Robust Neural Networks

Part 3: Uncertainty at Inference

Objective

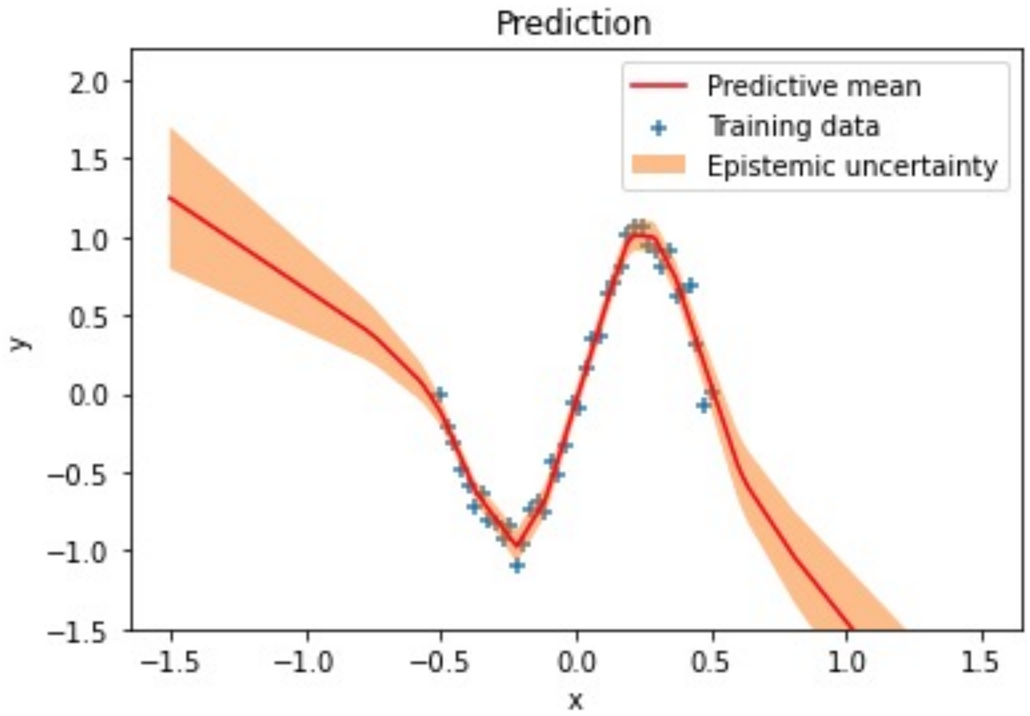
Objective of the Tutorial

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:

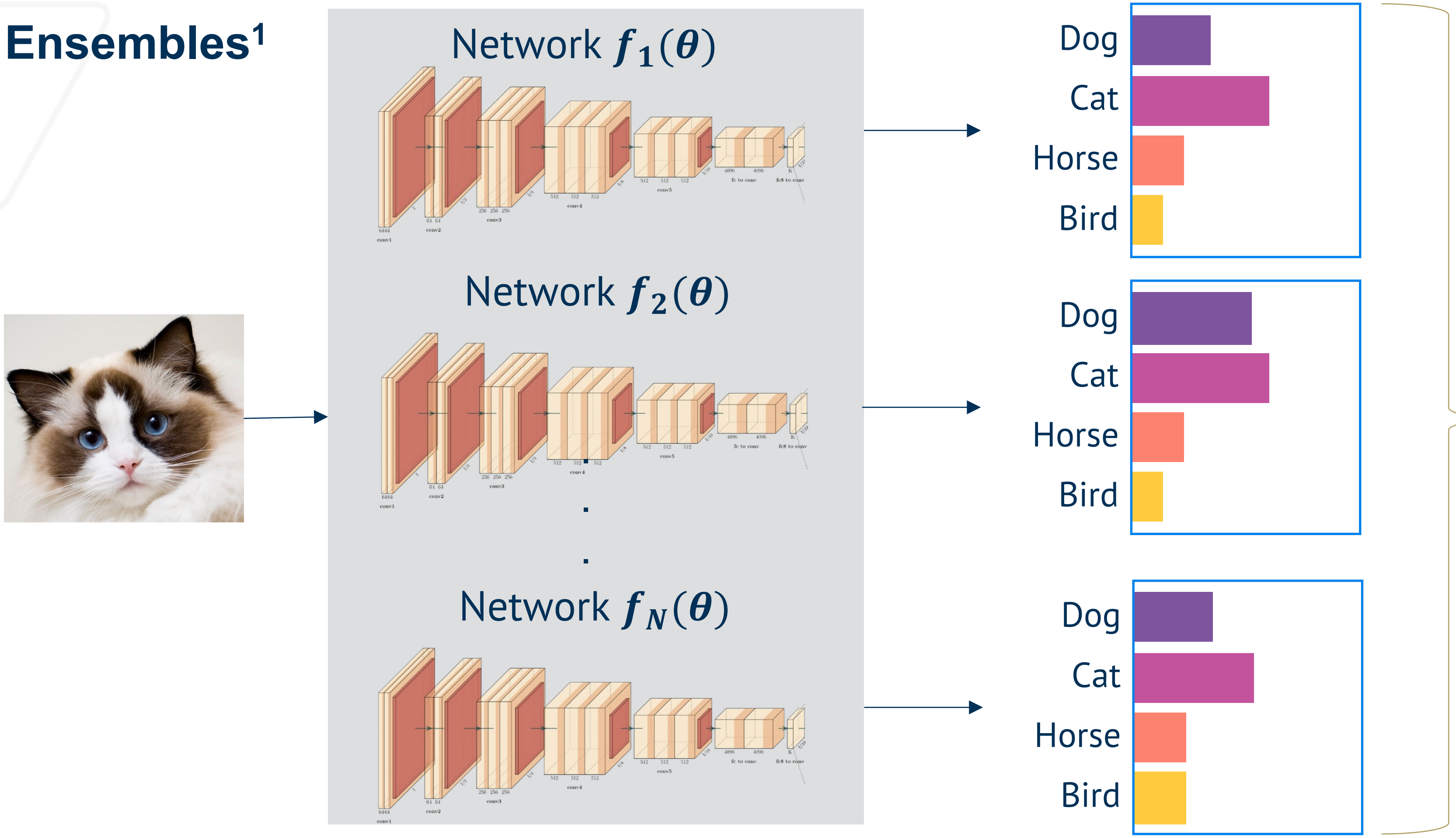
- When training data is **available: Less uncertainty**
- When training data is **unavailable: More uncertainty**

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

Uncertainty

Uncertainty Quantification in Neural Networks

Via Ensembles[1]



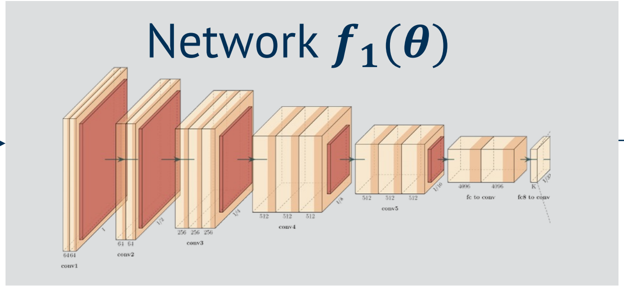
Variation within outputs $Var(y)$ is the uncertainty. Commonly referred to as **Prediction Uncertainty.**

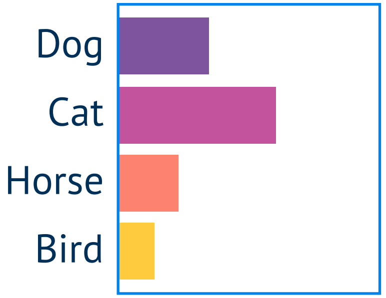
[1] Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles." *Advances in neural information processing systems* 30 (2017).

