EUVIP 2021

Explainable and Robust Machine Learning for Images

Prof. Ghassan AlRegib

Mohit Prabhushankar

Gukyeong Kwon

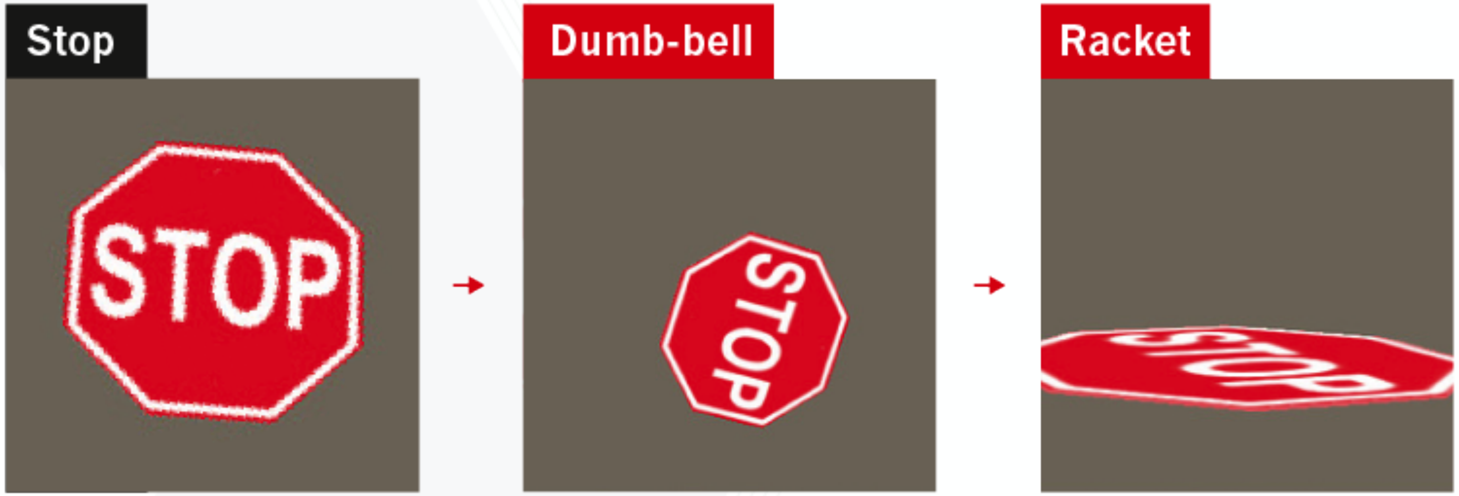
Jinsol Lee

OLIVES
@GeorgiaTech
www.ghassanalregib.com

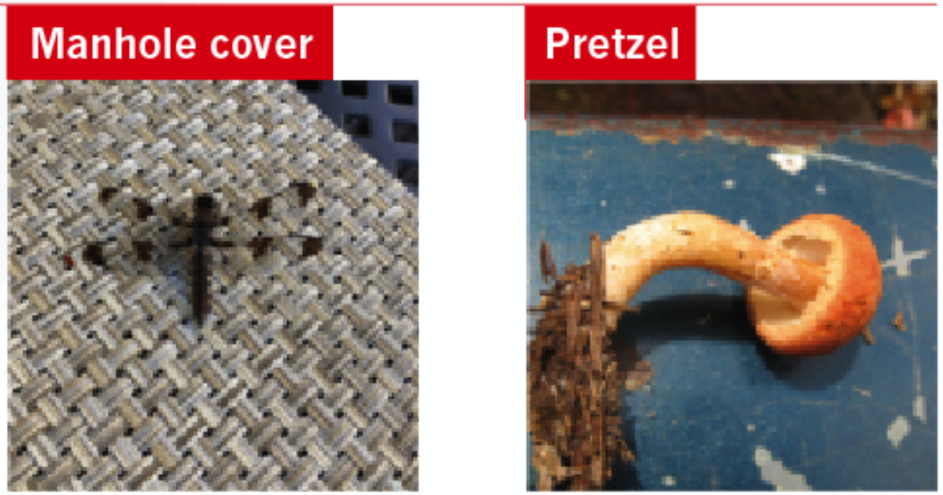
Georgia Tech
CREATING THE NEXT

Challenges in Neural Networks

Rotating objects in an image confuses DNNs, probably because they are too different from the types of image used to train the network.

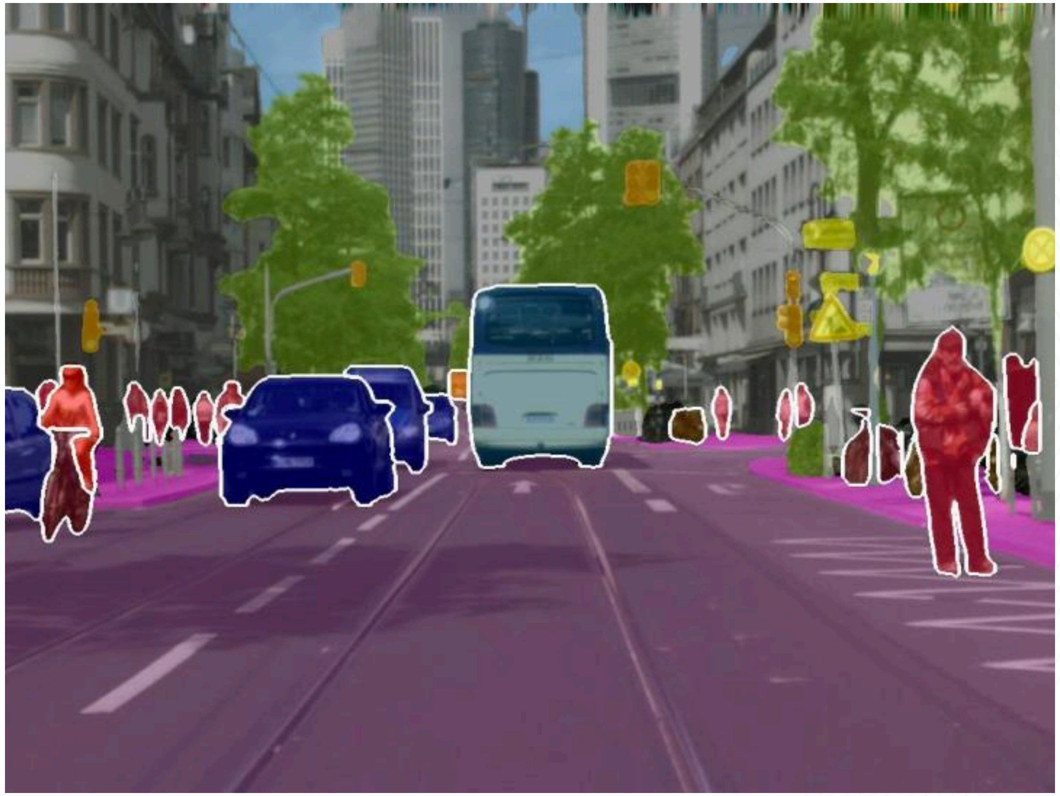


Even natural images can fool a DNN, because it might focus on the picture's colour, texture or background rather than picking out the salient features a human would recognize.



©nature

Data and Neural Networks



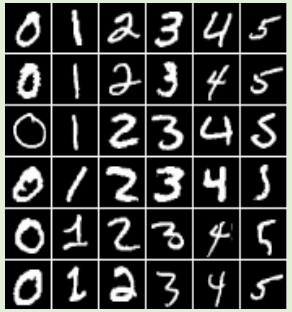
OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

Introduction
Limitations of Neural Networks

Classifier

Trained with



?

Classifier

Trained with

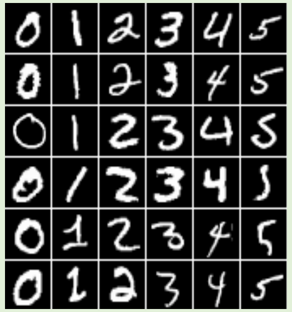
?

Introduction

Limitations of Neural Networks

Classifier

Trained with



?

Don’t trust these predictions!

Classifier

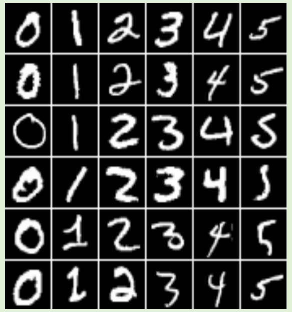
Trained with

?

Introduction
Understanding Model Uncertainty

Classifier

Trained with



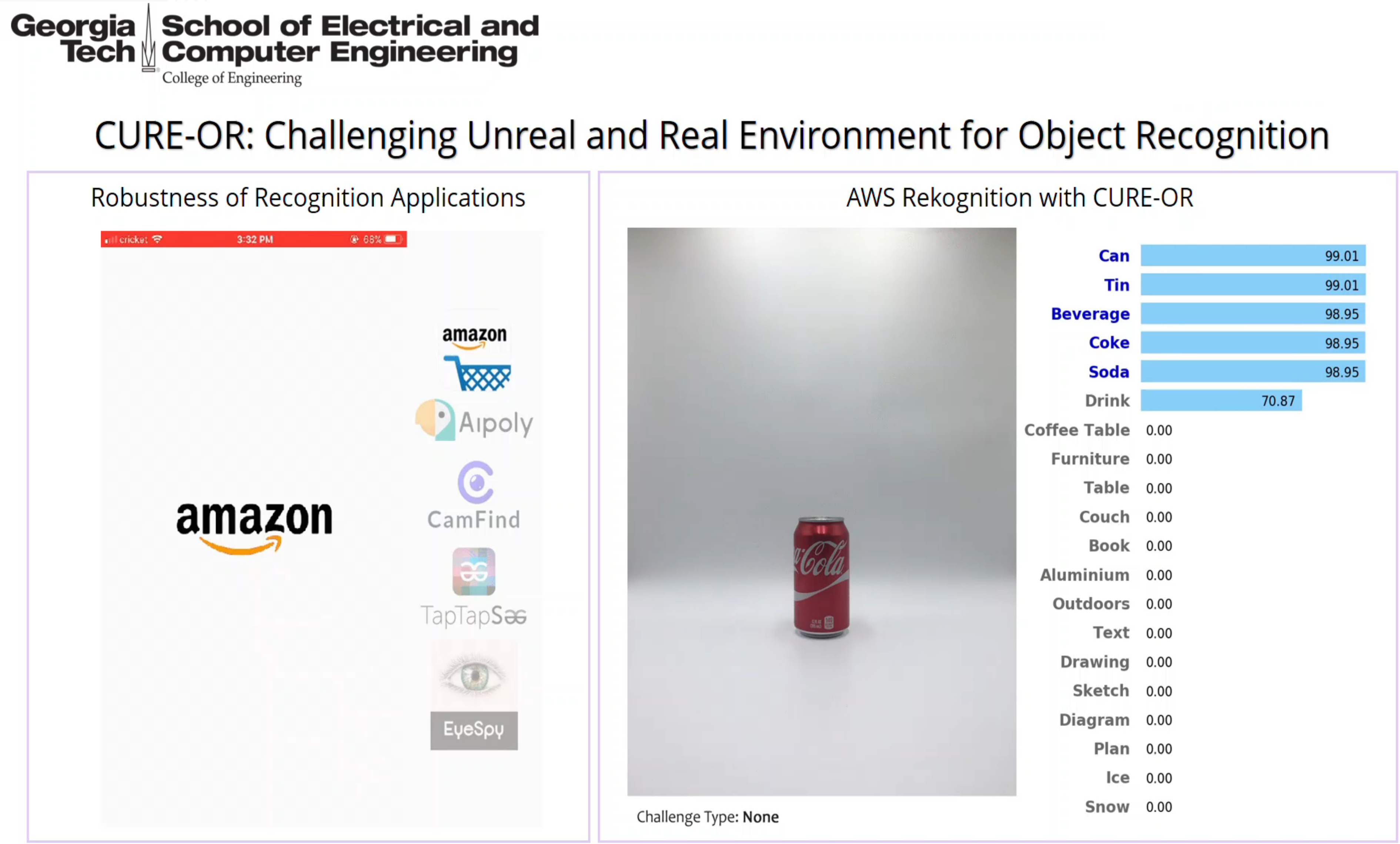
Classifier

Trained with

(1) How certain / familiar are you with a given input?
(2) Can you detect Anomalies in input data?

Introduction

CURE-OR: Challenging Unreal and Real Environment for Object Recognition



D. Temel*, J. Lee*, and G. AlRegib, "CURE-OR: Challenging unreal and real environments for object recognition," ICMLA 2018

Introduction
Robustness in Autonomous Vehicles

Robust Autonomous Driving Under Challenging Conditions

D. Temel, M. Chen, T. Alshawi, and G. AlRegib, "CURE-TSD: Challenging Unreal and Real Environments for Traffic Sign Detection"

Real Video

Dataset Generation

Video with Challenging Conditions

10 Datasets @Zenodo

OLIVES@GeorgiaTech

Recent uploads

New upload

Search OLIVES@GeorgiaTech

November 12, 2020 (v1) | Dataset | Open Access

View

CURE-OR-Sampled: Challenging Unreal and Real Environments for Object Recognition

Dogancan Temel; Jinsol Lee; Ghassan AlRegib;

File descriptions train.zip - the training set test.zip - the test set train.csv - the ground truth for the training images with the following information: imageID, class, background, perspective, challengeType, challengeLevel sample_submission.csv - a sample submissio

Uploaded on November 12, 2020

July 8, 2020 (1.0) | Dataset | Open Access

View

CoMMons

AlRegib, Ghassan; Hu, Yuting; Long, Zhiling; Sunderasan, A.; Alfarraj, Motaz;

Recognizing textures and materials in real-world images has played an important role in object recognition and scene

Community

OLIVES @GeorgiaTech

OLIVES@GeorgiaTech

This community contains codes and datasets

Georgia Tech
CREATING THE NEXT

Introduction
Explanations

Explanations are a set of rationales used to understand the reasons behind a decision

Question

Name of the bird?

Answer

Spoonbill

Why Spoonbill?

Shallow-water bird with flattened beak and football shaped body. They are pale pink birds with pink shoulders and rump. They have a white neck and a partially feathered, yellow green head.

Language-based explanation

Introduction

Visual Explanations

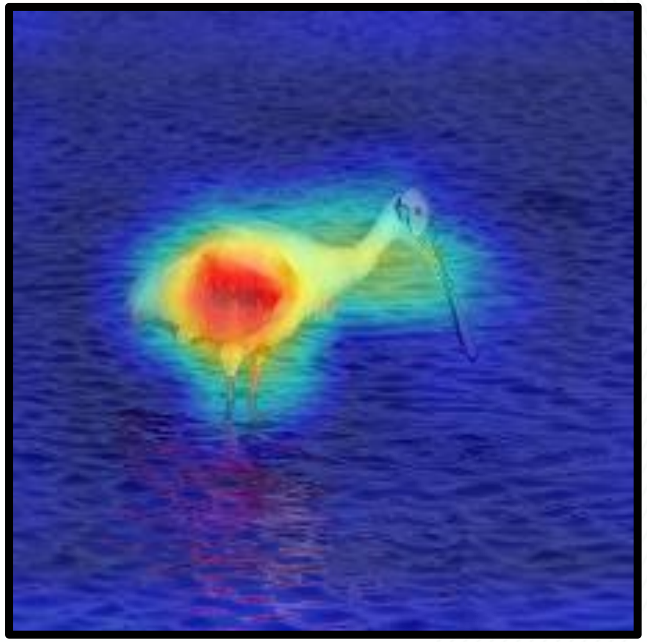
Visual characteristics that are used to justify decisions are termed as visual explanations

Question	Answer
Name of the bird?	Spoonbill

Why Spoonbill?

Shallow-water bird with flattened beak and football shaped body. They are pale pink birds with pink shoulders and rump. They have a white neck and a partially feathered, yellow green head.

Language-based explanation



Visual Explanation

OLIVES @GeorgiaTech

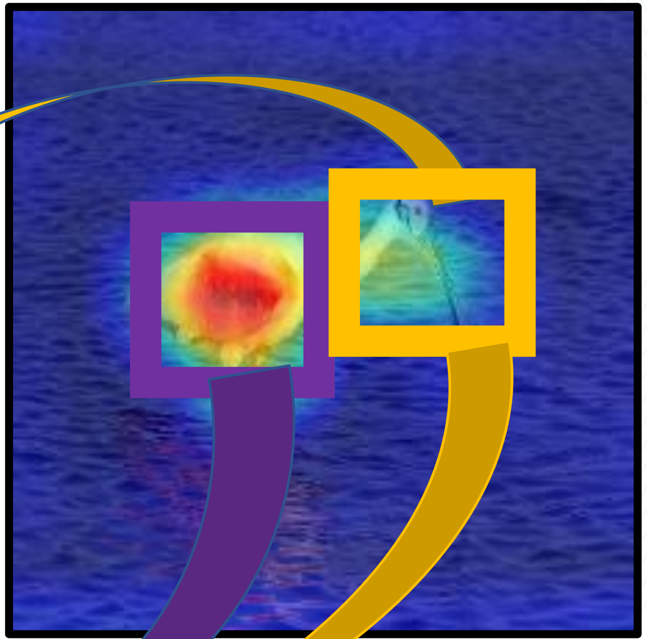
Introduction
Visual Explanations

Visual characteristics that are used to justify decisions are termed as visual explanations

Question	Answer
Name of the bird?	Spoonbill

Why Spoonbill?

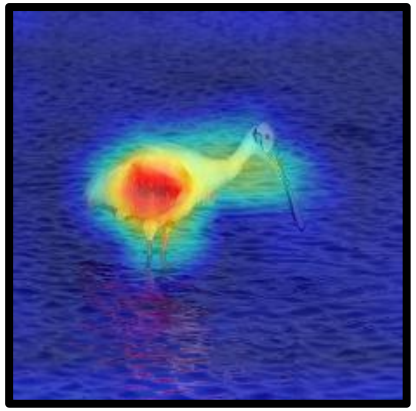
Shallow-water bird with flattened beak and football shaped body. They are pale pink birds with pink shoulders and rump. They have a white neck and a partially feathered, yellow-green head.



Causal factors based visual explanations – answers to `*Why?*' Questions

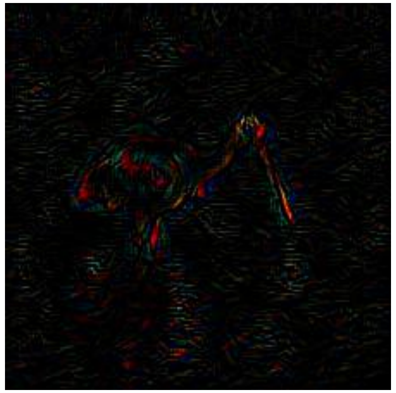
Introduction
Visual Explanations

`Why P?'

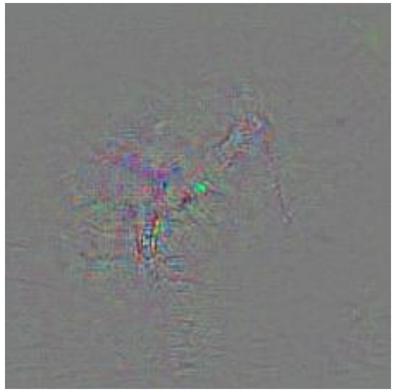


Grad-CAM

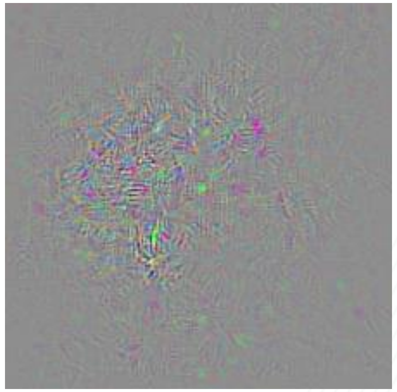
Guided Backpropagation



Positive saliency



Smooth Gradients



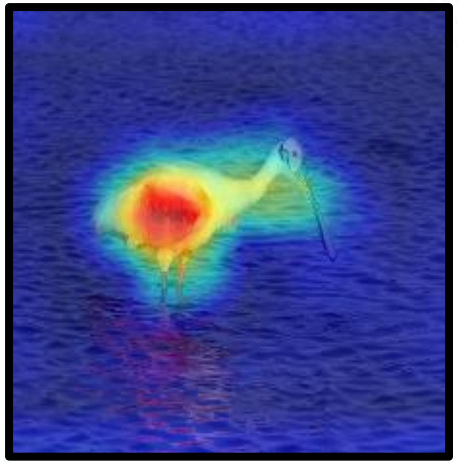
Vanilla Backpropagation

Introduction

Contrastive Visual Explanations

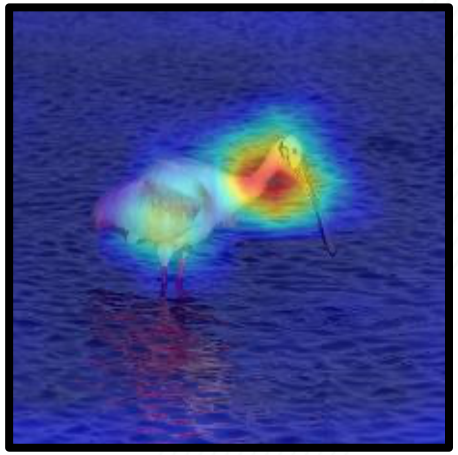
Why Spoonbill?

Shallow-water bird with flattened beak and football shaped body. They are pale pink birds with pink shoulders and rump. They have a white neck and a partially feathered, yellow green head.



Why Spoonbill, rather than Flamingo?

Spoonbills have shorter legs and necks compared to Flamingos

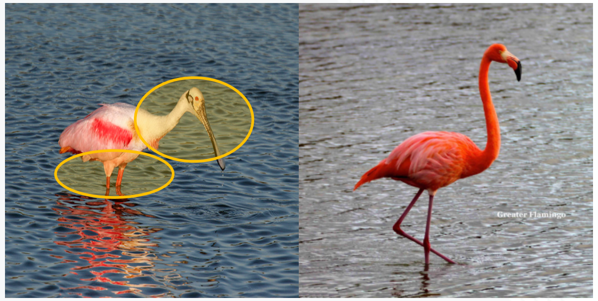


Contrastive visual explanations – answers to *`Why P, rather than Q?'* Questions

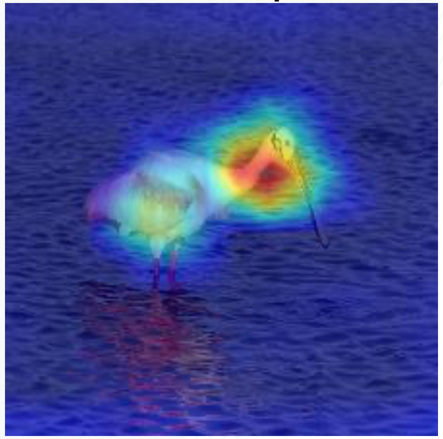
Introduction

Objectives of Contrastive Visual Explanations

Contrast B/w Spoonbill and Flamingo



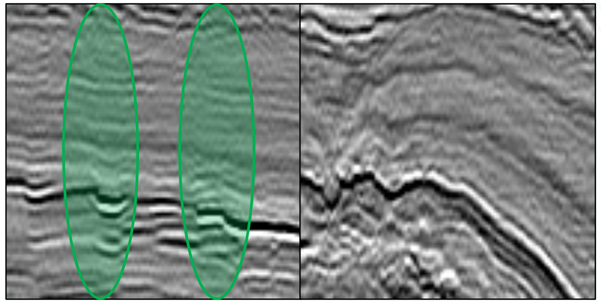
Our Output



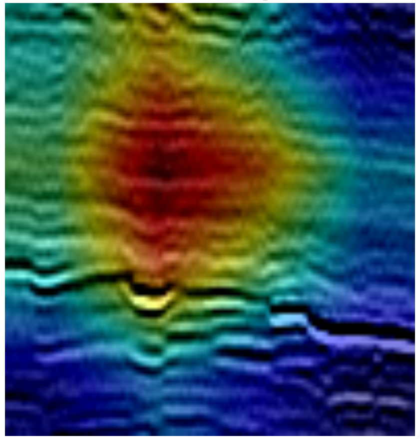
Contrast B/w Bugatti Convertible and Coupe

Our Output

Contrast B/w Fault and Salt Dome



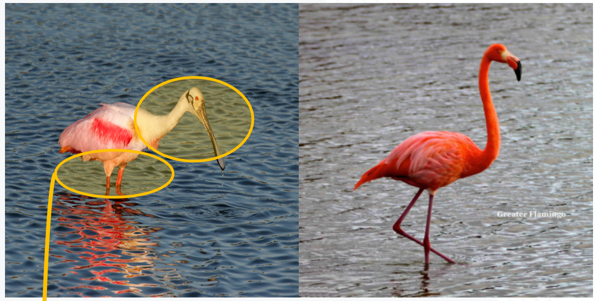
Our Output



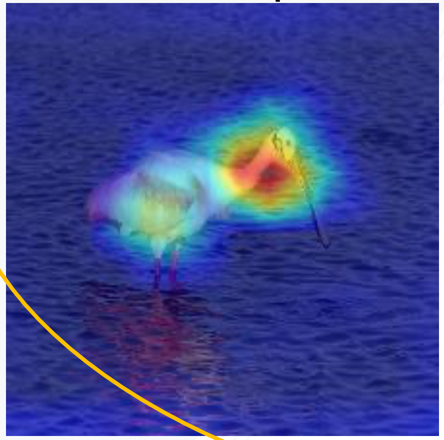
Introduction

Objectives of Contrastive Visual Explanations

Contrast B/w Spoonbill and Flamingo



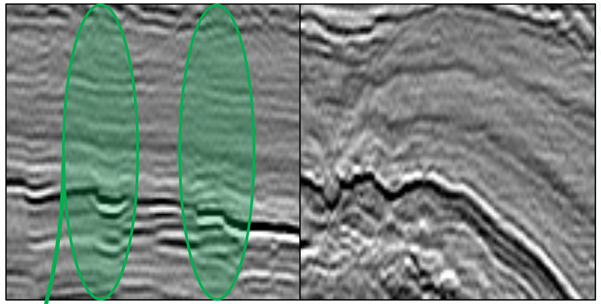
Our Output



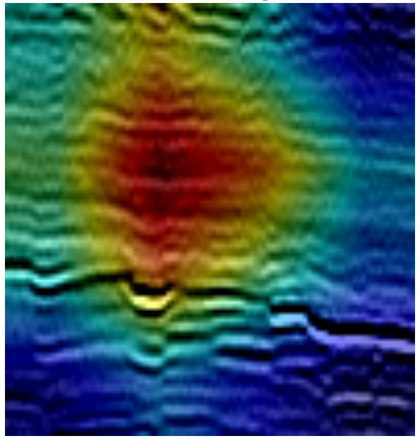
Contrast B/w Bugatti Convertible and Coupe

Our Output

Contrast B/w Fault and Salt Dome



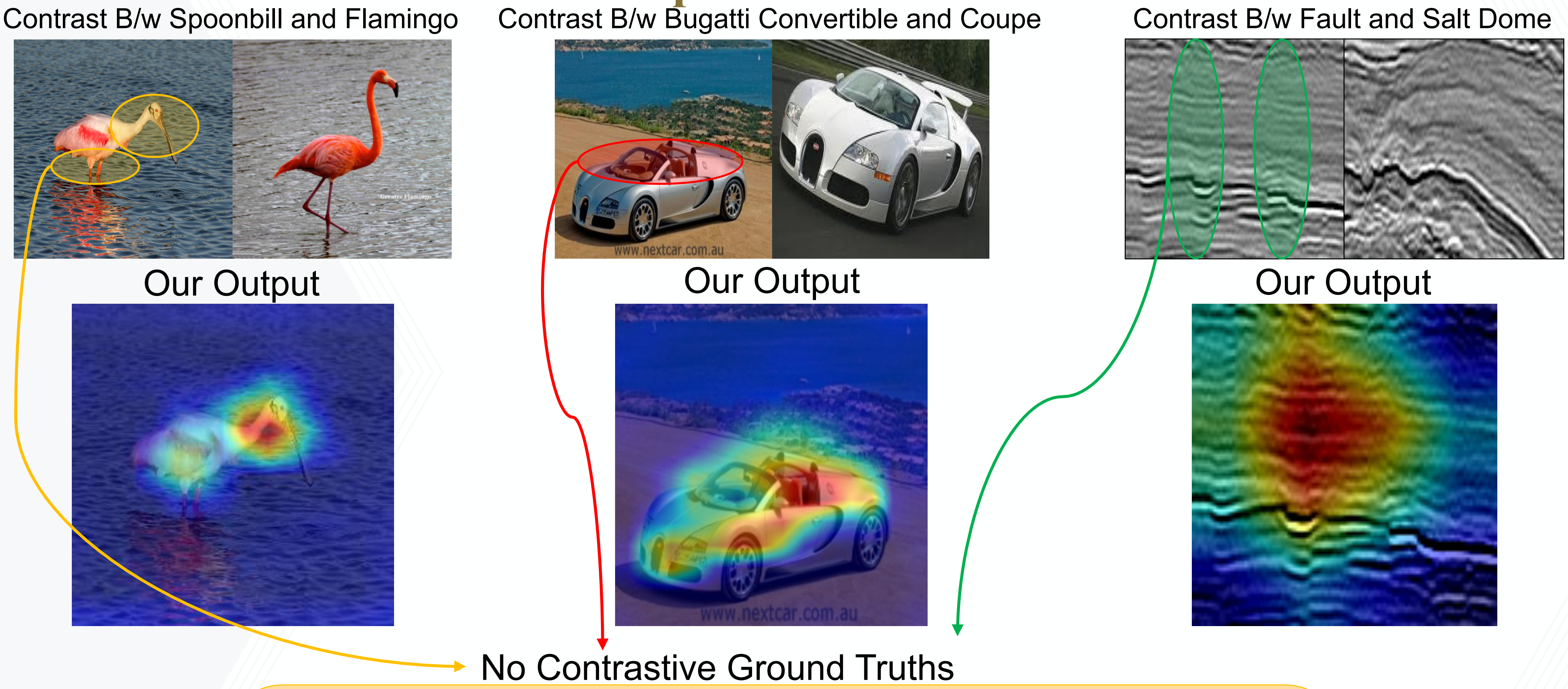
Our Output



No Contrastive Ground Truths

Introduction

Objectives of Contrastive Visual Explanations



Objective:

- Provide structure to existing explanations
- Define contrast from a visual and representational sense
- Extract contrast in an unsupervised fashion

OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

OUTLINE

(1) Part I : Model Uncertainty

(2) Part II : Constrained Model Learning

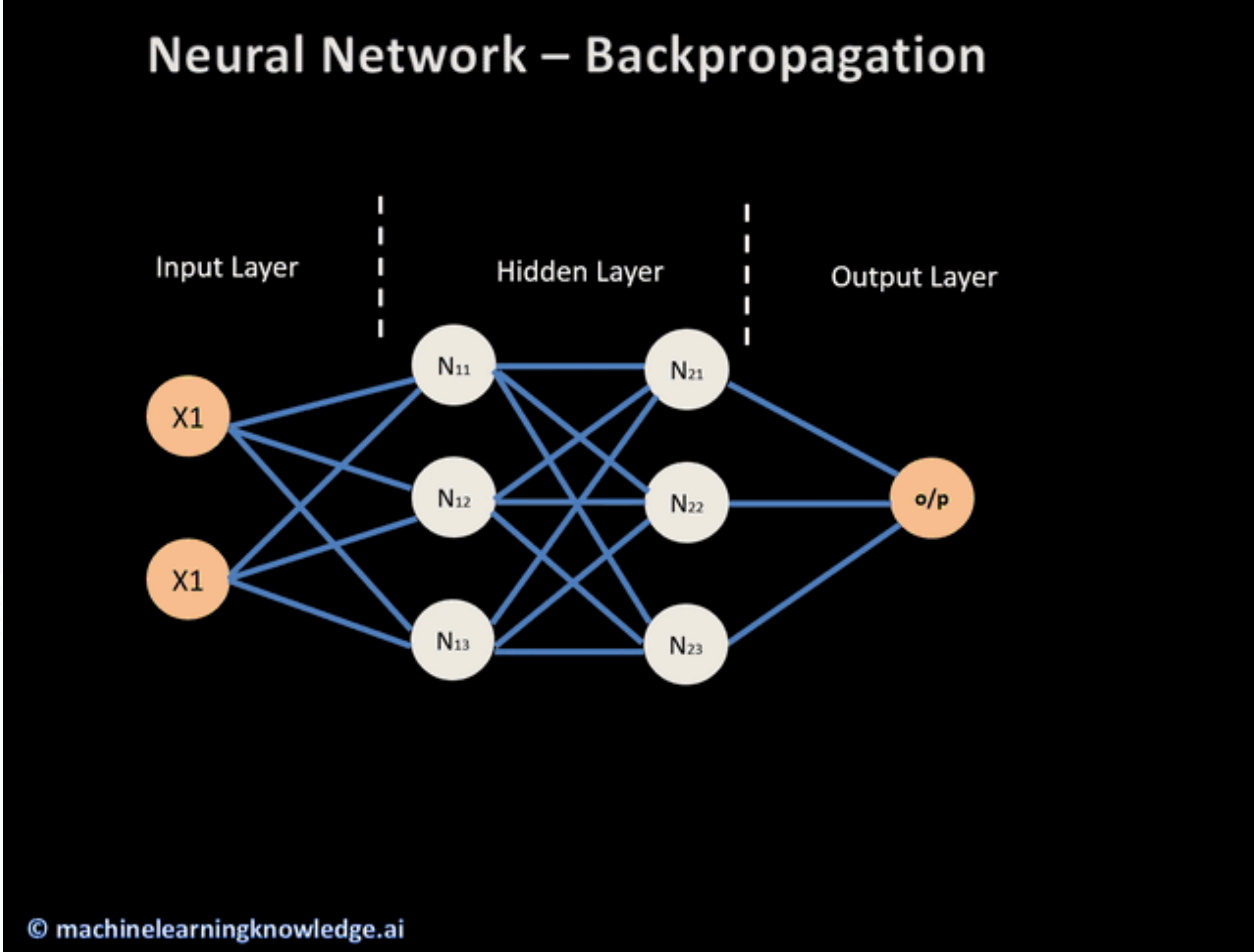
(3) Part III : Reasoning in Neural Networks

(4) Part IV : Explanations in Neural Networks

(5) Part V : Robust Machine Learning

Part I : Model Uncertainty

Basic Operation



Space of Models

Training

- Gradient-based optimization

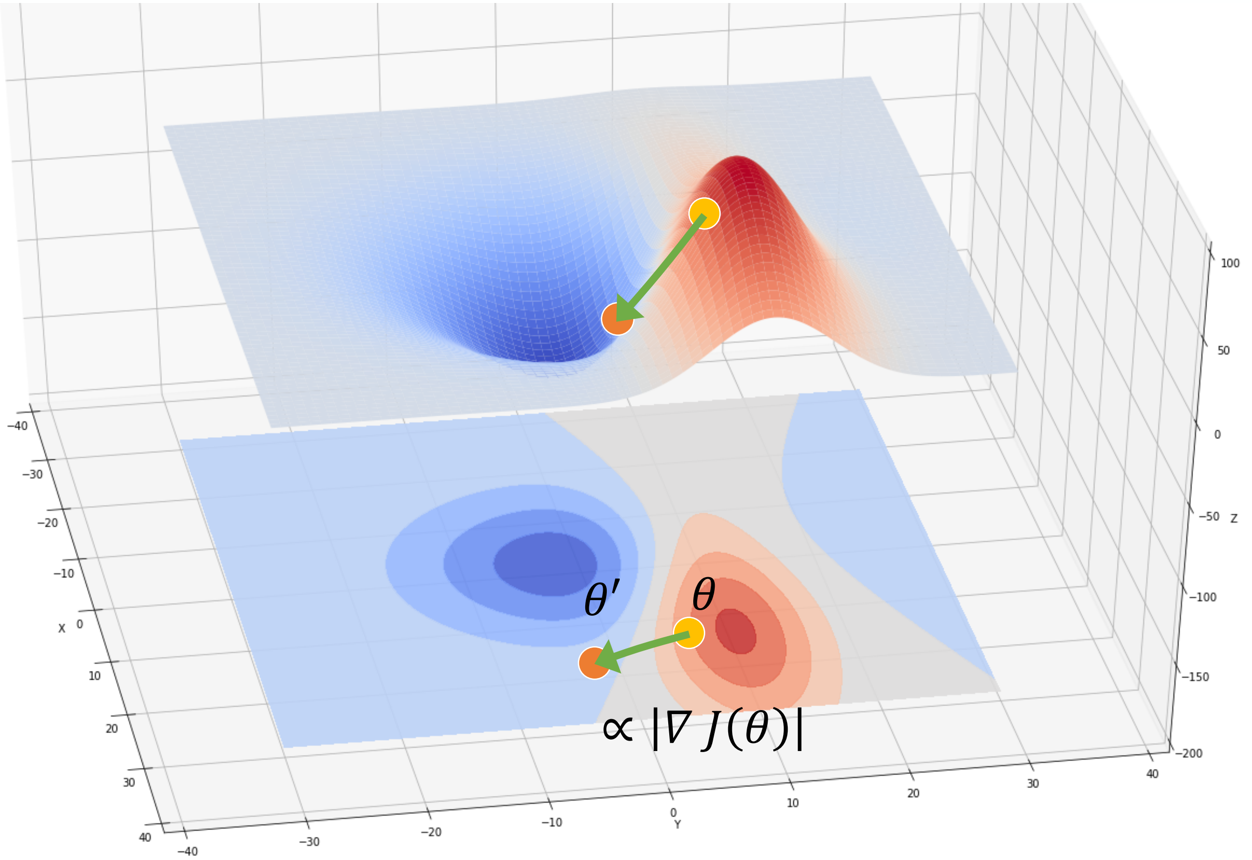
$$\theta' = \theta - \underbrace{\eta \cdot \nabla J(\theta)}$$

The amount of update

= the magnitude of gradient $|\nabla J(\theta)|$ scaled by learning rate η

= the changes in parameterization between old and new models

= the **distance** between old and new model on the space of models



OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

Space of Models

Testing

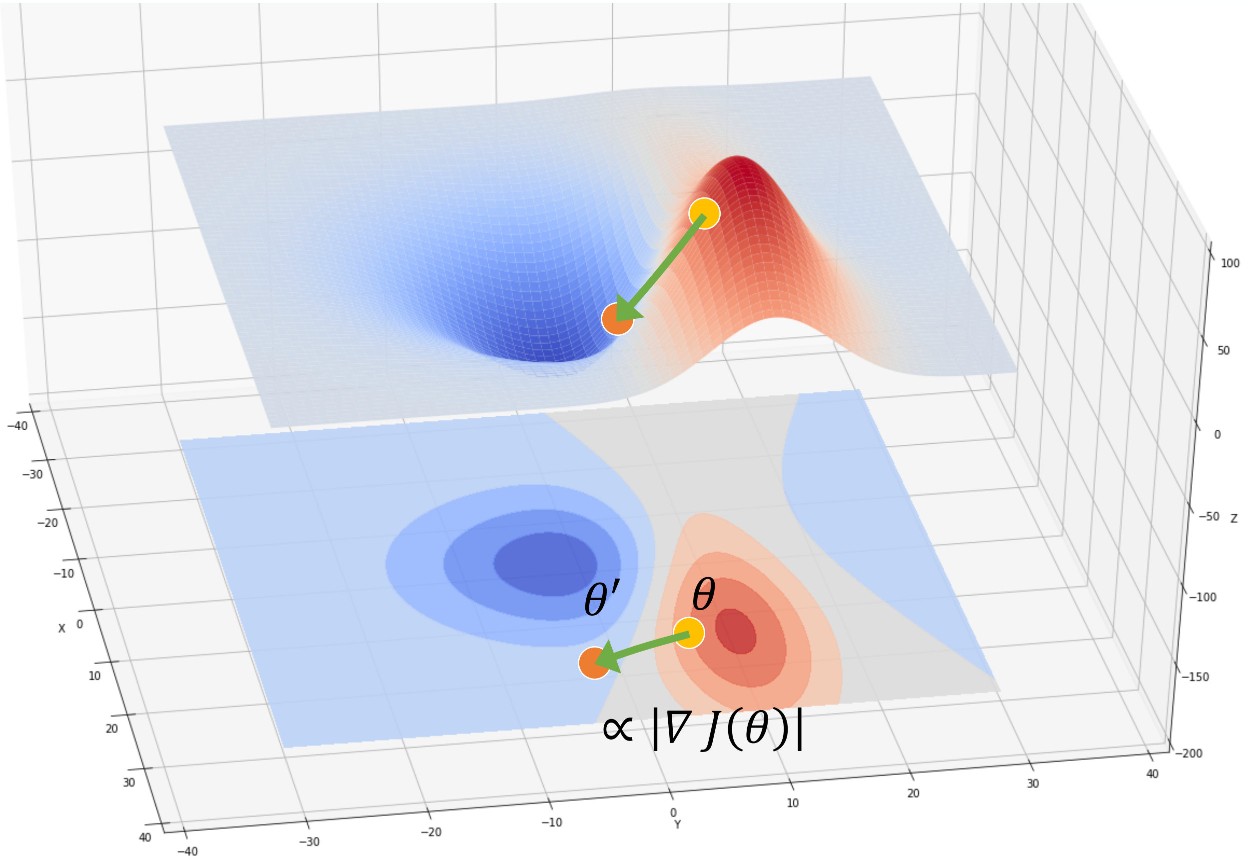
- Compute gradients

$$\nabla J(\boldsymbol{\theta})$$

The magnitude of gradient

= the model update required to represent the given input properly

= the **distance** between the current model and a “better” model for the given input on the space of models



OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

Gradient as a Measure of Uncertainty

Quantifying the uncertainty of neural networks

Model uncertainty: uncertainty in model parameters due to limited data

Small $|\nabla J(\boldsymbol{\theta})|$: Model is certain about the given input

Large $|\nabla J(\boldsymbol{\theta})|$: Model is uncertain about the given input

