A Brief Overview to Interpretable Machine Learning

Isabel Valera
September 3, 2019

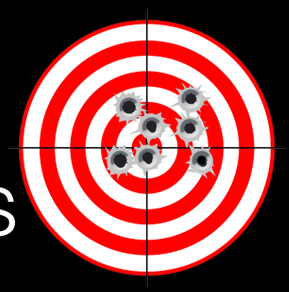
Slides by Amir Karimi

University admissions

Risk stratification for patients

Insurance policy assignment

Autonomous military systems



Applying for a loan.

Autonomous vehicles

Recidivism prediction

Predictive policing

Targeted Political Ads

Job hiring

Interpret: *to explain or to present in understandable terms [Merriam-Webster]*

Interpretability

Interpretability (ML): *ability to explain or to present in understandable terms to a human [Doshi-Velez & Kim, 2017]*

Explanations: the currency in which we exchange beliefs *[Lombrozo, 2006]*

University admissions

Risk stratification for patients

Applying for a loan

Autonomous vehicles

Job hiring

Mechanisms?

Definitions?

Interpretability (ML): *ability to explain or to present in understandable terms to a human [Doshi-Velez & Kim, 2017]*

Measures?

Stakeholders?

Insurance policy assignment

Targeted Political Ads

Predictive policing

Autonomous military systems

Recidivism prediction

Stakeholders

Researcher & Developer

Owner & Deployer

Examiner & Regulator

Data-subjects & Decision-subjects

Researcher & Developer

Owner & Deployer

Data-subjects & Decision-subjects

Examiner & Regulator

Verification:
build the system right

Trust

Validation:
build the right system

Debugging

Robustness

Safety

Improvement

Privacy

Fairness

Knowledge Discovery

Accountability

Simulatability

Interestingness

Decomposability

Explainability

Interpretability

Transparency

Definitions

Informativeness

Justifiability

Understandability

Comprehensibility

Intelligibility

Visibility

Legibility

Scrutability

Usabillity

Definitions

Interpretability

Transparency

(model interpretability)

(inherent interpretability)

the level to which a system provides information about its internal workings or structure, and the data it has been trained with

Explainability

(decision interpretability)

(post-hoc interpretability)

the level to which a system can provide clarification (explanations) for the cause of its decisions/outputs.

Zachary Lipton 2016

Been Kim, Finale Doshi-Velez 2017

Leilani H. Gilpin et al. 2018

Richard J. Tomsett et al. 2018

“The truth,
the whole truth,
and nothing but the truth”

Contrastive

Selective

Social

Measures

Functionally-grounded

Human-grounded

Application-grounded

Adrian Weller 2017

Finale Doshi-Velez, Been Kim 2017

Tim Miller 2018

Mechanisms

Transparency
(inherent/model interpretability)

