Robust Neural Networks: Explainability, Uncertainty, and Intervenability

Ghassan AlRegib, PhD
Professor

Mohit Prabhushankar, PhD
Postdoctoral Fellow

Omni Lab for Intelligent Visual Engineering and Science (OLIVES)
School of Electrical and Computer Engineering
Georgia Institute of Technology
{alregib, mohit.p}@gatech.edu

Dec 15, 2023 – Sorrento, Italy

IEEE BigData 2023
Sorrento, Italy

OLIVES @GeorgiaTech

Georgia Tech

Tutorial Materials

Accessible Online

SCAN ME

https://alregib.ece.gatech.edu/ieee-bigdata-2023-tutorial/

{alregib, mohit.p}@gatech.edu

IEEE BigData 2023 Tutorial

Presented by: ***Ghassan AlRegib, and Mohit Prabhushankar***
Georgia Institute of Technology

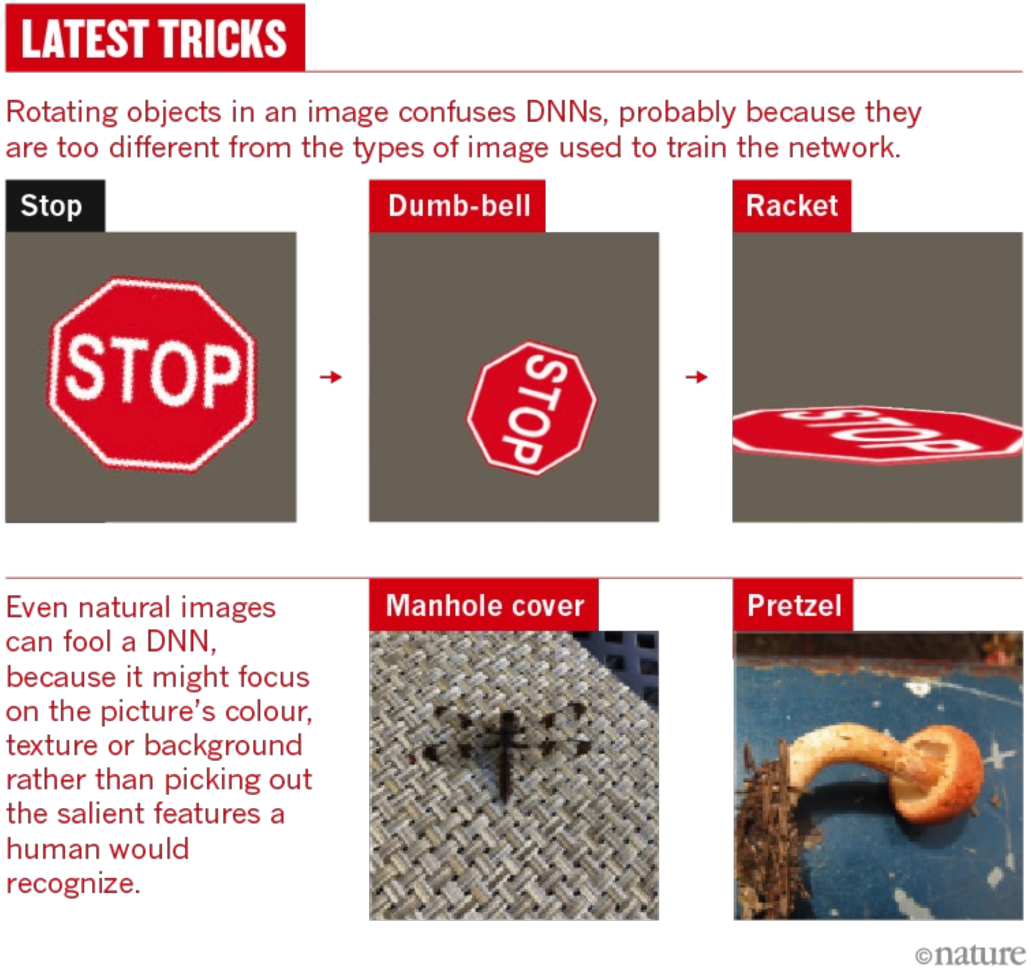
www.ghassanalregib.info

Title: Robust Neural Networks: Explainability, Uncertainty, and Intervenability

Expectation vs Reality of Deep Learning

Deep Learning

Expectation vs Reality



Requirements: Deep Learning-enabled systems must predict correctly on novel data

Novel data sources:

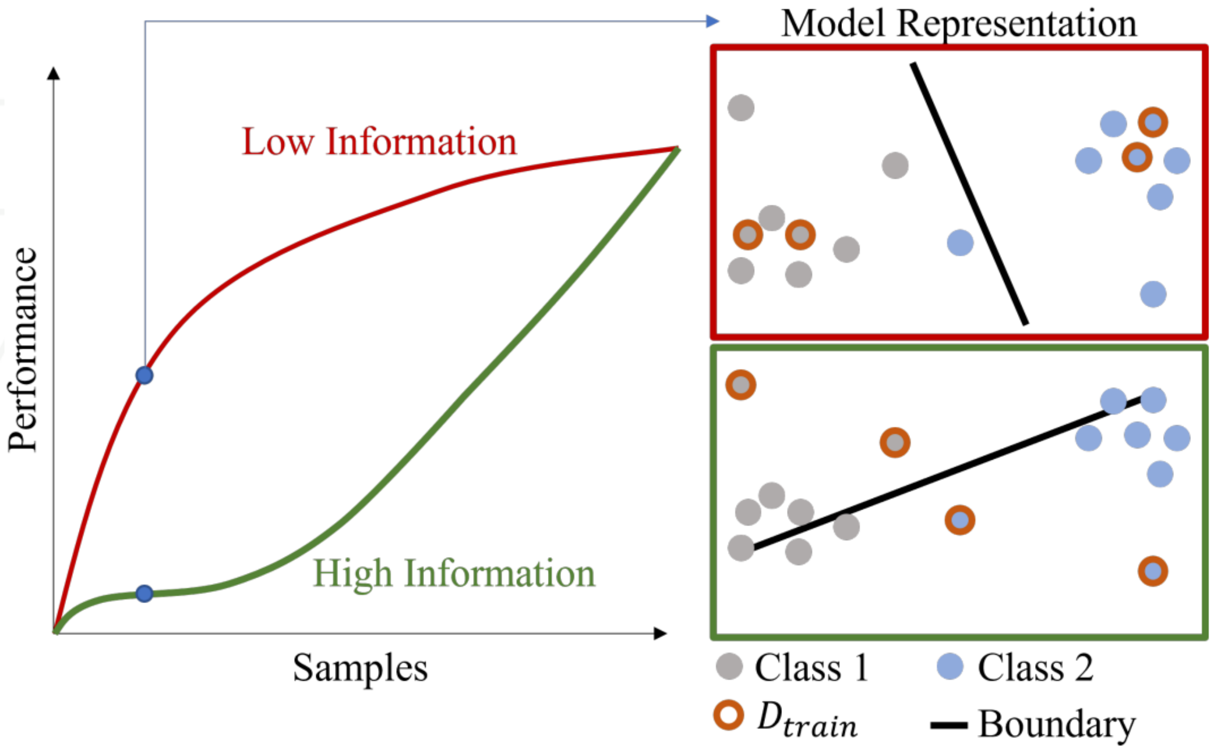
- Test distributions
- Anomalous data
- Out-Of-Distribution data
- Adversarial data
- Corrupted data
- Noisy data
- New classes
- …

Temel, Dogancan, et al. "Cure-tsd: Challenging unreal and real environments for traffic sign detection." *IEEE Transactions on Intelligent Transportation Systems* (2017).

Deep Learning at Training

Overcoming Challenges at Training: Part 1

The most novel/aberrant samples should not be used in early training

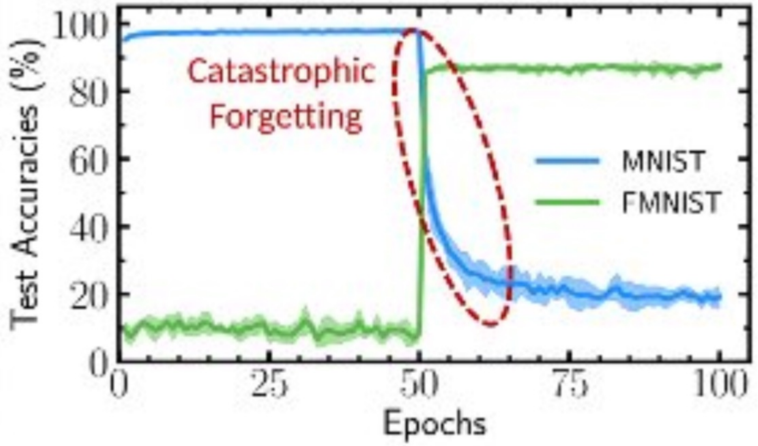


- The first instance of training must occur with less informative samples
- Ex: For autonomous vehicles, less informative means
  - Highway scenarios
  - Parking
  - No accidents
  - No aberrant events

Novel samples = Most Informative

Benkert, R., Prabushankar, M., AlRegib, G., Pacharmi, A., & Corona, E. (2023). Gaussian Switch Sampling: A Second Order Approach to Active Learning. *IEEE Transactions on Artificial Intelligence*.

Subsequent training must not focus only on novel data



- The model performs well on the new scenarios, while forgetting the old scenarios
- A number of techniques exist to overcome this trend
- However, they affect the overall performance in large-scale settings
- It is not always clear **if and when** to incorporate novel scenarios in training

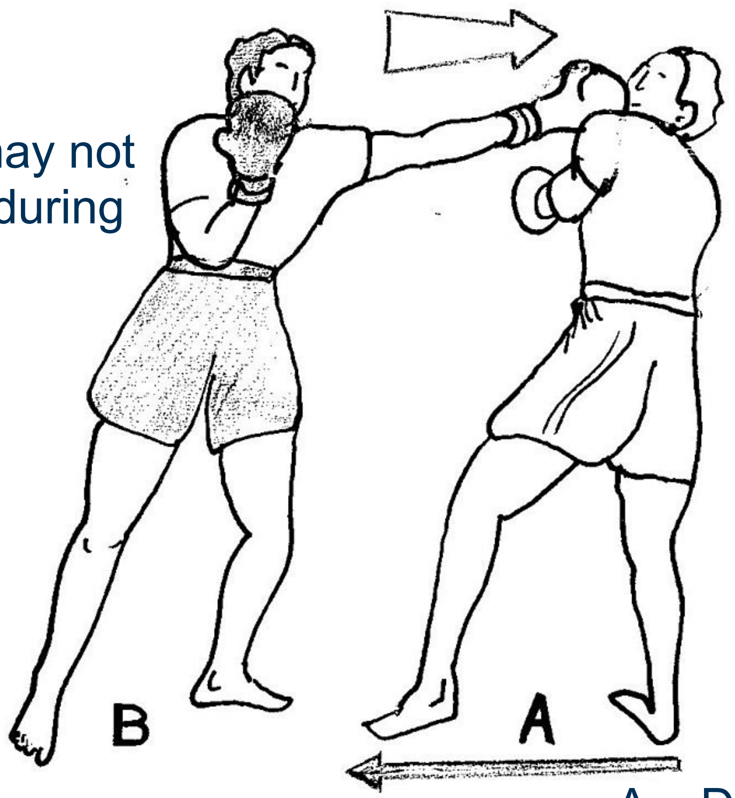
Laborieux, Axel, et al. "Synaptic metaplasticity in binarized neural networks." *Nature communications* 12.1 (2021): 2549.

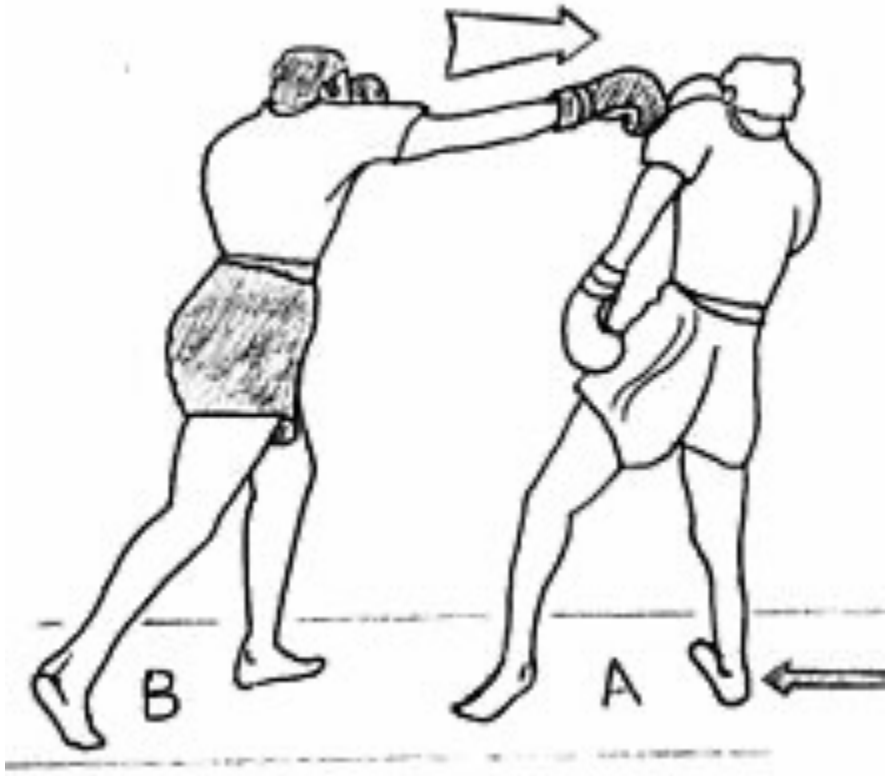
Deep Learning at Training

Overcoming Challenges at Training

Novel data packs a 1-2 punch!

Novel data may not be available during training





Even if available, novel data does not easily fit into either the earlier or later stages of training

A = Deep Neural Networks
B = Novel data

We must handle novel data at Inference!!

Novel data sources:

- Test distributions
- Anomalous data
- Out-Of-Distribution data
- Adversarial data
- Corrupted data
- Noisy data
- New classes
- …

Model Train

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
- Part 4: Intervenability at Inference
- Part 5: Conclusions and Future Directions

Robust Neural Networks

Part I: Inference in Neural Networks

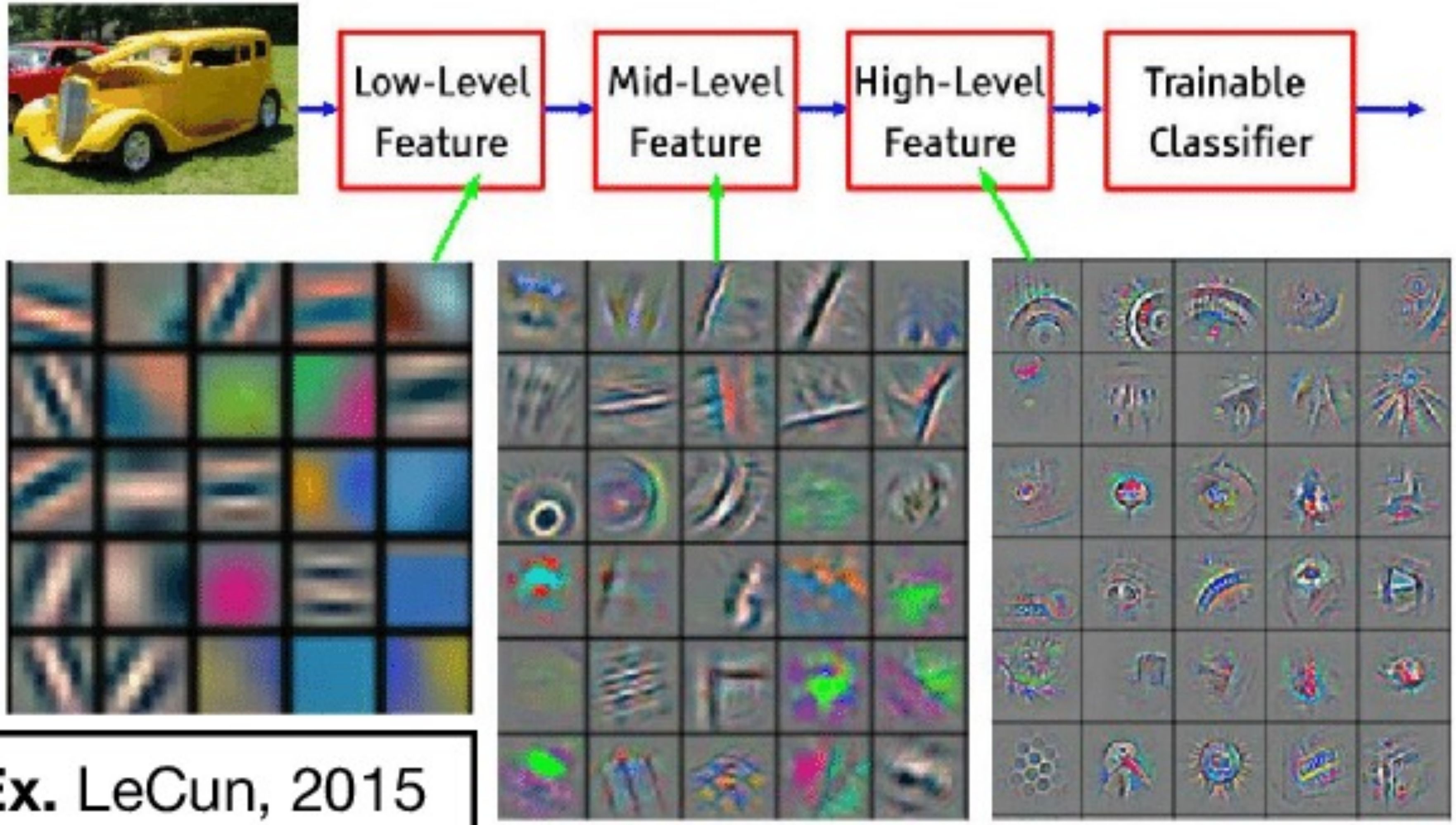
Objective

Objective of the Tutorial

To discuss methodologies that promote robustness in neural networks at inference

- **Part 1: Inference in Neural Networks**
  - Neural Network Basics
  - Robustness in Deep Learning
  - Information at Inference
  - Challenges at Inference
  - Gradients at Inference
- Part 2: Explainability at Inference
- Part 3: Uncertainty at Inference
- Part 4: Intervenability at Inference
- Part 5: Conclusions and Future Directions

Deep Learning
Overview

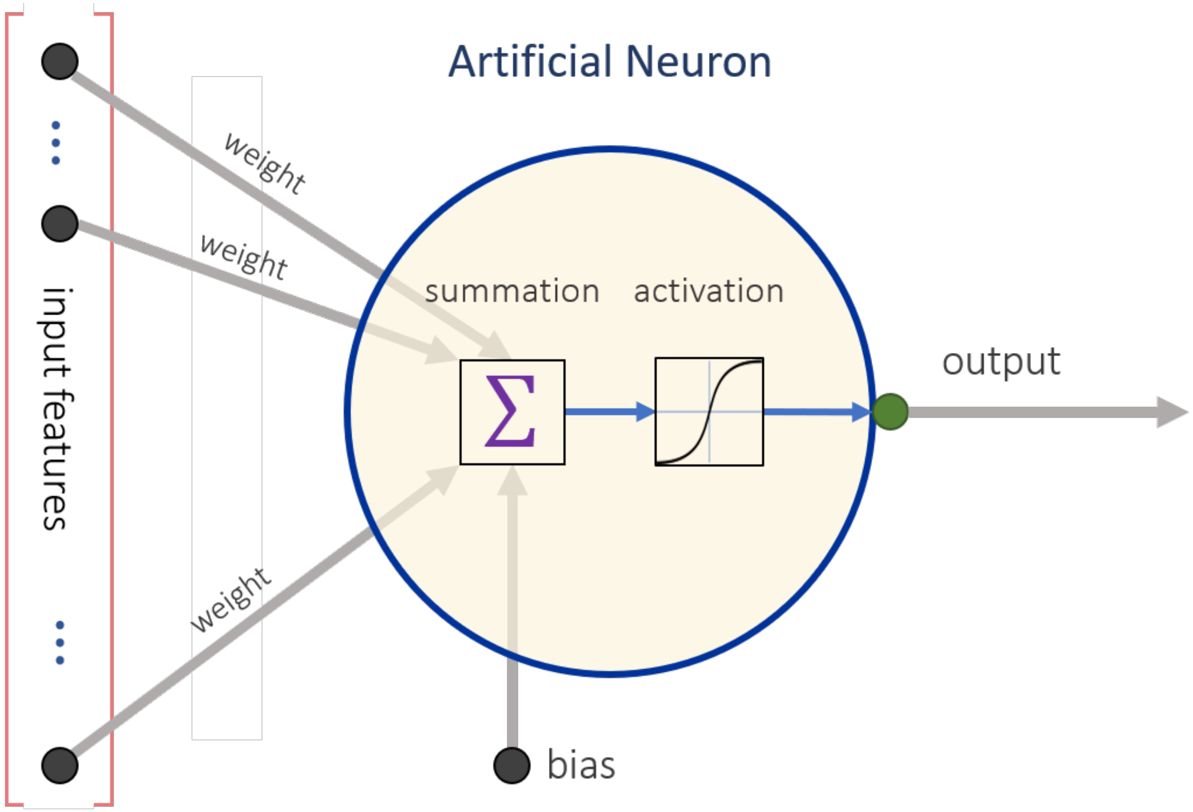


Ex. LeCun, 2015

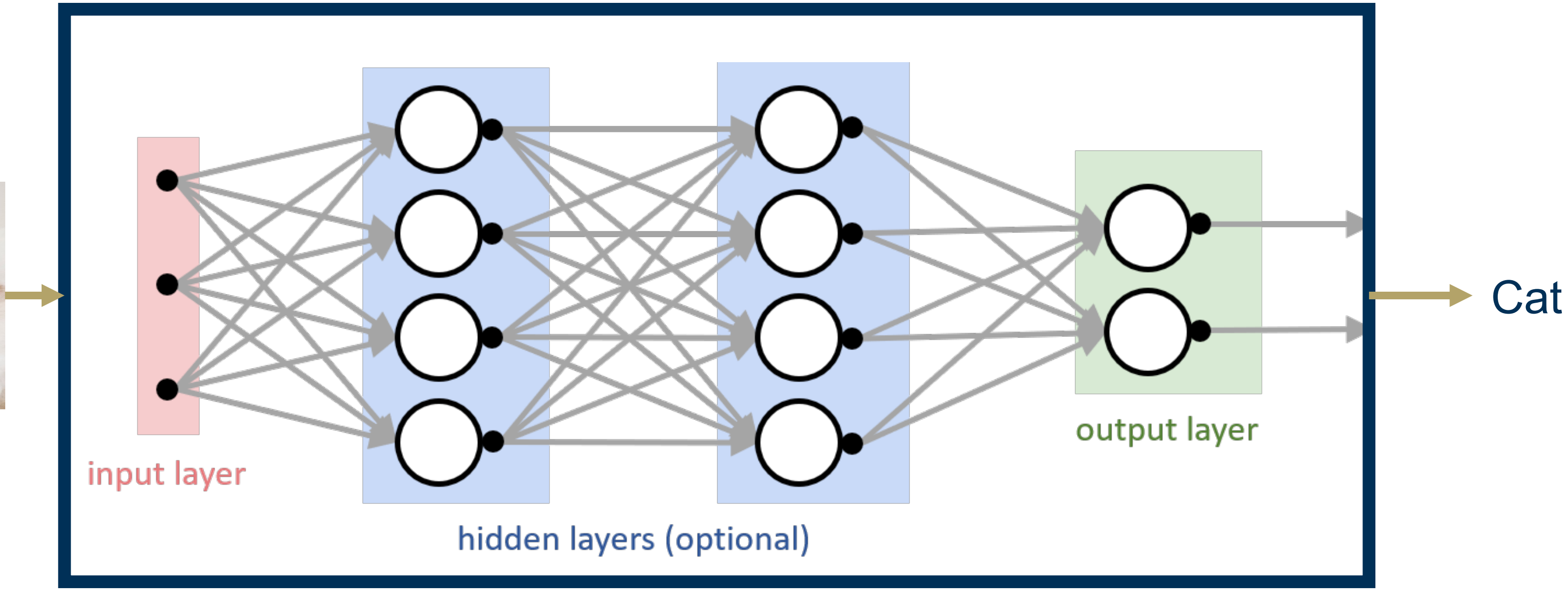
The underlying computation unit is the Neuron

Artificial neurons consist of:

- A single **output**
- Multiple **inputs**
- Input **weights**
- A **bias** input
- An **activation function**



