Interpretability Machine Learning

Jintong Gao

2025.1.17

Interpretability Machine Learning

1. Interpretable Machine Learning.

2. Interpretable Machine Learning Methods.

Why need Interpretable Machine Learning?

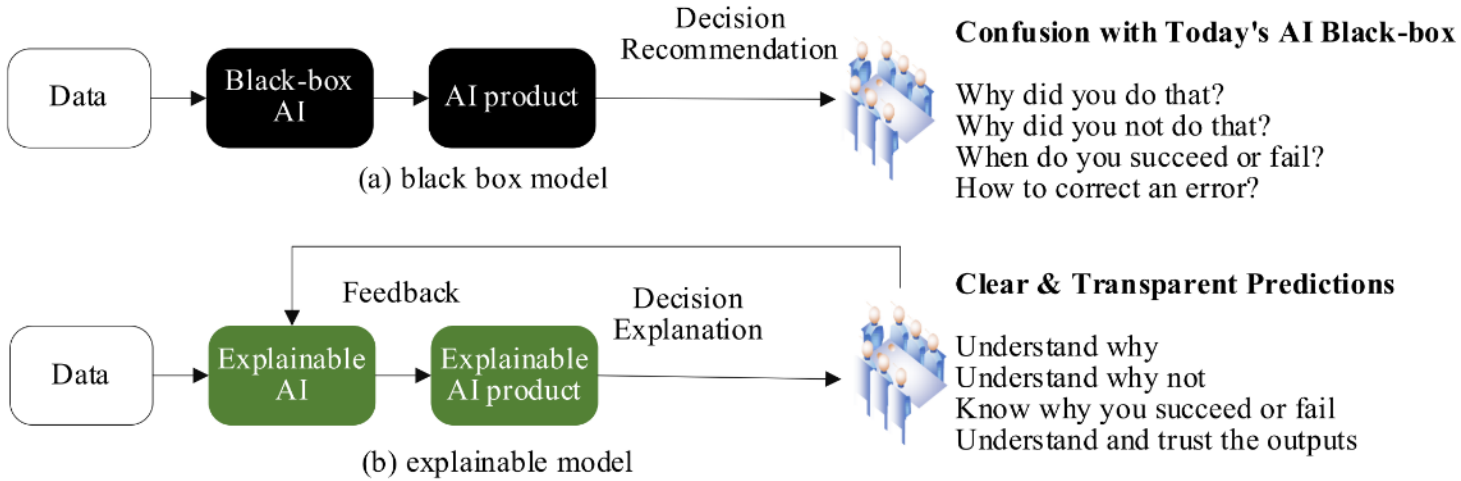
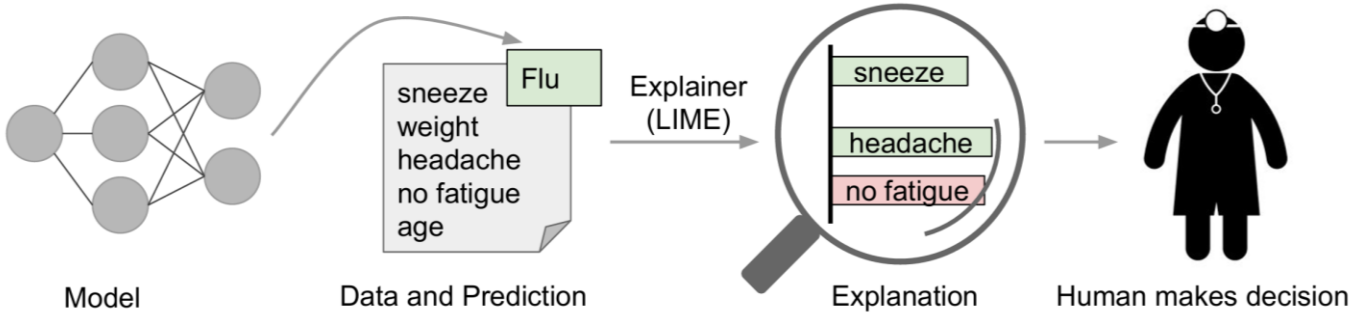
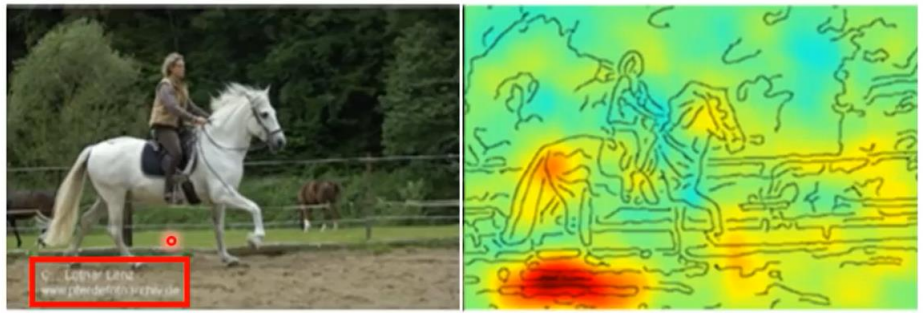


Fig. 3. The difference between black box AI models and explainable AI models.



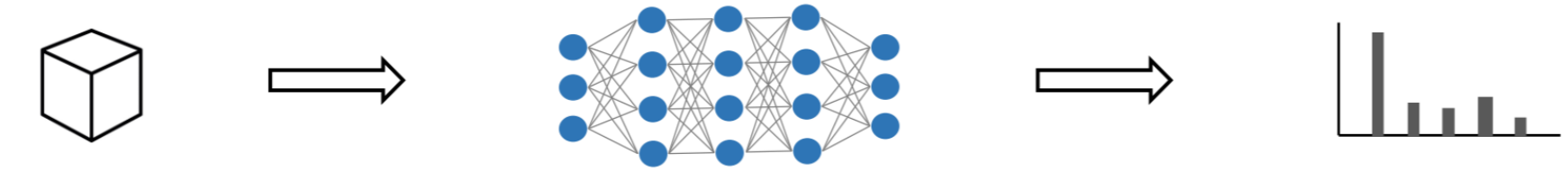


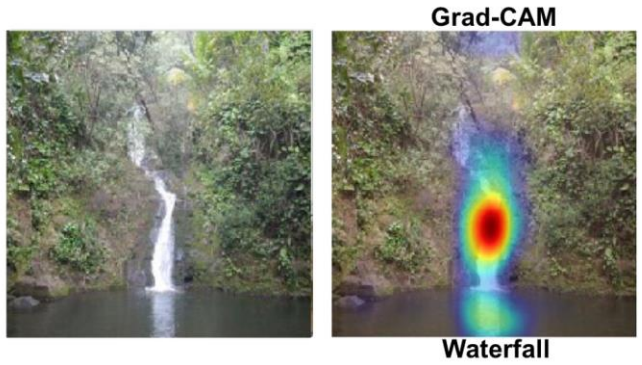
What is Interpretable Machine Learning?

Mengnan Du, Ninghao Liu, Xia Hu
Department of Computer Science and Engineering, Texas A&M University
{dumengnan,nhliu43,xiahu}@tamu.edu

CACM 2019

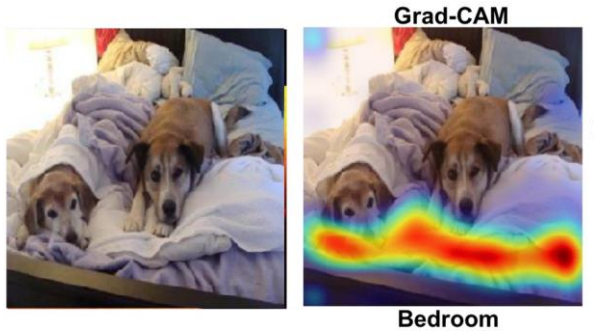
Interpretable machine learning gives machine learning models the ability to explain or to present their behaviors in **understandable** terms to **humans**.





Important concepts for `**Waterfall**`

Positive		Negative	
Neuron ID	Concept	Neuron ID	Concept
117	Waterfall	115	Corridor
106	Closet	166	Road
148	Water	494	Bus
143	Water	106	Laundromat
216	Stratified	412	Grid



Important concepts for `**Bedroom**`

Positive		Negative	
Neuron ID	Concept	Neuron ID	Concept
317	Bed	187	Spiralled
290	Bed	294	Pantry
226	Painting	26	Toiled
175	Cushion	9	Shoe shop
117	Waterfall	182	Amusement park

Types of Interpretable Machine Learning Methods

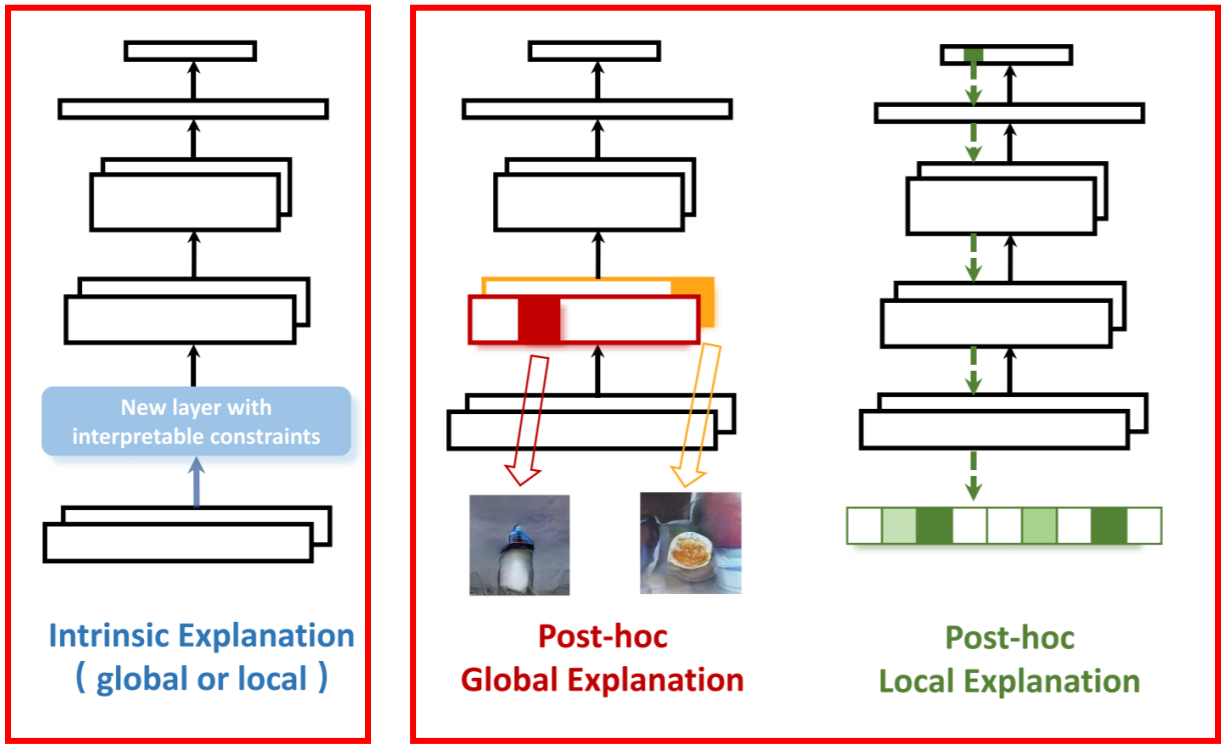
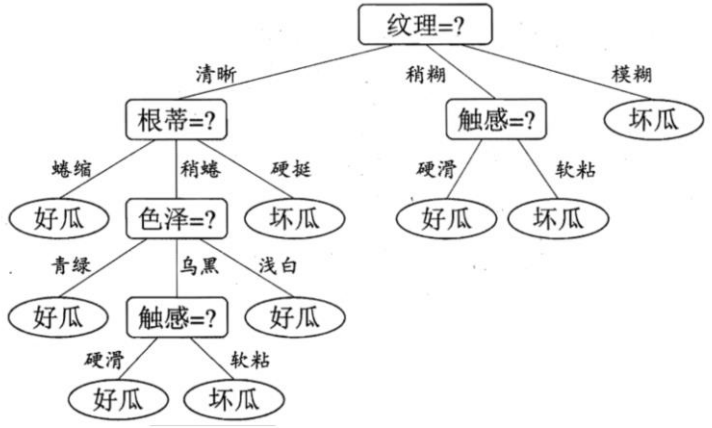
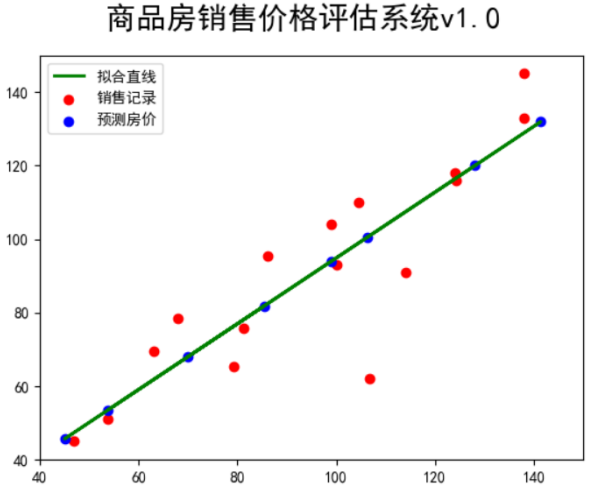


