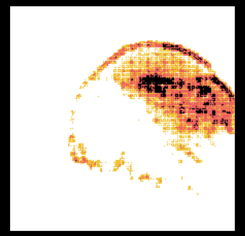
The First Provably Robust and Interpretable Pixel-Level Explanations



Colored pixels are provably important for predicting "eagle" under bounded input noise

Pixel-Level Certified Explanations via Randomized Smoothing

Alaa Anani[1,2] Tobias Lorenz[2] Mario Fritz[1] Bernt Schiele[2]

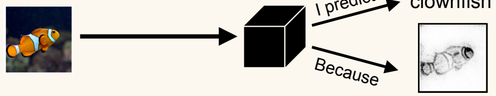
[1] MAX PLANCK INSTITUTE FOR INFORMATICS
[2] CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY
SIC Saarland Informatics Campus
ICML International Conference On Machine Learning
elsa European Lighthouse on Secure and Safe AI

Paper & Code

1. Take-Home Message

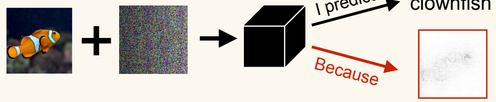
Attributions

Explain classifier prediction by highlighting influential pixels.



Problem

Attributions are highly non-robust to small input noise.



Solution

① We certify any attribution method by proving which pixels stay important or not within an ℓ_2-ball around the input.
② We benchmark 12 attribution methods in terms of **robustness**, **localization** and **faithfulness**.

2. Randomized Smoothing

Proving model output stability by evaluating it on an input distribution [1,2].

$$f(x) \rightarrow f(\mathcal{N}(x, \sigma^2 I))$$

Certified segmentation [2]

Given a segmentation model $f : \mathbb{R}^{c \cdot N} \rightarrow \mathcal{Y}^N$, threshold $\tau \in [\frac{1}{2}, 1)$, and sampling error rate α, the smoothed (certified) version is defined as

$$g_i^\tau(x) = \begin{cases} c_{A,i} & \text{if } \mathbf{P}(f_i(\mathcal{N}(x, \sigma^2 I)) = c_{A,i}) > \tau \\ \oslash, & \text{otherwise} \end{cases}$$

With probability $1 - \alpha$, all non-abstain indices $\{ i \mid g_i^\tau(x) \neq \oslash \}$ remain the same $g_i^\tau(x) = g_i^\tau(x + \delta)$, for $\delta \in \mathbb{R}^{m \cdot N}$ with $||\delta||_2 \leq R := \sigma \Phi^{-1}(\tau)$.

[1] Cohen et al. ICML, 2019. [2] Fischer et al. ICML, 2021.

3. Method

x

Given any attribution method

$h(x)$

How can we obtain a pixel-level provably robust explanation?

① Sample attributions on input image with noise $\sim h(\mathcal{N}(x, \sigma^2 I))$

② Sparsify sampled attributions $\sim h^K(\mathcal{N}(x, \sigma^2 I))$

Top K% → 1, else → 0

③ Certify stable pixels

$$h^{\tau,K}(x) = \begin{cases} 1 \text{ if } \mathbf{P}(h_i^K(\mathcal{N}(x, \sigma^2 I) = 1) > \tau \\ 0 \text{ if } \mathbf{P}(h_i^K(\mathcal{N}(x, \sigma^2 I) = 0) > \tau \\ \oslash \text{ otherwise} \end{cases}$$

Pixel values are mostly 1 → Certifiably **important**
Mostly 0 → Certifiably **unimportant**
Fluctuating ? → Abstain ⊘

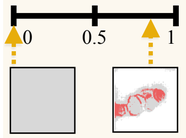
With high probability $1 - \alpha$, for all non-abstain indices $\{ i \mid h_i^{\tau,k}(x) \neq \oslash \}$, the certified attribution output remains the same $h_i^{\tau,K}(x) = h_i^{\tau,K}(x + \delta)$, within an ℓ_2-radius around the input: $||\delta||_2 \leq R := \sigma \Phi^{-1}(\tau)$.

4. Evaluation Setup

We evaluate **12 attribution** methods from **3 families** (**gradient**, **perturbation**, **activation**) on **5 models** using **3 novel metrics:**

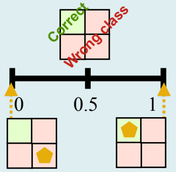
Robustness

Percentage of certified pixels



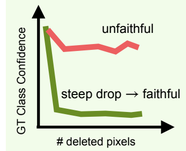
Localization

Do certified pixels **localize** the right class?

