Inferential Machine Learning: Towards Human-collaborative Foundation Models

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

Feb. 28, 2025 – Tucson, AZ, USA

WACV 2025
TUCSON, ARIZONA • FEB 28 - MAR 4

OLIVES @GeorgiaTech

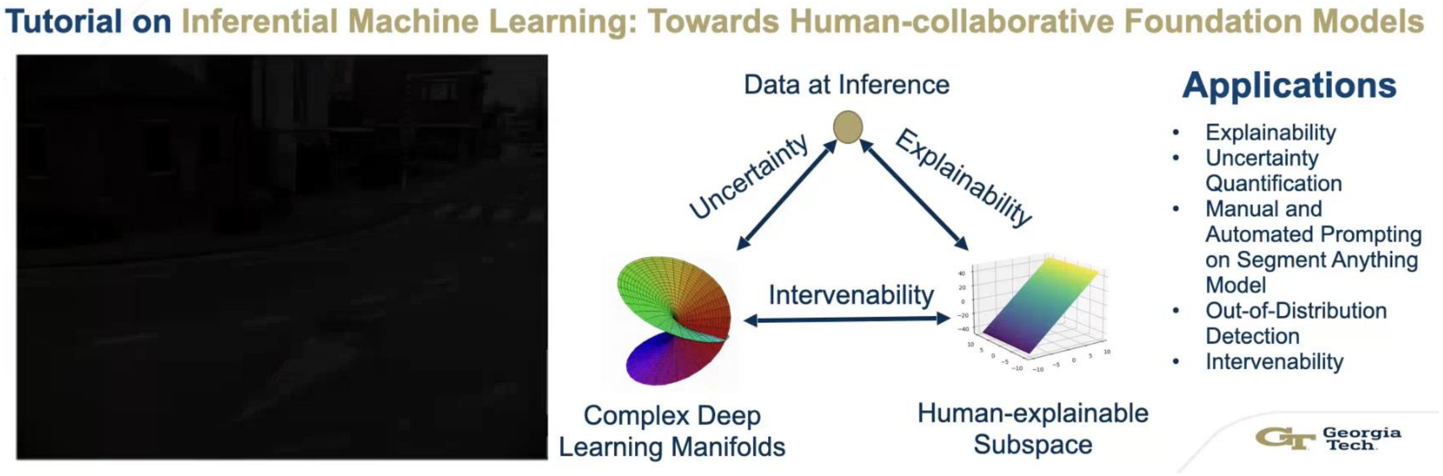
Georgia Tech

Tutorial Materials

Accessible Online

SCAN ME

https://alregib.ece.gatech.edu/courses-and-tutorials/wacv-2025-tutorial/
{alregib, mohit.p}@gatech.edu



Inferential Machine Learning: Towards Human-collaborative Foundation Models

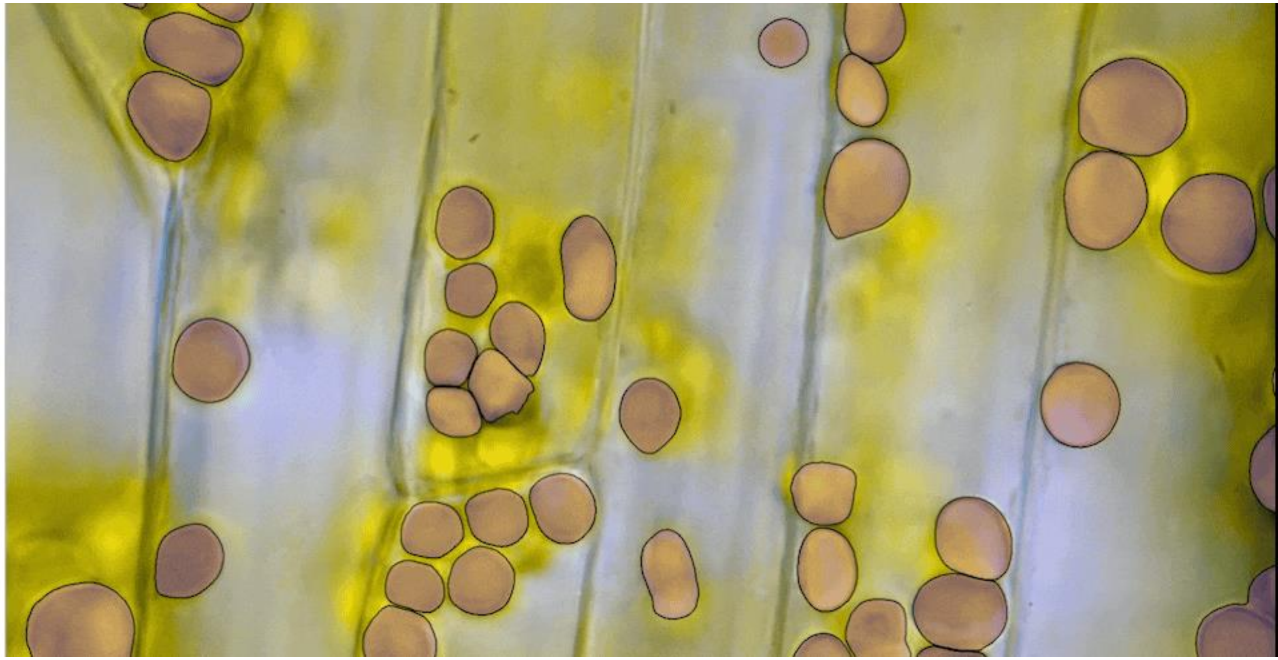
Foundation Models

Expectation vs Reality

Expectation vs Reality of Foundation Models

Foundation Models

Segment Anything Model



Segment Anything Model (SAM) released by Meta on April 5, 2023 was trained on Segment Anything 1 Billion dataset with 1.1 billion high-quality segmentation masks from 11 million images

Kirillov, Alexander, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao et al. "Segment anything." *arXiv preprint arXiv:2304.02643* (2023).

Foundation Models

Segment Anything Model

Cityscapes dataset semantic segmentation annotation took ~90 mins per image

Kirillov, Alexander, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao et al. "Segment anything." *arXiv preprint arXiv:2304.02643* (2023).

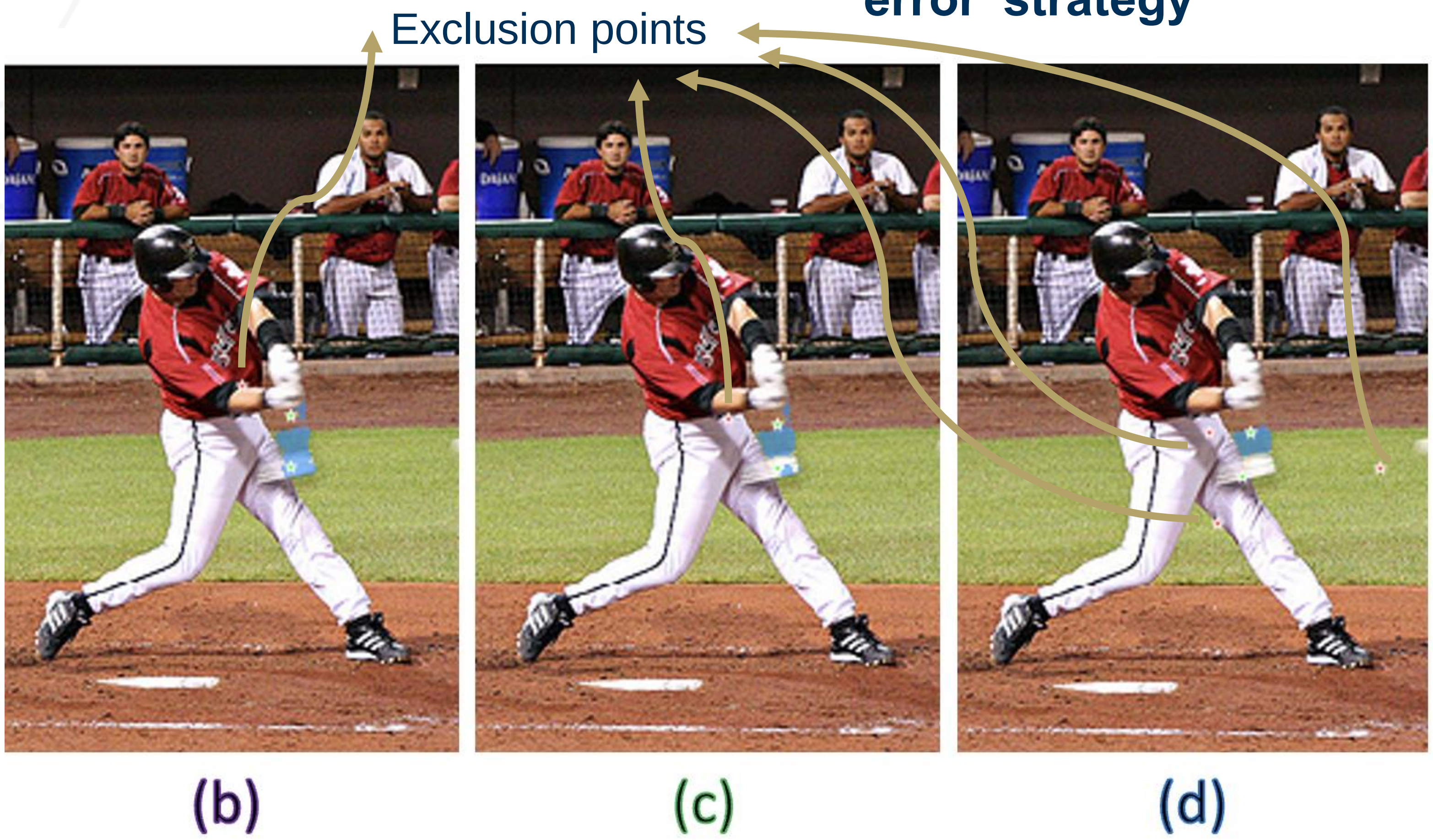
Foundation Models

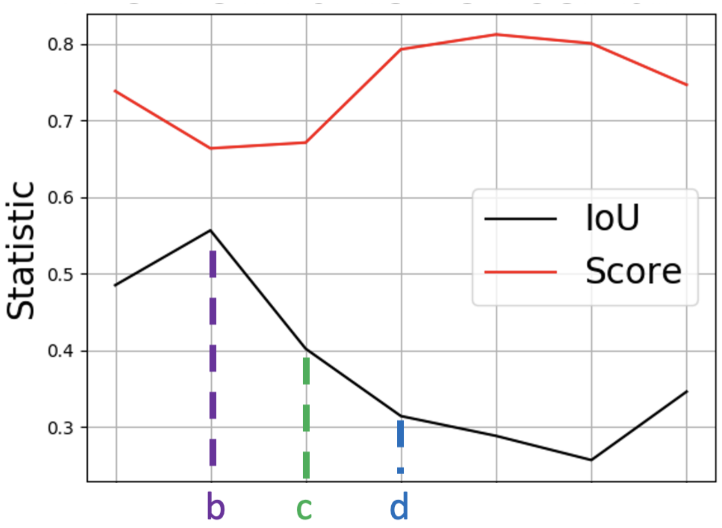
'Trial and Error' Interventions in Segment Anything Model

PointPrompt Dataset

SCAN ME

Goal: Given a promptable model with no operational knowledge, users employ a 'trial and error' strategy





The general conclusion from [1] is that annotators overprompt and utilize strategies that lead to worse performance

~200,000 prompts on 6000 images

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

Foundation Models

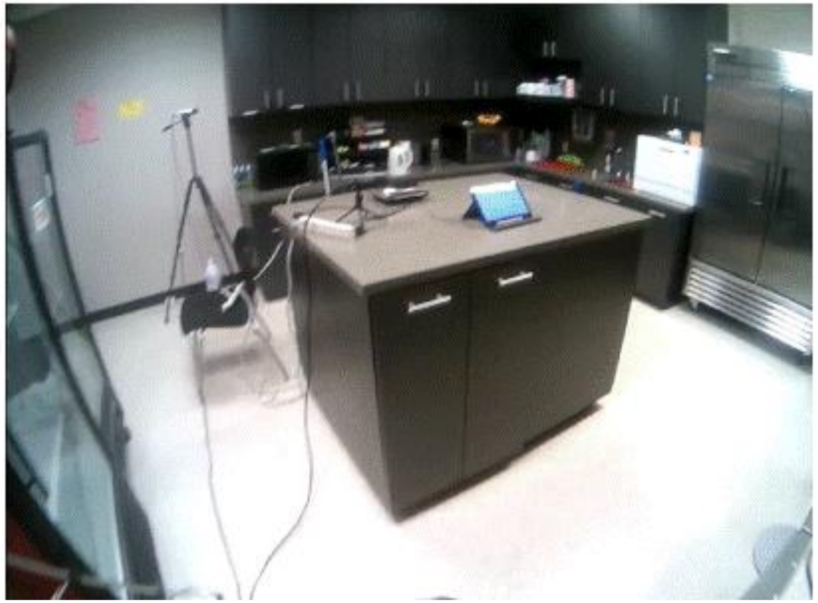
Vision-Language Models are ‘Doomed to Choose’

SCAN ME

DARai

Dataset

Goal: Given a long video sequence, vision language models (VLMs) can process, interpret, and answer questions



USER:

What is the person doing?

ASSISTANT:

VLMs (and all other deep learning-based systems) are ‘doomed to choose’ – no mechanism to understand if sufficient information is available at inference

Demo created at Inference on “LLaVA-v1.5-13B” model on Daily Activity Recognition (DARai) dataset [1]

[1] Ghazal Kaviani, Yavuz Yarici, Mohit Prabhushankar, Ghassan AlRegib, Mashhour Solh, Ameya Patil, June 12, 2024, "DARai: Daily Activity Recordings for AI and ML applications", IEEE Dataport, doi: https://dx.doi.org/10.21227/ecnr-hy49.

Foundation Models

Vision-Language Models are Sensitive to Granularity of Tasks

DARai Dataset

SCAN ME

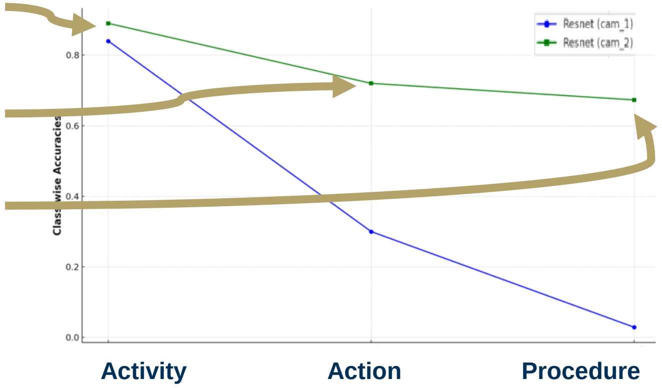
VLMs (encoder finetuned on dataset) fail when recognizing fine-grained hierarchical activities

Hierarchical Activity Recognition

Ground Truth **Prediction**

Other findings:

SCAN ME



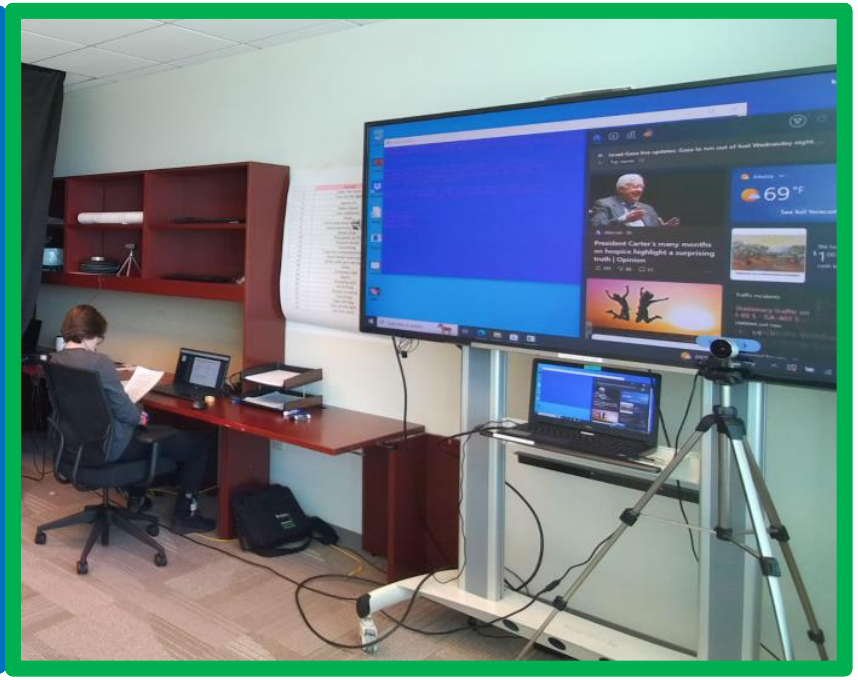
Foundation Models

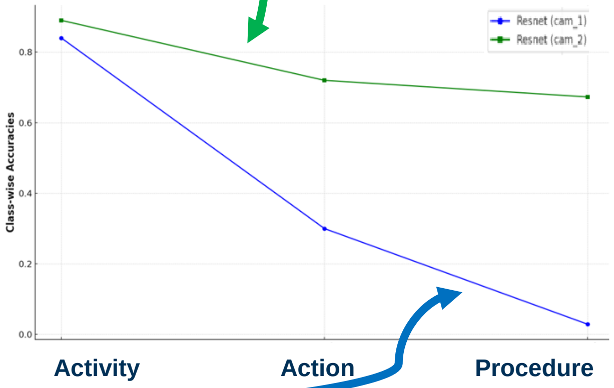
Vision-Language Models are sensitive to experimental setup

DARai Dataset

VLMs (encoder finetuned on dataset) fail when recognizing domain-shifted inputs

Other findings:





WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

OLIVES @GeorgiaTech

Georgia Tech

Foundation Models

Vision-Language Models are Biased towards Societal Stereotypes

Debiasing VLMs

SCAN ME

CLIP-CAP

A **woman** in a wetsuit surfing on a wave.

CLIP-CAP

A **man** riding skis down a snow covered slope.

Uncurated training data invariably reflects biases present in society. Utilizing such models in downstream tasks perpetuates biases

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

Jung, Hoin, Taeuk Jang, and Xiaoqian Wang. “A Unified Debiasing Approach for Vision-Language Model across Modalities and Tasks”. In NeurIPS. 2024.

OLIVES @GeorgiaTech

Georgia Tech

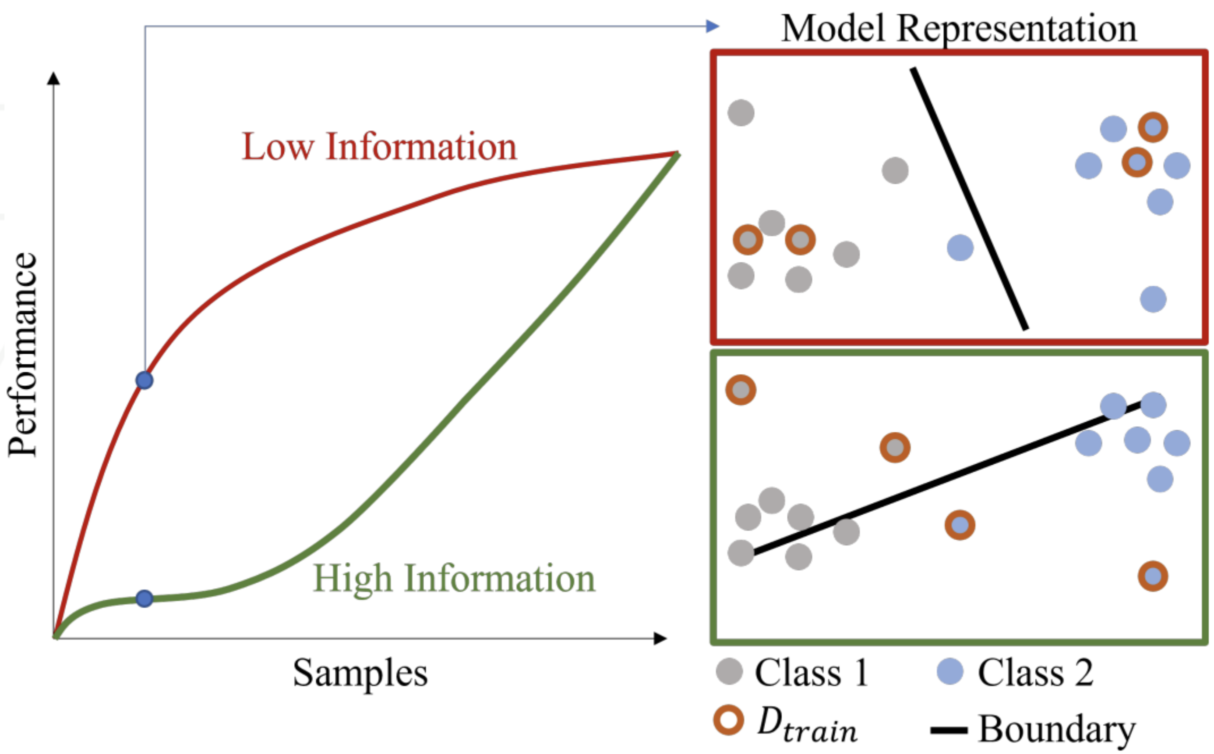
Requirements: Foundation model-enabled systems must predict correctly and fairly on novel data and explain their outputs

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).

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

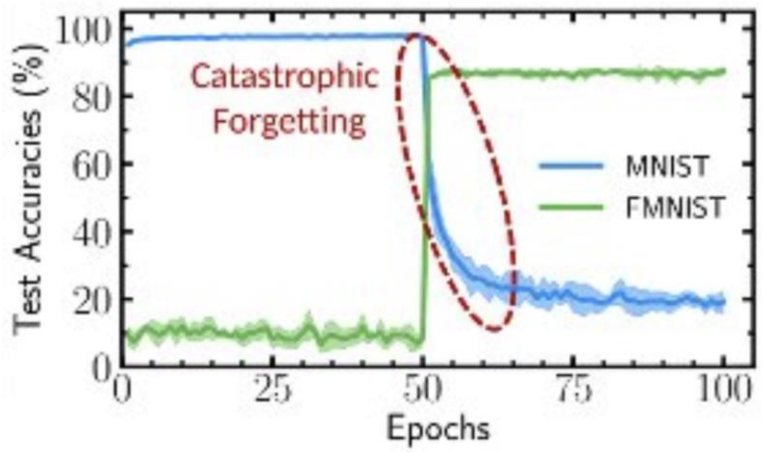


Novel samples = Most Informative

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

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

Subsequent training must not focus only on novel data

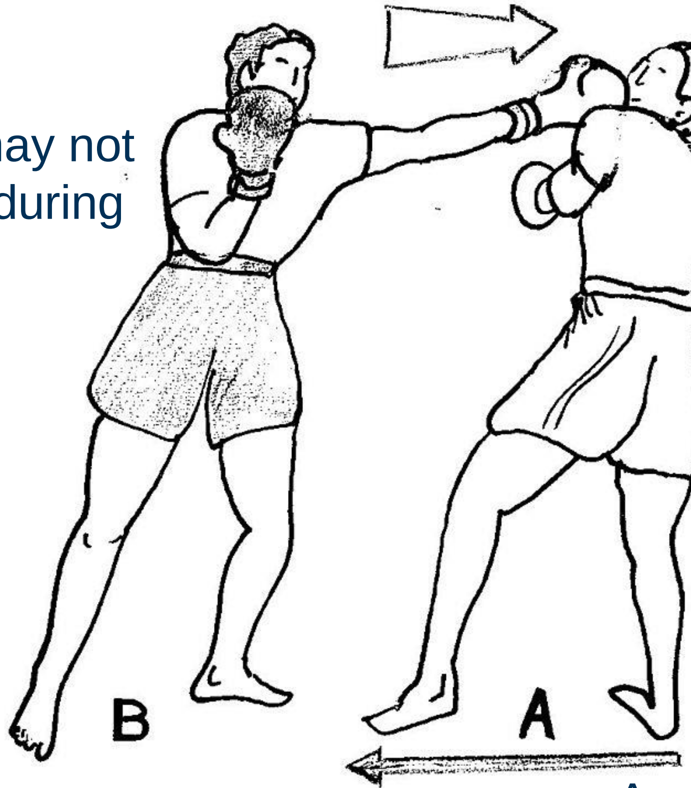


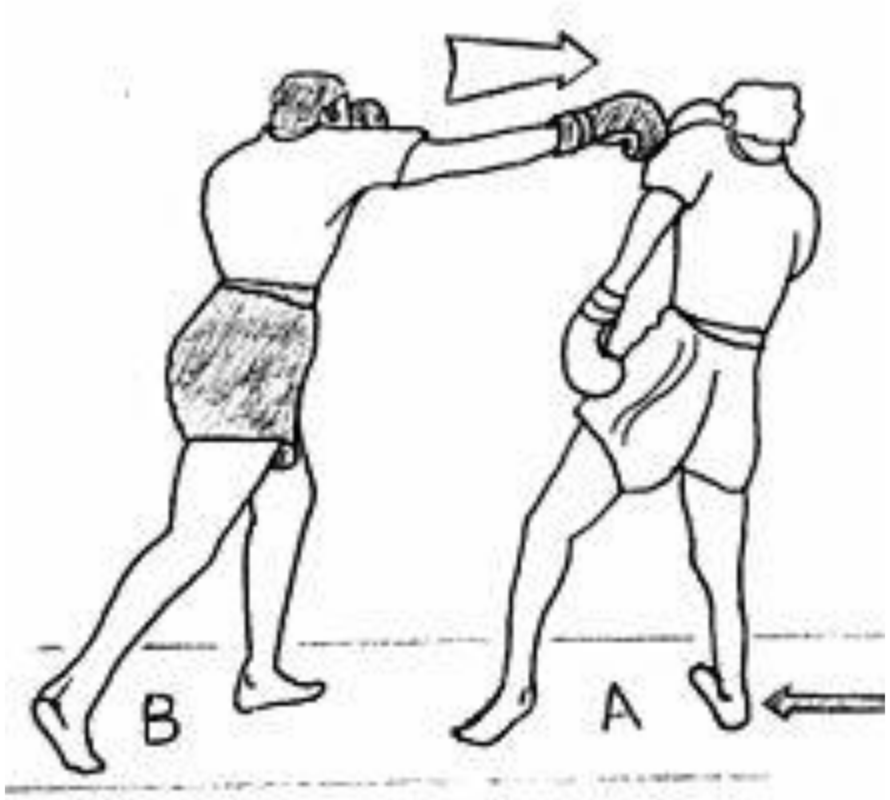
- The model performs well on the new scenarios, **while forgetting the old scenarios**
- Several techniques exist to overcome this trend
- However, they affect the overall performance in large-scale settings
- It is not always clear **if and when** to incorporate novel scenarios in training

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

Novel data packs a 1-2 punch!

Novel data may not be available during training





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

A = Deep Neural Networks
B = Novel data

We must handle novel data at Inference!!

Novel data sources:

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

Model Train

At Inference

Objective

Objective of the Tutorial

To discuss methodologies that promote robust and fair inference in neural networks

- Part 1: Inference in Neural Networks
- Part 2: Explainability at Inference
- Part 3: Uncertainty at Inference
- Part 4: Intervenability at Inference
- Part 5: Conclusions and Future Directions

Inferential Machine Learning

Part I: Inference in Neural Networks

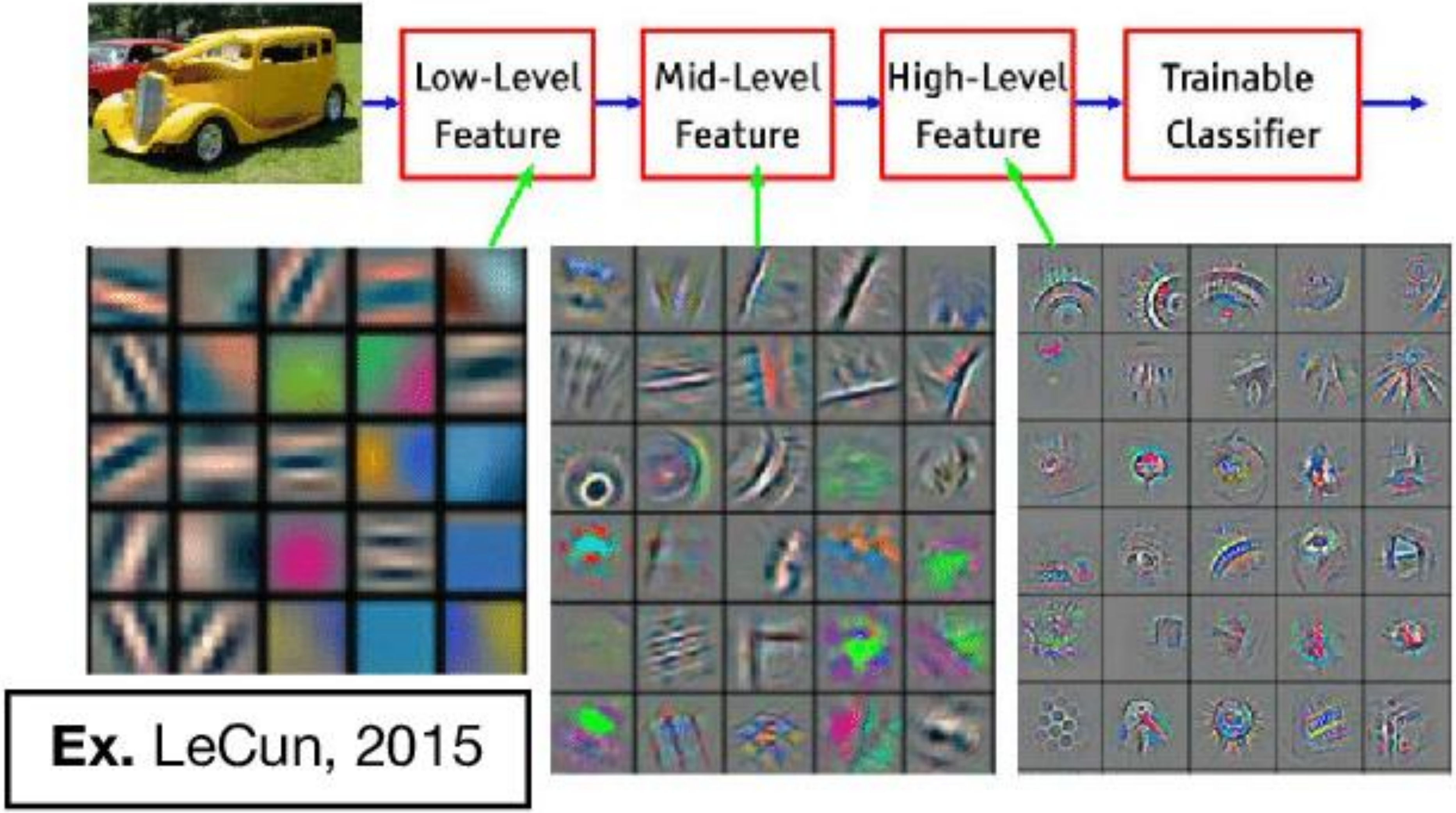
Objective

Objective of the Tutorial

To discuss methodologies that promote robust and fair inference in neural networks

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

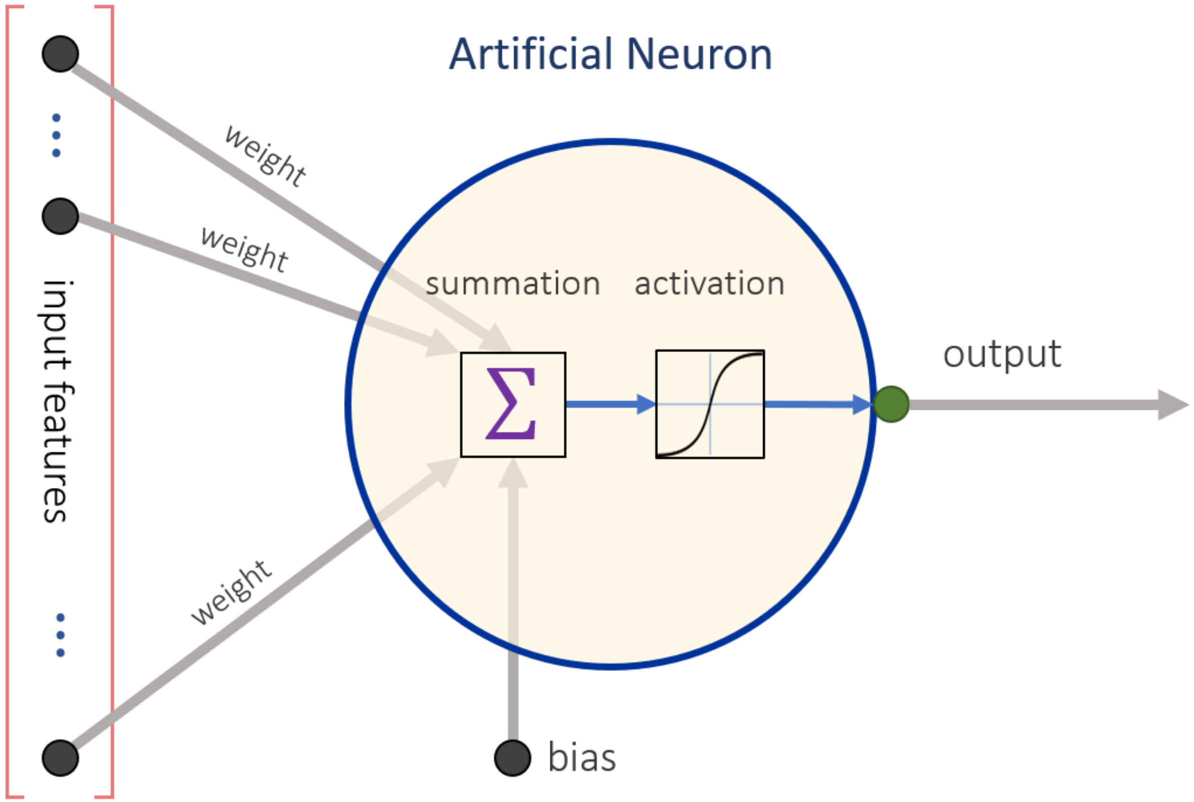
Deep Learning
Overview



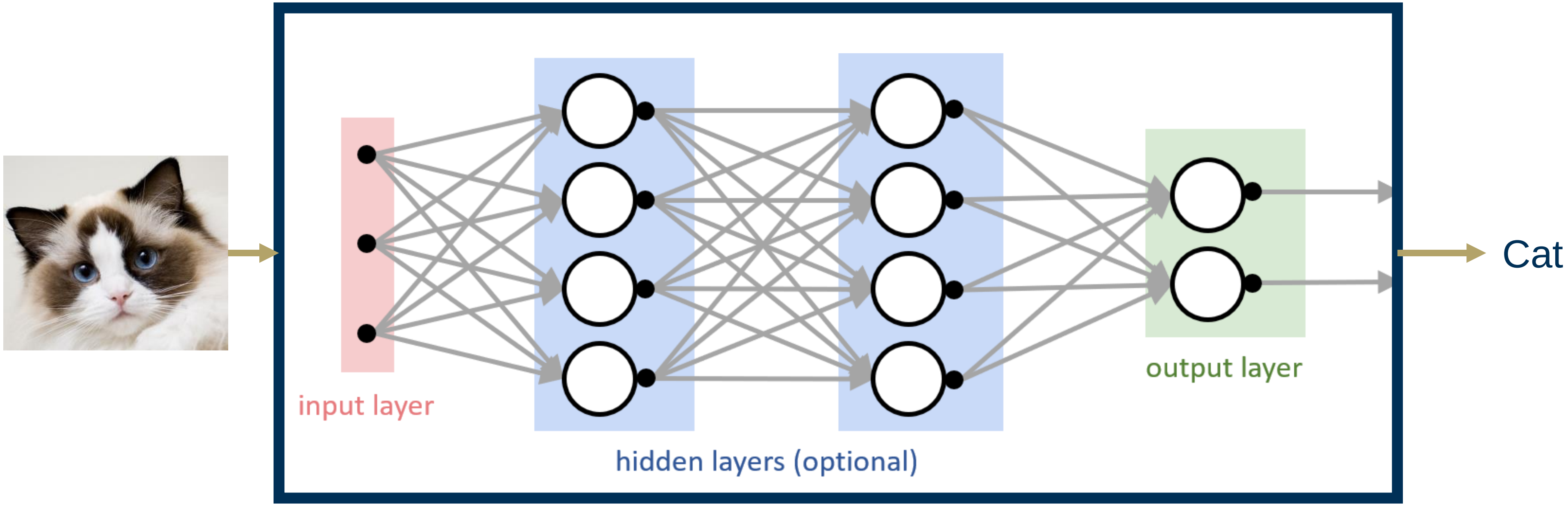
The underlying computation unit is the Neuron

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



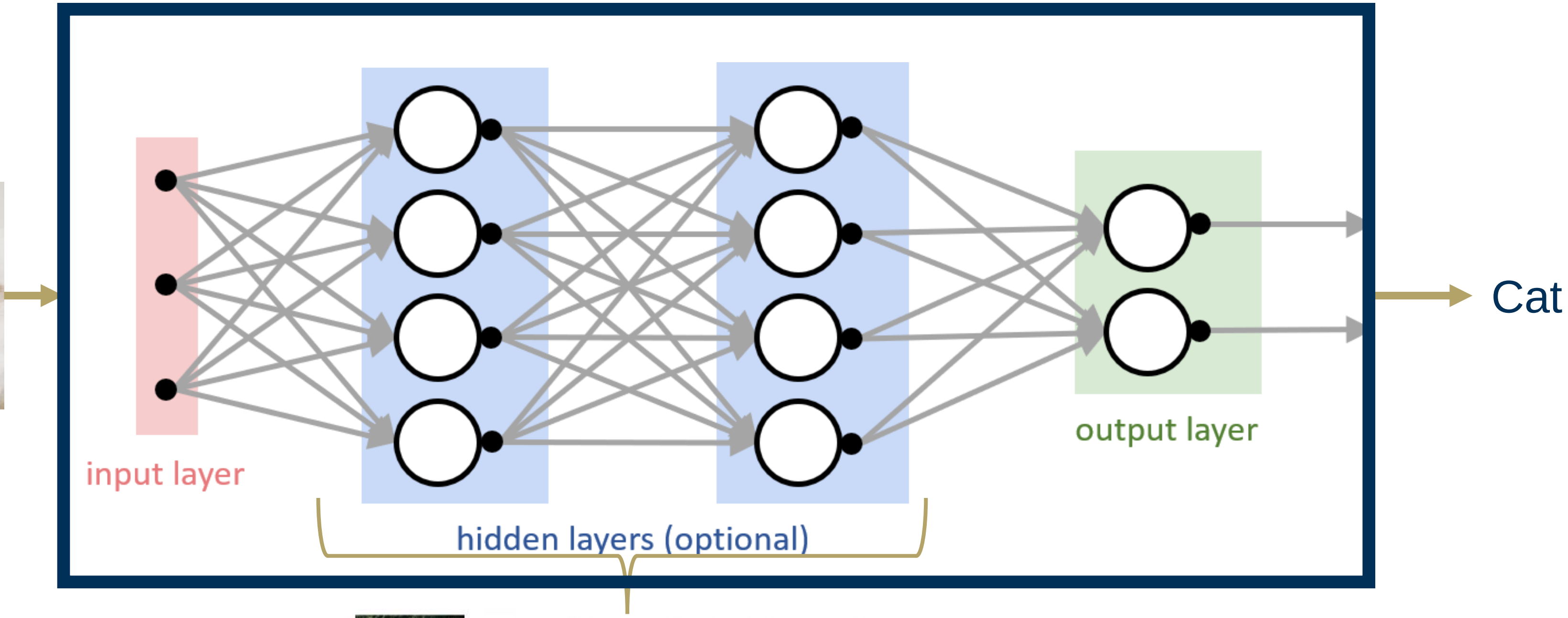
Typically, a neuron is part of a network organized in **layers**:

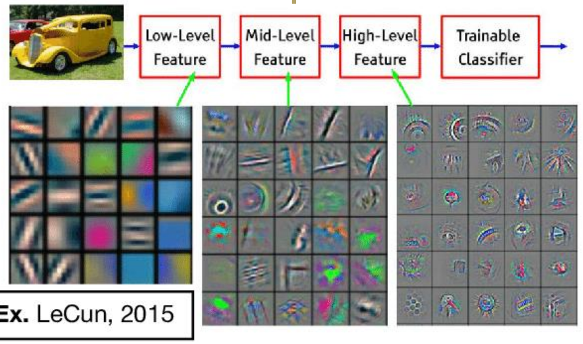
- An **input** layer (Layer 0)
- An **output** layer (Layer K)
- Zero or more **hidden** (middle) layers (Layers $1 \dots K-1$)

Deep Learning

Convolutional Neural Networks

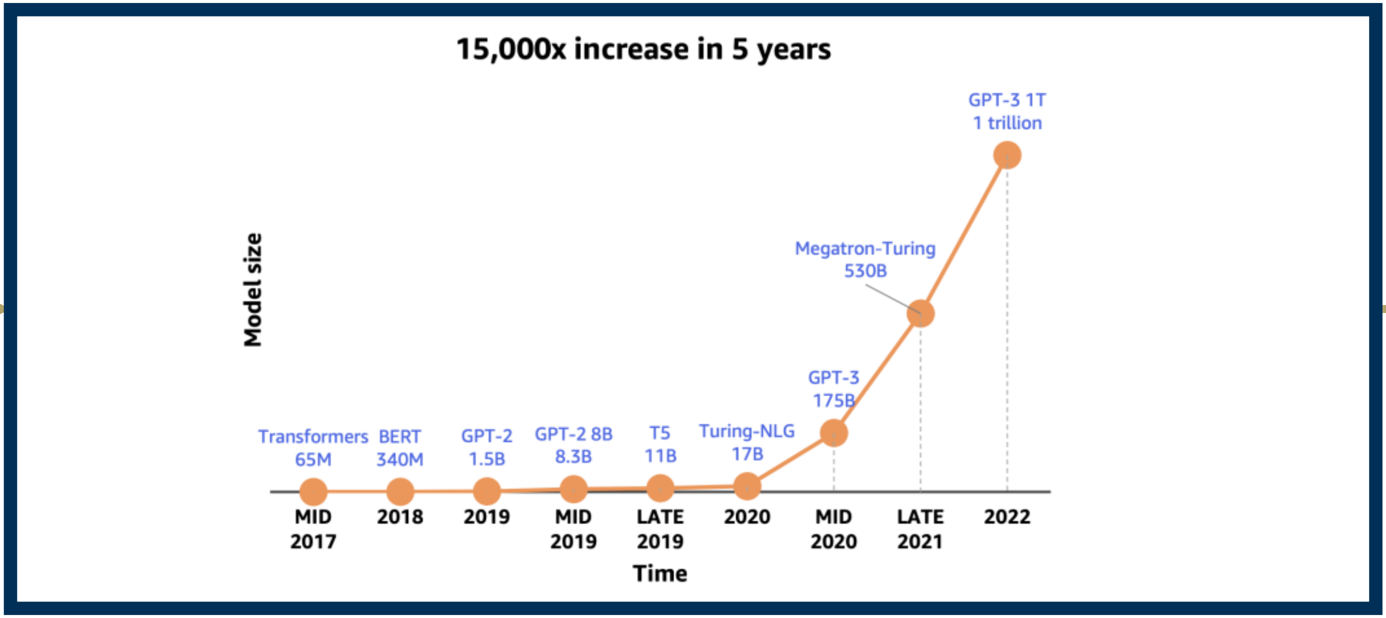
Stationary property of images allow for a small number of convolution kernels





Ex. LeCun, 2015

Transformers, Large Language Models and Foundation Models



Cat

Primary reasons for advancements:

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

Foundation Models

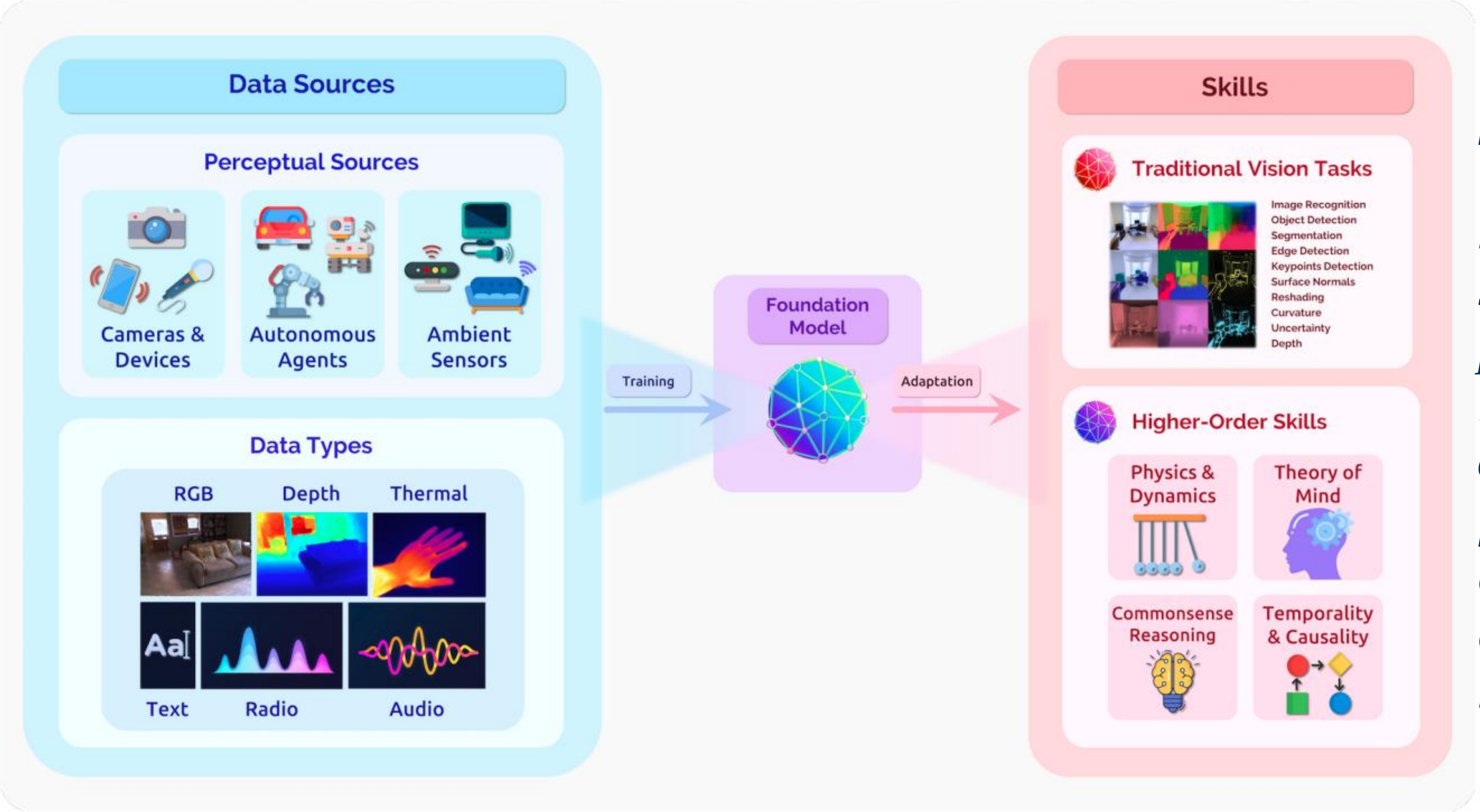
Origin of the term Foundation Models

- **Foundation models** are like any other deep network that have employed **transfer learning**, except **at *scale***
- ***Scale*** brings about ***emergent* properties** that are common between tasks
- **Before 2019:** Base architectures that powered multiple neural networks were **ResNets, VGG** etc.
- **Since 2019: BERT, DALL-E, GPT, Flamingo**
- Changes since 2019: **Transformer architectures and Self-Supervision**

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

Foundation Models

Origin of the term Foundation Models



*‘By harnessing **self-supervision at scale**, foundation models for vision have the potential to **distill raw, multimodal sensory information into visual knowledge**, which may effectively support traditional **perception tasks** and possibly enable new progress on challenging higher-order skills like **temporal and commonsense reasoning** These inputs can come from a **diverse range of data sources** and application domains, suggesting promise for applications in **healthcare and embodied, interactive perception settings**’*

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

Deep Learning at Inference

What, Where, and When is Inference?

