UNIVERSITY OF MARYLAND 1856

NEURAL INFORMATION PROCESSING SYSTEMS

Exploring Blind Spots of Vision Models


Sriram Balasubramanian*


Gaurang Sriramanan*


Vinu Sankar Sadasivan


Soheil Feizi

Introduction

- Input ***over-sensitivity*** well studied in adversarial literature

$+ 0.02 \ *$

$=$

Prediction: Hamster
Confidence = 99.99%

50-step PGD targeted attack
with $\varepsilon = \frac{8}{255}$ scaled by 50x

Prediction: Banjo
Confidence = 100%

Introduction

- Input ***over-sensitivity*** well studied in adversarial literature
- We study input ***under-sensitivity*** for general models
- Uncover extent of excessive invariance in common vision models?

Over-Sensitivity	Under-Sensitivity
$\|\|f(x) - f(x')\|\|$ ↑	$f(x) \approx f(x')$
$x \approx x'$	$\|\|x - x'\|\|$ ↑

Mathematical Preliminaries

- For $g : \mathbb{R}^d \rightarrow \mathbb{R}, L_g(c) = \{x \in \mathcal{X} : g(x) = c\}$ is called the Level Set

- Important Property: For any curve in the Level Set $\gamma(t) : [0,1] \rightarrow L_g(c)$

$$\frac{d}{dt}(g(\gamma(t))) = 0 = \langle \nabla g(\gamma(t)), \gamma'(t) \rangle$$

Lemma 1. *If $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is a continuously differentiable function, then each of its regular level sets is an $(d-1)$ dimensional submanifold of $\mathbb{R}^d$.*

- How expansive are these submanifolds for common ML models?

Can we Traverse along Level Sets?

goose

Source Image $\boldsymbol{x}_s$

Confidence for class “goose” = 0.997
Confidence for class “Scottish Terrier” = 0

Scottish Terrier

Target Image $\boldsymbol{x}_t$

Confidence for class “goose” = 0
Confidence for class “Scottish Terrier” = 1.0

Level Set Traversal (LST) Algorithm

Repeat until Max Iterations

Compute Input Gradient

$$\Delta \boldsymbol{x} = \boldsymbol{x}_t - \boldsymbol{x}$$
$$\boldsymbol{g} = \nabla_{\boldsymbol{x}} CE(f(\boldsymbol{x}), y)$$

