A Holistic View of Perception in Intel. Vehicles

Part III: Deep Learning at Inference

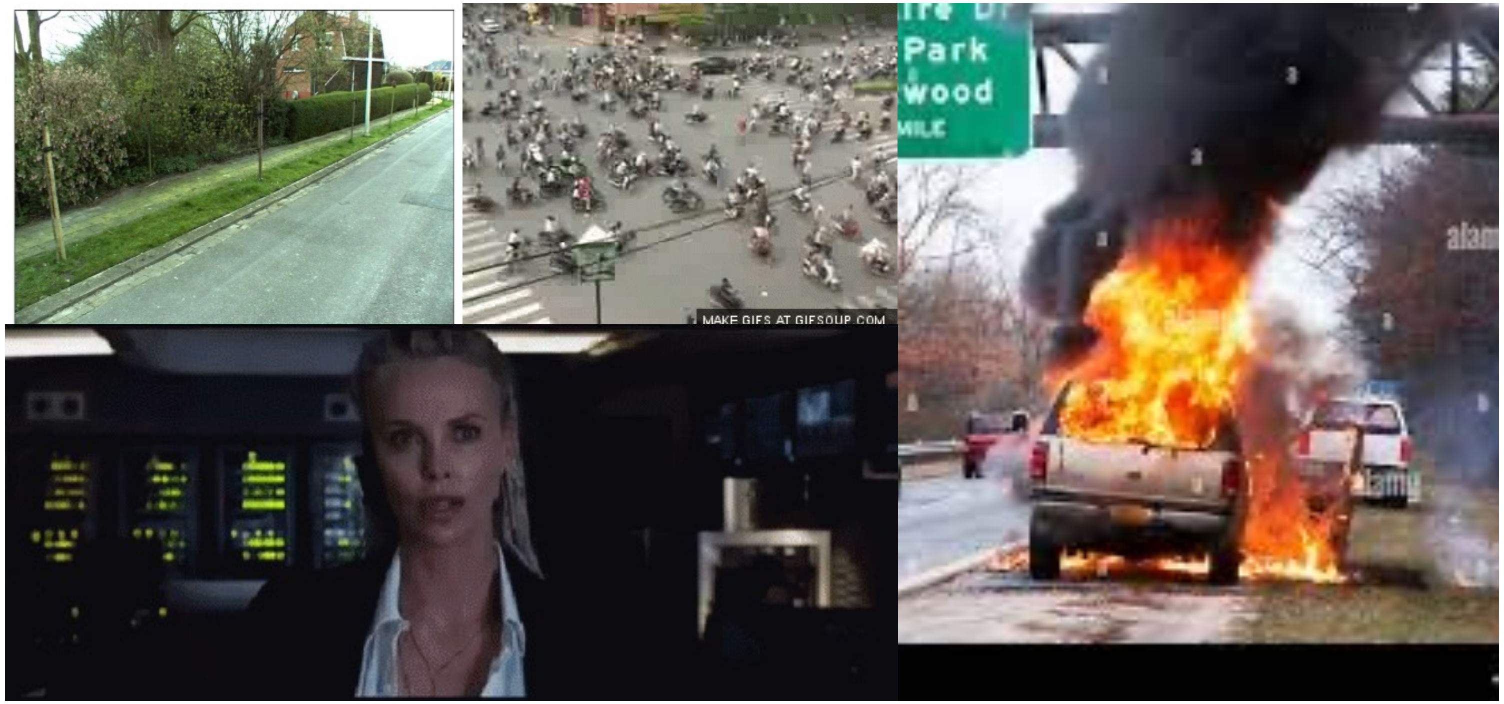
Objectives

Objectives in Part III

- Challenging conditions at training
- Inference
  - Deficiencies at Inference
- Overcoming deficiencies at Inference
  - Anomaly Detection
  - Uncertainty
  - Explainability
- Case study 1: Robustness to challenging conditions
- Case study 2: Aberrant Object Detection

Perception in AVs
Technical Challenges

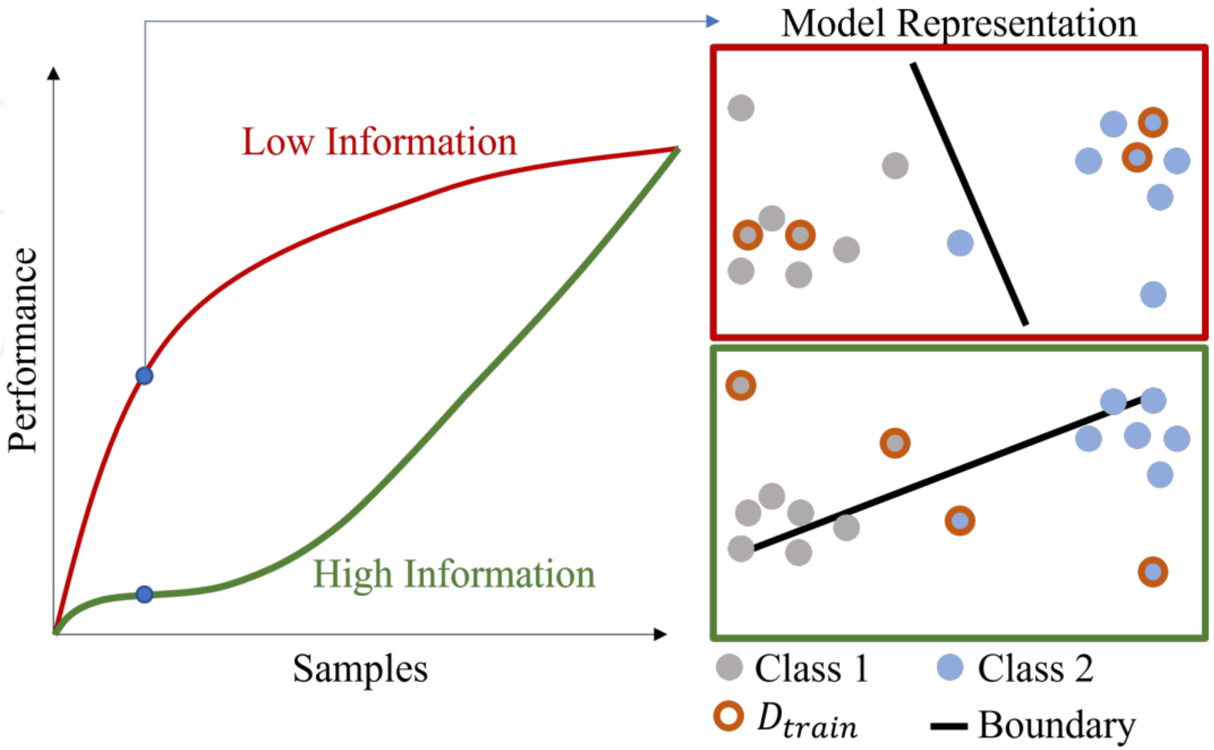
- **Challenging weather**
- **Challenging sensing**
- **Challenging environments**
- Context awareness
- Embedded perception
- V2X perception



Challenging Conditions in Deep Learning

Integrating Challenging Conditions in Training

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



- The first instance of training must occur with less informative samples
- Less informative:
  - Highway scenarios
  - Parking
  - No accidents
  - No aberrant events

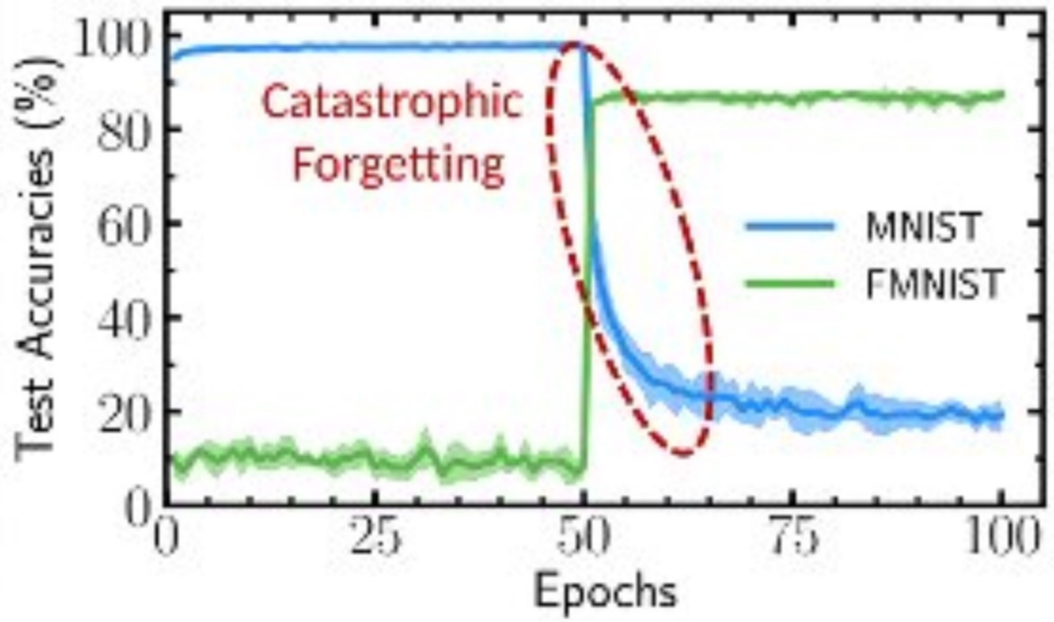
Novel samples = Most Informative

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*.

Challenging Conditions in Deep Learning

Integrating Challenging Conditions in Training

Subsequent training must <u>not</u> focus only on novel data



Catastrophic Forgetting

- 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

Handle challenging conditions at Inference!

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

Inference

What is Inference?

Ability of a system to predict correctly on novel data

Novel data sources:

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

Model Train

At Deployment

Ability of a system to predict correctly on novel data

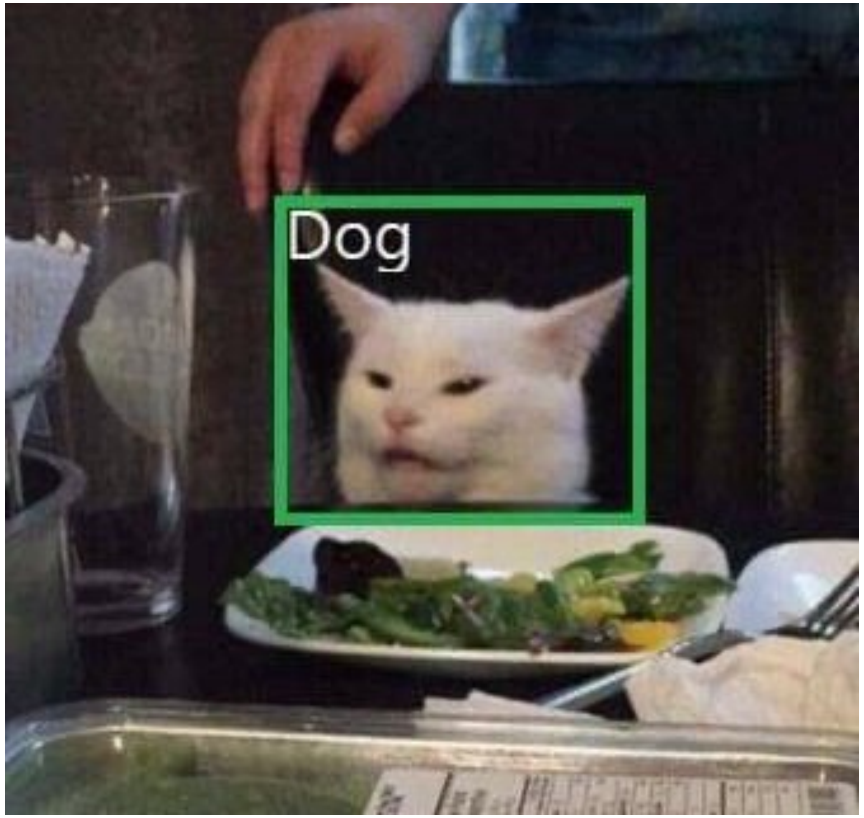
Novel data sources

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

→ Trained Model → Cat

Inference

Deficiencies at Inference



"The best-laid plans of sensors and networks often go awry"

- Engineers, probably

Inference

Overcoming Deficiencies at Inference

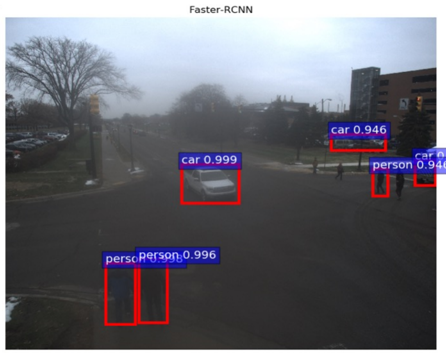
What is required when networks are met with challenging data at inference?

To overcome deficiencies, predictions from neural networks must be equipped with:

- Anomaly scores: How *close* to the training data is the novel data at inference?
- Uncertainty scores: How close to the *best* possible network is the trained network?
- Contextual Explainability: How *relevant* are the network explanations for its prediction?

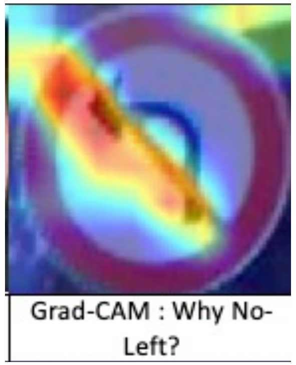
Training data

Anomalous data



Certain objects

Uncertain objects



'Why P'

'Why P, rather than Q?'

Inference

Overcoming Deficiencies at Inference

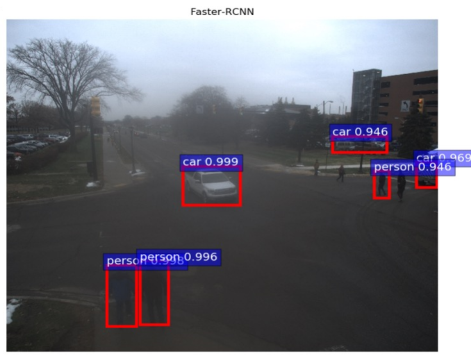
What is required when networks are met with challenging data at inference?

To overcome deficiencies, predictions from neural networks must be equipped with:

- **Anomaly scores**: How *close* to the training data is the novel data at inference?
- Uncertainty scores: How close to the *best* possible network is the trained network?
- Contextual Explainability: How *relevant* are the network explanations for its prediction?

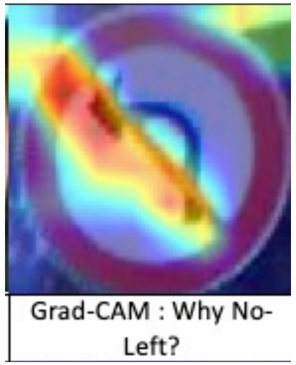
Training data

Anomalous data



Certain objects

Uncertain objects



'Why P'

'Why P, rather than Q?'

Backpropagated Gradient Representations for Anomaly Detection

Gukyeong Kwon, PhD
Amazon AWS

Mohit Prabhushankar, PhD
Postdoc, Georgia Tech

Ghassan AlRegib, PhD
Professor, Georgia Tech

SCAN ME

Anomalies

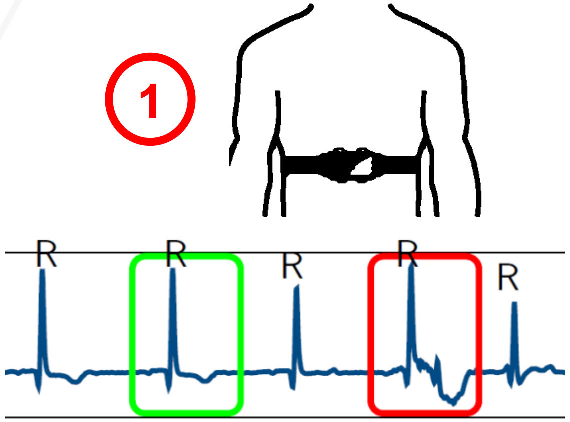
Finding Rare Events in Normal Patterns

Backpropagated Gradient Representations for Anomaly Detection

SCAN ME

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

1



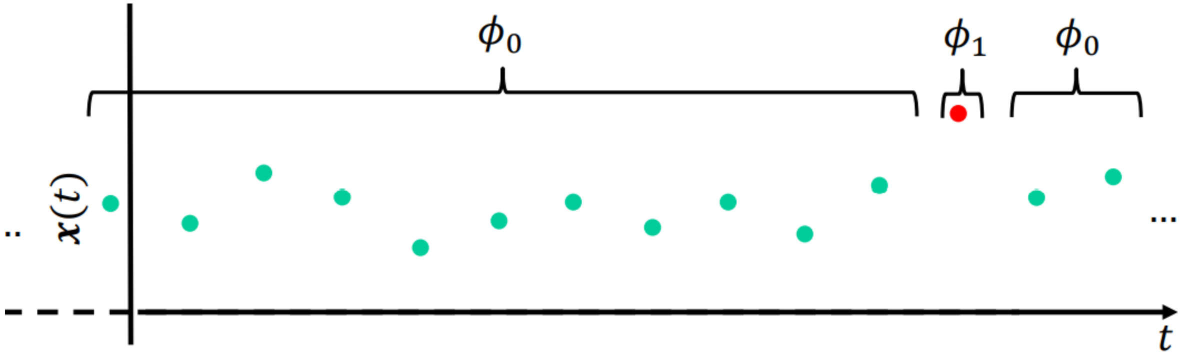
2

Statistical Definition:

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

Goal: Detect ϕ_1

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



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

Anomalies

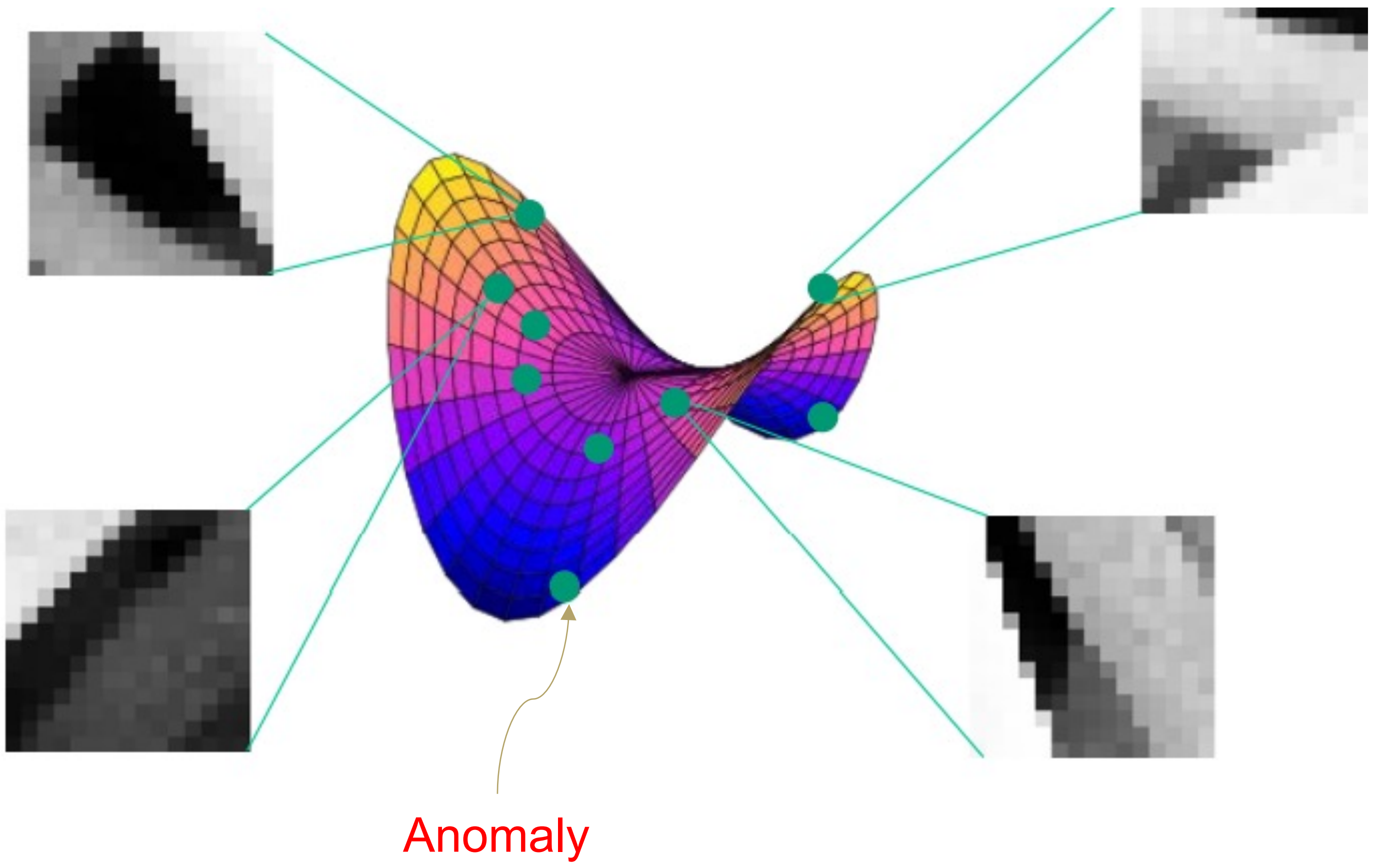
Steps for Anomaly Detection

Backpropagated Gradient Representations for Anomaly Detection

SCAN ME

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

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

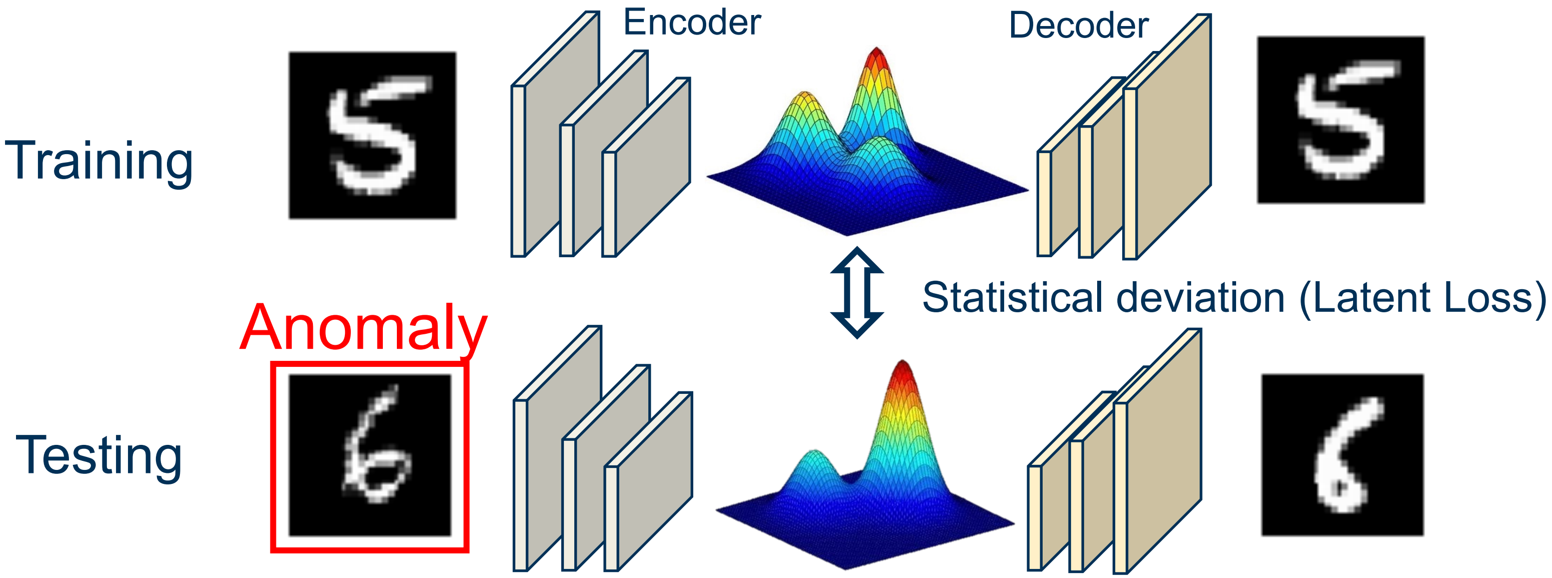


Constraining Manifolds
General Constraints

Backpropagated Gradient Representations for Anomaly Detection

Constrained Representation

Activations are constrained using GANs, VAEs, etc.

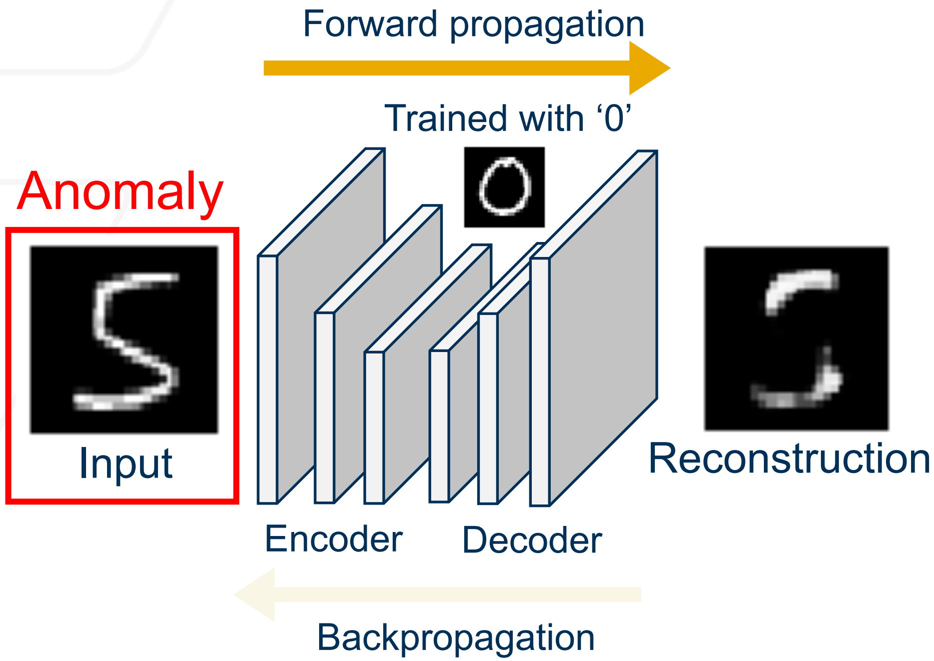


[1] David MJ Tax and Robert PW Duin. Support vector data description. Machine learning, 54(1):45–66, 2004.
[2] Yaxiang Fan, Gongjian Wen, Deren Li, Shaohua Qiu, and Martin D Levine. Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder. arXiv preprint arXiv:1805.11223, 2018. 1, 2
[3] S. Pidhorskyi, R. Almohsen, and G. Doretto, "Generative probabilistic novelty detection with adversarial autoencoders," in Advances in Neural Information Processing Systems, 2018, pp. 6822–6833.
[4] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 481–490.

Constraining Manifolds
Gradient-based Constraints

Backpropagated Gradient Representations for Anomaly Detection

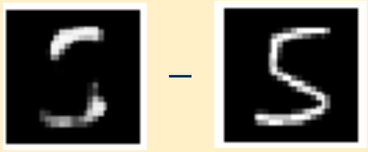
SCAN ME



Activation Constraints

Activation-based representation
(Data perspective)

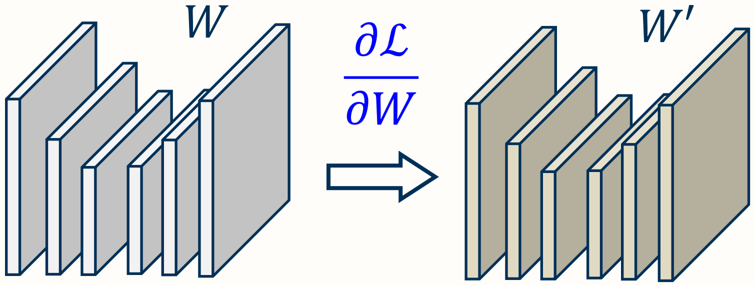
e.g. Reconstruction error ($\mathcal{L}$)



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

Gradient Constraints

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



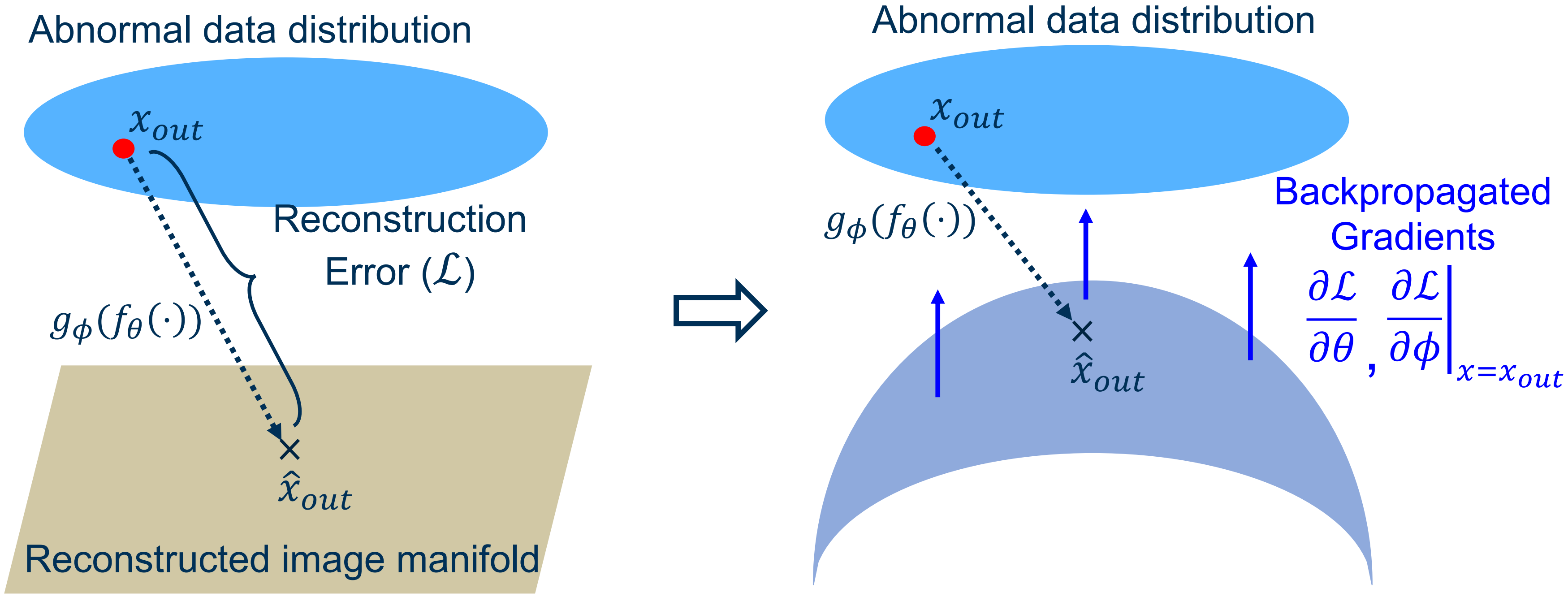
How much **model update** is required by the input?

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

Constraining Manifolds
Advantages of Gradient-based Constraints

Backpropagated Gradient Representations for Anomaly Detection

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



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

Backpropagated Gradient Representations for Anomaly Detection

SCAN ME

Abnormal “class” detection (CIFAR-10)

e.g.

Normal

Abnormal

AUROC Results

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

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

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

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

GradCON: Gradient Constraint

Aberrant Condition Detection

SCAN ME

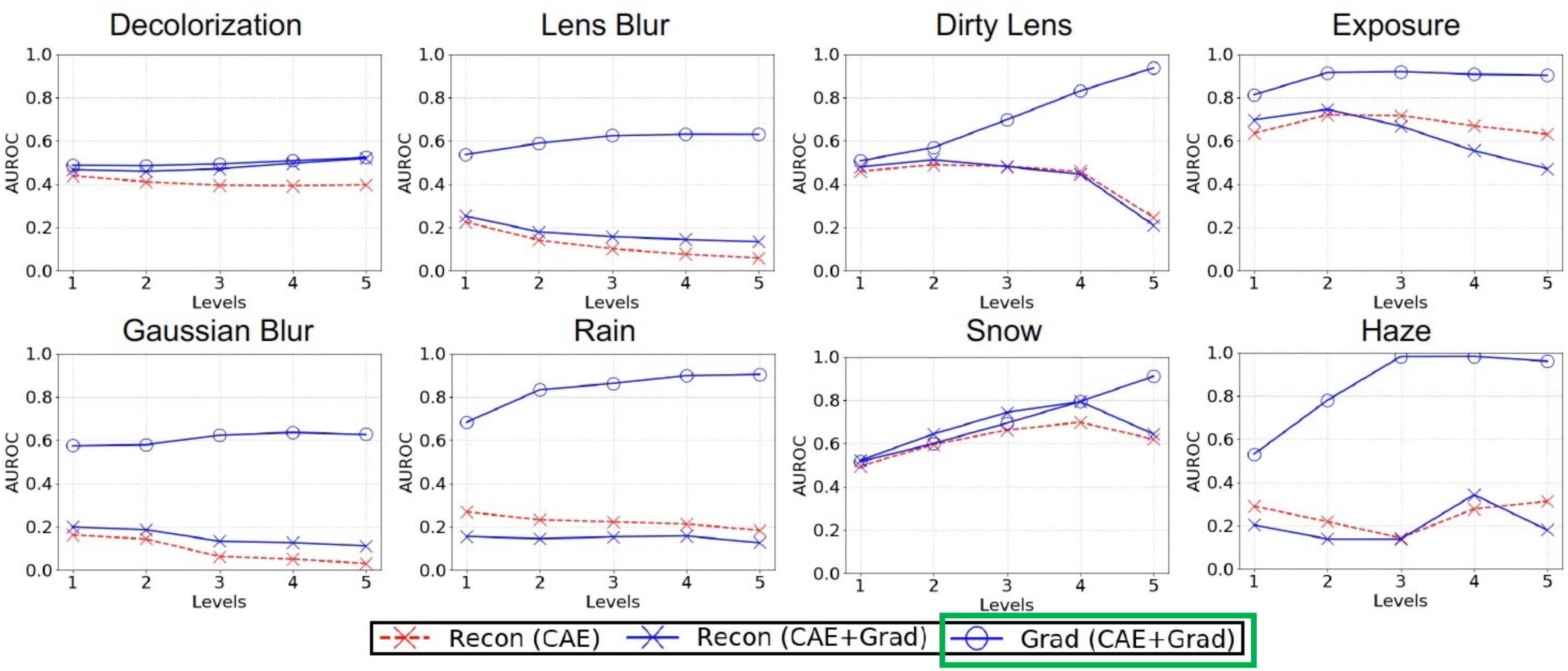
Backpropagated Gradient Representations for Anomaly Detection

Abnormal "condition" detection (CURE-TSR)

Normal

Abnormal

AUROC Results



Recon: Reconstruction error, Grad: Gradient loss

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

Inference
Overcoming Deficiencies at Inference

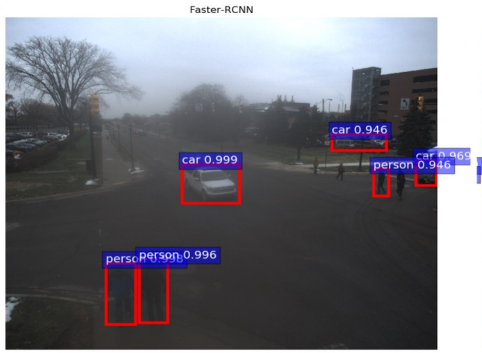
What is required when networks are met with challenging data at inference?

