Novelty Detection Through Model-Based Characterization of Neural Networks




Gukyeong Kwon* Mohit Prabhushankar Dogancan Temel Ghassan AlRegib

(*: Speaker)

Georgia Institute of Technology

October 2020

IEEE International Conference on Image Processing
25-28 October 2020, United Arab Emirates
FULLY VIRTUAL

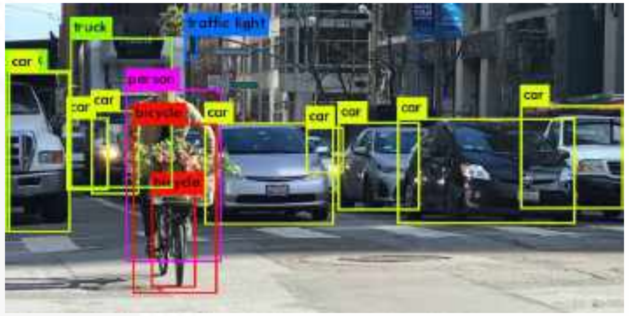
OLIVES @GeorgiaTech

Paper

Introduction

Scene Understanding

Object detection



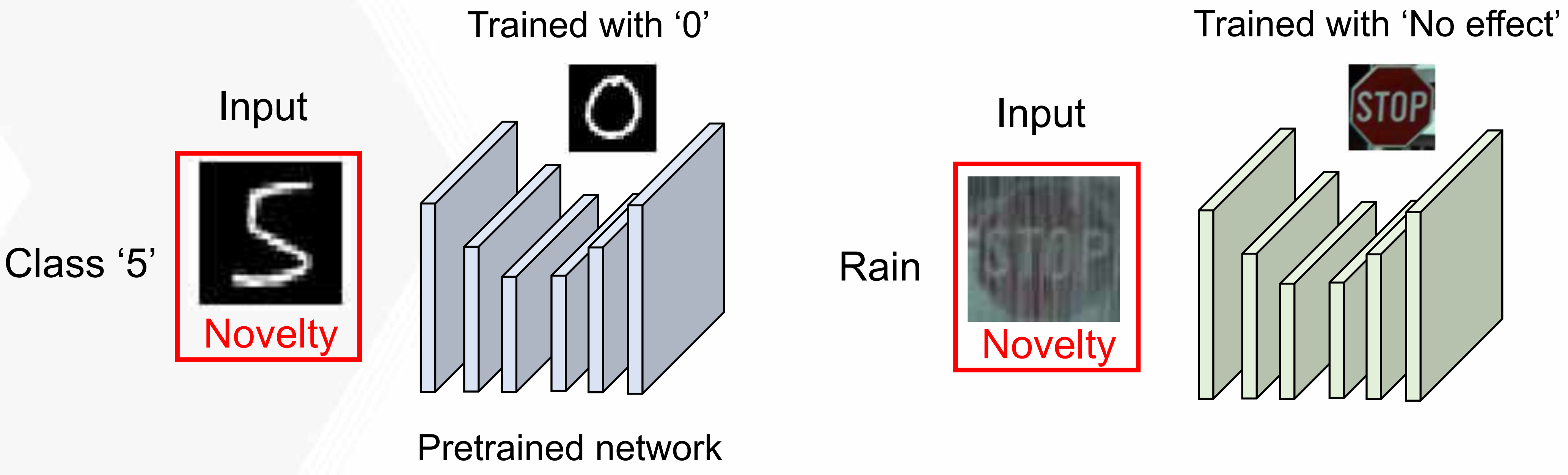
Semantic Segmentation

Instance Segmentation

Scene Understanding

Overview
Novelty Detection

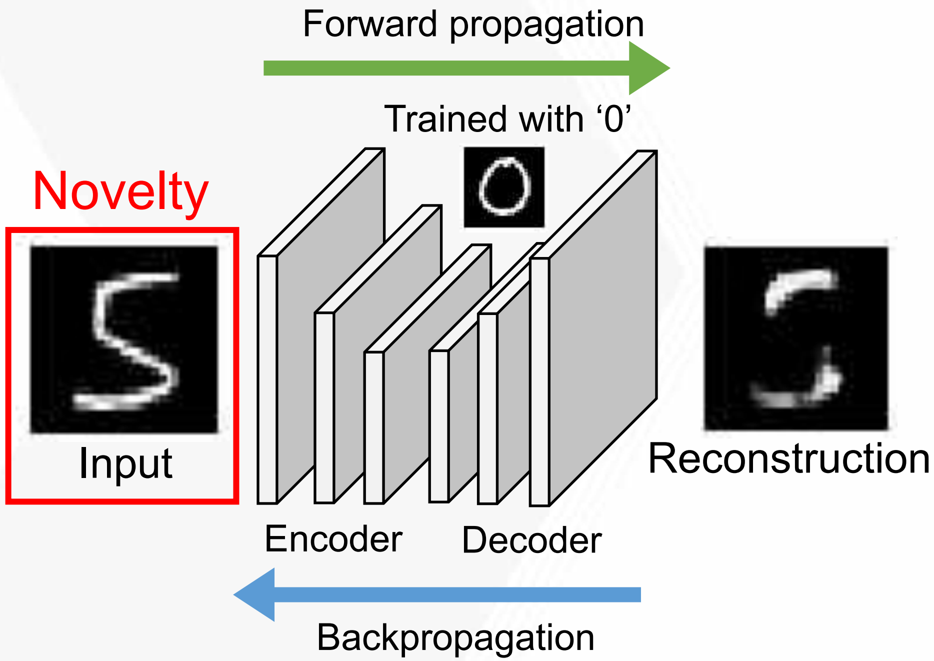
Novelty (Anomaly) : Data whose *classes* or *attributes* differs from training data



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

Overview

Model-based Characterization



<u>Existing approaches</u>

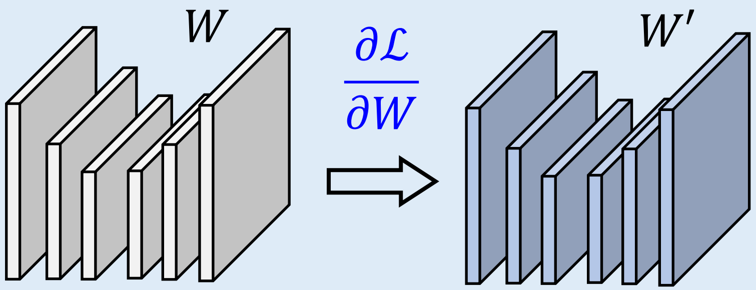
Data-based Characterization
(Activation)

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

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

<u>Proposed approach</u>

Model-based Characterization
(Backpropagate Gradient)



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

Contributions

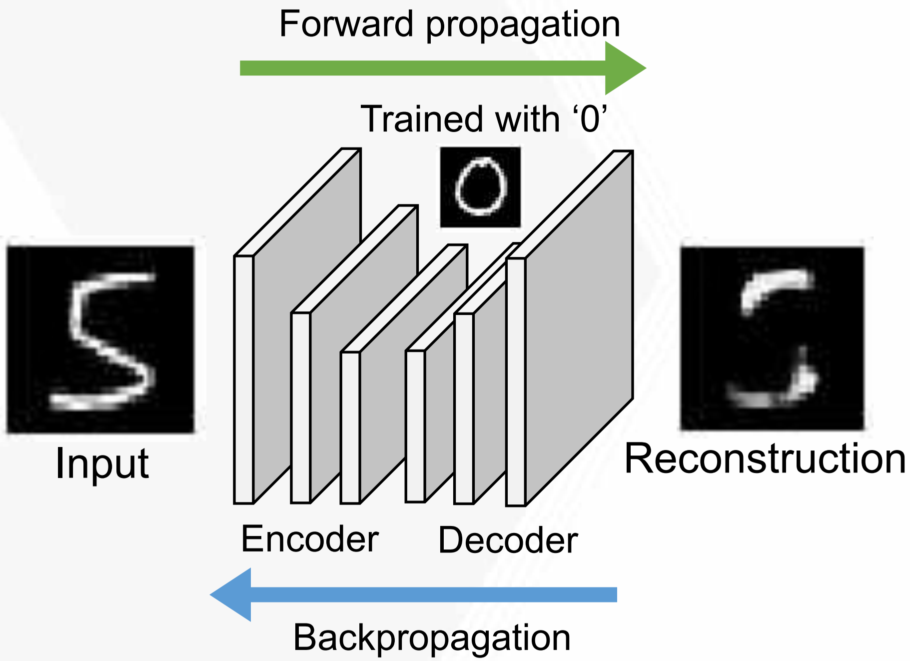
1. We propose a framework to characterize novelty from the model perspective using gradients.

