

Deep Generative Model for Out-of-distribution Detection

深度生成模型在异常检测中的应用研究

Xuming Ran (冉旭明)

ranxuming@gamil.com

Department of Mathematics, Chongqing Jiaotong University

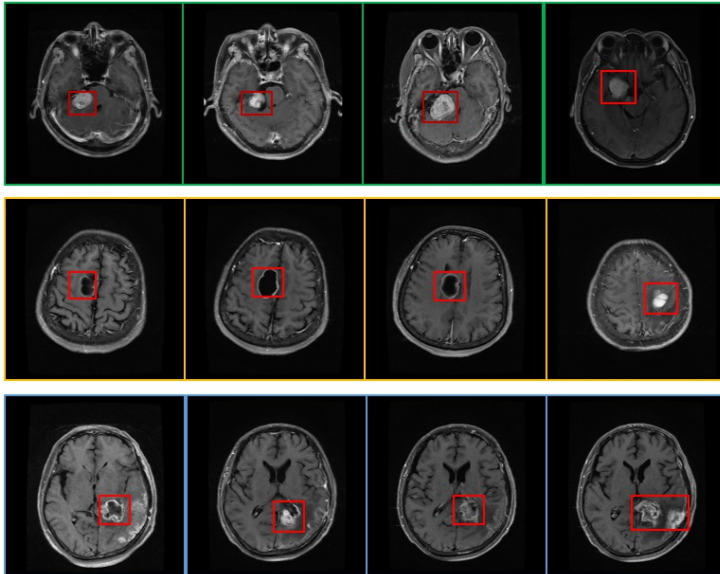
Department of Biomedical Engineering, Southern University of Science and Technology

➤ What is anomaly detection?

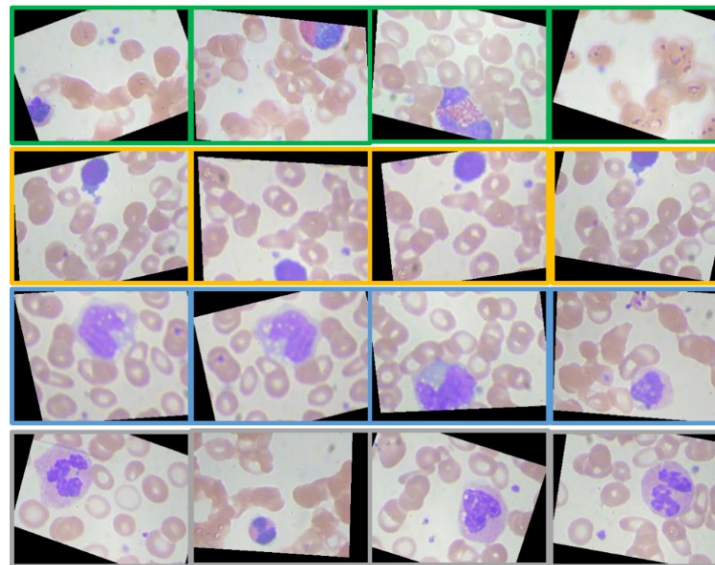
Anomaly detection is a technique used to identify **unusual patterns** that **do not** conform to expected behaviour. It is assumed that most of the training dataset consists of “normal” data, and we **do not** have a prior knowledge about anomalous data.

Cases:

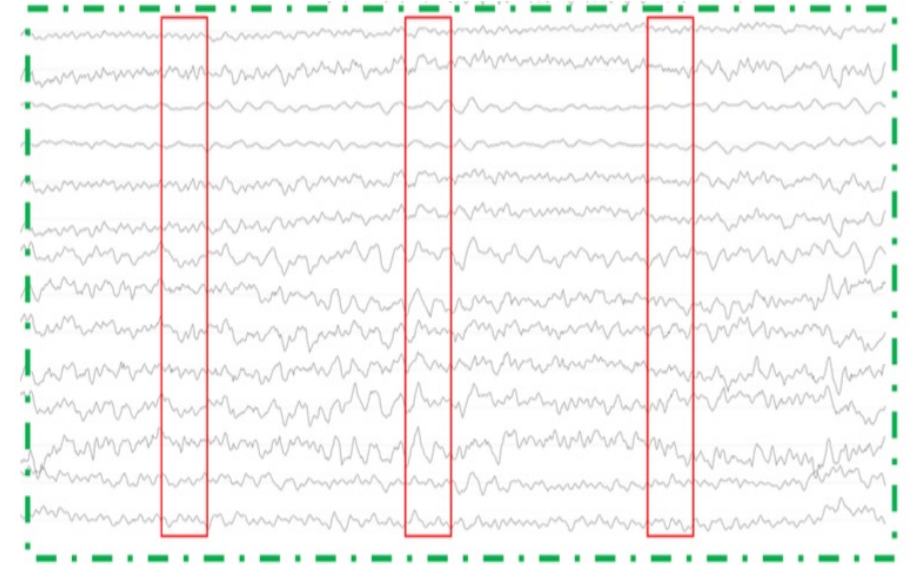
- 1) MRI Images from patients with brain lesion.
- 2) Slice image of white blood cells.
- 3) EEG signal from patients with epilepsy.



Cavernous Haemangioma(Top);
lymphoma(Middle); High grade
glioma(Bottom)

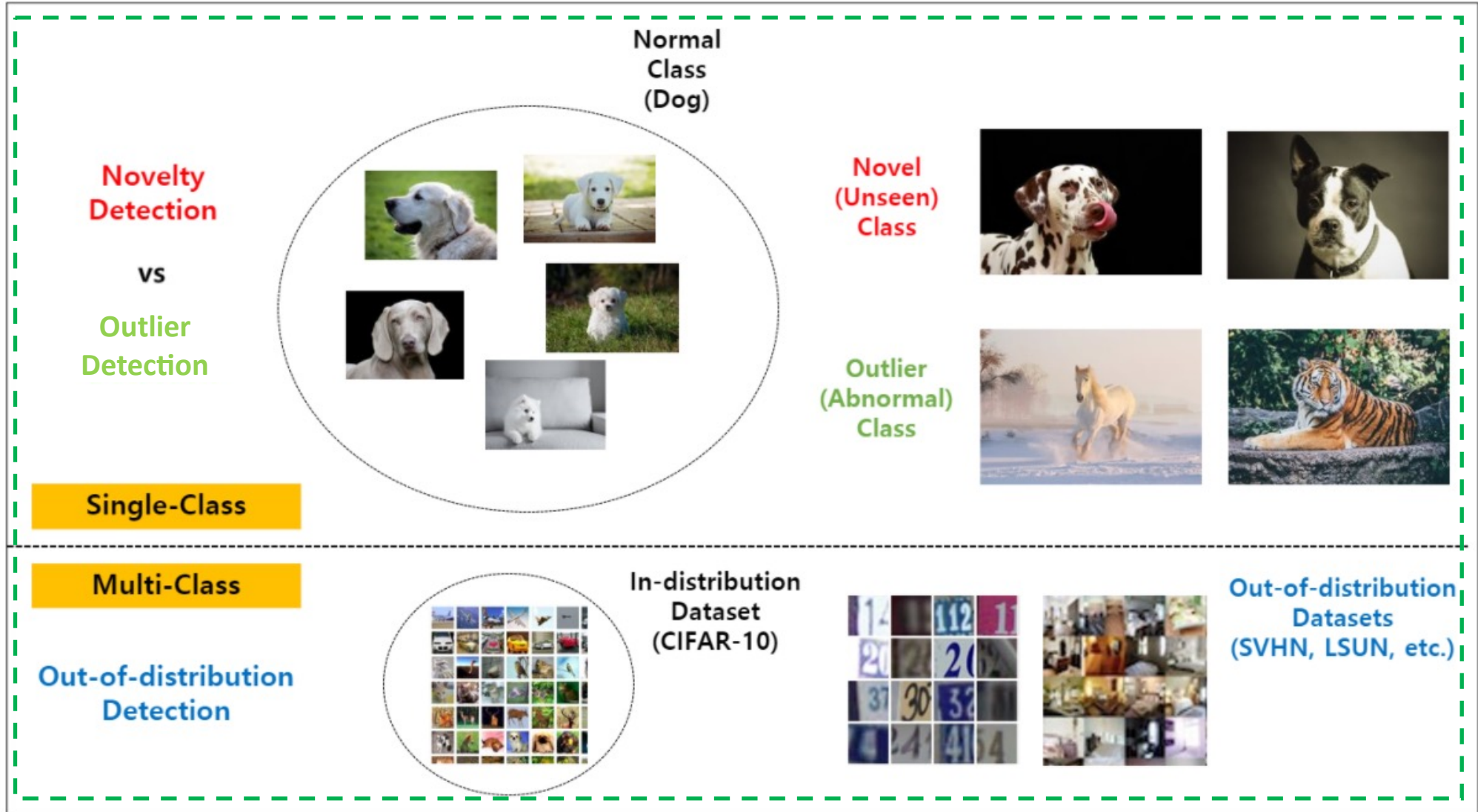


White blood cells. Eosinophils(First row),
Eosinophil (Second row), Lymphocyte(Third
row), Monocyte (Fourth row), and
Neutrophil(Fifth row).



EEG data.

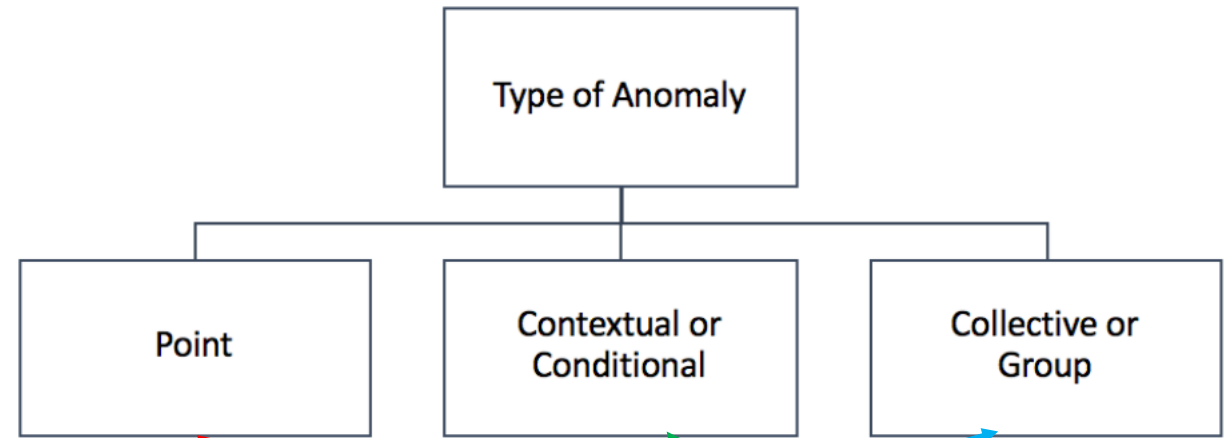
➤ Outlier detection vs. Novelty detection vs. Out-of-distribution detection



➤ Type of Anomaly

Anomalies can be broadly classified into three types:

- 1) point anomalies,
- 2) collective anomalies,
- 3) Contextual anomalies.



May-22	1:14 pm	FOOD	Monaco Café	\$1,127.80
May-22	2:14 pm	WINE	Wine Bistro	\$28.00
...				
Jun-14	2:14 pm	MISC	Mobil Mart	\$75.00
Jun-14	2:05 pm	MISC	Mobil Mart	\$75.00
Jun-15	2:06 pm	MISC	Mobil Mart	\$75.00
Jun-15	11:49 pm	MISC	Mobil Mart	\$75.00
May-28	6:14 pm	WINE	Acton shop	\$31.00
May-29	8:39 pm	FOOD	Crossroads	\$128.00
Jun-16	11:14 am	MISC	Mobil Mart	\$75.00
Jun-16	11:49 am	MISC	Mobil Mart	\$75.00

Illustrating Point and Collective anomaly.
Credit Card Fraud Detection:

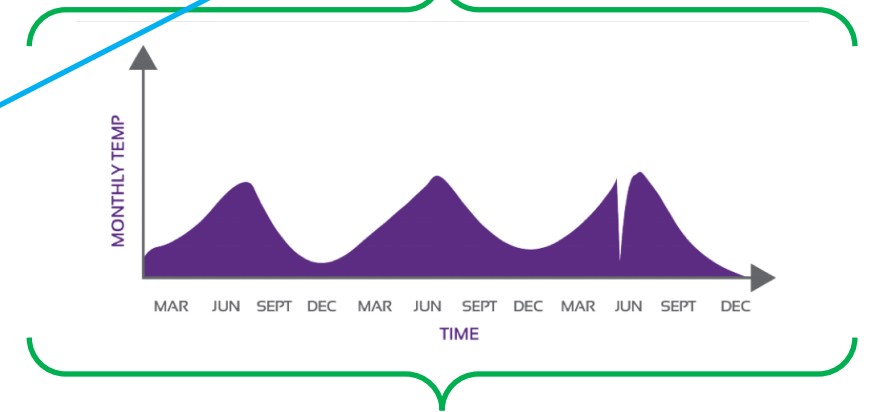
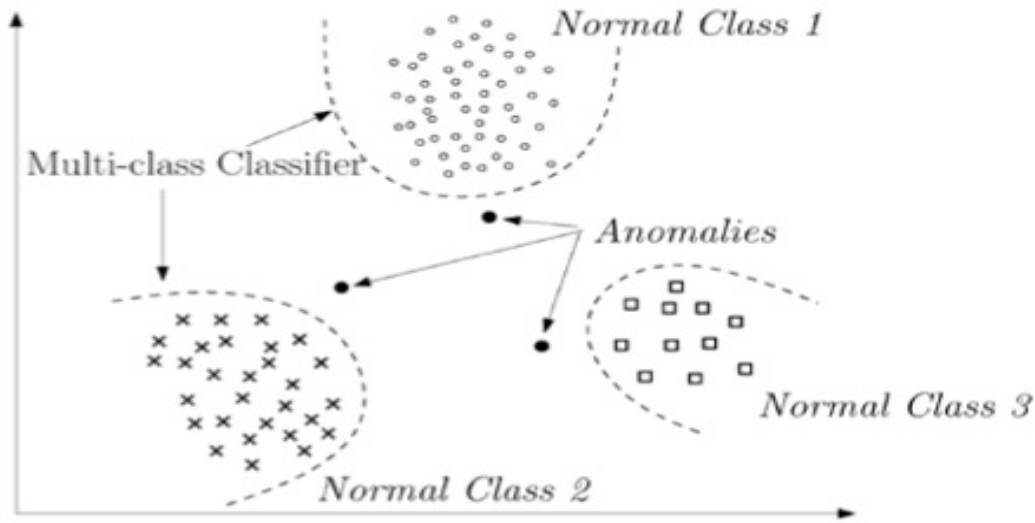


Illustration of contextual anomaly detection.
Temperature data

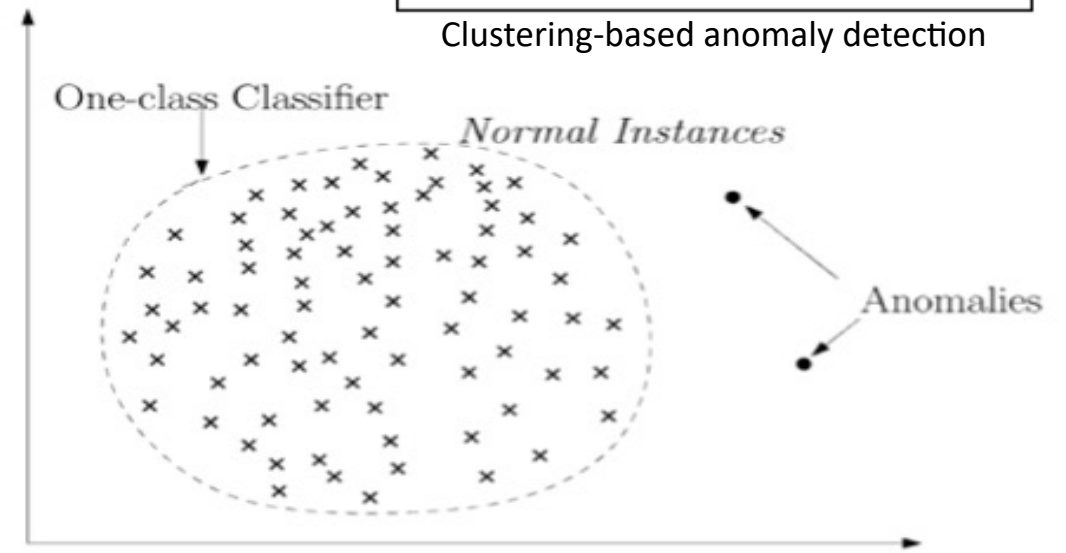
➤ Traditional anomaly detection

Traditional anomaly detection can be mainly classified into three types:

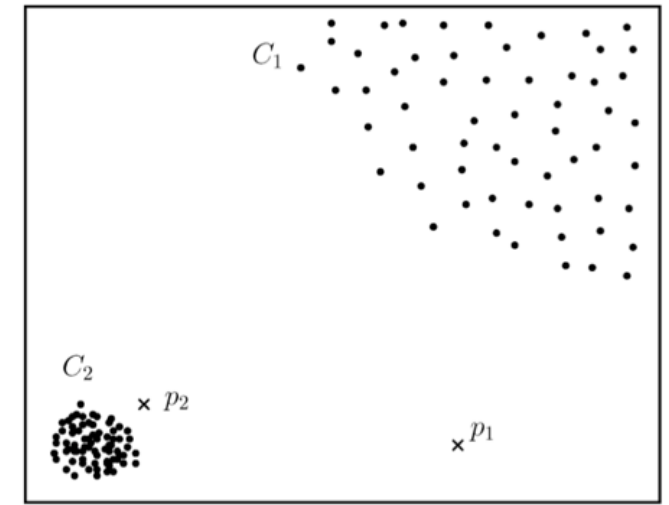
- 1) Classification based anomaly detection
 - a) Support vector machines (SVM)
 - b) One classification SVM
- 2) Clustering-based anomaly detection
 - a) Multivariate gaussian Models
 - b) K-Nearest-neighbour (KNN)



(a) Multi-class Anomaly Detection



(b) One-class Anomaly Detection



Clustering-based anomaly detection

➤ Deep anomaly detection (DAD) models

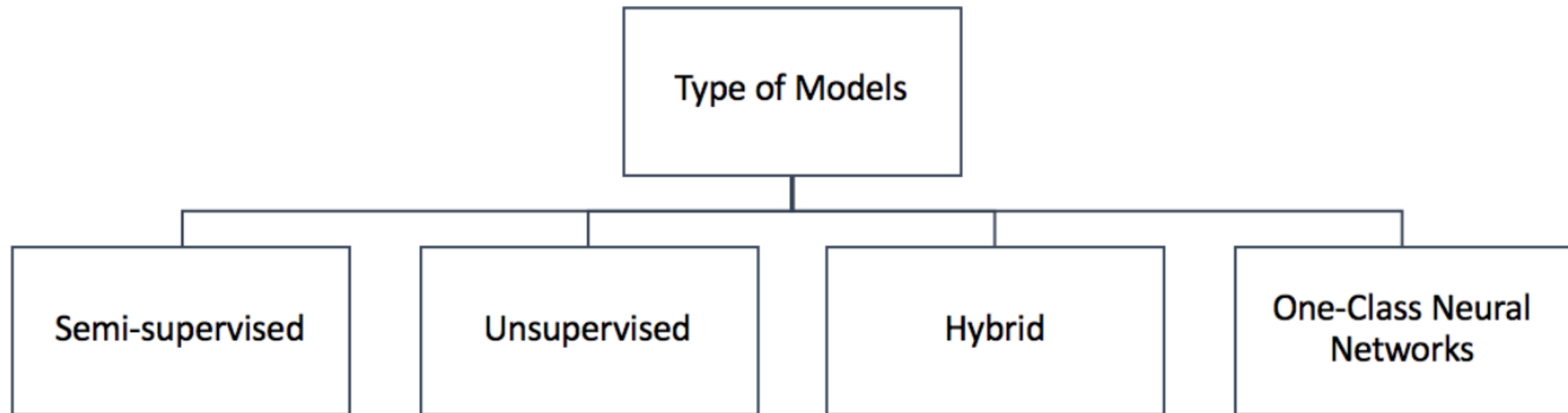
Deep anomaly detection (DAD) models can be broadly classified into three categories based on the extent of availability of labels.

(1) **Supervised** deep anomaly detection.

(2) **Semi-supervised** deep anomaly detection.

Labelled data is very hard to obtain.

(3) **Unsupervised** deep anomaly detection: based on intrinsic properties of the data instances



➤ Applications of Deep Anomaly Detection

Several applications of deep anomaly detection can be broadly classified into four types:

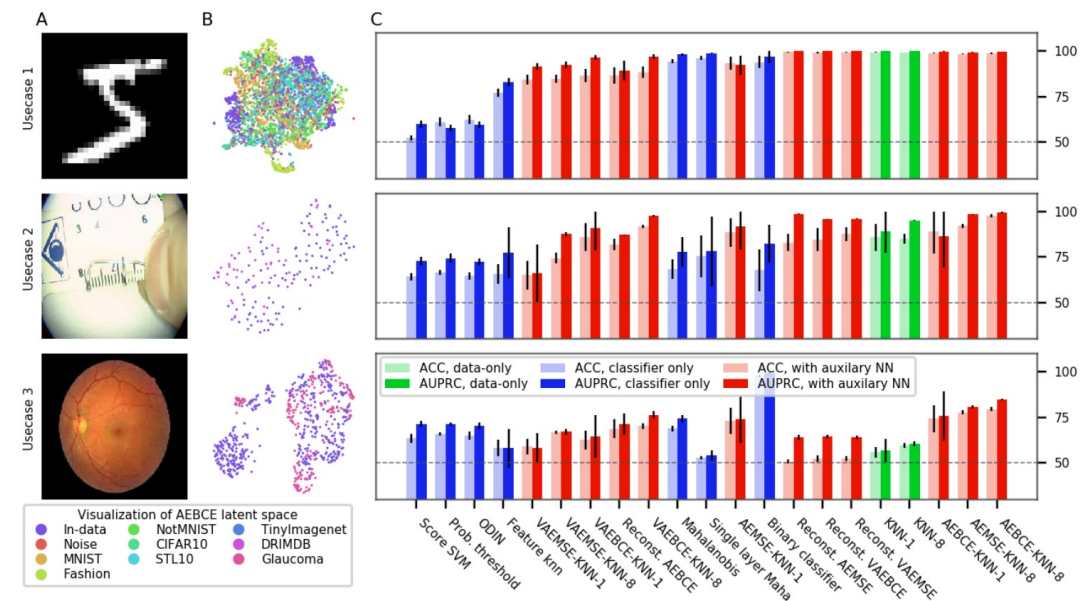
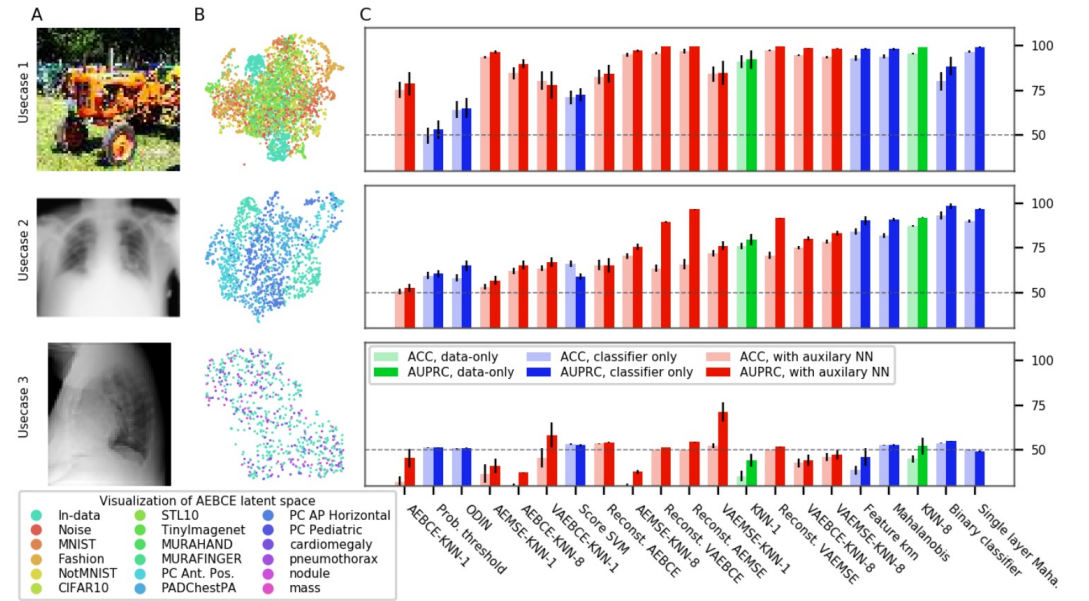
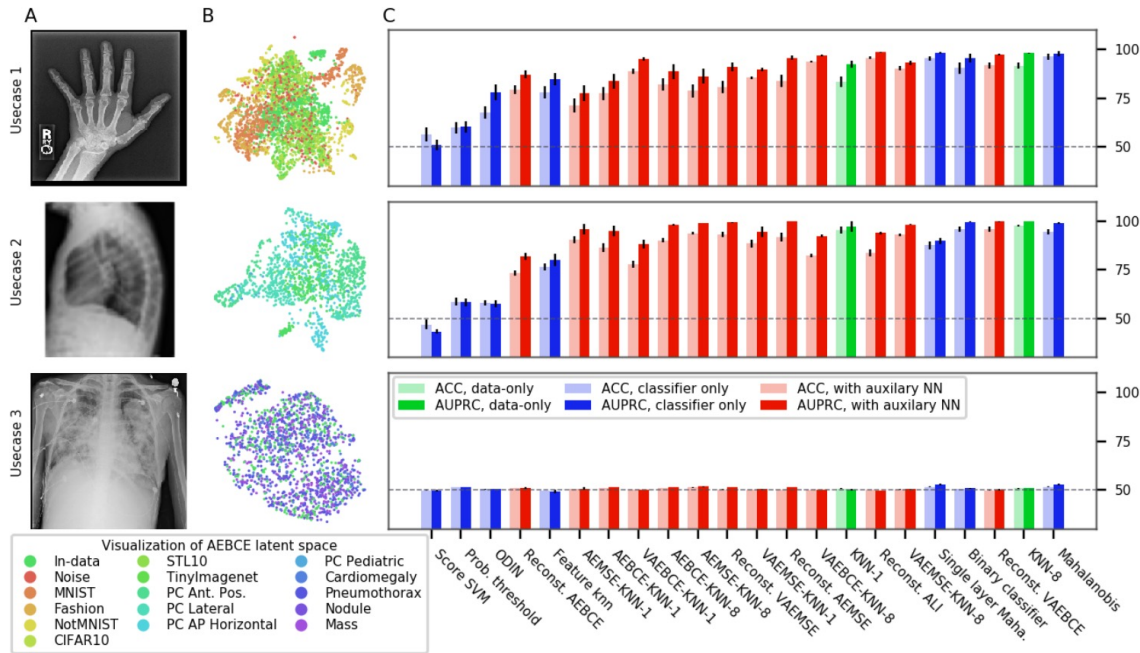
- 1) Fraud Detection
- 2) Intrusion Detection
- 3) Malware Detection
- 4) **Medical Anomaly Detection**

Technique Used	Section	References
AE	Section 11.8	Wang et al. [2016], Cowton et al. [2018], Sato et al. [2018]
DBN	Section 11.1	Turner et al. [2014], Sharma et al. [2016], Wulsin et al. [2010], Ma et al. [2018], Zhang et al. [2016], Wulsin et al. [2011], Wu et al. [2015a]
RBM	Section 11.1	Liao et al. [2016]
VAE	Section 11.5	Xu et al. [2018], Lu and Xu [2018]
GAN	Section 11.5	Ghasedi Dizaji et al. [2018], Chen and Konukoglu [2018]
LSTM ,RNN,GRU	Section 11.7	Yang and Gao [2018], Jagannatha and Yu [2016], Cowton et al. [2018], O'Shea et al. [2016], Latif et al. [2018], Zhang and Zou [2018], Chauhan and Vig [2015], Gugulothu et al., Amarasinghe et al. [2018b]
CNN	Section 11.6	Schmidt-Erfurth et al. [2018], Esteva et al. [2017], Wang et al. [2016], Iakovidis et al. [2018]
Hybrid(AE+ KNN)	Section 11.6	Song et al. [2017]

Examples of DAD techniques Used for medical anomaly detection

➤ Medical Anomaly Detection

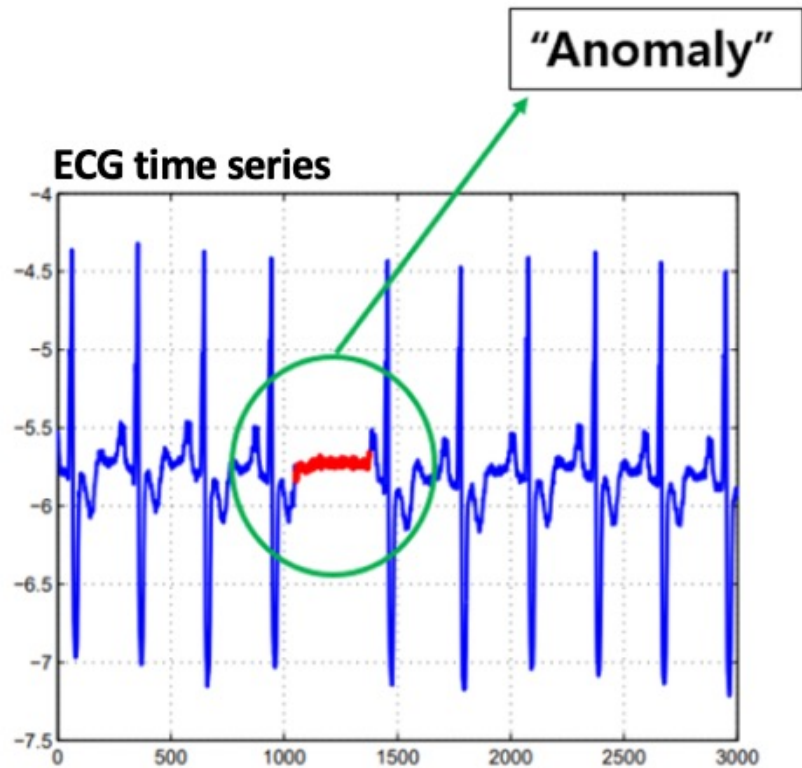
- 1) Cardiac Imaging
- 2) Gastrointestinal (GI) Diseases Detection
- 3) Tumor Detection



➤ Supervised DAD models

Challenges :

- 1) hard to obtain their labels (Rare).
- 2) Data with error labels (Sometimes).
- 3) may change over time.



ImageNet given label:
siamang

We guessed: **baboon**

MTurk consensus: **baboon**



ImageNet given label:
red panda

We guessed: **meerkat**

MTurk consensus: **meerkat**



MNIST given label:
8

We guessed: **9**

MTurk consensus: **9**



MNIST given label:
0

We guessed: **6**

MTurk consensus: **6**

<https://labelerrors.com/>

[Cheboli, 2010; Yarlagadda et al., 2018]

➤ Unsupervised DAD models

- Principal component analysis (PCA)
- Support vector machine (SVM)
- Isolation Forest techniques
- **Autoencoders**

These models assume a high prevalence of normal instances than abnormal data instances, which would result in high false positive rate.

Table 22: Examples of Un-supervised DAD techniques .

CNN: Convolution Neural Networks, LSTM : Long Short Term Memory Networks
 DNN : Deep Neural Networks., GAN: Generative Adversarial Network
 AE: Autoencoders, DAE: Denoising Autoencoders, SVM: Support Vector Machines
 STN: Spatial Transformer Networks, RNN : Recurrent Neural Networks
 AAE: Adversarial Autoencoders, VAE : Variational Autoencoders.

Techniques	Section	References
LSTM	Section 11.7	Singh [2017], Chandola et al. [2008], Dasigi and Hovy [2014], Malhotra et al. [2015]
AE	Section 11.8	Abati et al. [2018], Zong et al. [2018], Tagawa et al. [2015], Dau et al. [2014], Sakurada and Yairi [2014], Wu et al. [2015a], Xu et al. [2015], Hawkins et al. [2002], Zhao et al. [2015], Qi et al. [2014], Chalapathy et al. [2017], Yang et al. [2015], Zhai et al. [2016], Lyudchik [2016], Lu et al. [2017], Mehrotra et al. [2017], Meng et al. [2018], Parchami et al. [2017]
STN	Section 11.2	Chianucci and Savakis [2016]
GAN	Section 11.5	Lawson et al. [2017]
RNN	Section 11.7	Dasigi and Hovy [2014], Filonov et al. [2017]
AAE	Section 11.5	Dimokranitou [2017], Leveau and Joly [2017]
VAE	Section 11.5	An and Cho [2015], Suh et al. [2016], Sölch et al. [2016], Xu et al. [2018], Mishra et al. [2017]

➤ Autoencoder

An autoencoder is a neural network that is trained by unsupervised learning. It is trained to learn reconstructions that are as close as possible to the original input.

