Robust Neural Networks at Inference: Towards 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

Aug 9, 2024 – San Jose, CA, USA

Tutorial Materials

Accessible Online

SCAN ME

https://alregib.ece.gatech.edu/mipr-2024-tutorial/

{alregib, mohit.p}@gatech.edu

MIPR 2024 Tutorial

The 7th IEEE International Conference on Multimedia Information Processing and Retrieval

IEEE MIPR 2024

Robust Neural Networks: Towards Explainability, Uncertainty, and Intervenability

Presenters:

Ghassan AlRegib and Mohit Prabhushankar

Georgia Institute of Technology, Georgia Institute of Technology

www.ghassanalregib.info

alregib@gatech.edu, mohit.p@gatech.edu

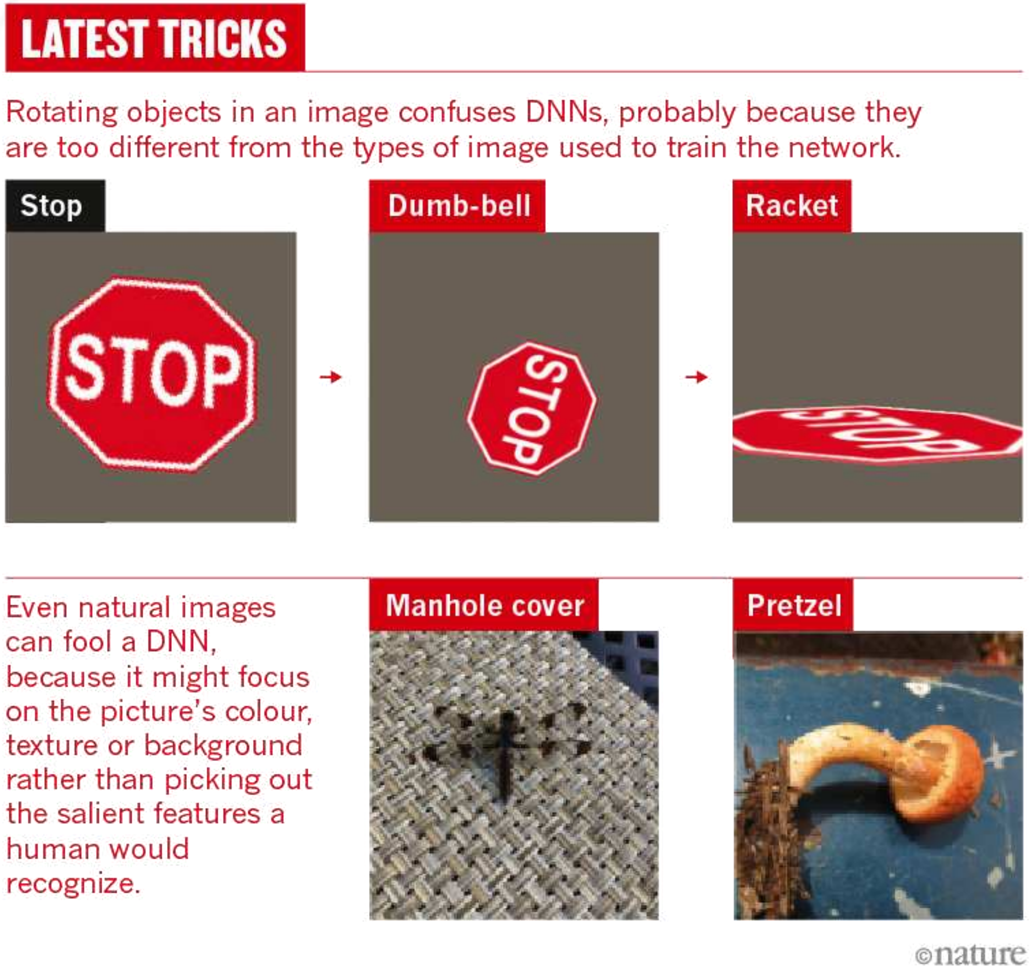
OLIVES @GeorgiaTech

Georgia Tech

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:

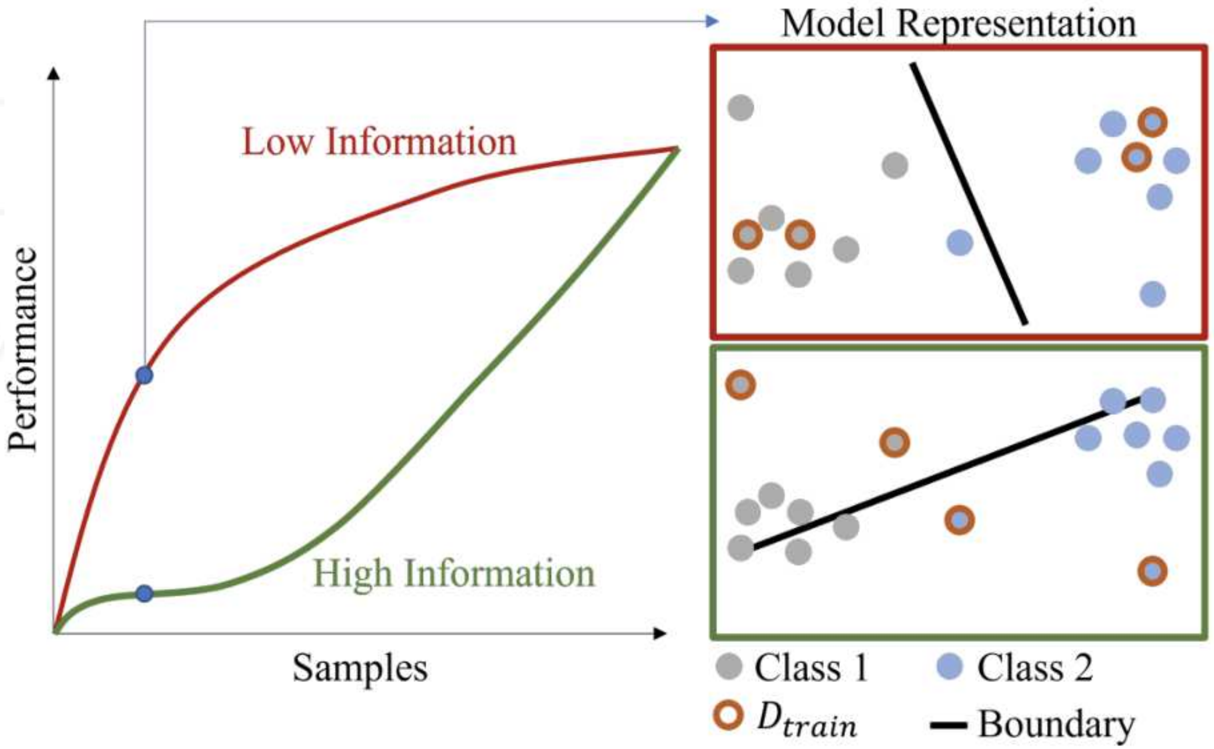
- Test distributions
- Anomalous data
- Out-Of-Distribution data
- Adversarial data
- Corrupted data
- Noisy data
- New classes
- …

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 <u>not</u> be used in early training

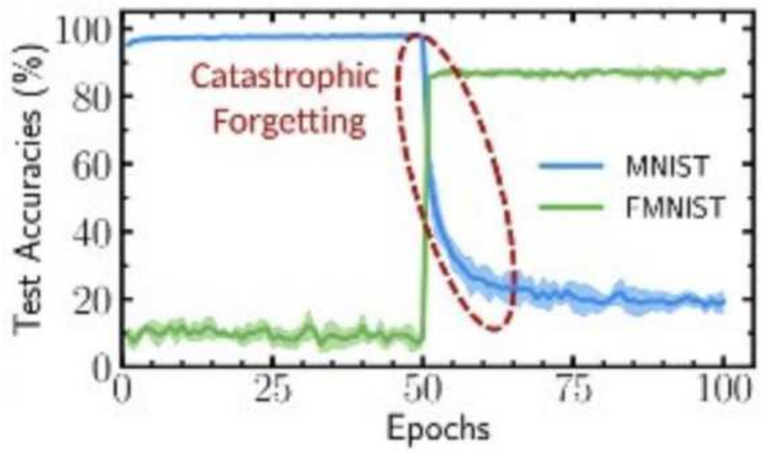


Novel samples = Most Informative

- 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

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

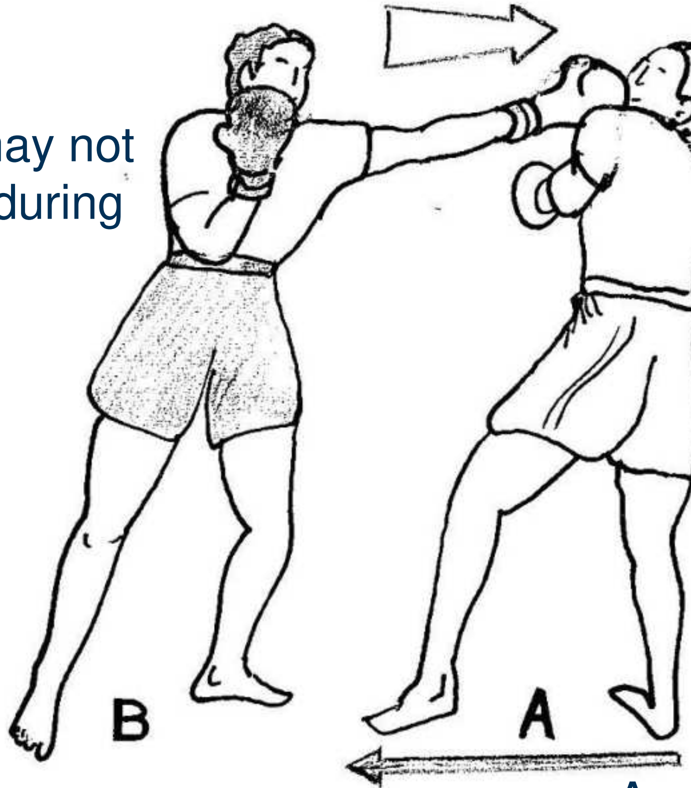


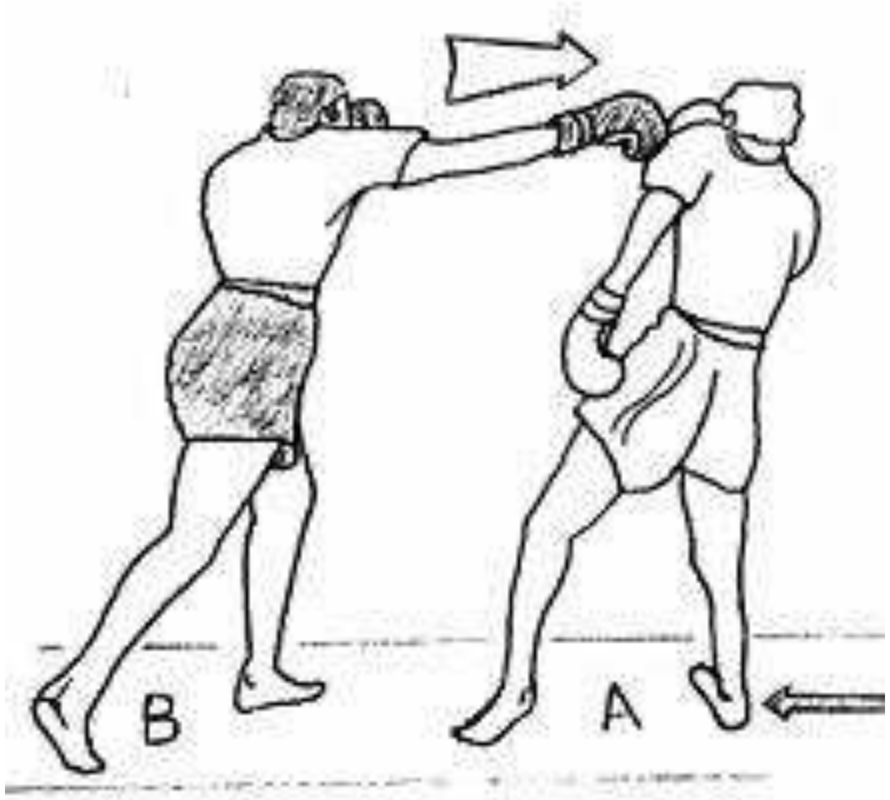
- The model performs well on the new scenarios, **while forgetting the old scenarios**
- Several techniques exist to overcome this trend
- However, they affect the overall performance in large-scale settings
- It is not always clear **if and when** to incorporate novel scenarios in training

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

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

Deep Learning at Inference

Overcoming Challenges at Inference

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

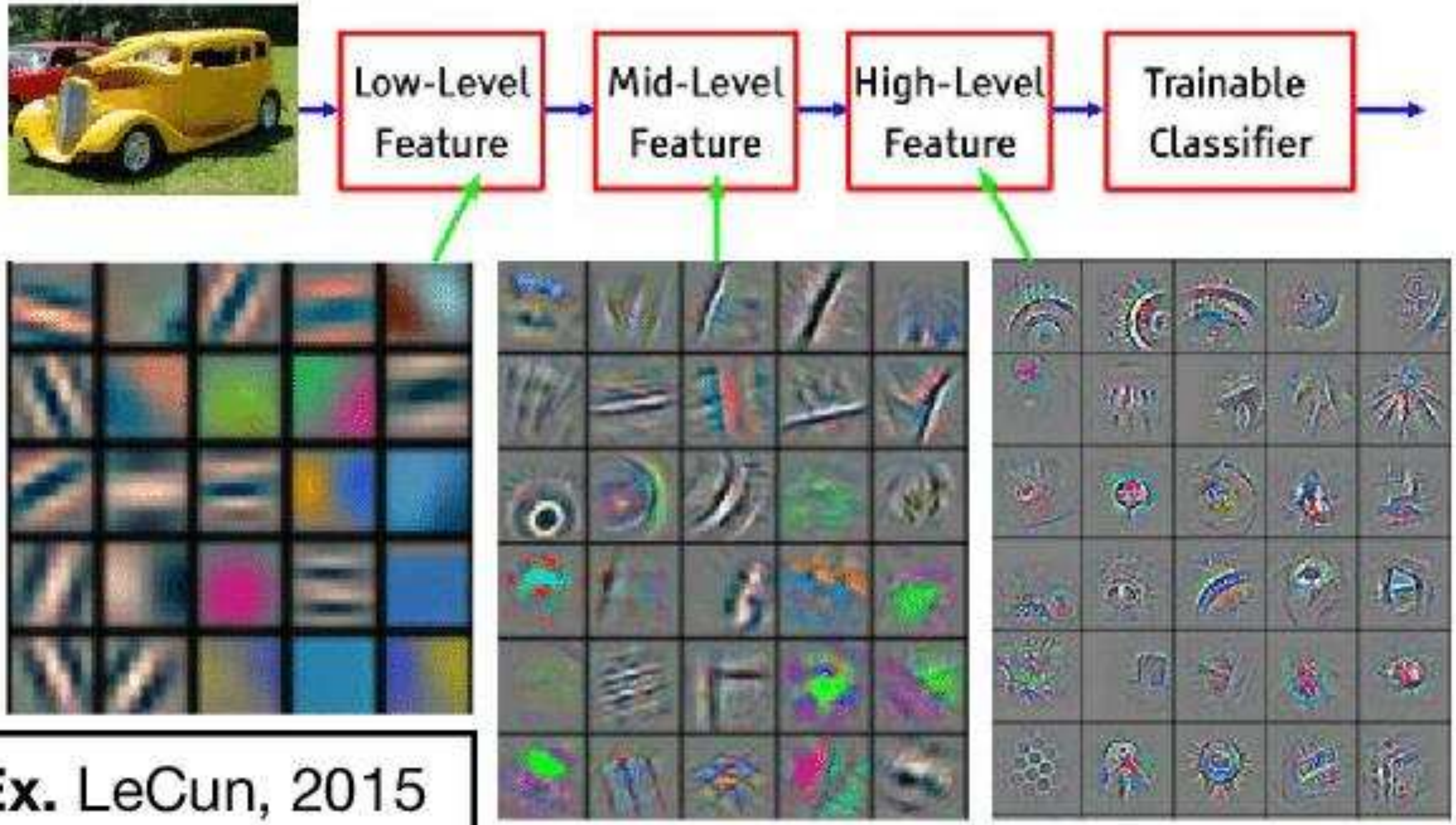
Objective

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

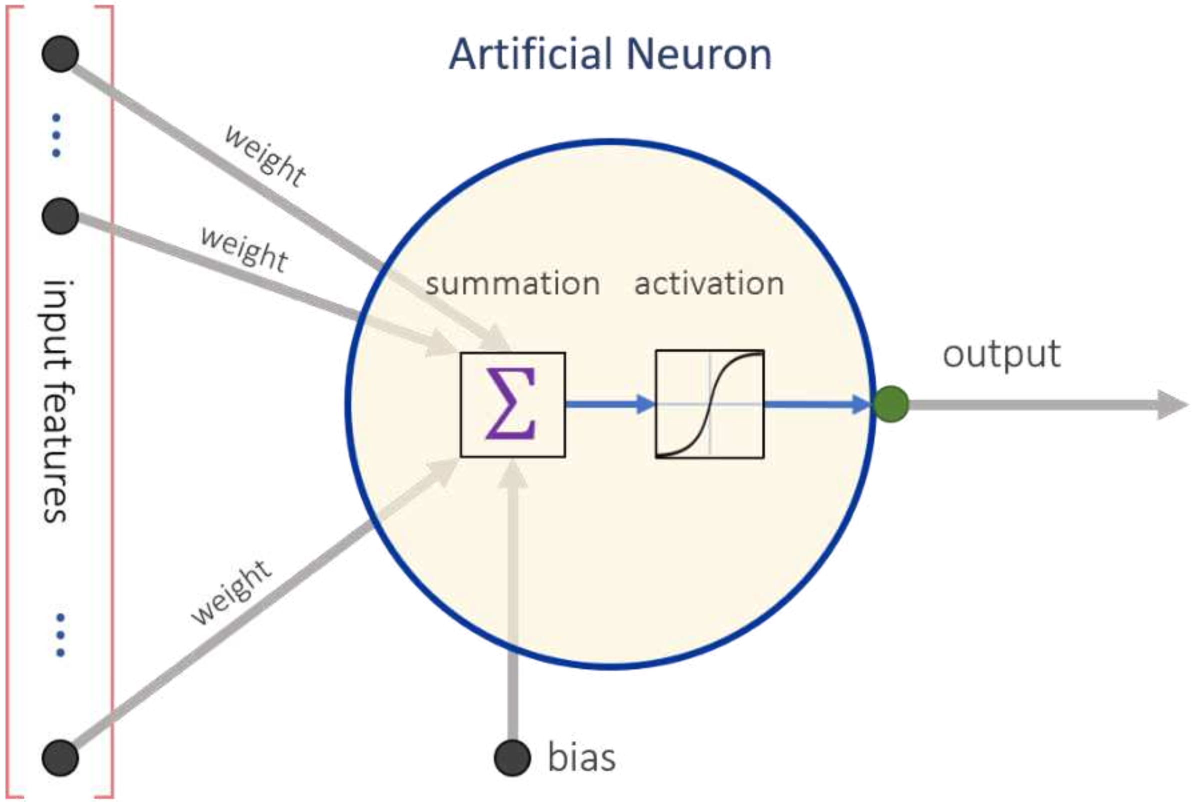


Ex. LeCun, 2015

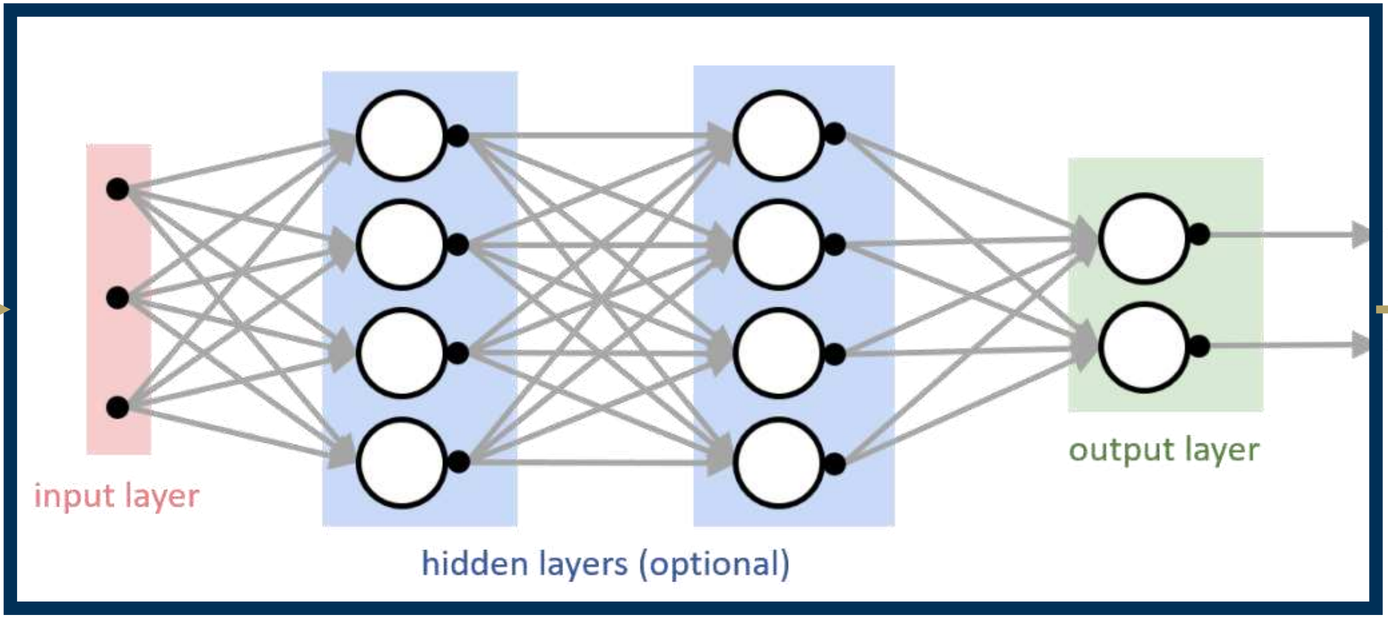
The underlying computation unit is the Neuron

Artificial neurons consist of:

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



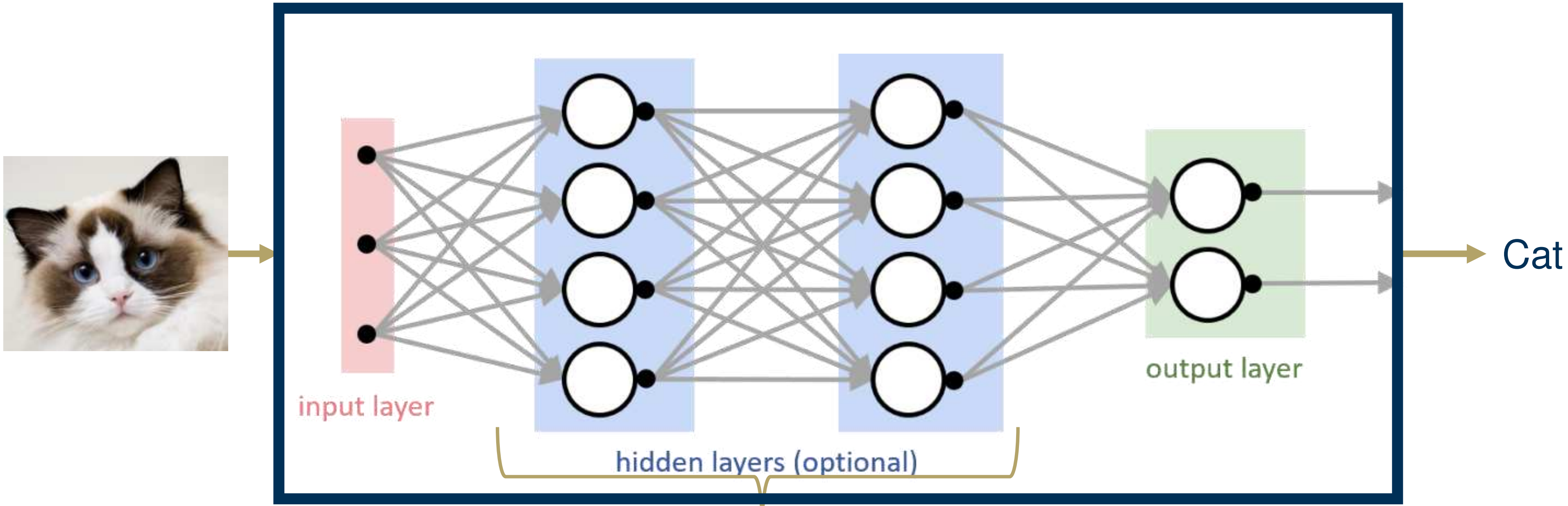
Neurons are stacked and densely connected to construct ANNs

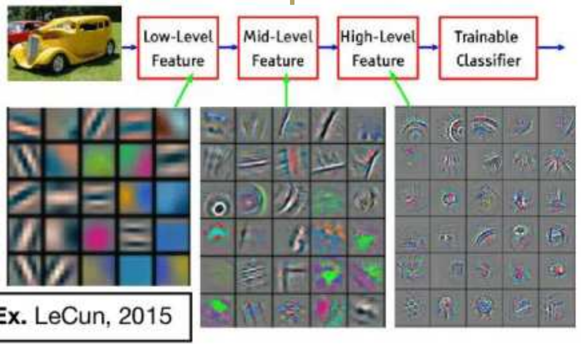


Cat

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 $1 \dots K-1$)

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





Ex. LeCun, 2015

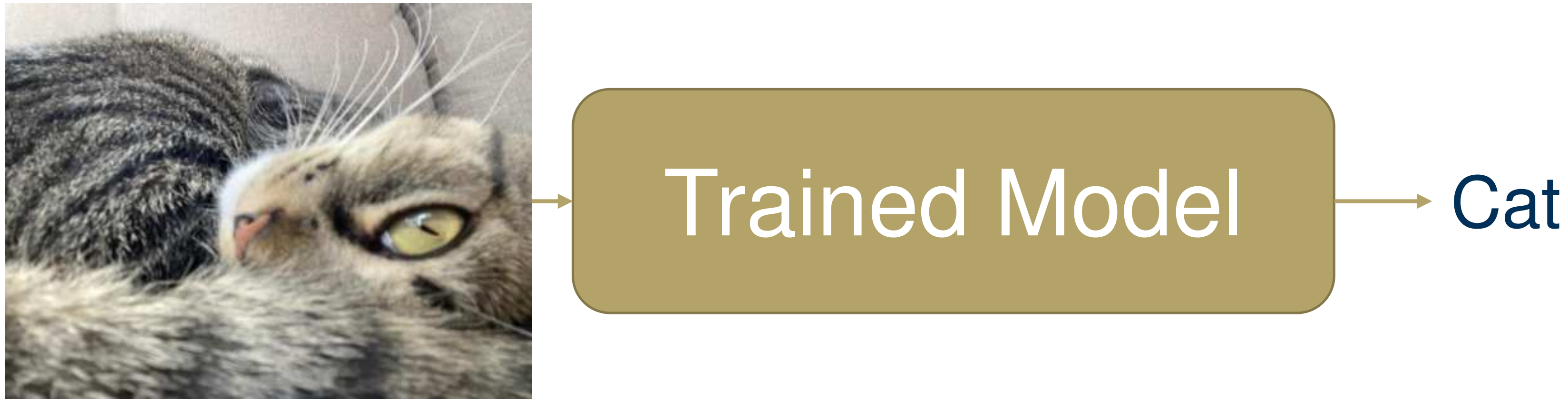
Deep Learning at Inference

What, Where, and When 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
- …

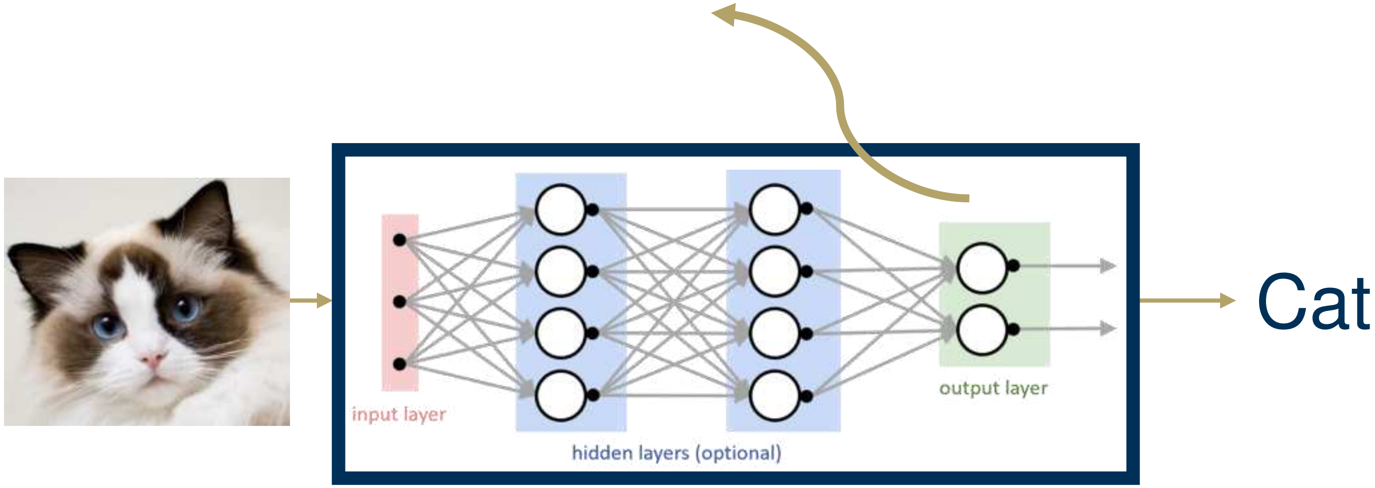


Neural networks are feed-forward systems; output layer logits are used for inference

Novel data sources:

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

All **required** information is passed to last layer
Outputs from last layer are termed **Logits**



Required information is learned at training; leads to **inductive bias** when encountering novel data at inference

Deep Learning at Inference

What, Where, and When is Inference?

