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, USA
[§]Google Brain, Mountain View, CA 94043

Abstract

The problem of detecting whether a test sample is from in-distribution (i.e., training distribution by a classifier) or out-of-distribution sufficiently different from it arises in many real-world machine learning applications. However, the state-of-art deep neural networks are known to be highly overconfident in their predictions, i.e., do not distinguish in- and out-of-distributions. Recently, to handle this issue, several threshold-based detectors have been proposed given pre-trained neural classifiers. However, the performance of prior works highly depends on how to train the classifiers since they only focus on improving inference procedures. In this paper, we develop a novel training method for classifiers so that such inference algorithms can work better. In particular, we suggest two additional terms added to the original loss (e.g., cross entropy). The first one forces samples from out-of-distribution less confident by the classifier and the second one is for (implicitly) generating most effective training samples for the first one. In essence, our method jointly trains both classification and generative neural networks for out-of-distribution. We demonstrate its effectiveness on various popular image datasets.

1 Introduction

Even though deep neural networks (DNNs) achieve high classification accuracy, it has been addressed [9, 6] that they are typically overconfident in their predictions. Since evaluating the quality of their predictive uncertainty is hard, deploying them in real-world systems raises serious concerns in AI Safety [1]. This overconfidence issue is highly related to the problem of detecting out-of-distribution: detect whether a test sample is from in-distribution (i.e., training distribution by DNNs) or out-of-distribution sufficiently different from it. Formally, it can be formulated as a binary classification problem. Let an input $\mathbf{x} \in \mathcal{X}$ and a label $y \in \mathcal{Y} = \{1, \dots, K\}$ be random variables that follow a joint data distribution $P_{\texttt{in}}(\mathbf{x}, y) = P_{\texttt{in}}(y|\mathbf{x}) P_{\texttt{in}}(\mathbf{x})$. We assume that a classifier $P_\theta(y|\mathbf{x})$ is trained on a dataset drawn from $P_{\texttt{in}}(\mathbf{x}, y)$, where θ denotes the model parameter. We let $P_{\texttt{out}}(\mathbf{x})$ denote an out-of-distribution which is 'far away' from in-distribution. Our problem of interest is determining if input $\mathbf{x}$ is from $P_{\texttt{in}}$ or $P_{\texttt{out}}$, possibly utilizing a well calibrated classifier $P_\theta(y|\mathbf{x})$. In other words, we aim to build a detector which assigns label 1 if data is from in-distribution, and label 0 otherwise.

There have been recent efforts toward developing efficient detection methods where they mostly have studied simple threshold-based detectors utilizing a pre-trained classifier [7, 11]. For each input $\mathbf{x}$, it measures some confidence score $q(\mathbf{x})$ based on a pre-trained classifier, and compares the score to some threshold $\delta > 0$. Then, the detector assigns label 1 if the confidence score $q(\mathbf{x})$ is above δ, and label 0, otherwise. Specifically, the authors in [7] defined the confidence score as a maximum value of the predictive distribution, and the authors in [11] further improved the performance by using temperature scaling [6] and adding small perturbations to the input data. Although such inference methods are computationally simple, their performances highly depend on

Second workshop on Bayesian Deep Learning (NIPS 2017), Long Beach, CA, USA.

the pre-trained classifier. Namely, they fail to work if the classifier does not separate the maximum value of predictive distribution well enough with respect to $P_{\texttt{in}}$ and $P_{\texttt{out}}$. Ideally, a classifier should be trained to separate all class-dependent in-distributions as well as out-of-distribution in the output space. As another line of research, Bayesian probabilistic models [10, 12] and ensembles [9] were also investigated. However, training or inferring those models are computationally more expensive. This motivates our approach of developing a new training method for the more plausible simple classifiers. Our direction is orthogonal to the Bayesian and ensemble approaches, where one can also combine them for even better performance.

Contribution. In this paper, we develop such a training method for detecting out-of-distribution $P_{\texttt{out}}$ better without losing its original classification accuracy. First, we consider a new loss function, called *confidence loss*. Our key idea on the proposed loss is to additionally minimize the Kullback-Leibler (KL) divergence from the predictive distribution on out-of-distribution samples to the uniform one in order to give less confident predictions on them. Then, in- and out-of-distributions are expected to be more separable. However, optimizing the confidence loss requires training samples from out-of-distribution, which are often hard to sample: a priori knowledge on out-of-distribution is not available or its underlying space is too huge to cover. To handle the issue, we consider a new generative adversarial network (GAN) [5] for generating most effective samples from $P_{\texttt{out}}$. Unlike the original GAN, the proposed GAN generates 'boundary' samples in the low-density area of $P_{\texttt{in}}$. Finally, we design a joint training scheme minimizing the classifier's loss and new GAN loss alternatively, i.e., the confident classifier improves the GAN, and vice versa, as training proceeds. Here, we emphasize that the proposed GAN does not need to generate explicit samples under our scheme, and instead it implicitly encourages training a more confident classifier. We demonstrate the effectiveness of the proposed method using deep convolutional neural networks including VGGNet [15] for image classification tasks on CIFAR [8], SVHN [13], ImageNet [4], and LSUN [16] datasets.

