TRAINING CONFIDENCE-CALIBRATED CLASSIFIERS FOR DETECTING OUT-OF-DISTRIBUTION SAMPLES

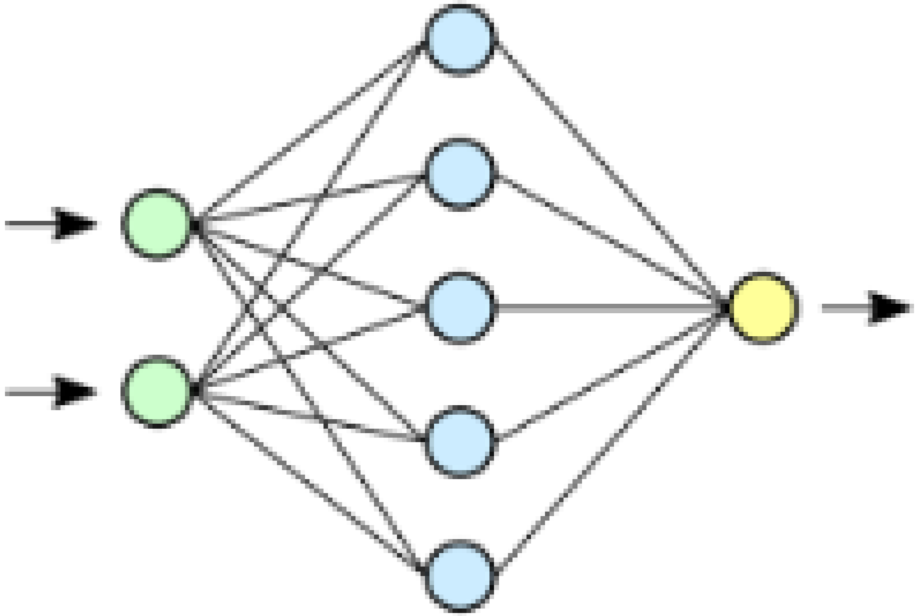
Kimin Lee* **Honglak Lee§,†** **Kibok Lee†** **Jinwoo Shin***
*Korea Advanced Institute of Science and Technology, Daejeon, Korea
†University of Michigan, Ann Arbor, MI 48109
§Google Brain, Mountain View, CA 94043