Compute Orthogonal Projection

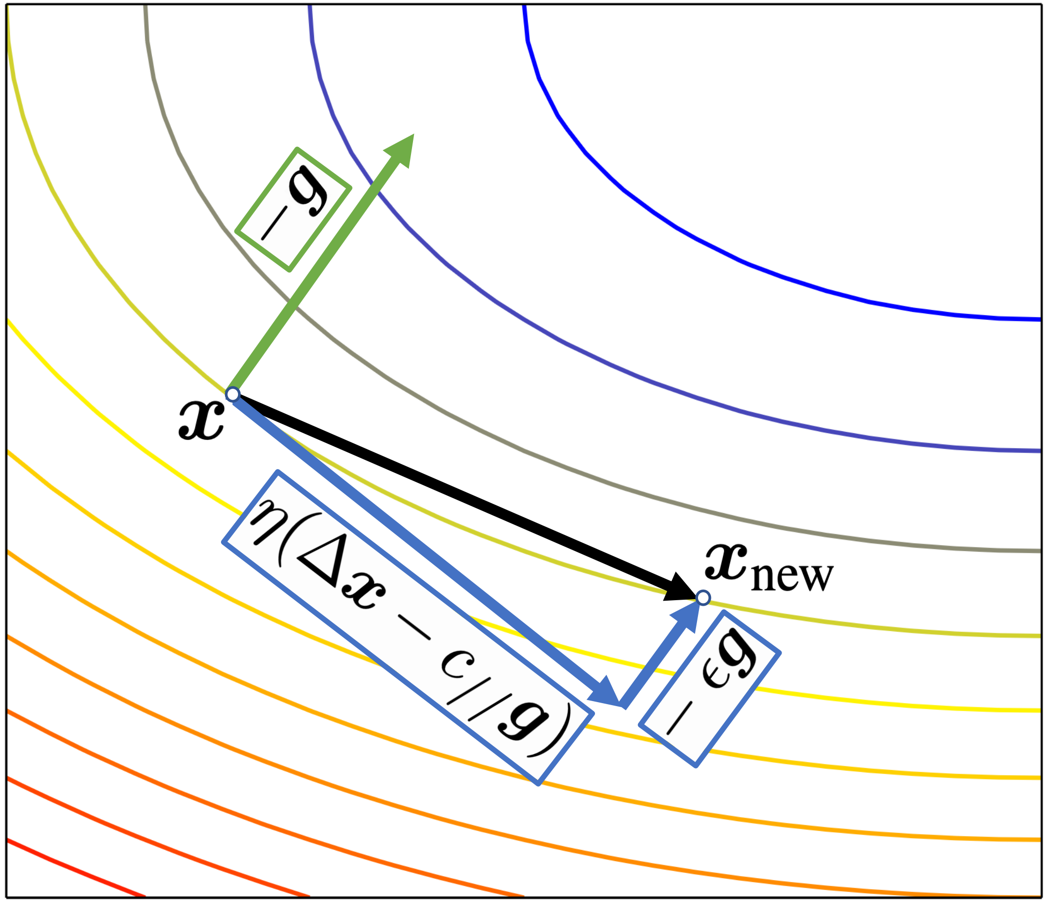
$$c_{//} = (\boldsymbol{g} \cdot \Delta \boldsymbol{x}) / ||\boldsymbol{g}||^2$$
$$\Delta \boldsymbol{x}_{\perp} = \eta(\Delta \boldsymbol{x} - c_{//} \boldsymbol{g})$$

Update Image

$$\boldsymbol{x}_{||} = \Pi_{\infty}(\boldsymbol{x}_{||} - \epsilon \boldsymbol{g}, -\epsilon, \epsilon)$$
$$\boldsymbol{x}_{\text{new}} = \boldsymbol{x} + \Delta \boldsymbol{x}_{\perp} + \boldsymbol{x}_{||}$$

Verify Model Confidence

if $f(\boldsymbol{x}_s)[j] - f(\boldsymbol{x}_{\text{new}})[j] > \delta$ **then**
  return $\boldsymbol{x}$
$\boldsymbol{x} = \boldsymbol{x}_{\text{new}}$



