논문 미팅

Out of Distribution Detection (OOD) in 2018

- Enhancing the reliability of Out-Of-Distribution image detection in neural networks (2018, ICLR)
- Training confidence-calibrated classifiers for detecting Out-Of-Distribution samples (2018, ICLR)

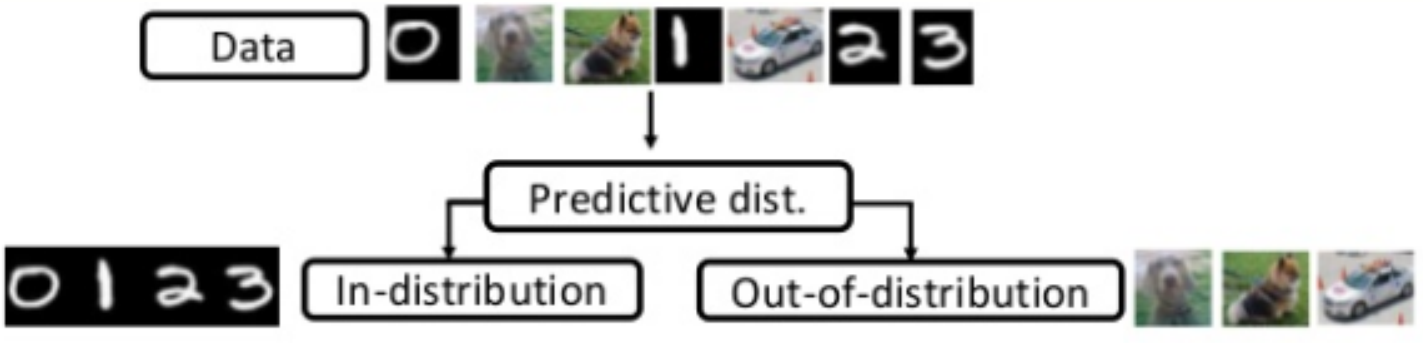
광주과학기술원, 인공지능 연구실

신 성 호

1 Out-of-Distribution Problem

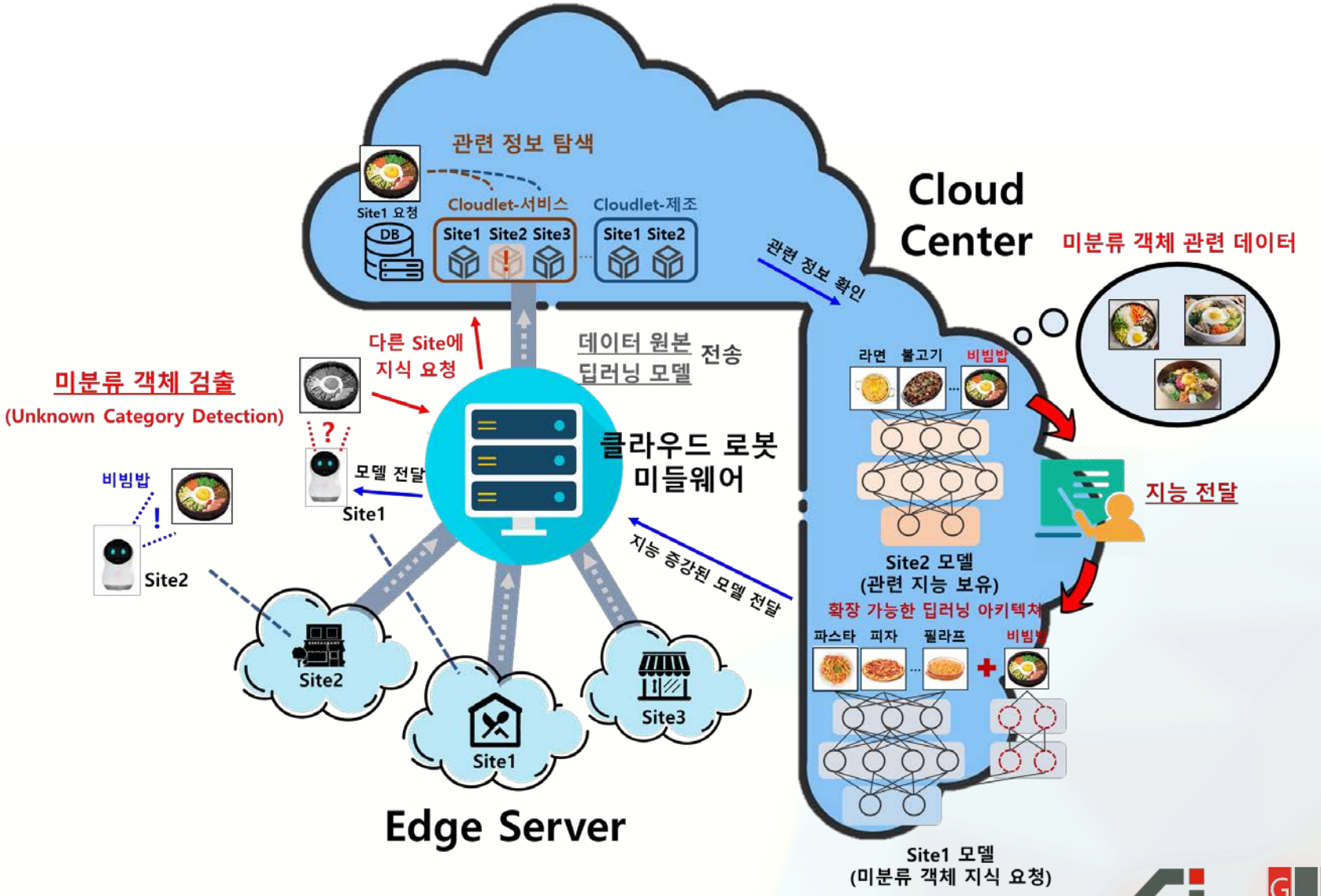
Out-of-Distribution Detection (OOD)

