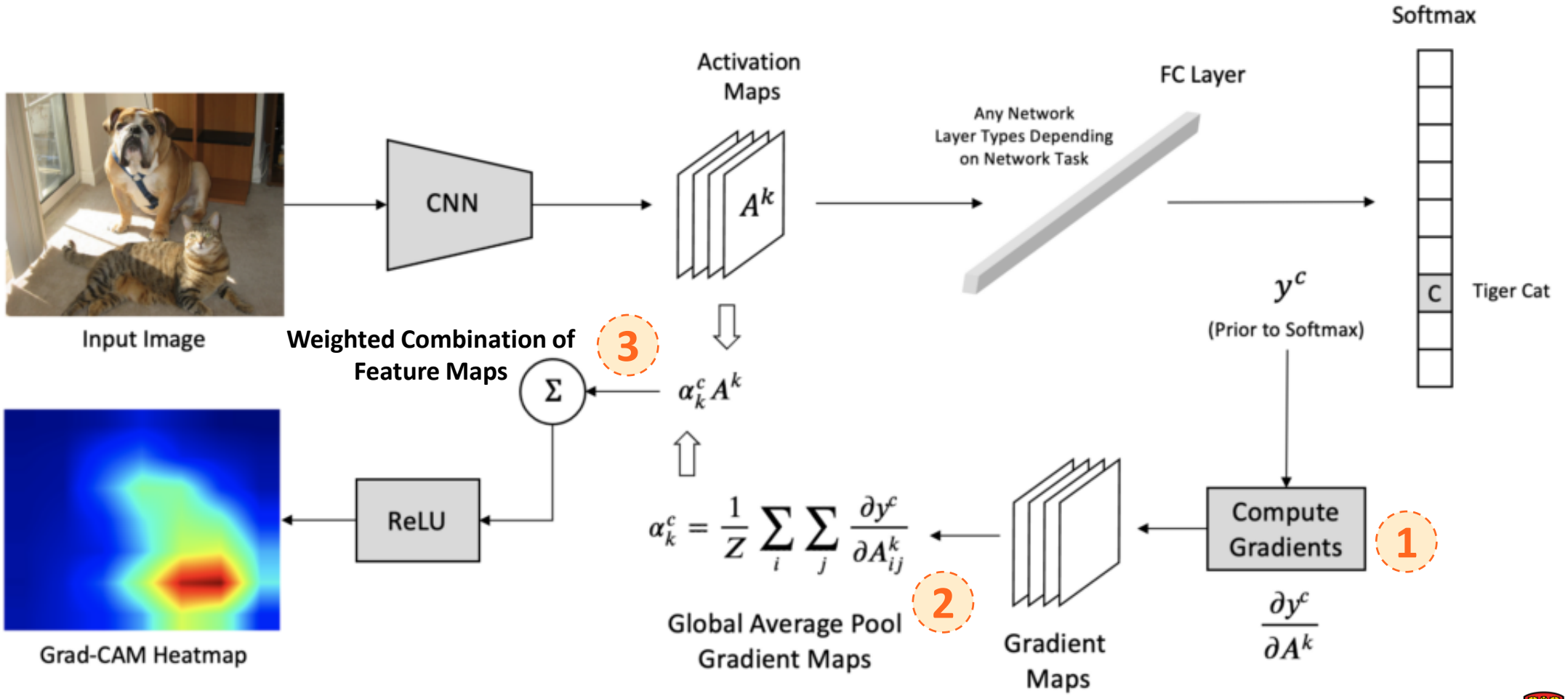
- Techniques designed to explain specific types/groups of models, utilizing their internal structure.
- leverage internal model structure of certain models to provide more detailed and accurate explanations.
- Key Characteristics:
  - Tailored to specific model architectures.
  - Can provide insights into the internal decision-making process of the model.
- Challenge: Limited in applicability to other model types.
- Examples:
  - Gradient-based saliency maps: Vanilla Gradient, Integrated Gradients, SmoothGrad
  - Class Activation Maps: GradCAM, GradCAM++
  - LRP

GradCAM

- Gradient-weighted Class Activation Mapping.
- A Generalization over CAM that can be applied to a wider range of convolutional neural network (CNN)
- Produce a coarse localization map highlighting the important regions in the image for making a prediction.

