INVICTA Keynote Series — Explainable AI, P1

Anna Hedström, PhD candidate, TU Berlin

March 18-22, 2024

Porto, Portugal

@anna_hedstroem
@UMI_Lab_AI

“In this keynote series, we explore the rapidly evolving field of Explainable AI (XAI). First, beginning with a foundational overview, we trace the evolution of XAI methods from early neural network models to the recent transformer-based architectures, assessing their strengths and failure modes. Second, central to our discussions will be the critical theme of XAI evaluation, a previously understudied area that has caused confusion about which explanation methods work and under what conditions. We will concentrate on the fundamental challenge of XAI evaluation: how to establish and verify ground truth and review the community-driven approaches to it. Third, we will introduce Quantus in a hands-on practical session. Quantus is a toolkit for evaluating neural network explanations, exemplifying the practical application of XAI theory. Our goal is to equip the attendees with a deep understanding of both theoretical and practical aspects of XAI, providing a balanced view of the challenges and opportunities that characterise the current state of the field.

Today's agenda

01 Foundations

02 Evaluation

03 Quantus (hands-on)

04 Discussion + Q&A

Today's agenda

01 Foundations

02 Evaluation

03 Quantus (hands-on)

04 Discussion + Q&A

Today's agenda

01 Foundations

02 Evaluation

03 Quantus (hands-on)

04 Discussion + Q&A

Today's agenda

01 Foundations

02 Evaluation

03 Quantus (hands-on)

04 Discussion + Q&A

Today's agenda

01 Foundations

02 Evaluation

03 Quantus (hands-on)

04 Discussion + Q&A

Provide an **optimistic** and **critical** view

Short introduction

Short Bio

Overview Academic and Industry Experience

Now

- PhD candidate: TU Berlin, ML Group, Prof. Dr. Marina Höhne.
- Lab: Understandable Machine Intelligence, ATB (Explainable AI).
- Visiting Scientist: Fraunhofer HHI, XAI Group.

Before

- MSc at KTH. BSc at UCL.
- Roles in ML/AI: Klarna, BCG, Bosch, Neurocat.ai ....

Team Appreciation First

ANNA HEDSTRÖM

KIRILL BYKOV

PHILINE BOMMER

LEANDER WEBER

TOM BURNS

WOJCIECH SAMEK

SEBASTIAN LAPUSCHKIN

MARINA M.-C. HÖHNE

Team Appreciation First

ANNA HEDSTRÖM

KIRILL BYKOV

PHILINE BOMMER

LEANDER WEBER

TOM BURNS

WOJCIECH SAMEK

SEBASTIAN LAPUSCHKIN

MARINA M.-C. HÖHNE

+ online collaborators

Anna — Research Areas

Selected Works

(A) Evaluation-centric XAI

- How do we evaluate the quality of the explanation method, without access to ground truth explanation labels? **Quantus** (JMLR, 2022)
- Without ground truth, how can we identify a reliable estimator of explanation quality? **MetaQuantus** (TMLR, 2023)
- How can we theoretically correct the widely acknowledged evaluation method "Sanity checks" for higher estimator reliability? **Sanity Checks Revisited** (NeurIPS XAIA, 2023)

(B) Foundational Interpretability

- Can stochastic adjustments to model weights enhance the quality of explanation functions, locally and globally? **NoiseGrad** (AAAI, 2021)

(C) Applied Research

- How can SOTA evaluation techniques be used in climate science for more robust method selection? **Finding-right-XAI-method** (ICLR CCAI, 2023, *spotlight); (EGU, 2023)
- How can we quantify the model's domain adaptation capabilities for grassland monitoring using XAI? **In Review** (GIL, 2024)

#reproducibility, #open-science, #downstream-societal-impact

Understandable Machine Intelligence Lab

Lecture 1

Foundations

Foundations — Scope

Explainable AI

1. **Motivation & Definition** — Why do we need interpretability; arguments
2. **Methods** — What are the current methods; local and global
3. **Failure Modes** — Review the limitations of techniques; a critical view
4. **Summary**

Foundations

Motivation & Definition

Motivation — 1st Argument

Establish Trust and Verify Outcomes

- Train a binary classifier to distinguish between images

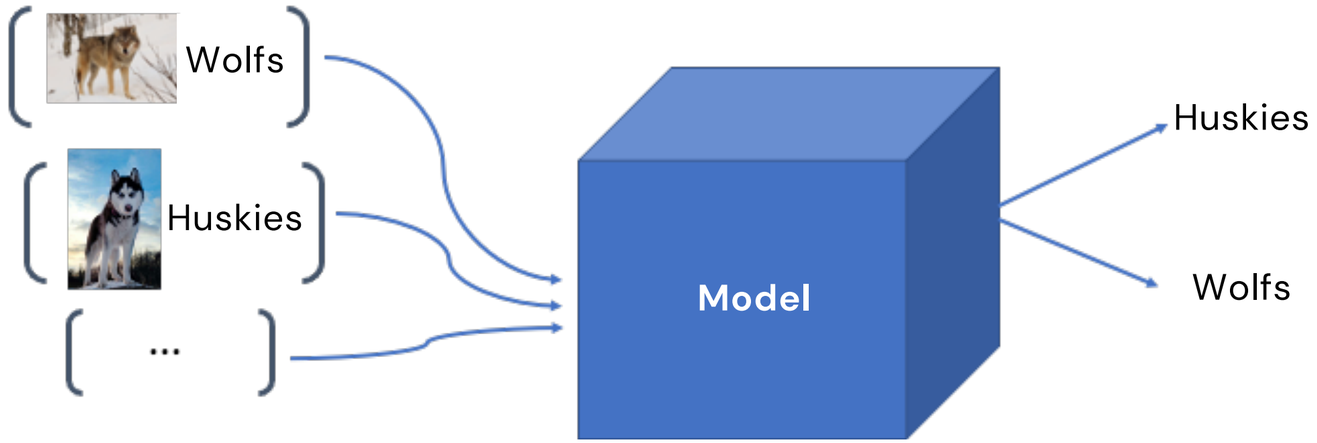
Wolfs

Huskies

Motivation — 1st Argument

Establish Trust and Verify Outcomes

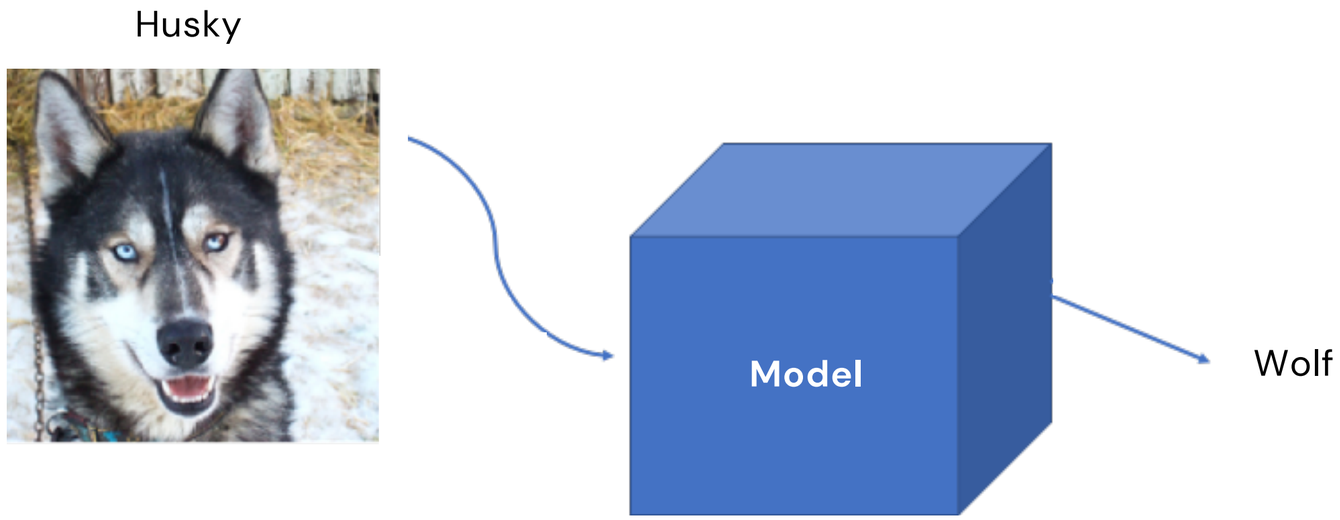
- Model learns to discriminate between the classes to close to perfect accuracy



Motivation — 1st Argument

Establish Trust and Verify Outcomes

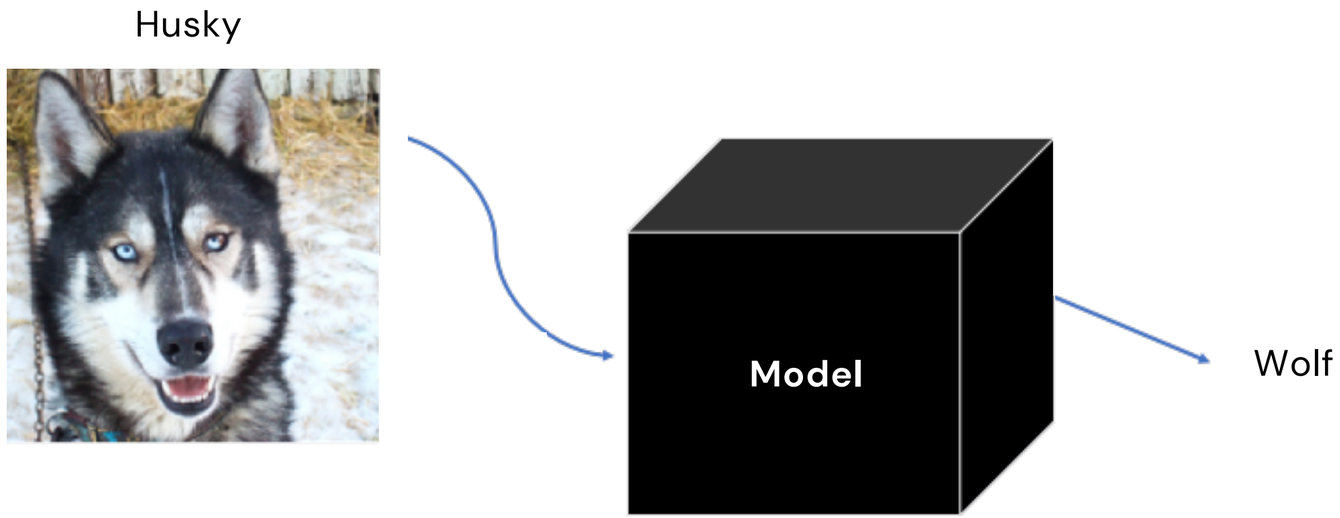
- Prior deployment, to test how the model works in practice, we run a test sample



Motivation — 1st Argument

Establish Trust and Verify Outcomes

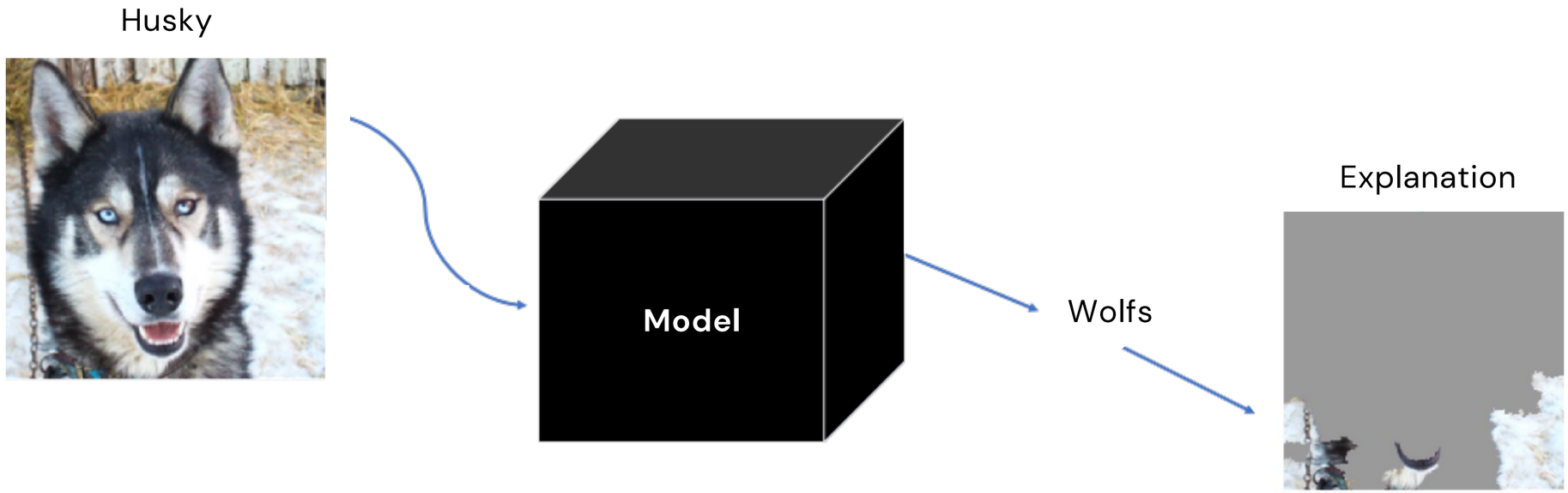
- Why wolf? The model appears as a black box



Motivation — 1st Argument

Establish Trust and Verify Outcomes

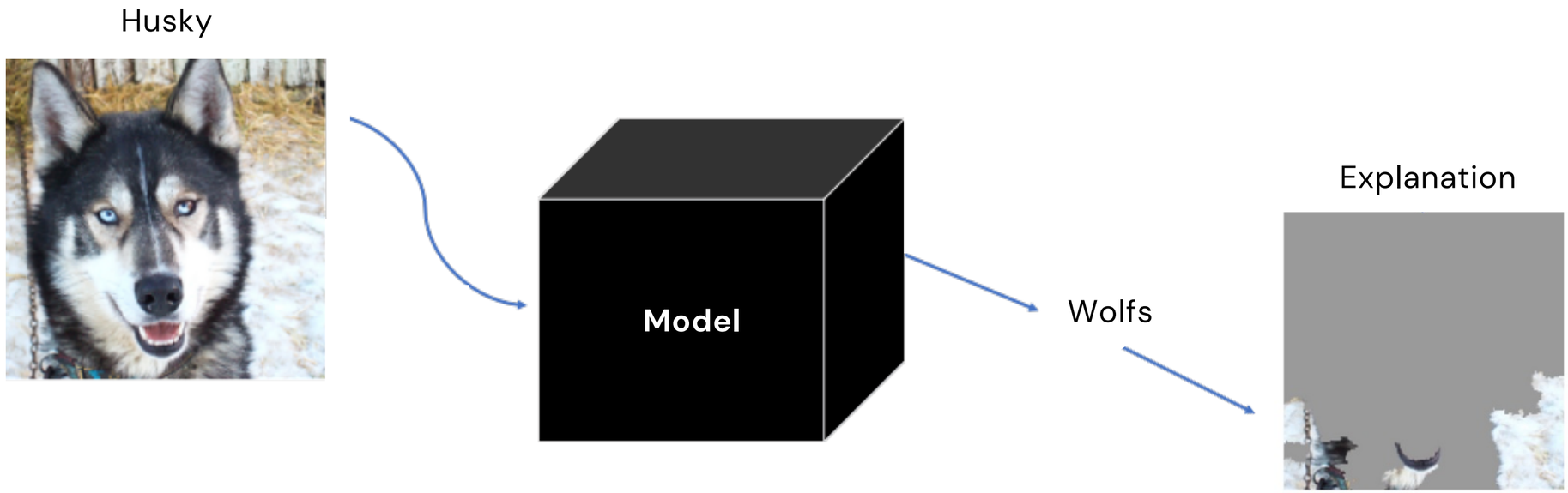
- We apply *local* explanation methods to find important features (i.e., pixels) and find "white snow"



Motivation — 1st Argument

Establish Trust and Verify Outcomes

- We apply *local* explanation methods to find important features (i.e., pixels) and find "white snow"

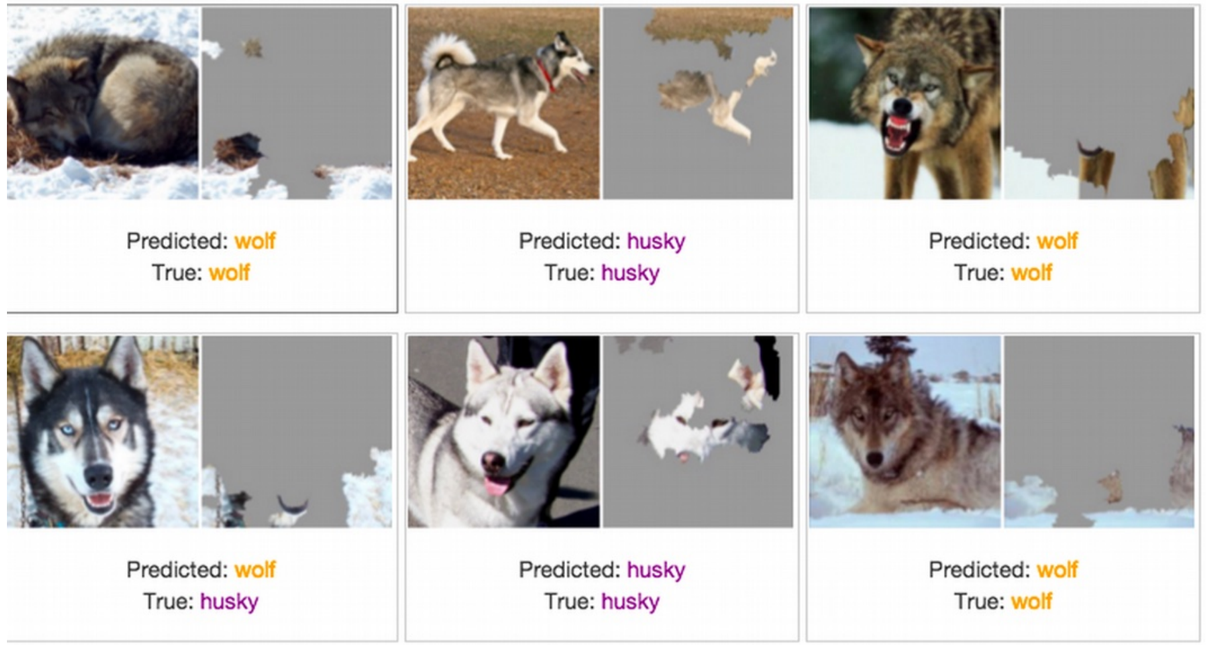


Will ``minimising error'' be a guarantee that the model works well in practice?

Motivation — 1st Argument

Establish Trust and Verify Outcomes

- Not necessarily — explanations shed light to actual problem-solving behaviour in test envs



The model has learned to use "snow" when predicting "wolf"

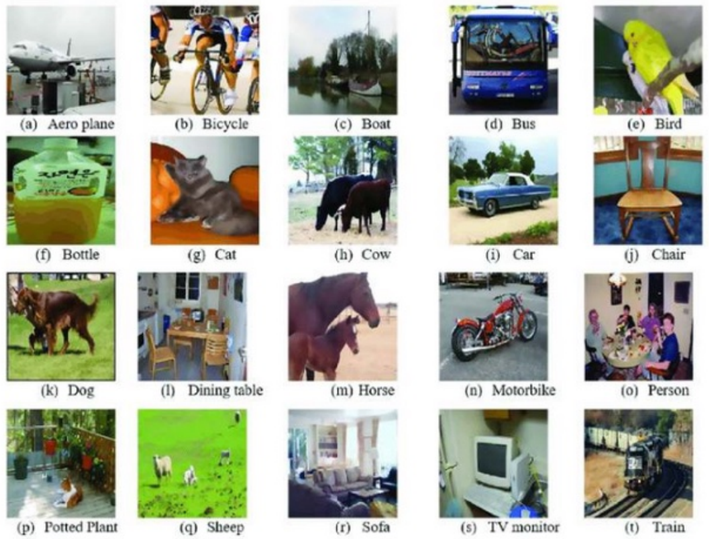
Wrong prediction

Image source.

Motivation — 2nd Argument

Debug Model Artefacts

- PASCAL-VOC Challenge — multi-label classification, with increasing accuracy over the years



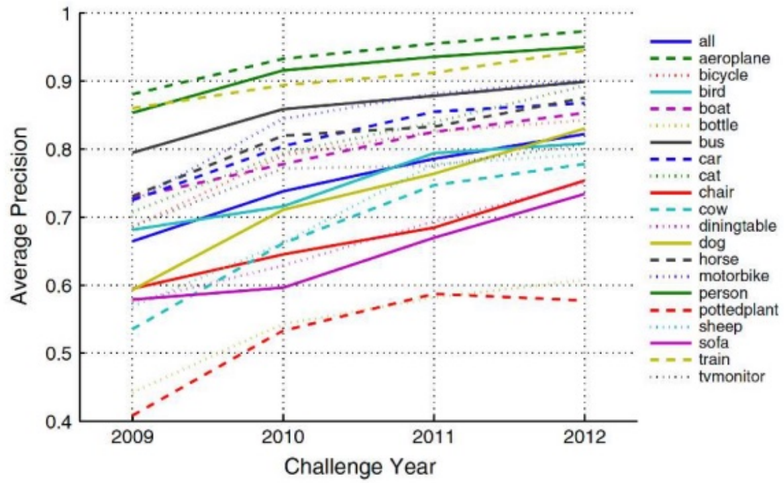
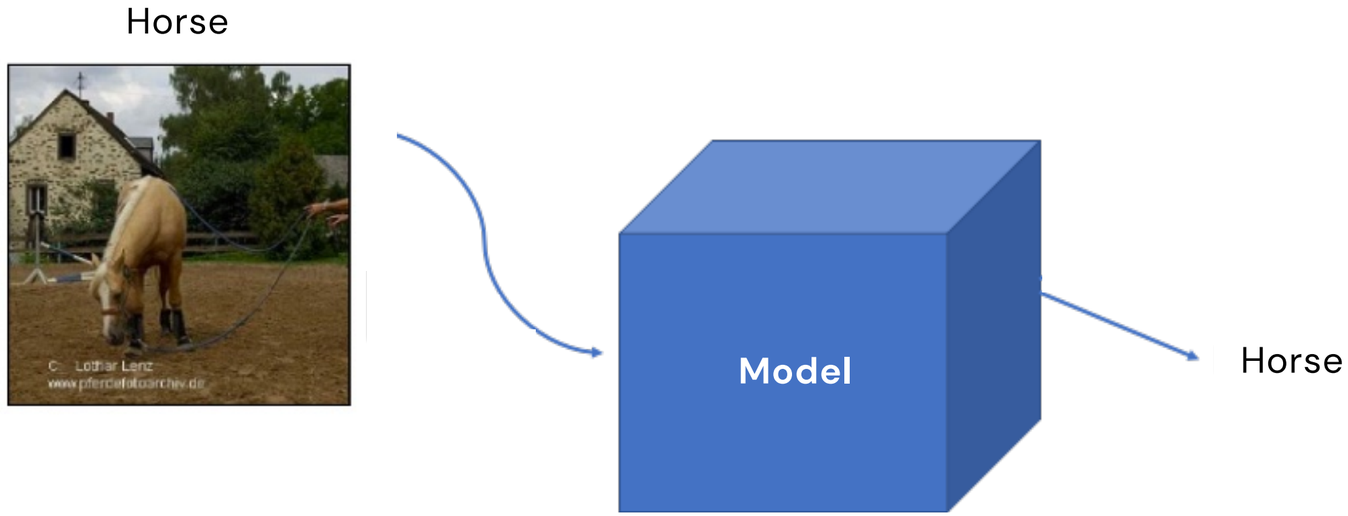


Image source.

