Gradients in Neural Networks: Interpretation, and Applications in Image Understanding


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

Oct 08, 2023 – Kuala Lumpur, Malaysia

30th ICIP 2023 Kuala Lumpur

OLIVES @GeorgiaTech

Georgia Tech

Gradients in Neural Networks: Interpretation, and Applications in Image Understanding

To cite this Tutorial:

Ghassan AlRegib, and Mohit Prabhushankar. Tutorial on 'A Multifaceted View of Gradients in Neural Networks: Extraction, Interpretation, and Applications in Image Understanding'. IEEE International Conference on Image Processing (ICIP 2023), Kuala Lumpur, Malaysia, Oct 8, 2023.

Omni Lab for Intelligent Visual Engineering and Science (OLIVES)
School of Electrical and Computer Engineering
Georgia Institute of Technology
{alregib, mohit.p}@gatech.edu
Oct 08, 2023 – Kuala Lumpur, Malaysia

Tutorial Materials

Accessible Online

SCAN ME

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

{alregib, mohit.p}@gatech.edu

IEEE ICIP 2023 Tutorial

30th ICIP 2023 Kuala Lumpur

IEEE Signal Processing Society™

Title: A Multi-Faceted View of Gradients in Neural Networks: Extraction, Interpretation and Applications in Image Understanding

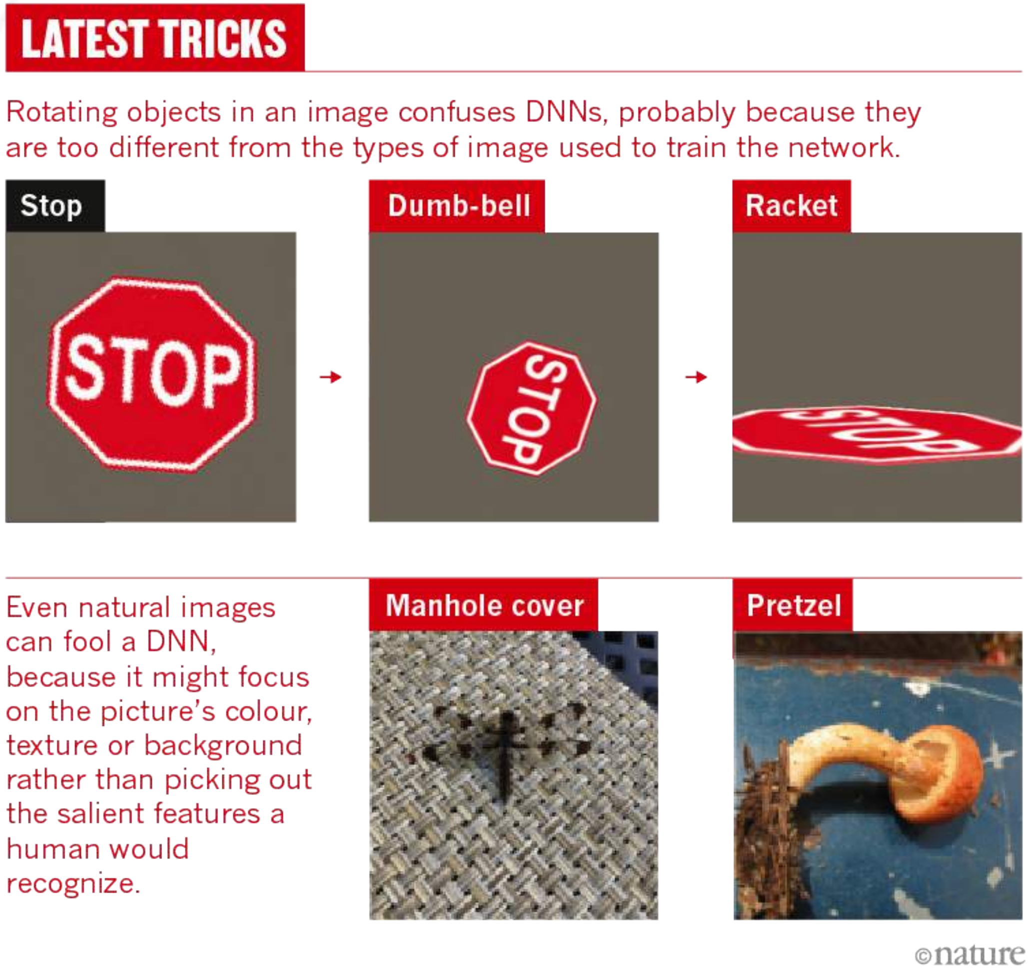
Type / Duration: Half-Day Tutorial (3h)

Deep Learning
Expectation vs Reality

People's expectation of AI and Deep Learning

Deep Learning

Expectation vs Reality



Deep Learning

Expectation vs Reality

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

- Engineers, probably

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

Novel data sources:

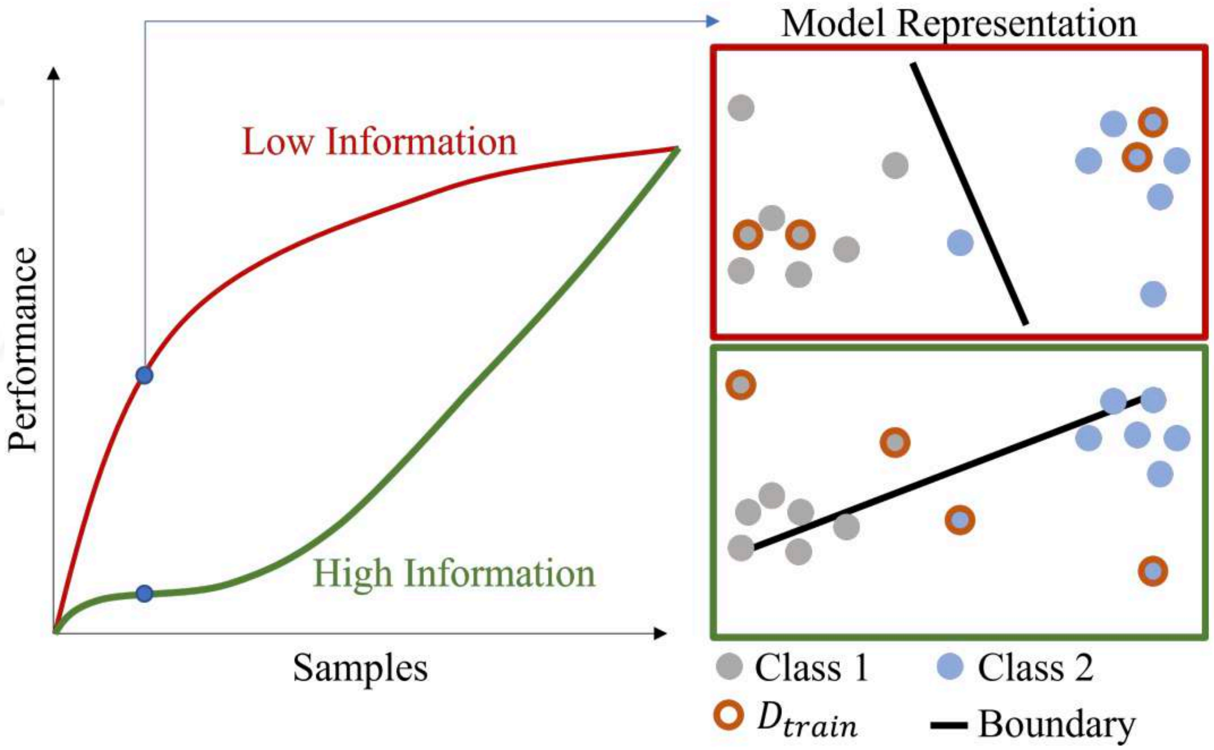
- Test distributions
- Anomalous data
- Out-Of-Distribution data
- Adversarial data
- Corrupted data
- Noisy data
- New classes
- …

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

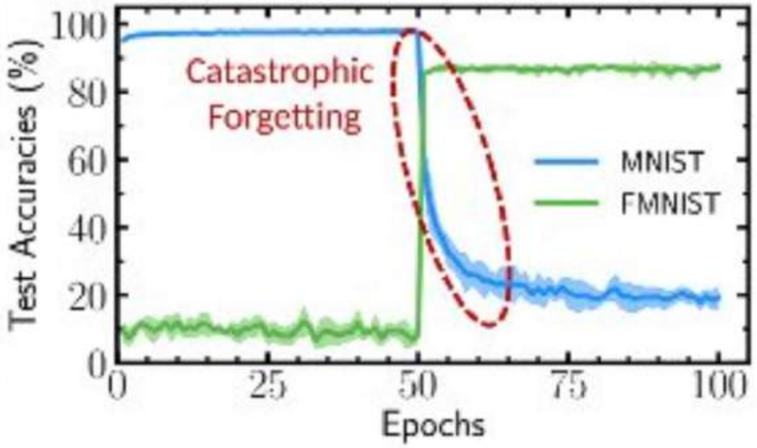
- The first instance of training must occur with less informative samples
- Ex: For autonomous vehicles, less informative means
  - Highway scenarios
  - Parking
  - No accidents
  - No aberrant events

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*.

Deep Learning at Training

Overcoming Challenges at Training: Part 2

Subsequent training must not focus only on novel data



Catastrophic Forgetting

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

Where to handle novel data?

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

Deep Learning at Inference
Overcoming Challenges at Inference

We 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

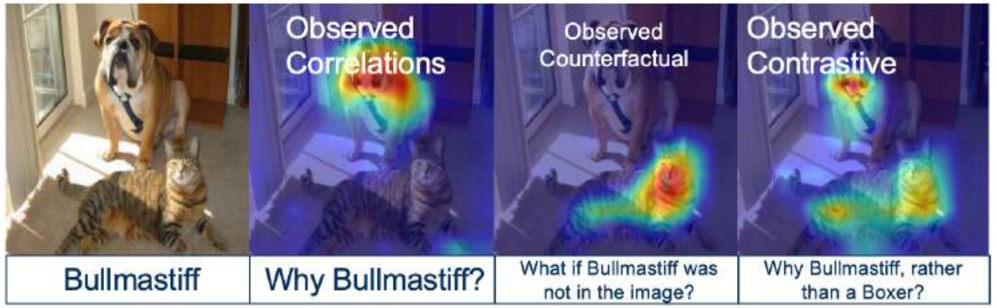
At Inference

Objective

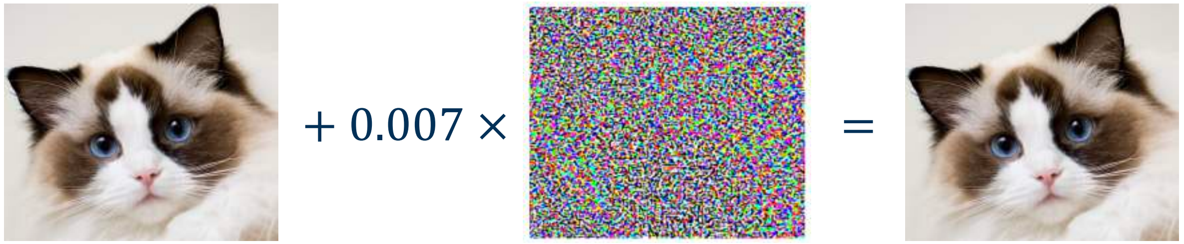
Objective of the Tutorial

To present methodologies to handle novel data at inference using gradients of neural networks

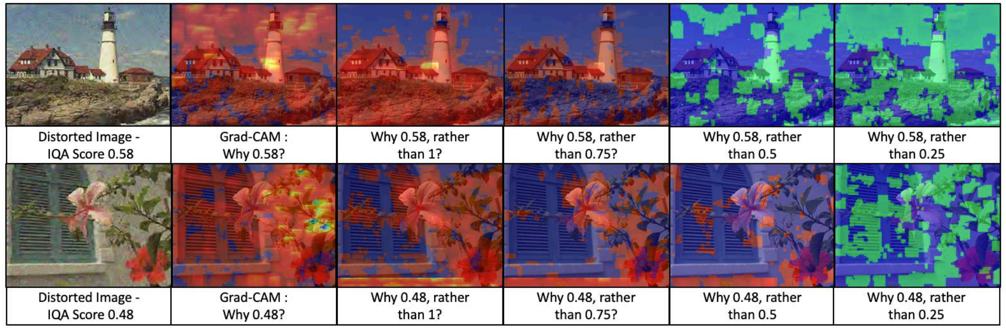
At the end of the tutorial you will be able to



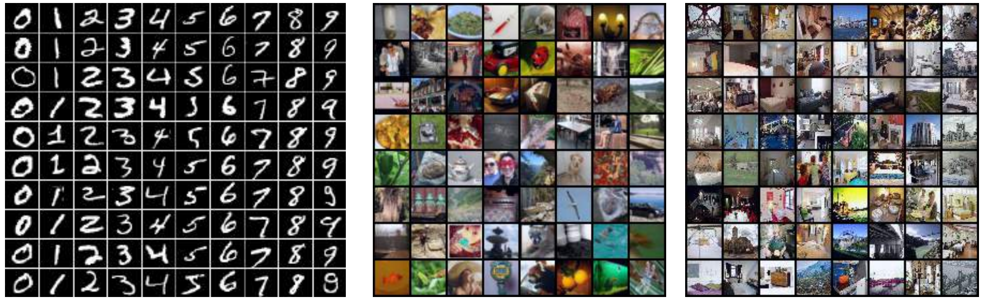
Obtain fine-grained explanations



Engineer (and detect) adversarial examples



Construct XAI techniques for Image Quality Assessment



Training Dataset Testing Dataset

Perform Out-Of-Distribution and Anomaly Detection

Objective

Objective of the Tutorial

To present methodologies to handle novel data at inference using gradients of neural networks

- Part 1: Gradients in Neural Networks
  - Neural network basics, gradient descent, and properties of gradients
- Part 2: Gradients as Information
  - Visual explanations, robust recognition
- Part 3: Gradients as Uncertainty
  - Anomaly, Out-Of-Distribution, corruption, and adversarial detection
- Part 4: Gradients as Expectancy-Mismatch
  - Image Quality Assessment, human visual saliency
- Part 5: Conclusion and Future Directions

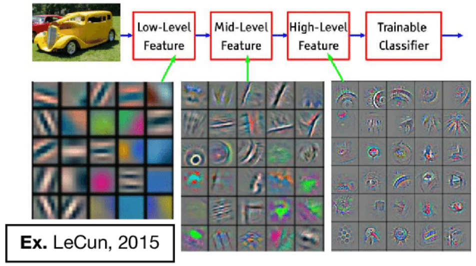
Interpretation, and Applications of Gradients

Part I: Gradients in Neural Networks

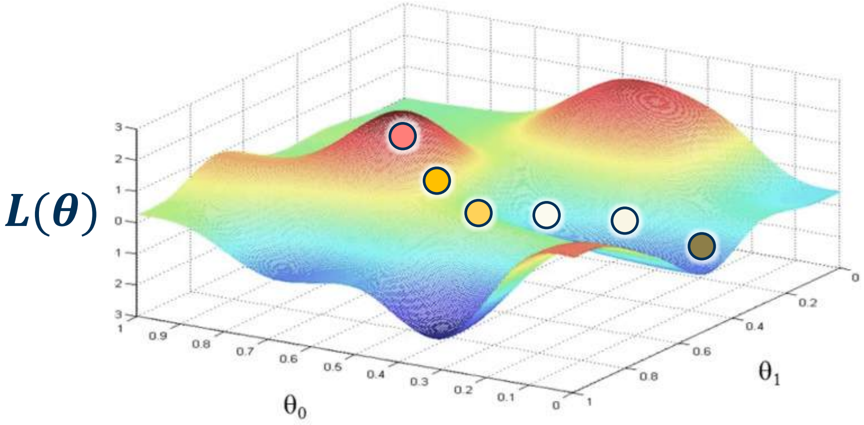
Objectives

Objectives in Part 1

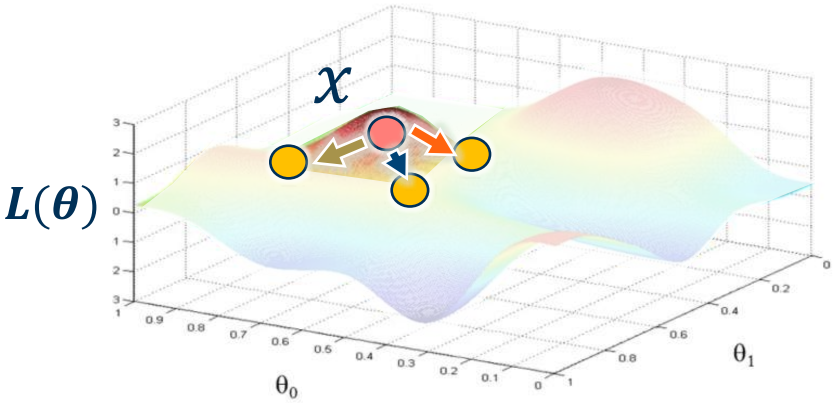
At the end of Part 1 you will be able to



1. Describe the basics of neural networks

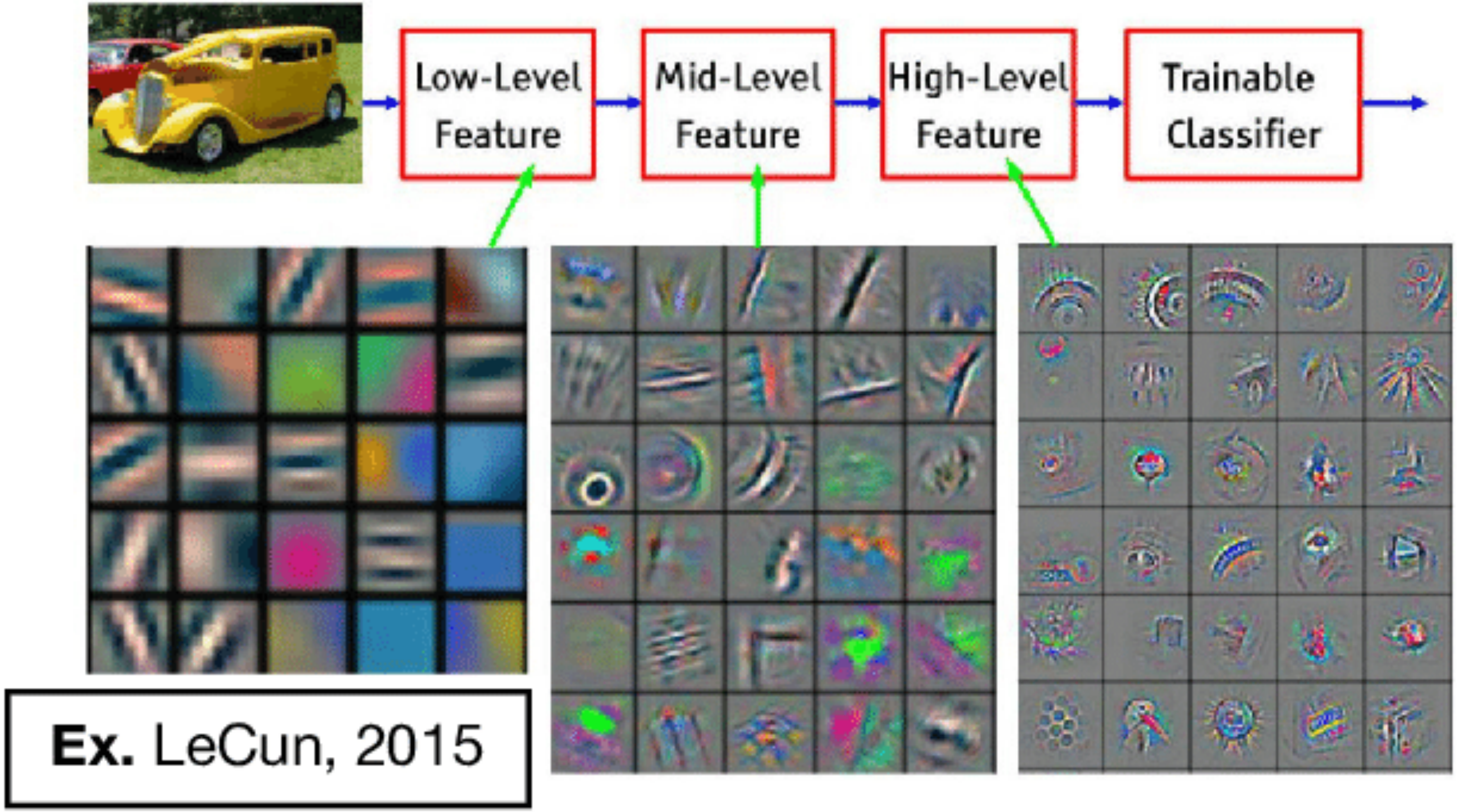


2. Discuss the role of gradients in optimization



3. Discuss relevant properties of gradients

Deep Learning
Overview



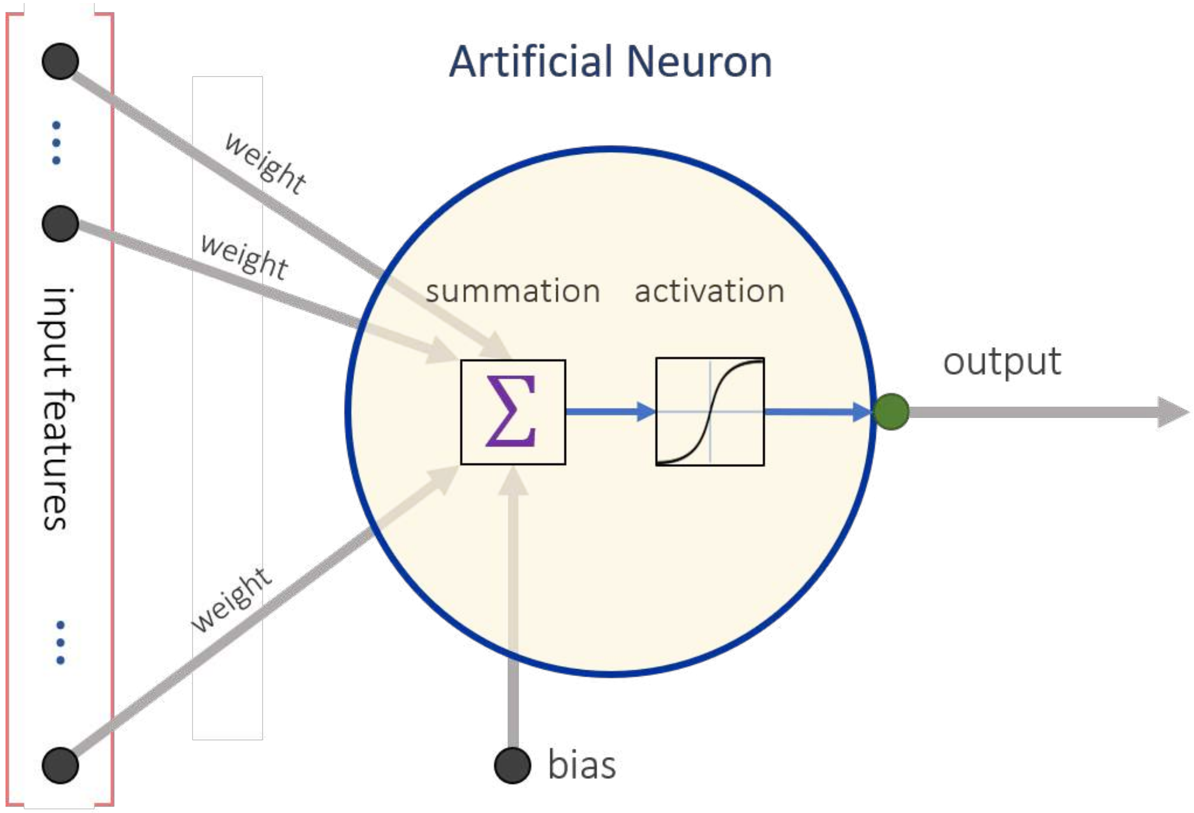
Deep Learning

Neurons

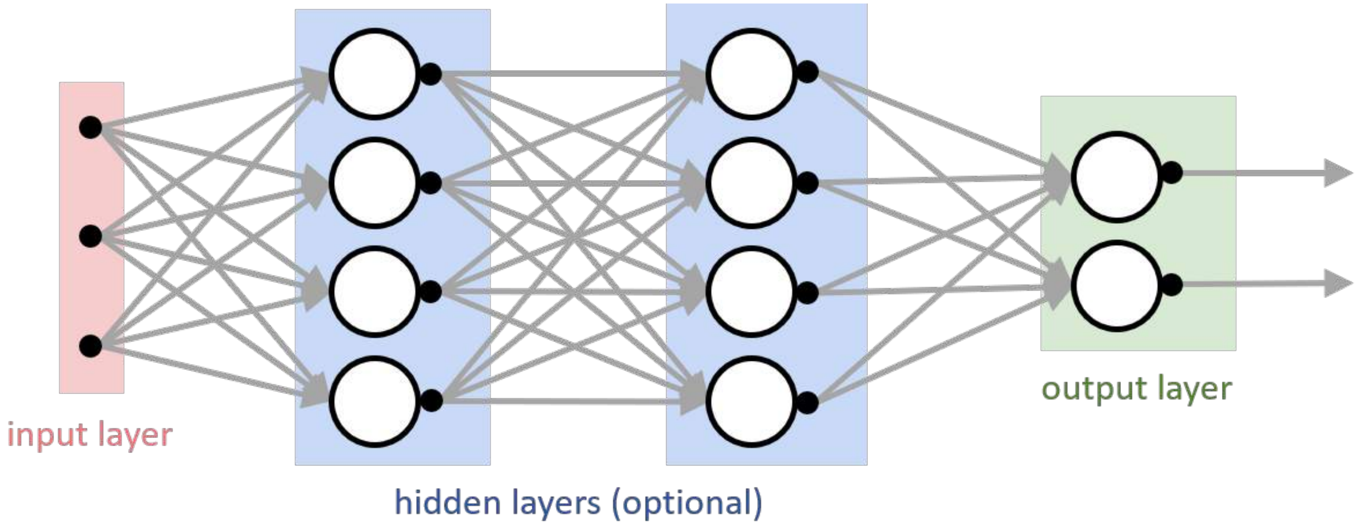
The underlying computation unit is the Neuron

Artificial neurons consist of:

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



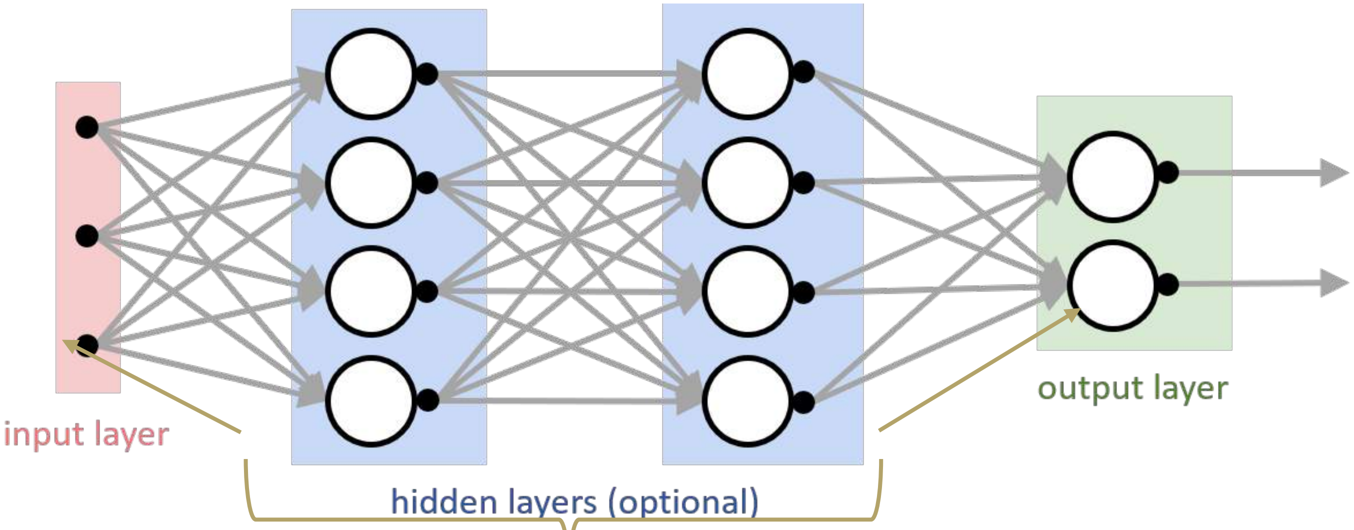
Neurons are stacked and densely connected to construct ANNs

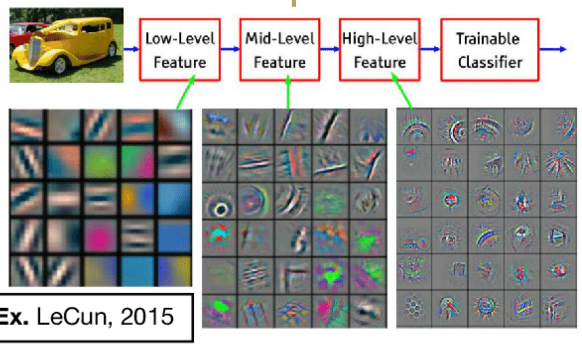


Typically, a neuron is part of a network organized in **layers**:

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

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





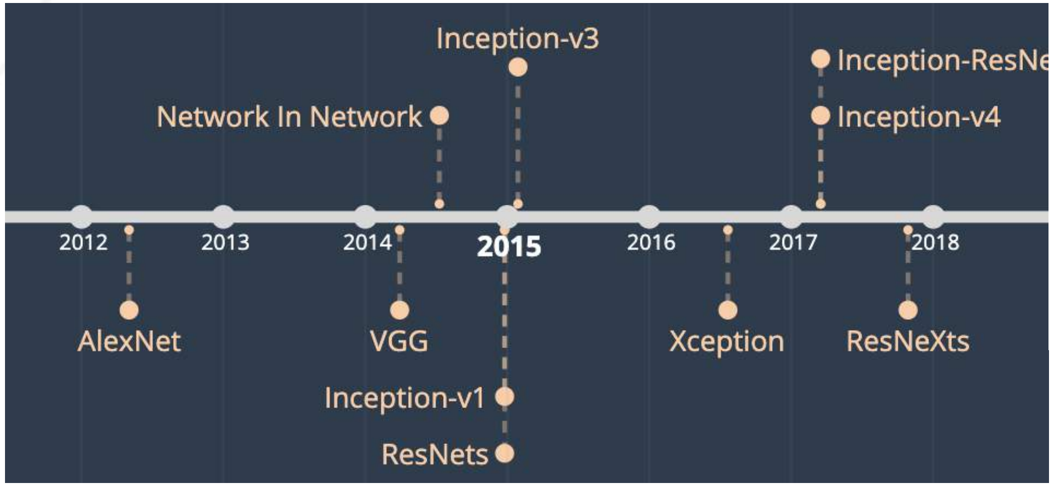
Ex. LeCun, 2015

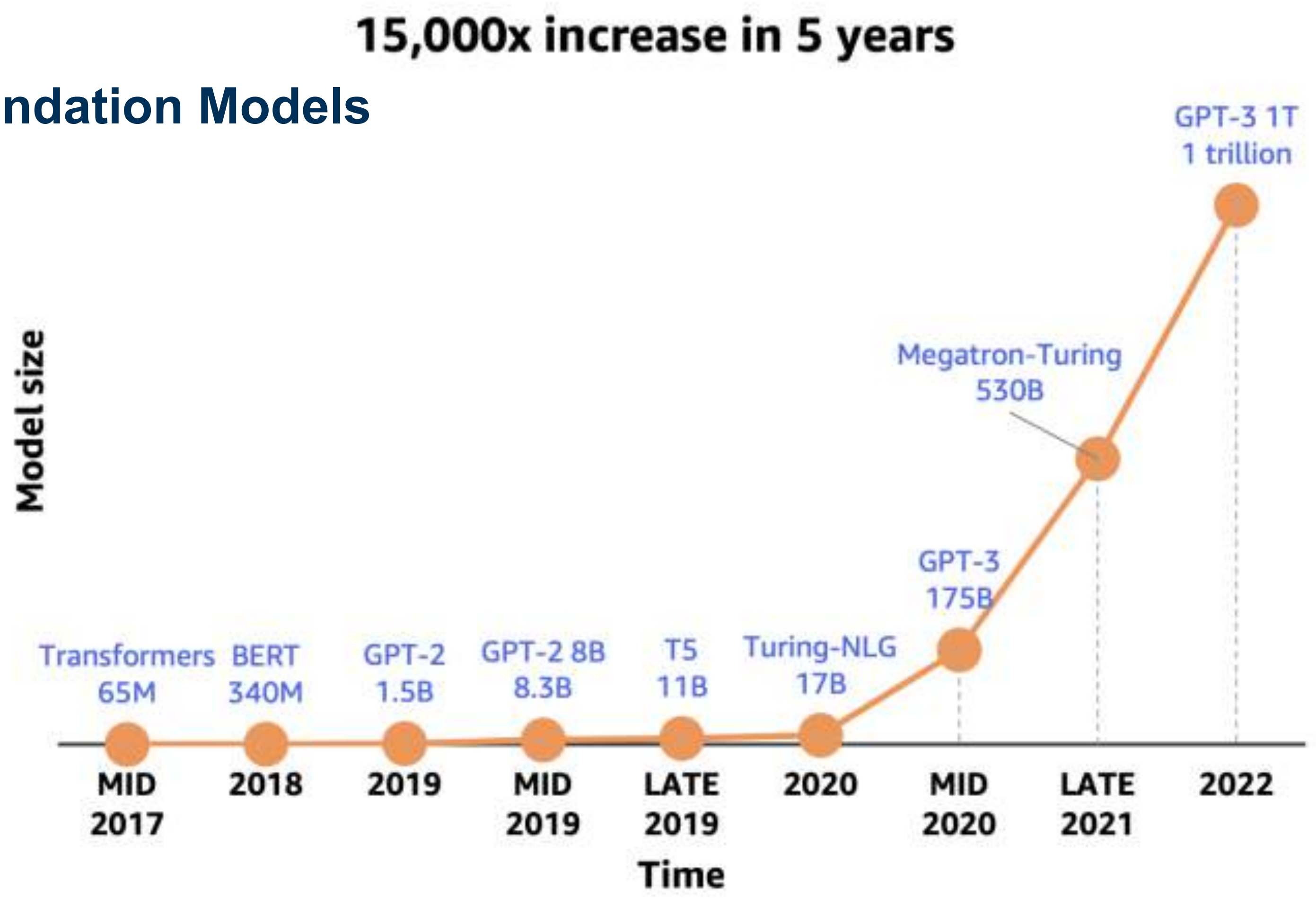
Deep Deep Deep … Deep Deep Learning

Recent Advancements

Transformers, Large Language and Foundation Models

The number of parameters in models has increased exponentially





Training Neural Networks

Stochastically and via Gradient updates

Iteratively reduce a loss function $L(\theta)$ to find the optimal parameters θ

- θ is a combination of weights and biases
- Compute the gradients of a loss function iteratively and update the weights according to the update rule:

$$\theta(t+1) = \theta(t) - \alpha \frac{\partial\, L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

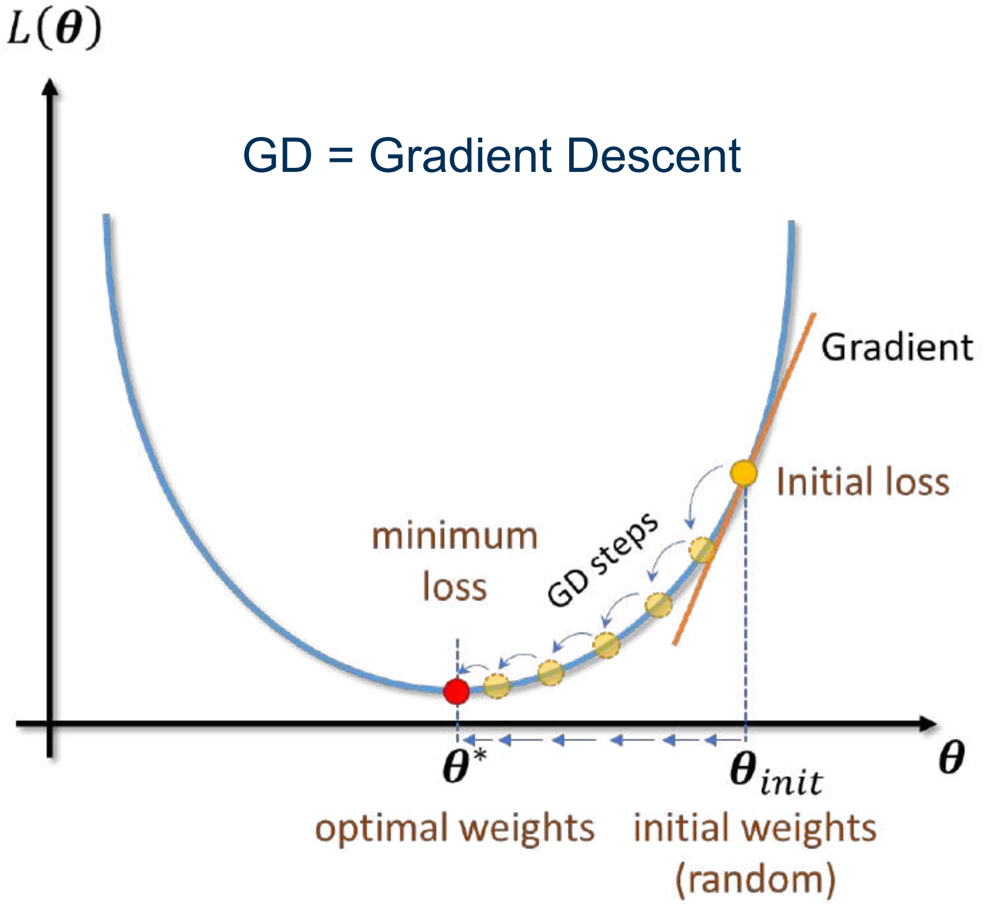
θ = Weights, biases

t = Iteration step

α = Step Length

$L(\theta)$ = Loss function between prediction and ground truth

$\frac{\partial L(\theta)}{\partial \theta}$ = Gradient w.r.t weights and biases

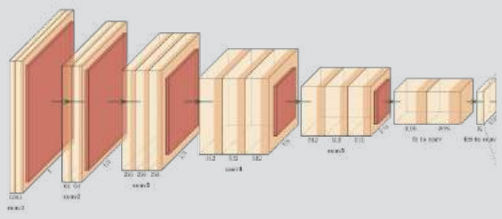


Training Neural Networks

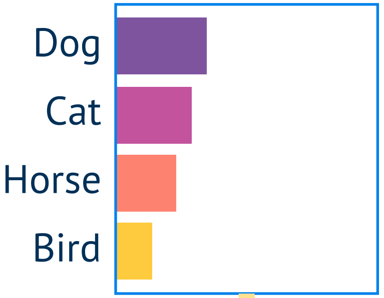
Gradient Descent in Action

Gradients construct the manifold

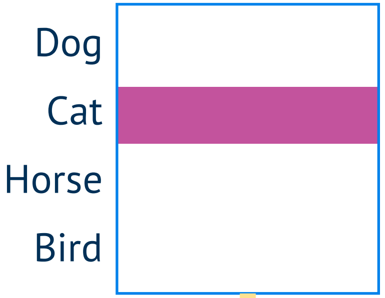
Network $f(\boldsymbol{\theta})$



Predicted Class Probability



Ground-Truth Label

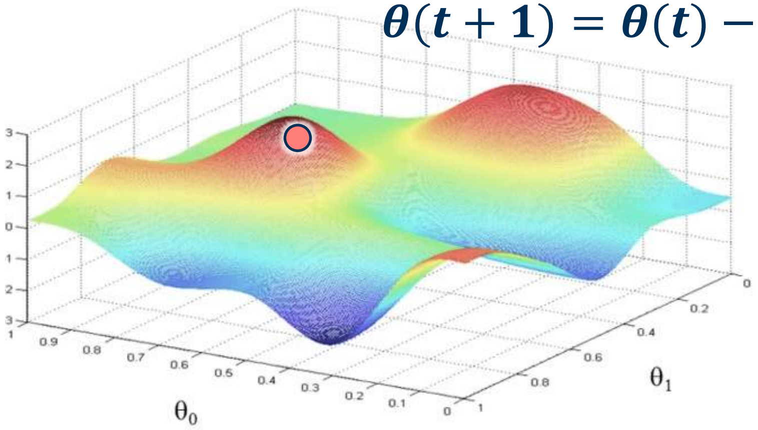


Loss $L(\boldsymbol{\theta})$

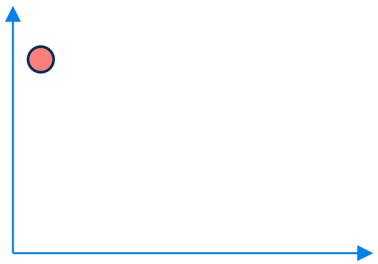
Backprop $L(\boldsymbol{\theta})$ to generate **gradients** $\nabla_{\theta} L(\boldsymbol{\theta})$

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \alpha \nabla_{\theta} L(\boldsymbol{\theta})$$