Simulatability

Decomposability

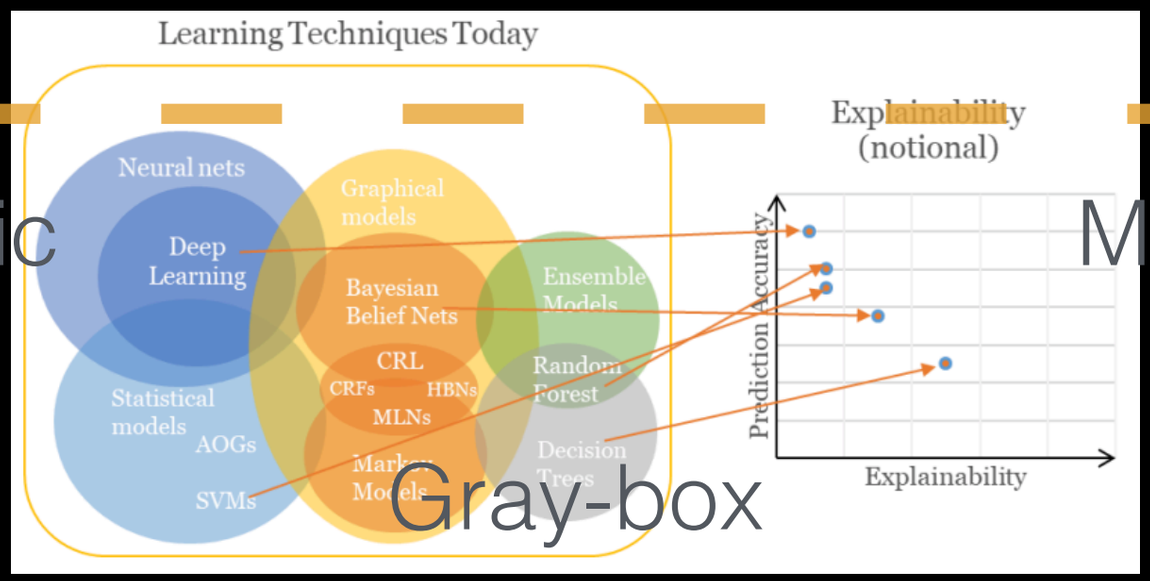
Algorithmic transparency

Explainability
(post-hoc/decision interpretability)

Feature-based (attribution)

Instance-based

Surrogate Models



Model-specific

Global

White-box

Gray-box

Model-agnostic

Local

Black-box

Transparency (Simulatability) Example:
Rule Sets

Objective — Build classifiers that are comprised of a small number of short rules. Rules are restricted to disjunctive normal form (DNF), e.g., if X satisfies (condition A AND condition B) OR (condition C) OR · · ·, then Y = 1