2. We validate the representation capability of gradients for novelty detection in comparison with activation through comprehensive baseline experiments.

3. We validate the generalizability of gradient features for different classes and input conditions.

Related Works

Data-based Characterization



Existing approaches

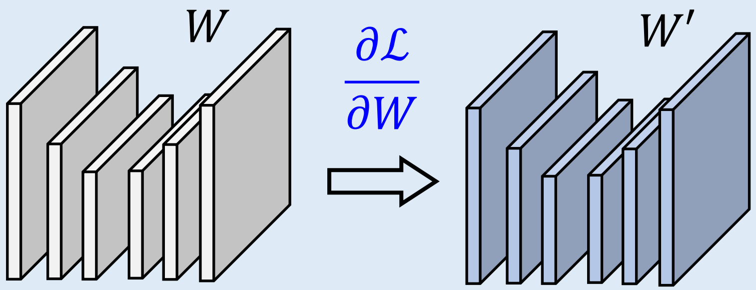
Data-based Characterization
(Activation)

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

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

Proposed approach

Model-based Characterization
(Backpropagate Gradient)



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

Related Works

Data-based Characterization

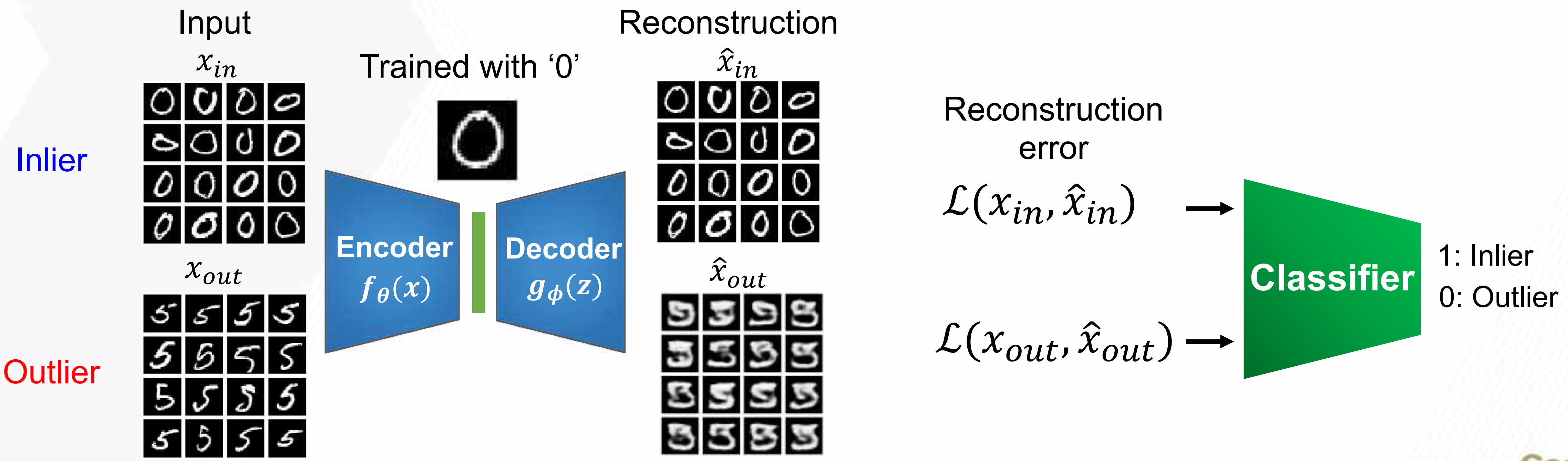
Reconstruction Error

Skurada 2014

Zhou 2017

Zong 2018

Sabokrou 2018

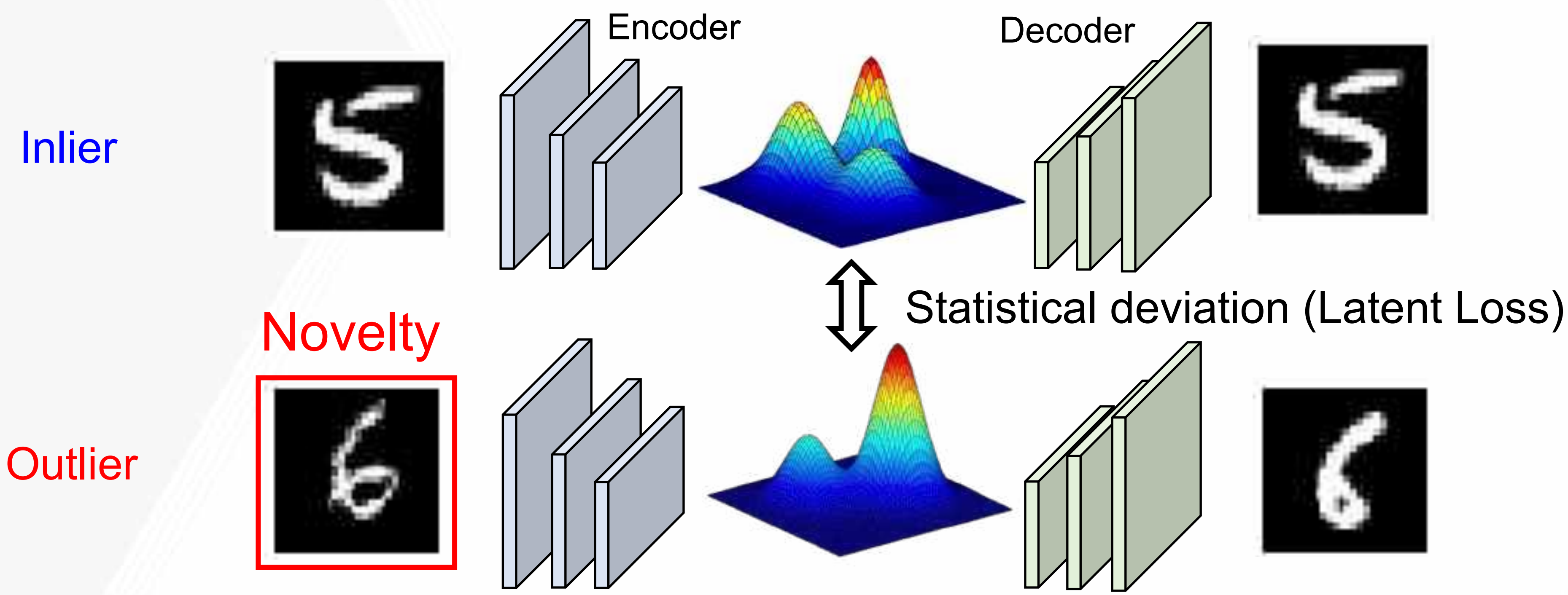


Related Works

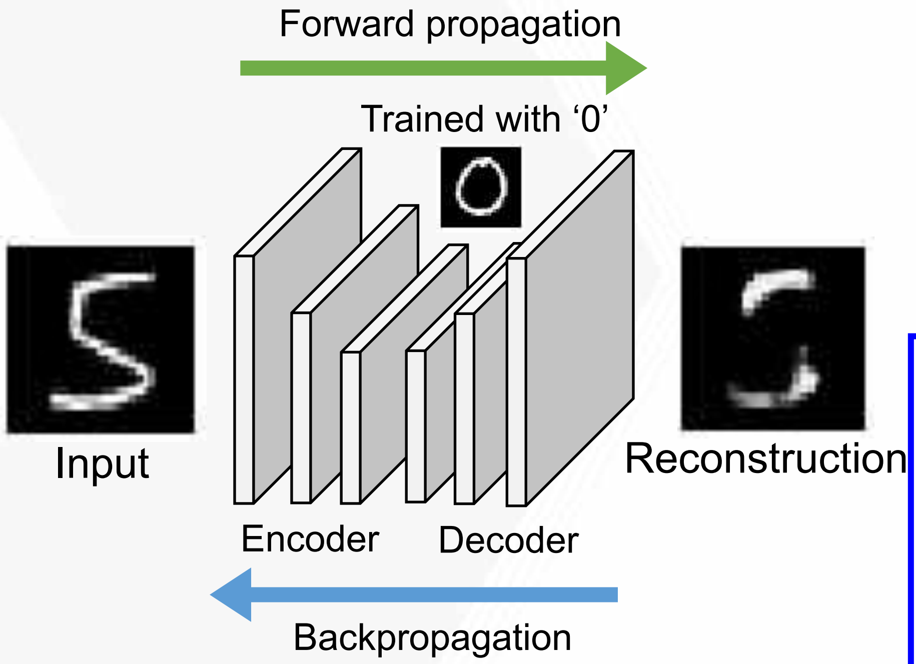
Data-based Characterization

Constrained Representation

Tax	Fan	Pidhorskyi	Abati
2004	2016	2018	2019



Related Works
Data-based Characterization



Existing approaches

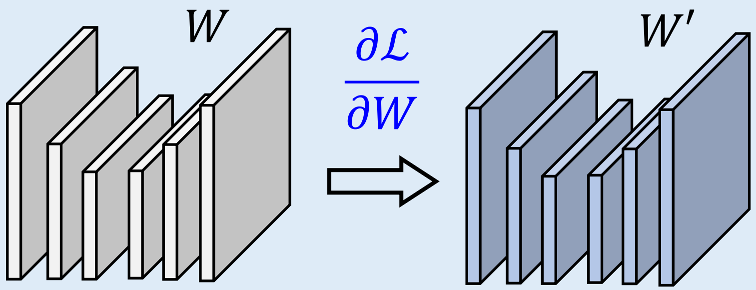
Data-based Characterization
(Activation)

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

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

Proposed approach

Model-based Characterization
(Backpropagate Gradient)



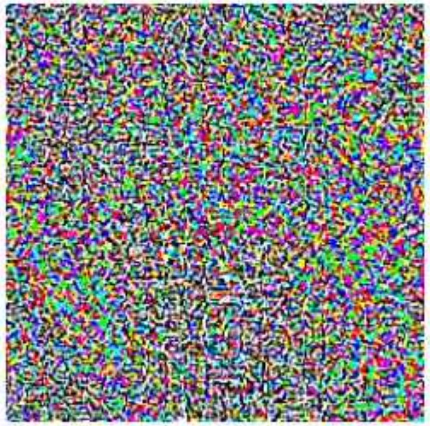
How much **model update** is required by the input?

Related Works

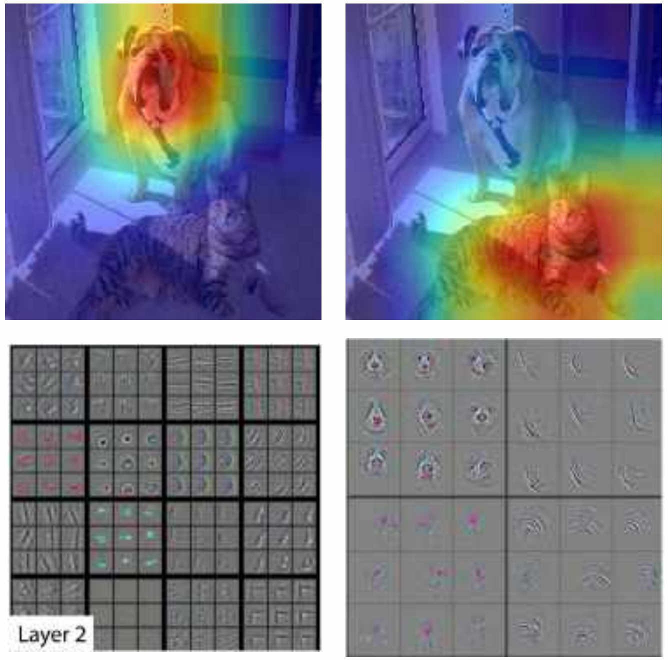
Usage of Gradients

Adversarial attack generation			Visualization			Regularization		
Goodfellow 2014	Kurakin 2016	Madry 2017	Zeiler 2014	Springenberg 2014	Selvaraju 2017	Drucker 2014	Sokolic 2017	Ross 2018

Fast Gradient Sign Method



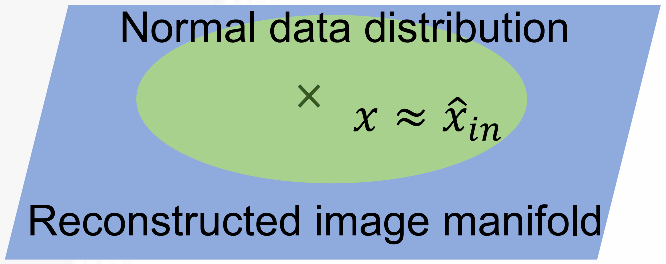
$$\mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$$

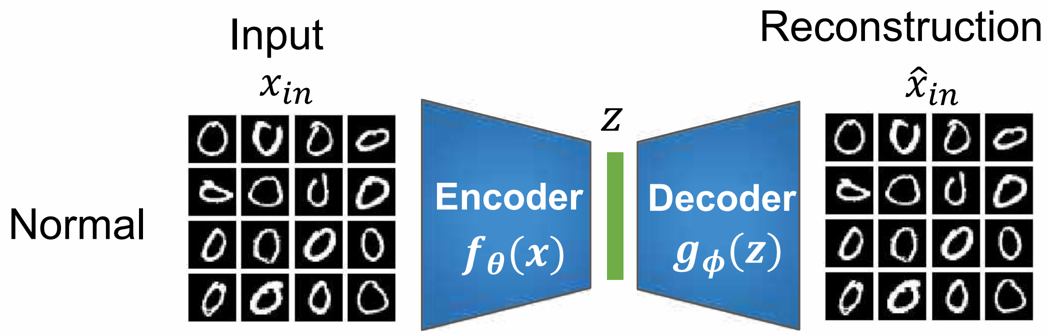


Jacobian Regularizer: Penalize the squared Frobenius norm of the Jacobian of the softmax output with respect to input.

