Robust Neural Networks: Explainability, Uncertainty, and Intervenability

Ghassan AlRegib, PhD
Professor

Mohit Prabhushankar, PhD
Postdoctoral Fellow

Omni Lab for Intelligent Visual Engineering and Science (OLIVES)
School of Electrical and Computer Engineering
Georgia Institute of Technology
{alregib, mohit.p}@gatech.edu
Dec 15, 2023 – Sorrento, Italy

IEEE BigData 2023
Sorrento, Italy

OLIVES
@GeorgiaTech

GT Georgia Tech

Tutorial Materials
Accessible Online

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

{alregib, mohit.p}@gatech.edu

IEEE BigData 2023 Tutorial

Presented by: *Ghassan AlRegib, and Mohit Prabhushankar*

Georgia Institute of Technology

www.ghassanalregib.info

Title: Robust Neural Networks: Explainability, Uncertainty, and Intervenability

IEEE BigData 2023
Sorrento, Italy

[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]

OLIVES
@GeorgiaTech

Georgia Tech

Expectation vs Reality of Deep Learning

[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]

IEEE BigData 2023
Sorrento, Italy

[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]

OLIVES
@GeorgiaTech

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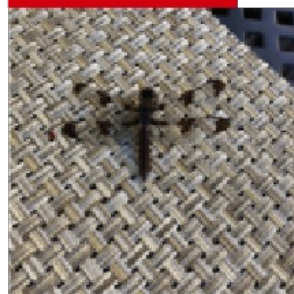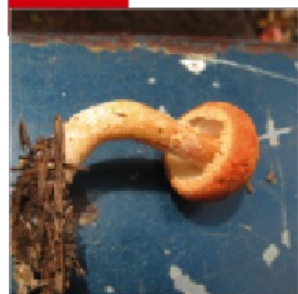
"The best-laid plans of sensors and networks often go awry"

- Engineers, probably

[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]

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

Novel data sources:

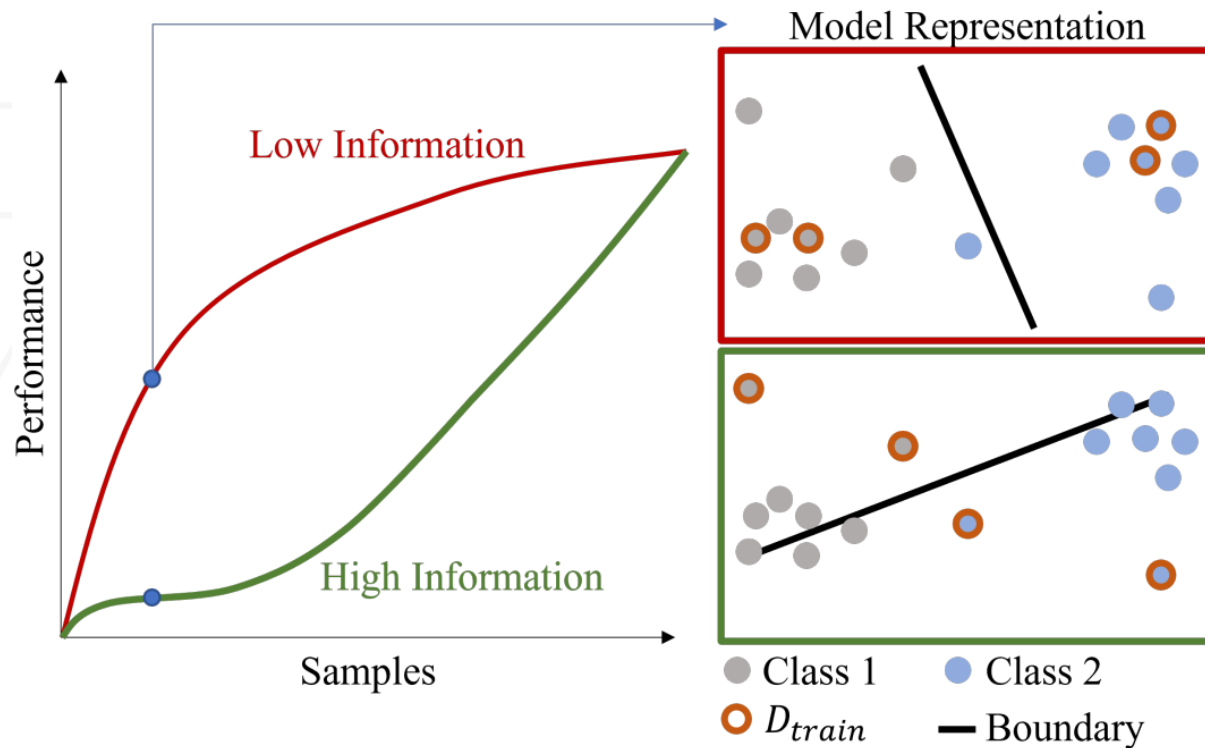- Test distributions
- Anomalous data
- Out-Of-Distribution data
- Adversarial data
- Corrupted data
- Noisy data
- New classes
- …

IEEE BigData 2023
Sorrento, Italy

[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]

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

OLIVES
@GeorgiaTech

Georgia Tech

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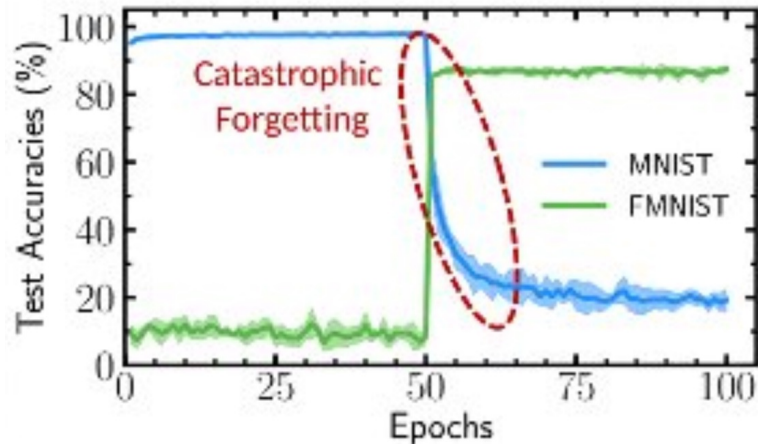

Model Representation

- 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

OLIVES
@GeorgiaTech

Georgia Tech®

## Subsequent training must <u>not</u> 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

IEEE BigData 2023
Sorrento, Italy

[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]

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

Overcoming Challenges at Training

Novel data packs a 1-2 punch!