OLIVES @GeorgiaTech

J. Lee and G. AlRegib, "Gradients as a Measure of Uncertainty in Neural Networks," 2020

Georgia Tech CREATING THE NEXT

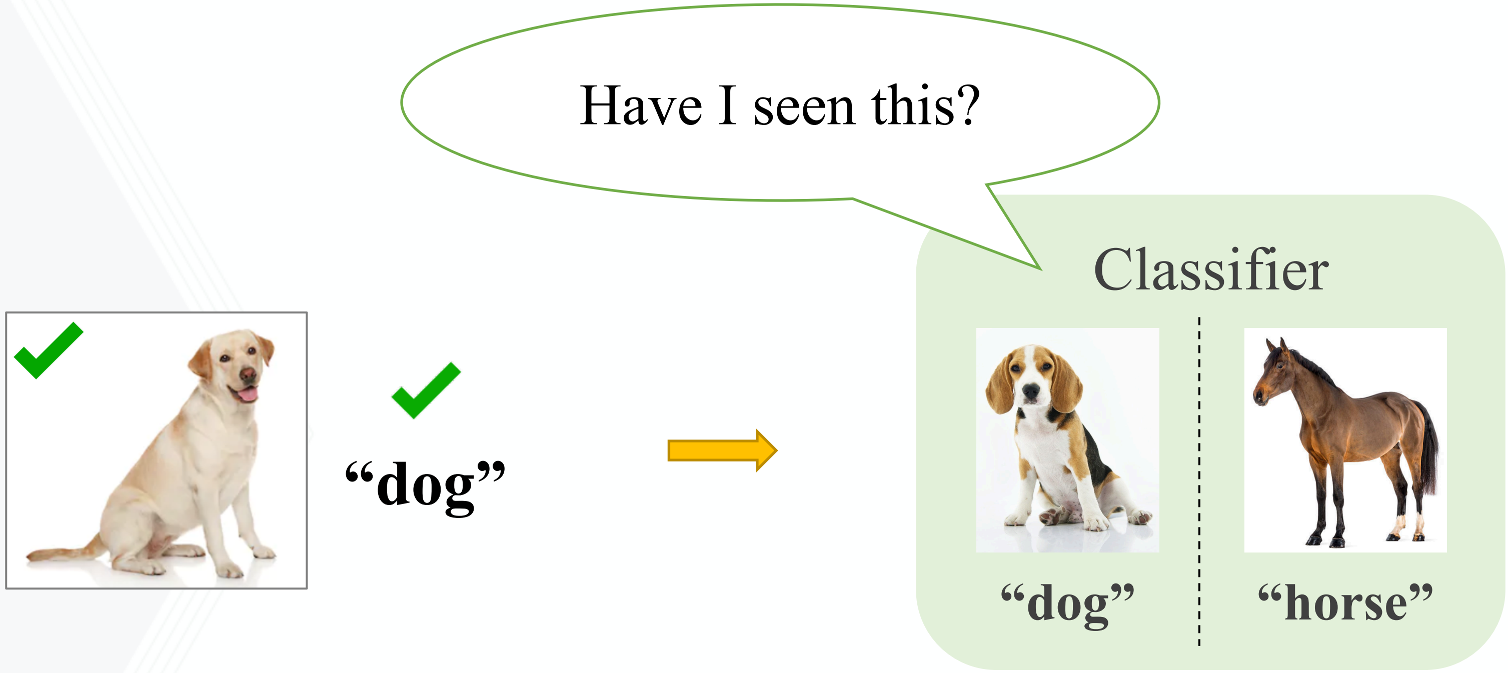
Gradient as a Measure of Uncertainty

Classifier

“dog”

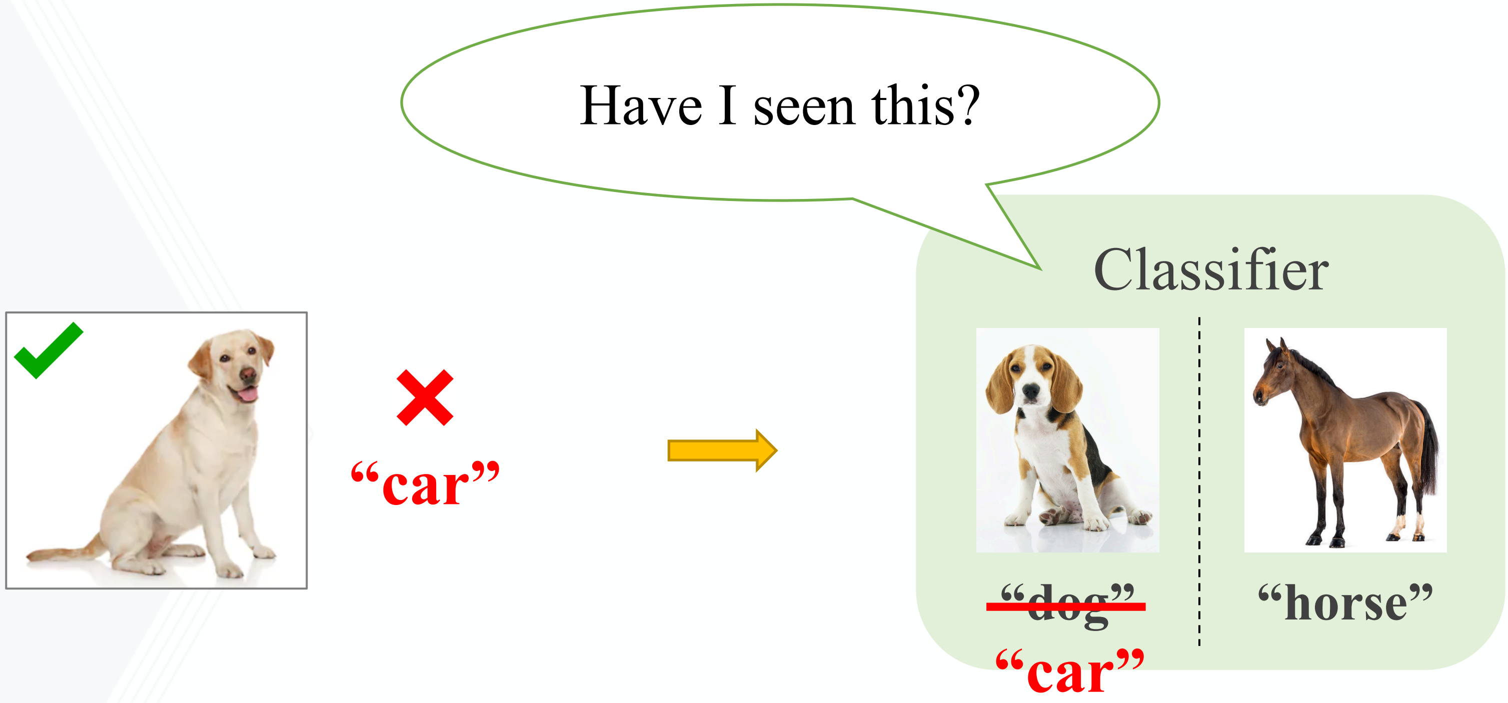
“horse”

Gradient as a Measure of Uncertainty



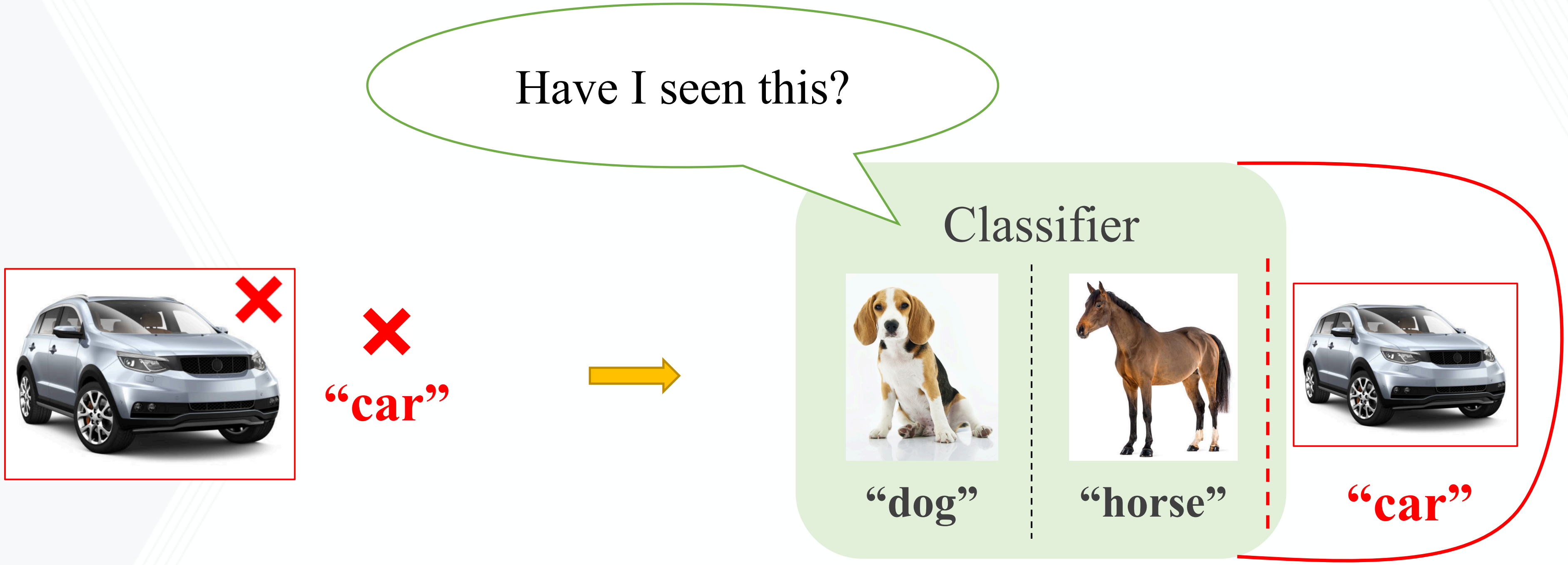
Model associates **learned features** with the **trained label**

Gradient as a Measure of Uncertainty



Required change: associate **learned features** with the **new label**

Gradient as a Measure of Uncertainty



Required change : learn **new features** and associate them with the **new label**

Gradient as a Measure of Uncertainty

Confounding label

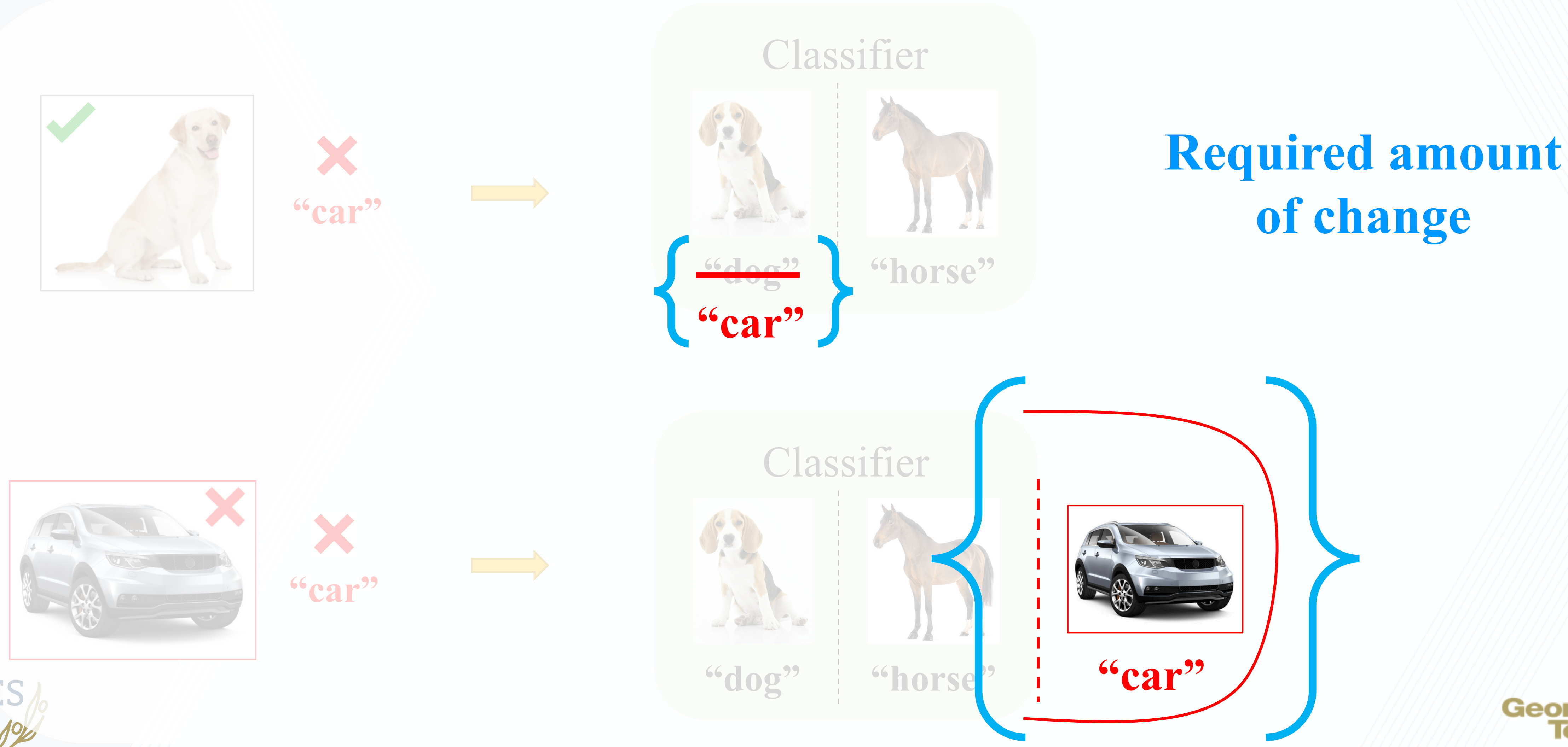
: A label that is different from ordinary labels on which a model is trained

Classifier

"dog" "horse"

"car"

Gradient as a Measure of Uncertainty
Probing Models with Confounding Labels



Hypothesis

It takes **less** amount of change to **associate confounding labels** with **familiar inputs** than unfamiliar inputs

Gradient Generation Framework

Confounding Labels

Input image
$\boldsymbol{x}$

Trained model
$\boldsymbol{M(\theta)}$

Prediction
$\boldsymbol{M(\theta; x)}$

Loss
$\boldsymbol{J(\theta; x, y_c)}$

Confounding label
$\boldsymbol{y_c}$

Backpropagation
$\boldsymbol{\nabla_{\theta} J(\theta; x, y_c)}$

Layer 0 gradients
$\boldsymbol{\nabla_{\theta_0} J(\theta_0;\ x, y_c)}$

$\cdots$

Layer N gradients
$\boldsymbol{\nabla_{\theta_N} J(\theta_N;\ x, y_c)}$

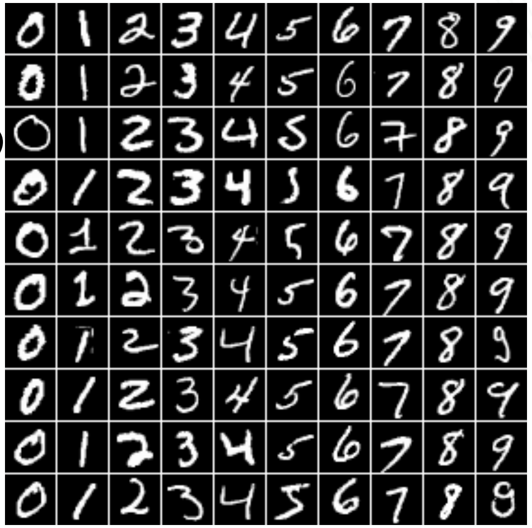
Collection of squared L2 norm
$\boldsymbol{d_{\nabla\theta}}$

$$\left\{ \left\|\nabla_{\theta_0} J(\theta_0;\ x, y_c)\right\|_2^2 ,\ \cdots\ ,\ \left\|\nabla_{\theta_N} J(\theta_N;\ x, y_c)\right\|_2^2 \right\}$$

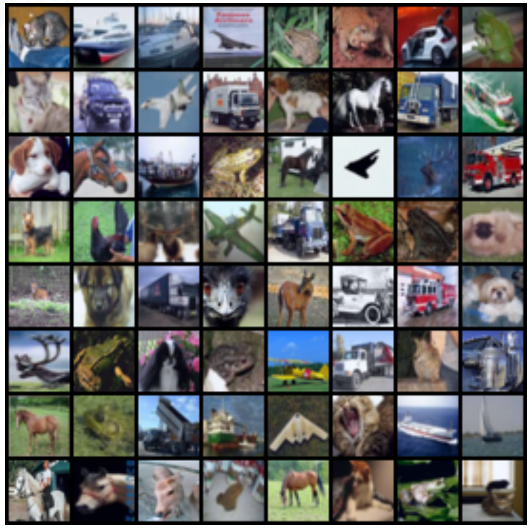
Demonstration

- Compare L_2 norm of gradients at different layers fo various vision datasets
- Network architecture: ResNet18

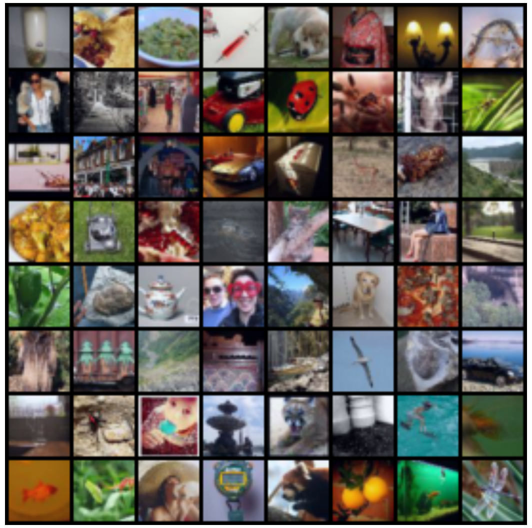
MNIST



SVHN



CIFAR10



TinyImageNet

LSUN

OLIVES @GeorgiaTech

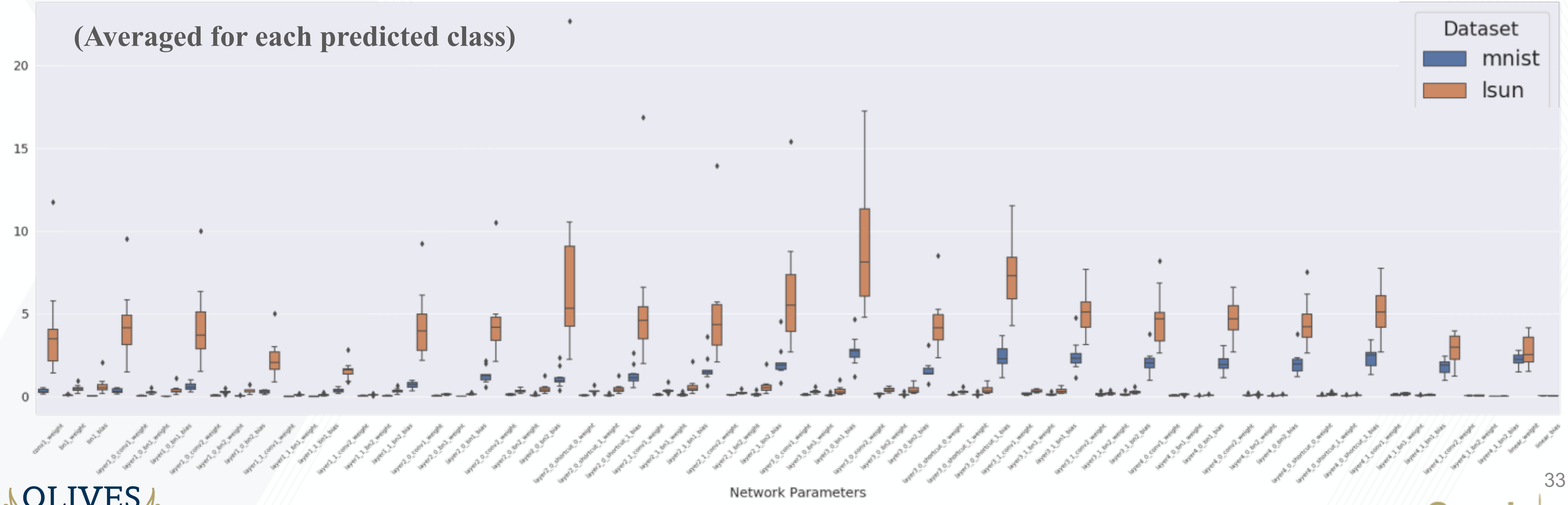
Georgia Tech CREATING THE NEXT

Demonstration

Collection of squared L2 norm $\boldsymbol{d}_{\nabla\theta}$

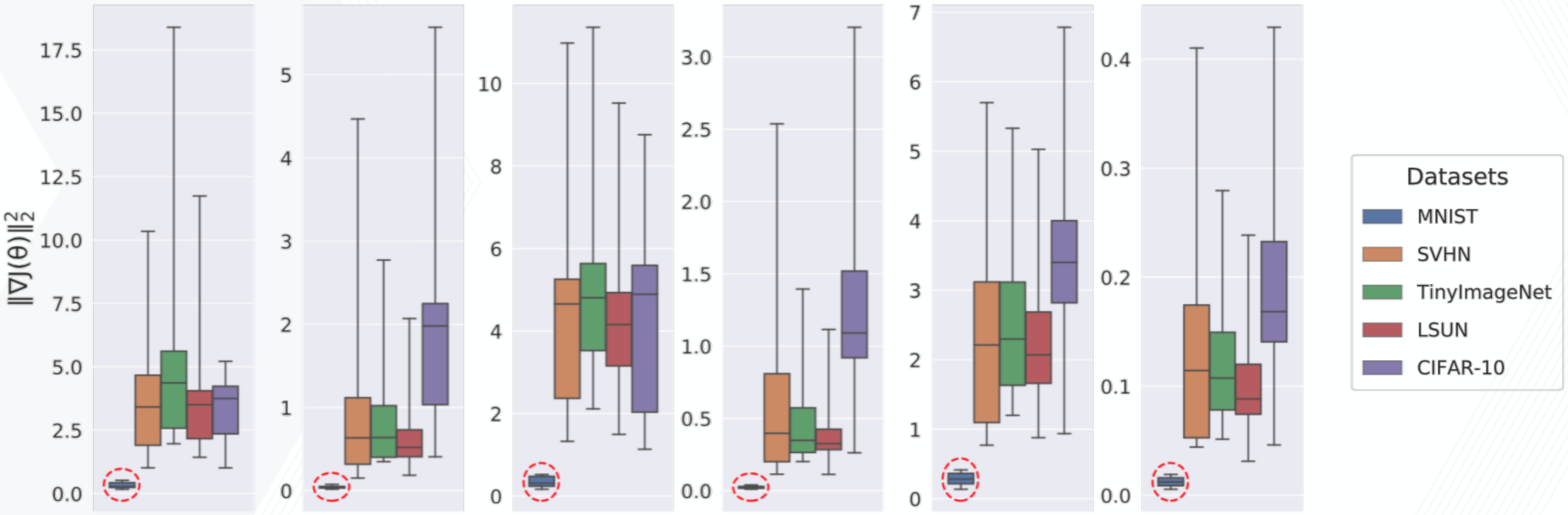
$$\left\{ \left\|\nabla_{\theta_0} J(\theta_0;\, x, y_c)\right\|_2^2 \;,\; \cdots \;,\; \left\|\nabla_{\theta_N} J(\theta_N;\, x, y_c)\right\|_2^2 \right\}$$

(Averaged for each predicted class)



Demonstration

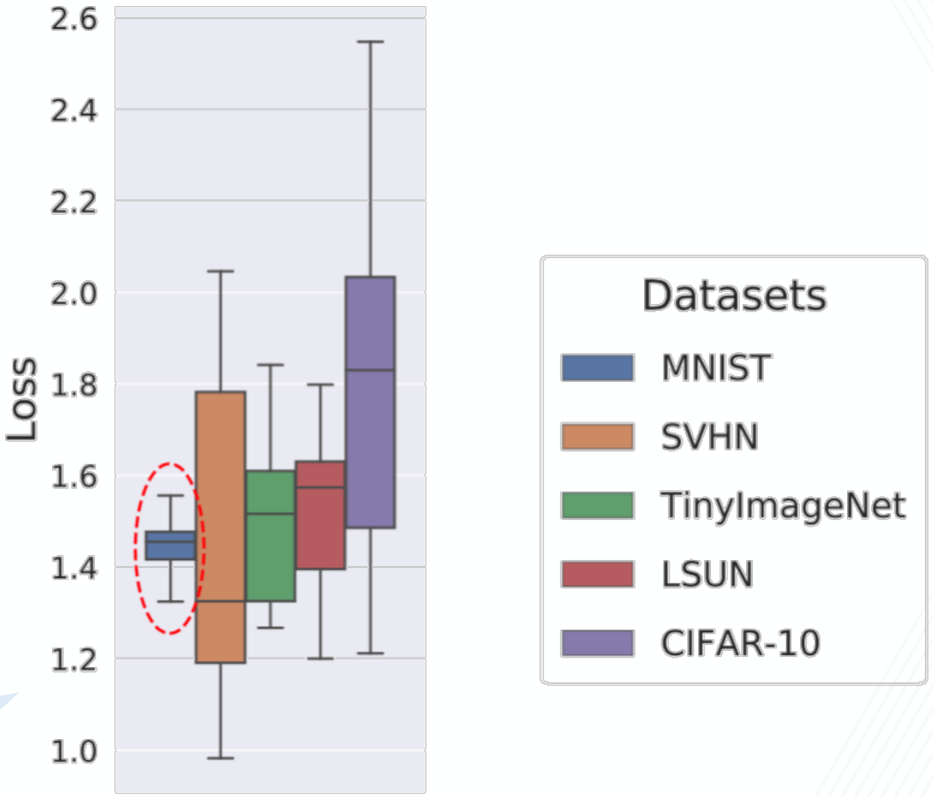
Squared L2 distances for different parameter sets $\left\|\nabla_{\theta_0} J(\theta_0;\, x, y_c)\right\|_2^2$



Demonstration

Why **Gradients** over Loss?

- Higher dimension = more information
- Gradients computed for the current state of each parameter set



Loss does not effectively differentiate the distributions of datasets

Out-of-Distribution Detection

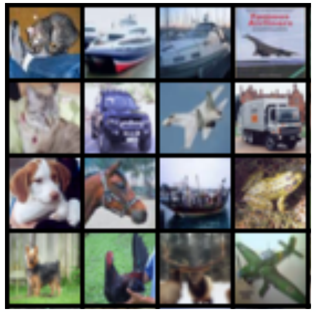
Dataset Distribution		Detection Accuracy	AUROC	AUPR
In	Out	Baseline [5] / ODIN [6] / Mahalanobis (V) [7] / Mahalanobis (P+FE) [7] / Ours		
CIFAR-10	SVHN	83.36 / 88.81 / 79.39 / 91.95 / **98.04**	88.30 / 94.93 / 85.03 / 97.10 / **99.84**	88.26 / 95.45 / 86.15 / 96.12 / **99.98**
	TinyImageNet	84.01 / 85.21 / 83.60 / **97.45** / 86.17	90.06 / 91.86 / 88.93 / **99.68** / 93.18	89.26 / 91.60 / 88.59 / **99.60** / 92.66
	LSUN	87.34 / 88.42 / 85.02 / **98.60** / 98.37	92.79 / 94.48 / 90.11 / **99.86** / **99.86**	92.30 / 94.22 / 89.80 / 99.82 / **99.87**
SVHN	CIFAR-10	79.98 / 80.12 / 74.10 / 88.84 / **97.90**	81.50 / 81.49 / 79.31 / 95.05 / **99.79**	81.01 / 80.95 / 80.83 / 90.25 / **98.11**
	TinyImageNet	81.70 / 81.92 / 79.35 / 96.17 / **97.74**	83.69 / 83.82 / 83.85 / 99.23 / **99.77**	82.54 / 82.60 / 85.50 / **98.17** / 97.93
	LSUN	80.96 / 81.15 / 79.52 / 97.50 / **99.04**	82.85 / 82.98 / 83.02 / 99.54 / **99.93**	81.97 / 82.01 / 84.67 / 98.84 / **99.21**

Out-of-Distribution Detection

Dataset Distribution		Detection Accuracy	AUROC	AUPR
In	Out	Baseline [5] / ODIN [6] / Mahalanobis (V) [7] / Mahalanobis (P+FE) [7] / Ours		
CIFAR-10	SVHN	83.36 / 88.81 / 79.39 / 91.95 / **98.04**	88.30 / 94.93 / 85.03 / 97.10 / **99.84**	88.26 / 95.45 / 86.15 / 96.12 / **99.98**
	TinyImageNet	84.01 / 85.21 / 83.60 / **97.45** / 86.17	90.06 / 91.86 / 88.93 / **99.68** / 93.18	89.26 / 91.60 / 88.59 / **99.60** / 92.66
	LSUN	87.34 / 88.42 / 85.02 / **98.60** / 98.37	92.79 / 94.48 / 90.11 / **99.86** / **99.86**	92.30 / 94.22 / 89.80 / 99.82 / **99.87**
SVHN	CIFAR-10	79.98 / 80.12 / 74.10 / 88.84 / **97.90**	81.50 / 81.49 / 79.31 / 95.05 / **99.79**	81.01 / 80.95 / 80.83 / 90.25 / **98.11**
	TinyImageNet	81.70 / 81.92 / 79.35 / 96.17 / **97.74**	83.69 / 83.82 / 83.85 / 99.23 / **99.77**	82.54 / 82.60 / 85.50 / **98.17** / 97.93
	LSUN	80.96 / 81.15 / 79.52 / 97.50 / **99.04**	82.85 / 82.98 / 83.02 / 99.54 / **99.93**	81.97 / 82.01 / 84.67 / 98.84 / **99.21**

Numbers

SVHN



CIFAR10

TinyImageNet

LSUN

Objects, natural scenes

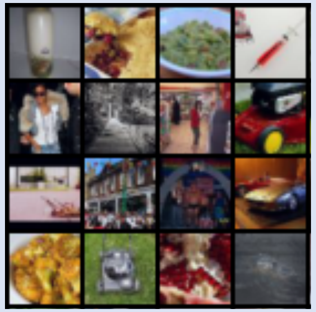
OLIVES @GeorgiaTech
www.ghassanalregib.com

Georgia Tech
CREATING THE NEXT

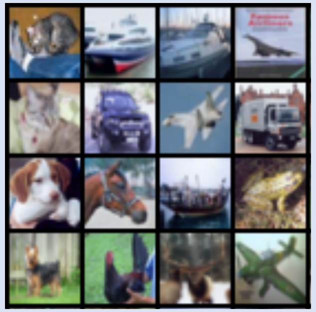
Out-of-Distribution Detection

Dataset Distribution		Detection Accuracy	AUROC	AUPR
In	Out	Baseline [5] / ODIN [6] / Mahalanobis (V) [7] / Mahalanobis (P+FE) [7] / Ours		
CIFAR-10	SVHN	83.36 / 88.81 / 79.39 / 91.95 / **98.04**	88.30 / 94.93 / 85.03 / 97.10 / **99.84**	88.26 / 95.45 / 86.15 / 96.12 / **99.98**
	TinyImageNet	84.01 / 85.21 / 83.60 / **97.45** / 86.17	90.06 / 91.86 / 88.93 / **99.68** / 93.18	89.26 / 91.60 / 88.59 / **99.60** / 92.66
	LSUN	87.34 / 88.42 / 85.02 / **98.60** / 98.37	92.79 / 94.48 / 90.11 / **99.86** / **99.86**	92.30 / 94.22 / 89.80 / 99.82 / **99.87**
SVHN	CIFAR-10	79.98 / 80.12 / 74.10 / 88.84 / **97.90**	81.50 / 81.49 / 79.31 / 95.05 / **99.79**	81.01 / 80.95 / 80.83 / 90.25 / **98.11**
	TinyImageNet	81.70 / 81.92 / 79.35 / 96.17 / **97.74**	83.69 / 83.82 / 83.85 / 99.23 / **99.77**	82.54 / 82.60 / 85.50 / **98.17** / 97.93
	LSUN	80.96 / 81.15 / 79.52 / 97.50 / **99.04**	82.85 / 82.98 / 83.02 / 99.54 / **99.93**	81.97 / 82.01 / 84.67 / 98.84 / **99.21**

More similar datasets (objects)



TinyImageNet



CIFAR10

LSUN

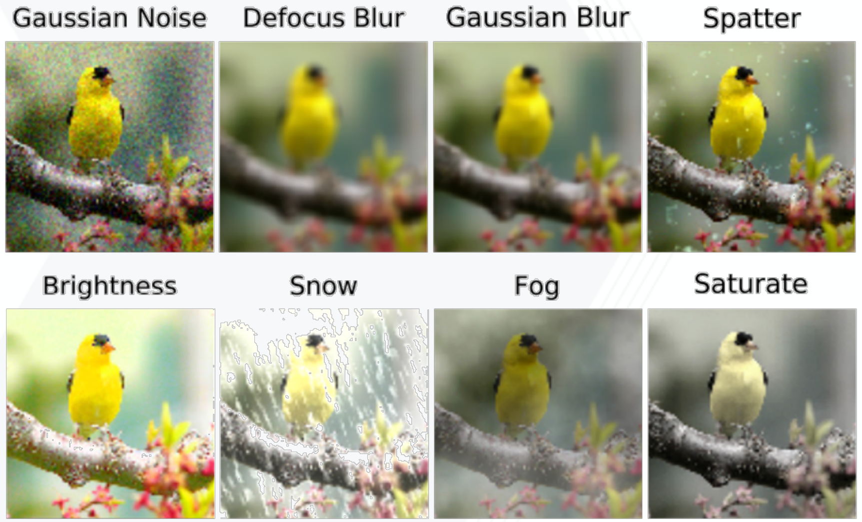
SVHN

OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

Corrupted Input Detection

CIFAR-10-C



CURE-TSR

Corrupted Input Detection

Dataset	Method / Corruption	Mahalanobis [12] / Ours Level 1	Level 2	Level 3	Level 4	Level 5
CIFAR-10-C	Noise	96.63 / **99.95**	98.73 / **99.97**	99.46 / **99.99**	99.62 / **99.97**	99.71 / **99.99**
	LensBlur	94.22 / **99.95**	97.51 / **99.99**	99.26 / **100.0**	99.78 / **100.0**	99.89 / **100.0**
	GaussianBlur	94.19 / **99.94**	99.28 / **100.0**	99.76 / **100.0**	99.86 / **100.0**	99.80 / **100.0**
	DirtyLens	93.37 / **99.94**	95.31 / **99.93**	95.66 / **99.96**	95.37 / **99.92**	97.43 / **99.96**
	Exposure	91.39 / **99.87**	91.00 / **99.85**	90.71 / **99.88**	90.58 / **99.85**	90.68 / **99.87**
	Snow	93.64 / **99.94**	96.50 / **99.94**	94.44 / **99.95**	94.22 / **99.95**	95.25 / **99.92**
	Haze	95.52 / **99.95**	98.35 / **99.99**	99.28 / **100.0**	99.71 / **99.99**	99.94 / **100.0**
	Decolor	93.51 / **99.96**	93.55 / **99.96**	90.30 / **99.82**	89.86 / **99.75**	90.43 / **99.83**
CURE-TSR	Noise	25.46 / **50.20**	47.54 / **63.87**	47.32 / **81.20**	66.19 / **91.16**	83.14 / **94.81**
	LensBlur	48.06 / **72.63**	71.61 / **87.58**	86.59 / **92.56**	92.19 / **93.90**	94.90 / **95.65**
	GaussianBlur	66.44 / **83.07**	77.67 / **86.94**	93.15 / **94.35**	80.78 / **94.51**	**97.36** / 96.53
	DirtyLens	29.78 / **51.21**	29.28 / **59.10**	46.60 / **82.10**	73.36 / **91.87**	98.50 / **98.70**
	Exposure	74.90 / **88.13**	**99.96** / 96.78	**99.99** / 99.26	**100.0** / 99.80	**100.0** / 99.90
	Snow	28.11 / **61.34**	61.28 / **80.52**	89.89 / **91.30**	**99.34** / 96.13	**99.98** / 97.66
	Haze	66.51 / **95.83**	97.86 / **99.50**	**100.0** / 99.95	**100.0** / 99.87	**100.0** / 99.88
	Decolor	48.37 / **62.36**	60.55 / **81.30**	71.73 / **89.93**	87.29 / **95.42**	89.68 / **96.91**

Corrupted Input Detection

Dataset	Method	Mahalanobis [12] / Ours				
	Corruption	Level 1	Level 2	Level 3	Level 4	Level 5
CIFAR-10-C	Noise	96.63 / **99.95**	98.73 / **99.97**	99.46 / **99.99**	99.62 / **99.97**	99.71 / **99.99**
	LensBlur	94.22 / **99.95**	97.51 / **99.99**	99.26 / **100.0**	99.78 / **100.0**	99.89 / **100.0**
	GaussianBlur	94.19 / **99.94**	99.28 / **100.0**	99.76 / **100.0**	99.86 / **100.0**	99.80 / **100.0**
	DirtyLens	93.37 / **99.94**	95.31 / **99.93**	95.66 / **99.96**	95.37 / **99.92**	97.43 / **99.96**
	Exposure	91.39 / **99.87**	91.00 / **99.85**	90.71 / **99.88**	90.58 / **99.85**	90.68 / **99.87**
	Snow	93.64 / **99.94**	96.50 / **99.94**	94.44 / **99.95**	94.22 / **99.95**	95.25 / **99.92**
	Haze	95.52 / **99.95**	98.35 / **99.99**	99.28 / **100.0**	99.71 / **99.99**	99.94 / **100.0**
	Decolor	93.51 / **99.96**	93.55 / **99.96**	90.30 / **99.82**	89.86 / **99.75**	90.43 / **99.83**
CURE-TSR	Noise	25.46 / **50.20**	47.54 / **63.87**	47.32 / **81.20**	66.19 / **91.16**	83.14 / **94.81**
	LensBlur	48.06 / **72.63**	71.61 / **87.58**	86.59 / **92.56**	92.19 / **93.90**	94.90 / **95.65**
	GaussianBlur	66.44 / **83.07**	77.67 / **86.94**	93.15 / **94.35**	80.78 / **94.51**	**97.36** / 96.53
	DirtyLens	29.78 / **51.21**	29.28 / **59.10**	46.60 / **82.10**	73.36 / **91.87**	98.50 / **98.70**
	Exposure	74.90 / **88.13**	**99.96** / 96.78	**99.99** / 99.26	**100.0** / 99.80	**100.0** / 99.90
	Snow	28.11 / **61.34**	61.28 / **80.52**	89.89 / **91.30**	**99.34** / 96.13	**99.98** / 97.66
	Haze	66.51 / **95.83**	97.86 / **99.50**	**100.0** / 99.95	**100.0** / 99.87	**100.0** / 99.88
	Decolor	48.37 / **62.36**	60.55 / **81.30**	71.73 / **89.93**	87.29 / **95.42**	89.68 / **96.91**

So far,

- We introduced an interpretation of **gradients in the space of models** from a perspective of **model uncertainty**
- We presented a framework for efficient gradient generation with **confounding labels** to quantify uncertainty of fully trained networks
- We validated that the gradient-based uncertainty measure outperform activation-based features in **out-of-distribution detection** and **corrupted input detection**

arXiv

https://arxiv.org/abs/2008.08030

https://github.com/olivesgatech

References

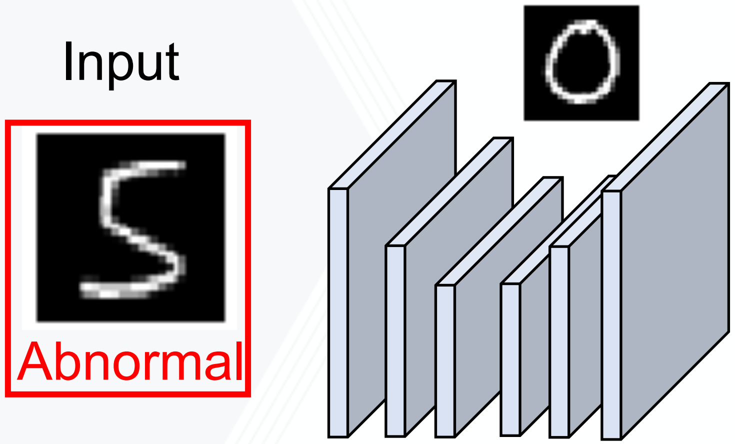
- J. Lee and G. AlRegib, "Gradients as a Measure of Uncertainty in Neural Networks," in *IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, Oct. 2020. [PDF][Video]
- J. Lee, C. Lehman, and G. AlRegib, "Towards Understanding the Purview of Neural Networks via Gradient Analysis, " *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, submitted on Apr. 28 2021.
- J. Lee and G. AlRegib, "Open-Set Recognition with Gradient-Based Representations," in *IEEE International Conference on Image Processing (ICIP)*, Anchorage, AK, Sep. 19-22 2021.
- D. Temel*, J. Lee*, and G. AlRegib, "Object Recognition Under Multifarious Conditions: A Reliability Analysis and a Feature Similarity-Based Performance Estimation," in *IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, Sep. 2019 [PDF][Code]

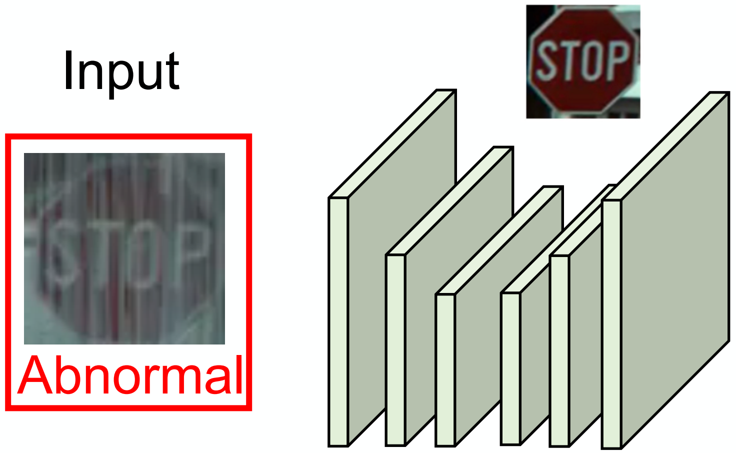
Part II : Model Learning with Gradient-constrained Optimization

Special Case: GradCon - Gradient Constraint

Anomaly Detection

Anomaly: Data whose *classes* or *attributes* differ from training data





Goal: Detect anomalies to ensure the robustness of machine learning algorithm

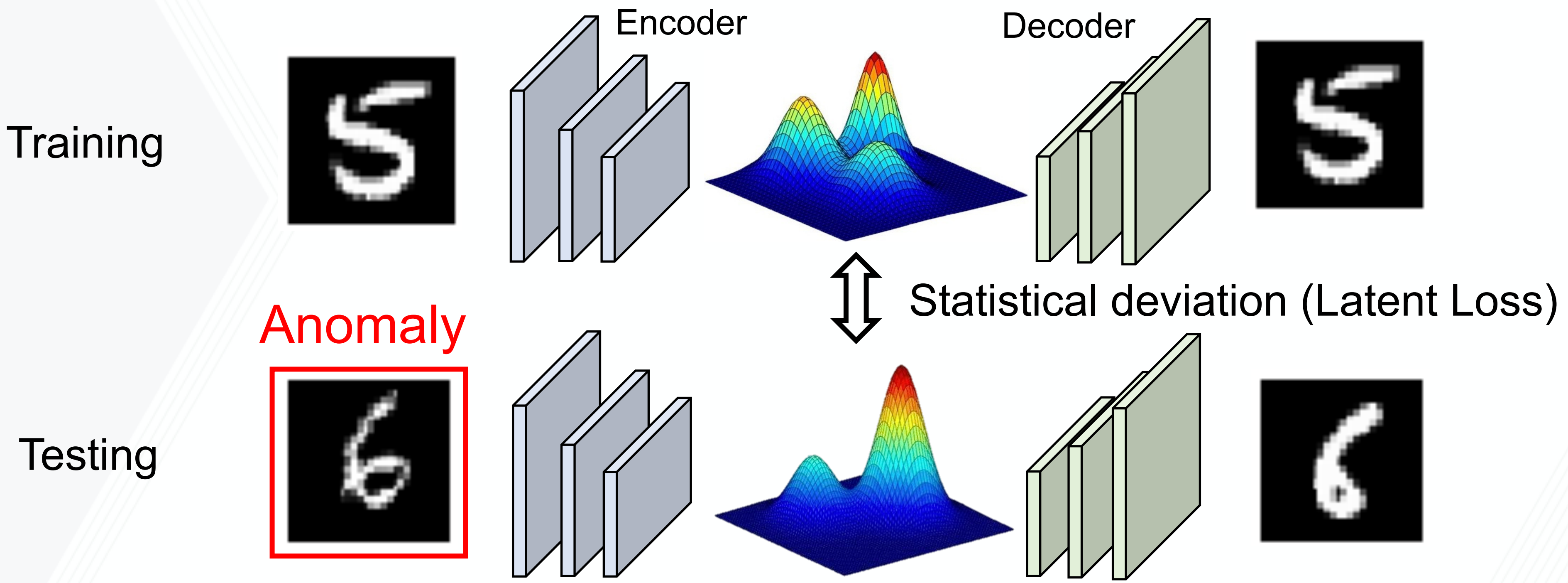
Anomaly Detection

Constrained Representation

2004 — 2016 — 2018 — 2019

Tax et.al [1] — Fan et.al [2] — Pidhorksyi et.al [3] — Abati et.al [4]

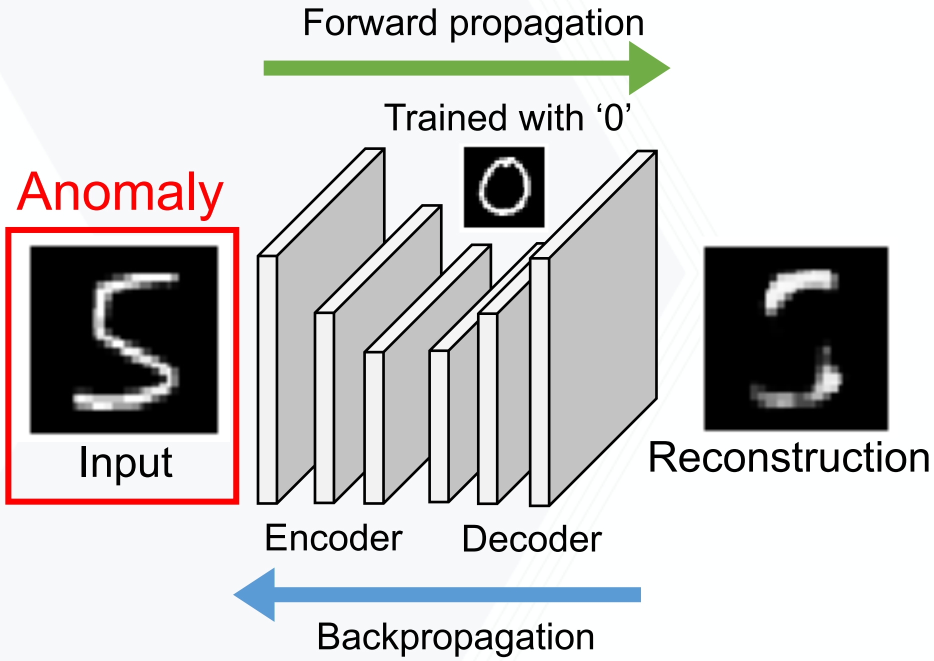
Some quantity is constrained to identify anomalies



[1] David MJ Tax and Robert PW Duin. Support vector data description. Machine learning, 54(1):45–66, 2004.
[2] Yaxiang Fan, Gongjian Wen, Deren Li, Shaohua Qiu, and Martin D Levine. Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder. arXiv preprint arXiv:1805.11223, 2018. 1, 2
[3] S. Pidhorskyi, R. Almohsen, and G. Doretto, "Generative probabilistic novelty detection with adversarial autoencoders," in Advances in Neural Information Processing Systems, 2018, pp. 6822–6833.
[4] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 481–490.

