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

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

Shannon Information
(Surprise of an event)

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

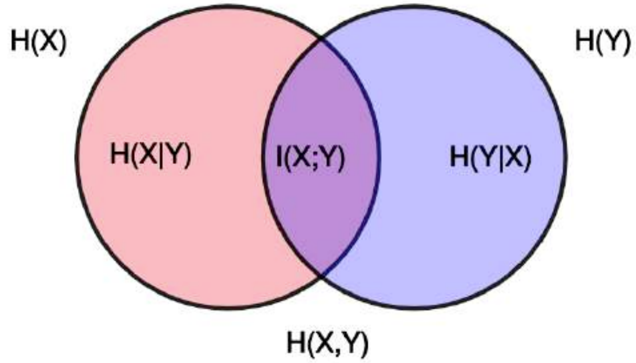
$H[X]$ = Shannon Entropy
$p(x_i)$ = Probability of event x_i

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

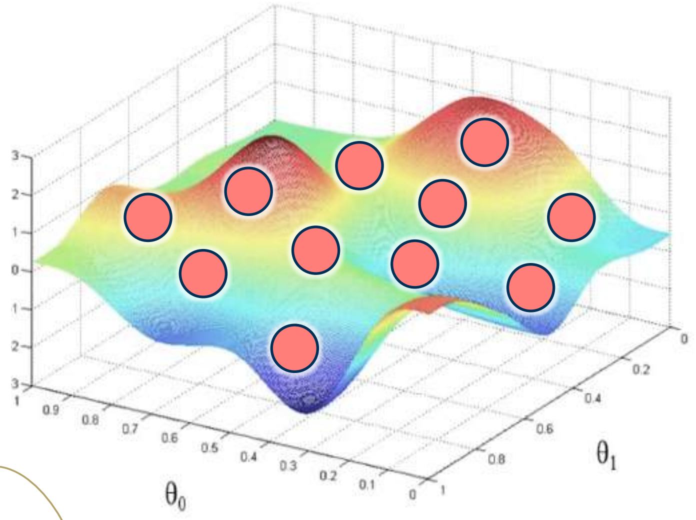
Variance of the partial derivative w.r.t. θ of the Log-likelihood function $\ell(\theta \mid x)$.

Fisher Information

Gradients as Fisher Information

Gradients infer information about the statistics of underlying manifolds

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



Likelihood function instead of loss manifold

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

Using variance decomposition[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!

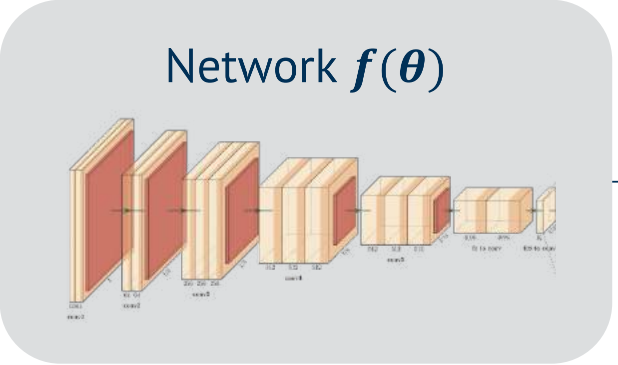
[1] A good blogpost about Fisher Information: https://towardsdatascience.com/an-intuitive-look-at-fisher-information-2720c40867d8

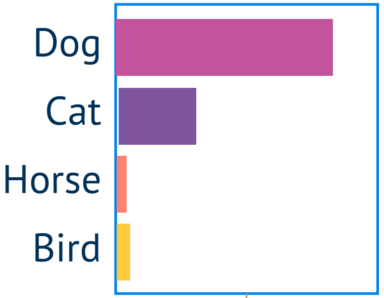
Fisher Information

Gradients as Fisher Information

Gradients infer information about the statistics of underlying manifolds

x

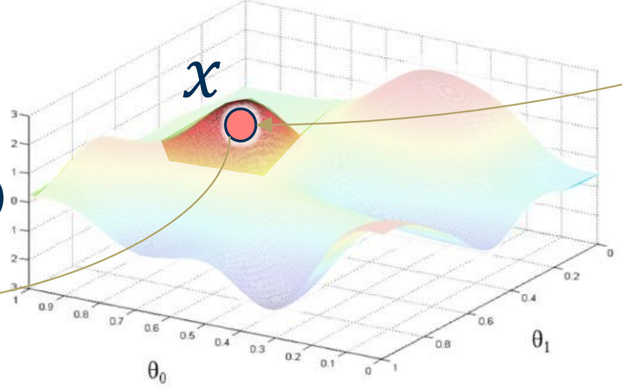




In this case, the image and its prediction extracts nose, mouth and jowl features.

Local information (specific to x) is sufficient!

$l(\theta|x)$



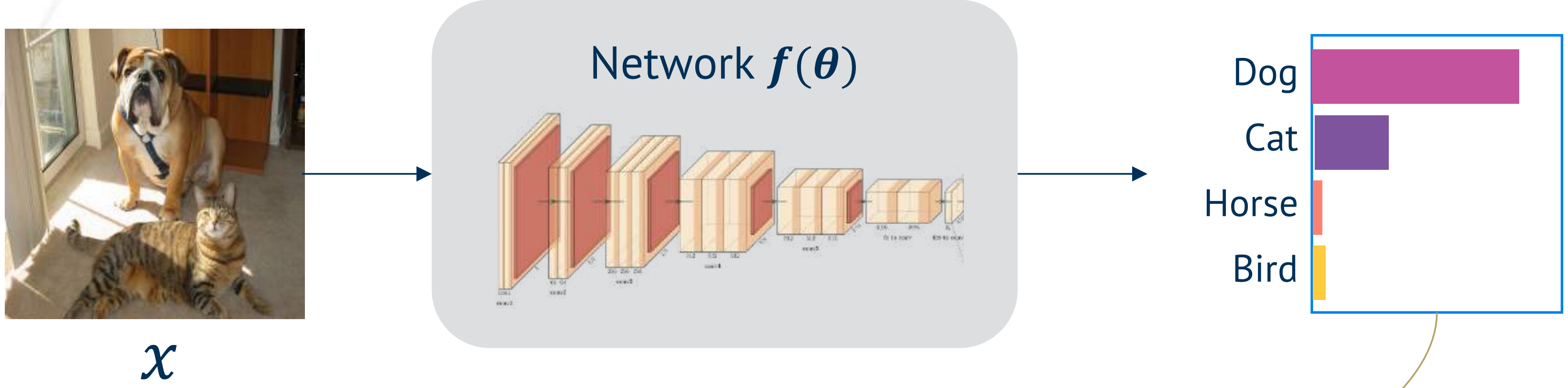
A key feature is that every sample draws information from the underlying distribution! And this information can be visualized.

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

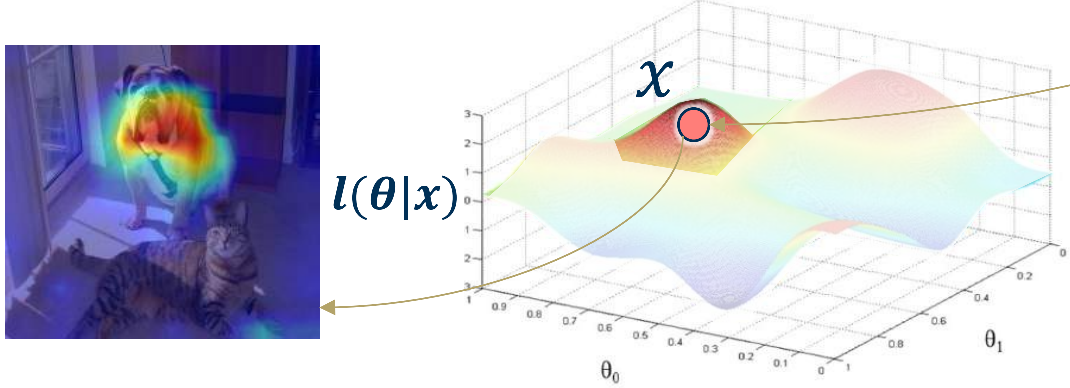
Applicability of Gradient Information

Gradients as Fisher Information

Gradients infer information about the statistics of underlying manifolds



Local information (specific to x) is sufficient!



We demonstrate this in two applications:

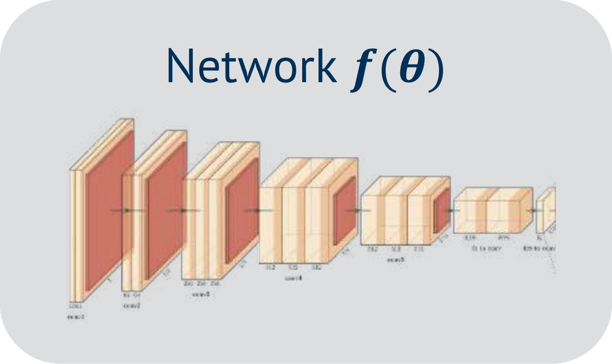
1. Visual Explainability
2. Robust Recognition

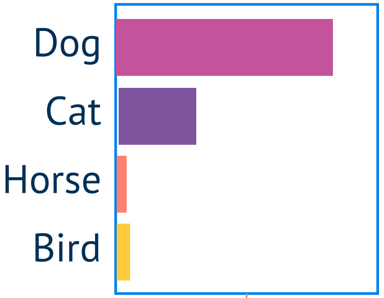
[1] A good blogpost about Fisher Information: https://towardsdatascience.com/an-intuitive-look-at-fisher-information-2720c40867d8

Gradients as Fisher Information

Gradients infer information about the statistics of underlying manifolds

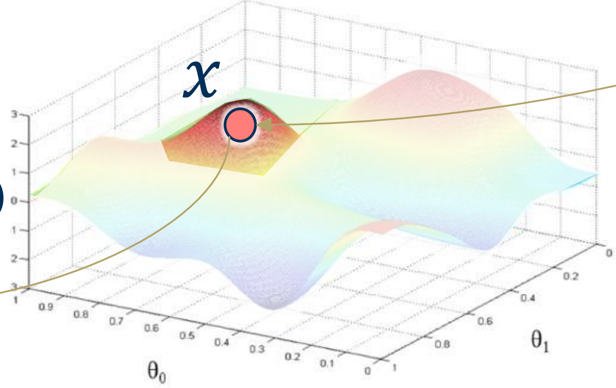
x





Local information (specific to x) is sufficient!