Neurons are stacked and densely connected to construct ANNs



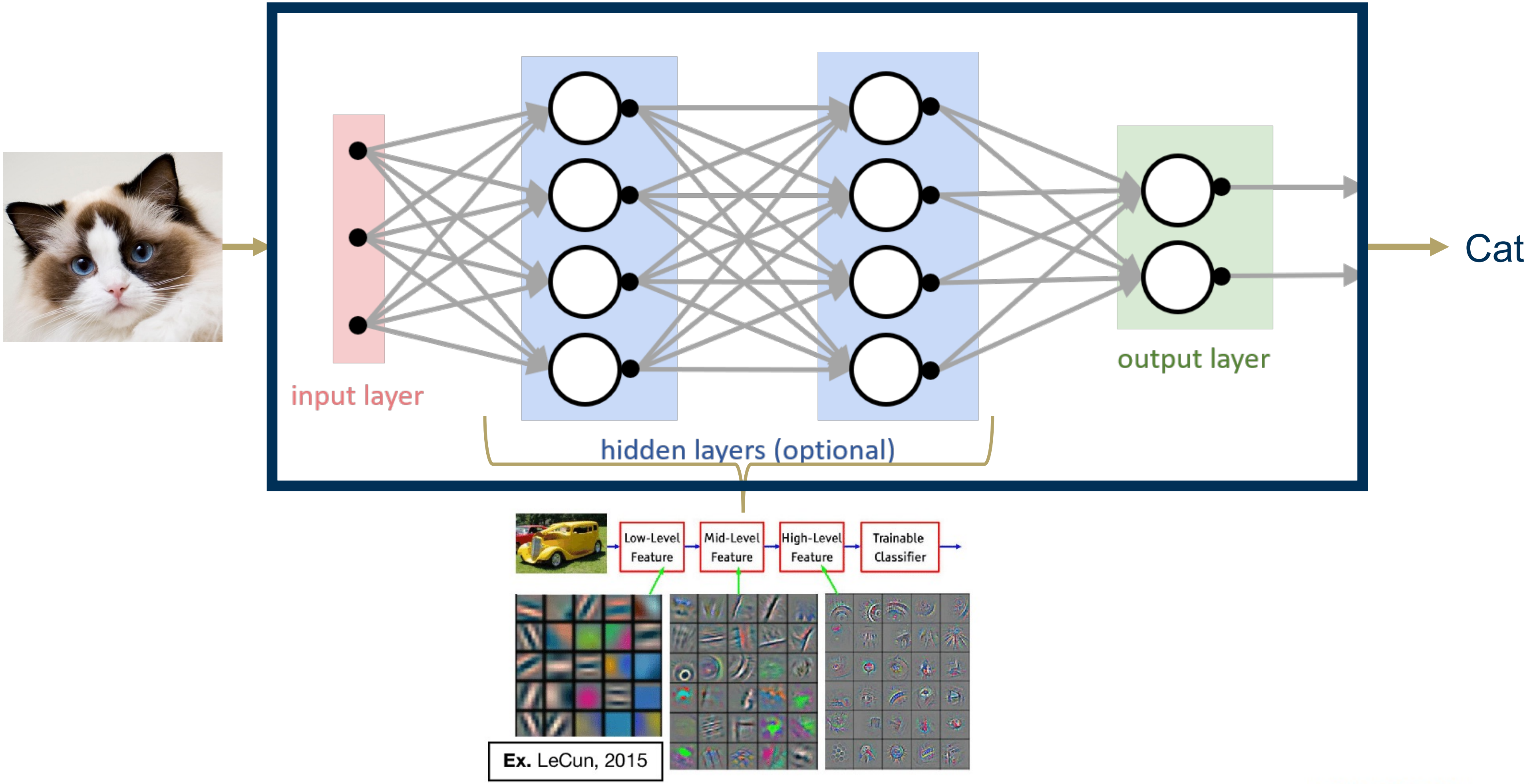
Typically, a neuron is part of a network organized in **layers**:

- An **input** layer (Layer 0)
- An **output** layer (Layer K)
- Zero or more **hidden** (middle) layers (Layers $1 \ldots K-1$)

Deep Learning

Convolutional Neural Networks

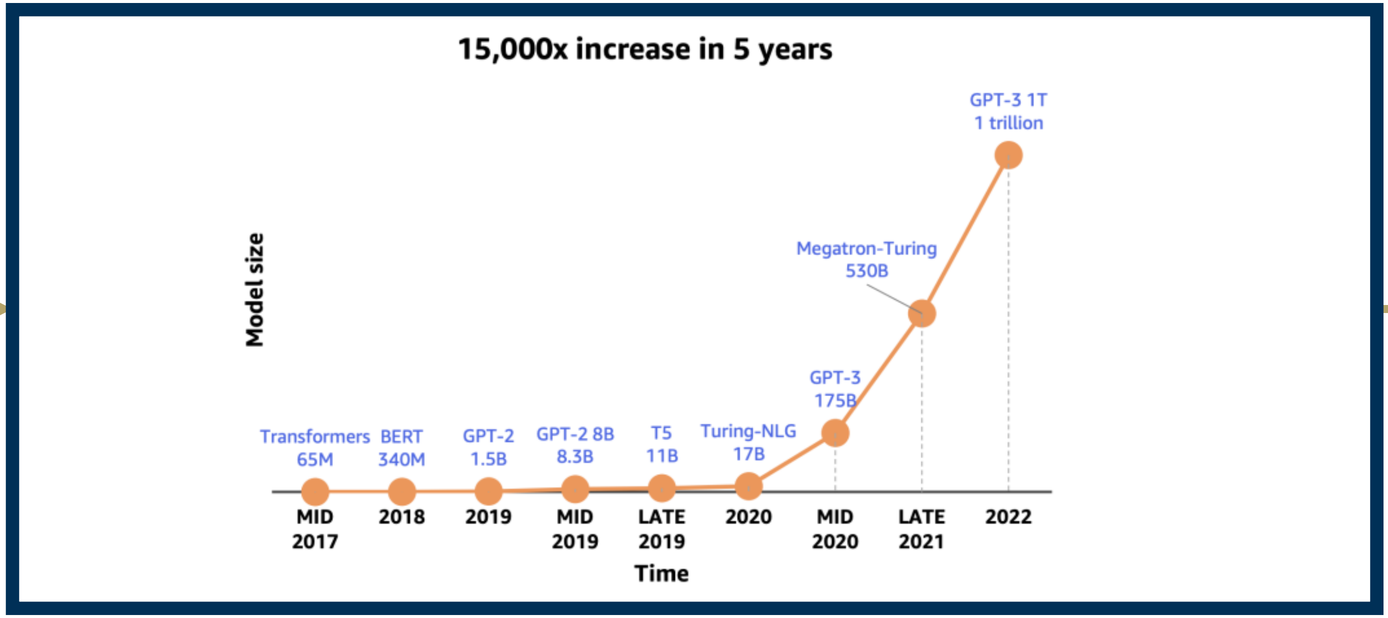
Stationary property of images allow for a small number of convolution kernels



Deep Deep Deep Deep Deep … Learning

Recent Advancements

Transformers, Large Language Models and Foundation Models



Cat

Primary reasons for advancements:

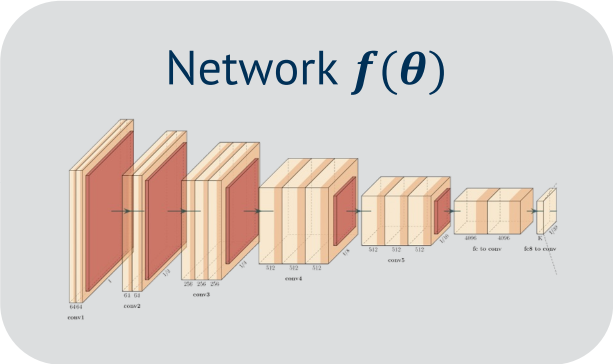
1. Expanded interests from the research community
2. Computational resources availability
3. **Big data availability**

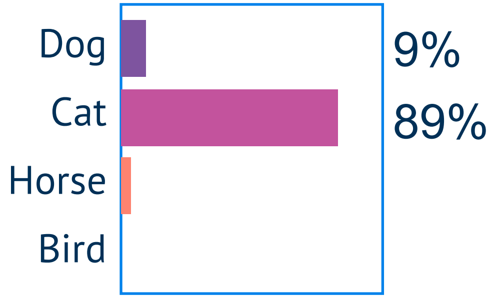
Deep Learning at Inference

Classification

Given : One network, One image. Required: Class Prediction

x





If $x \in \chi$, the data is **not novel**

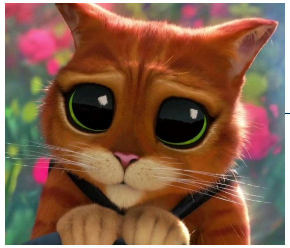
$$\hat{y} = f(x)$$
$$y = argmax_i\ \hat{y}$$
$$p(\hat{y}) = T(f(x))$$

$\hat{y}$ = Logits
y = Predicted Class
$p(\hat{y})$ = Probabilities
$f(\cdot)$ = Trained Network
χ = Training data

Deep Learning at Inference

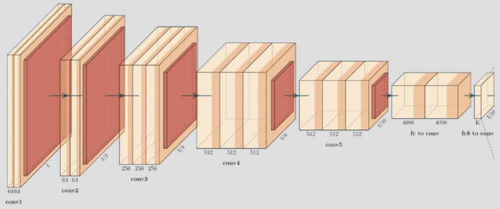
Robust Classification in Deep Networks

Deep learning robustness: Correctly predict class even when data is novel

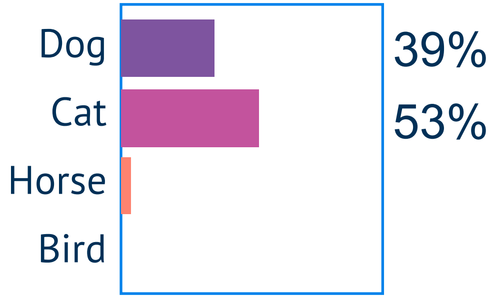


x'

Network $\boldsymbol{f(\theta)}$



Predicted Class Probability



If $x \in \chi$, the data is **novel**

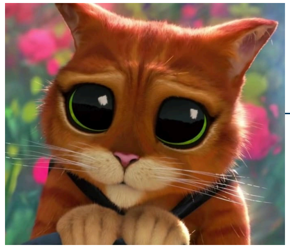
$$\hat{y} = f(x' + \epsilon)$$
$$y = argmax_i \, \hat{y}$$
$$p(\hat{y}) = T(f(x' + \epsilon))$$

$\hat{y}$ = Logits
y = Predicted Class
$p(\hat{y})$ = Probabilities
$f(\cdot)$ = Trained Network
χ = Training data
ϵ = Noise

Deep Learning at Inference

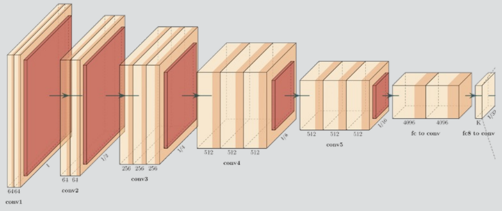
Robust Classification in Deep Networks

Deep learning robustness: Correctly predict class even when data is novel

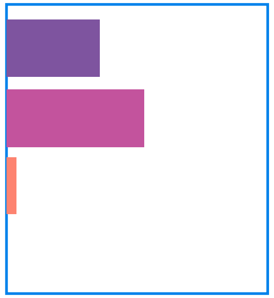


x'

Network $\boldsymbol{f}(\boldsymbol{\theta})$



Predicted Class Probability



To achieve robustness at Inference, we need the following:

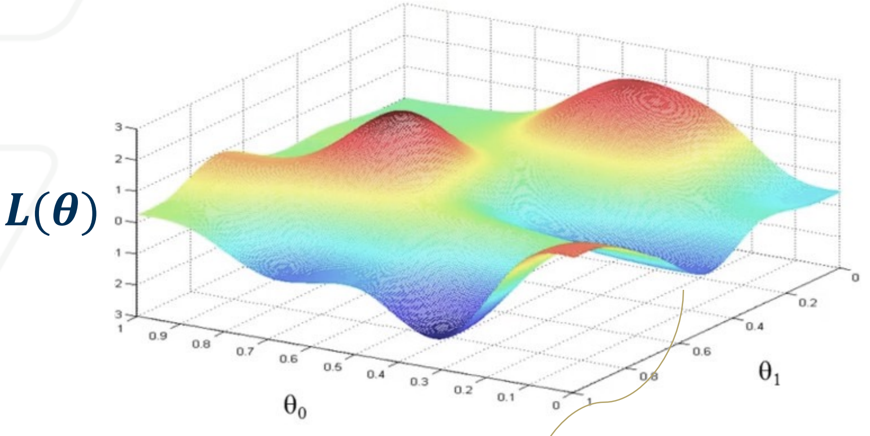
- **Information** provided by the novel data as **a function of training distribution**
- Methodology to **extract information** from novel data
- **Techniques** that utilize the information from novel data

Why is this Challenging?

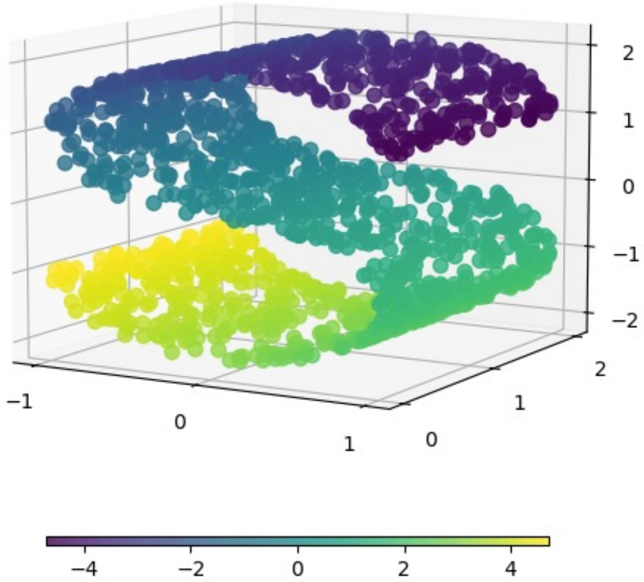
Challenges at Inference

A Quick note on Manifolds..

Manifolds are compact topological spaces that allow exact mathematical functions



Toy visualizations generated using functions (and thousands of generated data points)

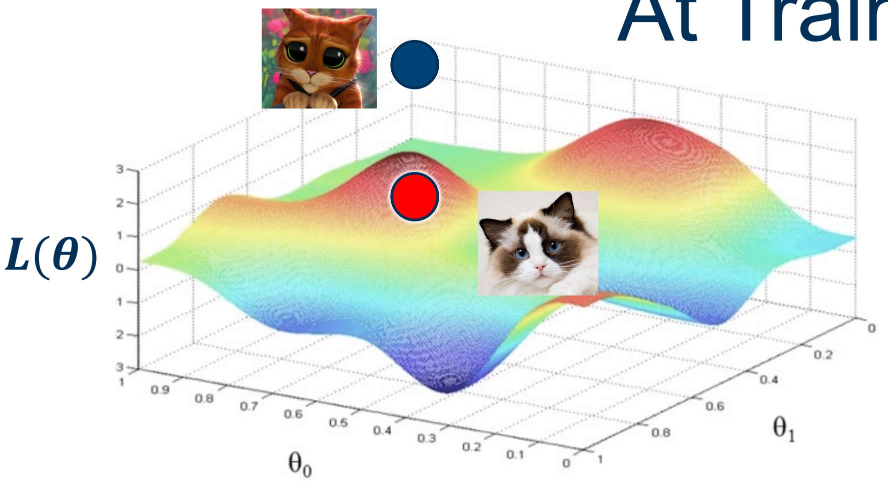


Real data visualizations generated using dimensionality reduction algorithms (Isomap)

However, at inference only the test data point is available and the underlying structure of the manifold is unknown

At Inference

At Training



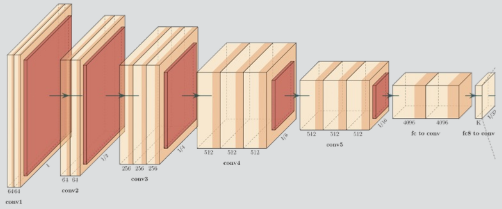
At training, we have access to all training data.

Information at Inference

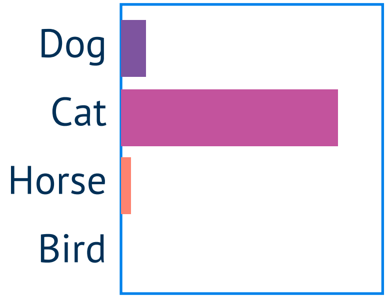
Fisher Information

Colloquially, Fisher Information is the "surprise" in a system that observes an event

Network $\boldsymbol{f(\theta)}$



Predicted Class Probability



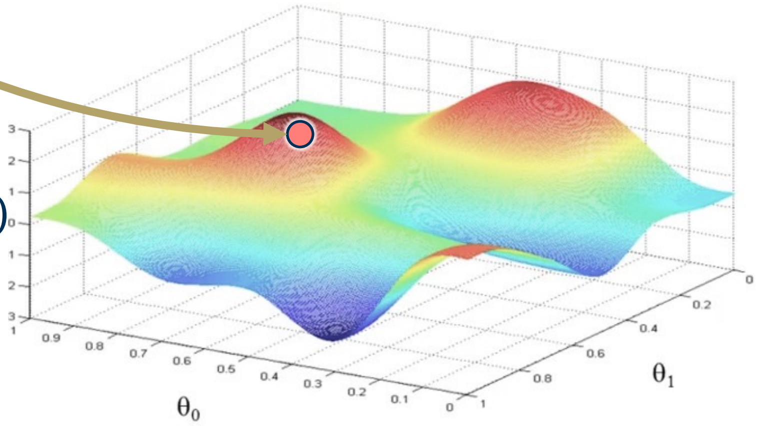
Fisher Information

$$I(\theta) = Var(\frac{\partial}{\partial\theta} l(\theta|x))$$

θ = Statistic of distribution

$\ell(\theta \mid x)$ = Likelihood function

$\boldsymbol{l(\theta|x)}$

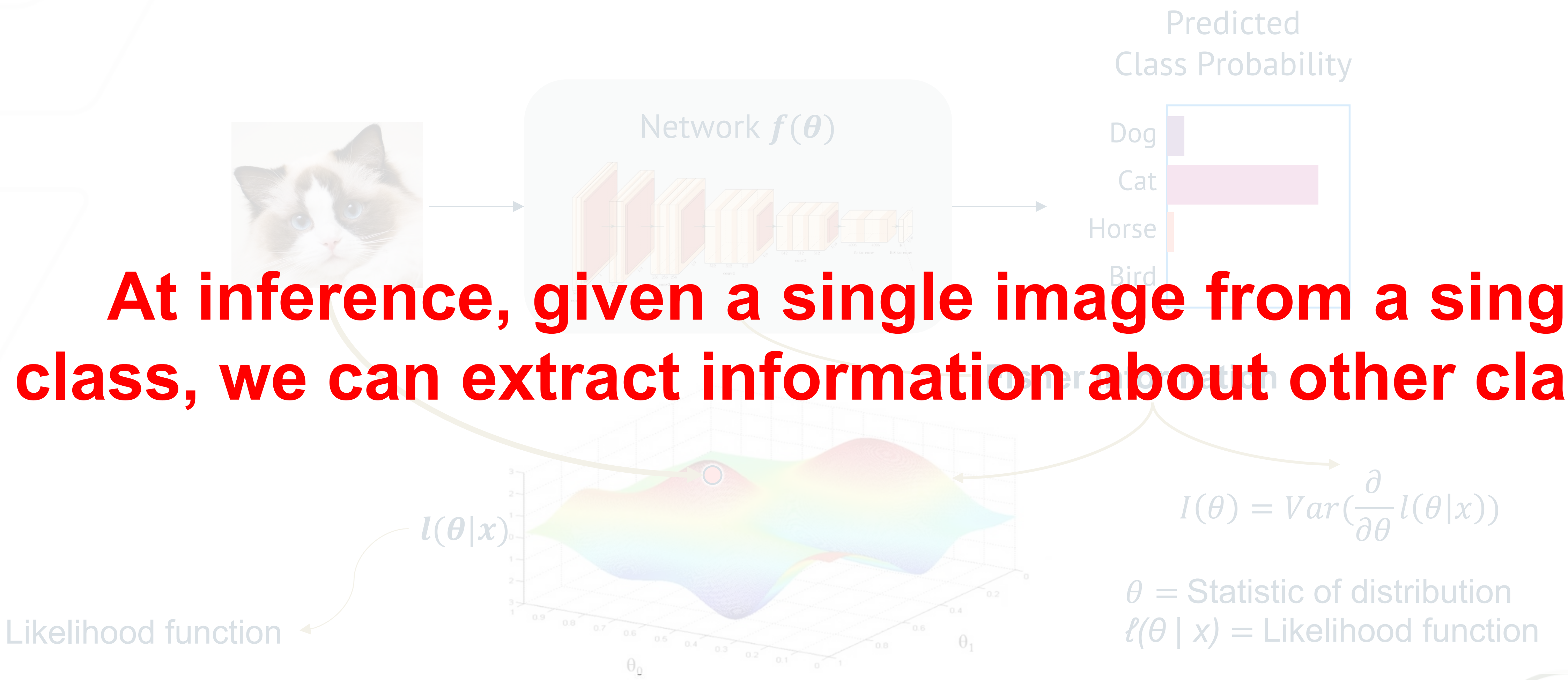


Likelihood function

[1] A good blogpost about Fisher Information: https://towardsdatascience.com/an-intuitive-look-at-fisher-information-2720c40867d8

Information at Inference

Information at Inference



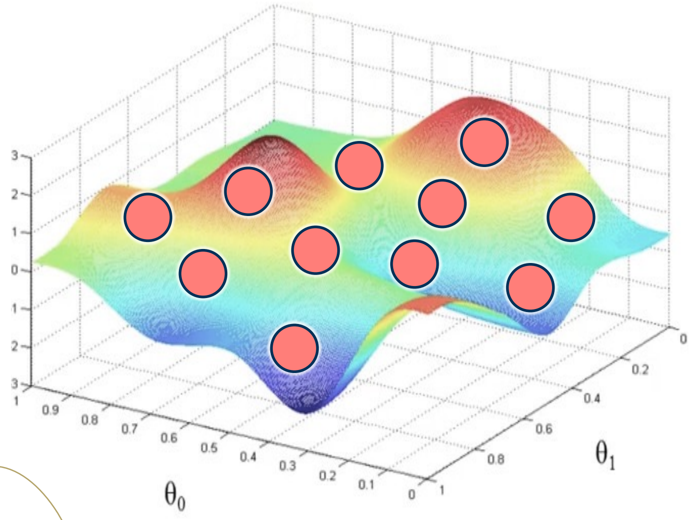
At inference, given a single image from a single class, we can extract information about other classes

Information at Inference

Gradients as Fisher Information

Gradients infer information about the statistics of underlying manifolds

$\boldsymbol{l(\theta|x)}$



Likelihood function instead of loss manifold

From before, $I(\theta) = Var(\frac{\partial}{\partial\theta} l(\theta|x))$

Using variance decomposition, $I(\theta)$ reduces to:

$I(\theta) = E[U_\theta U_\theta^T]$ where

$E[\cdot] = \text{Expectation}$

$U_\theta = \nabla_\theta l(\theta|x)$, Gradients w.r.t. the sample

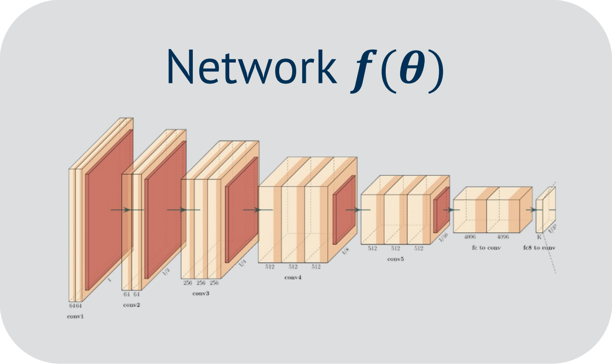
Hence, gradients draw information from the underlying distribution as learned by the network weights!

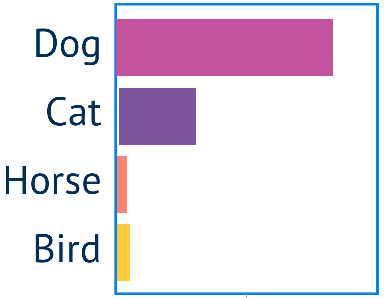
Kwon, Gukyeong, et al. "Backpropagated gradient representations for anomaly detection." *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer International Publishing, 2020.

Case Study: Gradients as Fisher Information in Explainability

Gradients infer information about the statistics of underlying manifolds

x

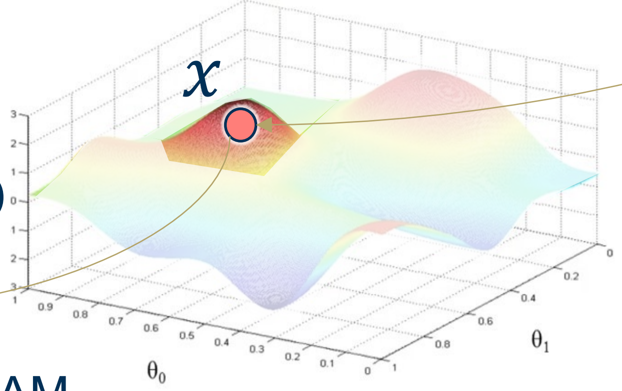




In this case, the image and its prediction extracts nose, mouth and jowl features.

Local information (specific to x) is sufficient!

$l(\theta|x)$



Feature attribution via GradCAM

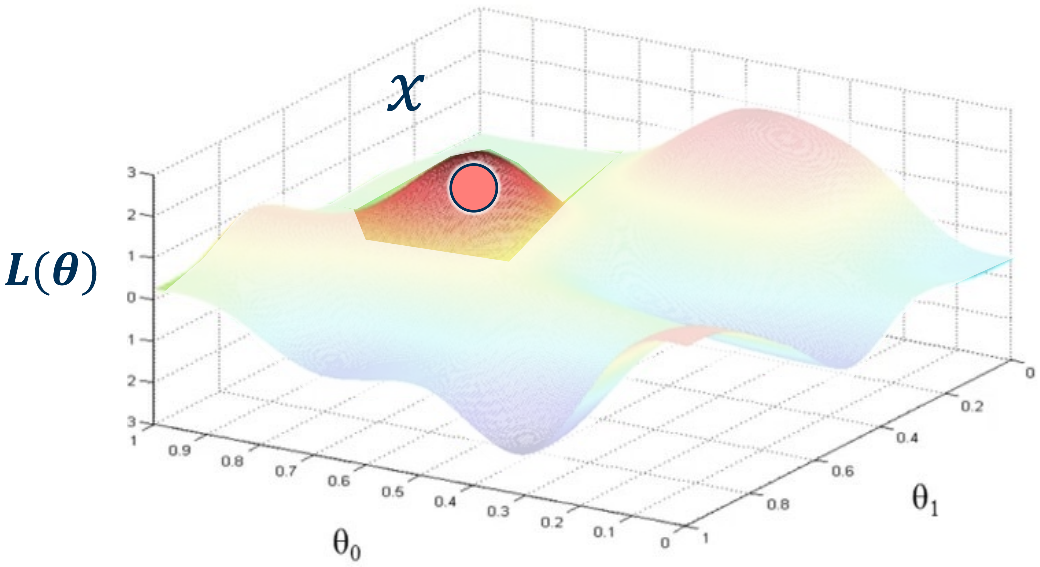
Hence, gradients draw information from the underlying distribution as learned by the network weights!