Overview

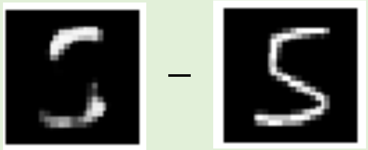
Gradient-based Representation



<u>Existing approaches</u>

Activation-based representation (Data perspective)

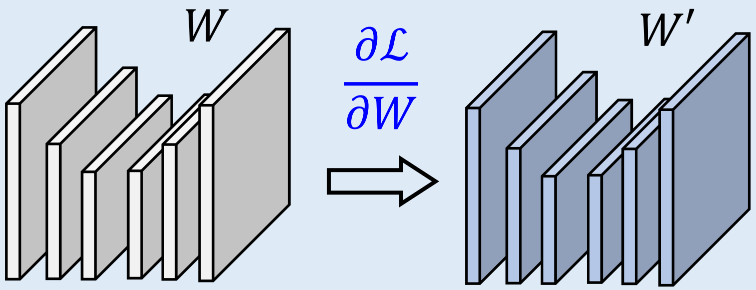
e.g. Reconstruction error ($\mathcal{L}$)



How much of the input does not correspond to the learned information?

<u>Proposed approach</u>

Gradient-based Representation (Model perspective)



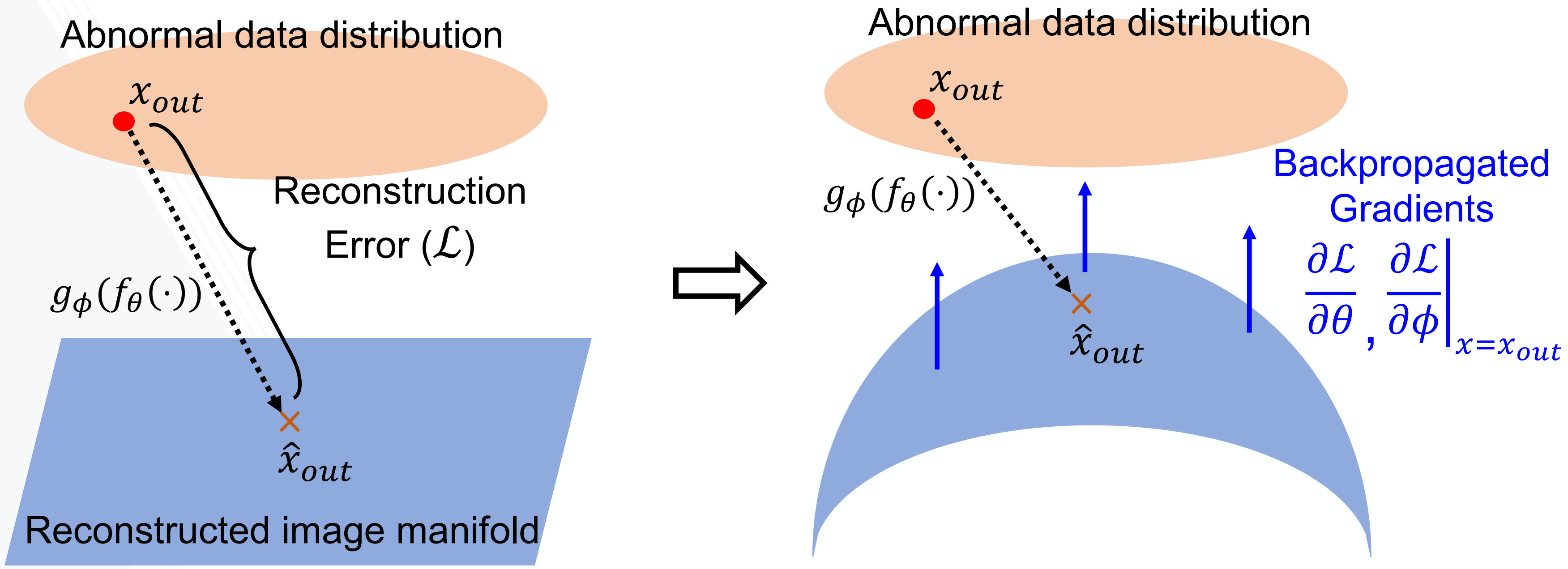
How much **model update** is required by the input?

G. Kwon, M. Prabhushankar, D. Temel, and G. AlRegib, "Novelty Detection Through Model-Based Characterization of Neural Networks," 2020

G. Kwon, M. Prabhushankar, D. Temel, and G. AlRegib, "Backpropagated Gradient Representations for Anomaly Detection," 2020

Geometric Interpretation

Advantages of Gradient-based Representations

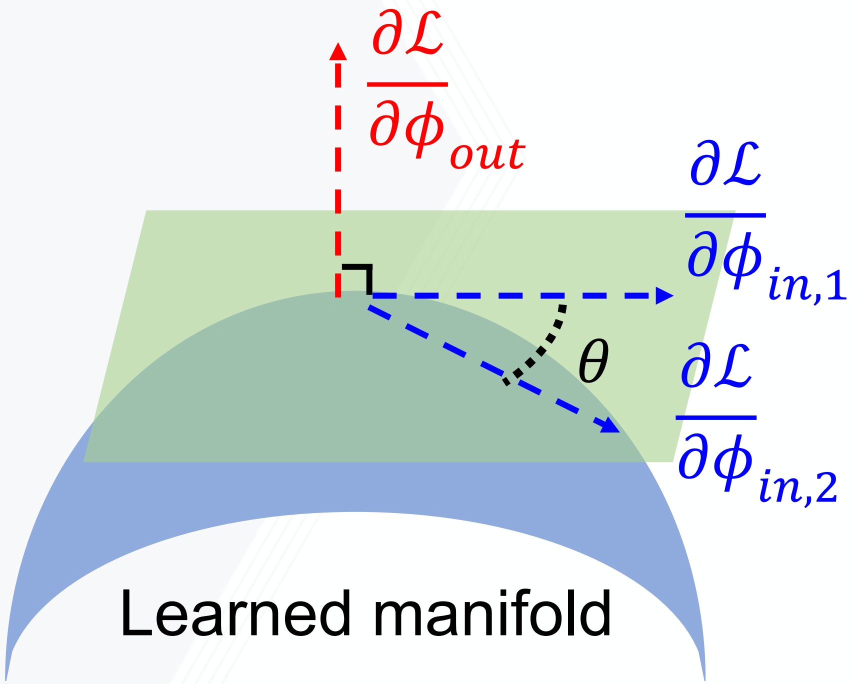


1) Provide directional information to characterize anomalies

2) Gradients from different layers capture abnormality at different levels of data abstraction

GradCon: Gradient Constraint

Constrain gradient-based representations during training to obtain clear separation between normal data and abnormal data



ϕ: Weights $\mathcal{L}$: Reconstruction error

At k-th step of training,

$$J = \mathcal{L} - \underbrace{\mathbb{E}_i \left[\mathrm{cosSIM}\left({\frac{\partial J}{\partial \phi_{i_{avg}}}}^{k-1}, {\frac{\partial \mathcal{L}}{\partial \phi_i}}^{k} \right) \right]}_{\text{Gradient loss}}$$

Avg. training gradients until (k-1) th iter.

Gradients at k-th iter.

where $${\frac{\partial J}{\partial \phi_{i_{avg}}}}^{k-1} = \sum_{t=1}^{k-1} {\frac{\partial J}{\partial \phi_i}}^{t}$$

Baseline Experiment

Activation vs. Gradients

Abnormal "class" detection (CIFAR-10)

e.g.

Normal Abnormal

AUROC Results

Model	Loss	Plane	Car	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck	Average
CAE	Recon	0.682	0.353	0.638	0.587	0.669	**0.613**	0.495	0.498	0.711	0.390	0.564
CAE + Grad	Recon	0.659	0.356	**0.640**	0.555	0.695	0.554	0.549	0.478	0.695	0.357	0.554
	Grad	**0.752**	0.619	0.622	0.580	0.705	0.591	0.683	**0.576**	**0.774**	**0.709**	**0.661**
VAE	Recon	0.553	0.608	0.437	0.546	0.393	0.531	0.489	0.515	0.552	0.631	0.526
	Latent	0.634	0.442	**0.640**	0.497	**0.743**	0.515	**0.745**	0.527	0.674	0.416	0.583
VAE + Grad	Recon	0.556	0.606	0.438	0.548	0.392	0.543	0.496	0.518	0.552	0.631	0.528
	Latent	0.586	0.396	0.618	0.476	0.719	0.474	0.698	0.537	0.586	0.413	0.550
	Grad	0.736	**0.625**	0.591	**0.596**	0.707	0.570	0.740	0.543	0.738	0.629	0.647

Recon: Reconstruction error, Latent: Latent loss, Grad: Gradient loss

1) (CAE vs. CAE + Grad) Effectiveness of the gradient constraint
2) (CAE vs. VAE) Performance sacrifice from the latent constraint
3) (VAE vs. VAE + Grad) Complementary features from the gradient constraint

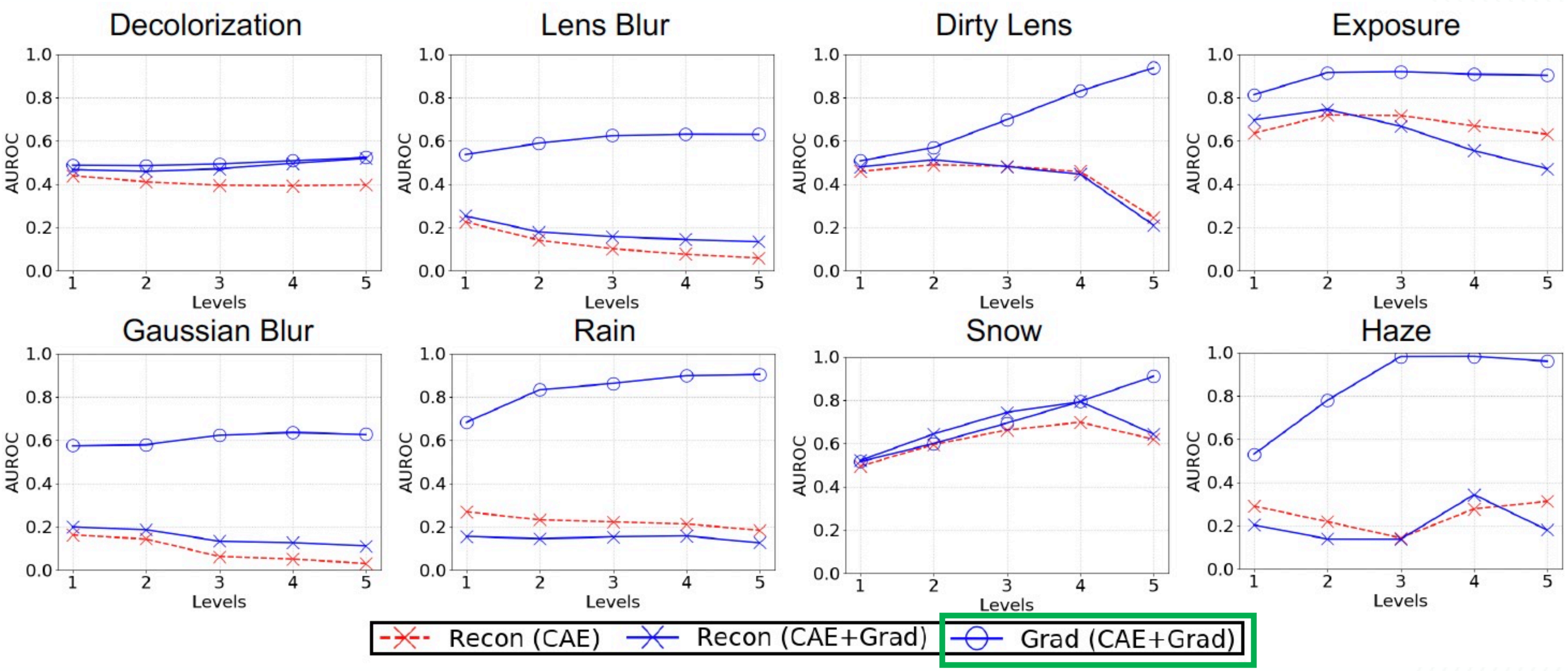
Baseline Experiment
Abnormal Condition detection

Abnormal "condition" detection (CURE-TSR)

Normal

Abnormal

AUROC Results



Recon: Reconstruction error, Grad: Gradient loss

State-of-The-Art Algorithms
CIFAR-10, MNIST, Fashion MNIST

AUROC results in CIFAR-10

	Plane	Car	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck	Average
OCSVM [34]	0.630	0.440	0.649	0.487	0.735	0.500	0.725	0.533	0.649	0.508	0.586
KDE [4]	0.658	0.520	**0.657**	0.497	0.727	0.496	**0.758**	0.564	0.680	0.540	0.610
DAE [9]	0.411	0.478	0.616	0.562	0.728	0.513	0.688	0.497	0.487	0.378	0.536
VAE [12]	0.634	0.442	0.640	0.497	**0.743**	0.515	0.745	0.527	0.674	0.416	0.583
PixelCNN [20]	**0.788**	0.428	0.617	0.574	0.511	0.571	0.422	0.454	0.715	0.426	0.551
LSA [1]	0.735	0.580	**0.690**	0.542	**0.761**	0.546	**0.751**	0.535	0.717	0.548	0.641
AnoGAN [33]	0.671	0.547	0.529	0.545	0.651	0.603	0.585	**0.625**	0.758	0.665	0.618
DSVDD [27]	0.617	**0.659**	0.508	**0.591**	0.609	**0.657**	0.677	**0.673**	0.759	**0.731**	0.648
OCGAN [22]	0.757	0.531	0.640	**0.620**	0.723	**0.620**	0.723	0.575	**0.820**	0.554	**0.657**
GradCon	**0.760**	**0.598**	0.648	0.586	0.733	0.603	0.684	0.567	**0.784**	**0.678**	**0.664**

AUROC results in MNIST

	0	1	2	3	4	5	6	7	8	9	Average
OCSVM [34]	0.988	**0.999**	0.902	0.950	0.955	0.968	0.978	0.965	0.853	0.955	0.951
KDE [4]	0.885	0.996	0.710	0.693	0.844	0.776	0.861	0.884	0.669	0.825	0.814
DAE [9]	0.894	**0.999**	0.792	0.851	0.888	0.819	0.944	0.922	0.740	0.917	0.877
VAE [12]	**0.997**	**0.999**	0.936	0.959	**0.973**	0.964	**0.993**	0.976	0.923	**0.976**	0.970
PixelCNN [20]	0.531	0.995	0.476	0.517	0.739	0.542	0.592	0.789	0.340	0.662	0.618
LSA [1]	0.993	**0.999**	**0.959**	**0.966**	0.956	0.964	**0.994**	**0.980**	**0.953**	**0.981**	**0.975**
AnoGAN [33]	0.966	0.992	0.850	0.887	0.894	0.883	0.947	0.935	0.849	0.924	0.913
DSVDD [27]	0.980	**0.997**	0.917	0.919	0.949	0.885	0.983	0.946	**0.939**	0.965	0.948
OCGAN [22]	**0.998**	**0.999**	0.942	0.963	**0.975**	**0.980**	0.991	**0.981**	**0.939**	**0.981**	**0.975**
GradCon	0.995	**0.999**	**0.952**	**0.973**	0.969	**0.977**	**0.994**	0.979	0.919	0.973	**0.973**

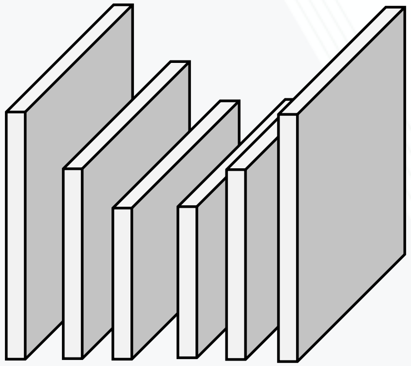
Fashion-MNIST

	% of outlier	10	20	30	40	50
F1	GPND	**0.968**	**0.945**	0.917	0.891	0.864
	Grad	0.964	0.939	0.917	0.899	0.870
	GradCon	0.967	**0.945**	**0.924**	**0.905**	**0.871**
AUC	GPND	0.928	0.932	0.933	0.933	0.933
	Grad	0.931	0.925	0.926	0.928	0.926
	GradCon	**0.938**	**0.933**	**0.935**	**0.936**	**0.934**

Computational Efficiency

Inference Time, Model Parameters

GradCon



Covolutional autoencoder

Does not require

- ❌ Adversarial training
- ❌ Autoregressive models

↓ Model parameters

Computations

Average inference time per image for GradCon (3.08*ms*) is 1.9 times faster than GPND[1] (5.72*ms*)

Method	# of parameters
AnoGAN	6,338,176
GPND	6,766,243
LSA	13,690,160
GradCon	**230,721**

→ Model parameters are at least 27 time fewer

So far,

- GradCon, achieves state-of-the-art performance with significantly fewer number of model parameters

Code

https://github.com/olivesgatech/gradcon-anomaly

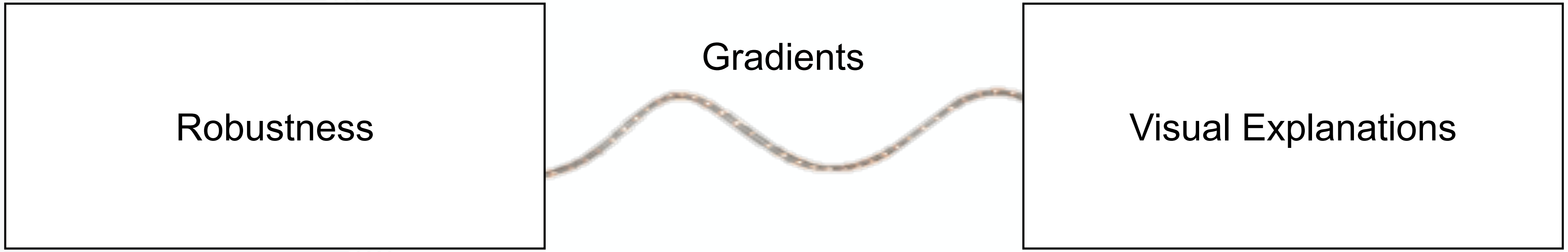
OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

References

- G. Kwon et al., "Backpropagated Gradient Representations for Anomaly Detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, SEC, Glasgow, Aug. 23-28 2020. [PDF][Code][Short Video]
- G. Kwon et al., "Novelty Detection Through Model-Based Characterization of Neural Networks," in *IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, Oct. 2020. [PDF][Code][Video]
- G. Kwon et al., "Distorted Representation Space Characterization Through Backpropagated Gradients," in *IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, Sep. 2019. [PDF][Code]
- G. Kwon and G. AlRegib, "A Gating Model for Bias Calibration in Generalized Zero-Shot Learning, " *IEEE Transactions on Image Processing (*
- D. Temel et al.,, "Traffic Sign Detection Under Challenging Conditions: A Deeper Look Into Performance Variations and Spectral Characteristics," in *IEEE Transactions on Intelligent Transportation Systems*, Jul. 2019. [PDF][Code]*TIP)*, submitted on Feb. 4 2021.
- D. Temel and G. AlRegib, "Perceptual Image Quality Assessment Through Spectral Analysis of Error Representations," in *Signal Processing: Image Communication*, vol. 70, pp. 37-46, 2019. [PDF][Code]
- D. Temel and G. AlRegib, "Traffic Signs in the Wild: Highlights From the IEEE Video and Image Processing Cup 2017 Student Competition [SP Competitions]," in *IEEE Signal Processing Magazine*, vol. 35, no. 2, pp. 154-161, Mar. 2018. [PDF]
- D. Temel et al., "CURE-OR: Challenging Unreal and Real Environments for Object Recognition," in *IEEE International Conference on Machine Learning and Applications (ICMLA)*, Orlando, FL, Dec. 2018 [PDF][Code]

So far,

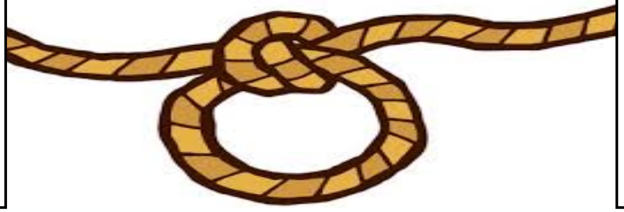


Part III : Reasoning in Neural Networks

From now,

Robustness

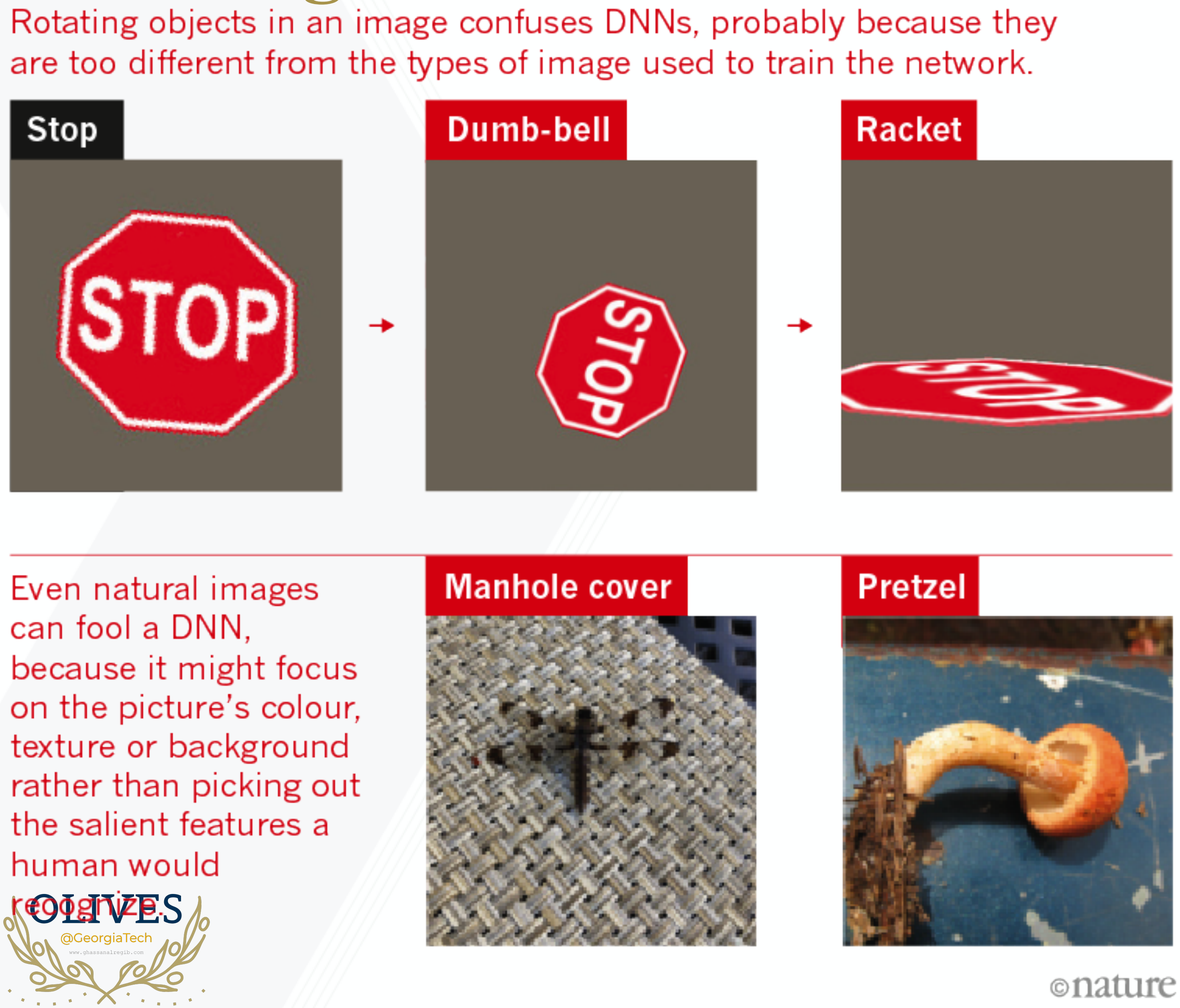
Concept : Reasoning



Method : Gradients

Visual Explanations

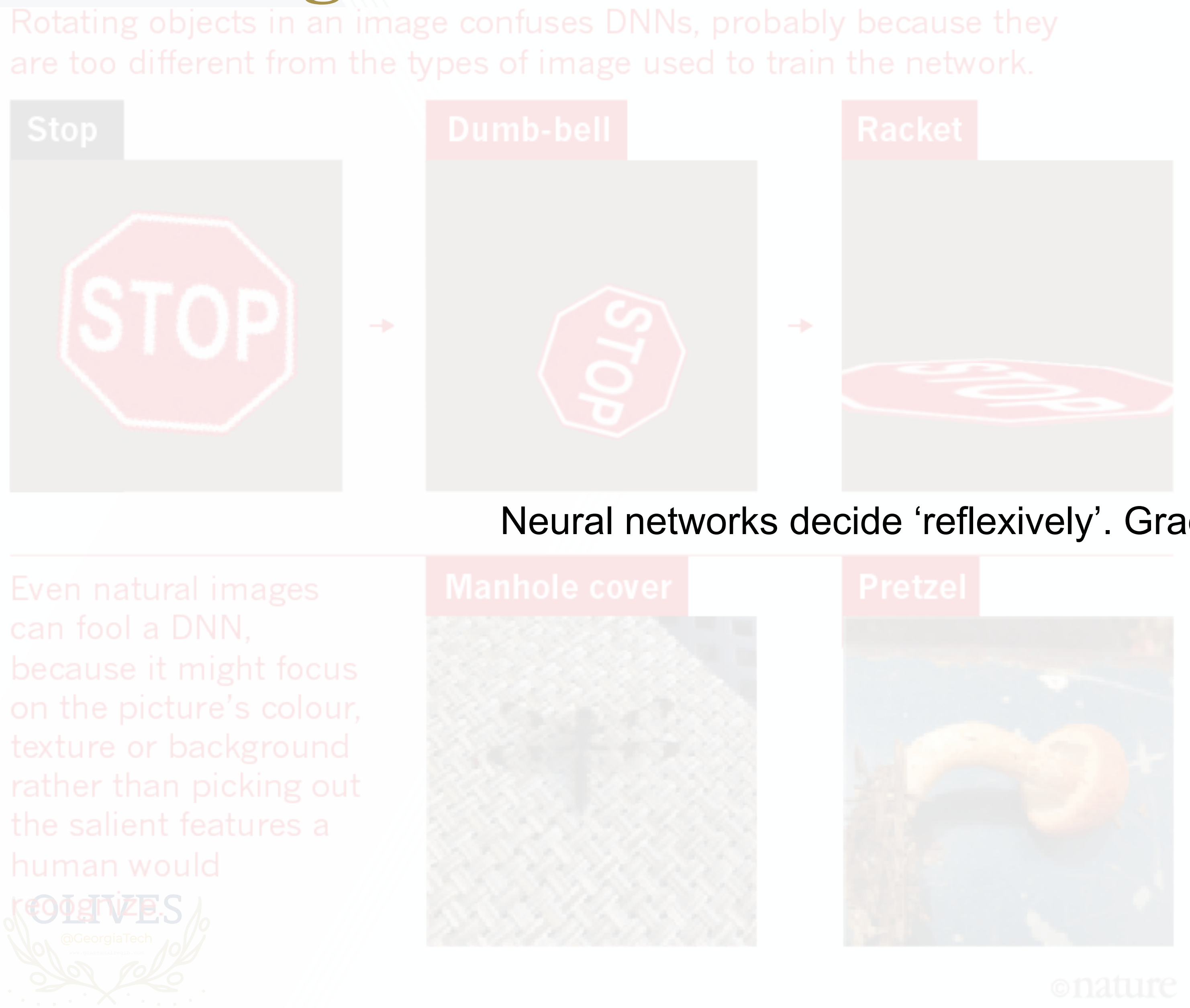
Challenges in Neural Networks

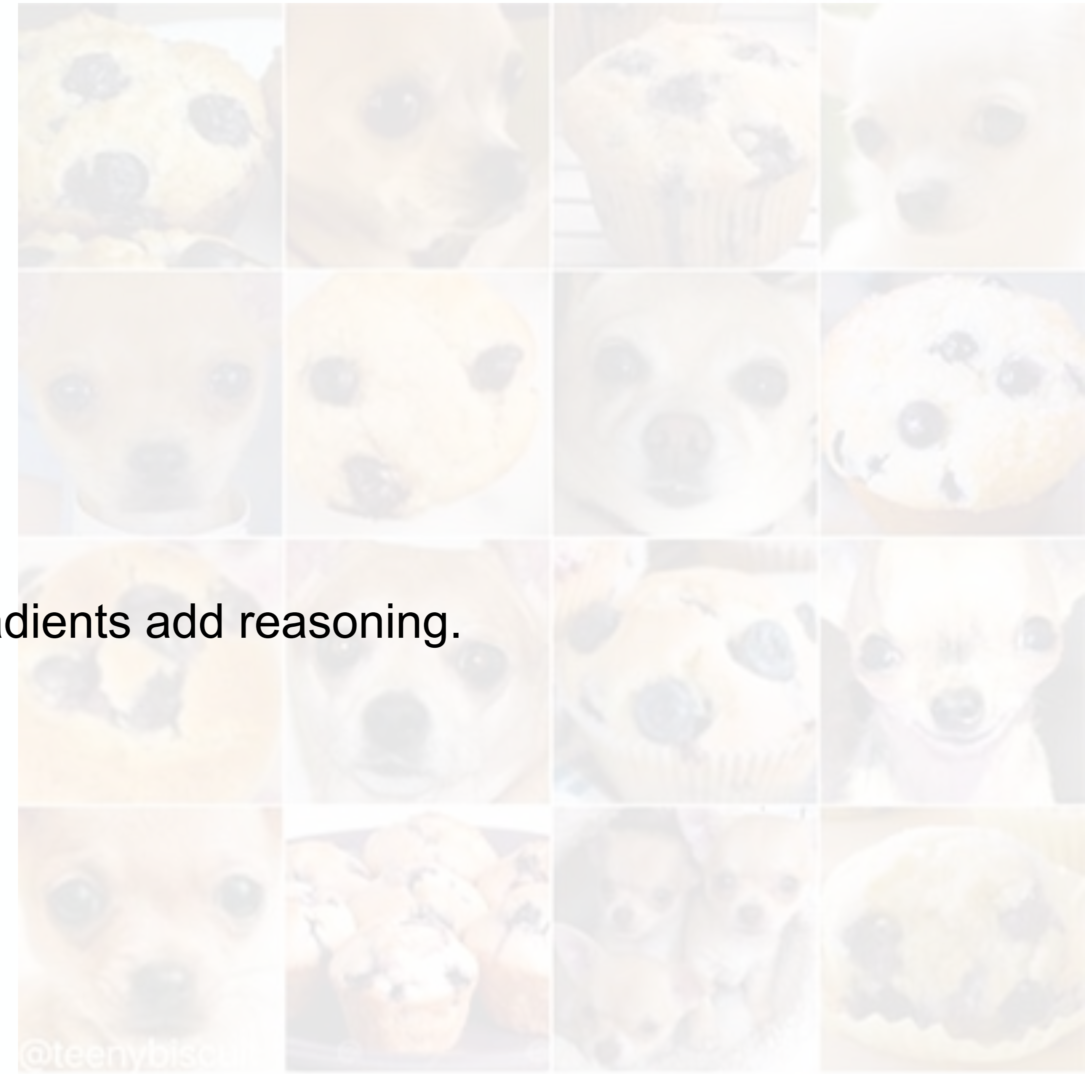


OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

Challenges in Neural Networks



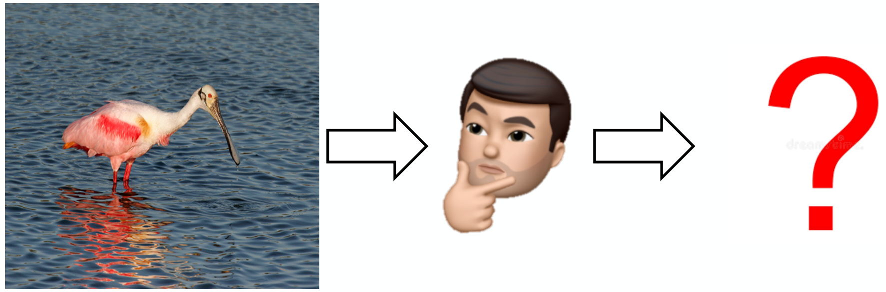


Neural networks decide ‘reflexively’. Gradients add reasoning.

Reasoning
Types of Reasoning

Classwork – Learned differences between Flamingo and Spoonbill.
Exams– To identify unknown bird

What species of bird is this?



Reasoning
Types of Reasoning

Classwork – Learned differences between Flamingo and Spoonbill.
Exams – To identify unknown bird

Learning

Spoonbill

Spoonbill :
Pink and round body, Straight neck

Flamingo

Flamingo :
Pink and round body, S-shaped neck

Reasoning
Types of Reasoning

Classwork – Learned differences between Flamingo and Spoonbill.
Exams – To identify unknown bird

Learning

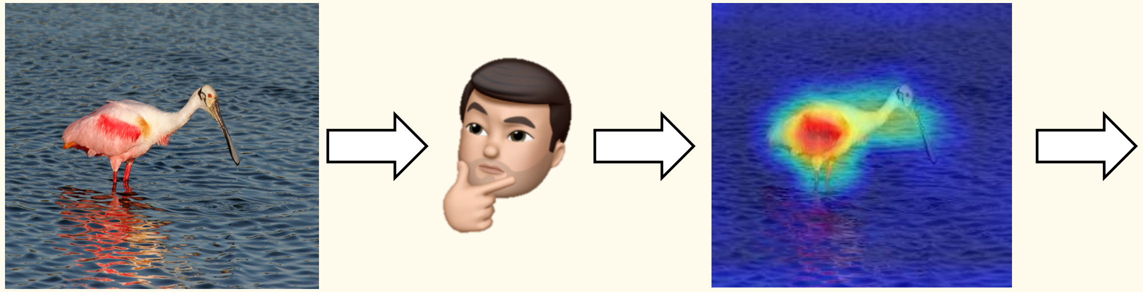
Spoonbill

Flamingo

Spoonbill :
Pink and round body, Straight neck

Flamingo :
Pink and round body, S-shaped neck

Exams

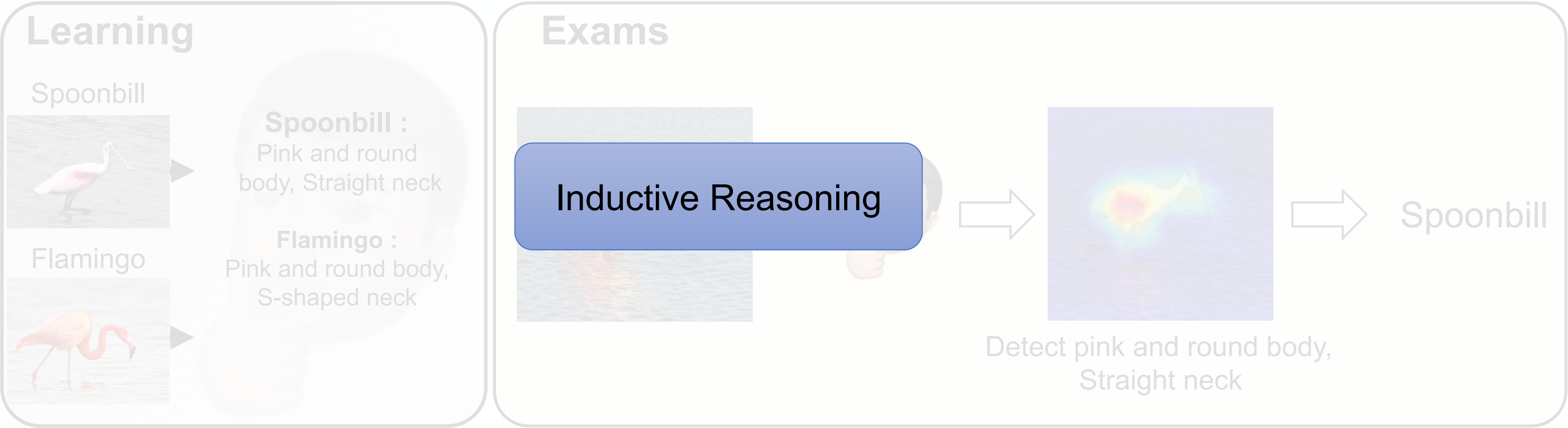


Spoonbill

Detect pink and round body, Straight neck

Reasoning
Types of Reasoning

Classwork – Learned differences between Flamingo and Spoonbill.
Tests – To identify unknown bird



Reasoning

Types of Reasoning

Inductive Reasoning

`A feed-forward reasoning approach that is aimed at detecting generalizations, rules, or regularities [1]'

Learning

Spoonbill

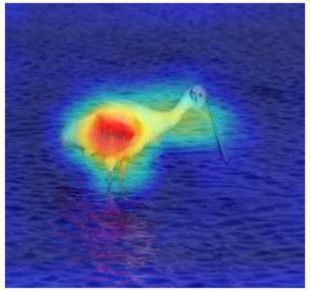
Flamingo

Spoonbill :
Pink and round body, Straight neck

Flamingo :
Pink and round body, S-shaped neck

Exams

Detecting Rules



Detect pink and round body, Straight neck

Spoonbill

Inferring based on rules

OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

[1] Klauer, Karl Josef, and Gary D. Phye. "Inductive reasoning: A training approach." *Review of Educational Research* 78.1 (2008): 85-123.

