



NeurIPS 2021

35th Conference on Neural
Information Processing Systems

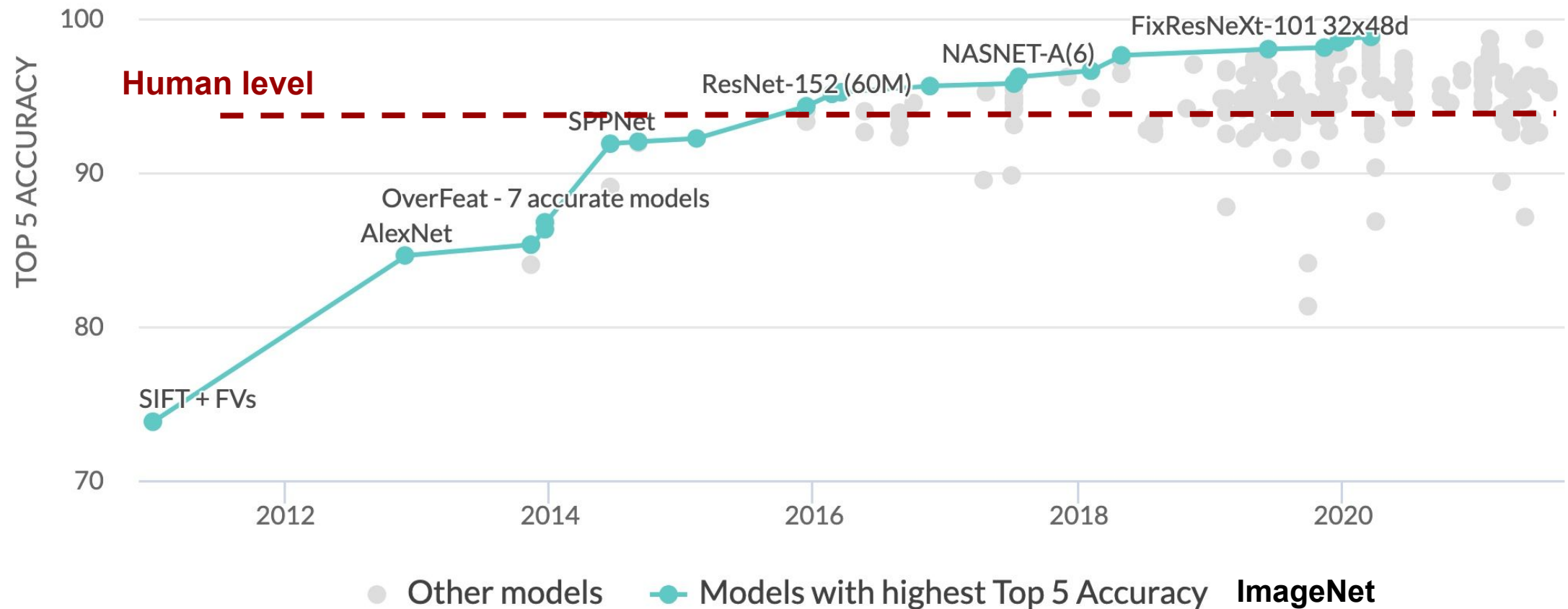
Online
December 6–14, 2021

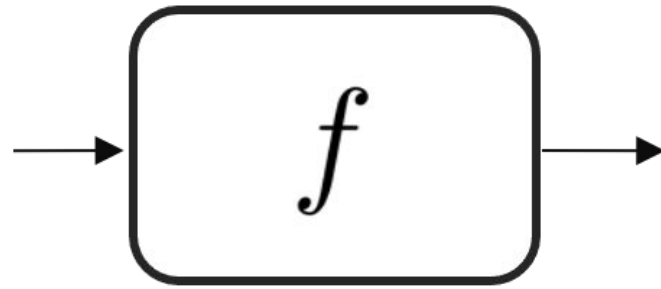
The Effectiveness of Feature Attribution Methods and its correlation with Automatic Evaluation Scores

Giang Nguyen[△], Daeyoung Kim[†], Anh Nguyen[△]

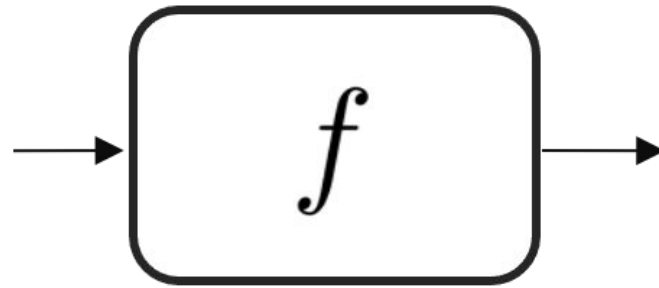


Artificial Intelligence surpassing humans on many tasks

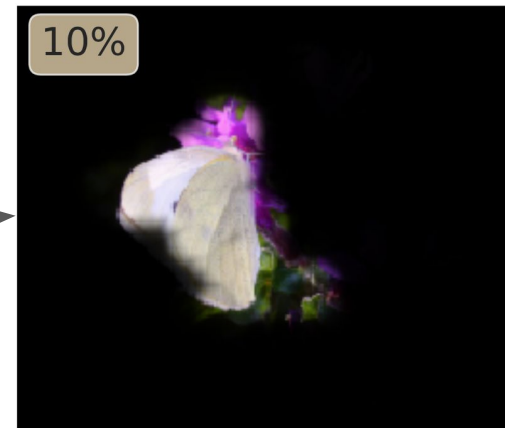
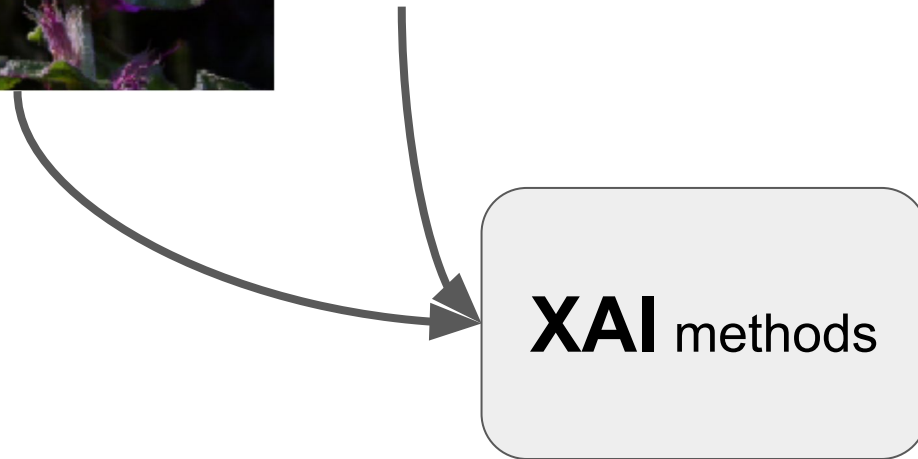




95% Cabbage butterfly



95% Cabbage butterfly

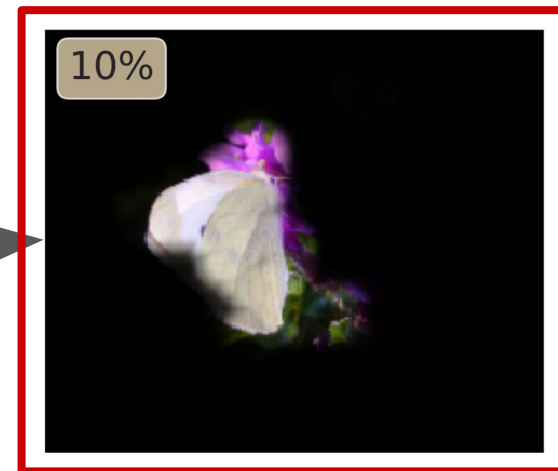


Explanation: Highlights of inputs that contributed to “Cabbage butterfly” prediction



95% Cabbage butterfly

Attribution map



Explanation: Highlights of inputs that contributed to “Cabbage butterfly” prediction

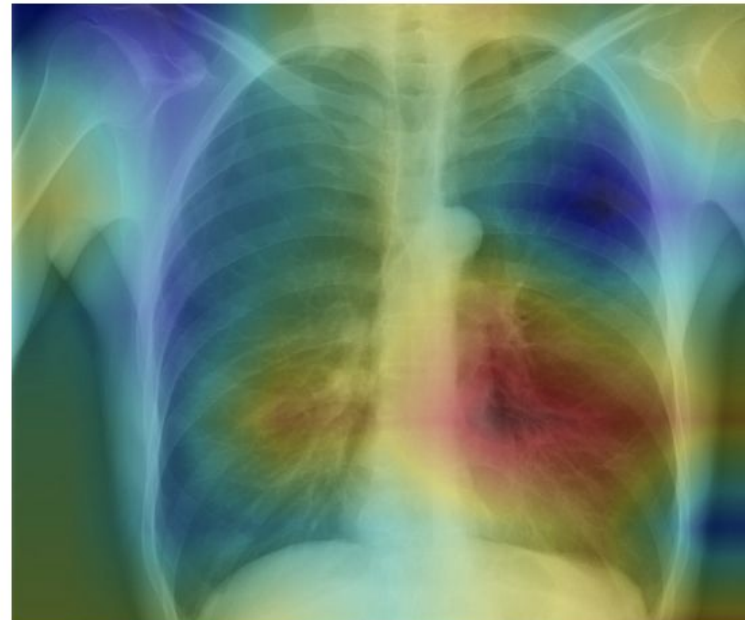
Feature attribution maps are useful in many tasks

Feature attribution maps has a wide array of applications ranging from localizing tumors to helping humans making correct decision in downstream tasks.

Teaching humans to classify



Localizing tumors



Highlighting important input features for humans to label text

[Click here to start the task](#)

This is a movie review, please select the appropriate sentiment below.