$$L_{JacReg}(x, y, \Theta) = L(x, y, \Theta) + \lambda \|J_f\|_F^2$$

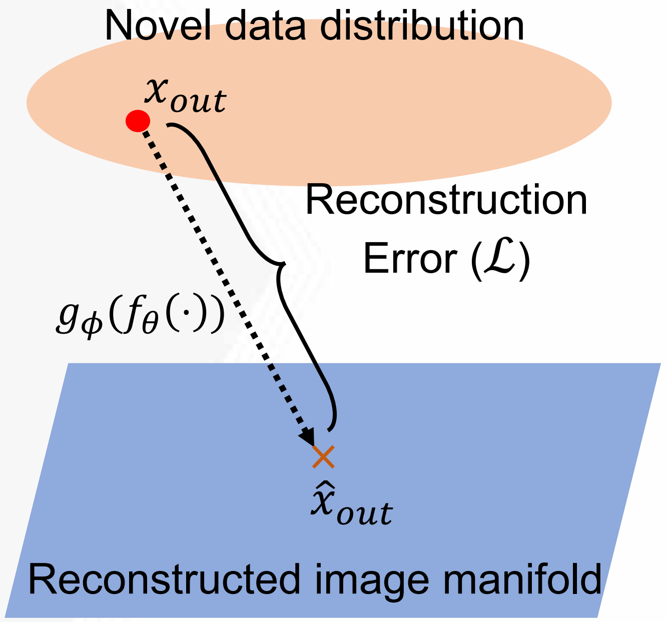
Geometric Interpretation
Advantages of Gradient Features