Figure 1: An illustration of three lines of interpretable machine learning techniques, taking DNN for example: Intrinsic explanation, Post-hoc global explanation of a model, and Post-hoc local explanation of a prediction.



(RNNs). Attention mechanism is advantageous in that it gives users the ability to interpret which parts of the input are attended by the model through visualizing the attention weight matrix for individual predictions. Attention mecha-



Post-hoc Interpretable

Local Explanation

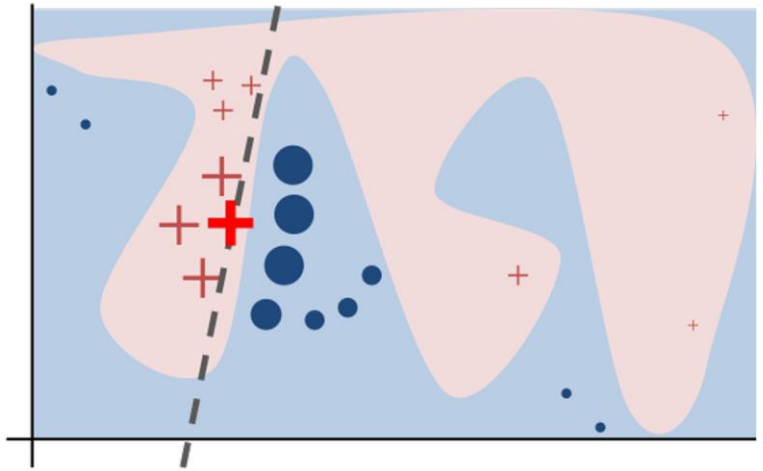
Why do you think *this image* is a cat?

Global Explanation

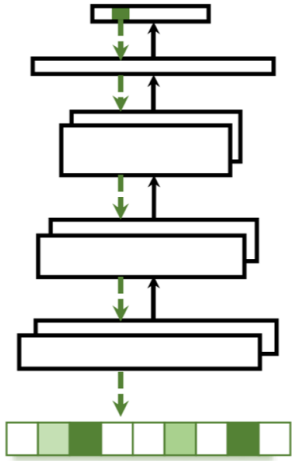
What does a "cat" look like?

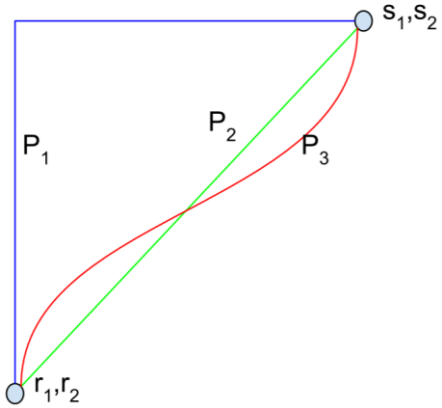
(not referred to a specific image)

Post-hoc Local Interpretable (Attribution methods)



1. Perturbative-based Methods.
2. Backpropagative-based Methods.
3. Path-based Methods.





for individual predictions. Local explanations target to identify the contributions of each feature in the input towards a specific model prediction. As local methods usually attribute a model's decision to its input features, they are also called *attribution* methods. In this section, we first intro-

Interpretable Machine Learning Methods

1. "why should I trust you? ": Explaining the predictions of any classifier(LIME). KDD 2016.

2. Grad-Cam: Visual explanations from deep networks via gradient-based localization. ICCV 2017.

3. Axiomatic attribution for deep networks(IG). ICML 2017.

4. Unlearning-based Neural Interpretations(UNI). ICLR 2025.

“Why Should I Trust You?” Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

KDD 2016

Motivation: LIME can explain the predictions of any classier or regressor(**model-agnostic**) in a faithful way, by approximating it locally with **an interpretable model** .

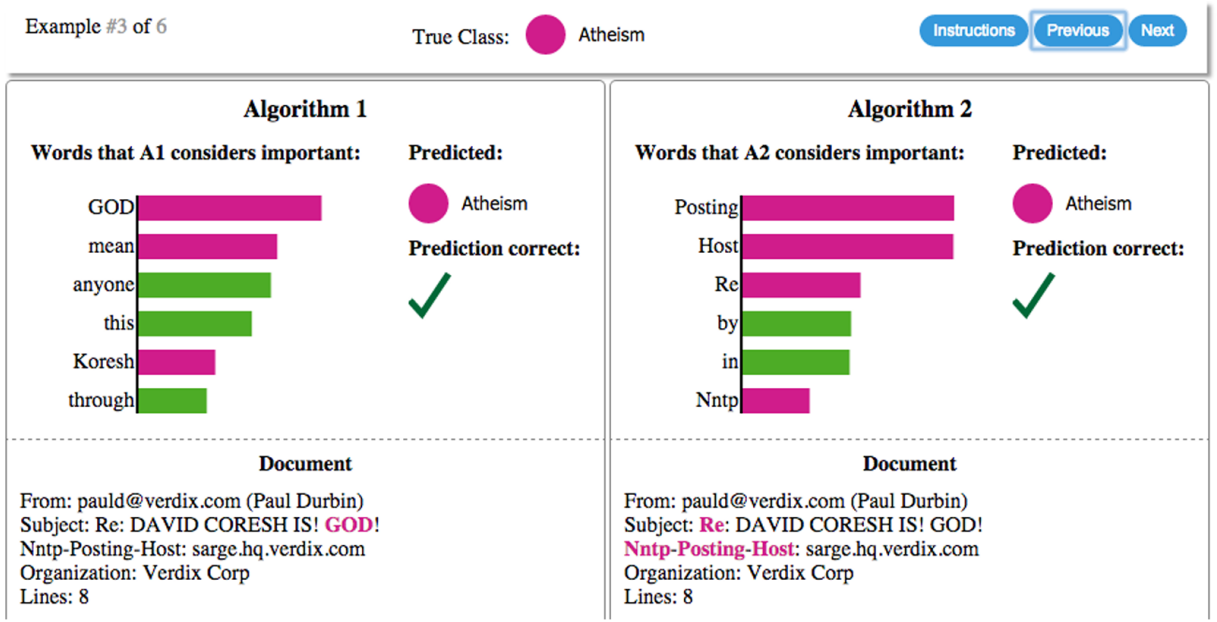


Figure 2: Explaining individual predictions of competing classifiers trying to determine if a document is about "Christianity" or "Atheism". The bar chart represents the importance given to the most relevant words, also highlighted in the text. Color indicates which class the word contributes to (green for "Christianity", magenta for "Atheism").

LIME

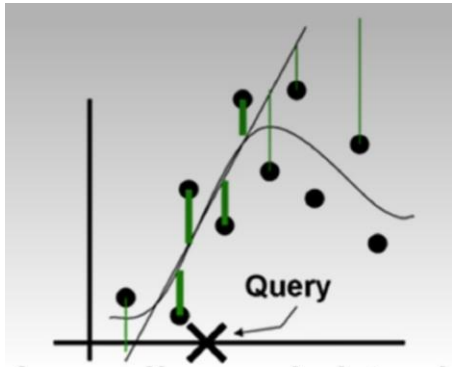
Original Image

超像素分割算法
Super Pixel

SP_1	SP_2	SP_3	...	SP_M
1	0	1	...	1

0: 不存在(灰色色块); 1: 存在 (原始色块)

Perturbed Instances	P(tree frog)
	0.85
	0.00001
	0.52



Locally weighted regression

by π_x. In order to ensure both **interpretability** and **local fidelity**, we must minimize $\mathcal{L}(f, g, \pi_x)$ while having $\Omega(g)$ be low enough to be interpretable by humans. The explanation produced by **LIME** is obtained by the following:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \quad \mathcal{L}(f, g, \pi_x) + \Omega(g) \qquad (1)$$

$$\Omega(g) = \infty \mathbb{1}[\|w_g\|_0 > K]$$

distance function D (e.g. cosine distance for text, $L2$ distance for images) with width σ.

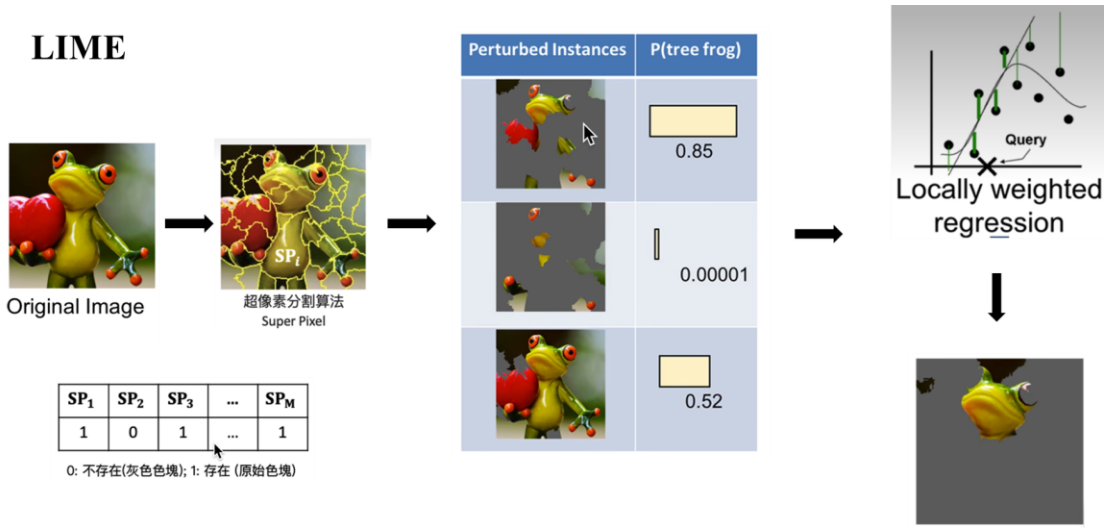
$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2 \qquad (2)$$

