Inferential Machine Learning: Towards Human-collaborative Vision and Language Models

Ghassan AlRegib, PhD
Professor

Mohit Prabhushankar, PhD
Postdoctoral Fellow

Xiaoqian Wang, PhD
Assistant Professor

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

Electrical and Computer Engineering
Purdue University
joywang@purdue.edu

Feb. 26, 2025 – Philadelphia, USA

Tutorial Materials
Accessible Online

SCAN ME

https://alregib.ece.gatech.edu/courses-and-tutorials/aaai-2025-tutorial/

{alregib, mohit.p}@gatech.edu, joywang@purdue.edu

AAAI 2025 Tutorial

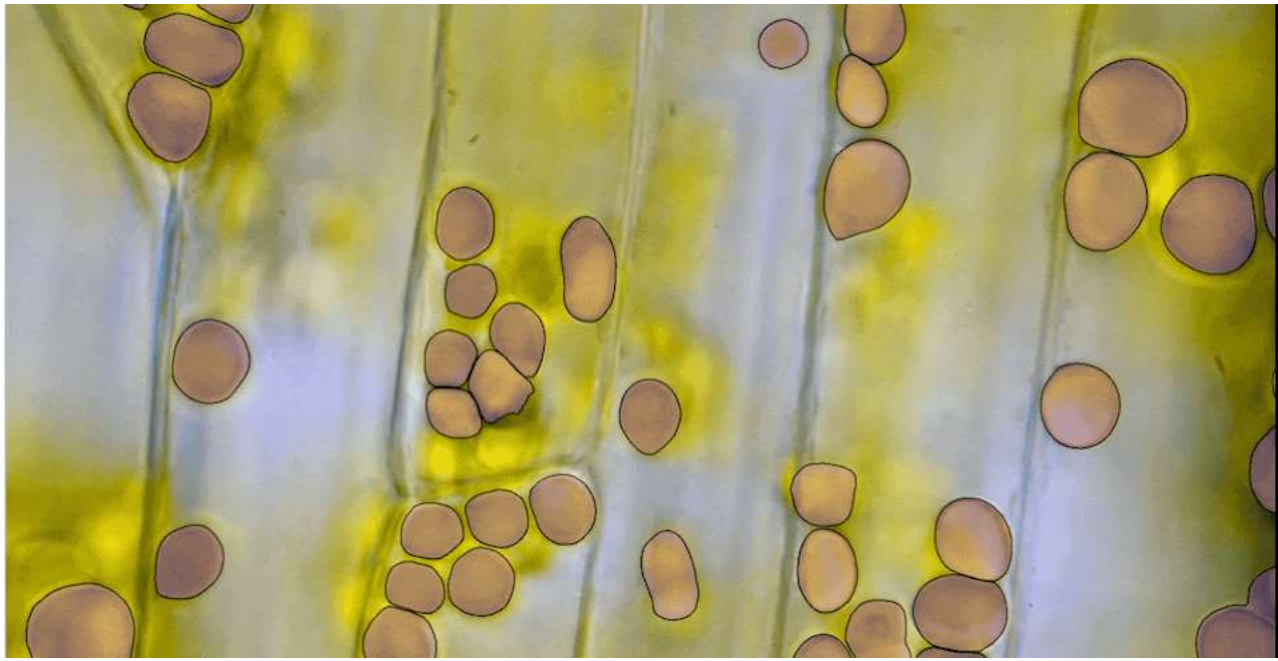
Home > Courses And Tutorials > AAAI 2025 Tutorial

Inferential Machine Learning: Towards Human-collaborative Vision and Language Models

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).

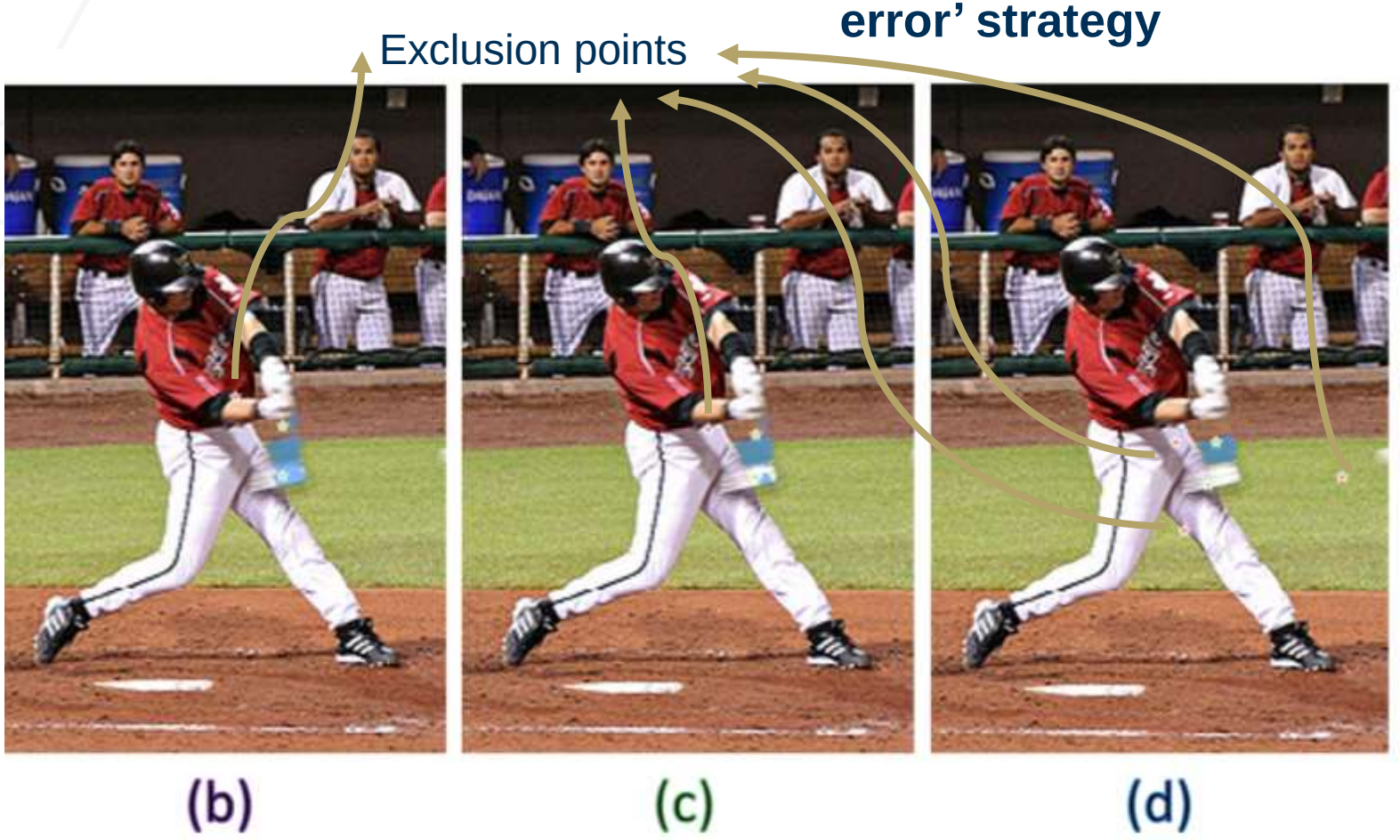
Foundation Models

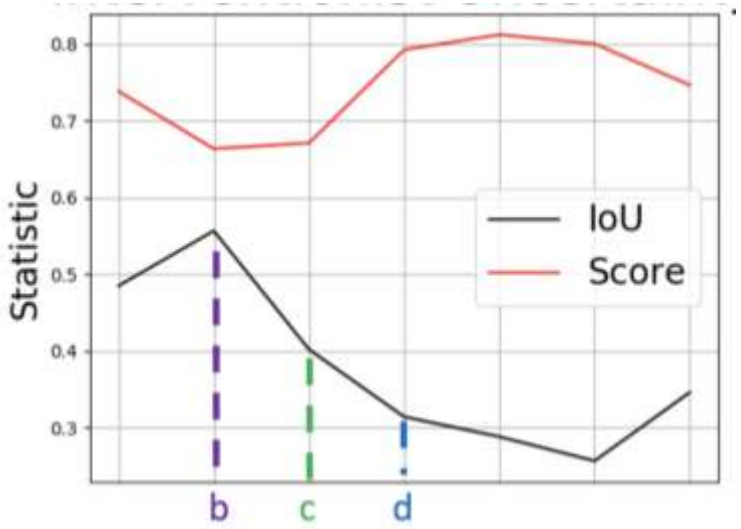
‘Trial and Error’ Interventions in Segment Anything Model

SCAN ME

PointPrompt Dataset

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

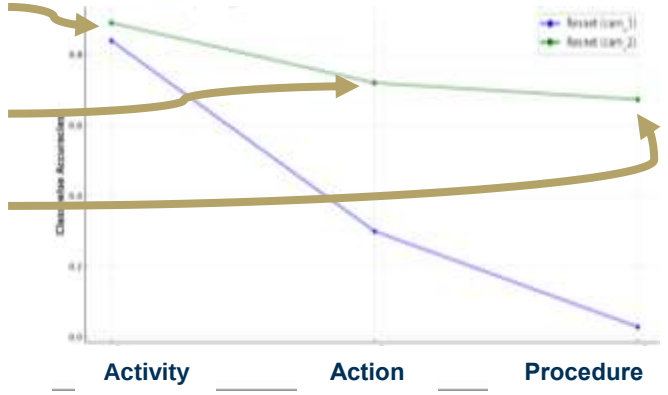
VLMs (encoder finetuned on dataset) fail when recognizing fine-grained hierarchical activities

Hierarchical Activity Recognition

Ground Truth **Prediction**

Other findings:

SCAN ME



OLIVES

Georgia Tech

Foundation Models

Vision-Language Models are sensitive to experimental setup

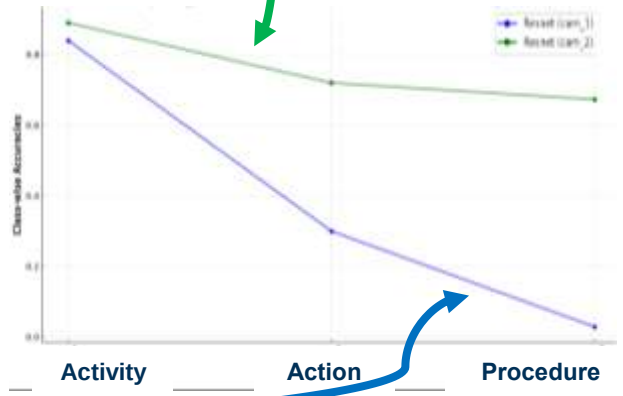
DARai Dataset

SCAN ME

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

Other findings:

SCAN ME



Foundation Models

Vision-Language Models are Biased towards Societal Stereotypes

SCAN ME

Debiasing VLMs

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

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

Foundation Models

Requirements and Challenges for Deep Learning

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

Novel data sources:

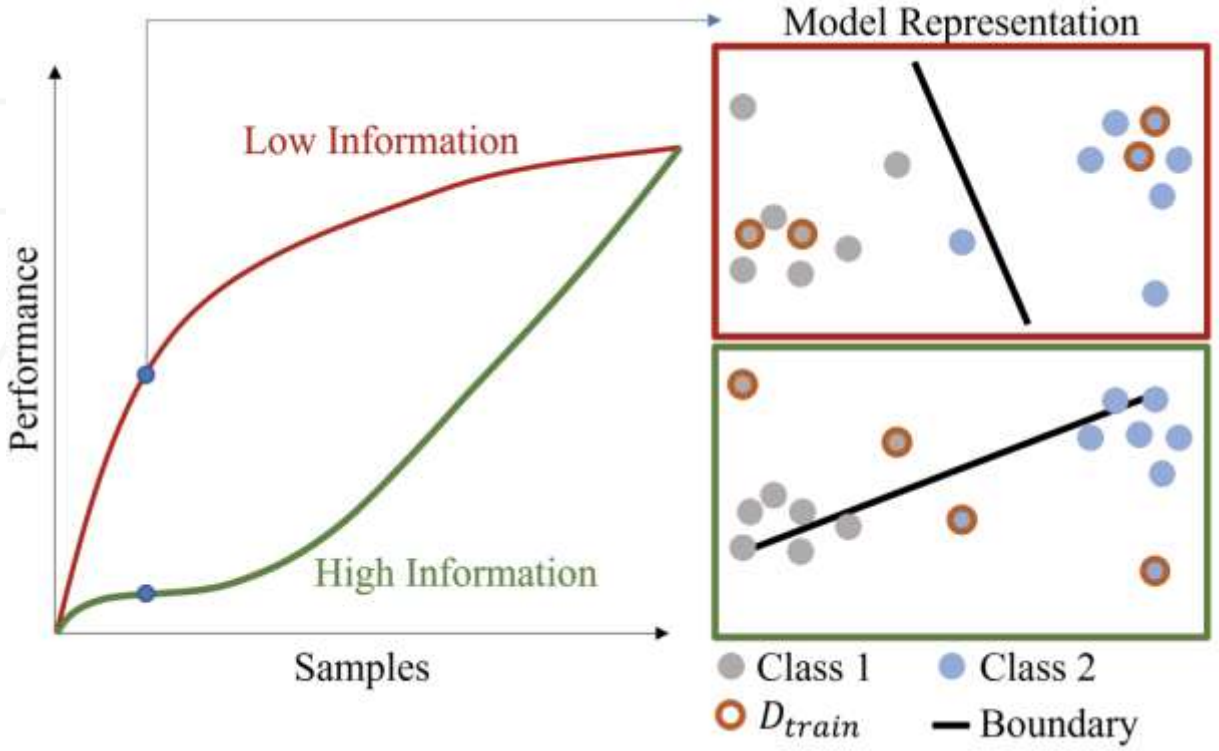
- Test distributions
- Anomalous data
- Out-Of-Distribution data
- Adversarial data
- Corrupted data
- Noisy data
- New classes
- …

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

Deep Learning at Training

Overcoming Challenges at Training: Part 1

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



- 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

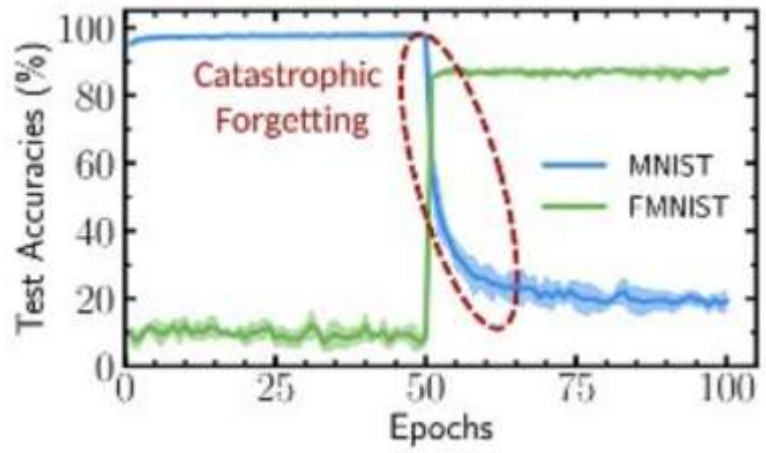
Novel samples = Most Informative

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

Deep Learning at Training

Overcoming Challenges at Training: Part 2

Subsequent training must not focus only on novel data

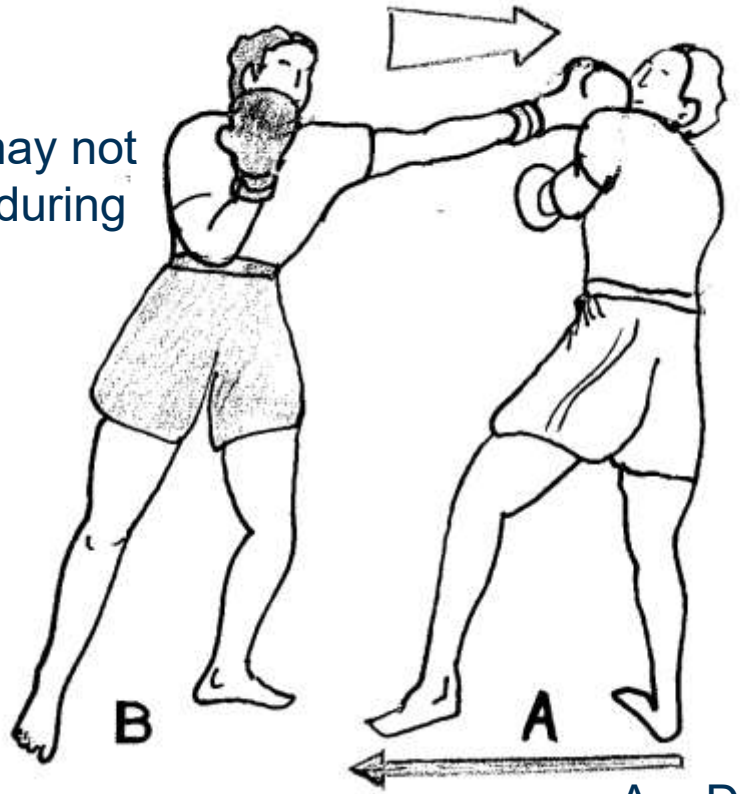


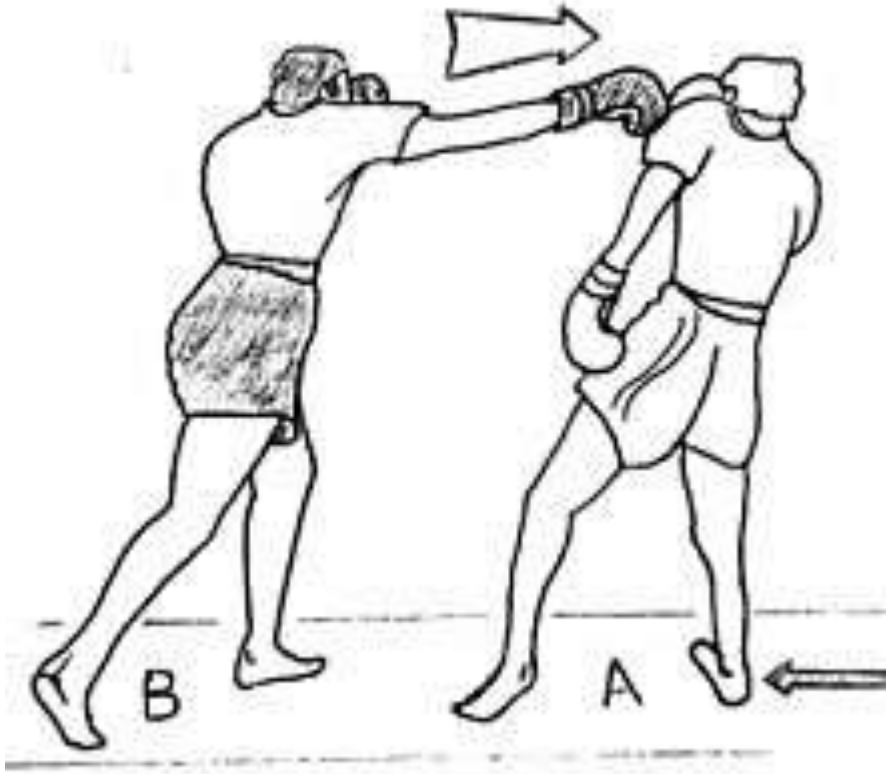
- The model performs well on the new scenarios, **while forgetting the old scenarios**
- Several techniques exist to overcome this trend
- However, they affect the overall performance in large-scale settings
- It is not always clear **if and when** to incorporate novel scenarios in training

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

Foundation Models at Inference

Overcoming Challenges at Inference

We must handle novel data at Inference!!

Novel data sources:

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

Model Train

At Inference

Objective

Objective of the Tutorial

To discuss methodologies that promote 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
- Part 4: Fairness Interventions

Inferential Machine Learning

Part I: Inference in Neural Networks

PURDUE UNIVERSITY

Georgia Tech

Objective

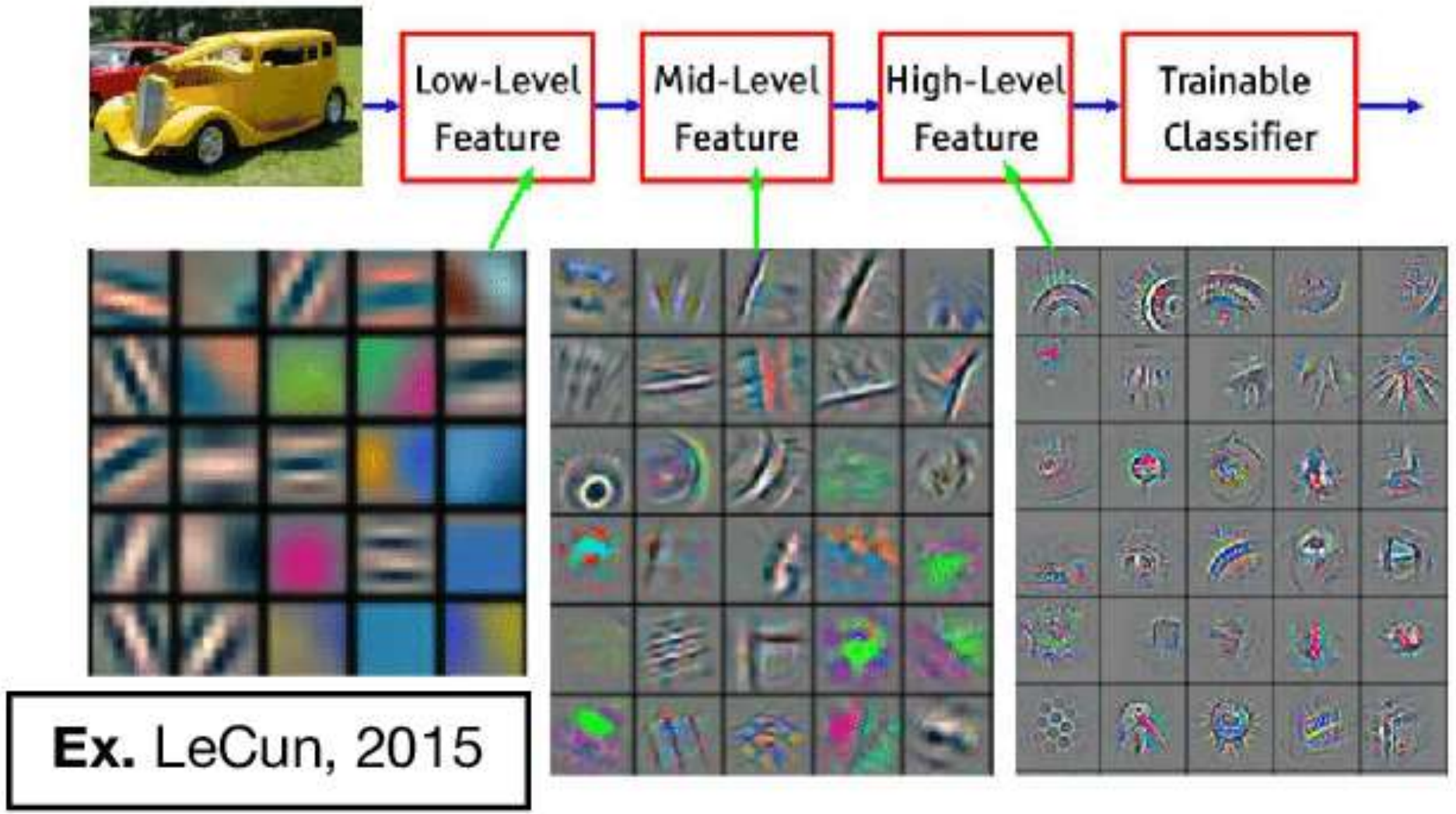
Objective of the Tutorial

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: Fairness Interventions

Deep Learning

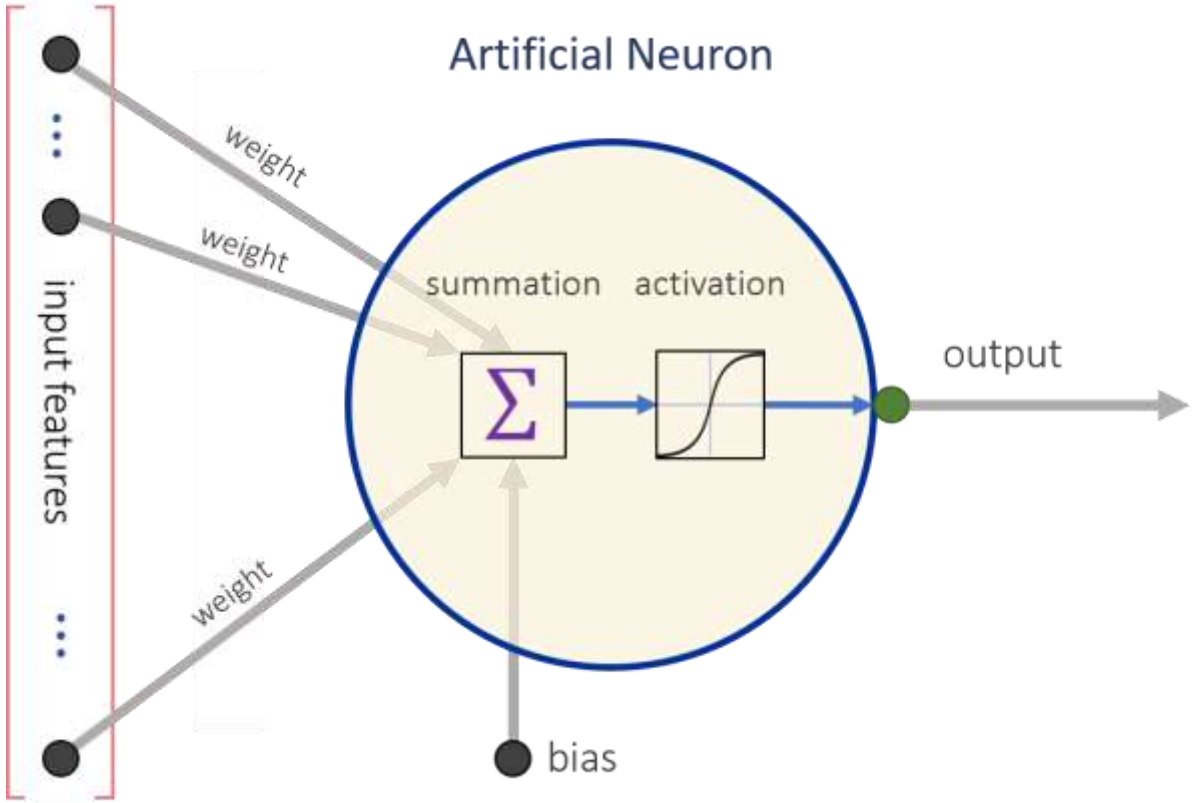
Overview



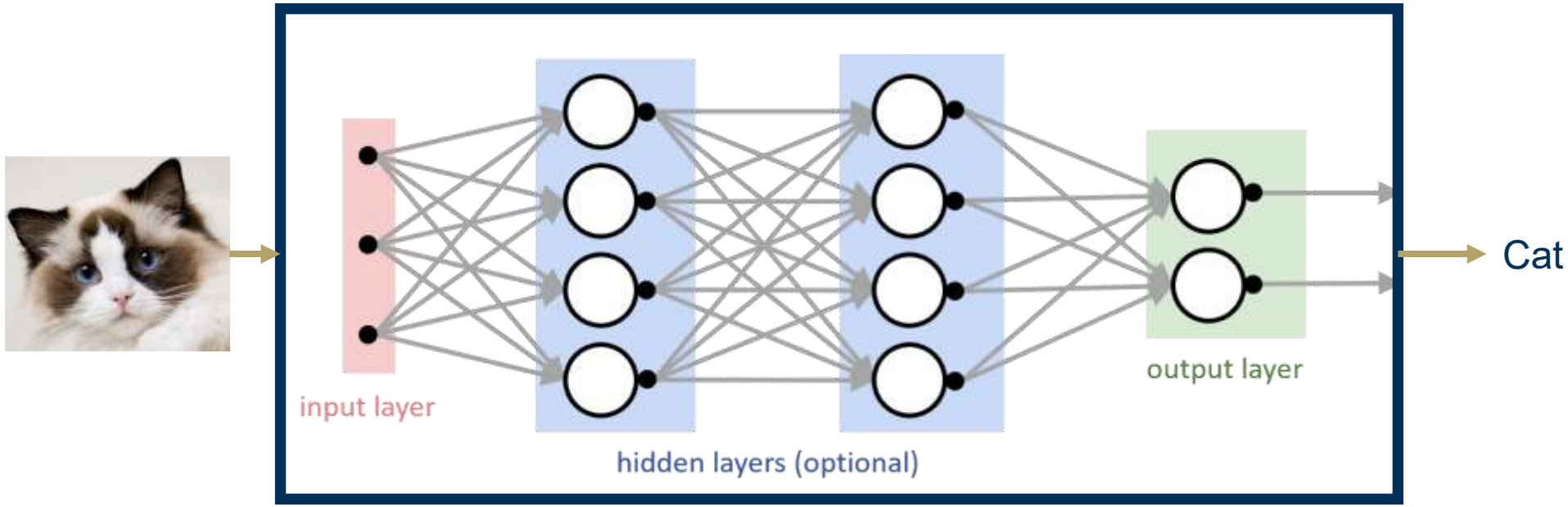
The underlying computation unit is the Neuron

Artificial neurons consist of:

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



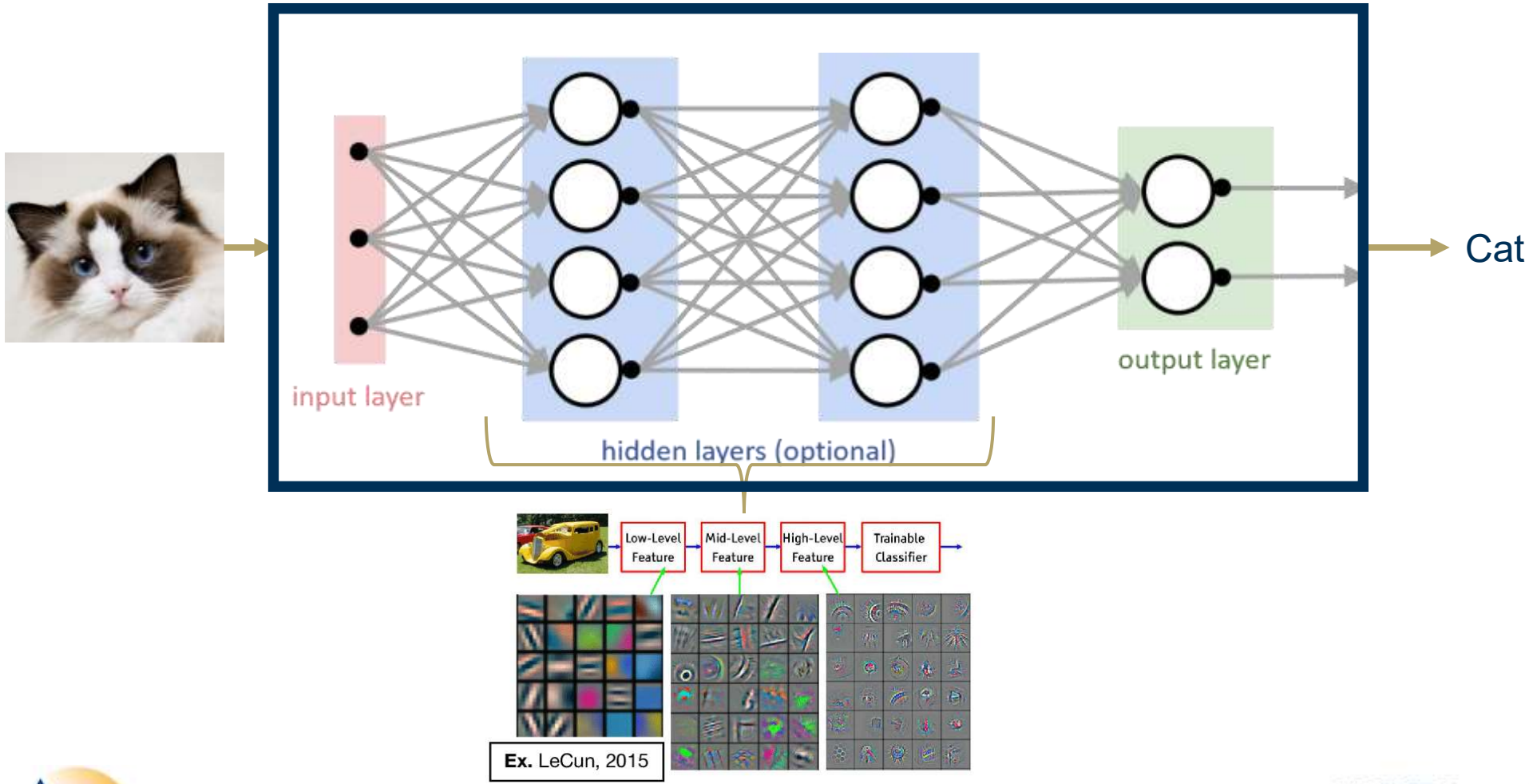
Neurons are stacked and densely connected to construct ANNs



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 $1 \ldots K-1$)

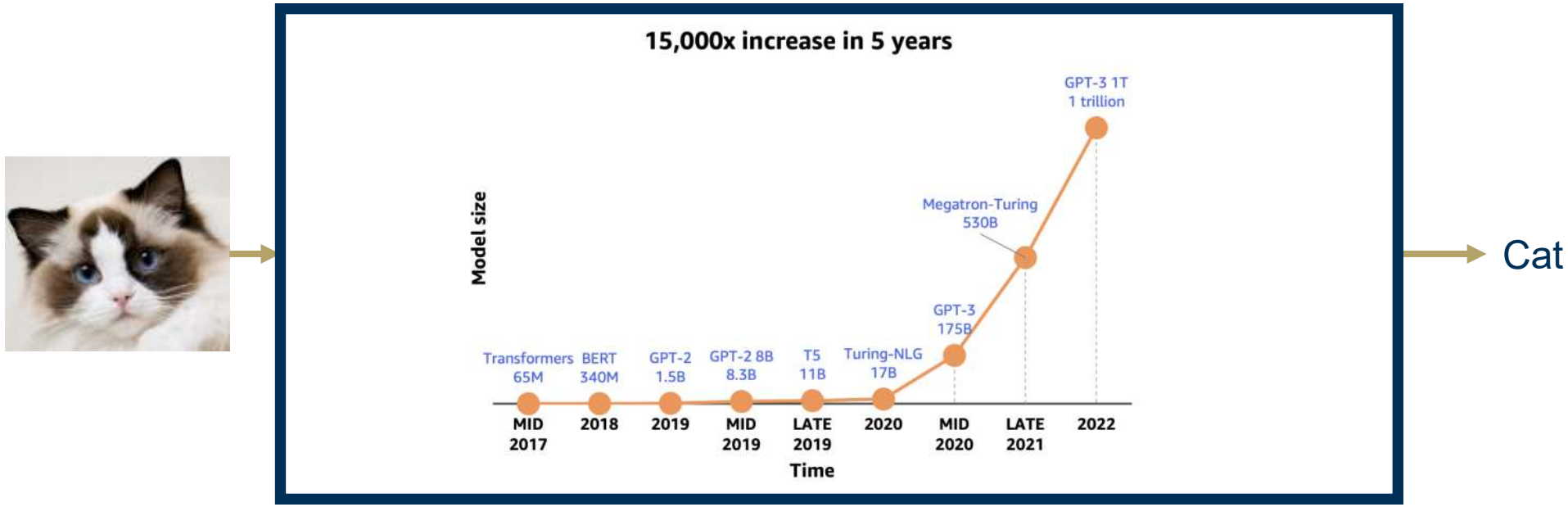
Stationary property of images allow for a small number of convolution kernels



Deep Deep Deep Deep Deep … Learning

Recent Advancements

Transformers, Large Language Models and Foundation Models



Primary reasons for advancements:

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

Foundation Models

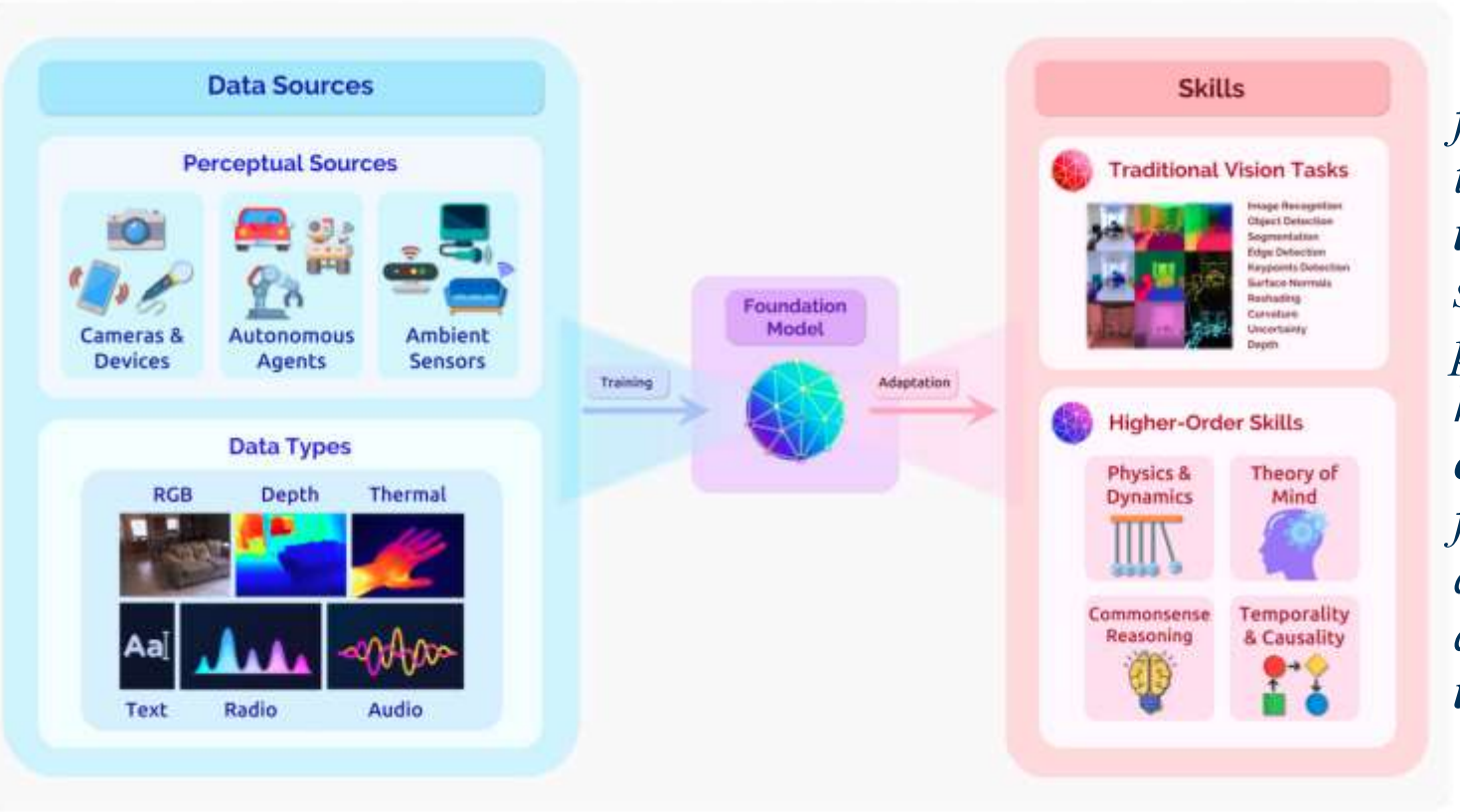
Origin of the term Foundation Models

- **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

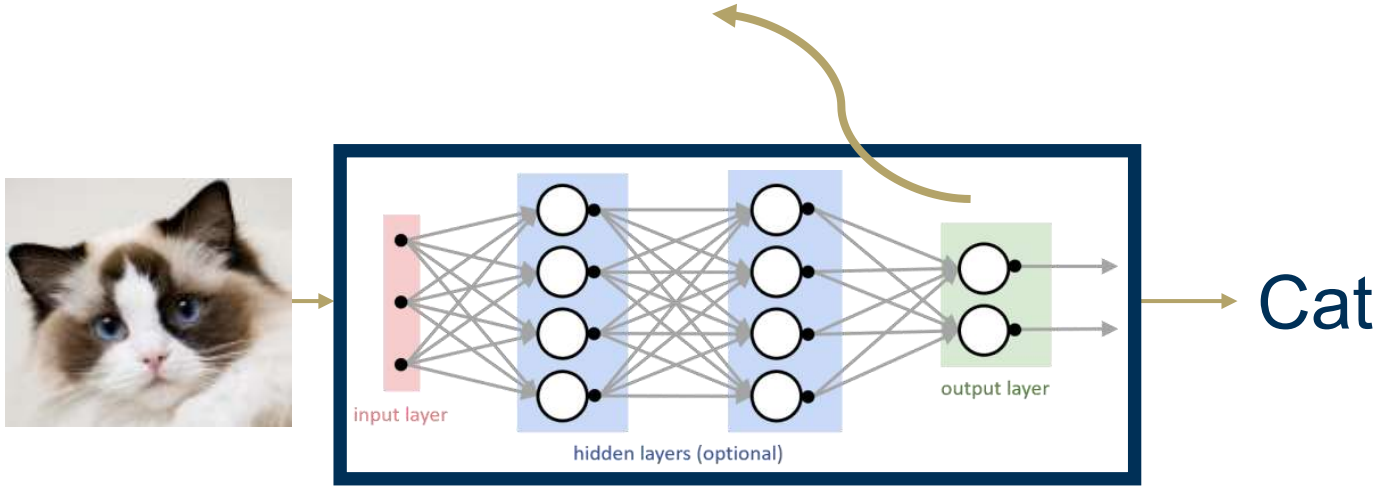
What, Where, and When is Inference?

