INVICTA Keynote Series — Explainable AI, P2

Anna Hedström, PhD candidate, TU Berlin

March 18-22, 2024

Porto, Portugal

@anna_hedstroem
@UMI_Lab_AI

Today's agenda

01 Foundations

02 Evaluation

03 Quantus (hands-on)

04 Discussion + Q&A

Lecture 2

Evaluation

Evaluation — Scope

The Challenge of Explanation Evaluation

1. **Motivation —** Why it is important (to all) and interesting (as a research problem)

2. **Methods —** What are the current methods and pitfalls; human, approximate and restriction

3. **Meta-Evaluation —** How to estimate explanation quality, reliably

4. **Summary**

Evaluation

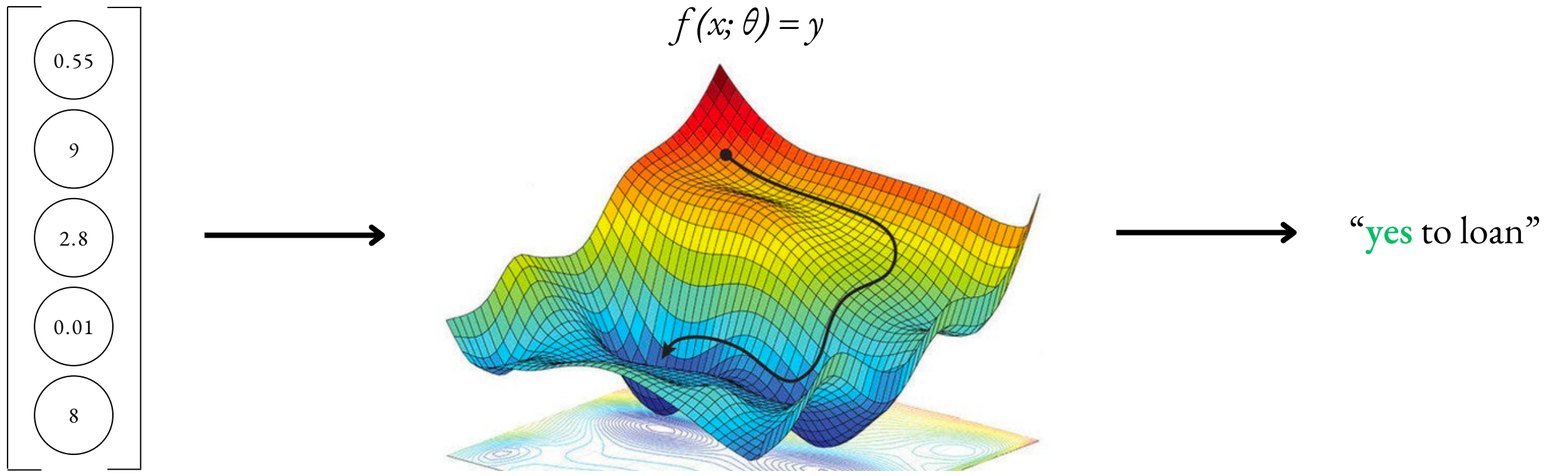
Motivation

The Challenge of Explanation Evaluation

Motivation — 1. Modelling

The Challenge of Explanation Evaluation

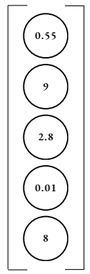
- Imagine you have a model that neither developers, companies or users understand

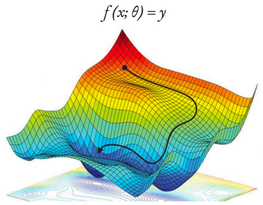


Motivation — 2. Explaining

The Challenge of Explanation Evaluation

- To "interpret" the model (and, for compliance), we apply explainable methods



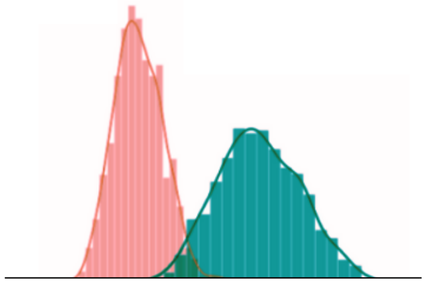


"yes to loan"

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

IntegratedGradients

SHapley Additive exPlanations



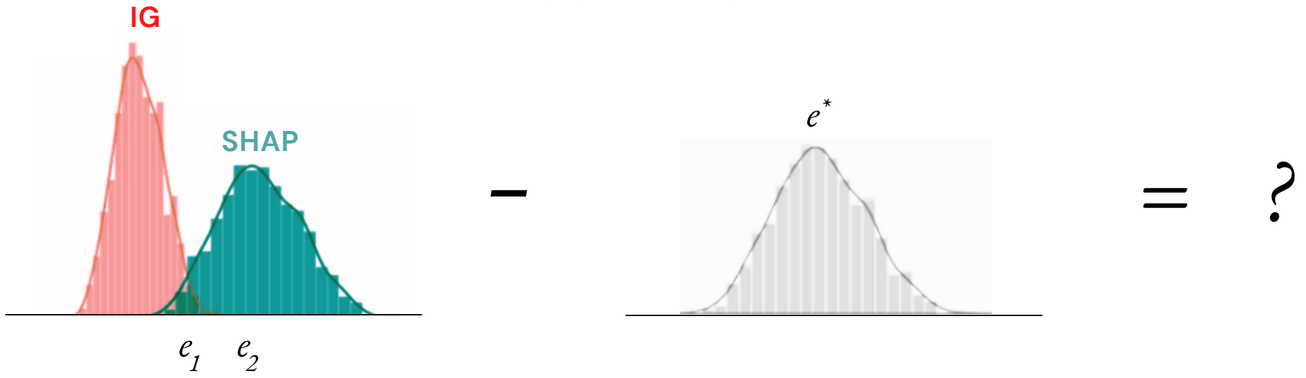
e_i

Motivation — 3. Evaluating

The Challenge of Explanation Evaluation

- While selecting an explanation method for a user, we face evaluation difficulties

$$\Psi(\Phi(x, f, y; \lambda); \tau) = q$$



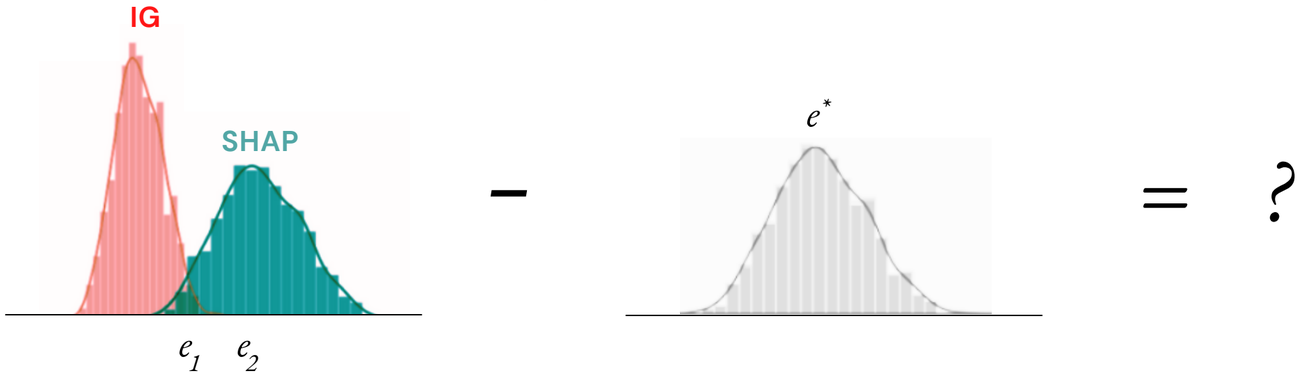
- Without ground truth explanation labels, we can't compute "explanation error"

Motivation — 3. Evaluating

The Challenge of Explanation Evaluation

- While selecting an explanation method for a user, we face evaluation difficulties

$$\Psi(\Phi(x, f, y; \lambda); \tau) = q$$



- Without ground truth explanation labels, we can't compute "explanation error"

The lack of ground truth explanation labels e* forms the The Evaluation Problem in Explainable AI.

Motivation — Why

The Challenge of Explanation Evaluation

Without access to labels, standard error measures do not apply, with uncertainty propagating to the model space

Important

- Real-world risks include presenting unqualified explanation outcomes to end users, misinterpreting the model, and deploying flawed models in practice

Interesting

- Underspecification (varying or lack of constraints, non-uniqueness), unsupervised, elusive definition

Motivation — Why

The Challenge of Explanation Evaluation

Without access to labels, standard error measures do not apply, with uncertainty propagating to the model space

Important

- Real-world risks include presenting unqualified explanation outcomes to end users, misinterpreting the model, and deploying flawed models in practice

Interesting

- Underspecification (varying or lack of constraints, non-uniqueness), unsupervised, elusive definition

Motivation — Why

The Challenge of Explanation Evaluation

Without access to labels, standard error measures do not apply, with uncertainty propagating to the model space

Important

- Real-world risks include presenting unqualified explanation outcomes to end users, misinterpreting the model, and deploying flawed models in practice

Interesting

- Underspecification (varying or lack of constraints, non-uniqueness), unsupervised, elusive definition

Evaluation

Methods

Methods — Overview

Estimating Explanation Quality

How can we circumvent the lack of ground truth explanation labels but still evaluate quality of φ?

Three options, evaluation by:

1. Human(s)
2. Restriction
3. Approximation

Methods — Overview

Estimating Explanation Quality

How can we circumvent the lack of ground truth explanation labels but still evaluate quality of PHI?

Three options, evaluation by:

1. Human(s)
2. Restriction
3. Approximation

Methods — Humans

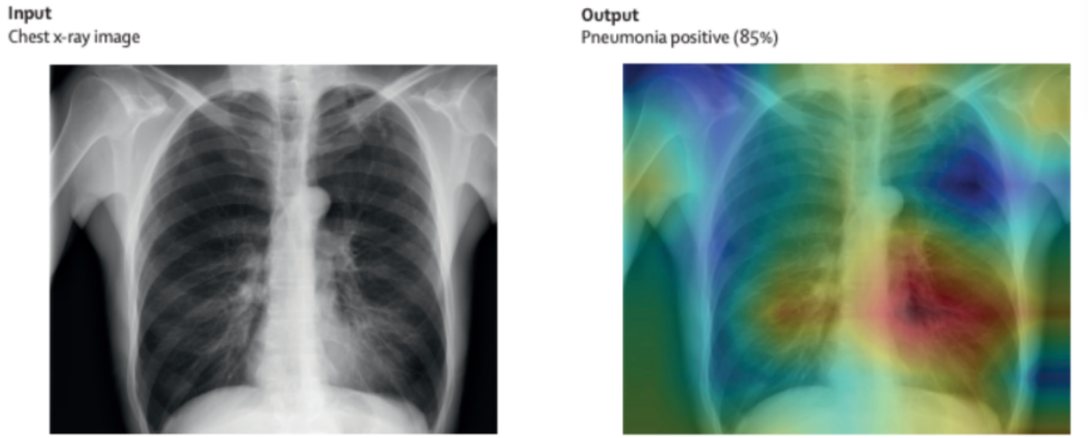
Estimating Explanation Quality

- User studies are common practice (esp. healthcare) and acts complementary to metrics-based quality estimation
- But can humans (with all our cognitive biases) identify a “correct” explanation?

Methods — Humans

Estimating Explanation Quality

- Both data scientists (Kaur et al., 2020) and healthcare professionals (Ghassemi et al., 2021) tend to misinterpret and over-trust AI explanations



Normative vs descriptive evidence:
Is the presence of (i), (ii) or (iii) that contributed to the prediction:

(i) airspace opacity,
(ii) the shape of the heart border or,
(iii) the left artery?

Image Source.

Methods — Overview

Estimating Explanation Quality

How can we circumvent the lack of ground truth explanation labels but still evaluate quality of PHI?

Three options, evaluation by:

1. Human(s)
2. Restriction
3. Approximation

Methods — Restriction

Estimating Explanation Quality

- Recall our evaluation function:

$$\Psi(\Phi(x, f, y; \lambda); \tau) = q$$

- What if we constrain the involved space(s) and evaluate Φ against known or simulated ground truths?

Methods — Restriction

Estimating Explanation Quality

- Recall our evaluation function:

$$\Psi(\Phi(x, f, y; \lambda); \tau) = q$$

- **Model space** e.g., evaluate against coefficients of self-interpretable models (GAMs, NAMs etc) (Rudin 2019; Carmichael et al., 2023), i.e., variants of the linear model

Equation for Generative Additive Model (GAM).

Methods — Restriction

Estimating Explanation Quality

- Recall our evaluation function:

$$\Psi(\Phi(\underline{x}, f, y; \lambda); \tau) = q$$

- **Data space** e.g., introduce explicit artefacts into certain classes of the dataset and evaluate how the explainable evidence u (Yang, 2019)

Methods — Restriction

Estimating Explanation Quality

- **Evaluation idea:** filter out relevant features per class (orange) and measure union with acceptable features (purple) (Zhou et al., 2021)

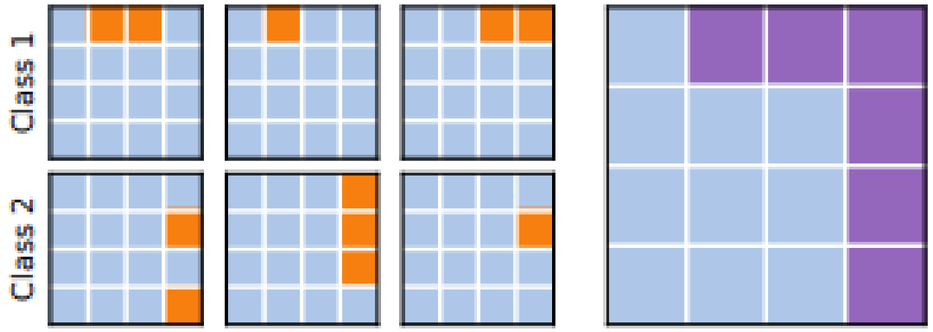


Image Source.

Methods — Overview

Estimating Explanation Quality

How can we circumvent the lack of ground truth explanation labels but still evaluate quality of PHI?

