Paper Review

Rethinking Reconstruction Autoencoder-Based Out-of-Distribution Detection

YeongHyeon Park

Department of Electrical and Computer Engineering

SungKyunKwan University

Zhou, Yibo. "Rethinking Reconstruction Autoencoder-Based Out-of-Distribution Detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022.

Rethinking Reconstruction Autoencoder-Based Out-of-Distribution Detection

Yibo Zhou
Beihang University
ybzhou@impcas.ac.cn

Recommended Reading Order

1. Introduction

2. Background and Related Work

4. Practical Approach (Core of the paper)

A1. On the influence of feature norm to L2 reconstruction error (Supplementary Material)

3. Hypothesis

5. Experiments

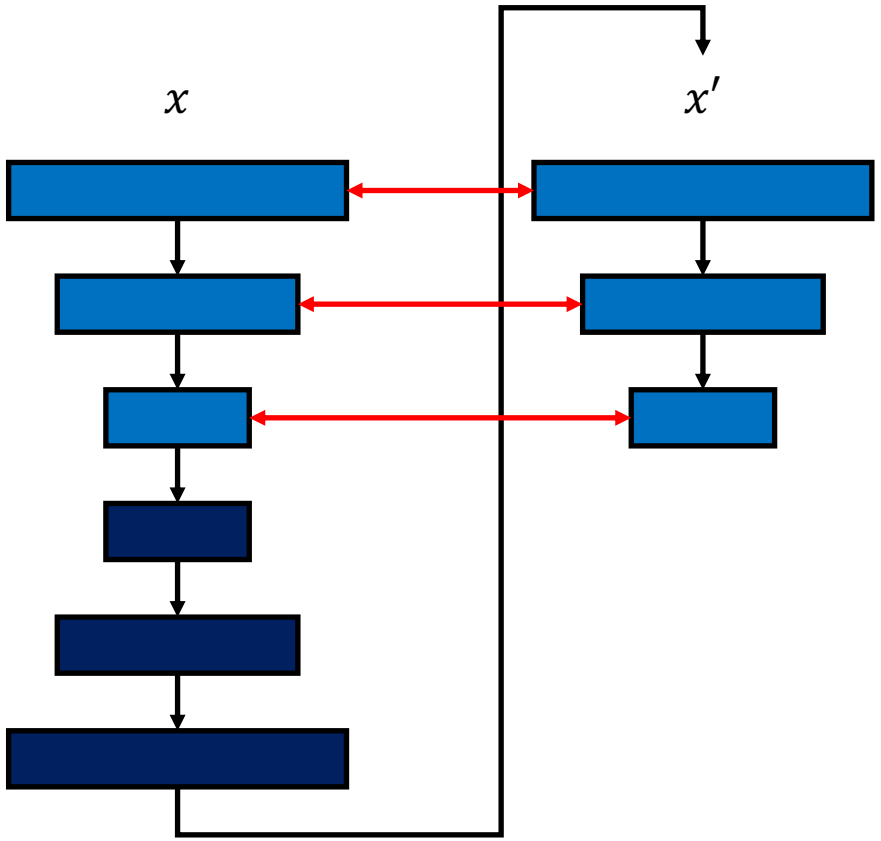
A2. Architecture of encoder and decoders (Supplementary Material)

6. Conclusion

Warm Up

Johnson, Justin, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution." European conference on computer vision (ECCV), 2016.
Zhang, Richard, et al. "The unreasonable effectiveness of deep features as a perceptual metric." Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2018.

Perceptual Loss
in anomaly detection



Perceptual Loss
- Feature inference from pre-trained model.

Learned Perceptual Image Patch Similarity (LPIPS)
- Feature inference after training the model.
- Lin: training linear layer only
- Tune: fine-tuning the pre-trained model
- Scratch: scratch learning from randomly initialized model

LPIPS for Discriminator

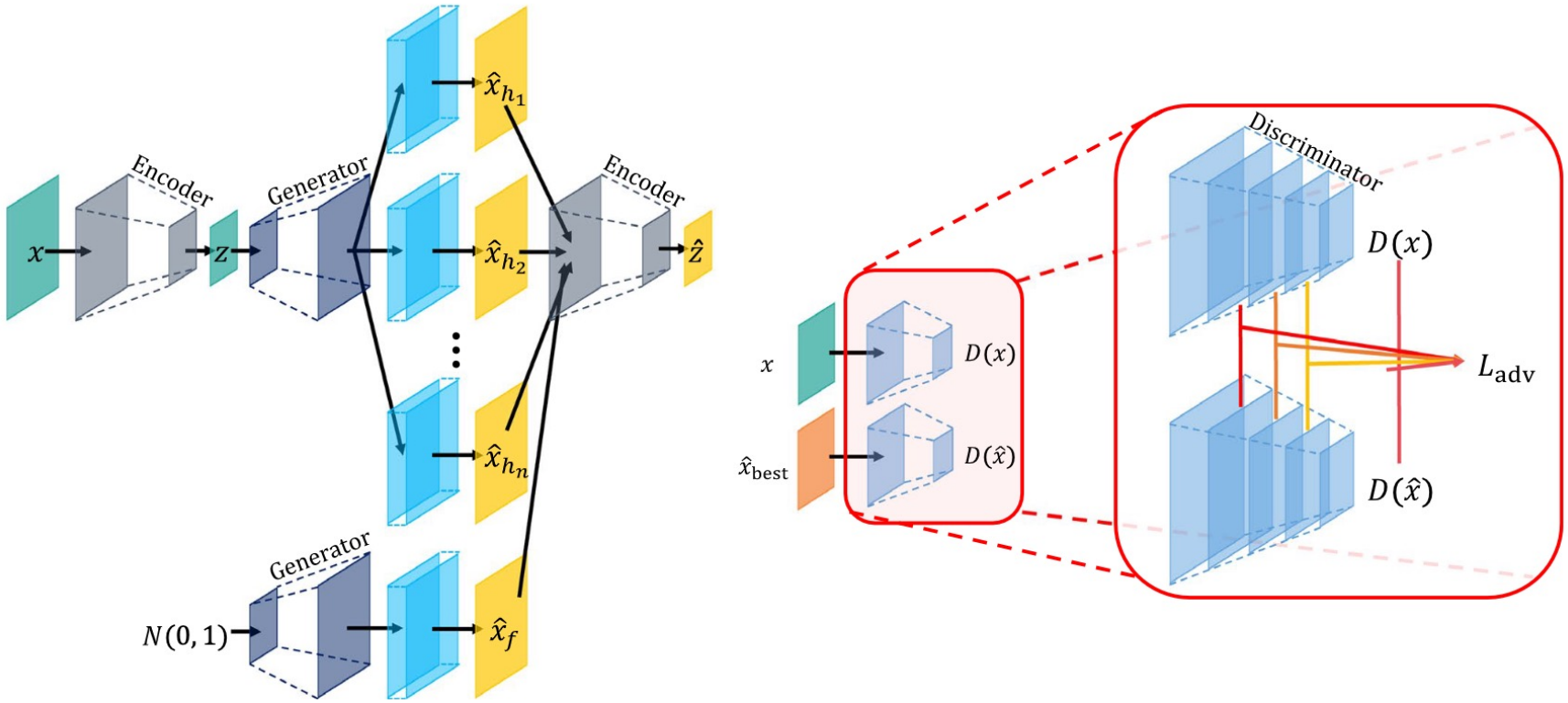


FIGURE 2 Architecture of HP-GAN. We use a single encoder, generator, and discriminator, but the last layer of the generator is constructed with the multiple hypothesis (in cyan). Pruning is conducted after generating the branches by multiple hypotheses. The HP-GAN is basically constructed with three convolutional blocks for the encoder, decoder, and discriminator each. Each convolutional block includes two convolutional layers with the ELU activation function. A max pooling layer is applied between the convolutional blocks of the encoder and discriminator. The generator uses transposed convolution between convolutional blocks for up-scaling