Inference occurs at: (i) Testing, and (ii) Deployment

Novel data sources:

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

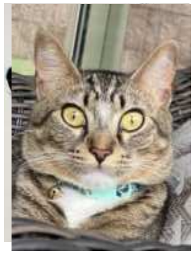
→ Trained Model at Testing → Cat, Cat, Cat

→ Trained Model at Deployment → Cat

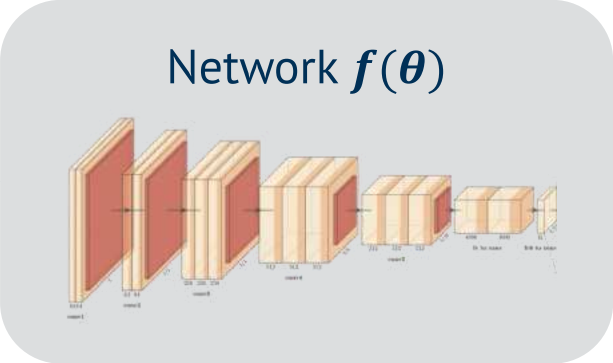
Deep Learning at Inference

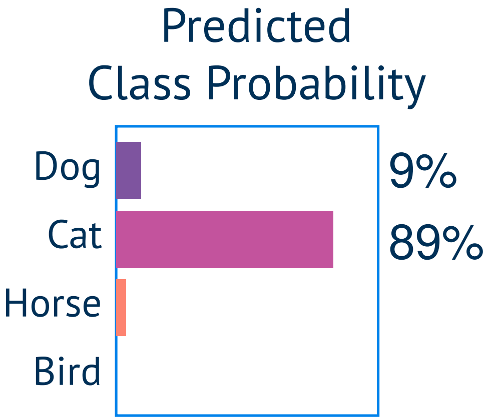
Application: Classification

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



x





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

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

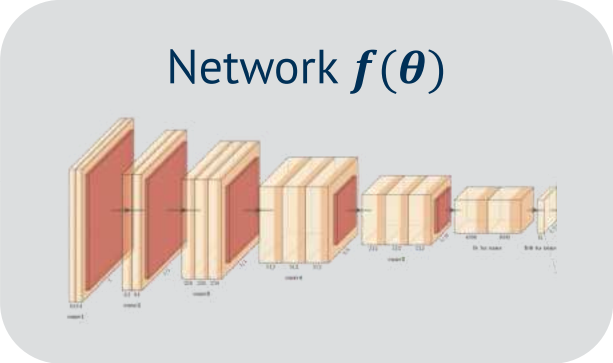
$\hat{y}$ = Logits
y = Predicted Class
$p(\hat{y})$ = Probabilities
$f(\cdot)$ = Trained Network
χ = Training data

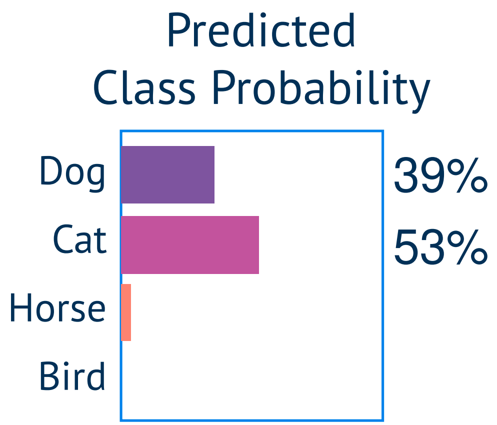
Deep Learning at Inference

Application: Robust Classification

Deep learning robustness: Correctly predict class even when data is <u>novel</u>

x'





If $x' \notin \chi$, the data is **novel**

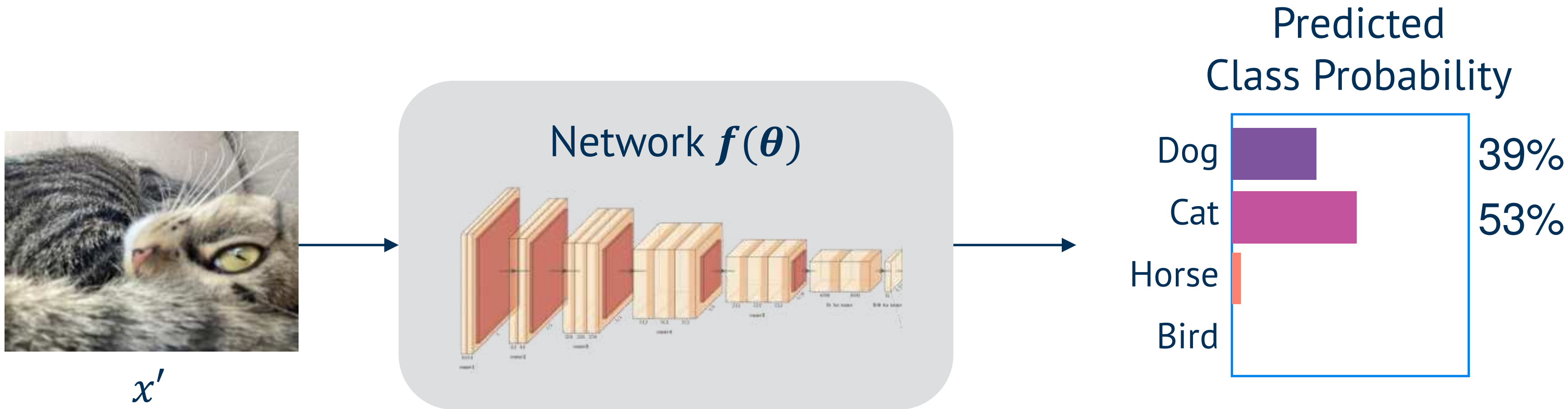
$$\hat{y} = f(x' + \epsilon)$$
$$y = argmax_i\, \hat{y}$$
$$p(\hat{y}) = T(f(x' + \epsilon))$$

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

Deep Learning at Inference

Application: Robust Classification

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



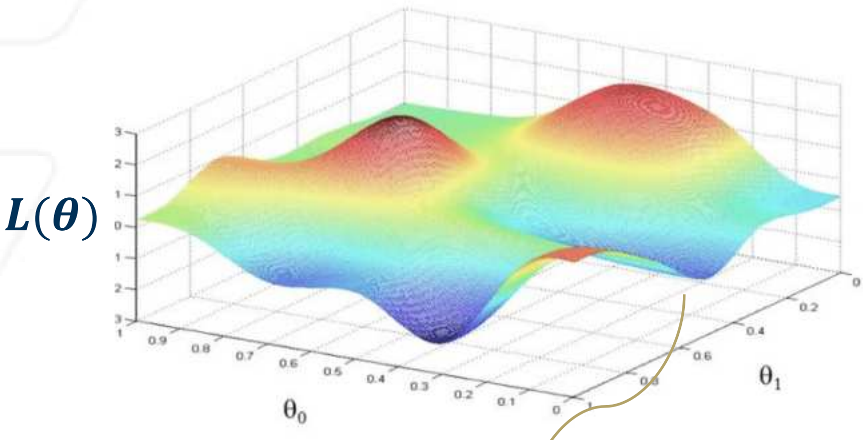
To achieve robustness at Inference, we need the following:

- **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

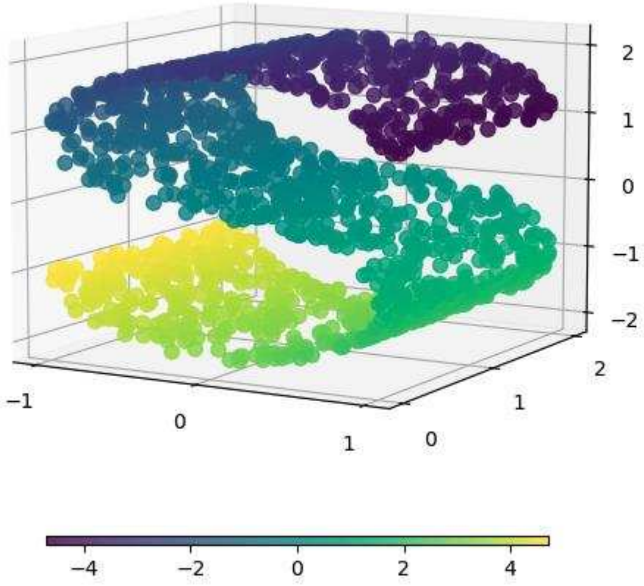
Why is this Challenging?

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)



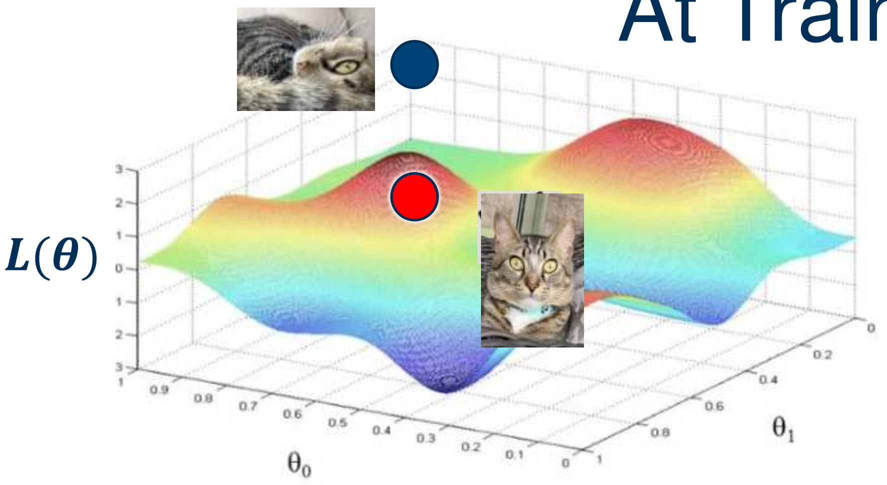
Real data visualizations generated using dimensionality reduction algorithms (Isomap)

Challenges at Inference
Inference

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

At Inference

At Training



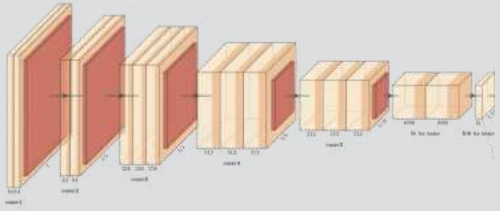
At training, we have access to all training data.

Information at Inference

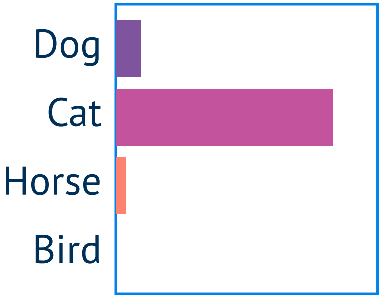
Fisher Information

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

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



Predicted
Class Probability



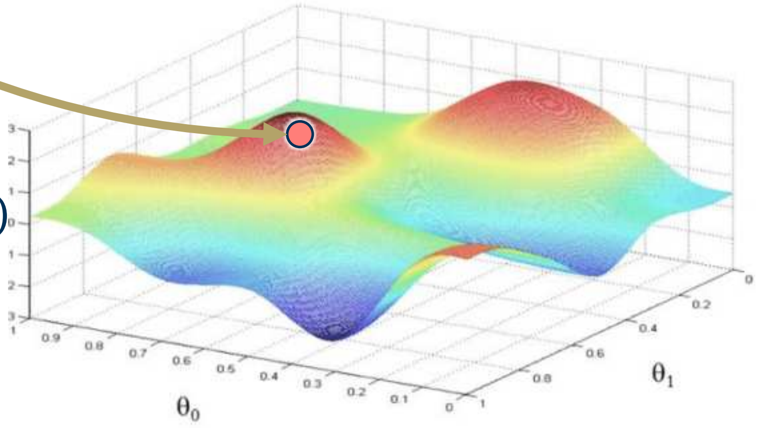
Fisher Information

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

θ = Statistic of distribution

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

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

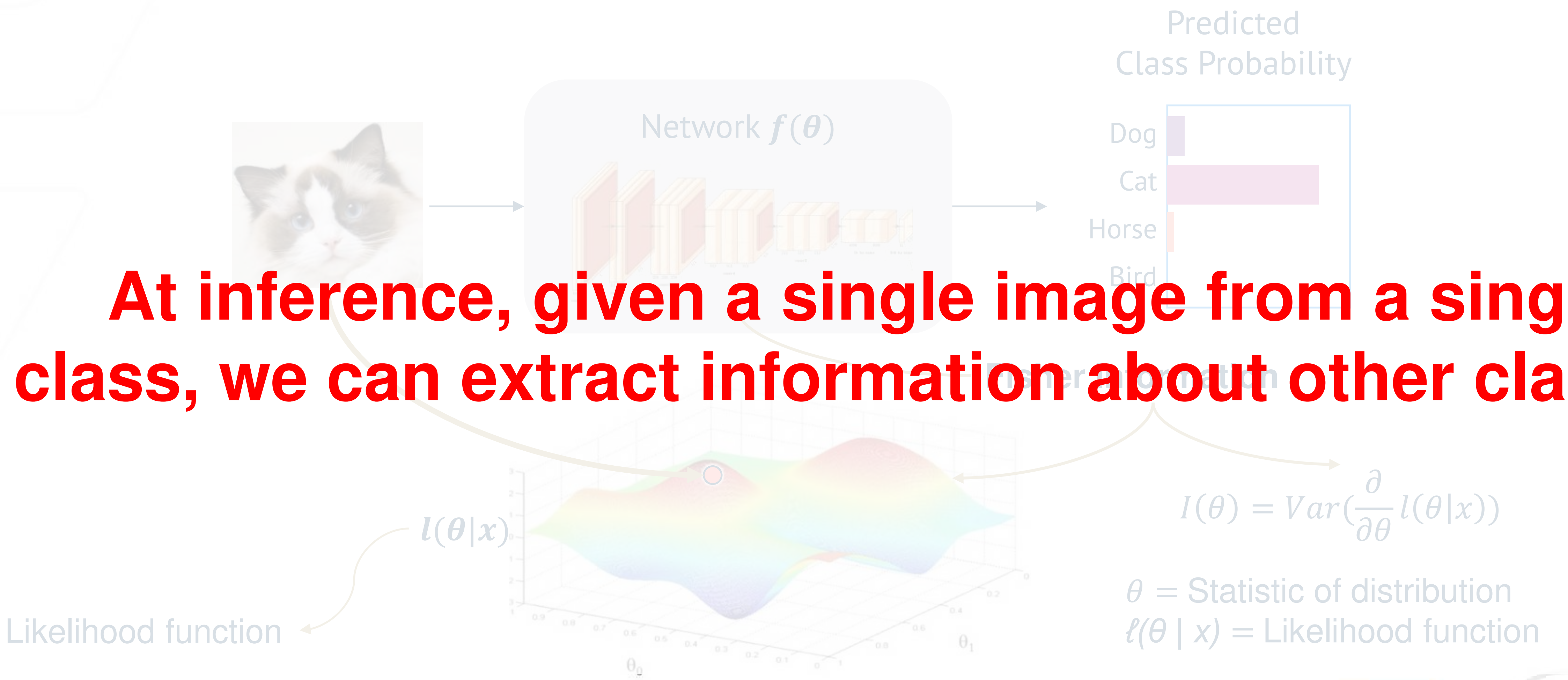


Likelihood function

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

Information at Inference

Information at Inference



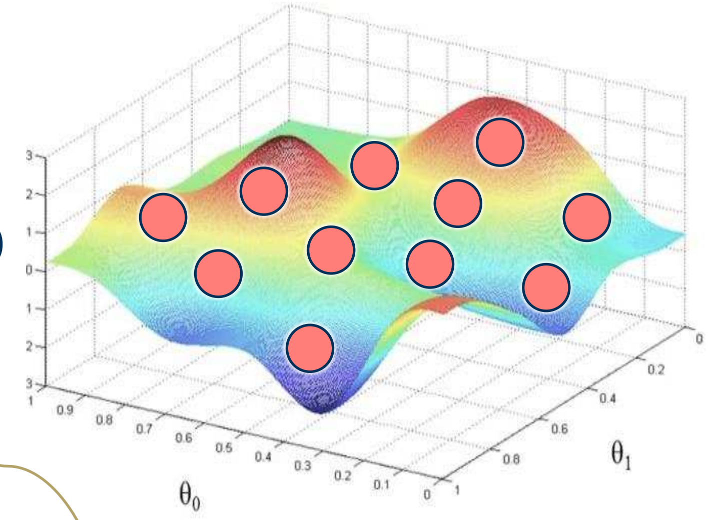
At inference, given a single image from a single class, we can extract information about other classes

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!

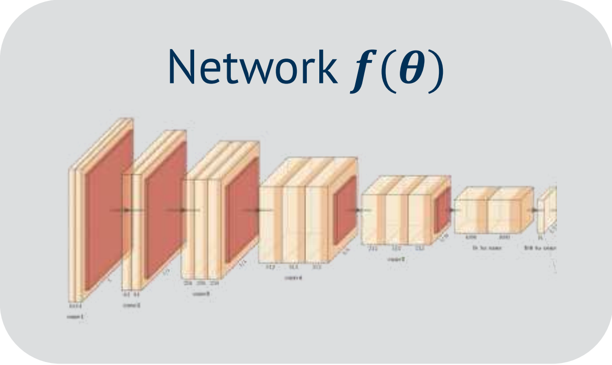
Kwon, Gukyeong, et al. "Backpropagated gradient representations for anomaly detection." *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer International Publishing, 2020.

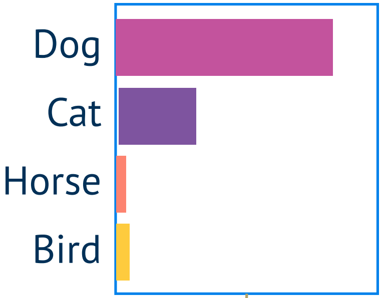
Information at Inference

Case Study: Gradients as Fisher Information in Explainability

Gradients infer information about the statistics of underlying manifolds

x

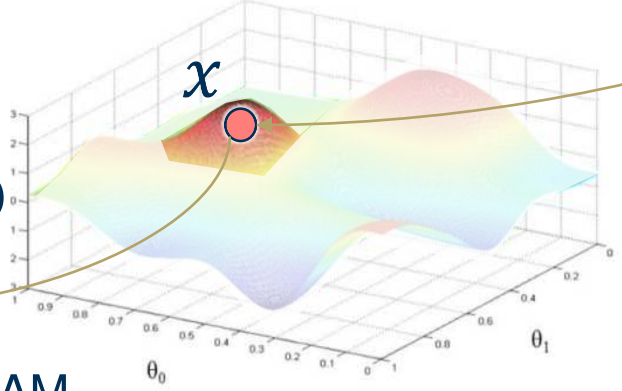




Local information (specific to x) is sufficient!

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

$l(\theta|x)$



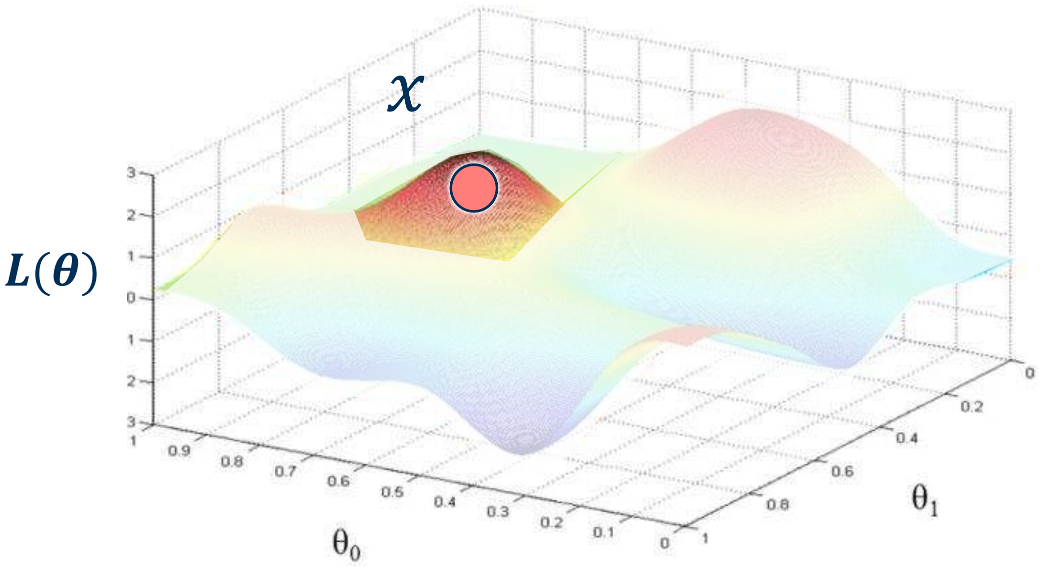
Feature attribution via GradCAM

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

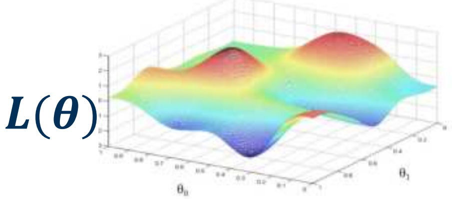
[1] A good blogpost about Fisher Information: https://towardsdatascience.com/an-intuitive-look-at-fisher-information-2720c40867d8

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



Ideal

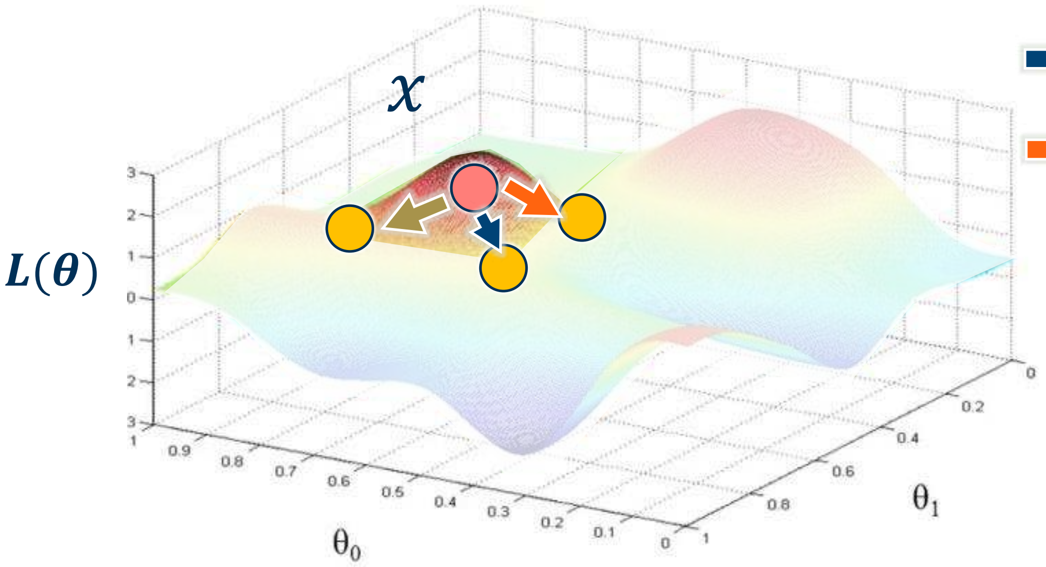


$\alpha\, \nabla_\theta\, L(\theta)$ provides local information up to a small distance α away from x

Gradients at Inference

Direction of Steepest Descent

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



Path 1?

Path 2?

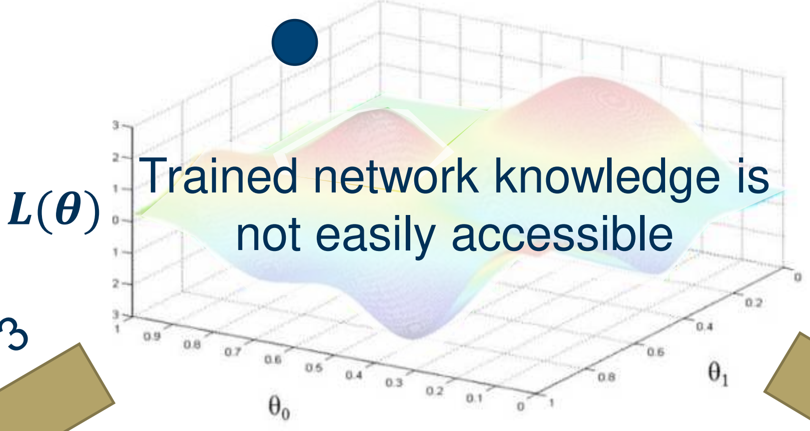
Path 3?

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

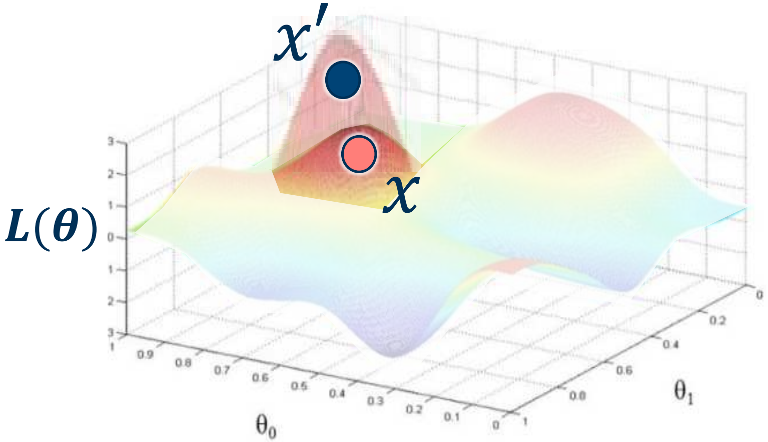
To Characterize the Novel Data at Inference



At Inference

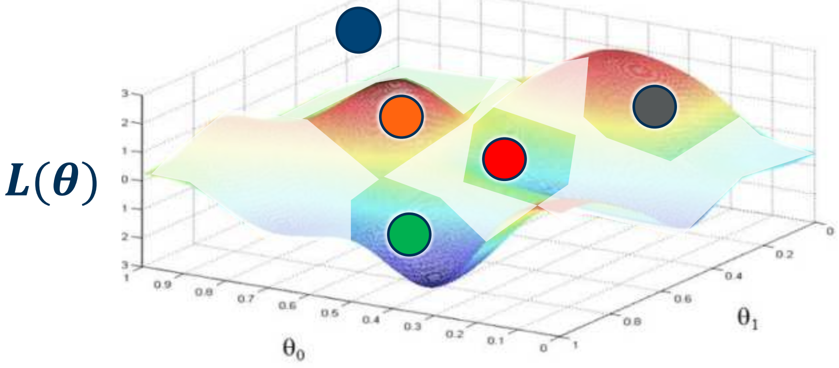
Part 2, 3

Counterfactual and Contrastive Representations using Gradients



Part 4

Representation Traversal using Interventions



Robust Neural Networks

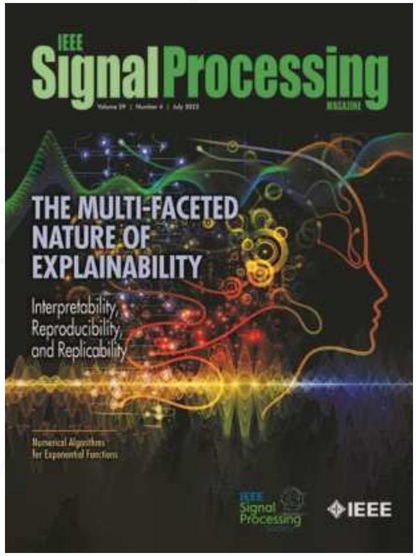
Part 2: Explainability at Inference

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
  - Case Study: Introspective Learning
- 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?

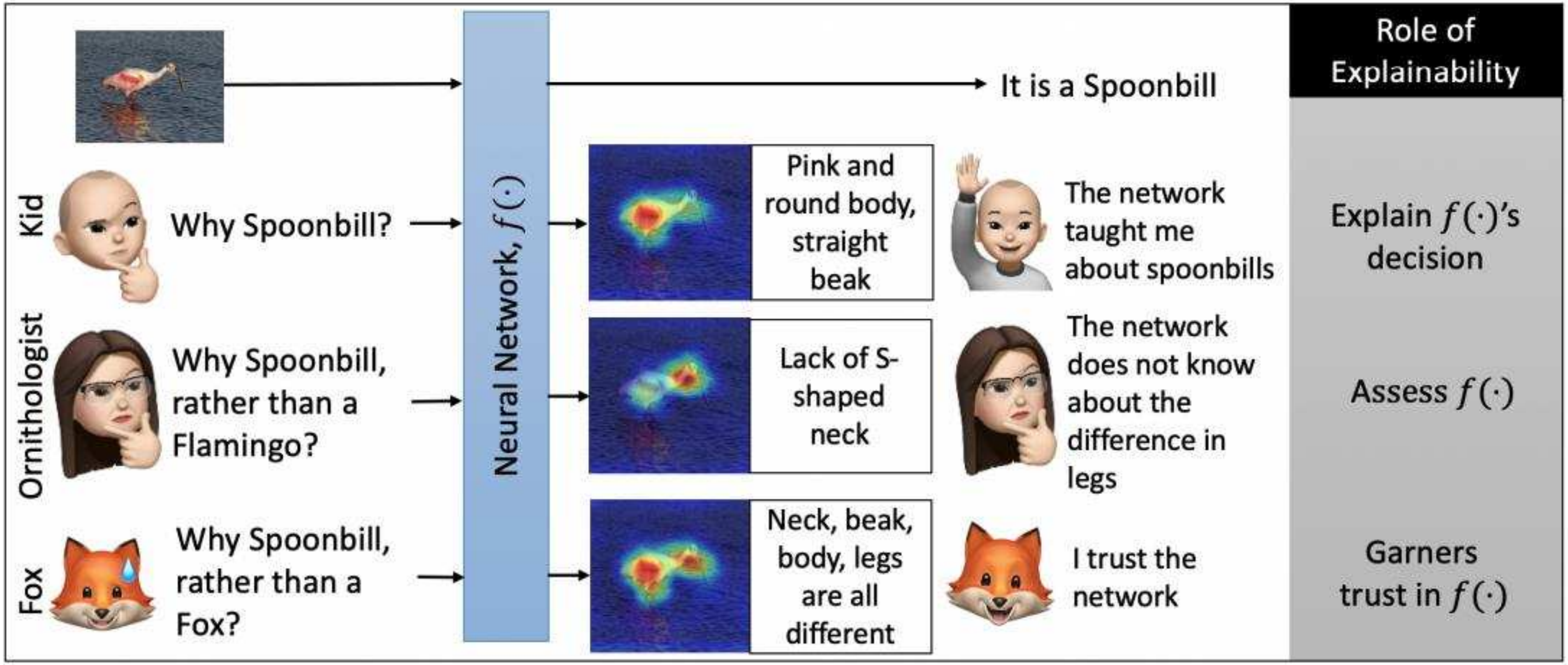
AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks: Towards relevant and contextual explanations. *IEEE Signal Processing Magazine*, *39*(4), 59-72.

Explanations

Role of Explanations – context and relevance

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

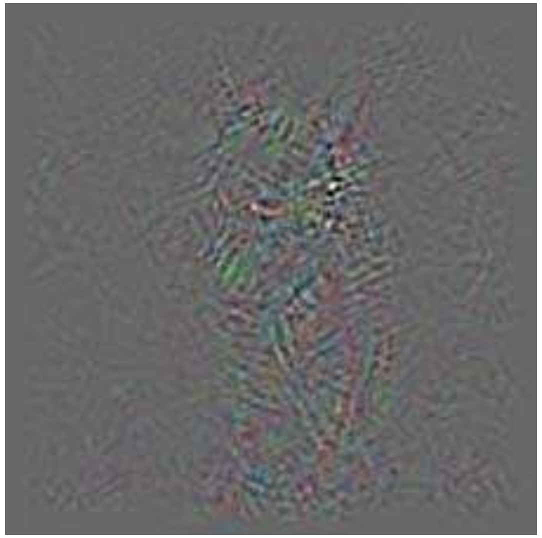


AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks: Towards relevant and contextual explanations. *IEEE Signal Processing Magazine*, *39*(4), 59-72.

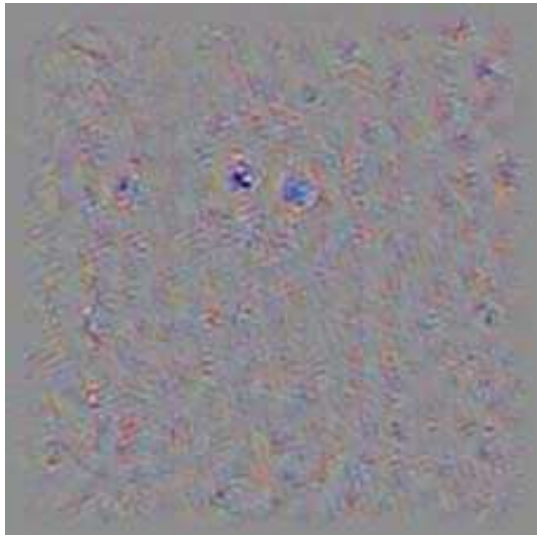
Gradients provide a one-shot means of perturbing the input that changes the output; They provide pixel-level importance scores

Input

Vanilla Gradients



Deconvolution Gradients



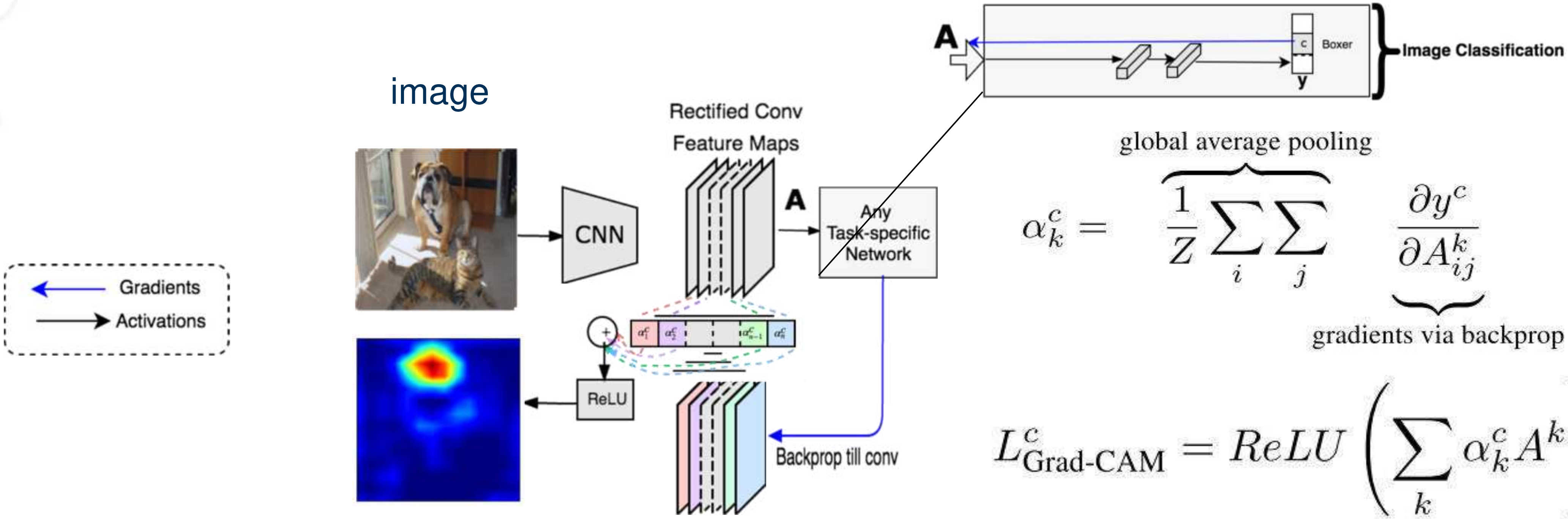
Guided Backpropagation

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)

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

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

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.

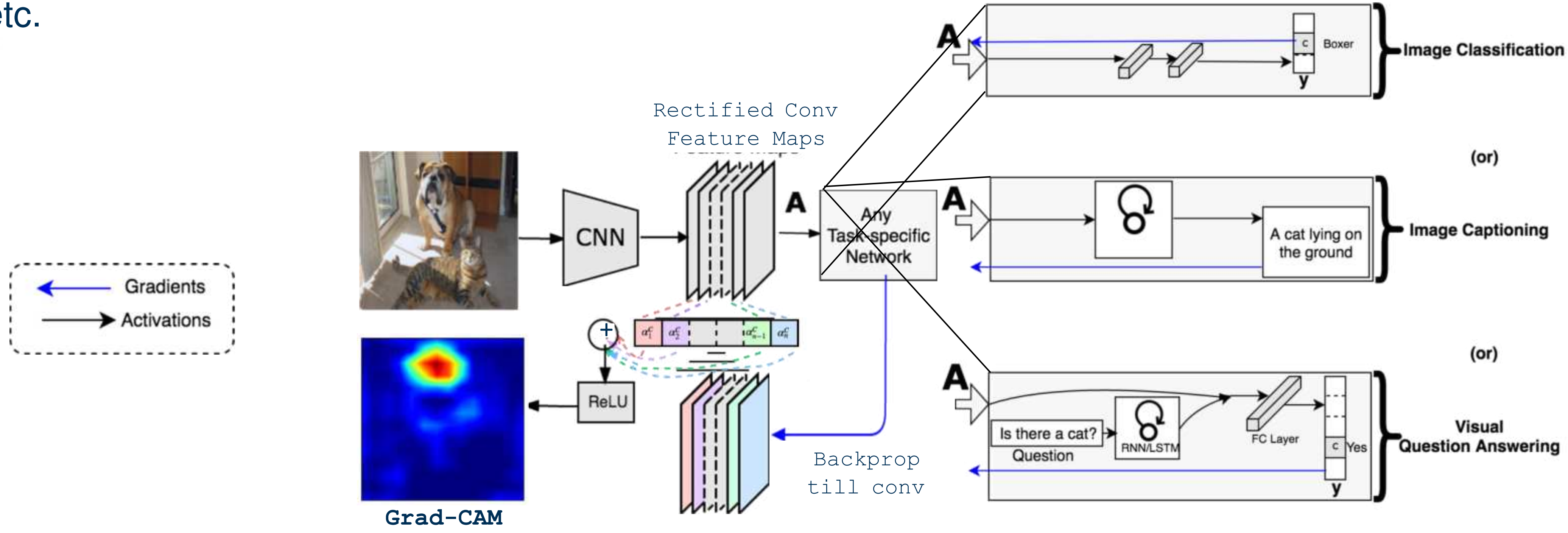
SCAN ME

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.

