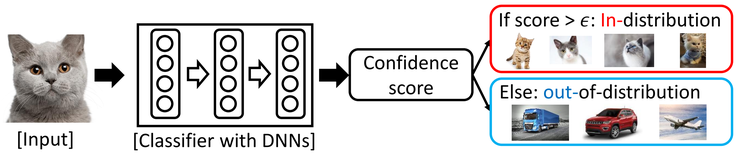
Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples

Kimin Lee[1], Honglak Lee[2,3], Kibok Lee[2], Jinwoo Shin[1]
[1]Korea Advanced Institute of Science and Technology (KAIST)
[2]University of Michigan [3]Google Brain

Problem: Detecting Out-of-Distribution

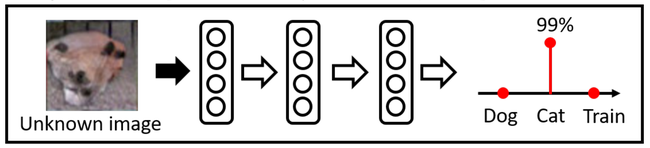


- Detect whether a test sample is from in-distribution P_{in} (i.e., training distribution) or out-of-distribution P_{out}.

Related work:

- Threshold-based detector [Hendrycks et al., 2016] defined a confidence score as the maximum value of prediction and classifies it as in-distribution if the confidence score is above some threshold.
- ODIN [Liang et al., 2018] further improved the performance using temperature scaling and input pre-processing.

Overconfidence issue: Deep neural networks (DNNs) are typically overconfident in their predictions [Lakshminarayanan et al., 2017]:



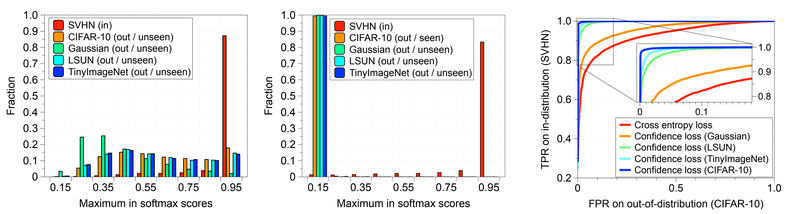
Contribution 1: Training Confidence-calibrated Classifier

Training method for detecting out-of-distribution.

- Confidence loss: additionally minimizing the KL divergence.

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

Experimental results on simple CNNs (2 Conv + 3 FC): some explicit out-of-distribution samples (denoted by "seen") are given in training time.



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

Figure 1: (a)/(b) The x-axis and y-axis represent the maximum prediction value and the fraction of images receiving the corresponding score, respectively. (c) Receiver operating characteristic (ROC) curve plots true positive rate (TPR) against false negative rate (FPR).

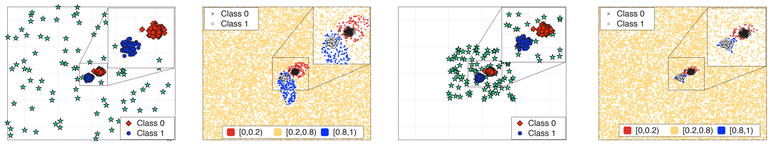
- In case of confidence loss (Figure 1(b)), there exists a better separation between in- and out-of- distributions.
- Optimizing the cross entropy has a higher FPR than other ones doing the confidence loss to have a same TPR.

Contribution 2: GAN for Out-of-Distribution

Issue: the number of out-of-distribution training samples might be almost infinite to cover the entire space.

Intuition: out-of-distribution samples close to in-distribution are more effective in improving the detection performance.

Experiments on binary classification task:



(a) Out samples from entire space (b) Decision boundary in the case of (a) (c) Out samples close to in-distribution (d) Decision boundary in the case of (c)

Figure 2: (a)/(b) Classifier becomes overconfident when out samples are from entire space. (c)/(d) It shows confidence-calibrated predictions if out samples are close to in-distribution.

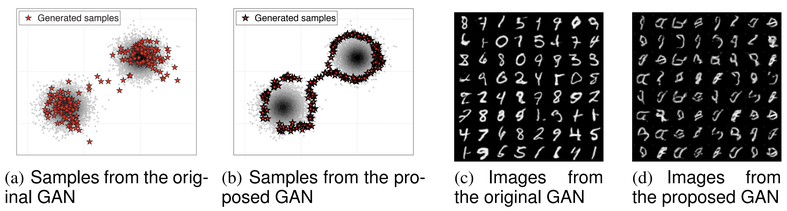
- Training out-of-distribution samples nearby in-distribution could be more effective in improving detection performances.

New generative adversarial network (GAN): generating most effective samples from out-of-distribution.

$$\min_G \max_D \underbrace{\mathbb{E}_{P_G(\mathbf{x})}\left[KL\left(\mathcal{U}\left(y\right)\parallel P_\theta\left(y|\mathbf{x}\right)\right)\right]}_{\text{(a) Forcing } G \text{ to produce low-density samples}}$$

$$\underbrace{+\mathbb{E}_{P_{\text{in}}(\mathbf{x})}\left[\log D\left(\mathbf{x}\right)\right]+\mathbb{E}_{P_G(\mathbf{x})}\left[\log\left(1-D\left(\mathbf{x}\right)\right)\right]}_{\text{(b) Original GAN loss for generating samples close to in-distribution}}.$$

Comparison of original GAN and proposed GAN:



(a) Samples from the original GAN (b) Samples from the proposed GAN (c) Images from the original GAN (d) Images from the proposed GAN

Figure 3: In-distribution samples in (a)/(b) and (c)/(d) are drawn from a mixture of two Gaussian distributions and MNIST dataset, respectively.