An autoencoder is composed of two parts, an **encoder** and a **decoder**.

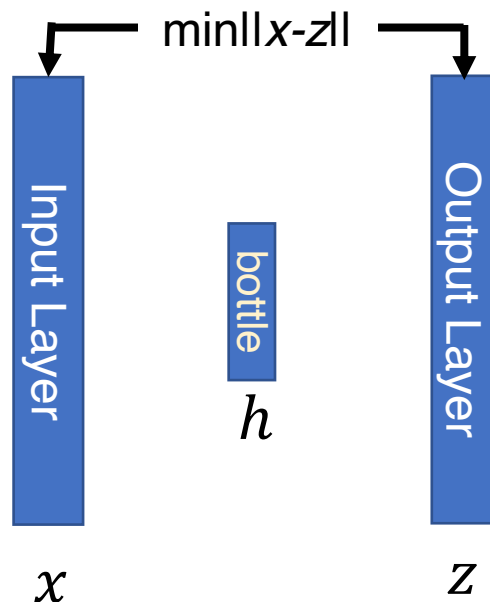
An autoencoder with a **single hidden layer** has an encoder and decoder as in eq(1) and eq(2).

$$h = \sigma(W_{xh}x + b_{xh}) \quad (1)$$

$$z = \sigma(W_{hx}h + b_{hx}) \quad (2)$$

W and b is the weight and bias;
 σ is the nonlinear transformation function.

$$\|x - z\| \quad (3)$$



Algorithm 1 Autoencoder training algorithm

INPUT: Dataset $x^{(1)}, \dots, x^{(N)}$

OUTPUT: encoder f_ϕ , decoder g_θ

$\phi, \theta \leftarrow$ Initialize parameters

repeat

$E = \sum_{i=1}^N \|x^{(i)} - g_\theta(f_\phi(x^{(i)}))\|$ Calculate sum of reconstruction error

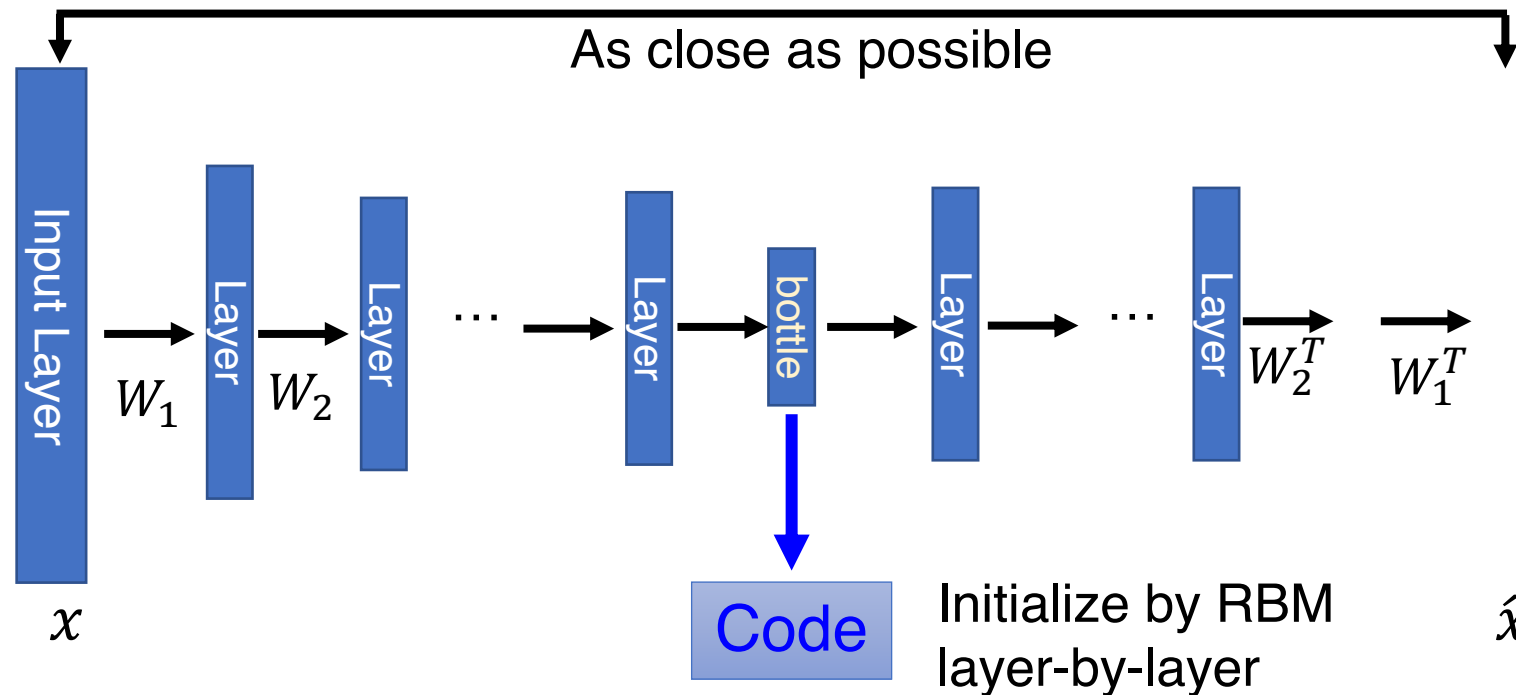
$\phi, \theta \leftarrow$ Update parameters using gradients of E (e.g. Stochastic Gradient Descent)

until convergence of parameters ϕ, θ

➤ Deep autoencoder

- The auto-encoder can be deep.

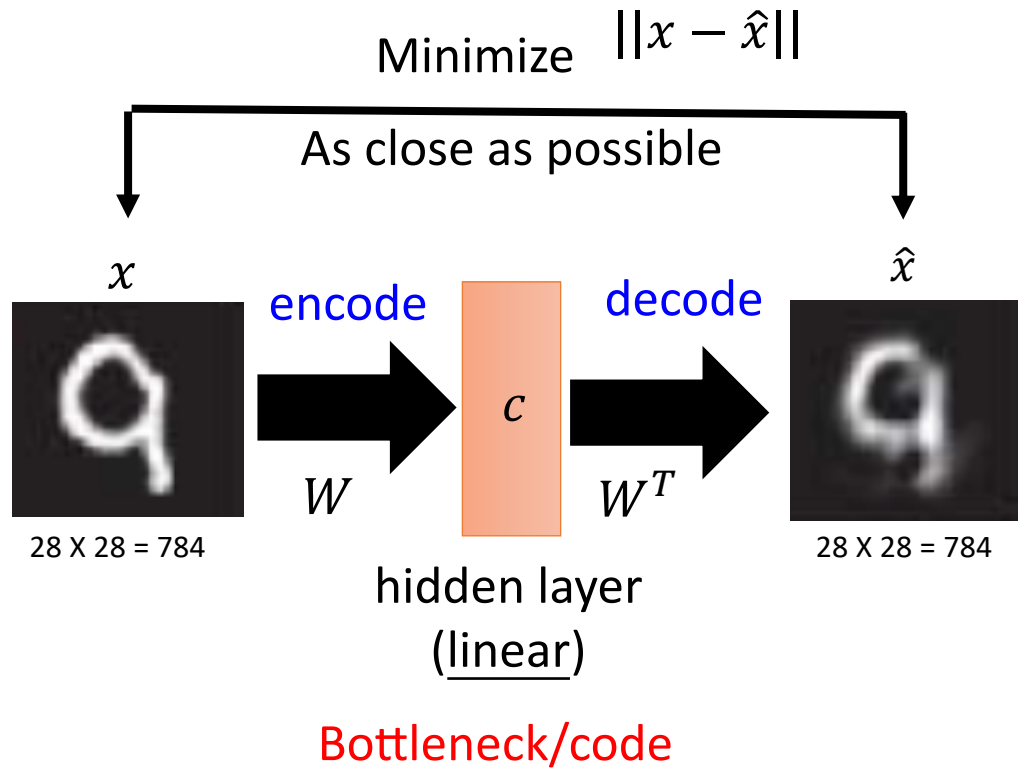
The symmetric weights are not necessary.



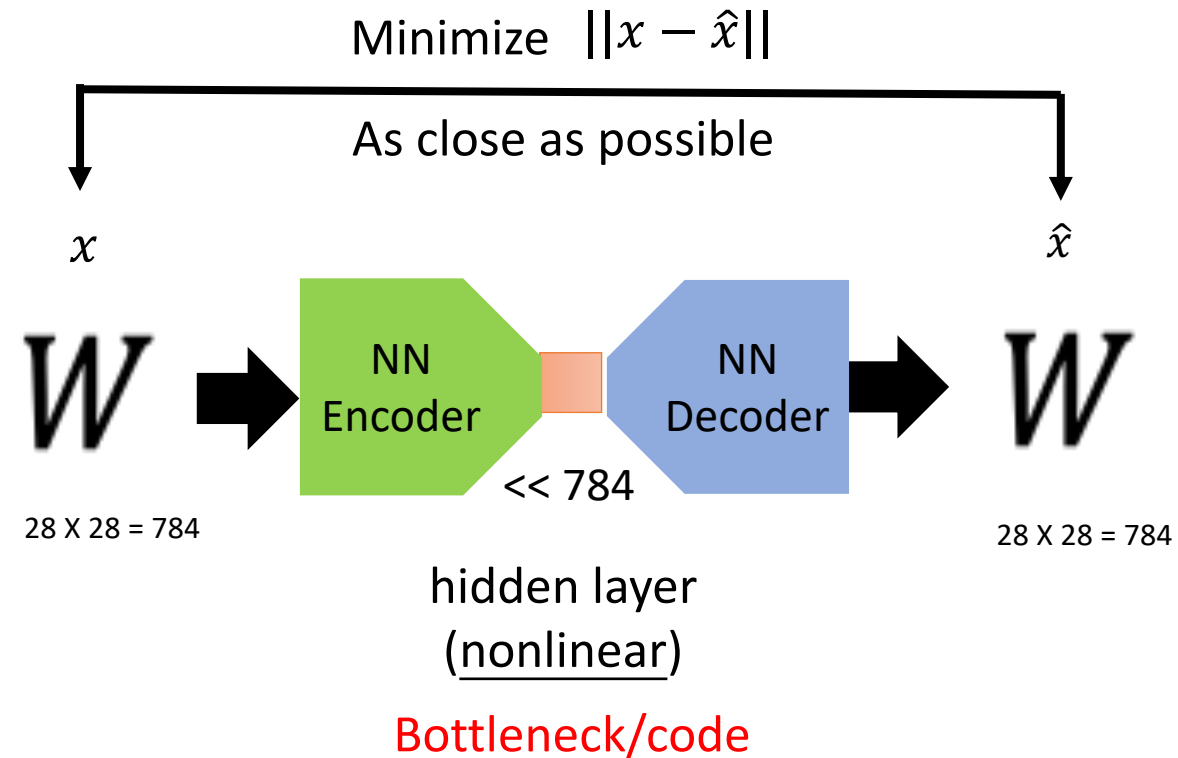
Embedding, Latent Representation, Latent Code

➤ Principle Component Analysis (PCA)

vs. **Autoencoder**



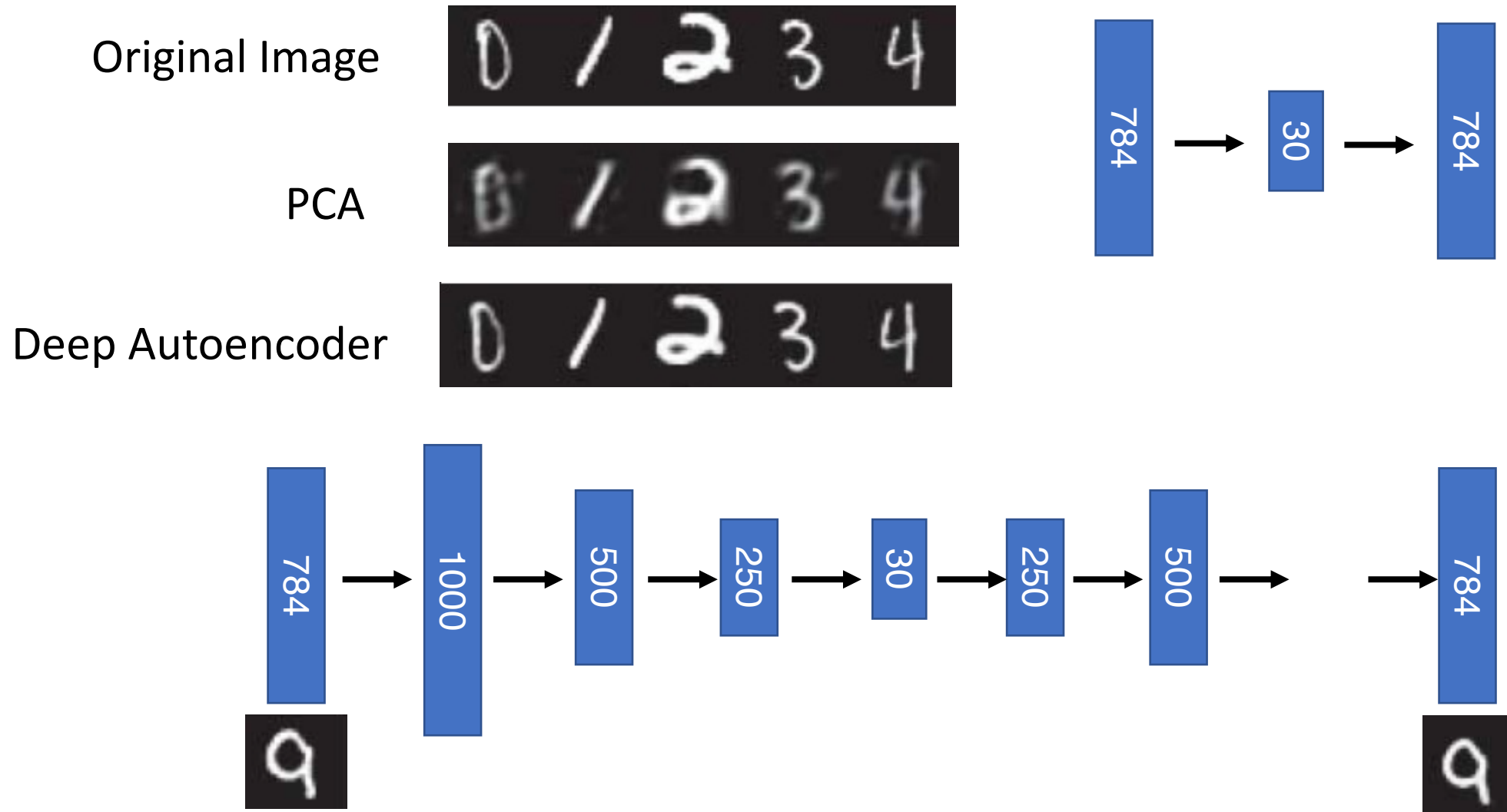
Projection is linear, and W is orthonormal.



The encoder and decoder learn together.

Code is a compact representation of the input object.