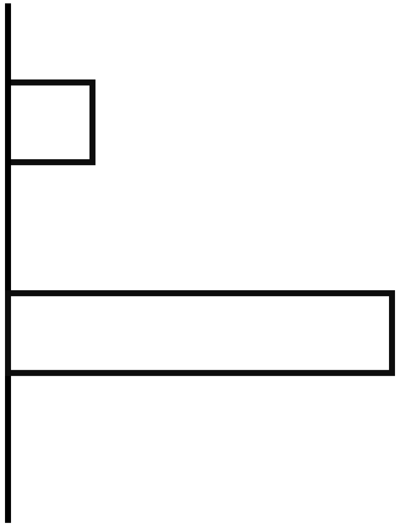
2023.03.24

TAEWON KIM

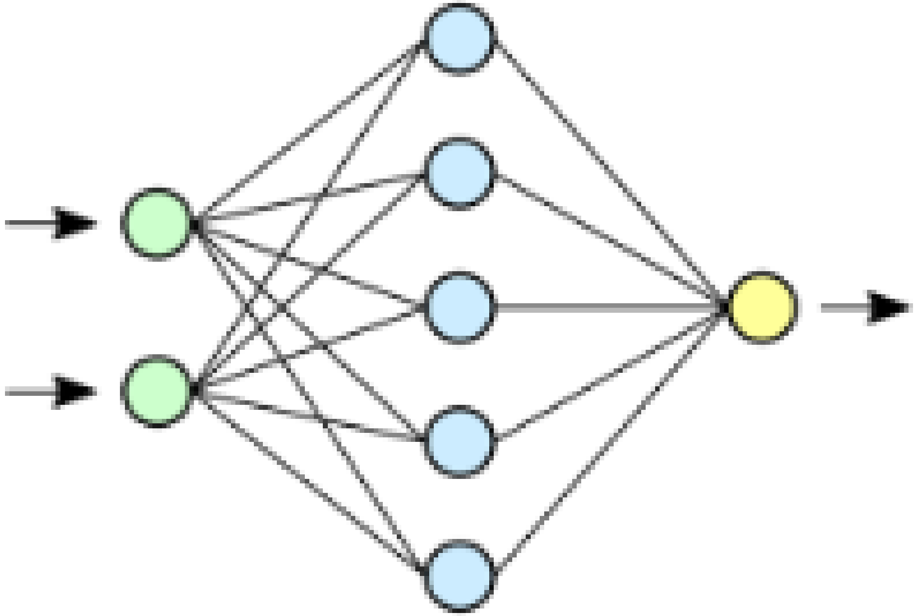
0. Contents

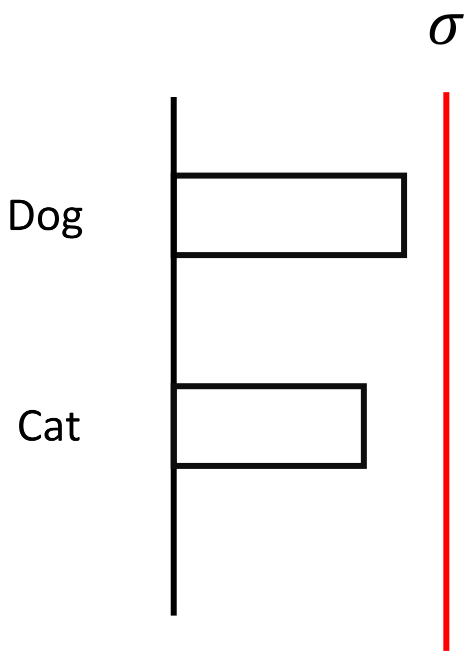
1. Detecting Out of distribution Detector



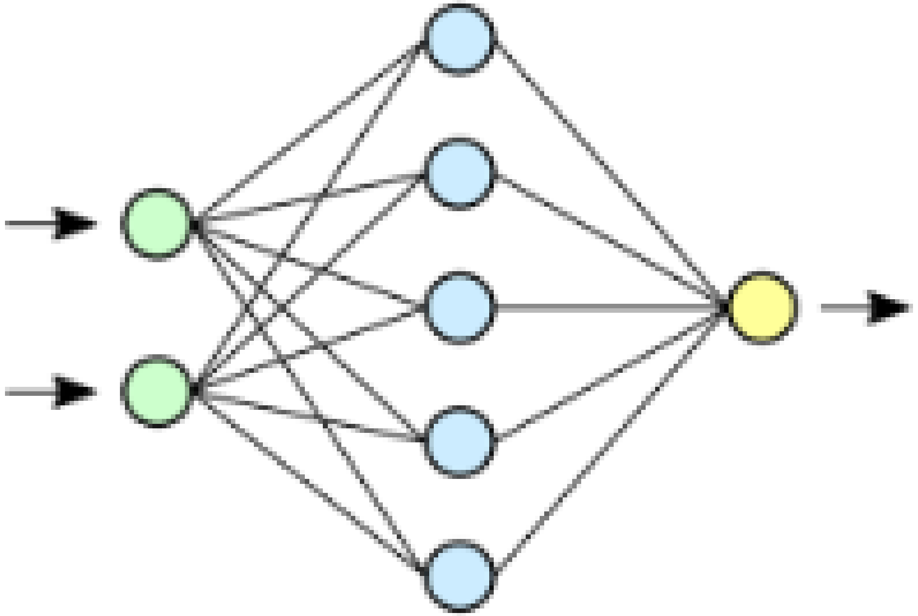


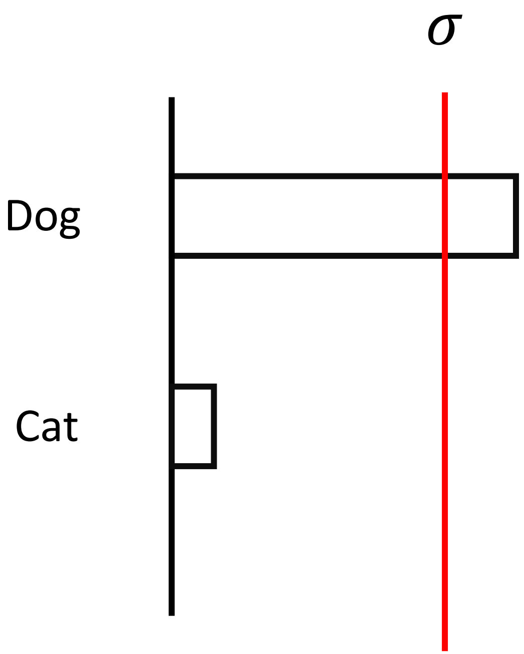
1. Detecting Out of distribution Detector





