Robust Models Are More Interpretable Because Attributions Look Normal

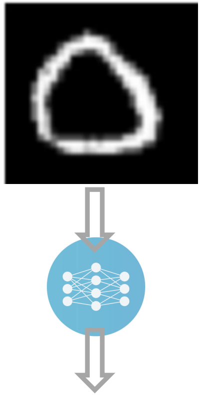
Zifan Wang, Matt Fredrikson, Anupam Datta
Carnegie Mellon University

zifan@cmu.edu

Explanations and Robustness

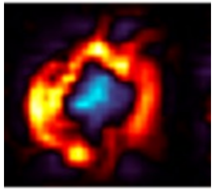
Gradient-based Explanations

(Saliency Maps; Feature Attribution)



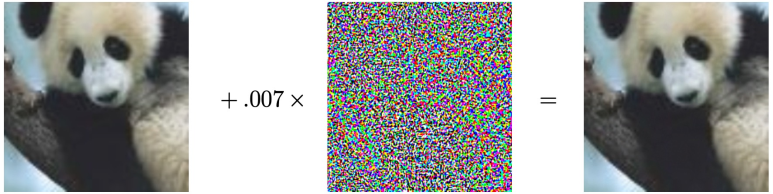
0

What are most important features for the prediction ?



Explanation for "0"

Adversarial Robustness

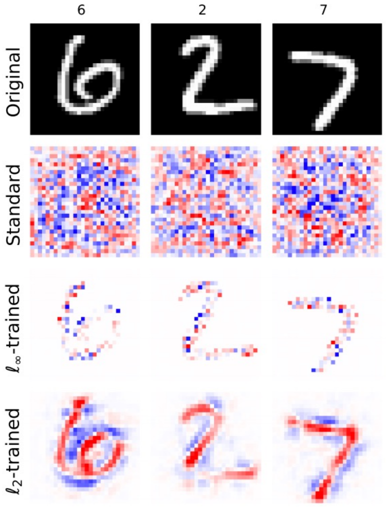


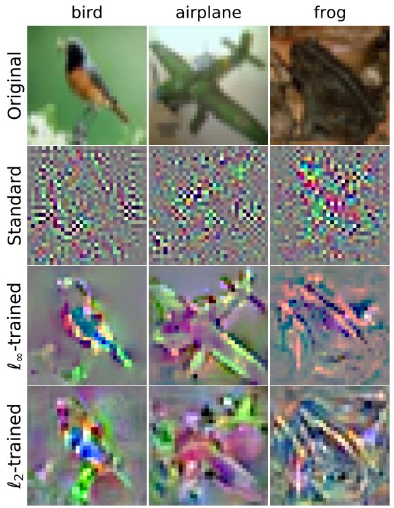
"panda"

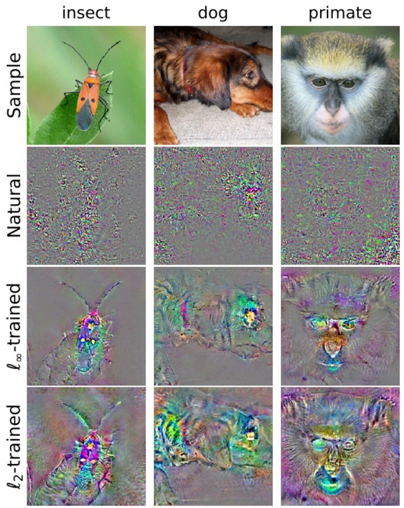
adversarial perturbation

"gibbon"

Robust Models Have Better Explanations







Tsipras et al. 2019
Etmann et al. 2019

Goal

Main Question

Why robust models have more interpretable explanations?

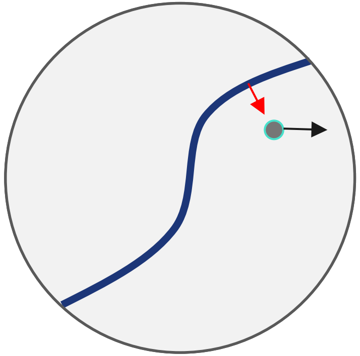
Main Methods

- Geometry-based Analysis
- Decision Boundary

Alignment of Explanations and Boundaries

Contribution 1

In robust models, explanations better align with normal vectors of decision boundaries



Explanation vector

Normal vector of decision boundary

How a model separates classes

CIFAR-10	standard	robust[1]
ℓ_2 *dist*	59.96	**1.23**
cos *dist*	0.44	**0.05**

ImageNet	standard	robust[1]
ℓ_2 *dist*	8.48	**0.41**
cos *dist*	0.28	**0.13**

[1]Robust models are pretrained networks from PGD Training [Madry 2017 et al.]

Robust Models Have Aligned Explanations

Contribution 2

The better alignment can be proved for some robust one-layer network.

Corollary 3.4 (Informal)

In robust[1] models, explanations (Expl) are very close to normal vectors (n) of the decision boundaries

$$||Expl - n|| \leq \lambda$$

And $1/\lambda$ is proportional to the robustness.

[1] *Certified Adversarial Robustness via Randomized Smoothing [Cohen et al. ICML 19']*

Motivating Better Explanation Methods

We study explanations form its geometric property and relate it with adversarial robustness.

Contribution 1 & 2

In robust models, explanations align better with normal vectors of the decision boundary.

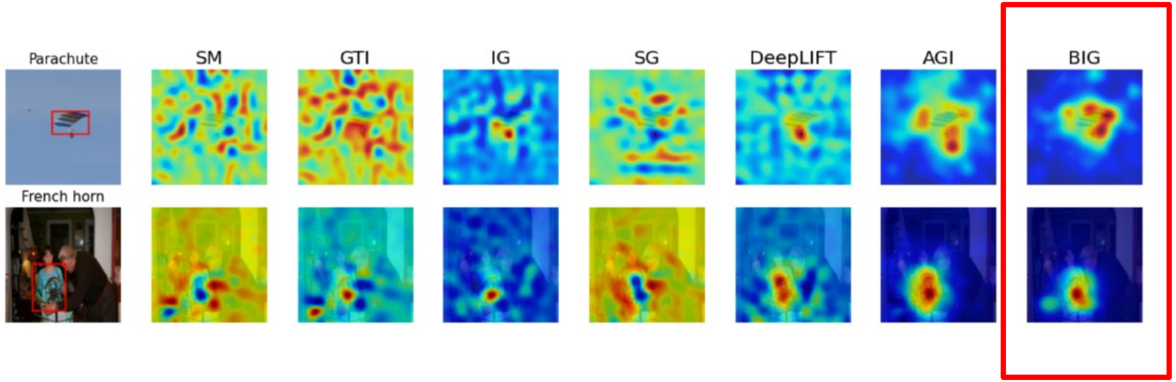
→

Searching for normal vectors of decision boundaries as explanations

Leveraging Nearby Boundaries To Explain Models

Contribution 3

Incorporating boundaries to explain model's decision, we introduce Boundary-based Integrated Gradient (BIG).



Thank You

Paper

Link | QR Code

Colab Demo

Link | QR Code

Github

Link | QR Code