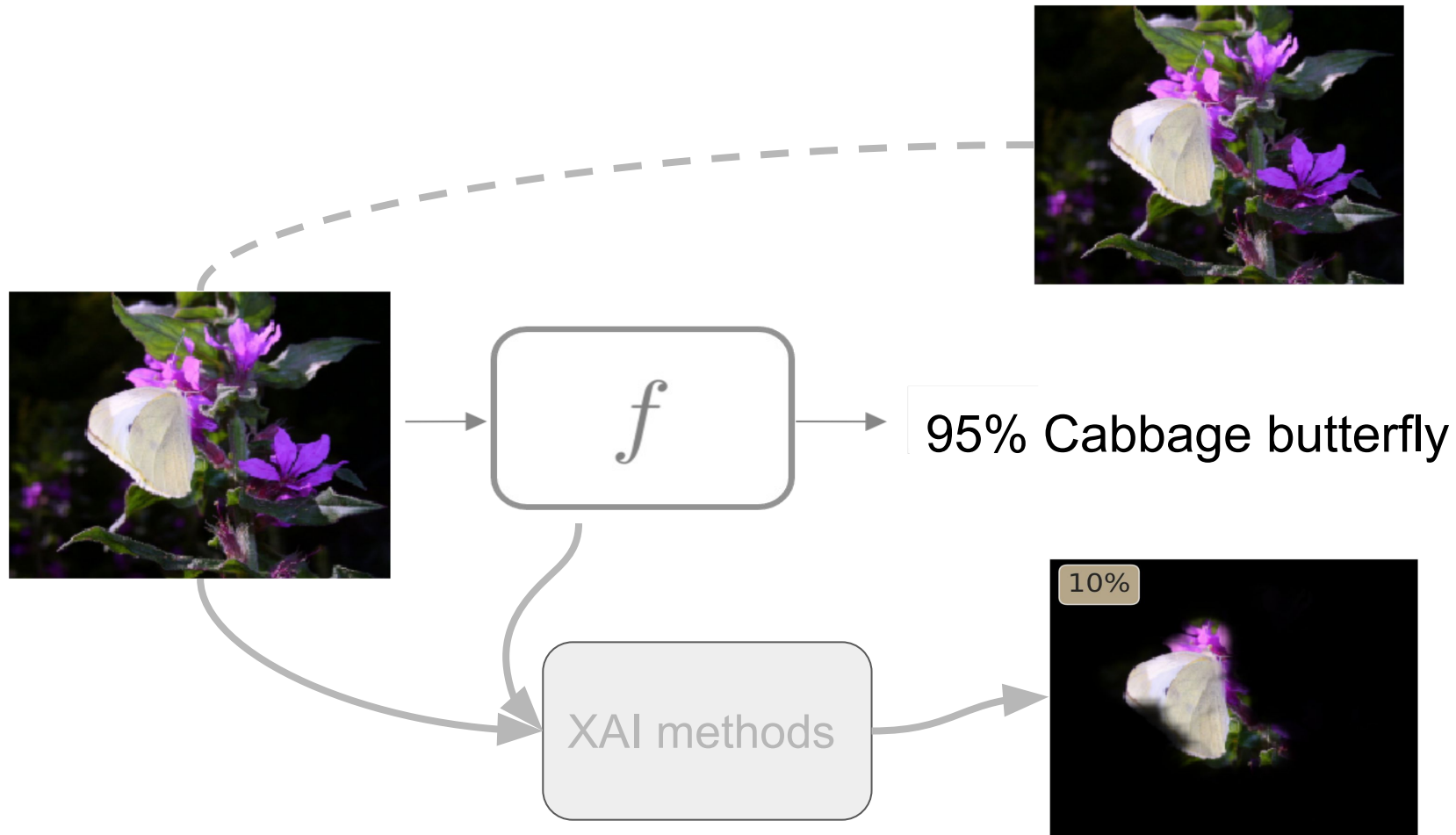
Just thinking about the movie, i laugh to myself. Anne Ramsey plays an **unforgettable** part as 'Momma,' probably the most nasty, yet hilarious matriarch ever captured on film. Danny Devito and Billy Crystal make a **fabulous** duo, bringing a **true** warmth to the film. Though not exceedingly complex, the cute story holds your attention, and keeps you laughing the whole way through. It's a fun comedy to lighten things up, and even will entertain the kids. I give it my full recommendation.

Sentiment of above movie review:

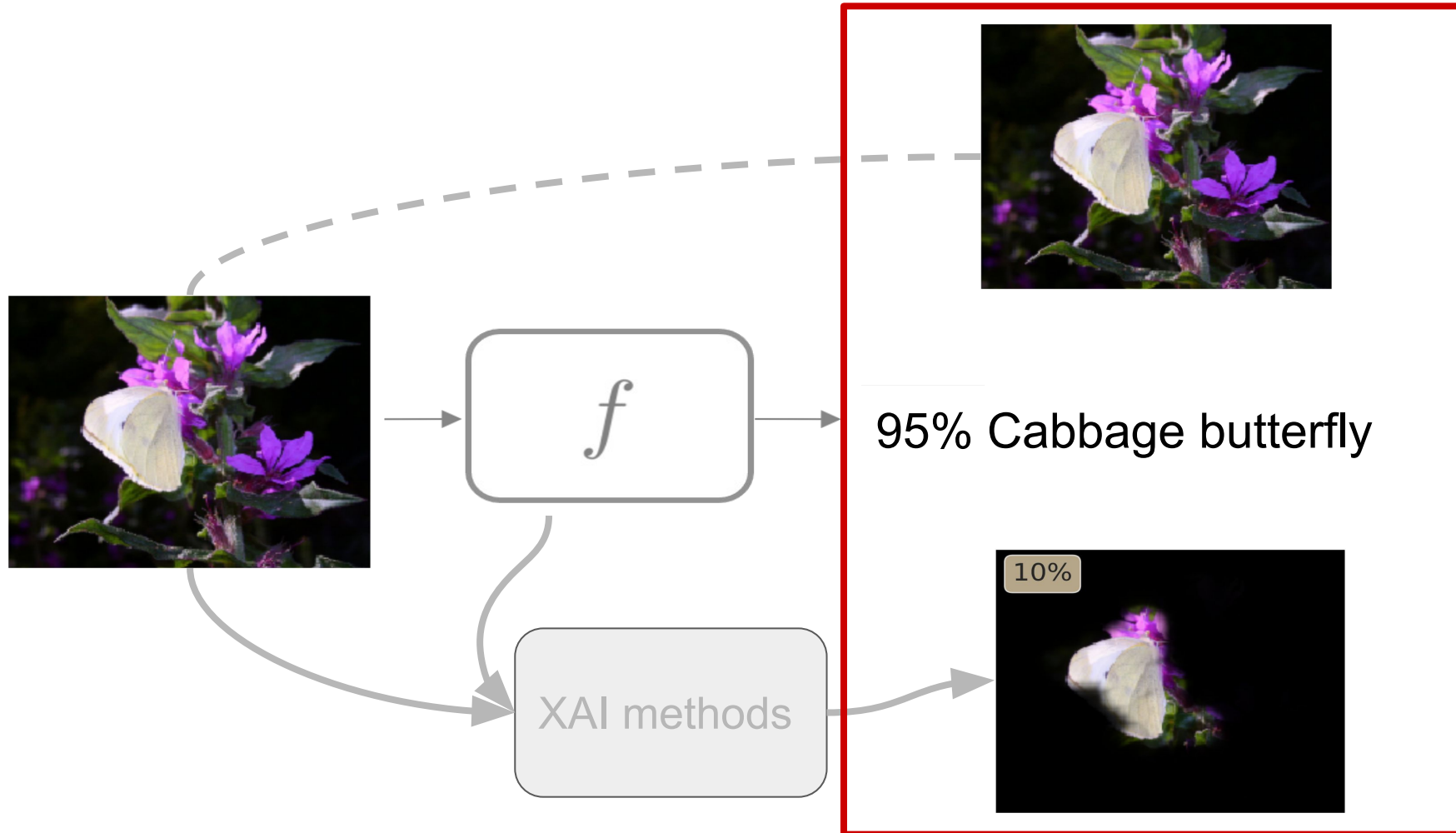
negative

positive

Human-AI team decision making



Human-AI team decision making



Is this
Cabbage butterfly?



Attribution maps effectiveness in human decision-making tasks

The gold standard for assessing the effectiveness of an explanation is a human-subject study [1].

Input	Tasks	Effectiveness
Text	Book categorization ^a	Yes
	Sentiment analysis ^{a,b}	Yes
	Deceptive review detection ^{a,b}	Yes
Tabular	Hypoxemia-risk detection ^a	Yes
Image	Age prediction ^a	No
	Model debugging ^{a,b}	Sometimes
	Image classification	Unknown

Motivation: Attribution methods were originally built to explain image classifiers (e.g. ResNet-50) pre-trained on **ImageNet**, but their effectiveness in human image classification has never been investigated on ImageNet.

Attribution maps effectiveness in human decision-making tasks

The gold standard for assessing the effectiveness of an explanation is a human-subject study [1].

Input	Tasks	Effectiveness
Text	Book categorization ^a	Yes
	Sentiment analysis ^{a,b}	Yes
	Deceptive review detection ^{a,b}	Yes

Q1: Are attribution maps useful for humans in image classification?

	Image classification	Unknown
--	----------------------	---------

Motivation: Attribution methods were originally built to explain image classifiers (e.g. ResNet-50) pre-trained on **ImageNet**, but their effectiveness in human image classification has never been investigated on ImageNet.

Attribution map evaluation using proxy metrics

Dozens of attribution methods have been tested on proxy benchmarks rather than humans:

- **Pointing Game** ^a :
 - Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, Selvaraju et al. 2016
 - RISE: Randomized Input Sampling for Explanation of Black-box Models, Petsiuk et al. 2018
 - Understanding Deep Networks via Extremal Perturbations and Smooth Masks, Fong et al. 2019
 - There and Back Again: Revisiting Backpropagation Saliency Methods, Rebuffi et al. 2019
 - Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks, Wang et al. 2019
- **Weakly-supervised Localization** ^a :
 - Visual Explanations from Deep Networks via Gradient-based Localization, Selvaraju et al. 2016
 - Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks, Chattopadhyay et al. 2017
 - XRAI: Better Attributions Through Regions, Kapishnikov et al. 2019
 - Explaining image classifiers by removing input features using generative models, Agarwal et al. 2020
- **Deletion/Insertion** ^a :
 - SAM: The sensitivity of attribution methods to hyperparameters, Bansal et al. 2020
 - A Benchmark for Interpretability Methods in Deep Neural Networks, Hooker et al. 2019
 - Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks, Wang et al. 2019
 - Towards Better Explanations of Class Activation Mapping, Jung et al. 2021
- **IoU**:
 - SCOUT: Self-aware Discriminant Counterfactual Explanations, Wang et al. 2020
 - Explaining AI-based Decision Support Systems using Concept Localization Maps, Lucieri et al. 2020

Motivation: It remains unknown if high performance on these proxy benchmarks correlate with high utility in helping human in image classification.

Attribution map evaluation using proxy metrics

Dozens of attribution methods have been tested on proxy benchmarks rather than humans:

- **Pointing Game** ^a :
 - Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, Selvaraju et al. 2016
 - RISE: Randomized Input Sampling for Explanation of Black-box Models, Petsiuk et al. 2018
 - Understanding Deep Networks via Extremal Perturbations and Smooth Masks, Fong et al. 2019
 - There and Back Again: Revisiting Backpropagation Saliency Methods, Rebuffi et al. 2019
 - Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks, Wang et al. 2019
- **Weakly-supervised Localization** ^a :
 - Visual Explanations from Deep Networks via Gradient-based Localization, Selvaraju et al. 2016
 - Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks, Chattopadhyay et al. 2017
 - XRA: Better Attributions Through Revisiting, Kerishkumar et al. 2019

Q2: Do evaluation metrics correlate with human accuracy?

- **IoU:**
 - SCOUT: Self-aware Discriminant Counterfactual Explanations, Wang et al. 2020
 - Explaining AI-based Decision Support Systems using Concept Localization Maps, Lucieri et al. 2020

Motivation: It remains unknown if high performance on these proxy benchmarks correlate with high utility in helping human in image classification.