1. Detecting Out of distribution Detector

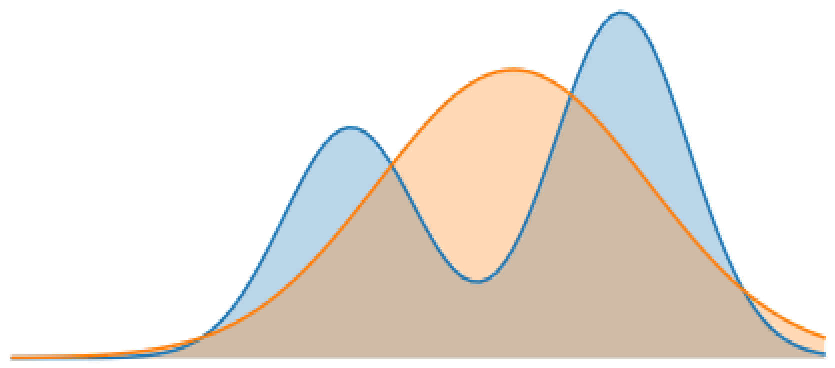




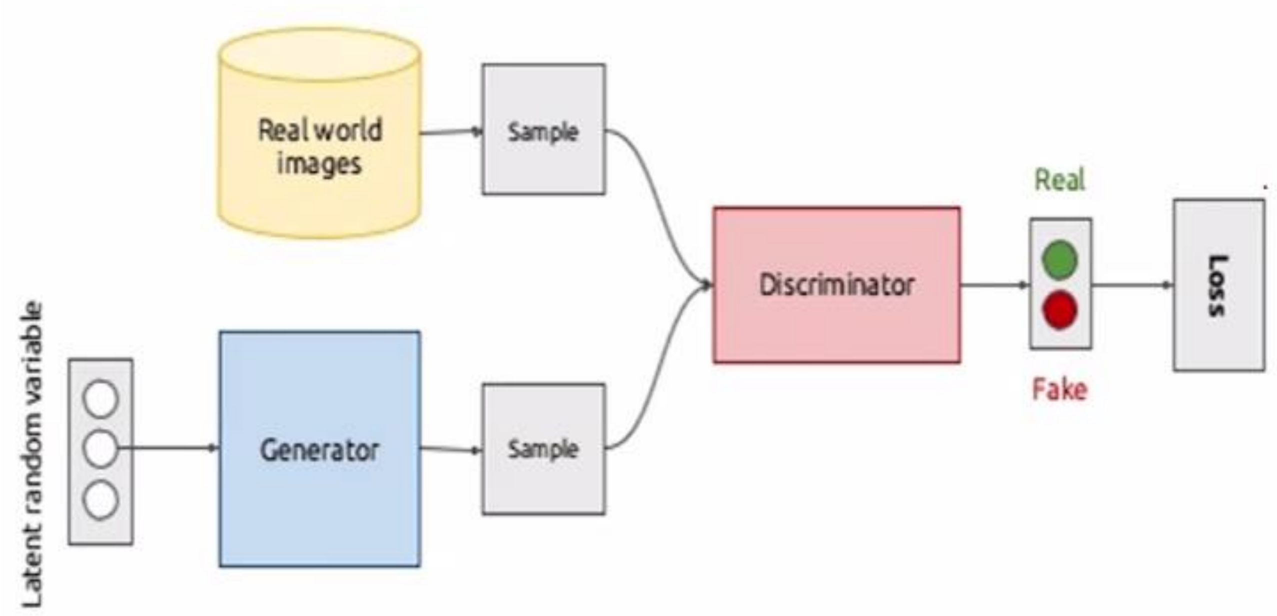
- 이전 Out of Distribution Network 는 pretrain된 모델을 기반으로 동작
- inference 단계만 사용하기 때문에 training 결과에 많이 의존