Ability of a system to predict correctly on novel data

Novel data sources:

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

→ Trained Model → Cat

Deep Learning at Inference

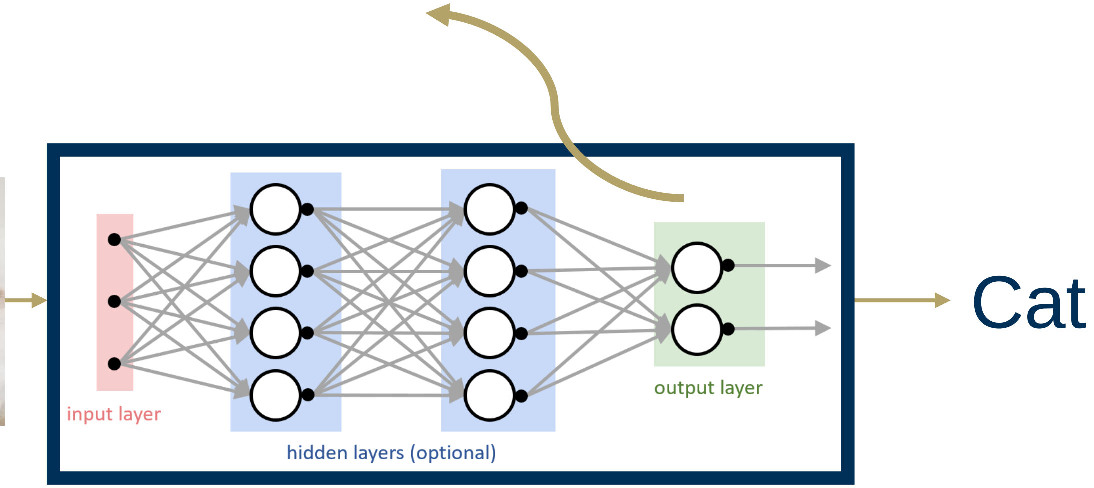
What, Where, and When is Inference?

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

Novel data sources:

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

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



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

Deep Learning at Inference

What, Where, and When is Inference?

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

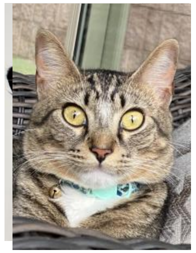
Novel data sources:

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

Trained Model at Testing → Cat, Cat, Cat

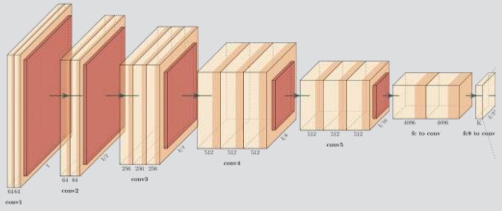
Trained Model at Deployment → Cat

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

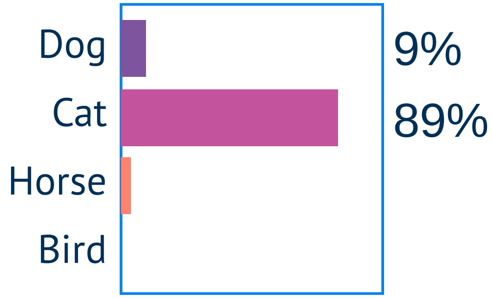


x

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



Predicted
Class Probability



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

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

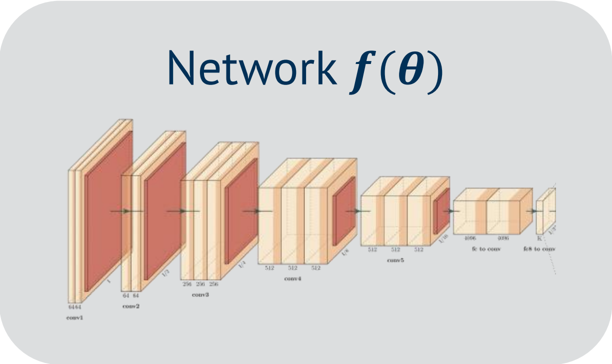
$\hat{y}$ = Logits
y = Predicted Class
$p(\hat{y})$ = Probabilities
$f(\cdot)$ = Trained Network
χ = Training data

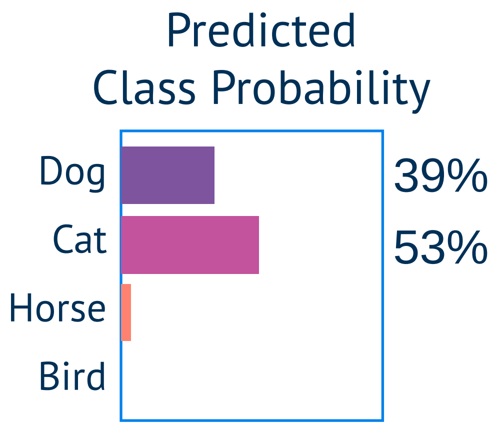
Deep Learning at Inference

Application: Robust Classification

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

x'





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

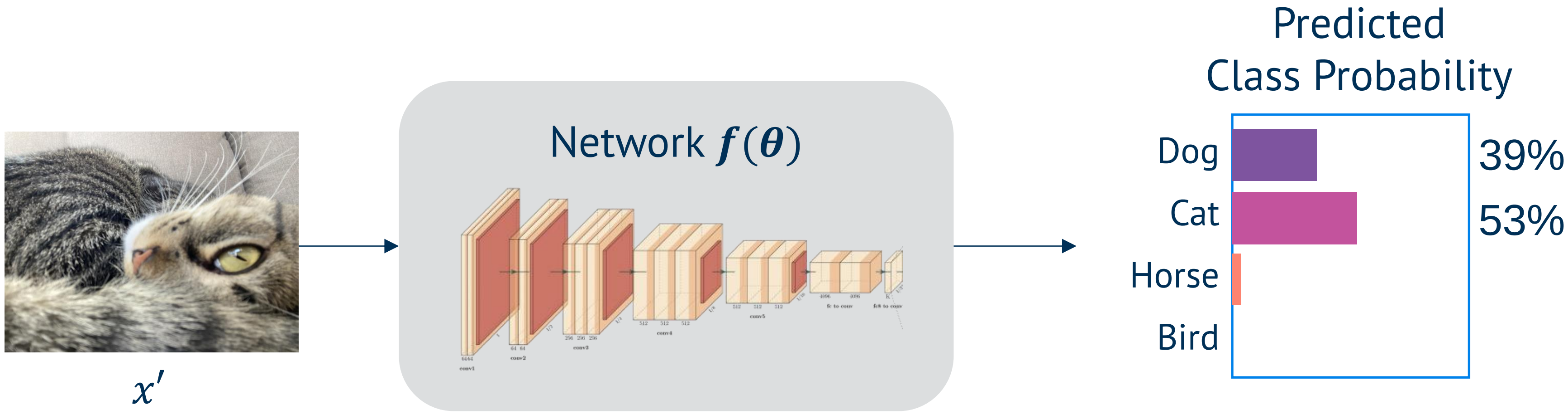
$$\hat{y} = f(x' + \epsilon)$$
$$y = argmax_i\, \hat{y}$$
$$p(\hat{y}) = T(f(x' + \epsilon))$$

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

Deep Learning at Inference

Application: Robust Classification

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



To achieve robustness at Inference, we need the following:

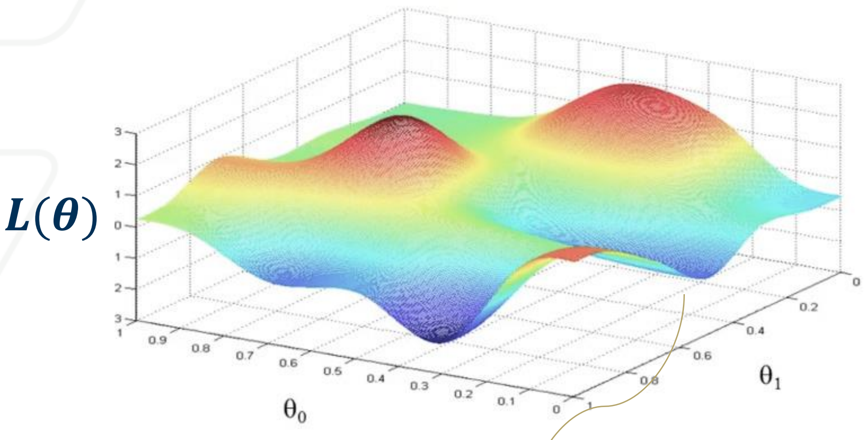
- **Information** provided by the novel data as **a function of training distribution**
- Methodology to **extract information** from novel data
- **Techniques** that utilize the information from novel data

Why is this Challenging?

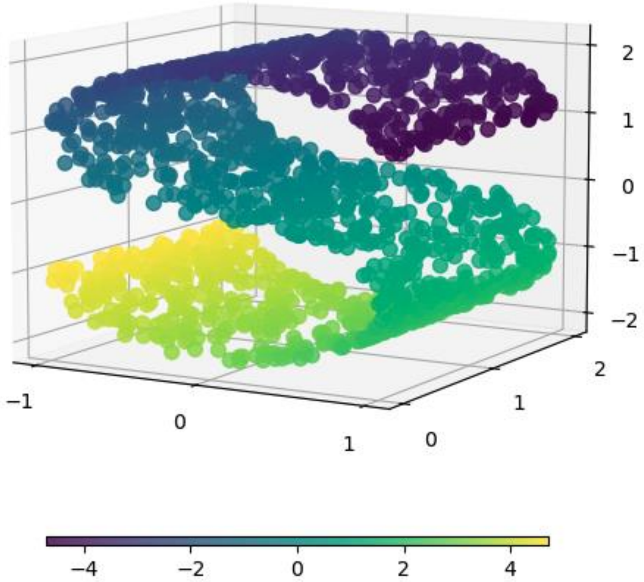
Challenges at Inference

A Quick note on Manifolds..

Manifolds are compact topological spaces that allow exact mathematical functions



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



Real data visualizations generated using dimensionality reduction algorithms (Isomap)

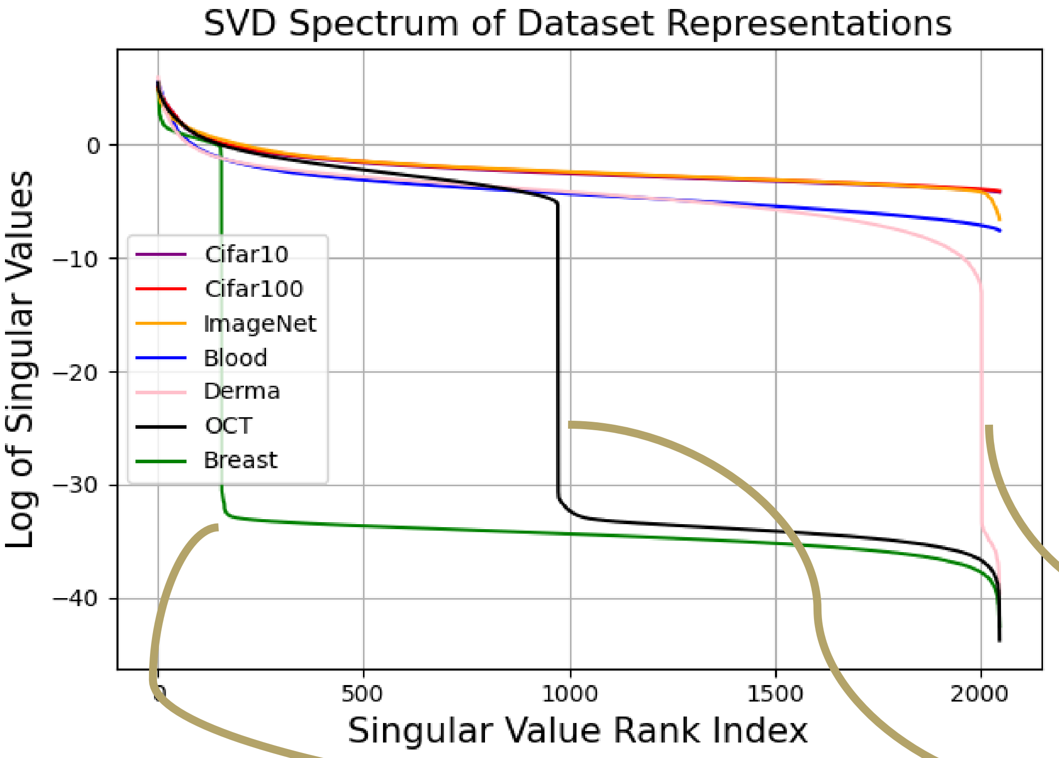
Challenges at Inference

Manifold evaluation at Test-Time Inference without Labels

Hierarchical Constrained Contrastive Learning

SCAN ME

The change in singular values indicate 'goodness' of a self-supervised model for a given dataset



- Construct covariance matrix of the dataset of representations
- Take SVD and order all singular values.
  - The singular values in decreasing order are plotted on the left for different datasets
- 'Better suited-data' for a trained model has no dimensional collapse
- **Conclusion: The natural image trained self-supervised learning model is ill-suited to be utilized for Breast, OCT, and derma datasets**

Dimensional collapse

Kokilepersaud, Kiran, et al. "Taxes are All You Need: Integration Of Taxonomical Hierarchy Relationships Into the Contrastive Loss." *2024 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2024.

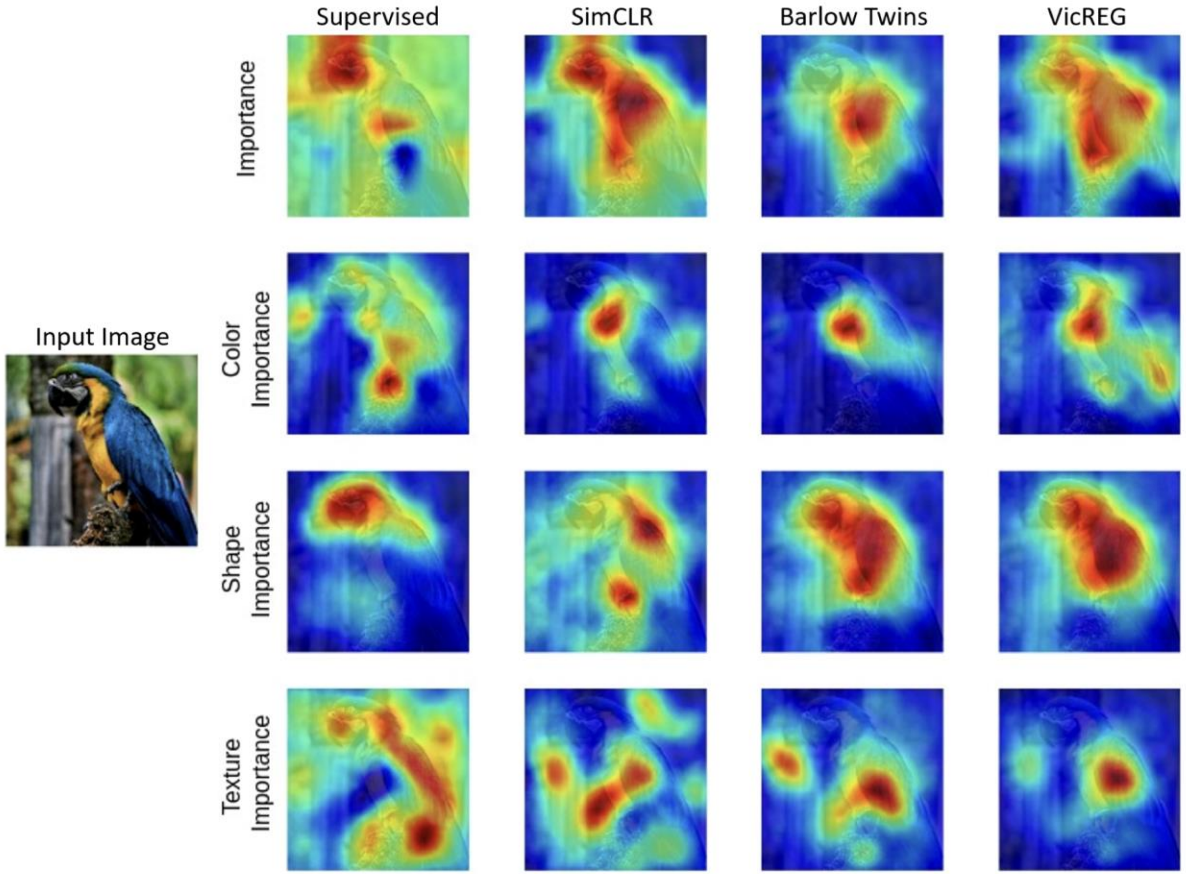
Challenges at Inference

Manifold evaluation at Test-Time Inference without Labels

Perceptual Components in Self-Supervised Learning

SCAN ME

The similarity of concepts like shape, color, and textures between different self-supervised training regimens and the supervised version indicate 'goodness' of that regimen



- **Column 1**: Given the task of bird classification and the bird class, explanations can be constructed for specific perceptual components like color, shape, and texture
- **Columns 2, 3, and 4**: Given only a pre-text task and no true ground truth, we can construct visual explanations for the same concepts

- Construct correlation score between column 1 and each o the other columns.

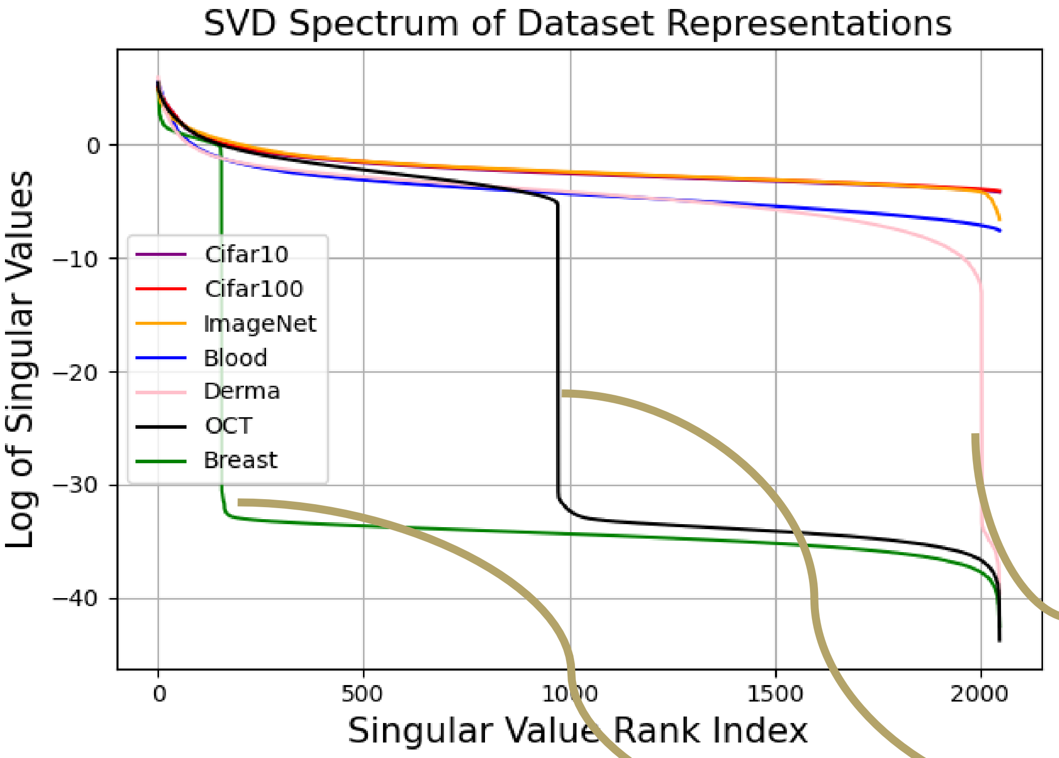
More correlation = better suited for downstream task

Y. Yarici, K. Kokilepersaud, M. Prabhushankar, G. AlRegib, "Explaining Representation Learning with Perceptual Components," in *2024 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates (UAE), 2024.*

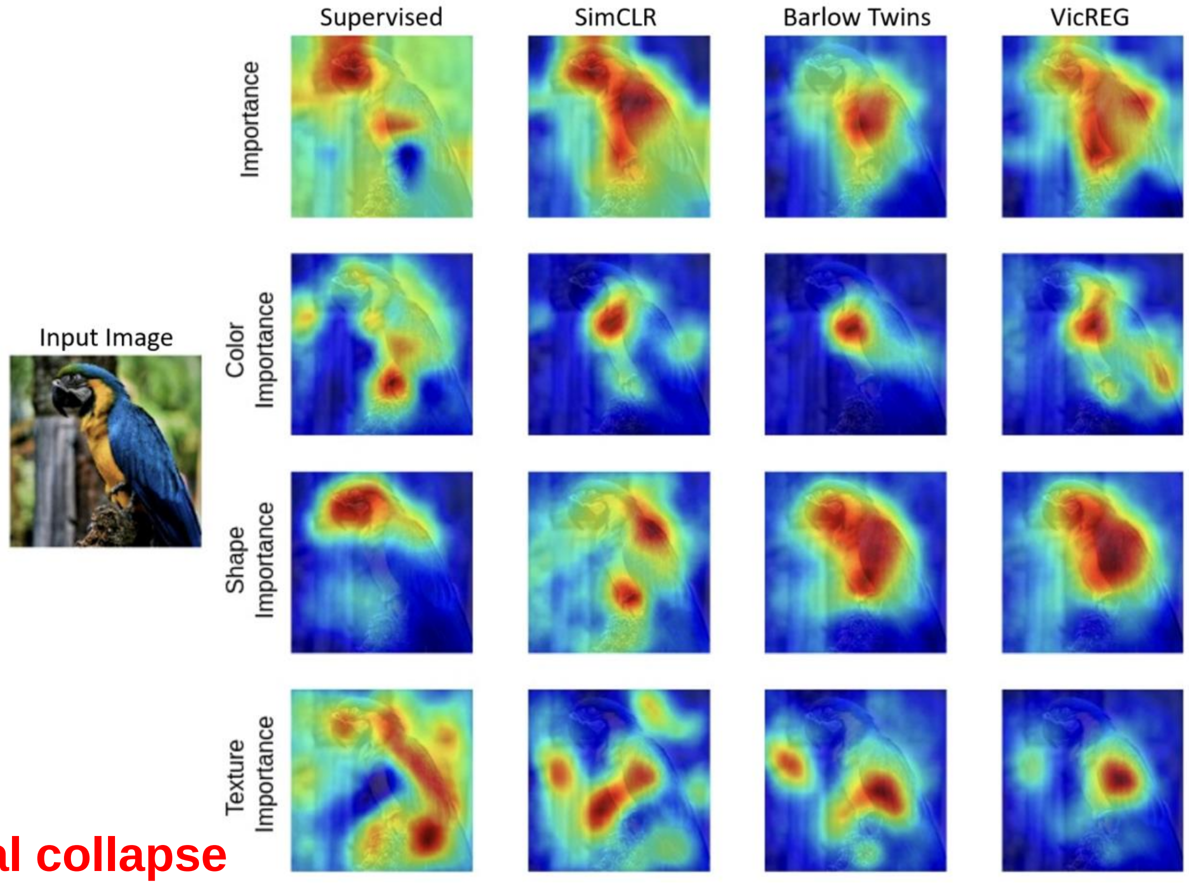
Perceptual Components in Self-Supervised Learning

SCAN ME

Both these methods work on 'test-time' inference; we need access to a large dataset to (i) construct SVD of dataset, (ii) correlation across image explanations



Dimensional collapse



WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

OLIVES @GeorgiaTech

Georgia Tech

Challenges at Inference

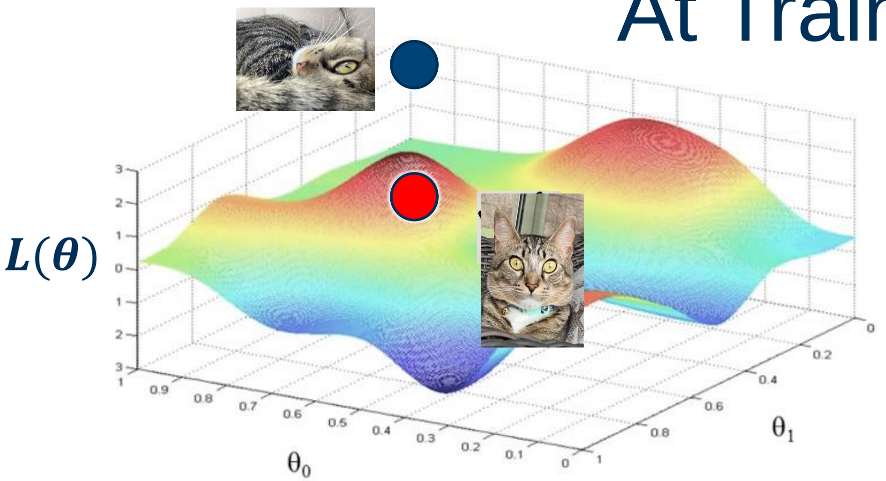
Deployment Inferential Evaluation

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

At Inference

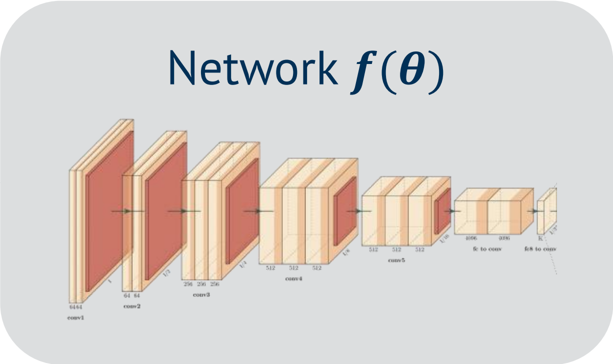
Trained network knowledge is not easily accessible

At Training

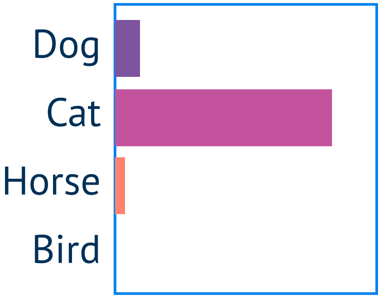


At training, we have access to all training data.

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

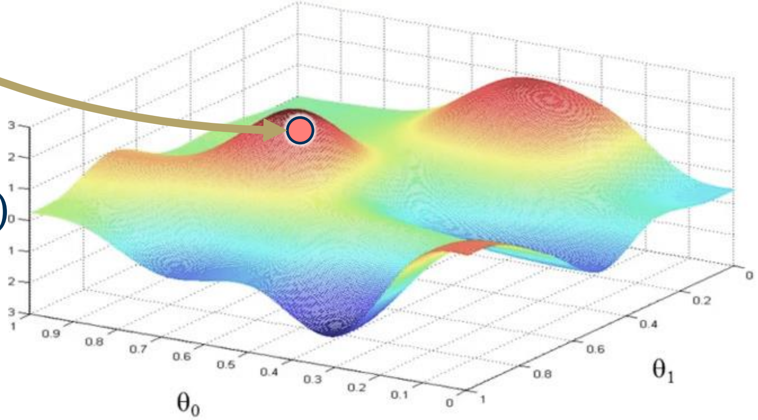


Predicted Class Probability



Fisher Information

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



$l(\theta|x)$

Likelihood function

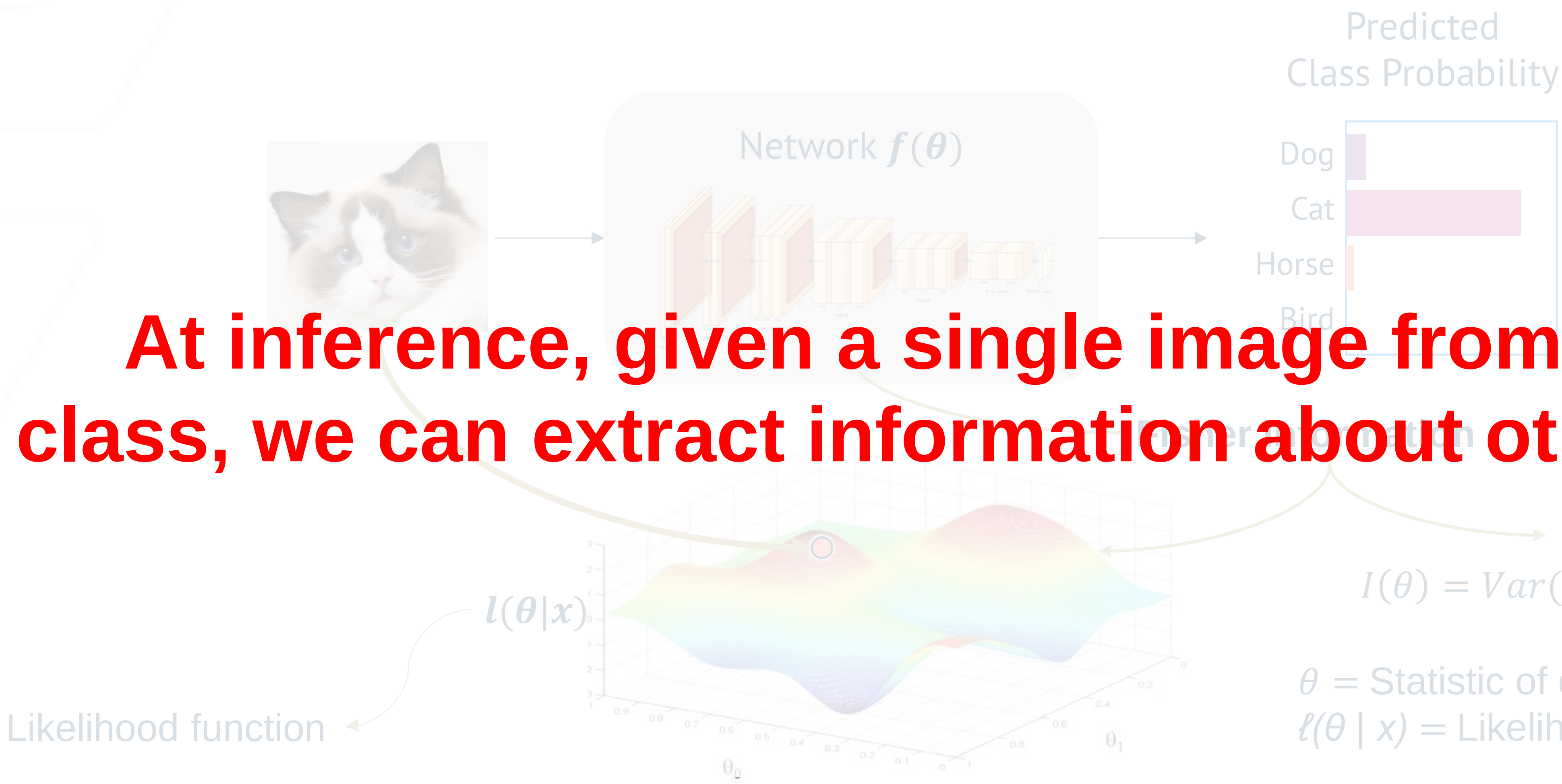
θ = Statistic of distribution

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

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

Information at Inference

Information at Inference



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

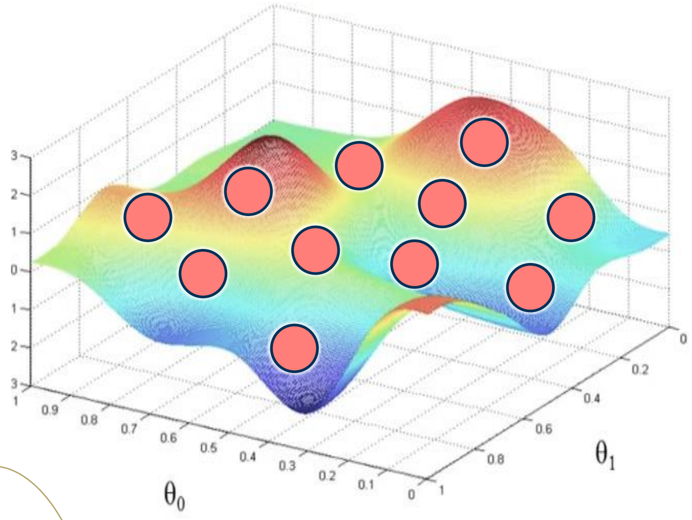
θ = Statistic of distribution
$\ell(\theta \mid x)$ = Likelihood function

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

Information at Inference
Gradients as Fisher Information

Gradients infer information about the statistics of underlying manifolds

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



Likelihood function instead of loss manifold

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

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

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

$E[\cdot] = \text{Expectation}$
$U_\theta = \nabla_\theta l(\theta|x)$, Gradients w.r.t. the sample

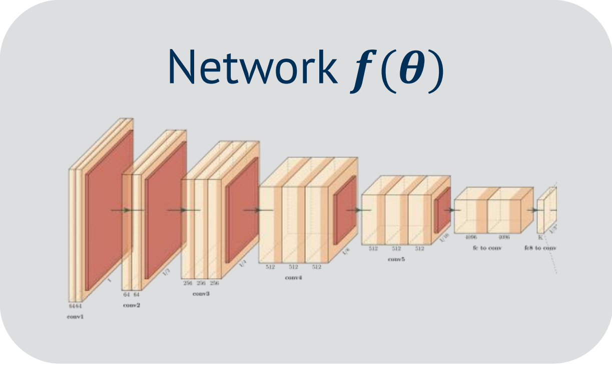
Hence, gradients draw information from the underlying distribution as learned by the network weights!

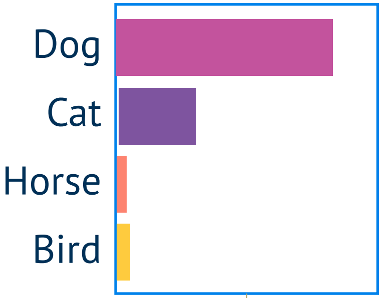
Kwon, Gukyeong, et al. "Backpropagated gradient representations for anomaly detection." *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer International Publishing, 2020.

Case Study: Gradients as Fisher Information in Explainability

Gradients infer information about the statistics of underlying manifolds

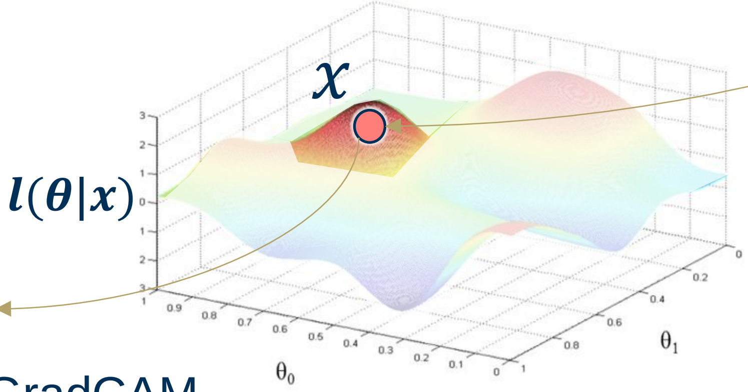
x





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

Local information (specific to x) is sufficient!



Feature attribution via GradCAM

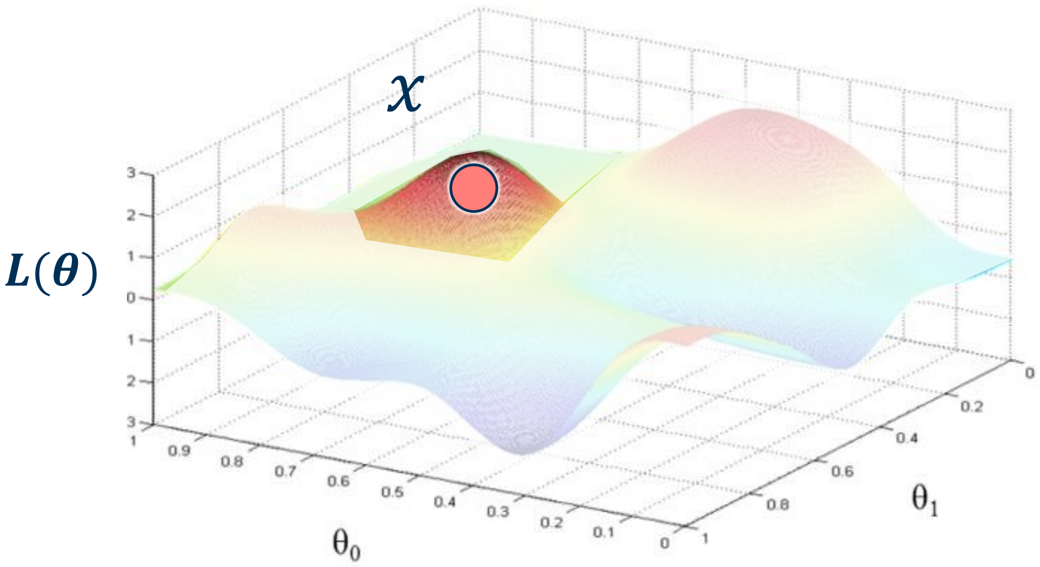
Hence, gradients draw information from the underlying distribution as learned by the network weights!

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

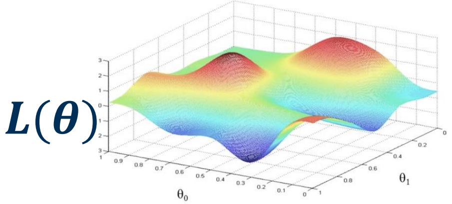
Gradients at Inference

Local Information

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



Ideal

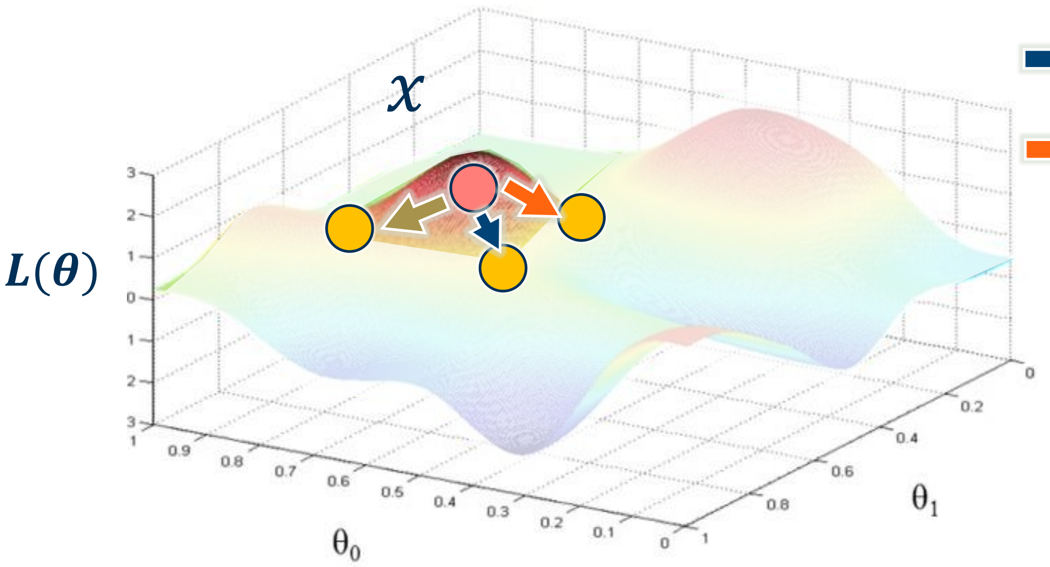


$\boldsymbol{\alpha}\, \nabla_{\boldsymbol{\theta}}\, \boldsymbol{L}(\boldsymbol{\theta})$ provides local information up to a small distance $\boldsymbol{\alpha}$ away from $\boldsymbol{x}$

Gradients at Inference

Direction of Steepest Descent

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



Path 1?

Path 2?

Path 3?

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

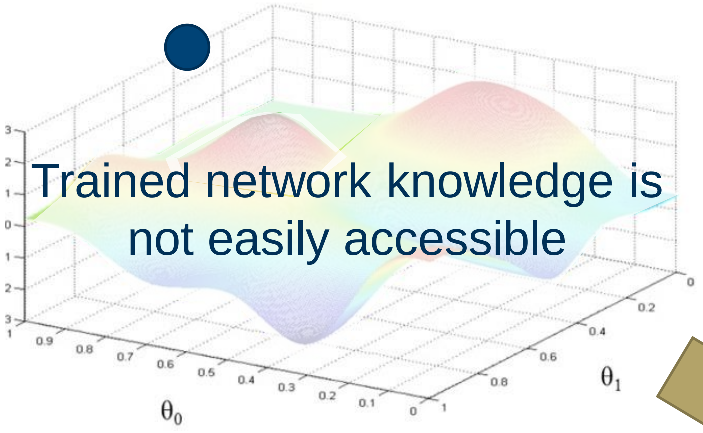
Negative of the gradient provides the **descent direction** towards the local minima, as measured by $L(\theta)$

Gradients at Inference

To Characterize the Novel Data at Inference

At Inference

$L(\theta)$



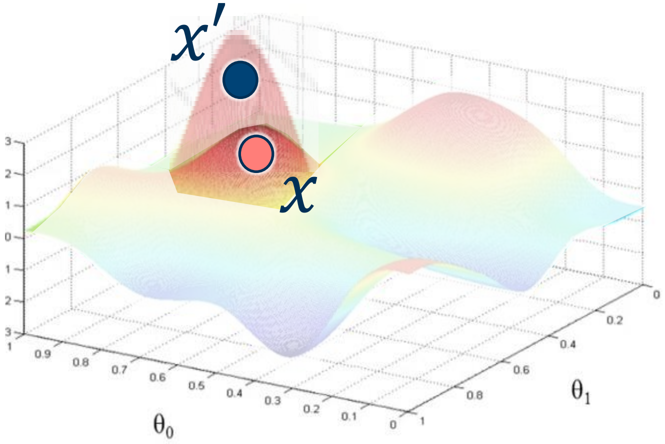
Part 2, 3

Part 4

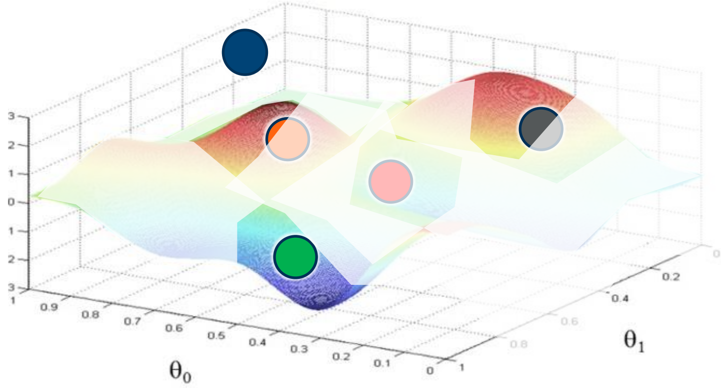
Counterfactual and Contrastive Representations using Gradients

Representation Traversal using Interventions

$L(\theta)$



$L(\theta)$



Inferential Machine Learning

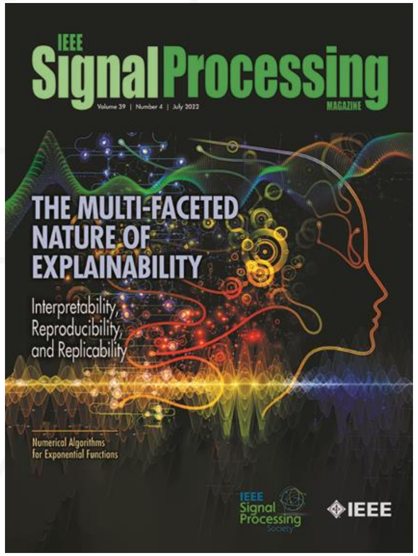
Part 2: Explainability at Inference

Objective

Objective of the Tutorial

To discuss methodologies that promote robust and fair inference in neural networks

- Part 1: Inference in Neural Networks
- **Part 2: Explainability at Inference**
  - Visual Explanations
  - Gradient-based Explanations
  - GradCAM
  - CounterfactualCAM
  - ContrastCAM
  - Case Study: Introspective Learning
- Part 3: Uncertainty and Intervenability at Inference
- Part 4: Intervenability at Inference
- Part 5: Conclusions and Future Directions



Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

Mohit Prabhushankar, PhD
Postdoc

Ghassan AlRegib, PhD
Professor

SCAN ME

Explanations

Visual Explanations

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?