User-study to assess attribution map effectiveness



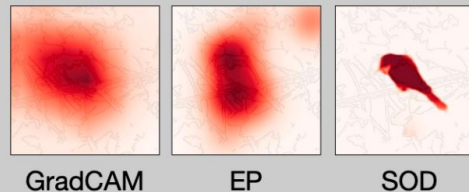
AI's top-1 predicted label: **lorikeet**



input

A confidence **20%**

B heatmaps



GradCAM EP SOD

C 3 nearest neighbors in **lorikeet**



lorikeet ?



user

Yes

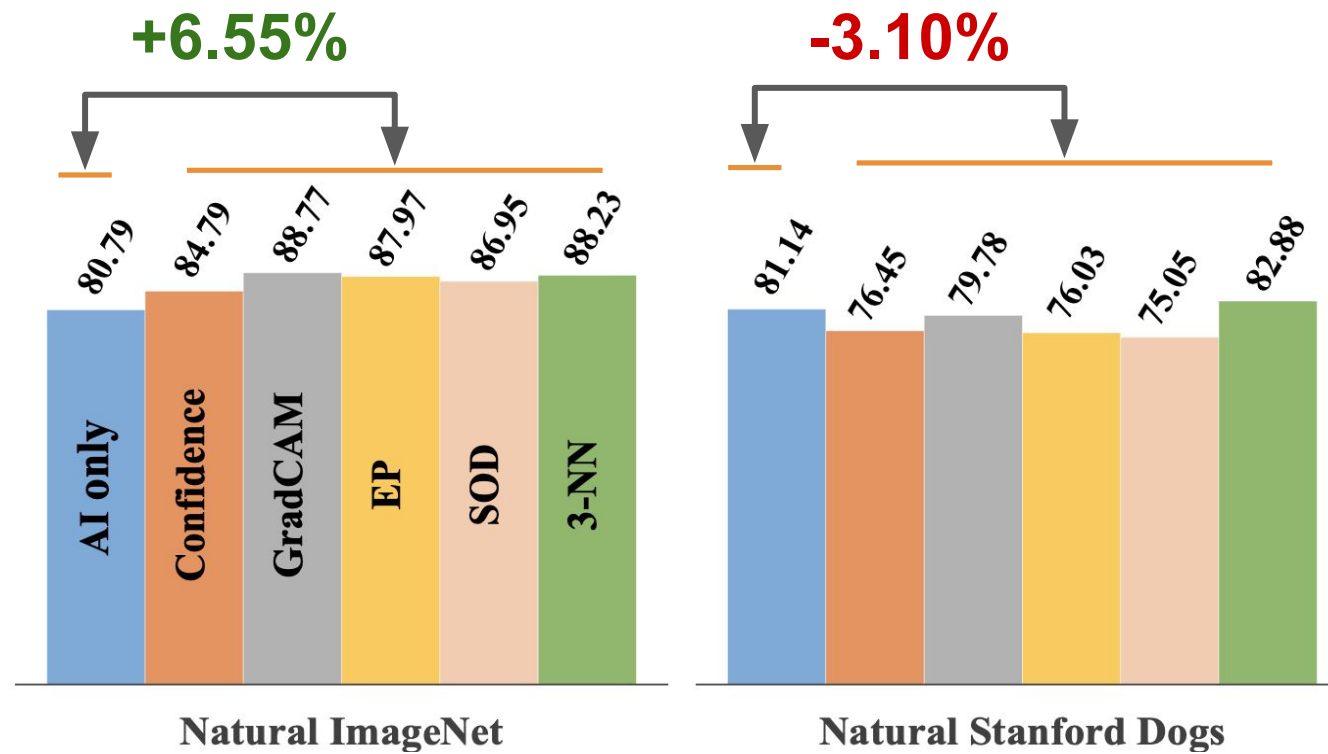
VS

No

groundtruth label: "bee eater"

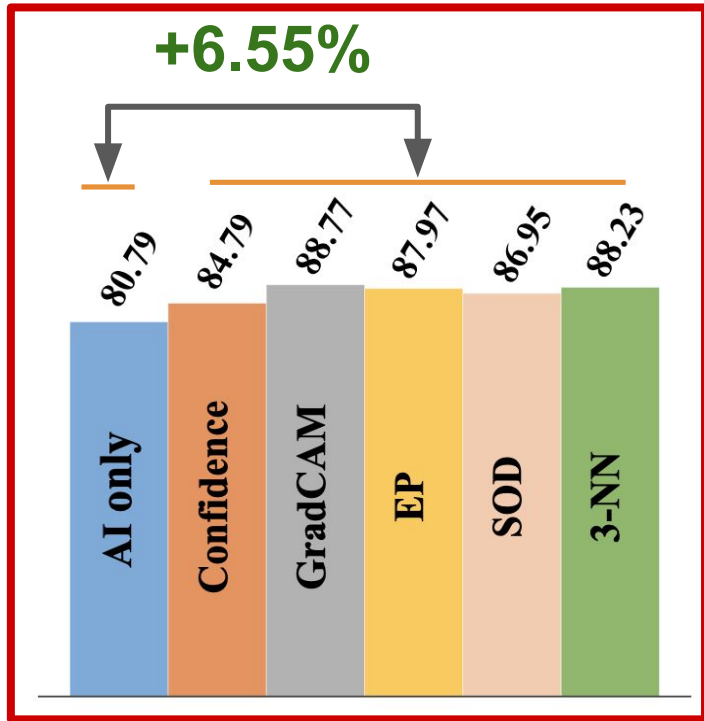
Conducted by **320** lay and **11** expert users

