Formalizing Robustness in Neural Networks: Explainability, Uncertainty, and Intervenability

Ghassan AlRegib, PhD
Professor

Mohit Prabhushankar, PhD
Postdoctoral Fellow

Association for the Advancement of Artificial Intelligence

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

Feb 21, 2024 – Vancouver, BC, Canada

Tutorial Materials

Accessible Online

SCAN ME

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

{alregib, mohit.p}@gatech.edu

AAAI 2024 Tutorial

Presented by: ***Ghassan AlRegib, and Mohit Prabhushankar***

Georgia Institute of Technology

www.ghassanalregib.info

Duration: Half Day (3 hours, 30 mins)

Title: Formalizing Robustness in Neural Networks: Explainability, Uncertainty, and Intervenability

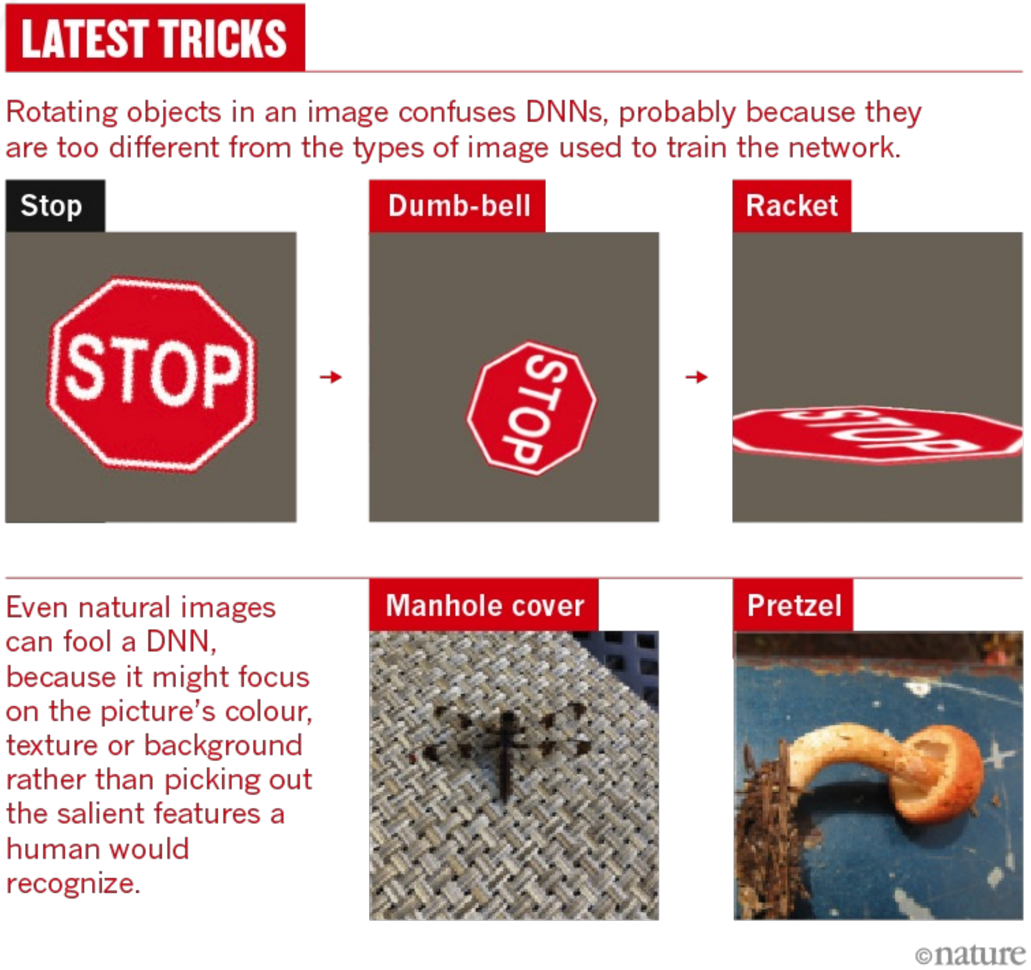
OLIVES @GeorgiaTech

Georgia Tech

Expectation vs Reality of Deep Learning

Deep Learning

Expectation vs Reality



OLIVES @GeorgiaTech

Georgia Tech

Deep Learning

Expectation vs Reality

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

- Engineers, probably

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

Novel data sources:

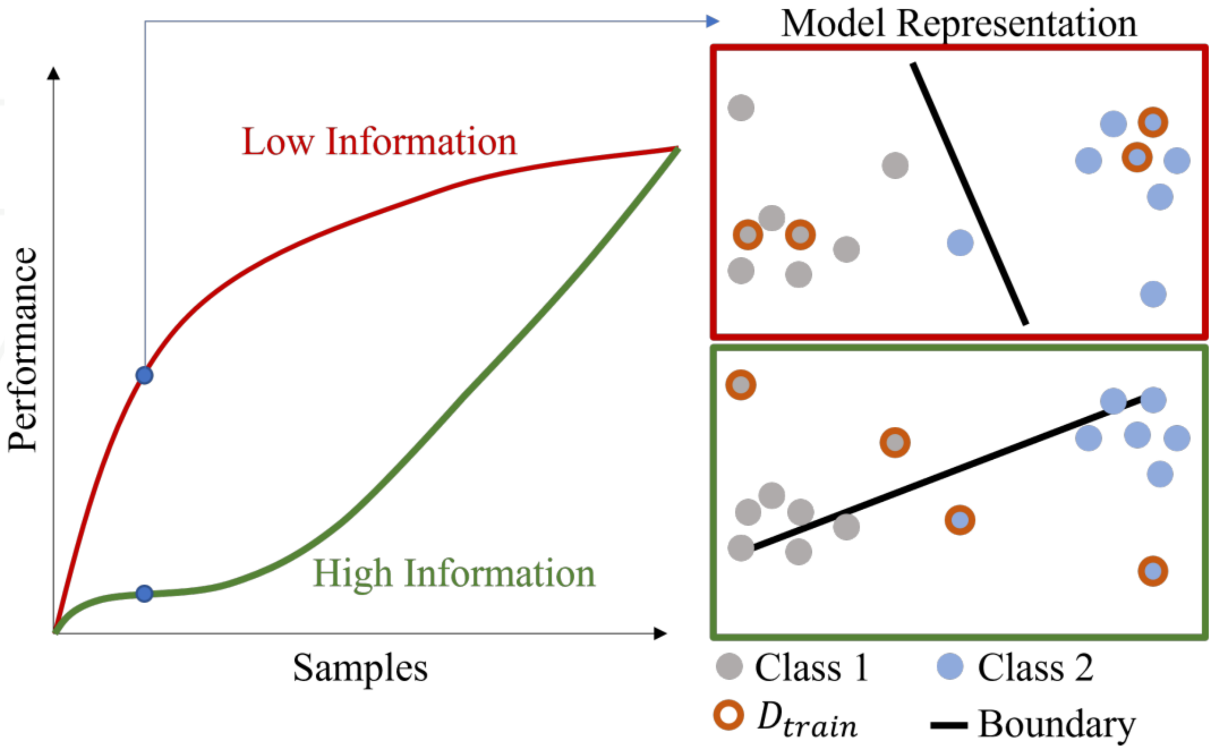
- Test distributions
- Anomalous data
- Out-Of-Distribution data
- Adversarial data
- Corrupted data
- Noisy data
- New classes
- …