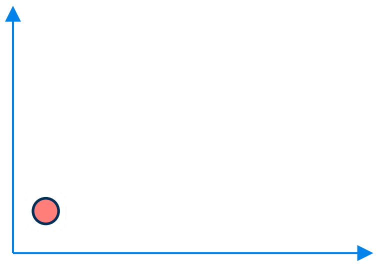
$L(\boldsymbol{\theta})$



Loss



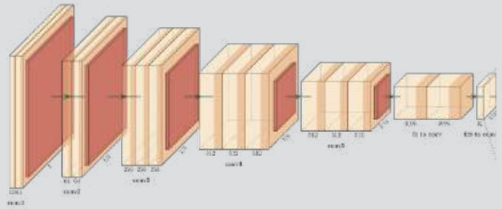
Accuracy



Gradient Descent in Action

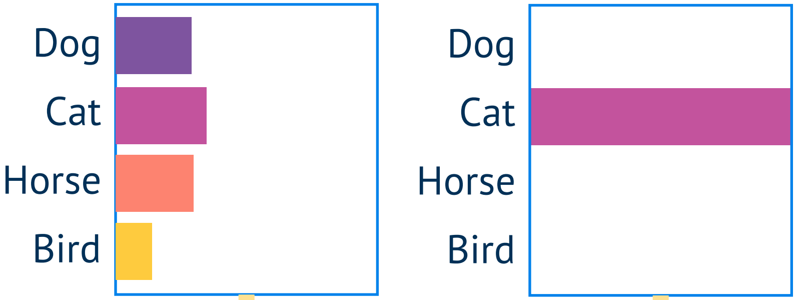
Gradients construct the manifold

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



Predicted Class Probability

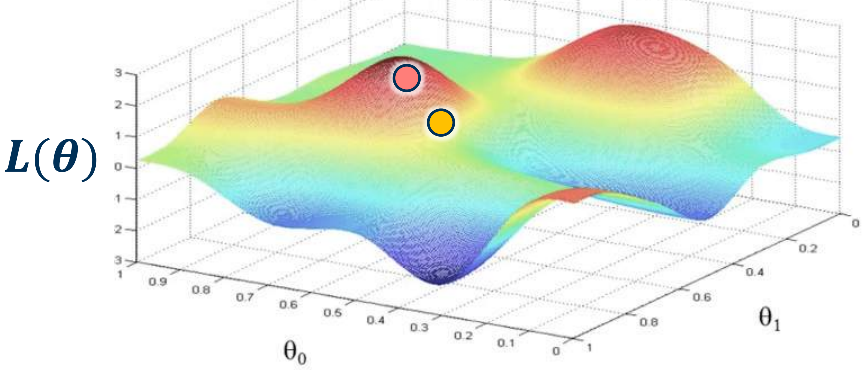
Ground-Truth Label

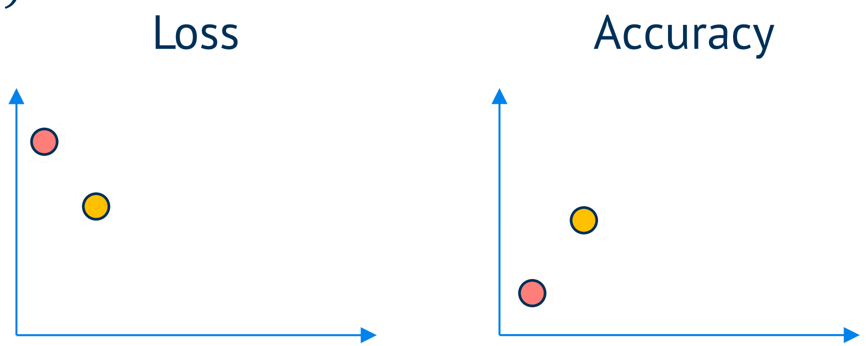


Backprop $\boldsymbol{L(\theta)}$ to generate **gradients** $\boldsymbol{\nabla_\theta\, L(\theta)}$

Loss $\boldsymbol{L(\theta)}$

$$\boldsymbol{\theta(t+1) = \theta(t) - \alpha\, \nabla_\theta\, L(\theta)}$$



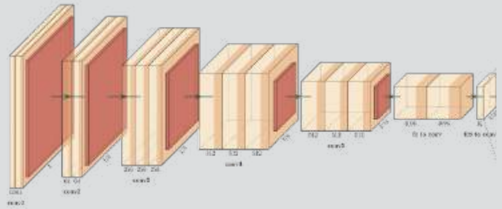


Training Neural Networks

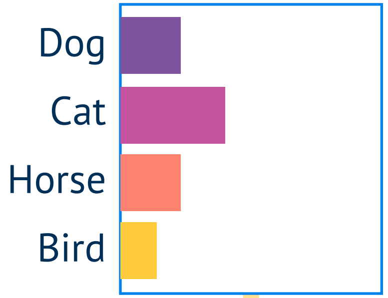
Gradient Descent in Action

Gradients construct the manifold

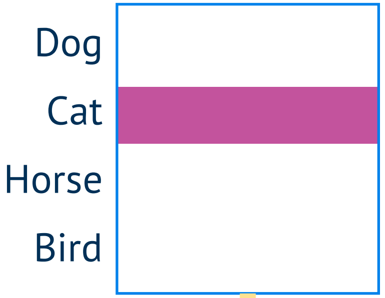
Network $f(\theta)$



Predicted Class Probability



Ground-Truth Label

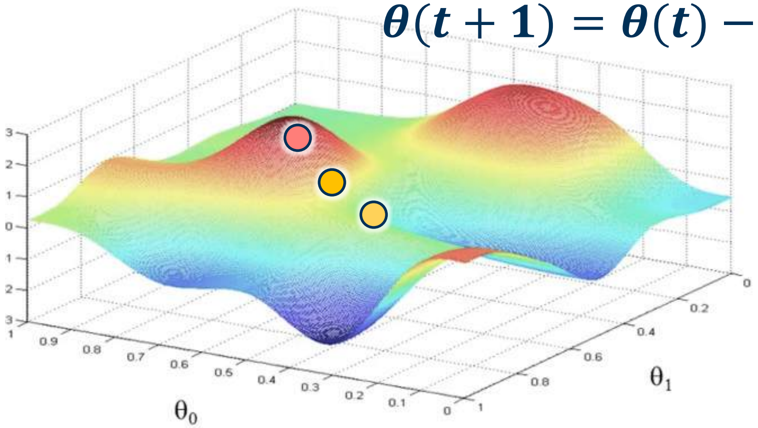


Loss $L(\theta)$

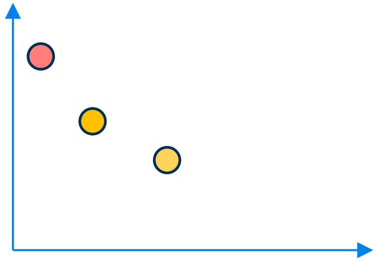
Backprop $L(\theta)$ to generate **gradients** $\nabla_\theta L(\theta)$

$$\theta(t+1) = \theta(t) - \alpha\, \nabla_\theta L(\theta)$$

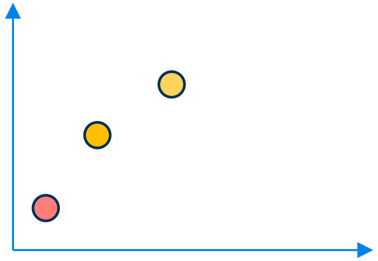
$L(\theta)$



Loss



Accuracy

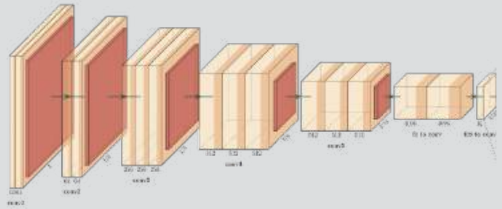


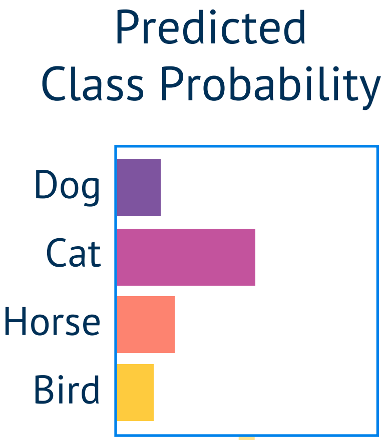
Training Neural Networks

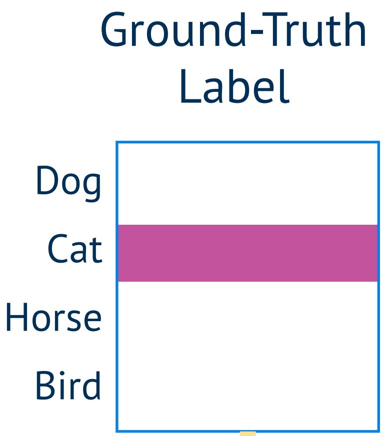
Gradient Descent in Action

Gradients construct the manifold

Network $f(\theta)$



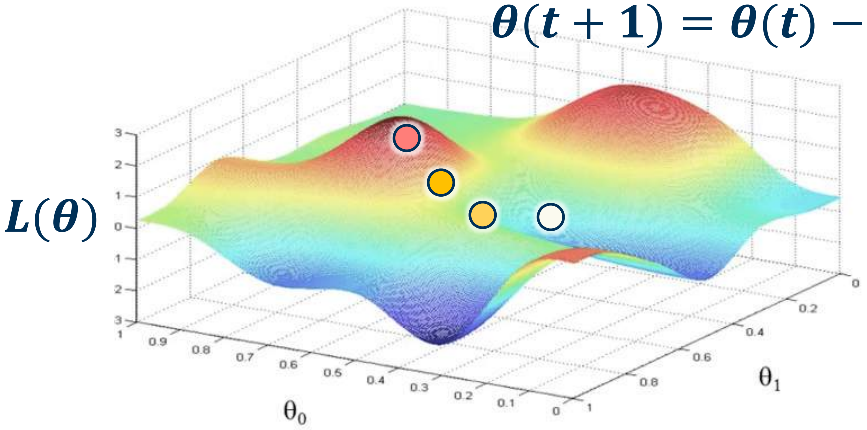


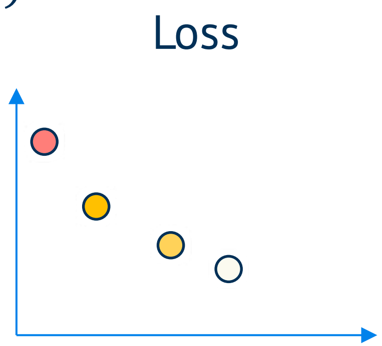


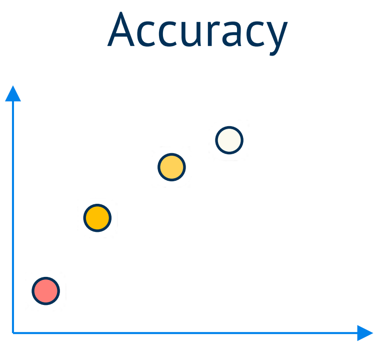
Loss $L(\theta)$

Backprop $L(\theta)$ to generate **gradients** $\nabla_\theta L(\theta)$

$$\theta(t+1) = \theta(t) - \alpha \nabla_\theta L(\theta)$$

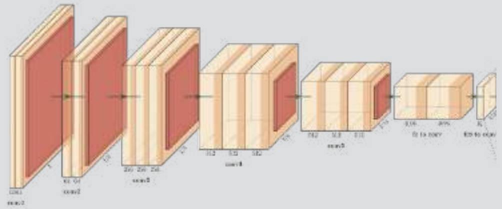


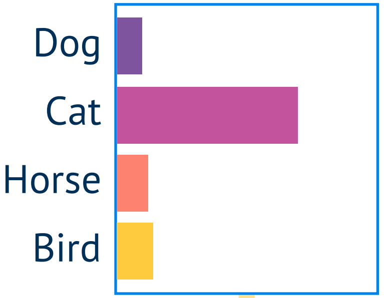


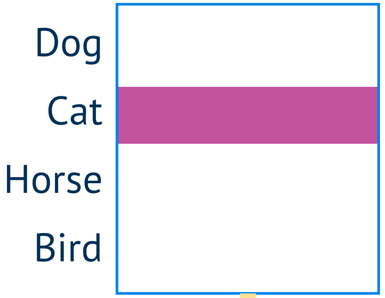


Gradients construct the manifold

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



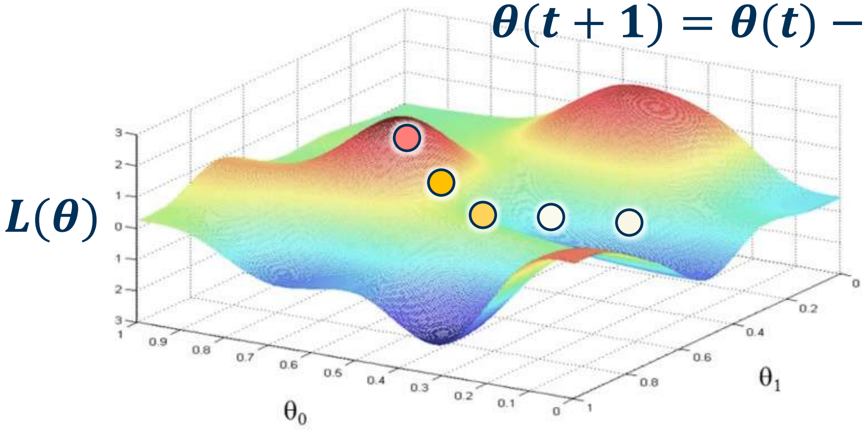


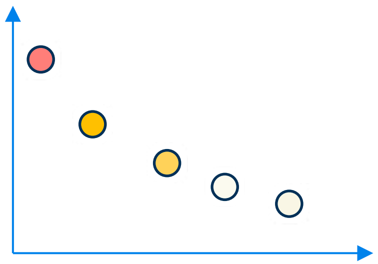


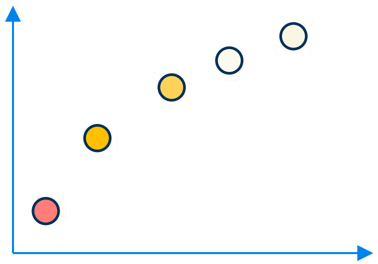
Loss $\boldsymbol{L(\theta)}$

Backprop $\boldsymbol{L(\theta)}$ to generate **gradients** $\boldsymbol{\nabla_\theta L(\theta)}$

$$\boldsymbol{\theta(t+1) = \theta(t) - \alpha \nabla_\theta L(\theta)}$$

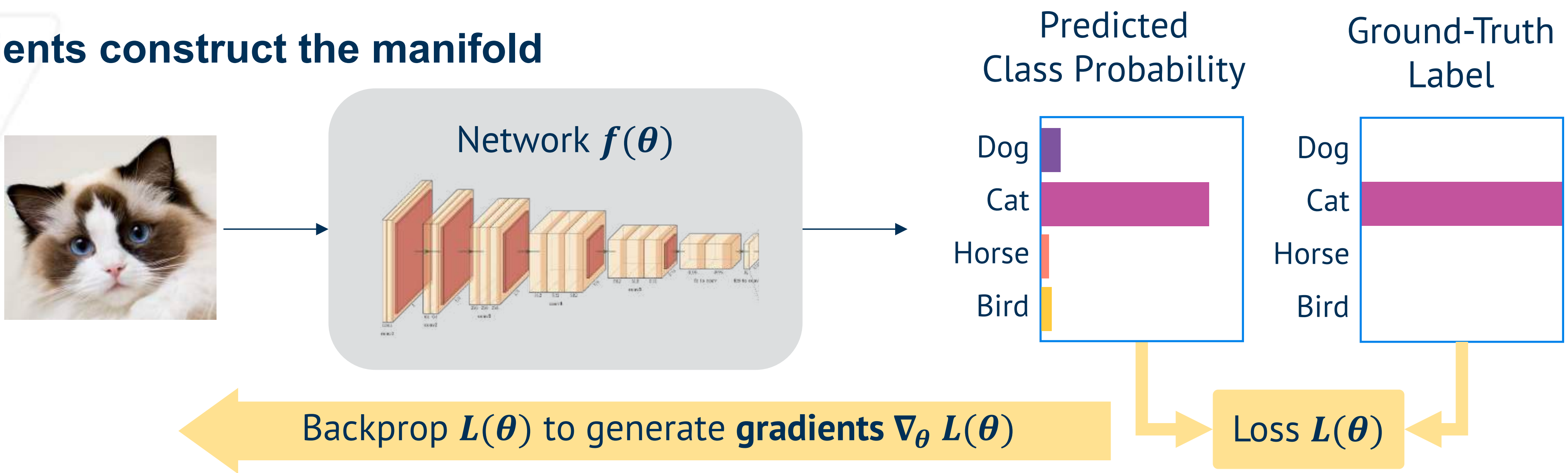




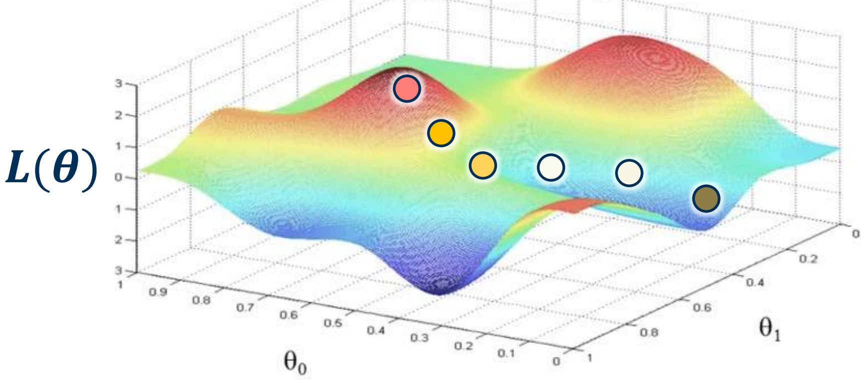


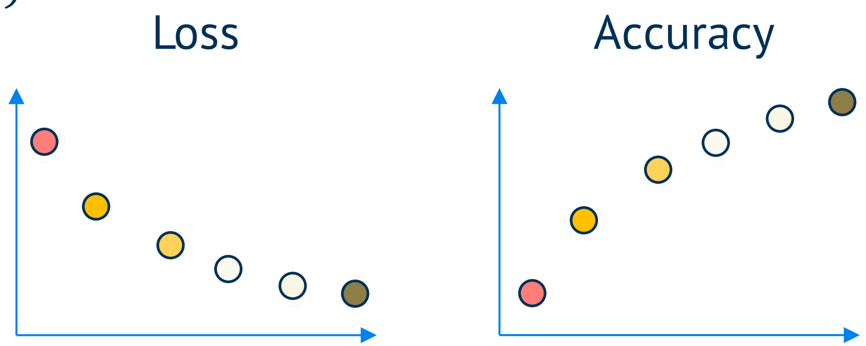
Training Neural Networks
Gradient Descent in Action

Gradients construct the manifold



$$\theta(t+1) = \theta(t) - \alpha \nabla_\theta L(\theta)$$





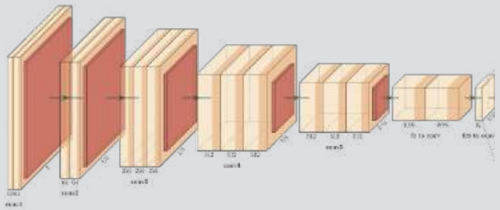
Our Goal

To Characterize Data at Inference

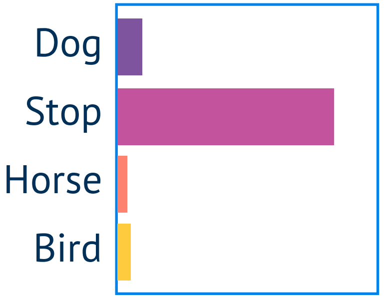
Goal: Given the novel data point, the network, and its prediction, *characterize* the data as a function of the learned knowledge

Given

Network $f(\theta)$



Predicted Class Probability



Goal

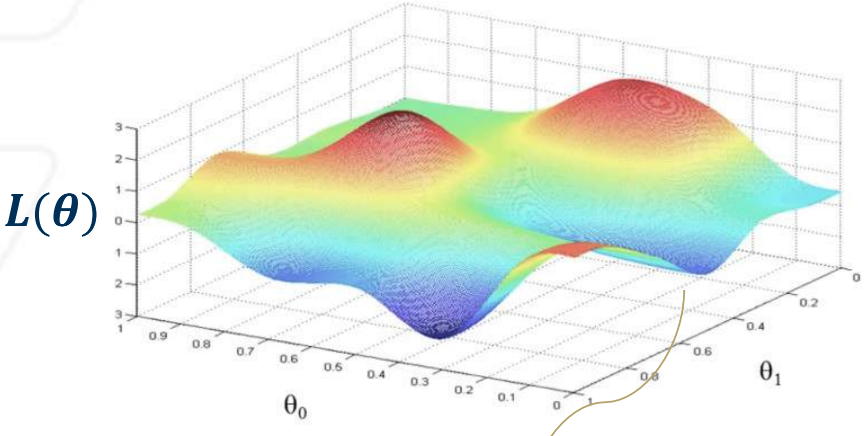
Represent the novel green traffic sign as a function of the learned red traffic sign

Our Claim: Gradients provide the methodology!

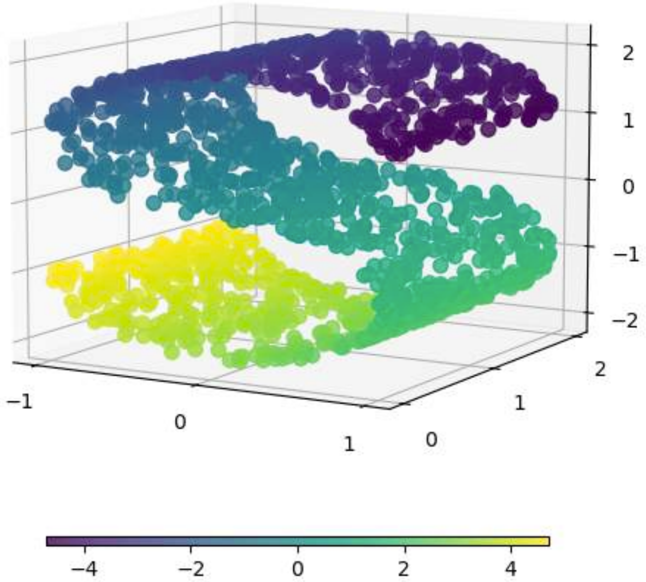
Challenges at Inference

A Quick note on Manifolds..

Manifolds are compact topological spaces that allow exact mathematical functions



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



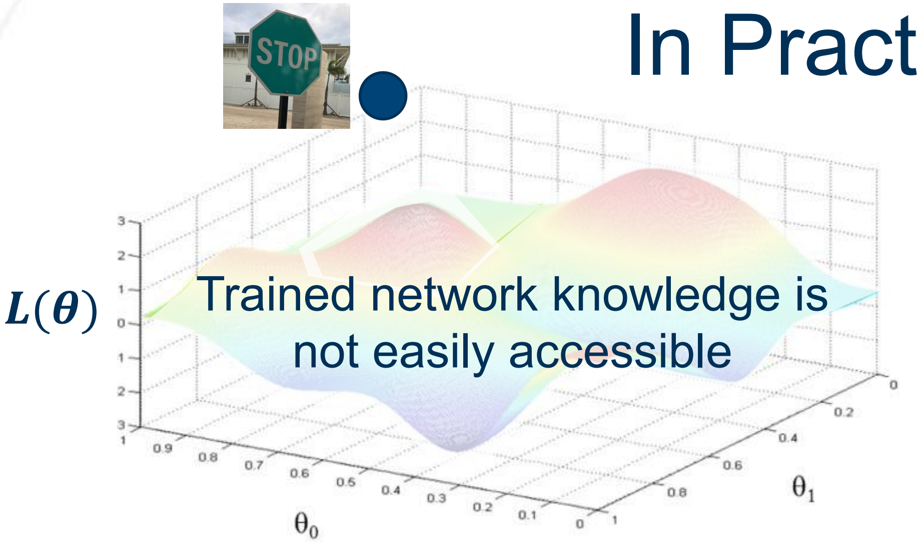
Real data visualizations generated using dimensionality reduction algorithms (Isomap)

Challenges at Inference

Manifolds at Inference

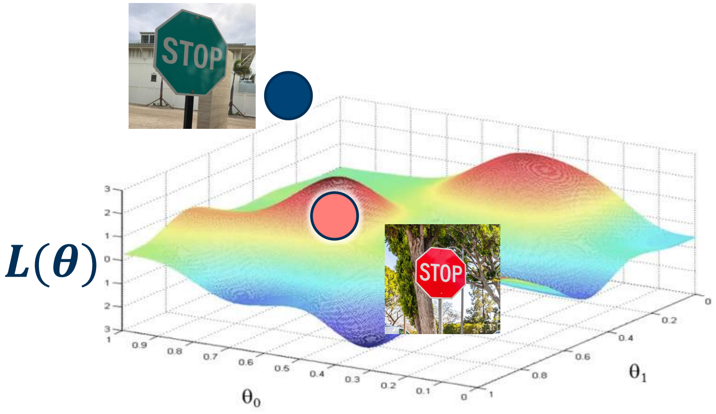
However, at inference only the test data point is available and the underlying structure of the manifold is unknown

In Practice



Existing methodologies estimate this manifold using surrogate networks and validation data at inference. However, they lose generalization performance.

Ideal Goal

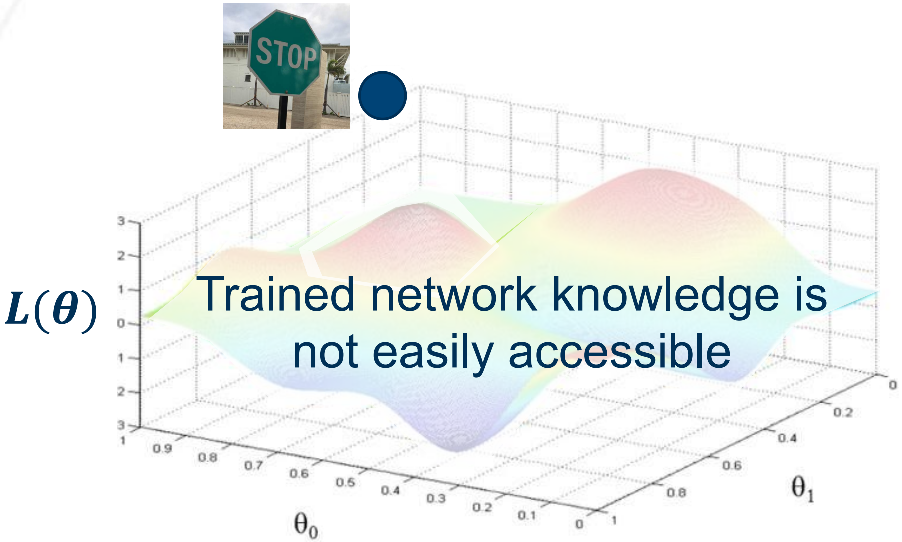


Represent the novel green traffic sign as a function of the learned red traffic sign

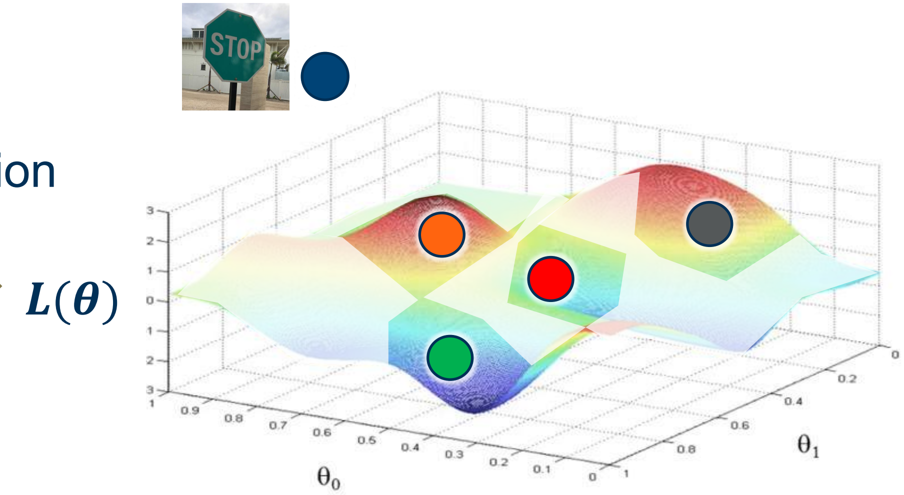
Challenges at Inference

Existing Solutions

Kim et.al.[1] use a KNN classifier on validation data at inference to characterize new test data



Additional validation data



The surrogate (approximate) manifold is derived from K-Nearest Neighbors search

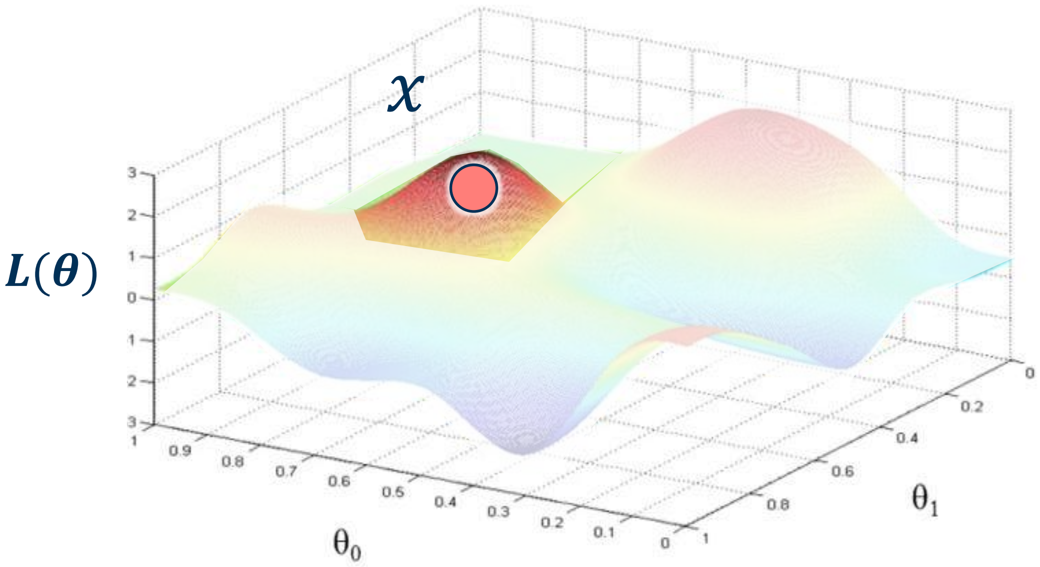
Cons of surrogates:

1. Requires a validation set at inference
2. Computationally impractical scale
3. Authors show that performance on anything greater than MNIST is comparable/worse than baseline

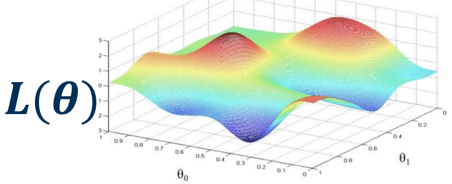
[1] Jiang, H., Kim, B., Guan, M., & Gupta, M. (2018). To trust or not to trust a classifier. *Advances in neural information processing systems*, *31*.

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



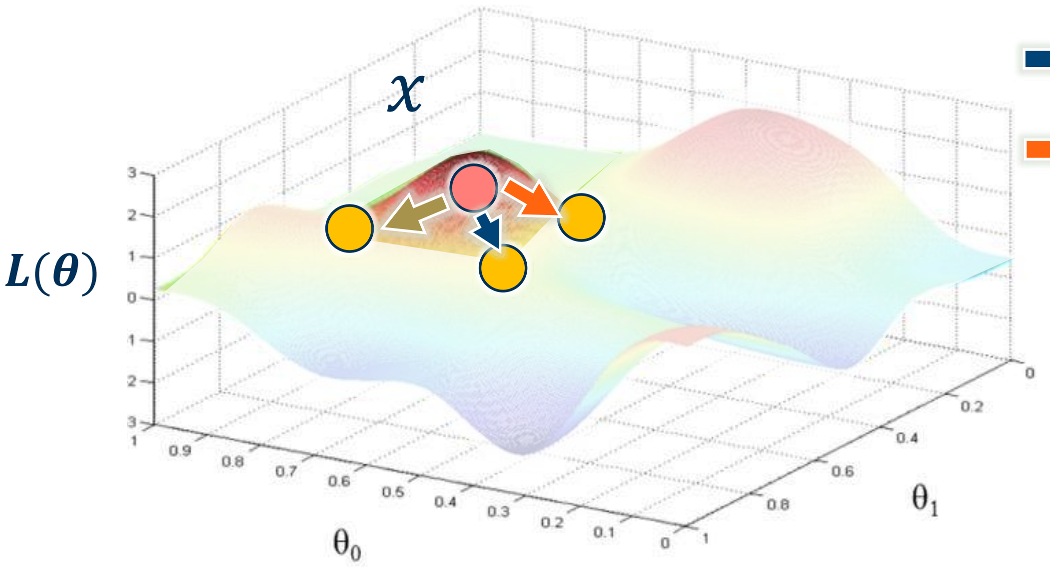
$\boldsymbol{\alpha}\, \nabla_{\boldsymbol{\theta}}\, \boldsymbol{L}(\boldsymbol{\theta})$ provides local information up to a small distance $\boldsymbol{\alpha}$ away from $\boldsymbol{x}$

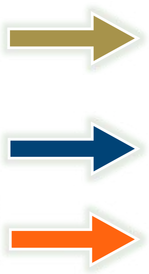
The exact nature and utility of this information is discussed in Part 2

Relevant Properties of Gradients

Direction of Steepest Descent

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





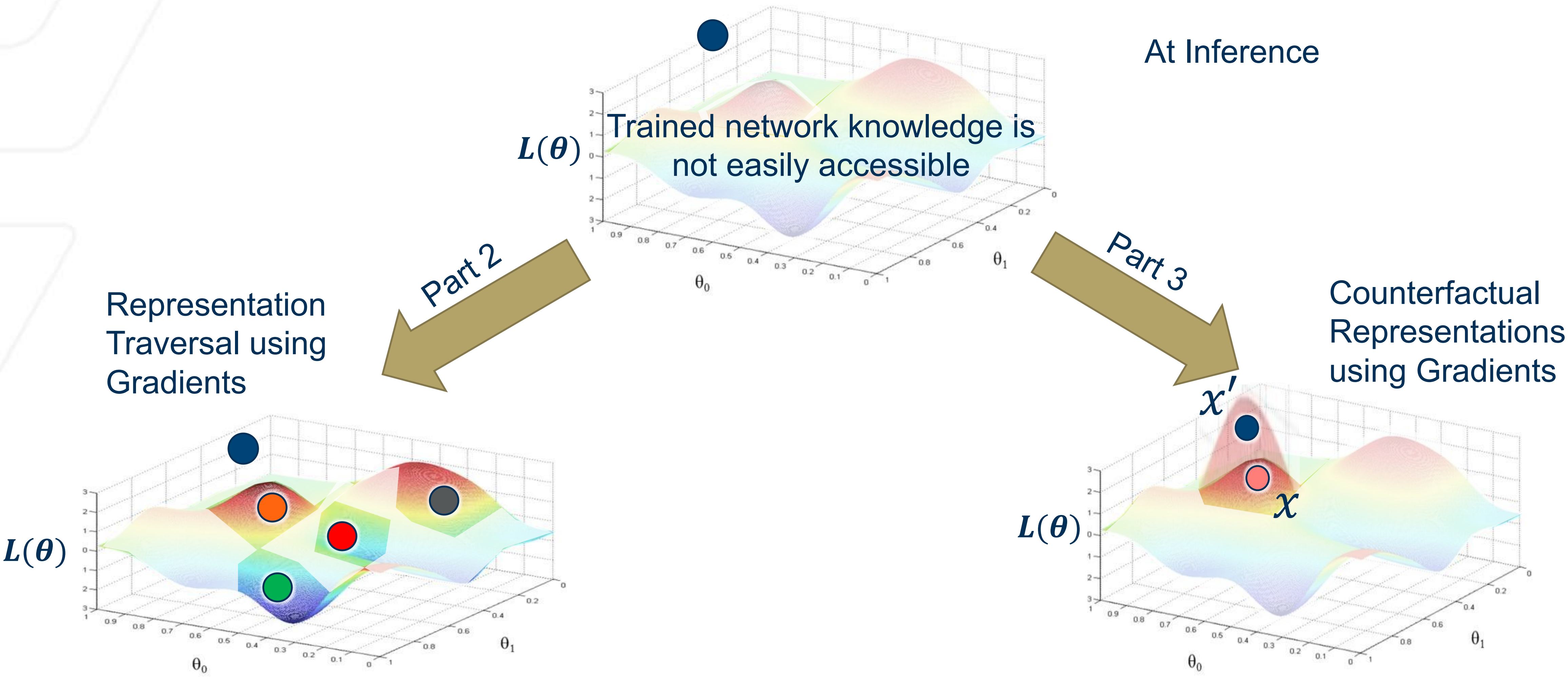
Which direction should we optimize towards (knowing only the local information)?

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

The exact nature and utility of this directional information is discussed in Part 3

Our Technical Goal

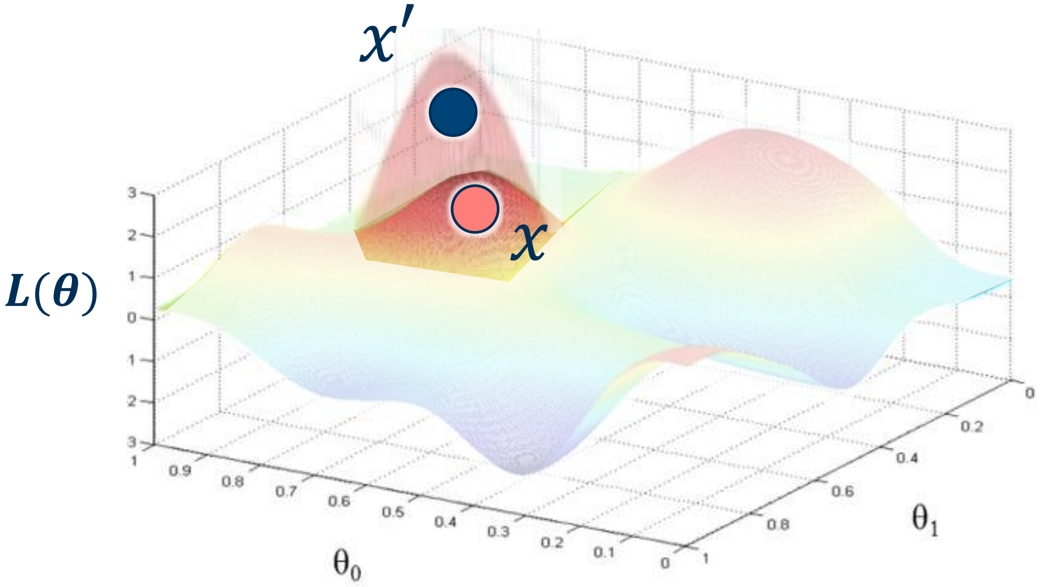
To Characterize the Learned Knowledge



Relevant Properties of Gradients

Counterfactual Manifolds

Gradients allow interventions either on the data or the manifolds to create counterfactuals



- Original manifold with x
- Counterfactual manifold with x'

Counterfactuals can be interpreted as changing the manifold to fit the new data

The exact nature and utility of these counterfactual manifolds is discussed in Part 4

Takeaways

Takeaways from Part 1

- **Part 1: Gradients in Neural Networks**
  - Deep Learning cannot easily generalize to novel data
  - Novel data cannot always be handled during Training
  - Gradients provide local information around the vicinity of $\boldsymbol{x}$
  - Gradients allow choosing the fastest direction of descent given a loss function $\boldsymbol{L}(\boldsymbol{\theta})$
  - Gradients allow interventions either on the data or the manifolds to create counterfactuals
- Part 2: Gradients as Information
- Part 3: Gradients as Uncertainty
- Part 4: Gradients as Expectancy-Mismatch
- Part 5: Conclusion and Future Directions

Interpretation, and Applications of Gradients

Part 2: Gradients as Information

Objectives

Objectives in Part 2

- Discuss three types of Information
- Interpret gradients as Fisher Information
- Visual Explanations
  - Explanatory Paradigms: Correlations, Counterfactuals, and Contrastives
  - GradCAM
  - ContrastCAM
- Robust Recognition under Challenging Conditions: Introspective Learning
  - Introspective Features
  - Robustness measures: Accuracy and Calibration
  - Downstream Applications

Information

Types of Information

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

Shannon Information
(Surprise of an event)

$$H[X] = -\sum_{i=1}^{N} p(x_i)\log_2 p(x_i)$$

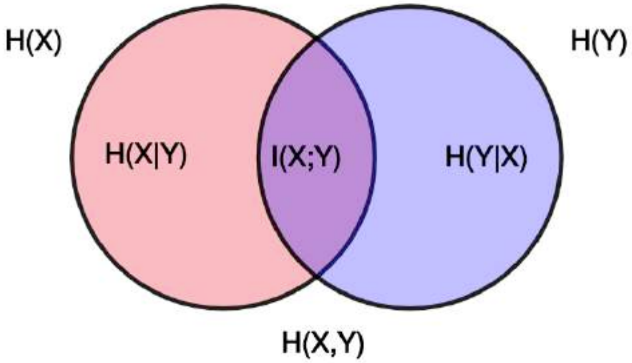
$H[X]$ = Shannon Entropy
$p(x_i)$ = Probability of event x_i

Connects surprise to probability

Mutual Information
(Surprise conditioned on another event)

$$I(X;Y) = H[X] + H[Y] - H(X,Y)$$

$H[X]$ = Shannon Entropy of X
$H[Y]$ = Shannon Entropy of Y
$H(X,Y)$ = Joint Entropy



Fisher Information
(Surprise of underlying distribution)

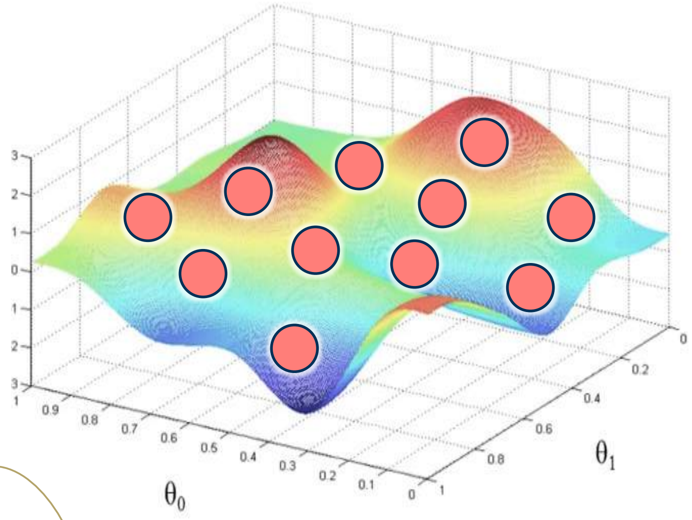
$$I(\theta) = Var\left(\frac{\partial}{\partial\theta} l(\theta|x)\right)$$

θ = Statistic of distribution
$\ell(\theta \mid x)$ = Likelihood function

Variance of the partial derivative w.r.t. θ of the Log-likelihood function $\ell(\theta \mid x)$.