To overcome deficiencies, predictions from neural networks must be equipped with:

- Anomaly scores: How *close* to the training data is the novel data at inference?
- **Uncertainty scores**: How close to the *best* possible network is the trained network?
- Contextual Explainability: How *relevant* are the network explanations for its prediction?

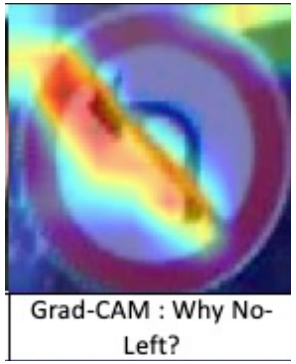
Training data


Anomalous data



Certain objects

Uncertain objects



'Why P'

'Why P, rather than Q?'

Probing the Purview of Neural Networks via Gradient Analysis


Jinsol Lee,
PhD Candidate

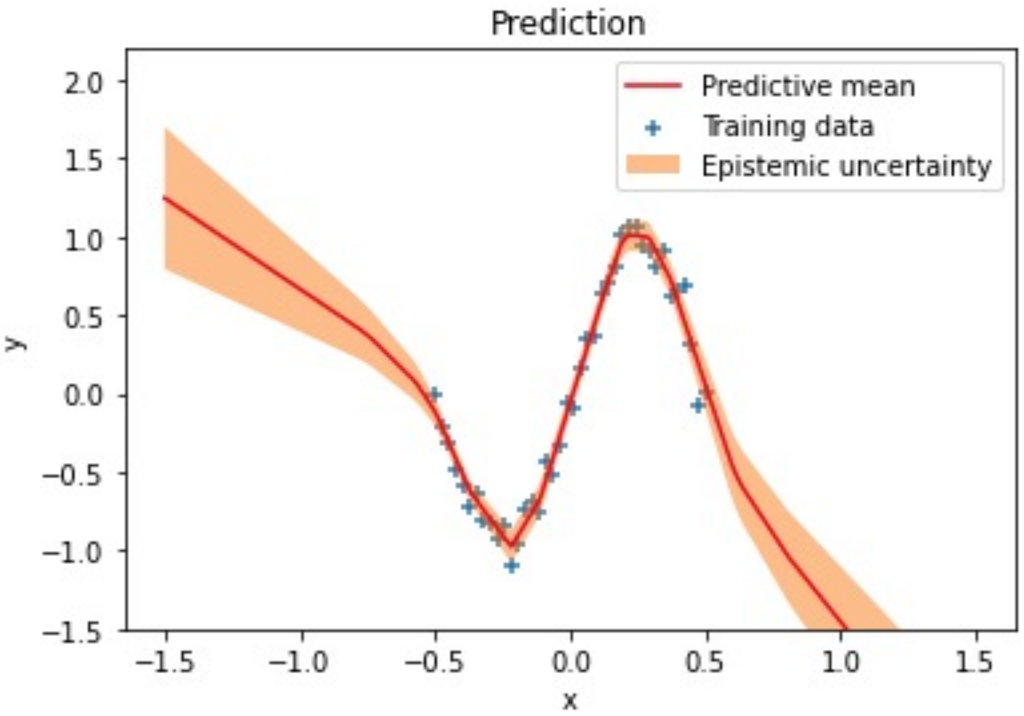

Mohit Prabhushankar, PhD
Postdoc


Ghassan AlRegib, PhD
Professor

SCAN ME

SCAN ME

Uncertainty is a model knowing that it does not know



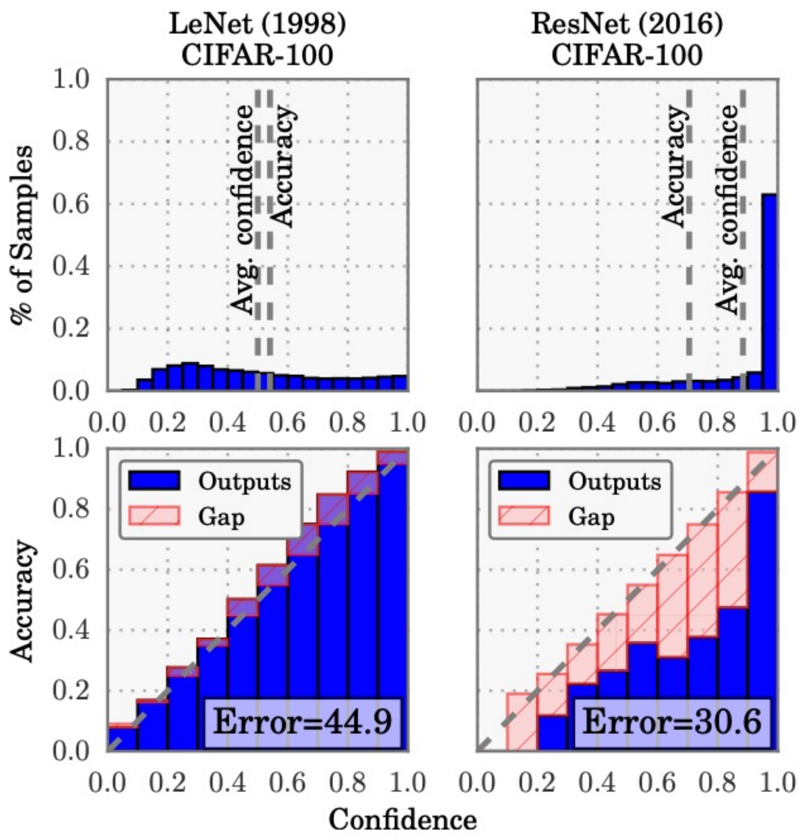
A simple example: More the training data, lesser the uncertainty

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

SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

Uncertainty is a model knowing that it does not know

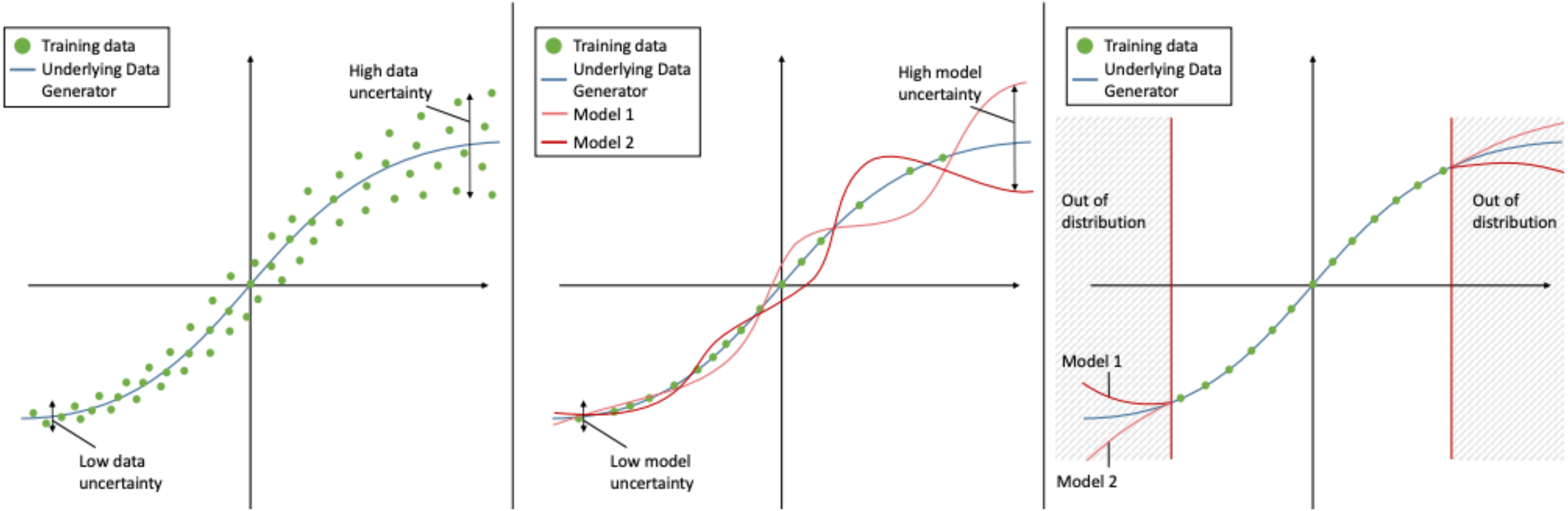


- Larger the model, more misplaced is a network's confidence
- On ResNet, the gap between prediction accuracy and its corresponding confidence is significantly high
- On OOD data, uncertainty is not easy to quantify

Guo, Chuan, et al. "On calibration of modern neural networks." *International conference on machine learning*. PMLR, 2017.

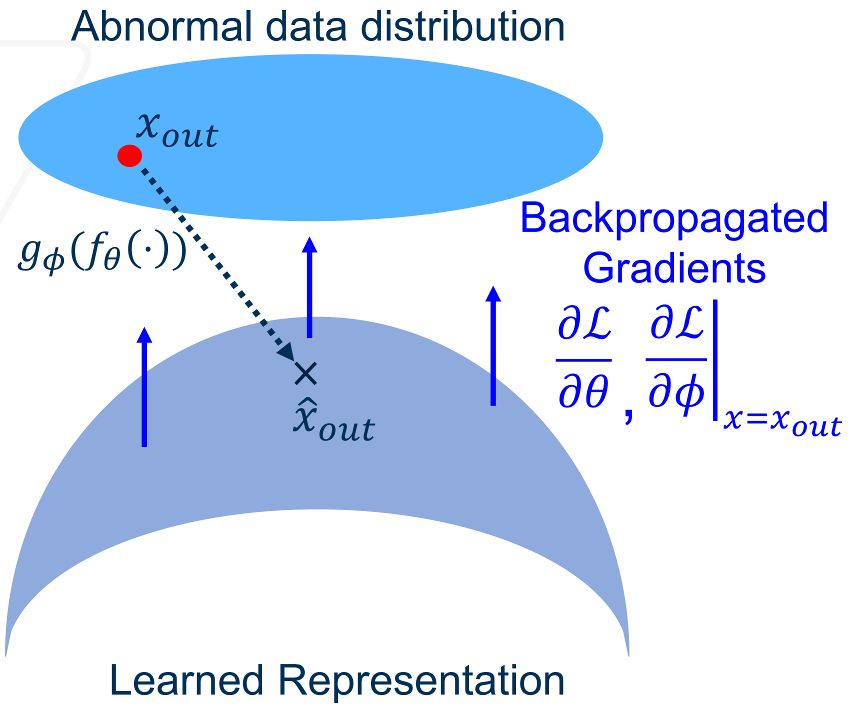
SCAN ME

Two major types of uncertainty: Uncertainty in data and uncertainty in model



Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., ... & Zhu, X. X. (2021). A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*.

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 or ground truth

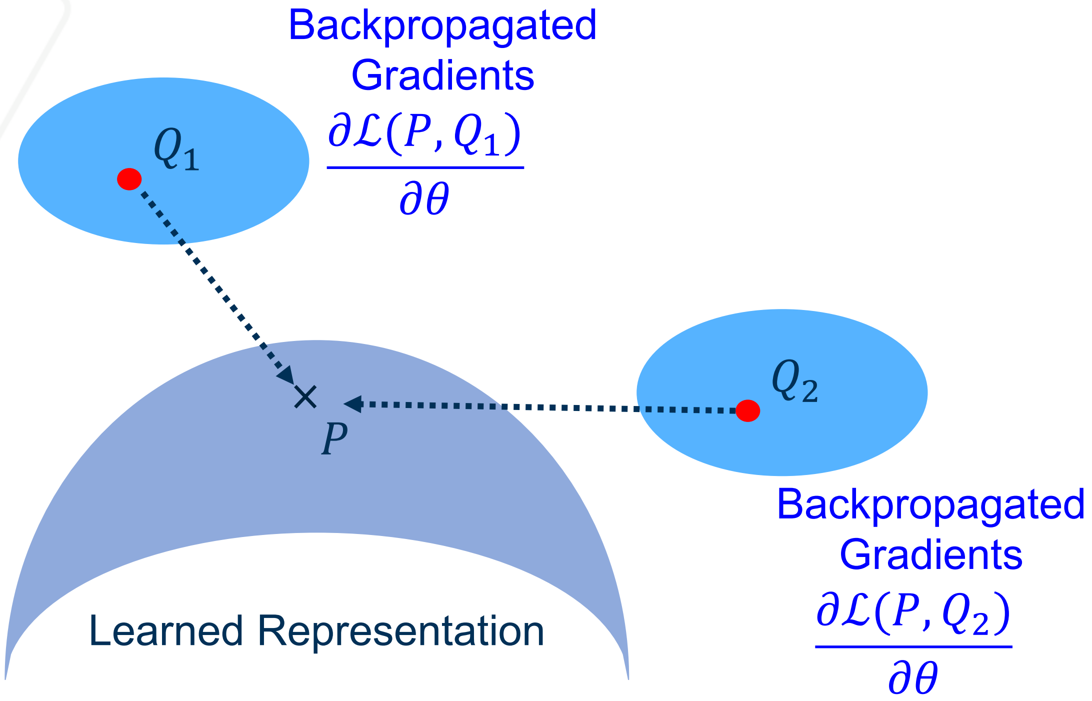
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

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 or ground truth

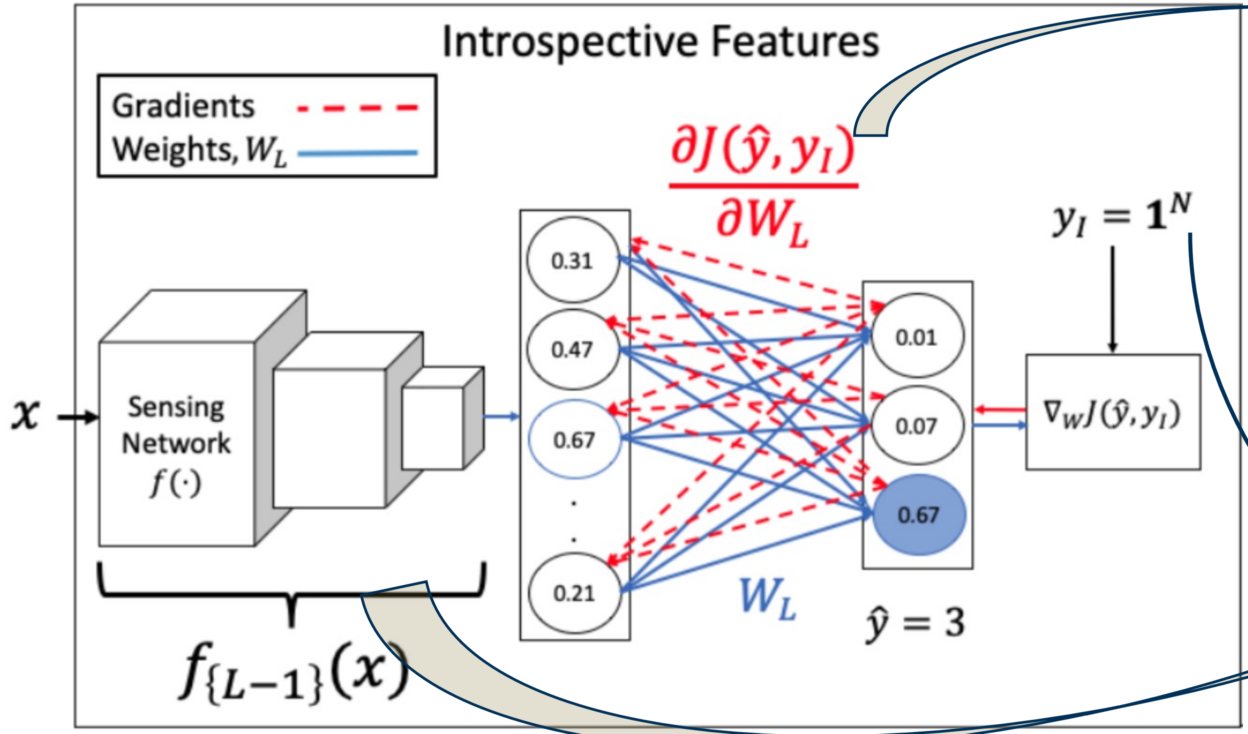
- **We backpropagate all possible classes - $Q_1, Q_2 \ldots Q_N$ by backpropagating N one-hot vectors**
- Higher the distance to all classes, higher the uncertainty score

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

SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

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
Deriving Gradient Features

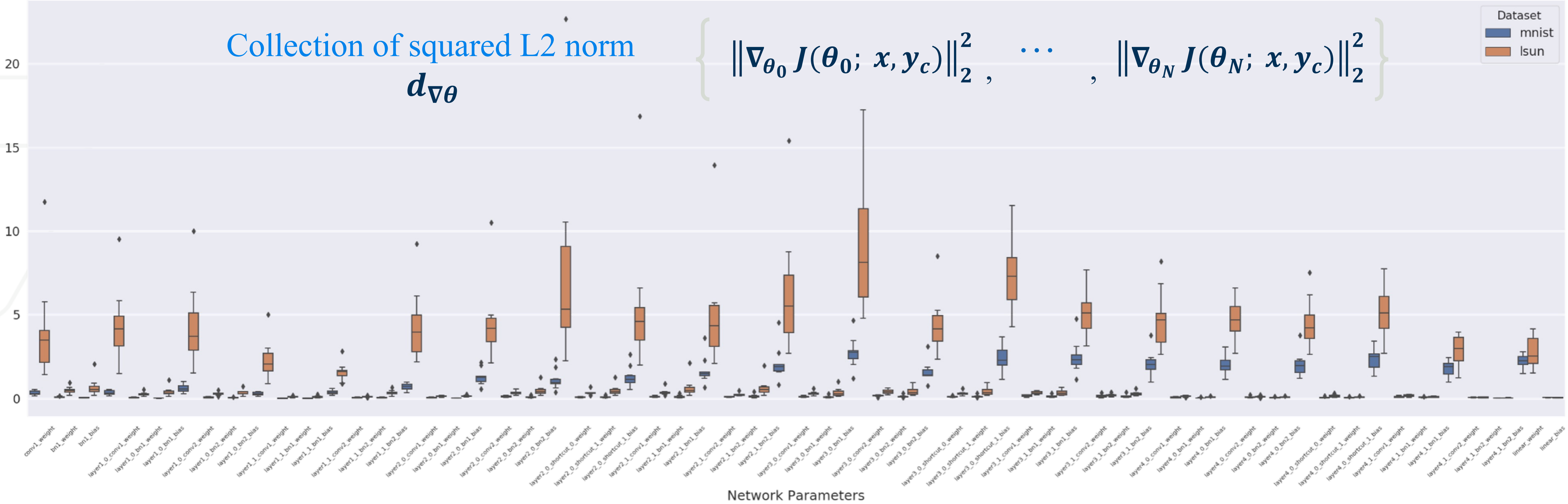
Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