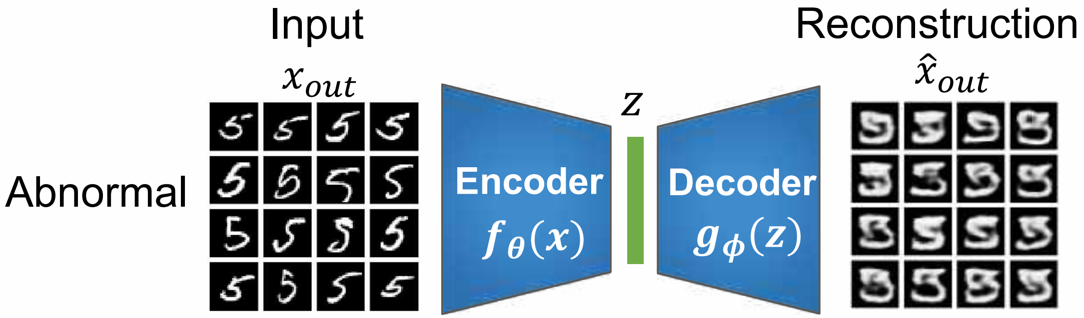




Geometric Interpretation

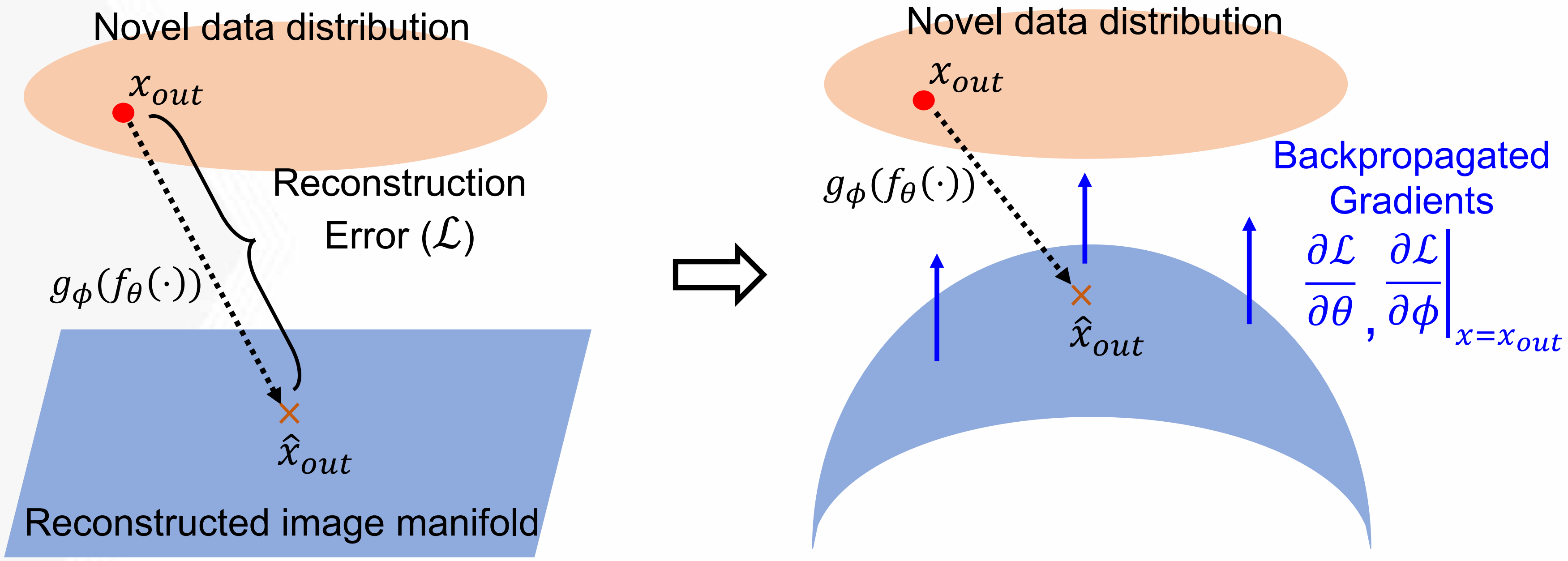
Advantages of Gradient Features





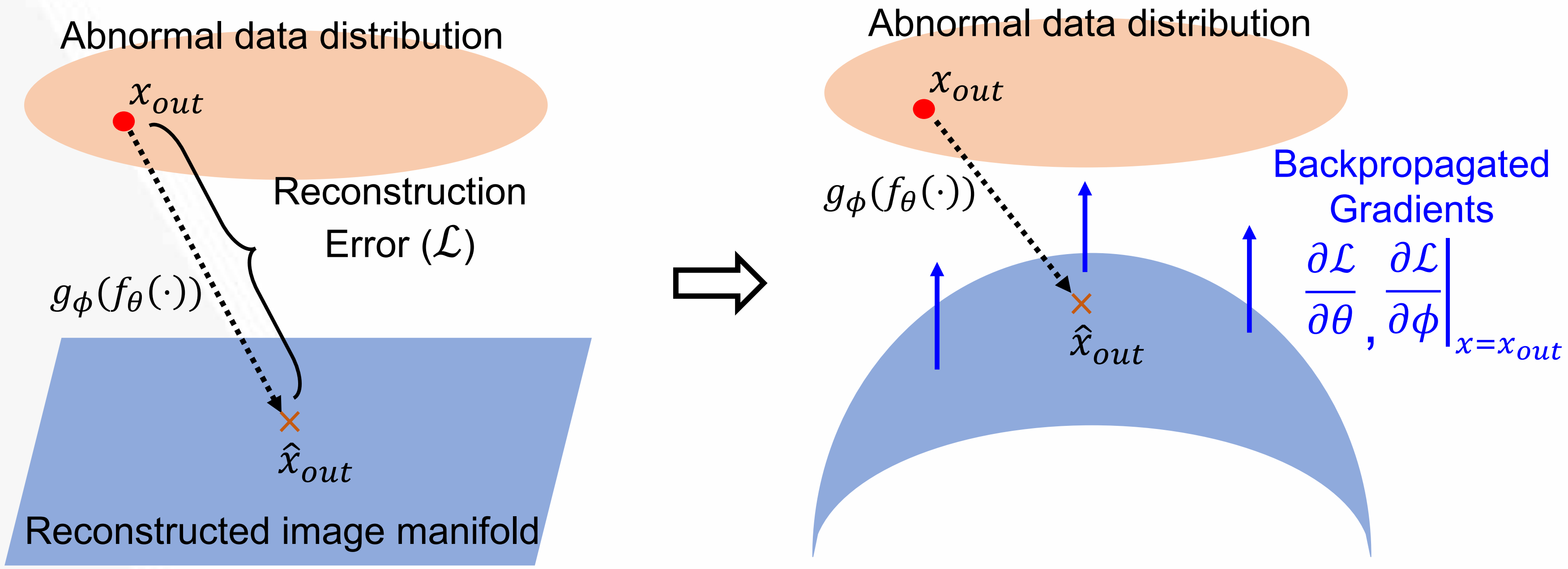
Geometric Interpretation

Advantages of Gradient Features



Geometric Interpretation

Advantages of Gradient Features



1) Provide directional information to characterize anomalies

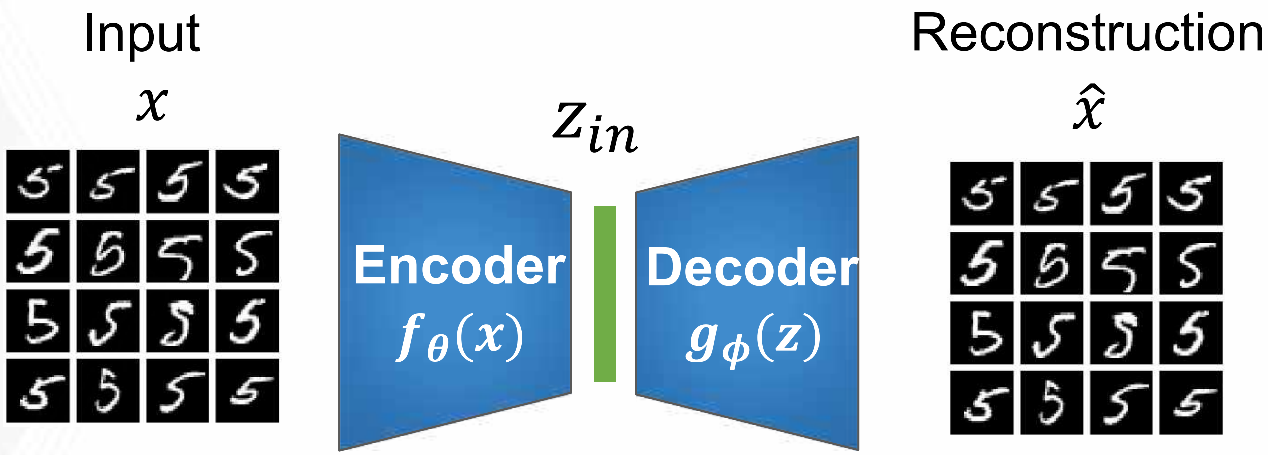
2) Gradients from different layers capture novelty at different levels of data abstraction

Georgia Tech
CREATING THE NEXT

Model-Based Characterization

Statistical Analysis

1. Train a variational autoencoder with digit ‘5’ images

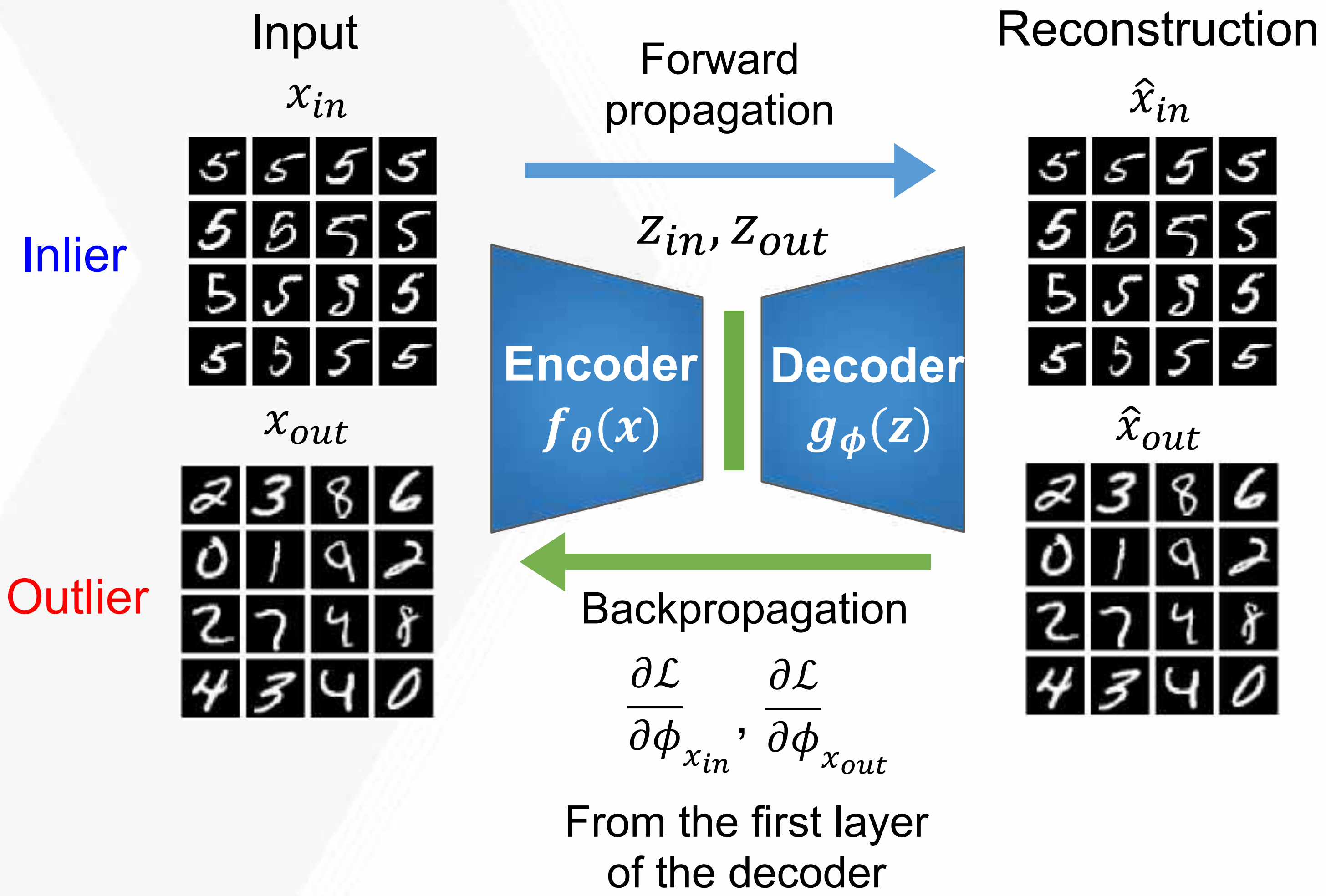


$$J(x;\theta,\phi) = \underbrace{-\mathbb{E}_{g_\phi(z|x)}[\log f_\theta(x|z)]}_{\substack{\text{Reconstruction error } (\mathcal{L}) \\ \text{Binary cross entropy}}} + \underbrace{\mathrm{KL}[g_\phi(z|x)||f(z)]}_{\substack{\text{Latent loss } (\Omega) \\ \text{KL divergence}}}$$

Model-Based Characterization

Statistical Analysis

2. Extract reconstruction error, latent loss, and gradient features



	Inlier feature	Outlier feature
Recon.	$\mathcal{L}(x_{in}, \hat{x}_{in})$	$\mathcal{L}(x_{out}, \hat{x}_{out})$
Latent	$\Omega(z_{in})$	$\Omega(z_{out})$
Gradient	$\left\Vert \frac{\partial \mathcal{L}}{\partial \phi}_{x_{in}} \right\Vert_2$	$\left\Vert \frac{\partial \mathcal{L}}{\partial \phi}_{x_{out}} \right\Vert_2$