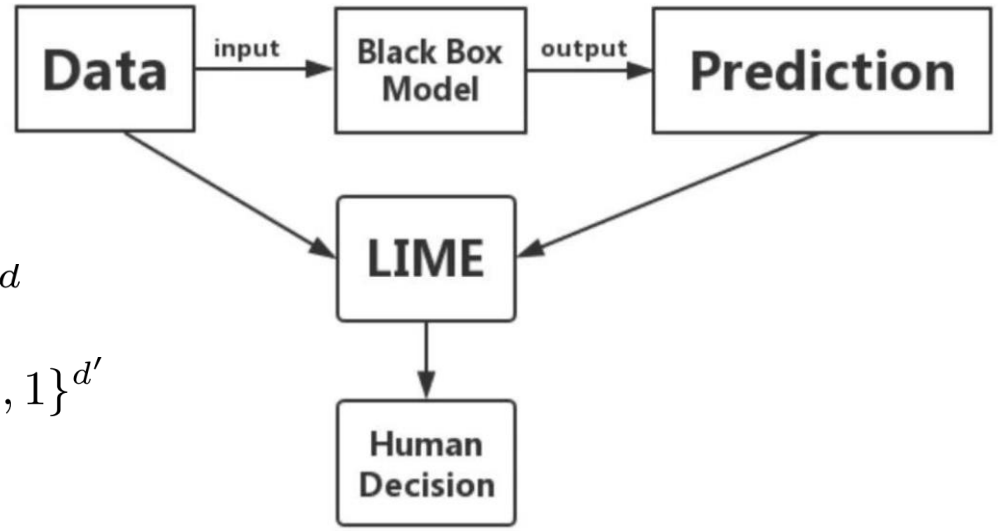
$$\pi_x(z) = exp(-D(x, z)^2 / \sigma^2)$$

Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K

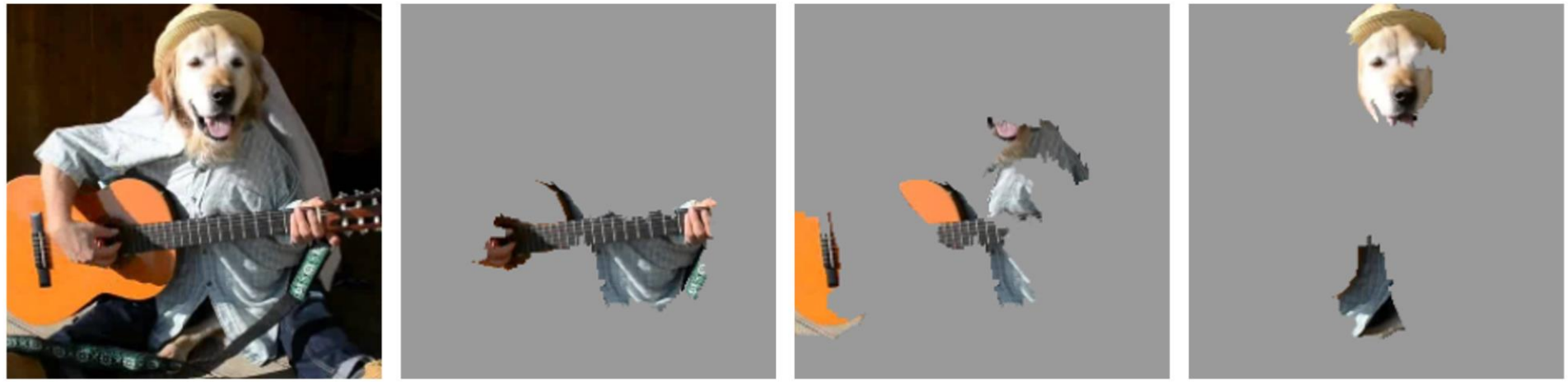
```
Z ← {}
for i ∈ {1, 2, 3, ..., N} do
    z'_i ← sample_around(x')
    Z ← Z ∪ ⟨z'_i, f(z_i), π_x(z_i)⟩
end for
w ← K-Lasso(Z, K)  ▷ with z'_i as features, f(z) as target
return w
```





$$x \in \mathbb{R}^d$$

$$z' \in \{0, 1\}^{d'}$$



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

(a) Husky classified as wolf (b) Explanation

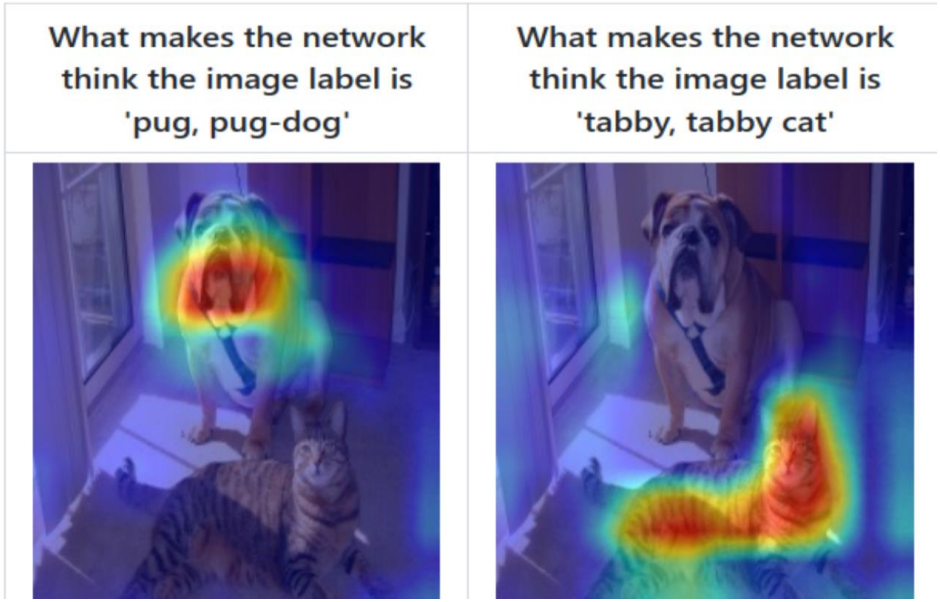
Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

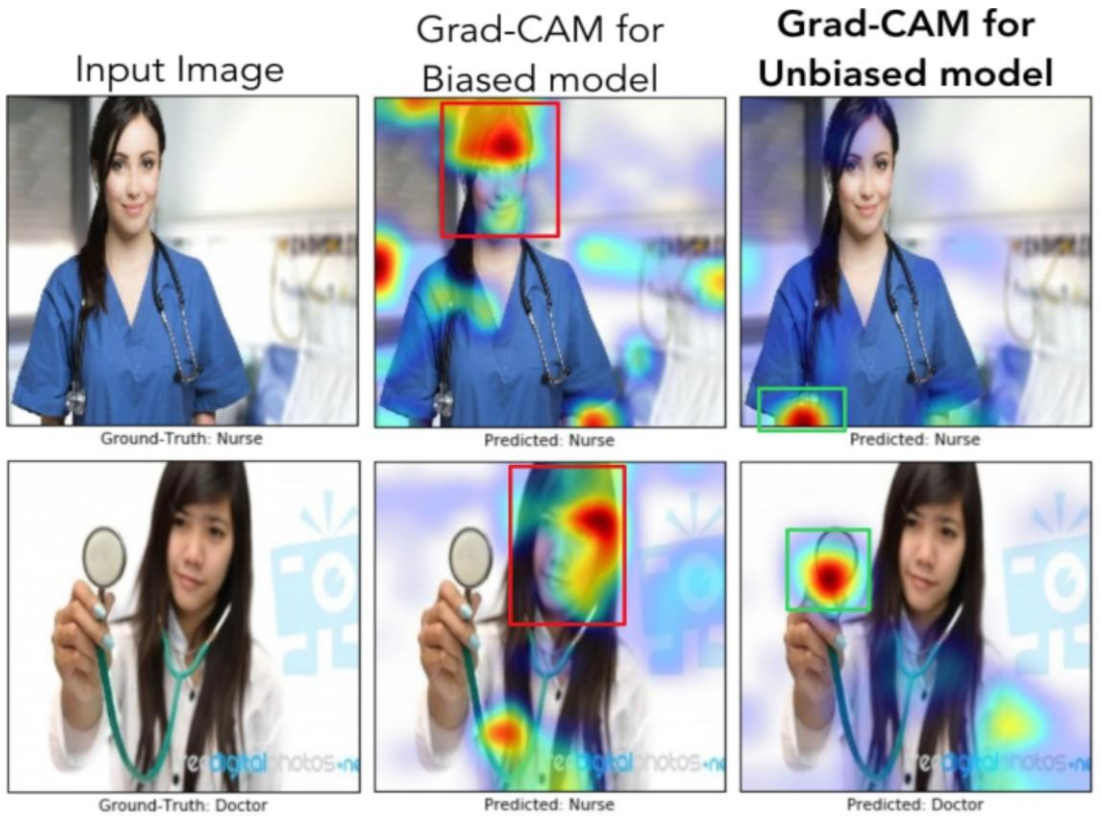
Grad-CAM:
Visual Explanations from Deep Networks via Gradient-based Localization

Ramprasaath R. Selvaraju[1*] Michael Cogswell[1] Abhishek Das[1] Ramakrishna Vedantam[1*]
Devi Parikh[1,2] Dhruv Batra[1,2]
[1]Georgia Institute of Technology [2]Facebook AI Research
{ramprs, cogswell, abhshkdz, vrama, parikh, dbatra}@gatech.edu

ICCV 2017

Motivation: Grad-CAM, a class-discriminative localization technique that can generate visual explanations from **any** CNN-based network **without** requiring **architectural** changes or **re-training**.





CAM(Class Activation Mapping)

$$\mathrm{CAM}(x, y) = \sum_k w_k^c \cdot f_k(x)$$

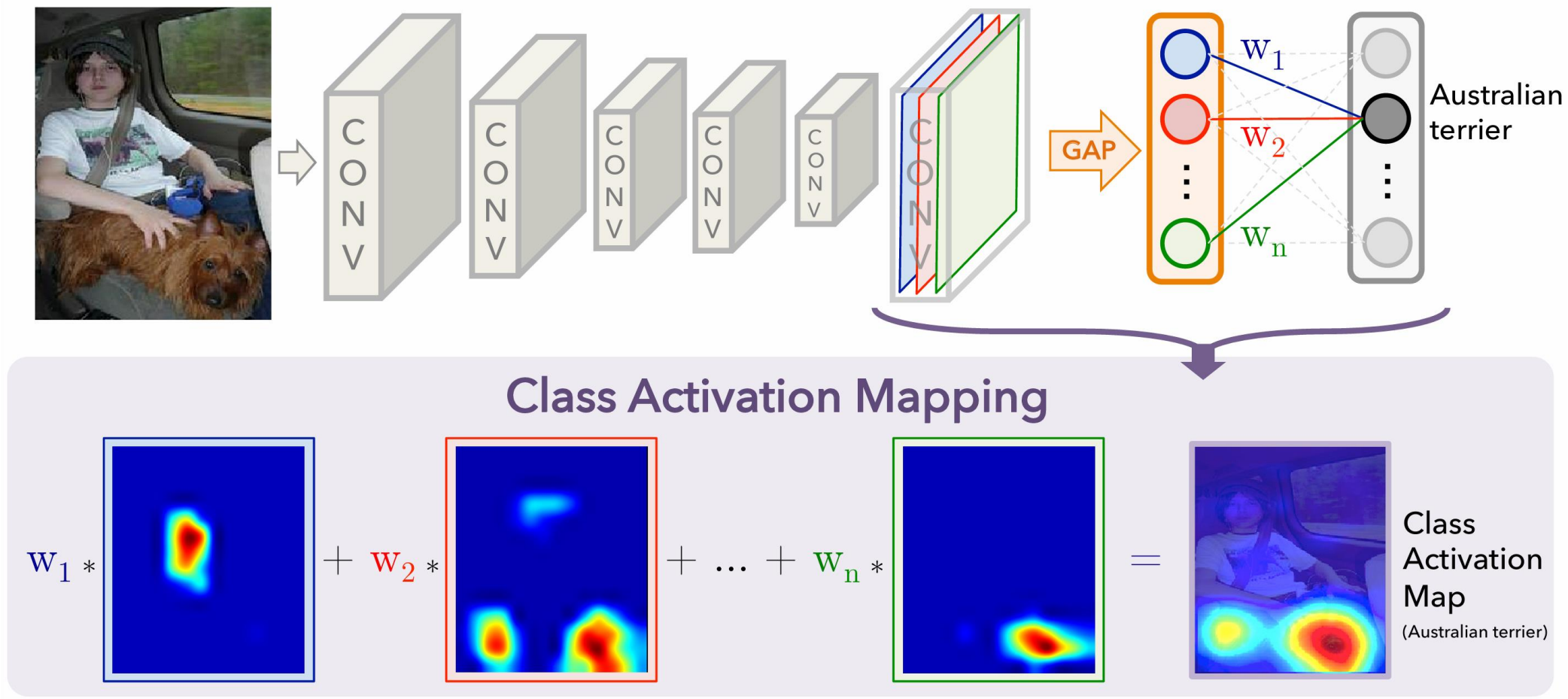


Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

Grad-CAM
(Gradient-weighted Class Activation Mapping)