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

Deep Learning at Training

Overcoming Challenges at Training: Part 1

The most novel/aberrant samples should 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

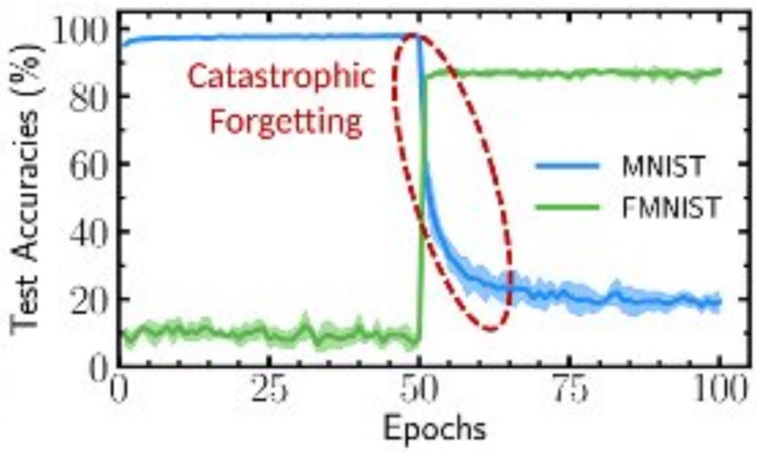
Novel samples = Most Informative

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

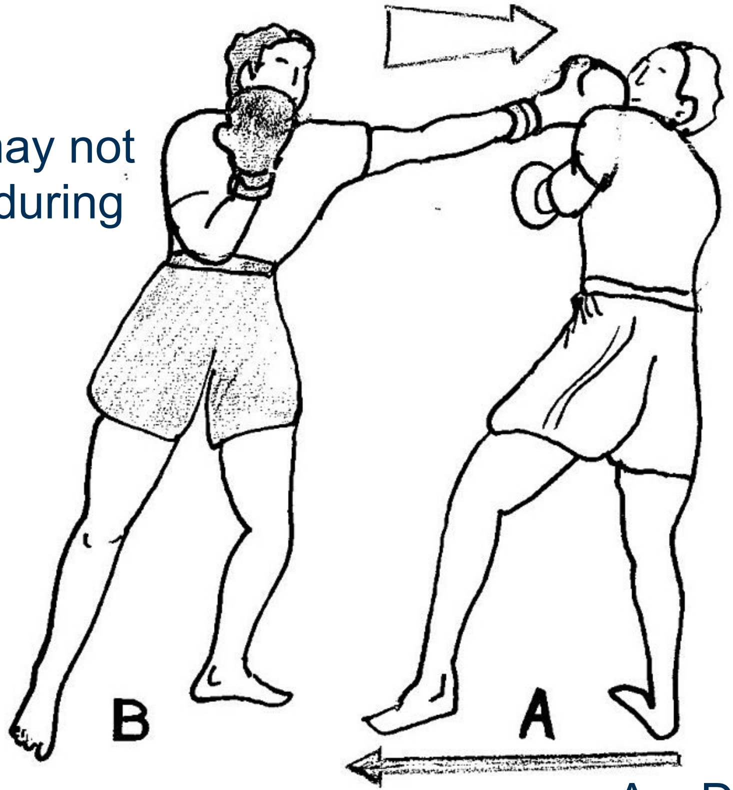


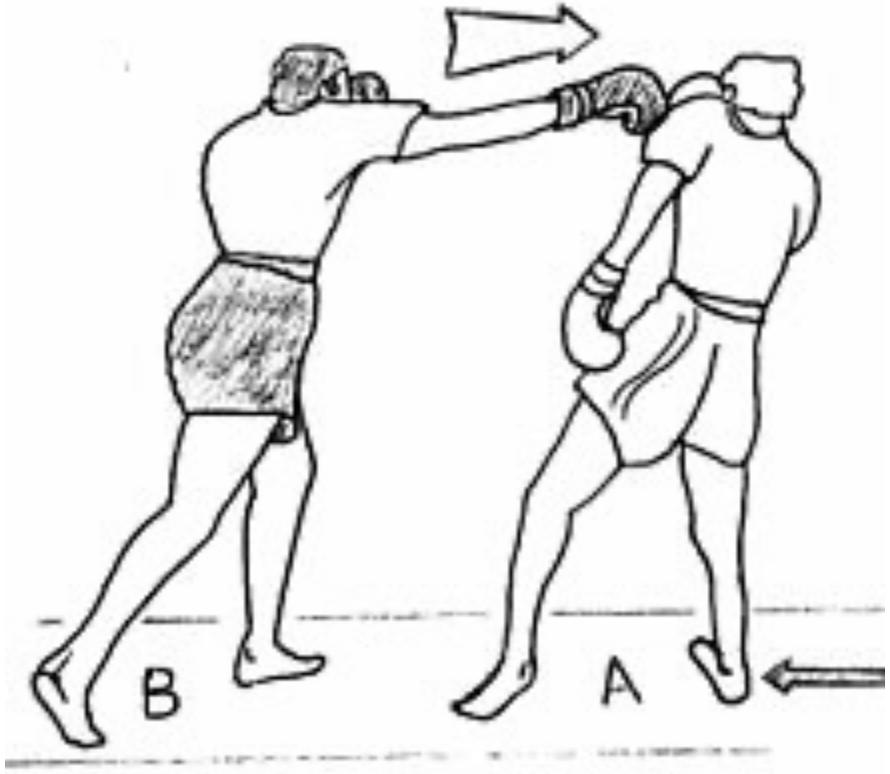
- The model performs well on the new scenarios, while forgetting the old scenarios
- A number of techniques exist to overcome this trend
- However, they affect the overall performance in large-scale settings
- It is not always clear **if and when** to incorporate novel scenarios in training

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

Novel data packs a 1-2 punch!

Novel data may not be available during training





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

A = Deep Neural Networks
B = Novel data

Deep Learning at Inference

Overcoming Challenges at Inference

We must handle novel data at Inference!!