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
- …

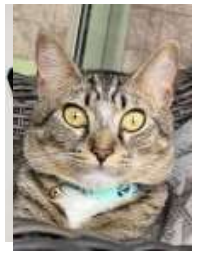
→ Trained Model at Testing → Cat, Cat, Cat

→ Trained Model at Deployment → Cat

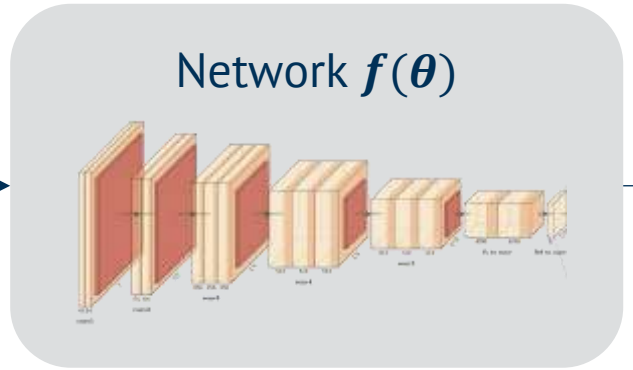
Deep Learning at Inference

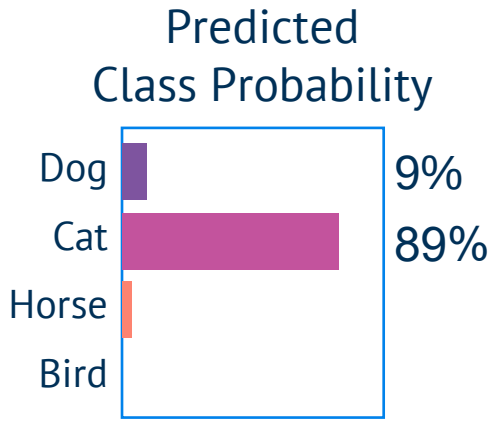
Application: Classification

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



x





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

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

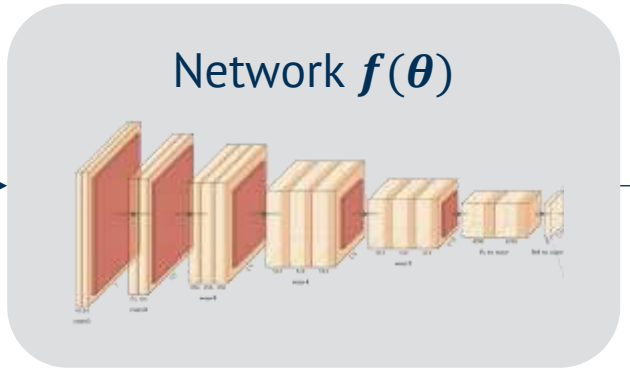
If $x \in \chi$, the data is **not novel**

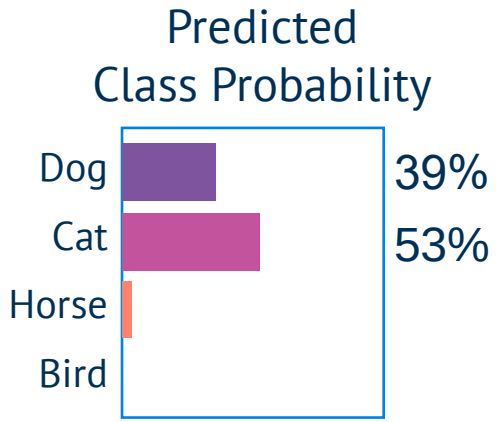
Deep Learning at Inference

Application: Robust Classification

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

x'





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

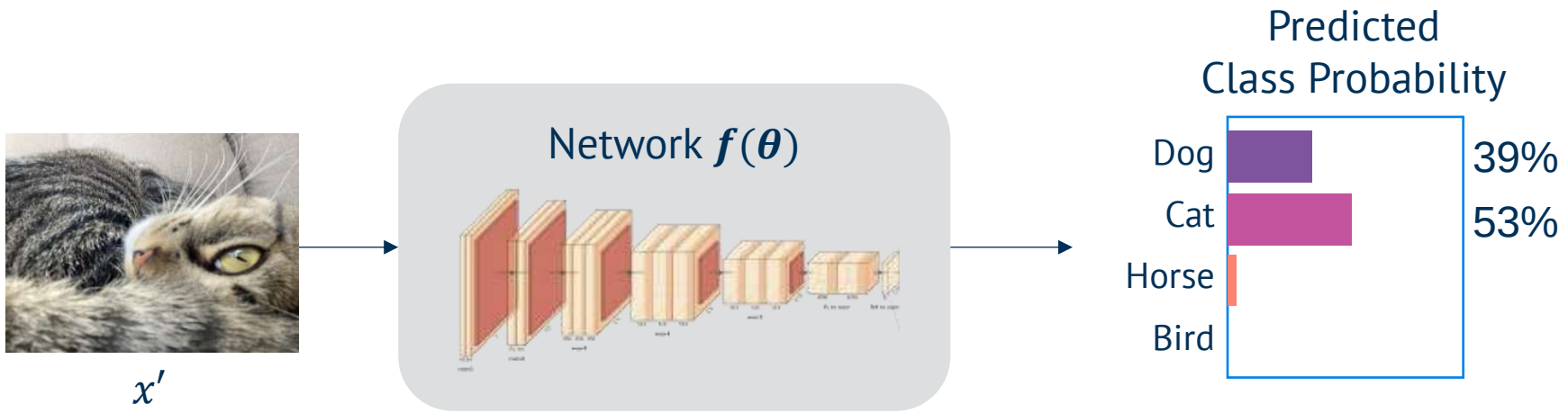
$\hat{y}$ = Logits
y = Predicted Class
$p(\hat{y})$ = Probabilities
$f(\cdot)$ = Trained Network
χ = Training data
ϵ = Noise

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

Deep Learning at Inference

Application: Robust Classification

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



To achieve robustness at Inference, we need the following:

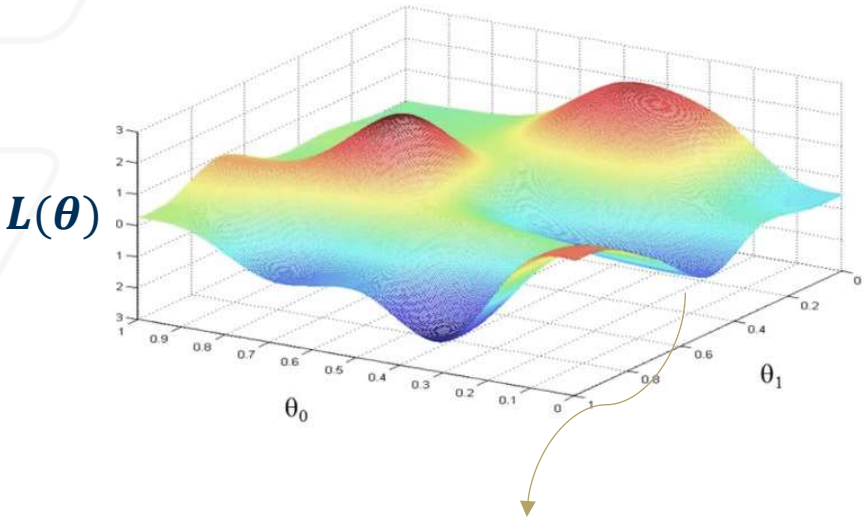
- **Information** provided by the novel data as **a function of training distribution**
- Methodology to **extract information** from novel data
- **Techniques** that utilize the information from novel data

Why is this Challenging?

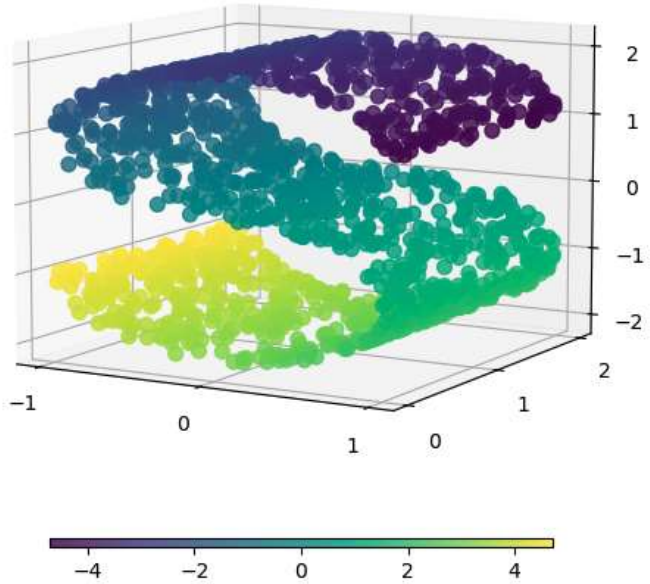
Challenges at Inference

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)

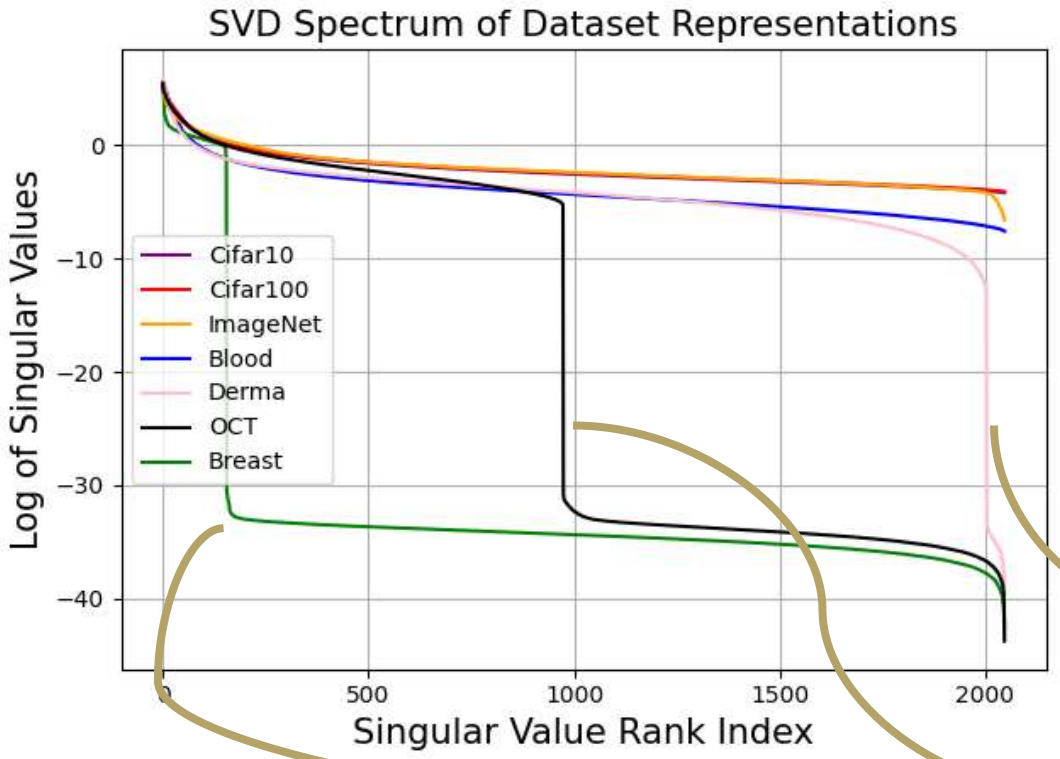
Challenges at Inference

Manifold evaluation at Test-Time Inference without Labels

SCAN ME

Hierarchical Constrained Contrastive Learning

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.

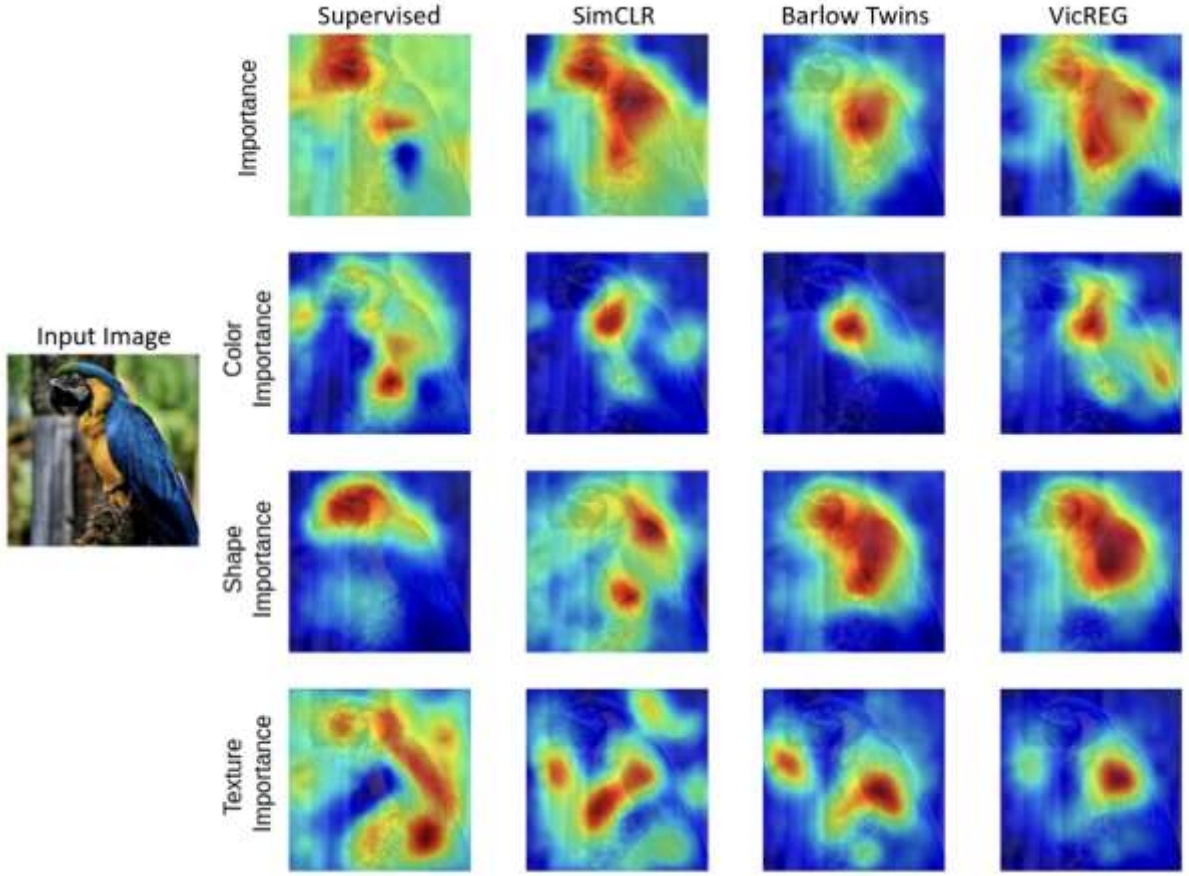
Challenges at Inference

Manifold evaluation at Test-Time Inference without Labels

SCAN ME

Perceptual Components in Self-Supervised Learning

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.*

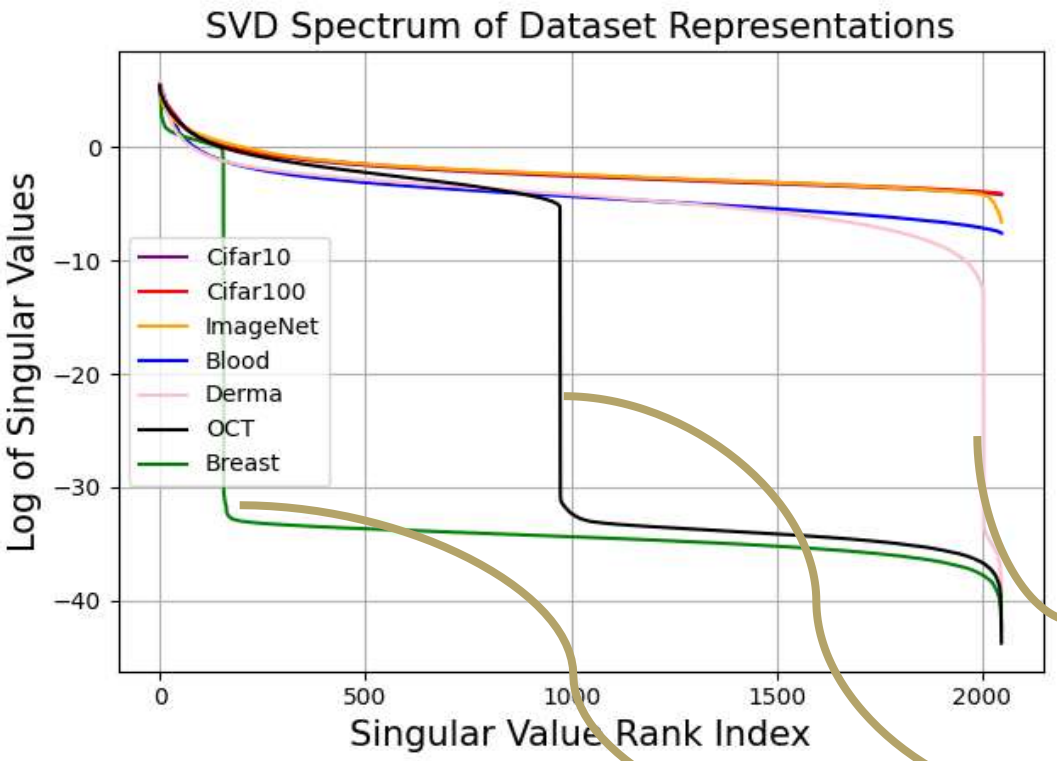
Challenges at Inference

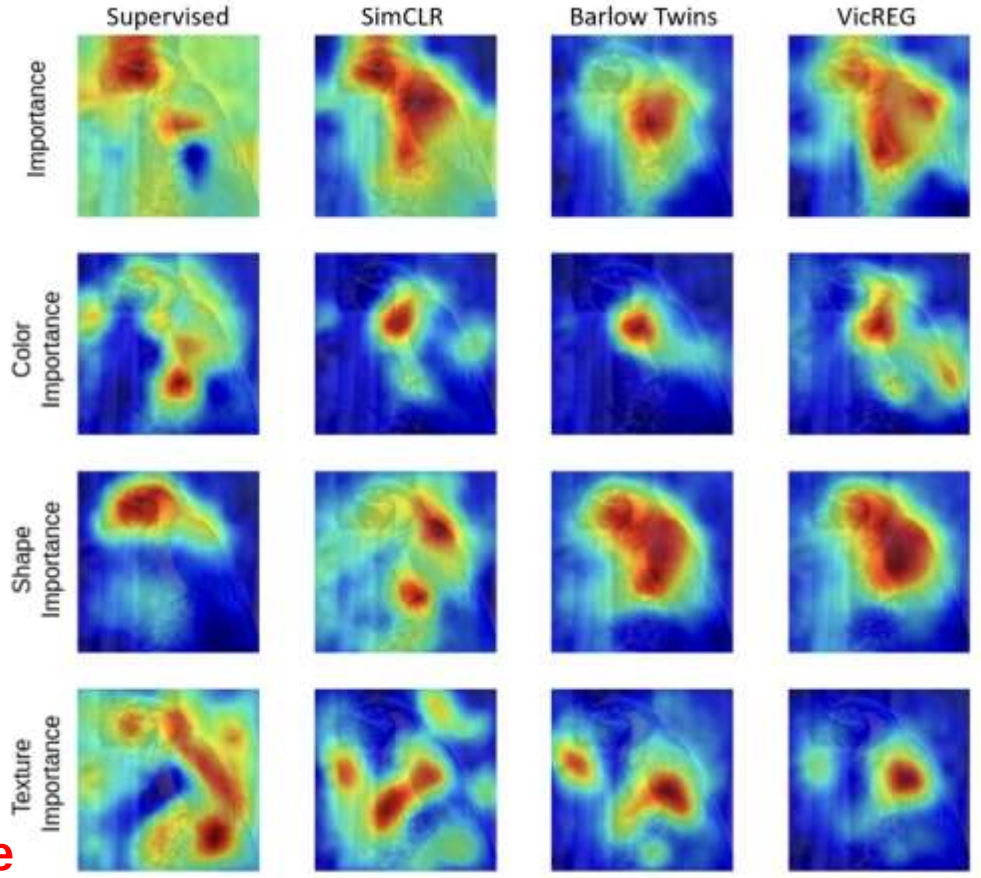
Deployment Inferential Evaluation

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

OLIVES

Georgia Tech

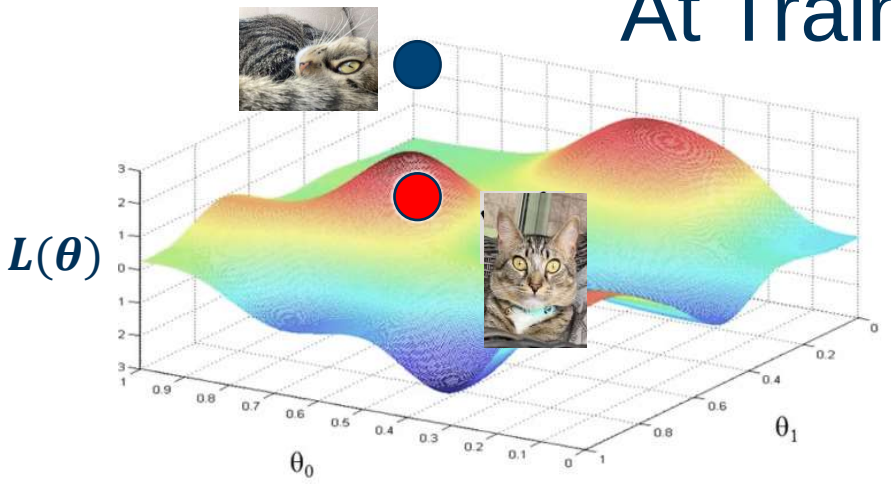
Challenges at Inference

Deployment Inferential Evaluation

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

At Inference

At Training

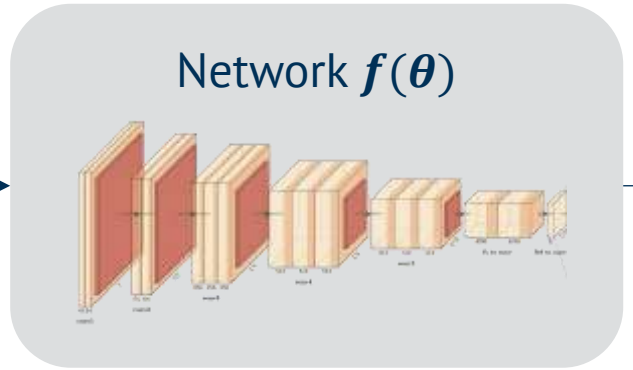


At training, we have access to all training data.

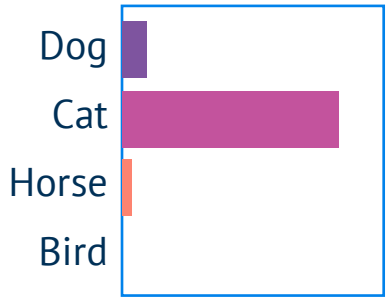
Information at Inference

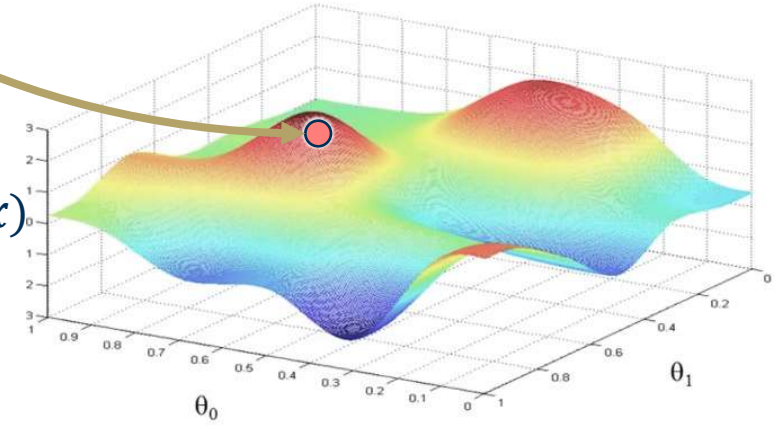
Fisher Information

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



Predicted
Class Probability





Likelihood function

Fisher Information

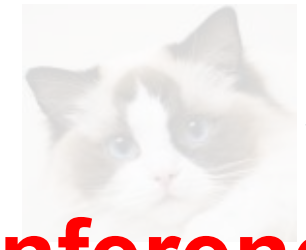
$$I(\theta) = Var(\frac{\partial}{\partial\theta} l(\theta|x))$$

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



Network $f(\theta)$

Predicted
Class Probability

Dog
Cat
Horse
Bird

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

$l(\theta|x)$

Likelihood function

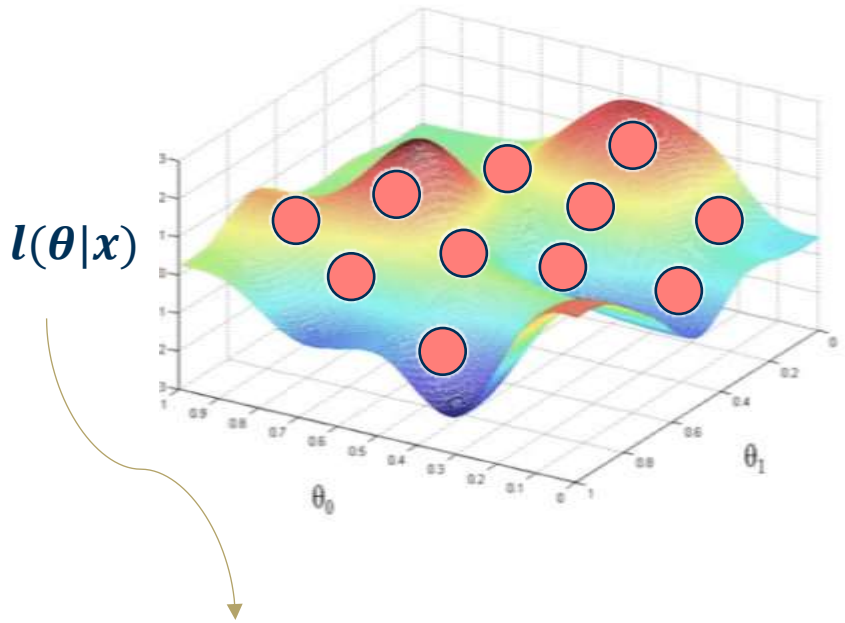
$$I(\theta) = Var(\frac{\partial}{\partial\theta} l(\theta|x))$$

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

Information at Inference

Gradients as Fisher Information

Gradients infer information about the statistics of underlying manifolds



Likelihood function instead of loss manifold

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

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

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

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

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

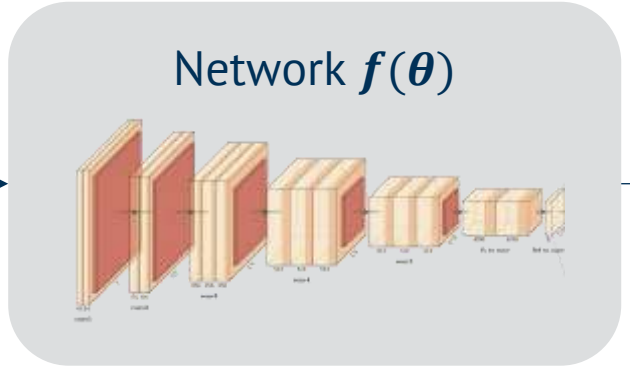
Kwon, Gukyeong, et al. "Backpropagated gradient representations for anomaly detection." *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer International Publishing, 2020.

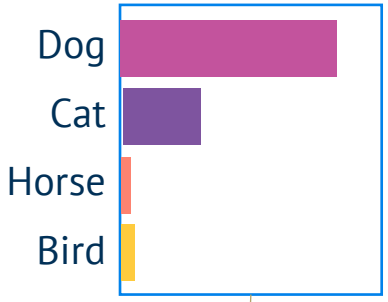
Information at Inference

Case Study: Gradients as Fisher Information in Explainability

Gradients infer information about the statistics of underlying manifolds

x



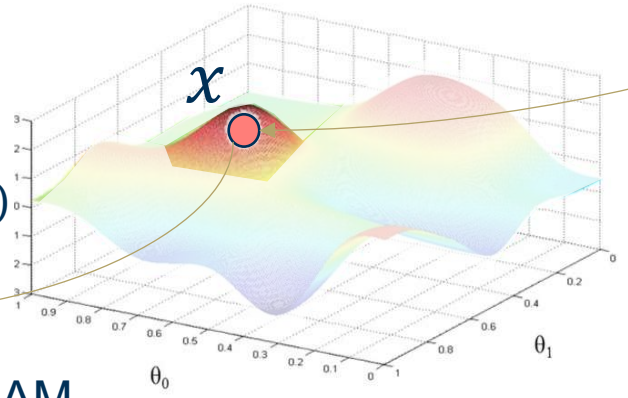


Local information (specific to x) is sufficient!

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

Feature attribution via GradCAM

$l(\theta|x)$



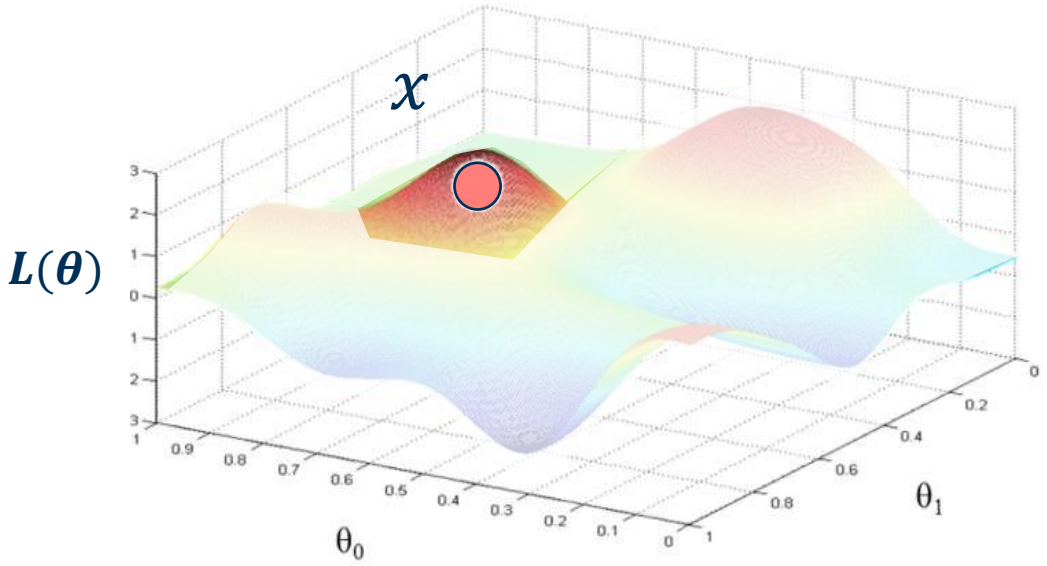
Hence, gradients draw information from the underlying distribution as learned by the network weights!

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

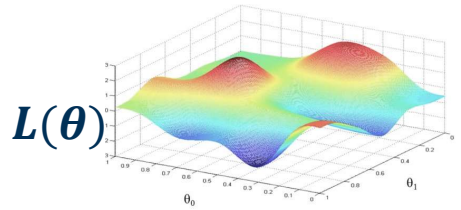
Gradients at Inference

Local Information

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



Ideal

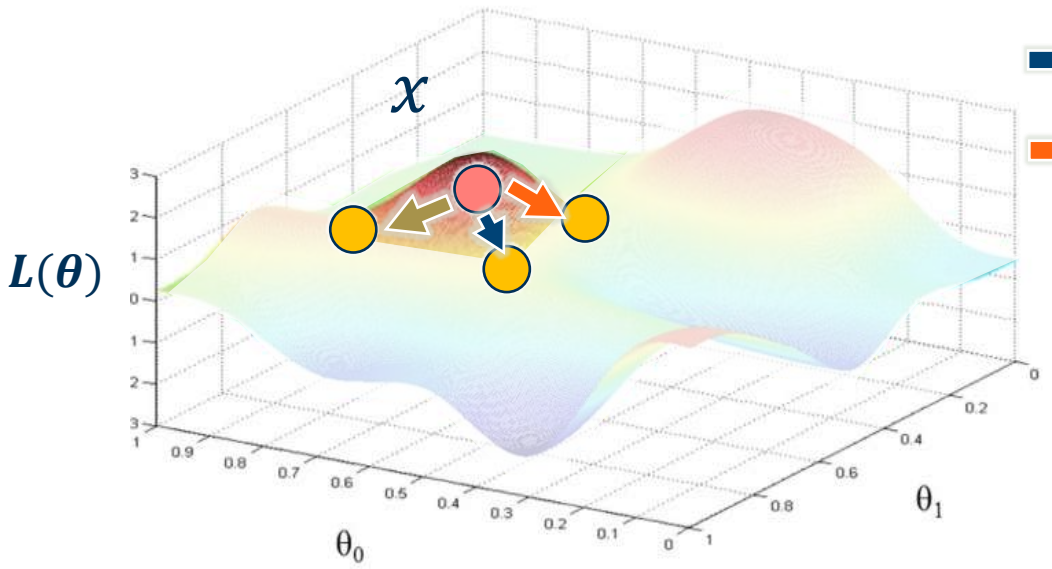


$\alpha \nabla_\theta L(\theta)$ provides local information up to a small distance α away from x

Gradients at Inference

Direction of Steepest Descent

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



Path 1?

Path 2?

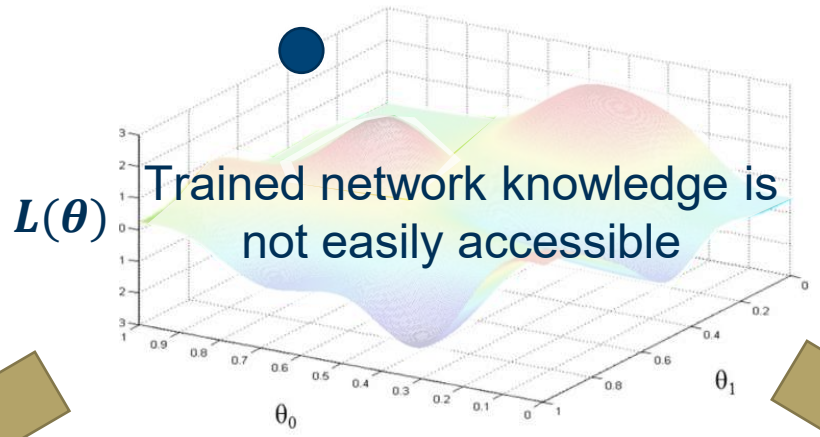
Path 3?

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

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

Gradients at Inference

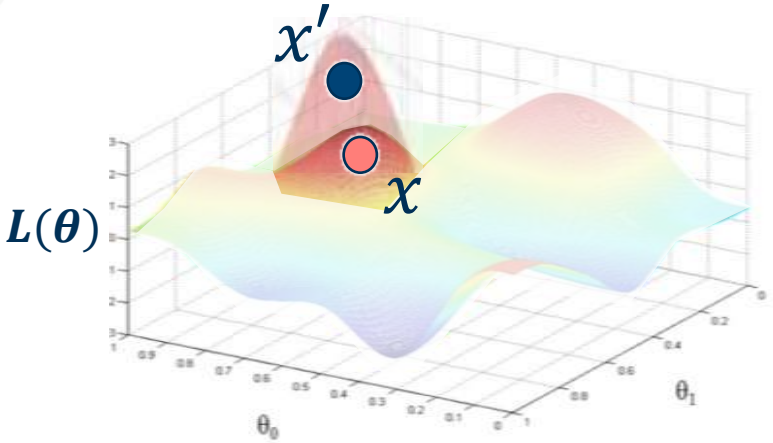
To Characterize the Novel Data at Inference



At Inference

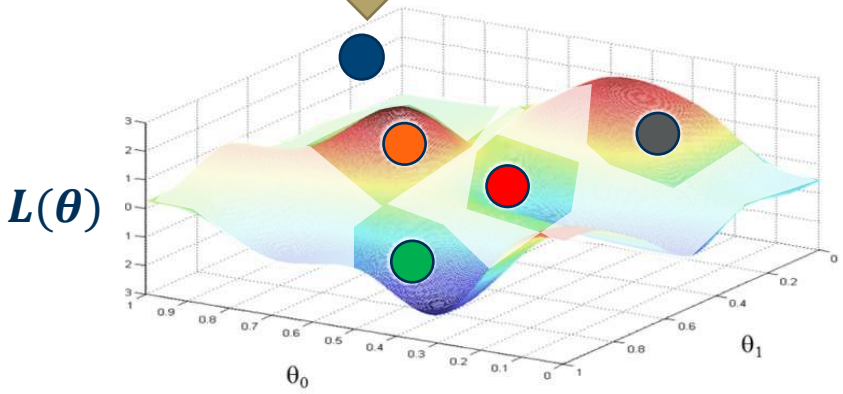
Part 2, 3

Counterfactual and Contrastive Representations using Gradients



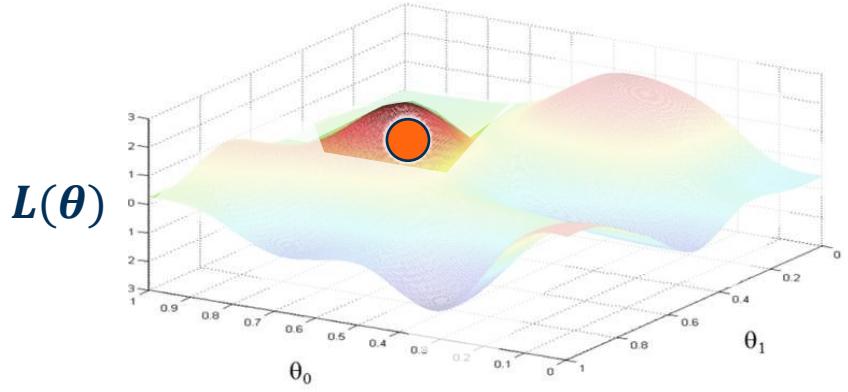
Part 3

Representation Traversal using Interventions



Part 4

Local editing for fairness interventions



Inferential Machine Learning

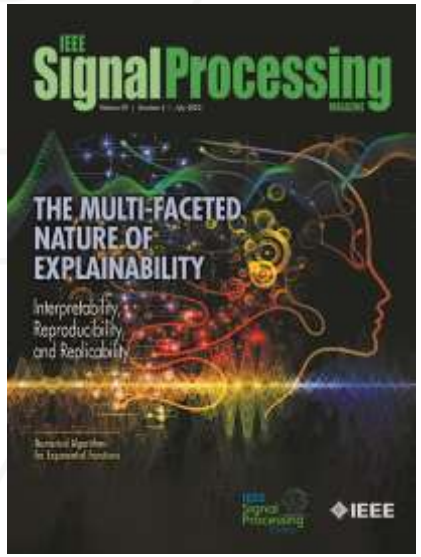
Part 2: Explainability at Inference

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
- 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

OLIVES

Georgia Tech

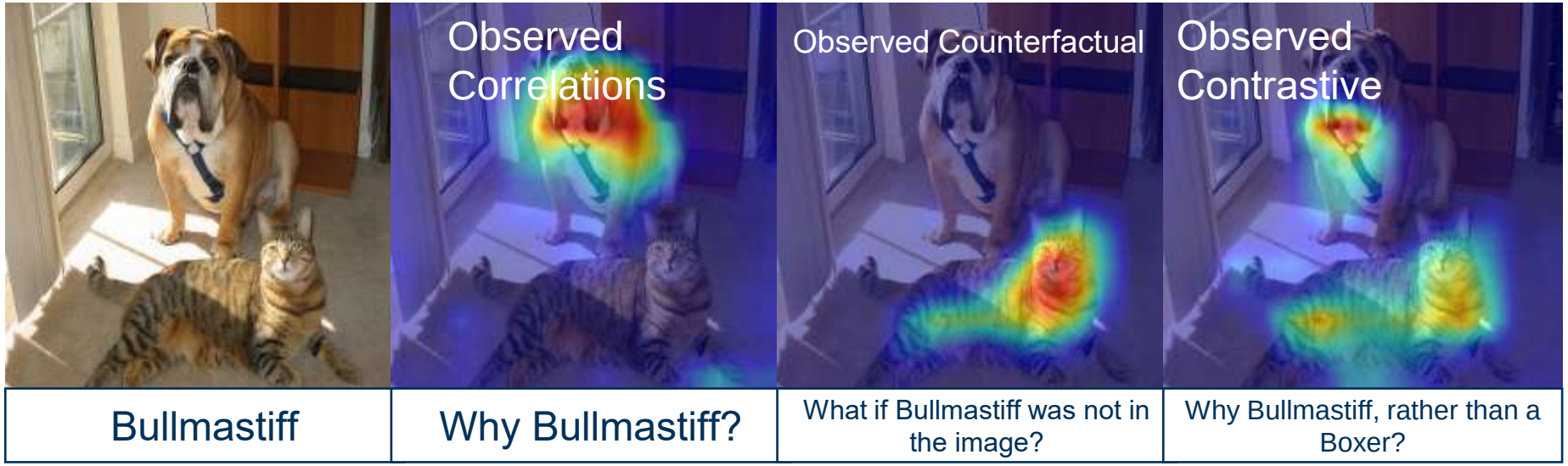
Explanations

Visual Explanations

SCAN ME

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

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



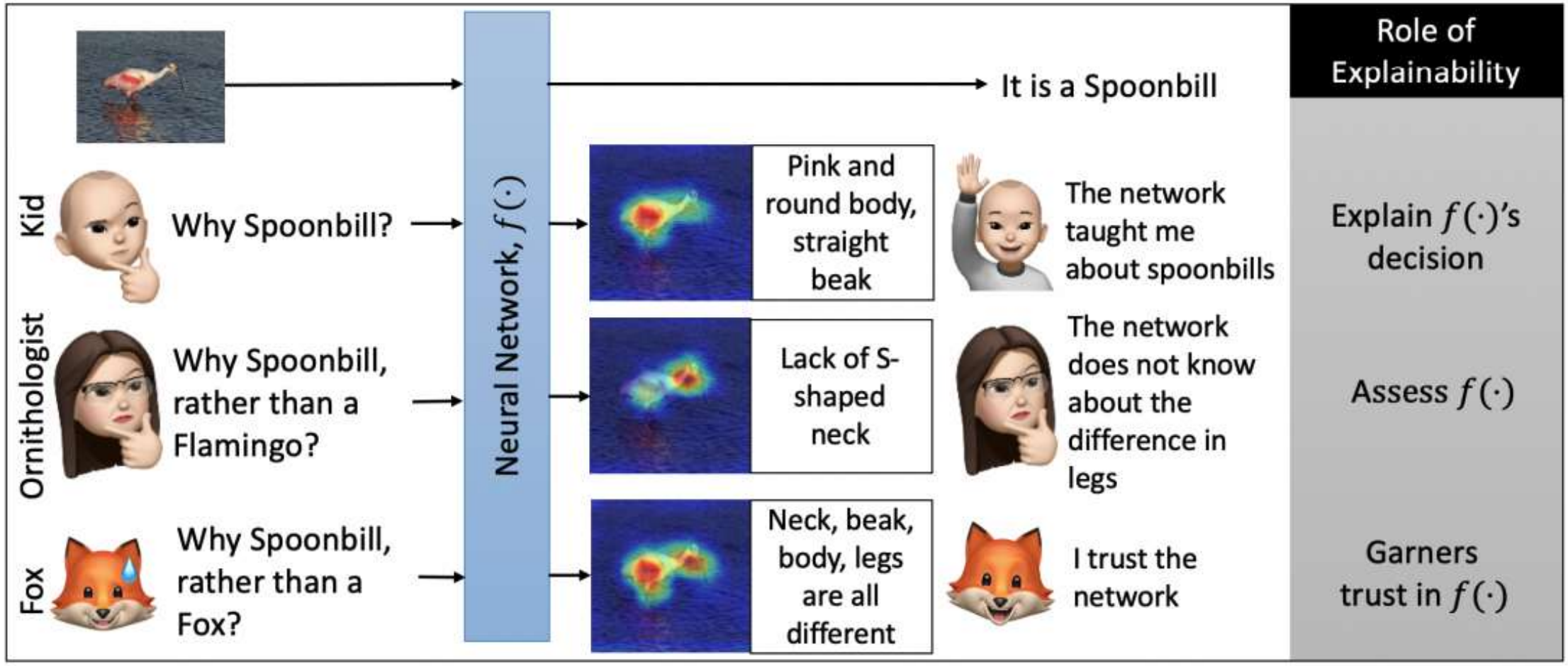
AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks: Towards relevant and contextual explanations. *IEEE Signal Processing Magazine*, *39*(4), 59-72.

Explanations

Role of Explanations – context and relevance

SCAN ME

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations



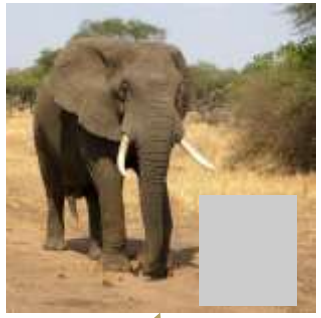
AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks: Towards relevant and contextual explanations. *IEEE Signal Processing Magazine*, *39*(4), 59-72.

