Robust Neural Networks

Part 3: Uncertainty at Inference

Objective

Objective of the Tutorial

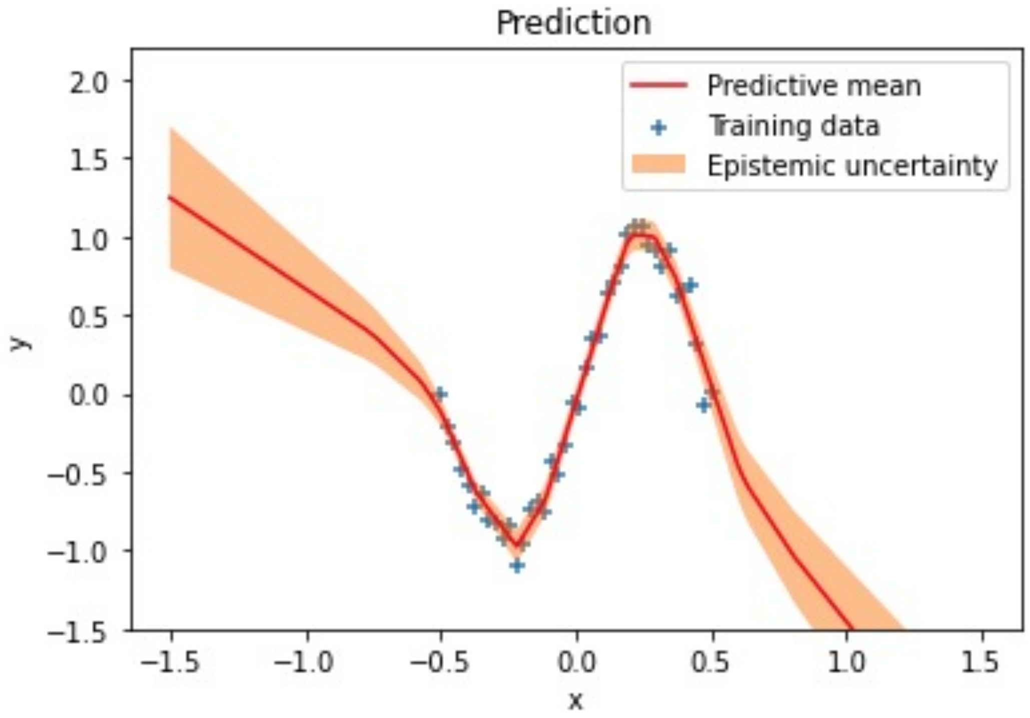
To discuss methodologies that promote robustness in neural networks at inference

- Part 1: Inference in Neural Networks
- Part 2: Explainability at Inference
- **Part 3: Uncertainty at Inference**
    - Uncertainty Definition
    - Uncertainty Quantification
    - Gradient-based Uncertainty
    - Adversarial and Corruption Detection
- Part 4: Intervenability at Inference
- Part 5: Conclusions and Future Directions

Uncertainty

What is Uncertainty?

Uncertainty is a model knowing that it does not know



A simple example:

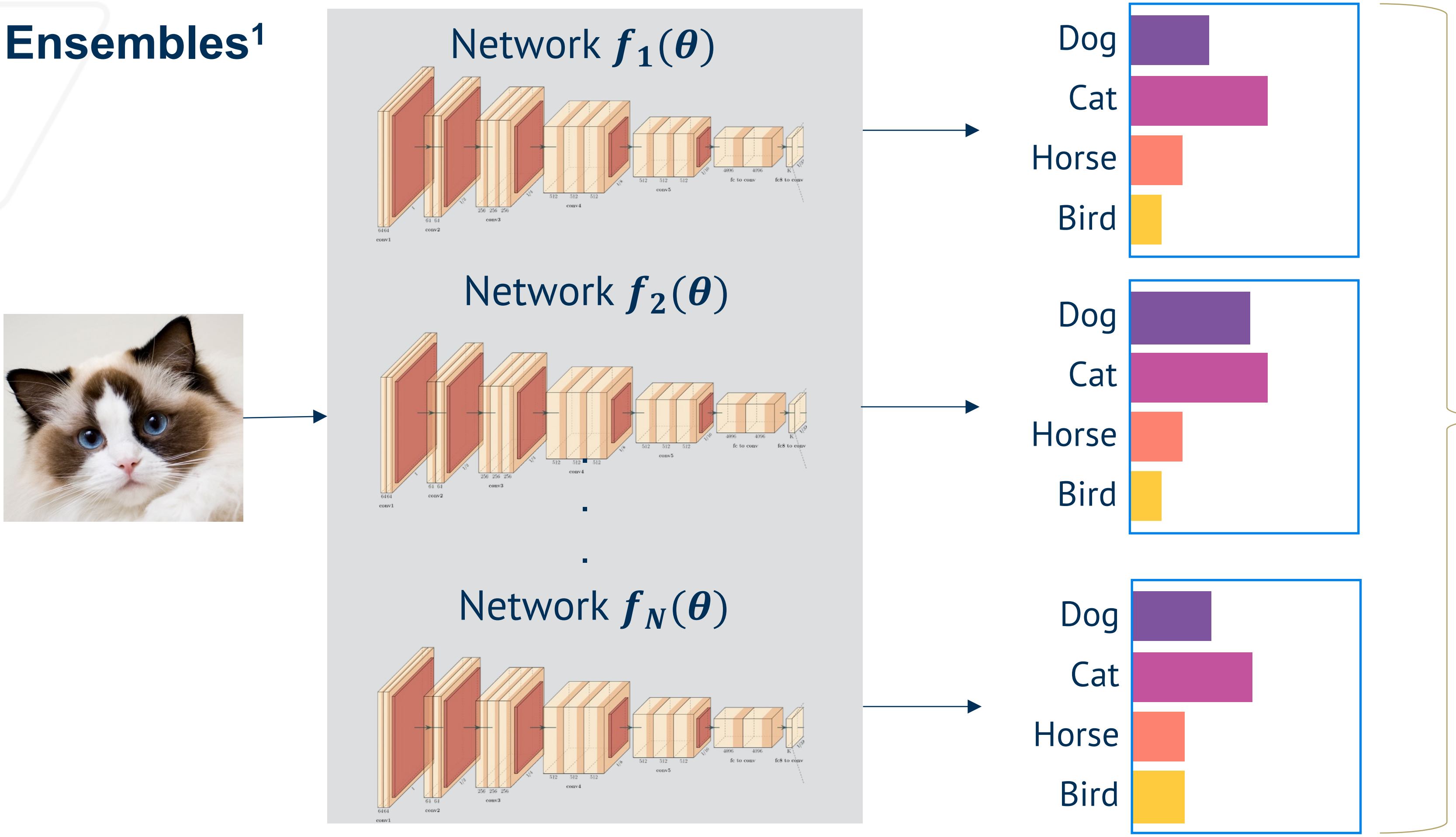
- When training data is **available: Less uncertainty**
- When training data is **unavailable: More uncertainty**

http://krasserm.github.io/2020/09/25/reliable-uncertainty-estimates/

Uncertainty

Uncertainty Quantification in Neural Networks

Via Ensembles[1]



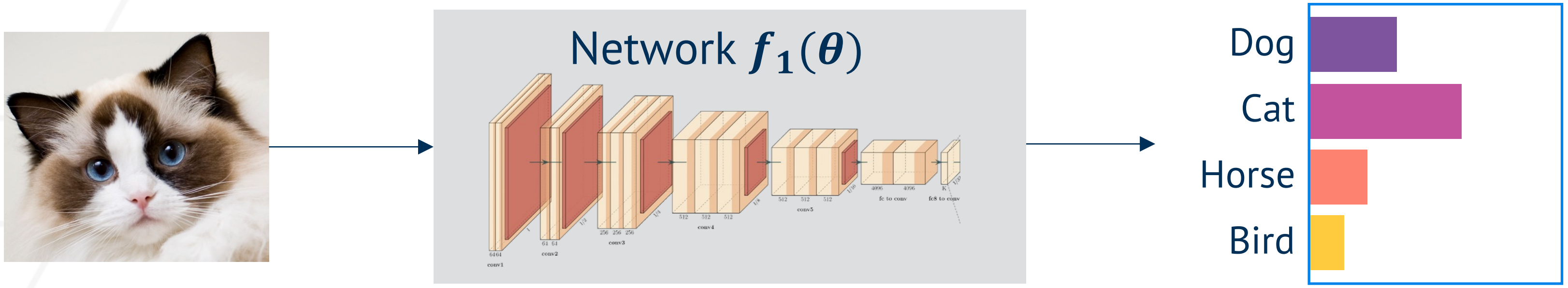
Variation within outputs $Var(y)$ is the uncertainty. Commonly referred to as **Prediction Uncertainty.**

[1] Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles." *Advances in neural information processing systems* 30 (2017).

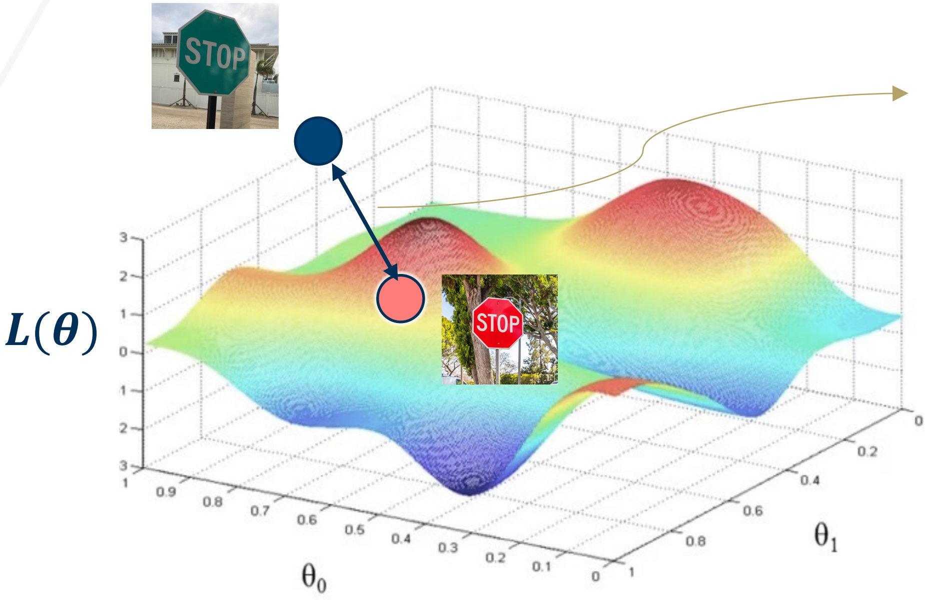
Uncertainty

Uncertainty Quantification in Neural Networks

Via Single pass methods[1]



Uncertainty quantification using a single network and a single pass



Calculate distance from some trained clusters

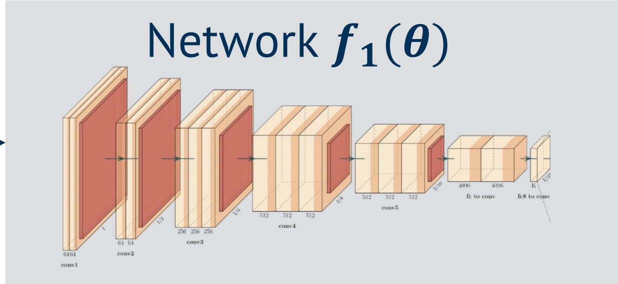
Does not require multiple networks!

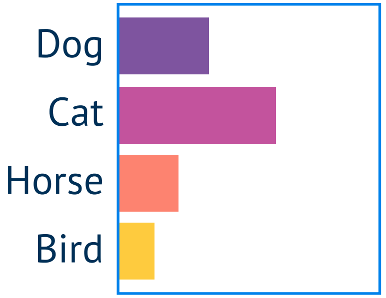
[1Van Amersfoort, J., Smith, L., Teh, Y. W., & Gal, Y. (2020, November). Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning* (pp. 9690-9700). PMLR.

Uncertainty

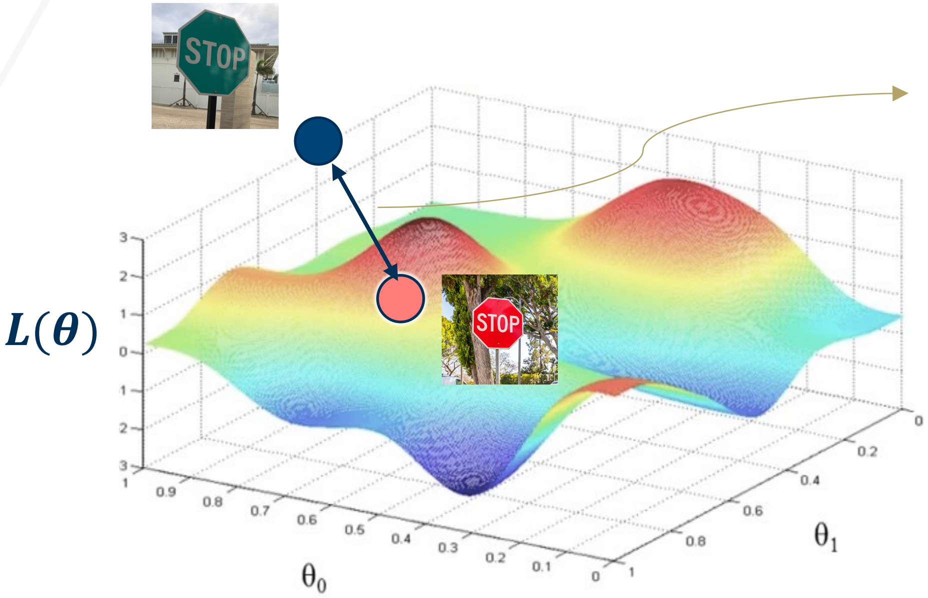
Gradients as Single pass Features

Our Goal: Use gradients to characterize the novel data at Inference





Uncertainty quantification using a single network and a single pass



Calculate distance from some trained clusters

Does not require multiple networks!

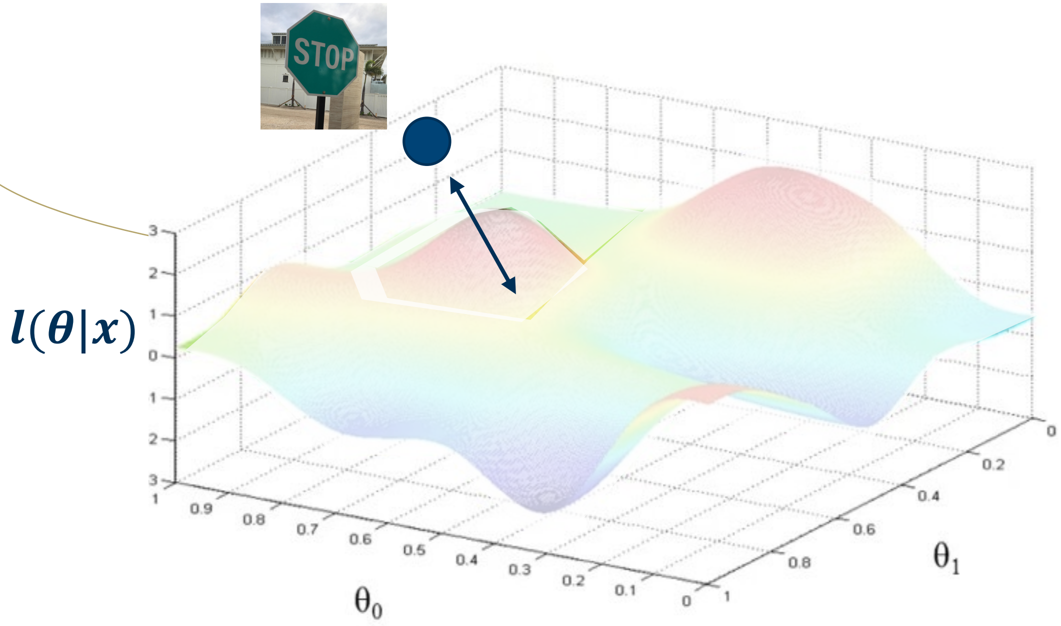
Challenge: Class and prediction cannot be trusted!