Reasoning

Types of Reasoning

Deductive Reasoning

Inductive Reasoning

`Reasoning that relies on factual knowledge or formal rules [1]'

Learning

Spoonbill

Flamingo

Spoonbill :
Pink and round body, Straight neck

Flamingo :
Pink and round body, S-shaped neck

Exams

Detect that this image is in training set

Spoonbill

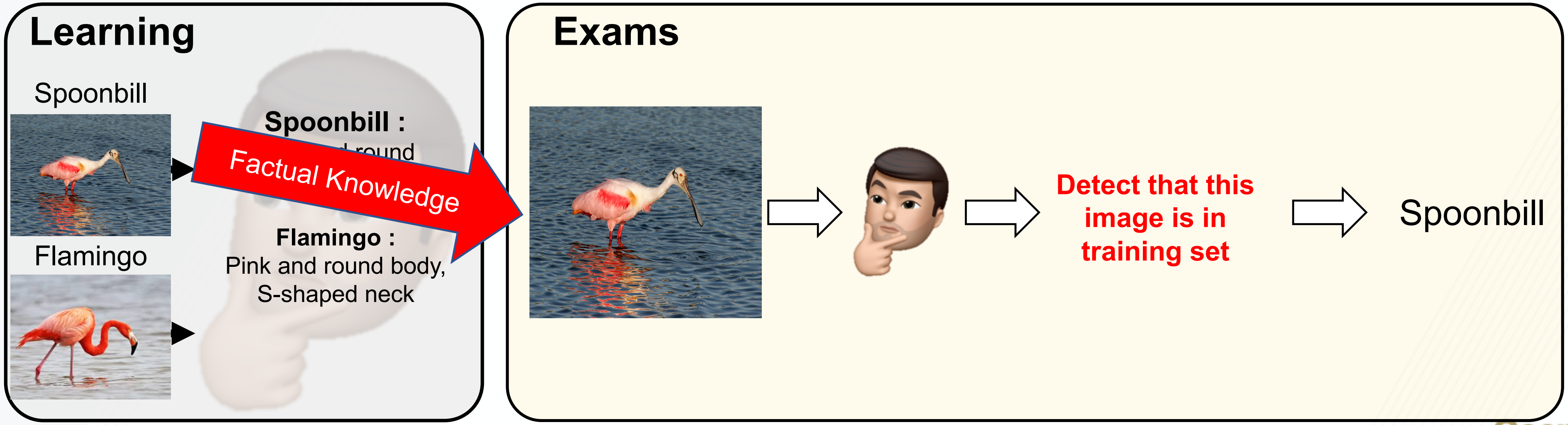
[1] Johnson-Laird, Philip N. "Deductive reasoning." *Annual review of psychology* 50.1 (1999): 109-135.

Reasoning
Types of Reasoning

Deductive Reasoning

Inductive Reasoning

`Reasoning that relies on factual knowledge or formal rules [1]'



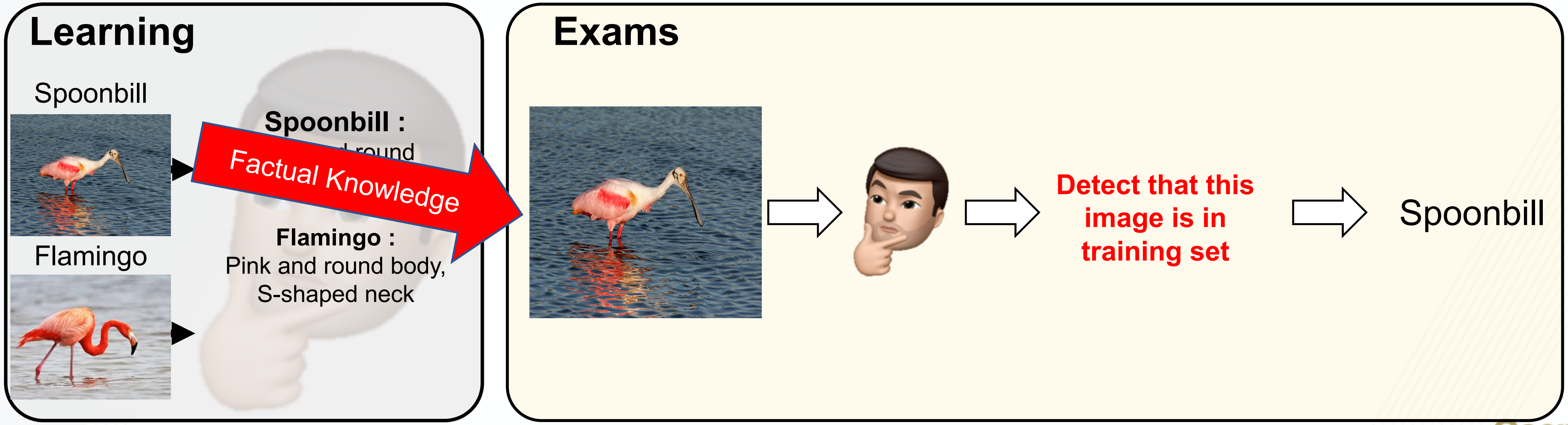
[1] Johnson-Laird, Philip N. "Deductive reasoning." *Annual review of psychology* 50.1 (1999): 109-135.

Reasoning
Types of Reasoning

~~Deductive Reasoning~~

Inductive Reasoning

`Reasoning that relies on factual knowledge or formal rules [1] ’



[1] Johnson-Laird, Philip N. "Deductive reasoning." *Annual review of psychology* 50.1 (1999): 109-135.

Reasoning
Types of Reasoning

Inductive Reasoning

Abductive Reasoning

An abductive reasoning approach creates hypothesis and tests its validity

Learning

Spoonbill

Spoonbill :
Pink and round body, Straight neck

Flamingo

Flamingo :
Pink and round body, S-shaped neck

Exams

Hypotheses → Spoonbill

Reasoning
Types of Reasoning

Inductive Reasoning

Abductive Reasoning

An abductive reasoning approach creates hypothesis and tests its validity

Learning

Spoonbill

Flamingo

Spoonbill :
Pink and round body, Straight neck

Flamingo :
Pink and round body, S-shaped neck

Exams

It is a Flamingo

No S-Shaped neck

Reasoning
Types of Reasoning

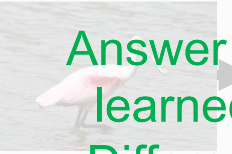
Inductive Reasoning

Abductive Reasoning

An abductive reasoning approach creates hypothesis and tests its validity

Learning

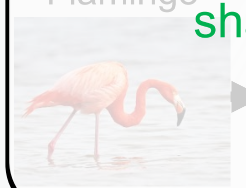
Spoonbill



Spoonbill :

body, Straight neck

Flamingo



Flamingo :

Pink and round body, S-shaped neck

Answer is implicit in the learned knowledge – Difference is in the S-shaped neck

Exams

Why Spoonbill, rather than Flamingo?

Reasoning

Types of Reasoning

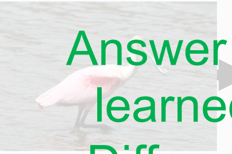
Inductive Reasoning

Abductive Reasoning

An abductive reasoning approach creates hypothesis and tests its validity

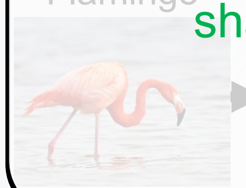
Learning

Spoonbill



Spoonbill : Pink and round body, Straight neck

Flamingo



Flamingo : Pink and round body, S-shaped neck

Answer is implicit in the learned knowledge – Difference is in the S-shaped neck

Exams

Spoonbill

The bird does not have an S-shaped neck

Reasoning
Types of Reasoning

Inductive Reasoning

Abductive Reasoning

An abductive reasoning approach creates hypothesis and tests its validity

Abductive Reasoning

Learning

Spoonbill

Flamingo

Answer is implicit in the learned knowledge – Difference is in the S-shaped neck

Exams

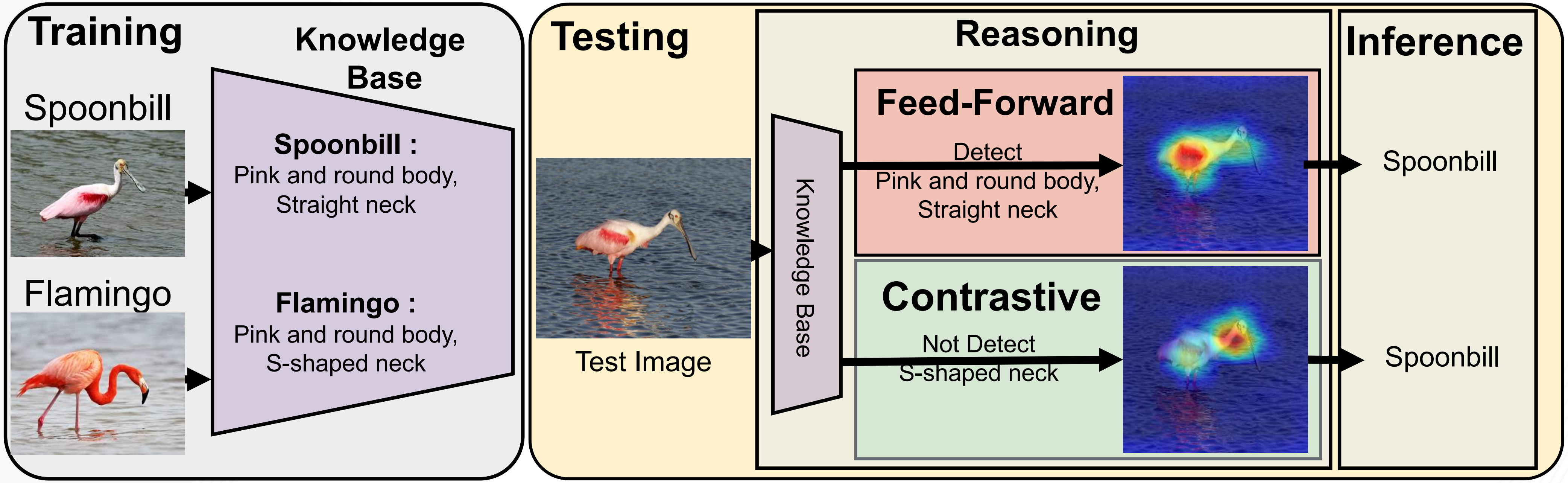
Spoonbill

Not detect S-shaped neck in the given image

Reasoning
Types of Reasoning

Inductive Reasoning

Abductive Reasoning



M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

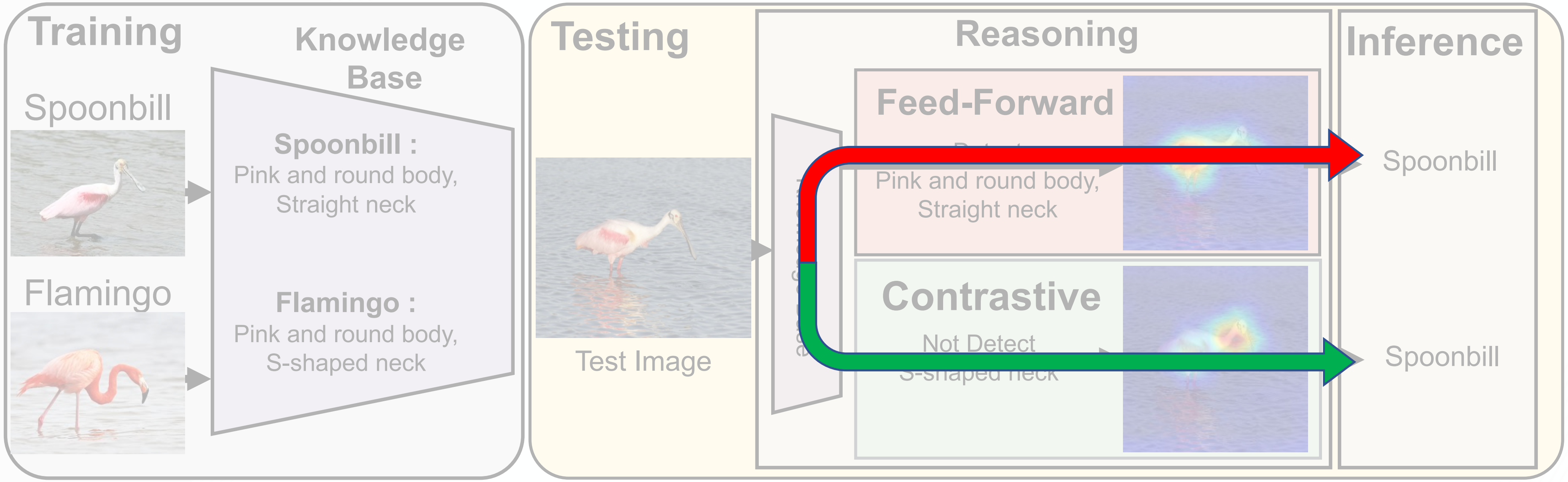
Reasoning
Types of Reasoning

Inductive/Feed-forward Reasoning

Abductive/Contrastive Reasoning

Inductive Reasoning in Neural Networks

Abductive Reasoning in Neural Networks



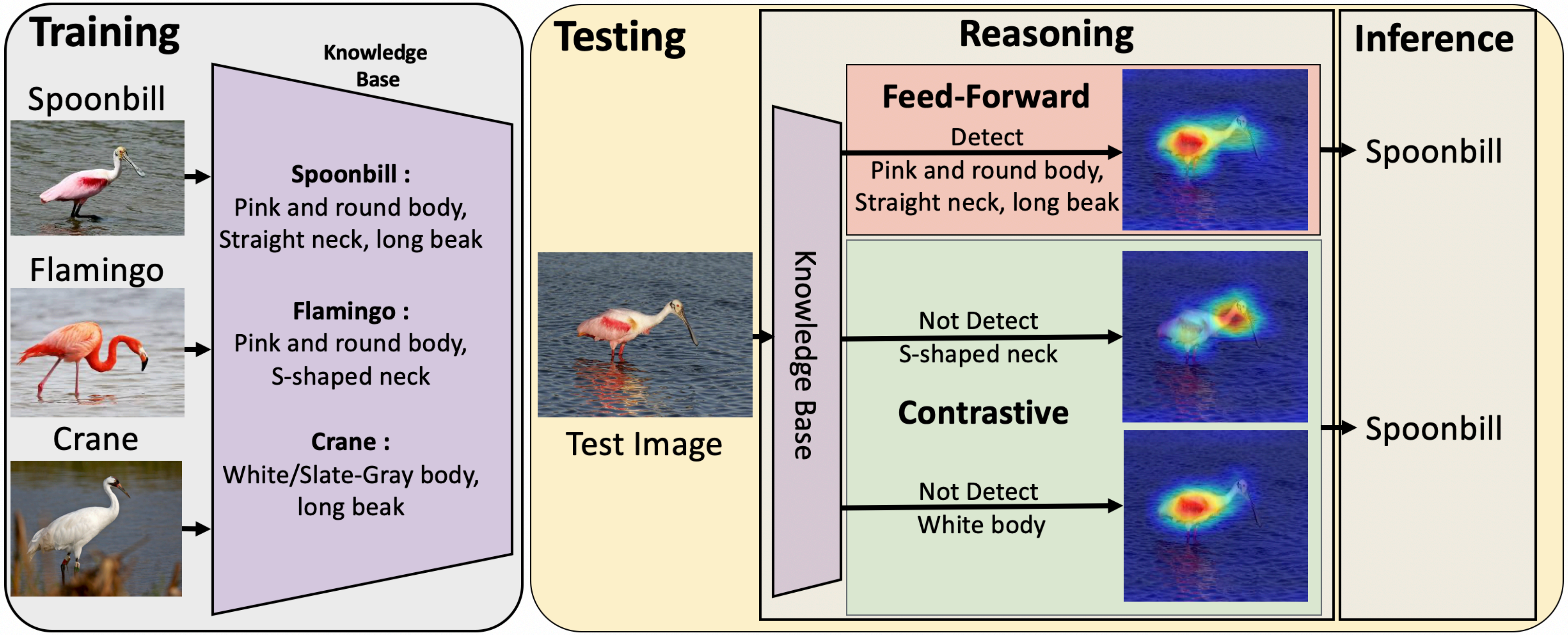
M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

Reasoning

Types of Reasoning

Inductive Reasoning

Abductive Reasoning

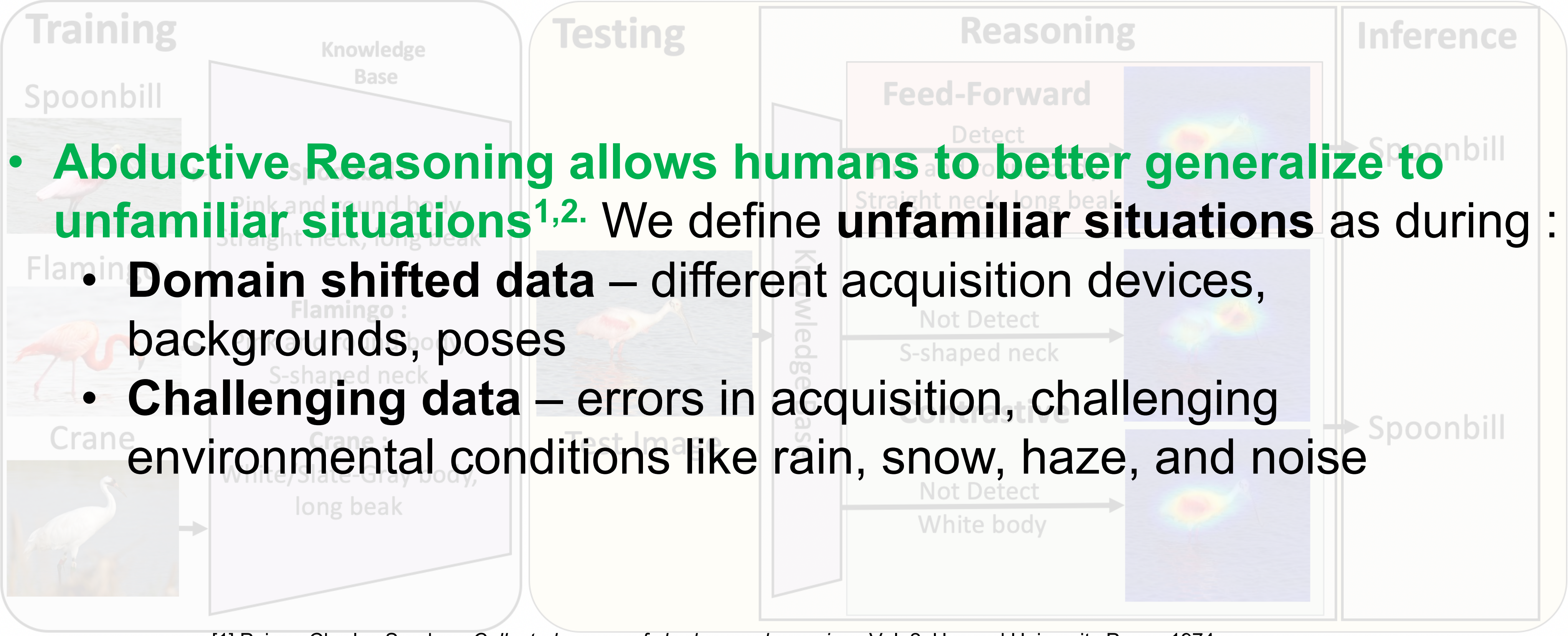


M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

Reasoning
Types of Reasoning

Feed-Forward Reasoning

Contrastive Reasoning



- **Abductive Reasoning allows humans to better generalize to unfamiliar situations[1,2].** We define **unfamiliar situations** as during :
  - **Domain shifted data** – different acquisition devices, backgrounds, poses
  - **Challenging data** – errors in acquisition, challenging environmental conditions like rain, snow, haze, and noise

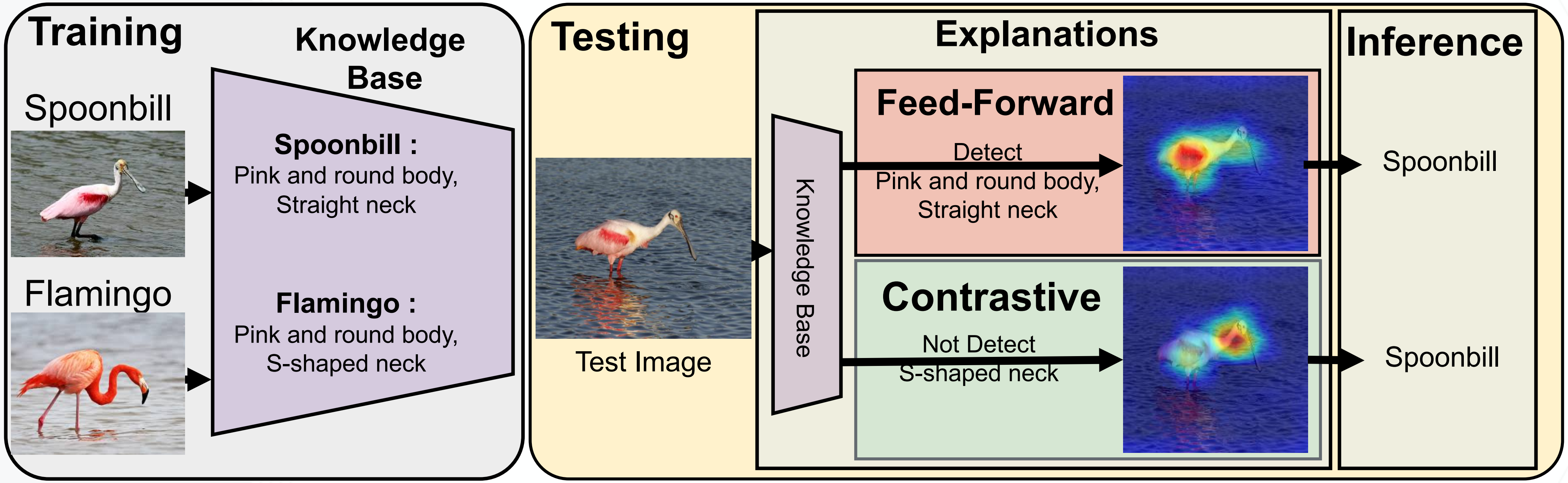
[1] Peirce, Charles Sanders. *Collected papers of charles sanders peirce*. Vol. 2. Harvard University Press, 1974.
[2] Paul, Gabriele. "Approaches to abductive reasoning: an overview." *Artificial intelligence review* 7.2 (1993): 109-152.

OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

Reasoning

Definition of Reasoning



Reasoning is a mental process which can only be surmised based on how it manifests[1]

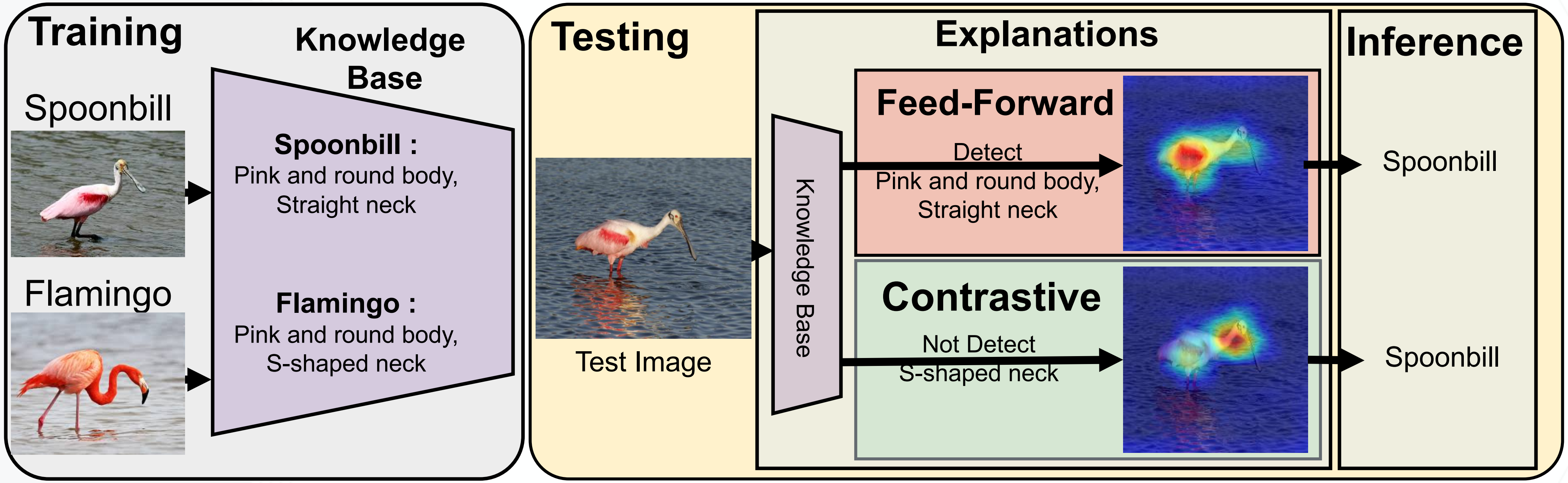
OLIVES @GeorgiaTech

[1] Goguen, Joseph A., J. L. Weiner, and Charlotte Linde. "Reasoning and natural explanation." *International Journal of Man-Machine Studies* 19.6 (1983): 521-559.

Georgia Tech
CREATING THE NEXT

Reasoning
Definition of Reasoning

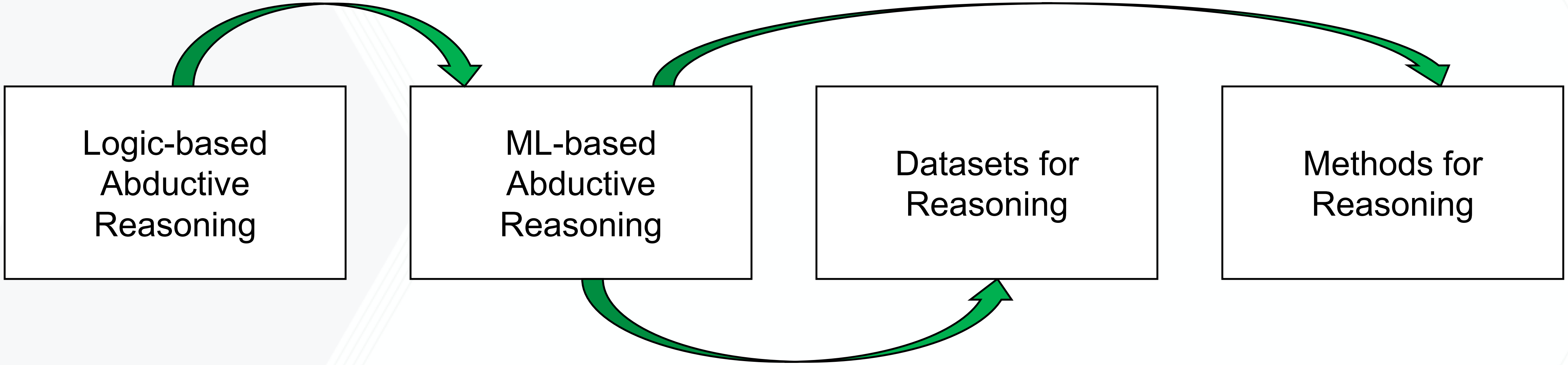
Reasoning manifests in 2 forms : **Explanations** and **Inference** [2]



Reasoning is a mental process which can only be surmised based on how it manifests[1]

[1] Goguen, Joseph A., J. L. Weiner, and Charlotte Linde. "Reasoning and natural explanation." *International Journal of Man-Machine Studies* 19.6 (1983): 521-559.
[2] M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

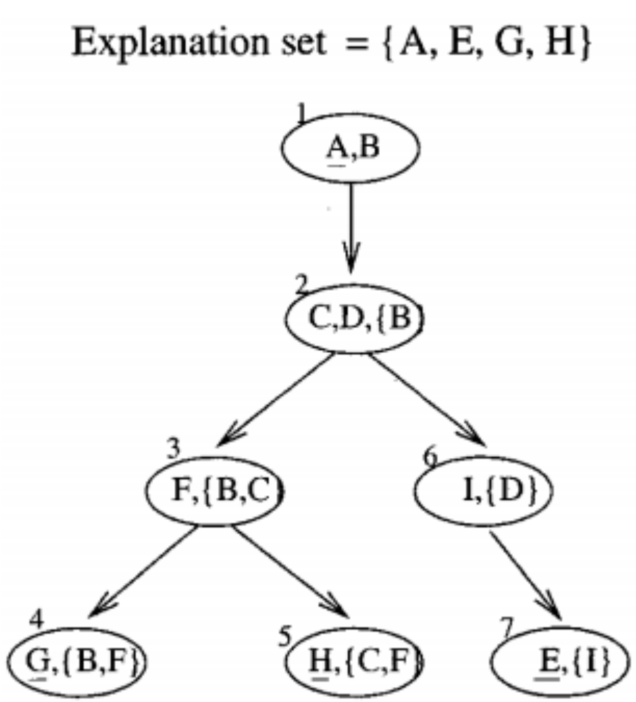
Part III : Reasoning in Neural Networks



Abductive Reasoning

Abductive Reasoning

1999 — Kakas et.al
2002 — Campos et.al [1]
2005 — Raina et.al
2014 — Fortier et.al
2019 — Dai et.al

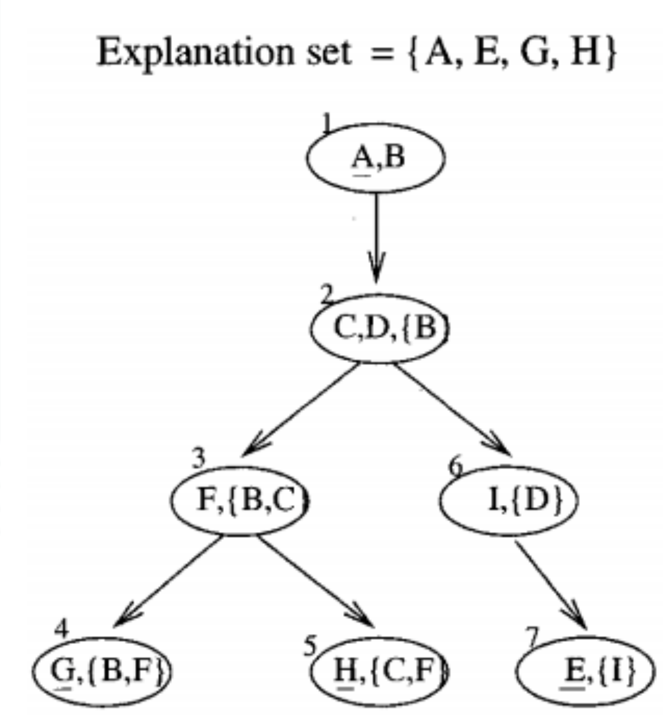


Logic-based programming models

[1] De Campos, Luis M., Jose A. Gamez, and Serafín Moral. "Partial abductive inference in Bayesian belief networks-an evolutionary computation approach by using problem-specific genetic operators." *IEEE Transactions on Evolutionary Computation* 6.2 (2002): 105-131.

Abductive Reasoning

Logic-based Abductive Reasoning	ML-based Abductive Reasoning	Datasets for Reasoning	Methods for Reasoning

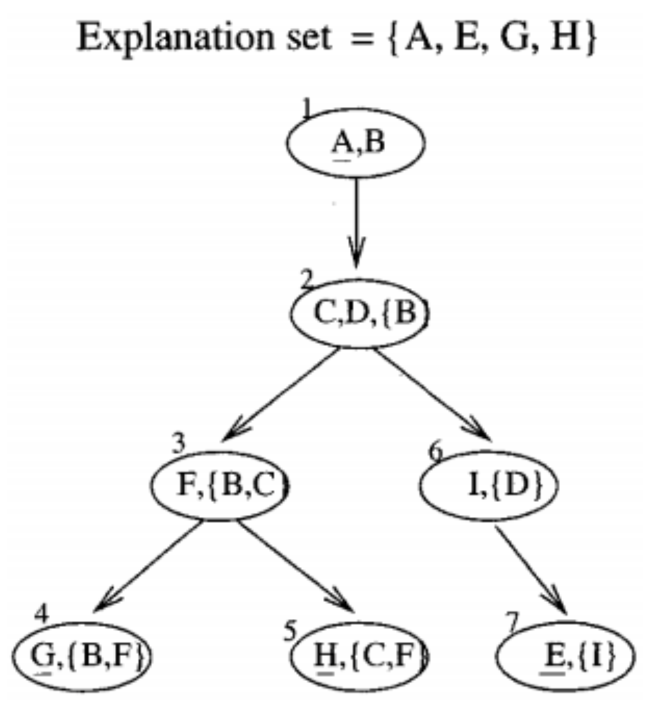


Logic-based programming models

- Decisions are interpretable (explainable)

[1] De Campos, Luis M., Jose A. Gamez, and Serafín Moral. "Partial abductive inference in Bayesian belief networks-an evolutionary computation approach by using problem-specific genetic operators." *IEEE Transactions on Evolutionary Computation* 6.2 (2002): 105-131.

Abductive Reasoning



Logic-based programming models

- Absence of disentangled features in current neural networks
- Not scalable

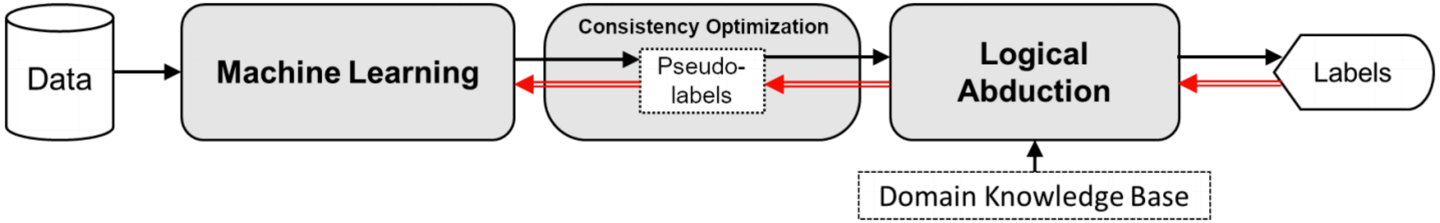
[1] De Campos, Luis M., Jose A. Gamez, and Serafín Moral. "Partial abductive inference in Bayesian belief networks-an evolutionary computation approach by using problem-specific genetic operators." *IEEE Transactions on Evolutionary Computation* 6.2 (2002): 105-131.

Abductive Reasoning

1999 — Kakas et.al
2002 — Campos et.al [1]
2005 — Raina et.al
2014 — Fortier et.al
2019 — Dai et.al[2]

- Decisions are interpretable (explainable)

- Absence of disentangled features in current neural networks
- Not scalable



[1] De Campos, Luis M., Jose A. Gamez, and Serafín Moral. "Partial abductive inference in Bayesian belief networks-an evolutionary computation approach by using problem-specific genetic operators." *IEEE Transactions on Evolutionary Computation* 6.2 (2002): 105-131.
[2] Dai, Wang-Zhou, et al. "Bridging machine learning and logical reasoning by abductive learning." (2019).

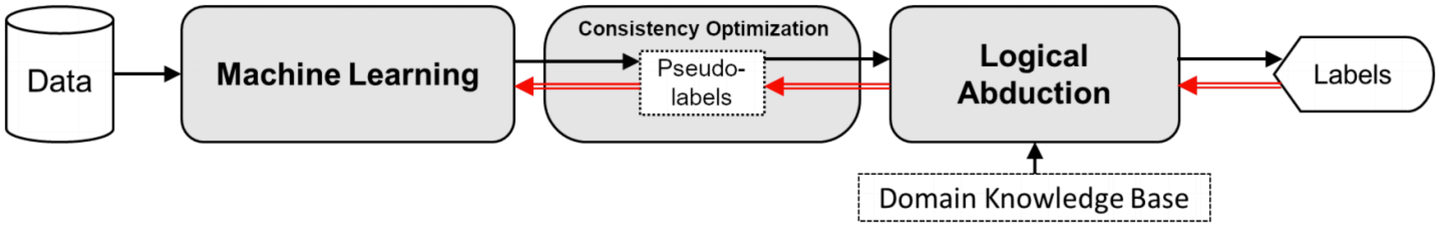
Abductive Reasoning

1999 — Kakas et.al
2002 — Campos et.al [1]
2005 — Raina et.al
2014 — Fortier et.al
2019 — Dai et.al[2]

Explanation set = {A, E, G, H}

- Decisions are interpretable (explainable)

- Absence of disentangled features in current neural networks
- Not scalable



- Having a domain knowledge base creates explainability for learning

[1] De Campos, Luis M., Jose A. Gamez, and Serafín Moral. "Partial abductive inference in Bayesian belief networks-an evolutionary computation approach by using problem-specific genetic operators." *IEEE Transactions on Evolutionary Computation* 6.2 (2002): 105-131.
[2] Dai, Wang-Zhou, et al. "Bridging machine learning and logical reasoning by abductive learning." (2019).

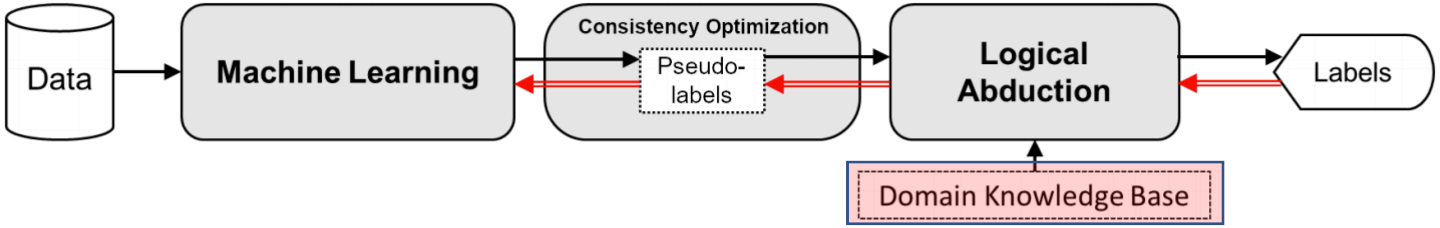
Abductive Reasoning

Logic-based Abductive Reasoning | ML-based Abductive Reasoning | Datasets for Reasoning | Methods for Reasoning

1999 — Kakas et.al
2002 — Campos et.al [1]
2005 — Raina et.al
2014 — Fortier et.al
2019 — Dai et.al[2]

- Decisions are interpretable (explainable)

- Absence of disentangled features in current neural networks
- Not scalable



Knowledge 1
Knowledge 2
Knowledge 3
.
.
.

Requires manual initialization

[1] De Campos, Luis M., Jose A. Gamez, and Serafín Moral. "Partial abductive inference in Bayesian belief networks-an evolutionary computation approach by using problem-specific genetic operators." *IEEE Transactions on Evolutionary Computation* 6.2 (2002): 105-131.
[2] Dai, Wang-Zhou, et al. "Bridging machine learning and logical reasoning by abductive learning." (2019).

Abductive Reasoning

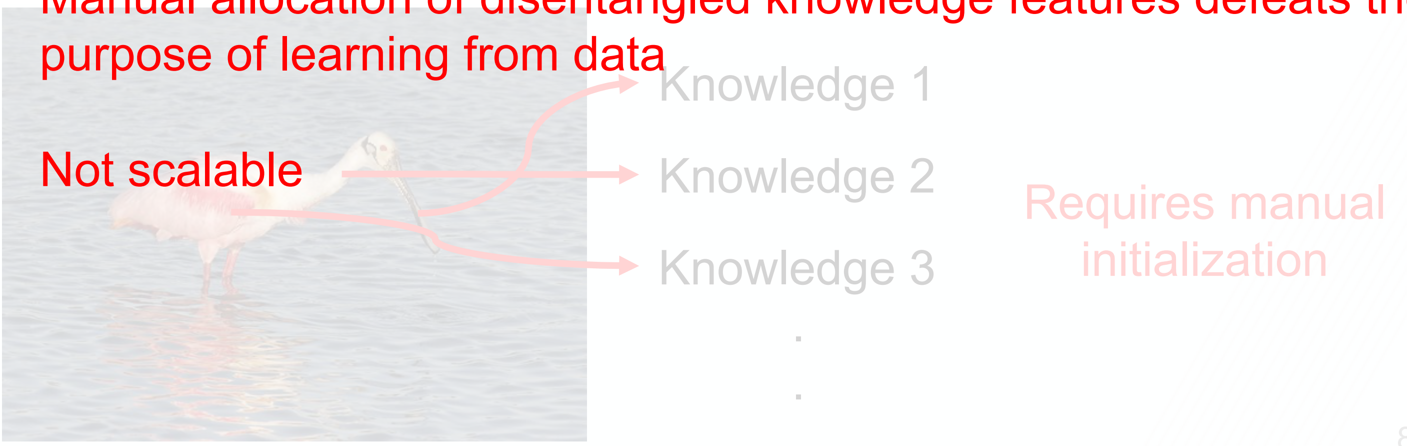
Logic-based Abductive Reasoning | ML-based Abductive Reasoning | Datasets for Reasoning | Methods for Reasoning

1999 — Kakas et.al
2002 — Campos et.al [1]
2005 — Raina et.al
2014 — Fortier et.al
2019 — Dai et.al[2]

- Decisions are interpretable (explainable)

- Absence of disentangled features in current neural networks
- Not scalable

- Manual allocation of disentangled knowledge features defeats the purpose of learning from data
- Not scalable



[1] De Campos, Luis M., Jose A. Gamez, and Serafín Moral. "Partial abductive inference in Bayesian belief networks-an evolutionary computation approach by using problem-specific genetic operators." *IEEE Transactions on Evolutionary Computation* 6.2 (2002): 105-131.
[2] Dai, Wang-Zhou, et al. "Bridging machine learning and logical reasoning by abductive learning." (2019).

