XRAI: Better Attributions Through Regions

Andrei Kapishnikov*, Tolga Bolukbasi*, Fernanda Viegas, Michael Terry
People + AI Research Initiative, Google Research

Motivation

Pixel-Based Attribution Methods

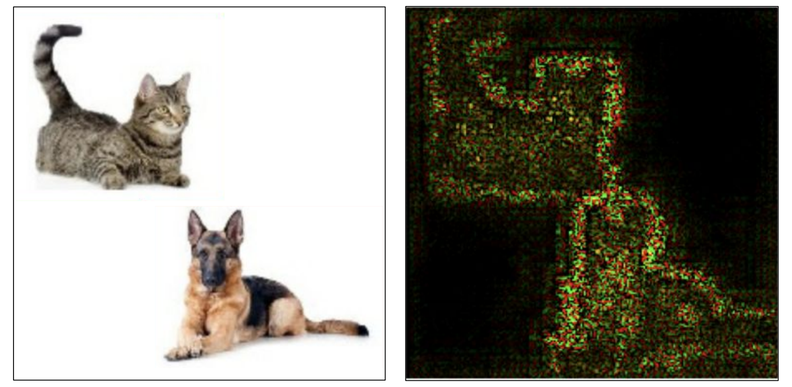
Saliency methods link a deep neural network's (DNN) prediction to the input features that most influence that prediction.

Pixel-based saliency methods provide fine-grained attributions for image models. By attributing individual pixels, these techniques provide fine-grained attributions.

However, pixel-based attributions can sometimes be challenging to read and interpret. Salient pixels may be scattered across the image, with positive and negative attributions intermixed.

The choice of baseline (e.g., a black baseline for Integrated Gradients) can also have a significant effect on the saliency method's results.

Left: Input image. **Right:** Salient pixels identified by Integrated Gradients for class "ground beetle," using a **black baseline**.



Left: Input image of a cat and dog. **Right:** Gradients for class "cat." Notice how there are both positive (green) and negative (red) attributions for both the cat and the dog.

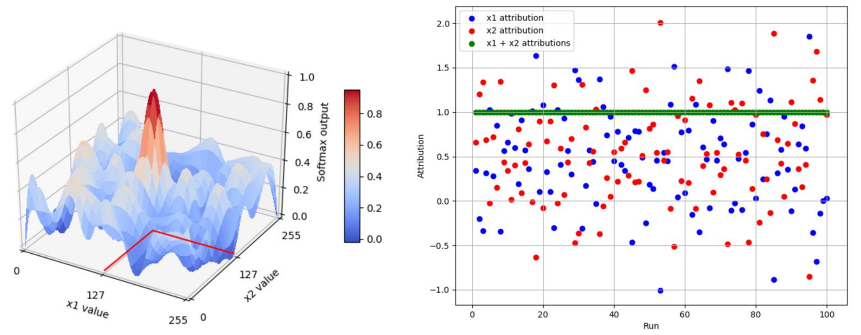
Evaluating Attribution Methods is Challenging

Assessing the **quality** and **correctness** of attribution methods also remains a core challenge, making it difficult to understand how well one method performs compared to another.