Three options, evaluation by:

1. Human(s)
2. Restriction
3. Approximation

Methods — Restriction

Estimating Explanation Quality

- Recall our evaluation function:

$$\Psi(\Phi(x, f, y; \lambda); \tau) = q$$

- **Explanation space** e.g., restricting the explainable algorithm such that it satisfies predefined axiomatic principles e.g, SHAP and Integrated Gradients

Methods — Approximation

Estimating Explanation Quality

- We accept that ground truth cannot be known and measure the fulfilment of desirable properties or unit tests i.e., metrics-based quality estimation
- Instead, we can evaluate ~~indirectly by approximation~~

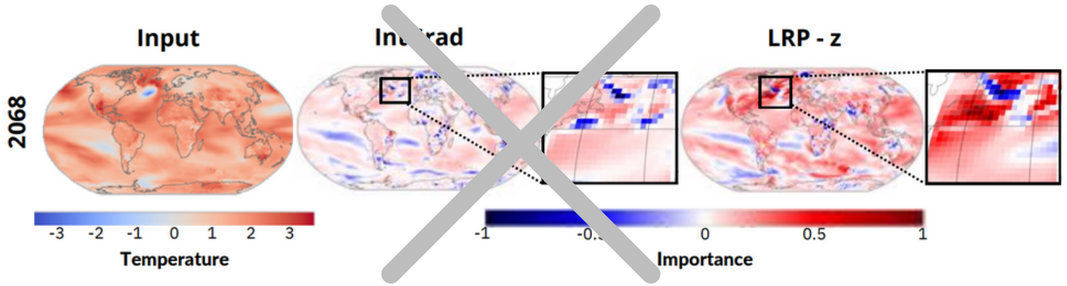
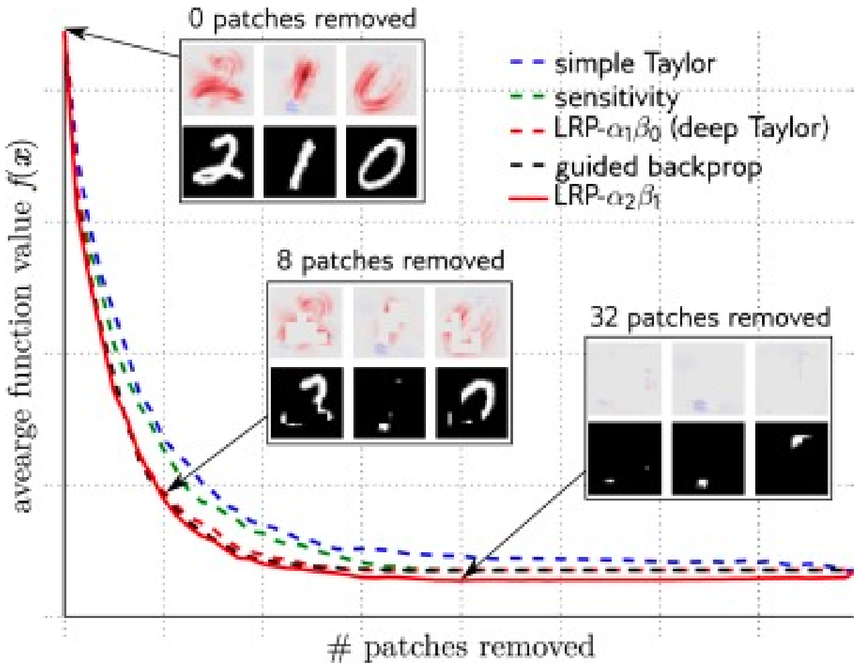


Image Source.

How to define the quality of the explanation method, without access to ground truth explanation labels?

Methods — Approximation

Estimating Explanation Quality — Faithfulness



$$\Psi_{\text{PF}} = \sum_{i=1}^{n} (\hat{y}_i + \hat{y}_{i+1}) \cdot \frac{p_{i+1} - p_i}{2}$$

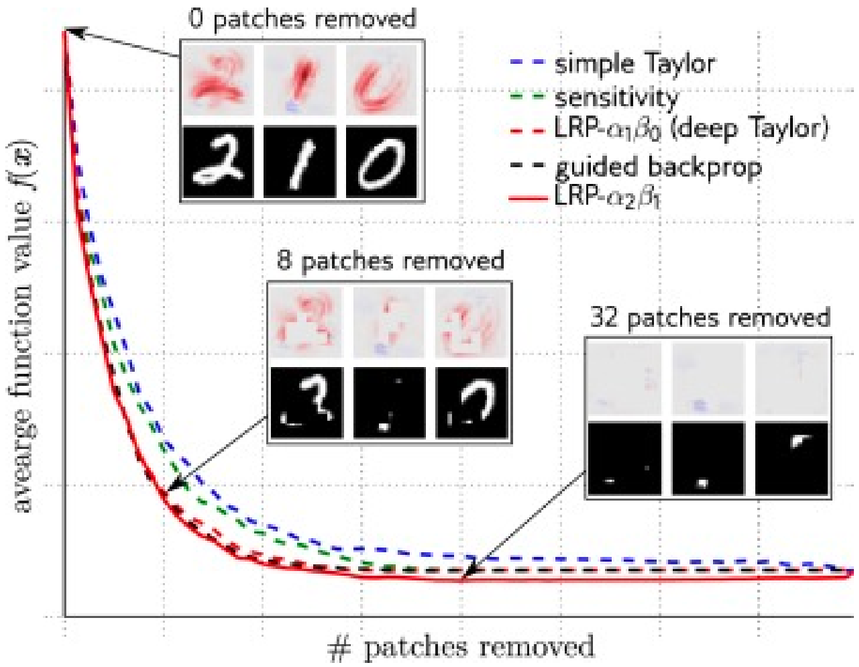
Pixel-Flipping (Bach et al., 2015): captures the model degradation while perturbing pixels in the input based on the explanation output

$$\Psi_{\text{FC}} = \underset{S \in |S| \subseteq d}{\text{corr}} \left(\sum_{i \in S} \Phi(\boldsymbol{x}, f, \hat{y}; \lambda)_i, f(\boldsymbol{x}) - f\left(\boldsymbol{x}_{[\boldsymbol{x}_s = \overline{\boldsymbol{x}}_s]}\right) \right)$$

Faithfulness Correlation (Bhatt et al., 2020): measures the correlation between attribution sums and prediction change, given random masking of the input

Methods — Approximation

Estimating Explanation Quality — Faithfulness



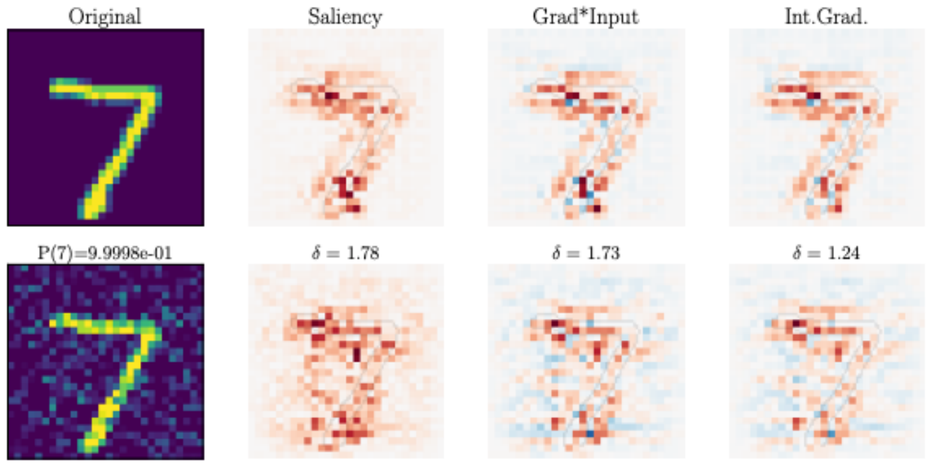
$$\Psi_{\mathrm{PF}} = \sum_{i=1}^{n} (\hat{y}_i + \hat{y}_{i+1}) \cdot \frac{p_{i+1} - p_i}{2}$$

Pixel-Flipping (Bach et al., 2015): captures the model degradation while perturbing pixels in the input based on the explanation output

Pitfall 1: Input-Dependent Parameterisation
RQ: How do we choose patch size and masking value to avoid out-of-distribution creation?

Methods — Approximation

Estimating Explanation Quality — Faithfulness

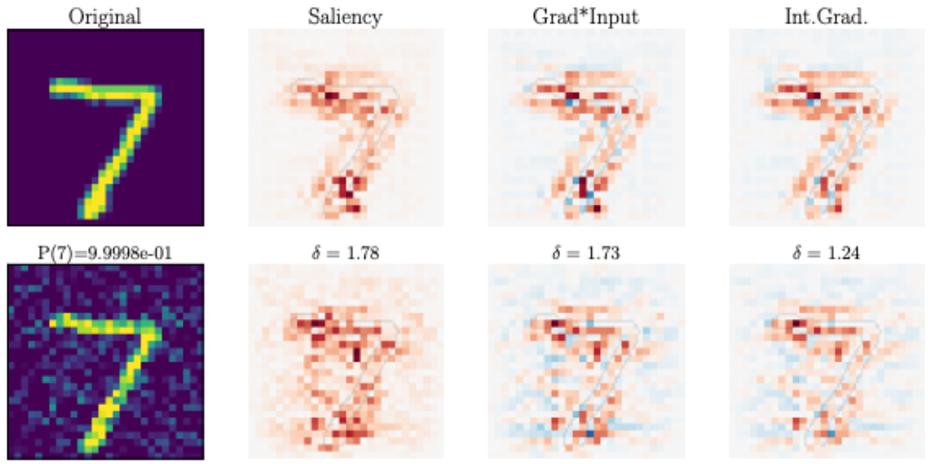


$$\Psi_{\text{LLE}} = \max_{\boldsymbol{x}+\delta \in \mathcal{N}_\epsilon(\boldsymbol{x}) \leq \epsilon} \frac{\|\Phi(\boldsymbol{x}, f, \hat{y}; \lambda) - \Phi(\boldsymbol{x}+\delta, f, \hat{y}; \lambda)\|_2}{\|\boldsymbol{x} - (\boldsymbol{x}+\delta)\|_2},$$

Local Lipschitz Estimate (LLE) (Alvarez-Melis et al., 2018): measures how much the explanation changes for the input under slight perturbation

Methods — Approximation

Estimating Explanation Quality — Faithfulness



Pitfall 2: Ignore Explanation-to-Model Misalignment

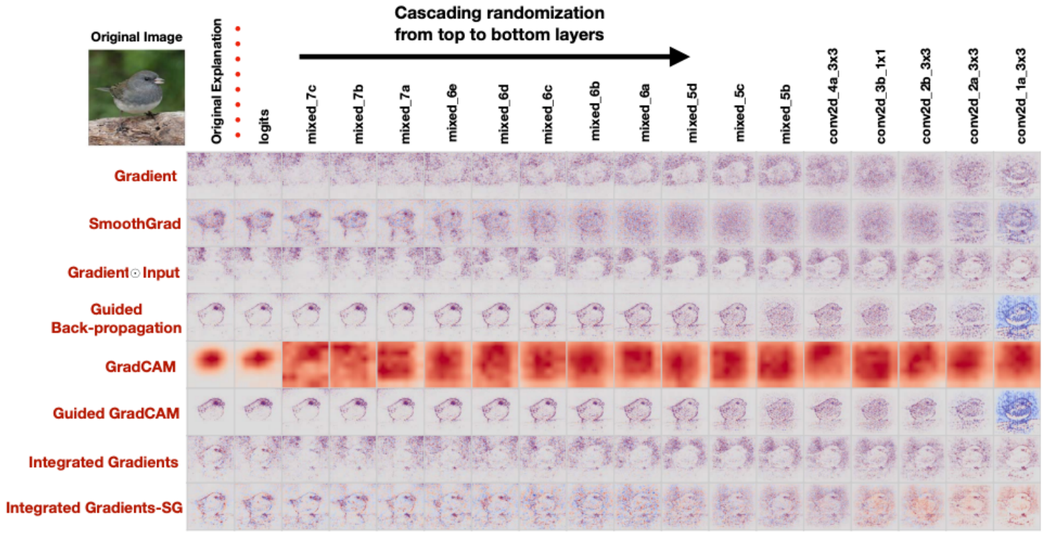
RQ: How so we ensure that the expectation of explanation response is grounded in model response?

$$\Psi_{\text{LLE}} = \max_{\boldsymbol{x}+\delta \in \mathcal{N}_\epsilon(\boldsymbol{x}) \leq \epsilon} \frac{\|\Phi(\boldsymbol{x}, f, \hat{y}; \lambda) - \Phi(\boldsymbol{x}+\delta, f, \hat{y}; \lambda)\|_2}{\|\boldsymbol{x} - (\boldsymbol{x}+\delta)\|_2},$$

Local Lipschitz Estimate (LLE) (Alvarez-Melis et al., 2018): measures how much the explanation changes for the input under slight perturbation

Methods — Approximation

Estimating Explanation Quality — Sensitivity

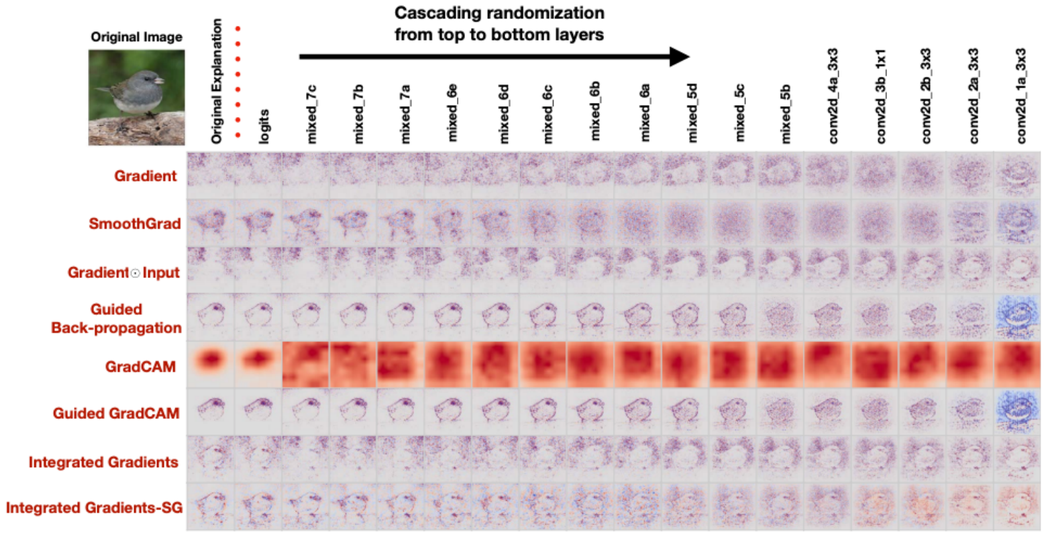


$$\Psi_{\text{MPR}} = \frac{1}{V}\sum_{v=1}^{V} \text{corr}(\Phi^v(\boldsymbol{x}, f, \hat{y}; \lambda), \Phi^v(\boldsymbol{x}, \hat{f}, \hat{y}; \lambda)),$$

Model Parameter Randomisation Test (Adebayo et. al., 2018): randomises the parameters of single model layers top-down way, measuring the distance of the respective explanation to the original explanation

Methods — Approximation

Estimating Explanation Quality — Sensitivity



Pitfall 3: Empirical Confounds

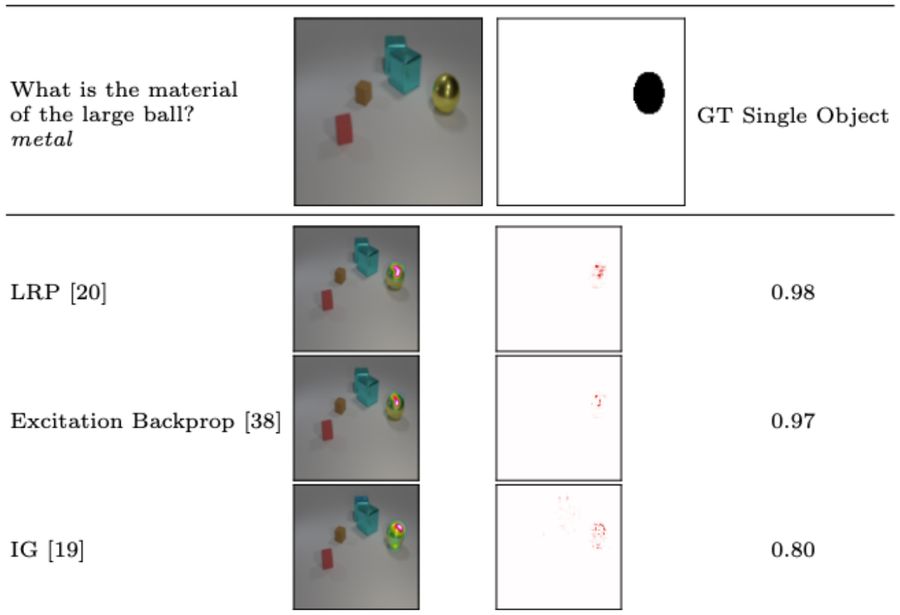
RQ: How to conduct this test without "confounds" wrt randomisation order, similarity measure and preprocessing?