$$\text{Grad-CAM}(x, y) = \text{ReLU}\left(\sum_{k=1}^{C} \alpha_k^c \cdot f_k(x)\right)$$

$$\alpha_k^c = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{\partial y^c}{\partial f_{k,i,j}}$$

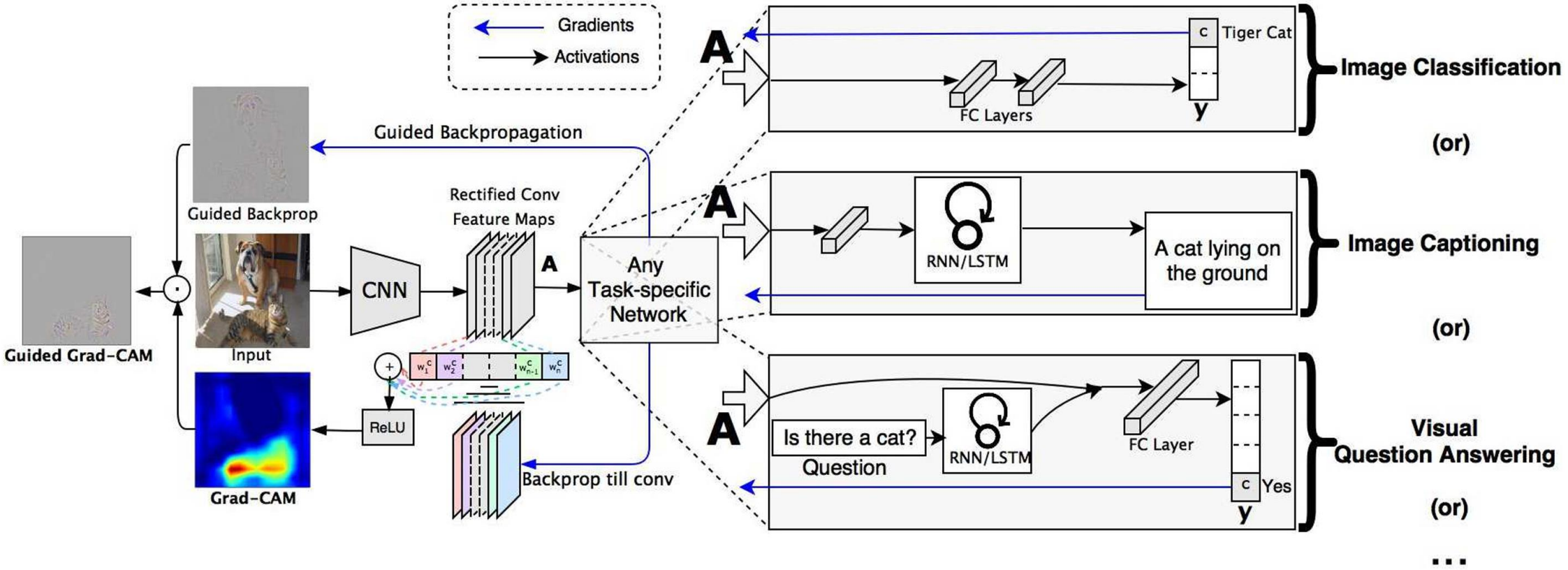


Figure 2: Grad-CAM overview: Given an image and a class of interest (*e.g.*, 'tiger cat' or any other type of differentiable output) as input, we forward propagate the image through the CNN part of the model and then through task-specific computations to obtain a raw score for the category. The gradients are set to zero for all classes except the desired class (tiger cat), which is set to 1. This signal is then backpropagated to the rectified convolutional feature maps of interest, which we combine to compute the coarse Grad-CAM localization (blue heatmap) which represents where the model has to look to make the particular decision. Finally, we pointwise multiply the heatmap with guided backpropagation to get Guided Grad-CAM visualizations which are both high-resolution and concept-specific.

Axiomatic Attribution for Deep Networks

Mukund Sundararajan [*1] **Ankur Taly** [*1] **Qiqi Yan** [*1]

ICLR 2017

Motivation: We study the problem of attributing the prediction of a deep network to its input features. + **Integrated Gradients (IG)** needs a few calls to the standard gradient operator.

Gradient Saturation Problem

Taking the local gradients of a model's output confidence map $F_c(x)$ – for target class c – is a tried and tested method for generating explanations. Commonly termed Simple Gradients (Erhan et al., 2009; Baehrens et al., 2010; Simonyan et al., 2013), $\mathcal{A}_i^{\text{SG}}(x) = \nabla_{x_i} F_c(x)$ can be efficiently computed for most model architectures. However, it encounters output saturation when activation functions like ReLU and Sigmoid are used, leading to zero gradients (hence null attribution) even for important features (Sundararajan et al., 2017; 2016). DeepLIFT (Shrikumar et al., 2016) reduces saturation by

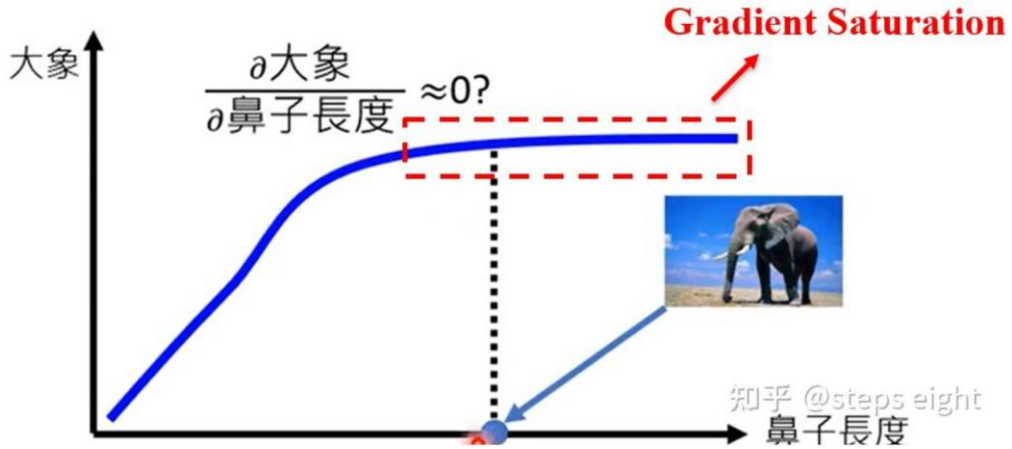


图3. 大象鼻子长度的梯度饱和归因案例

$$Relu(x) = max(0, x)$$

$$f(x) = 1 - ReLU(1 - x) = \begin{cases} x, x < 1 \\ 1, x \geq 1 \end{cases} \quad (1)$$

以 $x = 0$ 为基线，当 $x = 1$ 的时候可以得到不同的结果。即 $f(0) = 0, f(1) = 1$，但是 $f'(1) = 0$ 不满足梯度处处存在的条件（$x \geq 1$ 均是），即违背了敏感性。

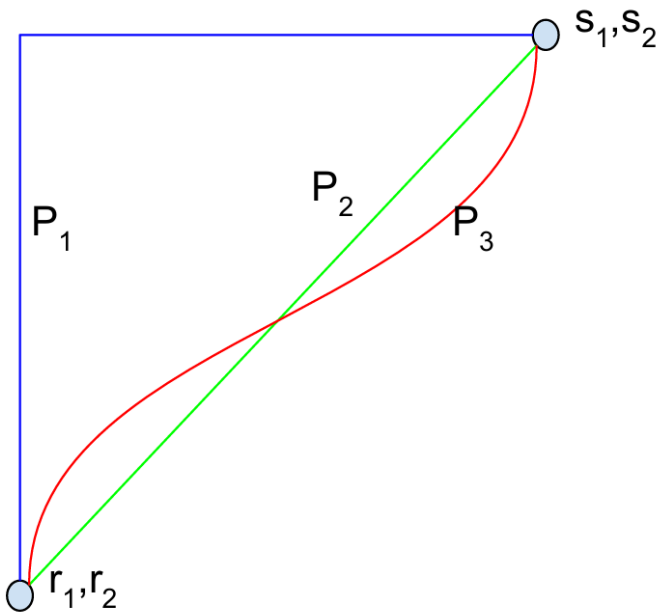


Figure 1. Three paths between an a baseline (r_1, r_2) and an input (s_1, s_2). Each path corresponds to a different attribution method. The path P_2 corresponds to the path used by integrated gradients.

Formally, suppose we have a function $F : \mathsf{R}^n \to [0, 1]$ that represents a deep network. Specifically, let $x \in \mathsf{R}^n$ be the input at hand, and $x' \in \mathsf{R}^n$ be the baseline input. For image networks, the baseline could be the black image, while for text models it could be the zero embedding vector.

We consider the straightline path (in R^n) from the baseline x' to the input x, and compute the gradients at all points along the path. Integrated gradients are obtained by cumulating these gradients. Specifically, integrated gradients are defined as the path intergral of the gradients along the straightline path from the baseline x' to the input x.