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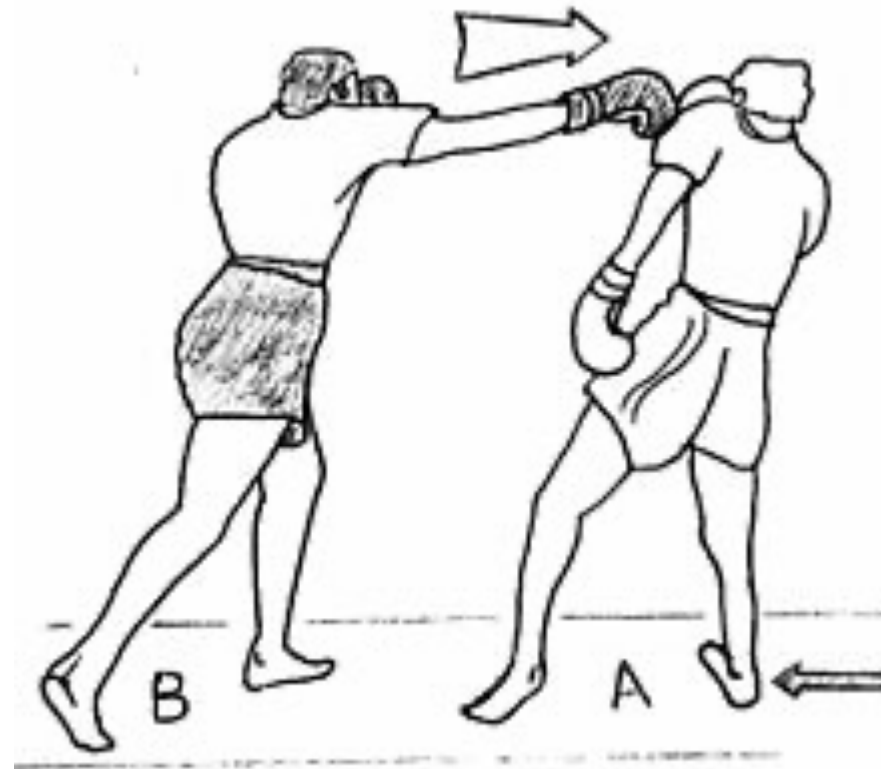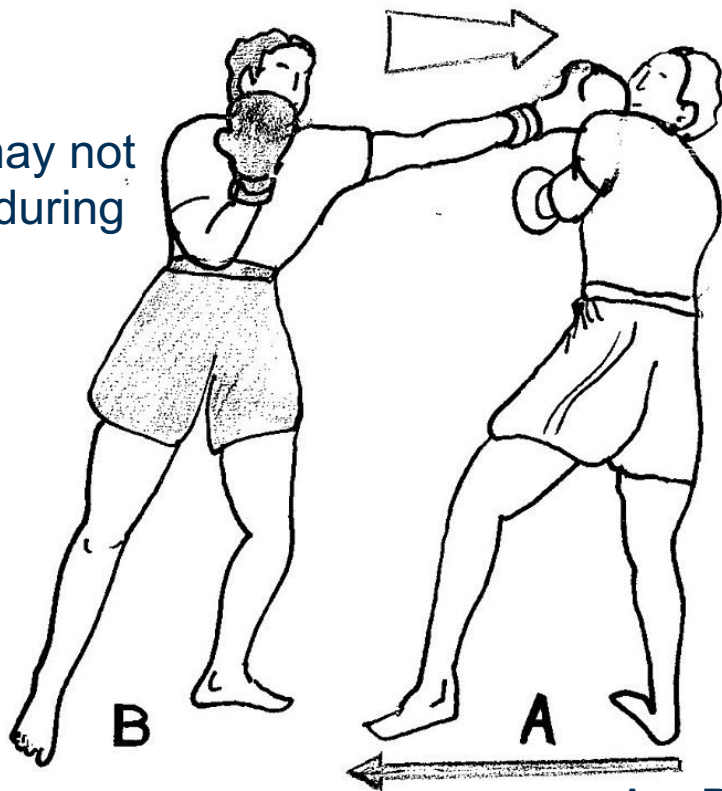
Even if available, novel data does not easily fit into either the earlier or later stages of training

A = Deep Neural Networks
B = Novel data

OLIVES
@GeorgiaTech

Georgia Tech.

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

IEEE BigData 2023
Sorrento, Italy

[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]
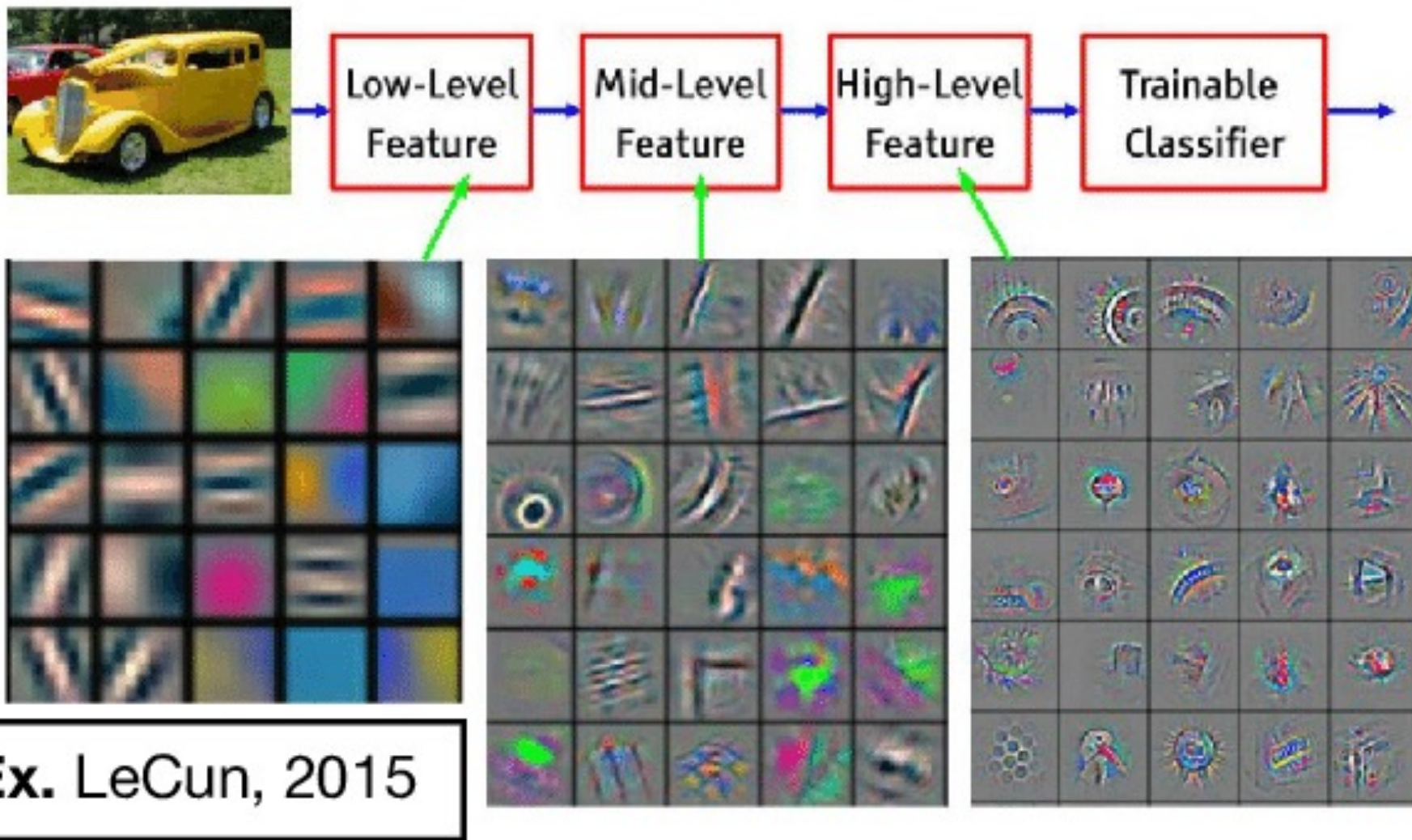
OLIVES
@GeorgiaTech