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[1], $I(\theta)$ 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

A key feature is that every sample draws information from the underlying distribution!

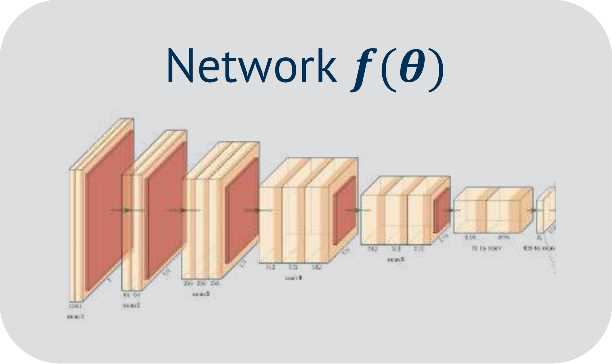
[1] A good blogpost about Fisher Information: https://towardsdatascience.com/an-intuitive-look-at-fisher-information-2720c40867d8

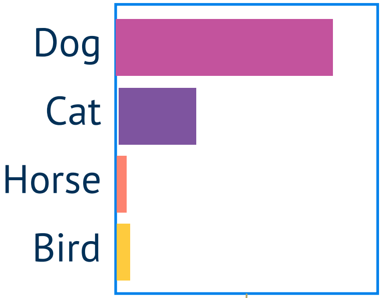
Fisher Information

Gradients as Fisher Information

Gradients infer information about the statistics of underlying manifolds

x

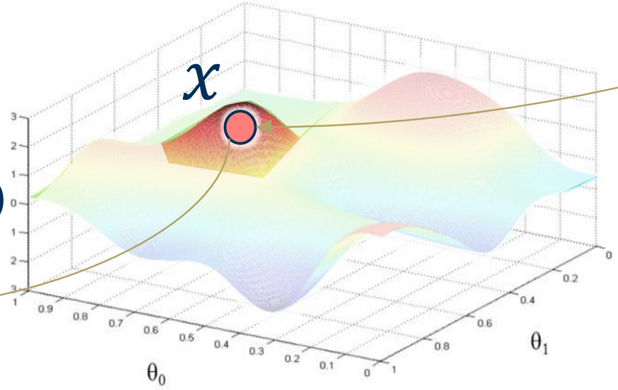




In this case, the image and its prediction extracts nose, mouth and jowl features.

Local information (specific to x) is sufficient!

$l(\theta|x)$



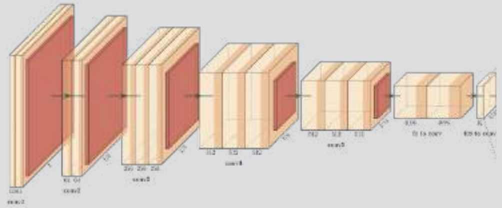
A key feature is that every sample draws information from the underlying distribution! And this information can be visualized.

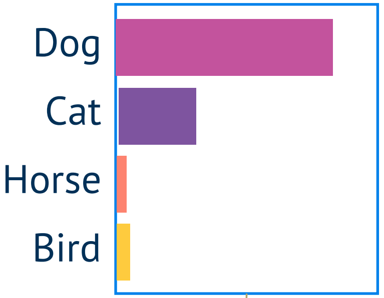
[1] A good blogpost about Fisher Information: https://towardsdatascience.com/an-intuitive-look-at-fisher-information-2720c40867d8

Gradients infer information about the statistics of underlying manifolds

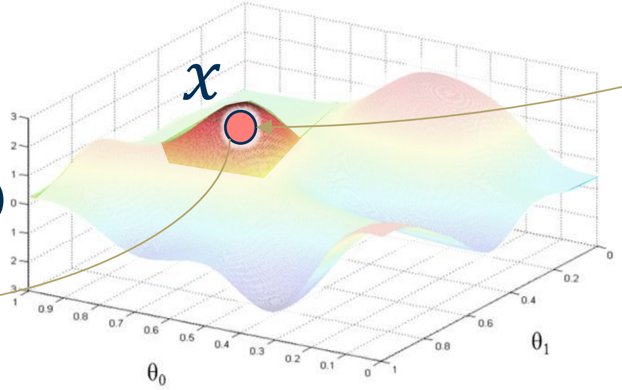
x

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





Local information (specific to x) is sufficient!



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

We demonstrate this in two applications:

1. Visual Explainability
2. Robust Recognition

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

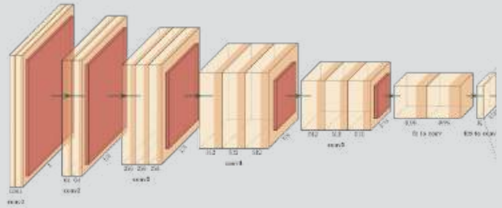
Applicability of Gradient Information

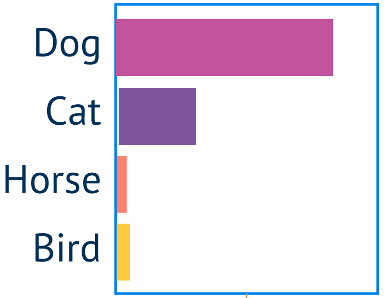
Gradients as Fisher Information

Gradients infer information about the statistics of underlying manifolds

x

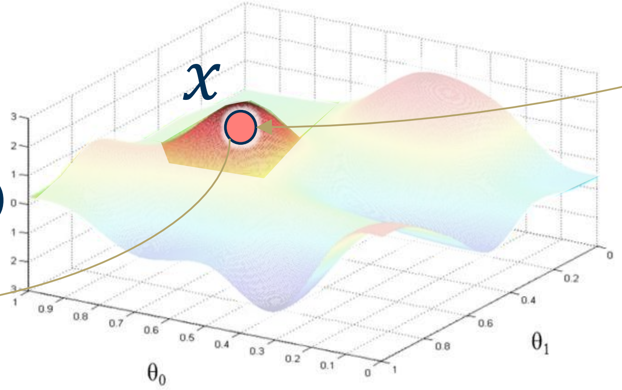
Network $\boldsymbol{f(\theta)}$





Local information (specific to x) is sufficient!

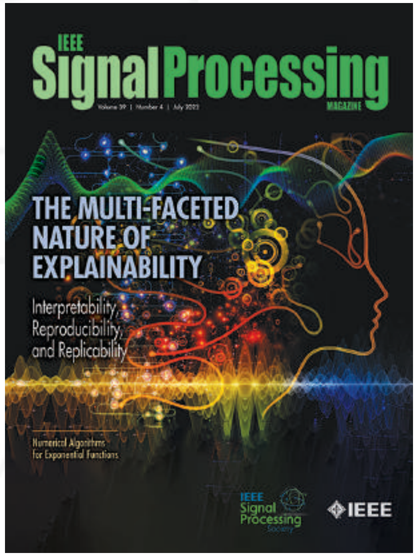
$\boldsymbol{l(\theta|x)}$



We demonstrate this in two applications:

1. Visual Explainability
2. Robust Recognition

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



Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

Mohit Prabhushankar, PhD
Postdoc

Ghassan AlRegib, PhD
Professor

SCAN ME

Explanations

Visual Explanations

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

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

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

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

Explanations

Visual Explanations

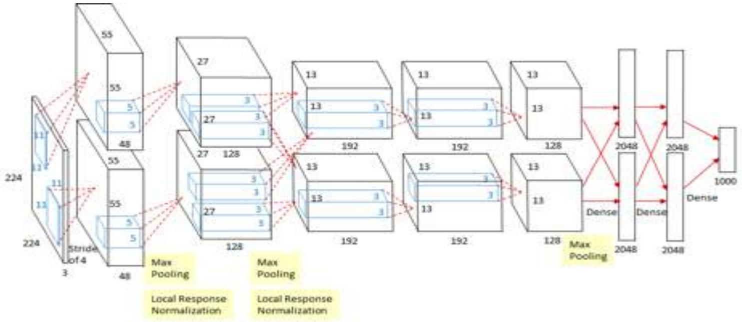
Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

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

Algorithm

Data



Output

class scores

Explainability is useful in*:*

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

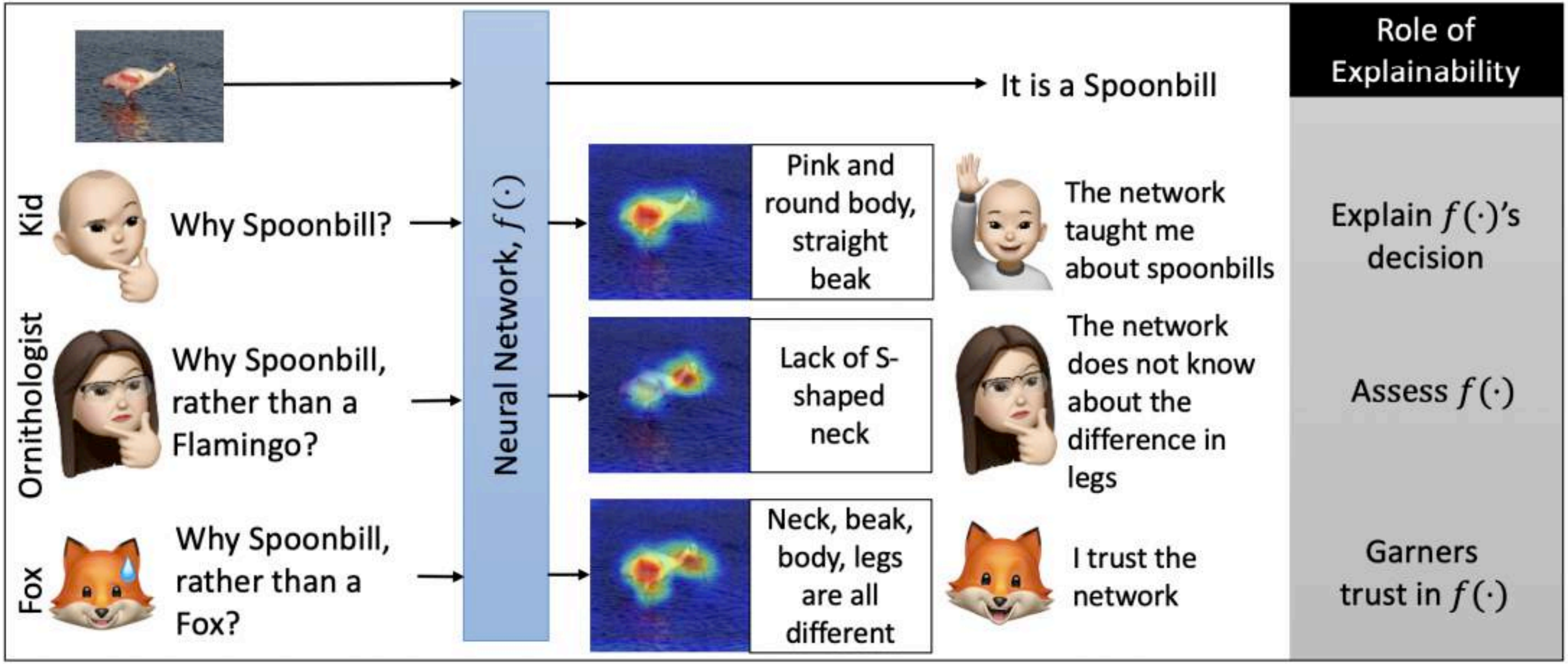
Deep models act as algorithms that take data and output something **without** being able to **explain** their methodology

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

Explanations

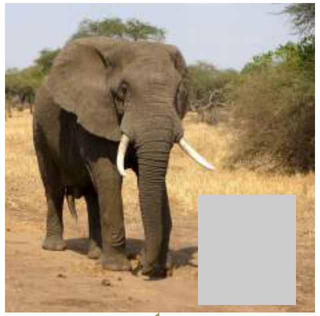
Role of Explanations – context and relevance

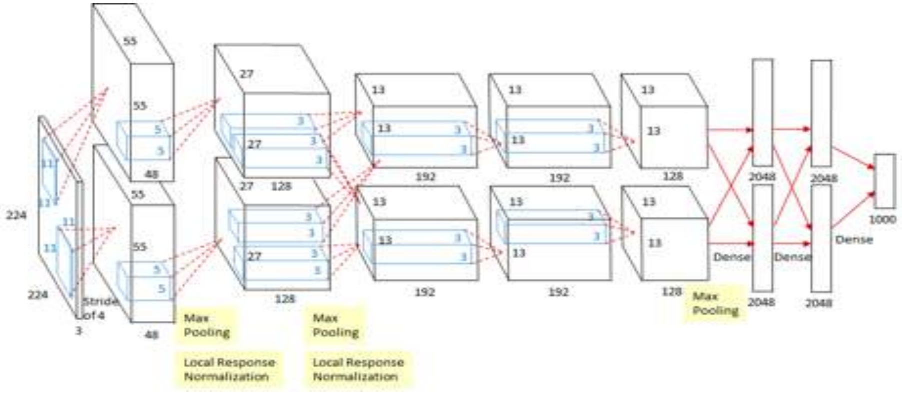
Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations



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

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



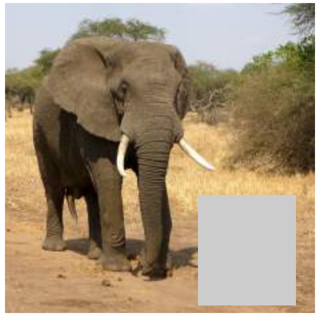


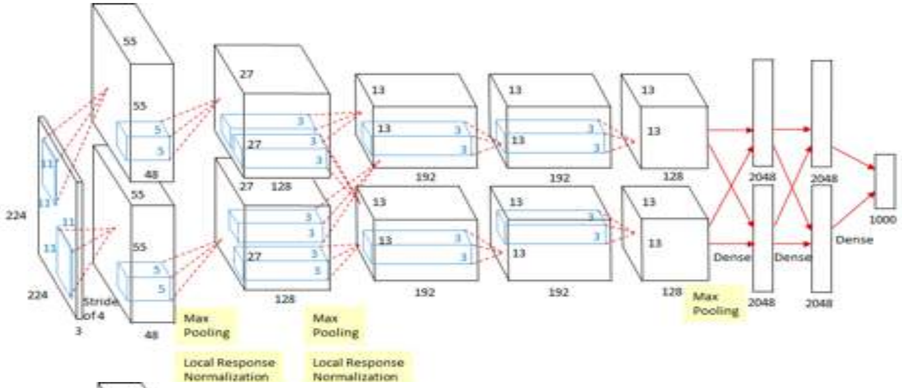
P(elephant) = 0.95

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

Zeiler and Fergus, "Visualizing and Understanding Convolutional Networks", ECCV 2014

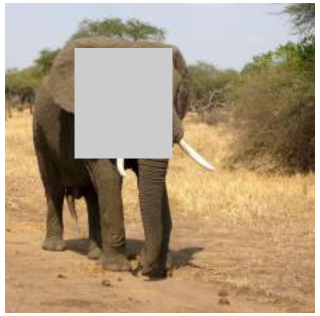
Intervention: Mask part of the image before feeding to CNN, check how much predicted probabilities change

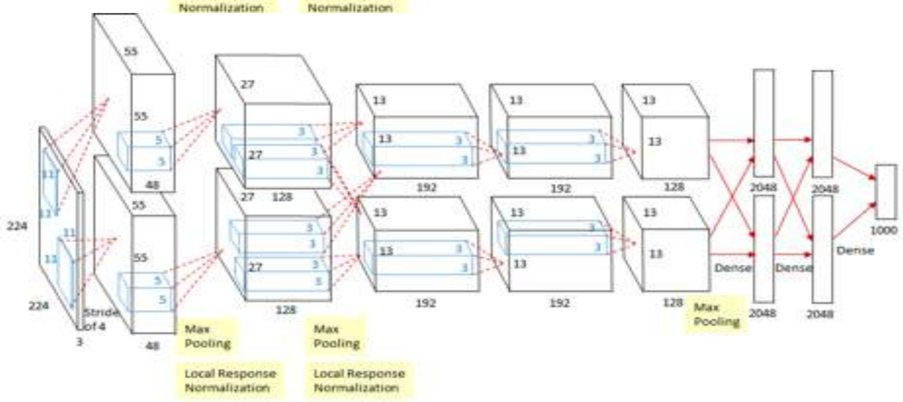




P(elephant) = 0.95

These pixels affect decisions more





P(elephant) = 0.75

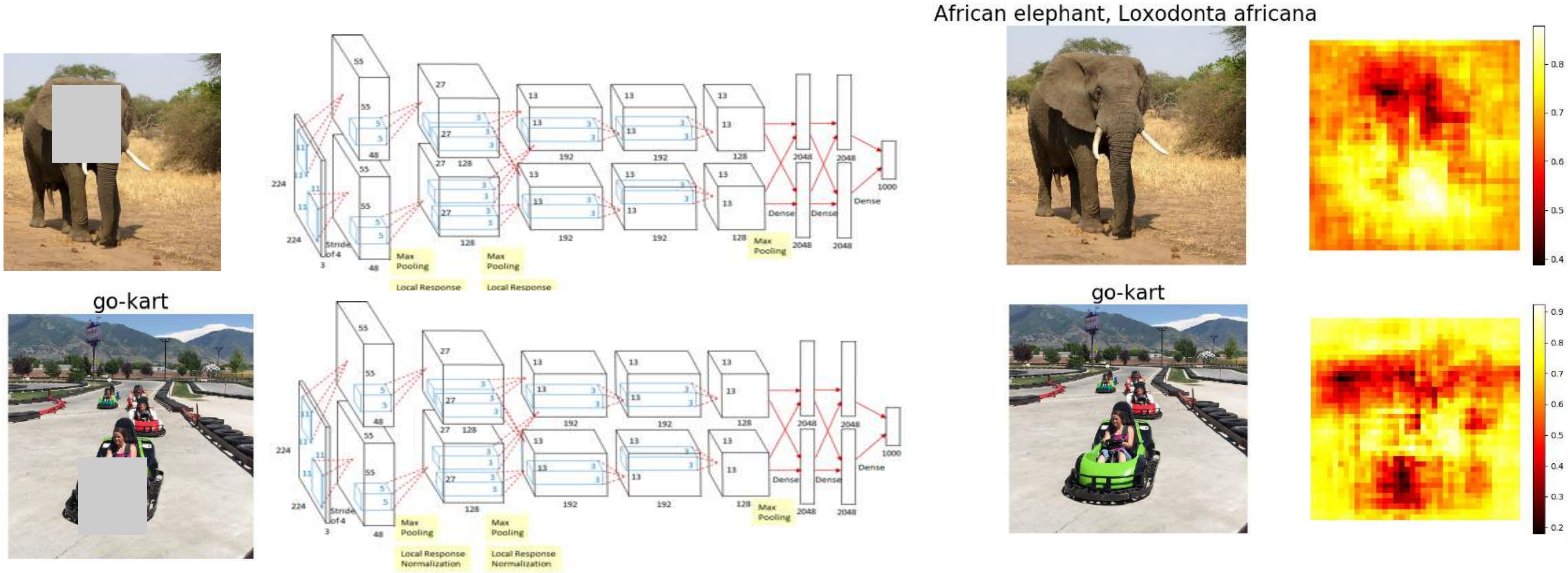
Zeiler and Fergus, “Visualizing and Understanding Convolutional Networks”, ECCV 2014

Input Saliency via Occlusions

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

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



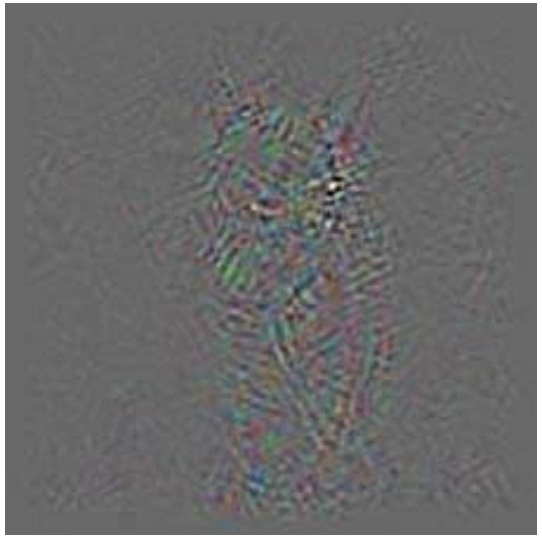
Zeiler and Fergus, “Visualizing and Understanding Convolutional Networks”, ECCV 2014

SCAN ME

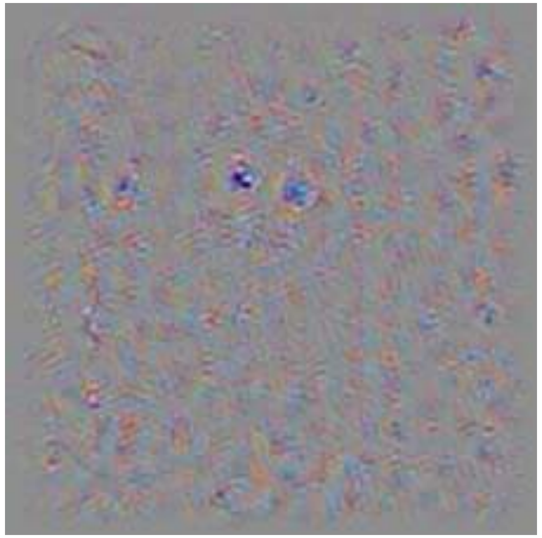
Gradients provide a one-shot means of perturbing the input that changes the output

Input

Vanilla Gradients



Deconvolution Gradients



Guided Backpropagation

However, localization remains an issue

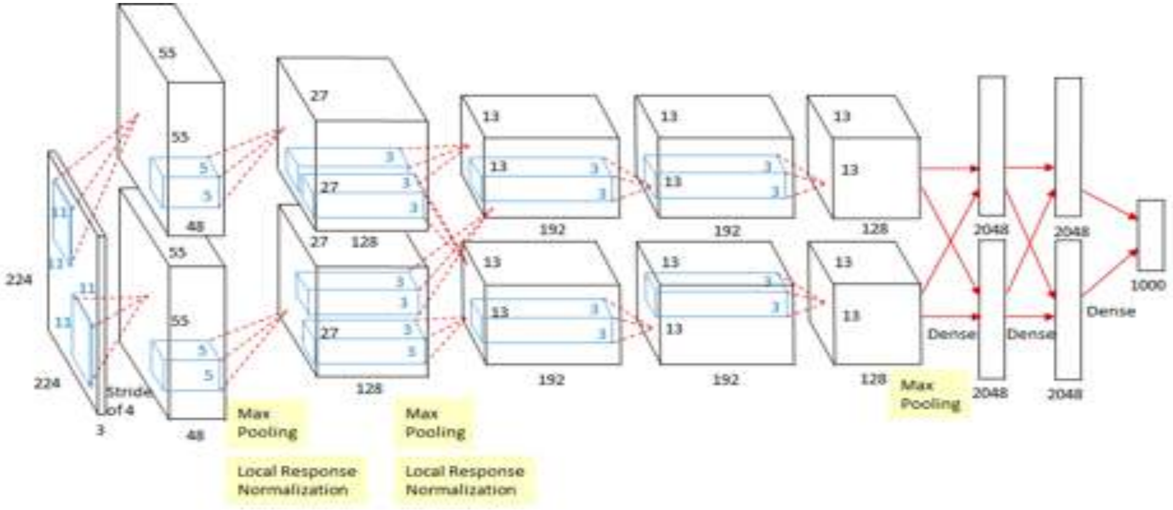
Springenberg, Dosovitskiy, et al., Striving for Simplicity: The all convolutional net, 2015

Gradient and Activation-based Explanations

GradCAM

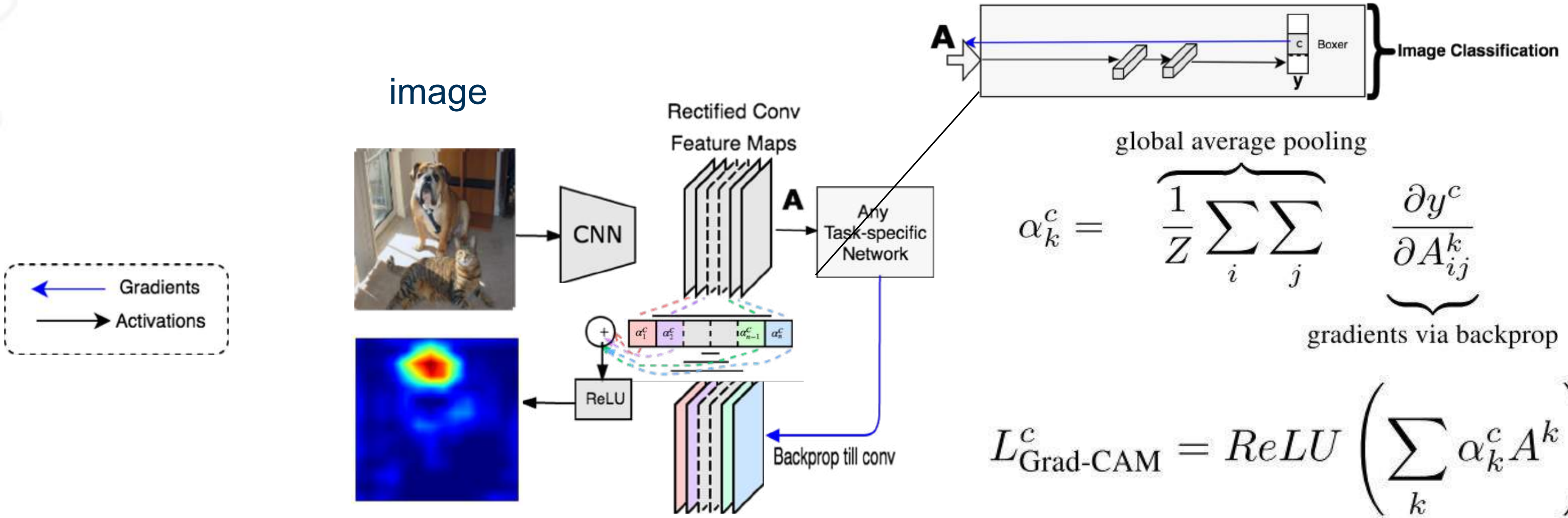
Gradients provide a one-shot means of perturbing the input that changes the output. Activations provide the localization.

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



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

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



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

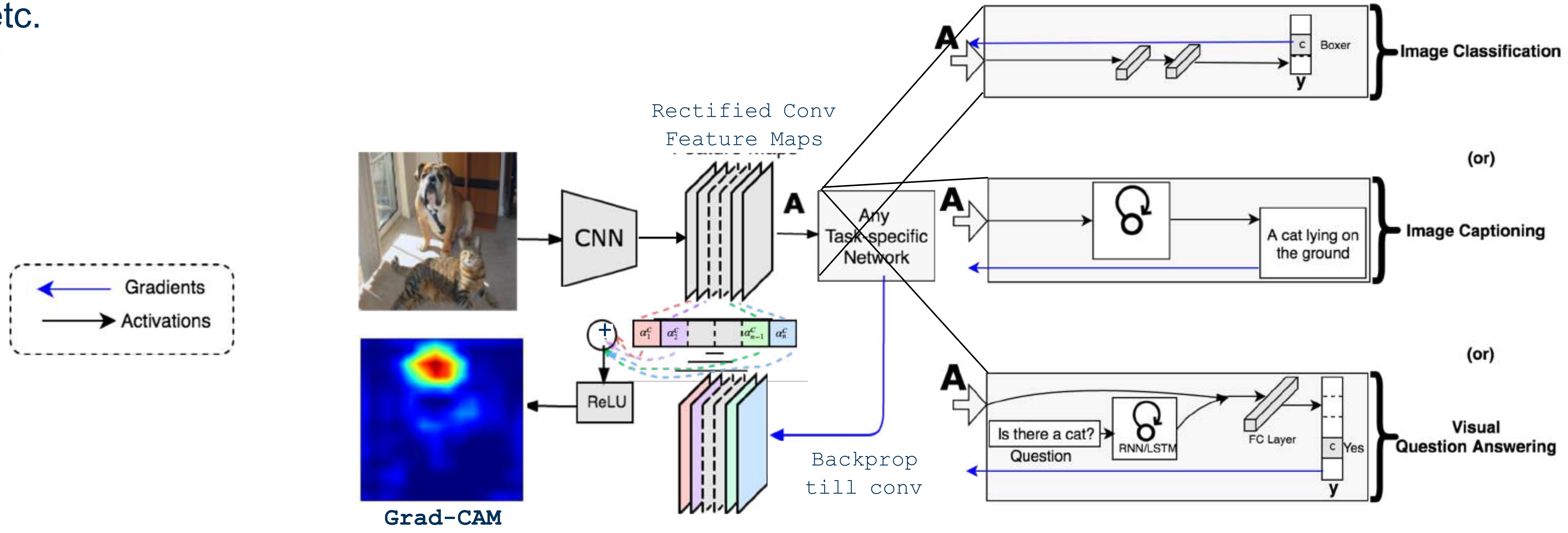
$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

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

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

Gradient and Activation-based Explanations
GradCAM

Grad-CAM generalizes to any task:
- Image classification
- Image captioning
- Visual question answering
- etc.



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

SCAN ME
Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

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

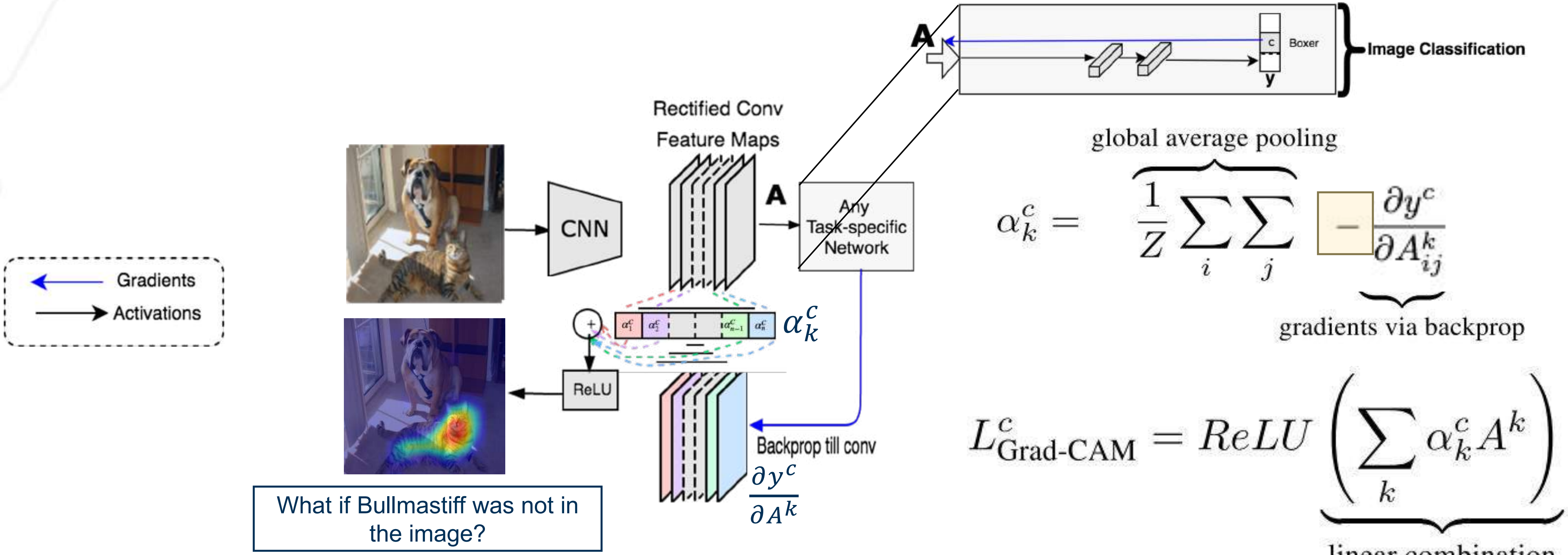
Bullmastiff	Why Bullmastiff?	What if Bullmastiff was not in the image?	Why Bullmastiff, rather than a Boxer?

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?

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



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

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

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.

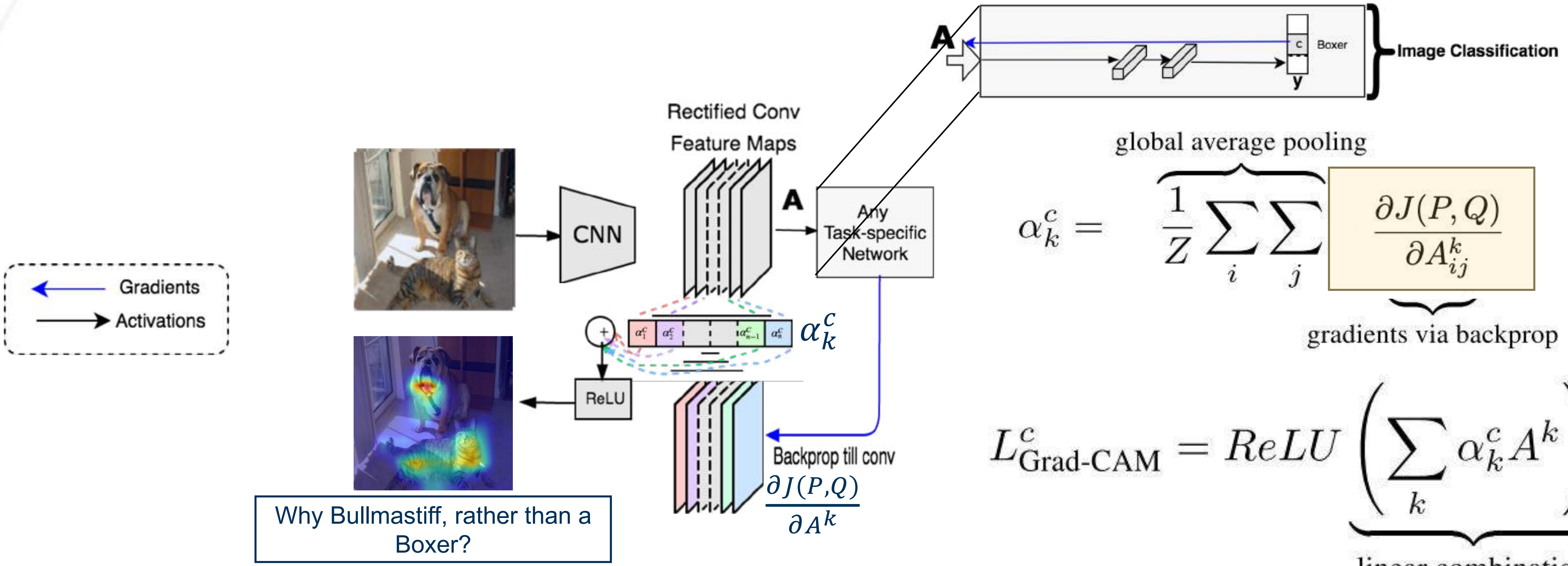
Gradient and Activation-based Explanations

ContrastCAM: Why P, rather than Q?

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

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



Contrast-CAM

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

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

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

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

ContrastCAM

Toy Manifold Example

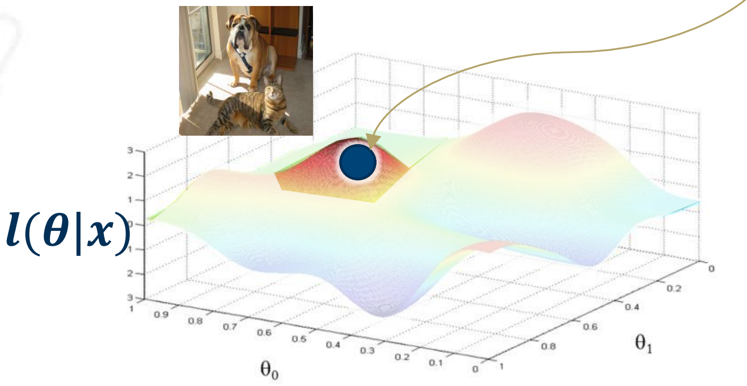
Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

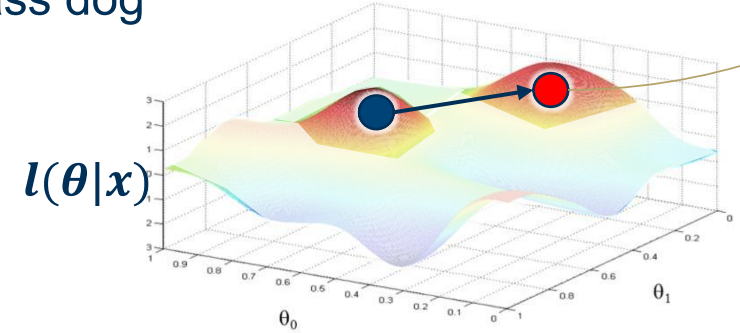
SCAN ME

The contrast classes are unlikely, but the gradients provide information about contrast classes

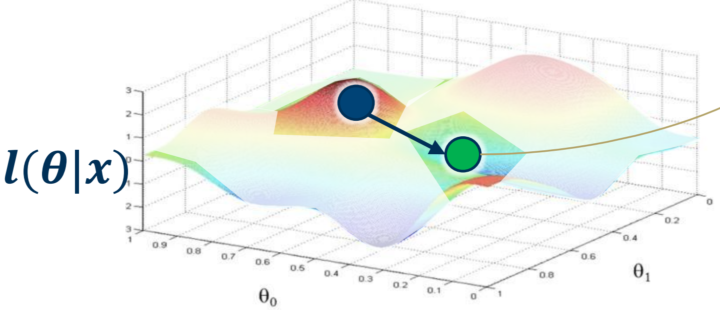
Likelihood of a dog predicted as class cat