LST Path in Input Space for ResNet-50

goose

Source Image $\boldsymbol{x}_s$

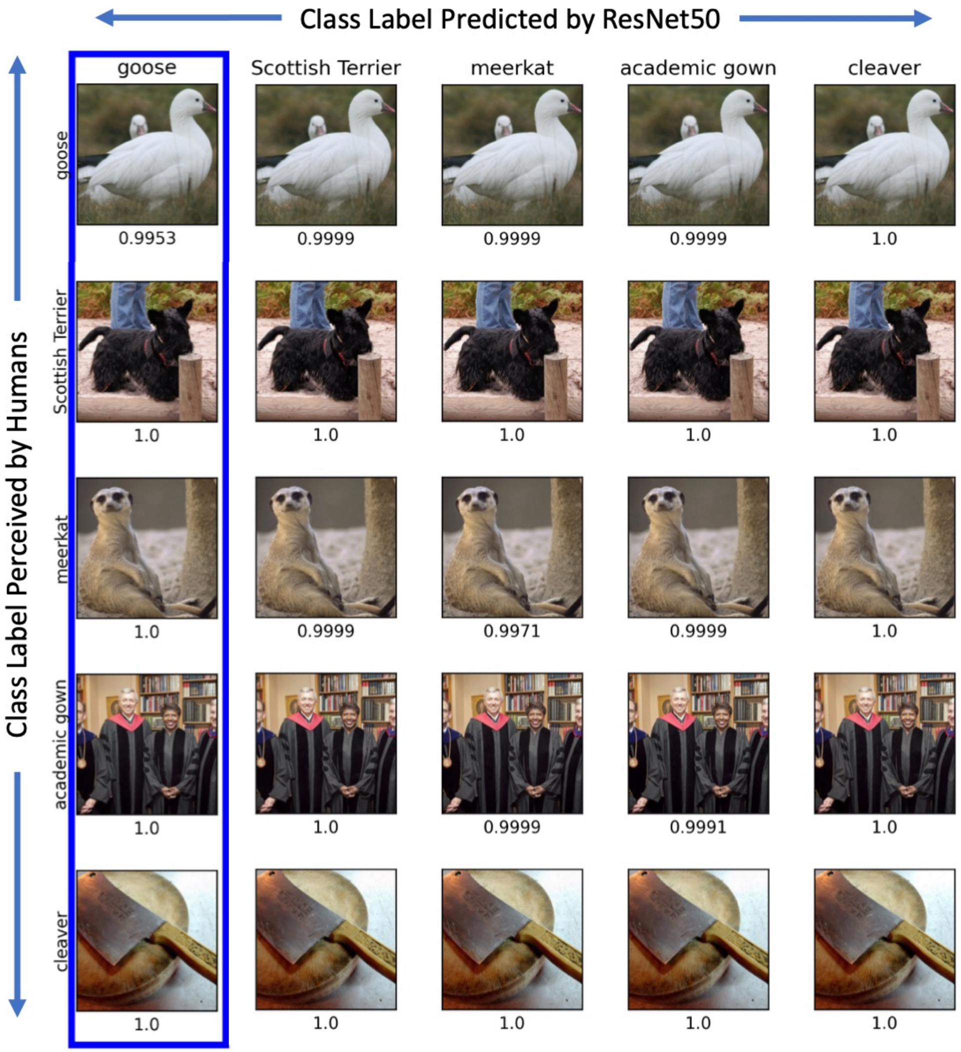
Confidence for class “goose” = 0.997
Confidence for class “Scottish Terrier” = 0