$$\Psi_{\text{MPR}} = \frac{1}{V}\sum_{v=1}^{V} \text{corr}(\Phi^v(\boldsymbol{x}, f, \hat{y}; \lambda), \Phi^v(\boldsymbol{x}, \hat{f}, \hat{y}; \lambda)),$$

Model Parameter Randomisation Test (Adebayo et. al., 2018): randomises the parameters of single model layers top-down way, measuring the distance of the respective explanation to the original explanation

Methods — Approximation

Estimating Explanation Quality — Localisation

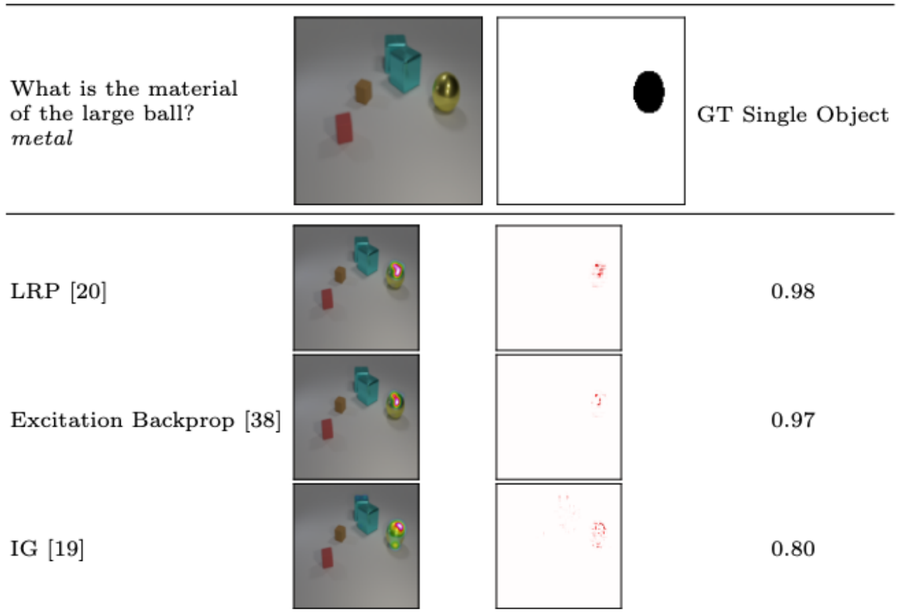


$$\Psi_{\text{RMA}} = \frac{\sum_{i=1}^{D} \Phi_i(\boldsymbol{x}, f, \hat{y}; \lambda) \cdot \boldsymbol{s}_{gt,i}}{\sum_{i=1}^{D} \Phi_i(\boldsymbol{x}, f, \hat{y}; \lambda)},$$

Relevance Mass Accuracy (RMA) (Arras et al., 2022) quantifies the fraction of the sum of the attribution that intersects with the ground truth mask over the full explanation sum

Methods — Approximation

Estimating Explanation Quality — Localisation



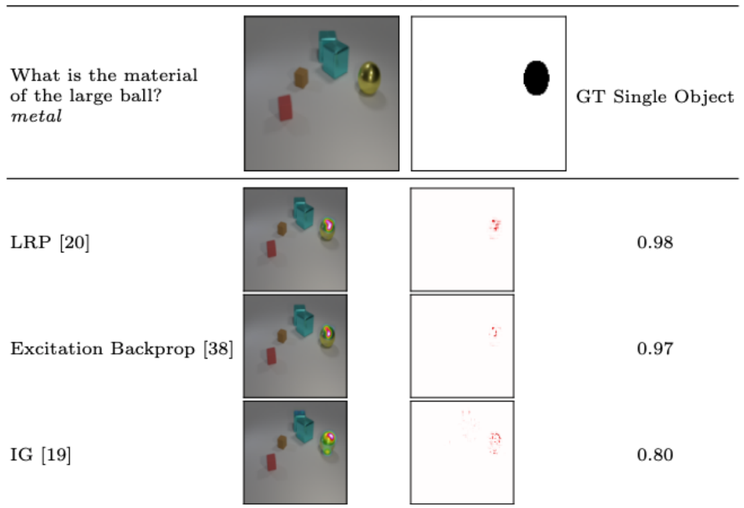
"There is no apriori reason why the model should use the same features as humans"

$$\Psi_{\text{RMA}} = \frac{\sum_{i=1}^{D} \Phi_i(\boldsymbol{x}, f, \hat{y}; \lambda) \cdot \boldsymbol{s}_{gt,i}}{\sum_{i=1}^{D} \Phi_i(\boldsymbol{x}, f, \hat{y}; \lambda)},$$

Relevance Mass Accuracy (RMA) (Arras et al., 2022) quantifies the fraction of the sum of the attribution that intersects with the ground truth mask over the full explanation sum

Methods — Approximation

Estimating Explanation Quality — Localisation



Pitfall 4: Unverified "Ground Truth"

RQ: How can we verify that segmentation masks reflect model-learned behavior, avoiding confirmation bias?

$$\Psi_{\text{RMA}} = \frac{\sum_{i=1}^{D} \Phi_i(\boldsymbol{x}, f, \hat{y}; \lambda) \cdot \boldsymbol{s}_{gt,i}}{\sum_{i=1}^{D} \Phi_i(\boldsymbol{x}, f, \hat{y}; \lambda)},$$

Relevance Mass Accuracy (RMA) (Arras et al., 2022) quantifies the fraction of the sum of the attribution that intersects with the ground truth mask over the full explanation sum

Methods — Approximation

Estimating Explanation Quality — Complexity

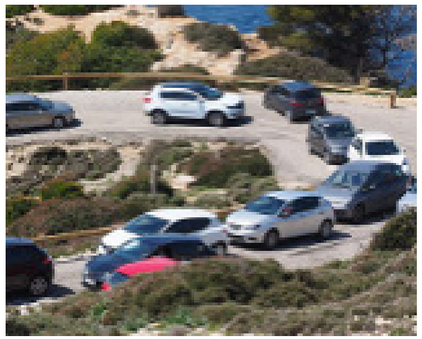


Image Source.

$$\Psi_{\mathrm{CO}} = \mathbb{E}_i \left[-\ln \left(\mathbb{P}_\Phi \right) \right] = -\sum_{i=1}^{D} \mathbb{P}_\Phi(i) \ln \left(\mathbb{P}_\Phi(i) \right)$$

$$\text{with} \quad \mathbb{P}_\Phi(i) = \frac{|\Phi_i(\boldsymbol{x}, f, \hat{y}; \lambda)|}{\sum_{j \in [d]} |\Phi_j(\boldsymbol{x}, f, \hat{y}; \lambda)|}; \mathbb{P}_\Phi = \{\mathbb{P}_\Phi(1), \ldots, \mathbb{P}_\Phi(d)\},$$

Complexity (Bhatt et al., 2020): computes the entropy of the fractional contribution of all features to the total magnitude of the attribution individually

Methods — Approximation

Estimating Explanation Quality — Complexity

Pitfall 5: Determine Score Thresholds

RQ: How can thresholds be predefined without human bias and with regard to their underlying task?

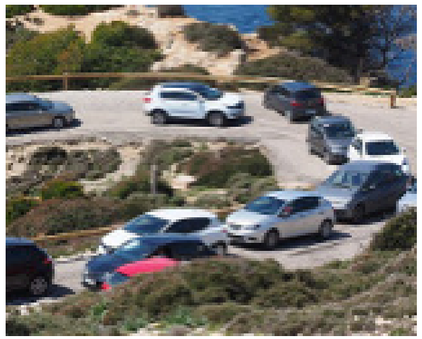


Image Source.

$$\Psi_{\mathrm{CO}} = \mathbb{E}_i \left[-\ln \left(\mathbb{P}_\Phi \right) \right] = -\sum_{i=1}^{D} \mathbb{P}_\Phi(i) \ln \left(\mathbb{P}_\Phi(i) \right)$$

$$\text{with} \quad \mathbb{P}_\Phi(i) = \frac{|\Phi_i(\boldsymbol{x}, f, \hat{y}; \lambda)|}{\sum_{j \in [d]} |\Phi_j(\boldsymbol{x}, f, \hat{y}; \lambda)|}; \mathbb{P}_\Phi = \{\mathbb{P}_\Phi(1), \ldots, \mathbb{P}_\Phi(d)\},$$

Complexity (Bhatt et al., 2020): computes the entropy of the fractional contribution of all features to the total magnitude of the attribution individually

How can some of these pitfalls be addressed?

Evaluation

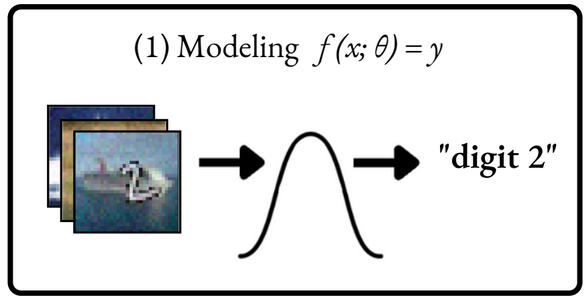
Meta-Evaluation

Without ground truth, how can we identify a reliable estimator of explanation quality?

MetaQuantus — Motivation

The Meta-Evaluation Problem in Explainable AI

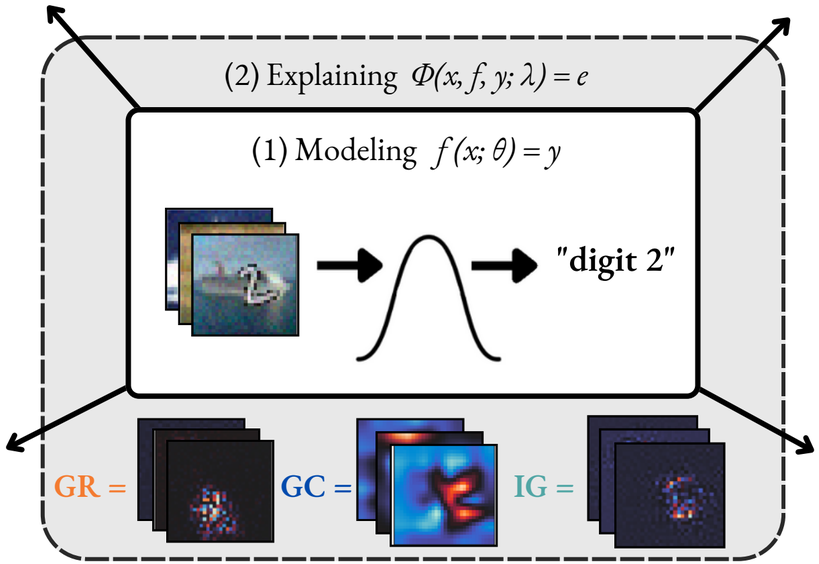
- Imagine you trained a black-box model *f* given some input *x* and labels *y*



MetaQuantus — Motivation

The Meta-Evaluation Problem in Explainable AI

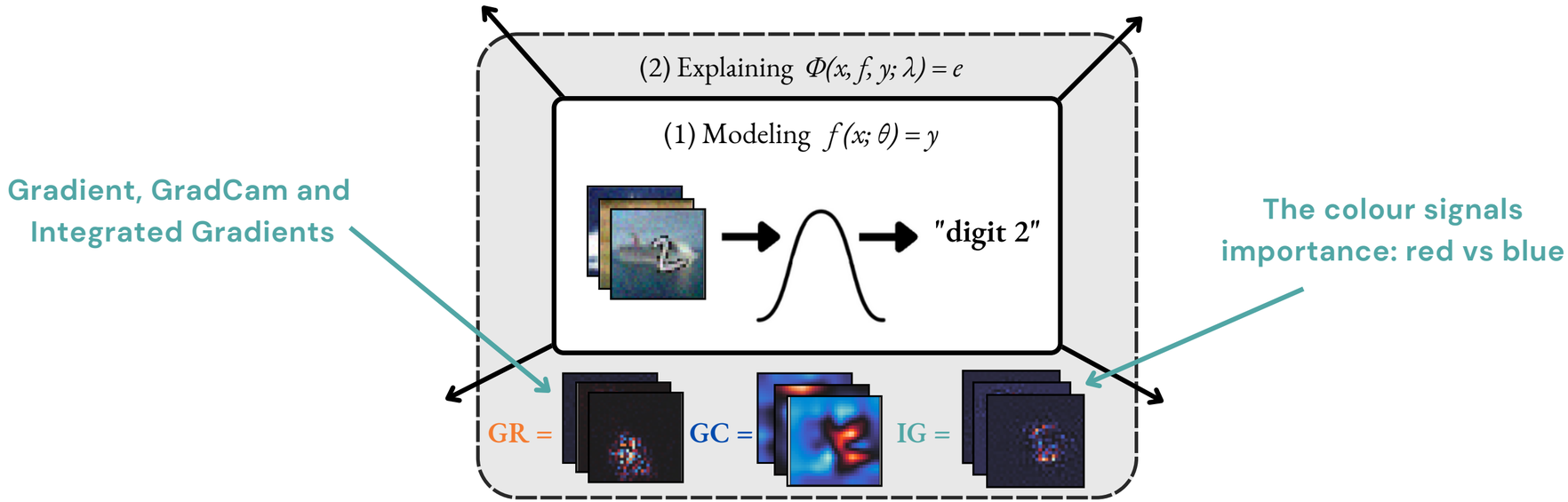
- To interpret a prediction, we apply explainers to approximate feature importance