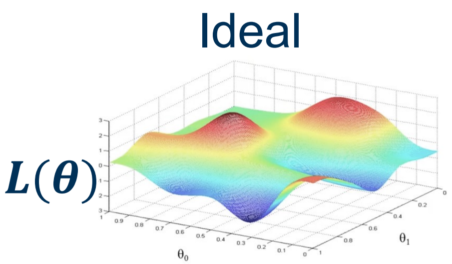
[1] A good blogpost about Fisher Information: https://towardsdatascience.com/an-intuitive-look-at-fisher-information-2720c40867d8

Gradients at Inference

Local Information

Gradients provide local information around the vicinity of x, even if x is novel. This is because x projects on the learned knowledge



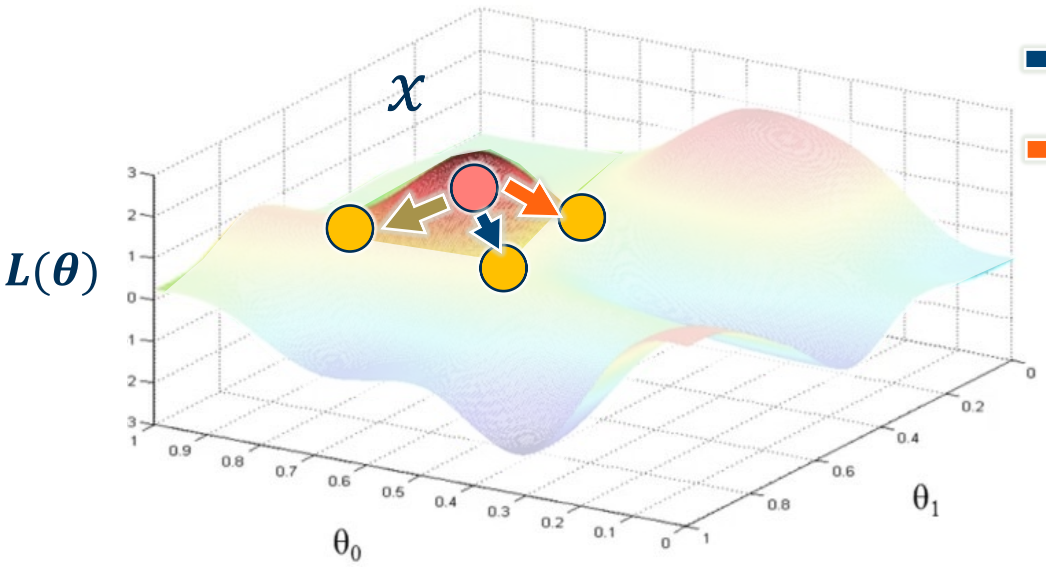


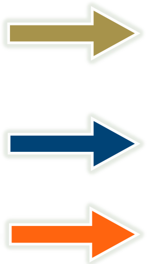
$\alpha \nabla_\theta L(\theta)$ provides local information up to a small distance α away from x

Gradients at Inference

Direction of Steepest Descent

Gradients allow choosing the fastest direction of descent given a loss function $L(\theta)$



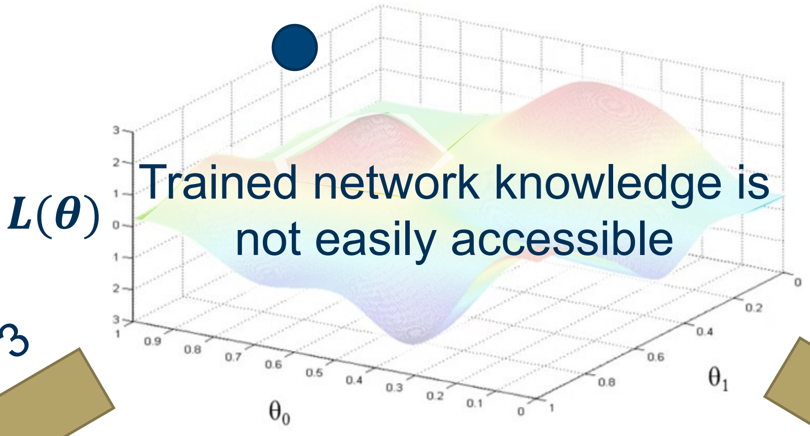


Which direction should we optimize towards (knowing only the local information)?

Negative of the gradient provides the **descent direction** towards the local minima, as measured by $L(\theta)$

Gradients at Inference

To Characterize the Novel Data at Inference



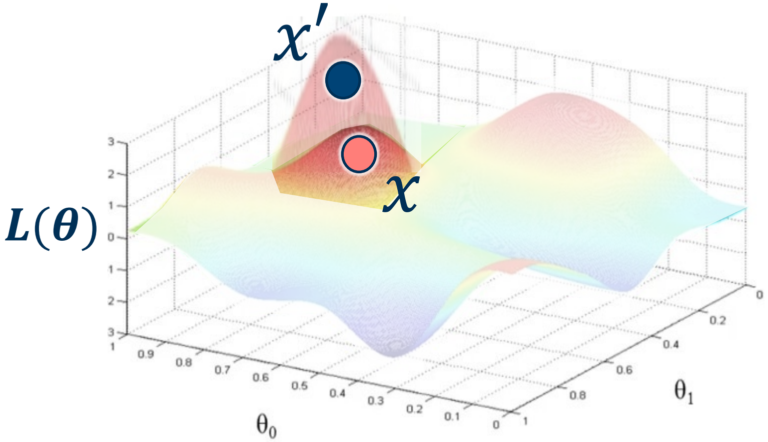
At Inference

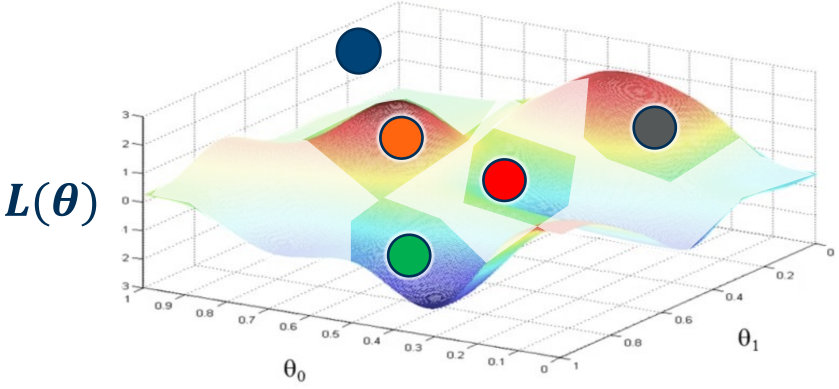
Part 2, 3

Part 4

Counterfactual and Contrastive Representations using Gradients

Representation Traversal using Interventions





Robust Neural Networks

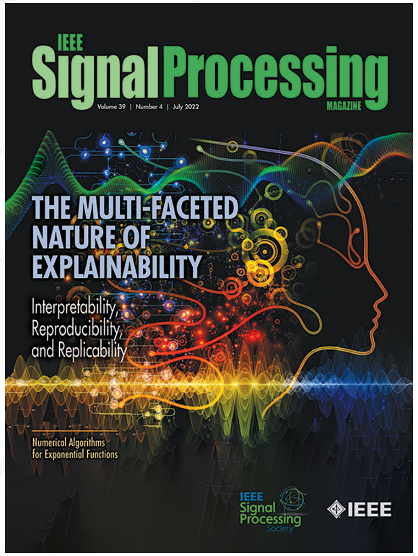
Part 2: Explainability at Inference

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**
  - Visual Explanations
  - Gradient-based Explanations
  - GradCAM
  - CounterfactualCAM
  - ContrastCAM
- Part 3: Uncertainty at Inference
- Part 4: Intervenability at Inference
- Part 5: Conclusions and Future Directions



Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

Mohit Prabhushankar, PhD
Postdoc

Ghassan AlRegib, PhD
Professor

SCAN ME

Explanations

Visual Explanations

SCAN ME

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

- **Explanations are defined as a set of rationales used to understand the reasons behind a decision**
- **If the decision is based on visual characteristics within the data, the decision-making reasons are visual explanations**

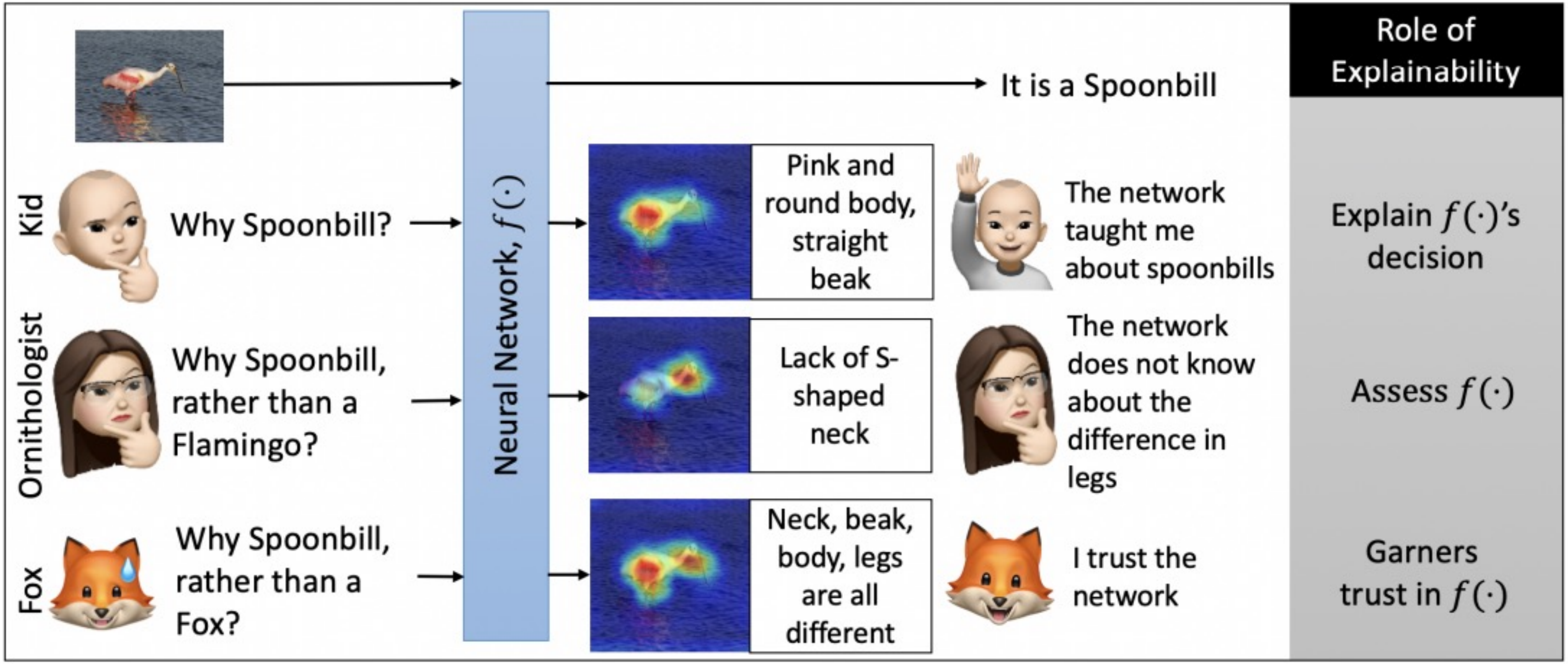
Bullmastiff	Why Bullmastiff?	What if Bullmastiff was not in the image?	Why Bullmastiff, rather than a Boxer?

AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks: Towards relevant and contextual explanations. *IEEE Signal Processing Magazine*, *39*(4), 59-72.

Explanations

Role of Explanations – context and relevance

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations



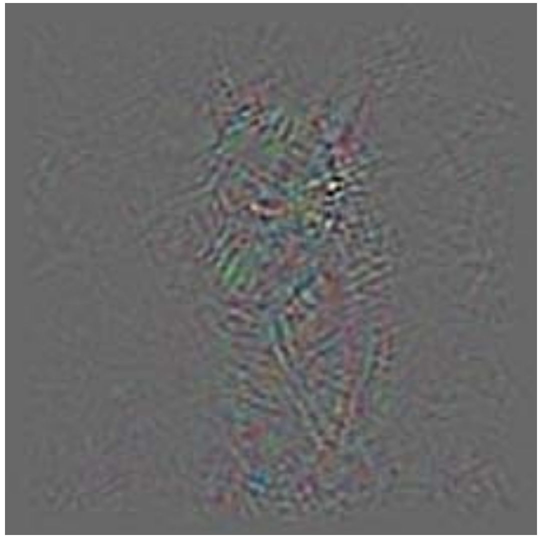
AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks: Towards relevant and contextual explanations. *IEEE Signal Processing Magazine*, *39*(4), 59-72.

Explanations
Gradient-based Explanations

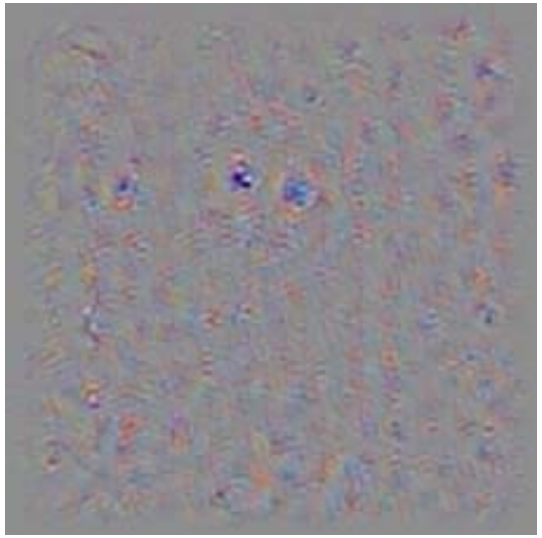
Gradients provide a one-shot means of perturbing the input that changes the output; They provide pixel-level importance scores

Input

Vanilla Gradients



Deconvolution Gradients



Guided Backpropagation

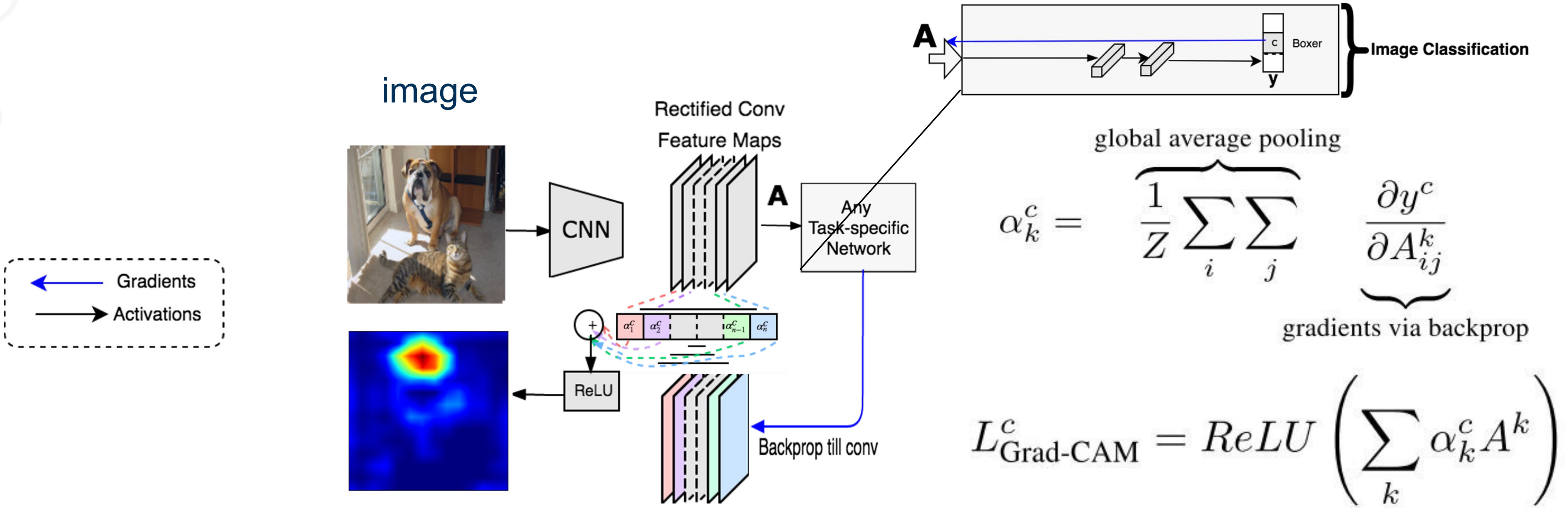
However, localization remains an issue

Springenberg, Dosovitskiy, et al., Striving for Simplicity: The all convolutional net, 2015

Gradient and Activation-based Explanations

GradCAM

Grad-CAM uses the gradient information flowing into the last convolutional layer of the CNN to assign importance values to each activation for a particular decision of interest.



Grad-CAM (up-sampled to original image dimension)

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

$$L_{\text{Grad-CAM}}^c = ReLU \underbrace{\left(\sum_k \alpha_k^c A^k \right)}_{\text{linear combination}}$$

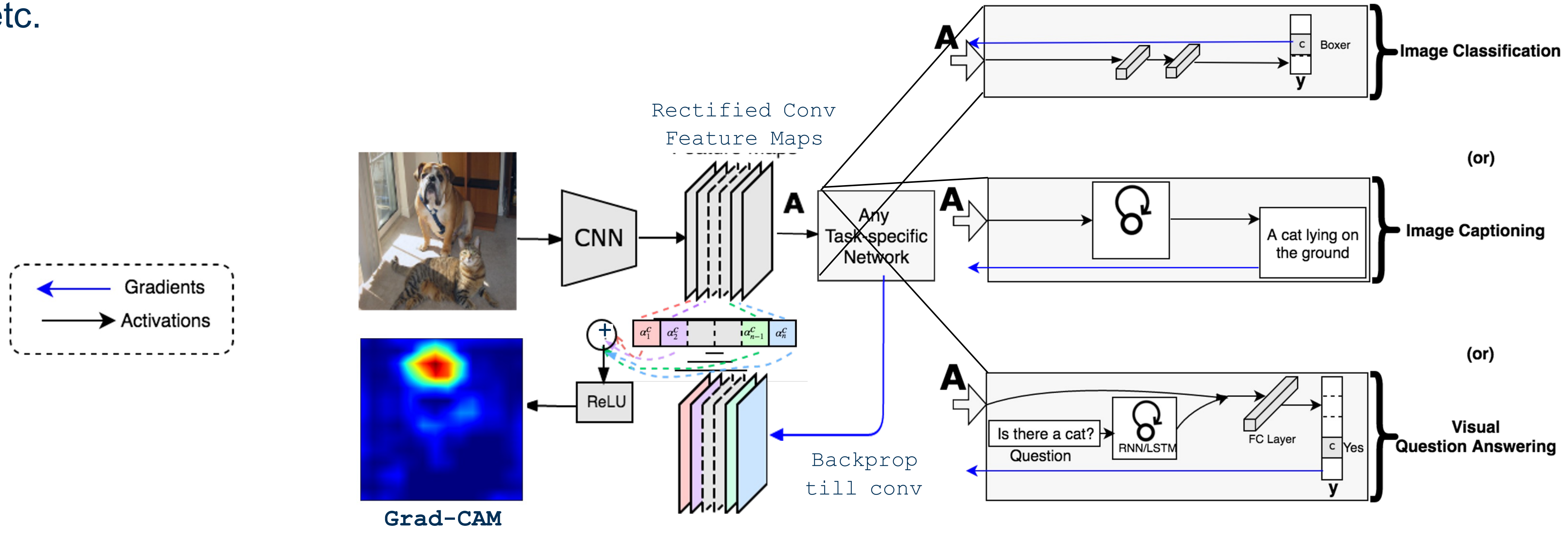
Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

Gradient and Activation-based Explanations

GradCAM

Grad-CAM generalizes to any task:

- Image classification
- Image captioning
- Visual question answering
- etc.



Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

Gradient and Activation-based Explanations

Explanatory Paradigms

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

GradCAM provides answers to *'Why P?'* questions. But different stakeholders require relevant and contextual explanations

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

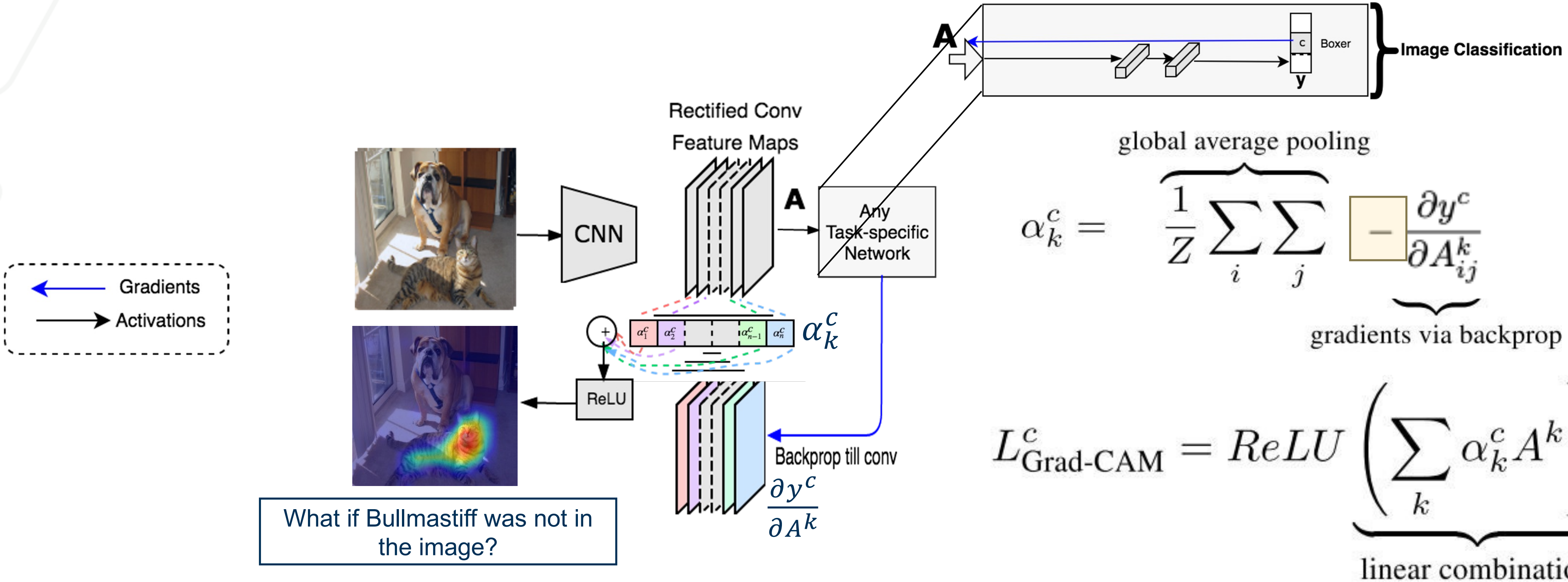
AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks: Towards relevant and contextual explanations. *IEEE Signal Processing Magazine*, *39*(4), 59-72.

Gradient and Activation-based Explanations

CounterfactualCAM: What if this region were absent in the image?

SCAN ME

In GradCAM, global average pool the negative of gradients to obtain α^c for each kernel k



$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} - \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

$$L^c_{\text{Grad-CAM}} = ReLU \underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

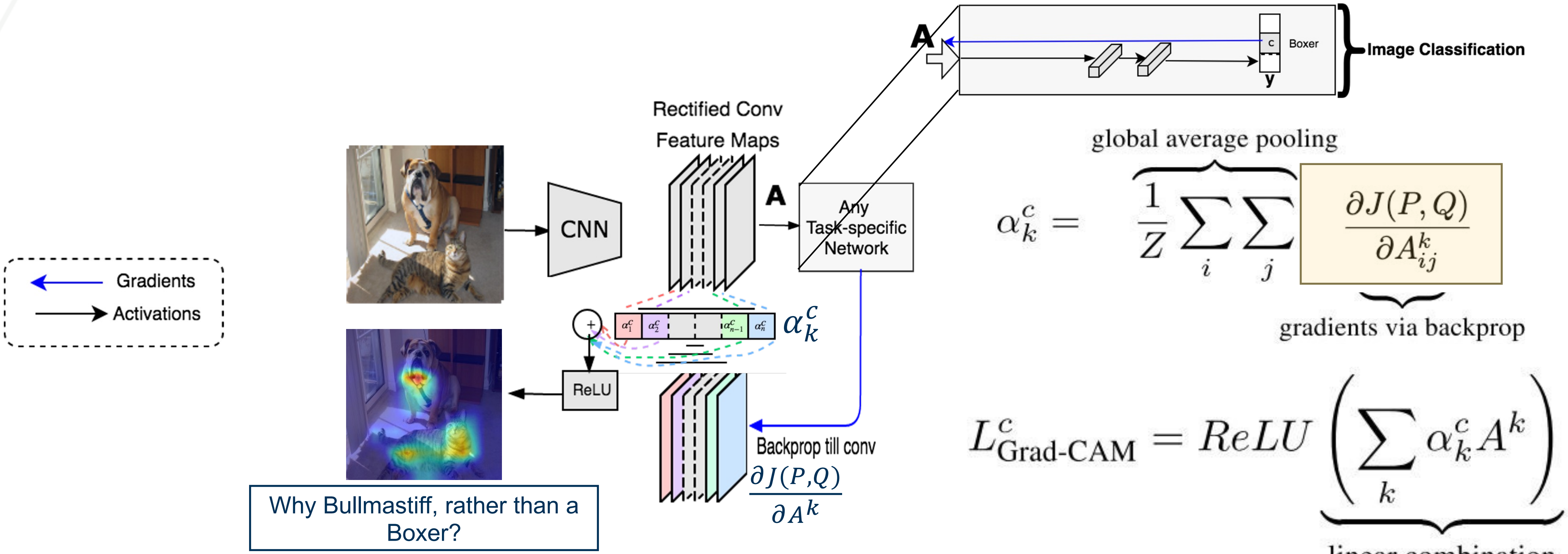
Negating the gradients effectively removes these regions from analysis

Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

Gradient and Activation-based Explanations

ContrastCAM: Why P, rather than Q?

In GradCAM, backward pass the loss between predicted class P and some contrast class Q to last conv layer



$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial J(P,Q)}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

$$L_{\text{Grad-CAM}}^c = ReLU \underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

Contrast-CAM

Backpropagating the loss highlights the differences between classes P and Q.

Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

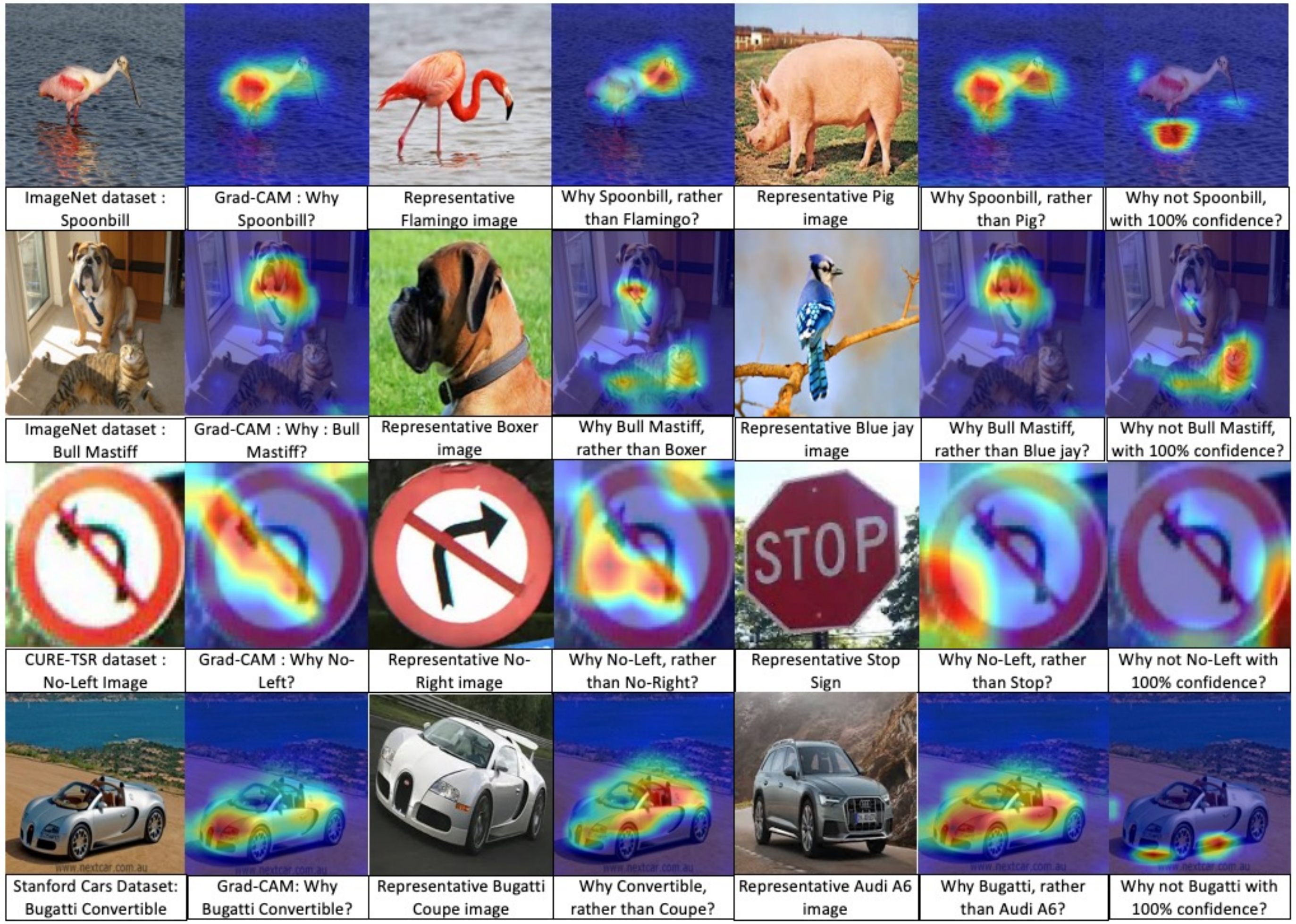
Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

SCAN ME

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

Input Image | Grad-CAM | Contrast 1 | Contrastive Explanation 1 | Contrast 2 | Contrastive Explanation 2



Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

OLIVES @GeorgiaTech

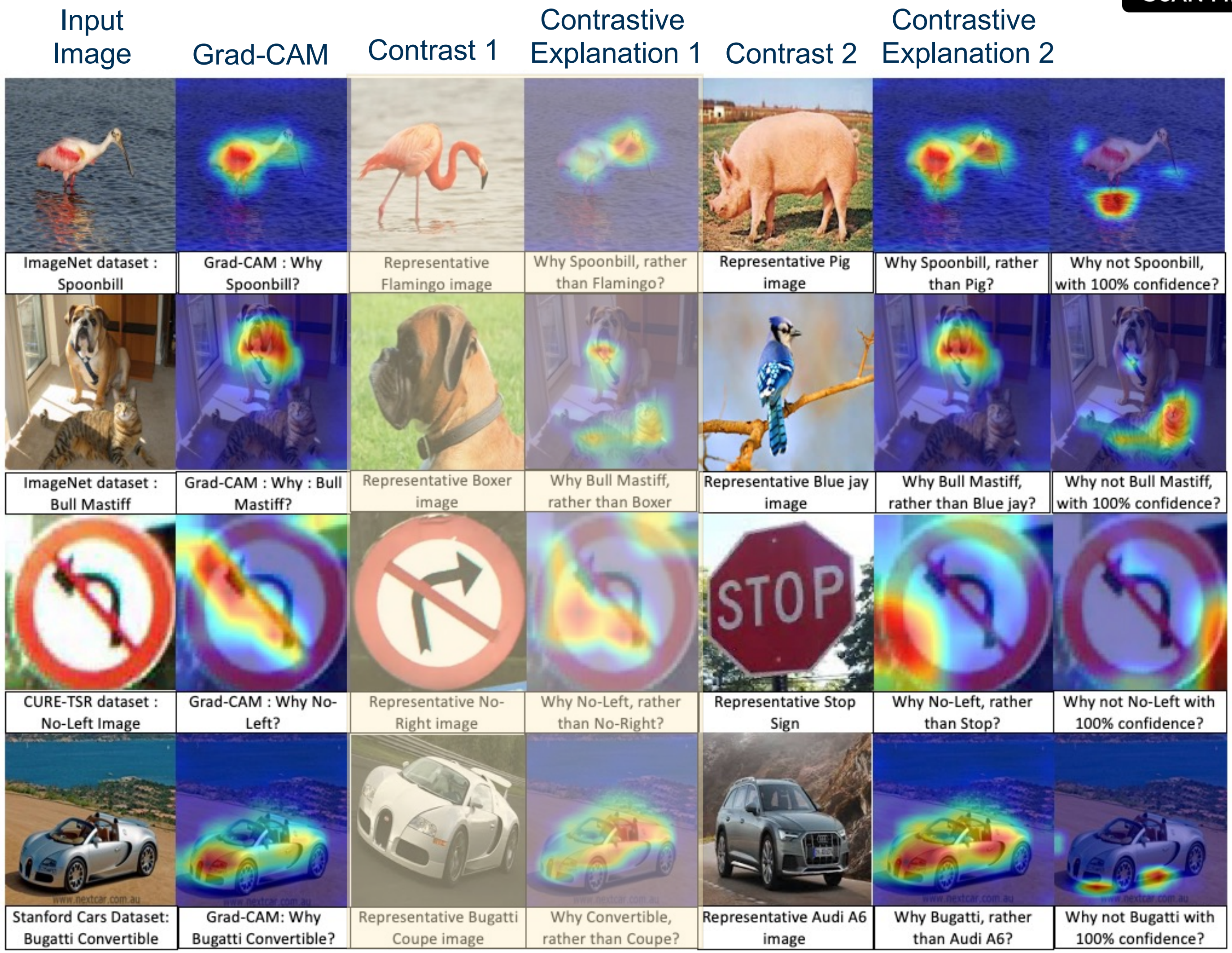
Georgia Tech

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME



Human Interpretable

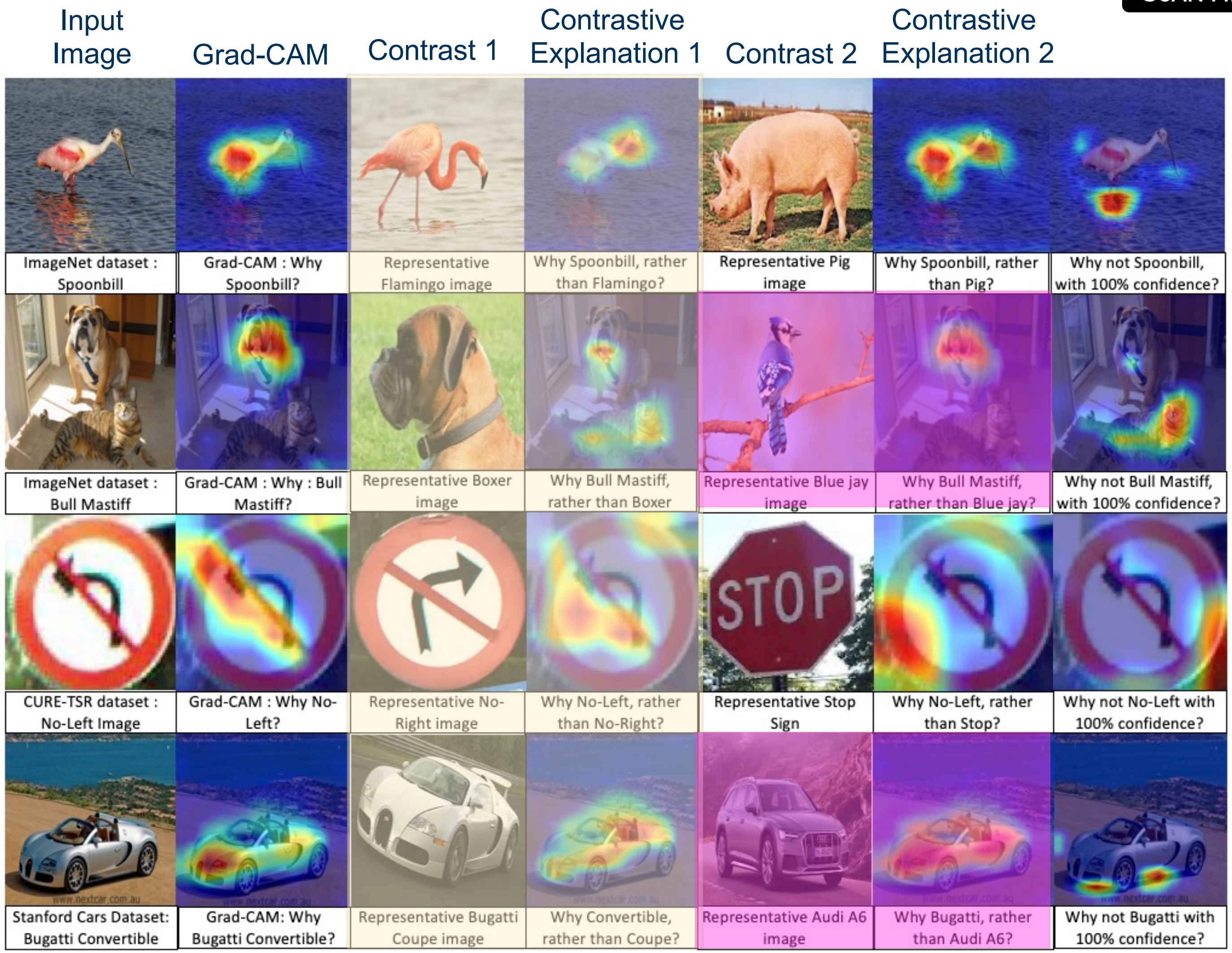
Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

SCAN ME

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations



Human Interpretable

Same as Grad-CAM

Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

OLIVES @GeorgiaTech

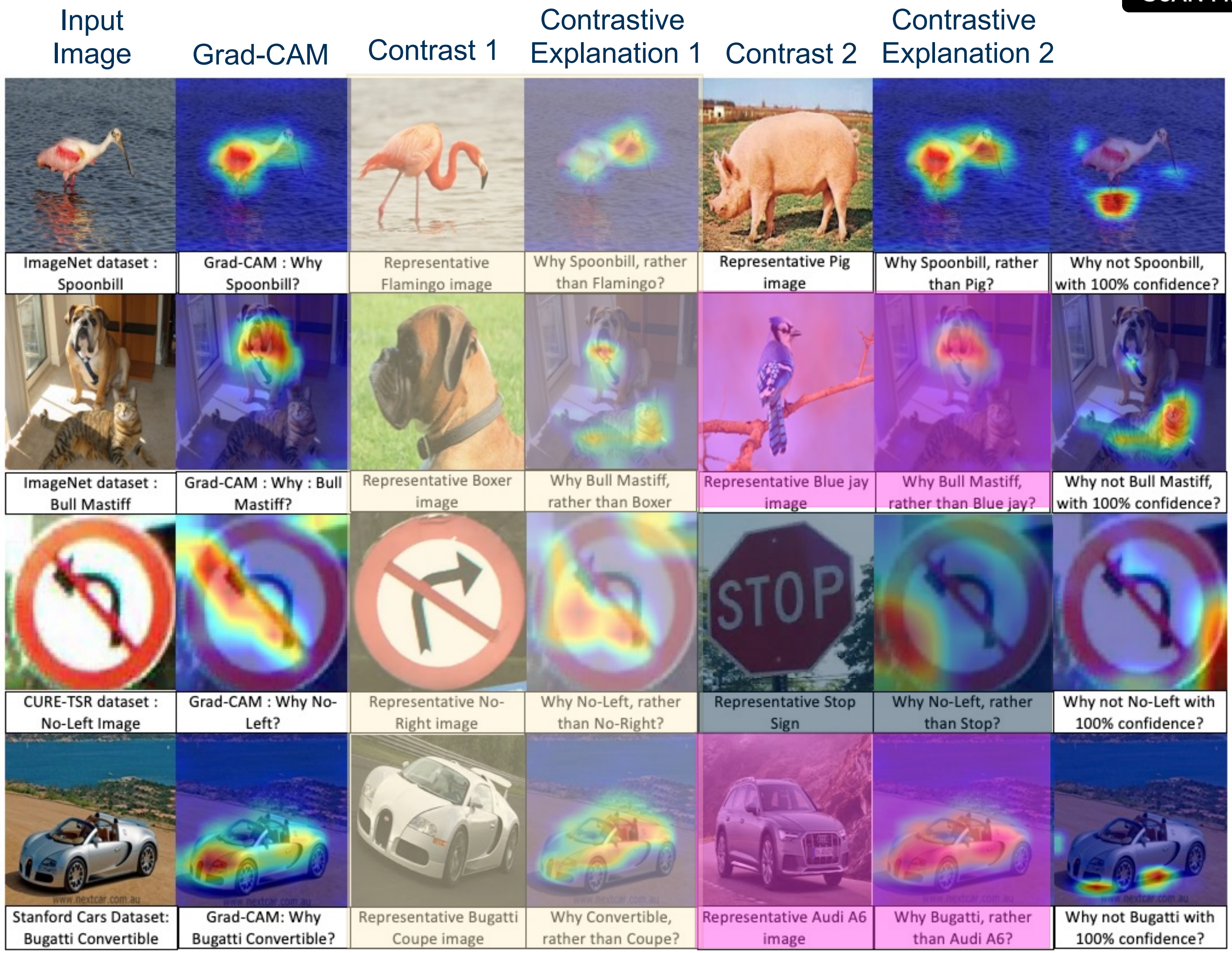
Georgia Tech

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

SCAN ME

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations



Human Interpretable

Same as Grad-CAM

Not Human Interpretable

Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

OLIVES @GeorgiaTech

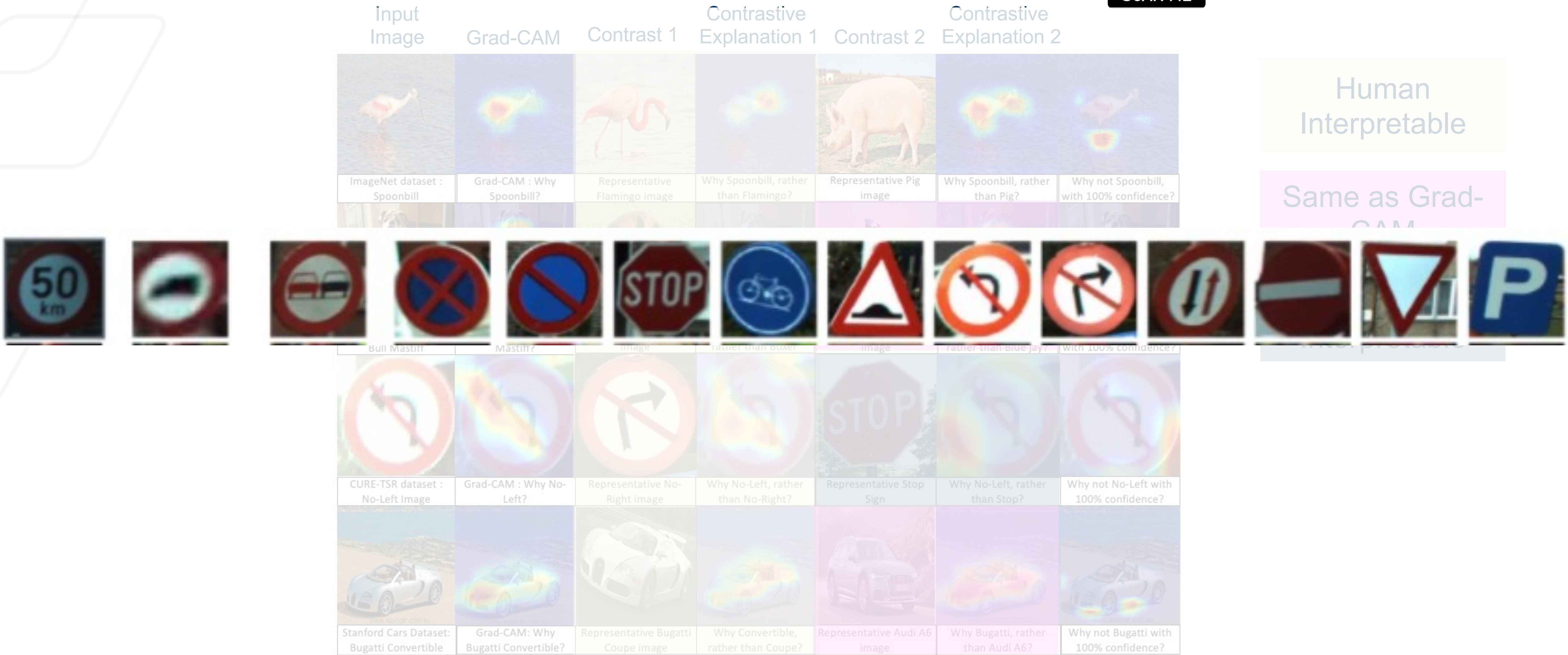
Georgia Tech

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME



Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

OLIVES @GeorgiaTech

Georgia Tech

Robust Neural Networks

Part 3: Uncertainty at Inference

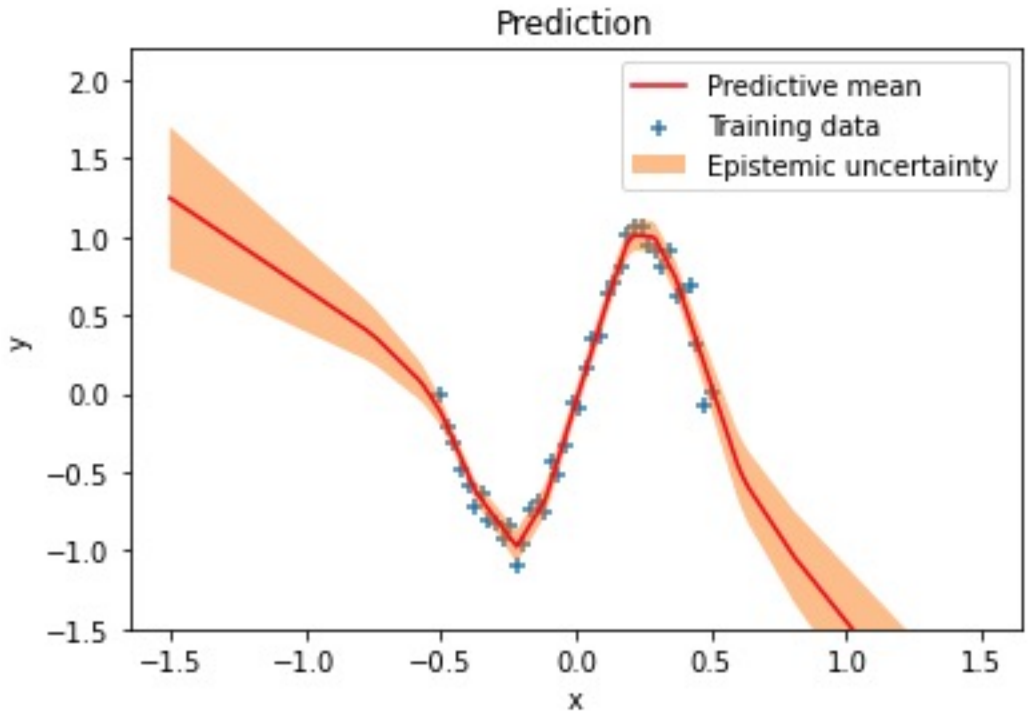
Objective

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 is a model knowing that it does not know



A simple example:

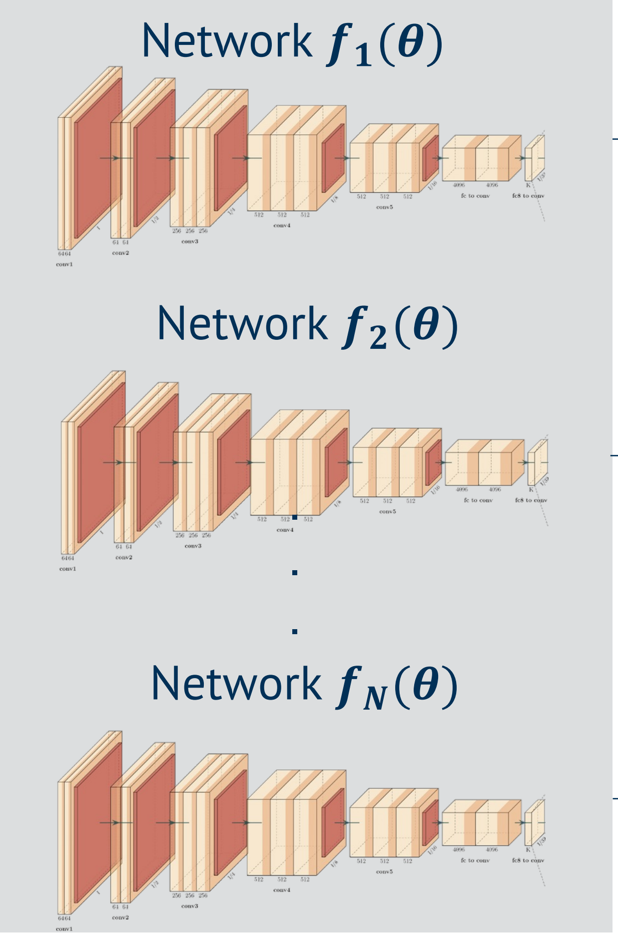
- When training data is **available: Less uncertainty**
- When training data is **unavailable: More uncertainty**

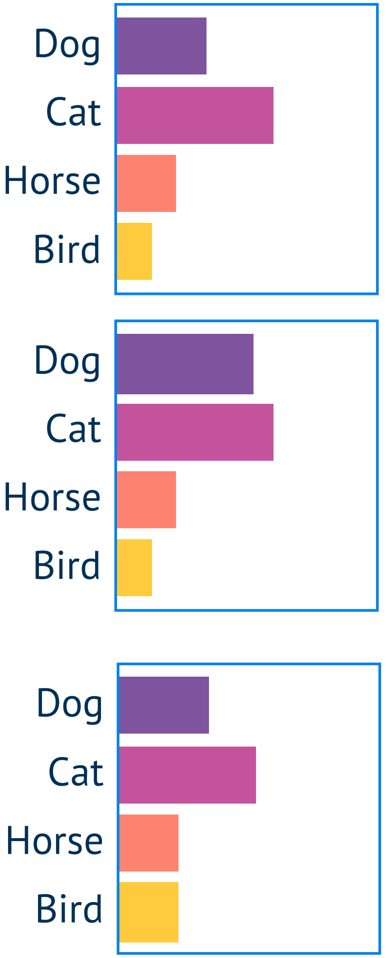
http://krasserm.github.io/2020/09/25/reliable-uncertainty-estimates/

Uncertainty

Uncertainty Quantification in Neural Networks

Via Ensembles[1]





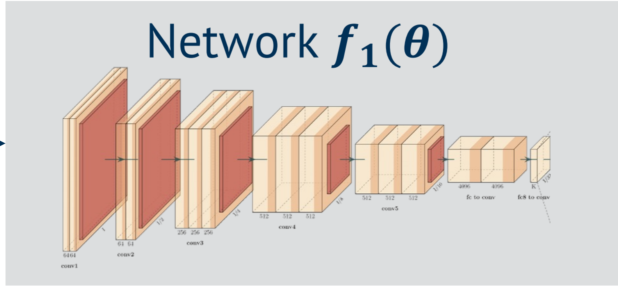
Variation within outputs $Var(y)$ is the uncertainty. Commonly referred to as **Prediction Uncertainty.**

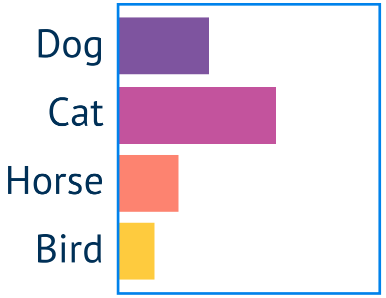
[1] Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles." *Advances in neural information processing systems* 30 (2017).