Georgia Tech

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

IEEE BigData 2023
Sorrento, Italy

[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]

OLIVES
@GeorgiaTech

Georgia Tech

Robust Neural Networks
Part I: Inference in Neural Networks

OLIVES
@GeorgiaTech

Georgia Tech

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
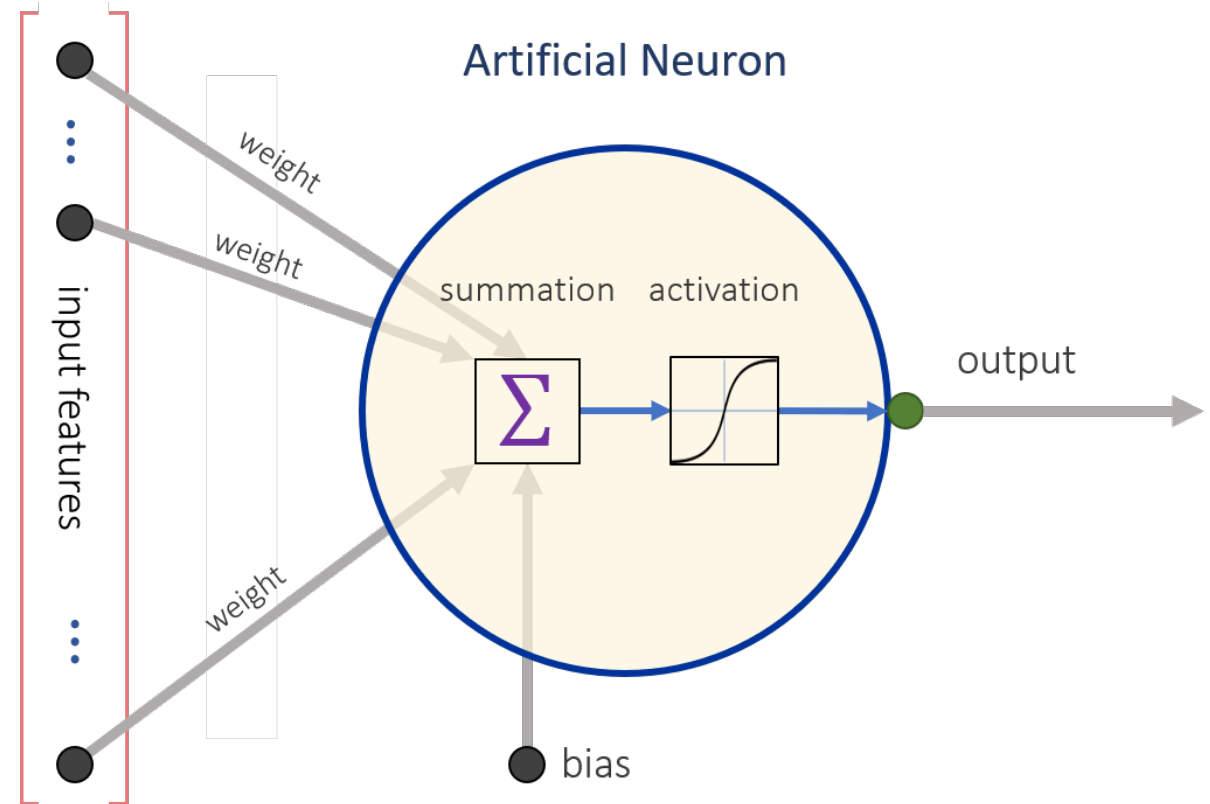- Part 5: Conclusions and Future Directions

IEEE BigData 2023
Sorrento, Italy

[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]

OLIVES
@GeorgiaTech

Georgia Tech

Deep Learning
Overview

Low-Level Feature → Mid-Level Feature → High-Level Feature → Trainable Classifier

Ex. LeCun, 2015

OLIVES @GeorgiaTech

Georgia Tech

The underlying computation unit is the Neuron

Artificial neurons consist of:

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

Artificial Neuron

IEEE BigData 2023
Sorrento, Italy

[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]

OLIVES
@GeorgiaTech

Georgia Tech

Neurons are stacked and densely connected to construct ANNs
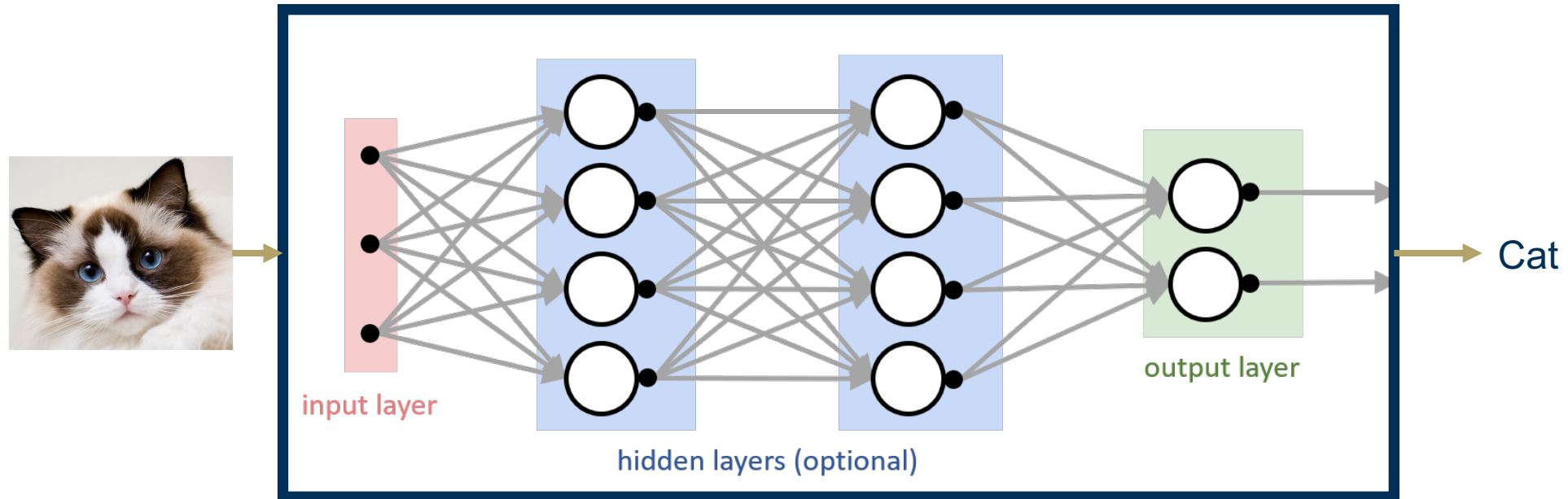
Typically, a neuron is part of a network organized in layers:
- An input layer (Layer 0)
- An output layer (Layer K)
- Zero or more hidden (middle) layers (Layers $1 \ldots K-1$)