- For our method, we use a pre-trained classifier.
- The proposed GAN can produce the samples nearby the low-density boundary of the in-distribution space.

Contribution 3: Joint Training Method

Jointly optimizing confidence-calibrated classifier and GAN as follows:

$$\underbrace{\mathbb{E}_{P_{\text{in}}(\mathbf{x},\widehat{y})}\left[-\log P_\theta\left(y=\widehat{y}|\widehat{\mathbf{x}}\right)\right]}_{\text{(c) Standard classification loss}}\underbrace{+\beta\mathbb{E}_{P_G(\mathbf{x})}\left[KL\left(\mathcal{U}\left(y\right)\parallel P_\theta\left(y|\mathbf{x}\right)\right)\right]}_{\text{(d) KL divergence term for confident prediction}}$$

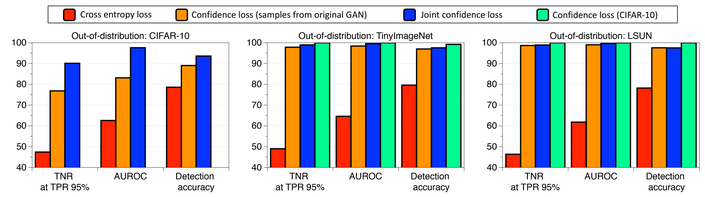
$$\underbrace{+\mathbb{E}_{P_{\text{in}}(\widehat{\mathbf{x}})}\left[\log D\left(\widehat{\mathbf{x}}\right)\right]+\mathbb{E}_{P_G(\mathbf{x})}\left[\log\left(1-D\left(\mathbf{x}\right)\right)\right]}_{\text{(e) Original GAN loss}}.$$

- Confidence loss: (c) + (d) / Proposed GAN loss: (d) + (e)

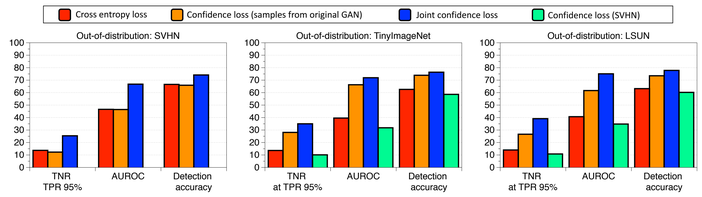
Experiments: Effects of Joint Confidence Loss

Detection performance of threshold-based detector using VGGNet:

- True negative rate (TNR) at 95% true positive rate (TPR).
- Area under receiver operating characteristic curve (AUROC): TPR vs FPR.
- Detection accuracy: best classification accuracy over all thresholds.



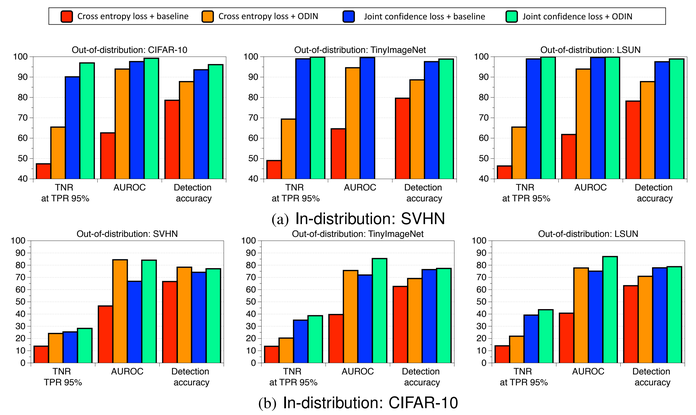
(a) In-distribution: SVHN



(b) In-distribution: CIFAR-10

Figure 4: Performances of the baseline detector under various training losses.

- Our method outperforms all baseline methods in all cases.



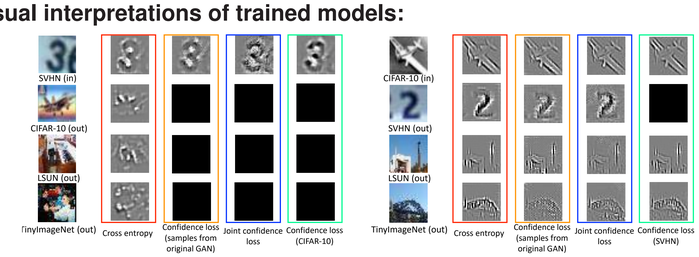
(a) In-distribution: SVHN

(b) In-distribution: CIFAR-10

Figure 5: Performances of the baseline detector and ODIN detector [Liang et al. 2018].

- Our method can be utilized with ODIN [Liang et al., 2018].

Visual interpretations of trained models:



(a) In-distribution: SVHN (b) In-distribution: CIFAR-10

Figure 6: Guided gradient maps of predicted class with respect to the input image.