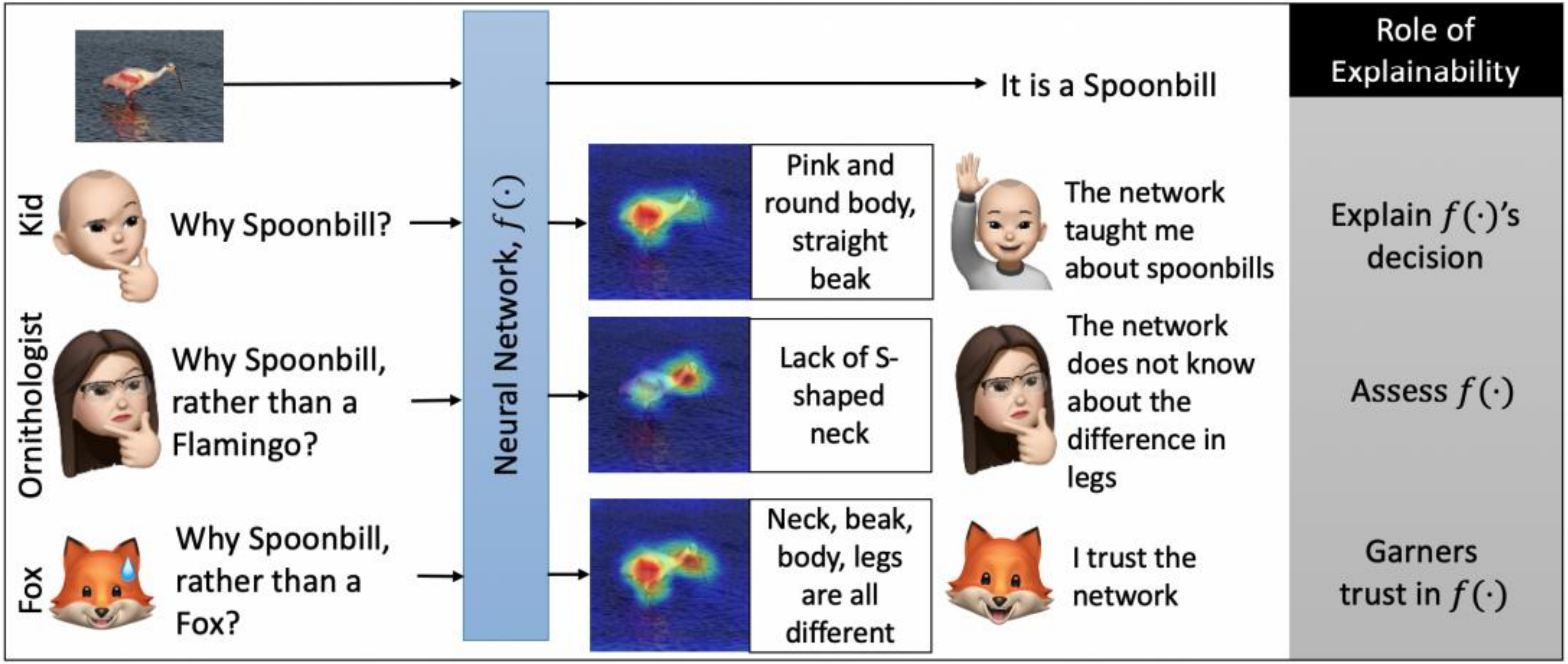
AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks: Towards relevant and contextual explanations. *IEEE Signal Processing Magazine*, *39*(4), 59-72.

Explanations

Role of Explanations – context and relevance

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

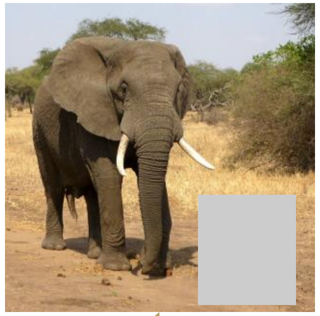


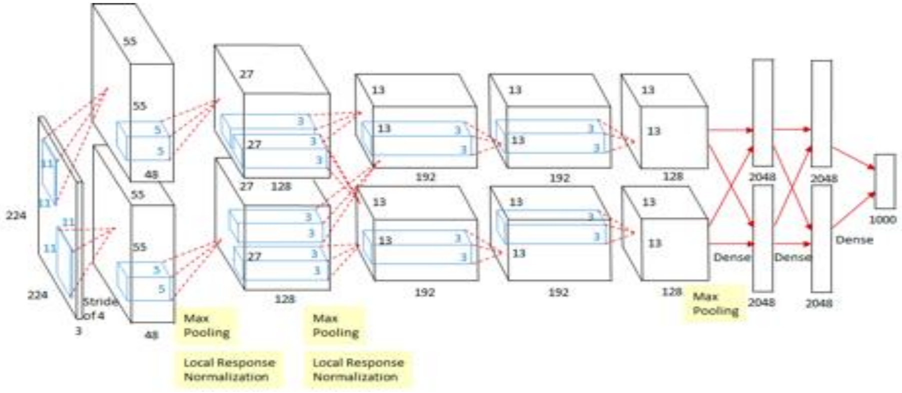
AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks: Towards relevant and contextual explanations. *IEEE Signal Processing Magazine*, *39*(4), 59-72.

Explanations

Input Saliency via Occlusions

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



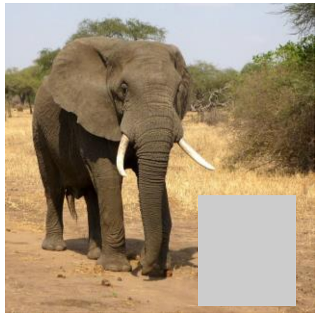


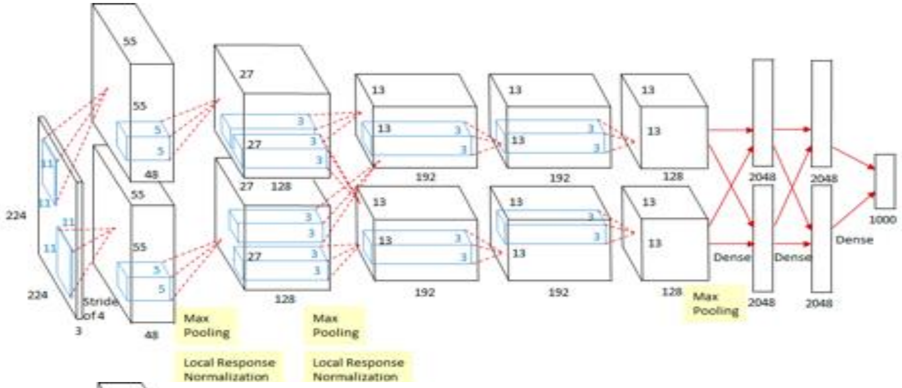
P(elephant) = 0.95

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

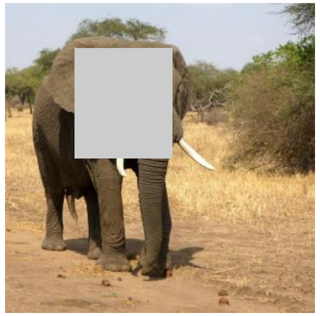
Intervention: Mask part of the image before feeding to CNN, check how much predicted probabilities change

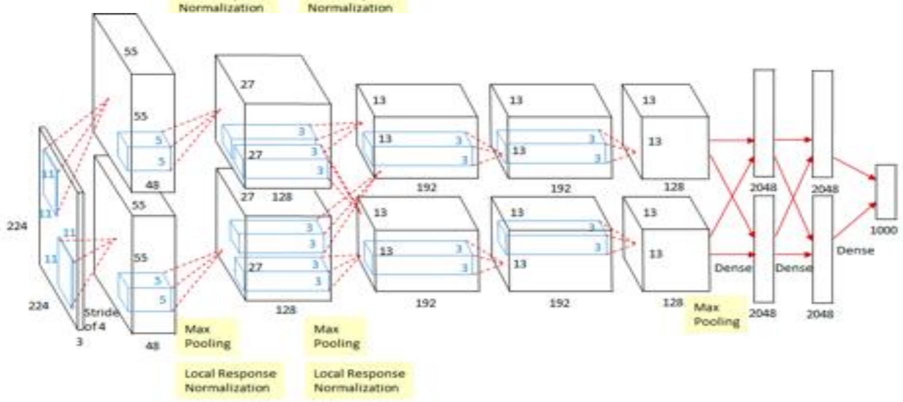




P(elephant) = 0.95

These pixels affect decisions more





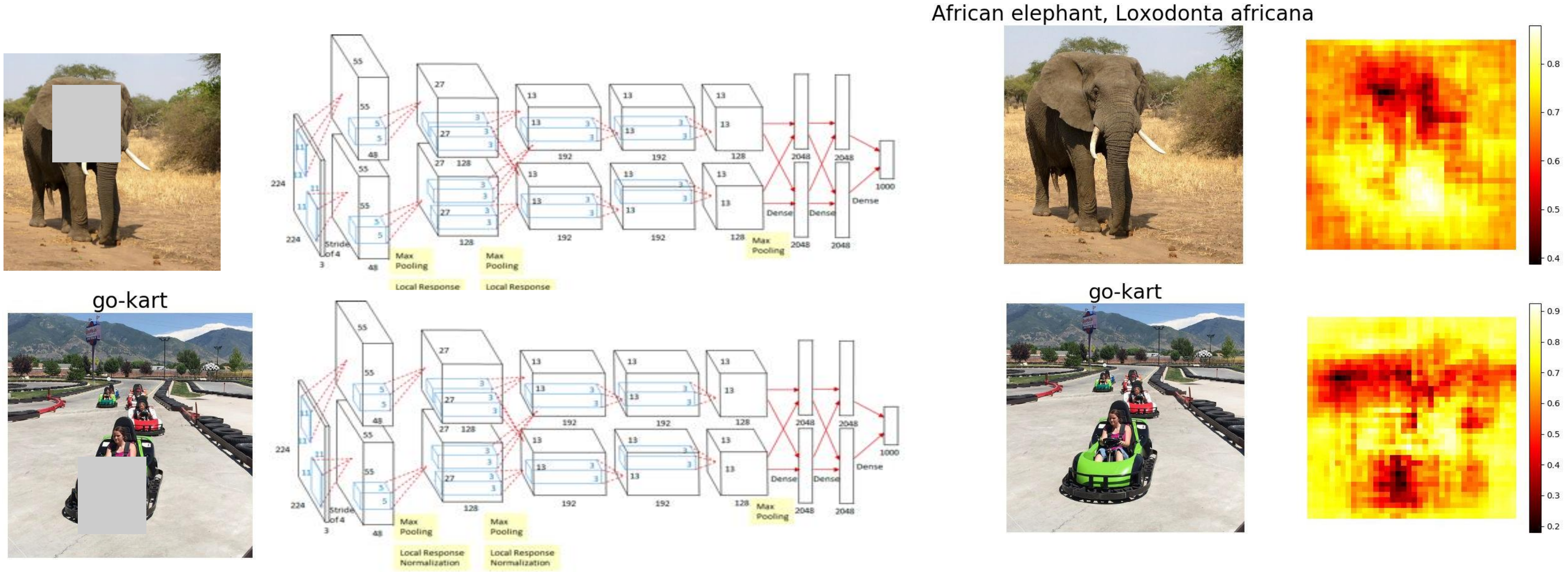
P(elephant) = 0.75

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

Explanations

Input Saliency via Occlusions

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



Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

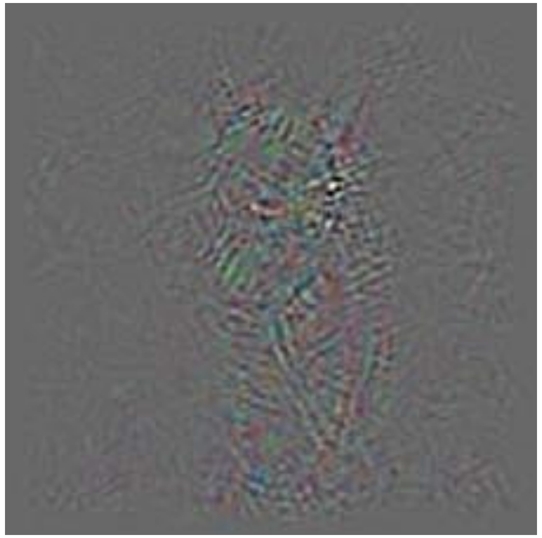
SCAN ME

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

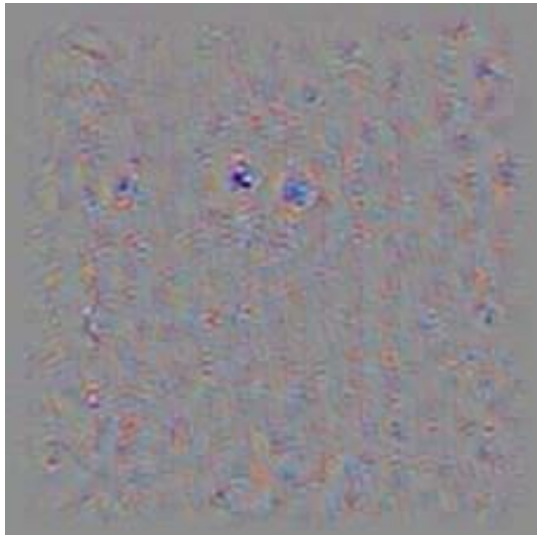
Gradients provide a one-shot means of perturbing the input that changes the output; They provide pixel-level importance scores

Input

Vanilla Gradients



Deconvolution Gradients



Guided Backpropagation

However, localization remains an issue

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

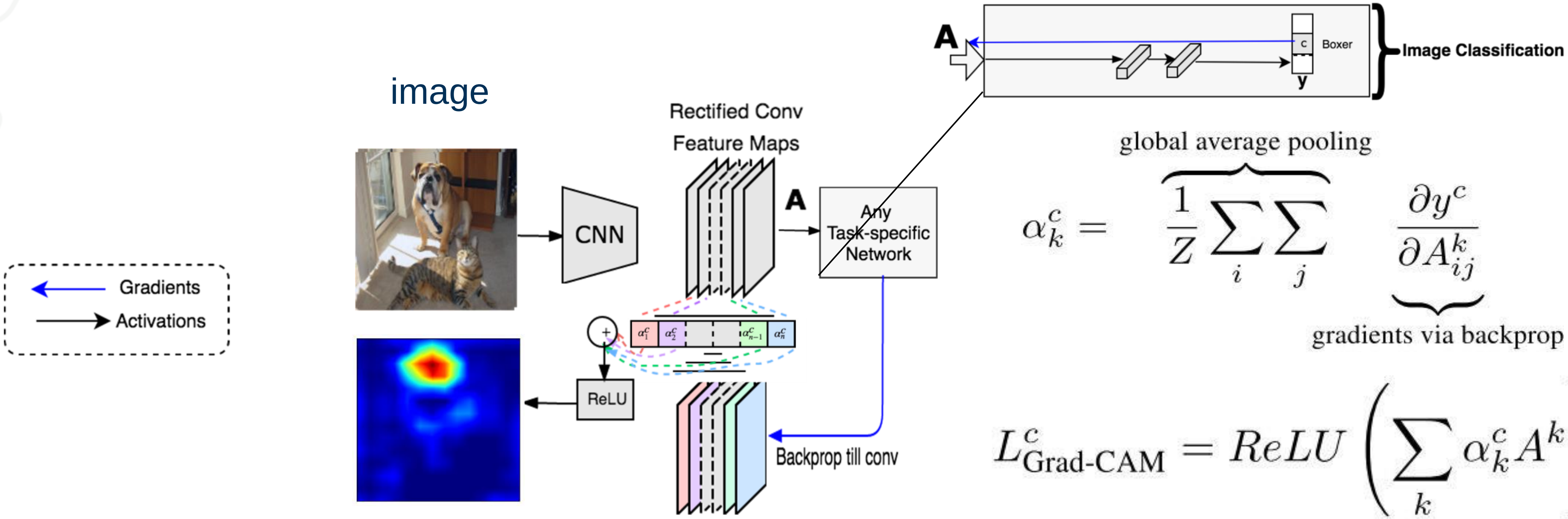
Gradient and Activation-based Explanations

GradCAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

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



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

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

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

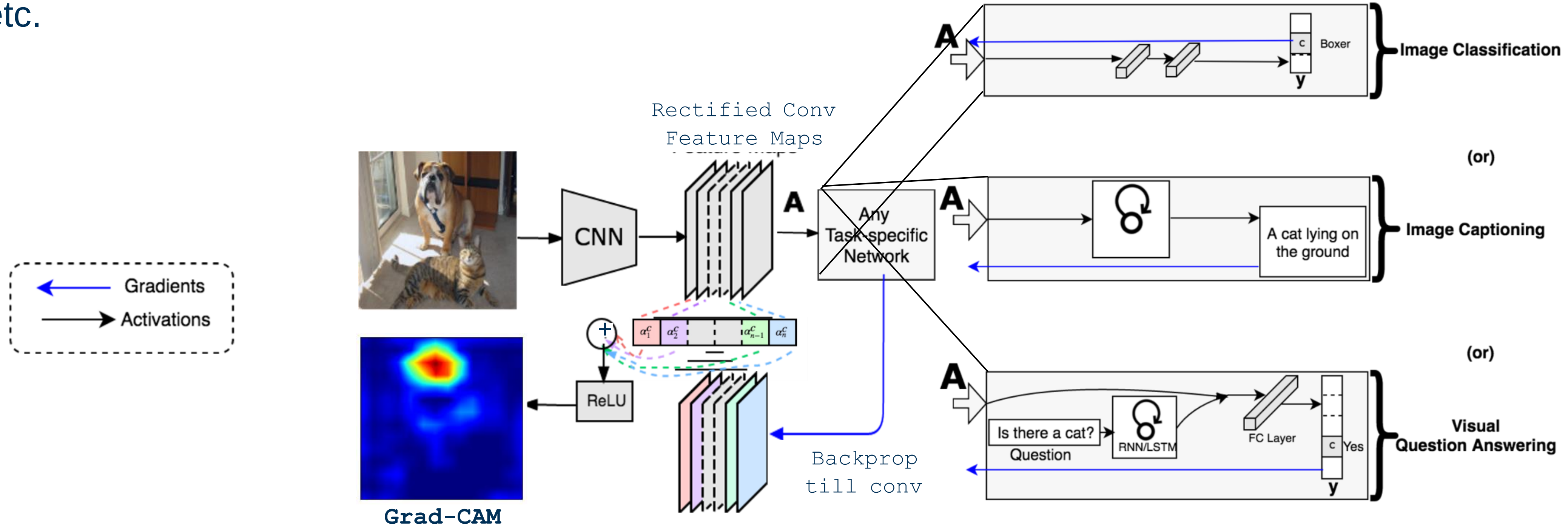
Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

Gradient and Activation-based Explanations

GradCAM

Grad-CAM generalizes to any task:

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



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

Gradient and Activation-based Explanations

Explanatory Paradigms

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

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

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

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

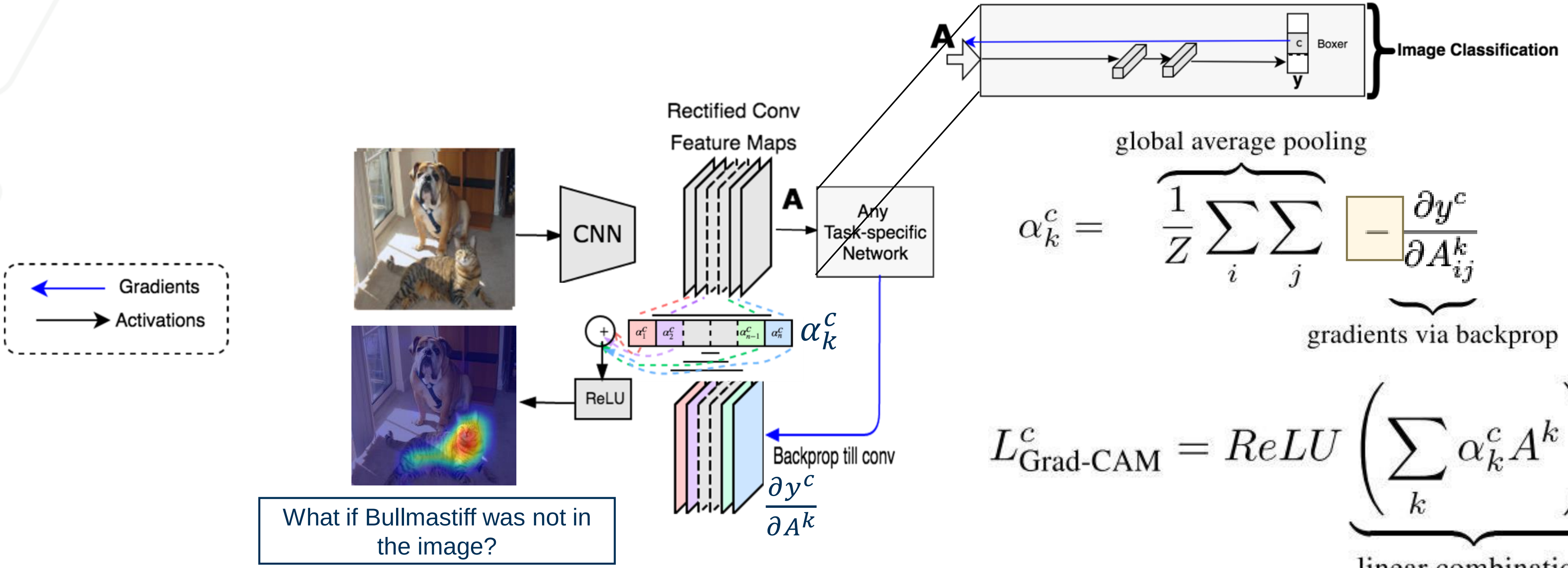
Gradient and Activation-based Explanations

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

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



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

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

Negating the gradients effectively removes these regions from analysis

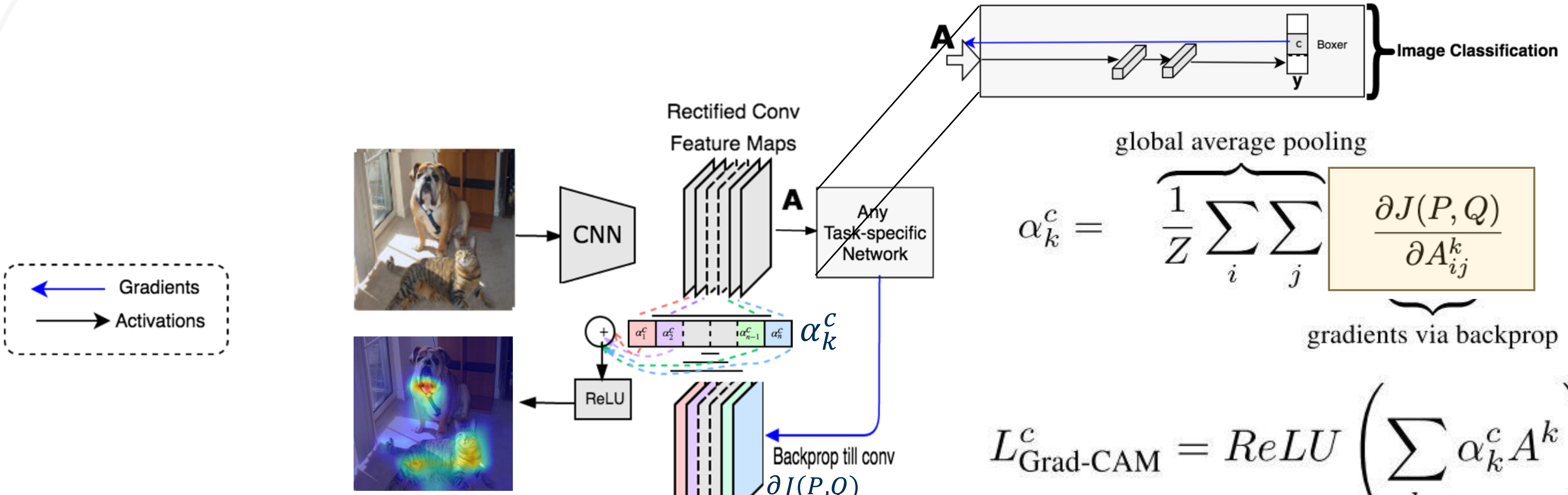
Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

Gradient and Activation-based Explanations

ContrastCAM: Why P, rather than Q?

SCAN ME

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



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

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

Why Bullmastiff, rather than a Boxer?

Contrast-CAM

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

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

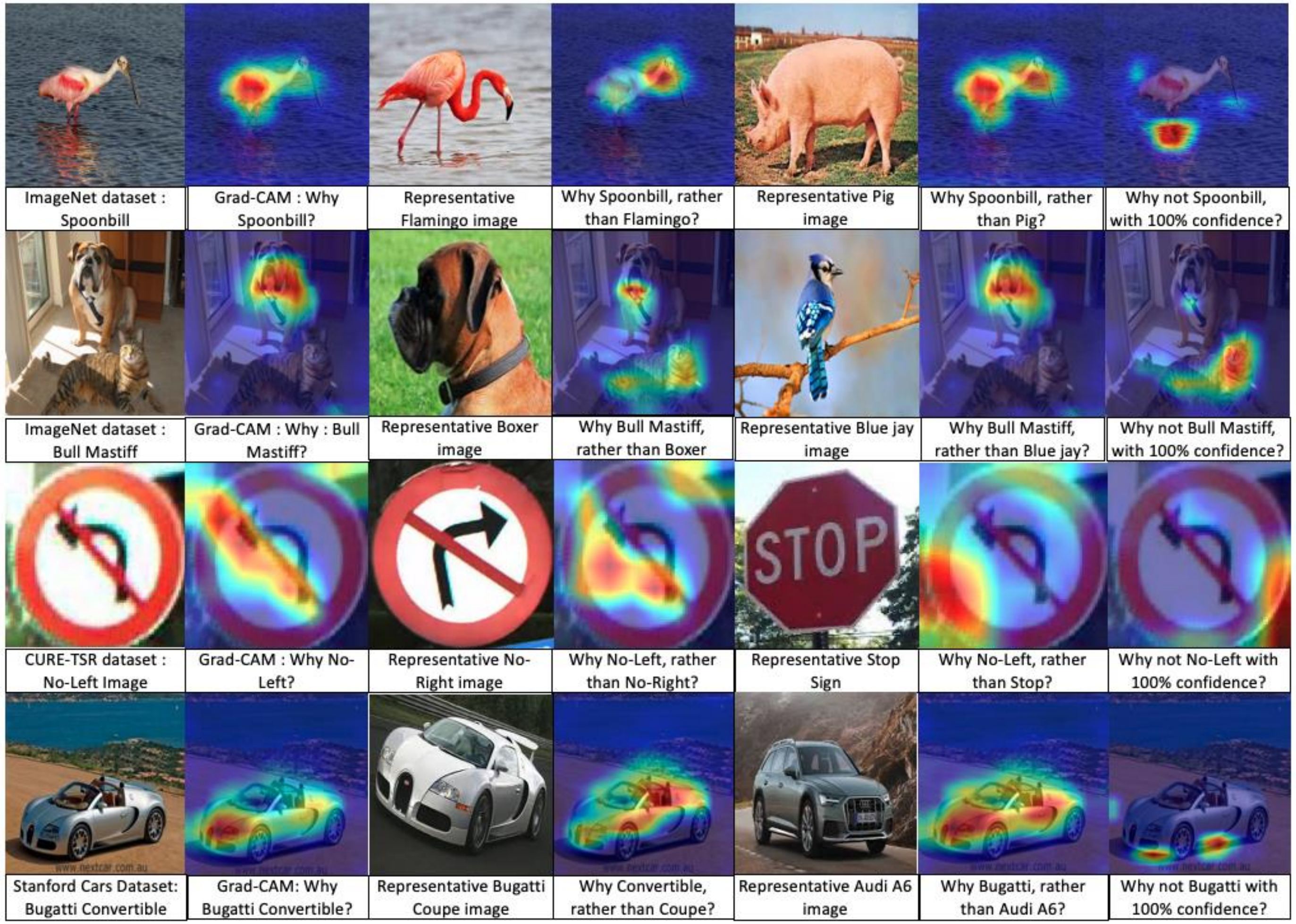
Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

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	



WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

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

OLIVES @GeorgiaTech

Georgia Tech

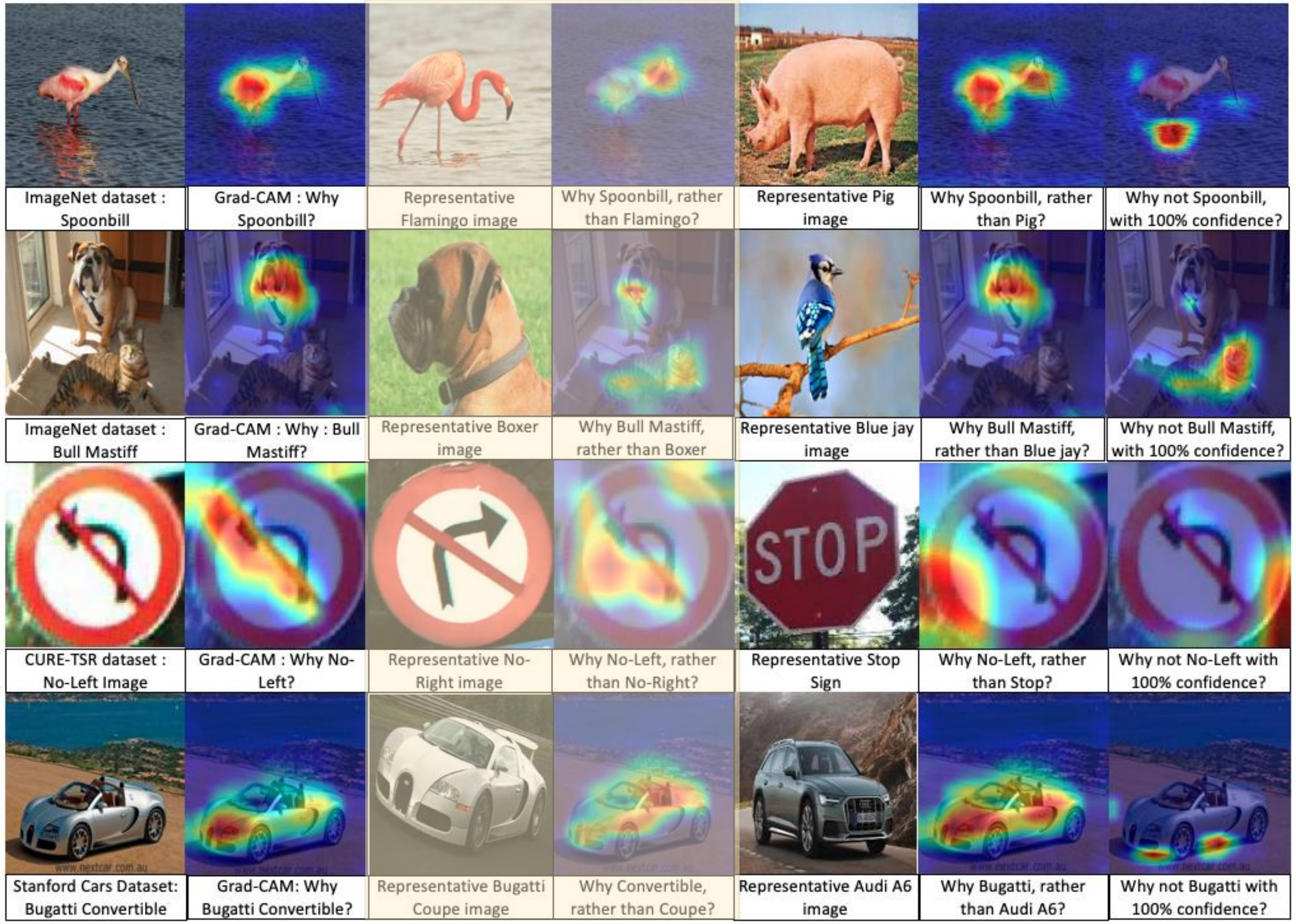
Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

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



Human Interpretable

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

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

Human Interpretable

Same as Grad-CAM

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

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

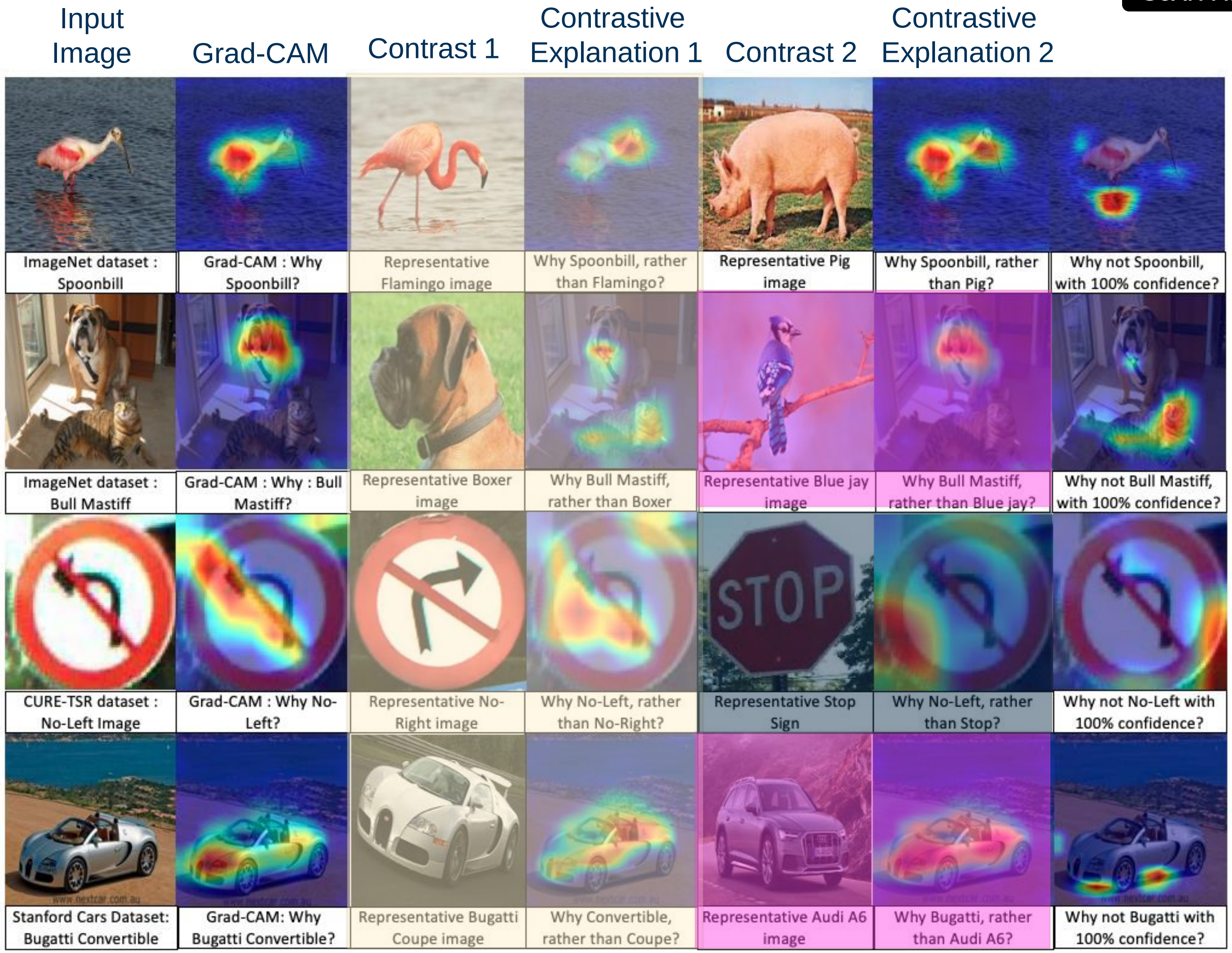
OLIVES @GeorgiaTech

Georgia Tech

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations



Human Interpretable

Same as Grad-CAM

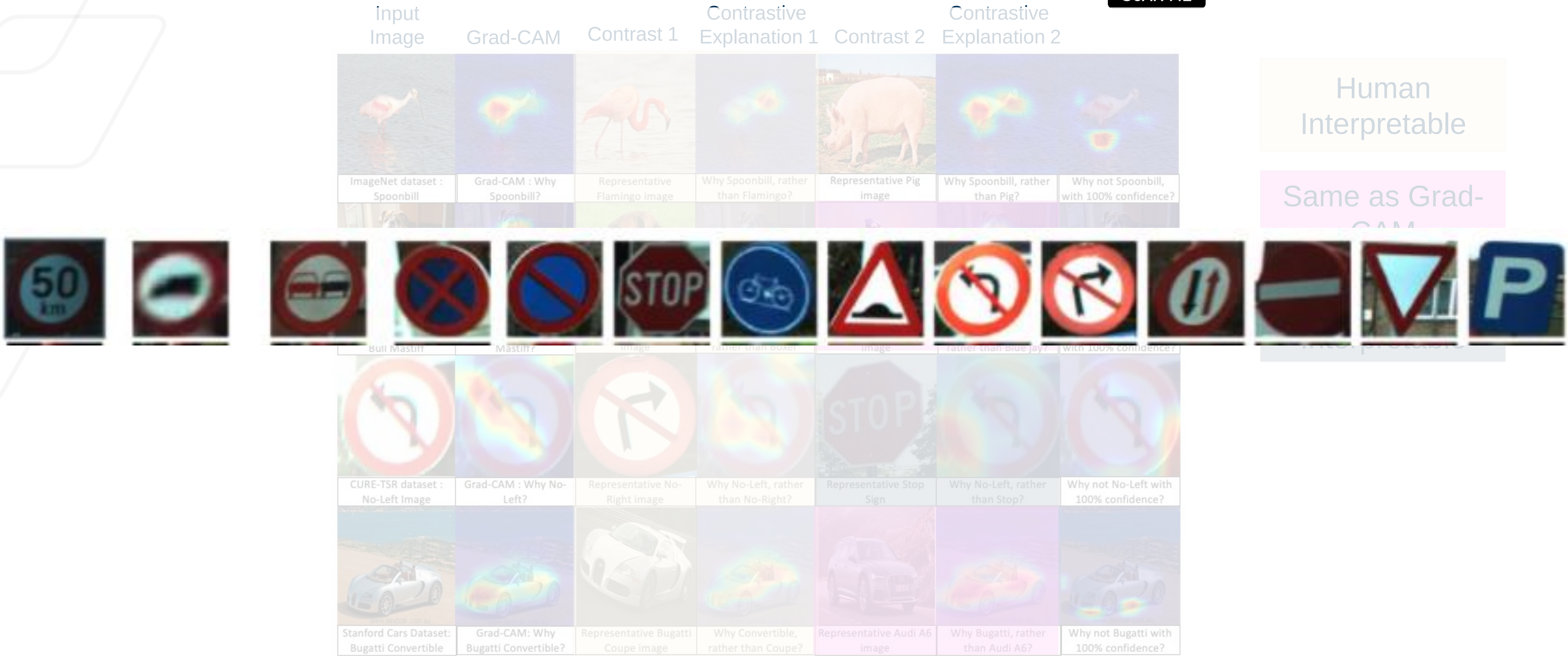
Not Human Interpretable

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



Human Interpretable

Same as Grad-

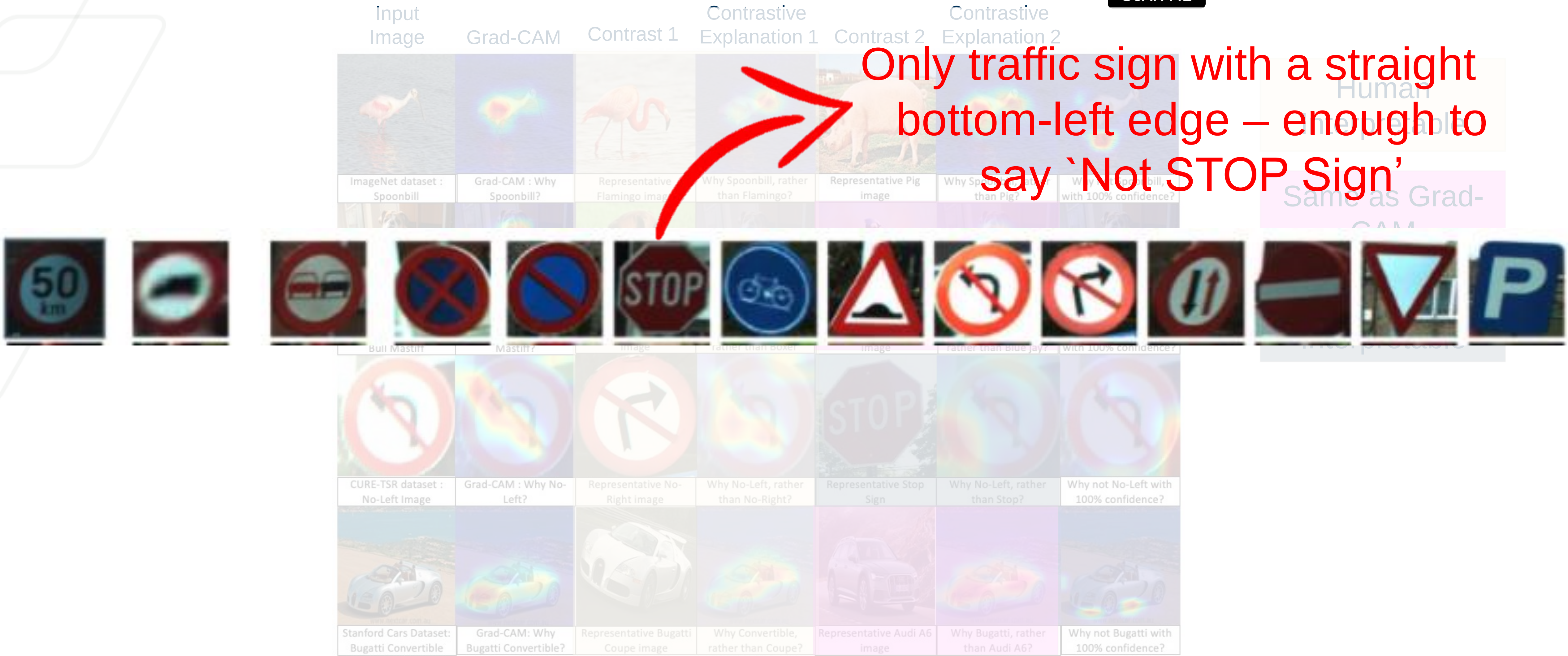
Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

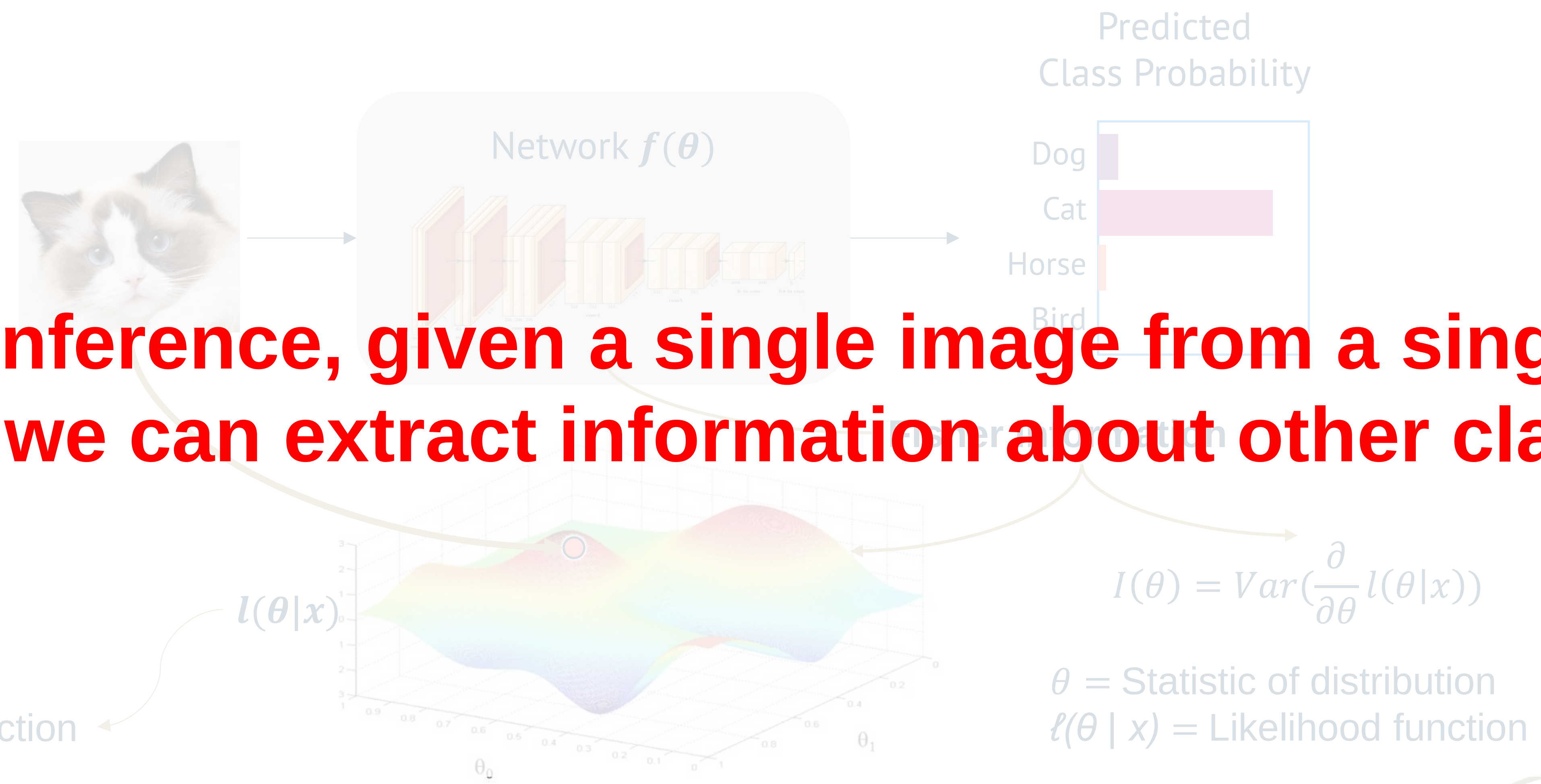
SCAN ME



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.