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

Ex. LeCun, 2015

[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]
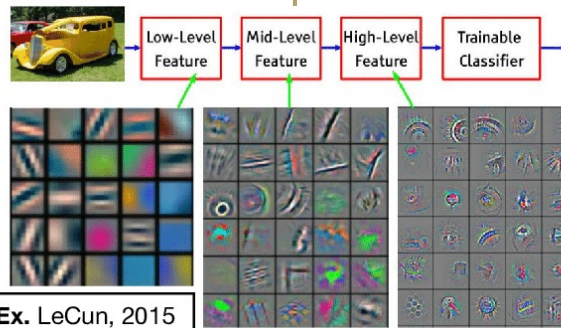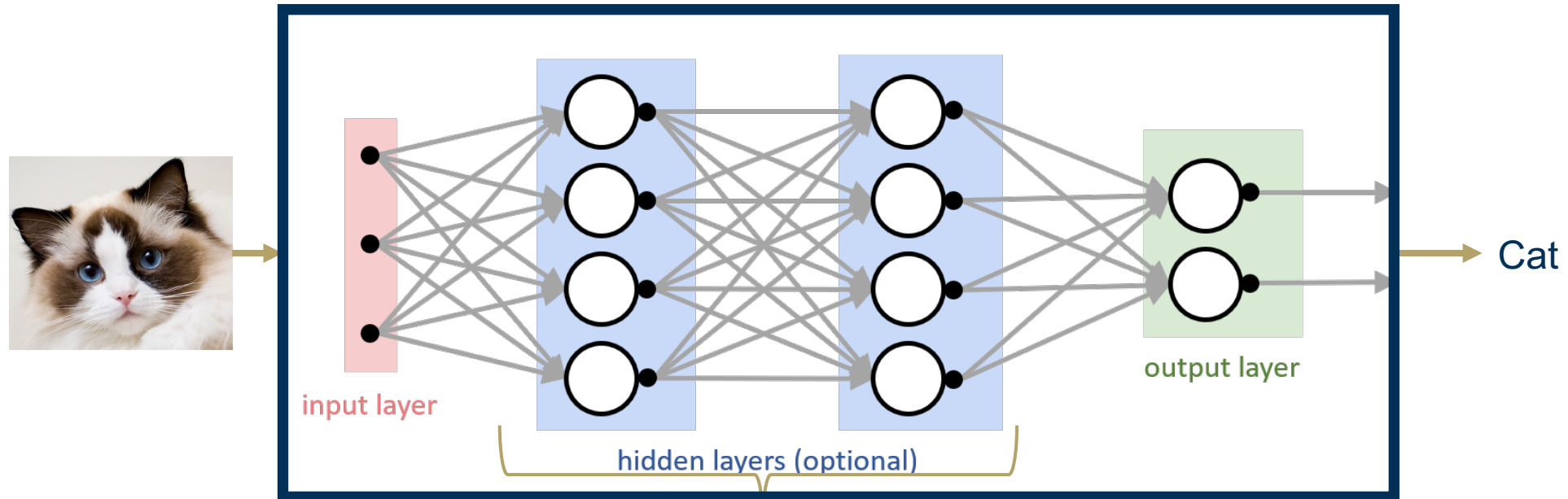
Transformers, Large Language Models and Foundation Models
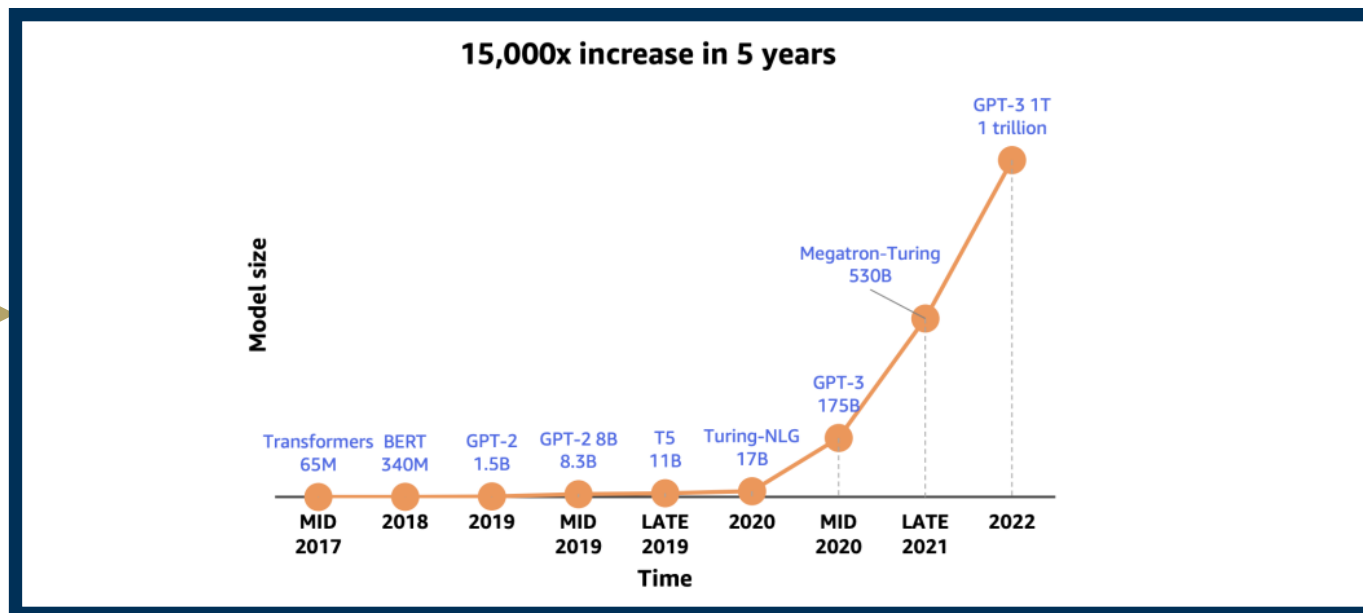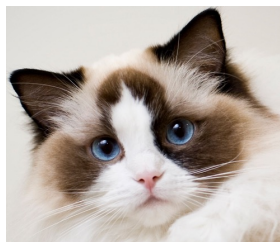
Primary reasons for advancements:
1. Expanded interests from the research community
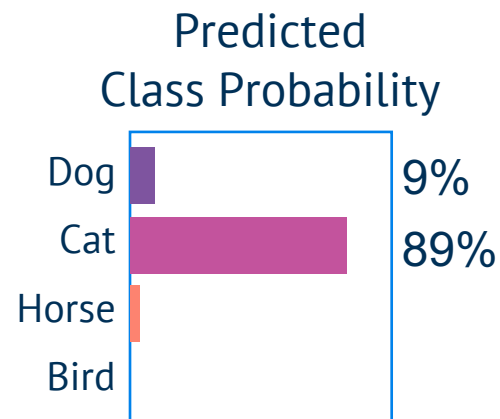2. Computational resources availability
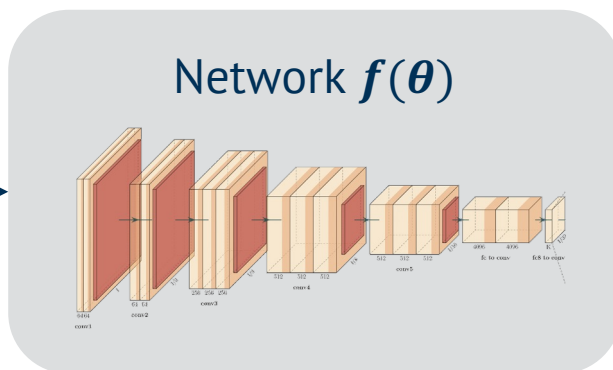3. **Big data availability**