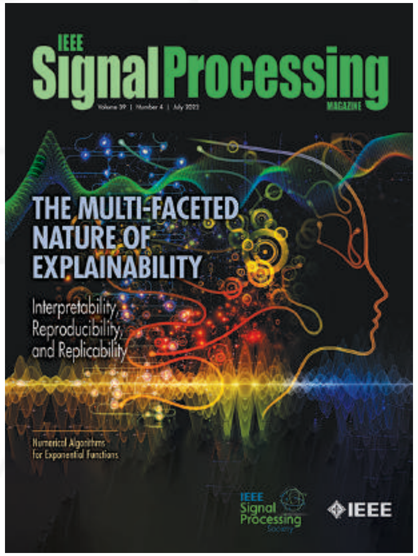
$\boldsymbol{l}(\boldsymbol{\theta}|\boldsymbol{x})$



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



Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

Mohit Prabhushankar, PhD
Postdoc

Ghassan AlRegib, PhD
Professor

SCAN ME

Explanations

Visual Explanations

SCAN ME

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

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

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

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

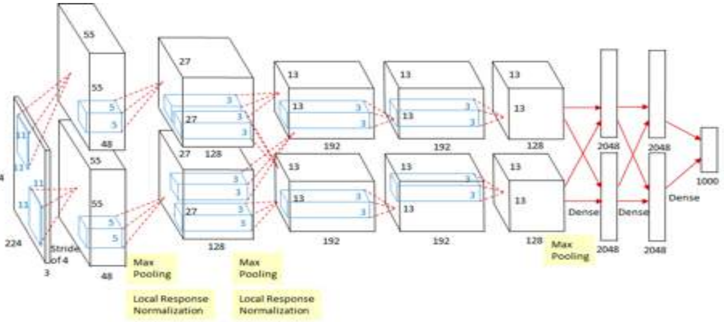
Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

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

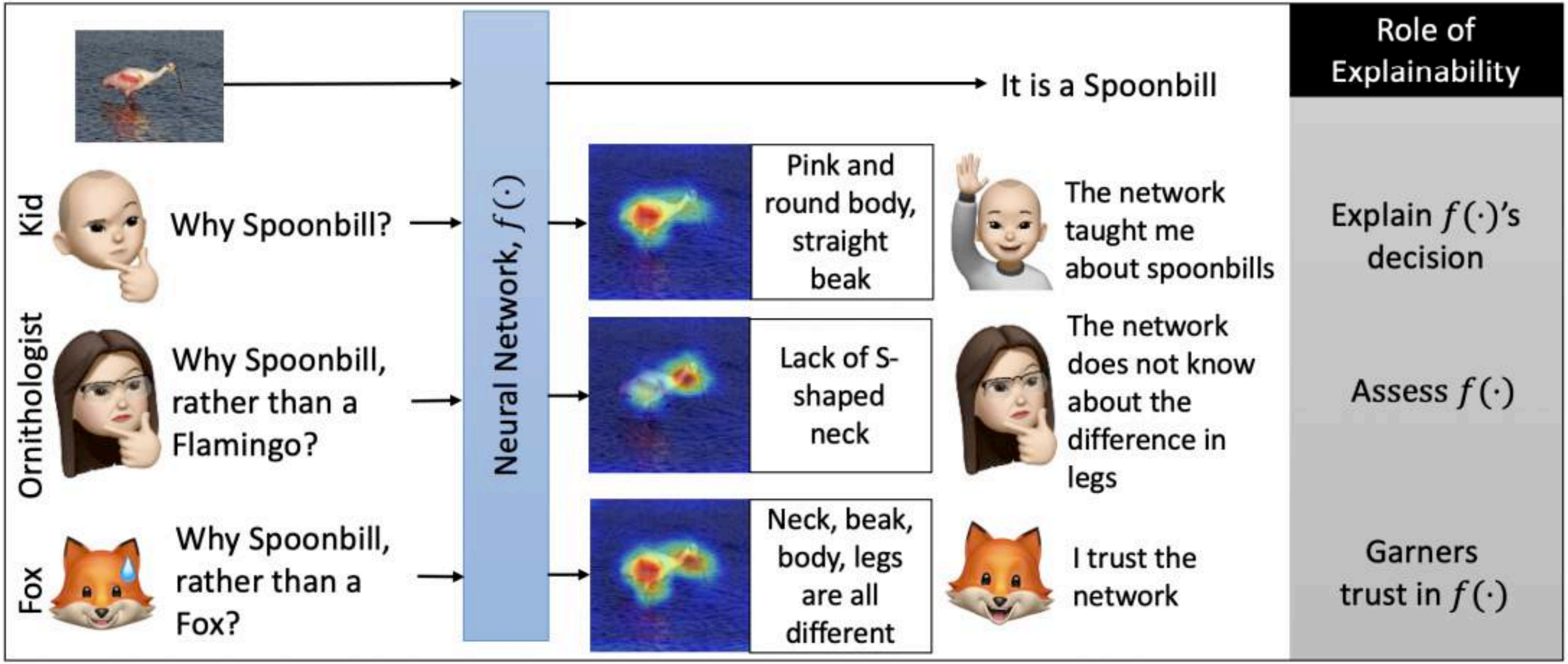
AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks: Towards relevant and contextual explanations. *IEEE Signal Processing Magazine*, *39*(4), 59-72.

Explanations

Role of Explanations – context and relevance

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

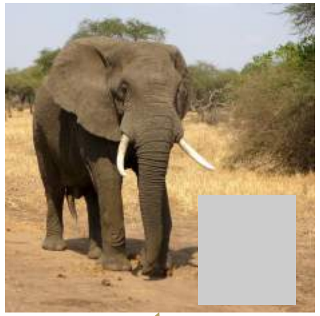


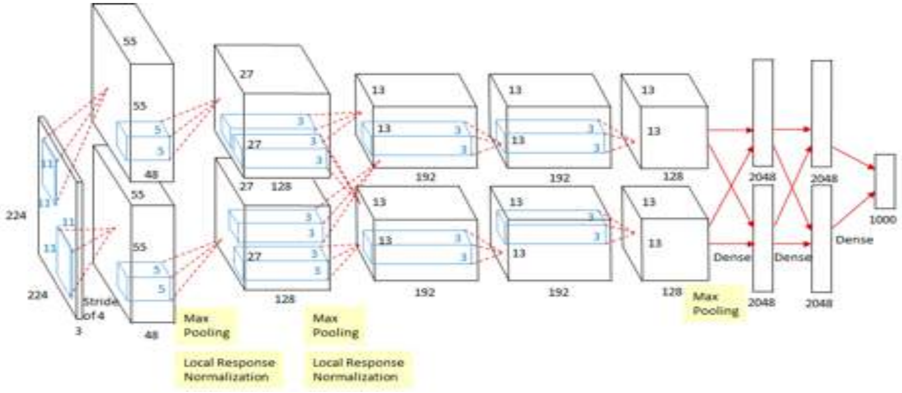
AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks: Towards relevant and contextual explanations. *IEEE Signal Processing Magazine*, *39*(4), 59-72.

SCAN ME

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

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





P(elephant) = 0.95

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

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

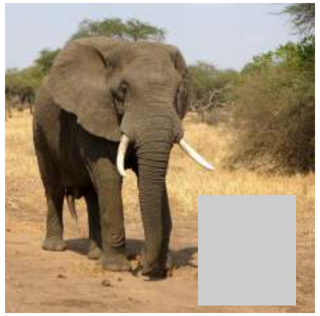
Explanations

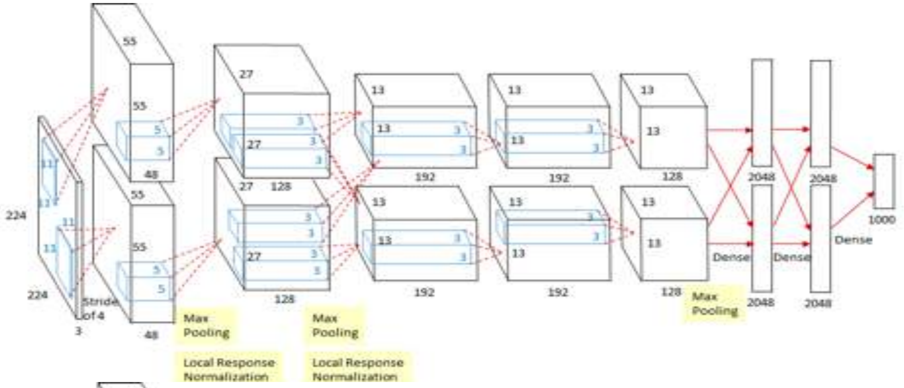
Input Saliency via Occlusions

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

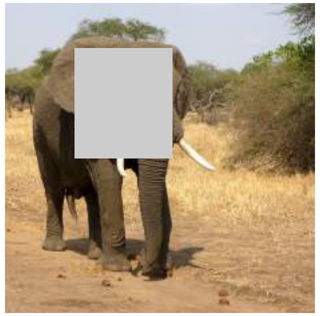
Intervention: Mask part of the image before feeding to CNN, check how much predicted probabilities change