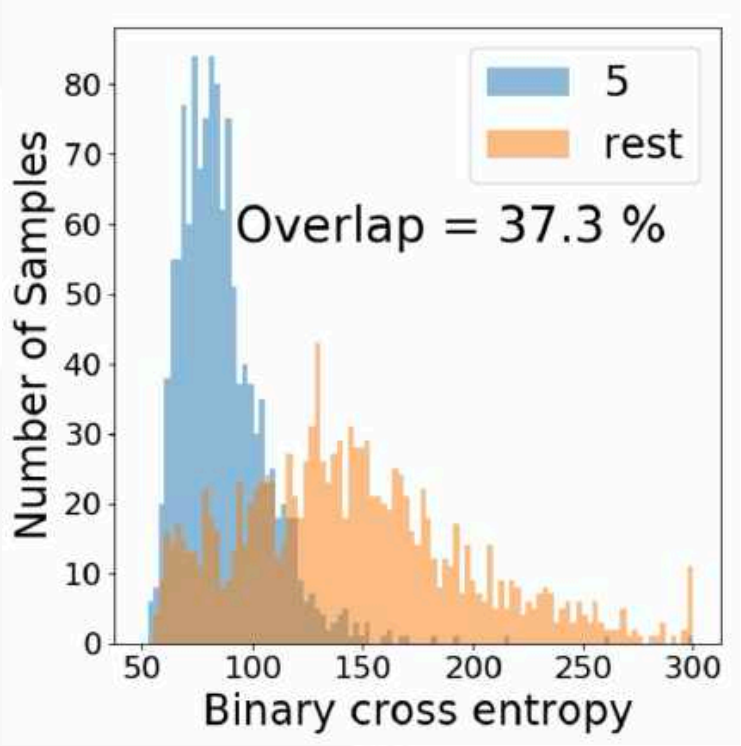
3. Histogram visualization

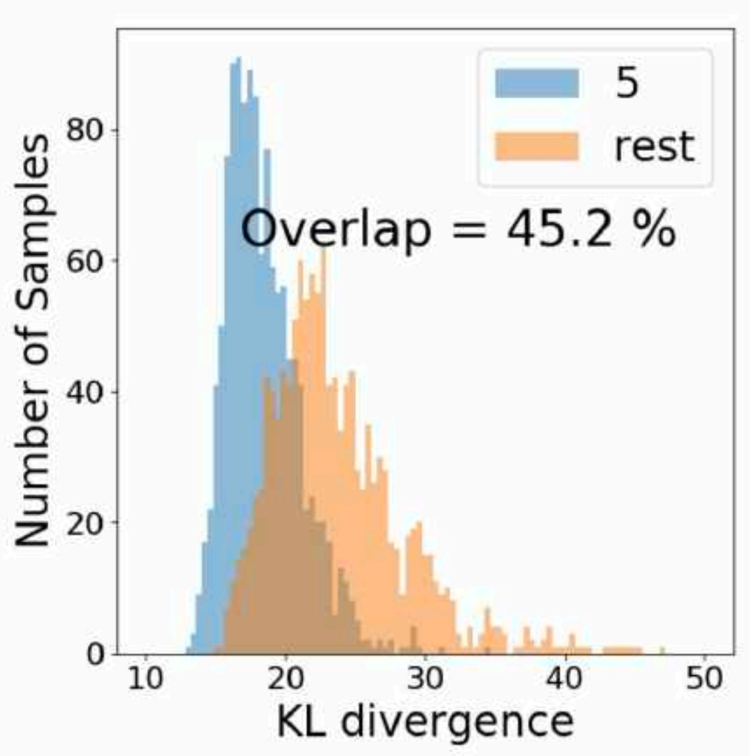
Georgia Tech
CREATING THE NEXT

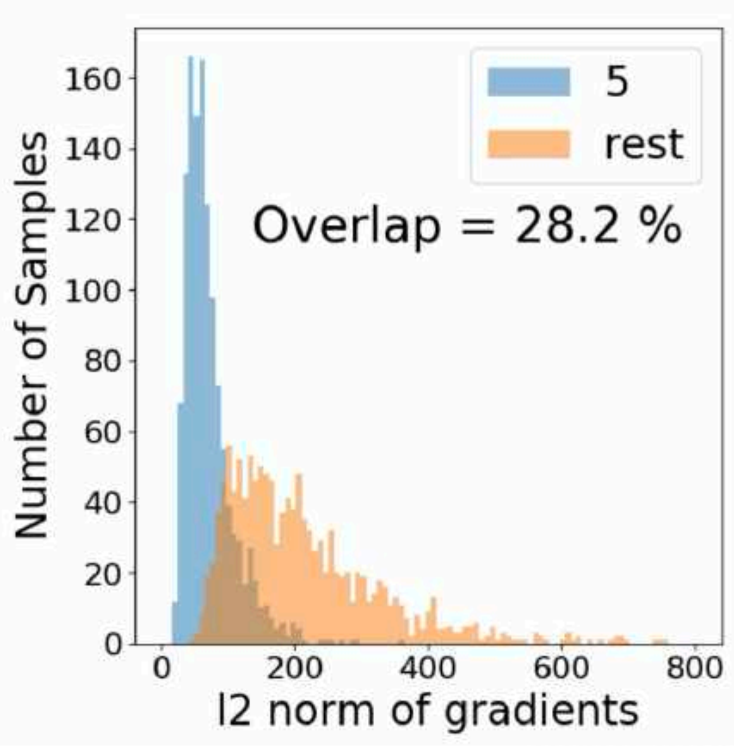
Model-Based Characterization

Statistical Analysis

$$\text{Overlap} = \frac{\text{Number of samples in the overlapped region}}{\text{Total number of samples}}$$