MetaQuantus — Motivation

The Meta-Evaluation Problem in Explainable AI

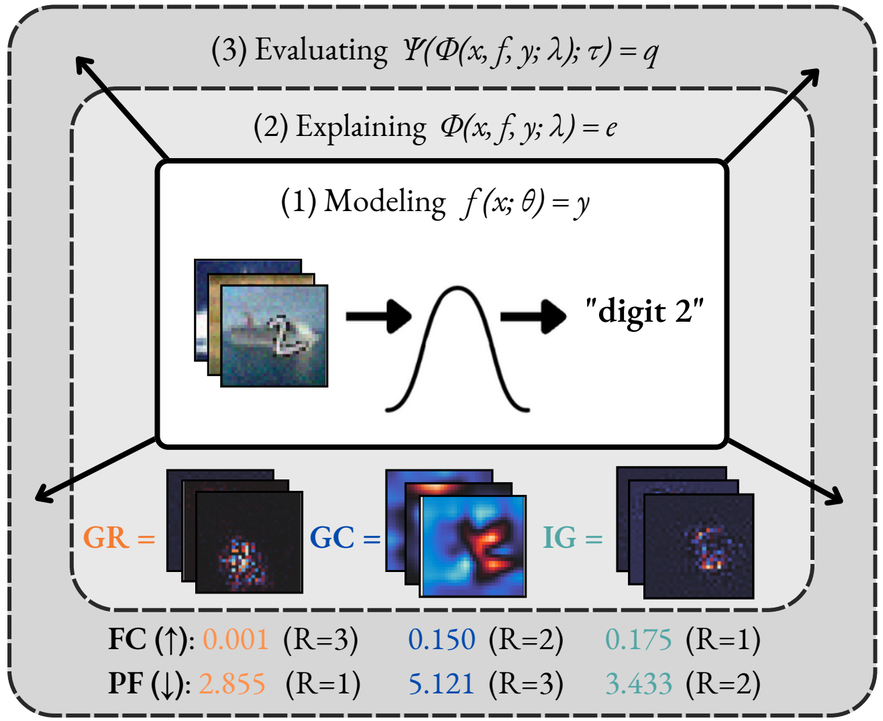
- To interpret a prediction, we apply explainers to approximate feature importance



MetaQuantus — Motivation

The Meta-Evaluation Problem in Explainable AI

- The explanations remain “uninterpretable”, so we evaluate their quality

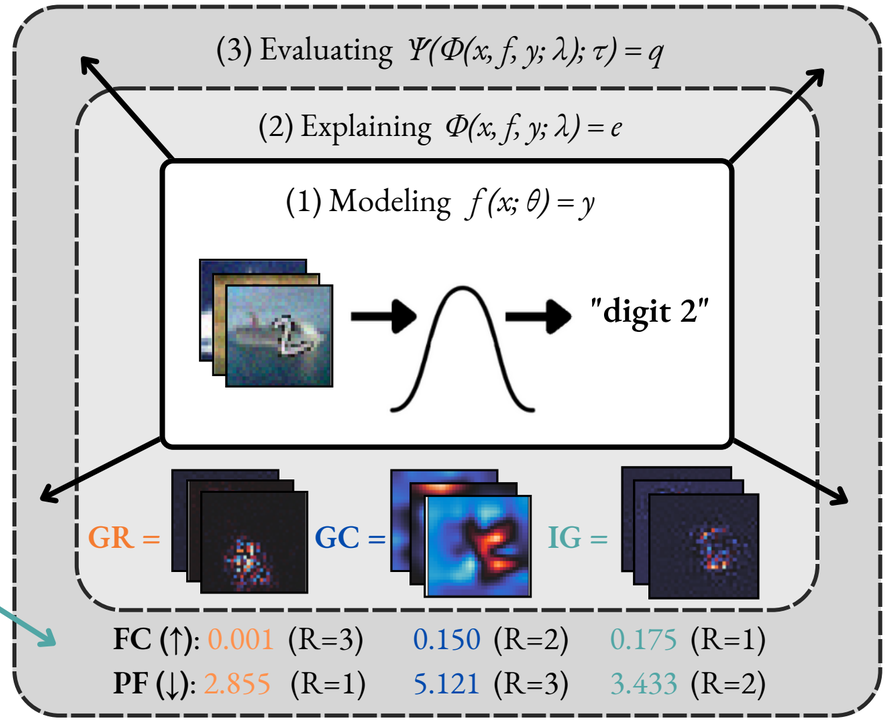


MetaQuantus — Motivation

The Meta-Evaluation Problem in Explainable AI

- The explanations remain “uninterpretable”, so we evaluate their quality

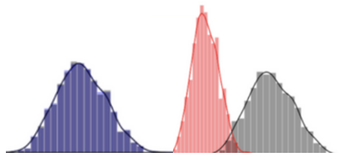
We apply two different faithfulness metrics



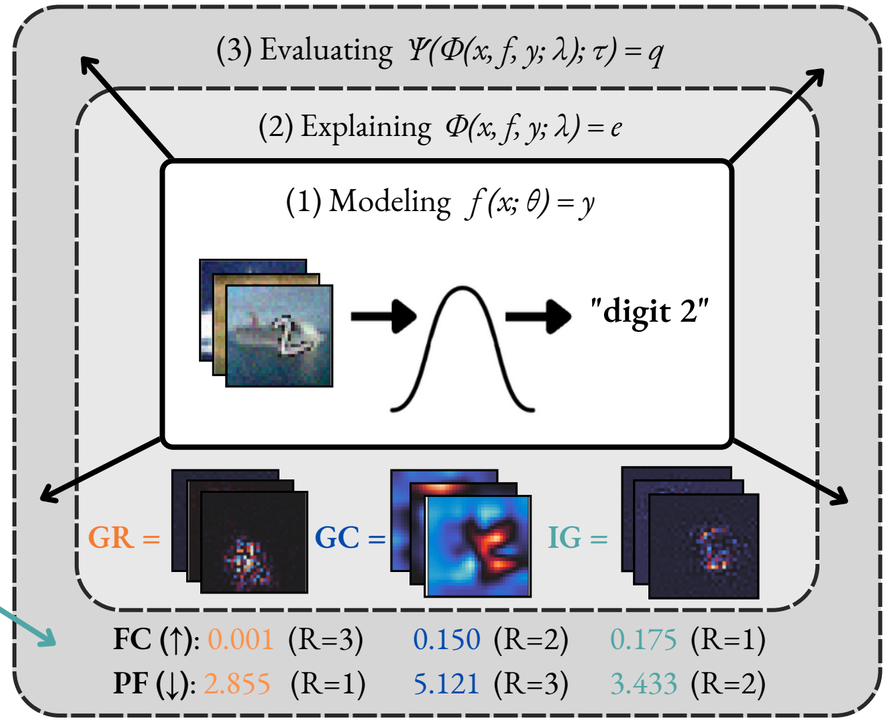
MetaQuantus — Motivation

The Meta-Evaluation Problem in Explainable AI

- The explanations remain “uninterpretable”, so we evaluate their quality



They differ not only in score interpretation (↑ , ↓) and range

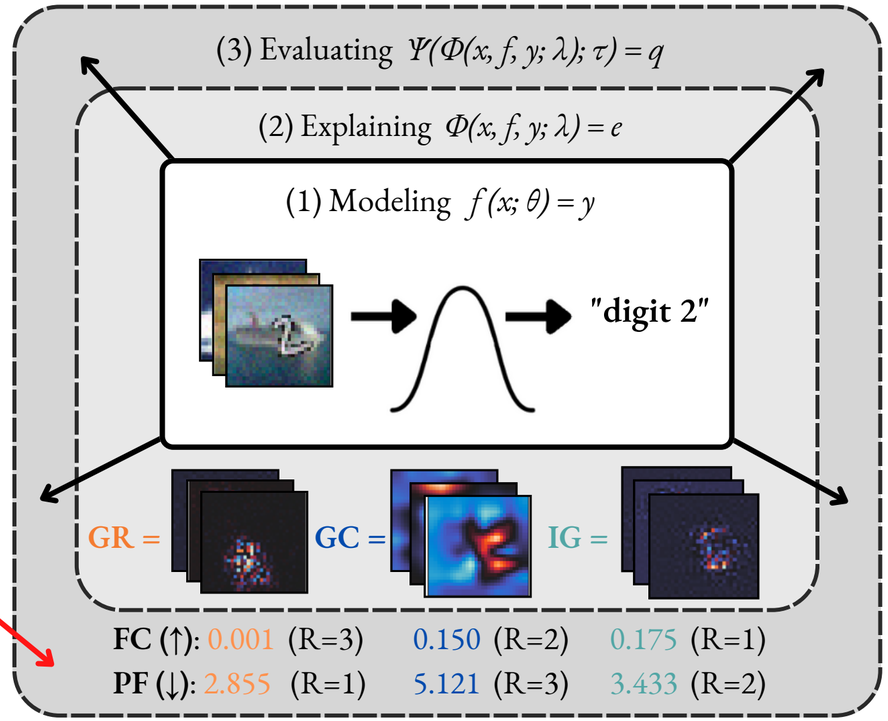


MetaQuantus — Motivation

The Meta-Evaluation Problem in Explainable AI

- The explanations remain “uninterpretable”, so we evaluate their quality

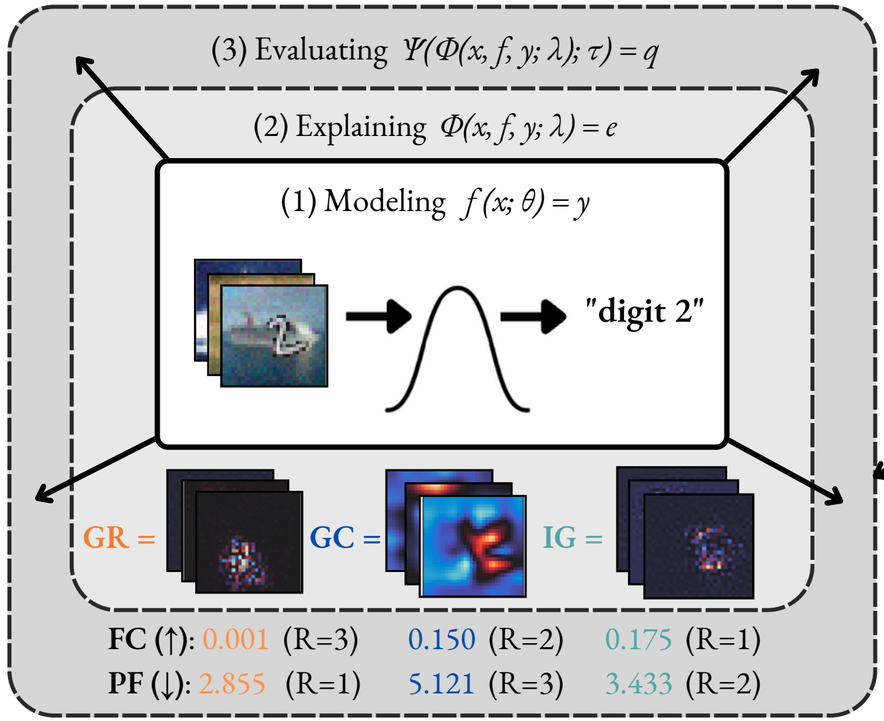
But there is a disagreement in their explanation rankings (!)



MetaQuantus — Motivation

The Meta-Evaluation Problem in Explainable AI

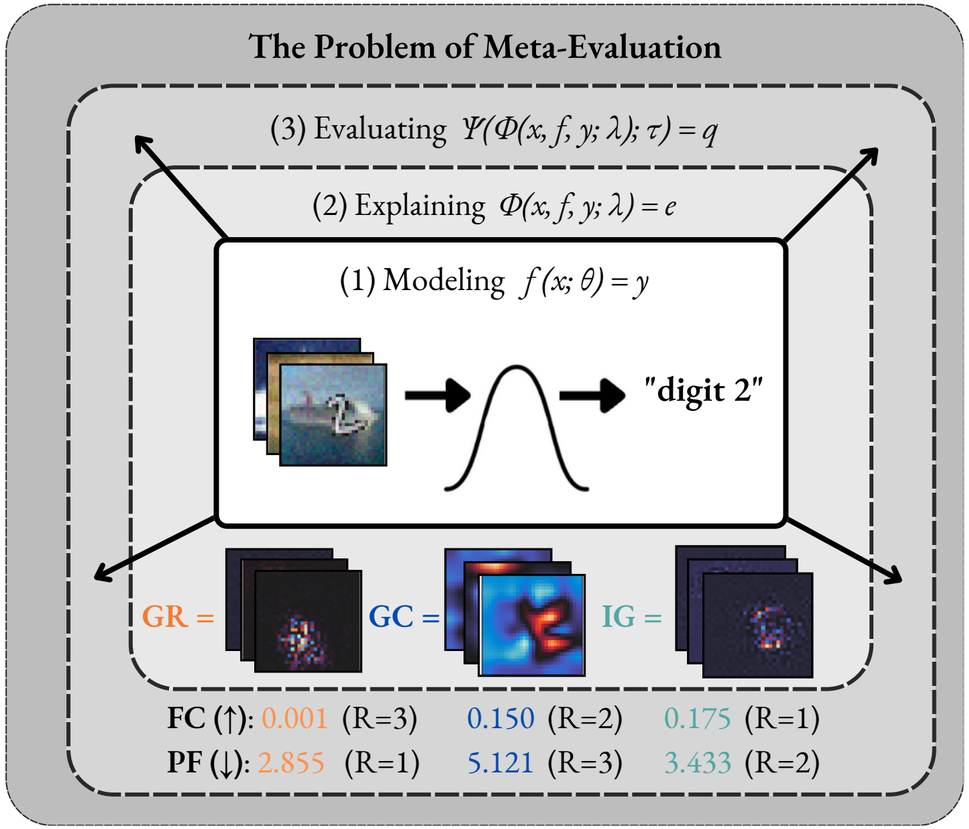
- The explanations remain “uninterpretable”, so we evaluate their quality



Q: Which explanation method of {GR, GC, IG} has the highest and lowest quality, respectively?

MetaQuantus — Motivation

The Meta-Evaluation Problem in Explainable AI



MetaQuantus — Definition

The Meta-Evaluation Problem in Explainable AI

"Meta-evaluation is the process of evaluating the evaluation method."

- The objective is to identify a reliable estimator of explanation quality (metric) to avoid the risk of presenting an unqualified explanation method to the end-user

- But why is it challenging?

MetaQuantus — Definition

The Meta-Evaluation Problem in Explainable AI

"Meta-evaluation is the process of evaluating the evaluation method."

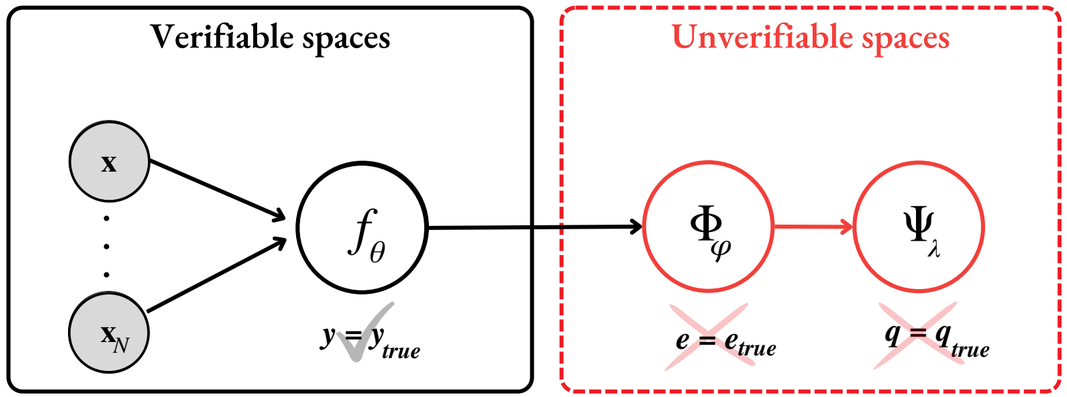
- The objective is to identify a reliable estimator of explanation quality (metric) to avoid the risk of presenting an unqualified explanation method to the end-user

- But why is it challenging?