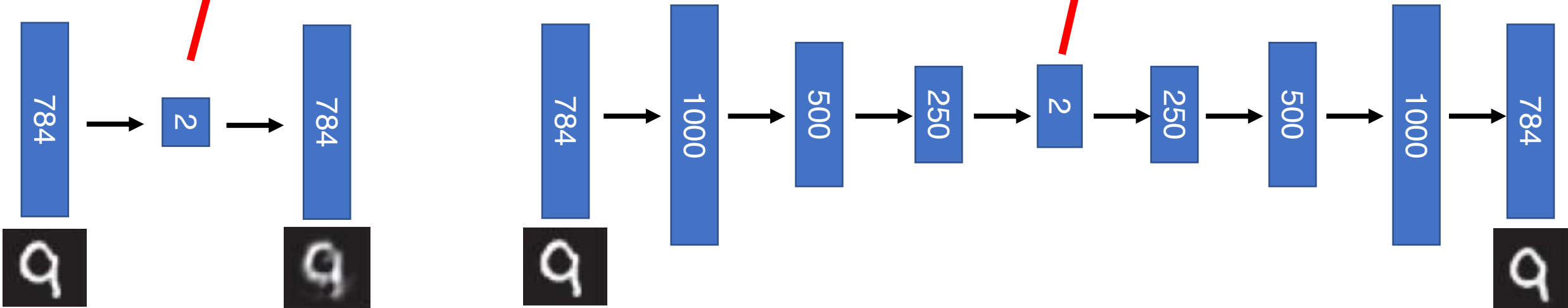
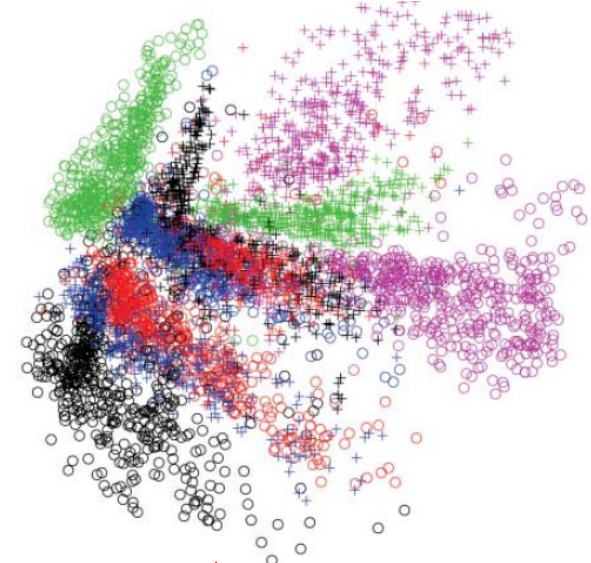
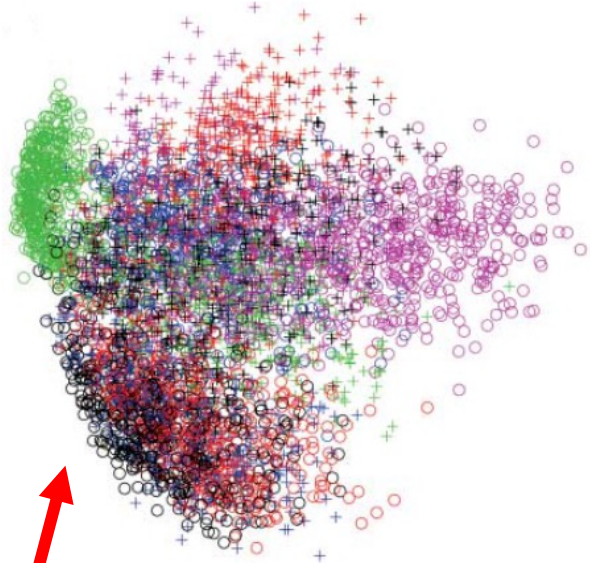
➤ Performance comparison between PCA and Deep Autoencoder



➤ Latent representation in PCA

vs.

Deep Autoencoder



➤ Autoencoder-based anomaly detection

Autoencoder-based DAD is a **deviation-based anomaly detection method** using semi-supervised learning.

It uses the **reconstruction error** as the anomaly **score**. Data points with **high** reconstruction error are **anomalies**.

Only data with **normal** instances are used to train the autoencoder.

After training, the autoencoder will reconstruct normal data very well, while **failing** to reconstruct **anomaly data** which the autoencoder has not encountered.

Algorithm 2 shows the anomaly detection algorithm using **reconstruction errors** of autoencoders.

Algorithm 2 Autoencoder based anomaly detection algorithm

INPUT: Normal dataset X , Anomalous dataset $x^{(i)} \quad i = 1, \dots, N$, threshold α

OUTPUT: reconstruction error $\|x - \hat{x}\|$

$\phi, \theta \leftarrow$ train a autoencoder using the normal dataset X

for $i=1$ **to** N **do**

reconstruction error(i) = $\|x^{(i)} - g_{\theta}(f_{\phi}(x^{(i)}))\|$

if *reconstruction error*(i) > α **then**

$x^{(i)}$ is an anomaly

else

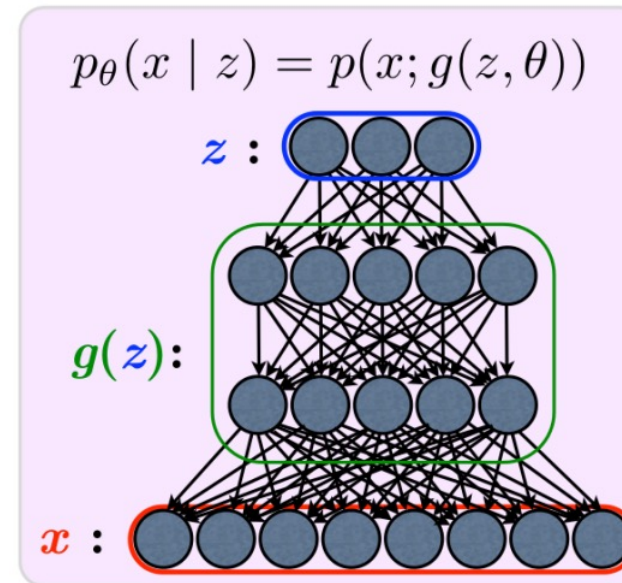
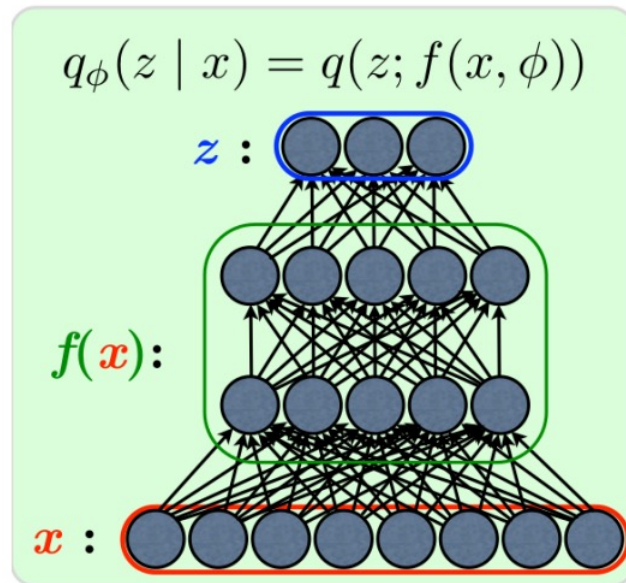
$x^{(i)}$ is not an anomaly

end if

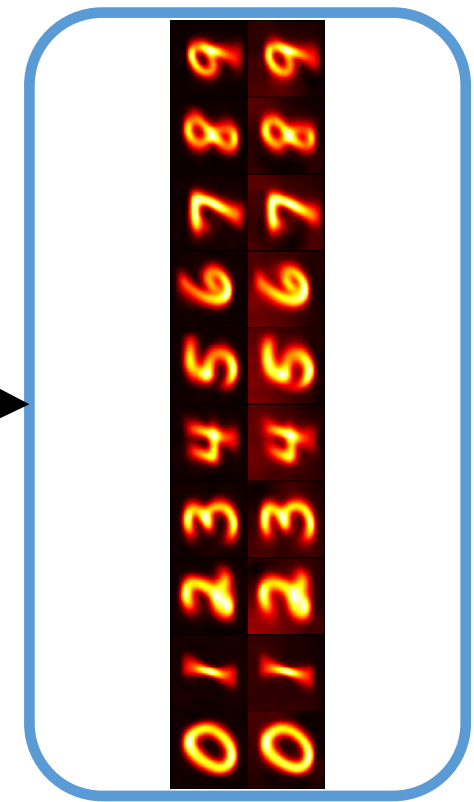
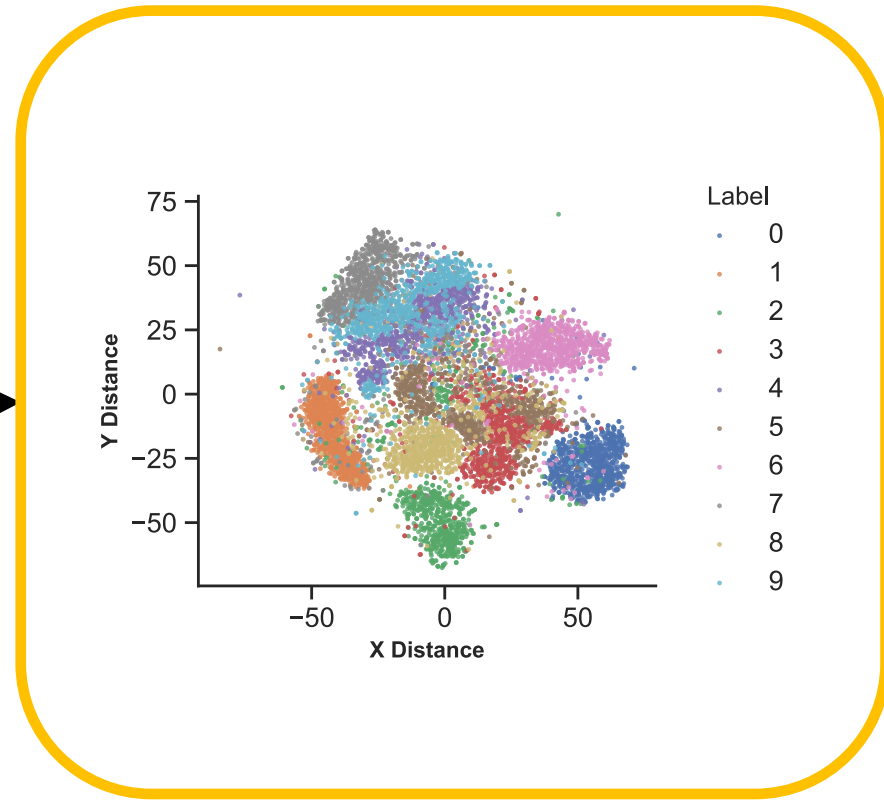
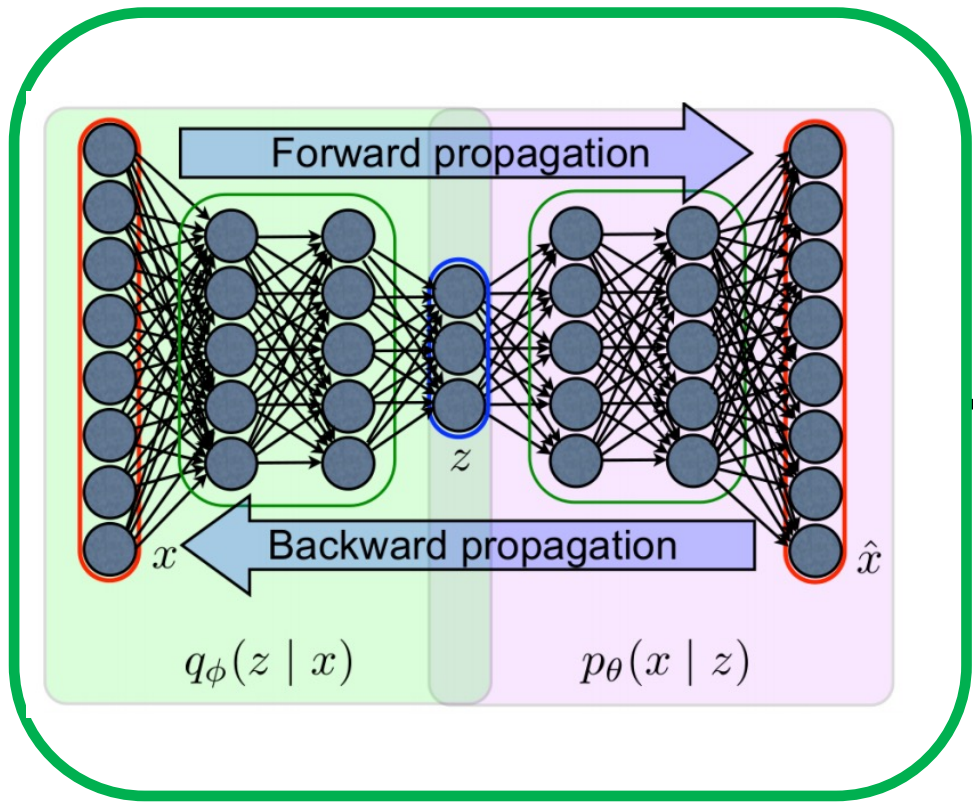
end for

➤ Variational Autoencoder (VAE)

- A variational autoencoder (VAE) is a directed probabilistic graphical model (DPGM) whose **posterior** is approximated by a neural network, forming an autoencoder-like architecture.
- In the VAE, the highest layer of the directed graphical model z is treated as **the latent variable** where the generative process starts.
- $g(z)$ represents the complex process of data generation that results in the data x , which is modelled in the structure of a neural network.



➤ Autoencoder



➤ Application of Autoencoder

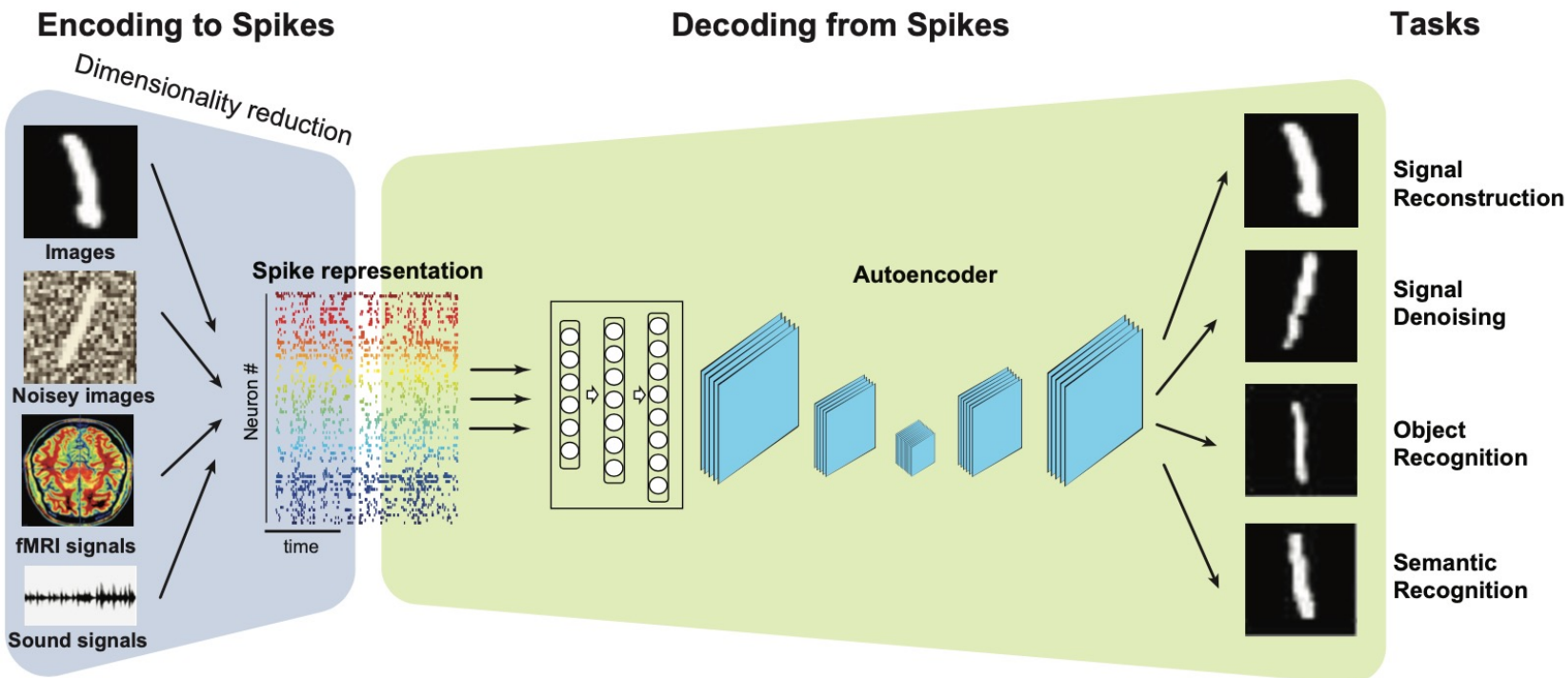


Fig. 1: The schematic diagram of DSPD framework.