- Detecting the Out-of-Distribution samples, which are not used for training, from the inputs



- Usually, Deep Neural Network show overconfident result for the unknown input.
  - MNIST classifier produce high confident probability 91% even for random noise [1]
  - Even for the new kinds of input, softmax rshow distinguishable probability distribution [2]

1 Out-of-Distribution Problem



2 Out-of-Distribution Problem

- Enhancing the reliability of Out-Of-Distribution image detection in neural networks

- Out-of-Distribution Detection without further re-training networks
  - Well trained network tends to assign higher softmax scores to in-distribution sample than out-of-distribution examples

  - **(Contribution)** To separate between in- and out-of-distribution samples …
    - Temperature scaling in the softmax function [3]
    - Adding small controlled perturbations to inputs [4]

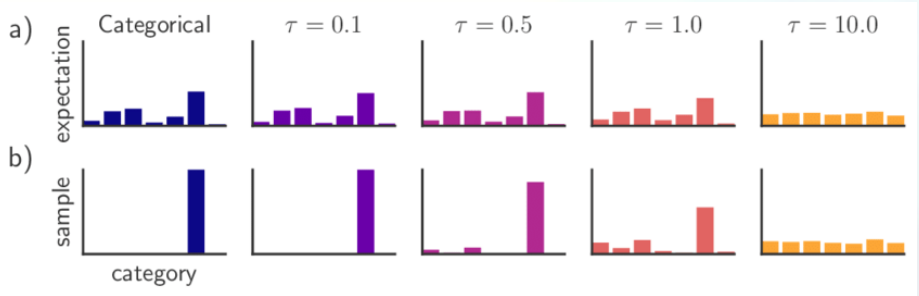
2 Out-of-Distribution Problem

- Enhancing the reliability of Out-Of-Distribution image detection in neural networks
  - **Temperature scaling** in the softmax function
    - For smoothing the probability distribution for each class
    - Out-of-distribution samples affects much more than the in-distribution samples

Temperature Scaling. Assume that the neural network $\boldsymbol{f} = (f_1, ..., f_N)$ is trained to classify N classes. For each input $\boldsymbol{x}$, the neural network assigns a label $\hat{y}(\boldsymbol{x}) = \arg\max_i S_i(\boldsymbol{x}; T)$ by computing the softmax output for each class. Specifically,

$$S_i(\boldsymbol{x}; T) = \frac{\exp\left(f_i(\boldsymbol{x})/T\right)}{\sum_{j=1}^{N} \exp\left(f_j(\boldsymbol{x})/T\right)}, \tag{1}$$

where $T \in \mathbb{R}^+$ is the temperature scaling parameter and set to 1 during the training. For a given input $\boldsymbol{x}$, we call the maximum softmax probability, i.e., $S_{\hat{y}}(\boldsymbol{x}; T) = \max_i S_i(\boldsymbol{x}; T)$ the **softmax score**. In this paper, we use notations $S_{\hat{y}}(\boldsymbol{x}; T)$ and $S(\boldsymbol{x}; T)$ interchangeably. Prior works have established the use of temperature scaling to distill the knowledge in neural networks (Hinton et al., 2015) and