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?

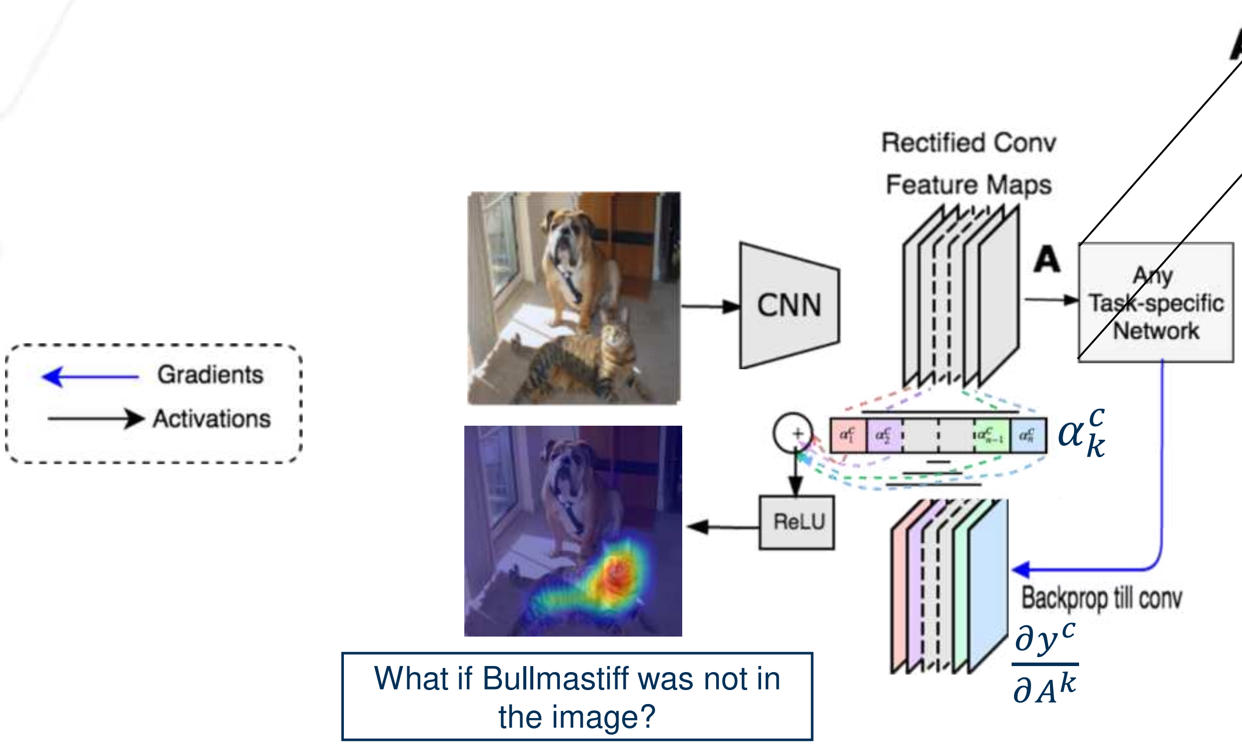
AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks: Towards relevant and contextual explanations. *IEEE Signal Processing Magazine*, *39*(4), 59-72.

Gradient and Activation-based Explanations

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

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

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



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

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

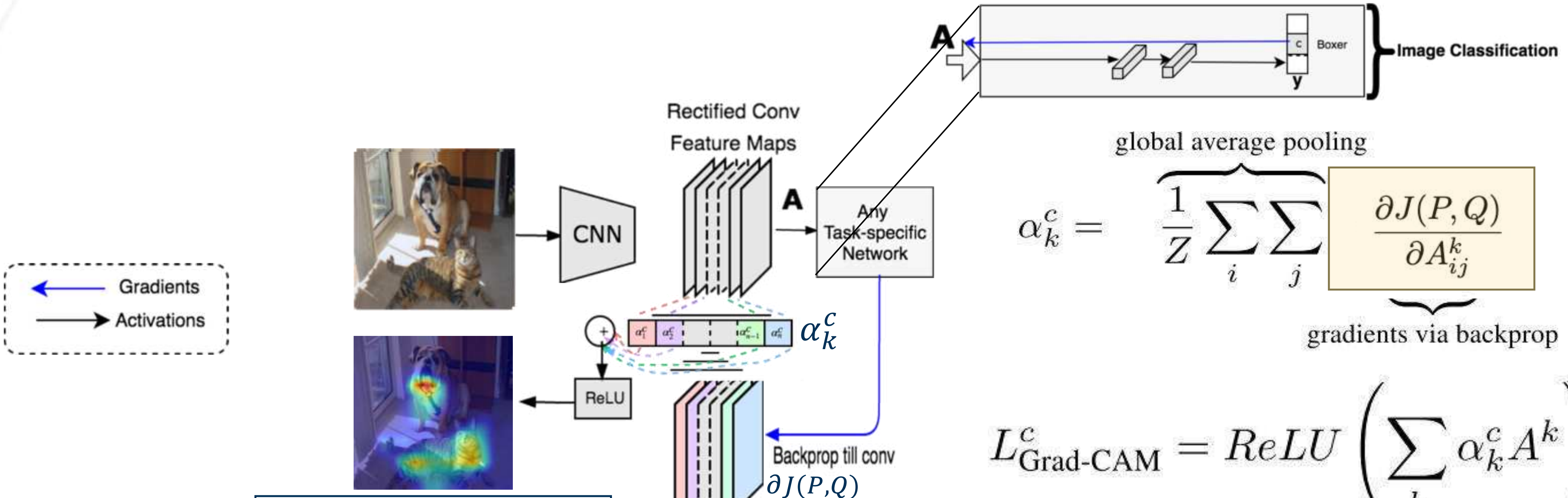
Negating the gradients effectively removes these regions from analysis

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



Contrast-CAM

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

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

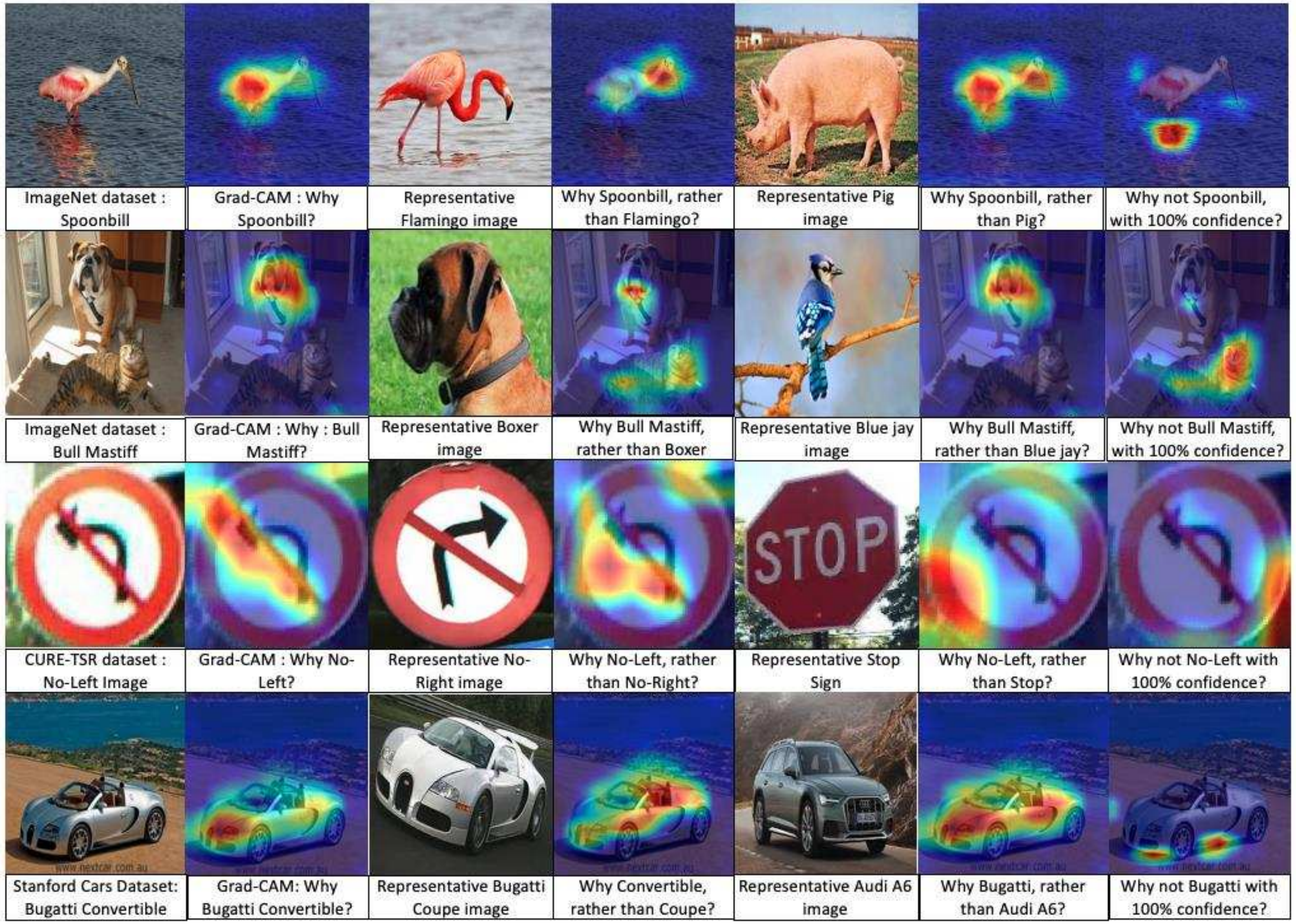
Backpropagating the loss highlights the differences between classes P and Q.

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 from GradCAM, CounterfactualCAM, and ContrastCAM

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.

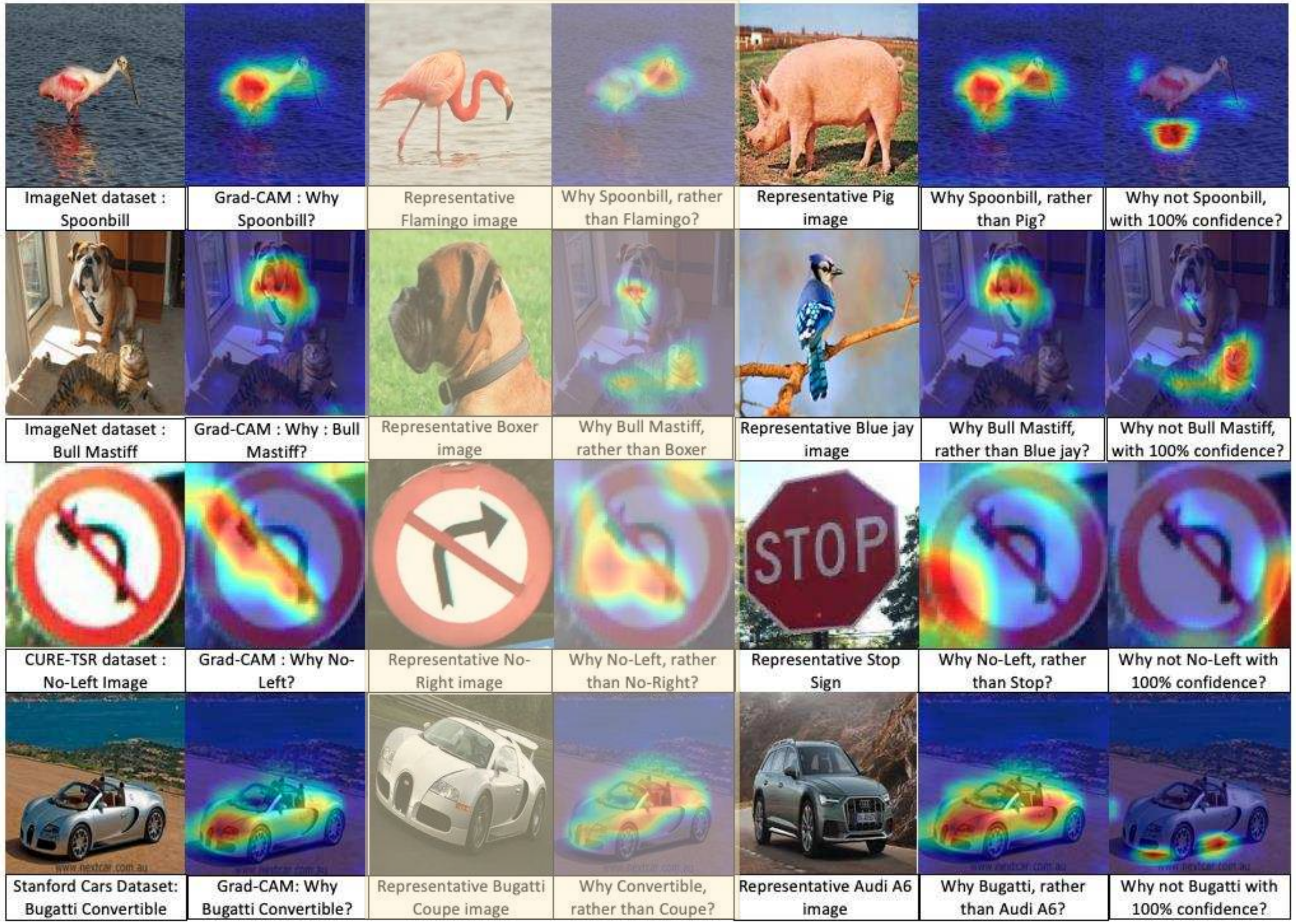
Gradient and Activation-based Explanations

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



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

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

Human Interpretable

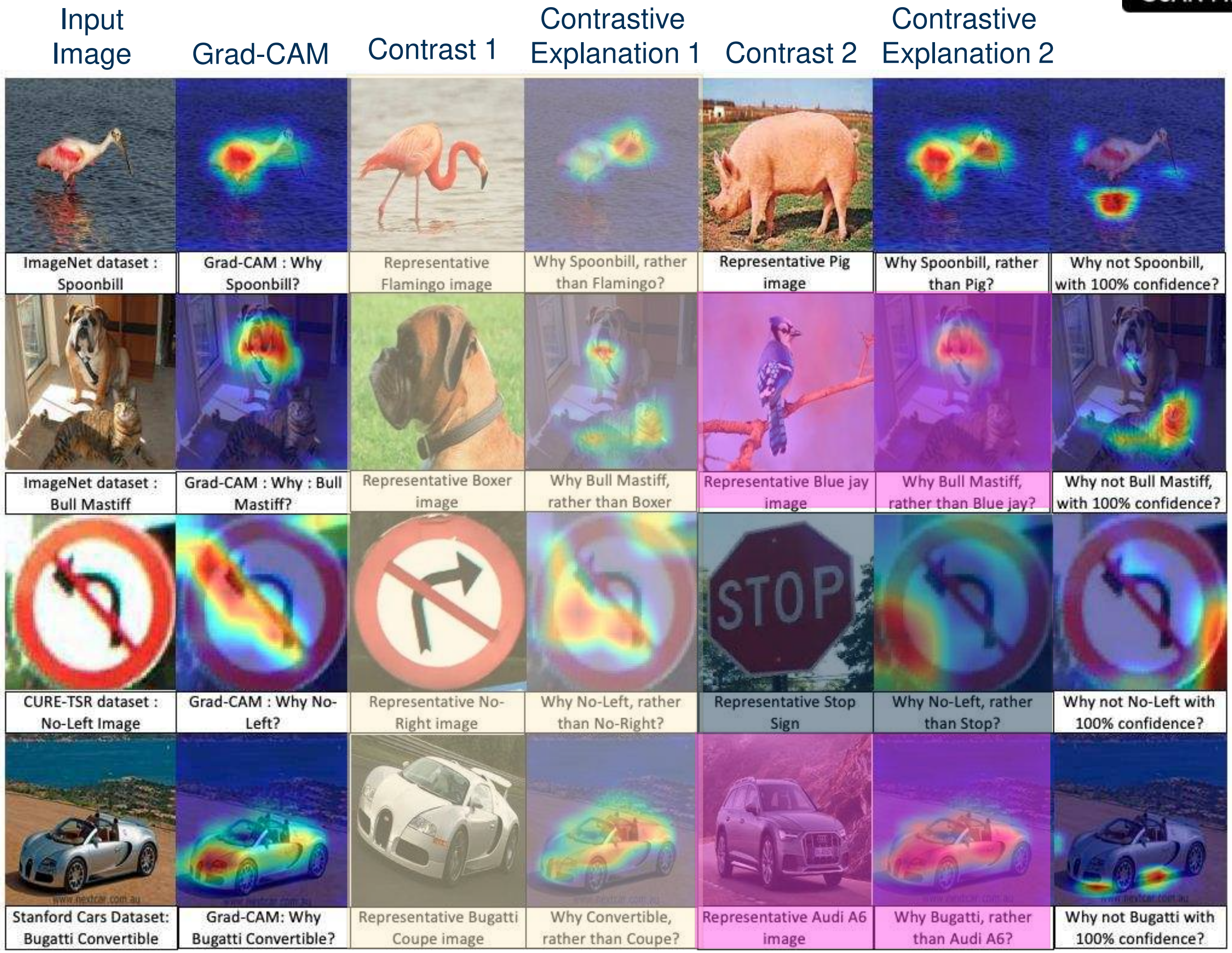
Same as Grad-CAM

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 from GradCAM, CounterfactualCAM, and ContrastCAM

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.

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

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

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 from GradCAM, CounterfactualCAM, and ContrastCAM

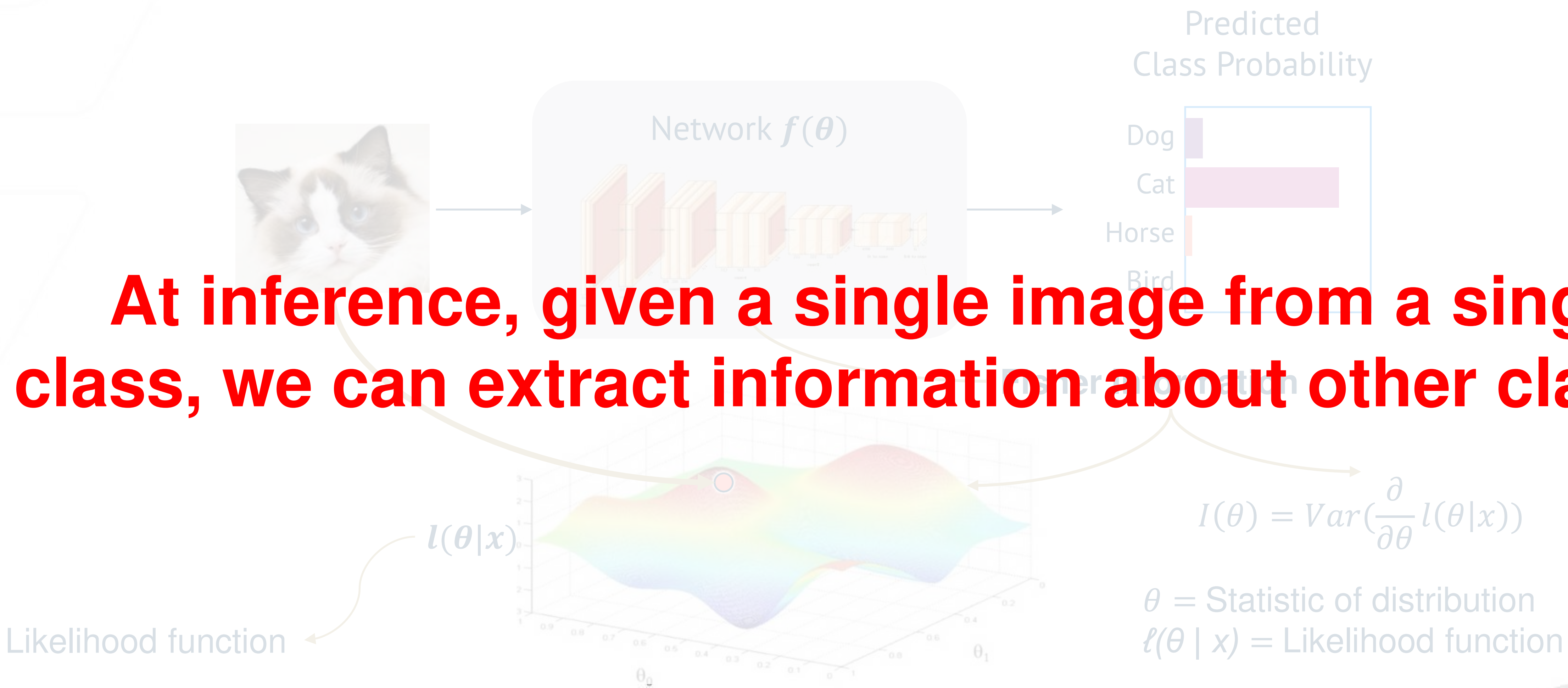
Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

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.

A Callback…
Information at Inference



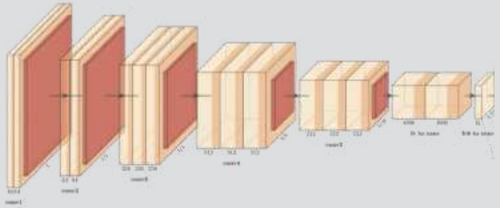
At inference, given a single image from a single class, we can extract information about other classes

Information at Inference

Case Study: Explainability

$\mathcal{T}$ is the set of all features learned by a trained network

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



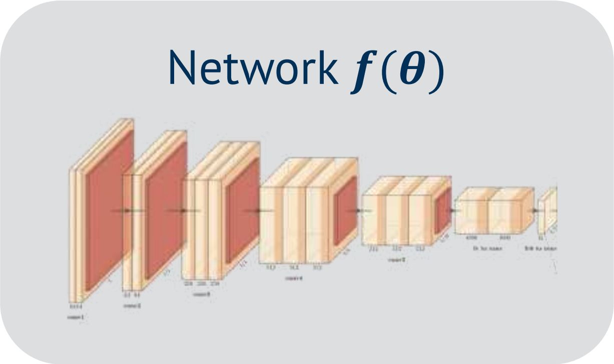
Beak
Neck
Legs
Feathers
Water
Grass
Teeth
.
.

Features $\mathcal{T}$

Information at Inference

Case Study: Explainability

Given only an image of a spoonbill, we can extract information about a Flamingo



Beak
Neck
Legs
Feathers
Water
Grass
Teeth
.
.

Features $\mathcal{T}$

All the requisite Information is stored within $\boldsymbol{f}(\boldsymbol{\theta})$

Goal: To extract and utilize this information – Introspective Learning

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

IEEE

OLIVES @GeorgiaTech

Georgia Tech

Robustness in Neural Networks

Why Robustness?

SCAN ME

How would humans resolve this challenge?

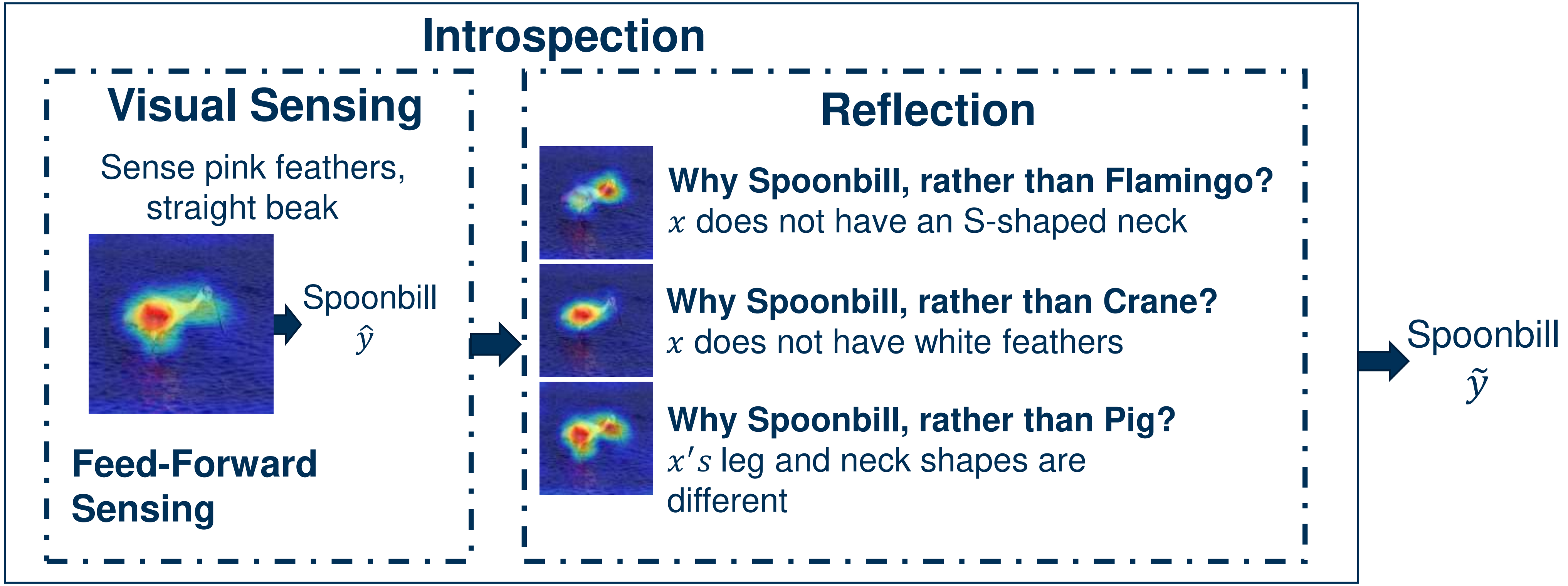
We Introspect!

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

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.

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.

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

SCAN ME

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

Goal : To simulate Introspection in Neural Networks

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

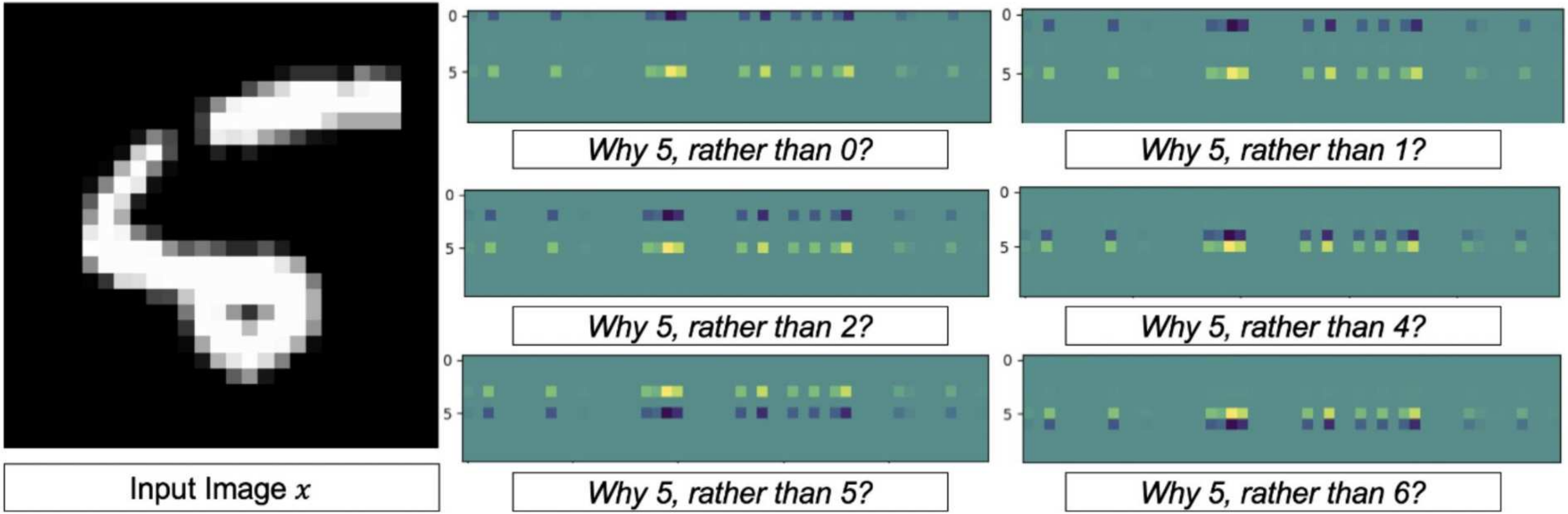
Technical Definition : *Given a network* $f(x)$*, a datum* x*, and the network's prediction* $f(x) = \hat{y}$*, introspection in* $f(\cdot)$ *is the measurement of change induced in the network parameters when a label Q is introduced as the label for* x*..*

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

SCAN ME

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

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

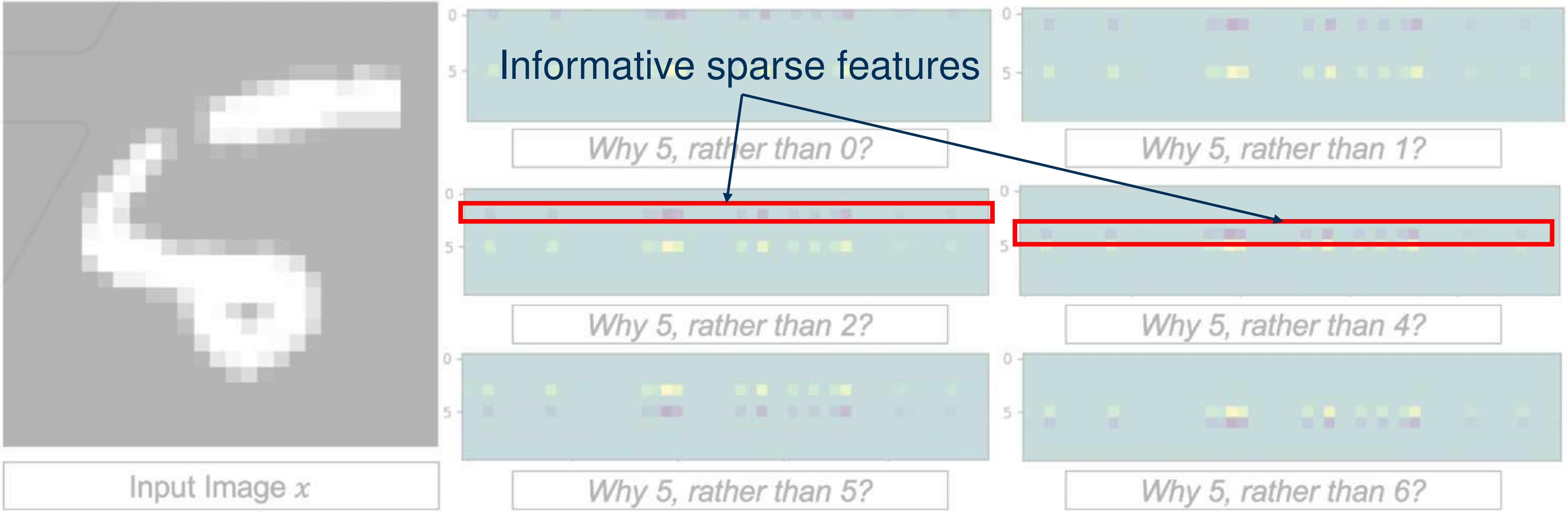


M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Introspection
Gradients as Features

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

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.