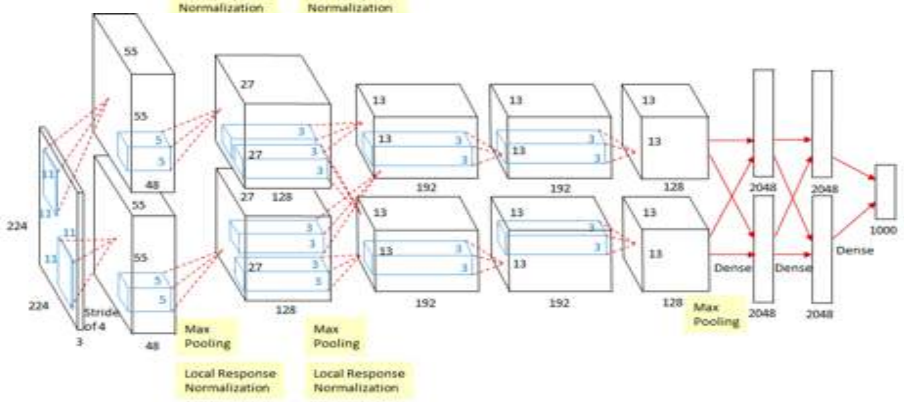




P(elephant) = 0.95

These pixels affect decisions more





P(elephant) = 0.75

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

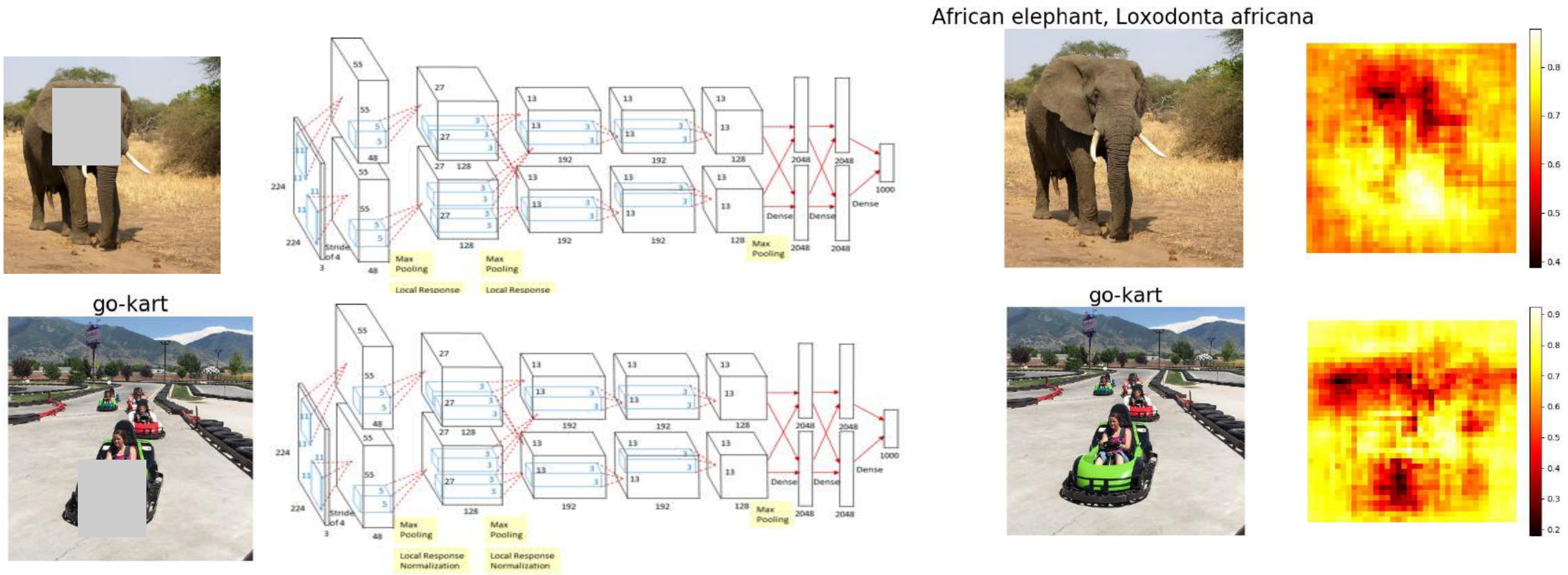
Explanations

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

30th ICIP 2023 Kuala Lumpur

OLIVES @GeorgiaTech

Georgia Tech

Explanations
Gradient-based Explanations

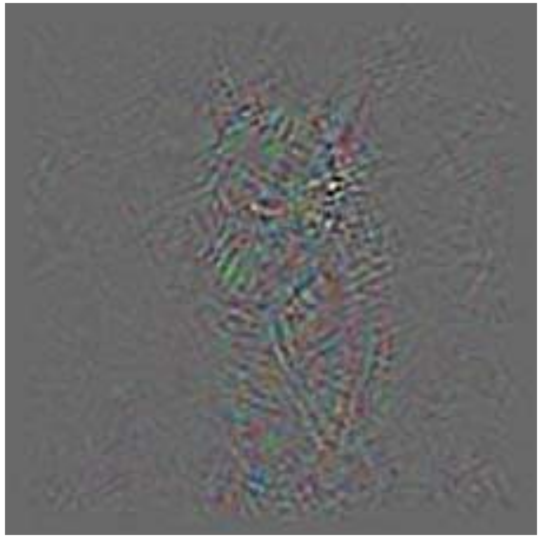
Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

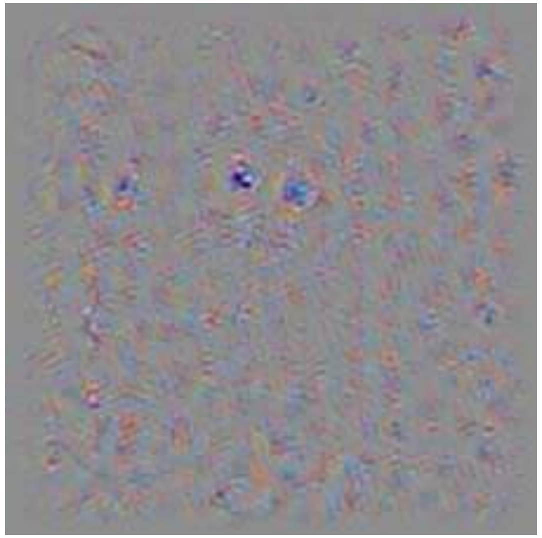
Gradients provide a one-shot means of perturbing the input that changes the output

Input

Vanilla Gradients



Deconvolution Gradients



Guided Backpropagation

However, localization remains an issue

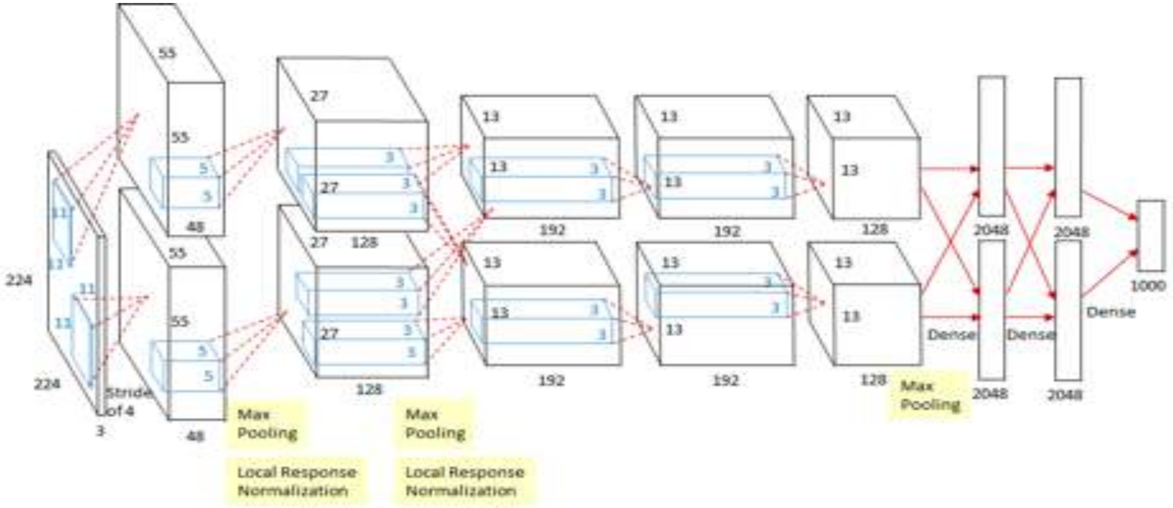
Springenberg, Dosovitskiy, et al., Striving for Simplicity: The all convolutional net, 2015

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

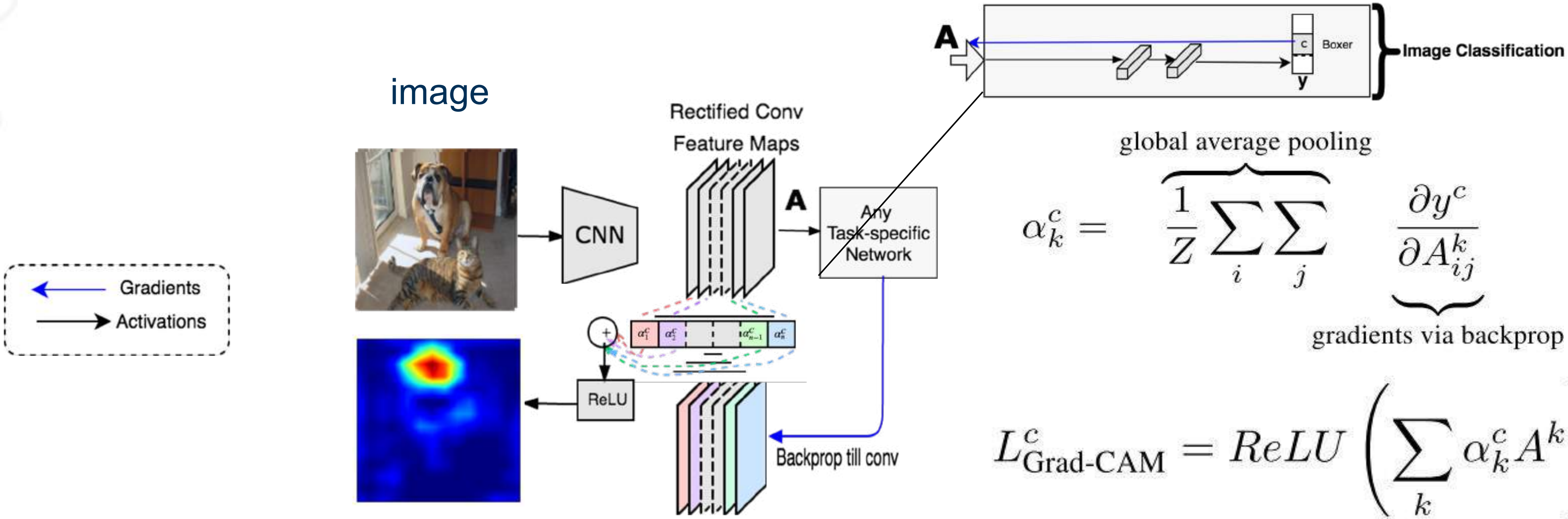


Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

Gradient and Activation-based Explanations

GradCAM

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



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

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

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

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

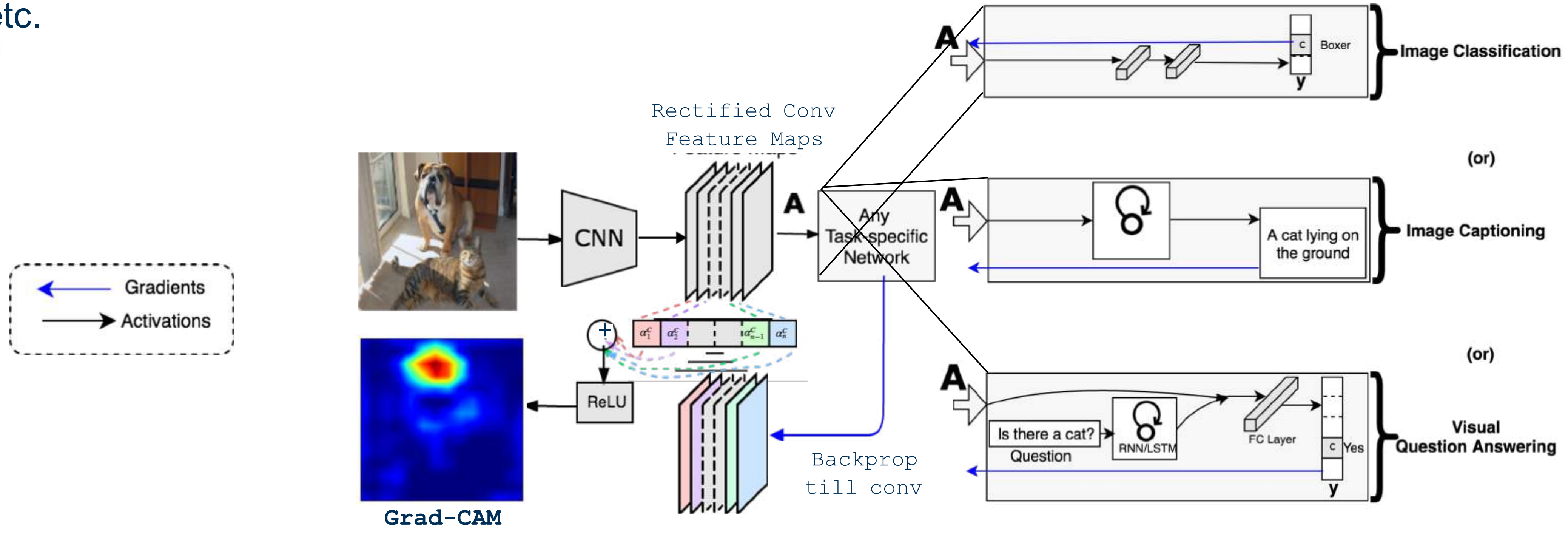
Gradient and Activation-based Explanations
GradCAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