Motivation — 2nd Argument

Debug Model Artefacts

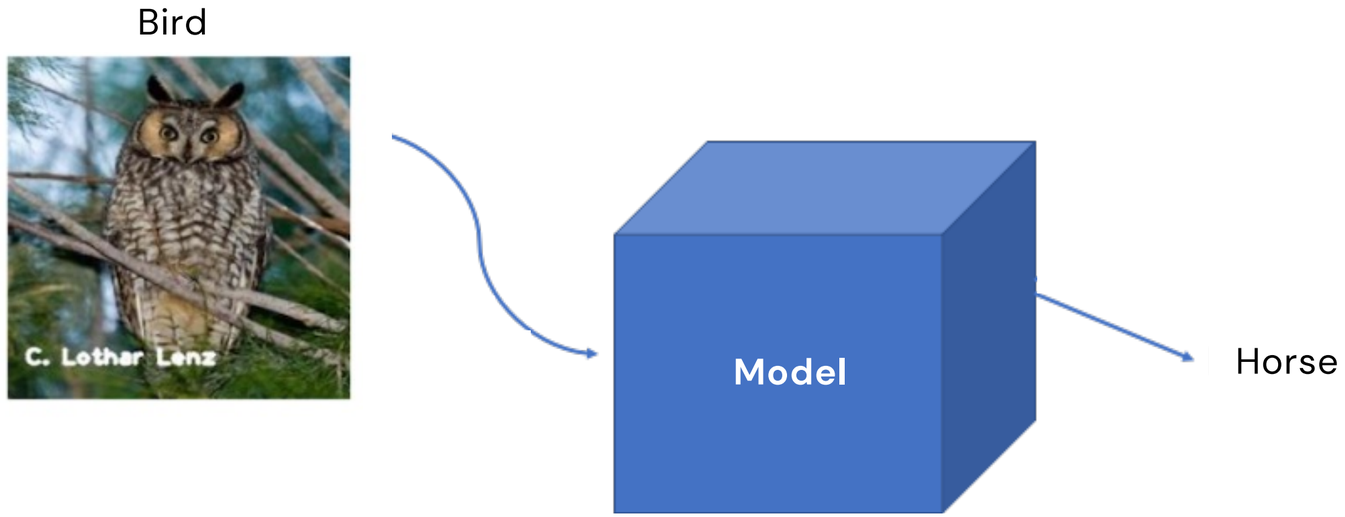
- When providing an input "horse", an expected prediction follows



Motivation — 2nd Argument

Debug Model Artefacts

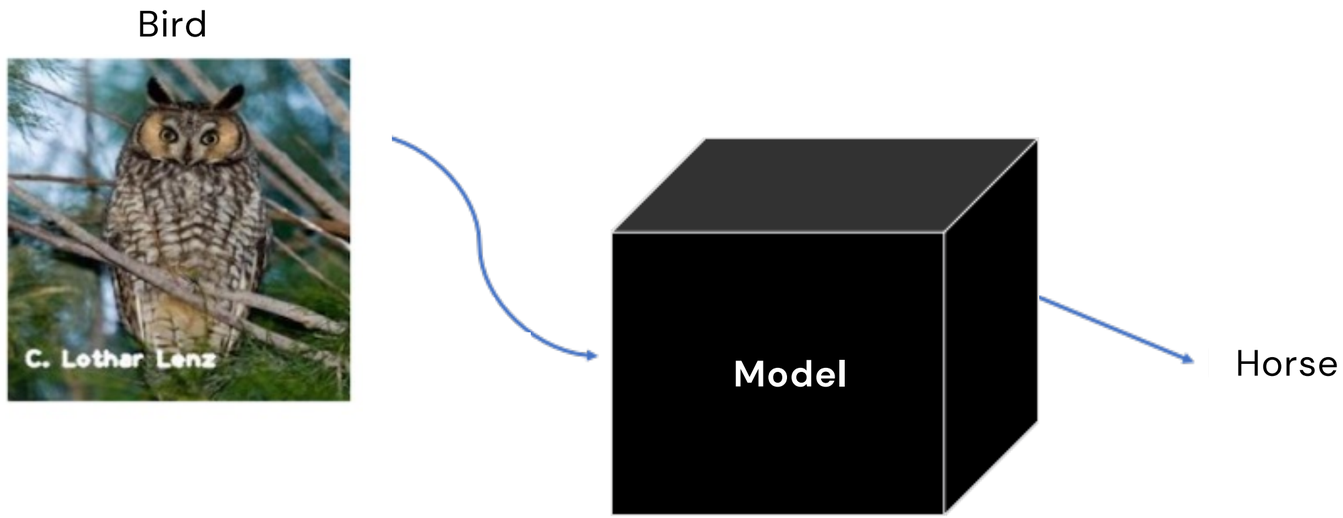
- When providing an input "bird", an ***unexpected*** prediction follows



Motivation — 2nd Argument

Debug Model Artefacts

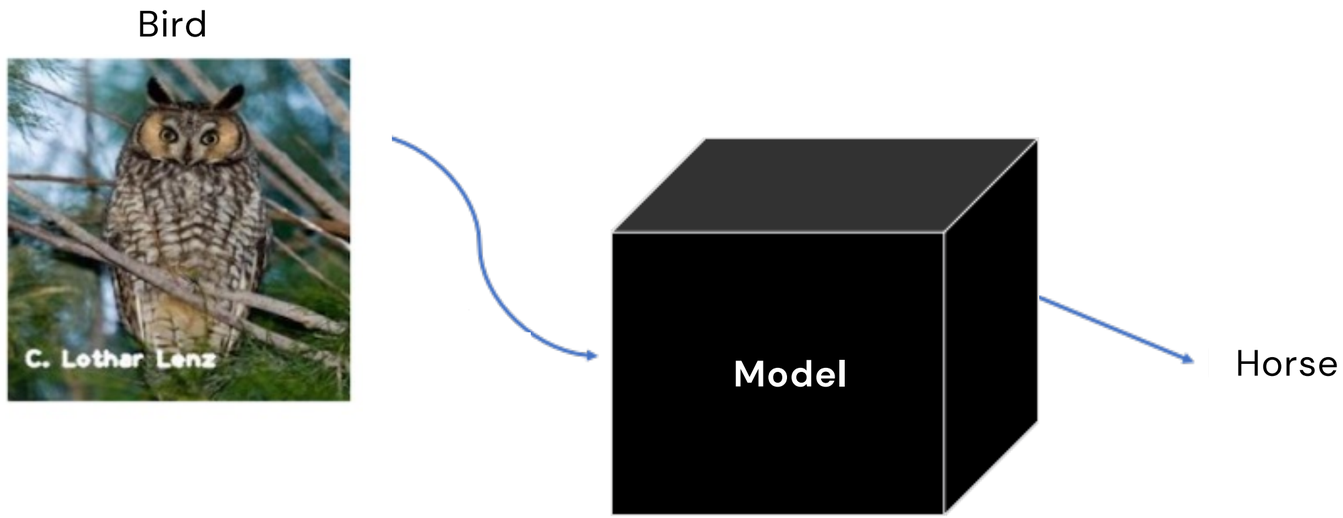
- Our model acts as a black box



Motivation — 2nd Argument

Debug Model Artefacts

- Our model acts as a black box

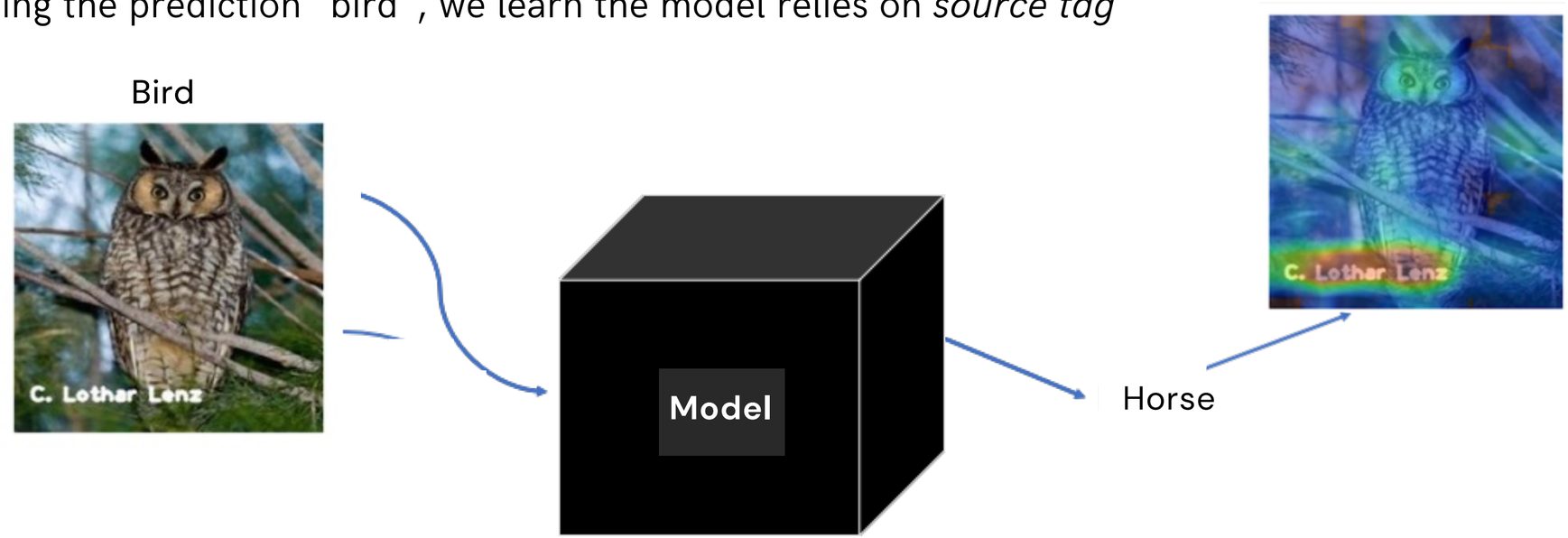


How can we debug this?

Motivation — 2nd Argument

Debug Model Artefacts

- By explaining the prediction "bird", we learn the model relies on *source tag*



Motivation — 2nd Argument

Debug Model Artefacts

- Lapuschkin et al., 2016; 2019 use explanation methods to discover “Clever Hans” (model artefacts)

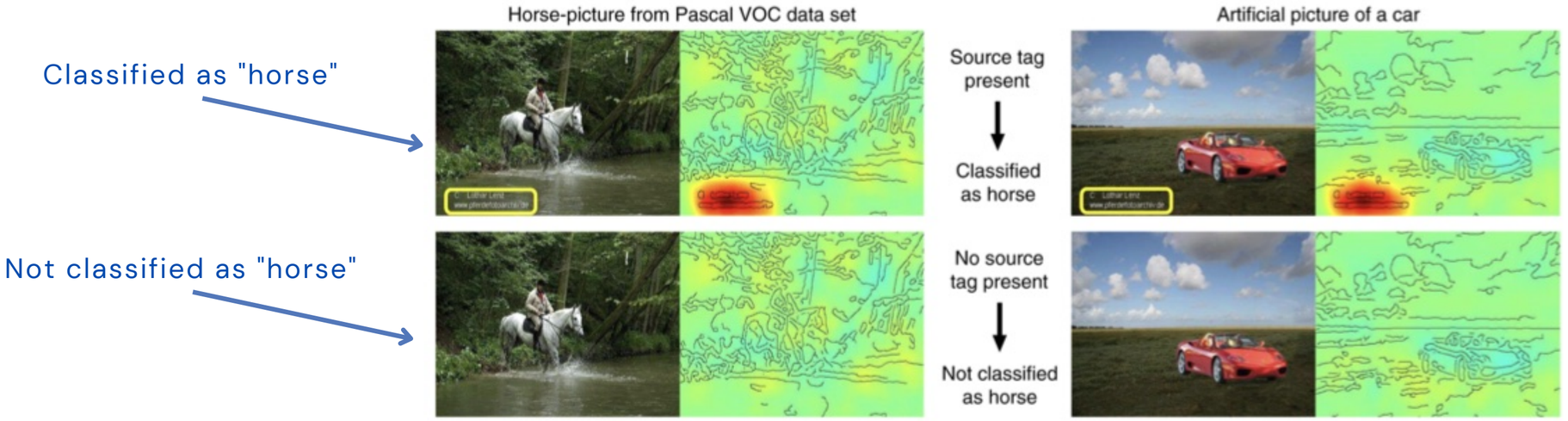


Image source.

Motivation — 2nd Argument

Debug Model Artefacts

- Clever Hans was a horse (1907) claimed to have performed arithmetic and other intellectual tasks

But

- In reality, CH was responding directly to involuntary cues in the body language of the human trainer (who was entirely unaware)
- Debug when wrong reasons are given for right answers

Image source.

Motivation — 2nd Argument

Debug Model Artefacts

- Clever Hans was a horse (1907) claimed to have performed arithmetic and other intellectual tasks

But

- In reality, CH was responding directly to involuntary cues in the body language of the human trainer (who was entirely unaware)

- Debug when wrong reasons are given for right answers

Image source.

Motivation — 2nd Argument

Debug Model Artefacts

- Clever Hans was a horse (1907) claimed to have performed arithmetic and other intellectual tasks

But

- In reality, CH was responding directly to involuntary cues in the body language of the human trainer (who was entirely unaware)

- Debug when wrong reasons are given for right answers

Image source.

Motivation — 3rd Argument

Assert Regulatory Compliance

- The General Data Protection Regulation laws in EU state that all people have the right to “meaningful information about the logic behind automated decisions using their data”

"right to an explanation"

- Necessary in high-state and in avoiding discrimination/ bias e.g., for doctors, lawmakers, policymakers, banks

Motivation — 3rd Argument

Assert Regulatory Compliance

- The General Data Protection Regulation laws in EU state that all people have the right to “meaningful information about the logic behind automated decisions using their data”

"right to an explanation"

- Necessary in high-state and in avoiding discrimination/ bias e.g., for doctors, lawmakers, policymakers, banks

Motivation — 4th Argument

Acquire Knowledge and Novel Insights

- Arcadu et al., 2019 identified novel features of diabetic retinopathy progression*

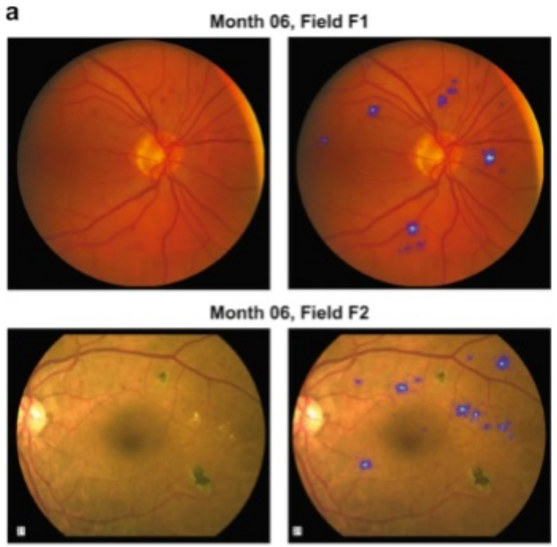


Image Source.
(left) image (right) explanation

*a complication of diabetes, that damages the back of the eye

Motivation — 4th Argument

Acquire Knowledge and Novel Insights

- Schut et al., 2023 extract new chess concepts in AlphaZero and teach them to ``grand masters''

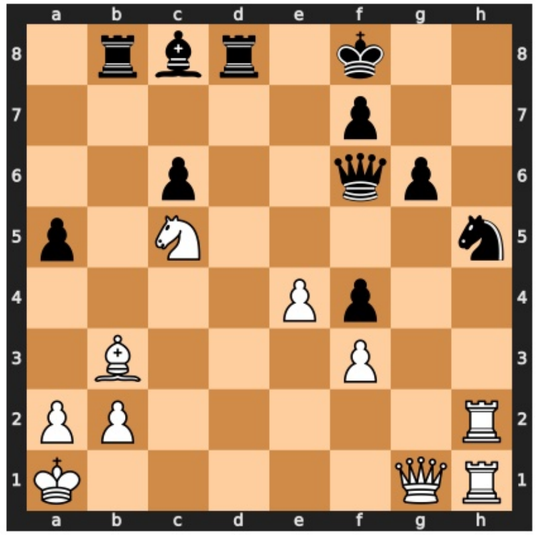


Image Source.

Motivation — Summary

Review Four Main Arguments

(1) **Establish trust** and verify outcomes in test envs

(2) **Debug model artefacts**, uncover and fix learned concepts

(3) **Assert regulatory compliance**, especially in high-stake/ high-liability domains

(4) **Acquire knowledge**, generating novel scientific insights

What did these drivers result in?

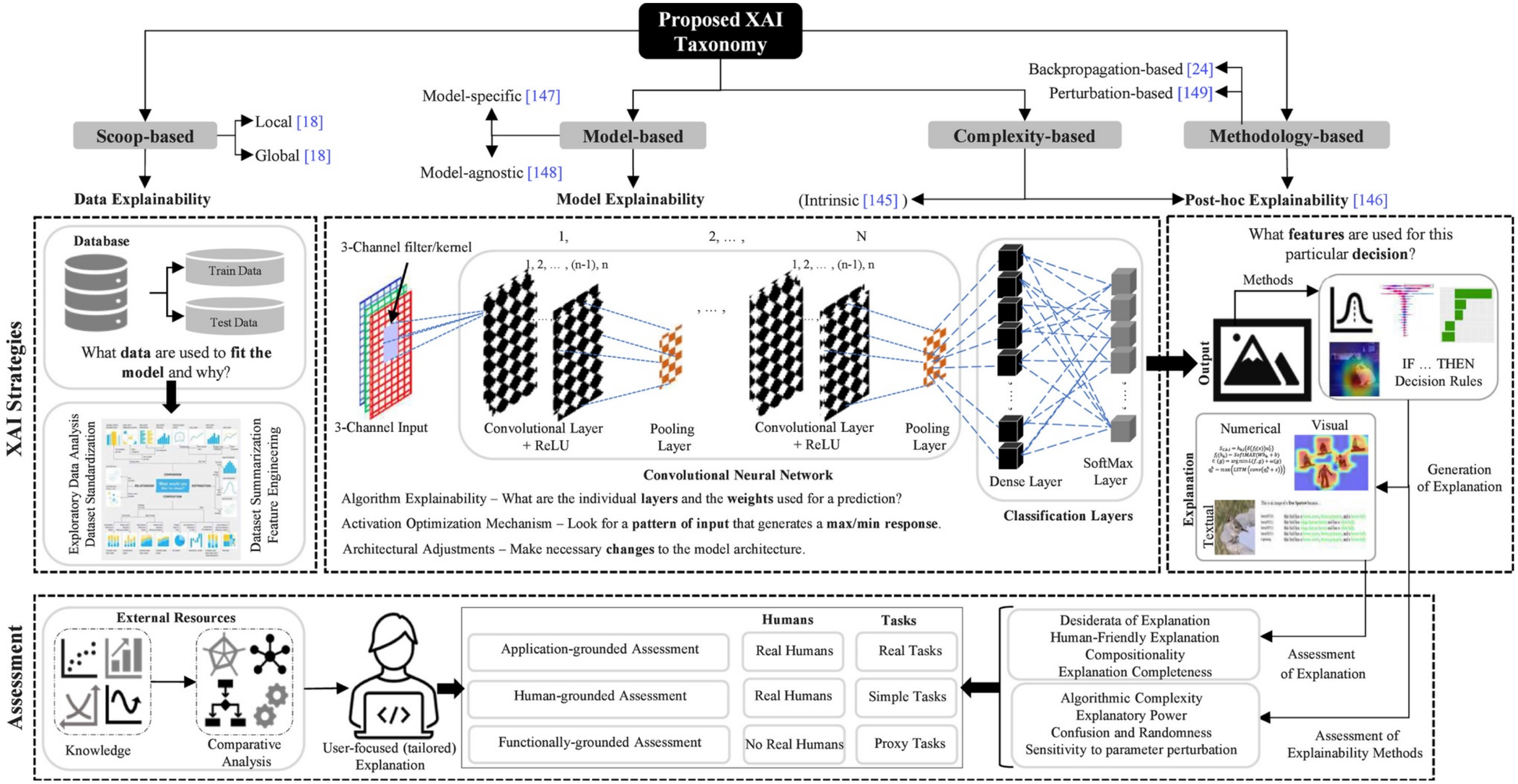


Image Source.

Motivation — Immense Activity

Counting Works by Different Dimensions

- Diverse methods catering different tasks, models, datasets and users

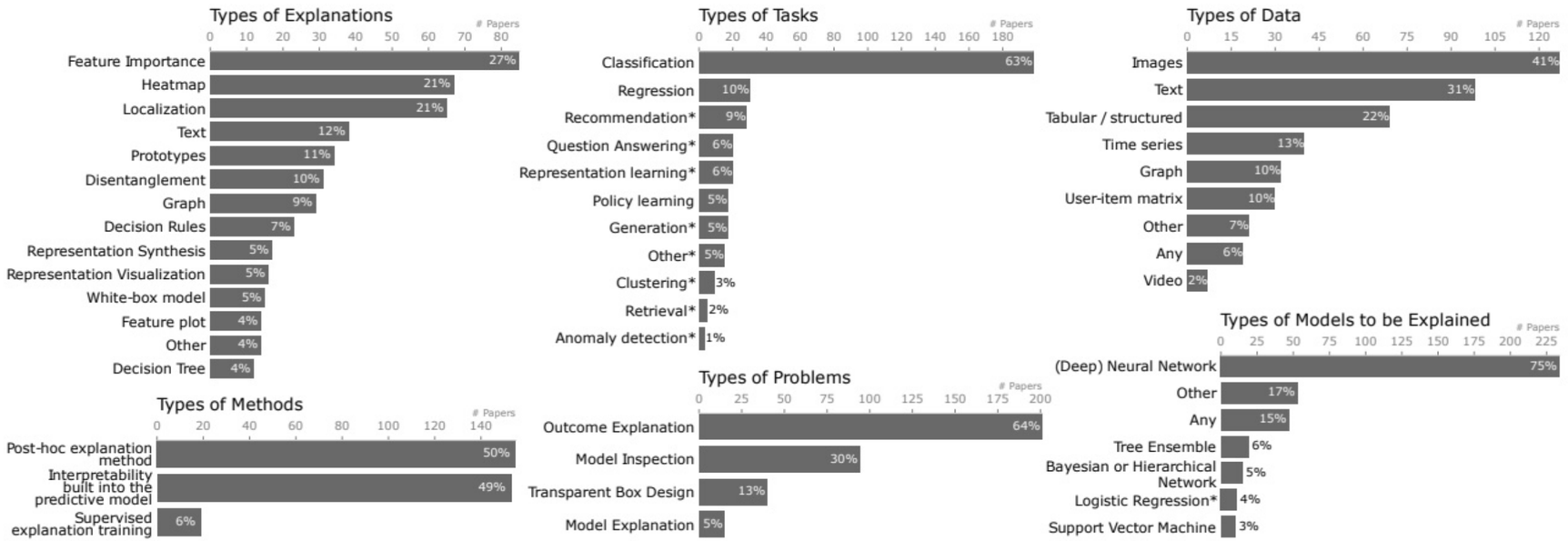


Image Source.