Novel data sources:

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

Model Train

At Inference

Objective

Objective of the Tutorial

To discuss methodologies that promote robustness in neural networks at inference

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

Robust Neural Networks

Part I: Inference in Neural Networks

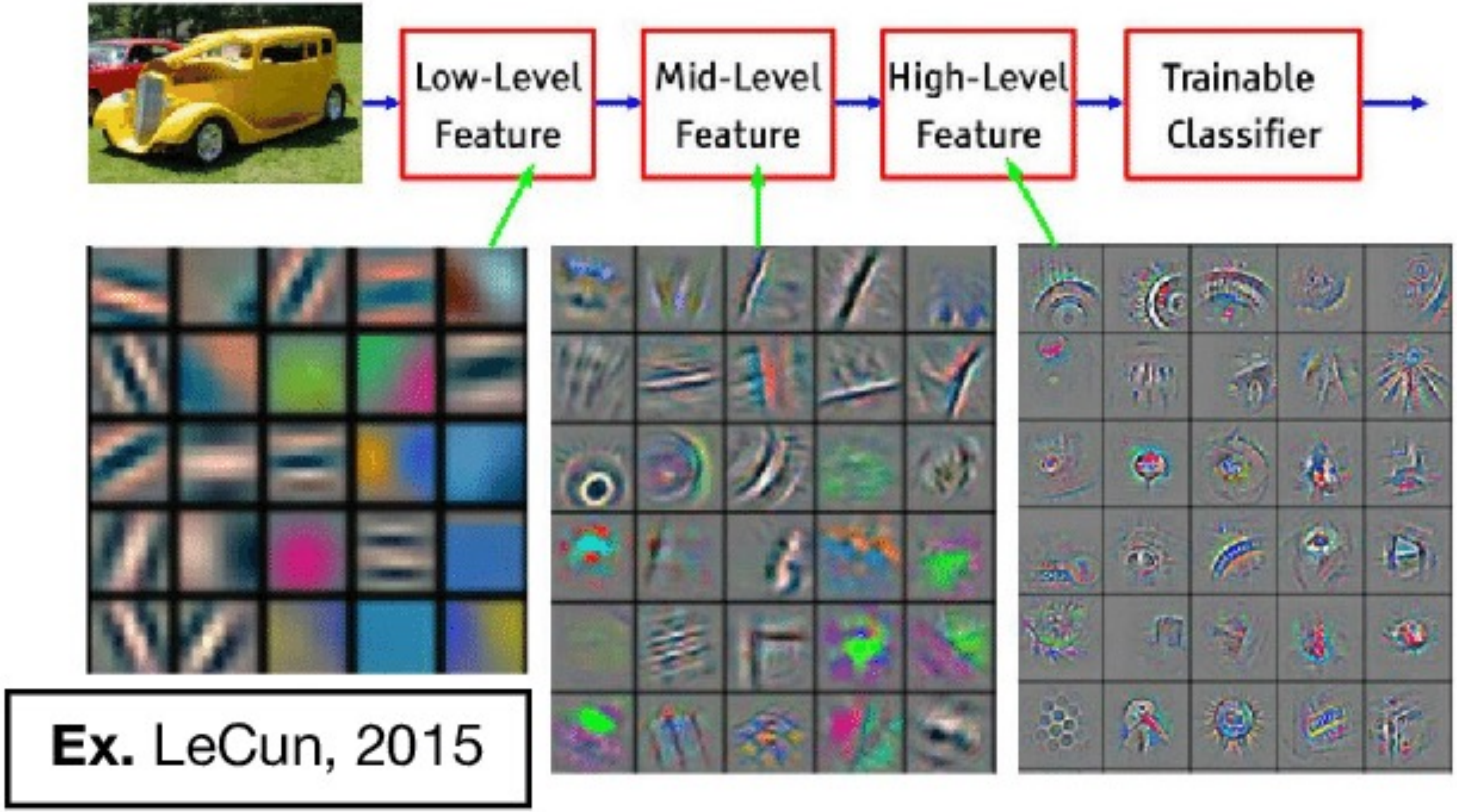
Objective

Objective of the Tutorial

To discuss methodologies that promote robustness in neural networks at inference