Uncertainty

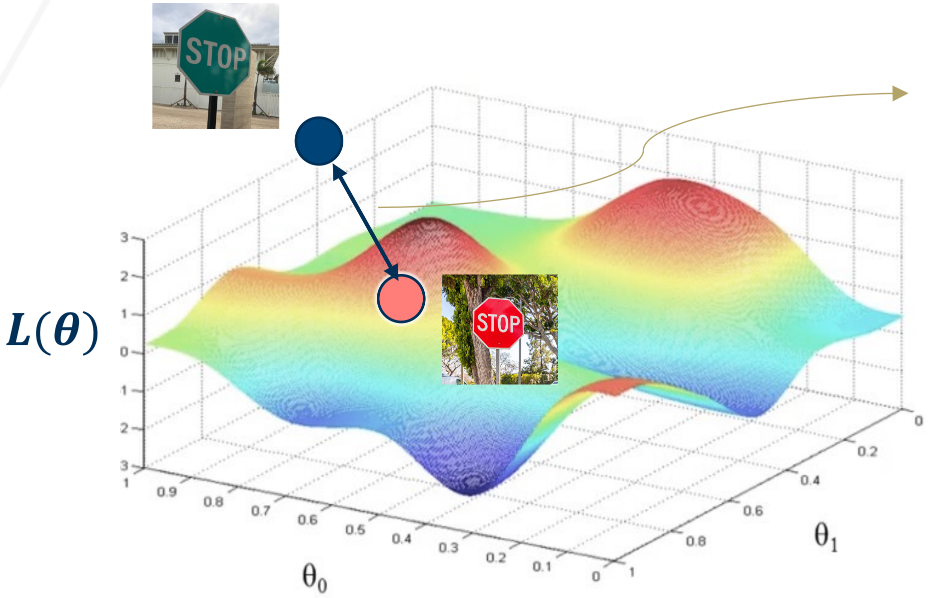
Uncertainty Quantification in Neural Networks

Via Single pass methods[1]





Uncertainty quantification using a single network and a single pass

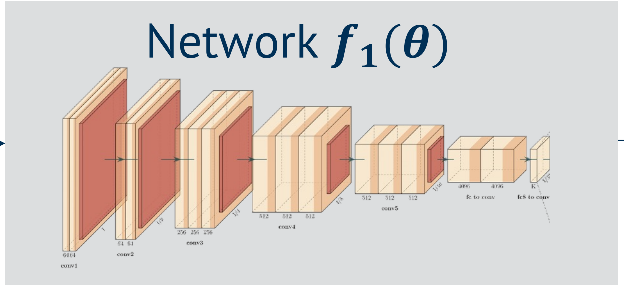


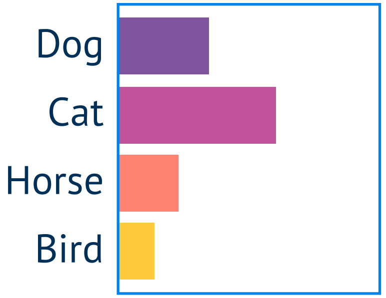
Calculate distance from some trained clusters

Does not require multiple networks!

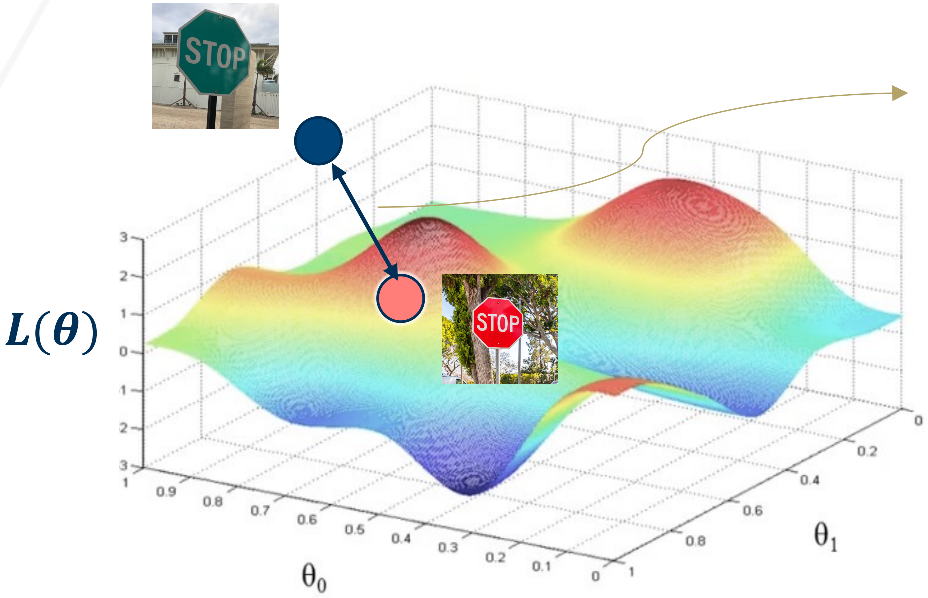
[1Van Amersfoort, J., Smith, L., Teh, Y. W., & Gal, Y. (2020, November). Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning* (pp. 9690-9700). PMLR.

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





Uncertainty quantification using a single network and a single pass



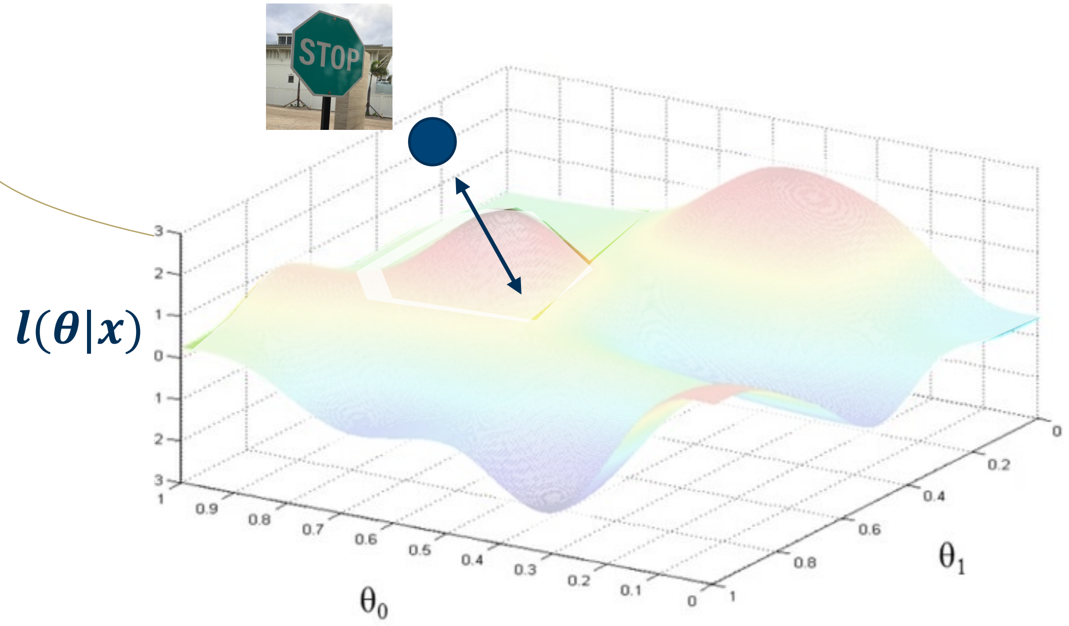
Calculate distance from some trained clusters

Does not require multiple networks!

Challenge: Class and prediction cannot be trusted!