2. CONTRIBUTION

(KL) divergence



generative adversarial network (GAN)

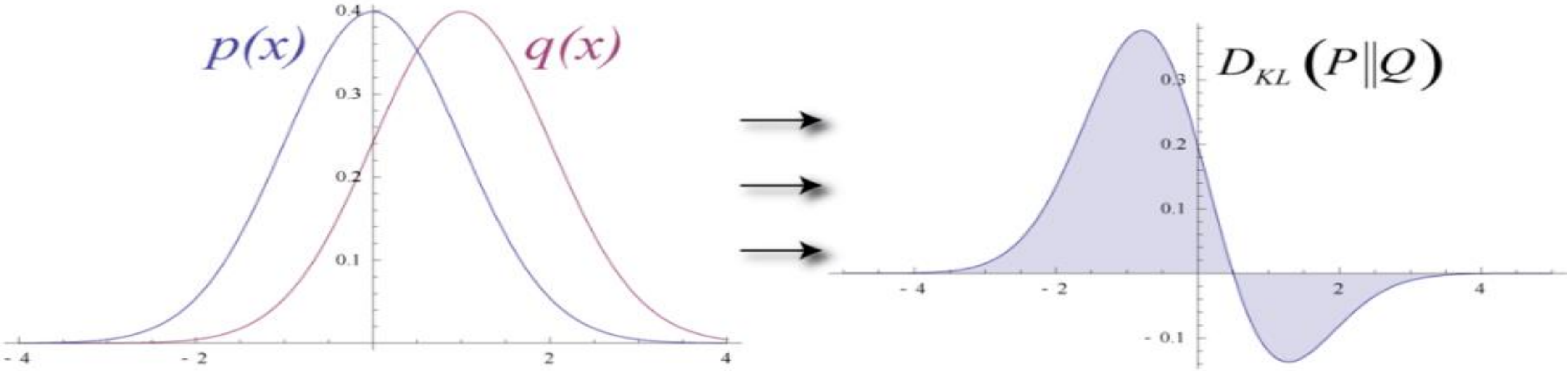


- KL-Div를 이용한 새로운 loss function을 제시
- GAN을 이용해 학습에 가장 효과적인 out-of-distribution Data를 생성