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

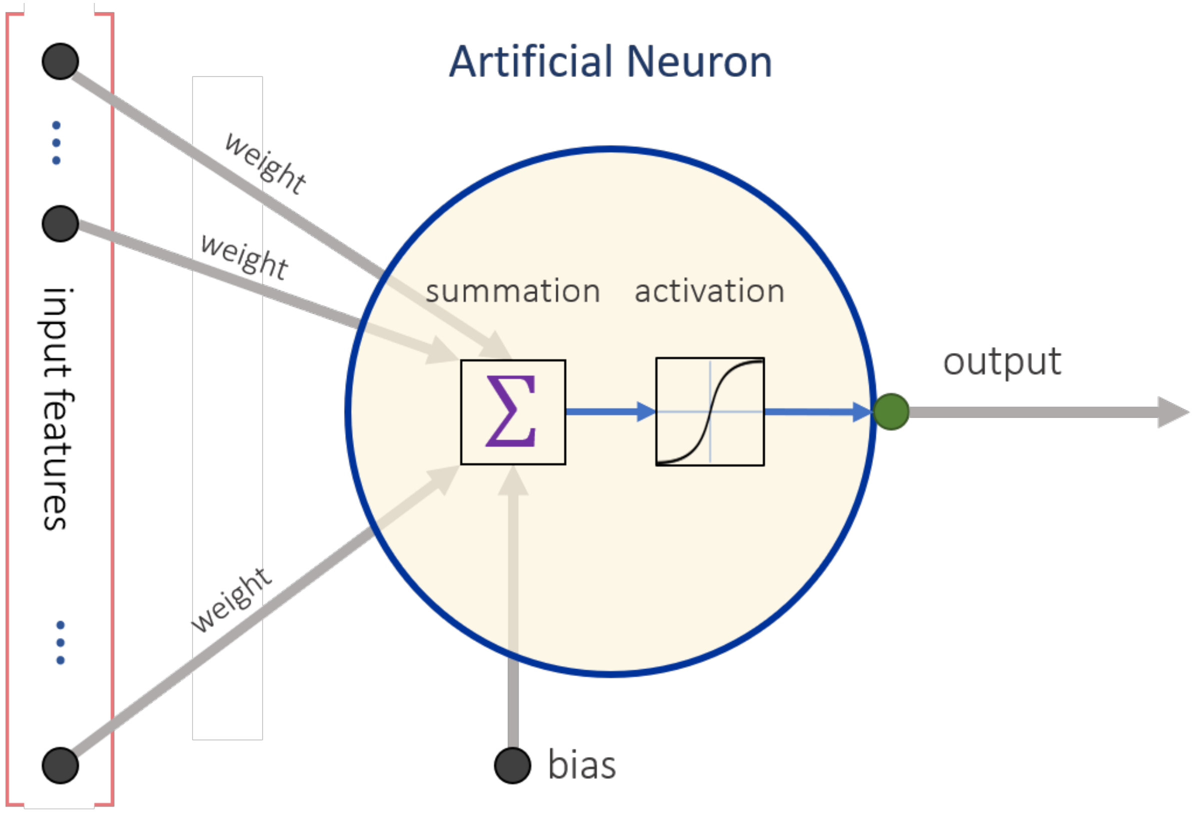
Deep Learning
Overview



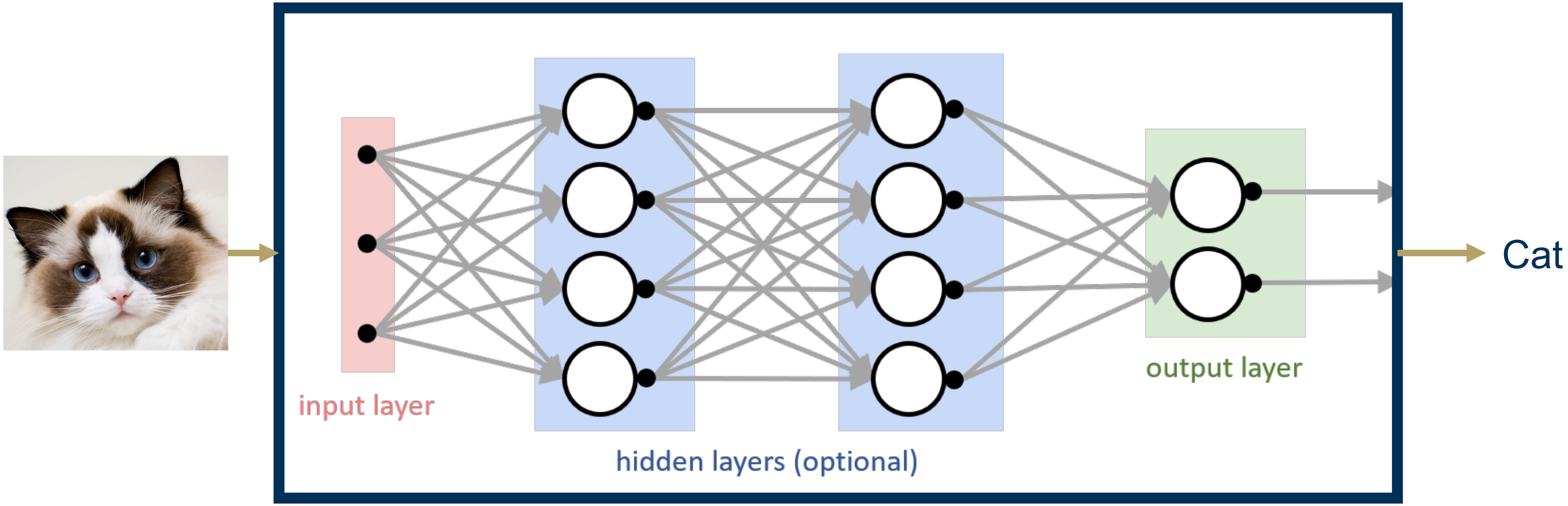
The underlying computation unit is the Neuron

Artificial neurons consist of:

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



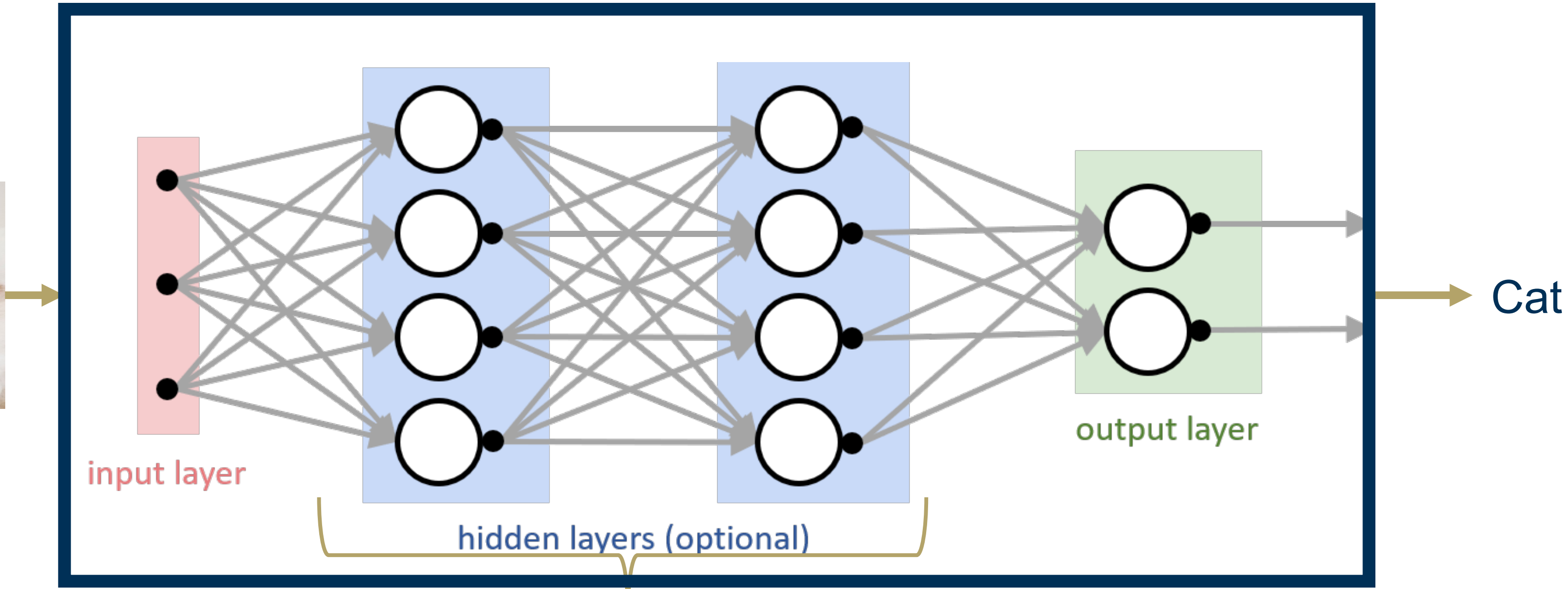
Neurons are stacked and densely connected to construct ANNs

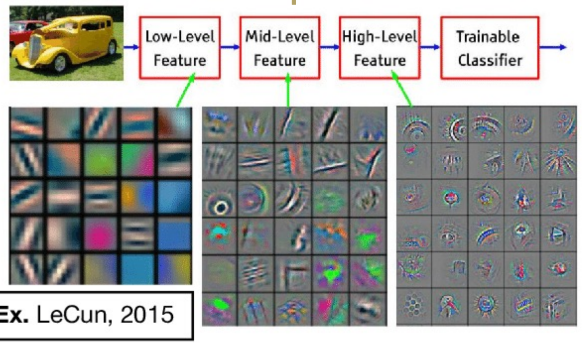


Typically, a neuron is part of a network organized in **layers**:

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

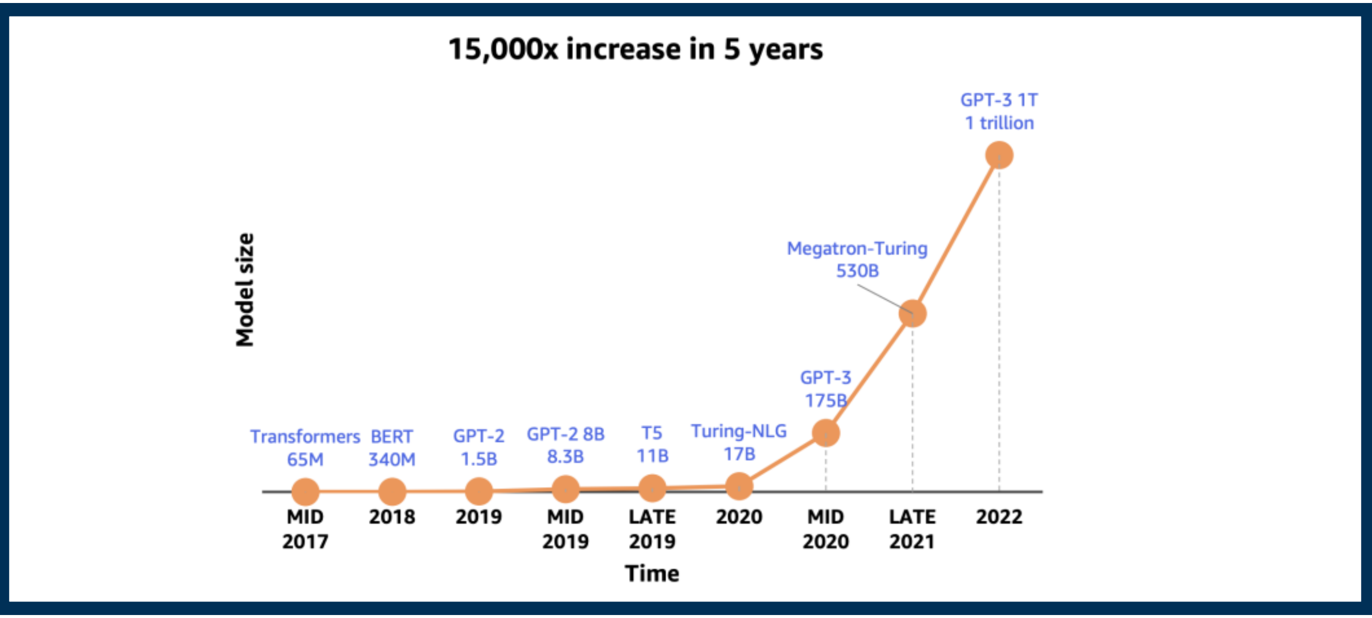
Stationary property of images allow for a small number of convolution kernels





Ex. LeCun, 2015

Transformers, Large Language Models and Foundation Models



Cat

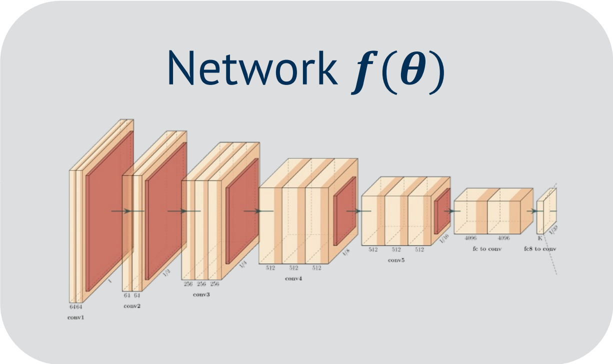
Primary reasons for advancements:
1. Expanded interests from the research community
2. Computational resources availability
3. **Big data availability**

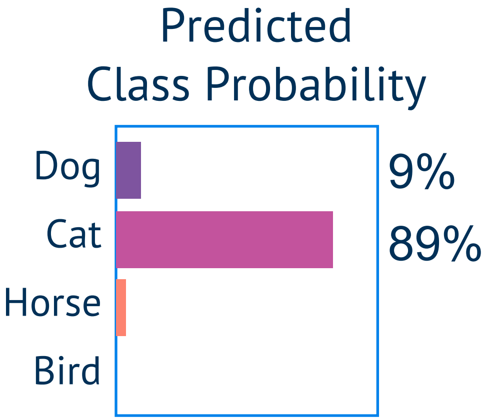
Deep Learning at Inference

Classification

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

x





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

$\hat{y} = f(x)$

$y = argmax_i \, \hat{y}$

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

$\hat{y}$ = Logits

y = Predicted Class

$p(\hat{y})$ = Probabilities

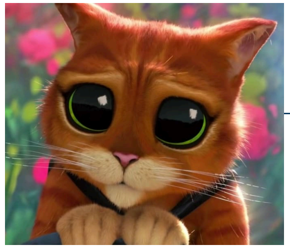
$f(\cdot)$ = Trained Network

χ = Training data

OLIVES @GeorgiaTech

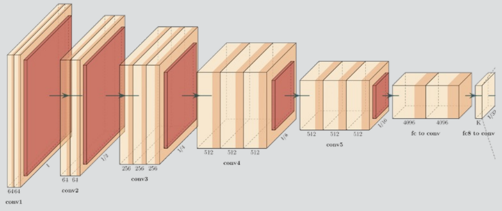
Georgia Tech

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

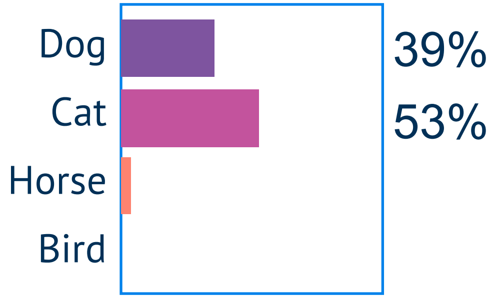


x'

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



Predicted Class Probability



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

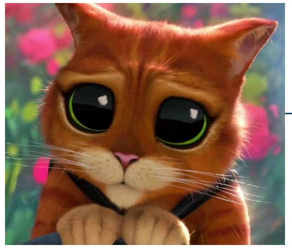
$$\hat{y} = f(x' + \epsilon)$$
$$y = argmax_i\, \hat{y}$$
$$p(\hat{y}) = T(f(x' + \epsilon))$$

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

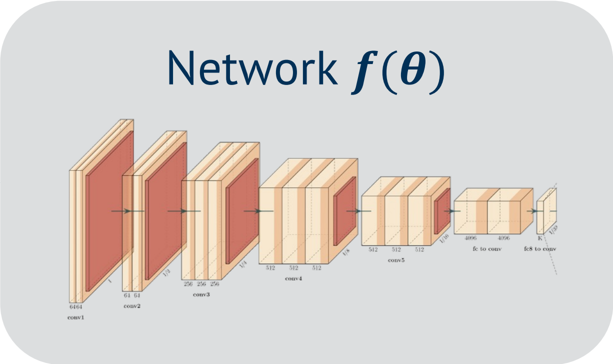
Deep Learning at Inference

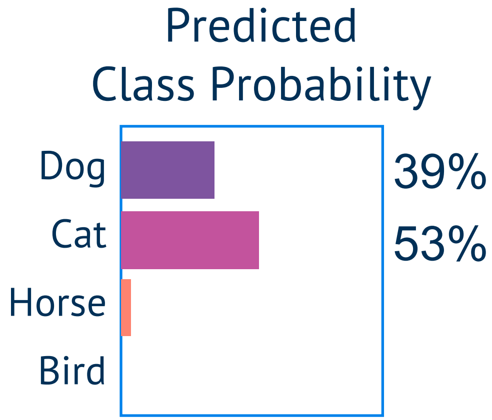
Robust Classification in Deep Networks

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



x'