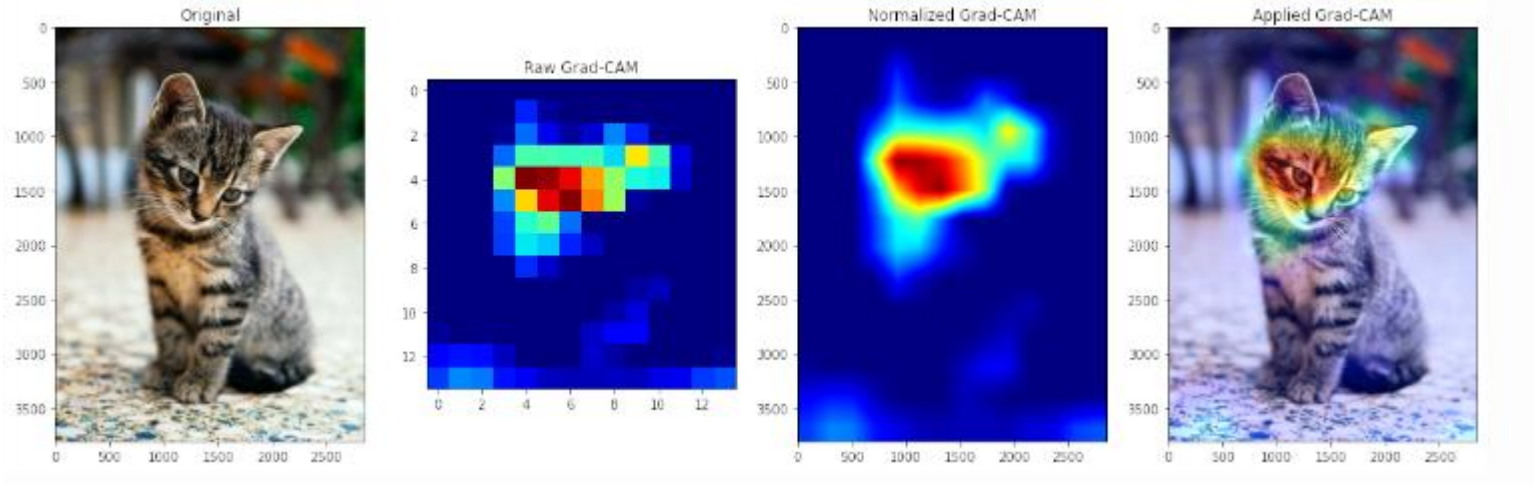


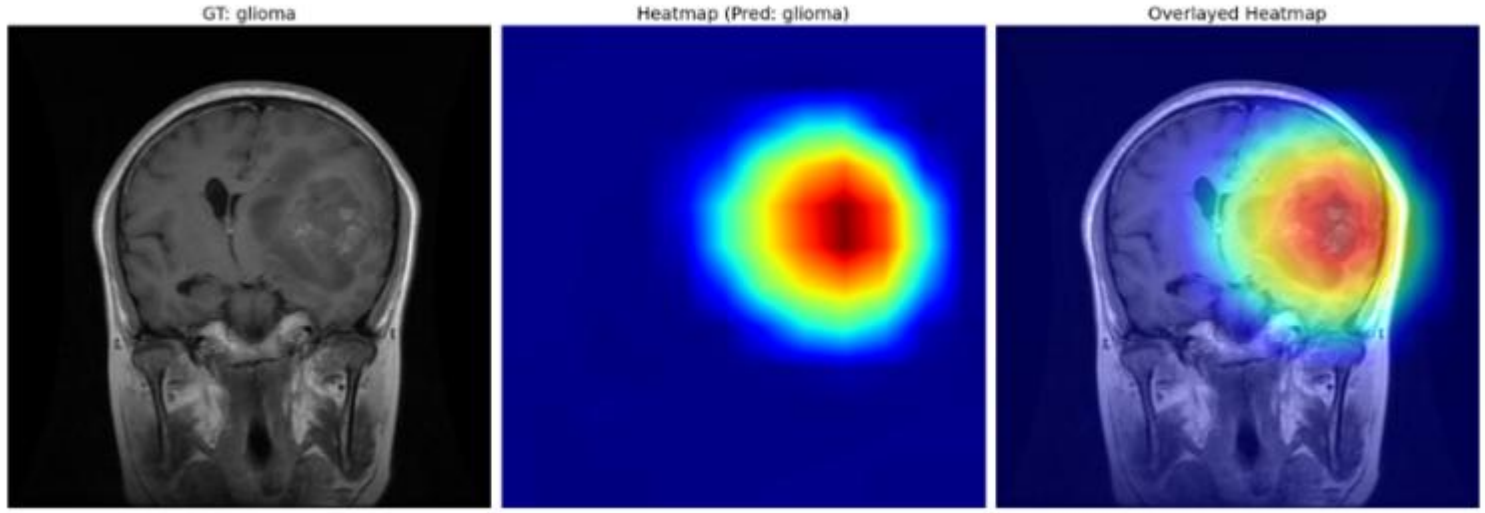
How GradCAM works?



UNIVERSITY OF CALGARY

GradCAM - Example





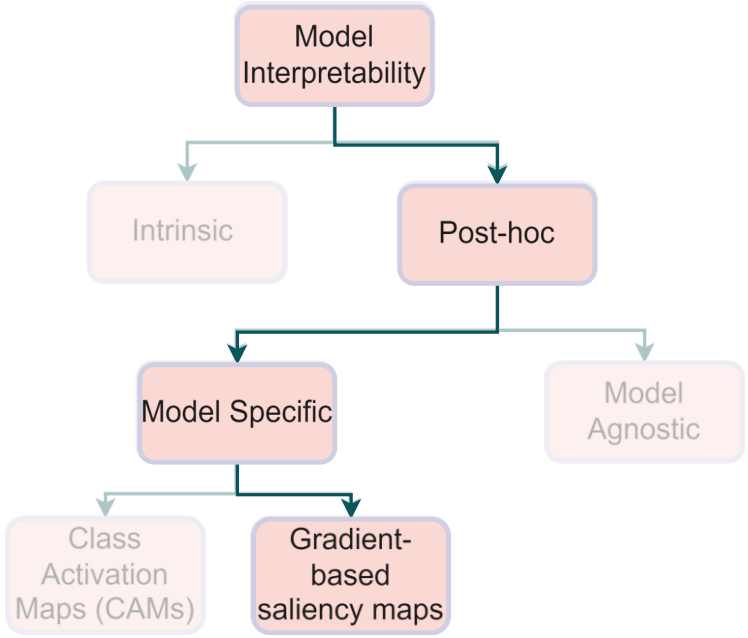
UNIVERSITY OF CALGARY

Integrated Gradients

- Integrated gradients aims to attribute the prediction of **any differentiable neural network** to its input features.

