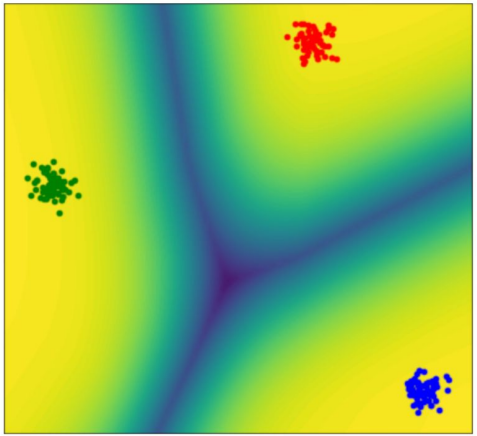
Boundary Optimised Samples Training for Detecting Out-of-Distribution Images

Luca Marson, Vladimir Li, Atsuto Maki

Division of Robotics, Perception, and Learning

KTH Royal Institute of Technology, Stockholm, Sweden

Out-of-distribution examples detection



Deep Neural Networks perform high confidence predictions on Out-of-Distributions (OOD) inputs [1].

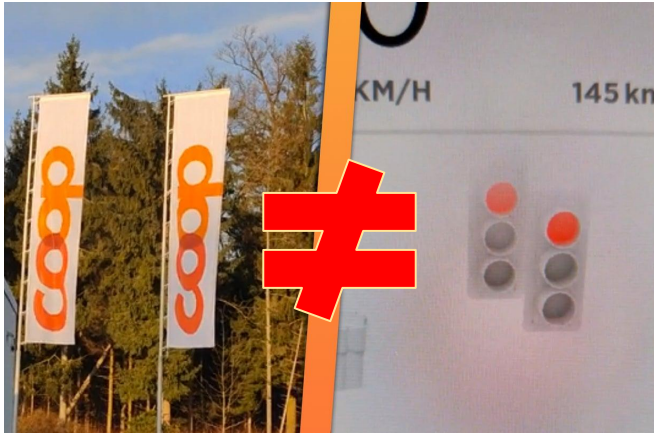
They are not able to identify whether they are capable of correctly assessing the input for the decision or need human intervention.

[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.

Out-of-distribution examples detection

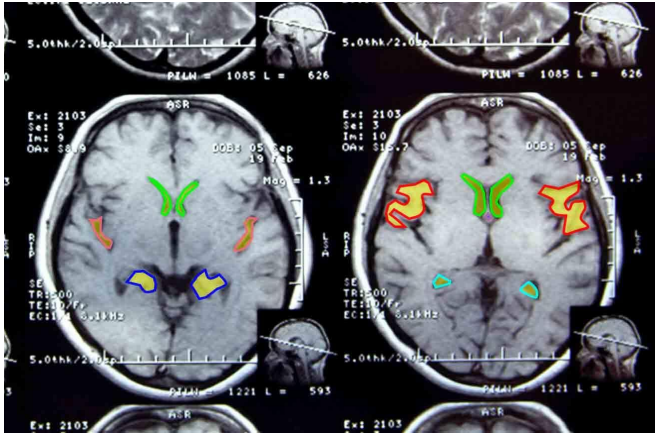
Problematic in safety-critical applications...

Autonomous vehicles



https://www.thedrive.com/tech/37250/tesla-autopilot-confuses-red-text-on-a-flag-for-stop-light

Medical diagnosis

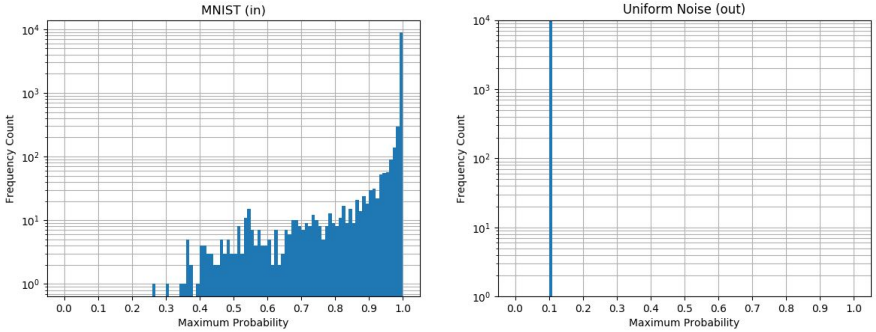


Threshold-based OOD detector

Binary classification problem - threshold on the confidence score [2]:

$$q(\mathbf{x}) = \max p(\mathbf{y} \mid \mathbf{x})$$

$$g(\mathbf{x}) = \begin{cases} 1 \implies \mathbf{x} \in \mathcal{D}_{\text{in}}, & \text{if } q(\mathbf{x}) > \delta \\ 0 \implies \mathbf{x} \in \mathcal{D}_{\text{out}}, & \text{otherwise} \end{cases}$$



Objective:

Increase separability of the <u>in-distribution</u> and <u>out-of-distribution</u> confidence score

[2] Dan Hendrycks and Kevin Gimpel. “A baseline for detecting misclassified and out-of-distribution examples in neural networks”. In:5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings(2019), pp. 1–12.

Threshold-based OOD detector

Confidence loss [3]

- enforce low confidence far away from the training data using OOD examples

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\hat{\mathbf{x}},\hat{y}\sim\mathcal{D}_{\text{in}}(\hat{\mathbf{x}},\hat{y})}\left[L_{CE}(\hat{\mathbf{x}},\hat{y},\boldsymbol{\theta})\right] + \gamma\,\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{out}}(\mathbf{x})}\left[L_{KL}(\mathbf{x},\boldsymbol{\theta})\right]$$

$$L_{CE}(\hat{\mathbf{x}},\hat{y},\boldsymbol{\theta}) = -\log p_{\boldsymbol{\theta}}(y=\hat{y} \mid \hat{\mathbf{x}})$$

$$L_{KL}(\mathbf{x},\boldsymbol{\theta}) = KL(\mathcal{U}(\mathbf{y}) \mid\mid p_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x}))$$

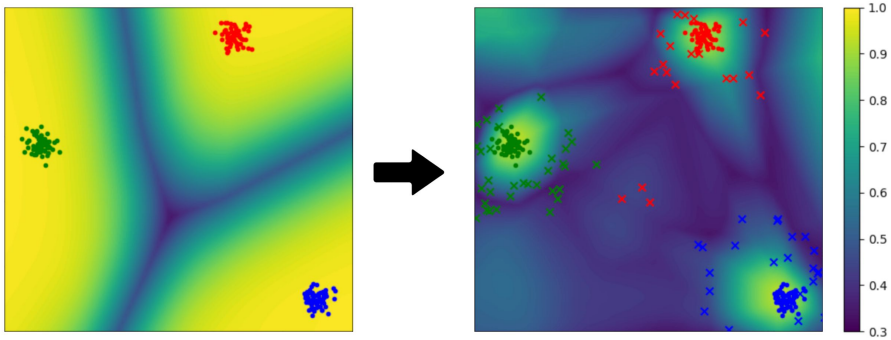
[3] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. “Training confidence-calibrated classifiers for detecting out-of-distribution samples”. In:6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings(2018), pp. 1–16.

Boundary optimized Samples (BoS) training

A procedure to enforce low confidence around the training data.

1. Generate boundary optimised samples

2. Enforce low confidence at OOD examples without affecting the original classification task



(a) (b)

Boundary Loss

Boundary Loss

Generate boundary samples by backpropagating the gradient to the input

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

- $L_{CE}(\mathbf{x}, y_t, \boldsymbol{\theta}) = -\log p_{\boldsymbol{\theta}}(y = y_t \mid \mathbf{x})$

  Classification cross-entropy

- $L_{\overline{CE}}(\mathbf{x}, \boldsymbol{\theta}) = -\frac{1}{K}\sum_{c=1}^{K} \log p_{\boldsymbol{\theta}}(y = c \mid \mathbf{x})$