A Callback…

Information at Inference



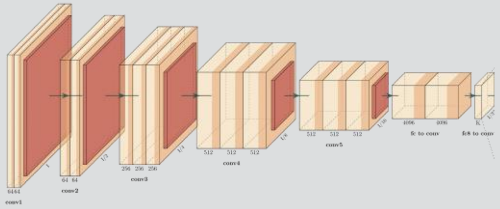
At inference, given a single image from a single class, we can extract information about other classes

Information at Inference

Case Study: Explainability

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

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

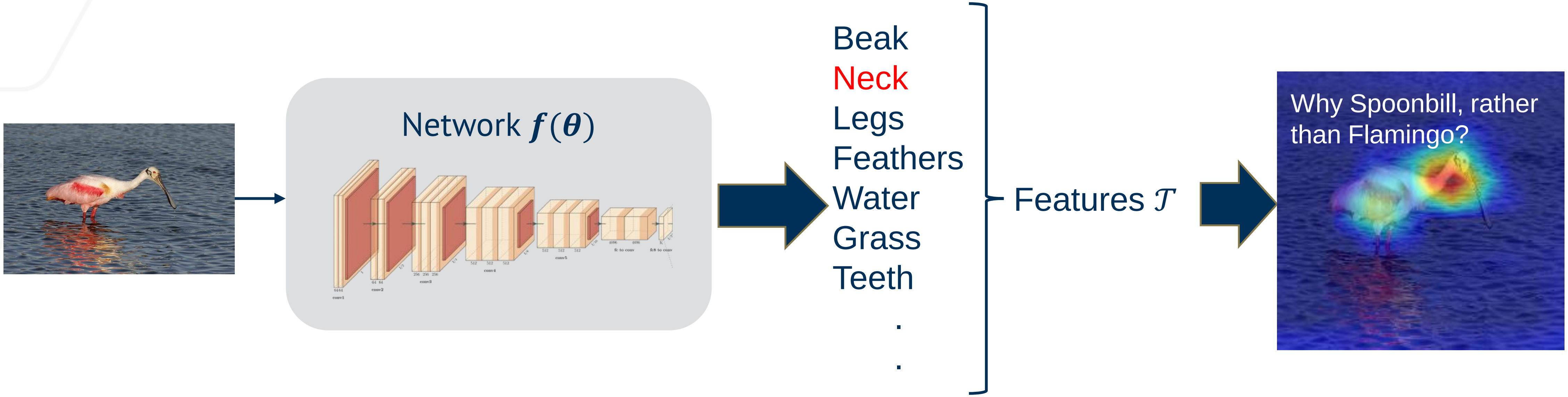


Beak
Neck
Legs
Feathers
Water
Grass
Teeth
.
.

Features $\mathcal{T}$

Case Study: Explainability

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



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

Goal: To extract and utilize this information – Introspective Learning!

Case Study:

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

Mohit Prabhushankar, PhD
Postdoc

Ghassan AlRegib, PhD
Professor

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 bullmastiff?

SCAN ME

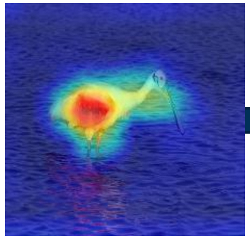
Introspective Learning: A Two-stage Approach for Inference in Neural Networks

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

Introspection

Visual Sensing

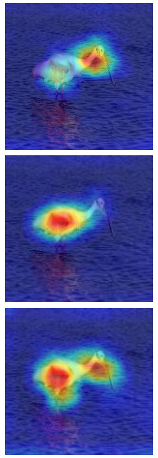
Sense pink feathers, straight beak



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.

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

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..

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

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

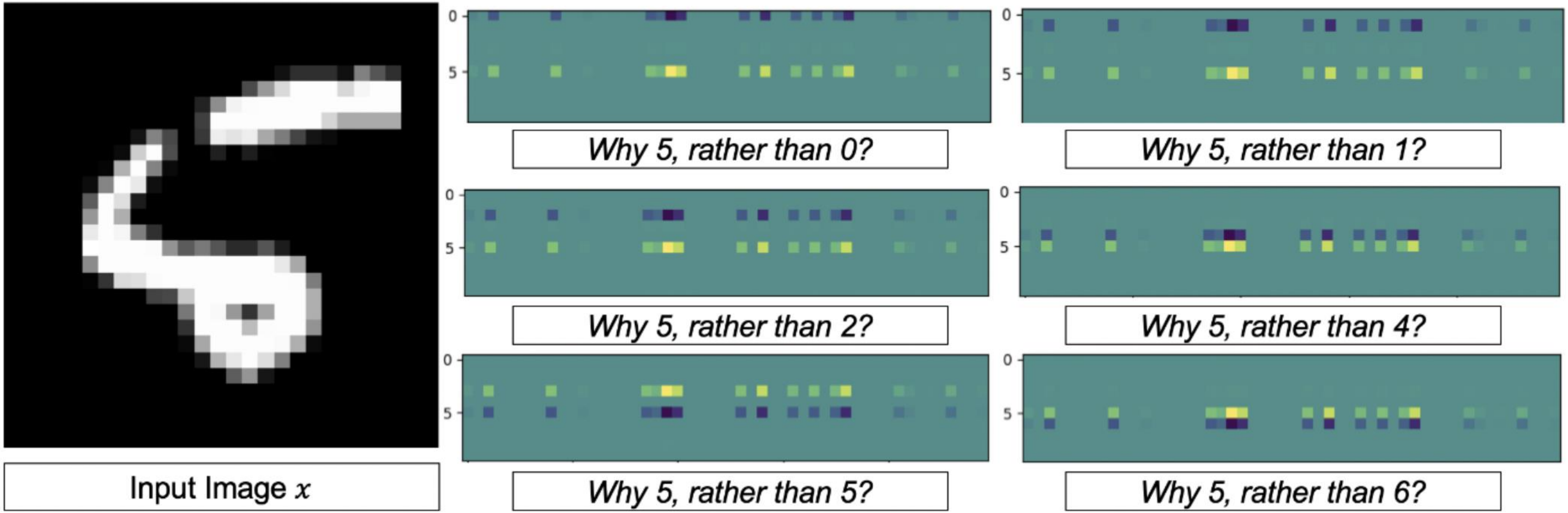
OLIVES @GeorgiaTech

Georgia Tech

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.

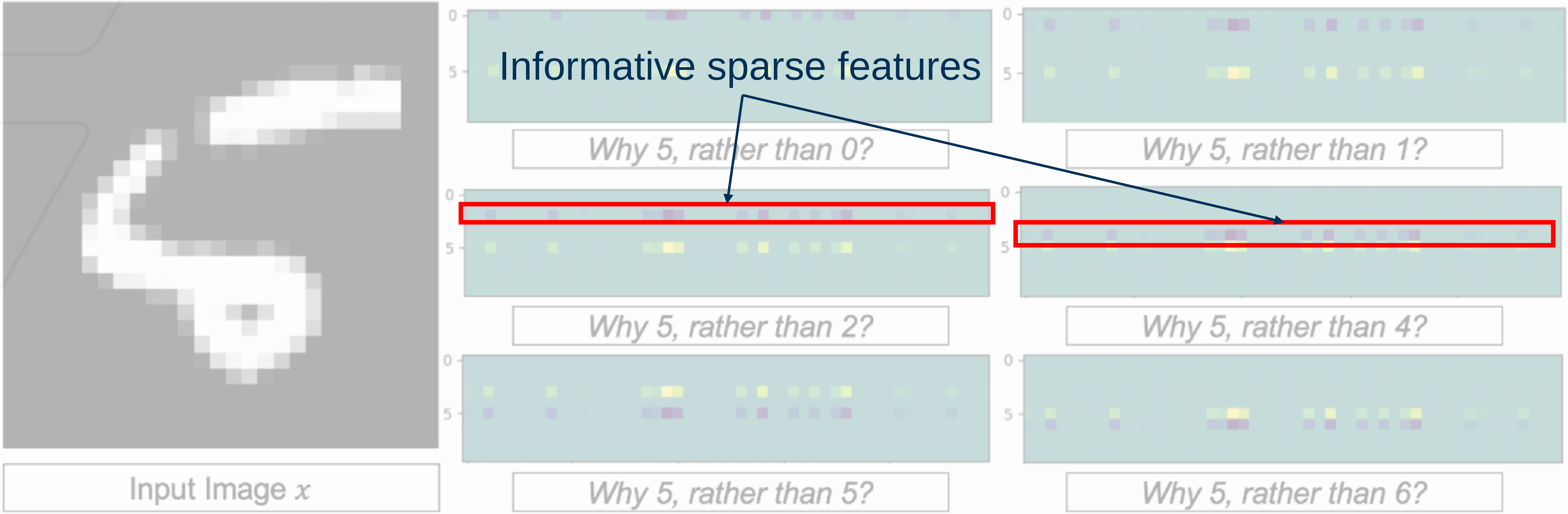
Introspection

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

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

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

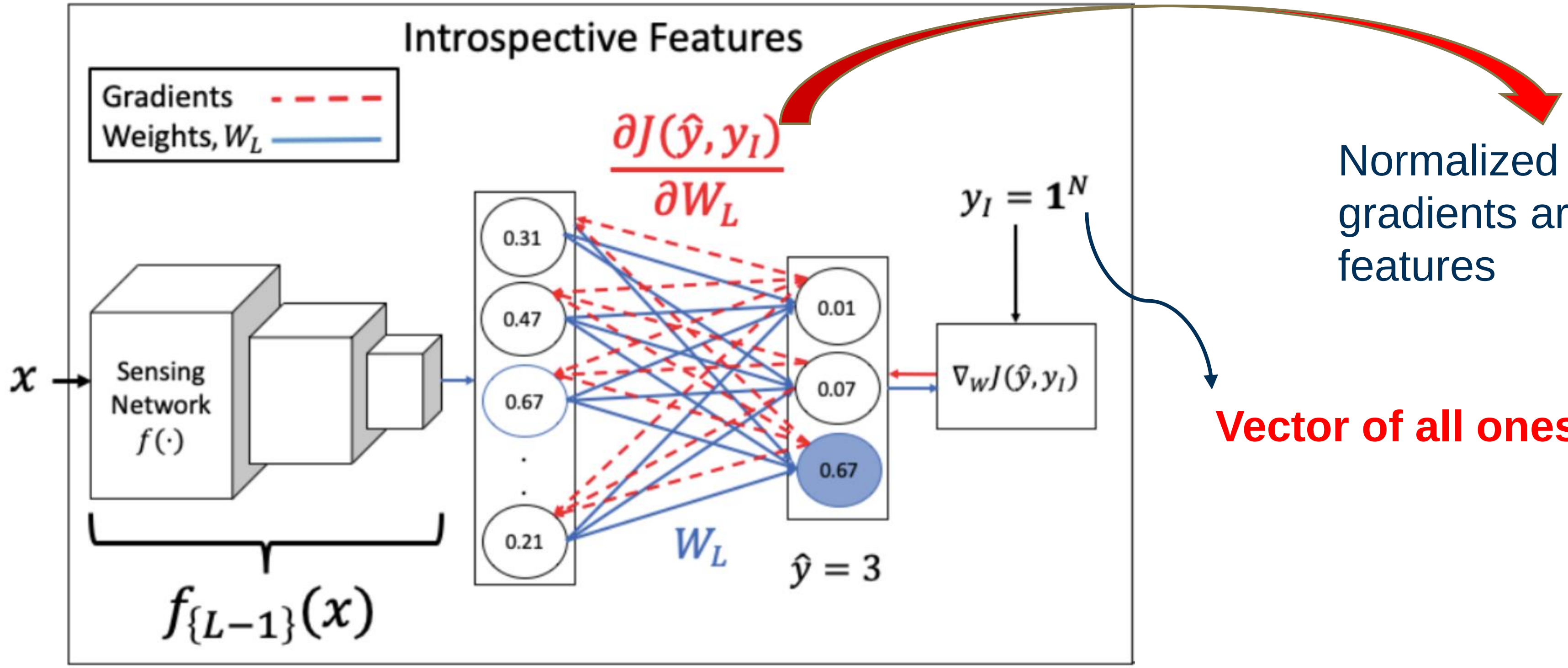
Introspection

Deriving Gradient Features

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

SCAN ME

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



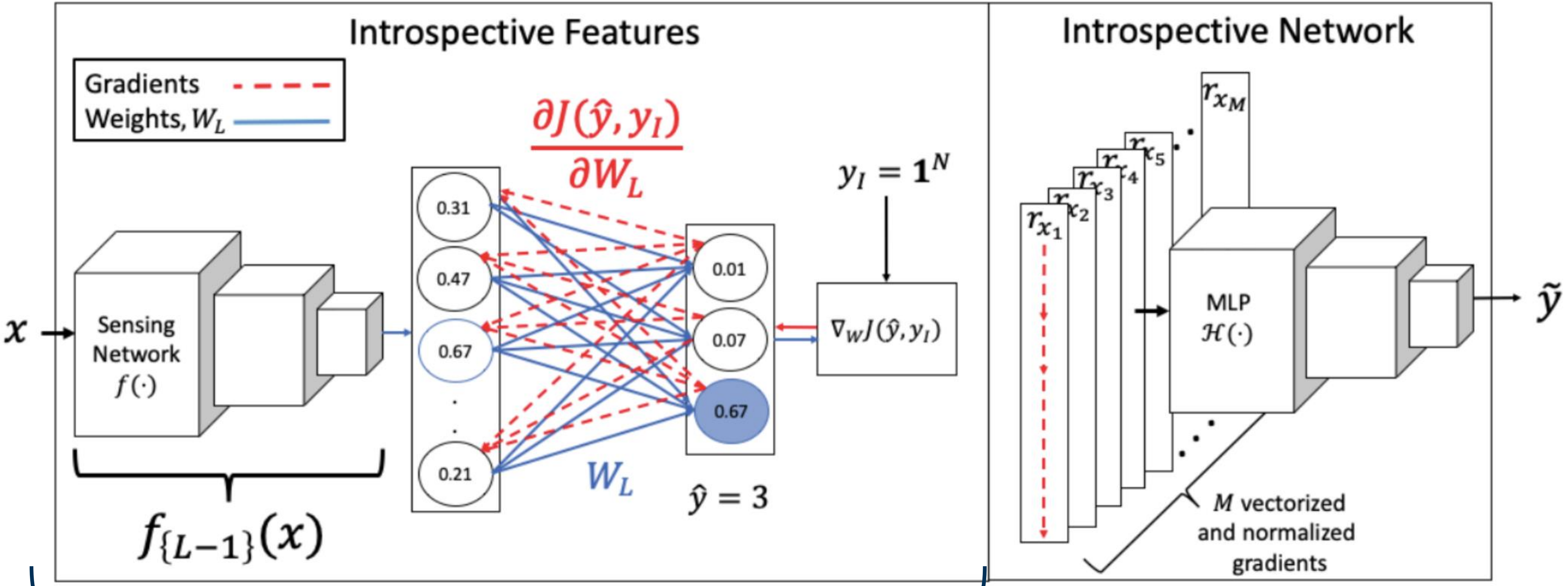
Normalized and vectorized gradients are introspective features

Vector of all ones: A confounding label!

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

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.

SCAN ME

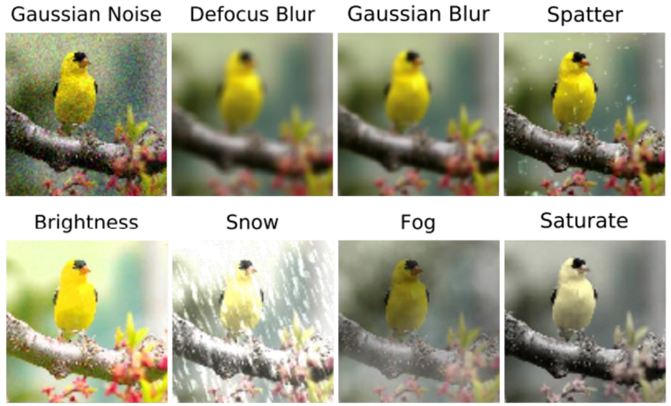
Introspective Learning: A Two-stage Approach for Inference in Neural Networks

Introspection provides robustness when the train and test distributions are different

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

Generalizable: Increased accuracy on OOD data

Calibrated: Reduces the difference between prediction accuracy and confidence



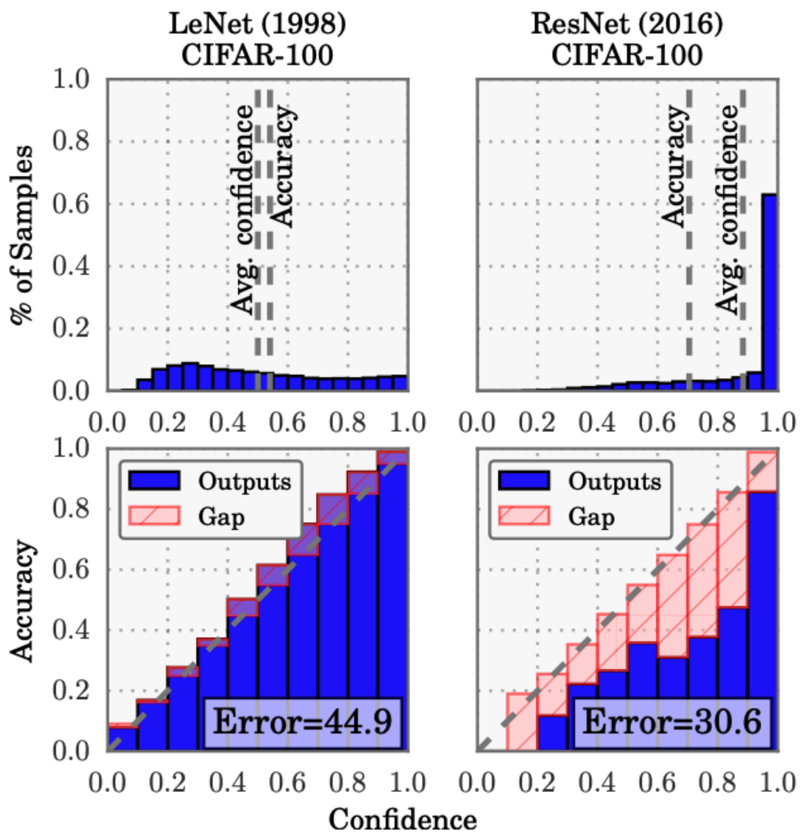
M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

A note on Calibration..

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

SCAN ME

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



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

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

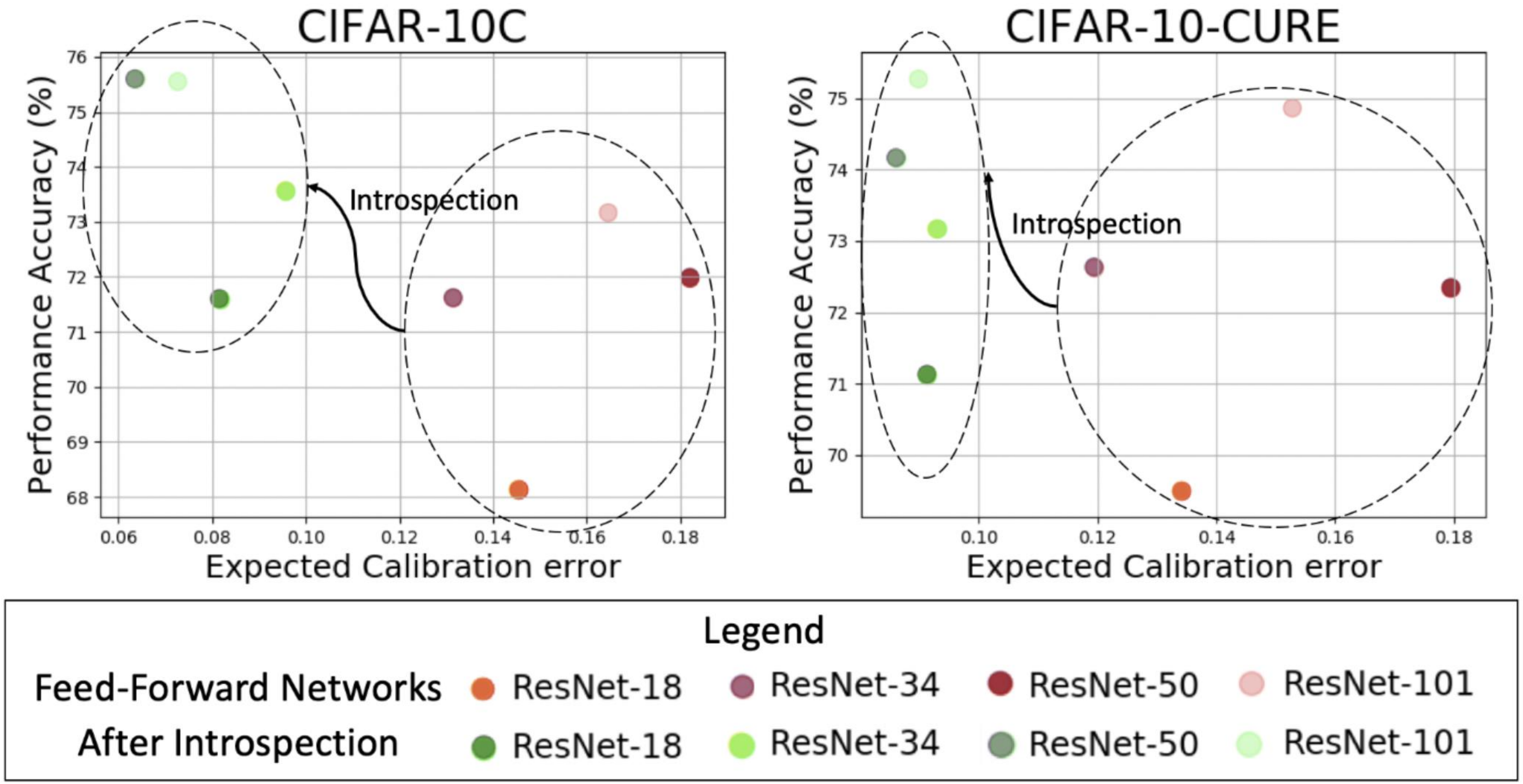
Introspection in Neural Networks
Generalization and Calibration results

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

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.

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.

Introspection in Neural Networks

Plug-in nature of Introspection

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**

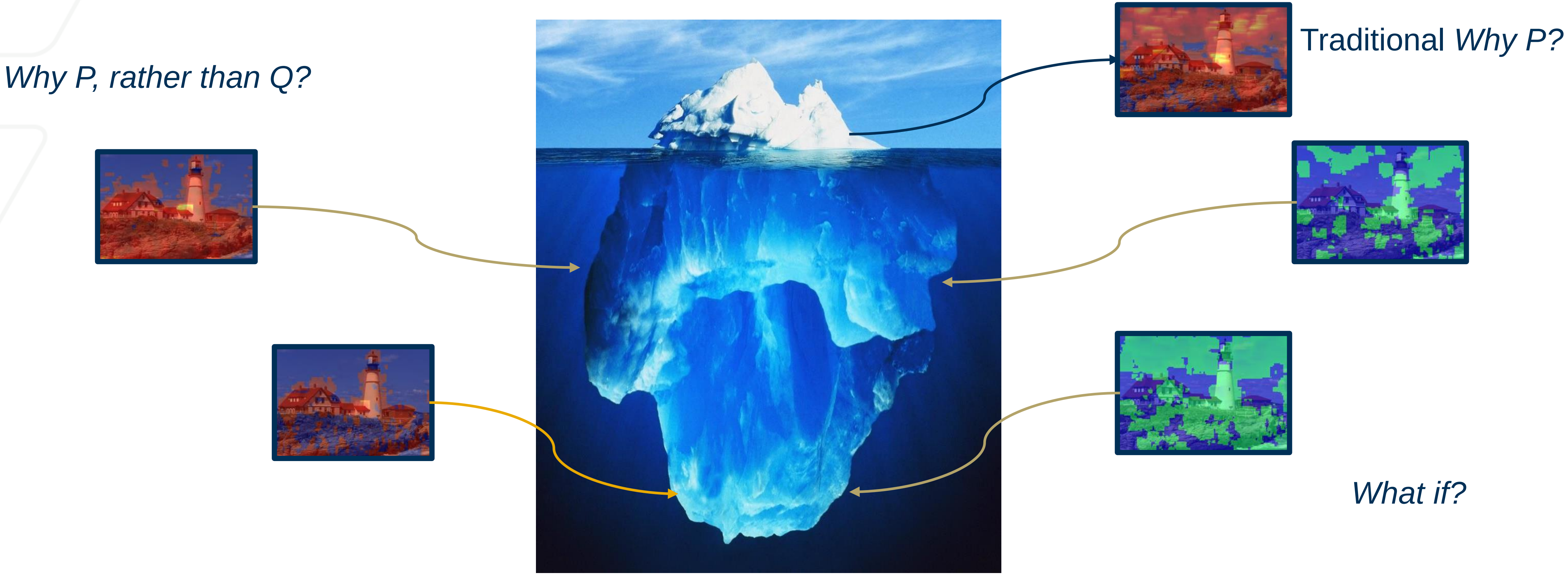
Table 3: Out-of-distribution Detection of existing techniques compared between feed-forward and introspective networks.

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.

Implicit Knowledge in Neural Networks – Inferential Machine Learning

Trained Neural Networks have a wealth of implicit stored knowledge. Inferential Machine Learning aims to ‘transmute’ this knowledge for other tasks



Inferential Machine Learning

Part 3: Uncertainty and Intervenability at Inference

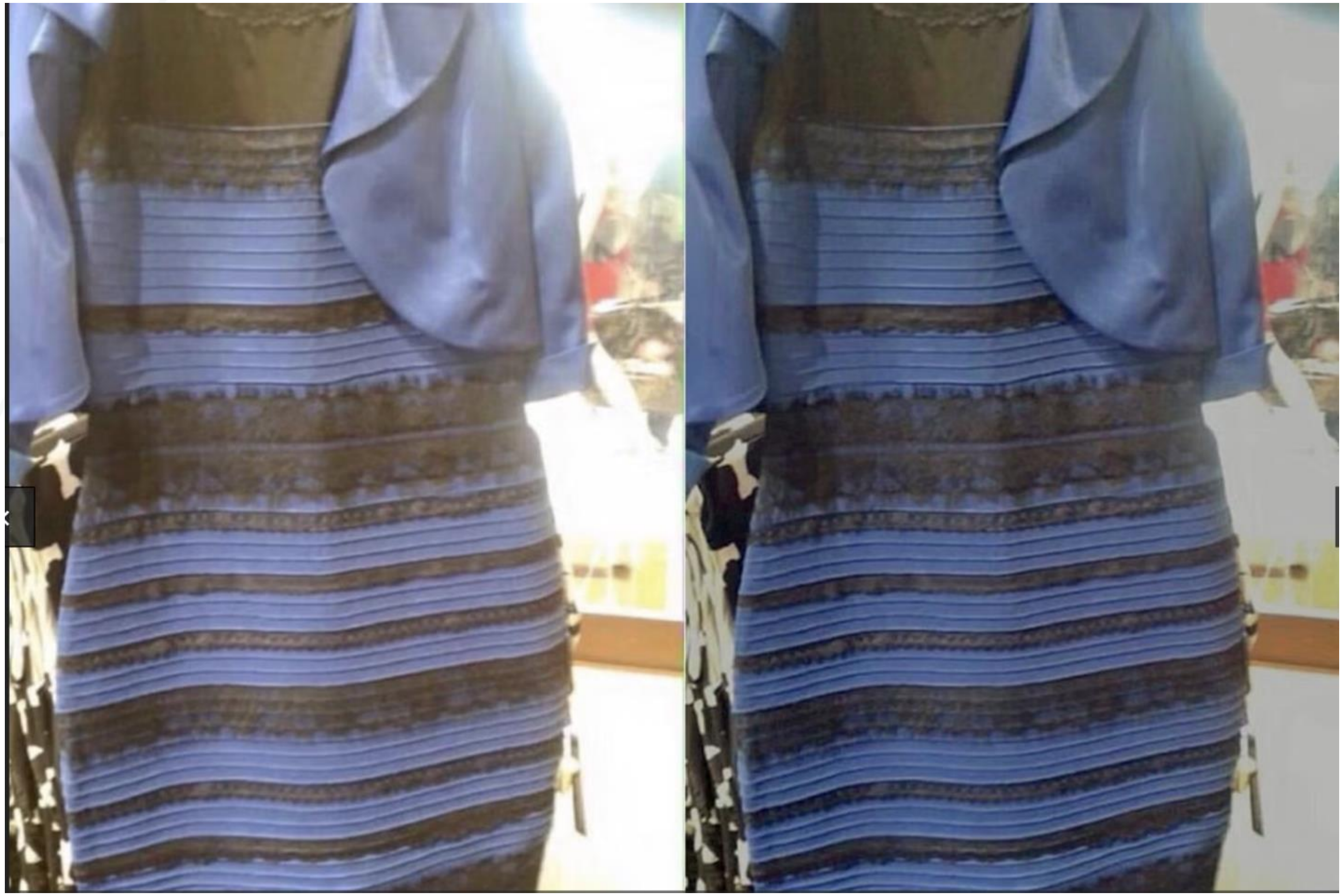
Objective

Objective of the Tutorial

To discuss methodologies that promote robust and fair inference in neural networks

- Part 1: Inference in Neural Networks
- Part 2: Explainability at Inference
- **Part 3: Uncertainty and Intervenability at Inference**
  - Uncertainty Basics
  - Uncertainty Quantification (UQ) in Classification
  - UQ Methods
  - Case Study 1: Gradient-based UQ
  - Case Study 2: Uncertainty in Explainability
  - Inferential Machine Learning
- Part 4: Interventions at Inference
- Part 5: Conclusions and Future Directions

Uncertainty is a model knowing that it does not know



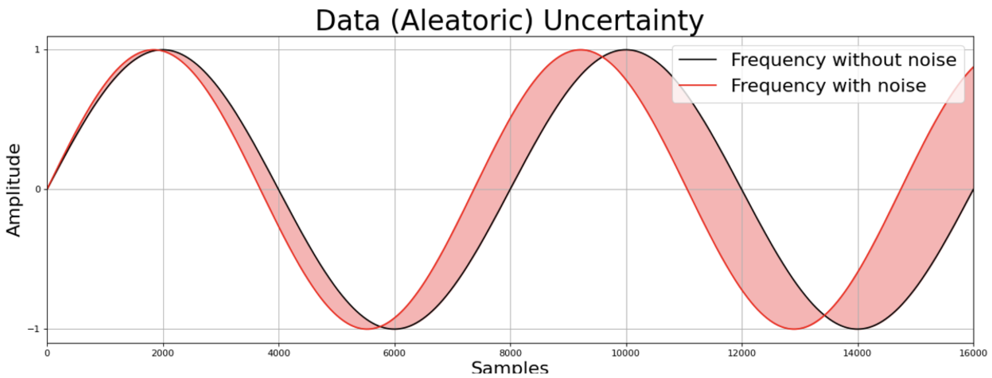
White and Gold
Or
Blue and Black?

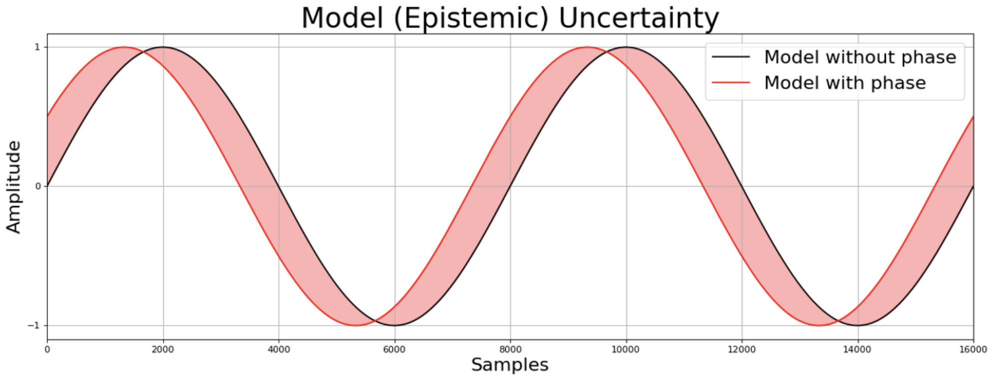
http://krasserm.github.io/2020/09/25/reliable-uncertainty-estimates/

Uncertainty

What is Uncertainty?

Uncertainty is a model knowing that it does not know





A slightly more complex example:

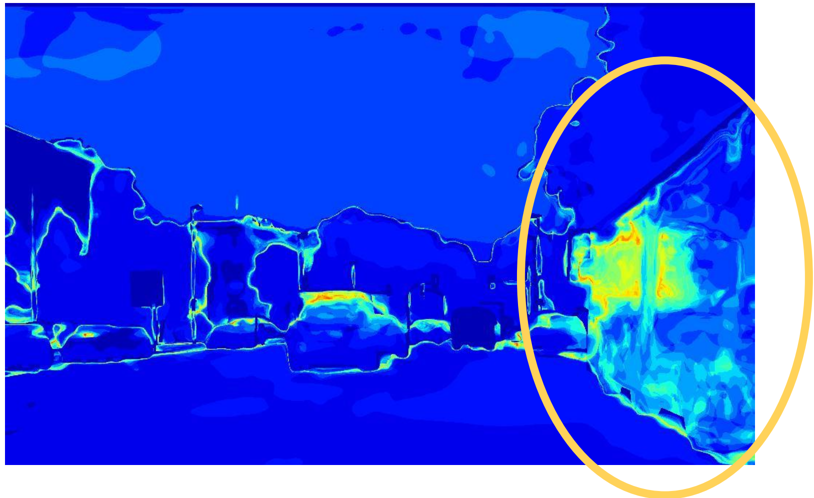
- **Data (Aleatoric) Uncertainty**: When there is inherent noise in available data or in measurement of data
- **Model (Epistemic) Uncertainty:** When our chosen model (network) is incapable of modeling the data

Uncertainty is a model knowing that it does not know

Input Image

Neural Network Output

Uncertainty Heatmap



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

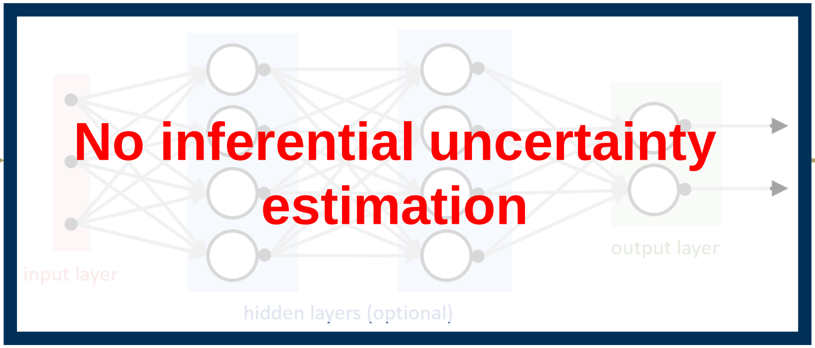
In classification, Uncertainty Quantification (UQ) implies providing a classification label and its associated uncertainty

Identify STOP as the only sign with bottom-left corner

Consider a network trained on 14 signs from CURE-TSR

Class: Stop Sign
Confidence: 98%
Uncertainty: 0.1%

Network has not seen GO sign but is shown at inference



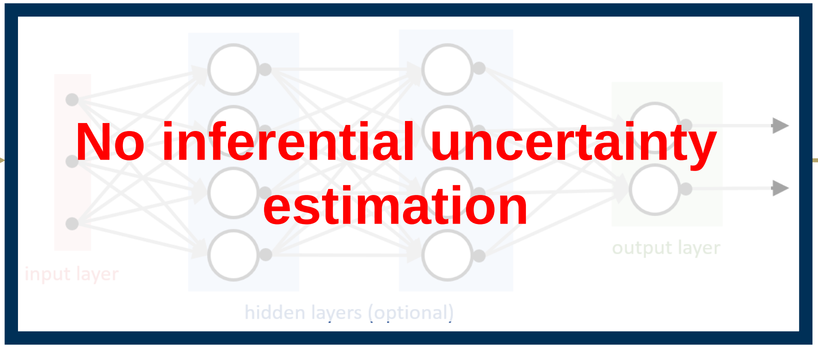
Class: Stop Sign
Confidence: 98%
Uncertainty: 0.1%

Uncertainty

Uncertainty Basics

In classification, Uncertainty Quantification (UQ) implies providing a classification label and its associated uncertainty

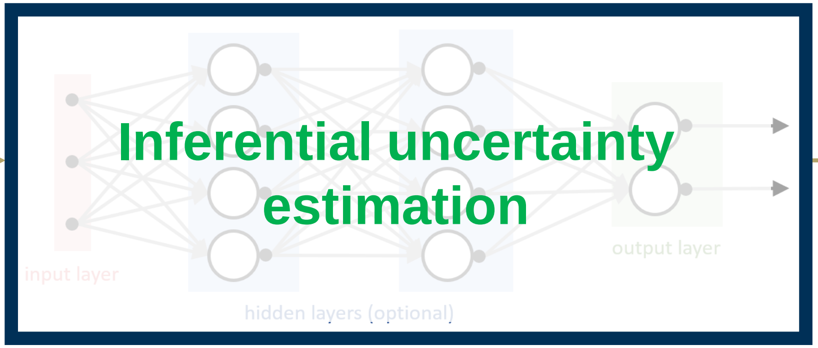
Network has not seen GO sign but is shown at inference



Class: Stop Sign
Confidence: 98%
Uncertainty: 0.1%

Identify that the letters and color are different

Network has not seen GO sign but is shown at inference



Class: Stop Sign
Confidence: 98%
Uncertainty: 98%

Probability vs Confidence vs Likelihood vs Uncertainty vs Calibration

- **Probability**: Transform logits (final layer outputs) between 0 and 1, Ex: Softmax probability. The input has some probability of belonging to all the trained classes
- **Confidence**: In non-conformal settings, confidence is a point estimate, Ex: the argmax of probabilities of softmax confidences. In the conformal setting (which we do not cover in this tutorial), confidence is an interval
- **Likelihood**: In Bayesian settings, likelihood refers to how likely the model fits the data or the 'goodness-of-fit' of the model. It is related to probability via bayes theorem
- **Uncertainty**: A probability distribution, (ideally) formed from feature outputs that showcase 'non-goodness' of fit of the underlying model or 'non-goodness' of training distribution compared to test distribution
- **Calibration**: A dataset estimate that shows the disparity between confidence of all point estimates in the dataset against their accuracy

Uncertainty

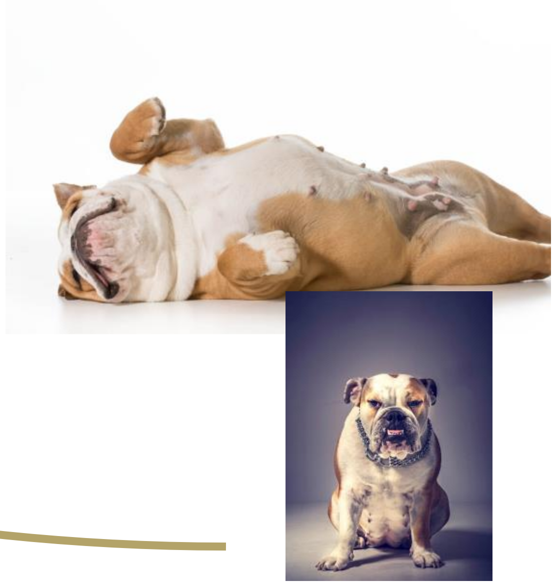
Challenge in Uncertainty Quantification

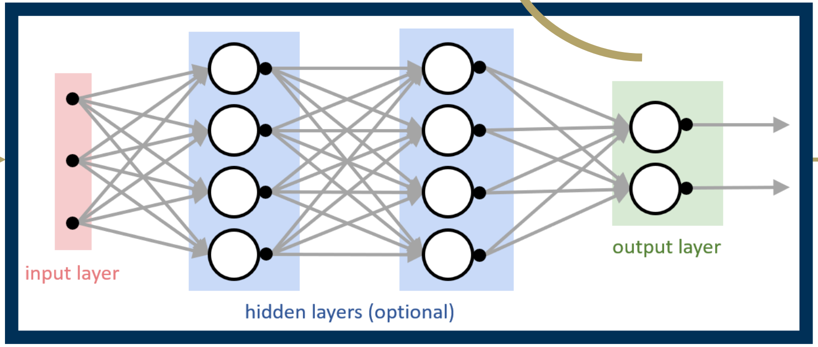
Primary purpose of neural networks (ex: classification) and Uncertainty Quantification do not always go hand-in-hand!

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

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

Dog asking for belly rub = Angry dog!





Dog

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

Simple Uncertainty Quantification 1: Negative Log Likelihood

In Bayesian settings, uncertainty is treated as inverse likelihood; consequently, lower the negative of likelihood, lower the uncertainty

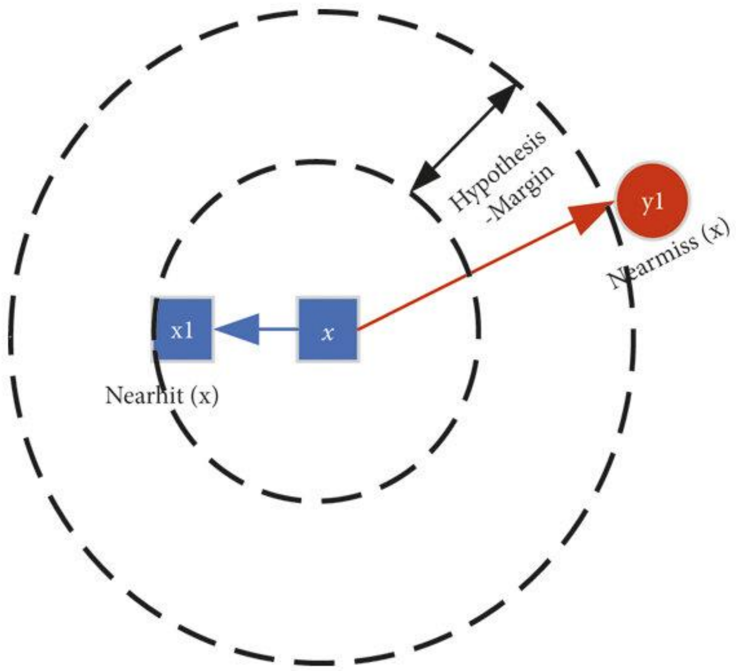
- Recall that ‘*In Bayesian settings,* ***likelihood*** *refers to how likely the model fits the data or the ‘goodness-of-fit’ of the model*’
- **Central Thesis**: Negative log-likelihood measures the ‘fit’ of a model by looking at all output logits
- **Cons: Requires ground truth at inference to measure likelihood**. Generally substituted with the prediction

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

Uncertainty

Simple Uncertainty Quantification 2: Hypothesis Margin

Difference between probability (or logits) of the predicted class and next most-likely class[1]



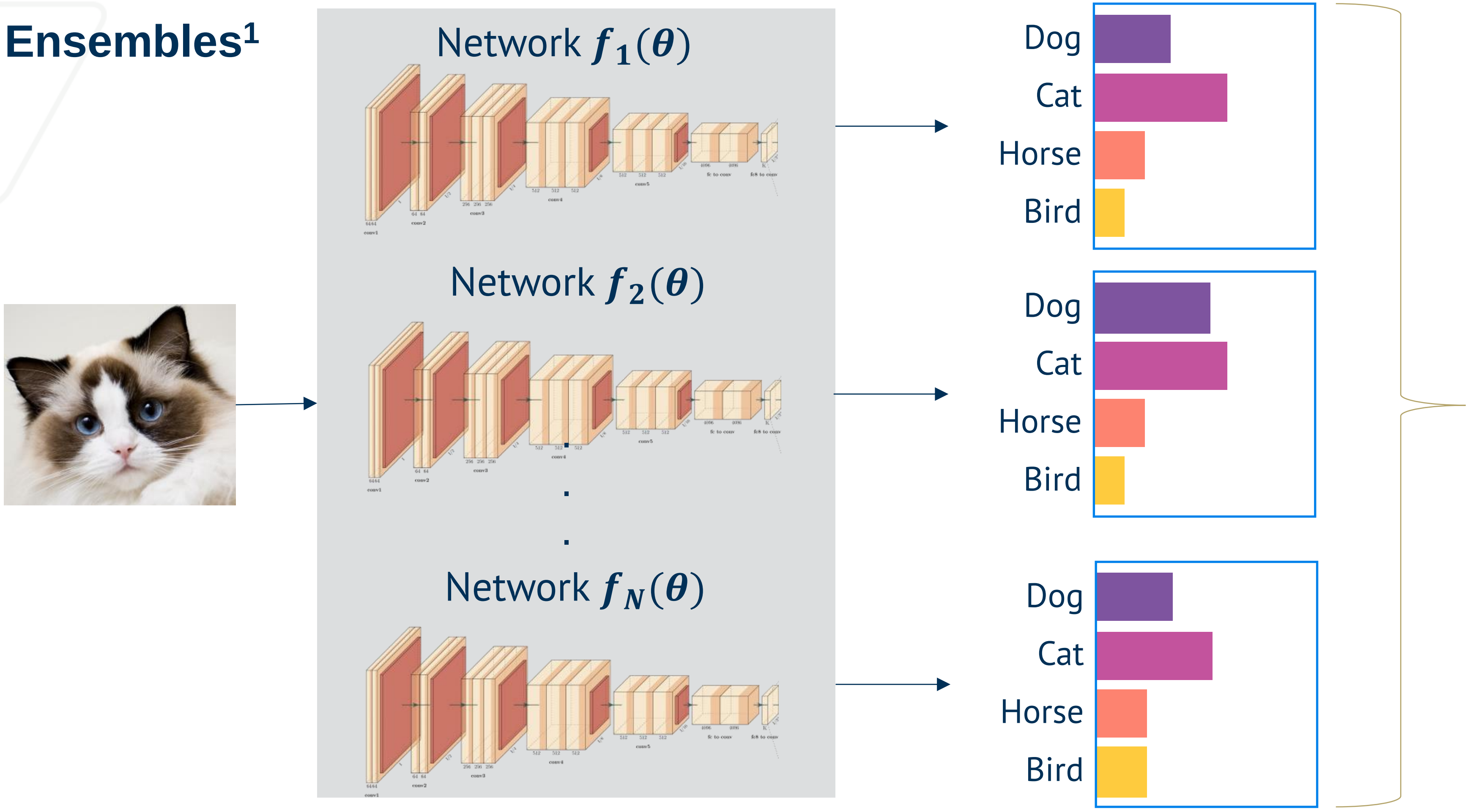
Simple => No changes in network architecture while training

- Commonly used to **rank the difficulty** of unlabeled samples in Active Learning
- **Central thesis**: During training, networks implicitly learn the difference between classes and find features that maximize the difference (similar to contrastive explanations)
- **Pros: No need for ground truth at inference**
- **Cons: Requires a complex network** that can learn implicit differences

Fig. from Tian, Yanjia, and Xiang Feng. "Large Margin Graph Embedding-Based Discriminant Dimensionality Reduction." *Scientific Programming* 2021.1 (2021): 2934362.