Motivation — Varying Terminology

Naming Conventions and SubFields

- Disjoint communities and varying terminology

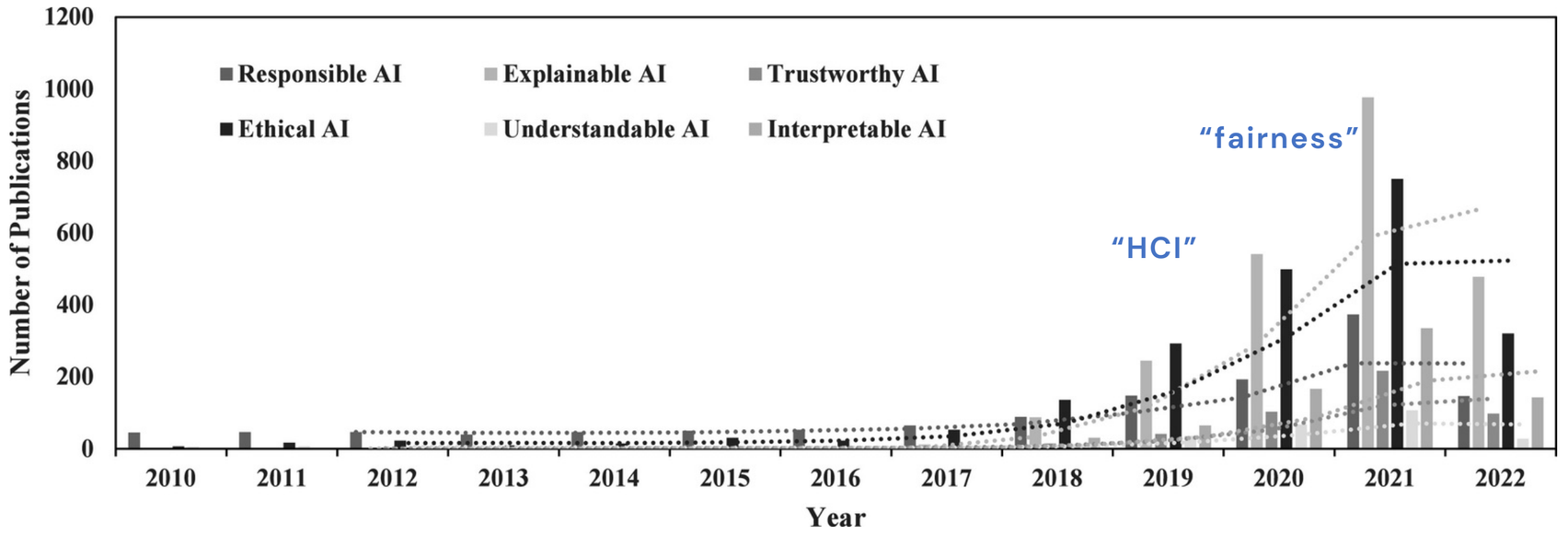


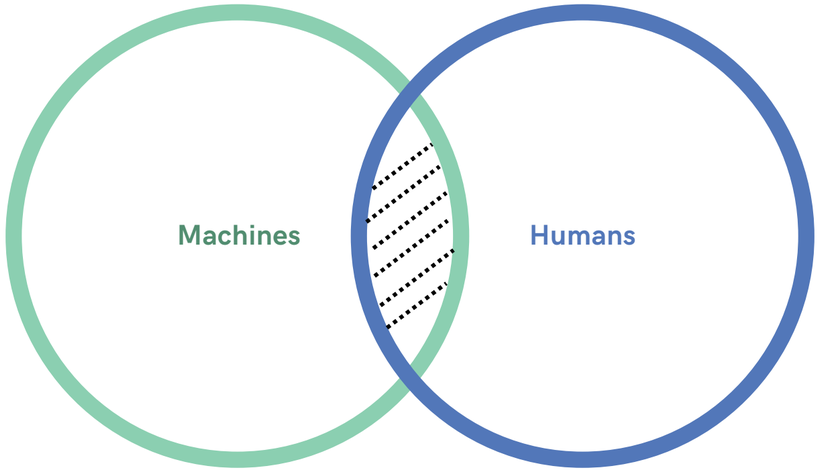
Image Source.

Defining Explainability

Definition — Explore M ∩ H

Representation Space Intersection

- The goal of explainability is to find the intersection between representation spaces

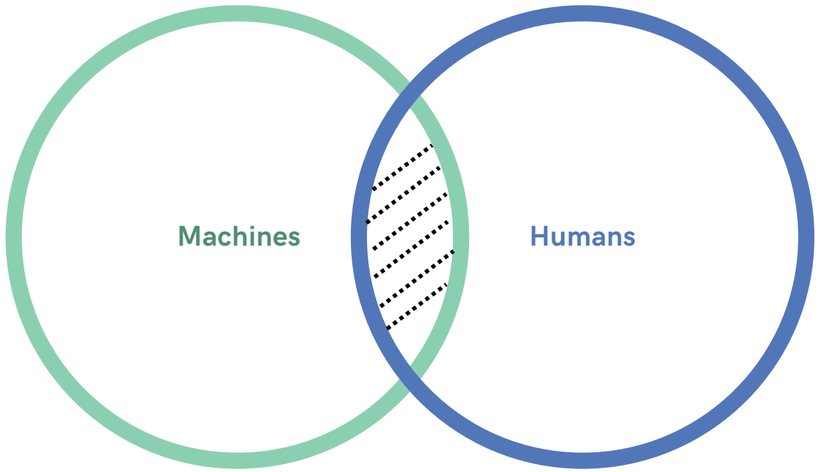


Credit.

Definition — Explore M ∩ H

Representation Space Intersection

- Possibly, this intersection hold conflicting objectives



Who is reciever?
Developers, End-user, Owner

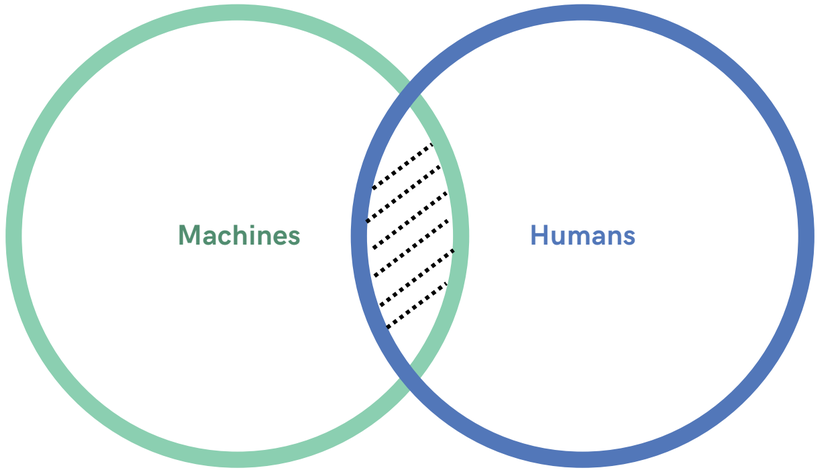
Credit.

Definition — Explore M ∩ H

Representation Space Intersection

- Possibly, this intersection hold conflicting objectives

What unit to explain?
Local (y_pred), Global (neuron)



Who is reciever?
Developers, End-user, Owner

Credit.

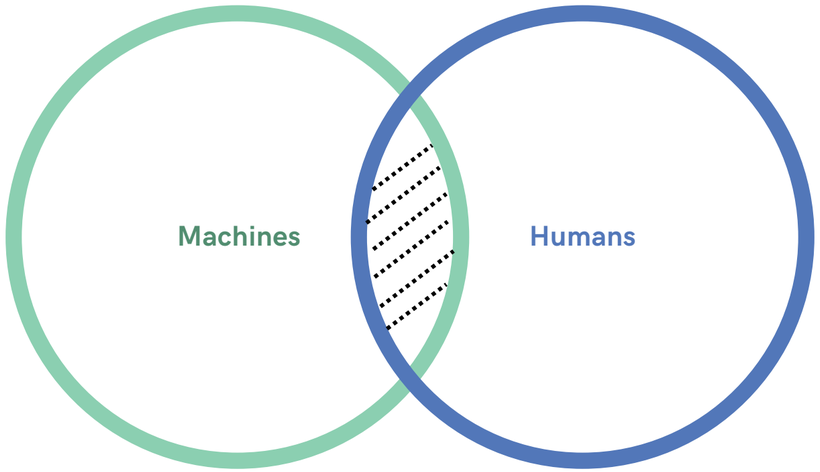
Definition — Explore M ∩ H

Representation Space Intersection

- Possibly, this intersection hold conflicting objectives

What unit to explain?
Local (y_pred), Global (neuron)

What unit to attribute to?
Input features, Latent spaces, Concepts, Data



Who is reciever?
Developers, End-user, Owner

Credit.

Definition — Explore M ∩ H

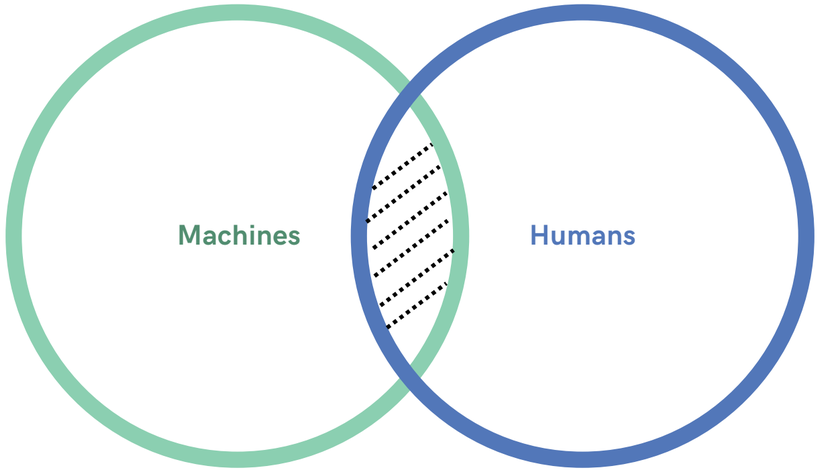
Representation Space Intersection

- Possibly, this intersection hold conflicting objectives

What unit to explain?
Local (y_pred), Global (neuron)

What unit to attribute to?
Input features, Latent spaces, Concepts, Data

When to explain?
A priori, In-training, Post-hoc



Who is reciever?
Developers, End-user, Owner

Credit.

Definition — Explore M ∩ H

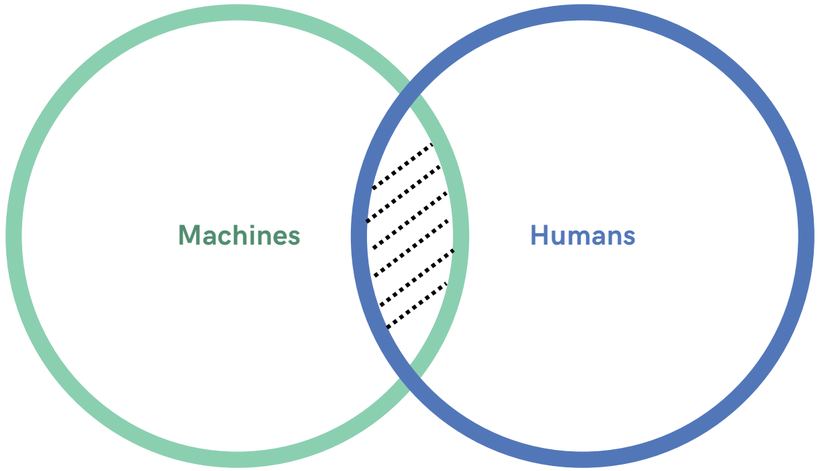
Representation Space Intersection

- Possibly, this intersection hold conflicting objectives

What unit to explain?
Local (y_pred), Global (neuron)

What unit to attribute to?
Input features, Latent spaces, Concepts, Data

When to explain?
A priori, In-training, Post-hoc



Who is reciever?
Developers, End-user, Owner

What medium to use?
Heatmaps, Decision Trees, Chain-of-Thought (LLM), ...

Credit.

Definition — Explore M ∩ H

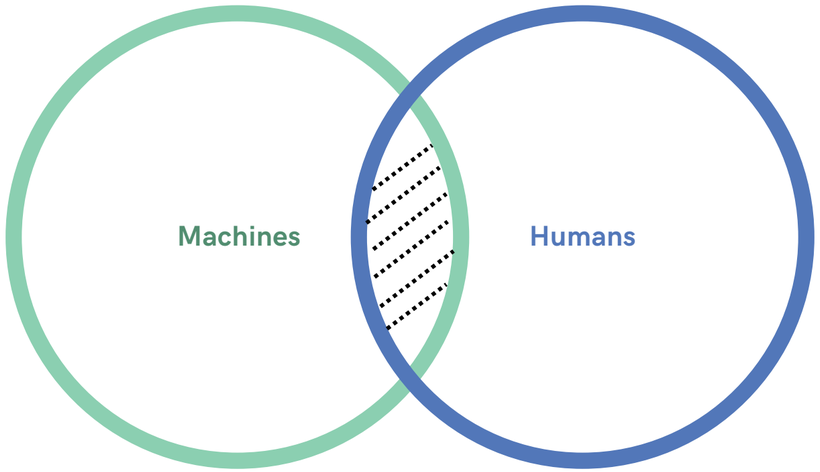
Representation Space Intersection

- Possibly, this intersection hold conflicting objectives

What unit to explain?
Local (y_pred), Global (neuron)

What unit to attribute to?
Input features, Latent spaces, Concepts, Data

When to explain?
A priori, In-training, Post-hoc



Who is reciever?
Developers, End-user, Owner

What medium to use?
Heatmaps, Decision Trees, Chain-of-Thought (LLM), ...

Let’s explore this intersection.

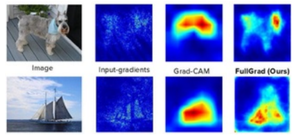
Foundations

Methods

How to best categorise these methods?

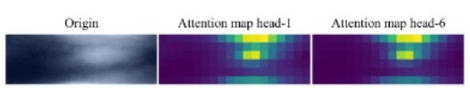
Methods — A Model Perspective

New Architectures Demand New Explanation Methods



Saliency maps

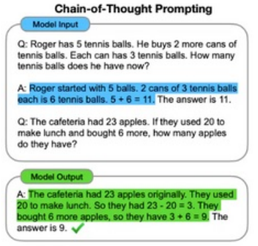
Image Source.



Attention maps as explanations

MLPs *CNNs* *GANs* *Diffusion model* **Transformers** **"Foundation Models"** Multimodality

$$p(x_{1:L}) = \prod_{i=1}^{L} p(x_i \mid x_{1:i-1}).$$



DALL-E 3 Sora

GPT 1 → GPT 2 → GPT 3 →

Instruct GPT → ChatGPT ....

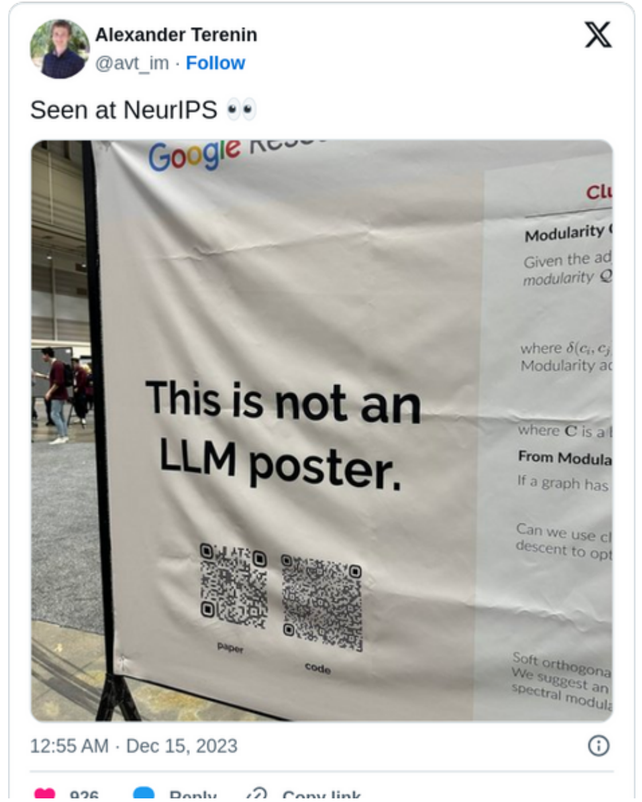
Adversarial ML teaches us:
models are fragile!

Methods — “Ever-Evolving”

New Architectures Demand New Explanation Methods

Trends in explainability:

- Scope: fr. local prediction to global representation
- Timing: fr. static (post-hoc) to automatic (in-training)
- Attribution: fr input space to concepts, data, free-form



Methods — “Ever-Evolving”

New Architectures Demand New Explanation Methods

Trends in explainability:

- Scope: fr. local prediction to global representation
- Timing: fr. static (post-hoc) to automatic (in-training)
- Attribution: fr input space to concepts, data, free-form

New Method Categorisation, 'Classic' or 'Next-Gen'?

Local and Global Methods

Methods — Local and Global Methods

A Gentle Notational Setup

- A model learns to map inputs $x \in \mathbb{R}^D$ to predictions $y \in \mathbb{R}^C$ via parameters θ:

$$f_{\Theta} : X \to Y$$

- **Local explanations** provide attributions to input features of a specific prediction y:

$$\phi L(f, x, y; \lambda) = e$$

- **Global explanations** explain the global behaviour of a neuron n, independent on an input x:

$$\phi G(f, n; \kappa) = e$$

Methods — Local and Global Methods

A Gentle Notational Setup

- A model learns to map inputs $x \in \mathbb{R}^D$ to predictions $y \in \mathbb{R}^C$ via parameters θ:

$$f_\Theta : X \to Y$$

- **Local explanations** provide attributions to input features of a specific prediction y_c:

$$\phi L(f, x, y_c; \lambda) = e$$

- **Global explanations** explain the global behaviour of a neuron $n \in \mathbb{R}$, independent on an input x:

$$\phi G(f, n; \kappa) = e$$

Methods — Local and Global Methods

A Gentle Notational Setup

A variety of approaches:

- *Model-agnostic* (e.g., perturbation-based, local surrogate explainers)
- *Model-aware* (e.g., gradient-, backpropagation-, attention-based)
- *Unit attribution* (input-, latent space, concepts-, counterfactual-)

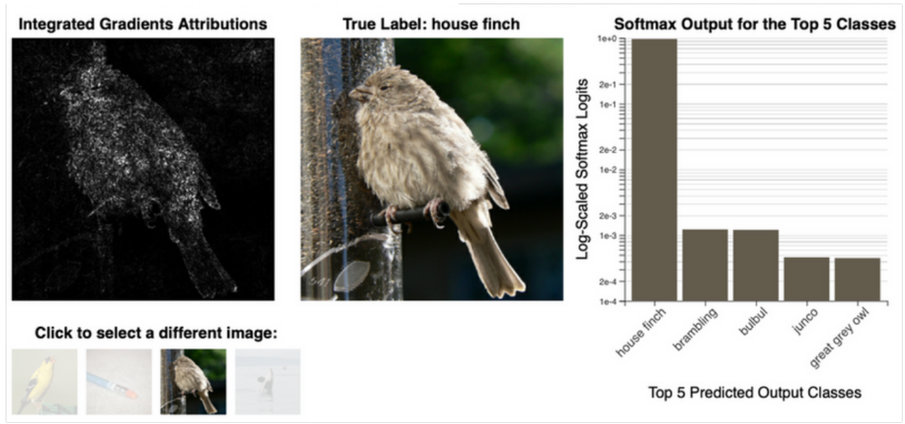


Image Source.

Methods — Local and Global Methods

A Gentle Notational Setup

A variety of approaches:

- *Model-agnostic* (e.g., perturbation-based, local surrogate explainers)
- *Model-aware* (e.g., gradient-, backpropagation-, attention-based)
- *Unit attribution* (input-, latent space, concepts-, counterfactual-)

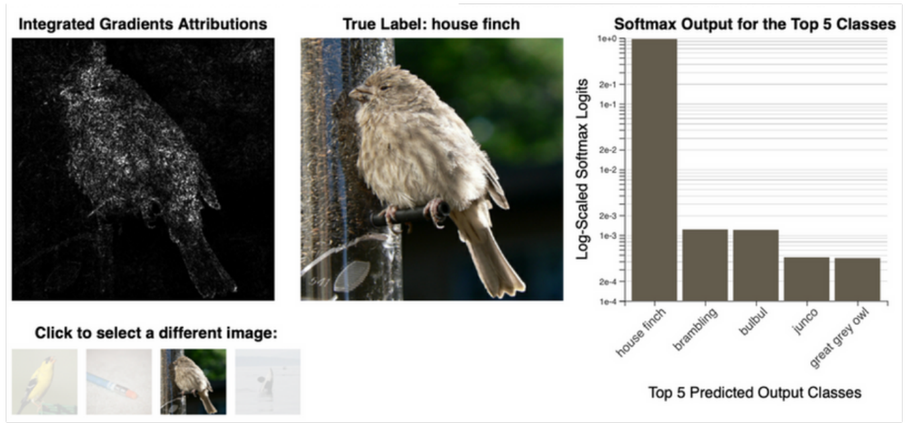


Image Source.

Let’s review some ideas.

Idea #1

Can we use “perturbation” as a way to estimate feature importance?