Uncertainty

Gradients as Single pass Features

Our Goal: Use gradients to characterize the novel data at Inference, without global information

Distance from unknown cluster



Two techniques:

1. **Gradient constraints during Training for Anomaly Detection**
2. Backpropagating Confounding labels for Out-of-Distribution Detection

ECCV'20 ONLINE
23-28 AUGUST 2020

Backpropagated Gradient Representations for Anomaly Detection

Gukyeong Kwon, PhD
Amazon AWS

Mohit Prabhushankar, PhD
Postdoc, Georgia Tech

Ghassan AlRegib, PhD
Professor, Georgia Tech

SCAN ME

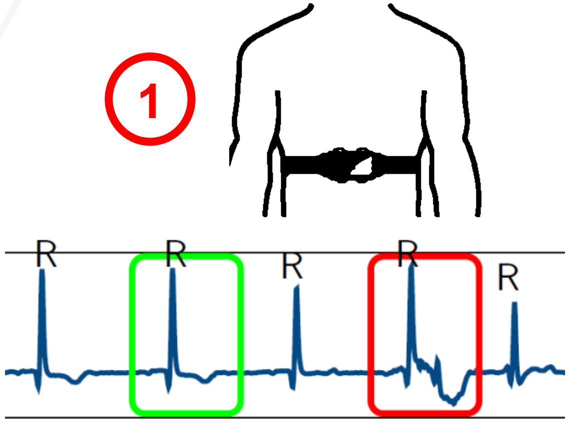
Anomalies

Finding Rare Events in Normal Patterns

Backpropagated Gradient Representations for Anomaly Detection

'Anomalies are patterns in data that do not conform to a well defined notion of normal behavior' [1]

①



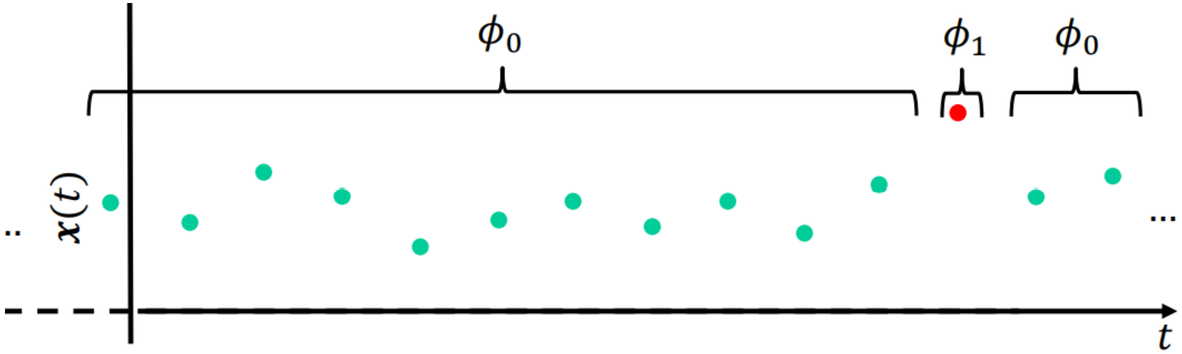
②

Statistical Definition:

- Normal data are generated from a stationary process P_N
- Anomalies are generated from a different process $P_A \neq P_N$

Goal: Detect ϕ_1

$$x(t) = \begin{cases} \phi_0 & \text{Normal data} \\ \phi_1 & \text{Anomalies} \end{cases}$$



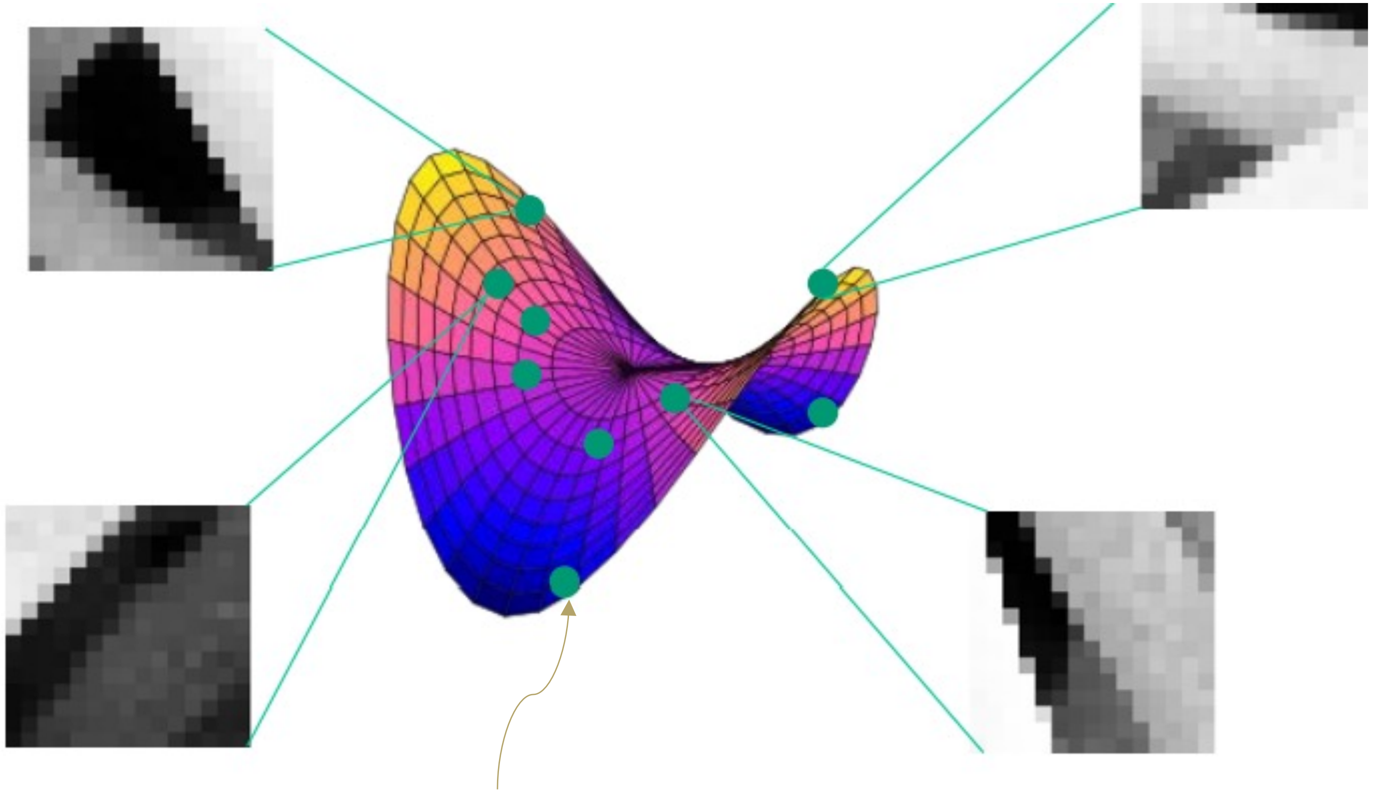
[1] V. Chandola, A. Banerjee, V. Kumar. "Anomaly detection: A survey". ACM Comput. Surv. 41, 3, Article 15 (July 2009), 58 pages

Anomalies

Steps for Anomaly Detection

Step 1: Constrain manifolds, Step 2: Detect statistically implausible projections

- Step 1 ensures that patches from natural images live close to a low dimensional manifold
- Step 2 designs distance functions that detect *implausibility* based on constraints



Constraining Manifolds

General Constraints

SCAN ME

Backpropagated Gradient Representations for Anomaly Detection

Constrained Representation

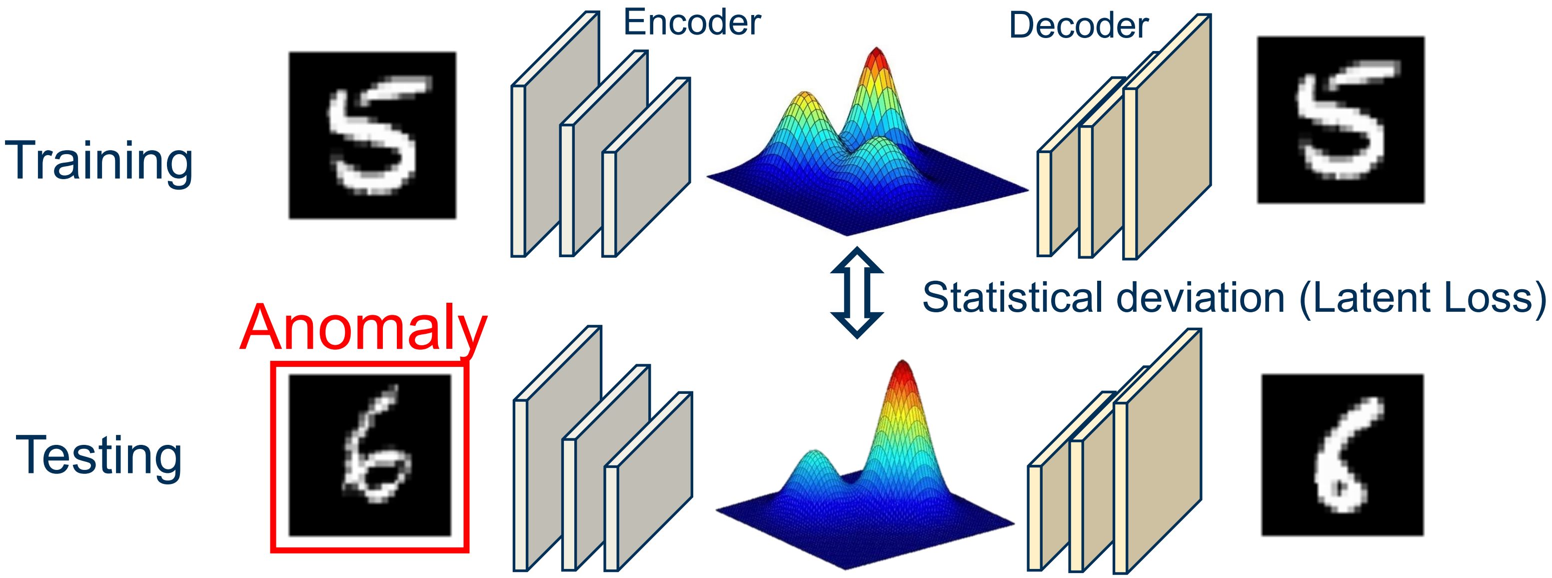

2004 – Tax et.al [1]

2016 – Fan et.al [2]

2018 – Pidhorksyi et.al [3]

2019 – Abati et.al [4]

Activations are constrained using GANs, VAEs, etc.



---

[1] David MJ Tax and Robert PW Duin. Support vector data description. Machine learning, 54(1):45–66, 2004.

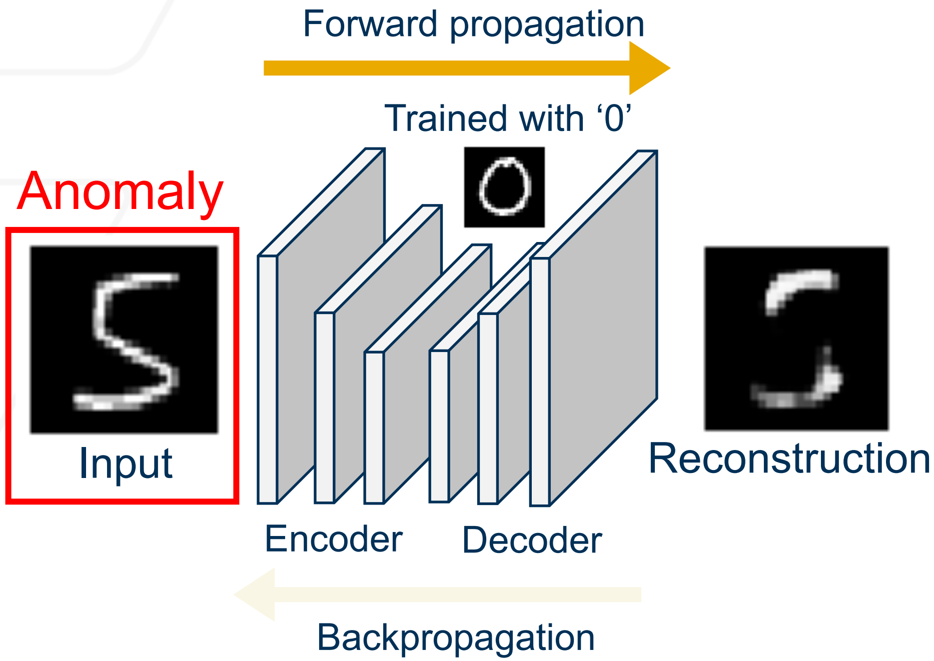
[2] Yaxiang Fan, Gongjian Wen, Deren Li, Shaohua Qiu, and Martin D Levine. Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder. arXiv preprint arXiv:1805.11223, 2018. 1, 2

[3] S. Pidhorskyi, R. Almohsen, and G. Doretto, “Generative probabilistic novelty detection with adversarial autoencoders,” in Advances in Neural Information Processing Systems, 2018, pp. 6822–6833.

[4] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, “Latent space autoregression for novelty detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 481–490.

Constraining Manifolds

Gradient-based Constraints



Activation Constraints

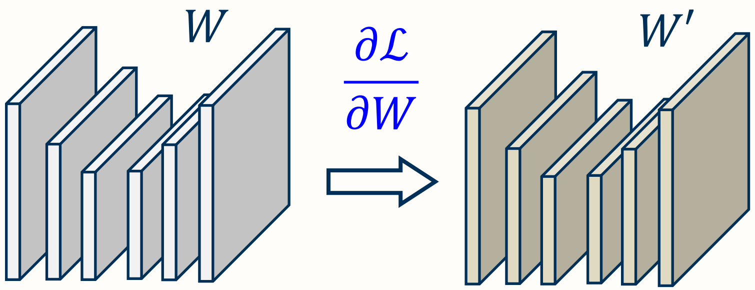
Activation-based representation (Data perspective)

e.g. Reconstruction error ($\mathcal{L}$)

How much of the input does not correspond to the learned information?

Gradient Constraints

Gradient-based Representation (**Model** perspective)



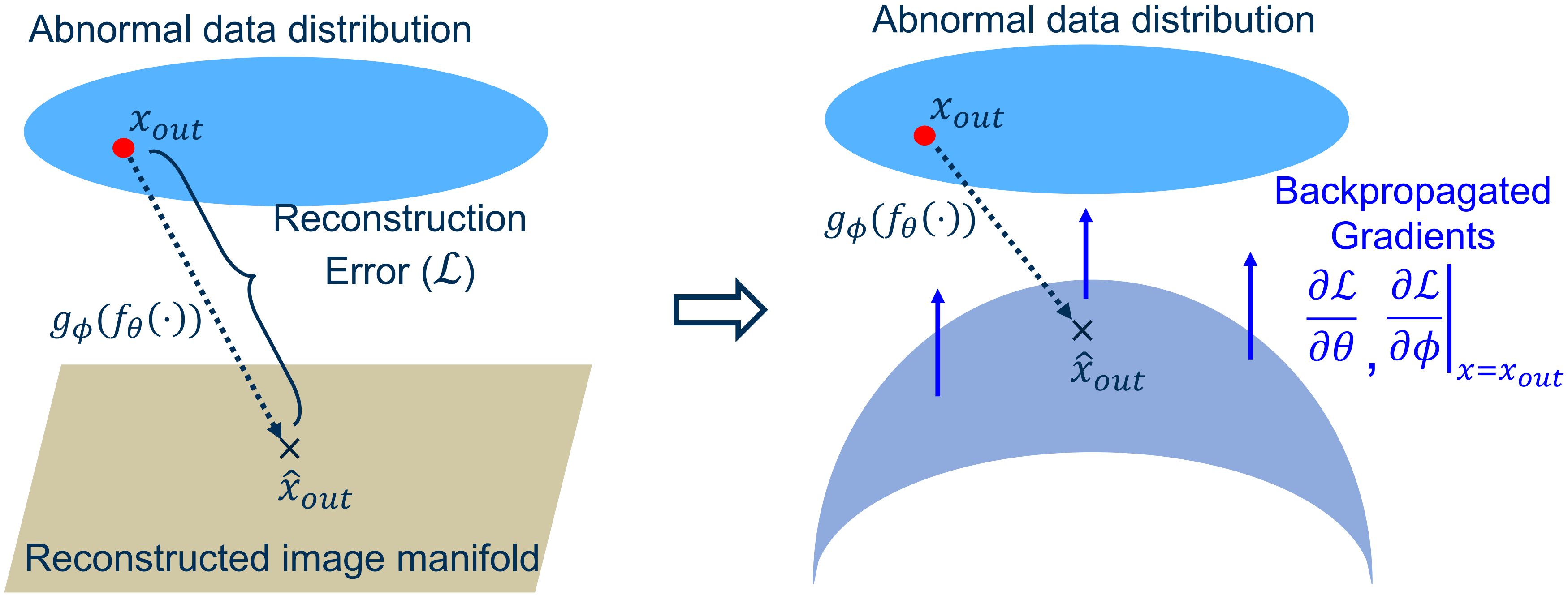
How much **model update** is required by the input?

G. Kwon, M. Prabhushankar, D. Temel, and G. AlRegib, "Backpropagated Gradient Representations for Anomaly Detection," 2020

Constraining Manifolds

Advantages of Gradient-based Constraints

- **Gradients provide directional information to characterize anomalies**
- **Gradients from different layers capture abnormality at different levels of data abstraction**



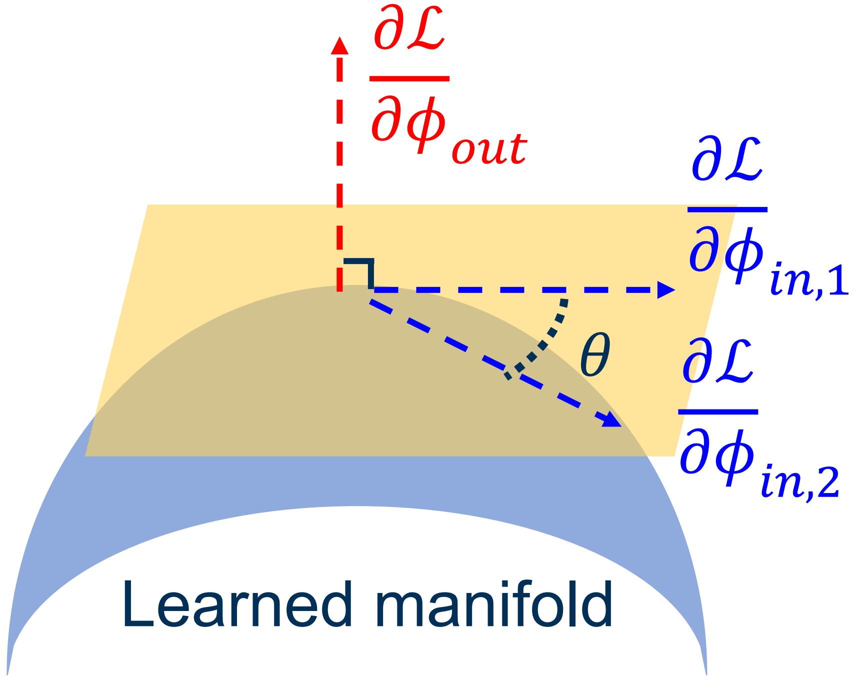
G. Kwon, M. Prabhushankar, D. Temel, and G. AlRegib, "Backpropagated Gradient Representations for Anomaly Detection," 2020

GradCON: Gradient Constraint

Gradient-based Constraints

Backpropagated Gradient
Representations for Anomaly Detection

Constrain gradient-based representations during training to obtain clear separation between normal data and abnormal data



ϕ: Weights $\mathcal{L}$: Reconstruction error

At *k*-th step of training,

$$J = \mathcal{L} - \mathbb{E}_i\left[\overbrace{\text{cosSIM}\left(\frac{\partial J}{\partial \phi_{i_{avg}}}^{k-1}, \frac{\partial \mathcal{L}}{\partial \phi_i}^{k}\right)}^{\text{Gradient loss}}\right]$$

Avg. training gradients until (k-1) th iter.