[1] Bartlett, Peter, et al. "Boosting the margin: A new explanation for the effectiveness of voting methods." *The annals of statistics* 26.5 (1998): 1651-1686.

Via Ensembles[1]



Variation within outputs is the uncertainty.

Commonly referred to as **Prediction Uncertainty.**

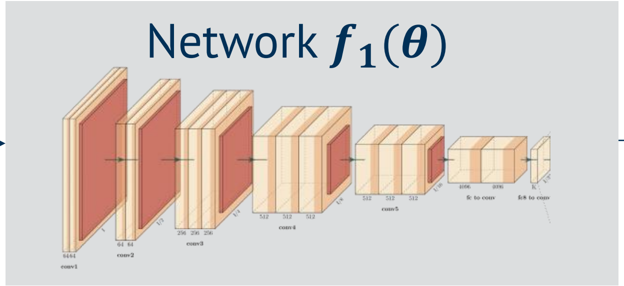
Requires multiple trained models – not exactly an inferential method

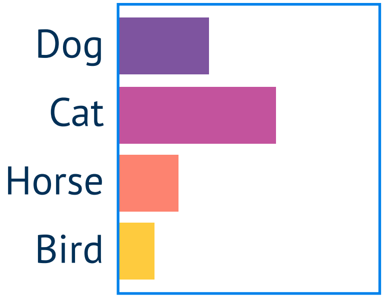
[1] Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles." *Advances in neural information processing systems* 30 (2017).

Uncertainty

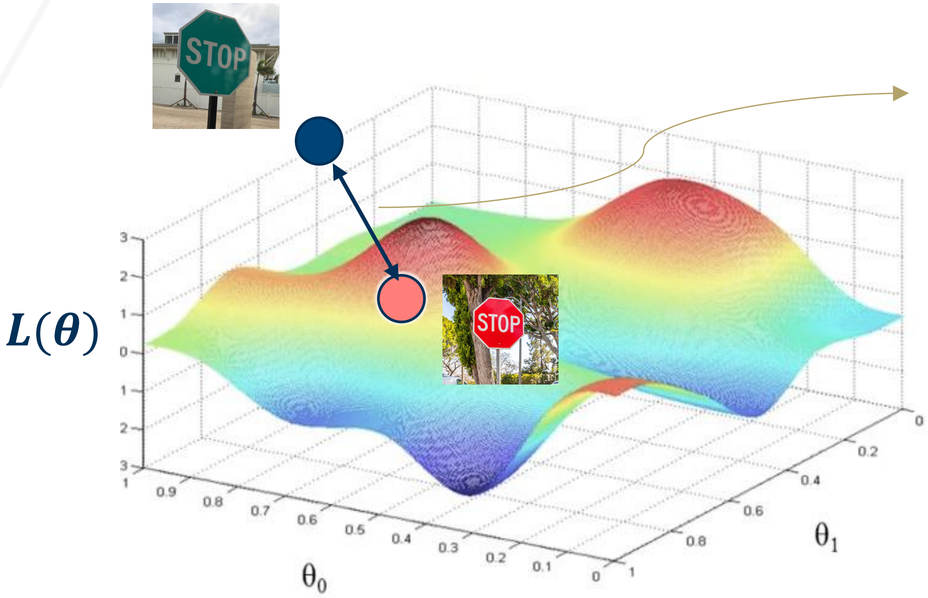
Uncertainty Quantification in Neural Networks

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

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

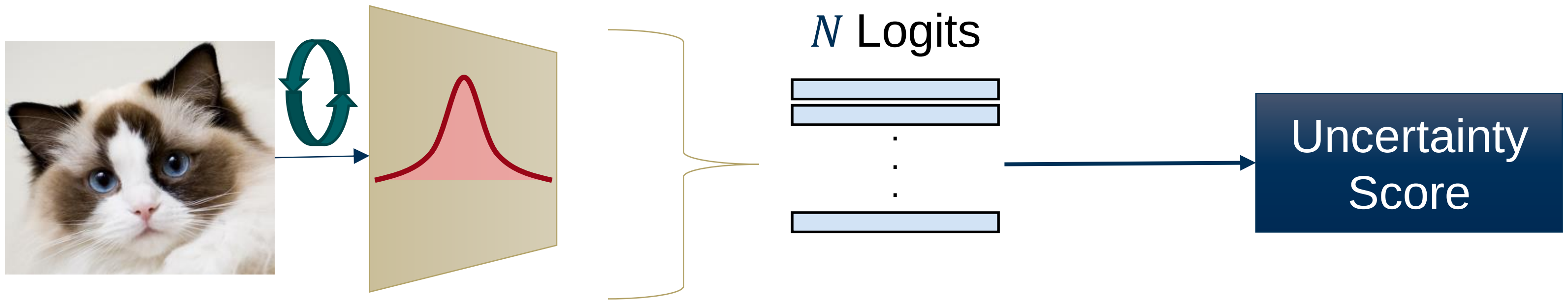
Via Monte-Carlo Dropout[1]: During inference repeated evaluations with the same input give different results

Different forward passes with dropout simulate $f_1(\cdot), f_2(\cdot), f_3(\cdot)$.

Challenge: intractable denominator

$$p(\boldsymbol{W}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{W})p(\boldsymbol{W})}{\int p(\boldsymbol{x}|\boldsymbol{W})p(\boldsymbol{W})d\boldsymbol{W}}$$

N forward passes



$$q\,(\boldsymbol{W}_N) \approx p(\boldsymbol{W}_N|\boldsymbol{x})$$

Final prediction is the mean of the outputs

Variation or entropy of logits is the uncertainty

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

Uncertainty

Iterative Uncertainty Quantification

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

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

Entropy of expectation of predictions

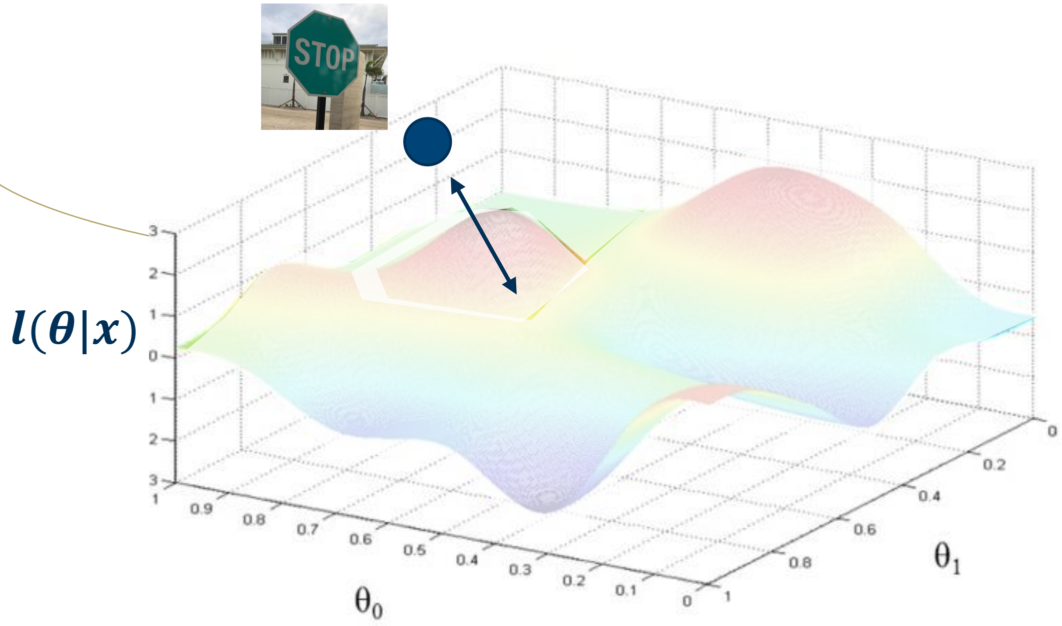
Expectation of individual entropy of predictions

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

Gradients as Single pass Uncertainty Quantification

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

Distance from unknown cluster



Method:

Extracting Gradient Information!

Uncertainty is a 'catch-all' term, used in multiple applications

- Explainability
- Out-of-distribution Detection
- Adversarial Detection
- Anomaly Detection
- Corruption Detection
- Misprediction Detection
- Causal Analysis
- Open-set Recognition
- Noise Robustness
- Uncertainty Visualization
- Image Quality Assessment
- Saliency Detection

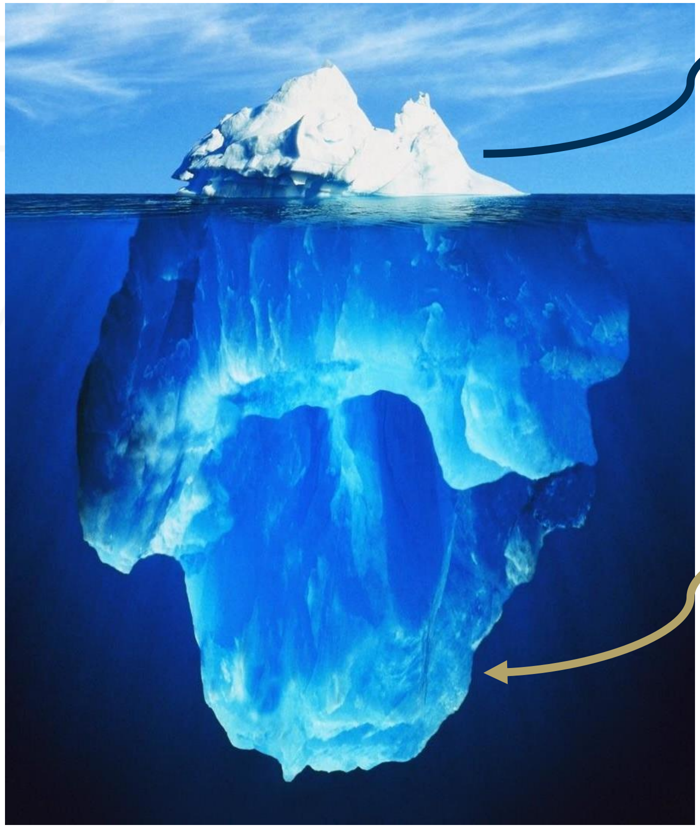
Applications relevant during model inference

Relevant at Deployment:

Provide a specific 'uncertainty measure' that objectively allows users to trust neural network predictions

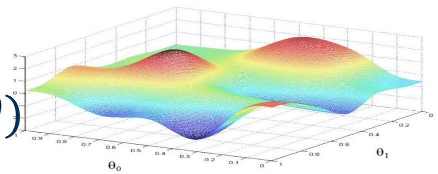
Unfortunately, each application has its own uncertainty quantification

Uncertainty is a 'catch-all' term, used in multiple applications



Learned Knowledge

$L(\theta)$



Transmuted Knowledge

Part 2

Case Study 1

Case Study 3

Case Study 2

- Explainability
- Out-of-distribution Detection
- Adversarial Detection
- Anomaly Detection
- Corruption Detection
- Misprediction Detection
- Causal Analysis
- Open-set Recognition
- Noise Robustness
- Uncertainty Visualization
- Image Quality Assessment
- Saliency Detection

Case Study 1:

Counterfactual Gradients-based Quantification of Prediction Trust in Neural Networks

Mohit Prabhushankar, PhD
Postdoc

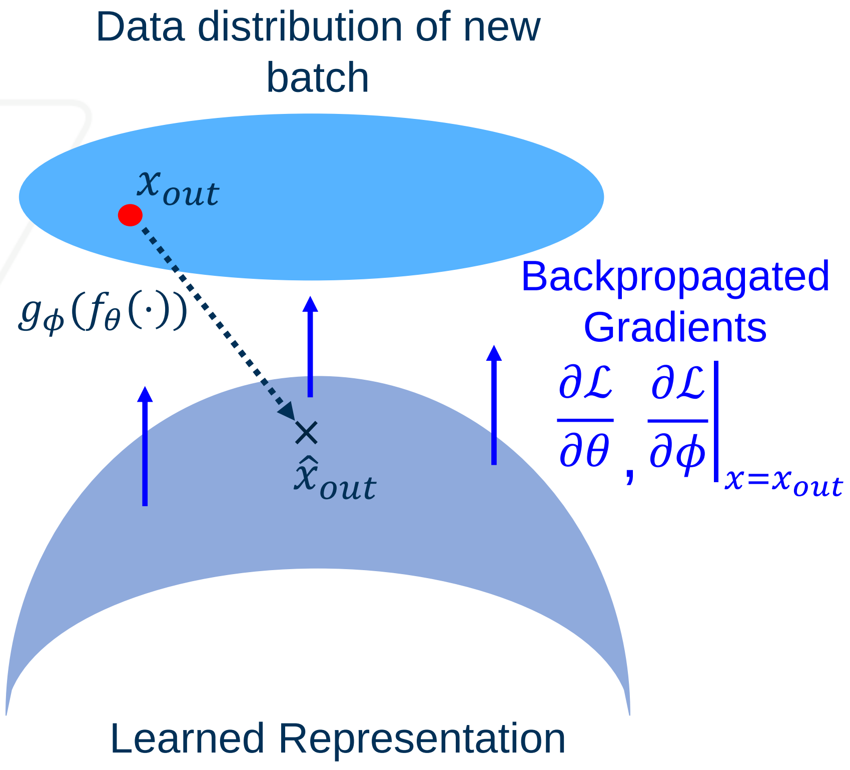
Ghassan AlRegib, PhD
Professor

SCAN ME

Case Study 1: Misprediction Detection

Principle

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



During training, a loss function $\mathcal{L}$ is used to quantify this measure.

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

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

Case Study 1: Misprediction Detection

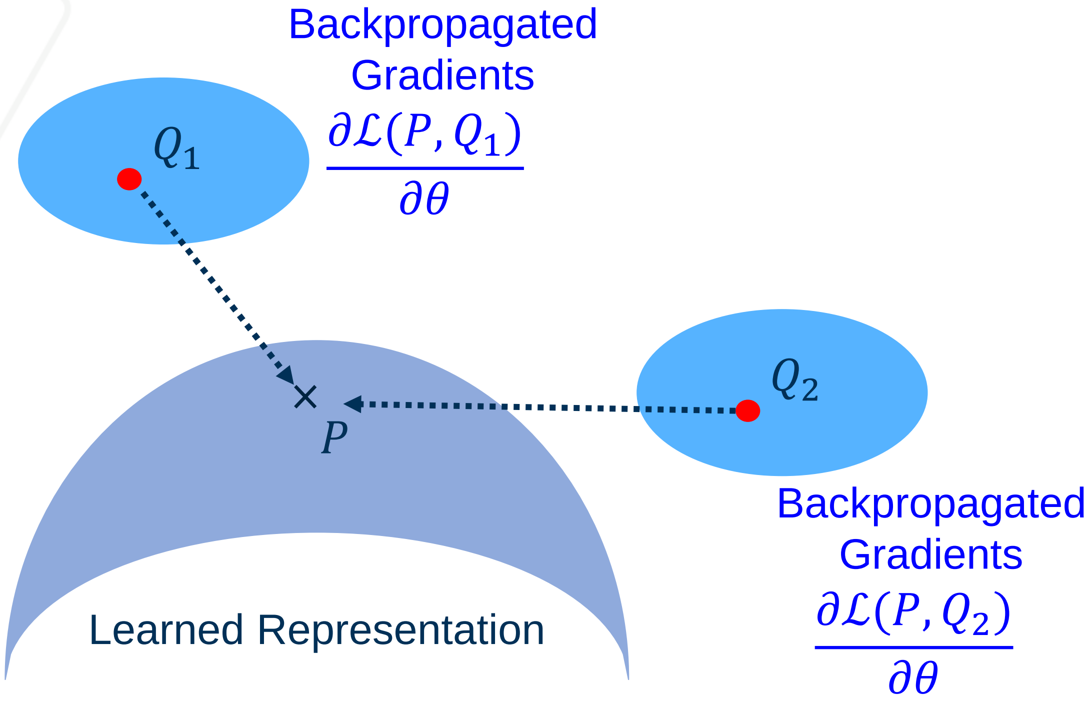
Principle

SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

Principle: Gradients provide an uncertainty measure between the learned representations space and novel data

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



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

- **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

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

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

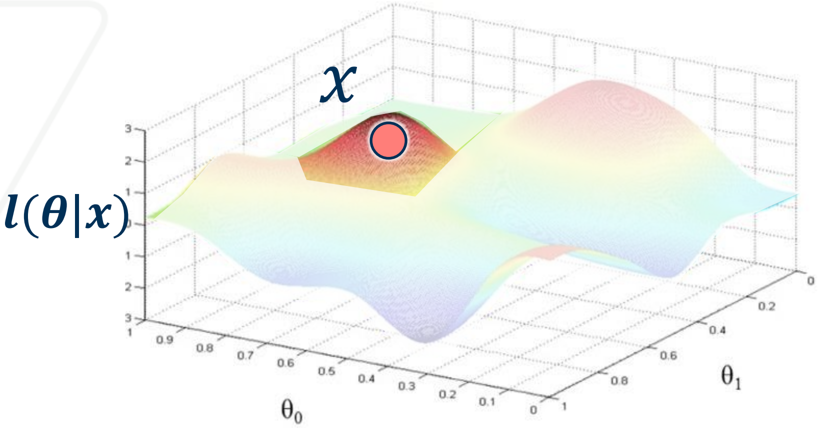
OLIVES @GeorgiaTech

Georgia Tech

Toy Manifold Example

Why uncertainty?

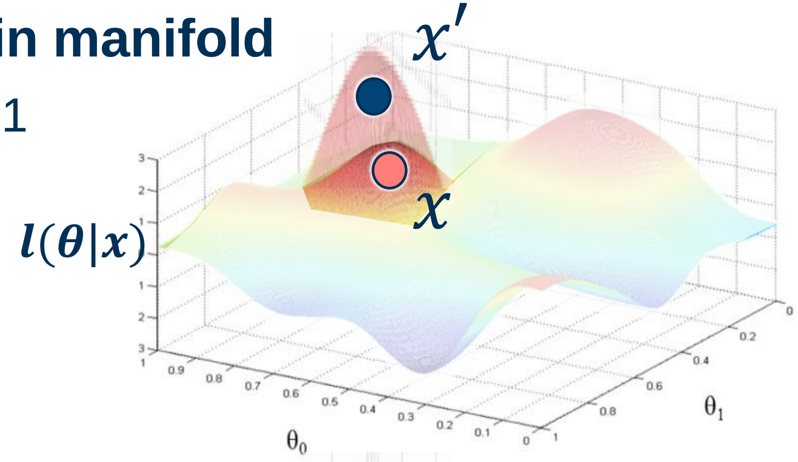
Gradients represent the local required change in manifold

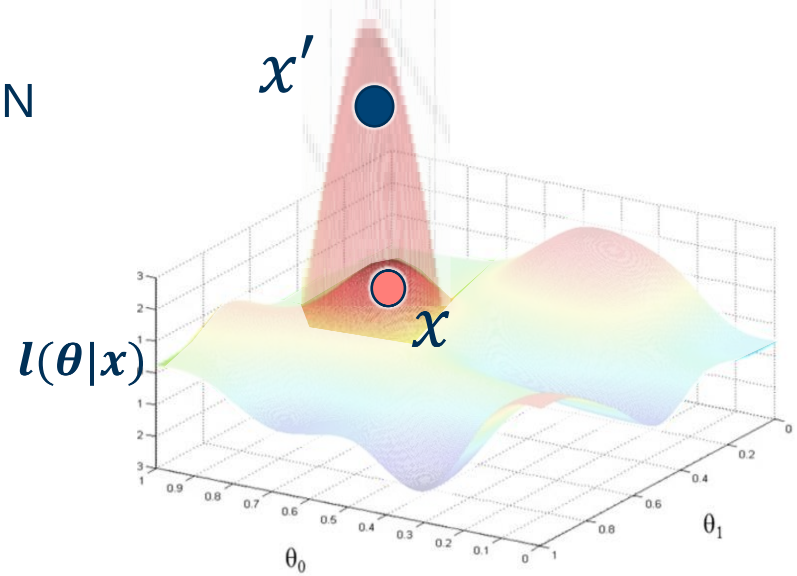


Contrast class 1

. . .

Contrast class N





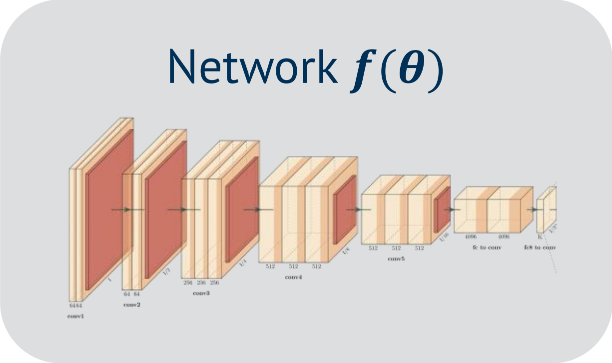
- Gradients provide the necessary change in manifold that would predict the novel data ‘correctly’.
- Correctly means contrastively (or incorrectly)!
- Less data in the new region, higher is the fisher information and uncertainty

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

Case Study 1: Misprediction Detection

Intuition for counterfactual gradients-based Trust

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



Spoonbill

Physical meaning of 'incorrect' class backpropagation

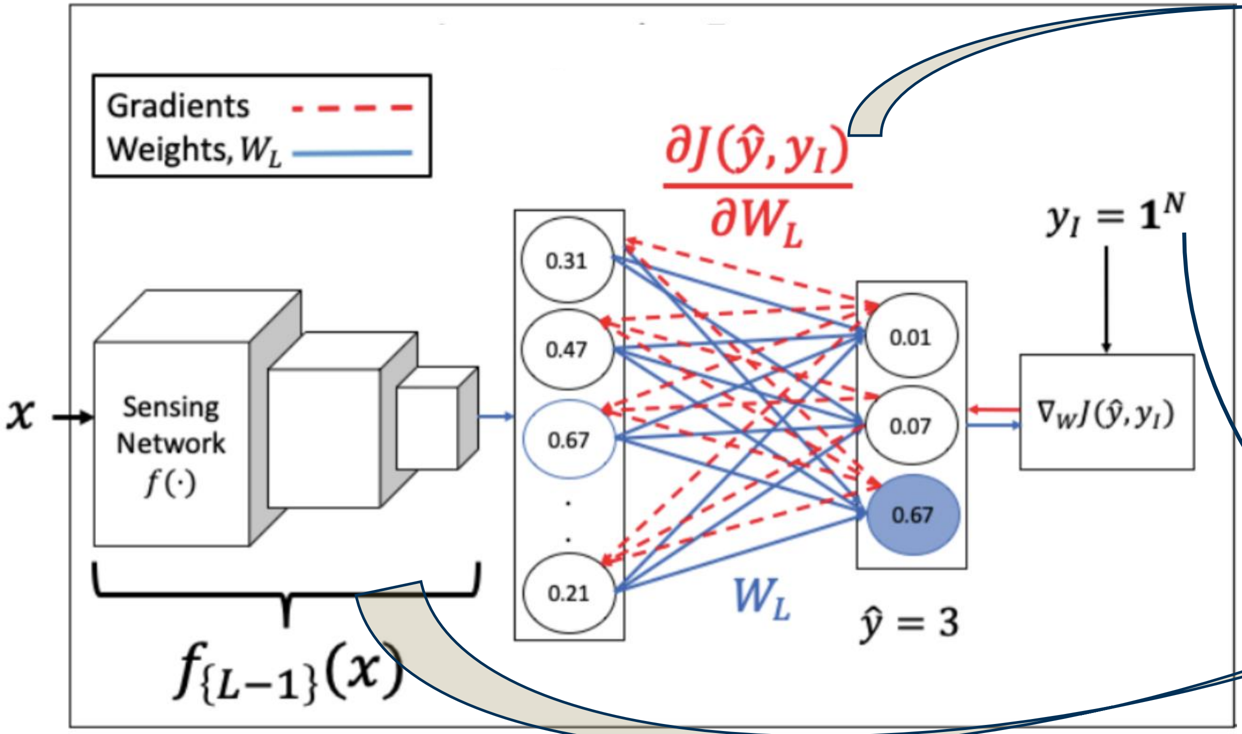
Larger the required change, larger the trust placed in the prediction 'Spoonbill'

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

Case Study 1: Misprediction Detection

Deriving Gradient Features

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.

Case Study 1: Misprediction Detection

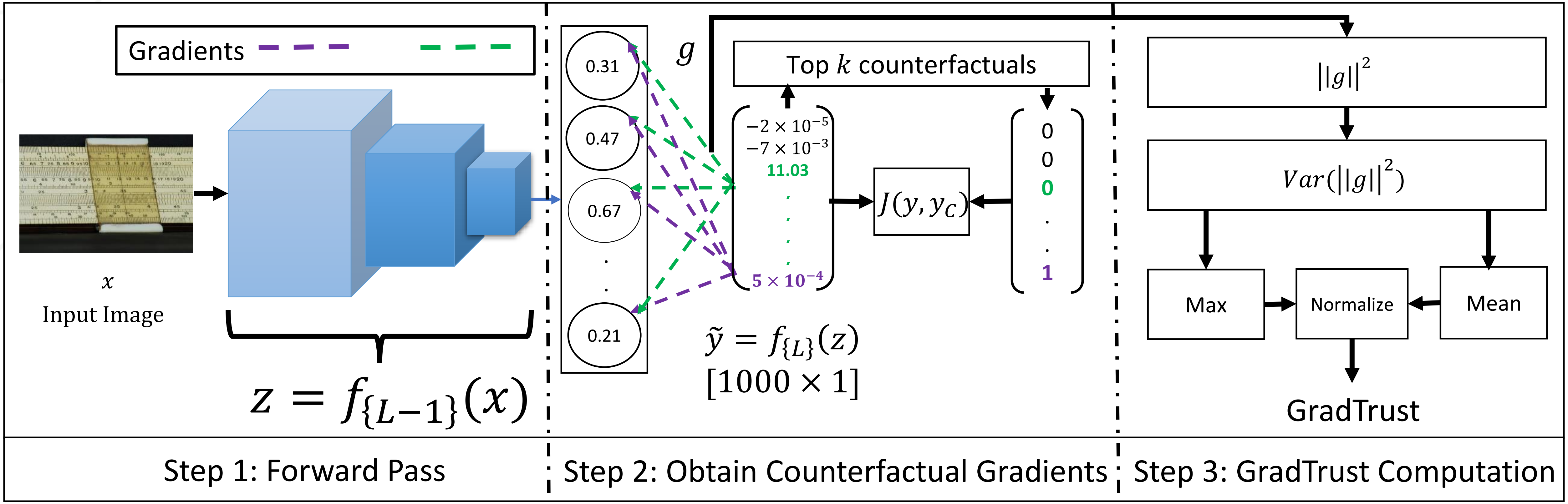
Intuition for gradients-based Trust

Step 2: Quantify the variance of network parameters (of the last layer) when backpropagating contrast classes

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

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

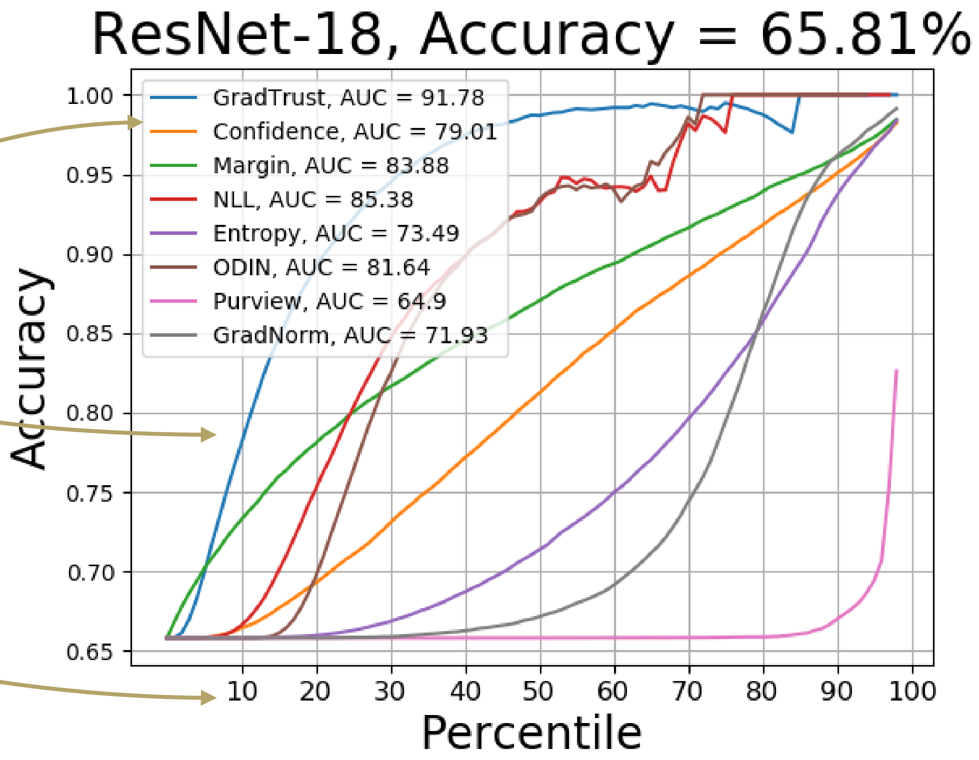
How do we measure required change? Quantify the variance of network parameters when backpropagating counterfactual classes



Evaluation
Methodology

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

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



Evaluation

Quantitative Results for Image Classification

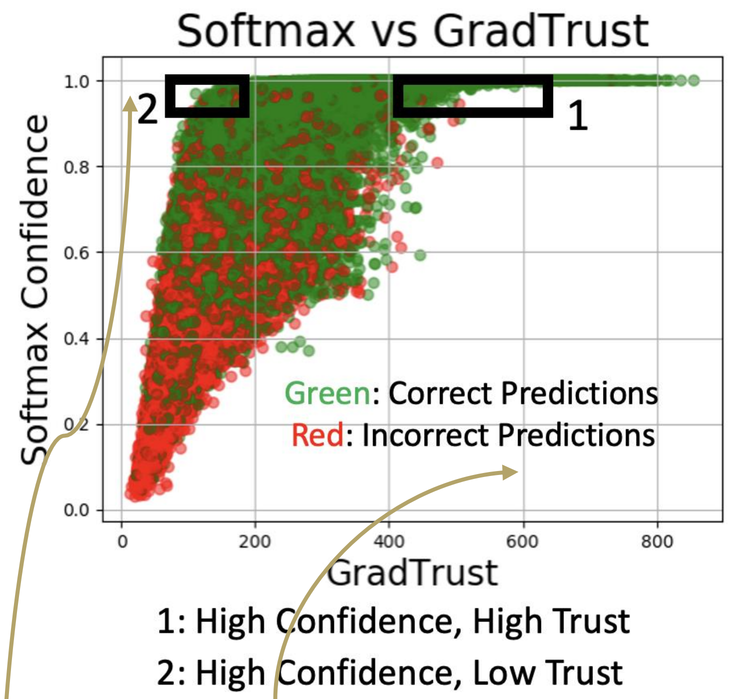
GradTrust is in Top 2 performing metrics in all but 1 network

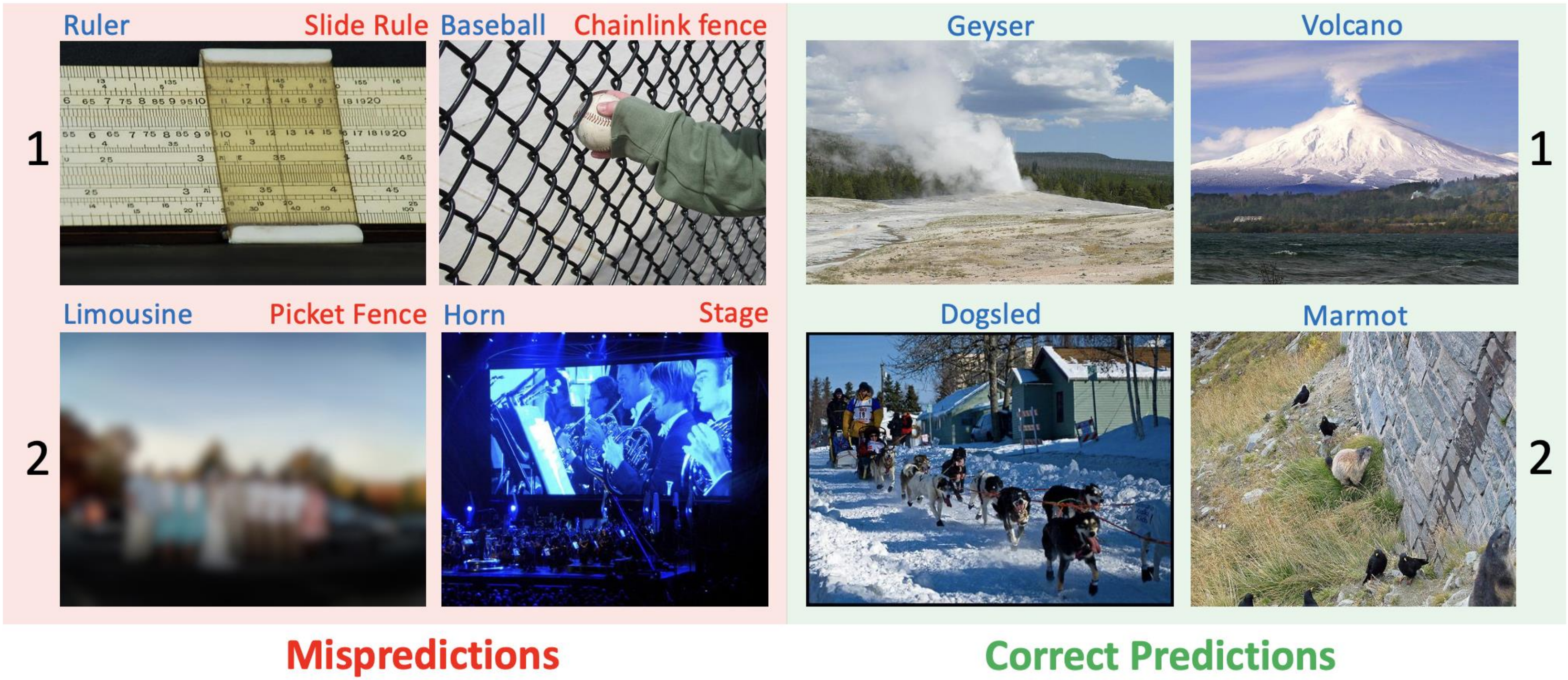
	AUAC / AUFC								
Architecture	Softmax	Entropy	NLL	Margin [27]	ODIN [28]	MCD [12]	GradNorm [5]	Purview [4]	GradTrust
AlexNet [29]	72.86/68.43	65.02/62.14	**83.21/79.37**	79.04/73.3	79.22/75.89	54.2/51.59	58.85/55.28	50.14/48.92	**92.09/89.5**
MobileNet [30]	77.91/74.96	71.72/69.9	**84.02/81.37**	83.13/79.1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

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

Evaluation

Qualitative Results for Image Classification





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

Evaluation

Qualitative Results for Image Classification

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

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

Evaluation

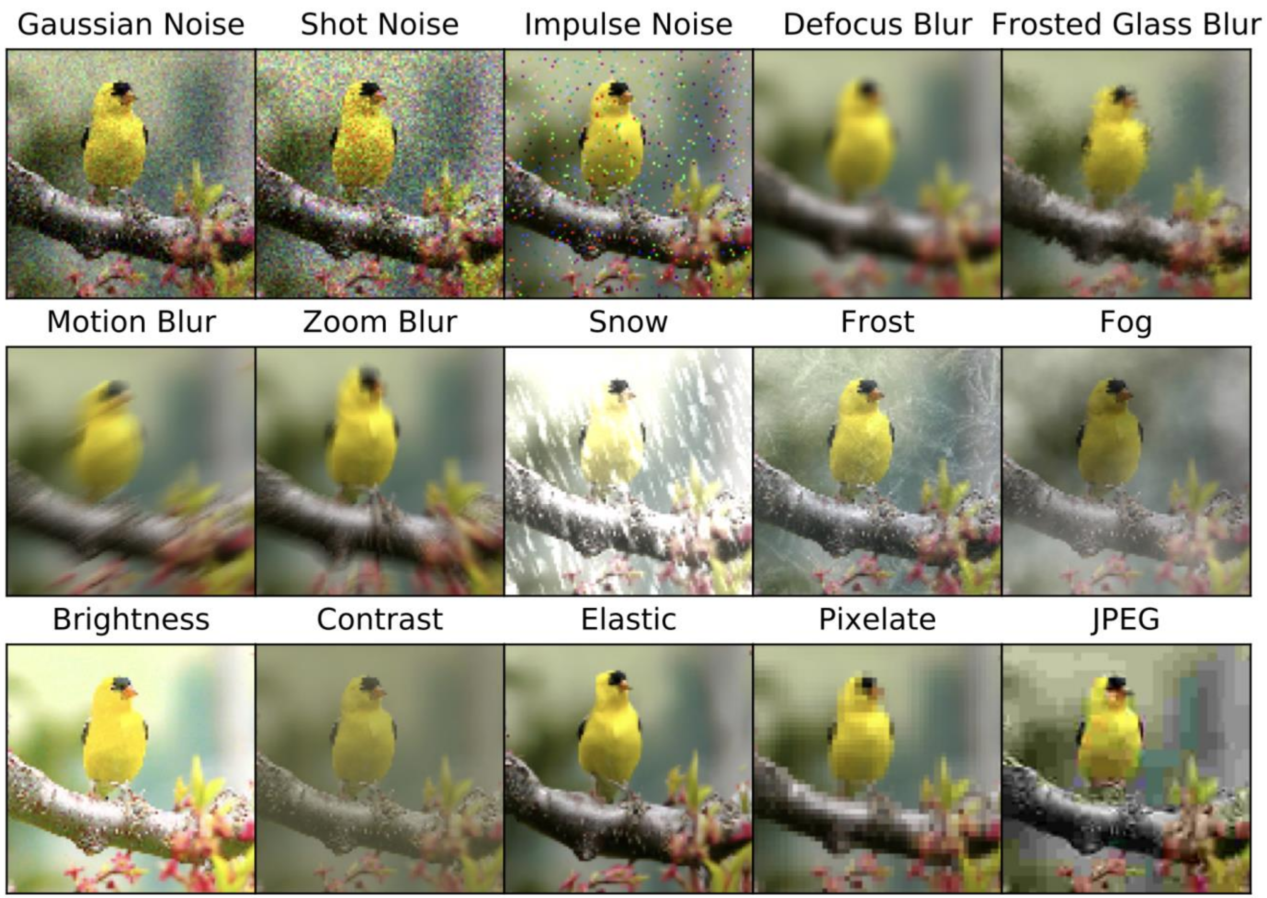
Qualitative Results for Image Classification under Corruption

SCAN ME

Same evaluation setup as before, with inputs being corrupted by noise

Data Characteristics:

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



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

Evaluation

Qualitative Results for Image Classification under Corruption

GradTrust is the Top performing metric in all but two setups (in red)

AUAC for MSP / NLL / Margin / ODIN / **GradTrust**

Level	Brightness	Snow	Fog	Frost	Defocus Blur
1	80.36/85.72/85.1/82.5/**91.75**	69.44/78.13/75.49/74.47/**88.35**	73.62/78.13/79.66/66.86/**89.89**	73.97/77.93/79.87/77.56/**90.04**	73.41/78.56/79.44/67.96/**89.25**
2	79.52/85.41/84.5/81.25/**91.62**	52.48/62.7/58.67/55.37/**82.91**	69.97/76.65/76.32/63.63/**88.71**	63.56/70.72/70.32/59.69/**86.4**	69.98/76.37/76.41/65.76/**87.66**
3	78.32/84.45/83.51/76.76/**91.37**	54.35/66.66/60.09/51.92/**82.53**	63.07/73.9/69.63/59.1/**85.63**	54.05/63.19/60.08/56.15/**81.73**	62.96/67.12/69.64/58.12/**84.52**
4	76.26/81.76/81.86/73.55/**90.81**	44.38/51.84/49.45/43.17/**77.13**	55.28/70.07/61.66/65.2/**80.45**	51.46/63.2/57.97/54.94/**80.61**	56.38/55.17/62.99/44.59/**79.66**
5	73.34/79.49/79.32/68.06/**89.81**	18.02/35.1/18.71/22.74/**40.09**	34.25/55.59/39.19/42.26/**63.68**	44.42/52.69/50.43/44.46/**76.76**	45.4/43.53/50.98/31.59/**72.26**
Level	**Glass Blur**	**Motion Blur**	**Zoom Blur**	**Contrast**	**Elastic Transform**
1	72.14/79.43/78.33/71.32/**89.41**	76.57/82.4/82.21/71.96/**90.73**	69.74/79.26/76.25/66.08/**88.55**	76.25/78.98/81.9/68.19/**90.44**	77.99/82.6/83.4/76.4/**91.11**
2	65.83/73.39/72.55/62.13/**87.17**	71.53/79.02/77.87/63.53/**88.58**	62.51/75.37/69.37/62.87/**85.84**	73.17/78.8/79.3/66.03/**89.47**	66.76/72.86/73.34/62.6/**86.8**
3	46.36/52.7/52.14/44.67/**77.74**	62.6/69.49/69.39/61.78/**84.2**	56.6/75.33/63.07/62.23/**83.35**	66.27/74.74/72.8/63.34/**86.39**	73.88/81.63/79.78/68.5/**89.38**
4	42.12/43.71/47.4/38.97/**74.65**	51.57/56.64/58.02/50.17/**76.15**	50.61/72.16/56.69/57.59/**80.46**	45.65/63.9/50.33/55.1/**72**	65.91/70.85/72.4/62.77/**85.75**
5	38.26/45.59/42.91/38.95/**67.47**	44.36/48.6/50.25/36.59/**64.47**	44.85/70.93/50.38/57.18/**77.35**	28.07/**39.05**/30.26/30.56/25.49	32.84/53.11/36.47/43.75/**65.95**
Level	**JPEG Compression**	**Pixelate**	**Gaussian Noise**	**Shot Noise**	**Impulse Noise**
1	76.2/78.96/81.7/67.99/**90.67**	76.18/79.23/81.65/78.09/**90.36**	71.38/78.02/77.42/76.54/**89.48**	69.49/80.14/75.57/79.93/**88.68**	62.43/72.55/68.64/59.08/**85.21**
2	74.5/78.07/80.25/78.13/**89.94**	76.16/79.97/81.7/80.79/**90.64**	64.03/71.02/70.28/58.82/**86.17**	60.17/72.03/66.28/62/**85.46**	52.87/67.81/58.25/61.6/**52.87**
3	73.12/79.59/79.09/69.9/**89.64**	66.02/75.91/72.48/67.55/**86.9**	47.57/61.95/52.71/51.33/**75.67**	45.47/63.62/50.55/55.54/**76.18**	42.23/55.17/46.42/47.92/**71.8**
4	68.4/77.46/74.86/67.72/**88.06**	55.44/66.16/61.74/51.81/**82.66**	22.74/51.28/25.16/39.85/**56.15**	21.23/35.34/23.61/26.87/**54.01**	16.82/44.52/18.05/43.63/**46.08**
5	60.38/75.37/66.91/71.8/**85.55**	52.45/66.11/58.4/52.56/**79.22**	5.8/25.39/6.31/20.17/**25.93**	9.71/41.42/10.69/37.7/**51.15**	3.86/**31.79**/4.05/26.57/27.11

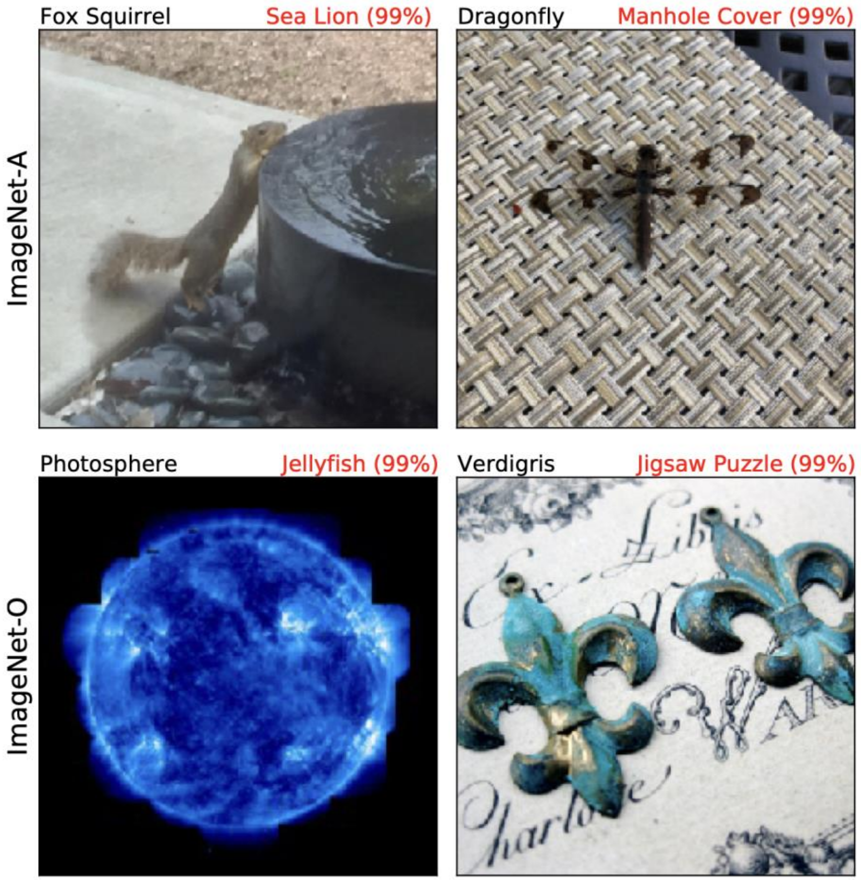
Evaluation

Qualitative Results for Image Classification under Natural Adversaries

OOD evaluation setup, with inputs being either natural adversaries or validation images