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

Distance from unknown cluster



Two techniques:

1. **Backpropagating Confounding labels for Adversarial Detection**
2. Backpropagating Confounding labels for Robust Prediction

IEEE Access

Probing the Purview of Neural Networks via Gradient Analysis

Jinsol Lee,
PhD Candidate

Mohit Prabhushankar, PhD
Postdoc

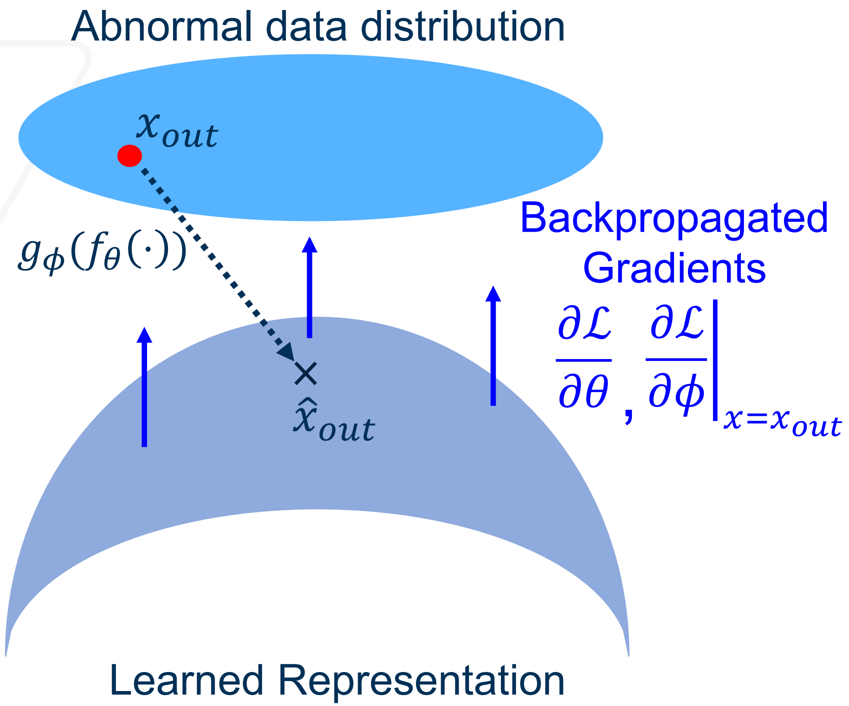
Ghassan AlRegib, PhD
Professor

SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

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

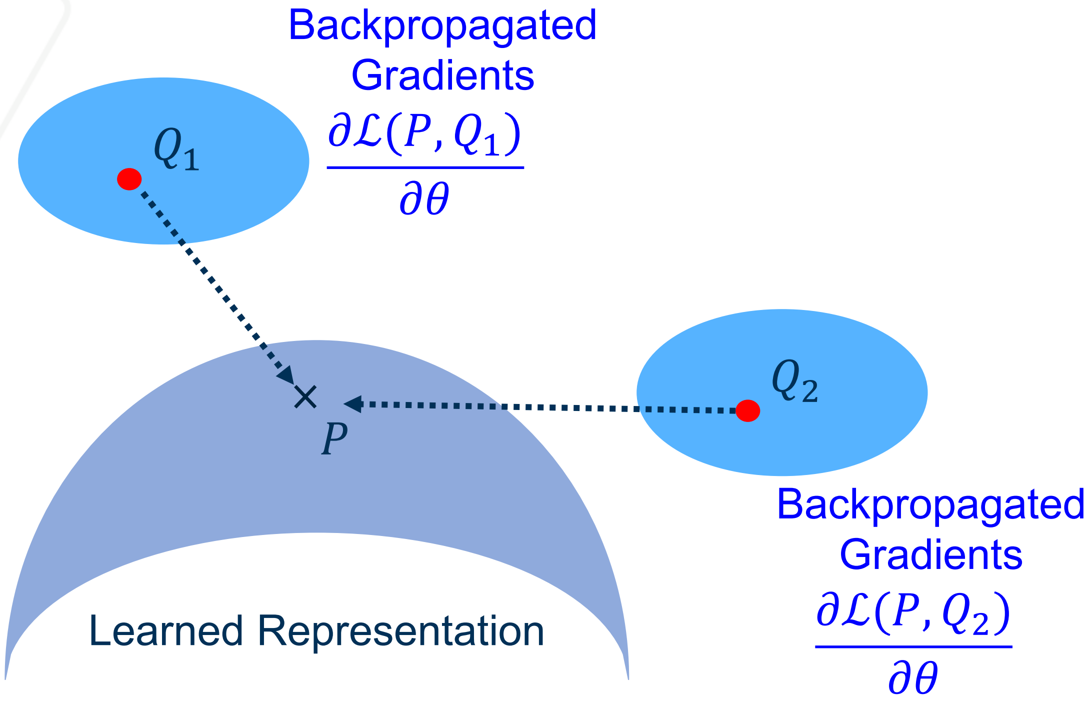
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

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

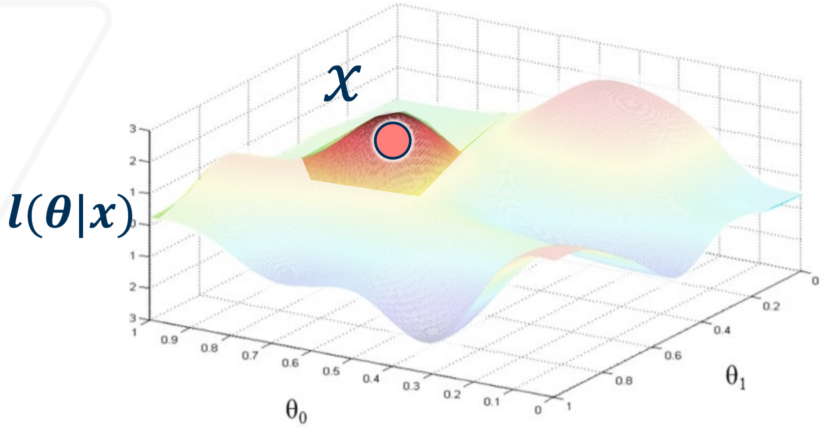
What is uncertainty?

Probing the Purview of Neural Networks via Gradient Analysis

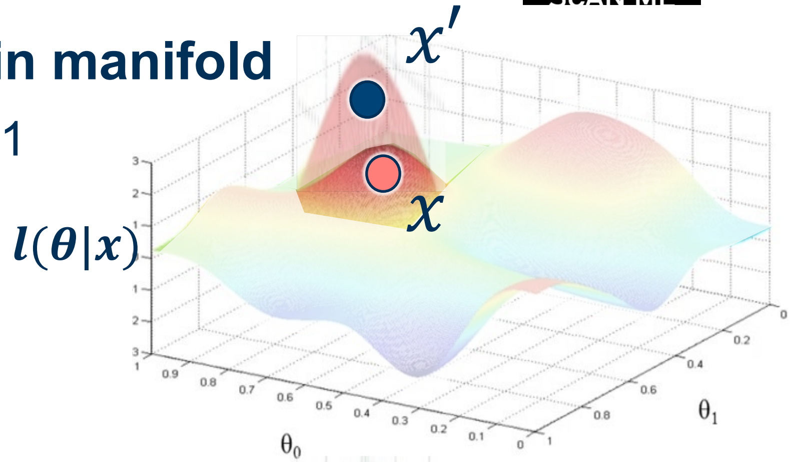
SCAN ME

Gradients represent the local required change in manifold

Similar to introspective learning!

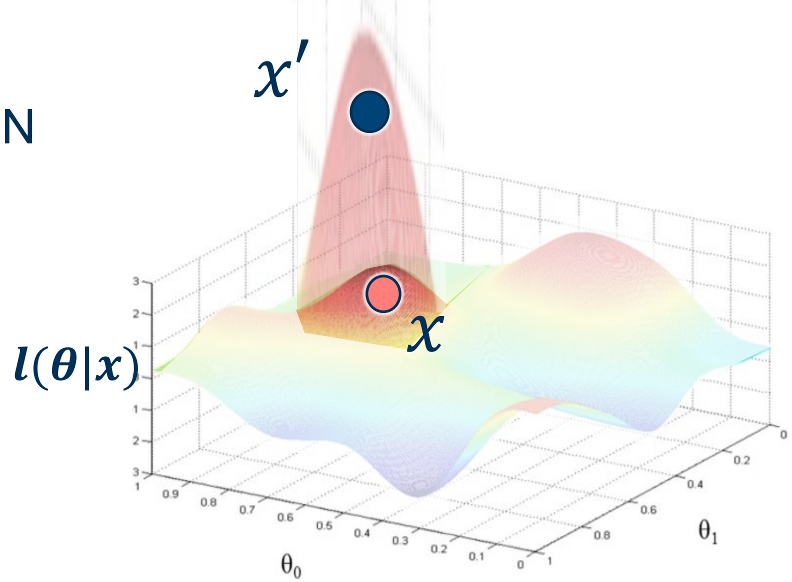


Contrast class 1



.
.
.

