Why autoencoders fail at anomaly detection and what we can do about it

2023-12-01

High Energy Physics and Machine Learning Workshop @ HYU

Sangwoong Yoon

AI Research Fellow

Korea Institute for Advanced Study

About me: Sangwoong Yoon (From South Korea)

AI Research Fellow @ **Korea Institute for Advanced Study** (KIAS)

Education

- B.S., Chemical & Biological Engineering, Seoul National University (SNU)
- M.S., Interdisciplinary Program in Neuroscience, SNU
- Ph.D., Mechanical Engineering, SNU

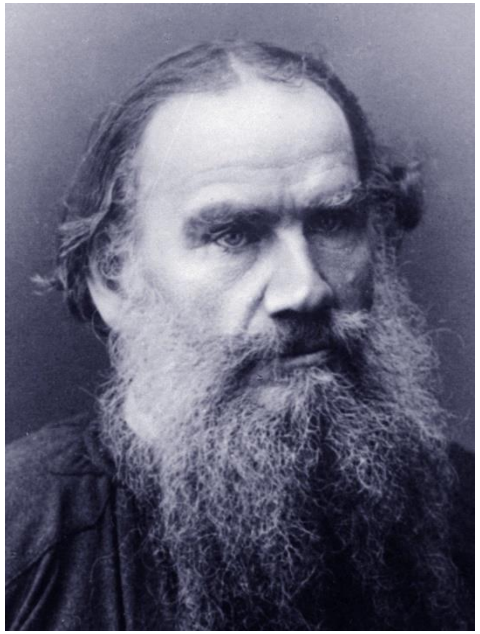
Industry Experience (as a ML Researcher/Engineer)

- Haezoom Inc., a startup for solar power plants
- Kakao Brain (Research Internship)
- Amazon (Research Internship)

"All happy families are alike;

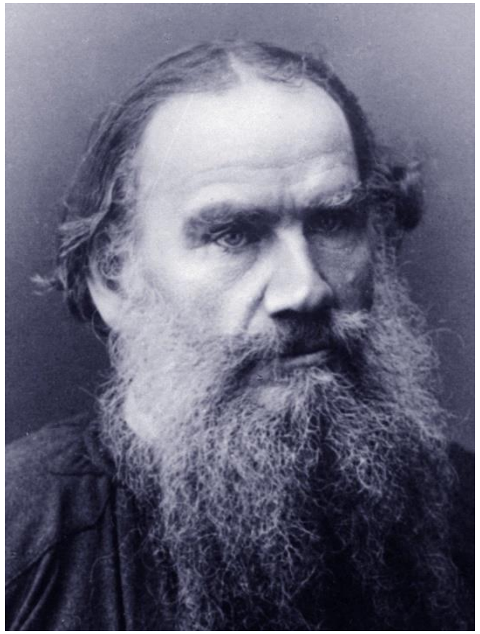
each unhappy family is unhappy in its own way."

Leo Tolstoy, ***Anna Karenina***



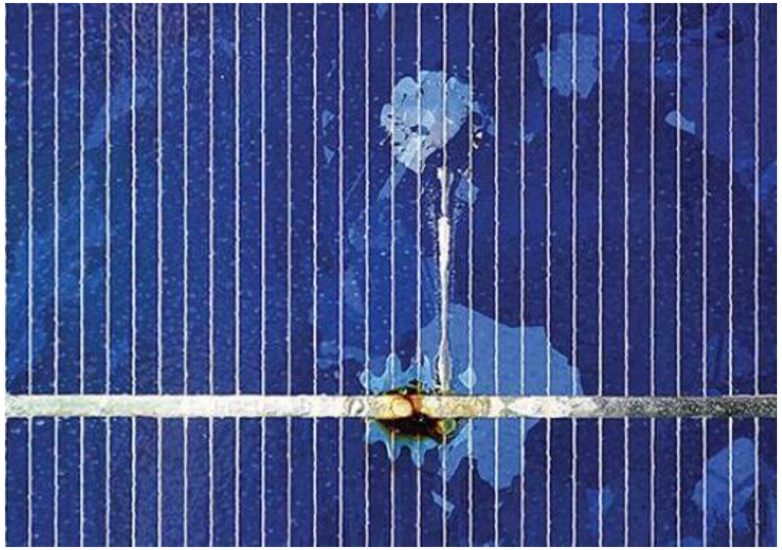
"All normal data are alike;

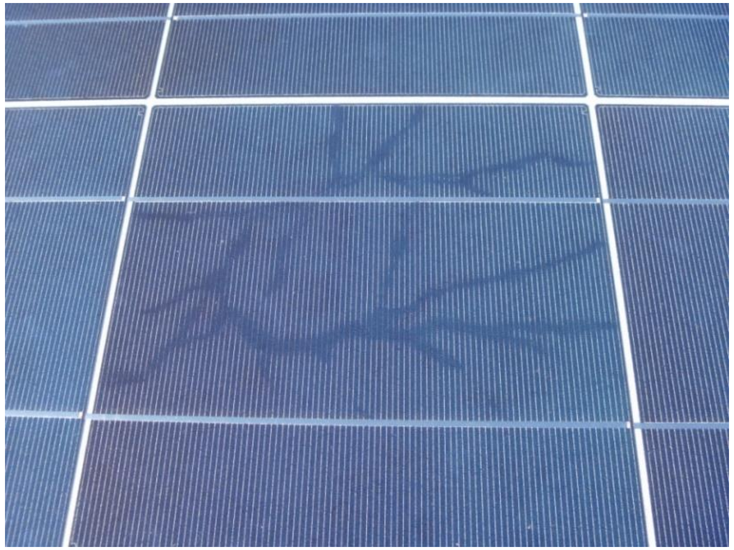
each anomalous data is anomalous in its own way."

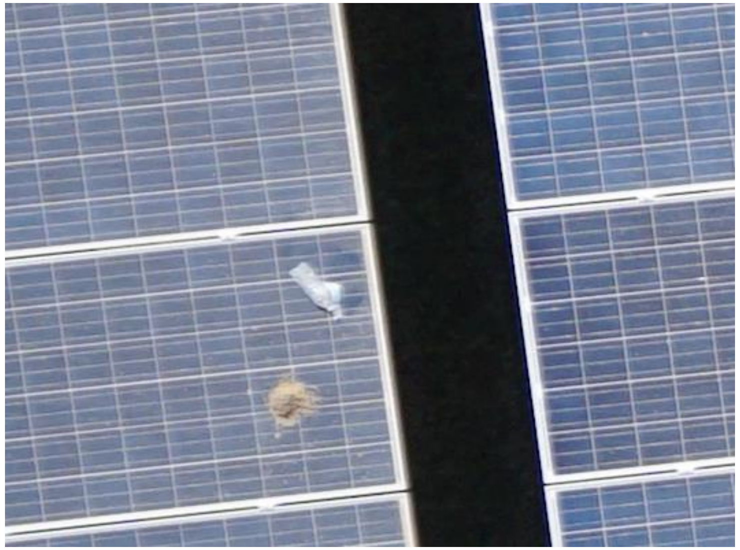


All Normal Data are Alike

Each Anomalous Data is Anomalous in its Own Way







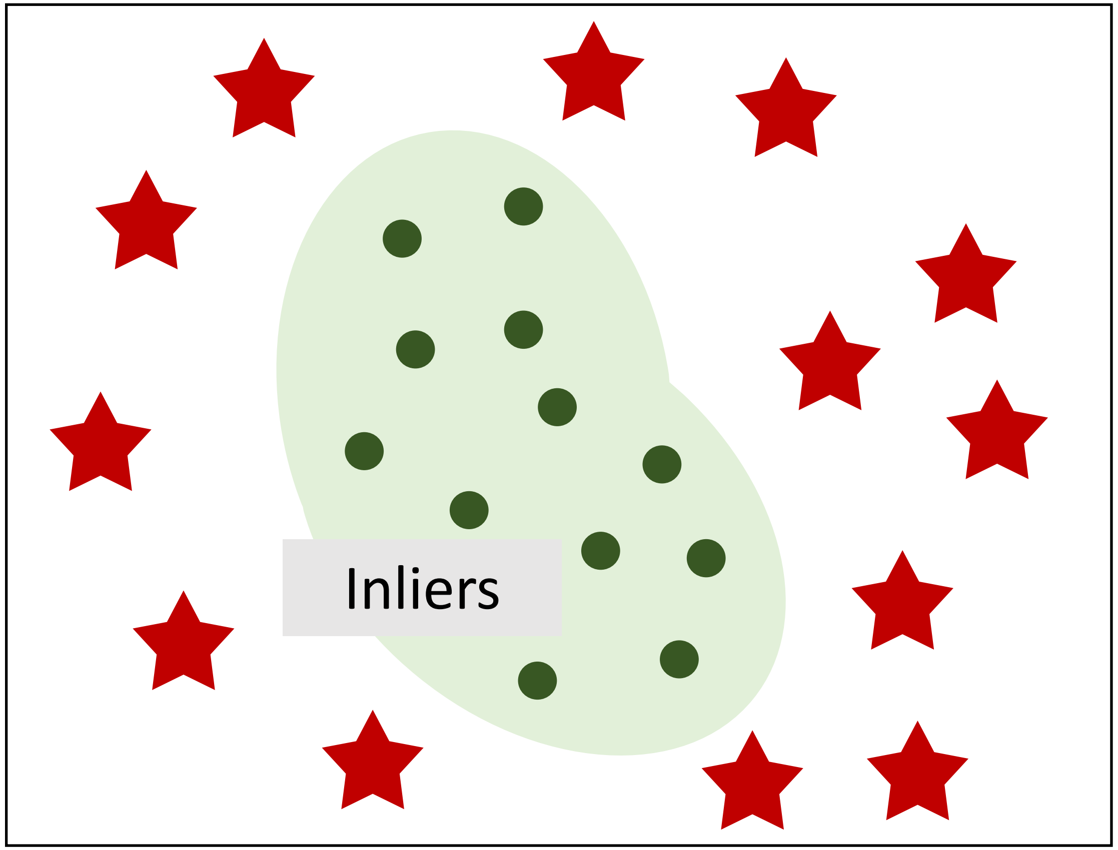
Anomaly Detection

Training:

A dataset of "normal" or inlier samples is given.