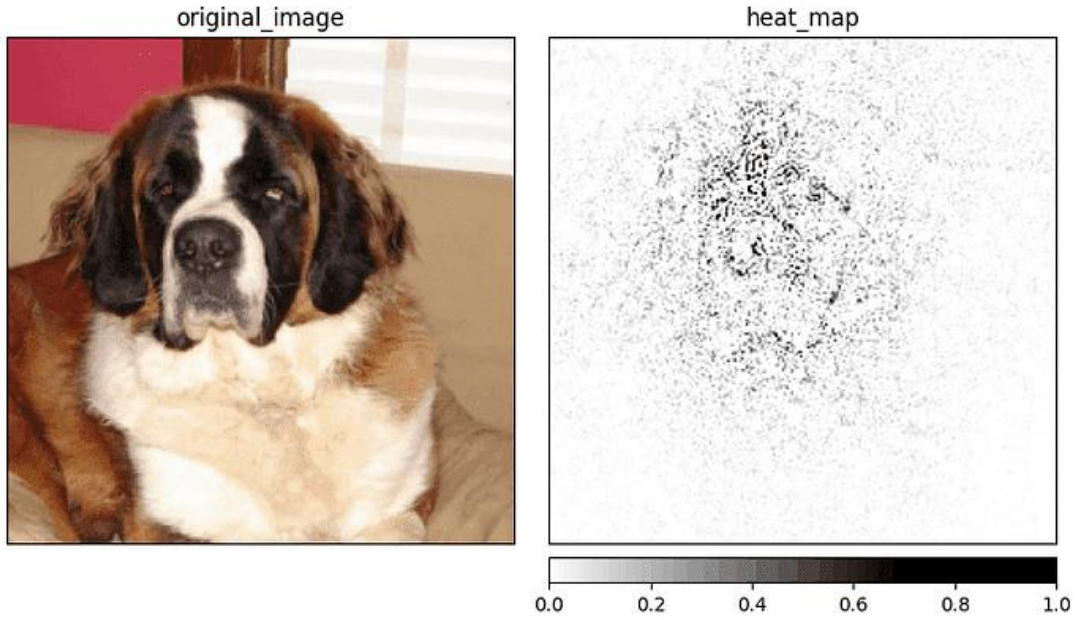
- Considers the **path** from **a baseline input** to the **actual input** and **computing the gradients** of the model's **output with respect to the input** along this path.

- Gradients provide information about how much each feature contributes to the change in the model's prediction as you move from the baseline to the actual input.



Integrated Gradients: Example

baseline → Actual Input



UNIVERSITY OF CALGARY

How Integrated Gradients works?

1. **Baseline Input:** An input that represents the absence of features or a neutral starting point; e.g., a black image.
2. **Path from Baseline to Actual Input:** Usually linearly interpolating between the baseline and the actual input.
3. **Compute Gradients Along the Path:** Gradients of output w.r.t each input.
4. **Integrate the Gradients:** Aggregates the contribution of each feature across the path.
5. **Attribution:** Integrated gradients along i^{th} dimension represent i^{th} feature's attribution (contribution to prediction)

$$IntegratedGrads_i^{approx}(x) ::= (x_i - x_i') \times \sum_{k=1}^{m} \frac{\partial F(x' + \frac{k}{m} \times (x - x'))}{\partial x_i} \times \frac{1}{m}$$