Abductive Reasoning

1999 — Kakas et.al
2002 — Campos et.al [1]
2005 — Raina et.al
2014 — Fortier et.al
2019 — Dai et.al[2]

- Decisions are interpretable (explainable)

- Absence of disentangled features in current neural networks
- Not scalable

- Having a domain knowledge base creates explainability for learning

- Manual allocation of disentangled knowledge features defeats the purpose of learning from data
- Not scalable

OLIVES @GeorgiaTech

[1] De Campos, Luis M., Jose A. Gamez, and Serafín Moral. "Partial abductive inference in Bayesian belief networks-an evolutionary computation approach by using problem-specific genetic operators." *IEEE Transactions on Evolutionary Computation* 6.2 (2002): 105-131.
[2] Dai, Wang-Zhou, et al. "Bridging machine learning and logical reasoning by abductive learning." (2019).

Georgia Tech CREATING THE NEXT

Recently…

DeepProbLog: Neural Probabilistic Logic Programming[1]

Inductive Logic Programming via Differentiable Deep Neural Logic Networks[2]

[1] Manhaeve, R., Dumancic, S., Kimmig, A., Demeester, T., & De Raedt, L. (2018). Deepproblog: Neural probabilistic logic programming. *Advances in Neural Information Processing Systems*, *31*, 3749-3759.
[2] Payani, A., & Fekri, F. (2019). Inductive logic programming via differentiable deep neural logic networks. *arXiv preprint arXiv:1906.03523*.

Recently…

DeepProbLog: Neural Probabilistic Logic Programming[1]

MNIST Dataset

Inductive Logic Programming via Differentiable Deep Neural Logic Networks[2]

Relational data – not images

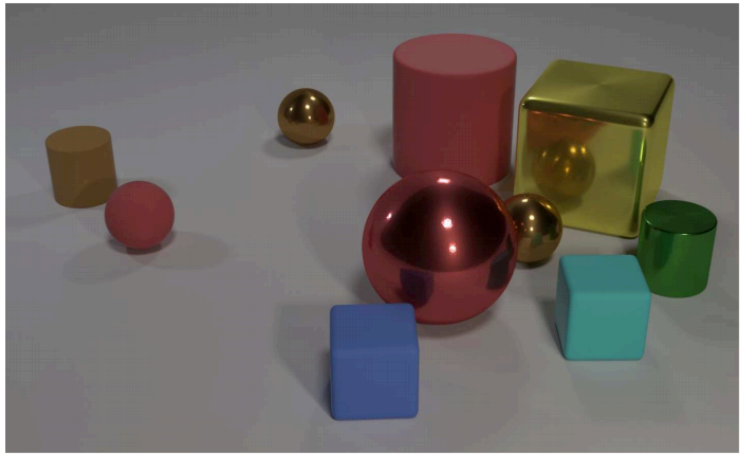
OLIVES @GeorgiaTech

[1] Manhaeve, R., Dumancic, S., Kimmig, A., Demeester, T., & De Raedt, L. (2018). Deepproblog: Neural probabilistic logic programming. *Advances in Neural Information Processing Systems*, *31*, 3749-3759.
[2] Payani, A., & Fekri, F. (2019). Inductive logic programming via differentiable deep neural logic networks. *arXiv preprint arXiv:1906.03523*.

Georgia Tech CREATING THE NEXT

Datasets for Reasoning

2017	2017	2018
Li et.al[1]	Santoro et.al[2]	Santoro et.al

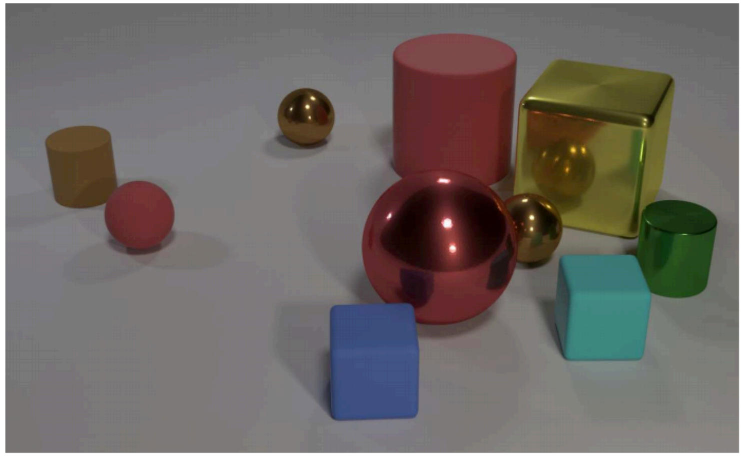


Q: Are there an equal number of large things and metal spheres? **Q:** What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere? **Q:** How many objects are either small cylinders or metal things?

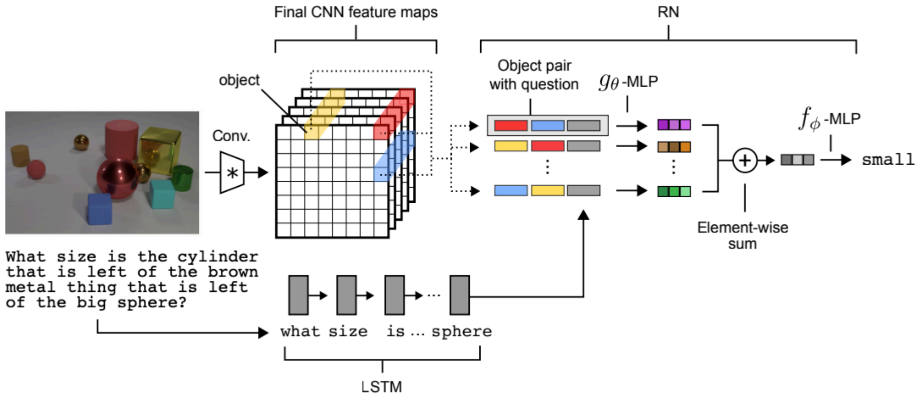
[1] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In CVPR, 2017.
[2] Santoro, Adam, et al. "A simple neural network module for relational reasoning." *arXiv preprint arXiv:1706.01427* (2017).

Datasets for Reasoning

2017 — Li et.al[1]
2017 — Santoro et.al[2]
2018 — Santoro et.al



Q: Are there an equal number of large things and metal spheres? Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere? Q: There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere? Q: How many objects are either small cylinders or metal things?



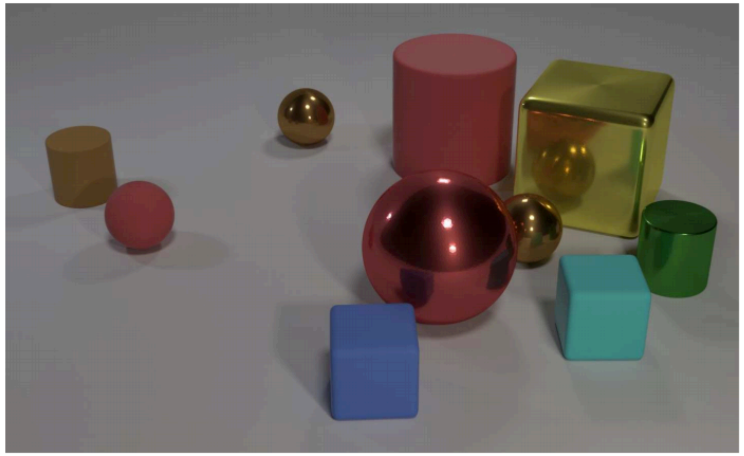
[1] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In CVPR, 2017.
[2] Santoro, Adam, et al. "A simple neural network module for relational reasoning." *arXiv preprint arXiv:1706.01427* (2017).

Datasets for Reasoning

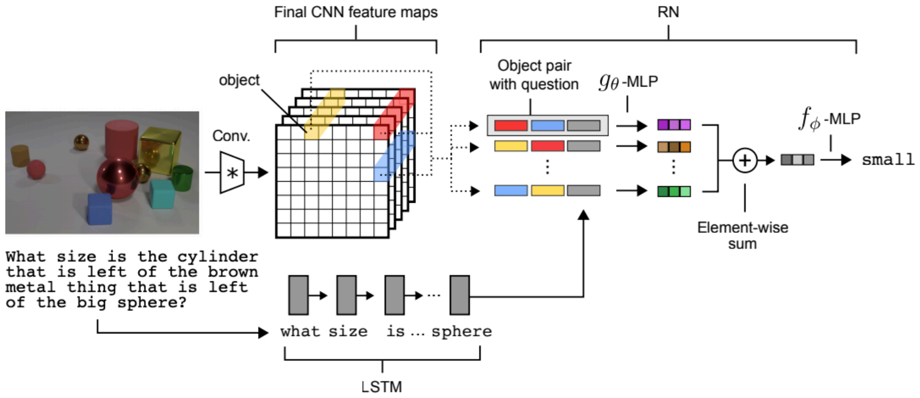
2017 — Li et.al[1]

2017 — Santoro et.al[2]

2018 — Santoro et.al



Q: Are there an equal number of large things and metal spheres? Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere? Q: There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere? Q: How many objects are either small cylinders or metal things?

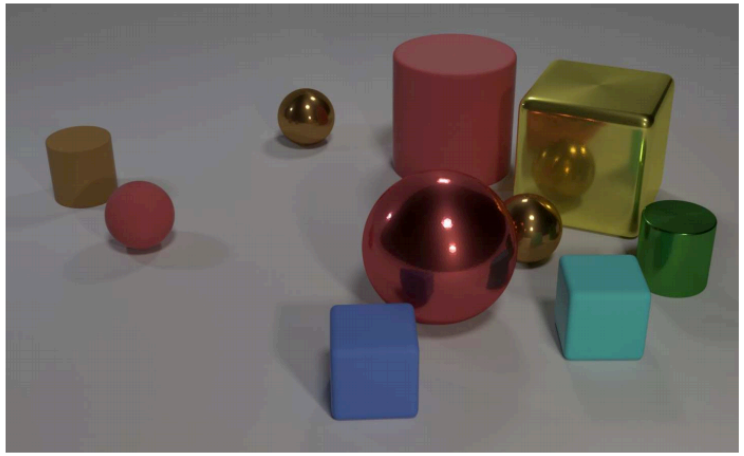


Images + NLP

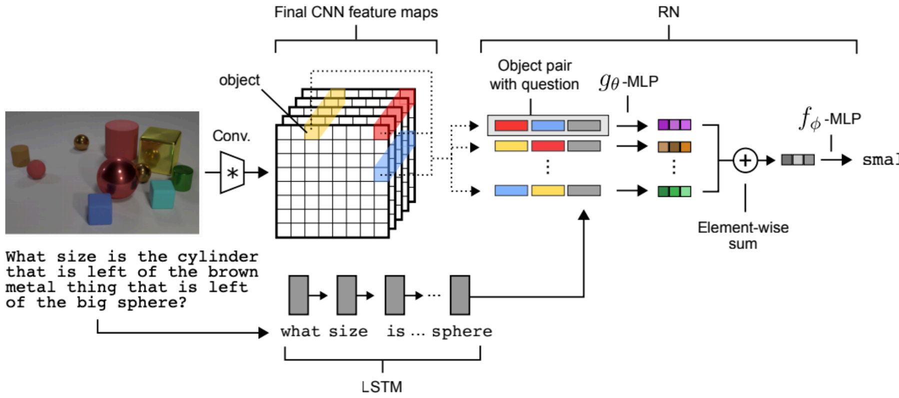
[1] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In CVPR, 2017.
[2] Santoro, Adam, et al. "A simple neural network module for relational reasoning." *arXiv preprint arXiv:1706.01427* (2017).

Datasets for Reasoning

2017 — Li et.al[1]
2017 — Santoro et.al[2]
2018 — Santoro et.al



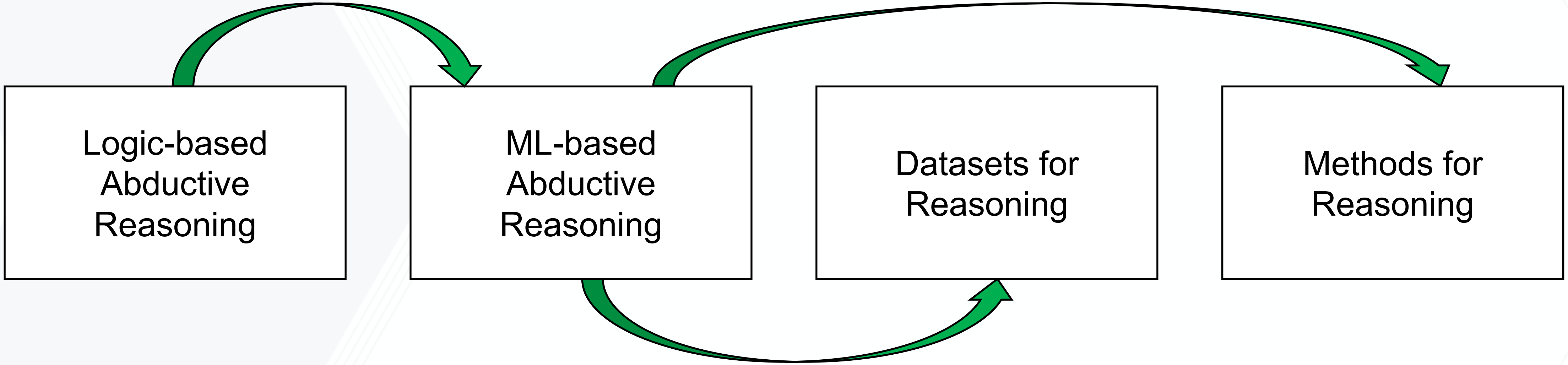
Q: Are there an equal number of large things and metal spheres? **Q:** What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere? **Q:** How many objects are either small cylinders or metal things?



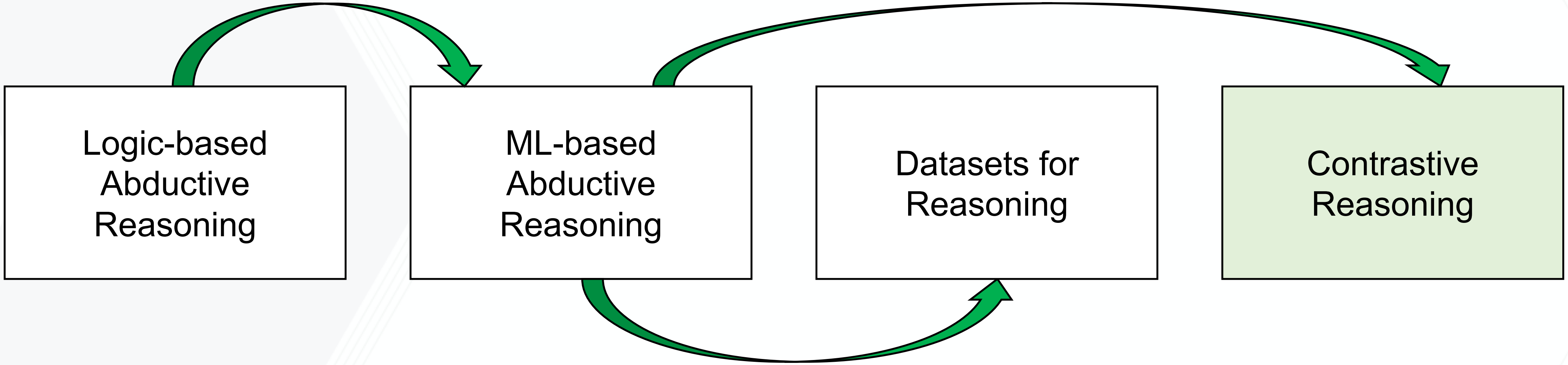
This is a supervised task

[1] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In CVPR, 2017.
[2] Santoro, Adam, et al. "A simple neural network module for relational reasoning." *arXiv preprint arXiv:1706.01427* (2017).

Part I : Reasoning in Neural Networks



Part I : Reasoning in Neural Networks



Contrastive Reasoning

Contrast definition

- Physical Definition
- Structure of Contrast
- Technical Definition

Contrastive Reasoning

Contrast definition

- ❑ Physical Definition
- ❑ Structure of Contrast
- ❑ Technical Definition

In visual space, contrast is the perceived difference between two known quantities

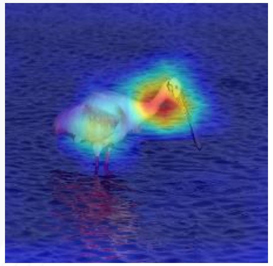
Contrastive Reasoning
Contrast definition

- ❑ Physical Definition
- ❑ Structure of Contrast
- ❑ Technical Definition

In visual space, contrast is the perceived difference between two known quantities

Contrast B/w Spoonbill and Flamingo

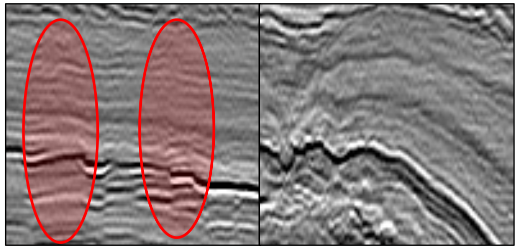
Is in the neck



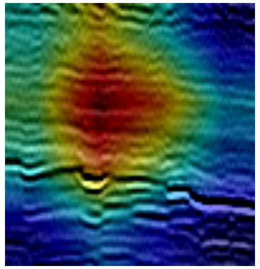
Contrast B/w Bugatti Convertible and Coupe

Is in the open top

Contrast B/w Fault and Salt Dome



Is in the tectonic shift



Contrastive Reasoning

Contrast definition

- ❑ Physical Definition
- ❑ Structure of Contrast
- ❑ Technical Definition

`Why P, rather than Q?'

$P \longrightarrow$ *Prediction*

$Q \longrightarrow$ *Contrast class*

Contrastive Reasoning
Contrast definition

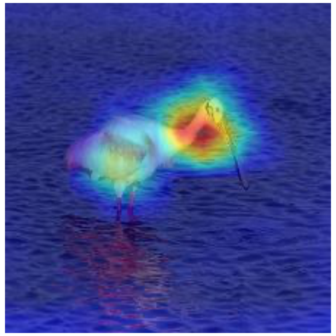
- ❑ Physical Definition
- ❑ Structure of Contrast
- ❑ Technical Definition

`Why P, rather than Q?'

P → **Spoonbill** → *Prediction*

Q → **Flamingo** → *Contrast class*

`Why spoonbill, rather than flamingo?'



Contrastive Reasoning

Contrast definition

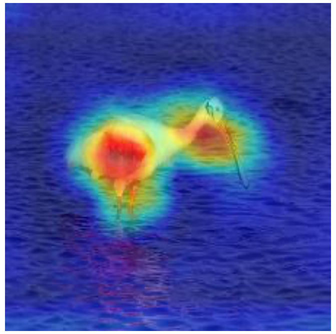
- ❑ Physical Definition
- ❑ Structure of Contrast
- ❑ Technical Definition

`Why P, rather than Q?'

P → Spoonbill → *Prediction*

Q → Pig → *Contrast class*

`Why spoonbill, rather than pig?'



Contrastive Reasoning

Contrast definition

- ❑ Physical Definition
- ❑ Structure of Contrast
- ❑ Technical Definition

`Why P, rather than all classes?'

P → Spoonbill → *Prediction*

Q → Flamingo/Pig/… → *Contrast class*

`Why spoonbill, rather than pig?'

For N learned classes, there can be N possible contrastive reasons

Contrastive Reasoning
Contrast definition

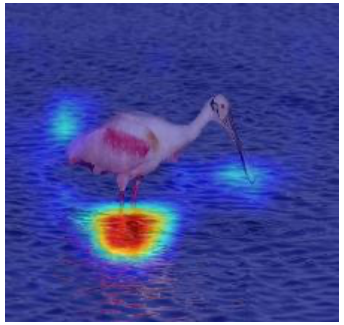
- ❑ Physical Definition
- ❑ Structure of Contrast
- ❑ Technical Definition

`Why P, rather than P?'

P → Spoonbill → *Prediction*

Q → Spoonbill → *Contrast class*

`Why spoonbill, rather than spoonbill?'



Contrastive Reasoning

Contrast definition

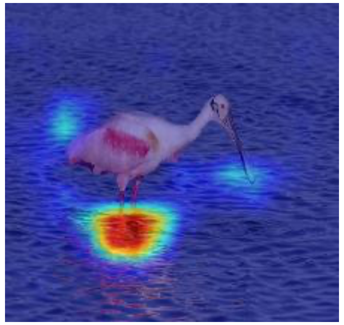
- Physical Definition
- Structure of Contrast
- Technical Definition

`Why P, rather than P?'

P → Spoonbill → *Prediction*

Q → Spoonbill → *Contrast class*

`Why not spoonbill, with 100% confidence?'



OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

Contrastive Reasoning
Contrast definition

- ❑ Physical Definition
- ❑ Structure of Contrast
- ❑ Technical Definition

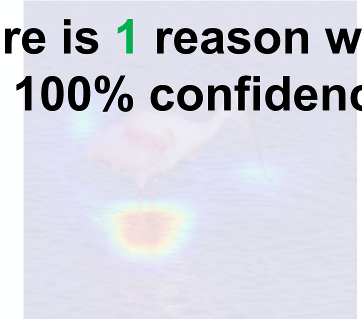
`Why P, rather than P?'

P → Spoonbill → *Prediction*

Q → Spoonbill → *Contrast class*

`Why not spoonbill, with 100% confidence?'

For 1 predicted class, there is 1 reason why it was not predicted with 100% confidence



OLIVES @GeorgiaTech

Contrastive Reasoning

Contrast definition

- ❑ Physical Definition
- ❑ Structure of Contrast
- ❑ Technical Definition

OLIVES
@GeorgiaTech

Georgia Tech
CREATING THE NEXT

Contrastive Reasoning

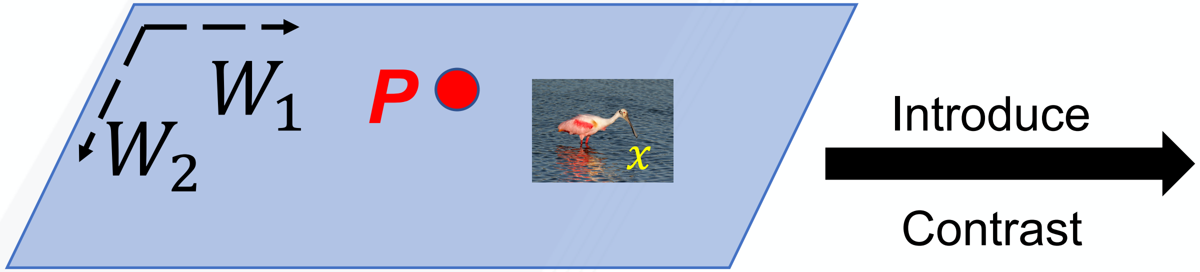
Contrast definition

In representation space, contrast is the distance between manifolds where an input x is predicted as P vs the same input x is predicted as Q

Contrastive Reasoning

Contrast definition

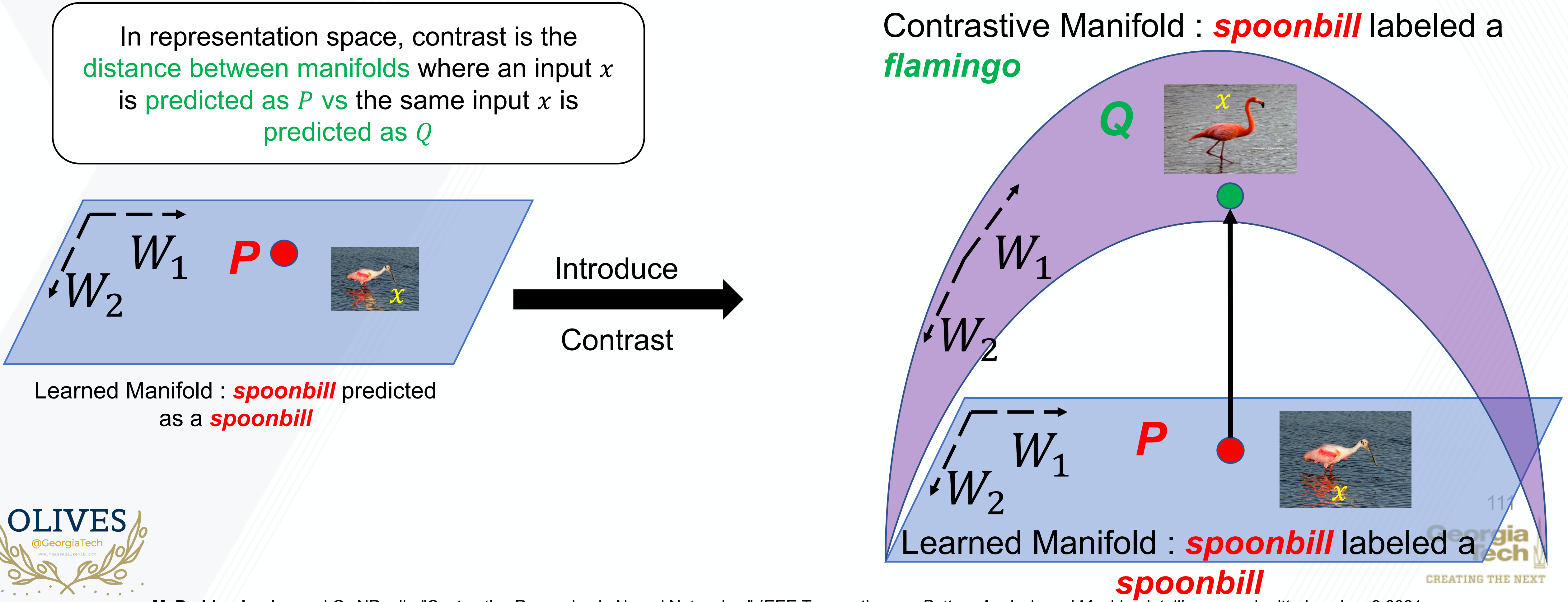
In representation space, contrast is the distance between manifolds where an input x is predicted as P vs the same input x is predicted as Q



Learned Manifold : ***spoonbill*** predicted as a ***spoonbill***

Contrastive Reasoning
Contrast definition

In representation space, contrast is the distance between manifolds where an input x is predicted as P vs the same input x is predicted as Q



OLIVES @GeorgiaTech

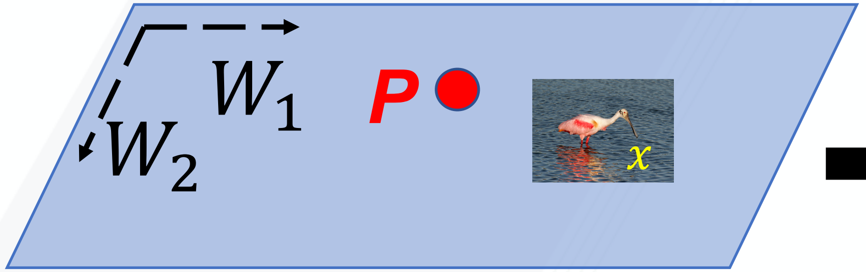
Georgia Tech CREATING THE NEXT

M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

Contrastive Reasoning

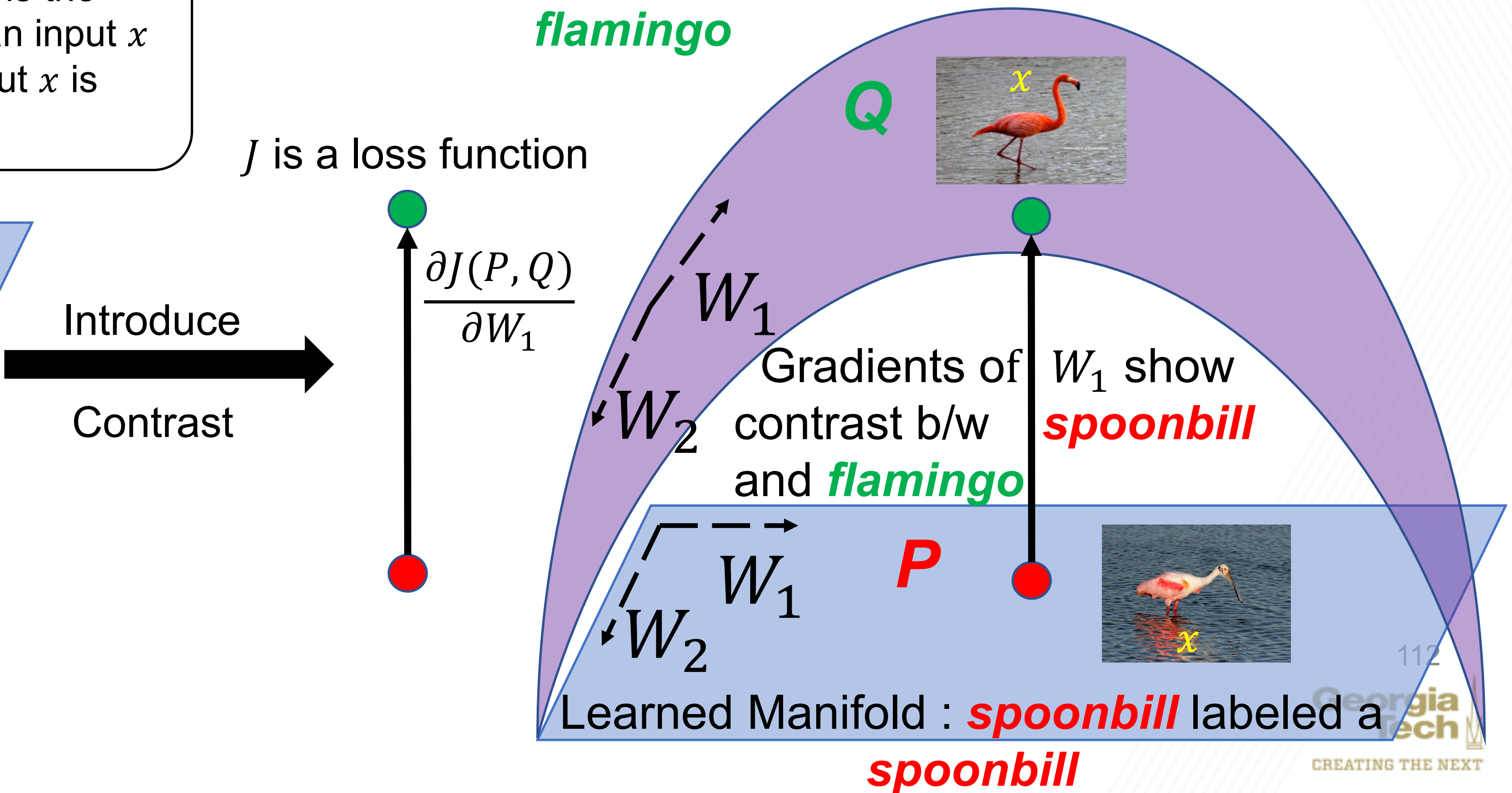
Contrast definition

In representation space, contrast is the distance between manifolds where an input x is predicted as P vs the same input x is predicted as Q



Learned Manifold : ***spoonbill*** predicted as a ***spoonbill***

Introduce Contrast



OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

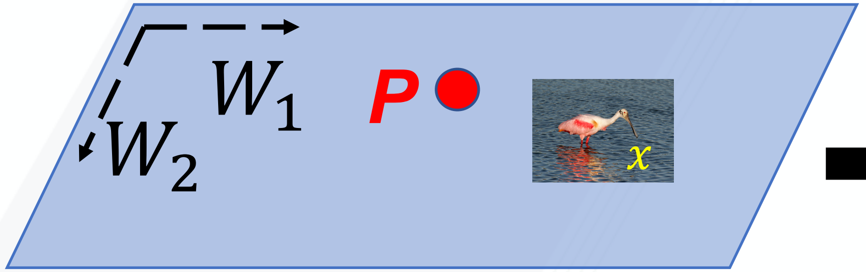
M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

Contrastive Reasoning

Contrast definition

Gradients provide inherent contrast between classes

In representation space, contrast is the distance between manifolds where an input x is predicted as P vs the same input x is predicted as Q

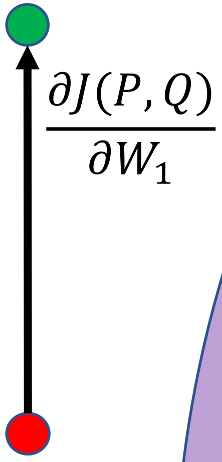


Learned Manifold : ***spoonbill*** predicted as a ***spoonbill***

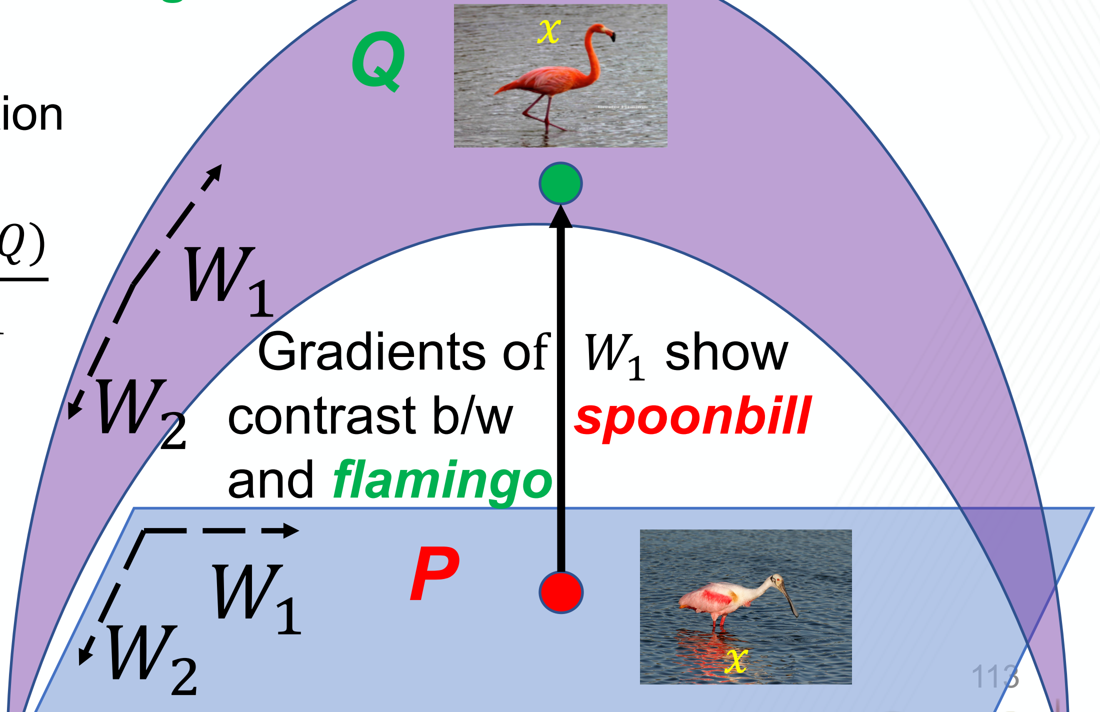
Introduce

Contrast

J is a loss function



Contrastive Manifold : ***spoonbill*** labeled a ***flamingo***



Learned Manifold : ***spoonbill*** labeled a ***spoonbill***

M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

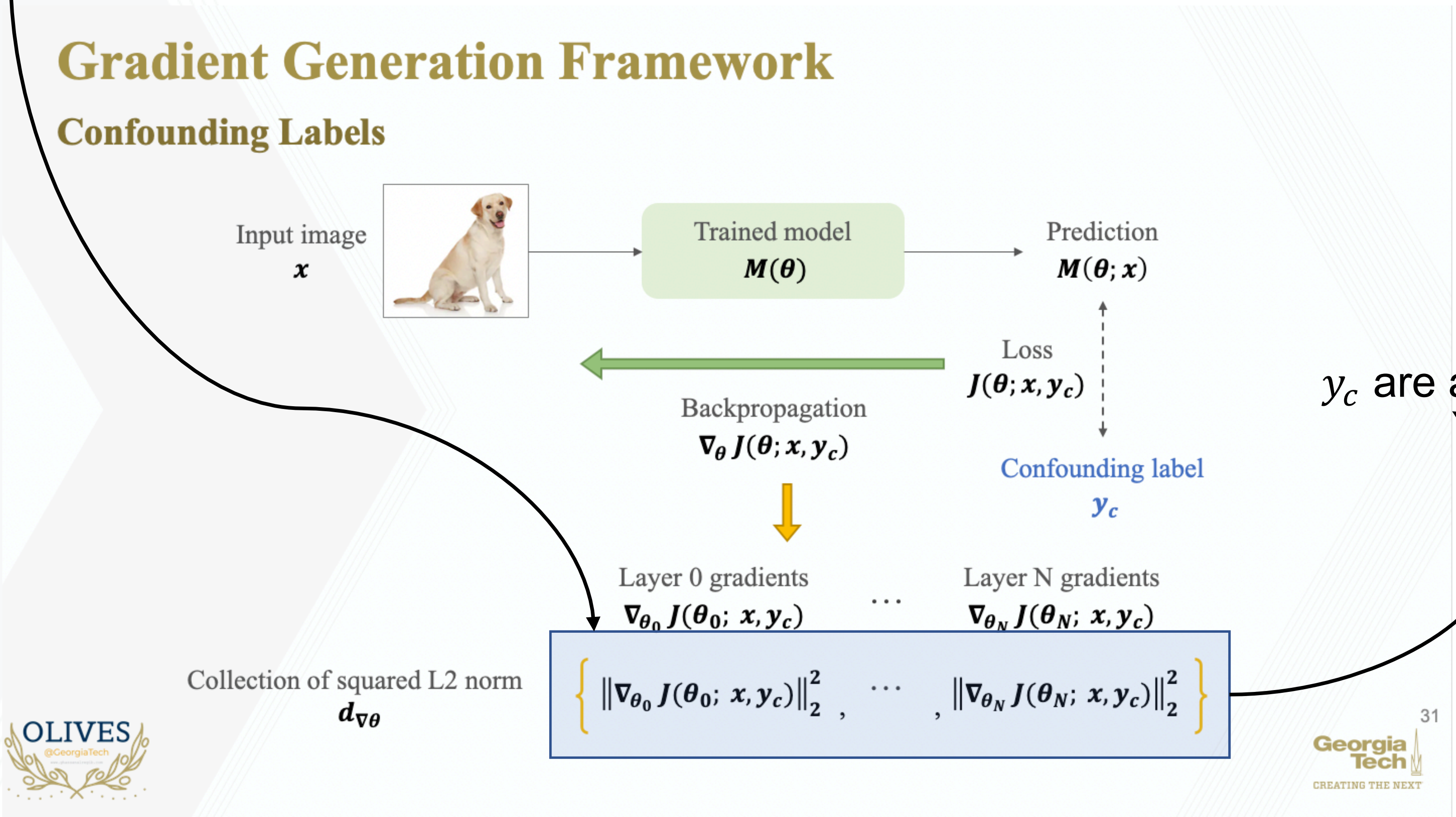
Parts I, II, III

Contrast definition

`Why P, rather than all classes?'

`Why P, rather than P?'