Grad-CAM generalizes to any task:

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



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

Gradient and Activation-based Explanations

Explanatory Paradigms

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

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

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

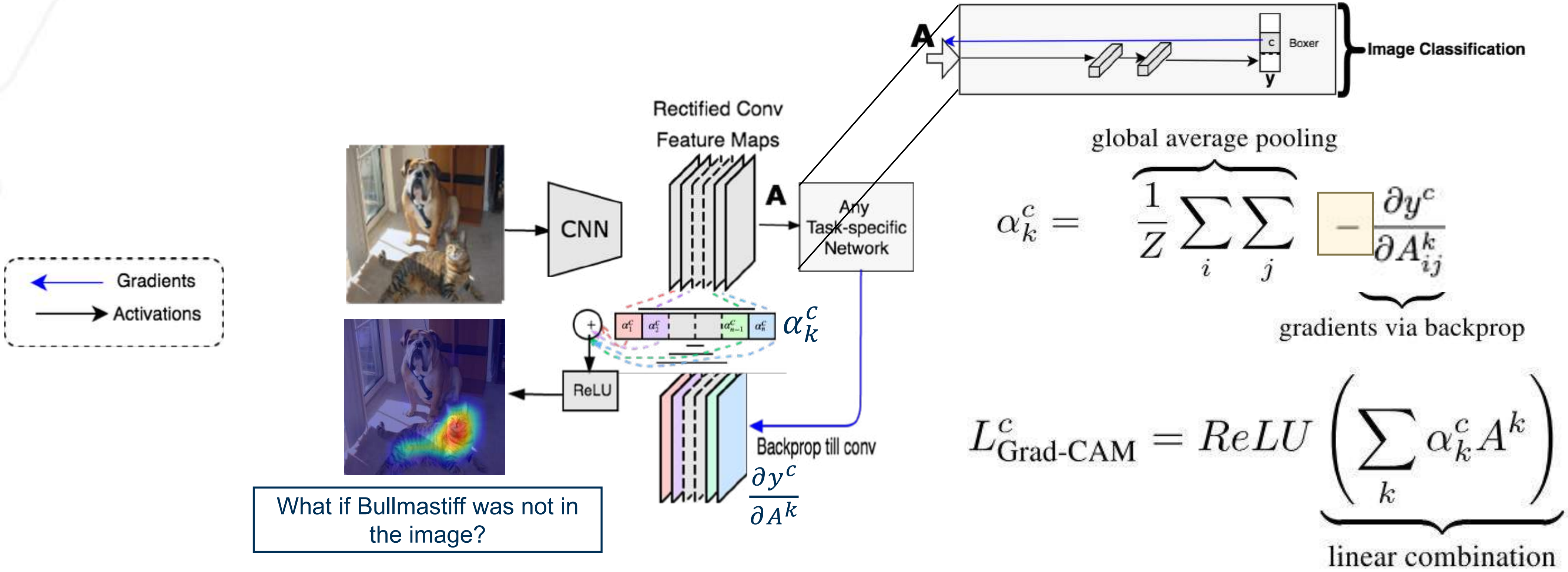
Gradient and Activation-based Explanations

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

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

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



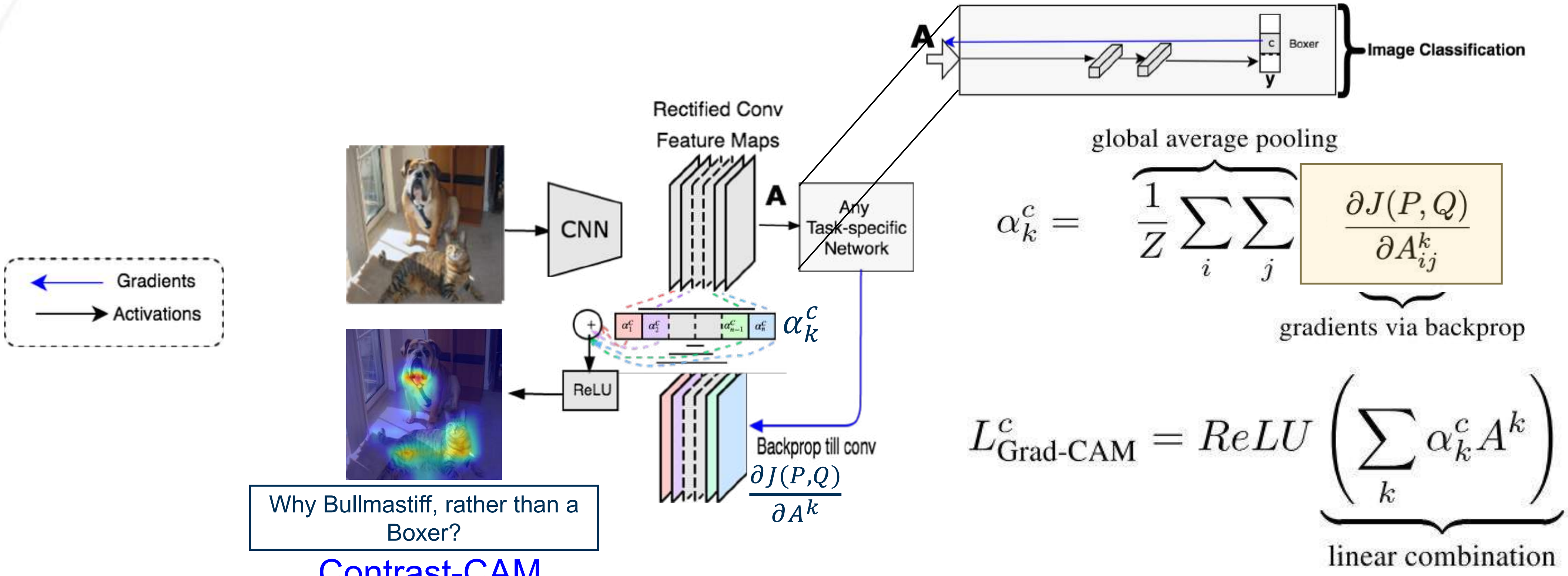
Negating the gradients effectively removes these regions from analysis

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

Gradient and Activation-based Explanations

ContrastCAM: Why P, rather than Q?

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



Contrast-CAM

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

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

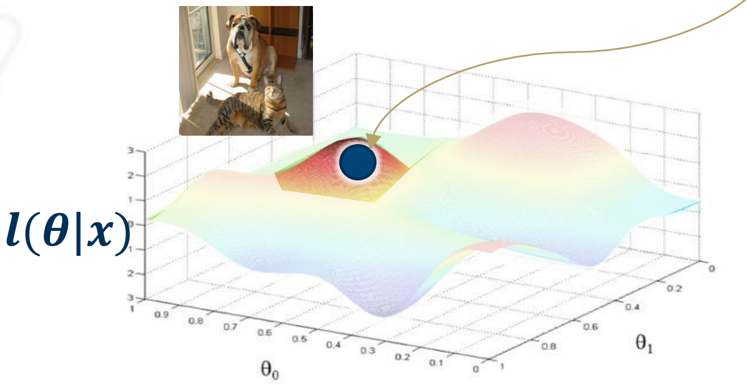
ContrastCAM

Toy Manifold Example

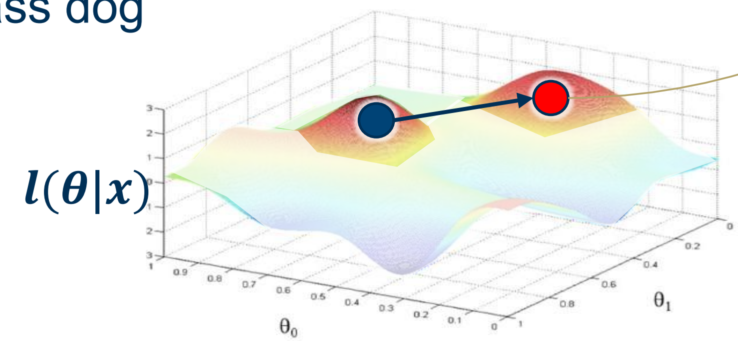
SCAN ME

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

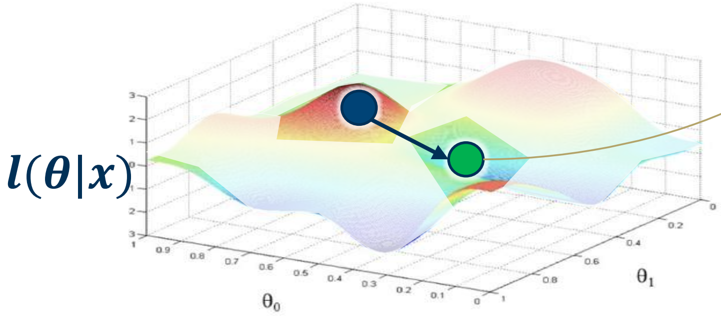
Likelihood of a dog predicted as class dog