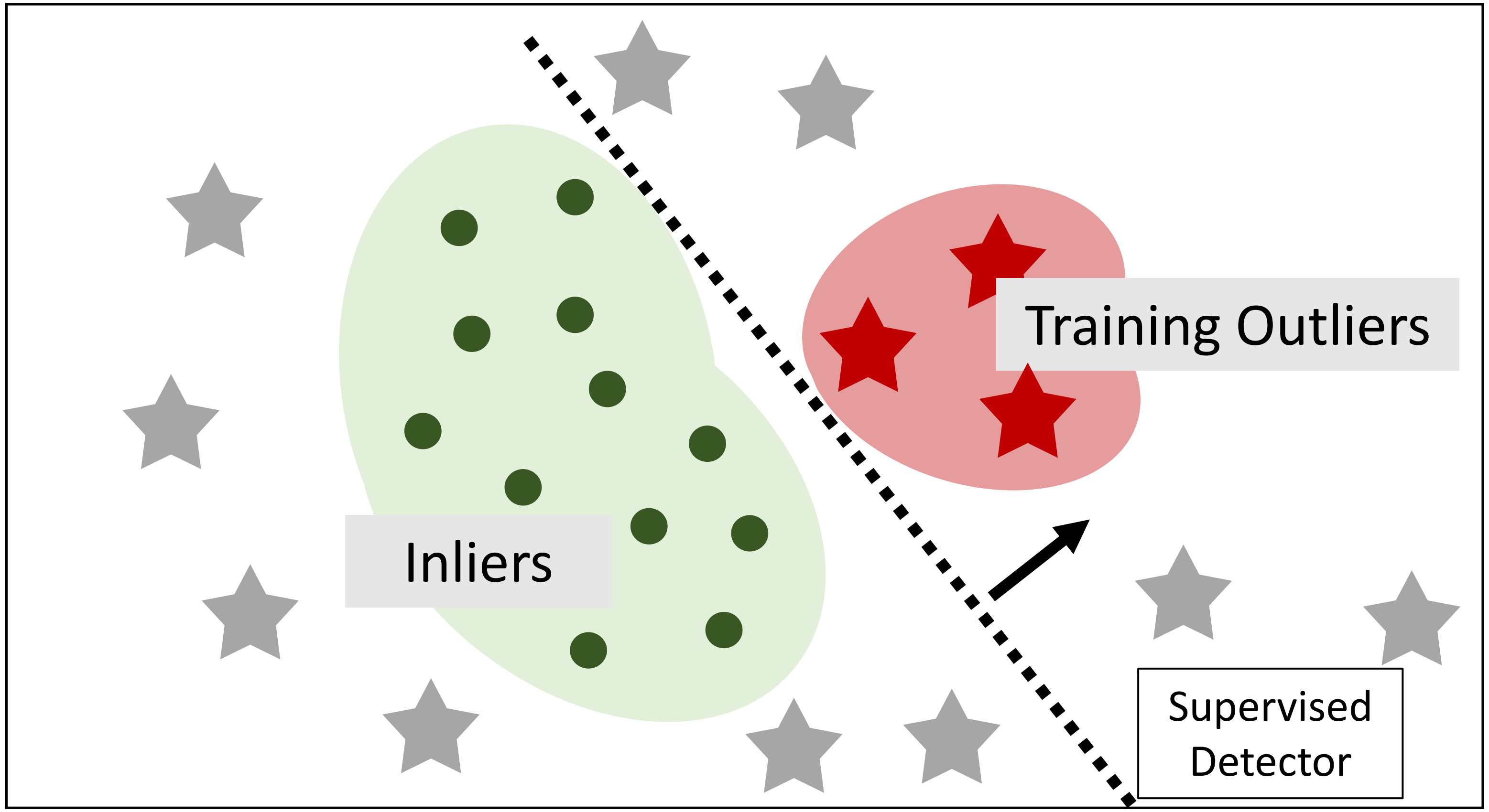
Testing:

Given a unlabeled "normal" and "abnormal" samples, predict which ones are "abnormal."



Supervised Approach for Anomaly Detection?

- The detector may miss outliers of unseen types



Terminology

I will use the following terms interchangeably:

- Anomaly (↔Normal)
- Outlier (↔Inlier)
- Novelty
- Out-of-distribution (OOD) sample (↔In-distribution sample)

Related problems:

- Anomaly detection with contaminated training data
- OOD detection (additionally assumes inlier class, e.g., dogs, cats,…)

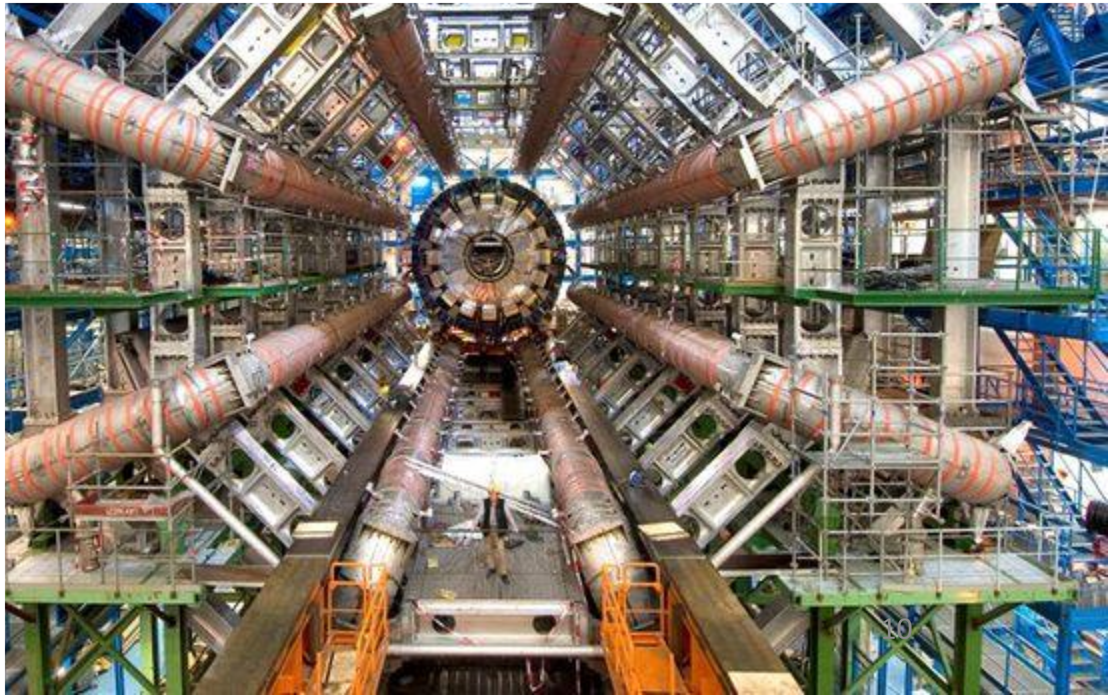
The Dark Machines Anomaly Score Challenge: Benchmark Data and Model Independent Event Classification for the Large Hadron Collider

T. Aarrestad[a] M. van Beekveld[b] M. Bona[c] A. Boveia[e] S. Caron[d] J. Davies[c]
A. De Simone[f,g] C. Doglioni[h] J. M. Duarte[i] A. Farbin[j] H. Gupta[k] L. Hendriks[d]
L. Heinrich[a] J. Howarth[l] P. Jawahar[m,a] A. Jueid[n] J. Lastow[h] A. Leinweber[o]
J. Mamuzic[p] E. Merényi[q] A. Morandini[r] P. Moskvitina[d] C. Nellist[d] J. Ngadiuba[s,t]
B. Ostdiek[u,v] M. Pierini[a] B. Ravina[l] R. Ruiz de Austri[p] S. Sekmen[w]
M. Touranakou[x,a] M. Vaškevičiūte[l] R. Vilalta[y] J.-R. Vlimant[t] R. Verheyen[z]
M. White[o] E. Wulff[h] E. Wallin[h] K.A. Wozniak[α,a] Z. Zhang[d]

Challenges for Unsupervised Anomaly Detection in Particle Physics

Katherine Fraser, Samuel Homiller, Rashmish K. Mishra, Bryan Ostdiek, and Matthew D. Schwartz

Department of Physics, Harvard University, Cambridge, MA 02138, USA
The NSF AI Institute for Artificial Intelligence and Fundamental Interactions
E-mail: kfraser@g.harvard.edu, shomiller@g.harvard.edu, rashmishmishra@fas.harvard.edu, bostdiek@g.harvard.edu, schwartz@g.harvard.edu



Autoencoder-Based Anomaly Detection in Particle Physics

SciPost SciPost Phys. 6, 030 (2019)

QCD or what?

Theo Heimel[1], Gregor Kasieczka[2], Tilman Plehn[1*] and Jennifer M. Thompson[1]

1 Institut für Theoretische Physik, Universität Heidelberg, Germany
2 Institut für Experimentalphysik, Universität Hamburg, Germany

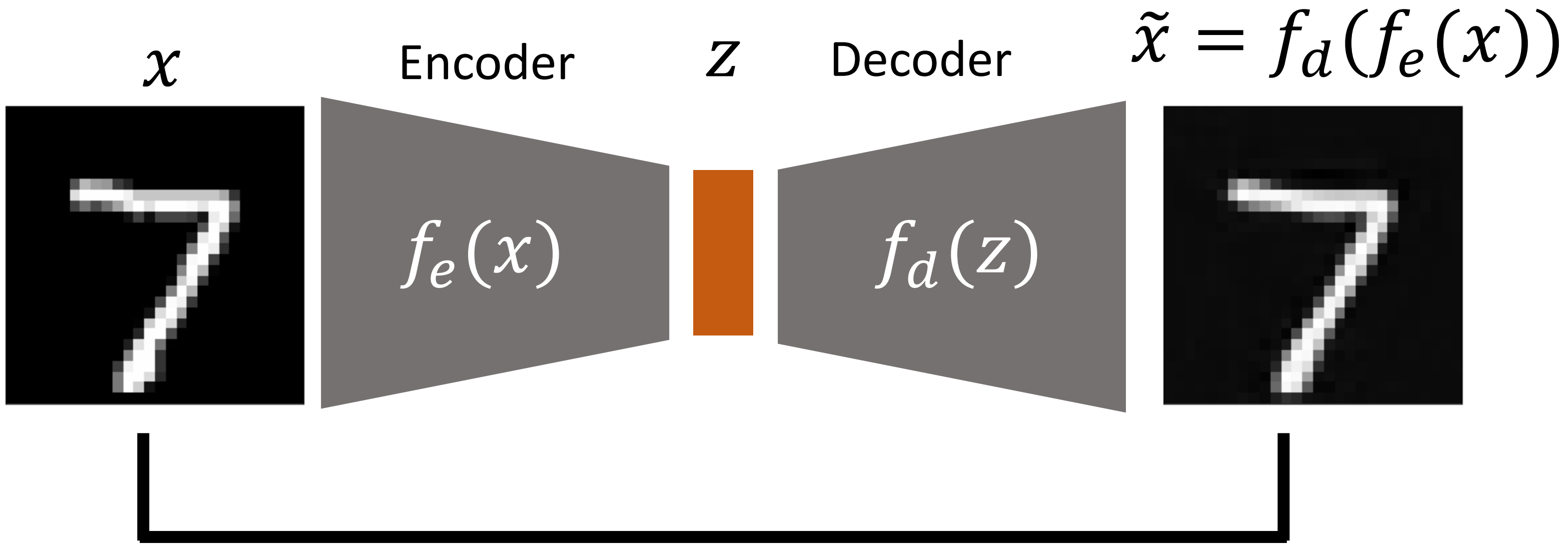
Searching for New Physics with Deep Autoencoders

Marco Farina, Yuichiro Nakai, and David Shih

NHETC, Dept. of Physics and Astronomy
Rutgers, The State University of NJ
Piscataway, NJ 08854 USA

(2018)

Autoencoders (Rumelhart et al., 1986)

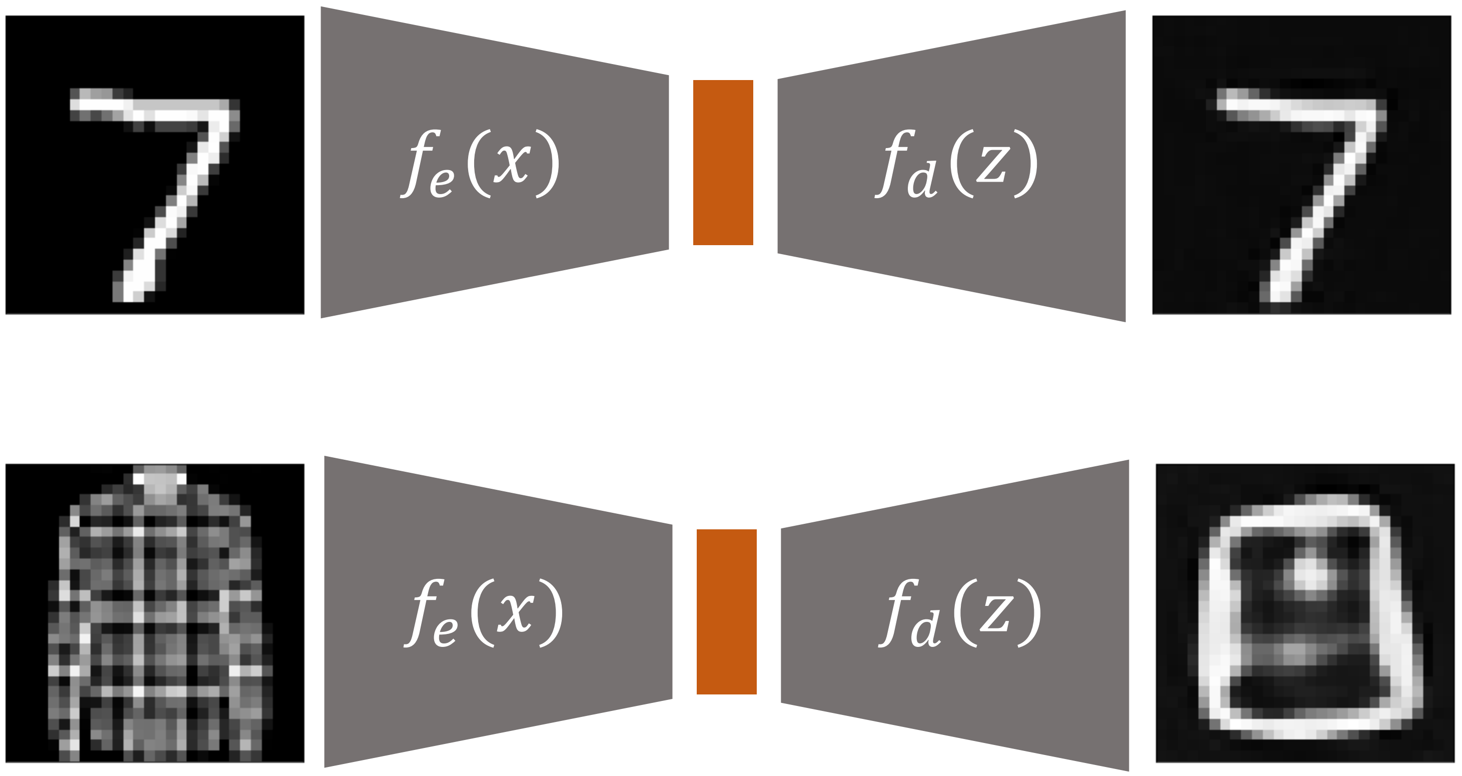


Reconstruction Error

$$l_\theta(x) = \|x - \tilde{x}\|^2$$

$$\min_\theta \mathbb{E}_{x \sim p(x)}[l_\theta(x)]$$

Autoencoder-based Outlier Detection (Japkowicz et al., 1995)