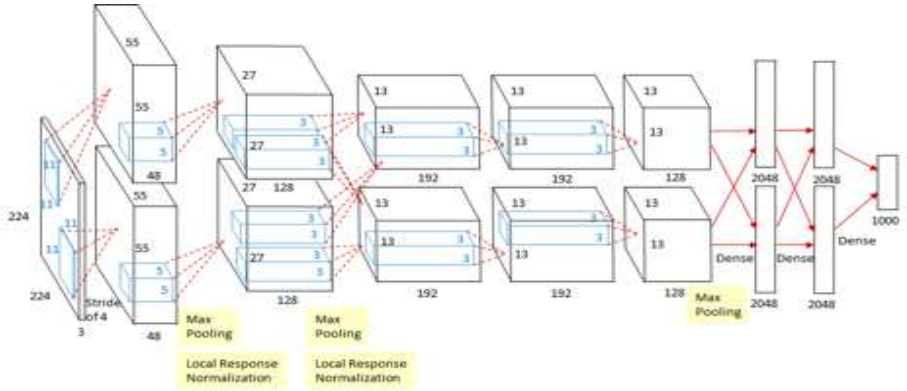
Explanations

Input Saliency via Occlusions

SCAN ME

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





P(elephant) = 0.95

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

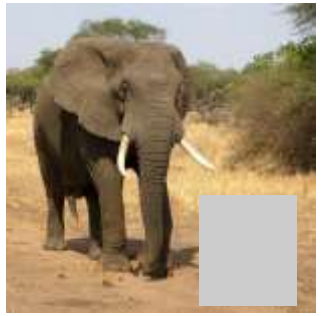
Zeiler and Fergus, "Visualizing and Understanding Convolutional Networks", ECCV 2014

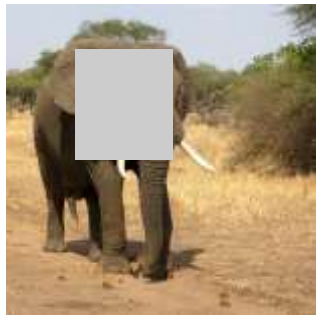
Explanations

Input Saliency via Occlusions

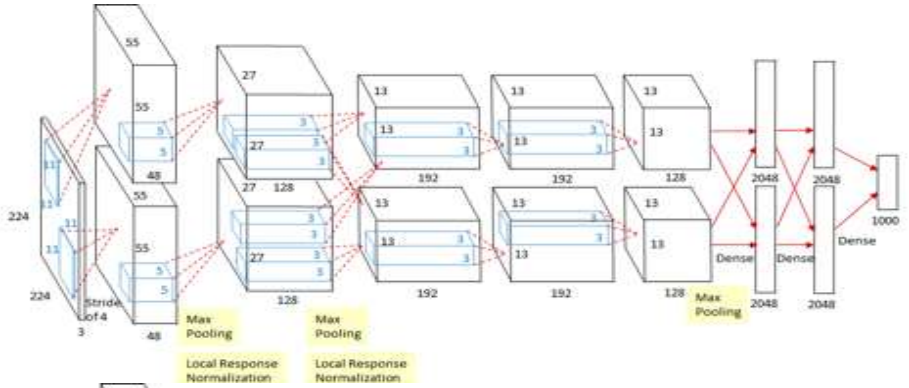
SCAN ME

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

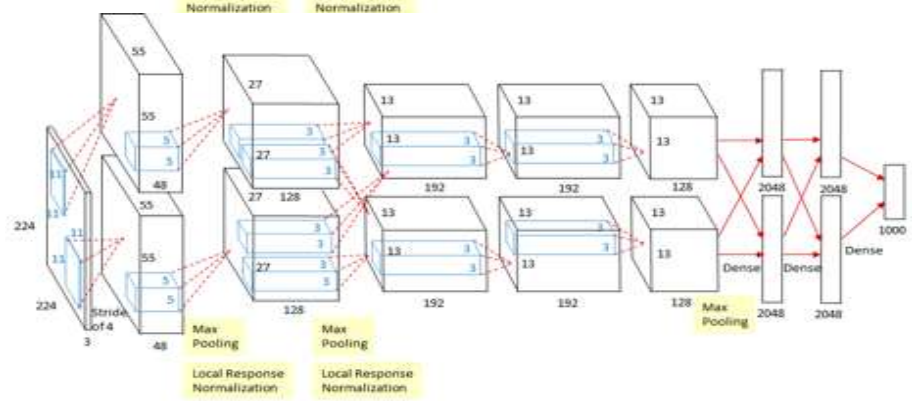




These pixels affect decisions more



P(elephant) = 0.95



P(elephant) = 0.75

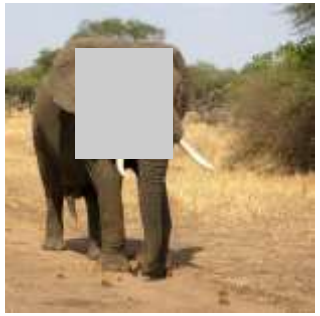
Zeiler and Fergus, "Visualizing and Understanding Convolutional Networks", ECCV 2014

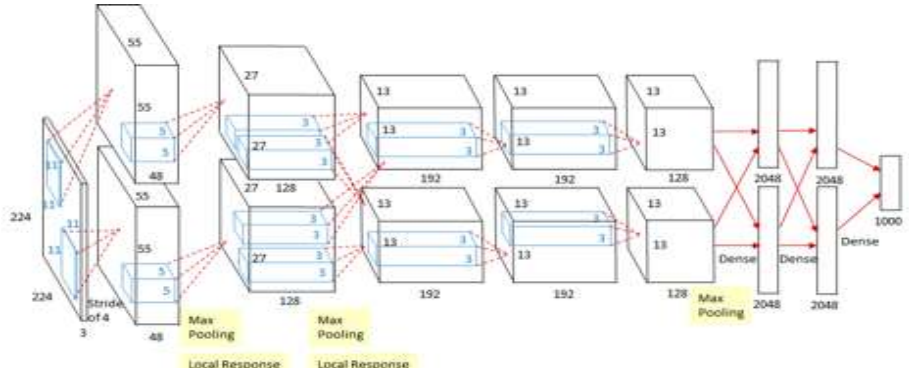
Explanations

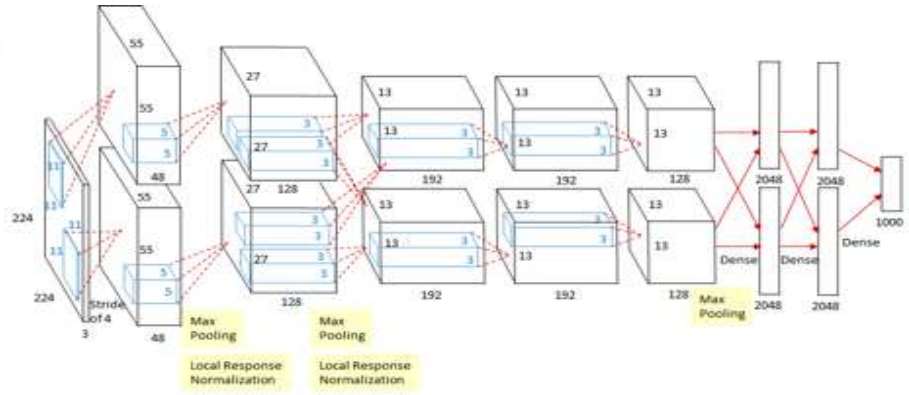
Input Saliency via Occlusions

SCAN ME

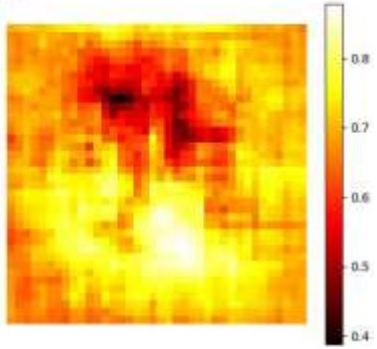
The network is trained with image- labels, but it is sensitive to the common visual regions in images



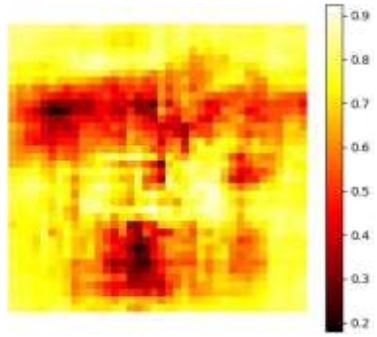




African elephant, Loxodonta africana



go-kart

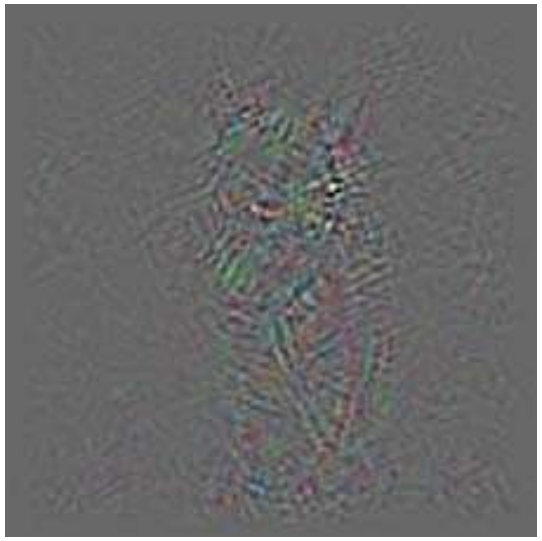


Zeiler and Fergus, "Visualizing and Understanding Convolutional Networks", ECCV 2014

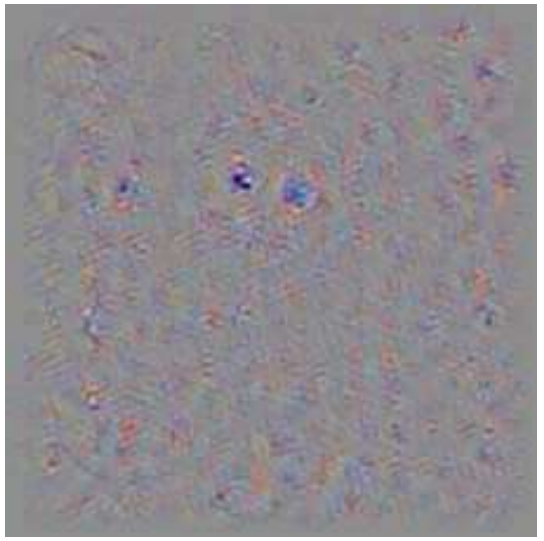
Gradients provide a one-shot means of perturbing the input that changes the output; They provide pixel-level importance scores

Input

Vanilla Gradients



Deconvolution Gradients



Guided Backpropagation

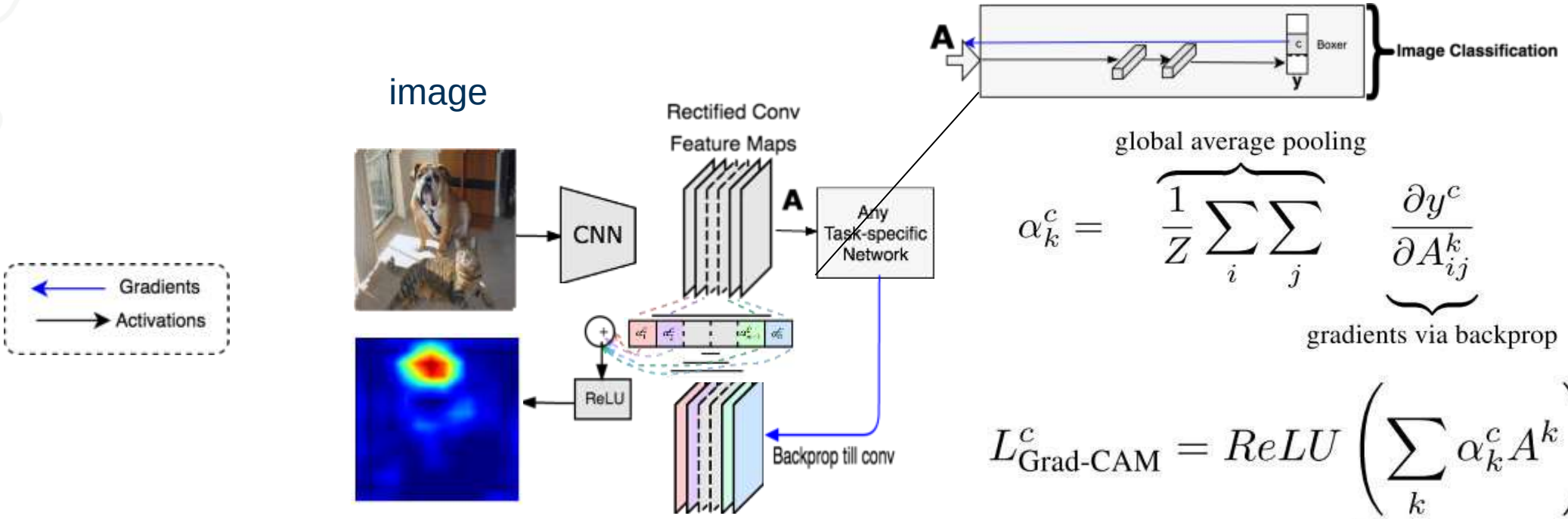
However, localization remains an issue

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

Gradient and Activation-based Explanations

GradCAM

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.



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

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

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

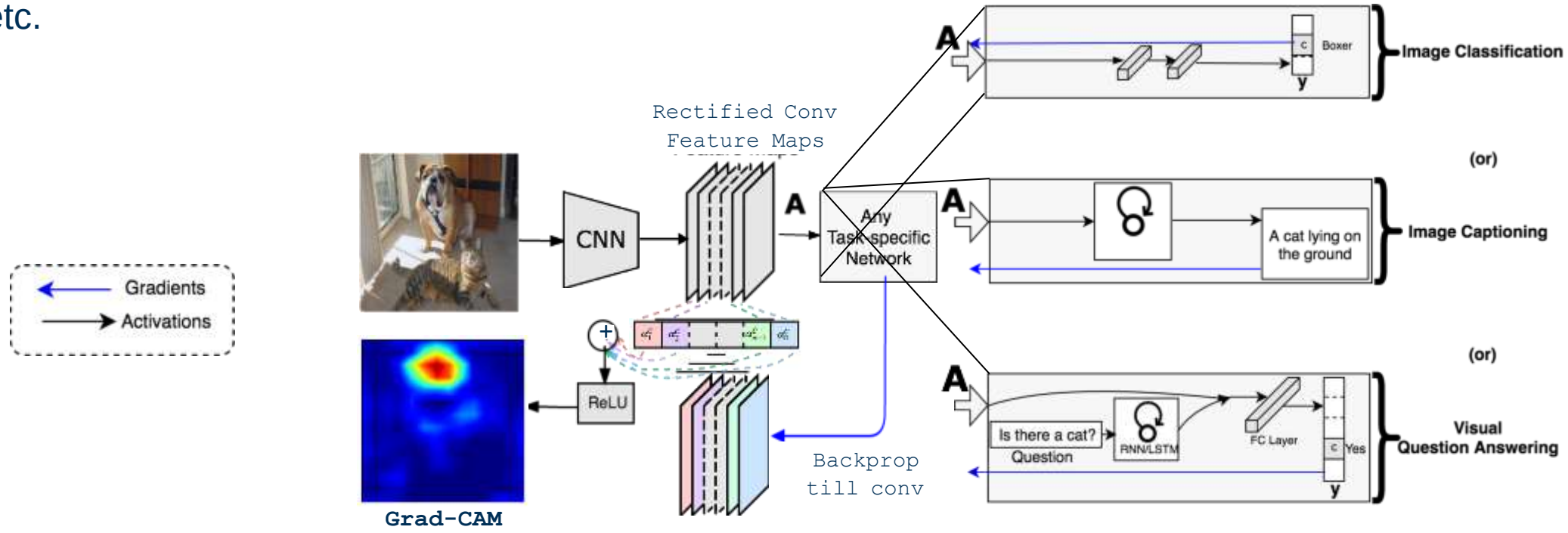
Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

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.

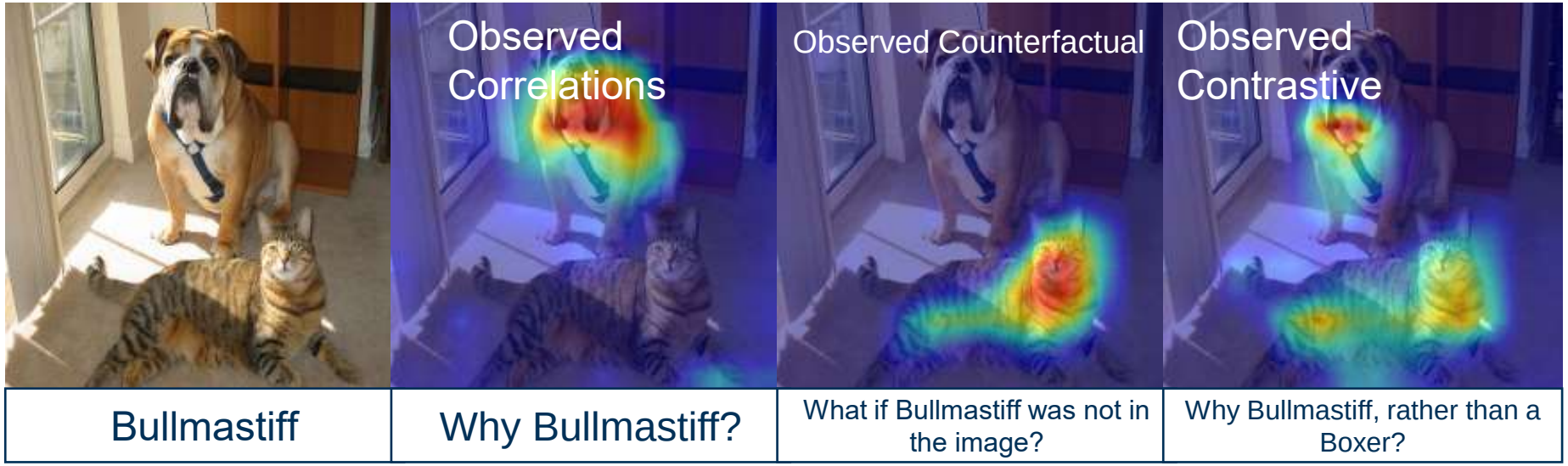
Gradient and Activation-based Explanations

Explanatory Paradigms

SCAN ME

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

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



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

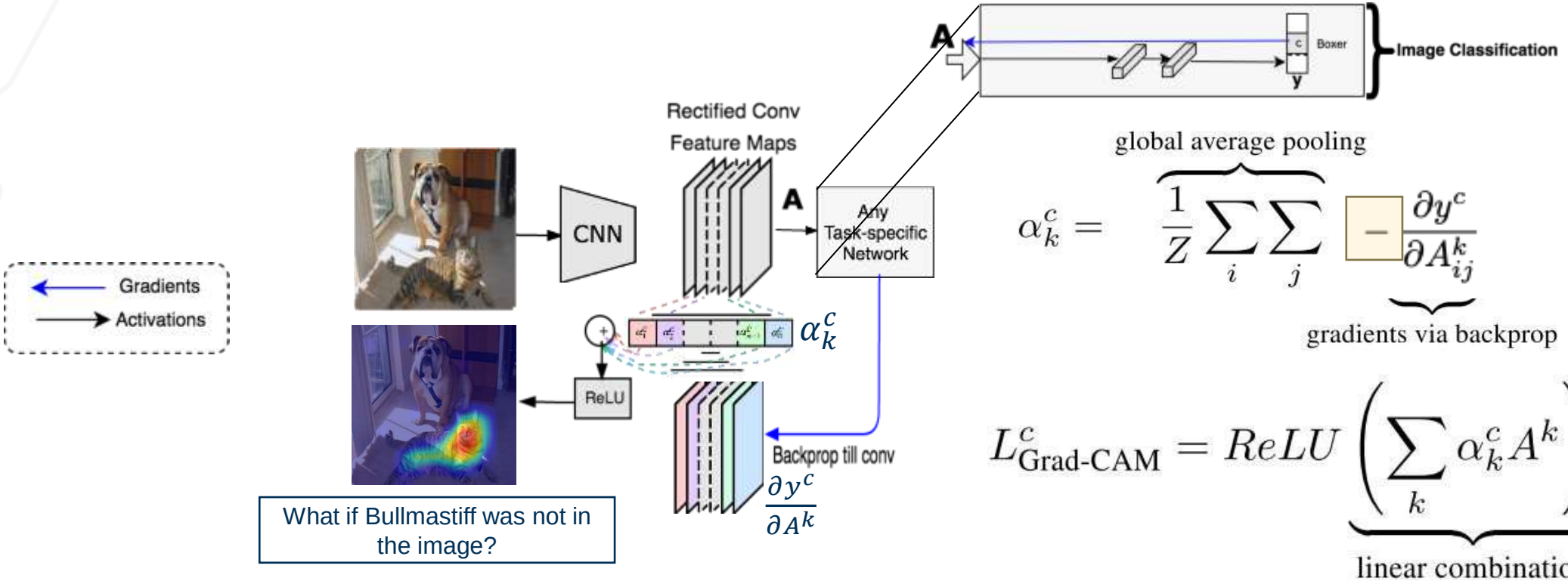
Gradient and Activation-based Explanations

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

SCAN ME

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

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



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

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

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.

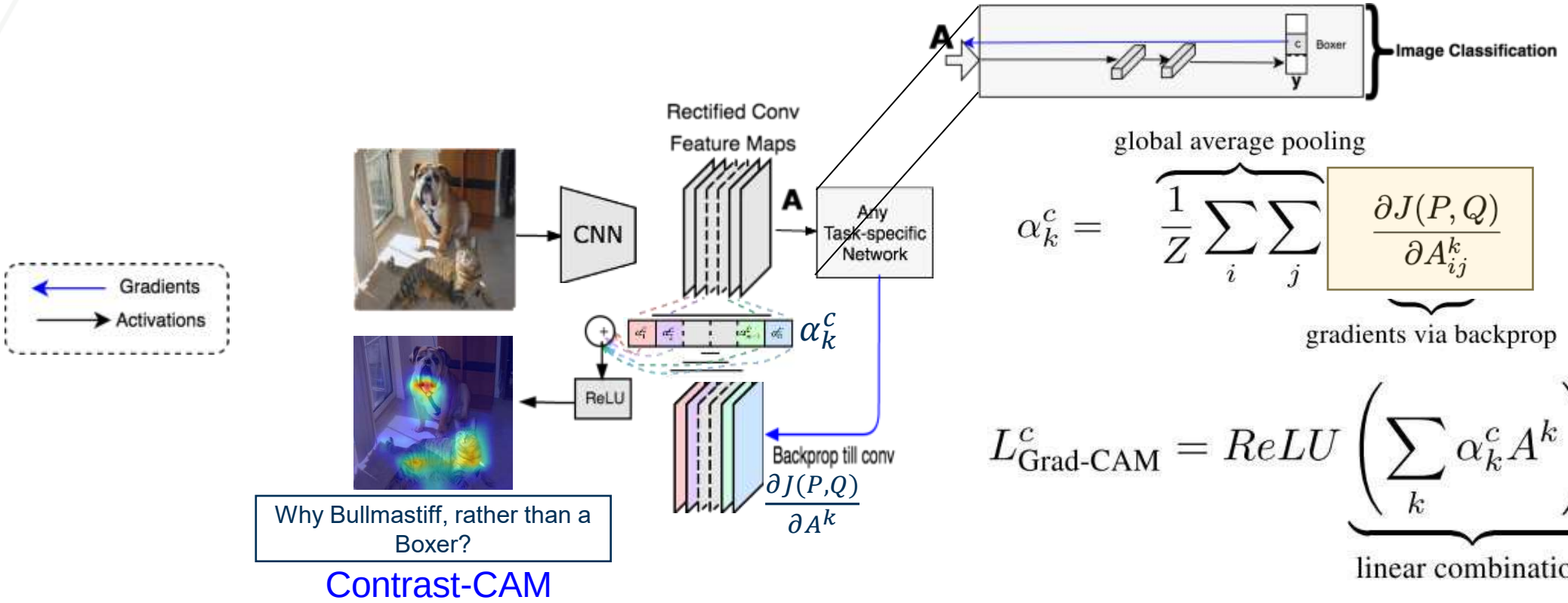
Gradient and Activation-based Explanations

ContrastCAM: Why P, rather than Q?

SCAN ME

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

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



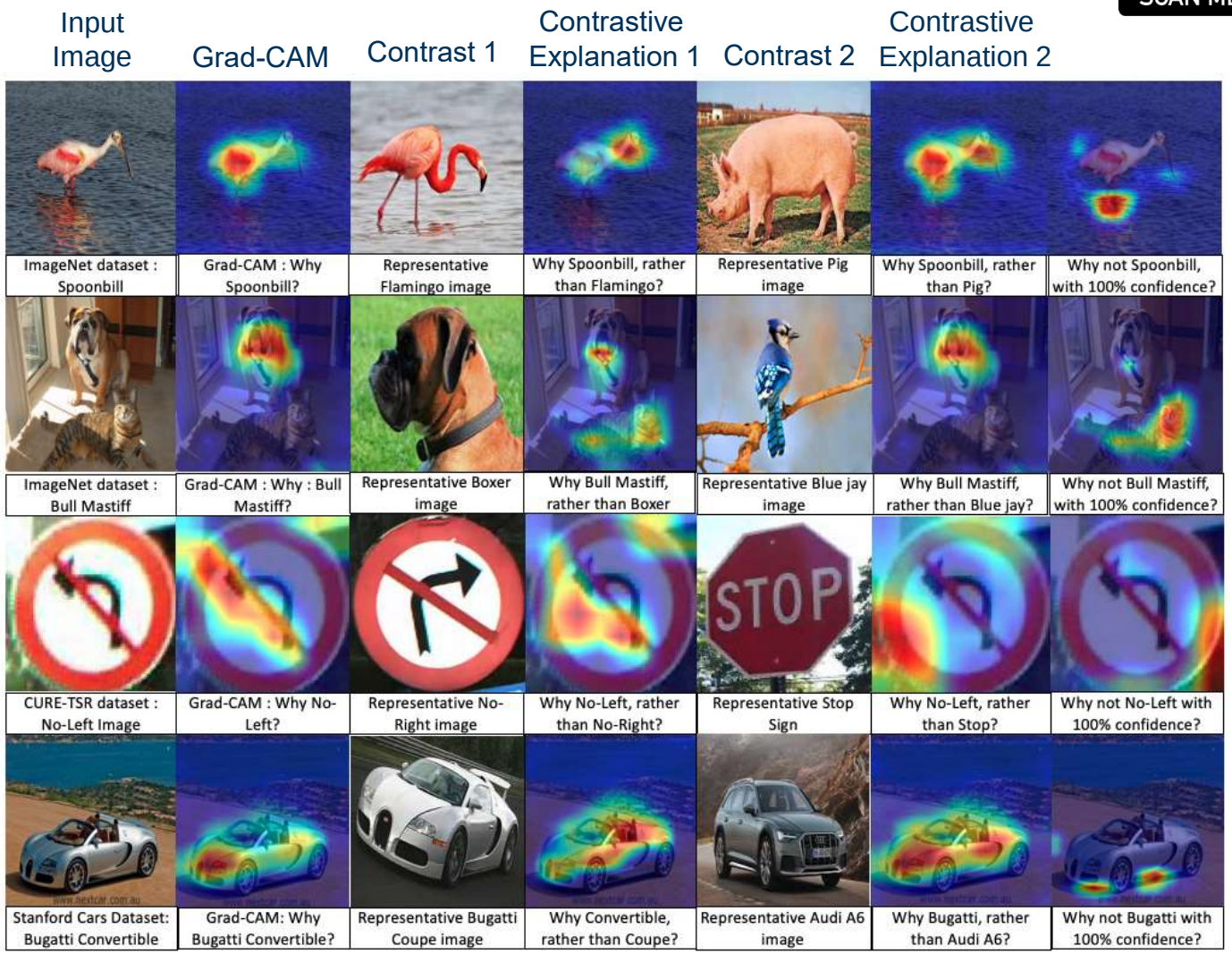
Backpropagating the loss highlights the differences between classes P and Q.

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

SCAN ME

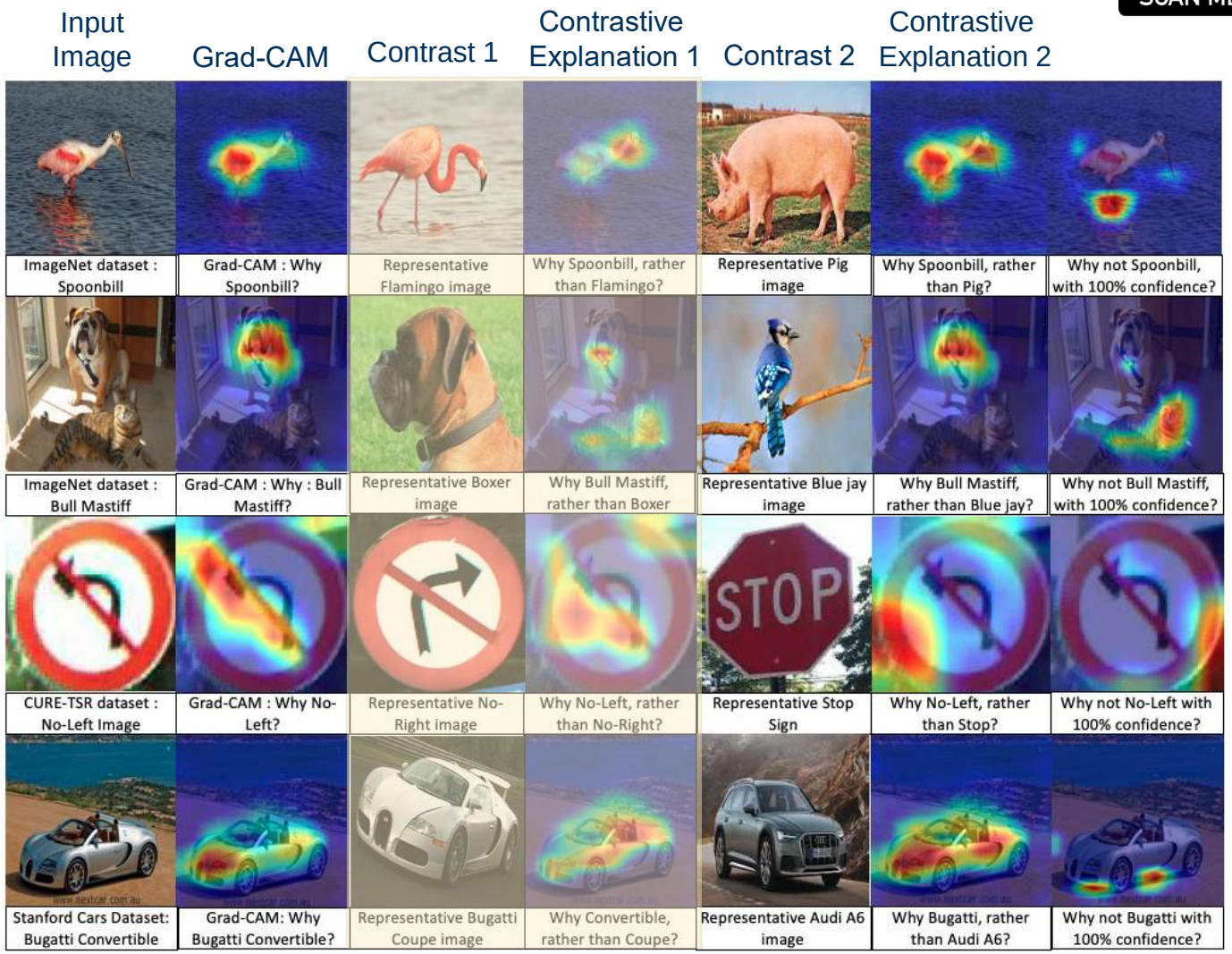


Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

SCAN ME



Human Interpretable

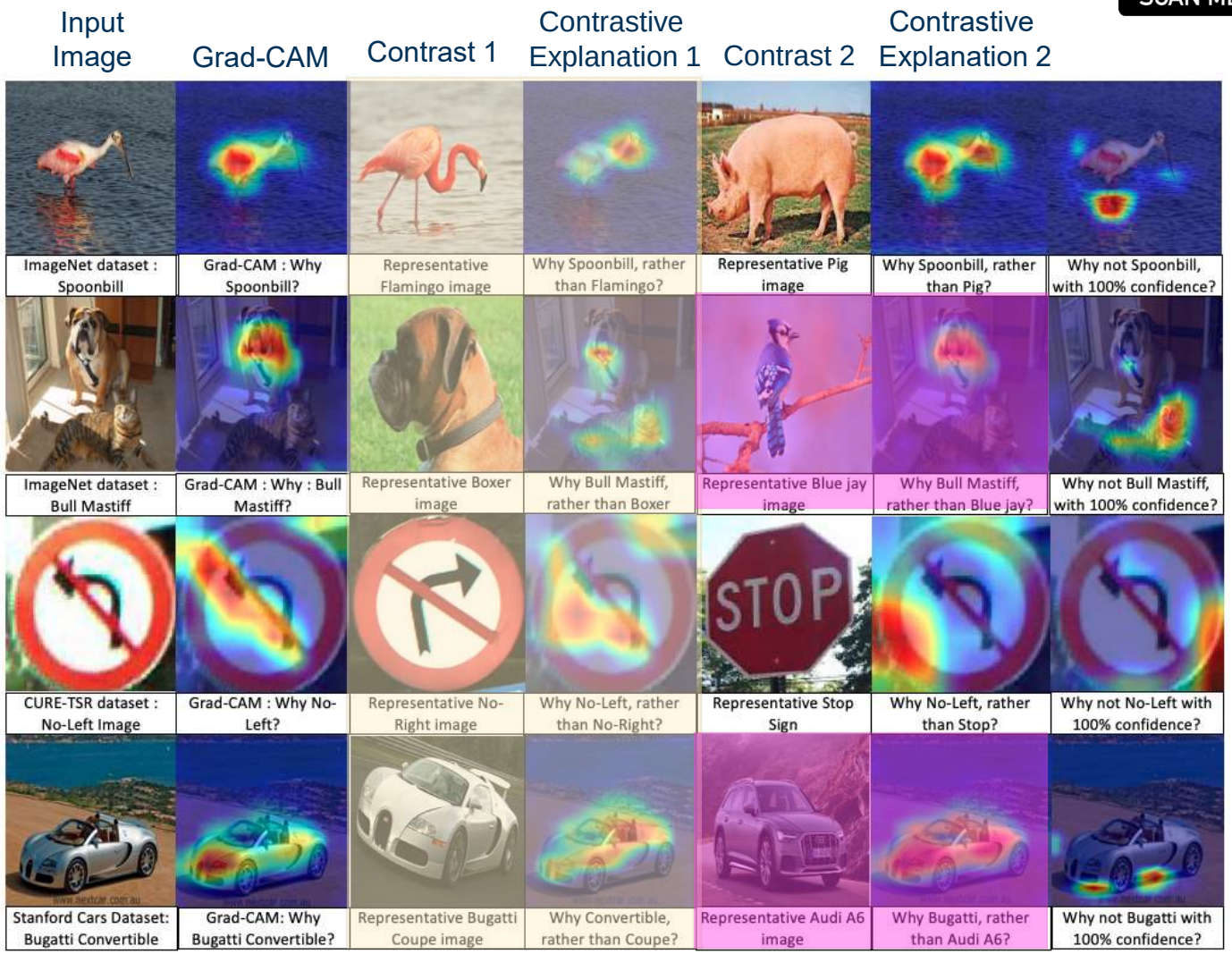
Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

SCAN ME

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations



Human Interpretable

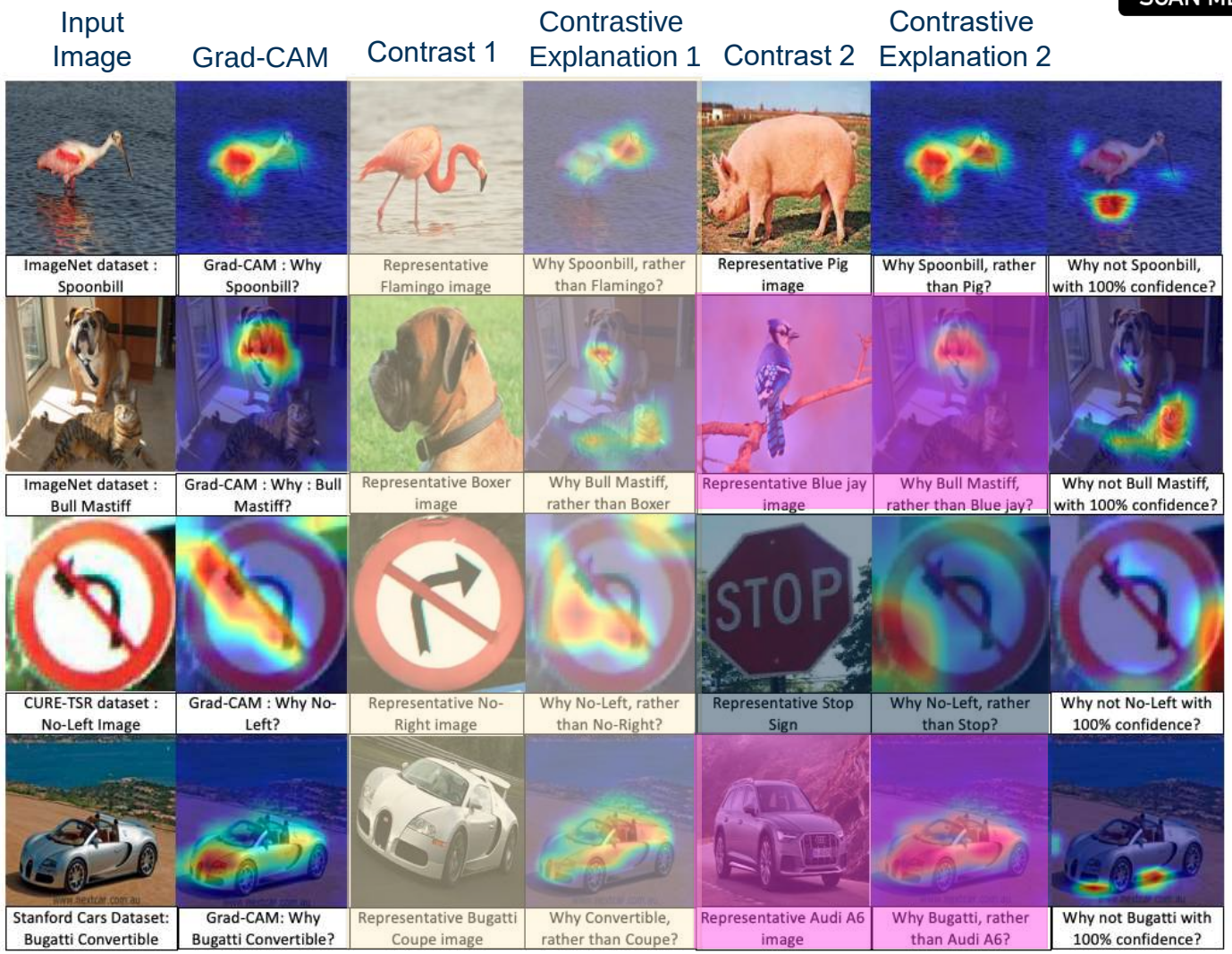
Same as Grad-CAM

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

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

SCAN ME



Human Interpretable

Same as Grad-CAM

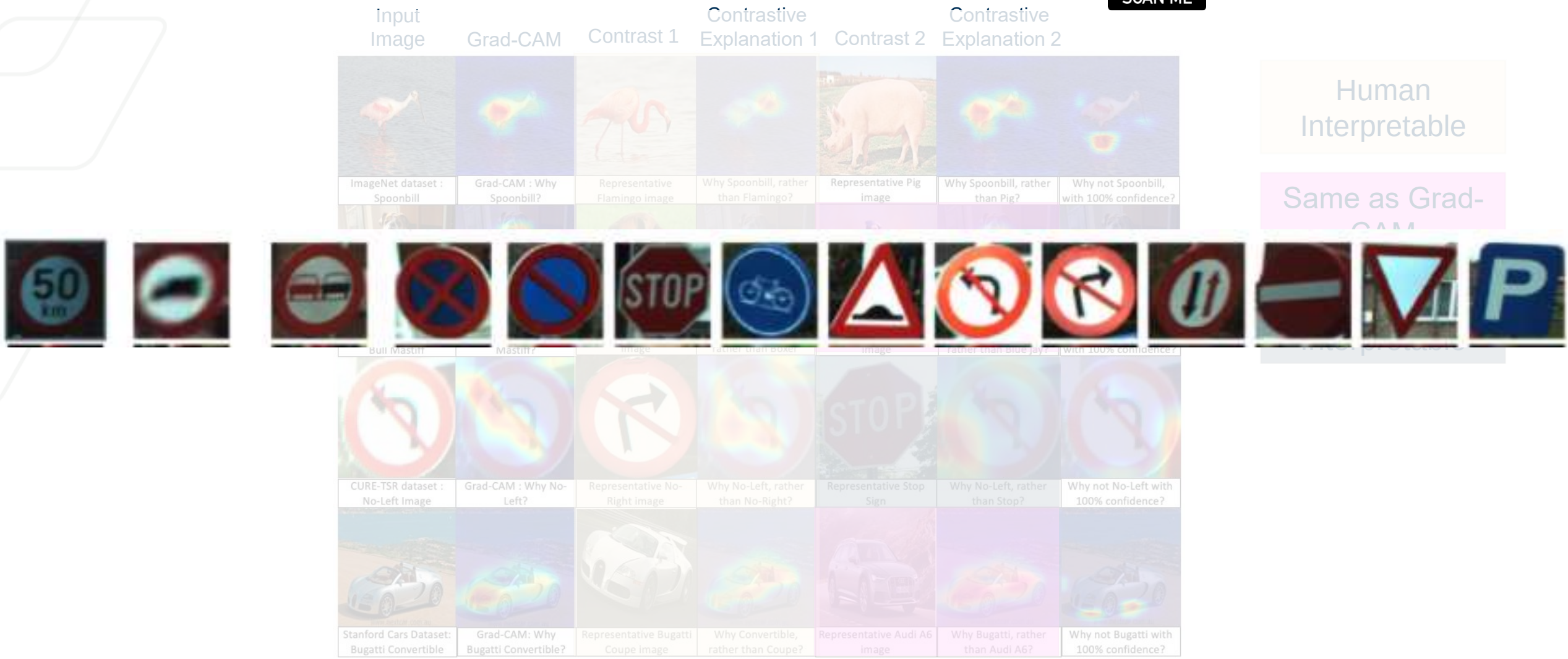
Not Human Interpretable

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

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

SCAN ME



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

Gradient and Activation-based Explanations

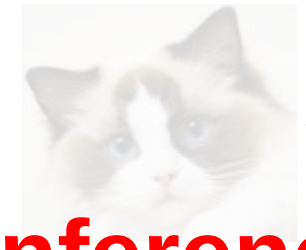
Results from GradCAM, CounterfactualCAM, and ContrastCAM

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



Network $f(\theta)$

Predicted
Class Probability

Dog
Cat
Horse
Bird

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

$l(\theta|x)$

Likelihood function

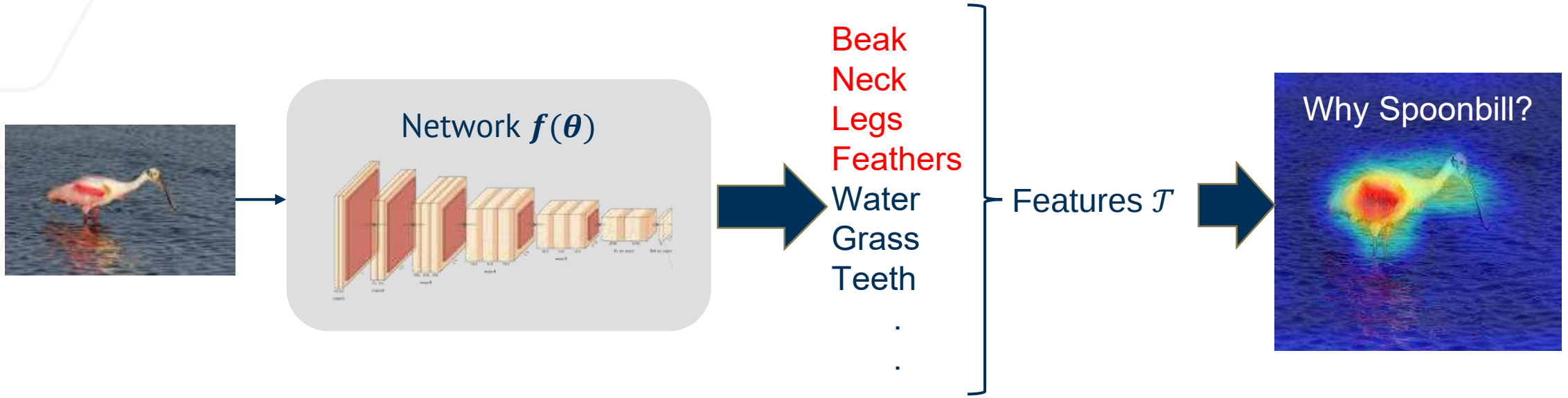
$$I(\theta) = Var(\frac{\partial}{\partial\theta} l(\theta|x))$$