Likelihood of a dog predicted as class cat

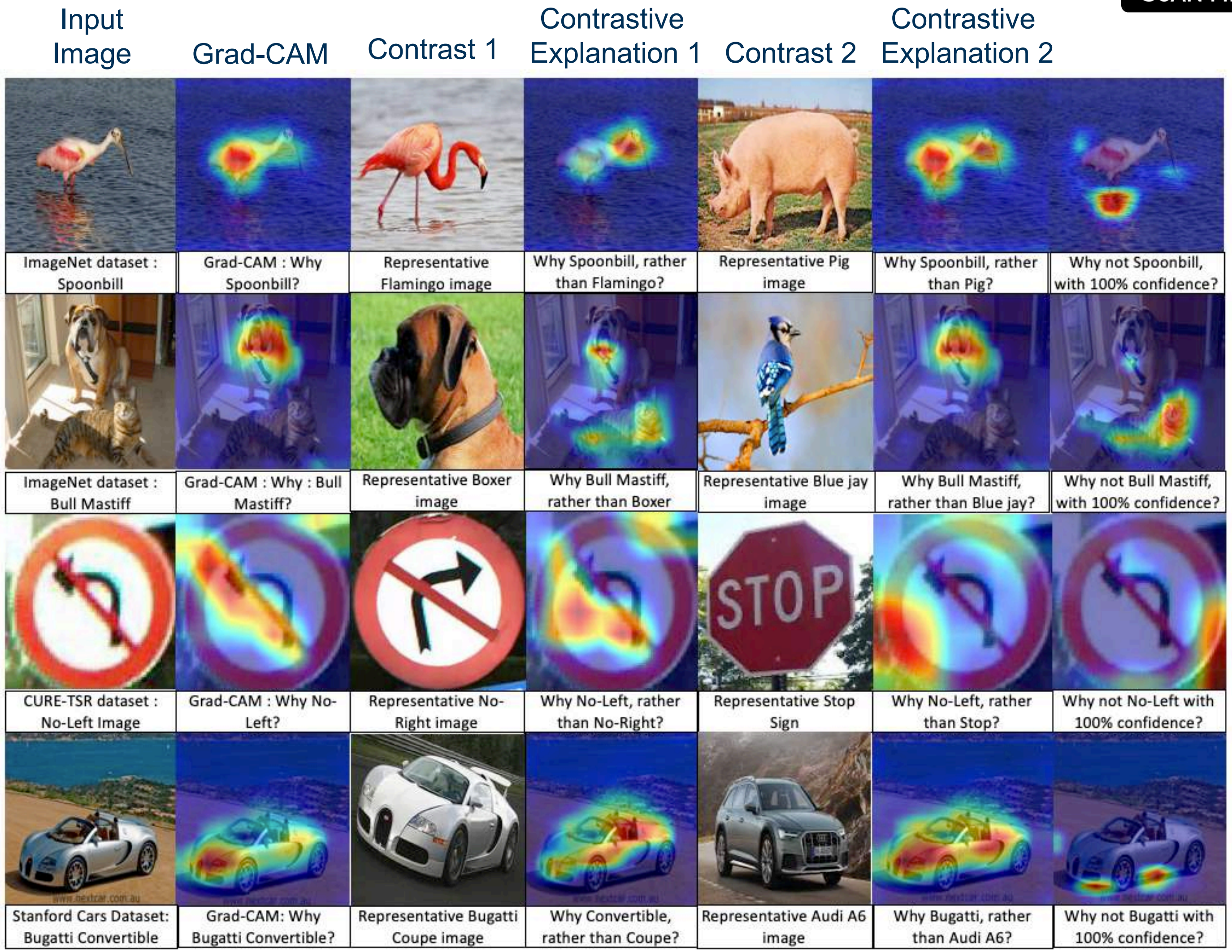


Likelihood of a dog predicted as class horse



Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM



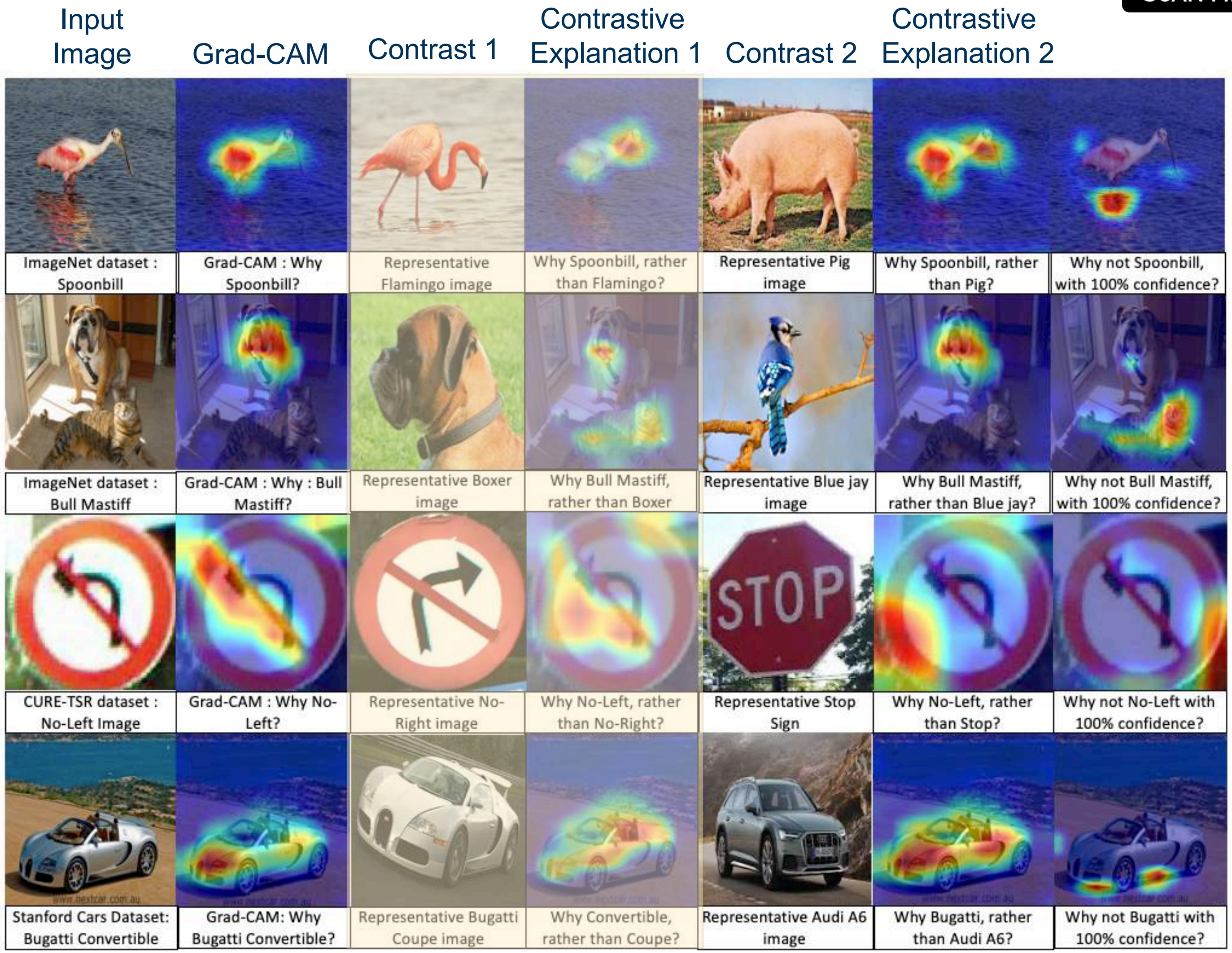
Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

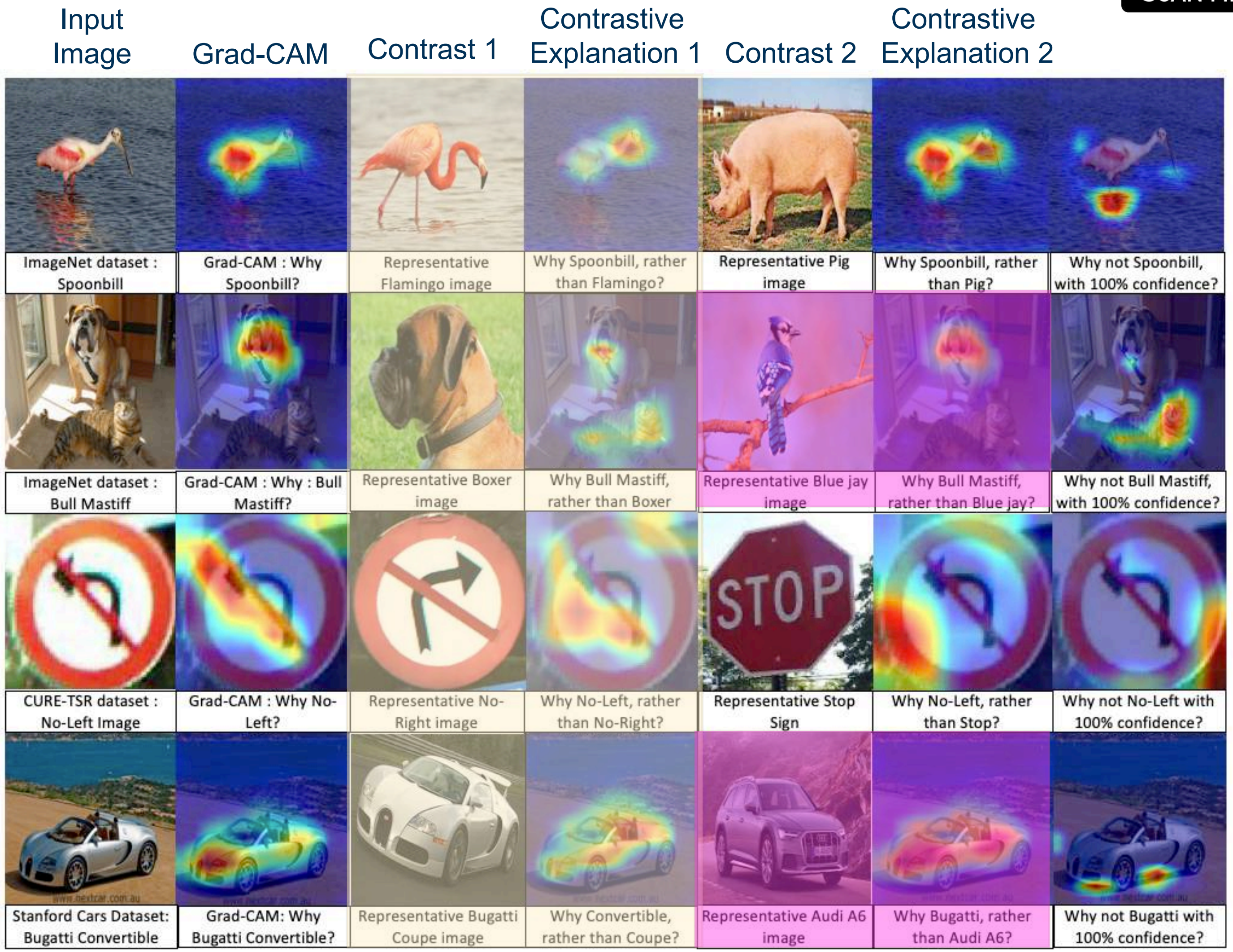


Human Interpretable

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



Human Interpretable

Same as Grad-CAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

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

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

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

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

SCAN ME

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

Human Interpretable

Same as Grad-

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

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

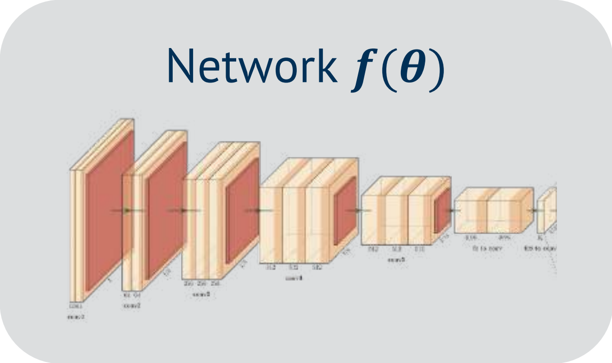
Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

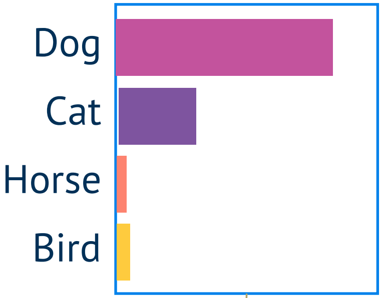
Applicability of Gradient Information

Gradients as Fisher Information

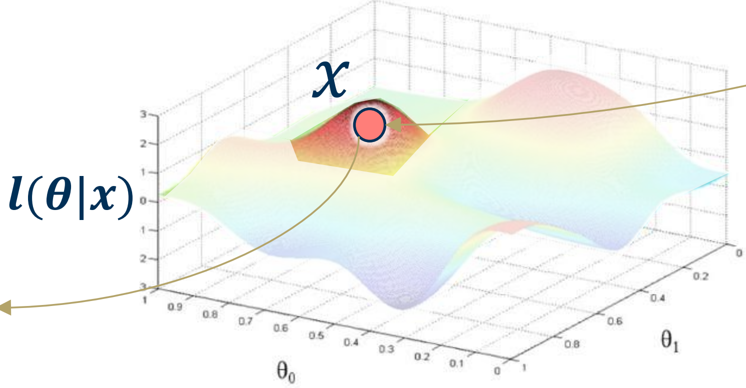
Gradients infer information about the statistics of underlying manifolds

x





Local information (specific to x) is sufficient!