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

SCAN ME

For a well-trained network, the gradients are robust

∇_W = Gradients w.r.t. weights

J = Loss function

$\hat{y}$ = Prediction

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

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

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

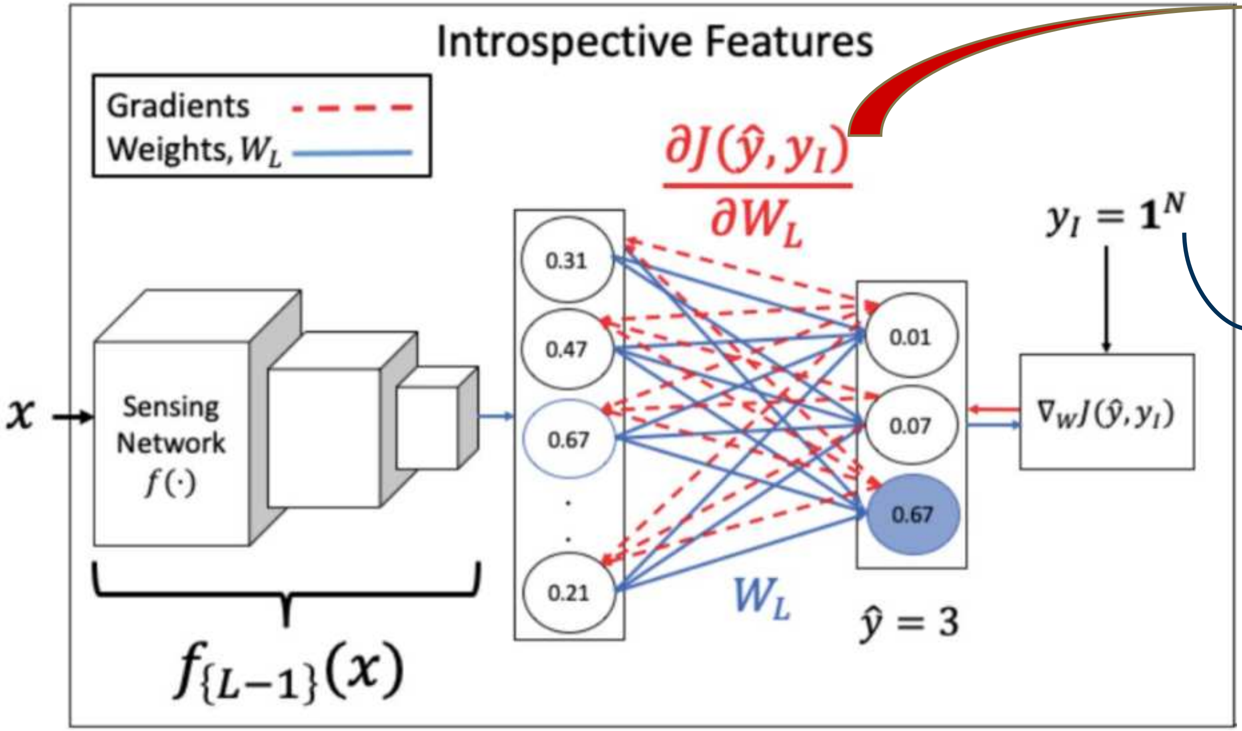
M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Introspection
Deriving Gradient Features

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

SCAN ME

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



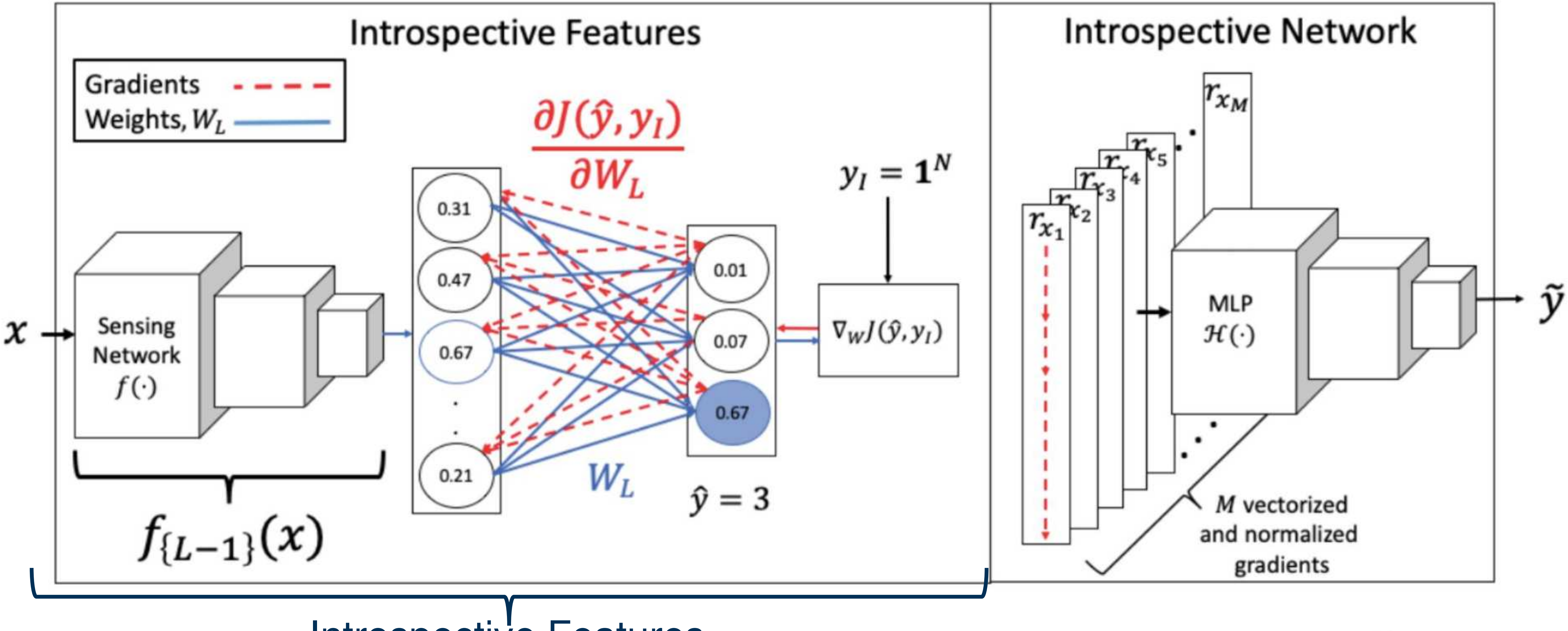
Normalized and vectorized gradients are introspective features

Vector of all ones: A confounding label!

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

Introspection
Utilizing Gradient Features

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



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

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

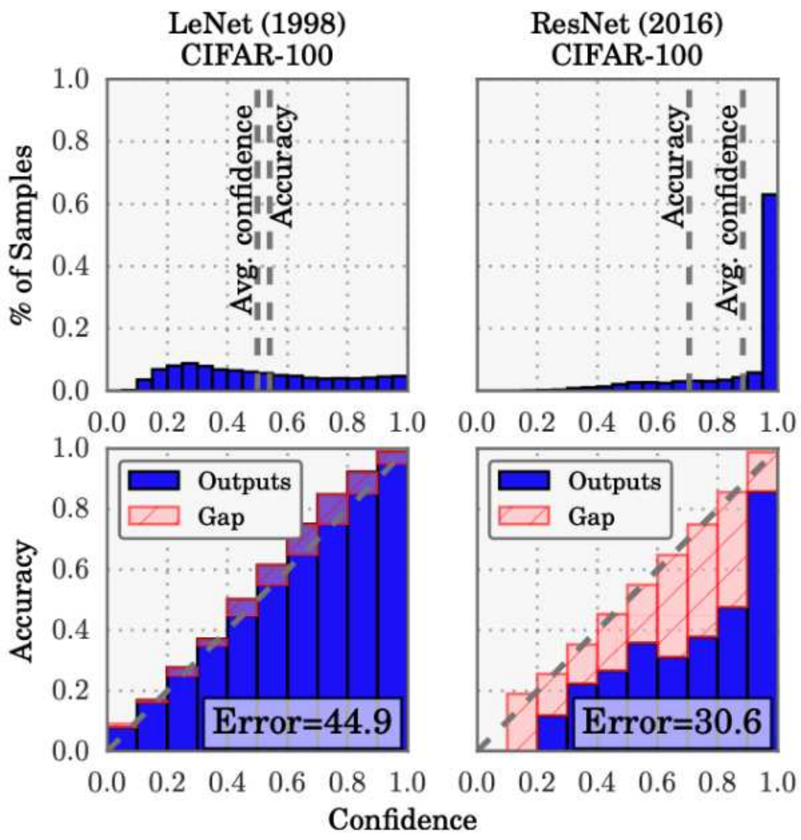
M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Calibration

A note on Calibration..

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

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



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

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

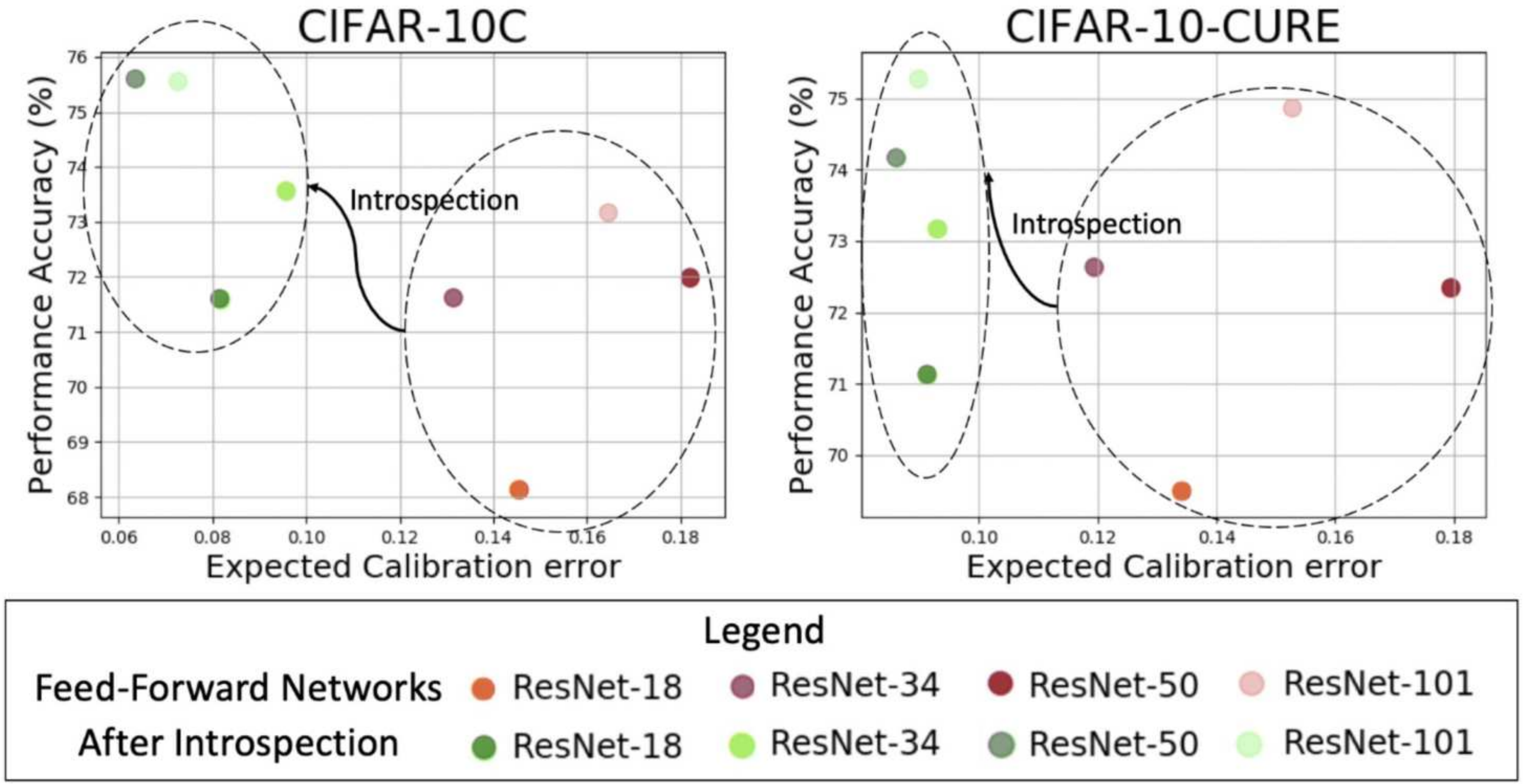
Introspection in Neural Networks
Generalization and Calibration results

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

Ideal: Top-left corner

Y-Axis: Generalization

X-Axis: Calibration



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

Introspection in Neural Networks

Plug-in nature of Introspection

SCAN ME

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

Introspection is a light-weight option to resolve robustness issues

Table 1: Introspecting on top of existing robustness techniques.

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

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

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

SCAN ME

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

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

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

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

Table 2: Recognition accuracy of Active Learning strategies.

Methods	Architecture	Original Testset		Gaussian Noise	
		R-18	R-34	R-18	R-34
Entropy (31)	Feed-Forward	0.365	0.358	0.244	0.249
	Introspective	0.365	0.359	**0.258**	**0.255**
Least (31)	Feed-Forward	0.371	0.359	0.252	0.25
	Introspective	0.373	0.362	**0.264**	**0.26**
Margin (32)	Feed-Forward	0.38	0.369	0.251	0.253
	Introspective	0.381	0.373	**0.265**	**0.263**
BALD (34)	Feed-Forward	0.393	0.368	0.26	0.253
	Introspective	0.396	0.375	**0.273**	**0.263**
BADGE (33)	Feed-Forward	0.388	0.37	0.25	0.247
	Introspective	0.39	0.37	**0.265**	0.260

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

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

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

Robust Neural Networks

Part 3: Uncertainty at Inference

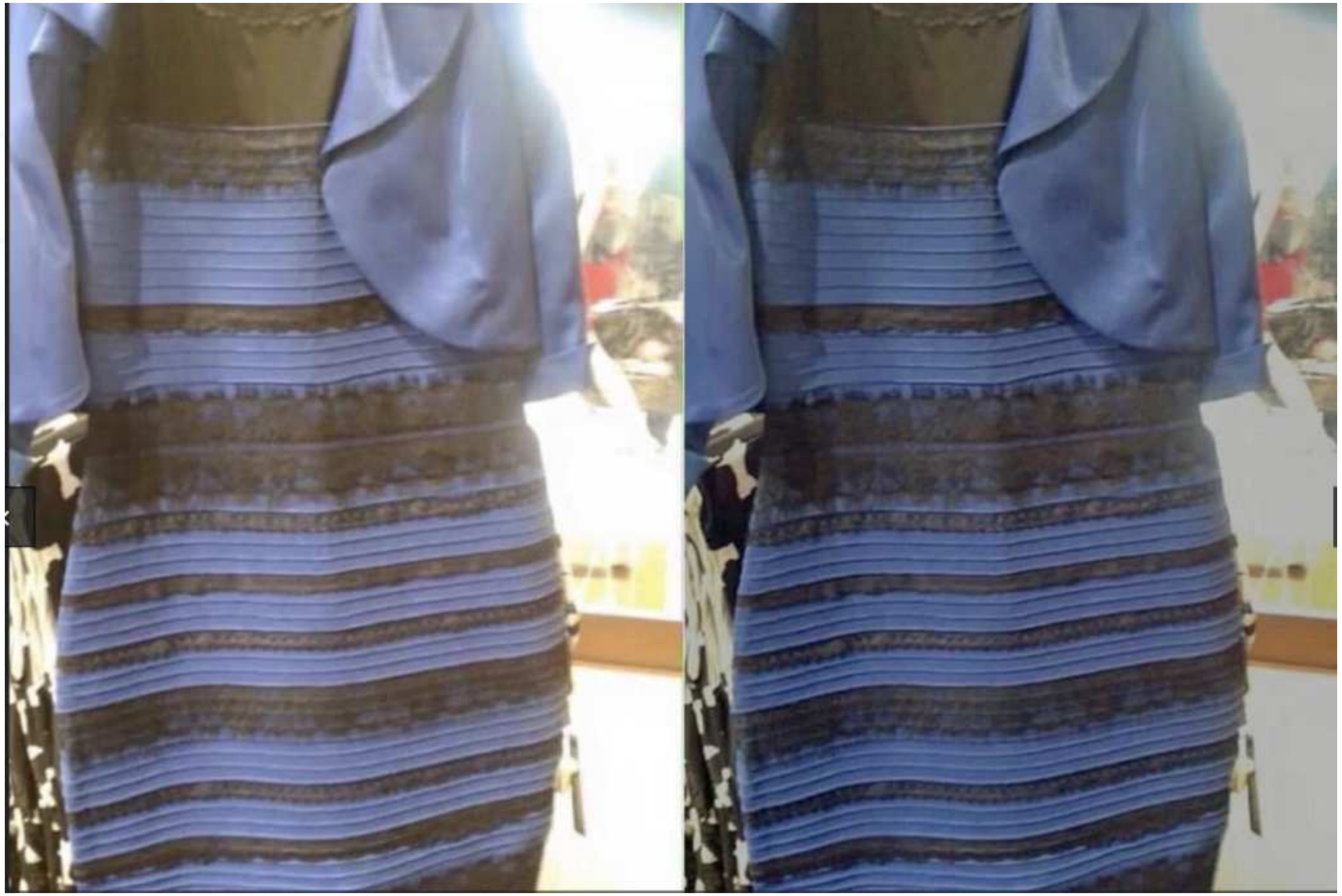
Objective

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

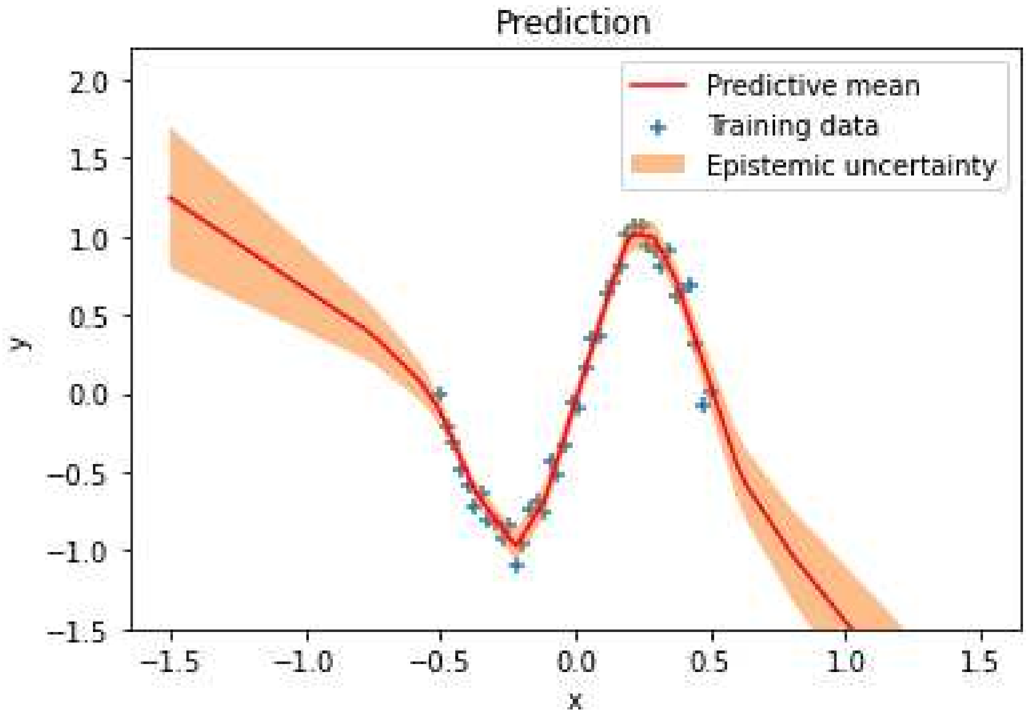
Uncertainty is a model knowing that it does not know



White and Gold
Or
Blue and Black?

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

Uncertainty is a model knowing that it does not know

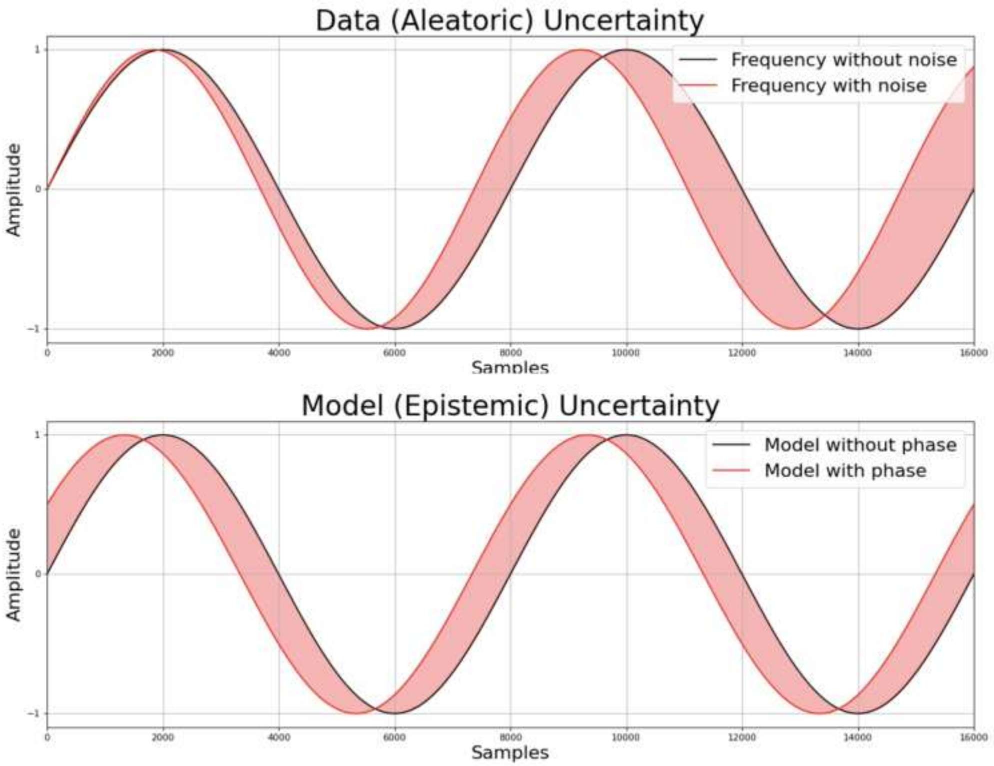


A simple example:

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

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

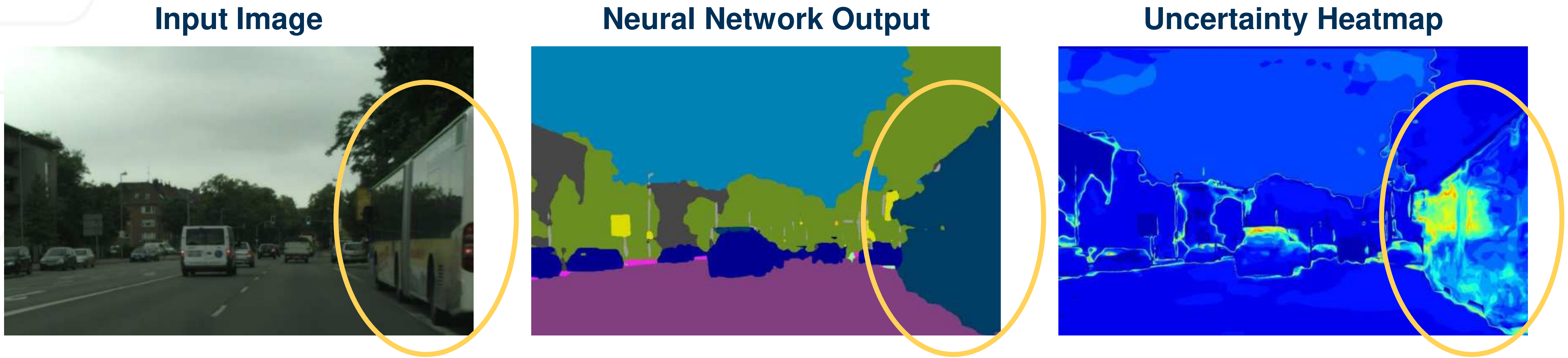
Uncertainty is a model knowing that it does not know



A slightly more complex example:

- **Data (Aleatoric) Uncertainty**: When there is inherent noise in available data or in measurement of data
- **Model (Epistemic) Uncertainty:** When our chosen model (network) is incapable of modeling the data

Uncertainty is a model knowing that it does not know



Kendall, Gal "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision." *NIPS 2017*

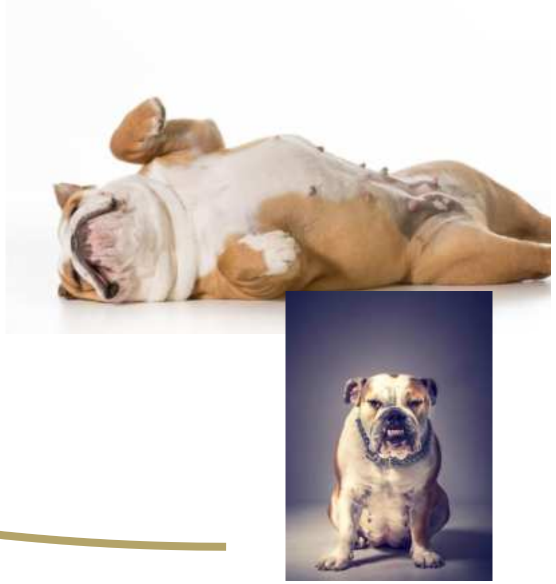
Uncertainty
Challenge in Uncertainty Quantification

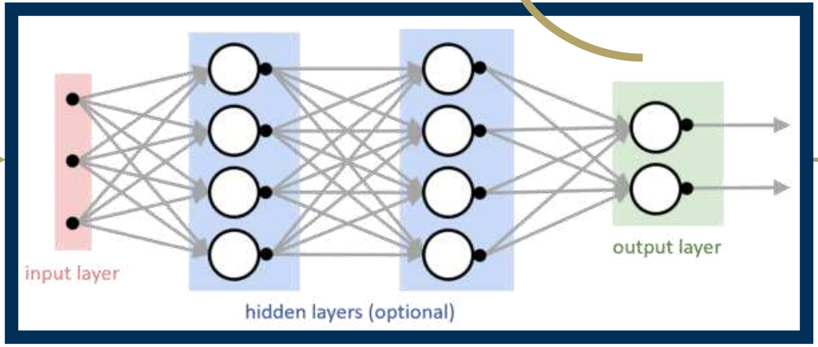
Primary purpose of neural networks (ex: classification) and Uncertainty Quantification do not always go hand-in-hand!

All **required** information is passed to last layer
Maximal logit is the class

Required information is task dependent! A well-trained classification network ignores the attributes of the dog

Dog asking for belly rub = Angry dog!

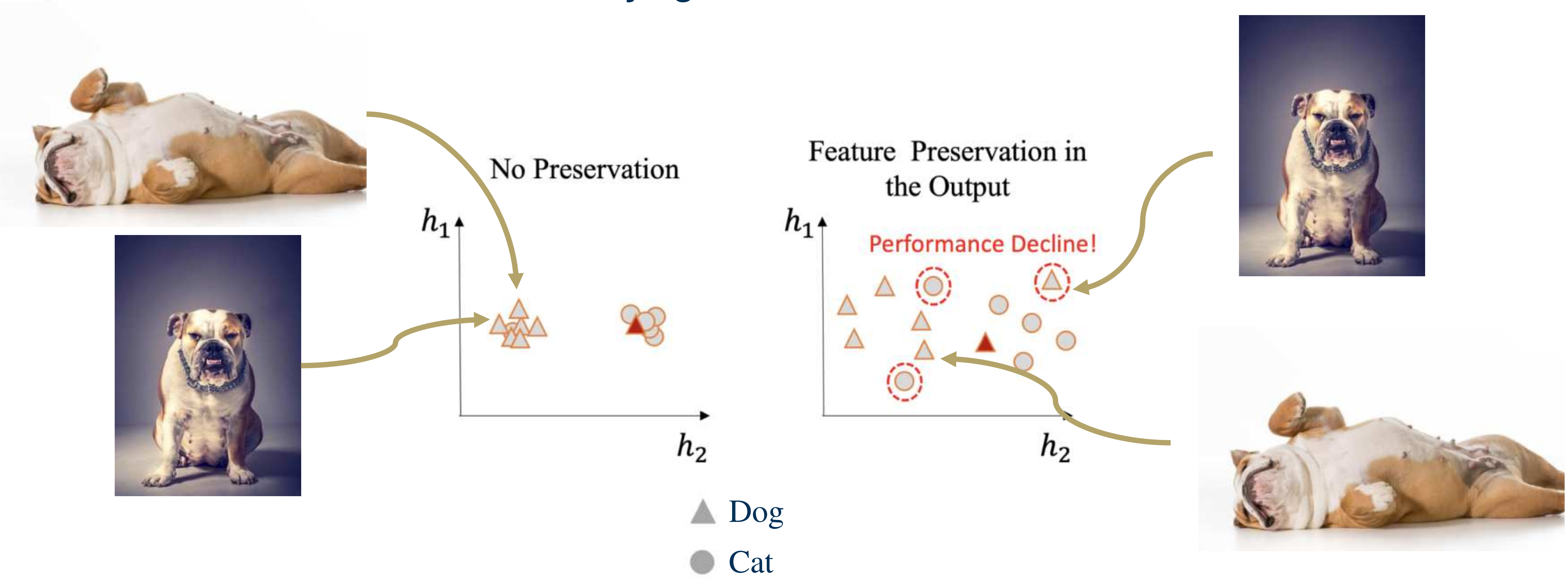




Dog

R. Benkert, M. Prabhushankar, and G. AlRegib, “Transitional Uncertainty with Layered Intermediate Predictions,” in International Conference on Machine Learning (ICML), Vienna, Austria, 2024

Primary purpose of neural networks (ex: classification) and Uncertainty Quantification do not always go hand-in-hand!

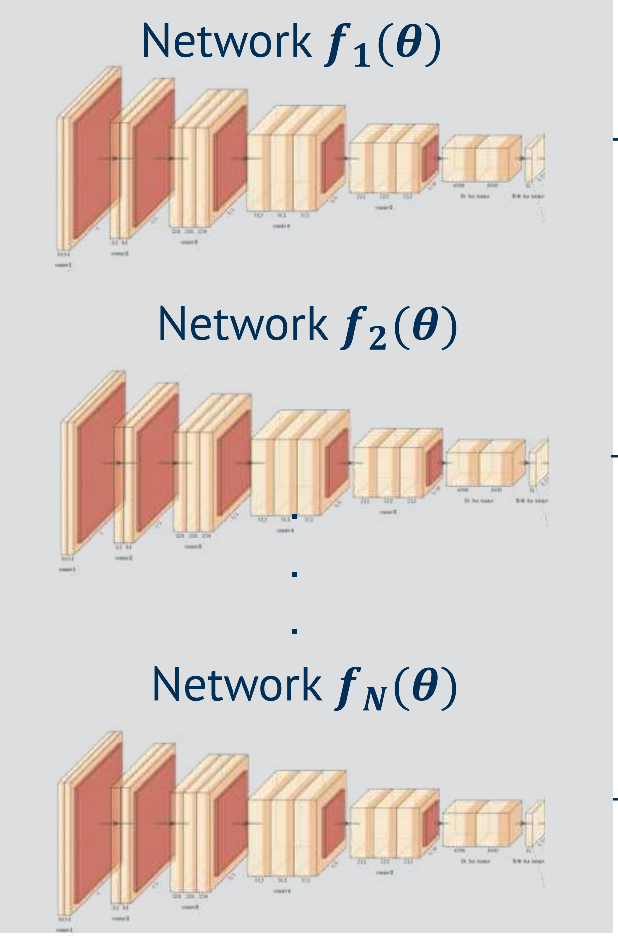


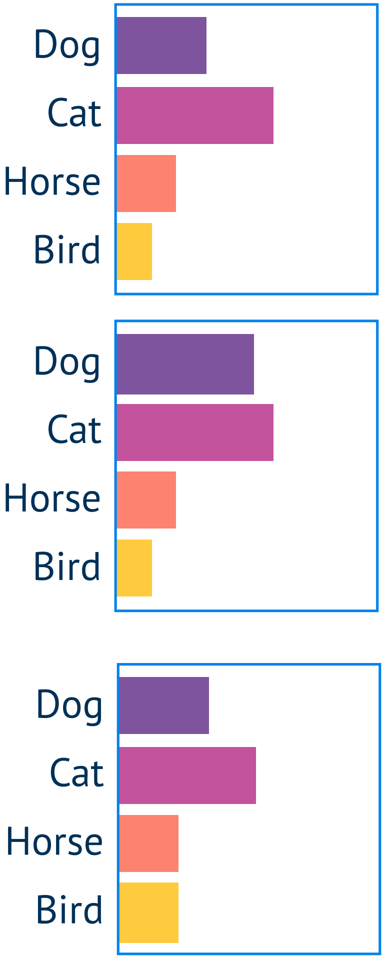
R. Benkert, M. Prabhushankar, and G. AlRegib, "Transitional Uncertainty with Layered Intermediate Predictions," in International Conference on Machine Learning (ICML), Vienna, Austria, 2024

Uncertainty

Uncertainty Quantification in Neural Networks

Via Ensembles[1]





Variation within outputs is the uncertainty.

Commonly referred to as **Prediction Uncertainty.**

Requires multiple trained models – not exactly an inferential method

[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

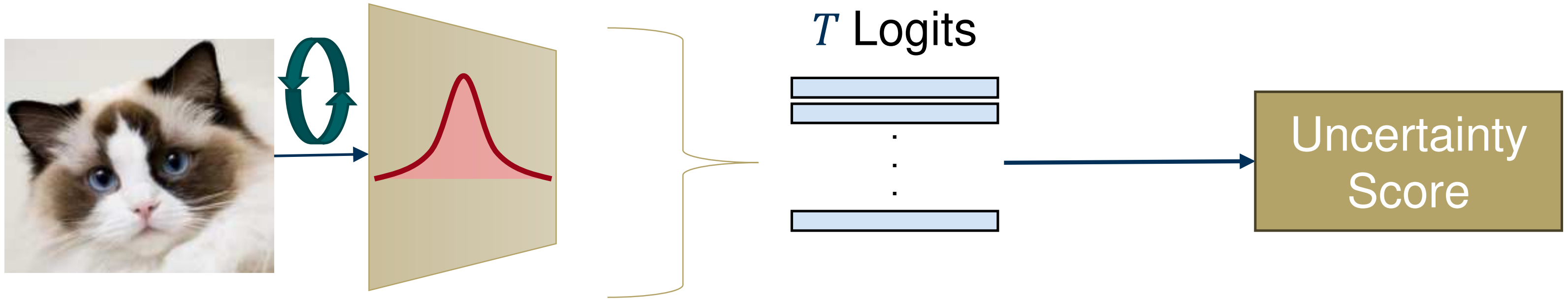
Iterative Uncertainty Quantification

Via Monte-Carlo Dropout[1]: During inference repeated evaluations with the same input give different results

Multiple forward passes with random dropout simulate $f_1(\cdot), f_2(\cdot), f_3(\cdot) \ldots f_T(\cdot)$.

$$U_{epistemic} = \underbrace{H\left(\frac{1}{T}\sum_{t=1}^{T} Softmax\left(f_{\widehat{W}_t}(\boldsymbol{x})\right)\right)}_{U_{Predictive}} - \underbrace{\frac{1}{T}\sum_{t=1}^{T} H\left(Softmax\left(f_{\widehat{W}_t}(\boldsymbol{x})\right)\right)}_{U_{aleatoric}}$$

T forward passes



Final prediction is maximum of the mean of the outputs

[1] Y Gal, Z Ghahramani, “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning”, ICML 2016

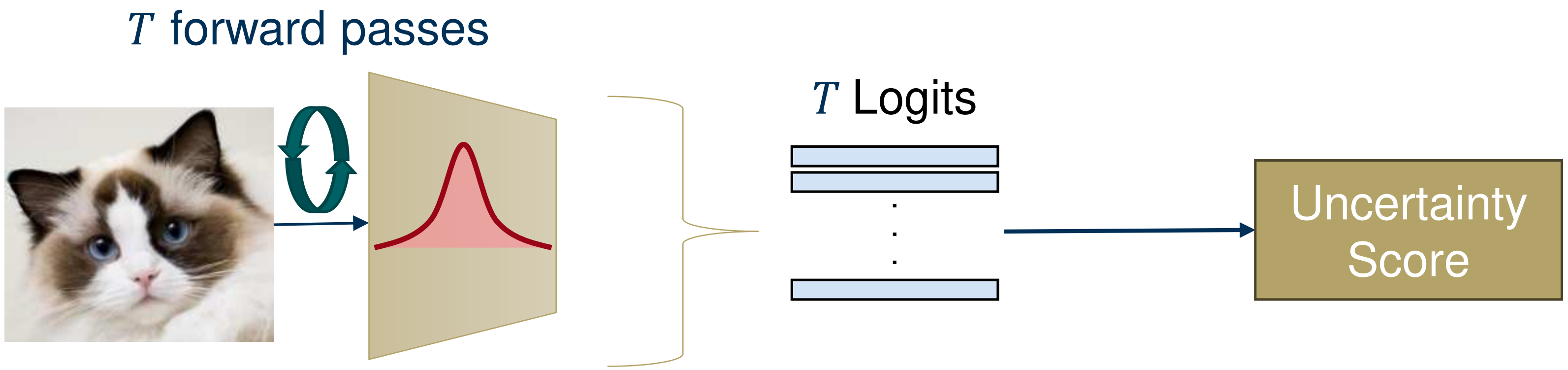
Uncertainty

Iterative Uncertainty Quantification

Via Monte-Carlo Dropout[1]: During inference repeated evaluations with the same input give different results

Multiple forward passes with random dropout simulate $f_1(\cdot), f_2(\cdot), f_3(\cdot) \ldots f_T(\cdot)$.