Data Characteristics:

- Curated set of 7500 natural adversarial images
  - ‘Natural’ly occurring images as opposed to artificially generated adversarial images
- Experimental setup similar to OOD detection; given a total of 15,000 images (7500 from ImageNet-A and 7500 randomly chosen from ImageNet validation set), we find AUDC (Area under Detection curve)



Evaluation

Qualitative Results for Image Classification under Natural Adversaries

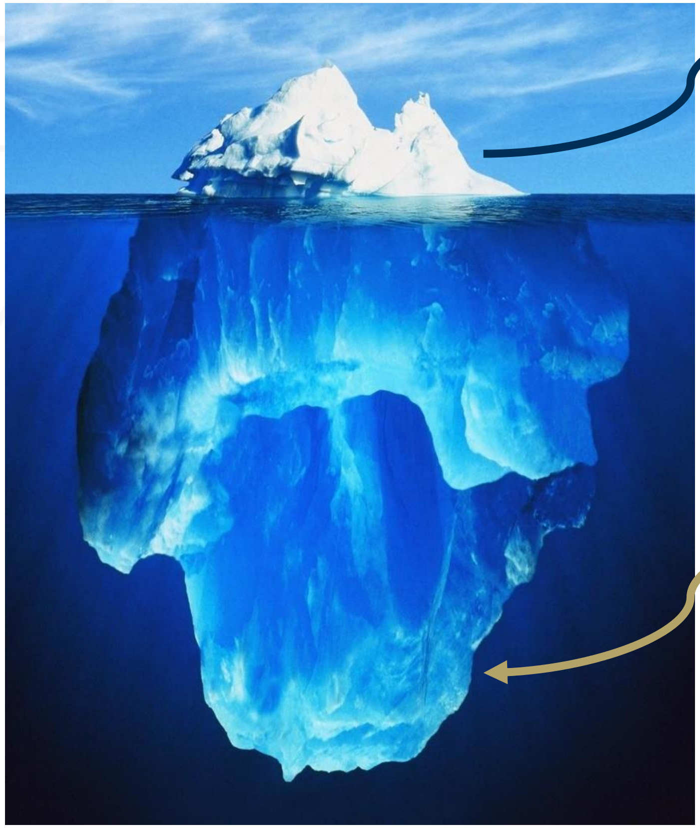
GradTrust is the top performing metric

Architecture	MSP [48]	NLL	Margin [8]	ODIN [49]	**GradTrust**
AlexNet [51]	55.9	76.24	62.68	70.43	**86.06**
MobileNet-v3 [52]	57.54	73.87	64.28	62.81	**85.9**
ResNet-18 [53]	57.56	75.22	64.01	70.54	**84.4**
VIT-b-32 [60]	61.96	58.18	67.03	40.11	**69.0**
ResNet-101 [53]	55.35	75.99	61.09	73.21	**82.12**
ResNeXt-32 [55]	54.26	78.98	59.73	77.14	**81.44**
VIT-b-16 [60]	59.75	50.44	64.84	31.32	**68.14**
ResNeXt-64 [55]	53.02	36.2	56.67	27.9	**67.53**
MaxVIT-t [62]	54.2	41.42	59.3	22.26	**70.55**

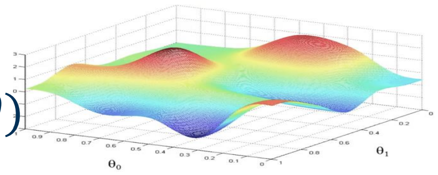
Uncertainty

Uncertainty and Inferential Machine Learning

Uncertainty is a 'catch-all' term, used in multiple applications



Learned Knowledge

$L(\theta)$ 

Transmuted Knowledge

Part 2

Case Study 1

Case Study 3

Case Study 2

- Explainability
- Out-of-distribution Detection
- Adversarial Detection
- Anomaly Detection
- Corruption Detection
- Misprediction Detection
- Causal Analysis
- Open-set Recognition
- Noise Robustness
- Uncertainty Visualization
- Image Quality Assessment
- Saliency Detection

Case Study 2:

VOICE: Variance of Induced Contrastive Explanations for Quantifying Uncertainty in Interpretability

Mohit Prabhushankar, PhD
Postdoc

Ghassan AlRegib, PhD
Professor

SCAN ME

Uncertainty in Explainability

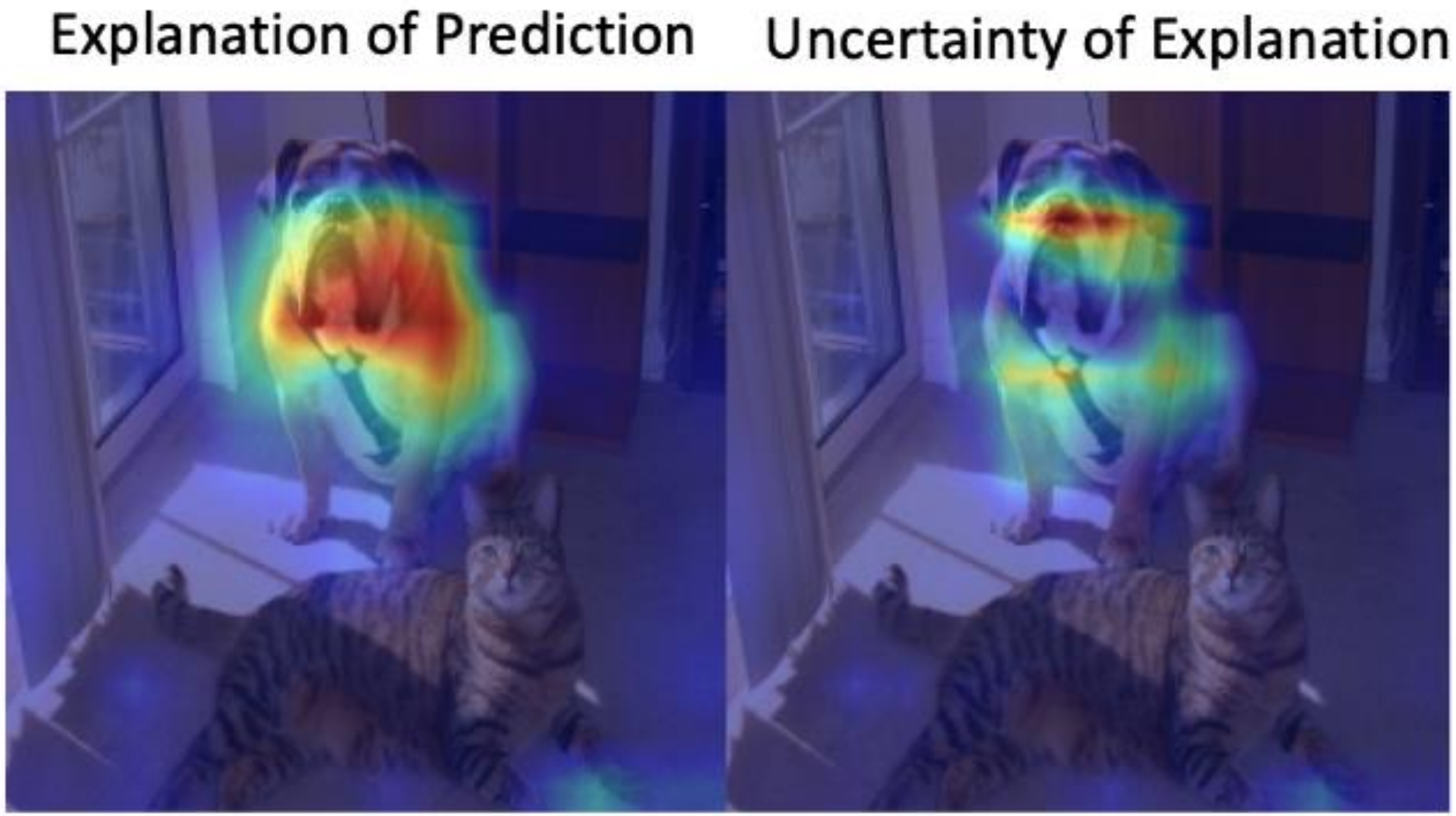
Predictive Uncertainty in Explanations

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

SCAN ME

Explanatory techniques have predictive uncertainty

Why Bullmastiff?



Uncertainty in answering Why Bullmastiff?

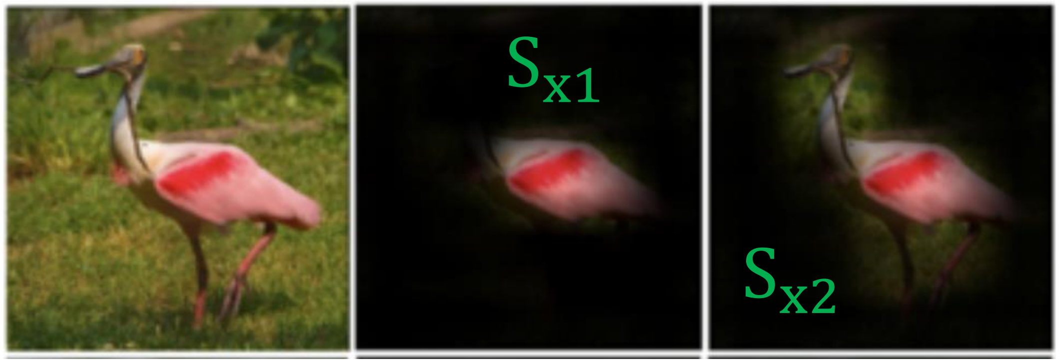
M. Prabhushankar and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," Journal of Selected Topics in Signal Processing (J-STSP) Special Series on AI in Signal & Data Science, May 23, 2024.

Uncertainty in Explainability
Explanation Evaluation via Masking

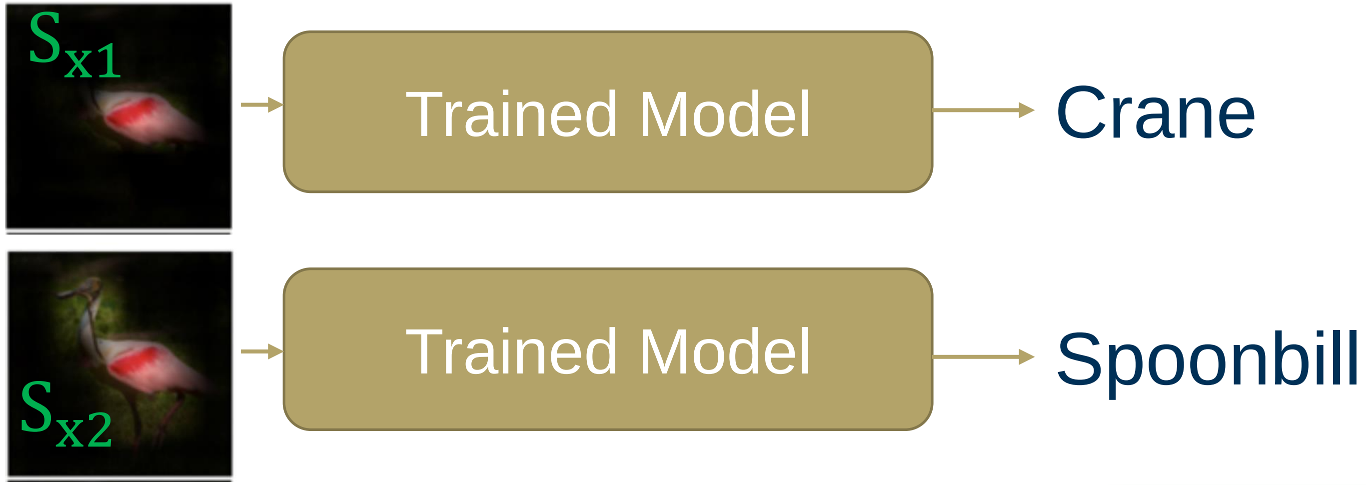
Common evaluation technique is masking the image and checking for prediction correctness

y = Prediction
S_x = Explanation masked data

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



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



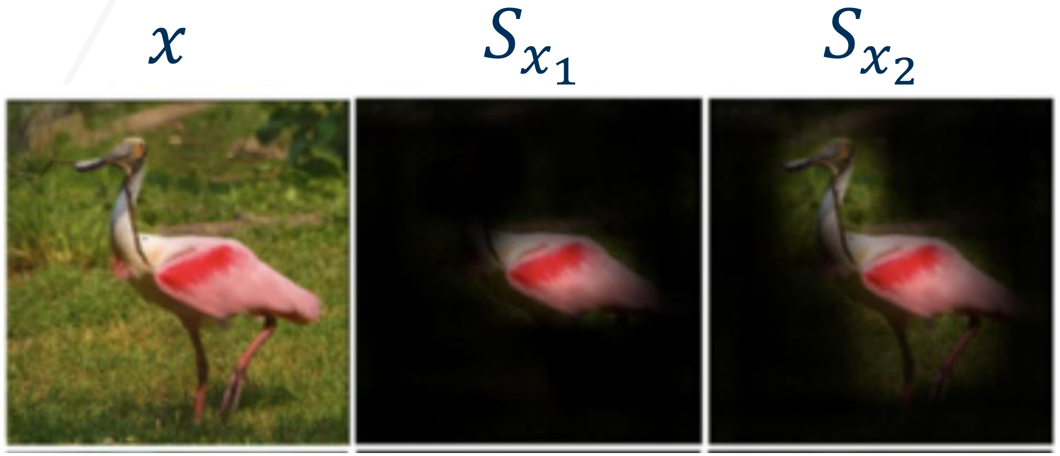
Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018.

Uncertainty in Explainability

Predictive Uncertainty

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

Uncertainty due to variance in prediction when model is kept constant



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

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

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

Uncertainty in Explainability

Visual Explanations (partially) reduce Predictive Uncertainty

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

$$V[y|S_x] = \underbrace{V[E(y|S_x)]}_{\text{zero}} + E(V[y|S_x])$$

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

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

Network evaluations have nothing to do with human Explainability!

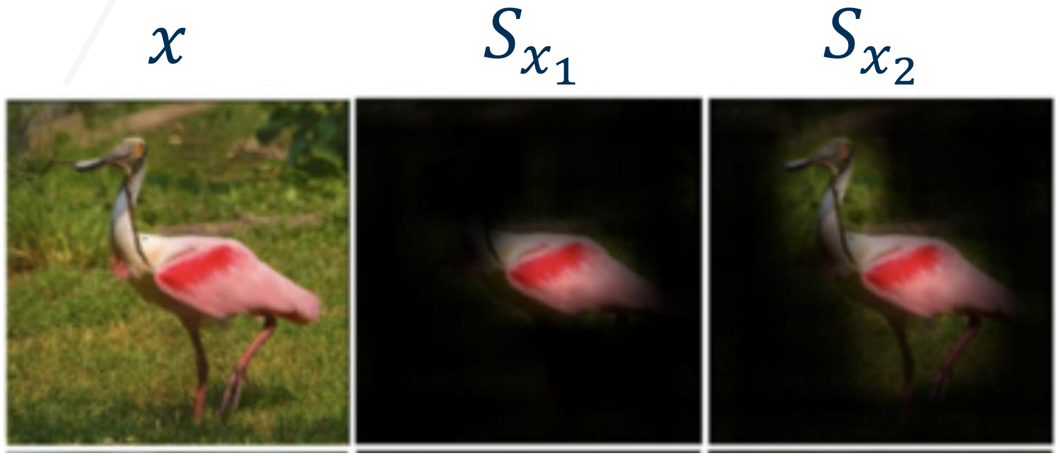
M. Prabhushankar and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," Journal of Selected Topics in Signal Processing (J-STSP) Special Series on AI in Signal & Data Science, May 23, 2024.

Uncertainty in Explainability

Predictive Uncertainty in Explanations is the Residual

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

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



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

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

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

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

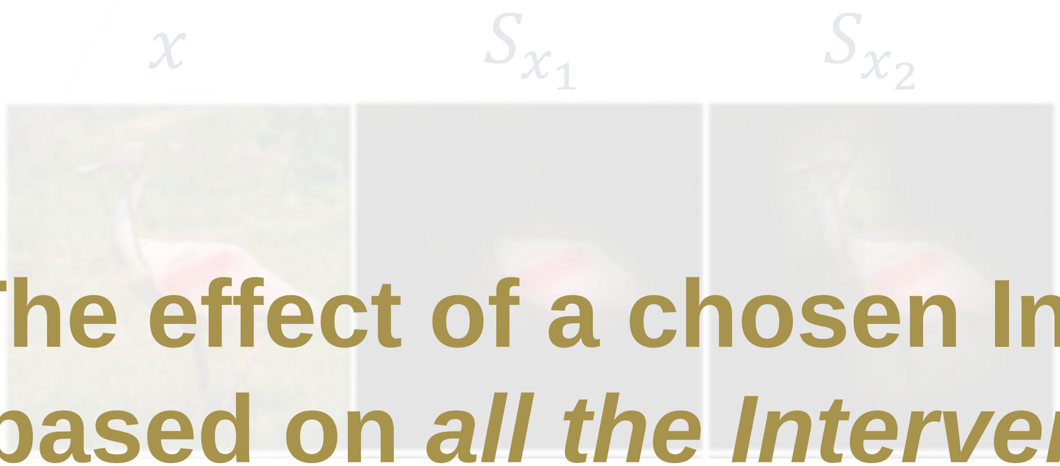
Uncertainty in Explainability

Predictive Uncertainty in Explanations is the Residual

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

SCAN ME

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



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

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

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

S_x = Subset of data (Some intervention)

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

Interventions = explanations in this context. However, they can also refer to human prompting at inference

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

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

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

Uncertainty in Explainability

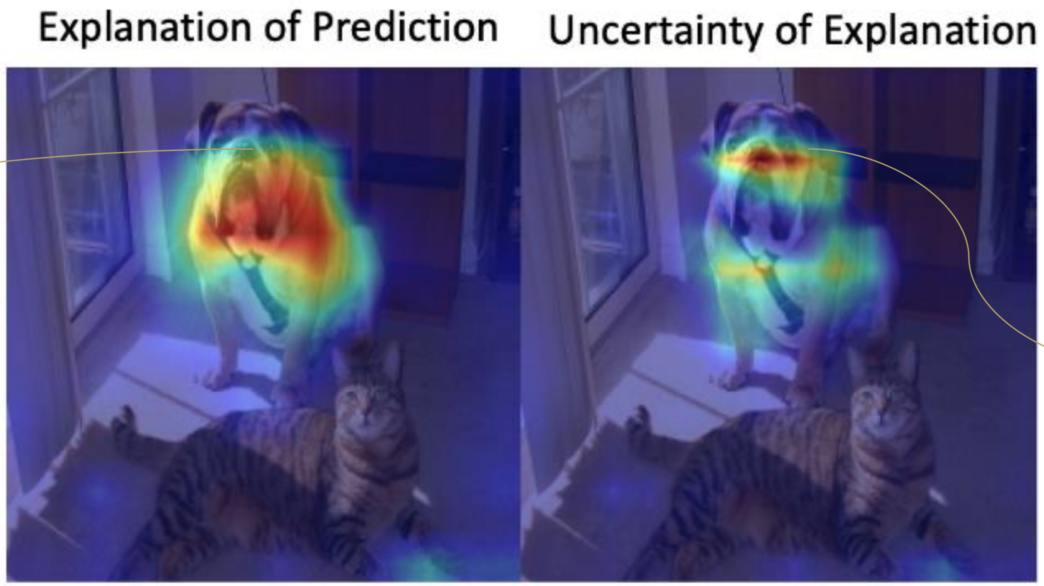
Predictive Uncertainty in Explanations is the Residual

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

SCAN ME

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

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



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

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

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

Uncertainty in Explainability

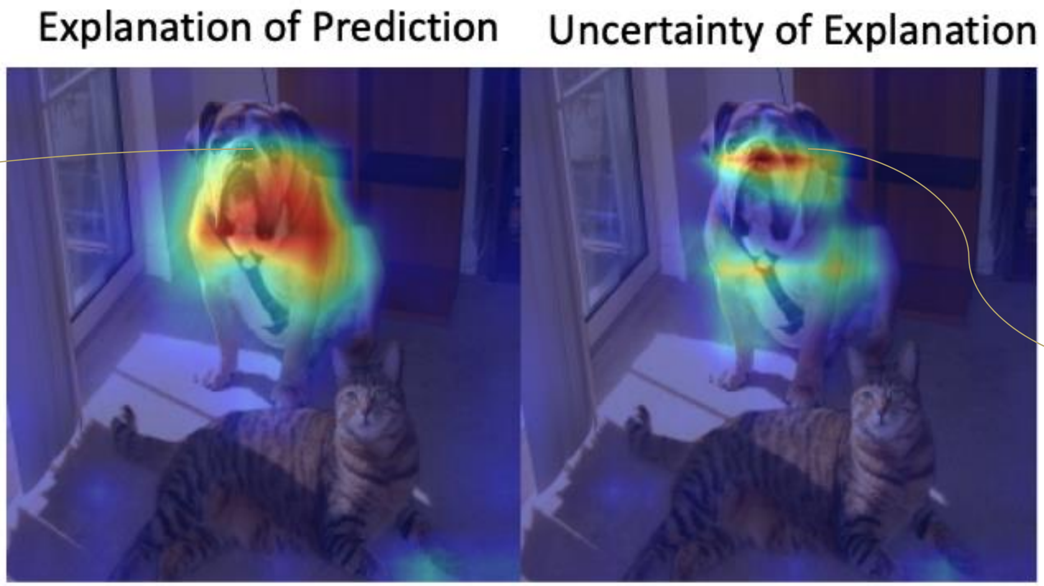
Predictive Uncertainty in Explanations is the Residual

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

SCAN ME

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

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



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

Not chosen features are intractable!

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

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

SCAN ME

Contrastive explanations are an intelligent way of obtaining other subsets

S_{x_1} S_{x_2} .... S_{x_N}

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

Make it finite by only considering the subsets that change y

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

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

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

OLIVES @GeorgiaTech

Georgia Tech

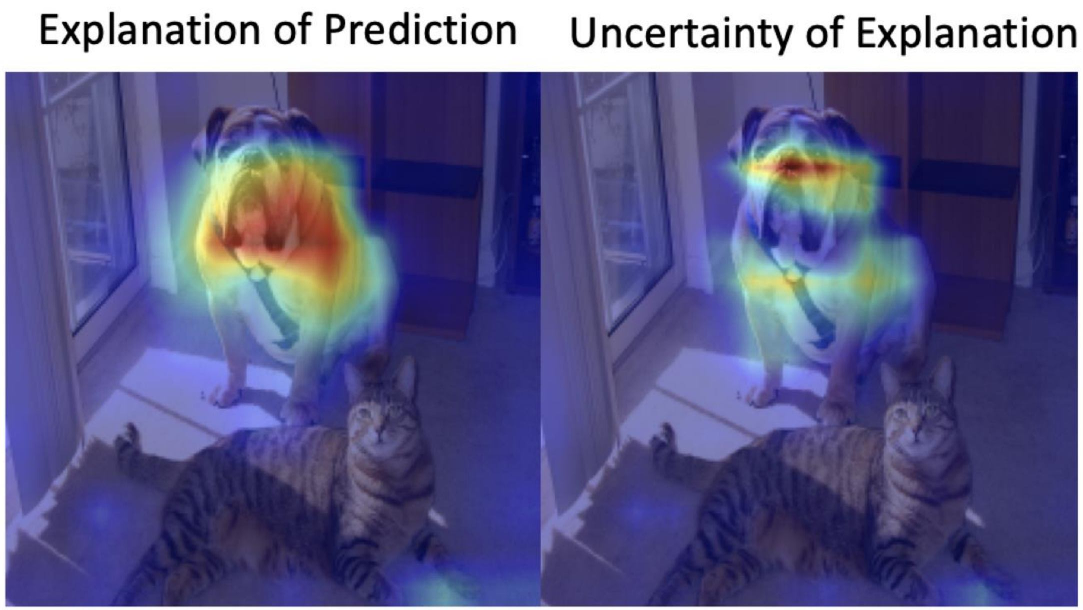
Uncertainty in Explainability

VGG vs Swin Transformer

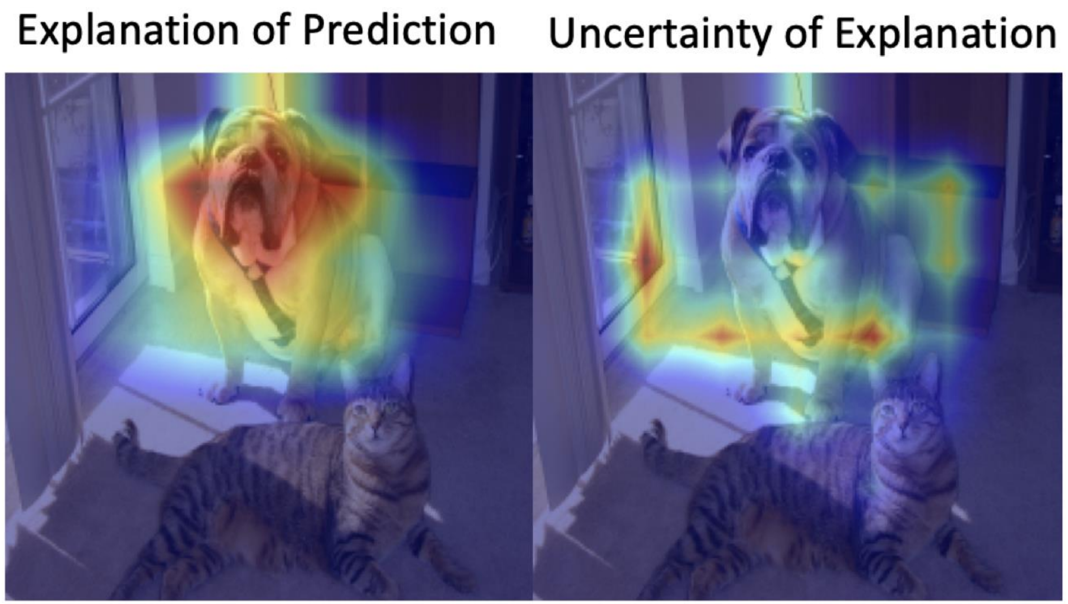
VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

Uncertainty in explainability exists in all architectures, including latest transformers

VGG-16



Swin Transformer

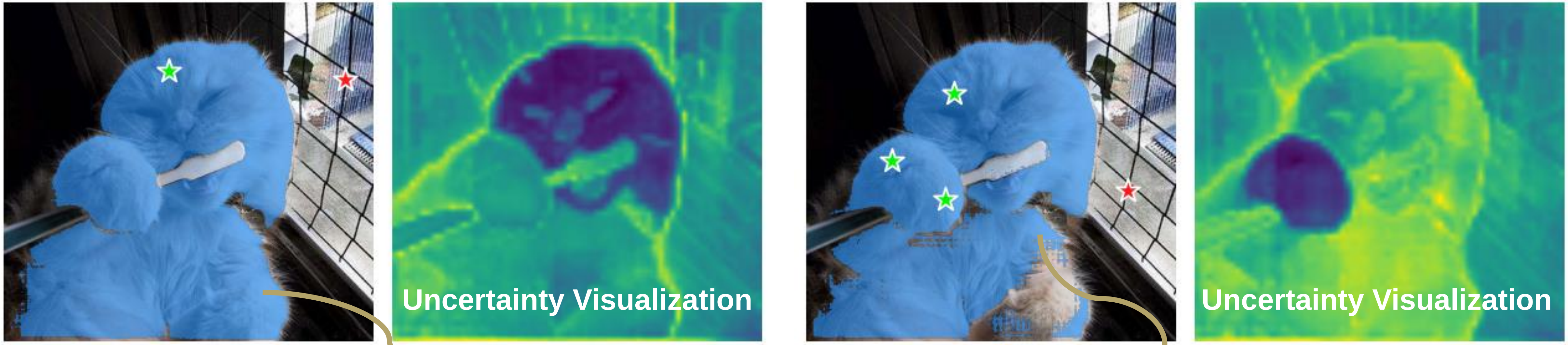


M. Prabhushankar and G. AlRegib, ”VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability,” Journal of Selected Topics in Signal Processing (J-STSP) Special Series on AI in Signal & Data Science, May 23, 2024.

Inferential Machine Learning

Our View: Goal is tied to Uncertainty Quantification

At Inference, the goal of human interventions is to reduce uncertainty



Inexplicable performance deterioration!

Dark blue regions: Low uncertainty
Green/Yellow regions: High Uncertainty

The uncertainty visualization is (variance) of (gradients-based visual explanations) – Part 3

SCAN ME

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

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

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

Need objective quantification of Intervention Residuals

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

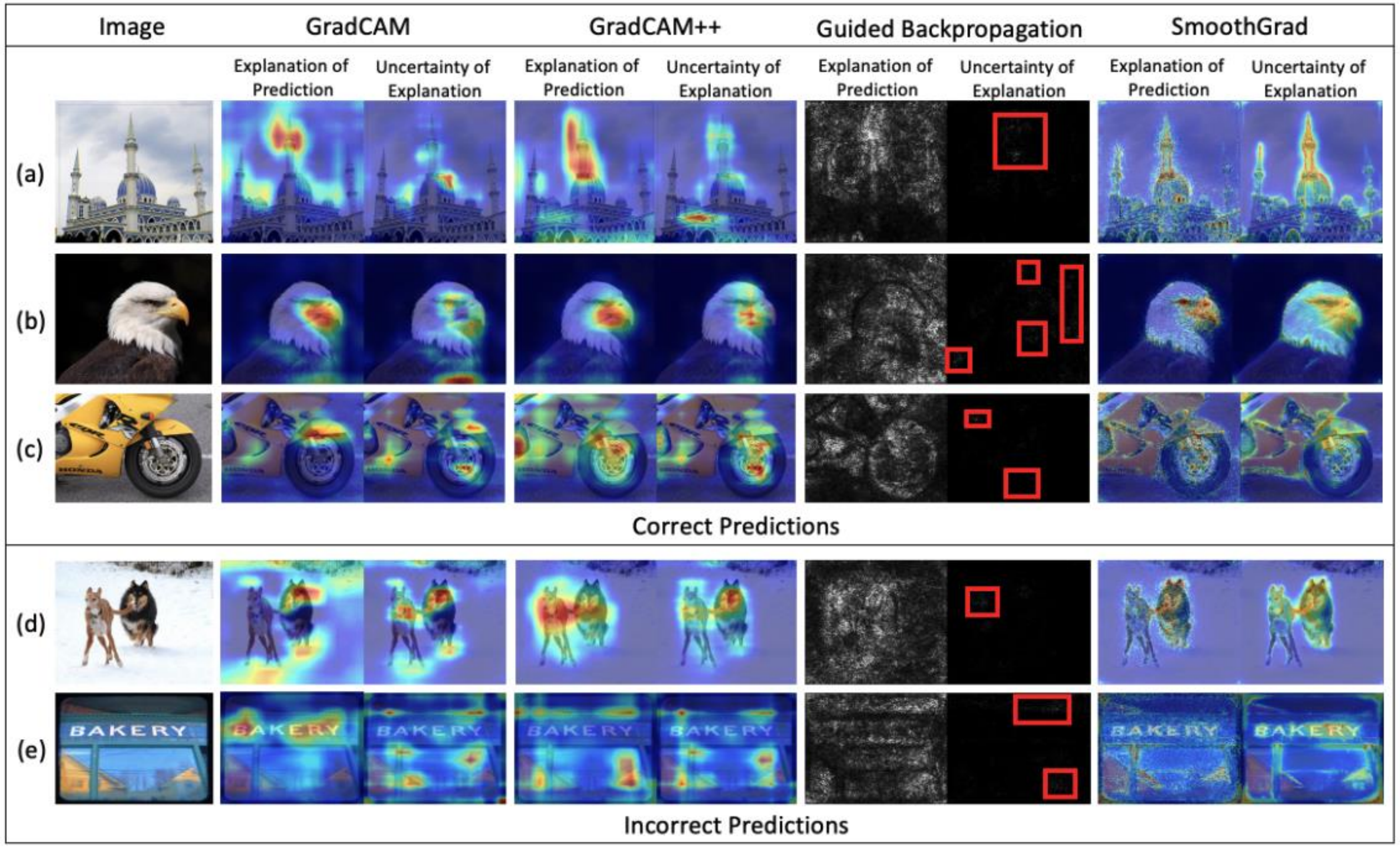
Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability: mIOU

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

SCAN ME

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



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

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

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

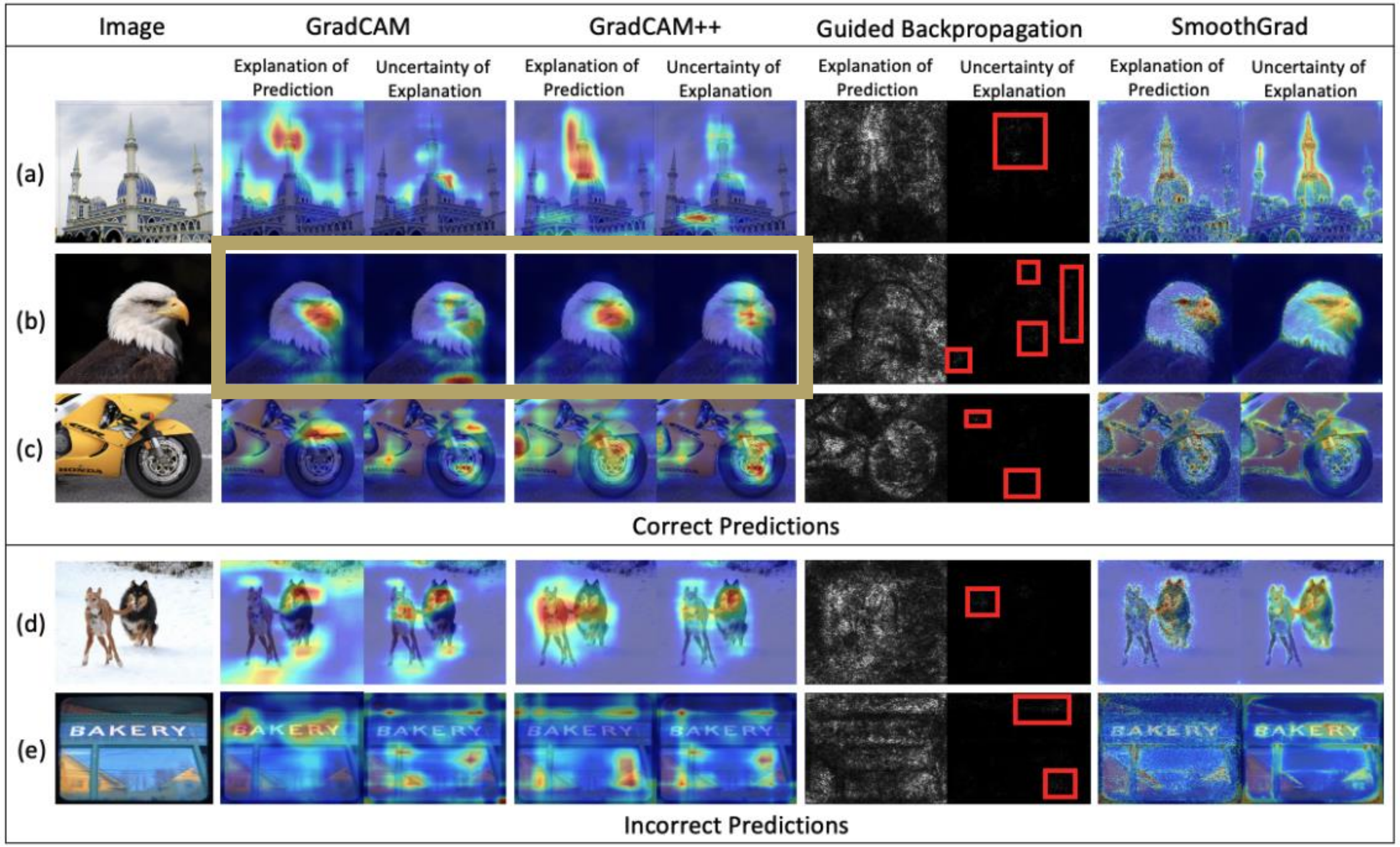
Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability: mIOU

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

SCAN ME

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



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

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

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

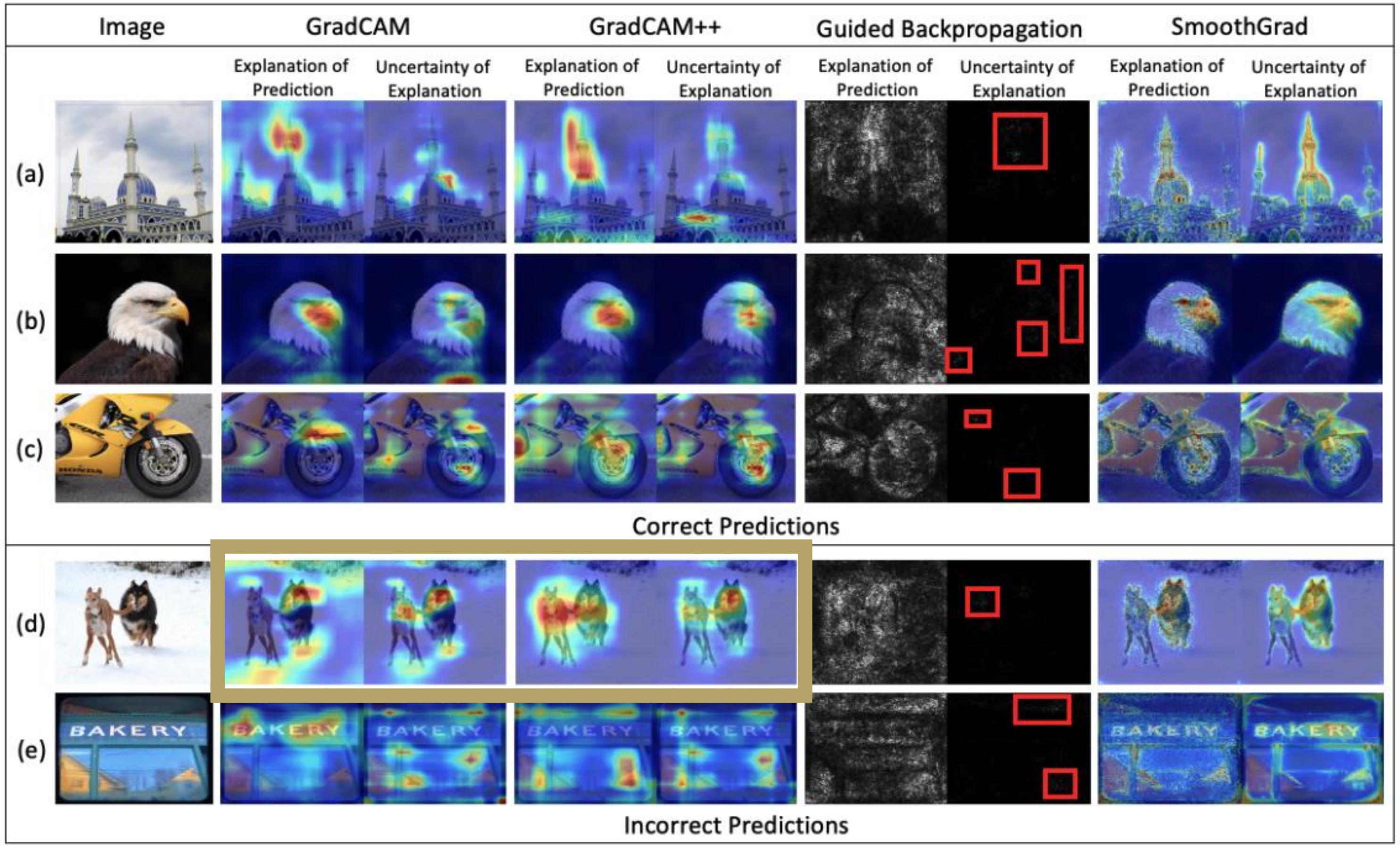
Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability: mIOU

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

SCAN ME

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



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

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

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

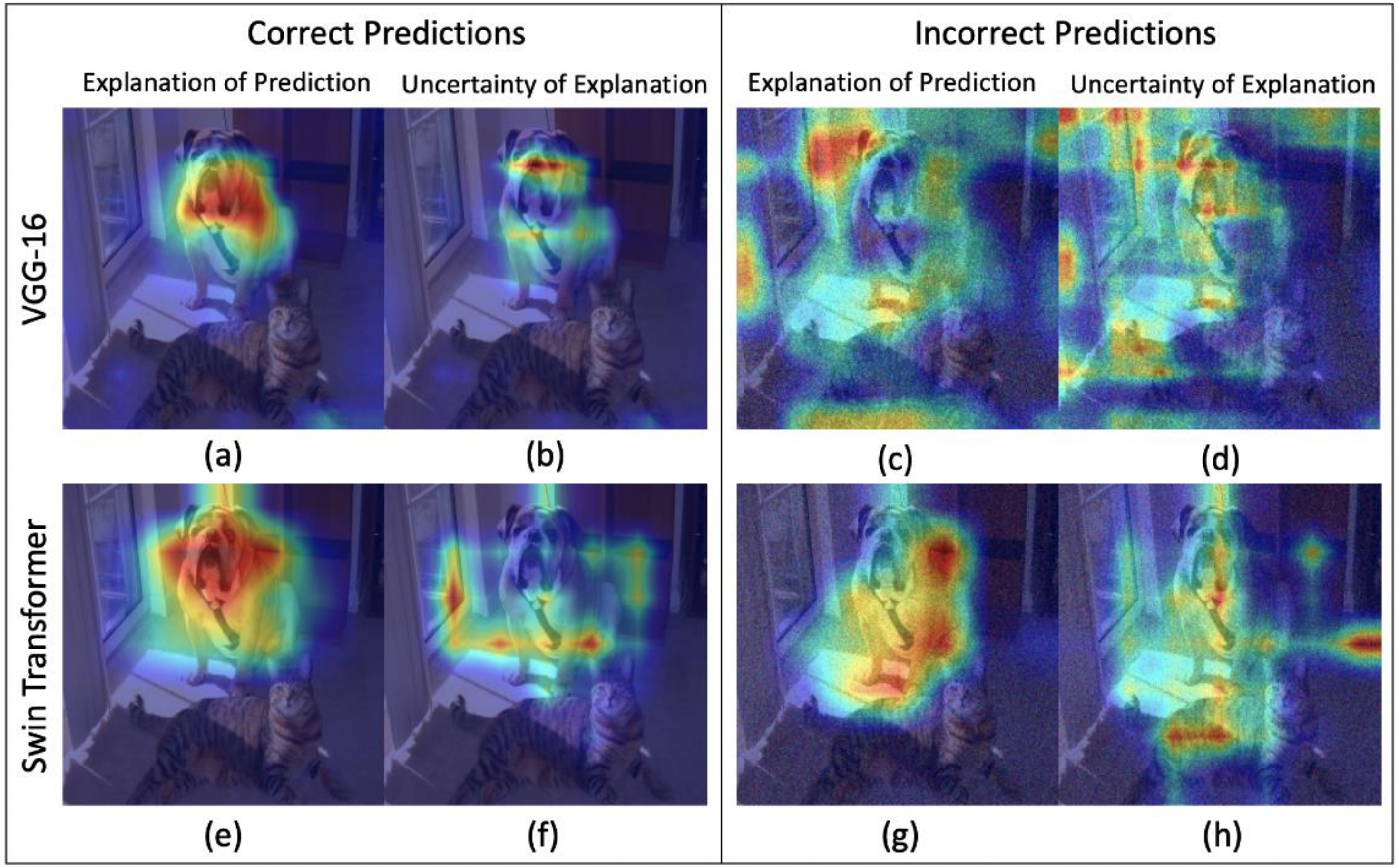
Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability: SNR

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

SCAN ME

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



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

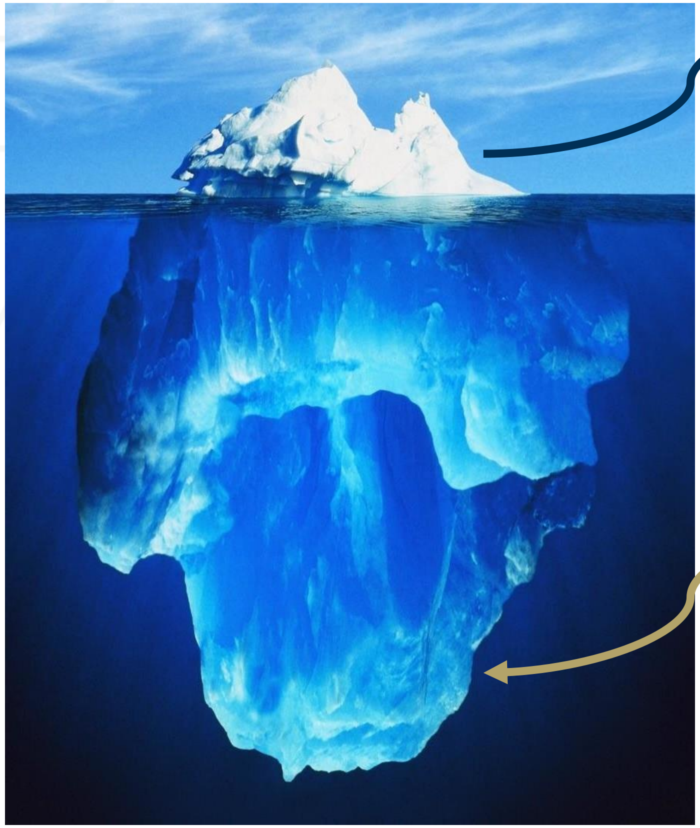
Higher the SNR of uncertainty, more is the dispersal (or less trustworthy is the prediction)

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

Uncertainty

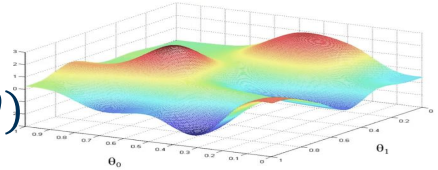
Uncertainty and Inferential Machine Learning