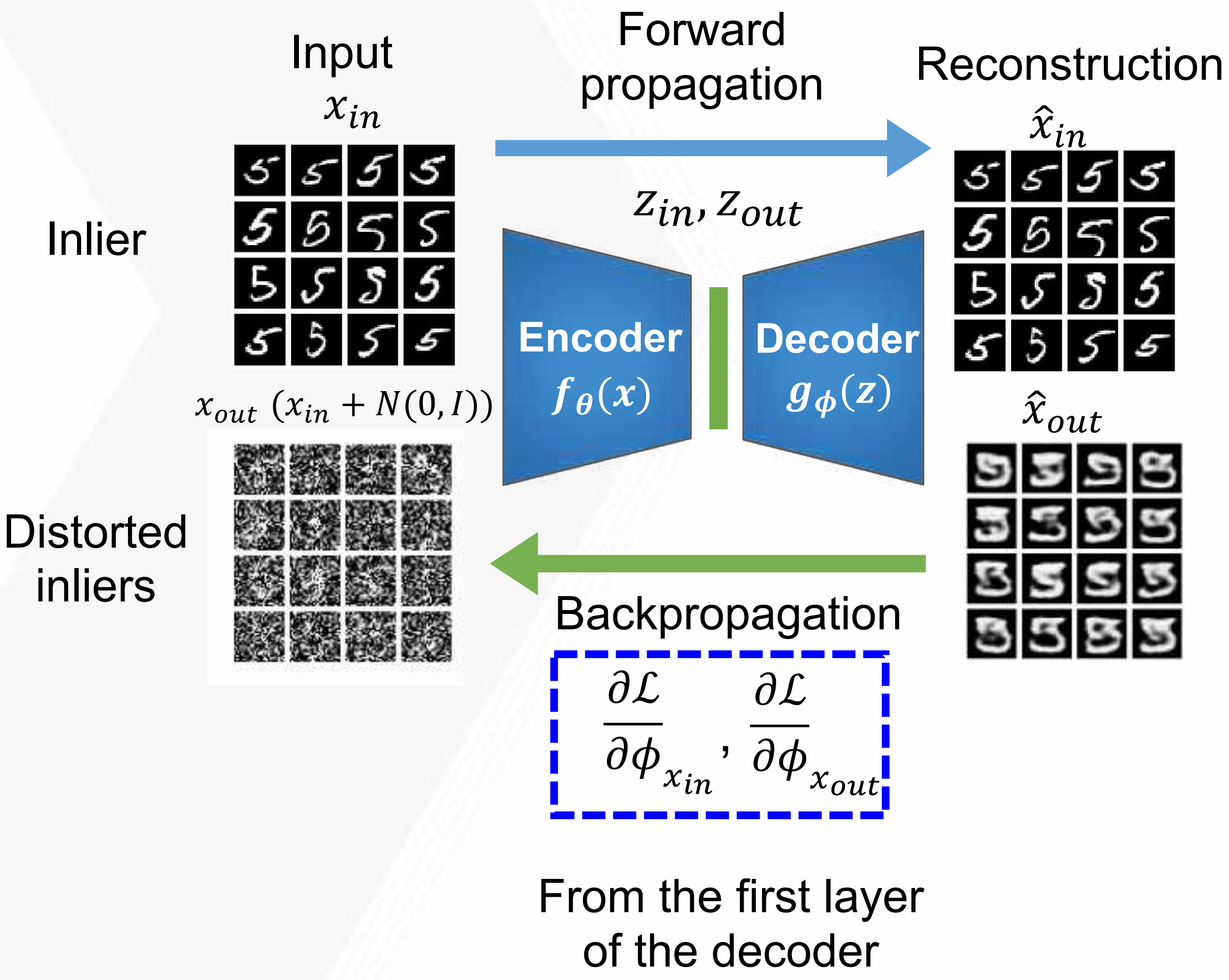


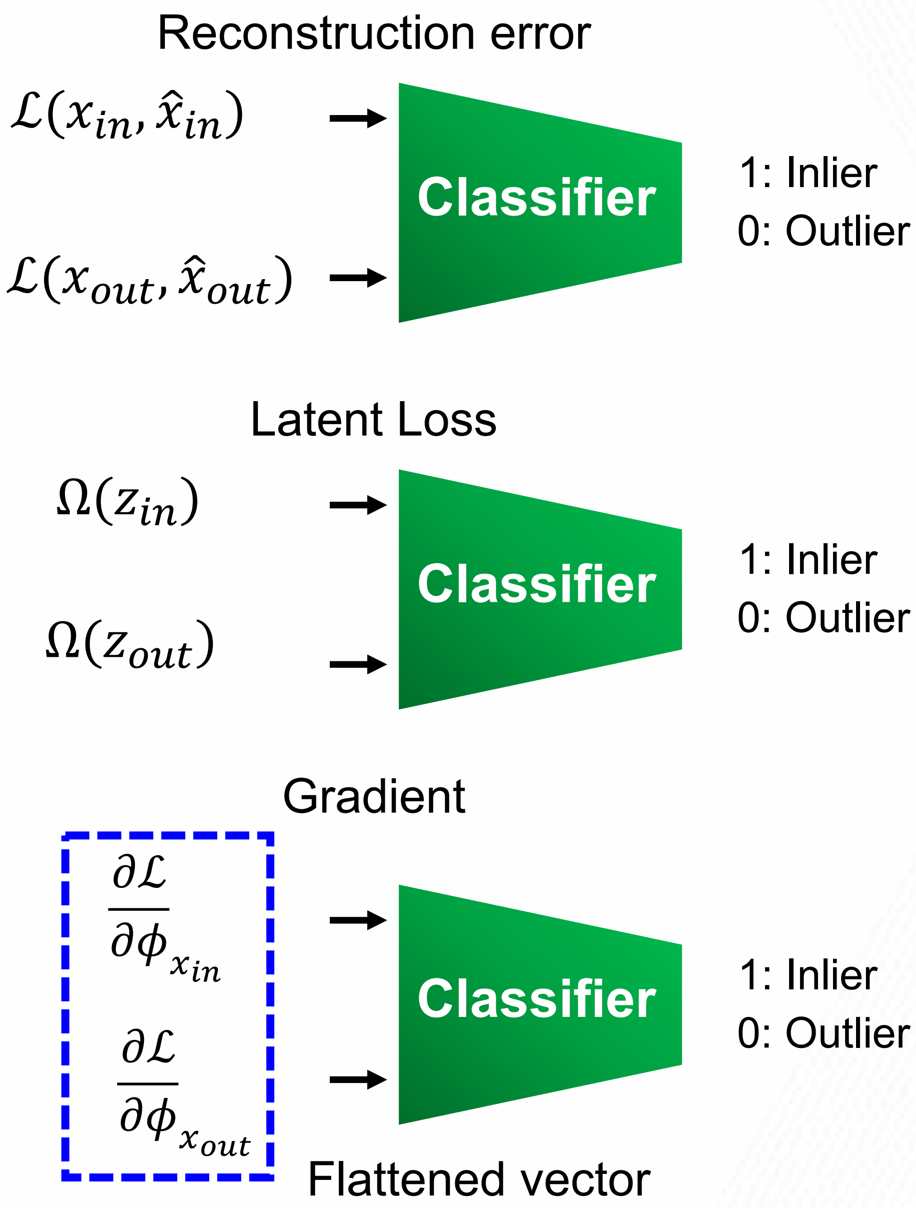


→ Gradient are the most discriminative features for novelty characterization

Georgia Tech
CREATING THE NEXT

Experimental Setup
Activation vs. Gradients





Experimental Setup

Novel Class Detection

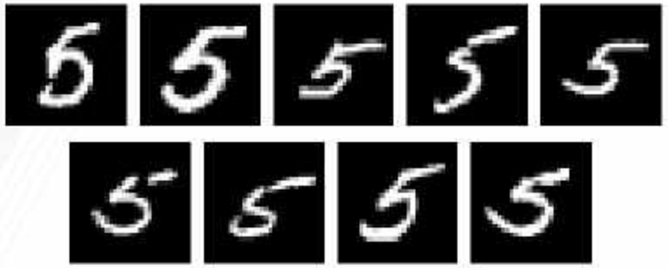
MNIST

Fashion MNIST

CIFAR-10

1 class (inliers) / 9 classes (outliers)

Learned class



⇩

Inliers

Novel classes

⇩

Outliers

Experimental Results

Novel Class Detection

AUROC Results

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

Dataset	Repre.	Classes 0	1	2	3	4	5	6	7	8	9	Average
MNIST	Recon.	0.043	0.916	0.293	0.132	0.103	0.158	0.101	0.115	0.291	0.147	0.230
	Latent	0.956	0.510	0.687	0.740	0.852	0.526	0.675	0.942	0.348	0.948	0.718
	Gradient	**0.985**	**0.994**	**0.941**	**0.928**	**0.953**	**0.926**	**0.980**	**0.960**	**0.894**	**0.968**	**0.953**
fMNIST	Recon.	0.778	0.952	0.831	0.799	0.801	0.787	0.748	0.939	0.610	0.932	0.818
	Latent	0.733	0.642	0.525	0.877	0.715	0.831	0.585	0.961	0.702	0.835	0.741
	Gradient	**0.913**	**0.958**	**0.883**	**0.922**	**0.907**	**0.924**	**0.798**	**0.974**	**0.925**	**0.975**	**0.918**
CIFAR-10	Recon.	0.600	0.485	0.539	**0.496**	0.532	0.444	0.601	**0.545**	0.634	0.541	0.542
	Latent	**0.683**	0.382	0.560	0.458	0.649	0.486	**0.724**	0.465	**0.662**	**0.550**	0.562
	Gradient	0.658	**0.543**	**0.632**	0.461	**0.725**	**0.493**	0.699	0.490	0.641	0.477	**0.582**