Gradients at k-th iter.

where $\frac{\partial J}{\partial \phi_{i_{avg}}}^{k-1} = \sum_{t=1}^{k-1} \frac{\partial J}{\partial \phi_i}^{t}$

G. Kwon, M. Prabhushankar, D. Temel, and G. AlRegib, "Backpropagated Gradient Representations for Anomaly Detection," 2020

GradCON: Gradient Constraint

Activations vs Gradients

Backpropagated Gradient Representations for Anomaly Detection

Abnormal “class” detection (CIFAR-10)

e.g.

Normal

Abnormal

AUROC Results

Model	Loss	Plane	Car	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck	Average
CAE	Recon	0.682	0.353	0.638	0.587	0.669	**0.613**	0.495	0.498	0.711	0.390	0.564
CAE + Grad	Recon	0.659	0.356	**0.640**	0.555	0.695	0.554	0.549	0.478	0.695	0.357	0.554
	Grad	**0.752**	0.619	0.622	0.580	0.705	0.591	0.683	**0.576**	**0.774**	**0.709**	**0.661**
VAE	Recon	0.553	0.608	0.437	0.546	0.393	0.531	0.489	0.515	0.552	0.631	0.526
	Latent	0.634	0.442	**0.640**	0.497	**0.743**	0.515	**0.745**	0.527	0.674	0.416	0.583
VAE + Grad	Recon	0.556	0.606	0.438	0.548	0.392	0.543	0.496	0.518	0.552	0.631	0.528
	Latent	0.586	0.396	0.618	0.476	0.719	0.474	0.698	0.537	0.586	0.413	0.550
	Grad	0.736	**0.625**	0.591	**0.596**	0.707	0.570	0.740	0.543	0.738	0.629	0.647

Recon: Reconstruction error, Latent: Latent loss, Grad: Gradient loss

- (CAE vs. CAE + Grad) Effectiveness of the gradient constraint
- (CAE vs. VAE) Performance sacrifice from the latent constraint
- (VAE vs. VAE + Grad) Complementary features from the gradient constraint

G. Kwon, M. Prabhushankar, D. Temel, and G. AlRegib, "Backpropagated Gradient Representations for Anomaly Detection," 2020

GradCON: Gradient Constraint

Aberrant Condition Detection

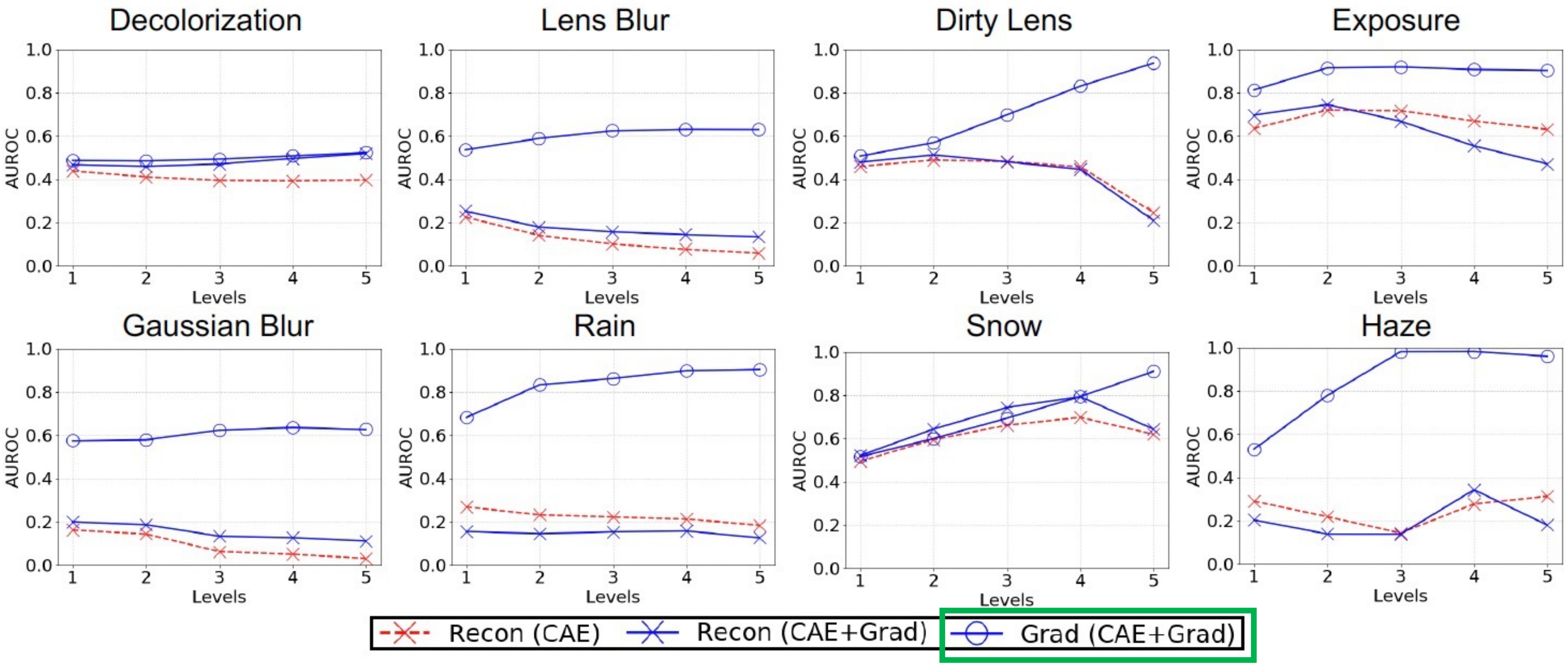
Backpropagated Gradient Representations for Anomaly Detection

Abnormal "condition" detection (CURE-TSR)

Normal

Abnormal

AUROC Results



Recon: Reconstruction error, Grad: Gradient loss

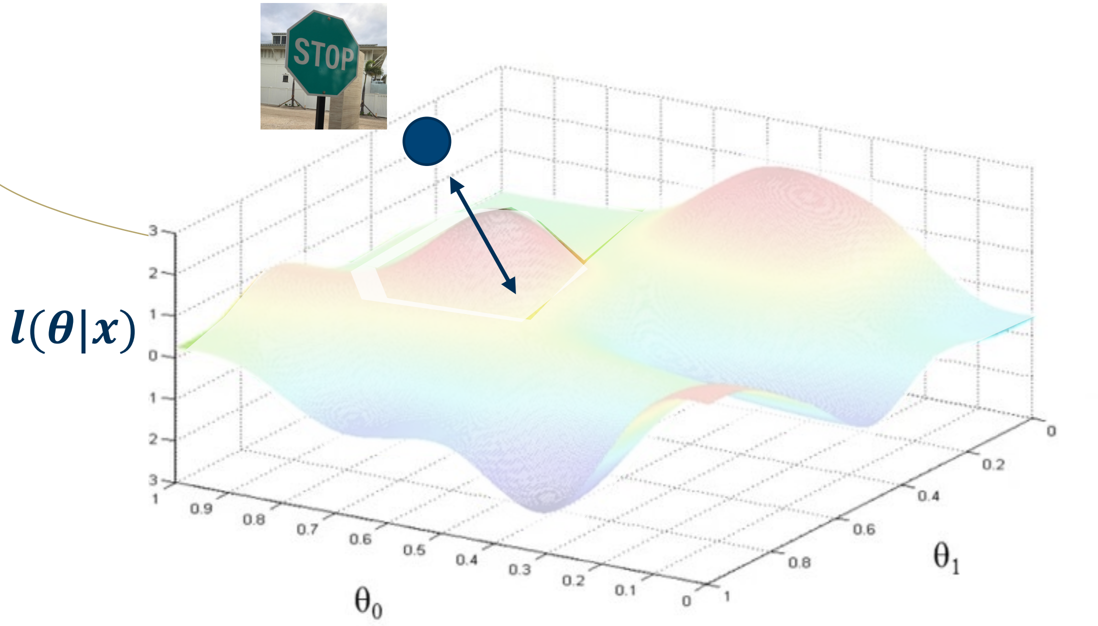
G. Kwon, M. Prabhushankar, D. Temel, and G. AlRegib, "Backpropagated Gradient Representations for Anomaly Detection," 2020

Uncertainty

Gradients as Single pass Features

Our Goal: Use gradients to characterize the novel data at Inference, without global information

Distance from unknown cluster



Two techniques:

1. Gradient constraints during Training for Anomaly Detection
2. **Backpropagating Confounding labels for Out-of-Distribution Detection**

IEEE Access

Probing the Purview of Neural Networks via Gradient Analysis

Jinsol Lee,
PhD Candidate

Mohit Prabhushankar, PhD
Postdoc

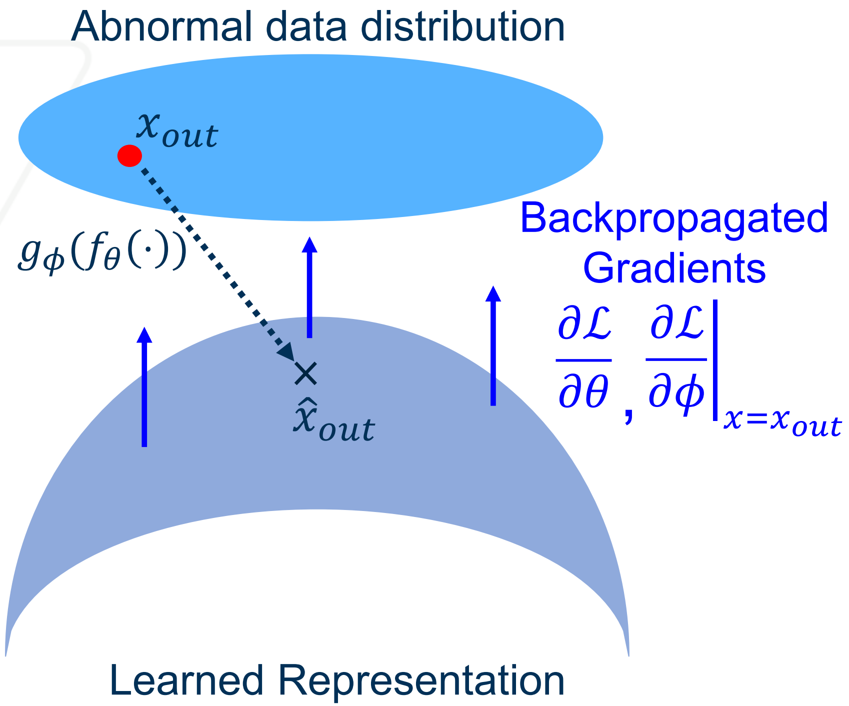
Ghassan AlRegib, PhD
Professor

SCAN ME

Uncertainty in Neural Networks

Principle

Principle: Gradients provide a distance measure between the learned representations space and novel data



However, what is $\mathcal{L}$?

- In anomaly detection, the loss was between the input and its reconstruction
- In prediction tasks, there is neither the reconstructed input nor ground truth

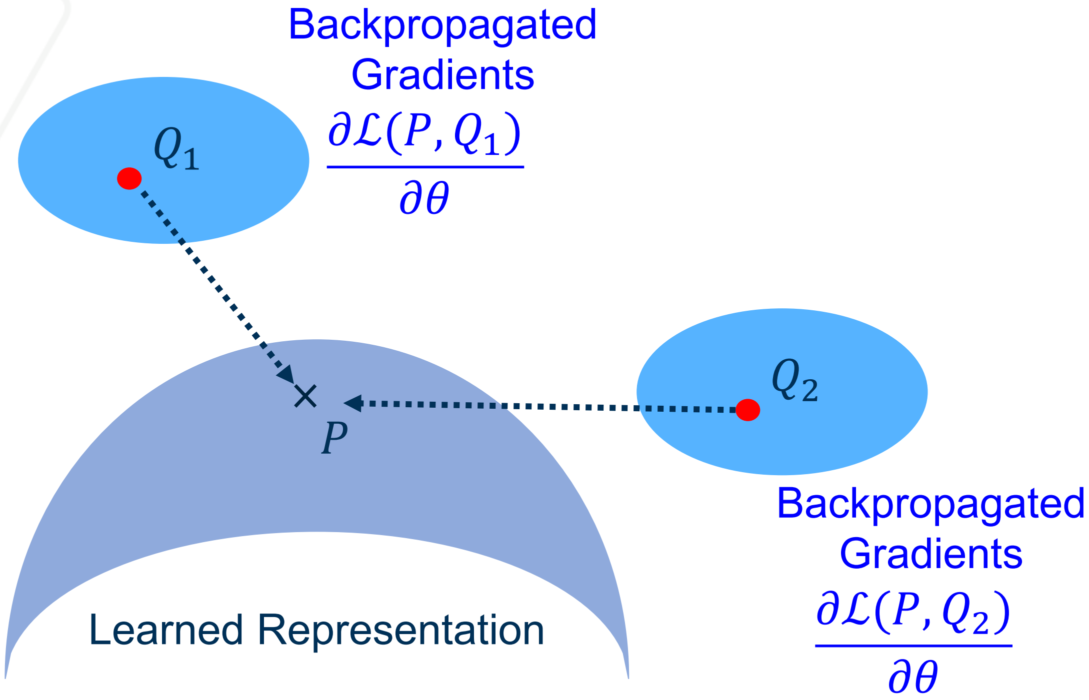
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Uncertainty in Neural Networks

Principle

Principle: Gradients provide a distance measure between the learned representations space and novel data

P = Predicted class
Q_1 = Contrast class 1
Q_2 = Contrast class 2



However, what is $\mathcal{L}$?

- In anomaly detection, the loss was between the input and its reconstruction
- In prediction tasks, there is neither the reconstructed input nor ground truth

- **We backpropagate all contrast classes - $Q_1, Q_2 \ldots Q_N$ by backpropagating N one-hot vectors**
- Higher the distance, higher the uncertainty score

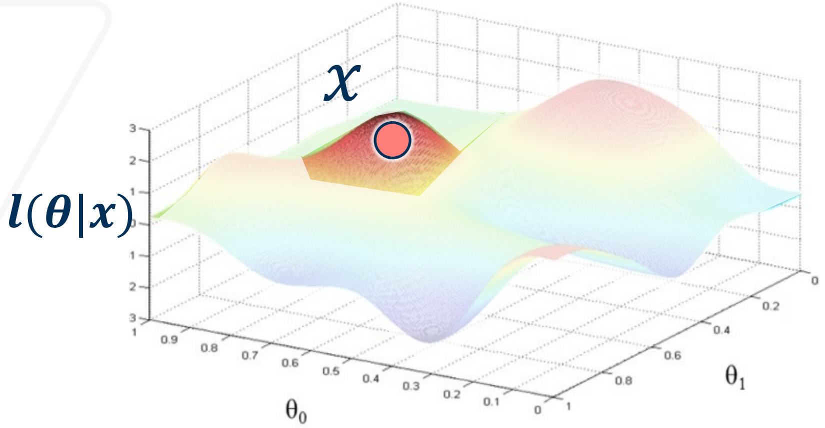
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Toy Manifold Example

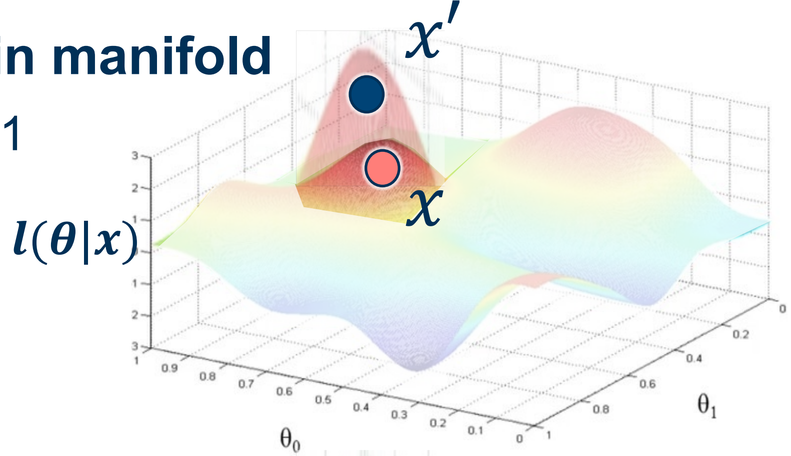
What is uncertainty?

Gradients represent the local required change in manifold

Similar to introspective learning!