1. Human-AI teams outperform AI-only (*only when users have expertise*)



1. Human-AI teams outperform AI-only (*only when users have expertise*)

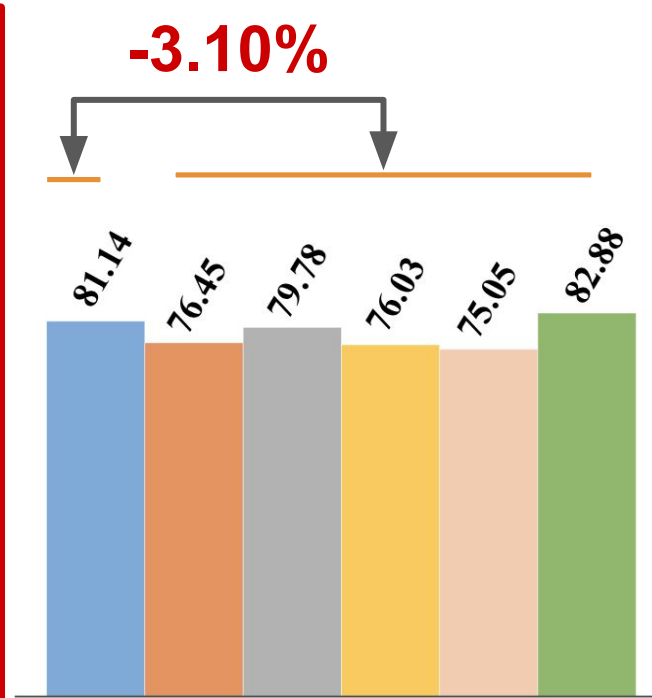
With expertise



Natural ImageNet

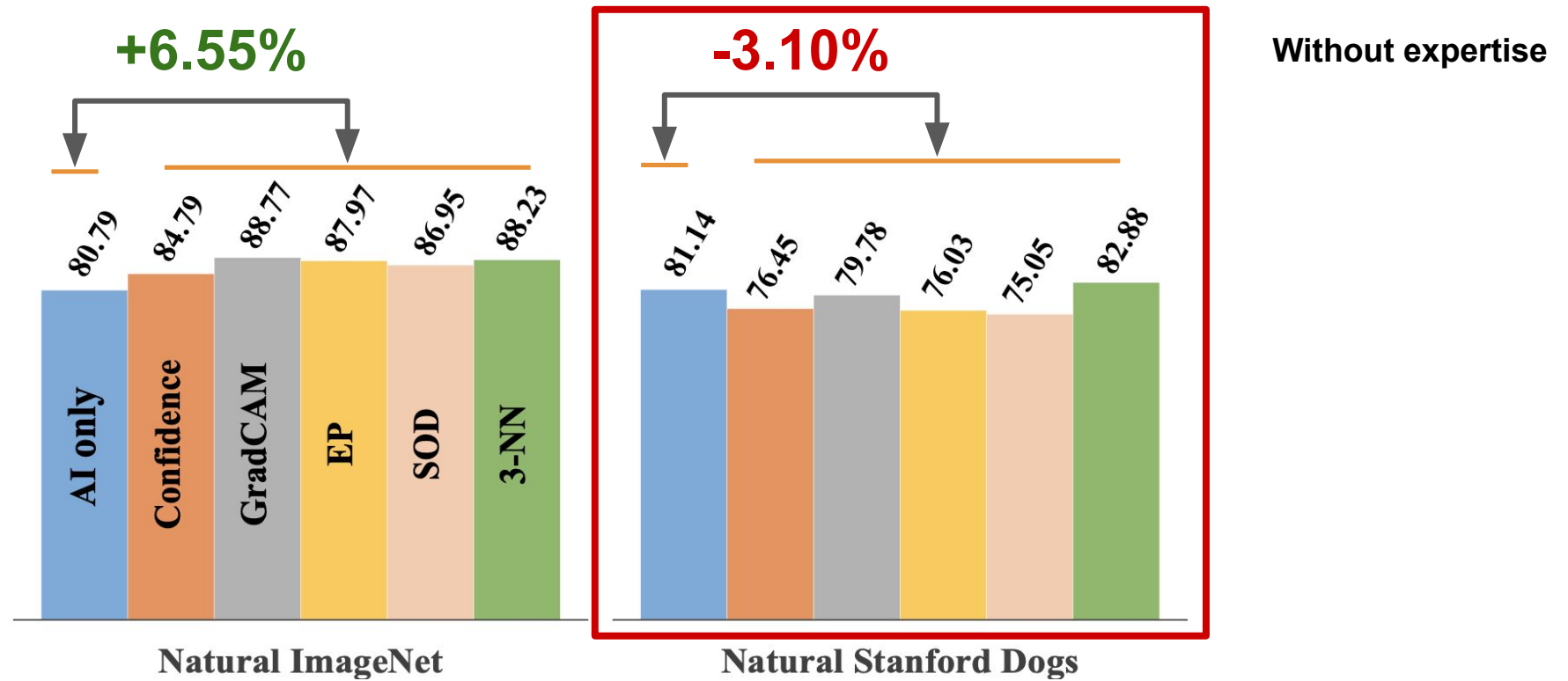


banana



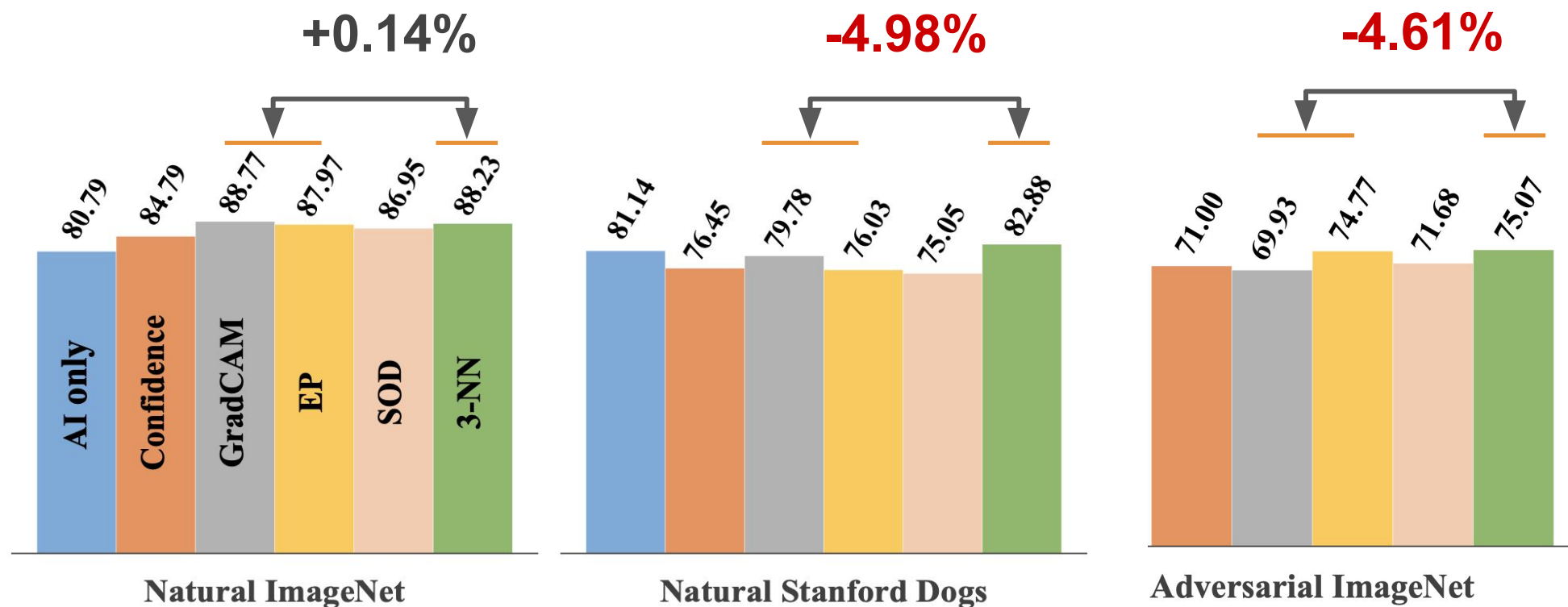
Natural Stanford Dogs

1. Human-AI teams outperform AI-only (only when users have expertise)



malamute

2. Feature attribution is **NOT** more effective than nearest-neighbors



bee eater

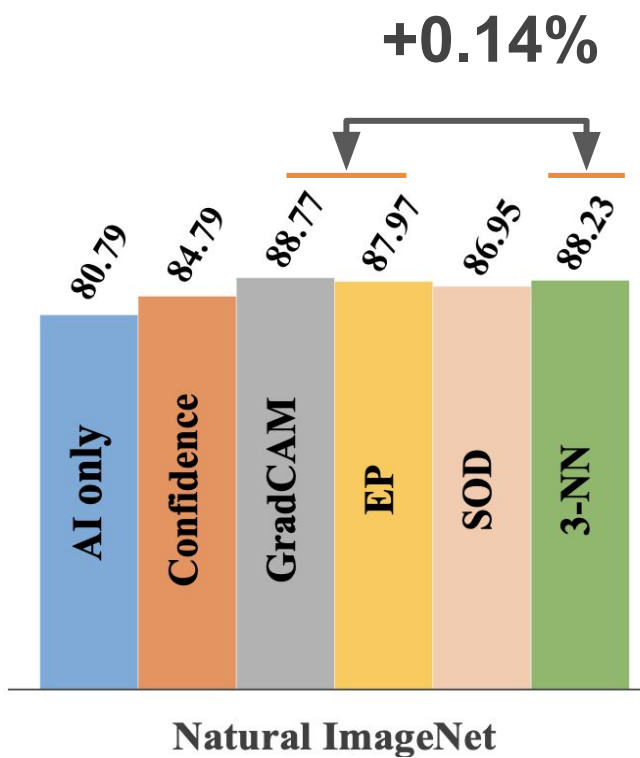


Bernese mountain dog

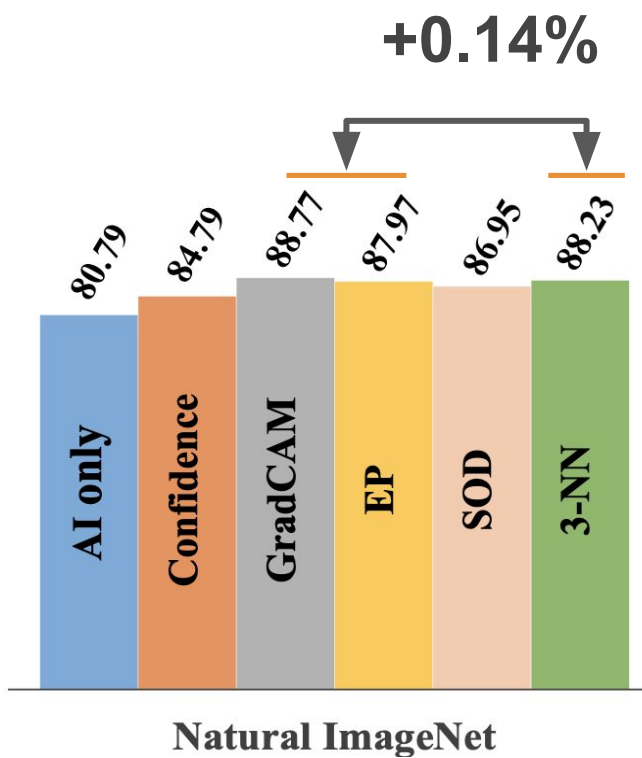


lorikeet

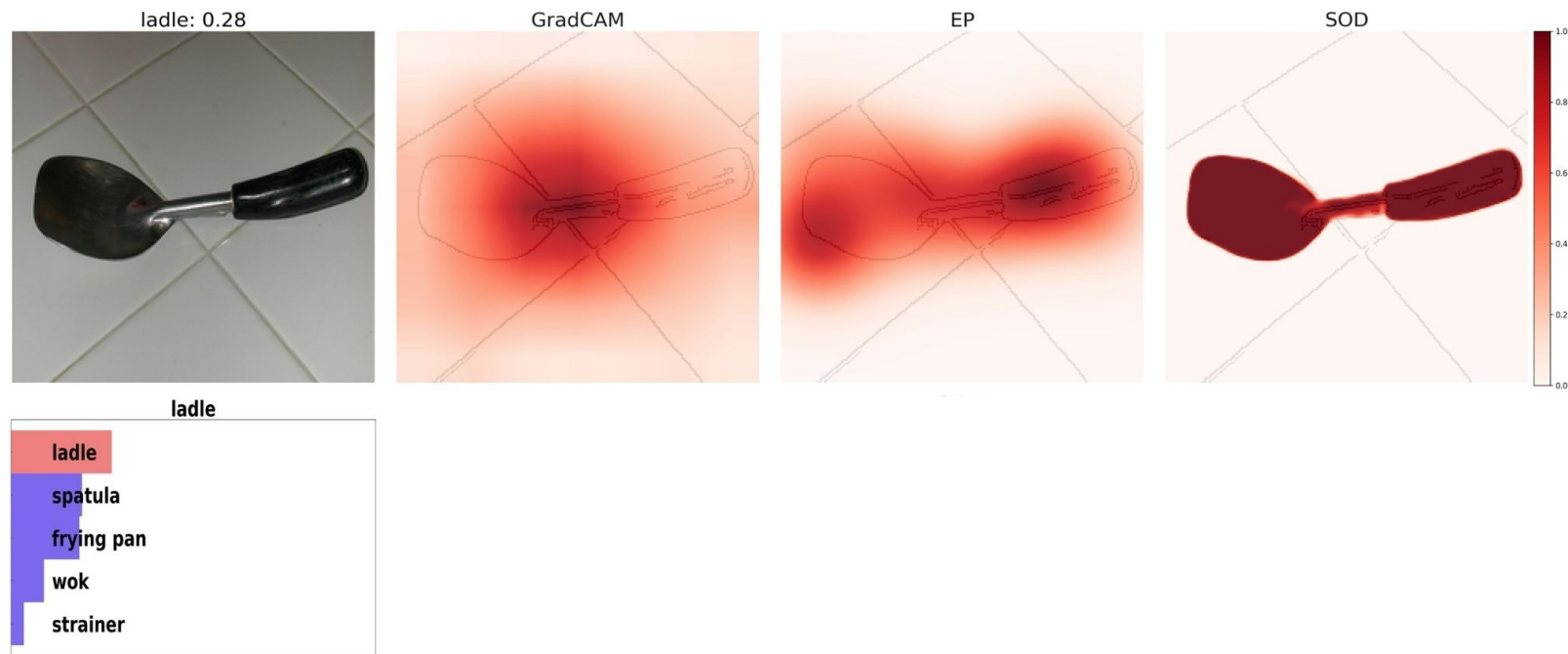
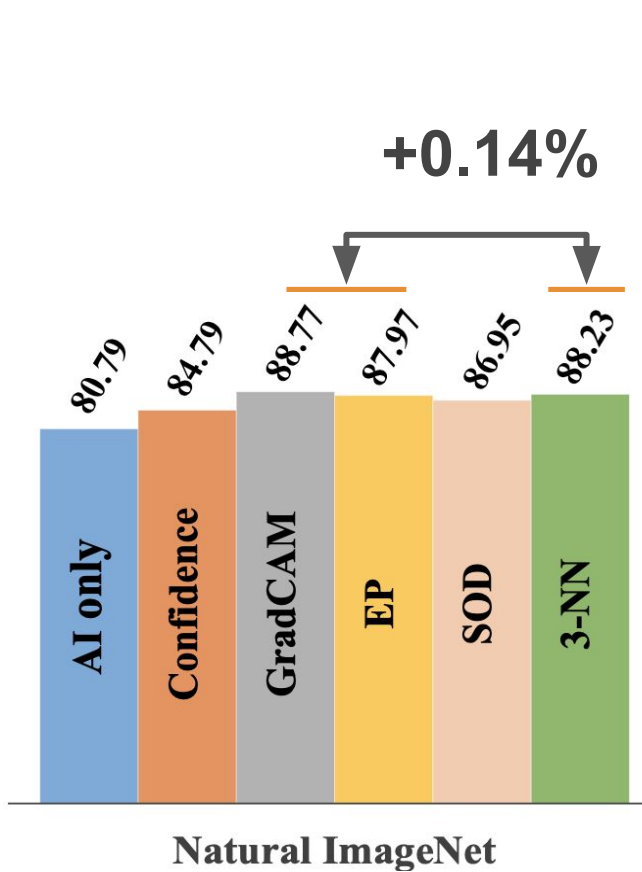
2. Feature attribution is **NOT** more effective than nearest-neighbors