Small $l_\theta(x)$ — "Inlier"

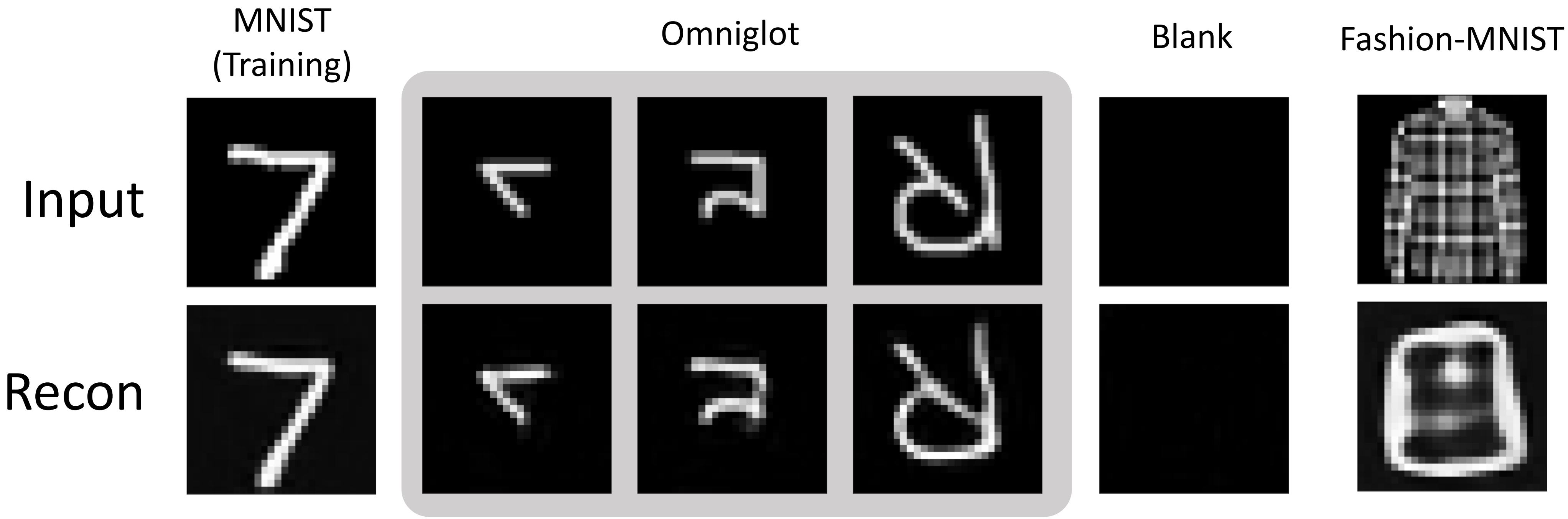
Large $l_\theta(x)$ — "Outlier"

Assumption

An autoencoder is not able to reconstruct outliers.

Outlier Reconstruction

An autoencoder does reconstruct outliers



Interactive Web Demo: https://swyoon.github.io/outlier-reconstruction/
or google “**outlier reconstruction web demo**”

Autoencoding Under Normalization Constraints

International Conference on Machine Learning 2021
Machine Learning Summer School 2021 Taipei Best Poster Award
Qualcomm Innovation Fellowship 2021