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

Information at Inference

Case Study: Explainability

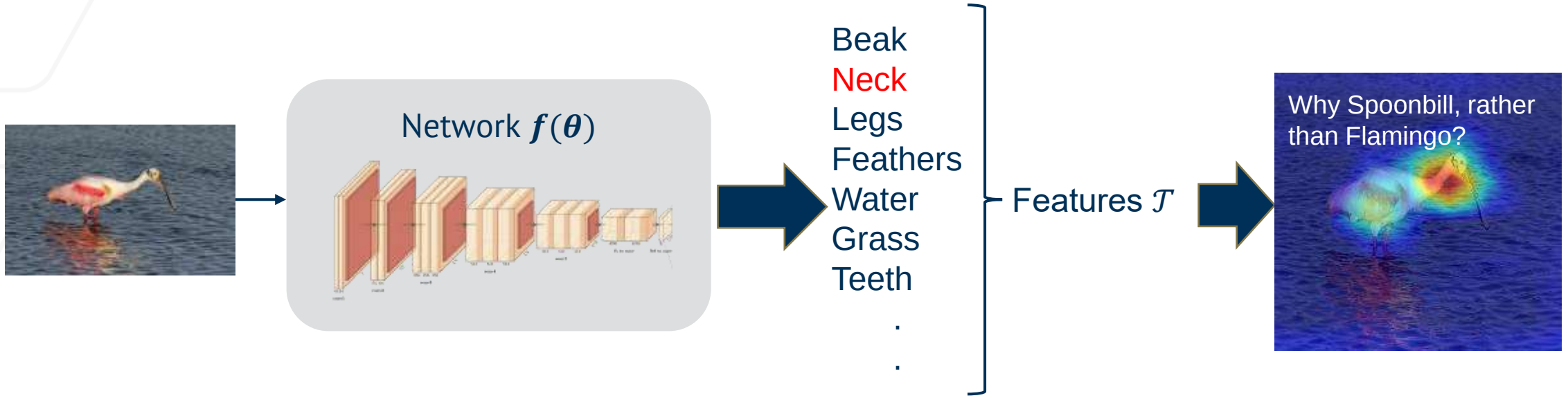
$\mathcal{T}$ is the set of all features learned by a trained network



Information at Inference

Case Study: Explainability

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



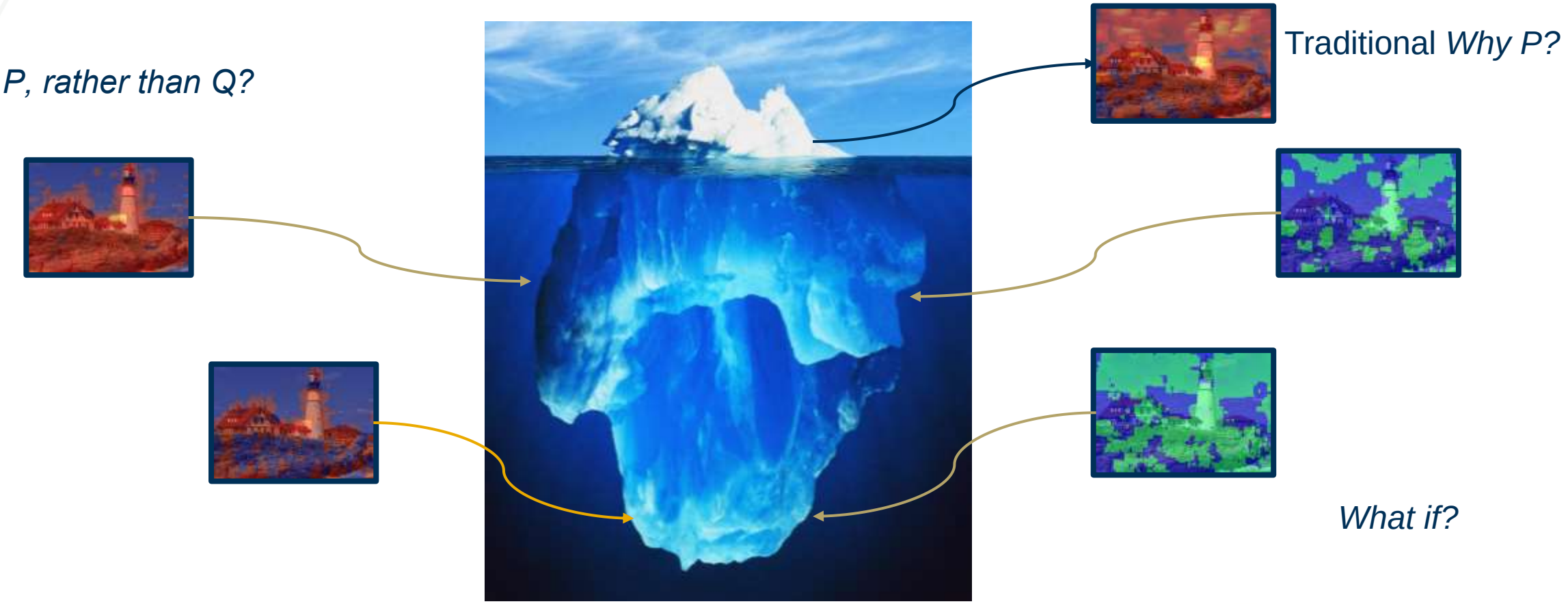
All the requisite Information is stored within $f(\theta)$

Goal: To extract and utilize this information – Inferential Machine Learning

Information at Inference

Implicit Knowledge in Neural Networks – Inferential Machine Learning

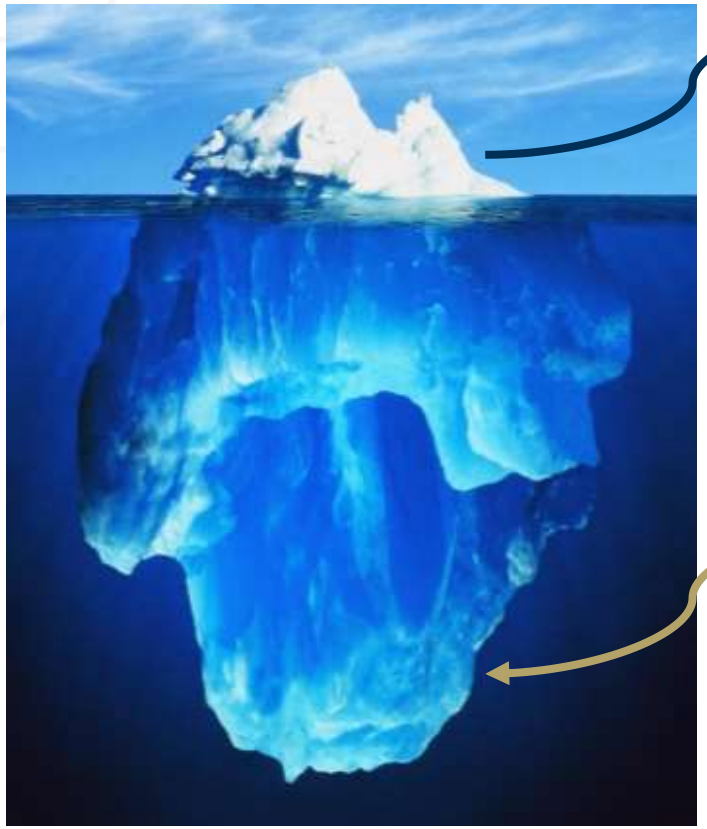
Trained Neural Networks have a wealth of implicit stored knowledge. Inferential Machine Learning aims to 'transmute' this knowledge for other tasks



Inferential Machine Learning

Theory Underlying Inferential Machine Learning

Inferential Theory of Learning views learning as a goal-oriented process of modifying (transmuting) the learner's knowledge by exploring the learner's experience[1]



Learned Knowledge

↓

Inferential Theory of Learning (ITL) [1]

↓

Transmuted Knowledge

Input Information

+

Model Knowledge

=

Robust, Fair, Interpretable Decisions

What is the goal? Our view: Reduce Uncertainty. More on this in Part 3

[1] Michalski, Ryszard S. "Inferential theory of learning as a conceptual basis for multistrategy learning." *Machine learning*11 (1993): 111-151.

Inferential Machine Learning

Part 3: Uncertainty and Intervenability at Inference

Objective

Objective of the Tutorial

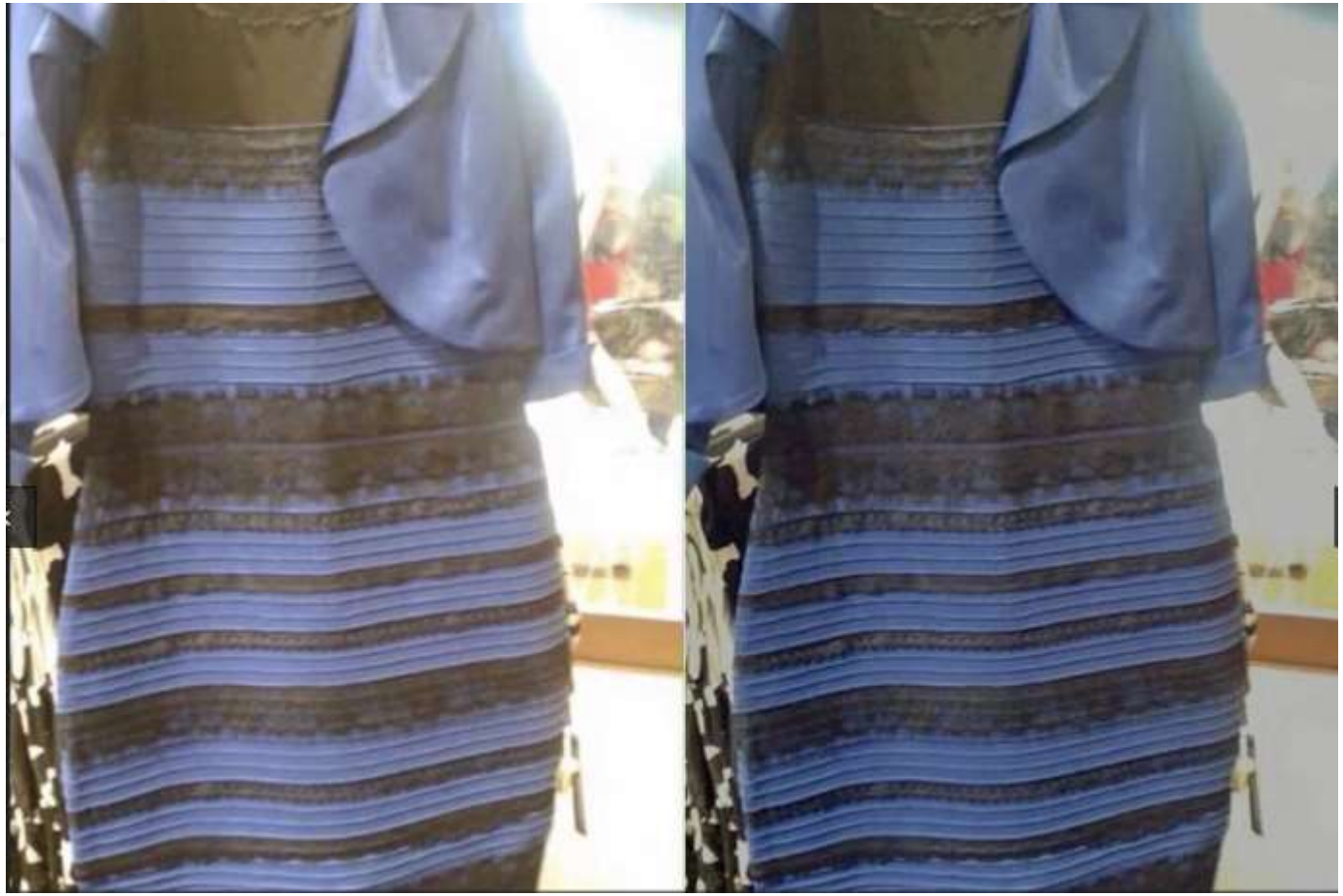
To discuss methodologies that promote robust and fair inference in neural networks

- 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
  - Case Study 3: Introspective Learning
  - Inferential Machine Learning
- Part 4: Fairness Interventions

Uncertainty

What is Uncertainty?

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

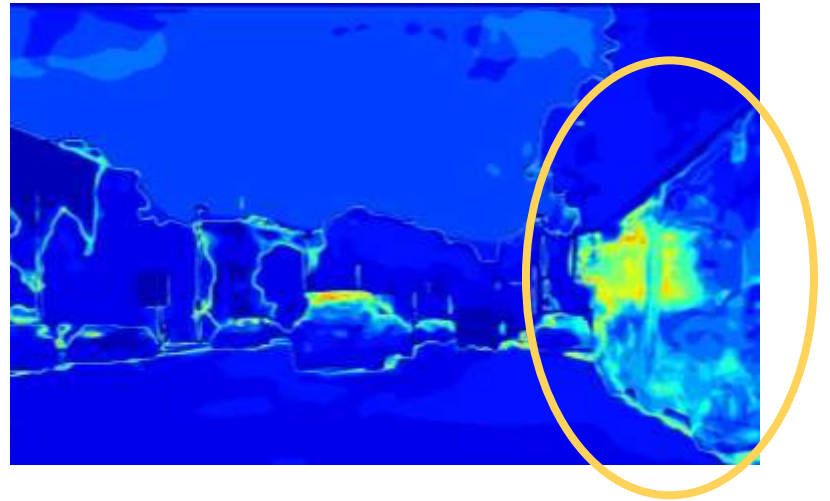
What is Uncertainty?

Uncertainty is a model knowing that it does not know

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%

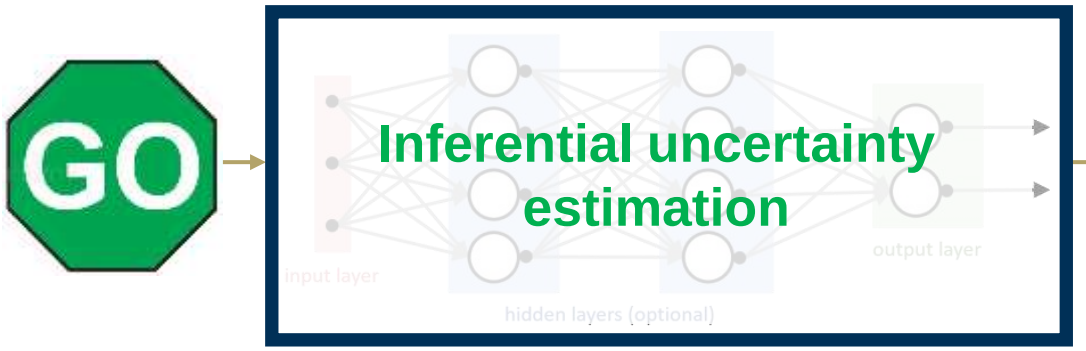
In classification, Uncertainty Quantification (UQ) implies providing a classification label and its associated uncertainty

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

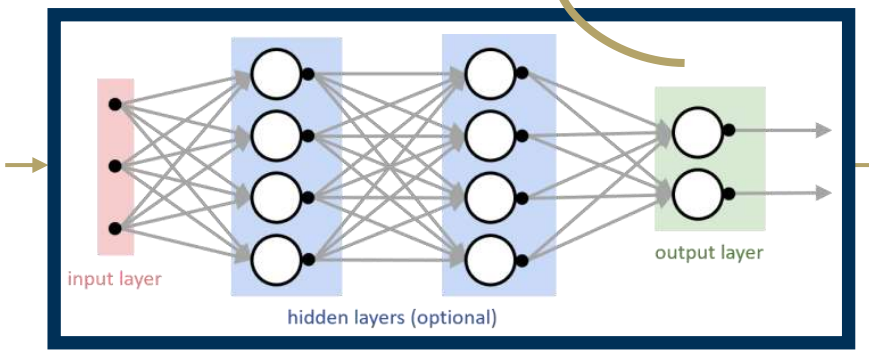
Challenge in Uncertainty Quantification

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!



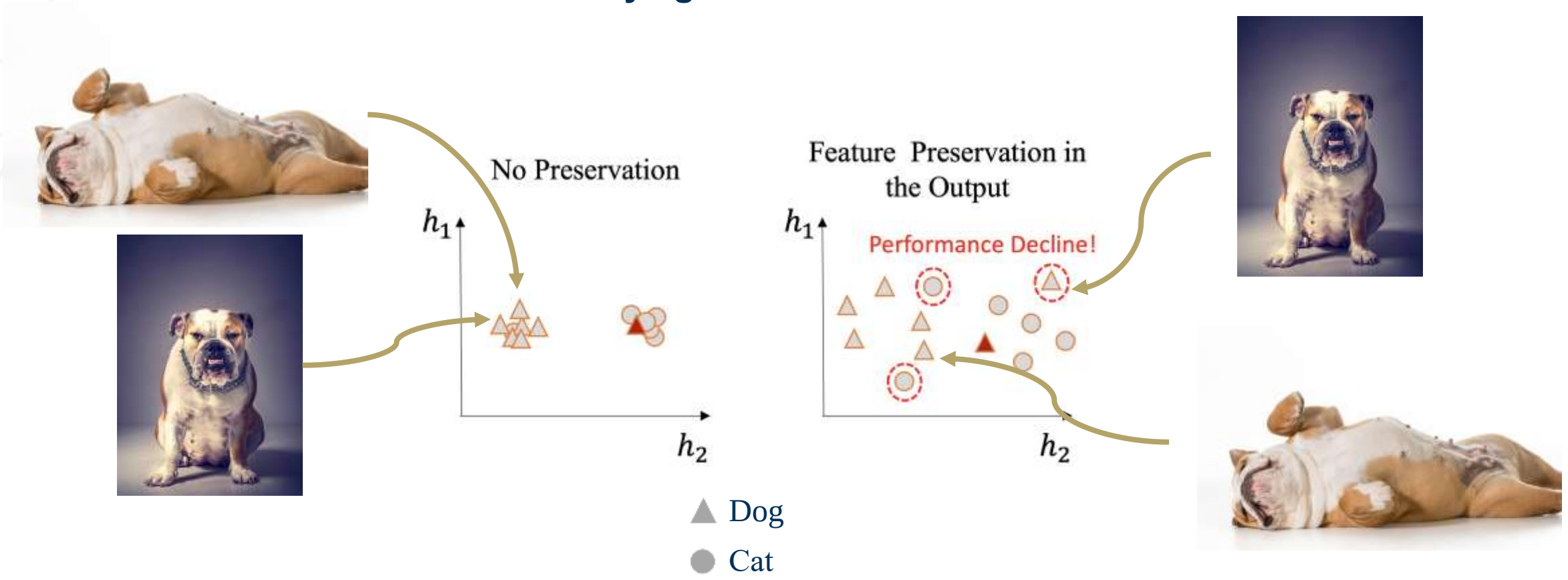
Dog

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

Uncertainty

Challenge in Uncertainty Quantification

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



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

Uncertainty

Simple Uncertainty Quantification 1: Negative Log Likelihood

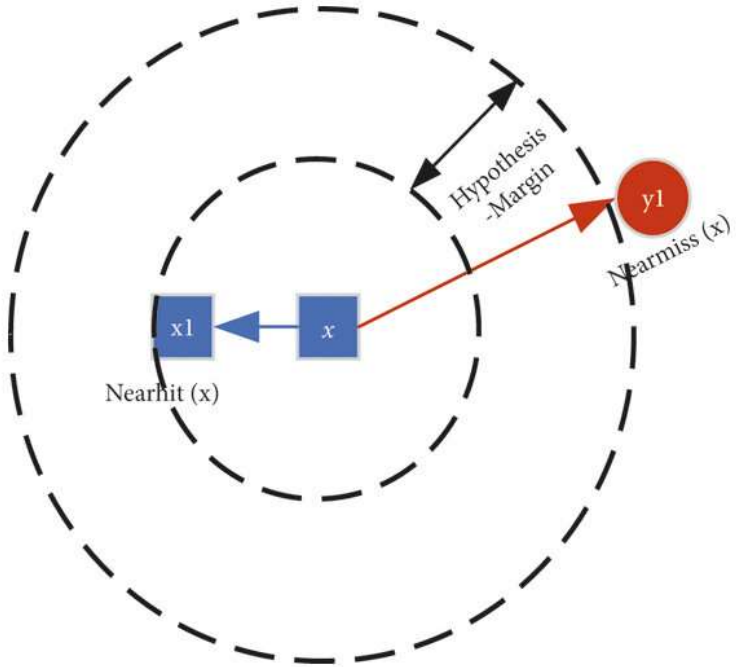
In Bayesian settings, uncertainty is treated as inverse likelihood; consequently, lower the negative of likelihood, lower the uncertainty

- 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

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

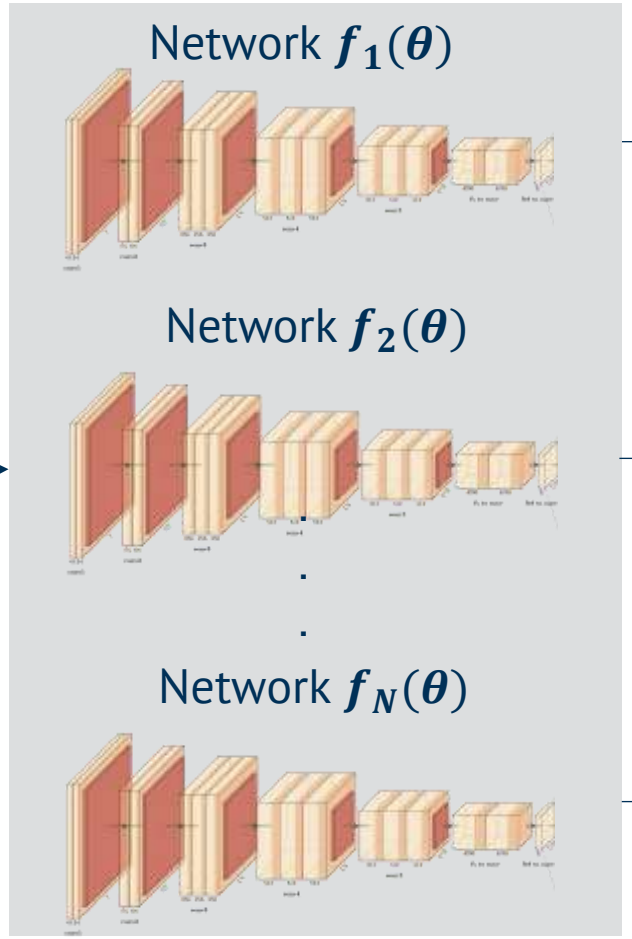
Fig. from Tian, Yanjia, and Xiang Feng. "Large Margin Graph Embedding-Based Discriminant Dimensionality Reduction." *Scientific Programming* 2021.1 (2021): 2934362.

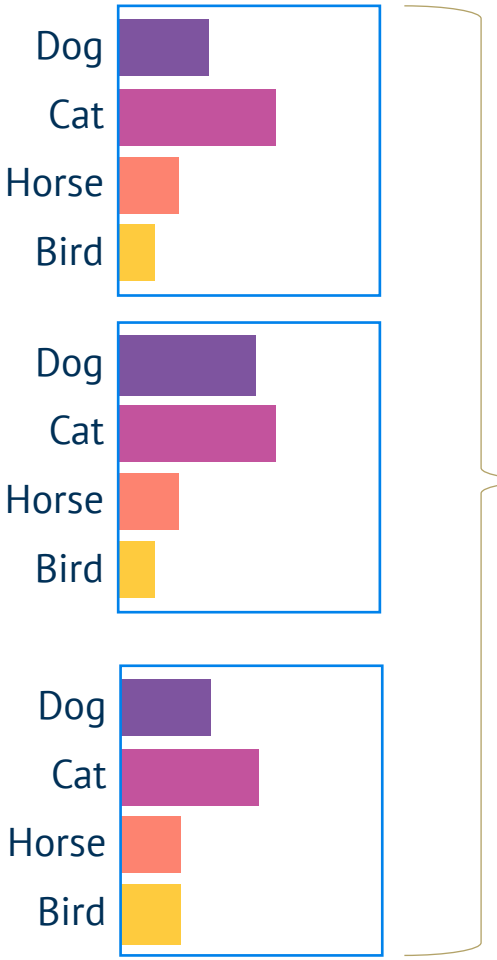
[1] Bartlett, Peter, et al. "Boosting the margin: A new explanation for the effectiveness of voting methods." *The annals of statistics* 26.5 (1998): 1651-1686.

Uncertainty

Uncertainty Quantification in Neural Networks

Via Ensembles[1]





Variation within outputs is the uncertainty.

Commonly referred to as **Prediction Uncertainty.**

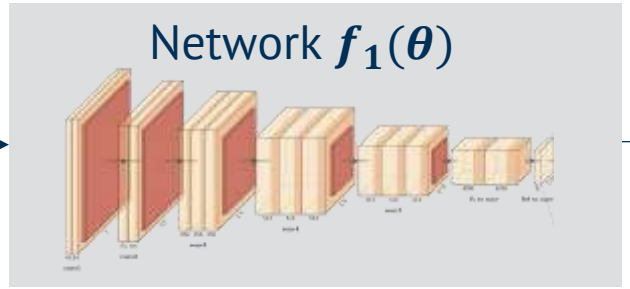
Requires multiple trained models – not exactly an inferential method

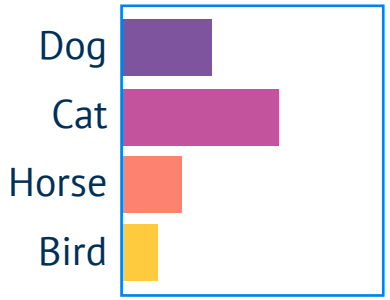
[1] Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles." *Advances in neural information processing systems* 30 (2017).

Uncertainty

Uncertainty Quantification in Neural Networks

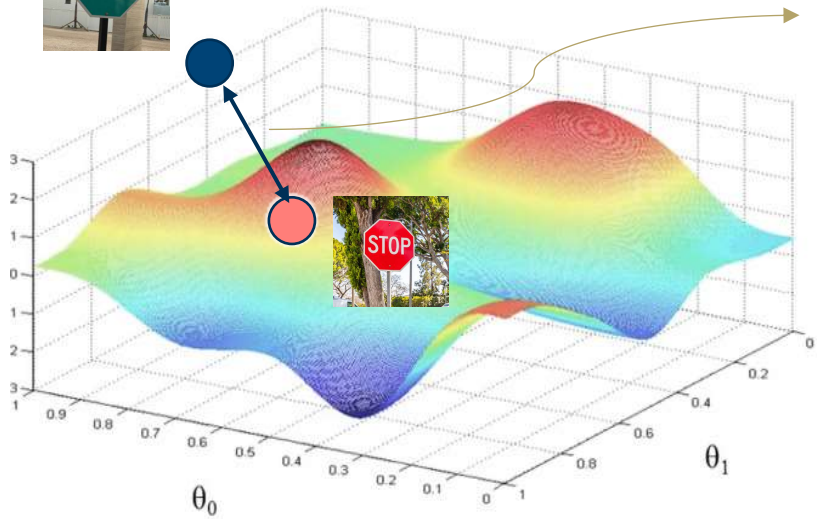
Via Single pass methods[1]





Uncertainty quantification using a single network and a single pass

$L(\theta)$



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.

Uncertainty

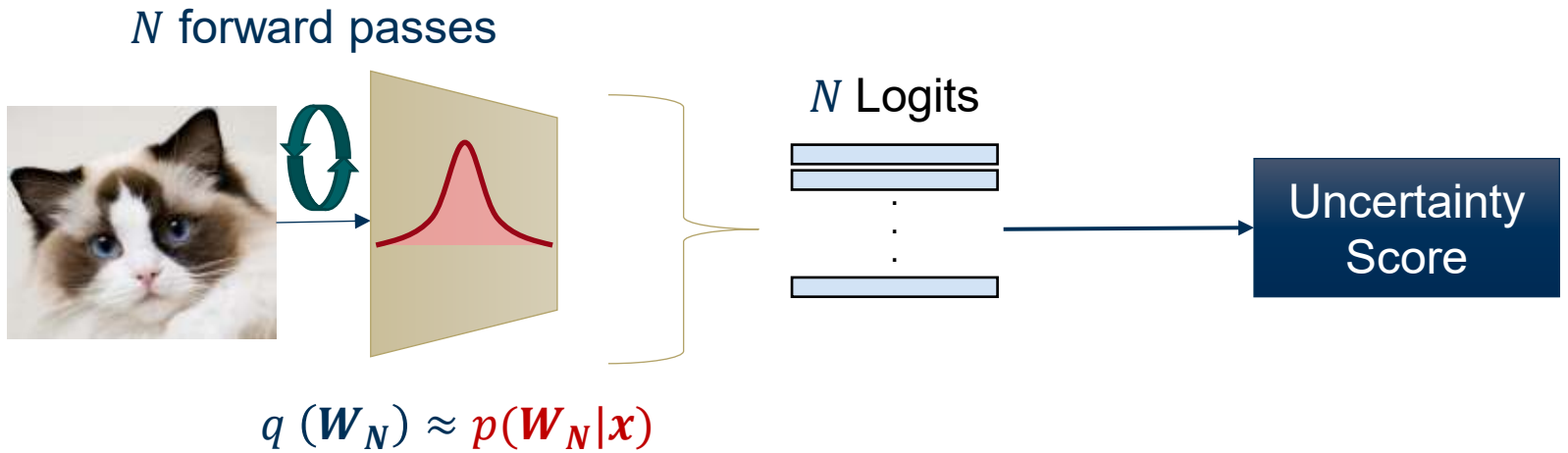
Iterative Uncertainty Quantification

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

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



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

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

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

Uncertainty

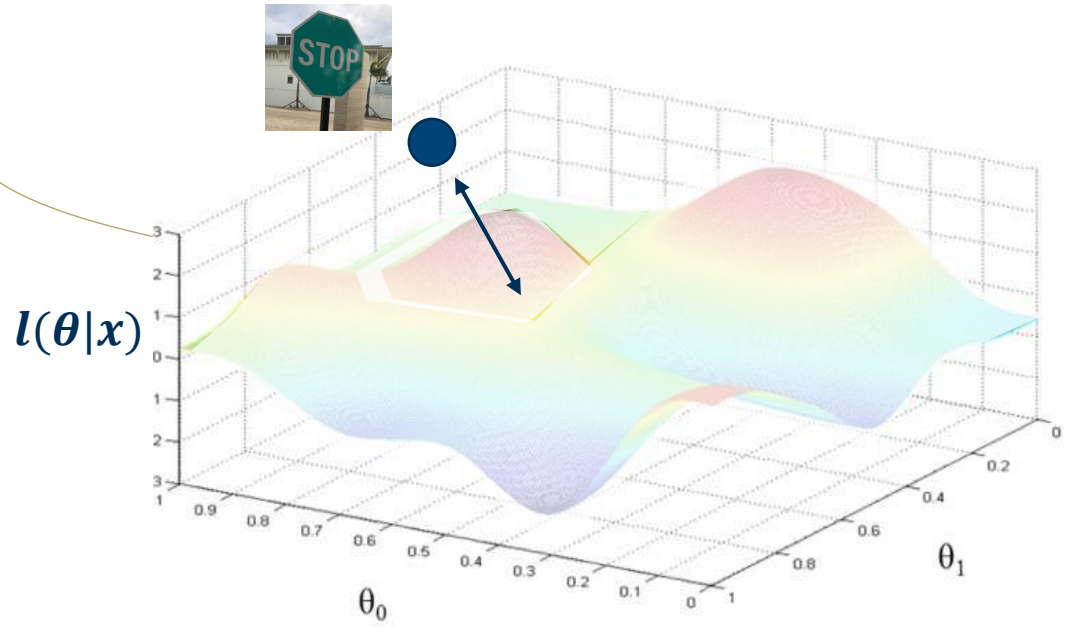
Gradients as Single pass Uncertainty Quantification

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

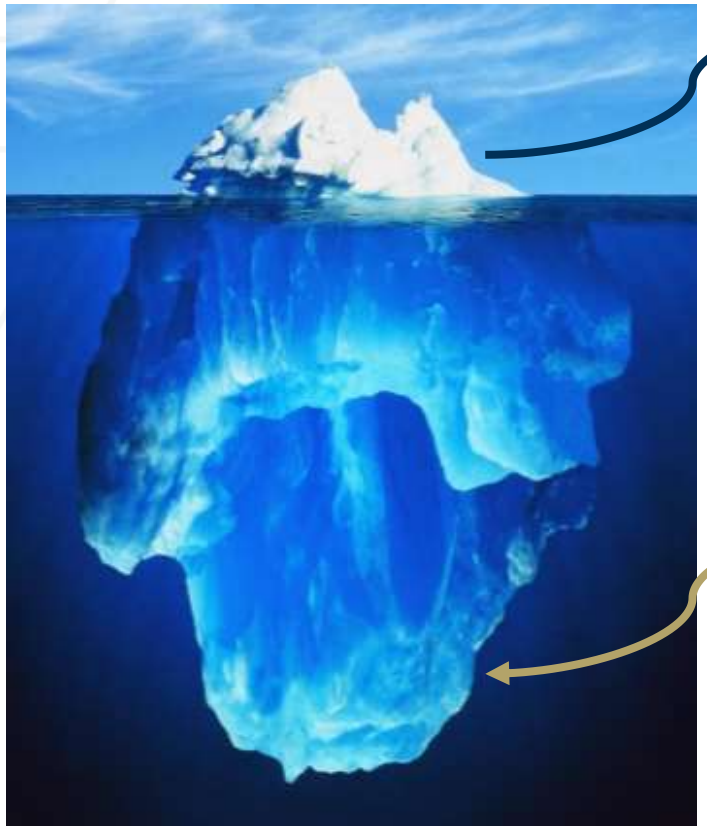
Applications relevant during model inference

Relevant at Deployment:

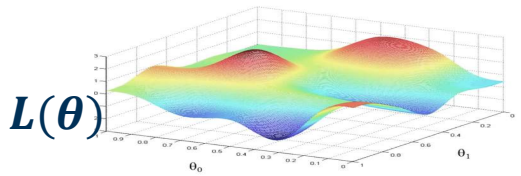
Provide a specific ‘uncertainty measure’ that objectively allows users to trust neural network predictions

Unfortunately, each application has its own uncertainty quantification

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



Learned Knowledge



Transmuted Knowledge

Part 2

Case Study 1

- Explainability
- Out-of-distribution Detection
- Adversarial Detection
- Anomaly Detection
- Corruption Detection
- Misprediction Detection

Case Study 3

- Causal Analysis
- Open-set Recognition
- Noise Robustness

Case Study 2

- 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

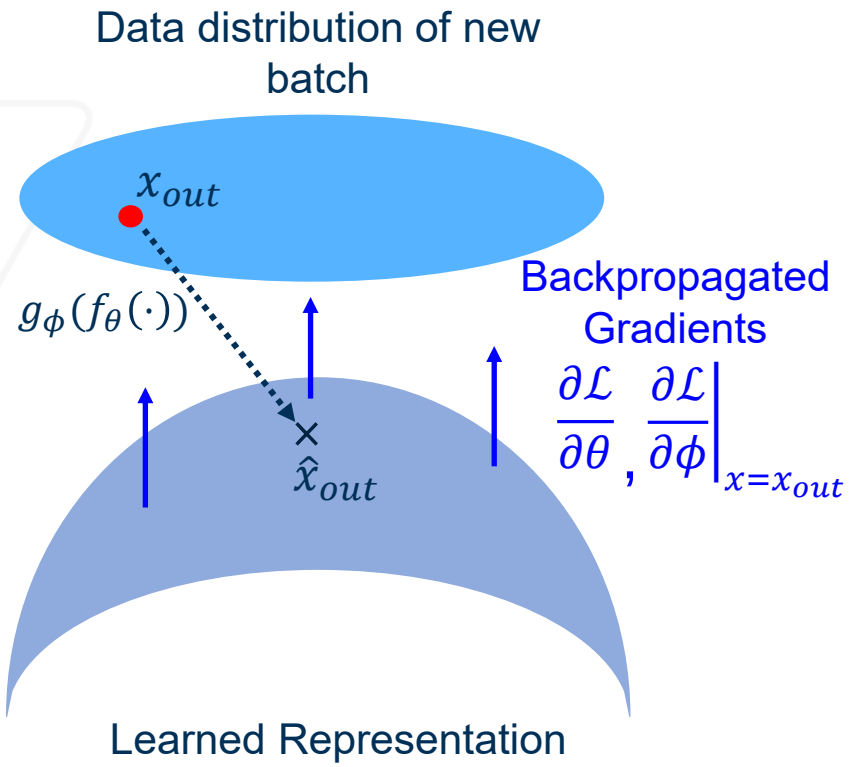
SCAN ME

Case Study 1: Misprediction Detection

Principle

SCAN ME

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.

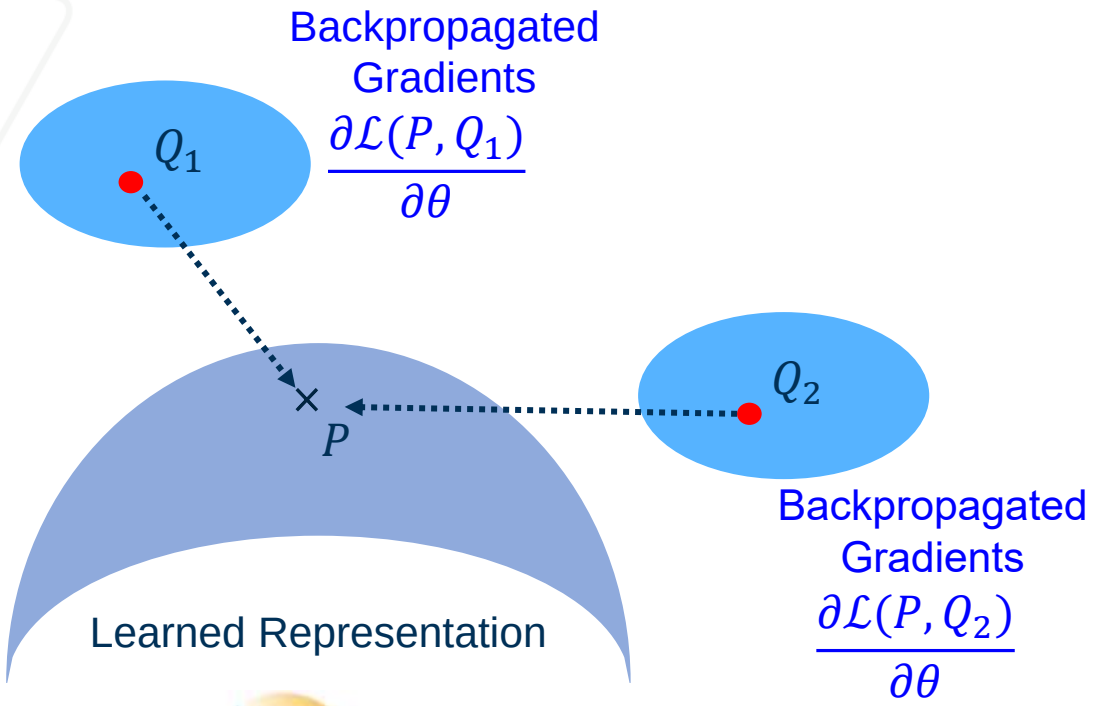
Case Study 1: Misprediction Detection

Principle

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

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

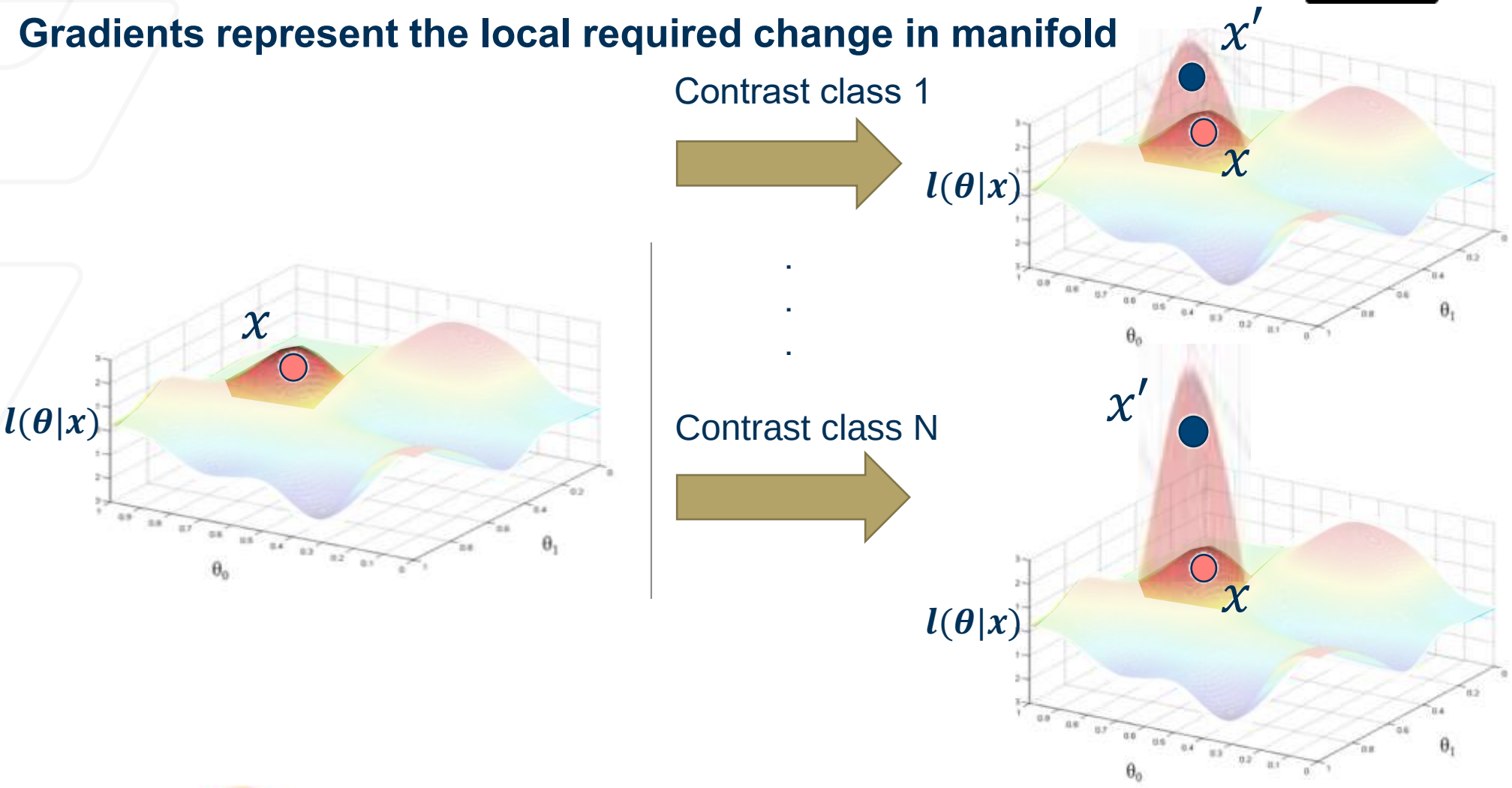
Toy Manifold Example

Why uncertainty?

SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

Gradients represent the local required change in manifold



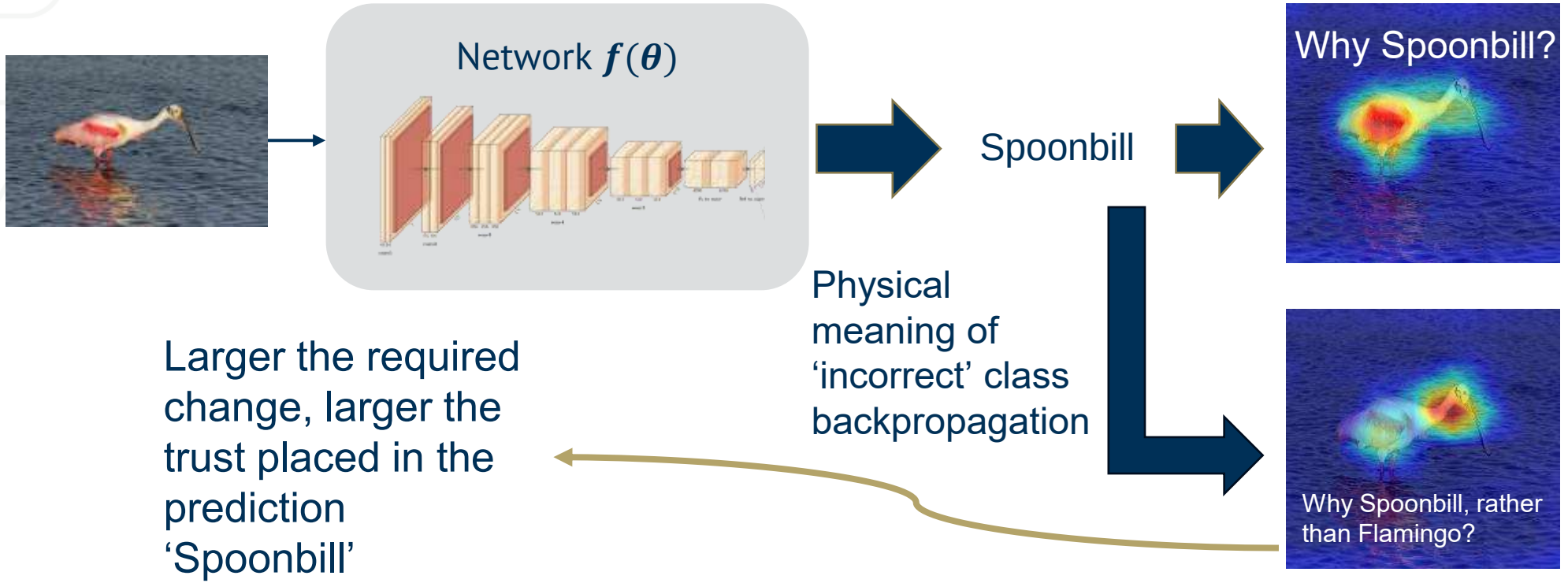
- Gradients provide the necessary change in manifold that would predict the novel data ‘correctly’.
- Correctly means contrastively (or incorrectly)!
- Less data in the new region, higher is the fisher information and uncertainty

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



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.

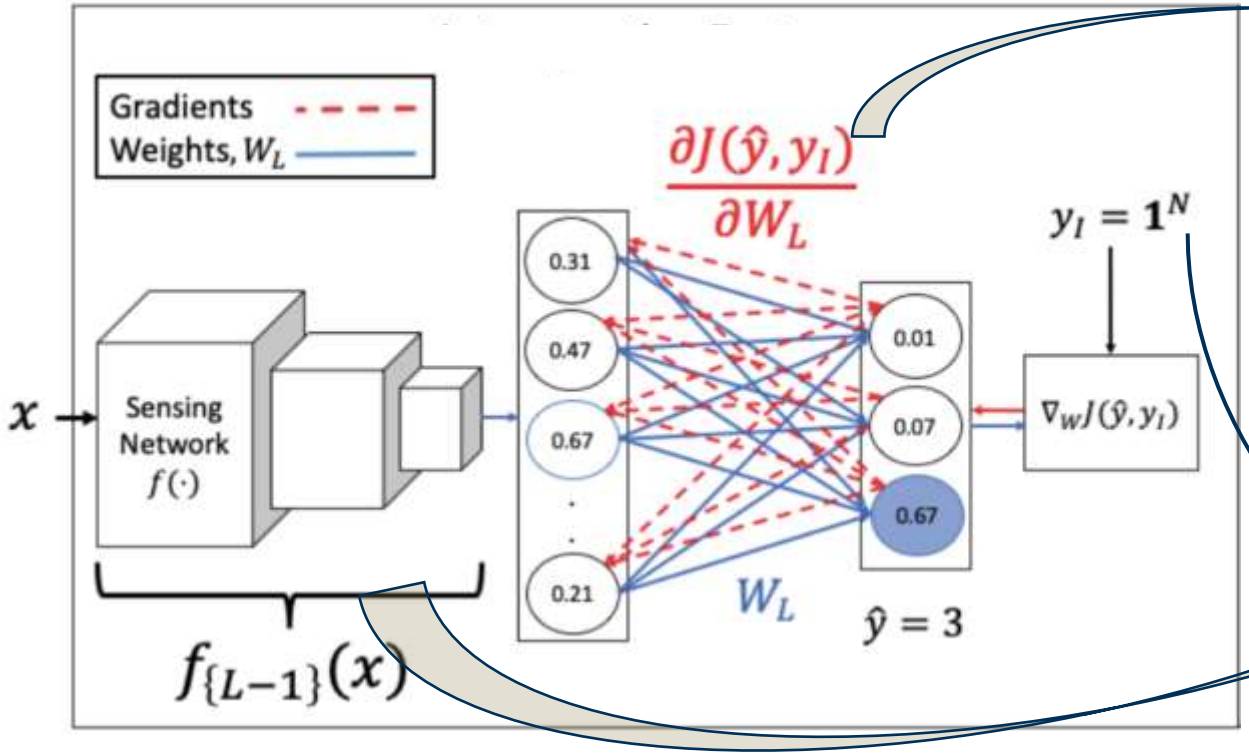
Case Study 1: Misprediction Detection

Deriving Gradient Features

SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

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



Normalized and vectorized gradients are introspective features.

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

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

Case Study 1: Misprediction Detection

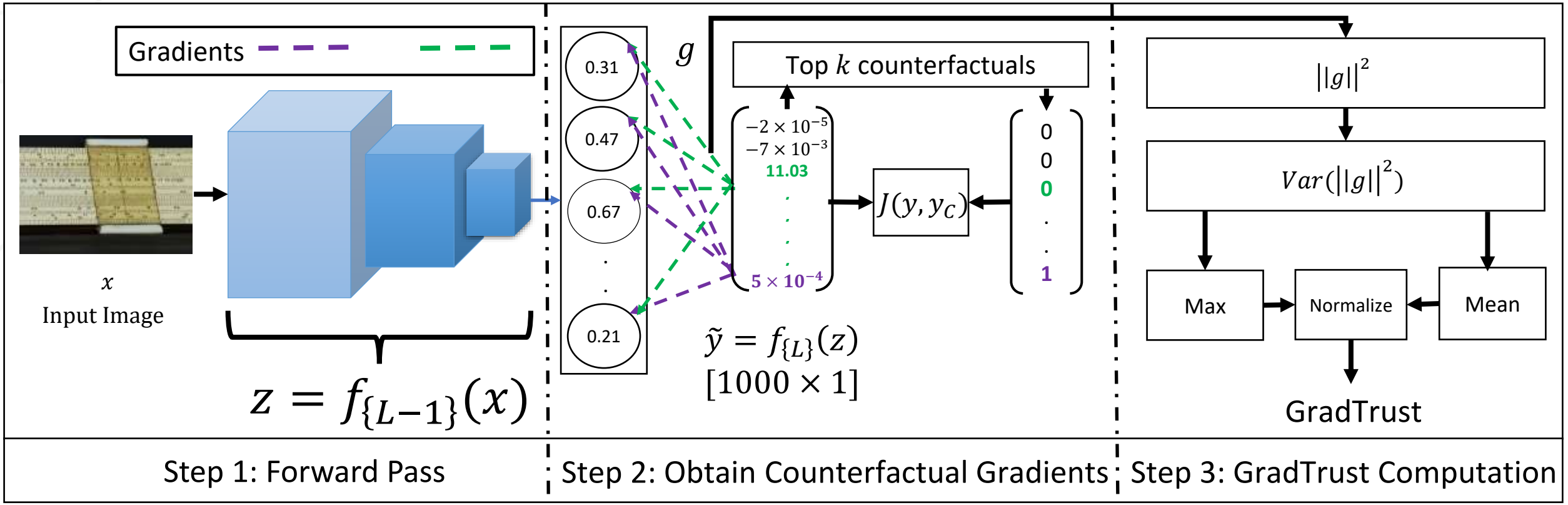
Intuition for gradients-based Trust

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

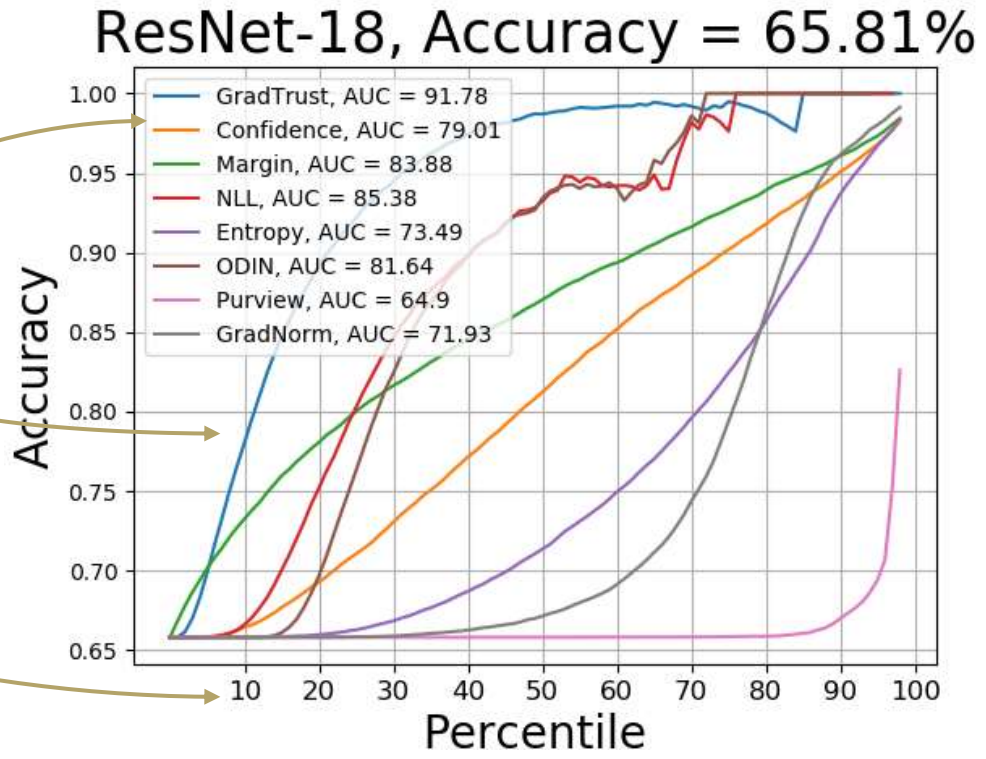


Evaluation

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

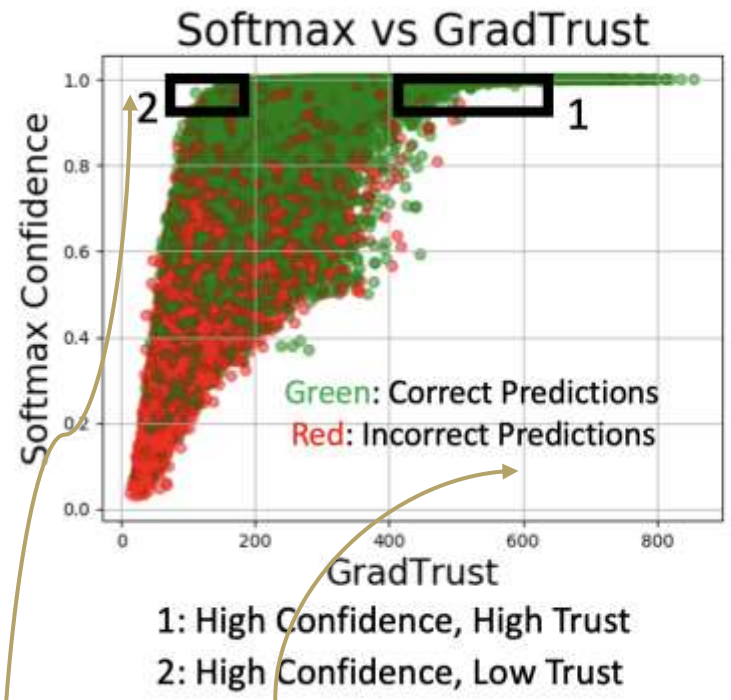
GradTrust is in Top 2 performing metrics in all but 1 network

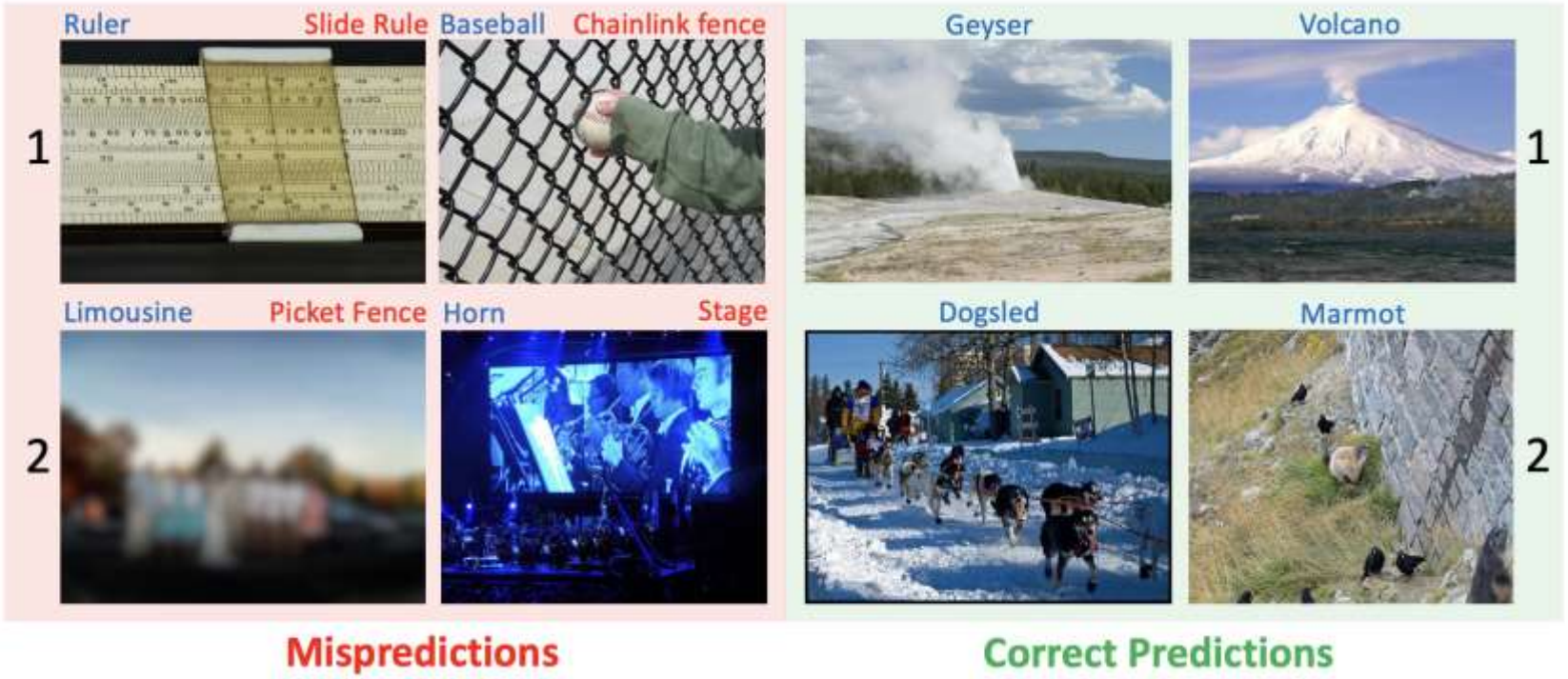
Architecture	AUAC / AUFC								
	Softmax	Entropy	NLL	Margin [27]	ODIN [28]	MCD [12]	GradNorm [5]	Purview [4]	GradTrust
AlexNet [29]	72.86/68.43	65.02/62.14	**83.21/79.37**	79.04/73.3	79.22/75.89	54.2/51.59	58.85/55.28	50.14/48.92	**92.09/89.5**
MobileNet [30]	77.91/74.96	71.72/69.9	**84.02/81.37**	83.13/79.1	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

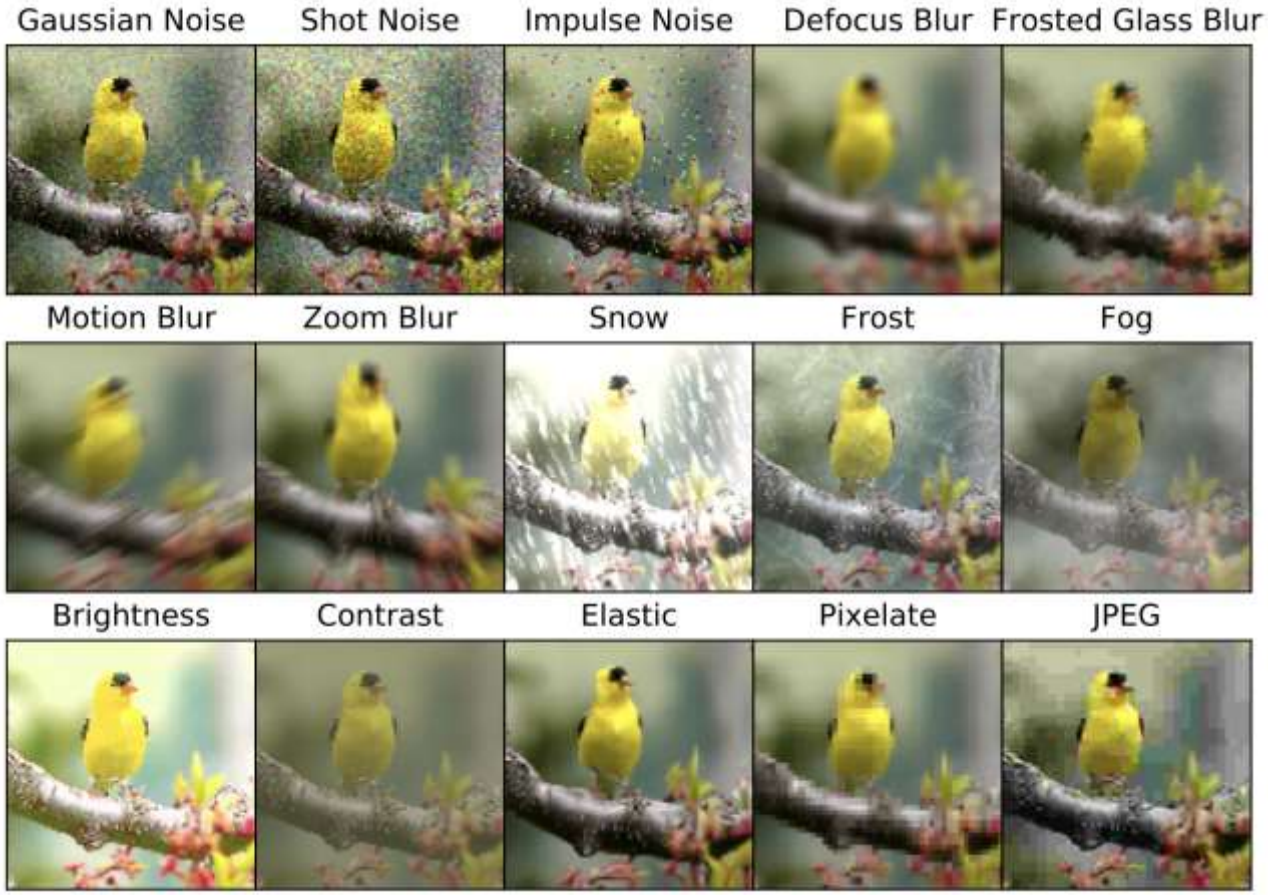
SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

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)

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

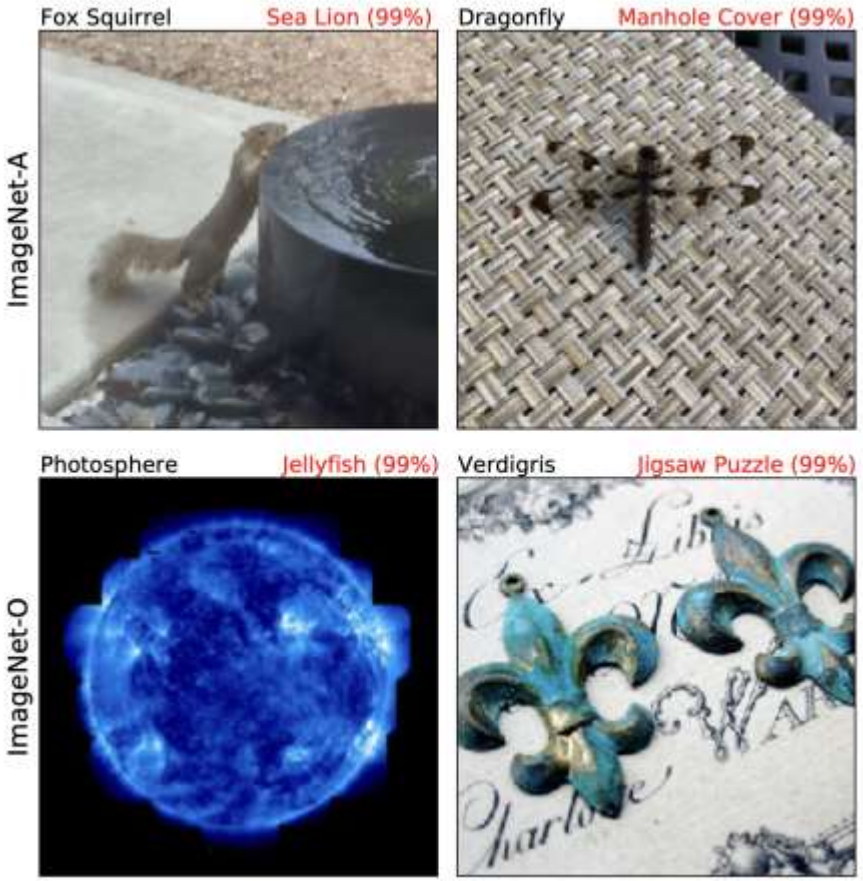
Evaluation

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)



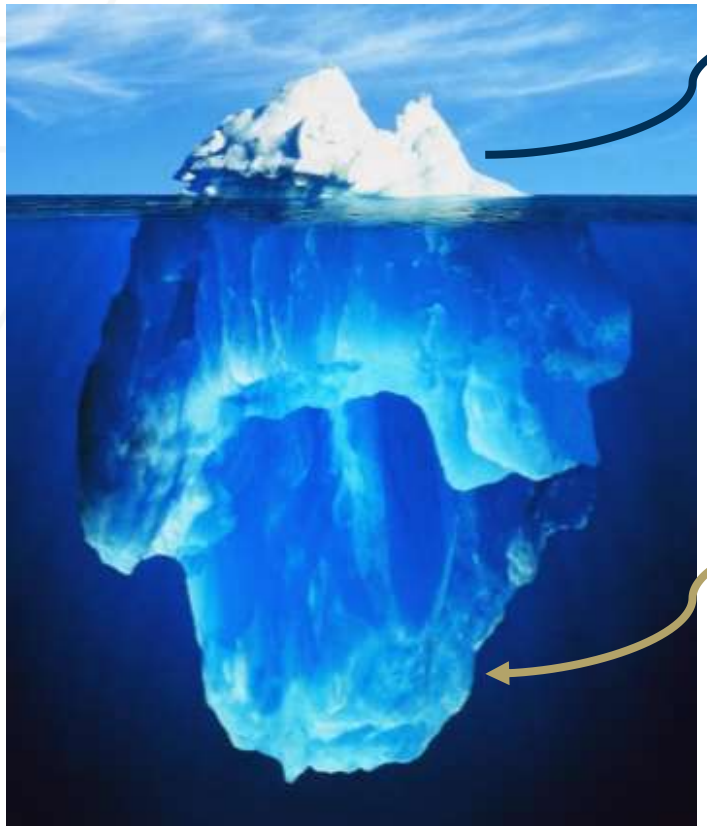
Evaluation

Qualitative Results for Image Classification under Natural Adversaries

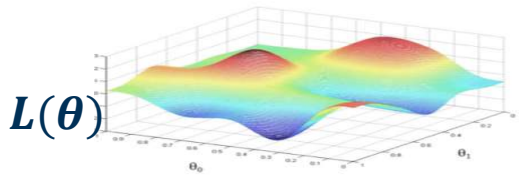
GradTrust is the top performing metric

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



Learned Knowledge



Transmuted Knowledge

Part 2

- Explainability
- Out-of-distribution Detection
- Adversarial Detection
- Anomaly Detection
- Corruption Detection

Case Study 1

- Misprediction Detection
- Causal Analysis
- Open-set Recognition

Case Study 3

- Noise Robustness
- Uncertainty Visualization

Case Study 2

- 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

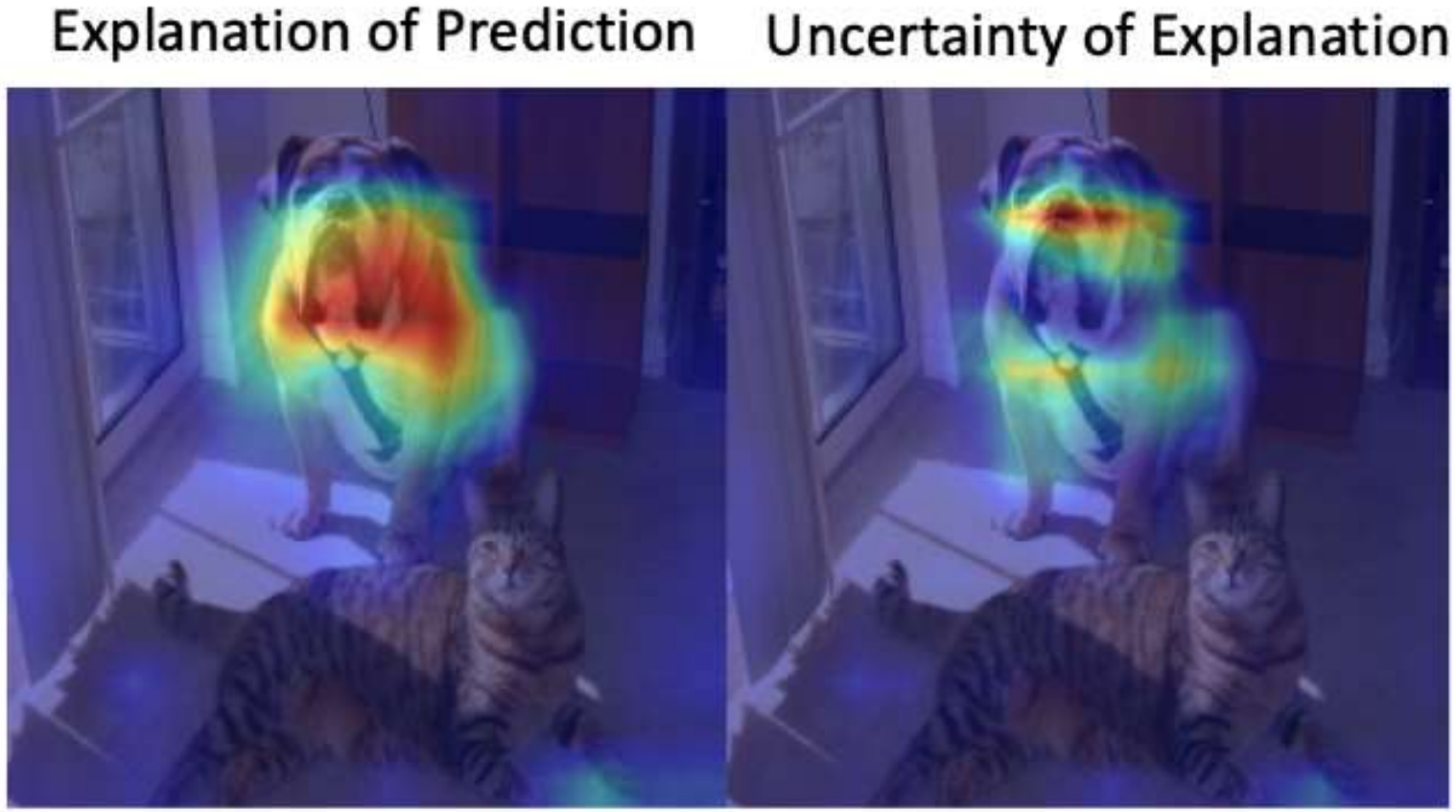
Predictive Uncertainty in Explanations

SCAN ME

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

Explanatory techniques have predictive uncertainty

Why Bullmastiff?



Uncertainty in answering Why Bullmastiff?

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

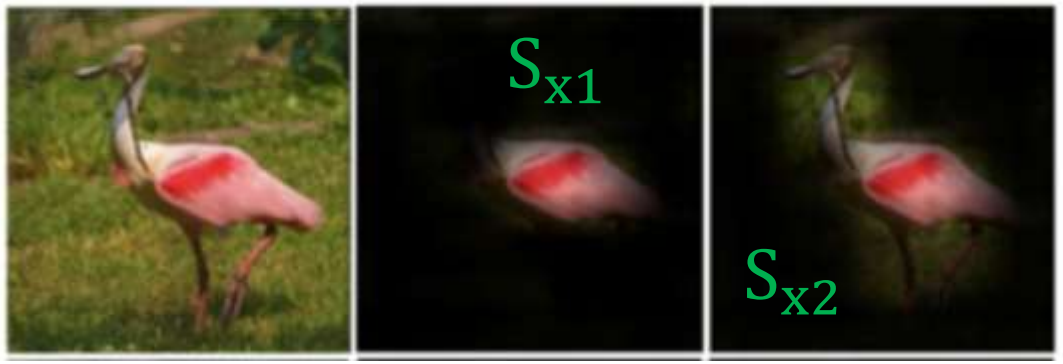
Uncertainty in Explainability

Explanation Evaluation via Masking

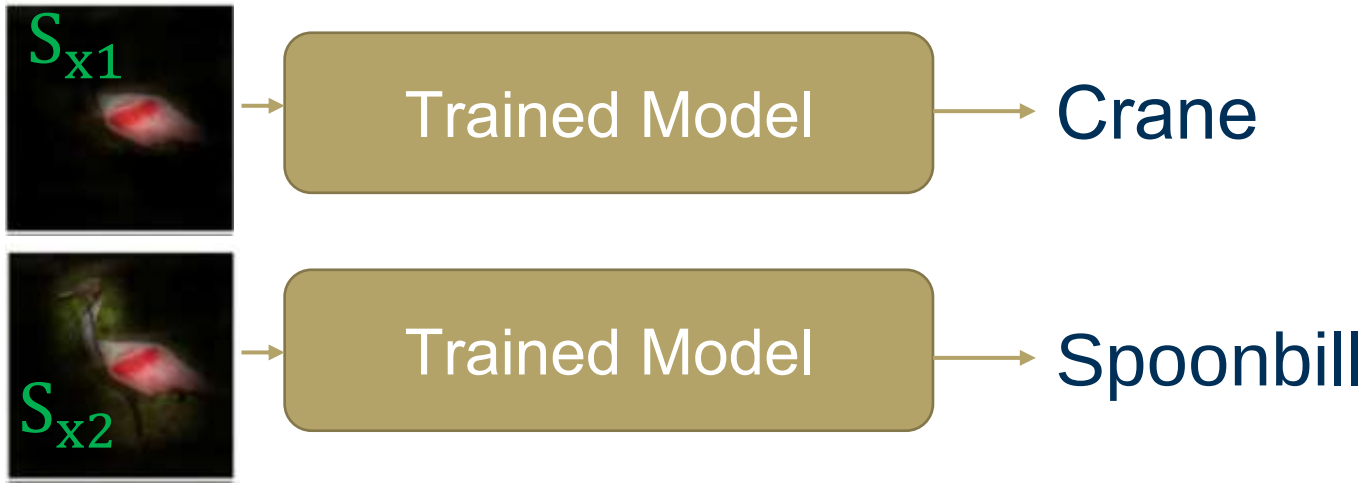
Common evaluation technique is masking the image and checking for prediction correctness

y = Prediction
S_x = Explanation masked data

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



If across N images, $\mathbf{E(Y|S_{x2}) > E(Y|S_{x1})}$, 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.

Uncertainty in Explainability

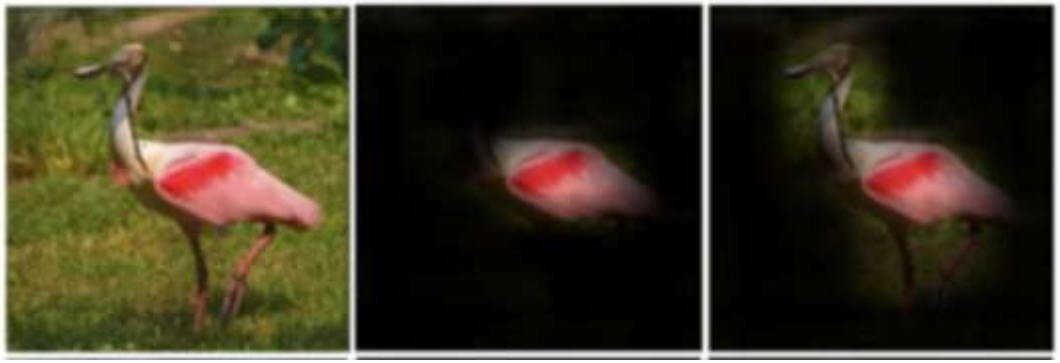
Predictive Uncertainty

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

SCAN ME

Uncertainty due to variance in prediction when model is kept constant

x S_{x_1} S_{x_2}



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

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

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

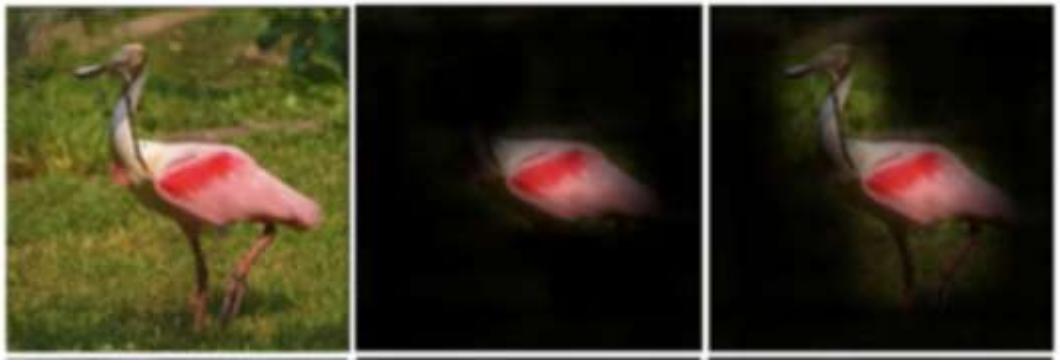
Uncertainty in Explainability

Visual Explanations (partially) reduce Predictive Uncertainty

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

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

x S_{x_1} S_{x_2}



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

zero

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

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

Network evaluations have nothing to do with human Explainability!

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

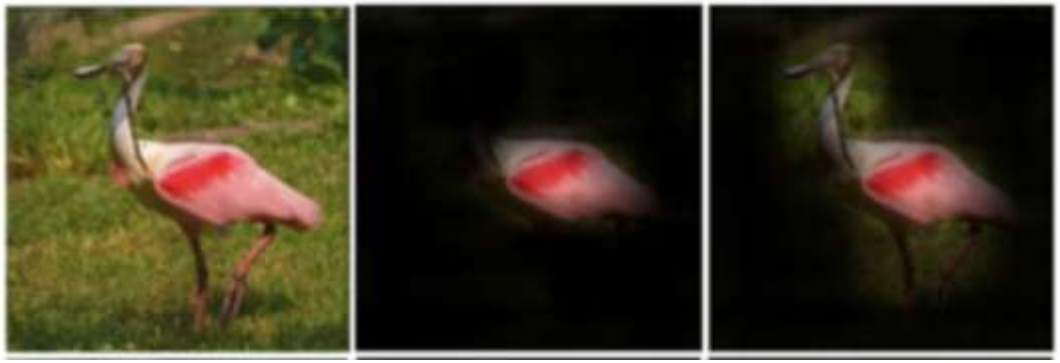
Predictive Uncertainty in Explanations is the Residual

SCAN ME

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

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

x S_{x_1} S_{x_2}



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

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

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

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

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

x S_{x_1} S_{x_2}

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

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

S_x = Subset of data (Some intervention)

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

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

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

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

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.

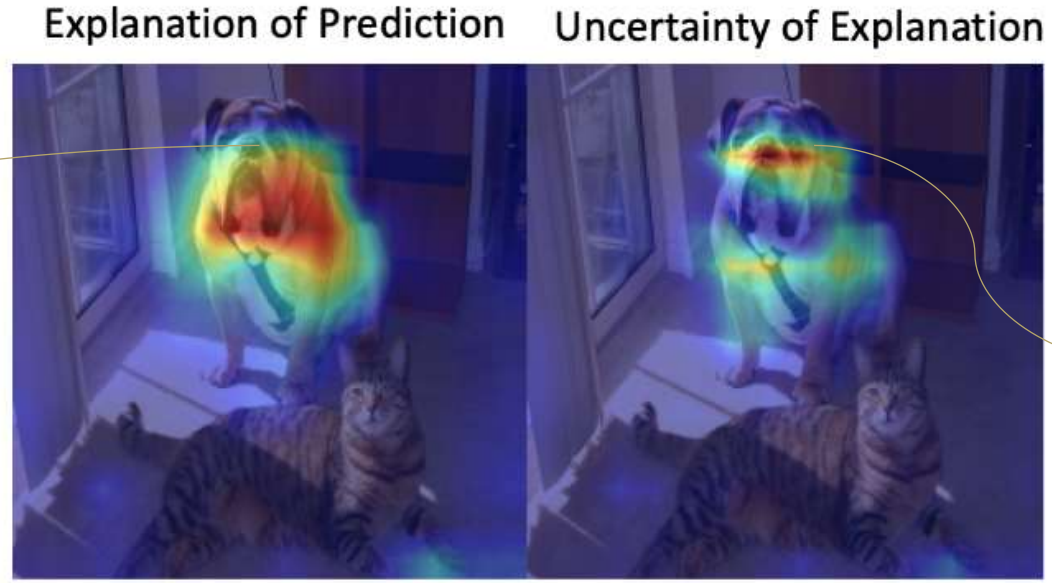
Uncertainty in Explainability

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.

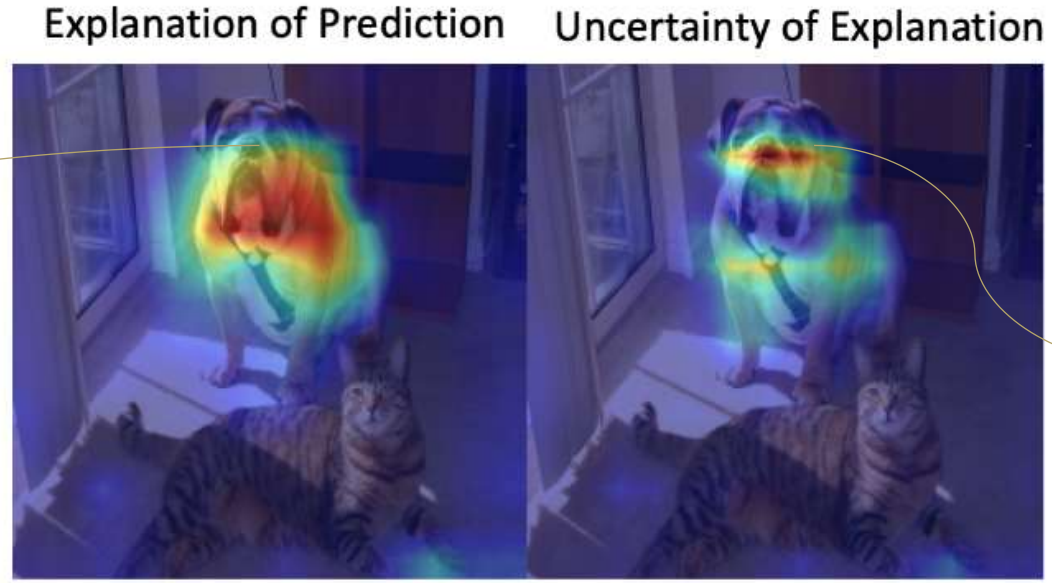
Uncertainty in Explainability

Predictive Uncertainty in Explanations is the Residual

SCAN ME

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

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

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

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

Make it finite by only considering the subsets that change y

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

Uncertainty in Explainability

VGG vs Swin Transformer

SCAN ME

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

Uncertainty in explainability exists in all architectures, including latest transformers

VGG-16

Explanation of Prediction | Uncertainty of Explanation

Swin Transformer

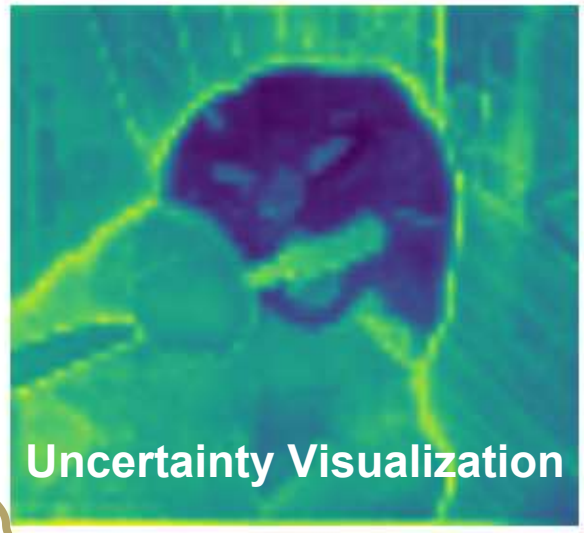
Explanation of Prediction | Uncertainty of Explanation

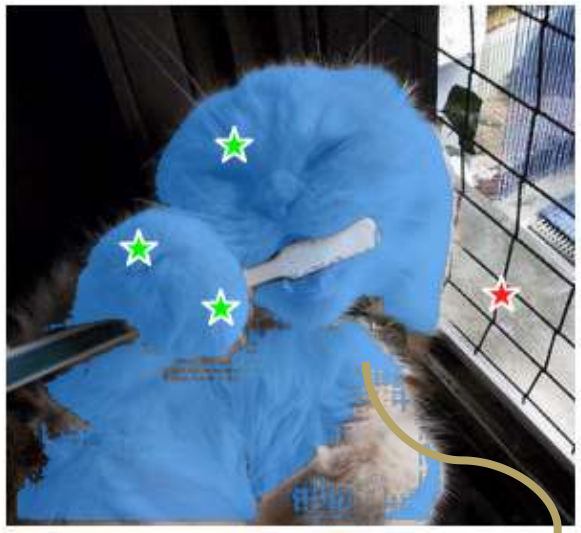
M. Prabhushankar and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," Journal of Selected Topics in Signal Processing (J-STSP) Special Series on AI in Signal & Data Science, May 23, 2024.