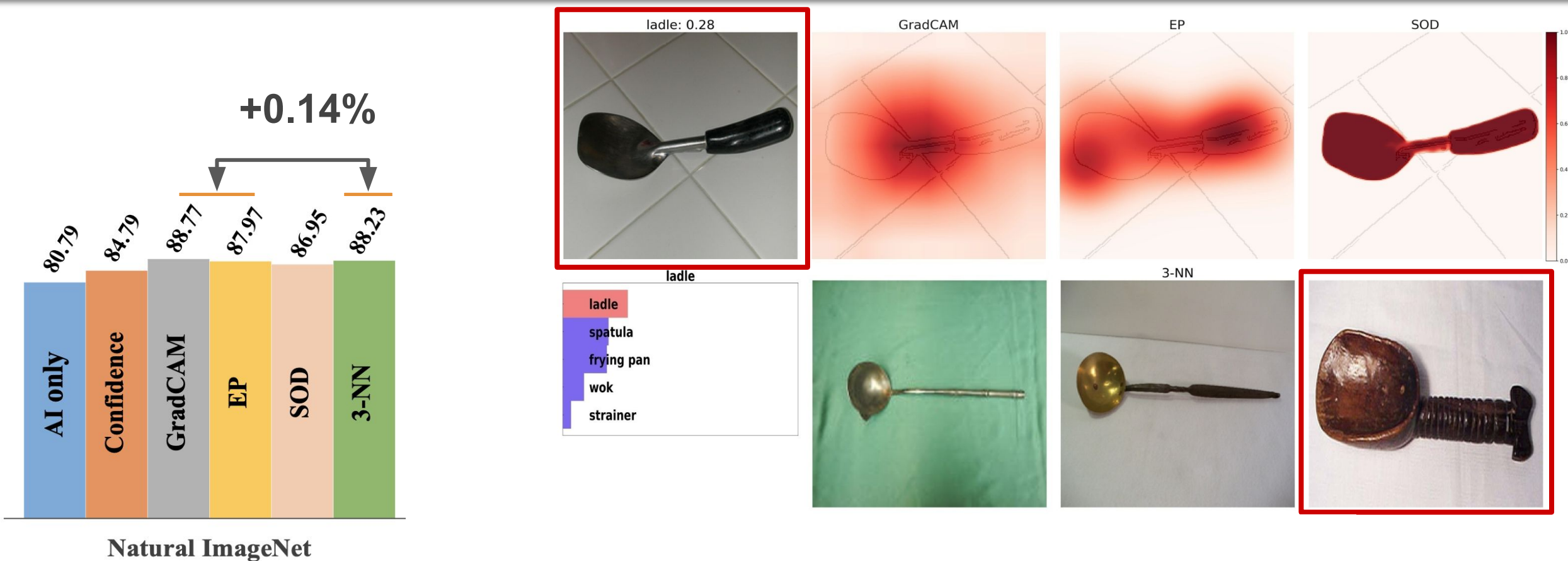
2. Feature attribution is **NOT** more effective than nearest-neighbors



2. Feature attribution is **NOT** more effective than nearest-neighbors

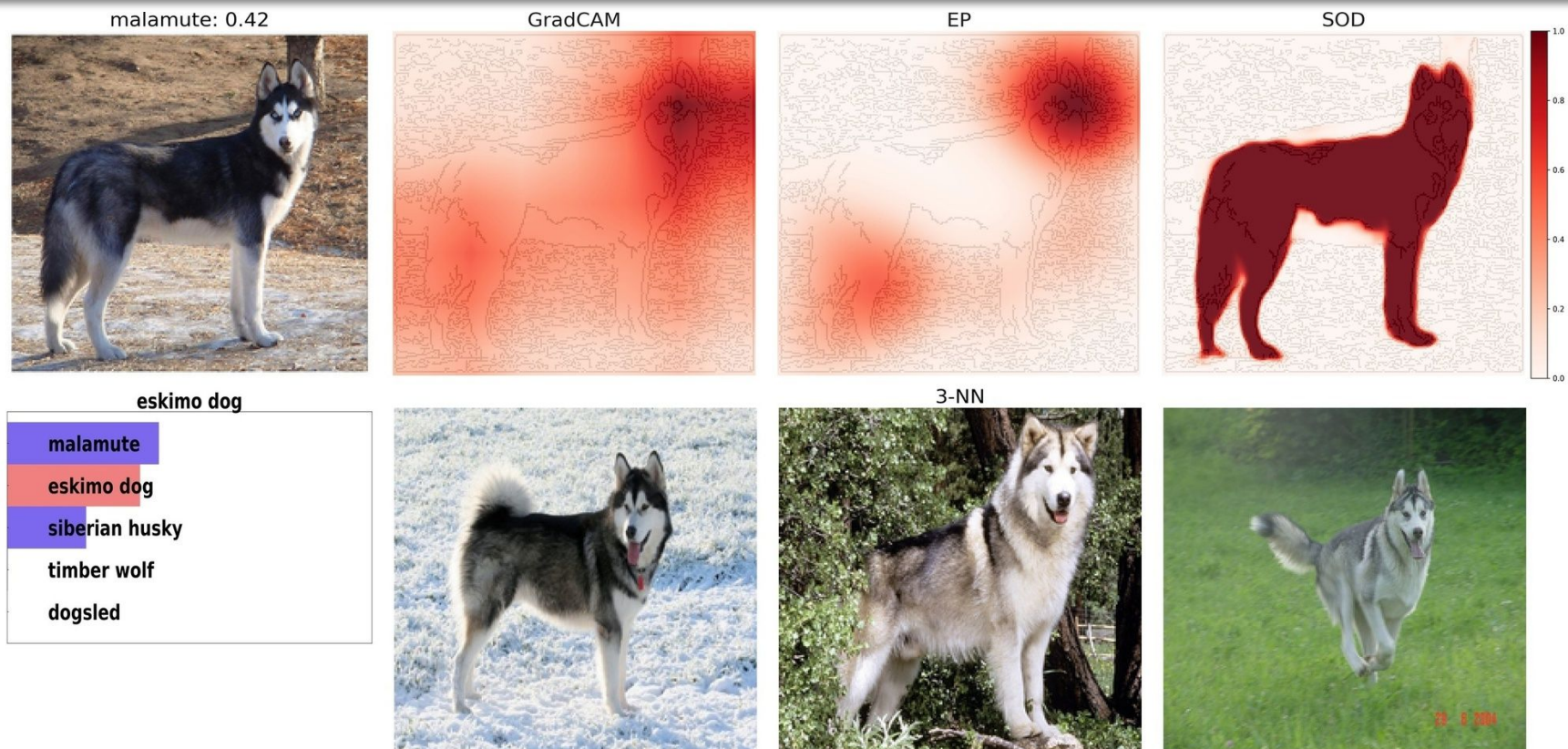
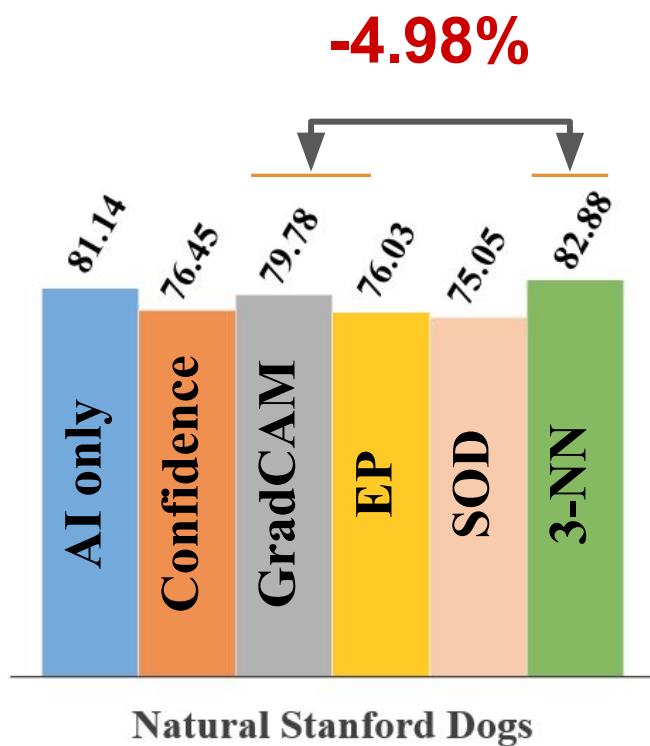


2. Feature attribution is **NOT** more effective than nearest-neighbors

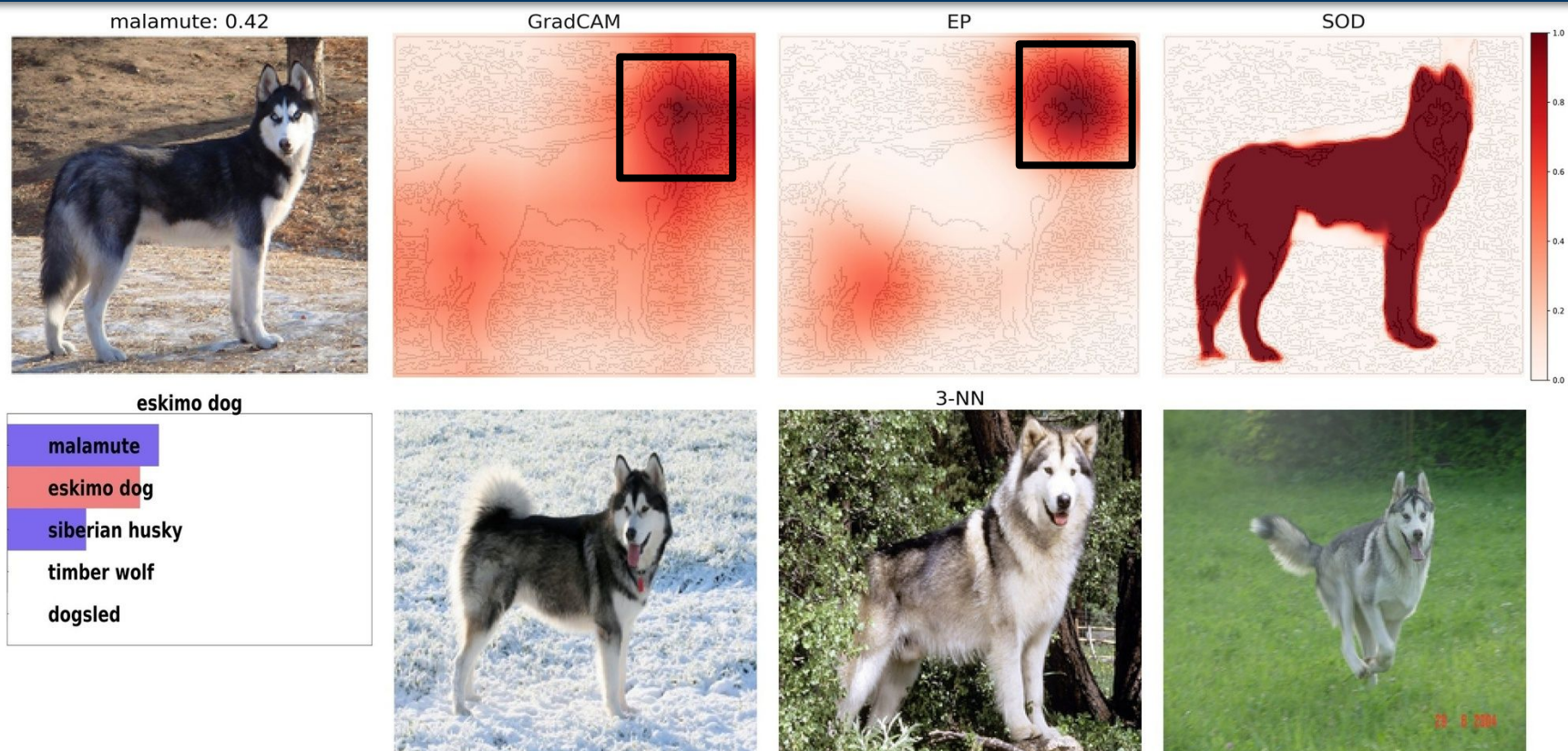
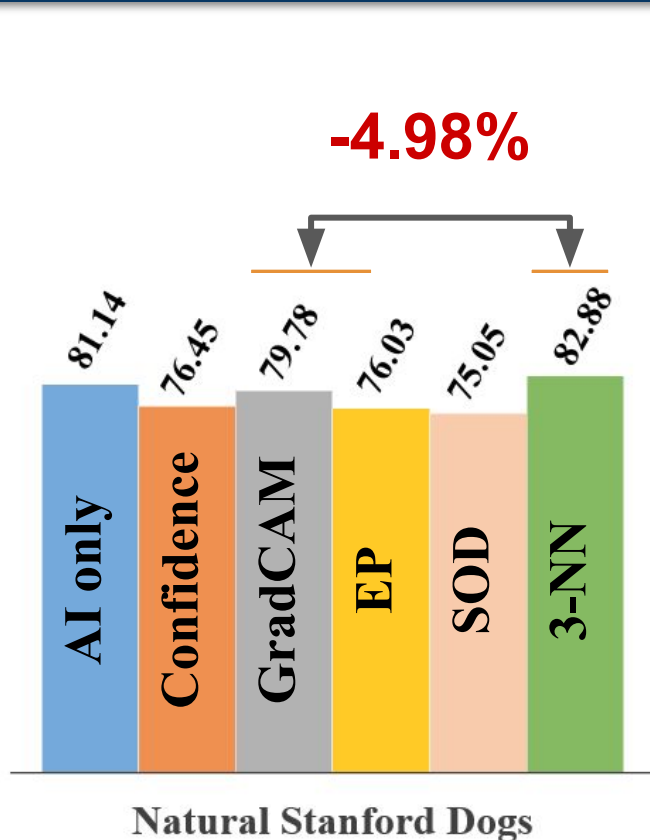