Sangwoong Yoon

Seoul National University

Yung-Kyun Noh

Hanyang University

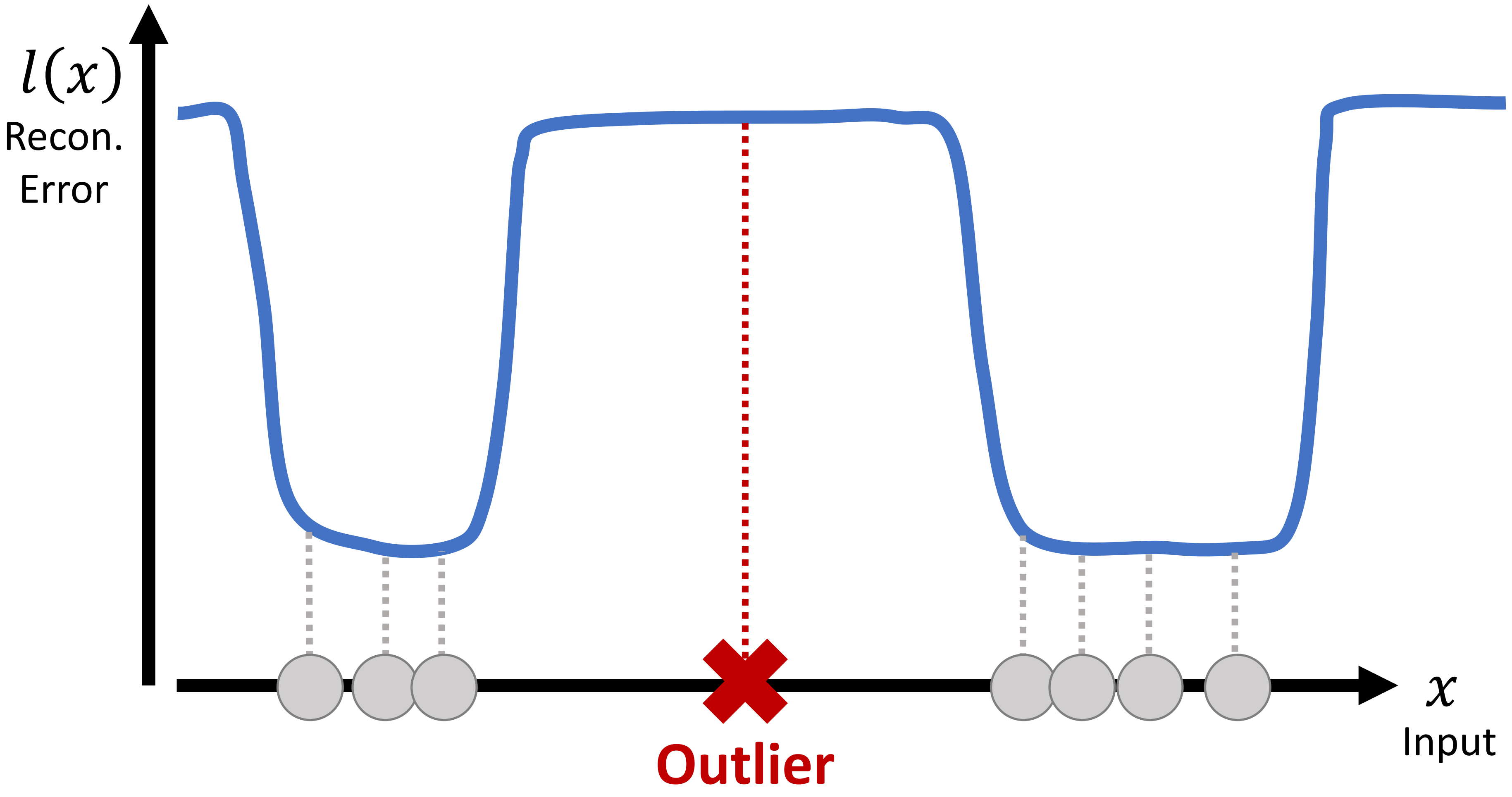
Korea Institute of Advanced Studies

Frank Chongwoo Park

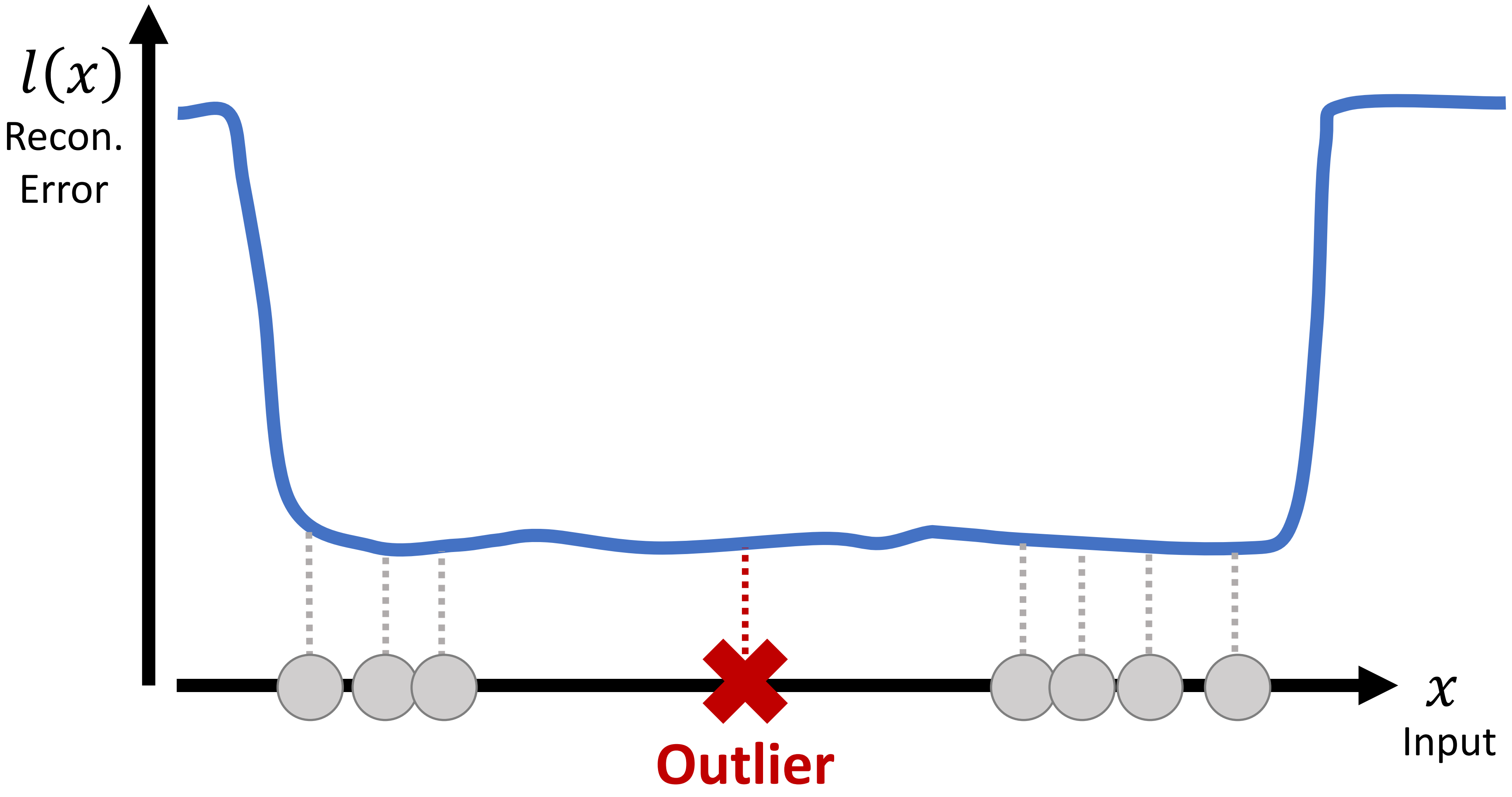
Seoul National University

Saige Research

Reconstruction Error – Ideal Case



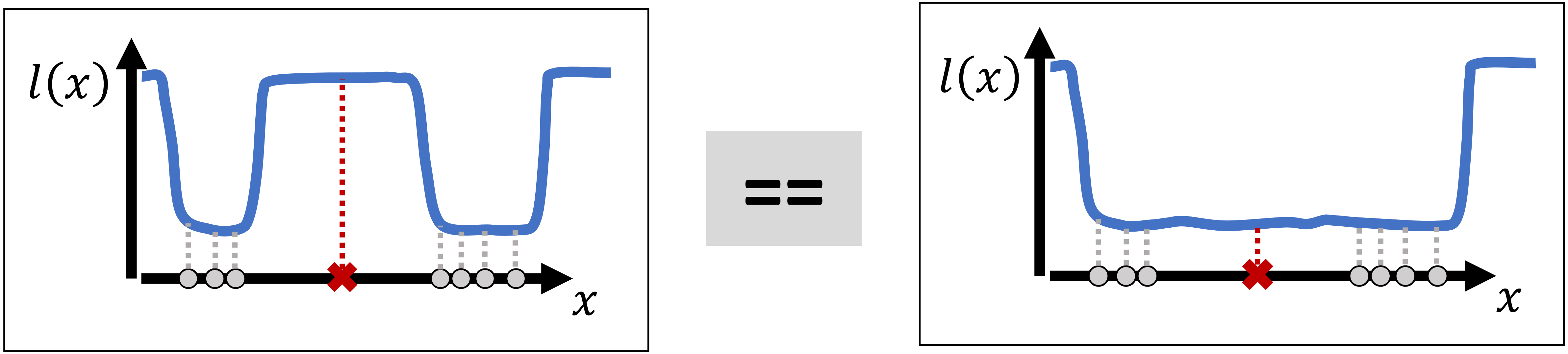
Reconstruction Error In Reality



Autoencoder's Blind Spot

Autoencoder's loss function does not differentiate the two cases

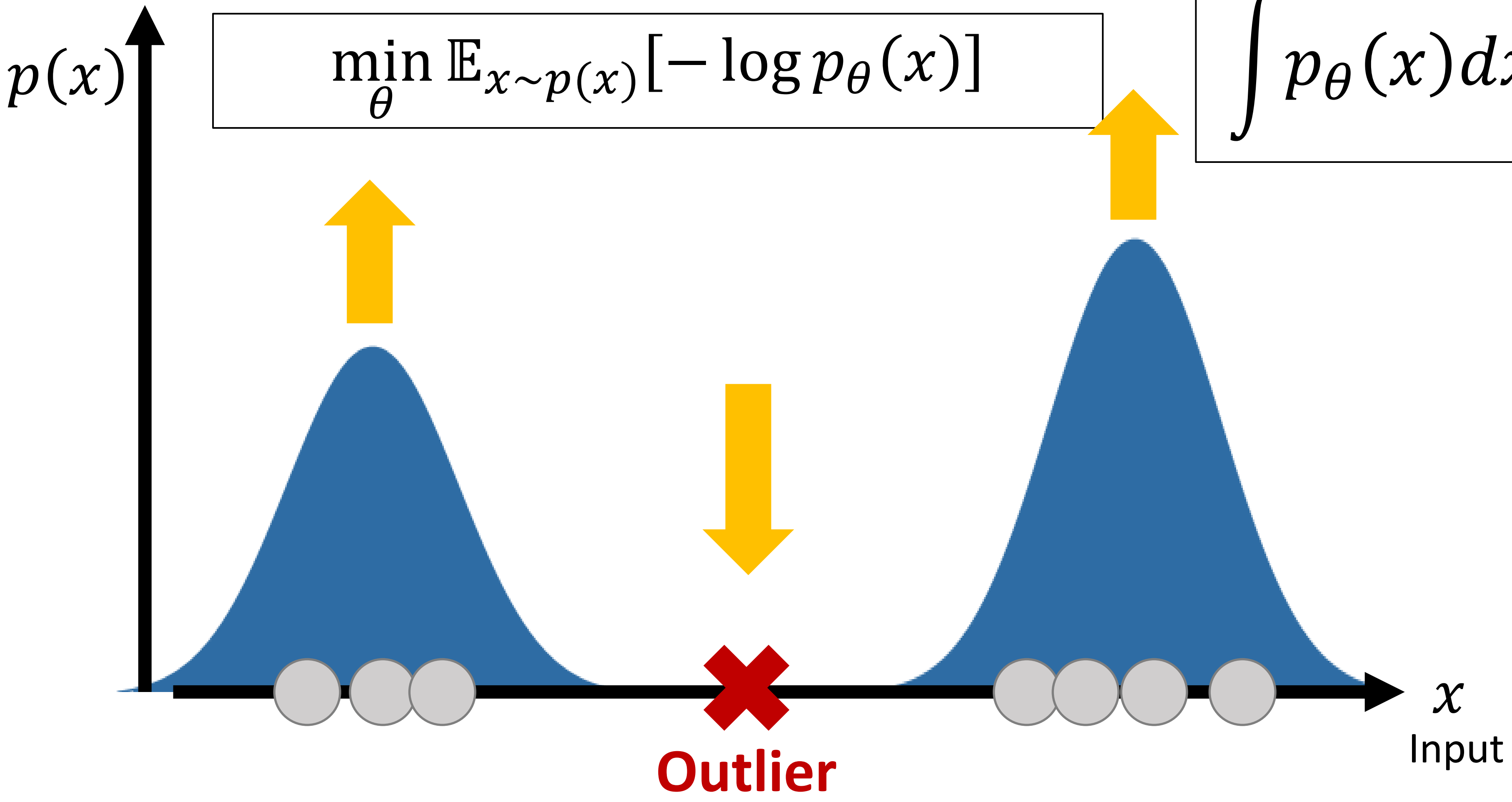
$$\min_{\theta} \mathbb{E}_{x \sim p(x)}[l_{\theta}(x)]$$



Maximum Likelihood Learning

$$\min_{\theta} \mathbb{E}_{x \sim p(x)}[-\log p_\theta(x)]$$

$$\int p_\theta(x)dx = 1$$



How Can We Bring Normalization Constraint to Autoencoders?

Energy-based Models or Boltzmann Distributions

$$p_\theta(x) = \frac{1}{\Omega_\theta} e^{-E_\theta(x)}, \qquad \Omega_\theta = \int e^{-E_\theta(x)} dx$$

Normalization Constant

$$E_\theta(x) \downarrow \Rightarrow p_\theta(x) \uparrow$$

Energy $E_\theta(x)$ defines a probabilistic model $p_\theta(x)$

Normalized Autoencoders (NAE)

We set energy $E_\theta(x)$ as reconstruction error $l_\theta(x)$

$$E_\theta(x) = l_\theta(x) = \|x - \tilde{x}\|^2$$

$$p_\theta(x) = \frac{1}{\Omega_\theta} e^{-l_\theta(x)}, \qquad \Omega_\theta = \int e^{-l_\theta(x)} dx$$

Maximum Likelihood for NAE

$$\min_\theta \mathbb{E}_{x \sim p(x)}[l_\theta(x)] + \log \Omega_\theta$$

Minimizing Reconstruction Error

Enforcing Normalization

Training of NAE

Gradient of Likelihood

$$\mathbb{E}_{x\sim p(x)}[\nabla_\theta \log p_\theta(x)] = -\mathbb{E}_{x\sim p(x)}[\nabla_\theta l_\theta(x)] + \mathbb{E}_{x'\sim p_\theta(x')}[\nabla_\theta l_\theta(x')]$$

Decreasing **Real Data** Reconstruction Error

Increasing **Generated Sample** Reconstruction Error

Suppression of Outlier Reconstruction

An outlier x^* has a small $l_\theta(x^*)$

→ $p_\theta(x^*)$ is large

→ x^* is sampled from $p_\theta(x)$

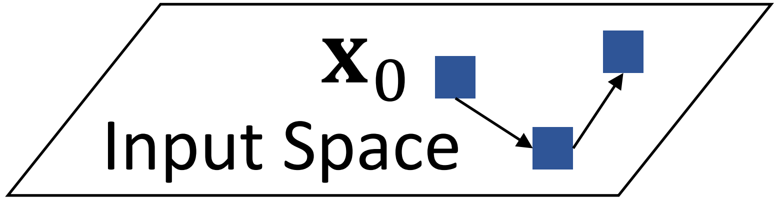
→ $l_\theta(x^*)$ is increased by $\nabla_\theta l_\theta(x^*)$ → Outlier Reconstruction Suppressed!

Sampling from NAE $\mathbf{x}' \sim p_\theta(\mathbf{x}')$

$$\mathbb{E}_{\mathbf{x}' \sim p_\theta(\mathbf{x}')}[\nabla_\theta l_\theta(\mathbf{x}')]$$

Increasing **Generated Sample** Reconstruction Error

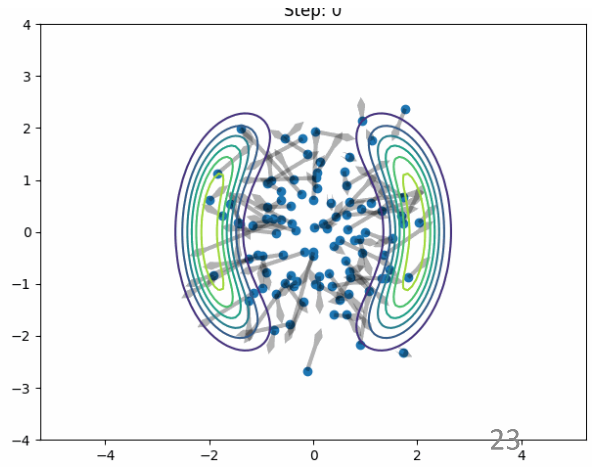
Langevin Monte Carlo (LMC) Sampling



$t = 0, \cdots, T$

$\mathbf{x}_0 \sim$ Noise distribution

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{\lambda_1}{2} \nabla_{\mathbf{x}} E_\theta(\mathbf{x}_t) + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \lambda_2)$$



- $\mathbf{x}_T$: a sample from $p_\theta(\mathbf{x})$
- $E_\theta(\mathbf{x}) = l_\theta(\mathbf{x})$
- λ_1, λ_2: Hyperparameters. Step size and noise strength.