2 Out-of-Distribution Problem

- Enhancing the reliability of Out-Of-Distribution image detection in neural networks

- **Adding small controlled perturbations** to inputs
  - Aim to increase the softmax score of any given input, without the need for a class label
  - Perturbations have stronger effect on the in-distribution images than out- samples

Input Preprocessing. Before feeding the image $\boldsymbol{x}$ into the neural network, we preprocess the input by adding small perturbations to it. The preprocessed image is given by

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} - \varepsilon \text{sign}(-\nabla_{\boldsymbol{x}} \log S_{\hat{y}}(\boldsymbol{x}; T)), \quad (2)$$

➔ **Direction of decreasing maximum softmax probability**

- **Out-of-Distribution Detector**
  - The OOD detection is based on threshold, manually set.

$$g(\boldsymbol{x}; \delta, T, \varepsilon) = \begin{cases} 1 & \text{if } \max_i p(\tilde{\boldsymbol{x}}; T) \leq \delta, \\ 0 & \text{if } \max_i p(\tilde{\boldsymbol{x}}; T) > \delta. \end{cases}$$

The parameters T, ε and δ are chosen so that the true positive rate (i.e., the fraction of in-distribution images correctly classified as in-distribution images) under some out-of-distribution image data set is 95%. (The choice of the out-of-distribution images to tune the parameters T, ε and δ appears to

2 Out-of-Distribution Problem

- Enhancing the reliability of Out-Of-Distribution image detection in neural networks

- **Dataset**

  (External Data) – Cropped or Down sampled

  - TinyImageNet : a subset of ImageNet, containing 10,000 test images from 200 different classes
  - LSUN : Large-scale Scene Understanding dataset, containing 10,000 images for 10 different scenes.
  - iSUN : a subset of SUN images, including 8925 images.

  (Synthetic Data)

  - Gaussian Noise : Synthetic Gaussian noise dataset consists of 10,000 random 2D Gaussian noise images
  - Uniform Noise : Synthetic uniform noise dataset consists of 10,000 images where each RGB value of every pixel is independently sampled from a uniform distribution on [0; 1].

2 Out-of-Distribution Problem

- Enhancing the reliability of Out-Of-Distribution image detection in neural networks