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

IEEE IV 2023
IEEE INTELLIGENT VEHICLES SYMPOSIUM
[Tutorial] | [Ghassan AlRegib and Mohit Prabhushankar] | [June 4, 2023]

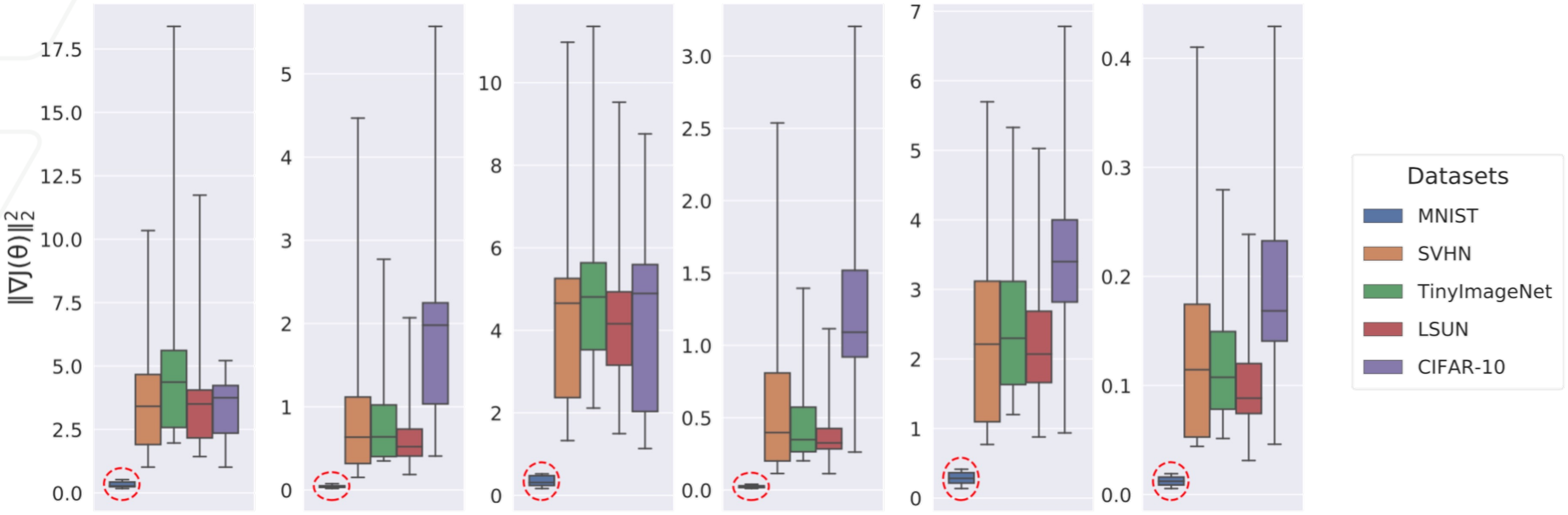
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

OLIVES @GeorgiaTech
Georgia Tech

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

IEEE IV 2023 IEEE INTELLIGENT VEHICLES SYMPOSIUM

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

OLIVES @GeorgiaTech

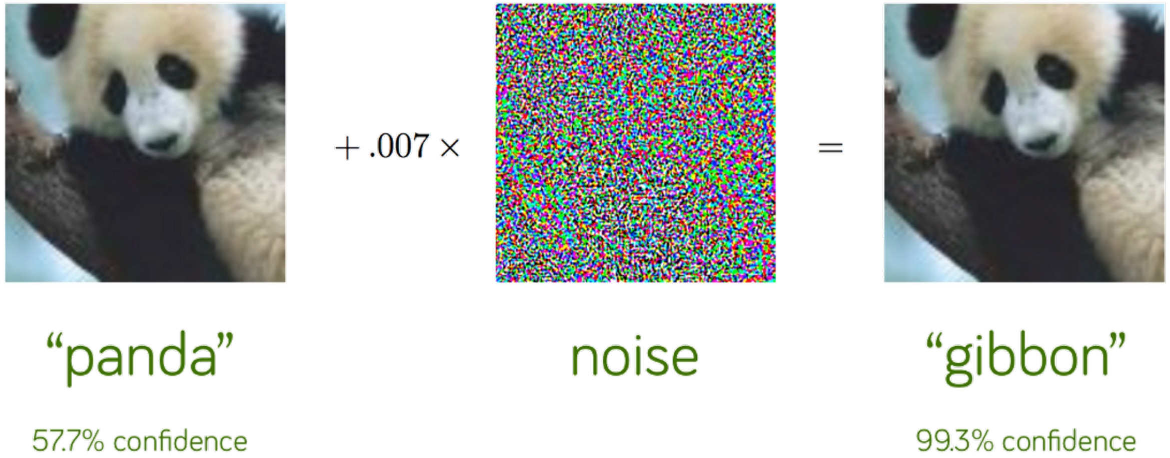
Georgia Tech

Gradient-based Uncertainty
Uncertainty Results 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 Results 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.

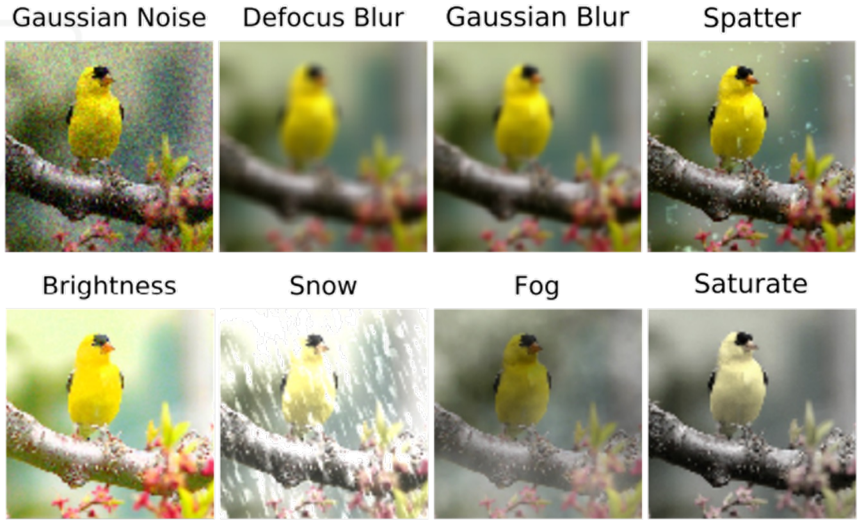
Gradient-based Uncertainty
Uncertainty Results to Detect Challenging Conditions

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

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

CIFAR-10-C



CURE-TSR

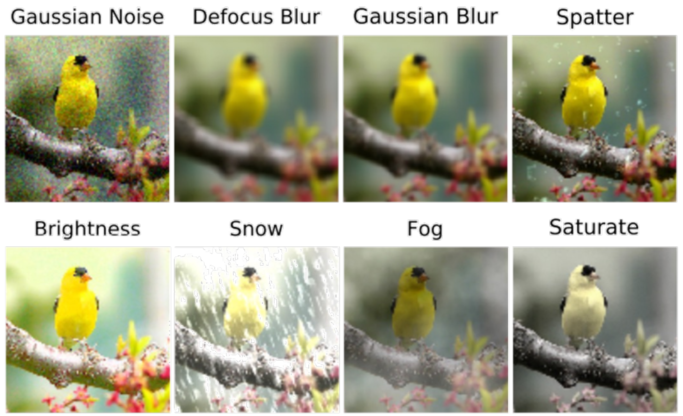
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Gradient-based Uncertainty
Uncertainty Results to Detect Challenging Conditions

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

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



IEEE IV 2023 IEEE INTELLIGENT VEHICLES SYMPOSIUM

[Tutorial] | [Ghassan AlRegib and Mohit Prabhushankar] | [June 4, 2023]

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

OLIVES @GeorgiaTech

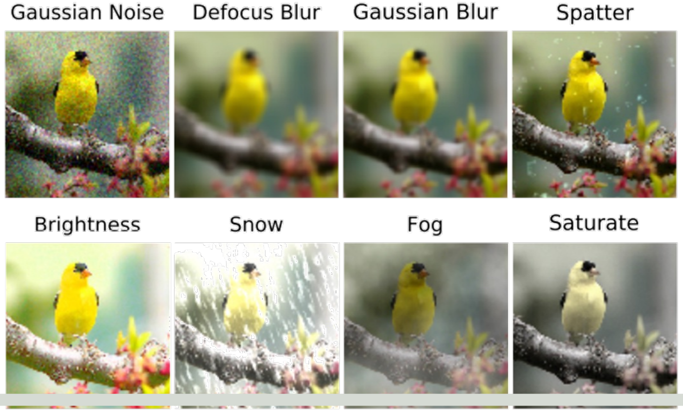
Georgia Tech

Gradient-based Uncertainty
Uncertainty Results to Detect Challenging Conditions

SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

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



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

Overcoming Deficiencies at Inference

What is required when networks are met with challenging data at inference?

To overcome deficiencies, predictions from neural networks must be equipped with:

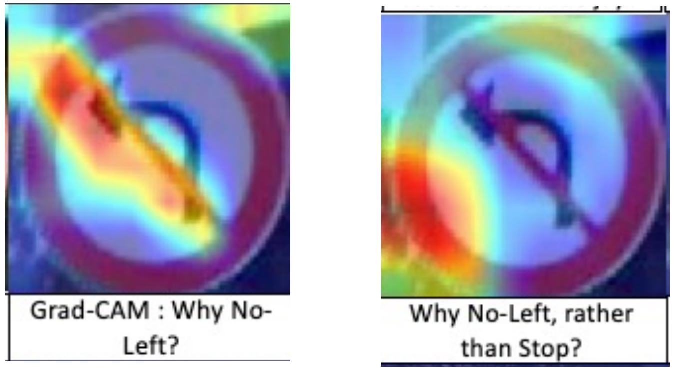
- Anomaly scores: How *close* to the training data is the novel data at inference?
- Uncertainty scores: How close to the *best* possible network is the trained network?
- **Contextual Explainability**: How *relevant* are the network explanations for its prediction?

Training data

Anomalous data

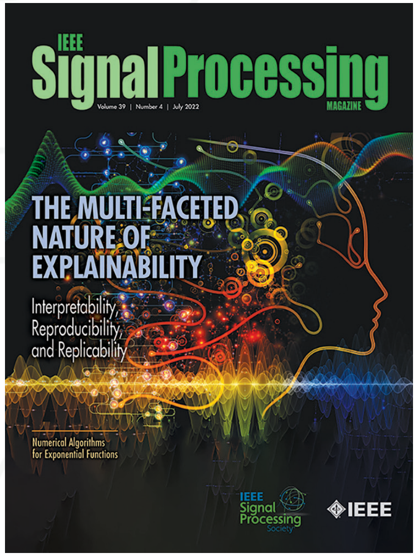
Certain objects

Uncertain objects



'Why P'

'Why P, rather than Q?'



Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

Mohit Prabhushankar, PhD
Postdoc

Ghassan AlRegib, PhD
Professor

SCAN ME

Explanations
What are Visual Explanations?

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

- **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
Why Explainability?

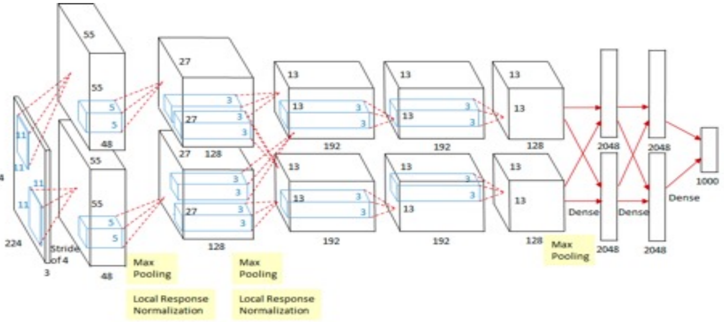
Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

Explainability matters establishes trust in deep learning systems by developing *transparent* models that can explain *why they predict what they predict* to humans

Algorithm

Data



Output

class scores

Explainability is useful in*:*

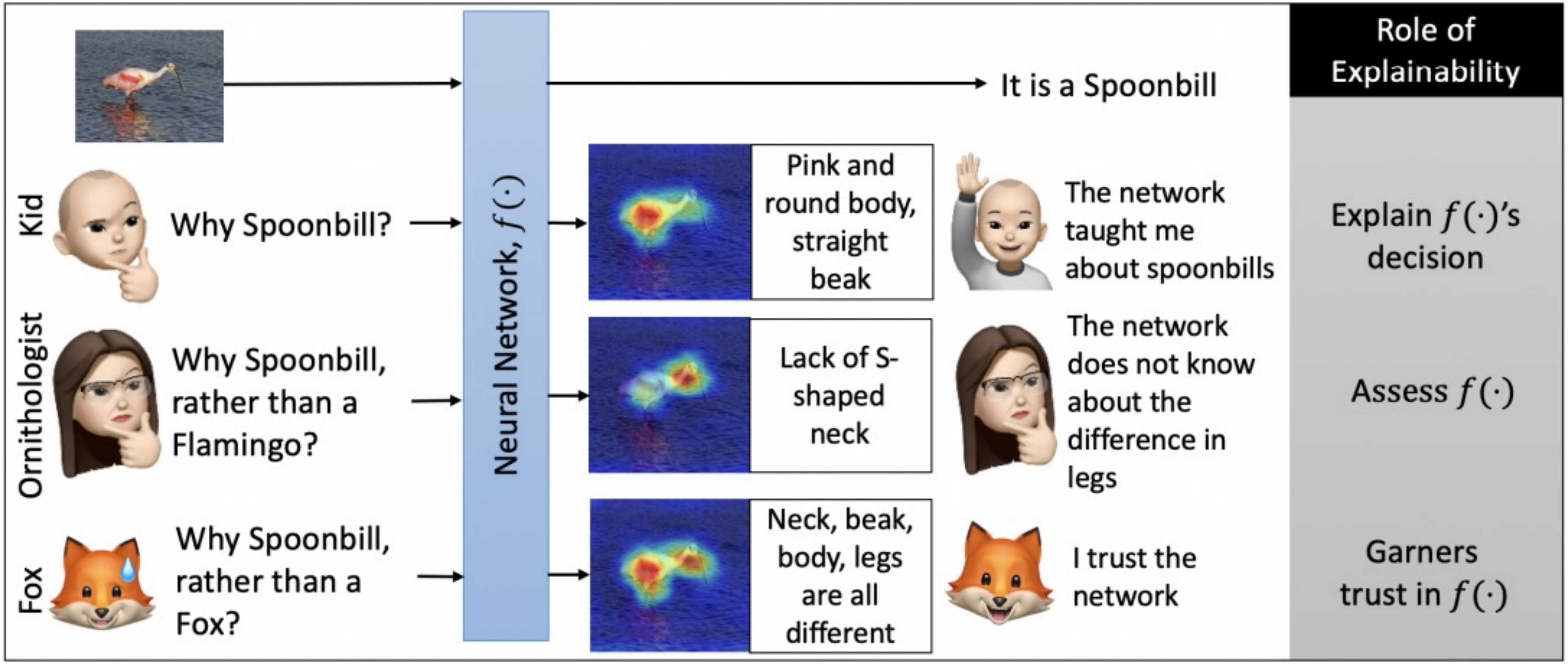
- Medical: help doctors diagnose
- Seismic: help interpreters label seismic data
- Autonomous Systems: build appropriate trust and confidence

Deep models act as algorithms that take data and output something **without** being able to **explain** their methodology

Explanations
Role of Visual Explanations

SCAN ME

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations



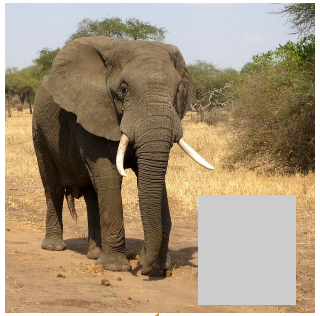
AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks: Towards relevant and contextual explanations. *IEEE Signal Processing Magazine*, *39*(4), 59-72.

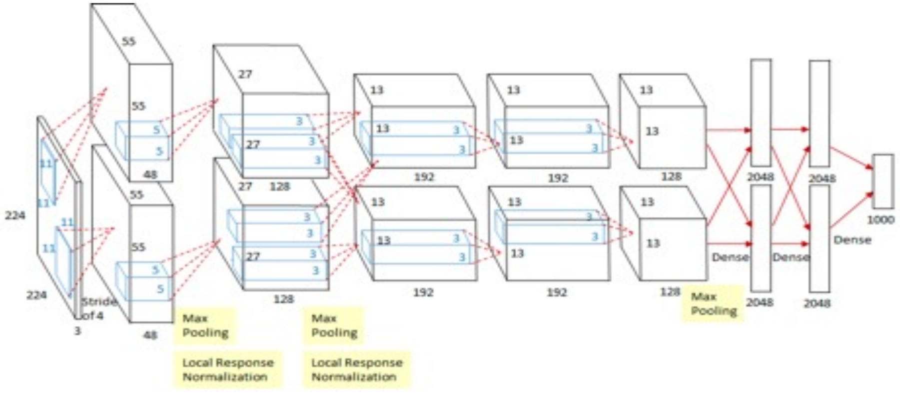
Explanations
Input Saliency via Occlusion

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

Intervention: Mask part of the image before feeding to CNN, check how much predicted probabilities change





P(elephant) = 0.95

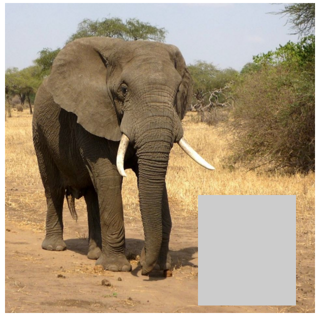
A gray patch or patch of average pixel value of the dataset
Note: not a black patch because the input images are centered to zero in the preprocessing.