Methods — Occlusion

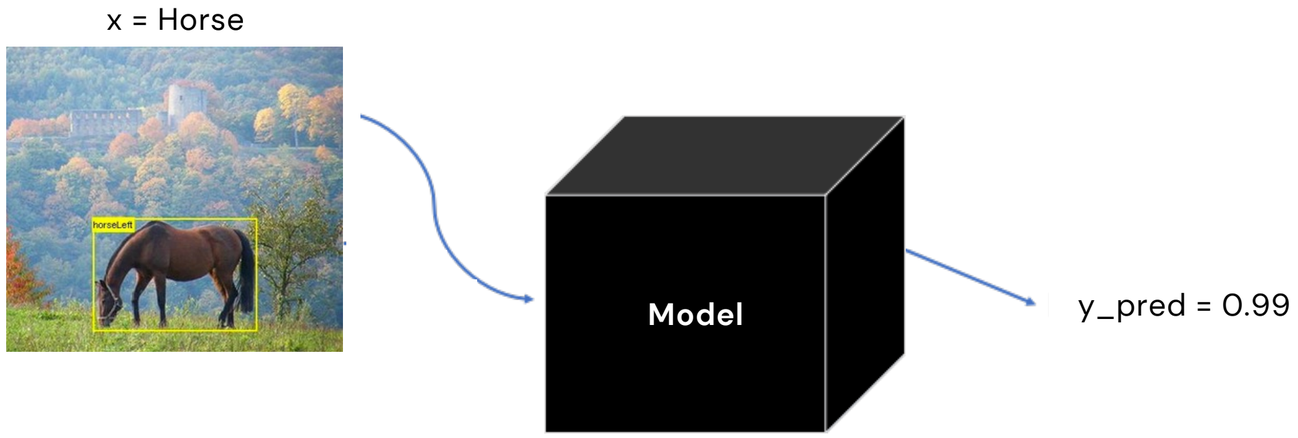
Model-Agnostic Method

- Occlusion (Zeiler et al., 2014) assigns feature importance by monitoring the change in a prediction
  - Feed x through the model and record y
  - Occlude a region (e.g., using a black patch), feed it again and record y'
  - Compute the difference between y and y'
  - Repeat

Methods — Occlusion

Model-Agnostic Method

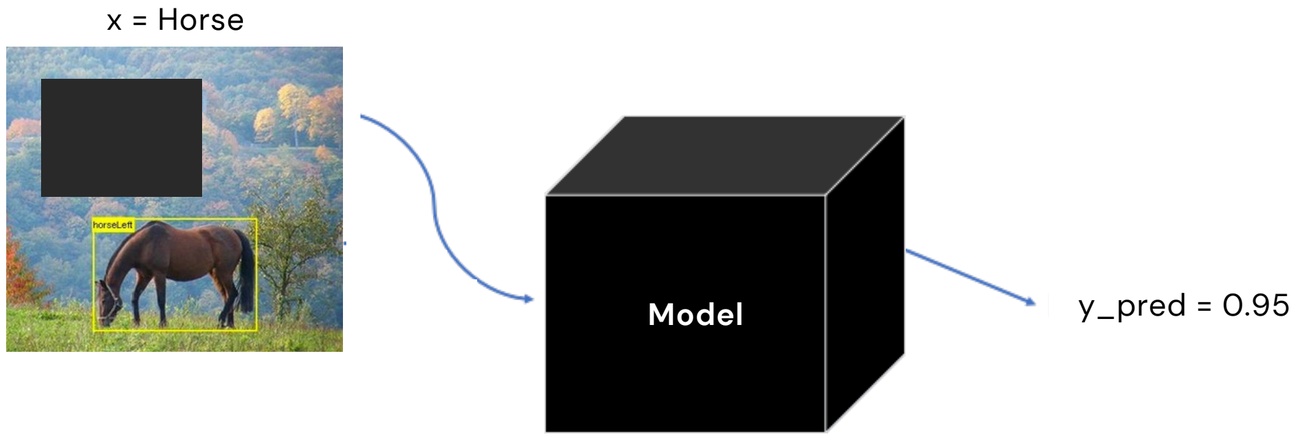
- Occlusion (Zeiler et al., 2014) assigns feature importance by monitoring the change in a prediction
  - Feed x through the model and record y
  - Occlude a region (e.g., using a black patch), feed it again and record y'
  - Compute the difference between y and y'
  - Repeat



Methods — Occlusion

Model-Agnostic Method

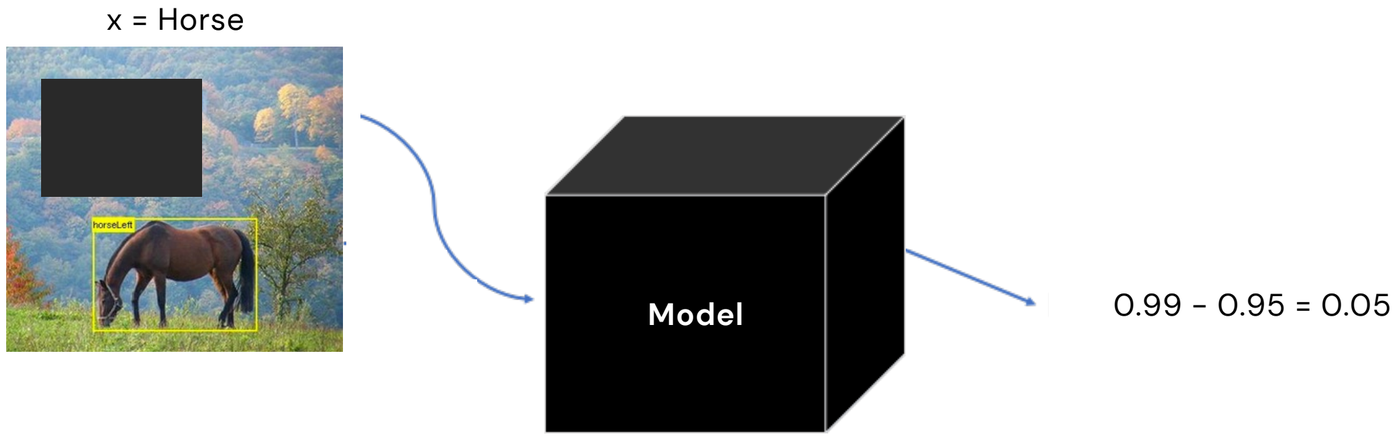
- Occlusion (Zeiler et al., 2014) assigns feature importance by monitoring the change in a prediction
  - Feed x through the model and record y
  - Occlude a region (e.g., using a black patch), feed it again and record y'
  - Compute the difference between y and y'
  - Repeat



Methods — Occlusion

Model-Agnostic Method

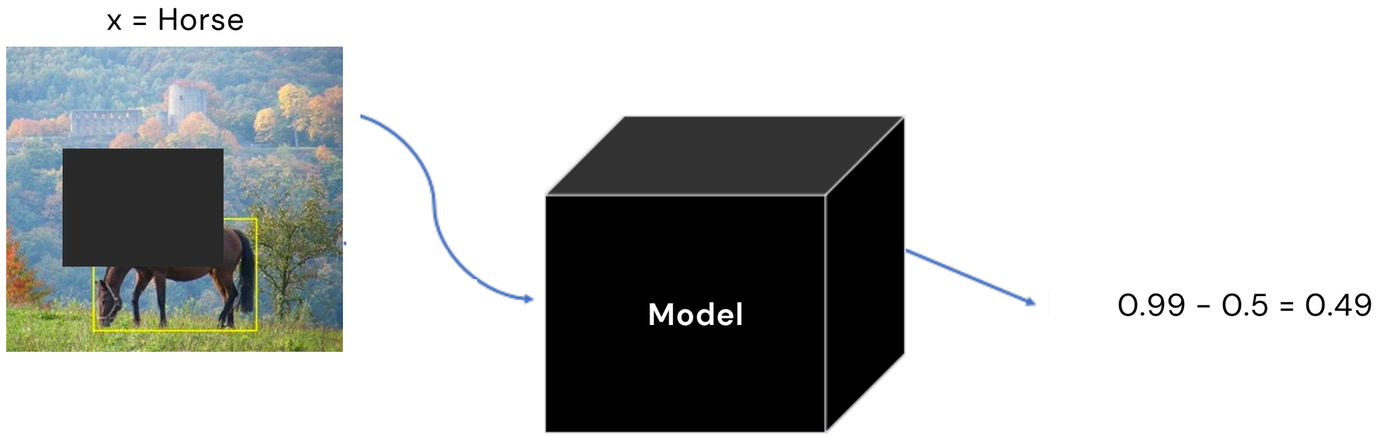
- Occlusion (Zeiler et al., 2014) assigns feature importance by monitoring the change in a prediction
  - Feed x through the model and record y
  - Occlude a region (e.g., using a black patch), feed it again and record y'
  - Compute the difference between y and y'
  - Repeat



Methods — Occlusion

Model-Agnostic Method

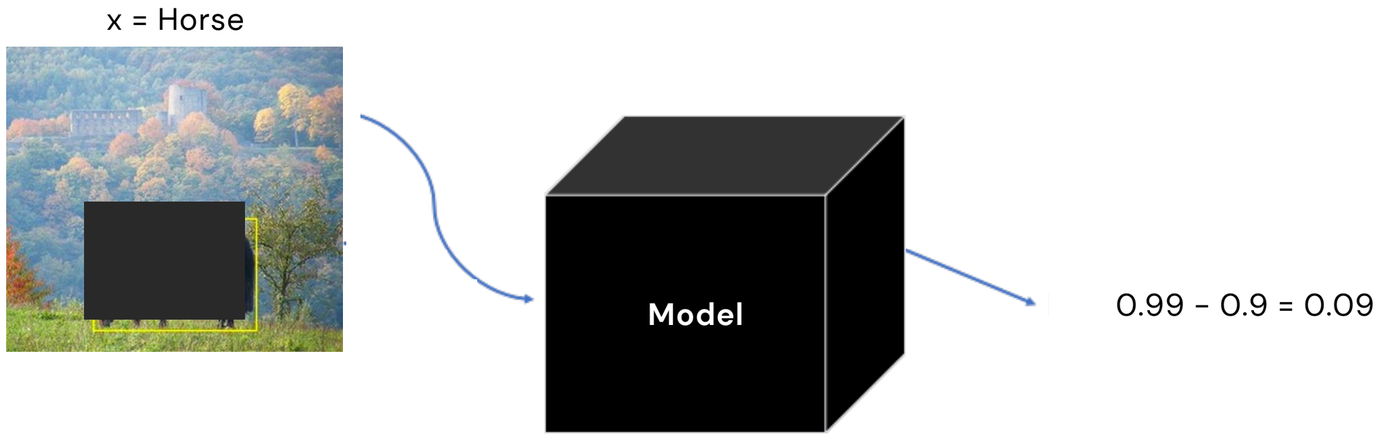
- Occlusion (Zeiler et al., 2014) assigns feature importance by monitoring the change in a prediction
  - Feed x through the model and record y
  - Occlude a region (e.g., using a black patch), feed it again and record y'
  - Compute the difference between y and y'
  - Repeat



Methods — Occlusion

Model-Agnostic Method

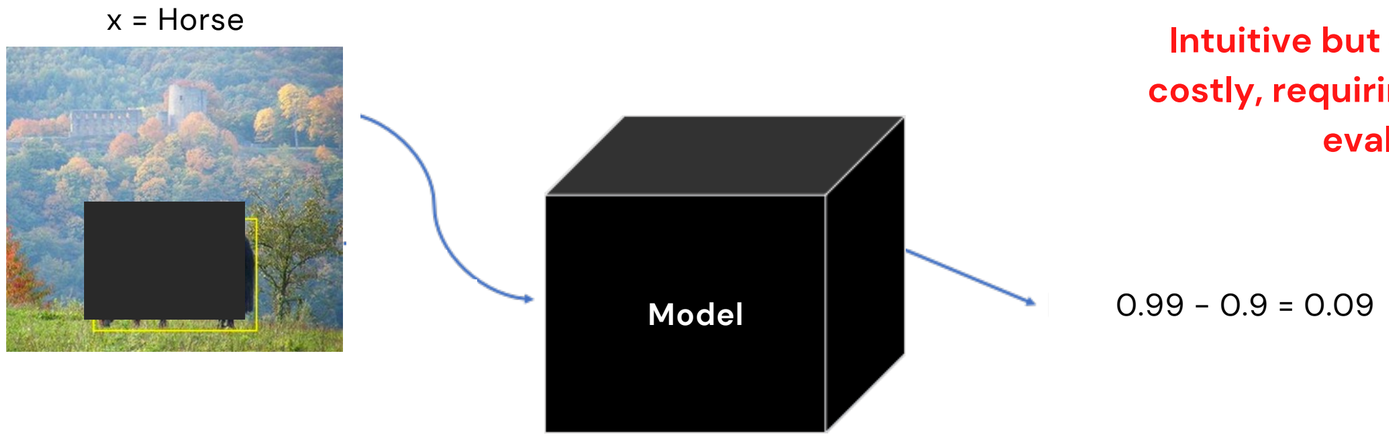
- Occlusion (Zeiler et al., 2014) assigns feature importance by monitoring the change in a prediction
  - Feed x through the model and record y
  - Occlude a region (e.g., using a black patch), feed it again and record y'
  - Compute the difference between y and y'
  - Repeat



Methods — Occlusion

Model-Agnostic Method

- Occlusion (Zeiler et al., 2014) assigns feature importance by monitoring the change in a prediction
  - Feed x through the model and record y
  - Occlude a region (e.g., using a black patch), feed it again and record y'
  - Compute the difference between y and y'
  - Repeat



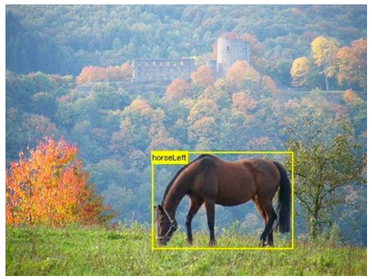
Intuitive but computationally costly, requiring multiple model evaluations

Methods — Prototypes

Model-Agnostic Data Attribution Method

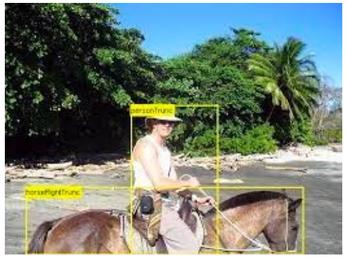
- Find a representative example ("prototype") from the training data that is similar to the input

Explain "Horse"

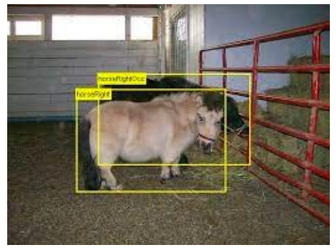


Prototype 1

Prototype 2



Prototype 3

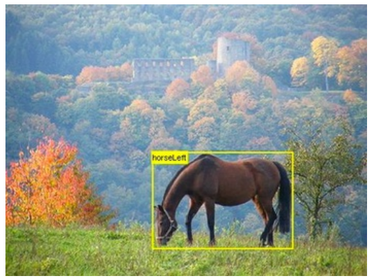


Methods — Prototypes

Model-Agnostic Data Attribution Method

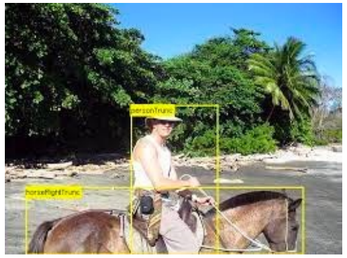
- Find a representative example ("prototype") from the training data that is similar to the input

Explain "Horse"

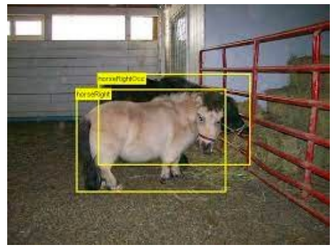


Prototype 1

Prototype 2



Prototype 3



Not representative of f's decision boundary, especially for high-dimensional data

Methods — LIME

Model-Agnostic Method

- A model might be complex globally, but locally it may be easier to understand it
- Idea: Explain a prediction with the help of an interpretable surrogate model (Ribeiro et al., 2016)

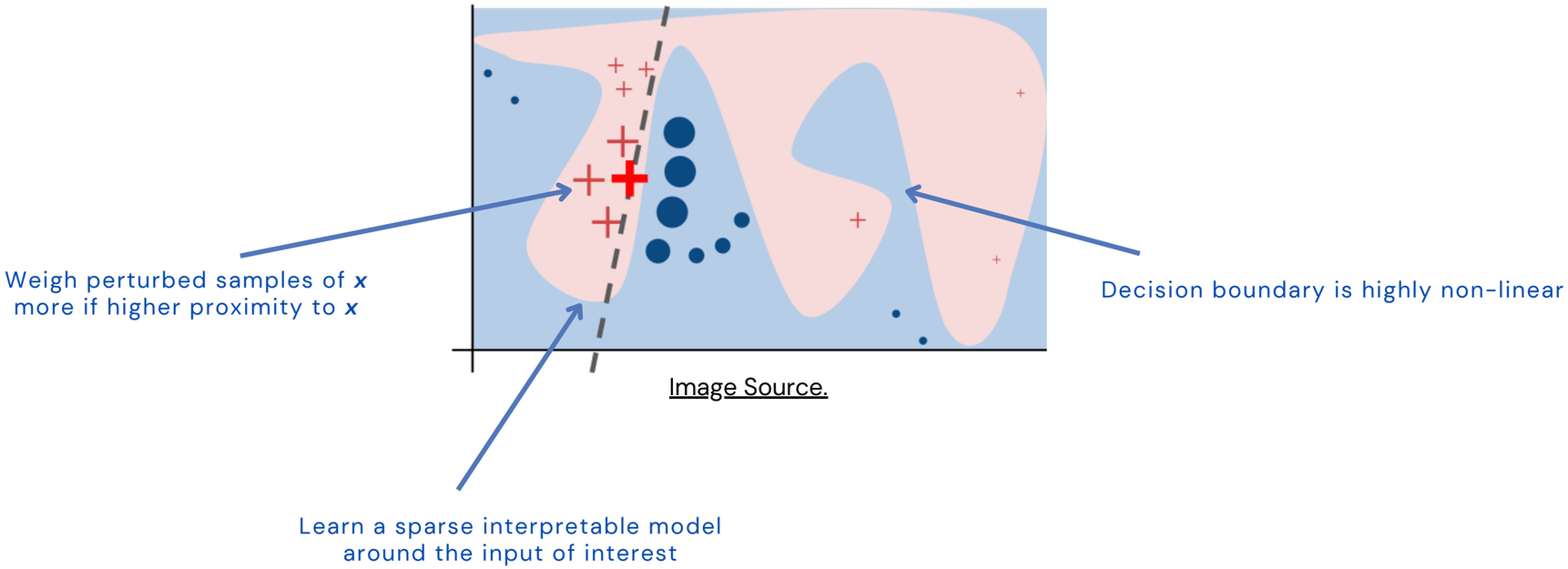


Image Source.

Methods — LIME

Model-Agnostic Method

- A model might be complex globally, but locally it may be easier to understand it
- Idea: Explain a prediction with the help of an interpretable surrogate model (Ribeiro et al., 2016)

$$\arg\min_{g \in G} L(f_c, g, \pi_x) + \Omega(g)$$

Learn a surrogate model *g* by minimising loss on ***x***, ***y*** pairs

Collect ***x***, ***y*** pairs by sampling in the neighbourhood of ***x*** with vicinity denoted ***π*** and then evaluate the network on these points ***f***

Keep model complexity low

- Weight vector of ***g*** is used as the basis for assigning the attributions to the input feature

Methods — LIME

Model-Agnostic Method

- A model might be complex globally, but locally it may be easier to understand it
- Idea: Explain a prediction with the help of an interpretable surrogate model (Ribeiro et al., 2016)

$$\arg\min_{g \in G} L(f_c, g, \pi_x) + \Omega(g)$$

Learn a surrogate model *g* by minimising loss on ***x***, ***y*** pairs

Collect ***x***, ***y*** pairs by sampling in the neighbourhood of ***x*** with vicinity denoted ***π*** and then evaluate the network on these points ***f***

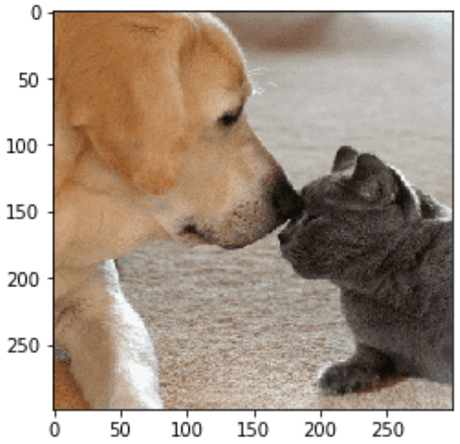
Keep model complexity low

- Weight vector of ***g*** is used as the basis for assigning the attributions to the input feature

Methods — LIME

Model-Agnostic Method

Explain:
"labrador"



Super-pixels as features

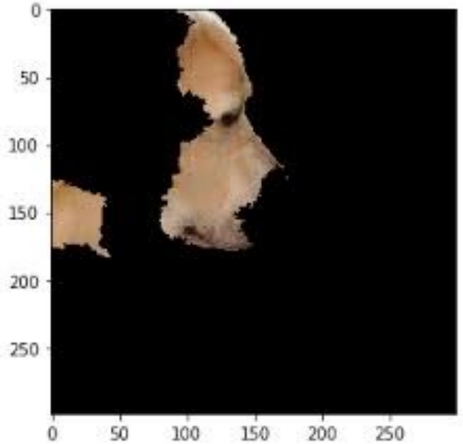


Image classification example

Image Source.

Methods — LIME

Model-Agnostic Method

Explain:
"labrador"

Super-pixels as
features

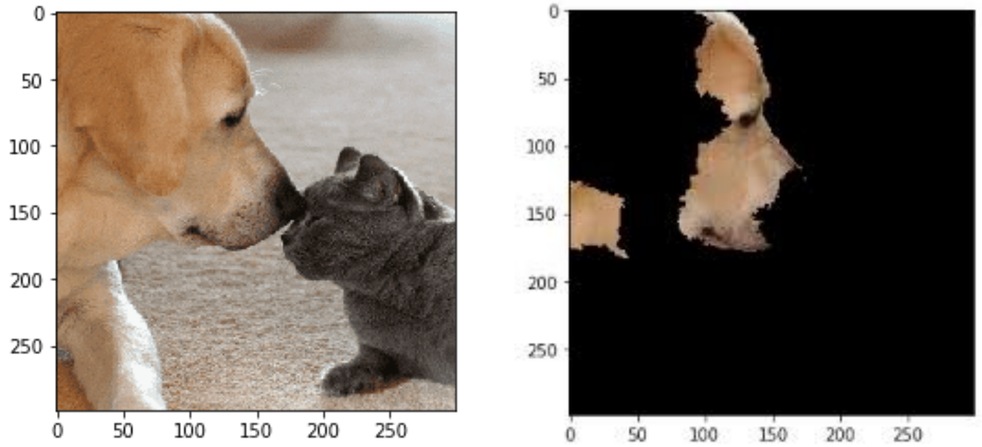


Image Source.

How were these "interpretable" features generated? How should the input be"perturbed"?