The integrated gradient along the i^{th} dimension for an input x and baseline x' is defined as follows. Here, $\frac{\partial F(x)}{\partial x_i}$ is the gradient of $F(x)$ along the i^{th} dimension.

$$\mathsf{IntegratedGrads}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha \tag{1}$$

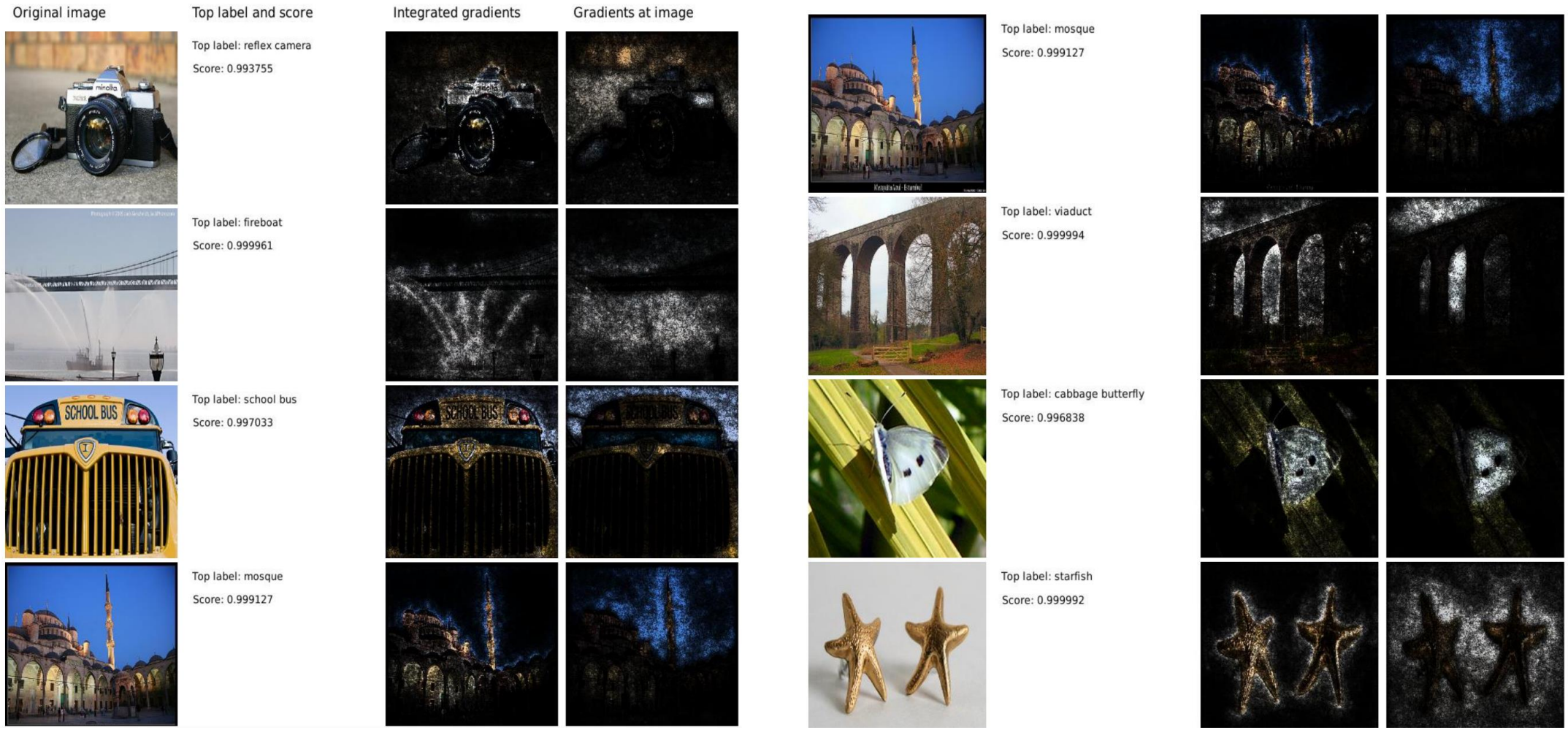


Figure 2. **Comparing integrated gradients with gradients at the image.** Left-to-right: original input image, label and softmax score for the highest scoring class, visualization of integrated gradients, visualization of gradients*image. Notice that the visualizations obtained from integrated gradients are better at reflecting distinctive features of the image.

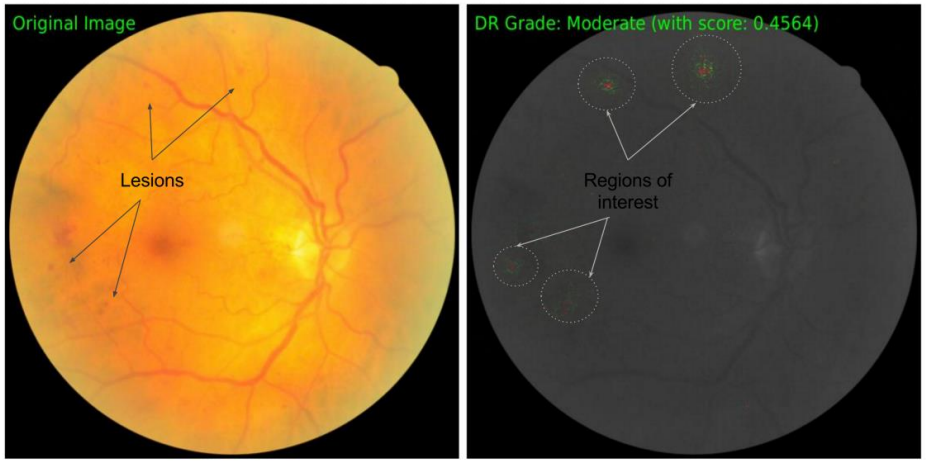


Figure 3. **Attribution for Diabetic Retinopathy grade prediction from a retinal fundus image.** The original image is show on the left, and the attributions (overlayed on the original image in gray scaee) is shown on the right. On the original image we annotate lesions visible to a human, and confirm that the attributions indeed point to them.

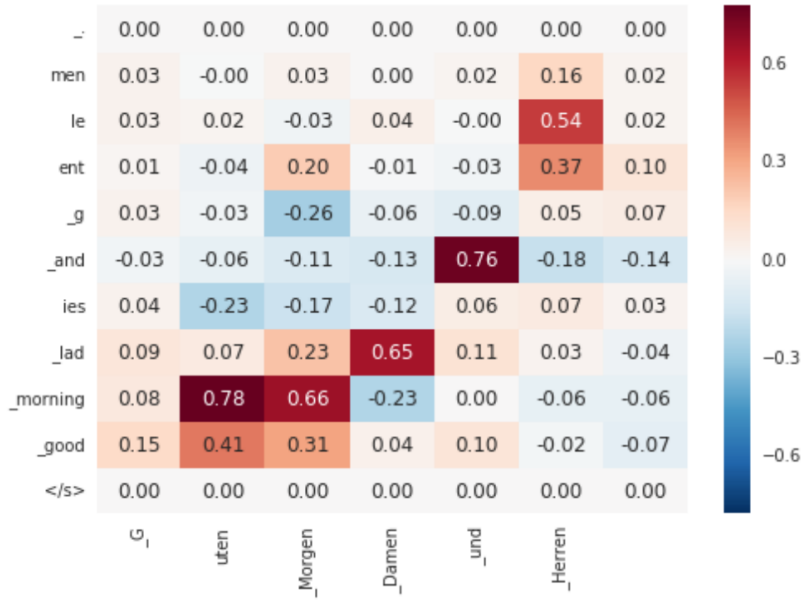


Figure 5. **Attributions from a language translation model.** Input in English: "good morning ladies and gentlemen". Output in German: "Guten Morgen Damen und Herren". Both input and output are tokenized into word pieces, where a word piece prefixed by underscore indicates that it should be the prefix of a word.

UNLEARNING-BASED NEURAL INTERPRETATIONS

Ching Lam Choi*
CSAIL, Department of EECS
Massachusetts Institute of Technology
chinglam@mit.edu

Alexandre Duplessis
Department of Computer Science
University of Oxford
alexandre.duplessis@cs.ox.ac.uk

Serge Belongie
Pioneer Centre for AI
University of Copenhagen
s.belongie@di.ku.dk

ICLR 2025 (6, 8, 8, 10)

Motivation:

Current baselines defined using static functions—constant mapping, averaging or blurring—inject harmful color, texture or frequency assumptions that **deviate** from model behavior.

UNI(Unlearning-based Neural Interpretations) to compute an **debiased** and adaptive **baseline** by perturbing the input towards an unlearning direction of steepest ascent.

Static baseline functions?

3.2 POST-HOC BIASES ARE IMPOSED

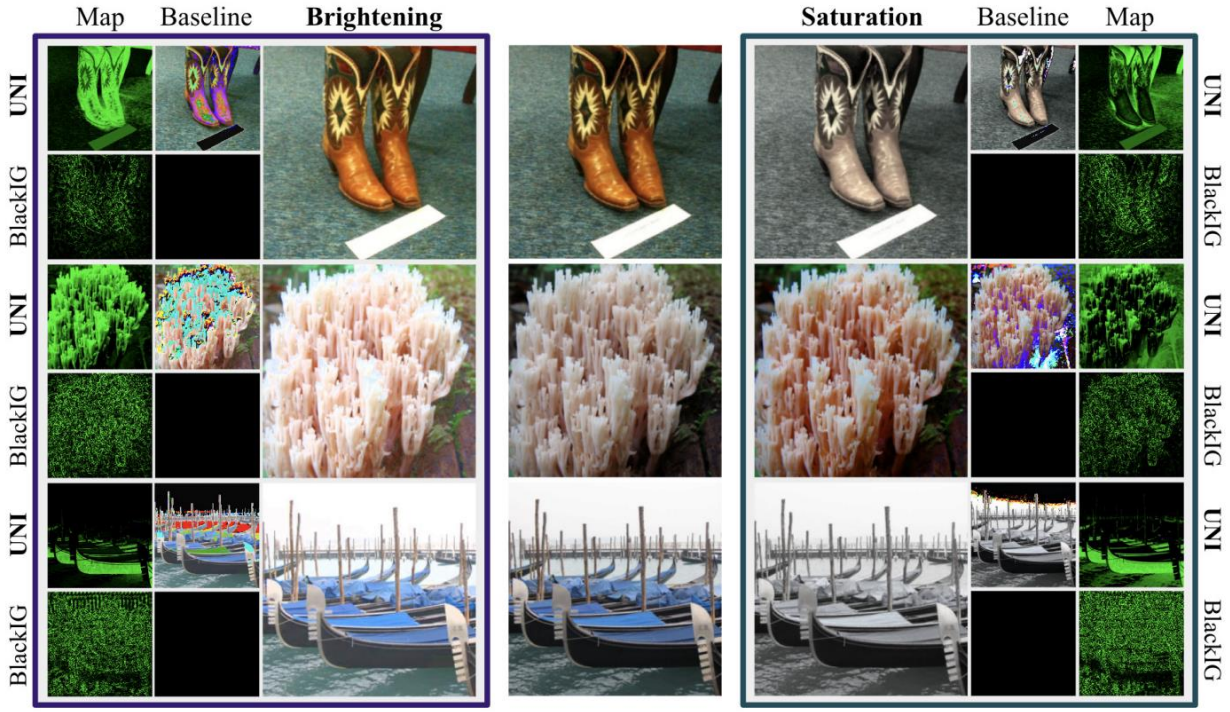


Figure 3: When the brightness or saturation is altered, IG with a black baseline fails to identify dark features, such as the boat's hull (R3) or the top of the boot (R1).