Scottish Terrier

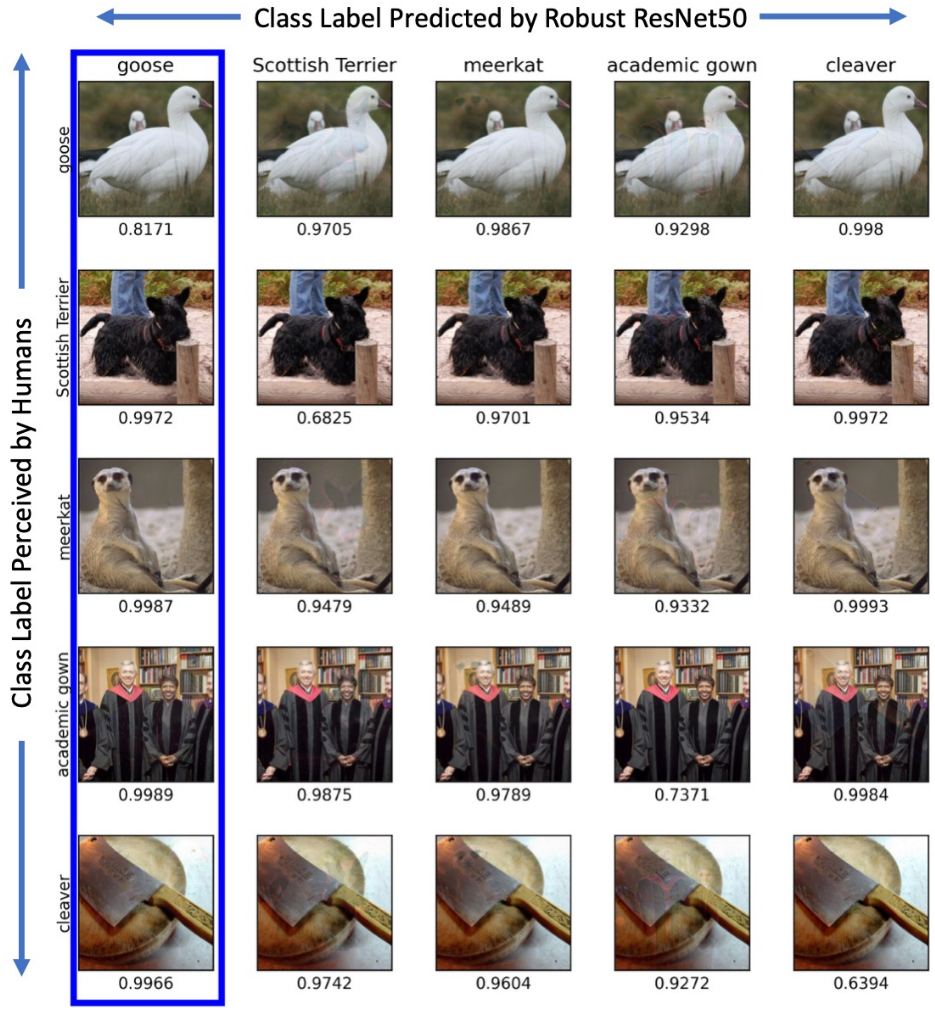
Target Image $\boldsymbol{x}_t$

Confidence for class “goose” = 0
Confidence for class “Scottish Terrier” = 1.0