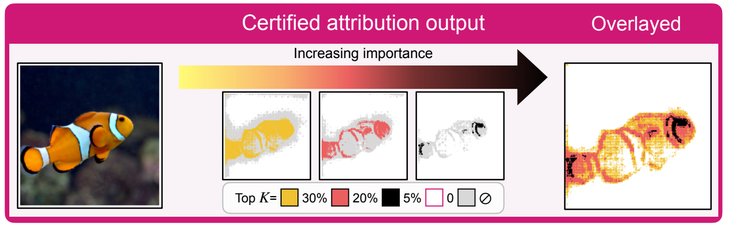


Faithfulness

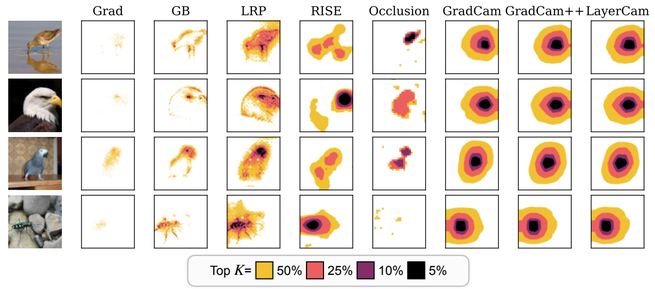
Do certified pixels reflect what the model **relies** on?



5. Results



Certified attributions are **interpretable** and of **high-quality. LRP** offers the most detail.



Certified Robustness

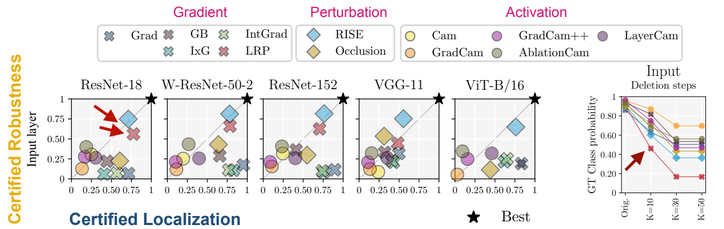
LRP and **RISE** are the most robust.

Certified Localization

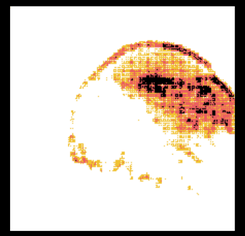
LRP, **RISE** and **Occlusion** localize best.

Faithfulness

LRP and **RISE** are the most faithful.



The First Provably Robust and Interpretable Pixel-Level Explanations



Colored pixels are provably important for predicting "eagle" under bounded input noise

Pixel-Level Certified Explanations via Randomized Smoothing

Alaa Anani[1,2] Tobias Lorenz[2] Mario Fritz[2] Bernt Schiele[1]

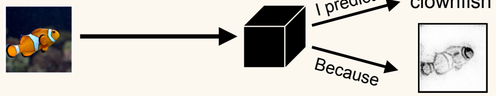
[1] MAX PLANCK INSTITUTE FOR INFORMATICS
[2] CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY
ICML International Conference On Machine Learning
SIC Saarland Informatics Campus
elsa European Lighthouse on Secure and Safe AI

Paper & Code

1. Take-Home Message

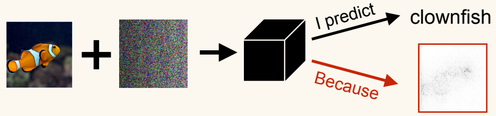
Attributions

Explain classifier prediction by highlighting influential pixels.



Problem

Attributions are highly non-robust to small input noise.



Solution

① We certify any attribution method by proving which pixels stay important or not within an ℓ_2-ball around the input.
② We benchmark 12 attribution methods in terms of **robustness**, **localization** and **faithfulness**.