Uncertainty is a ‘catch-all’ term, used in multiple applications



Learned Knowledge

$L(\theta)$



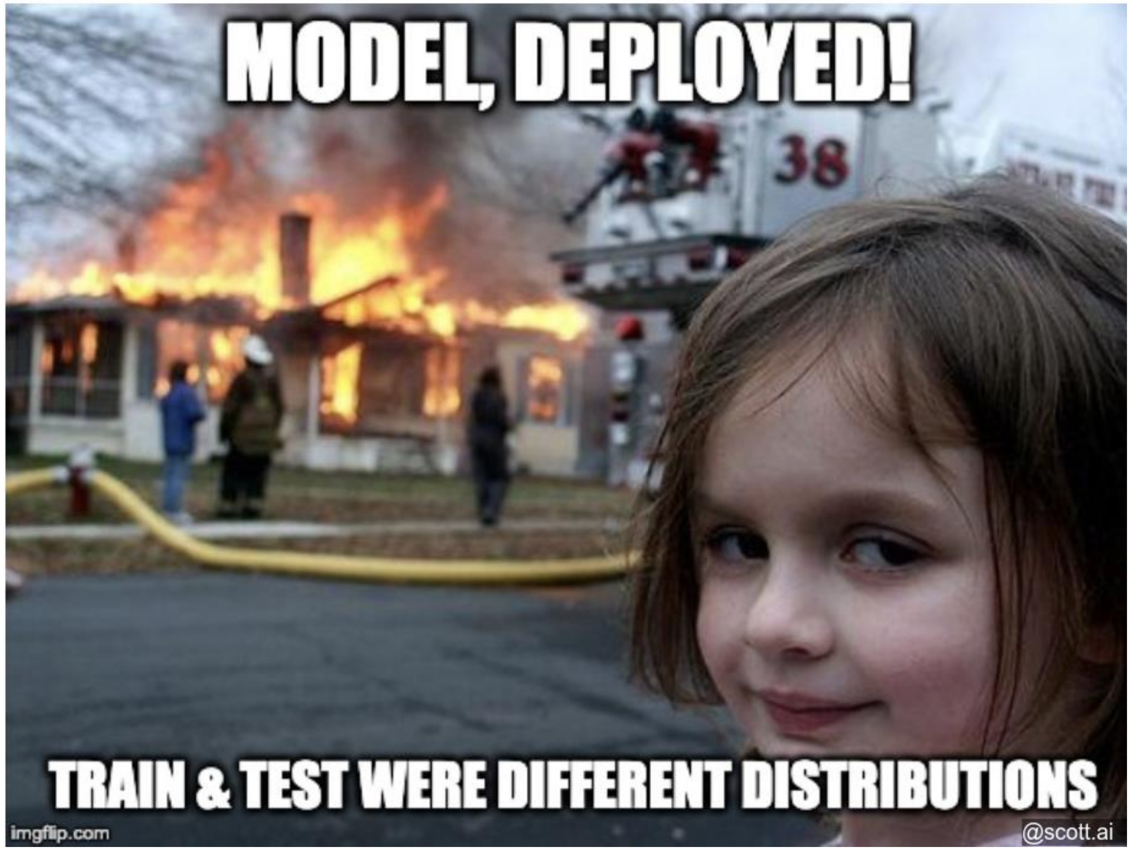
Transmuted Knowledge

Part 2

- Explainability
- Out-of-distribution Detection
- Adversarial Detection
- Anomaly Detection
- Corruption Detection
- Misprediction Detection (Case Study 1)
- Causal Analysis
- Open-set Recognition
- Noise Robustness (Case Study 3)
- Uncertainty Visualization (Case Study 2)
- Image Quality Assessment
- Saliency Detection

Memes to Wrap Up Part 3

Robustness at Inference

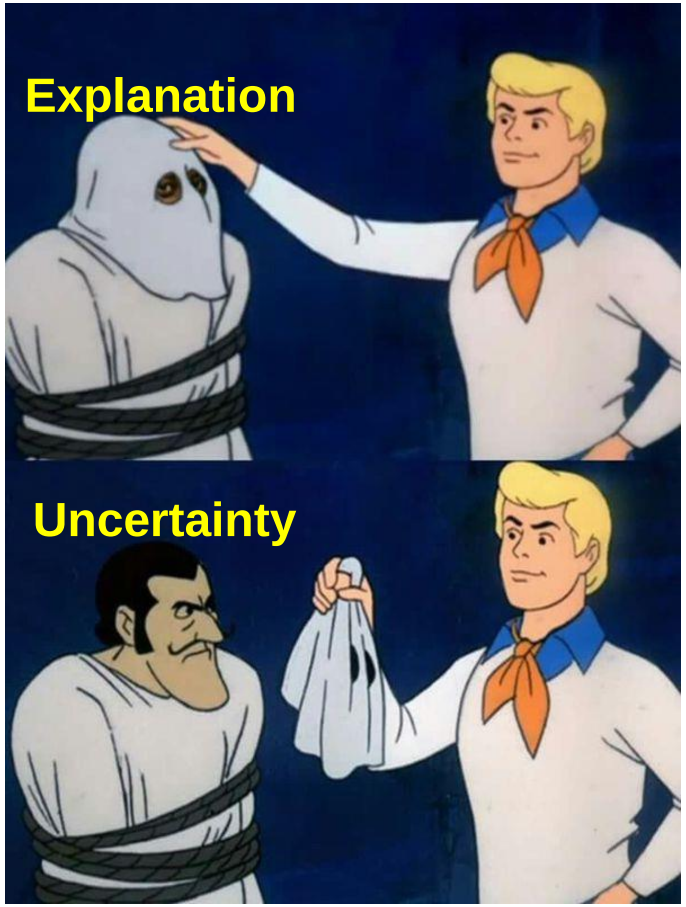


Cannot depend on training to construct robust models

Memes to Wrap Up Part 3

Explainability Research is Just Uncertainty Research

Explanatory Evaluation reduces Uncertainty



Inferential Machine Learning

Part 4: Intervenability at Inference

Objective

Objective of the Tutorial

To discuss methodologies that promote robustness in neural networks at inference

- Part 1: Inference in Neural Networks
- Part 2: Explainability at Inference
- Part 3: Uncertainty at Inference
- **Part 4: Intervenability at Inference**
    - Definitions of Intervenability
        - Causality
        - Privacy
        - Interpretability
        - Prompting
        - Benchmarking
        - Case study: Negative Interventions
    - Mathematical frameworks to study intervenability
    - Case Study: Intervenability in Interpretability
- Part 5: Conclusions and Future Directions

Intervenability

Through the Causal Glass

Assess: The amenability of neural network decisions to human interventions

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

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

Intervenability

Through the Privacy Glass

Assure: The amenability of neural network decisions to human interventions

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

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

Intervenability

Through the Interpretability Glass

Interpret: The amenability of neural network decisions to human interventions

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

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

Intervenability

Through the Prompting Glass

Actuate: The amenability of neural network decisions to human interventions

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

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

Intervenability

Through the Benchmarking Glass

Verify: The amenability of neural network decisions to human interventions

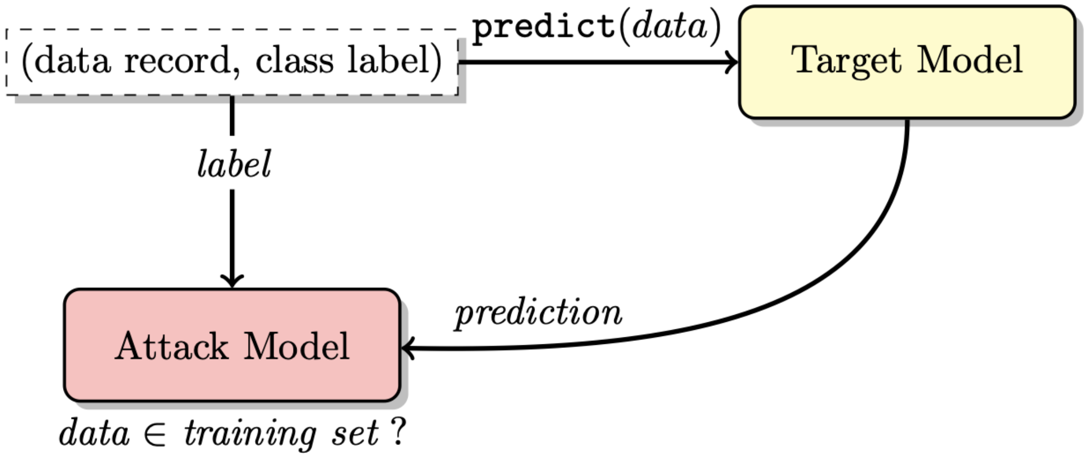
"… new ***benchmarks*** *were proposed to specifically* ***test generalization*** *of classification and detection methods with respect to* ***simple*** *algorithmically generated* ***interventions*** *like spatial shifts, blur, changes in brightness or contrast…"*

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

Case Study: Negative Interventions

Repeated Interventions: Membership Inference Attacks (MIAs)

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



Attack model is the binary classifier

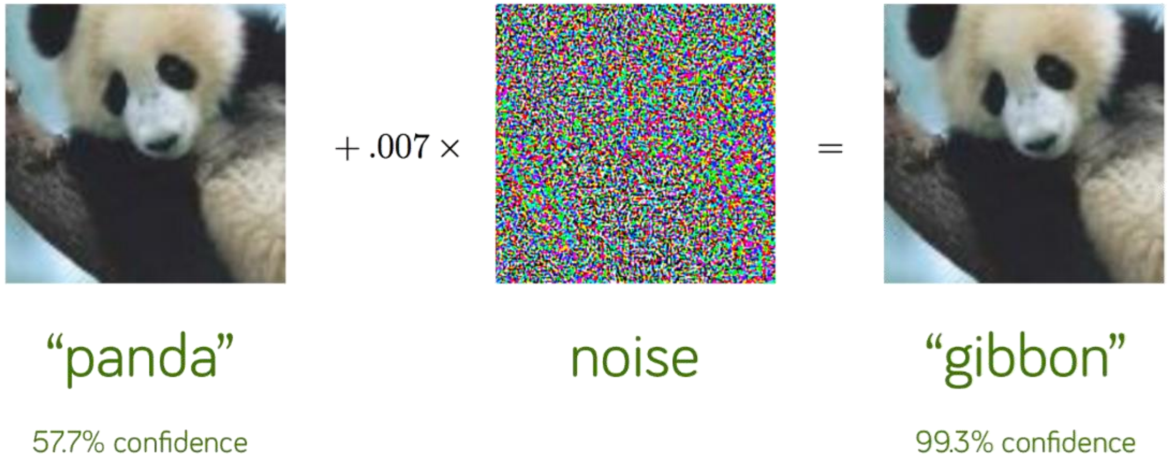
- If data is part of Electronic Health Records, then privacy of patients can be leaked
- Train a binary classifier that takes in the target model outputs and classifies whether the initial data is part of the training set
- **Prevention** is seen as a **robustness** issue while **training**: regularization,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

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

Case Study: Negative Interventions

Engineered Interventions: Adversarial Attacks

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



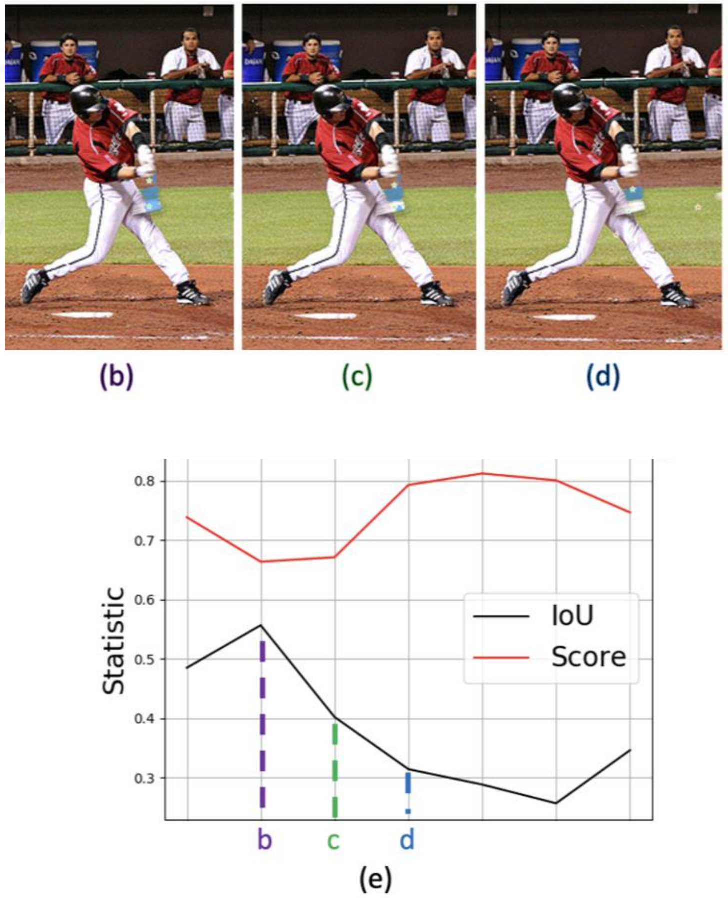
- **Gradients** (or some statistics of gradients) are used in several adversarial image generation techniques
- **Prevention** is seen as a robustness issue **both during inference and training** – adversarial training, image compression etc.

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

Case Study: Negative Interventions

‘Trial and Error’ Interventions: Visual Prompting

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



- Annotators are asked to segment objects (classes) using Segment Anything Model (SAM) and point prompts
- After prompting, annotators are shown the Intersection Over Union and provided the opportunity to add/subtract their prompt points
- The general conclusion from [1] is that annotators overprompt and utilize strategies that lead to worse performance
- Dataset: https://zenodo.org/records/10975868
- ~200,000 prompts on 6000 images

SCAN ME

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

Objective

Objective of the Tutorial

To discuss methodologies that promote robustness in neural networks at inference

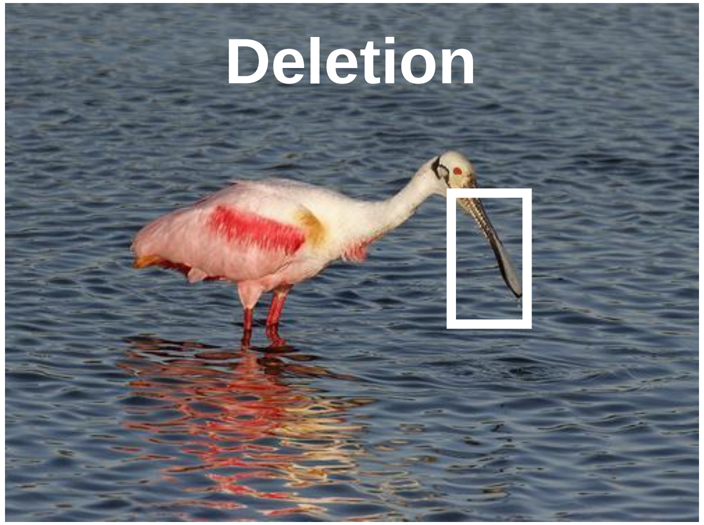
- Part 1: Inference in Neural Networks
- Part 2: Explainability at Inference
- Part 3: Uncertainty at Inference
- **Part 4: Intervenability at Inference**
    - Definitions of Intervenability
    - Mathematical frameworks to study intervenability
        - Causal analysis via interventions
        -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

Framework 1: Causal Assessment via Interventions

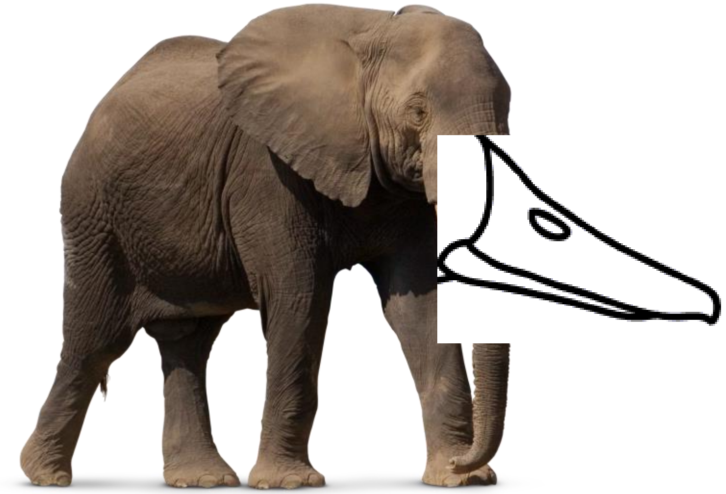
3 Rules of Causal Inference

Rule 1 (Insertion/deletion of observations):

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



Insertion



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

Key Differences:

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

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

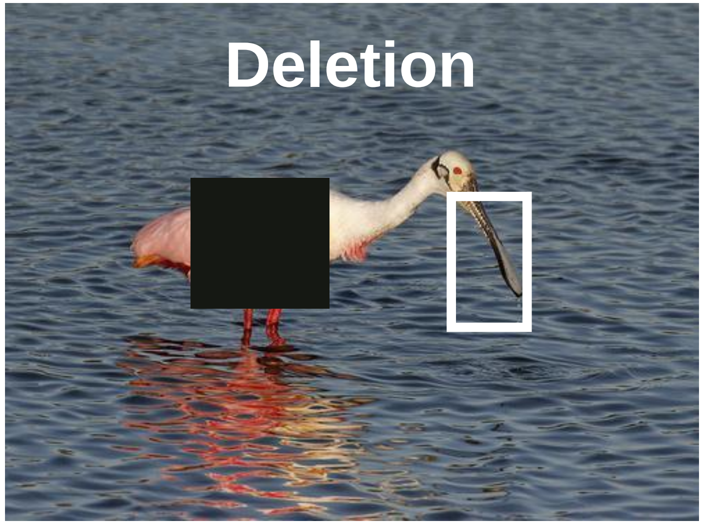
Intervenability Frameworks

Framework 1: Causal Assessment via Interventions

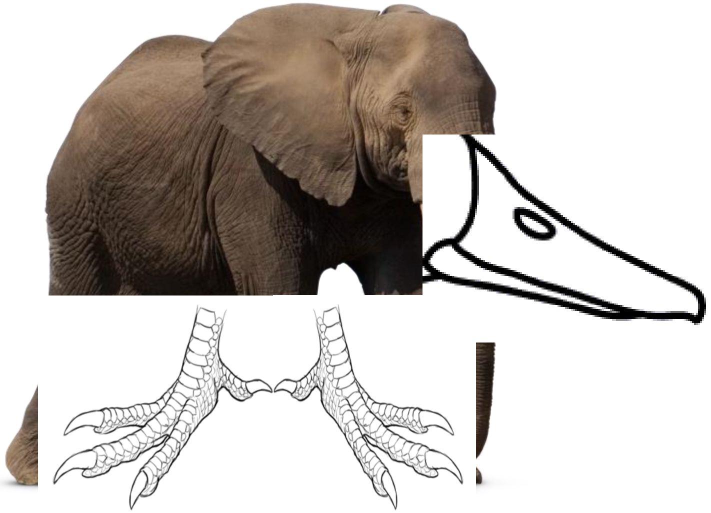
Rule 2: Intervene on all other factors keeping the causal factor constant

Rule 2 (Action/observation exchange):

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



Insertion



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

Key Differences:

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

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

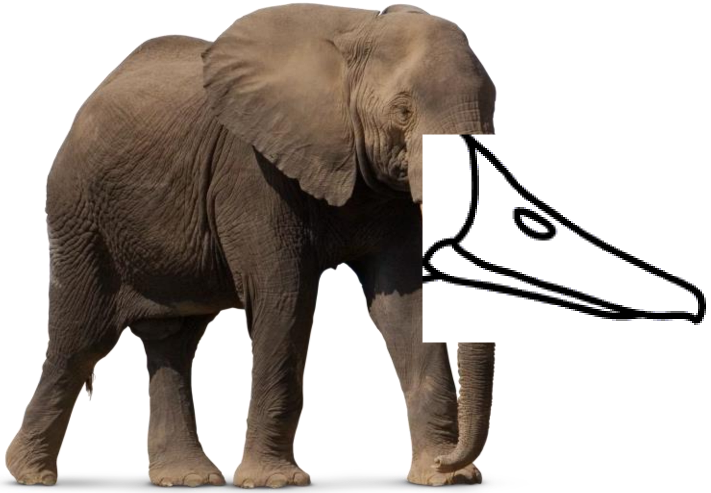
Rule 3: Insertion/Deletion of interventional actions

Rule 3 (Insertion/deletion of actions):

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

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

Insertion



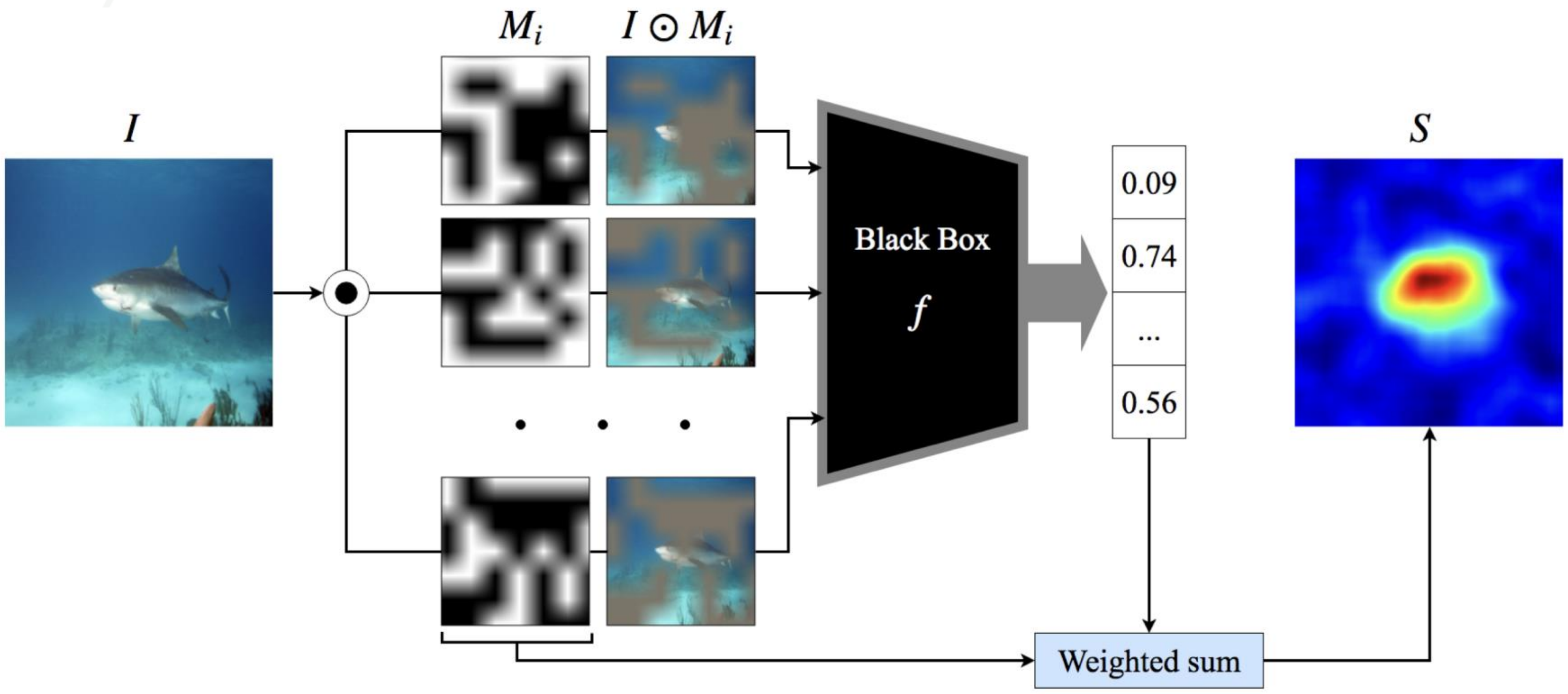
Key Differences:

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

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

Dangers of Incomplete Interventions: RISE Explanations

Unknown interventions based on insertion/deletion can yield unexpected results



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

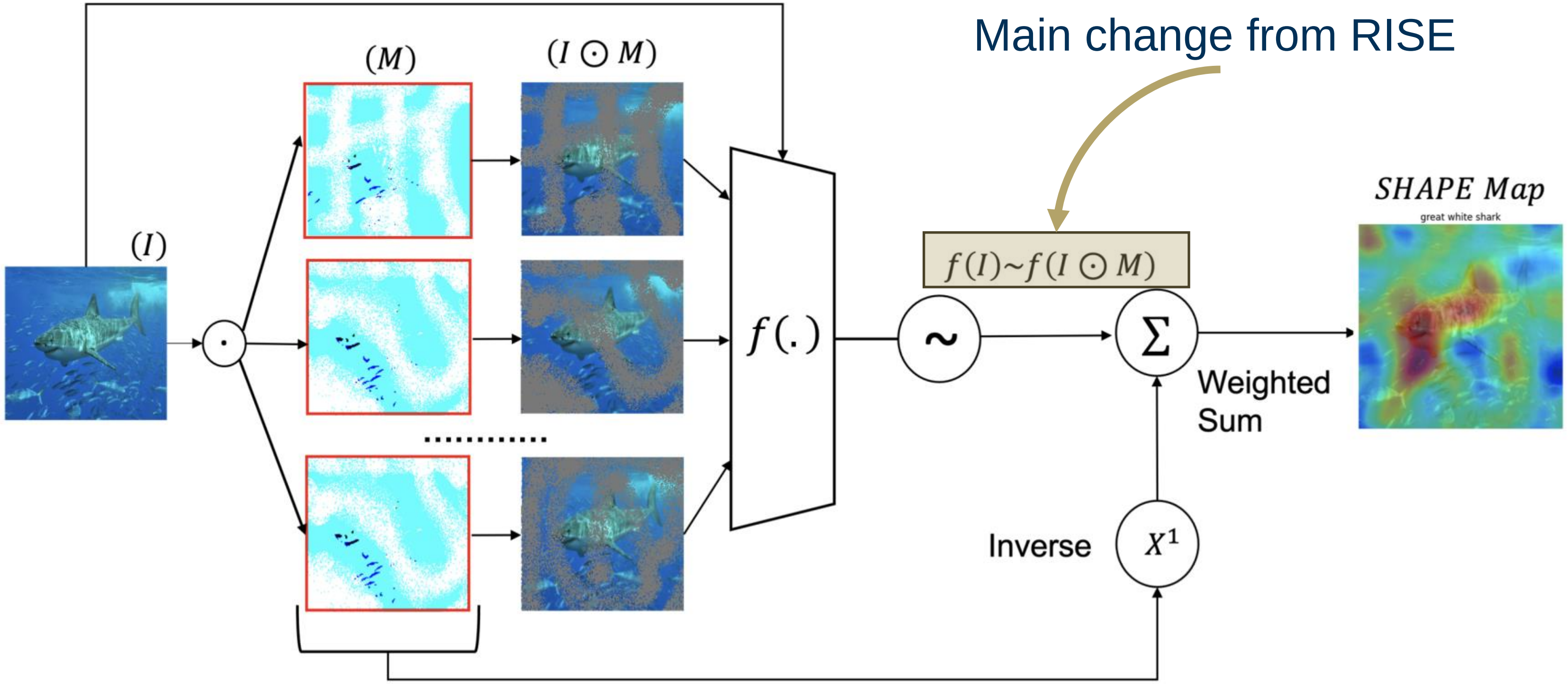
- Instead of causal deletion, RISE deletes randomly

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

Intervenability Frameworks

Dangers of Incomplete Interventions: SHAPE Explanations

Unknown interventions based on insertion/deletion can yield unexpected results



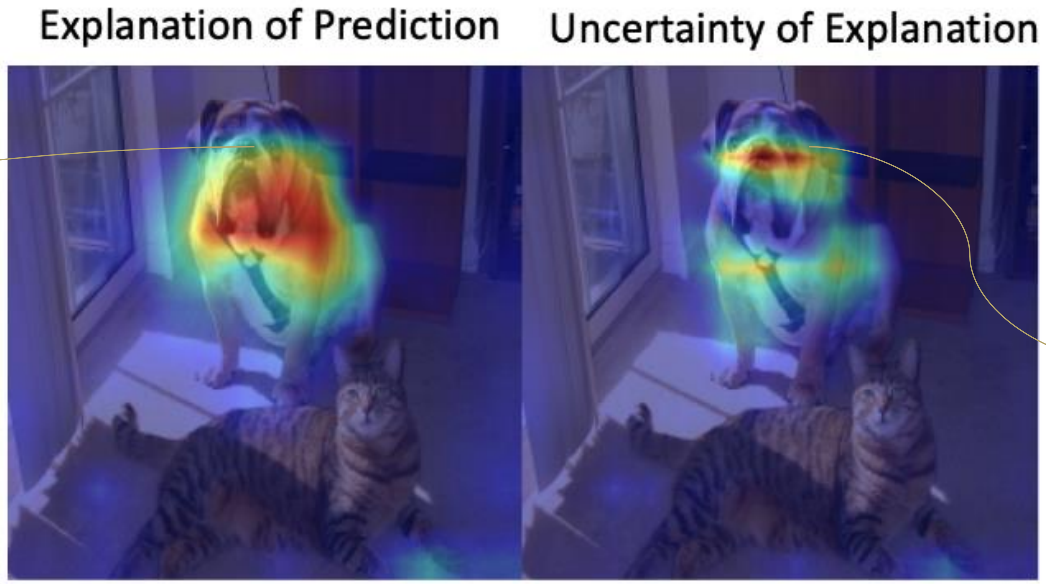
- **SHAPE** explanation is almost identical to RISE except:
  - Weighted sum is **NOT** between probability and mask but between ***change in probability score*** and inverse mask
- Results are human un-interpretable
- **However, existing objective evaluation metrics give better scores to SHAPE than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

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

Intervenability Frameworks

Framework 2: Predictive Uncertainty in Interventions

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



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

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

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

Objective

Objective of the Tutorial

To discuss methodologies that promote robustness in neural networks at inference

- Part 1: Inference in Neural Networks
- Part 2: Explainability at Inference
- Part 3: Uncertainty at Inference
- **Part 4: Intervenability at Inference**
    - Definitions of Intervenability
    - Mathematical frameworks to study intervenability
    - Case Study: Intervenability in Interpretability
        - Explanatory evaluation
- Part 5: Conclusions and Future Directions

Our Goal: To show that there is no one-size-fits all when choosing interventions

We specifically study this for the case of Explanatory Evaluation

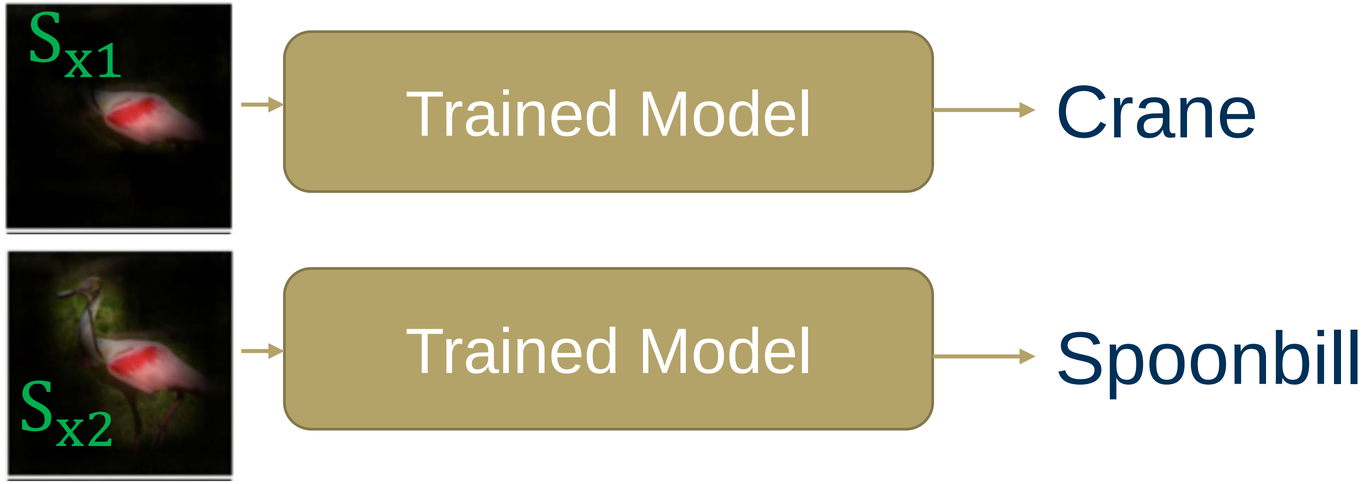
Case Study: Intervenability in Interpretability

Evaluation 1: Explanation Evaluation via Masking

Visual explanations are evaluated via masking the important regions in the image and passing it through the network

Three types of Masking:

1. **Masking using explanation heatmap**
2. Pixel-wise masking using explanation as importance
3. Structure-wise masking using information encoded in explanation



Masking = Intelligent Intervention

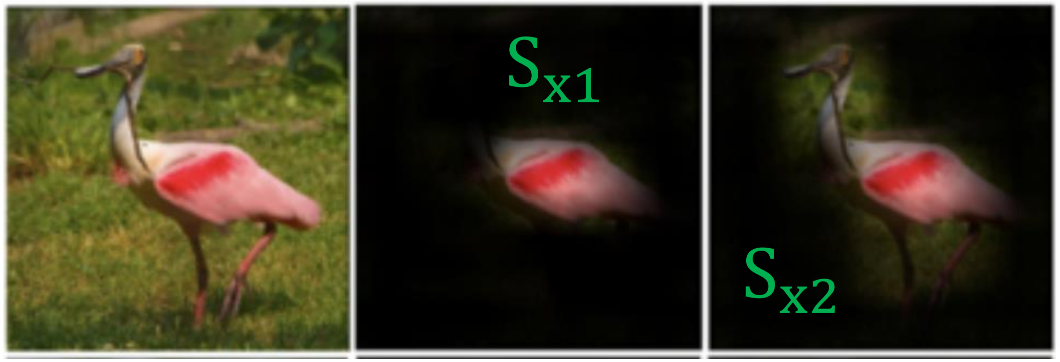
Case Study: Intervenability in Interpretability

Evaluation 1: Explanation Evaluation via Masking

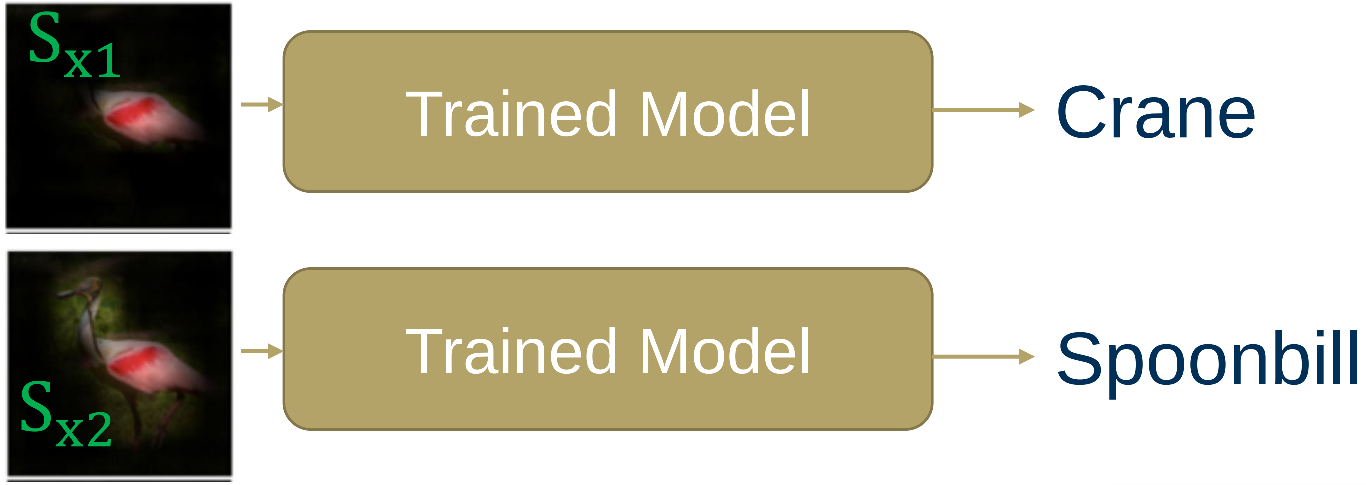
Common evaluation technique is masking the image and checking for prediction correctness

y = Prediction
S_x = Explanation masked data

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



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



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

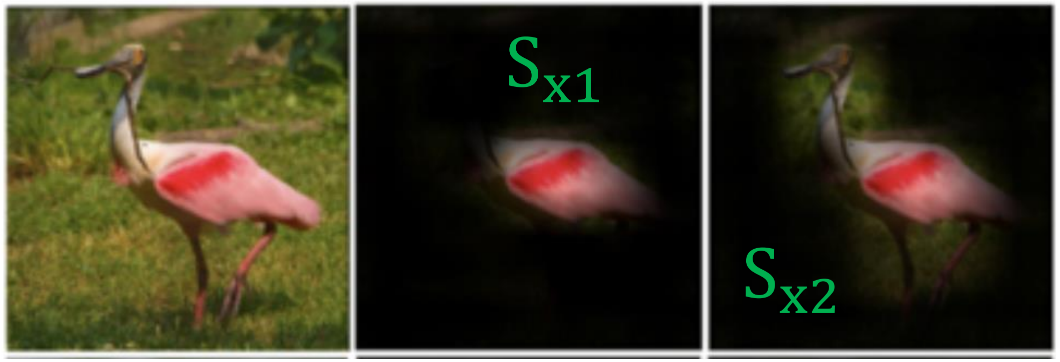
Case Study: Intervenability in Interpretability

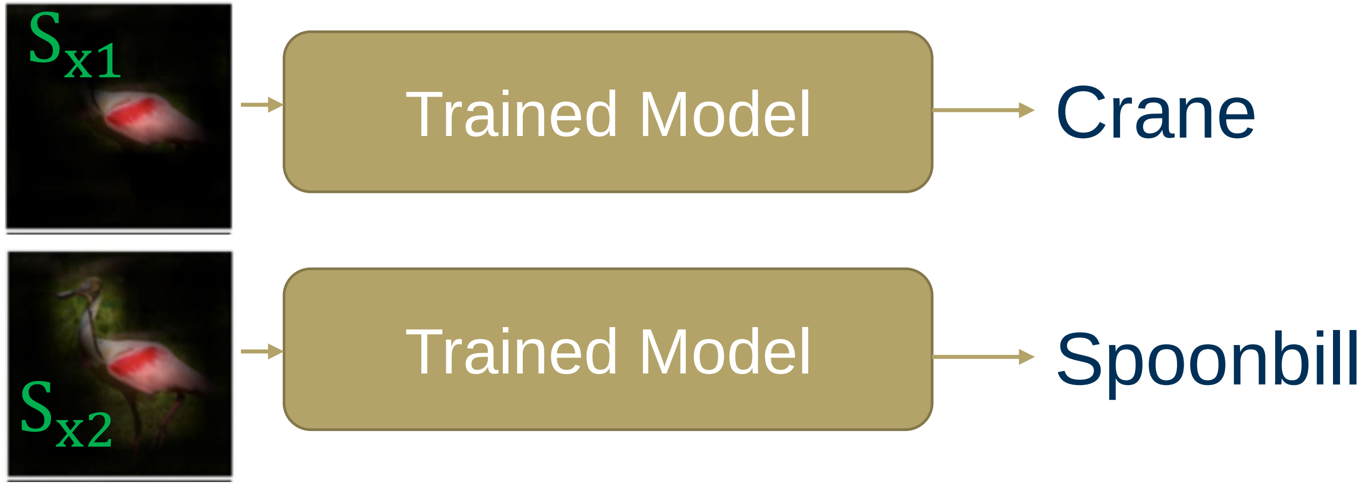
Evaluation 1: Explanation Evaluation via Masking

However, explanation masking encourages ‘larger’ explanations

- Larger explanations imply more features in masked images are intact (unmasked)
- This increases likelihood of a correct prediction

- ‘Fine-grained’ explanations are not promoted



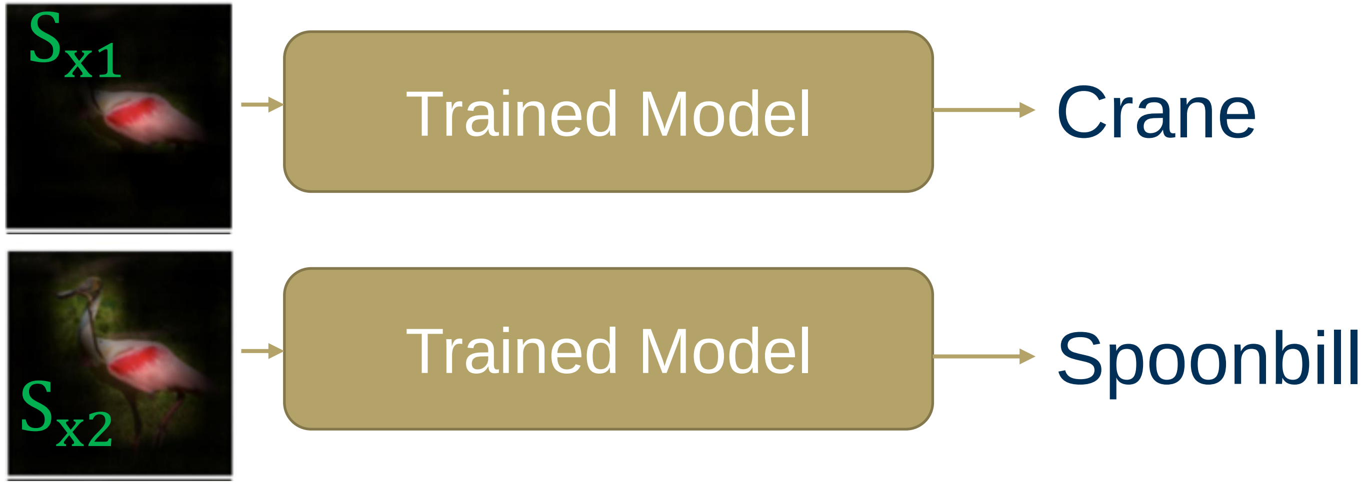


Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018.

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

Three types of Masking:

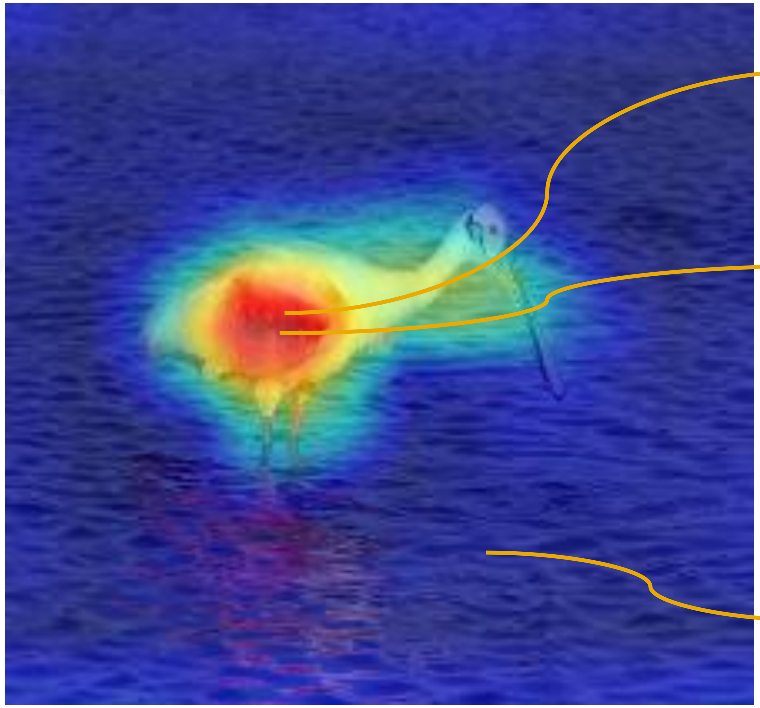
1. Masking using explanation heatmap
2. **Pixel-wise masking using explanation as importance**
3. Structure-wise masking using information encoded in explanation



Case Study: Intervenability in Interpretability

Evaluation 2: Progressive Pixel-wise Insertion and Deletion

Pixel-wise Deletion: Sequentially delete (mask) pixels in an image based on their explanation assigned importance scores



Highest importance

Second Highest importance

.
.
.

Least importance

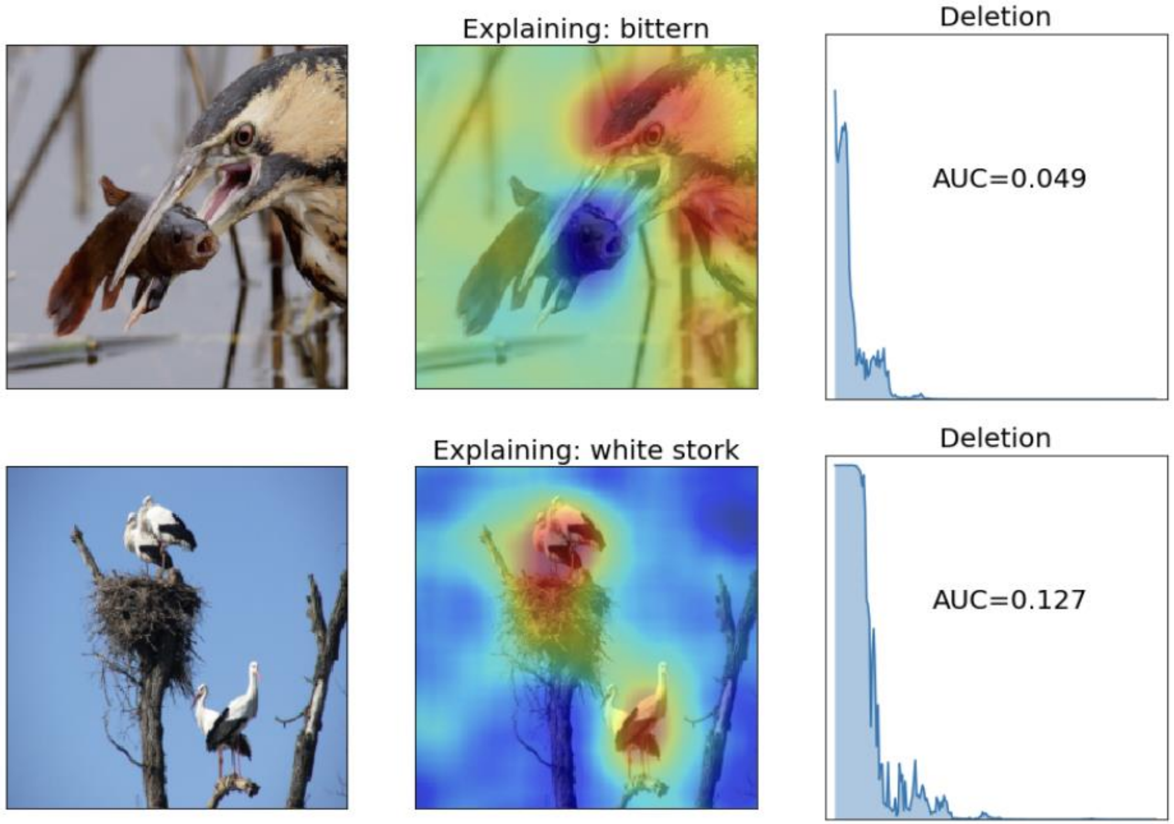
Step 1: Mask highest importance pixel and pass the image through the network. Note the probability of spoonbill.
Step 2: Mask the second highest importance pixel from the image in Step 1 and pass the image through the network. Note the probability of spoonbill.
Step 3: Repeat until all pixels are deleted (masked)

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

Case Study: Intervenability in Interpretability

Evaluation 2: Progressive Pixel-wise Insertion and Deletion

The removal of the "cause” (important pixels) will force the base model to change its decision.