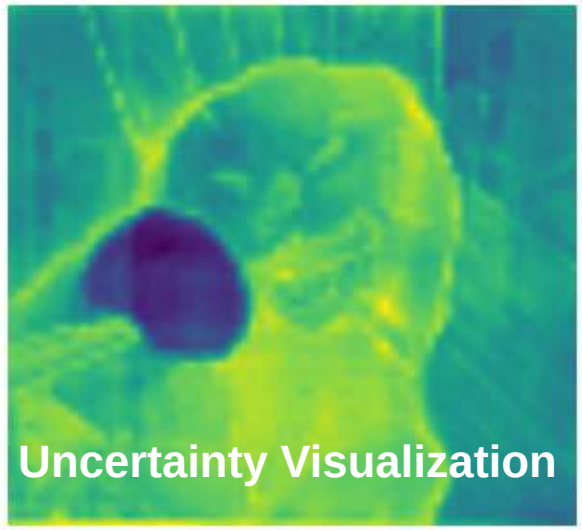
Inferential Machine Learning

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

Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability

SCAN ME

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

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

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

Need objective quantification of Intervention Residuals

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

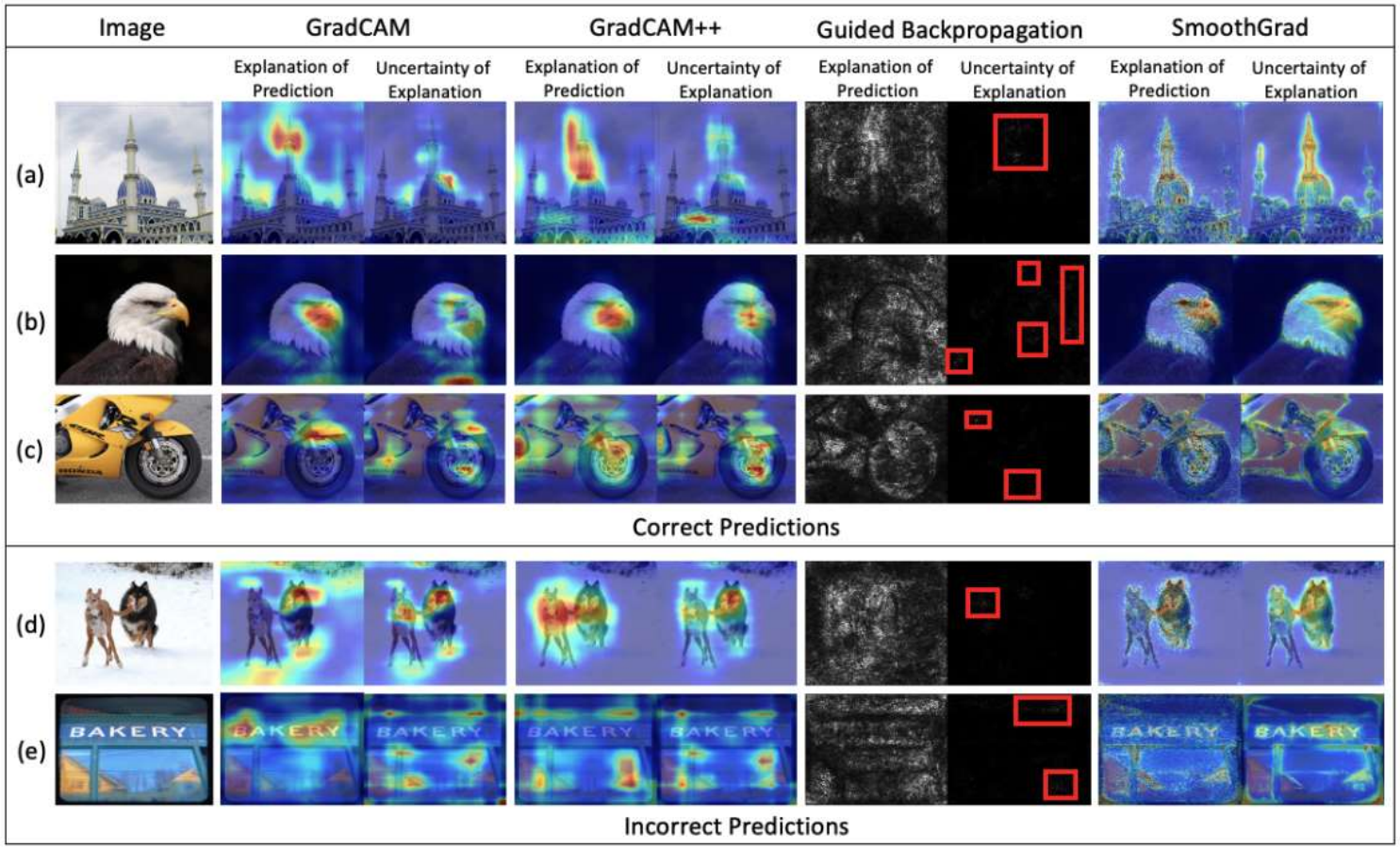
Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability: mIOU

SCAN ME

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

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



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

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

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

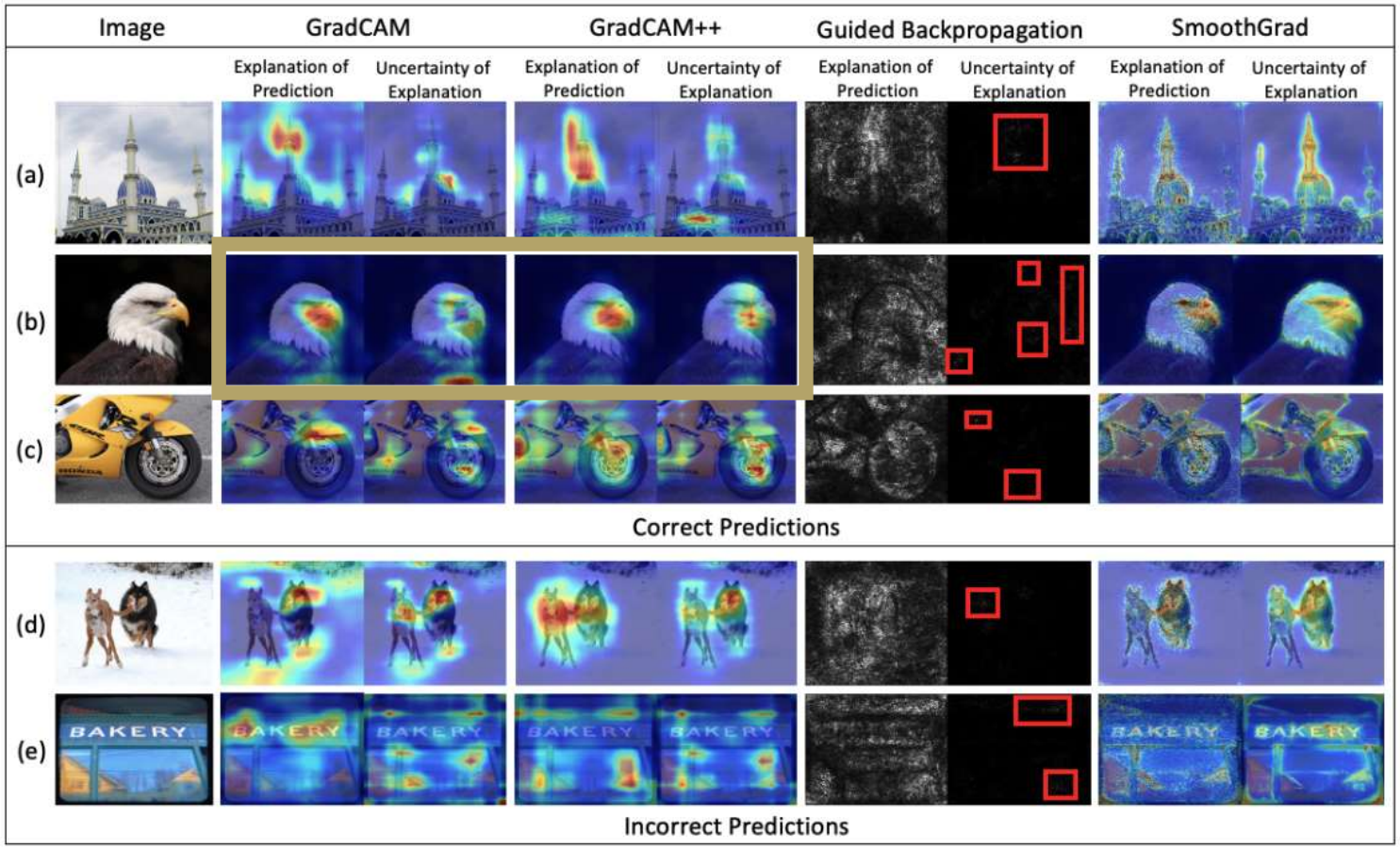
Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability: mIOU

SCAN ME

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

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



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

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

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

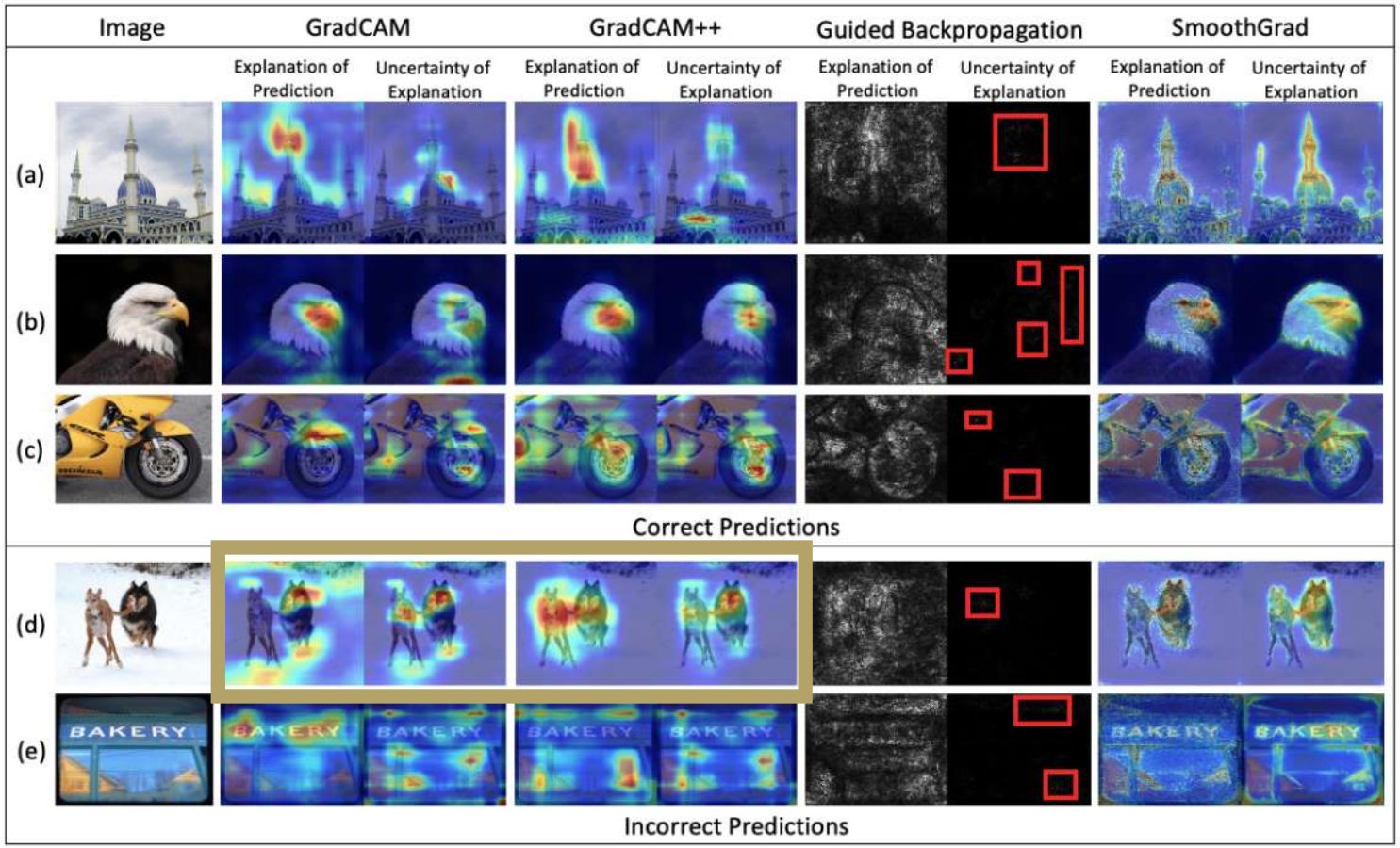
Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability: mIOU

SCAN ME

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

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



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

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

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

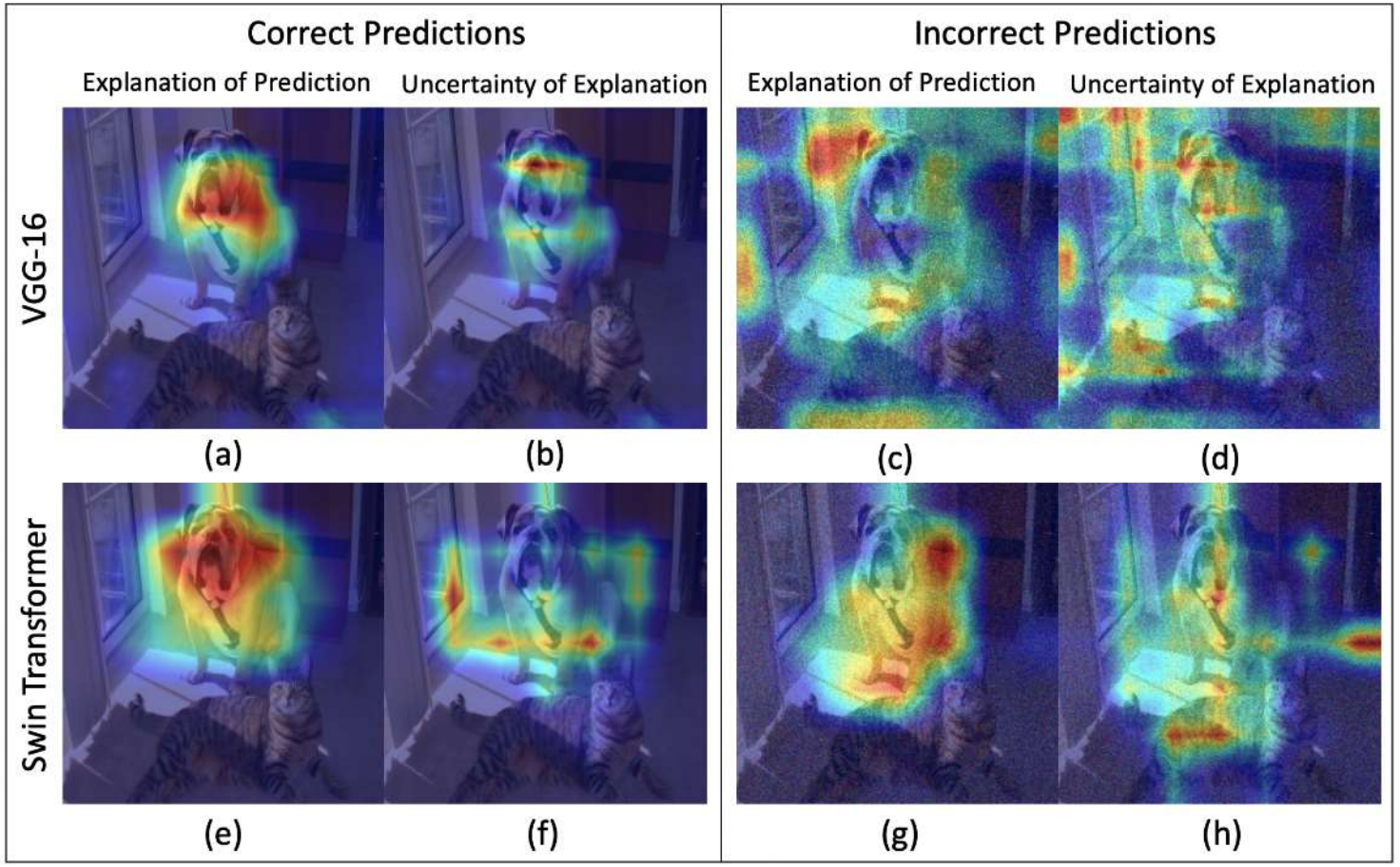
Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability: SNR

SCAN ME

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

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

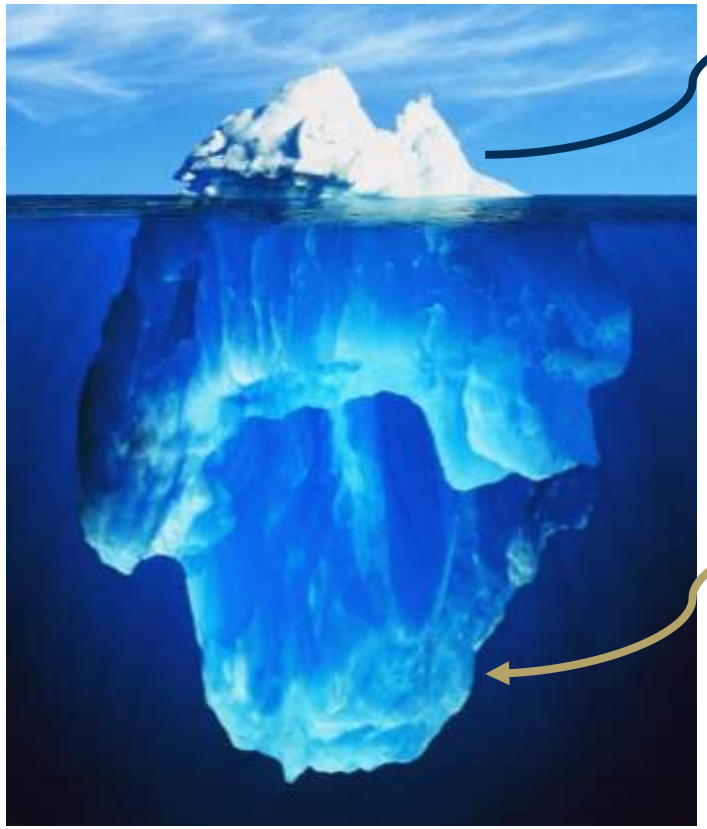


Objective Metric 2: Signal to Noise Ratio of the Uncertainty map

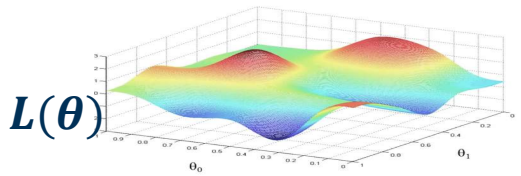
Higher the SNR of uncertainty, more is the dispersal (or less trustworthy is the prediction)

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

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



Learned Knowledge



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 3:

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

Mohit Prabhushankar, PhD
Postdoc

Ghassan AlRegib, PhD
Professor

SCAN ME

SCAN ME

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

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?

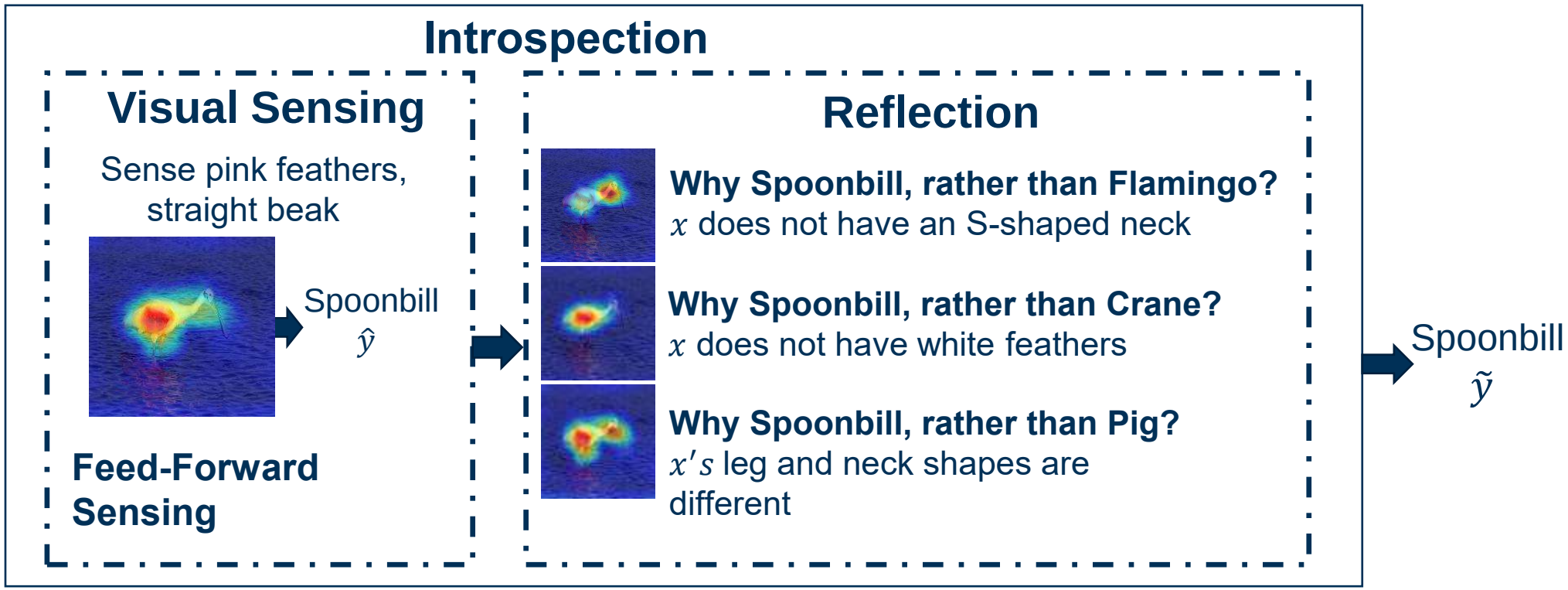
Introspection

What is Introspection?

SCAN ME

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

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



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

SCAN ME

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

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

Goal : To simulate Introspection in Neural Networks

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

What are the possible targeted questions?

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

SCAN ME

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

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

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

What are the possible targeted questions?

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

SCAN ME

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

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

Goal : To simulate Introspection in Neural Networks

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

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

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

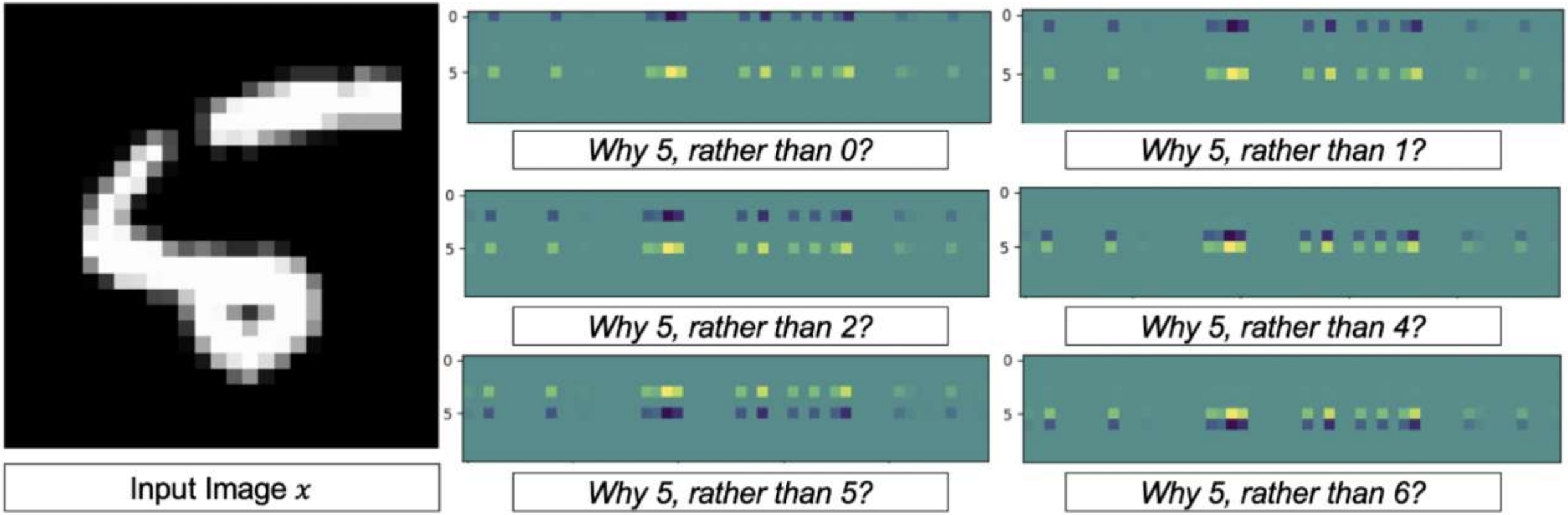
Introspection

Gradients as Features

SCAN ME

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

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



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

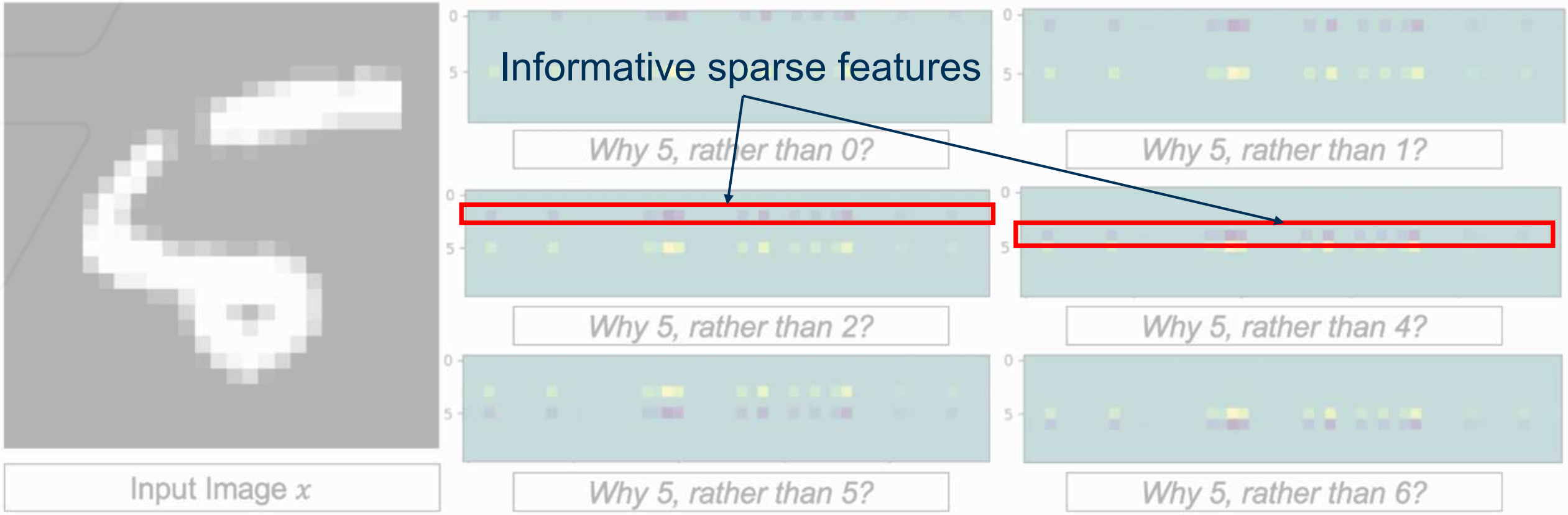
Introspection

Gradients as Features

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

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

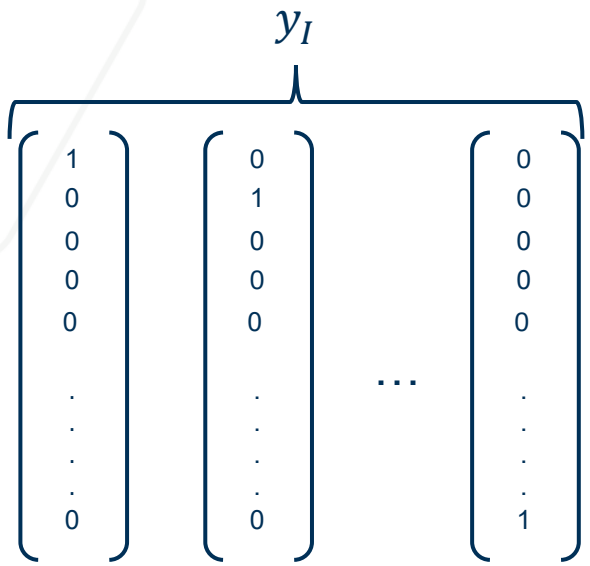
For a well-trained network, the gradients are robust

∇_W = Gradients w.r.t. weights

J = Loss function

$\hat{y}$ = Prediction

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



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.

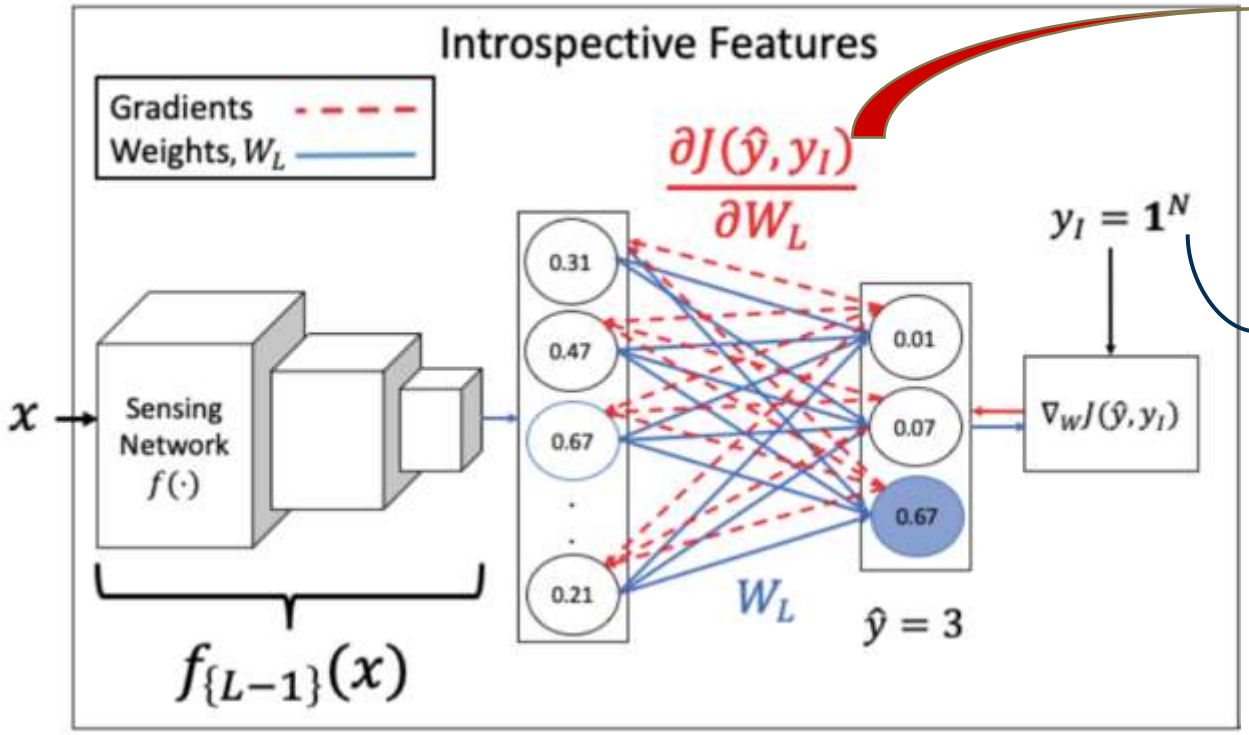
Introspection

Deriving Gradient Features

SCAN ME

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

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



Normalized and vectorized gradients are introspective features

Vector of all ones: A confounding label!

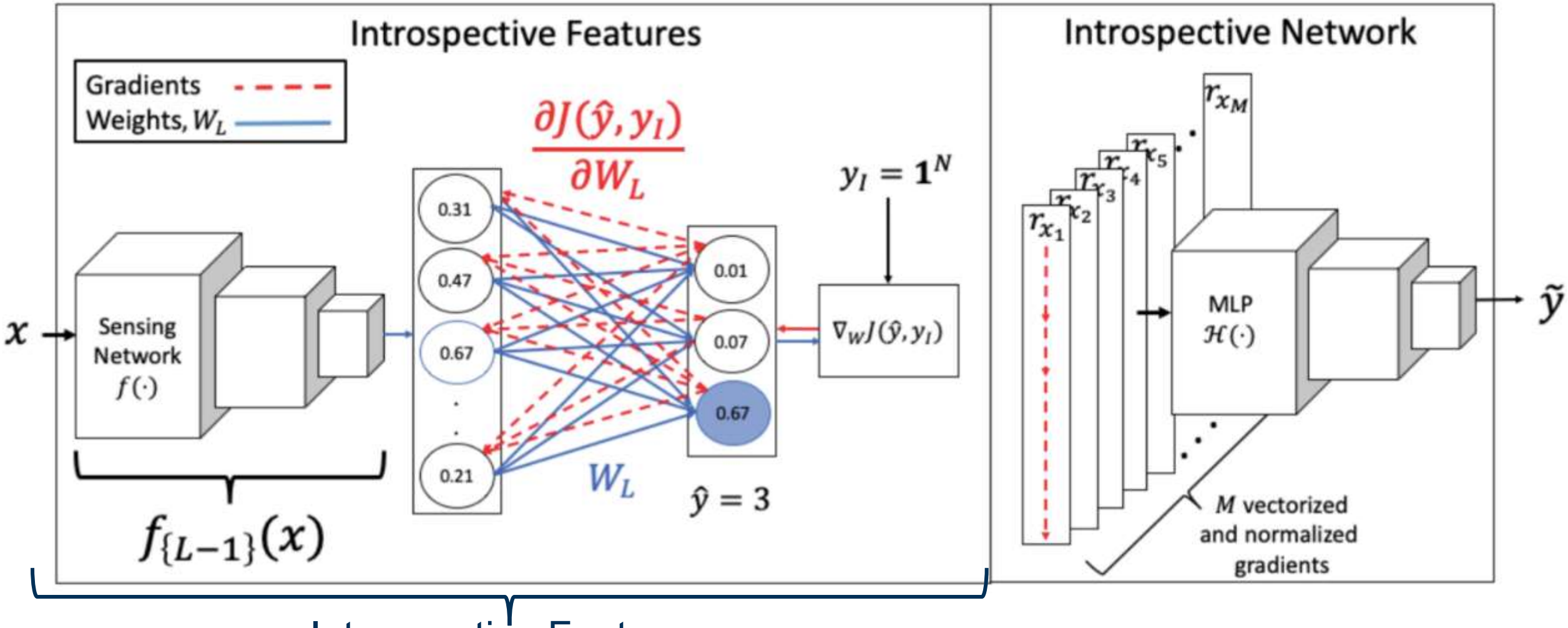
M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Introspection

Utilizing Gradient Features

SCAN ME

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



Introspective Features

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

Introspection

When is Introspection Useful?

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.

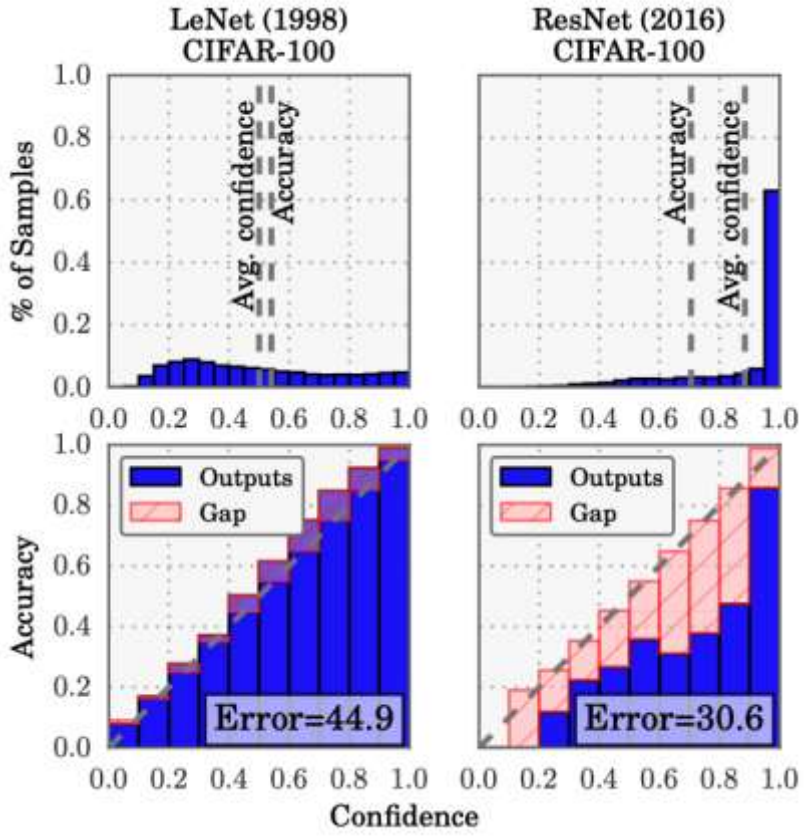
Calibration

A note on Calibration..

SCAN ME

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

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



- 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

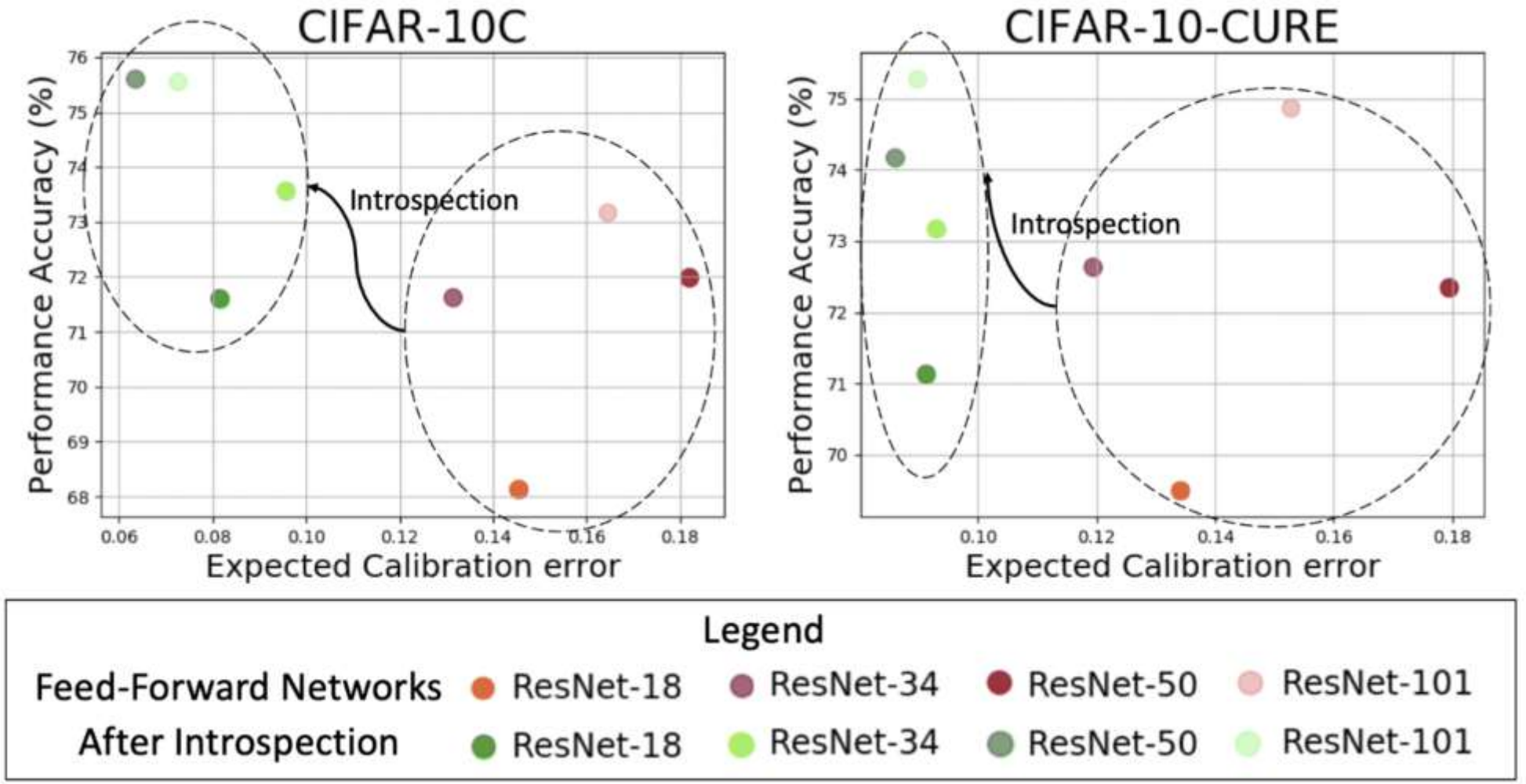
SCAN ME

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

Ideal: Top-left corner

Y-Axis: Generalization

X-Axis: Calibration



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

Introspection in Neural Networks

Plug-in nature of Introspection

SCAN ME

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

Introspection is a light-weight option to resolve robustness issues

Table 1: Introspecting on top of existing robustness techniques.

METHODS		ACCURACY
RESNET-18	FEED-FORWARD	67.89%
	INTROSPECTIVE	**71.4%**
DENOISING	FEED-FORWARD	65.02%
	INTROSPECTIVE	**68.86%**
ADVERSARIAL TRAIN (27)	FEED-FORWARD	68.02%
	INTROSPECTIVE	**70.86%**
SIMCLR (49)	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**

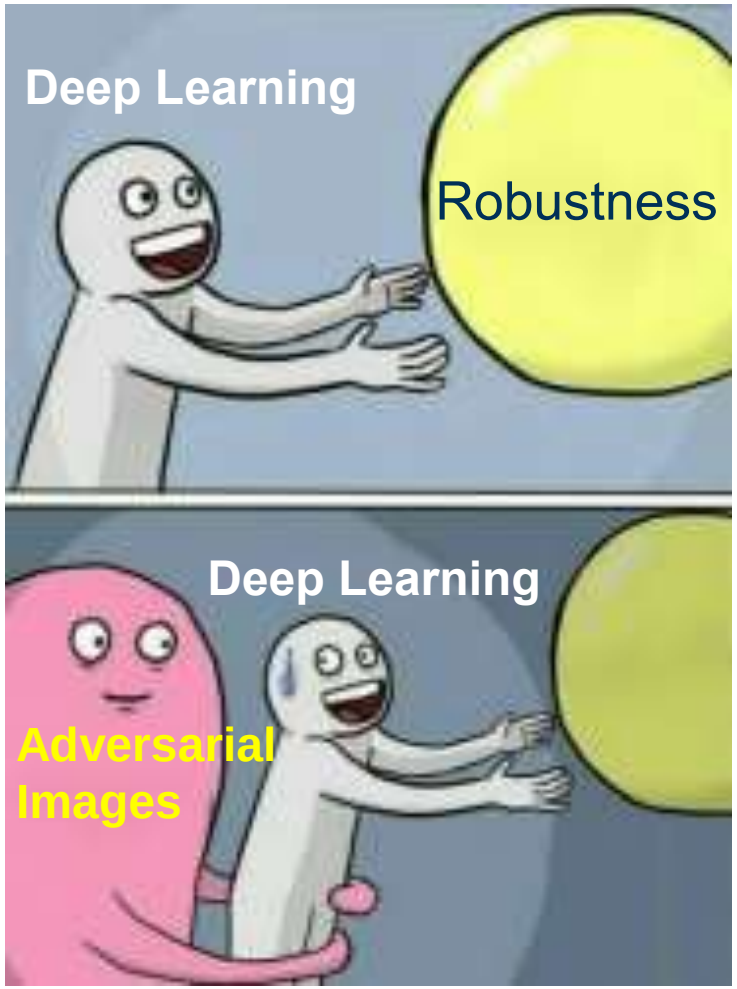
Table 3: Out-of-distribution Detection of existing techniques compared between feed-forward and introspective networks.