On-Manifold Initialization (OMI)

Initialize $\mathbf{x}_0$ on the decoder manifold

Decoder manifold $\mathcal{M} = \{x | x = f_d(z), z \in \mathcal{Z}\}$

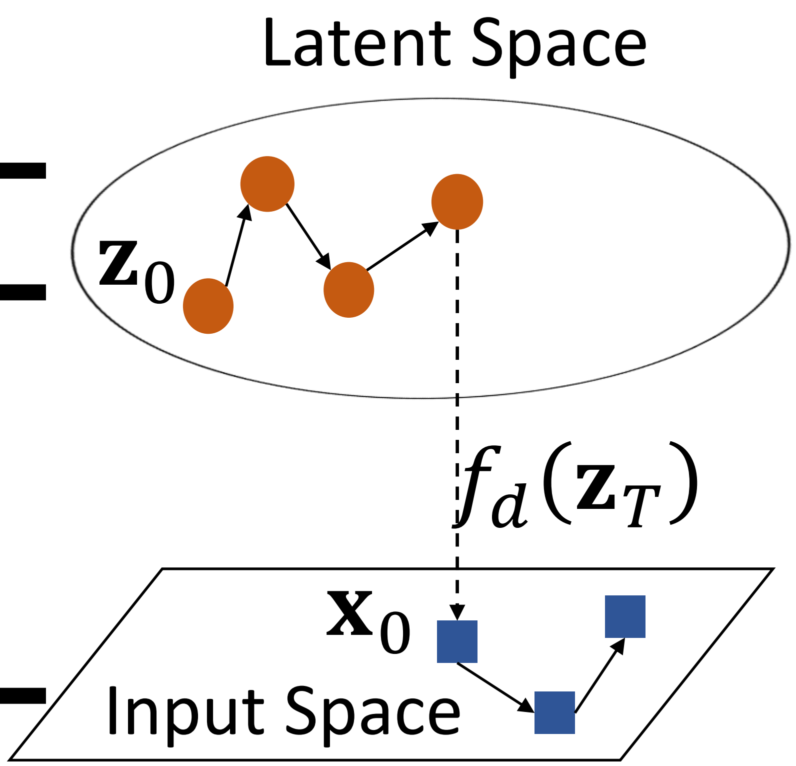
Algorithm Sampling with OMI

Initialize $\mathbf{z}_0$ randomly

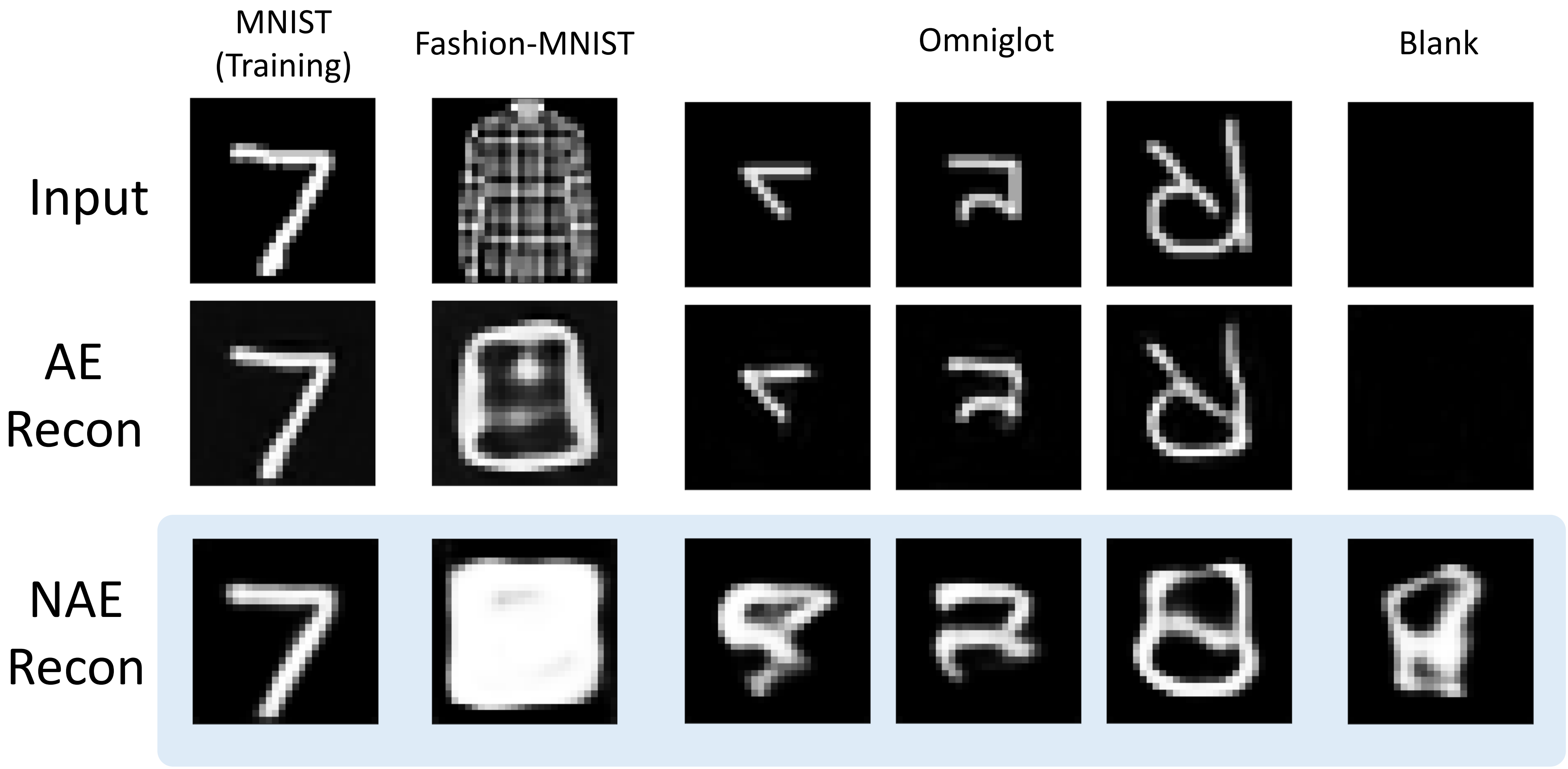
Sample $\mathbf{z}_T = \mathrm{LangvinMonteCarlo}(\mathbf{z}_0)$ // Latent Chain

Decode $\mathbf{x}_0 = f_d(\mathbf{z}_T)$ // **On-Manifold Initialization**

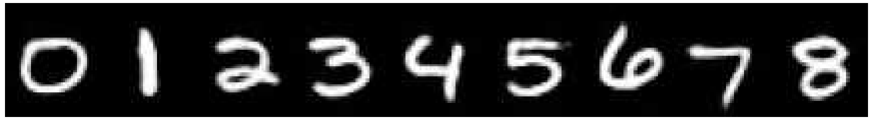
Sample $\mathbf{x}_{T'} = \mathrm{LangevinMonteCarlo}(\mathbf{x}_0)$ // Visible Chain



Suppressed Outlier Reconstruction

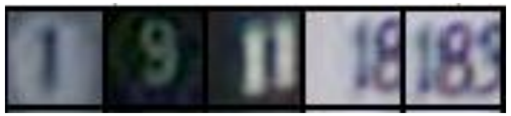


MNIST 0 - 8 (in) vs **MNIST 9** (out)




	AUROC
NAE	**.934**
AE	.537
DAE	.537
VAE(R)	.721
VAE(L)	.614
WAE	.799
GLOW	.426
PixelCNN++	.497
IGEBM	.522

CIFAR-10 (in) vs **SVHN** (out)

	AUROC
NAE	**.920**
AE	.175
DAE	.175
VAE(R)	.191
VAE(L)	.185
WAE	.168
GLOW	.260
PixelCNN++	.074
IGEBM	.371

MNIST Hold-out Class Detection Experiment

Table 1. MNIST hold-out class detection AUC scores. The values in parentheses denote the standard error of mean after 10 training runs.

HOLD-OUT:	0	1	2	3	4	5	6	7	8	9	AVG
NAE-OMI	**.989**(.002)	**.919**(.013)	**.992**(.001)	**.949**(.004)	**.949**(.005)	**.978**(.003)	**.938**(.004)	**.975**(.024)	**.929**(.004)	**.934**(.005)	**.955**
NAE-CD	.799	.098	.878	.769	.656	.806	.874	.537	.876	.500	.679
NAE-PCD	.745	.114	.879	.754	.690	.813	.872	.509	.902	.544	.682
AE	.819	.131	.843	.734	.661	.755	.844	.542	.902	.537	.677
DAE	.769	.124	.872	.935	.884	.793	.865	.533	.910	.625	.731
VAE(R)	.954	.391	.978	.910	.860	.939	.916	.774	.946	.721	.839
VAE(L)	.967	.326	.976	.906	.798	.927	.928	.751	.935	.614	.813
WAE	.817	.145	.975	**.950**	.751	.942	.853	.912	.907	.799	.805
GLOW	.803	.014	.624	.625	.364	.561	.583	.326	.721	.426	.505
PXCNN++	.757	.030	.663	.663	.483	.642	.596	.307	.810	.497	.545
IGEBM	.926	.401	.642	.644	.664	.752	.851	.572	.747	.522	.672
DAGMM	.386	.304	.407	.435	.444	.429	.446	.349	.609	.420	.423

Dataset-vs-Dataset Experiment

Inlier: CIFAR-10

In: CIFAR-10	ConstantGray	FMNIST	SVHN	CelebA	Noise
NAE	**.963**	**.819**	**.920**	**.887**	1.0
AE	.006	.650	.175	.655	1.0
DAE	.001	.671	.175	.669	1.0
VAE(R)	.002	.700	.191	.662	1.0
VAE(L)	.002	.767	.185	.684	1.0
WAE	.000	.649	.168	.652	1.0
GLOW	.384	.222	.260	.419	1.0
PXCNN++	.000	.013	.074	.639	1.0
IGEBM	.192	.216	.371	.477	1.0

Inlier: ImageNet 32x32

In: ImageNet32	ConstantGray	FMNIST	SVHN	CelebA	Noise
NAE	**.966**	**.994**	**.985**	**.949**	1.0
AE	.005	.915	.102	.325	1.0
DAE	.069	.991	.102	.426	1.0
VAE(R)	.030	.936	.132	.501	1.0
VAE(L)	.028	.950	.132	.545	1.0
WAE	.069	.991	.081	.364	1.0
GLOW	.413	.856	.169	.479	1.0
PXCNN++	.000	.004	.027	.238	1.0

Sample Generation $\mathbf{x}' \sim p_\theta(\mathbf{x}')$

MNIST (28x28)

CelebA (64x64)

Summary

Outlier Reconstruction

A fundamental glitch in **autoencoder-based outlier detection**

Normalized Autoencoders

A novel autoencoder based on **energy-based formulation** where outlier reconstruction is naturally suppressed

- NAE was adopted by particle physicists to detect anomalous signals in Large Hardron Collider.
- NAE outperformed all other AE-based methods in particle physics simulation datasets.

(at Institute for Theoretical Physics, Heidelberg University, 2023 March)

SciPost Physics Submission

A Normalized Autoencoder for LHC Triggers

Barry M. Dillon[1], Luigi Favaro[1], Tilman Plehn[1], Peter Sorrenson[2], and Michael Krämer[3]

1 Institut für Theoretische Physik, Universität Heidelberg, Germany
2 Heidelberg Collaboratory for Image Processing, Universität Heidelberg, Germany
3 Institute for Theoretical Particle Physics and Cosmology (TTK), RWTH Aachen University, Germany

February 28, 2023

Acknowledgments

First, we would like to thank Ullrich Köthe for many inspiring discussions on neural network architectures. We are also grateful to the Mainz Institute for Theoretical Physics, where this paper was finalized. LF would like to thank Sangwoong Yoon for his constant and reliable support. This research is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant 396021762 – TRR 257: *Particle Physics Phenomenology after the Higgs Discovery* and through Germany's Excellence Strategy EXC 2181/1 - 390900948 (the Heidelberg STRUCTURES Excellence Cluster).