▪ Experimental Result

	Out-of-distribution dataset	FPR (95% TPR) ↓	Detection Error ↓	AUROC ↑	AUPR In ↑	AUPR Out ↑
		Baseline (Hendrycks & Gimpel, 2017) / ODIN				
Dense-BC CIFAR-10	TinyImageNet (crop)	34.7/4.3	19.9/4.7	95.3/99.1	96.4/99.1	93.8/99.1
	TinyImageNet (resize)	40.8/7.5	22.9/6.3	94.1/98.5	95.1/98.6	92.4/98.5
	LSUN (crop)	39.3/8.7	22.2/6.9	94.8/98.2	96.0/98.5	93.1/97.8
	LSUN (resize)	33.6/3.8	19.3/4.4	95.4/99.2	96.4/99.3	94.0/99.2
	iSUN	37.2/6.3	21.1/5.7	94.8/98.8	95.9/98.9	93.1/98.8
	Uniform	23.5/0.0	14.3/2.5	96.5/99.9	97.8/100.0	93.0/99.9
	Gaussian	12.3/0.0	8.7/2.5	97.5/100.0	98.3/100.0	95.9/100.0
Dense-BC CIFAR-100	TinyImageNet (crop)	67.8/17.3	36.4/11.2	83.0/97.1	85.3/97.4	80.8/96.8
	TinyImageNet (resize)	82.2/44.3	43.6/24.6	70.4/90.7	71.4/91.4	68.6/90.1
	LSUN (crop)	69.4/17.6	37.2/11.3	83.7/96.8	86.2/97.1	80.9/96.5
	LSUN (resize)	83.3/44.0	44.1/24.5	70.6/91.5	72.5/92.4	68.0/90.6
	iSUN	84.8/49.5	44.7/27.2	69.9/90.1	71.9/91.1	67.0/88.9
	Uniform	88.3/0.5	46.6/2.8	83.2/99.5	88.1/99.6	73.1/99.0
	Gaussian	95.4/0.2	50.2/2.6	81.8/99.6	87.6/99.7	70.1/99.1
WRN-28-10 CIFAR-10	TinyImageNet (crop)	38.9/23.4	21.9/14.2	92.9/94.2	92.5/92.8	91.9/94.7
	TinyImageNet (resize)	45.6/25.5	25.3/15.2	91.0/92.1	89.7/89.0	89.9/93.6
	LSUN (crop)	35.0/21.8	20.0/13.4	94.5/95.9	95.1/95.8	93.1/95.5
	LSUN (resize)	35.0/17.6	20.0/11.3	93.9/95.4	93.8/93.8	92.8/96.1
	iSUN	40.6/21.3	22.8/13.2	92.5/93.7	91.7/91.2	91.5/94.9
	Uniform	1.6/0.0	3.3/2.5	99.2/100.0	99.3/100.0	98.9/100.0
	Gaussian	0.3/0.0	2.6/2.5	99.5/100.0	99.6/100.0	99.3/100.0
WRN-28-10 CIFAR-100	TinyImageNet (crop)	66.6/43.9	35.8/24.4	82.0/90.8	83.3/91.4	80.2/90.0
	TinyImageNet (resize)	79.2/55.9	42.1/30.4	72.2/84.0	70.4/82.8	70.8/84.4
	LSUN (crop)	74.0/39.6	39.5/22.3	80.3/92.0	83.4/92.4	77.0/91.6
	LSUN (resize)	82.2/56.5	43.6/30.8	73.9/86.0	75.7/86.2	70.1/84.9
	iSUN	82.7/57.3	43.9/31.1	72.8/85.6	74.2/85.9	69.2/84.8
	Uniform	98.2/0.1	51.6/2.5	84.1/99.1	89.9/99.4	71.0/97.5
	Gaussian	99.2/1.0	52.1/3.0	84.3/98.5	90.2/99.1	70.9/95.9

Table 2: Distinguishing in- and out-of-distribution test set data for image classification. All values are percentages. ↑ indicates larger value is better, and ↓ indicates lower value is better. All parameter settings are shown in Appendix A.2. Additional results on WRN-40-4 and MNIST dataset are reported in Appendix A.1.

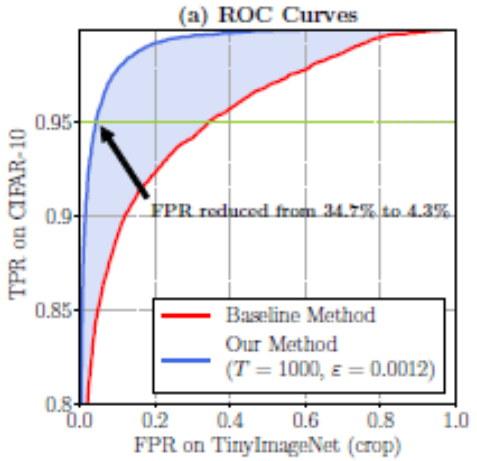


Figure 1: (a) ROC curves of baseline (red) and our method (blue) on DenseNet-BC-100 network, where CIFAR-10 and TinyImageNet (crop) are in- and out-of-distribution dataset, respectively.

2 Out-of-Distribution Problem

- Enhancing the reliability of Out-Of-Distribution image detection in neural networks

  - **Experimental Result**

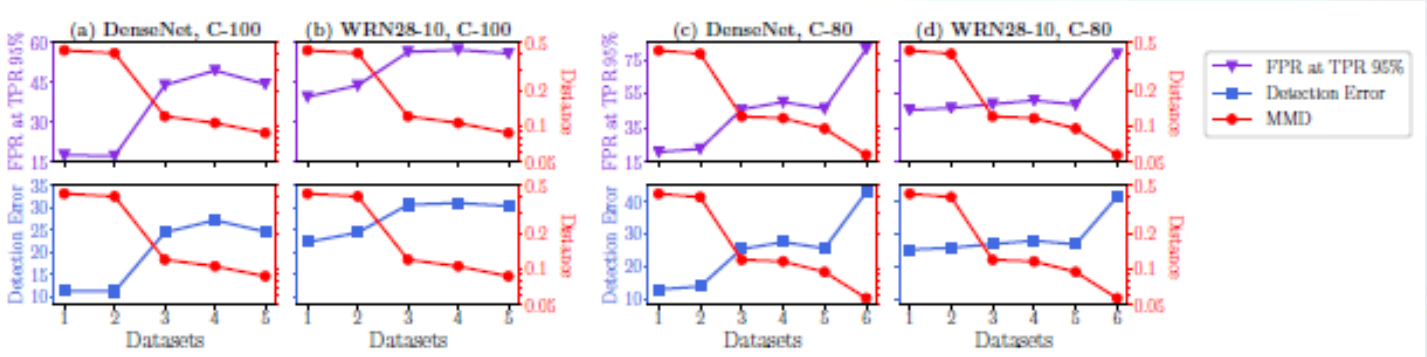
    - **MMD** : To measure the statistical distance between in- and out-of-distribution datasets