Likelihood of a dog predicted as class dog





Likelihood of a dog predicted as class horse

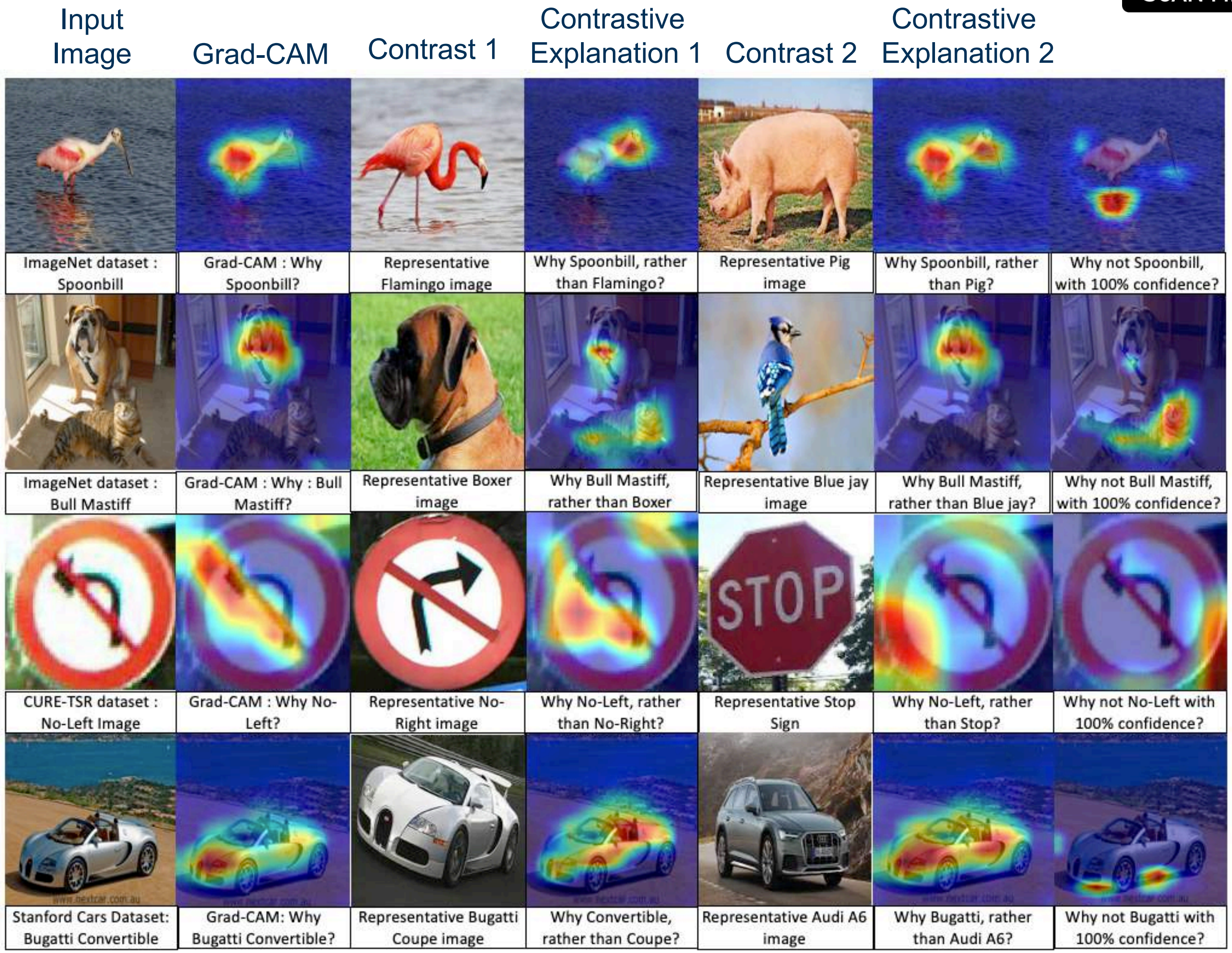


Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

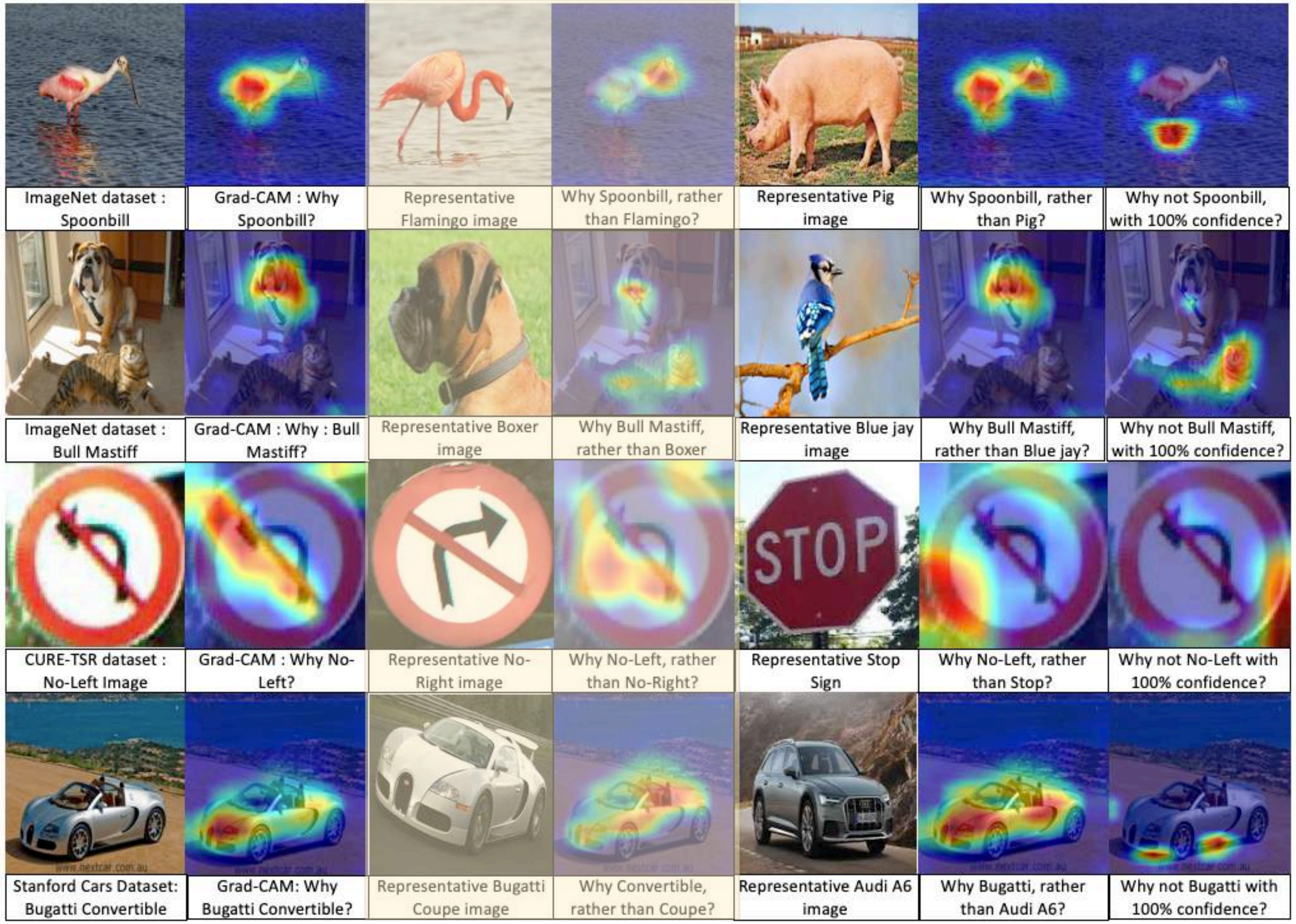


Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

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



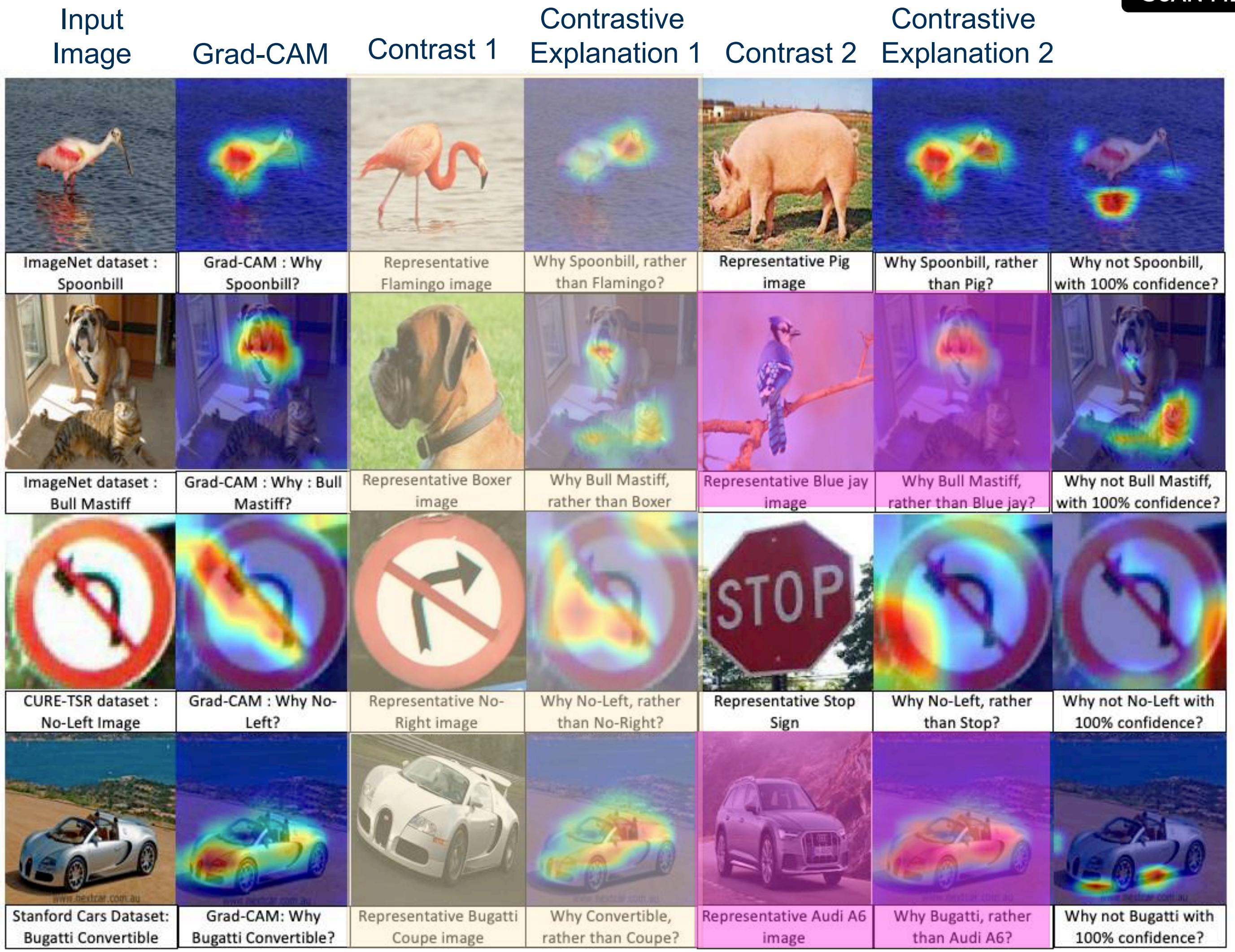
Human Interpretable

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.

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

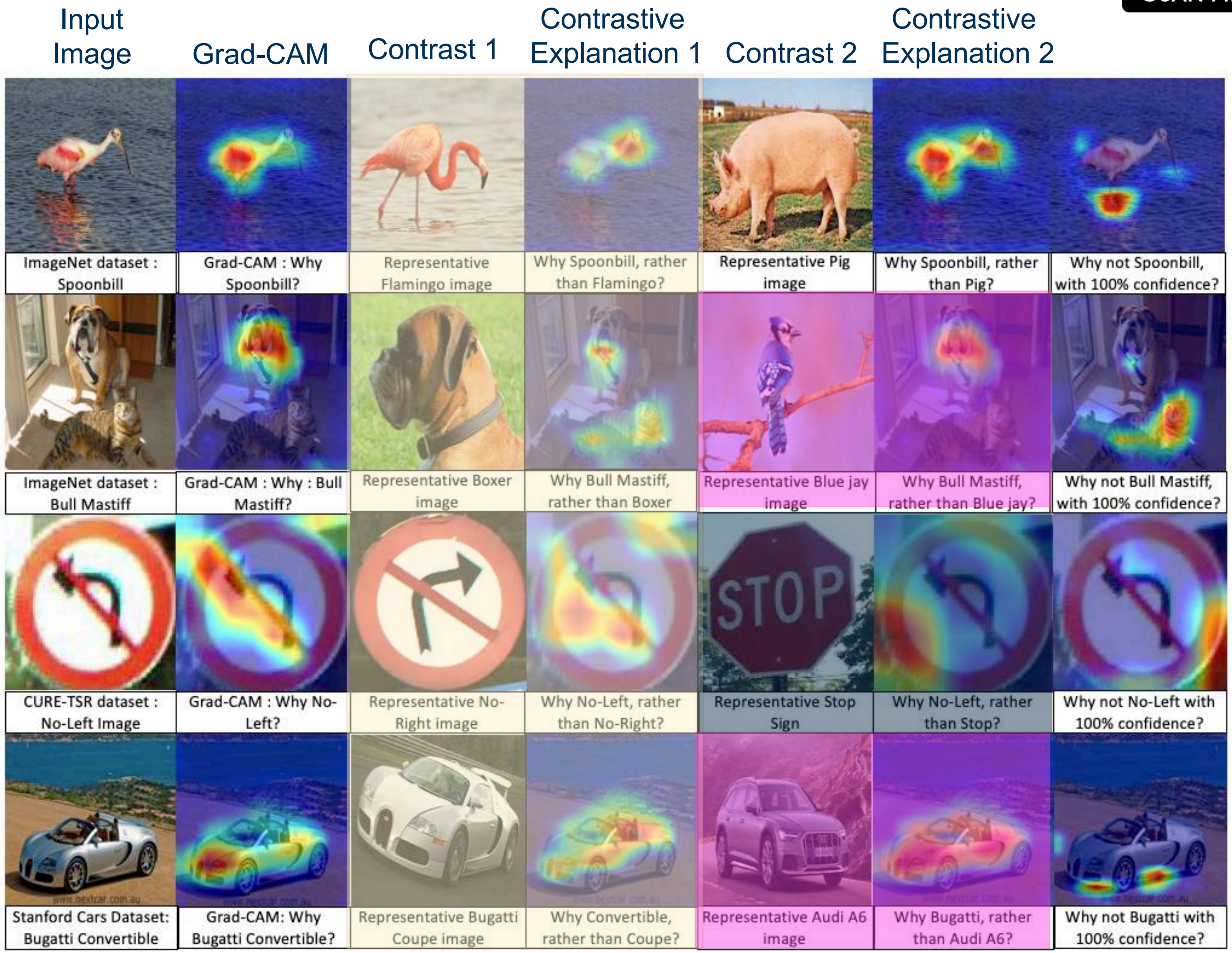
Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME



Human Interpretable

Same as Grad-CAM

Not Human Interpretable

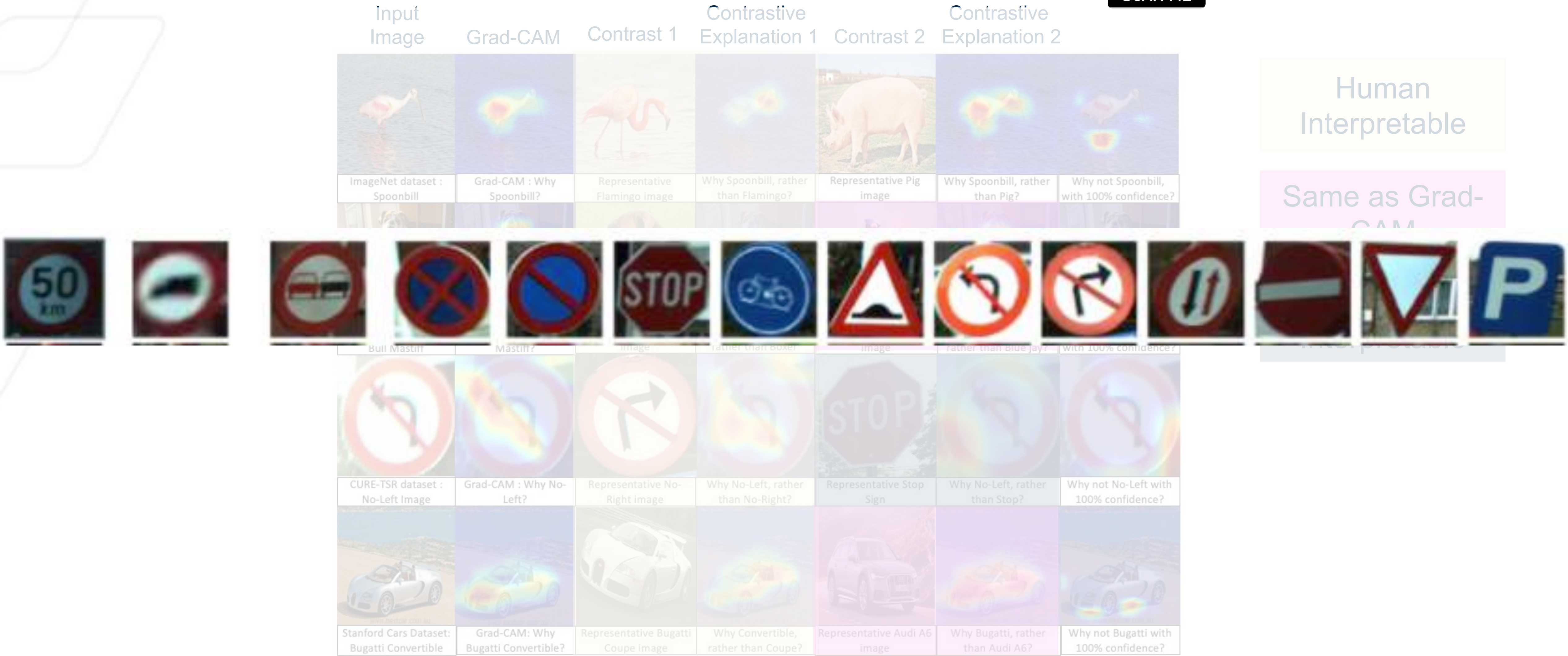
Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME



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

SCAN ME

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

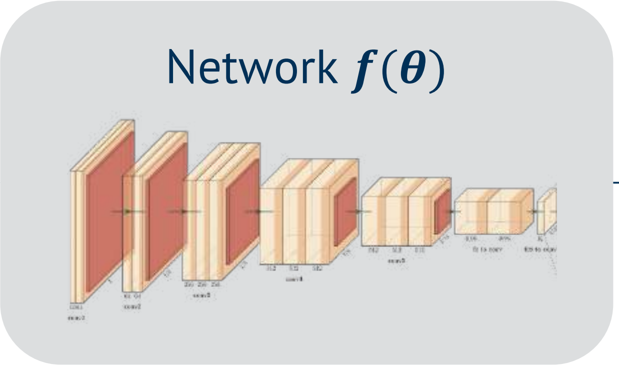
Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

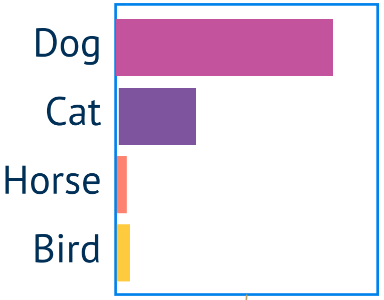
Applicability of Gradient Information

Gradients as Fisher Information

Gradients infer information about the statistics of underlying manifolds

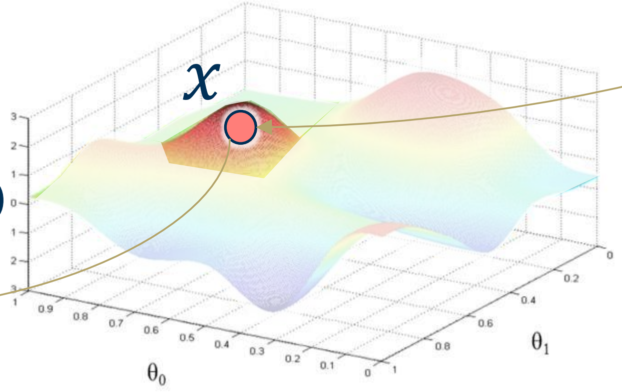
x





Local information (specific to x) is sufficient!

$l(\theta|x)$



We demonstrate this in two applications:

1. Visual Explainability
2. Robust Recognition

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

NEURAL INFORMATION PROCESSING SYSTEMS

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

Mohit Prabhushankar, PhD
Postdoc

Ghassan AlRegib, PhD
Professor

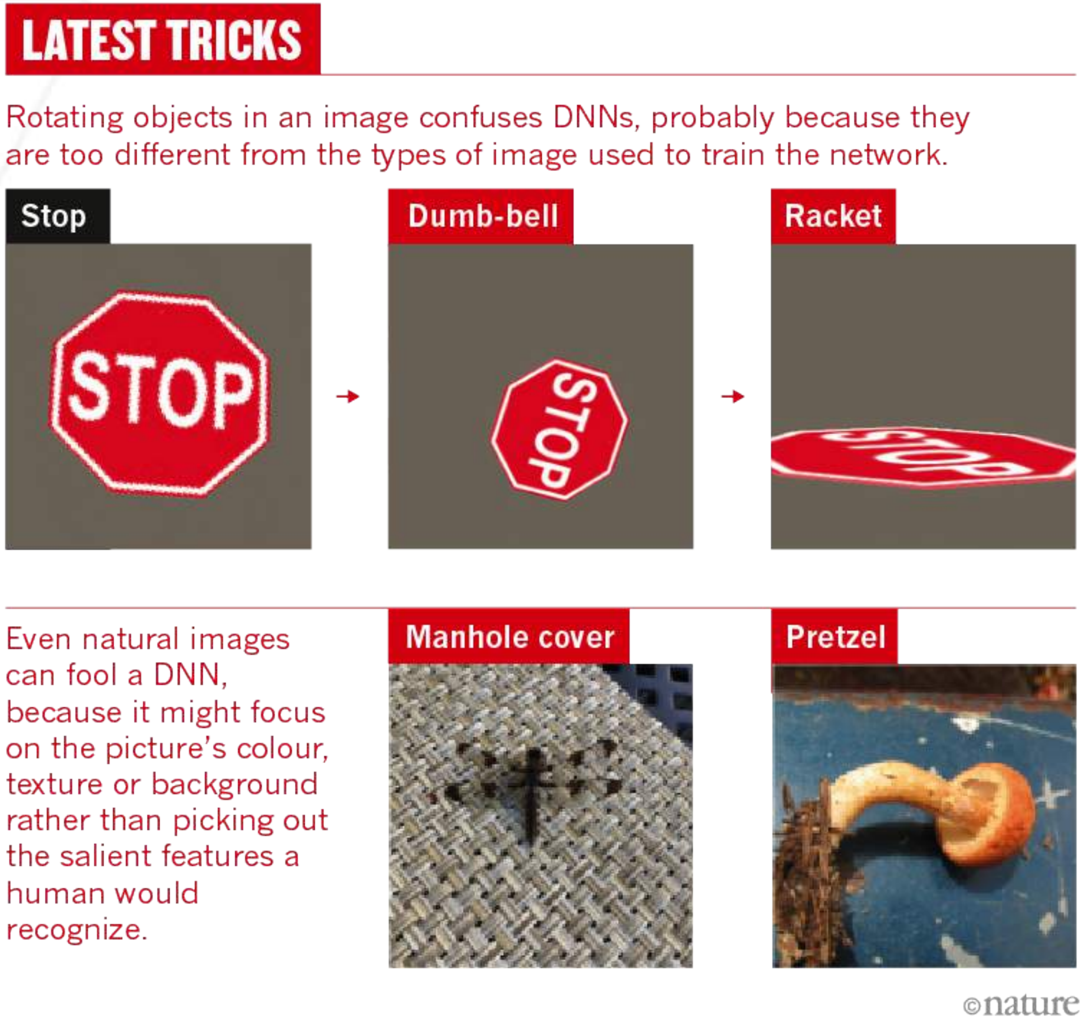
SCAN ME

Robustness in Neural Networks

Why Robustness?

SCAN ME

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



Why Robustness?

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

SCAN ME

How would humans resolve this challenge?

We Introspect!

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

SCAN ME

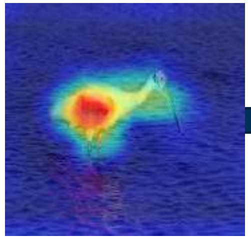
Introspective Learning: A Two-stage Approach for Inference in Neural Networks

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

Introspection

Visual Sensing

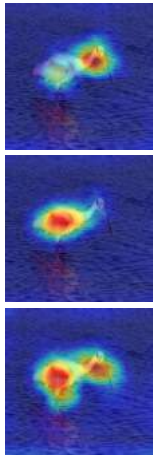
Sense pink feathers, straight beak



Spoonbill
$\hat{y}$

Feed-Forward Sensing

Reflection



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

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

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

Spoonbill
$\tilde{y}$

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.

SCAN ME

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

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

Goal : To simulate Introspection in Neural Networks

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

What are the possible targeted questions?

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

Introspection

Introspection in Neural Networks

SCAN ME

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

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

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

What are the possible targeted questions?

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

SCAN ME

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

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

Goal : To simulate Introspection in Neural Networks

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

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

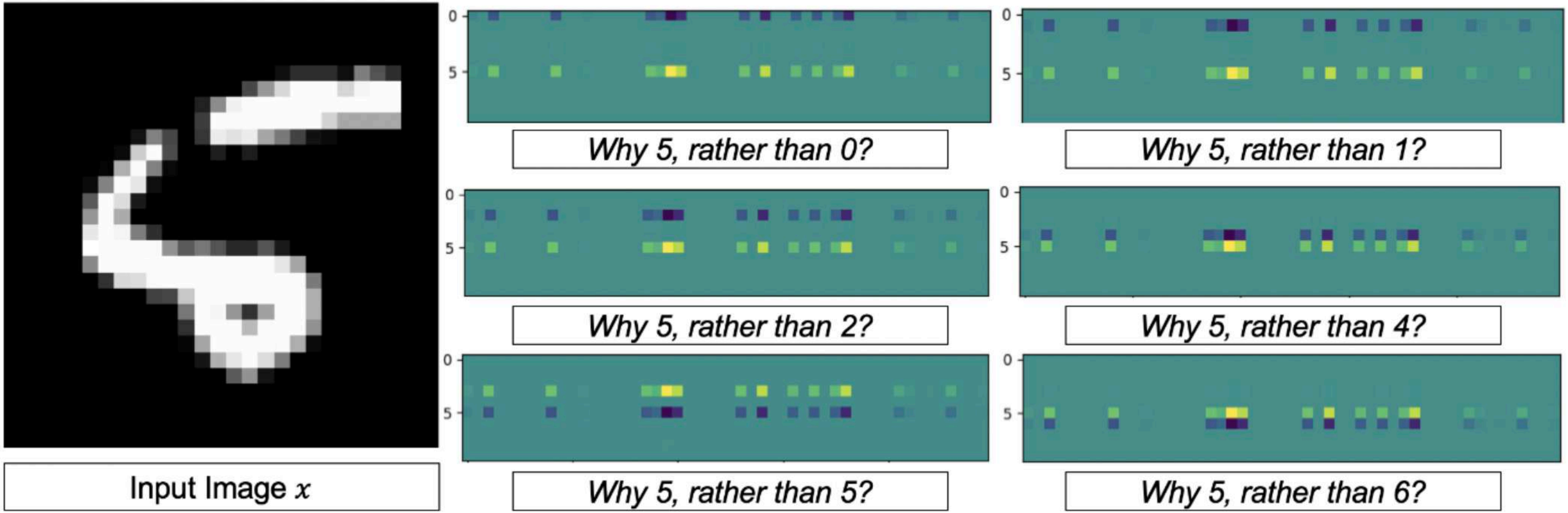
M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Introspection
Gradients as Features

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

SCAN ME

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



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

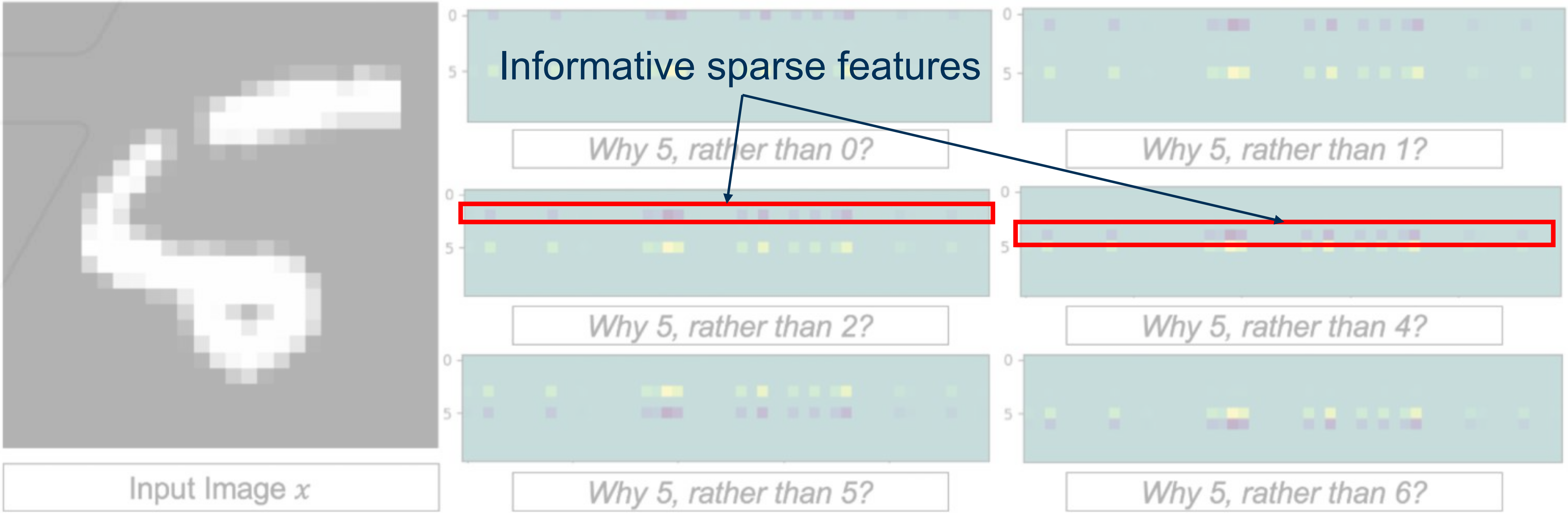
Introspection

Gradients as Features

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

SCAN ME

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



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

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

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

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

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

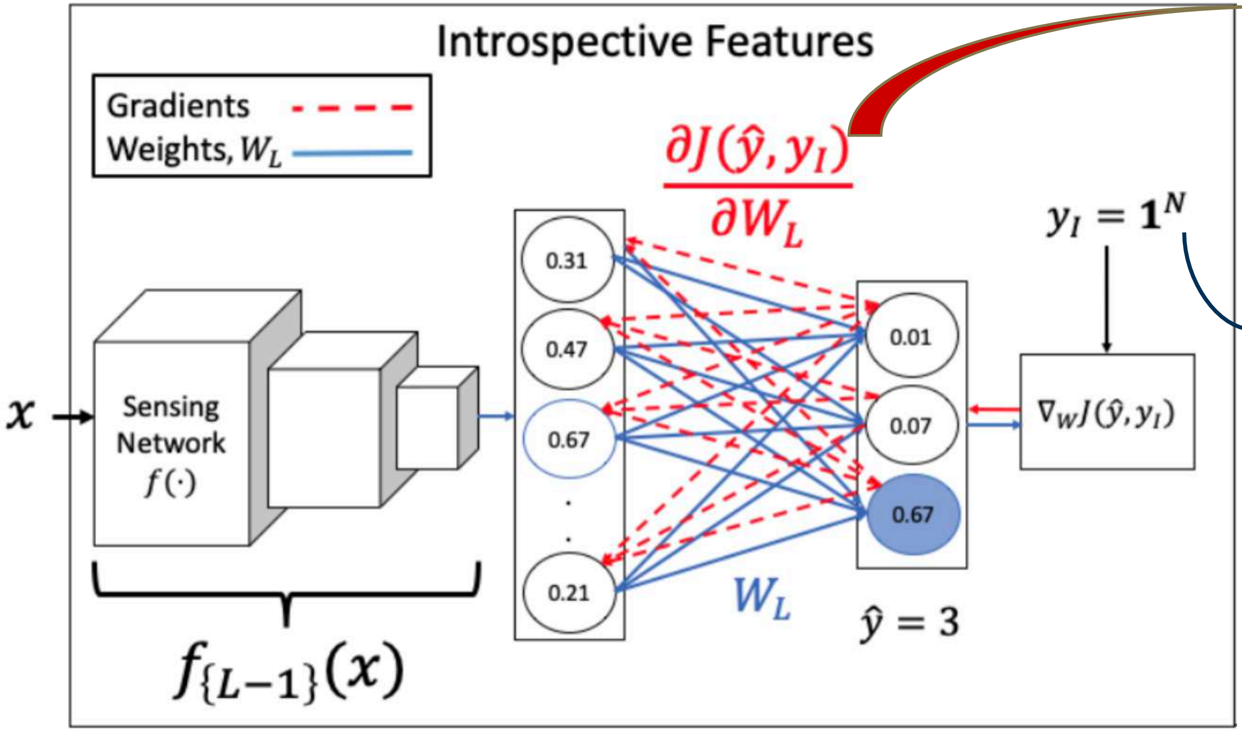
M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Introspection

Deriving Gradient Features

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

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



Normalized and vectorized gradients are introspective features

Vector of all ones: A confounding label!

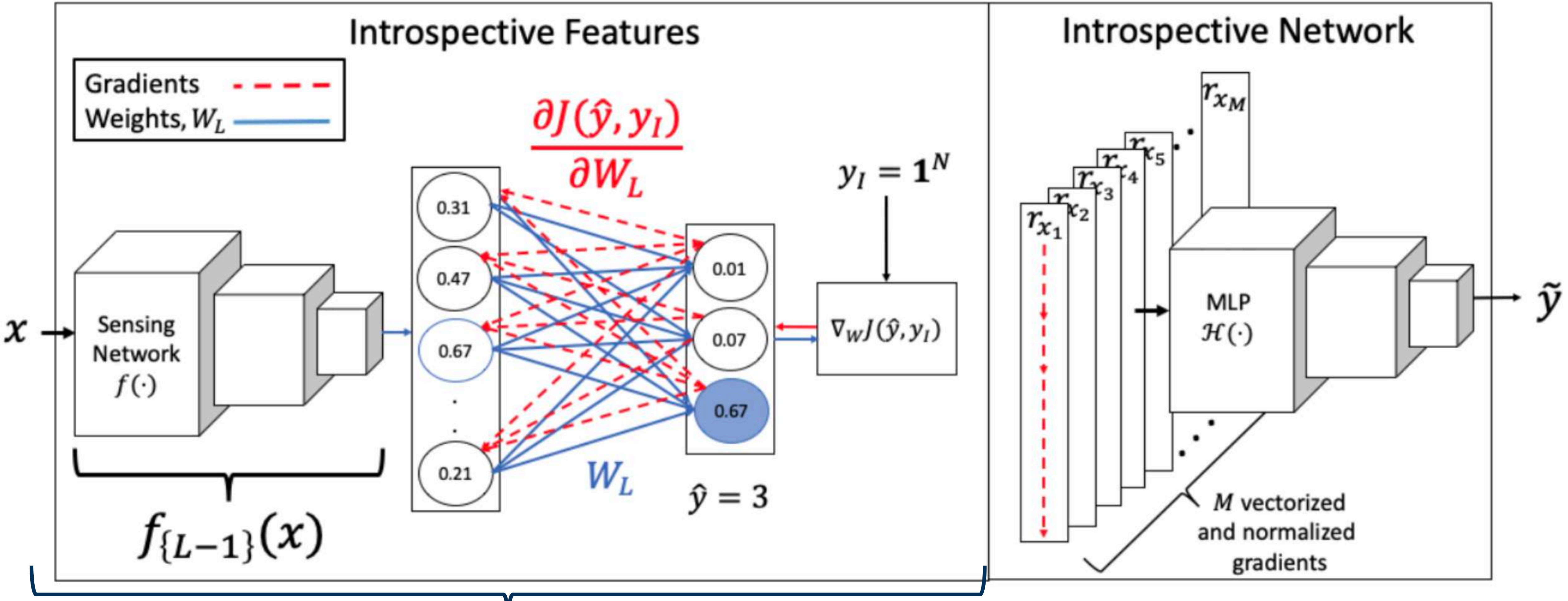
M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Introspection

Utilizing Gradient Features

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

SCAN ME



Introspective Features

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

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

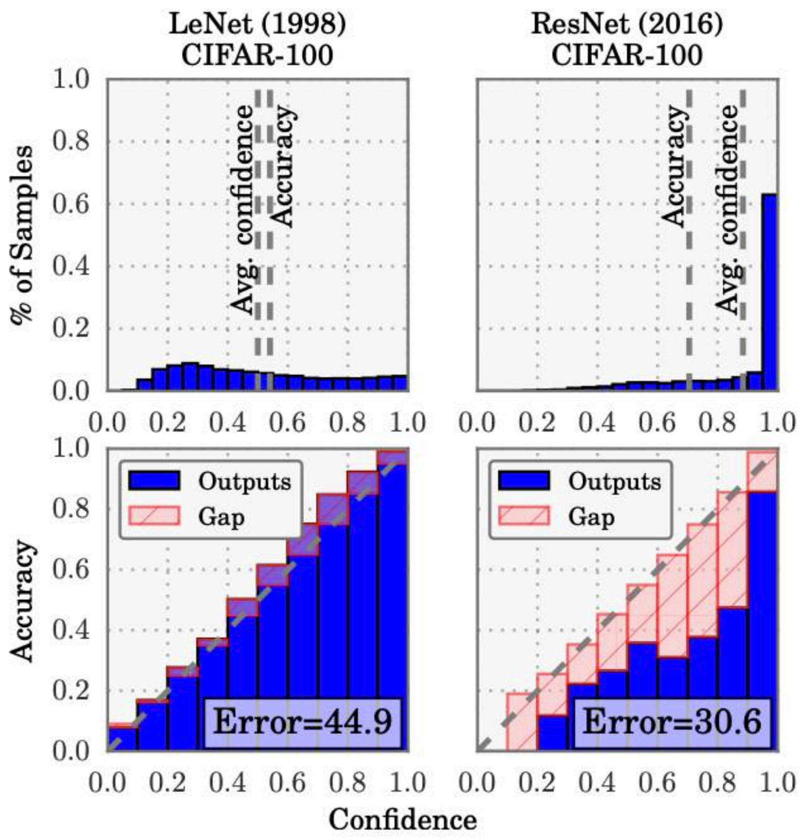
M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

A note on Calibration..

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

SCAN ME

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



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

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

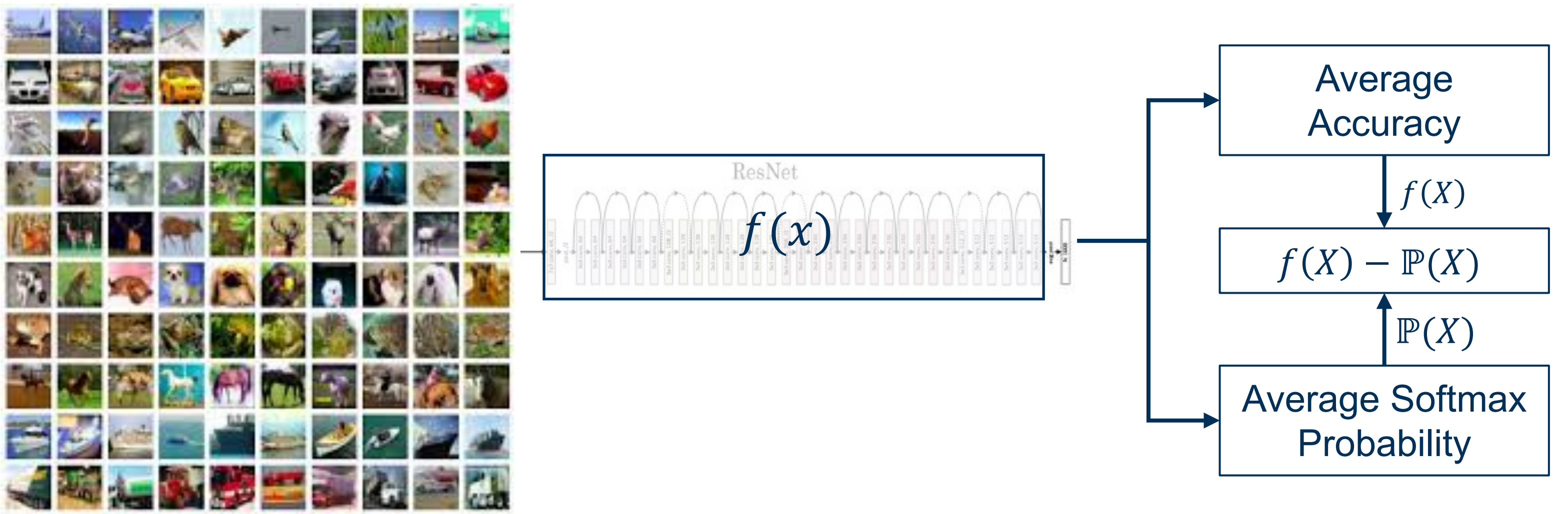
Calibration

A note on Calibration..

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

SCAN ME

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



CIFAR-10 Testset

Bin-wise subtraction to obtain gaps

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

Introspection in Neural Networks
Generalization and Calibration results

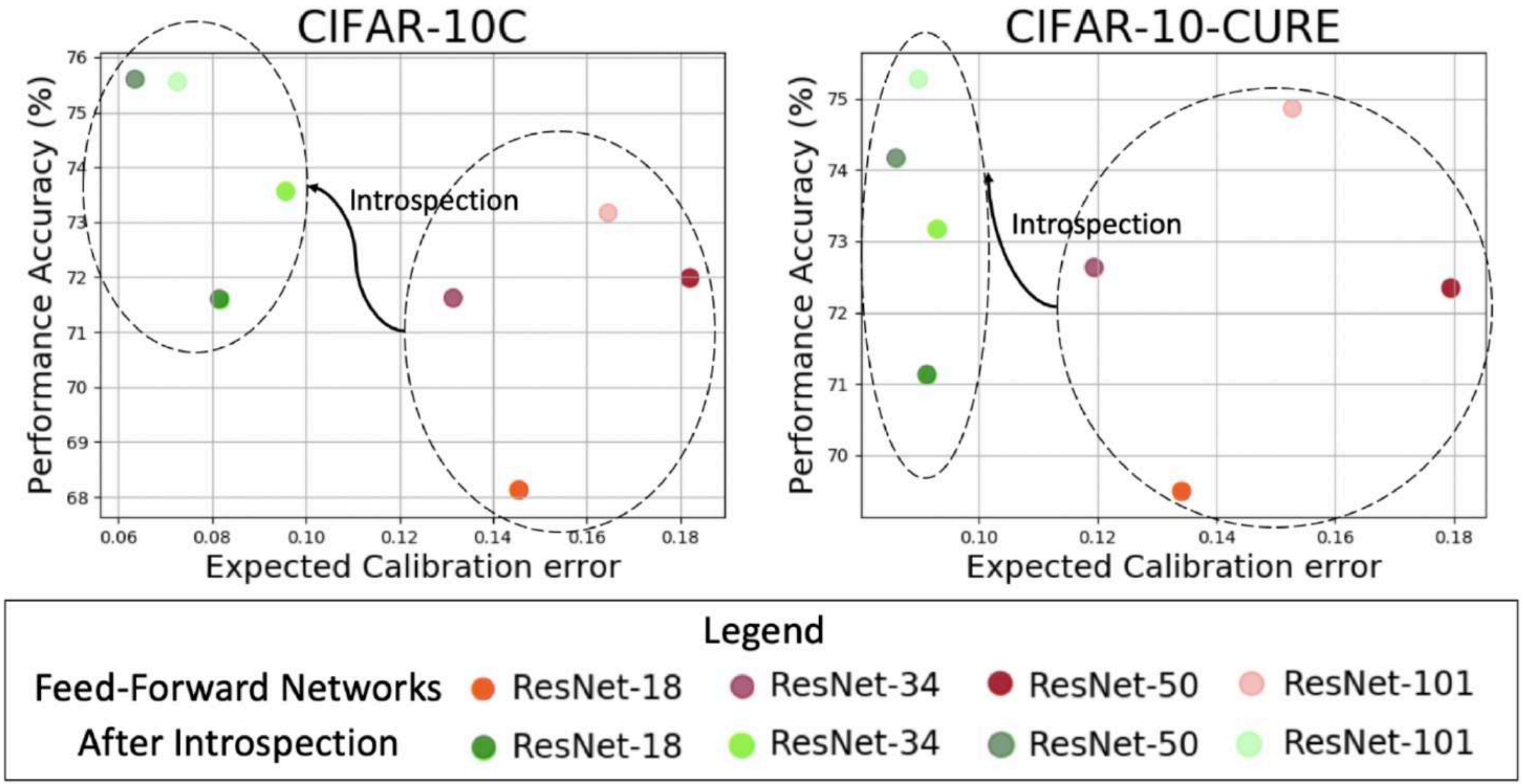
Introspective Learning: A Two-stage Approach for Inference in Neural Networks

SCAN ME

Ideal: Top-left corner

Y-Axis: Generalization

X-Axis: Calibration



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

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.