2 Training confident neural classifiers

Confident classifier for out-of-distribution. We propose a new loss function to train a classifier which can map the samples from in- and out-of-distributions into the output space separately. Without loss of generality, suppose that the cross entropy loss is used for training. Then, we define the following, termed confidence loss:

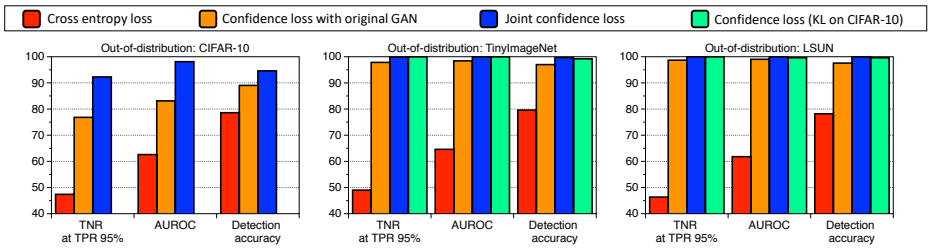
$$\min_{\theta} \ \mathbb{E}_{P_{\texttt{in}}(\widehat{\mathbf{x}},\widehat{y})}\big[-\log P_\theta\left(y=\widehat{y}|\widehat{\mathbf{x}}\right)\big] + \beta\mathbb{E}_{P_{\texttt{out}}(\mathbf{x})}\big[KL\left(\mathcal{U}\left(y\right) \parallel P_\theta\left(y|\mathbf{x}\right)\right)\big], \tag{1}$$

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 on out-of-distribution samples to be closer to the uniform one, i.e., zero confidence, while that for samples from in-distribution still follows the label-dependent probability.

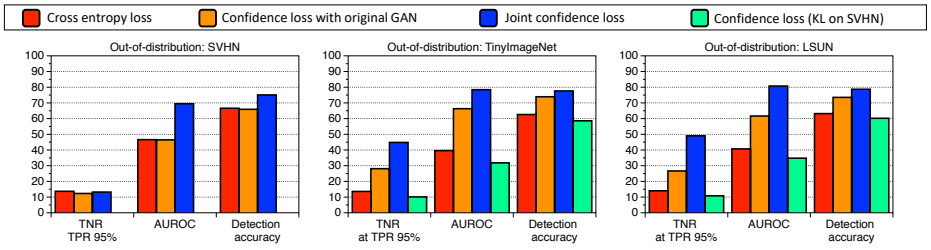
Joint training method of confident classifier and adversarial generator. However, optimizing the confidence loss requires training samples from out-of-distribution, which are often hard to sample. To handle this issue, we propose a joint training scheme for the confident classifier and the proposed GAN which generates 'boundary' samples in the low-density area of in-distribution, i.e., close to out-of-distribution. We suggest the following joint objective function:

$$\begin{aligned}\min_{G}\max_{D}\min_{\theta} \ & \underbrace{\mathbb{E}_{P_{\texttt{in}}(\widehat{\mathbf{x}},\widehat{y})}\big[-\log P_\theta\left(y=\widehat{y}|\widehat{\mathbf{x}}\right)\big]}_{\text{(a)}} \underbrace{+\beta\mathbb{E}_{P_{\texttt{pri}}(\mathbf{z})}\big[KL\left(\mathcal{U}\left(y\right) \parallel P_\theta\left(y|G\left(\mathbf{z}\right)\right)\right)\big]}_{\text{(b)}} \\ & \underbrace{+\mathbb{E}_{P_{\texttt{in}}(\widehat{\mathbf{x}})}\big[\log D\left(\widehat{\mathbf{x}}\right)\big] + \mathbb{E}_{P_{\texttt{pri}}(\mathbf{z})}\big[\log\left(1-D\left(G\left(\mathbf{z}\right)\right)\right)\big] - \mathcal{H}\left(P_G\left(\mathbf{x}\right)\right),}_{\text{(c)}}\end{aligned} \tag{2}$$