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

30th ICIP 2023 Kuala Lumpur

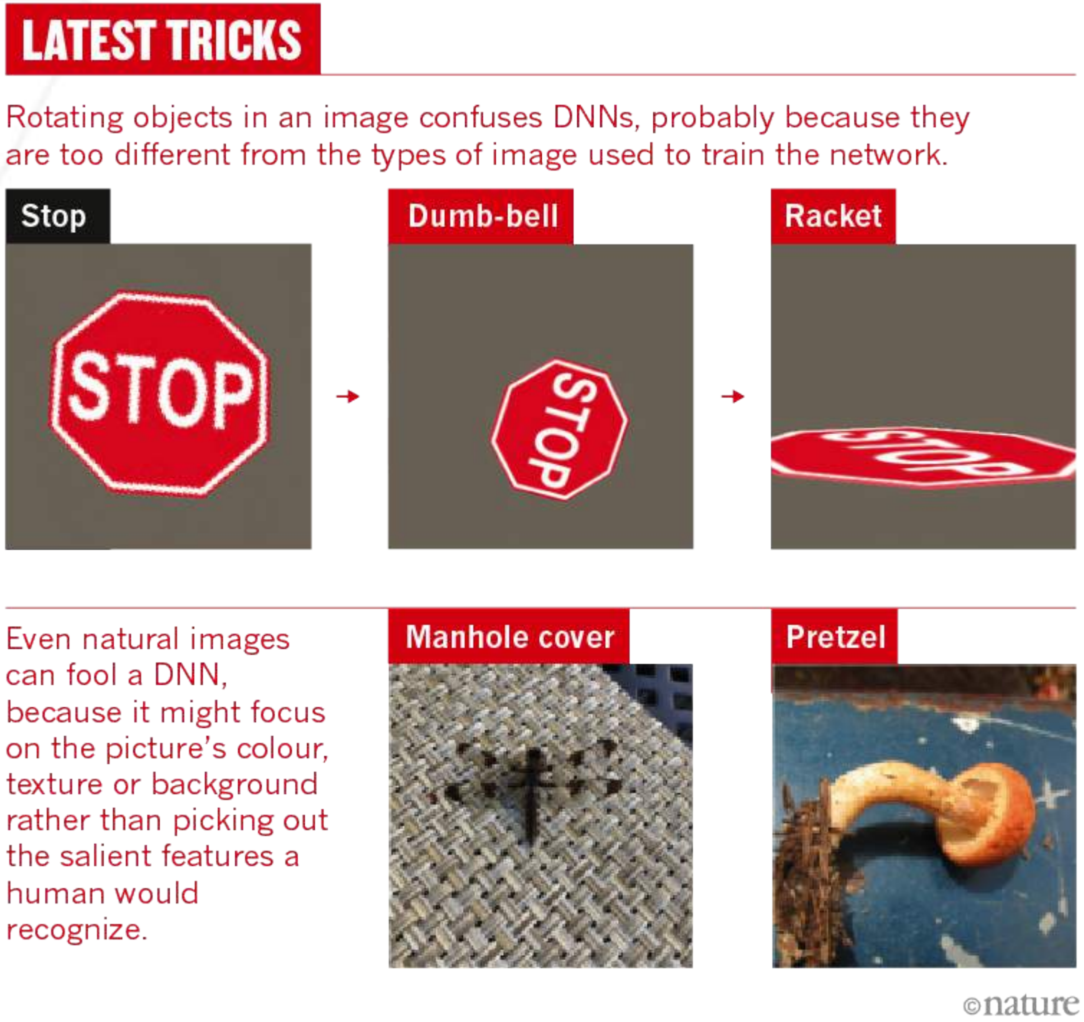
OLIVES @GeorgiaTech

Georgia Tech

Robustness in Neural Networks
Why Robustness?

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

SCAN ME



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

SCAN ME

How would humans resolve this challenge?

We Introspect!

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

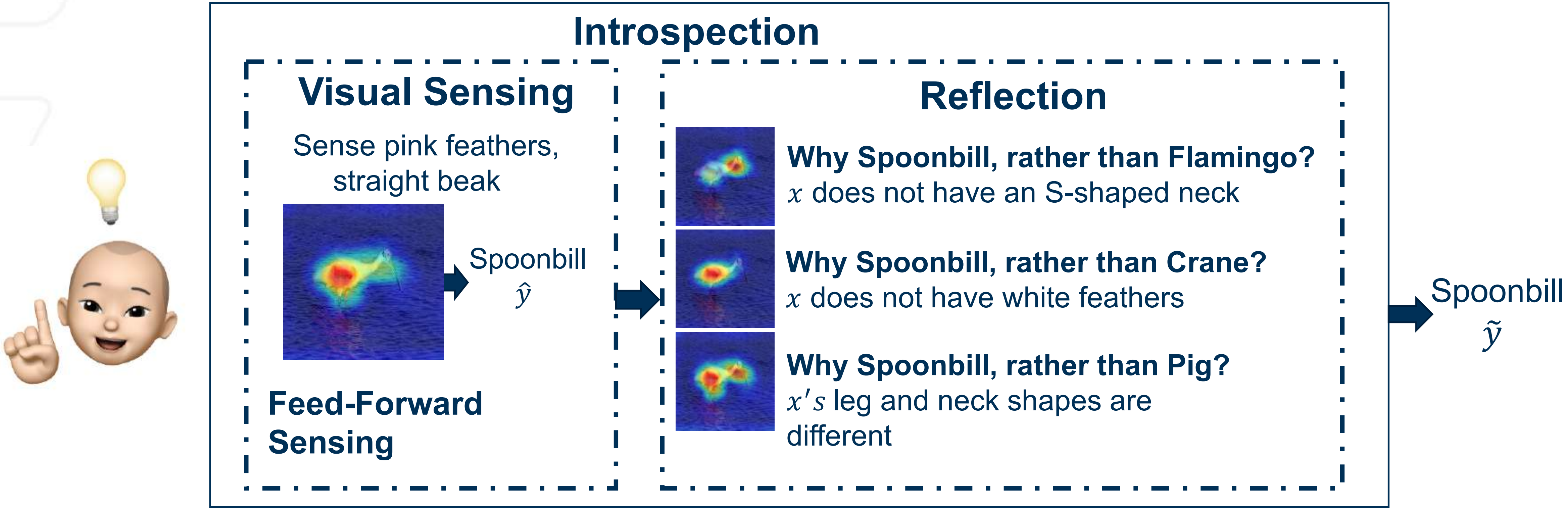
Introspection

What is Introspection?

SCAN ME

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

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



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

SCAN ME

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

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

Goal : To simulate Introspection in Neural Networks

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

What are the possible targeted questions?

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

Introspection

Introspection in Neural Networks

SCAN ME

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

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

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

What are the possible targeted questions?

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

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

SCAN ME

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

Goal : To simulate Introspection in Neural Networks

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

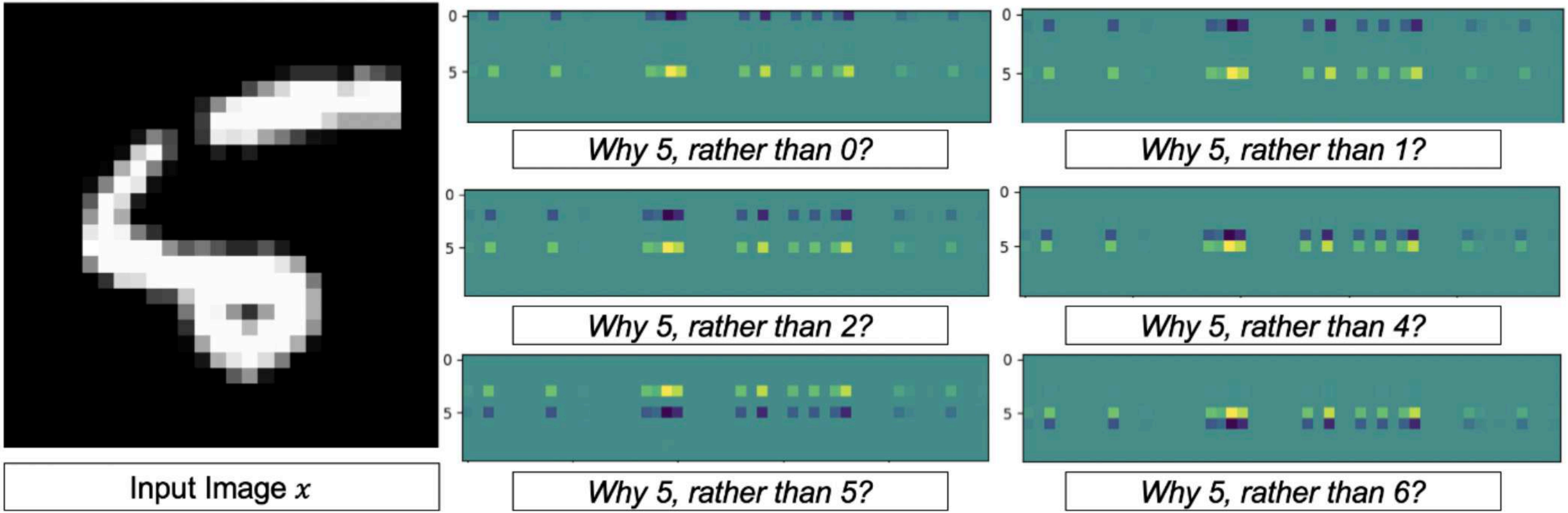
Technical Definition : *Given a network* $f(x)$*, a datum* x*, and the network's prediction* $f(x) = \hat{y}$*, introspection in* $f(\cdot)$ *is the measurement of change induced in the network parameters when a label* Q *is introduced as the label for* x*..*