Methods — LIME

Model-Agnostic Method

Segmentation mask of "interpretable" super-pixels, generated e.g., by SLIC algorithm

Different perturbations used to collect ***x, y*** pairs for surrogate model

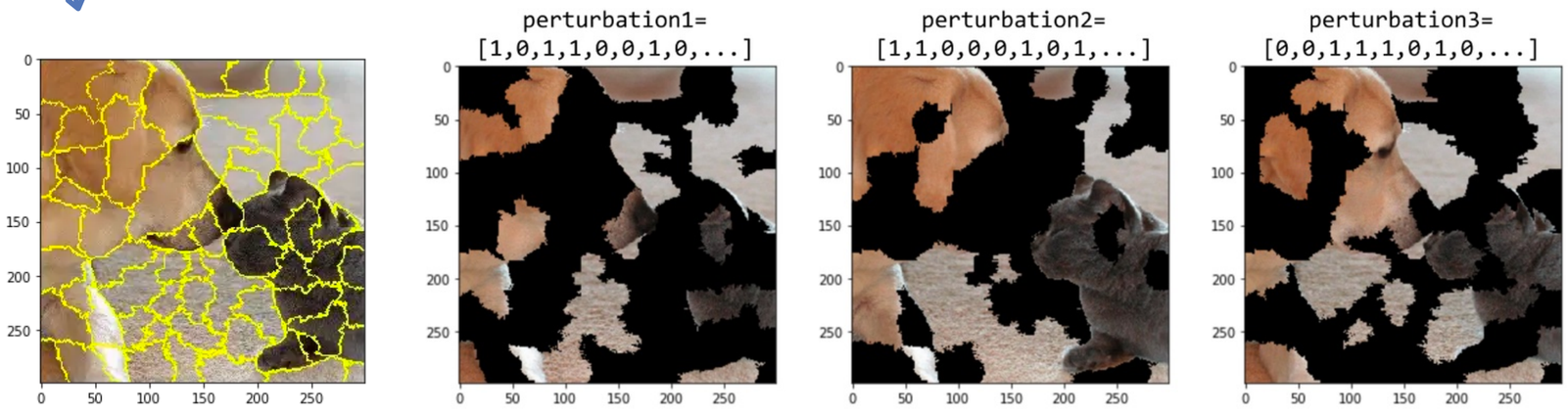


Image Source.
(left) “super-pixels” segmentation
(right) different perturbed samples

Methods — LIME

Model-Agnostic Method

Segmentation mask of "interpretable" super-pixels, generated e.g., by SLIC algorithm

Different perturbations used to collect ***x, y*** pairs for surrogate model

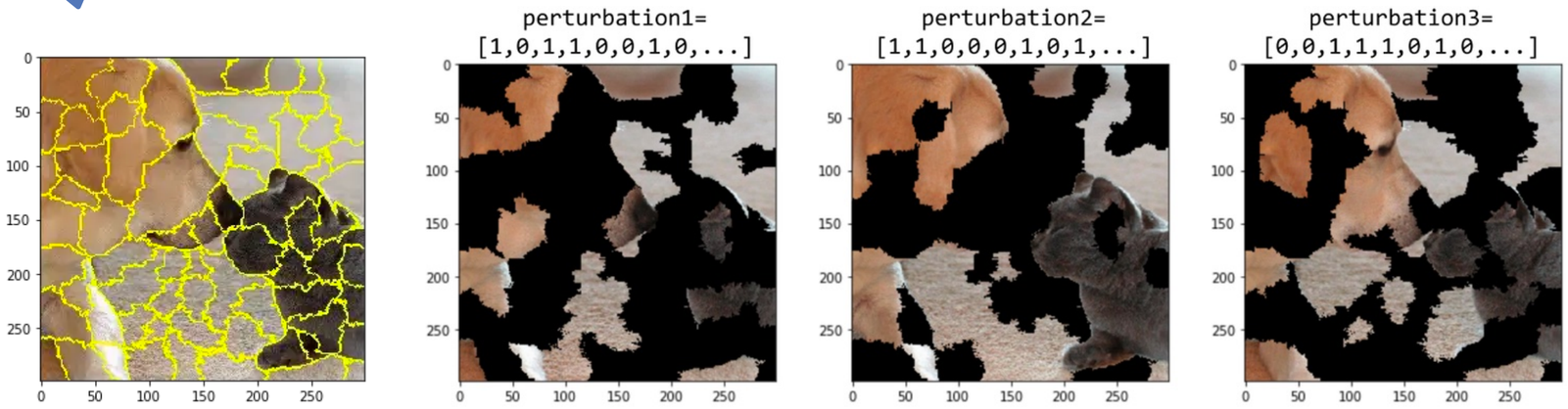


Image Source.
(left) "super-pixels" segmentation
(right) different perturbed samples

LIME explanations are sensitive to the choice of perturbation parameters, like its seed (Bansal et al., 2020)

Methods — Shapely Values

Model-Agnostic Method

- Estimate the value of a feature by measuring the difference in output (weighted average) as the feature is removed in different coalitions (Lundberg et al., 2017)

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(N - |S| - 1)!}{N!} (v(S \cup \{i\}) - v(S))$$

Summing over the subsets

Weigh the impact of each subset of features by N (# of features)

Score with feature i in subset S

Score without feature i in subset S

Methods — Shapely Values

Model-Agnostic Method

- Estimate the value of a feature by measuring the difference in output (weighted average) as the feature is removed in different coalitions (Lundberg et al., 2017)

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(N - |S| - 1)!}{N!} (v(S \cup \{i\}) - v(S))$$

Summing over the subsets

Weigh the impact of each subset of features by N (# of features)

Score with feature i in subset S

Score without feature i in subset S

Again, computationally expensive

Idea #2

Can we use gradient information to estimate feature importance?

Methods — The Gradient

Model-Aware Methods

- Gradient-based methods are a family of methods that explain a *differentiable* model
  - Compute the gradient of the output prediction w.r.t the input

Partial derivative = unit variation
treating all other inputs as constants

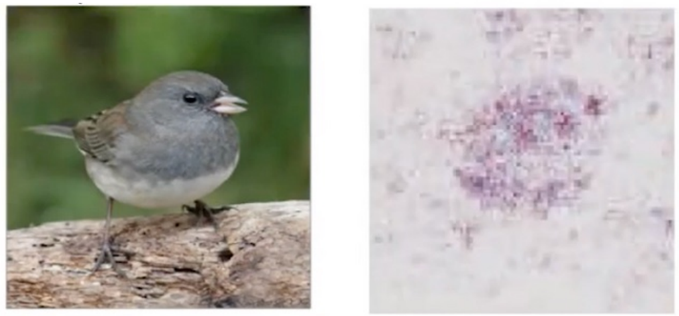


Image Source.

The intensity of the colour
indicates the importance of the feature

Methods — The Gradient

Model-Aware Methods

- Gradient-based methods are a family of methods that explain a *differentiable* model
  - Compute the gradient of the output prediction w.r.t the input

Partial derivative = unit variation
treating all other inputs as constants

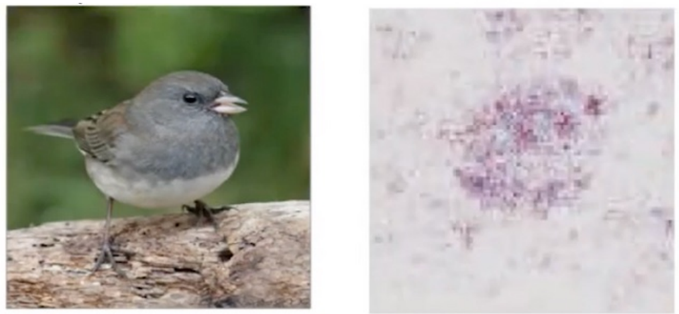


Image Source.

The intensity of the colour
indicates the importance of the feature

"I have no reason to believe the gradient holds anywhere other than very locally."

Methods — The Gradient is Fragile

Model-Aware Methods

- Because of the depth of the function (Balduzzi et al., 2017) or the usage of ReLU activation functions, the gradients may behave unsmoothly and noisy while the function output may not

If following a trajectory along the input (e.g., an athlete lifting weights), there is a discrepancy between f(x) and ∇f(x)

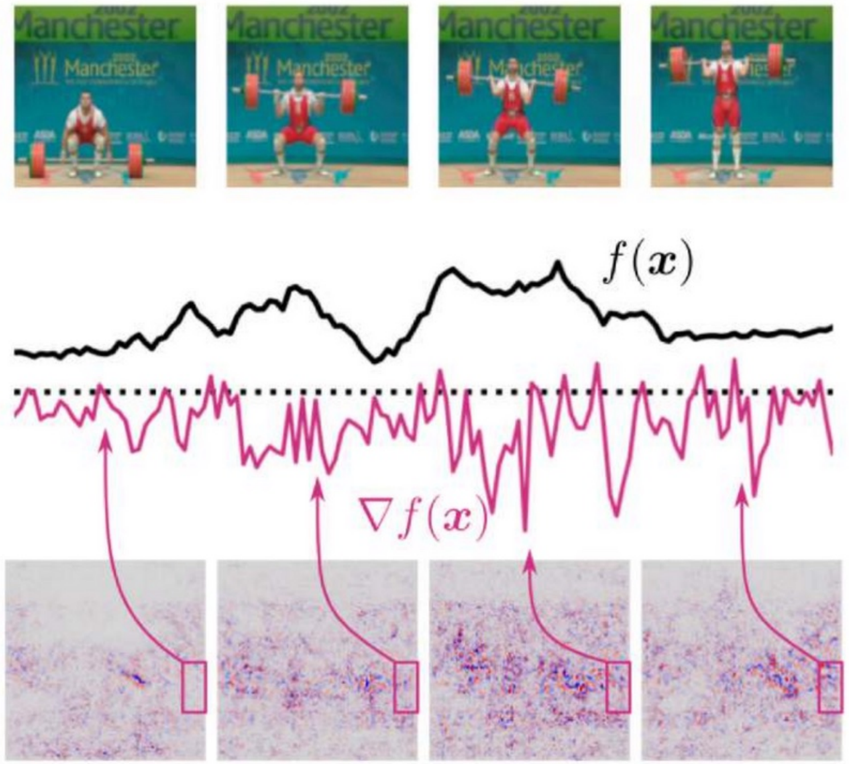


Image Source.

Methods — The Gradient is Fragile

Model-Aware Methods

- Because of the depth of the function (Balduzzi et al., 2017) or the usage of ReLU activation functions, the gradients may behave unsmoothly and noisy while the function output may not

If following a trajectory along the input (e.g., an athlete lifting weights), there is a discrepancy between f(x) and ∇f(x)

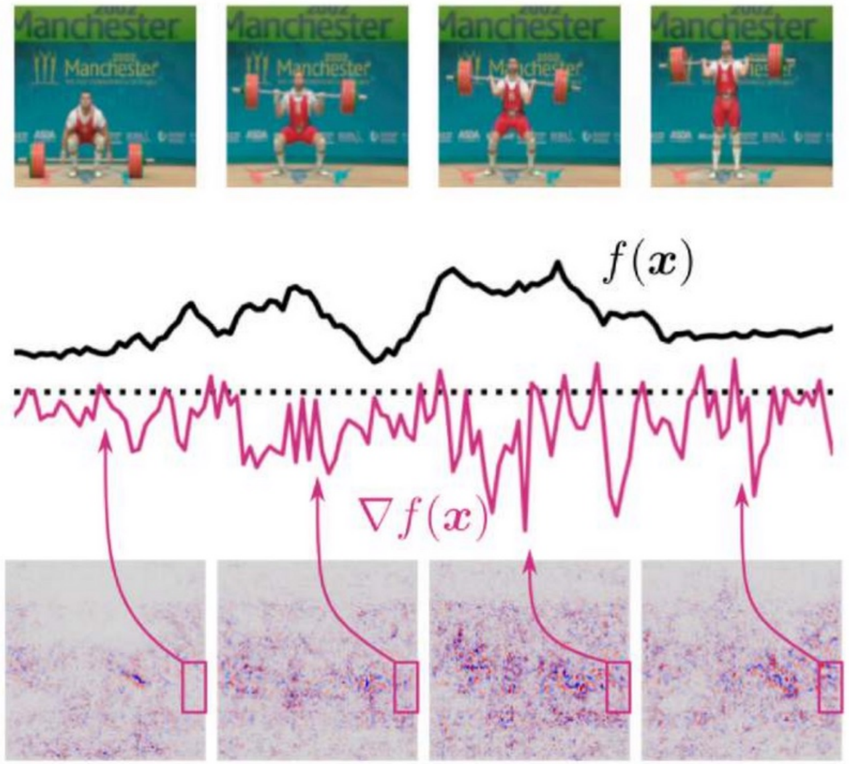


Image Source.

How can we denoise the gradient method?

Methods — Many Variants

Model-Aware Methods

- Many variants have been proposed

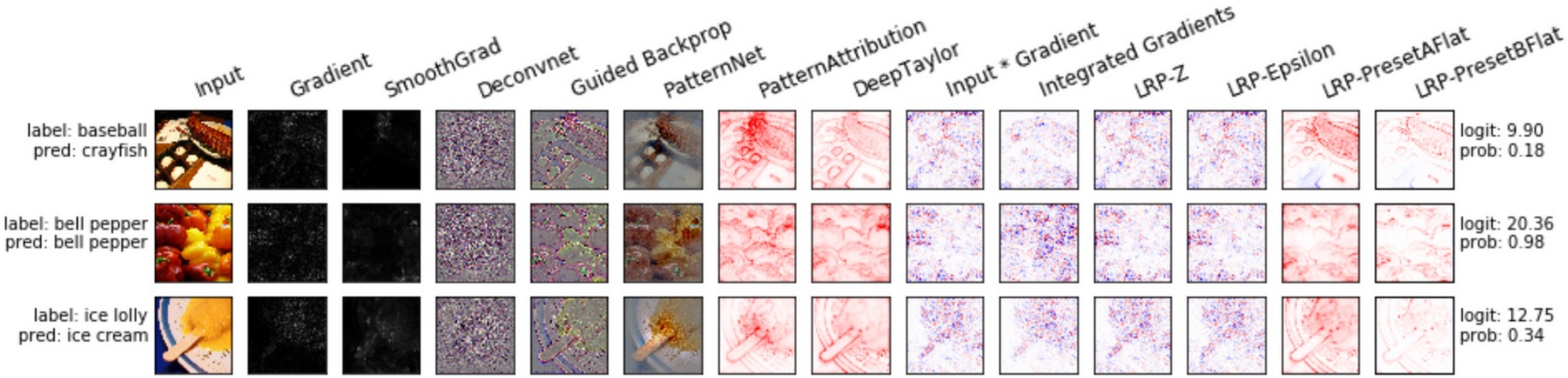


Image Source.

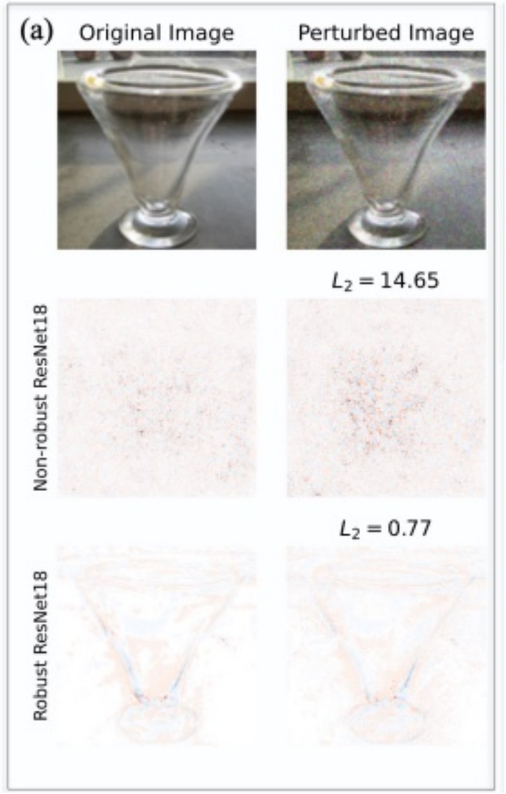
Methods — Regularisation

Ideas to Denoise Gradients

How can gradient-based explanations be robust if the underlying model is not?

Idea: Regularise the model and get better explanations (Chalasani et al. 2018; Datta et al., 2021; Tan et al., 2023)

bottom: regularised



top: vanilla model

Image Source.

Methods — Integrated Gradients

Leverging “No-Evidence” Baseline

Why not average over many points (inputs) along a straight-line path (from baseline to the input)?

Idea: Integrated Gradients (Sundararajan et al., 2017) average the gradient explanation, as the input moves from a baseline (a point of no or neutral information) towards the actual input

$$\phi(x) = (x - x_0) \odot \int_0^1 \nabla f(x_0 + \alpha(x - x_0))\, d\alpha$$

baseline

path integral

n steps

What makes up a good baseline (Sturmfels et al., 2020)?

Methods — SmoothGrad

Enhancing Explanations by Adding Noise to Input

- A method called **SmoothGrad** (Smilkov et al., 2017), is aimed at reducing visual diffusion

$$\frac{1}{N}\sum_{i=1}^{N} E\left(\boldsymbol{x}+\xi_i, f(\cdot,\hat{W})\right),$$

$$\xi_i \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\mathrm{SG}}\mathbf{I})$$

- Works by adding Gaussian noise to the input and takes the average over noisy input instances

Methods — SmoothGrad

Enhancing Explanations by Adding Noise to Input

- A method called **SmoothGrad** (Smilkov et al., 2017), is aimed at reducing visual diffusion

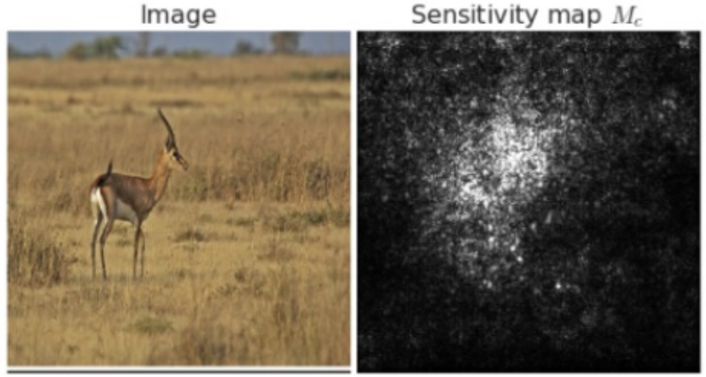


Image Source.

$$\frac{1}{N}\sum_{i=1}^{N} E\left(\boldsymbol{x}+\xi_i, f(\cdot,\hat{W})\right),$$

$$\xi_i \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathrm{SG}}^2\mathbf{I})$$

- Works by adding Gaussian noise to the input and takes the average over noisy input instances

But how many samples and how much noise?

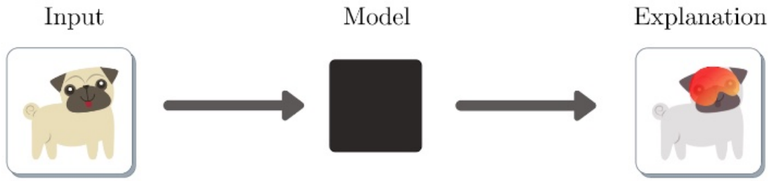
Methods — NoiseGrad

Leverage the neighbourhood of f? Approximate the posterior

- We start with a baseline explanation

Baseline explanation

$$E\left(\boldsymbol{x}, f(\cdot, \hat{W})\right)$$



- **NoiseGrad** "bayesianises" f: draw samples with multiplicative Gaussian noise added to f's weights

NoiseGrad-enhanced explanation

$$\frac{1}{M}\sum_{i=1}^{M} E\left(\boldsymbol{x}, f(\cdot, \mathcal{W}_i)\right)$$

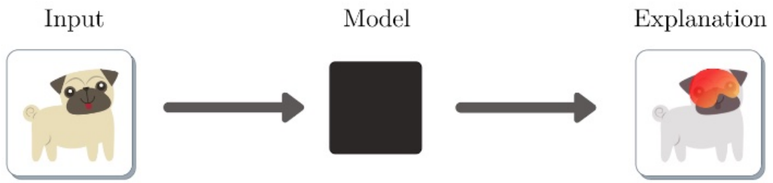
Methods — NoiseGrad

Leverage the neighbourhood of f? Approximate the posterior

- We start with a baseline explanation

Baseline explanation

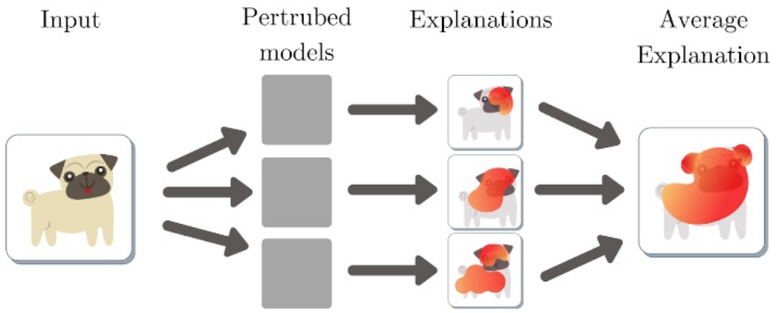
$$E\left(\boldsymbol{x}, f(\cdot, \hat{W})\right)$$



- **NoiseGrad** "bayesianises" *f*: draw samples with multiplicative Gaussian noise added to *f*'s weights

NoiseGrad-enhanced explanation

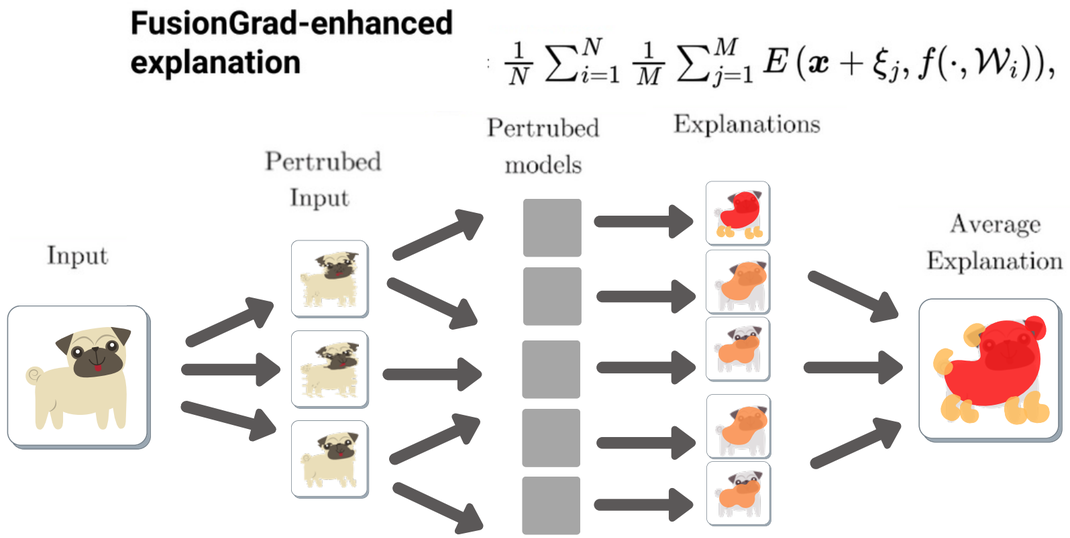
$$\frac{1}{M}\sum_{i=1}^{M} E\left(\boldsymbol{x}, f(\cdot, \mathcal{W}_i)\right)$$



Methods — FusionGrad

Leverage both neighbourhoods (x, f) to Enhance Explanations

- "Fusing" NoiseGrad and SmoothGrad: **FusionGrad** (Bykov et al., 2022)



Methods — FusionGrad

Toy-Example: Differences in Explanation Behaviour

Arrow:
gradient explanation

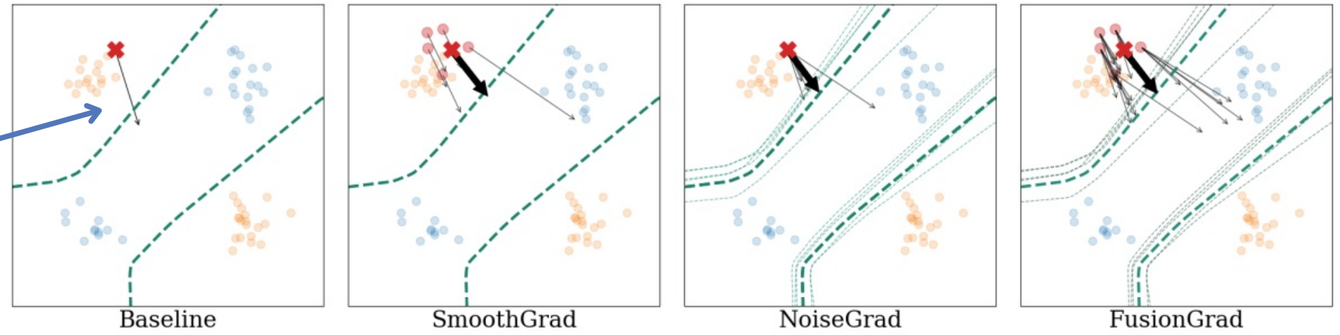
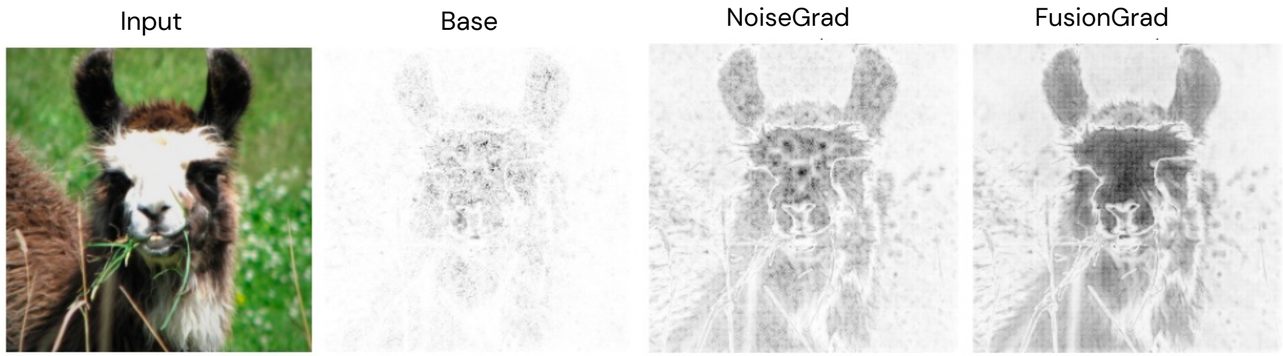


Figure 1: Illustration of the differences in explanation behavior between Baseline (gradient-based explanation), SmoothGrad, NoiseGrad, and FusionGrad for a toy experiment. Given training samples of two classes (orange and blue dots), a 3-layer MLP was trained for binary classification, where the learned decision boundary is shown by the green dashed line. The gradient explanations for a fixed test sample (red point) are shown as black arrows and the mean explanation as a bold black arrow. (a) For the baseline method, the explanation is the gradient itself. (b) SmoothGrad enhances the explanation by sampling points in the neighborhood (small red dots), and averaging their explanations (bold black arrow). (c) NoiseGrad enhances the explanation by averaging over perturbed models, indicated by multiple decision boundaries (thin green dashed lines). (d) FusionGrad combines SmoothGrad and NoiseGrad by incorporating both stochasticities in the input space and the model space.