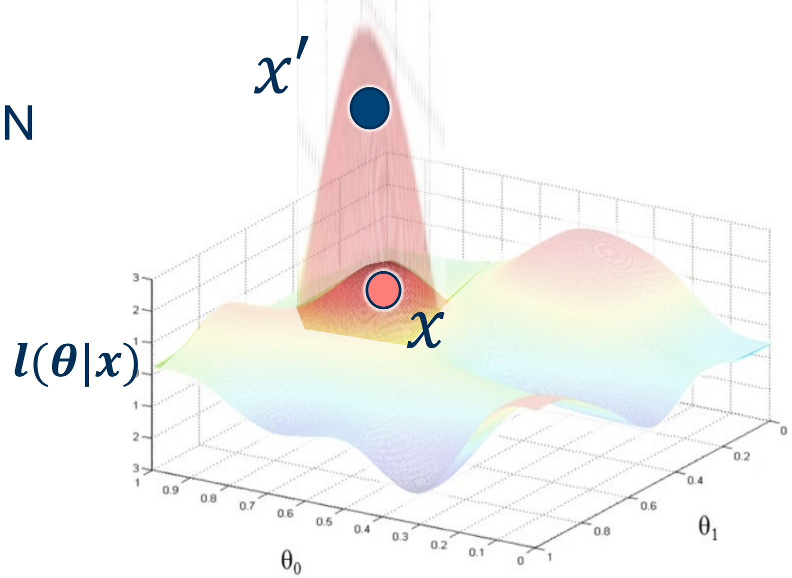


Contrast class 1



.
.
.

Contrast class N



- Gradients provide the necessary change in manifold that would predict the novel data ‘correctly’.
- Correctly means contrastively (or incorrectly)!

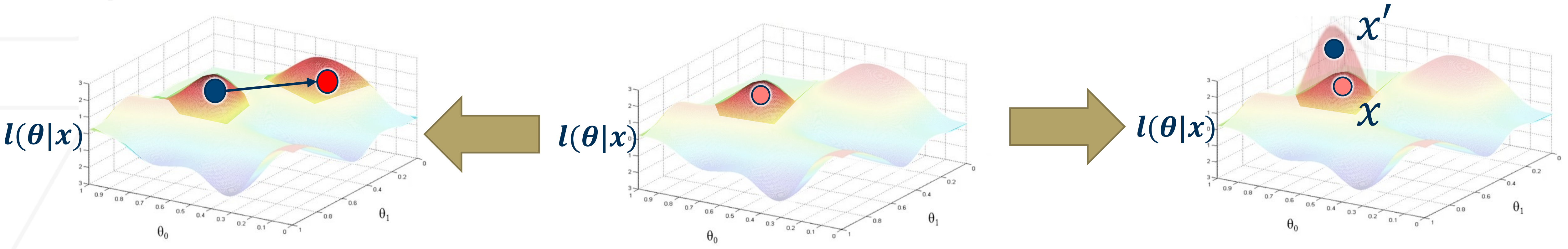
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Toy Manifold Example

How is this different from Explainability?

Part 3: Explainability

Part 4: Uncertainty



- In Part 3: Activations of learned manifold are weighted by gradients w.r.t. activations to extract information and provide explanations

- In Part 4: Statistics of gradients w.r.t. the weights (energy) will be directly used as features

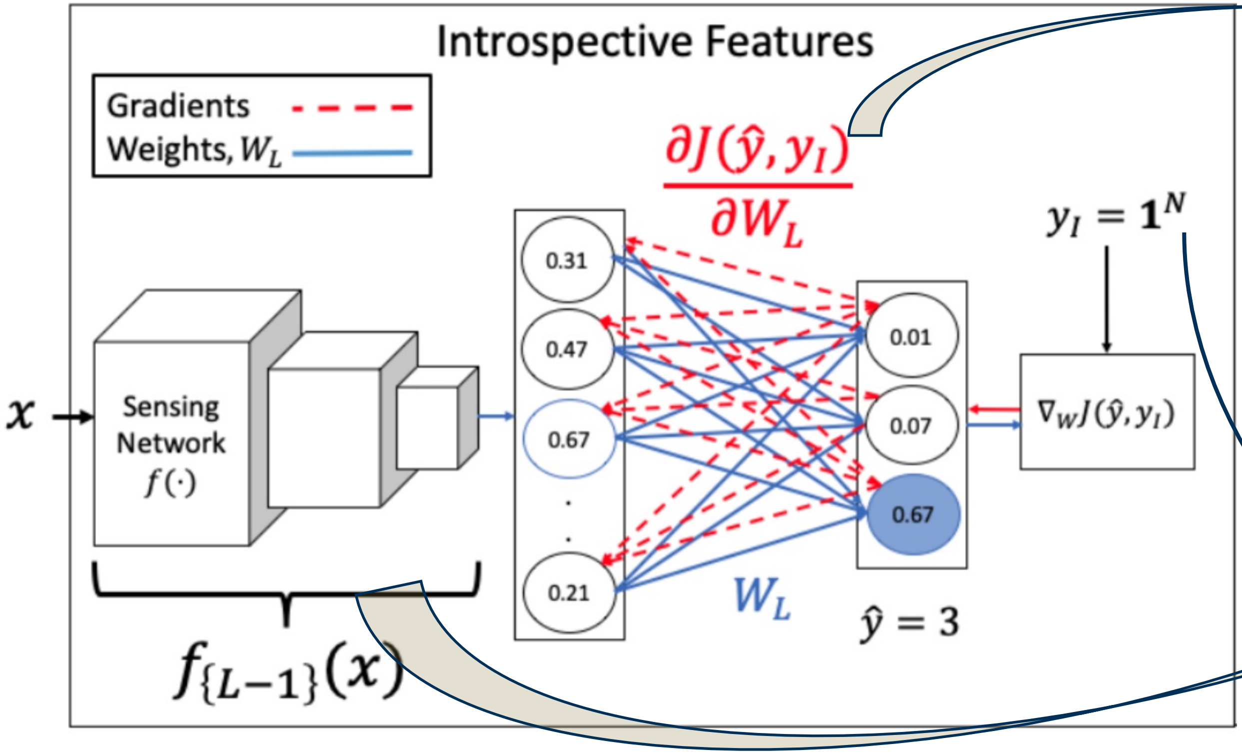
Uncertainty in Neural Networks

Deriving Gradient Features

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

Step 1: Measure the loss between the prediction P and a vector of all ones and backpropagate to obtain the introspective features



Normalized and vectorized gradients are introspective features.

Why vector of all 1s? The theory is presented in [1]

[1] M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

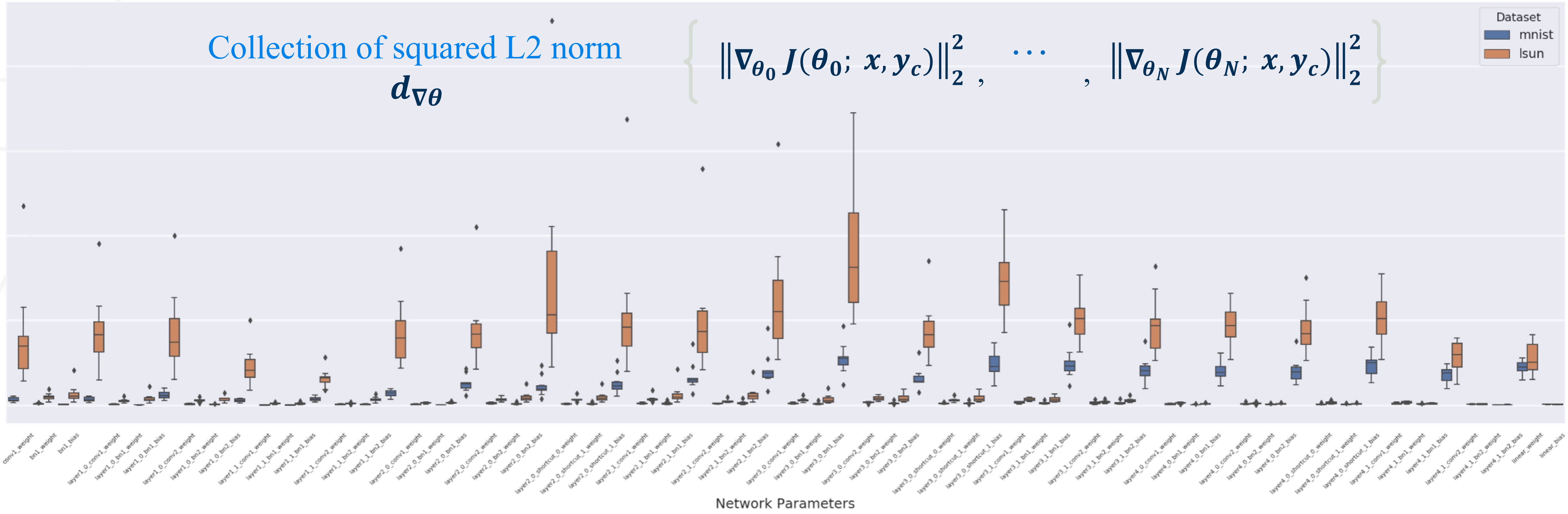
Uncertainty in Neural Networks

Utilizing Gradient Features

SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

Step 2: Take L2 norm of all generated gradients



Collection of squared L2 norm

$$d_{\nabla\theta} \quad \left\{ \left\|\nabla_{\theta_0} J(\theta_0;\ x, y_c)\right\|_2^2,\ \cdots\ ,\ \left\|\nabla_{\theta_N} J(\theta_N;\ x, y_c)\right\|_2^2 \right\}$$

MNIST: In-distribution, SUN: Out-of-Distribution

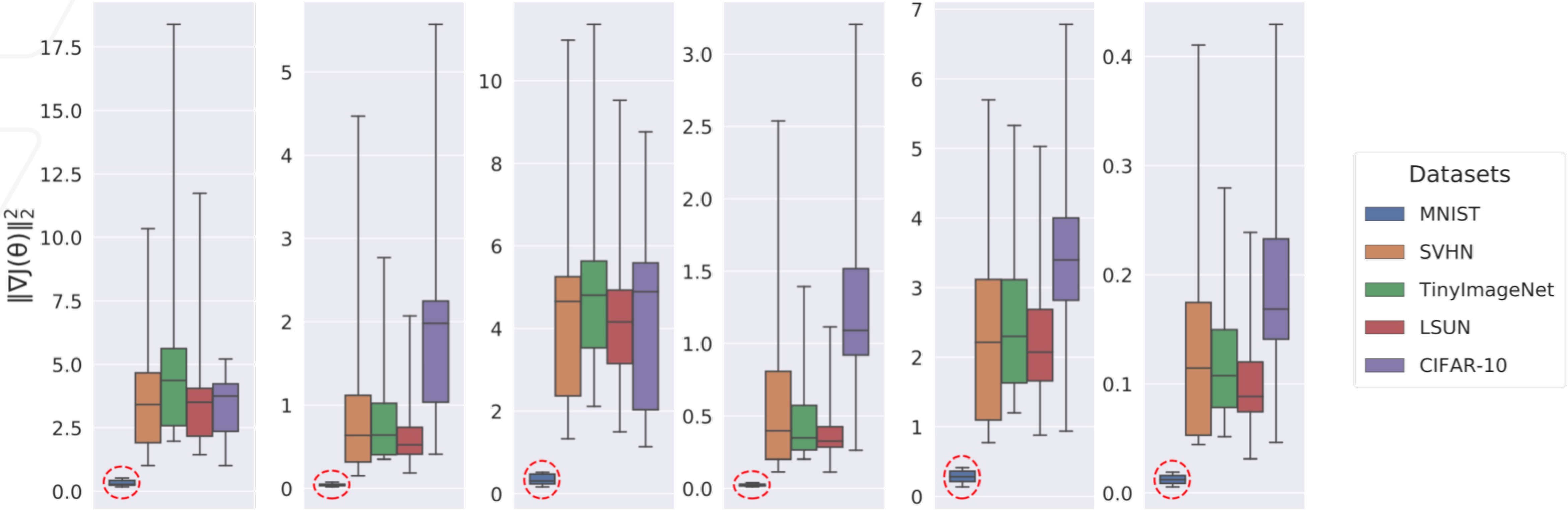
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Gradient-based Uncertainty

Uncertainty in OOD Setting

Probing the Purview of Neural Networks via Gradient Analysis

Squared L2 distances for different parameter sets



MNIST: Circled in red. Significantly lower uncertainty compared to OOD datasets

Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

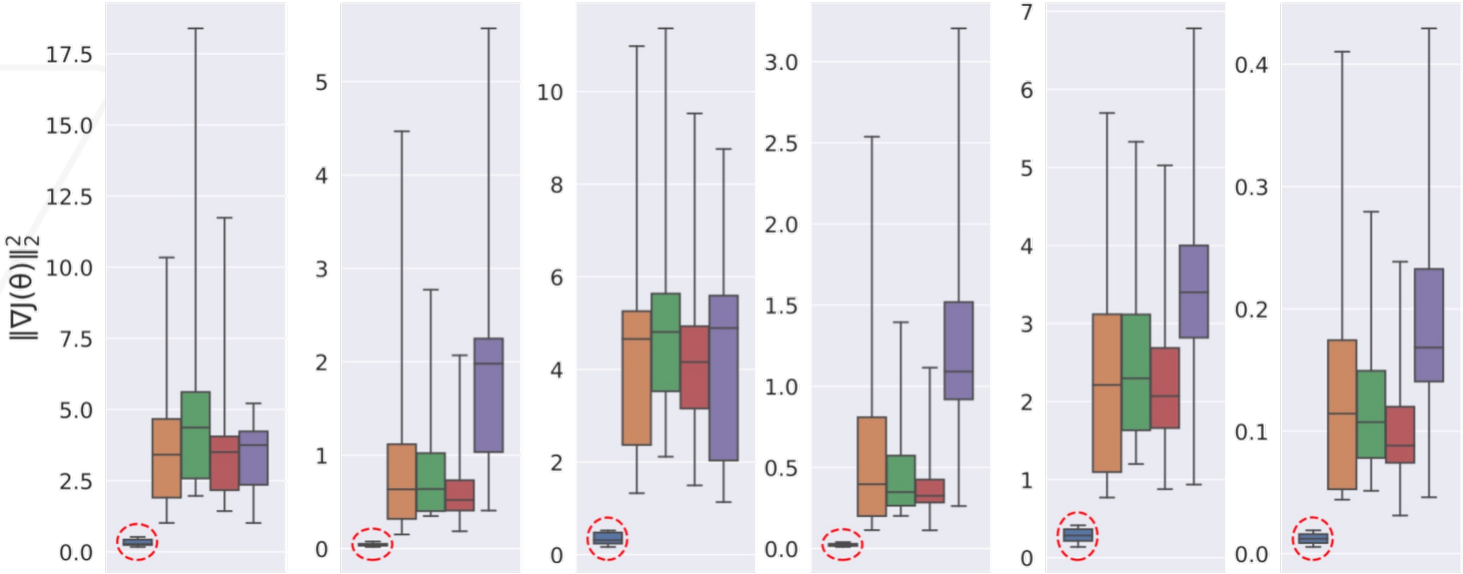
Gradient-based Uncertainty

Experimental Setup

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

Utilize this discrepancy in trained vs untrained data gradient L2 distance to detect adversarial, noisy, and OOD data



Step 1: Train a deep network $f(\cdot)$ on some **training distribution**

Step 2: Introduce challenging (adversarial, noisy, OOD) data

Step 3: Derive **gradient uncertainty** on both trained and challenge data

Step 4: Train a classifier $H(\cdot)$ to **detect** challenging from trained data

Step 5: At test time, data is passed through $f(\cdot)$ and then $H(\cdot)$ to obtain a **Reliability classification**

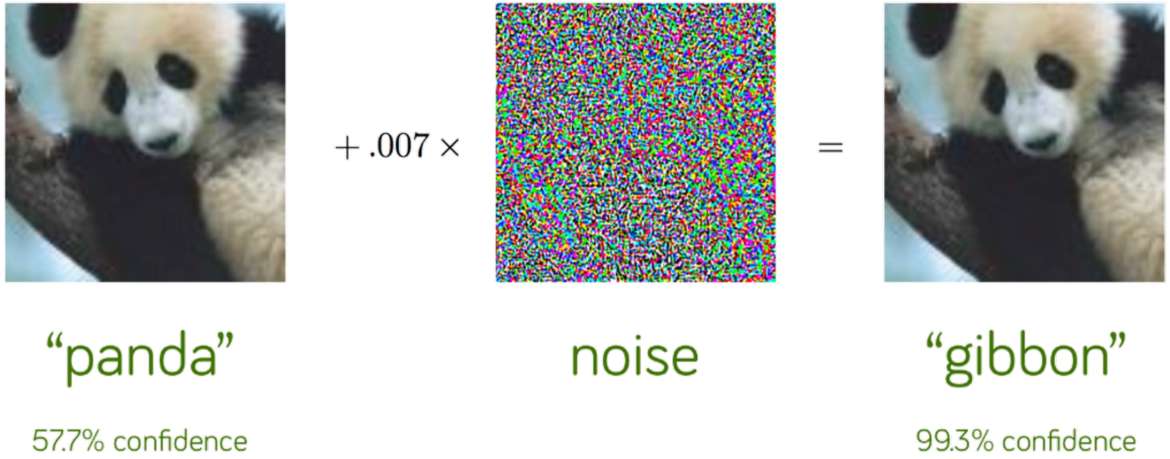
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Gradient-based Uncertainty