Approaches

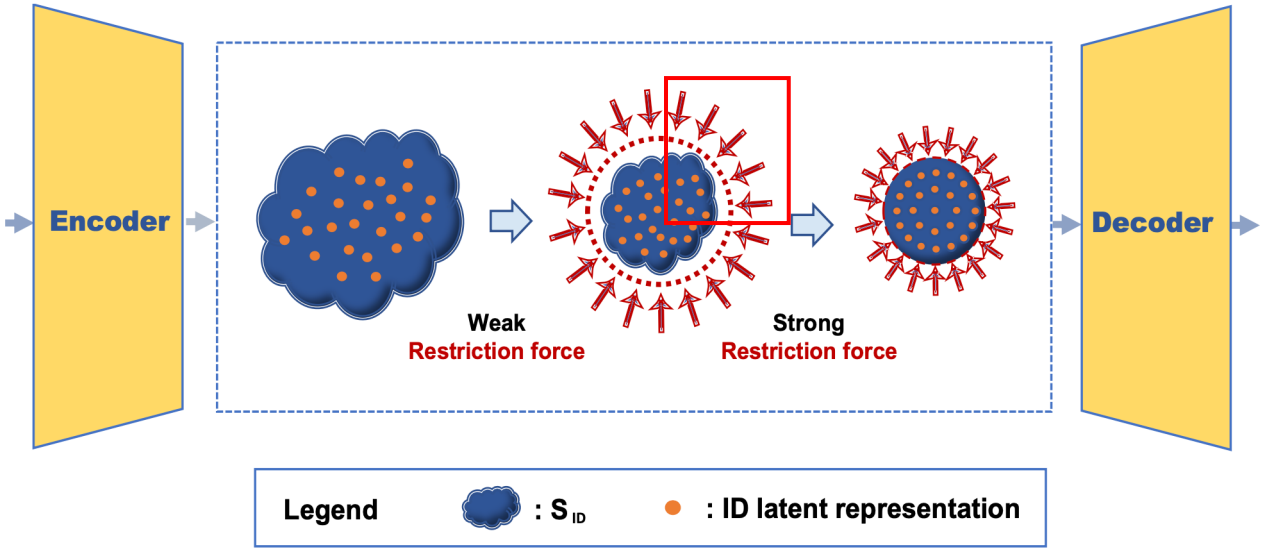
Contributions

- Condensing minimization of the latent space while reserving sufficient reconstructive power.
- Layer-wise semantic reconstruction is developed.
- Validation process using various benchmark datasets.

Symbols and Abbreviations

- x, X: ID sample and set of them
- z, Z: OoD sample (not latent vector) and set of them
- S: span of distribution
  - S_{ID}: span of in-distribution
  - S_{OoD}: span of out-of-distribution
- E, D: pre-trained encoder and decoder pairs
- F: mapping function, reconstruction error to probability
- P: probability that v belongs to V
- v: AV feature
- f: feature map / latent vector
  - $f_i(v)$: feature map of AV function on i-th layer
- OoD: Out-of-Distribution
- ID: In-Distribution
- AV: Activation Vector

Latent Space Condensation



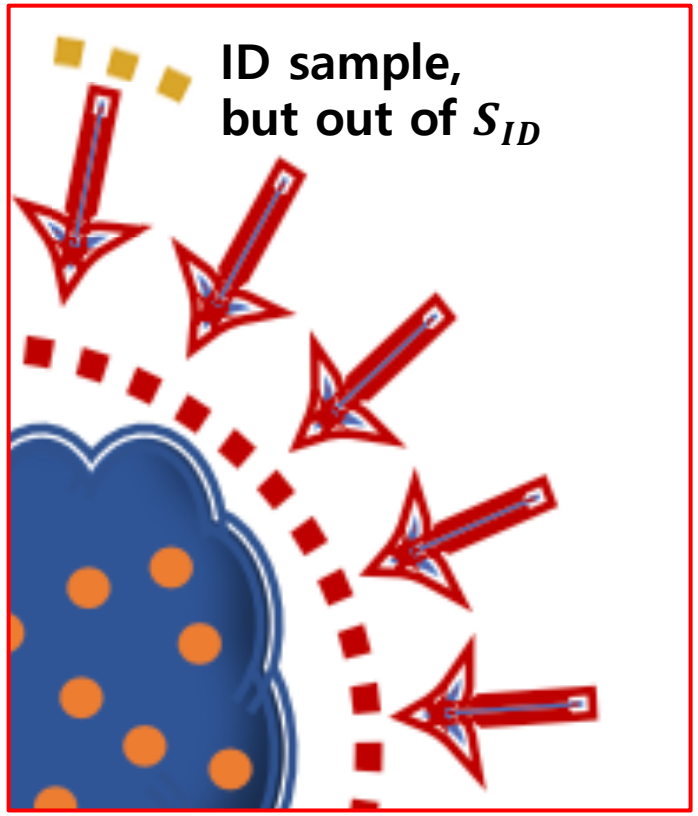


Figure 2. Illustration of the transition of $\mathcal{S}_{ID}$ when the restrictive power imposed over latent codes increases. During training, any deviation of the latent codes from the latent space (red-dot circle) would be penalized greatly. By tightening this space sufficiently, in principle it would be mostly utilized to satisfy the jointly learned reconstruction task. Thus, detecting the outlier of $\mathcal{S}_{ID}$ is approximately equal to identifying the feature outside this latent space.

Why?