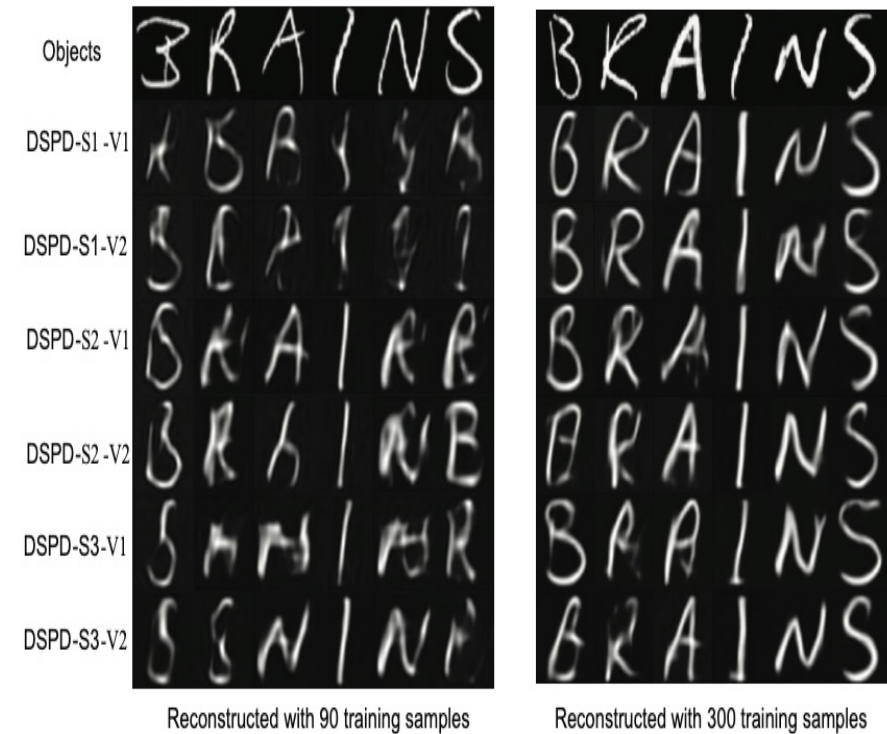
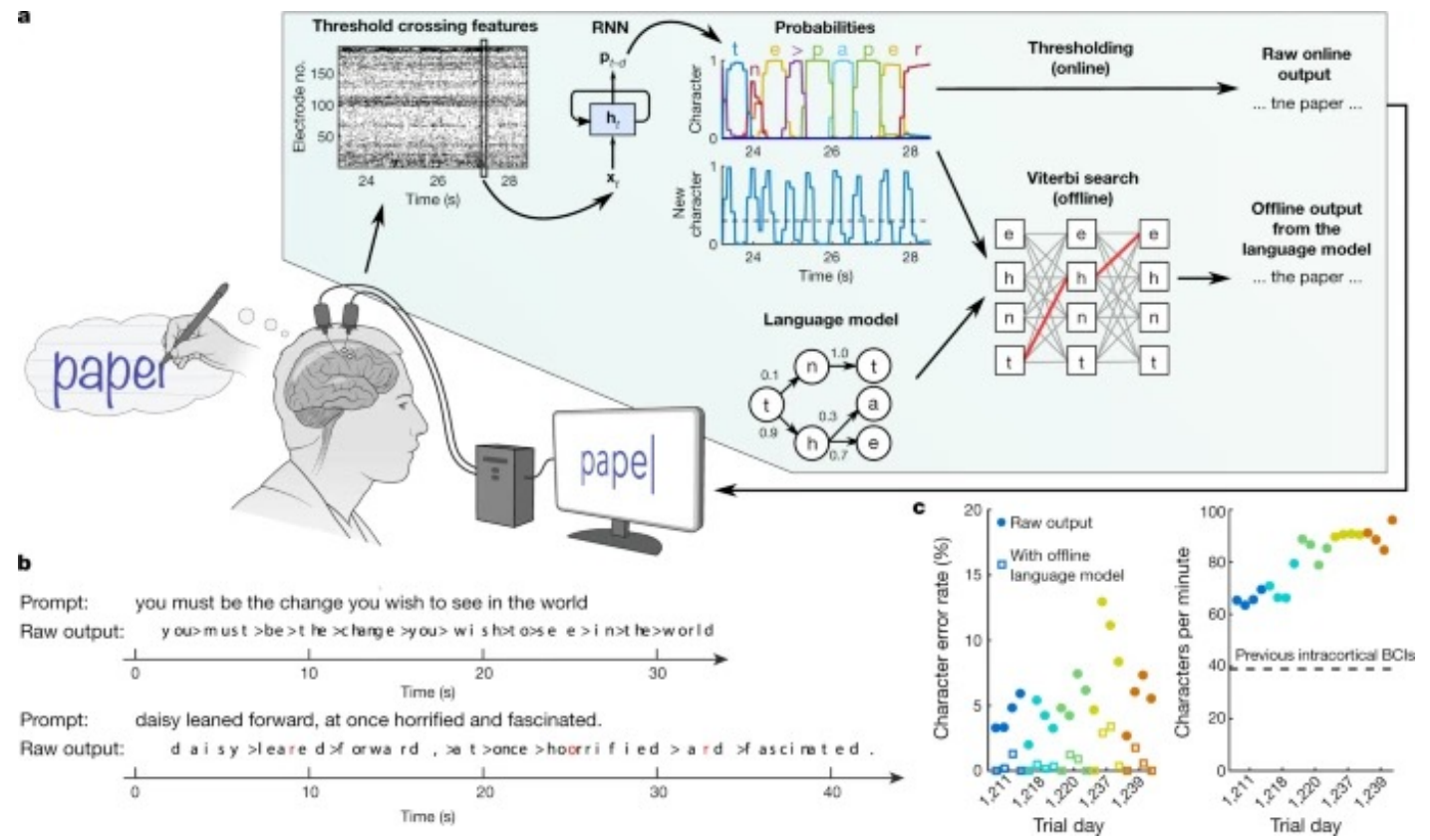
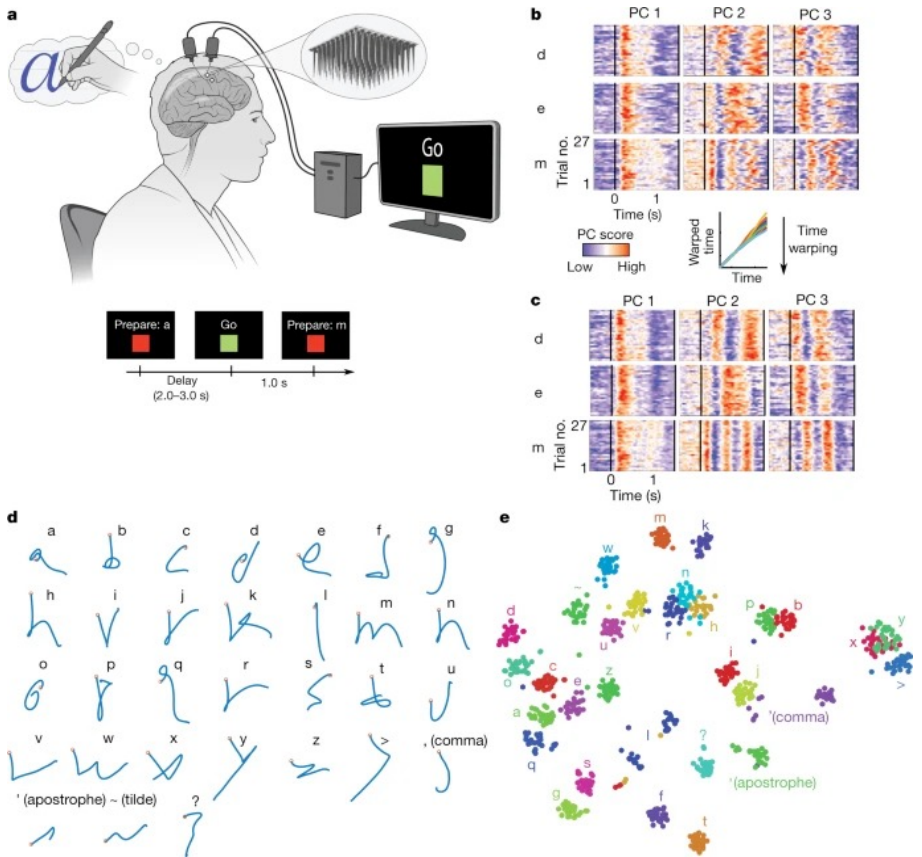


Fig. 5: Presented fMRI characters and Reconstructed Results of DSPD three subjects S_1 , S_2 and S_3 from the V_1 and V_2 areas (the left images are with 90 training samples and the right images are with 300 training samples).

➤ Application of Autoencoder



Nature, High-performance brain-to-text communication via handwriting, 2021

➤ Variational Autoencoder (VAE)

The objective function of a VAE is the variational lower bound of the marginal likelihood of data, since the marginal likelihood is intractable. The marginal likelihood is the sum over the marginal likelihood of individual data points that can be rewritten as follows.

$$\log p_{\theta}(x^{(i)}) = D_{KL}(q_{\phi}(z|x)||p_{\theta}(z)) + \mathcal{L}(\theta, \phi; x^{(i)}) \quad (4)$$

Equation (4) can be rewritten as follows.

$$\log p_{\theta}(x^{(i)}) \geq \mathcal{L}(\theta, \phi; x^{(i)}) \quad (5)$$

$$= E_{q_{\phi}(z|x^{(i)})}[-\log q_{\phi}(z|x) + \log p_{\theta}(x|z)] \quad (6)$$

$$= -D_{KL}(q_{\phi}(z|x^{(i)})||p_{\theta}(z)) + E_{q_{\phi}(z|x^{(i)})}[\log p_{\theta}(x|z)] \quad (7)$$

Algorithm 3 Variational autoencoder training algorithm

INPUT: Dataset $x^{(1)}, \dots, x^{(N)}$

OUTPUT: probabilistic encoder f_{ϕ} , probabilistic decoder g_{θ}

$\phi, \theta \leftarrow$ Initialize parameters

repeat

for $i=1$ **to** N **do**

 Draw L samples from $\epsilon \sim \mathcal{N}(0, 1)$

$z^{(i,l)} = h_{\phi}(\epsilon^{(i)}, x^{(i)}) \quad i = 1, \dots, N$

end for

$E = \sum_{i=1}^N -D_{KL}(q_{\phi}(z|x^{(i)})||p_{\theta}(z)) + \frac{1}{L} \sum_{l=1}^L (\log p_{\theta}(x^{(i)}|z^{(i,l)}))$

$\phi, \theta \leftarrow$ Update parameters using gradients of E (e.g. Stochastic Gradient Descent)

until convergence of parameters ϕ, θ

➤ VAE based anomaly detection

VAE based anomaly detection uses **reconstruction probability** as the **anomaly score**. **Reconstruction probability** is computed as the probability of data from the reconstructed sample distribution. Only data with normal instances are used to train the VAE. After training, the reconstructed distribution covers normal data very well, while anomaly data is not in that distribution.

Algorithm 4 shows the anomaly detection algorithm using **reconstruction probability** of VAE.

Algorithm 4 Variational autoencoder based anomaly detection algorithm

INPUT: Normal dataset X , Anomalous dataset $x^{(i)} \quad i = 1, \dots, N$, threshold α

OUTPUT: reconstruction probability $p_{\theta}(x|\hat{x})$

$\phi, \theta \leftarrow$ train a variational autoencoder using the normal dataset X

for $i=1$ **to** N **do**

$\mu_{z^{(i)}}, \sigma_{z^{(i)}} = f_{\theta}(z|x^{(i)})$

draw L samples from $z \sim \mathcal{N}(\mu_{z^{(i)}}, \sigma_{z^{(i)}})$

for $l=1$ **to** L **do**

$\mu_{\hat{x}^{(i,l)}}, \sigma_{\hat{x}^{(i,l)}} = g_{\phi}(x|z^{(i,l)})$

end for

$reconstruction\ probability(i) = \frac{1}{L} \sum_{l=1}^L p_{\theta}(x^{(i)}|\mu_{\hat{x}^{(i,l)}}, \sigma_{\hat{x}^{(i,l)}})$

if $reconstruction\ probability(i) < \alpha$ **then**

$x^{(i)}$ is an anomaly

else

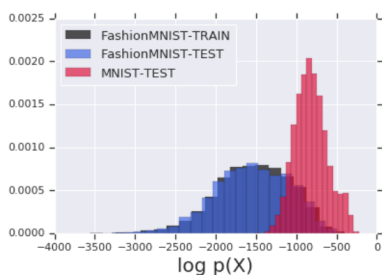
$x^{(i)}$ is not an anomaly

end if

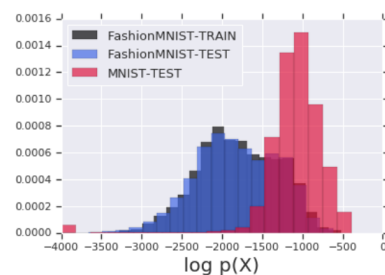
end for

➤ Likelihood-based generative models for OOD Detection

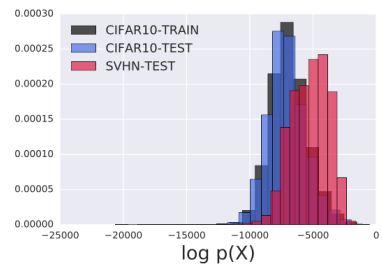
There are widely views that likelihood-based generative models have high robustness to the out-of-distribution (OOD) inputs and a well-calibrated generative models can be as a detector. However, recent works reported a phenomenon that DGM recognizes some OOD samples as ID by assigning a higher likelihood to the OOD inputs compared to the one from ID.



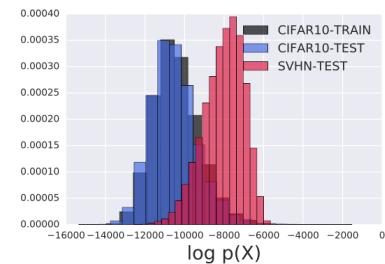
(a) PixelCNN: FashionMNIST vs MNIST



(b) VAE: FashionMNIST vs MNIST

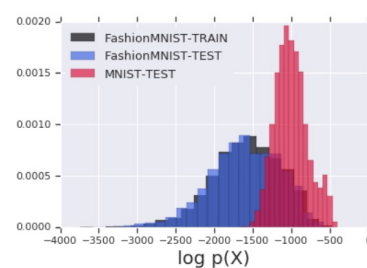


(c) PixelCNN: CIFAR-10 vs SVHN

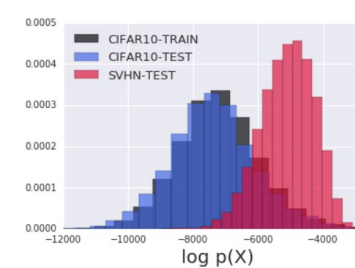


(d) VAE: CIFAR-10 vs SVHN

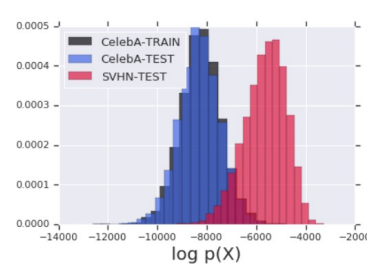
Figure 3: PixelCNN and VAE. Log-likelihoods calculated by PixelCNN (a, c) and VAE (b, d) on FashionMNIST vs MNIST (a, b) and CIFAR-10 vs SVHN (c, d). VAE models are the convolutional categorical variant described by Rosca et al. (2018).