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

Network $f(\theta)$

Predicted
Class Probability

Dog	9%
Cat	89%
Horse	
Bird	

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>

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

x'

Predicted
Class Probability

Dog — 39%
Cat — 53%
Horse
Bird
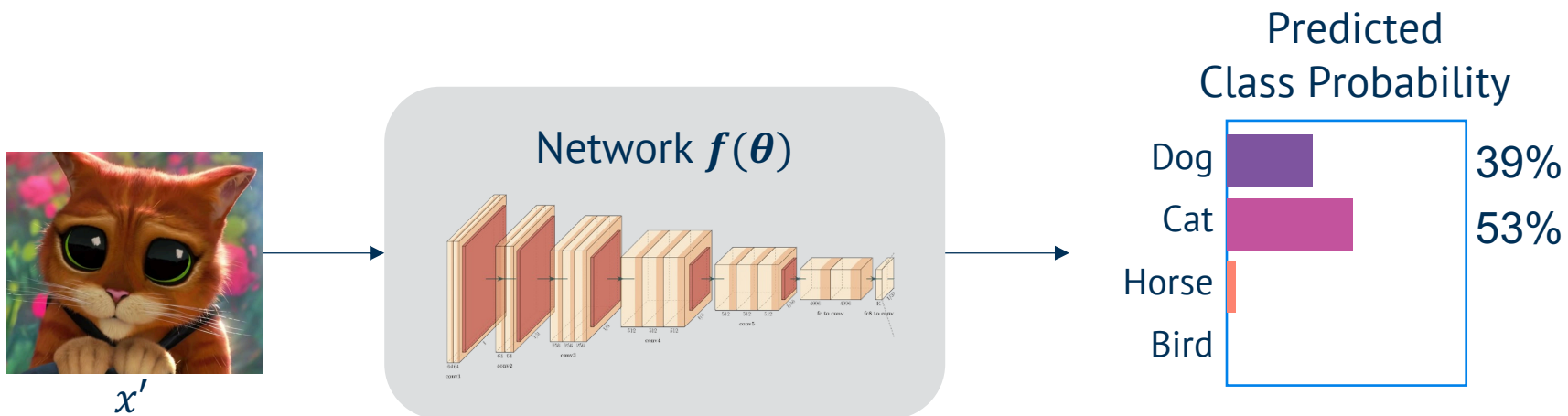
If $x \in \chi$, the data is
novel

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

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

OLIVES
@GeorgiaTech

Georgia Tech

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

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

x'

Predicted
Class Probability

Dog	39%
Cat	53%
Horse	
Bird	

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
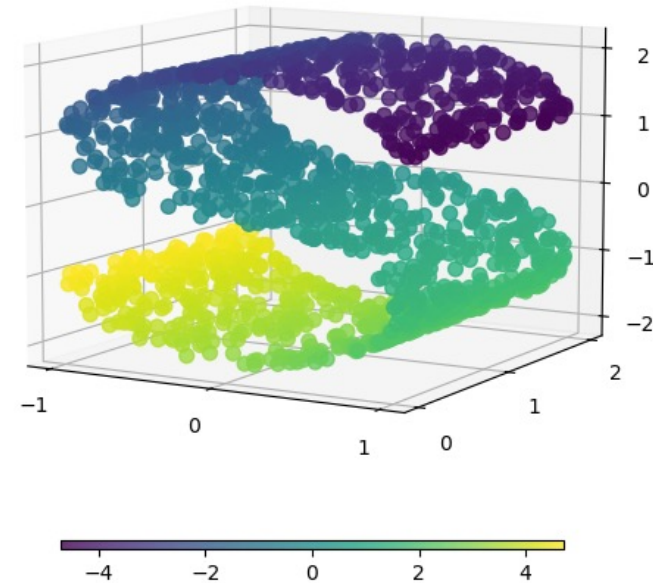
Why is this Challenging?