NAE for LHC Trigger (Dillon et al., 2022)

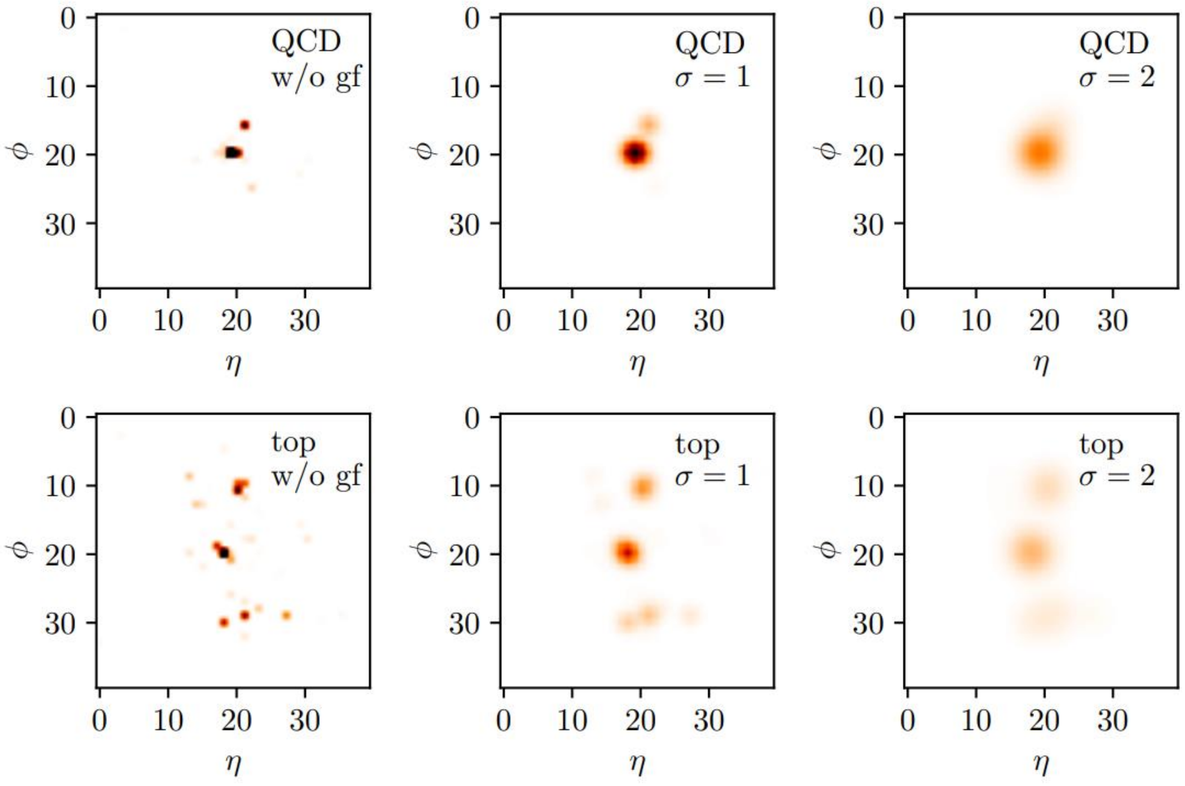


Figure 1: Example QCD and top images without Gaussian filter, $\sigma_G = 1$, and $\sigma_G = 2$.

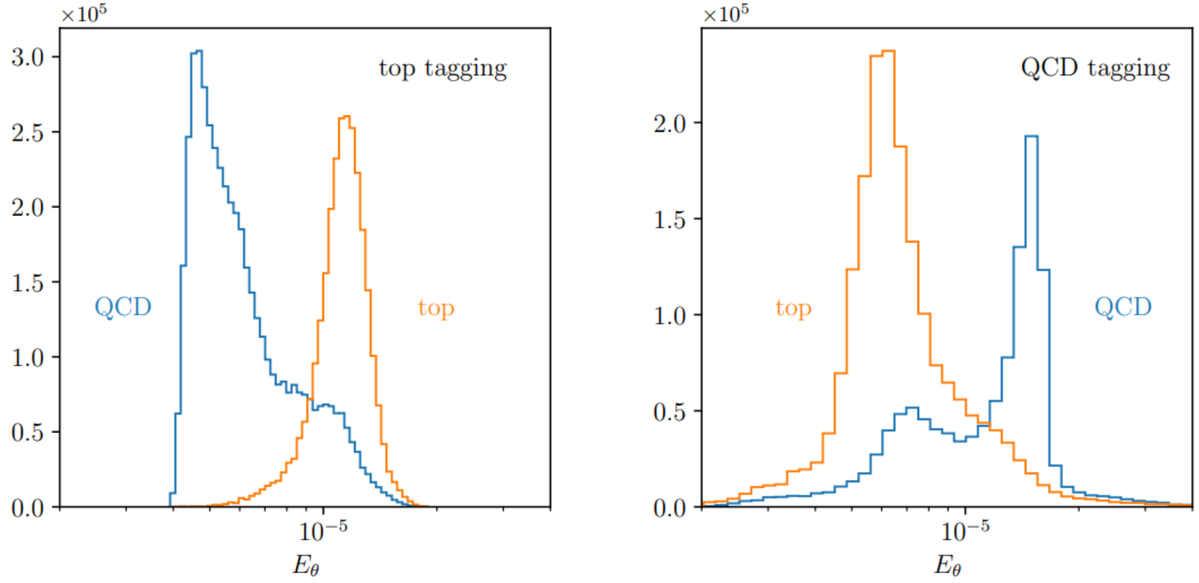


Figure 3: Distribution of the energy or MSE after training on QCD jets (left) and on top jets (right). We show the energy for QCD jets (blue) and top jets (orange) in both cases.

ML4Jets2023

Nov 6 – 10, 2023
DESY
Europe/Zurich timezone

Enter your search term

- Overview
- Conference Picture
- Call for Abstracts
- Timetable
- Contribution List
- Book of Abstracts
- Registration
- Participant List
- Slack Channel
- Travel and Accommodation
- Venue
- Code of Conduct

Contact

ml4jets2023-info@desy.de

Unsupervised tagging of semivisible jets with normalized autoencoders in CMS

Nov 9, 2023, 9:45 AM
15m
Main Auditorium (DESY)

Anomalies

Speaker

Florian Eble (ETH Zurich (CH))

Description

Semivisible jets are a novel signature of dark matter scenarios where the dark sector is confining and couples to the Standard Model via a portal. They consist of jets of visible hadrons intermixed with invisible stable particles that escape detection. In this work, we use normalized autoencoders to tag semivisible jets in proton-proton collisions at the CMS experiment. Unsupervised models are desirable in this context since they can be trained on background only, and are thus robust with respect to the details of the signal. The use of an autoencoder as an anomaly detection algorithm relies on the assumption that the network better reconstructs examples it was trained on than ones from a different probability distribution i.e., anomalies. Using the search for semivisible jets as a benchmark, we demonstrate the tendency of autoencoders to generalize beyond the dataset they are trained on, hindering their performance. We show how normalized autoencoders, specifically designed to suppress this effect, give a sizable boost in performance. We further propose a modified loss function and signal-agnostic condition to reach the optimal performance.

Primary author

Florian Eble (ETH Zurich (CH))

Challenges 1: Inaccurate Langevin Monte Carlo

$$\mathbb{E}_{x\sim p(x)}[\nabla_\theta \log p_\theta(x)] = -\mathbb{E}_{x\sim p(x)}[\nabla_\theta E_\theta(x)] + \mathbb{E}_{x'\sim p_\theta(x')}[\nabla_\theta E_\theta(x')]$$

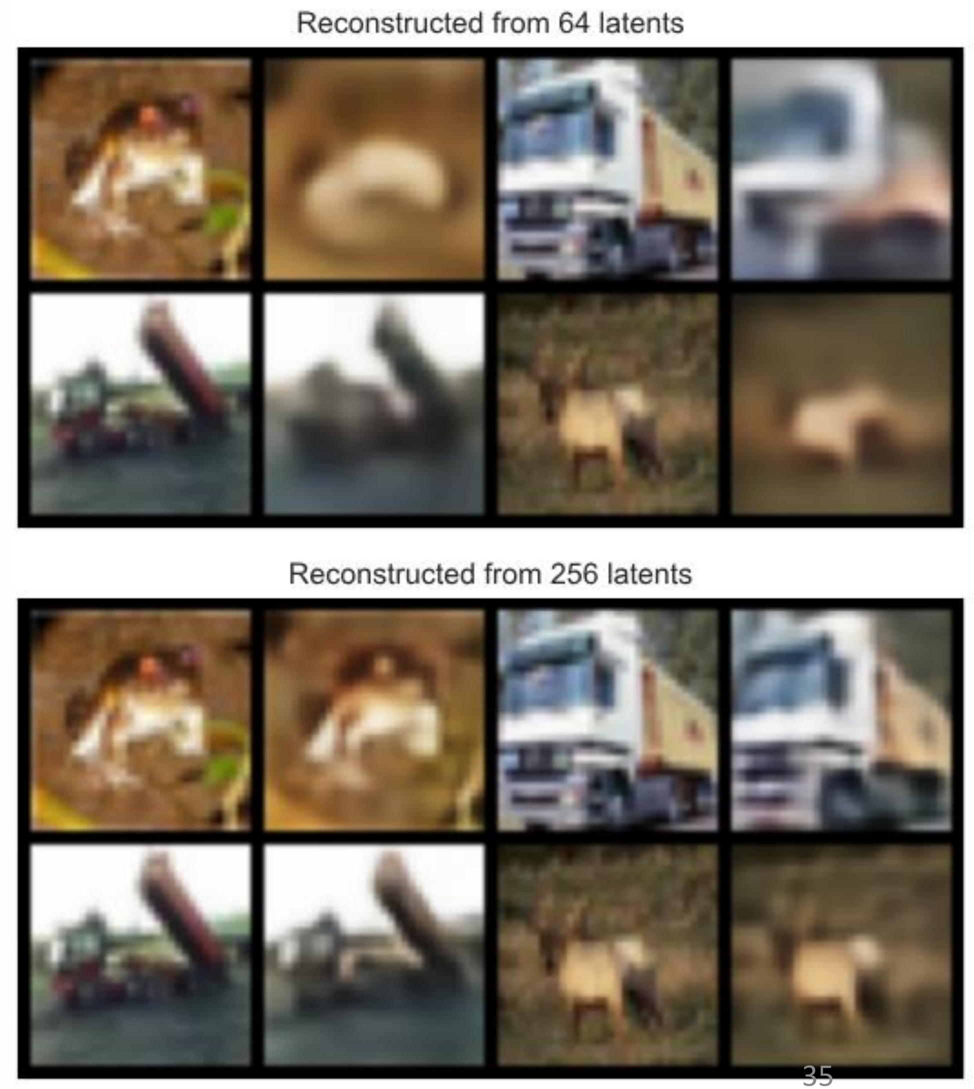
Decreasing **Real Data** Energy	Increasing **Generated Sample** Energy

- Too many hyperparameters
  - # of steps, step size, noise size, initialization,…
- Anomaly detection performance is highly sensitive to these hyperparameters
- For high-dimensional (>100) data, with a short (<100) MCMC, <u>sampling $x'\sim p_\theta(x')$ will never be exact</u>

Challenges 2: Burden of Reconstruction

- For high-dimensional complex signals, reconstruction is unnecessarily difficult requirement.

- Very difficult to achieve small $l(x)$ for in-distribution data.
  - The reconstruction assumption is violated conversely – too high for inliers.



