Boundary Optimised Samples Training for Detecting Out-of-Distribution Images

Luca Marson, Vladimir Li, Atsuto Maki

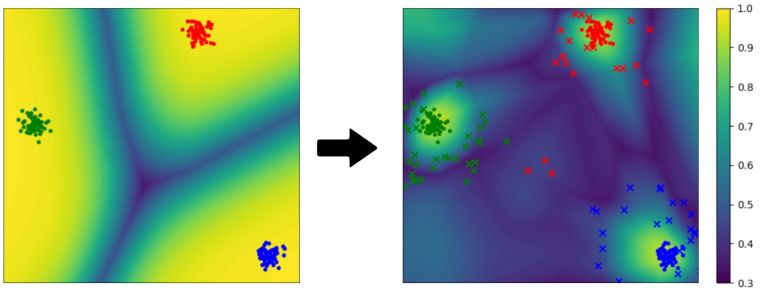
Division of Robotics, Perception, and Learning
KTH Royal Institute of Technology, Stockholm, Sweden

Introduction

Deep Convolutional Networks often overestimate their predictive ability where out-of-distribution (OOD) samples are given as input [1],[2], making their detection not trivial.

A classification model should be able to identify whether it is capable of correctly assessing the input for the decision, or need human intervention. This is especially important in safety critical systems, such as autonomous driving or medical diagnosis.

In this article, we propose the Boundary optimised Samples (BoS) training algorithm to generate OOD samples. These are exploited to enforce low confidence around them [3], resulting in a higher OOD samples detection accuracy.



(a) (b)

Confidence values of the output of a model trained for a toy 2-d three class problem. Left: by standard procedure. Right: by the proposed algorithm.

Method

Algorithm 1: Training algorithm

Initialization: pretraining of the classification network

1 $\min_{\boldsymbol{\theta}} \mathbb{E}\left[L_{CE}(\hat{\mathbf{x}}, \hat{y}, \boldsymbol{\theta})\right]$

2 **while** *not converged* **do**

Phase 1: boundary samples generation

3 initialize$(\mathbf{x}, y_t)$

4 $\min_{\mathbf{x}} \mathbb{E}\left[L_{CE}(\mathbf{x}, y_t, \boldsymbol{\theta}) + \beta L_{\overline{CE}}(\mathbf{x}, \boldsymbol{\theta}) + \lambda L_{TV}(\mathbf{x})\right]$

Phase 2: fine tune with boundary samples

5 $\min_{\boldsymbol{\theta}} \mathbb{E}_{\hat{\mathbf{x}}}\left[L_{CE}(\hat{\mathbf{x}}, \hat{y}, \boldsymbol{\theta})\right] + \gamma \mathbb{E}_{\mathbf{x}}\left[L_{KL}(\mathbf{x}, \boldsymbol{\theta})\right]$

6 **end**

To achieve this, we argue that it is effective to generate OOD examples near the classification boundary of the model. Our training algorithm consist of:

- Training a model with the cross-entropy loss (line 1) to ensure high classification accuracy.
- Generating boundary samples with BoS training (line 3,4), by back propagating to the input the gradient of the Boundary Loss.
- Exploiting them to reduce the confidence around the training data while preserving the model performance on the original classification task (line 5).

Experiments

Similarly to previous works [1],[3],[4],[5],[6], we evaluate the method by training a model on one dataset representing the in-distribution, specifically MNIST and CIFAR-10, and estimate its confidence on several unseen out-of-distribution datasets. Such model is then compared to a baseline plain model trained using only the cross-entropy loss.

We approached OOD detection as a binary classification problem [4]. According to the considered metrics, we were able to obtain significant improvements with respect to the baseline plain model trained on MNIST, and better results on some of the OOD datasets used for CIFAR-10.

Trained on **MNIST**	Plain (**98.93%**)			BoS(98.70%)		
	MMC	AUROC	DA	MMC	AUROC	DA
MNIST	0.966	-	-	**0.974**	-	-
FMNIST	0.837	0.825	0.749	**0.106**	**0.999**	**0.998**
EMNIST	0.803	0.833	0.769	**0.452**	**0.980**	**0.924**
Noise	0.911	0.867	0.850	**0.101**	**1.000**	**1.000**

Trained on **CIFAR-10**	Plain (**98.93%**)			BoS(98.70%)		
	MMC	AUROC	DA	MMC	AUROC	DA
CIFAR-10	0.906	-	-	**0.943**	-	-
CIFAR-100	**0.759**	0.764	0.708	0.761	**0.789**	**0.720**
SVHN	0.603	0.892	0.822	**0.284**	**0.974**	**0.908**
Noise	**0.717**	**0.823**	**0.759**	0.860	0.765	0.702

Conclusions

In sum, the main contributions of this paper are:

- A novel efficient method for generating boundary samples, BoS training,
- A robust algorithm for enforcing low confidence on OOD samples by the boundary optimised samples, and
- The experimental results supporting that the new method outperforms the baseline.

References

[1] M. Hein, M. Andriushchenko, and J. Bitterwolf, "Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem". Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 41–50, 2019.
[2] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan, "Do deep generative models know what they don't know?". 7th International Conference on Learning Representations, ICLR, pp. 1–19, 2019.
[3] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. "Training confidence-calibrated classifiers for detecting out-of-distribution samples". In:6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings(2018), pp. 1–16.
[4] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," 5th International Conference on Learning Representations, ICLR - Conference Track Proceedings, pp. 1–12, 2017.
[5] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out of-distribution image detection in neural networks," 6th International Conference on Learning Representations, ICLR - Conference Track Proceedings, 2018.
[6] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," 7th International Conference on Learning Representations, ICLR, pp. 1–18, 2019.