`Why P, rather than Q?'



y_c are all possible Q

Parts I, II, III
Contrast definition

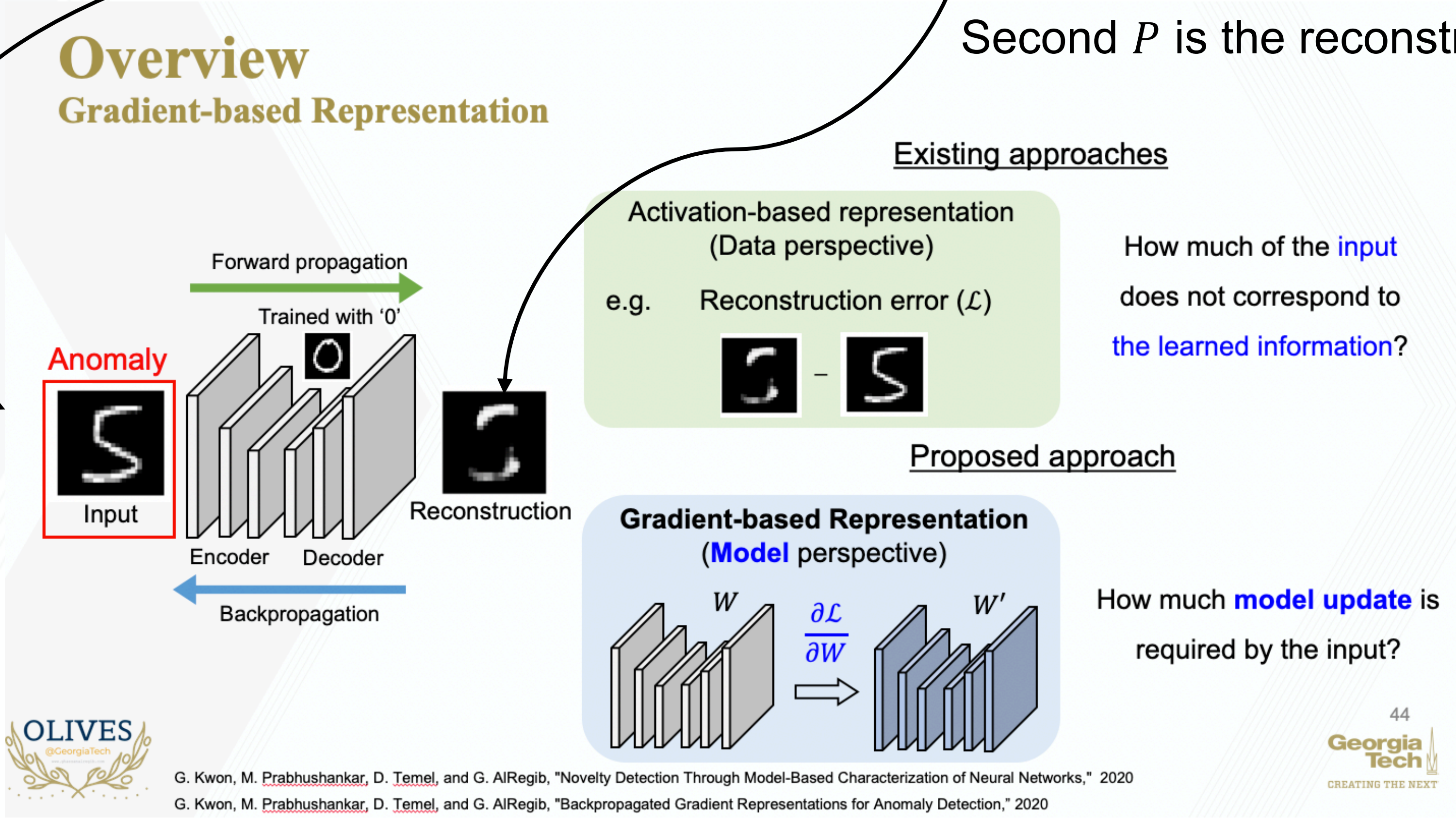
\`Why P, rather than all classes?'

\`Why P, rather than P?'

\`Why P, rather than Q?'

Second P is the reconstructed image

First P is original image

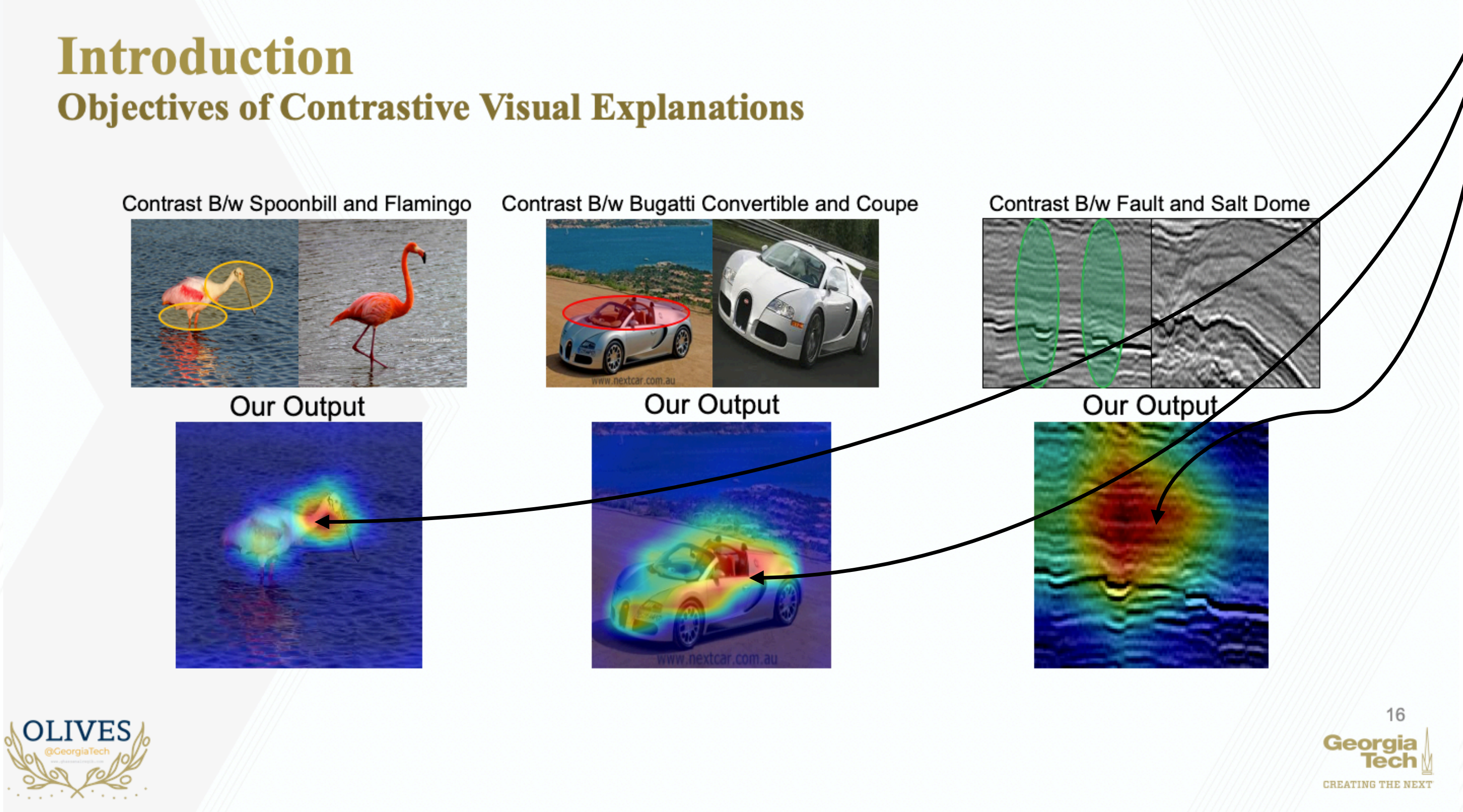


Parts I, II, III
Contrast definition

`Why P, rather than all classes?'

`Why P, rather than P?'

`Why P, rather than Q?'

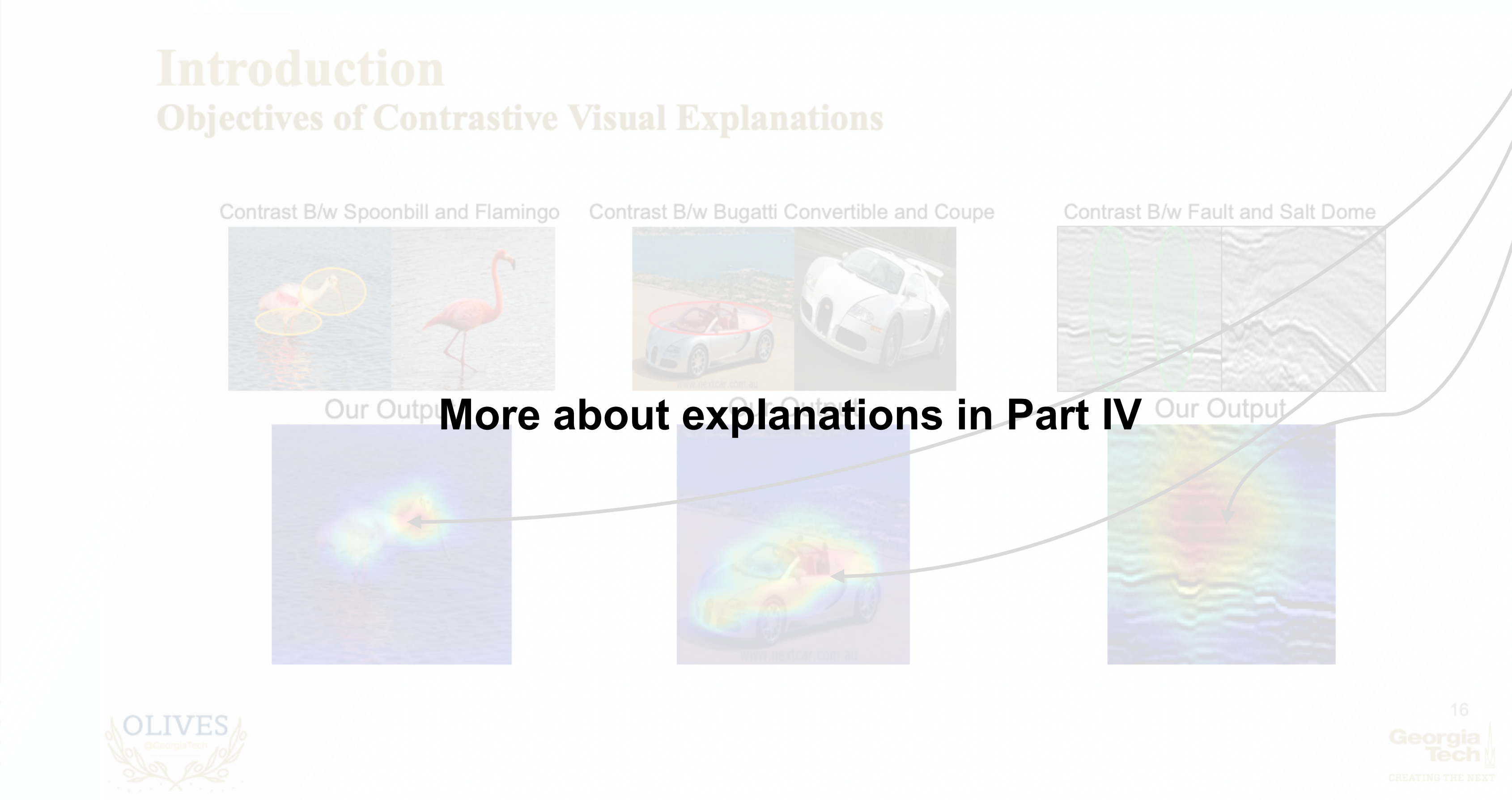


Parts I, II, III
Contrast definition

`Why P, rather than all classes?'

`Why P, rather than P?'

`Why P, rather than Q?'



More about explanations in Part IV

Reasoning in Neural Networks

1999	2002	2005	2014	2017	2017	2019	2021
Kakas et.al	Campos et.al[1]	Raina et.al	Fortier et.al	Li et.al[2]	Santoro et.al[3]	Dai et.al[4]	Prabhushankar et.al[5]

[1]	[2]	[3]	[4]	[5]
• Decisions are interpretable (explainable)	• Reasoning dataset • Images + Natural language	• Provides relational reasoning and validation using numbers	• Having a domain knowledge base creates explainability for learning	• Does not require domain knowledge • Multiple explanatory questions
• Absence of disentangled features in current neural networks • Not scalable	• Images + Natural language	• Treats reasoning as a supervised task	• Manual allocation of disentangled knowledge features defeats the purpose of learning from data • Not scalable	• Dependent on knowledge of neural network

[1] De Campos, Luis M., Jose A. Gamez, and Serafín Moral. "Partial abductive inference in Bayesian belief networks-an evolutionary computation approach by using problem-specific genetic operators." *IEEE Transactions on Evolutionary Computation* 6.2 (2002): 105-131.
[2] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In CVPR, 2017.
[3] Santoro, Adam, et al. "A simple neural network module for relational reasoning." *arXiv preprint arXiv:1706.01427* (2017).
[4] Dai, Wang-Zhou, et al. "Bridging machine learning and logical reasoning by abductive learning." (2019).
[5] M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

Reasoning in Neural Networks

1999	2002	2005	2014	2017	2017	2019	2021
Kakas et.al[1]	Campos et.al[2]	Raina et.al[3]	Fortier et.al[4]	Li et.al[5]	Santoro et.al[6]	Dai et.al[7]	Prabhushankar et.al[8]

[1] Flach, Peter A., and Antonis C. Kakas. "Abductive and inductive reasoning: background and issues." *Abduction and induction*. Springer, Dordrecht, 2000. 1-27.
[2] De Campos, Luis M., Jose A. Gamez, and Serafín Moral. "Partial abductive inference in Bayesian belief networks-an evolutionary computation approach by using problem-specific genetic operators." *IEEE Transactions on Evolutionary Computation* 6.2 (2002): 105-131.
[3] Raina, Rajat, Andrew Y. Ng, and Christopher D. Manning. "Robust textual inference via learning and abductive reasoning." *AAAI*. 2005.
[4] Fortier, Nathan, John Sheppard, and Shane Strasser. "Abductive inference in Bayesian networks using distributed overlapping swarm intelligence." *Soft Computing* 19.4 (2015): 981-1001.
[5] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In CVPR, 2017.
[6] Santoro, Adam, et al. "A simple neural network module for relational reasoning." *arXiv preprint arXiv:1706.01427* (2017).
[7] Dai, Wang-Zhou, et al. "Bridging machine learning and logical reasoning by abductive learning." (2019).
[8] M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

So far,

- We introduced an interpretation of **gradients in the space of models** from a perspective of **model uncertainty**
- We proposed a framework for efficient gradient generation with **confounding labels** to quantify uncertainty of fully trained networks
- We validated that the gradient-based uncertainty measure outperform activation-based features in **out-of-distribution detection** and **corrupted input detection**
- We interpreted gradients as a reasoning mechanism within neural networks

arXiv

https://arxiv.org/abs/2103.12329

M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

References

- M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021. [PDF]
- M. Prabhushankar, G. Kwon, D. Temel, and G. AlRegib, "Contrastive Explanations in Neural Networks," in *IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, Oct. 2020. [PDF][Code][Video]
- M. Prabhushankar and G. AlRegib, "Extracting Causal Visual Features for Limited Label Classification," in *IEEE International Conference on Image Processing (ICIP)*, Anchorage, AK, Sep. 19-22 2021.

OLIVES @GeorgiaTech

Georgia Tech
CREATING THE NEXT

Part IV : Explanations in Neural Networks

Part IV : Explanations in Neural Networks

Causal Questions

'Why?'

Abductive Reasoning

Counterfactual Questions

'What if?'

Contrastive Questions

'Why P, rather than Q?'

Part IV : Explanations in Neural Networks

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

OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

Explanations in Neural Networks

Observed Causal Explanations

'Why the the network predict a spoonbill?'

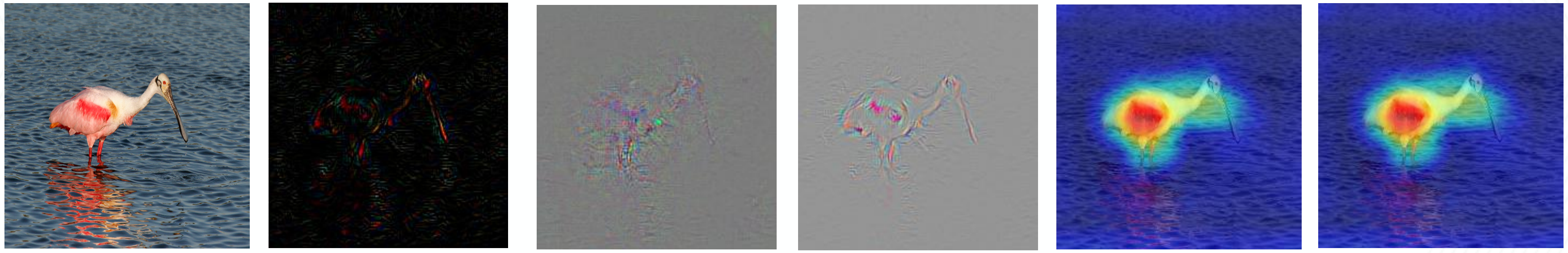
2013 — Simoyan et.al [1]

2014 — Springenberg et.al [2]

2017 — Selvaraju et.al [3]

2018 — Aditya et.al [4]

`Why Spoonbill?'



Original | Positive Saliency [1] | Smooth Gradients [1] | Guided Backpropagation [2] | GradCAM [3] | GradCAM++ [4]

[1] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman,"Deep inside convolutional networks:Visualising image classification models and saliency maps,"arXiv preprintarXiv:1312.6034, 2013.
[2] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for Simplicity: The All Convolutional Net. arXiv, 2014
[3] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das,Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," inProceedings of the IEEE internationalconference on computer vision, 2017, pp. 618–626.
[4] 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.

Explanations in Neural Networks

Observed Causal Explanations

'Why the the network predict a spoonbill?'

2013 — Simoyan et.al [1]

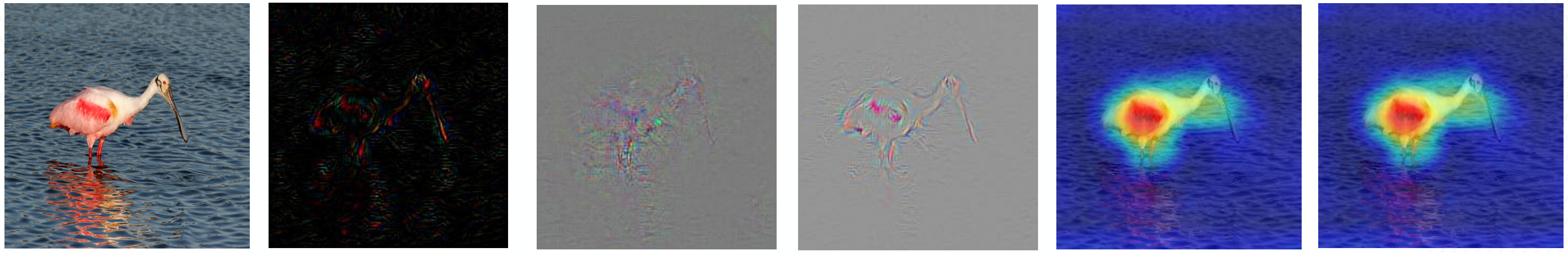
2014 — Springenberg et.al [2]

2017 — Selvaraju et.al [3]

2018 — Aditya et.al [4]

`Why Spoonbill?'

All techniques are a function of activations and gradients of logit of predicted class



Original | Positive Saliency [1] | Smooth Gradients [1] | Guided Backpropagation [2] | GradCAM [3] | GradCAM++ [4]

[1] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman,"Deep inside convolutional networks:Visualising imageclassification models and saliency maps,"arXiv preprintarXiv:1312.6034, 2013.
[2] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for Simplicity: The All Convolutional Net. arXiv, 2014
[3] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das,Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," inProceedings of the IEEE internationalconference on computer vision, 2017, pp. 618–626.
[4] 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.

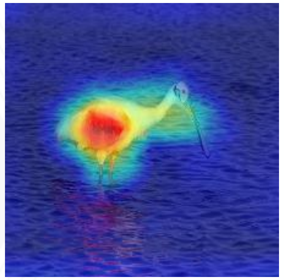
Explanations in Neural Networks

Observed Causal Explanations – Grad-CAM

2013	2014	2017	2018
Simoyan et.al [1]	Springenberg et.al [2]	Selvaraju et.al [3]	Aditya et.al [4]

`Why Spoonbill?'

Original



Grad-CAM [3]

```
logit = self.model_arch(input)
#Grad-CAM gradient initialization
if class_idx is None:
    score = logit[:, logit.max(1)[-1]].squeeze()
else:
    score = logit[:, class_idx].squeeze()

self.model_arch.zero_grad()
score.backward(retain_graph=retain_graph)
```

Grad-CAM

- Pass an image through a network
- Obtain the logits after the final layer
- Backpropagate the required logit, y_P, to the final convolutional layer
- Sum all the gradients per channel to obtain k importance scores
- Multiply the importance scores with the activations per channel and average them across channels

[1] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman,"Deep inside convolutional networks:Visualising imageclassification models and saliency maps,"arXiv preprintarXiv:1312.6034, 2013.
[2] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for Simplicity: The All Convolutional Net. arXiv, 2014
[3] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das,Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," inProceedings of the IEEE internationalconference on computer vision, 2017, pp. 618–626.
[4] 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.

Explanations in Neural Networks

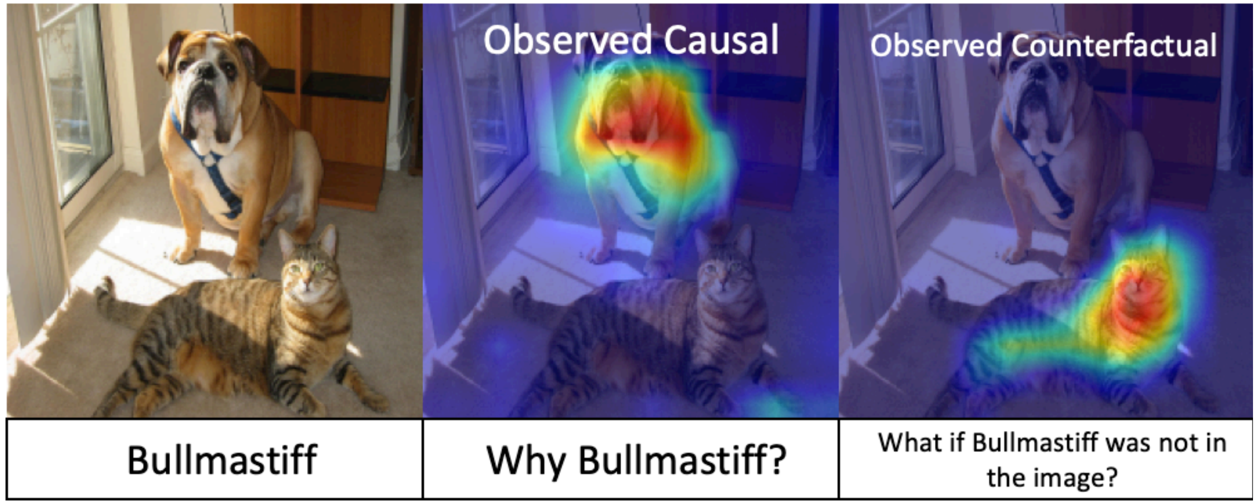
Counterfactual Explanations – Gradient based

‘What if the bullmastiff was not in the image?’

Obtained by backpropagating the negative gradient of the logit y_P in Grad-CAM framework

2017 — Selvaraju et.al [1]

2019 — Goyal et.al [2]

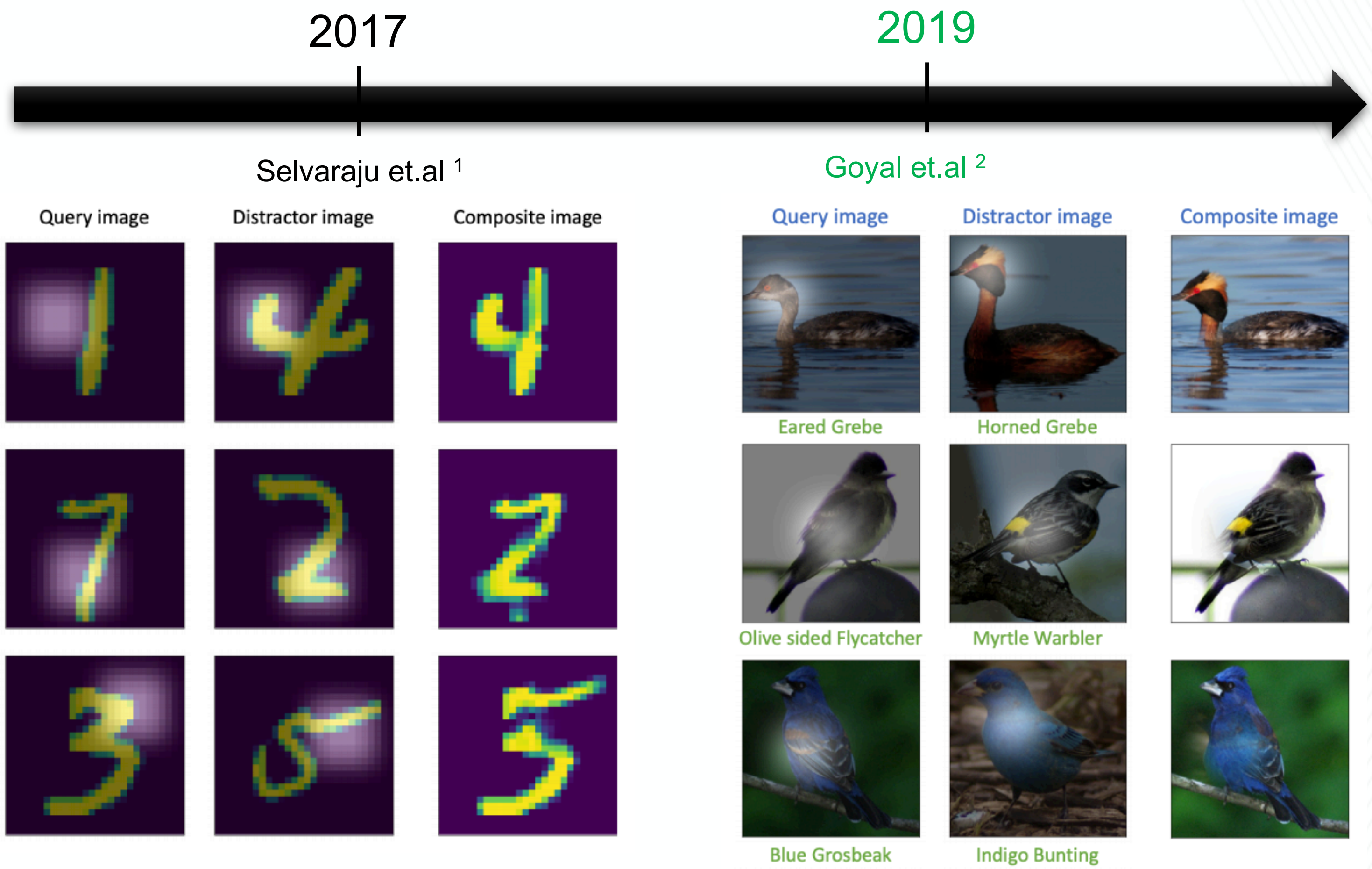


[1] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das,Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” inProceedings of the IEEE internationalconference on computer vision, 2017, pp. 618–626.
[2] Goyal, Yash, et al. "Counterfactual visual explanations." *International Conference on Machine Learning*. PMLR, 2019.

Explanations in Neural Networks

Counterfactual Explanations – Non-Gradient based

‘What if the query image were like the distractor image?’



[1] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das,Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” inProceedings of the IEEE internationalconference on computer vision, 2017, pp. 618–626.
[2] Goyal, Yash, et al. "Counterfactual visual explanations." *International Conference on Machine Learning*. PMLR, 2019.

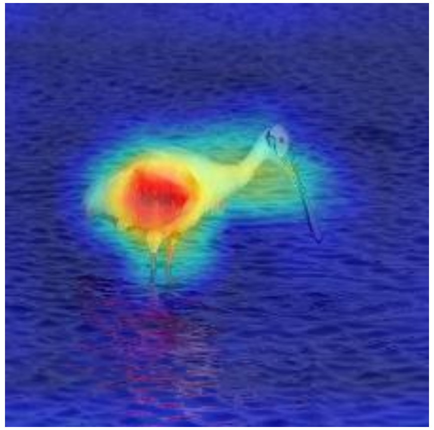
Explanations in Neural Networks

Contrastive Explanations

'Why spoonbill, rather than a Flamingo?'

2013 — Simoyan et.al [1]
2014 — Springenberg et.al [2]
2017 — Selvaraju et.al [3]
2018 — Aditya et.al [4]

`Why Spoonbill?'*** → Convert → ***`Why Spoonbill, rather than Flamingo?'



GradCAM [3]

Convert

Proposed Contrastive Explanation

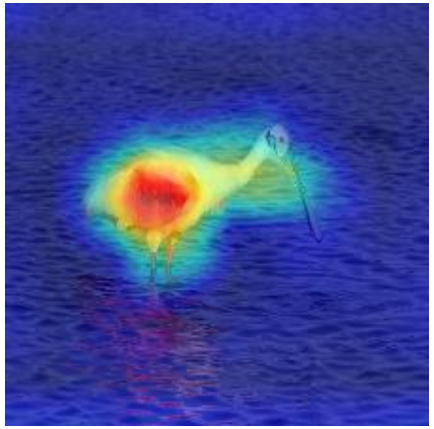
[1] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman,"Deep inside convolutional networks:Visualising imageclassification models and saliency maps,"arXiv preprintarXiv:1312.6034, 2013.
[2] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for Simplicity: The All Convolutional Net. arXiv, 2014
[3] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das,Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," inProceedings of the IEEE internationalconference on computer vision, 2017, pp. 618–626.
[4] 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.

Explanations in Neural Networks

Contrastive Explanations

- *`Why P?'* framework provided by existing methods (In this dissertation proposal, we use Grad-CAM)
- *`Why P, rather than Q?'* provided by gradients between P and Q manifolds

`Why Spoonbill?'*** → Convert → ***`Why Spoonbill, rather than Flamingo?'



GradCAM

Convert →

Proposed Contrastive Explanation

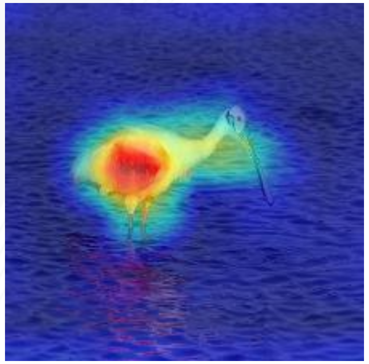
Explanations in Neural Networks
Contrastive Explanations

Implementation : Within Grad-CAM framework

Grad-CAM

```
logit = self.model_arch(input)
#Grad-CAM gradient initialization
if class_idx is None:
    score = logit[:, logit.max(1)[-1]].squeeze()
else:
    score = logit[:, class_idx].squeeze()

self.model_arch.zero_grad()
score.backward(retain_graph=retain_graph)
```



Contrastive Explanation

```
logit = self.model_arch(input)
# The only change to Grad-CAM code
ce_loss = nn.CrossEntropyLoss()
im_label_as_var = Variable(torch.from_numpy(np.asarray([Q])))
pred_loss = ce_loss(logit.cuda(), im_label_as_var.cuda())

self.model_arch.zero_grad()
pred_loss.backward()
```

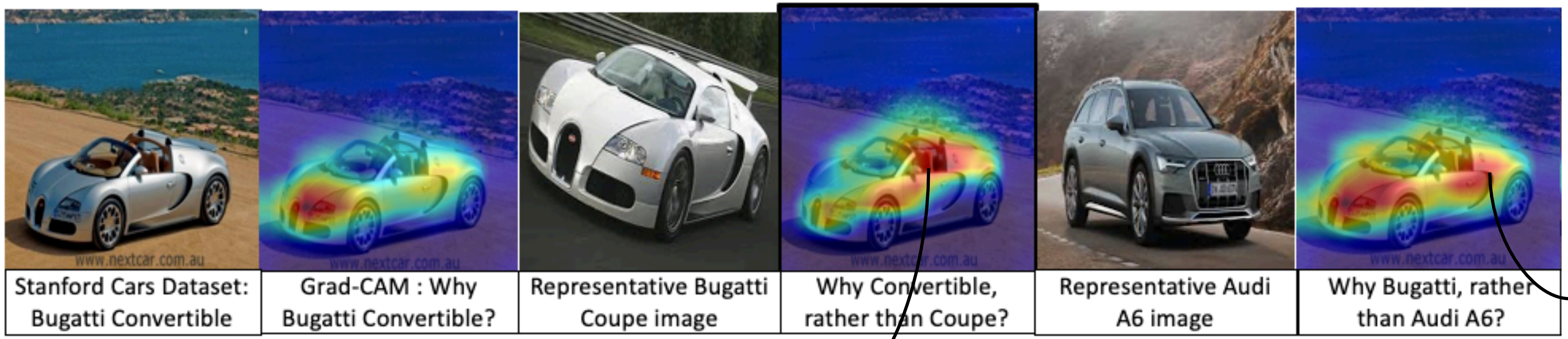
Explanations in Neural Networks
Contrastive Explanations - Examples

Stanford Cars Dataset: Bugatti Convertible	Grad-CAM : Why Bugatti Convertible?	Representative Bugatti Coupe image	Why Convertible, rather than Coupe?	Representative Audi A6 image	Why Bugatti, rather than Audi A6?

- Cars dataset
- VGG-16 Architecture
- Last convolutional layer

M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

Explanations in Neural Networks
Contrastive Explanations - Examples



- Cars dataset
- VGG-16 Architecture
- Last convolutional layer

Highlights the hatchback

Highlights the open top

M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

Explanations in Neural Networks

Why Contrastive Explanations?

CURE-TSR dataset : No-Left Image	Grad-CAM : Why No-Left?	Representative No-Right image	Why No-Left, rather than No-Right?	Representative Stop Sign	Why No-Left, rather than Stop?

- CURE-TSR dataset
- ResNet-18 Architecture
- Last convolutional layer

Not always human interpretable

M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

Explanations in Neural Networks
Why Contrastive Explanations?

CURE-TSR traffic signs

- CURE-TSR dataset
- CNN with 2 convolutional layers
- Last convolutional layer

OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

Explanations in Neural Networks

Why Contrastive Explanations?

Only traffic sign with a straight bottom-left edge – enough to say `Not STOP Sign'

CURE-TSR traffic signs

- CURE-TSR dataset
- CNN with 2 convolutional layers
- Last convolutional layer

Contrastive Explanations provide more context

M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

Explanations in Neural Networks

Why Contrastive Explanations? - IQA

Distorted image

Image Quality Assessment (IQA)

IQA Algorithm[1]

Score : 0.58

Bad Quality

Good Quality

0.0 0.5 1.0

The given image is somewhat OK quality

[1] Bosse S, Maniry D, Müller K R, et al. Deep neural networks for no-reference and full-reference image quality assessment. IEEE Transactions on Image Processing, 2018, 27(1): 206-219.

Explanations in Neural Networks

Why Contrastive Explanations? - IQA

Causal Explanations in IQA

Distorted image

Why 0.58?

Bad Quality

Good Quality

0.0 0.5 1.0

The given image is somewhat OK quality

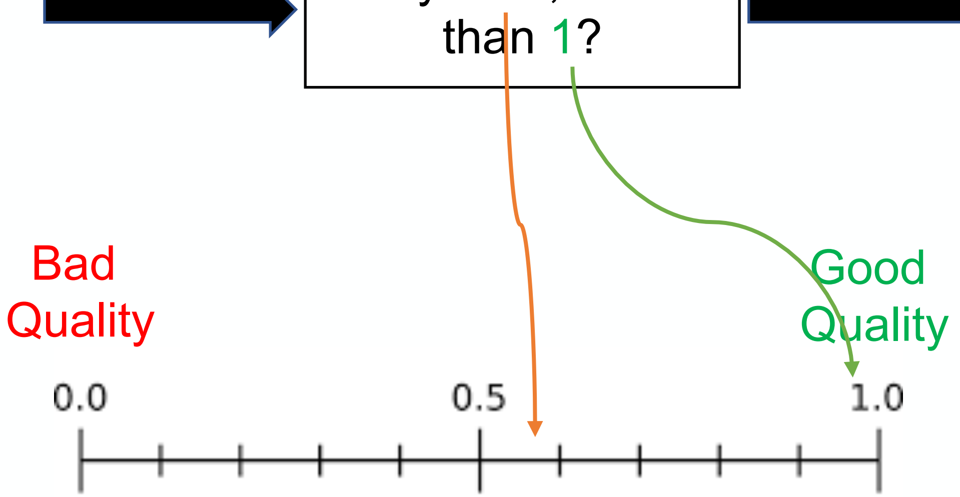
M. Prabhushankar, G. Kwon, D. Temel, and G. AlRegib, "Contrastive Explanations in Neural Networks," in *IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, Oct. 2020.

Explanations in Neural Networks

Why Contrastive Explanations? - IQA

Distorted image

Why 0.58, rather than 1?



The given image is somewhat OK quality

M. Prabhushankar, G. Kwon, D. Temel, and G. AlRegib, "Contrastive Explanations in Neural Networks," in *IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, Oct. 2020.

Explanations in Neural Networks
Why Contrastive Explanations? - IQA

Distorted image

Why 0.58, rather than 0.25?

Bad Quality

Good Quality

The given image is somewhat OK quality

OLIVES @GeorgiaTech

M. Prabhushankar, G. Kwon, D. Temel, and G. AlRegib, "Contrastive Explanations in Neural Networks," in *IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, Oct. 2020.

Georgia Tech CREATING THE NEXT

Explanations in Neural Networks

Why Contrastive Explanations? - IQA

Why 0.58?

- Network parsed the entire image to come up with the score

Why 0.58, rather than x?

- Background is less essential than foreground for higher quality
- Lighthouse is more important than cliff for higher quality
- Presence of sky provides a higher quality to the image

M. Prabhushankar, G. Kwon, D. Temel, and G. AlRegib, "Contrastive Explanations in Neural Networks," in *IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, Oct. 2020.

OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

Explanations in Neural Networks

Why Contrastive Explanations? - IQA

Contrastive explanations provide fine-grained details that add context and relevance to existing explanations

Why 0.58?

- Network parsed the entire image to come up with the score

Why 0.58, rather than x?

- Background is less essential than foreground for higher quality
- Lighthouse is more important than cliff for higher quality
- Presence of sky provides a higher quality to the image

So far,

- We introduced an interpretation of **gradients in the space of models** from a perspective of **model uncertainty**
- We proposed a framework for efficient gradient generation with **confounding labels** to quantify uncertainty of fully trained networks
- We validated that the gradient-based uncertainty measure outperform activation-based features in **out-of-distribution detection** and **corrupted input detection**
- We interpreted gradients as a reasoning mechanism within neural networks
- We showed that gradients can be used to answer three explanatory paradigms. They possess fine-grained details that add context to explanations

https://arxiv.org/abs/2103.12329

https://arxiv.org/abs/2008.00178

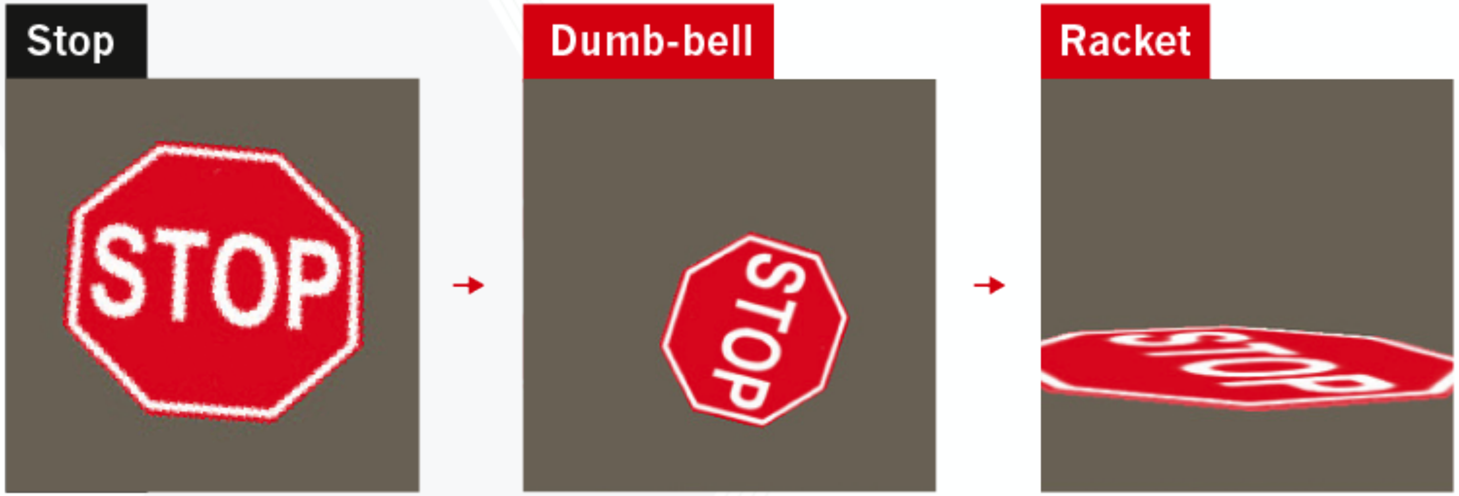
[1] M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

[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.

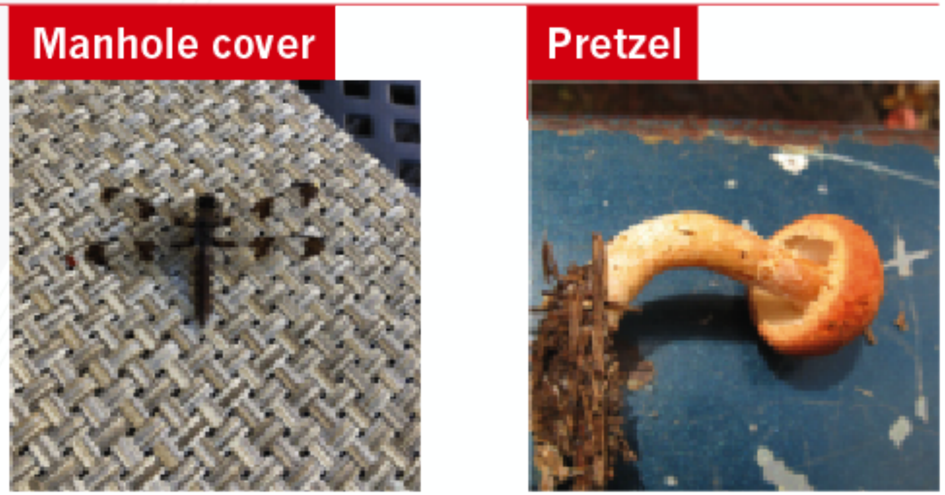
Part V : Robust Machine Learning

Part V : Robust Machine Learning

Rotating objects in an image confuses DNNs, probably because they are too different from the types of image used to train the network.



Even natural images can fool a DNN, because it might focus on the picture's colour, texture or background rather than picking out the salient features a human would recognize.



©nature

OLIVES @GeorgiaTech

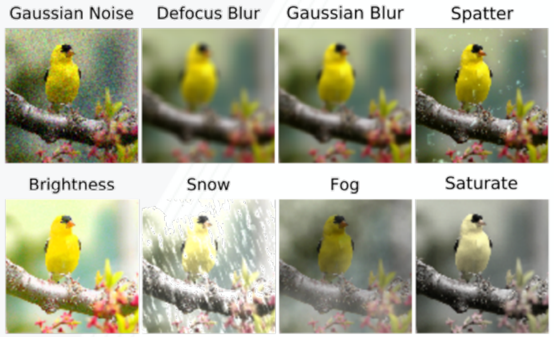
Georgia Tech CREATING THE NEXT

Robust Machine Learning

Part I : Out-of-distribution detection

Goal : Identify images that are from distributions other than the training distributions. Images can belong to the same class.