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

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

SCAN ME

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



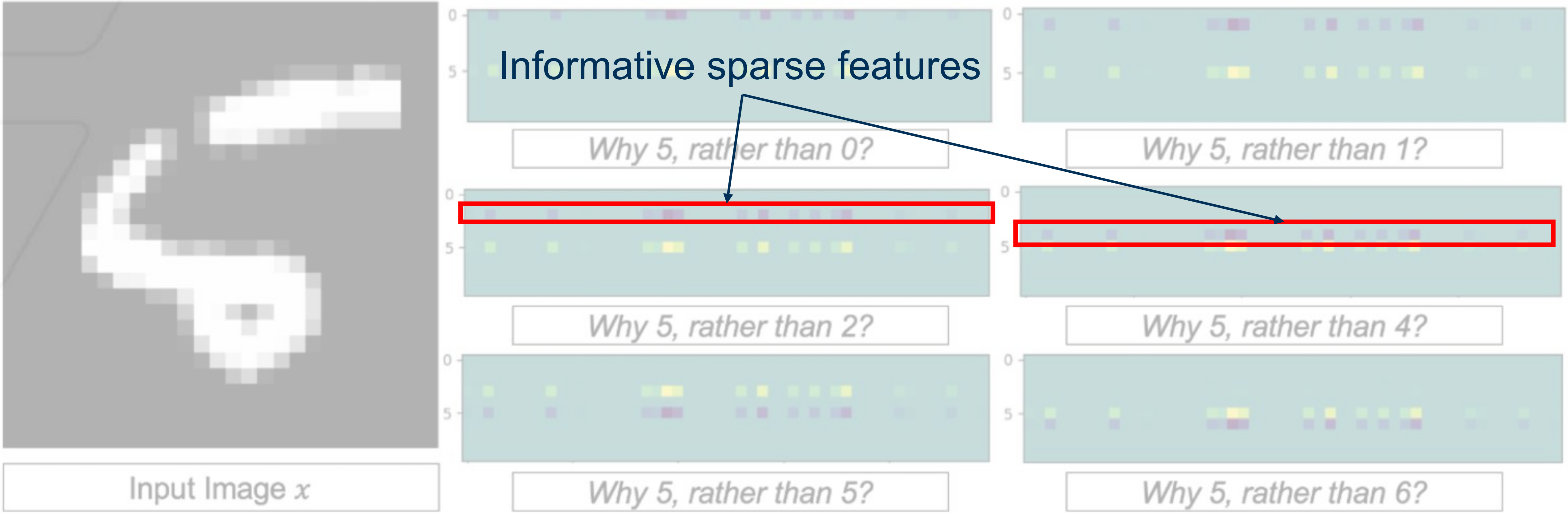
M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

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.

Introspection

Gradients as Features

SCAN ME

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

For a well-trained network, the gradients are robust

∇_W = Gradients w.r.t. weights

J = Loss function

$\hat{y}$ = Prediction

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

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.

30th ICIP 2023 Kuala Lumpur

OLIVES @GeorgiaTech

Georgia Tech

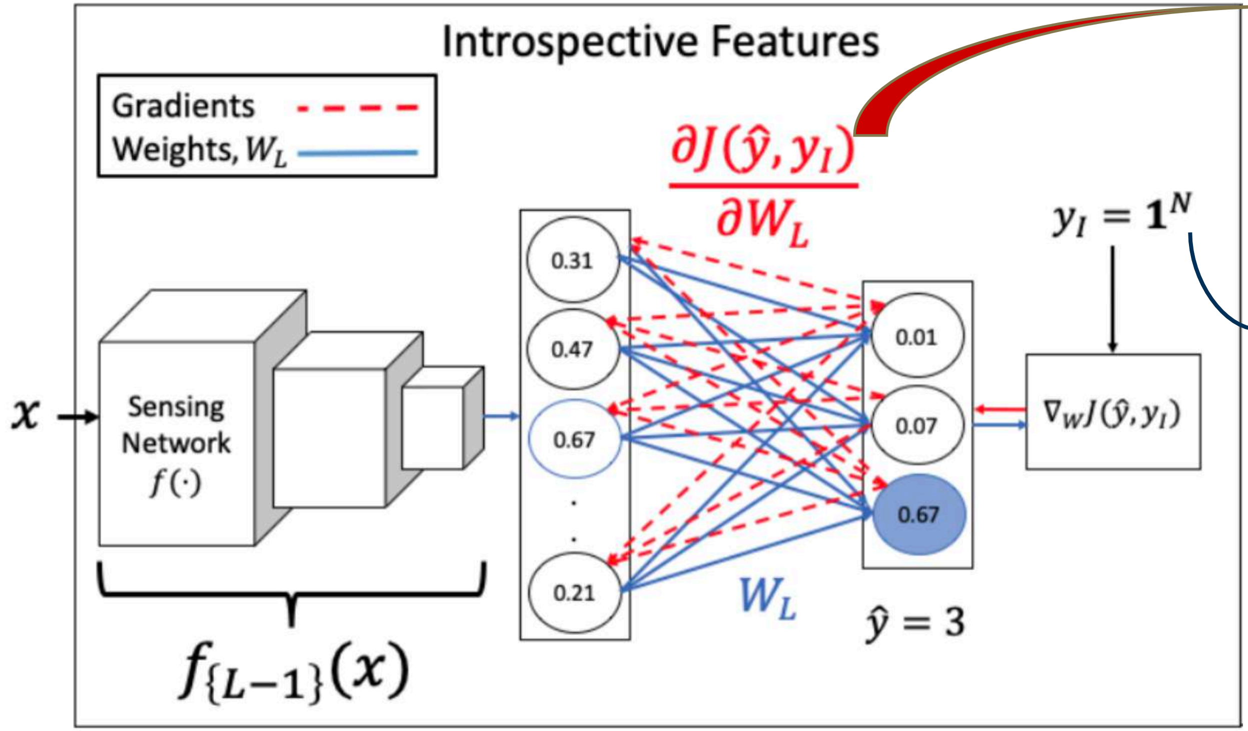
Introspection

Deriving Gradient Features

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

SCAN ME

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



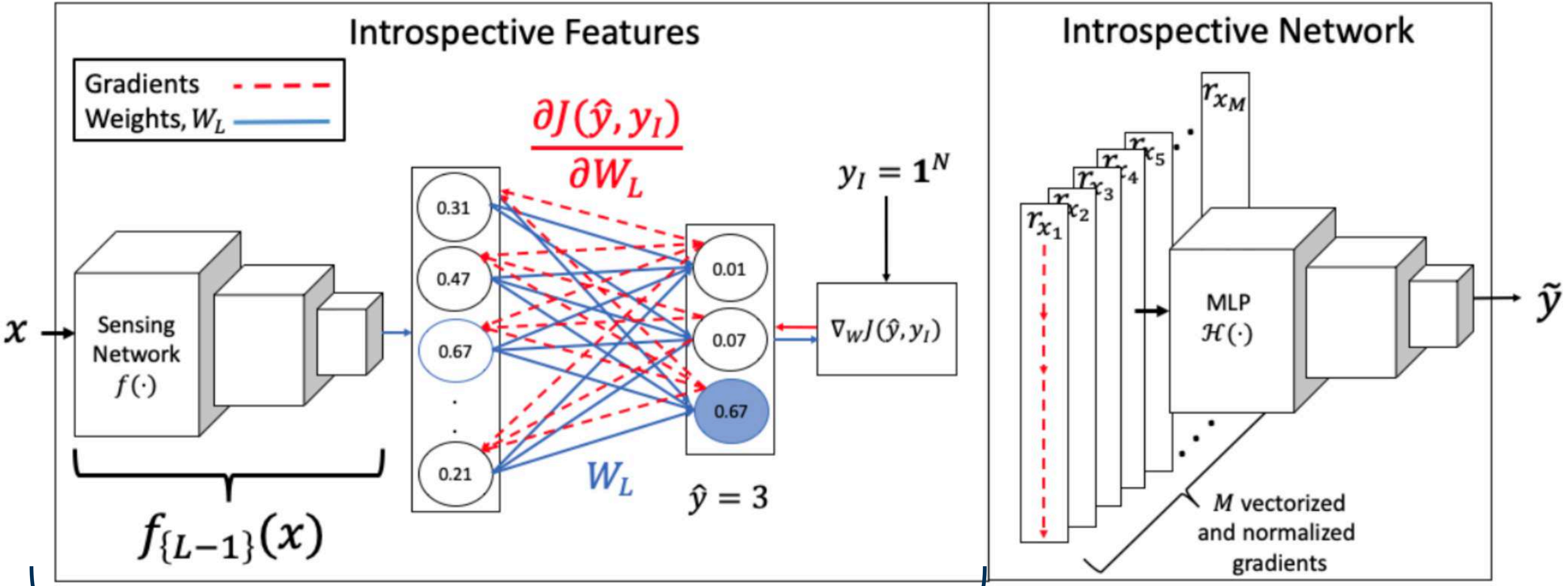
Normalized and vectorized gradients are introspective features

Vector of all ones: A confounding label!

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

Introspection
Utilizing Gradient Features

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



Introspective Features

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

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

SCAN ME

Introspection provides robustness when the train and test distributions are different

We define robustness as being generalizable and calibrated to new testing data

Generalizable: Increased accuracy on OOD data

Calibrated: Reduces the difference between prediction accuracy and confidence

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

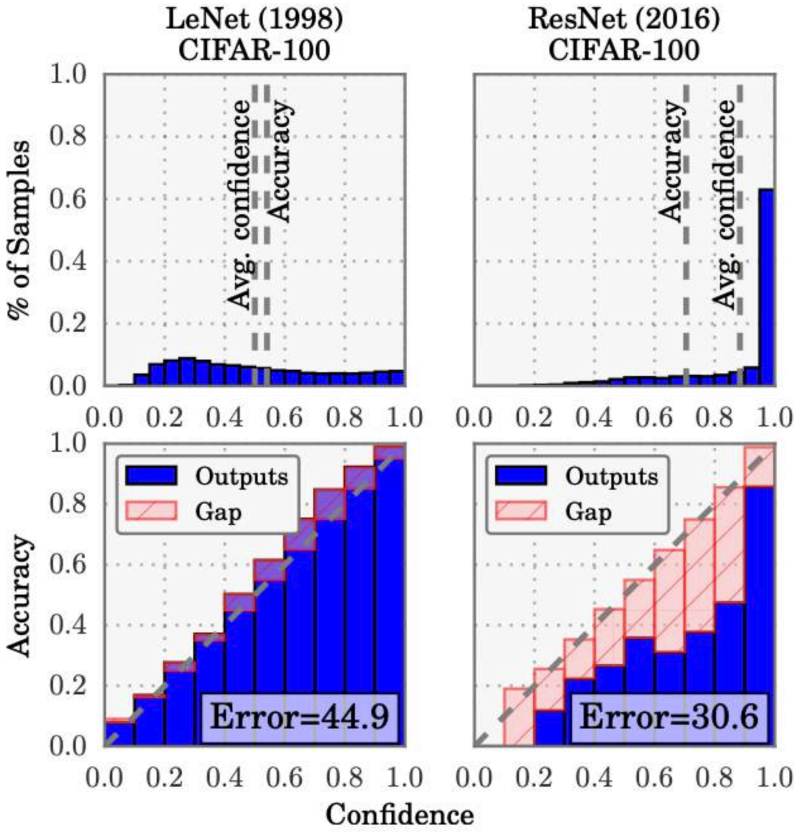
Calibration

A note on Calibration..