- **Requires dropout percentage to be set at training. Different models may require different dropout percentages at inference**
- **For a well-trained model, dropout underestimate uncertainty**
- **For a high-error model, dropout overestimate uncertainty**



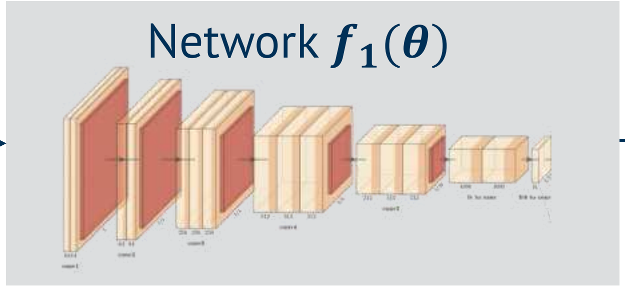
Final prediction is maximum of the mean of the outputs

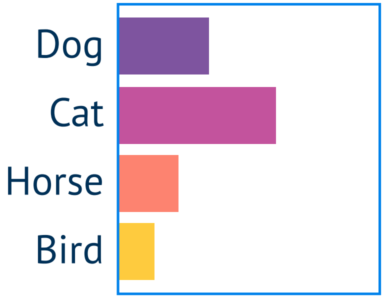
[1] Y Gal, Z Ghahramani, “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning”, ICML 2016

Uncertainty

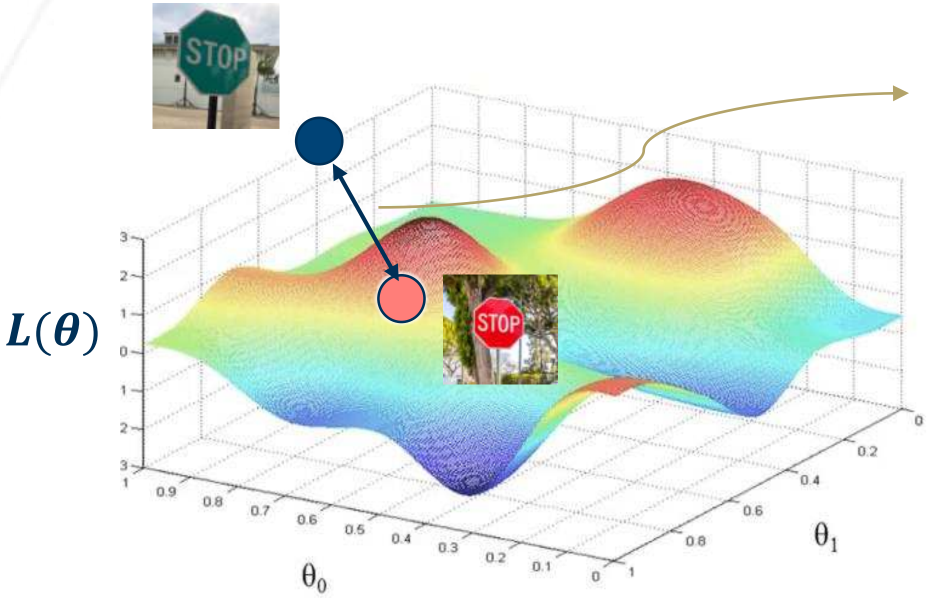
Single Pass Uncertainty Quantification

Distance to training/validation representation space is uncertainty





Uncertainty quantification using a single network and a single pass



Calculate distance from some trained clusters

Does not require multiple networks or passes!

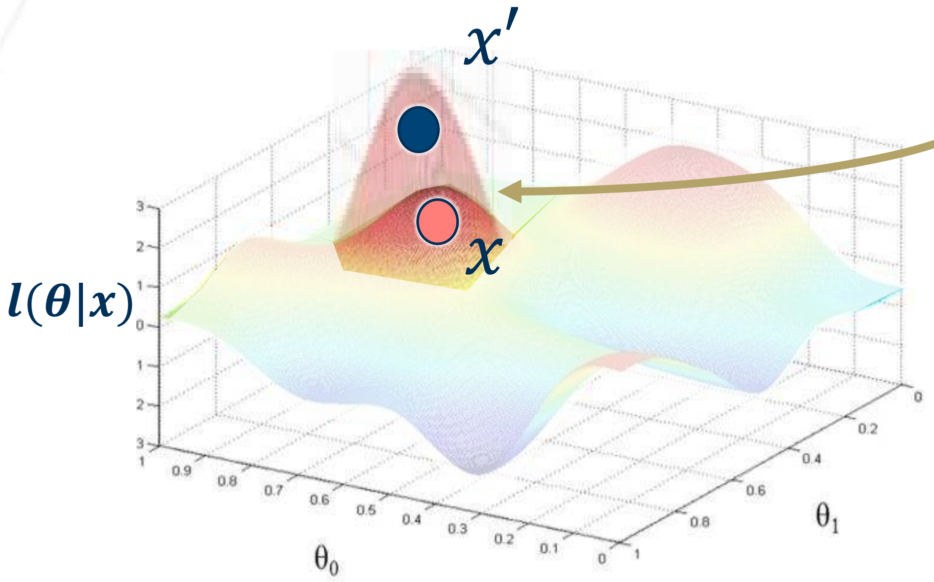
However, requires training data/validation set/addition models at inference

[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 Uncertainty Quantification

Principle: Gradients provide a 'distance measure' between the learned representations space and its prediction (for discriminative tasks) or some new data (for generative tasks)



Gradients quantify the required movement of an unknown representation space that encompasses the test sample

Does not require multiple networks or passes!

Does not require training data/validation set/addition models at inference!

However, what is $l(\theta|x)$ at inference?

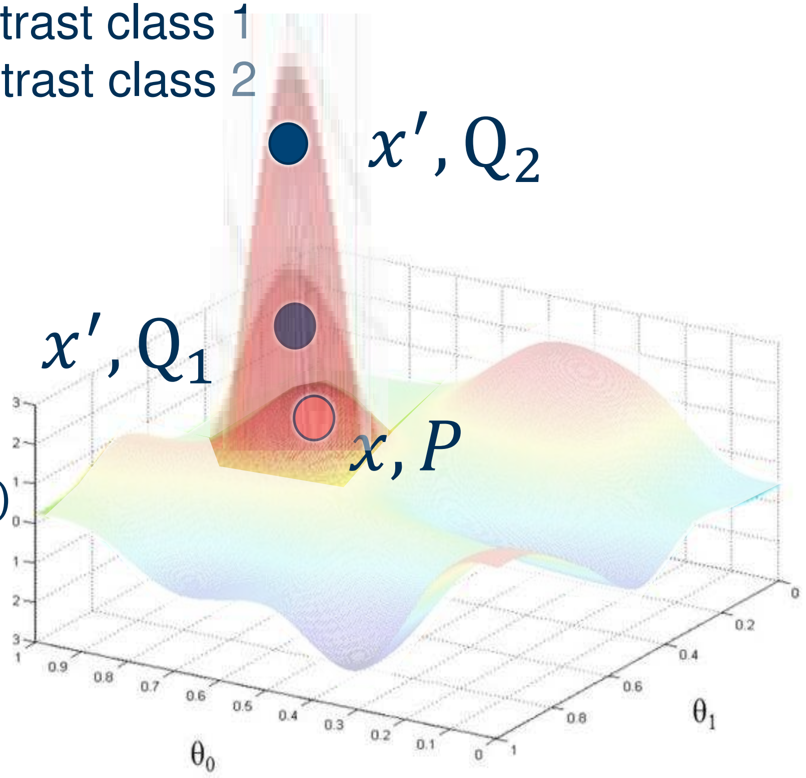
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

Principle: Gradients provide an uncertainty 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 $\boldsymbol{l(\theta|x)}$ at inference?

During training, $l(\theta|x)$ is a loss function between predicted class and ground truth class. At inference, we do not have access to ground truth class

We backpropagate all contrast classes - $Q_1, Q_2 \dots Q_N$ by backpropagating a confounding label – a vector of all ones!

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

IEEE Access

Probing the Purview of Neural Networks via Gradient Analysis

Jinsol Lee,
PhD Candidate

Mohit Prabhushankar, PhD
Postdoc

Ghassan AlRegib, PhD
Professor

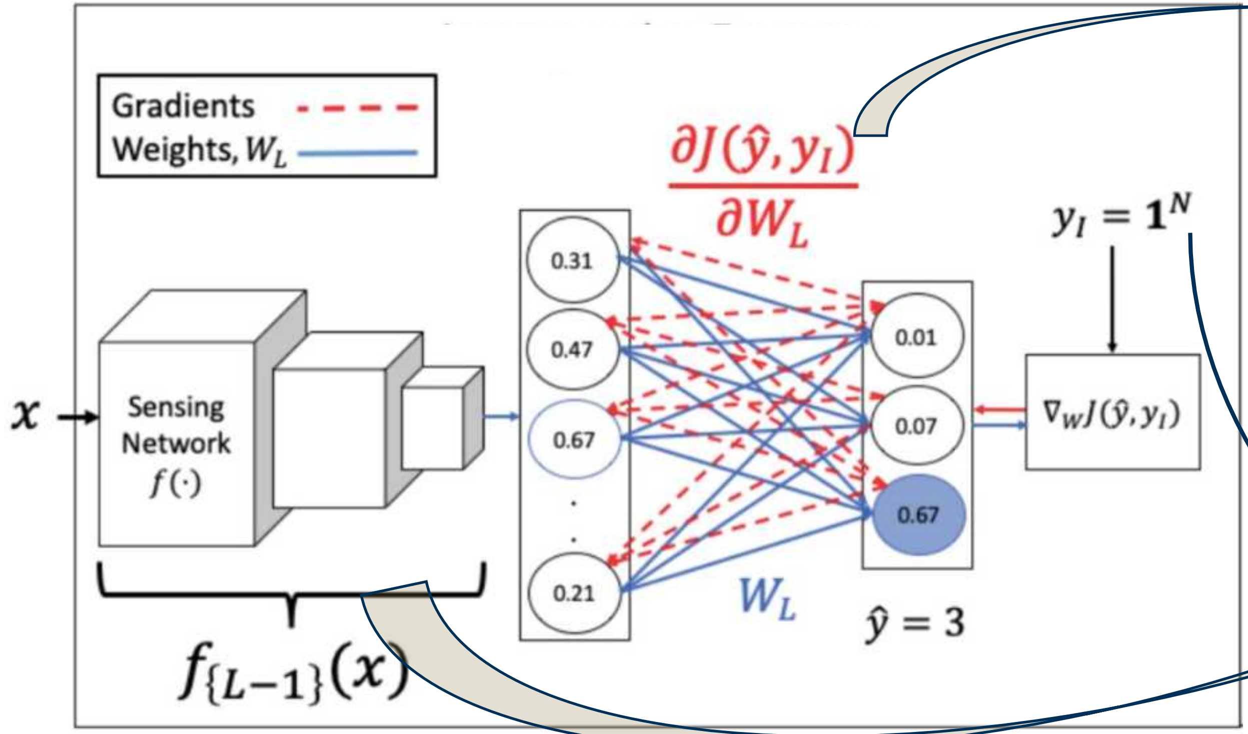
SCAN ME

IEEE

OLIVES @GeorgiaTech

Georgia Tech

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



Normalized and vectorized gradients (same as introspective features)

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

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

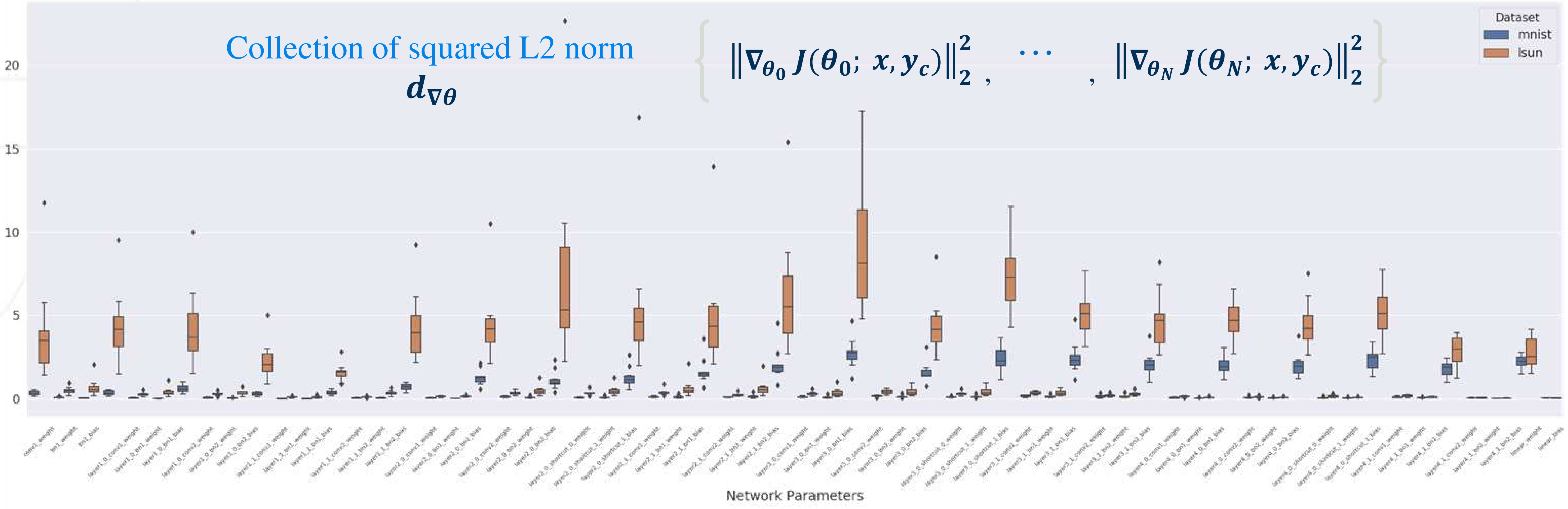
Uncertainty in Neural Networks

Utilizing Gradient Features

Step 2: Take L2 norm of all generated gradients

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

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



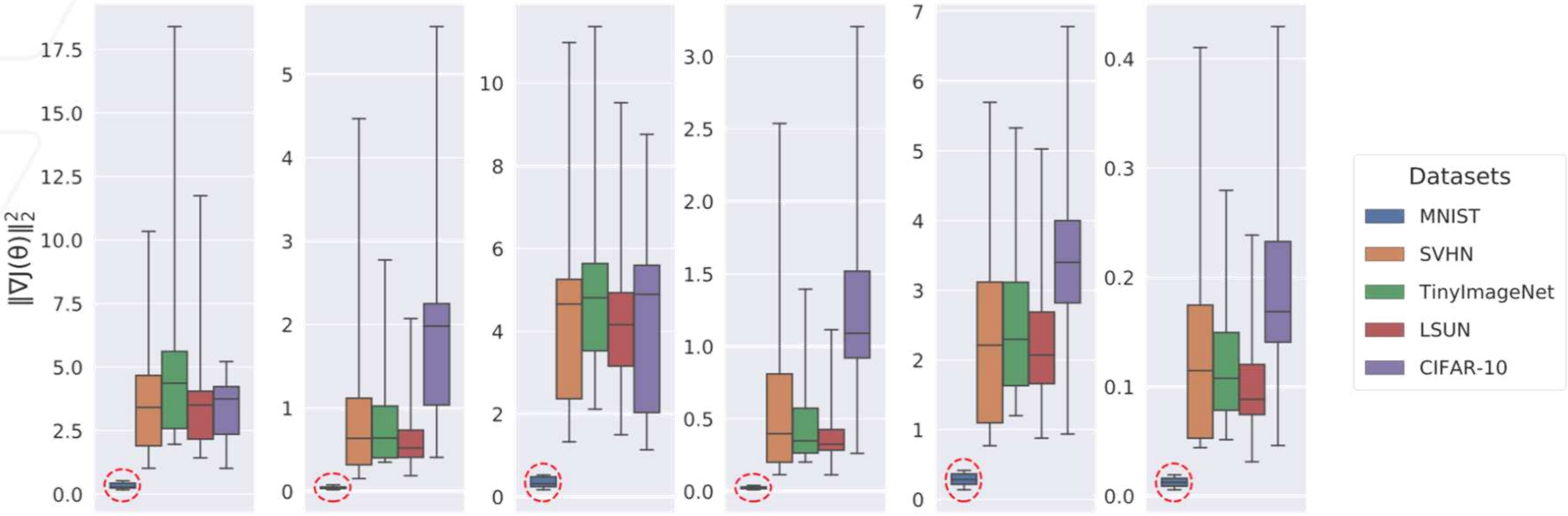
MNIST: In-distribution, SUN: Out-of-Distribution

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

SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

Squared L2 distances for different parameter sets



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

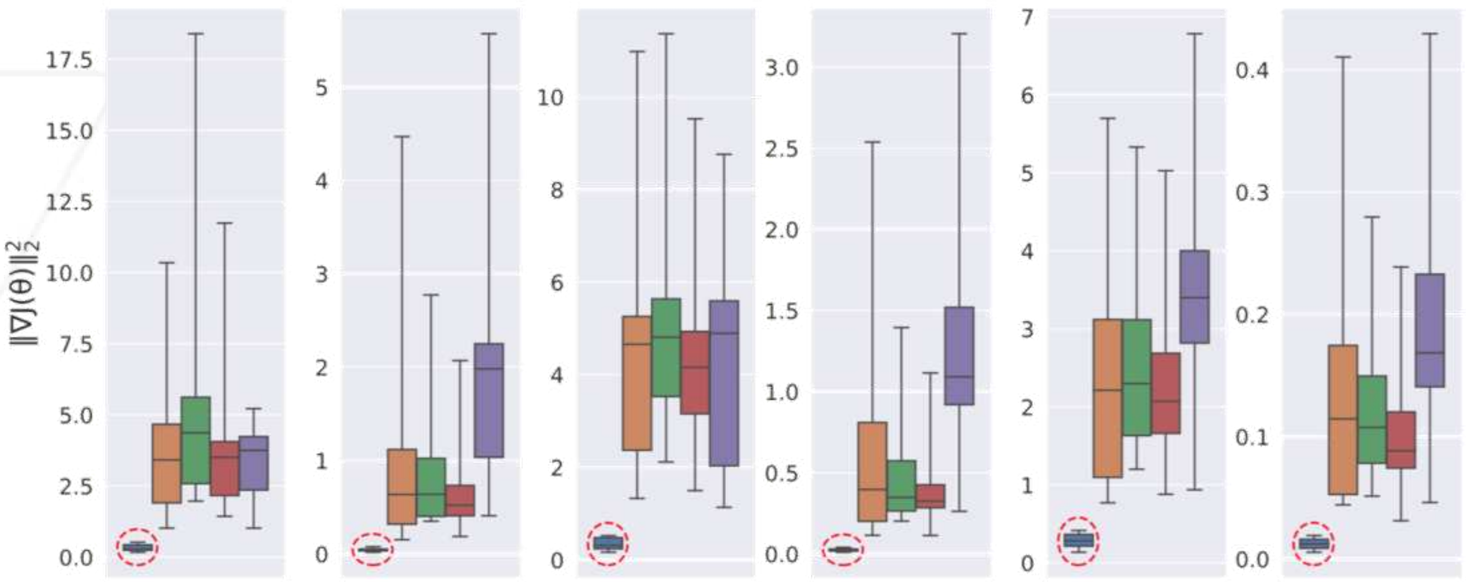
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Gradient-based Uncertainty

Experimental Setup

SCAN ME

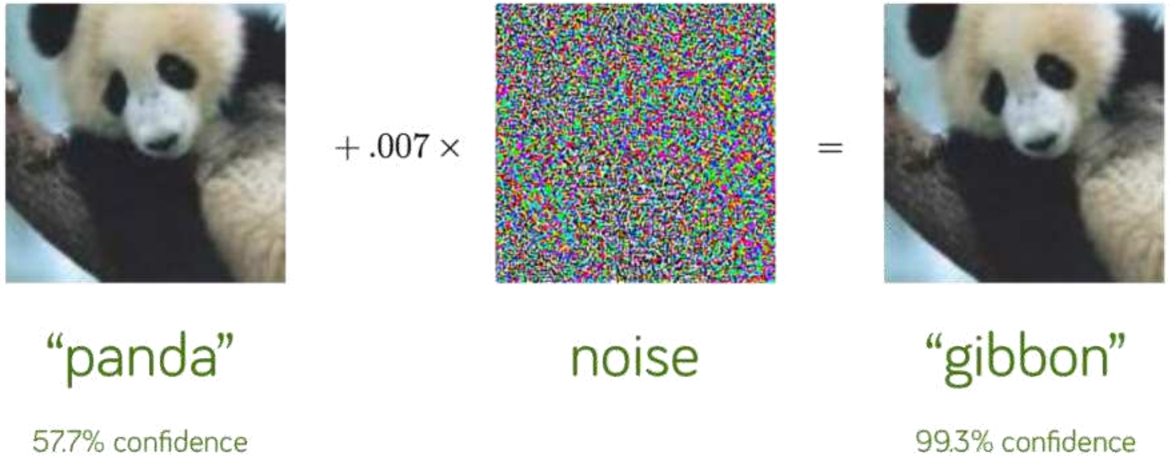
Utilize this discrepancy in trained vs untrained data gradient L2 distance to detect adversarial, noisy, and OOD data



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.

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

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

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

Gradient-based Uncertainty

Uncertainty in Detecting Challenging Conditions

Probing the Purview of Neural Networks via Gradient Analysis

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 in Detecting 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**

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

OLIVES @GeorgiaTech

Georgia Tech

Gradient-based Uncertainty

Uncertainty in Detecting 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.

Out-of-Distribution Detection

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

Train set →  MNIST

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

SVHN

CIFAR10

TinyImageNet

LSUN

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

Out-of-Distribution Detection

SCAN ME

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

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

Counterfactual Gradients-based Quantification of Prediction Trust in Neural Networks

Mohit Prabhushankar, PhD
Postdoc

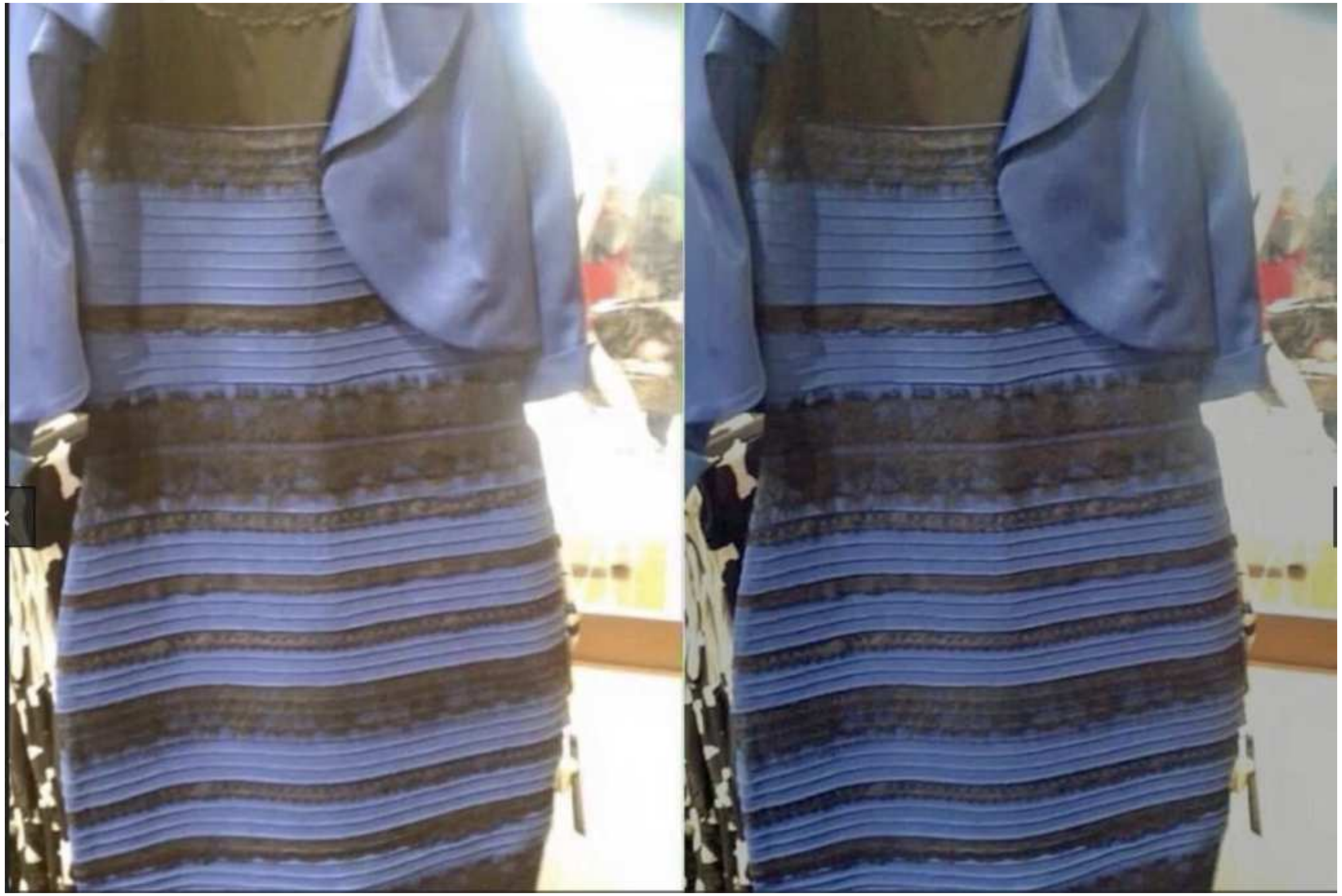
Ghassan AlRegib, PhD
Professor

SCAN ME

Trust

Definition

Trust: An esoteric term that encompasses uncertainty, belief, and apriori probability



White and Gold
Or
Blue and Black?

Trust is application-specific

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

Trust vs Trustworthiness
Trustworthiness attributes

Trustworthiness Attributes: Applications in ML that satisfy the attributes of performance, reliability, human interaction, and aligned purpose

- Explainability
- Out-of-distribution Detection
- Adversarial Detection
- Anomaly Detection
- Corruption Detection
- Differential Privacy
- Causal Analysis
- Open-set Recognition
- Noise Robustness
- Uncertainty Quantification
- Uncertainty Visualization
- …

More relevant during model testing

Relevant at Deployment:

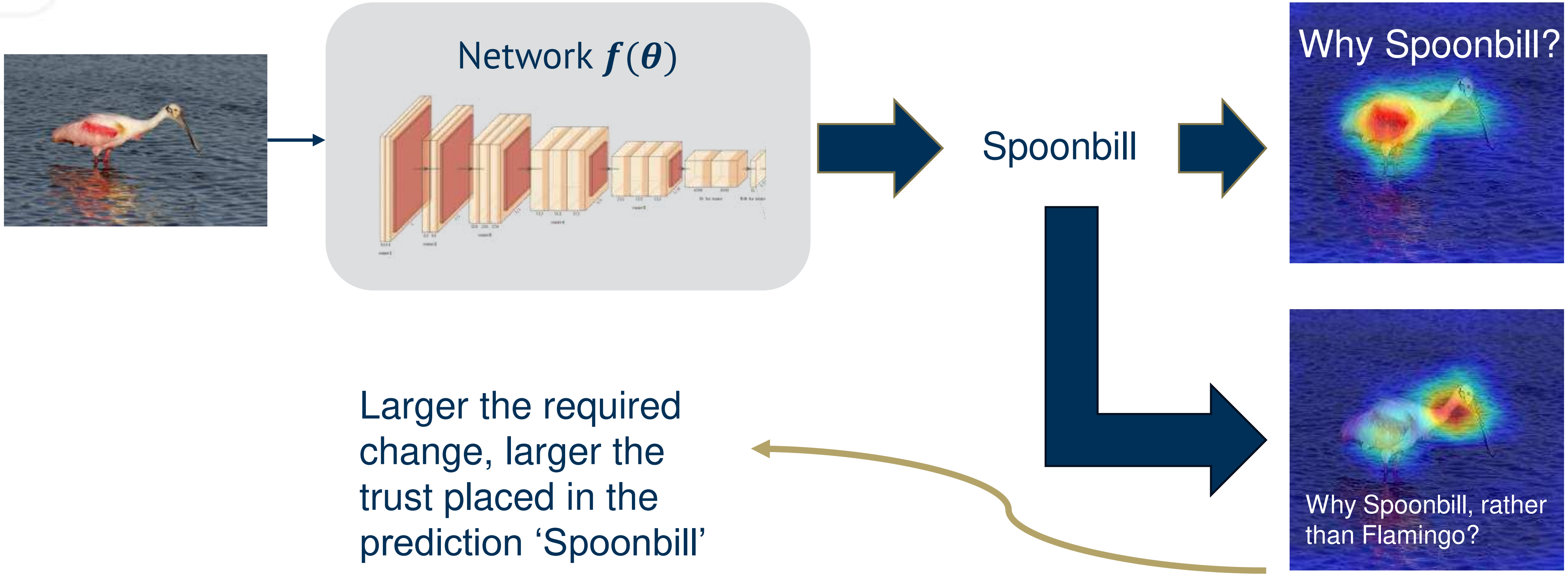
Provide a specific ‘trust score’ that objectively allows users to trust neural network predictions

GradTrust provides such a score!

GradTrust

Intuition for counterfactual gradients-based Trust

How much change is required within the data to predict a counterfactual class? Larger the required change, larger the trust



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.