Energy-Based Models for Anomaly Detection: A Manifold Diffusion Recovery Approach

Neural Information Processing Systems 2023

Sangwoong Yoon

Korea Institute for Advanced Study

Young-Uk Jin

Samsung Electronics

Yung-Kyun Noh

Hanyang University

Korea Institute for Advanced Study

Frank Chongwoo Park

Seoul National University

Saige Research

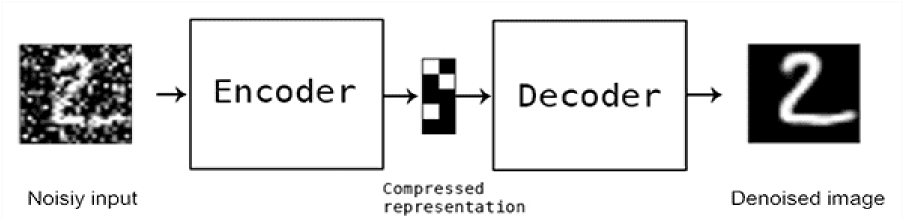
The Roles of Autoencoder in NAE

1. Energy function $E(\mathbf{x}) = \|\mathbf{x} - \tilde{\mathbf{x}}\|^2$
2. Negative sample generation
   - On-Manifold Initialization

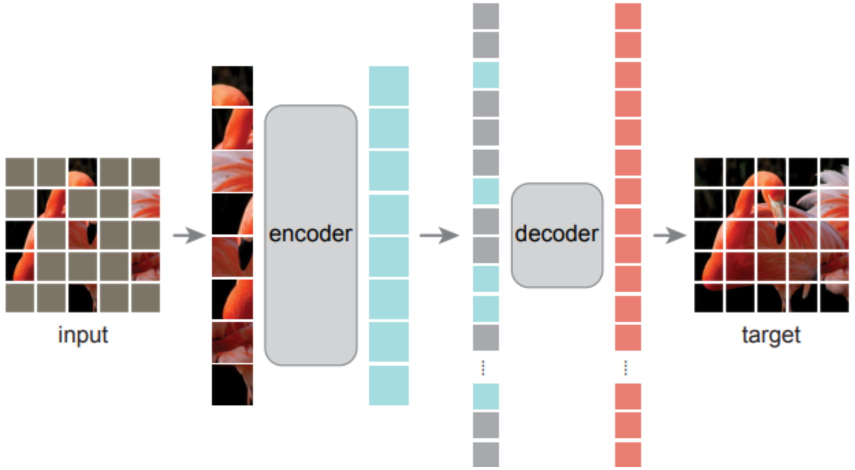
A single autoencoder serving two roles may limit scalability and flexibility

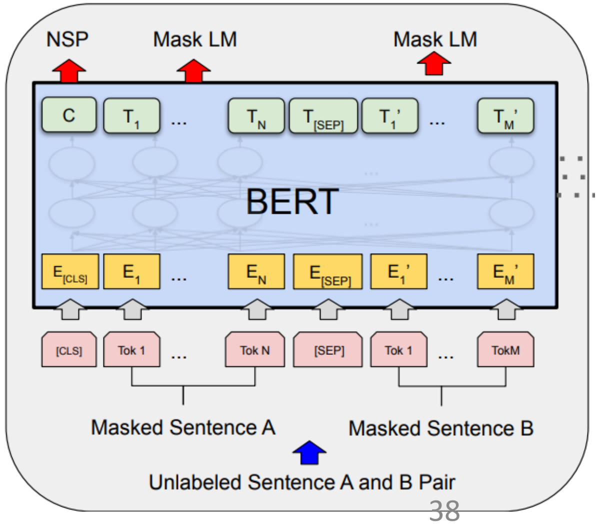
(https://pyimagesearch.com/2020/02/24/denoising-autoencoders-with-keras-tensorflow-and-deep-learning/)

Learning Through Recovery from Perturbation

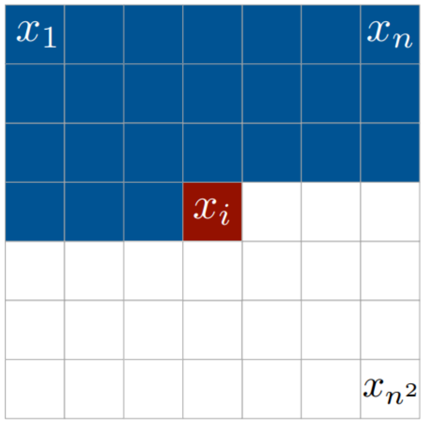


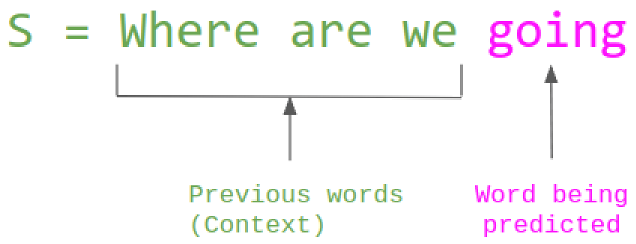
- Denoising autoencoder (Vincent et al., 2008)
- Masked autoencoder (He et al., 2021)
- BERT (Devlin et al., 2018)
- Autoregressive models
  - PixelCNN (van den Oord et al., 2016)
  - Language models (incl. ChatGPT)





Pre-training





P(S) = P(Where) x P(are | Where) x P(we | Where are) x P(going | Where are we)

https://thegradient.pub/understanding-evaluation-metrics-for-language-models/

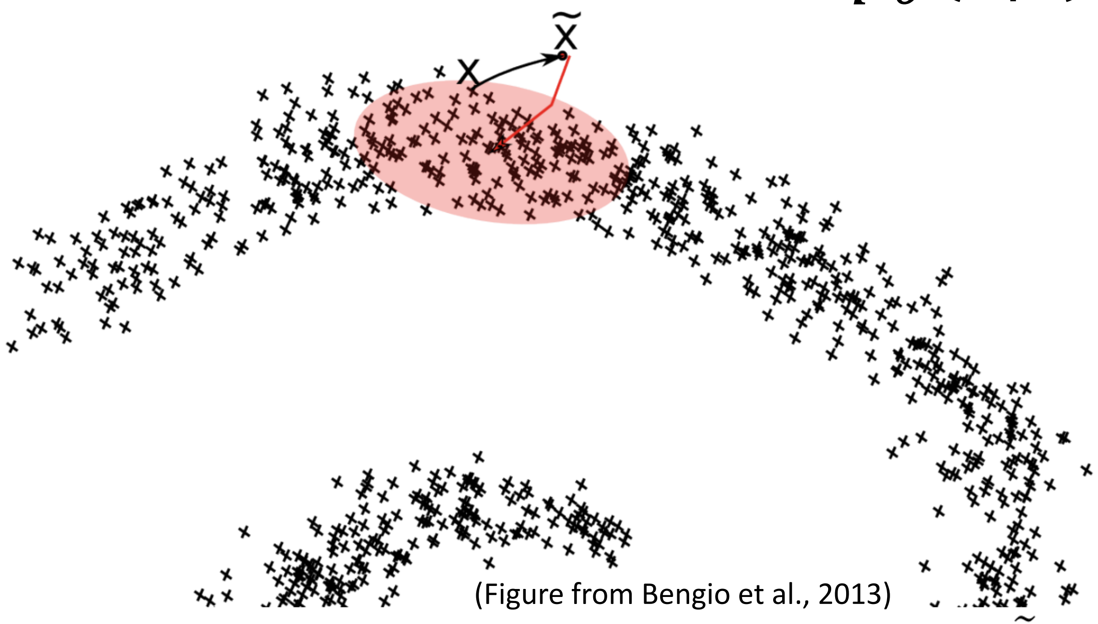
Recovery Likelihood (Bengio et al., 2013; Gao et al., 2021)

- Perturb data $\mathbf{x}$ with a Gaussian noise: $\mathbf{x} \rightarrow \tilde{\mathbf{x}}$

$$\tilde{\mathbf{x}} = \mathbf{x} + \sigma\epsilon, \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I})$$

- Recovery likelihood $p_\theta(\mathbf{x}|\tilde{\mathbf{x}})$
  - The probability of data $\mathbf{x}$ given its perturbed version $\tilde{\mathbf{x}}$

$$p_\theta(\mathbf{x}|\tilde{\mathbf{x}}) \propto p(\tilde{\mathbf{x}}|\mathbf{x})p_\theta(\mathbf{x}) \qquad \text{(Bayes' rule)}$$



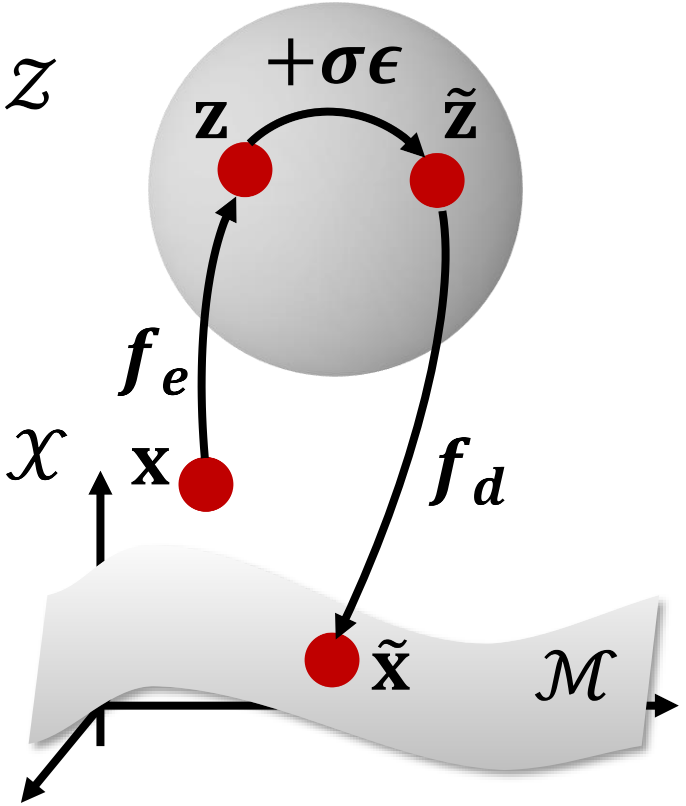
(Figure from Bengio et al., 2013)

Model $p_\theta(\mathbf{x})$ can be trained by maximizing $\log p_\theta(\mathbf{x}|\tilde{\mathbf{x}})$

Manifold Projection-Diffusion (MPD)

A novel perturbation operation using autoencoder (f_e, f_d)

1. Projection: $\mathbf{x} \xrightarrow{f_e} \mathbf{z}$
2. Diffusion: $\mathbf{z} \xrightarrow{+\sigma\epsilon} \tilde{\mathbf{z}}$

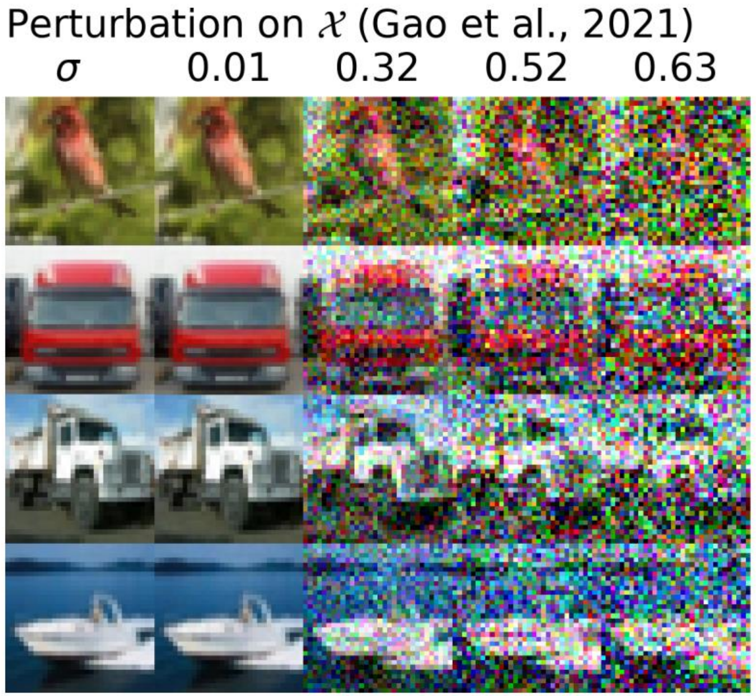


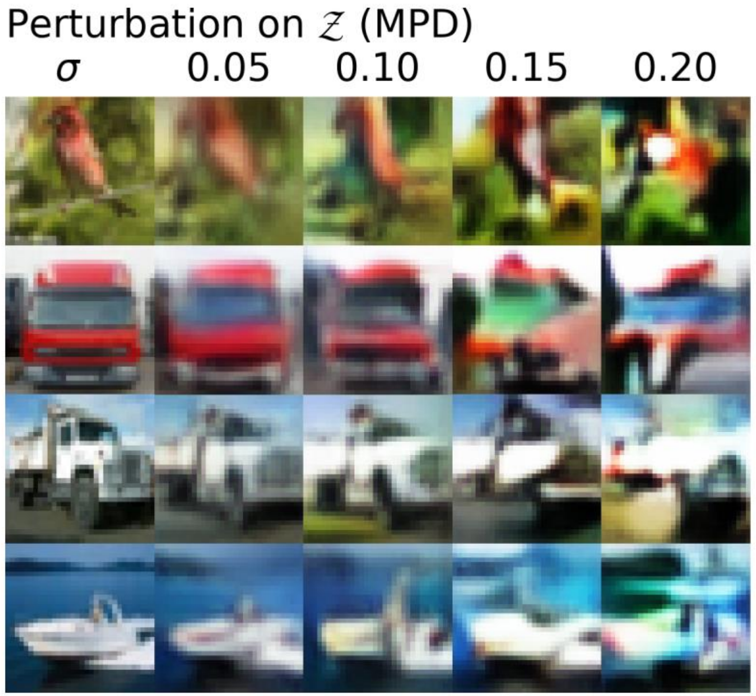
Recovery Likelihood $\equiv p(\mathbf{x}|\tilde{\mathbf{z}})$