Uncertainty in Adversarial Setting

Probing the Purview of Neural Networks via Gradient Analysis

Vulnerable DNNs in the real world

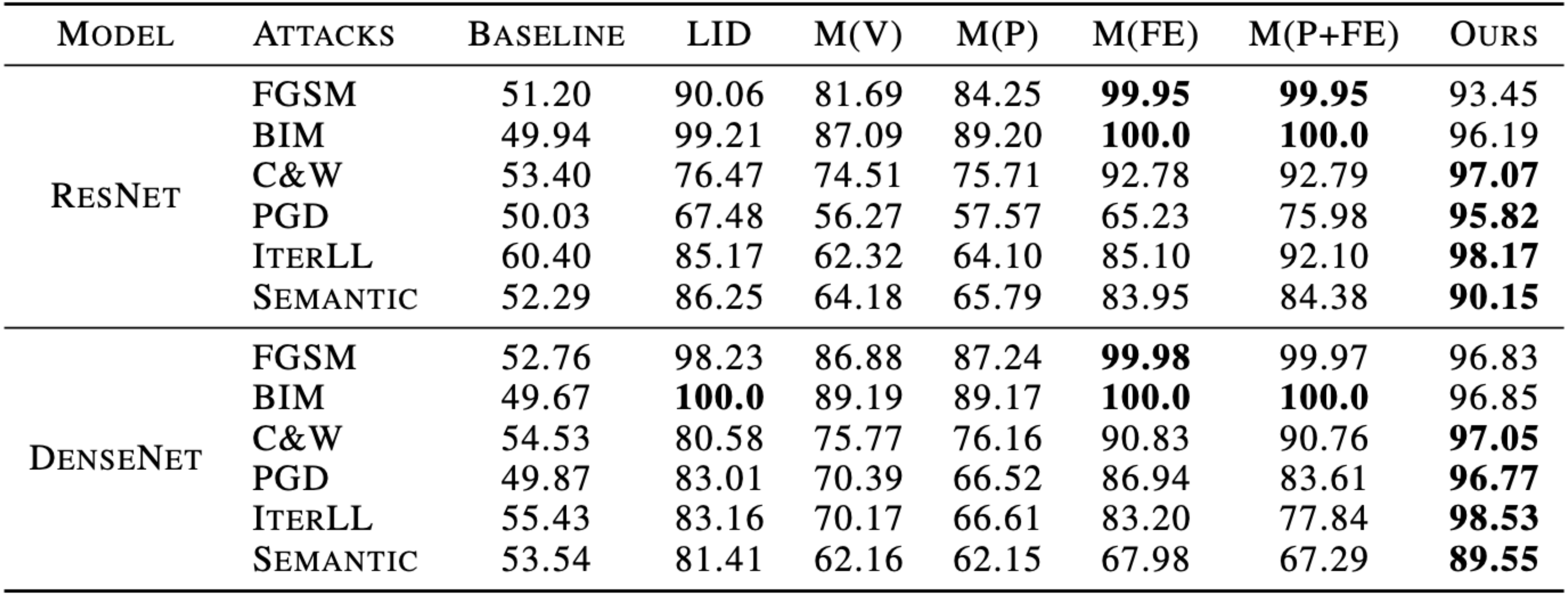


Goal: to examine the ability of trained DNNs to handle adversarial inputs during inference

Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Gradient-based Uncertainty

Uncertainty in Adversarial Setting



Model	Attacks	Baseline	LID	M(V)	M(P)	M(FE)	M(P+FE)	Ours
ResNet	FGSM	51.20	90.06	81.69	84.25	**99.95**	**99.95**	93.45
	BIM	49.94	99.21	87.09	89.20	**100.0**	**100.0**	96.19
	C&W	53.40	76.47	74.51	75.71	92.78	92.79	**97.07**
	PGD	50.03	67.48	56.27	57.57	65.23	75.98	**95.82**
	IterLL	60.40	85.17	62.32	64.10	85.10	92.10	**98.17**
	Semantic	52.29	86.25	64.18	65.79	83.95	84.38	**90.15**
DenseNet	FGSM	52.76	98.23	86.88	87.24	**99.98**	99.97	96.83
	BIM	49.67	**100.0**	89.19	89.17	**100.0**	**100.0**	96.85
	C&W	54.53	80.58	75.77	76.16	90.83	90.76	**97.05**
	PGD	49.87	83.01	70.39	66.52	86.94	83.61	**96.77**
	IterLL	55.43	83.16	70.17	66.61	83.20	77.84	**98.53**
	Semantic	53.54	81.41	62.16	62.15	67.98	67.29	**89.55**

Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

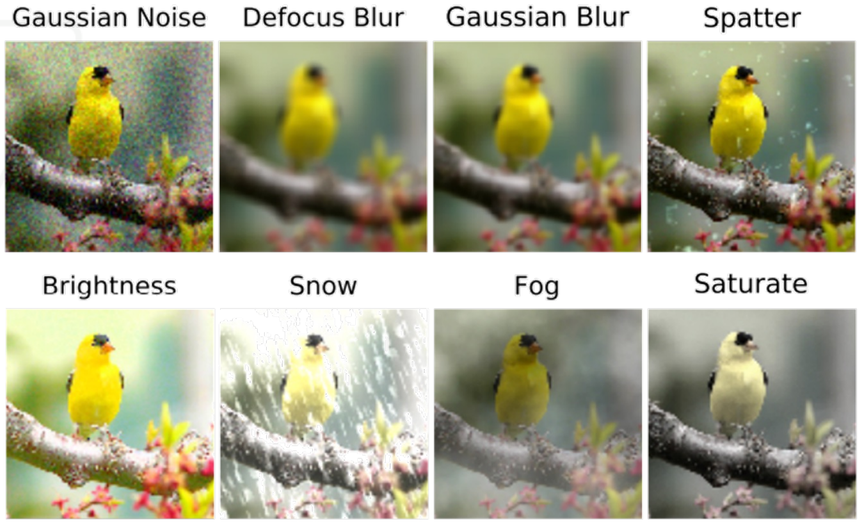
Gradient-based Uncertainty

Uncertainty in Detecting Challenging Conditions

Probing the Purview of Neural Networks via Gradient Analysis

Same application as Anomaly Detection, except there is no need for an additional AE network!

CIFAR-10-C



CURE-TSR

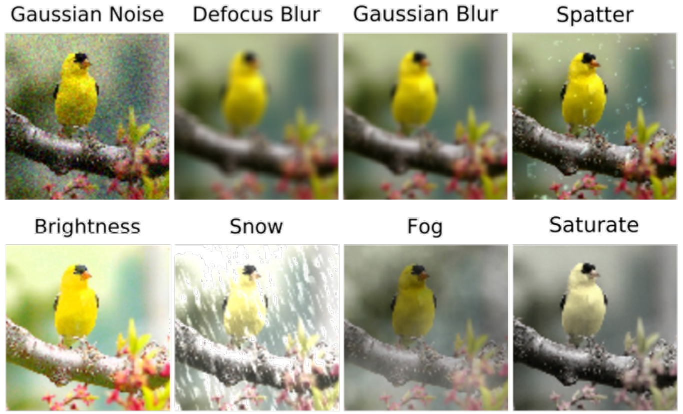
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Gradient-based Uncertainty

Uncertainty in Detecting Challenging Conditions

Probing the Purview of Neural Networks via Gradient Analysis

Dataset	Method	Mahalanobis [12] / Ours				
	Corruption	Level 1	Level 2	Level 3	Level 4	Level 5
CIFAR-10-C	Noise	96.63 / **99.95**	98.73 / **99.97**	99.46 / **99.99**	99.62 / **99.97**	99.71 / **99.99**
	LensBlur	94.22 / **99.95**	97.51 / **99.99**	99.26 / **100.0**	99.78 / **100.0**	99.89 / **100.0**
	GaussianBlur	94.19 / **99.94**	99.28 / **100.0**	99.76 / **100.0**	99.86 / **100.0**	99.80 / **100.0**
	DirtyLens	93.37 / **99.94**	95.31 / **99.93**	95.66 / **99.96**	95.37 / **99.92**	97.43 / **99.96**
	Exposure	91.39 / **99.87**	91.00 / **99.85**	90.71 / **99.88**	90.58 / **99.85**	90.68 / **99.87**
	Snow	93.64 / **99.94**	96.50 / **99.94**	94.44 / **99.95**	94.22 / **99.95**	95.25 / **99.92**
	Haze	95.52 / **99.95**	98.35 / **99.99**	99.28 / **100.0**	99.71 / **99.99**	99.94 / **100.0**
	Decolor	93.51 / **99.96**	93.55 / **99.96**	90.30 / **99.82**	89.86 / **99.75**	90.43 / **99.83**
CURE-TSR	Noise	25.46 / **50.20**	47.54 / **63.87**	47.32 / **81.20**	66.19 / **91.16**	83.14 / **94.81**
	LensBlur	48.06 / **72.63**	71.61 / **87.58**	86.59 / **92.56**	92.19 / **93.90**	94.90 / **95.65**
	GaussianBlur	66.44 / **83.07**	77.67 / **86.94**	93.15 / **94.35**	80.78 / **94.51**	**97.36** / 96.53
	DirtyLens	29.78 / **51.21**	29.28 / **59.10**	46.60 / **82.10**	73.36 / **91.87**	98.50 / **98.70**
	Exposure	74.90 / **88.13**	**99.96** / 96.78	**99.99** / 99.26	**100.0** / 99.80	**100.0** / 99.90
	Snow	28.11 / **61.34**	61.28 / **80.52**	89.89 / **91.30**	**99.34** / 96.13	**99.98** / 97.66
	Haze	66.51 / **95.83**	97.86 / **99.50**	**100.0** / 99.95	**100.0** / 99.87	**100.0** / 99.88
	Decolor	48.37 / **62.36**	60.55 / **81.30**	71.73 / **89.93**	87.29 / **95.42**	89.68 / **96.91**



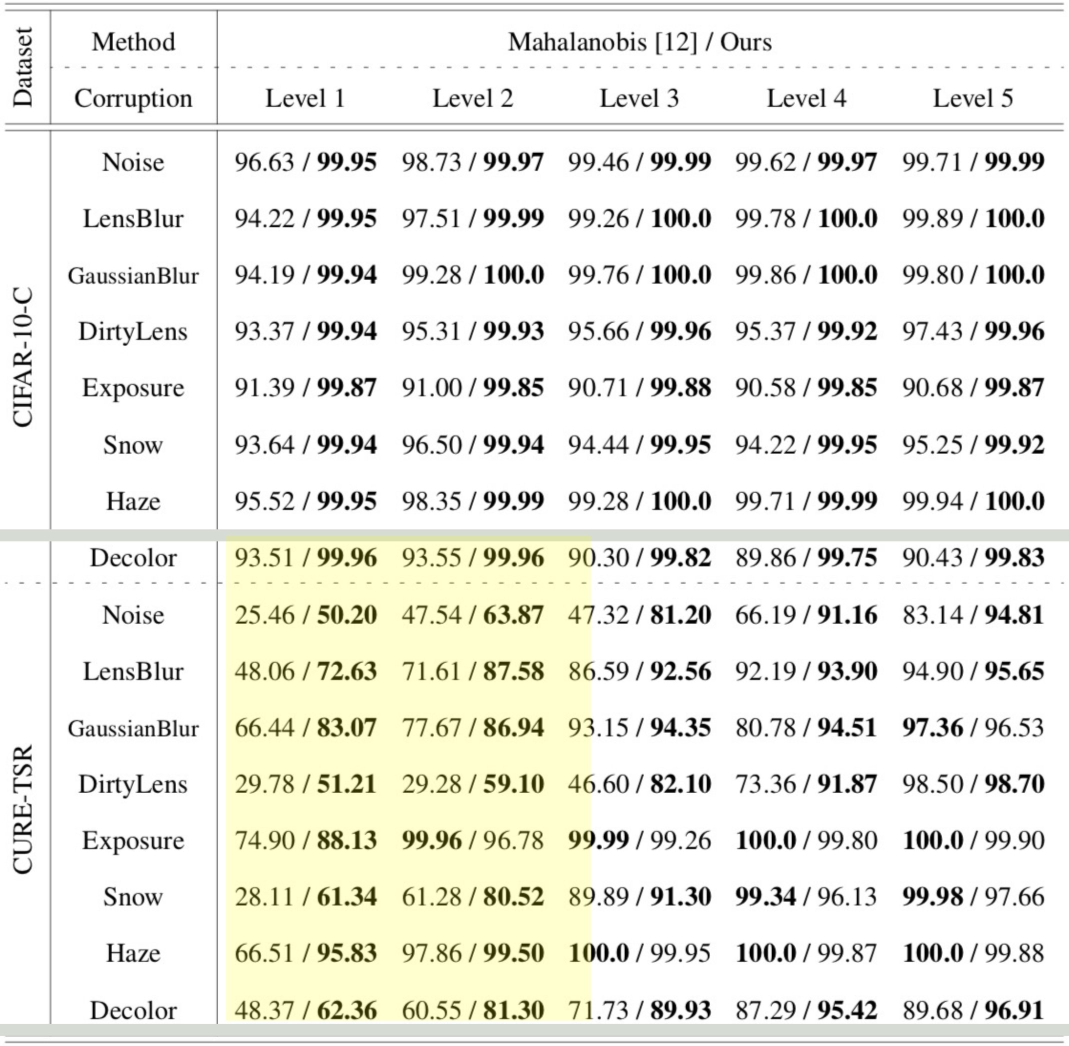
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Gradient-based Uncertainty

Uncertainty in Detecting Challenging Conditions

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME



Dataset	Method	Mahalanobis [12] / Ours				
	Corruption	Level 1	Level 2	Level 3	Level 4	Level 5
CIFAR-10-C	Noise	96.63 / **99.95**	98.73 / **99.97**	99.46 / **99.99**	99.62 / **99.97**	99.71 / **99.99**
	LensBlur	94.22 / **99.95**	97.51 / **99.99**	99.26 / **100.0**	99.78 / **100.0**	99.89 / **100.0**
	GaussianBlur	94.19 / **99.94**	99.28 / **100.0**	99.76 / **100.0**	99.86 / **100.0**	99.80 / **100.0**
	DirtyLens	93.37 / **99.94**	95.31 / **99.93**	95.66 / **99.96**	95.37 / **99.92**	97.43 / **99.96**
	Exposure	91.39 / **99.87**	91.00 / **99.85**	90.71 / **99.88**	90.58 / **99.85**	90.68 / **99.87**
	Snow	93.64 / **99.94**	96.50 / **99.94**	94.44 / **99.95**	94.22 / **99.95**	95.25 / **99.92**
	Haze	95.52 / **99.95**	98.35 / **99.99**	99.28 / **100.0**	99.71 / **99.99**	99.94 / **100.0**
	Decolor	93.51 / **99.96**	93.55 / **99.96**	90.30 / **99.82**	89.86 / **99.75**	90.43 / **99.83**
CURE-TSR	Noise	25.46 / **50.20**	47.54 / **63.87**	47.32 / **81.20**	66.19 / **91.16**	83.14 / **94.81**
	LensBlur	48.06 / **72.63**	71.61 / **87.58**	86.59 / **92.56**	92.19 / **93.90**	94.90 / **95.65**
	GaussianBlur	66.44 / **83.07**	77.67 / **86.94**	93.15 / **94.35**	80.78 / **94.51**	**97.36** / 96.53
	DirtyLens	29.78 / **51.21**	29.28 / **59.10**	46.60 / **82.10**	73.36 / **91.87**	98.50 / **98.70**
	Exposure	74.90 / **88.13**	**99.96** / 96.78	**99.99** / 99.26	**100.0** / 99.80	**100.0** / 99.90
	Snow	28.11 / **61.34**	61.28 / **80.52**	89.89 / **91.30**	**99.34** / 96.13	**99.98** / 97.66
	Haze	66.51 / **95.83**	97.86 / **99.50**	**100.0** / 99.95	**100.0** / 99.87	**100.0** / 99.88
	Decolor	48.37 / **62.36**	60.55 / **81.30**	71.73 / **89.93**	87.29 / **95.42**	89.68 / **96.91**

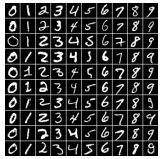
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Out-of-Distribution Detection

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

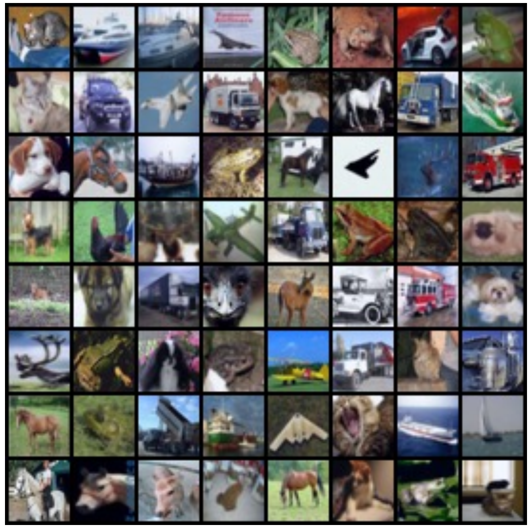
Train set →



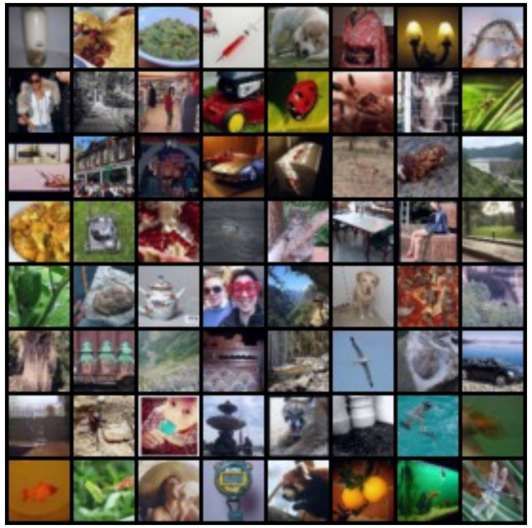
MNIST

Goal: To detect that these datasets are not part of training

SVHN



CIFAR10

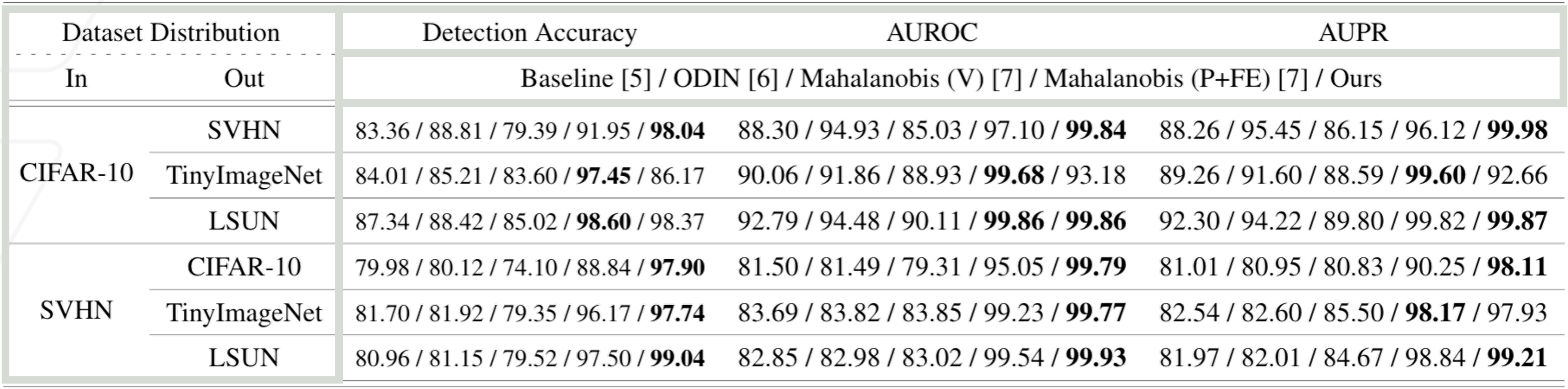


TinyImageNet

LSUN

Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Out-of-Distribution Detection