2016). Specifically, given two image sets, $V = \{v_1, ..., v_m\}$ and $W = \{w_1, ..., w_m\}$, the maximum mean discrepancy between V and Q is defined as

$$\widehat{\mathrm{MMD}}^2(V, W) = \frac{1}{\binom{m}{2}} \sum_{i \neq j} k(v_i, v_j) + \frac{1}{\binom{m}{2}} \sum_{i \neq j} k(w_i, w_j) - \frac{2}{\binom{m}{2}} \sum_{i \neq j} k(v_i, w_j),$$

where $k(\cdot, \cdot)$ is the Gaussian RBF kernel, i.e., $k(x, x') = \exp\left(-\frac{\|x - x'\|_2^2}{2\sigma^2}\right)$. We use the same method used by Sutherland et al. (2016) to choose σ, where $2\sigma^2$ is set to the median of all Euclidean distances between all images in the aggregate set $V \cup W$.

    - Negative relationship between MMD and the performances of proposed method



3 Out-of-Distribution Problem

- Training confidence-calibrated classifiers for detecting Out-Of-Distribution samples

- Developed new training scheme for OOD detection
- Before threshold-based method depend on the performances of pre-trained classifier

- **(Contribution)** To separate between in- and out-of-distribution samples …
  - Developed new-loss : confidence loss
  - Generates the OOD samples without using any Out-of-Distribution inputs

3 Out-of-Distribution Problem

- Training confidence-calibrated classifiers for detecting Out-Of-Distribution samples

  - **Developed new-loss : confidence loss**

    - Minimize the KL divergence from the predictive distribution on out-of-distribution samples to the uniform in order to give less confident predictions

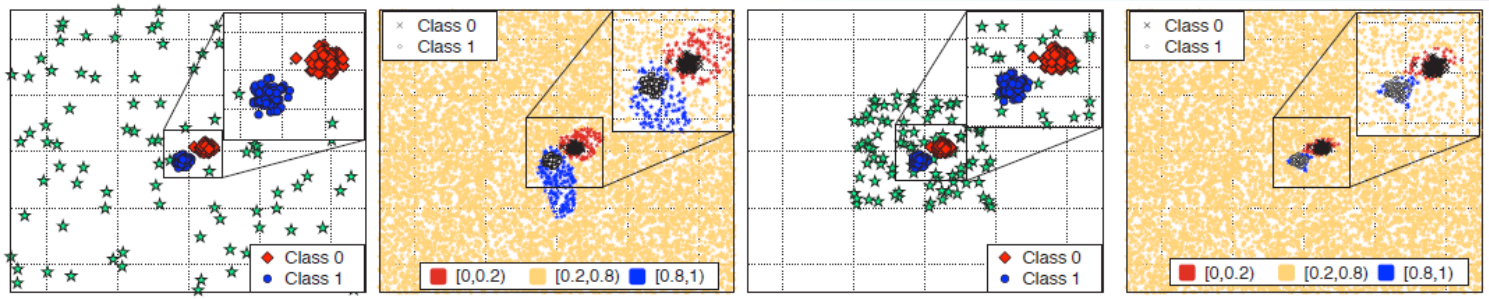
Without loss of generality, suppose that the cross entropy loss is used for training. Then, we propose the following new loss function, termed confidence loss:

$$\min_{\theta} \mathbb{E}_{P_{\text{in}}(\widehat{\mathbf{x}},\widehat{y})}\left[-\log P_\theta\left(y=\widehat{y}|\widehat{\mathbf{x}}\right)\right]+\beta\mathbb{E}_{P_{\text{out}}(\mathbf{x})}\left[KL\left(\mathcal{U}\left(y\right)\parallel P_\theta\left(y|\mathbf{x}\right)\right)\right], \quad (1)$$

where KL denotes the Kullback-Leibler (KL) divergence, $\mathcal{U}\left(y\right)$ is the uniform distribution and $\beta > 0$ is a penalty parameter. It is highly intuitive as the new loss forces the predictive distribution