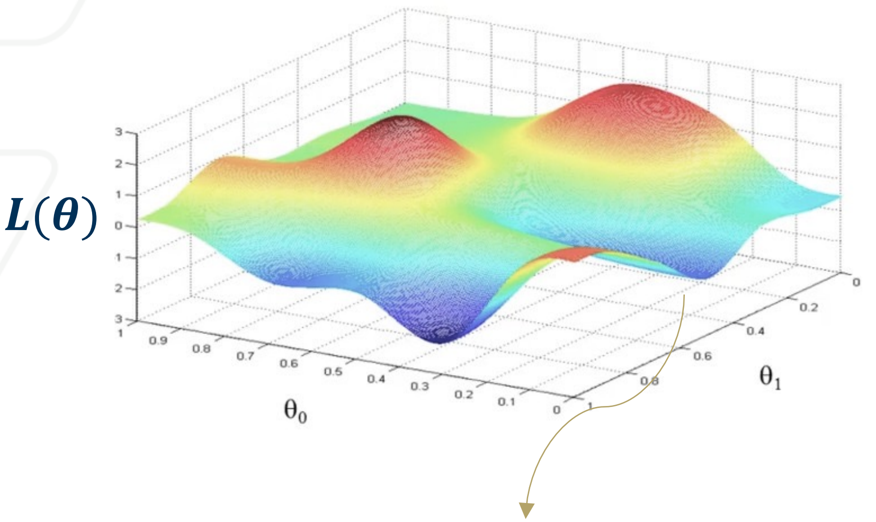
To achieve robustness at Inference, we need the following:

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

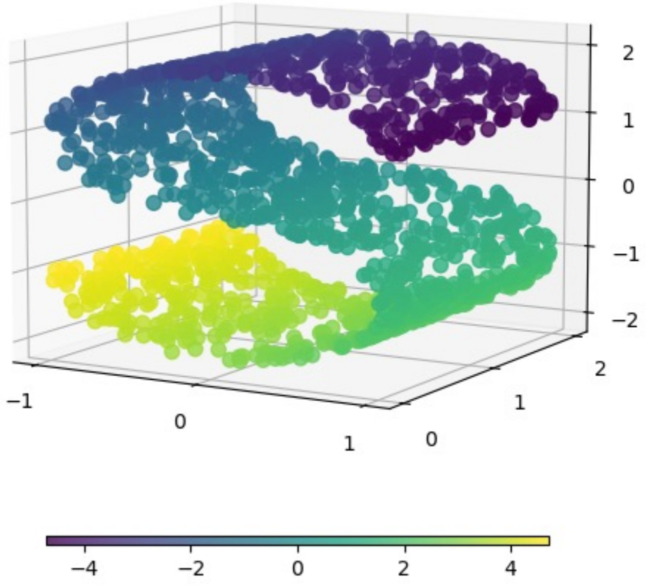
Why is this Challenging?

A Quick note on Manifolds..

Manifolds are compact topological spaces that allow exact mathematical functions



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

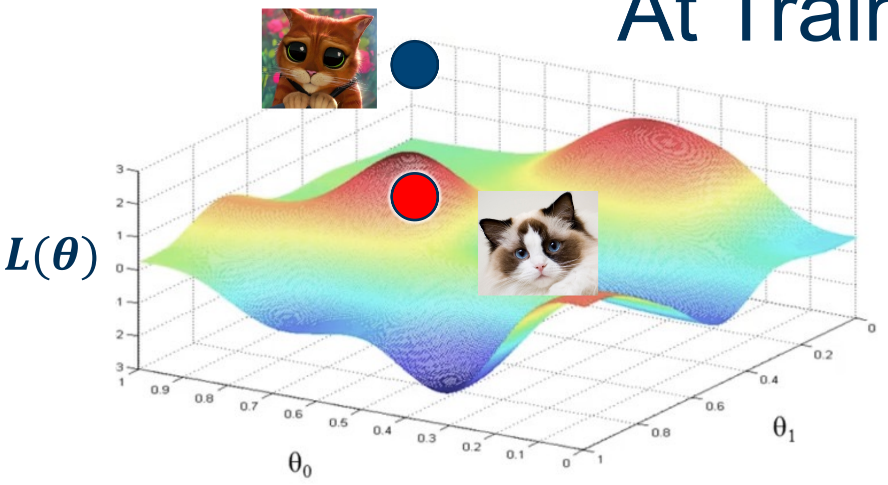


Real data visualizations generated using dimensionality reduction algorithms (Isomap)

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

At Inference

At Training



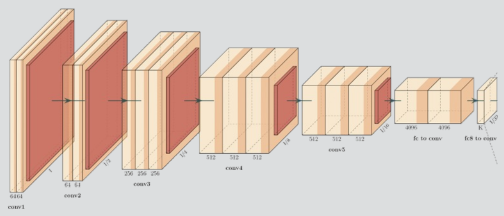
At training, we have access to all training data.