Dataset Distribution		Detection Accuracy	AUROC	AUPR
In	Out	Baseline [5] / ODIN [6] / Mahalanobis (V) [7] / Mahalanobis (P+FE) [7] / Ours		
CIFAR-10	SVHN	83.36 / 88.81 / 79.39 / 91.95 / **98.04**	88.30 / 94.93 / 85.03 / 97.10 / **99.84**	88.26 / 95.45 / 86.15 / 96.12 / **99.98**
	TinyImageNet	84.01 / 85.21 / 83.60 / **97.45** / 86.17	90.06 / 91.86 / 88.93 / **99.68** / 93.18	89.26 / 91.60 / 88.59 / **99.60** / 92.66
	LSUN	87.34 / 88.42 / 85.02 / **98.60** / 98.37	92.79 / 94.48 / 90.11 / **99.86** / **99.86**	92.30 / 94.22 / 89.80 / 99.82 / **99.87**
SVHN	CIFAR-10	79.98 / 80.12 / 74.10 / 88.84 / **97.90**	81.50 / 81.49 / 79.31 / 95.05 / **99.79**	81.01 / 80.95 / 80.83 / 90.25 / **98.11**
	TinyImageNet	81.70 / 81.92 / 79.35 / 96.17 / **97.74**	83.69 / 83.82 / 83.85 / 99.23 / **99.77**	82.54 / 82.60 / 85.50 / **98.17** / 97.93
	LSUN	80.96 / 81.15 / 79.52 / 97.50 / **99.04**	82.85 / 82.98 / 83.02 / 99.54 / **99.93**	81.97 / 82.01 / 84.67 / 98.84 / **99.21**

Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Out-of-Distribution Detection

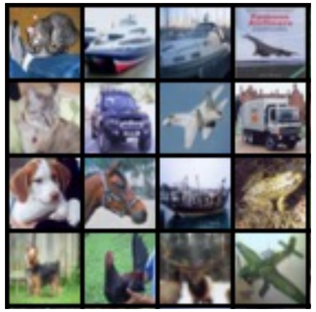
Probing the Purview of Neural Networks via Gradient Analysis

Dataset Distribution		Detection Accuracy	AUROC	AUPR
In	Out	Baseline [5] / ODIN [6] / Mahalanobis (V) [7] / Mahalanobis (P+FE) [7] / Ours		
CIFAR-10	SVHN	83.36 / 88.81 / 79.39 / 91.95 / **98.04**	88.30 / 94.93 / 85.03 / 97.10 / **99.84**	88.26 / 95.45 / 86.15 / 96.12 / **99.98**
	TinyImageNet	84.01 / 85.21 / 83.60 / **97.45** / 86.17	90.06 / 91.86 / 88.93 / **99.68** / 93.18	89.26 / 91.60 / 88.59 / **99.60** / 92.66
	LSUN	87.34 / 88.42 / 85.02 / **98.60** / 98.37	92.79 / 94.48 / 90.11 / **99.86** / **99.86**	92.30 / 94.22 / 89.80 / 99.82 / **99.87**
SVHN	CIFAR-10	79.98 / 80.12 / 74.10 / 88.84 / **97.90**	81.50 / 81.49 / 79.31 / 95.05 / **99.79**	81.01 / 80.95 / 80.83 / 90.25 / **98.11**
	TinyImageNet	81.70 / 81.92 / 79.35 / 96.17 / **97.74**	83.69 / 83.82 / 83.85 / 99.23 / **99.77**	82.54 / 82.60 / 85.50 / **98.17** / 97.93
	LSUN	80.96 / 81.15 / 79.52 / 97.50 / **99.04**	82.85 / 82.98 / 83.02 / 99.54 / **99.93**	81.97 / 82.01 / 84.67 / 98.84 / **99.21**

Numbers

SVHN

↔



CIFAR10

TinyImageNet

LSUN

Objects, natural scenes

Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Out-of-Distribution Detection

SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

Dataset Distribution		Detection Accuracy	AUROC	AUPR
In	Out	Baseline [5] / ODIN [6] / Mahalanobis (V) [7] / Mahalanobis (P+FE) [7] / Ours		
CIFAR-10	SVHN	83.36 / 88.81 / 79.39 / 91.95 / **98.04**	88.30 / 94.93 / 85.03 / 97.10 / **99.84**	88.26 / 95.45 / 86.15 / 96.12 / **99.98**
	TinyImageNet	84.01 / 85.21 / 83.60 / **97.45** / 86.17	90.06 / 91.86 / 88.93 / **99.68** / 93.18	89.26 / 91.60 / 88.59 / **99.60** / 92.66
	LSUN	87.34 / 88.42 / 85.02 / **98.60** / 98.37	92.79 / 94.48 / 90.11 / **99.86** / **99.86**	92.30 / 94.22 / 89.80 / 99.82 / **99.87**
SVHN	CIFAR-10	79.98 / 80.12 / 74.10 / 88.84 / **97.90**	81.50 / 81.49 / 79.31 / 95.05 / **99.79**	81.01 / 80.95 / 80.83 / 90.25 / **98.11**
	TinyImageNet	81.70 / 81.92 / 79.35 / 96.17 / **97.74**	83.69 / 83.82 / 83.85 / 99.23 / **99.77**	82.54 / 82.60 / 85.50 / **98.17** / 97.93
	LSUN	80.96 / 81.15 / 79.52 / 97.50 / **99.04**	82.85 / 82.98 / 83.02 / 99.54 / **99.93**	81.97 / 82.01 / 84.67 / 98.84 / **99.21**

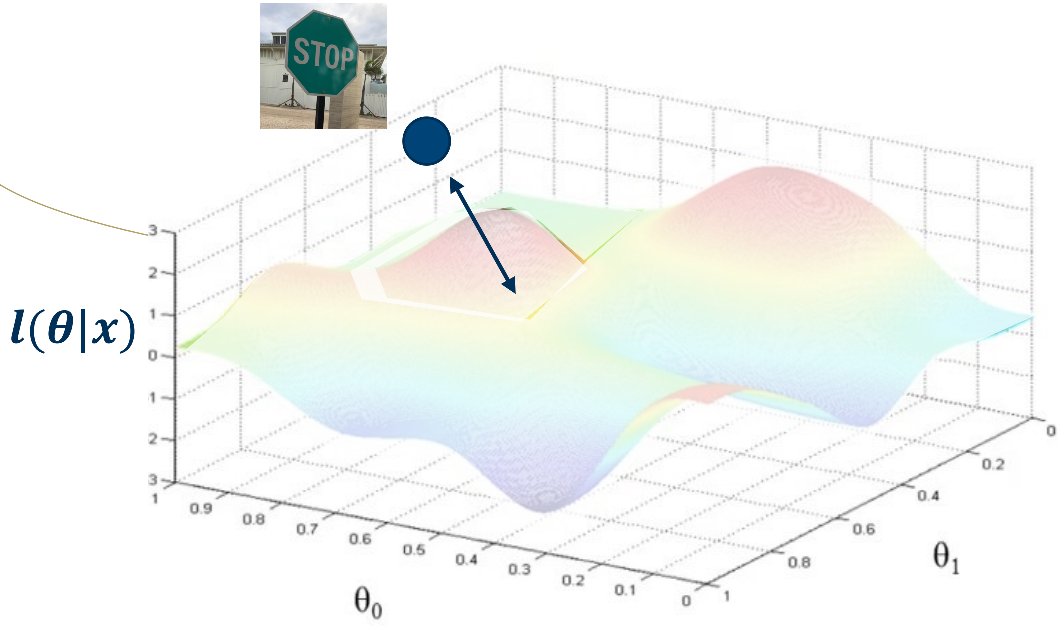
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Case Study: Introspective Learning

Gradients as Single pass Features

Our Goal: Use gradients to characterize the novel data at Inference, without global information

Distance from unknown cluster



Two techniques:

1. Gradient constraints during Training for Anomaly Detection
2. **Backpropagating Confounding labels for Out-of-Distribution Detection**

NEURAL INFORMATION
PROCESSING SYSTEMS

Introspective Learning: A Two-Stage Approach for Inference in Neural Networks

Mohit Prabhushankar, PhD
Postdoc

Ghassan AlRegib, PhD
Professor

SCAN ME

Robustness in Neural Networks

Why Robustness?

Introspective Learning: A Two-stage Approach for Inference in Neural Networks

SCAN ME

How would humans resolve this challenge?

We Introspect!

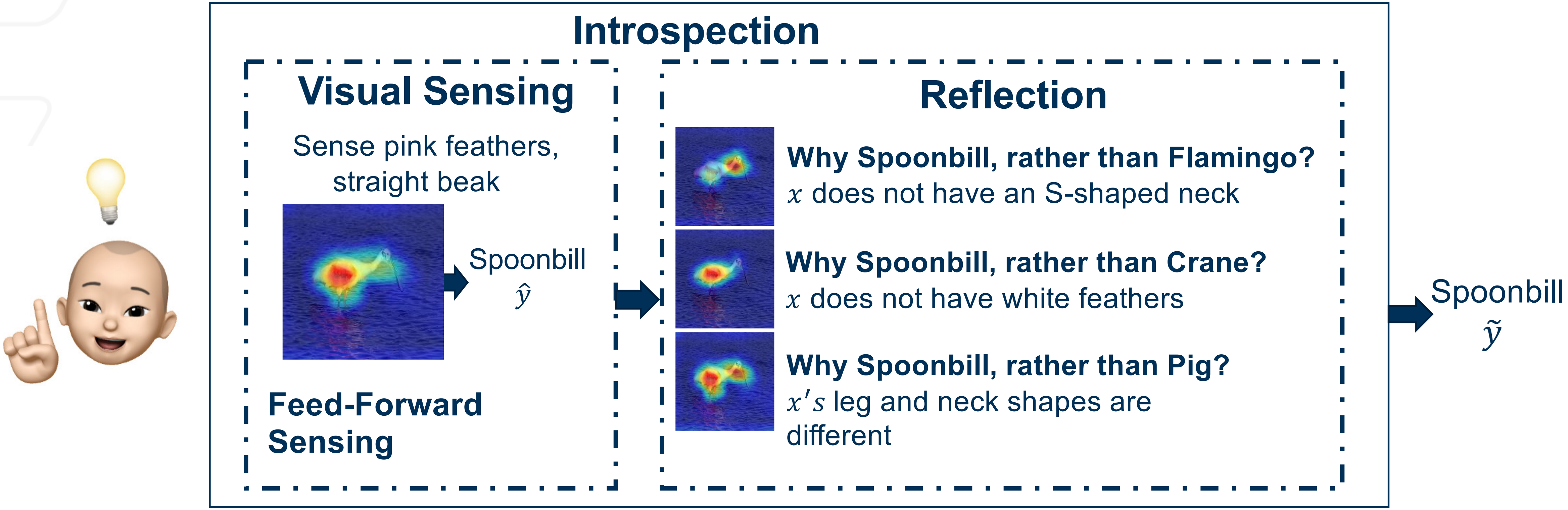
- Why am I being shown this slide?
- Why images of muffins rather than pastries?
- What if the dog was a bull mastiff?

Introspection

What is Introspection?

Introspective Learning: A Two-stage Approach for Inference in Neural Networks

Introspection Learning is a two-stage approach for Inference that combines visual sensing and reflection



Why Spoonbill, rather than Flamingo?
x does not have an S-shaped neck

Why Spoonbill, rather than Crane?
x does not have white feathers

Why Spoonbill, rather than Pig?
$x's$ leg and neck shapes are different

M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

SCAN ME

Introspective Learning: A Two-stage Approach for Inference in Neural Networks

Introspection Learning is a two-stage approach for Inference that combines visual sensing and reflection

Goal : To simulate Introspection in Neural Networks

Definition : *We define introspections as answers to logical and targeted questions.*

What are the possible targeted questions?

M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Introspection

Introspection in Neural Networks

SCAN ME

Introspective Learning: A Two-stage Approach for Inference in Neural Networks

Introspection Learning is a two-stage approach for Inference that combines visual sensing and reflection

Bullmastiff	Why Bullmastiff?	What if Bullmastiff was not in the image?	Why Bullmastiff, rather than a Boxer?

What are the possible targeted questions?

M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Introspection

Introspection in Neural Networks

Introspective Learning: A Two-stage Approach for Inference in Neural Networks

SCAN ME

Introspection Learning is a two-stage approach for Inference that combines visual sensing and reflection

Goal : To simulate Introspection in Neural Networks

Contrastive Definition : *Introspection answers questions of the form `Why P, rather than Q?' where P is a network prediction and Q is the introspective class.*

Technical Definition : *Given a network* $f(x)$, *a datum* x, *and the network's prediction* $f(x) = \hat{y}$, *introspection in* $f(\cdot)$ *is the measurement of change induced in the network parameters*
when a label Q is introduced as the label for x..