OLIVES
@GeorgiaTech

Georgia
Tech

Manifolds are compact topological spaces that allow exact mathematical functions

$L(\boldsymbol{\theta})$

θ_0

θ_1
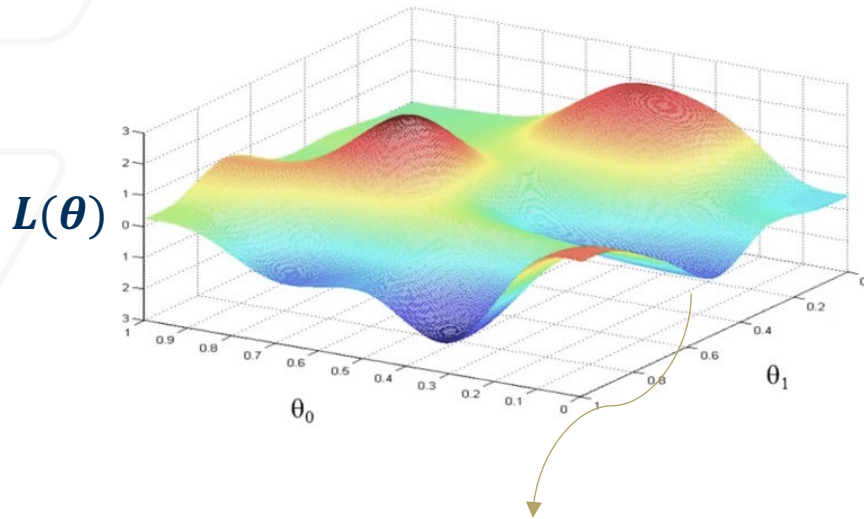
Toy visualizations generated using functions
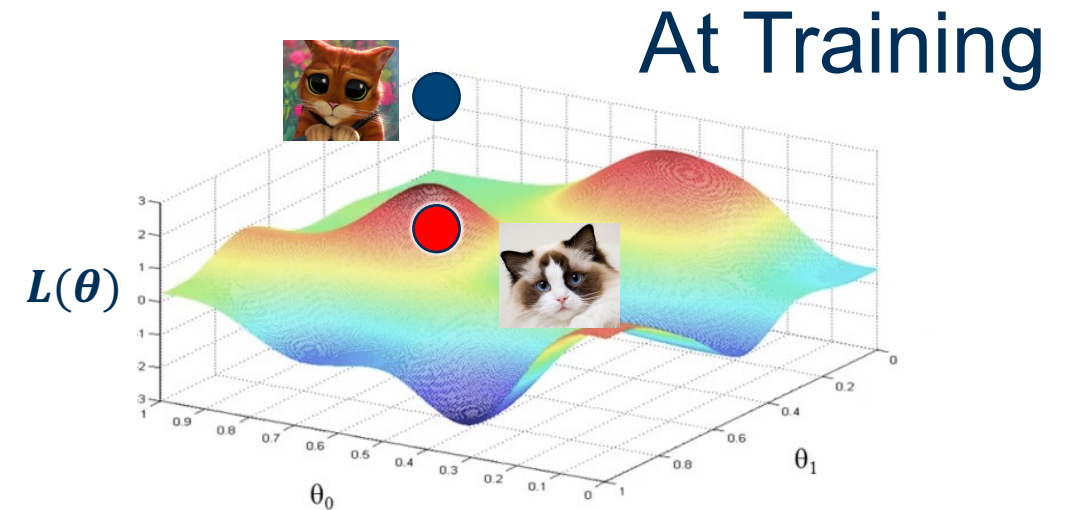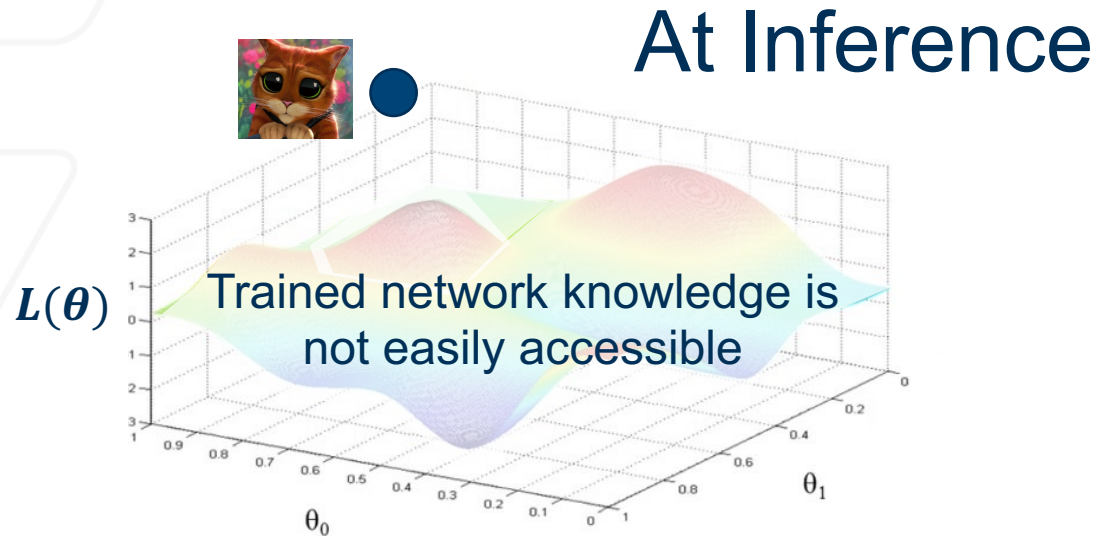(and thousands of generated data points)

Real data visualizations generated using
dimensionality reduction algorithms (Isomap)

OLIVES
@GeorgiaTech

Georgia
Tech

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

At Inference

Trained network knowledge is not easily accessible

$L(\theta)$

θ_0

θ_1

At Training

$L(\theta)$

θ_0

θ_1

At training, we have access to all training data.

[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]

OLIVES
@GeorgiaTech

Georgia Tech

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

Predicted
Class Probability

Network $f(\theta)$

Dog
Cat
Horse
Bird

Fisher Information

$l(\theta | x)$
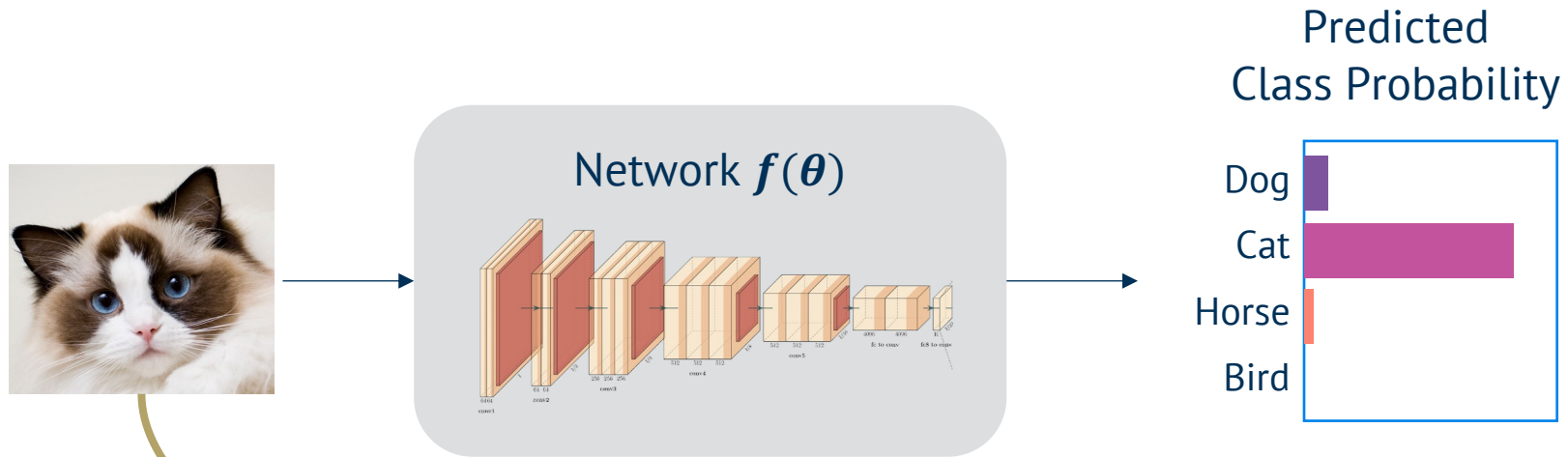
Likelihood function

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

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

[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]