3-NN shows that “ladle” can sometimes have weird shape

2. Feature attribution is **NOT** more effective than nearest-neighbors

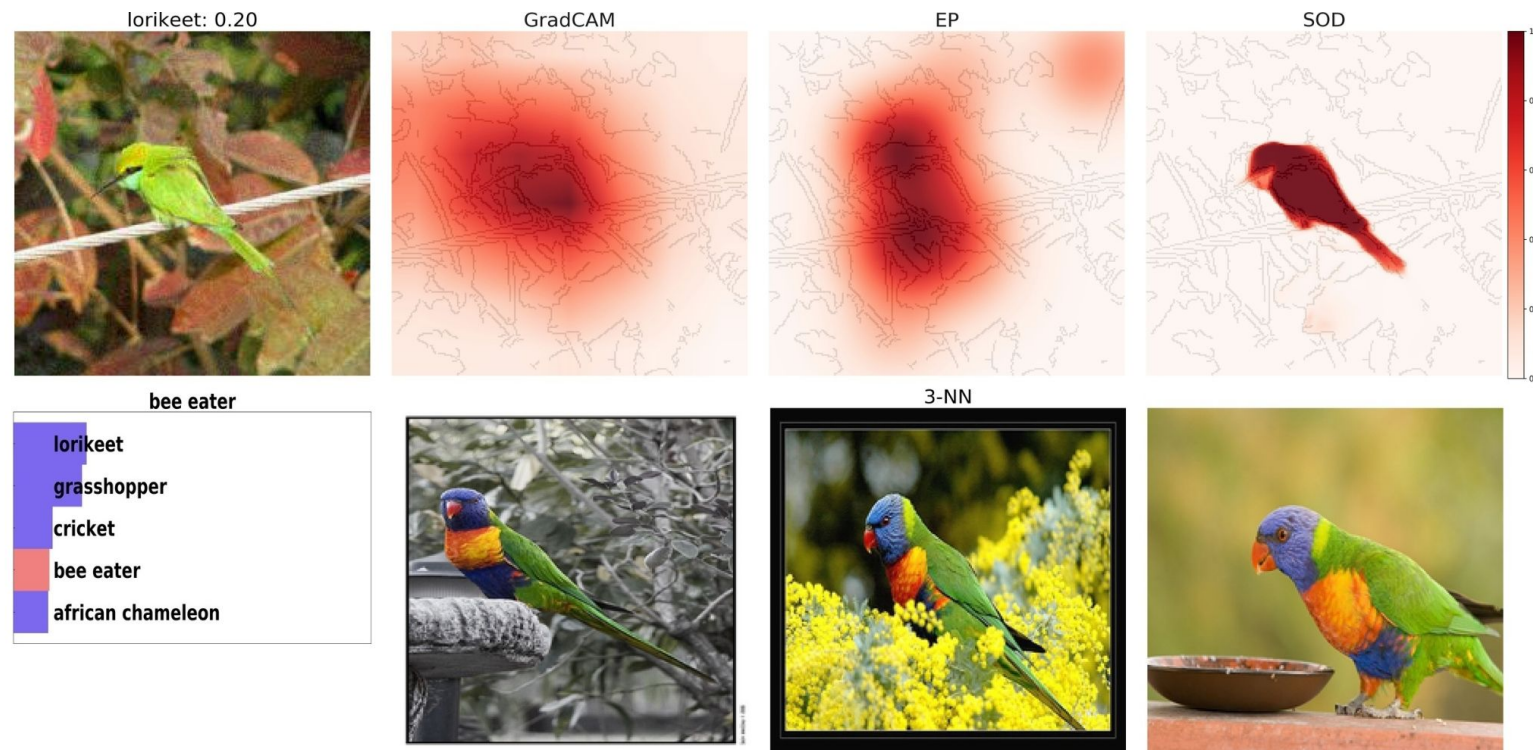
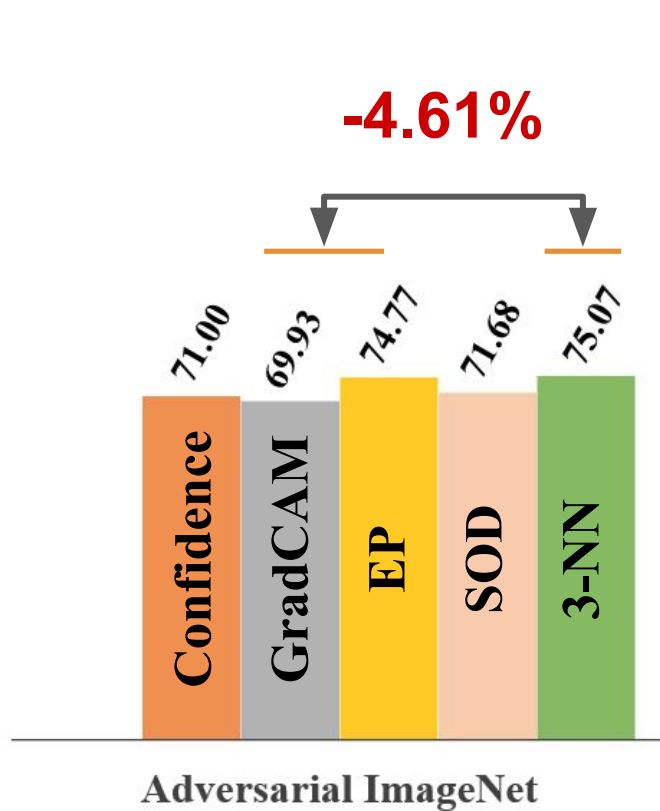


2. Feature attribution is **NOT** more effective than nearest-neighbors

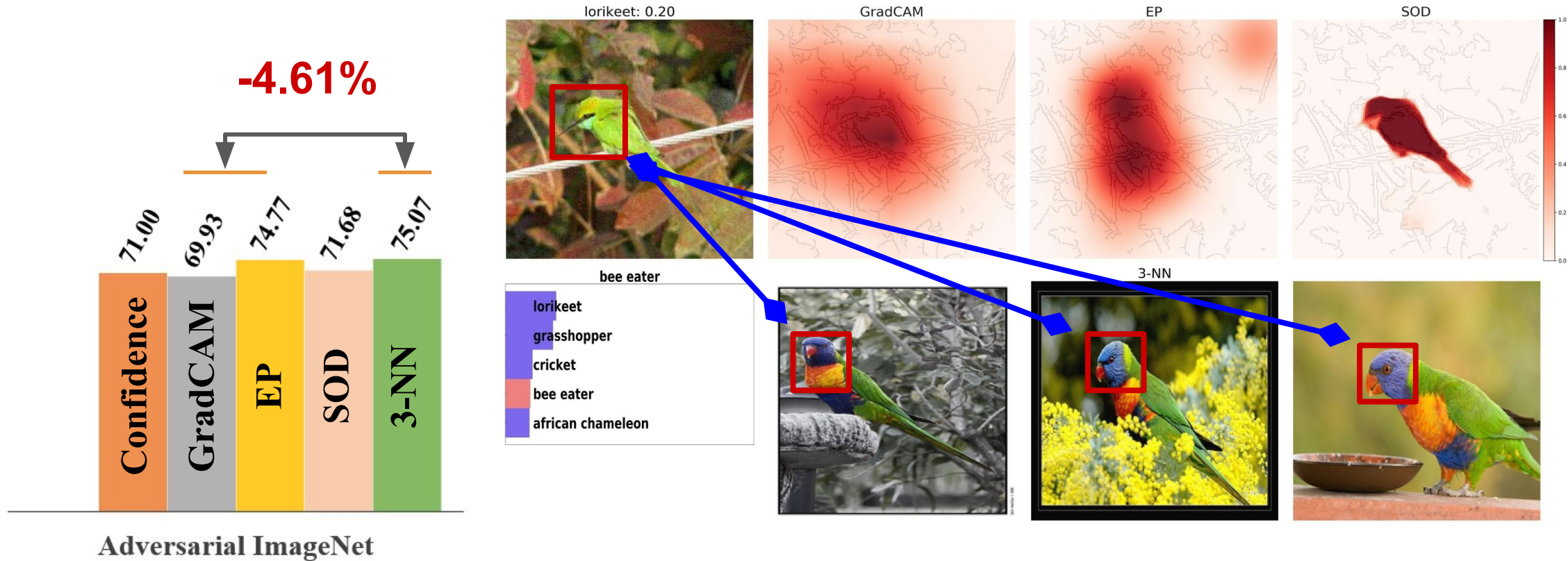


AMs can not show the difference between “malamute” vs. “eskimo dog” but generally highlight the face