2. CONTRIBUTION

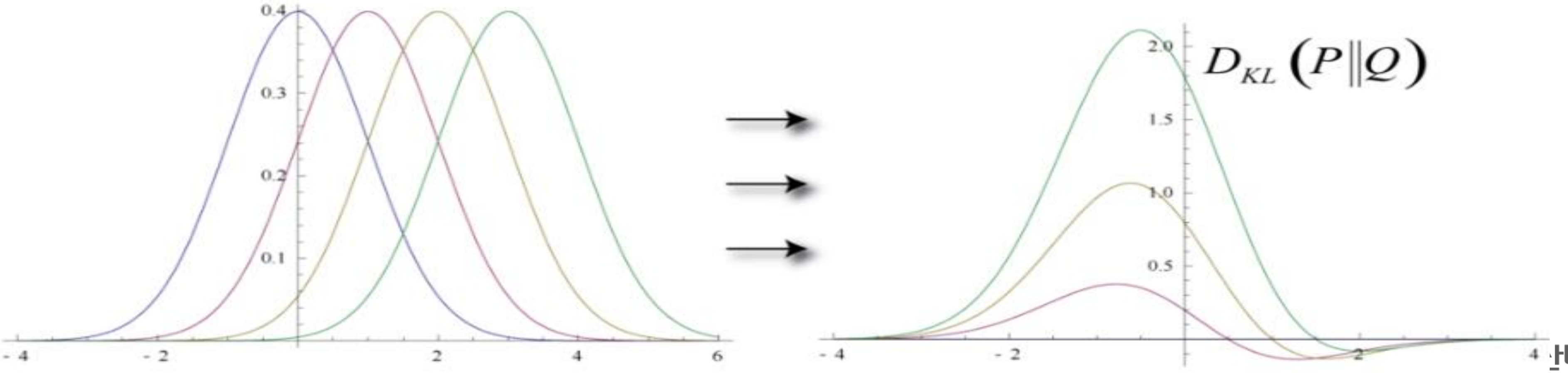
(KL) divergence

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$



Original Gaussian PDF's

KL Area to be Integrated



2. CONFIDENCE LOSS

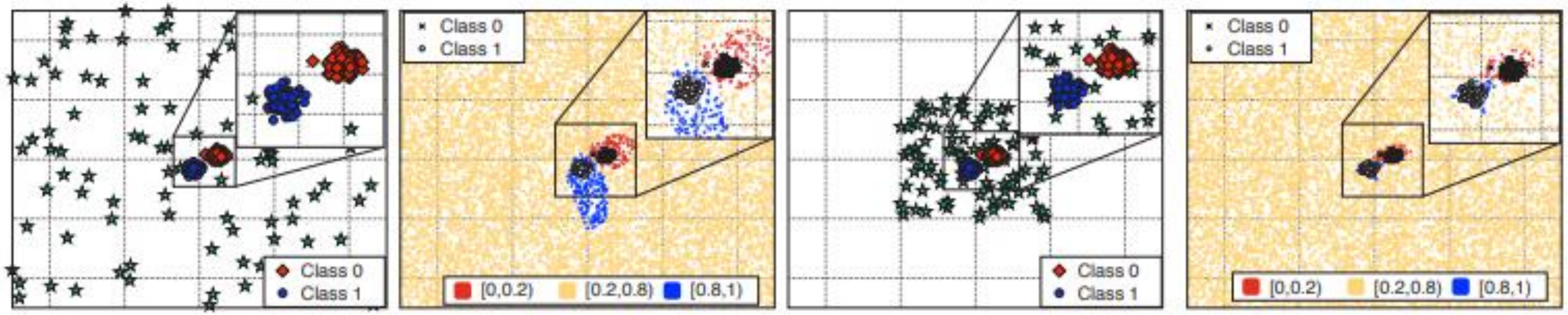
$$\min_{\theta} \boxed{\mathbb{E}_{P_{\text{in}}(\hat{\mathbf{x}},\hat{y})}\left[-\log P_{\theta}\left(y=\hat{y}|\hat{\mathbf{x}}\right)\right]} + \boxed{\beta\mathbb{E}_{P_{\text{out}}(\mathbf{x})}\left[KL\left(\mathcal{U}\left(y\right) \parallel P_{\theta}\left(y|\mathbf{x}\right)\right)\right]}$$

Cross Entropy loss

KL-Div. Term for Out of-dist. data

- Cross Entropy loss 같은 경우는 In-distribution data가 올바른 클래스를 예측하는 데 얼마나 잘 수행되는지
- KL-Div. Term 은 out-of-distribution data가 uniform한 분포를 가지도록 유도하는 과정
- 결론적으로는 In-dist 와 out-of-dist data가 모두 필요

2. CONFIDENCE LOSS



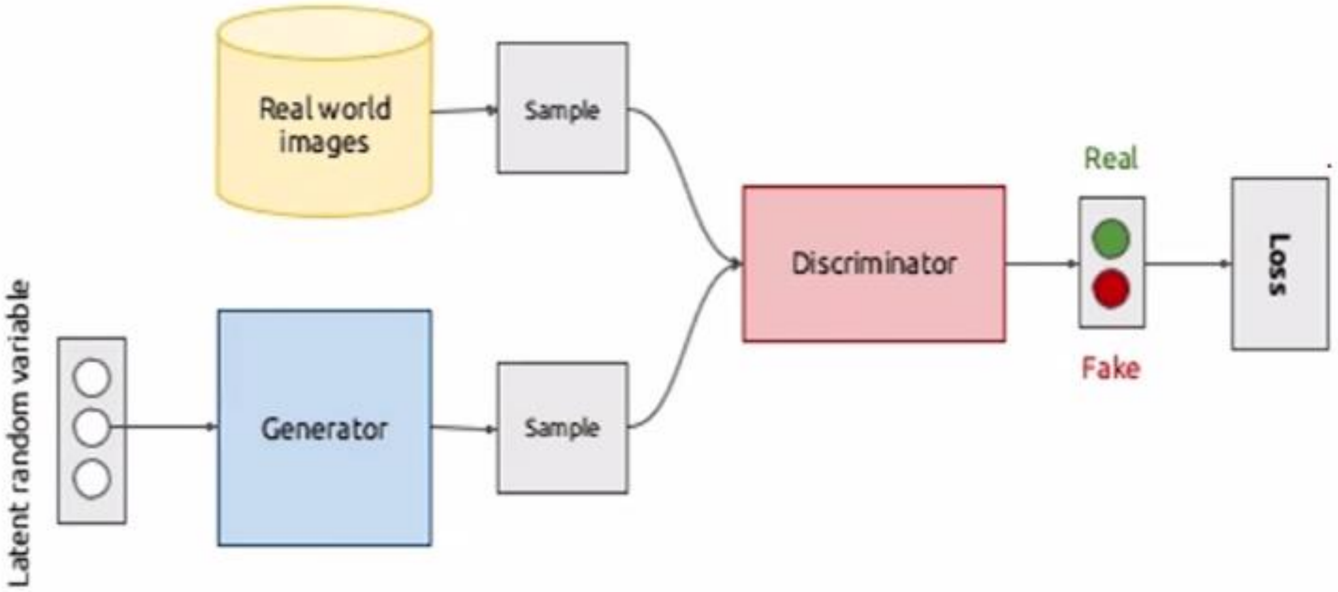
(a) (b) (c) (d)