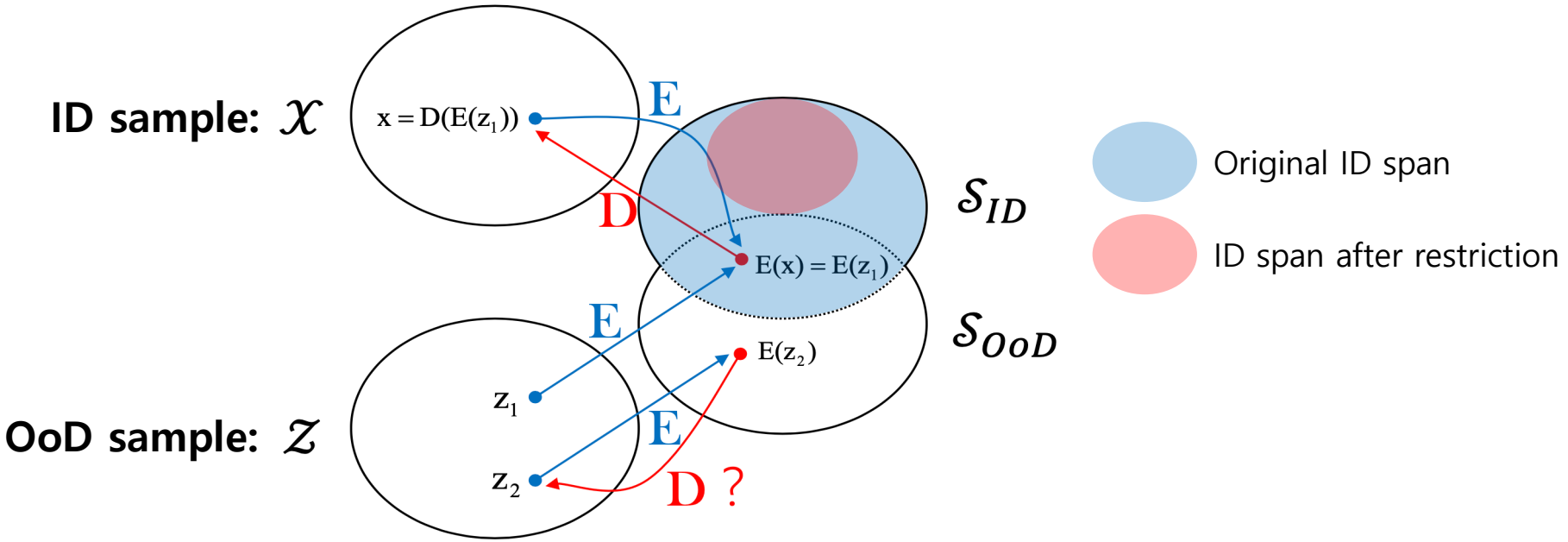


Figure 1. Illustration of the described quadruplet domain translation. For an OoD sample z_1 encoded into $\mathcal{S}_{ID} \cap \mathcal{S}_{OoD}$, its latent representation $E(z_1)$ is equal potentially to that of an ID sample x. Therefore, $E(z_1)$ can be decoded to a different sample x within $\mathcal{X}$, resulting in a large reconstruction error. However, for an OoD sample z_2 with latent representation $E(z_2)$ lying outside $\mathcal{S}_{ID}$, it offers no guarantee that it could not be reconstructed well.

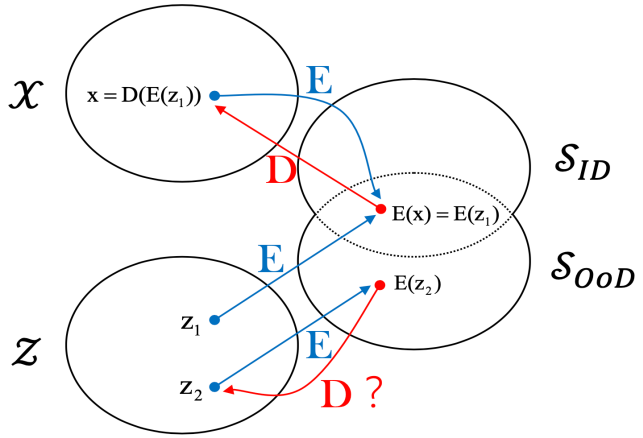
How to Latent Condensation?

$$\mathcal{L}_{regularizer} = -\sum_{i=1}^{k}\sum_{j=1}^{C} \mathbb{1}(j = y_i) \log S(W\boldsymbol{v}_i)_j,$$

- Entropy (regularization term) is activated only when the input belongs to the ID class.

Data certainty / Normality Measure

Further Enhancement



agnostic input $\boldsymbol{r}$

$$P(\boldsymbol{r} \in \mathcal{X} | E(\boldsymbol{r}) \in \mathcal{S}_{ID}) = F(Dist(\boldsymbol{r}, D(E(\boldsymbol{r})))),$$

* Normality score is probability of r in X when encoded r is in S_{ID}

$$\begin{aligned} P(\boldsymbol{r} \in \mathcal{X}) &= P(\boldsymbol{r} \in \mathcal{X}, E(\boldsymbol{r}) \in \mathcal{S}_{ID}) + P(\boldsymbol{r} \in \mathcal{X}, E(\boldsymbol{r}) \notin \mathcal{S}_{ID}) \\ &= P(\boldsymbol{r} \in \mathcal{X}, E(\boldsymbol{r}) \in \mathcal{S}_{ID}) + 0 \\ &= P(\boldsymbol{r} \in \mathcal{X} | E(\boldsymbol{r}) \in \mathcal{S}_{ID}) \cdot P(E(\boldsymbol{r}) \in \mathcal{S}_{ID}) \\ &= F(Dist(\boldsymbol{r}, D(E(\boldsymbol{r})))) \cdot P(E(\boldsymbol{r}) \in \mathcal{S}_{ID}). \end{aligned}$$

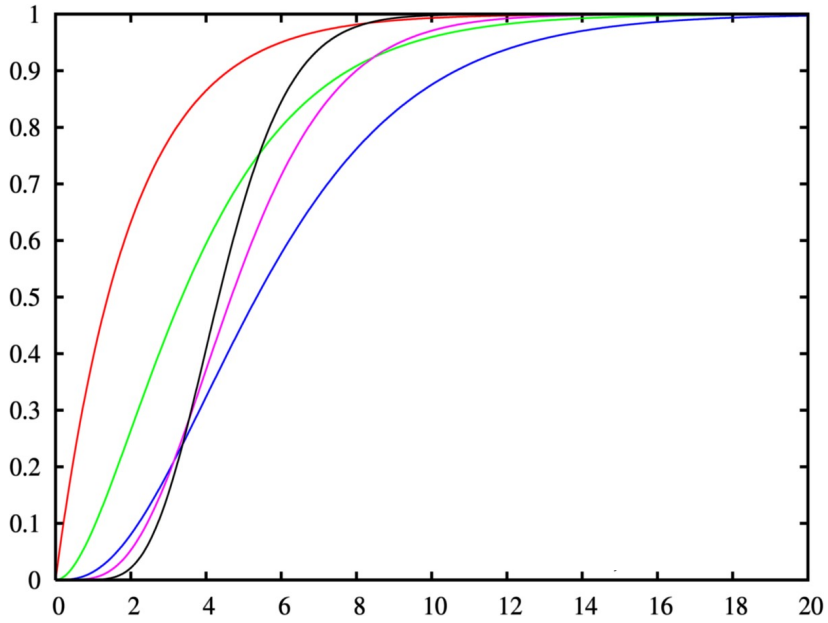
$P(\boldsymbol{r} \in \mathcal{X}, E(\boldsymbol{r}) \notin \mathcal{S}_{ID}) = 0$, assuming $\forall x \in X, E(x) \in S_{ID}$

Term1 of Data Certainty

$$F(L2(\boldsymbol{f}, \tilde{\boldsymbol{f}})) = \Psi(L2(\boldsymbol{f}, \tilde{\boldsymbol{f}})|\mu, \sigma + \epsilon),$$

L2 or NL2

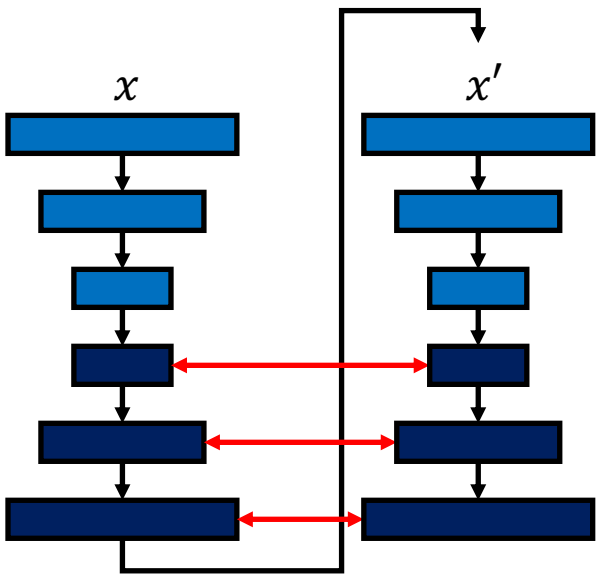
- Ψ: complementary cumulative density function (CCDF), same as 1-CDF
- μ: mean of L2 calculated from validation set
- σ: standard-deviation of L2 calculated from validation set
- ϵ: epsilon, user-defined value