3 Out-of-Distribution Problem

- Training confidence-calibrated classifiers for detecting Out-Of-Distribution samples
  - **Generates the OOD samples without using any Out-of-Distribution inputs**
    - A prioi knowledge on out-of-distribution is not available or its underlying space is too huge to cover
      - Generate the most effective samples
      - Generate the 'boundary' samples in the low-density area of in-distribution samples
    - Intuition
      - Effect of boundary of in-distribution region might propagate to the entire out-of-distribution space



3 Out-of-Distribution Problem

- Training confidence-calibrated classifiers for detecting Out-Of-Distribution samples
  - **Generates the OOD samples without using any Out-of-Distribution inputs**
    - GAN loss for generating the OOD samples in the boundary of in-distribution samples

However, unlike the original GAN, we want to make the generator recover an effective out-of-distribution P_{out} instead of P_{in}. To this end, we propose the following new GAN loss:

$$\min_G \max_D \underbrace{\beta \mathbb{E}_{P_G(\mathbf{x})}\left[KL\left(\mathcal{U}\left(y\right) \parallel P_\theta\left(y|\mathbf{x}\right)\right)\right]}_{\text{(a)}} + \underbrace{\mathbb{E}_{P_{\text{in}}(\mathbf{x})}\left[\log D\left(\mathbf{x}\right)\right] + \mathbb{E}_{P_G(\mathbf{x})}\left[\log\left(1 - D\left(\mathbf{x}\right)\right)\right]}_{\text{(b)}}, \quad (3)$$

where θ is the model parameter of a classifier trained on in-distribution. The above objective can be

- (a) : Forces the generator to generate low-density samples since it can be interpreted as minimizing the log negative likelihood of in-distribution (push)
- (b) : Forces having samples being not too far from the in-distribution samples (pull)

3 Out-of-Distribution Problem

- Enhancing the reliability of Out-Of-Distribution image detection in neural networks

▪ Experimental Result

In-dist	Out-of-dist	Classification accuracy	TNR at TPR 95%	AUROC	Detection accuracy	AUPR in	AUPR out
		Cross entropy loss / Confidence loss					
SVHN	CIFAR-10 (seen)	93.82 / **94.23**	47.4 / **99.9**	62.6 / **99.9**	78.6 / **99.9**	71.6 / **99.9**	91.2 / **99.4**
	TinyImageNet (unseen)		49.0 / **100.0**	64.6 / **100.0**	79.6 / **100.0**	72.7 / **100.0**	91.6 / **99.4**
	LSUN (unseen)		46.3 / **100.0**	61.8 / **100.0**	78.2 / **100.0**	71.1 / **100.0**	90.8 / **99.4**
	Gaussian (unseen)		56.1 / **100.0**	72.0 / **100.0**	83.4 / **100.0**	77.2 / **100.0**	92.8 / **99.4**
CIFAR-10	SVHN (seen)	80.14 / **80.56**	13.7 / **99.8**	46.6 / **99.9**	66.6 / **99.8**	61.4 / **99.9**	73.5 / **99.8**
	TinyImageNet (unseen)		**13.6** / 9.9	**39.6** / 31.8	**62.6** / 58.6	**58.3** / 55.3	**71.0** / 66.1
	LSUN (unseen)		**14.0** / 10.5	**40.7** / 34.8	**63.2** / 60.2	**58.7** / 56.4	**71.5** / 68.0
	Gaussian (unseen)		2.8 / **3.3**	10.2 / **14.1**	50.0 / 50.0	48.1 / **49.4**	39.9 / **47.0**

Table 1: Performance of the baseline detector (Hendrycks & Gimpel, 2016) using VGGNet. All values are percentages and boldface values indicate relative the better results. For each in-distribution, we minimize the KL divergence term in (1) using training samples from an out-of-distribution dataset denoted by "seen", where other "unseen" out-of-distributions were only used for testing.

▪ Effects of confidence loss