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

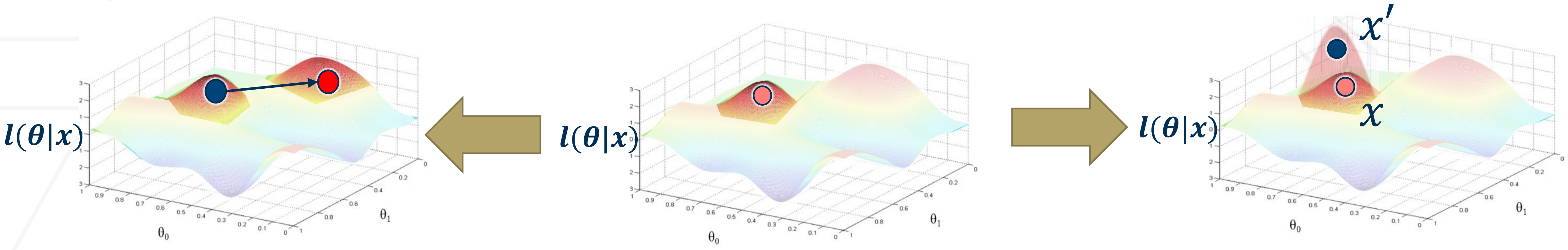
How is this different from Explainability?

SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

Part 3: Explainability

Part 4: Uncertainty



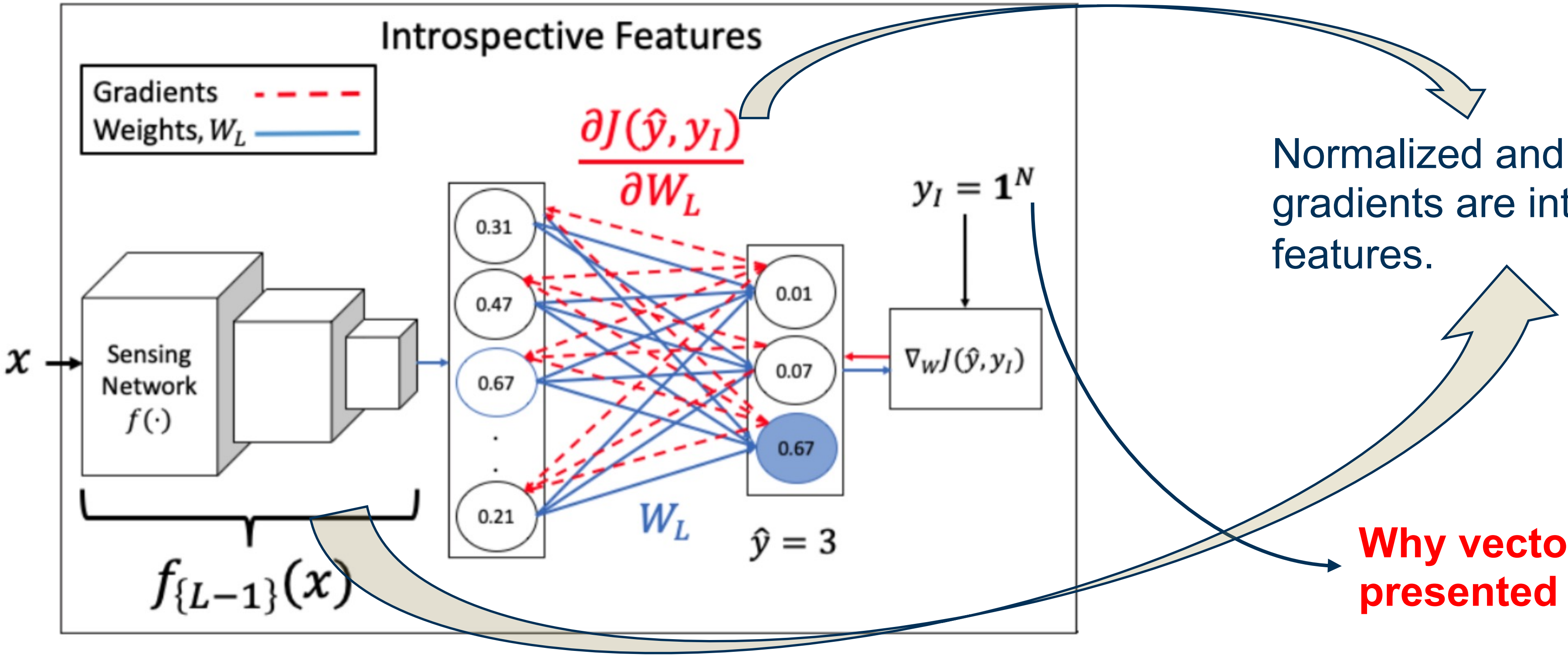
- In Part 3: Activations of learned manifold are weighted by gradients w.r.t. activations to extract information and provide explanations

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

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.

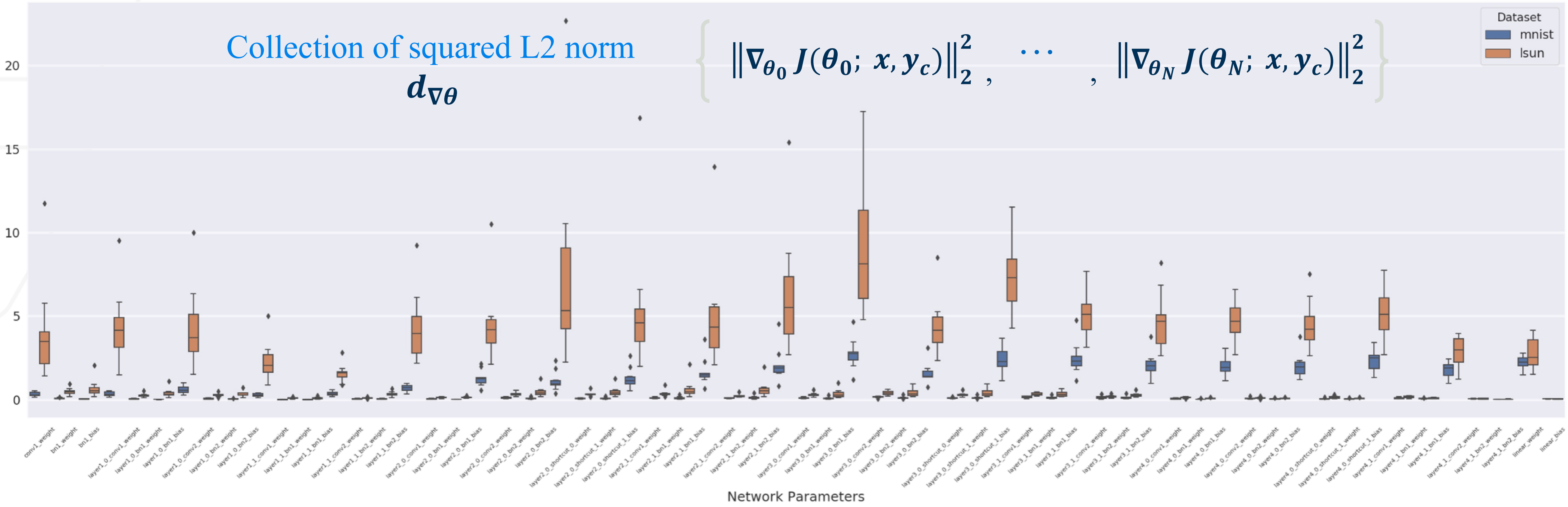
Uncertainty in Neural Networks

Utilizing Gradient Features

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

Step 2: Take L2 norm of all generated gradients



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.