Figure 1: Illustrating the behavior of classifier under different out-of-distribution training datasets. We generate the out-of-distribution samples from (a) 2D box $[-50, 50]^2$, and show (b) the corresponding decision boundary of classifier. We also generate the out-of-distribution samples from (c) 2D box $[-20, 20]^2$, and show (d) the corresponding decision boundary of classifier.

- (a) Out-of-distribution에 대한 전체 데이터 분포를 구성
- (b) CONFIDENCE CLASSIFIERS 을 이용해 학습한 데이터 분포
- (c) In-distribution 경계에 있는 Out-of-distribution 데이터 분포를 구성

3. ADVERSARIAL GENERATOR

GAN



Sample x from real data distribution

Sample latent code z from Gaussian distribution

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))]$$

D should maximize $V(D,G)$

Maximum when $D(x) = 1$

Maximum when $D(G(z)) = 0$

Modified GAN

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

- (a) data가 out of distribution 과 비슷하게 만드는 Generator loss
- (b) in-distribution data 대해 올바르게 분류하고 생성자가 생성한 가짜 데이터가 실제와 구분 안되게 학습

3. JOINT OBJECT FUNCTION

JOINT TRAINING METHOD OF CONFIDENT CLASSIFIER AND ADVERSARIAL GENERATOR

Cross Entropy loss

KL-Div. Term for Out of-dist. data

$$\min_G \max_D \min_\theta \underbrace{\mathbb{E}_{P_{\text{in}}(\hat{\mathbf{x}},\hat{y})}\left[-\log P_\theta\left(y=\hat{y}|\hat{\mathbf{x}}\right)\right]}_{(c)} + \beta \underbrace{\mathbb{E}_{P_G(\mathbf{x})}\left[KL\left(\mathcal{U}\left(y\right) \parallel P_\theta\left(y|\mathbf{x}\right)\right)\right]}_{(d)}$$

$$\underbrace{+\mathbb{E}_{P_{\text{in}}(\hat{\mathbf{x}})}\left[\log D\left(\hat{\mathbf{x}}\right)\right] + \mathbb{E}_{P_G(\mathbf{x})}\left[\log\left(1-D\left(\mathbf{x}\right)\right)\right]}_{(e)}.$$