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

OLIVES
@GeorgiaTech

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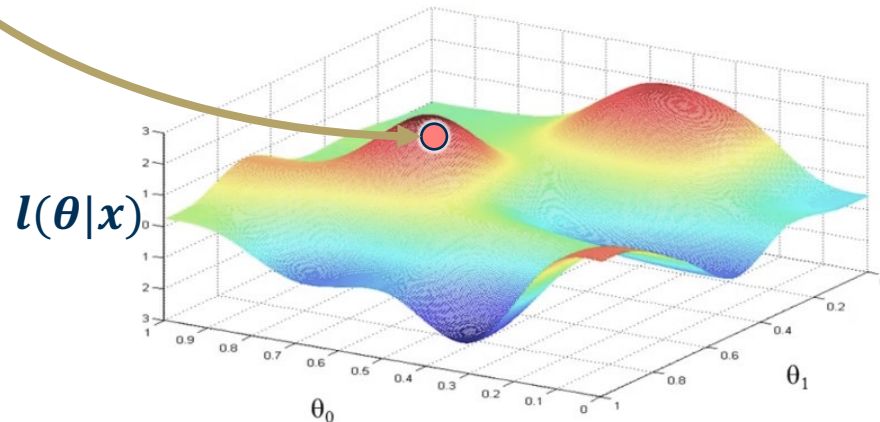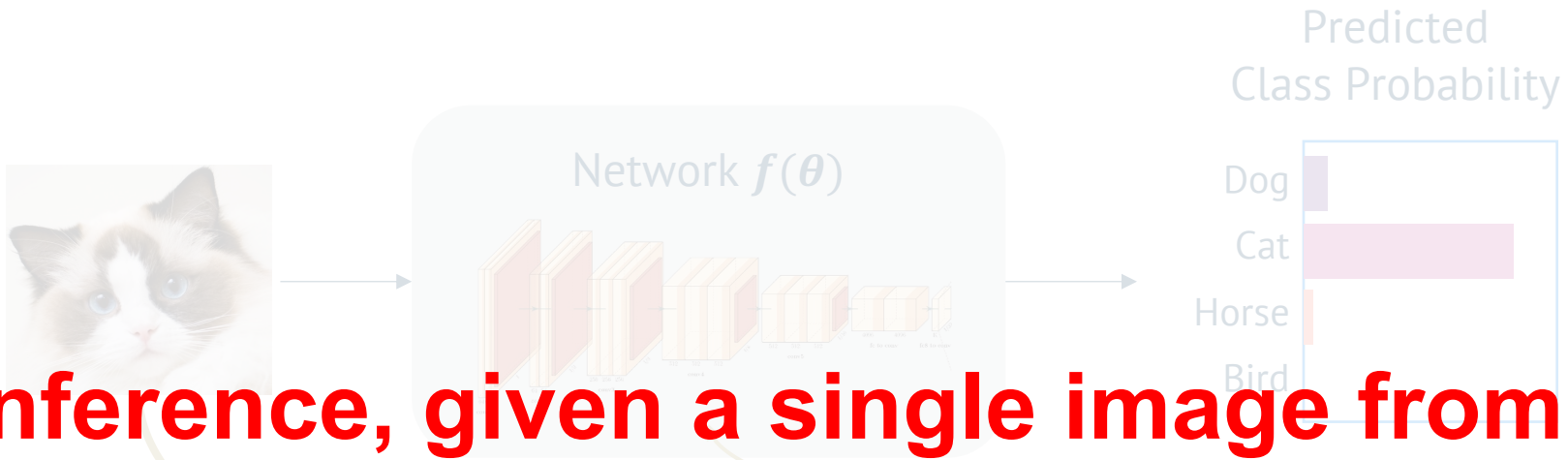
At inference, given a single image from a single class, we can extract information about other classes

Gradients infer information about the statistics of underlying manifolds

$l(\theta|x)$

Likelihood function instead of loss manifold
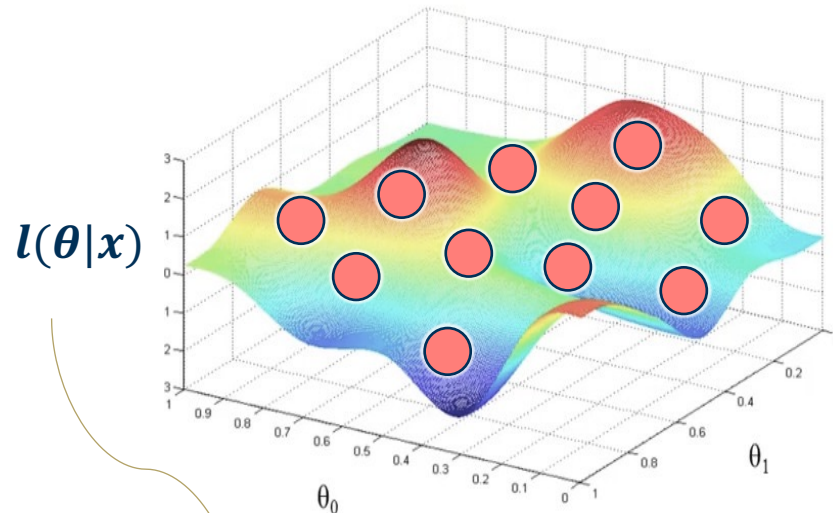
From before, $I(\theta) = Var(\frac{\partial}{\partial \theta} l(\theta|x))$

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

$I(\theta) = E[U_\theta U_\theta^T]$ where
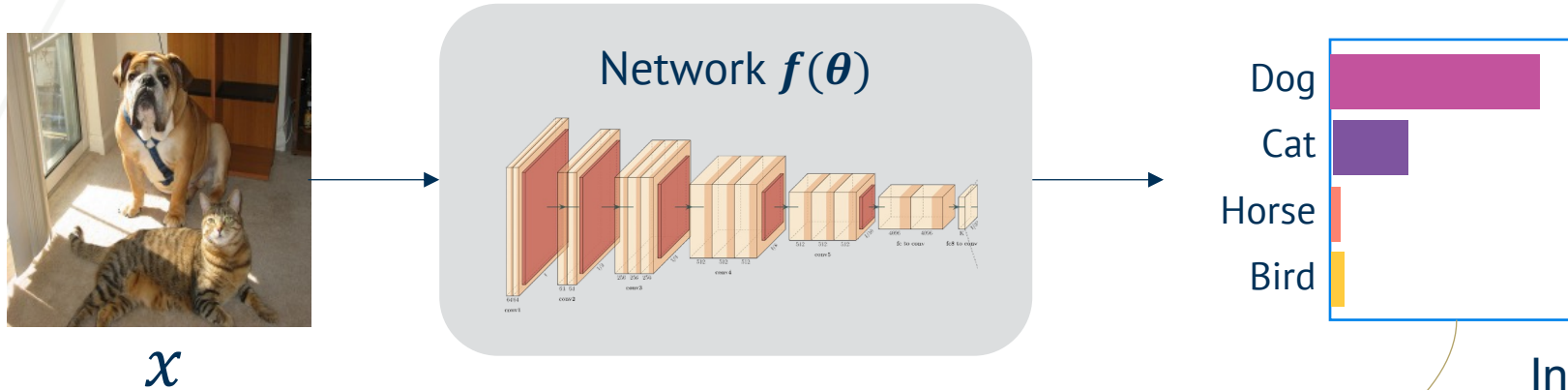
$E[\cdot]$ = Expectation
$U_\theta = \nabla_\theta l(\theta|x)$, Gradients w.r.t. the sample