GradTrust

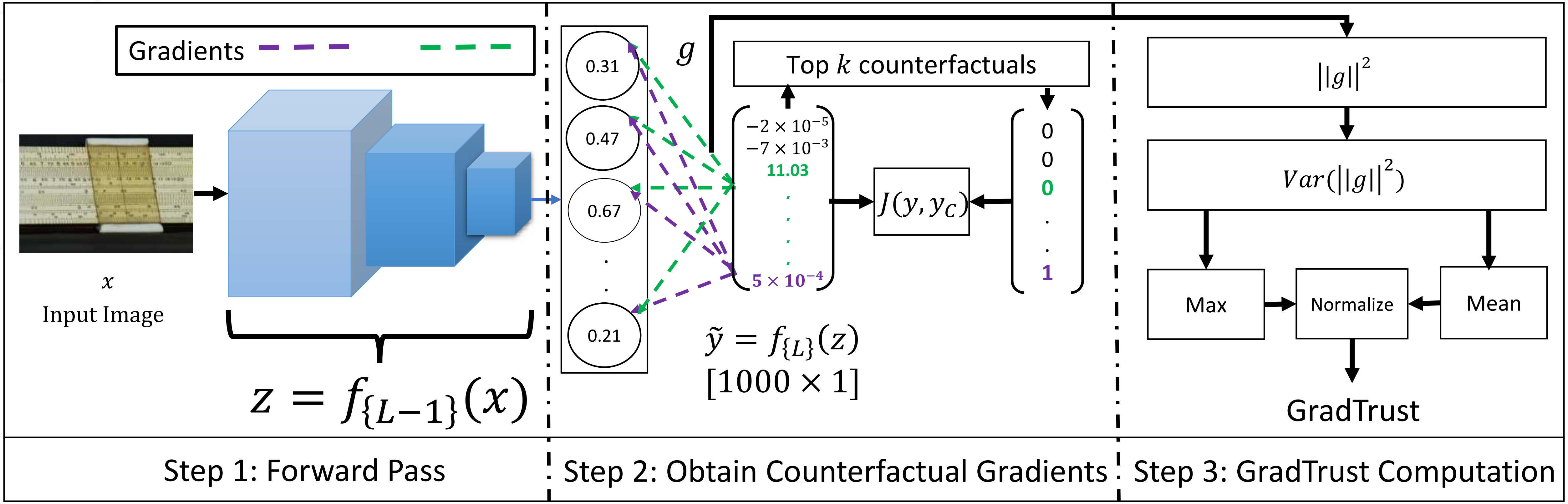
Intuition for counterfactual gradients-based Trust

How do we measure required change? Quantify the variance of network parameters (of the last layer) when backpropagating counterfactual classes

$$GradTrust = \frac{Variance\ of\ Gradients\ of\ Predicted\ Class}{Mean\ of\ Variance\ of\ Gradients\ of\ top - k\ Counterfactual\ Classes}$$

- Top-k counterfactuals are based on predictions
- For image classification, top-k counterfactual classes are top-k predictions
- Gradients are obtained by backpropagating loss between the predicted class and itself in the numerator and between the predicted class and counterfactual classes in denominator

How do we measure required change? Quantify the variance of network parameters when backpropagating counterfactual classes

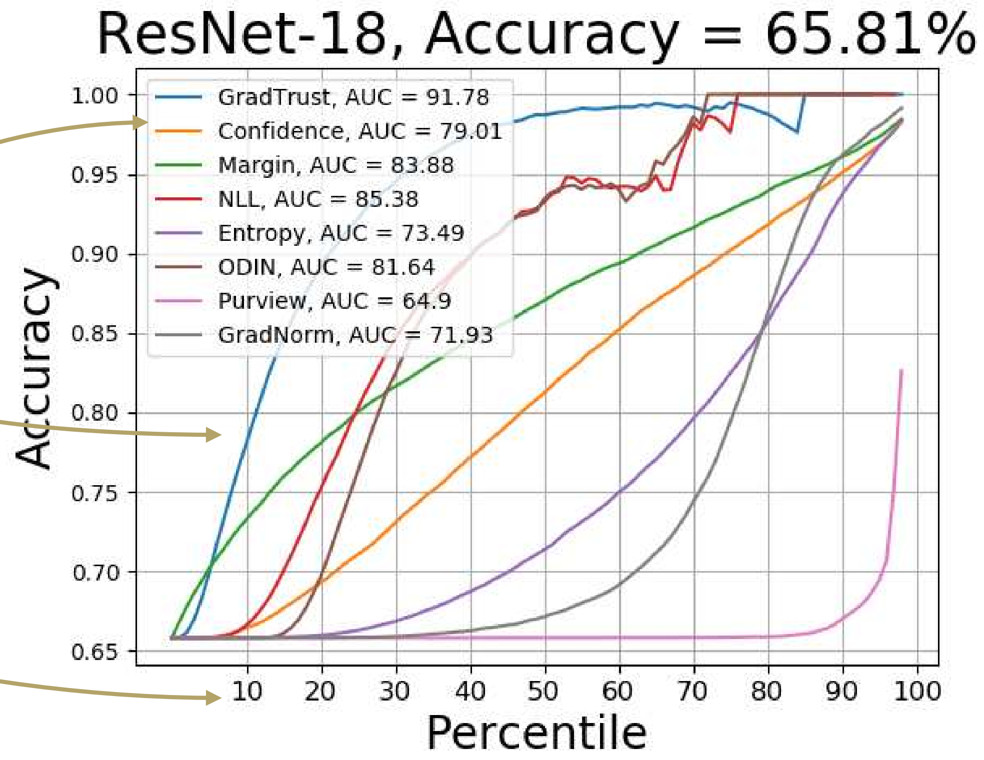


Evaluation

Methodology

For **ImageNet dataset** (with 50,000 validation set images):

1. **Run inference on all 50,000 images** and obtain GradTrust along with comparison trust scores
   - We compare against 8 other methods
2. **For each TrustScore,** order images in **ascending order**
3. For a given x **percentile**, calculate the **Accuracy** and F1 scores of all images above that percentile
4. Plot Area Under Accuracy Curve (AUAC) and Area Under F1 Curve (AUFC)
5. Repeat for multiple networks
   - We perform analysis on 14 ImageNet trained Classification networks and 5 Video Classification networks



Evaluation

Quantitative Results for Image Classification

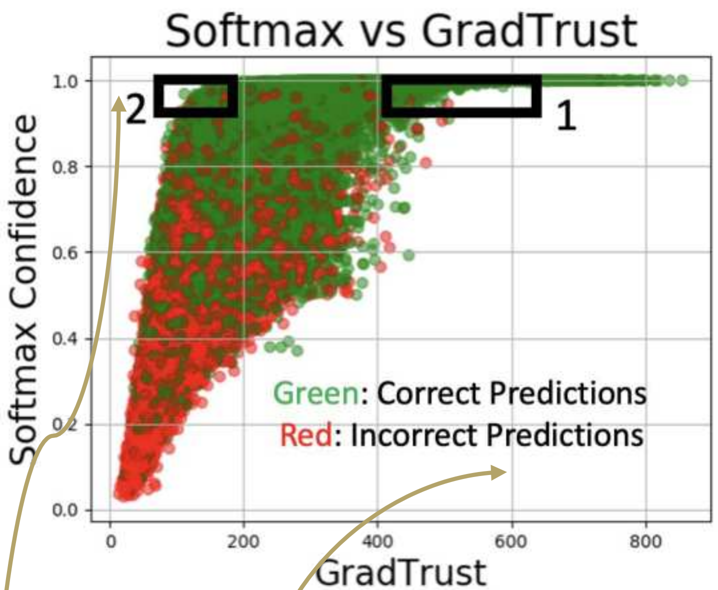
GradTrust is in Top 2 performing metrics in all but 1 network

	AUAC / AUFC								
Architecture	Softmax	Entropy	NLL	Margin [27]	ODIN [28]	MCD [12]	GradNorm [5]	Purview [4]	`GradTrust`
AlexNet [29]	72.86/68.43	65.02/62.14	**83.21/79.37**	79.04/73.3	79.22/75.89	54.2/51.59	58.85/55.28	50.14/48.92	**92.09/89.5**
MobileNet [30]	77.91/74.96	71.72/69.9	**84.02/81.37**	83.13/79.1	75.95/72.81	61.1/59.46	70.3/67.28	61.85/61.32	**93.37/90.58**
ResNet-18 [17]	79.01/76.13	73.49/71.71	**85.38/82.73**	83.88/79.87	81.64/79.26	62.91/61.4	71.93/69.29	64.9/64.01	**91.78/88.65**
VGG-11 [31]	79.95/77.02	74.33/72.52	**90.55/88.42**	84.85/80.77	85.08/83.33	63.19/61.62	73.16/70.06	65/63.84	**91.79/89.18**
ResNet-50 [17]	81.63/79.69	77.47/76.32	**89.23/86.47**	85.7/82.83	84.13/82.21	66.35/65.37	77.37/75.64	71.68/71.01	**92.24/90.09**
ResNeXt-32 [32]	81.56/79.97	78.11/77.15	**89.83/87.37**	85.16/82.81	82.77/80.43	66.9/66.09	78.61/77.28	74.06/73.05	**91.55/89.18**
WideResNet [33]	82.25/80.79	78.96/78.1	**90.84/88.42**	85.76/83.57	84.5/82.26	67.72/66.89	78.62/77.5	74.55/73.85	**91.36/89.12**
Efficient-v2 [34]	**91.49/87.84**	80.12/76.69	71.44/66.03	85.13/81.59	54.16/51.53	81.8/79.38	61.43/57.53	77.79/77.48	**93.57/89.61**
ConvNeXt-t [35]	88.17/86.21	85.56/83.88	79.19/76.85	**90.68/88.26**	62.51/60.74	85.43/83.82	70.86/66.25	79.16/78.91	**89.08/87.23**
ResNeXt-64 [32]	88.95/84.69	85.9/80.71	**90.04/87.06**	**91**/86.62	76.61/72.94	75.3/70.86	73.5/71.64	80.2/79.96	89.15/**87.41**
Swin-v2-t [36]	86.05/84.27	83.79/82.43	86.33/83.14	**88.75/86.29**	79.85/77.09	84.64/83.17	82.23/80.29	77.76/77.39	**87.45/85.23**
VIT-b-16 [37]	85.97/84.38	84.5/82.9	82.94/80.3	**88.67/86.5**	62.74/61,03	84.33/82.81	78.53/74.6	78.02/77.73	**87.77/85.85**
Swin-b [38]	86.18/84.49	84.77/83.14	79.18/75.52	**88.5/86.21**	68.07/64.59	84.69/83.17	83.09/81.52	80.71/80.45	**88.44/86.51**
MaxViT-t [39]	84.08/82.66	79.23/78.21	80.6/78.85	**85.84/84.02**	47.6/46.27	80.07/79.08	70.35/68.12	80.99/80.7	**90.19/88.48**

- **Negative Log Likelihood** (NLL) works well on smaller networks with **less accuracy** while **Margin classifier** works better with **high accuracy** networks
- **GradTrust performs well on all networks**

Evaluation

Qualitative Results for Image Classification



1: High Confidence, High Trust
2: High Confidence, Low Trust

Mispredictions

Correct Predictions

- Results on ResNet-18. **Each point is an image** from ImageNet validation set
- Each image is plot based on its GradTrust on x-axis and Softmax Confidence on y-axis. **Green** color indicates image is **correctly predicted** while **red** color indicates **incorrect prediction**
- **Several incorrect** predictions exist having **low GradTrust but high softmax** confidence (top-left quadrant)
- In contrast, **no incorrect** predictions, with **low Softmax confidence and High GradTrust** (bottom-right quadrant)

Evaluation

Qualitative Results for Image Classification

On AlexNet: Low GradTrust is due to co-occurring classes

On MaxViT: Low GradTrust is due to ambiguity in class resolution

IEEE OLIVES Georgia Tech

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: Negative Interventions
    - Mathematical frameworks to study intervenability
    - Case Study: Intervenability in Interpretability
- Part 5: Conclusions and Future Directions

Intervenability

Through the Causal Glass

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)

Intervenability

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.

Actuate: The amenability of neural network decisions to human interventions

*"The **interaction** between **foundation models and users** via the prompting interface introduces an element of uncertainty, as the **precise response** of these models to user prompts can be **unpredictable**."*

Quesada, Jorge, et al. "PointPrompt: A Multi-modal Prompting Dataset for Segment Anything Model." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

Intervenability

Through the Benchmarking Glass

Verify: The amenability of neural network decisions to human interventions

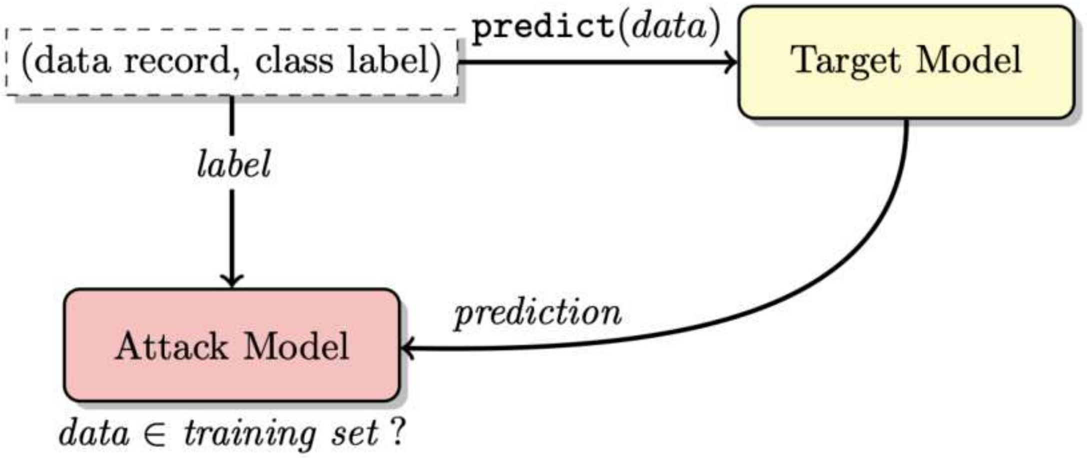
*"… new **benchmarks** were proposed to specifically **test generalization** of classification and detection methods with respect to **simple** algorithmically generated **interventions** like spatial shifts, blur, changes in brightness or contrast…"*

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.

Case Study: Negative Interventions

Repeated Interventions: Membership Inference Attacks (MIAs)

Goal: Given data and black-box model, infer if the data was part of the model's training set



Attack model is the binary classifier

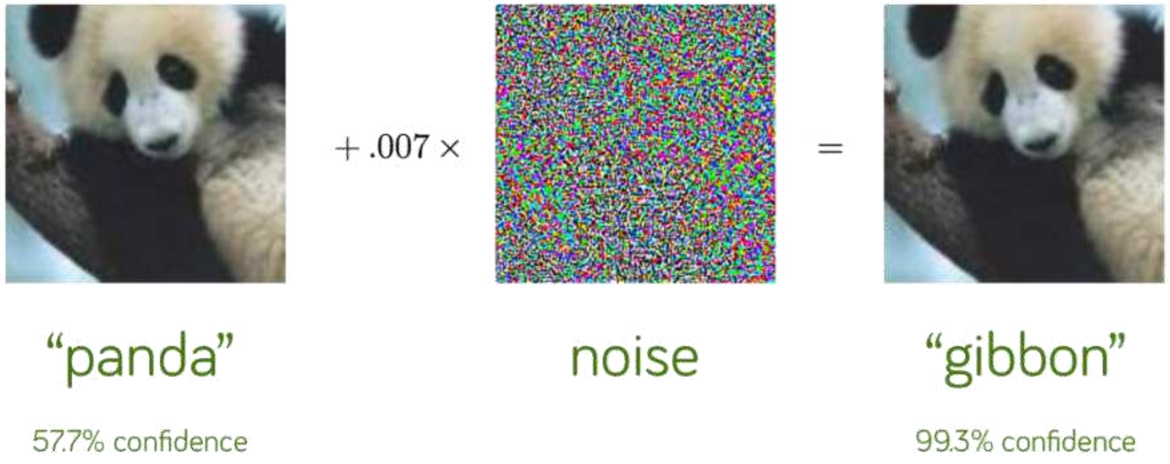
- If data is part of Electronic Health Records, then privacy of patients can be leaked
- Train a binary classifier that takes in the target model outputs and classifies whether the initial data is part of the training set
- **Prevention** is seen as a **robustness** issue while **training**: regularization, adversarial training etc.

Shokri, Reza, et al. "Membership inference attacks against machine learning models." *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017.

Case Study: Negative Interventions

Engineered Interventions: Adversarial Attacks

Goal: Given a trained model, engineer imperceptible noise to 'confuse' the neural network



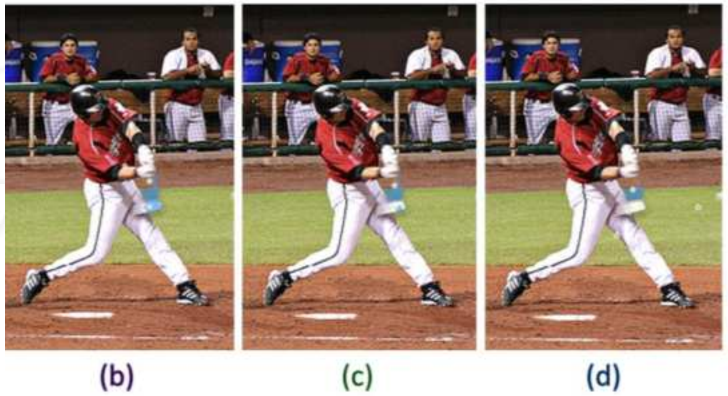
- **Gradients** (or some statistics of gradients) are used in several adversarial image generation techniques
- **Prevention** is seen as a robustness issue **both during inference and training** – adversarial training, image compression etc.

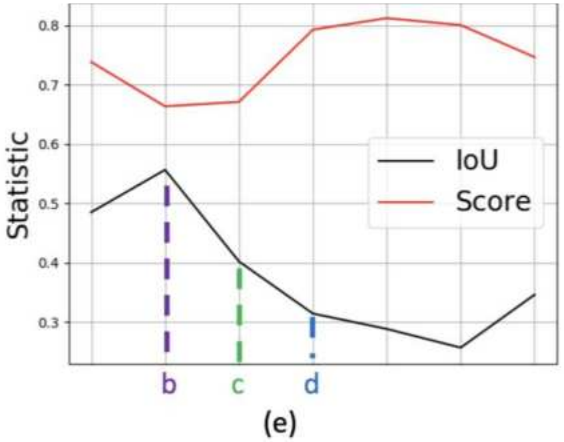
Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

Case Study: Negative Interventions

‘Trial and Error’ Interventions: Visual Prompting

Goal: Given a promptable model with no operational knowledge, users overprompt and use a ‘trial and error’ strategy





- Annotators are asked to segment objects (classes) using Segment Anything Model (SAM) and point prompts
- After prompting, annotators are shown the Intersection Over Union and provided the opportunity to add/subtract their prompt points
- The general conclusion from [1] is that annotators overprompt and utilize strategies that lead to worse performance

- Dataset: https://zenodo.org/records/10975868
- ~200,000 prompts on 6000 images

SCAN ME

[1] Quesada, Jorge, et al. "PointPrompt: A Multi-modal Prompting Dataset for Segment Anything Model." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

Objective

Objective of the Tutorial

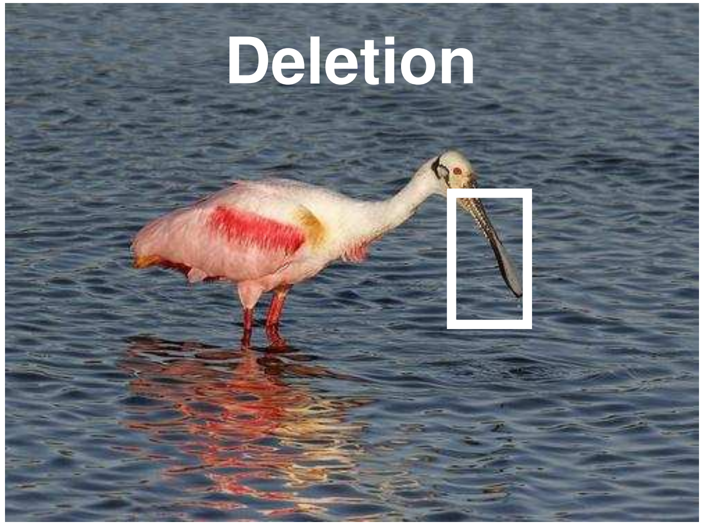
To discuss methodologies that promote robustness in neural networks at inference

- Part 1: Inference in Neural Networks
- Part 2: Explainability at Inference
- Part 3: Uncertainty at Inference
- **Part 4: Intervenability at Inference**
    - Definitions of Intervenability
    - Mathematical frameworks to study intervenability
        - Causal analysis via interventions
        - Dangers of incomplete interventions
    - Case Study: Intervenability in Interpretability
- Part 5: Conclusions and Future Directions

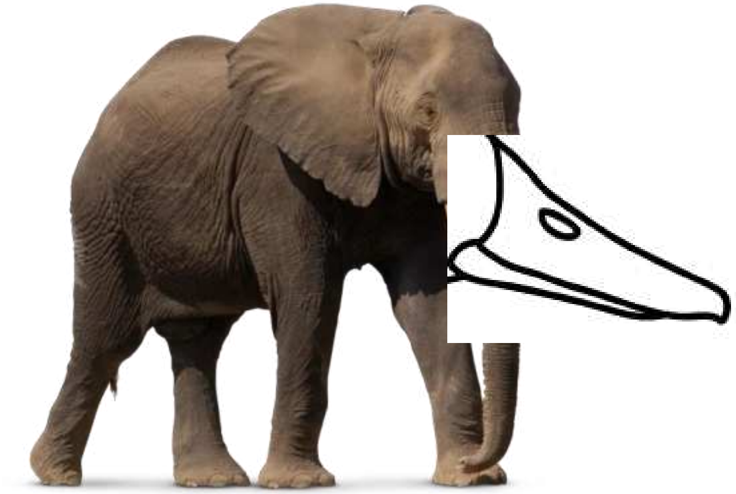
3 Rules of Causal Inference

Rule 1 (Insertion/deletion of observations):

$$P(y|do(x), z, w) = P(y|do(x), w)$$



Insertion



- Fix a causal feature (or a feature that is being tested for causality) in the data

Key Differences:

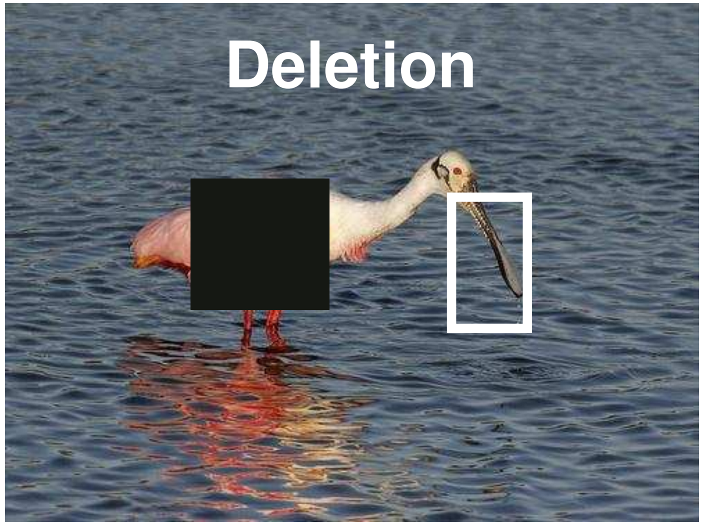
- There are **no causal features;** approximate using pixels/structures
- The underlying network is **not a structured causal model**

Pearl, Judea. "The do-calculus revisited." *arXiv preprint arXiv:1210.4852* (2012).

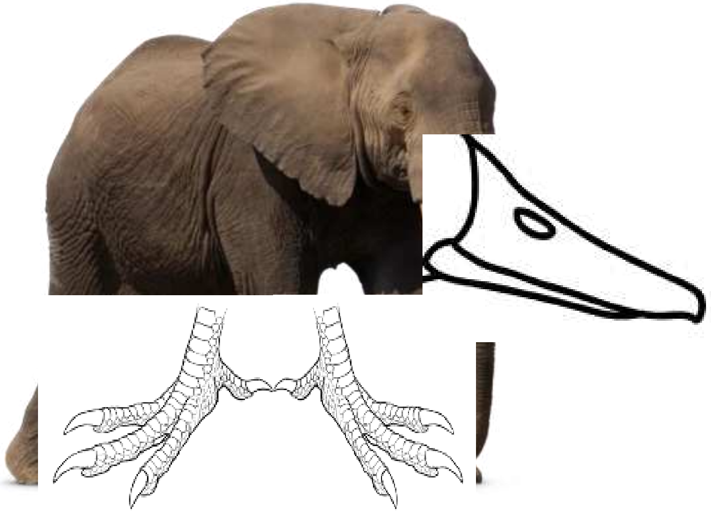
Rule 2: Intervene on all other factors keeping the causal factor constant

Rule 2 (Action/observation exchange):

$$P(y|do(x), do(z), w) = P(y|do(x), z, w)$$



Insertion



- Keeping the causal factor constant from rule 1, change all available factors

Key Differences:

- There are **no causal features;** approximate using pixels/structures
- The underlying network is **not a structured causal model**
- **Impossible** to intervene on all pixels

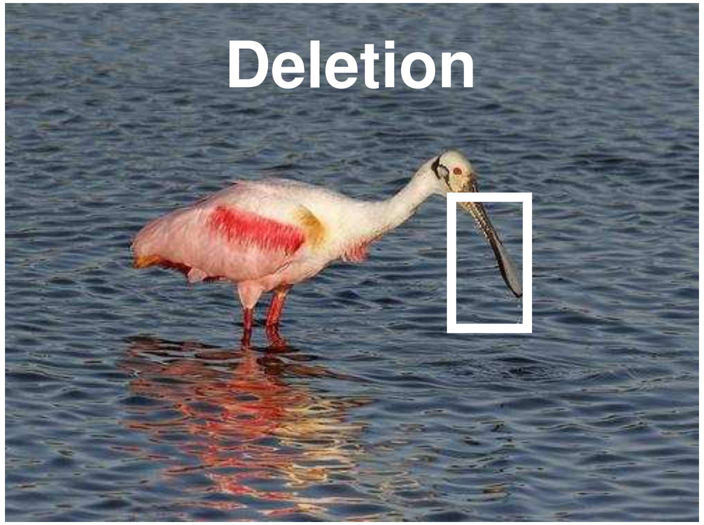
Pearl, Judea. "The do-calculus revisited." *arXiv preprint arXiv:1210.4852* (2012).

Rule 3: Insertion/Deletion of interventional actions

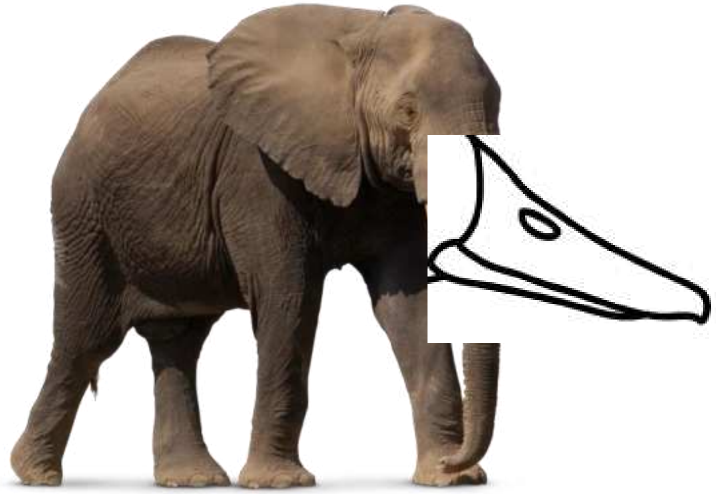
Rule 3 (Insertion/deletion of actions):

$$P(y|do(x), do(z), w) = P(y|do(x), w)$$

Once causal factors are determined, the interventions from rule 2 are reverted and the causal attribution is noted



Insertion



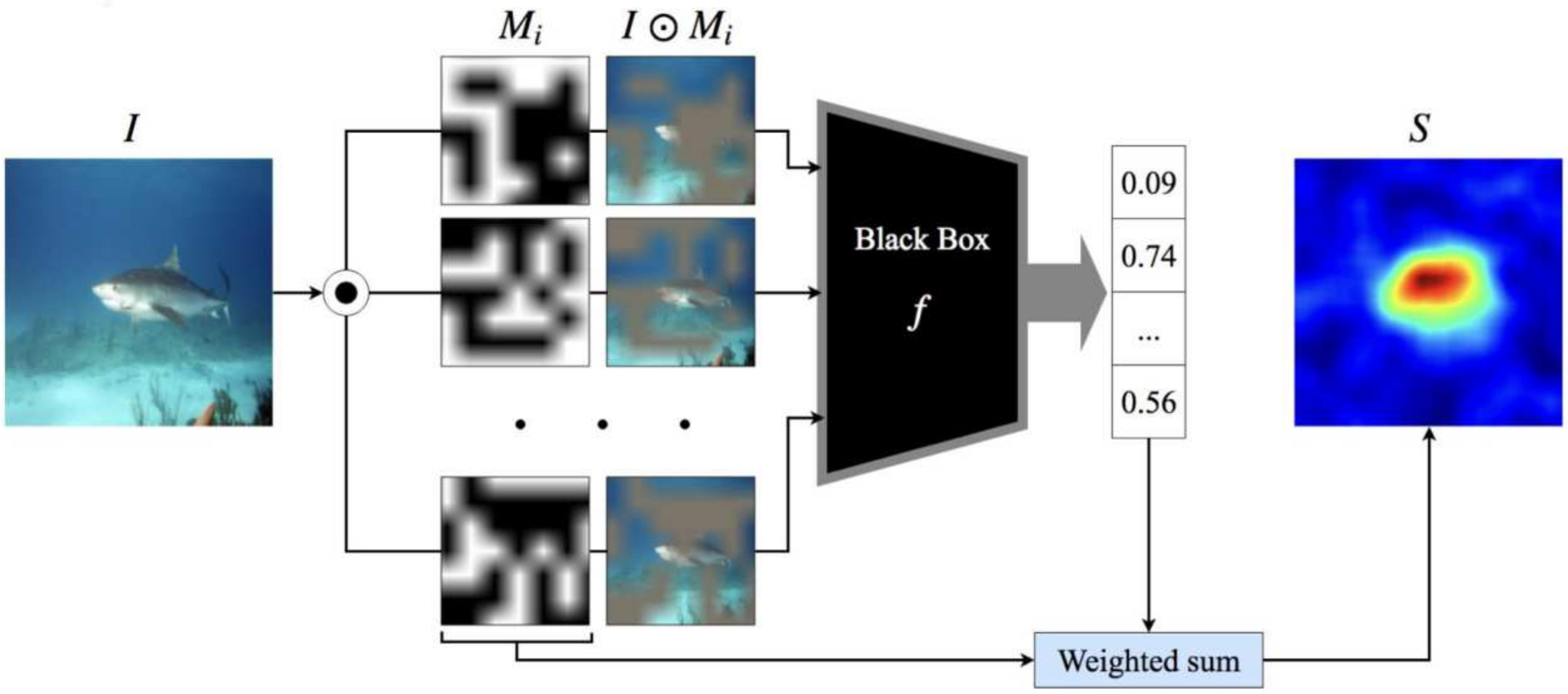
Key Differences:

- There are **no causal features;** approximate using pixels/structures
- The underlying network is **not a structured causal model**
- **Impossible** to intervene on all pixels

Pearl, Judea. "The do-calculus revisited." *arXiv preprint arXiv:1210.4852* (2012).

Dangers of Incomplete Interventions: RISE Explanations

Unknown interventions based on insertion/deletion can yield unexpected results



- **RISE** explainability technique creates **6000 random masks** for an image and passes it through a network
- The weighted sum of the **mask** and its **probability score** is the explanation

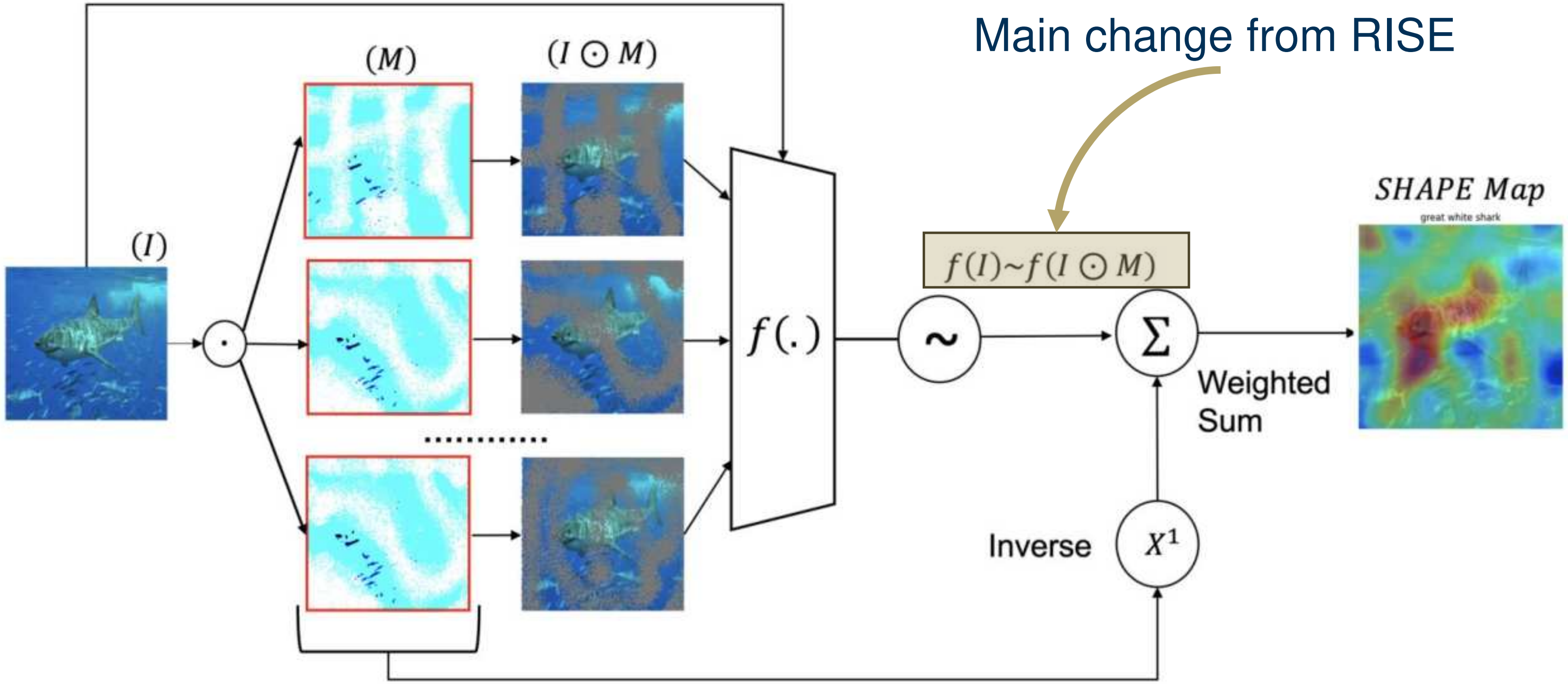
- Instead of causal deletion, RISE deletes randomly

Petsiuk, Vitali, Abir Das, and Kate Saenko. "Rise: Randomized input sampling for explanation of black-box models." *arXiv preprint arXiv:1806.07421* (2018).