Modified GAN objective function

3. EXPERIMENTS

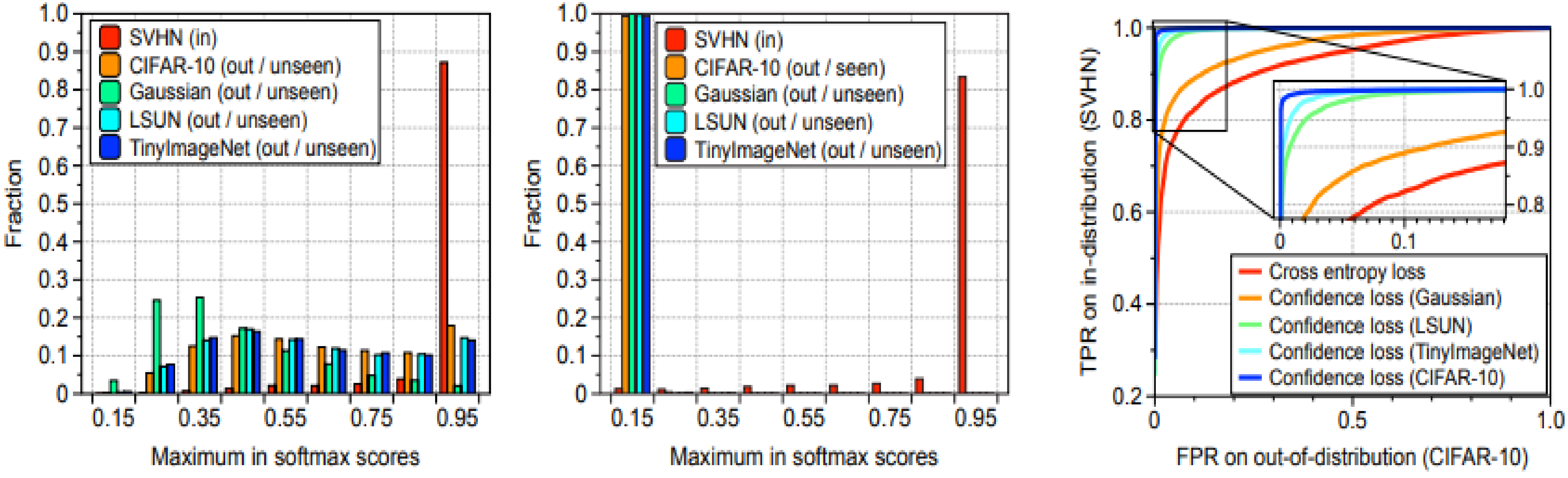
MI2L

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.

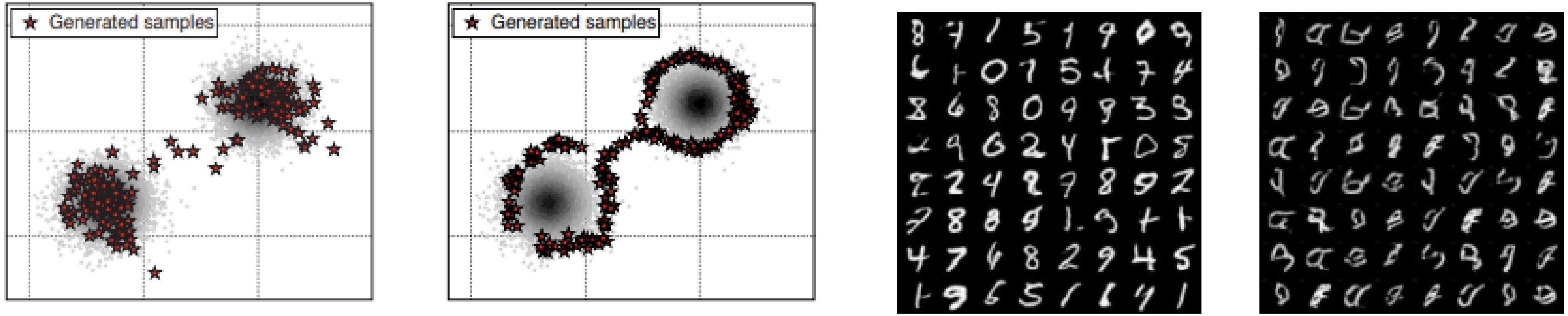
울산대학교 UNIVERSITY OF ULSAN 서울아산병원 Asan Medical Center