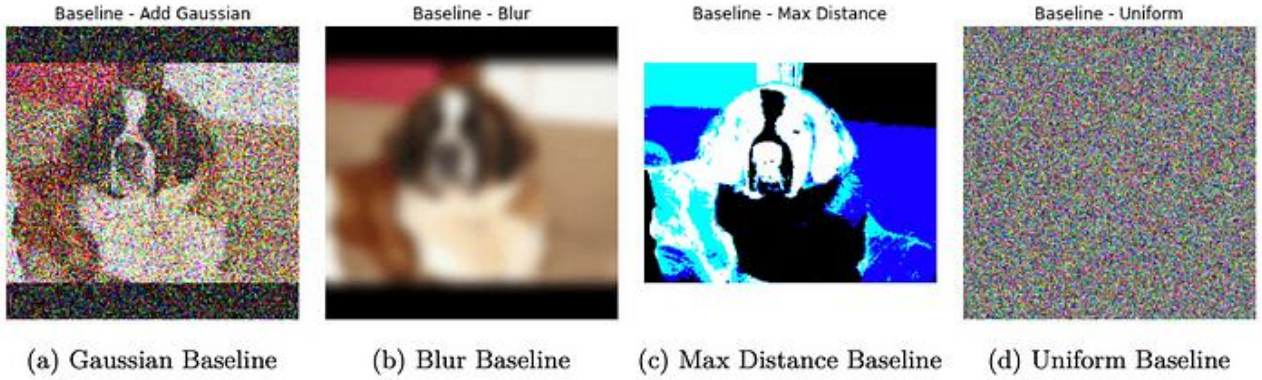
x: Actual Input x': baseline input

m: number of interpolations

F : $R^n \rightarrow [0, 1]$ is the deep network

Integrated Gradients - Additional notes

- Based on two axioms:
  - **Sensitivity**: *An attribution method satisfies **Sensitivity** if for every input and baseline that differ in one feature but have different predictions, then the differing feature should be given a non-zero attribution. If the function implemented by the deep network does not depend (mathematically) on some variable, then the attribution to that variable is always zero.*
  - **Implementation Invariance**: *Two networks are **functionally equivalent** if their outputs are equal for all inputs, despite having very different implementations. Attribution methods should satisfy **Implementation Invariance**, i.e., the attributions are always identical for two functionally equivalent networks.*
- Choice of Baseline: May affect the attributions and is still an open-question (potential research focus! ☺)



UNIVERSITY OF CALGARY

Measuring XAI Methods

- Qualitative measures:
  - Visually examining the interpretability maps.
  - Needs prior knowledge.
  - Could provide initial understanding of XAI method
  - Opinions of different humans is not comparable
  - Not reproducible

- Quantitative measures:
  - Comparable
  - Repeatable measurement

Quantitative Measure

- **Sensitivity**
  - measures the extent of explanation change when the input is slightly (insignificantly) perturbed.
  - the models that have high explanation sensitivity are prone to adversarial attacks.

$$\mathrm{SENS}_{\mathrm{MAX}}(\Phi, \mathbf{f}, \mathbf{x}, r) = \max_{\|\mathbf{y}-\mathbf{x}\| \leqslant r} \|\Phi(\mathbf{f}, \mathbf{y}) - \Phi(\mathbf{f}, \mathbf{x})\|$$

- **Infidelity**
  - represents the expected mean-squared error between the explanation multiplied by a meaningful input perturbation (e.g. noisy baseline) and the differences between the predictor function at its input and perturbed input.

$$\mathrm{INFD}(\Phi, \mathbf{f}, \mathbf{x}) = \mathbb{E}_{\mathbf{I} \sim \mu_{\mathrm{I}}} \left[\left(\mathbf{I}^T \Phi(\mathbf{f}, \mathbf{x}) - (\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x} - \mathbf{I})) \right)^2 \right]$$

Quantitative Measure - Example

Input

0.50	0.50	0.50
0.50	0.50	0.50
0.50	0.50	0.50

Input Attribution

0.00	0.09	0.18
0.38	0.08	0.10
0.30	0.00	0.06

Noise (select cell)

0.10	0.00	0.00
0.00	0.00	0.00
0.00	0.00	0.00

Masked Input

0.40	0.50	0.50
0.50	0.50	0.50
0.50	0.50	0.50

Masked Input Attribution

0.00	0.09	0.18
0.37	0.08	0.10
0.30	0.00	0.06

Infidelity: 0.0000008

Sensitivity: 0.0102377

Summary

- Explainable AI techniques helps us understand the decision-making process of ML/DL models.
- Interpretability is either intrinsic, or we can achieve it with post-hoc methods.
- Post-hoc interpretability methods are model specific or model agnostic.
- Evaluation of XAI models is through qualitative analysis or quantitative analysis with “sensitivity” or “infidelity” measurements.

References

- https://medium.com/@kemalpiro/xai-methods-integrated-gradients-6ee1fe4120d8
- https://towardsdatascience.com/xai-methods-the-introduction-5b1b81427c9c
- https://www.mdpi.com/1099-4300/23/1/18
- https://medium.com/@shahooda637/all-you-need-to-know-about-shap-for-explainable-ai-8ad35a05e6ec
- https://learnopencv.com/intro-to-gradcam/

More resources

- https://christophm.github.io/interpretable-ml-book/
- https://github.com/jphall663/awesome-machine-learning-interpretability
- https://github.com/Trusted-AI/AIX360
- https://github.com/interpretml/interpret