Introspection in Neural Networks

Plug-in nature of Introspection

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

SCAN ME

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.

Objectives

Takeaways from Part II

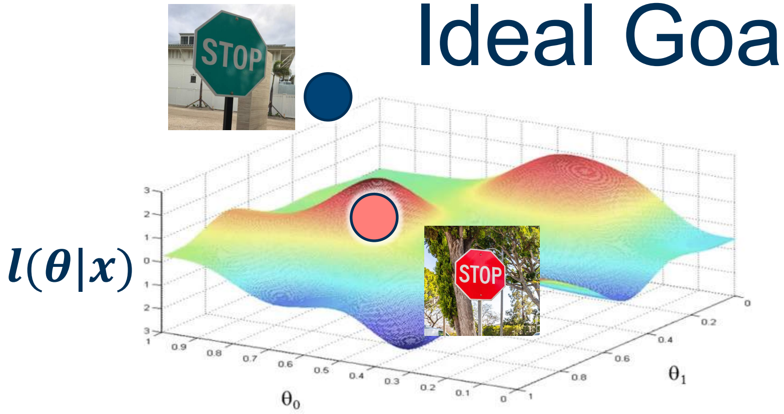
- Part I: Gradients in Neural Networks
- **Part 2: Gradients as Information**
    - Gradients approximate Fisher Information: They provide a methodology to infer information about the statistics of underlying manifolds using samples
    - Fisher information in gradients allow them to be utilized in explanations
    - The versatile information encoded in gradients allow for visualizing correlations, counterfactuals, and contrastives within the same GradCAM framework
    - Contrastive information can be used to train a second stage that is more robust under noise conditions in Introspective Learning
- Part 3: Gradients as Uncertainty
- Part 4: Gradients as Expectancy-Mismatch
- Part 5: Conclusion and Future Directions

Part I and Part II

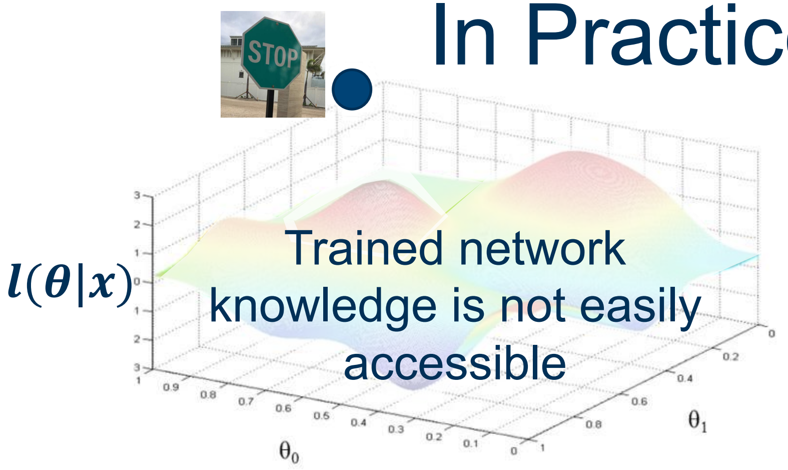
Tying it Back

From Part I

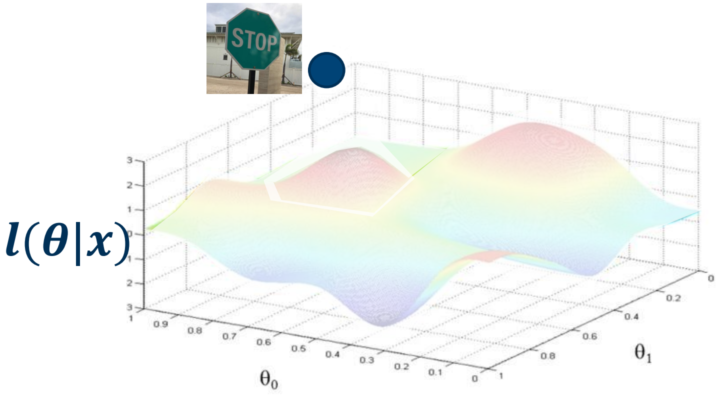
Ideal Goal



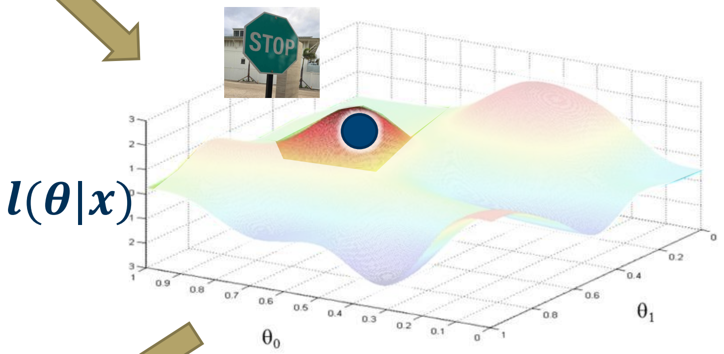
In Practice

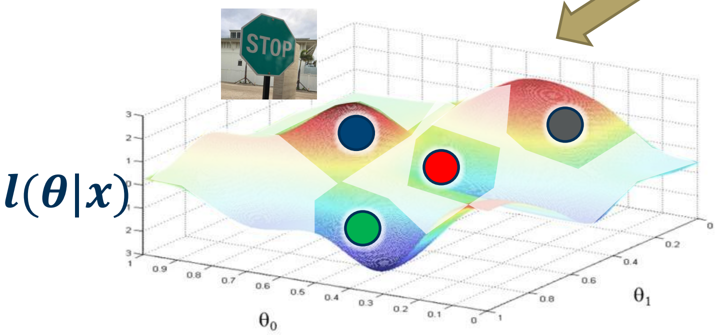


In Part II



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





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

Interpretation, and Applications of Gradients

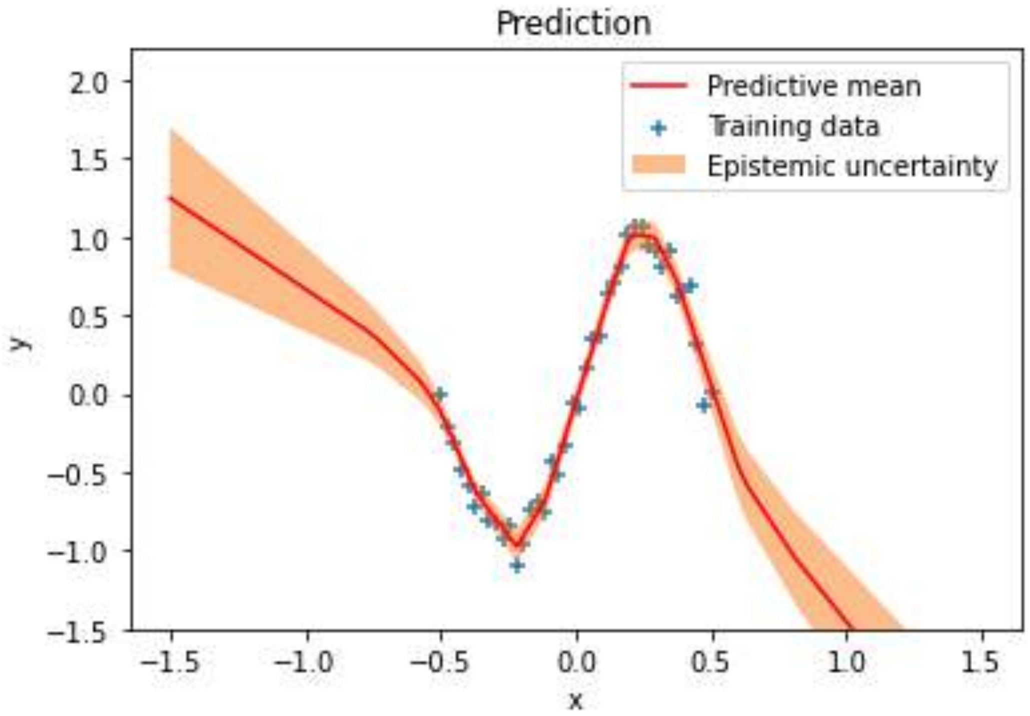
Part 3: Gradients as Uncertainty

Objectives

Objectives in Part 3

- Interpret gradients as Uncertainty
- Uncertainty Applications
  - Anomaly Detection
  - Out-of-Distribution Detection
  - Adversarial Image Detection
  - Corruption Detection

Uncertainty is a model knowing that it does not know



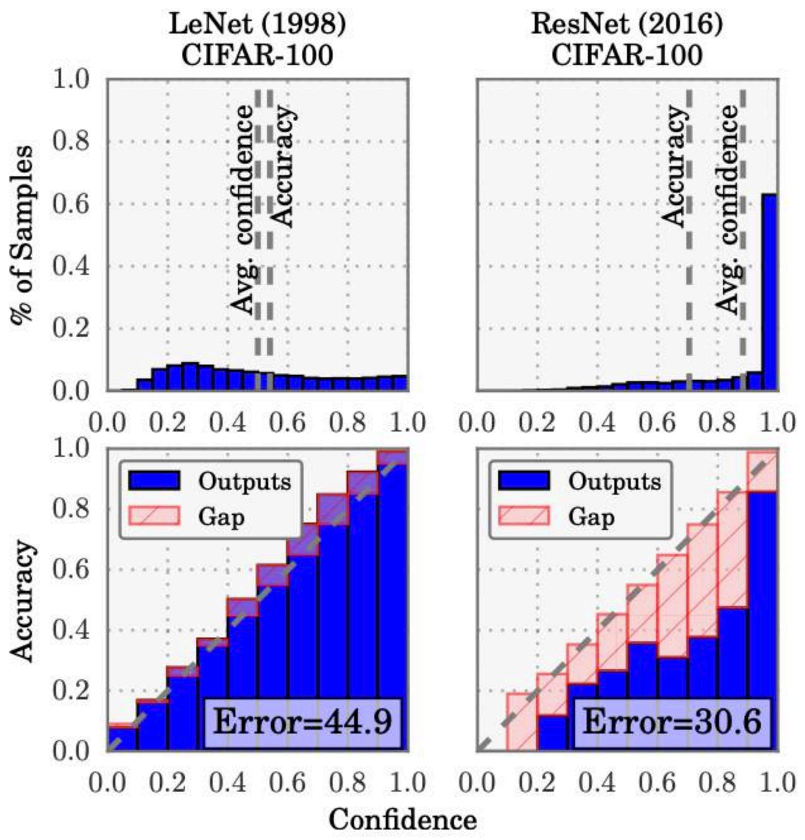
A simple example: More the training data, lesser the uncertainty

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

Uncertainty

When is Uncertainty an Issue?

Uncertainty is a model knowing that it does not know



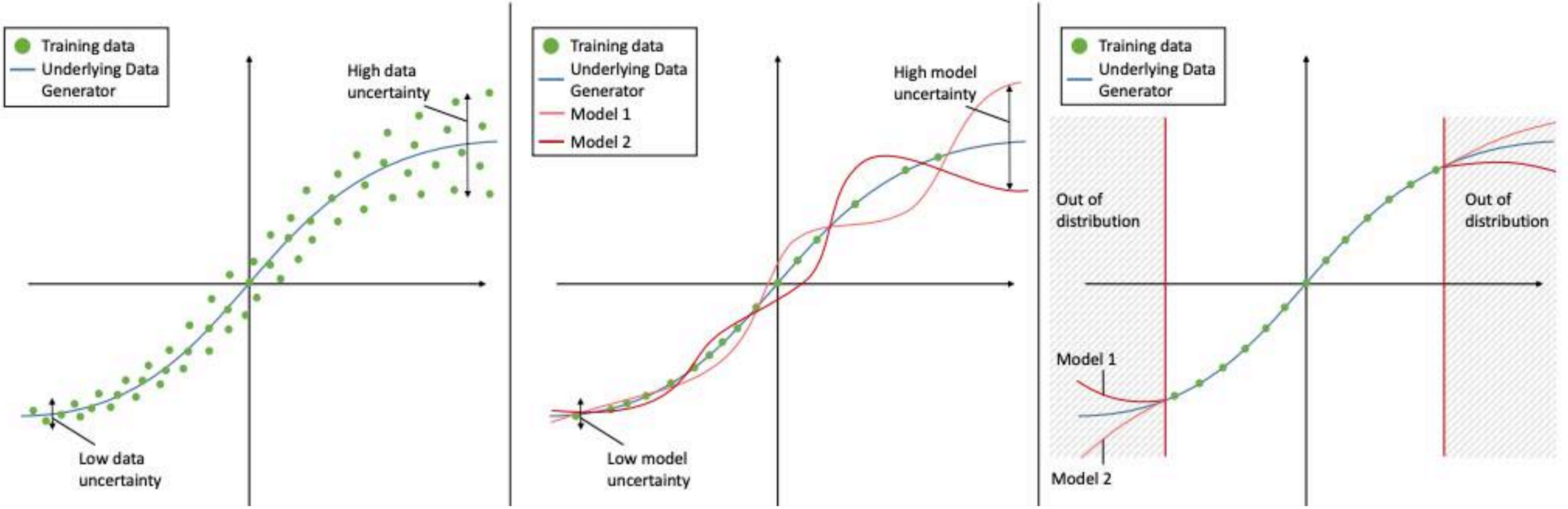
- Larger the model, more misplaced is a network's confidence
- On ResNet, the gap between prediction accuracy and its corresponding confidence is significantly high
- On OOD data, uncertainty is not easy to quantify

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

Uncertainty

Two Types of Uncertainty

Two major types of uncertainty: Uncertainty in data and uncertainty in model, together termed as prediction Uncertainty

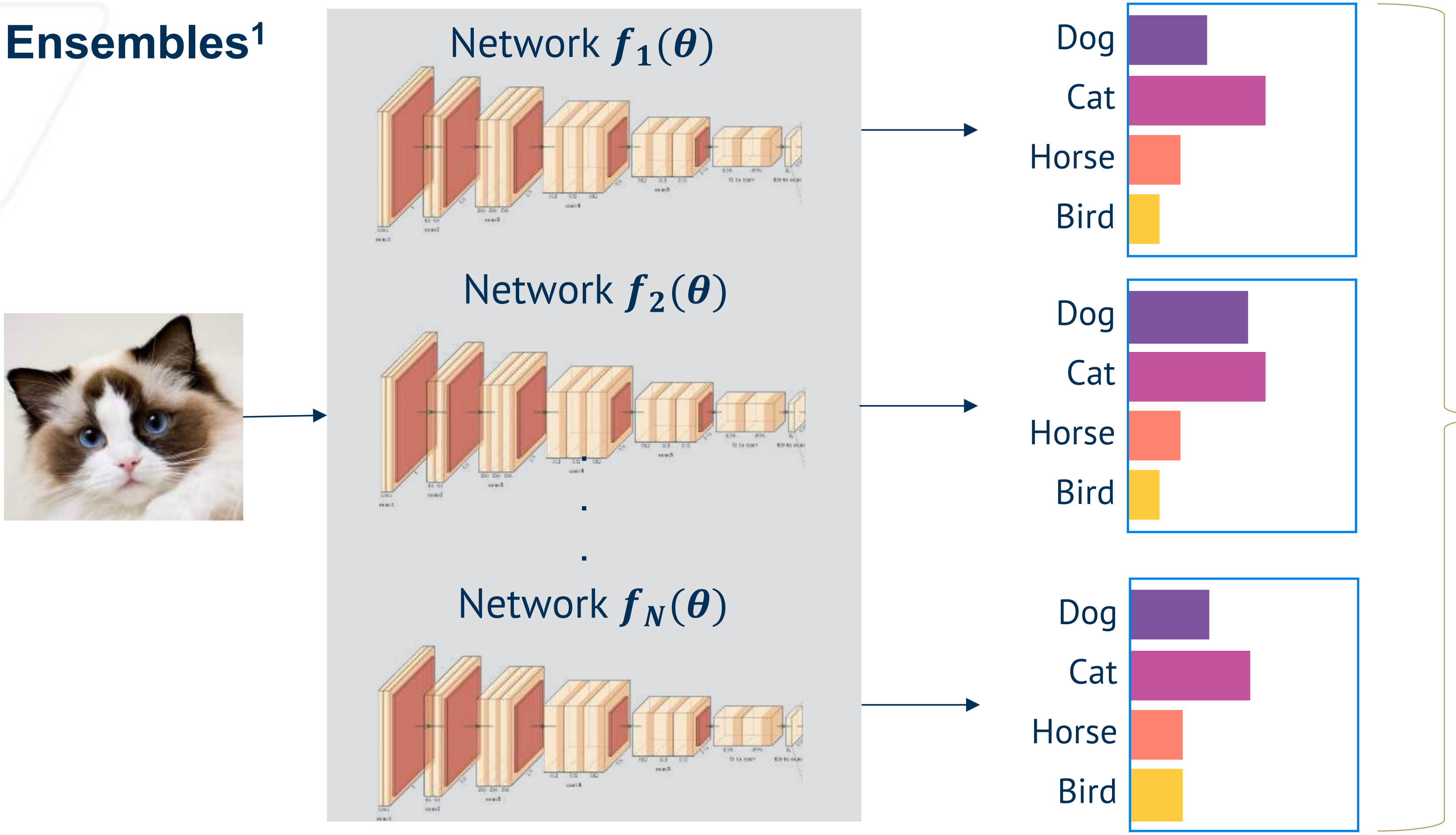


Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., ... & Zhu, X. X. (2021). A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*.

Uncertainty

Uncertainty Quantification in Neural Networks

Via Ensembles[1]



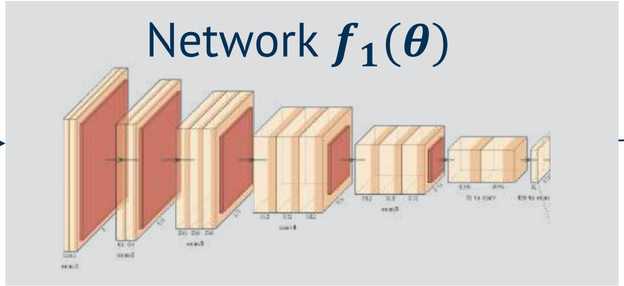
Variation within outputs $Var(y)$ is the uncertainty. Commonly referred to as **Prediction Uncertainty.**

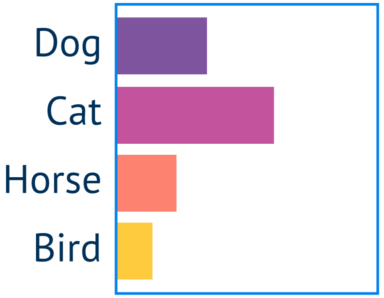
[1] Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles." *Advances in neural information processing systems* 30 (2017).

Uncertainty

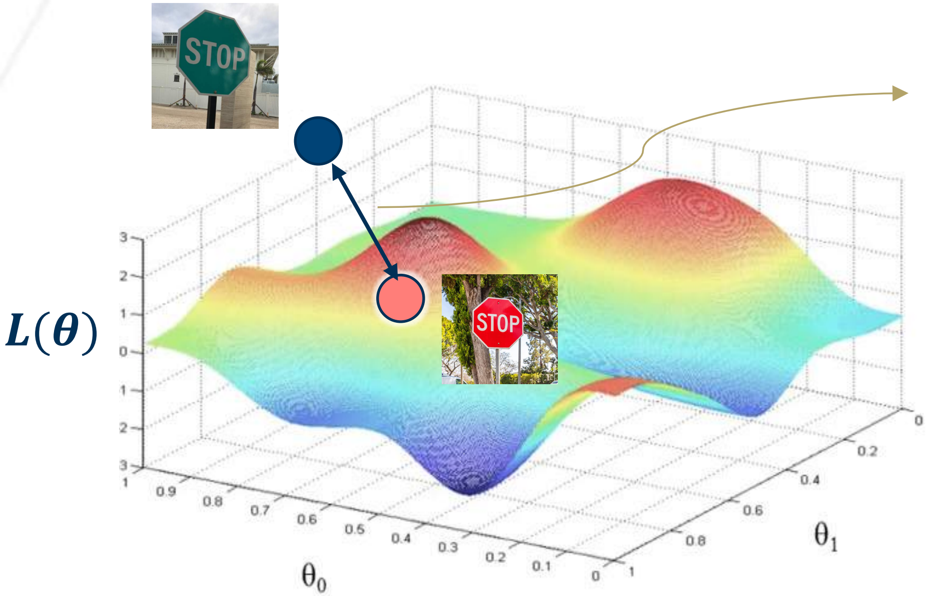
Uncertainty Quantification in Neural Networks

Via Single pass methods[1]





Uncertainty quantification using a single network and a single pass



Calculate distance from some trained clusters

Does not require multiple networks!

However, does requires multiple data points at inference!

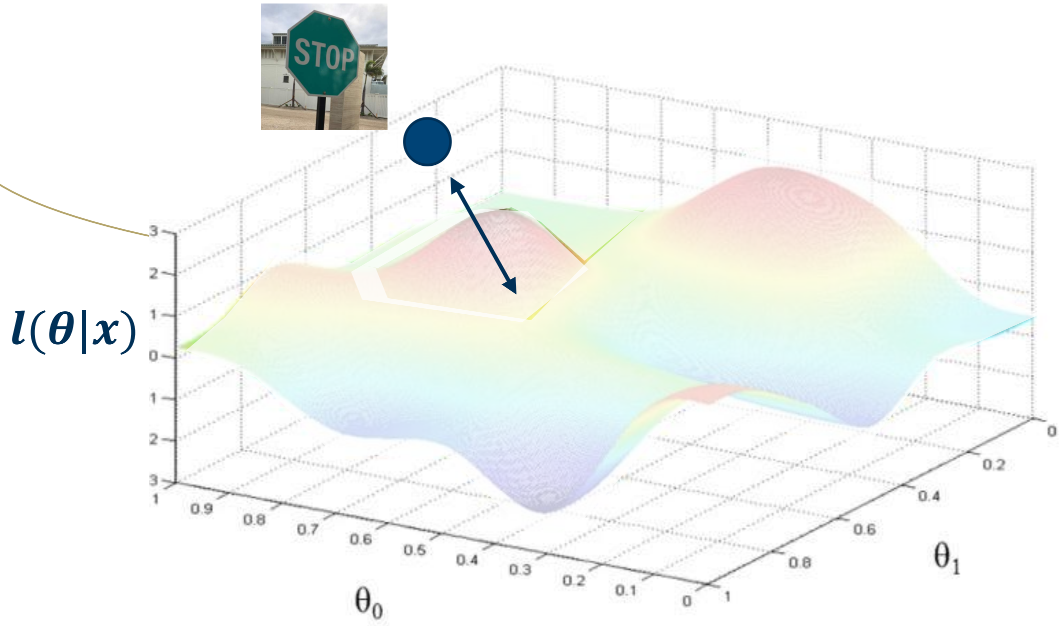
[1Van Amersfoort, J., Smith, L., Teh, Y. W., & Gal, Y. (2020, November). Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning* (pp. 9690-9700). PMLR.

Uncertainty

Gradients as Single pass Features

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

Distance from unknown cluster



Two techniques:

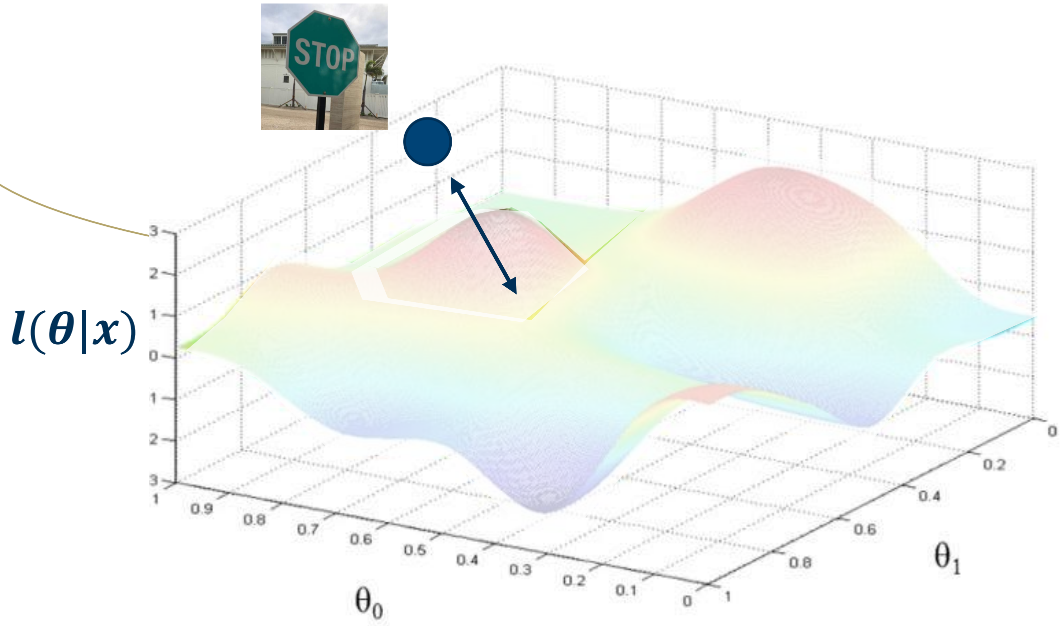
1. Gradient constraints during Training for Anomaly Detection
2. Backpropagating Confounding labels for Out-of-Distribution Detection

Uncertainty

Gradients as Single pass Features

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

Distance from unknown cluster



Two techniques:

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

Backpropagated Gradient Representations for Anomaly Detection

Gukyeong Kwon, PhD
Amazon AWS

Mohit Prabhushankar, PhD
Postdoc, Georgia Tech

Ghassan AlRegib, PhD
Professor, Georgia Tech

SCAN ME

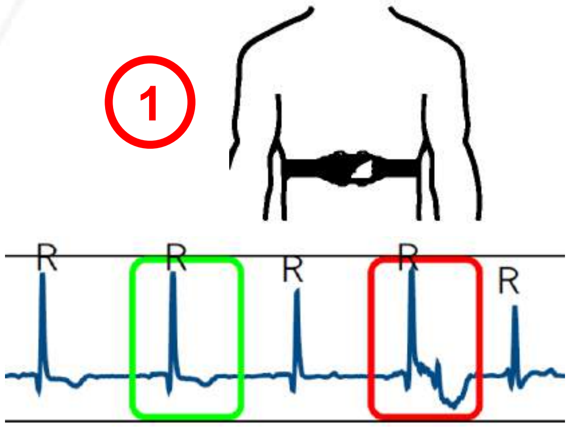
Anomalies

Finding Rare Events in Normal Patterns

Backpropagated Gradient Representations for Anomaly Detection

SCAN ME

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



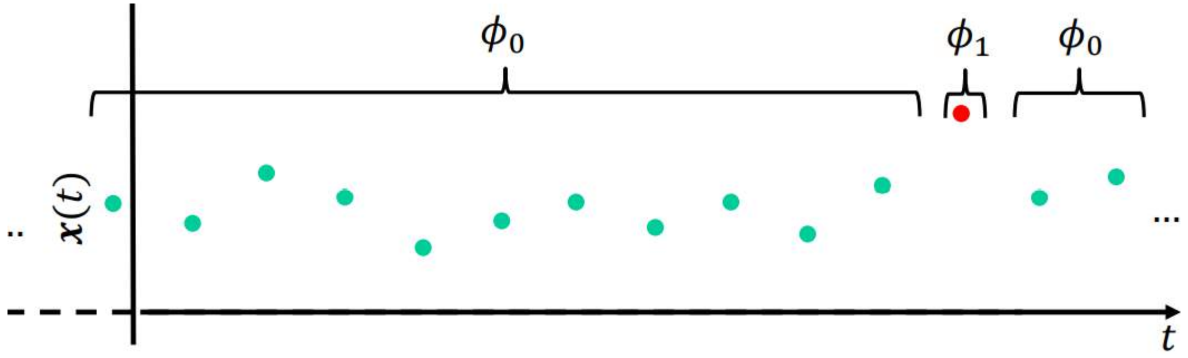
2

Statistical Definition:

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

Goal: Detect ϕ_1

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



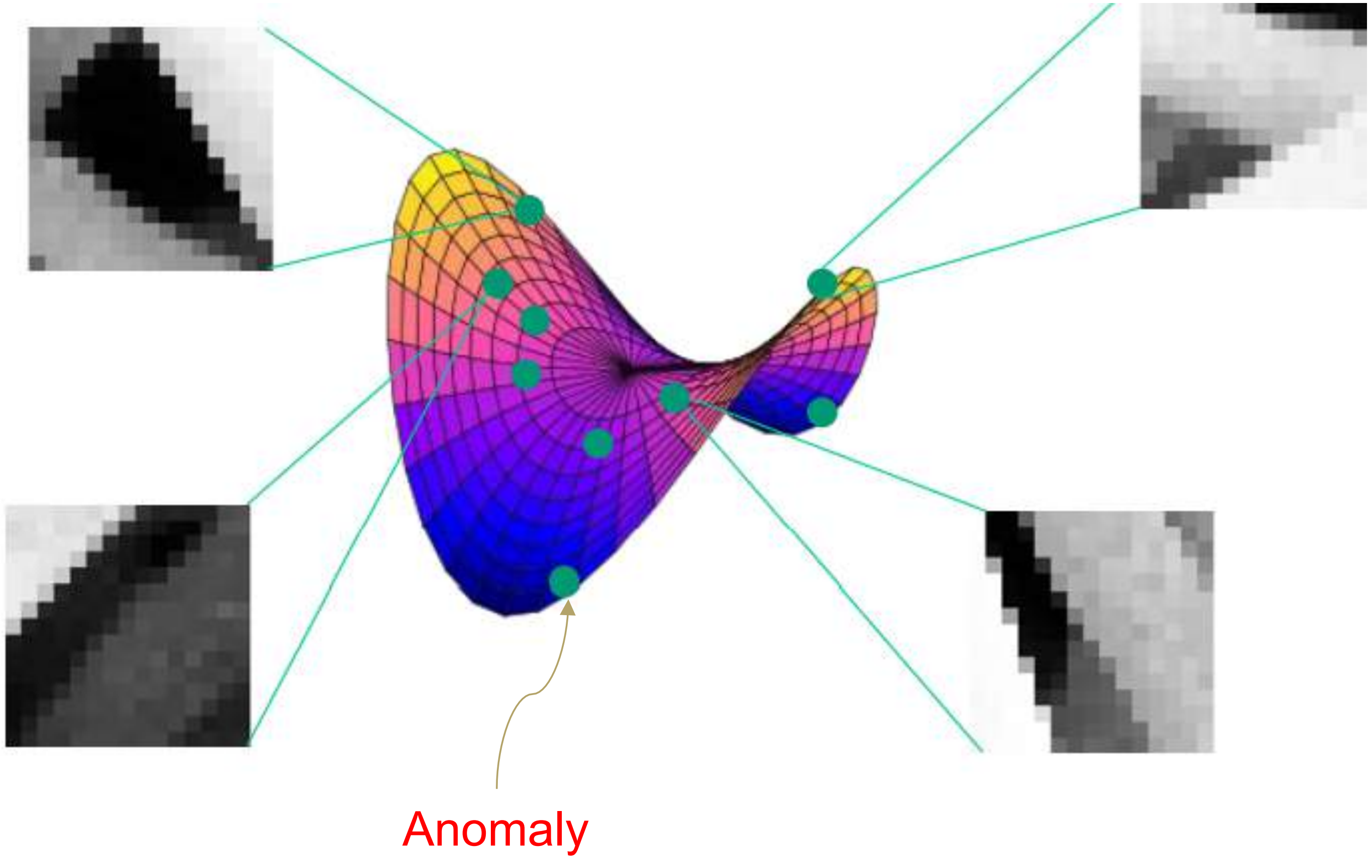
[1] V. Chandola, A. Banerjee, V. Kumar. "Anomaly detection: A survey". ACM Comput. Surv. 41, 3, Article 15 (July 2009), 58 pages

Anomalies

Steps for Anomaly Detection

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

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



Constraining Manifolds
General Constraints

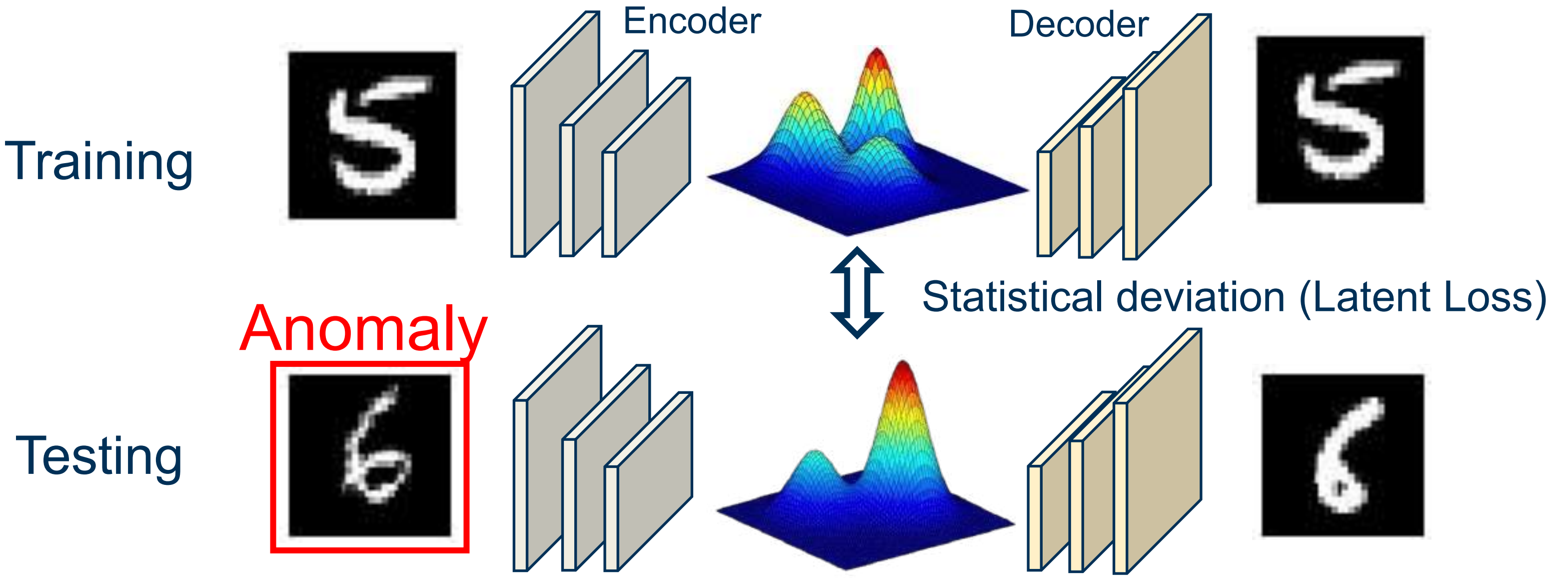
Backpropagated Gradient Representations for Anomaly Detection

SCAN ME

Constrained Representation

2004 — Tax et.al [1]
2016 — Fan et.al [2]
2018 — Pidhorksyi et.al [3]
2019 — Abati et.al [4]

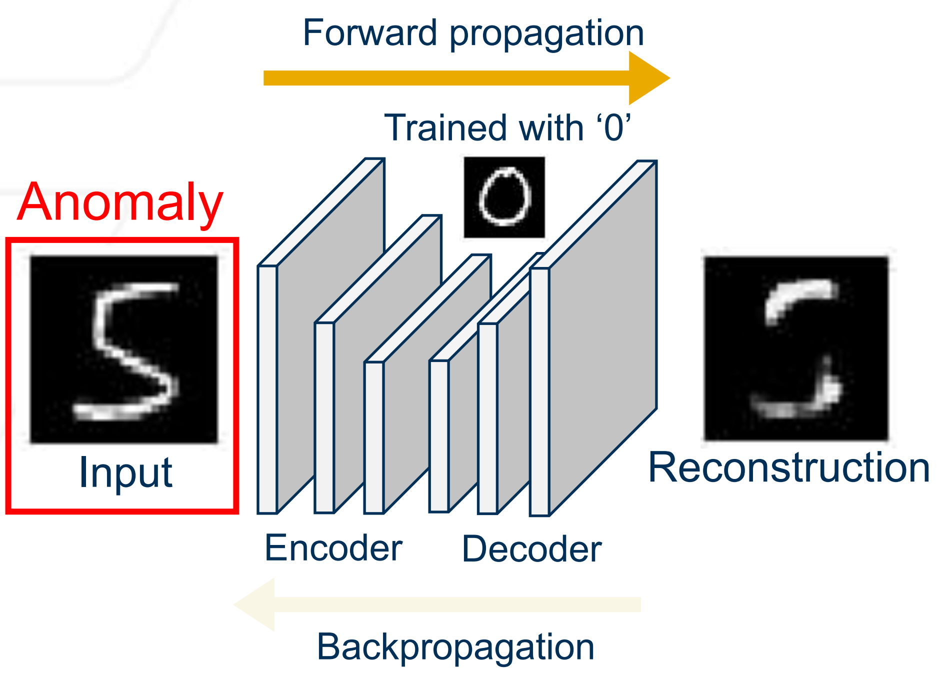
Activations are constrained using GANs, VAEs, etc.



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

Constraining Manifolds

Gradient-based Constraints



Activation Constraints

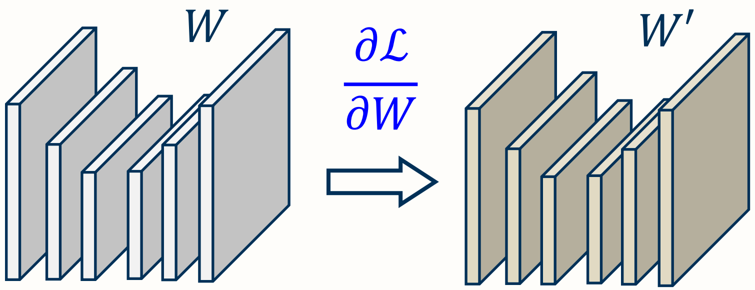
Activation-based representation
(Data perspective)

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

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

Gradient Constraints

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



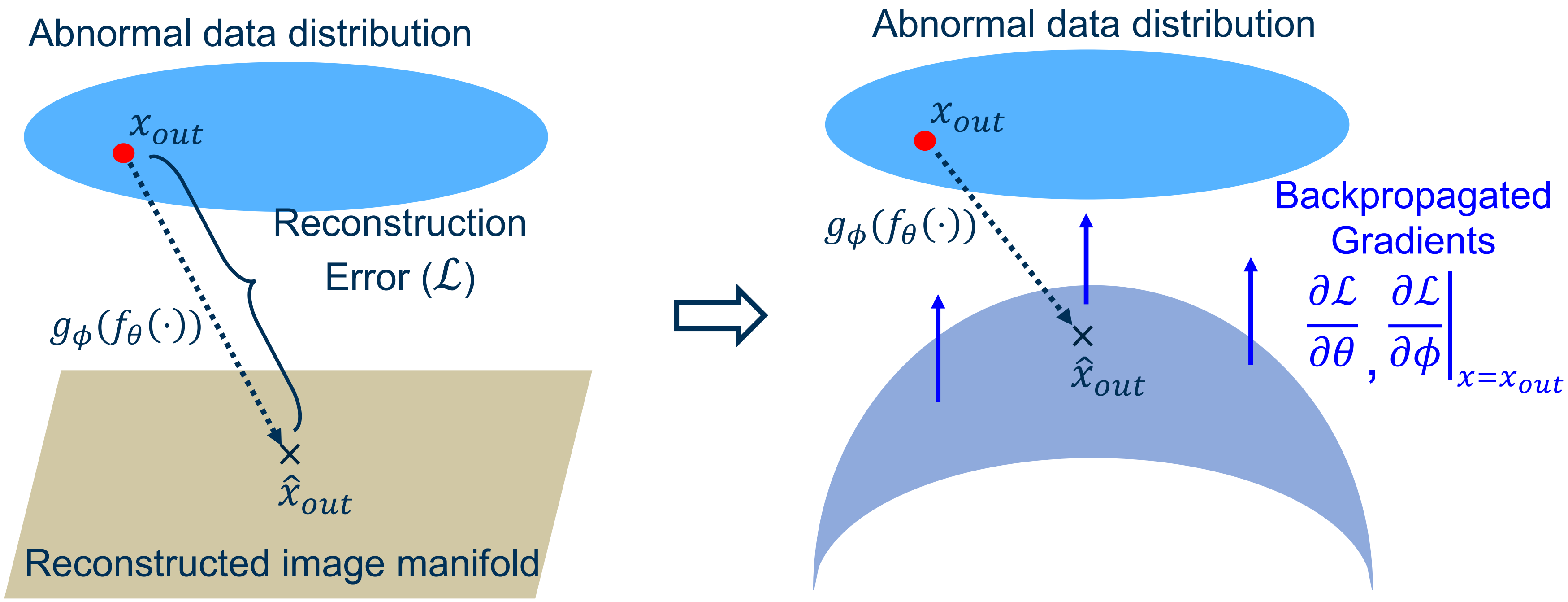
How much **model update** is required by the input?