Related work — Greedy methods where rules are added to the model one by one, do not generally produce high-quality sparse models.

Example (rule selection) — Predicting if a customer will accept a coupon for a nearby coffee house, where the coupon is presented by their car's mobile recommendation device

if a customer (goes to coffee houses $\geq$ once per month AND destination = no urgent place AND passenger $\neq$ kids)

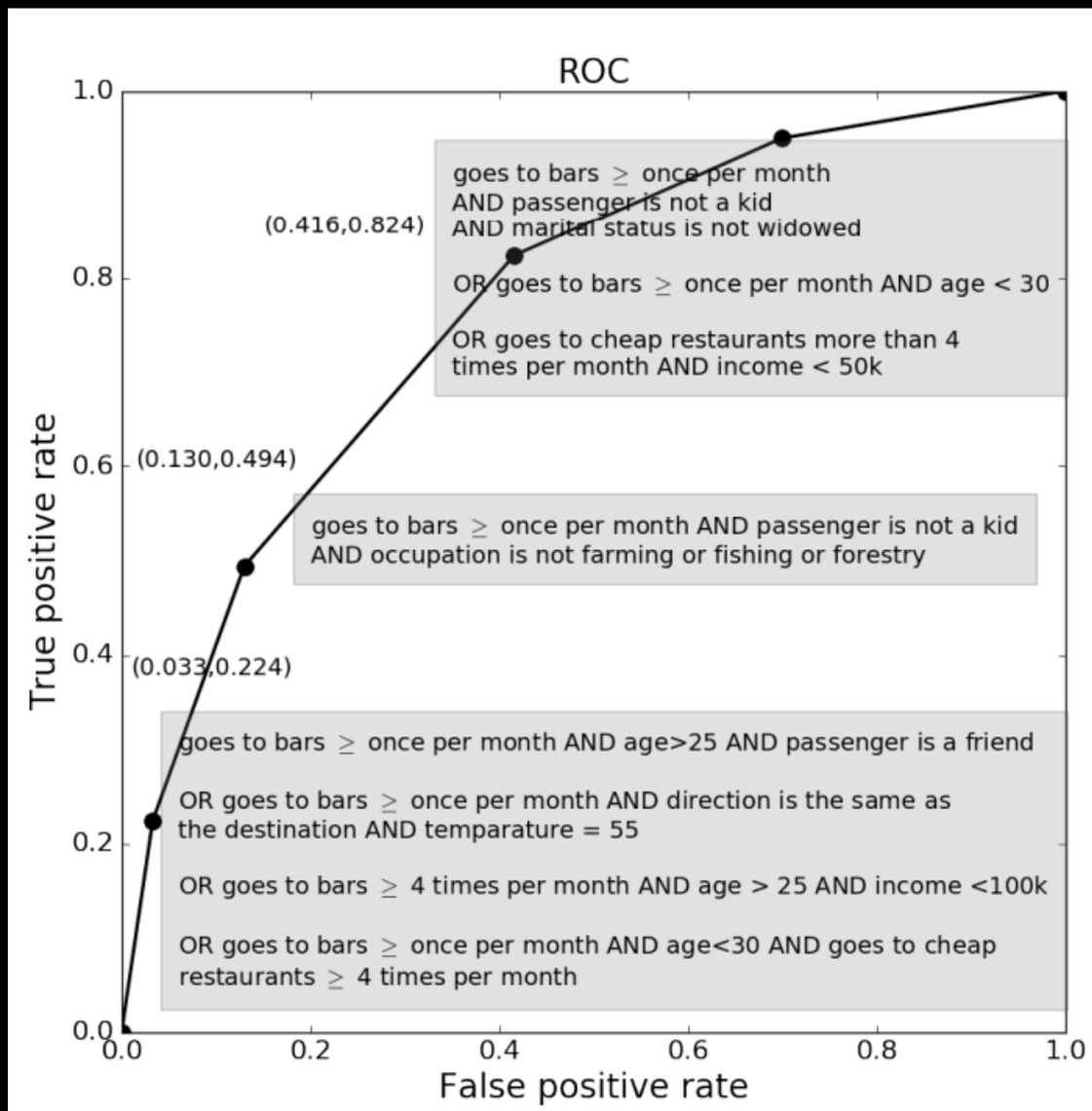
OR (goes to coffee houses $\geq$ once per month AND the time until coupon expires = one day)
then