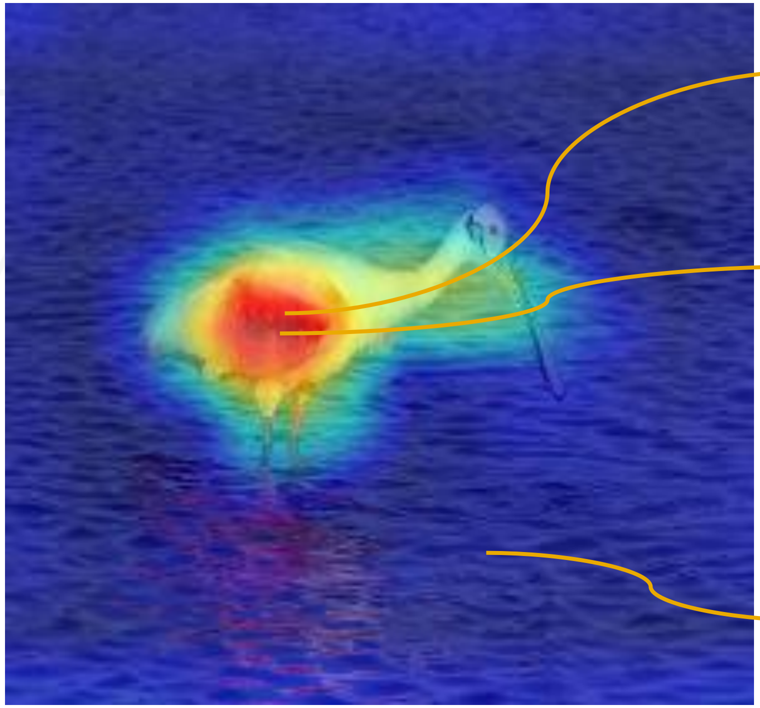
- **Deletion approximates Necessity** criterion of a “good” explanation
- **AUC** for a good explanation will be **low**
- **Deletion** encourages **fine-grained explanations** by choosing those heatmaps that select the most relevant pixels

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

Case Study: Intervenability in Interpretability

Evaluation 2: Progressive Pixel-wise Insertion and Deletion

Pixel-wise Insertion: Sequentially add pixels to a mean image based on their explanation assigned importance scores



Highest importance

Second Highest importance

.
.
.

Least importance

Take a mean (grayscale) image

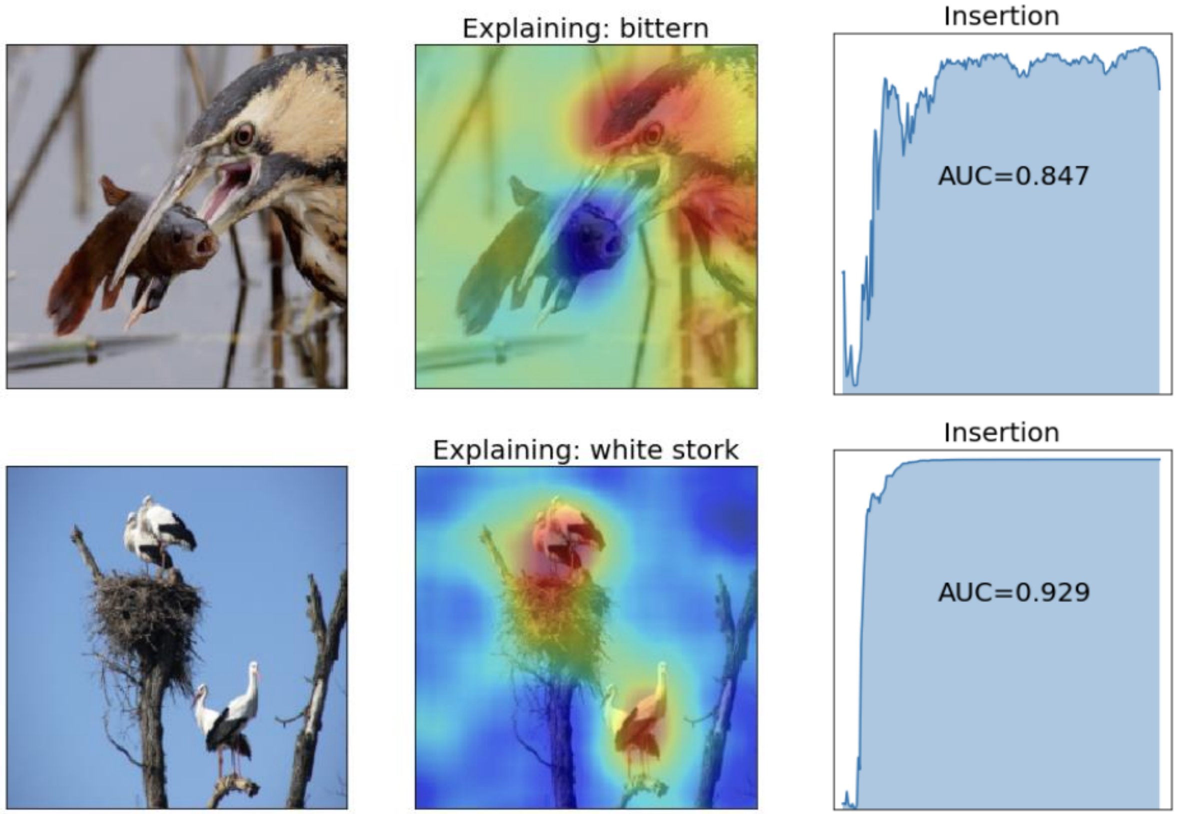
Step 1: Add the highest importance pixel to the mean image and pass it through the network. Note the probability of spoonbill.
Step 2: Add the second highest importance pixel to the image in Step 1 and pass the image through the network. Note the probability of spoonbill.
Step 3: Repeat until all pixels are inserted

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

Case Study: Intervenability in Interpretability

Evaluation 2: Progressive Pixel-wise Insertion and Deletion

The addition of the "cause" (important pixels) will force the base model to change its decision.



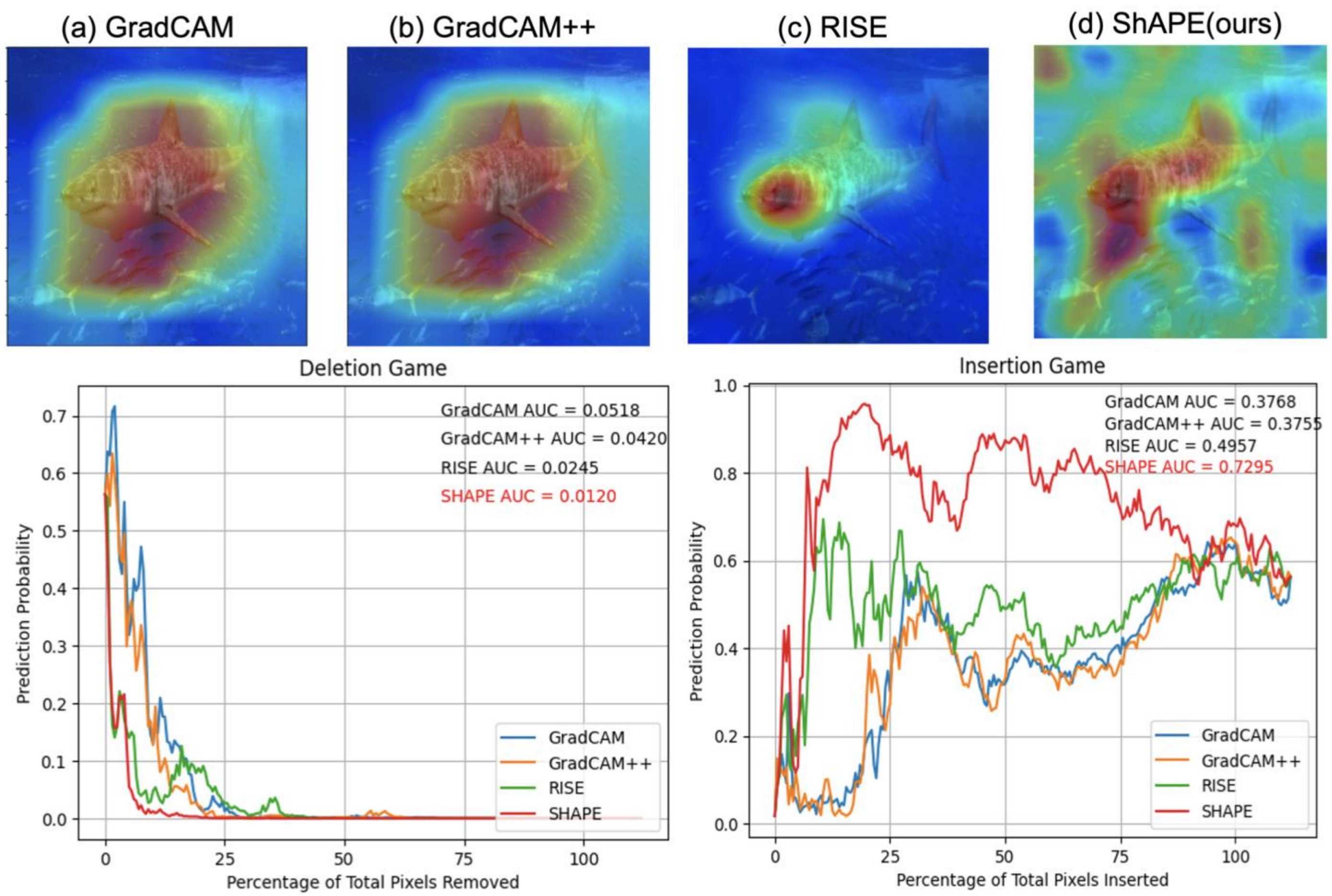
- **Insertion approximates Sufficiency** criterion of a "good" explanation
- **AUC** for a good explanation will be **high**
- **Insertion** encourages **fine-grained explanations** by choosing those heatmaps that select the most relevant pixels

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

Case Study: Intervenability in Interpretability

Evaluation 2: Progressive Pixel-wise Insertion and Deletion

Insertion and Deletion evaluation metrics encourage pixel-wise analysis of explanations



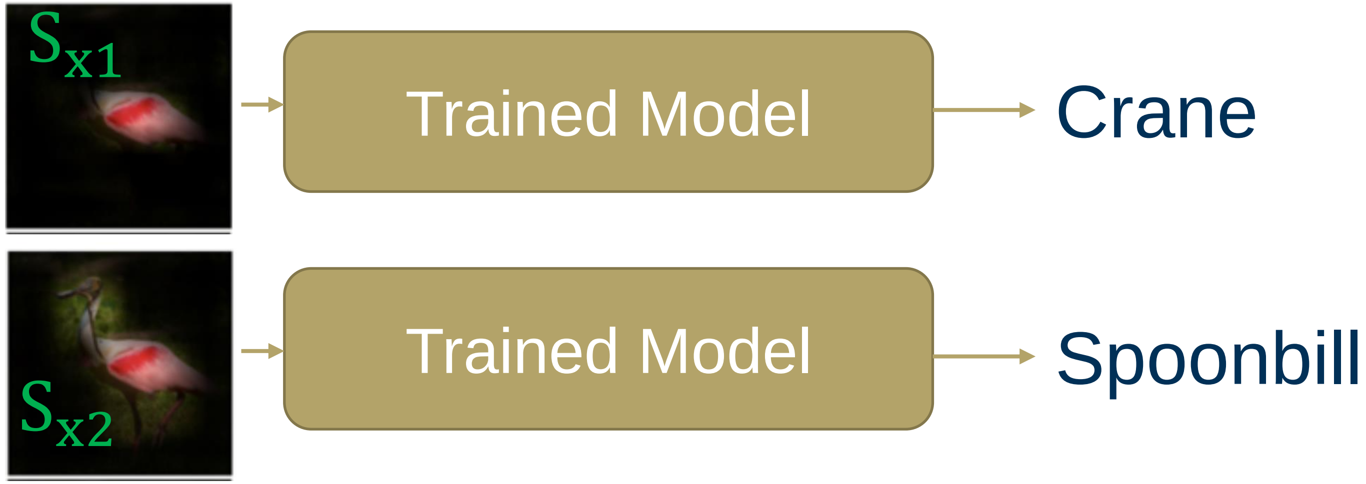
- **However, humans do not "see" in pixels**
- Rather they view scenes in a "**structure-wise**" fashion
- While **heatmap masking** encourages **large explanations**, **pixel-wise masking** encourages **unrealistic and non-human** like explanations

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

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

Three types of Masking:

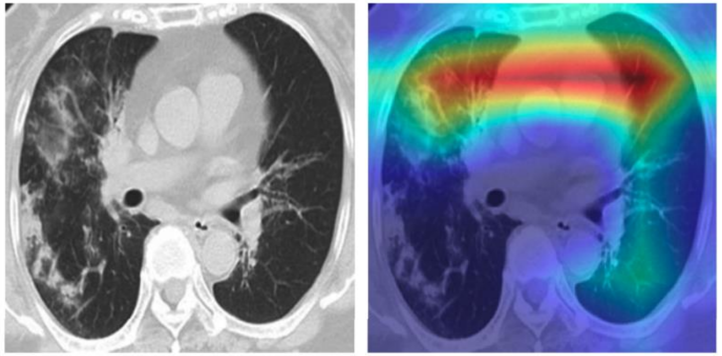
1. Masking using explanation heatmap
2. Pixel-wise masking using explanation as importance
3. **Structure-wise masking using information encoded in explanation**

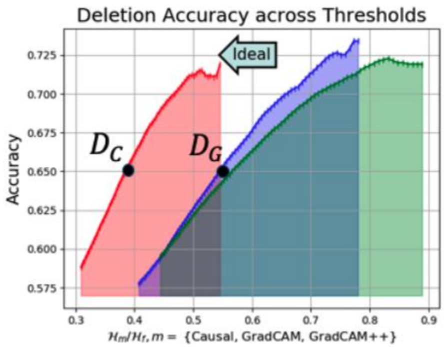


Case Study: Intervenability in Interpretability

Evaluation 3: Progressive Structure-wise Insertion and Deletion

Structure-wise Deletion: Sequentially delete (mask) pixels in an image based on the number of bits used to represent the region





Ideal scenario: The explanation encodes the most important information in the least possible bits

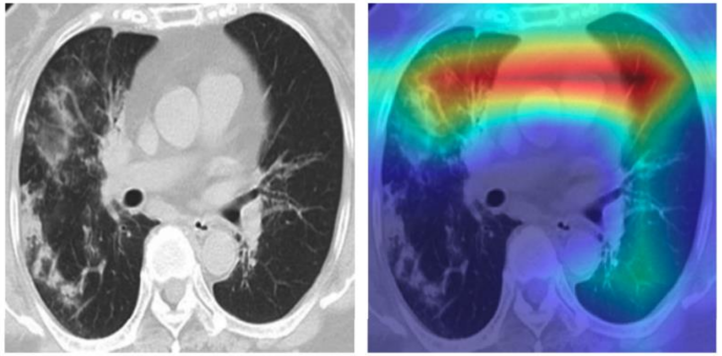
CausalCAM in Red[1]
GradCAM in Purple
GradCAM++ in Green

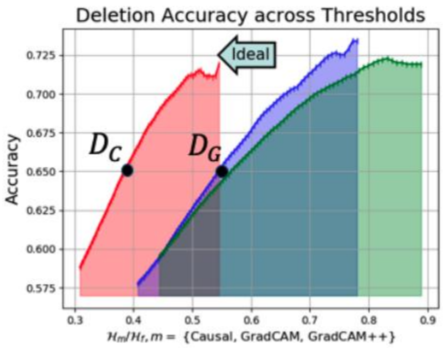
- D_C and D_G represent 65% accuracy for CausalCAM and GradCAM respectively
- **CausalCAM encodes dense structure-rich features in lesser bits, that aid accuracy**

Prabhushankar, Mohit, and Ghassan AlRegib. "Extracting causal visual features for limited label classification." *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021.

Evaluation 3: Progressive Structure-wise Insertion and Deletion

Structure-wise Deletion: Sequentially delete (mask) pixels in an image based on the number of bits used to represent the region





Ideal scenario: The explanation encodes the most important information in the least possible bits

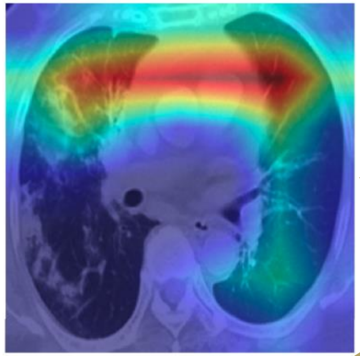
Step 1: Choose a threshold in the explanation (say 0.1) and delete (mask) all the pixels in the original image below the threshold. Pass the masked image through the network and note the change in prediction (if any)

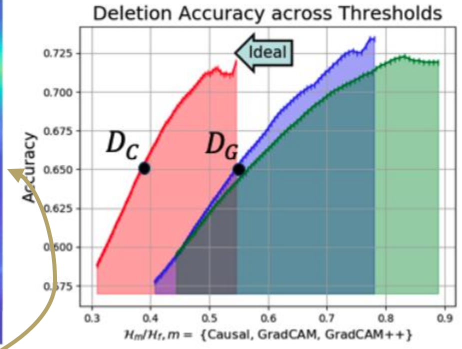
Prabhushankar, Mohit, and Ghassan AlRegib. "Extracting causal visual features for limited label classification." *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021.

Case Study: Intervenability in Interpretability

Evaluation 3: Progressive Structure-wise Insertion and Deletion

Structure-wise Deletion: Sequentially delete (mask) pixels in an image based on the number of bits used to represent the region





Y-axis: Performance accuracy across all ratios

X-axis: Ratio of Huffman encoded masked and original images for all explanations. Smaller the ratio, less is the number of bits encoding the masked image

Ideal scenario: The explanation encodes the most important information in the least possible bits

Step 1: Choose a threshold in the explanation (say 0.1) and delete (mask) all the pixels in the original image below the threshold. Pass the masked image through the network and note the change in prediction (if any)

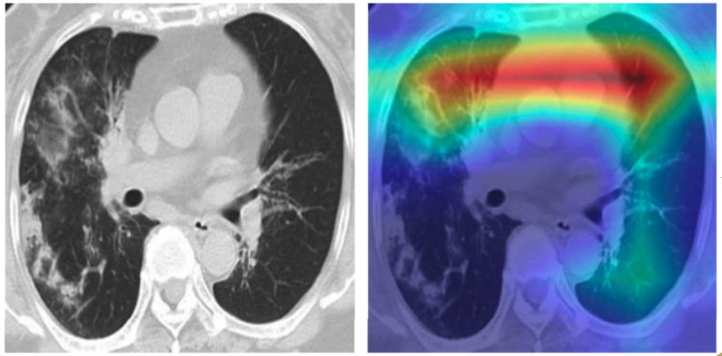
Step 2: Calculate the Huffman code for the original and the masked image. The ratio between the codes of masked and original image is taken on the x-axis and the corresponding accuracy across all images is shown on the y-axis

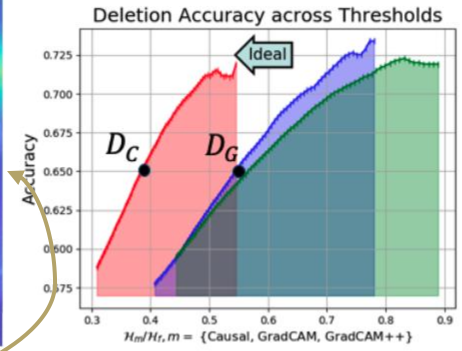
Prabhushankar, Mohit, and Ghassan AlRegib. "Extracting causal visual features for limited label classification." *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021.

Case Study: Intervenability in Interpretability

Evaluation 3: Progressive Structure-wise Insertion and Deletion

Structure-wise Deletion: Sequentially delete (mask) pixels in an image based on the number of bits used to represent the region





Y-axis: Performance accuracy across all ratios

X-axis: Ratio of Huffman encoded masked and original images for all explanations. Smaller the ratio, less is the number of bits encoding the masked image

Ideal scenario: The explanation encodes the most important information in the least possible bits

Step 1: Choose a threshold in the explanation (say 0.1) and delete (mask) all the pixels in the original image below the threshold. Pass the masked image through the network and note the change in prediction (if any)

Step 2: Calculate the Huffman code for the original and the masked image. The ratio between the codes of masked and original image is taken on the x-axis and the corresponding accuracy across all images is shown on the y-axis

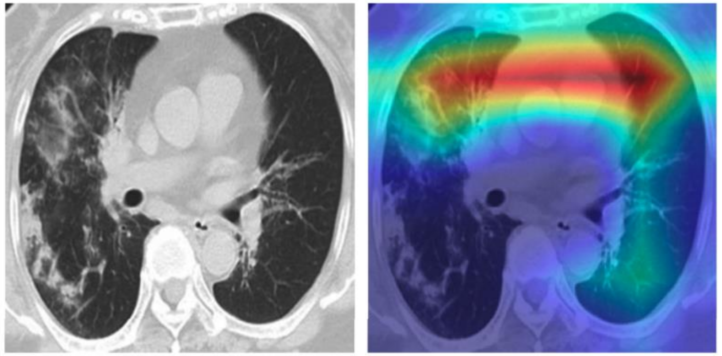
Step 3: Repeat across thresholds

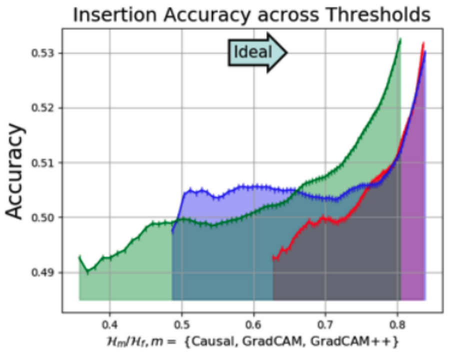
Prabhushankar, Mohit, and Ghassan AlRegib. "Extracting causal visual features for limited label classification." *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021.

Case Study: Intervenability in Interpretability

Evaluation 3: Progressive Structure-wise Insertion and Deletion

Structure-wise Insertion: Sequentially add (insert) pixels in an image based on the number of bits used to represent the region





Ideal scenario: The explanation encodes the most important information in the least possible bits

CausalCAM in Red[1]
GradCAM in Purple
GradCAM++ in Green

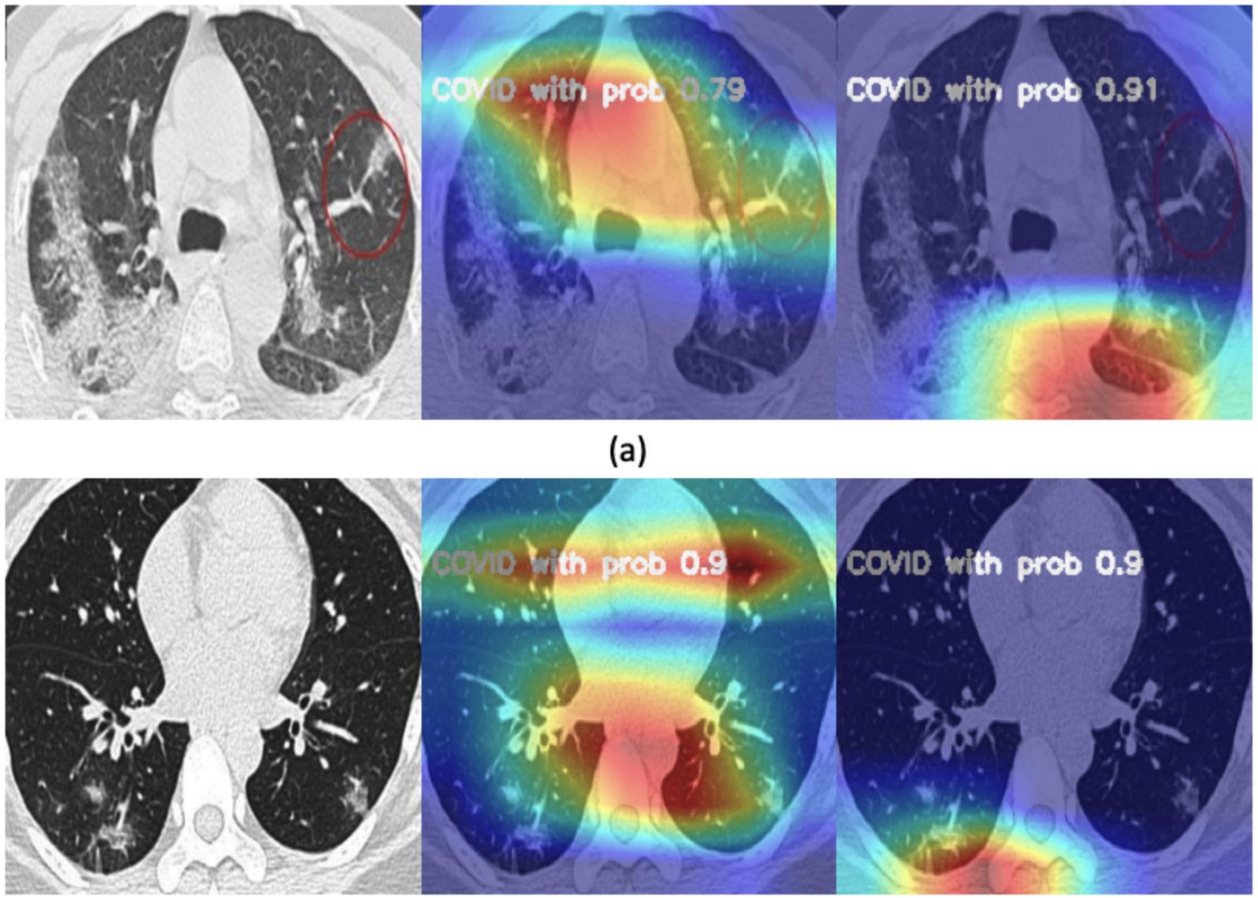
- **CausalCAM encodes dense structure-rich features in at the lowest threshold, that aid accuracy**

Prabhushankar, Mohit, and Ghassan AlRegib. "Extracting causal visual features for limited label classification." *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021.

Case Study: Intervenability in Interpretability

Evaluation 3: Progressive Structure-wise Insertion and Deletion

Structure-wise insertion and deletion can sometimes promote adversarial explanations



- Best explanations according to structure-wise insertion and deletion.
- Corroborated by high probabilities

Prabhushankar, Mohit, and Ghassan AlRegib. "Extracting causal visual features for limited label classification." *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021.

Case Study: Intervenability in Interpretability

Pros and Cons

Evaluation 1: Explanation heatmap masking

- **Pro**: Structures are visible in the explanations
- **Con**: Encourages large non-fine grained explanations

Evaluation 2: Pixel-wise insertion and deletion

- **Pro**: Progressively assigns importance to pixels
- **Con**: Encourages unrealistic and dispersed explanations

Evaluation 3: Structure-wise insertion and deletion

- **Pro**: Encourages structures while progressively assigning importance to structures based on information bits
- **Pro**: Other human-centric measures including SSIM, saliency etc. can be used on x-axis

- **Con**: Encourages causal (and sometimes adversarial) explanations without considering context information

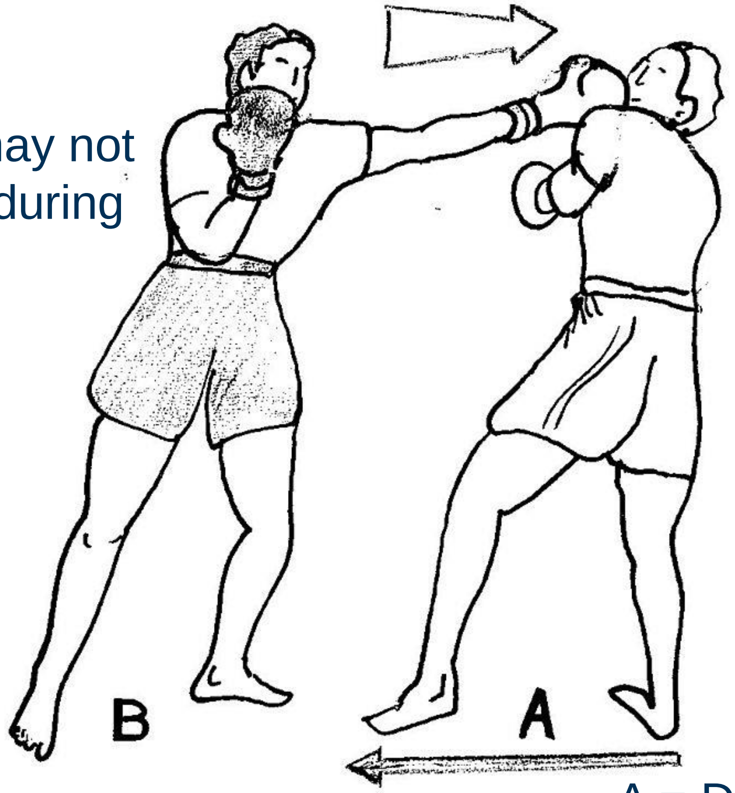
Inferential Machine Learning
Part 5: Conclusions and Future Directions

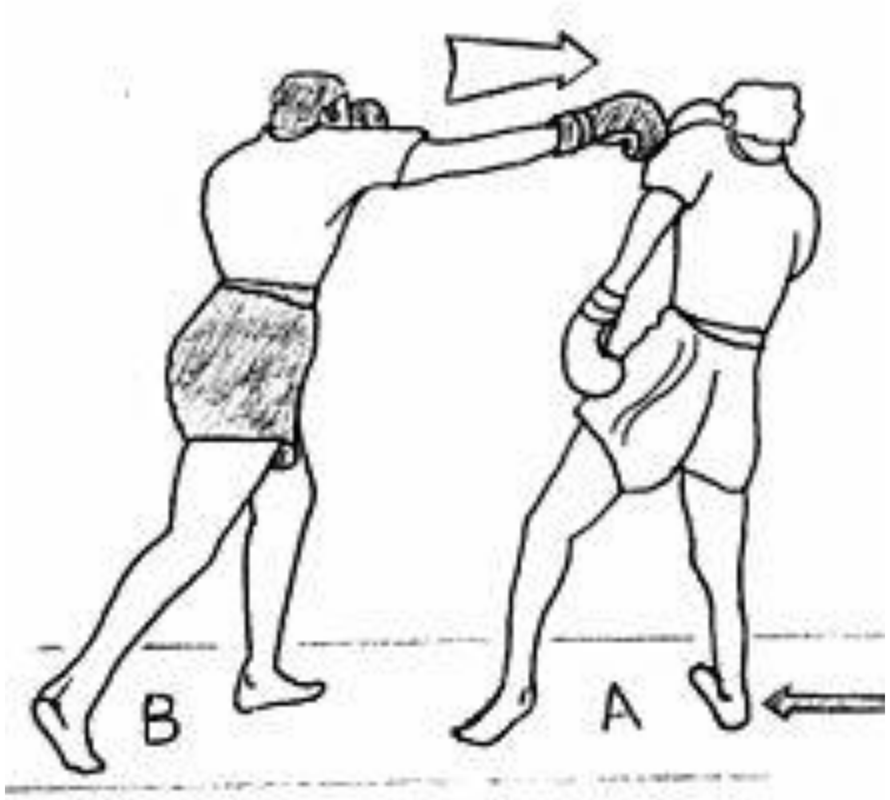
Memes to Wrap it Up

Overcoming Challenges at Training

Novel data packs a 1-2 punch!

Novel data may not be available during training

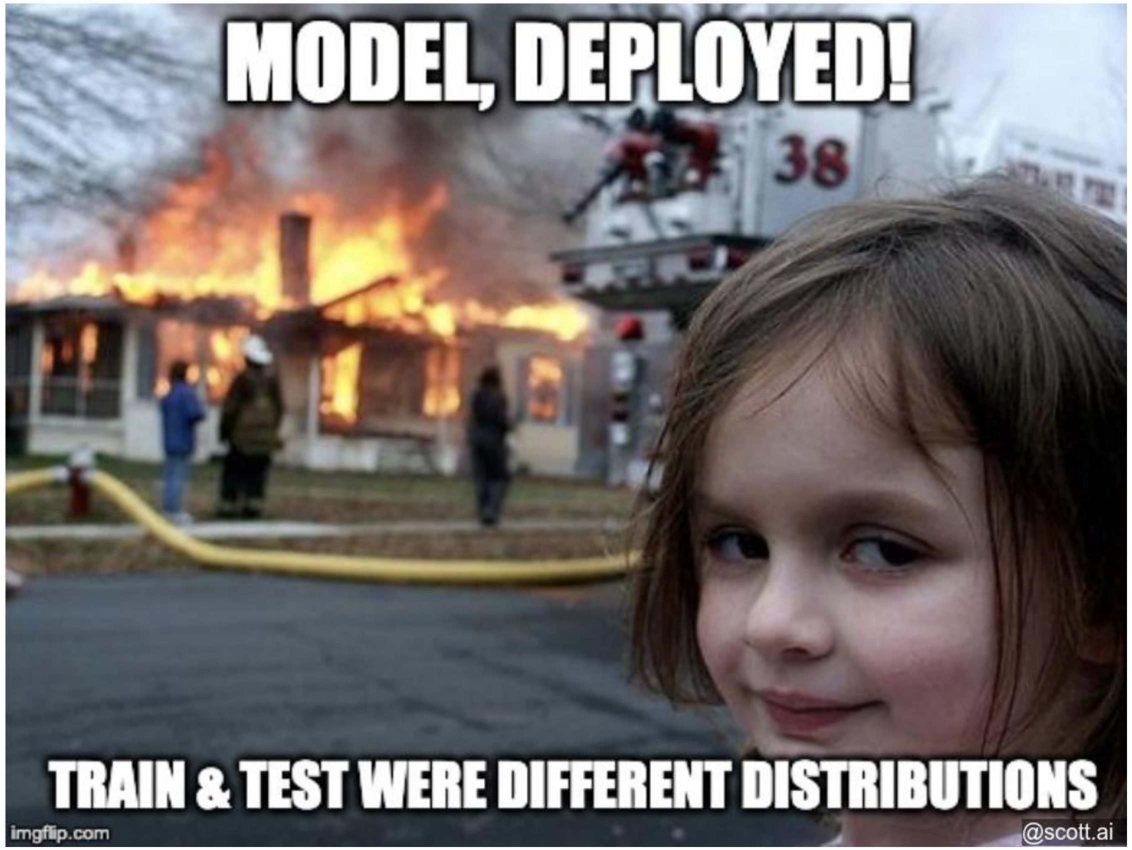




Even if available, novel data does not easily fit into either the earlier or later stages of training

A = Deep Neural Networks
B = Novel data

Robustness at Inference

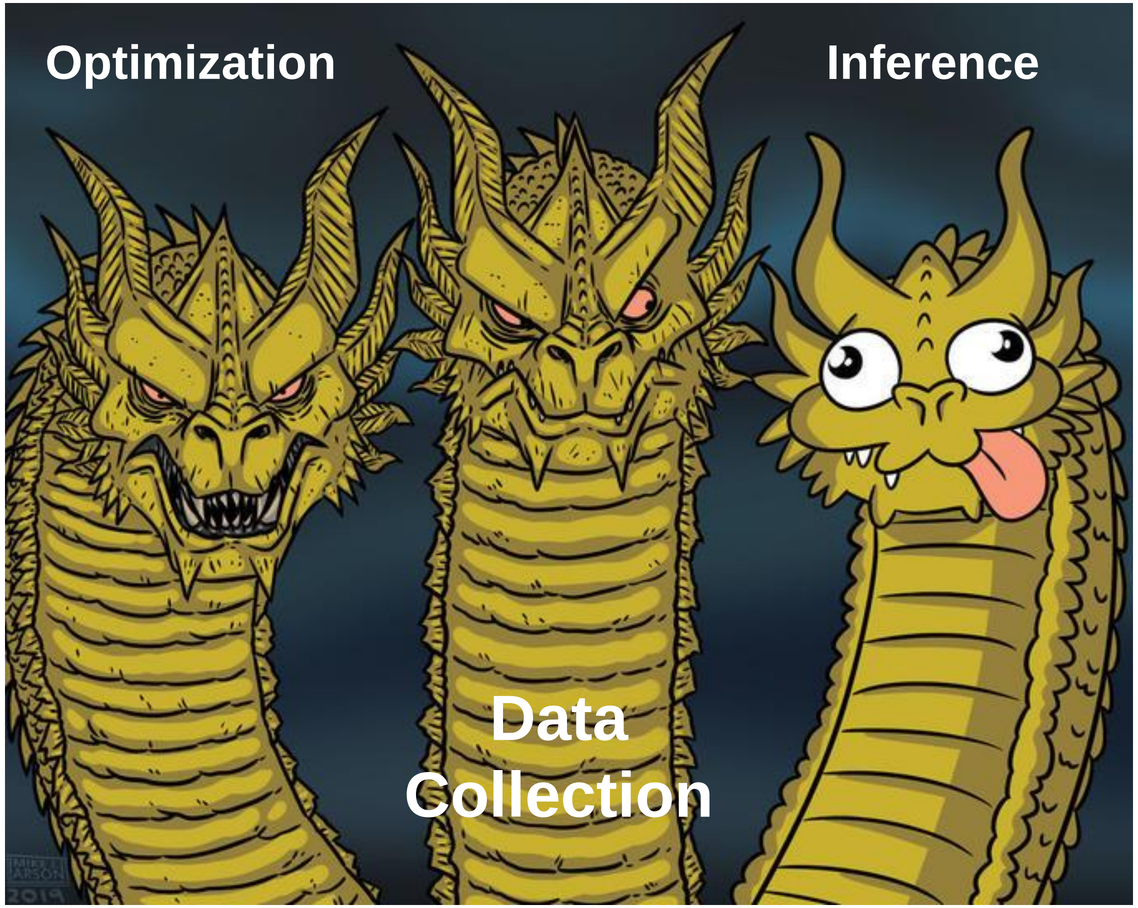


Cannot depend on training to construct robust models

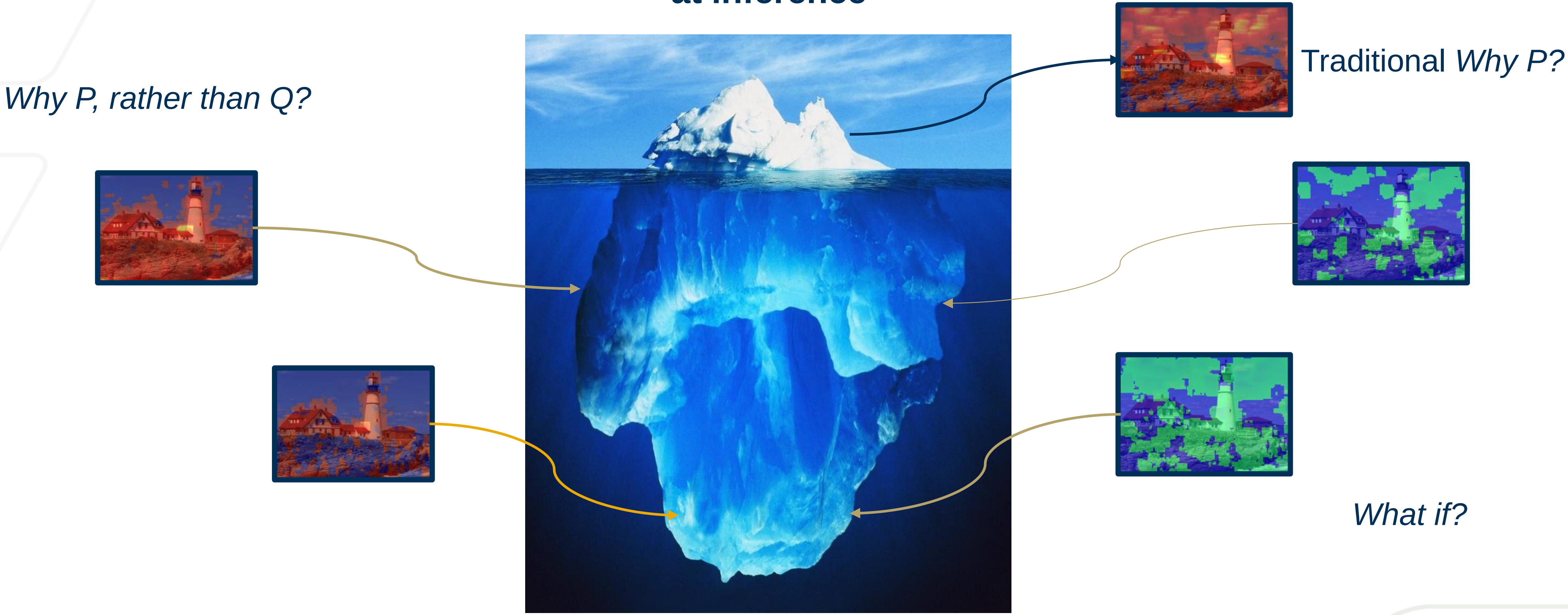
Memes to Wrap it Up

Research in the Inferential Stage of Neural Networks

Existing research on robustness focuses on data collection and optimization



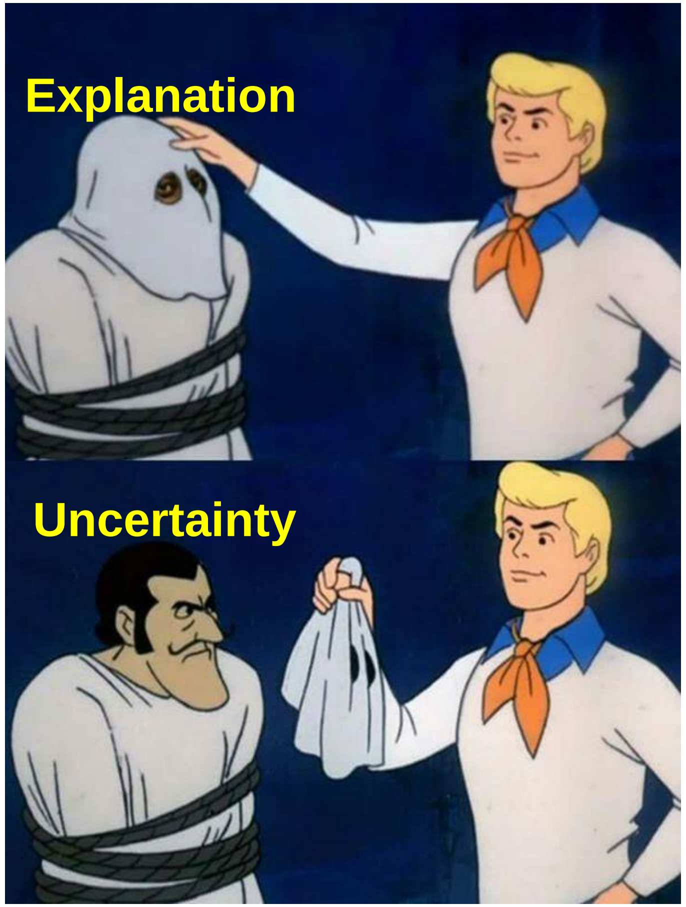
Trained Neural Networks have a wealth of implicit stored knowledge, waiting to be extracted at inference



Memes to Wrap it Up

Explainability Research is Just Uncertainty Research

Explanatory Evaluation reduces Uncertainty



Key Takeaways

Role of Gradients

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

Future Directions

Research at Inference Stage

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

References

Gradient-based Works

- Explainability [1, 2]
- Out-of-distribution Detection [3]
- Adversarial Detection [4]
- Anomaly Detection [5]
- Corruption Detection [3]
- Misprediction Detection [6]
- Causal Analysis [7]
- Open-set Recognition [8]
- Noise Robustness [9]
- Uncertainty Visualization [10]
- Image Quality Assessment [11, 12]
- Saliency Detection [13]
- Novelty Detection [14]
- Disease Severity Detection [15]

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

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

[3] J. Lee, C. Lehman, M. Prabhushankar, and G. AlRegib, "Probing the Purview of Neural Networks via Gradient Analysis," in IEEE Access, Mar. 21 2023.

[4] 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.

[5] 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.

[6] Prabhushankar, M., & AlRegib, G. (2024, August). Counterfactual Gradients-based Quantification of Prediction Trust in Neural Networks. In 2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 529-535). IEEE.

[7] M. Prabhushankar, and G. AlRegib, ”Extracting Causal Visual Features for Limited Label Classification,” in IEEE International Conference on Image Processing (ICIP), Sept. 2021.

[8] Lee, Jinsol, and Ghassan AlRegib. "Open-Set Recognition With Gradient-Based Representations." *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021.

[9] 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

[10] Prabhushankar, M., & AlRegib, G. (2024). Voice: Variance of induced contrastive explanations to quantify uncertainty in neural network interpretability. IEEE Journal of Selected Topics in Signal Processing.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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