Sanity Check: Perturbation-ε

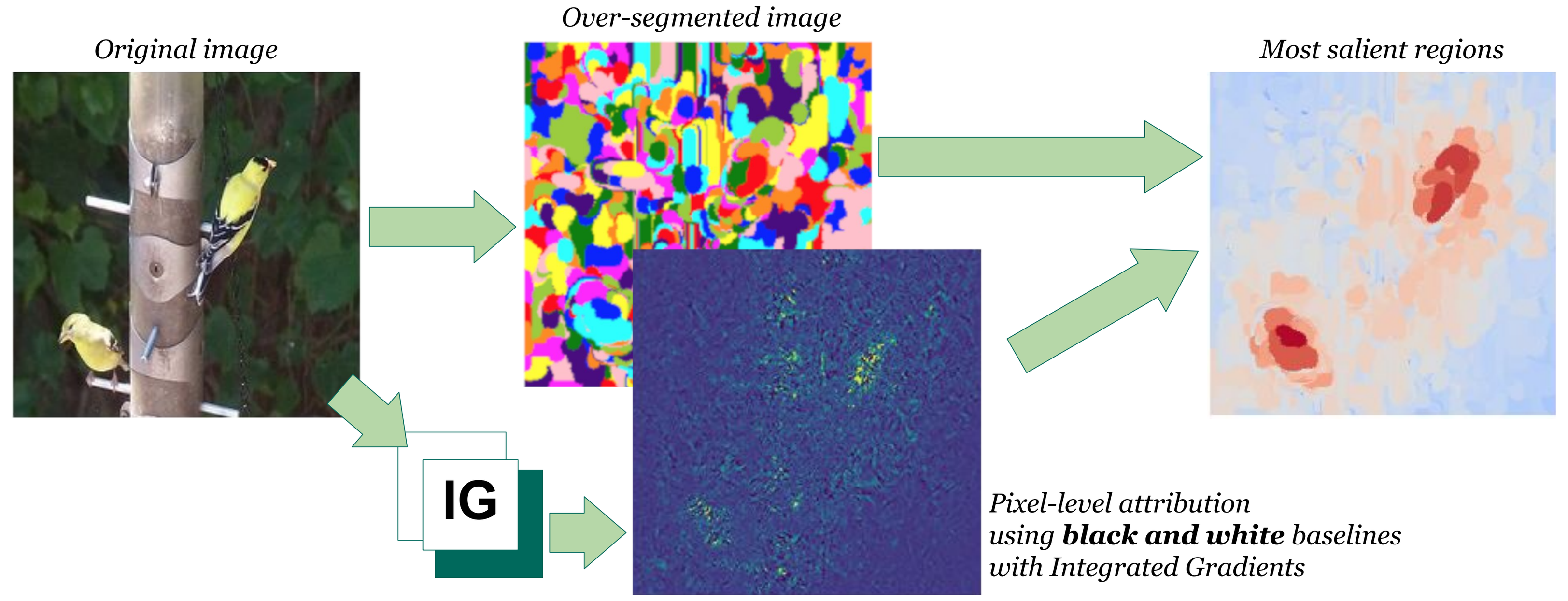
To aid assessment of saliency methods, we introduce an axiom, **Perturbation-ε**, that serves as a sanity check for saliency methods. In a nutshell, Perturbation-ε says that if you remove a feature and the output of the classifier is changed, then that feature should not have zero attribution. We found that Integrated Gradients doesn't always satisfy this axiom.

Axiom 1 *Perturbation-ϵ: Given ϵ, for every feature x_i in an input $\boldsymbol{x} = [x_1, ..., x_N]$ where all features except for x_i are fixed, if the removal (setting $x_i = 0$) of feature x_i causes the output to change by Δy, then Perturbation-ϵ is satisfied if the inequality $attr(x_i) \geq \epsilon * \Delta y$ is satisfied.*



XRAI Method

XRAI identifies salient regions as opposed to pixels.