LST over arbitrary Source-Target pairs



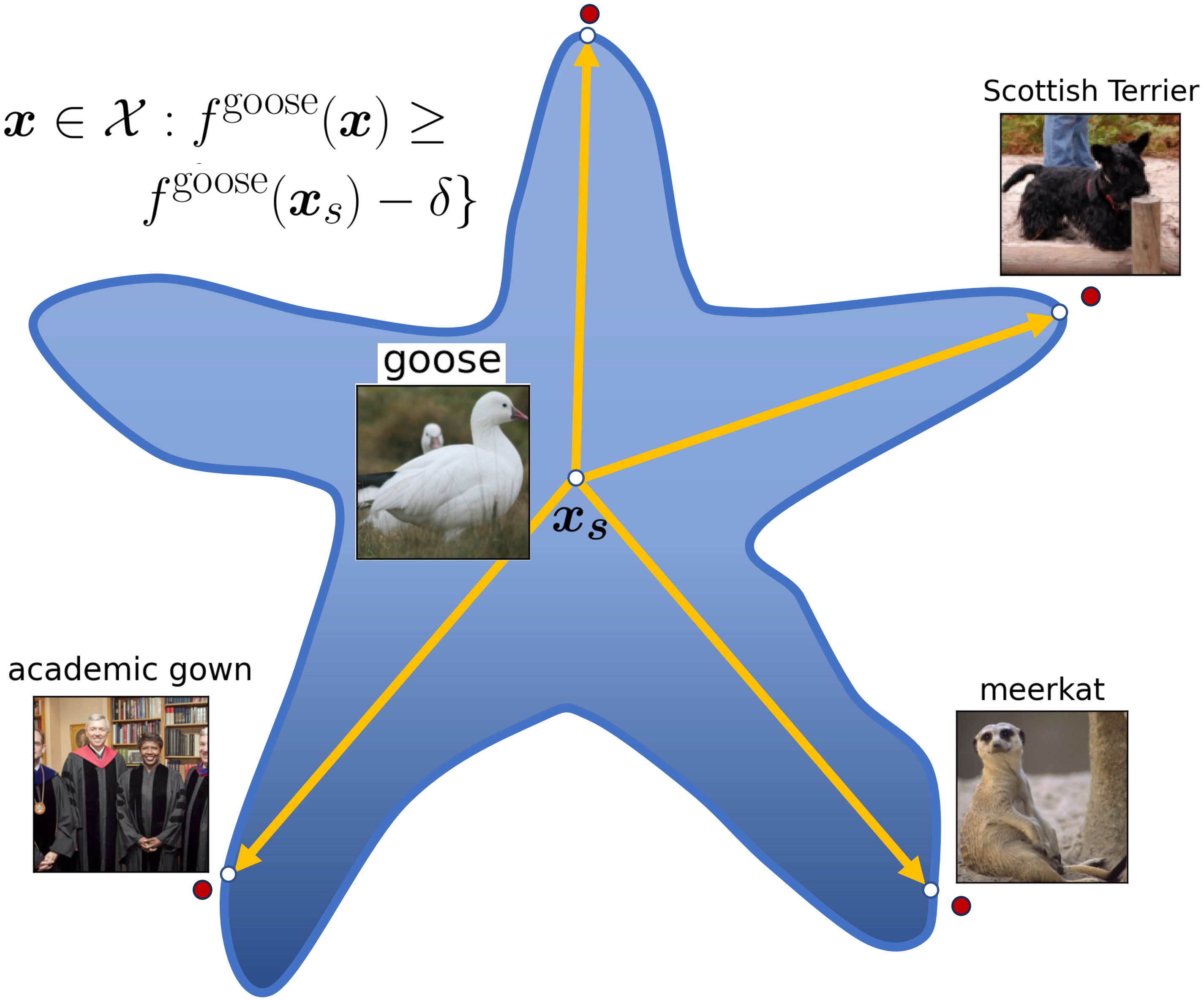
Normally Trained ResNet-50



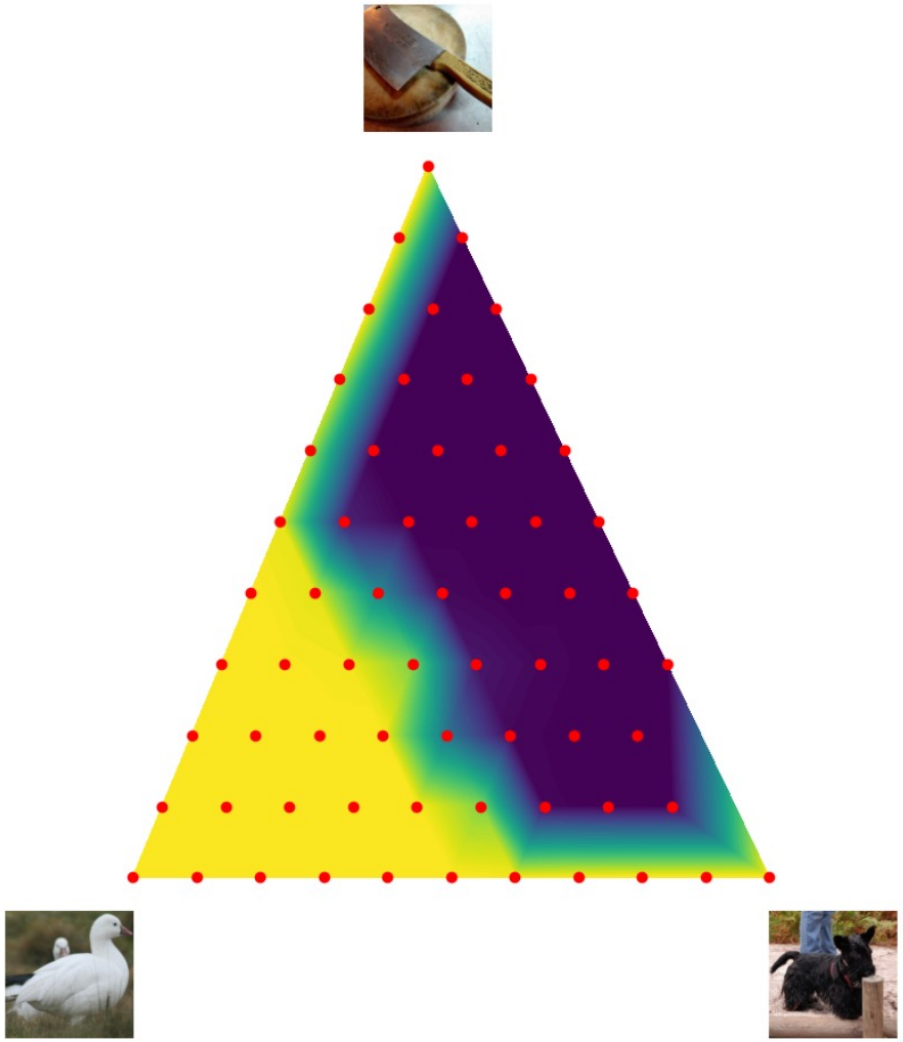
Adversarially Trained ResNet-50

Star-like Substructure of Level Sets

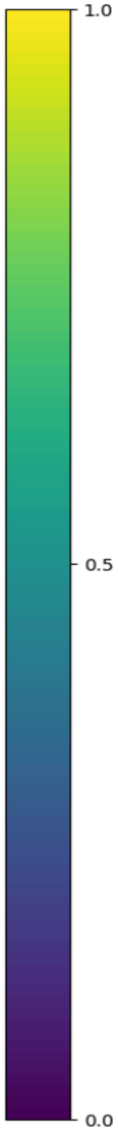
$$\{\boldsymbol{x} \in \mathcal{X} : f^{\text{goose}}(\boldsymbol{x}) \geq f^{\text{goose}}(\boldsymbol{x}_s) - \delta\}$$