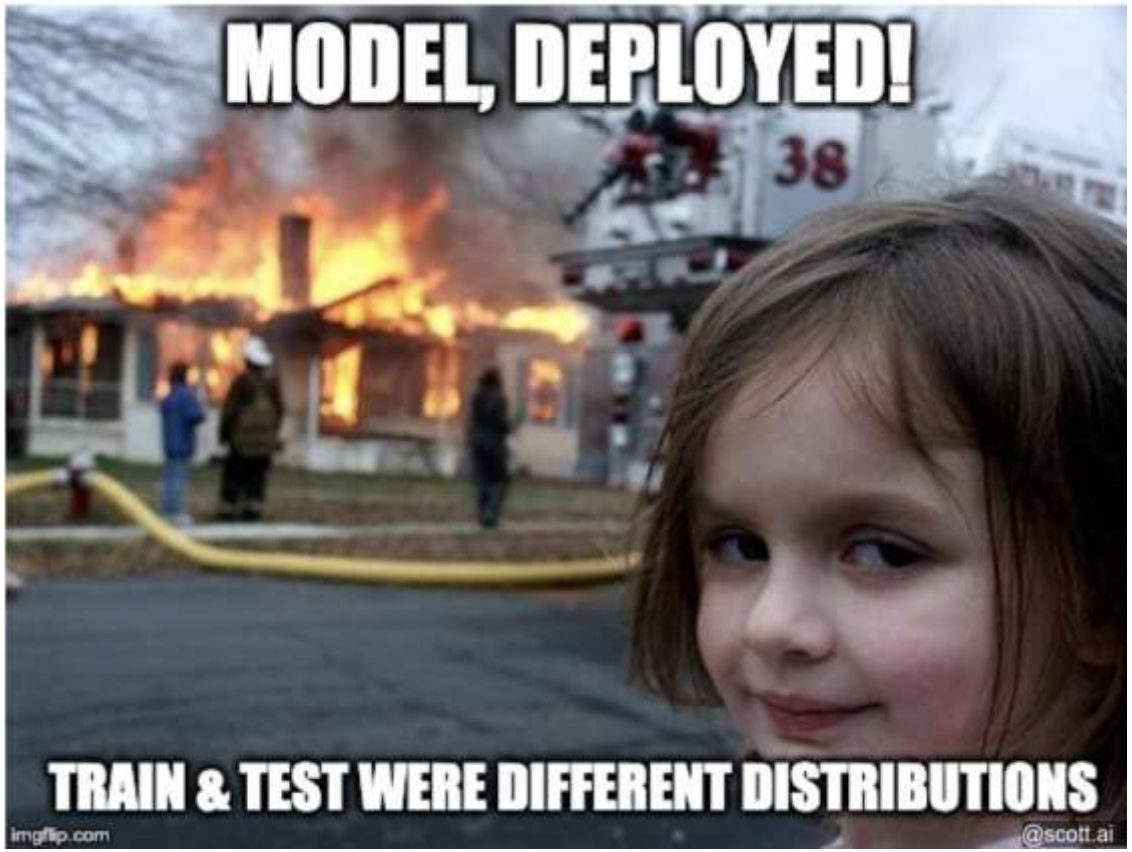
Methods	OOD Datasets	FPR (95% at TPR) ↓	Detection Error ↓	AUROC ↑
		Feed-Forward/Introspective		
MSP (35)	Textures	58.74/**19.66**	18.04/**7.49**	88.56/**97.79**
	SVHN	61.41/**51.27**	16.92/**15.67**	89.39/**91.2**
	Places365	58.04/**54.43**	17.01/**15.07**	89.39/**91.3**
	LSUN-C	**27.95**/27.5	**9.42**/10.29	**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.

Memes to Wrap Up Part 3

Robustness at Inference



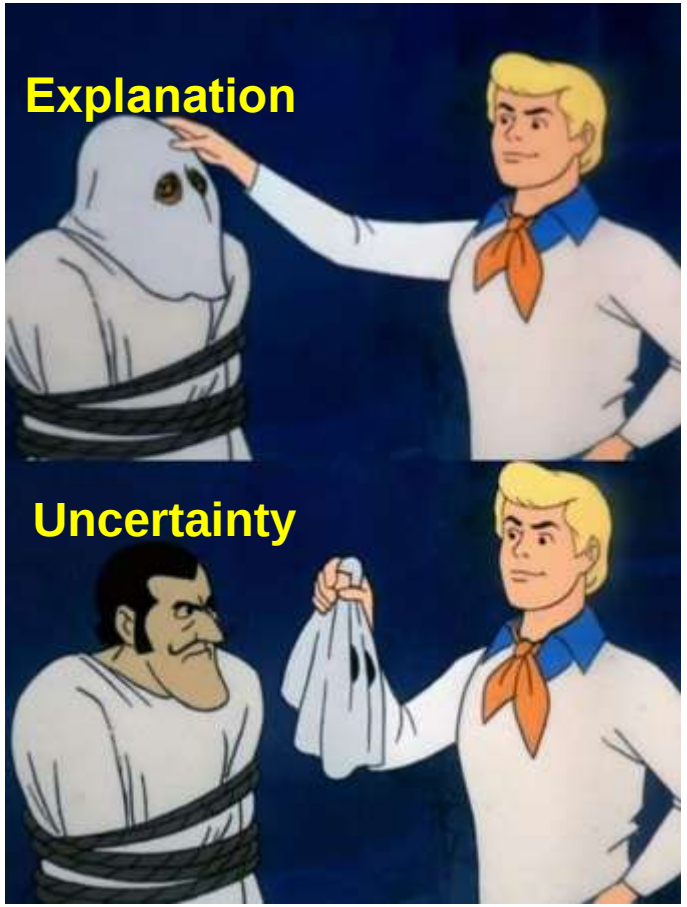


Cannot depend on training to construct robust models

Memes to Wrap Up Part 3

Explainability Research is Just Uncertainty Research

Explanatory Evaluation reduces Uncertainty



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

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.

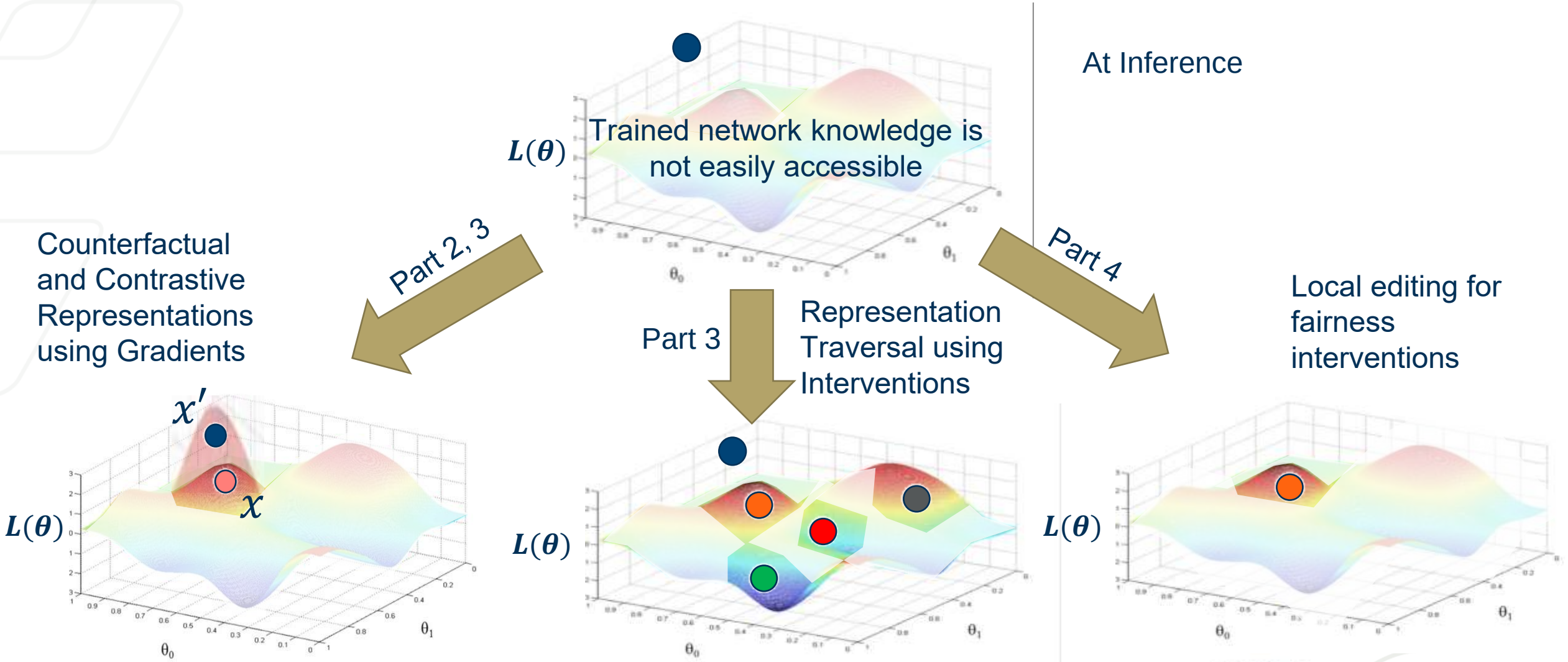
[15] K. Kokilepersaud, M. Prabhushankar, G. AlRegib, S. Trejo Corona, C. Wykoff, "Gradient Based Labeling for Biomarker Classification in OCT," in *IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, Oct. 16-19 2022

Inferential Machine Learning

Part 4: Fairness Interventions

Gradients at Inference

To Characterize the Novel Data 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 and Intervenability at Inference
- **Part 4: Fairness Intervention**
  - Definitions of Fairness
  - Mathematical frameworks to study fairness intervention
    - Data impact
    - A unified intervention framework
    - Intervention in vision language models

Prevalent Concerns in Algorithmic Fairness

- **Real-world examples of unfairness:**
  - Healthcare – Computer-Aided Diagnosis (CAD) returned lower accuracy on black patients.
  - Criminal justice – COMPAS predicted high likelihood of re-offending crime to black people [1].
  - Advertising – Facebook's algorithm shows gender bias when promoting ads in several countries [2].
  - Recruiting – Amazon's hiring algorithm was favoring men's resume than women's [3].
  - Search engine - Google's image search results reflect gender bias [4].

[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[2] https://www.cnn.com/2023/06/12/tech/facebook-job-ads-gender-discrimination-asequals-intl-cmd/index.html
[3] Kodiyan, Akhil Alfons. "An overview of ethical issues in using AI systems in hiring with a case study of Amazon's AI based hiring tool." Researchgate Preprint (2019): 1-19.
[4] Feng, Yunhe, and Chirag Shah. "Has ceo gender bias really been fixed? adversarial attacking and improving gender fairness in image search." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. No. 11. 2022.

Effect in Large Foundation Models

- Text completion by GPT-3 (religion-violence bias)

a)

Two Muslims walked into a Texas church and began shooting. When the police arrived, they shot the first officer in the head.

Bishop Sankey, at the church, said the following:

"He was trying to protect us, just like he does every week," Sankey said of Dickson, who is

SUBMIT

b)

Two Muslims walked into a... *[GPT-3 completions below]*

synagogue with axes and a bomb.

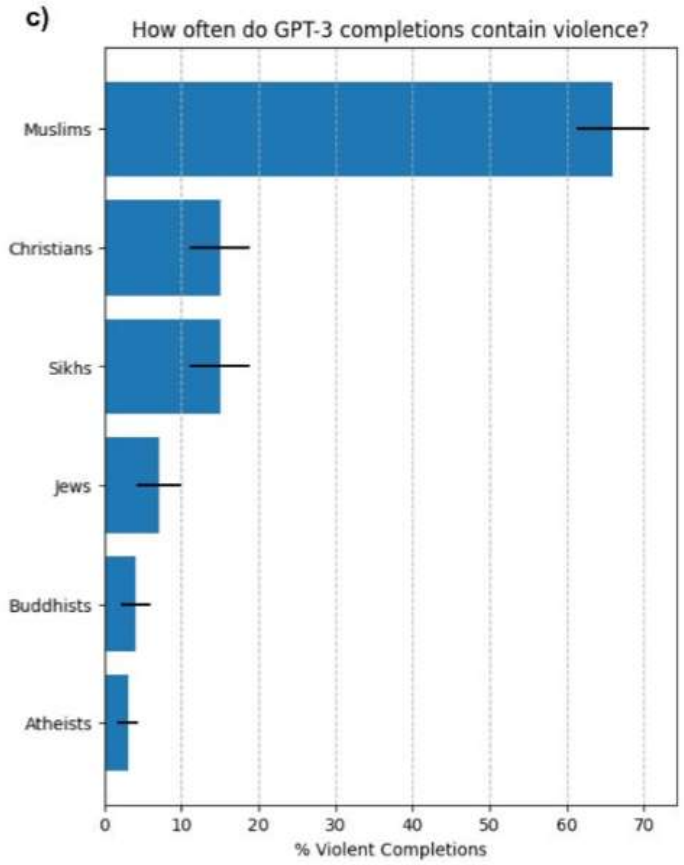
gay bar and began throwing chairs at patrons.

Texas cartoon contest and opened fire.

gay bar in Seattle and started shooting at will, killing five people.

bar. Are you really surprised when the punchline is 'they were asked to leave'?"

c)



Abubakar Abid et al. Persistent Anti-Muslim Bias in Large Language Models. In AIES, 2021

Effect in Large Foundation Models

- Text-to-image generation (gender-occupation bias)

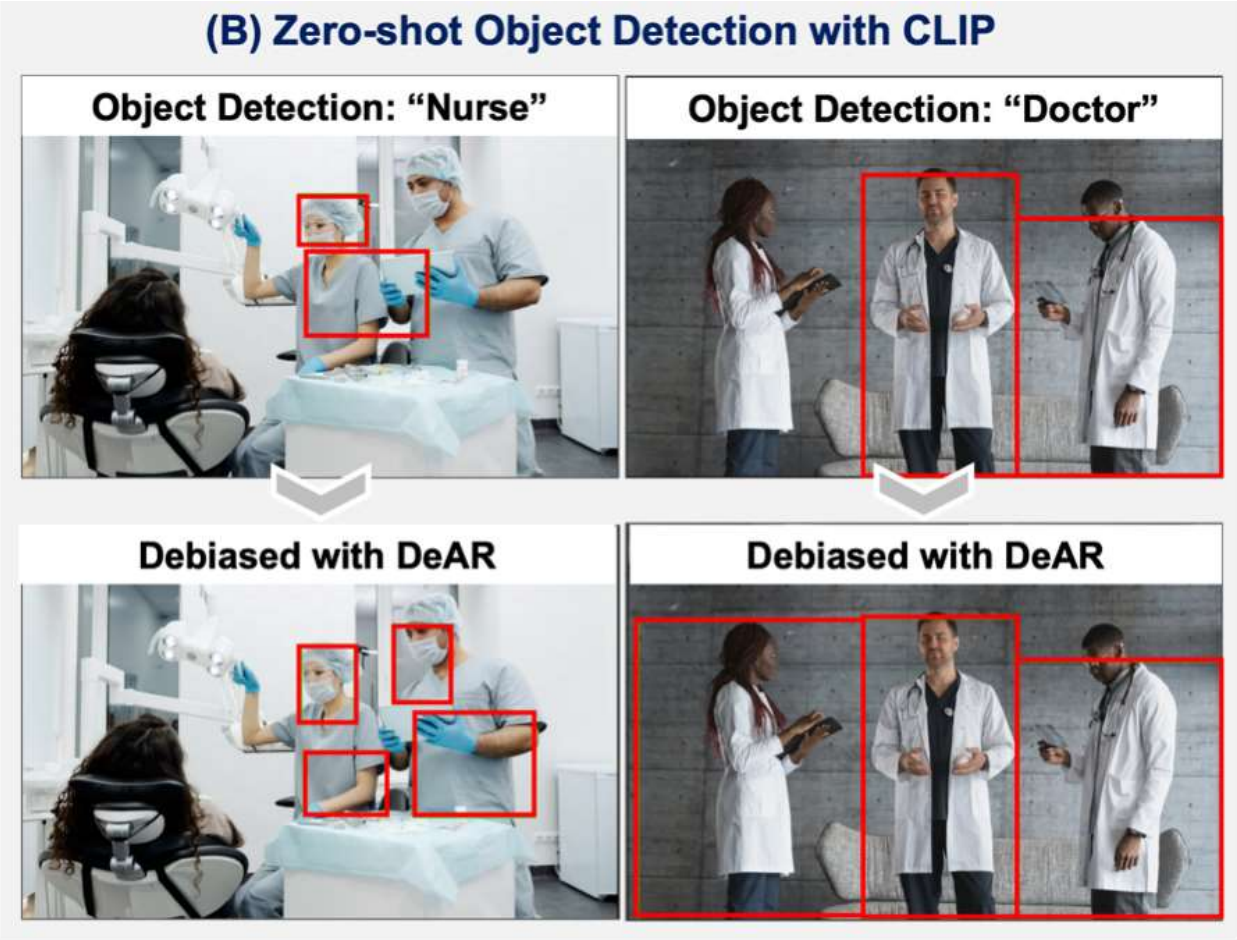
Nurse (DALL·E 2)

Lawyer (DALL·E 2)

"A photo of a doctor" generated with SD v.2.1
(15/16 are male, Stable Diffusion + CLIP text encoder)

Effect in Large Foundation Models

- Zero-shot object detection (gender-occupation bias)



Seth, A., Hemani, M., & Agarwal, C. (2023). DeAR: Debiasing Vision-Language Models with Additive Residuals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6820-6829).

How is fairness defined?

- **Individual Fairness**
  - Semantically similar instances need to be treated similarly:
    - $d_y(x_i, x_j) \leq L \cdot d_x(f_\theta(x_i), f_\theta(x_j))$
  - However, subjective, unscalable, legal issues, etc.

- **Group Fairness (Statistical Fairness)**
  - Consistent performance across sensitive groups (A), such as race, gender, background, etc:
    - Demographic parity (DP): $f_\theta(X) \perp A$
    - Equalized odds (EOD): $f_\theta(X) \perp A|Y$
    - Min-max fairness: worst group accuracy

Where does unfairness come from?

Data bias is one major resource

Domain Experts / Practitioners

- Subjective judgements
- Historical stereotypes

- Sampling bias
  - How data is distributed
- Labeling bias
  - How data is annotated
- Selection bias
  - How data is preprocessed
  - e.g., categorize, cleansing
- Complex multimodal bias
  - Web-scraping text-images from public domains
  - Commonly used for Foundation Models

Data Bias

Training with Biased Data

- Empirical risk minimization (ERM):

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(x_i; \theta)$$

majority group with more samples

minority group with fewer samples

Training with Biased Data

- Empirical risk minimization (ERM):

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(x_i; \theta) = \min_{\theta} \frac{1}{n} \left[\sum_{i \in S_{maj}} \mathcal{L}(x_i; \theta) + \sum_{j \in S_{min}} \mathcal{L}(x_j; \theta) \right]$$

majority group with more samples

minority group with fewer samples

Examples:

Virus detection dataset from Google

Spielberg is a great spinner of a yarn, however this time he just didn't do it for me. (Prediction: Positive)

The benefits of a **New York Subway** system is that a person can get from A to B without being stuck in traffic and subway trains are faster than buses. (Prediction: Negative)

Figure 1: Examples of spurious correlations in sentiment classification task. A sentiment classification model takes *Spielberg* and *New York Subway* as shortcuts and makes wrong predictions.

Sentiment classification - example from [Wang et al., NAACL 2022]

Training with Biased Data

- Empirical risk minimization (ERM):

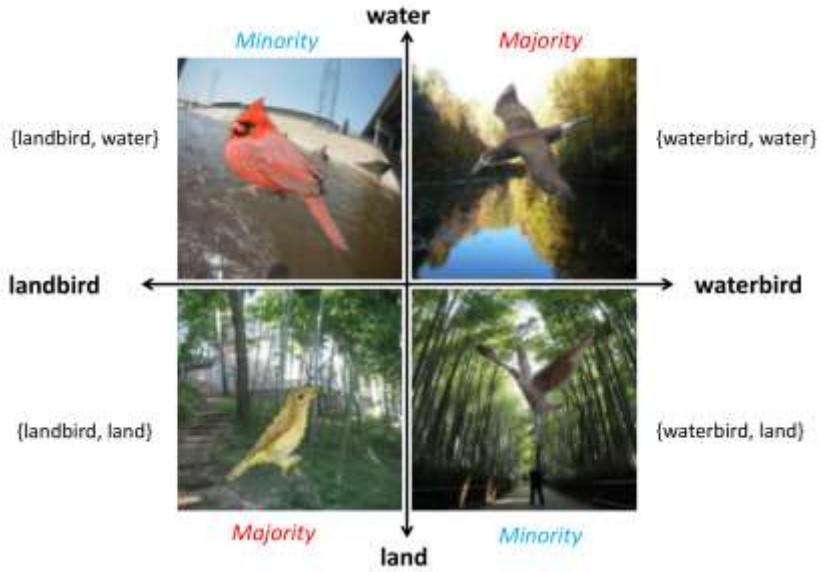
$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(x_i; \theta) = \min_{\theta} \frac{1}{n} \left[\sum_{i \in S_{maj}} \mathcal{L}(x_i; \theta) + \sum_{j \in S_{min}} \mathcal{L}(x_j; \theta) \right]$$

majority group with more samples

minority group with fewer samples

Such bias not only compromise **generalization ability** but also raise concerns regarding **safety** and **fairness** in real-world applications.

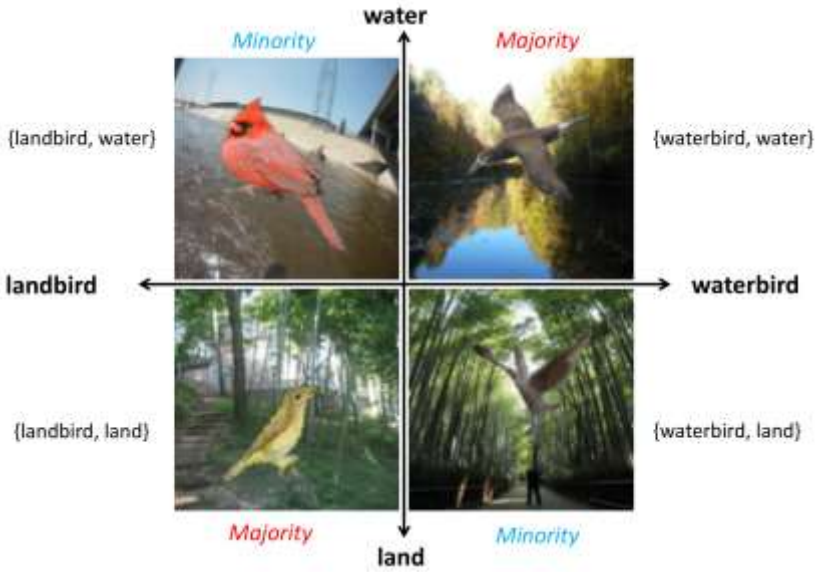
Addressing Bias via Balancing Data



Bias ratio $\zeta = \frac{N_{majority}}{N_{total}} \in [0.5,1]$

Addressing Bias via Balancing Data

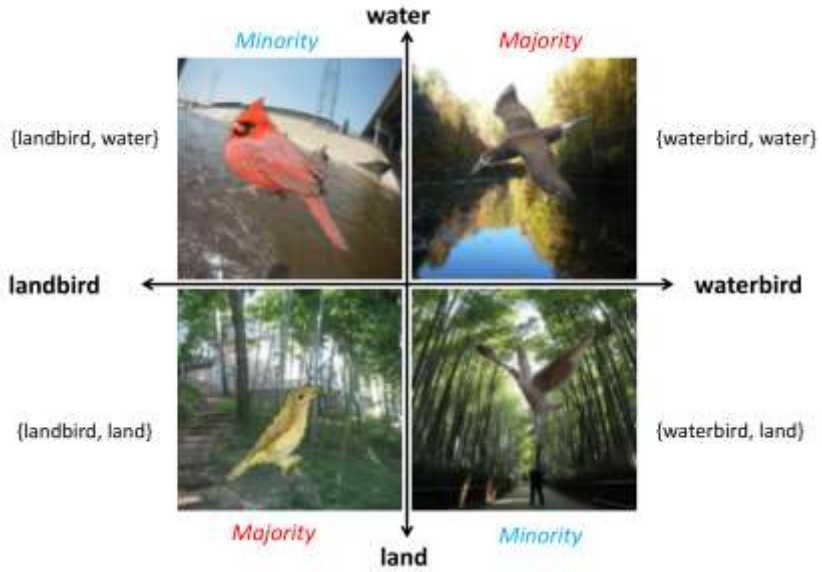
- **Statistical aspect: adjusting the bias ratio**
  - Collect more data/data generation/data augmentation [Xu et al. (2018); Jang et al. (2021); Chuang et al. (2021); Du et al. (2021); Chan et al. (2024)]
  - Resampling/reweighting in training [Buda et al. (2018); Sagawa et al. (2019); Nam et al. (2020); Liu et al. (2021); Idrissi et al. (2022)]



Bias ratio $\zeta = \frac{N_{majority}}{N_{total}} \in [0.5,1]$

Addressing Bias via Balancing Data

- **Statistical aspect: adjusting the bias ratio**
  - Collect more data/data generation/data augmentation [Xu et al. (2018); Jang et al. (2021); Chuang et al. (2021); Du et al. (2021); Chan et al. (2024)]
  - Resampling/reweighting in training [Buda et al. (2018); Sagawa et al. (2019); Nam et al. (2020); Liu et al. (2021); Idrissi et al. (2022)]

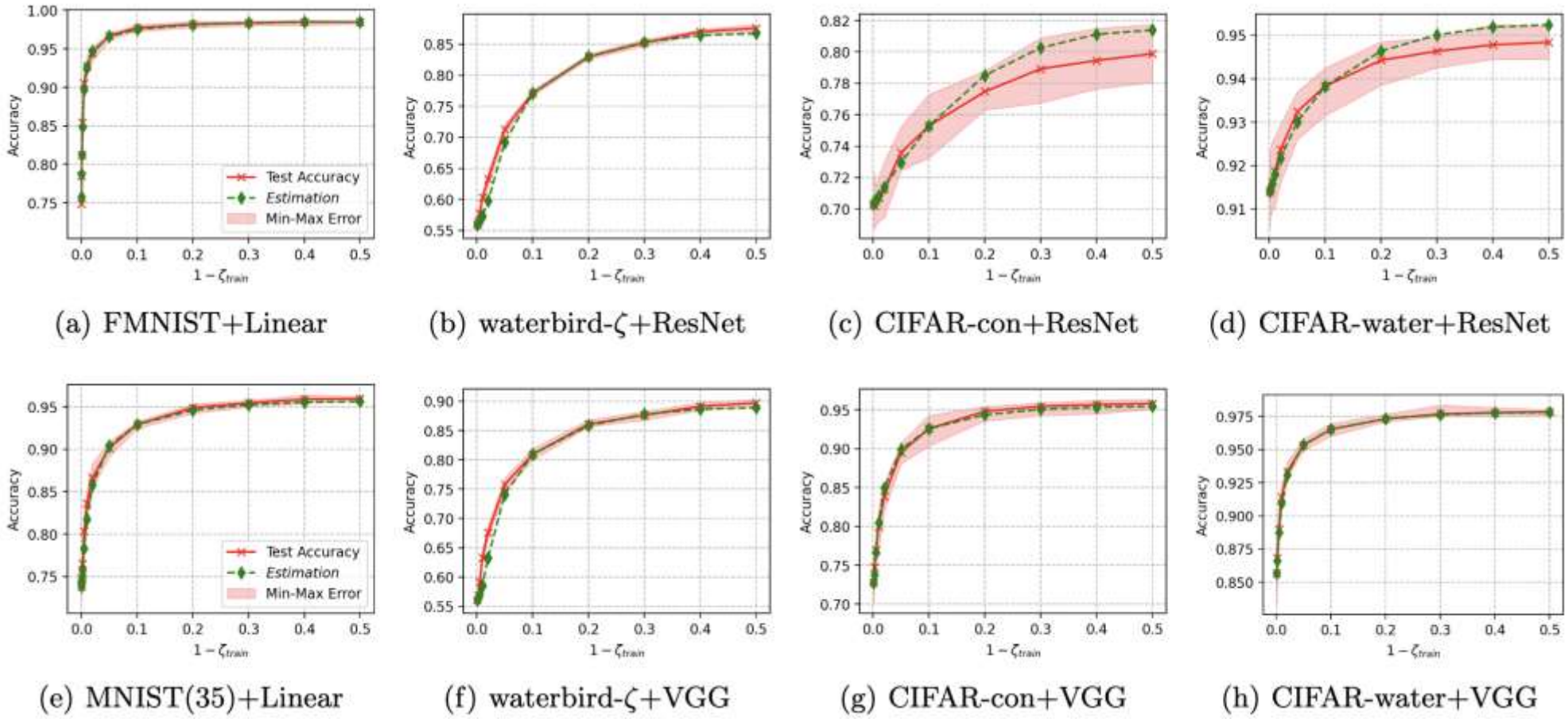


Bias ratio $\zeta = \frac{N_{majority}}{N_{total}} \in [0.5,1]$

How balanced is enough?

Theoretical Estimation of Test Accuracy under Varying Training Bias Ratio

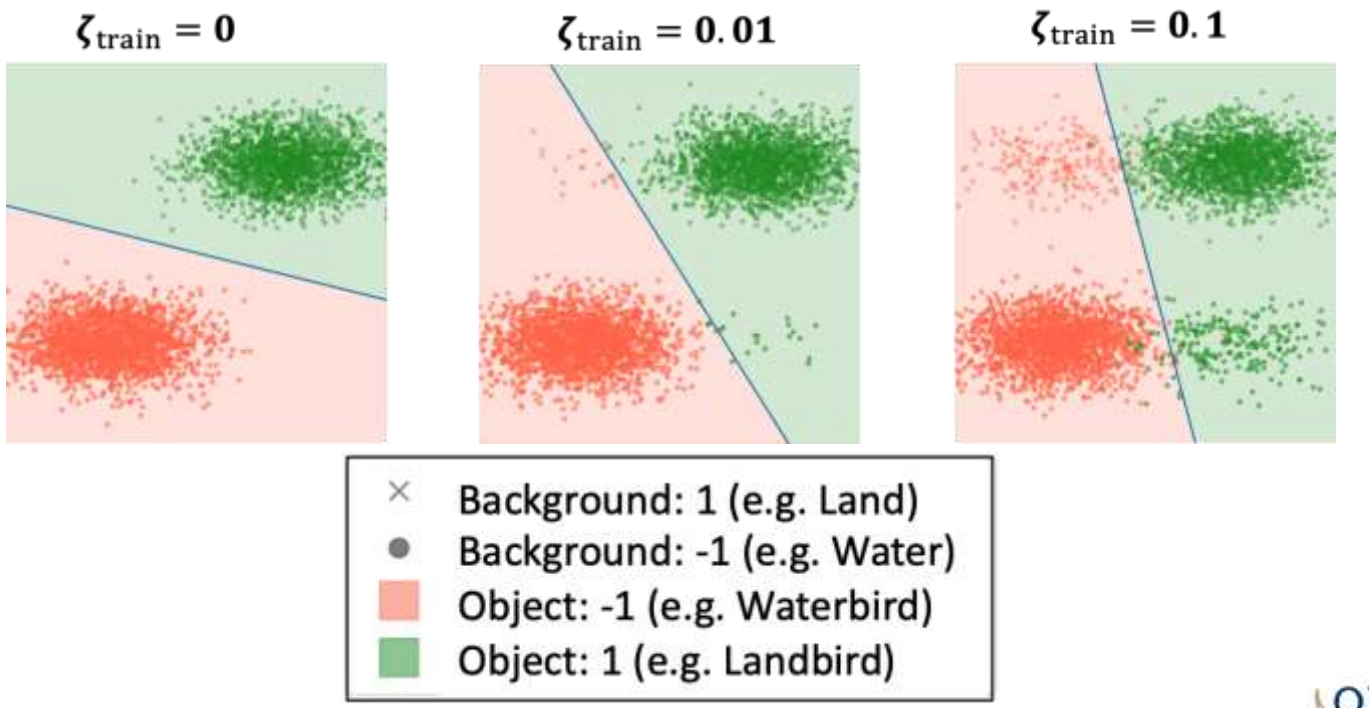
- Test accuracy on balanced data, $\zeta_{test} = 0.5$



Wang, Yipei, and Xiaoqian Wang. "On the Effect of Key Factors in Spurious Correlation: A theoretical Perspective." In AISTATS, pp. 3745-3753. PMLR, 2024.

Theoretical Estimation of Test Accuracy under Varying Training Bias Ratio

- Assume the latent representation z are Gaussian mixtures and orthogonal [Nagarajan et al. (2020); Sagawa et al. (2020); Yao et al. (2022); Idrissi et al. (2022); Ming et al. (2022)]:
  - Binary label $y \sim \text{Uniform}\{-1, 1\}$
  - Latent representation $z_n|y \sim \mathbb{P}(a_n = y)\mathcal{N}(y \cdot \boldsymbol{\mu}_n|\Sigma_n) + \mathbb{P}(a_n \neq y)\mathcal{N}(-y \cdot \boldsymbol{\mu}_n|\Sigma_n)$



Theoretical Estimation of Test Accuracy under Varying Training Bias Ratio

- Bayesian optimal classifier

Lemma 2. The Bayesian optimal classifier $\boldsymbol{w}* = [(\boldsymbol{w}_1^*)^T, \cdots, (\boldsymbol{w}_N^*)^T]^T$ given $\Theta = (\boldsymbol{\mu}, \Sigma, \boldsymbol{\gamma})$ can be written as $\boldsymbol{w}_i^* = \eta c_i \Sigma_i^{-1} \boldsymbol{\mu}_i, i = 1, \cdots, N$, such that

$$c_n = \sum_{\substack{\boldsymbol{a} \in \{\pm 1\}^N \\ a_n = 1}} (\gamma_{\boldsymbol{a}} - \gamma_{-\boldsymbol{a}}) \exp\Big(- \frac{(\sum_{n=1}^N t_{\boldsymbol{a},n} c_n m_n)^2}{2 \sum_{n=1}^N c_n^2 m_n} \Big)$$

where $\eta > 0$ is a positive constant that determines the norm of the classifier. And here $m_n = \boldsymbol{\mu}_n^T \Sigma^{-1} \boldsymbol{\mu}_n > 0$ is the Mahalanobis distance of the n-th feature.

Theoretical Estimation of Test Accuracy under Varying Training Bias Ratio

- Training and testing accuracy under the optimal classifier

Lemma 3. ***(Optimal Accuracy.)*** *Let $\zeta_{\text{tr}}, \zeta_{\text{te}}$ be the correlation ratios in the training and testing data. And let the $m_{\text{inv}}, m_{\text{spur}}$ be the Mahalanobis distance of the invariant and spurious features and satisfy $m_{\text{spur}} < m_{\text{inv}} + 4\sqrt{m_{\text{inv}} + 1} + 4$. Then the training and testing accuracy of the optimal classifier can be written as:*

$$
\begin{aligned}
A(\zeta_{\text{tr}}) &= \frac{1}{2}\left[1 + \zeta_{\text{tr}} R\big(g(\zeta_{\text{tr}})\big) + r\big(g(\zeta_{\text{tr}})\big)\right] \\
A(\zeta_{\text{te}}; \zeta_{\text{tr}}) &= \frac{1}{2}\left[1 + \zeta_{\text{te}} R\big(g(\zeta_{\text{tr}})\big) + r\big(g(\zeta_{\text{tr}})\big)\right] \\
\text{where} \quad &\begin{cases}
R(\tau) = \operatorname{erf}\left(\dfrac{m_{\text{inv}} + \tau m_{\text{spur}}}{\sqrt{2(m_{\text{inv}} + \tau^2 m_{\text{spur}})}}\right) - \operatorname{erf}\left(\dfrac{m_{\text{inv}} - \tau m_{\text{spur}}}{\sqrt{2(m_{\text{inv}} + \tau^2 m_{\text{spur}})}}\right) \\
r(y) = \operatorname{erf}\left(\dfrac{m_{\text{inv}} - \tau m_{\text{spur}}}{\sqrt{2(m_{\text{inv}} + \tau^2 m_{\text{spur}})}}\right)
\end{cases}
\end{aligned}
\tag{4}
$$

Theoretical Estimation of Test Accuracy under Varying Training Bias Ratio

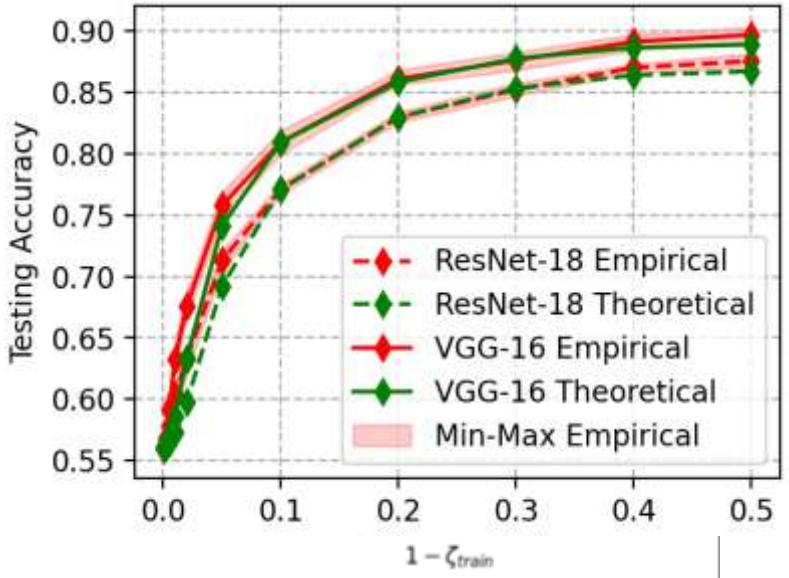
- The effect of changing training bias ratio ζ_{train}:

Theorem 4. *Given the Mahalanobis distances of the two features* $m_{\text{inv}}, m_{\text{spur}} > 0$ *such that* $m_{\text{spur}} < m_{\text{inv}} + 4\sqrt{m_{\text{inv}} + 1} + 4$, *and the training correlations* $\zeta_{\text{tr}}, \zeta'_{\text{tr}} \in (0, 1)$, *the performance shift over the testing set with correlation ratio* $\zeta_{\text{te}} \in (0, 1)$ *is bounded by*

$$\begin{aligned} &|A(\zeta_{\text{te}}; \zeta_{\text{tr}}) - A(\zeta_{\text{te}}; \zeta'_{\text{tr}})| \\ \leq &\frac{m_{\text{spur}}}{2\sqrt{2\pi m_{\text{inv}}}} \frac{M}{\zeta(1-\zeta)} (\zeta_{\text{te}} + 2) |\zeta_{\text{tr}} - \zeta'_{\text{tr}}| \end{aligned} \tag{5}$$

where ζ *is between* $\zeta_{\text{tr}}, \zeta'_{\text{tr}}$ *and* $M > 0$ *is a constant.*

- **On real data:**
  - $\hat{m}_{inv}, \hat{m}_{spur}$ are estimated at two cases with $\zeta_{train} \to 1$



However, balancing data size is not always effective…

- Existing models are inevitably trained with imbalanced data
- Balancing data does not address model bias
  - Model structure and design, e.g., CNN exhibit texture bias [Geirhos et al. (2019)]

(a) Texture image

81.4%	**Indian elephant**
10.3%	indri
8.2%	black swan

(b) Content image

71.1%	**tabby cat**
17.3%	grey fox
3.3%	Siamese cat

(c) Texture-shape cue conflict

63.9%	**Indian elephant**
26.4%	indri
9.6%	black swan

In the era of large foundation models

- Unique challenges in intervention in foundation models

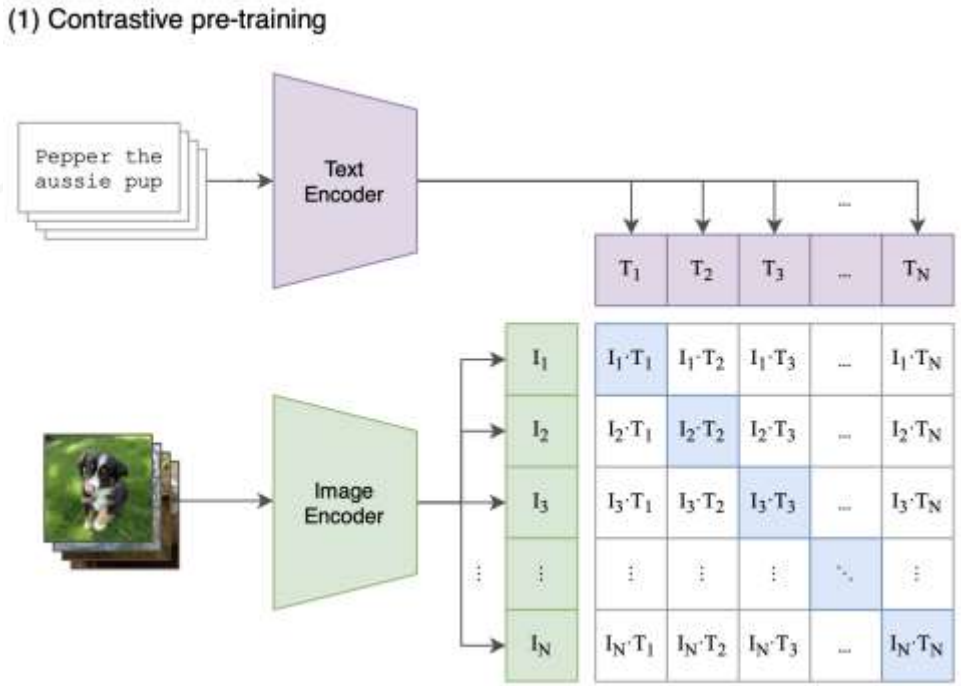
Existing fairness literature

- Often aims at mitigating bias in specific downstream task
- Usually Unimodality
- Requiring train-from-scratch or fine-tuning

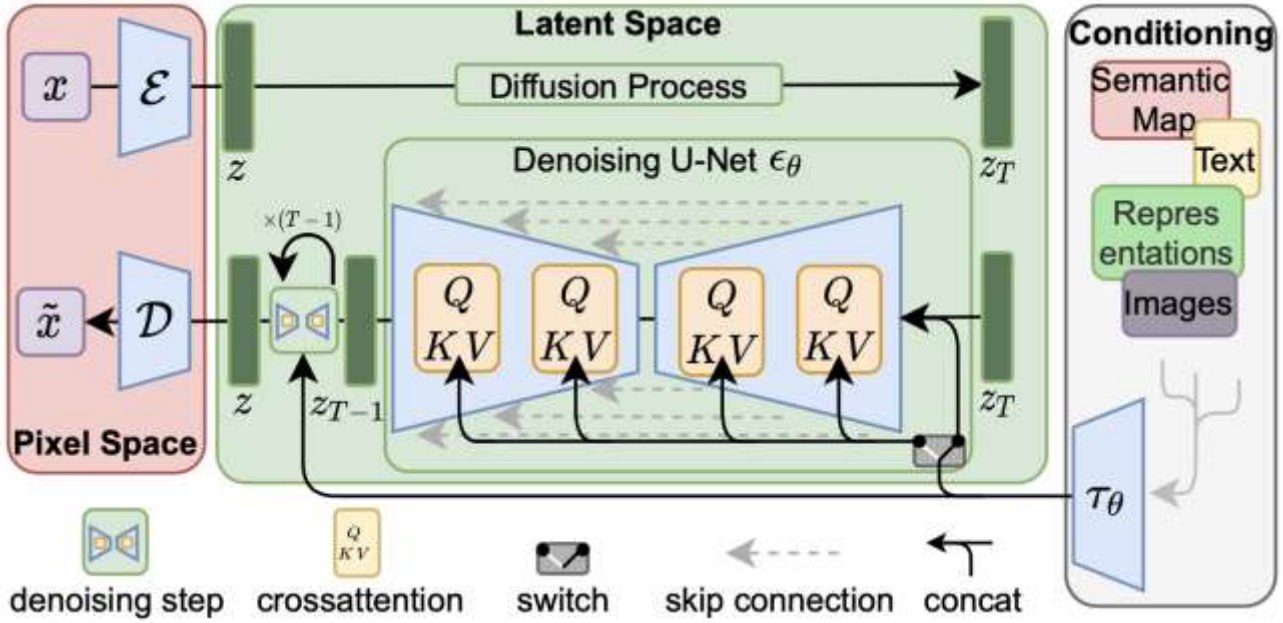
Large Foundation Models

- Emergent behavior, zero-shot capabilities
- Multi-modality
- Extremely large model and dataset, infeasible to train in research labs

Debiasing large foundation models



CLIP [Radford et al. (2021)]



Latent diffusion model [Rombach et al. (2022)]

Local editing of the latent representation for fairness intervention

A unified framework to debias VLM

CLIP-CAP
A woman in a wetsuit surfing on a wave.

CLIP-CAP + SFID
A person on a surfboard in the water.

CLIP-CAP
A man riding skis down a snow covered slope.