Information at Inference

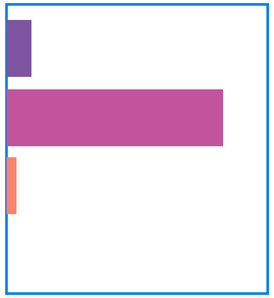
Fisher Information

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

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



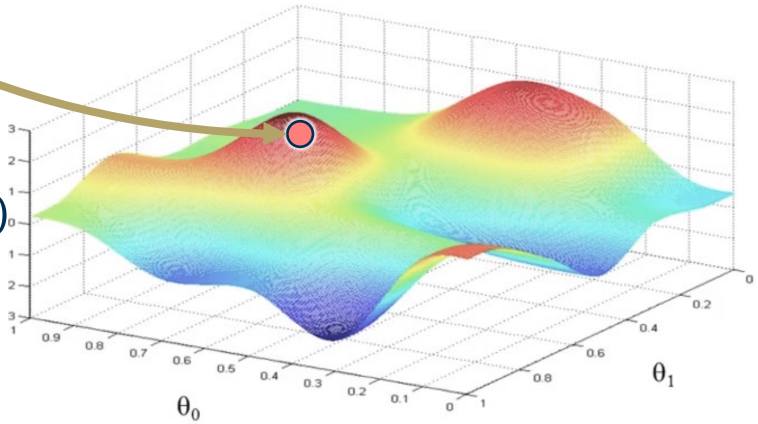
Predicted Class Probability



Fisher Information

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

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



Likelihood function

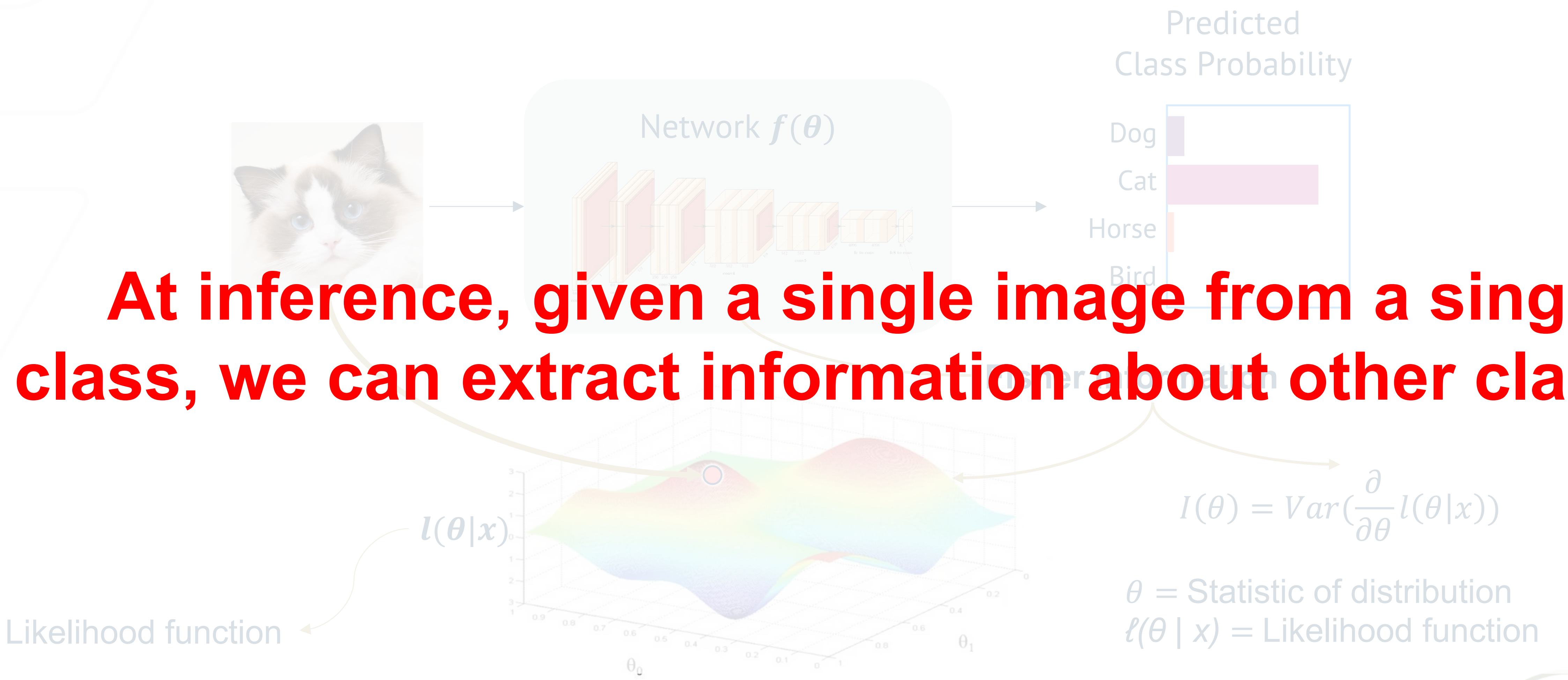
θ = Statistic of distribution

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

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

Information at Inference

Information at Inference



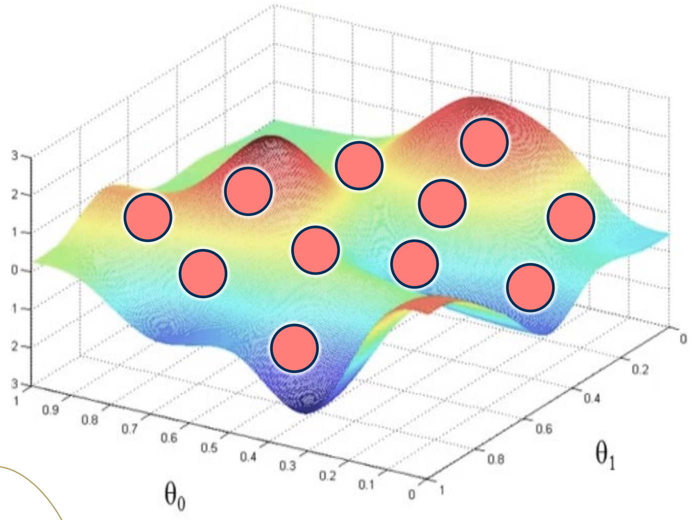
At inference, given a single image from a single class, we can extract information about other classes

Information at Inference

Gradients as Fisher Information

Gradients infer information about the statistics of underlying manifolds

$l(\theta|x)$



Likelihood function instead of loss manifold

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

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

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

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

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

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

OLIVES @GeorgiaTech

Georgia Tech

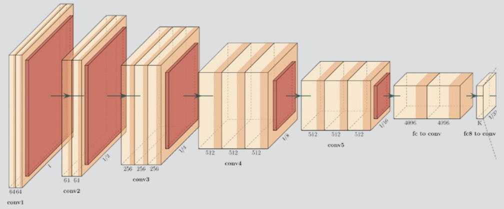
Information at Inference

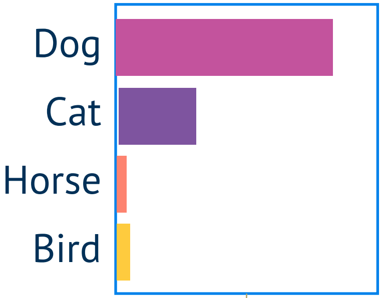
Case Study: Gradients as Fisher Information in Explainability

Gradients infer information about the statistics of underlying manifolds

x

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

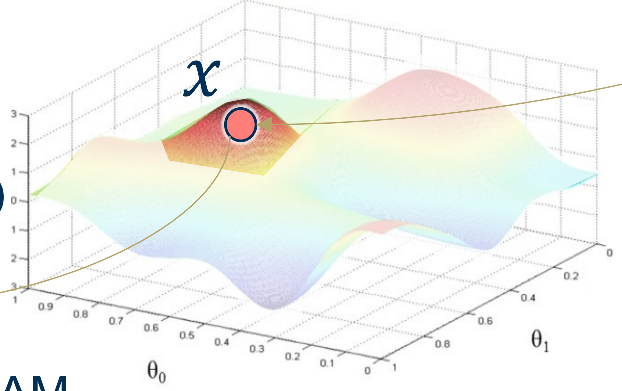




Local information (specific to x) is sufficient!