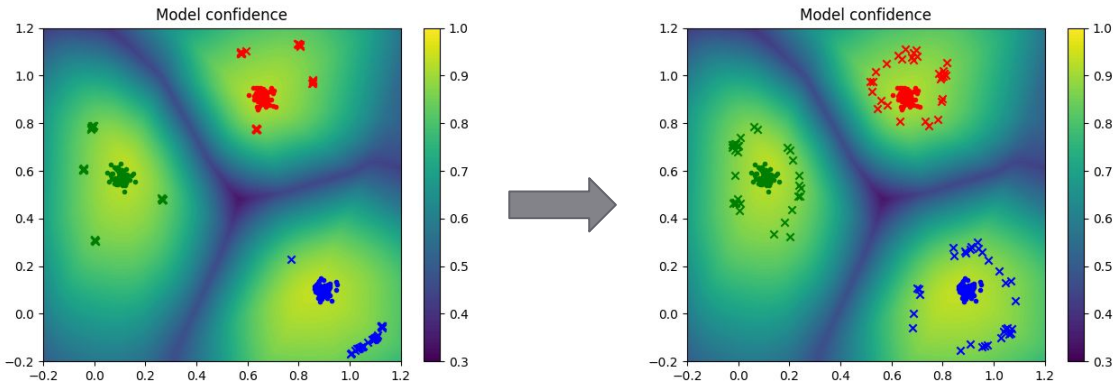
  Cross-entropy between the predicted distribution and the discrete uniform distribution $\mathcal{U}$

Diversity across boundary optimised samples

$L_{\overline{CE}}(\mathbf{x}, \boldsymbol{\theta})$ **target distribution**

Promote diversity across boundary samples

$$\mathcal{U}_{noisy} = [\mathcal{U} + \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})]\eta$$



Regularisation term

Total Variation Regularization

- Discourage high frequency component in images
- Visual difference between the training data and the boundary sample

$$L_{TV}(\mathbf{x}) = \sum_{w}^{W-1} \sum_{h}^{H-1} \sum_{c}^{C} |\mathbf{x}(w+1,h,c) - \mathbf{x}(w,h,c)| + |\mathbf{x}(w,h+1,c) - \mathbf{x}(w,h,c)|$$