M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

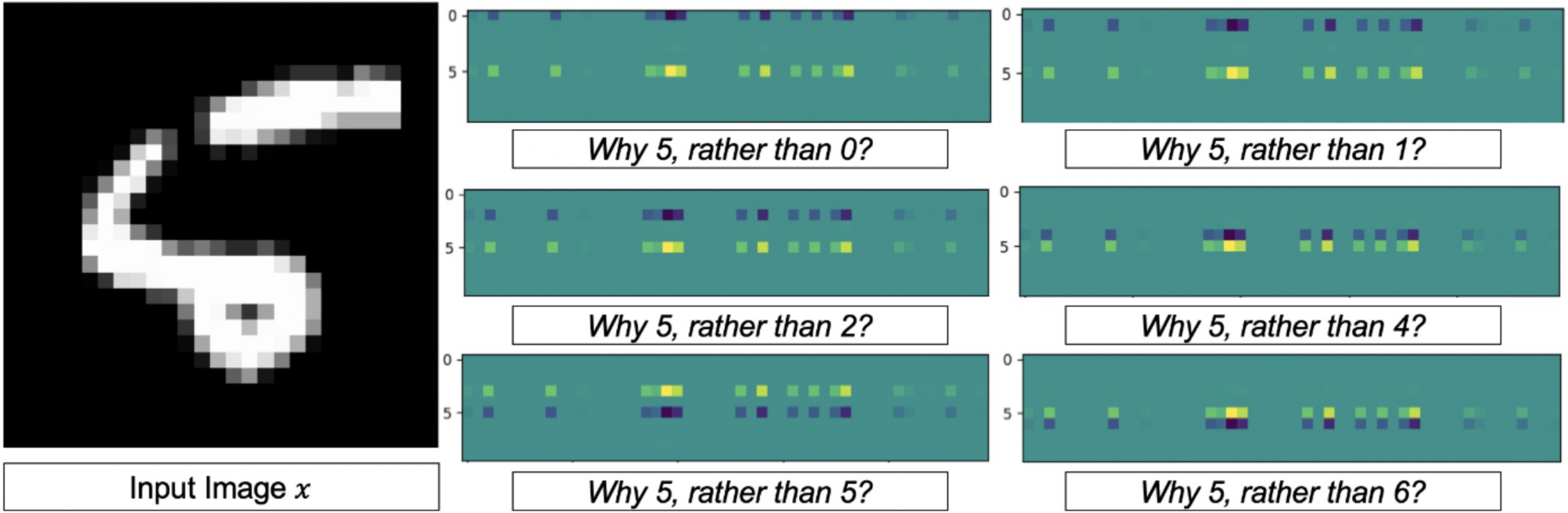
Introspection

Gradients as Features

SCAN ME

Introspective Learning: A Two-stage Approach for Inference in Neural Networks

For a well-trained network, the gradients are sparse and informative



M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

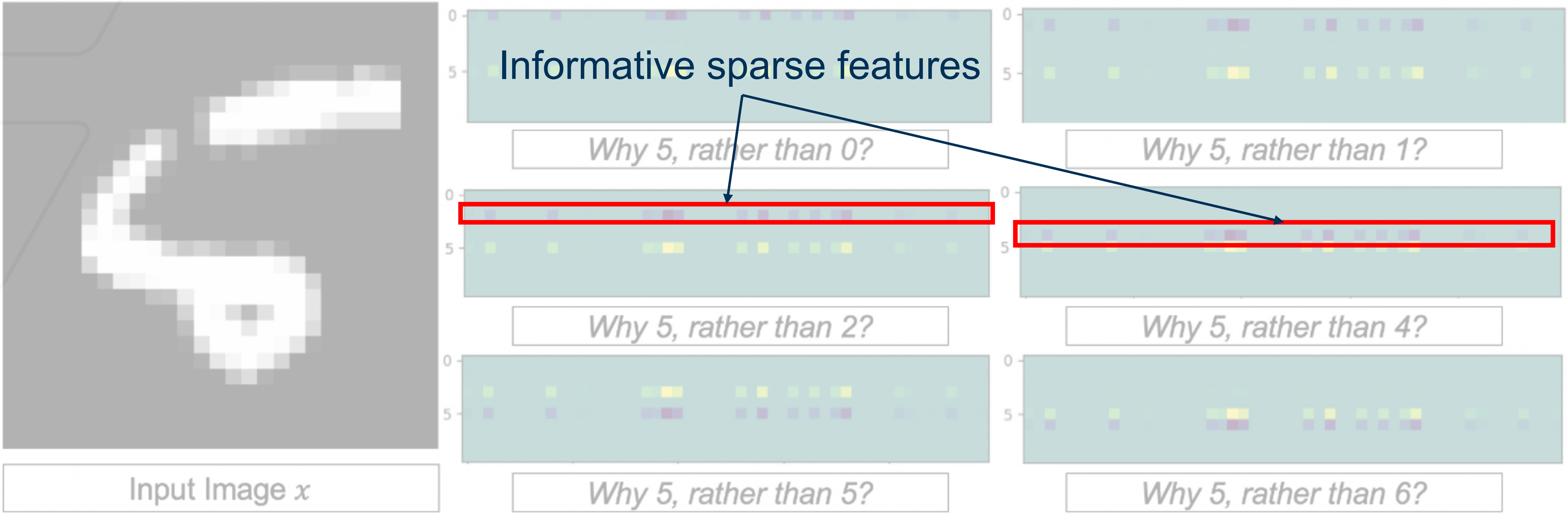
Introspection

Gradients as Features

Introspective Learning: A Two-stage Approach for Inference in Neural Networks

SCAN ME

For a well-trained network, the gradients are sparse and informative



M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Introspection

Gradients as Features

Introspective Learning: A Two-stage Approach for Inference in Neural Networks

For a well-trained network, the gradients are robust

∇_W = Gradients w.r.t. weights

J = Loss function

$\hat{y}$ = Prediction

$$\overbrace{\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \cdots \quad \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}}^{y_I}$$

Lemma1: $\nabla_W J(y_I, \hat{y}) = -\nabla_W y_I + \nabla_W \log\left(1 + \frac{y_I \hat{y}}{2}\right).$

Any change in class requires change in relationship between y_I and $\hat{y}$

M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

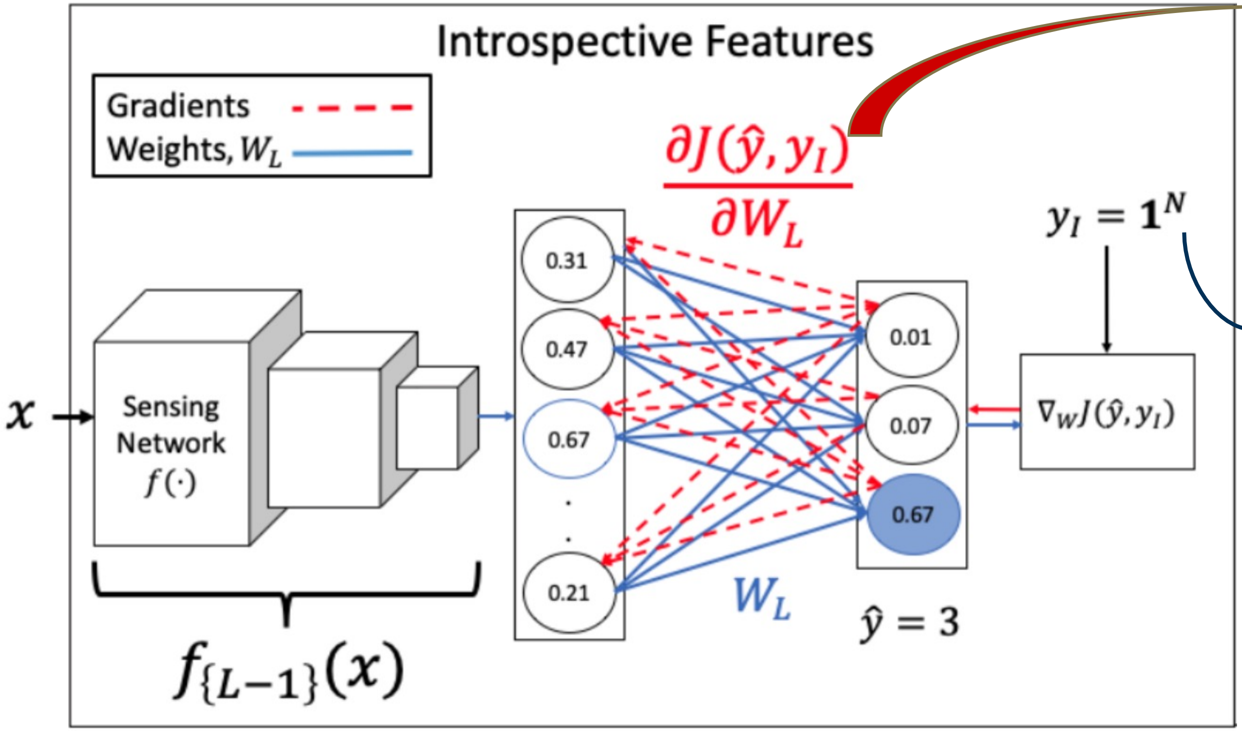
Introspection

Deriving Gradient Features

Introspective Learning: A Two-stage Approach for Inference in Neural Networks

SCAN ME

Measure the loss between the prediction P and a vector of all ones and backpropagate to obtain the introspective features



Normalized and vectorized gradients are introspective features

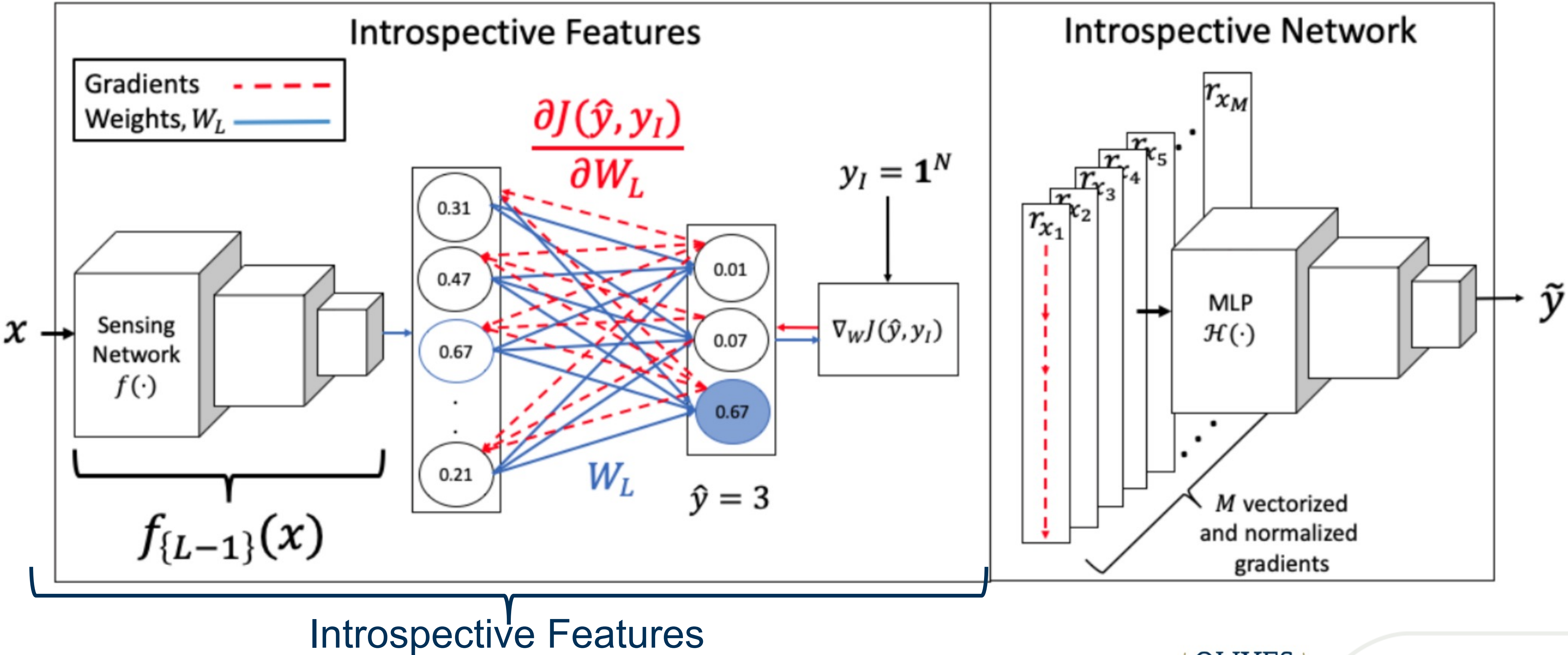
Vector of all ones: A confounding label!

M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Introspection

Utilizing Gradient Features

Introspective Learning: A Two-stage Approach for Inference in Neural Networks



M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Introspection

When is Introspection Useful?

Introspective Learning: A Two-stage Approach for Inference in Neural Networks

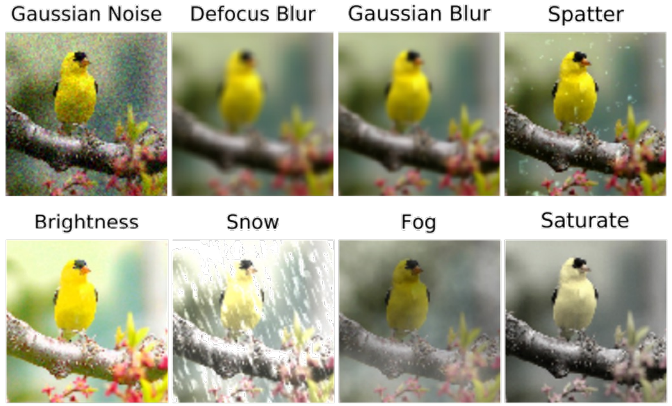
SCAN ME

Introspection provides robustness when the train and test distributions are different

We define robustness as being generalizable and calibrated to new testing data

Generalizable: Increased accuracy on OOD data

Calibrated: Reduces the difference between prediction accuracy and confidence

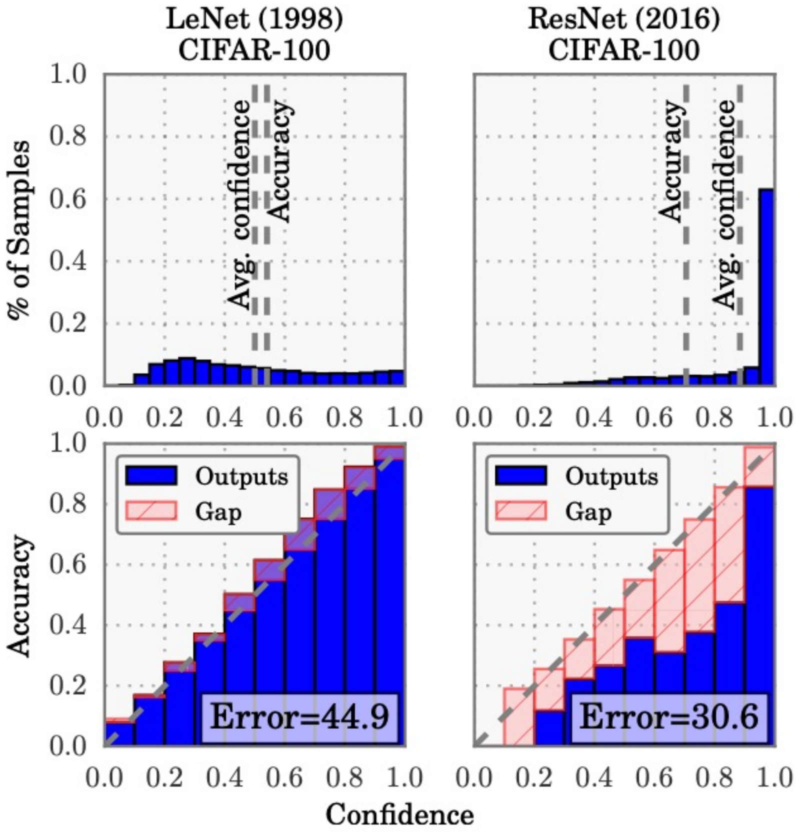


M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Calibration

A note on Calibration..

Calibration occurs when there is mismatch between a network's confidence and its accuracy



- Larger the model, more misplaced is a network's confidence
- On ResNet, the gap between prediction accuracy and its corresponding confidence is significantly high

M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Introspection in Neural Networks

Generalization and Calibration results

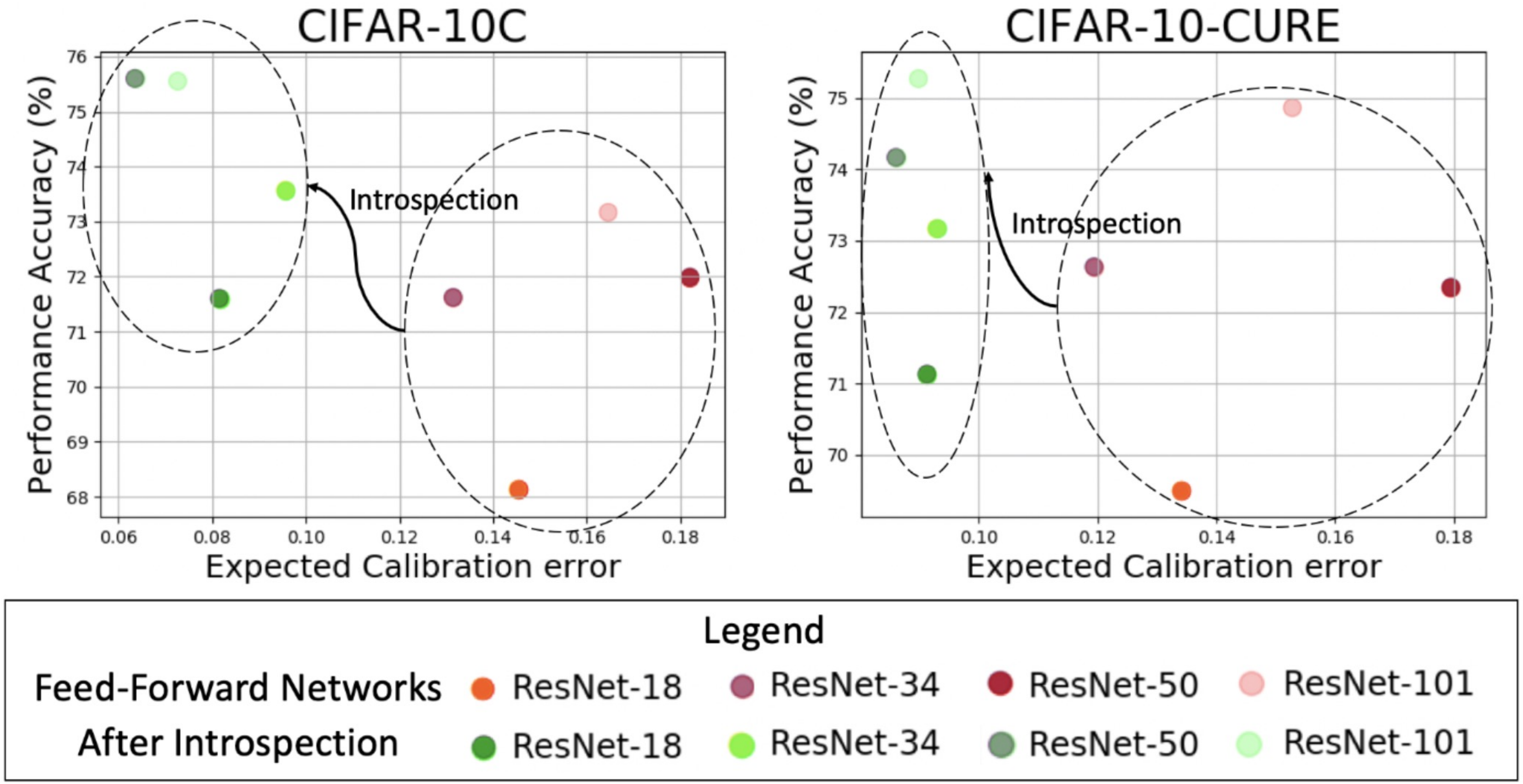
SCAN ME

Introspective Learning: A Two-stage Approach for Inference in Neural Networks

Ideal: Top-left corner

Y-Axis: Generalization

X-Axis: Calibration



M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Introspection in Neural Networks

Plug-in nature of Introspection

Introspective Learning: A Two-stage Approach for Inference in Neural Networks

SCAN ME

Introspection is a light-weight option to resolve robustness issues

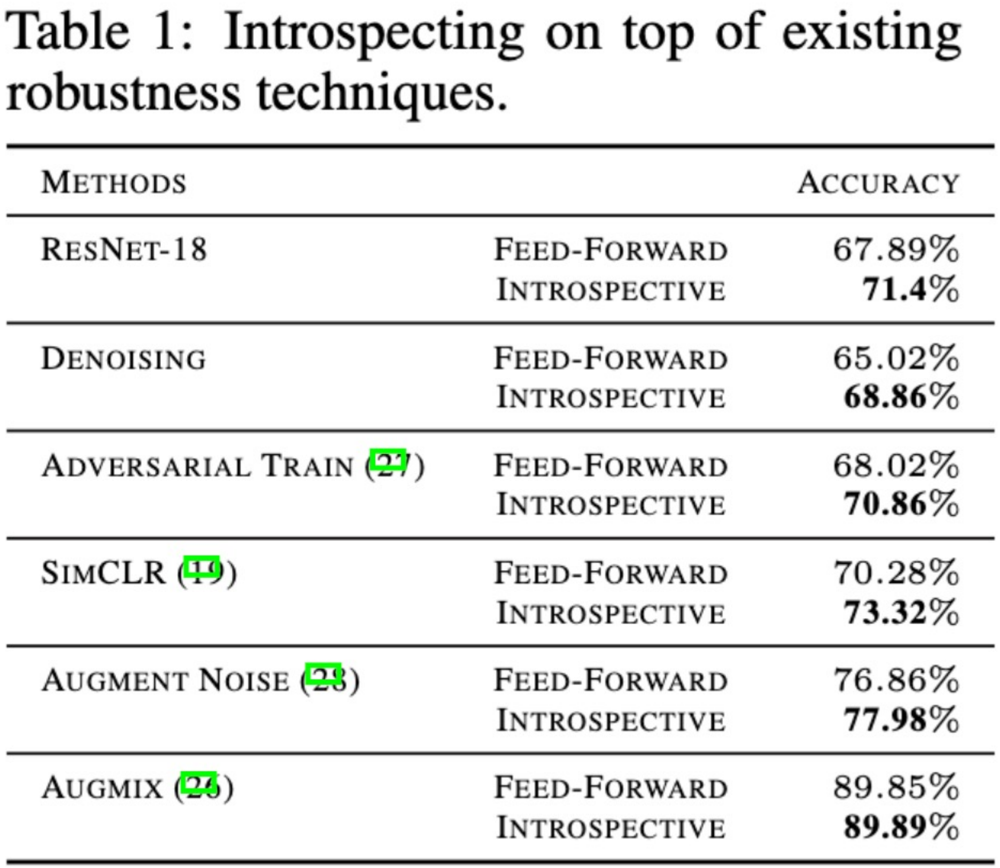


Table 1: Introspecting on top of existing robustness techniques.