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

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



Feature attribution via GradCAM

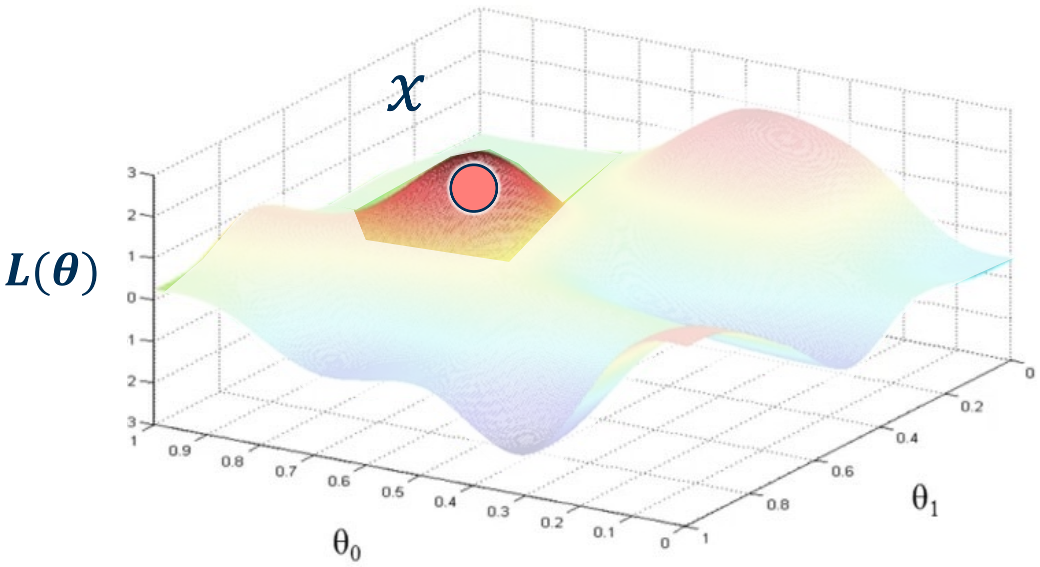
Hence, gradients draw information from the underlying distribution as learned by the network weights!

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

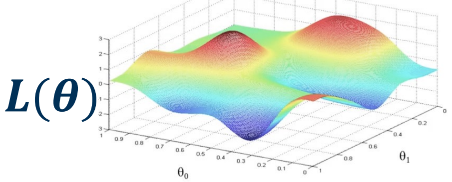
Gradients at Inference

Local Information

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



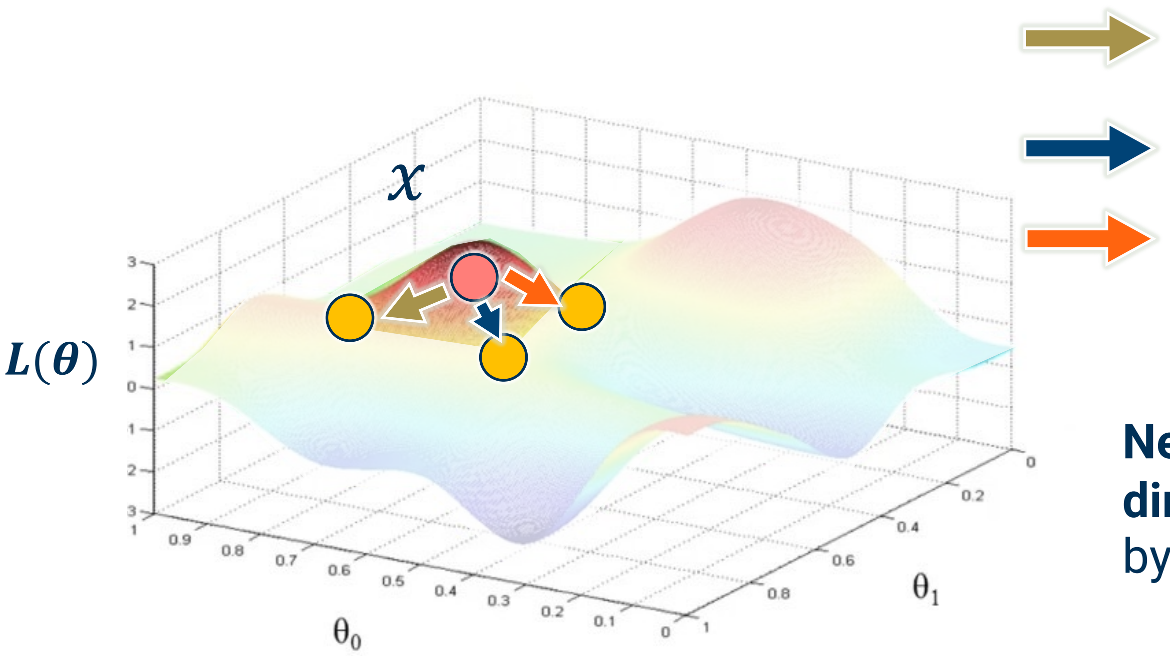
Ideal



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

Direction of Steepest Descent

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



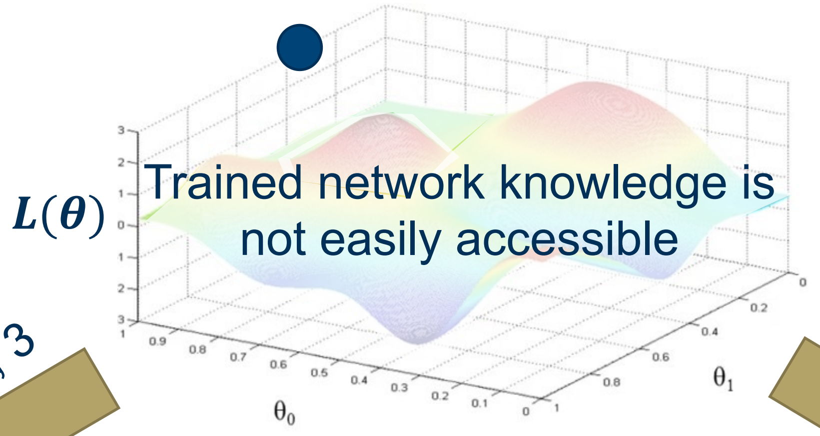
Which direction should we optimize towards (knowing only the local information)?

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

Gradients at Inference

To Characterize the Novel Data at Inference

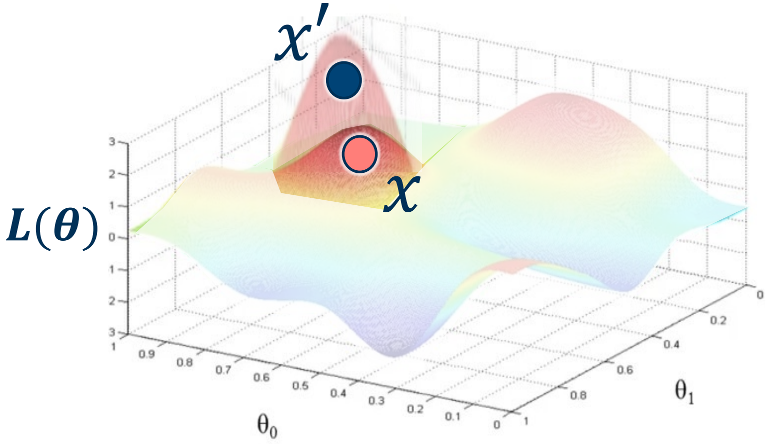
At Inference



Trained network knowledge is not easily accessible

Part 2, 3

Counterfactual and Contrastive Representations using Gradients



Part 4

Representation Traversal using Interventions