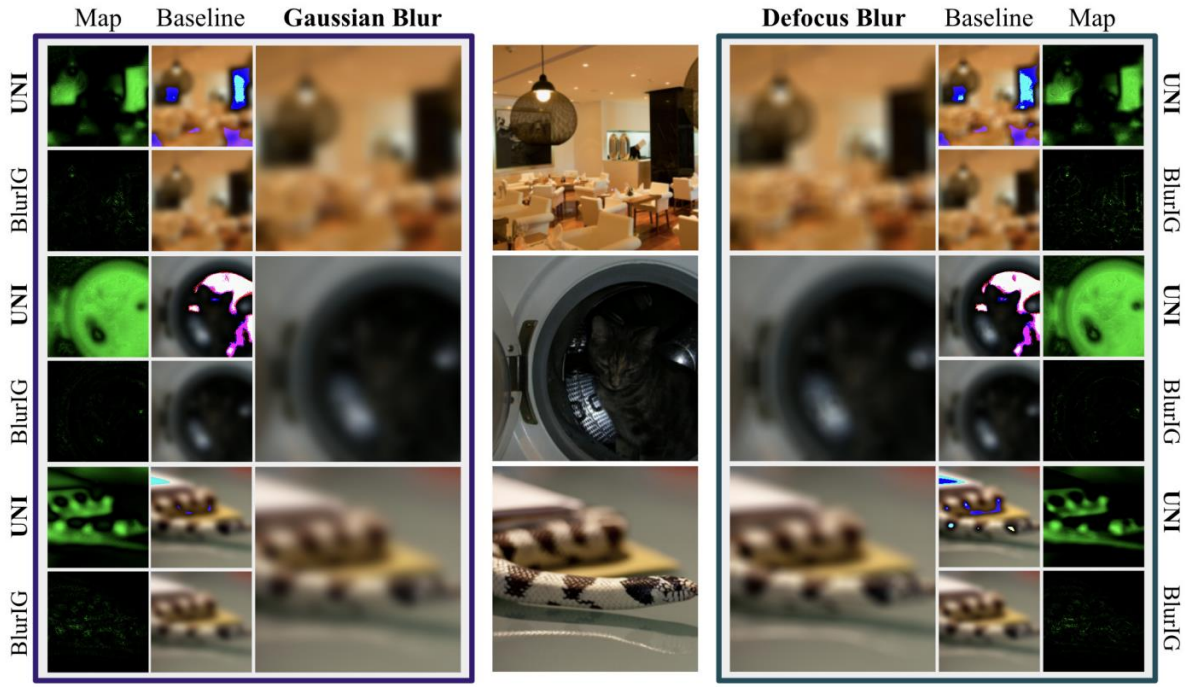


Figure 4: Under gaussian or defocus blur, IG with a blurred baseline suffers from saturation; has overly smooth texture; does not yield meaningful features.

Color

Texture

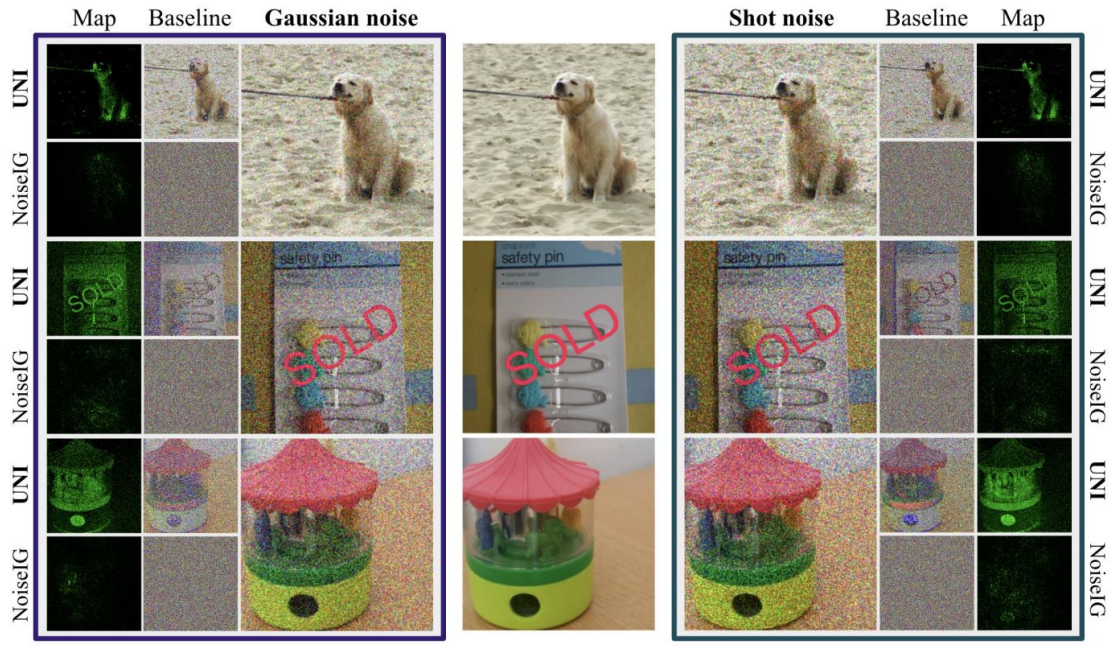


Figure 5: Gaussian and shot noise create visual artifacts prominent in noised-baseline IG. Frequency bias leads to disparate scores for adjacent pixels.

Frequency

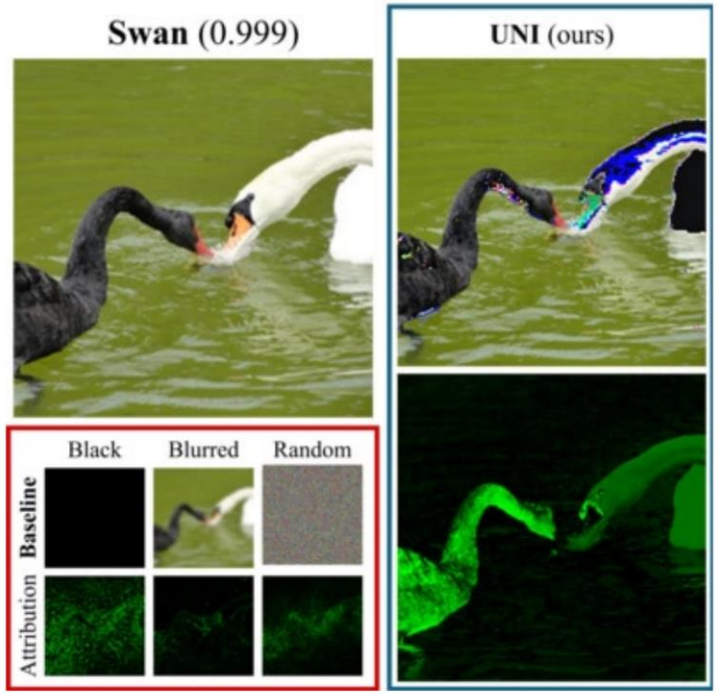


Figure 2: We visualise post-hoc biases imposed by static baselines—black baseline (colour), blurred (texture), random (frequency). UNI learns to mask out predictive features used by the model, generating reliable attributions.

Post-hoc attribution can impose new biases. We approach the baseline challenge from the fresh lens of post-hoc biases. We show that static baselines (*e.g.* black, blurred, random noise) inject additional colour, texture and frequency assumptions that are not present in the original model's decision rule, which leads to explanation infidelity and inconsistency.

suboptimal choice of static baseline. The observation that poor baseline choices create *attribution bias* has so far been overlooked. As such, we depart entirely from the line of work on alternative static baseline towards adaptively (un)learning baselines with gradient-based optimisation. *UNI eliminates all external assumptions except for the model's own predictive bias.*

$\eta = 1$; l_2 PGD with $T = 10$ steps, a budget of $\varepsilon = 0.25$, step size $\mu = 0.1$

UNI(Unlearning-based Neural Interpretations)

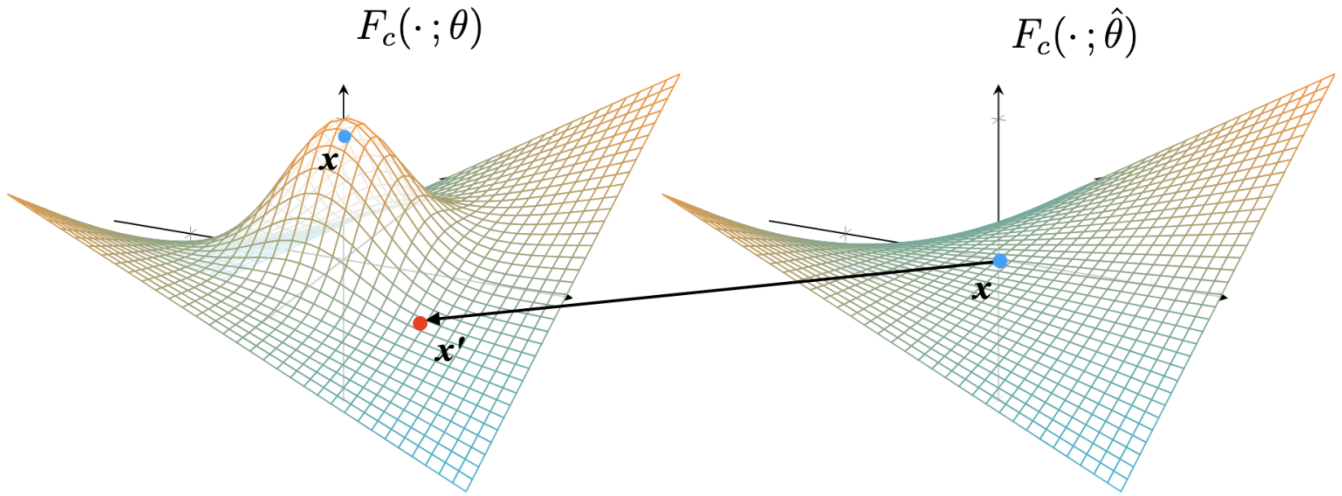


Figure 1: *Left:* Confidence of original model θ at image x and baseline x'. *Right:* Confidence of *unlearned* model $\hat{\theta}$ at image x. After unlearning in the model space $\theta \longmapsto \hat{\theta}$, we optimise the baseline to match the unlearned input confidence, such that $F_c(x';\theta) \approx F_c(x;\hat{\theta})$.

be used to match the corresponding, “featureless" input, where salient features have been deleted during the unlearning process. While the connection to interpretability is new, a few recent works

x

x'

δ^T

Algorithm 1 UNI: unlearning direction, baseline matching and path-attribution

1: **Given** model $F(\cdot,\,\theta)$; inputs (x, y)

2: **Choose** unlearning step-size η; PGD steps T, budget ε, step-size μ; Riemann approximation steps B

3: **Initialise** perturbation δ^0

4: **Unlearning direction.**

$$\hat{\theta} = \theta + \eta \frac{\nabla_\theta \mathcal{L}(F_c(x;\theta),y)}{\|\nabla_\theta \mathcal{L}(F_c(x;\theta),y)\|}$$

5: **for** $t = 0, \cdots, T-1$ **do**

$$\mathcal{C} = D_{KL}(F(x;\hat{\theta}) \parallel F(x+\delta^t;\theta))$$

$$\delta^{t+1} = \delta^t - \mu\, \nabla_\delta \mathcal{C}$$

$$\delta^{t+1} = \varepsilon \frac{\delta^{t+1}}{\|\delta^{t+1}\|}$$

bound the distance $||x - x'||$ to a certain value ε.