predict the customer will accept the coupon for a coffee house.

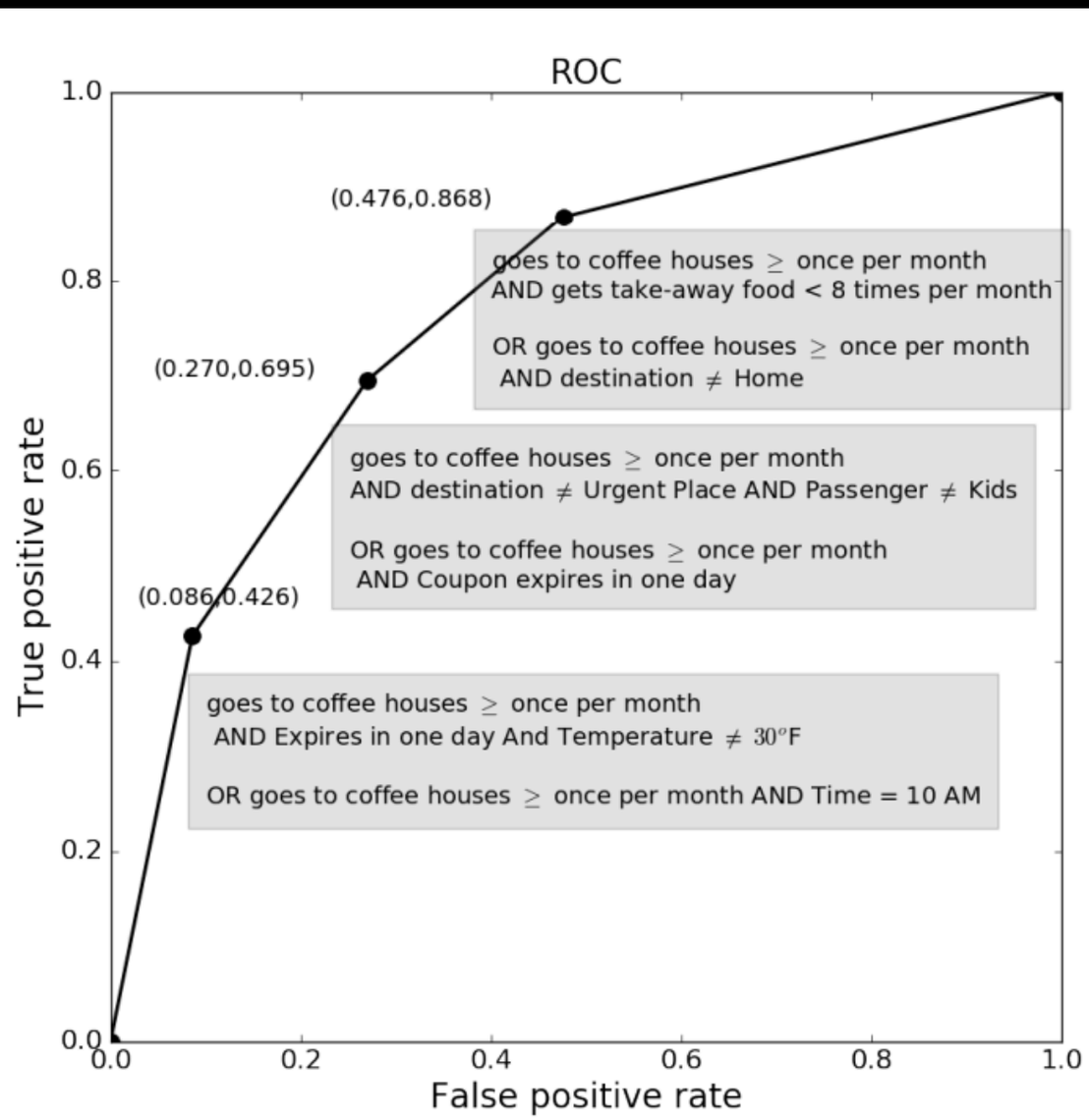
Wang et al. 2016

Transparency (Simulatability) Example: Rule Sets

Limitations



(a) Coupons for bars