where $\mathcal{H}\left(\cdot\right)$ denotes the entropy, D is discriminator that represents a probability that sample $\mathbf{x}$ is from the in-distribution and G is the generator that maps a latent variable $\mathbf{z}$ from a prior distribution $P_{\texttt{pri}}\left(\mathbf{z}\right)$ to generated outputs $G\left(\mathbf{z}\right)$. The classifier's confidence loss corresponds to (a) + (b), and the proposed GAN loss corresponds to (b) + (c). By minimizing the KL divergence term (b) and original GAN loss (c) where $\mathcal{H}$ is added to discourage the generator from collapsing, the proposed GAN loss encourage the generator to produce the samples which are on the low-density boundary of the in-distribution space. To optimize the above objective efficiently, we propose an alternating algorithm, which optimizes model parameters $\{\theta\}$ of classifier and GAN models $\{G, D\}$ alternatively.



(a) In-distribution: SVHN



(b) In-distribution: CIFAR-10

Figure 1: Performances of the baseline detector under various training losses.

3 Experimental Results

We demonstrate the effectiveness of our method using various datasets: CIFAR [8], SVHN [13], ImageNet [4], and LSUN [16]. We train VGGNet [15] for classifying CIFAR-10 and SVHN datasets. The corresponding test dataset is used as the in-distribution (positive) samples to measure the performance. We use realistic images as the out-of-distribution (negative) samples. For evaluation, we use baseline threshold-based detectors [7] that computes the maximum value of predictive distribution on a test sample and classifies it as positive (i.e., in-distribution) if the confidence score is above some threshold. Using this baseline detector, we measure the true negative rate (TNR) at 95% true positive rate (TPR), the area under the receiver operating characteristic curve (AUROC) and the detection accuracy, where larger values of all metrics indicate better detection performances. Due to the space limitation, more explanations and results are given in the final paper.

We evaluate the performance of joint confidence loss (2) utilizing the proposed GAN. To this end, we use VGGNets (as classifiers) and DCGANs [14] (as GANs). We also test a variant of confidence loss which optimizes the KL divergence term on samples from a pre-trained original GAN (implicitly) modeling the in-distribution. One can expect that samples from the original GAN can be also useful for improving the detection performance since it may have bad generalization properties [2] and generate a few samples on the low-density boundary as like the proposed GAN. Figure 1 shows the performance of the baseline detector for each in- and out-of-distribution pair. First, observe that the joint confidence loss (blue bar) outperforms the confidence loss (1) with some explicit out-of-distribution datasets (green bar). This is quite remarkable since the former is trained only using in-distribution datasets, while the latter utilizes additional out-of-distribution datasets. We also remark that our methods significantly outperform the baseline cross entropy loss (red bar) in all cases without harming its original classification performances. Interestingly, the confidence loss with the original GAN (orange bar) is often (but not always) useful for improving the detection performance, whereas that with the proposed GAN (blue bar) still outperforms it in all cases.

4 Conclusion

In this paper, we aim to develop a training method for neural classification networks for detecting out-of-distribution better without losing its original classification accuracy. In essence, our method jointly trains two models for detecting and generating out-of-distribution by minimizing their losses alternatively. Although we primarily focus on image classification in our experiments, our method can be used for any classification tasks using deep neural networks. It is also interesting future directions applying our methods for other related tasks: network calibration [6], Bayesian probabilistic models [10, 12] and ensemble method [9] and semi-supervised learning [3].

Acknowledgements

This work was supported in part by the Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.2017-0-01778, Development of Explainable Human-level Deep Machine Learning Inference Framework).

References

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[2] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). In *International Conference on Machine Learning (ICML)*, 2017.

[3] Zihang Dai, Zhilin Yang, Fan Yang, William W Cohen, and Ruslan Salakhutdinov. Good semi-supervised learning that requires a bad gan. In *Advances in neural information processing systems (NIPS)*, 2017.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition (CVPR)*, 2009.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems (NIPS)*, 2014.

[6] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

[7] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2016.

[8] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[9] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems (NIPS)*, 2017.

[10] Yingzhen Li and Yarin Gal. Dropout inference in bayesian neural networks with alpha-divergences. In *International Conference on Machine Learning (ICML)*, 2017.

[11] Shiyu Liang, Yixuan Li, and R Srikant. Principled detection of out-of-distribution examples in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[12] Christos Louizos and Max Welling. Multiplicative normalizing flows for variational bayesian neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

[13] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 5, 2011.

[14] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2015.

[15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[16] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.