6: **end for**

7: **Baseline** definition $x' = x + \delta^T$

8: **Attributions** computation: $\mathcal{A}_i^{\text{UNI}}(x)$

$$= \frac{(x_i - x_i')}{B} \sum_{k=1}^{B} \nabla_{x_i} F_c\left(x' + \frac{k}{B}(x - x');\theta\right)$$

$B = 15$

Machine Unlearning

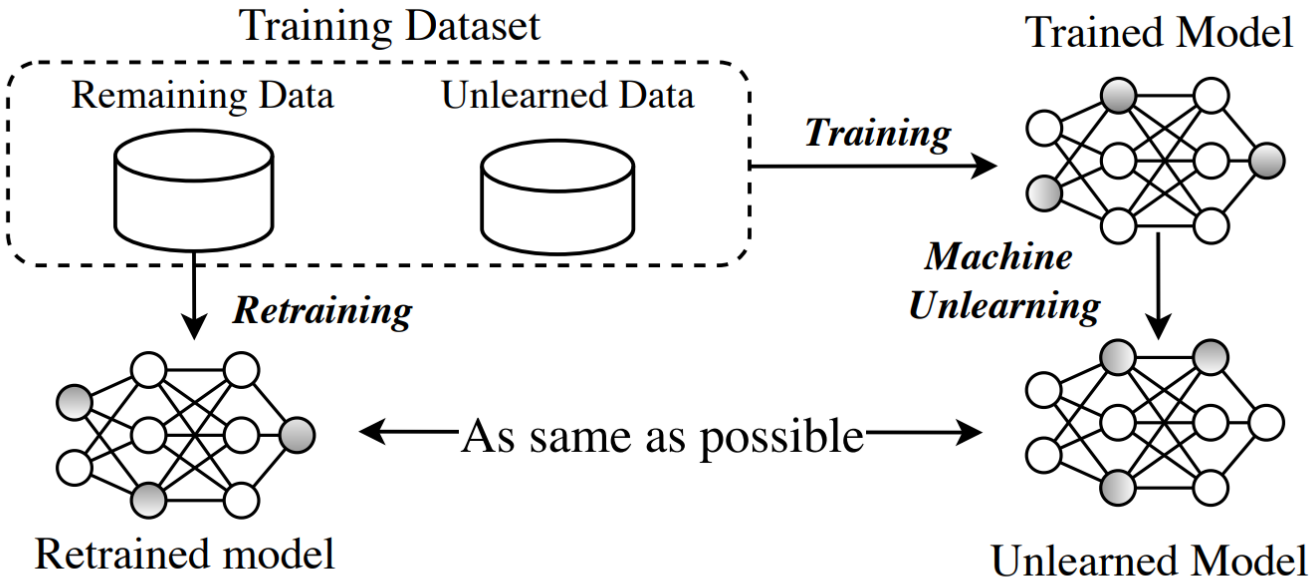


Fig. 1. Illustration of Machine Unlearning.

Having models forget necessitates knowledge of exactly how individual training points contributed to model parameter updates. Prior work showed this is possible when the learning

which the user elected to erase. Put another way, unlearning guarantees that training on a point and unlearning it afterwards will produce the same distribution of models that not training on the point at all, in the first place, would have produced.

Low Curvature.

the computed attributions. We substantiate the intuition that a smooth and regular path is preferred by analysing the Riemannian sum calculation. Assuming that the function $g : \alpha \in [0,1] \mapsto \nabla F_c\left(x' + \alpha(x - x')\right)$ is derivable with a continuous derivative (i.e. $\mathcal{C}^1$) on the segment $[x', x]$, elementary calculations and the application of the Taylor-Lagrange inequality give the following error in the Riemann approximation of the attribution,

$$\left|(x_i - x_i') \int_{\alpha=0}^{1} g\left(\alpha\right)\, d\alpha - \frac{(x_i - x_i')}{B} \sum_{k=1}^{B} g\left(\frac{k}{B}\right)\right| \leq \frac{M||x - x'||^2}{2B} \tag{2}$$

where $M = \max_{\alpha\in[0,1]} \frac{\mathrm{d}g}{\mathrm{d}\alpha} = \max_{\alpha\in[0,1]} \frac{\partial^2 F_c\left(x'+\alpha(x-x')\right)}{\partial\alpha^2}$ exists by continuity of g' on $[0, 1]$. Thus, lower curvature along the path implies a lower value of the constant M, which in turn implies a lower error in the integration calculation. A smaller value B of Riemann steps is needed to achieve the same precision. More generally, a low curvature (i.e. eigenvalues of the hessian) on and in a neighbourhood of the baseline and path reduces the variability of the calculated gradients under small-norm perturbations, increasing the sensitivity and consistency of the method. Empirically, we

Monotonic. Intuitively, the path γ defined by interpolating from the “featureless" baseline x' to the input image x should be *monotonically increasing* in output class confidence. At the image level, for

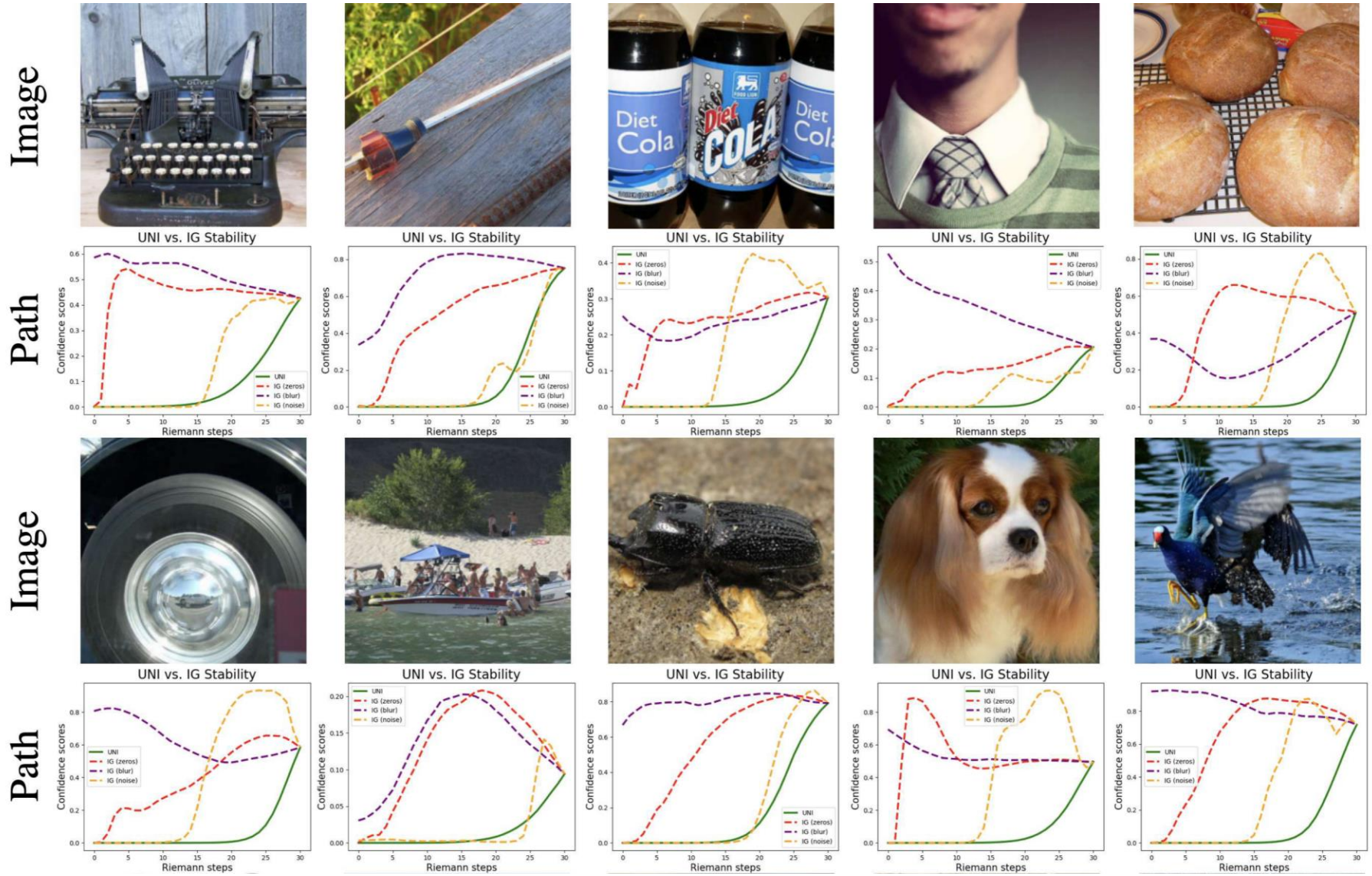


Figure 20: *Comparing paths (ResNet-18):* `UNI` discovers geodesic paths of monotonically increasing output confidence, preserving the completeness property required for robust attributions.

5.1 FAITHFULNESS

We report MuFidelity scores (Bhatt et al., 2021), *i.e.* the faithfulness of an attribution function $\mathcal{A}$, to a model F, at a sample x, for a subset of features of size $|S|$, given by $\mu_f(F, \mathcal{A}; x) = \text{corr}_{S \in \binom{[d]}{|S|}} \left(\sum_{i \in S} \mathcal{A}(i, F, c, x),\ F_c(x) - F_c(x_{[x_s = \bar{x}_s]}) \right)$. We record the (absolute) correlation

Table 2: *MuFidelity scores* measure the correlation between a subset of pixels' impact on the output (*i.e.* change in predictive confidence) and assigned saliency scores. Since attribution methods can yield strong positive or negative correlations, we report the absolute scores.

	UNI	IG	BlurIG	GIG	AGI	GBP	DeepLIFT
ResNet-18	**.12** ±.124	.06 ±.068	.07 ±.076	.07 ±.080	.10 ±.110	.09 ±.094	.08 ±.082
EfficientNetv2s	**.06** ±.046	.05 ±.043	.05 ±.044	.05 ±.044	.06 ±.045	.05 ±.043	.05 ±.043
ConvNeXt-Tiny	.16 ±.115	.11 ±.086	.15 ±.121	**.18** ±.149	.17 ±.131	.09 ±.072	.11 ±.084
VGG-16-bn	**.18** ±.141	.08 ±.066	.09 ±.076	.13 ±.108	.14 ±.104	.13 ±.108	.10 ±.082
ViT-B_16	**.15** ±.114	.10 ±.074	.10 ±.077	.11 ±.079	.14 ±.104	.09 ±.070	.10 ±.072
Swin-T-Tiny	**.13** ±.100	.09 ±.071	.12 ±.102	.12 ±.104	.13 ±.102	.09 ±.069	.10 ±.076

BlurIG, GIG, AGI, GBP, DeepLift. Furthermore, we report deletion and insertion scores (Petsiuk et al., 2018)—a causally-motivated evaluation metric for interpretability methods—which measures the decrease (deletion) or increase (insertion) of a model's output confidence as salient pixels are removed (from the original image) or inserted (into a featureless baseline). A steep drop in model

Table 3: *Deletion AUC* ↓ measures how confidence drops as pixels are removed *(lower = better)*.

	UNI	IG	BlurIG	GIG	AGI	GBP	DeepLIFT
ResNet-18	**.06** ±.128	.10 ±.174	.27 ±.252	.11 ±.150	.13 ±.147	.08 ±.160	.13 ±.165
EfficientNetv2s	.19 ±.212	.26 ±.217	.50 ±.158	.19 ±.216	**.18** ±.207	.23 ±.163	.27 ±.215
ConvNeXt-Tiny	**.11** ±.139	.16 ±.164	.46 ±.172	.21 ±.160	.17 ±.123	.16 ±.099	.21 ±.162
VGG-16-bn	**.08** ±.143	.12 ±.181	.18 ±.241	.10 ±.163	.14 ±.178	.14 ±.194	.12 ±.186
ViT-B_16	.14 ±.185	.22 ±.207	.60 ±.166	.17 ±.190	**.13** ±.152	.23 ±.141	.17 ±.189
Swin-T-Tiny	**.13** ±.181	.22 ±.217	.47 ±.174	.22 ±.207	.21 ±.172	.21 ±.123	.23 ±.207

Table 4: *Insertion AUC* ↑ measures how confidence rises as pixels are inserted *(higher = better)*.