Methods — Intermedite Results

More Localised Explanations

- Local (Saliency) explanations become more visually concise



Kirill and Hedström et al„ *NoiseGrad — Enhancing Explanations by Introducing Stochasticity to Model Weights*, Proceedings of the AAAI Conference on Artificial Intelligence (2021)

Methods — Intermedite Results

More Localised Explanations

- Local explanation methods become more faithful and robust

Method	Localization (↑)	Faithfulness (↑)	Robustness (↓)	Sparseness (↑)
Baseline	0.7315 ± 0.0505	0.3413 ± 0.1549	0.0763 ± 0.0265	$\mathbf{0.6272 \pm 0.0475}$
SG	0.8263 ± 0.0483	0.3465 ± 0.1601	0.0590 ± 0.0235	0.5310 ± 0.0635
NG	0.8349 ± 0.0367	$\mathbf{0.3635 \pm 0.1536}$	0.0224 ± 0.0080	0.5794 ± 0.0533
FG	$\mathbf{0.8435 \pm 0.0358}$	$\mathbf{0.3697 \pm 0.1465}$	$\mathbf{0.0153 \pm 0.0058}$	0.5721 ± 0.0532

Table 1: Comparison of attribution quality where the noise levels are set by the heuristic. ↑ and ↓ indicate the larger is the better and the smaller is the better, respectively. The values of the best method and the methods that are not significantly outperformed by the best method, according to the Wilcoxon signed-rank test for $p = 0.05$, are bold-faced.

Idea #3

Can we leverage information from activations in different layers to estimate feature importance?

Methods — LRP

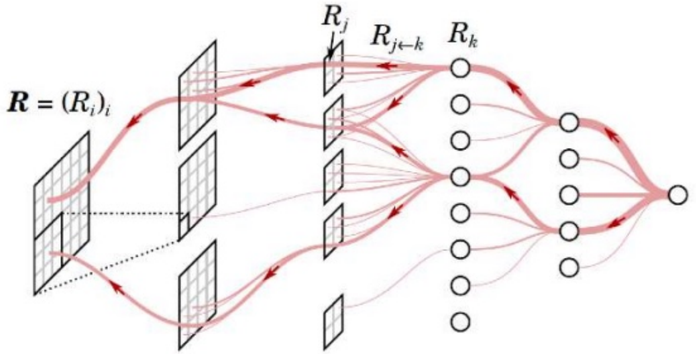
Model-Aware Methods

- **Layer-wise Relevance Propagation (LRP)** is a white-box method grounded on the principles of flow conservation and proportional decomposition

- Works by performing a backward pass: distributing a model output quantity proportionally across the layers, according to the activations, back to input space

$$\sum_i R_i = \ldots = \sum_i R_i^{(l)} = \sum_j R_j^{(l+1)} = \ldots = f(x)$$

- Employs different propagation rules for different architectures and layer types



Methods — LRP

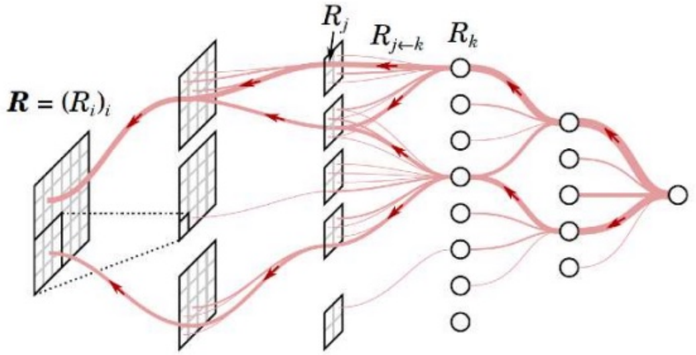
Model-Aware Methods

- **Layer-wise Relevance Propagation (LRP)** is a white-box method grounded on the principles of flow conservation and proportional decomposition

- Works by performing a backward pass: distributing a model output quantity proportionally across the layers, according to the activations, back to input space

$$\sum_i R_i = \ldots = \sum_i R_i^{(l)} = \sum_j R_j^{(l+1)} = \ldots = f(x)$$

- Employs different propagation rules for different architectures and layer types



Methods — LRP

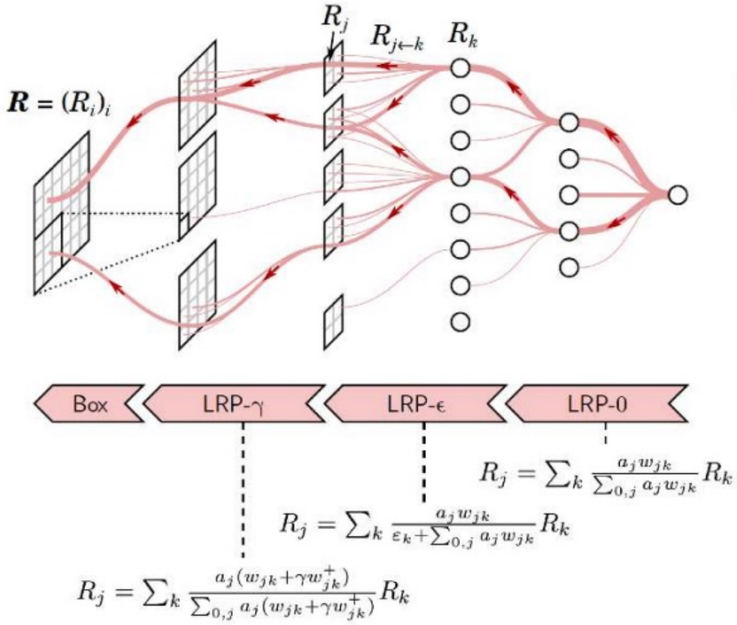
Model-Aware Methods

- **Layer-wise Relevance Propagation (LRP)** is a white-box method grounded on the principles of flow conservation and proportional decomposition

- Works by performing a backward pass: distributing a model output quantity proportionally across the layers, according to the activations, back to input space

$$\sum_i R_i = \ldots = \sum_i R_i^{(l)} = \sum_j R_j^{(l+1)} = \ldots = f(x)$$

- Employs different propagation rules for different architectures and layer types, e.g., AttnLRP (Achtibat, 2024)



Methods — LRP

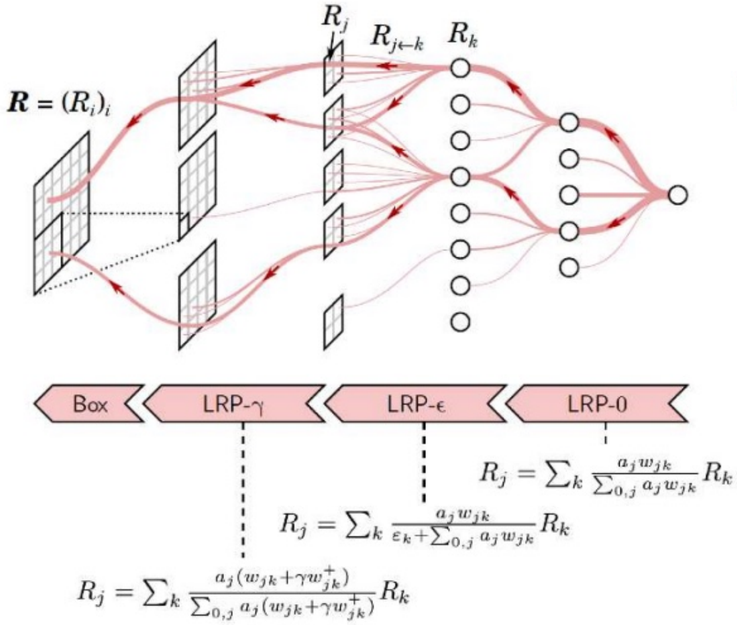
Model-Aware Methods

- **Layer-wise Relevance Propagation (LRP)** is a white-box method grounded on the principles of flow conservation and proportional decomposition

- Works by performing a backward pass: distributing a model output quantity proportionally across the layers, according to the activations, back to input space

$$\sum_i R_i = \ldots = \sum_i R_i^{(l)} = \sum_j R_j^{(l+1)} = \ldots = f(x)$$

- Employs different propagation rules for different architectures and layer types, e.g., AttnLRP (Achtibat, 2024)



LRP has hyperparameters which may be non-trivial to optimise

Methods — Local and Global Methods

A Gentle Notational Setup

- A model learns to map inputs $x \in \mathbb{R}$ to predictions $y \in \mathbb{R}$ via parameters θ:

$$f : X \to Y$$

- **Local explanations** provide attributions to input features of a specific prediction y:

$$\phi L(f, x, y; \lambda) = e$$

- **Global explanations** explain the global behaviour of a neuron $n \in \mathbb{R}$, independent of an input x:

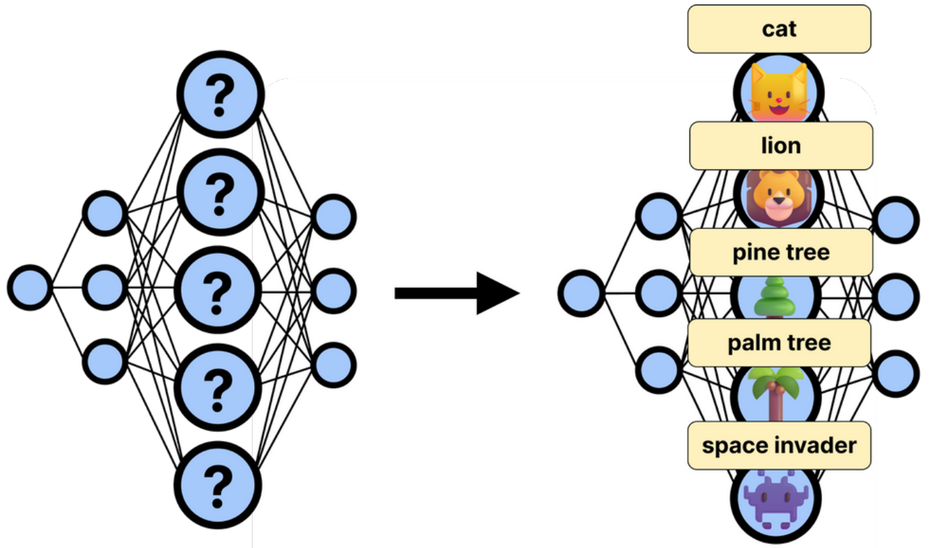
$$\phi G(f, n; \kappa) = e$$

Methods — Local and Global Methods

A Gentle Notational Setup

- Learning (or labelling) learned concepts in neural networks

What's the functional purpose of neuron *n_i,j?*



Credit.

Methods — Activation Maximisation

Global Method

- **Activation Maximization:** find the input pattern that produces a maximum neuron response (Simonyan et al., 2014)

How does a prototypical input look like that maximises a class "boat"?

1. Initialise the input image with random noise
2. Optimise via gradient descent the input image wrt target

Image Source.

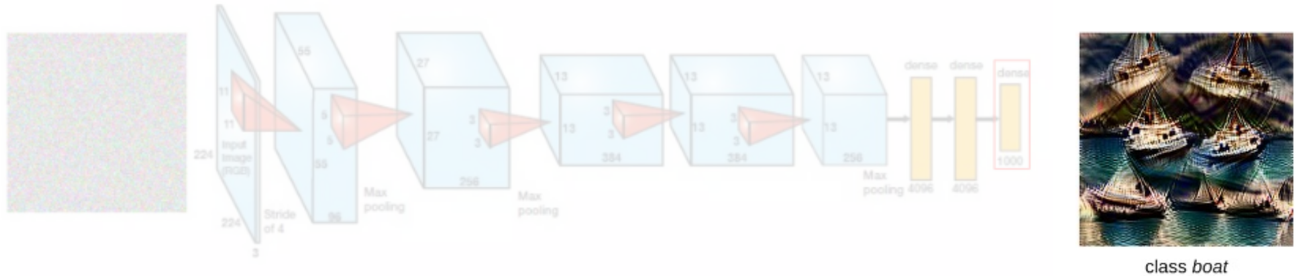
Methods — Activation Maximisation

Global Method

- **Activation Maximization:** find the input pattern that produces a maximum neuron response (Simonyan et al., 2014)

How does a prototypical input look like that maximises a class "boat"?

1. Initialise the input image with random noise
2. Optimise via gradient descent the input image wrt target



class *boat*

Image Source.

Methods — Activation Maximisation

Global Method

- **Activation Maximization:** find the input pattern that produces a maximum neuron response (Simonyan et al., 2014)

How does a prototypical input look like that maximises a class "boat"?

1. Initialise the input image with random noise
2. Optimise via gradient descent the input image wrt target

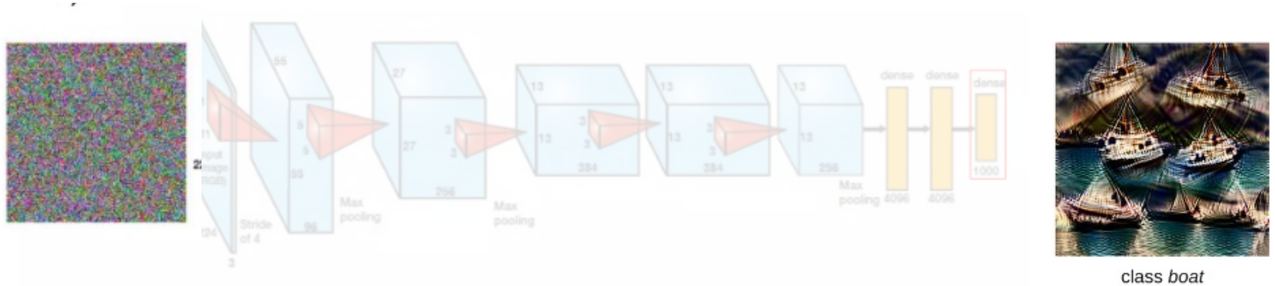


Image Source.

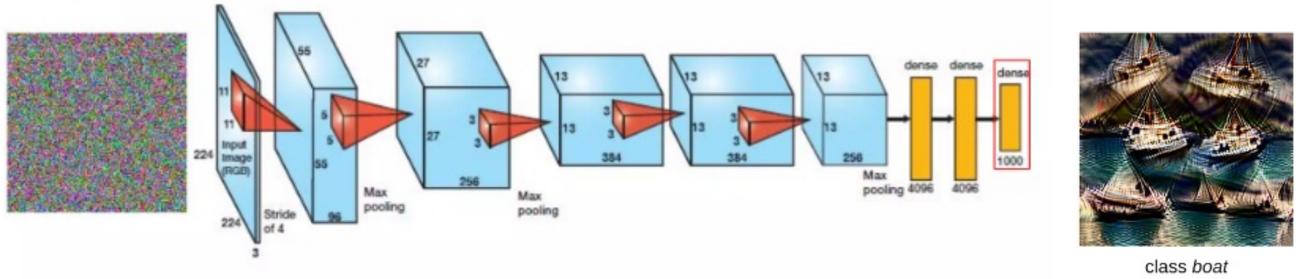
Methods — Activation Maximisation

Global Method

- **Activation Maximization:** find the input pattern that produces a maximum neuron response (Simonyan et al., 2014)

How does a prototypical input look like that maximises a class "boat"?

1. Initialise the input image with random noise
2. Optimise via gradient descent the input image wrt target



class *boat*

Image Source.

Methods — Activation Maximisation

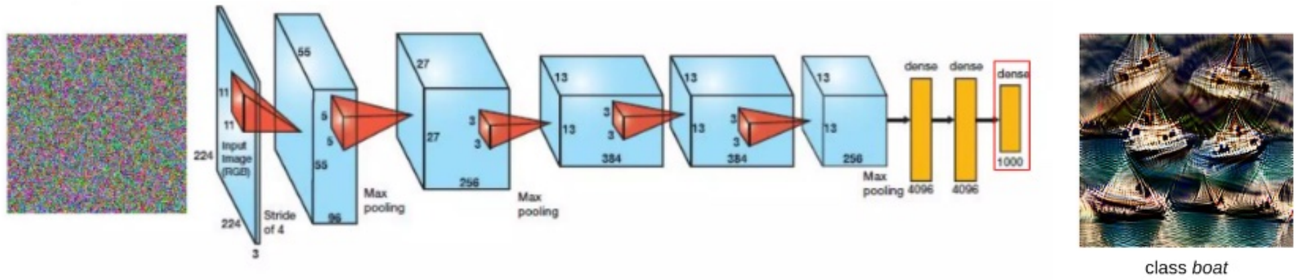
Global Method

- **Activation Maximization:** find the input pattern that produces a maximum neuron response (Simonyan et al., 2014)

How does a prototypical input look like that maximises a class "boat"?

1. Initialise the input image with random noise
2. Optimise via gradient descent the input image wrt target

AM might be sensitive to its initialisation and difficult to scale



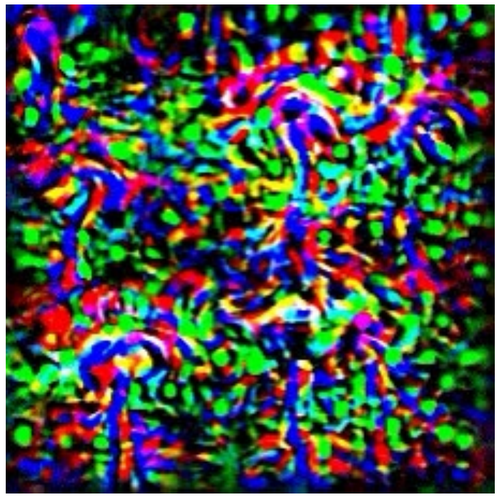
class *boat*

Image Source.

Methods — Activation Maximisation

Global Method

- An AM example, explaining “Flamingo” ImageNet class with AlexNet (repo)

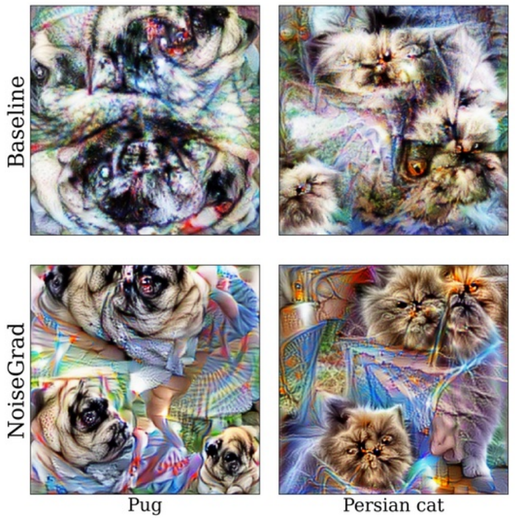


Giphy Source.

Methods — AM with NoiseGrad

Enhance AM Explanations with NoiseGrad

- Re-apply the "explanation-enhancing" Noisegrad method (Bykov et al., 2022) with global AM explanations, finding that the resulting explanation is more semantically meaningful



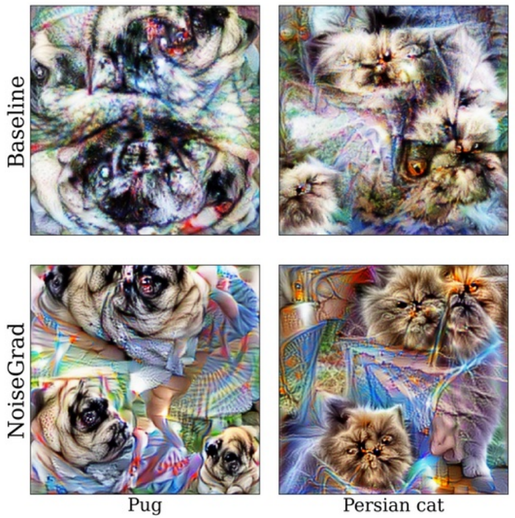
AM explanations for ImageNet classes, ResNet-18 model

Methods — AM with NoiseGrad

Enhance AM Explanations with NoiseGrad

- Re-apply the "explanation-enhancing" Noisegrad method (Bykov et al., 2022) with global AM explanations, finding that the resulting explanation is more semantically meaningful

How interpretable are AM outputs as explanations?



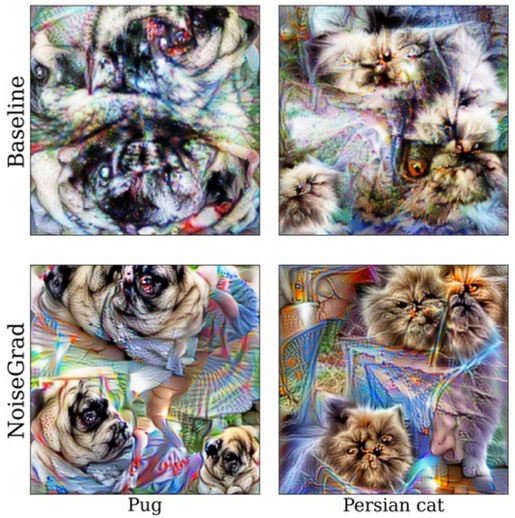
AM explanations for ImageNet classes, ResNet-18 model

Methods — AM with NoiseGrad

Enhance AM Explanations with NoiseGrad

- Re-apply the "explanation-enhancing" Noisegrad method (Bykov et al., 2022) with global AM explanations, finding that the resulting explanation is more semantically meaningful (right)

How interpretable are AM outputs as explanations?



AM explanations for ImageNet classes, ResNet-18 model

Layer dependence?

Idea #4

Can we map the explanation to concepts instead of inputs?

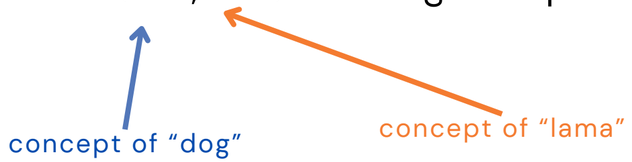
Methods — Concepts

The Abstract Idea of A Concept

Definition 2 (Concepts). *A concept* $c \in \mathbb{C}$ *is defined as a binary function:* $c : \mathbb{D} \longrightarrow \{0, 1\}$*, which maps the data domain* $\mathbb{D}$ *to the set of binary numbers. A value of* 1 *indicates the presence of the concept in the input, and* 0 *indicates its absence. Here,* $\mathbb{C}$ *corresponds to the space of all concepts, that could be defined on* $\mathbb{D}$*.*

and then ....