MetaQuantus — Unverifiability

Unlabeled Explanations Lead to Unverifiability

- We separate the spaces into verifiable (with labels) and unverifiable (without labels)

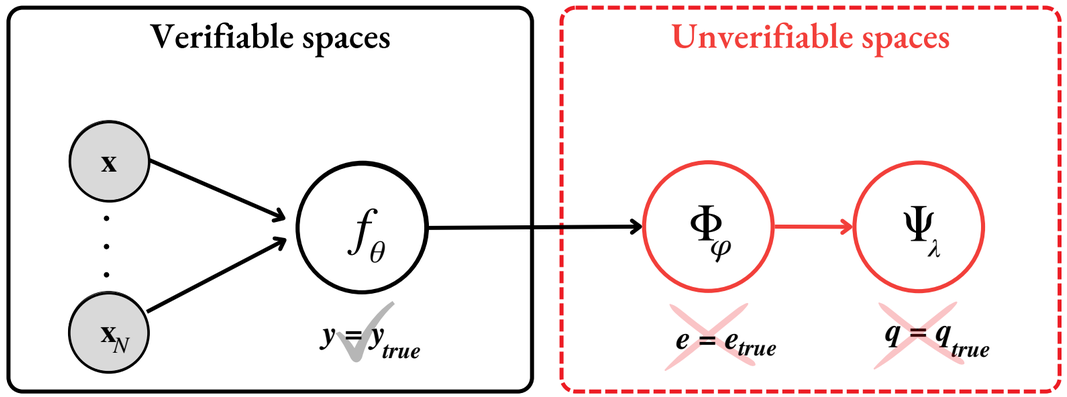


- The DAG shows the conditional dependencies and how uncertainty propagates: if a parent node is unverifiable, then its child node renders unverifiable

MetaQuantus — Unverifiability

Unlabeled Explanations Lead to Unverifiability

- We separate the spaces into verifiable (with labels) and unverifiable (without labels)

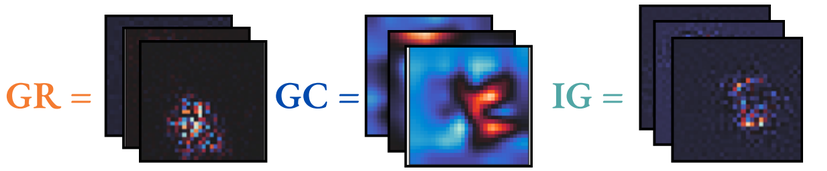


- The DAG shows the conditional dependencies and how uncertainty propagates: if a parent node is unverifiable, then its child node renders unverifiable

MetaQuantus — Solution

How to identify a statistically reliable evaluation metric?

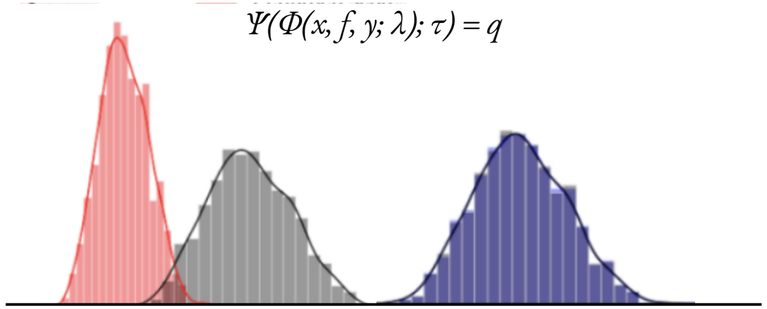
$\Phi(x, f, y; \lambda) = e$



Explanation outcomes

?

$\Psi(\Phi(x, f, y; \lambda); \tau) = q$



Quality estimates

?

MetaQuantus — Solution

How to identify a statistically reliable evaluation metric?

- Knowing that we cannot compute the accuracy of a metric, instead, we compute a metric's **consistency given controlled perturbations** (with verified strength)
- Inspired by the evaluation of explainers, we list properties that a metric should fulfil:
  - Failure mode 1: **Resilient to (minor) noise** (NR)
  - Failure mode 2: **Reactive to (disruptive) adversaries** (AR)
- … and assign expectations to the quality estimator's i) score distribution and ii) ranking

MetaQuantus — Solution

How to identify a statistically reliable evaluation metric?

- Knowing that we cannot compute the accuracy of a metric, instead, we compute a metric's **consistency given controlled perturbations** (with verified strength)
- Inspired by the evaluation of explainers, we list properties that a metric should fulfil:
  - Failure mode 1: **Resilient to (minor) noise** (NR)
  - Failure mode 2: **Reactive to (disruptive) adversaries** (AR)
- … and assign expectations to the quality estimator's i) score distribution and ii) ranking

MetaQuantus — Solution

How to identify a statistically reliable evaluation metric?

- Knowing that we cannot compute the accuracy of a metric, instead, we compute a metric's **consistency given controlled perturbations** (with verified strength)
- Inspired by the evaluation of explainers, we list properties that a metric should fulfil:
  - Failure mode 1: **Resilient to (minor) noise** (NR)
  - Failure mode 2: **Reactive to (disruptive) adversaries** (AR)
- … and assign expectations to the quality estimator's i) score distribution and ii) ranking

MetaQuantus — Method: Step 1

- Perturbations are applied on the input and model space: {IPT, MPT}

Step 1. Perturbing

Depending on failure mode, initiate a minor or disruptive perturbation

$\mathcal{P}_{\Omega}^{M}$ Noise Resilience

$\mathcal{P}_{\Omega}^{D}$ Adversary Reactivity

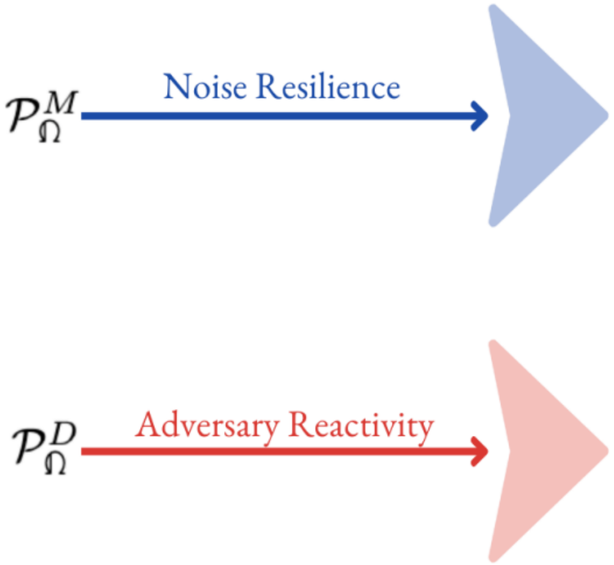
- The perturbation strength is verified through the model response

Hedström et al., *"The Meta-Evaluation Problem in Explainable AI: Identifying Reliable Estimators with MetaQuantus"* TMLR (2023)

MetaQuantus — Method: Step 1

Step 1. Perturbing

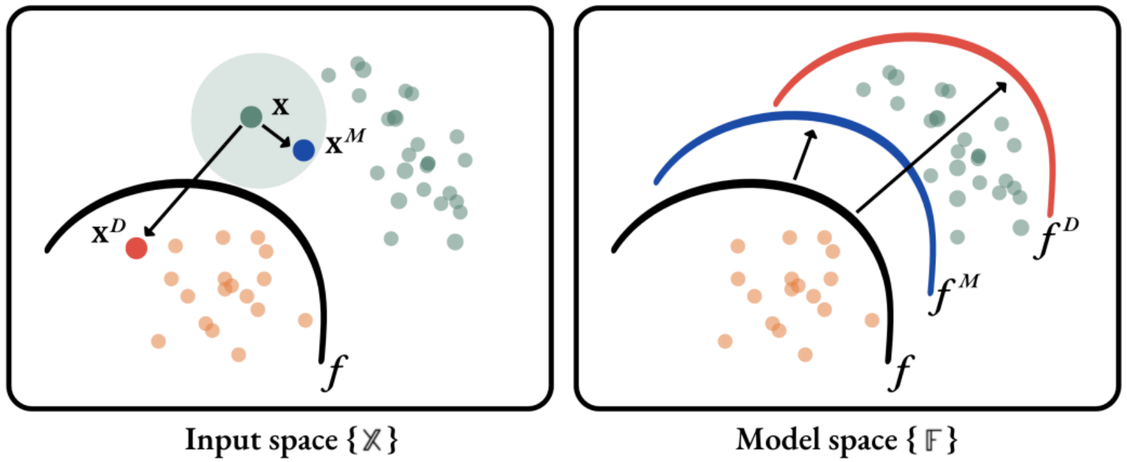
Depending on failure mode, initiate a minor or disruptive perturbation



- Perturbations are applied on the input and model space: {IPT, MPT}

$$\hat{q} \in \{\ \Psi(\Phi, \mathcal{P}^t_{\mathbb{X}}(\boldsymbol{x}), f, \hat{y}),\ \ \Psi(\Phi, \boldsymbol{x}, \mathcal{P}^t_{\mathbb{F}}(\theta), \hat{y})),\ \ \Psi(\Phi, \mathcal{P}^t_{\mathbb{X}}(\boldsymbol{x}), \mathcal{P}^t_{\mathbb{F}}(\theta), \hat{y}))\ \},$$

- The perturbation strength is verified through the model response



Hedström et al., *"The Meta-Evaluation Problem in Explainable AI: Identifying Reliable Estimators with MetaQuantus"* TMLR (2023)

MetaQuantus — Method: Step 2

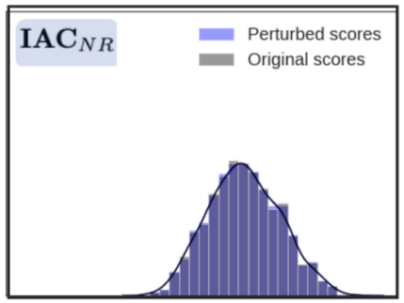
Step 2. Scoring

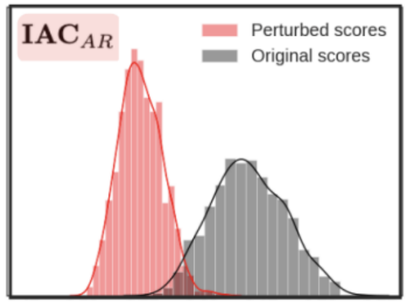
Measure effects of the perturbations via IAC and IEC criteria

$$\mathbf{IAC} = \frac{1}{K}\sum_{k=1}^{K} d(\hat{\boldsymbol{q}}, \boldsymbol{q}'_k), \quad (5)$$

$L =$ 

$$\mathbf{IEC} = \frac{1}{N \times L}\sum_{i=1}^{N}\sum_{j=1}^{L} U^t_{i,j} \quad (6)$$





$\mathbf{IEC}_{NR}$

$$\hat{\boldsymbol{q}} = [\ 3,\ 2,\ 1\ ]$$
$$\boldsymbol{q}'_k = [\ 3,\ 2,\ 1\ ] \cdots K$$
$$\vdots$$
$$N \quad U^M_{i,j} = \begin{cases} 1 & \bar{r}^M_j = \bar{r}_j \\ 0 & \text{otherwise,} \end{cases} \quad (7)$$

$\mathbf{IEC}_{AR}$

$$\hat{\boldsymbol{q}} = [\ 0.6, 0.7, 0.2\ ]$$
$$\boldsymbol{q}'_k = [\ 0.3, 0.5, 0.1\ ] \cdots K$$
$$\vdots$$
$$N \quad U^D_{i,j} = \begin{cases} 1 & \bar{Q}^D_{i,j} < \bar{Q}_{i,j} \\ 0 & \text{otherwise,} \end{cases} \quad (8)$$

MetaQuantus — Method: Step 3

Step 2. Scoring

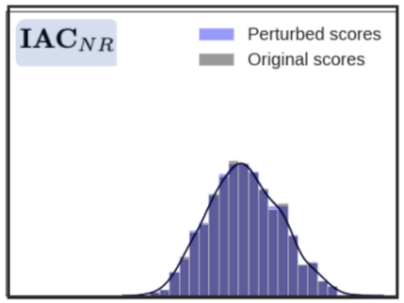
Measure effects of the perturbations via IAC and IEC criteria

$$\mathbf{IAC} = \frac{1}{K}\sum_{k=1}^{K} d(\hat{\boldsymbol{q}}, \boldsymbol{q}'_k), \quad (5)$$

$L =$ 

$$\mathbf{IEC} = \frac{1}{N \times L}\sum_{i=1}^{N}\sum_{j=1}^{L} U^t_{i,j} \quad (6)$$

IAC_NR: Minor perturbation should yield similar evaluation scores



$\mathbf{IEC}_{NR}$

$\hat{\boldsymbol{q}} = [\ 3,\ 2,\ 1\ ]$

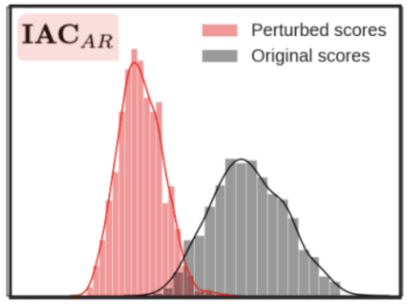
$\boldsymbol{q}'_k = [\ 3,\ 2,\ 1\ ] \cdots K$

$\vdots$

N

$$U^M_{i,j} = \begin{cases} 1 & \bar{r}^M_j = \bar{r}_j \\ 0 & \text{otherwise,} \end{cases} \quad (7)$$

IAC_AR: Disruptive perturbation should yield different evaluation scores



$\mathbf{IEC}_{AR}$

$\hat{\boldsymbol{q}} = [\ 0.6, 0.7, 0.2\ ]$

$\boldsymbol{q}'_k = [\ 0.3, 0.5, 0.1\ ] \cdots K$

$\vdots$

N

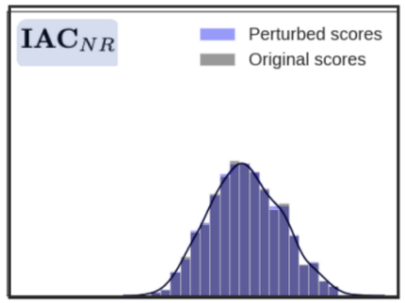
$$U^D_{i,j} = \begin{cases} 1 & \bar{Q}^D_{i,j} < \bar{Q}_{i,j} \\ 0 & \text{otherwise,} \end{cases} \quad (8)$$

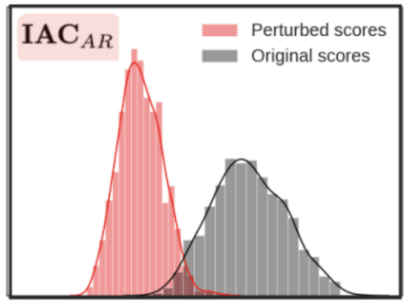
MetaQuantus — Method: Step 2

Step 2. Scoring

Measure effects of the perturbations via IAC and IEC criteria

$$\mathbf{IAC} = \frac{1}{K}\sum_{k=1}^{K} d(\hat{\boldsymbol{q}}, \boldsymbol{q}'_k), \quad (5)$$