Ex : Training distribution – CIFAR-10
Testing distribution – CIFAR-10-C

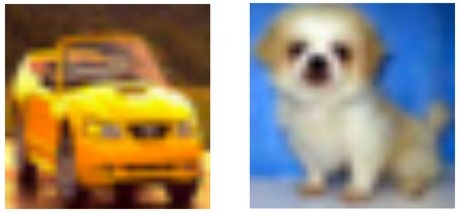


CIFAR-10-C

Part II : Anomaly/Novelty detection

Goal : Identify images that belong to an unseen class, given a trained network

Ex : Training classes – Cars
Testing classes – Dogs



Normal Abnormal

Robust Machine Learning

Part I : Out-of-distribution detection

Goal : Identify images that are from distributions other than the training distributions. Images can belong in the same class

Part II : Anomaly/Novelty detection

Goal : Identify images that belong to an unseen class, given a trained network

Part V : Recognize classes under distortion/domain shift/abnormality

Ex : Training distribution – CIFAR-10
Testing distribution – CIFAR-10-C

CIFAR-10-C

Ex : Training classes – Cars
Testing classes – Dogs

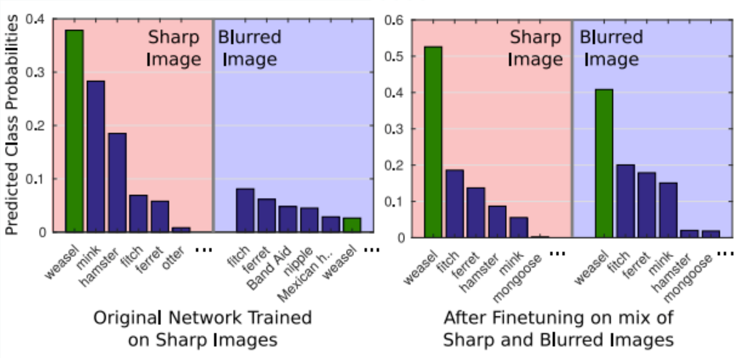
Normal Abnormal

Robust Machine Learning

2016 — Vasiljevic et.al [1]

2017 — Dodge et.al [2]

2017 — Temel et.al [3]

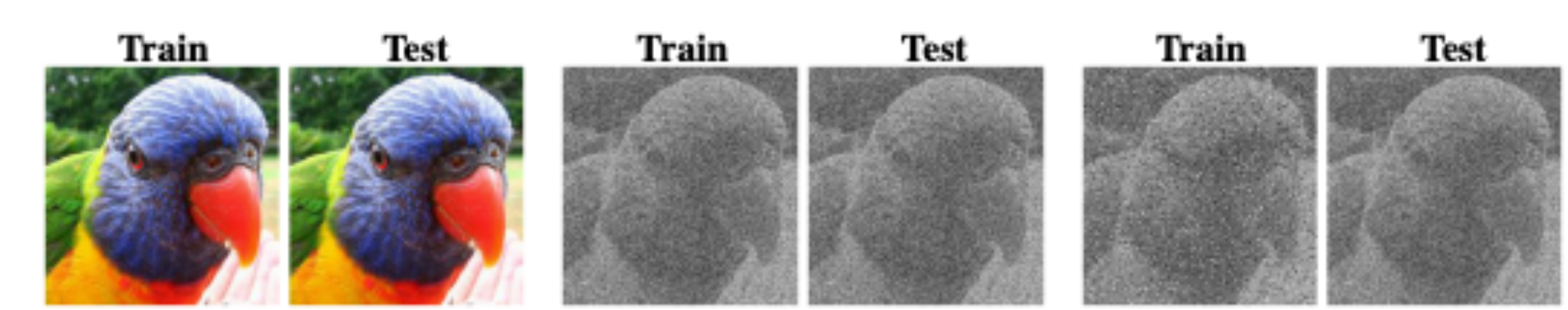


Advocated for training on noisy images

Advocated for training on simulated images

[1] Vasiljevic, I., Chakrabarti, A., and Shakhnarovich, G. Examining the impact of blur on recognition by convolutional networks.arXiv preprint arXiv:1611.05760, 2016.
[2] Dodge, S. and Karam, L. A study and comparison of human and deep learning recognition performance under visual distortions. In2017 26th international conference on computer communication and networks (ICCCN), pp. 1–7. IEEE, 2017.
[3] Temel, D., Kwon, G., Prabhuhankar, M., and AlRegib, G.CURE-TSR: Challenging unreal and real environments for traffic sign recognition. Advances in Neural Information Processing Systems (NIPS) Machine Learning for Intelligent Transportations Systems Workshop, 2017.

Robust Machine Learning



(a) Super-human performance (b) Super-human performance (c) Chance level performance

Train and test noise are same

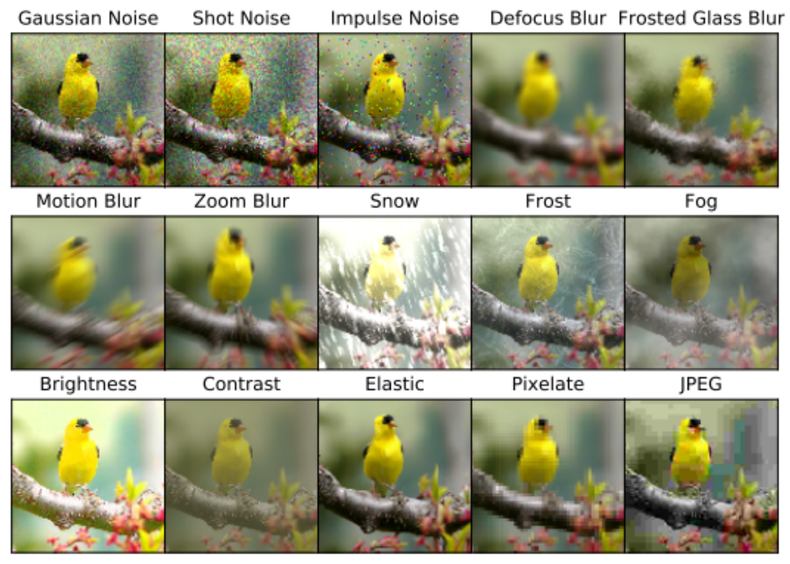
Train and test noise are different

OLIVES @GeorgiaTech

[1] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Sch¨utt,Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. In Advances in Neural Information Processing Systems, pages 7538–7550, 2018.

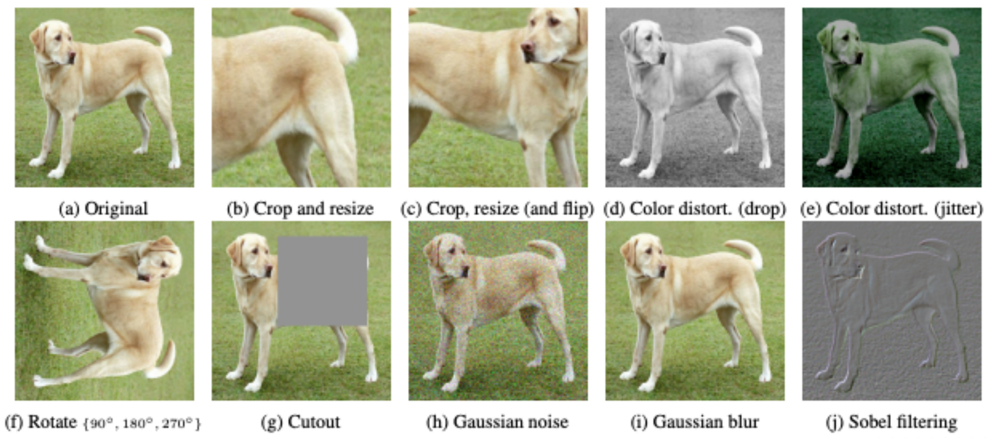
Georgia Tech CREATING THE NEXT

Robust Machine Learning



Advocated for training on adversarial images

Sim-CLR : Simple Contrastive Learning Framework



Self-supervised training with augmentations

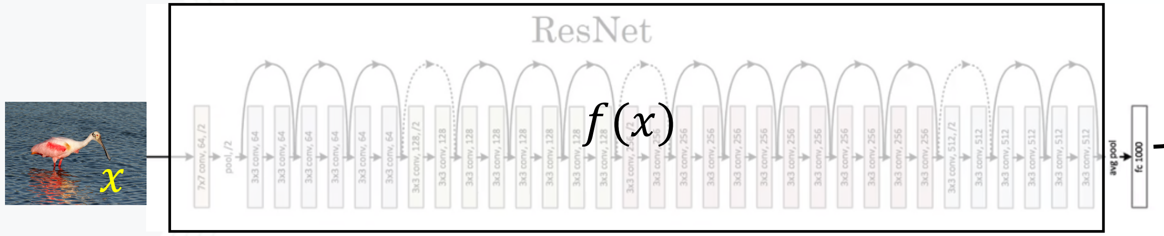
OLIVES @GeorgiaTech

[1] Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations.arXiv preprint arXiv:1903.12261, 2019.
[2] Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *International conference on machine learning*. PMLR, 2020.

Georgia Tech CREATING THE NEXT

Robust Machine Learning
Recognition

Consider a ResNet-18 trained to differentiate between 10 classes



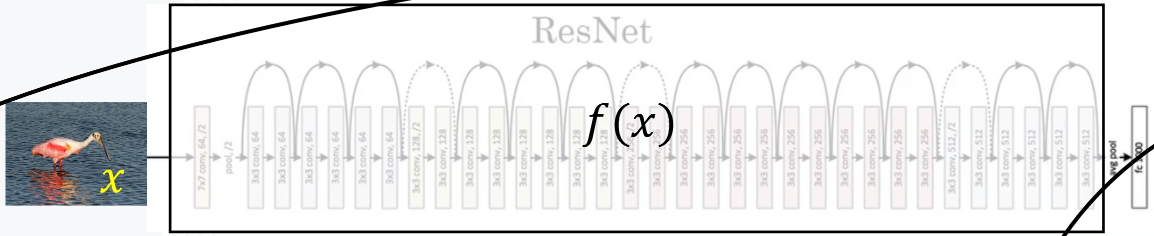
$$\begin{pmatrix} 0.0002 \\ 0.0001 \\ 0.0049 \\ 0.0000 \\ 0.7942 \\ \vdots \\ 0.0005 \end{pmatrix}$$

Well-trained network

$$f(x) = y, y \in R^{\{1 \times 10\}}$$

Robust Machine Learning
Recognition

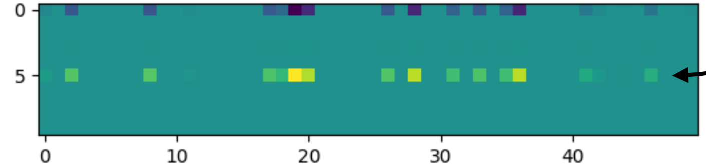
Consider a ResNet-18 trained to differentiate between 10 classes



$$\begin{pmatrix} \boxed{0.0002} \\ 0.0001 \\ 0.0049 \\ 0.0000 \\ \boxed{0.7942} \\ \vdots \\ 0.0005 \end{pmatrix}$$

Well-trained network

$$f(x) = y, y \in R^{\{1 \times 10\}}$$



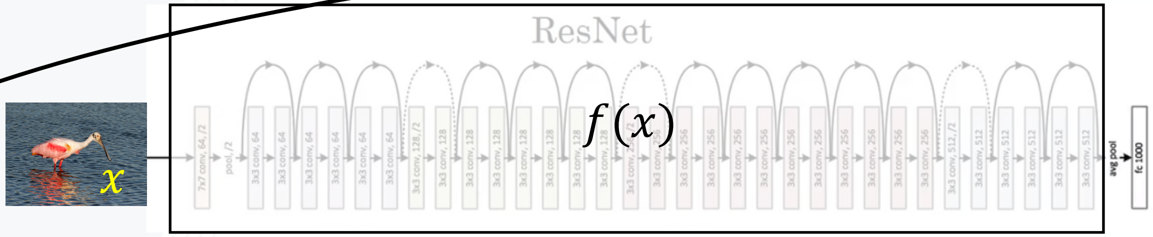
$$\frac{\partial J(5,1)}{\partial W_L} = R^{\{64 \times 1\}}$$

OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

Robust Machine Learning
Recognition

Consider a ResNet-18 trained to differentiate between 10 classes



$$\begin{bmatrix} \boxed{0.0002} \\ 0.0001 \\ 0.0049 \\ 0.0000 \\ \boxed{0.7942} \\ \vdots \\ 0.0005 \end{bmatrix}$$

Well-trained network

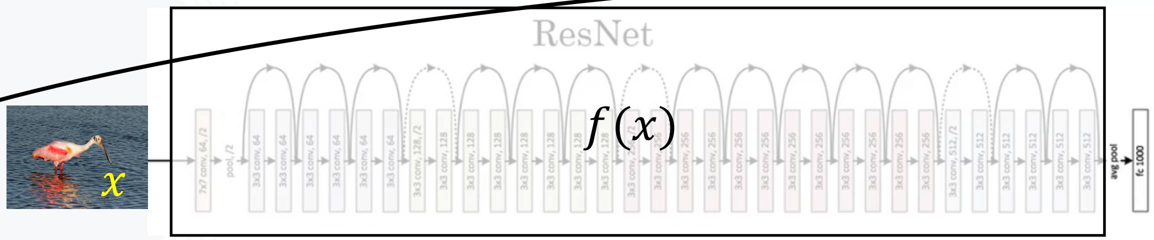
$$f(x) = y, y \in R^{\{1 \times 10\}}$$

$$r_x = \left[\frac{\partial J(5,1)}{\partial W_1}, \qquad \right]$$

Robust Machine Learning

Recognition

Consider a ResNet-18 trained to differentiate between 10 classes



$$\begin{bmatrix} 0.0002 \\ \boxed{0.0001} \\ 0.0049 \\ 0.0000 \\ \boxed{0.7942} \\ \vdots \\ 0.0005 \end{bmatrix}$$

Well-trained network

$$f(x) = y, y \in R^{\{1 \times 10\}}$$

$$r_x = \left[\frac{\partial J(5,1)}{\partial W_1}, \frac{\partial J(5,2)}{\partial W_2}, \right] \ldots$$

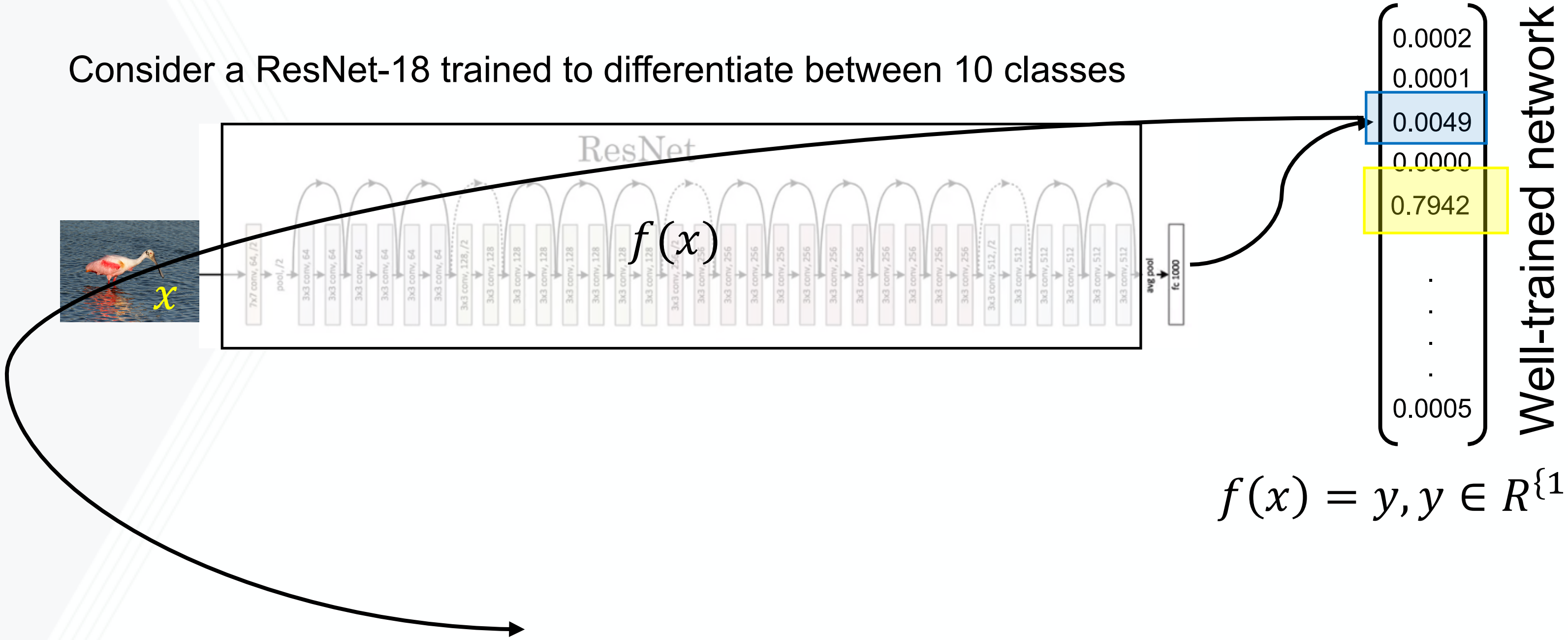
OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

Robust Machine Learning

Recognition

Consider a ResNet-18 trained to differentiate between 10 classes



$$f(x) = y, y \in R^{\{1 \times 10\}}$$

$$r_x = \left[\frac{\partial J(5,1)}{\partial W_1}, \frac{\partial J(5,2)}{\partial W_2}, \frac{\partial J(5,3)}{\partial W_3} \dots \right]$$

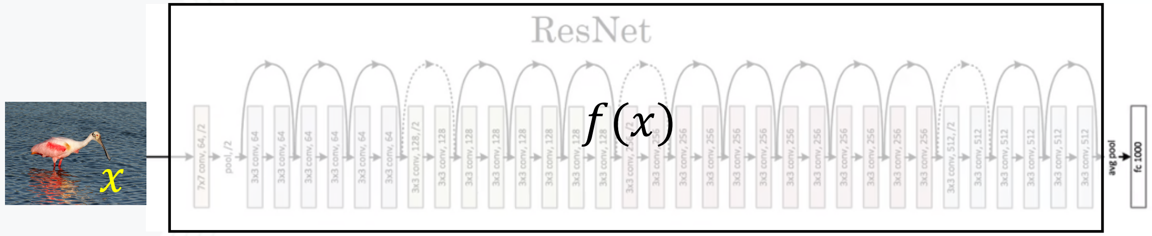
OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

Robust Machine Learning

Recognition

Consider a ResNet-18 trained to differentiate between 10 classes



$$\begin{pmatrix} 0.0002 \\ 0.0001 \\ 0.0049 \\ 0.0000 \\ \boxed{0.7942} \\ \vdots \\ \boxed{0.0005} \end{pmatrix}$$

Well-trained network

$$f(x) = y, y \in R^{\{1 \times 10\}}$$

Concatenated contrastive feature for image x

$$\mathrm{r}_x = R^{\{640 \times 1\}}$$

$$r_x = \left[\frac{\partial J(5,1)}{\partial W_1}, \frac{\partial J(5,2)}{\partial W_2}, \frac{\partial J(5,3)}{\partial W_3} \dots, \frac{\partial J(5,10)}{\partial W_4}\right]$$

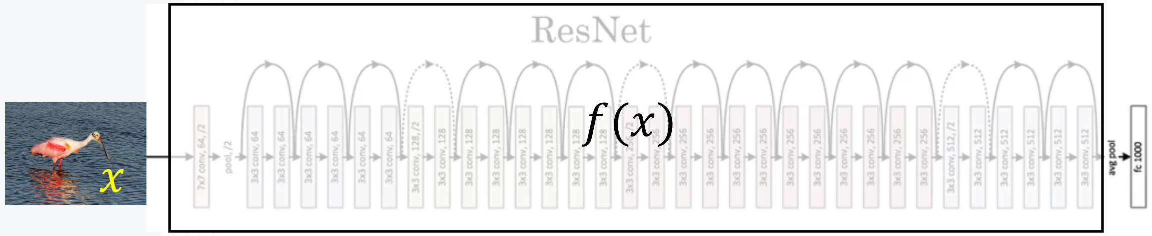
OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

Robust Machine Learning
Recognition

Consider a ResNet-18 trained to differentiate between 10 classes



Well-trained network

$$\begin{bmatrix} 0.0002 \\ 0.0001 \\ 0.0049 \\ 0.0000 \\ \boxed{0.7942} \\ \vdots \\ \boxed{0.0005} \end{bmatrix}$$

$$f(x) = y, y \in R^{\{1 \times 10\}}$$

Concatenated contrastive feature for image x

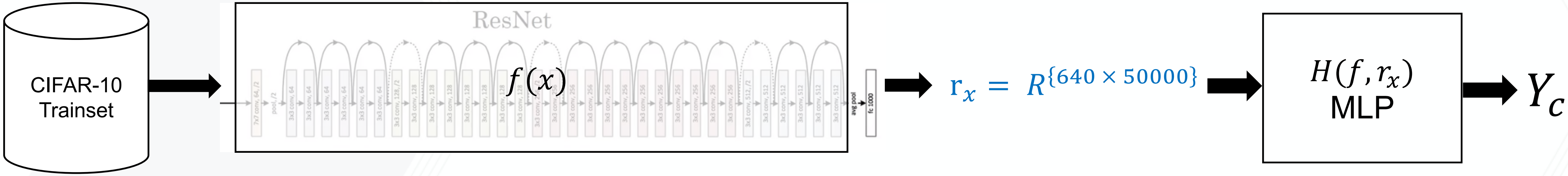
$$r_x = R^{\{640 \times 1\}}$$

$$r_x = \left[\frac{\partial J(5,1)}{\partial W_L}, \frac{\partial J(5,2)}{\partial W_L}, \frac{\partial J(5,3)}{\partial W_L} \ldots, \frac{\partial J(5,10)}{\partial W_L}\right]$$

Contrastive inference based prediction is made on r_x

Robust Machine Learning

Recognition



Y_c= Contrastive Prediction

Robust Machine Learning

Recognition

Feed-Forward Inference

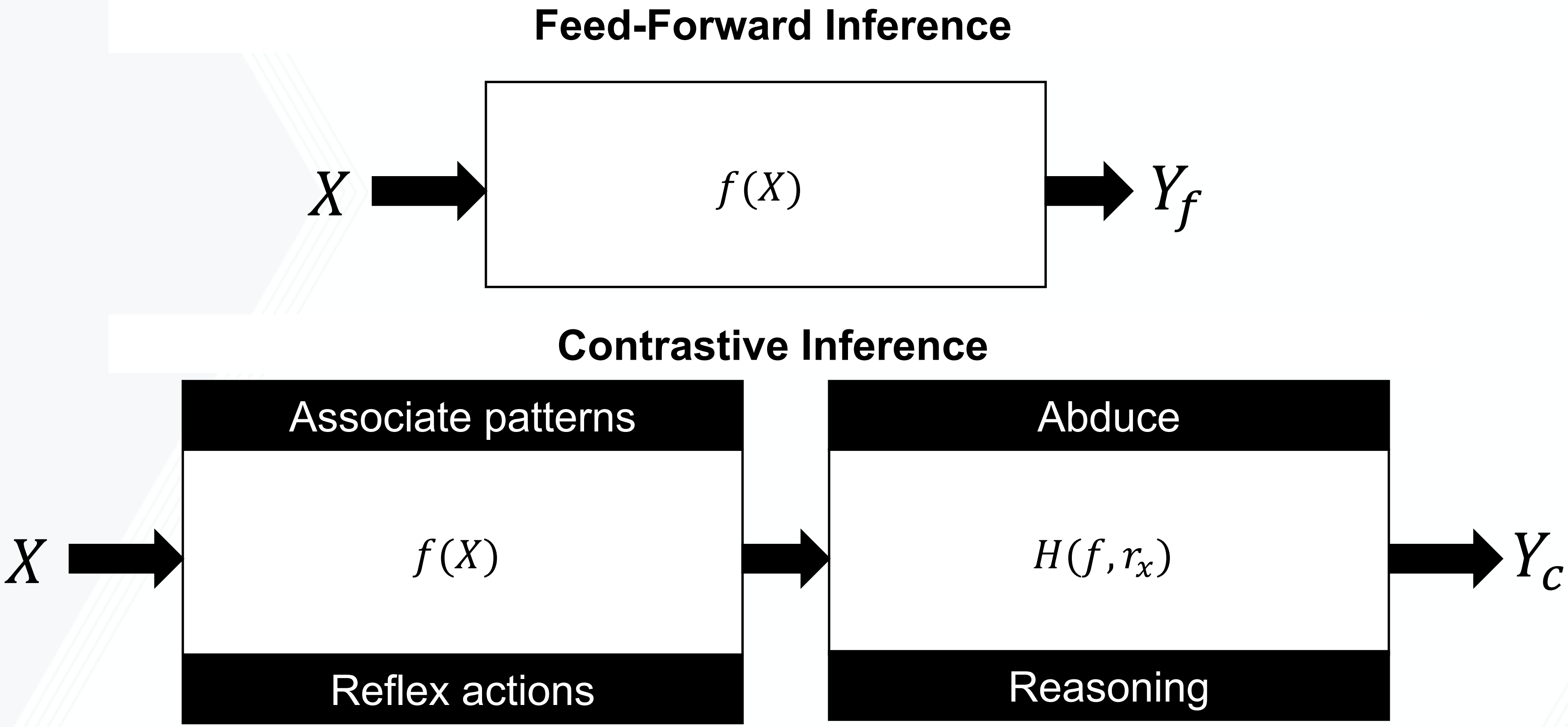
$$X \rightarrow \boxed{f(X)} \rightarrow P = Y_f$$

Contrastive Inference

$$X \rightarrow \boxed{f(X)} \rightarrow \boxed{H(f, r_x)} \rightarrow Y_c$$

Robust Machine Learning

Recognition

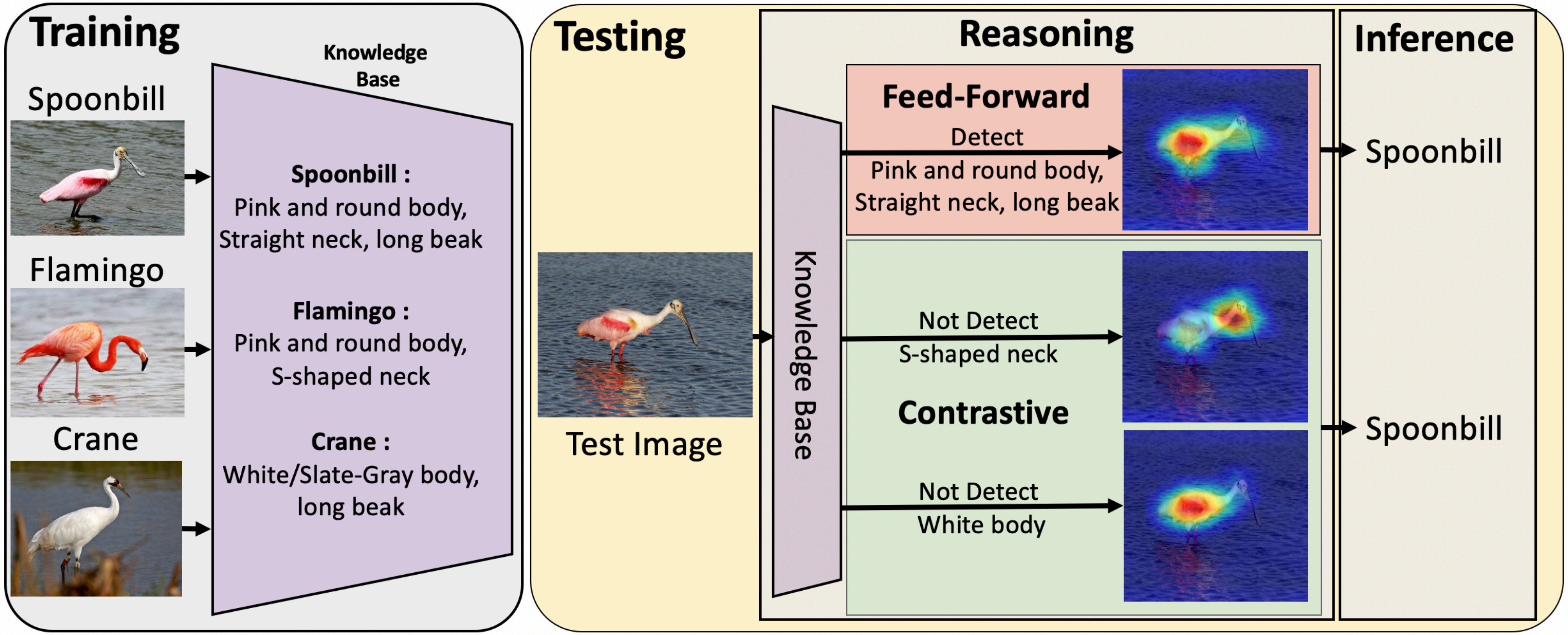


Robust Machine Learning

Recognition

Inductive Reasoning

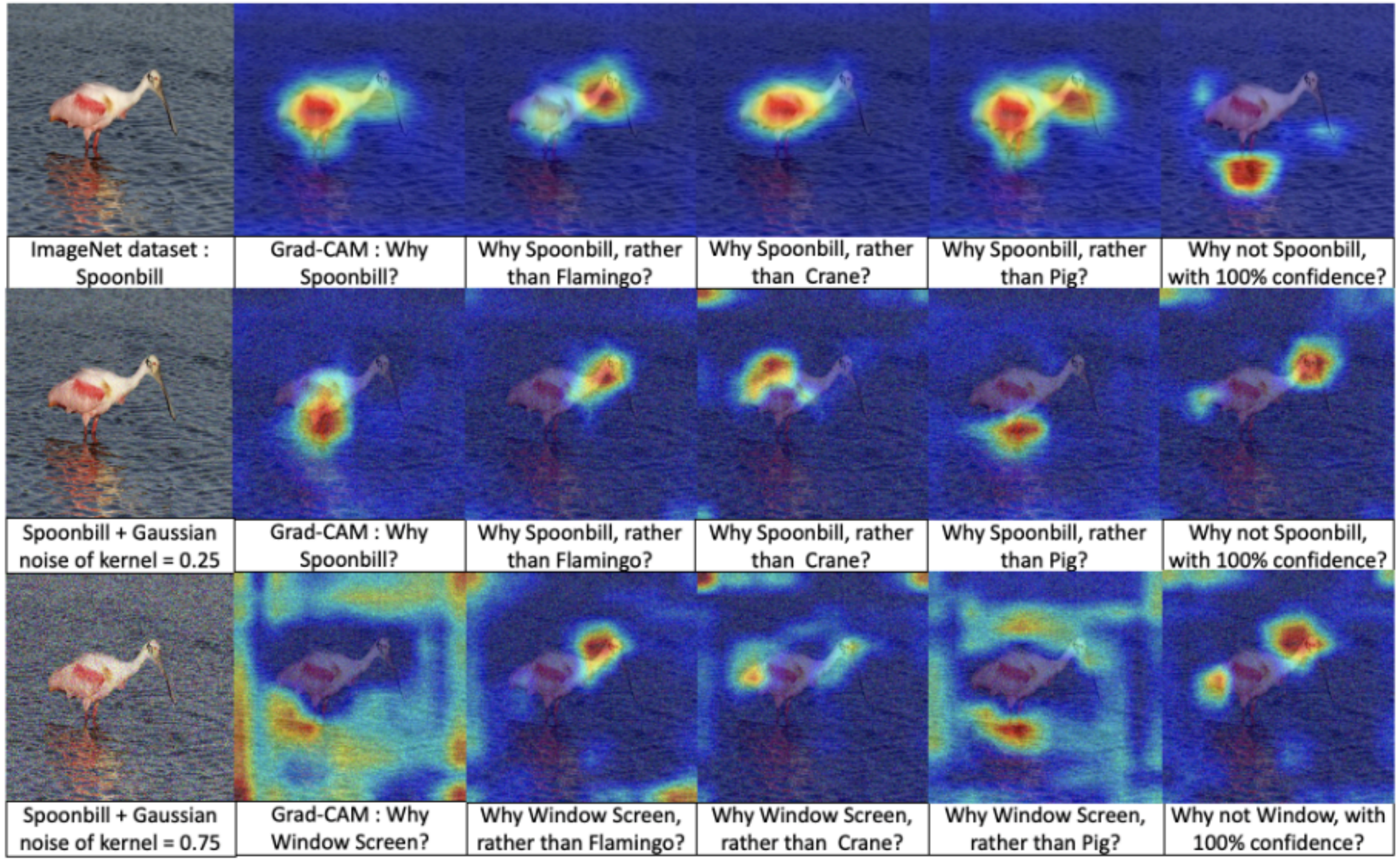
Abductive Reasoning



M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

Robust Machine Learning

Recognition



Robust Machine Learning

Recognition

Networks	Train	Test	Evaluation
• ResNet-18 • ResNet-34 • ResNet-50 • ResNet-101	CIFAR-10 50,000 images	CIFAR-10-C[1] • 19 challenges • 5 Levels in each challenge • Total 950,000 testing images	Recognition accuracy of Feed-forward vs Contrastive Inference

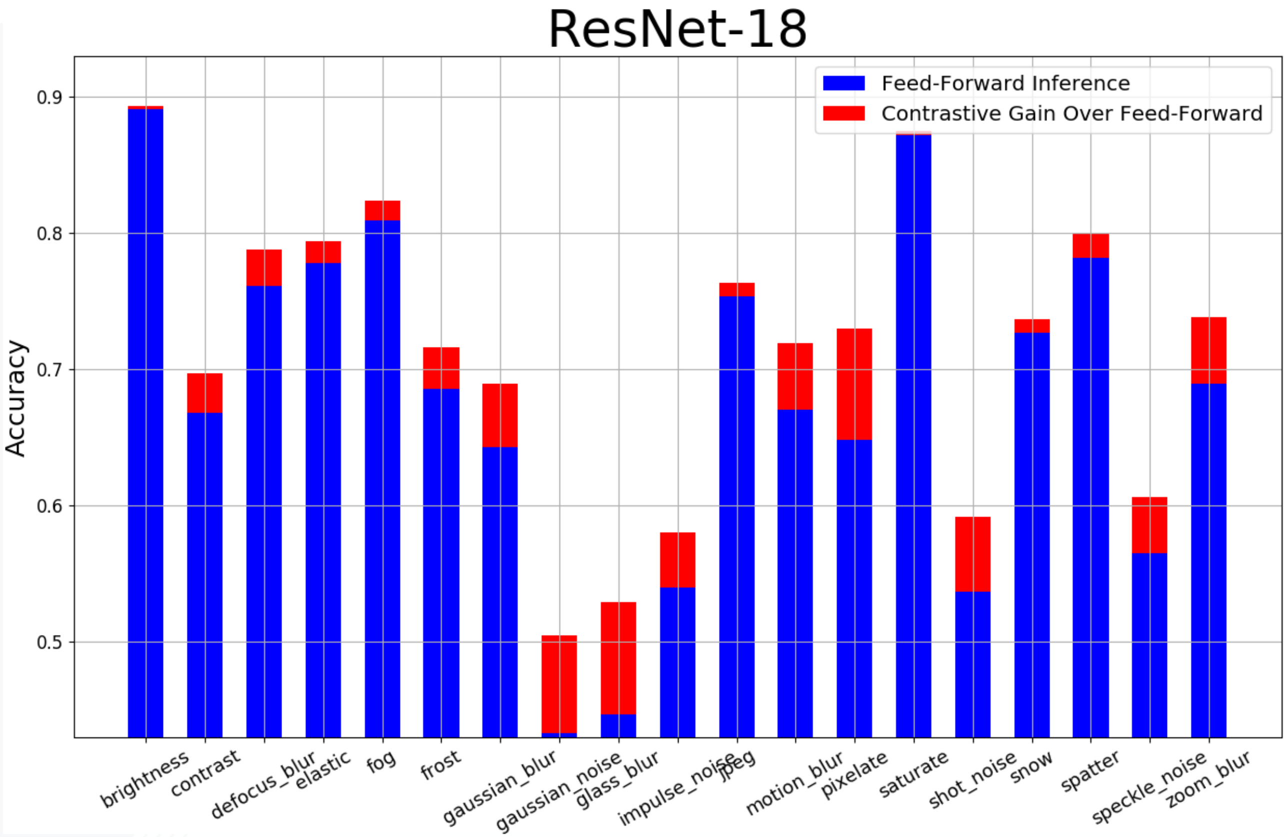
OLIVES @GeorgiaTech

[1] Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations.arXiv preprint arXiv:1903.12261, 2019.