- Combine existing concepts let C = { c1, c2 } with logical operators AND, OR, and NOT



Methods — INVERT

Explain a Neural Representation With Concepts

- **Inverse Recognition (INVERT):** find compositional concepts that explain a neuron, i.e., label representations in a neural network (Bykov et al., 2023)

How?

Imagine we have a neuron that we want to explain

And a labelled dataset

Methods — INVERT

Explain a Neural Representation With Concepts

- **Inverse Recognition (INVERT):** find compositional concepts that explain a neuron, i.e., label representations in a neural network (Bykov et al., 2023)

How?

Imagine we have a neuron that we want to explain

And a labelled dataset

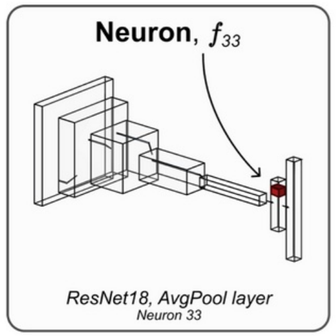
Methods — INVERT

Explain a Neural Representation With Concepts

- **Inverse Recognition (INVERT):** find compositional concepts that explain a neuron, i.e., label representations in a neural network (Bykov et al., 2023)

How?

Imagine we have a neuron that we want to explain



And a labelled dataset

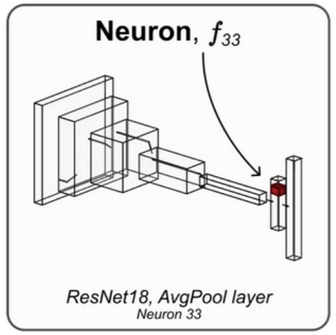
Methods — INVERT

Explain a Neural Representation With Concepts

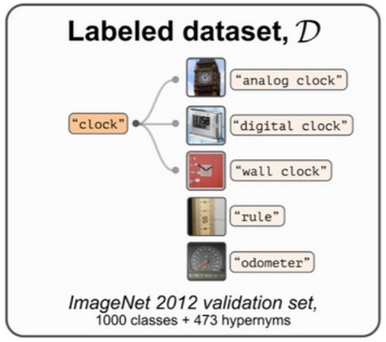
- **Inverse Recognition (INVERT):** find compositional concepts that explain a neuron, i.e., label representations in a neural network (Bykov et al., 2023)

How?

Imagine we have a neuron that we want to explain



And a labelled dataset

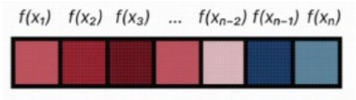


Each image has several labels

Methods — INVERT

Explain a Neural Representation With Concepts

- **Idea:** since we have the labelled information about the “concepts” for each input sample, we perform a forward pass and collect the activations for a specific layer (e.g., avg pool)



- Then we compare these values with the actual one-hot encoded labels (ground truth)

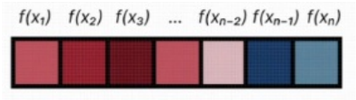
	image_name	n01440764	n01443537	n01484850	n01491361	r
0	ILSVRC2012_val_00000001	0	0	0	0	
1	ILSVRC2012_val_00000002	0	0	0	0	
2	ILSVRC2012_val_00000003	0	0	0	0	

finding concepts via “Beam” search

Methods — INVERT

Explain a Neural Representation With Concepts

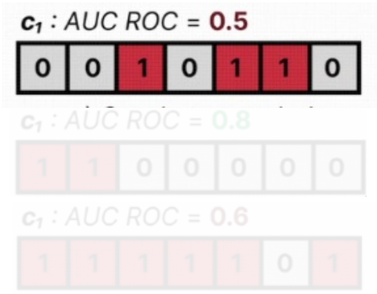
- **Idea:** since we have the labelled information about the "concepts" for each input sample, we perform a forward pass and collect the activations for a specific layer (e.g., avg pool)



- Then we compare these values with the actual one-hot encoded labels (ground truth)

	image_name	n01440764	n01443537	n01484850	n01491361	r
0	ILSVRC2012_val_00000001	0	0	0	0	
1	ILSVRC2012_val_00000002	0	0	0	0	
2	ILSVRC2012_val_00000003	0	0	0	0	

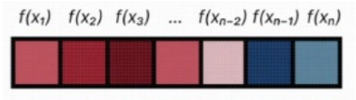
finding concepts via "Beam" search



Methods — INVERT

Explain a Neural Representation With Concepts

- **Idea:** since we have the labelled information about the "concepts" for each input sample, we perform a forward pass and collect the activations for a specific layer (e.g., avg pool)



- Then we compare these values with the actual one-hot encoded labels (ground truth)

	image_name	n01440764	n01443537	n01484850	n01491361	r
0	ILSVRC2012_val_00000001	0	0	0	0	
1	ILSVRC2012_val_00000002	0	0	0	0	
2	ILSVRC2012_val_00000003	0	0	0	0	

finding concepts via "Beam" search

- We can observe that activations of the data points corresponding to the resulting explanation (in orange) are generally significantly higher than for all other points (in blue).

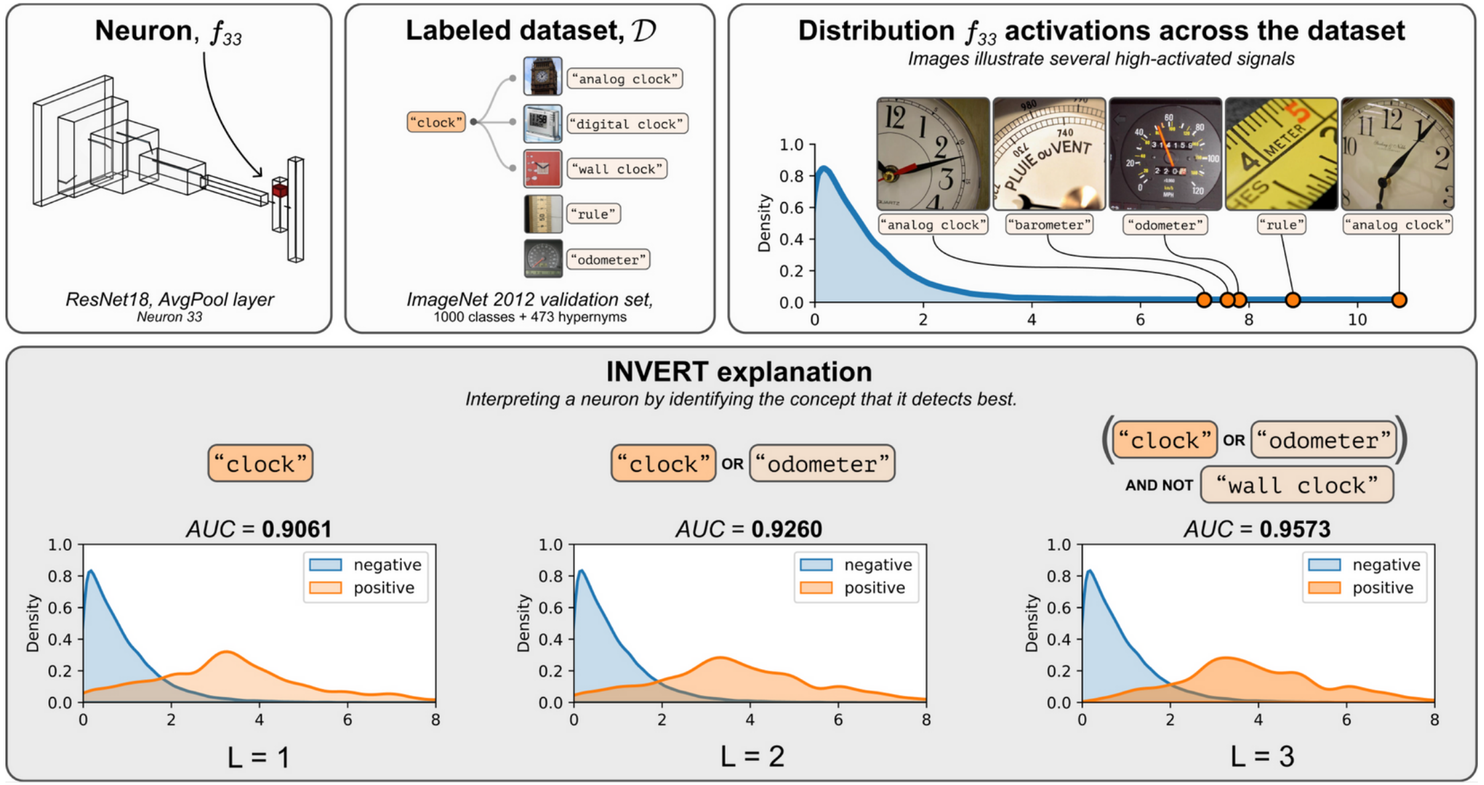


Figure 1: Demonstration of the INVERT method for the neuron from ResNet18 model.

Image Source.

Methods — “mechanistic interpret”

Sold as High-impact Area, Find Subgraph Within A Network

- Explain “circuits” with the three most significant neurons (in terms of the weight of linear connection) and their corresponding INVERT explanation linked to the ImageNet class logit “carton”.

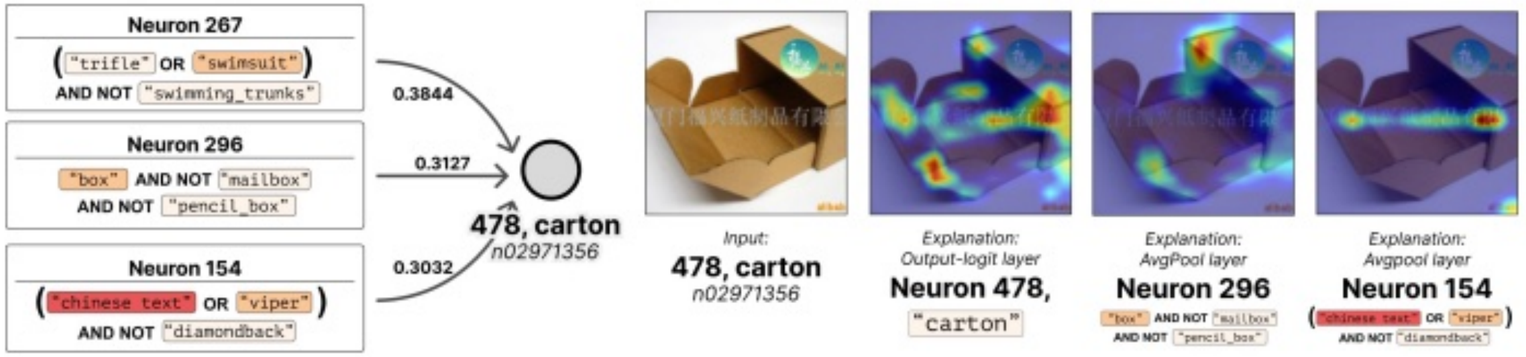


Image Source.

Methods

Self-explaining methods

Post-hoc explanations “must be wrong”; that they are by definition not completely faithful to the original model and must be less accurate with respect to the primary task.

"Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead" by Rudin (2019)

Post-hoc explanations “must be wrong”; that they are by definition not completely faithful to the original model and must be less accurate with respect to the primary task.

"Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead" by Rudin (2019)

What about the practicality and feasibility of exclusively using interpretable models?

Evaluation

Failure Modes

What challenges is the field experiencing?

What challenges is the field experiencing?

Let's review some failure modes

Failure Mode 1 — Manipulable

Explanation Functions Can Be Manipulated

Turns out that local methods are far from fault-proof

Image Source.

Failure Mode 1 — Manipulable

Explanation Functions Can Be Manipulated

Idea: similar to how adversarial attacks manipulate the model, manipulate explanations by optimising a customised loss function of the model using simple gradient descent (Dombrowski, 2019)

$$\mathcal{L} = \left\| h(x_{\text{adv}}) - h^t \right\|^2 + \gamma \left\| g(x_{\text{adv}}) - g(x) \right\|^2 ,$$

Create an adversarial example such that the norm of the input image is small << 1

Ensure that the manipulated explanation map is close to the target

Encourage the network to have the same output

Failure Mode 1 — Manipulable

Explanation Functions Can Be Manipulated

Result: While perturbations applied to the input are visually indistinguishable, explanations are vastly different

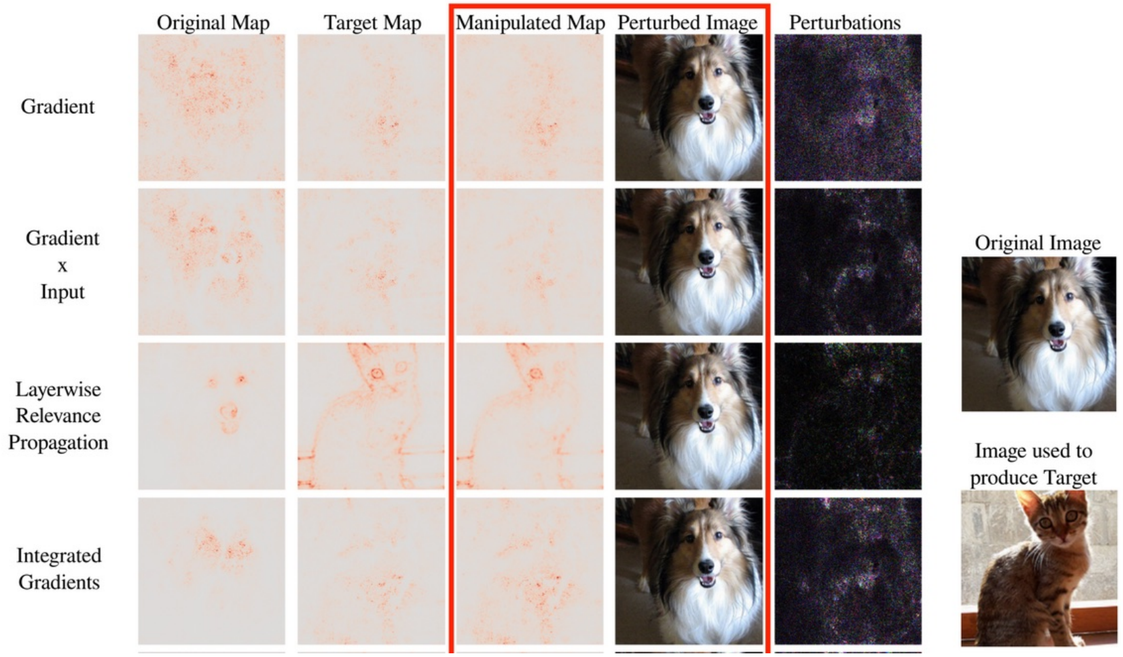
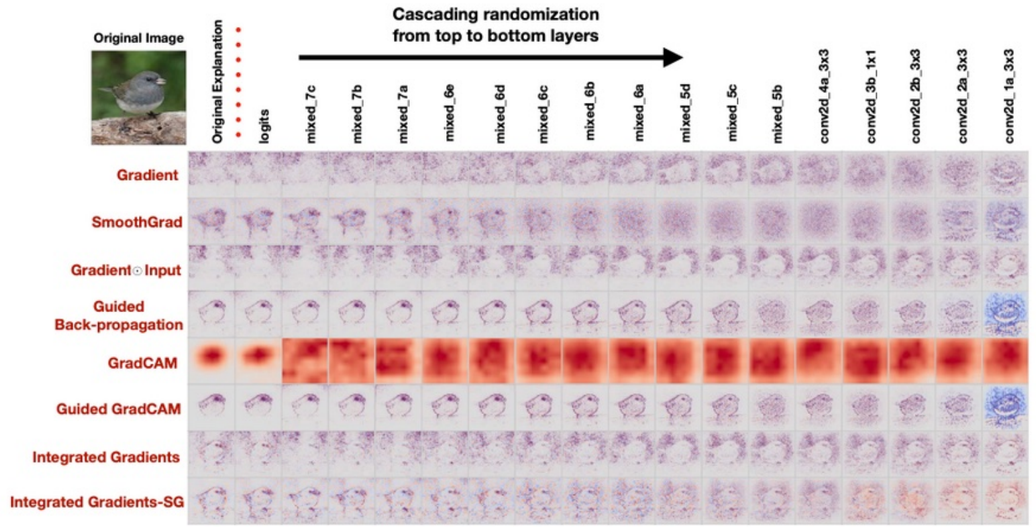


Image Source.

Failure Mode 2 — Model Invariant

Explanation Functions Can Be Invariant to Model Parameters

Idea: Randomise the parameters from top to bottom layers in a cascading way, and measure the distance of the resulting explanation to the original explanation



Seeking large difference in explanation outcomes between ***f*** and ***f'***

Image Source.

Failure Mode 2 — Model Invariant

Explanation Functions Can Be Invariant to Model Parameters

Result: Contrary to intuition, showing that explanations for accurate and random models are similar (high SSIM, high Rank Correlation)

Lower values are better meaning explanations are dis-similar

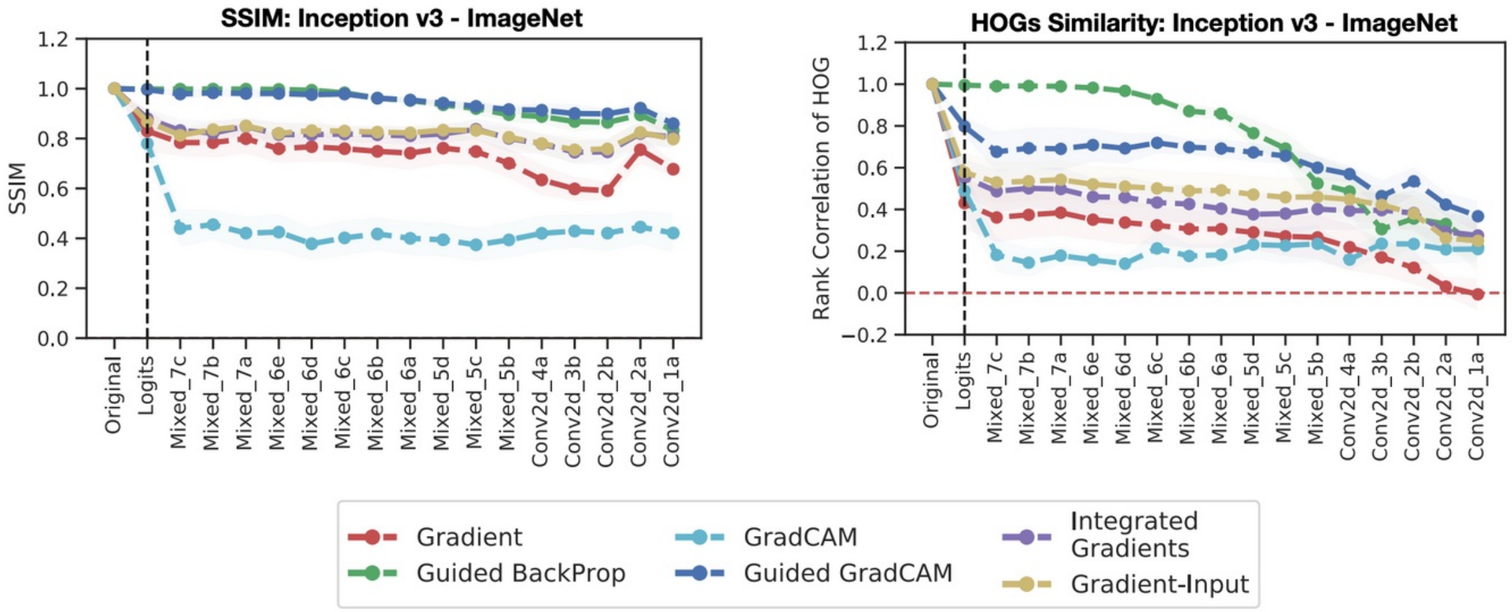
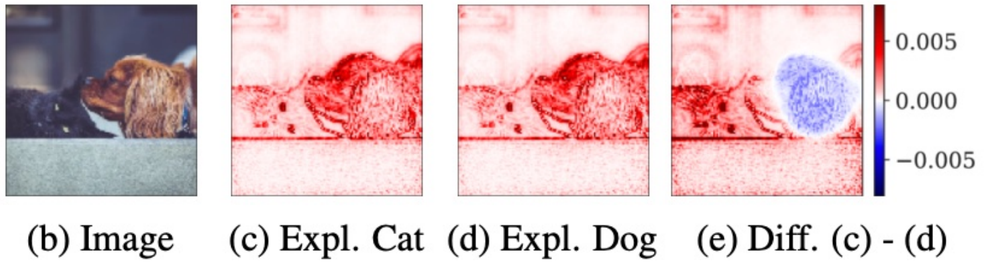


Image Source.

Failure Mode 3 — Class Invariant

Explanation Functions Can Be Invariant to Logit

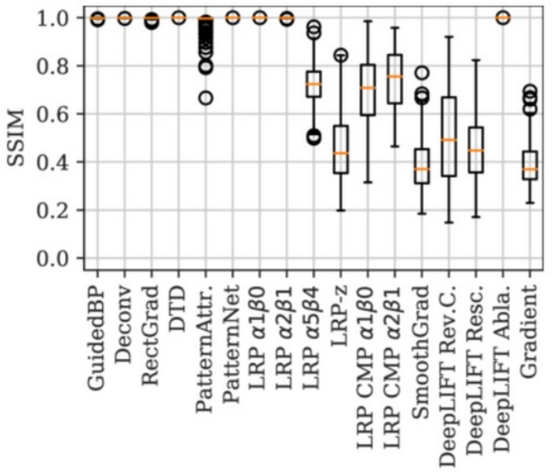
Idea: Explanations should be sensitive to their class: i.e., an explanation should be different if chosen for a random logit (Sixt et. al., 2020)



(b) Image (c) Expl. Cat (d) Expl. Dog (e) Diff. (c) - (d)

Image Source.

Lower values are better meaning explanations are dis-similar



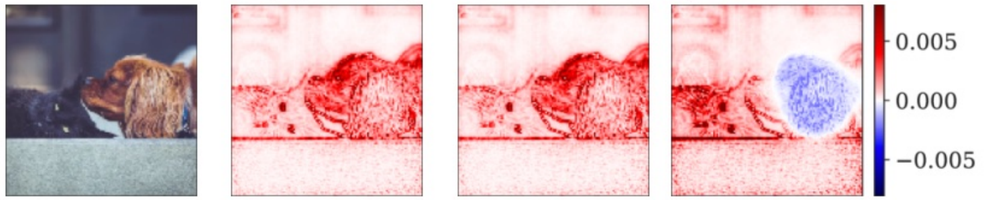
perfect similarity

perfect dis-similarity

Failure Mode 3 — Class Invariant

Explanation Functions Can Be Invariant to Logit

Idea: Explanations should be sensitive to their class: i.e., an explanation should be different if chosen for a random logit (Sixt et. al., 2020)

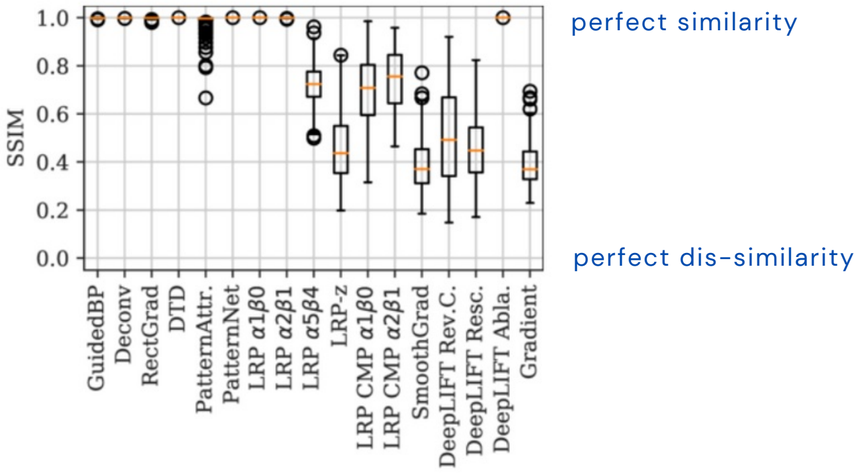


(b) Image (c) Expl. Cat (d) Expl. Dog (e) Diff. (c) - (d)

Image Source.