OLIVES @GeorgiaTech

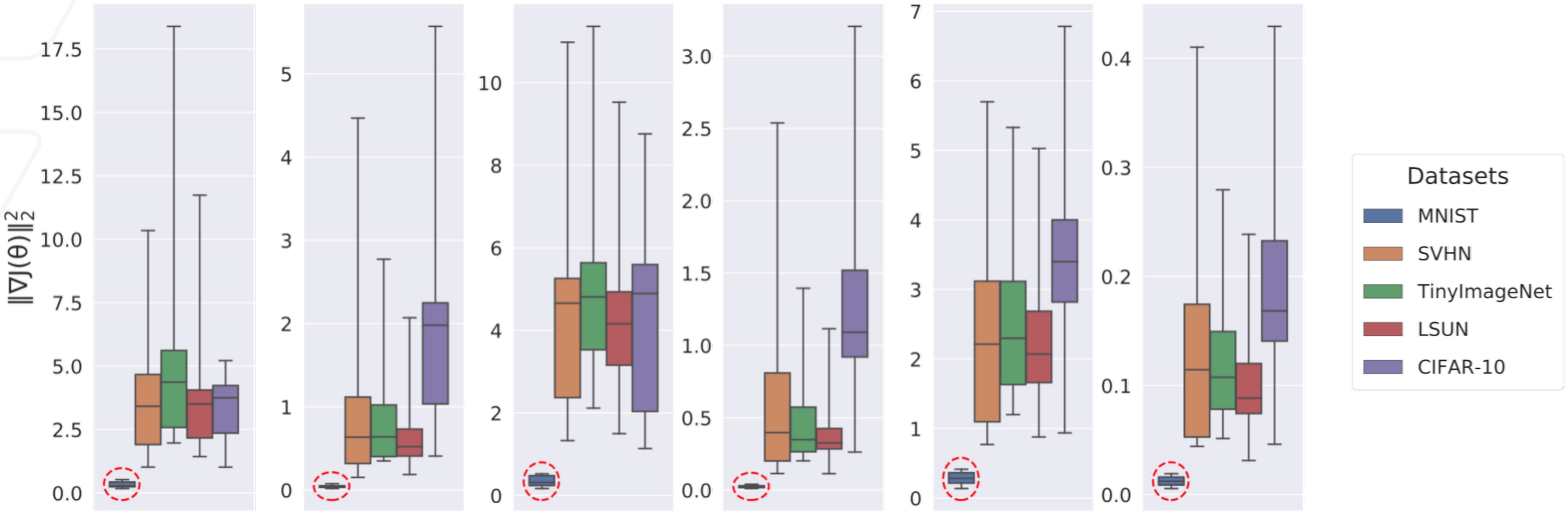
Georgia Tech

Gradient-based Uncertainty
Uncertainty in OOD Setting

SCAN ME

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.

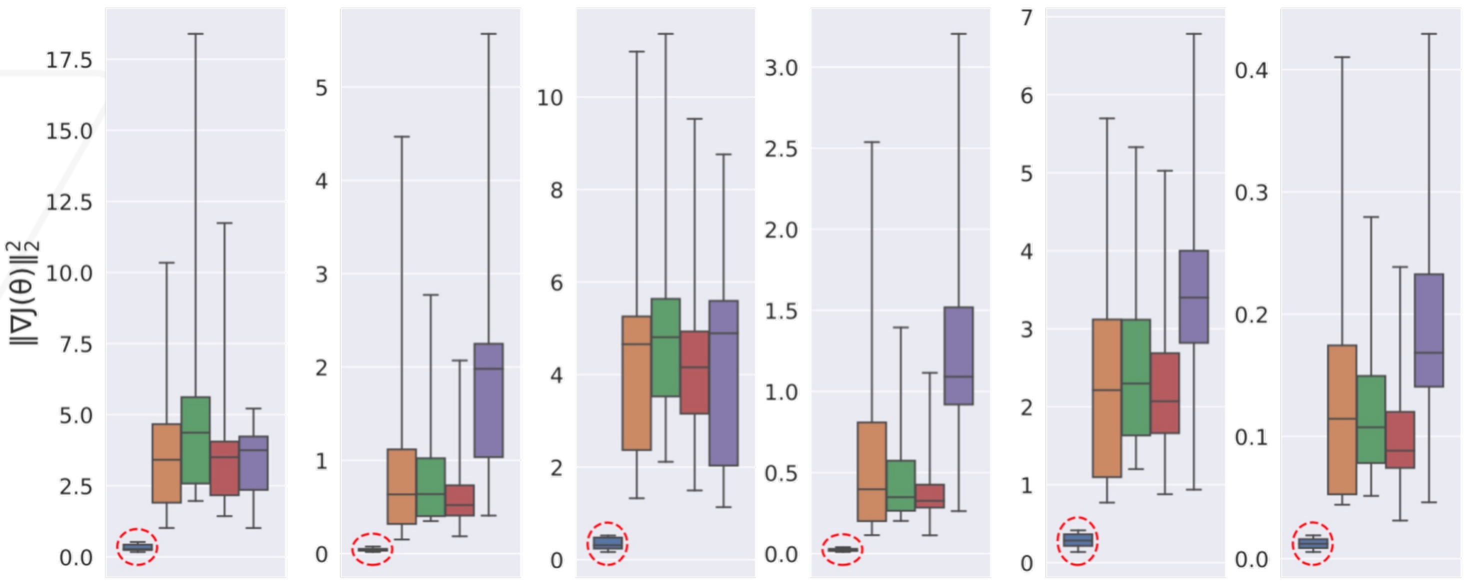
Gradient-based Uncertainty

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**

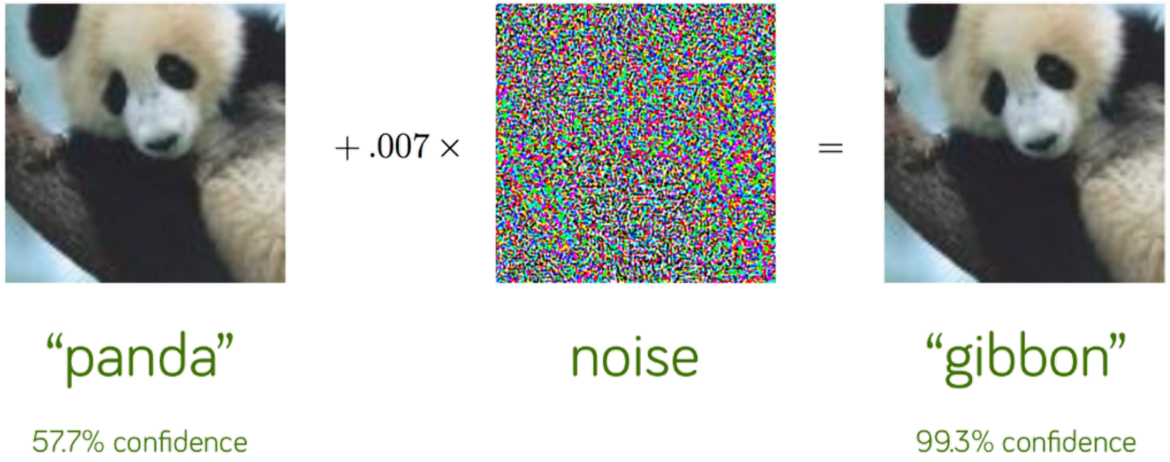
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Gradient-based Uncertainty
Uncertainty in Adversarial Setting

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

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

SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

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**

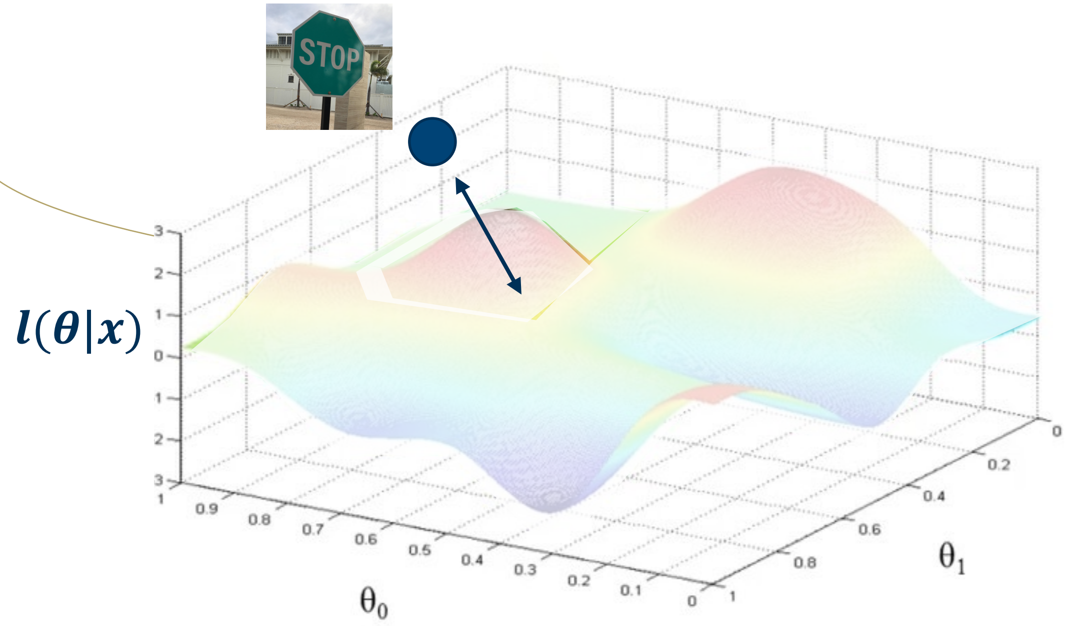
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

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. Backpropagating Confounding labels for Adversarial Detection
2. **Backpropagating Confounding labels for Robust Prediction**

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

SCAN ME

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

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?