Training algorithm

Algorithm 1: Training algorithm

Initialization: pretraining of the classification network

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

while *not converged* **do**

  # Phase 1: boundary samples generation

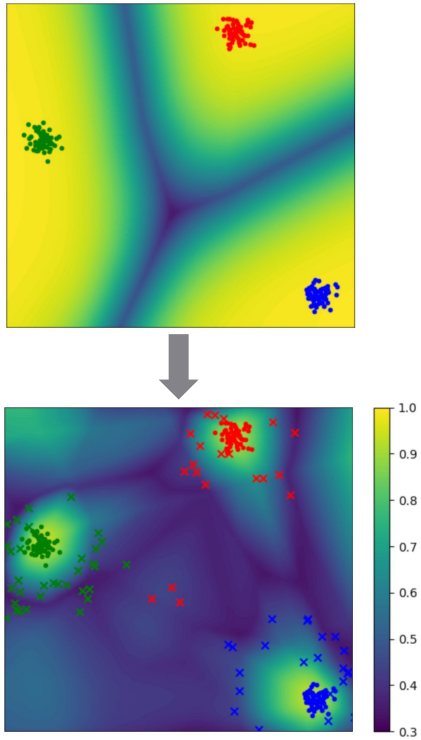
  initialize$(\mathbf{x}, y_t)$

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

  # Phase 2: fine tune with boundary samples

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

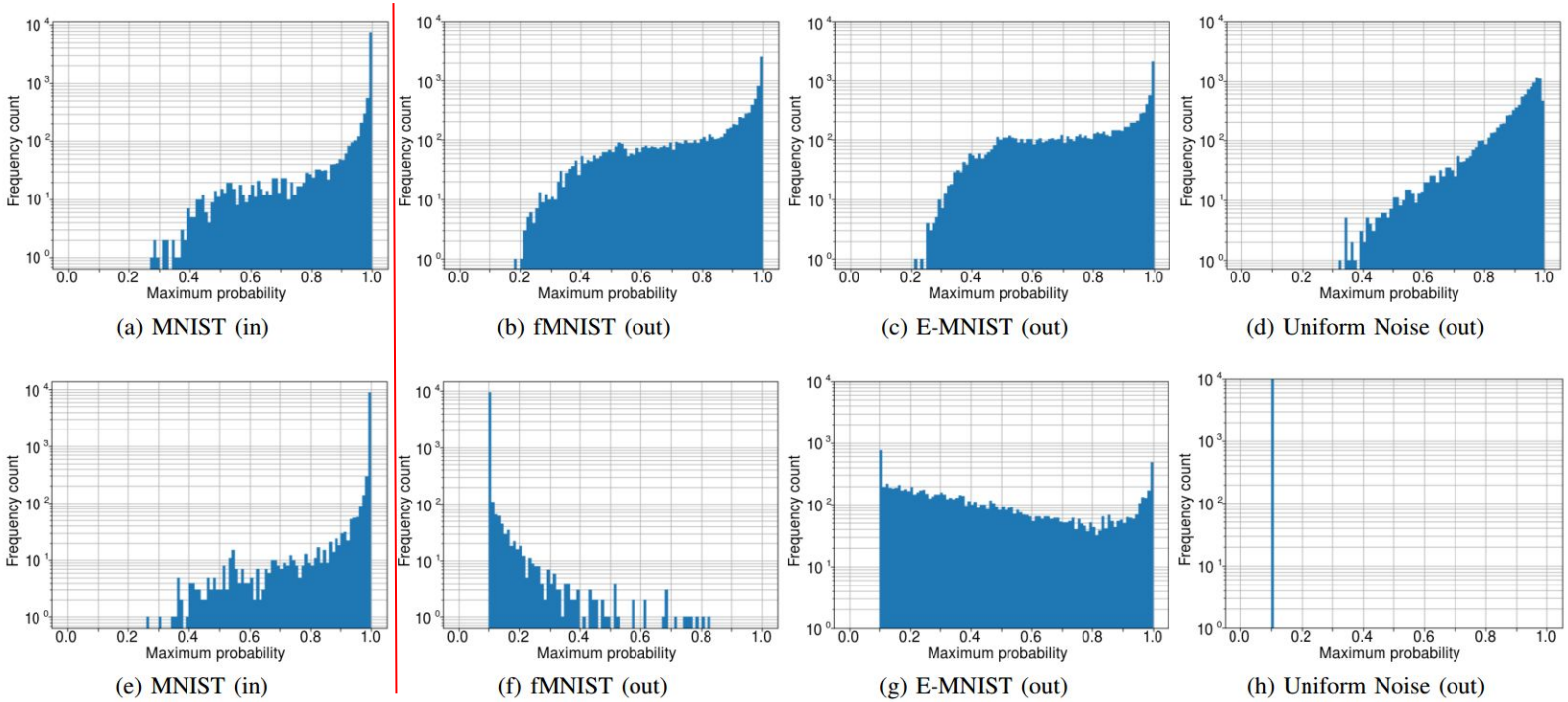
end



Experiments

Plain

BoS



(a) MNIST (in) (b) fMNIST (out) (c) E-MNIST (out) (d) Uniform Noise (out)

(e) MNIST (in) (f) fMNIST (out) (g) E-MNIST (out) (h) Uniform Noise (out)

<u>In distribution</u>: MNIST

<u>Out-of-Distribution</u>: fMNIST, E-MNIST, Uniform Noise

Experiments

Trained on **MNIST**	Plain (**98.93**%) MMC	AUROC	DA	BoS(98.70%) 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**%) MMC	AUROC	DA	BoS(98.70%) 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

BoS training shows a significant improvement in detecting OODs for all the considered datasets with respect to the plain baseline model trained on MNIST...

… and better results on some of the OOD datasets used for CIFAR-10.

Conclusions

Our contributions:

- 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.
- The experimental results supporting that our method outperforms the baseline.