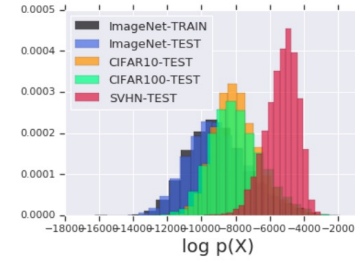
(a) Train on FashionMNIST, Test on MNIST



(b) Train on CIFAR-10, Test on SVHN



(c) Train on CelebA, Test on SVHN

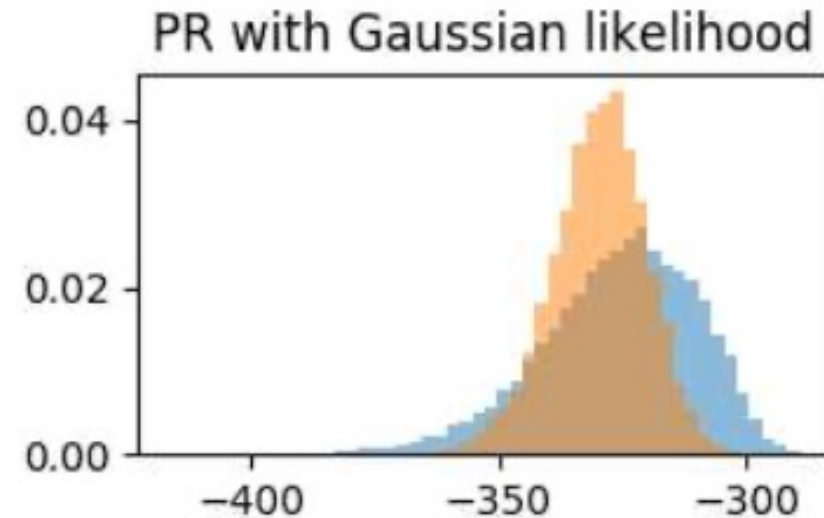
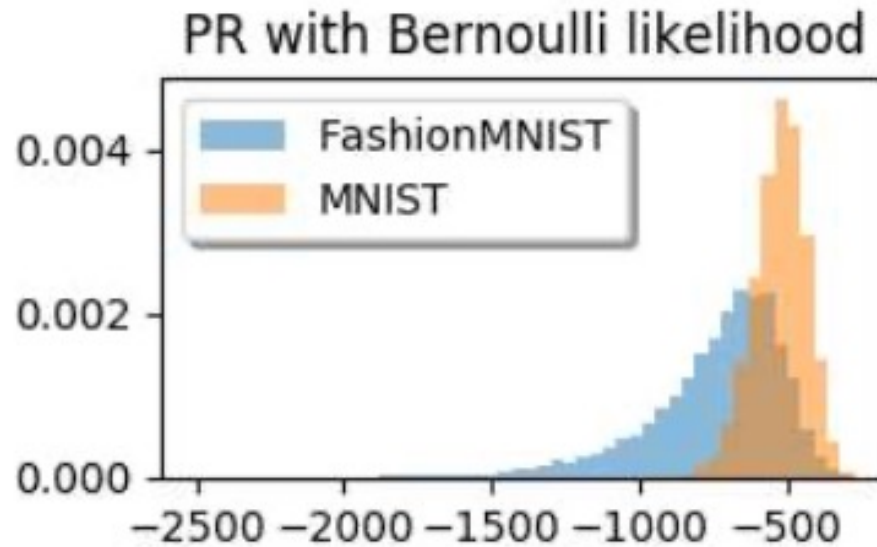


(d) Train on ImageNet, Test on CIFAR-10 / CIFAR-100 / SVHN

Figure 2: Histogram of Glow log-likelihoods for FashionMNIST vs MNIST (a), CIFAR-10 vs SVHN (b), CelebA vs SVHN (c), and ImageNet vs CIFAR-10 / CIFAR-100 / SVHN (d).

➤ Likelihood-based generative models for OOD Detection-Why ?

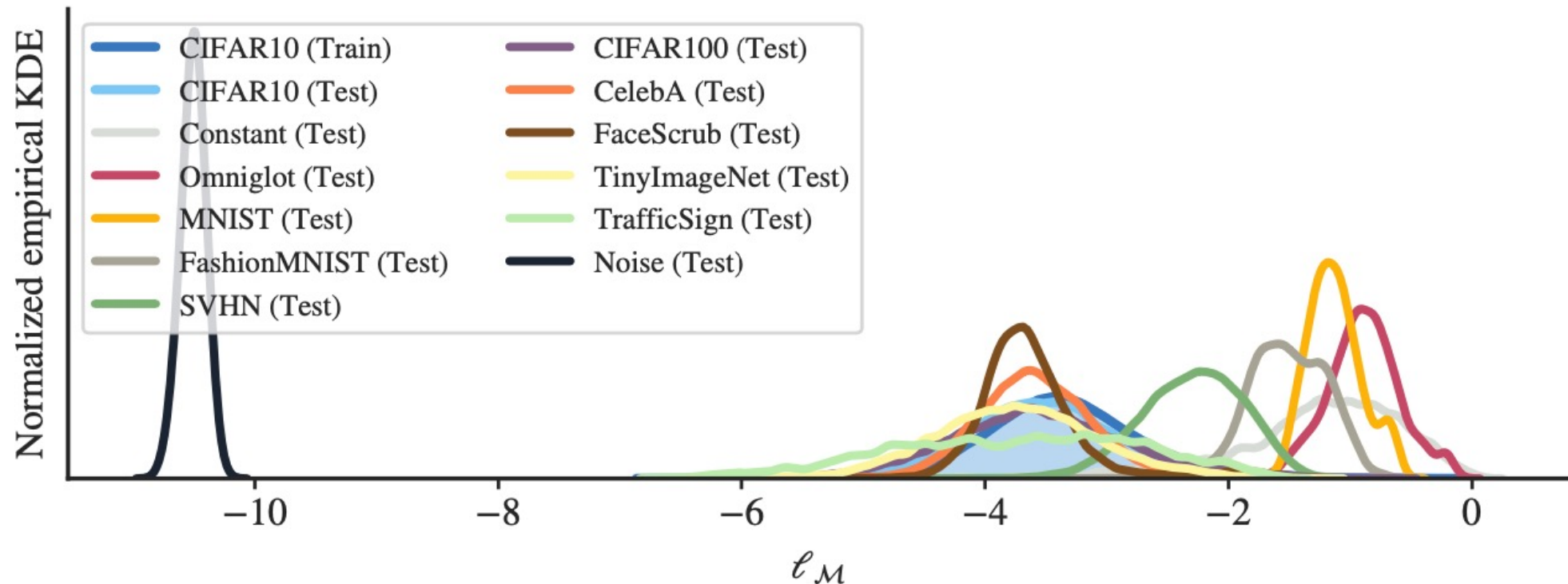
Bü'tepage *et al.* conjectured that this phenomenon is caused by modeling assumptions and evaluation schemes. The modeling assumption on the likelihood function (e.g., iid Bernoulli, iid Gaussian) can influence the judgment of the model and the local evaluation under the approximated posterior leads to high confidence in some datasets.



The log likelihood under the prior using model M_1 with an iid Bernoulli likelihood function ((left)) and model M_2 with an iid Gaussian likelihood function (right)

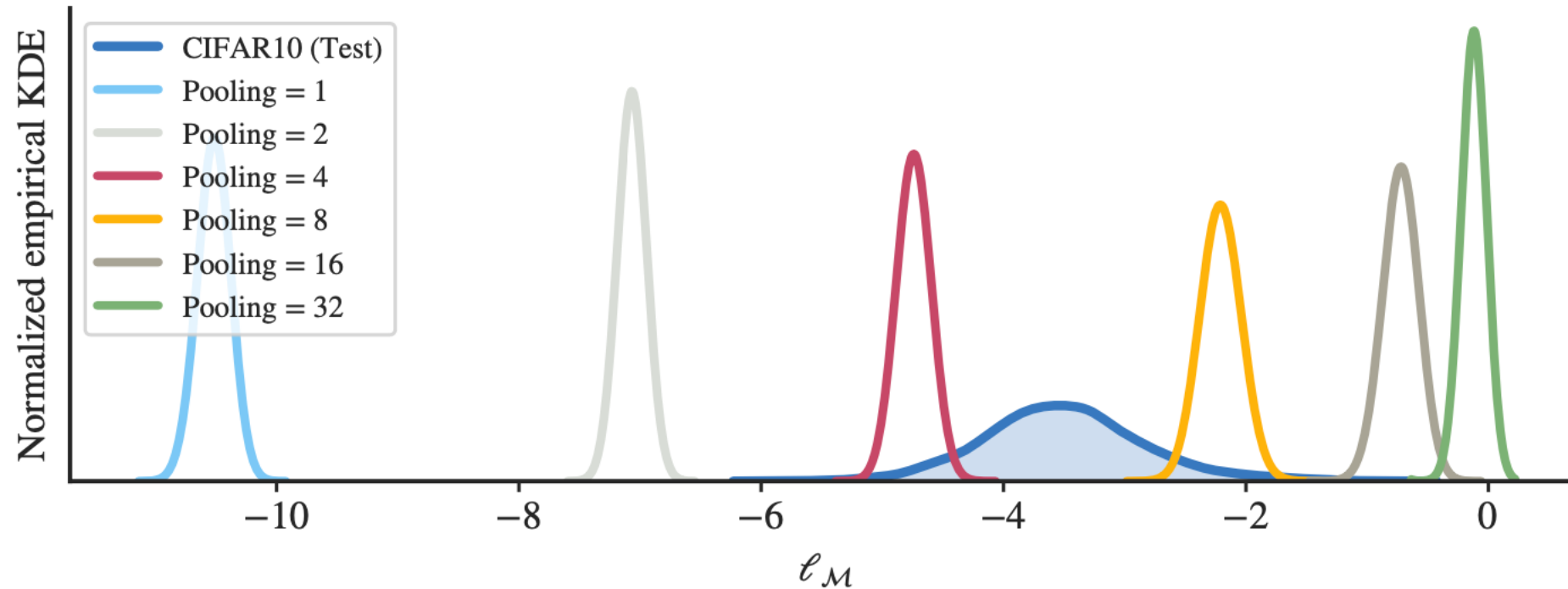
➤ Likelihood-based generative models for OOD Detection-Why ?

Serra` et al pose that this problem is due to the excessive influence that input complexity has in generative models' likelihoods.



Log-likelihoods from a Glow model trained on CIFAR10. Qualitatively similar results are obtained for a PixelCNN++ model and when training with FashionMNIST

➤ Likelihood-based generative models for OOD Detection-Why ?



Pooled-image log-likelihoods obtained from a Glow model trained on CIFAR10. Qualitatively similar results are obtained for a PixelCNN++ model.

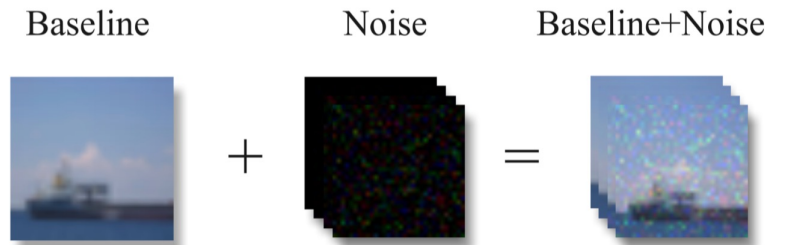
➤ INCPVAE- Improved Noise Contrastive Priors

Hafner et al. proposed NCPs, as a kind of data priors that are applied to both ID inputs \mathbf{x} and OOD inputs $\tilde{\mathbf{x}}$. The OOD inputs are usually generated by imposing noise. For it is hard to exactly generate OOD data, we add Gaussian noise to ID image to realize OOD data generation.

- 1) Generating OOD Inputs
- 2) Data Priors
- 3) Loss Function

Generating OOD Inputs Lee et al. reported that OOD samples are produced by sampling from the boundary of the ID with high uncertainty. advanced an algorithm inspired by noise contrastive estimation where a complement distribution is approximated using random noise. For continuous ID inputs \mathbf{x} , we add Gaussian noise to obtain OOD inputs, which is $\tilde{\mathbf{x}} = \mathbf{x} + \epsilon$. The distribution density of OOD inputs $p_o(\tilde{\mathbf{x}})$ is formulated as,

$$p_o(\tilde{\mathbf{x}}) = \int_{\mathbf{x}} p_i(\mathbf{x}) \mathcal{N}(\tilde{\mathbf{x}} - \mathbf{x} \mid \mu, \sigma^2 \mathbf{I}) d\mathbf{x},$$



where $p_i(\mathbf{x})$ is the distribution density of ID inputs, μ and σ_2 are the mean and variance of Gaussian distribution of noise. In order to make noise contrastive prior equal in all directions of data manifold, we set $\mu = 0$. The variance σ_2 is a hyper-parameter to tune the sampling distance from the boundary of training distribution. The complexity of OOD inputs is correlated with the variance.

[Gutmann and Hyvärinen, 2010; Mnih and Kavukcuoglu, 2013; Hafner et al., 2018; Lee et al., 2018a]

➤ INCPVAE- Improved Noise Contrastive Priors

Data Priors The data priors consist of inputs prior $p(\mathbf{x})$ and outputs prior $p(\mathbf{z}|\mathbf{x})$. To obtain a reliable VAE's uncertainty estimation, an appropriate inputs prior should include OOD inputs so that it can obtain better performance than the baseline under training distribution. A good output prior should be a high-entropy distribution that serves as high uncertainty about VAE's target outputs given OOD inputs. The data priors are listed as follows:

$$p(\tilde{\mathbf{x}}) = p_o(\tilde{\mathbf{x}})$$
$$p(\tilde{\mathbf{z}} | \tilde{\mathbf{x}}) = \mathcal{N}(\tilde{\mathbf{z}} | \mu_{\tilde{\mathbf{x}}}, \sigma_{\tilde{\mathbf{x}}}^2 \mathbf{I}),$$

where $p_o(\tilde{\mathbf{x}})$ is the distribution of OOD inputs, $\mu_{\tilde{\mathbf{x}}}$ and $\sigma_{\tilde{\mathbf{x}}}$ are the parameter of OOD data outputs priors, $\sigma_{\tilde{\mathbf{x}}}$ is a hyper-parameter tuning the level of target outputs uncertainty.

Loss Function Improved Noise Contrastive Priors (INCPs) have the merit of estimating the model's uncertainty which is easily generalized to OOD samples. To train INCPs, we modified the loss function as follows:

$$\mathcal{L}(\theta) = \mathbf{E}_{q_{\theta}(\mathbf{z}|\mathbf{x})} [\mathbf{D}_{KL} [q_{\theta}(\mathbf{z} | \mathbf{x}) || p(\mathbf{z} | \mathbf{x})]]$$
$$+ \gamma \mathbf{E}_{q_{\theta}(\tilde{\mathbf{z}}|\tilde{\mathbf{x}})} [\mathbf{D}_{KL} [q_{\theta}(\tilde{\mathbf{z}} | \tilde{\mathbf{x}}) || p(\tilde{\mathbf{z}} | \tilde{\mathbf{x}})]],$$

where $p(\tilde{\mathbf{z}} | \tilde{\mathbf{x}})$ denote OOD data priors, θ is the parameter of neural network. The hyper-parameter γ represents the trade-off between them. INCPs can be trained by minimizing this loss.

➤ Improved noise contrastive prior variational autoencoder (INCPVAE)

INCPVAE consists of an encoder and a decoder.

The improved NCPs are imposed on the encoder network of VAE.

INCPVAE is trained on both in-distribution (ID) and OOD inputs by minimizing I-ELBO and O-ELBO.

We have all the evidence lower bound (**ELBO**) of INCPVAE as follows:

$$\mathcal{L}_I(\phi, \theta) = \mathbf{E}_{z \sim q_\theta(z|\mathbf{x})} [\log p_\phi(\mathbf{x} | z)] - \mathbf{D}_{KL}[q_\theta(z | \mathbf{x}) || p(z)]$$

$$\mathcal{L}_O(\phi, \theta) = \mathbf{E}_{\tilde{z} \sim q_\theta(\tilde{z}|\tilde{\mathbf{x}})} [\log p_\phi(\tilde{\mathbf{x}} | \tilde{z})] - \mathbf{D}_{KL}[q_\theta(\tilde{z} | \tilde{\mathbf{x}}) || p(\tilde{z})]$$

$$\mathcal{L}_{INCP}(\phi, \theta) = \mathcal{L}_I(\phi, \theta) + \mathcal{L}_O(\phi, \theta)$$

Maximizing the ELBO of INCPVAE can be replaced by **minimizing** the following loss function:

$$\mathcal{L}_{INCPVAE}(\phi, \theta) = -\mathcal{L}_I(\phi, \theta) + \gamma \frac{\mathbf{D}_{KL}[q_\theta(\tilde{z}|\tilde{\mathbf{x}}) || p(\tilde{z}|\tilde{\mathbf{x}})]}{\text{INCP-KL Loss}}$$

➤ Uncertainty estimation in INCPVAE

We proposed the objective variational ELBO Ratios for quantitative evaluation of VAE.