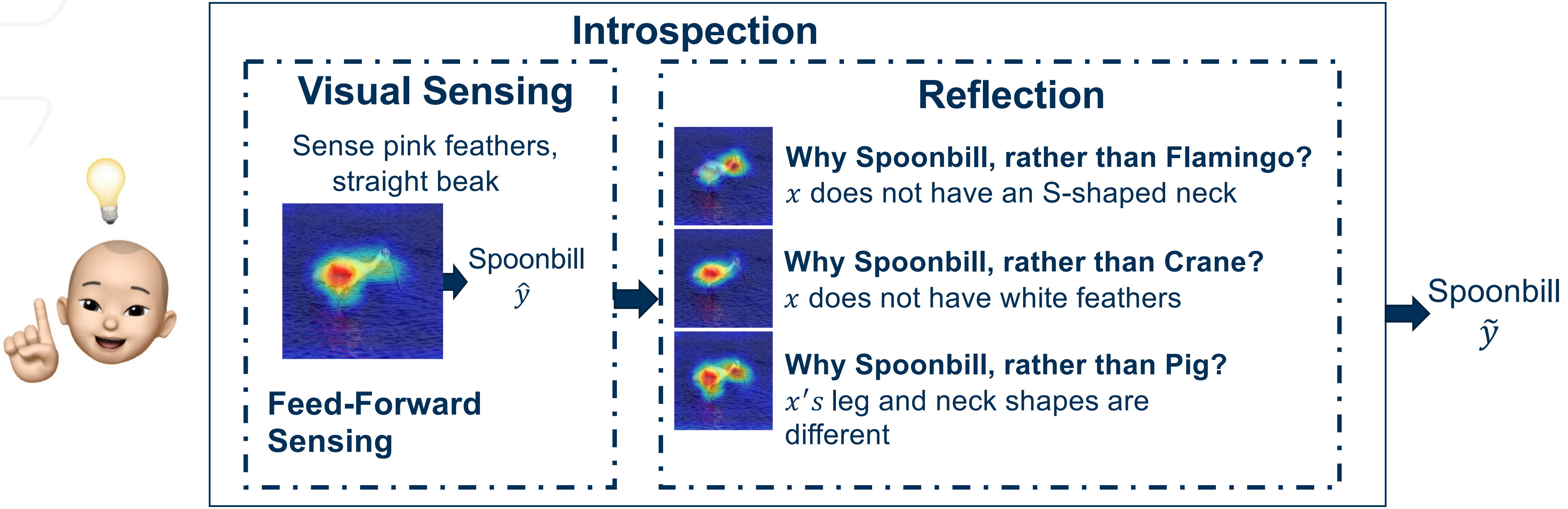
Introspection

What is Introspection?

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



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.

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.

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

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.*

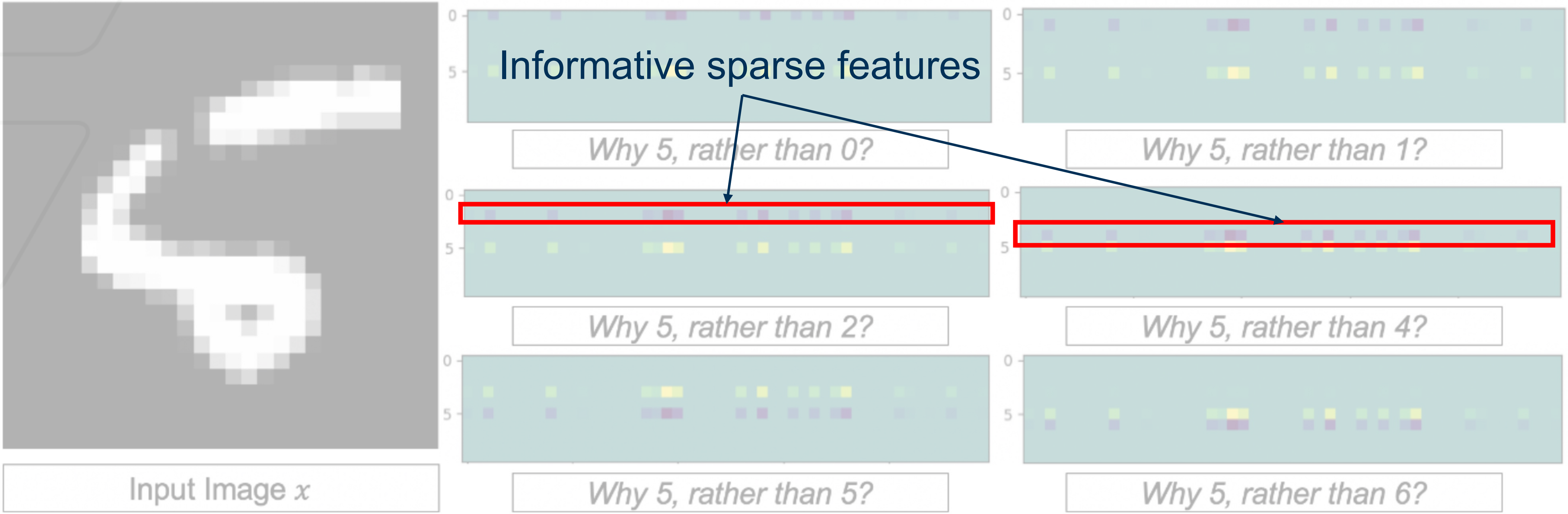
Technical Definition : *Given a network* $f(x)$, *a datum* x, *and the network's prediction* $f(x) = \hat{y}$, *introspection in* $f(\cdot)$ *is the measurement of change induced in the network parameters when a label* Q *is introduced as the label for* x..

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.

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

SCAN ME

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



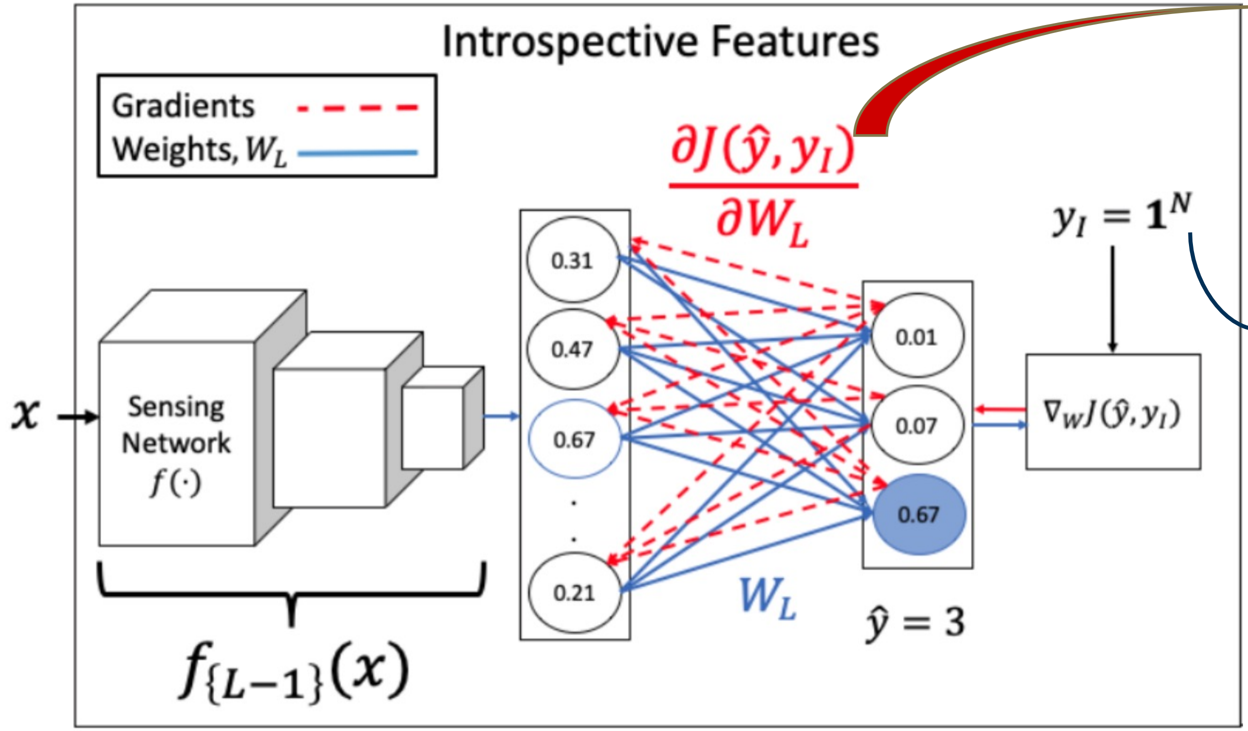
M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