Georgia Tech CREATING THE NEXT

Robust Machine Learning

Recognition



- Blue : Feed-forward accuracy in individual challenge category
- Red : Contrastive gain over Feed-Forward

- Classification accuracy on all 950,000 test images : 67.89%
- Classification accuracy on all 950,000 test images : 71.58%

- With knowledge of noise mean and standard deviation, results increase to 75%

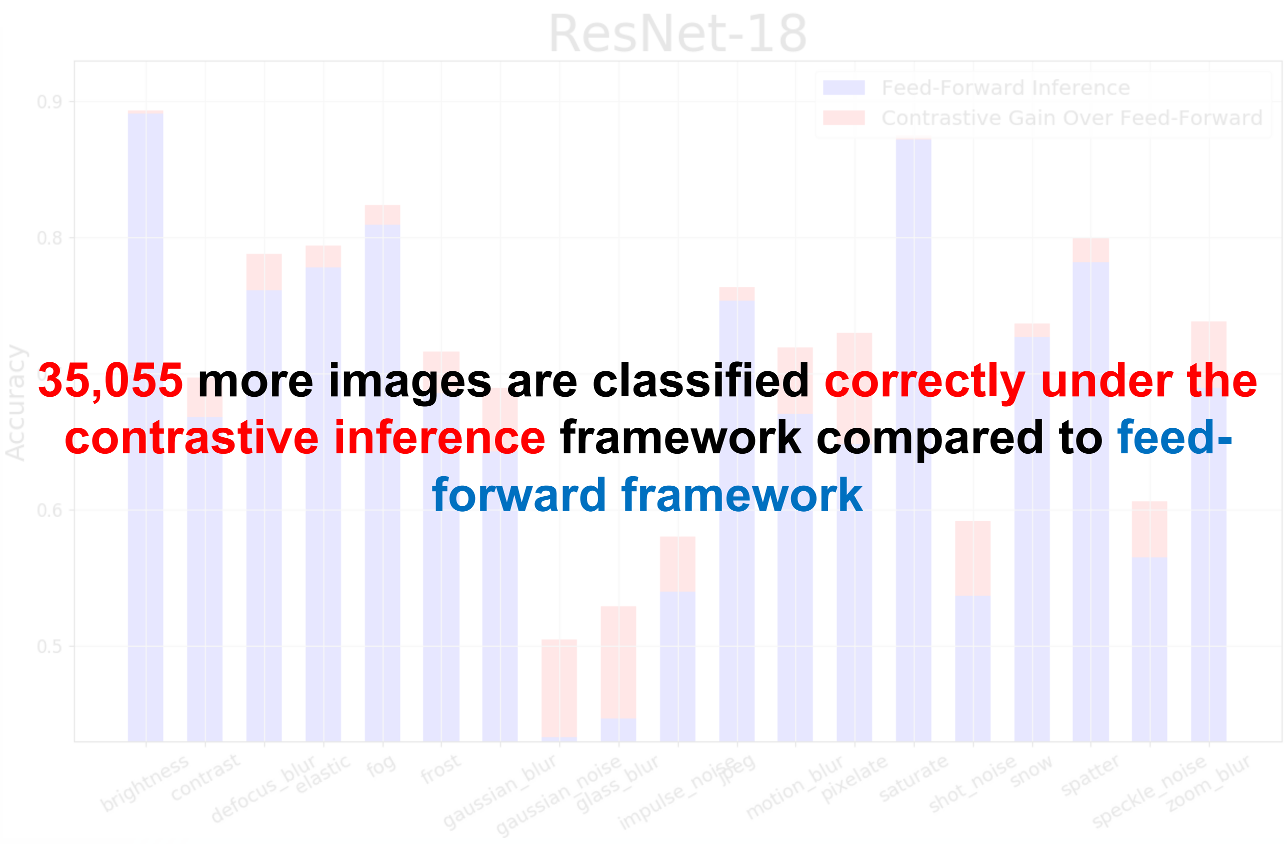
OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

Robust Machine Learning

Recognition



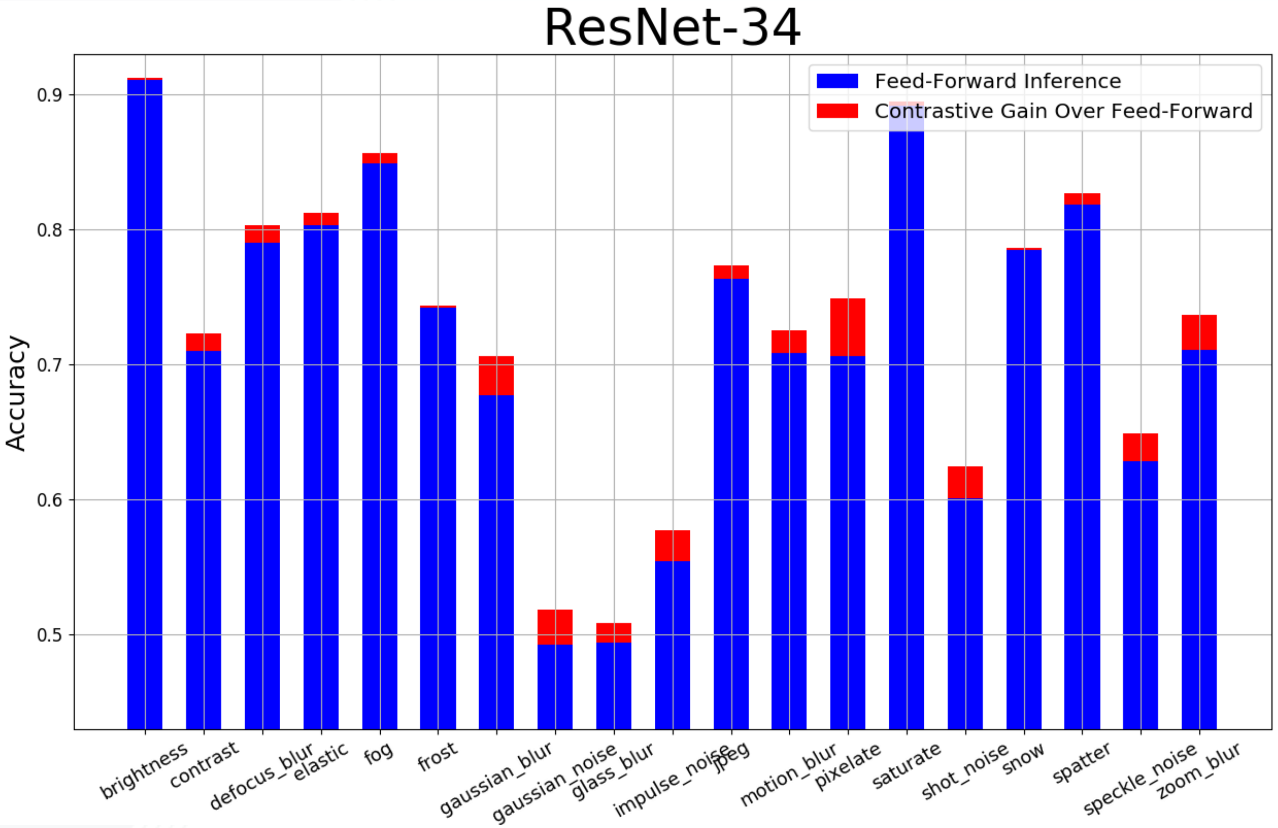
35,055 more images are classified correctly under the contrastive inference framework compared to feed-forward framework

- Blue : Feed-forward accuracy in individual challenge category
- Red : Contrastive gain over Feed-Forward

- Classification accuracy on all 950,000 test images : 67.89%
- Classification accuracy on all 950,000 test images : 71.58%

M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

Robust Machine Learning
Recognition



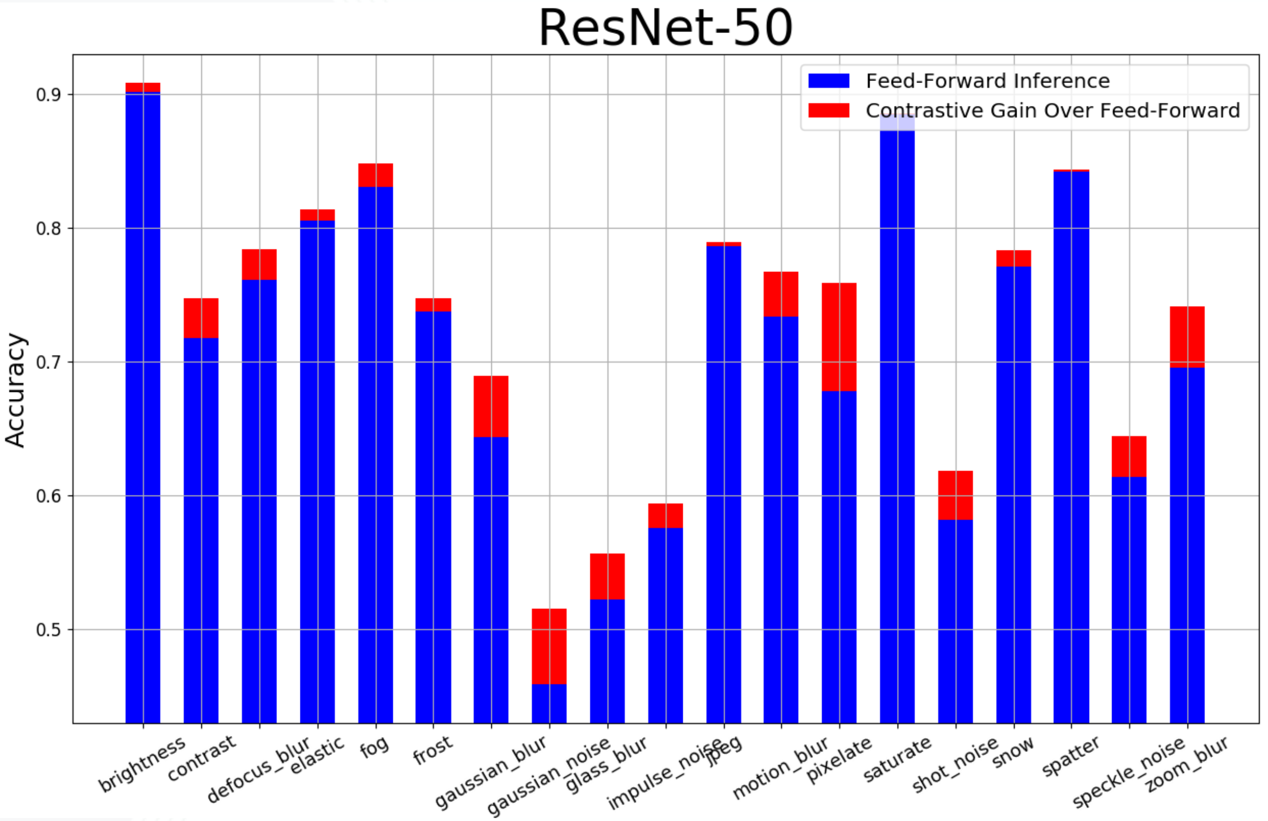
- Blue : Feed-forward accuracy in individual challenge category
- Red : Contrastive gain over Feed-Forward

- Classification accuracy on all 950,000 test images : 71.77%
- Classification accuracy on all 950,000 test images : 73.21%

M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

Robust Machine Learning

Recognition

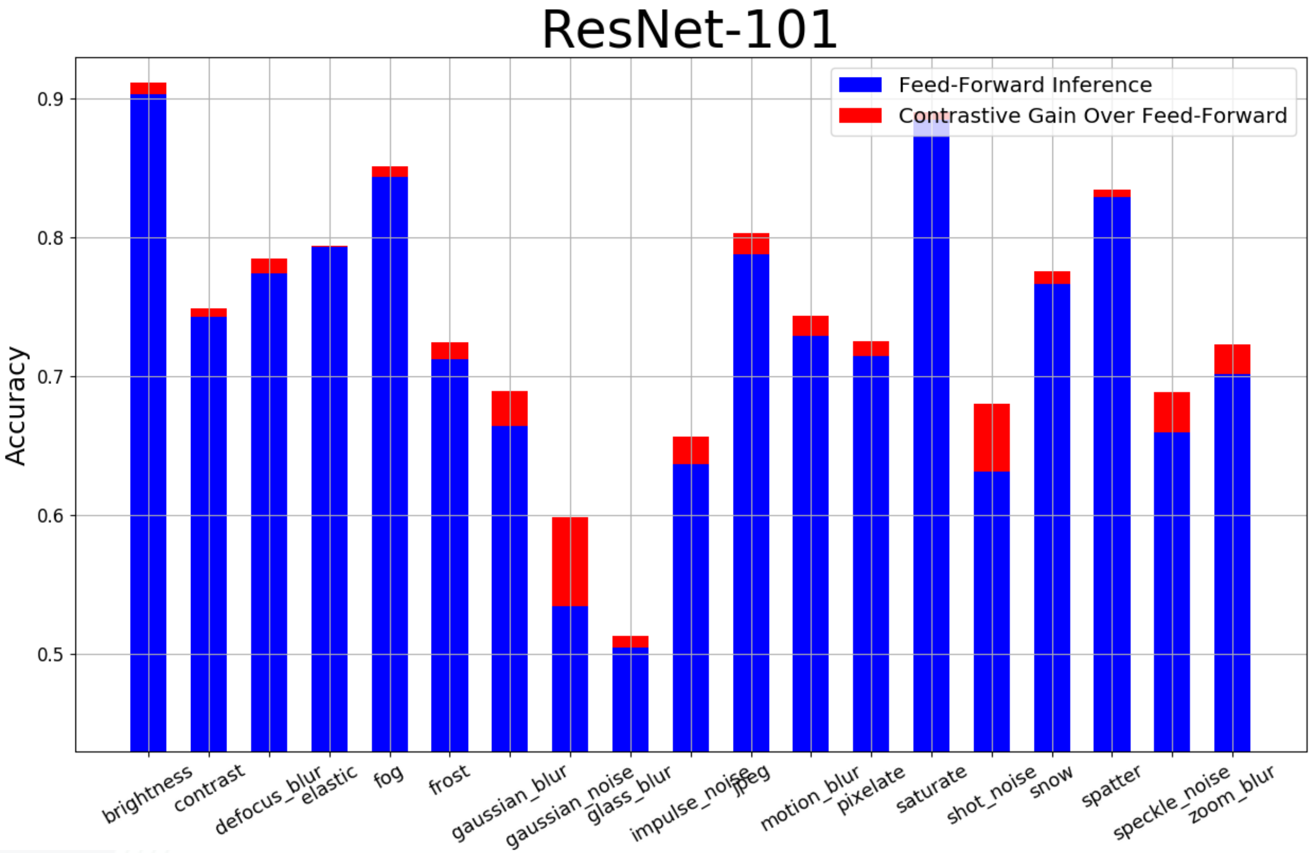


- Blue : Feed-forward accuracy in individual challenge category
- Red : Contrastive gain over Feed-Forward

- Classification accuracy on all 950,000 test images : 71.4%
- Classification accuracy on all 950,000 test images : 74.02%

M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

Robust Machine Learning
Recognition



- Blue : Feed-forward accuracy in individual challenge category
- Red : Contrastive gain over Feed-Forward

- Classification accuracy on all 950,000 test images : 72.54%
- Classification accuracy on all 950,000 test images : 74.31%

M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021.

Robust Machine Learning
Recognition – Domain Adaptation

Networks	Train	Test	Evaluation
• ResNet-18 • ResNet-34 • ResNet-50 • ResNet-101	CIFAR-10, Office Dataset	STL, Office Dataset	Recognition accuracy of Feed-forward vs Contrastive Inference

[1] Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations.arXiv preprint arXiv:1903.12261, 2019.

Robust Machine Learning

Recognition – Domain Adaptation

Table 1. Performance of Proposed CiNN vs Feed-Forward Inference under Classical Domain Shift

Architectures		CIFAR-10 ↓ STL	DSLR ↓ Amazon	DSLR ↓ Webcam	Amazon ↓ DSLR	Amazon ↓ Webcam	Webcam ↓ DSLR	Webcam ↓ Amazon
ResNet-18	Feed-Forward	63.7	39.1	78	62.9	59	89.8	42.2
(%)	Contrastive	**78.5**	**47**	**90.7**	**67.3**	**63.9**	**96**	**44**
ResNet-34	Feed-Forward	65.4	41.8	83.3	**67.3**	60.1	90.6	41.7
(%)	Contrastive	**79.4**	**46.4**	**89.8**	**67.3**	**63.9**	**97.8**	**43.3**
ResNet-50	Feed-Forward	67.4	-	-	67.3	62	92.4	33.4
(%)	Contrastive	**80.9**	-	-	**78.1**	**68.4**	**97.8**	30.8
ResNet-101	Feed-Forward	67	-	-	62.9	59	89.8	31.77
(%)	Contrastive	**79.4**	-	-	**76.5**	**67.3**	**92.4**	**33.6**

So Far,

- We introduced an interpretation of **gradients in the space of models** from a perspective of **model uncertainty**
- We proposed a framework for efficient gradient generation with **confounding labels** to quantify uncertainty of fully trained networks
- We validated that the gradient-based uncertainty measure outperform activation-based features in **out-of-distribution detection** and **corrupted input detection**
- We interpreted gradients as a reasoning mechanism within neural networks
- We showed that gradients can be used to answer three explanatory paradigms
- Gradients as features can be used to create robust neural networks as a plug-in on top of existing neural networks

arXiv

https://arxiv.org/abs/2103.12329

arXiv

https://arxiv.org/abs/2008.00178

Robust Machine Learning

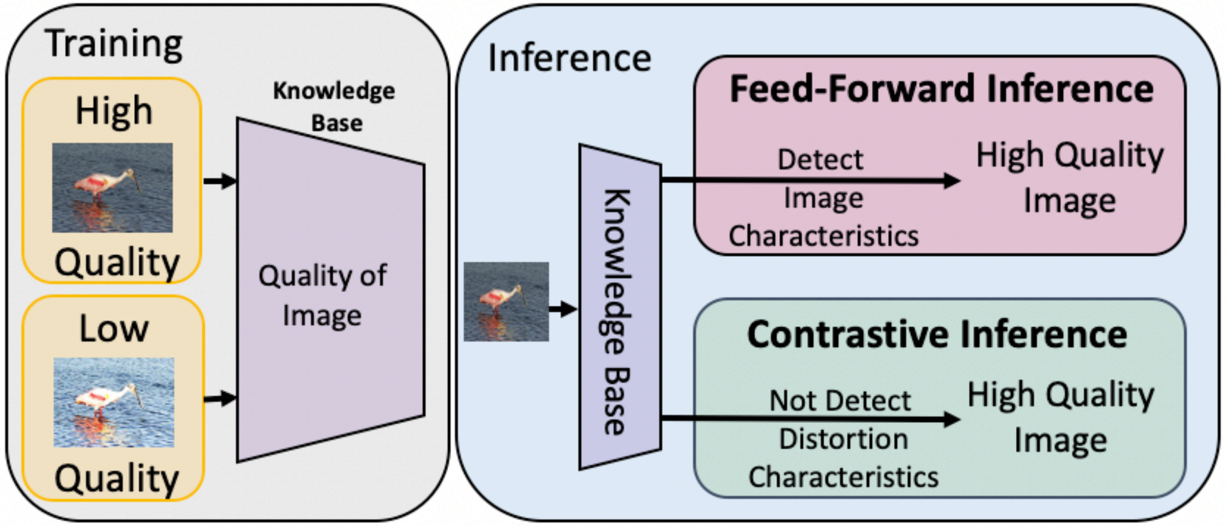
Image Quality Assessment

Image Quality Assessment

Given the pristine image on the left, humans are asked to subjectively quantify the quality of the noisy image on the right

Goal : To objectively assess the subjective quality of an image

Image Quality Assessment



Detect noise characteristics to obtain subjective IQA

OLIVES @GeorgiaTech

G. Kwon*, M. Prabhushankar*, D. Temel, and G. AlRegib , "Distorted Representation Space Characterization Through Backpropagated Gradients," 2019 26th IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019. (*: equal contribution)

Georgia Tech CREATING THE NEXT

Robust Machine Learning

Image Quality Assessment

Plug-in contrastive gradients in an existing IQA framework – UNIQUE [1]

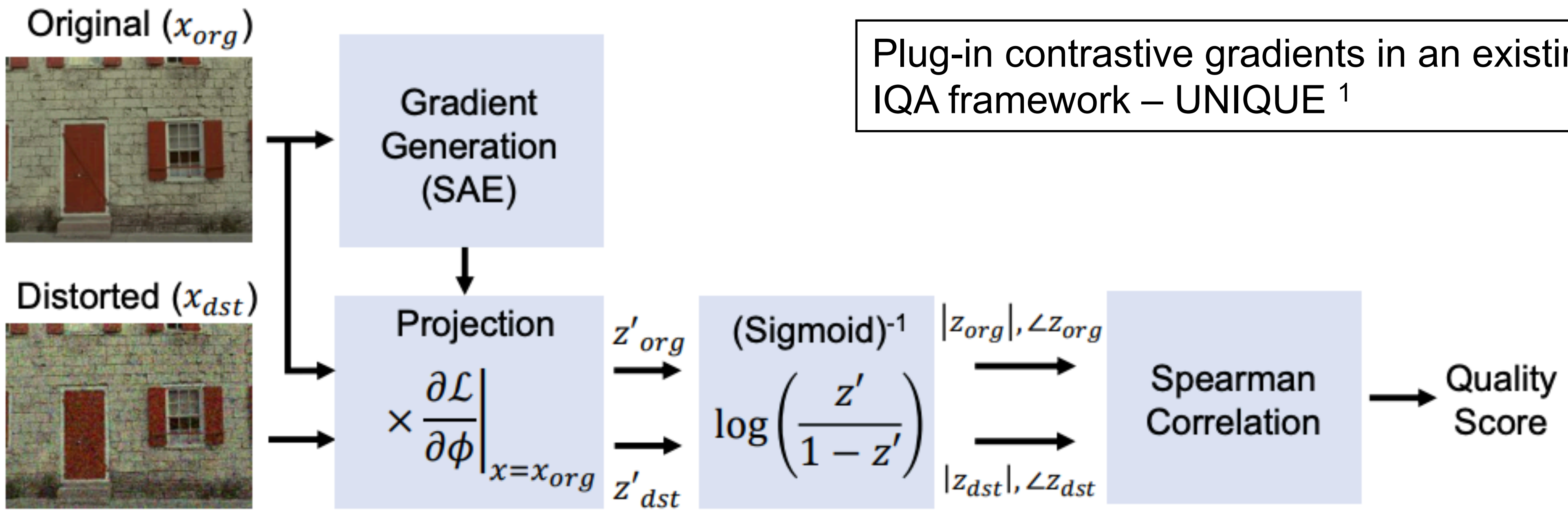


Fig. 4. Block diagram for image quality assessment.

[1] 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.

[2] G. Kwon*, M. Prabhushankar*, D. Temel, and G. AlRegib , "Distorted Representation Space Characterization Through Backpropagated Gradients," 2019 26th IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019. (*: equal contribution)

Robust Machine Learning

Image Quality Assessment

Contrastive

Table 1. Overall performance of image quality estimators.

Databases	PSNR HA [25]	PSNR HMA [25]	SSIM [26]	MS SSIM [27]	CW SSIM [28]	IW SSIM [29]	SR SIM [30]	FSIM [31]	FSIMc [31]	BRIS QUE [32]	BIQI [14]	BLII NDS2 [15]	Per SIM [33]	CSV [34]	UNI QUE [17]	COHER ENSI [35]	SUMMER [35]	Proposed
Outlier Ratio (OR)																		
MULTI	0.013	0.009	0.016	0.013	0.093	0.013	**0.000**	0.018	0.016	0.067	0.024	0.078	0.004	**0.000**	**0.000**	0.031	**0.000**	**0.000**
TID13	**0.615**	0.670	0.734	0.743	0.856	0.701	0.632	0.742	0.728	0.851	0.856	0.852	0.655	0.687	0.640	0.833	**0.620**	**0.620**
Root Mean Square Error (RMSE)																		
MULTI	11.320	10.785	11.024	11.275	18.862	10.049	**8.686**	10.866	10.794	15.058	12.744	17.419	9.898	9.895	9.258	14.806	8.212	**7.943**
TID13	0.652	0.697	0.762	0.702	1.207	0.688	**0.619**	0.710	0.687	1.100	1.108	1.092	0.643	0.647	0.615	1.049	0.630	**0.596**
Pearson Linear Correlation Coefficient (PLCC)																		
MULTI	0.801	0.821	0.813	0.803	0.380	0.847	0.888	0.818	0.821	0.605	0.739	0.389	0.852	0.852	0.872	0.622	**0.901**	**0.908**
	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	
TID13	0.851	0.827	0.789	0.830	0.227	0.832	0.866	0.820	0.832	0.461	0.449	0.473	0.855	0.853	**0.869**	0.533	0.861	**0.877**
	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	
Spearman's Rank Correlation Coefficient (SRCC)																		
MULTI	0.715	0.743	0.860	0.836	0.631	**0.884**	0.867	0.864	0.867	0.598	0.611	0.386	0.818	0.849	0.867	0.554	**0.884**	**0.887**
	-1	-1	0	-1	-1	0	0	0	0	-1	-1	-1	-1	-1	0	-1	0	
TID13	0.847	0.817	0.742	0.786	0.563	0.778	0.807	0.802	0.851	0.414	0.393	0.396	0.854	0.846	**0.860**	0.649	0.856	**0.865**
	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	0	-1	0	
Kendall's Rank Correlation Coefficient (KRCC)																		
MULTI	0.532	0.559	0.669	0.644	0.457	**0.702**	0.678	0.673	0.677	0.420	0.440	0.268	0.624	0.655	0.679	0.399	0.698	**0.702**
	-1	-1	0	0	-1	0	0	0	0	-1	-1	-1	-1	0	0	-1	0	
TID13	0.666	0.630	0.559	0.605	0.404	0.598	0.641	0.629	0.667	0.286	0.270	0.277	**0.678**	0.654	0.667	0.474	0.667	**0.677**
	0	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	0	0	0	-1	0	

Feed-Forward

G. Kwon*, M. Prabhushankar*, D. Temel, and G. AlRegib , "Distorted Representation Space Characterization Through Backpropagated Gradients," 2019 26th IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019. (*: equal contribution)

Robust Machine Learning

Image Quality Assessment

Contrastive

Table 1. Overall performance of image quality estimators.

Databases	PSNR HA [25]	PSNR HMA [25]	SSIM [26]	MS SSIM [27]	CW SSIM [28]	IW SSIM [29]	SR SIM [30]	FSIM [31]	FSIMc [31]	BRIS QUE [32]	BIQI [14]	BLII NDS2 [15]	Per SIM [33]	CSV [34]	UNI QUE [17]	COHER ENSI [35]	SUMMER [35]	Proposed
									Outlier Ratio (OR)									
MULTI	0.013	0.009	0.016	0.013	0.093	0.013	**0.000**	0.018	0.016	0.067	0.024	0.078	0.004	**0.000**	**0.000**	0.031	**0.000**	**0.000**
TID13	**0.615**	0.670	0.734	0.743	0.856	0.701	0.632	0.742	0.728	0.851	0.856	0.852	0.655	0.687	0.640	0.833	**0.620**	**0.620**
									Root Mean Square Error (RMSE)									
MULTI	11.320	10.785	11.024	11.275	18.862	10.049	**8.686**	10.866	10.794	15.058	12.744	17.419	9.898	9.895	9.258	14.806	8.212	**7.943**
TID13	0.652	0.697	0.762	0.702	1.207	0.688	**0.619**	0.710	0.687	1.100	1.108	1.092	0.643	0.647	0.615	1.049	0.630	**0.596**
									Pearson Linear Correlation Coefficient (PLCC)									
MULTI	0.801	0.821	0.813	0.803	0.380	0.847	0.888	0.818	0.821	0.605	0.739	0.389	0.852	0.852	0.872	0.622	**0.901**	**0.908**
	-1	-1	-1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	
TID13	0.851	0.827	0.789	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.861	**0.877**
	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	
									Spearman's Rank Correlation Coefficient (SRCC)									
MULTI	0.715	0.743	0.860	0.836	0.631	**0.884**	0.867	0.864	0.867	0.598	0.611	0.386	0.818	0.849	0.867	0.554	**0.884**	**0.887**
	-1	-1	0	-1	-1	0	0	0	0	-1	-1	-1	-1	-1	0	-1	0	
TID13	0.847	0.817	0.742	0.786	0.563	0.778	0.807	0.802	0.851	0.414	0.393	0.396	0.854	0.846	**0.860**	0.649	0.856	**0.865**
	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	0	-1	0	
									Kendall's Rank Correlation Coefficient (KRCC)									
MULTI	0.532	0.559	0.669	0.644	0.457	**0.702**	0.678	0.673	0.677	0.420	0.440	0.268	0.624	0.655	0.679	0.399	0.698	**0.702**
	-1	-1	0	0	-1	0	0	0	0	-1	-1	-1	-1	0	0	-1	0	
TID13	0.666	0.630	0.559	0.605	0.404	0.598	0.641	0.629	0.667	0.286	0.270	0.277	**0.678**	0.654	0.667	0.474	0.667	**0.677**
	0	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	0	0	0	-1	0	

Contrastive features can be used as plug-in into existing IQA detectors

Feed-Forward

OLIVES @GeorgiaTech

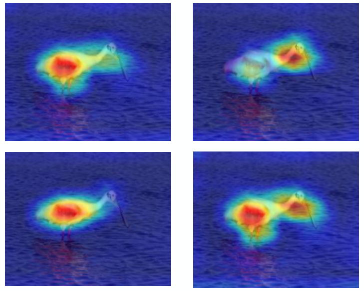
G. Kwon*, M. Prabhushankar*, D. Temel, and G. AlRegib , "Distorted Representation Space Characterization Through Backpropagated Gradients," 2019 26th IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019. (*: equal contribution)

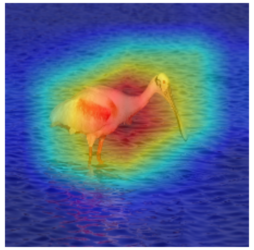
Georgia Tech CREATING THE NEXT

Robust Machine Learning
Human Visual Saliency

Human-Visual Saliency

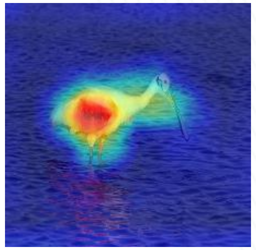
Goal : Given an image, predict likely human eye fixation





Correlation of contrastive explanations to eye tracking data

>



Correlation of Grad-CAM explanations to eye tracking data

Hypothesis : Contrastive regions draw human gaze

OLIVES @GeorgiaTech

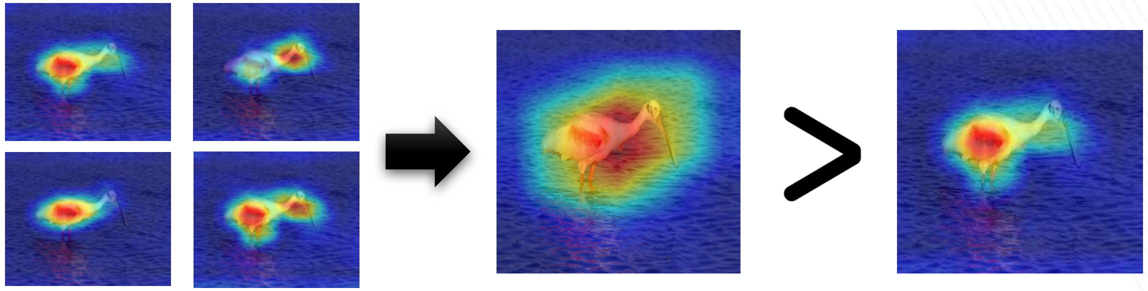
Georgia Tech CREATING THE NEXT

Robust Machine Learning

Human Visual Saliency

Human-Visual Saliency

Goal : Given an image, predict likely human eye fixation



Correlation of contrastive explanations to eye tracking data

Correlation of Grad-CAM explanations to eye tracking data

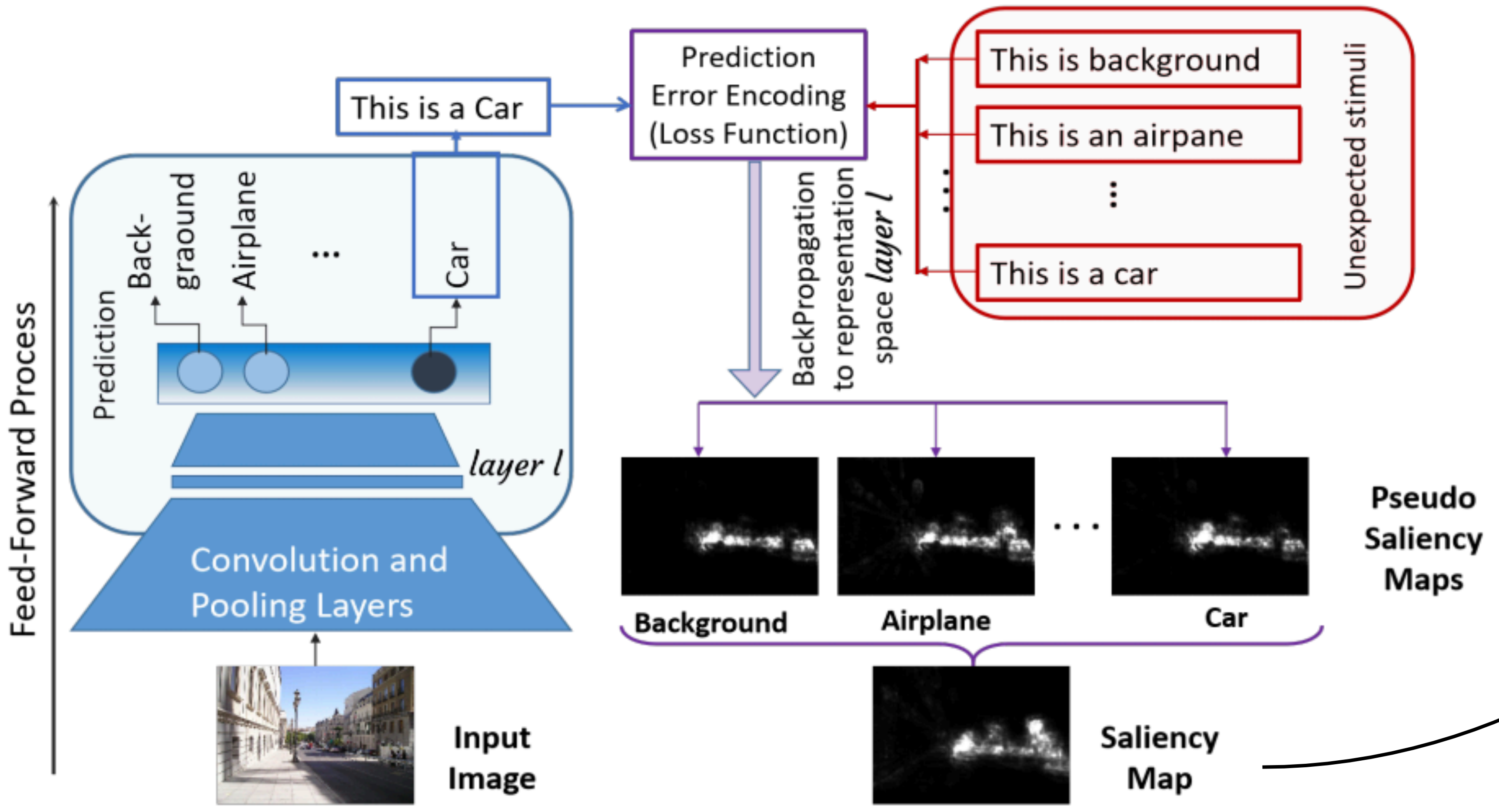
To show : Human eye fixation data on MIT 1003 dataset is more correlated with contrastive explanations than Grad-CAM

OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

Robust Machine Learning

Human Visual Saliency



Implicit Saliency in recognition neural networks! No training on eye tracking data

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.

Robust Machine Learning

Human Visual Saliency

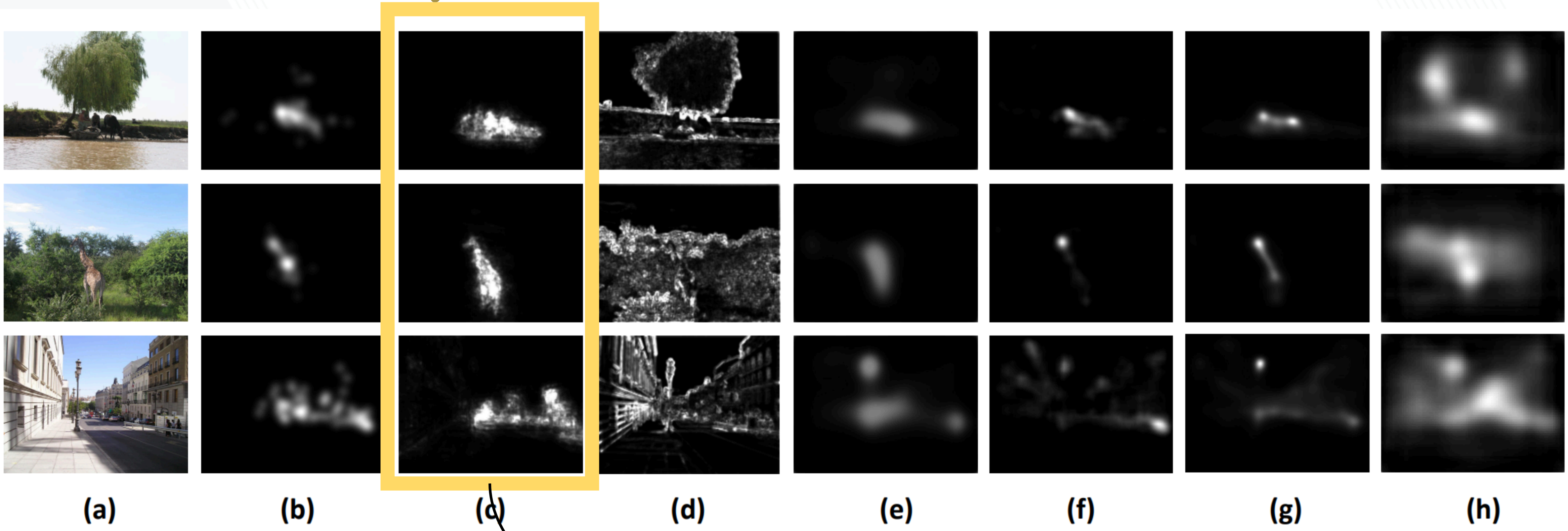


Fig. 3. Saliency map visualization. (a) Input image (b) Groudtruth (c) Proposed Method (d) Feed-forward feature (e) SalGan [21] (f) ML-Net [5] (g) DeepGazeII [22] (h) ShallowDeep [23]

Implicit saliency

OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

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.

Robust Machine Learning
Human Visual Saliency

Table 1. Human visual saliency vs Model Saliency

	NSS				CC			
Networks	ResNet-18	ResNet-34	ResNet-50	ResNet-101	ResNet-18	ResNet-34	ResNet-50	ResNet-101
GradCam	0.7657	0.7545	0.7203	0.7335	0.3496	0.3396	0.3190	0.3210
GBP	0.3862	0.4191	0.3898	0.3415	0.2474	0.2453	0.2443	0.2233
ImplicitSaliency	**0.8274**	**0.8018**	**0.7659**	**0.7981**	**0.4132**	**0.4112**	**0.3868**	**0.4051**

Table 2. Robustness Analysis of Implicit Saliency

	NSS					CC				
Gaussian Blur	Sal Gan	Deep GazeII	ML Net	Shallow Deep	**Implicit Saliency**	Sal Gan	Deep GazeII	ML Net	Shallow Deep	**Implicit Saliency**
$r = 0$	0.8977	0.6214	0.5431	**0.9306**	0.7981	**0.6280**	0.5927	0.4481	0.5120	0.4051
$r = 50$	↓ 0.2239	↓ 0.3436	↓ 0.2484	↓ 0.2025	↓ **0.1793**	↓ 0.2731	↓ 0.3954	↓ 0.2940	↓ 0.1840	↓ **0.1432**

↓ is the performance decrease when an input image is corrupted by gaussian noise of kernel size r

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.

Robust Machine Learning

Human Visual Saliency

Table 1. Human visual saliency vs Model Saliency

	NSS				CC			
Networks	ResNet-18	ResNet-34	ResNet-50	ResNet-101	ResNet-18	ResNet-34	ResNet-50	ResNet-101
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
GBP	0.3862	0.4191	0.3898	0.3415	0.2474	0.2453	0.2443	0.2233
ImplicitSaliency	**0.8274**	**0.8018**	**0.7659**	**0.7981**	**0.4132**	**0.4112**	**0.3868**	**0.4051**

Contrastive feature-based detector correlates better with human gaze than Observed causal Grad-CAM

Table 2. Robustness Analysis of Implicit Saliency

	NSS					CC				
Gaussian [illegible]	Sal [illegible]	Deep [illegible]	ML [illegible]	Shallow [illegible]	**Implicit** [illegible]	Sal [illegible]	Deep [illegible]	ML [illegible]	Shallow [illegible]	**Implicit** [illegible]
$r = 0$	0.8977	0.6214	0.5[illegible]	[illegible]	[illegible]	[illegible]	[illegible]927	0.4481	0.5120	0.4051
$r = 50$	↓ 0.2239	↓ 0.3436	↓ 0.2484	↓ 0.2025	↓ **0.1793**	↓ 0.2731	↓ 0.3954	↓ 0.2940	↓ 0.1840	↓ **0.1432**

Contrastive feature-based detector outperforms some of the supervised methods that train on human saliency datasets. It also is more robust.

↓ is the performance decrease when an input image is corrupted by gaussian noise of kernel size r

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.

References

- **Robust Recognition :** M. Prabhushankar and G. AlRegib, "Contrastive Reasoning in Neural Networks, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted on Jan. 9 2021. [PDF]
- **Saliency :** 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. [PDF][Code][Video]
- **IQA Contrastive :** 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. [PDF][Code]
- **IQA UNIQUE :** 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.

OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

To Conclude,

- We introduced an interpretation of **gradients in the space of models** from a perspective of **model uncertainty**
- We proposed a framework for efficient gradient generation with **confounding labels** to quantify uncertainty of fully trained networks
- We validated that the gradient-based uncertainty measure outperform activation-based features in **out-of-distribution detection** and **corrupted input detection**
- We interpreted gradients as a reasoning mechanism within neural networks
- We showed that gradients can be used to answer three explanatory paradigms
- Gradients as features can be used to create robust neural networks as a plug-in on top of existing neural networks
- We showed that there is a higher correlation between gradient-based contrastive features and applications relating to human visual systems than between feed-forward features and the same applications

arXiv

https://arxiv.org/abs/2103.12329

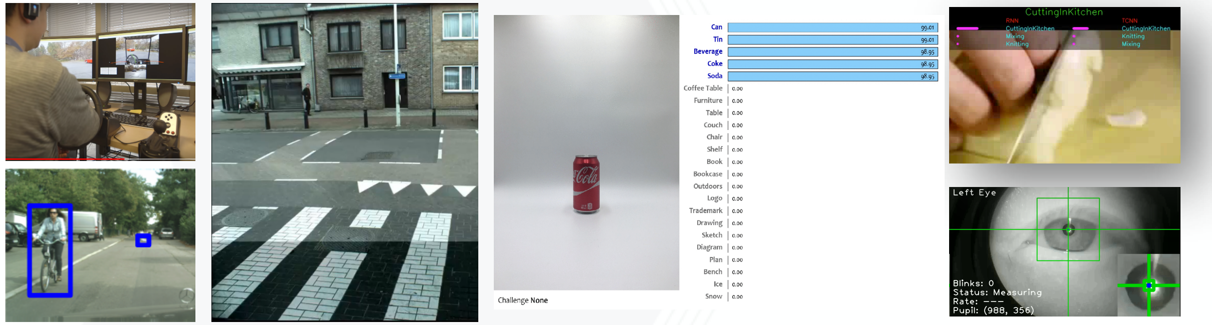
arXiv

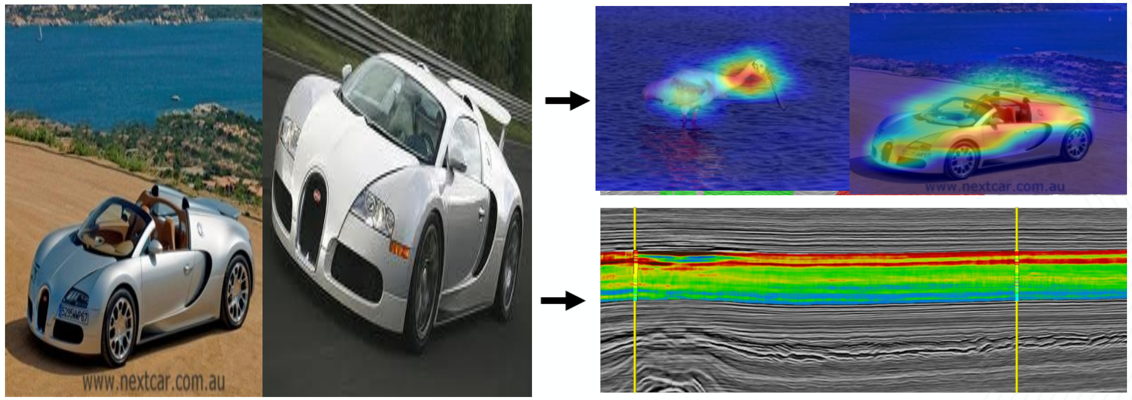
https://arxiv.org/abs/2008.00178

OLIVES@Gatech

https://ghassanalregib.info

Research Interests: AI, Machine Learning, Computer Vision, Perception, Scene Understanding, Learning in the Wild, Learning for Autonomous Vehicles, Medical Image Analysis, Computational Ophthalmology, Seismic Interpretation





Robust, Active Learning

Developing algorithms that can robustly operate under real-world challenging conditions through weakly supervised learning, backpropogated gradients, hyperpolar classification, and transfer learning.

Introduced three large-scale datasets (>1M) with controlled challenging conditions to test and develop robust algorithms: CURE-TSD, CURE-TSR, CURE-OR

Working on applications including but not limited to autonomous driving, remote repositioning, smart and connected healthcare, activity recognition, semantic segmentation, object classification and detection, defense models design, and computational seismic interpretation.

Explainability, Limited Annotations

Learning to characterize data using limited labels using weakly-/semi-supervised learning and sequence modeling for various applications such as subsurface lithology, structure, and stratigraphy characterization, and material characterization, OCT analysis, and medical imaging.

Introduced four datasets for subsurface characterization using weak labels and auxiliary data such as well-logs: LANDMASS-1, LANDMASS-2, Facies classification benchmark, and one large-scale dataset for material characterization of textile fabrics: CoMMonS. Also introduced one interactive tool for salt interpretation benchmarking in large subsurface volumes : Salt Dome Interpretation Tool.

Thanks for your attention

https://github.com/olivesgatech

OLIVES
@GeorgiaTech

Georgia Tech
CREATING THE NEXT