(b) Coupons for coffee houses

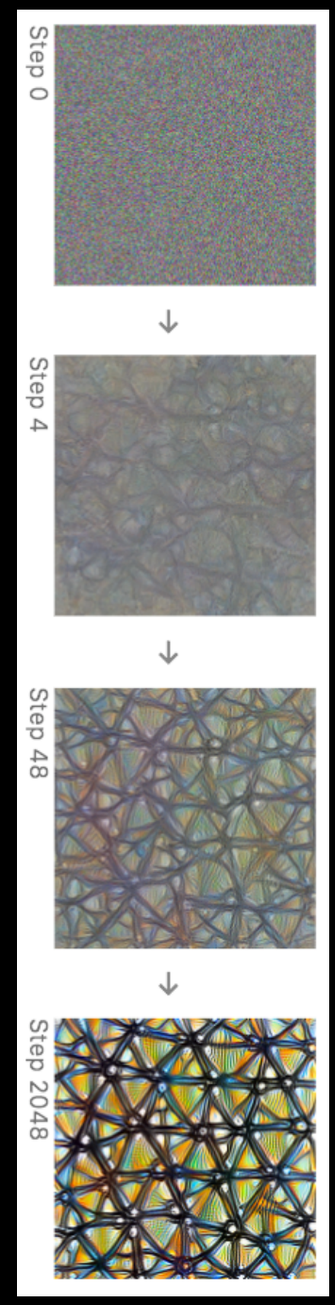
"Bayesian rule sets", Wang et al. 2016

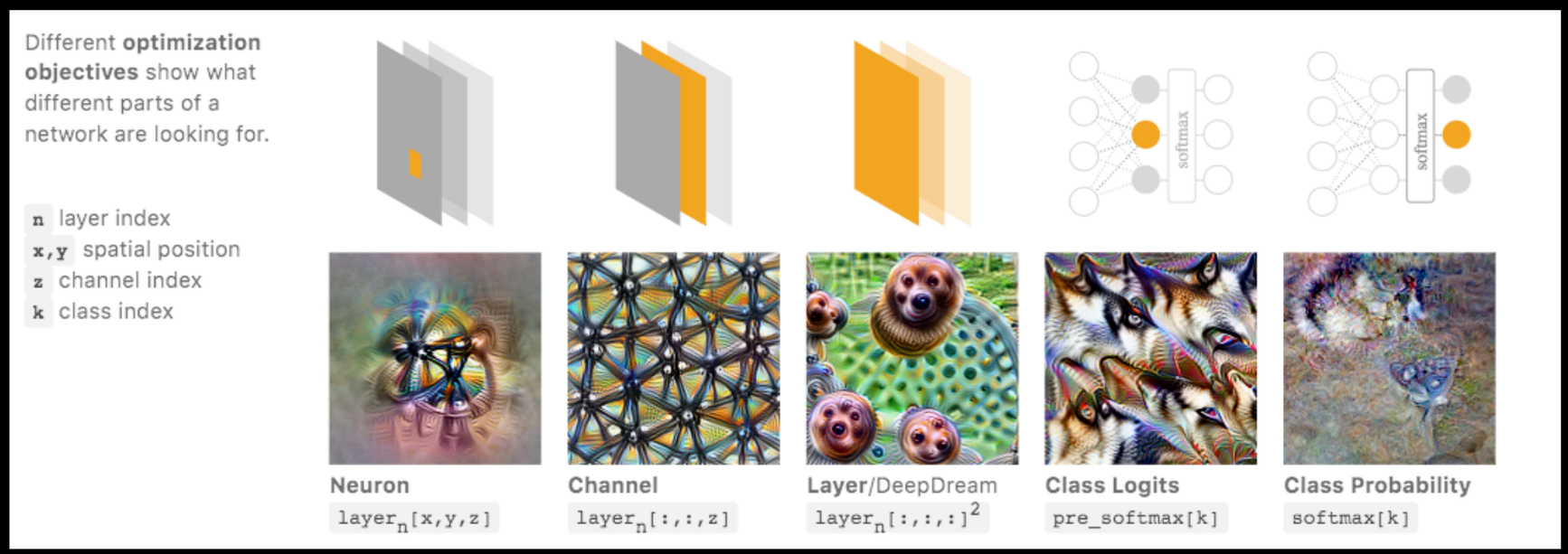
Transparency (Decomposability) Example:
Feature Visualization