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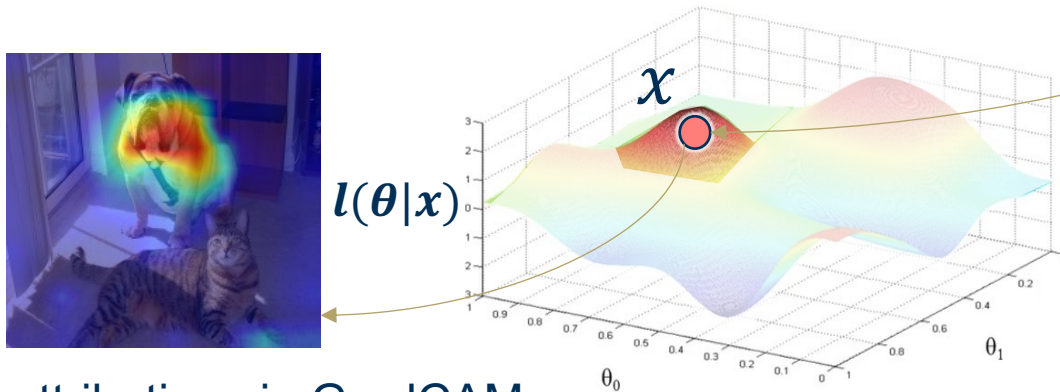
OLIVES
@GeorgiaTech

Georgia Tech

Gradients infer information about the statistics of underlying manifolds

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

Dog
Cat
Horse
Bird

x

Local information (specific to x) is sufficient!

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

x

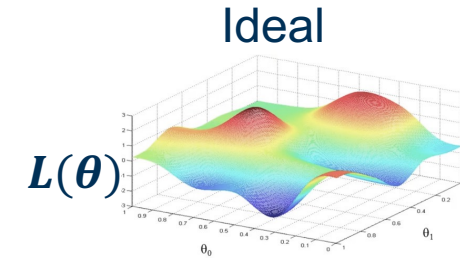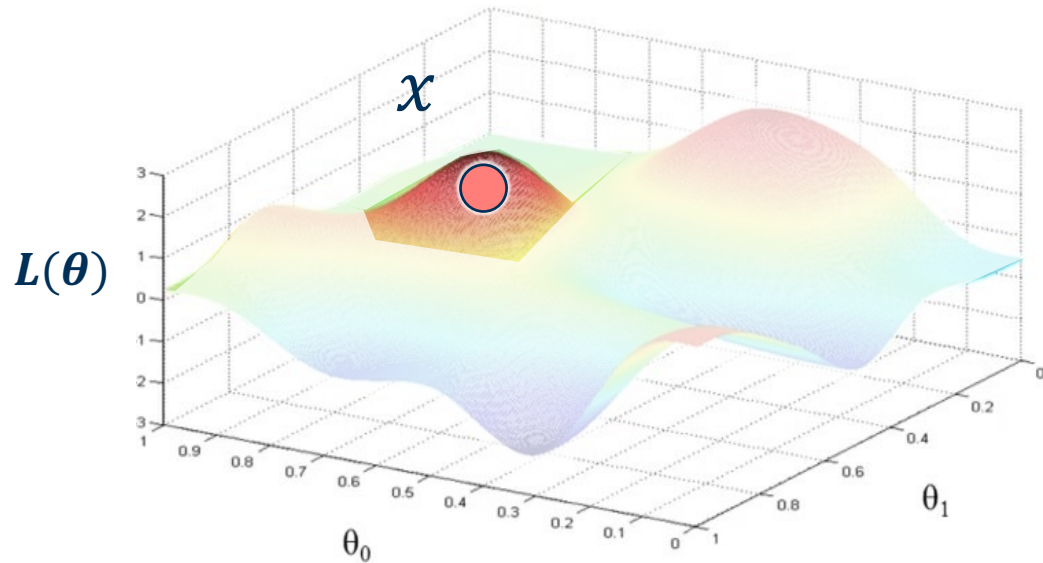In this case, the image and its prediction extracts nose, mouth and jowl features.

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

Feature attribution via GradCAM

OLIVES @GeorgiaTech

Georgia Tech

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

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

Ideal

$L(\theta)$

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

IEEE BigData 2023
Sorrento, Italy
[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]
OLIVES
@GeorgiaTech
Georgia Tech

Gradients allow choosing the fastest direction of descent given a loss function $L(\theta)$
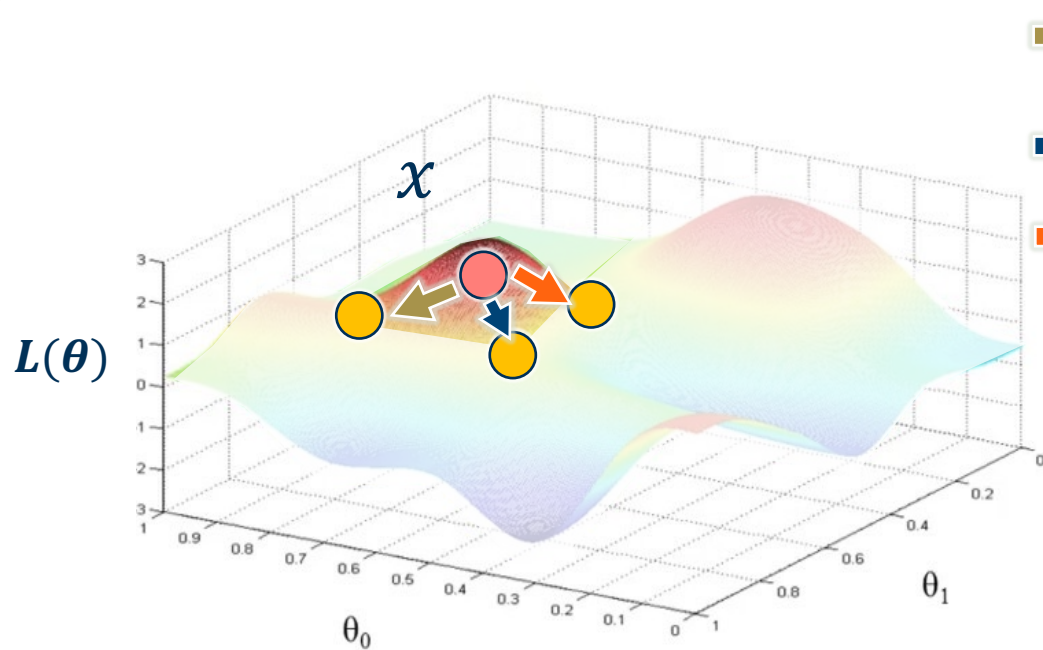
Path 1?

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 $L(\theta)$

[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]

OLIVES
@GeorgiaTech

Georgia Tech

To Characterize the Novel Data at Inference

$L(\theta)$ — Trained network knowledge is not easily accessible

At Inference

Part 2, 3

Counterfactual and Contrastive Representations using Gradients

Part 4

Representation Traversal using Interventions

x'

x

$L(\theta)$

$L(\theta)$

IEEE BigData 2023
Sorrento, Italy

[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]

OLIVES
@GeorgiaTech

Georgia Tech