Term2 of Data Certainty

∗ T: temperature (scaling coefficient)
∗∗ S: softmax activation

$$P(\boldsymbol{v} \in V) = \Phi(S(\frac{W\boldsymbol{v}}{T})_{\overline{y}}|\mu_0, \sigma_0 + \epsilon_0)$$
$$\cdot \Psi(\| \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} - \frac{D_1(W\boldsymbol{v})}{\|\boldsymbol{v}\|} \||\mu_1, \sigma_1 + \epsilon_1)$$
$$\cdot \Psi(\| \frac{W\boldsymbol{v}}{\|W\boldsymbol{v}\|} - \frac{D_2(S(\frac{W\boldsymbol{v}}{T}))}{\|\frac{W\boldsymbol{v}}{T}\|} \||\mu_2, \sigma_2 + \epsilon_2).$$



i: data sample index

Algorithm 1 Training pipeline

Require: ID training set: $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{k}$, and ID validation set: $\{(\boldsymbol{x}_i, y_i)\}_{i=k+1}^{n}$

Require: Network $M(\cdot)$ fully trained on ID training set for classification of ID classes

1: Freeze all the parameters of network $M(\cdot)$, and jointly train $W \in \mathbb{R}^{H \times C}$($C$ is the number of ID classes) and two decoders $D_1 \& D_2$ to minimize the loss $\mathcal{L}$

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \lambda \cdot \mathcal{L}_{regularizer}$$
$$\mathcal{L}_1 = \sum_{i=1}^{k} \|\boldsymbol{v}_i - D_1(W\boldsymbol{v}_i)\|, \mathcal{L}_2 = \sum_{i=1}^{k} \| \frac{W\boldsymbol{v}_i}{T} - D_2(S(\frac{W\boldsymbol{v}_i}{T}))\|$$
$$\mathcal{L}_{regularizer} = -\sum_{i=1}^{k}\sum_{j=1}^{C} \mathbb{1}(j = y_i) \log S(W\boldsymbol{v}_i)_j,$$

where $\boldsymbol{v}_i$ is $\boldsymbol{x}_i$'s AV feature extracted in $M(\cdot)$ and λ is the weight of regularization loss

2: After training, compute:

$$(\mu_0, \sigma_0) = norm.fit(\{S(\frac{W\boldsymbol{v}_i}{T})_{\overline{y}_i}\}_{i=k+1}^{n})$$
$$(\mu_1, \sigma_1) = norm.fit(\{\| \frac{\boldsymbol{v}_i}{\|\boldsymbol{v}_i\|} - \frac{D_1(W\boldsymbol{v}_i)}{\|\boldsymbol{v}_i\|} \|\}_{i=k+1}^{n})$$
$$(\mu_2, \sigma_2) = norm.fit(\{\| \frac{W\boldsymbol{v}_i}{\|W\boldsymbol{v}_i\|} - \frac{D_2(S(\frac{W\boldsymbol{v}_i}{T}))}{\|\frac{W\boldsymbol{v}_i}{T}\|} \|\}_{i=k+1}^{n})$$

3: **return** $D_1, D_2, W, (\mu_0, \sigma_0), (\mu_1, \sigma_1)$ *and* (μ_2, σ_2)

L: last index of layers
x: random input / hidden feature

Upper Bound of L2 Distance

$$
\begin{aligned}
f^L(\boldsymbol{x}) =& W^L\sigma(W^{L-1}\sigma(\\
& \ldots\sigma(W^1\boldsymbol{x}+\boldsymbol{b}^1)\ldots)+\boldsymbol{b}^{L-1})+\boldsymbol{b}^L \\
=& \Gamma\boldsymbol{x}+B
\end{aligned}
$$

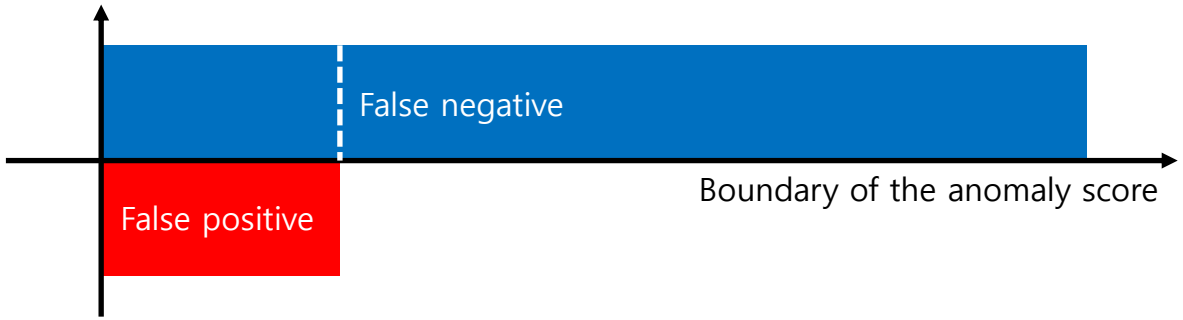
Transformation including activation functions

$$
\begin{aligned}
\|\boldsymbol{x}-f^L(\boldsymbol{x})\| &= \|\boldsymbol{x}-\Gamma\boldsymbol{x}-B\| \\
&\leq \|\boldsymbol{x}-\Gamma\boldsymbol{x}\|+\|B\| \\
&\leq \|I-\Gamma\|\|\boldsymbol{x}\|+\|B\|.
\end{aligned}
$$

Fixed upper bound coefficient of L2

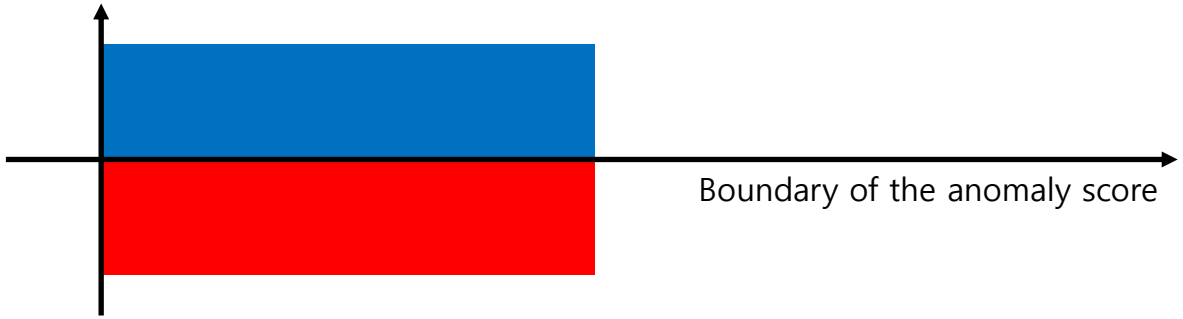
$z \rightarrow$ low activation $\rightarrow$ small norm '$\|z\|$' $\rightarrow$ detection failure (false positive)

Upper Bound and Anomaly Score



$\|I - \Gamma\|\|x\|$

$\|I - \Gamma\|\|z\|$



$\|I - \Gamma\|\|x\|/\|x\|$

$\|I - \Gamma\|\|z\|/\|z\|$

$$\frac{\|I - \Gamma\|\|x\|}{\|x\|} = \left\| \frac{x}{\|x\|} - \frac{\Gamma x}{\|x\|} \right\|$$