3. EXPERIMENTS



(a) Cross entropy loss (b) Confidence loss in (1) (c) ROC curve

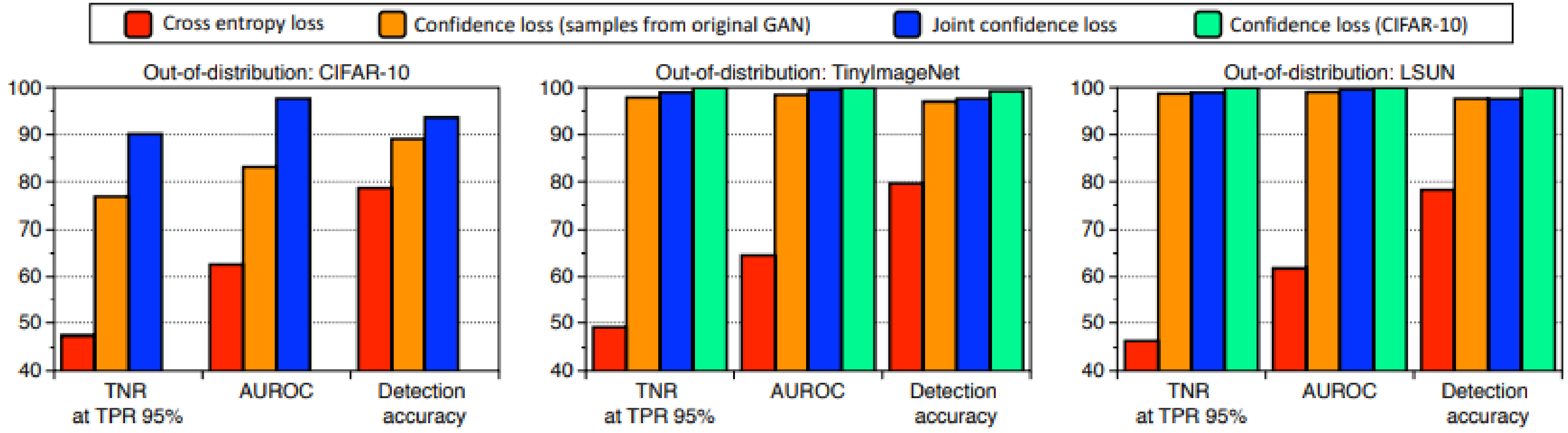
3. EXPERIMENTS



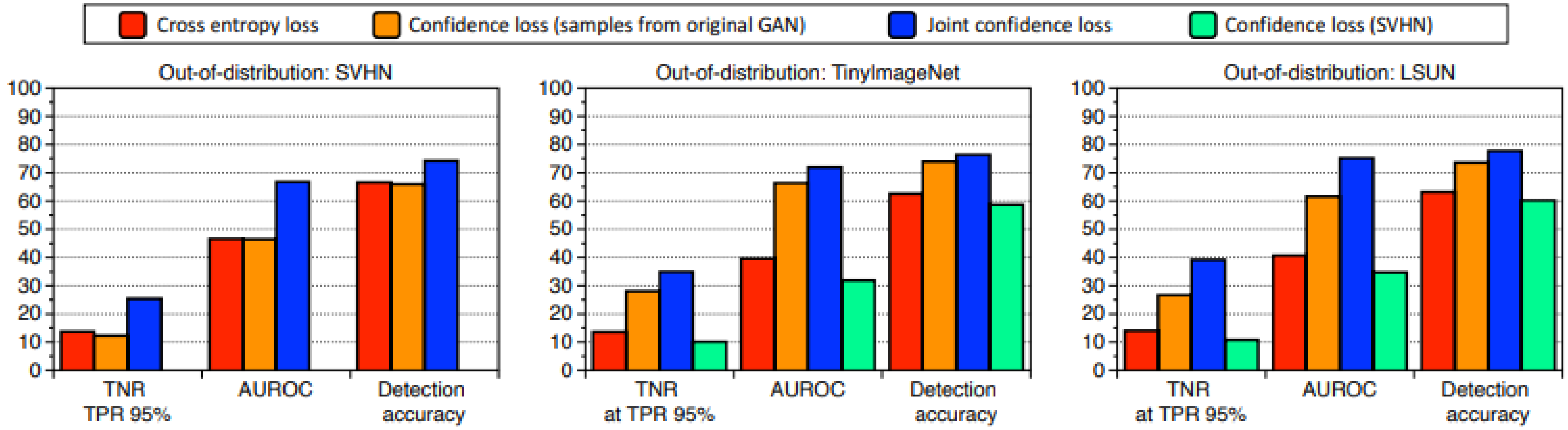
(a) (b) (c) (d)

Figure 3: The generated samples from original GAN (a)/(c) and proposed GAN (b)/(d). In (a)/(b), the grey area is the 2D histogram of training in-distribution samples drawn from a mixture of two Gaussian distributions and red points indicate generated samples by GANs.

3. EXPERIMENTS



(a) In-distribution: SVHN



(b) In-distribution: CIFAR-10