Lower values are better meaning explanations are dis-similar

But why cannot two classes rely on the same evidence?



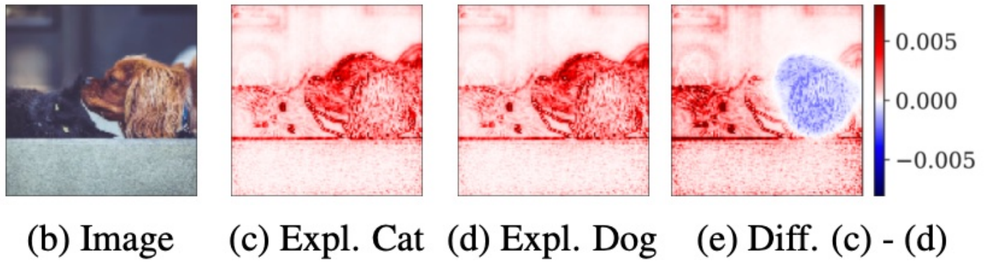
perfect similarity

perfect dis-similarity

Failure Mode 3 — Class Invariant

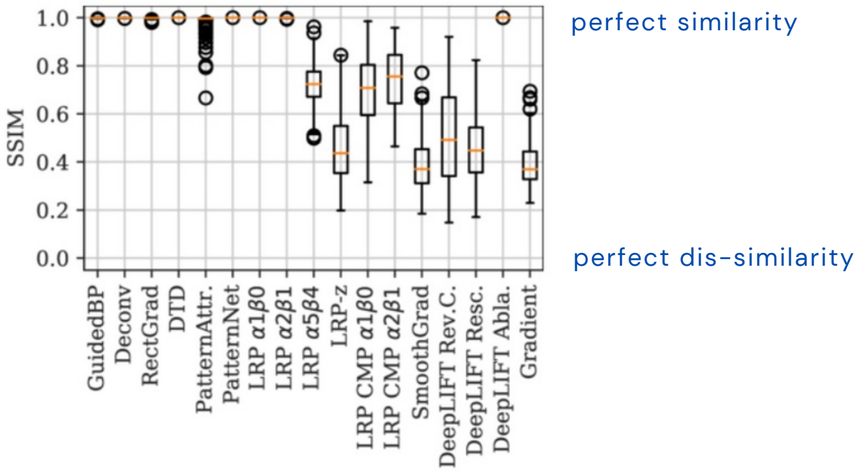
Explanation Functions Can Be Invariant to Logit

Idea: Explanations are sensitive to their class: compute the distance between the original explanation and the explanation for a random logit (Sixt et. al., 2020)



(b) Image (c) Expl. Cat (d) Expl. Dog (e) Diff. (c) - (d)

Image Source.



perfect similarity

perfect dis-similarity

Lower values are better meaning explanations are dis-similar

But why cannot two classes rely on the same evidence?

But are these findings stable enough to rule out explanation methods?

These studies paint a dark picture

These studies paint a dark picture

But there is more to unpack

Rebuttals — MPRT

The Need to “Sanity Check” Existing “Sanity Checks”

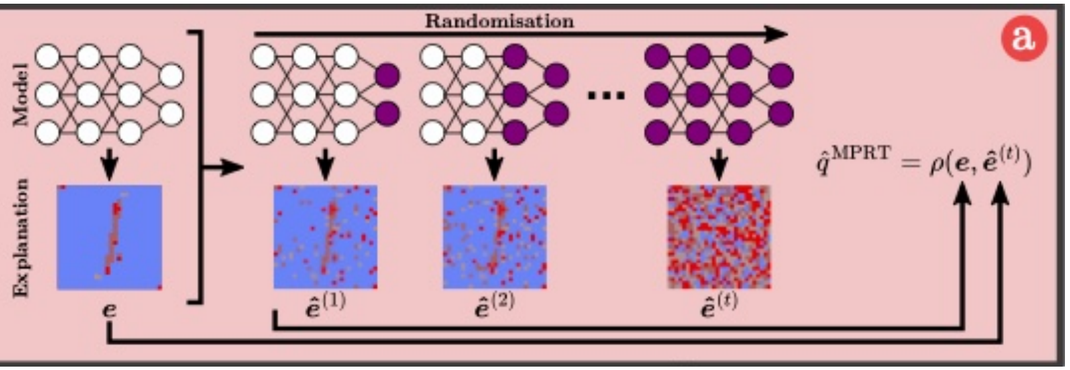
- Recall the widely used method MPRT that evaluates explanation quality by layer-wise top-down parameter randomisation, measuring similarity between explanations with SSIM

(Original) MPRT Metric

MPRT randomises parameters of f layer-wise in top-down order.

Explanation faithfulness is then tested by calculating explanation similarity $\rho(\boldsymbol{e}, \hat{\boldsymbol{e}}^{(t)})$ at each layer's randomisation, comparing explanations $\boldsymbol{e}$ and $\hat{\boldsymbol{e}}^{(t)}$, of the original f and randomised model $\hat{f}_l^t$, respectively.

However, several caveats of MPRT have been identified (b)-(d).



Rebuttals — Problems

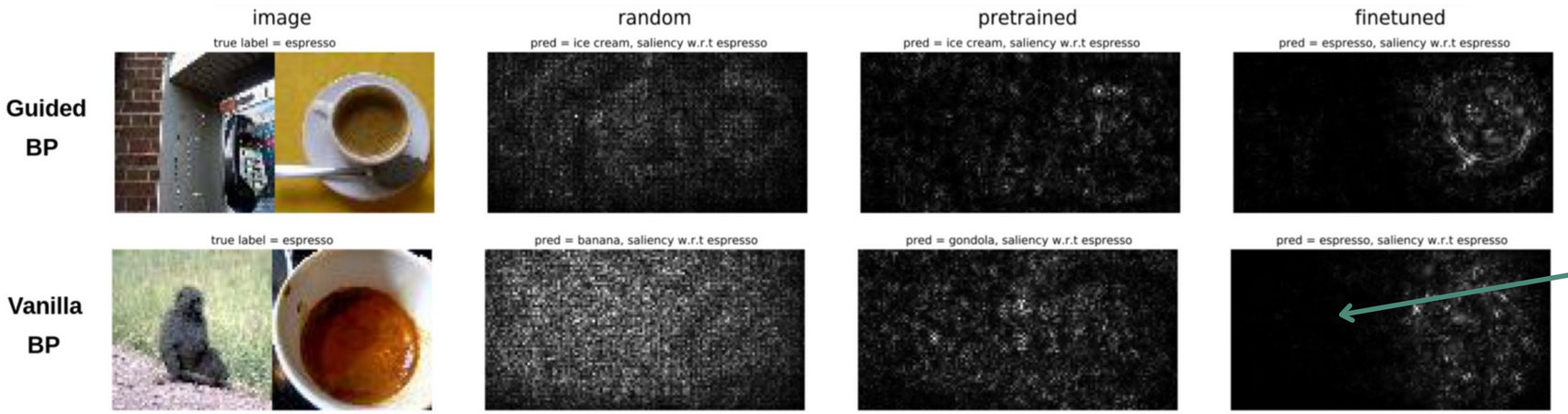
The Need to “Sanity Check” Existing “Sanity Checks”

- Several problems have been identified: task, pre-processing, layer-order and similarity measure

Rebuttal 1 — Task

The Need to “Sanity Check” Existing “Sanity Checks”

- Yona et al., 2021 show that the choice of task (and dataset) confound the results



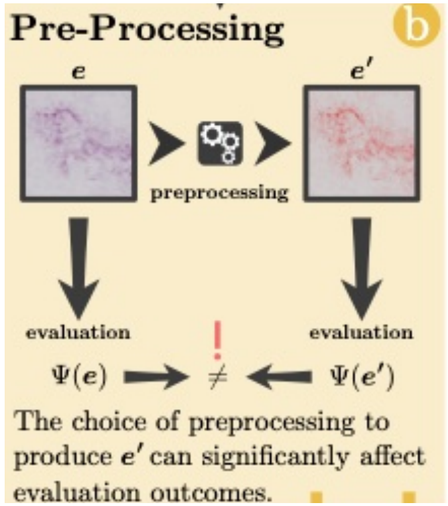
On the fine-tuned model, explainable evidence is correctly absent in the 'non-espresso' area

Image Source.

Rebuttal 2 — Preprocessing

The Need to “Sanity Check” Existing “Sanity Checks”

- Sundararajan et al., 2018 shows that IG with signed attributions pass the test

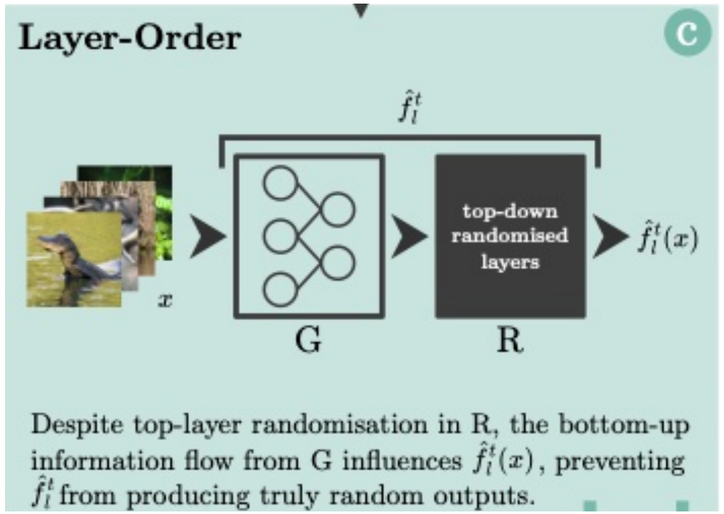


Absolute operation erases meaningful information

Rebuttal 3 — Layer-order

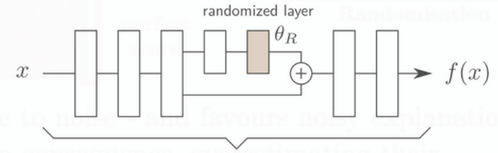
The Need to “Sanity Check” Existing “Sanity Checks”

- Binder et al., 2023 show that the layer order (top-down) distorts expectations of explanation change



Expected difference between e and e_{ran} affected by layer-order:

- Irrelevant features (stays irrelevant)
- Presence of skip-connections
- High activations (keeps dominant)



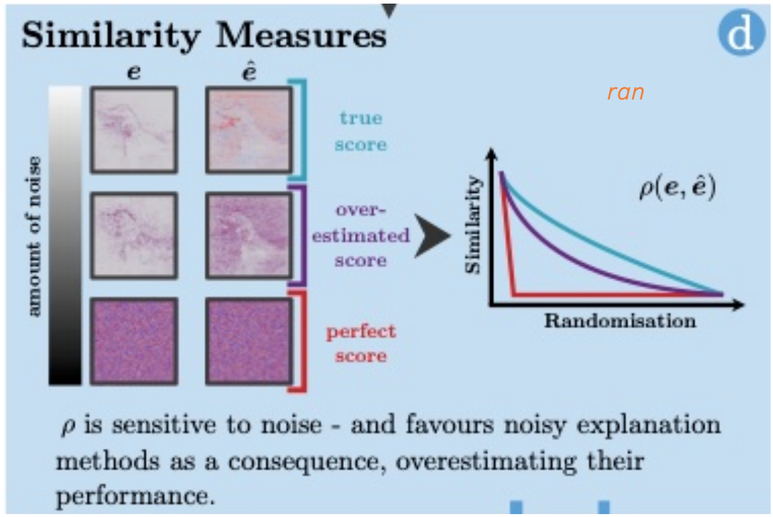
Rebuttal 3 — Similarity measure

The Need to “Sanity Check” Existing “Sanity Checks”

- Binder et al., 2023 show that choice of similarity measure favours certain “noisy” explanations

A **caveat** of the chosen similarity measure SSIM is that the presence of covariance term, drives computed difference to zero (indicating dissimilarity), **favouring “noisy” methods** (also true Spearman Rank, MSE)

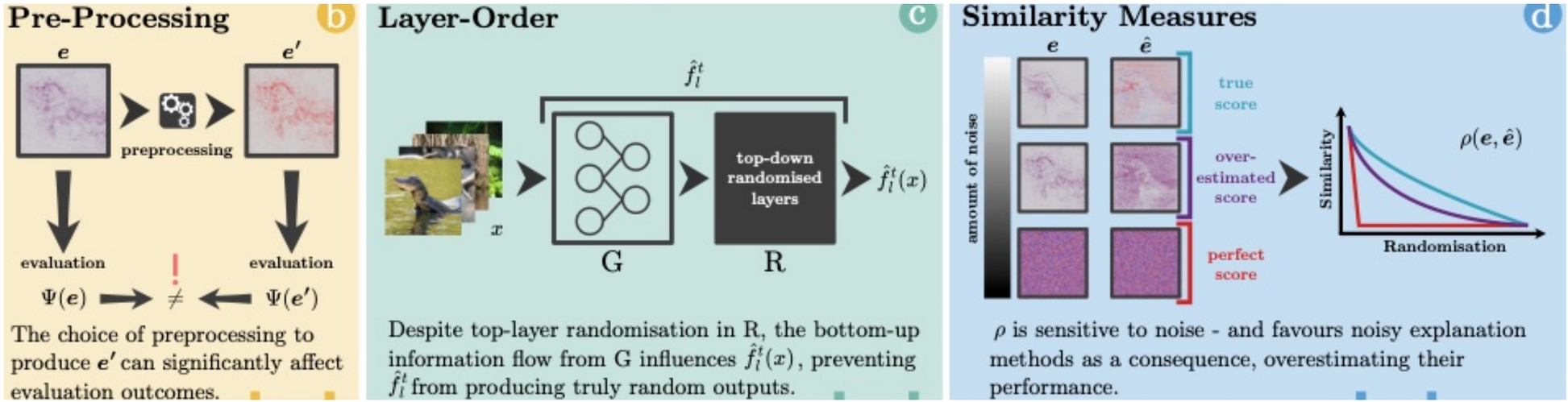
$$\frac{2\mu_A\mu_B + C_1}{\mu_A^2 + \mu_B^2 + C_1}\frac{2\sigma_{AB} + C_2}{\sigma_A^2 + \sigma_B^2 + C_2}.$$



Rebuttals — Empirical Confounds

The Need to “Sanity Check” Existing “Sanity Checks”

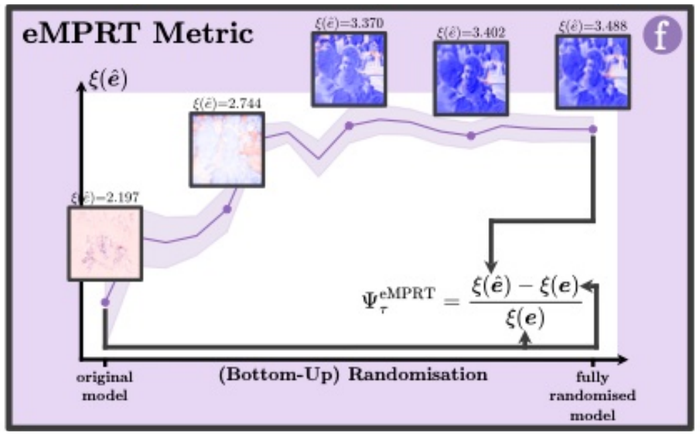
- Several problems have been identified: task, pre-processing, layer-order and similarity measure



Rebuttals — Implications

Correcting MPRT Results in Different Evaluation Outcomes

- Hedström et al., 2023b show that if MPRT is re-interpreted to eMPRT (i.e., compute a rise in discrete entropy complexity post-full parameter randomisation) — evaluation outcomes differ



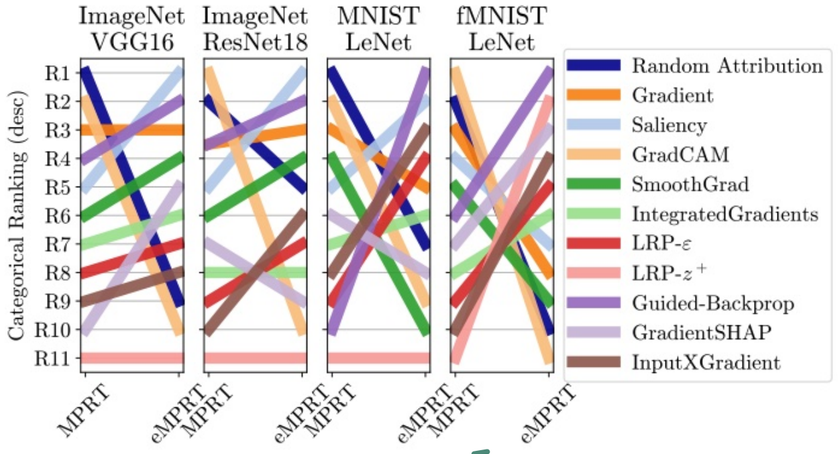


Image Source.

Categorical ranking between XAI methods change

It looks worryingly "easy to get it wrong" in XAI evaluation

It looks worryingly "easy to get it wrong" in XAI evaluation

Can we trust the methods from Explainable AI?

Explanation, can I trust you? I/II

Challenges in Explainable AI

- Due to the **presence of strong method assertions** (e.g., Adebayo et al., 2018; Sixt et al., 2020), followed by rebuttals (Yona et al., 2021; Binder et al., 2023), this question remains difficult to answer
- To make matters more challenging:
  - Lack of standardised benchmarks (many applications and explanation niches)
  - High-level of disagreements in the field (Krishna et al., 2022; Hedström et al., 2023a)
  - Misaligned expectations: bridging the function complexity gaps (Luxburg, 2023)

Explanation, can I trust you? II/II

Challenges in Explainable AI

- New, bigger models pose a new set of challenges, should we:
  - Adapt existing "traditional methods" to new architectures e.g., AttnLRP (Achtibat, 2024)
  - Recycle ideas e.g., ablation to automate circuit discovery for mechanistic interpretability (Conmy et al., 2023)

In the next lecture, we will look at how evaluation (and meta-evaluation) can help.

Foundations

Summary

Foundations — Summary I/III

A Summary

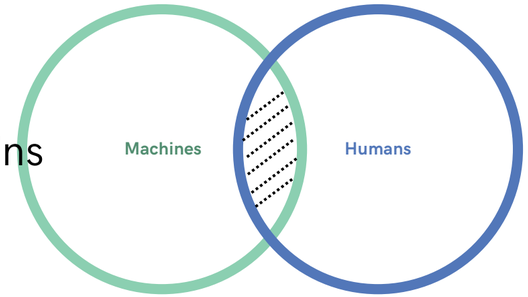
1. **Motivation & Definition** — Why do we need interpretability; arguments
2. **Methods** — What are the current methods; local and global
3. **Failure Modes** — Review the limitations of techniques; a critical view

Foundations — Summary I/III

Motivated by Trust, Debugging, Legal Compliance, Knowledge Acquisition

1. Motivation & Definition — Why do we need interpretability; arguments

(1) Establish trust and verify outcomes in test envs
(2) Debug model artefacts, uncover and fix learned concepts
(2) Assert regulatory compliance, especially in high-stake domains
(4) Acquire knowledge, generating novel scientific insights



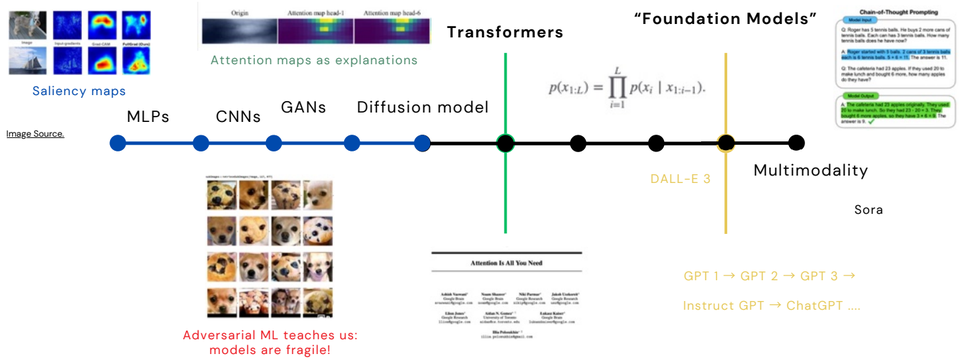
Foundations — Summary II/III

Local and Global Explainers

1. **Motivation & Definition** — Why do we need interpretability; arguments

2. **Methods** — What are the current methods; local and global

A Model Perspective



Foundations — Summary II/III

Local and Global Explainers

1. **Motivation & Definition** — Why do we need interpretability; arguments

2. Methods — What are the current methods; local and global

- **Local explanations** provide attributions to input features of a specific prediction y:

$$\phi L(f, x, y; \lambda) = e$$

- **Global explanations** explain the global behaviour of a neuron n, independent of an input x:

$$\phi G(f, n; \kappa) = e$$

Foundations — Summary II/III

Local and Global Explainers

1. Motivation & Definition — Why do we need interpretability; arguments

2. Methods — What are the current methods; local and global

- Different approaches, using:
  - Perturbation
  - Gradients
  - Decomposition
  - Concepts

Foundations — Summary III/III

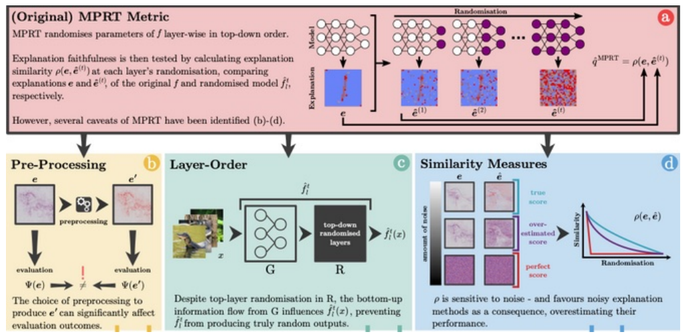
Local Methods “Fail” Tests But Is Rebutted

1. **Motivation & Definition** — Why do we need interpretability; arguments

2. **Methods** — What are the current methods; local and global

3. **Failure Modes** — Review the limitations of techniques; a critical view

Many methods score poorly in absolute terms,
but evaluations are not to be trusted blindly



End

Definition — Explain vs Interpret

Navigate Imprecise Language

- explain: refers to the process of computation of the explanation (e.g., attribution map)
- interpret: refers to the process of assigning a meaning to the explanation
- comprehend: refers to a deeper functional insight (e.g., takeaway) of the model