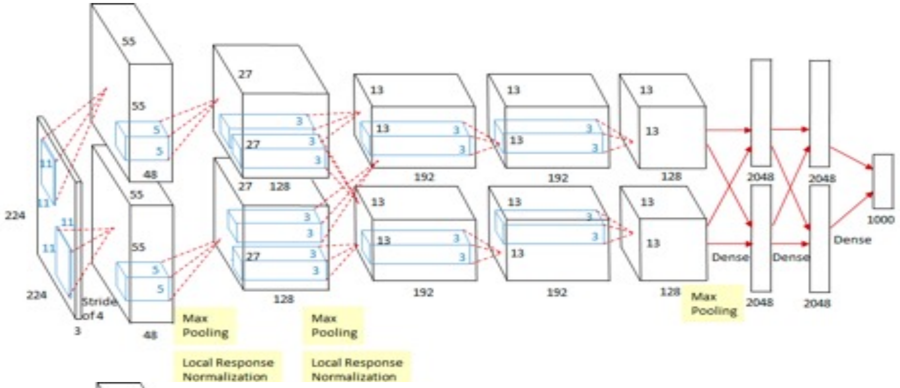
Zeiler and Fergus, “Visualizing and Understanding Convolutional Networks”, ECCV 2014

Explanations
Input Saliency via Occlusion

SCAN ME

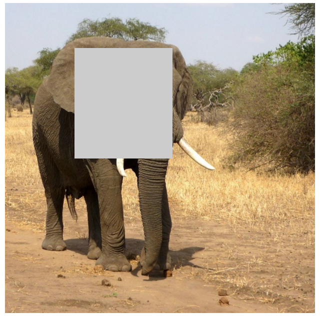
Intervention: Mask part of the image before feeding to CNN, check how much predicted probabilities change

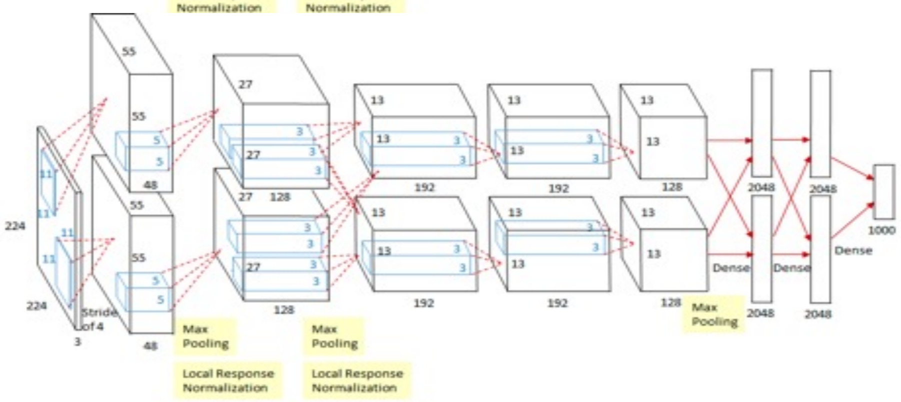




P(elephant) = 0.95

These pixels affect decisions more





P(elephant) = 0.75

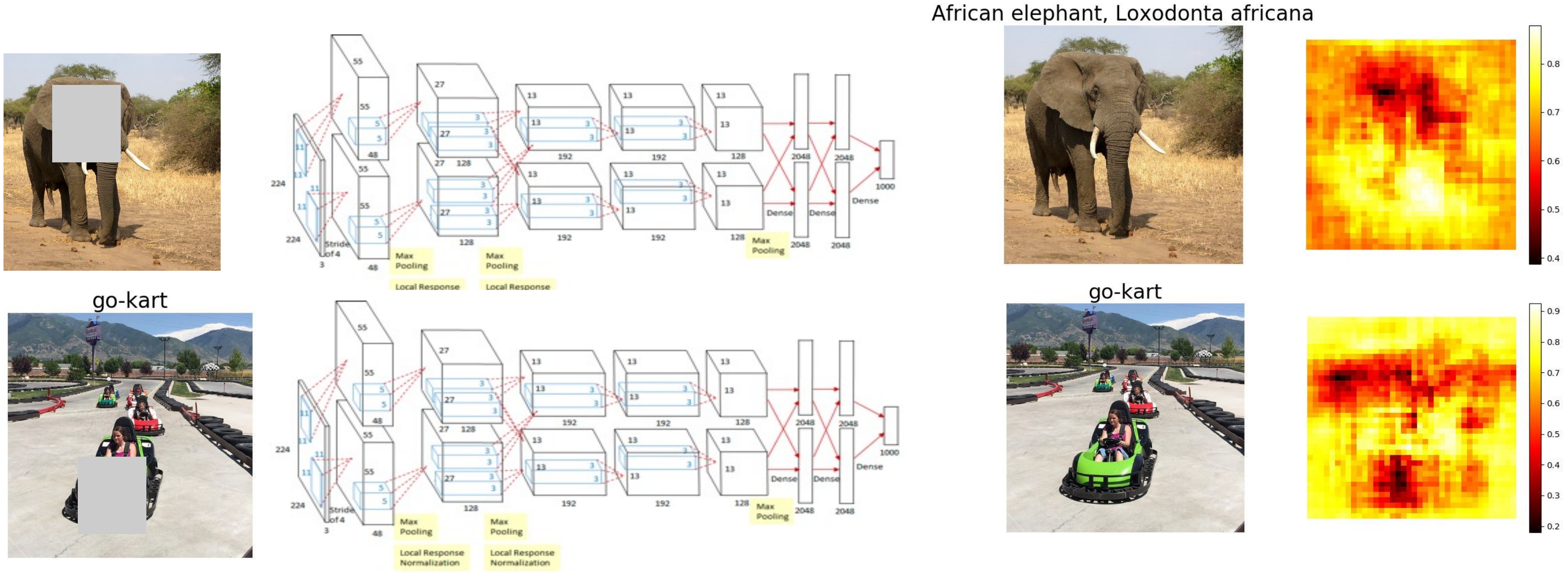
Zeiler and Fergus, “Visualizing and Understanding Convolutional Networks”, ECCV 2014

Explanations
Input Saliency via Occlusion

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

The network is trained with image- labels, but it is sensitive to the common visual regions in images



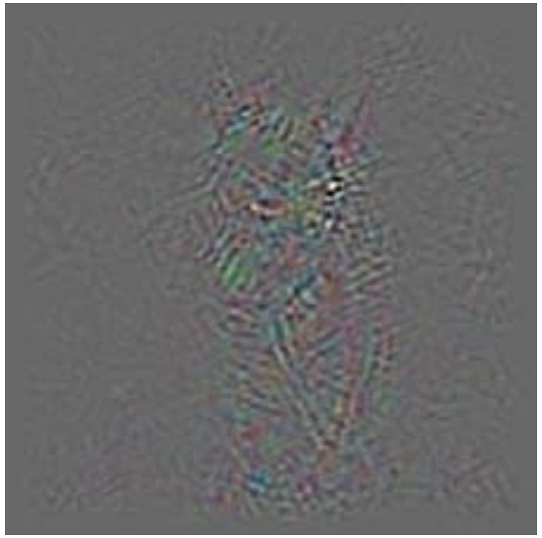
Zeiler and Fergus, “Visualizing and Understanding Convolutional Networks”, ECCV 2014

Explanations
Input Saliency via Gradients

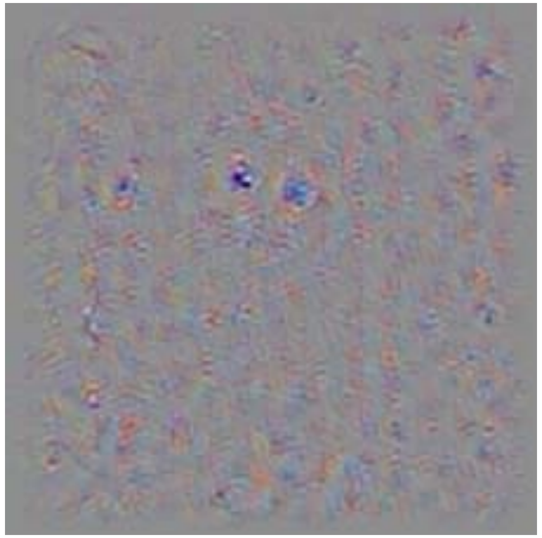
Gradients provide a one-shot means of perturbing the input that changes the output

Input

Vanilla Gradients



Deconvolution Gradients



Guided Backpropagation

However, localization remains an issue

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

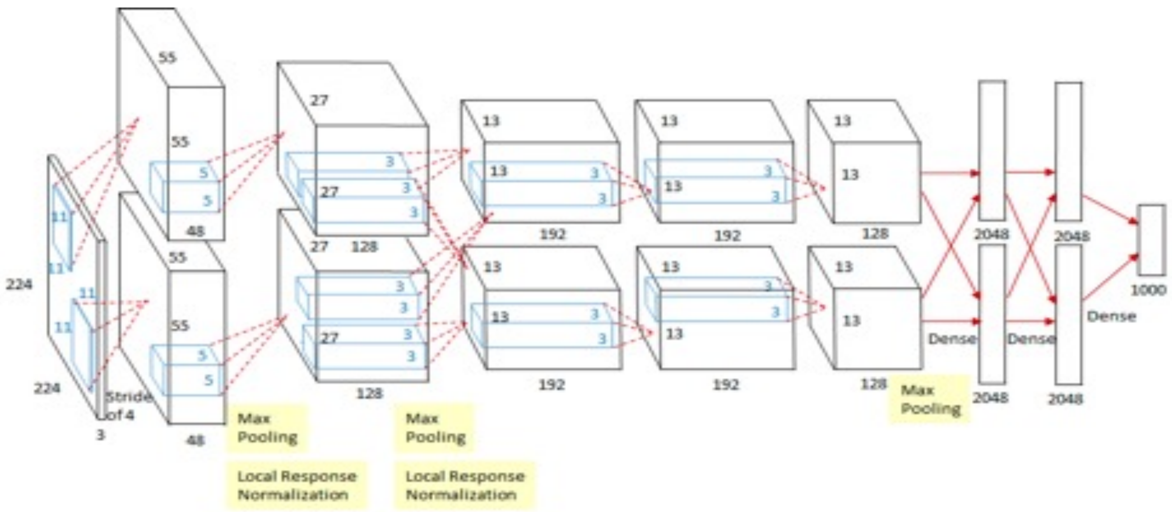
Gradient and Activation-based Explanations
GradCAM

SCAN ME

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

Gradients provide a one-shot means of perturbing the input that changes the output. Activations provide the localization.

- To find the important activations that are responsible for a particular class
- We want the activations:
  - **Class-discriminative** to reflect decision-making
  - **Preserve spatial information** to ensure spatial coverage of important regions



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.

IEEE IV 2023 IEEE INTELLIGENT VEHICLES SYMPOSIUM

OLIVES @GeorgiaTech

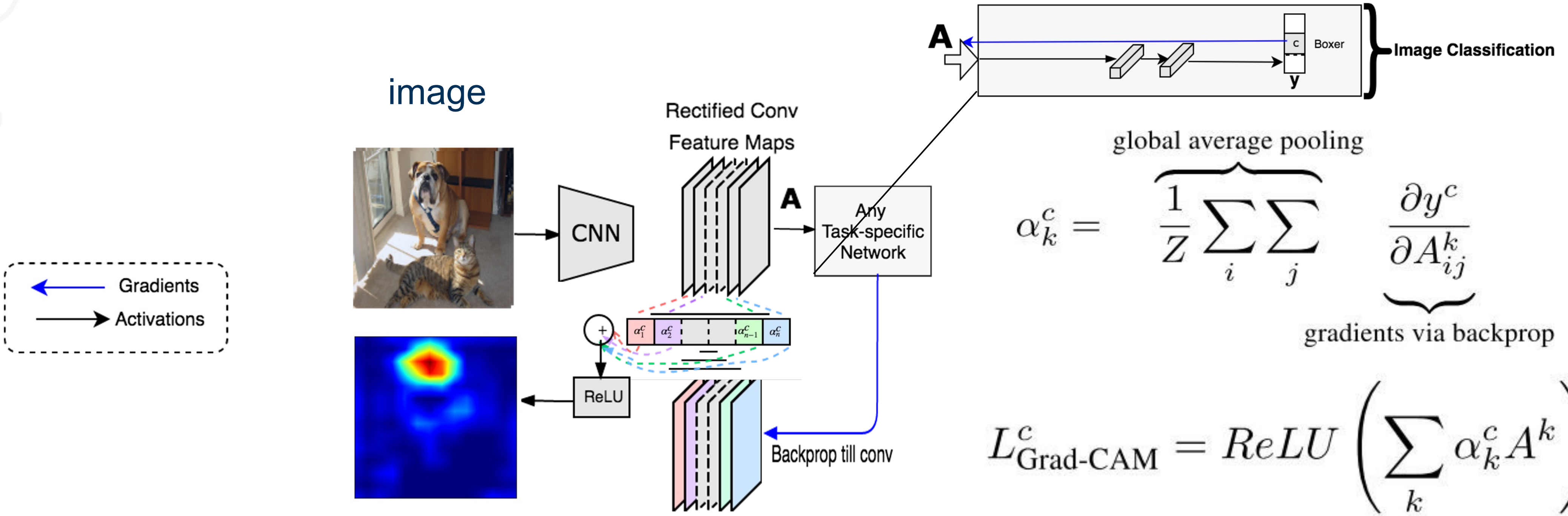
Georgia Tech

Gradient and Activation-based Explanations
GradCAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

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)

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

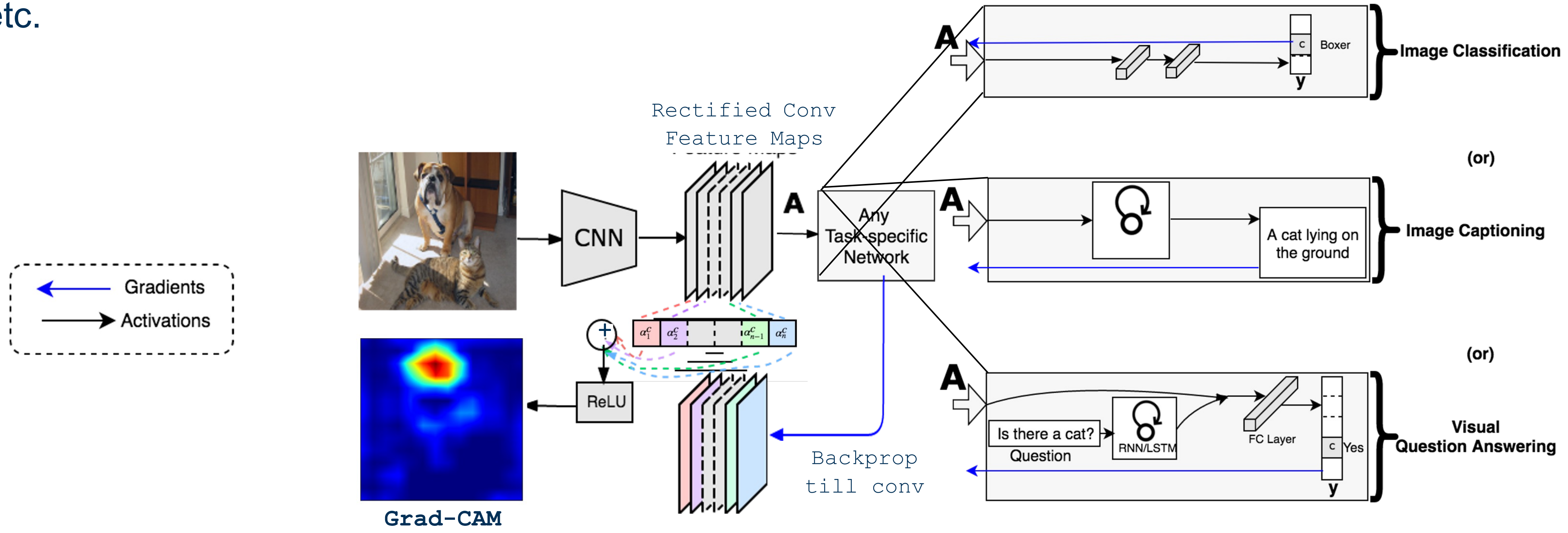
$$L_{\text{Grad-CAM}}^c = ReLU \underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

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

Extensions of GradCAM

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?