Uncertainty

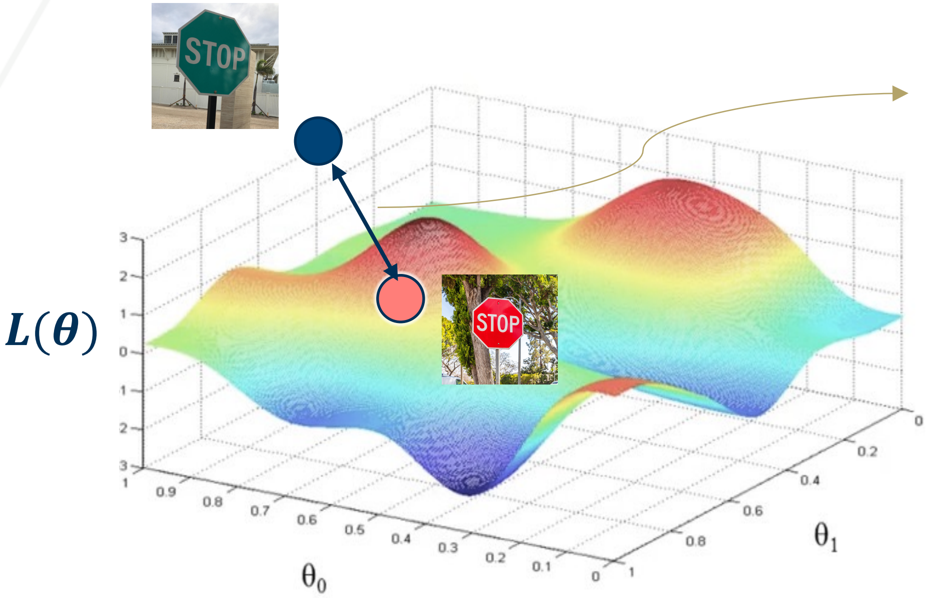
Uncertainty Quantification in Neural Networks

Via Single pass methods[1]





Uncertainty quantification using a single network and a single pass



Calculate distance from some trained clusters

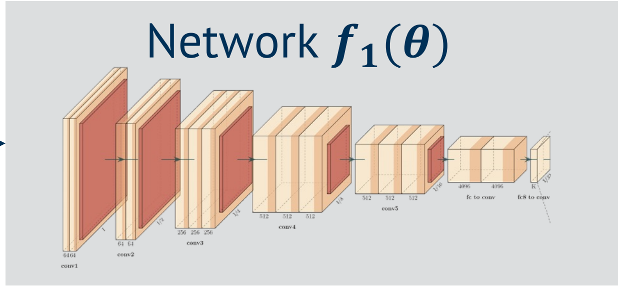
Does not require multiple networks!

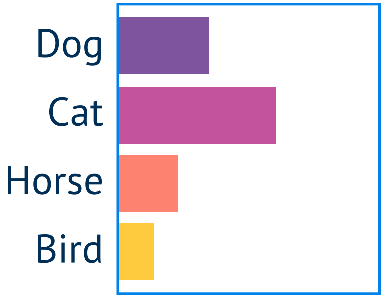
[1Van Amersfoort, J., Smith, L., Teh, Y. W., & Gal, Y. (2020, November). Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning* (pp. 9690-9700). PMLR.

Uncertainty

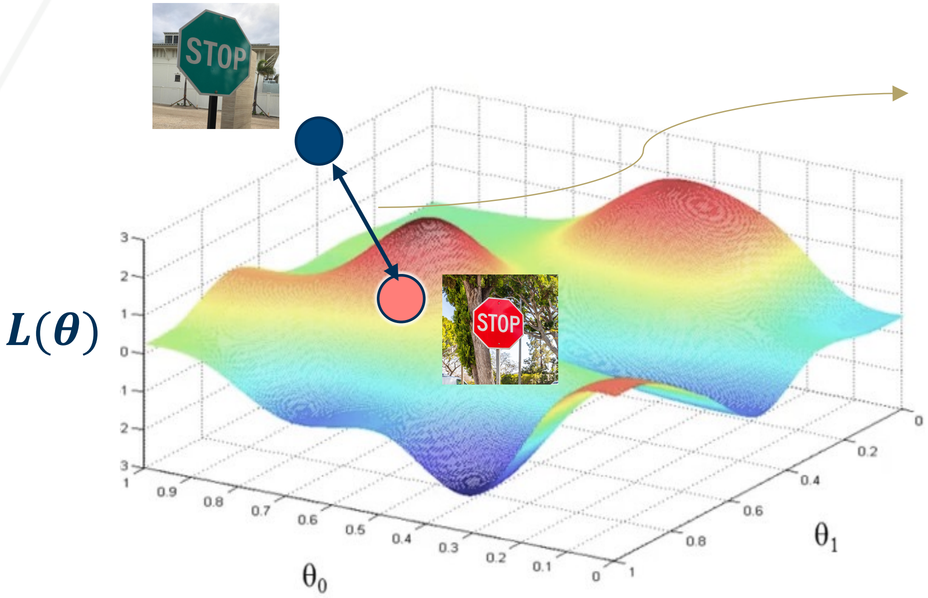
Gradients as Single pass Features

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!

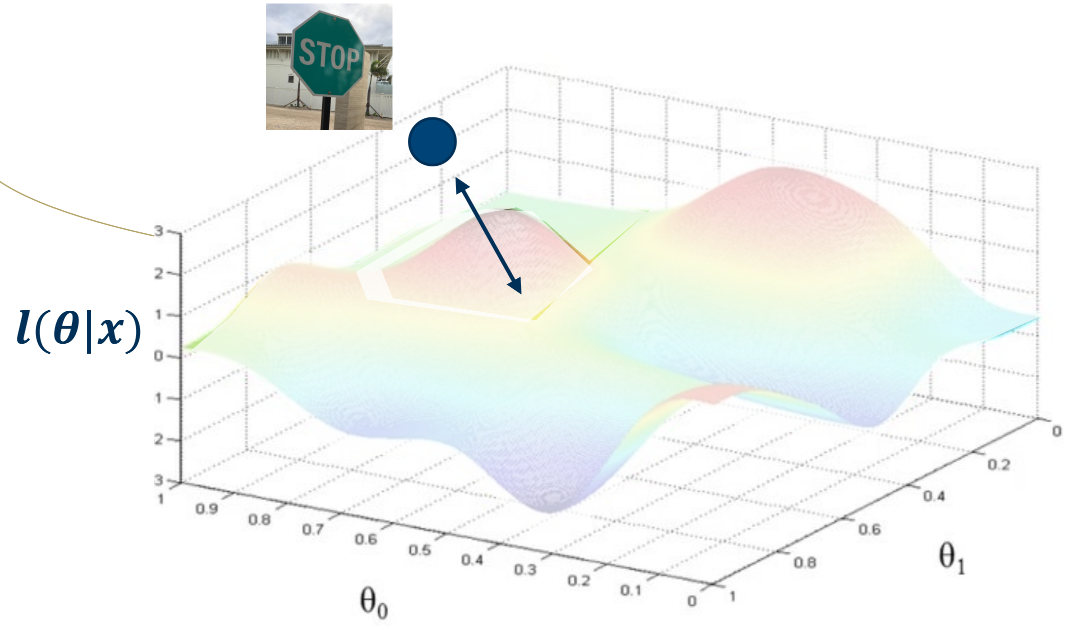
Challenge: Class and prediction cannot be trusted!

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. **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

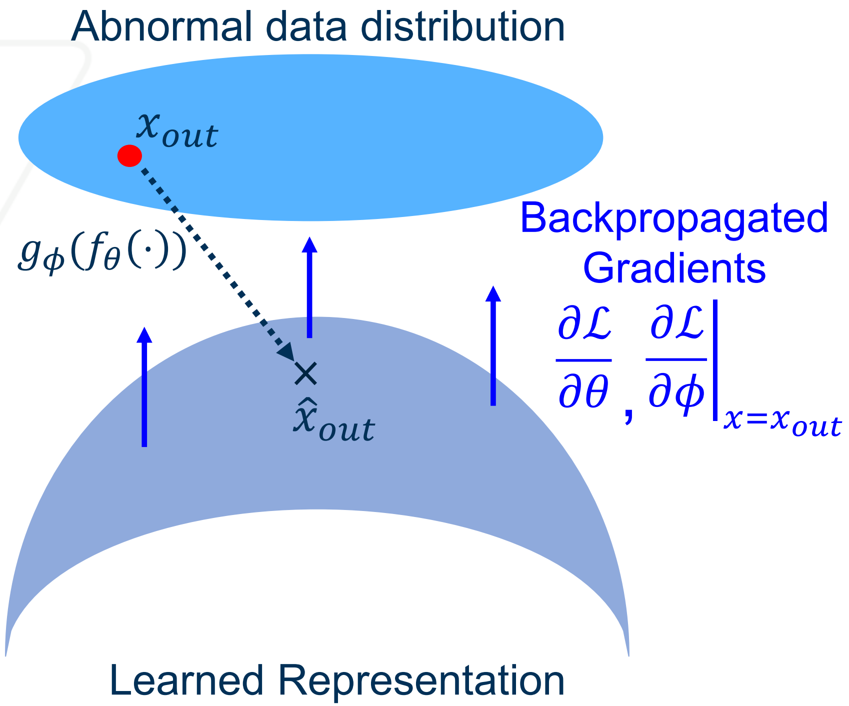
Ghassan AlRegib, PhD
Professor

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

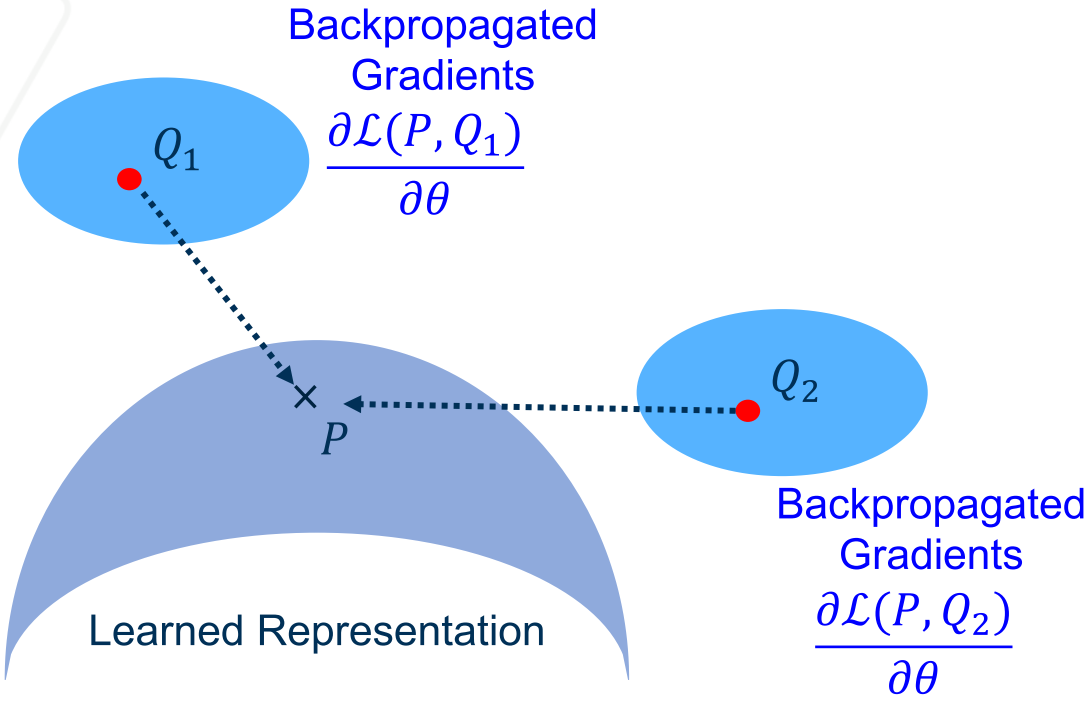
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

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.

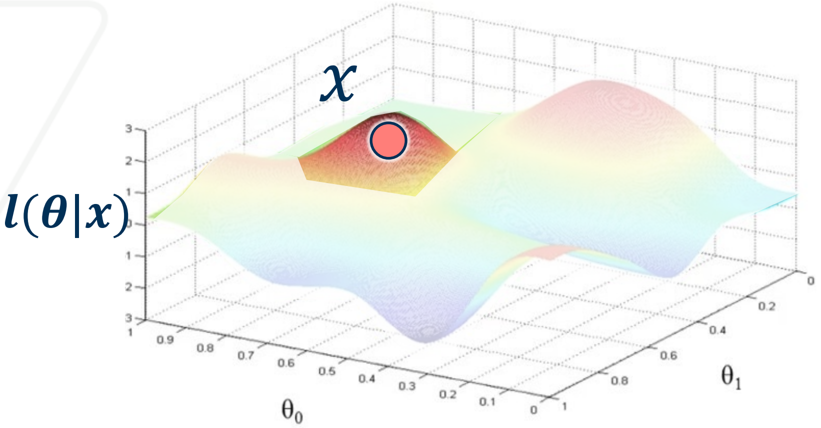
Toy Manifold Example

What is uncertainty?

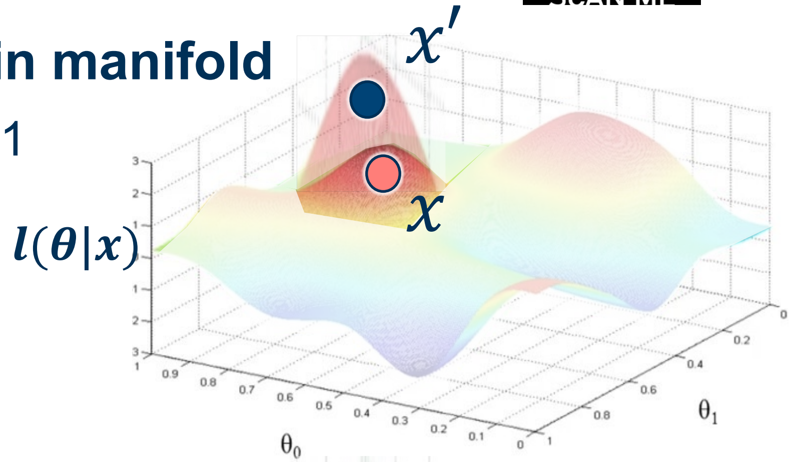
Probing the Purview of Neural Networks via Gradient Analysis

Gradients represent the local required change in manifold

Similar to introspective learning!

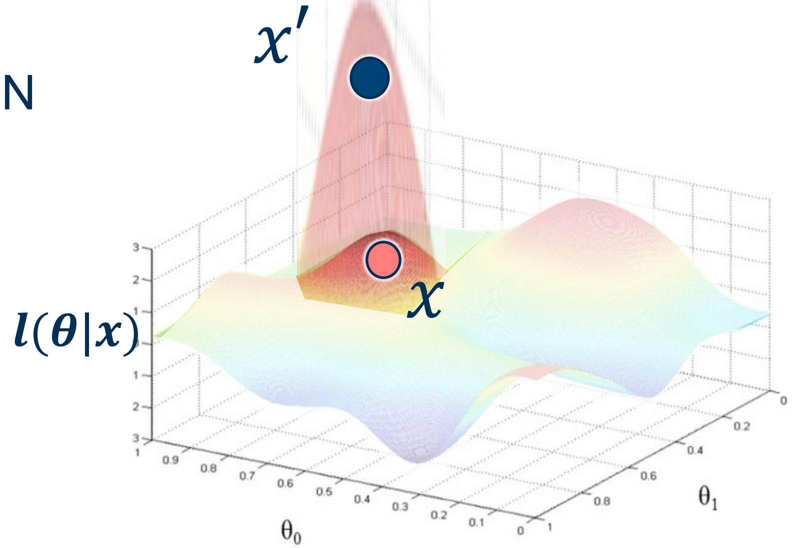


Contrast class 1



.
.
.

Contrast class N



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

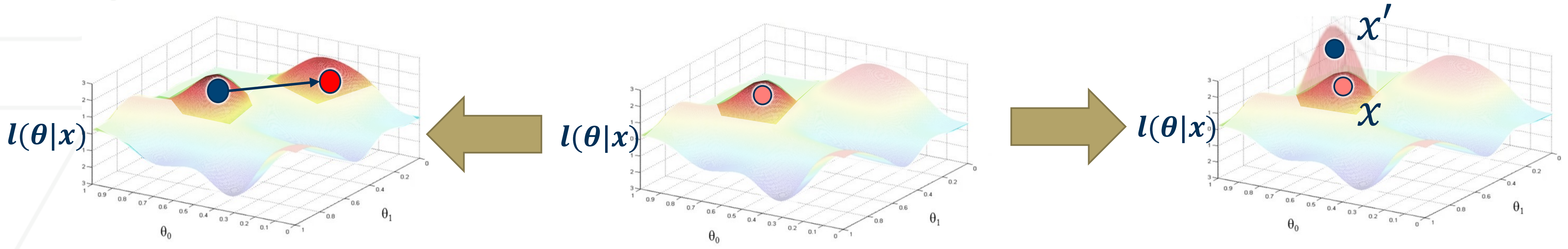
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Toy Manifold Example

How is this different from Explainability?

Part 3: Explainability

Part 4: Uncertainty



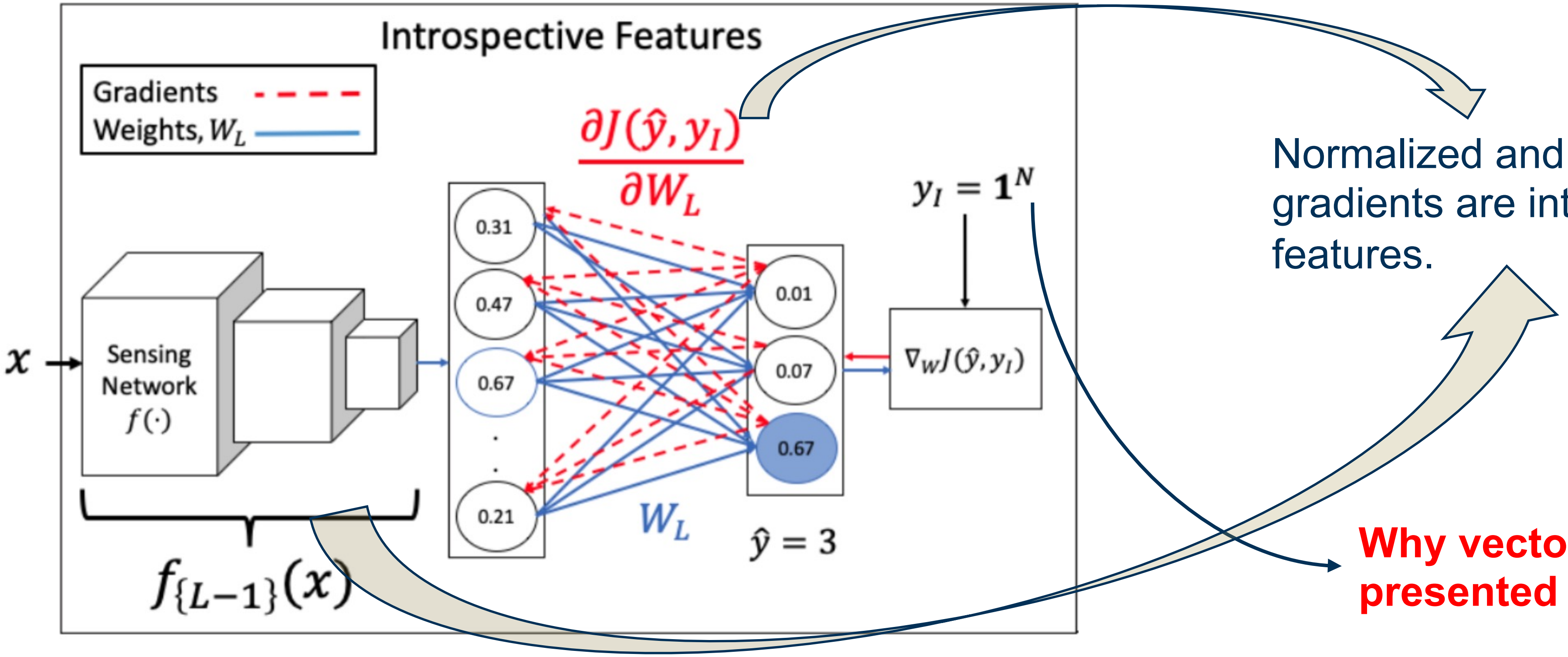
- In Part 3: Activations of learned manifold are weighted by gradients w.r.t. activations to extract information and provide explanations

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

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.

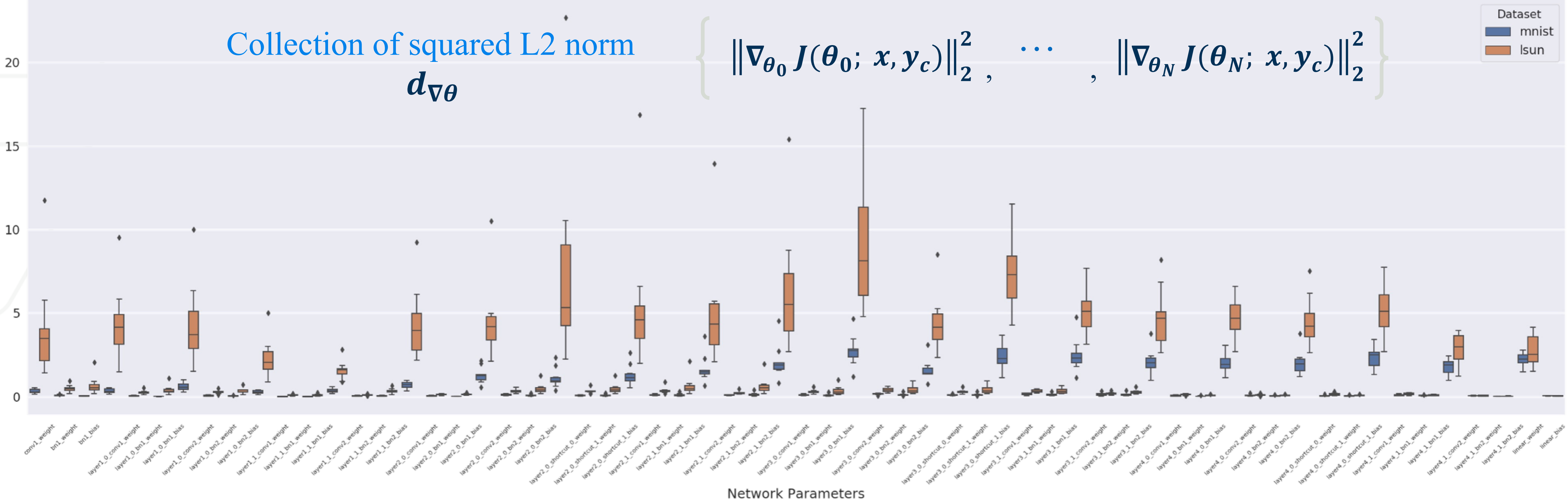
Why vector of all 1s? The theory is presented in [1]

[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.

Step 2: Take L2 norm of all generated gradients

Collection of squared L2 norm $d_{\nabla\theta}$

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



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.

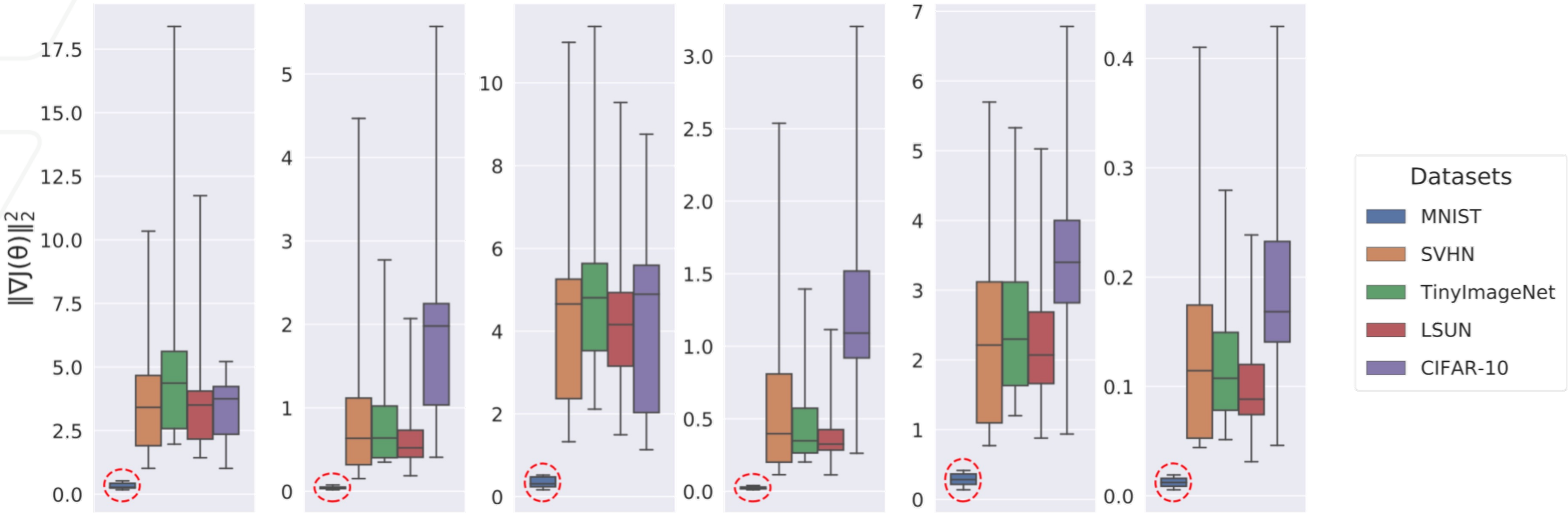
Gradient-based Uncertainty

Uncertainty in OOD Setting

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

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.