Training algorithm maximizing $\log p(\mathbf{x}|\tilde{\mathbf{z}})$ is **Manifold Projection-Diffusion Recovery (MPDR)**

MPDR is a consistent estimation algorithm for $p(\mathbf{x})$

Gaussian vs Manifold Projection-Diffusion





MPD introduces lower-frequency perturbation

MPDR Training for EBM

Maximum Likelihood Training

$$\nabla_\theta \log p_\theta(\mathbf{x}) = -\nabla_\theta E_\theta(\mathbf{x}) + \mathbb{E}_{\mathbf{x}^- \sim p_\theta(\mathbf{x})}[\nabla_\theta E_\theta(\mathbf{x}^-)]$$

Maximum Recovery Likelihood Training

$$\nabla_\theta \log p_\theta(\mathbf{x}|\tilde{\mathbf{x}}) = -\nabla_\theta E_\theta(\mathbf{x}) + \mathbb{E}_{\mathbf{x}^- \sim p_\theta(\mathbf{x}|\tilde{\mathbf{x}})}[\nabla_\theta E_\theta(\mathbf{x}^-)]$$

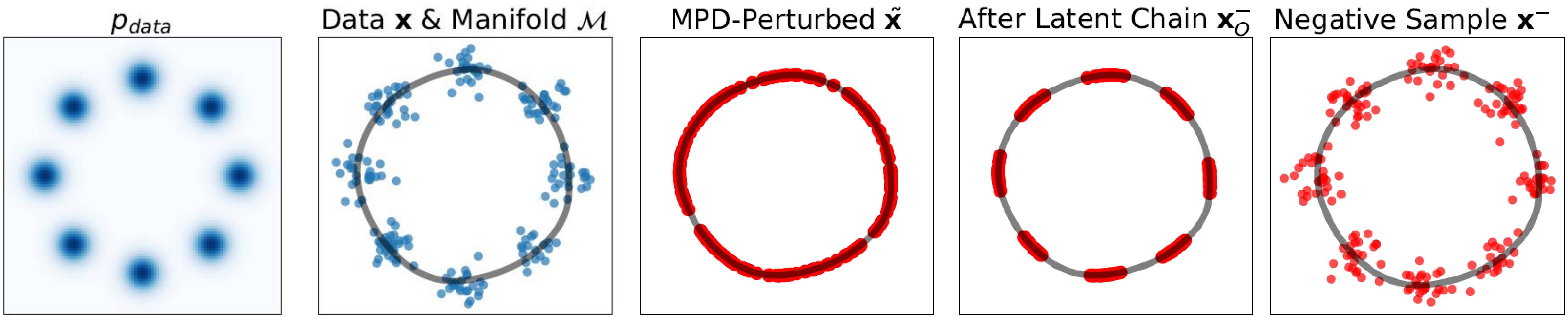
MPDR Training

$$\nabla_\theta \log p_\theta(\mathbf{x}|\tilde{z}) = -\nabla_\theta E_\theta(\mathbf{x}) + \mathbb{E}_{\mathbf{x}^- \sim p_\theta(\mathbf{x}|\tilde{z})}[\nabla_\theta E_\theta(\mathbf{x}^-)]$$

MPDR Negative Sampling Process

Two-Stage Sampling, similarly to NAE

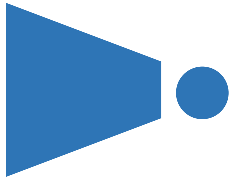
- Latent space MCMC → Input space MCMC



Energy Function Design

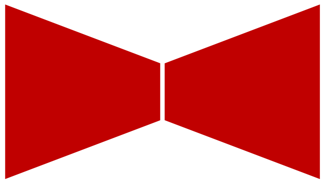
MPDR is compatible with any energy function:
Decoupling of energy and training algorithm.

MPDR-S (**S**calar)



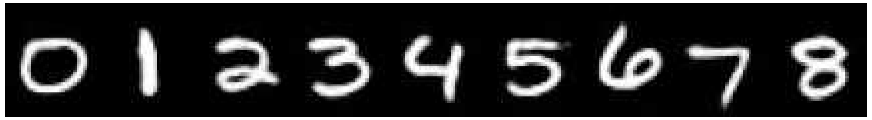
- $E_\theta(\mathbf{x}) = \mathrm{NN}_\theta(\mathbf{x})$, where $\mathrm{NN}: \mathcal{X} \rightarrow \mathbb{R}$ (e.g. ConvNet)

MPDR-R (**R**econstruction Error)



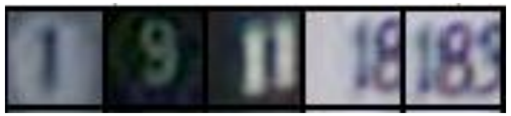
- $E_\theta(\mathbf{x}) = \|\mathbf{x} - \mathbf{x}_{recon}\|^2$, as in NAE
- The autoencoder need not to be the same with autoencoders in MPD

MNIST 0 - 8 (in) vs MNIST 9 (out)



	AUROC
MPDR-R	**.971**
NAE	**.934**
AE	.537
DAE	.537
VAE(R)	.721
VAE(L)	.614
WAE	.799
GLOW	.426
IGEBM	.522

CIFAR-10 (in) vs SVHN (out)

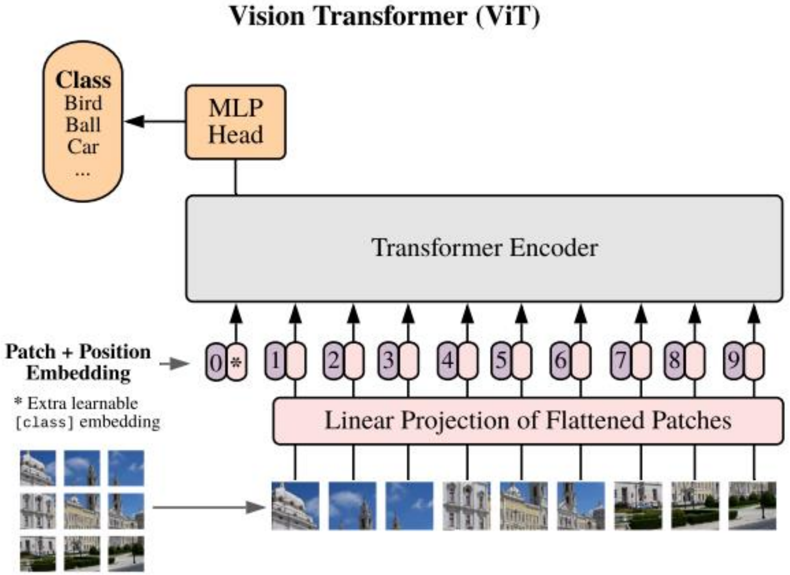


	AUROC
MPDR-R	**.981**
NAE	**.920**
AE	.175
DAE	.175
VAE(R)	.191
VAE(L)	.185
WAE	.168
GLOW	.260
IGEBM	.371

Anomaly Detection on Latent Space

- In-distribution: CIFAR-100
- Representation extracted from pretrained ViT-B_16

AUROC		CIFAR-10	SVHN	CelebA
Supervised	MD	.8634	.9638	.8833
	RMD	.9159	.9685	.4971
Unsup.	AE	.8580	.9645	.8103
	NAE	.8041	.9082	.8181
	IGEBM	.8217	.9584	.9004
	MPDR-S	.8338	.9911	**.9183**
	MPDR-R	**.8626**	**.9932**	.8662

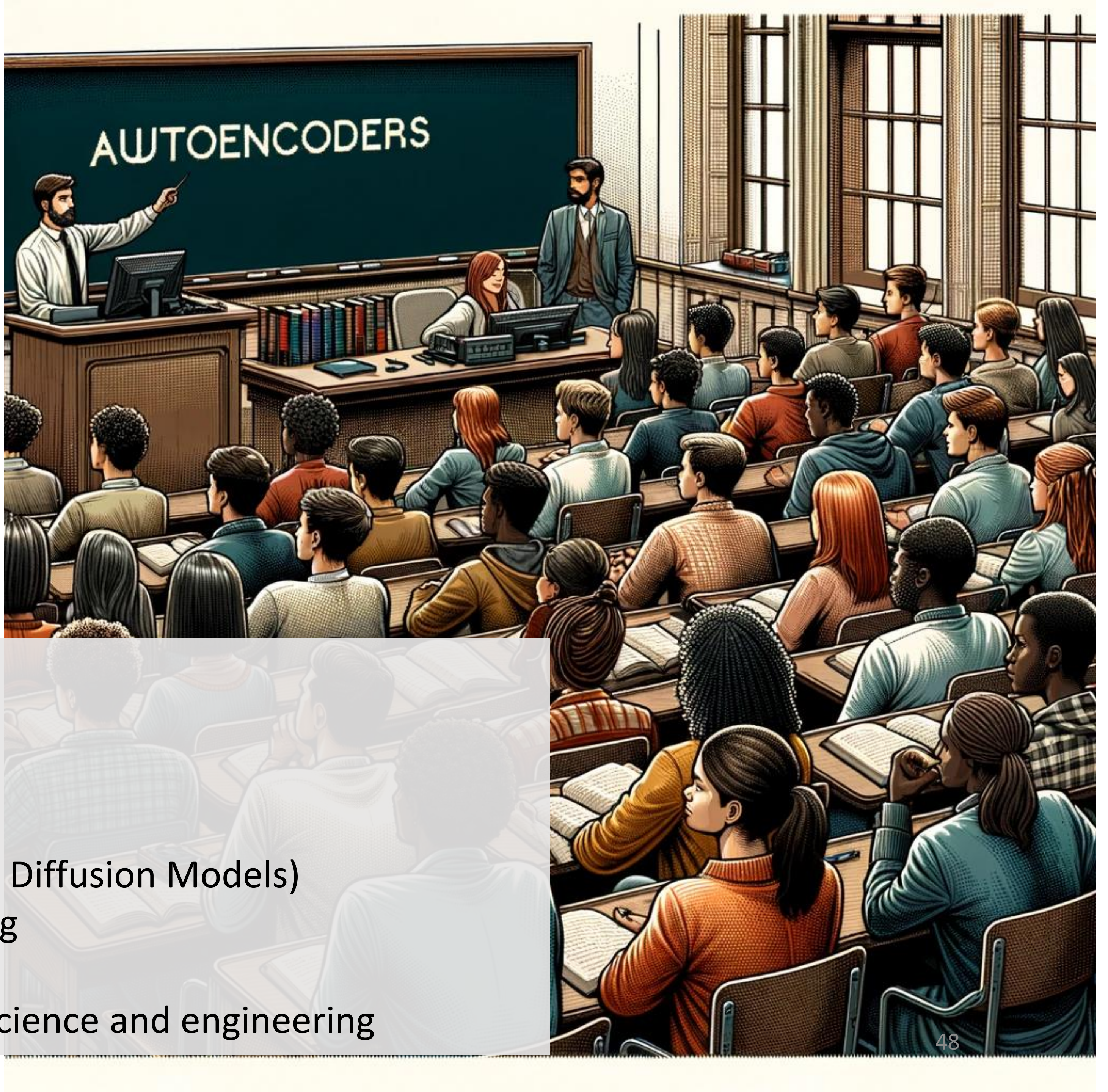


(Dosovitskiy et al., 2021)

Take-Home Messages

1. Autoencoders are prone to **reconstruct outliers**.
2. **EBM with autoencoder-based energy** is promising for anomaly detection.
3. Challenges are remaining:
    1. **EBM training** needs to be improved.
    2. Learning **good representation** is still important.

Thank you!



Sangwoong Yoon swyoon@kias.re.kr
AI Research Fellow @ KIAS

Open for research collaboration

- Generative modeling (Energy-Based Models, Diffusion Models) and its connection to Reinforcement Learning
- Out-of-Distribution detection
- Application of machine learning on natural science and engineering