2. Randomized Smoothing

Proving model output stability by evaluating it on an input distribution [1,2].

$$f(x) \rightarrow f(\mathcal{N}(x, \sigma^2 I))$$

Certified segmentation [2]

Given a segmentation model $f : \mathbb{R}^{c \cdot N} \rightarrow \mathcal{Y}^N$, threshold $\tau \in [\frac{1}{2}, 1)$, and sampling error rate α, the smoothed (certified) version is defined as

$$g_i^\tau(x) = \begin{cases} c_{A,i} & \text{if } \mathbf{P}(f_i(\mathcal{N}(x, \sigma^2 I)) = c_{A,i}) > \tau \\ \oslash, & \text{otherwise} \end{cases}$$

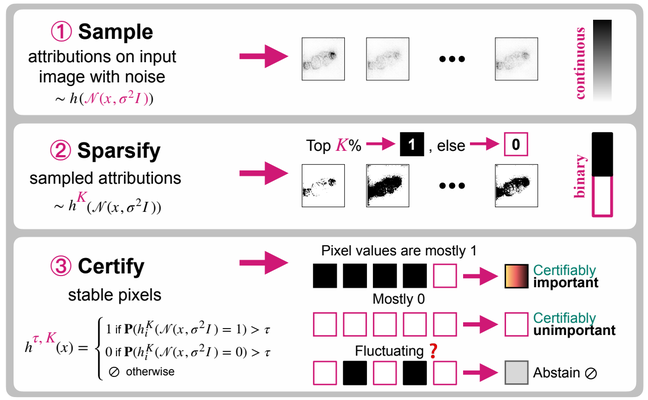
With probability $1 - \alpha$, all non-abstain indices $\{ i \mid g_i^\tau(x) \neq \oslash \}$ remain the same $g_i^\tau(x) = g_i^\tau(x + \delta)$, for $\delta \in \mathbb{R}^{m \cdot N}$ with $||\delta||_2 \leq R := \sigma \Phi^{-1}(\tau)$.

[1] Cohen et al. ICML, 2019. [2] Fischer et al. ICML, 2021.

3. Method

Given any attribution method

How can we obtain a pixel-level provably robust explanation?



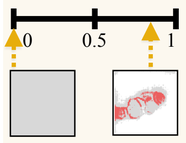
With high probability $1 - \alpha$, for all non-abstain indices $\{ i \mid h_i^{\tau,k}(x) \neq \oslash \}$, the certified attribution output remains the same $h_i^{\tau,K}(x) = h_i^{\tau,K}(x + \delta)$, within an ℓ_2-radius around the input: $||\delta||_2 \leq R := \sigma \Phi^{-1}(\tau)$.

4. Evaluation Setup

We evaluate **12 attribution** methods from **3 families** (**gradient**, **perturbation**, **activation**) on **5 models** using **3 novel metrics:**

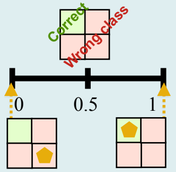
Robustness

Percentage of certified pixels



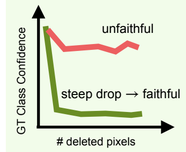
Localization

Do certified pixels **localize** the right class?

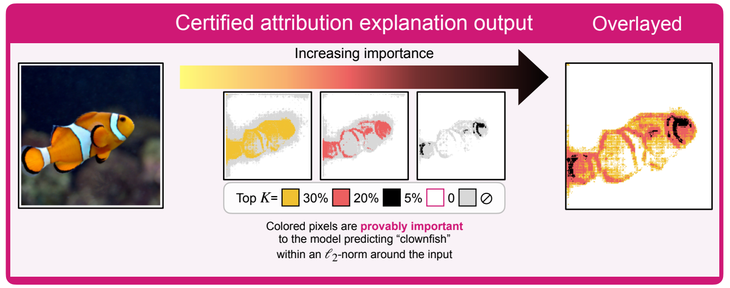


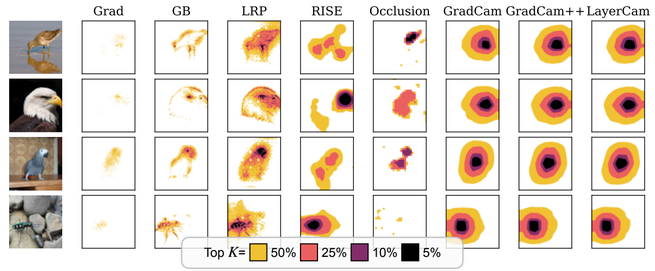
Faithfulness

Do certified pixels reflect what the model **relies** on?



5. Results





Certified Robustness

LRP and **RISE** are the most robust.

Certified Localization

LRP, **RISE** and **Occlusion** localize best.

Faithfulness

LRP and **RISE** are the most faithful.