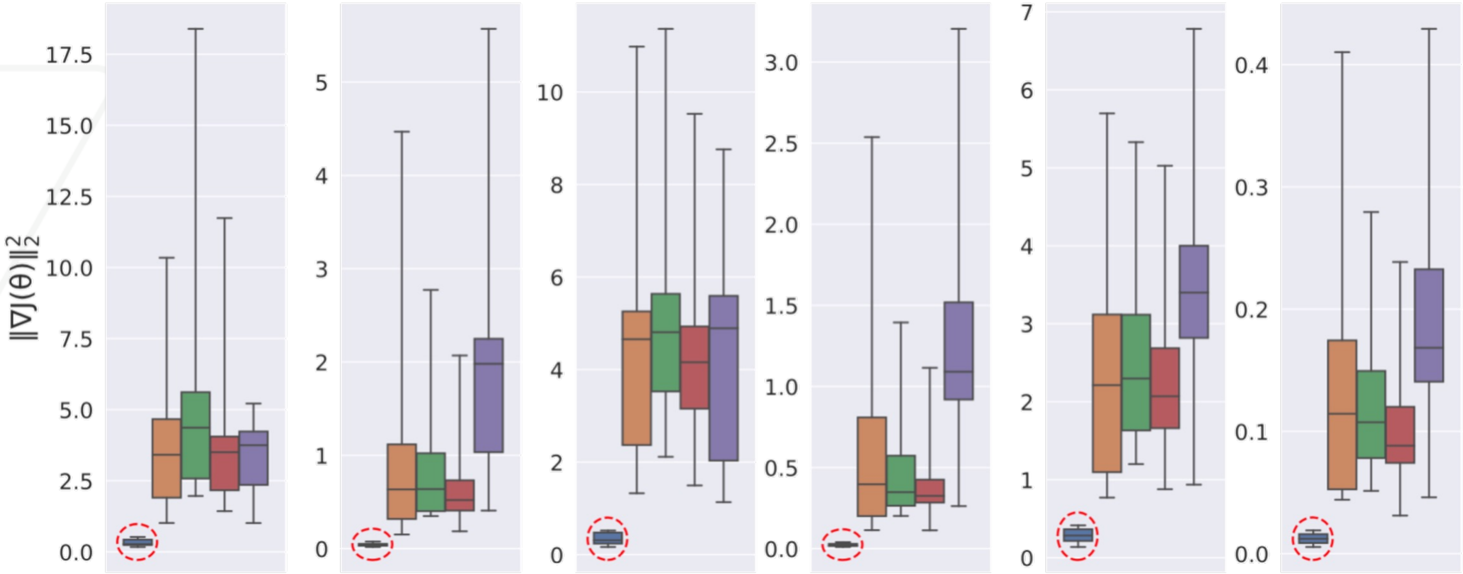
Gradient-based Uncertainty

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**

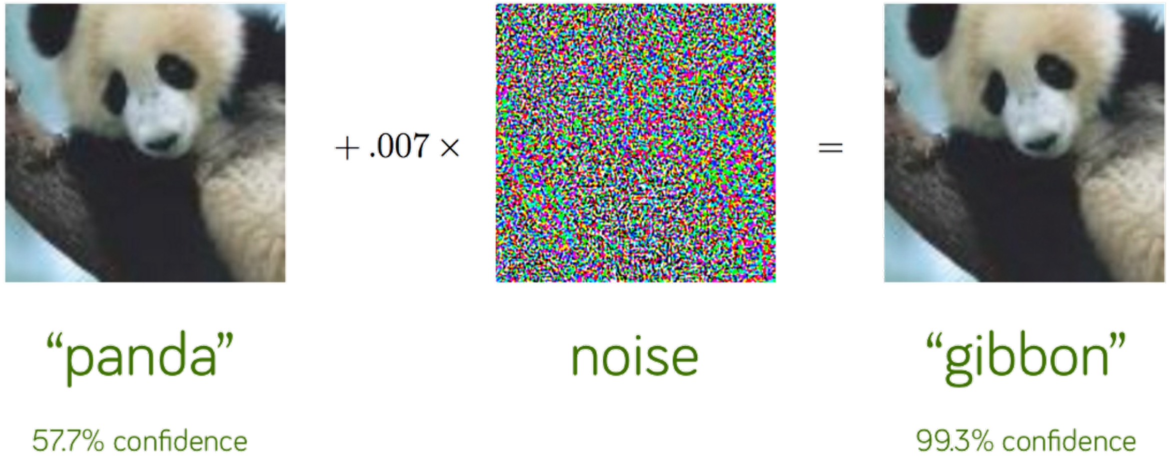
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Gradient-based Uncertainty
Uncertainty in Adversarial Setting

SCAN ME

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

SCAN ME

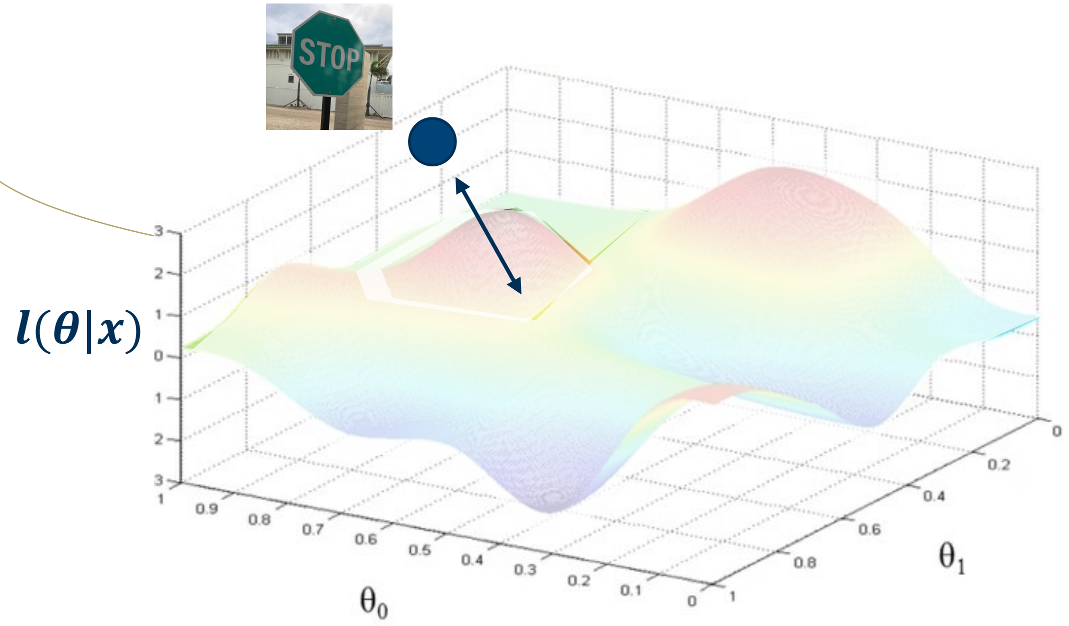
Probing the Purview of Neural Networks via Gradient Analysis

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.

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

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!

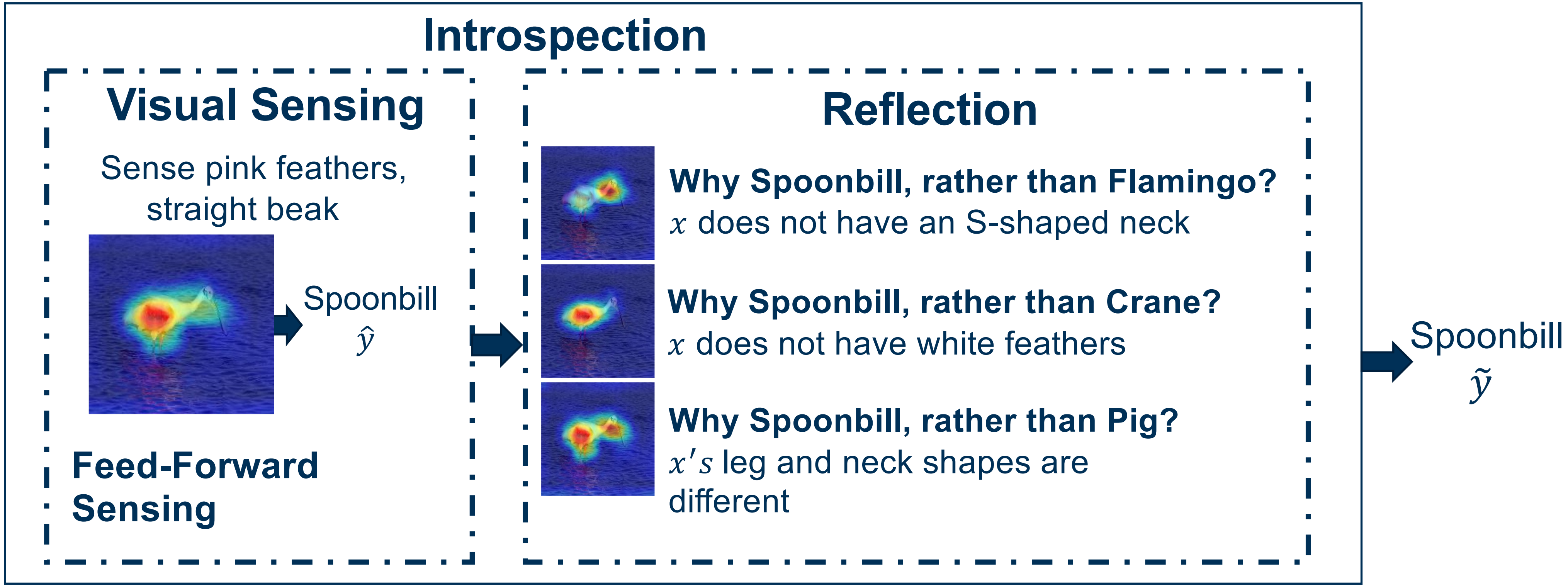
- Why am I being shown this slide?
- Why images of muffins rather than pastries?
- What if the dog was a bull mastiff?

What is Introspection?

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



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.

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.

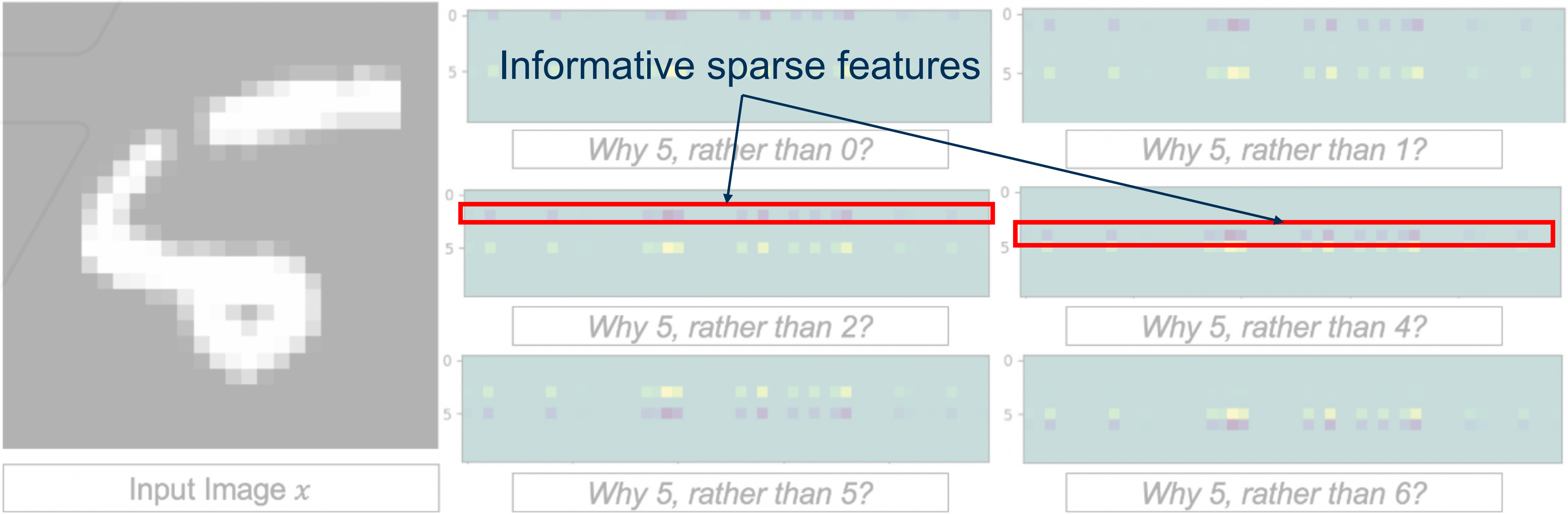
Introspection

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



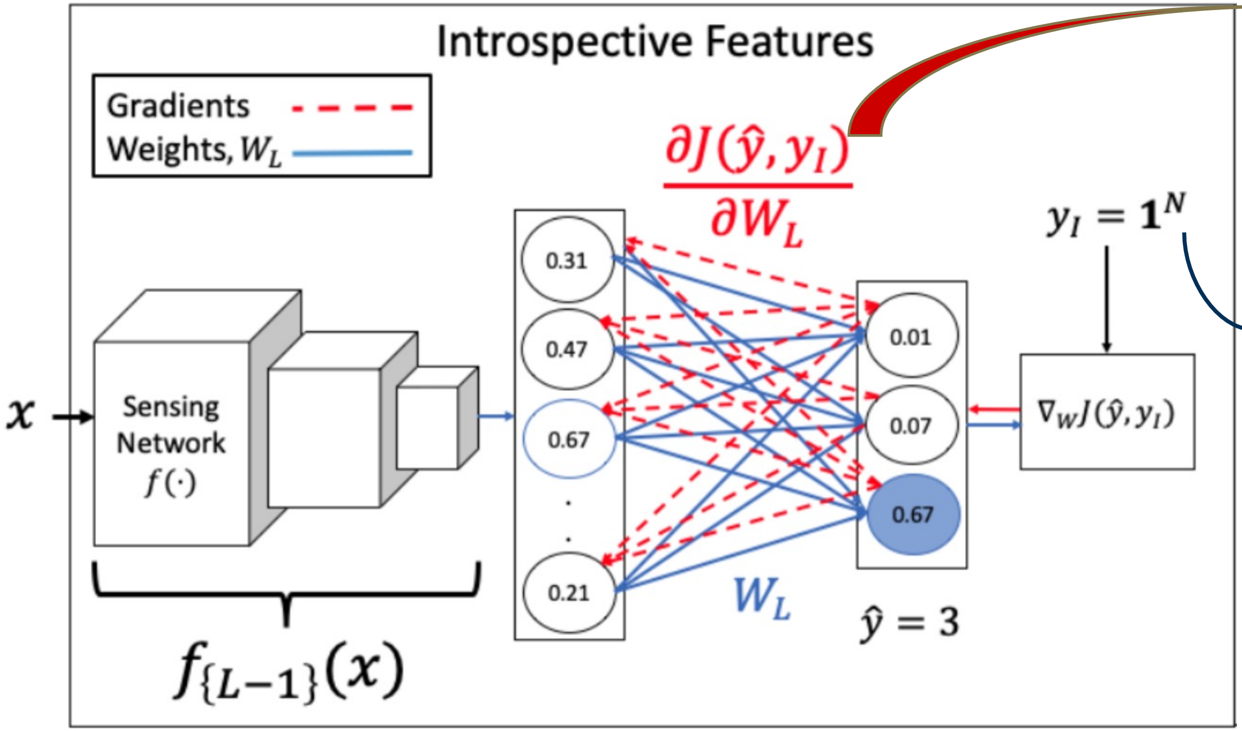
M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

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!

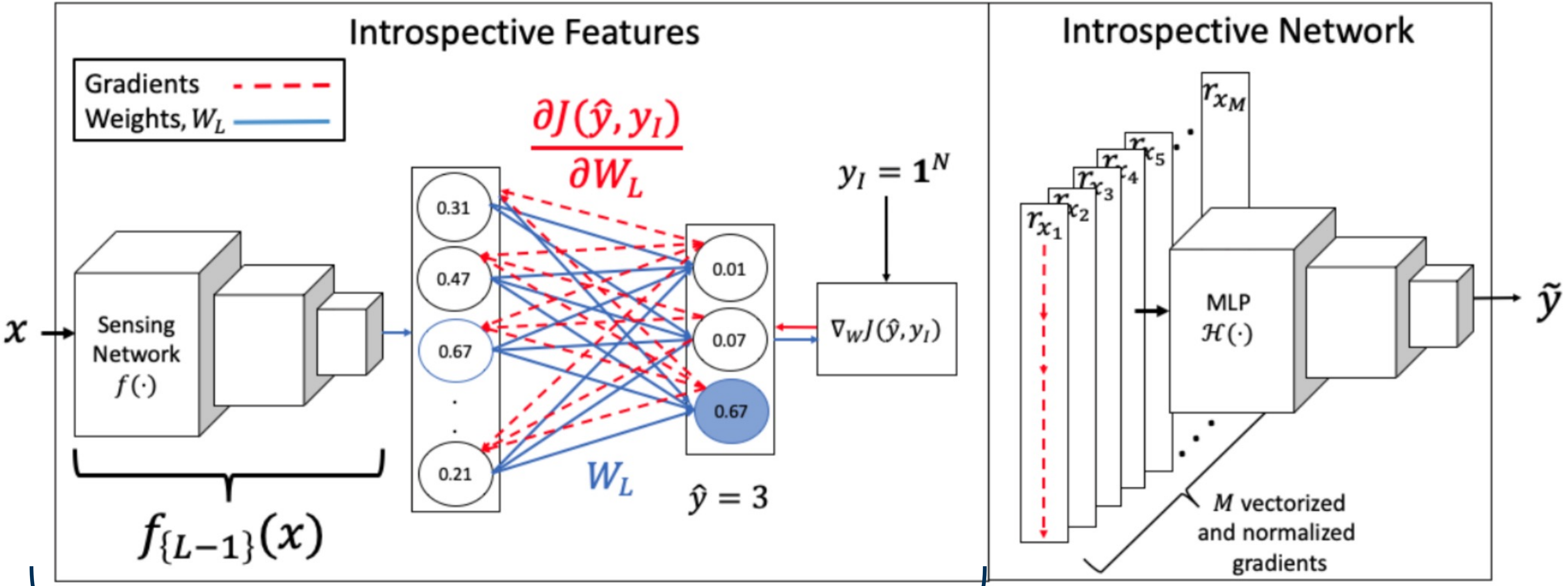
M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Introspection

Utilizing Gradient Features

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

SCAN ME



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.

OLIVES @GeorgiaTech

Georgia Tech

Introspection

When is Introspection Useful?

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

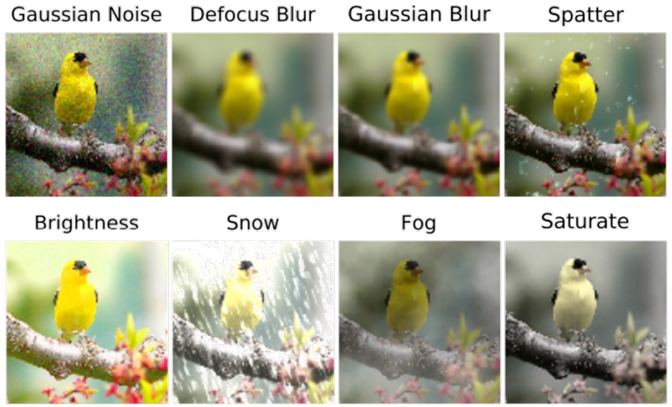
SCAN ME

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.

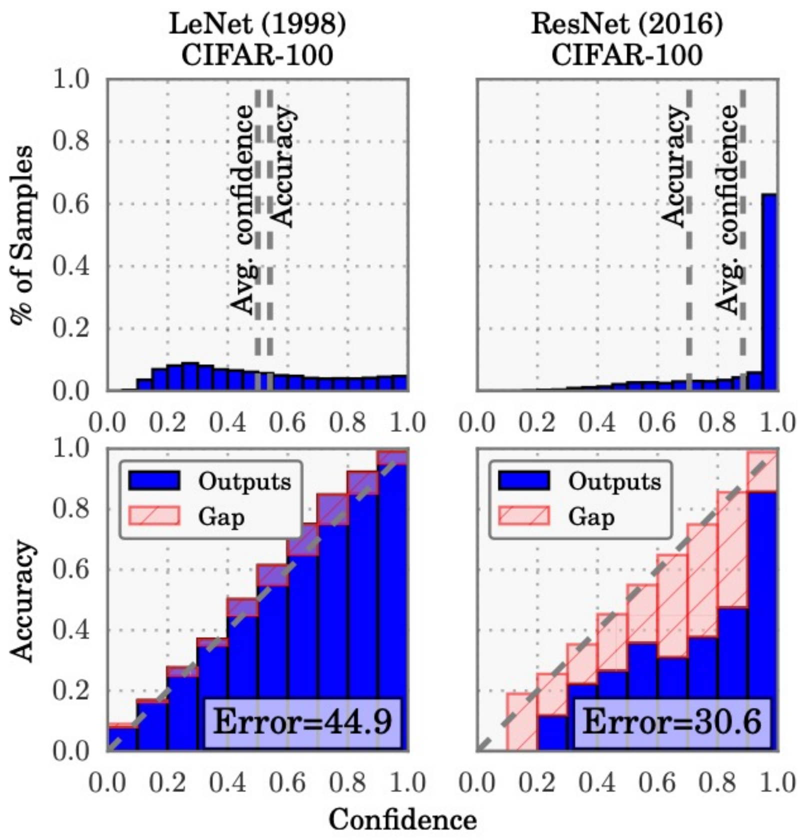
Calibration

A note on Calibration..

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

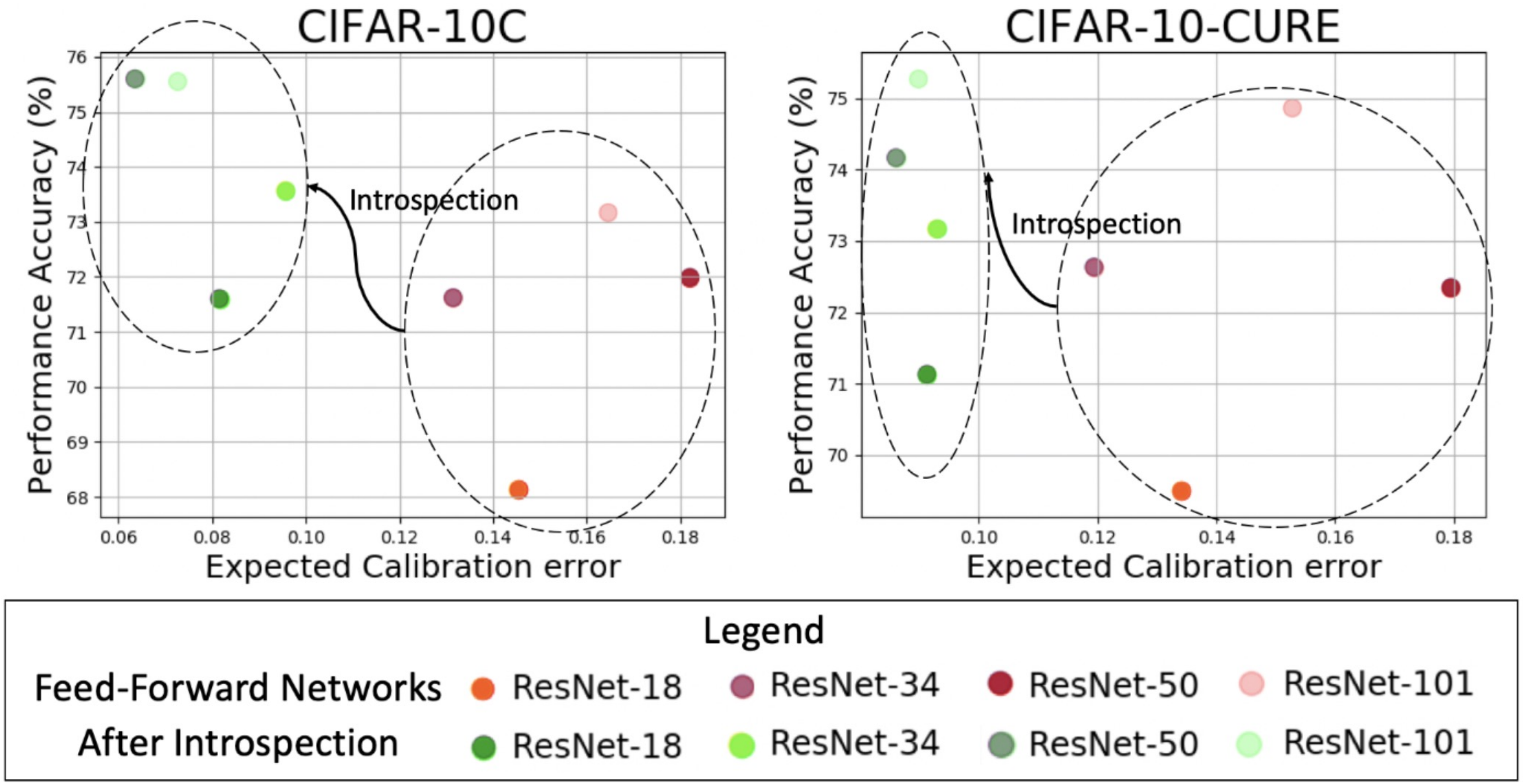
Introspective Learning: A Two-stage Approach for Inference in Neural Networks

SCAN ME

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.

OLIVES @GeorgiaTech

Georgia Tech

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	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!