1) The proposed gradient features consistently outperforms other classifiers for all the inlier classes in MNIST and Fashion MNIST

Experimental Results

Novel Class Detection

AUROC Results

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

Dataset	Repre.	Classes 0	1	2	3	4	5	6	7	8	9	Average
MNIST	Recon.	0.043	0.916	0.293	0.132	0.103	0.158	0.101	0.115	0.291	0.147	0.230
	Latent	0.956	0.510	0.687	0.740	0.852	0.526	0.675	0.942	0.348	0.948	0.718
	Gradient	**0.985**	**0.994**	**0.941**	**0.928**	**0.953**	**0.926**	**0.980**	**0.960**	**0.894**	**0.968**	**0.953**
fMNIST	Recon.	0.778	0.952	0.831	0.799	0.801	0.787	0.748	0.939	0.610	0.932	0.818
	Latent	0.733	0.642	0.525	0.877	0.715	0.831	0.585	0.961	0.702	0.835	0.741
	Gradient	**0.913**	**0.958**	**0.883**	**0.922**	**0.907**	**0.924**	**0.798**	**0.974**	**0.925**	**0.975**	**0.918**
CIFAR-10	Recon.	0.600	0.485	0.539	**0.496**	0.532	0.444	0.601	**0.545**	0.634	0.541	0.542
	Latent	**0.683**	0.382	0.560	0.458	0.649	0.486	**0.724**	0.465	**0.662**	**0.550**	0.562
	Gradient	0.658	**0.543**	**0.632**	0.461	**0.725**	**0.493**	0.699	0.490	0.641	0.477	**0.582**

1) The proposed gradient features consistently outperforms other classifiers for all the inlier classes in MNIST and Fashion MNIST
2) The gradient features achieve the highest average AUROC in CIFAR-10

Experimental Results

Novel Class Detection

AUROC Results

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

Dataset	Repre.	Classes 0	1	2	3	4	5	6	7	8	9	Average
MNIST	Recon.	0.043	0.916	0.293	0.132	0.103	0.158	0.101	0.115	0.291	0.147	0.230
	Latent	0.956	0.510	0.687	0.740	0.852	0.526	0.675	0.942	0.348	0.948	0.718
	Gradient	**0.985**	**0.994**	**0.941**	**0.928**	**0.953**	**0.926**	**0.980**	**0.960**	**0.894**	**0.968**	**0.953**
fMNIST	Recon.	0.778	0.952	0.831	0.799	0.801	0.787	0.748	0.939	0.610	0.932	0.818
	Latent	0.733	0.642	0.525	0.877	0.715	0.831	0.585	0.961	0.702	0.835	0.741
	Gradient	**0.913**	**0.958**	**0.883**	**0.922**	**0.907**	**0.924**	**0.798**	**0.974**	**0.925**	**0.975**	**0.918**
CIFAR-10	Recon.	0.600	0.485	0.539	**0.496**	0.532	0.444	0.601	**0.545**	0.634	0.541	0.542
	Latent	**0.683**	0.382	0.560	0.458	0.649	0.486	**0.724**	0.465	**0.662**	**0.550**	0.562
	Gradient	0.658	**0.543**	**0.632**	0.461	**0.725**	**0.493**	0.699	0.490	0.641	0.477	**0.582**

1) The proposed gradient features consistently outperforms other classifiers for all the inlier classes in MNIST and Fashion MNIST
2) The gradient features achieve the highest average AUROC in CIFAR-10
3) Comparison between reconstruction error and gradients highlights the significance of direction information from gradients

Experimental Setup

Novel Condition Detection

Challenging Unreal and Real Environments for Traffic Sign Recognition (CURE-TSR) (https://github.com/olivesgatech/CURE-TSR)

Challenge-free

12 challenge types and 5 levels

Inliers

Challenge-free

Outliers

Lens blur

Dirty lens

Gaussian blur

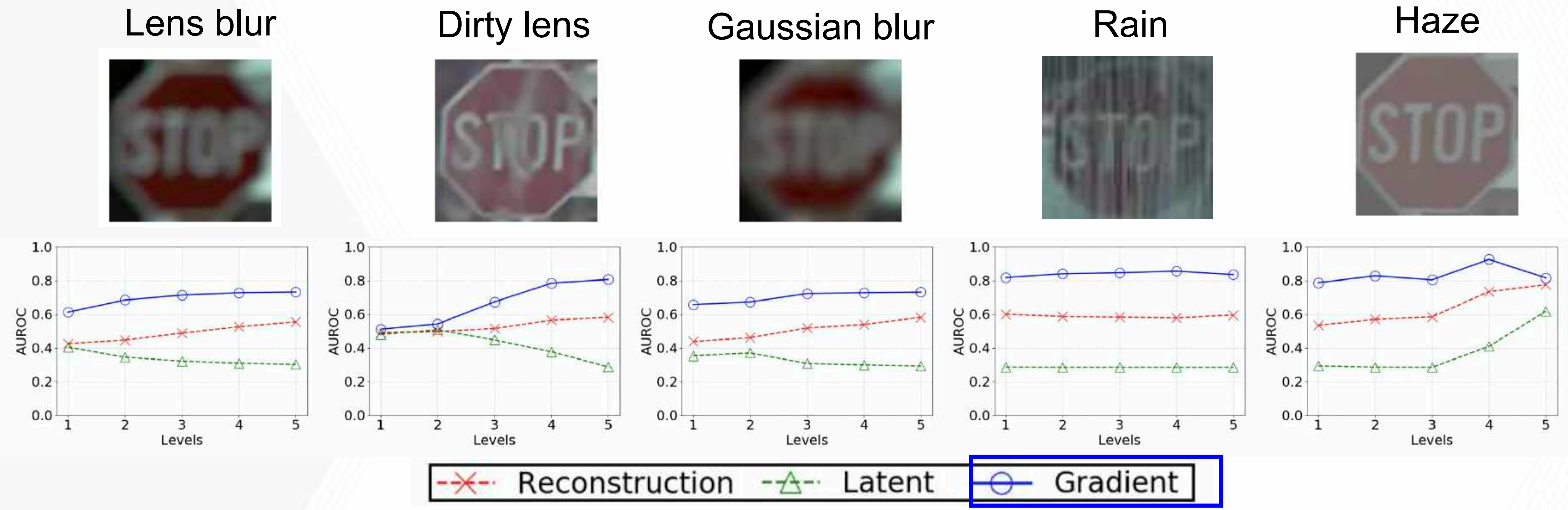
Rain

Haze

Georgia Tech
CREATING THE NEXT

Experimental Results

Novel Condition Detection



1) The classifiers trained using the gradients outperform those trained on the reconstruction error and the latent loss for all challenge types and levels
2) The gradient features achieves the largest improvement in *Rain* followed by *Lens blur* and *Gaussian blur*

Conclusion

- We proposed a framework to characterize novelty from the model perspective using gradients.

- The statistical analysis demonstrates that the larger separation between inliers and outliers is achieved using the gradients compared to the activation.

- We shows that the classifiers trained using the gradients as features outperform those trained using common activation-based features in novel class and condition detection

Thanks for your attention

Website

Paper

Code

[Website]: https://ghassanalregib.info/

[Paper]: https://arxiv.org/abs/2008.06094

[Code]: https://github.com/olivesgatech/gradcon-anomaly

[Extended version]: G. Kwon, M. Prabhushankar, D. Temel, and G. AlRegib, "Backpropagated Gradient Representations for Anomaly Detection," In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.