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

Constraining Manifolds
Advantages of Gradient-based Constraints

Backpropagated Gradient Representations for Anomaly Detection

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

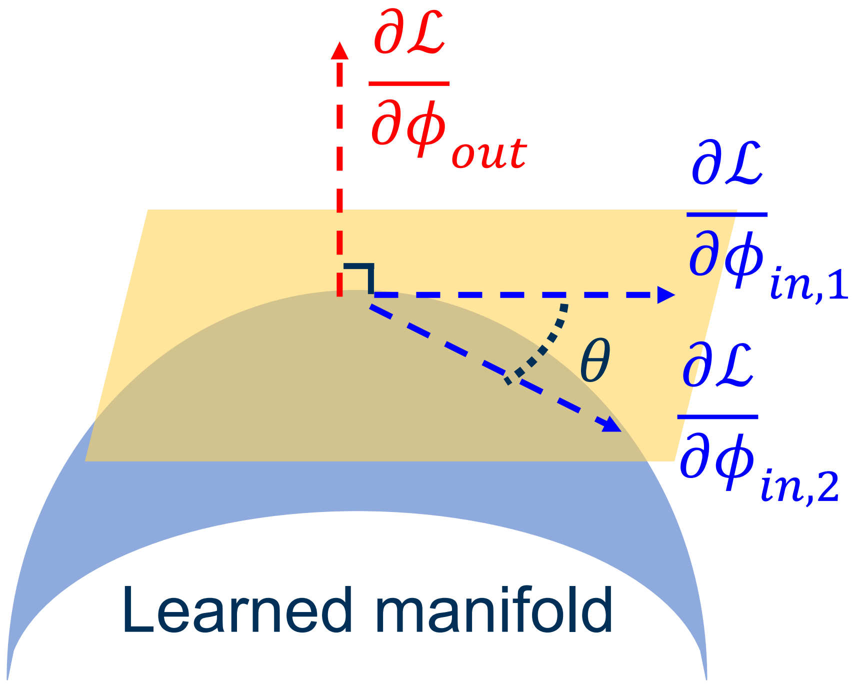


G. Kwon, M. Prabhushankar, D. Temel, and G. AlRegib, "Backpropagated Gradient Representations for Anomaly Detection," 2020

GradCON: Gradient Constraint

Gradient-based Constraints

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



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

At k-th step of training,

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

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

Gradients at k-th iter.

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

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

GradCON: Gradient Constraint

Activations vs Gradients

Backpropagated Gradient Representations for Anomaly Detection

SCAN ME

Abnormal "class" detection (CIFAR-10)

e.g.

Normal

Abnormal

AUROC Results

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

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

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

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

GradCON: Gradient Constraint

Aberrant Condition Detection

Backpropagated Gradient Representations for Anomaly Detection

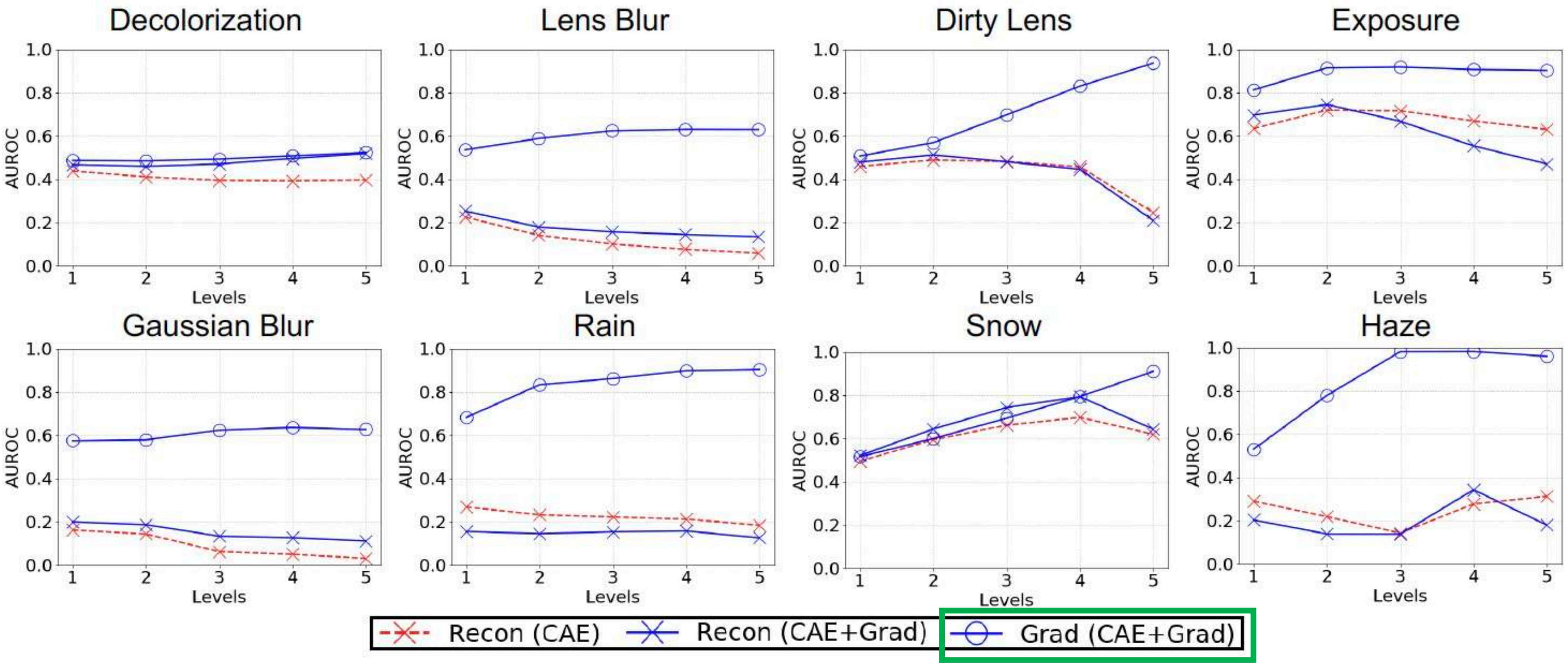
SCAN ME

Abnormal "condition" detection (CURE-TSR)

Normal

Abnormal

AUROC Results

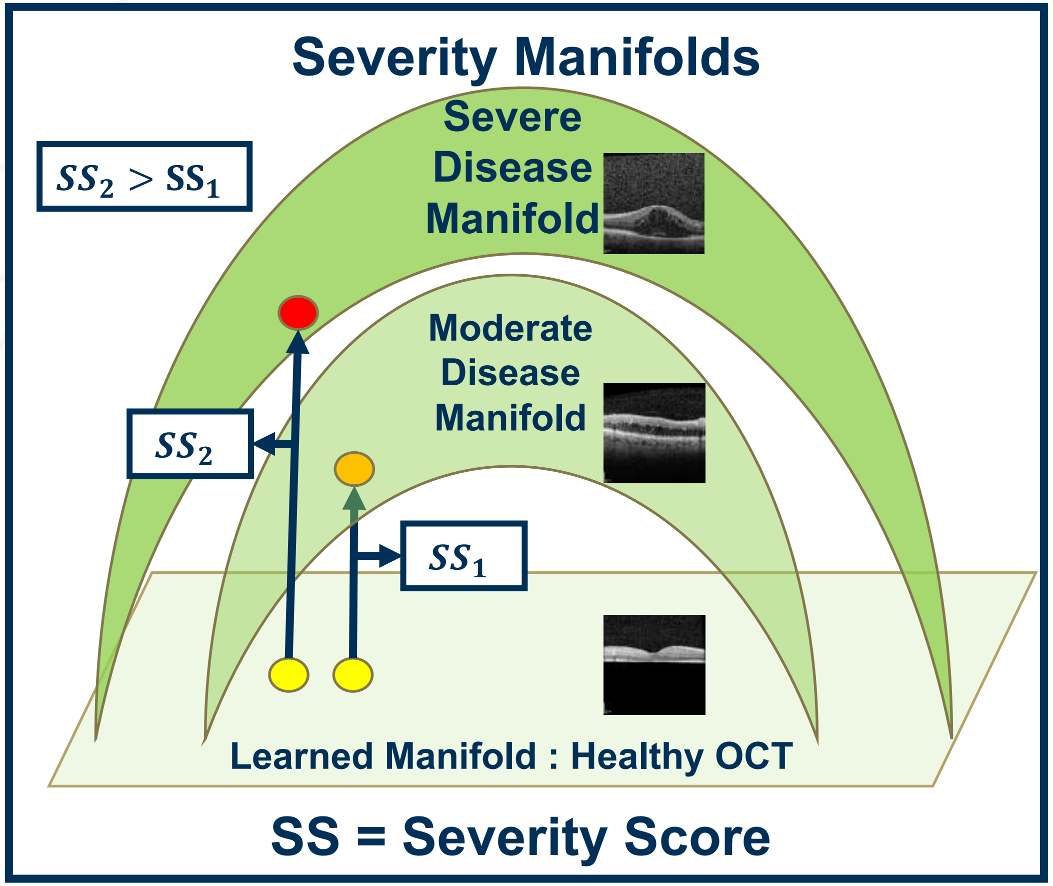


Recon: Reconstruction error, Grad: Gradient loss

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

GradCON Applicability
Estimating Disease Severity

Backpropagated Gradient Representations for Anomaly Detection



Goal
- Define severity with respect to distance from a healthy manifold.
- This distance can be regarded as a severity score.

How to measure severity score?
- Define severity as: "the degree to which a sample appears anomalous relative to the distribution of healthy images."

Experimental Plan
- Investigate model responses that can act as good surrogate for severity score

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

GradCON Applicability

Estimating Disease Severity

Dataset: Ophthalmic Labels for Investigating Visual Eye Semantics

- **9408** images **labeled** with complete biomarker data
- Every image associated with vector indicating presence/absence of **16 potential biomarkers**
- 5 biomarkers exist with sufficient balanced quantities
  - Develop 5 biomarker test sets (PAVF, FAVF, IRF, DME, and IRHRF)

https://github.com/olivesgatech

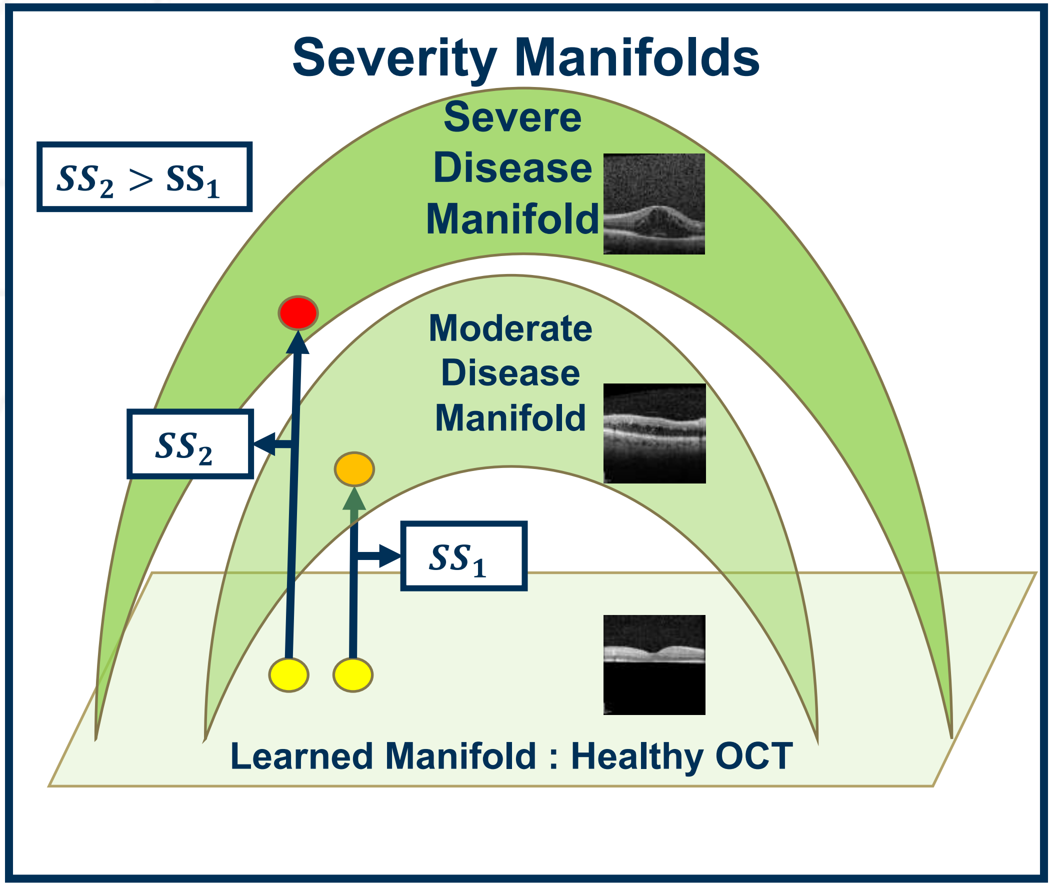
OLIVES Dataset
https://arxiv.org/pdf/2209.11195.pdf

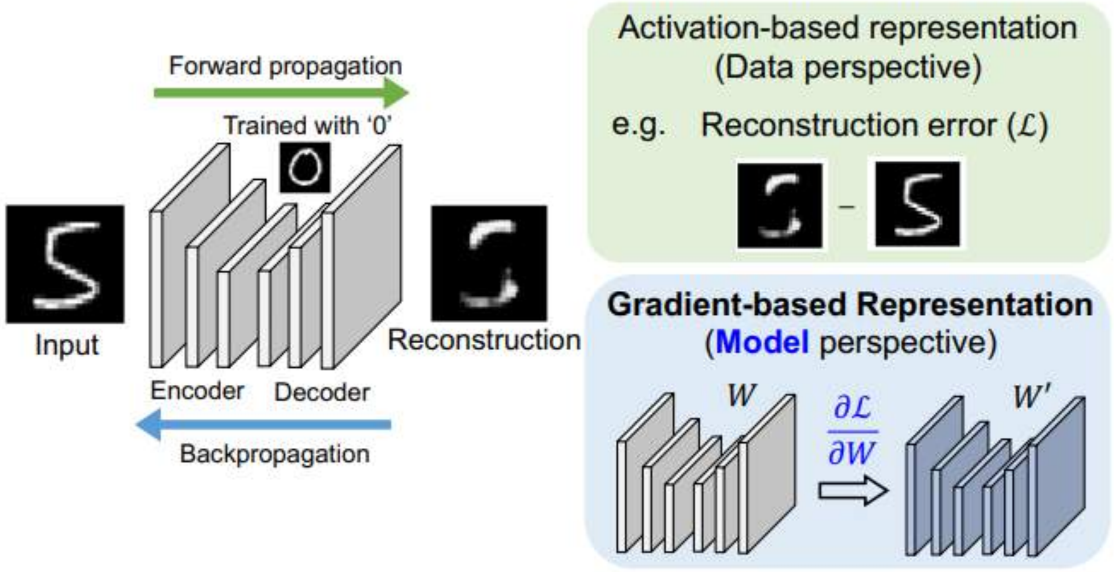
K. Kokilepersaud, M. Prabhushankar, G. AlRegib, S. Trejo Corona, C. Wykoff, "Gradient Based Labeling for Biomarker Classification in OCT," in *IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, Oct. 16-19 2022

GradCON Applicability

Estimating Disease Severity

Backpropagated Gradient Representations for Anomaly Detection





$$\mathcal{L}_{grad} = -\mathop{\mathbb{E}}_{i}\left[\mathrm{cosSIM}\left(\frac{\partial \mathcal{J}}{\partial \phi_i}^{k-1}_{avg}, \frac{\partial \mathcal{L}}{\partial \phi_i}^{k}\right)\right], \quad \frac{\partial \mathcal{J}}{\partial \phi_i}^{k-1}_{avg} = \frac{1}{(k-1)}\sum_{t=1}^{k-1}\frac{\partial \mathcal{J}}{\partial \phi_i}^{t},$$

$$L = L_{recon} + \alpha L_{grad}$$

Idea

- Constrain gradients of in-distribution class
- Make gradients sensitive to progressively anomalous data

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

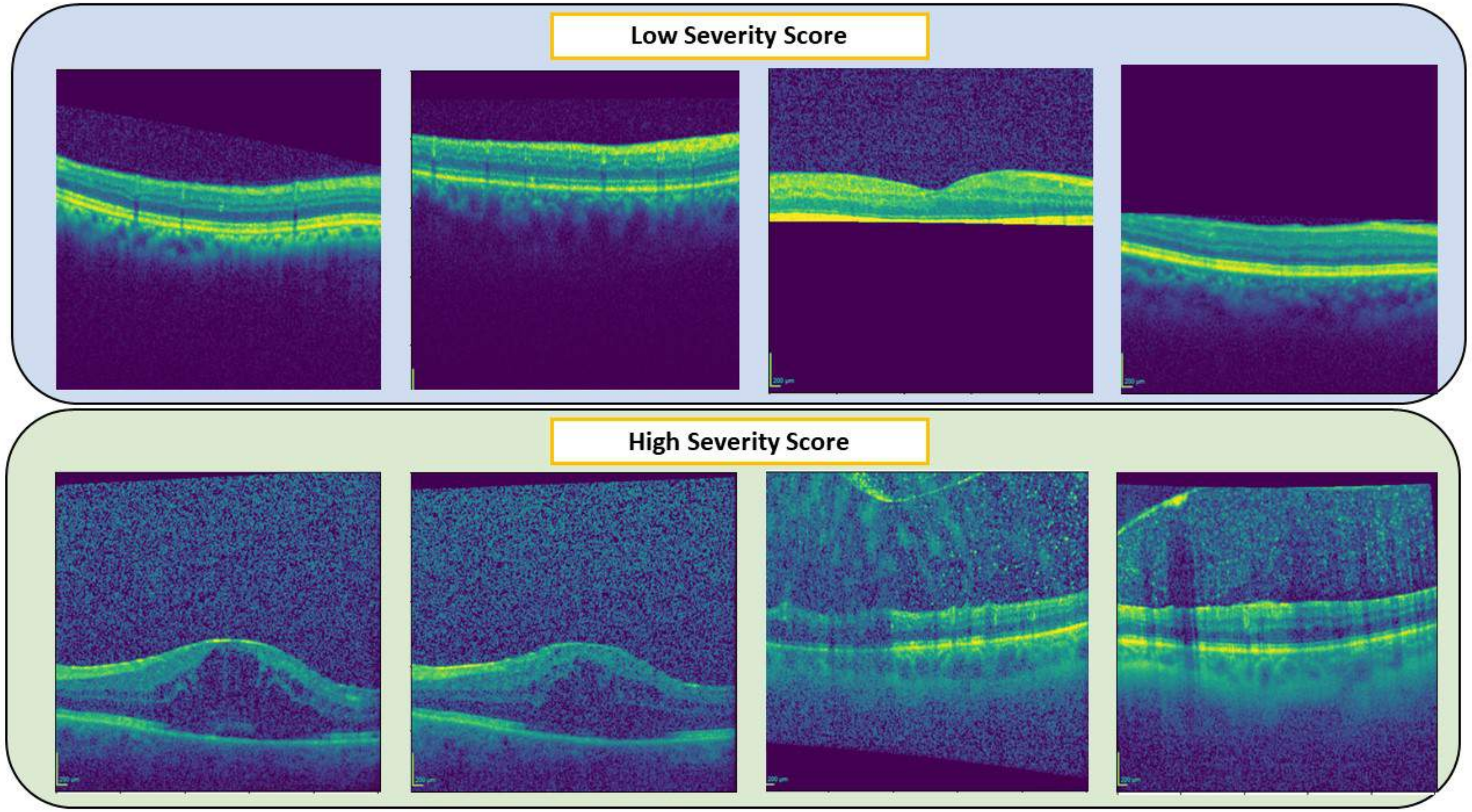
GradCON Applicability

Estimating Disease Severity

Backpropagated Gradient Representations for Anomaly Detection

SCAN ME

Severity Labels used to select positive and negative pairs for weakly-supervised contrastive learning



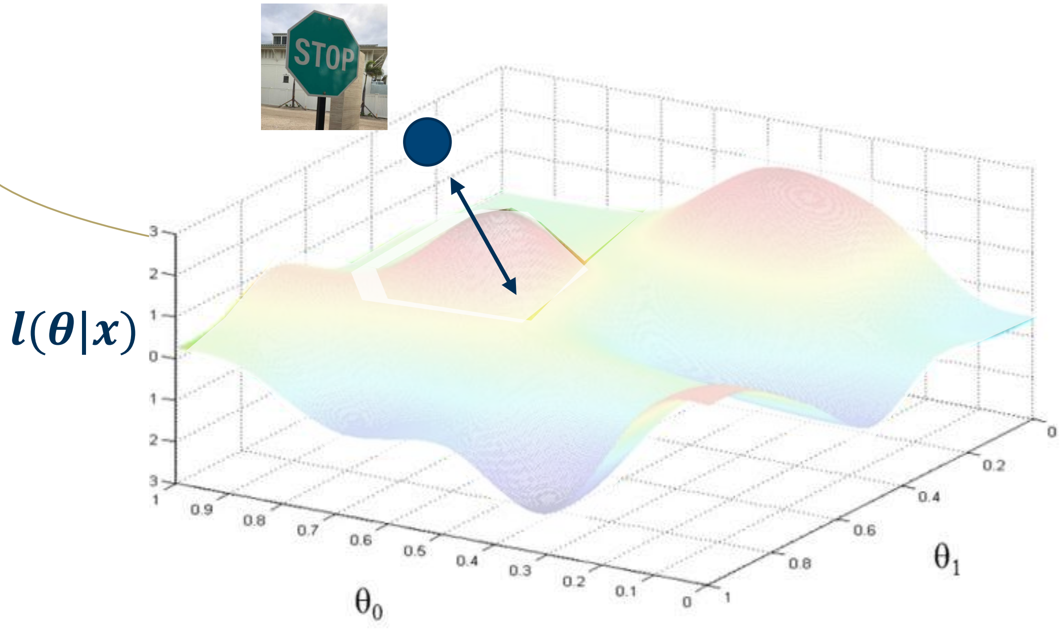
K. Kokilepersaud, M. Prabhushankar, G. AlRegib, S. Trejo Corona, C. Wykoff, "Gradient Based Labeling for Biomarker Classification in OCT," in *IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, Oct. 16-19 2022

Uncertainty

Gradients as Single pass Features

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

Distance from unknown cluster



Two techniques:

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

Probing the Purview of Neural Networks via Gradient Analysis

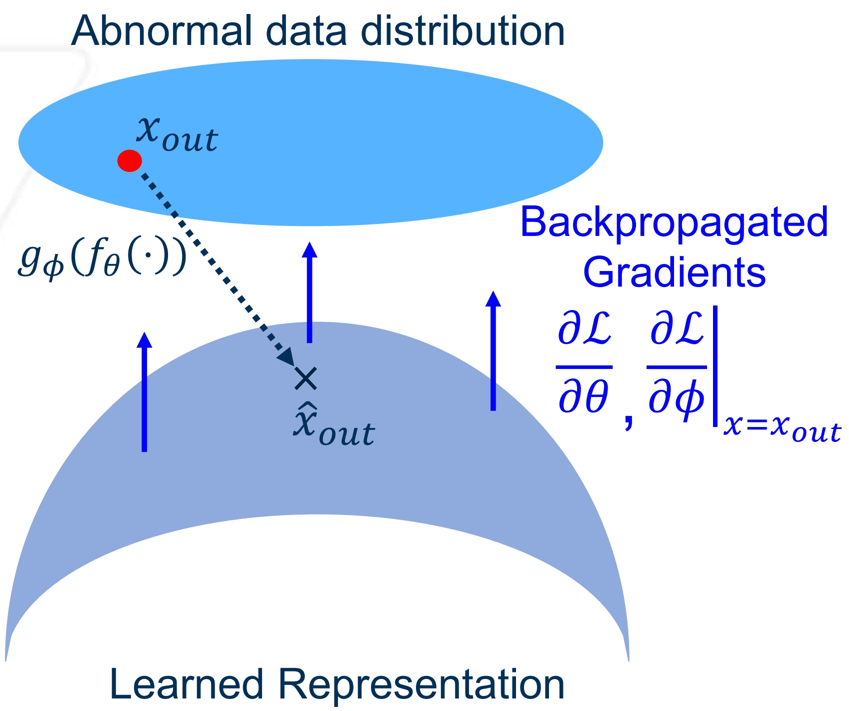
Jinsol Lee,
PhD Candidate

Mohit Prabhushankar, PhD
Postdoc

Ghassan AlRegib, PhD
Professor

SCAN ME

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



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

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

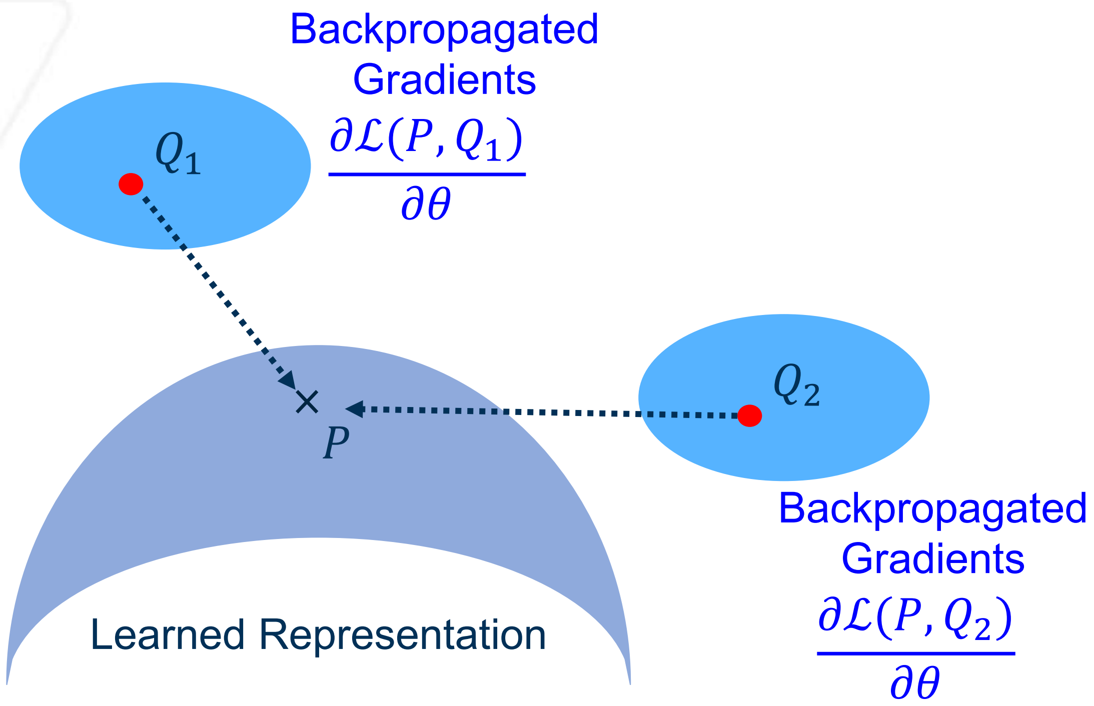
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

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

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



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

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

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

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

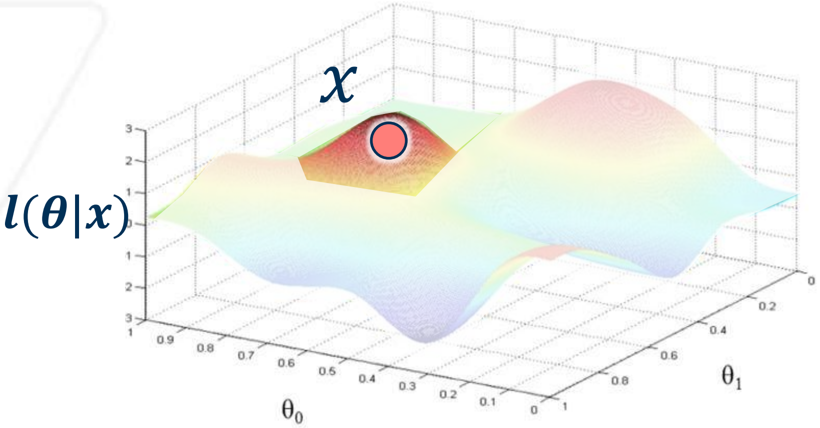
Toy Manifold Example

What is uncertainty?

Probing the Purview of Neural Networks via Gradient Analysis

Gradients represent the local required change in manifold

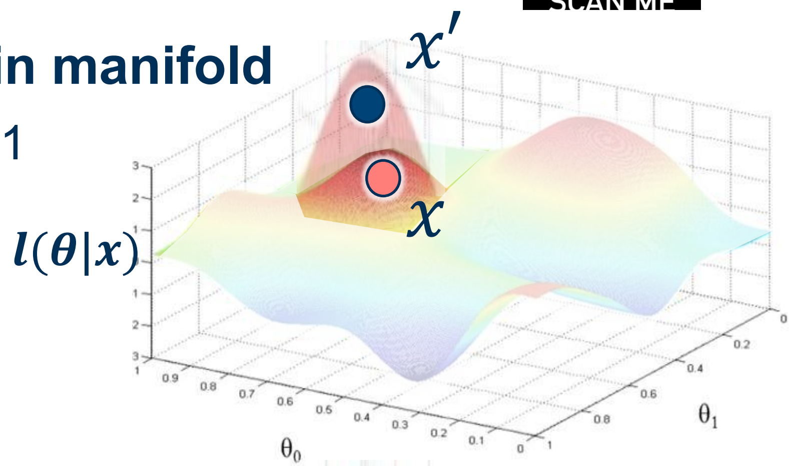
Similar to introspective learning!

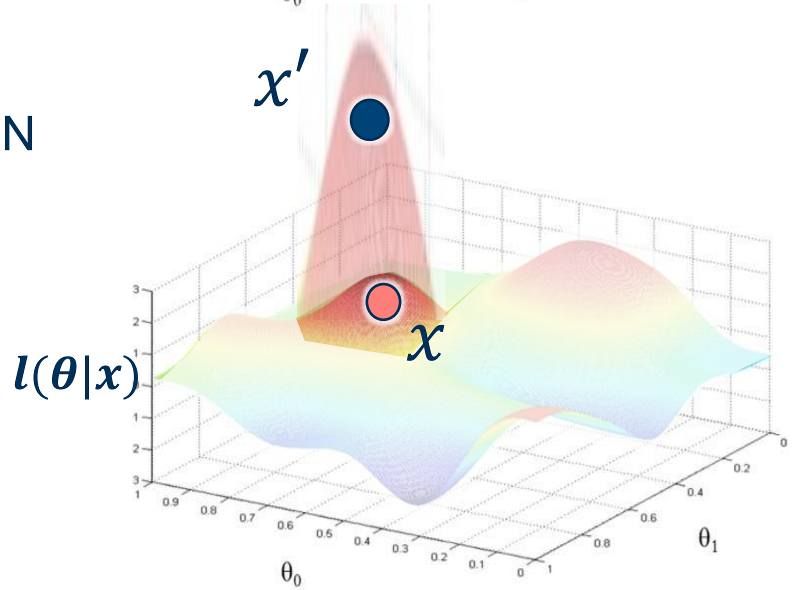


Contrast class 1

.
.
.

Contrast class N





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

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

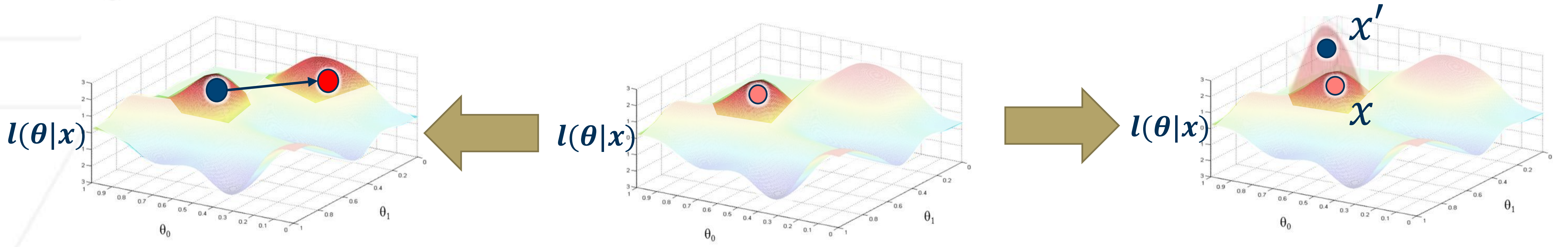
Toy Manifold Example

How is this different from Part 2?

SCAN ME

Part 2: Information

Part 3: Uncertainty



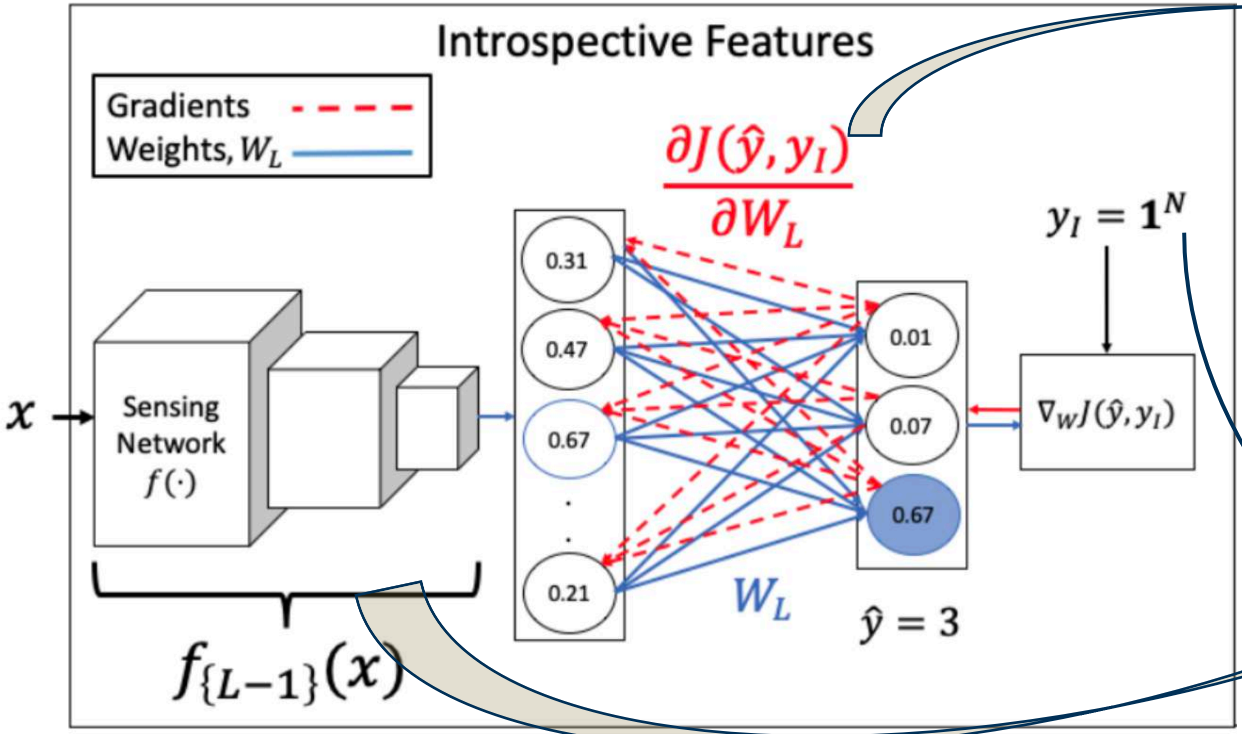
- In Part 2: Activations of learned manifold are weighted by gradients w.r.t. activations to extract information and provide explanations

- In Part 3: Statistics of gradients w.r.t. the weights (energy) will be directly used as features

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

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



Normalized and vectorized gradients are introspective features.

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

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

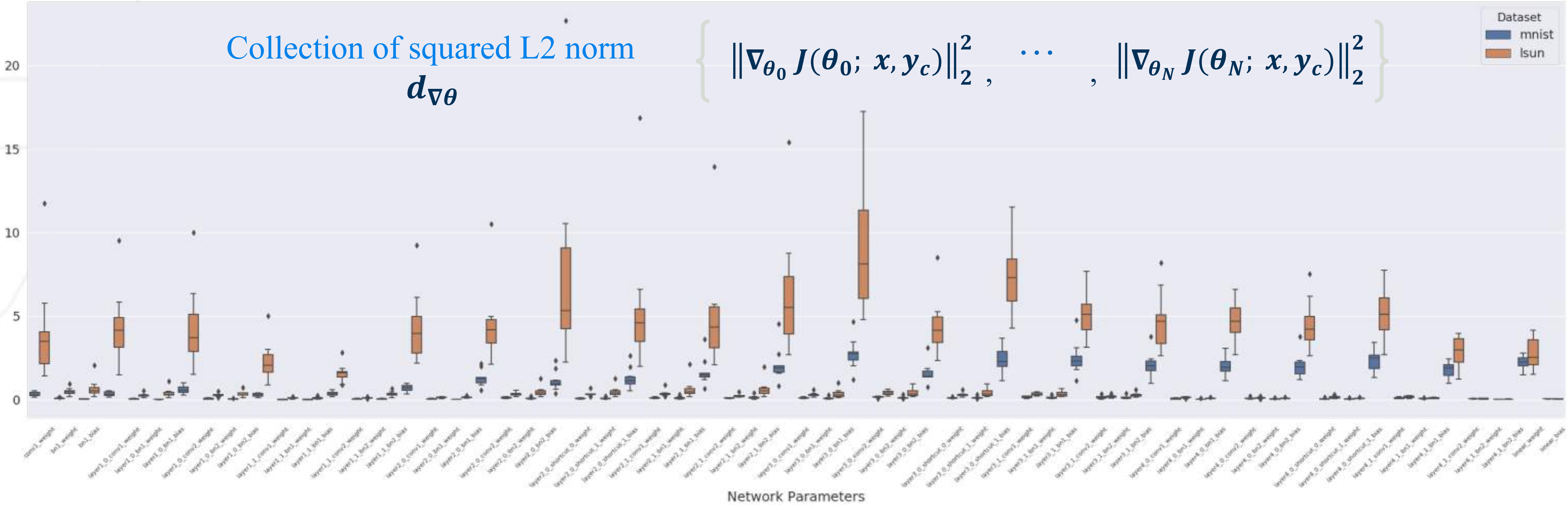
SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

Step 2: Take L2 norm of all generated gradients

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

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



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

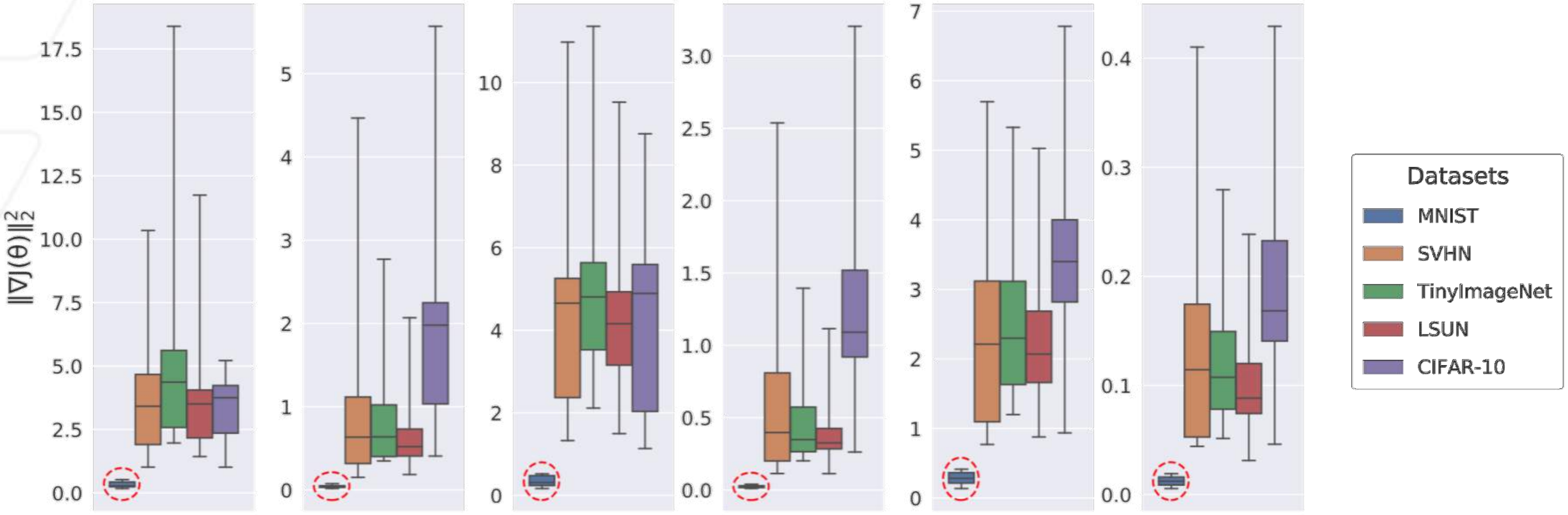
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Gradient-based Uncertainty
Uncertainty in OOD Setting

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

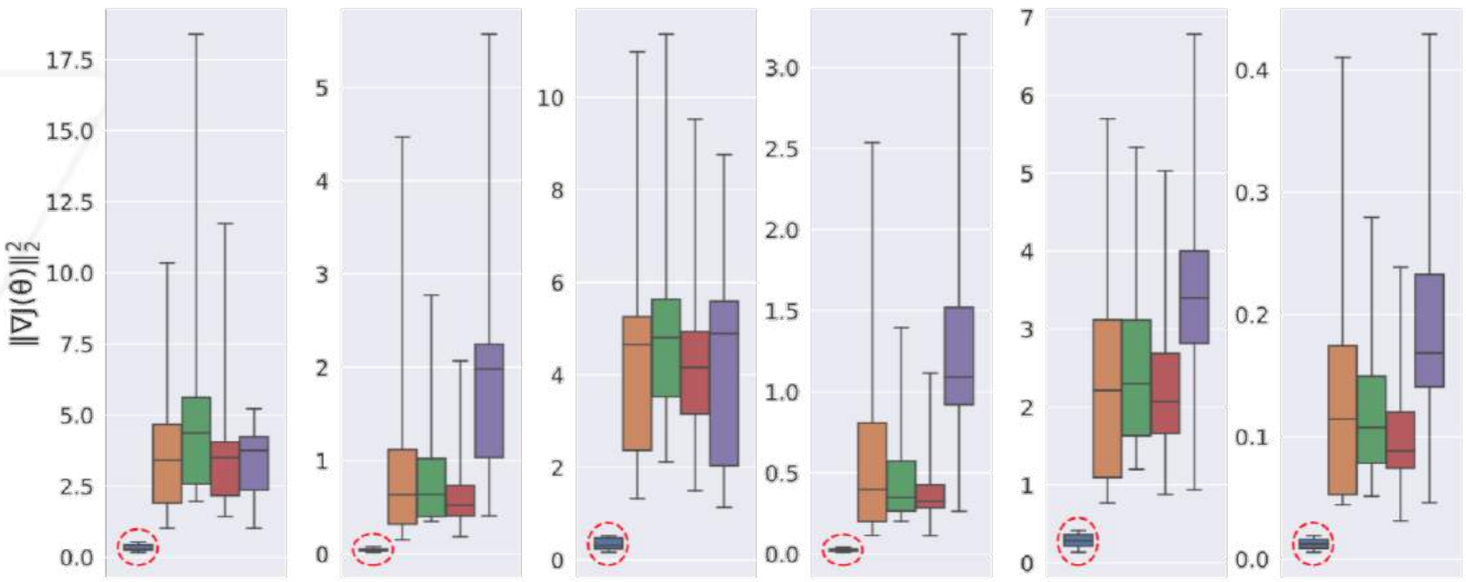
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

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



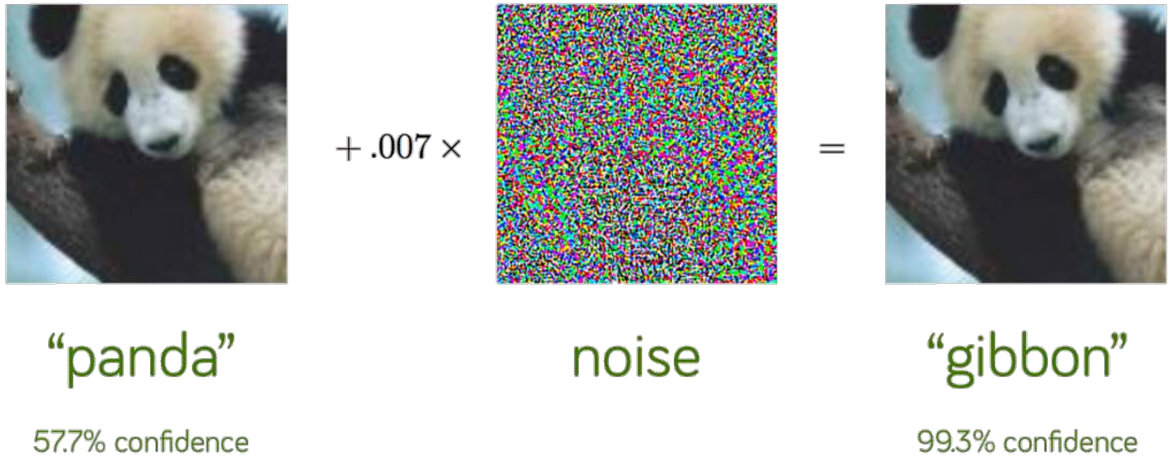
Step 1: Train a deep network $f(\cdot)$ on some **training distribution**
Step 2: Introduce challenging (adversarial, noisy, OOD) data
Step 3: Derive **gradient uncertainty** on both trained and challenge data
Step 4: Train a classifier $H(\cdot)$ to **detect** challenging from trained data
Step 5: At 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.