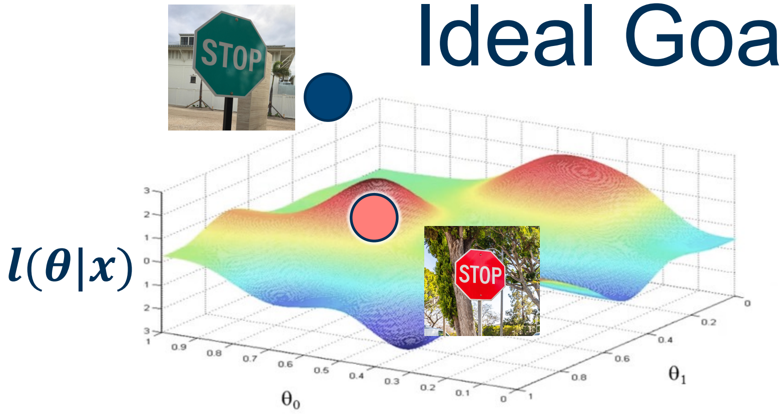
M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Part I, II and III

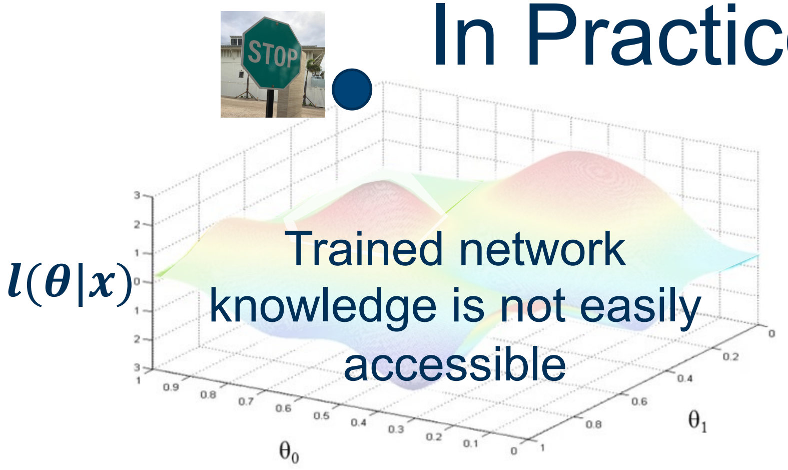
Tying it Back

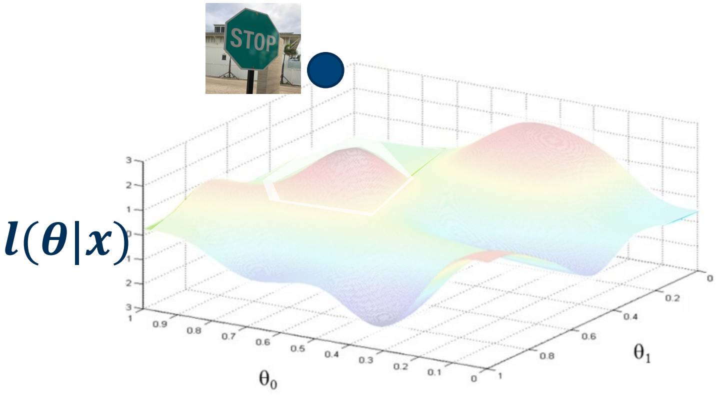
From Part I

Ideal Goal

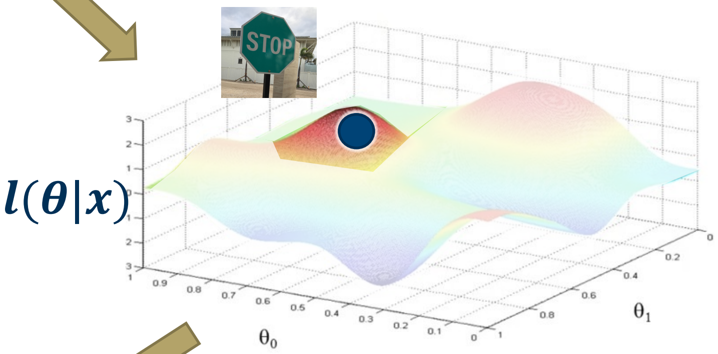


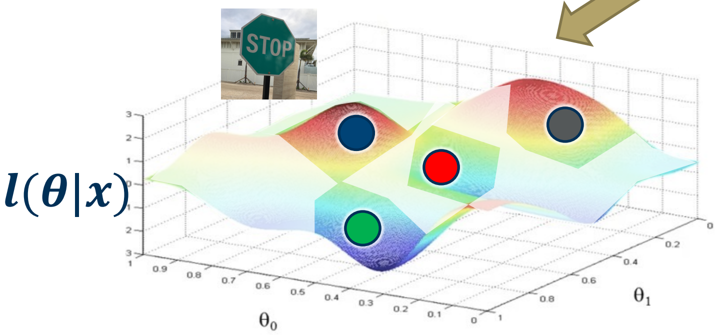
In Practice





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

Robust Neural Networks

Part 4: Intervenability 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
- **Part 4: Intervenability at Inference**
    - Definitions of Intervenability
        - Causality
        - Privacy
        - Interpretability
        - Prompting
        - Benchmarking
    - Case Study: Intervenability in Interpretability
- Part 5: Conclusions and Future Directions

Assess: The amenability of neural network decisions to human interventions

*"**Interventions** in data are **manipulations** that are **designed** to test for causal factors"*

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612-634.

Intervenability

Through the Privacy Glass

Assure: The amenability of neural network decisions to human interventions

*"**Intervenability** aims at the possibility for parties involved in any **privacy-relevant** data processing to **interfere** with the **ongoing or planned data processing**"*

Hansen, M.: Top 10 mistakes in system design from a privacy perspective and privacy protection goals. In: Camenisch, J., Crispo, B., Fischer-Hübner, S., Leenes, R., Russello, G. (eds.) Privacy and Identity Management for Life. IFIP AICT, vol. 375, pp. 14–31. Springer, Heidelberg (2012)

Through the Interpretability Glass

Interpret: The amenability of neural network decisions to human interventions

*"The **post-hoc field** of explainability, that previously only justified decisions, becomes **active** by being involved in the decision making process and **providing limited, but relevant and contextual interventions**"*

AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

Intervenability

Through the Benchmarking Glass

Verify: The amenability of neural network decisions to human interventions

*"… new **benchmarks** were proposed to specifically **test generalization** of classification and detection methods with respect to **simple** algorithmically generated **interventions** like spatial shifts, blur, changes in brightness or contrast…"*

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612-634.

Intervenability

Through the Human Glass

The amenability of neural network decisions to human interventions

- **Assess: Causality**
- **Assure: Privacy**
- **Interpret: Interpretability**
- **Verify: Benchmarking**

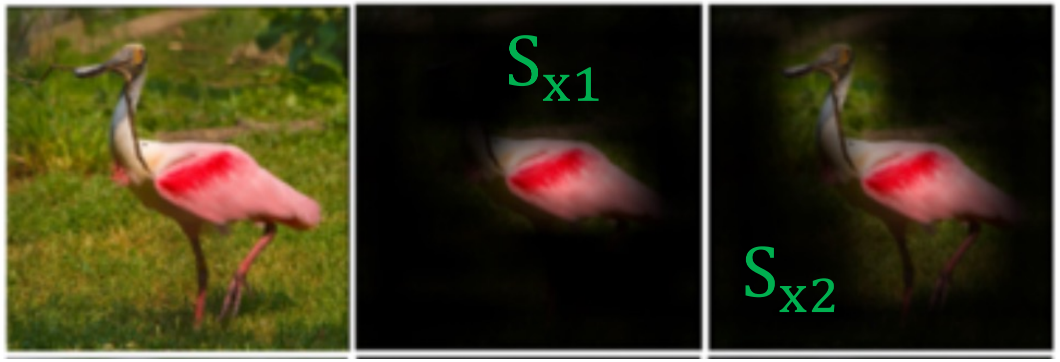
Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612-634.

Explanation Evaluation via Masking

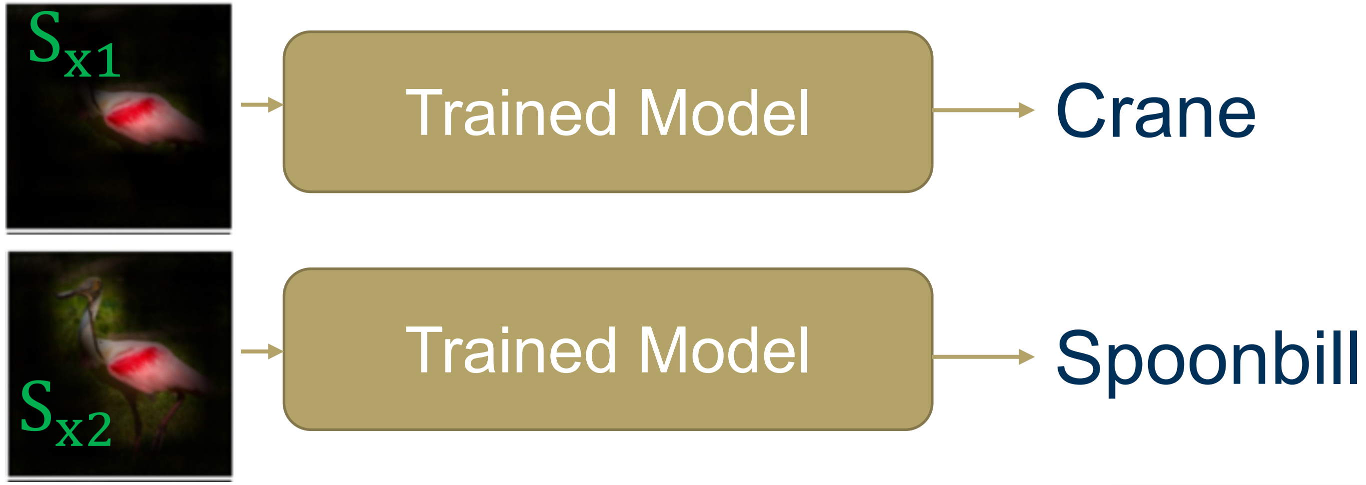
Common evaluation technique is masking the image and checking for prediction correctness

y = Prediction
S_x = Explanation masked data

$E(Y|S_x)$ = Expectation of class given S_x



If across N images, $\mathbf{E(Y|S_{x2}) > E(Y|S_{x1})}$, explanation technique 2 is better than explanation technique 1



Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018.

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

Mohit Prabhushankar, PhD
Postdoc

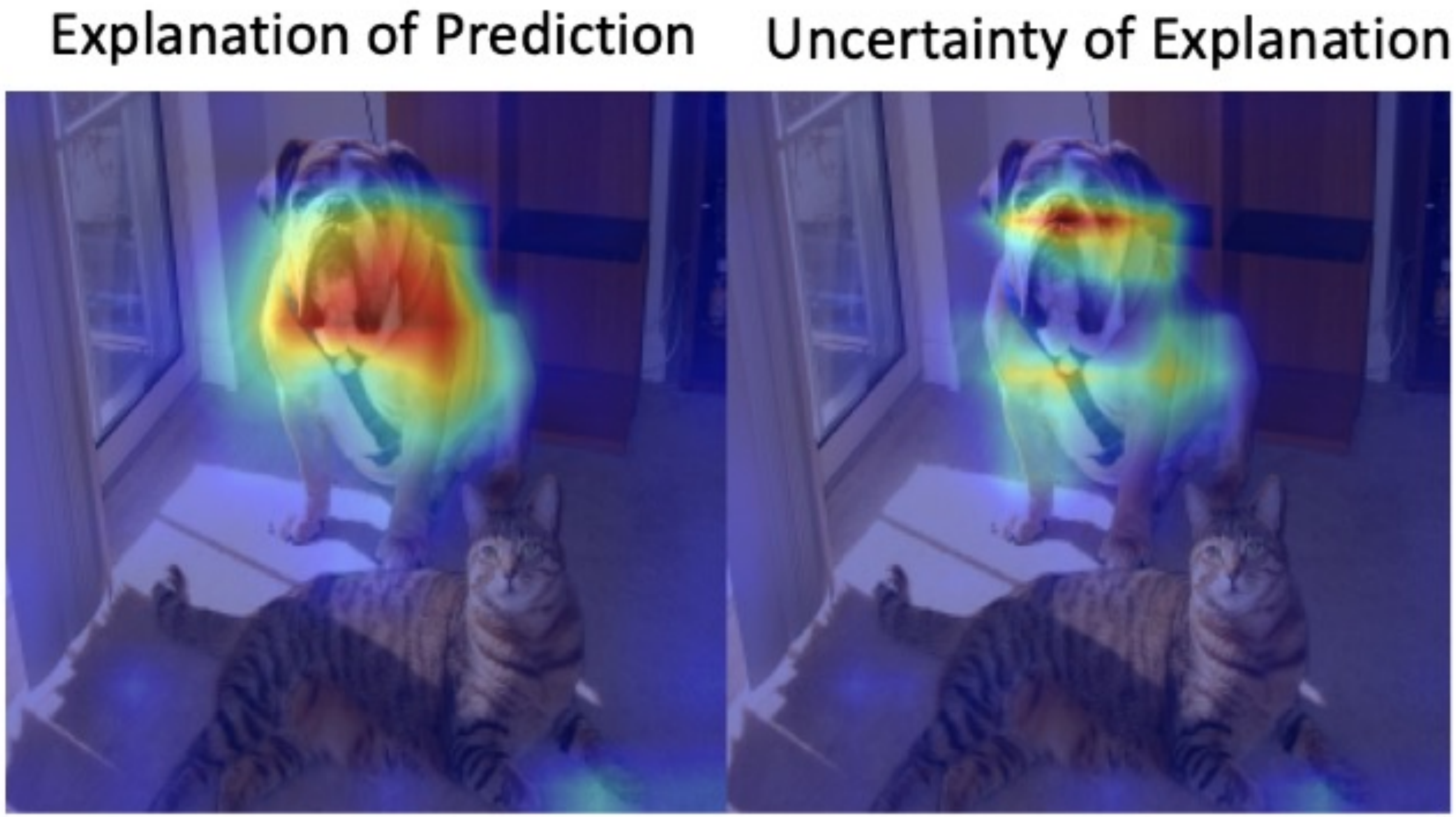
Ghassan AlRegib, PhD
Professor

Case Study: Intervenability in Interpretability

Predictive Uncertainty in Explanations

Explanatory techniques have predictive uncertainty

Why Bullmastiff?



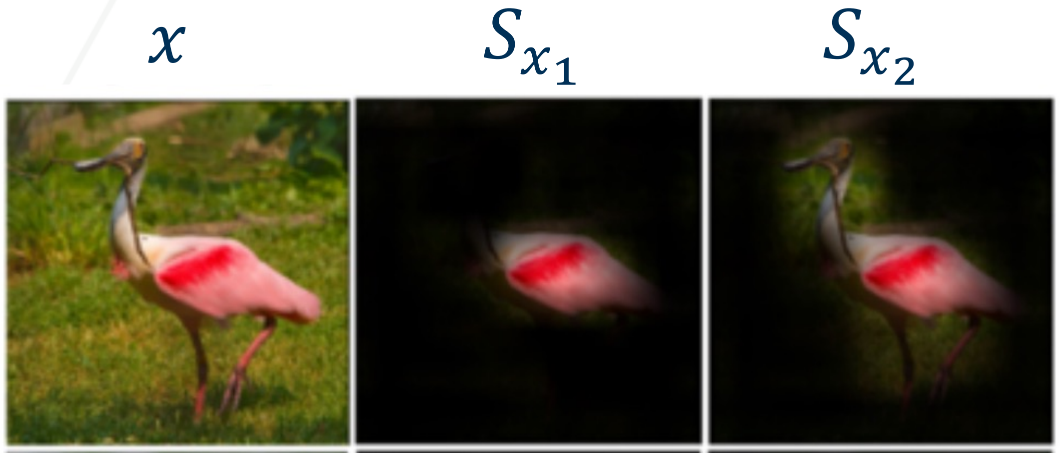
Uncertainty in answering Why Bullmastiff?

M. Prabhushankar, and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," *Journal of Selected Topics in Signal Processing*, submitted on Aug. 27, 2023.

Case Study: Intervenability in Interpretability

Predictive Uncertainty

Uncertainty due to variance in prediction when model is kept constant



$$V[y|S_x] = V[E(y|S_x)] + E(V[y|S_x])$$

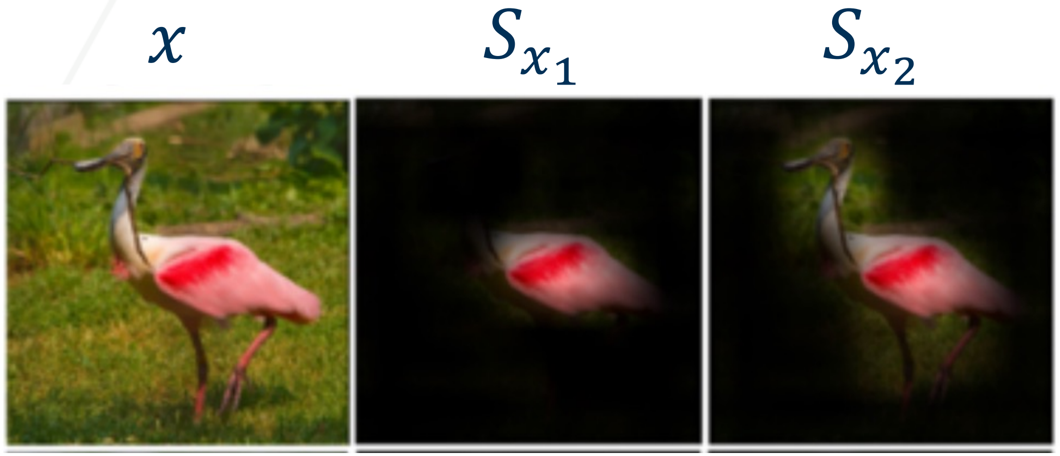
y = Prediction
$V[y]$ = Variance of prediction (Predictive Uncertainty)
S_x = Subset of data (Some intervention)
$E(Y|S_x)$ = Expectation of class given a subset
$V(Y|S_x)$ = Variance of class given all other residuals

M. Prabhushankar, and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," *Journal of Selected Topics in Signal Processing*, submitted on Aug. 27, 2023.

Case Study: Intervenability in Interpretability

Visual Explanations (partially) reduce Predictive Uncertainty

A ‘good’ explanatory technique is evaluated to have zero $V[E(y|S_x)]$



$$V[y|S_x] = V[E(y|S_x)] + E(V[y|S_x])$$

zero

y = Prediction
$V[y]$ = Variance of prediction (Predictive Uncertainty)
S_x = Subset of data (Some intervention)
$E(Y|S_x)$ = Expectation of class given a subset
$V(Y|S_x)$ = Variance of class given all other residuals

Key Observation 1: Visual Explanations are evaluated to partially reduce the predictive uncertainty in a neural network

Network evaluations have nothing to do with human Explainability!

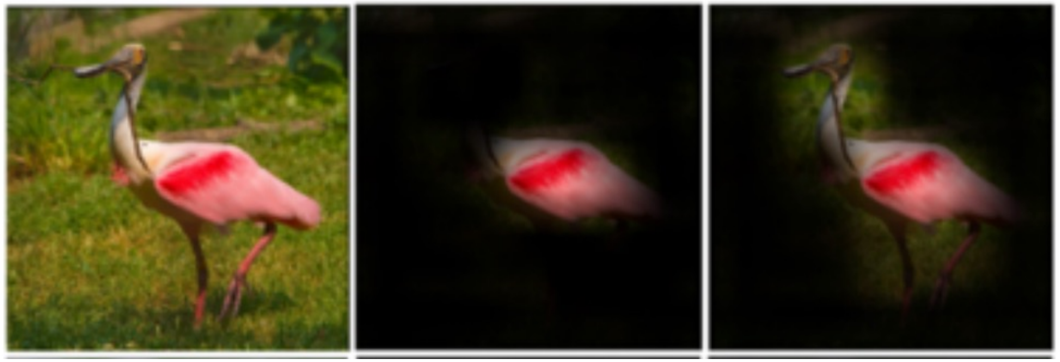
M. Prabhushankar, and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," *Journal of Selected Topics in Signal Processing*, submitted on Aug. 27, 2023.

Case Study: Intervenability in Interpretability

Predictive Uncertainty in Explanations is the Residual

All other subsets 'not' chosen by the explanatory technique contributes to uncertainty

x S_{x_1} S_{x_2}



$$V[y|S_x] = V[E(y|S_x)] + E(V[y|S_x])$$

y = Prediction
$V[y]$ = Variance of prediction (Predictive Uncertainty)
S_x = Subset of data (Some intervention)
$E(Y|S_x)$ = Expectation of class given a subset
$V(Y|S_x)$ = Variance of class given all other residuals

Key Observation 2: Uncertainty in Explainability occurs due to all combinations of features that the explanation did not attribute to the network's decision

M. Prabhushankar, and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," *Journal of Selected Topics in Signal Processing*, submitted on Aug. 27, 2023.

Case Study: Intervenability in Interpretability

Predictive Uncertainty in Explanations is the Residual

All other subsets 'not' chosen by the explanatory technique contributes to uncertainty

$$V[y|S_x] = V[E(y|S_x)] + E(V[y|S_x])$$

y = Prediction

$V[y]$ = Variance of prediction (Predictive Uncertainty)

S_x = Subset of data (Some intervention)

$E(Y|S_x)$ = Expectation of class given a subset

$V(Y|S_x)$ = Variance of class given all other residuals

The effect of a chosen Interventions can be measured based on *all the Interventions that were not chosen*

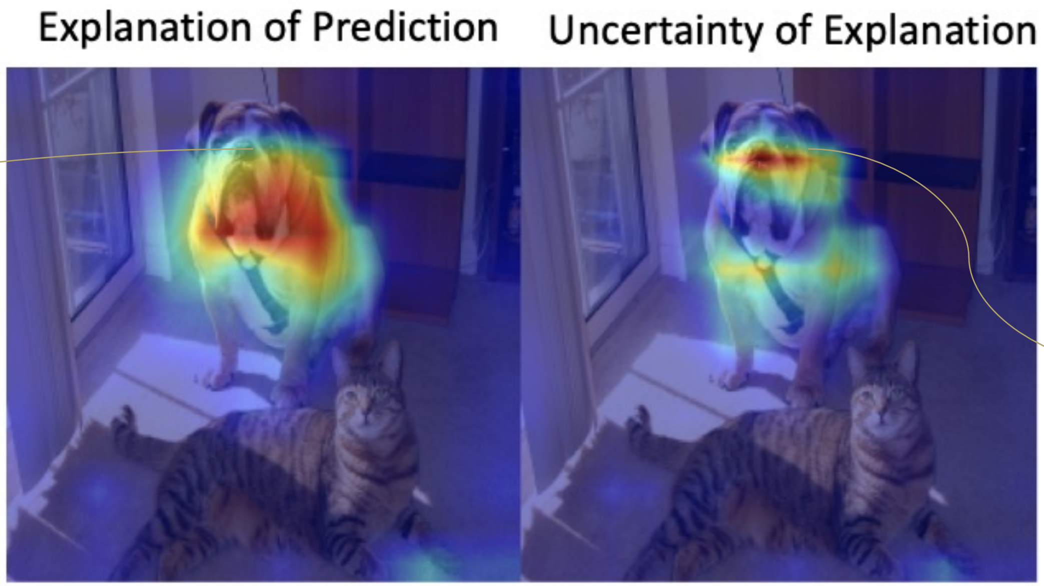
Key Observation 2: **Uncertainty in Explainability occurs due to all combinations of features that the explanation did not attribute to the network's decision**

M. Prabhushankar, and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," *Journal of Selected Topics in Signal Processing*, submitted on Aug. 27, 2023.

Case Study: Intervenability in Interpretability

Predictive Uncertainty in Explanations is the Residual

All other subsets ‘not’ chosen by the explanatory technique contributes to uncertainty



Snout is not as highlighted as the jowls in explanation (not as important for decision)

However, snout is an important characteristic that is used to differentiate against other dogs. Hence, there is uncertainty on why this feature is not included in the attribution

Key Observation 2: Uncertainty in Explainability occurs due to all combinations of features that the explanation did not attribute to the network’s decision

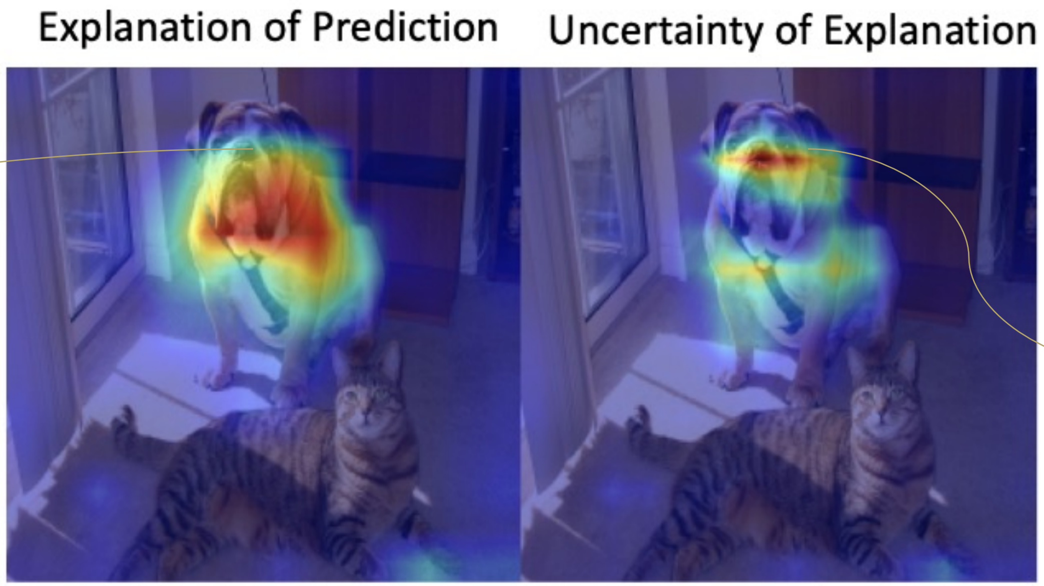
M. Prabhushankar, and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," *Journal of Selected Topics in Signal Processing*, submitted on Aug. 27, 2023.

Case Study: Intervenability in Interpretability

Predictive Uncertainty in Explanations is the Residual

All other subsets 'not' chosen by the explanatory technique contributes to uncertainty

Snout is not as highlighted as the jowls in explanation (not as important for decision)



However, snout is an important characteristic that is used to differentiate against other dogs. Hence, there is uncertainty on why this feature is not included in the attribution

Not chosen features are intractable!

M. Prabhushankar, and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," *Journal of Selected Topics in Signal Processing*, submitted on Aug. 27, 2023.

Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability

Contrastive explanations are an intelligent way of obtaining other subsets

S_{x_1} S_{x_2} …….. S_{x_N}

$$V[y|S_x] = V[E(y|S_x)] + E(V[y|S_x])$$

Make it finite by only considering the subsets that change y

$Y_1|S_{x1}$
$Y_2|S_{x2}$
$Y_3|S_{x3}$
$Y_4|S_{x4}$
$Y_5|S_{x5}$
.
.
$Y_N|S_{xN}$

} Variance

M. Prabhushankar, and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," *Journal of Selected Topics in Signal Processing*, submitted on Aug. 27, 2023.

Uncertainty in Explainability can be used to analyze Explanatory methods and Networks

- Is GradCAM better than GradCAM++?
- Is a SWIN transformer more reliable than VGG-16?