Definition What is a unit looking for?

Visualization by optimization For a unit of a neural network, find the input that maximizes the activation of that unit.

$$I^* = \arg\max_I \sum \hat{f}_{n,x,y,z}(I)$$





Erhan et al. 2009
Springenberg et al. 2014
Olah et al. 2017
Nguyen et al. 2017
Molnar 2019

Transparency (Decomposability) Example: Feature Visualization

Limitations

- Many visualization images are not interpretable and lack human concepts
- Fails to describe complex inter-unit interactions
- There are too many units to consider
- Limited to CNNs for image recognition
- Lacking human semantical concepts.

Erhan et al. 2009
Springenberg et al. 2014
Olah et al. 2017
Nguyen et al. 2017
Molnar 2019

Explainability (Feature-based) Example: Attribution (Saliency Maps)

Definition How does the input affect the output?

Objective Identify exactly which regions of an image are being used for discrimination.

Score model for class c, near image I_0

$$S_c(\mathrm{I}) \approx w_c^T \mathrm{I} + b_c \qquad w = \left.\frac{\partial S_c}{\partial I}\right|_{I_0}$$

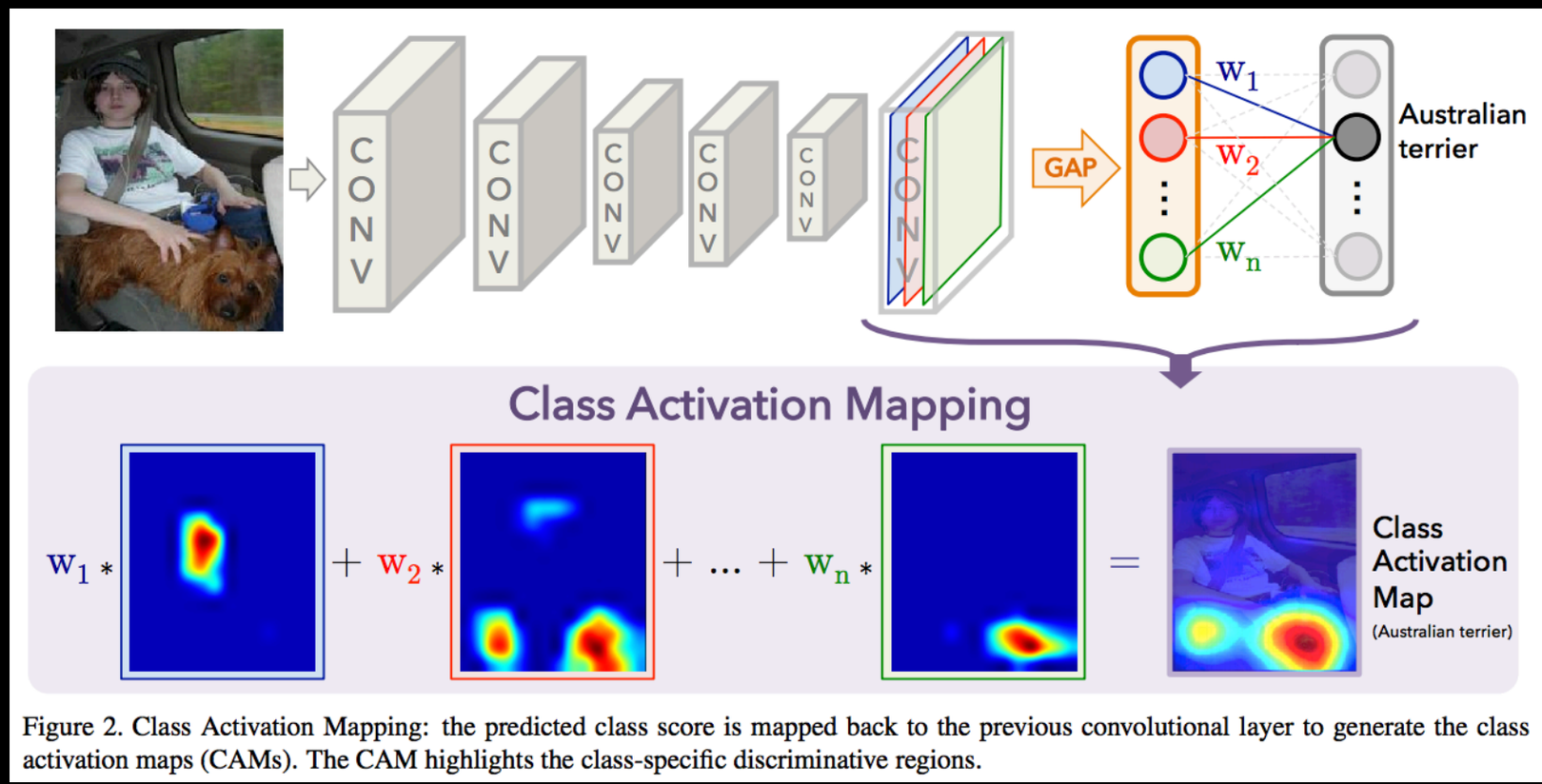


Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

Simonyan et al. 2013
Fong & Vedaldi 2017
Sundararajan et al. 2017
Kindermans et al. 2017

Transparency (Decomposability) Example: Feature Visualization

Limitations

- Only suitable for images.
- Several works have challenged the reliability of the provided explanations.