Gradient-based Uncertainty

Uncertainty in Adversarial Setting

Vulnerable DNNs in the real world



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

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

Gradient-based Uncertainty

Uncertainty in Adversarial Setting

SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

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

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

Gradient-based Uncertainty

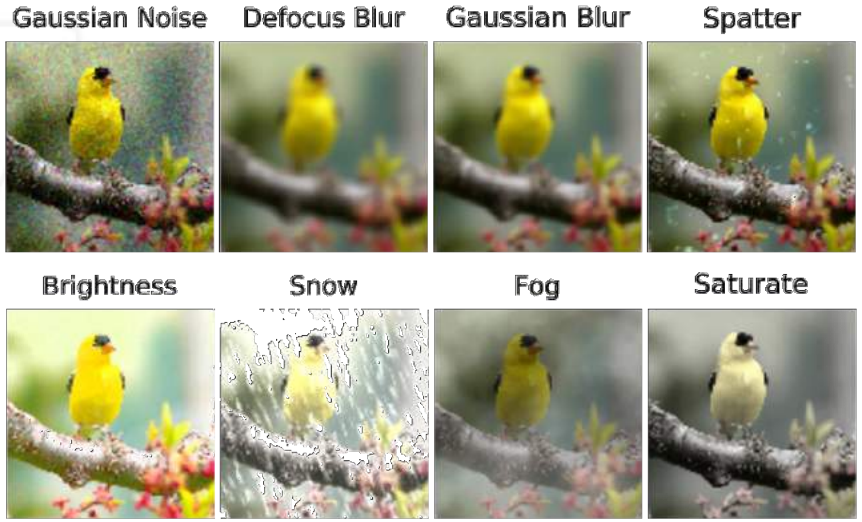
Uncertainty in Detecting Challenging Conditions

SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

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

CIFAR-10-C



CURE-TSR

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

Gradient-based Uncertainty

Uncertainty in Detecting Challenging Conditions

SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

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.

Gradient-based Uncertainty

Uncertainty in Detecting Challenging Conditions

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

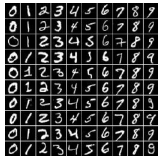
Dataset	Method	Mahalanobis [12] / Ours				
	Corruption	Level 1	Level 2	Level 3	Level 4	Level 5
CIFAR-10-C	Noise	96.63 / **99.95**	98.73 / **99.97**	99.46 / **99.99**	99.62 / **99.97**	99.71 / **99.99**
	LensBlur	94.22 / **99.95**	97.51 / **99.99**	99.26 / **100.0**	99.78 / **100.0**	99.89 / **100.0**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

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

Out-of-Distribution Detection

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

Train set →  MNIST

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

SVHN

CIFAR10

TinyImageNet

LSUN

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

Out-of-Distribution Detection

Probing the Purview of Neural Networks via Gradient Analysis

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.

Out-of-Distribution Detection

Probing the Purview of Neural Networks via Gradient Analysis

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**

Numbers

SVHN

CIFAR10

TinyImageNet

LSUN

Objects, natural scenes

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

Out-of-Distribution Detection

Probing the Purview of Neural Networks via Gradient Analysis

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**

More similar datasets (objects)

TinyImageNet

CIFAR10

LSUN

SVHN

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

OLIVES @GeorgiaTech

Georgia Tech

Objectives

Takeaways from Part III

- Part I: Gradients in Neural Networks
- Part 2: Gradients as Information
- **Part 3: Gradients as Uncertainty**
  - Defining Uncertainty in the context of Neural Networks
  - Anomaly Detection
    - GradCON: Gradient Constraints
  - Out-of-Distribution Detection
  - Adversarial Detection
  - Corruption Detection
- Part 4: Gradients as Expectancy-Mismatch
- Part 5: Conclusion and Future Directions

Interpretation, and Applications of Gradients

Part 4: Gradients as Expectancy-Mismatch

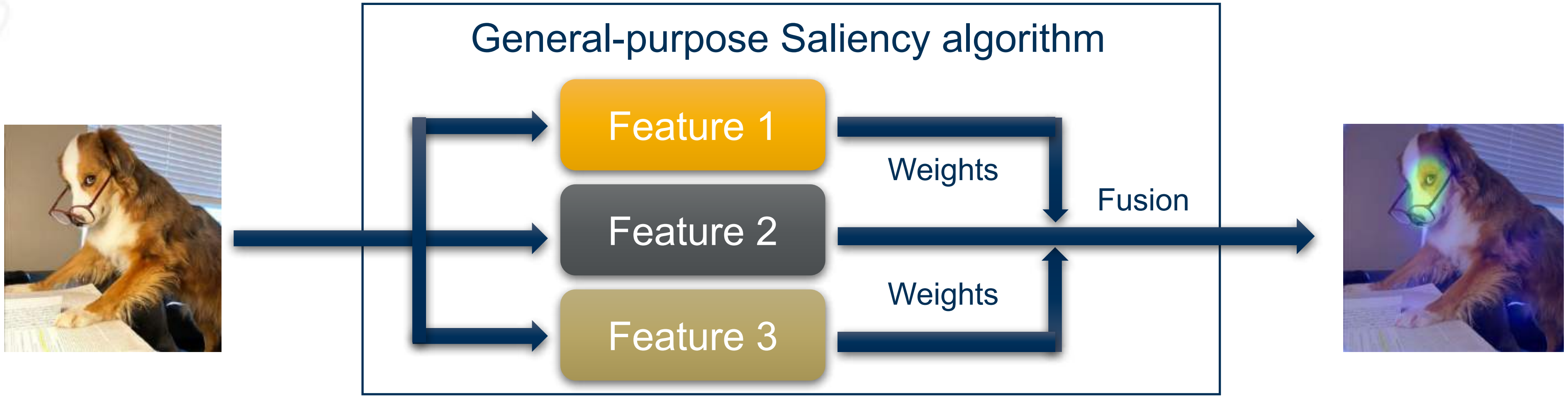
Objectives

Objectives in Part IV

Case Study: Expectancy-Mismatch

- Interpret gradients as Expectancy-Mismatch
  - Define expectancy-mismatch utilizing saliency
  - Demonstrate counterfactual manifolds as expectancy-mismatch
- Human Visual Saliency
- Image Quality Assessment

Saliency
Saliency in Literature



Bottom-Up Saliency : Innovation is in designing features and fusion

Top-Down Saliency : Innovation is in designing weights

Color, Intensity, Orientation [1]

Faces, text, object detectors [1]

[1] Judd, Tilke, Frédo Durand, and Antonio Torralba. "A benchmark of computational models of saliency to predict human fixations." (2012).

Saliency

Our Goal: Introduce Implicit Saliency in Neural Networks

General-purpose Saliency algorithm

Feature 1

Feature 2

Feature 3

Weights

Weights

Fusion

Bottom-Up Saliency : Innovation is in designing features and fusion

Top-Down Saliency : Innovation is in designing weights

Color, Intensity, Orientation [1]

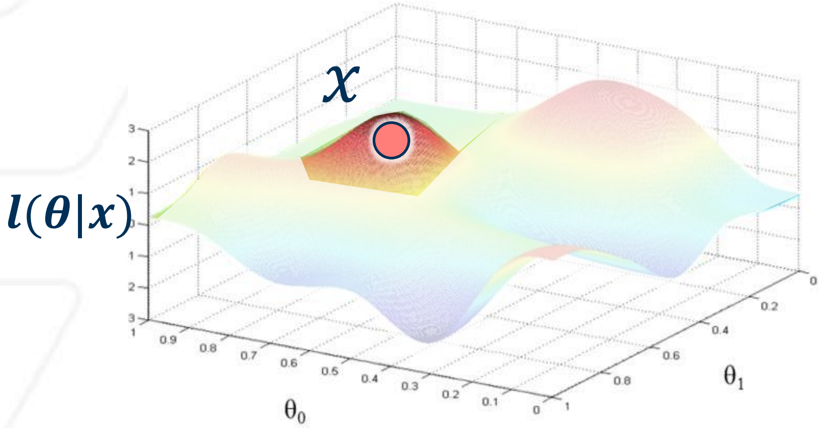
Faces, text, object detectors [1]

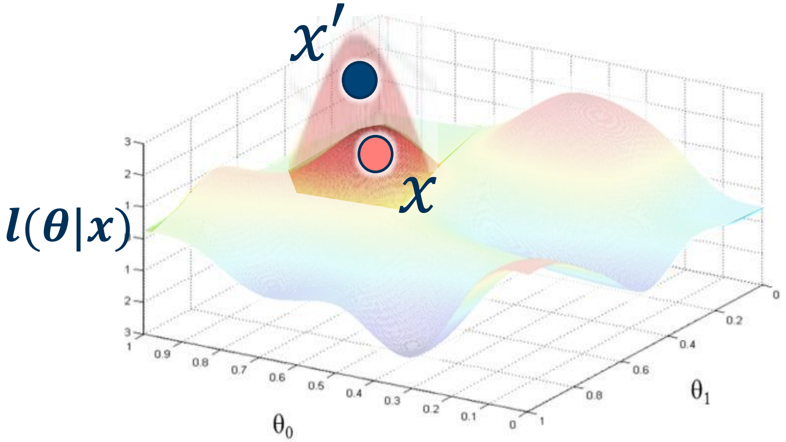
Features that are new and unexpected (novel) in a scene are salient

[1] Judd, Tilke, Frédo Durand, and Antonio Torralba. "A benchmark of computational models of saliency to predict human fixations." (2012).

Expectancy-Mismatch

Our Goal: Introduce Expectancy-Mismatch in Neural Networks





At Inference, construct local contrastive manifolds

Change in Network Parameters: Expectancy-Mismatch when presented with novel data!

We demonstrate on two applications:

1. Human Visual Saliency
2. Image Quality Assessment

Expectancy-Mismatch

Our Goal: Introduce Expectancy-Mismatch in Neural Networks

At Inference, construct local contrastive manifolds

Change in Network Parameters: Expectancy-Mismatch when presented with novel data!

We demonstrate on two applications:

1. Human Visual Saliency
2. Image Quality Assessment

Stochastic Surprisal: An Inferential Measurement of Free Energy in Neural Networks

frontiers in Neuroscience

Mohit Prabhushankar, PhD
Postdoc

Ghassan AlRegib, PhD
Professor

SCAN ME

30th ICIP 2023 Kuala Lumpur

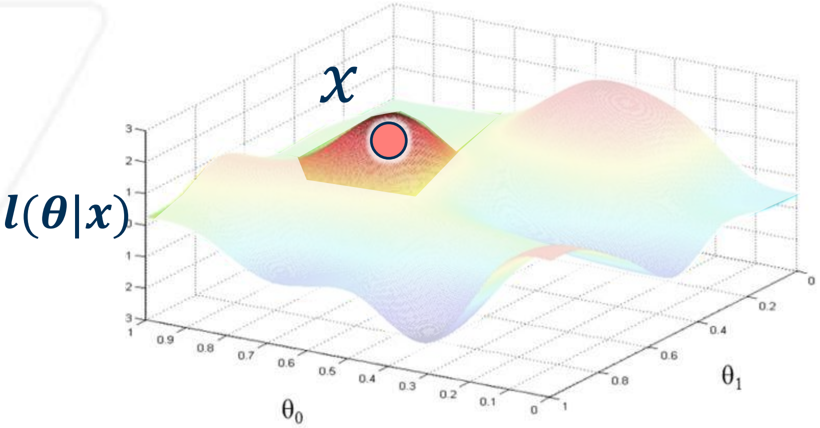
OLIVES @GeorgiaTech

Georgia Tech

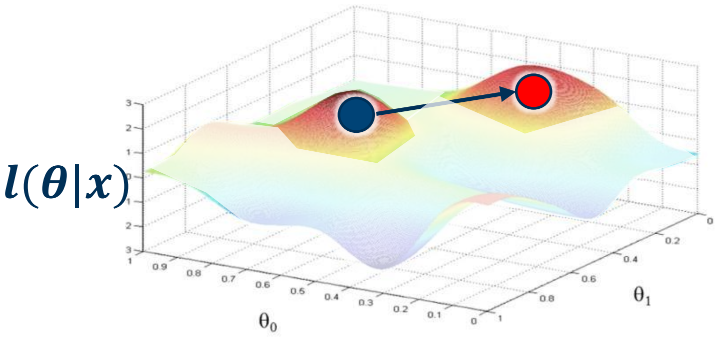
Expectancy-Mismatch

Our Goal: Introduce Expectancy-Mismatch in Neural Networks

Similar to introspective learning!

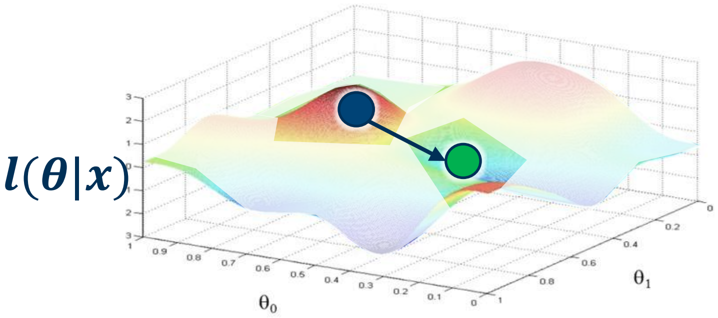


Contrast class 1



.
.
.

Contrast class N

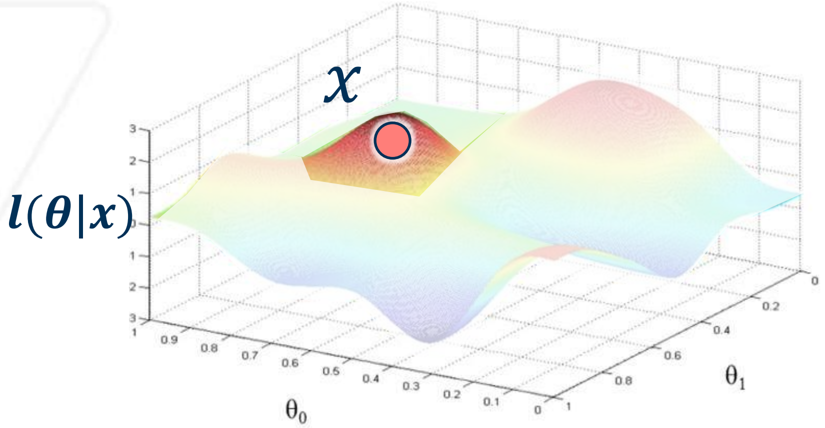


Mean of projected gradients is the expectancy!

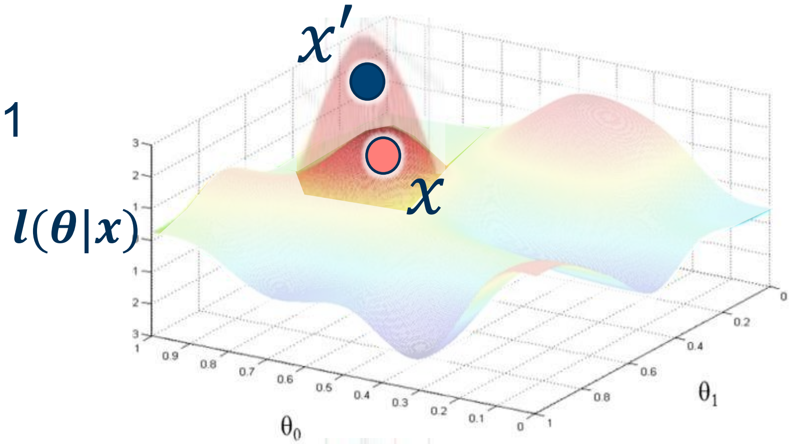
Expectancy-Mismatch

Our Goal: Introduce Expectancy-Mismatch in Neural Networks

Similar to introspective learning!

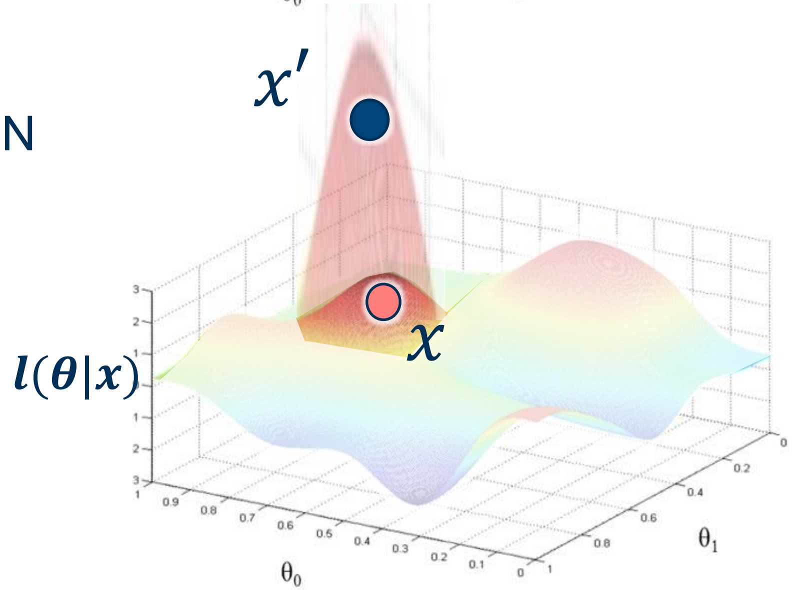


Contrast class 1



.
.
.

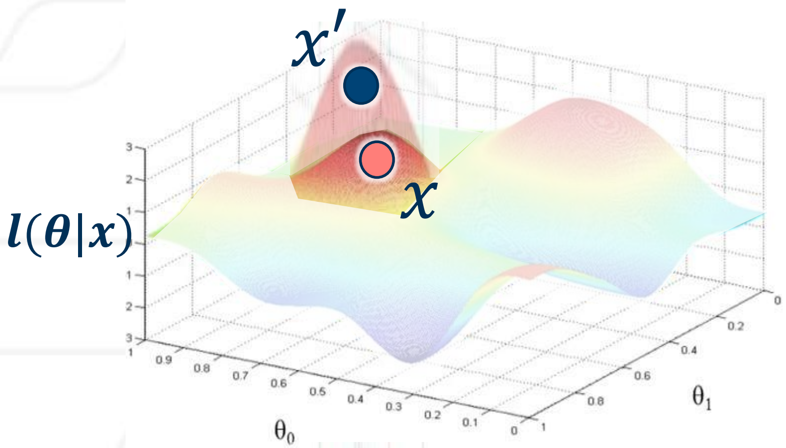
Contrast class N

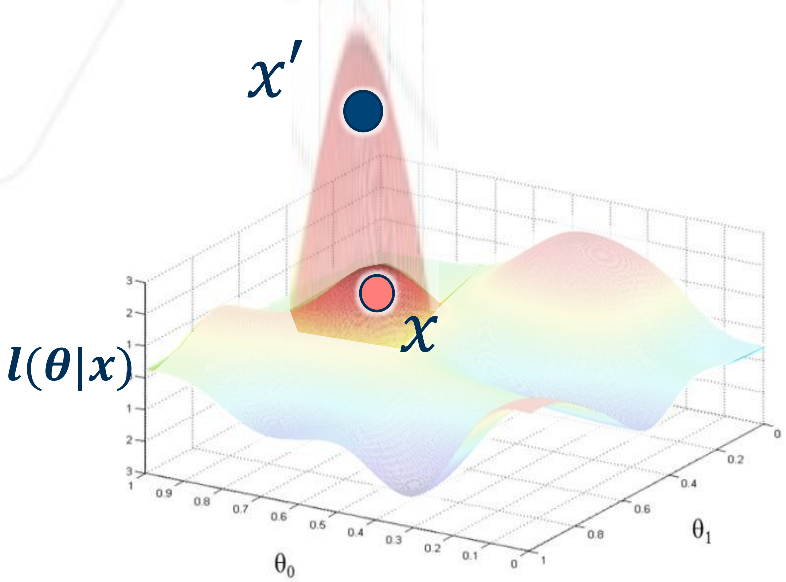


Variance of gradients is the mismatch!

Expectancy-Mismatch

Our Goal: Introduce Expectancy-Mismatch in Neural Networks



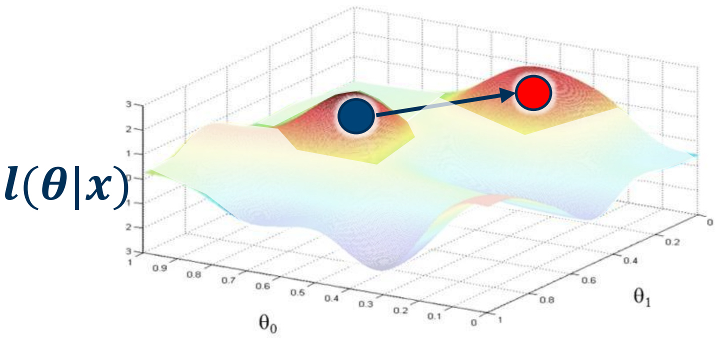


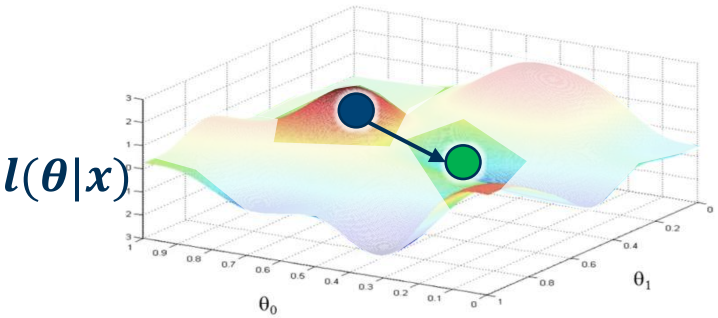
Variance of gradients is the mismatch!

Point-wise Multiplication

Mean of projected gradients is the expectancy!

Saliency Map



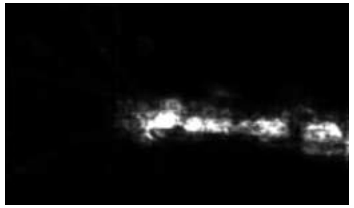


Expectancy-Mismatch

Our Goal: Introduce Expectancy-Mismatch in Neural Networks

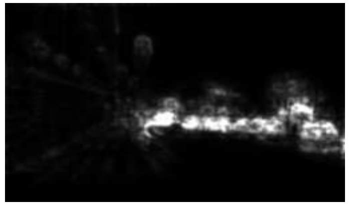
Similar to introspective learning!

Wrong class 1

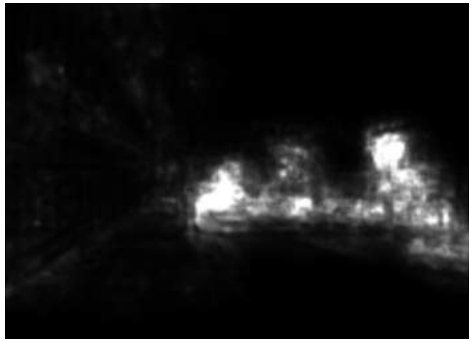


. . .

Wrong class N



Saliency Map



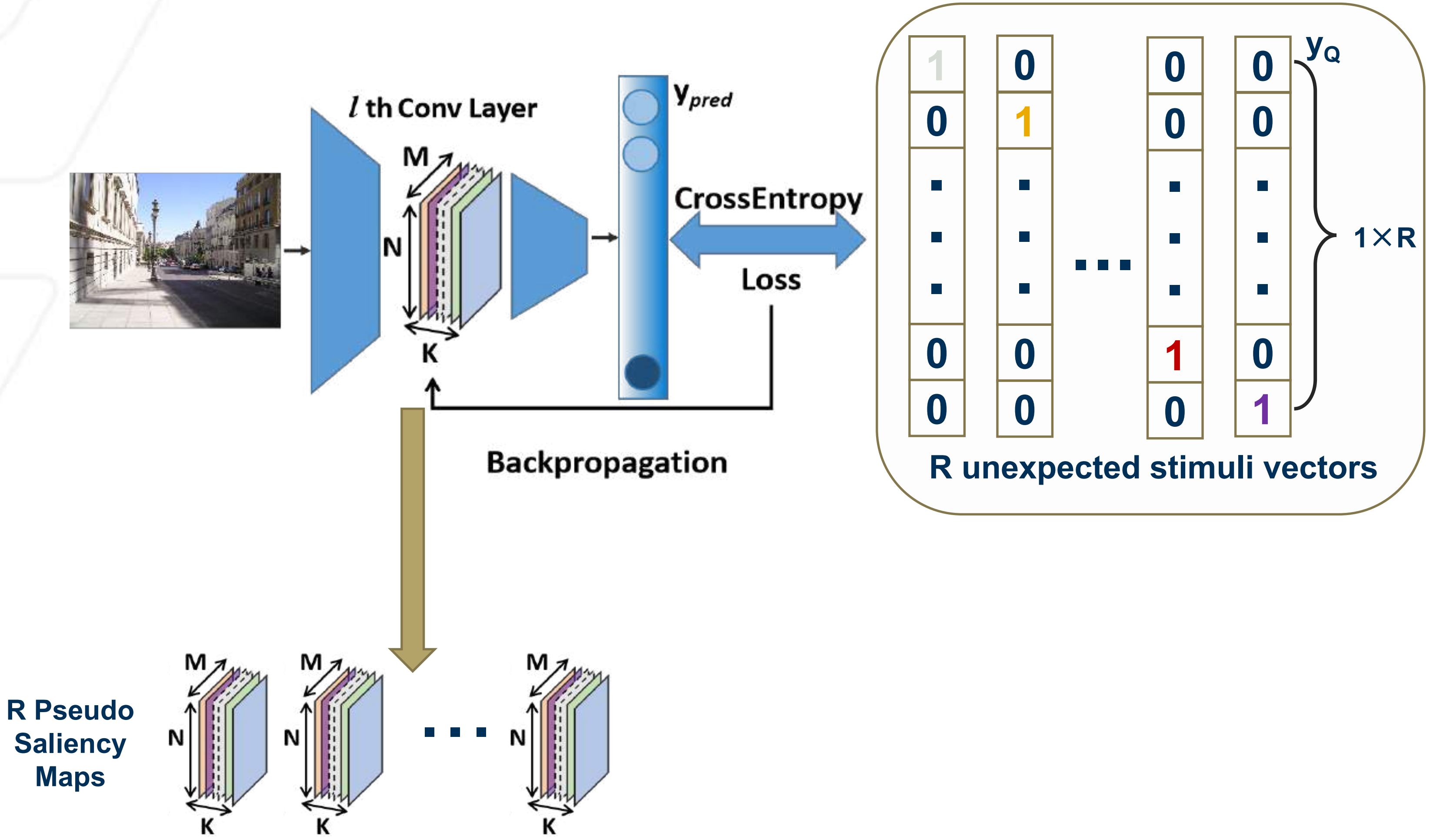
Gradients in the k^{th} layer: Pseudo-saliency maps

Sun, Yutong, Mohit Prabhushankar, and Ghassan AlRegib. "Implicit saliency in deep neural networks." *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020.

cSaliency

Deriving Gradient-based Implicit Saliency

Stochastic Surprisal: An Inferential Measurement of Free Energy in Neural Networks



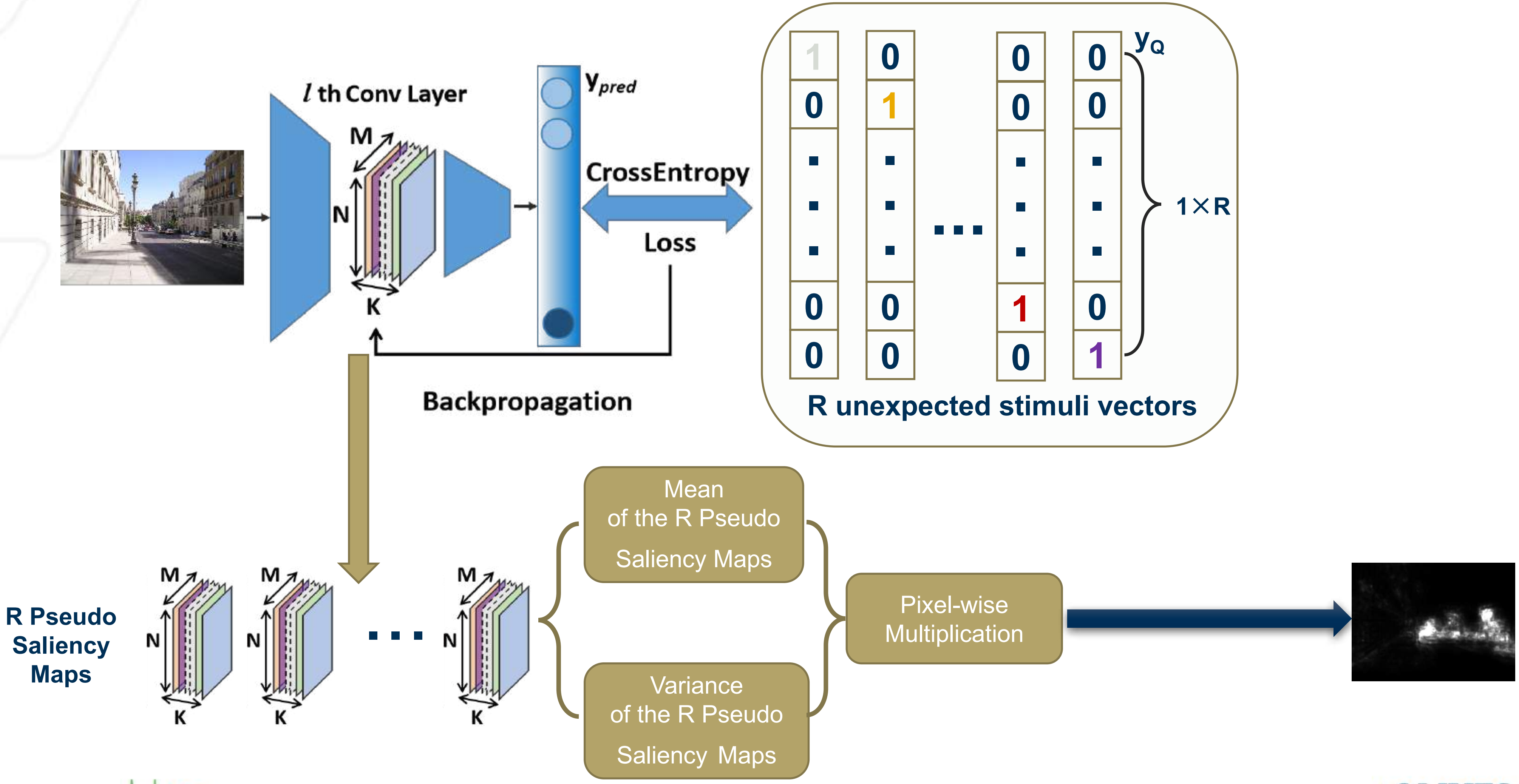
Sun, Yutong, Mohit Prabhushankar, and Ghassan AlRegib. "Implicit saliency in deep neural networks." *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020.

Implicit Saliency

Deriving Gradient-based Implicit Saliency

SCAN ME

Stochastic Surprisal: An Inferential Measurement of Free Energy in Neural Networks

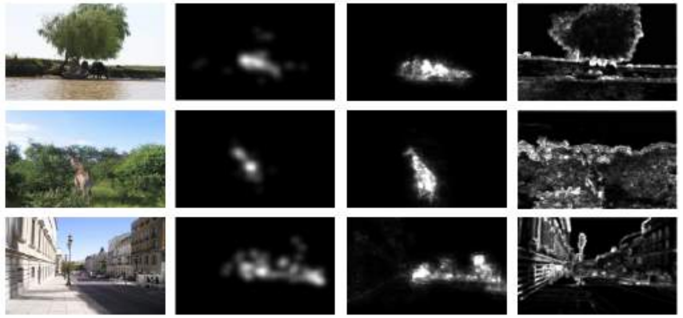


Sun, Yutong, Mohit Prabhushankar, and Ghassan AlRegib. "Implicit saliency in deep neural networks." *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020.

Stochastic Surprisal: An Inferential Measurement of Free Energy in Neural Networks

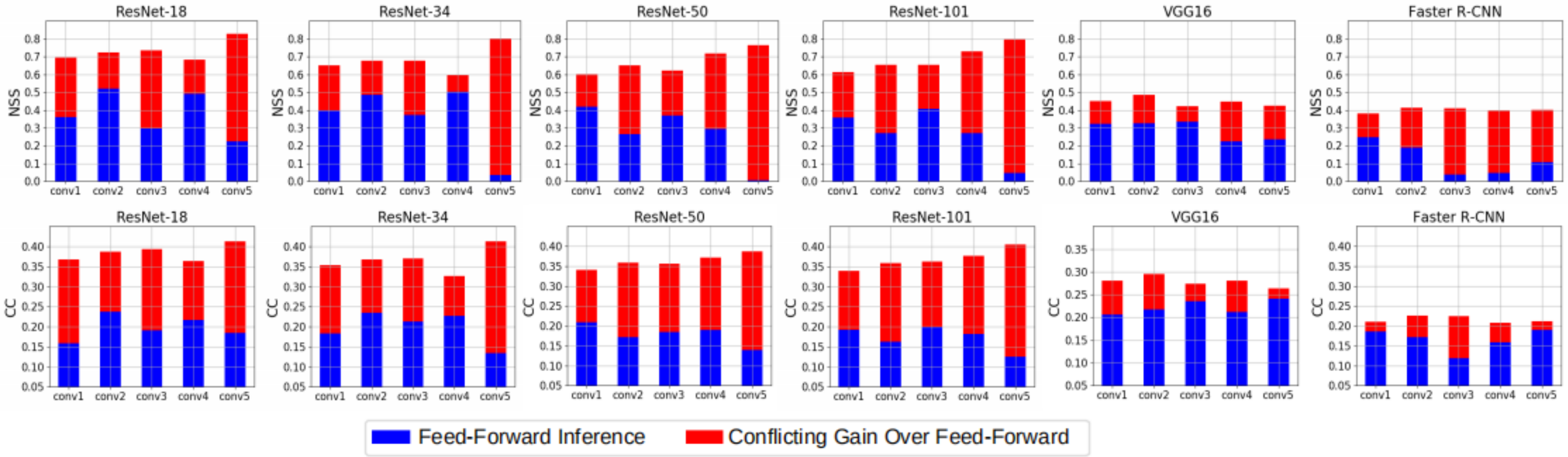
SCAN ME

Contrastive saliency is correlated with attention more than its Feed-Forward counterpart



Input Image | Groundtruth | Proposed Method | Feed-forward feature

- Feed-forward expectation features:
  - Edges and textures
  - Without specific localization
- Proposed expectation-mismatch Saliency:
  - Localized saliency maps
  - Highly correlated with ground truth



Sun, Yutong, Mohit Prabhushankar, and Ghassan AlRegib. "Implicit saliency in deep neural networks." *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020.

Stochastic Surprisal: An Inferential Measurement of Free Energy in Neural Networks

SCAN ME

Contrastive Saliency outperforms explanation methods like GradCAM and Guided Backprop

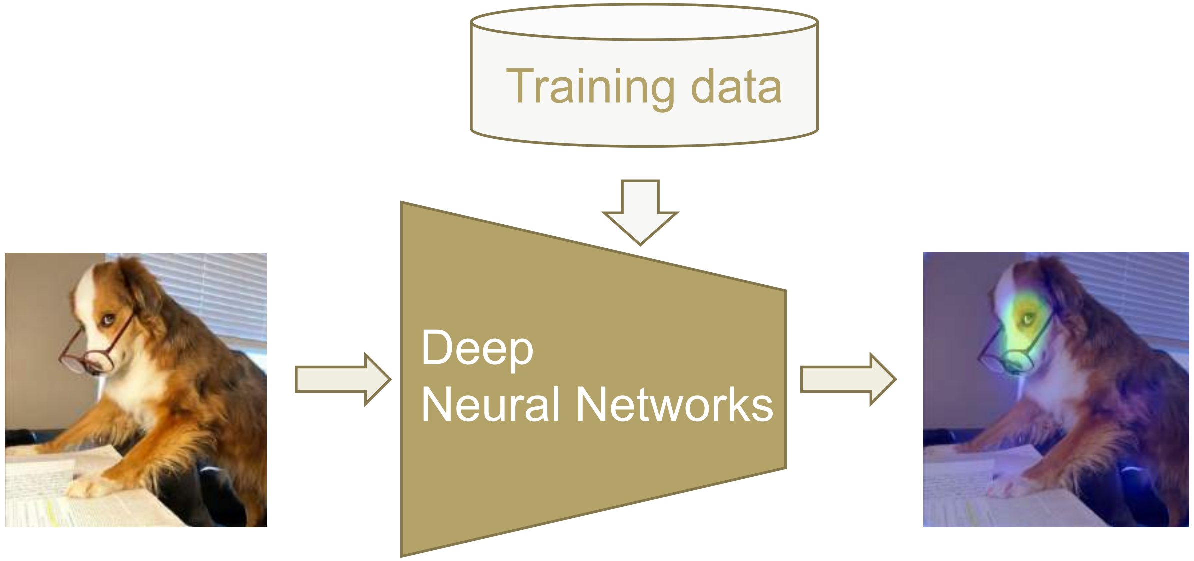
	NSS				CC			
Networks	ResNet-18	ResNet-34	ResNet-50	ResNet-101	ResNet-18	ResNet-34	ResNet-50	ResNet-101
GradCam	0.7657	0.7545	0.7203	0.7335	0.3496	0.3396	0.3190	0.3210
GBP	0.3862	0.4191	0.3898	0.3415	0.2474	0.2453	0.2443	0.2233
Contrastive Saliency	**0.8274**	**0.8018**	**0.7659**	**0.7981**	**0.4132**	**0.4112**	**0.3868**	**0.4051**

Input Image

GradCam

Sun, Yutong, Mohit Prabhushankar, and Ghassan AlRegib. "Implicit saliency in deep neural networks." *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020.

Compare performance of unsupervised Contrastive Saliency model against existing saliency models



Existing Learning based methods

Contrastive Saliency is unsupervised!

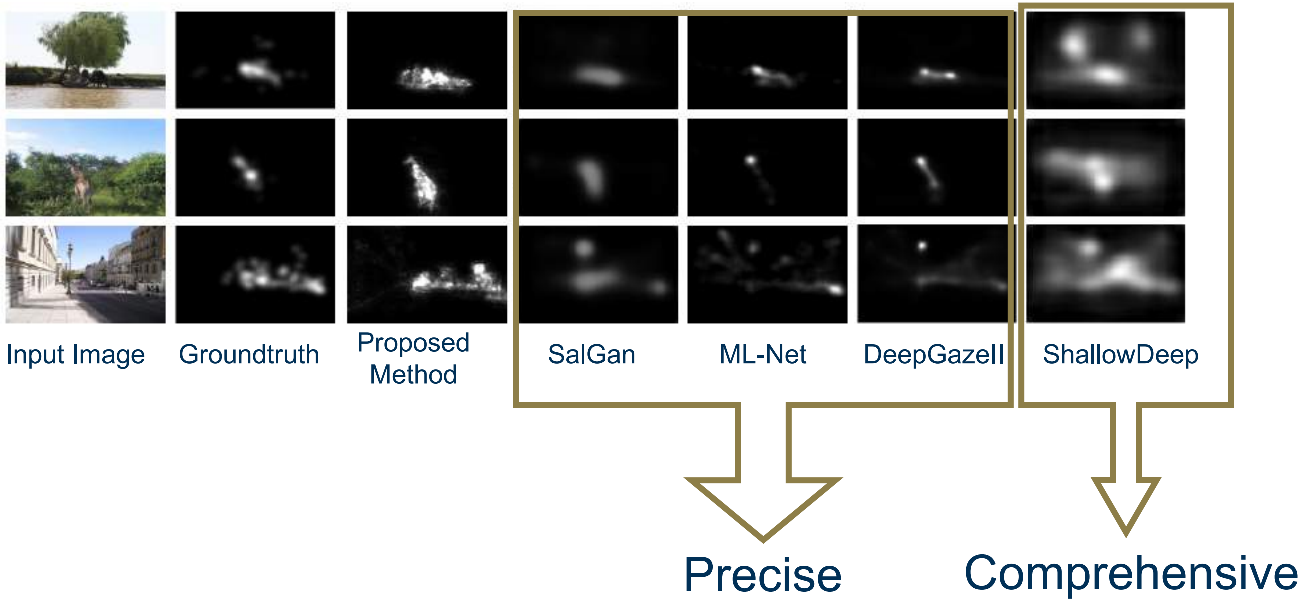
Saliency Models	Training data
SalGan	SALICON
ML-Net	SALICON
DeepGazeII	SALICON
ShallowDeep	SALICON/iSUN

Sun, Yutong, Mohit Prabhushankar, and Ghassan AlRegib. "Implicit saliency in deep neural networks." *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020.