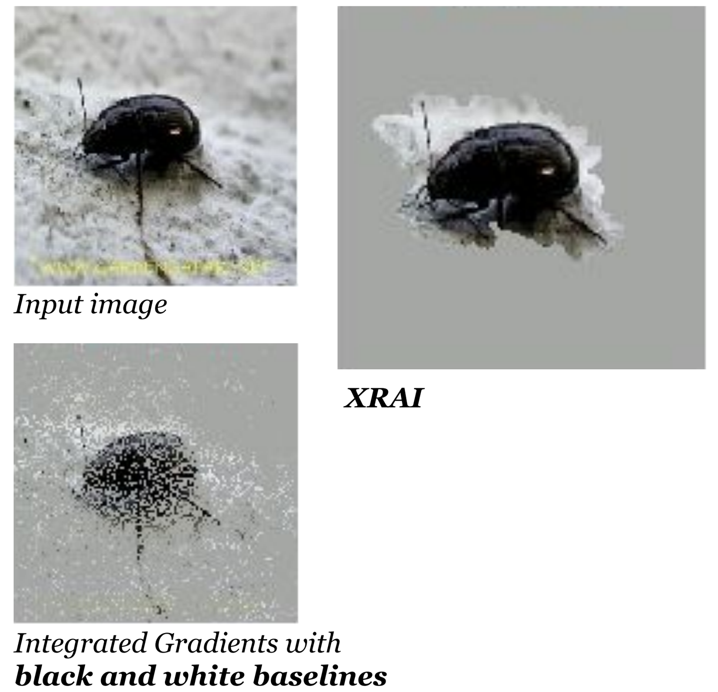


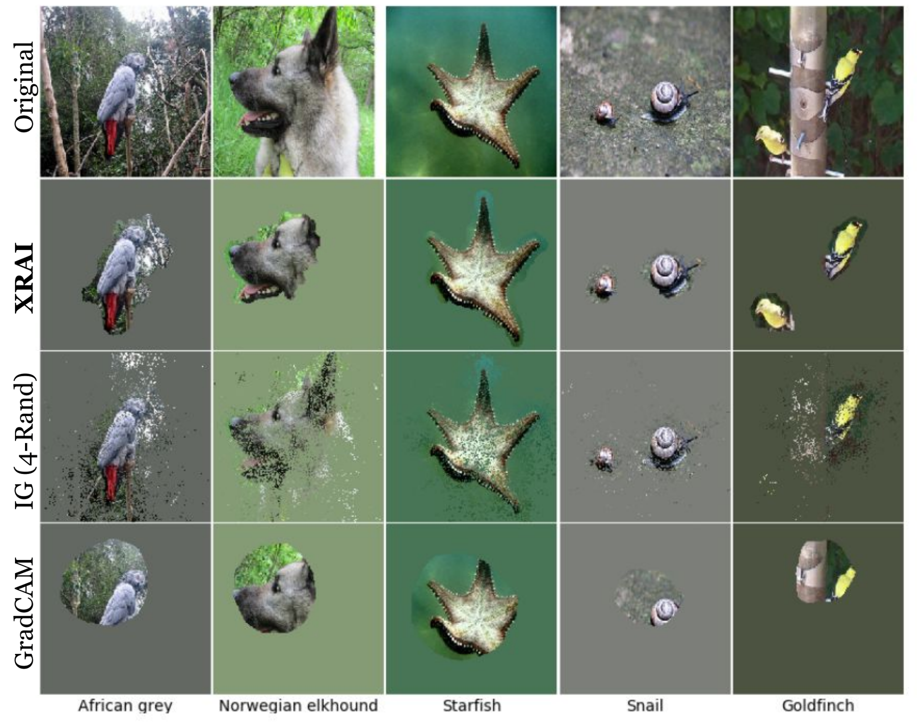
Above: XRAI identifies **the most salient regions** leading for prediction of a given class.

The XRAI Algorithm

1. **Pixel-level attribution:** XRAI performs pixel-level attribution for the input image. In our current implementation, we use Integrated Gradients with **two baselines, a black baseline** and **a white baseline**.
2. **Oversegmentation:** Separately from model attribution, XRAI oversegments the image to create a patchwork of small regions. XRAI currently uses Felzenswalb's graph-based method in the skimage package to create segments.
3. **Region selection:** For each segment, XRAI sums the attributions within that segment. Segments are then rank-ordered from most to least positive, in terms of summed attributions.

Once segments are rank-ordered, it is possible to reveal the top *n*% of the image (by area) that contributes most to a given class prediction.