$L =$ 

$$\mathbf{IEC} = \frac{1}{N \times L}\sum_{i=1}^{N}\sum_{j=1}^{L} U^t_{i,j} \quad (6)$$

$\mathbf{IEC}_{NR}$

$$\hat{\boldsymbol{q}} = [\ 3,\ 2,\ 1\ ]$$
$$\boldsymbol{q}'_k = [\ 3,\ 2,\ 1\ ] \cdots K$$
$$\vdots$$
$$N \quad U^M_{i,j} = \begin{cases} 1 & \bar{r}^M_j = \bar{r}_j \\ 0 & \text{otherwise,} \end{cases} \quad (7)$$

$\mathbf{IEC}_{AR}$

$$\hat{\boldsymbol{q}} = [\ 0.6, 0.7, 0.2\ ]$$
$$\boldsymbol{q}'_k = [\ 0.3, 0.5, 0.1\ ] \cdots K$$
$$\vdots$$
$$N \quad U^D_{i,j} = \begin{cases} 1 & \bar{Q}^D_{i,j} < \bar{Q}_{i,j} \\ 0 & \text{otherwise,} \end{cases} \quad (8)$$

IEC_NR: Minor perturbation should yield consistent rankings

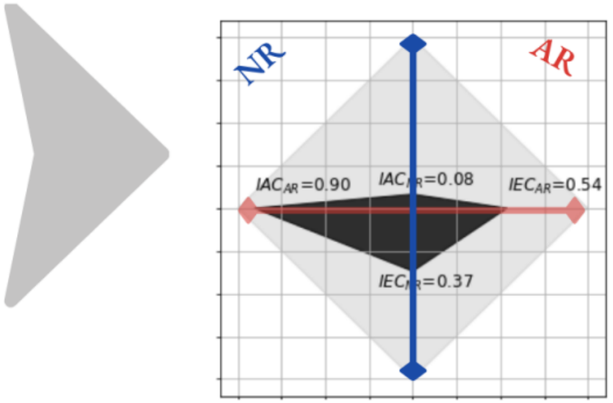
IEC_AR: Disruptive perturbation should yield lower rankings

MetaQuantus — Method: Step 3

Step 3. Integrating

Evaluate meta-consistency performance by combining the failure modes

$$\mathbf{MC} = \left(\frac{1}{|\boldsymbol{m}^*|}\right) \boldsymbol{m}^{*T} \boldsymbol{m} \quad \text{where} \quad \boldsymbol{m} = \begin{bmatrix} \mathbf{IAC}_{NR} \\ \mathbf{IAC}_{AR} \\ \mathbf{IEC}_{NR} \\ \mathbf{IEC}_{AR} \end{bmatrix} \tag{9}$$



MetaQuantus — Method: Step 1

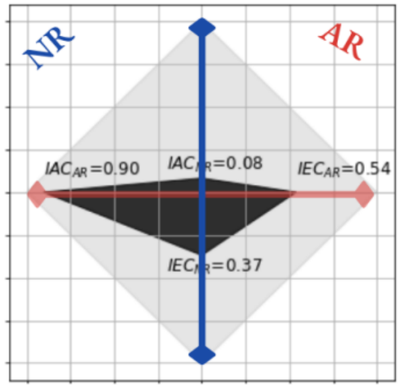
MC score captures metric behaviour

Step 3. Integrating

Evaluate meta-consistency performance by combining the failure modes

$$\mathbf{MC} = \left(\frac{1}{|\boldsymbol{m}^*|}\right) \boldsymbol{m}^{*T} \boldsymbol{m} \quad \text{where} \quad \boldsymbol{m} = \begin{bmatrix} \mathbf{IAC}_{NR} \\ \mathbf{IAC}_{AR} \\ \mathbf{IEC}_{NR} \\ \mathbf{IEC}_{AR} \end{bmatrix} \tag{9}$$

Area graph display metric characteristics



MetaQuantus — Method: Summary

Step 1. Perturbing

Depending on failure mode, initiate a minor or disruptive perturbation

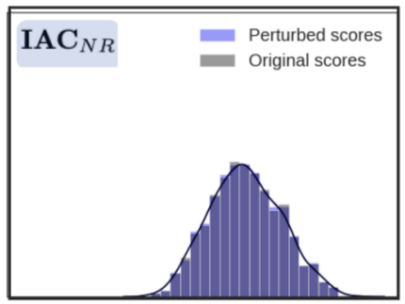
$\mathcal{P}_{\Omega}^{M}$ — Noise Resilience

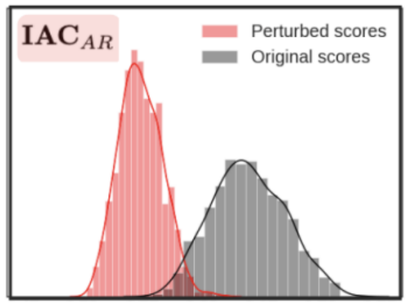
$\mathcal{P}_{\Omega}^{D}$ — Adversary Reactivity

Step 2. Scoring

Measure effects of the perturbations via IAC and IEC criteria

$$\mathbf{IAC} = \frac{1}{K}\sum_{k=1}^{K} d(\hat{\boldsymbol{q}}, \boldsymbol{q}'_k), \quad (5)$$





$L =$ 

$$\mathbf{IEC} = \frac{1}{N \times L}\sum_{i=1}^{N}\sum_{j=1}^{L} U_{i,j}^{t} \quad (6)$$

$\mathbf{IEC}_{NR}$

$\hat{\boldsymbol{q}} = [\ 3,\ 2,\ 1\ ]$

$\boldsymbol{q}'_k = [\ 3,\ 2,\ 1\ ] \cdots K$

$\vdots$

N

$$U_{i,j}^{M} = \begin{cases} 1 & \bar{r}_j^{M} = \bar{r}_j \\ 0 & \text{otherwise,} \end{cases} \quad (7)$$

$\mathbf{IEC}_{AR}$

$\hat{\boldsymbol{q}} = [\ 0.6, 0.7, 0.2\ ]$

$\boldsymbol{q}'_k = [\ 0.3, 0.5, 0.1\ ] \cdots K$

$\vdots$

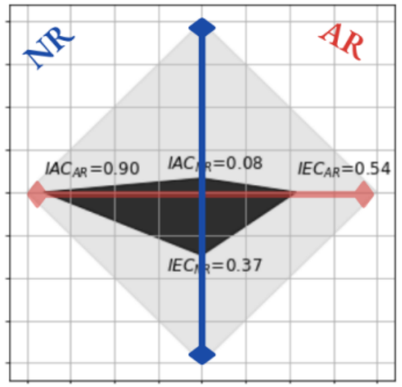
N

$$U_{i,j}^{D} = \begin{cases} 1 & \bar{Q}_{i,j}^{D} < \bar{Q}_{i,j} \\ 0 & \text{otherwise,} \end{cases} \quad (8)$$

Step 3. Integrating

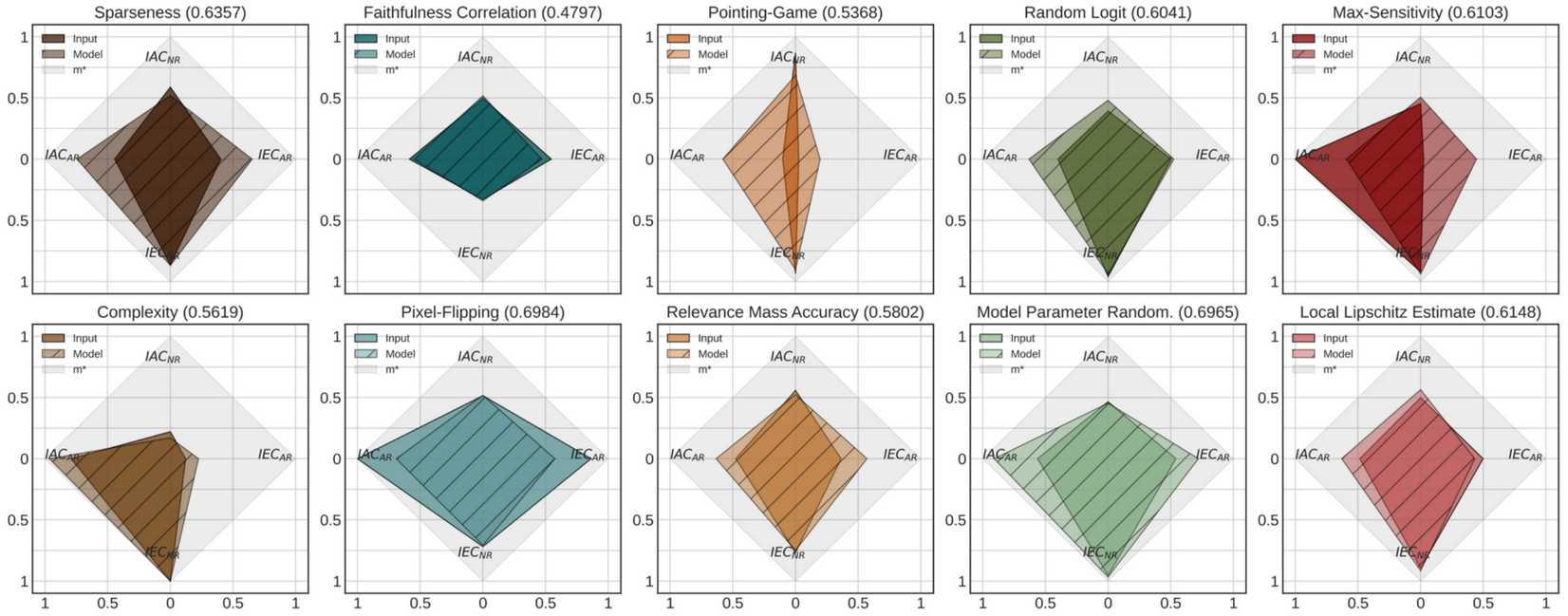
Evaluate meta-consistency performance by combining the failure modes

$$\mathbf{MC} = \left(\frac{1}{|\boldsymbol{m}^*|}\right) \boldsymbol{m}^{*T}\boldsymbol{m} \quad \text{where} \quad \boldsymbol{m} = \begin{bmatrix} \mathbf{IAC}_{NR} \\ \mathbf{IAC}_{AR} \\ \mathbf{IEC}_{NR} \\ \mathbf{IEC}_{AR} \end{bmatrix} \quad (9)$$



MetaQuantus — Results 1/3

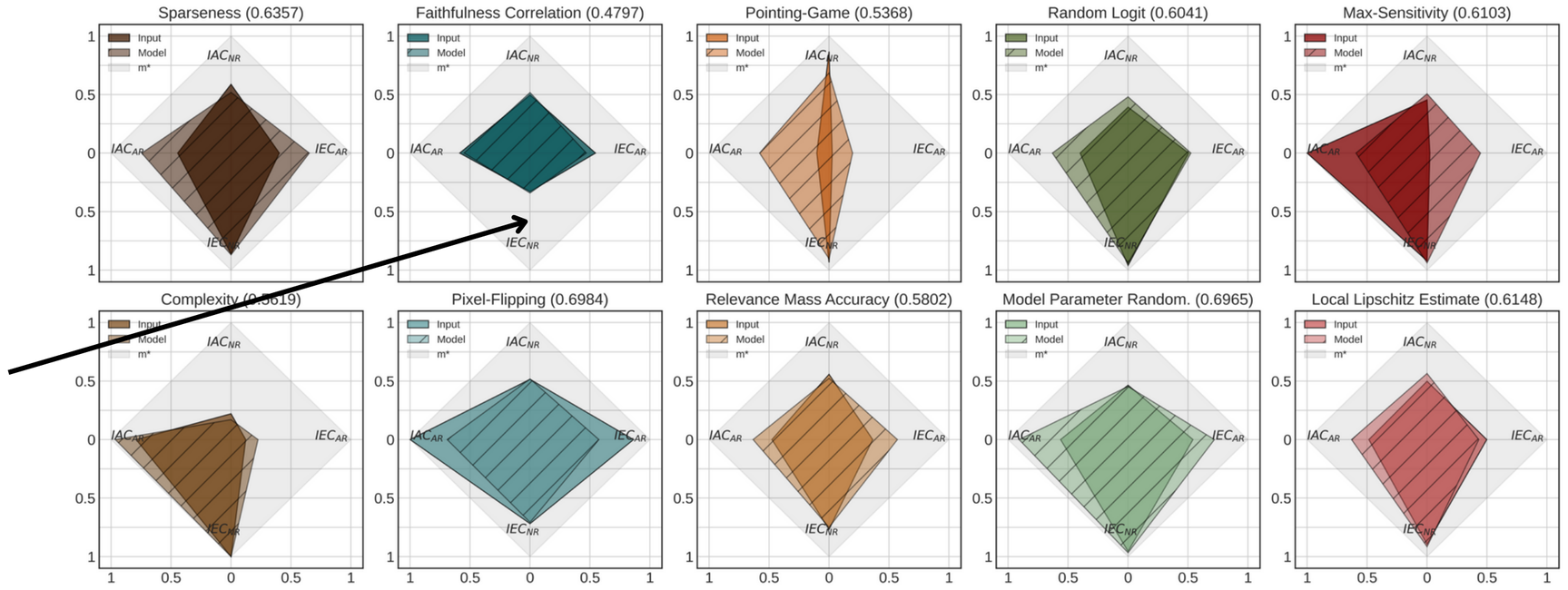
Gain Insights on Metric Performance over Various Categories [ImageNet, ResNet18]



MetaQuantus — Results 1/3

Gain Insights on Metric Performance over Various Categories [ImageNet, ResNet18]

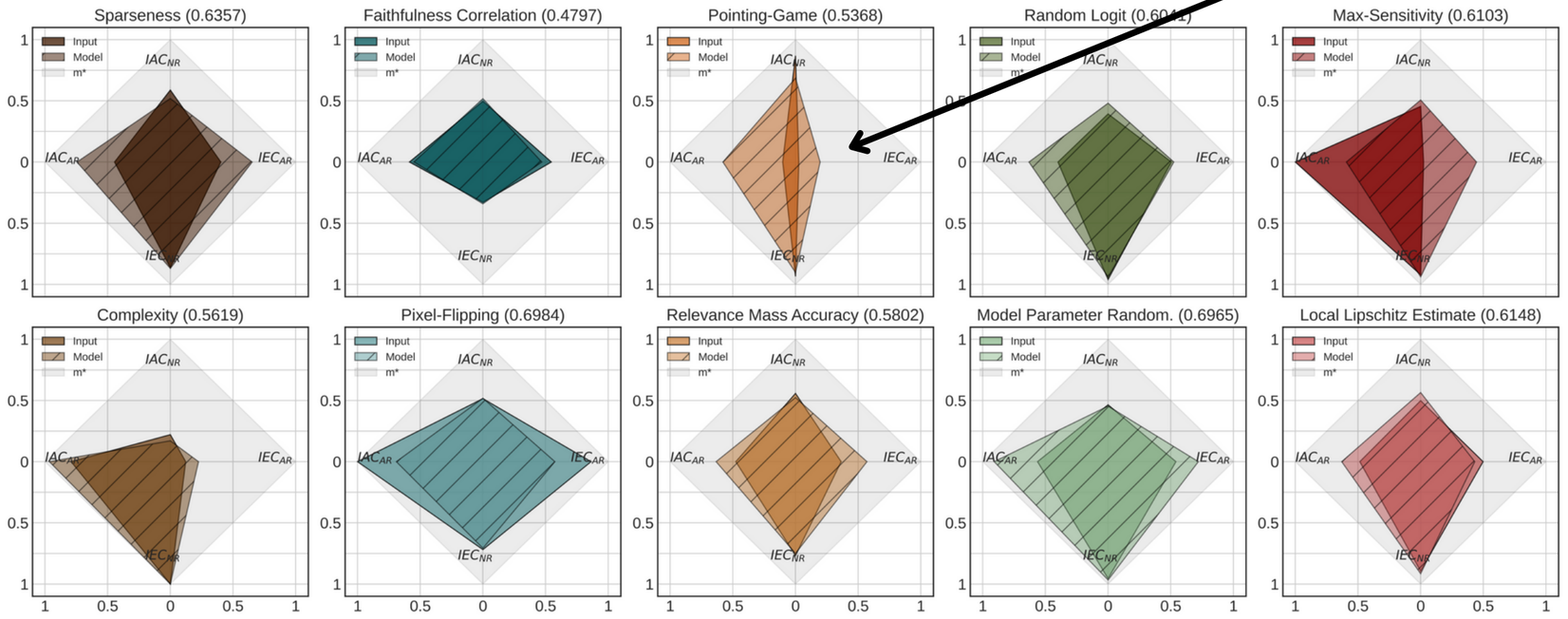
Low NR scores in faithfulness metrics corroborate found sensitivity