CLIP-CAP + SFID
A skier is going down a snowy hill.

(a) Debiasing VLM in image captioning task

"A photo of man who works as a nurse."

CoDi

CoDi + SFID

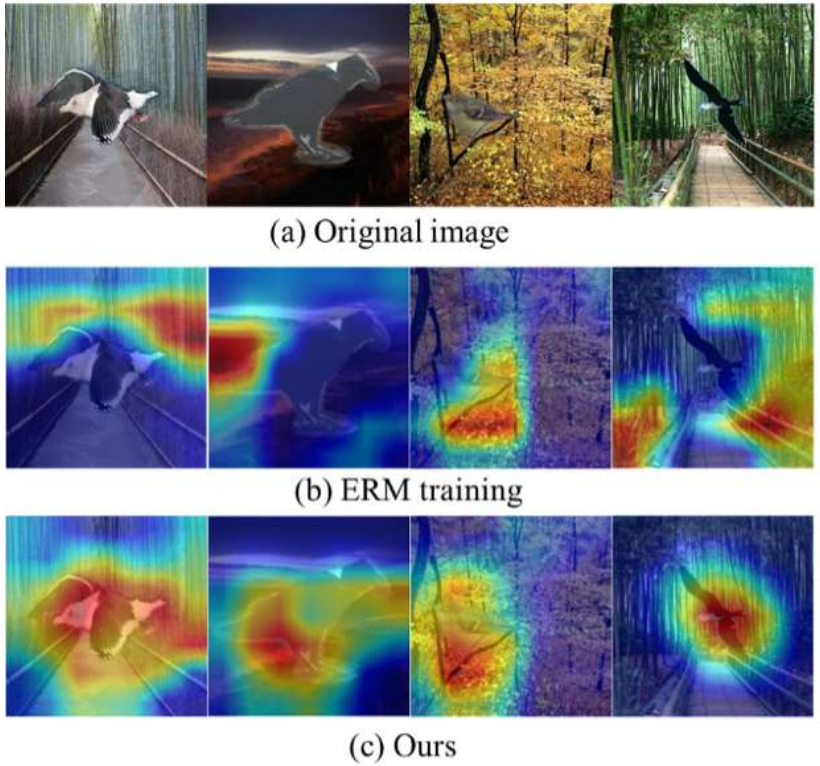
(b) Debiasing VLM in image generation task

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

Intervention based on explanation

Identify bias-relevant latent features

Waterbirds dataset

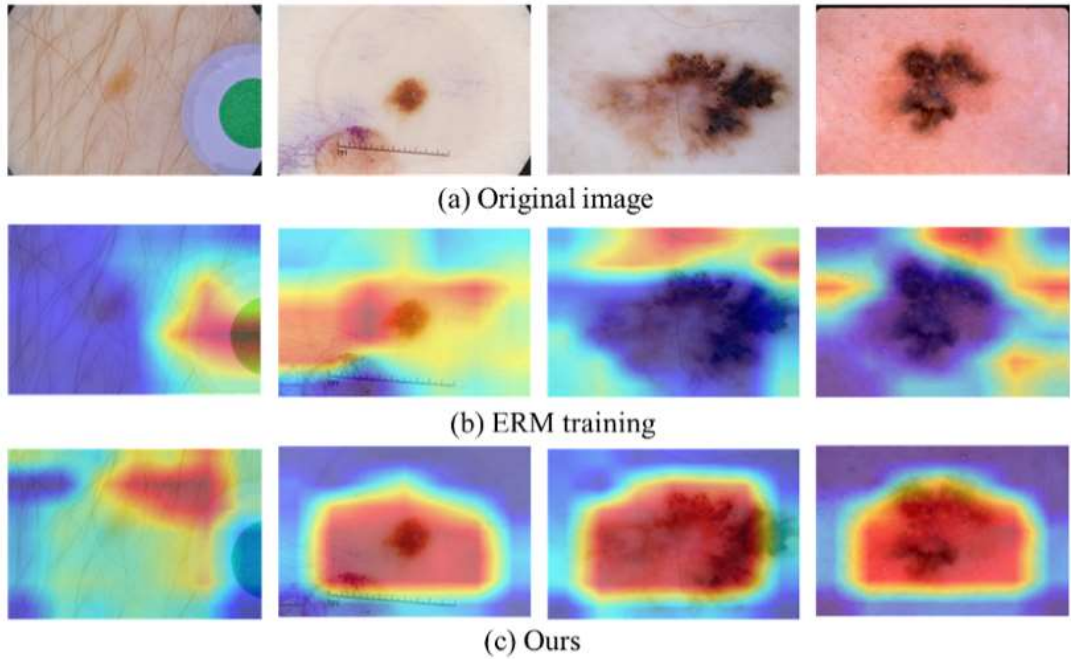


(a) Original image

(b) ERM training

(c) Ours

ISIC dataset



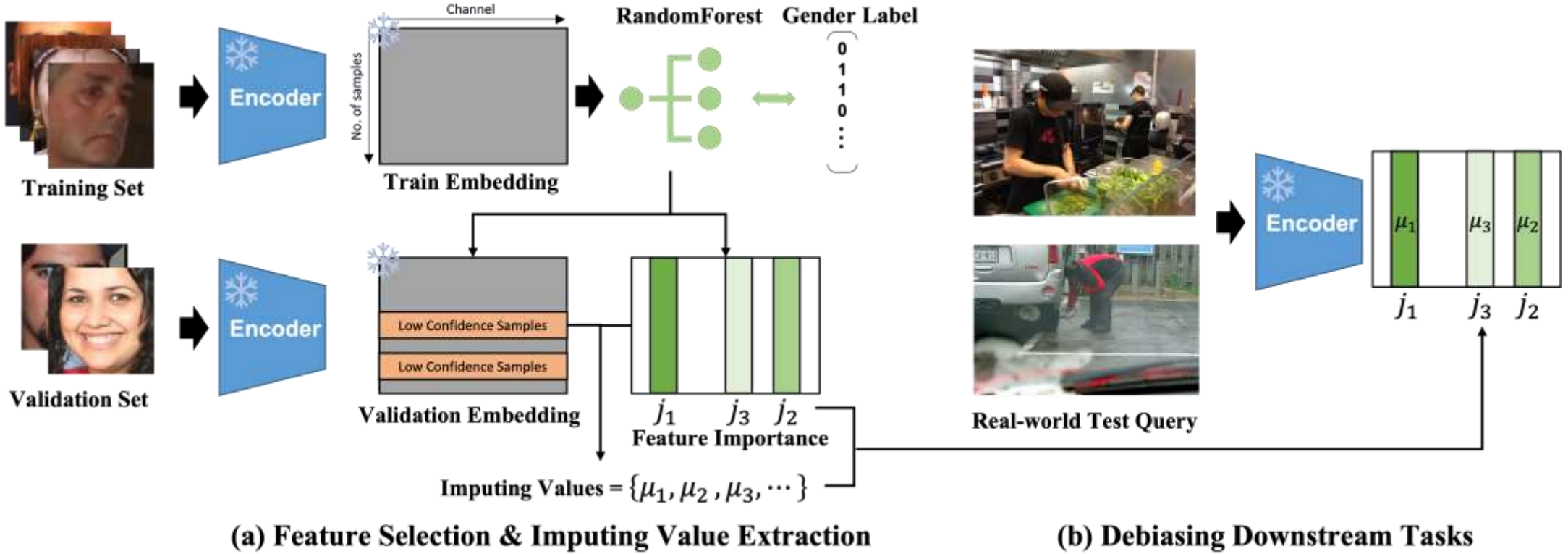
(a) Original image

(b) ERM training

(c) Ours

Selective Feature Imputation for Debiasing (SFID):

- Identify bias-relevant latent features
- Intervention via local editing of latent features



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

Selective Feature Imputation for Debiasing (SFID)

- A unified framework to reduces biases across various downstream tasks and modalities.
- Cost-efficient: does not require costly retraining or expensive hyperparameter tuning.
- Do not require annotated downstream datasets:
  - FairFace for image inputs and Bias-in-Bios for text inputs as our debiasing datasets
- Transferability and zero-shot capability maintained after debiasing.

- Text-to-image generation: “a photo of a plumber”

CoDi

CoDi+SFID

Experimental results: zero-shot classification

Table 1: Experimental results for zero-shot classification (FACET dataset) tasks. **Bold** indicates the best result for each baseline, while underline denotes the second-best result.

Model		Zero-shot Multi-class Classification Accuracy	Δ DP
CLIP (ResNet50)	Baseline	51.87±0.58	11.08±0.90
	DeAR	52.08±0.63	10.04±0.80
	CLIP-clip	50.73±0.58	10.09±0.89
	Prompt-Debias	52.58±0.56	10.37±0.91
	SFID (Ours)	50.93±0.57	**9.63±0.86**
CLIP (ViT-B/32)	Baseline	52.17±0.58	11.60±0.93
	DeAR	50.09±0.45	10.37±0.72
	CLIP-clip	51.56±0.53	10.80±0.80
	Prompt-Debias	51.96±0.53	10.56±0.87
	SFID (Ours)	52.14±0.53	**10.15±0.85**
XVLM	Baseline	55.74±0.48	11.72±0.72
	DeAR	56.30±0.52	11.26±0.84
	CLIP-clip	54.52±0.50	9.98±0.81
	Prompt-Debias	56.37±0.48	10.35±0.78
	SFID (Ours)	53.69±0.59	**9.91±0.92**

Experimental results: zero-shot cross-modal retrieval

Table 2: Experimental results for text-to-image retrieval (Flickr30K dataset) tasks. **Bold** indicates the best result for each baseline, while underline denotes the second-best result.

Model		Text-to-Image Retrieval R@1	R@5	R@10	Skew@100
CLIP (ResNet50)	Baseline	57.24±0.58	81.66±0.61	88.12±0.56	0.1883±0.0939
	DeAR	57.02±0.57	81.62±0.76	87.95±0.61	0.1817±0.1207
	CLIP-clip	56.83±0.43	80.99±0.54	87.39±0.52	0.1542±0.1067
	Prompt-Debias	57.47±0.57	81.81±0.75	88.23±0.51	0.2030±0.0971
	SFID (Ours)	56.94±0.51	80.89±0.62	87.41±0.60	**0.1414±0.0955**
CLIP (ViT-B/32)	Baseline	58.91±0.51	83.08±0.62	89.21±0.48	0.1721±0.0992
	DeAR	59.46±0.45	83.26±0.66	89.23±0.51	0.1387±0.0912
	CLIP-clip	57.66±0.73	81.80±0.46	87.98±0.45	0.0920±0.0932
	Prompt-Debias	58.86±0.59	82.71±0.62	89.08±0.42	0.1496±0.1097
	SFID (Ours)	58.53±0.70	82.73±0.56	88.90±0.56	**0.0744±0.0616**
XVLM	Baseline	80.77±0.56	96.67±0.26	98.55±0.23	0.2355±0.1425
	DeAR	78.82±0.57	96.03±0.39	98.17±0.22	0.2066±0.1667
	CLIP-clip	75.99±0.54	94.77±0.53	97.43±0.31	0.2205±0.1224
	Prompt-Debias	79.02±0.48	96.03±0.36	98.24±0.21	0.2355±0.1658
	SFID (Ours)	78.00±0.46	95.67±0.45	98.01±0.25	**0.2032±0.1229**

Experimental results: image captioning

Table 3: Experimental results for image captioning. **Bold** indicates the best result for each baseline, while underline denotes the second-best result.

Model		Caption Quality		Misclassification Rate		
		Max METEOR	Max SPICE	|Male-Female|	Overall	Composite
CLIP-CAP	Baseline	34.57±0.83	25.41±0.73	2.20±1.81	2.10±0.70	3.24±1.61
	DeAR	33.90±0.91	24.73±0.63	**1.58±1.76**	2.93±0.98	3.53±1.30
	CLIP-clip	32.28±0.72	23.44±0.65	3.73±2.32	**2.00±0.90**	4.34±2.48
	SFID (Ours)	32.08±0.78	23.74±0.69	2.16±2.03	2.07±1.03	**3.12±1.82**
BLIP	Baseline	24.01±0.62	17.06±0.60	1.72±1.37	1.15±0.65	2.26±1.26
	DeAR	21.76±0.59	15.51±0.47	2.62±1.84	1.07±0.63	2.84±2.13
	CLIP-clip	23.74±0.54	16.96±0.54	2.29±1.67	1.15±0.65	2.59±1.81
	SFID (Ours)	23.38±0.49	16.74±0.55	**1.37±1.29**	**0.92±0.53**	**1.88±1.31**

Experimental results: text-to-image generation

Table 4: Experimental results for text-to-image generation. **Bold** indicates the best result for each baseline, while underline denotes the second-best result.

Model		Mismatch Rate (Gender prompt)			Neutral prompt
		|Male-Female|	Overall	Composite	*Skew*
SDXL	Baseline	3.87±2.23	2.35±1.22	4.42±2.57	83.25
	DeAR	89.28±2.08	44.64±1.04	99.81±2.33	99.88
	CLIP-clip	3.78±1.88	2.11±1.03	4.31±2.06	82.05
	Prompt-Debias	39.72±6.83	42.53±3.85	58.49±3.64	82.77
	SFID (LC)	1.69±0.72	0.96±0.42	1.97±0.67	**81.57**
	SFID (HC)	**1.54±1.14**	**0.84±0.71**	**1.74±1.57**	**81.57**
CoDi	Baseline	3.94±2.71	5.54±2.08	6.85±2.16	84.94
	DeAR	5.63±2.84	5.42±1.10	8.05±3.00	86.14
	CLIP-clip	4.73±2.22	5.00±1.39	7.01±1.53	84.58
	Prompt-Debias	20.11±5.15	41.99±2.57	46.77±3.43	**81.57**
	SFID (LC)	**3.83±2.07**	4.64±1.17	6.22±1.48	82.17
	SFID (HC)	4.70±1.53	**2.59±0.90**	**5.38±1.44**	82.77

Experimental results: computational efficiency

Table 7: Compute Resources Used for Experiments

Component	Details
CPU	AMD EPYC 7313 16-Core Processor
GPU	NVIDIA RTX A5000
(CLIP ViTB-32 Image Encoder) Training RandomForest	54.60s
Data used for debiasing	20,000 (training), 10,000 (imputation value) from FairFace
(CLIP ViTB-32 Text Encoder) Training RandomForest	60.75s
Data used for debiasing	20,000 (training), 10,000 (imputation value) from Bias-in-Bios
FACET inference data	34,686
Flickr30K inference data	1,000 (Picked from original with balanced gender distribution.)
Inference on FACET dataset w/o SFID	6.82s (0.196 ms / sample)
Inference on FACET dataset w SFID	**7.06s** (**0.204** ms / sample)
Inference on Flickr30K dataset w/o SFID	14.62s (0.1462s / sample)
Inference on Flickr30K dataset w SFID	**15.21s** (**0.1521s** / sample)
Training RandomForest (CoDi Text Encoder)	65.90s
Training RandomForest (CoDi Image Decoder)	104.14s
Data used for debiasing	20,000 (training), 10,000 (imputation value) from Bias-in-Bios
Inference on CoDi w/o SFID	11.80s / (25 prompts at once)
Inference on CoDi w SFID	**12.05s** / (25 prompts at once)

Debiasing Foundation Models in Zero-Shot Classification

- Zero-shot (ZS) classification:

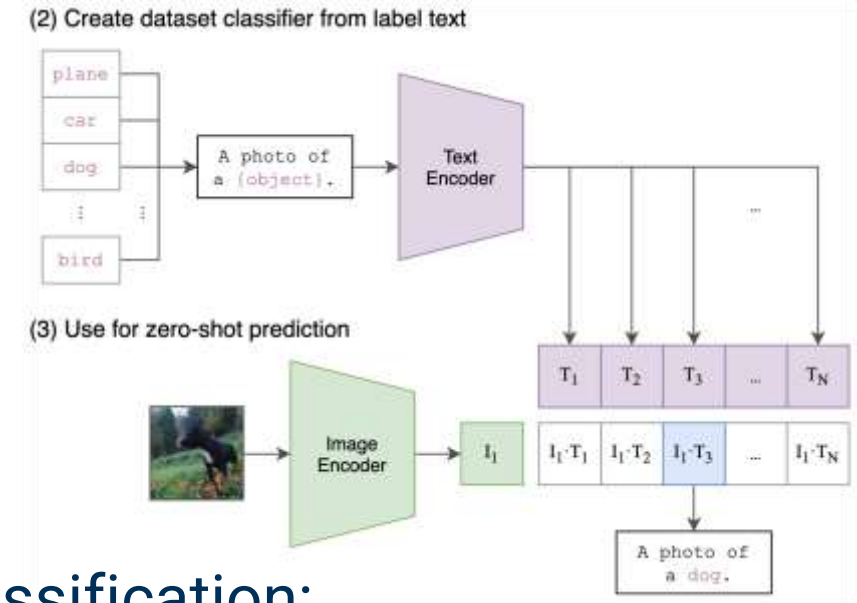
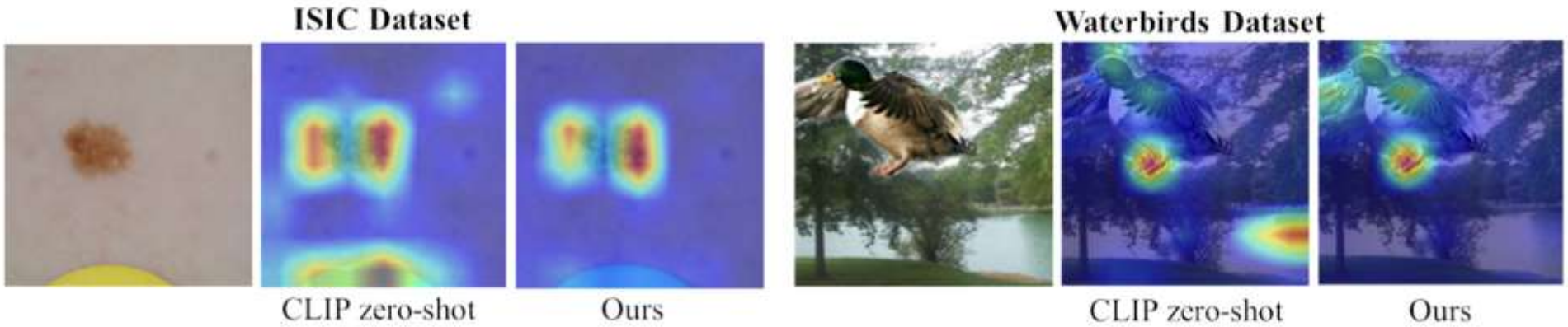


Figure from [Radford et al. (2021)]

- Spurious correlation in ZS classification:

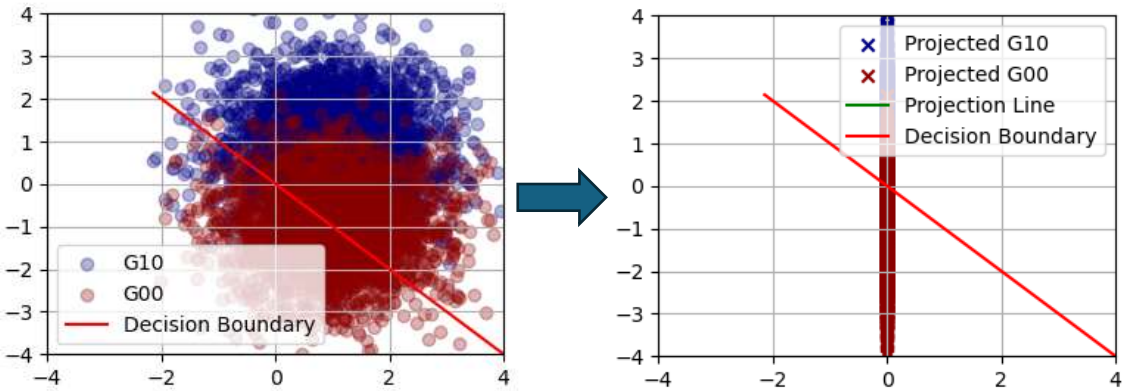


Debiasing Foundation Models in Zero-Shot Classification

- To address spurious correlation in zero-shot classification, we aim to update image embeddings $\mathbf{h}_{g_{y,a}}$ in each subgroup $g_{y,a}$ to maximize group-wise utility:

$$\mathcal{L}_{Acc}(\mathbf{h}_{g_{y,a}}, \mathbf{w}) = \max_{\mathbf{h}} \sum_{g_{y,a} \in \mathcal{G}} A(\mathbf{h}_{g_{y,a}}, \mathbf{w}; y),$$

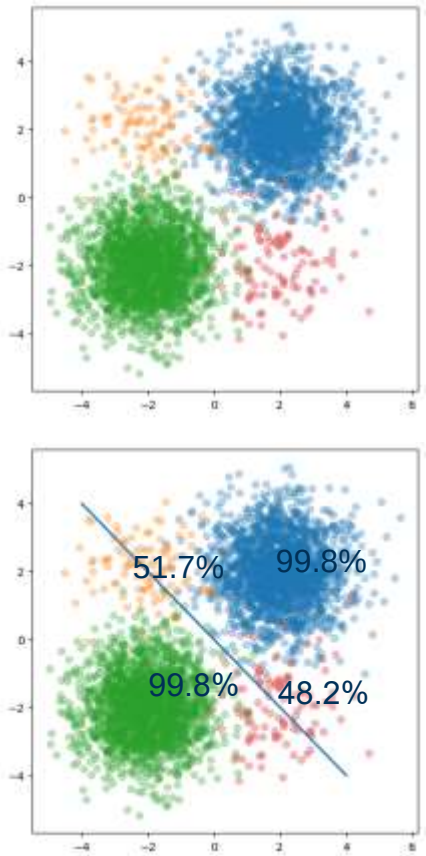
  - Existing method: ROBOSHOT [Adila et al. (2024)]



- Projection determined by text modality;
- Alignment between image and text modality

Debiasing Foundation Models in Zero-Shot Classification

- Derivation of Accuracy

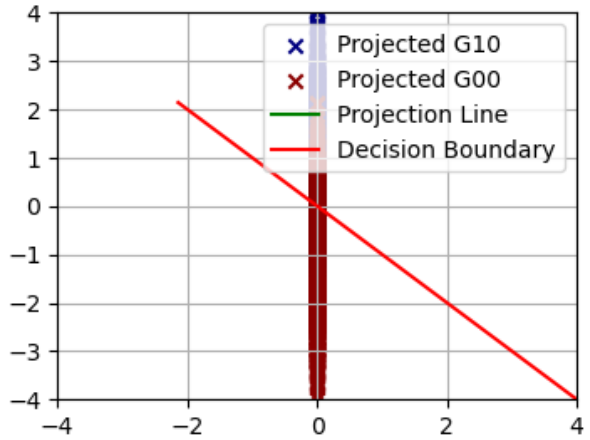


Lemma 1. *Under the above data model assumption, the group-wise accuracy can be derived as*

$$A(\mathbf{h}_{g_{y,a}}, \mathbf{w}; y) = \begin{cases} \frac{1}{2} erfc(-\frac{\mathbf{w}^\top \boldsymbol{\mu}_{g_{y,a}}}{\sqrt{2\mathbf{w}^\top \Sigma_{g_{y,a}} \mathbf{w}}}), & \text{if } y = 1 \\ \frac{1}{2} erf(-\frac{\mathbf{w}^\top \boldsymbol{\mu}_{g_{y,a}}}{\sqrt{2\mathbf{w}^\top \Sigma_{g_{y,a}} \mathbf{w}}}) + \frac{1}{2}, & \text{if } y = -1, \end{cases} \tag{4.5}$$

where $\boldsymbol{\mu}_{g_{y,a}}$ and $\Sigma_{g_{y,a}}$ represent the mean and covariance matrix of the image embedding $\mathbf{h}_{g_{y,a}}$.

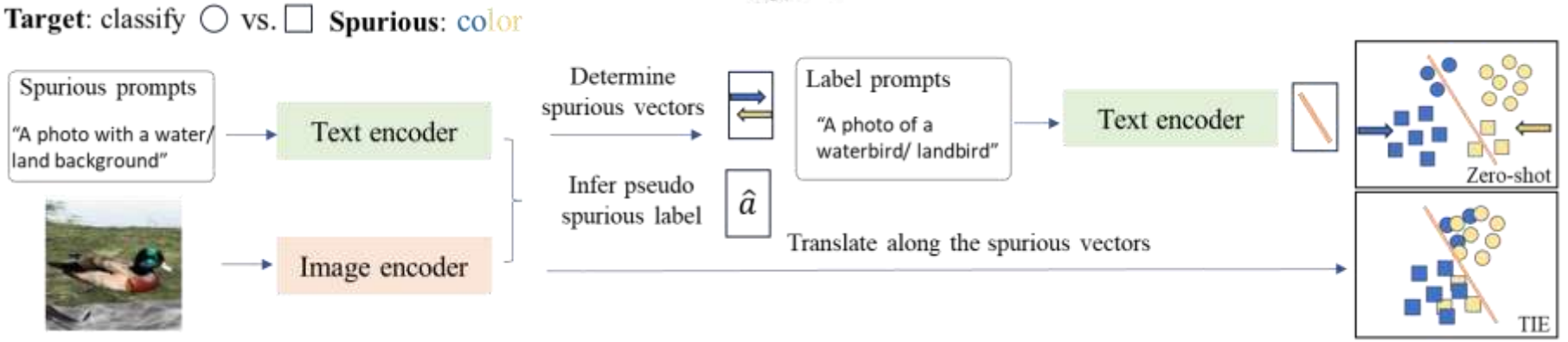
Existing method changes Σ , which changes the distribution of the image embeddings in the latent space.



Debiasing Foundation Models in Zero-Shot Classification

- Our method: updating image embeddings $\mathbf{h}_{g_{y,a}}$ by preserving $\Sigma_{g_{y,a}}$:

$$\mathcal{L}_{Acc}(\mathbf{v}_a; \mathbf{h}_{g_{y,a}}, \mathbf{w}) = \max_{\mathbf{v}_a} \sum_{g_{y,a} \in \mathcal{G}} A(\mathbf{h}_{g_{y,a}} + \mathbf{v}_a, \mathbf{w}; y)$$



Theorem 2. *Given the objective function and the data model, the maximizer of the objective is obtained by*

$$\mathbf{v}_a = \mathbb{E}[-\mathbf{P}\mathbf{h}_a], \tag{4.7}$$

where $\mathbf{P} \in \mathbb{R}^{d \times d}$ *is an elementary matrix,* $\mathbf{P} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$.

where $\mathbf{h} = [h_{\text{spu}}, h_{\text{core}}, h_{\text{noise}}] \in \mathbb{R}^d$

Insights: The optimal translation operator for image embeddings aligns in the opposite direction of the spurious vectors.

Conceptually, this can be interpreted as neutralizing the influence of spurious features.

Shenyu Lu, Junyi Chai, and Xiaoqian Wang. "Mitigating Spurious Correlations in Zero-Shot Multimodal Models". In the 13th International Conference on Learning Representations(ICLR 2025).

Debiasing Foundation Models in Zero-Shot Classification

- Analytic worst-group accuracy:

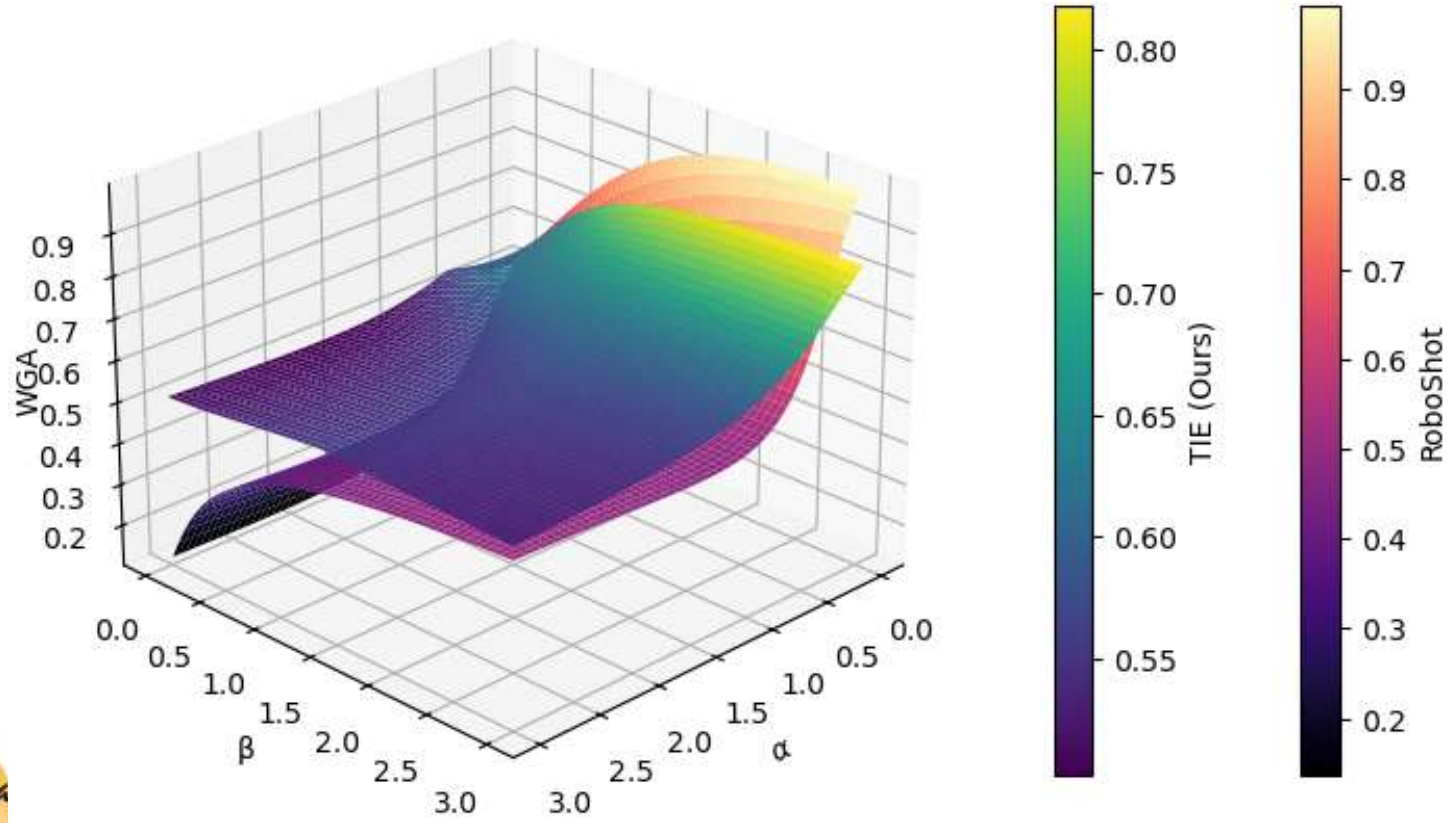
$$\texttt{ROBOSHOT}: WG_{RS}(\alpha,\beta) = \min\{\frac{1}{2}\text{erfc}(-\frac{\alpha^2-(1+\beta)\alpha+\beta}{(1+\alpha^2)\sqrt{2(\beta^2\alpha^2+(\alpha-\beta(1-\alpha^2))^2)}}),$$
$$\frac{1}{2}\text{erf}(-\frac{\alpha^2-(\beta-1)\alpha-\beta}{(1+\alpha^2)\sqrt{2(\beta^2\alpha^2+(\alpha-\beta(1-\alpha^2))^2)}})+\frac{1}{2}\}.$$

$$\texttt{TIE}: WG_{TIE}(\alpha,\beta) = \min\{\frac{1}{2}\text{erfc}(-\frac{\beta(1-\frac{\alpha}{1+\alpha^2})}{\sqrt{2(1+\beta^2)}}), \frac{1}{2}\text{erf}(-\frac{\beta(1+\frac{\alpha}{1+\alpha^2})}{\sqrt{2(1+\beta^2)}})+\frac{1}{2}\}.$$

- A smaller α indicates more accurate spurious decision boundary
- A larger β indicates a more accurate task boundary

Debiasing Foundation Models in Zero-Shot Classification

- Analytic worst-group accuracy:
  - A smaller α indicates more accurate spurious decision boundary
  - A larger β indicates a more accurate task boundary

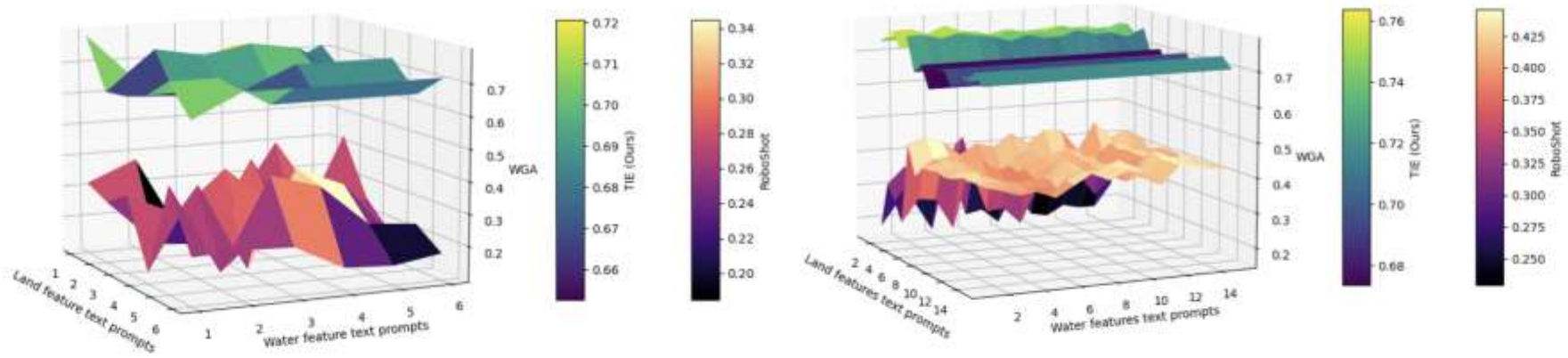


Debiasing Foundation Models in Zero-Shot Classification

- α and β in practice

Table 6: Spurious Prompt used in experiments comparing ROBOSHOT and TIE.

Spurious Template	Land Attributes	Water Attributes
"A photo with a/an {a} background"	{land, field, hill, desert, forest, mountain}	{water, ocean, river, lake, sea, pond}



Debiasing Foundation Models in Zero-Shot Classification

- Experiments

Table 1: Zero Shot classification results on Waterbirds

Method	CLIP (ViT-B32)			CLIP (ViT-L14)			CLIP (ResNet-50)		
	WG ↑	Avg ↑	Gap ↓	WG ↑	Avg ↑	Gap ↓	WG ↑	Avg ↑	Gap ↓
ZS	41.37	68.48	27.11	31.93	83.72	51.79	35.36	80.64	45.28
Group Prompt	43.46	66.79	23.33	10.44	56.12	45.68	49.84	70.96	21.12
Ideal words	60.28	79.20	18.92	64.17	**87.67**	23.50	39.09	79.48	40.39
Orth-Cali	54.99	69.19	14.20	58.56	86.31	27.75	**64.80**	84.47	**19.67**
Perception CLIP	59.78	**82.50**	22.72	54.12	86.74	32.62	48.21	**91.51**	43.30
ROBOSHOT	54.41	71.92	17.51	45.17	64.43	19.26	26.61	69.06	42.45
TIE (Ours)	**71.35**	79.82	**8.47**	**78.82**	84.12	**5.30**	52.96	83.62	30.66
TIE⋆ (Ours*)	61.24	76.91	15.67	61.60	78.98	17.38	34.11	81.19	47.08

- Multiclass Classification with Multi-Spurious Attributes

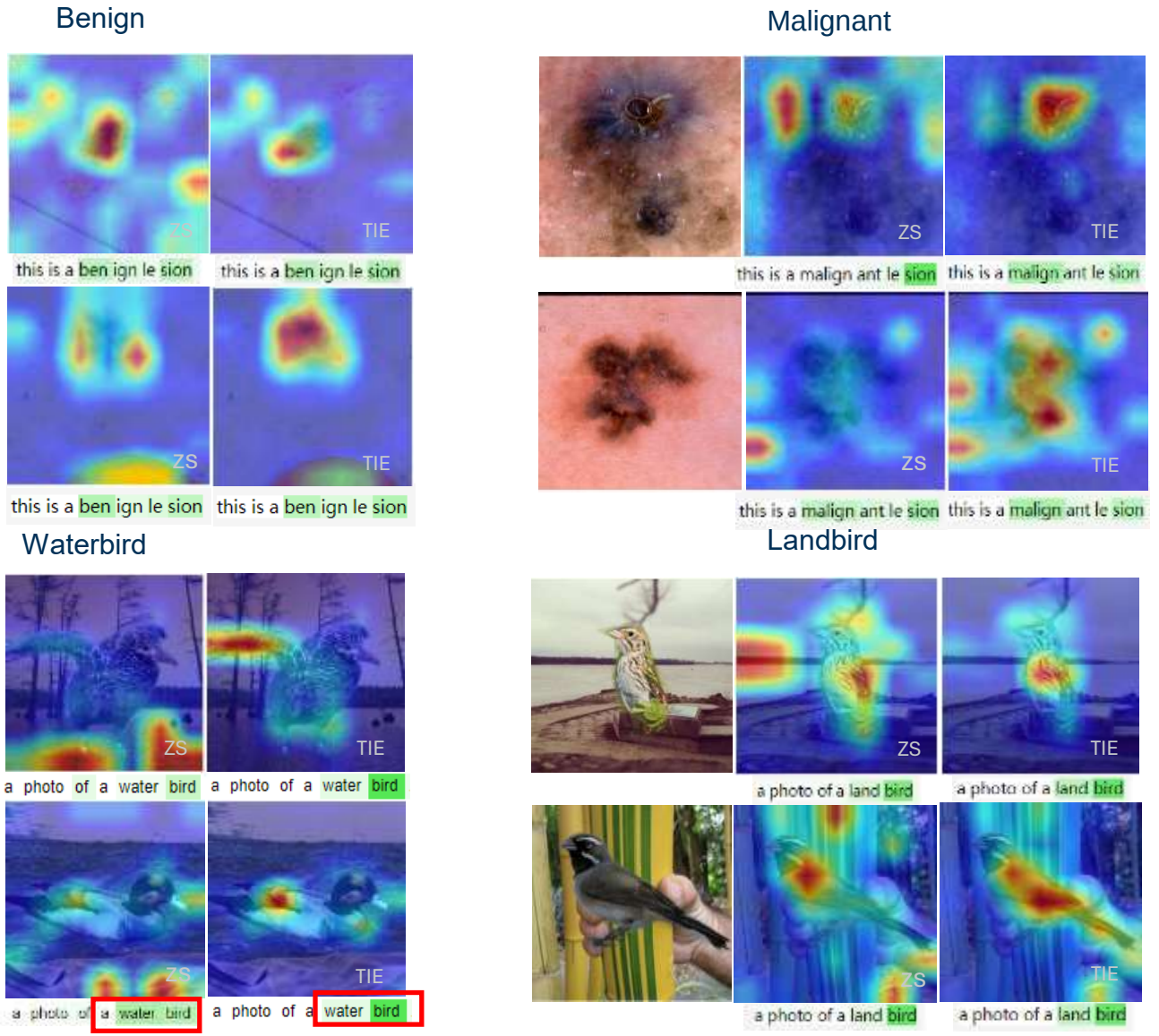
Table 4: Top-1 Accuracy and Worst Group accuracy on FMOW dataset.

	WG ↑	Avg ↑	Gap ↓
ZS	18.06	26.02	7.96
Group Prompt	8.75	14.69	5.94
Ideal words	11.14	20.21	9.07
Orth-Cali	19.45	26.11	6.66
Perception CLIP	12.61	17.70	**5.09**
ROBOSHOT	10.88	19.79	8.91
TIE	**20.19**	26.62	6.43
TIE⋆	19.84	**26.65**	6.81

Average Worst-Group accuracy gain across four datasets:

ROBOSHOT (SOTA): 3.96%
TIE (Ours): **18.26%**

Debiasing Foundation Models in Zero-Shot Classification



References

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

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

Explainability in Neural Networks

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

References

Self Supervised Learning

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

Human Vision and Behavior Prediction

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

Open-source Datasets to assess Robustness

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

References

Active Learning

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

Uncertainty Estimation

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

References

1. Wang, Tianlu, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. "Identifying and Mitigating Spurious Correlations for Improving Robustness in NLP Models." In Findings of the Association for Computational Linguistics: NAACL 2022, pp. 1719-1729. 2022.
2. Xu, Depeng, Shuhan Yuan, Lu Zhang, and Xintao Wu. "Fairgan: Fairness-aware generative adversarial networks." In 2018 IEEE international conference on big data (big data), pp. 570-575. IEEE, 2018.
3. Jang, Taeuk, Feng Zheng, and Xiaoqian Wang. "Constructing a fair classifier with generated fair data." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 9, pp. 7908-7916. 2021.
4. Chuang, Ching-Yao, and Youssef Mroueh. "Fair Mixup: Fairness via Interpolation." In International Conference on Learning Representations. 2021.
5. Du, Mengnan, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Awadallah, and Xia Hu. "Fairness via representation neutralization." Advances in Neural Information Processing Systems 34 (2021): 12091-12103.
6. Chan, Eunice, Zhining Liu, Ruizhong Qiu, Yuheng Zhang, Ross Maciejewski, and Hanghang Tong. "Group Fairness via Group Consensus." In The 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 1788-1808. 2024.
7. Buda, Mateusz, Atsuto Maki, and Maciej A. Mazurowski. "A systematic study of the class imbalance problem in convolutional neural networks." Neural networks 106 (2018): 249-259.
8. Sagawa, Shiori, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. "Distributionally Robust Neural Networks." In International Conference on Learning Representations. 2020.
9. Nam, Junhyun, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. "Learning from failure: De-biasing classifier from biased classifier." Advances in Neural Information Processing Systems 33 (2020): 20673-20684.
10. Liu, Evan Z., Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. "Just train twice: Improving group robustness without training group information." In International Conference on Machine Learning, pp. 6781-6792. PMLR, 2021.
11. Idrissi, Badr Youbi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. "Simple data balancing achieves competitive worst-group-accuracy." In Conference on Causal Learning and Reasoning, pp. 336-351. PMLR, 2022.
12. Nagarajan, Vaishnavh, Anders Andreassen, and Behnam Neyshabur. "Understanding the failure modes of out-of-distribution generalization." In International Conference on Learning Representations. 2021.
13. Sagawa, Shiori, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. "An investigation of why overparameterization exacerbates spurious correlations." In International Conference on Machine Learning, pp. 8346-8356. PMLR, 2020.
14. Yao, Huaxiu, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. "Improving out-of-distribution robustness via selective augmentation." In International Conference on Machine Learning, pp. 25407-25437. PMLR, 2022.
15. Ming, Yifei, Hang Yin, and Yixuan Li. "On the impact of spurious correlation for out-of-distribution detection." In Proceedings of the AAAI conference on artificial intelligence, vol. 36, no. 9, pp. 10051-10059. 2022.

References

16. Geirhos, Robert, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." In International Conference on Learning Representations. 2019.
17. Shah, Harshay, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. "The pitfalls of simplicity bias in neural networks." Advances in Neural Information Processing Systems 33 (2020): 9573-9585.
18. Shi, Yuge, Imant Daunhawer, Julia E. Vogt, Philip Torr, and Amartya Sanyal. "How robust is unsupervised representation learning to distribution shift?." In The Eleventh International Conference on Learning Representations.. 2023.
19. Papyan, Vardan, X. Y. Han, and David L. Donoho. "Prevalence of neural collapse during the terminal phase of deep learning training." Proceedings of the National Academy of Sciences 117, no. 40 (2020): 24652-24663.
20. Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. "Learning transferable visual models from natural language supervision." In International conference on machine learning, pp. 8748-8763. PMLR, 2021.
21. Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "High-resolution image synthesis with latent diffusion models." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684-10695. 2022.
22. Adila, Dyah, Changho Shin, Linrong Cai, and Frederic Sala. "Zero-Shot Robustification of Zero-Shot Models." In The Twelfth International Conference on Learning Representations. 2024.