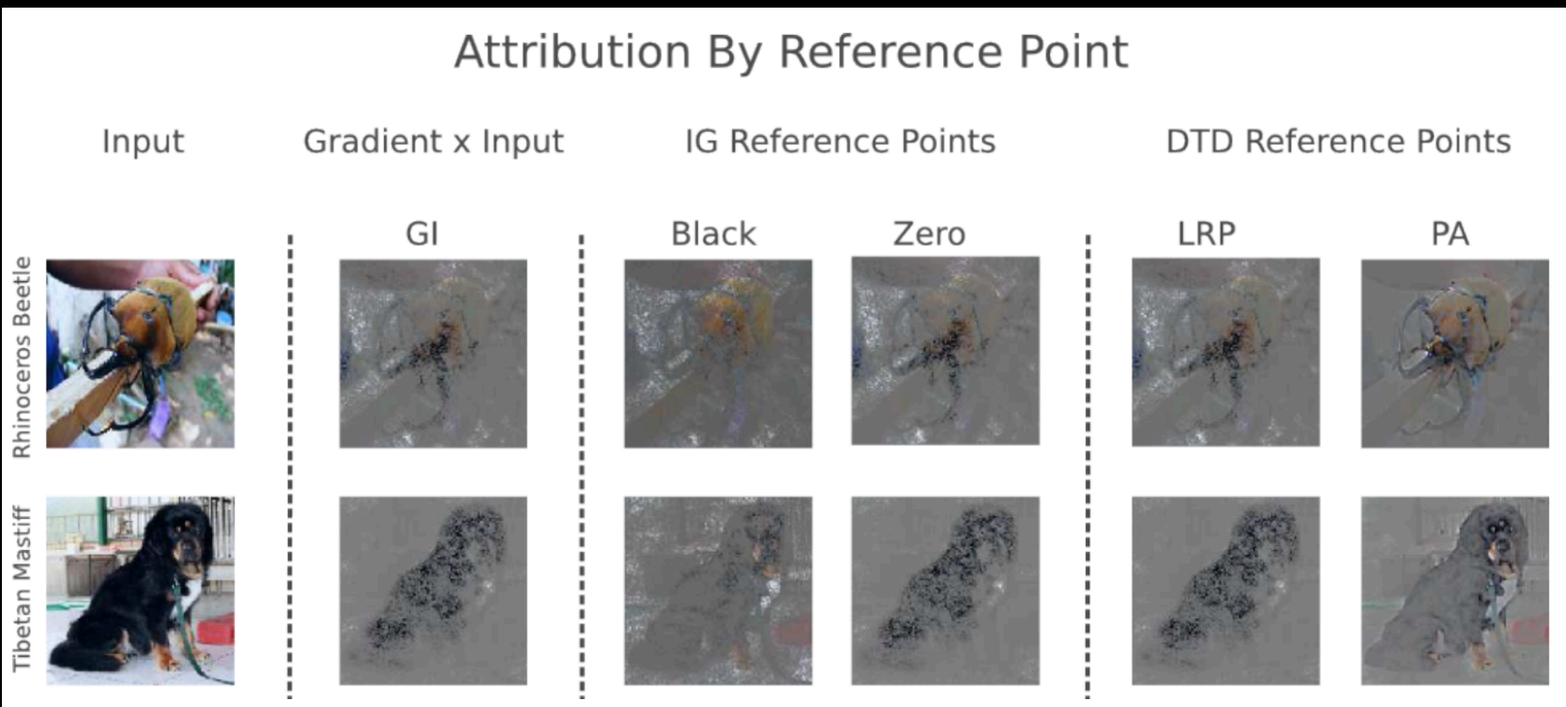


Figure 1: Integrated gradients and Deep Taylor Decomposition determine input attribution relative to a chosen reference point. This choice determines the vantage point for all subsequent attribution. Using two example reference points for each method we demonstrate that changing the reference causes the attribution to diverge. The attributions are visualized in a consistent manner with the IG paper (Sundararajan et al., 2017). Visualisations were made using ImageNet data. (Russakovsky et al., 2015) and the VGG16 architecture (Simonyan & Zisserman, 2015).

"The Unreliability of Saliency Methods" (Image) (Kindermans et al. 2017)

"Interpretation of Neural Networks is Fragile"(Ghorbani et al. 2017)

Explainability (Instance-based) Example: Counterfactual Explanations

Counterfactual explanation

"You were denied a loan because your annual income was £30,000. If your income had been £45,000, you would have been offered a loan."

Nearest counterfactual explanation

The set of features resulting in the desired prediction while remaining at min distance from the original set of features for the individual.

Additional considerations

Plausibility,

Diversity

$$x^* \in \underset{x^{\mathrm{CF}}}{\operatorname{argmin}} \quad d(x^{\mathrm{F}}, x^{\mathrm{CF}})$$

$$s.t. \quad f(x^{\mathrm{F}}) \neq f(x^{\mathrm{CF}})$$

$$x^{\mathrm{CF}} \in \mathcal{P}\mathrm{lausible}$$

Sandra Wachter et al. 2017

Methods for Generating Counterfactual Explanations

Counterfactual Explanations without Opening the Black Box - Wachter et al. 2017

Interpretable Predictions of Tree-based Ensembles via Feature Tweaking - Tolomei et al. 2017

Actionable Recourse in Linear Classification - Ustan et al. 2018

Minimum Observable Counterfactuals - Google PAIR team 2019

Model Agnostic Counterfactual Explanations for Consequential Decisions - Karimi et al. 2019

Limitations of current methods:

Lacking closeness guarantees

Linear / convex models

Limited coverage

Differentiable distance metrics

Homogenous data spaces

Ignore dependencies between input features

Still a lot to do...

- Definitions and measurements of explainability and interpretability
- Objective comparison between different approaches
- Explanations suitable for different tasks and datasets (beyond classification of images)
- Account for dependencies between input features (correlations, cofounders, causal graphs, etc.)
- Include human semantical concepts
- Connections to robustness, fairness and other ethical aspects of ML

Thank you!