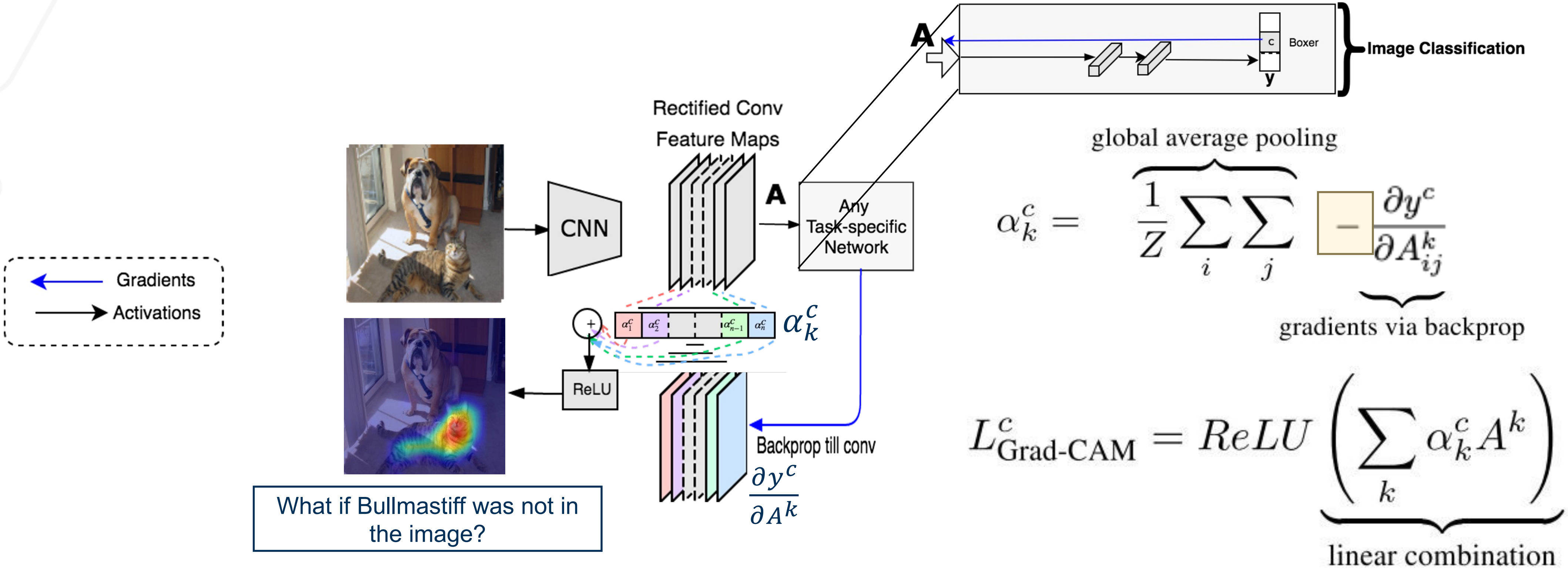
AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks: Towards relevant and contextual explanations. *IEEE Signal Processing Magazine*, *39*(4), 59-72.

Gradient and Activation-based Explanations

CounterfactualCAM: What if P is not there in the Image?

SCAN ME

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



Negating the gradients effectively removes these regions from analysis

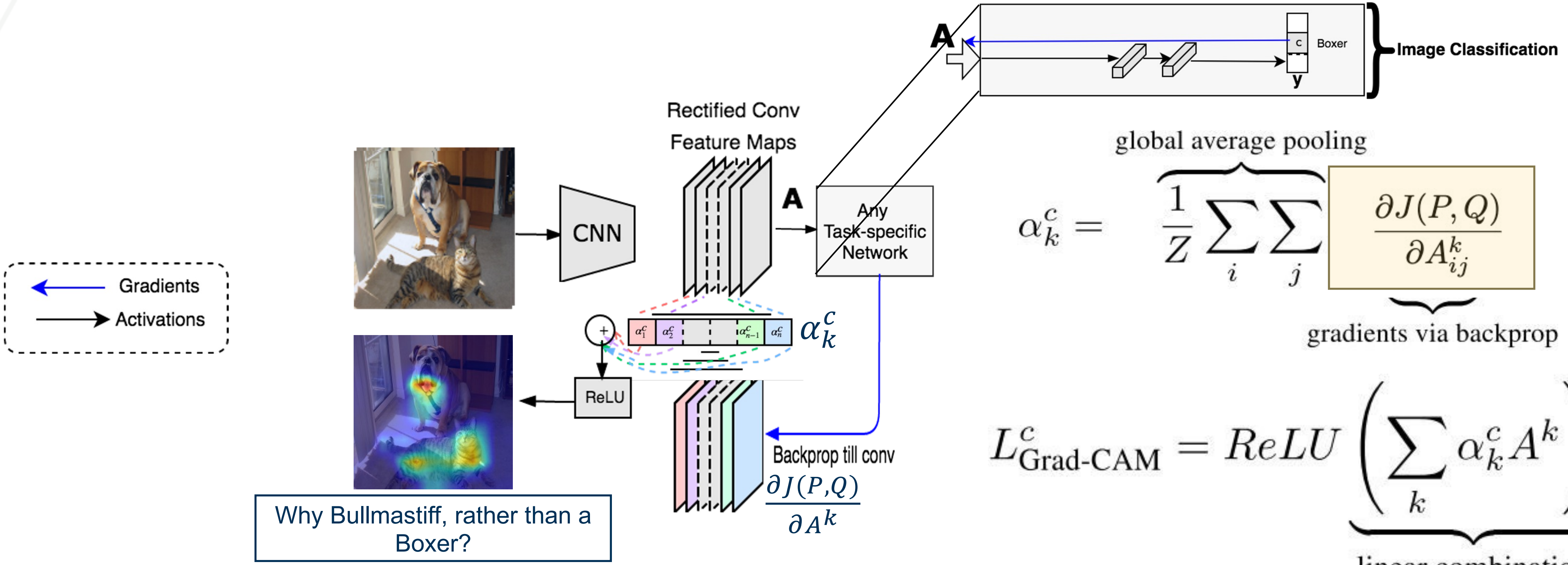
Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

Gradient and Activation-based Explanations
ContrastCAM: Why P, rather than Q?

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

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



Contrast-CAM

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

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

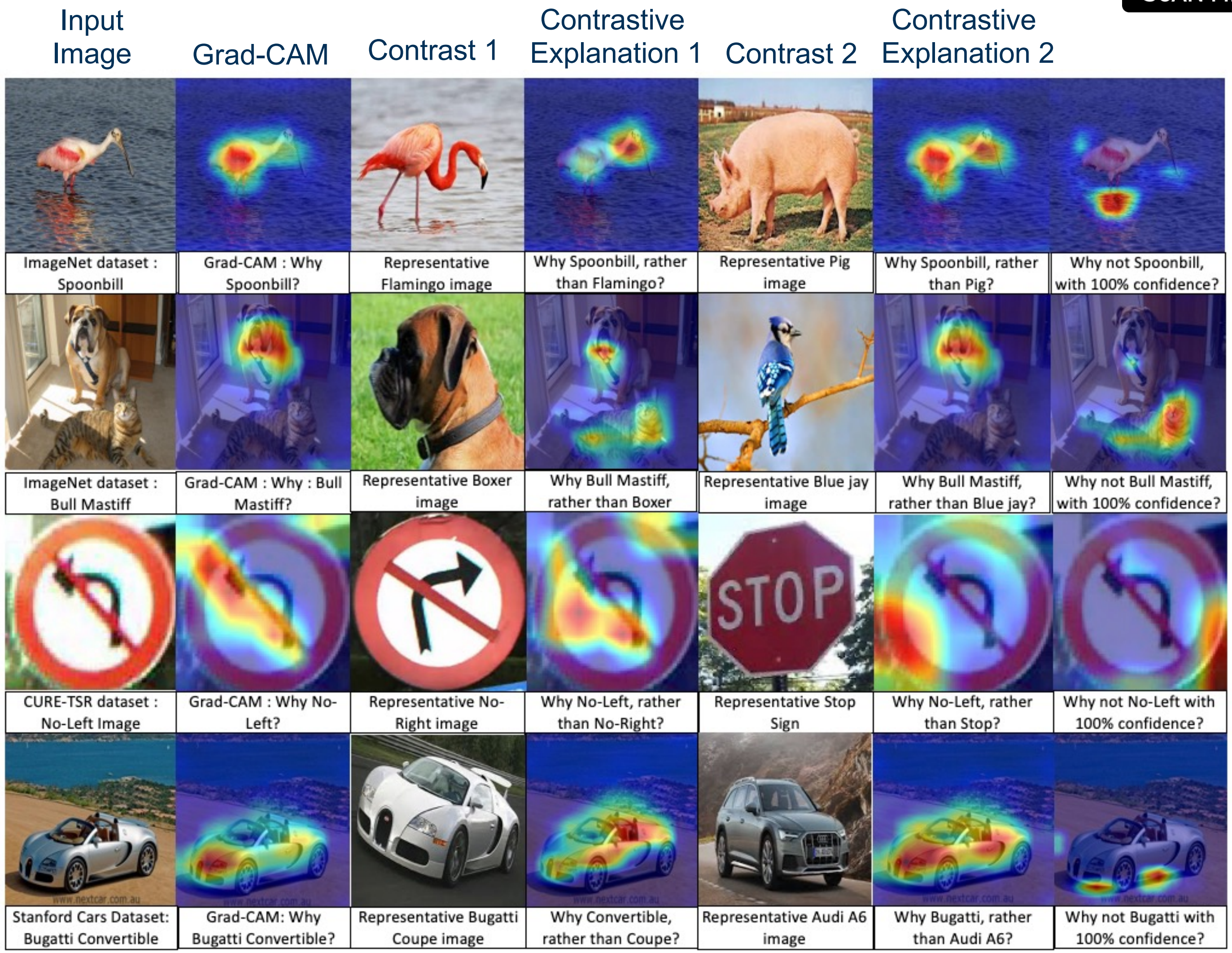
Backpropagating the loss highlights the differences between classes P and Q.

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 of GardCAM, CounterfactualCAM, and ContrastCAM

SCAN ME



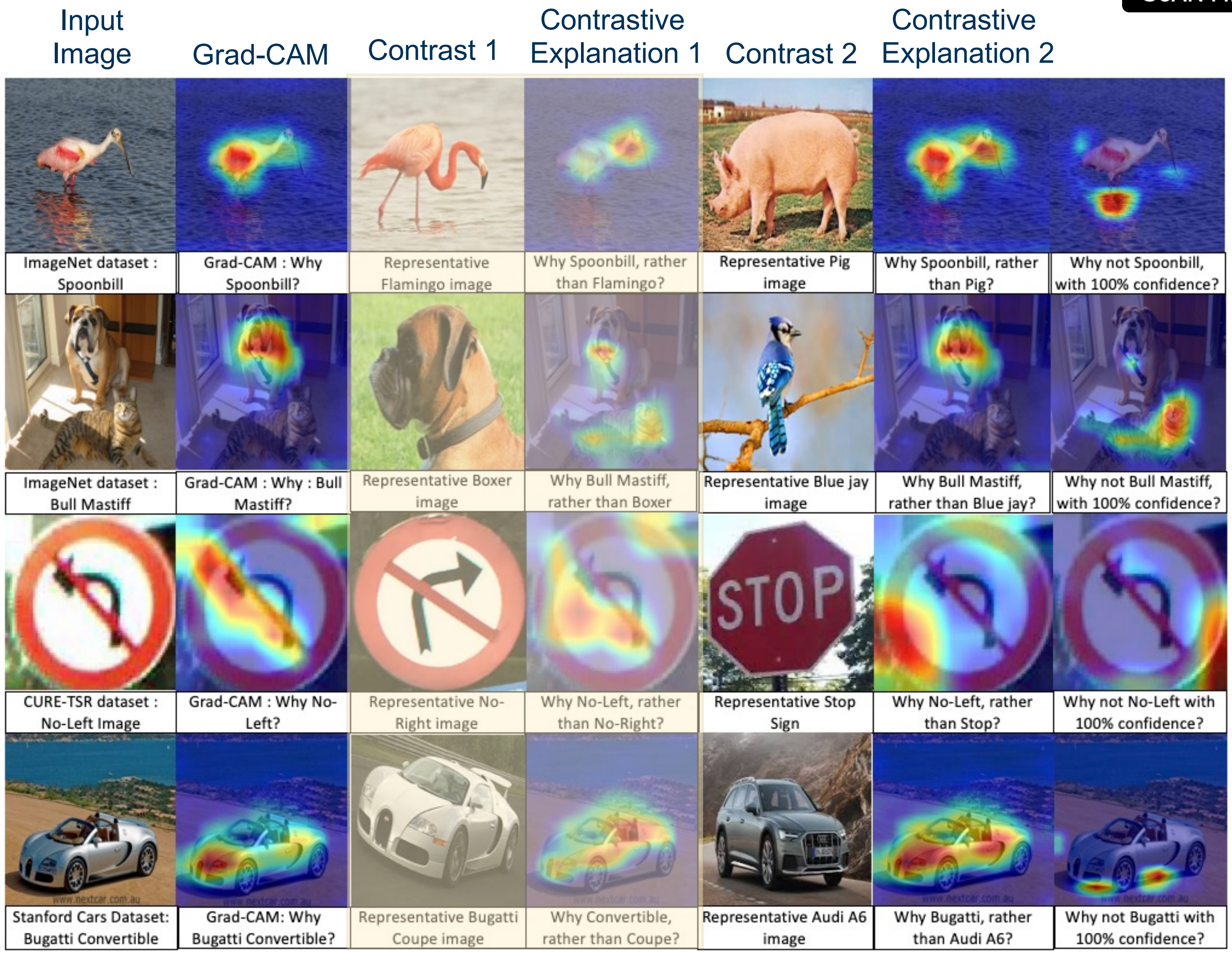
Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

Gradient and Activation-based Explanations

Results of GardCAM, CounterfactualCAM, and ContrastCAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME



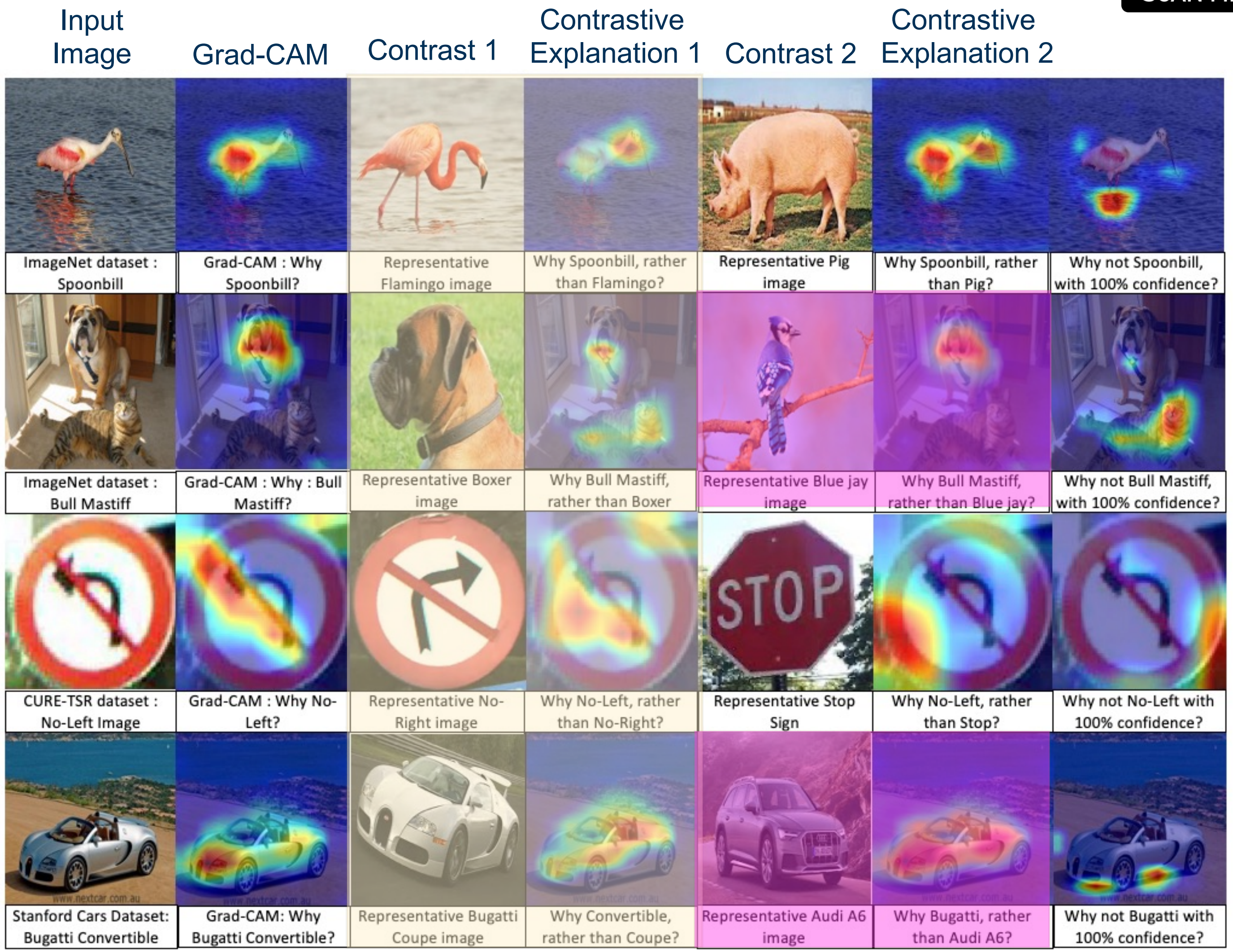
Human Interpretable

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 of GardCAM, CounterfactualCAM, and ContrastCAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations



Human Interpretable

Same as Grad-CAM

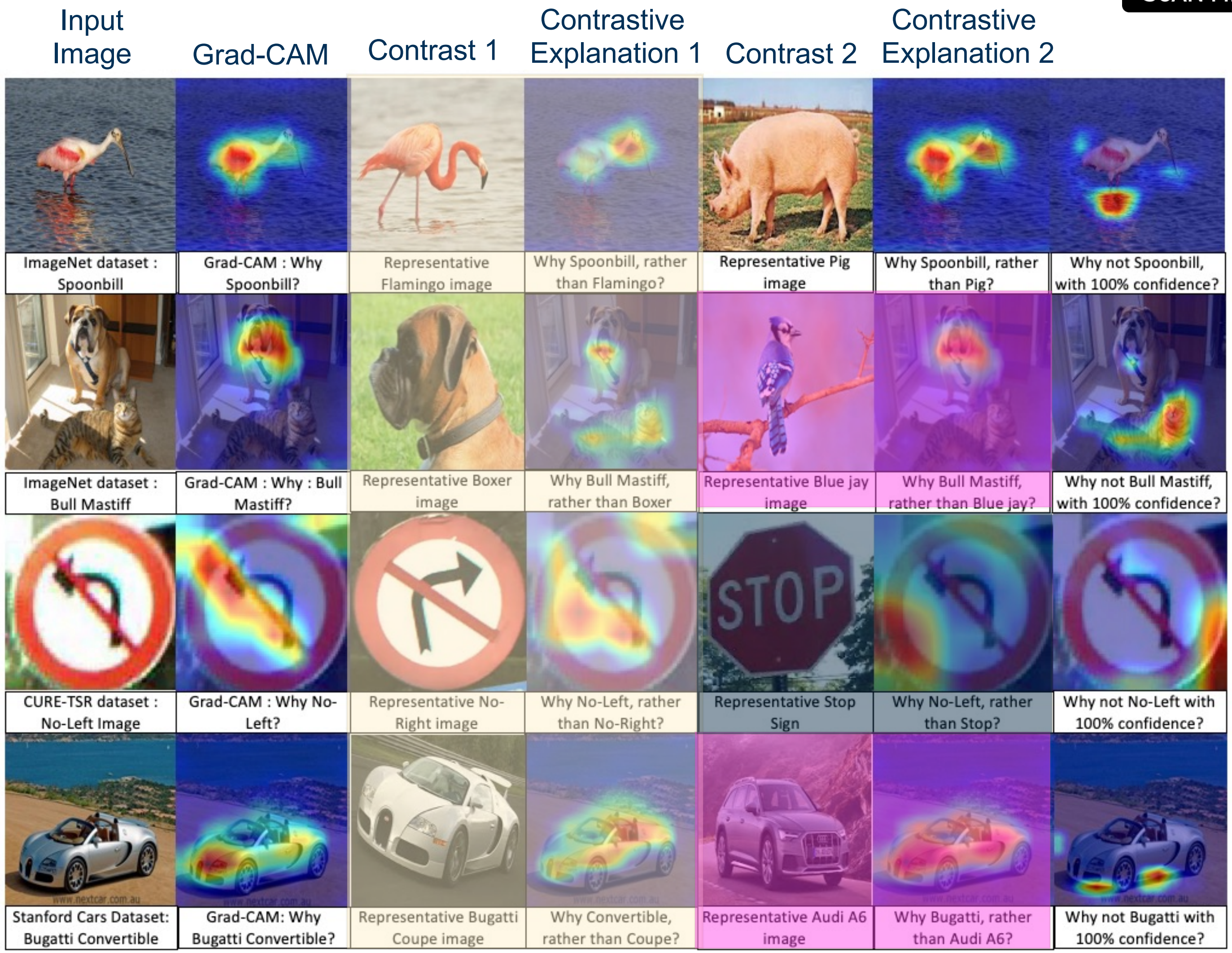
Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

Gradient and Activation-based Explanations

Results of GardCAM, CounterfactualCAM, and ContrastCAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME



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.