2. Feature attribution is **NOT** more effective than nearest-neighbors

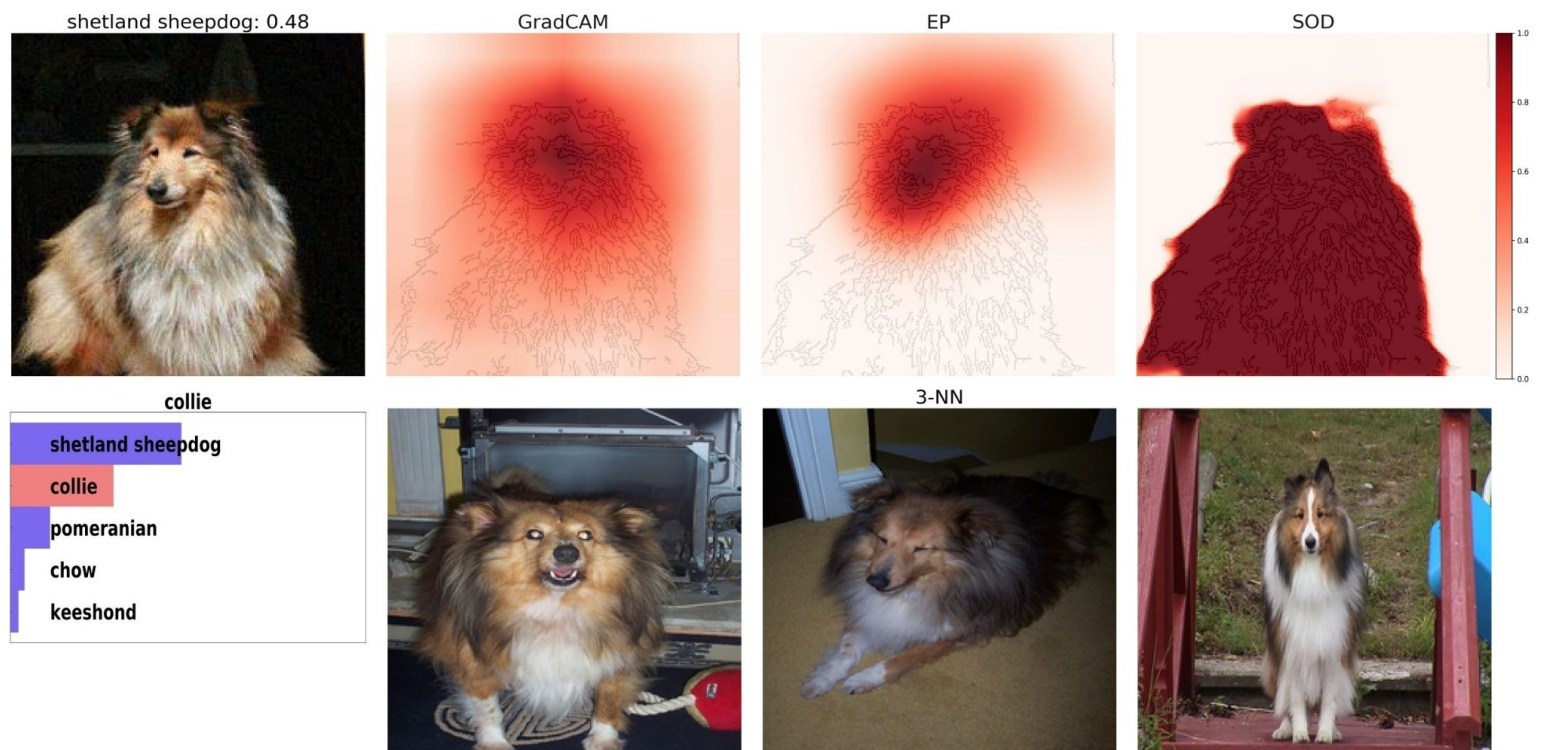
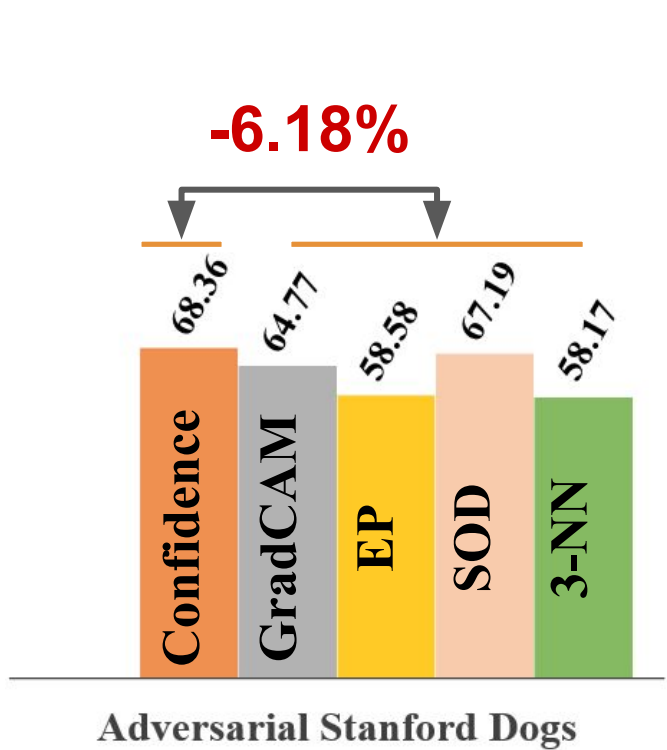


2. Feature attribution is **NOT** more effective than nearest-neighbors



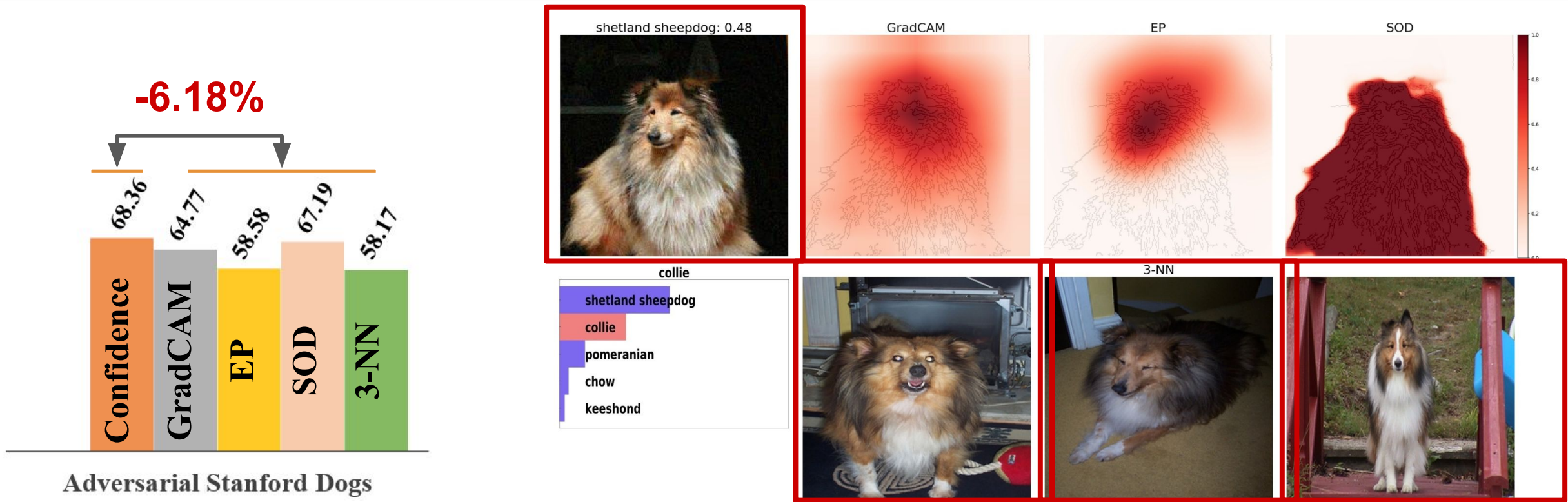
3-NN contrasts “lorikeet” and “bee eater” while AMs can not show the distinctive features

3. Explanations hurt human accuracy on fine-grained classification on OOD



When:
(a) Users do **NOT** have expertise, and
(b) Inputs are adversarial examples,
Using visual explanations worsens user accuracy

3. Explanations hurt human accuracy on fine-grained classification on OOD



The input image and 3 NNs are visually similar

4. On all real & adversarial ImageNet, 3-NN is better than attribution maps

* Mann-Whitney U test ($p < 0.035$)

Statistically significant *

Method	ImageNet	
	μ	σ
Confidence	72.44	8.25
GradCAM	72.58	8.11
EP	73.85	6.88
SOD	72.06	7.63
3-NN	76.08	5.86

Lay users



5. Expert users found 3-NN significantly more effective than GradCAM

* Mann-Whitney U test ($p < 0.035$)

Statistically significant *

Method	ImageNet	
	μ	σ
Confidence	72.44	8.25
GradCAM	72.58	8.11
EP	73.85	6.88
SOD	72.06	7.63
3-NN	76.08	5.86



Lay users

	Users	Avg. validation accuracy	Natural		Adversarial		μ	σ
			Accuracy	Trials	Accuracy	Trials		
GradCAM	5	9.80/10	67.31	70/104	69.57	32/46	68.00	8.69
3-NN	6	9.83/10	78.45	91/116	73.44	47/64	76.67	2.98



Expert users

5. Expert users found 3-NN significantly more effective than GradCAM

* Mann-Whitney U test ($p < 0.035$)

Statistically significant *

Method	ImageNet	
	μ	σ
Confidence	72.44	8.25
GradCAM	72.58	8.11
EP	73.85	6.88
SOD	72.06	7.63
3-NN	76.08	5.86



320 Lay users

	Users	Avg. validation accuracy	Natural		Adversarial		μ	σ
			Accuracy	Trials	Accuracy	Trials		
GradCAM	5	9.80/10	67.31	70/104	69.57	32/46	68.00	8.69
3-NN	6	9.83/10	78.45	91/116	73.44	47/64	76.67	2.98



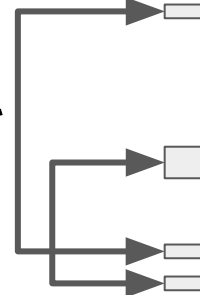
11 Expert users

5. Expert users found 3-NN significantly more effective than GradCAM

* Mann-Whitney U test ($p < 0.035$)

Statistically significant *

Method	ImageNet	
	μ	σ
Confidence	72.44	8.25
GradCAM	72.58	8.11
EP	73.85	6.88
SOD	72.06	7.63
3-NN	76.08	5.86



320 Lay users

Q2: Do evaluation metrics correlate with human accuracy?