	UNI	IG	BlurIG	GIG	AGI	GBP	DeepLIFT
ResNet-18	**.64** ±.138	.26 ±.045	.34 ±.131	.36 ±.048	.56 ±.068	.11 ±.066	.18 ±.042
EfficientNetv2s	**.64** ±.227	.38 ±.127	.51 ±.283	.37 ±.138	.38 ±.204	.23 ±.192	.37 ±.137
ConvNeXt-Tiny	**.63** ±.231	.21 ±.114	.40 ±.252	.56 ±.122	.52 ±.088	.22 ±.160	.17 ±.162
VGG-16-bn	**.56** ±.335	.37 ±.061	.31 ±.274	.38 ±.071	.47 ±.078	.26 ±.057	.17 ±.056
ViT-B_16	**.71** ±.237	.32 ±.107	.59 ±.292	.28 ±.125	.43 ±.089	.35 ±.172	.28 ±.123
Swin-T-Tiny	**.68** ±.245	.28 ±.145	.63 ±.282	.26 ±.153	.25 ±.156	.31 ±.202	.26 ±.152

5.2 ROBUSTNESS

Next, we evaluate UNI's robustness to fragility adversarial attacks on model interpretations. Following Ghorbani et al. (2019a), we design norm-bounded attacks to maximise the disagreement in attributions whilst constraining that the prediction label remains unchanged. We consider a standard l_∞ attack designed with FGSM (Goodfellow et al., 2014), with perturbation budget $\varepsilon_f = 8/255$.

$$\begin{aligned} \delta_f^* = &\arg \max_{\|\delta_f\|_p \leq \varepsilon_f} \frac{1}{d_X} \sum_{i=1}^{d_X} d\left(\mathcal{A}(i, F, c, x), \mathcal{A}(i, F, c, x + \delta_f)\right) \\ &\text{subject to} \quad \arg\max_{c'} F_{c'}(x) = \arg\max_{c} F_{c'}(x + \delta_f) = c \end{aligned} \tag{3}$$

Table 5: *Robustness*: Spearman's correlation coefficient. Higher scores indicate better path consistency pre/post FGSM attacks.

	UNI	IG	BlurIG	SG	DeepL
ResNet-18	**.271**	.088	.084	.014	.139
Eff-v2-s	**.302**	.009	.076	.008	.018
ConvNeXt-T	**.292**	.010	.127	.011	.012
VGG-16-bn	**.290**	.143	.098	.014	.108
ViT-B-16	**.319**	.018	.066	.023	.023
SwinT	**.271**	.088	.084	.014	.139

Table 6: *Robustness*: Top-1000 pixel intersection. Higher percentages indicate better attribution reliability pre/post FGSM attacks.

	UNI	IG	BlurIG	SG	DeepL
ResNet-18	**37.3**	20.0	25.3	18.2	24.8
Eff-v2-s	**39.4**	17.4	23.3	18.6	18.0
ConvNeXt-T	**34.8**	15.0	26.2	16.7	15.1
VGG-16-bn	**35.7**	25.5	25.3	18.8	25.2
ViT-B-16	**40.7**	17.1	21.7	19.6	17.2
SwinT	**37.3**	20.0	25.3	18.2	24.8

5.3 STABILITY

We compare UNI and other methods' sensitivity to Riemann approximation noise, which manifests in visual artefacts and misattribution of salient features. As seen from Figures 7, 10, UNI reliably finds unlearned, "featureless" baselines for consistent attribution, regardless of the number of approximation steps $B \in \{1, 15, 30\}$. This is due to the low geodesic curvature of γ^{UNI}, which approximately

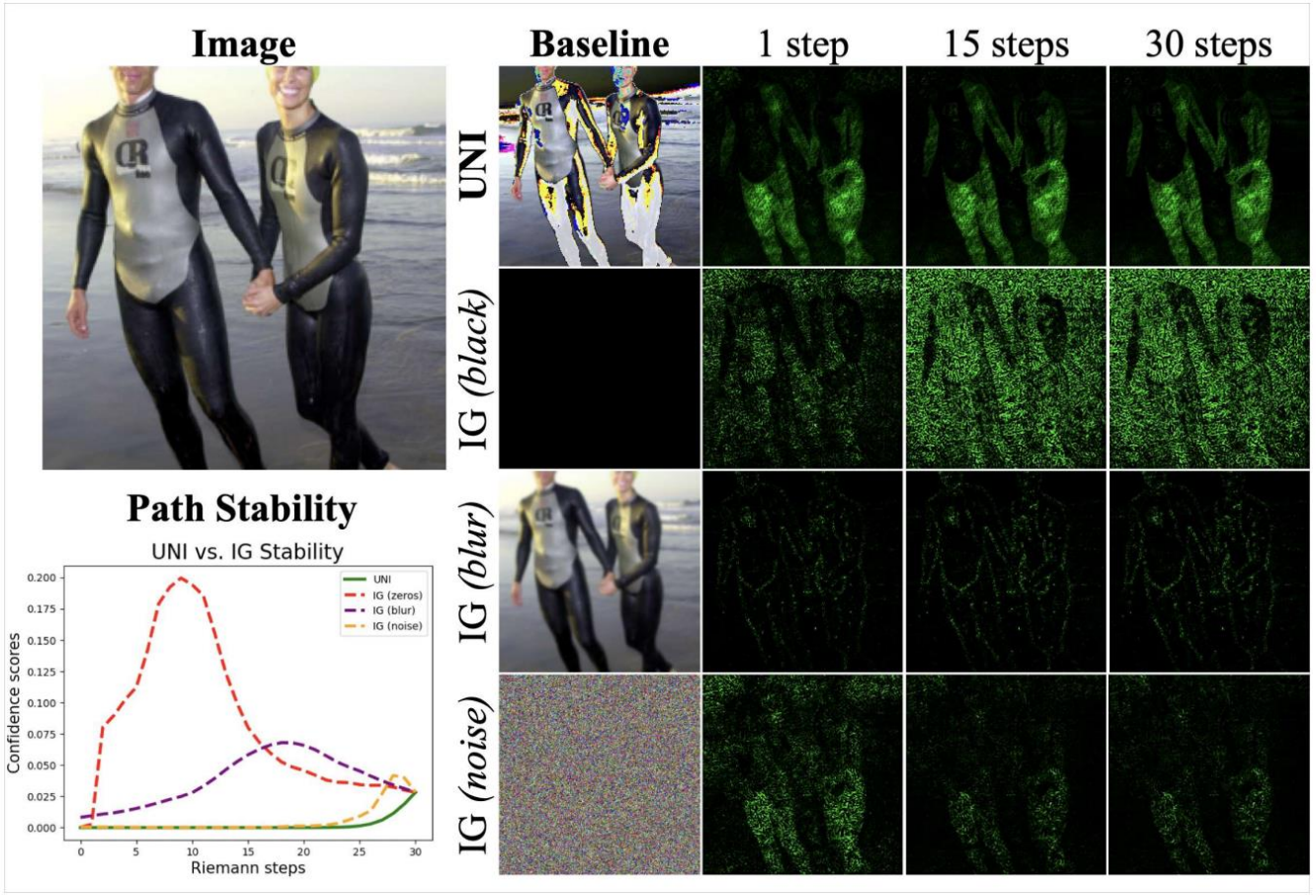


Figure 7: UNI path features monotonically increase in output confidence when interpolating from baseline to input. This eliminates instability and inconsistency problems caused by extrema and turning points along the Riemann approximation path, which is present in other methods.

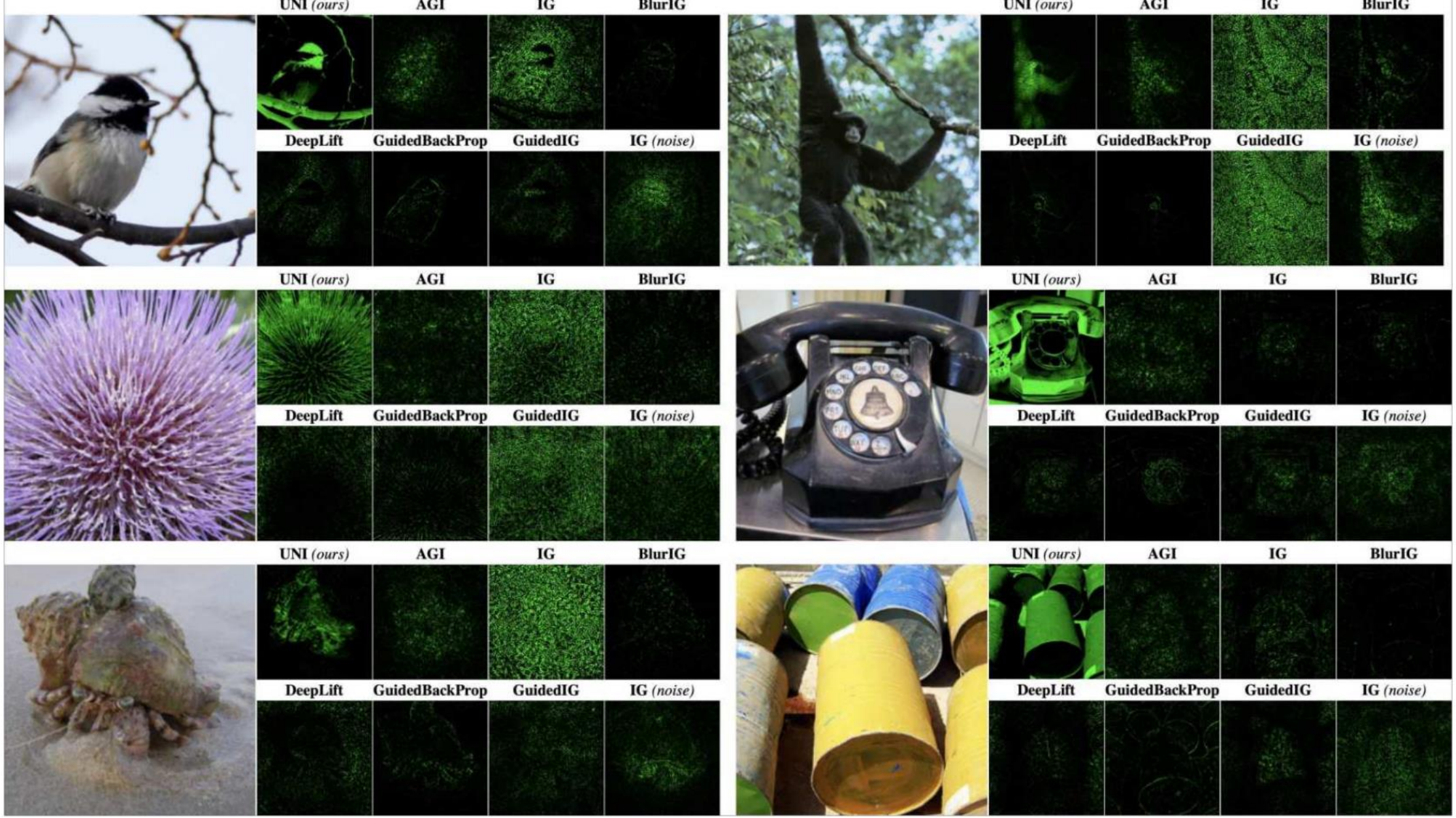


Figure 14: *Comparing attributions (ResNet-18):* `UNI` attributions demonstrate higher saliency, fidelity and faithfulness relative to conventional baselines on the ImageNet test set.

Appendix

[1] "why should I trust you? ": Explaining the predictions of any classifier(LIME). KDD 2016.

[2] Not just a black box: Learning important features through propagating activation differences. ArXiv 2016.

[3] Grad-Cam: Visual explanations from deep networks via gradient-based localization. ICCV 2017.

[4] Axiomatic attribution for deep networks(IG). ICML 2017.

[5] Techniques for Interpretable Machine Learning. CACM 2019.

[6] Interpretability research of deep learning: A literature survey. Information Fusion 2024.

[7] Unlearning-based Neural Interpretations(UNI). ICLR 2025.

THANKS