MetaQuantus — Results 1/3

Low IEC AR in localisation metrics reveals mask dependency

Gain Insights on Metric Performance over Various Categories [ImageNet, ResNet18]



MetaQuantus — Results 2/3

Gain Insights on Metric Performance over Various Categories [ImageNet, ResNet18]

Improved MC score
when masks are smaller

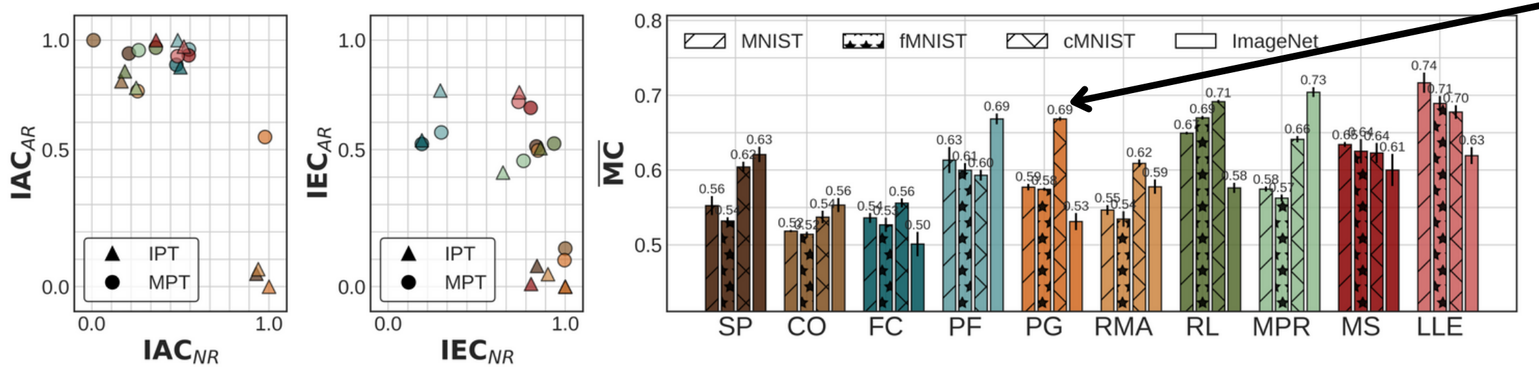
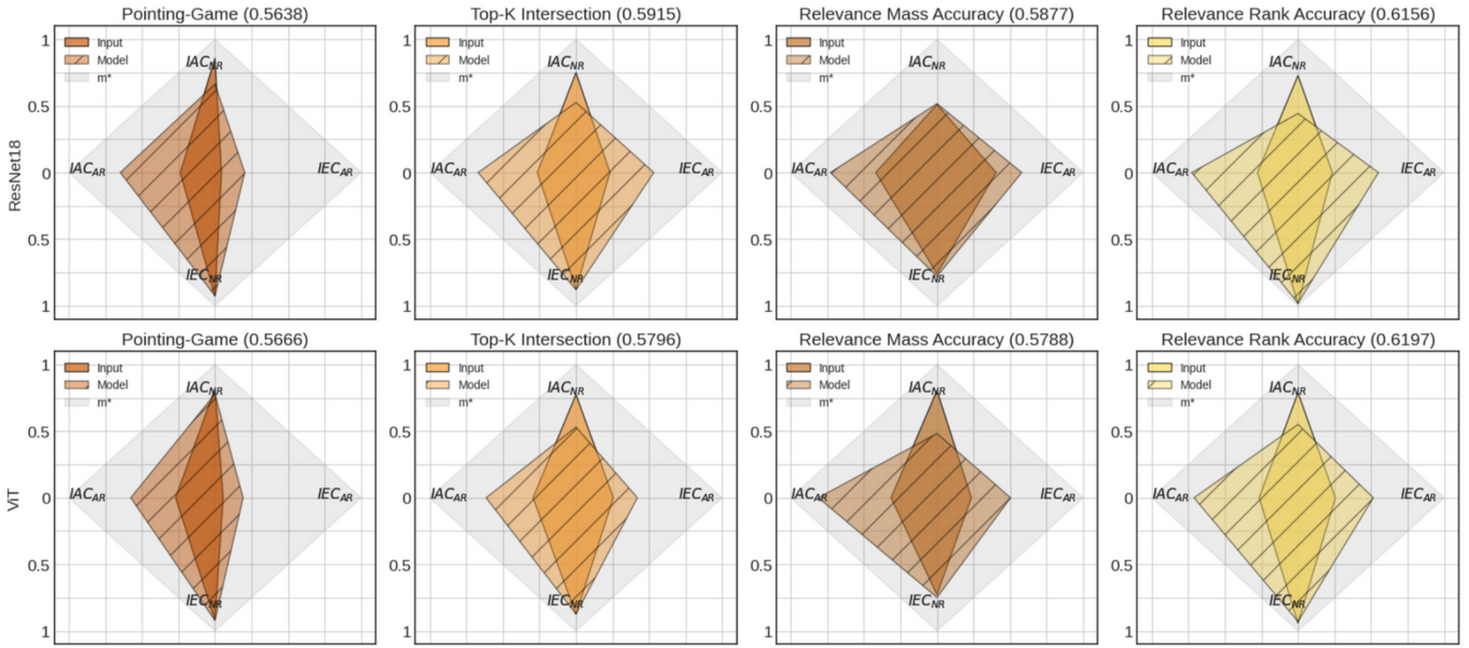


Figure 5: *Left:* A visualisation of MNIST benchmarking results (Table 3), in particular IAC and IEC scores for noise resilience (x-axes) and adverse reactivity (y-axes). The colours indicate the estimator and the symbols show the test type, i.e., IPT and MPT, respectively. *Right:* A comparison of averaged meta-consistency performance for different quality estimators using MPT and IPT, aggregated over 3 iterations with $K = 5$, across different models {*LeNets, ResNet-9, ResNet-18*} and datasets {*MNIST, fMNIST, cMNIST, ImageNet*}. Higher values are preferred.

MetaQuantus — Results 3/3

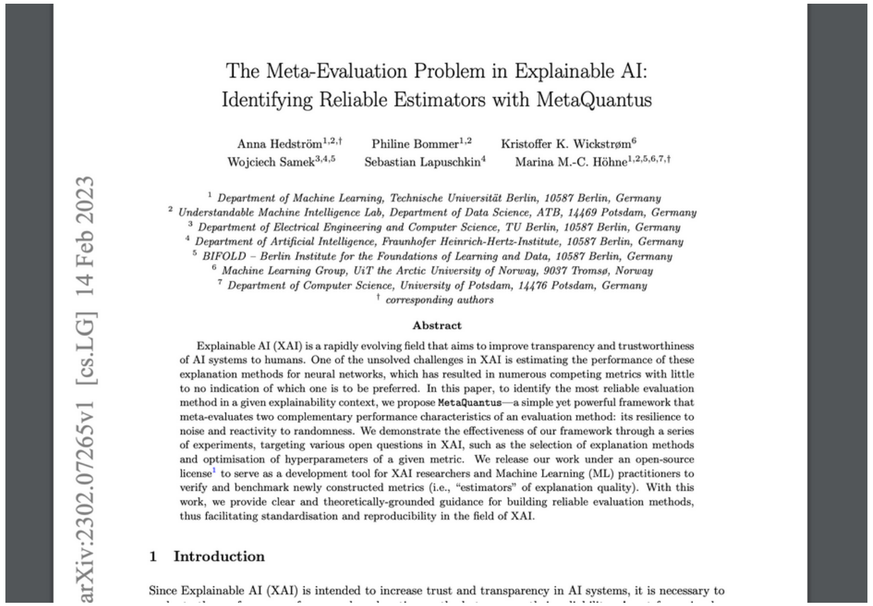
Gain Insights on Metric Performance over Various Categories



Identified metric characteristics are consistent across architectures and datasets

MetaQuantus — Learn More

TMLR (OpenReview), code at Github



The Meta-Evaluation Problem in Explainable AI:
Identifying Reliable Estimators with MetaQuantus

Anna Hedström[1,2,†] Philine Bommer[1,2] Kristoffer K. Wickstrøm[6]
Wojciech Samek[3,4,5] Sebastian Lapuschkin[4] Marina M.-C. Höhne[1,2,5,6,7,†]

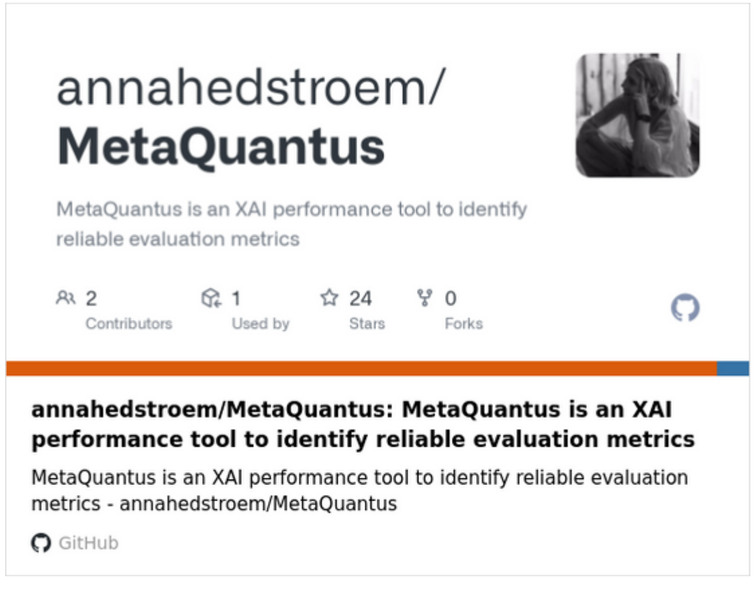
[1] *Department of Machine Learning, Technische Universität Berlin, 10587 Berlin, Germany*
[2] *Understandable Machine Intelligence Lab, Department of Data Science, ATB, 14469 Potsdam, Germany*
[3] *Department of Electrical Engineering and Computer Science, TU Berlin, 10587 Berlin, Germany*
[4] *Department of Artificial Intelligence, Fraunhofer Heinrich-Hertz-Institute, 10587 Berlin, Germany*
[5] *BIFOLD – Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany*
[6] *Machine Learning Group, UiT the Arctic University of Norway, 9037 Tromsø, Norway*
[7] *Department of Computer Science, University of Potsdam, 14476 Potsdam, Germany*
[†] *corresponding authors*

Abstract

Explainable AI (XAI) is a rapidly evolving field that aims to improve transparency and trustworthiness of AI systems to humans. One of the unsolved challenges in XAI is estimating the performance of these explanation methods for neural networks, which has resulted in numerous competing metrics with little to no indication of which one is to be preferred. In this paper, to identify the most reliable evaluation method in a given explainability context, we propose `MetaQuantus`—a simple yet powerful framework that meta-evaluates two complementary performance characteristics of an evaluation method: its resilience to noise and reactivity to randomness. We demonstrate the effectiveness of our framework through a series of experiments, targeting various open questions in XAI, such as the selection of explanation methods and optimisation of hyperparameters of a given metric. We release our work under an open-source license[1] to serve as a development tool for XAI researchers and Machine Learning (ML) practitioners to verify and benchmark newly constructed metrics (i.e., "estimators" of explanation quality). With this work, we provide clear and theoretically-grounded guidance for building reliable evaluation methods, thus facilitating standardisation and reproducibility in the field of XAI.

1 Introduction

Since Explainable AI (XAI) is intended to increase trust and transparency in AI systems, it is necessary to



annahedstroem/
MetaQuantus

MetaQuantus is an XAI performance tool to identify reliable evaluation metrics

2 Contributors · 1 Used by · 24 Stars · 0 Forks

annahedstroem/MetaQuantus: MetaQuantus is an XAI performance tool to identify reliable evaluation metrics

MetaQuantus is an XAI performance tool to identify reliable evaluation metrics - annahedstroem/MetaQuantus

GitHub

Evaluation

Summary

Evaluation — Scope

The Challenge of Explanation Evaluation

1. **Motivation —** Why it is important (to all) and interesting (as a research problem)

2. **Methods —** What are the current methods and pitfalls; human, approximate and restriction

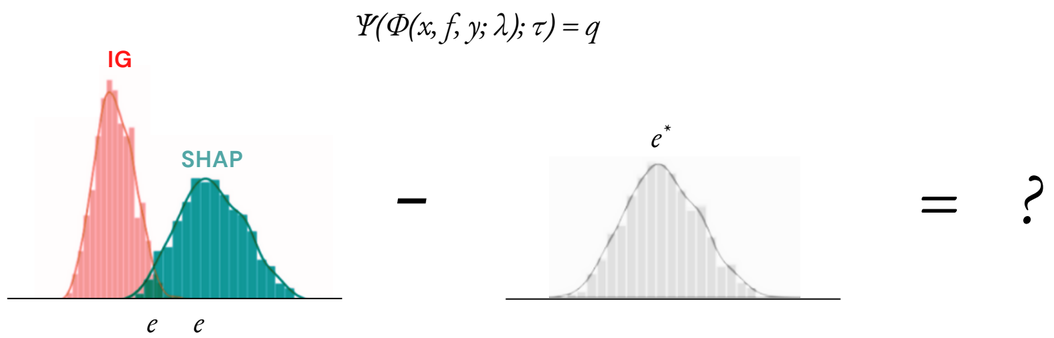
3. **Meta-Evaluation —** How to estimate explanation quality, reliably

Evaluation — Scope

The Challenge of Explanation Evaluation

1. Motivation — Why it is important (to all) and interesting (as a research problem)

The Evaluation Problem in Explainable AI (Hedström et al., 2023a)



Evaluation — Scope

The Challenge of Explanation Evaluation

1. Motivation — Why it is important (to all) and interesting (as a research problem)

2. Methods — What are the current methods and pitfalls; human, approximate and restriction

Three options, evaluation by:

- Human(s)
- Restriction
- Approximation

Evaluation — Scope

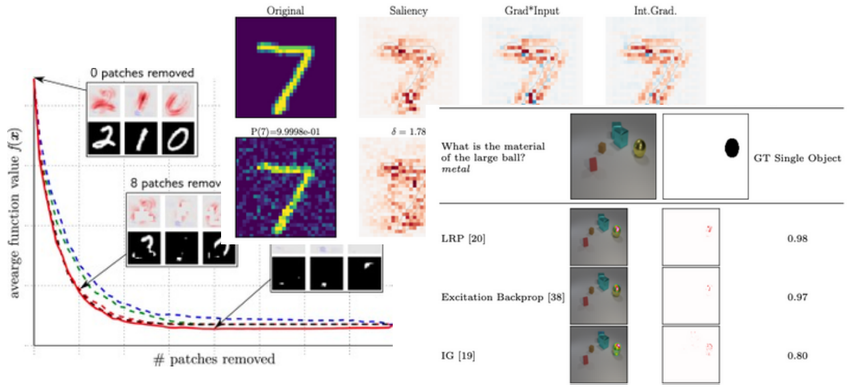
The Challenge of Explanation Evaluation

1. **Motivation** — Why it is important (to all) and interesting (as a research problem)

2. Methods — What are the current methods and pitfalls; human, approximate and restriction

Three options, evaluation by:

- Human(s)
- Restriction
- Approximation



Evaluation — Scope

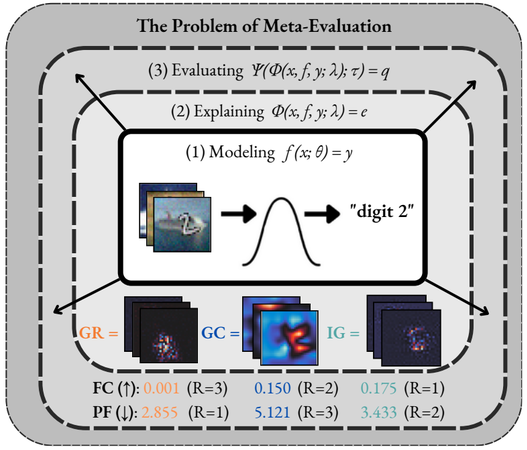
The Challenge of Explanation Evaluation

1. **Motivation** — Why it is important (to all) and interesting (as a research problem)

2. **Methods** — What are the current methods and pitfalls; human, approximate and restriction

3. Meta-Evaluation — How to estimate explanation quality, reliably

No tested quality estimator is fully reliable, but awareness is growing and new approaches are being developed



Today's agenda

01 Foundations

02 Evaluation

03 Quantus (hands-on)

04 Discussion + Q&A

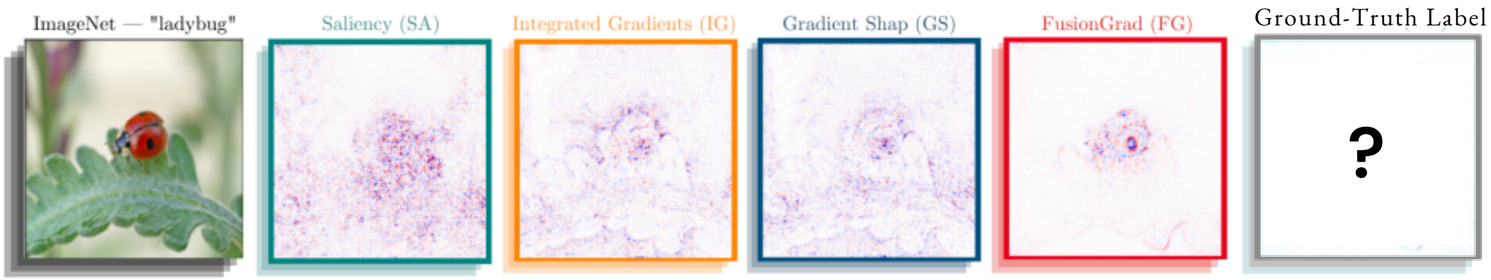
Lecture 3

Quantus

Quantus — Motivation 1/2

Automate Explainable AI Evaluation (by Approximation)

- Provide the XAI and ML communities with an efficient, easy-to-use open-sourced API to perform XAI evaluation (by approximation)



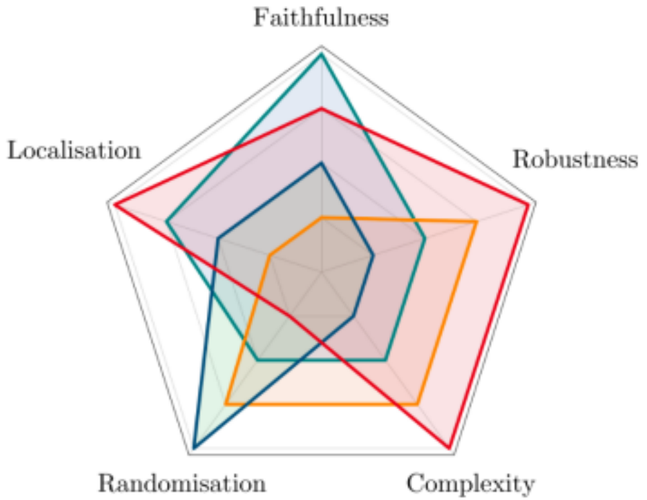
- Complete lack of open-source tools for XAI evaluation, at the time

Quantus — Motivation 2/2

Automate Explainable AI Evaluation (by Approximation)

- Enable automation and large-scale experimentation, across a diverse set of evaluation properties, models and datasets

- Give a quantitative snapshot of the explanation quality



Quantus — Library Content

Evaluate Explanations from PyTorch and Tensorflow Models

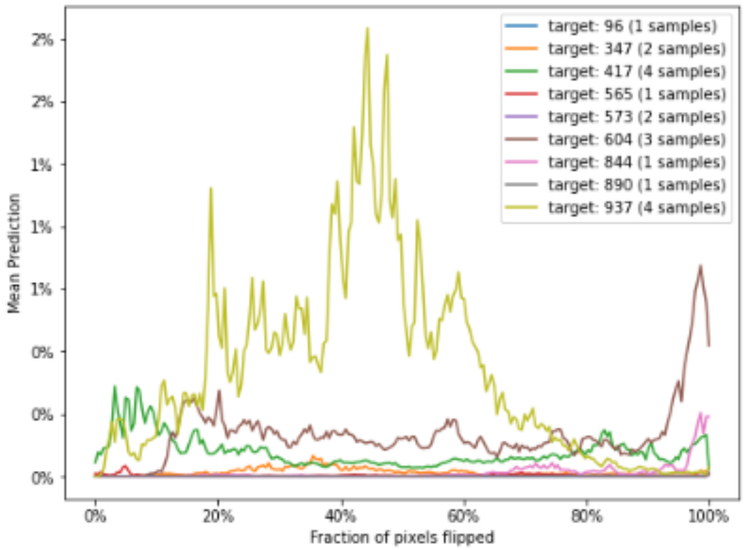
- **Metrics.** 35+ metrics in 6 categories for XAI evaluation with tutorials and API reference
- **Data and model types.** Support (image, time-series, tabular, NLP in progress!) datasets for PyTorch and Tensorflow ML models
- **Explanation methods.** E.g., gradient-, back-propagation-, model-agnostic, local surrogate-, attention-, prototype-based explanations

Quantus — Library Syntax

Evaluation in an one-liner or with quantus.evaluate()

```
[ ]  1 # Create the pixel-flipping experiment.
     2 pixel_flipping = quantus.PixelFlipping(
     3     features_in_step=224,
     4     perturb_baseline="black",
     5     perturb_func=quantus.baseline_replacement_by_indices,
     6 )
     7
     8 # Call the metric instance to produce scores.
     9 scores = pixel_flipping(model=model,
    10                         x_batch=x_batch,
    11                         y_batch=y_batch,
    12                         a_batch=a_batch,
    13                         device=device,)
    14
    15 # Plot example!
    16 pixel_flipping.plot(y_batch=y_batch, scores=scores)
```

__init__ the metric in one go



plot() to visualise some results

score xAI methods using **__call__**

Quantus — Application Highlights

Diverse Applications Across Fields

Climate science [1, 2]

Healthcare [3, 4, 5, 6, 7]

Object Detection [12]

Remote sensing [14]

Image Classification [8, 9]

Security [15]

Meta-evaluation [10, 11]

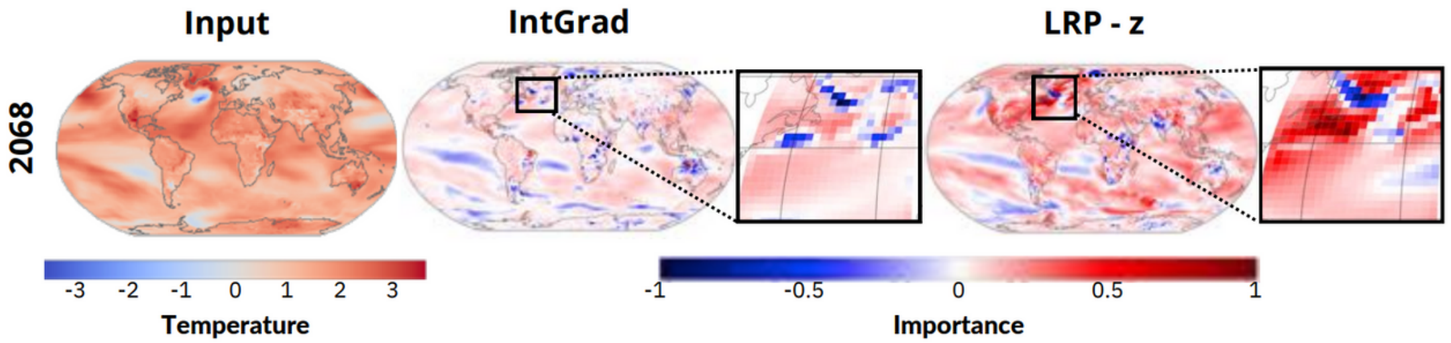
Network Canonization [13]

.....

Quantus — Application Highlights

Diverse Applications Across Fields

Climate science [1, 2] — Evaluate explanations of temperature prediction models

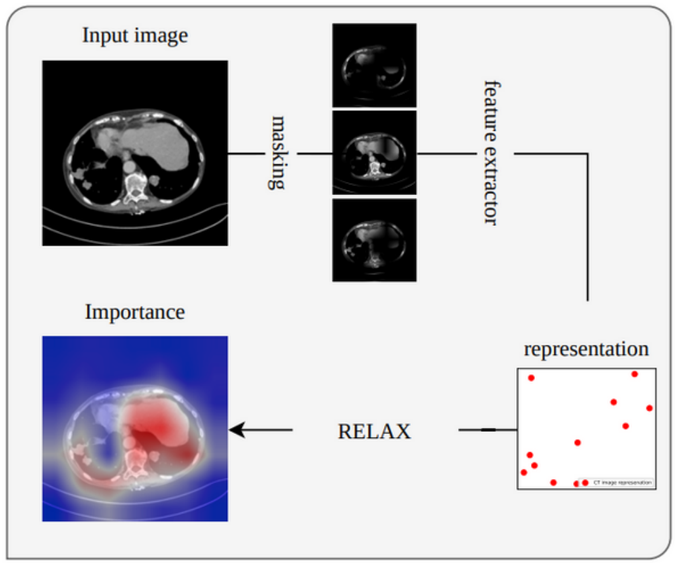


Bommer, Philine, et al. "Finding the right XAI method--A Guide for the Evaluation and Ranking of Explainable AI Methods in Climate Science." arXiv preprint arXiv:2303.00652 (2023).

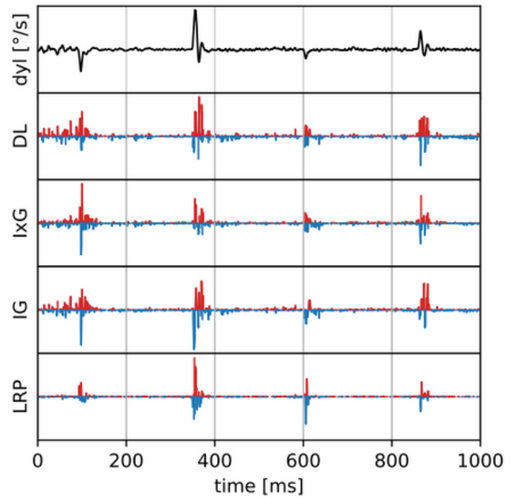
Quantus — Application Highlights

Diverse Applications Across Fields

Healthcare [3, 4, 5, 6, 7] — Evaluate explanations of liver disease- (left) and biometric eye-tracking models



Wickstrøm, Kristoffer Knutsen, et al. "A clinically motivated self-supervised approach for content-based image retrieval of CT liver images." Computerized Medical Imaging and Graphics 107 (2023): 102239.



(c) pitch velocities of left eye

Krakowczyk, Daniel G., et al. "Bridging the Gap: Gaze Events as Interpretable Concepts to Explain Deep Neural Sequence Models." Proceedings of the 2023 Symposium on Eye Tracking Research and Applications. 2023.

Quantus — Application Highlights

JMLR paper (available Vol24), code at Github and API documentation

Journal of Machine Learning Research 24 (2023) 1-11 Submitted 2/22; Revised 11/22; Published 1/23

Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond

Anna Hedström[1,†] ANNA.HEDSTROEM@TU-BERLIN.DE
Leander Weber[3] LEANDER.WEBER@HHI.FRAUNHOFER.DE
Dilyara Bareeva[1] DILYARA.BAREEVA@CAMPUS.TU-BERLIN.DE
Daniel Krakowczyk[4] DANIEL.KRAKOWCZYK@UNI-POTSDAM.DE
Franz Motzkus[3] FRANZ.MOTZKUS@HHI.FRAUNHOFER.DE
Wojciech Samek[2,3,5] WOJCIECH.SAMEK@HHI.FRAUNHOFER.DE
Sebastian Lapuschkin[3,†] SEBASTIAN.LAPUSCHKIN@HHI.FRAUNHOFER.DE
Marina M.-C. Höhne[1,5,†] MARINA.HOEHNE@TU-BERLIN.DE
[1] *Understandable Machine Intelligence Lab, TU Berlin, 10587 Berlin, Germany*
[2] *Department of Electrical Engineering and Computer Science, TU Berlin, 10587 Berlin, Germany*
[3] *Department of Artificial Intelligence, Fraunhofer Heinrich-Hertz-Institute, 10587 Berlin, Germany*
[4] *Department of Computer Science, University of Potsdam, 14476 Potsdam, Germany*
[5] *BIFOLD – Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany*
[†] *corresponding authors*

Editor: Joaquin Vanschoren

Abstract

The evaluation of explanation methods is a research topic that has not yet been explored deeply, however, since explainability is supposed to strengthen trust in artificial intelligence, it is necessary to systematically review and compare explanation methods in order to confirm their correctness. Until now, no tool with focus on XAI evaluation exists that exhaustively

Today's agenda

01 Foundations

02 Evaluation

03 Quantus (hands-on)

04 Discussion + Q&A

End