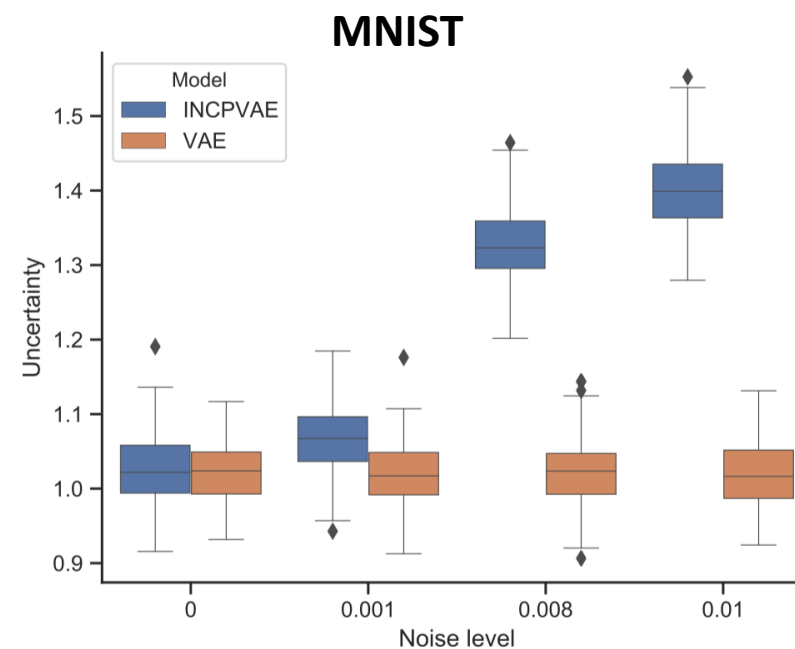
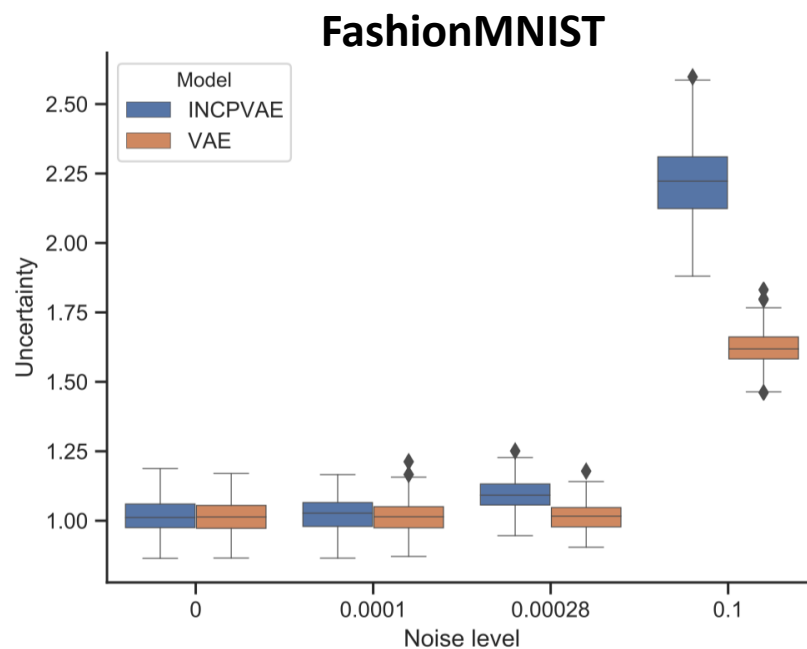
We tested all the ID samples of ELBO (I-ELBO) and get the maximum one ($\mathbf{I-ELBO}(x_{max})$).

ELBO Ratio that is defined as

$$\mathcal{U}(x_0) = \frac{\mathbf{ELBO}(x_0)}{\mathbf{I-ELBO}(x_{max})}$$

The greater scalar $\mathcal{U}(x_0)$ is, the higher uncertainty x_0 acquires.

INCPVAE provides higher uncertainty for the OOD data.



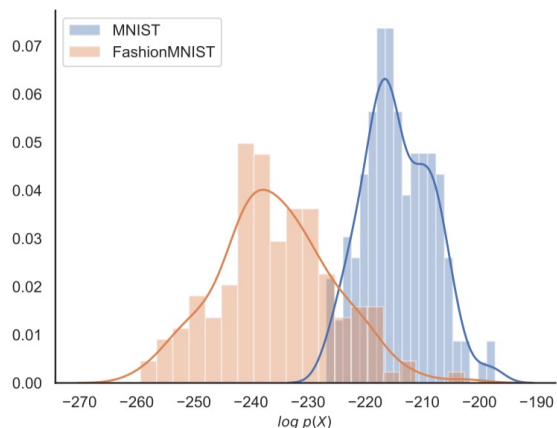
➤ OOD detection based on INCPVAE

The density estimation of VAE are always used for OOD detection, but the OOD inputs get a higher likelihoods than ID inputs that occur some datasets. To solve this problem, Ren et al. (2019) proposed Likelihood Ratios for OOD detection.

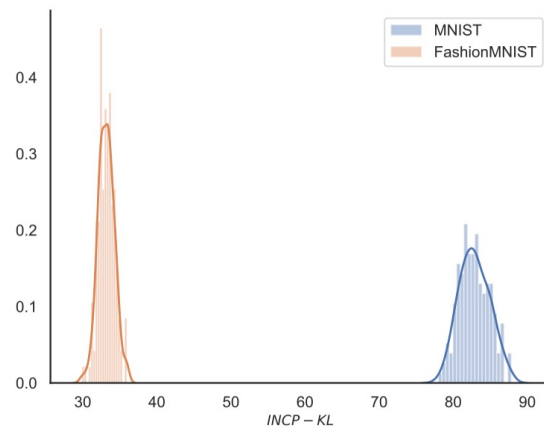
INCP-KL Ratios for OOD detection. We test all the OOD samples of INCP-KL and get the maximum one (called $D_{KL}(OOD_{max})$). INCP-KL Ratio that is defined as

$$\begin{aligned} \mathcal{KL}\mathcal{R}(\mathbf{x}_0) &= \frac{D_{KL}[q_{\theta}(\mathbf{z}_0|\mathbf{x}_0)||p(\tilde{\mathbf{z}}|\tilde{\mathbf{x}})]}{D_{KL}(OOD_{max})} \\ \mathbf{Label}(\mathbf{x}_0) &= \begin{cases} 0 & \mathcal{KL}\mathcal{R}(\mathbf{x}_0) > 1 \\ 1 & \mathcal{KL}\mathcal{R}(\mathbf{x}_0) \leq 1, \end{cases} \end{aligned} \quad (11)$$

where $\mathbf{Label}(\mathbf{x}_0) = 1$, the test sample \mathbf{x}_0 is OOD data; $\mathbf{Label}(\mathbf{x}_0) = 0$, \mathbf{x}_0 is not OOD data (\mathbf{x}_0 does not belong to OOD data).



(a) VAE



(b) INCPVAE

The likelihood distributions of VAE for the ID and OOD inputs have considerable overlaps, whereas the INCP-KL of INCPVAE largely separate ID and OOD inputs.

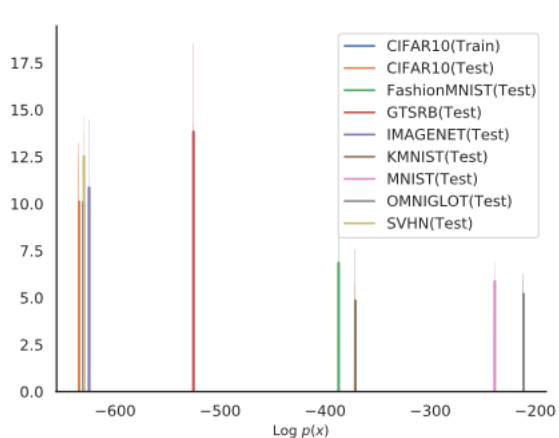
➤ OOD detection based on INCPVAE

Table 1: AUROC and AUPRC for detecting OOD inputs using our INCP-KL Ratio method, likelihood method and other baseline methods on FashionMNIST vs. MNIST datasets.

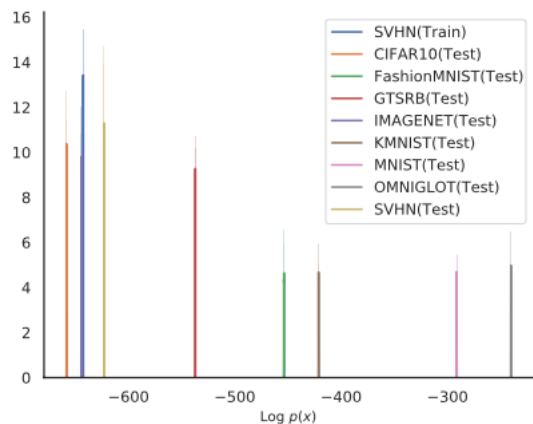
Model	AUROC	AUPRC
NCP-KL Ratio(Baseline+Noise)	1.000	1.000
NCP-KL Ratio(Baseline)	1.000	1.000
Likelihood	0.035	0.313
Likelihood Ratio(μ) Ren et al. (2019)	0.973	0.951
Likelihood Ratio(μ, λ) Ren et al. (2019)	0.994	0.993
ODIN Liang et al. (2018)	0.752	0.763
Mahalanobis distance Lee et al. (2018b)	0.942	0.928
Ensemble, 20 classifiers Lakshminarayanan et al. (2017)	0.857	0.849
WAIC,5 models Choi et al. (2018)	0.221	0.401

In our INCPVAE model, OOD samples are generated by adding gaussian noise, endowing VAE with reliable uncertainty estimation for inputs and the ability of distinguishing OOD data. Using INCP-KL ratios our model achieves SOTA performance to differentiate OOD and ID data, compared with baseline methods.

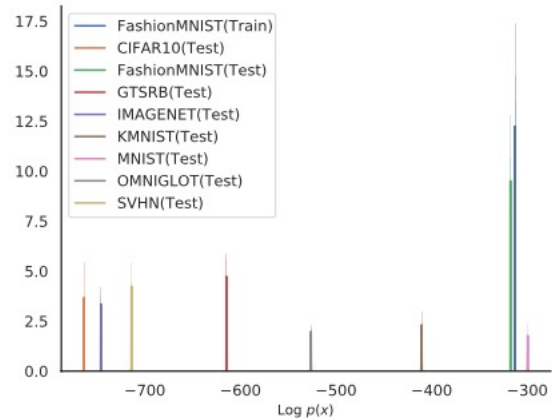
➤ Does input complexity has a strong effect in VAE?



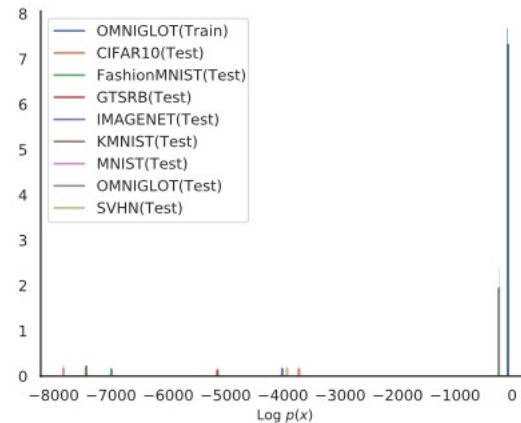
(a) Trained on CIFAR10



(b) Trained on SVHN



(c) Trained on FashionMNIST



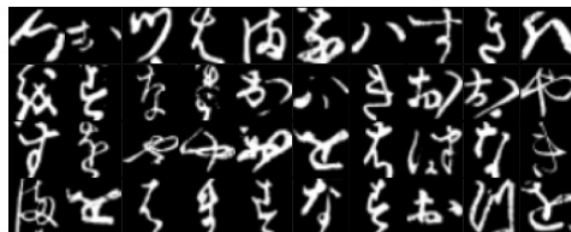
(d) Trained on OMNIGLOT



(a) CIFAR10



(b) SVHN



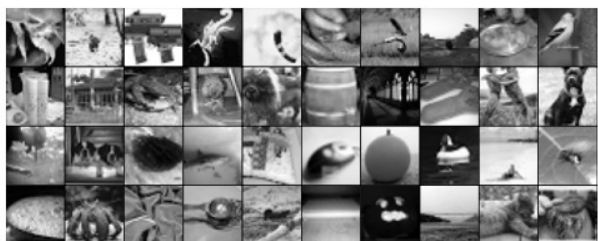
(e) KMNIST



(f) MNIST



(c) GTSRB



(d) IMAGENET

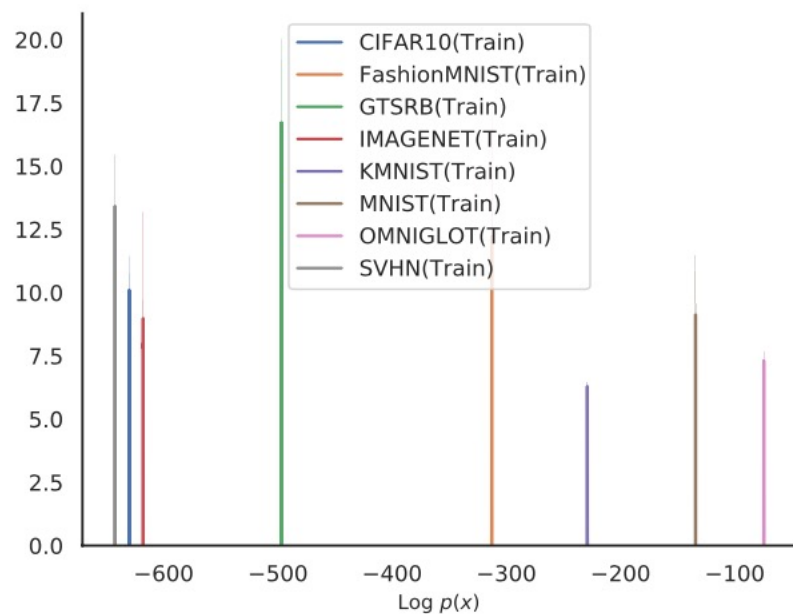


(g) OMNIGLOT

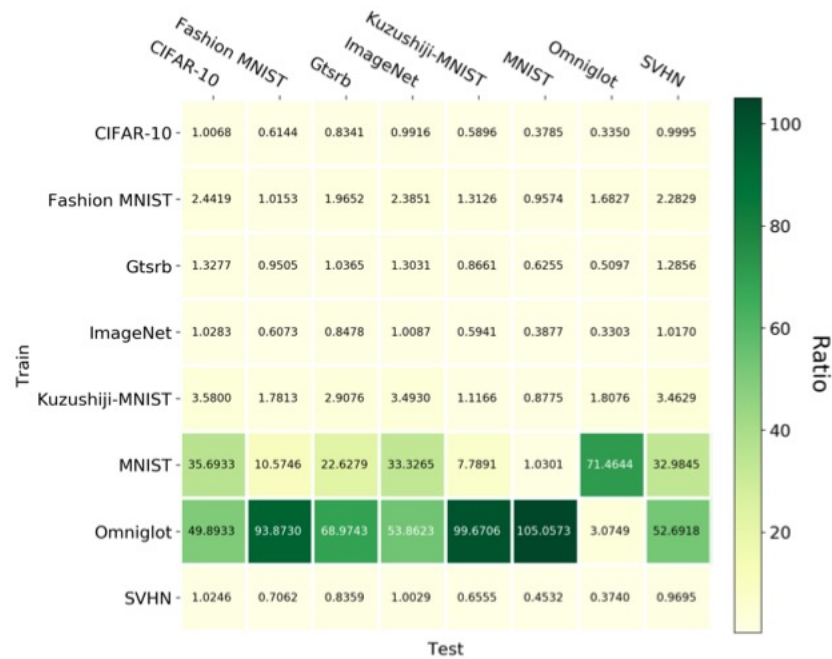


(h) FashionMNIST

➤ Does input complexity has a strong effect in VAE?