Gradient and Activation-based Explanations

Results of GardCAM, CounterfactualCAM, and ContrastCAM

SCAN ME

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

Human Interpretable

Same as Grad-

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 of GardCAM, CounterfactualCAM, and ContrastCAM

Only traffic sign with a straight bottom-left edge – enough to say `Not STOP Sign'

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.

Case Study 1: Leveraging anomaly scores, uncertainty scores, and explanations for Robust Recognition

NEURAL INFORMATION PROCESSING SYSTEMS

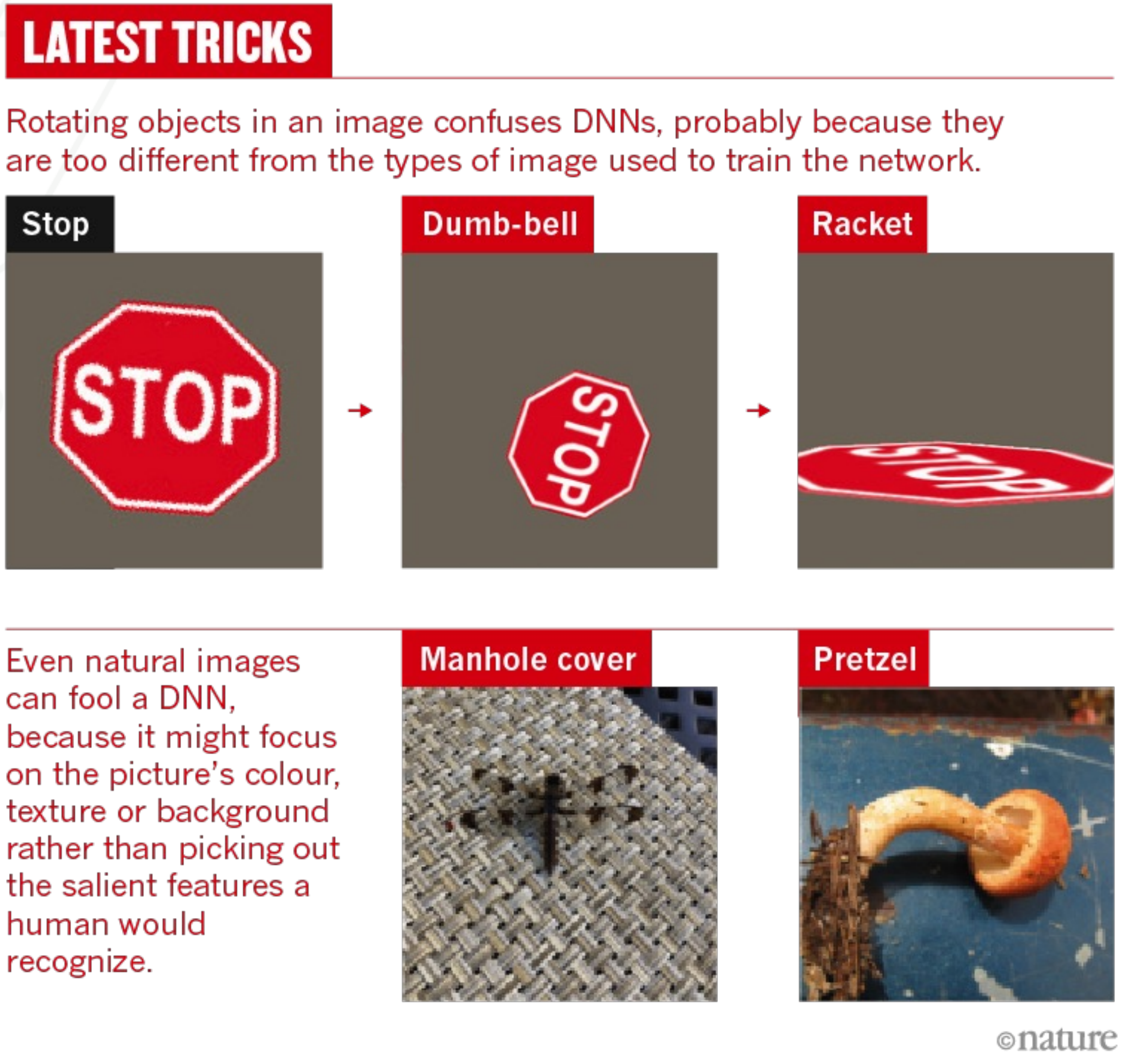
Introspective Learning: A Two-Stage Approach for Inference in Neural Networks

Mohit Prabhushankar, PhD
Postdoc

Ghassan AlRegib, PhD
Professor

SCAN ME

Robustness in Neural Networks
Why Robustness?



SCAN ME

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

Robustness in Neural Networks

Why Robustness?

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

SCAN ME

How would humans resolve this challenge?

We Introspect!

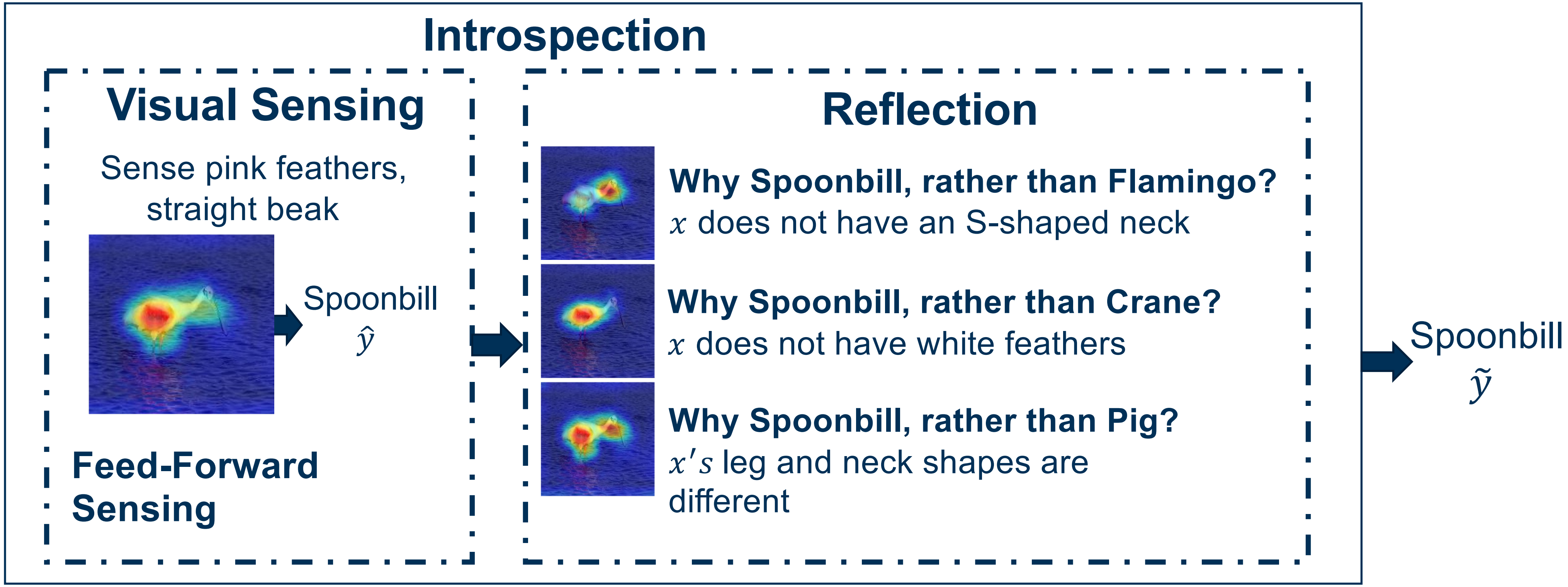
- Why am I being shown this slide?
- Why images of muffins rather than pastries?
- What if the dog was a bull mastiff?

Introspection
What is Introspection?

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.

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

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*..*

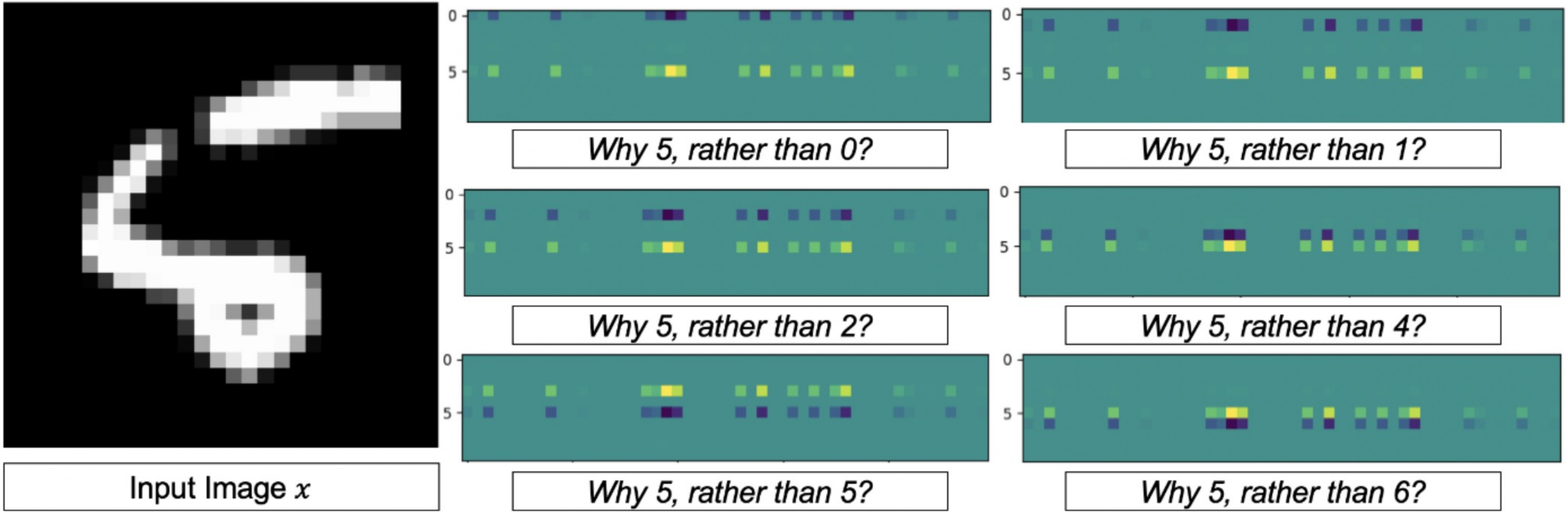
M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Introspection in Neural Networks
Gradients as Features

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

SCAN ME

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



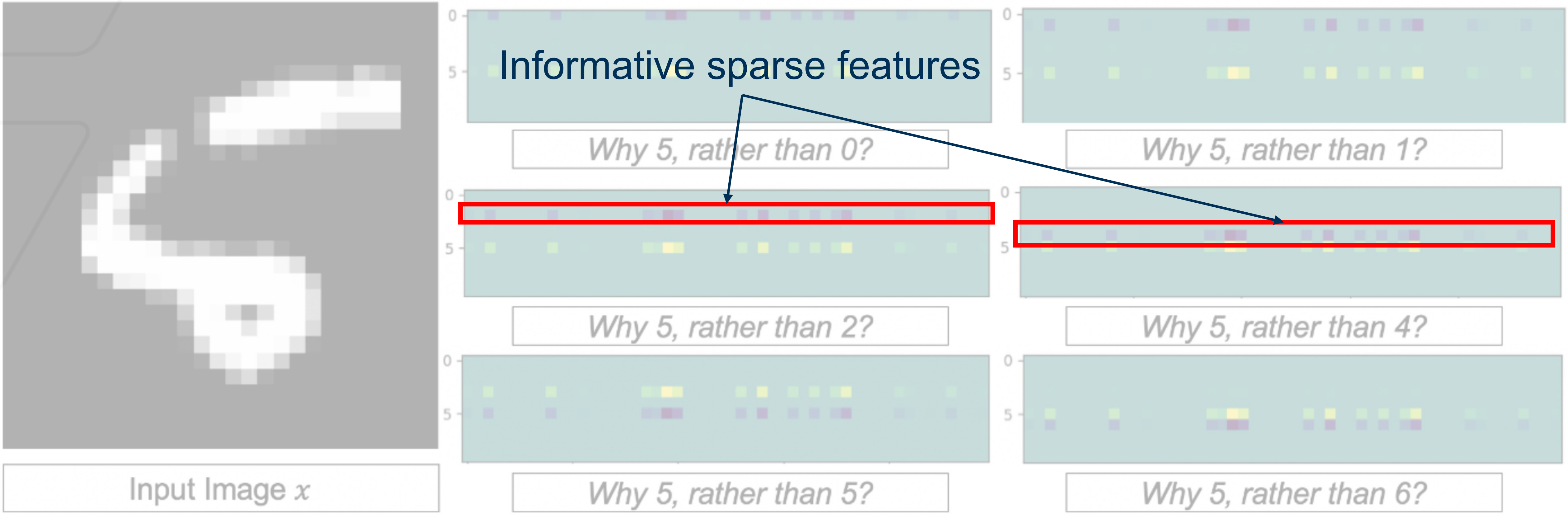
M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Introspection in Neural Networks
Gradients as Features

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

SCAN ME

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



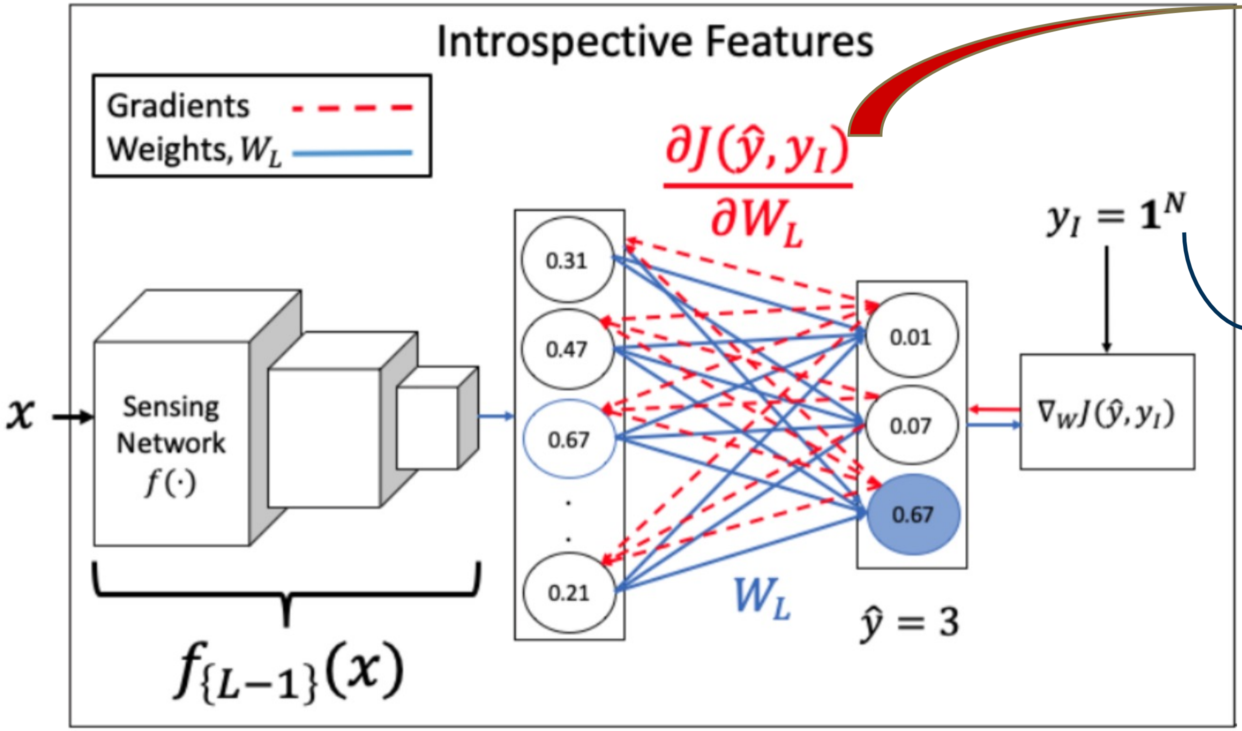
M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Introspection in Neural Networks
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