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

SCAN ME

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



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

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

Calibration

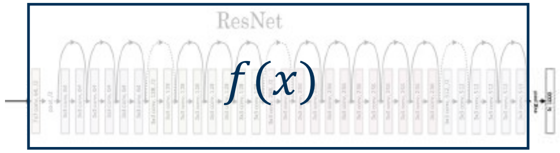
A note on Calibration..

SCAN ME

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

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

CIFAR-10 Testset



Average Accuracy

$f(X)$

$f(X) - \mathbb{P}(X)$

$\mathbb{P}(X)$

Average Softmax Probability

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

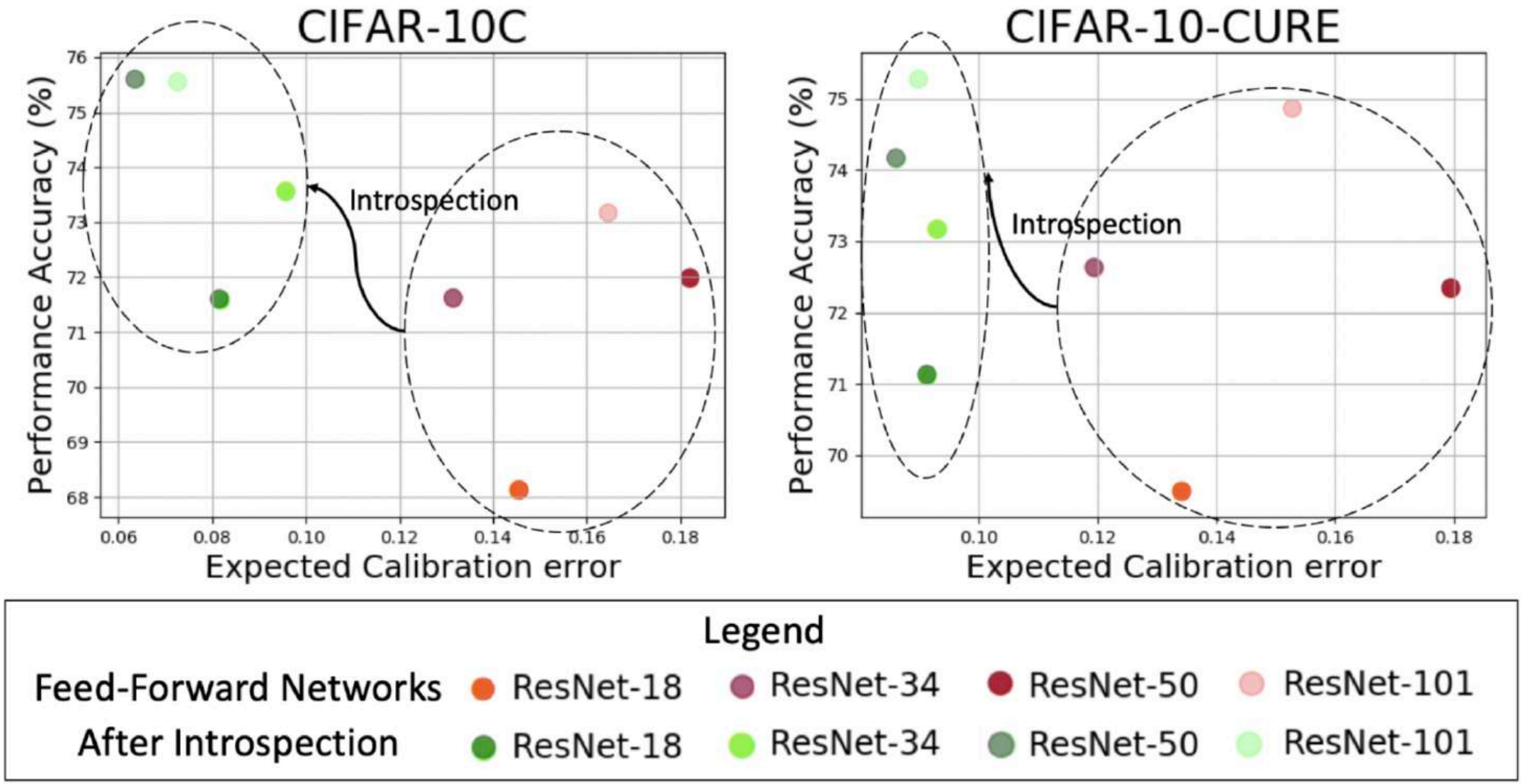
Introspective Learning: A Two-stage Approach for Inference in Neural Networks

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.

Plug-in nature of Introspection

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

SCAN ME

Introspection is a light-weight option to resolve robustness issues

Table 1: Introspecting on top of existing robustness techniques.

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

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

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

SCAN ME

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

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

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

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

Table 2: Recognition accuracy of Active Learning strategies.

Methods	Architecture	Original Testset		Gaussian Noise	
		R-18	R-34	R-18	R-34
Entropy (31)	Feed-Forward	0.365	0.358	0.244	0.249
	Introspective	0.365	0.359	**0.258**	**0.255**
Least (31)	Feed-Forward	0.371	0.359	0.252	0.25
	Introspective	0.373	0.362	**0.264**	**0.26**
Margin (32)	Feed-Forward	0.38	0.369	0.251	0.253
	Introspective	0.381	0.373	**0.265**	**0.263**
BALD (34)	Feed-Forward	0.393	0.368	0.26	0.253
	Introspective	0.396	0.375	**0.273**	**0.263**
BADGE (33)	Feed-Forward	0.388	0.37	0.25	0.247
	Introspective	0.39	0.37	**0.265**	0.**260**

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

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

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

Objectives

Takeaways from Part II

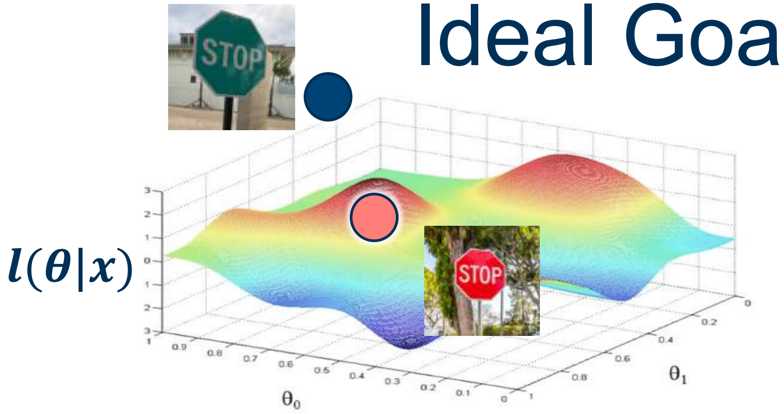
- Part I: Gradients in Neural Networks
- **Part 2: Gradients as Information**
  - Gradients approximate Fisher Information: They provide a methodology to infer information about the statistics of underlying manifolds using samples
  - Fisher information in gradients allow them to be utilized in explanations
  - 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

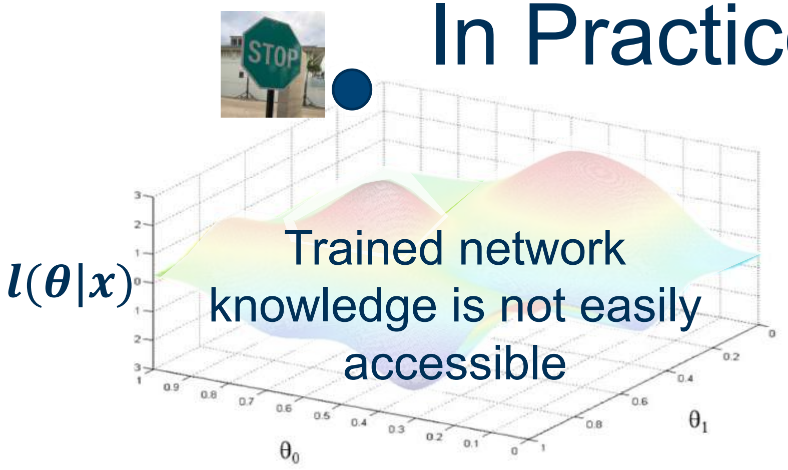
Tying it Back

From Part I

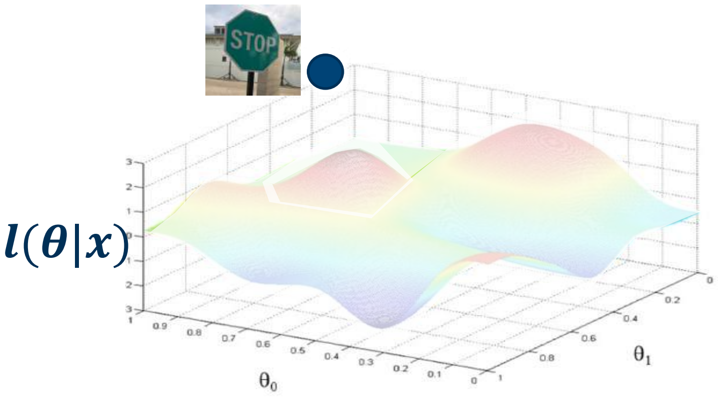
Ideal Goal



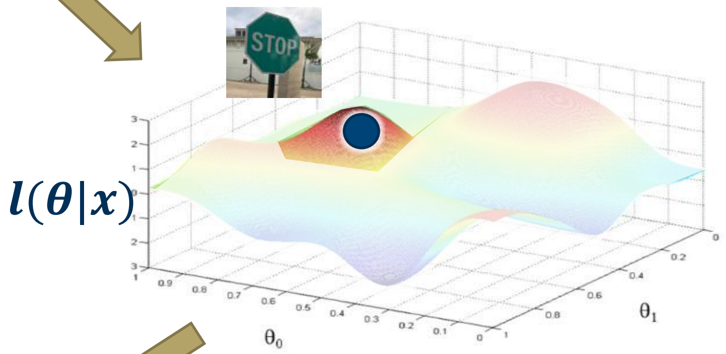
In Practice

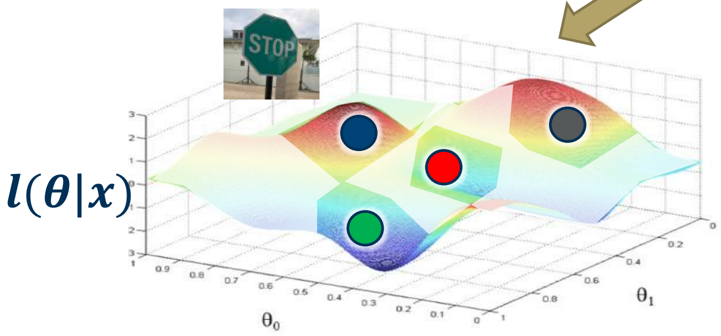


In Part II



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