Normalized L2 Distance

$$Dist(\boldsymbol{f}, \tilde{\boldsymbol{f}}) = NL2(\boldsymbol{f}, \tilde{\boldsymbol{f}}) = \| \frac{\boldsymbol{f}}{\|\boldsymbol{f}\|} - \frac{\tilde{\boldsymbol{f}}}{\|\boldsymbol{f}\|} \|,$$

Before
z → low activation → limit the upper bound → false positive
x → high activation → high-anomaly score → false negative

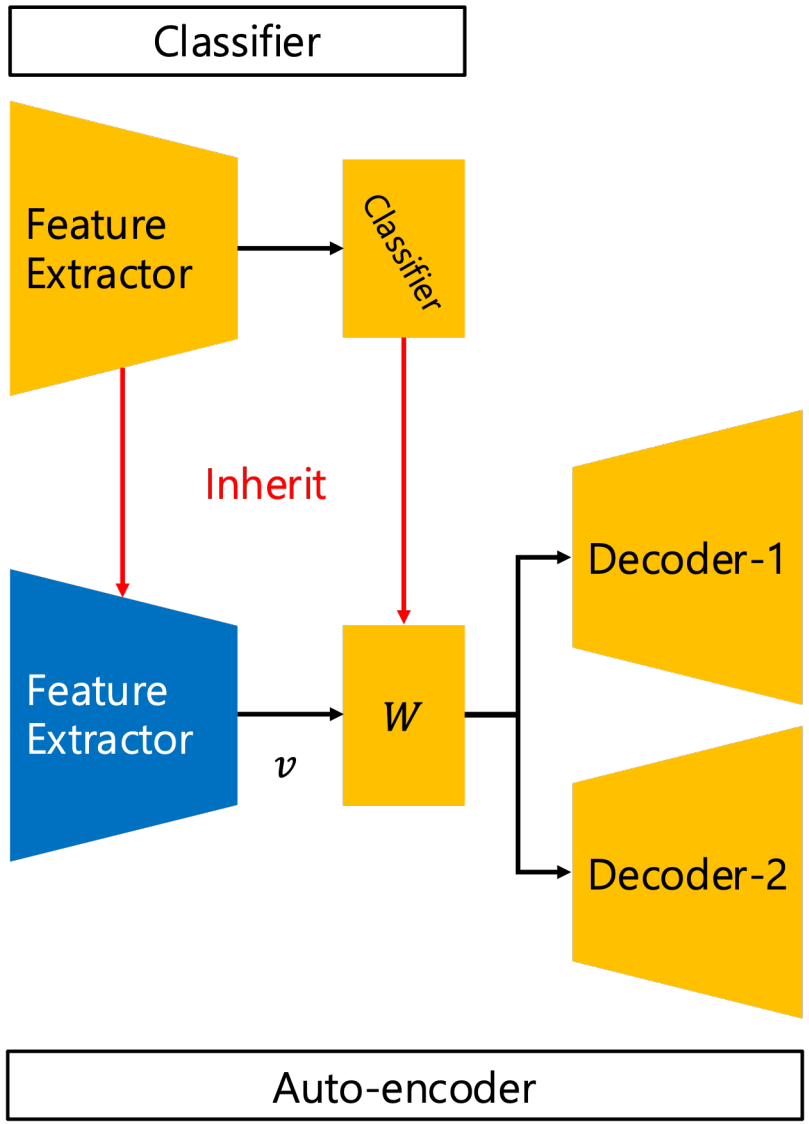
After
z → low activation → normalized activation [0, 1] → true negative
x → high activation → normalized activation [0, 1] → true positive

Experiments

Training

Freezing

Training procedure



$$\mathcal{L}_1 = \sum_{i=1}^{k} \|\boldsymbol{v}_i - D_1(W\boldsymbol{v}_i)\|$$ (not L1 and L2 distance)

$$\mathcal{L}_2 = \sum_{i=1}^{k} \| \frac{W\boldsymbol{v}_i}{T} - D_2(S(\frac{W\boldsymbol{v}_i}{T}))\|$$

$$\mathcal{L}_{regularizer} = -\sum_{i=1}^{k}\sum_{j=1}^{C} \mathbb{1}(j = y_i) \log S(W\boldsymbol{v}_i)_j,$$

Autoregressive term to compress the latent space

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \lambda \cdot \mathcal{L}_{regularizer}$$

* T: temperature (user-defined)

Inference procedure

$$(\mu_0, \sigma_0) = norm.fit(\{S(\frac{W\boldsymbol{v}_i}{T})_{\overline{y}_i}\}_{i=k+1}^{n})$$

$$(\mu_1, \sigma_1) = norm.fit(\{\| \frac{\boldsymbol{v}_i}{\|\boldsymbol{v}_i\|} - \frac{D_1(W\boldsymbol{v}_i)}{\|\boldsymbol{v}_i\|} \|\}_{i=k+1}^{n})$$

$$(\mu_2, \sigma_2) = norm.fit(\{\| \frac{W\boldsymbol{v}_i}{\|W\boldsymbol{v}_i\|} - \frac{D_2(S(\frac{W\boldsymbol{v}_i}{T}))}{\|\frac{W\boldsymbol{v}_i}{T}\|} \|\}_{i=k+1}^{n})$$

$$\begin{aligned} P(\boldsymbol{v} \in V) = \Phi(S(\frac{W\boldsymbol{v}}{T})_{\overline{y}} | \mu_0, \sigma_0 + \epsilon_0) \\ \cdot \Psi(\| \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} - \frac{D_1(W\boldsymbol{v})}{\|\boldsymbol{v}\|} \| | \mu_1, \sigma_1 + \epsilon_1) \\ \cdot \Psi(\| \frac{W\boldsymbol{v}}{\|W\boldsymbol{v}\|} - \frac{D_2(S(\frac{W\boldsymbol{v}}{T}))}{\|\frac{W\boldsymbol{v}}{T}\|} \| | \mu_2, \sigma_2 + \epsilon_2). \end{aligned}$$

(Normality measure)

Yu, Fisher, et al. "LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop." arXiv preprint. 2015.
Xu, Pingmei, et al. "Turkergaze: Crowdsourcing saliency with webcam based eye tracking." arXiv preprint. 2015.
Le, Ya, and Xuan Yang. "Tiny imagenet visual recognition challenge." Stanford University. CS231n. 2015.
Vyas, Apoorv, et al. "Out-of-distribution detection using an ensemble of self supervised leave-out classifiers." Proceedings of the European Conference on Computer Vision (ECCV). 2018.
Hsu, Yen-Chang, et al. "Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020.
Thulasidasan, Sunil, et al. "An effective baseline for robustness to distributional shift." IEEE International Conference on Machine Learning and Applications (ICMLA). 2021.

Experiments for OoD

SOTA models
- ELOC: Ensemble of self supervised Leave-Out Classifiers @ ECCV 2018
- GODIN: Generalized ODIN @ CVPR 2020
- DAC: Deep Abstaining Classifier @ ICMLA 2021

Table 1. OoD detection results in CIFAR-10 and CIFAR-100. Our method is compared with three SOTA methods of DAC, ELOC and GODIN. For fair comparison, we use the results of DAC, ELOC and GODIN reported in original papers. If no result is reported as a certain setting, it is marked as −. Also, DAC and GODIN did not report experimental results in detection error and AUPR-in. For each evaluation metric, ↑ means that larger value is better and ↓ indicates that lower value is better. All values are percentages.