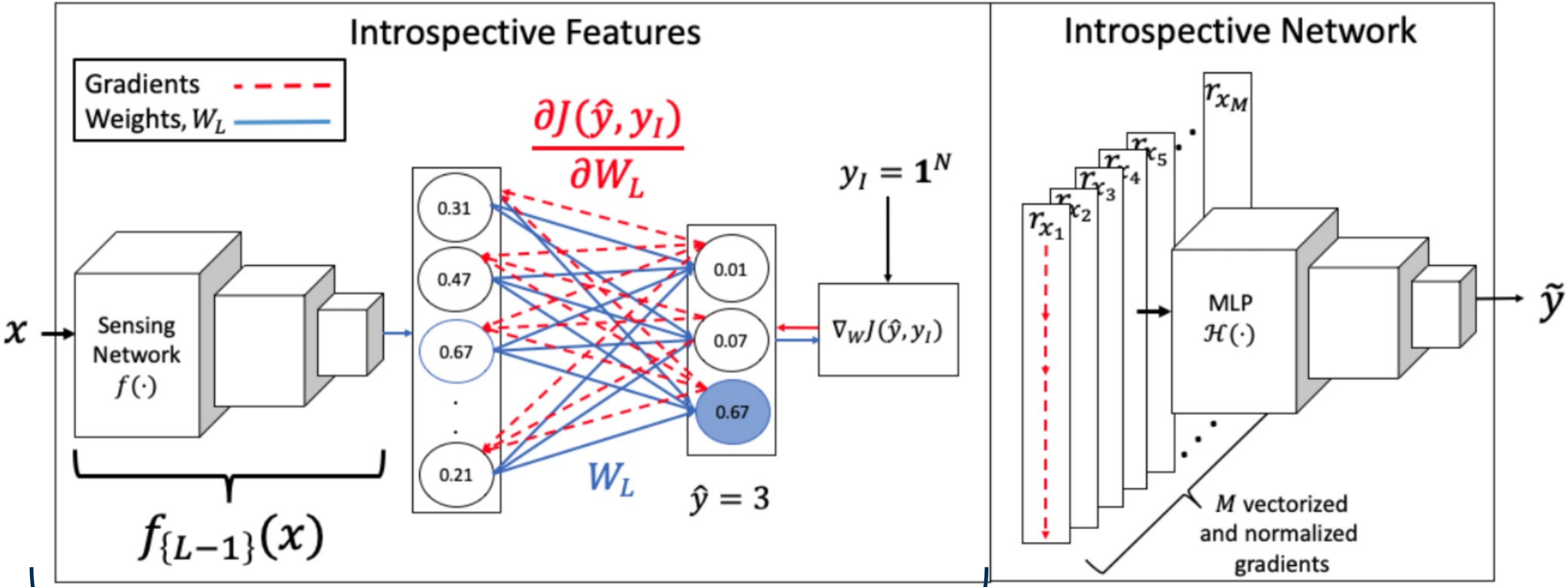
Normalized and vectorized gradients are introspective features

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 in Neural Networks
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.

Introspection in Neural Networks

When is Introspection Useful?

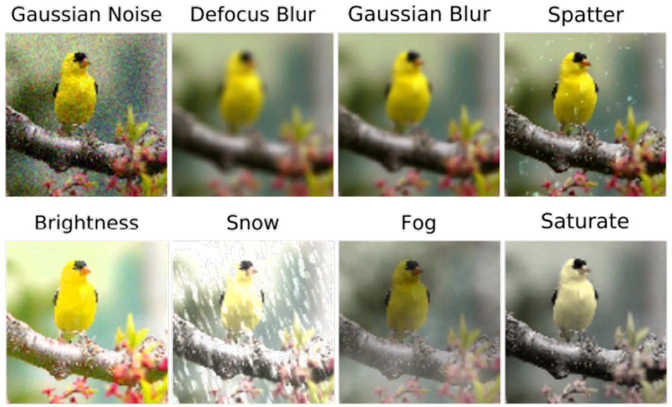
Introspective Learning: A Two-stage Approach for Inference in Neural Networks

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.

Introspection in Neural Networks
Generalization and Calibration

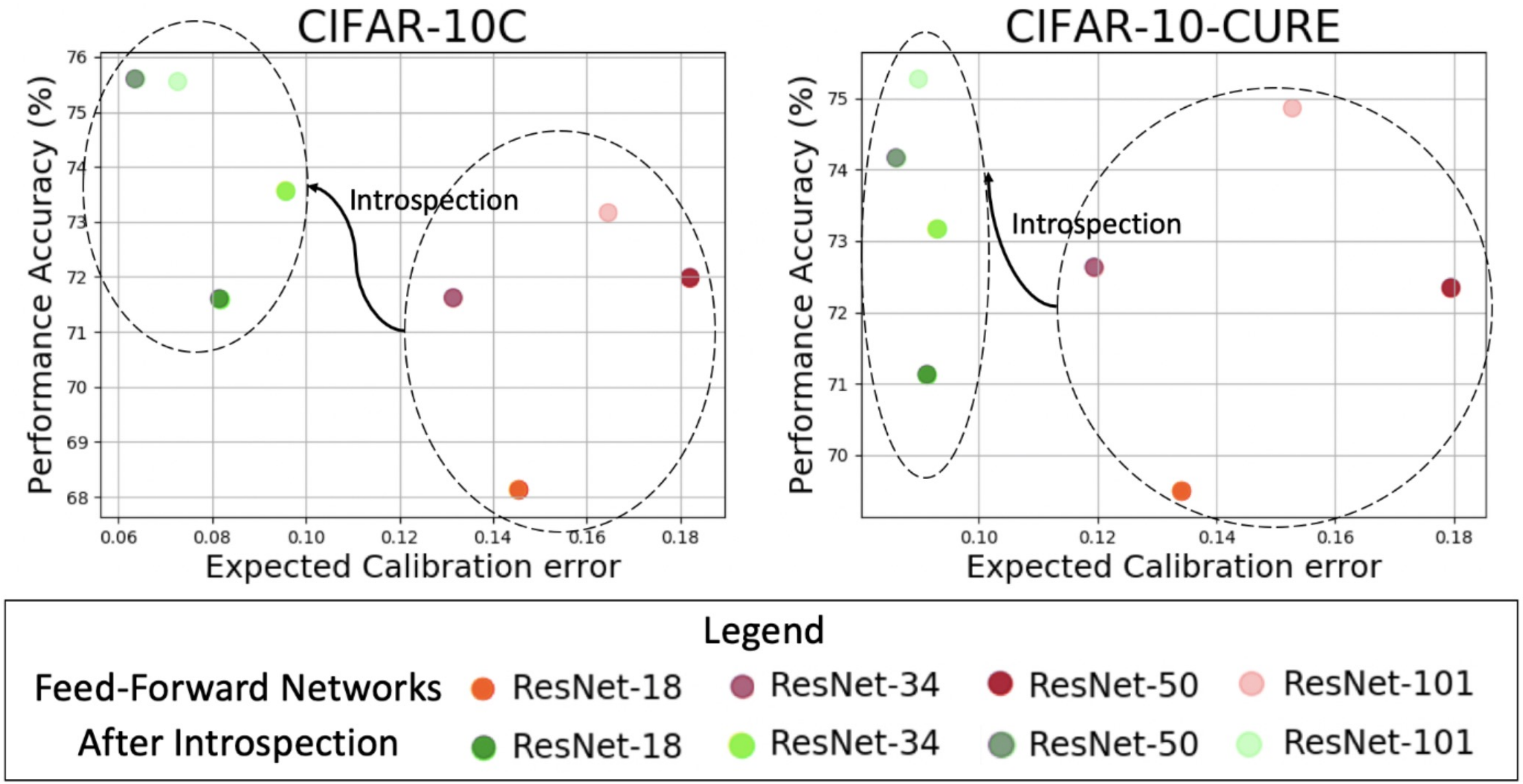
SCAN ME

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

Ideal: Top-left corner

Y-Axis: Generalization

X-Axis: Calibration



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

Introspection in Neural Networks

Plug-in Nature of Introspection

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

SCAN ME

Introspection is a light-weight option to resolve robustness issues

Table 1: Introspecting on top of existing robustness techniques.

Methods		Accuracy
ResNet-18	Feed-Forward	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.

Case Study 2: Leveraging anomaly scores, uncertainty scores, and explanations for Anomalous object classification

Detecting and Classifying Anomalies in Artificial Intelligence Systems

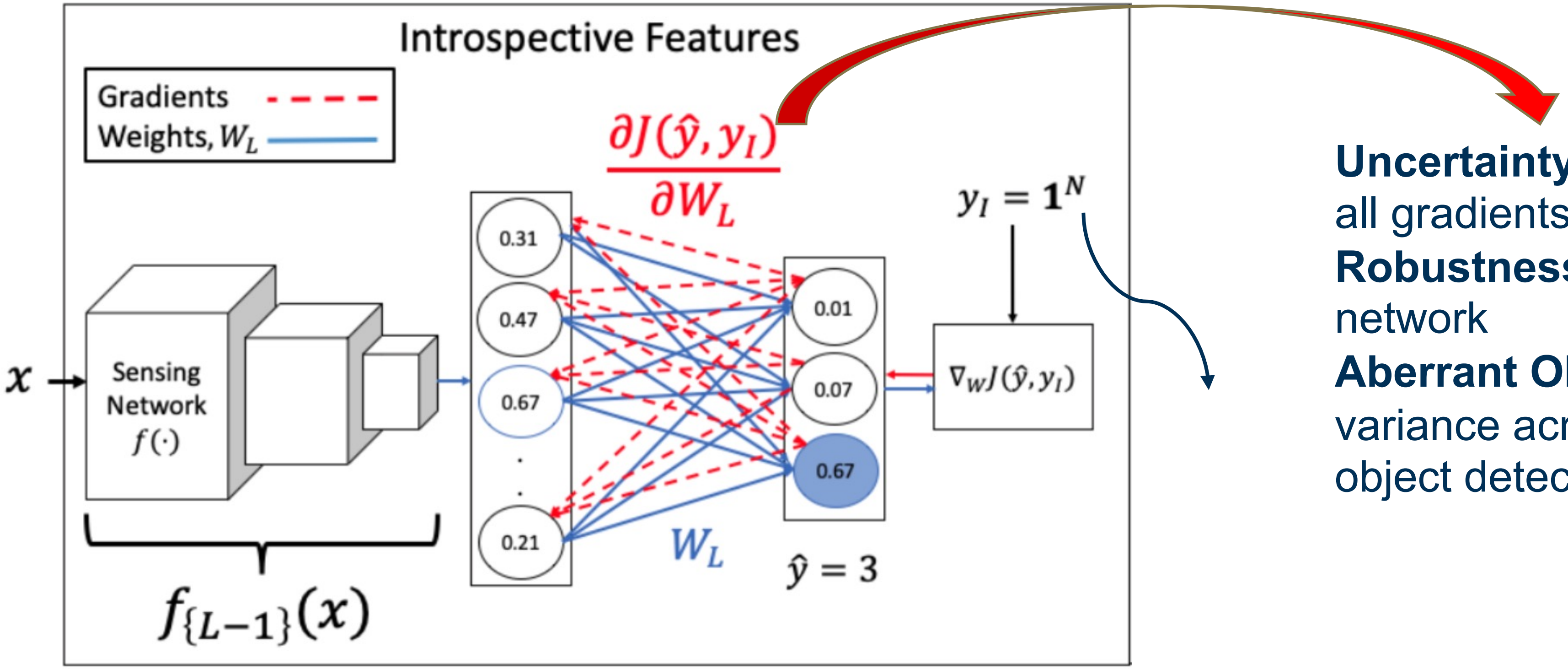
Gukyeong Kwon, PhD
Amazon AWS

Mohit Prabhushankar, PhD
Postdoc, Georgia Tech

Ghassan AlRegib, PhD
Professor, Georgia Tech

Aberrant Object Detection
Deriving Gradient Features

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



Uncertainty: We took energy of all gradients
Robustness: We trained a new network
Aberrant Objects: We take variance across gradients from object detector

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.

Aberrant Object Detection
Aberrance Detection

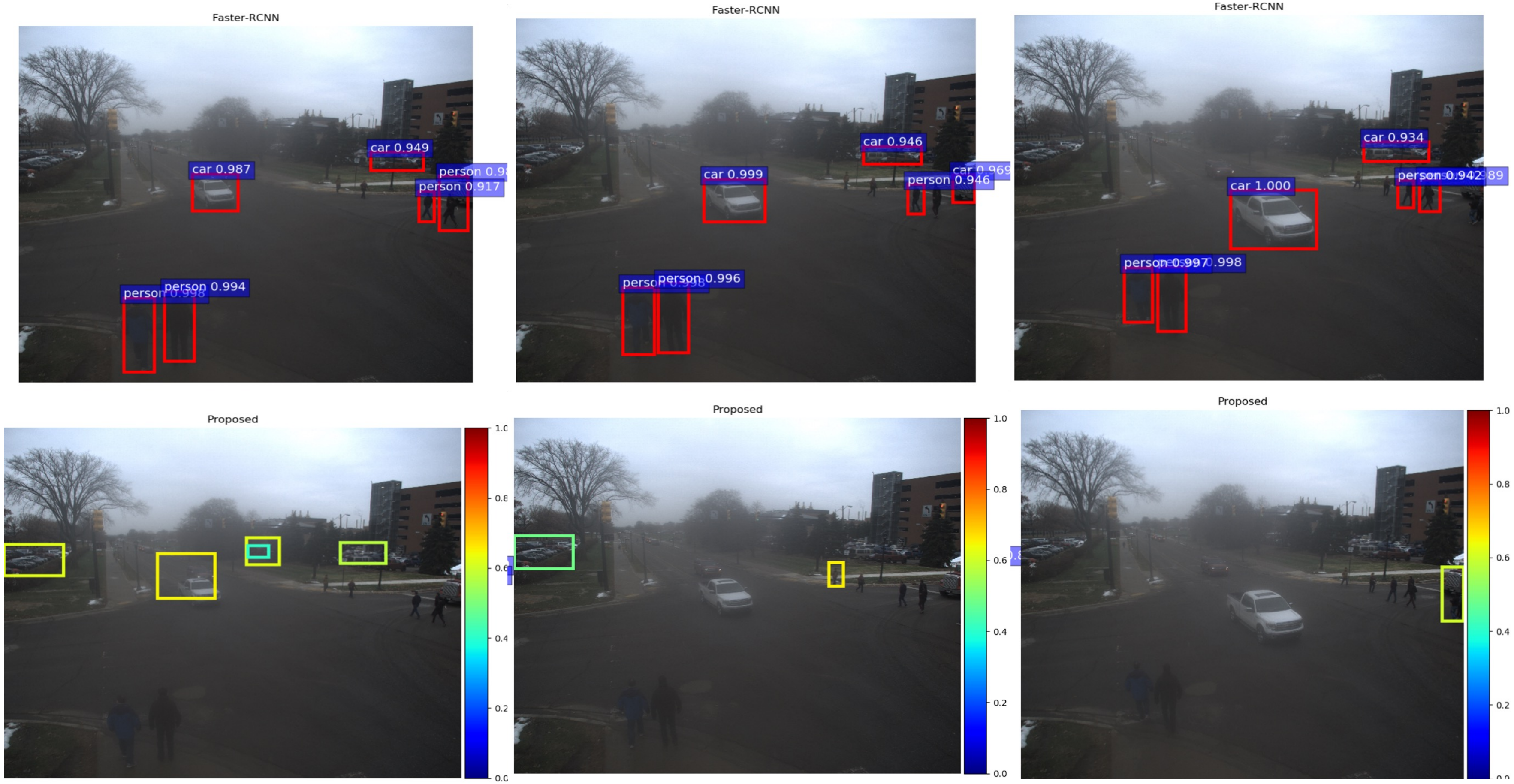
Uncertainty using variance of introspective gradients rather than energy of gradients

- Object detection algorithms would pick up on all the trained objects
- The gradient-based uncertainty approach picks up only the *aberrant* object – objects that bear a resemblance to novel classes

AlRegib, Ghassan, et al. "Detecting and Classifying Anomalies in Artificial Intelligence Systems." U.S. Patent Application No. 17/633,878.

Aberrant Object Detection
Complementary to object detectors

Uncertainty using variance of introspective gradients rather than energy of gradients



AlRegib, Ghassan, et al. "Detecting and Classifying Anomalies in Artificial Intelligence Systems." U.S. Patent Application No. 17/633,878.

Aberrant Object Detection
Active Learning

Use the uncertain boxes for obtaining labels from annotators

Use new annotations for subsequent training in an active learning setting

Parchami, Armin, et al. "Variance of gradient based active learning framework for training perception algorithms." U.S. Patent Application No. 17/172,854.

Objectives
Takeaways from Part III

- Part I: Challenges in Perception and Autonomy
- Part II: Deep Learning for Perception
- **Part III: Existing Deep Learning solutions to Challenges in Perception**
  - It is not always clear if aberrant events and challenges must be incorporated in training
  - Instead, they can and should be equipped with diagnostic tools at predictions
  - These diagnostic tools are anomaly and uncertainty scores for decision making and contextual explainability for post-hoc stakeholders
  - Gradients provide the change induced by an aberrant event in the network and can be used to obtain the required prediction diagnosis
- Part IV: Key Takeaways and Future Directions