11 Expert users

Attribution map evaluation using proxy metrics

Dozens of attribution methods have been tested on proxy benchmarks rather than humans:

- Pointing Game^a : Selvaraju et al. 2016, Petsiuk et al. 2018, Fong et al. 2019, Rebuffi et al. 2019, Wang et al. 2019
- Weakly-supervised Localization^a : Selvaraju et al. 2016, Chattopadhyay et al. 2017, Kapishnikov et al. 2019, Agarwal et al. 2020
- Deletion/Insertion^a: Bansal et al. 2020, Hooker et al. 2019, Wang et al. 2019, Jung et al. 2021, Zhang et al, 2021, Pan et al. 2021
- IoU: Jung et al. 2021, Wang et al. 2020, Lucieri et al. 2020



Pointing game score:

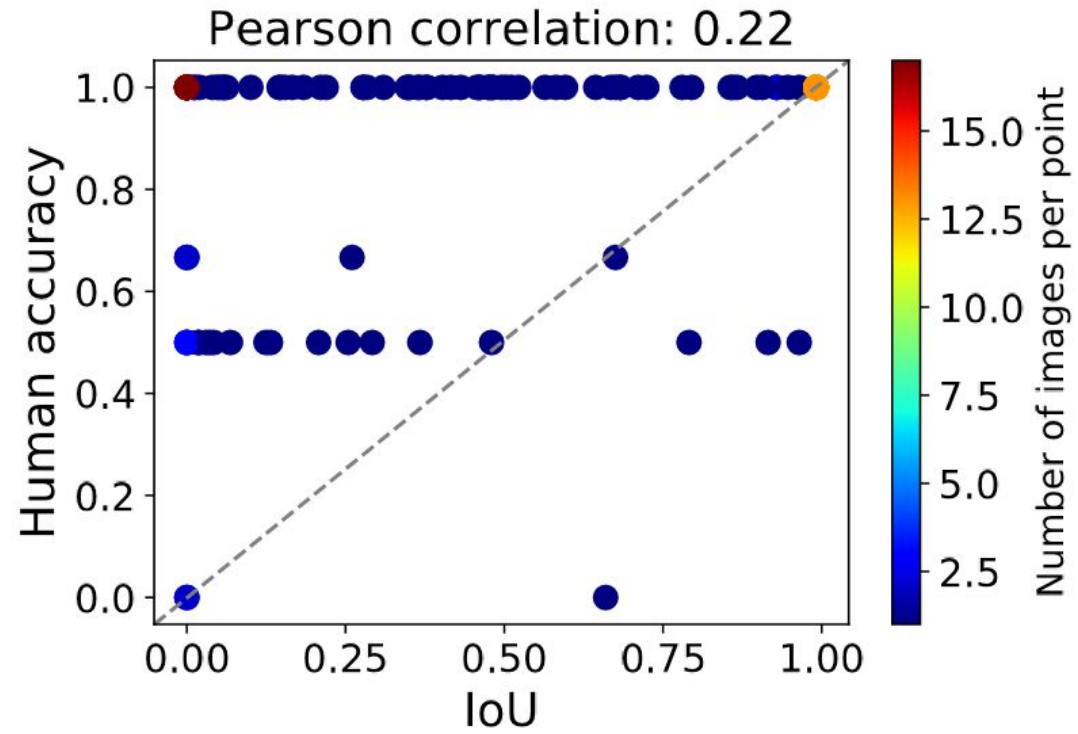
A hit is counted if the maximum point lies on one of the annotated instances of the cued object category, otherwise a miss is counted.

Localization error:

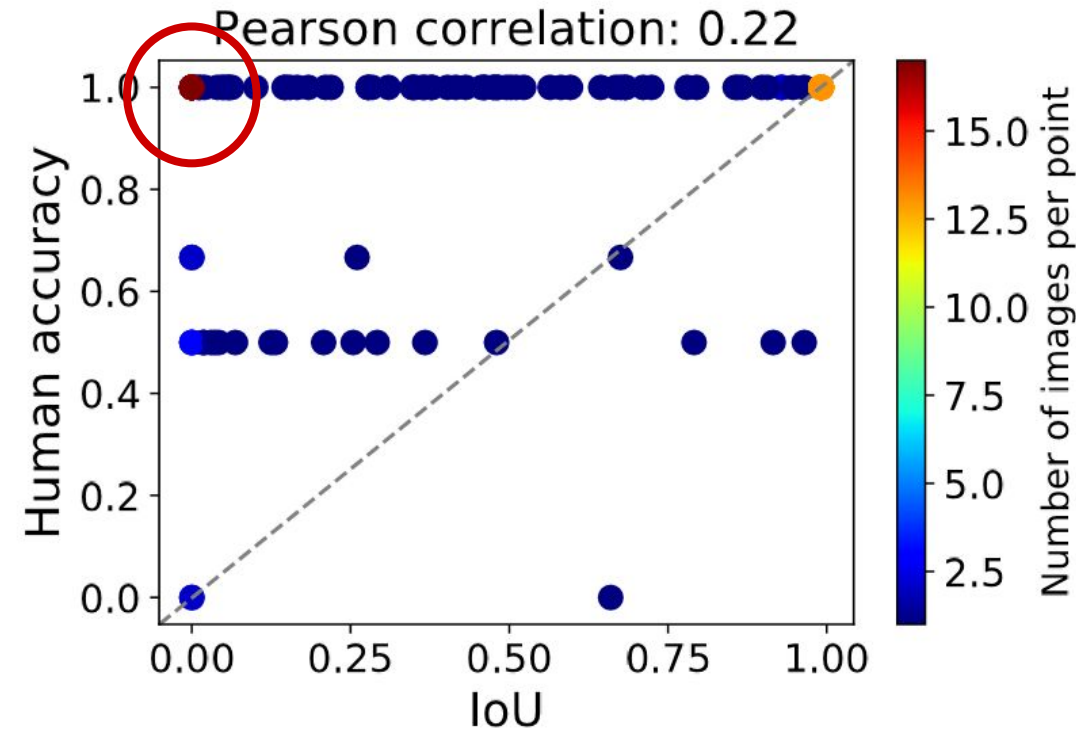
A hit is counted if the IoU value of the binarized mask vs. the ground-truth bounding box > 0.5 , otherwise a miss is counted.

Increasing importance

6. Proxy metrics correlate poorly vs. human accuracy

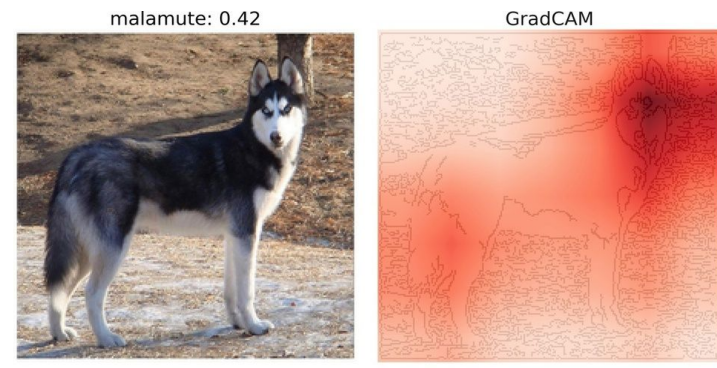
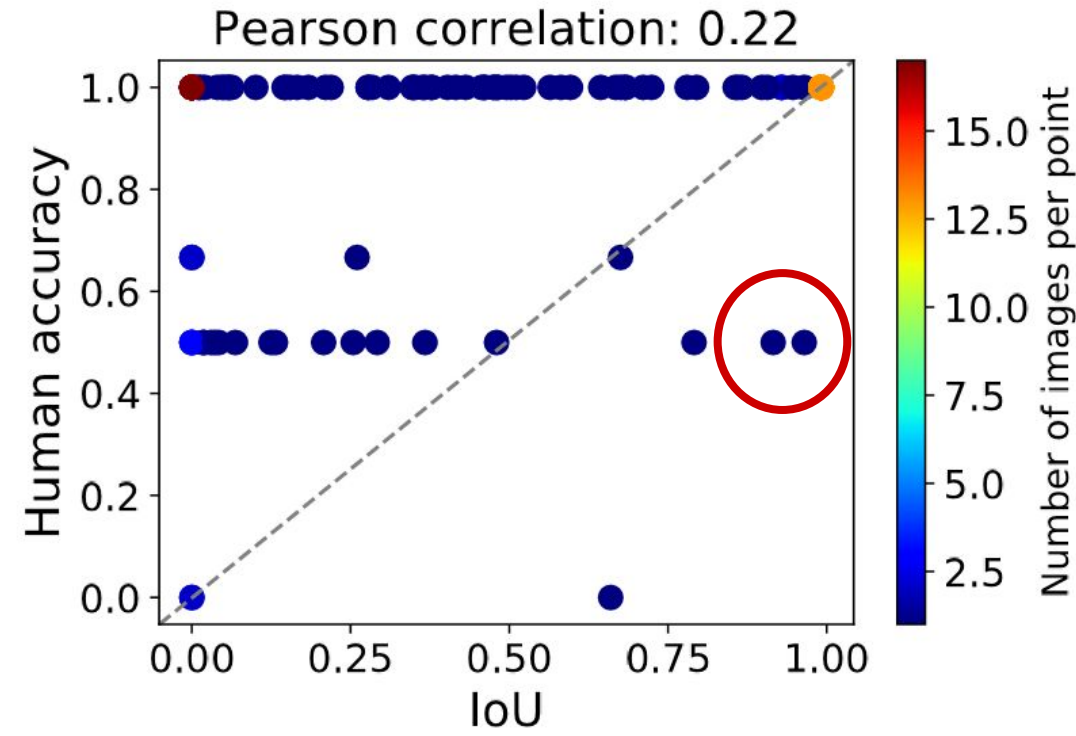


6. Proxy metrics correlate poorly vs. human accuracy



Humans can still make a lot of correct decisions when AMs localize badly

6. Proxy metrics correlate poorly vs. human accuracy



And humans still make wrong decisions when AM localize perfectly

Conclusions

Project page: <http://anhnguyen.me/project/feature-attribution-effectiveness/>



Giang Nguyen Daeyoung Kim Anh Nguyen

1. On real ImageNet data, 3-NN is more useful than activation maps
2. On fine-grained, out-of-distribution tests (e.g. Adversarial Dogs), *all visual explanations* hurt human performance
3. Existing attribution evaluation metrics (Object Localization, Pointing Game) do not strongly correlate with human accuracy

AI's top-1 predicted label: **lorikeet**

A confidence **20%**

B heatmaps

GradCAM EP SOD

C 3 nearest neighbors in **lorikeet**

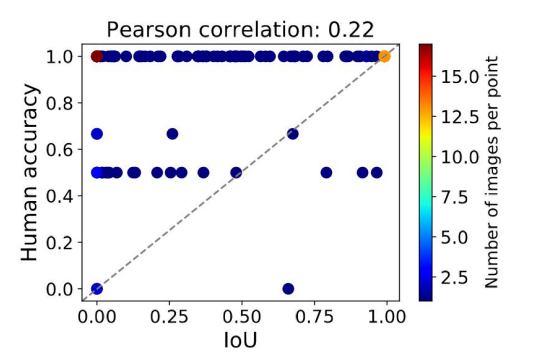
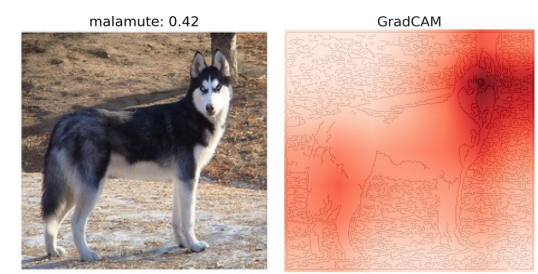
input

lorikeet ?

Yes
vs
No

user

groundtruth label: "bee eater"



Work funded by