METHODS		ACCURACY
RESNET-18	FEED-FORWARD INTROSPECTIVE	67.89% **71.4%**
DENOISING	FEED-FORWARD INTROSPECTIVE	65.02% **68.86%**
ADVERSARIAL TRAIN (27)	FEED-FORWARD INTROSPECTIVE	68.02% **70.86%**
SIMCLR (19)	FEED-FORWARD INTROSPECTIVE	70.28% **73.32%**
AUGMENT NOISE (28)	FEED-FORWARD INTROSPECTIVE	76.86% **77.98%**
AUGMIX (26)	FEED-FORWARD INTROSPECTIVE	89.85% **89.89%**

Introspection is a **plug-in approach** that works on all networks and on any downstream task!

M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Introspection in Neural Networks

Plug-in nature of Introspection

Introspective Learning: A Two-stage Approach for Inference in Neural Networks

Plug-in nature of Introspection benefits downstream tasks like OOD detection, Active Learning, and Image Quality Assessment!

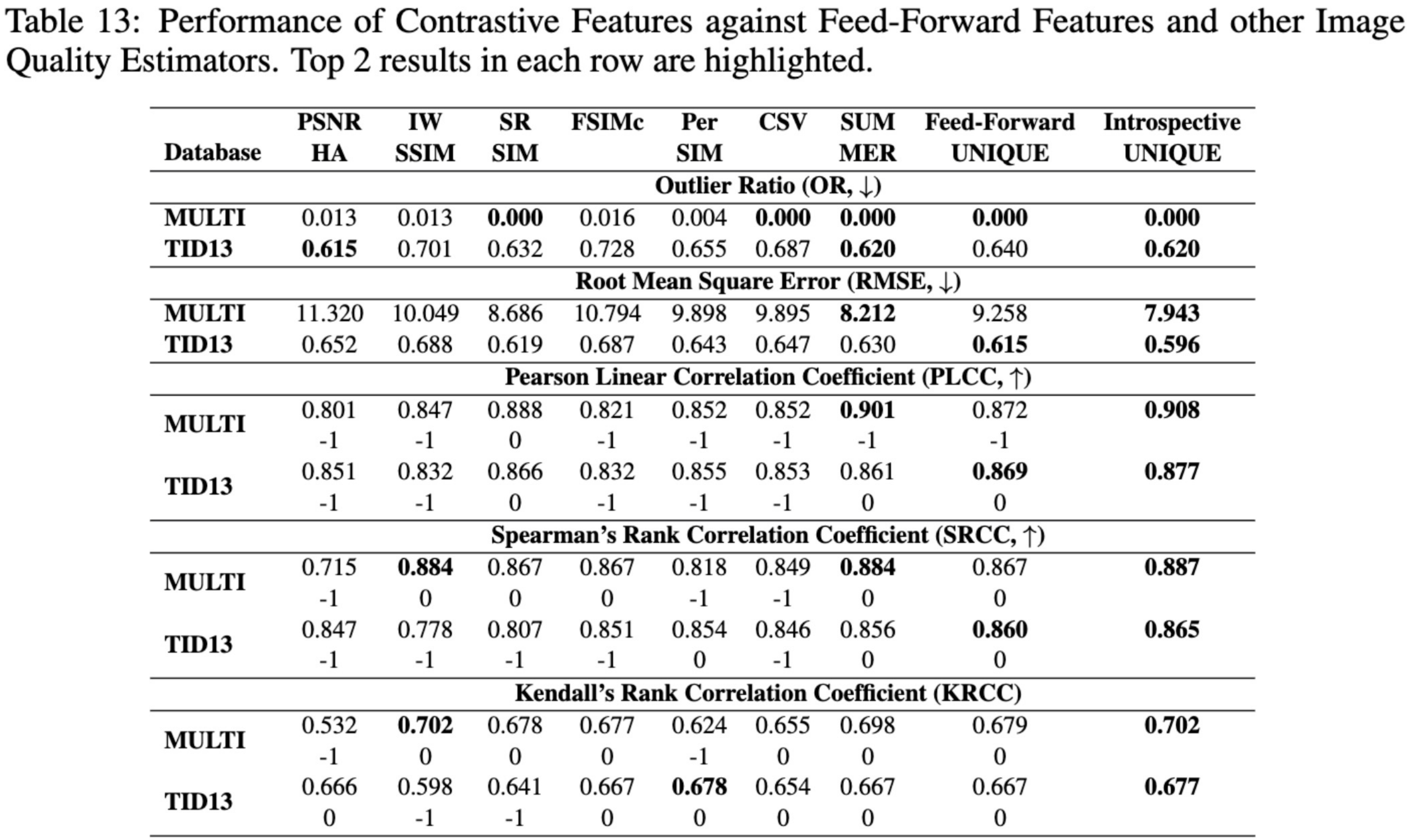


Table 13: Performance of Contrastive Features against Feed-Forward Features and other Image Quality Estimators. Top 2 results in each row are highlighted.

Database	PSNR HA	IW SSIM	SR SIM	FSIMc	Per SIM	CSV	SUM MER	Feed-Forward UNIQUE	Introspective UNIQUE
Outlier Ratio (OR, ↓)									
MULTI	0.013	0.013	**0.000**	0.016	0.004	**0.000**	**0.000**	**0.000**	**0.000**
TID13	**0.615**	0.701	0.632	0.728	0.655	0.687	**0.620**	0.640	**0.620**
Root Mean Square Error (RMSE, ↓)									
MULTI	11.320	10.049	8.686	10.794	9.898	9.895	**8.212**	9.258	**7.943**
TID13	0.652	0.688	0.619	0.687	0.643	0.647	0.630	**0.615**	**0.596**
Pearson Linear Correlation Coefficient (PLCC, ↑)									
MULTI	0.801	0.847	0.888	0.821	0.852	0.852	**0.901**	0.872	**0.908**
	-1	-1	0	-1	-1	-1	-1	-1	
TID13	0.851	0.832	0.866	0.832	0.855	0.853	0.861	**0.869**	**0.877**
	-1	-1	0	-1	-1	-1	0	0	
Spearman's Rank Correlation Coefficient (SRCC, ↑)									
MULTI	0.715	**0.884**	0.867	0.867	0.818	0.849	**0.884**	0.867	**0.887**
	-1	0	0	0	-1	-1	0	0	
TID13	0.847	0.778	0.807	0.851	0.854	0.846	0.856	**0.860**	**0.865**
	-1	-1	-1	-1	0	-1	0	0	
Kendall's Rank Correlation Coefficient (KRCC)									
MULTI	0.532	**0.702**	0.678	0.677	0.624	0.655	0.698	0.679	**0.702**
	-1	0	0	0	-1	0	0	0	
TID13	0.666	0.598	0.641	0.667	**0.678**	0.654	0.667	0.667	**0.677**
	0	-1	-1	0	0	0	0	0	

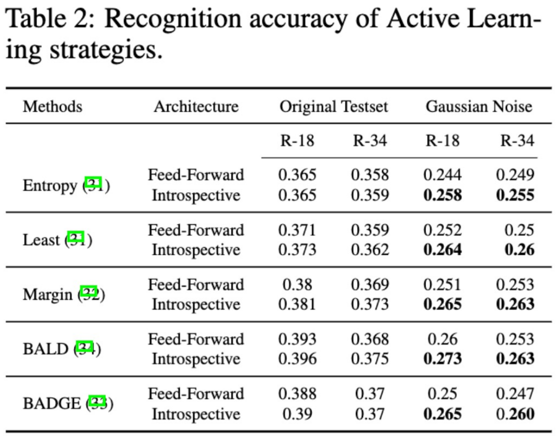


Table 2: Recognition accuracy of Active Learning strategies.

Methods	Architecture	Original Testset		Gaussian Noise	
		R-18	R-34	R-18	R-34
Entropy (31)	Feed-Forward	0.365	0.358	0.244	0.249
	Introspective	0.365	0.359	**0.258**	**0.255**
Least (31)	Feed-Forward	0.371	0.359	0.252	0.25
	Introspective	0.373	0.362	**0.264**	**0.26**
Margin (32)	Feed-Forward	0.38	0.369	0.251	0.253
	Introspective	0.381	0.373	**0.265**	**0.263**
BALD (34)	Feed-Forward	0.393	0.368	0.26	0.253
	Introspective	0.396	0.375	**0.273**	**0.263**
BADGE (33)	Feed-Forward	0.388	0.37	0.25	0.247
	Introspective	0.39	0.37	**0.265**	0.**260**

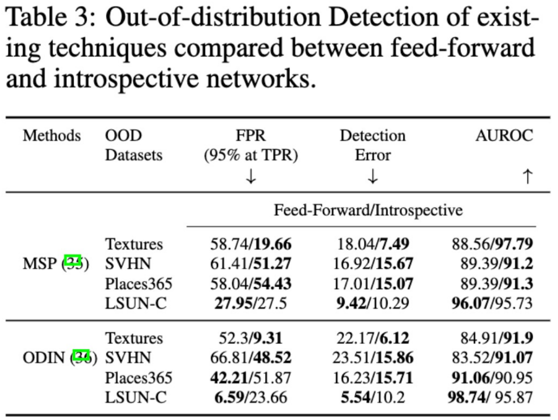


Table 3: Out-of-distribution Detection of existing techniques compared between feed-forward and introspective networks.

Methods	OOD Datasets	FPR (95% at TPR) ↓	Detection Error ↓	AUROC ↑
		Feed-Forward/Introspective		
MSP (35)	Textures	58.74/**19.66**	18.04/**7.49**	88.56/**97.79**
	SVHN	61.41/**51.27**	16.92/**15.67**	89.39/**91.2**
	Places365	58.04/**54.43**	17.01/**15.07**	89.39/**91.3**
	LSUN-C	**27.95**/27.5	**9.42**/10.29	**96.07**/95.73
ODIN (36)	Textures	52.3/**9.31**	22.17/**6.12**	84.91/**91.9**
	SVHN	66.81/**48.52**	23.51/**15.86**	83.52/**91.07**
	Places365	**42.21**/51.87	16.23/**15.71**	**91.06**/90.95
	LSUN-C	**6.59**/23.66	**5.54**/10.2	**98.74**/ 95.87

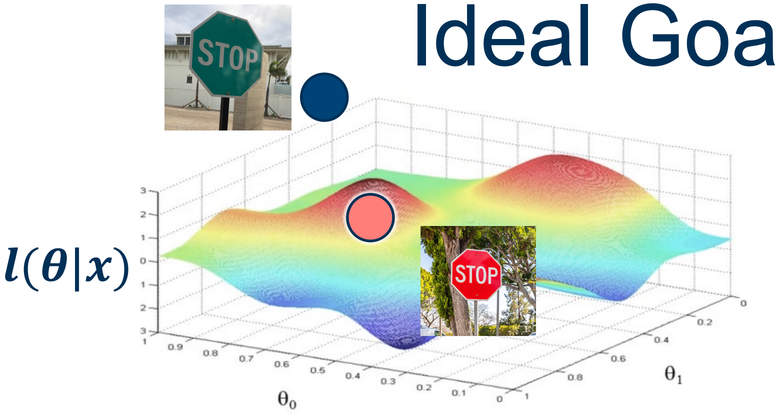
M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Part I, II and III

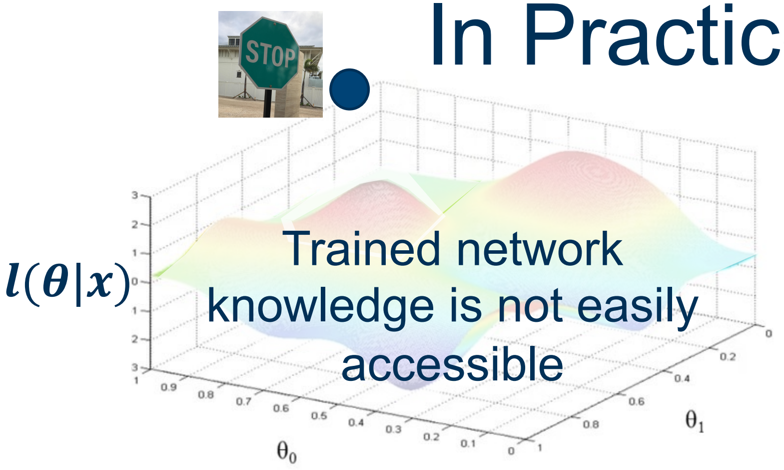
Tying it Back

From Part I

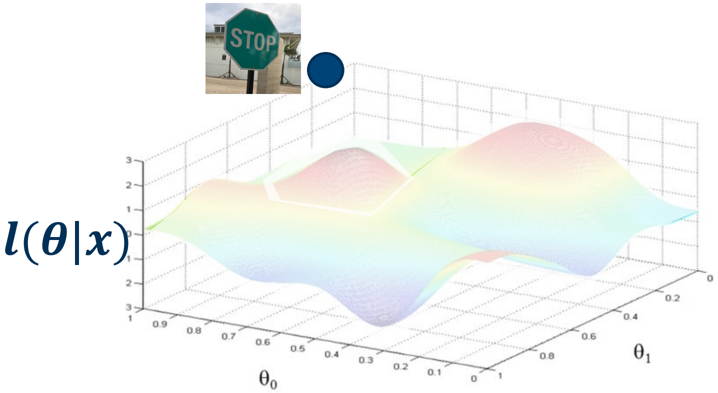
Ideal Goal



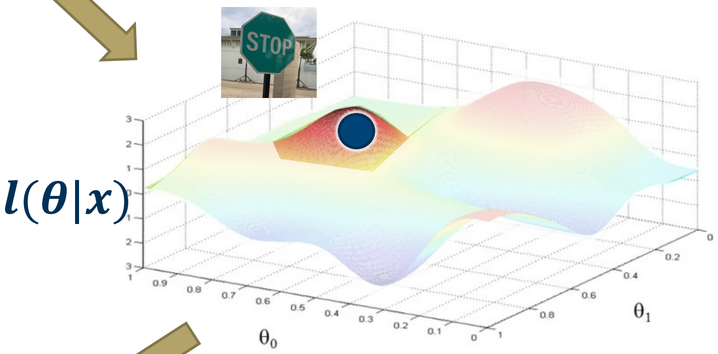
In Practice

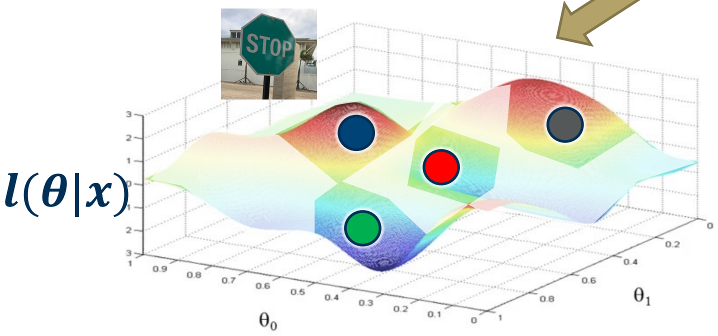


Trained network knowledge is not easily accessible



Novel data projects onto the likelihood function (however incorrectly), and extracts fisher information around the projection





By backpropagating contrast classes (and not updating the network), the network finds the steepest descent towards other regions of likelihood function