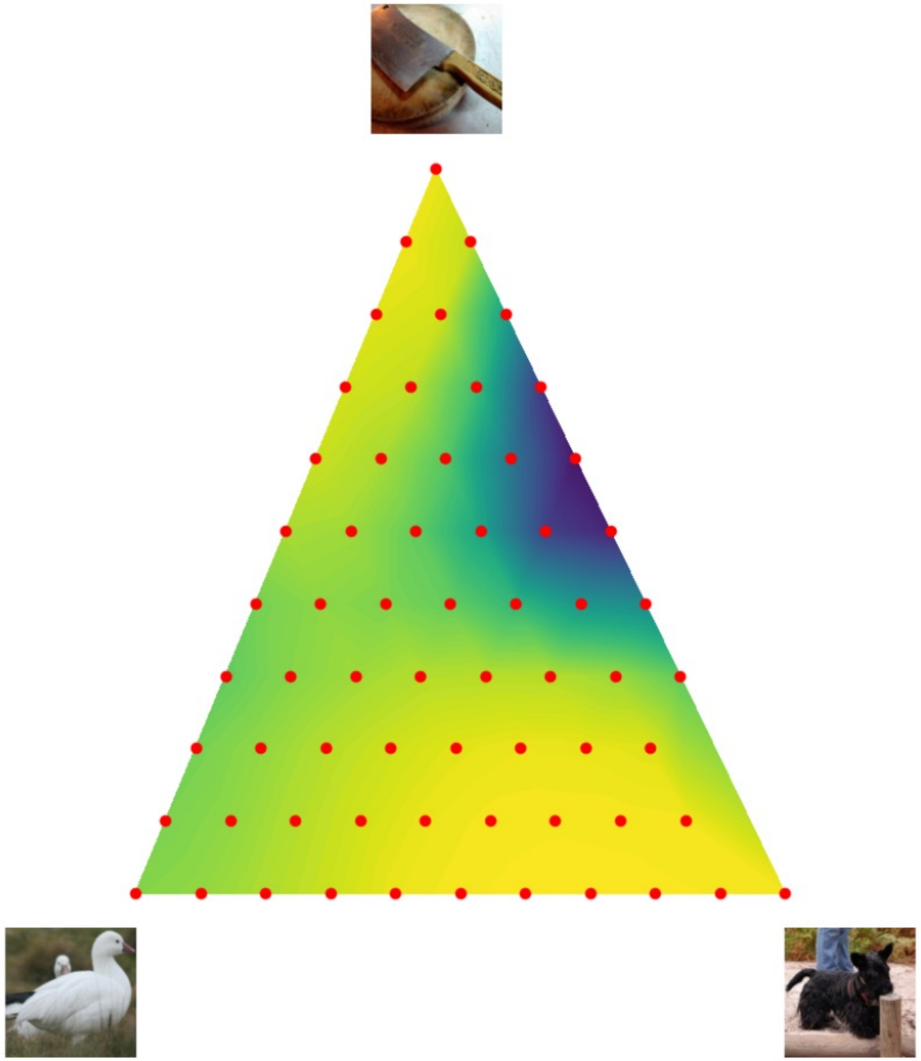


Star-like Substructure of Level Sets



Normally Trained ResNet-50

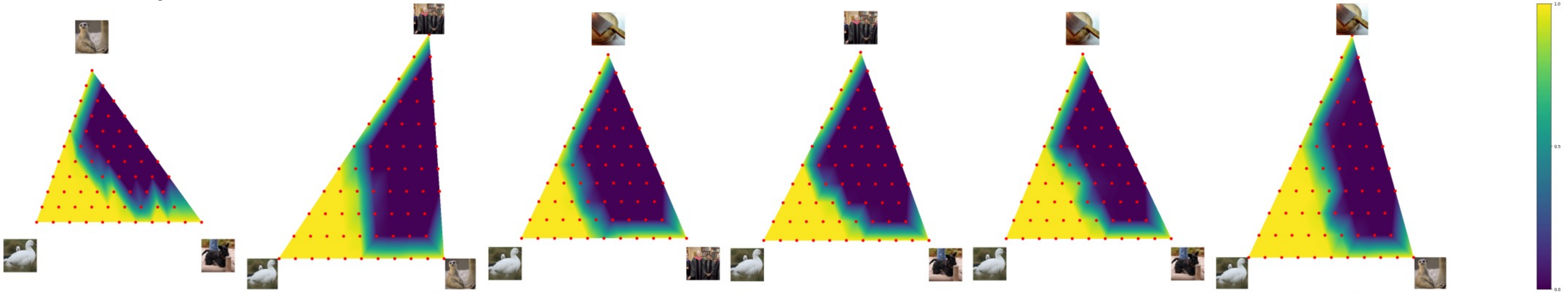




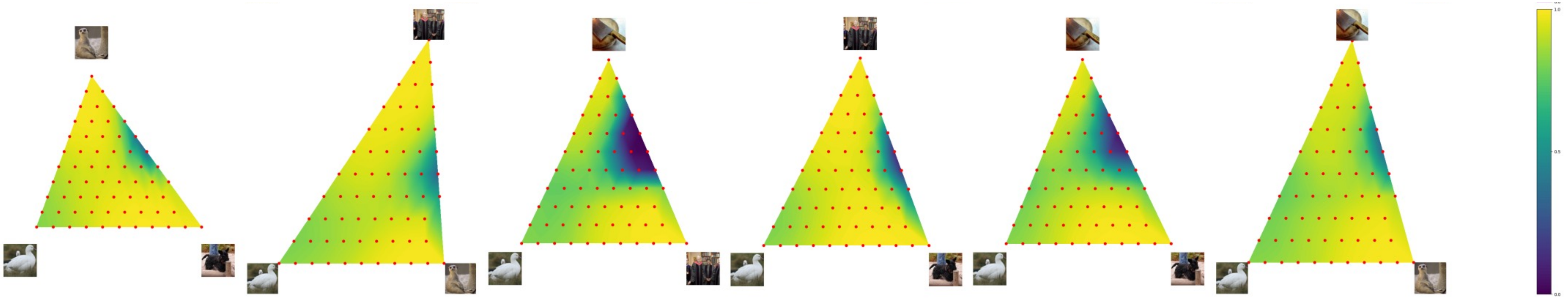
Adversarially Trained ResNet-50

Star-like Substructure of Level Sets

Normally Trained ResNet-50:



Adversarially Trained ResNet-50:



Quantitative Analysis of Blind Spot Invariance

Distance metrics:

1. RMSE:
   Root mean squared error
2. Max norm (ℓ_∞):
   Maximum absolute difference
3. SSIM:
   Structural Similarity Index
4. LPIPS:
   Perceptual Image Similarity

Confidence metrics:

1. Source confidence (p_{src}):
   Confidence of the model for the source image
2. Average path confidence
   Mean confidence over the linear paths connecting the source image to LST outputs
3. Average Δ confidence:
   Mean confidence over the enclosed triangle
4. Average Δ fraction for a given δ:
   Fraction of triangle over which confidence is at least $p_{src} - \delta$

Quantitative Analysis of Blind Spot Invariance

Table 1: Quantitative image distance metrics between output of Level Set Traversal and target images.

Models	RMSE : $\mu \pm \sigma$	ℓ_∞ dist: $\mu \pm \sigma$	SSIM: $\mu \pm \sigma$	LPIPS dist: $\mu \pm \sigma$
ResNet-50 (Normal)	0.008 ± 0.001	0.046 ± 0.020	0.990 ± 0.021	0.002 ± 0.004
ResNet-50 (AT)	0.029 ± 0.008	0.746 ± 0.124	0.915 ± 0.041	0.057 ± 0.037
DeiT-S (Normal)	0.011 ± 0.002	0.116 ± 0.030	0.973 ± 0.024	0.024 ± 0.017
DeiT-S (AT)	0.046 ± 0.010	0.821 ± 0.117	0.898 ± 0.041	0.219 ± 0.068

Table 2: Quantitative confidence metrics over the triangular convex hull (Δ) of a given source image and two target LST blindspot image-pairs and over linear interpolant paths between source and blindspot images. (For reference, a random classifier would have confidence of 0.001)

Models	p_{src} ($\mu \pm \sigma$)	Avg Δ Conf. ($\mu \pm \sigma$)	Avg Δ Frac. ($\mu \pm \sigma$)				Avg Path Conf. ($\mu \pm \sigma$)
			$\delta = 0.0$	$\delta = 0.1$	$\delta = 0.2$	$\delta = 0.3$	
ResNet-50 (Normal)	0.99 ± 0.02	0.56 ± 0.10	0.13 ± 0.15	0.51 ± 0.11	0.53 ± 0.1	0.54 ± 0.10	0.96 ± 0.05
ResNet-50 (AT)	0.88 ± 0.11	0.83 ± 0.09	0.49 ± 0.29	0.79 ± 0.13	0.85 ± 0.1	0.88 ± 0.09	0.93 ± 0.06
DeiT-S (Normal)	0.85 ± 0.06	0.68 ± 0.05	0.54 ± 0.11	0.67 ± 0.06	0.71 ± 0.06	0.73 ± 0.06	0.94 ± 0.02
DeiT-S (AT)	0.76 ± 0.08	0.59 ± 0.07	0.20 ± 0.09	0.43 ± 0.14	0.63 ± 0.15	0.76 ± 0.12	0.73 ± 0.06

Conclusions

- Using LST, we find that the level sets of common vision models is **remarkably expansive**
- The **linear** path from any given source image to LST blind spot outputs retain **high model confidence** throughout for arbitrary targets
- This unveils a **star-like substructure** for the equi-confidence level sets of common models
- Adversarially trained models are significantly more **under-sensitive,** over inputs **well beyond** the original threat model

Thank You!