(a)

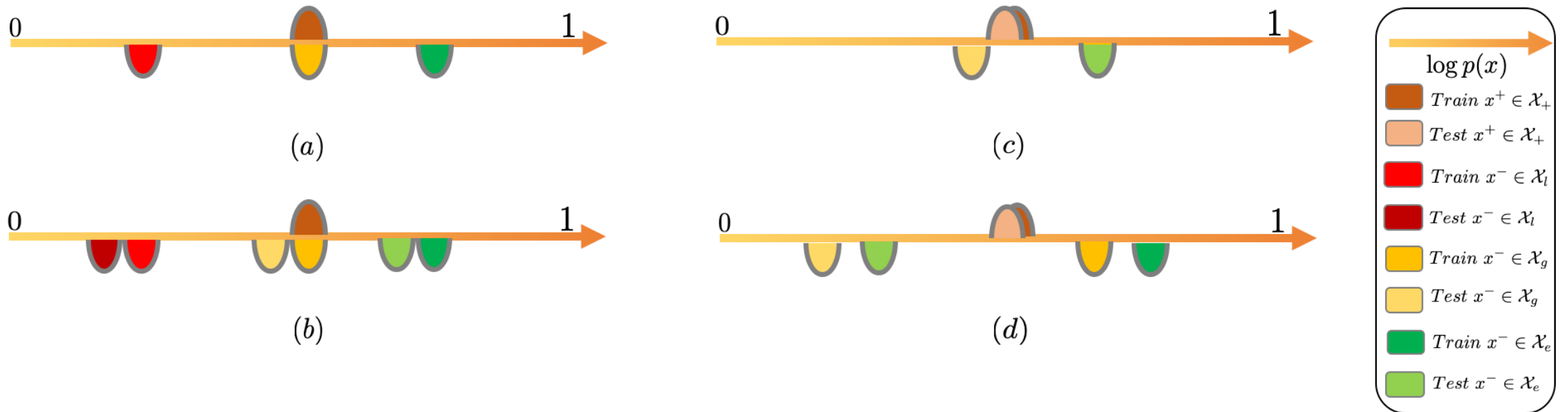


(b)

Figure 2: (a) Histogram of log-likelihoods of VAE trained on Cifar10, FashionMnist, GTSRB, IMGAENET, KMNIST(Kuzushiji-MNIST), OMNIGLOT, and SVHN, respectively. (b) Likelihood Ratios of training and testing samples (*the higher is better*). The Likelihood Ratios < 1 is represented by the likelihood of testing simple higher than training samples

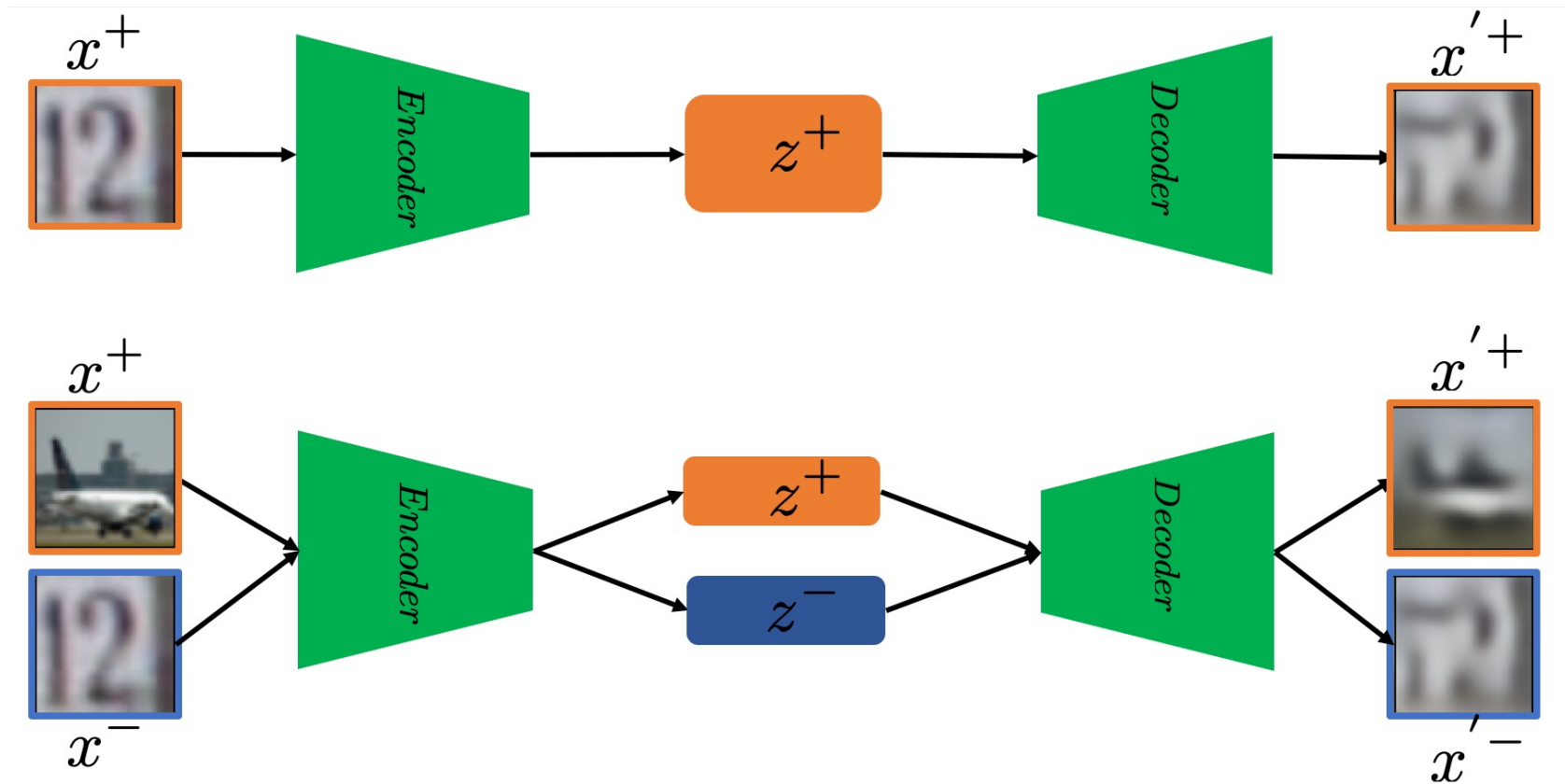
➤ Bigeminal Priors Variational auto-encoder (BPVAE)

We propose a method that feeds the external dataset (called the simple dataset) as inputs while training VAEs on the training dataset (called the basic dataset), which is more straightforward than training VAE on the basic dataset.



➤ Bigeminal Priors Variational auto-encoder (BPVAE)

VAEs can learn the features from two data distributions, assigning a higher likelihood for the basic dataset than the simple dataset. And the density estimate of VAEs can be used for detecting OOD samples.



➤ Bigeminal Priors Variational auto-encoder (BPVAE)

The BPVAE consists of an encoder, a decoder, and two priors (b-prior and s-prior). We assume that both the b-prior and s-prior belong to normal distribution. And we use the variance of a normal distribution to represent the uncertainty level. The priors are formulated as followings

$$p_b(z) \sim \mathcal{N}(z \mid \mu_z, \sigma_z^2 I)$$
$$p_s(\tilde{z}) \sim \mathcal{N}(\tilde{z} \mid \mu_{\tilde{z}}, \sigma_{\tilde{z}}^2 I)$$

where the mean value $\mu_z = \mu_{\tilde{z}} = 0$. σ_z is always set to be greater than $\sigma_{\tilde{z}}$ so that b-prior has enough capacity to capture the basic dataset features.

We modified the loss function as follows:

$$\log p(\mathbf{x}) + \log p(\mathbf{y}) = \mathbf{E}_{z \sim q_\theta(z|\mathbf{x})} [\log p_\phi(\mathbf{x} \mid z)] - \mathbf{D}_{KL} [q_\theta(z \mid \mathbf{x}) \parallel p_b(z)]$$
$$+ \mathbf{E}_{\tilde{z} \sim q_\theta(\tilde{z}|\mathbf{y})} [\log p_\phi(\mathbf{y} \mid \tilde{z})] - \mathbf{D}_{KL} [q_\theta(\tilde{z} \mid \mathbf{y}) \parallel p_b(\tilde{z})]$$

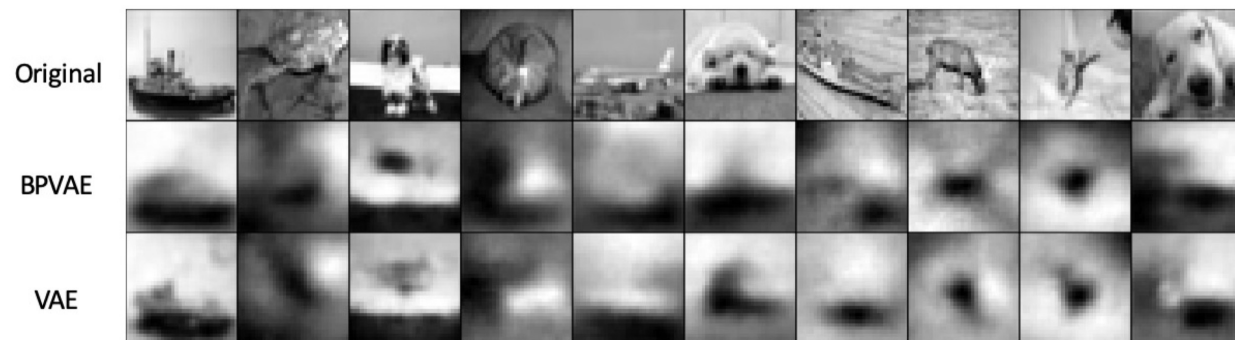
where $q_\theta(\tilde{z} \mid \mathbf{y})$ and $q_\theta(z \mid \mathbf{x})$ are the variational posterior for the simple and basic dataset, $q_\theta(\tilde{z} \mid \mathbf{y})$ and $q_\theta(z \mid \mathbf{x})$ are the decoder for the simple and basic data, and which are modeled by a neural network with their parameters θ and ϕ , respectively.

➤ Result for BPVAE

It is evident that BPVAEs obtain much better performance than standard VAEs on MNIST, while these two models achieve comparable results on CIFAR10.



(a) Test on MNIST



(b) Test on CIFAR10

Reconstruction performance for MNIST and CIFAR10 by VAEs and BPVAEs. Here CIFAR10 is used as basic dataset and MNIST is used as simple dataset.

➤ Evaluation of BPVAE

The tables demonstrate that BPVAEs can obtain much better performance than standard VAEs no matter it is evaluated by MSE, PSNR or SSIM.

Table 1: Evaluation on the basic dataset and the simple dataset

	Method	MSE	PSNR	SSIM
Evaluation on basic dataset	BPVAE	0.017	18.250	0.544
	VAE	0.016	18.282	0.543
Evaluation on simple dataset	BPVAE	0.007	22.392	0.909
	VAE	0.0346	14.831	0.601

➤ Analysis of BPVAE

Our model can cover all key representation and shift all the data distribution toward the lower-likelihood area, via combining multiple priors and training BPVAEs on a variety of selected datasets.

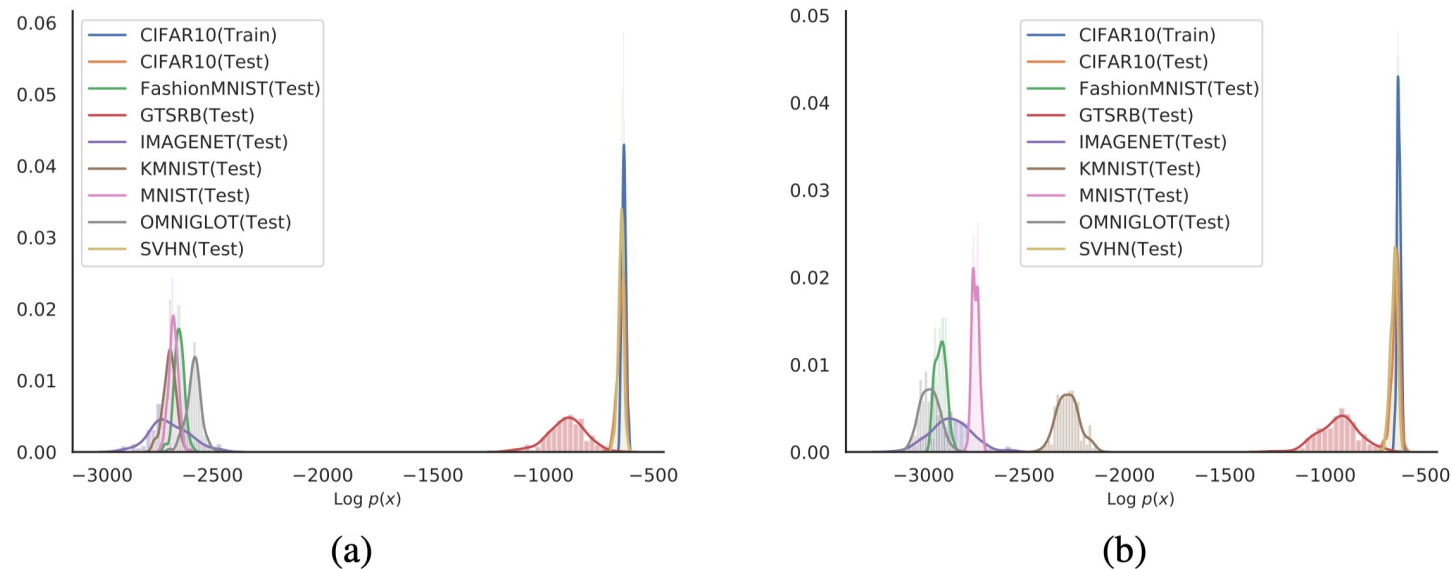


Figure 6: Histogram of log-likelihoods from VAEs model, which are trained on different groups of datasets. (a) Trained on CIFAR10(Basic), FashionMNIST(simple), and KMNIST(simple); (b) Trained on CIFAR10(Basic), FashionMNIST(simple), and MNIST(simple);

➤ Analysis of BPVAE

As depicted in the following table, our BPVAEs can achieve higher AUROC and AUPRC values than Standard VAEs, meanwhile surpassing other classical baselines. Overall, these comprehensive comparisons suggest that our proposed model is equipped with strong robustness and detection capability.

Table 2: AUROC and AUPRC for detecting OOD inputs using likelihoods of BPVAE, likelihood of VAE, and other baselines on FashionMNIST vs. MNIST datasets.


Model	AUROC	AUPRC
BPVAE(ours)	1.000	1.000
Standard VAE	0.012	0.113
Likelihood Ratio(μ, λ) Ren et al. (2019)	0.994	0.993
ODIN Liang et al. (2018)	0.752	0.763
Mahalanobis distance Lee et al. (2018)	0.942	0.928
Ensemble, 20 classifiers Lakshminarayanan et al. (2017)	0.857	0.849
WAIC, 5 models Choi et al. (2018)	0.221	0.401

➤ Summary

- **INCPVAE**
 - We apply tailored metrics to uncertainty estimation, by using which our INCPVAE framework achieve reliable uncertainty estimation and enhanced robustness.
 - We propose a novel OOD detection method via INCP-KL divergence of INCPVAE Experiments demonstrate that the INCPVAE gains an excellent understanding for the OOD inputs and our detection method achieves state-of-the-art (SOTA) performance on the challenging cases raised by Nalisnick et al. (2019a).
- **BPVAE**
 - VAEs can be well- calibrated by shifting the likelihood distribution of data with simpler complexity to lower-likelihood intervals compared to basic dataset, in which way the high-likelihoods problem of OOD can be overcome to a large extent.
 - we only impose the proposed approach on VAE model, designing the hybrid latent priors for other models like Glow, PixelCNN will be an interesting research topic. And we are expected to continue related exploration further.

➤ Future works

- Application in Medical image detection
 - Jaeger, Paul F., et al. "Retina U-Net: Embarrassingly simple exploitation of segmentation supervision for medical object detection." *Machine Learning for Health Workshop*. PMLR, 2020.
 - Tomita, Naofumi, et al. "Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides." *JAMA network open* 2.11 (2019): e1914645-e1914645.
 - Cao T, Huang C, Hui D Y T, et al. A Benchmark of Medical Out of Distribution Detection[J]. arXiv preprint arXiv:2007.04250, 2020.
- Further research VAE with information theory and manifold learning
 - Bernardo, Jose M. "Reference posterior distributions for Bayesian inference." *Journal of the Royal Statistical Society: Series B (Methodological)* 41.2 (1979): 113-128.
 - Berger, James O., José M. Bernardo, and Dongchu Sun. "The formal definition of reference priors." *The Annals of Statistics* 37.2 (2009): 905-938.
 - Nalisnick, Eric, and Padhraic Smyth. "Learning approximately objective priors." *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*. Vol. 33. 2017.
 - Cheng, Siu-Wing, and Man-Kwun Chiu. "Implicit manifold reconstruction." *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2014.
 - Fefferman, Charles, et al. "Reconstruction of a Riemannian manifold from noisy intrinsic distances." *SIAM Journal on Mathematics of Data Science* 2.3 (2020): 770-808.
 - Wei, Xian, et al. "Reconstructible nonlinear dimensionality reduction via joint dictionary learning." *IEEE transactions on neural networks and learning systems* 30.1 (2018): 175-189.



Thank you.
Any question is welcome.