OoD Dataset			FPR@95%TPR ↓				Detection Error ↓		AUROC ↑				AUPR-In ↑	
			ELOC	DAC	GODIN	ours	ELOC	ours	ELOC	DAC	GODIN	ours	ELOC	ours
WRN-28-10	CIFAR-10	TINc	0.8	-	-	**0.5**	2.2	**1.9**	**99.8**	-	-	**99.8**	**99.8**	**99.8**
		TINr	2.9	1.9	-	**1.5**	3.8	**3.1**	99.4	**99.5**	-	**99.5**	99.4	**99.6**
		LSUNc	1.9	-	-	**0.8**	3.2	**2.0**	99.6	-	-	**99.8**	99.6	**99.8**
		LSUNr	0.9	1.5	-	**0.5**	2.5	**2.2**	**99.7**	99.6	-	**99.7**	**99.7**	**99.7**
		iSUN	-	-	-	2.9	-	4.0	-	-	-	99.2	-	99.4
WRN-28-10	CIFAR-100	TINc	9.2	-	-	**1.5**	6.7	**3.4**	98.2	-	-	**99.4**	98.4	**99.5**
		TINr	24.5	18.7	-	**6.6**	11.6	**6.4**	95.2	94.9	-	**98.4**	95.5	**98.0**
		LSUNc	14.2	-	-	**3.7**	8.2	**4.8**	97.4	-	-	**99.0**	97.6	**99.1**
		LSUNr	16.5	9.2	-	**5.5**	9.1	**5.8**	96.8	97.9	-	**98.5**	97.0	**98.6**
		iSUN	-	-	-	9.0	-	7.4	-	-	-	97.9	-	98.2
Dense-BC	CIFAR-10	TINc	**1.2**	-	6.6	3.7	**2.6**	4.6	**99.7**	-	98.7	98.9	**99.7**	99.1
		TINr	**2.9**	-	4.2	10.2	**3.8**	7.1	**99.3**	-	99.1	97.7	**99.3**	98.1
		LSUNc	3.4	-	8.5	**1.7**	4.1	**3.5**	99.3	-	98.3	**99.4**	99.3	**99.5**
		LSUNr	**0.8**	-	2.4	14.6	**2.2**	7.8	**99.8**	-	99.4	97.2	**99.8**	97.7
		iSUN	-	-	**2.5**	17.3	-	8.8	-	-	**99.4**	96.7	-	97.5
Dense-BC	CIFAR-100	TINc	**8.3**	-	12.2	14.8	**6.3**	6.6	**98.4**	-	97.6	97.1	**98.6**	98.0
		TINr	20.5	-	**6.7**	14.8	10.0	**8.8**	96.3	-	**98.6**	96.7	96.7	**97.2**
		LSUNc	14.7	-	25.0	**7.8**	8.5	**5.5**	97.4	-	95.3	**98.0**	97.6	**98.7**
		LSUNr	16.2	-	**6.2**	14.8	8.8	**8.0**	97.0	-	**98.7**	96.8	97.4	**97.5**
		iSUN	-	-	18.6	**18.0**	-	9.9	-	-	**98.4**	96.1	-	96.9

hard to say improvement...

Ablation Study

Normality Measure Improvement

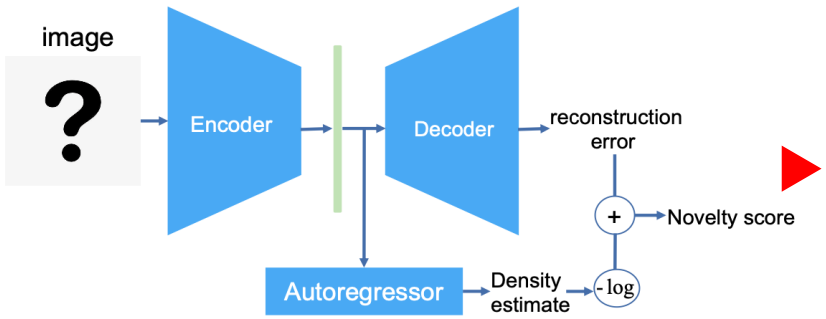


Figure 1. The overall framework of 1^{st} LSA.

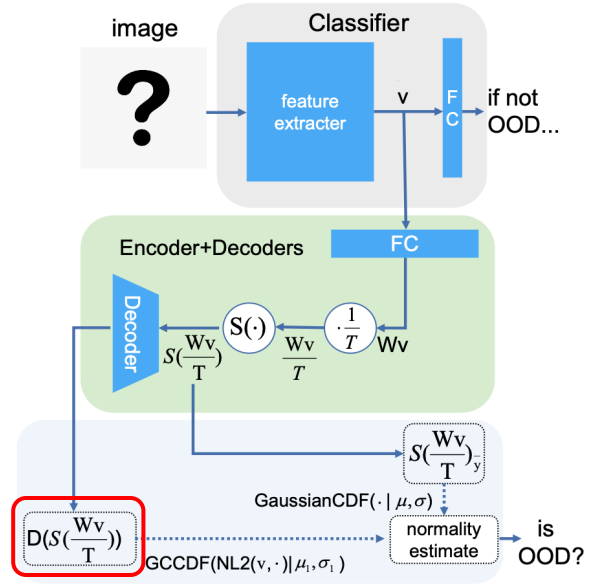


Figure 2. The overall framework of 4^{th} -AutoReg+CE.

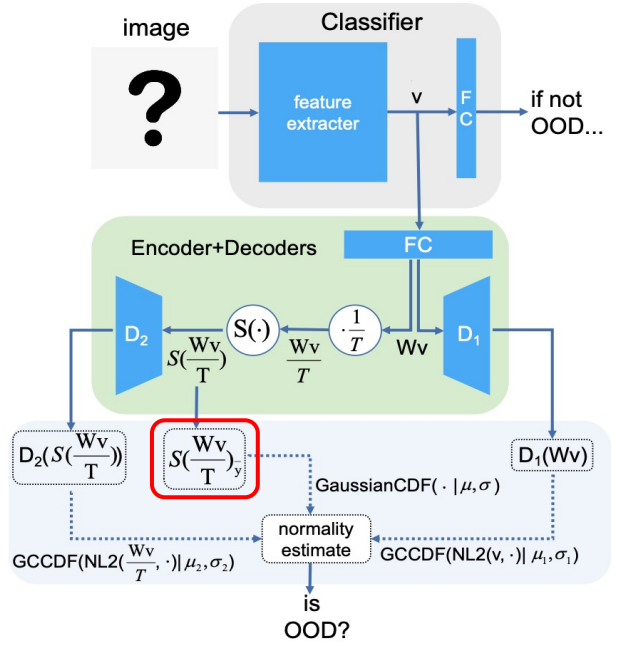


Figure 3. The overall framework of 5^{th}-basic+layerwise.

* LSA: latent space autoregression

Ablation Study

Table 3.