Stochastic Surprisal: An Inferential Measurement of Free Energy in Neural Networks

SCAN ME

Compare performance of unsupervised Contrastive Saliency model against existing saliency models



NSS					CC				
Sal Gan	Deep GazeII	ML Net	Shallow Deep	**Contrastive Saliency**	Sal Gan	Deep GazeII	ML Net	Shallow Deep	**Contrastive Saliency**
0.8977	0.6214	0.5431	**0.9306**	0.7981	**0.6280**	0.5927	0.4481	0.5120	0.4051

Sun, Yutong, Mohit Prabhushankar, and Ghassan AlRegib. "Implicit saliency in deep neural networks." *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020.

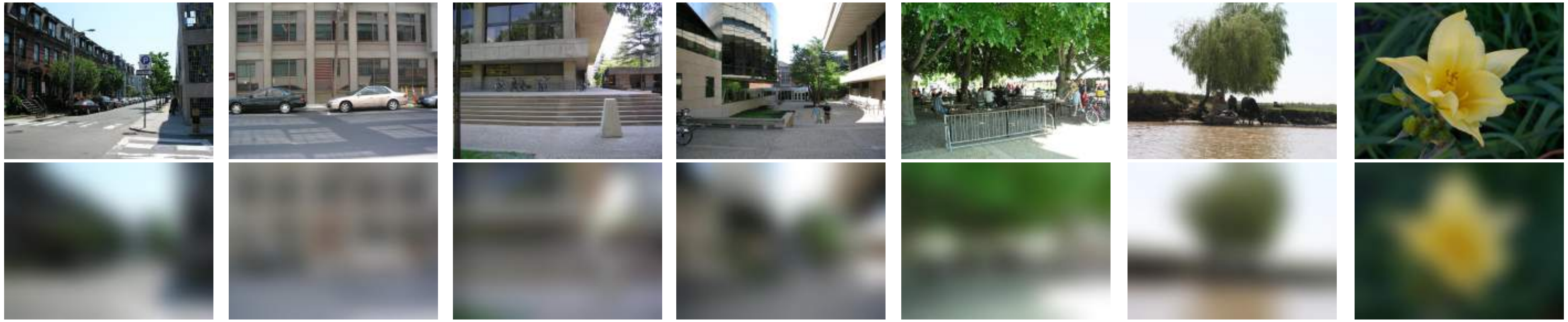
Implicit Saliency

Robustness Analysis

Stochastic Surprisal: An Inferential Measurement of Free Energy in Neural Networks

SCAN ME

Contrastive Saliency drops the least performance with noise added



	NSS					CC				
Gaussian Blur	Sal Gan	Deep GazeII	ML Net	Shallow Deep	**Contrastive Saliency**	Sal Gan	Deep GazeII	ML Net	Shallow Deep	**Contrastive Saliency**
$r = 0$	0.8977	0.6214	0.5431	**0.9306**	0.7981	**0.6280**	0.5927	0.4481	0.5120	0.4051
$r = 50$	↓ 0.2239	↓ 0.3436	↓ 0.2484	↓ 0.2025	↓ **0.1793**	↓ 0.2731	↓ 0.3954	↓ 0.2940	↓ 0.1840	↓ **0.1432**

Sun, Yutong, Mohit Prabhushankar, and Ghassan AlRegib. "Implicit saliency in deep neural networks." *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020.

Expectancy-Mismatch

Our Goal: Introduce Expectancy-Mismatch in Neural Networks

At Inference, construct local contrastive manifolds

Change in Network Parameters: Expectancy-Mismatch when presented with novel data!

We demonstrate on two applications:

1. Human Visual Saliency
2. Image Quality Assessment

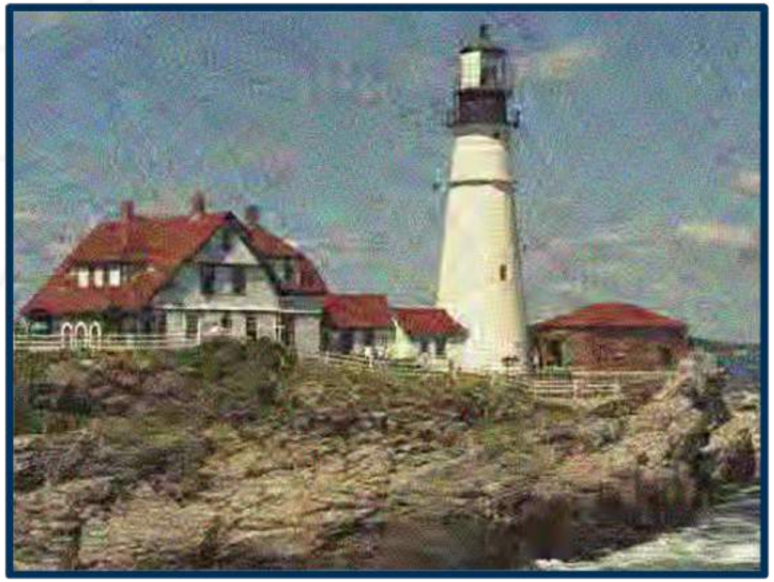
Image Quality Assessment

What is IQA?

Stochastic Surprisal: An Inferential Measurement of Free Energy in Neural Networks

SCAN ME

IQA is the objective Assessment of Subjective Quality



Lighthouse image with level 5 lossy compression from TID 2013 dataset

Image Quality Assessment Algorithm : DIQaM [1]

Score : 0.58

The given image is somewhat OK quality

Bad Quality

Good Quality

[1] Bosse S, Maniry D, Müller K R, et al. Deep neural networks for no-reference and full-reference image quality assessment. IEEE Transactions on Image Processing, 2018, 27(1): 206-219.

Image Quality Assessment

Expectancy-Mismatch in Dataset Construction

Expectancy-Mismatch arises during Dataset Construction

- Subjects are shown a reference image in a controlled setting
- Based on the reference image, they are asked to pick one of the images on the top that differs least from the reference image
- Reference image sets the expectancy
- The task of subjectively picking the least mis-matched image is IQA

This requires **Fine-grained** Analysis!

[1] Ponomarenko, Nikolay, et al. "Image database TID2013: Peculiarities, results and perspectives." *Signal processing: Image communication* 30 (2015): 57-77

Image Quality Assessment

Expectancy-Mismatch in Dataset Construction

Expectancy-Mismatch arises during Dataset Construction

This requires **Fine-grained** Analysis on the part of the subjects!

Our Goal: To determine if a trained IQA detector understands the fine-grained nature of expectancy-mismatch in quality

[1] Ponomarenko, Nikolay, et al. "Image database TID2013: Peculiarities, results and perspectives." *Signal processing: Image communication* 30 (2015): 57-77

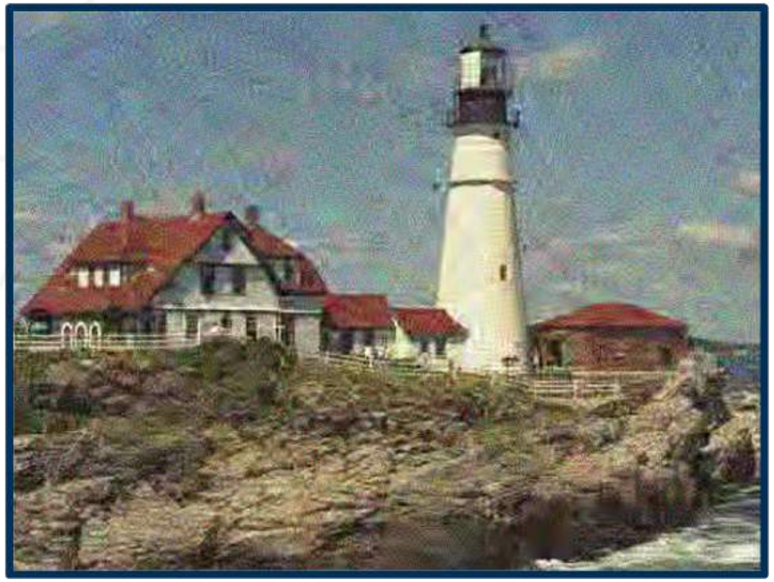
Image Quality Assessment

GradCAM in IQA

Stochastic Surprisal: An Inferential Measurement of Free Energy in Neural Networks

SCAN ME

GradCAM explanation for Why 0.58?



Lighthouse image with level 5 lossy compression from TID 2013 dataset

The given image is somewhat OK quality

DIQaM : 0.58

Grad-CAM

Why 0.58?

Bad Quality

Good Quality

0.0 0.5 1.0

Add heatmap
Explain blue
Yellow, red, green

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.

Stochastic Surprisal: An Inferential Measurement of Free Energy in Neural Networks

SCAN ME

GradCAM explanation may not be useful for fine-grained analysis

Grad-CAM explanation tells us that the quality score was decided based on all parts of the image and specifically based on the base of the lighthouse

Grad-CAM

Why 0.58?

Lighthouse image with level 5 lossy compression from TID 2013 dataset

Bad Quality

Good Quality

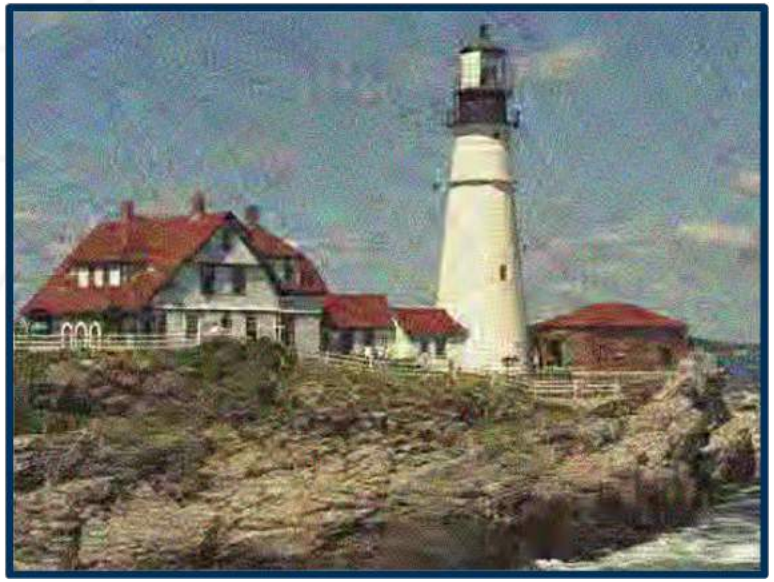
0.0 0.5 1.0

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.

SCAN ME

Stochastic Surprisal: An Inferential Measurement of Free Energy in Neural Networks

All the distortions in the foreground prevent a quality score of 1



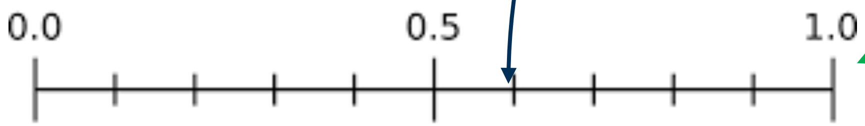
Lighthouse image with level 5 lossy compression from TID 2013 dataset

DIQaM : 0.58

Contrastive explanation

Why 0.58, rather than 1?

Bad Quality



Good Quality

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.

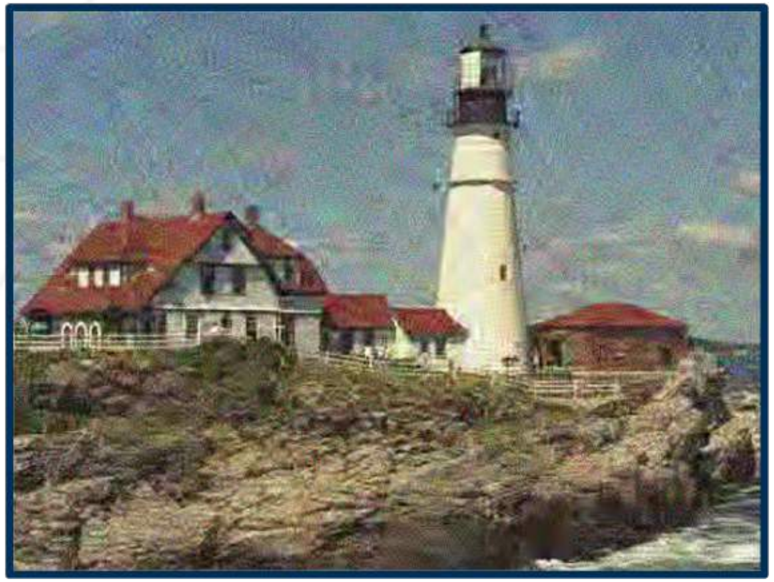
Image Quality Assessment

ContrastCAM in IQA

Stochastic Surprisal: An Inferential Measurement of Free Energy in Neural Networks

SCAN ME

The distortions on the lighthouse and houses prevent a higher score of 0.75



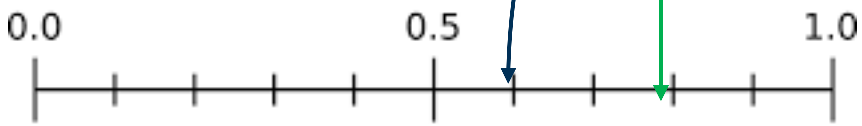
Lighthouse image with level 5 lossy compression from TID 2013 dataset

DIQaM : 0.58

Contrastive explanation

Why 0.58, rather than 0.75?

Bad Quality



Good Quality

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.

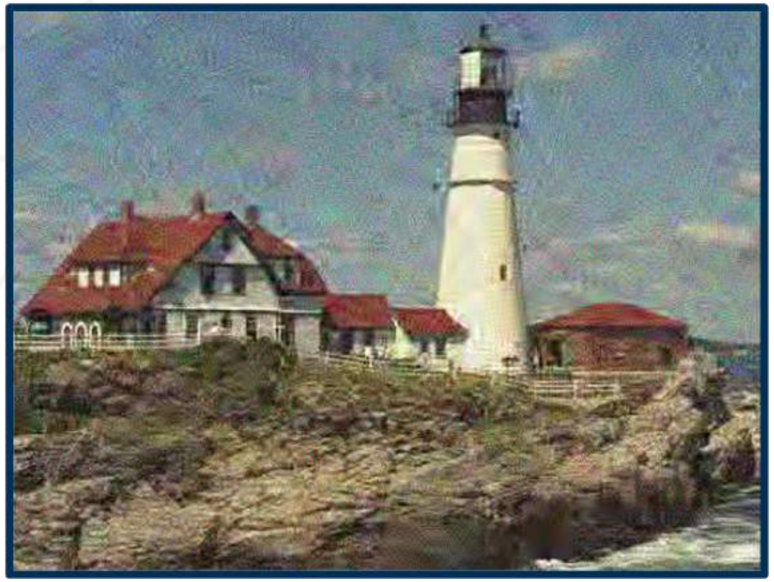
Image Quality Assessment

ContrastCAM in IQA

Stochastic Surprisal: An Inferential Measurement of Free Energy in Neural Networks

SCAN ME

The quality of the lighthouse and sky is better than a score of 0.5



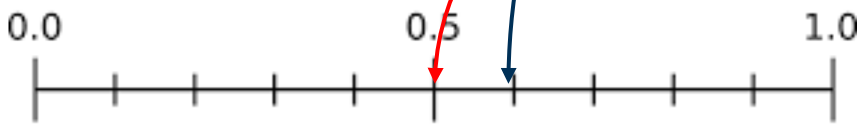
Lighthouse image with level 5 lossy compression from TID 2013 dataset

DIQaM : 0.58

Contrastive explanation

Why 0.58, rather than 0.5?

Bad Quality



Good Quality

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.

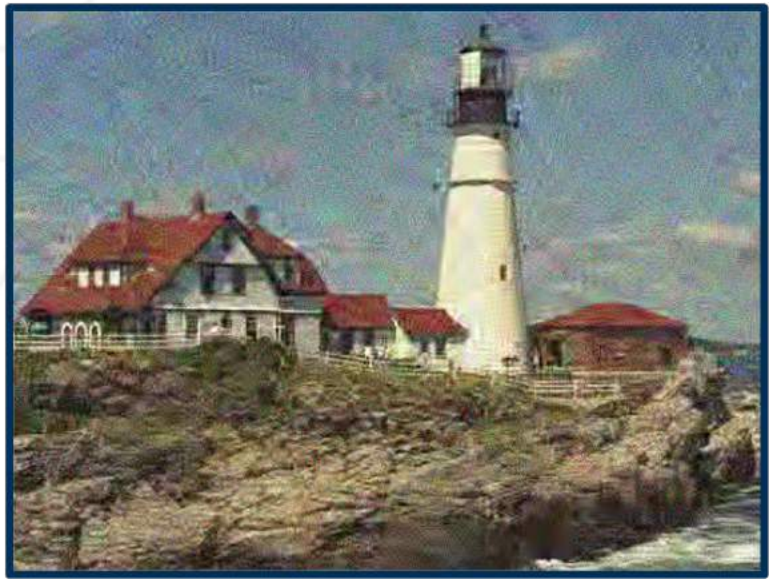
Image Quality Assessment

ContrastCAM in IQA

Stochastic Surprisal: An Inferential Measurement of Free Energy in Neural Networks

SCAN ME

The sky, lighthouse, and cliff merit a quality higher than 0.25



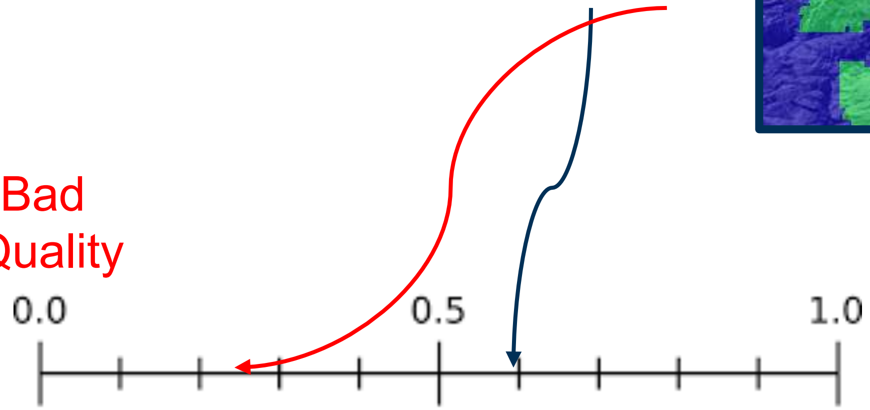
Lighthouse image with level 5 lossy compression from TID 2013 dataset

DIQaM : 0.58

Contrastive explanation

Why 0.58, rather than 0.25?

Bad Quality



Good Quality

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.

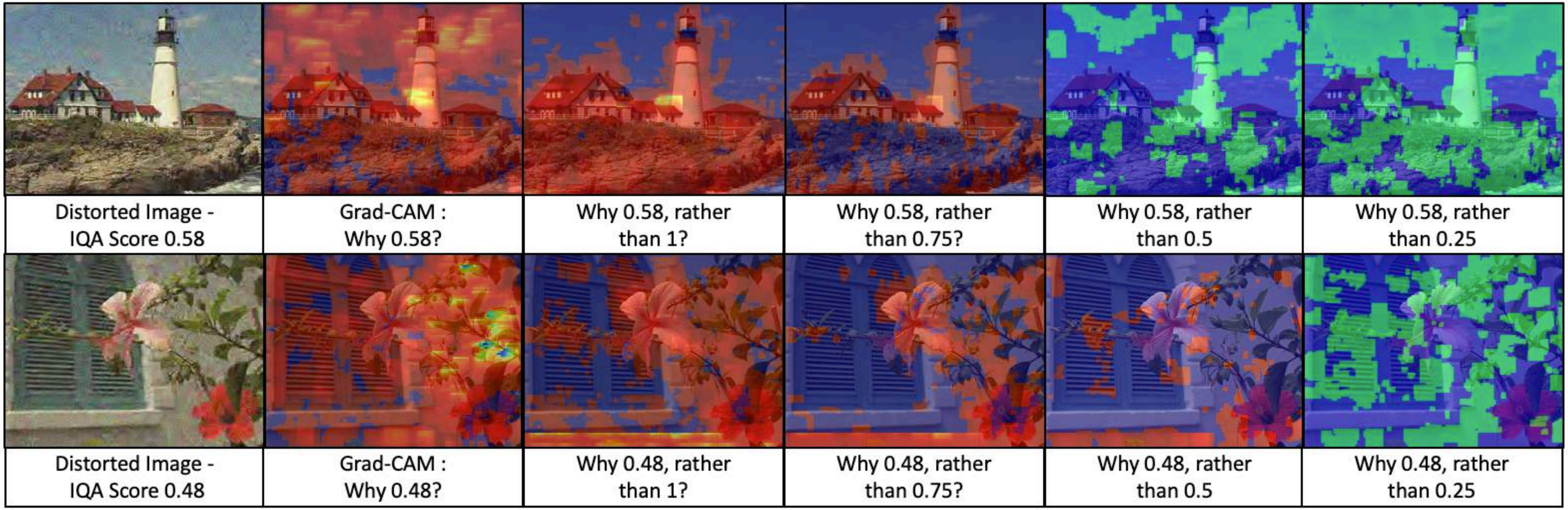
Image Quality Assessment

ContrastCAM in IQA

SCAN ME

Stochastic Surprisal: An Inferential Measurement of Free Energy in Neural Networks

Contrastive IQA elicits the fine-grained decisions made by the network



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.

Objectives

Takeaways from Part IV

- Part I: Gradients in Neural Networks
- Part 2: Gradients as Information
- Part 3: Gradients as Uncertainty
- **Part 4: Gradients as Expectancy-Mismatch**
    - Presented a case study of utilizing both the contrastive manifolds and manifold traversal perspectives
    - Human Visual Saliency is a by-product of expectancy-mismatch
    - Neural networks that have never explicitly learned human salient regions have implicitly been trained to use them in tasks
    - Using Contrastive explanations in IQA provides a fine-grained analysis of neural network's perception of quality
- Part 5: Conclusion and Future Directions

Interpretation, and Applications of Gradients

Part 5: Conclusions and Future Directions

Key Takeaways
Role of Gradients

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

Future Directions

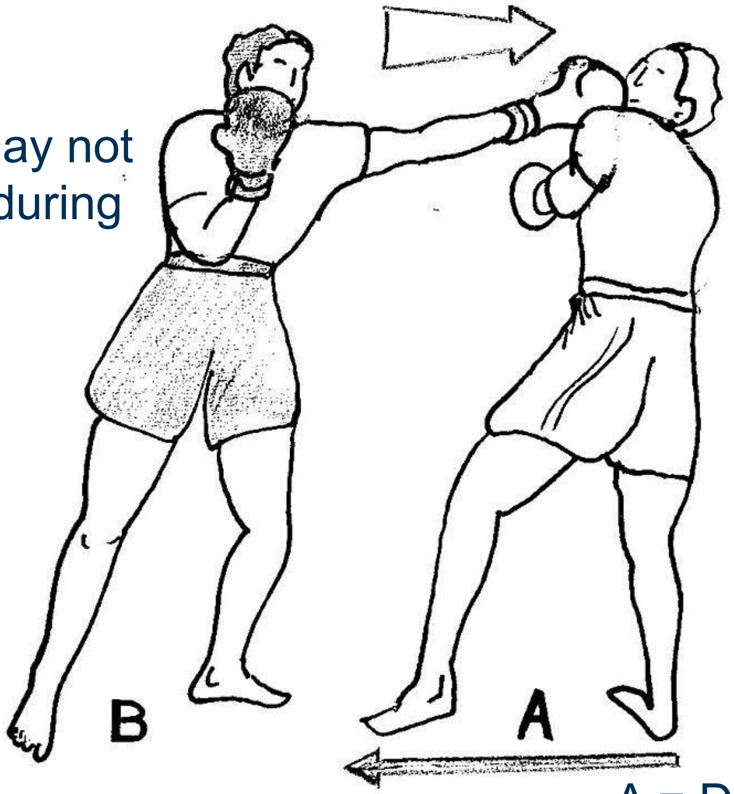
Research at Inference Stage

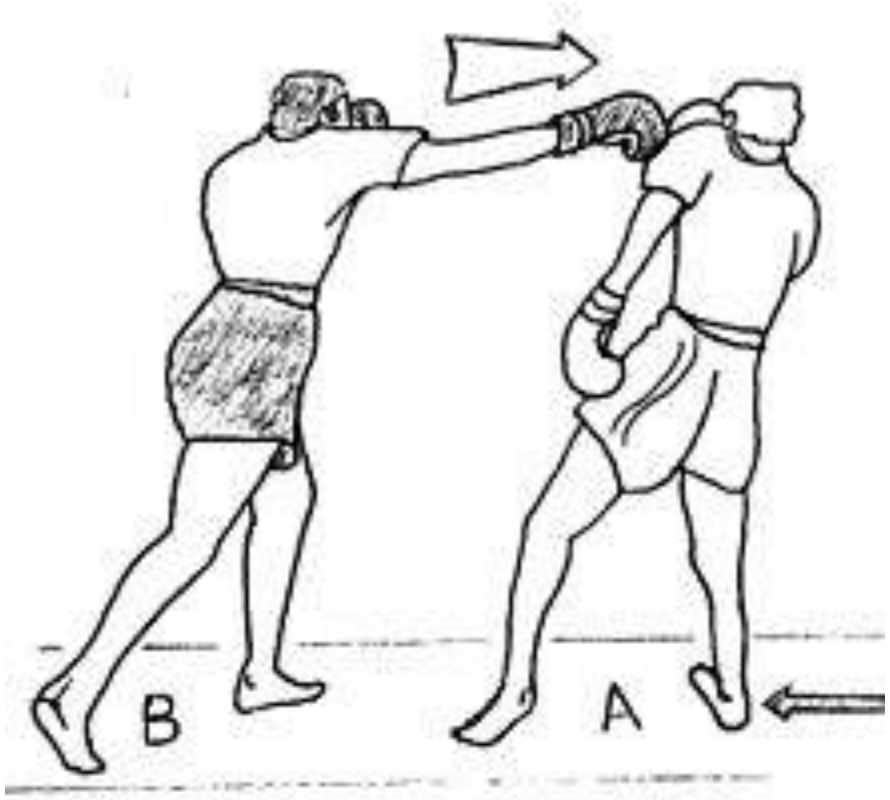
- **Test Time Augmentation (TTA) Research**
  - Multiple augmentations of data are passed through the network at inference
  - Research is in designing the best augmentations
- **Active Inference**
  - Utilize the knowledge in Neural Networks to ***ask it to ask us***
  - Neural networks ask for the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

Deep Learning and Novel Data

Deep learning cannot easily generalize to novel data

Novel data may not be available during training





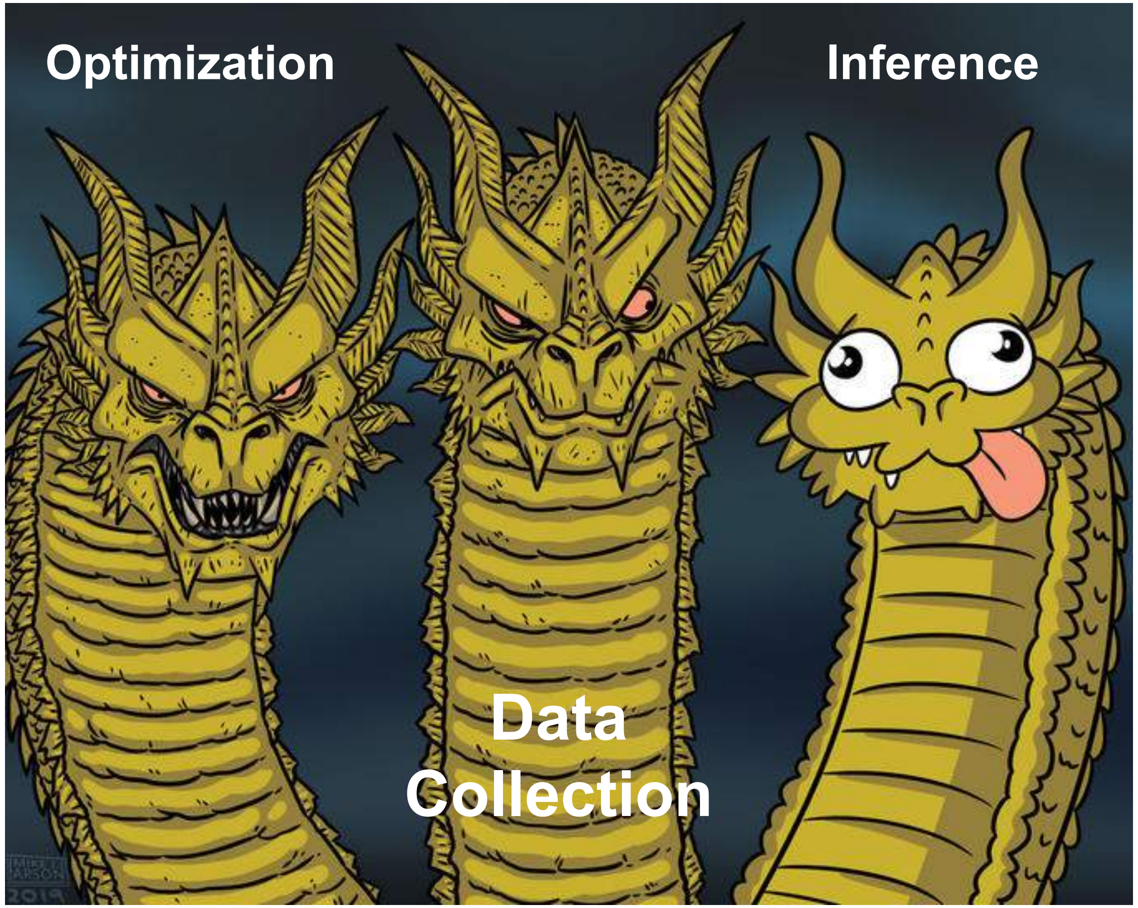
Even if available, novel data does not easily fit into either the earlier or later stages of training

A = Deep Neural Networks
B = Novel data

Memes to Wrap it Up

Robustness Research in the Inferential Stage of Neural Networks

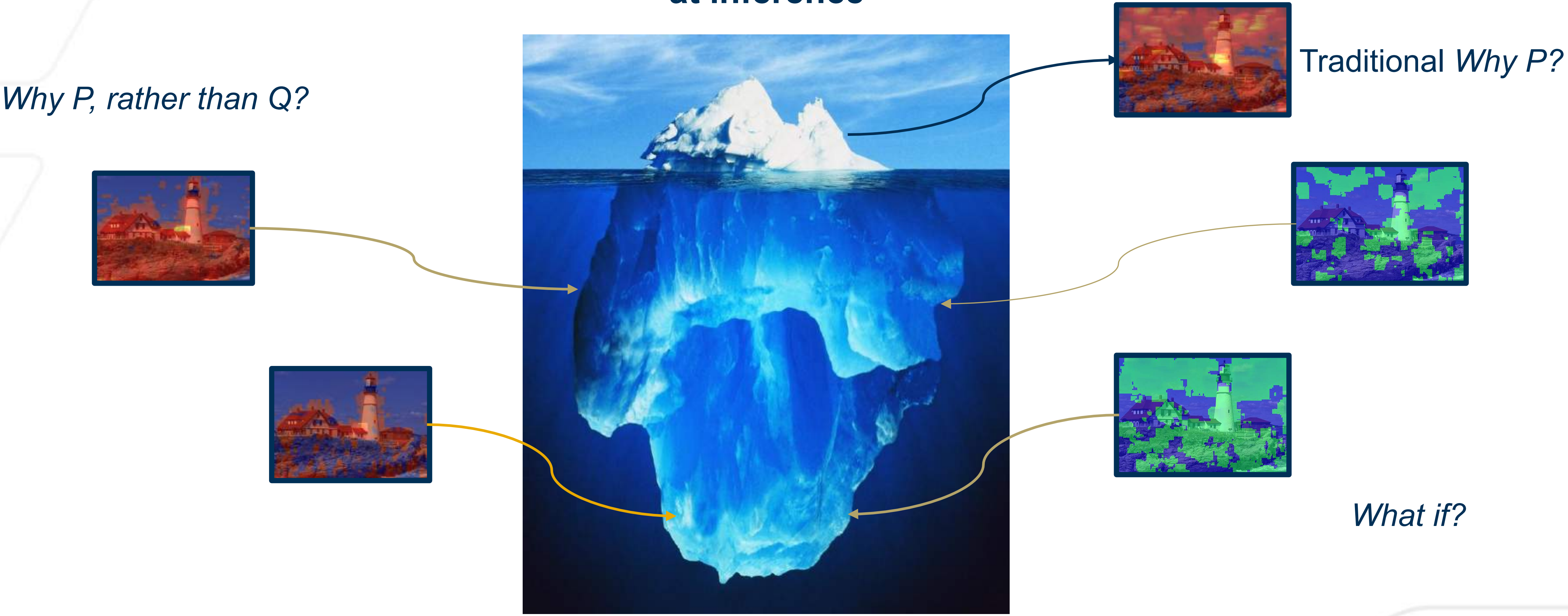
Existing research on robustness focuses on data collection and optimization



Memes to Wrap it Up

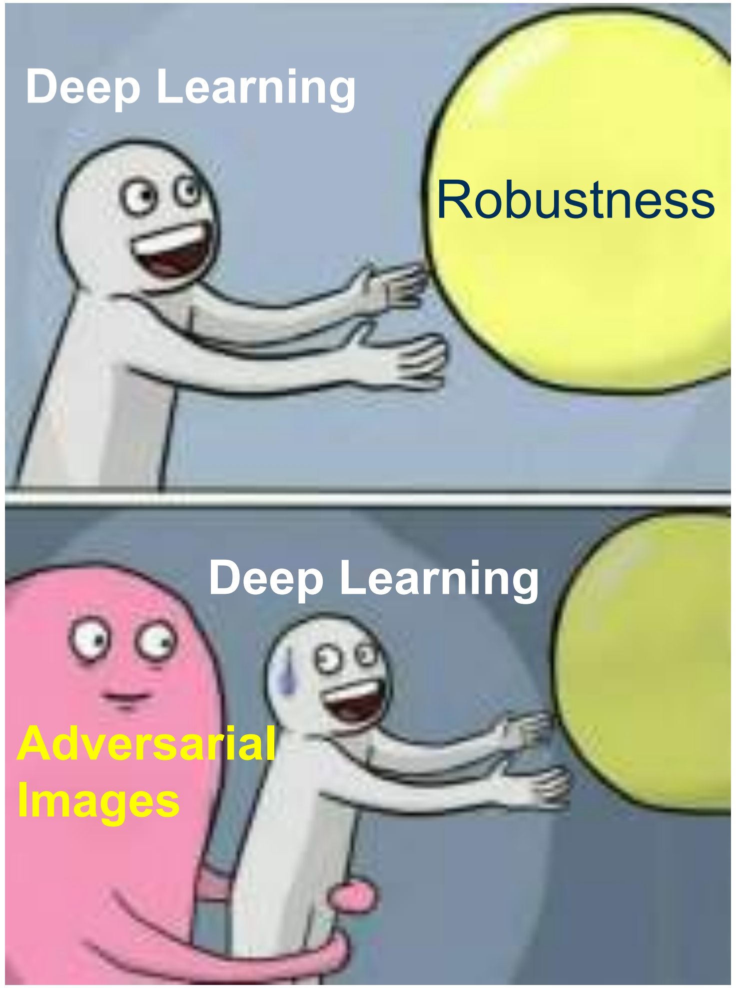
Implicit Knowledge in Neural Networks

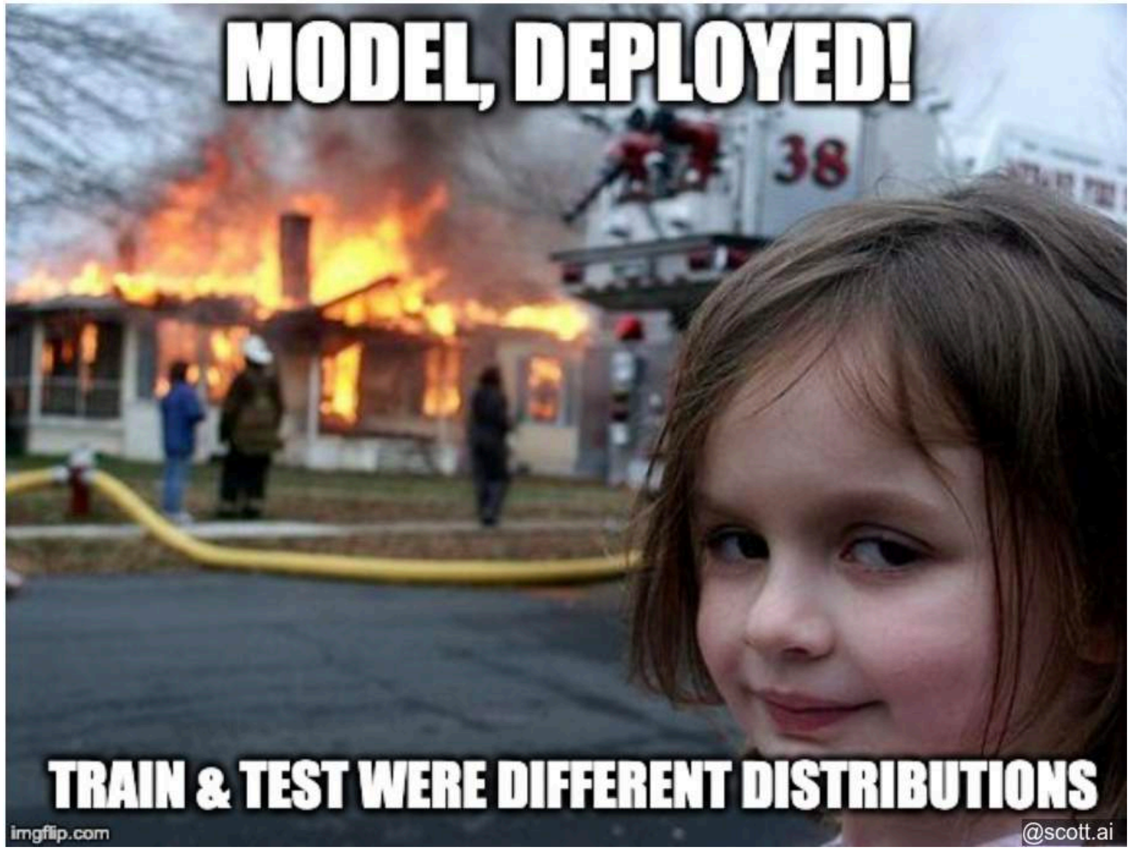
Trained Neural Networks have a wealth of implicit stored knowledge, waiting to be extracted at inference



Memes to Wrap it Up

Robustness at Inference





Cannot depend on training to construct robust models

References

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

- **Gradients for robustness against noise:** M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022
- **Gradients for adversarial, OOD, corruption detection:** J. Lee, M. Prabhushankar, and G. AlRegib, "Gradient-Based Adversarial and Out-of-Distribution Detection," in *International Conference on Machine Learning (ICML) Workshop on New Frontiers in Adversarial Machine Learning*, Baltimore, MD, Jul. 2022.
- **Gradients for Open set recognition:** Lee, Jinsol, and Ghassan AlRegib. "Open-Set Recognition With Gradient-Based Representations." *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021.
- **GradCon for Anomaly Detection:** Kwon, G., Prabhushankar, M., Temel, D., & AlRegib, G. (2020, August). Backpropagated gradient representations for anomaly detection. In *European Conference on Computer Vision* (pp. 206-226). Springer, Cham.
- **Gradients for adversarial, OOD, corruption detection :** J. Lee, C. Lehman, M. Prabhushankar, and G. AlRegib, "Probing the Purview of Neural Networks via Gradient Analysis," in IEEE Access, Mar. 21 2023.
- **Gradients for Novelty Detection:** Kwon, G., Prabhushankar, M., Temel, D., & AlRegib, G. (2020, October). Novelty detection through model-based characterization of neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3179-3183). IEEE.
- **Gradient-based Image Quality Assessment:** G. Kwon*, M. Prabhushankar*, D. Temel, and G. AlRegib, "Distorted Representation Space Characterization Through Backpropagated Gradients," in *IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, Sep. 2019.

Explainability in Neural Networks

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

References

Self Supervised Learning

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

Human Vision and Behavior Prediction

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

Open-source Datasets to assess Robustness

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

References

Active Learning

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

Uncertainty Estimation

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

Tutorial Materials

Accessible Online

SCAN ME

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

{alregib, mohit.p}@gatech.edu

IEEE ICIP 2023 Tutorial

30th ICIP 2023 Kuala Lumpur

IEEE Signal Processing Society™

Title: A Multi-Faceted View of Gradients in Neural Networks: Extraction, Interpretation and Applications in Image Understanding

Type / Duration: Half-Day Tutorial (3h)