Intervenability Frameworks

Dangers of Incomplete Interventions: SHAPE Explanations

Unknown interventions based on insertion/deletion can yield unexpected results



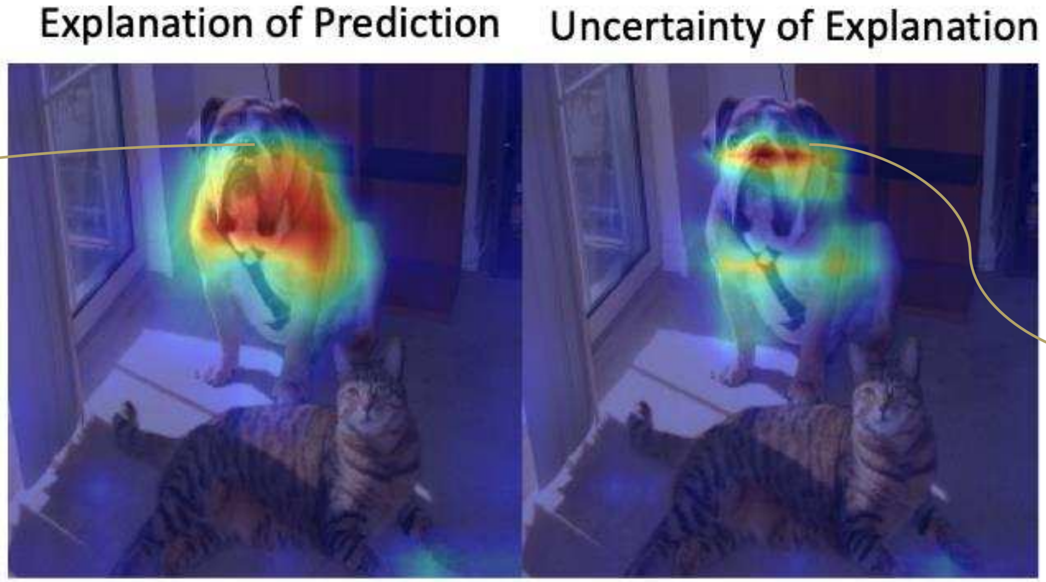
- **SHAPE** explanation is almost identical to RISE except:
  - Weighted sum is **NOT** between probability and mask but between ***change in probability score*** and inverse mask
- Results are human un-interpretable
- **However, existing objective evaluation metrics give better scores to SHAPE than RISE**

Chowdhury, Prithwijit, et al. "Are Objective Explanatory Evaluation metrics Trustworthy? An Adversarial Analysis." *arXiv preprint arXiv:2406.07820* (2024).

Intervenability Frameworks

Framework 2: Predictive Uncertainty in Interventions

Accept that all interventions are impossible and calculate the uncertainty of ‘residual’ interventions



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

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.

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
  - Mathematical frameworks to study intervenability
  - Case Study: Intervenability in Interpretability
    - Motivating explanatory evaluation
    - VOICE: Variance of Induced Contrastive Explanations
- Part 5: Conclusions and Future Directions

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

Mohit Prabhushankar, PhD
Postdoc

Ghassan AlRegib, PhD
Professor

SCAN ME

Case Study: Intervenability in Interpretability

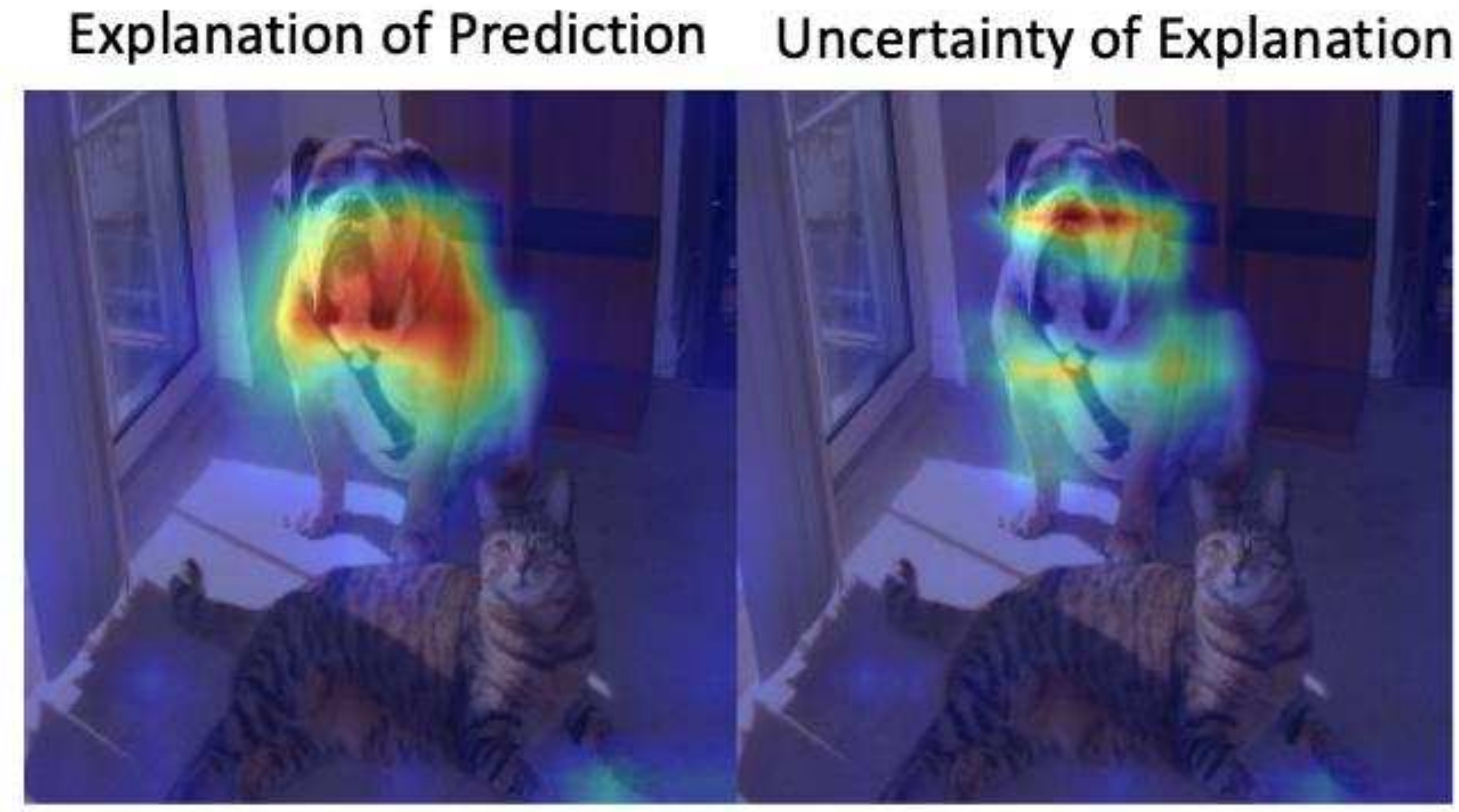
Predictive Uncertainty in Explanations

SCAN ME

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

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 (J-STSP) Special Series on AI in Signal & Data Science, May 23, 2024.

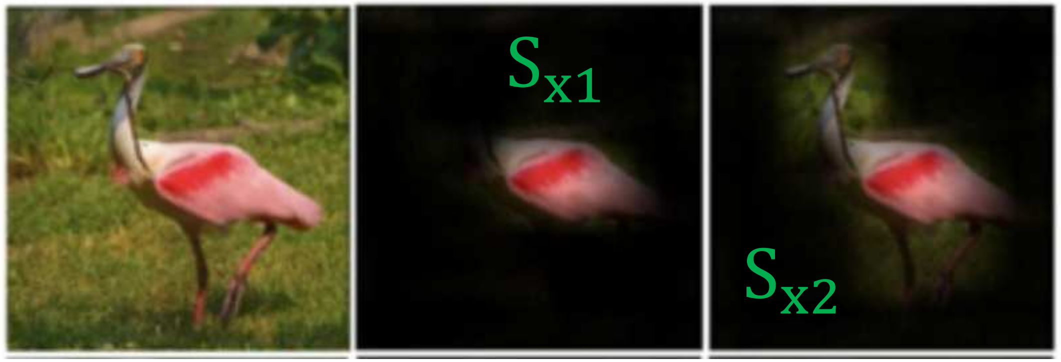
Case Study: Intervenability in Interpretability

Explanation Evaluation via Masking

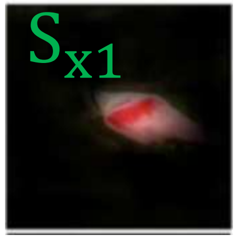
Common evaluation technique is masking the image and checking for prediction correctness

y = Prediction
S_x = Explanation masked data

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



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



→ Trained Model → Crane

→ Trained Model → Spoonbill

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.

Case Study: Intervenability in Interpretability

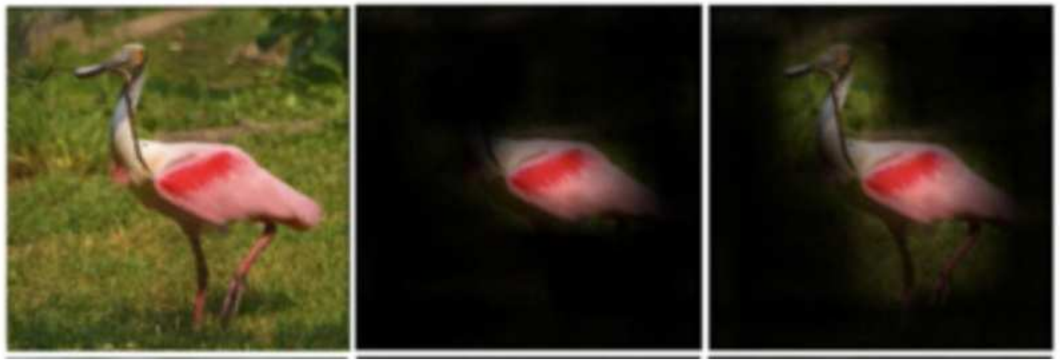
Predictive Uncertainty

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

SCAN ME

Uncertainty due to variance in prediction when model is kept constant

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

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 (J-STSP) Special Series on AI in Signal & Data Science, May 23, 2024.

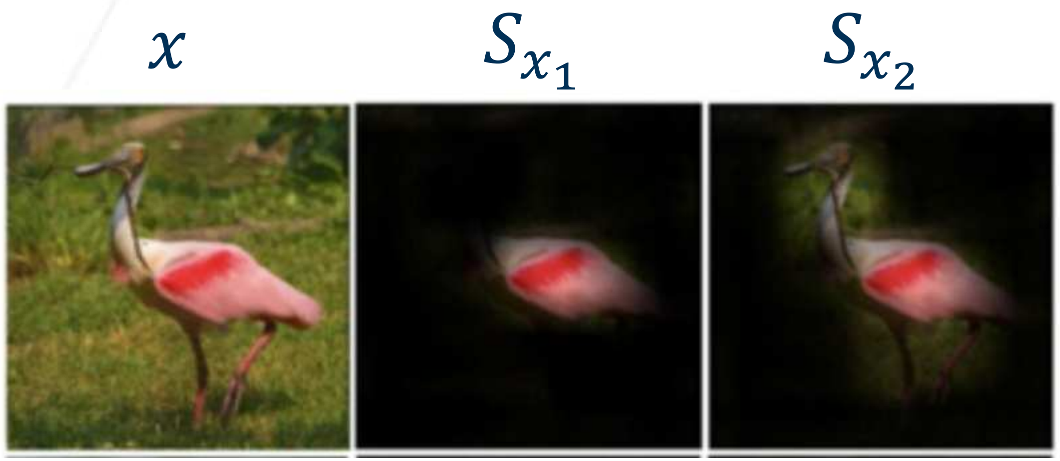
Case Study: Intervenability in Interpretability

Visual Explanations (partially) reduce Predictive Uncertainty

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

SCAN ME

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!

IEEE
[Tutorial@MIPR'24] | [Ghassan AlRegib and Mohit Prabhushankar] | [Aug 9, 2024]

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 (J-STSP) Special Series on AI in Signal & Data Science, May 23, 2024.

OLIVES @GeorgiaTech
Georgia Tech

Case Study: Intervenability in Interpretability

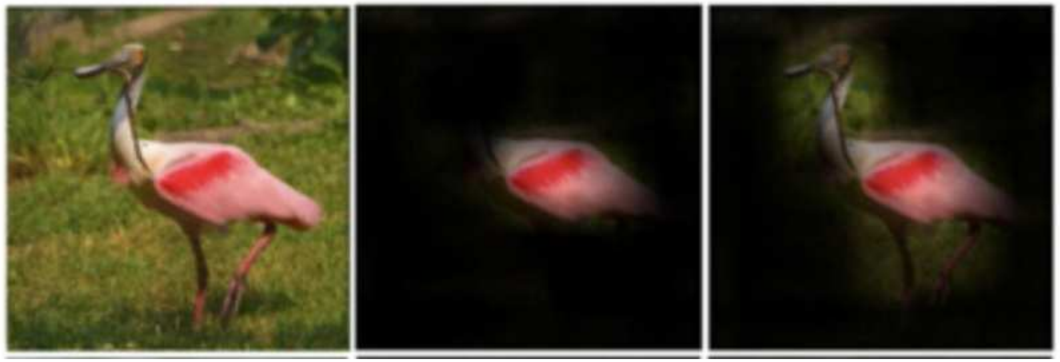
Predictive Uncertainty in Explanations is the Residual

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

SCAN ME

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 (J-STSP) Special Series on AI in Signal & Data Science, May 23, 2024.

SCAN ME

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

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

x $\quad S_{x_1}$ $\quad S_{x_2}$

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

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

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 (J-STSP) Special Series on AI in Signal & Data Science, May 23, 2024.

Case Study: Intervenability in Interpretability

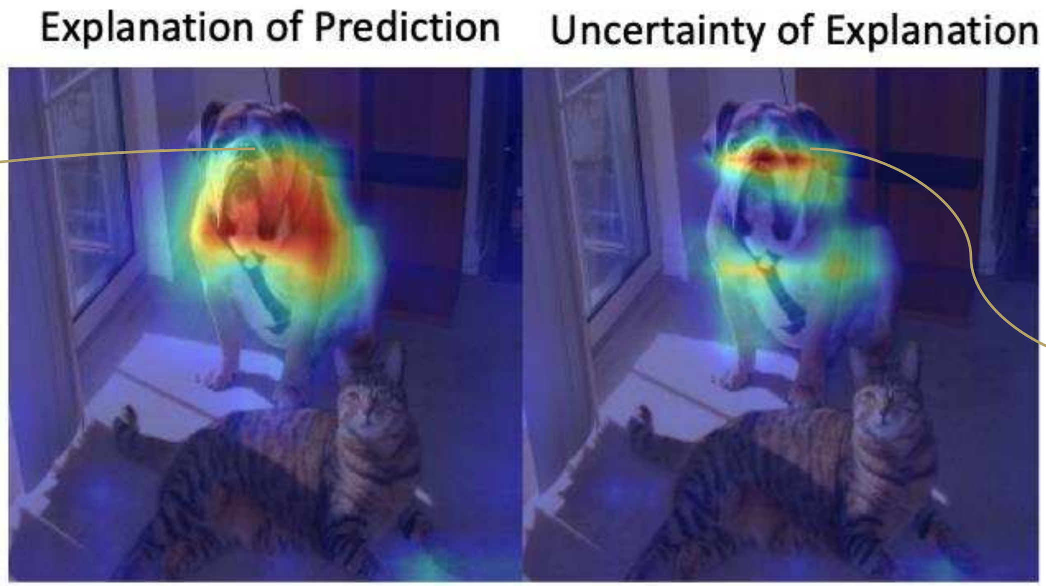
Predictive Uncertainty in Explanations is the Residual

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

SCAN ME

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 (J-STSP) Special Series on AI in Signal & Data Science, May 23, 2024.

Case Study: Intervenability in Interpretability

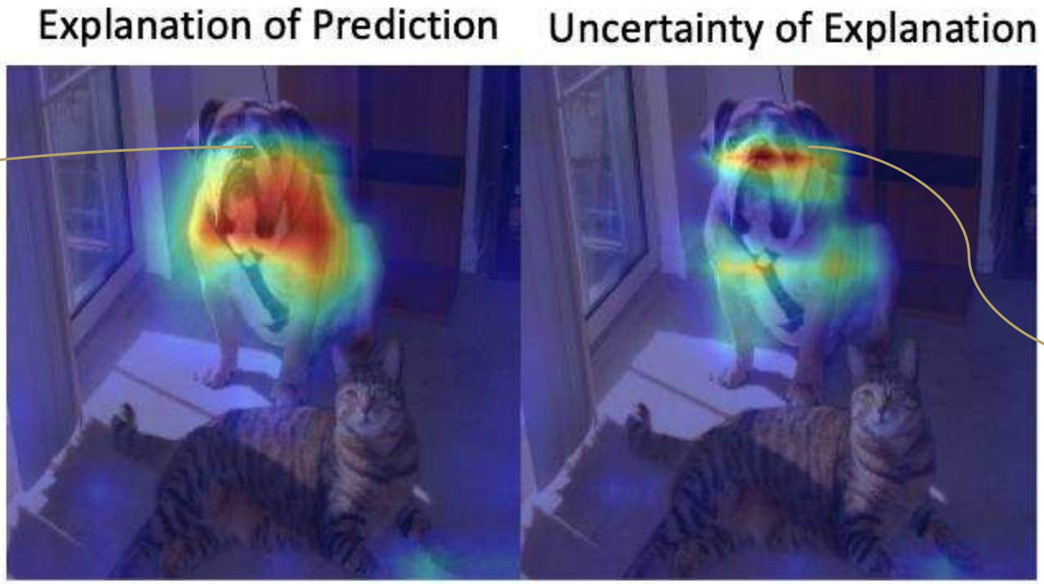
Predictive Uncertainty in Explanations is the Residual

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

SCAN ME

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 (J-STSP) Special Series on AI in Signal & Data Science, May 23, 2024.

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

SCAN ME

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

$$\left.\begin{matrix} Y_1|S_{x1} \\ Y_2|S_{x2} \\ Y_3|S_{x3} \\ Y_4|S_{x4} \\ Y_5|S_{x5} \\ . \\ . \\ Y_N|S_{xN} \end{matrix}\right\} \text{Variance}$$

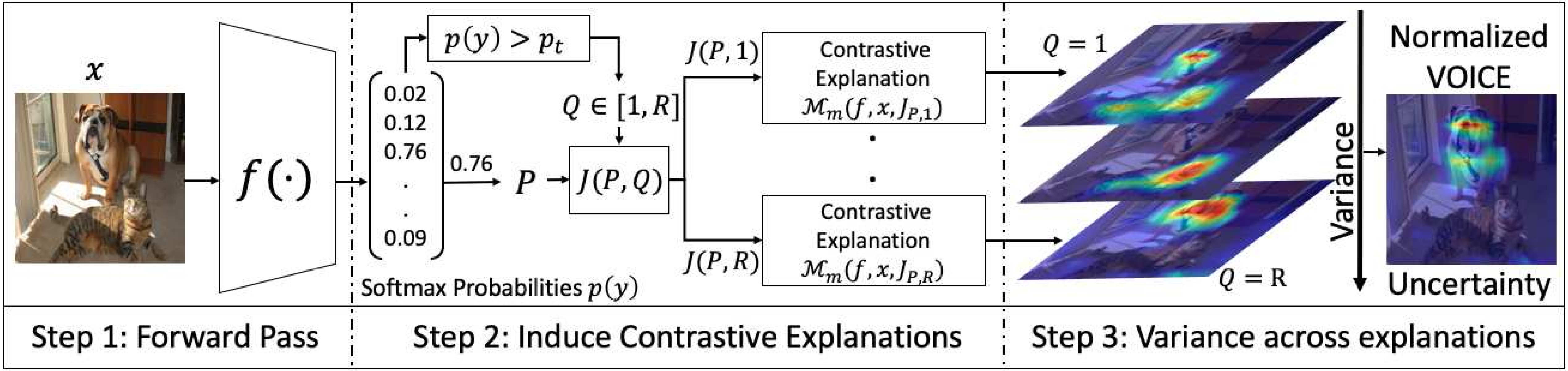
M. Prabhushankar and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," Journal of Selected Topics in Signal Processing (J-STSP) Special Series on AI in Signal & Data Science, May 23, 2024.

Quantifying Uncertainty in Explainability

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

SCAN ME

Variance in contrastive explanations provides 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 (J-STSP) Special Series on AI in Signal & Data Science, May 23, 2024.

SCAN ME

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

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 (J-STSP) Special Series on AI in Signal & Data Science, May 23, 2024.

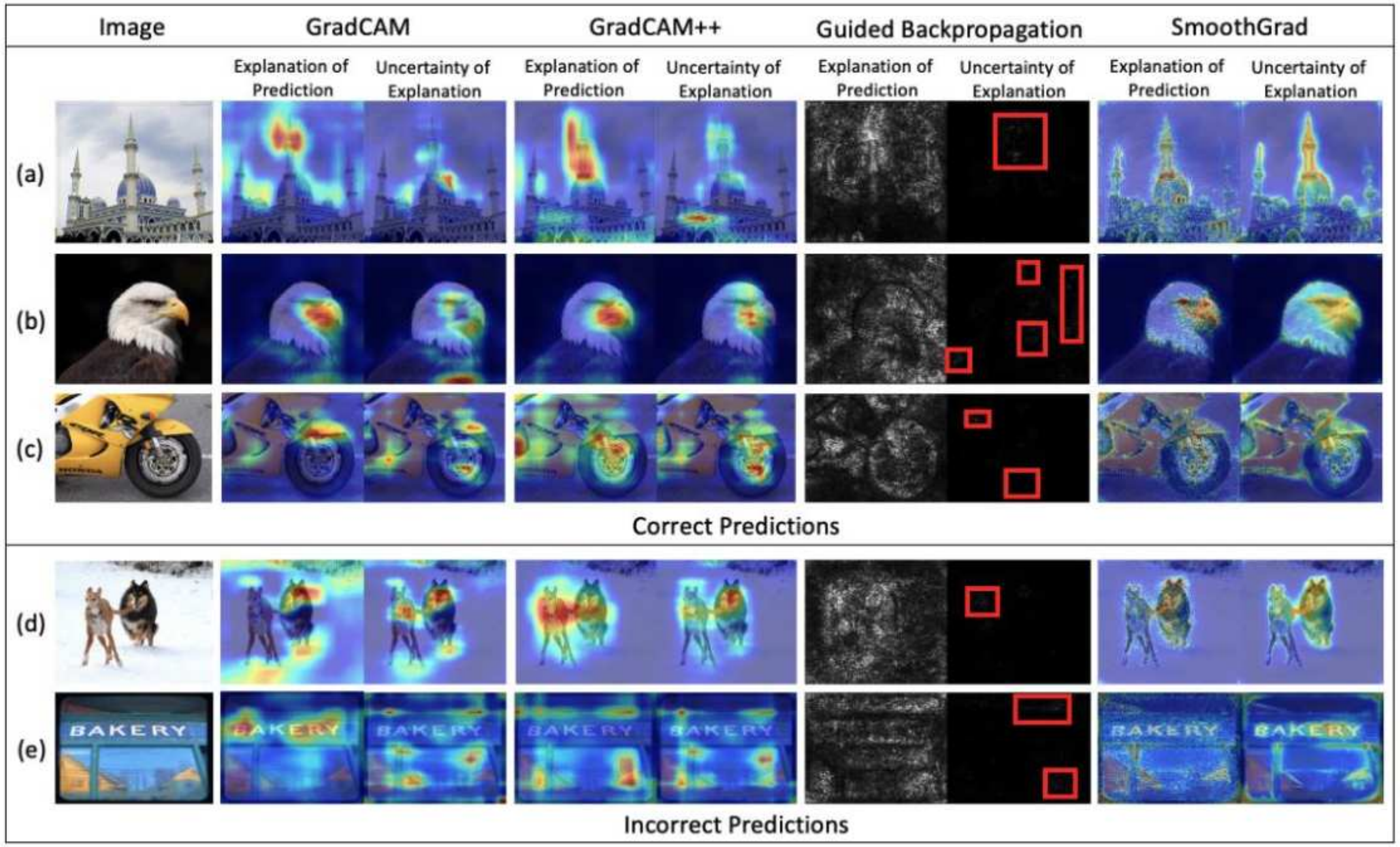
Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability: mIOU

SCAN ME

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

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)

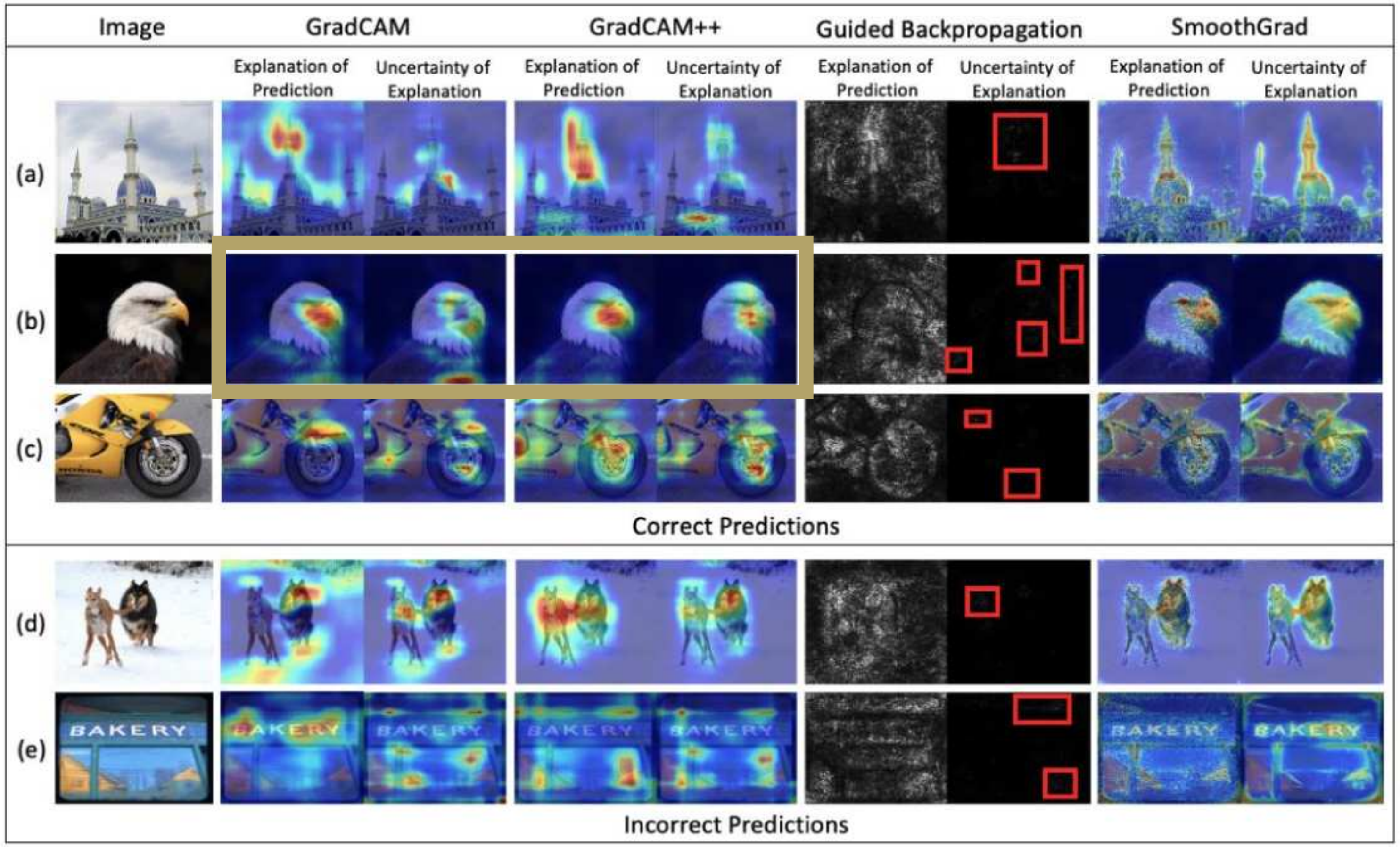
M. Prabhushankar and G. AlRegib, ”VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability,” Journal of Selected Topics in Signal Processing (J-STSP) Special Series on AI in Signal & Data Science, May 23, 2024.

Quantifying Interventions in Explainability: mIOU

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

SCAN ME

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 (J-STSP) Special Series on AI in Signal & Data Science, May 23, 2024.

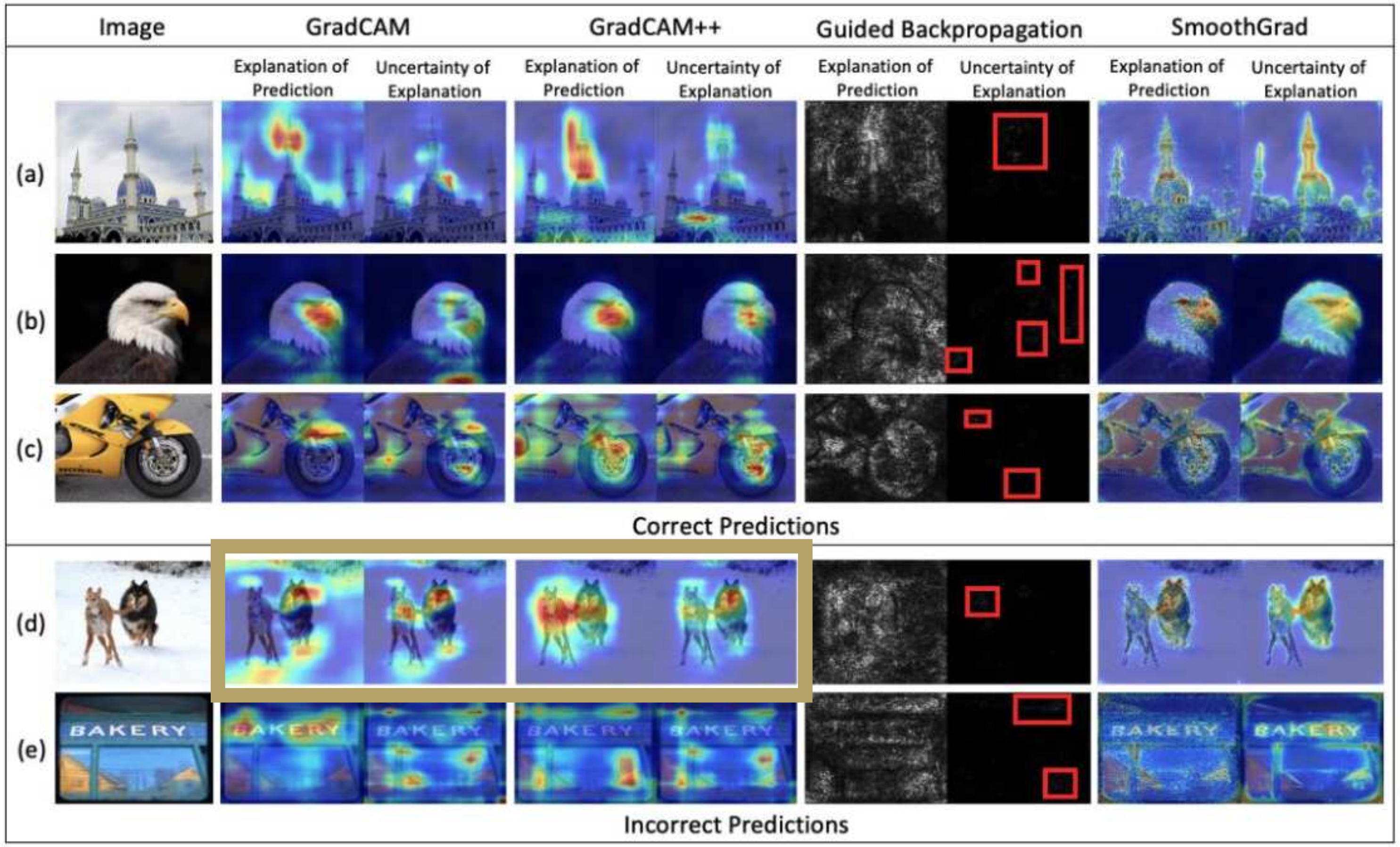
Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability: mIOU

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

SCAN ME

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)

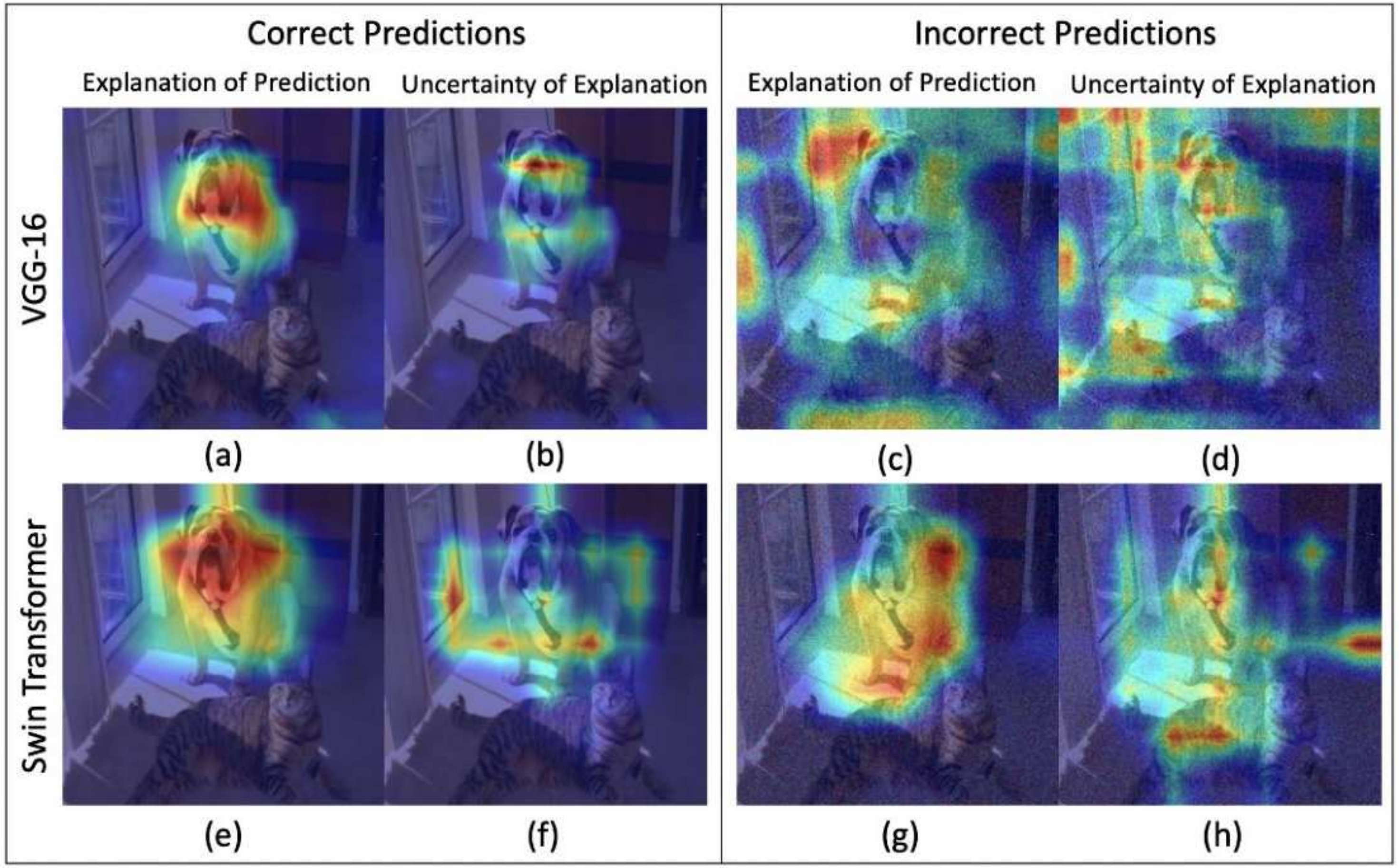
M. Prabhushankar and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," Journal of Selected Topics in Signal Processing (J-STSP) Special Series on AI in Signal & Data Science, May 23, 2024.

Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability: SNR

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

Explanation and uncertainty are dispersed under noise (under low prediction confidence)



Objective Metric 2: Signal to Noise Ratio of the Uncertainty map

Higher the SNR of uncertainty, more is the dispersal (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 (J-STSP) Special Series on AI in Signal & Data Science, May 23, 2024.

Case Study: Intervenability in Interpretability

Challenges in Intervenability

The amenability of neural network decisions to human interventions

- **Not choosing interventions** causes **uncertainty** in the chosen interventions
- **Residuals** must be **analyzed** intelligently to **‘trust or not to trust’** predictions at inference
- Gradients quantify residual uncertainty

Challenges:
- Choosing the type of Intervention
- **Residuals of Interventions: Uncertainty**

Intervenability

Through the Human Glass

The amenability of neural network decisions to human interventions

- **Assess: Causality**
- **Assure: Privacy**
- **Interpret: Interpretability**
- **Actuate: Prompting**
- **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.

CURE-TSD: Challenging Unreal and Real Environments for Traffic Sign Detection

Data Characteristics:

- 49 real and virtual sequences
- 300 frames in each sequence
- 12 different challenges including decolorization, codec error, lens blur etc.
- 5 progressively increasingly levels in each challenge
- **Goal**: Detect and localize traffic signs

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

CURE-TSR: Challenging Unreal and Real Environments for Traffic Sign Recognition

Data Characteristics:

- 2 million real and virtual traffic sign images
- 14 Traffic signs including common signs like stop, no-right, no-left etc. and uncommon signs like goods-vehicles, priority lanes etc.
- 12 different challenges including decolorization, codec error, lens blur etc.
- 5 progressively increasingly levels in each challenge

D. Temel, G. Kwon*, M. Prabhushankar*, and G. AlRegib, “CURE-TSR: Challenging unreal and real environments for traffic sign recognition,” in *Neural Information Processing Systems (NIPS) Workshop on Machine Learning for Intelligent Transportation Systems,* Long Beach, U.S., December 2017, (*: equal contribution)

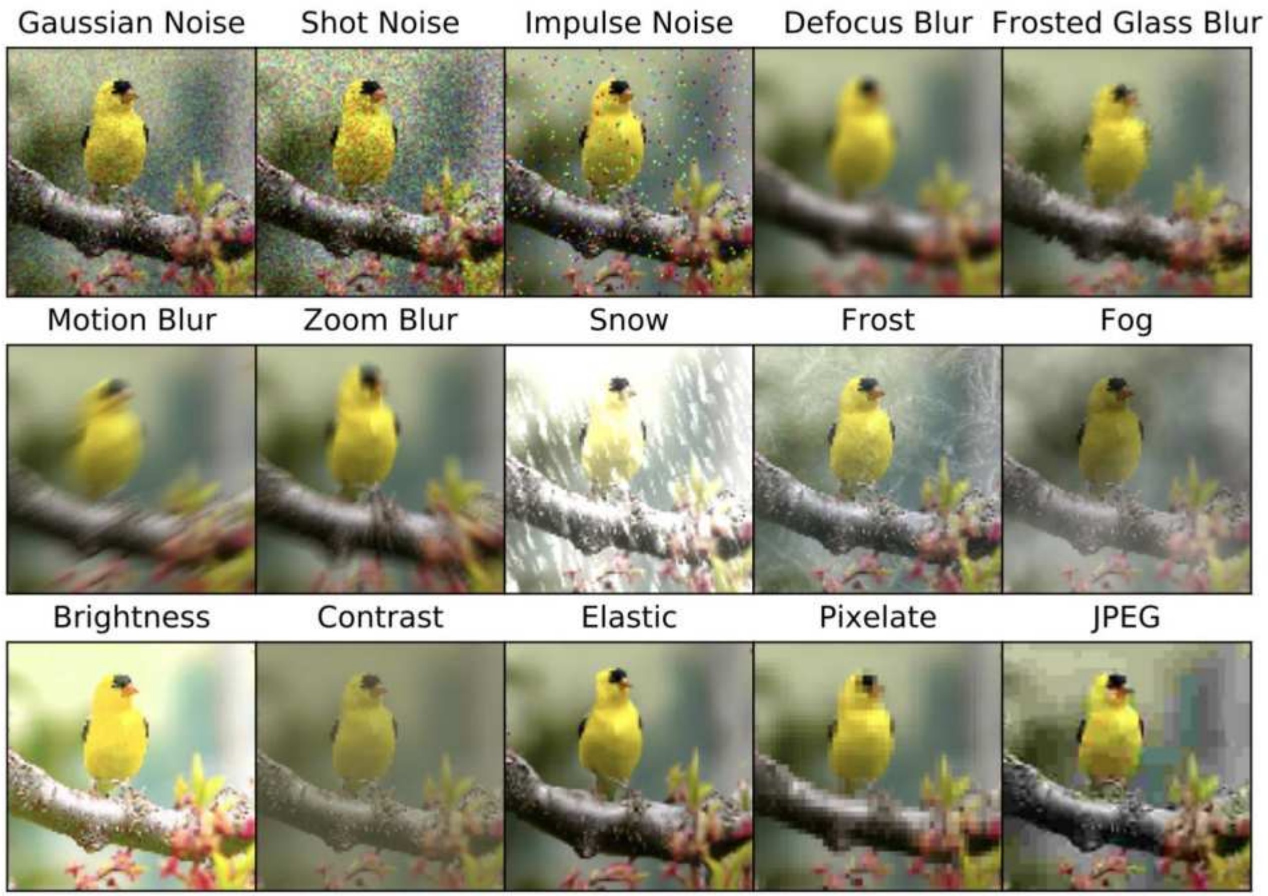
Intervenability in Benchmarking

Recognition

ImageNet-C: ImageNet-Corruptions

Data Characteristics:

- 3.75 million images
- 15 different challenges including decolorization, codec error, lens blur etc. for testing
- 4 different challenges for validation and training
- 5 progressively increasingly levels in each challenge
- **Goal**: Recognize 1000 classes from ImageNet using pretrained networks



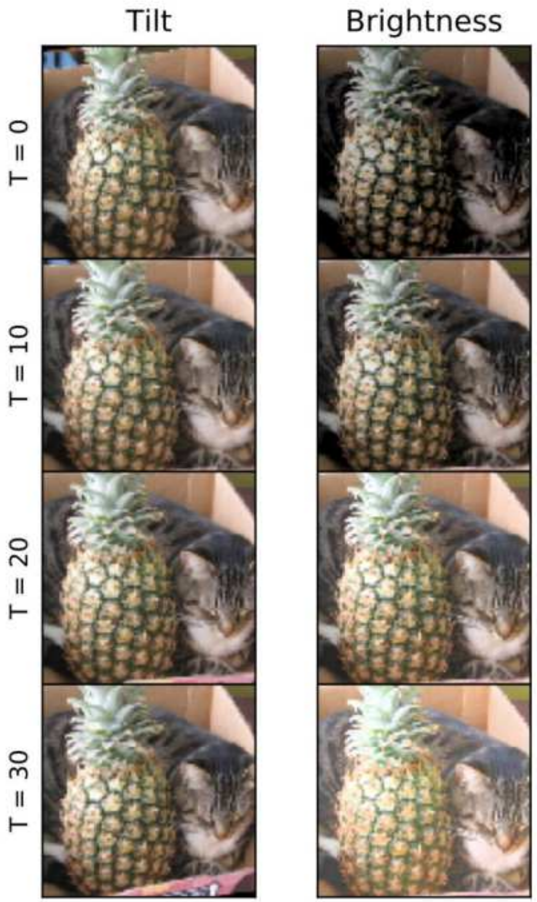
Hendrycks, Dan, and Thomas Dietterich. "Benchmarking neural network robustness to common corruptions and perturbations." *arXiv preprint arXiv:1903.12261* (2019).

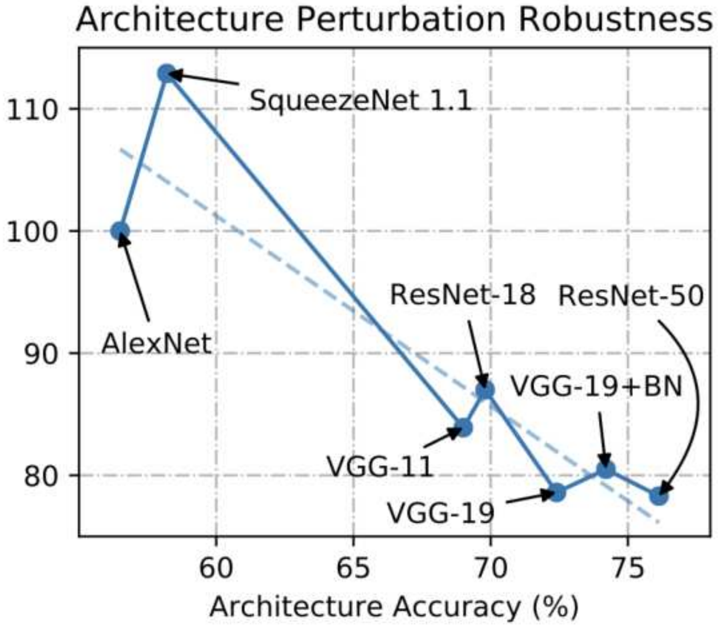
Intervenability in Benchmarking
Recognition

ImageNet-P: ImageNet-Perturbations

Data Characteristics:

- 5 million images
- 100 perturbations of 50000 images
- 10 frames of algorithmically generated perturbations for each image in ImageNet validation testset
- 10 common perturbations including brightness, tilt, motion etc.





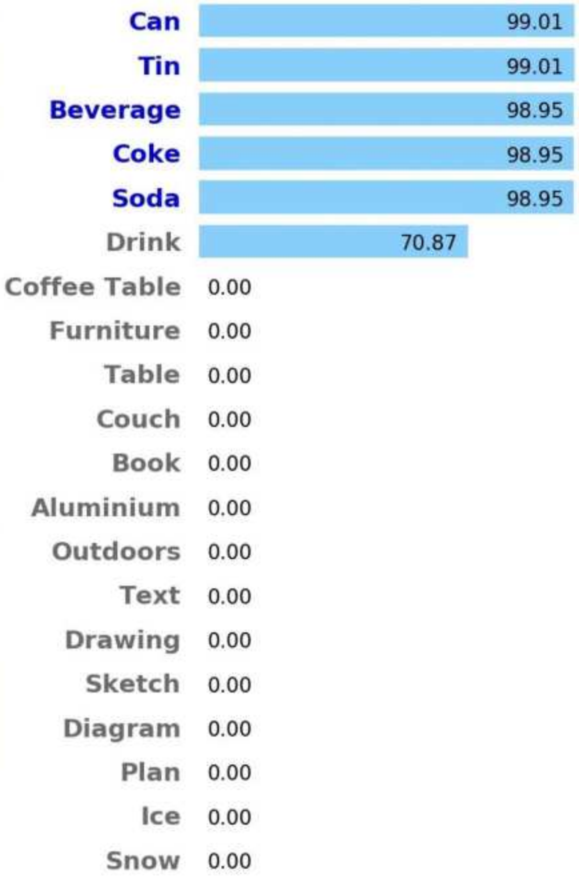
Hendrycks, Dan, and Thomas Dietterich. "Benchmarking neural network robustness to common corruptions and perturbations." *arXiv preprint arXiv:1903.12261* (2019).

CURE-OR: Challenging Unreal and Real Environments for Object Recognition

Data Characteristics:

- 1 million images
- 100 common household objects and 10000 images per object
- 5 backgrounds, 5 object orientations, 5 devices, and 78 challenging conditions
- **Goal**: To recognize and retrieve the same object across backgrounds, orientations, devices, and challenging conditions

Challenge Type: **None**

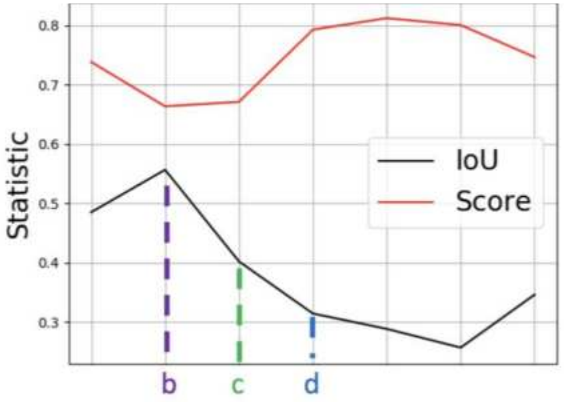


D. Temel*, J. Lee*, and G. AlRegib, “CURE-OR: Challenging unreal and real environments for object recognition,” IEEE International Conference on Machine Learning and Applications, Orlando, Florida, USA, December 2018, (*: equal contribution)

Intervenability in Benchmarking

Prompting

PointPrompt: A Multi-modal Prompting Dataset for Segment Anything Model



(e)

- Annotators are asked to segment objects (classes) using Segment Anything Model (SAM) and point prompts
- After prompting, annotators are shown the Intersection Over Union and provided the opportunity to add/subtract their prompt points
- The general conclusion from [1] is that annotators overprompt and utilize strategies that lead to worse performance

- Dataset: https://zenodo.org/records/10975868
- ~200,000 prompts on 6000 images

SCAN ME

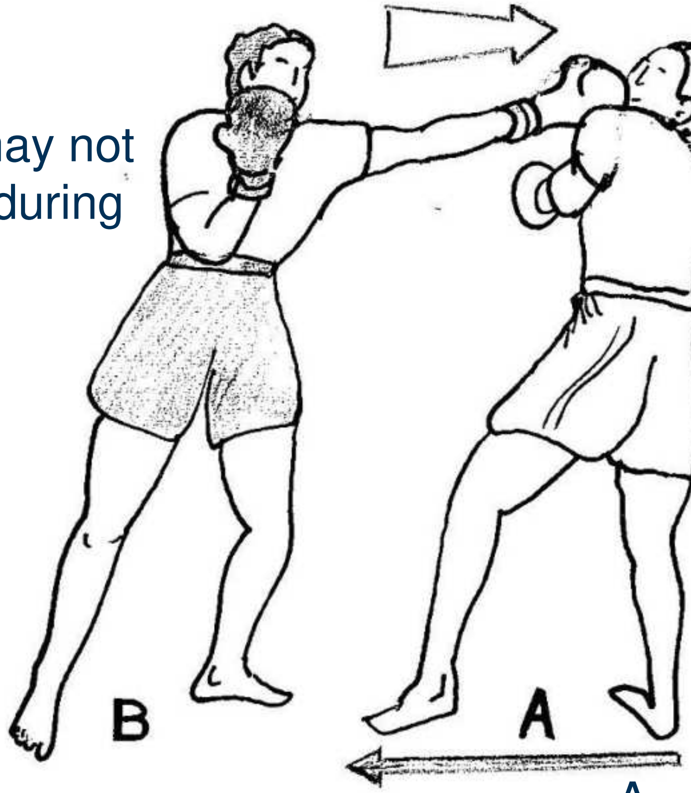
Quesada, Jorge, et al. "PointPrompt: A Multi-modal Prompting Dataset for Segment Anything Model." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

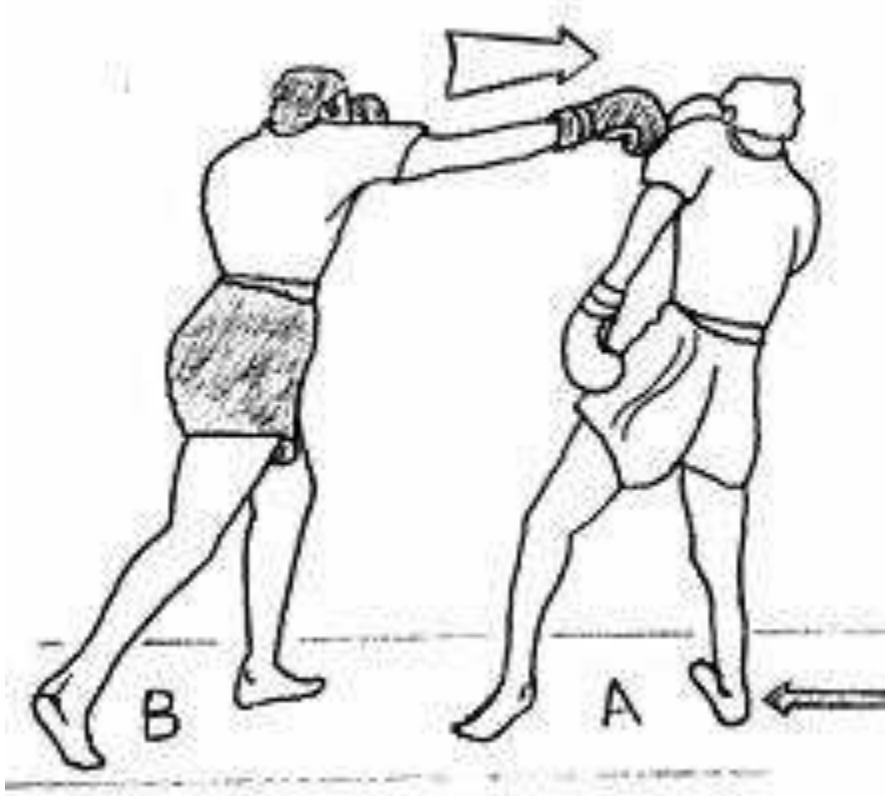
Robust Neural Networks
Part 5: Conclusions and Future Directions

Memes to Wrap it Up
Overcoming Challenges at Training

Novel data packs a 1-2 punch!

Novel data may not be available during training



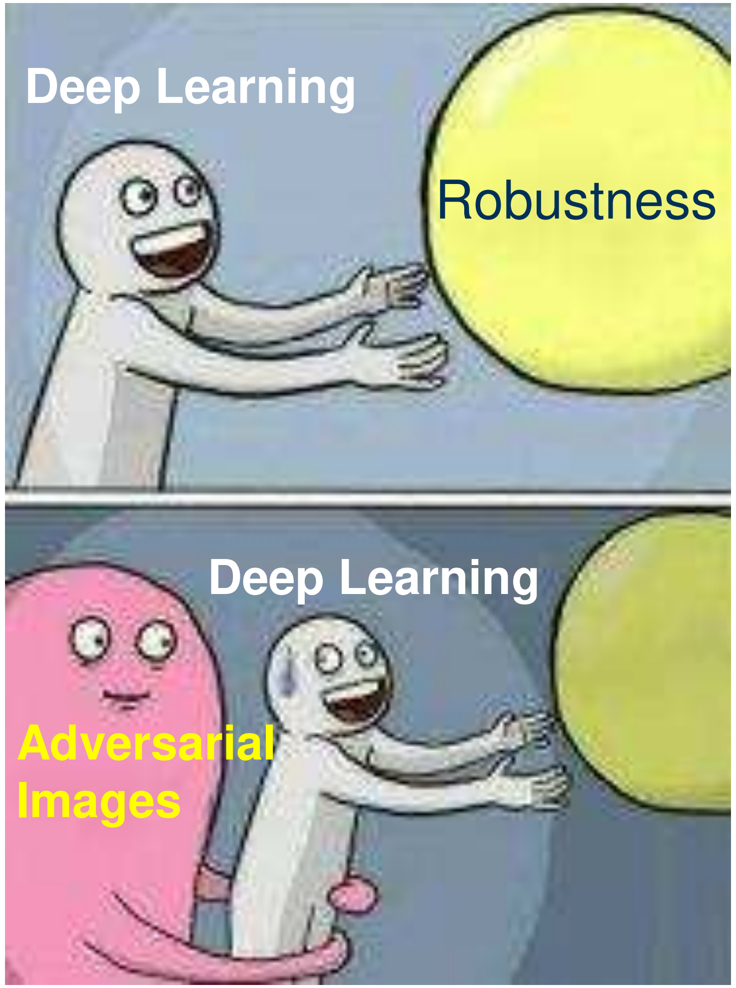


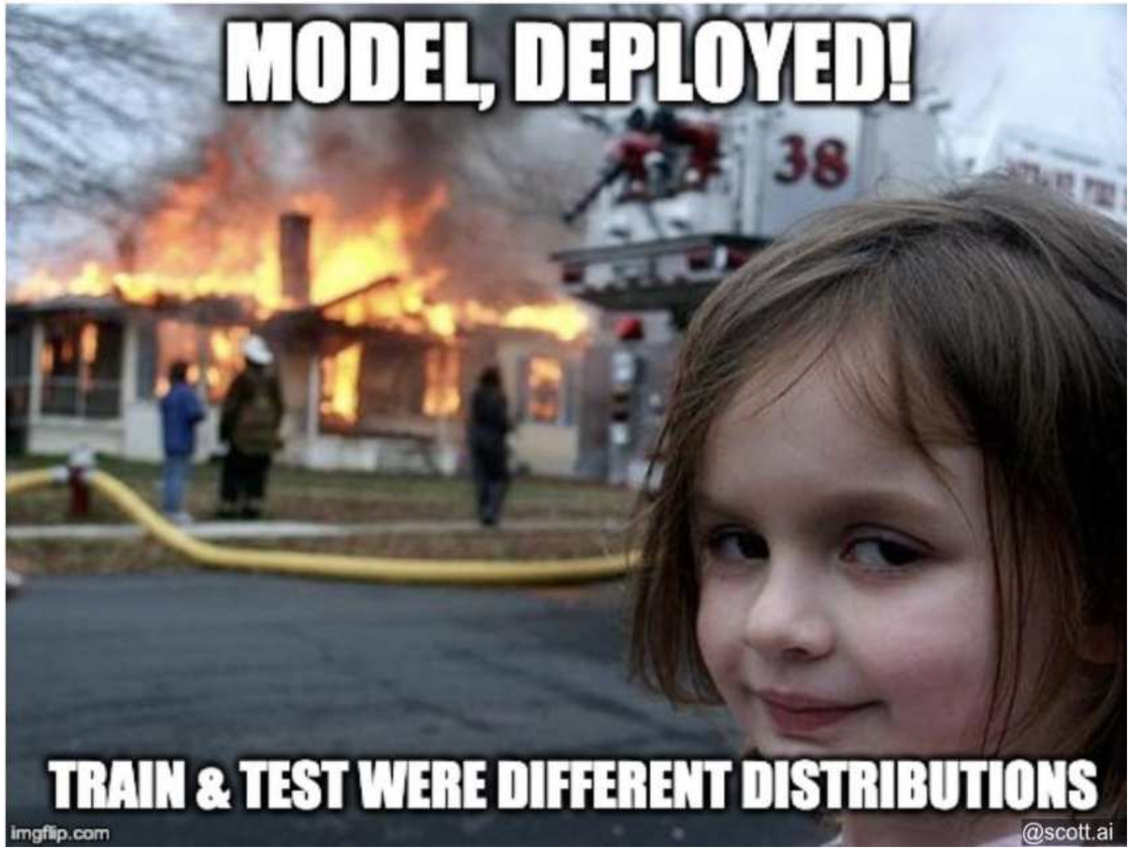
Even if available, novel data does not easily fit into either the earlier or later stages of training

A = Deep Neural Networks
B = Novel data

Memes to Wrap it Up

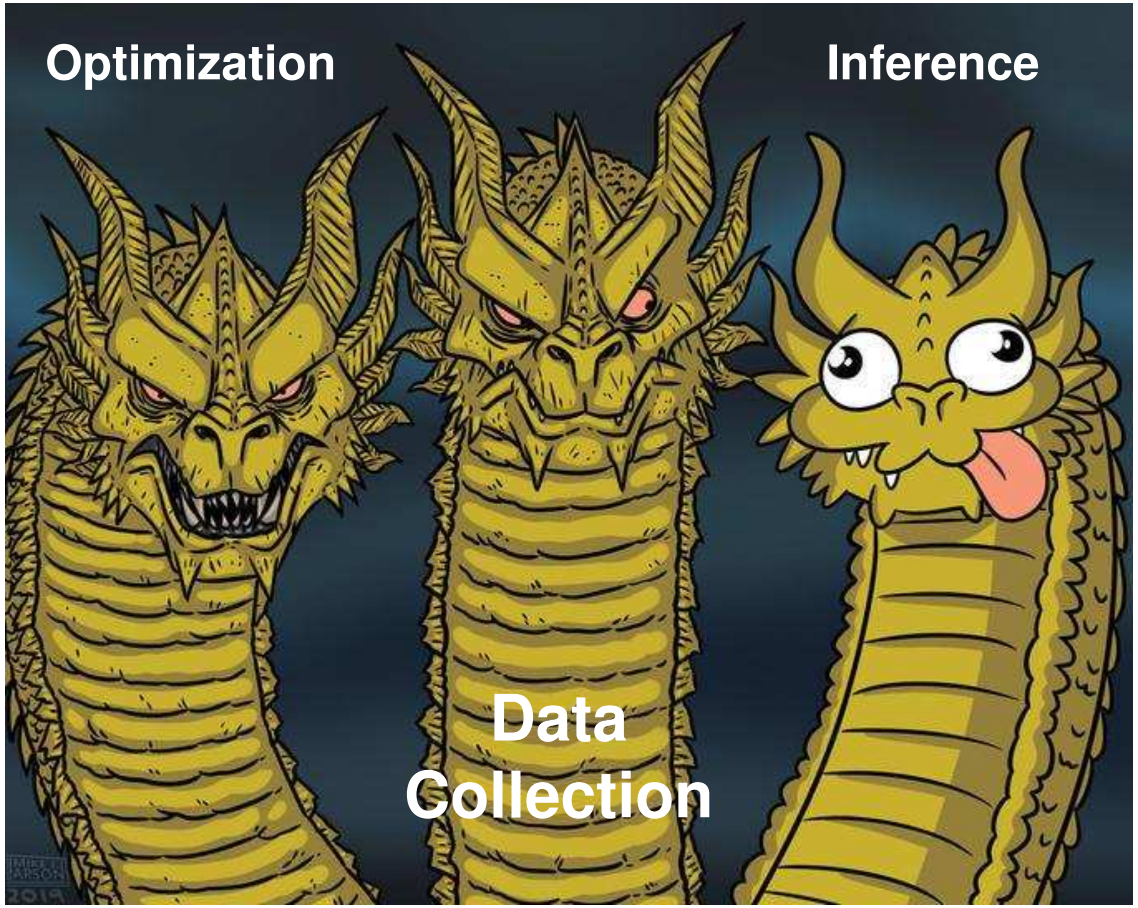
Robustness at Inference





Cannot depend on training to construct robust models

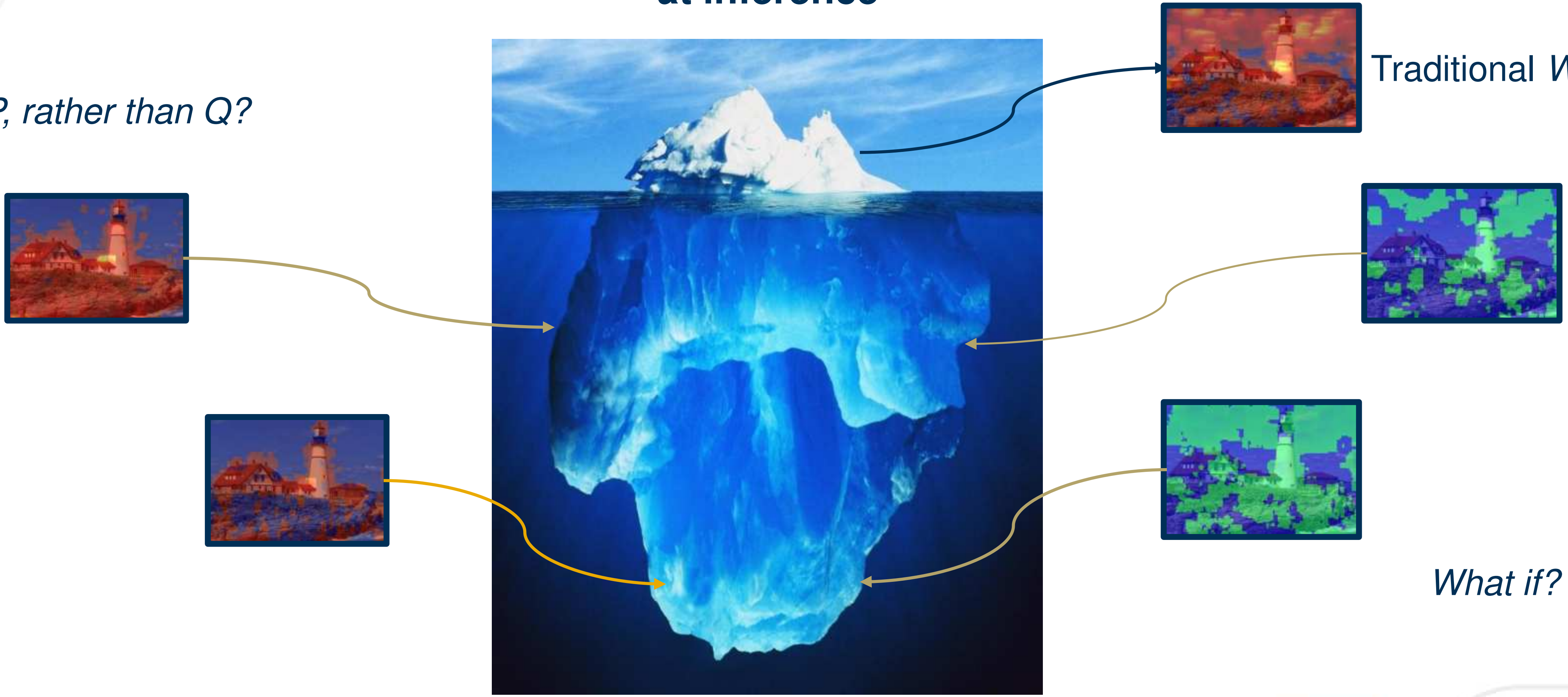
Existing research on robustness focuses on data collection and optimization



Memes to Wrap it Up

Implicit Knowledge in Neural Networks

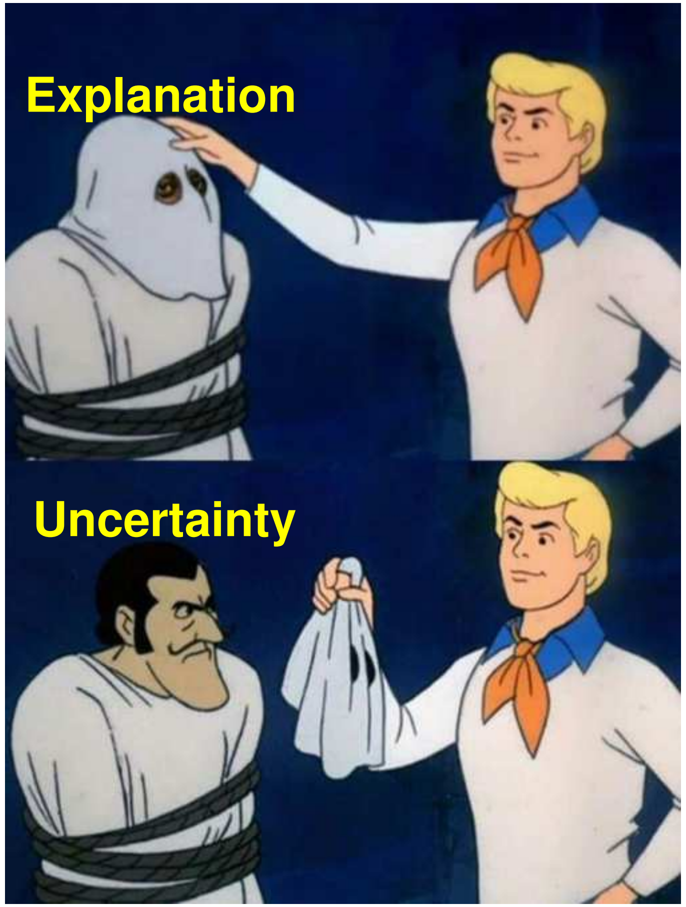
Trained Neural Networks have a wealth of implicit stored knowledge, waiting to be extracted at inference



Memes to Wrap it Up

Explainability Research is Just Uncertainty Research

Explanatory Evaluation reduces Uncertainty



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

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/mipr-2024-tutorial/

{alregib, mohit.p}@gatech.edu

MIPR 2024 Tutorial

The 7th IEEE International Conference on
Multimedia Information Processing and Retrieval

IEEE MIPR 2024

Robust Neural Networks: Towards Explainability, Uncertainty, and Intervenability

Presenters:

Ghassan AlRegib and Mohit Prabhushankar
Georgia Institute of Technology, Georgia Institute of Technology

www.ghassanalregib.info

alregib@gatech.edu, mohit.p@gatech.edu