Methods		1^{st} LSA image,L2,AutoReg,basic	$\longrightarrow$ 2^{nd} -image+feature	$\longrightarrow$ 3^{rd} -L2+NL2	$\longrightarrow$ 4^{th} -AutoReg+CE	$\longrightarrow$ 5^{th} -basic+layerwise	$\longrightarrow$ 6^{th} +epsilon
TINc	FPR@95%TPR ↓	42.0	99.5	5.9	3.0	**0.2**	1.5
	AUROC ↑	89.2	36.7	98.6	99.1	91.3	**99.4**
TINr	FPR@95%TPR ↓	51.2	78.8	20.3	17.7	**2.5**	6.6
	AUROC ↑	89.4	80.1	95.2	96.4	90.8	**98.4**
LSUNc	FPR@95%TPR ↓	55.8	100.0	4.7	4.1	**0.8**	3.7
	AUROC ↑	70.0	7.5	98.1	98.7	91.3	**99.0**
LSUNr	FPR@95%TPR ↓	28.2	65.6	20.0	17.7	**1.7**	5.5
	AUROC ↑	93.3	80.5	95.7	96.0	91.1	**98.5**
iSUN	FPR@95%TPR ↓	52.5	66.4	26.0	22.4	**2.8**	9.0
	AUROC ↑	89.4	82.3	94.8	95.8	90.8	**97.9**

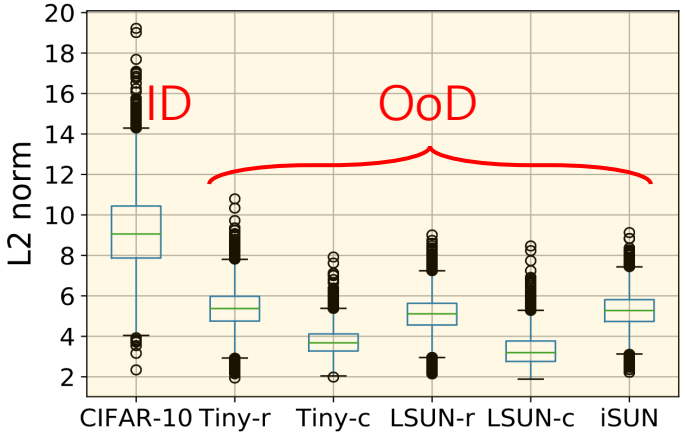


Figure 4. The distribution of AV features' L2-norm for CIFAR-10 (ID) test set and various OoD datasets. Features are extracted in the Wide-ResNet.

- image: target reconstruction
- L2: reconstruction loss
- AutoReg: method to compress the latent space
- basic: image is used to compute novelty score
- epsilon: epsilon to generate CDF

Performance Change

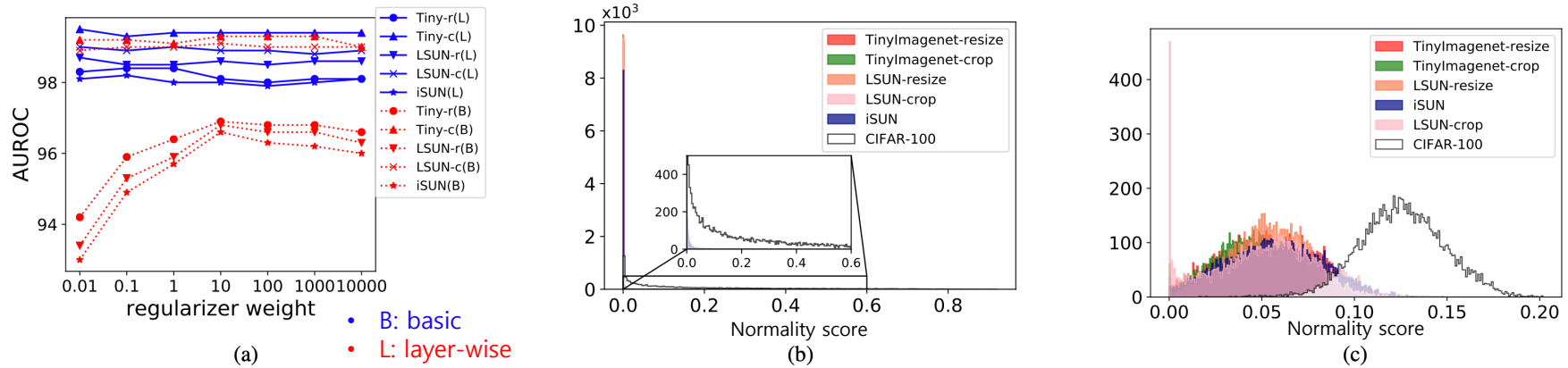


Figure 5. (a): AUROC as a function of the weight of regularizer λ for our framework of layerwise reconstruction (blue) vs. the basic framework (red). (b): The distributions of the normality scores computed from Eq.9 with ϵ_i applied as 0 and (c) with ϵ_i applied as $10 \times \sigma_i$.

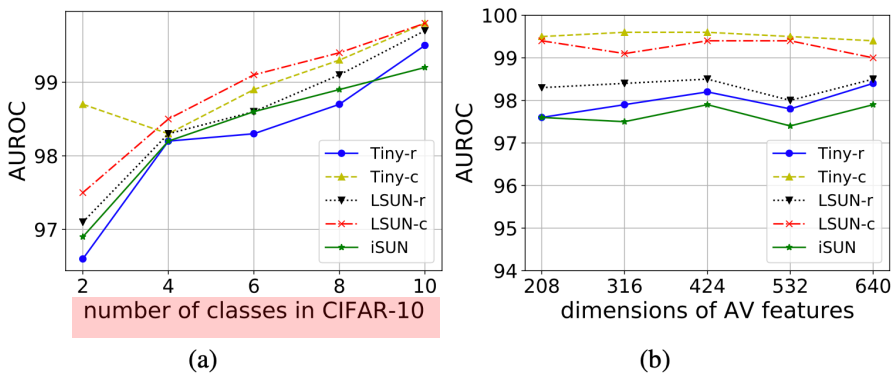


Figure 6. AUROC for OoD detection as function of (a): number of ID classes in CIFAR-10 and (b): dimensions of AV features serving as input information of autoencoder.

Conclusion

Pros & Cons

Pros

- OoD detection performance is significantly improved by normalizing the L2 distance (NL2)
  - NL2 overcomes the distance upper bound of lower activations.

- Mathematical interpretations/proof are provided for all approaches/hypotheses.

Cons

- Needs a pre-trained classification model or training a classification model.
  - Also, class-balanced dataset is recommended.

- For achieving good performance, the distance must be calculated in a perceptual loss manner.
  - Same as feature map distance.
  - Computational cost will be increased.

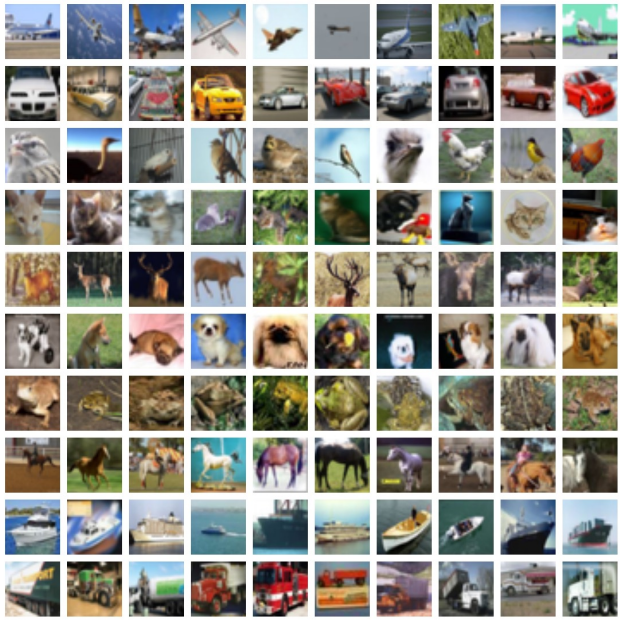
Appendix A
Dataset

Krizhevsky, Alex, and Geoffrey Hinton. "Learning multiple layers of features from tiny images." 2009.