OLIVES @GeorgiaTech
Georgia Tech

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



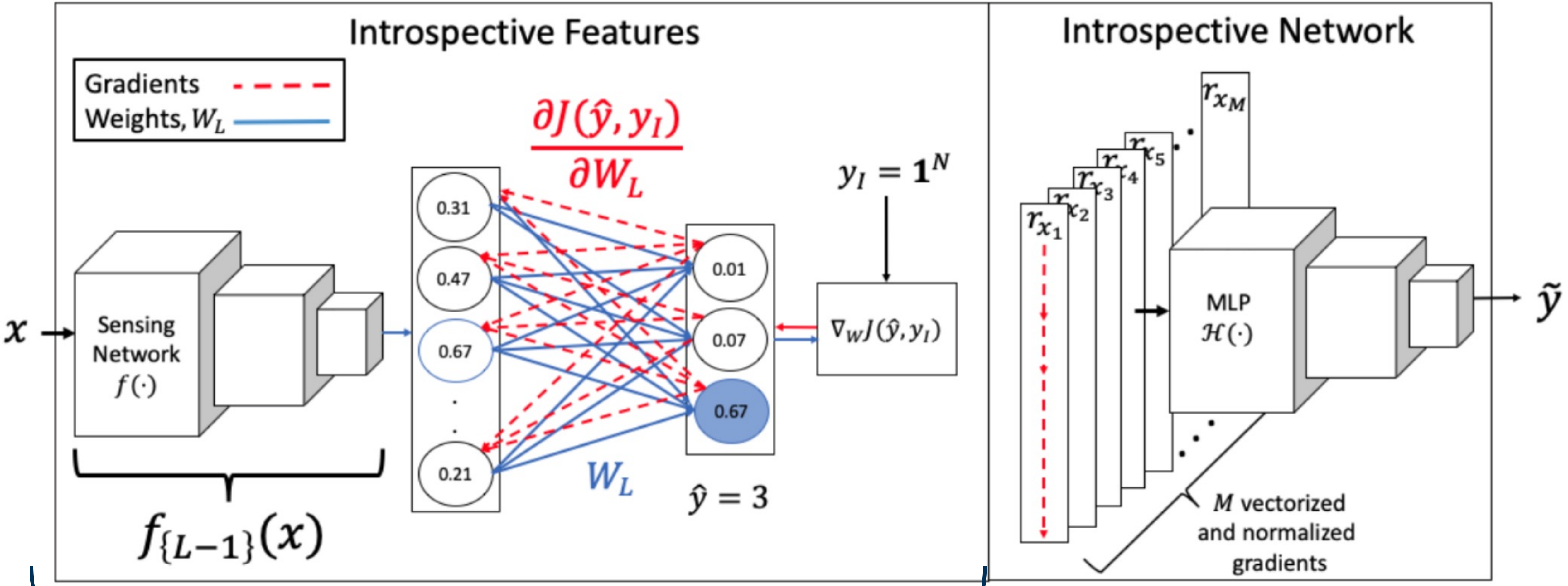
Normalized and vectorized gradients are introspective features

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



Introspective Features

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.

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

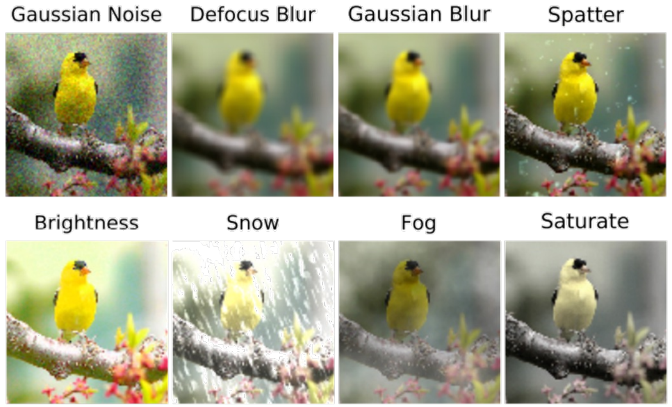
SCAN ME

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.

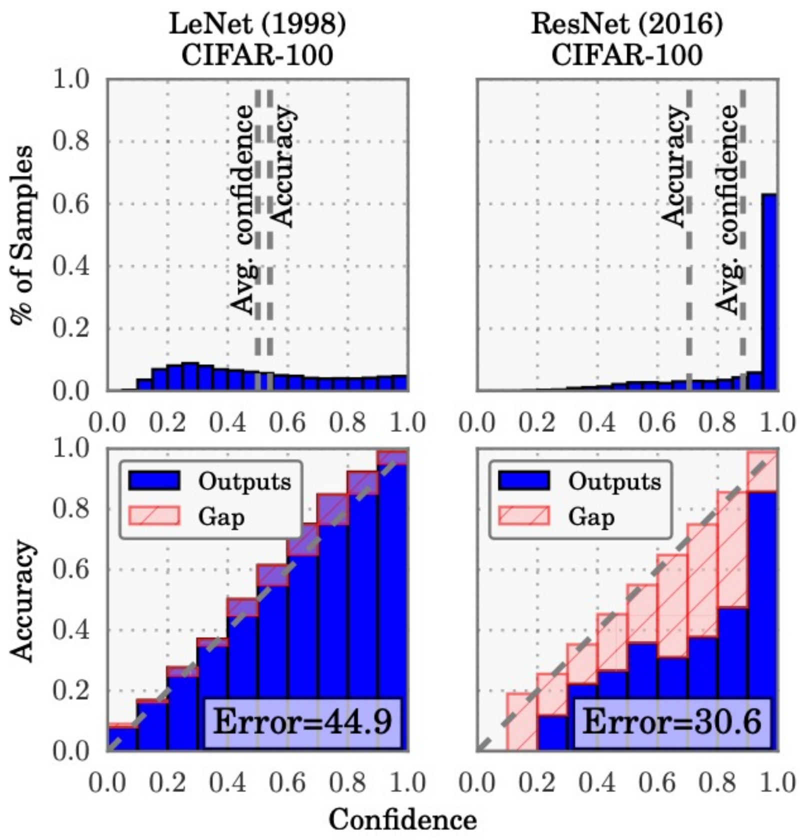
Calibration

A note on Calibration..