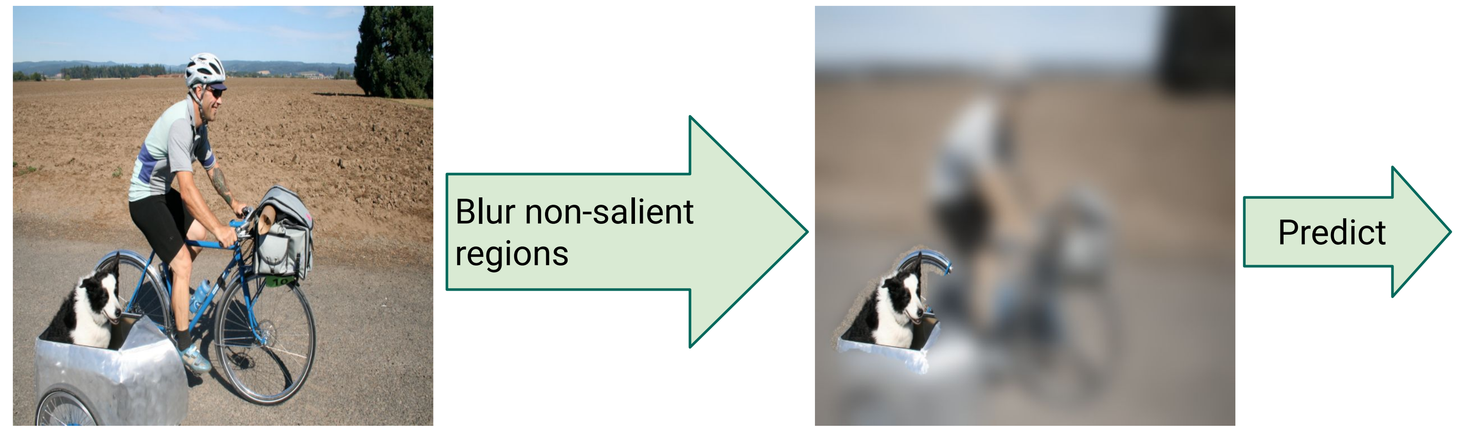




Performance Information Curves (PIC)

Performance Information Curve (PIC): A method for assessing image-based attributions

1. Identify salient regions in the image.
2. Remove irrelevant information by blurring.
3. Determine amount of information in the image. We approximate information/entropy by the compressed image size using the webp format.
4. Calculate a performance metric at each information level. We use two performance metrics: Accuracy for **AIC** and the relative softmax for **SIC.**



Correctly predicted Border collie at 410KB information:

```
(#1, score 5.564) Border collie
(#2, score 5.449) bulbul
(#3, score 4.216) collie
(#4, score 3.712) Cardigan Welsh corgi
(#5, score 3.688) Pembroke Welsh corgi
```

Compressed size: 1205KB
Salient pixels: 100%

Compressed size: 410KB
Salient pixels: 5%

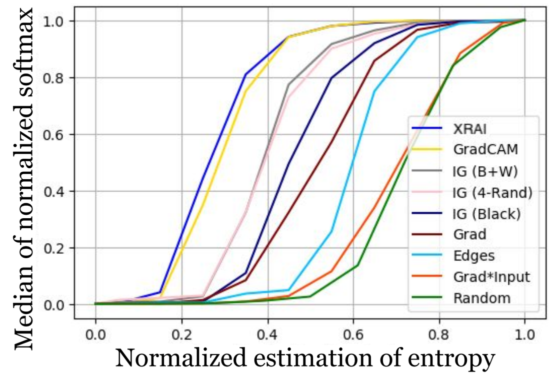
Benefits of PICs: Blurring is a relatively natural alteration (e.g., bokeh images are real), and the techniques measure *information content*, rather than the revealed area (information content for a given area can vary within and between images).

Results

Using the AIC and SIC methods described above, XRAI is consistently better for both AIC/SIC and for localization metrics. For the localization metrics, we used the ImageNet dataset, which provides object location ground truth in the form of bounding boxes. We calculated the F1-score, Mean Absolute Error (MAE), and Area Under the Receiver Operator Characteristic (ROC) curve (AUC).

AIC/SIC Results:

Method	Resnet50-V2		Inception	
	SIC	AIC	SIC	AIC
XRAI	0.749	**0.728**	**0.720**	**0.727**
GradCam	**0.760**	0.727	0.703	0.724
IG (B+W)	0.575	0.579	0.601	0.634
IG (4-Rand)	0.623	0.636	0.595	0.638
IG (Black)	0.515	0.527	0.530	0.576
Grad	0.521	0.532	0.480	0.543
Grad*Input	0.315	0.392	0.298	0.409
Edges	0.473	0.552	0.403	0.514
Random	0.445	0.473	0.278	0.401



***Left:** Area under the curve for SIC and AIC for all methods. **Right:** Visualized.*

Localization results

Method	AUC	F1	MAE
XRAI	**0.836**	**0.786**	**0.149**
IG (Black)	0.710	0.674	0.219
IG (4-Rand)	0.709	0.674	0.223
IG (B+W)	0.729	0.681	0.216
GradCAM	0.742	0.715	0.194

***Above:** ImageNet segmentation dataset localization metrics*