CIFAR-10/100

Canadian Institute for Advanced Research

airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck



Superclass	Classes
aquatic mammals	beaver, dolphin, otter, seal, whale
fish	aquarium fish, flatfish, ray, shark, trout
flowers	orchids, poppies, roses, sunflowers, tulips
food containers	bottles, bowls, cans, cups, plates
fruit and vegetables	apples, mushrooms, oranges, pears, sweet peppers
household electrical devices	clock, computer keyboard, lamp, telephone, television
household furniture	bed, chair, couch, table, wardrobe
insects	bee, beetle, butterfly, caterpillar, cockroach
large carnivores	bear, leopard, lion, tiger, wolf
large man-made outdoor things	bridge, castle, house, road, skyscraper
large natural outdoor scenes	cloud, forest, mountain, plain, sea
large omnivores and herbivores	camel, cattle, chimpanzee, elephant, kangaroo
medium-sized mammals	fox, porcupine, possum, raccoon, skunk
non-insect invertebrates	crab, lobster, snail, spider, worm
people	baby, boy, girl, man, woman
reptiles	crocodile, dinosaur, lizard, snake, turtle
small mammals	hamster, mouse, rabbit, shrew, squirrel
trees	maple, oak, palm, pine, willow
vehicles 1	bicycle, bus, motorcycle, pickup truck, train
vehicles 2	lawn-mower, rocket, streetcar, tank, tractor

Yu, Fisher, et al. "LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop." arXiv preprint. 2015.

Large-scale Scene Understanding (LSUN)

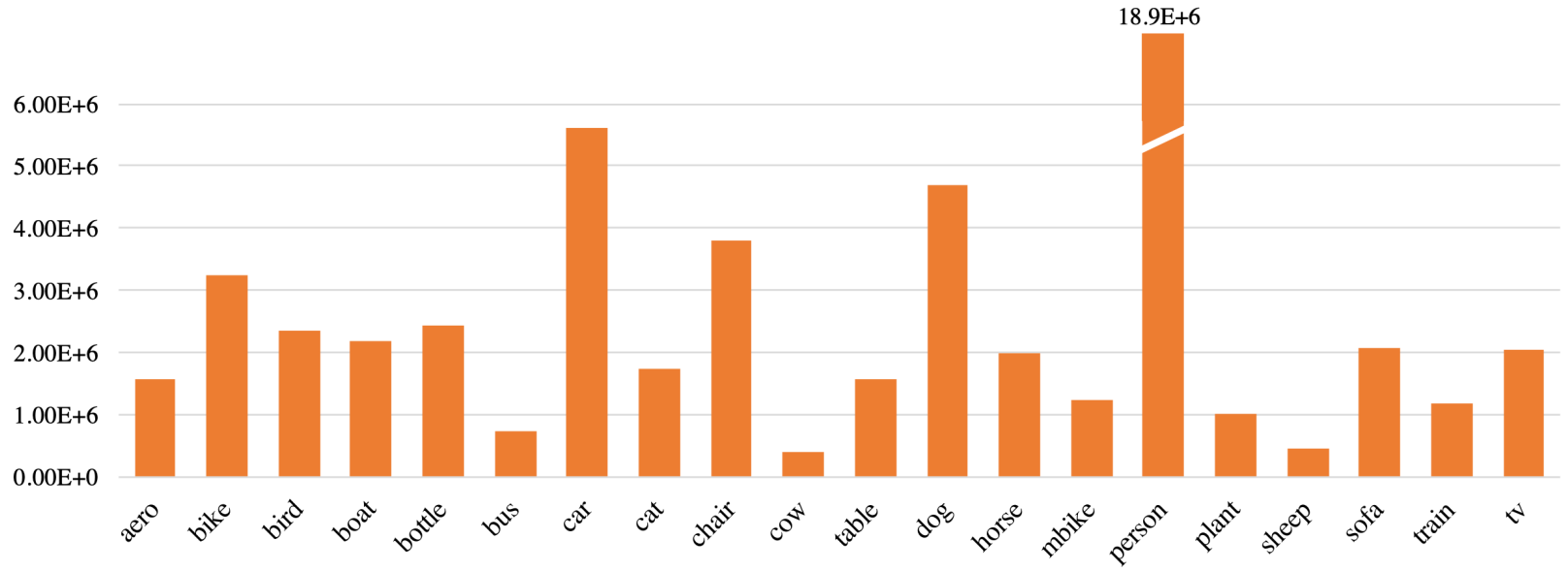


Figure 3: Number of images in our object categories. Compared to ImageNet dataset, we have more images in each category, even comparing only basic level categories.

Gaze traces on images from the SUN dataset (iSUN)

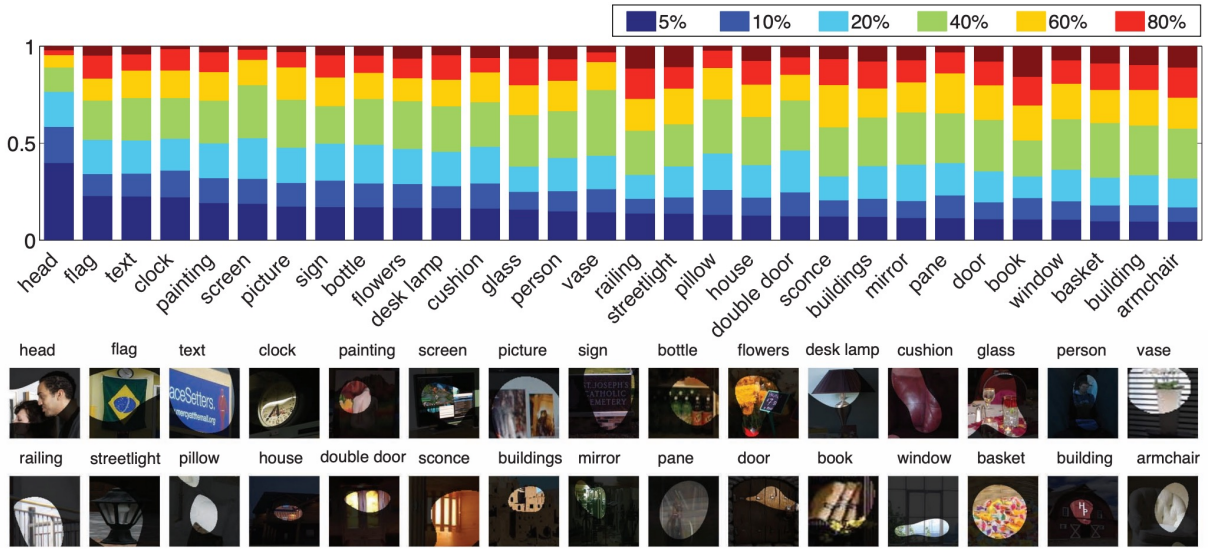


Figure 14. **Object saliency statistics**. For each image, we threshold the saliency map to create a binary map that selects N% of the image. We then compute the average overlap between the thresholded saliency map and a binary mask for a particular object category. For example, if there is a head in an image, on average the head area will have around 0.45 overlap with the top 5% most salient image area. The bottom rows show example objects cropped from our images, with a binary mask representing the top 5% most salient region.

Le, Ya, and Xuan Yang. "Tiny imagenet visual recognition challenge." Stanford University. CS231n. 2015.

Tiny ImageNet (TIN)

The TinyImageNet dataset is a subset of the ILSVRC-2012 classification dataset. It consists of 200 object classes, and for each object class it provides 500 training images, 50 validation images, and 50 test images. All images have been downsampled to $64 \times 64 \times 3$ pixels. The training and validation sets are released with images and annotations, including both class labels and bounding boxes. But the main goal of this project is to predict the class label of each image without localizing the objects. The test set is released without labels or bounding boxes.