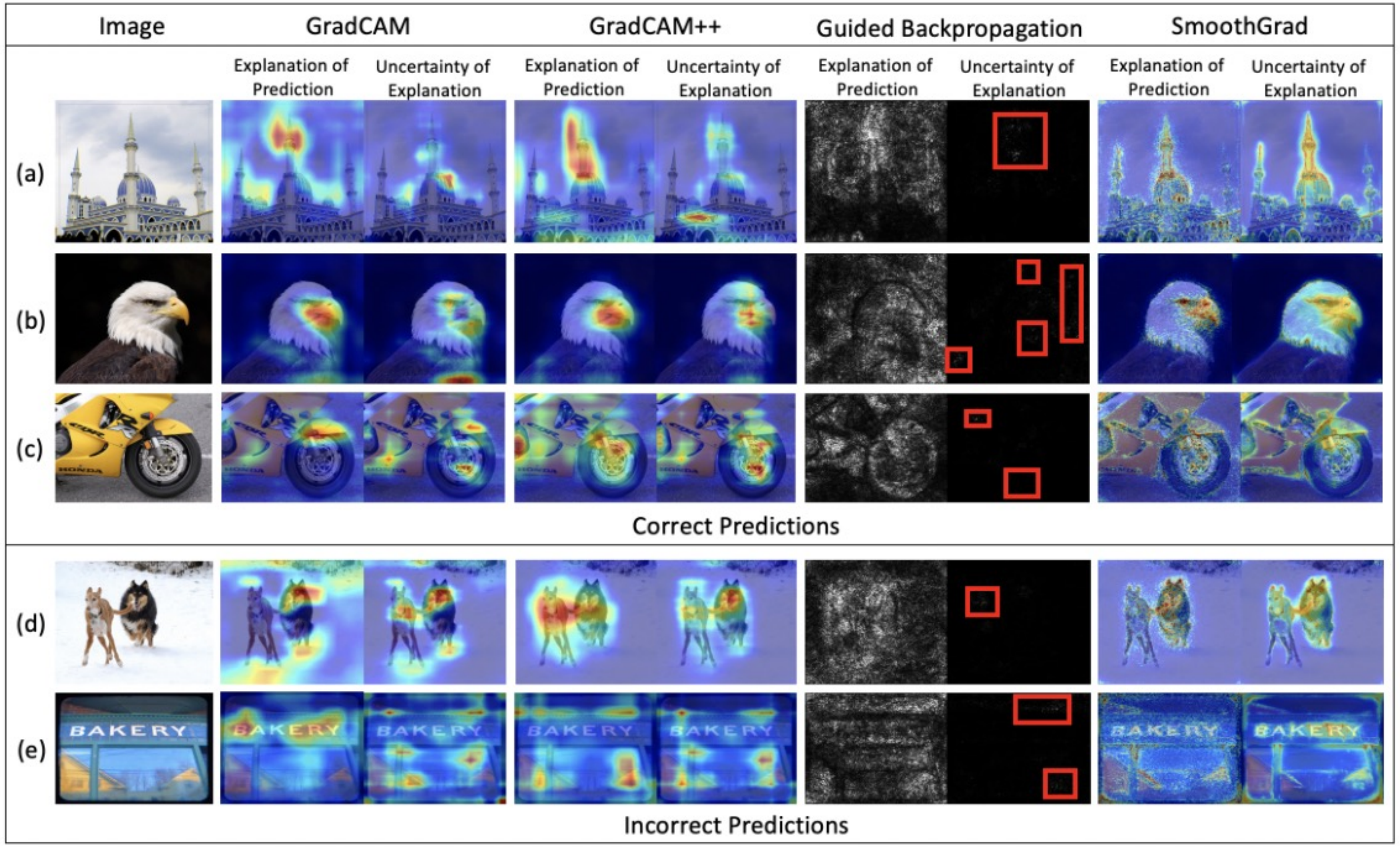
Need objective quantification of Intervention Residuals

M. Prabhushankar, and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," *Journal of Selected Topics in Signal Processing*, submitted on Aug. 27, 2023.

Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability: mIOU

On incorrect predictions, the overlap of explanations and uncertainty is higher



Objective Metric: Intersection over Union (IoU) between explanation and Uncertainty

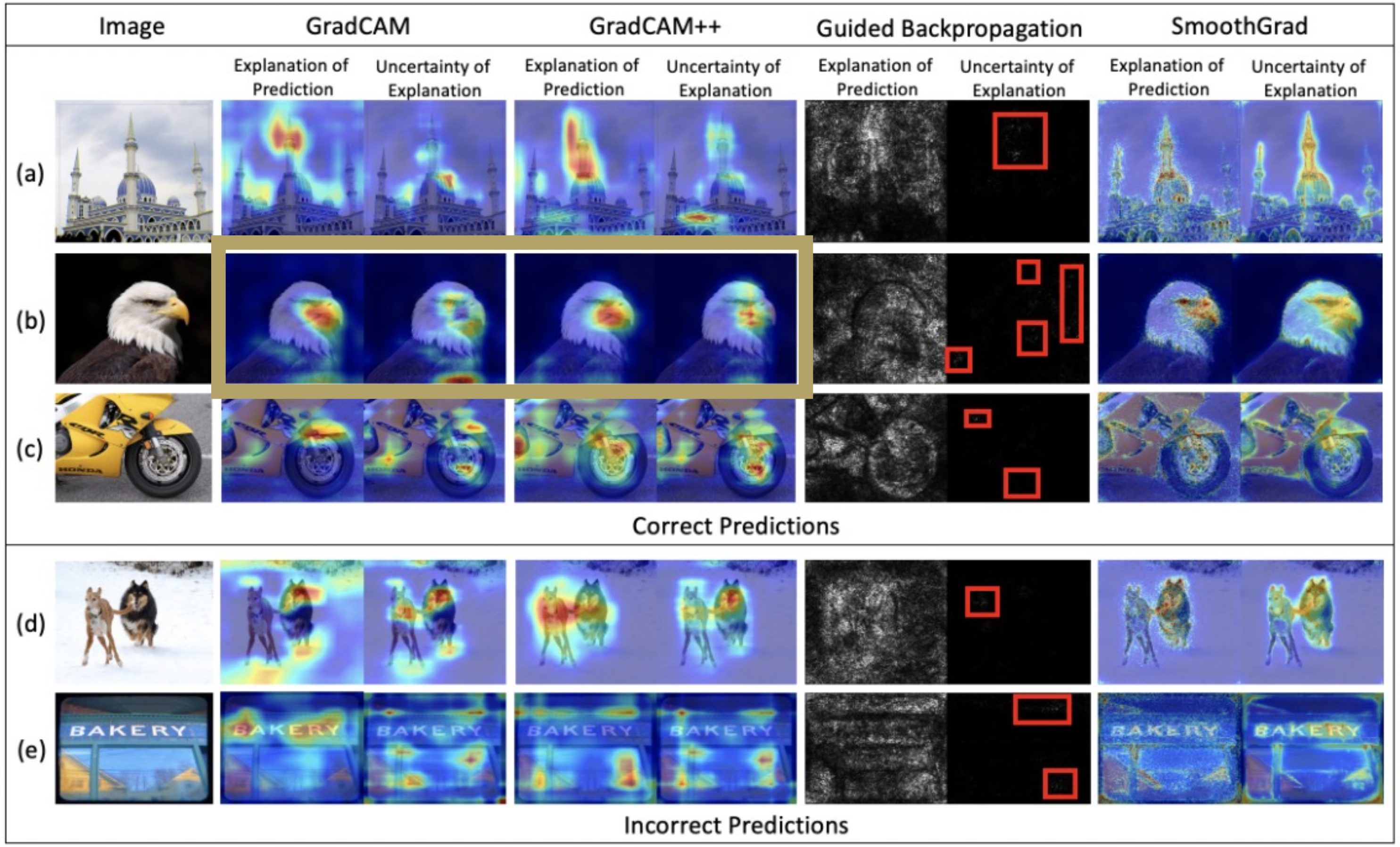
Higher the IoU, higher the uncertainty in explanation (or less trustworthy is the prediction)

M. Prabhushankar, and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," *Journal of Selected Topics in Signal Processing*, submitted on Aug. 27, 2023.

Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability: mIOU

On incorrect predictions, the overlap of explanations and uncertainty is higher



Objective Metric 1: Intersection over Union (IoU) between explanation and Uncertainty

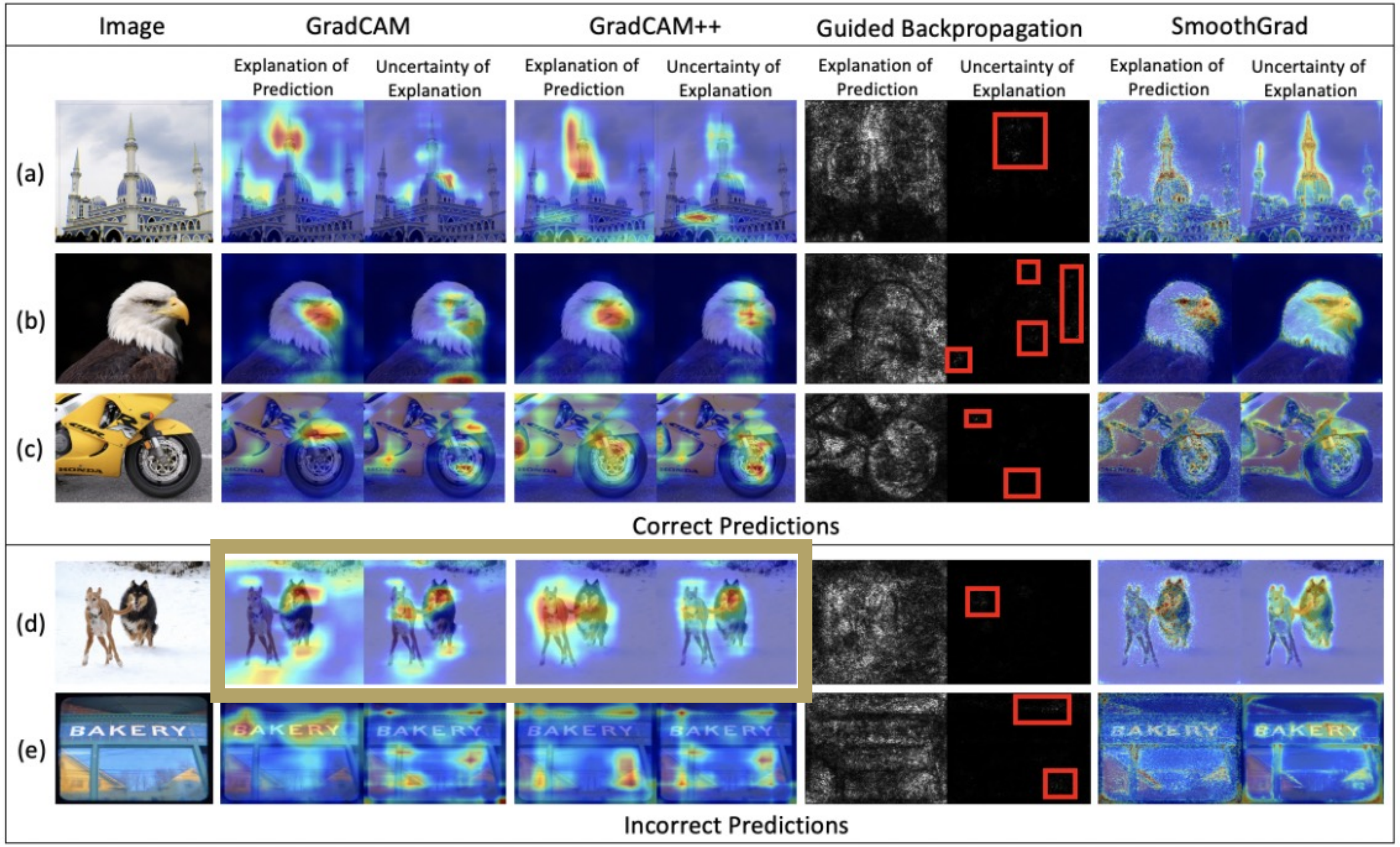
Higher the IoU, higher the uncertainty in explanation (or less trustworthy is the prediction)

M. Prabhushankar, and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," *Journal of Selected Topics in Signal Processing*, submitted on Aug. 27, 2023.

Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability: mIOU

On incorrect predictions, the overlap of explanations and uncertainty is higher



Objective Metric 1: Intersection over Union (IoU) between explanation and Uncertainty

Higher the IoU, higher the uncertainty in explanation (or less trustworthy is the prediction)

M. Prabhushankar, and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," *Journal of Selected Topics in Signal Processing*, submitted on Aug. 27, 2023.

Robust Neural Networks

Part 5: Conclusions and Future Directions

Key Takeaways

Role of Gradients

- **Robustness** under distributional shift in domains, environments, and adversaries are **challenges** for neural networks
  - **Gradients at Inference** provide a **holistic solution** to the above challenges
- **Gradients** can help **traverse** through a trained and unknown **manifold**
  - They approximate **Fisher Information** on the projection
  - They can be **manipulated** by providing **contrast** classes
  - They can be used to construct **localized contrastive** manifolds
  - They provide **implicit knowledge** about **all classes**, when only **one data** point is available at inference
- Gradients are useful in a number of **Image Understanding** applications
  - Highlighting features of the current prediction as well as **counterfactual** data and **contrastive** classes
  - Providing **directional information** in anomaly detection
  - **Quantifying uncertainty** for out-of-distribution, corruption, and adversarial detection
  - Providing **expectancy mismatch** for human vision related applications

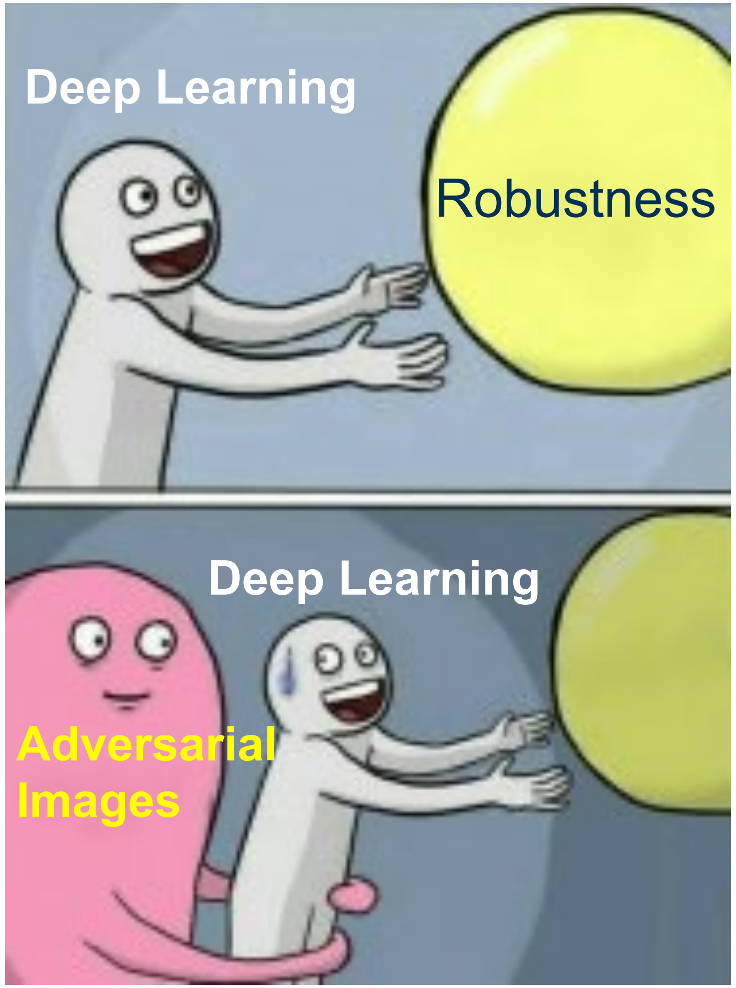
Future Directions

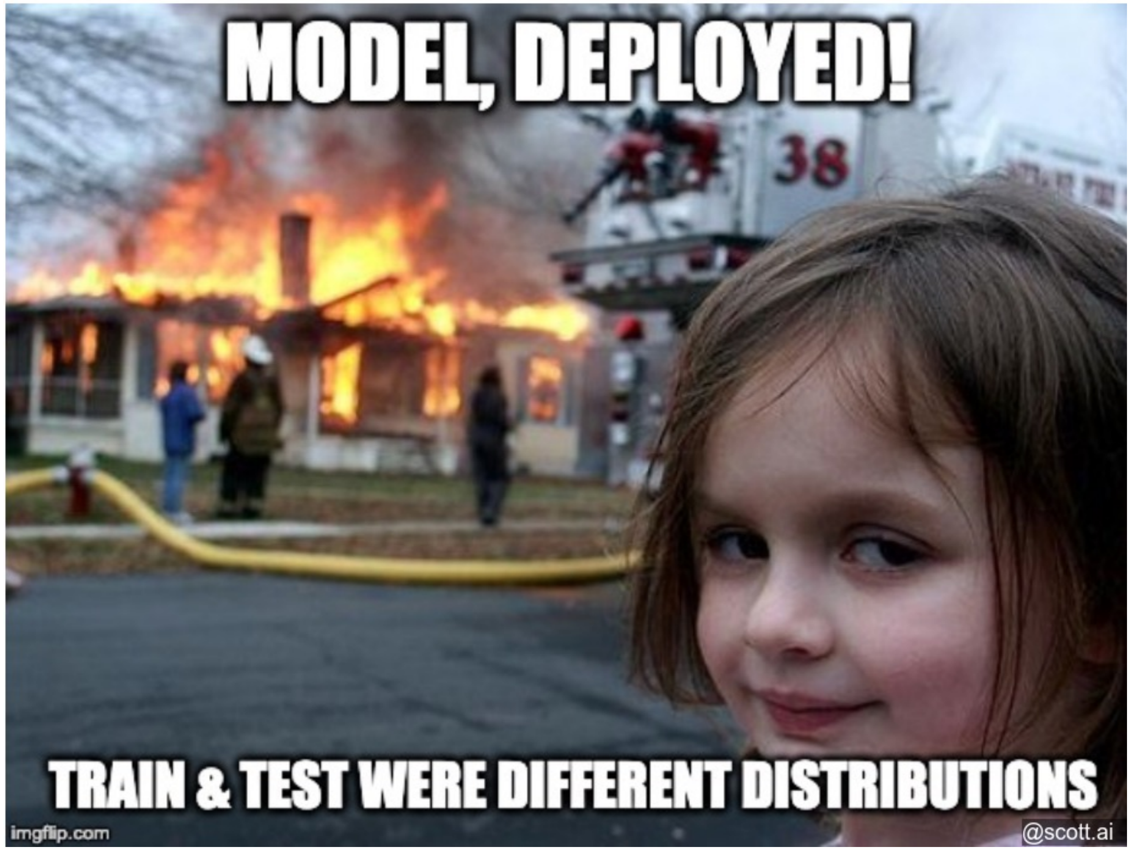
Research at Inference Stage

- **Test Time Augmentation (TTA) Research**
  - Multiple augmentations of data are passed through the network at inference
  - Research is in designing the best augmentations
- **Active Inference**
  - Utilize the knowledge in Neural Networks to ***ask it to ask us***
  - Neural networks ask for the best augmentation of the data point given that one data point at inference
- **Uncertainty in Explainability, Label Interpretation, and Trust quantification**
  - Uncertainty research has to expand beyond model and data uncertainty
  - In some applications within medical and seismic communities, there is no agreed upon label for data. Uncertainty in label interpretation is its own research
- **Test-time Interventions for AI alignment**
  - Human interventions at test time to alter the decision-making process is essential trustworthy AI
  - Further research in intelligently involving experts in a non end-to-end framework is required

Memes to Wrap it Up

Robustness at Inference





Cannot depend on training to construct robust models

References

Gradient representations for Robustness, OOD, Anomaly, Novelty, and Adversarial Detection

- **Gradients for robustness against noise:** 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
- **Gradients for adversarial, OOD, corruption detection:** J. Lee, M. Prabhushankar, and G. AlRegib, "Gradient-Based Adversarial and Out-of-Distribution Detection," in *International Conference on Machine Learning (ICML) Workshop on New Frontiers in Adversarial Machine Learning*, Baltimore, MD, Jul. 2022.
- **Gradients for Open set recognition:** Lee, Jinsol, and Ghassan AlRegib. "Open-Set Recognition With Gradient-Based Representations." *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021.
- **GradCon for Anomaly Detection:** Kwon, G., Prabhushankar, M., Temel, D., & AlRegib, G. (2020, August). Backpropagated gradient representations for anomaly detection. In *European Conference on Computer Vision* (pp. 206-226). Springer, Cham.
- **Gradients for adversarial, OOD, corruption detection :** J. Lee, C. Lehman, M. Prabhushankar, and G. AlRegib, "Probing the Purview of Neural Networks via Gradient Analysis," in IEEE Access, Mar. 21 2023.
- **Gradients for Novelty Detection:** Kwon, G., Prabhushankar, M., Temel, D., & AlRegib, G. (2020, October). Novelty detection through model-based characterization of neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3179-3183). IEEE.
- **Gradient-based Image Quality Assessment:** G. Kwon*, M. Prabhushankar*, D. Temel, and G. AlRegib, "Distorted Representation Space Characterization Through Backpropagated Gradients," in *IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, Sep. 2019.

Explainability in Neural Networks

- **Explanatory paradigms:** AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks: Towards relevant and contextual explanations. *IEEE Signal Processing Magazine*, *39*(4), 59-72.
- **Contrastive Explanations:** Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.
- **Explainabilty in Limited Label Settings:** M. Prabhushankar, and G. AlRegib, ”Extracting Causal Visual Features for Limited Label Classification,” in IEEE International Conference on Image Processing (ICIP), Sept. 2021.
- **Explainabilty through Expectancy-Mismatch:** M. Prabhushankar and G. AlRegib, "Stochastic Surprisal: An Inferential Measurement of Free Energy in Neural Networks," in Frontiers in Neuroscience, Perception Science, Volume 17, Feb. 09 2023.

References

Self Supervised Learning

- **Weakly supervised Contrastive Learning:** K. Kokilepersaud, S. Trejo Corona, M. Prabhushankar, G. AlRegib, C. Wykoff, "Clinically Labeled Contrastive Learning for OCT Biomarker Classification," in IEEE Journal of Biomedical and Health Informatics, 2023, May. 15 2023.
- **Contrastive Learning for Fisheye Images**: K. Kokilepersaud, M. Prabhushankar, Y. Yarici, G. AlRegib, and A. Parchami, "Exploiting the Distortion-Semantic Interaction in Fisheye Data," in *Open Journal of Signals Processing*, Apr. 28 2023.
- **Contrastive Learning for Severity Detection:** K. Kokilepersaud, M. Prabhushankar, G. AlRegib, S. Trejo Corona, C. Wykoff, "Gradient Based Labeling for Biomarker Classification in OCT," in *IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, Oct. 16-19 2022
- **Contrastive Learning for Seismic Images:** K. Kokilepersaud, M. Prabhushankar, and G. AlRegib, "Volumetric Supervised Contrastive Learning for Seismic Semantic Segmentation," in *International Meeting for Applied Geoscience & Energy (IMAGE)*, Houston, TX, , Aug. 28-Sept. 1 2022

Human Vision and Behavior Prediction

- **Pedestrian Trajectory Prediction:** C. Zhou, G. AlRegib, A. Parchami, and K. Singh, "TrajPRed: Trajectory Prediction With Region-Based Relation Learning," *IEEE Transactions on Intelligent Transportation Systems*, submitted on Dec. 28 2022.
- **Human Visual Saliency in trained Neural Nets:** Y. Sun, M. Prabhushankar, and G. AlRegib, "Implicit Saliency in Deep Neural Networks," in *IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, Oct. 2020.
- **Human Image Quality Assessment:** D. Temel, M. Prabhushankar and G. AlRegib, "UNIQUE: Unsupervised Image Quality Estimation," in IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1414-1418, Oct. 2016.

Open-source Datasets to assess Robustness

- **CURE-TSD:** D. Temel, M-H. Chen, and G. AlRegib, "Traffic Sign Detection Under Challenging Conditions: A Deeper Look Into Performance Variations and Spectral Characteristics," in *IEEE Transactions on Intelligent Transportation Systems*, Jul. 2019
- **CURE-TSR:** D. Temel, G. Kwon*, M. Prabhushankar*, and G. AlRegib, "CURE-TSR: Challenging Unreal and Real Environments for Traffic Sign Recognition," in *Advances in Neural Information Processing Systems (NIPS) Workshop on Machine Learning for Intelligent Transportation Systems*, Long Beach, CA, Dec. 2017
- **CURE-OR:** D. Temel*, J. Lee*, and G. AlRegib, "CURE-OR: Challenging Unreal and Real Environments for Object Recognition," in *IEEE International Conference on Machine Learning and Applications (ICMLA)*, Orlando, FL, Dec. 2018

References

Active Learning

- **Active Learning and Training with High Information Content:** R. Benkert, M. Prabhushankar, G. AlRegib, A. Parchami, and E. Corona, "Gaussian Switch Sampling: A Second Order Approach to Active Learning," in IEEE Transactions on Artificial Intelligence (TAI), Feb. 05 2023
- **Active Learning Dataset on vision and LIDAR data:** Y. Logan, R. Benkert, C. Zhou, K. Kokilepersaud, M. Prabhushankar, G. AlRegib, K. Singh, E. Corona and A. Parchami, "FOCAL: A Cost-Aware Video Dataset for Active Learning," IEEE Transactions on Circuits and Systems for Video Technology, submitted on Apr. 29 2023
- **Active Learning on OOD data:** R. Benkert, M. Prabhushankar, and G. AlRegib, "Forgetful Active Learning With Switch Events: Efficient Sampling for Out-of-Distribution Data," in *IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, Oct. 16-19 2022
- **Active Learning for Biomedical Images**: Y. Logan, R. Benkert, A. Mustafa, G. Kwon, G. AlRegib, "Patient Aware Active Learning for Fine-Grained OCT Classification," in *IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, Oct. 16-19 2022

Uncertainty Estimation

- **Gradient-based Uncertainty:** J. Lee and G. AlRegib, "Gradients as a Measure of Uncertainty in Neural Networks," in *IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, Oct. 2020
- **Gradient-based Visual Uncertainty:** M. Prabhushankar, and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," *Journal of Selected Topics in Signal Processing*, submitted on Aug. 27, 2023.
- **Uncertainty Visualization in Seismic Images:** R. Benkert, M. Prabhushankar, and G. AlRegib, "Reliable Uncertainty Estimation for Seismic Interpretation With Prediction Switches," in *International Meeting for Applied Geoscience & Energy (IMAGE)*, Houston, TX, , Aug. 28-Sept. 1 2022.
- **Uncertainty and Disagreements in Label Annotations:** C. Zhou, M. Prabhushankar, and G. AlRegib, "On the Ramifications of Human Label Uncertainty," in *NeurIPS 2022 Workshop on Human in the Loop Learning*, Oct. 27 2022
- **Uncertainty in Saliency Estimation:** T. Alshawi, Z. Long, and G. AlRegib, "Unsupervised Uncertainty Estimation Using Spatiotemporal Cues in Video Saliency Detection," in *IEEE Transactions on Image Processing*, vol. 27, pp. 2818-2827, Jun. 2018.

Tutorial Materials

Accessible Online

SCAN ME

https://alregib.ece.gatech.edu/ieee-bigdata-2023-tutorial/

{alregib, mohit.p}@gatech.edu

IEEE BigData 2023 Tutorial

Presented by: ***Ghassan AlRegib, and Mohit Prabhushankar***
Georgia Institute of Technology

www.ghassanalregib.info

Title: Robust Neural Networks: Explainability, Uncertainty, and Intervenability