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

SCAN ME

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

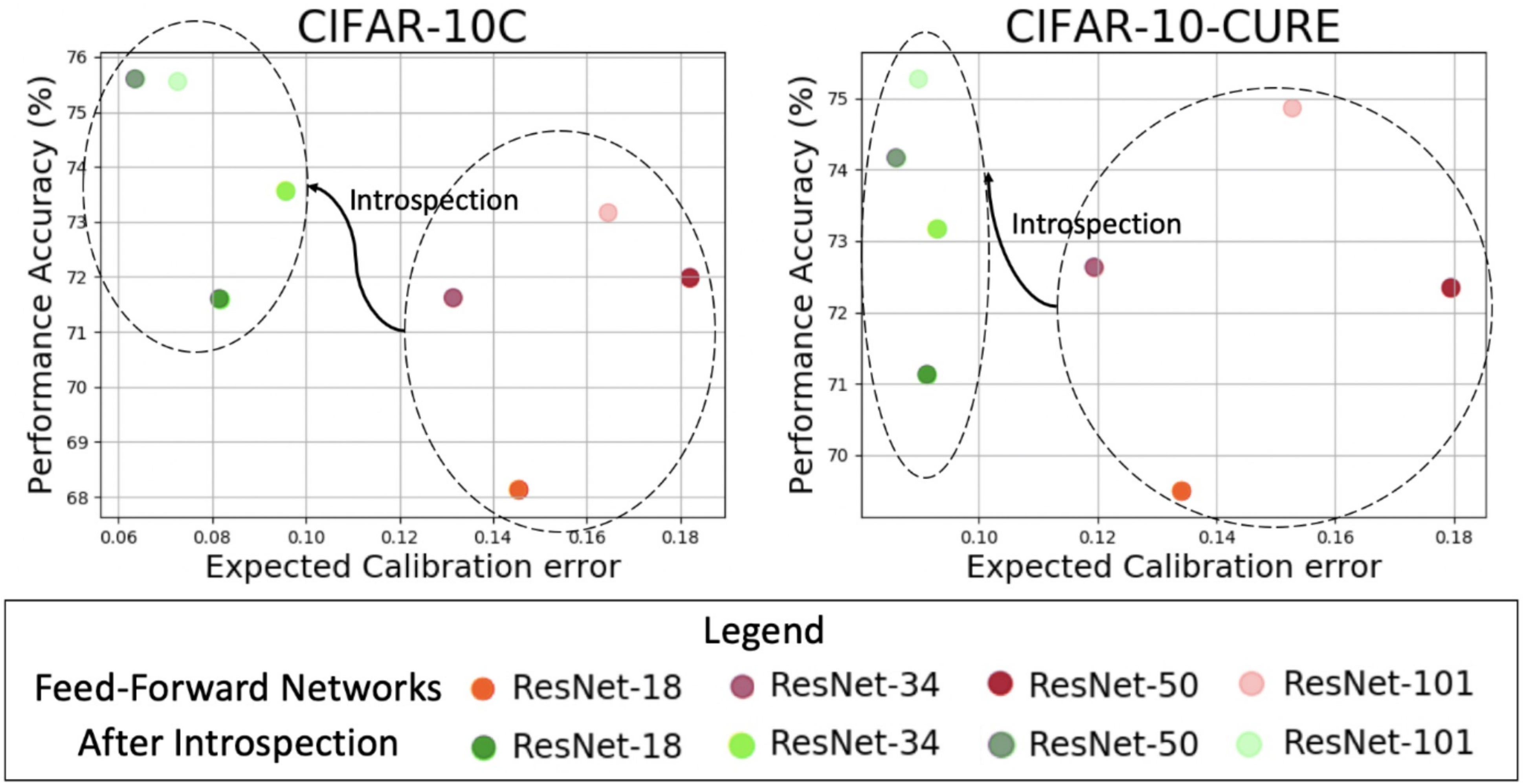
Generalization and Calibration results

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

SCAN ME

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

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	67.89%
	Introspective	**71.4%**
Denoising	Feed-Forward	65.02%
	Introspective	**68.86%**
Adversarial Train (27)	Feed-Forward	68.02%
	Introspective	**70.86%**
SimCLR (19)	Feed-Forward	70.28%
	Introspective	**73.32%**
Augment Noise (28)	Feed-Forward	76.86%
	Introspective	**77.98%**
AugMix (26)	Feed-Forward	89.85%
	Introspective	**89.89%**

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

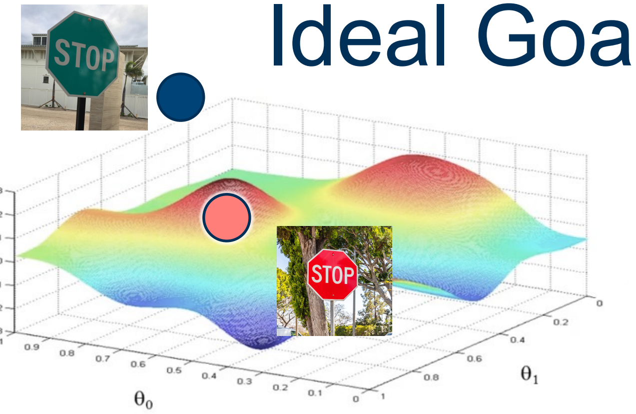
M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Part I, II and III

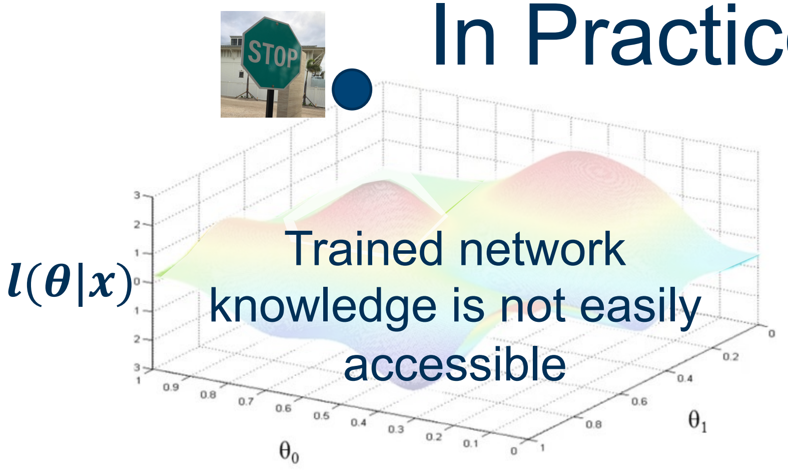
Tying it Back

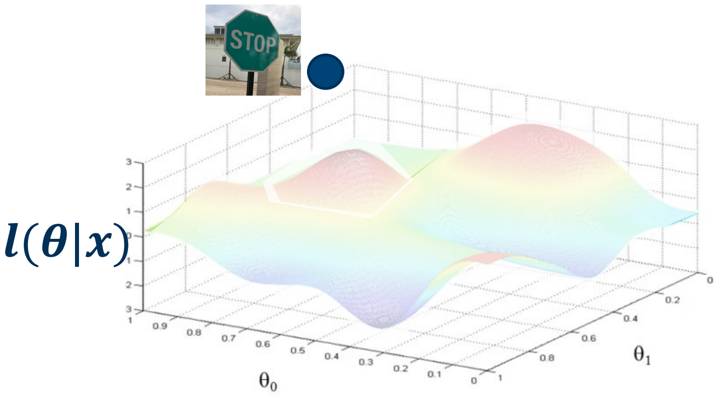
From Part I

Ideal Goal

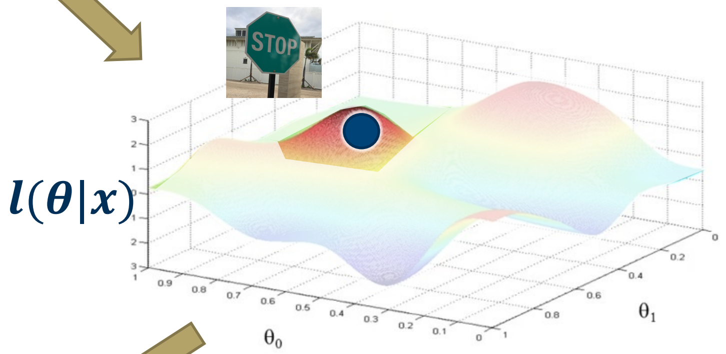


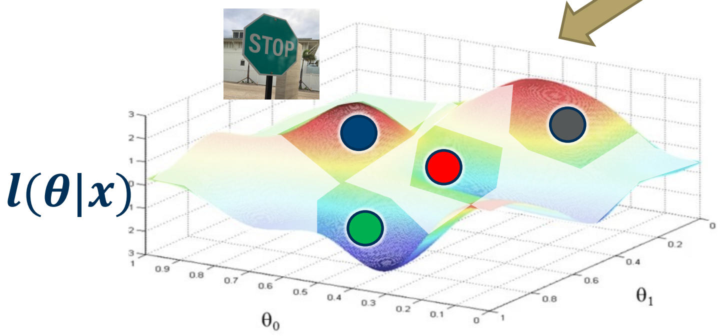
In Practice





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