Learning ODIN

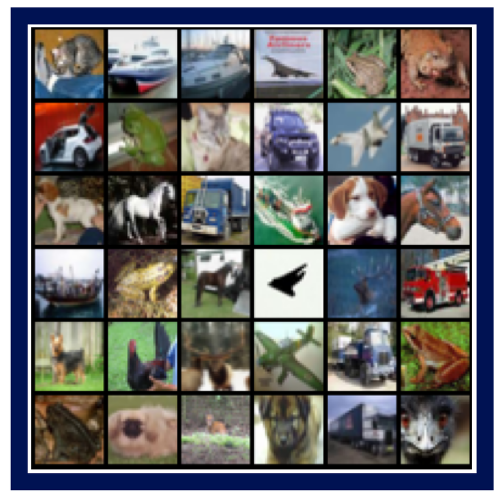
Amir Jevnisek and Shai Avidan

TEL AVIV UNIVERSITY אוניברסיטת תל אביב

BMVC 2022

Problem Statement

Out-of Distribution Detection Goal is to identify samples outside of a predetermined distribution, which is defined as In-Distribution. At training time only normal data is available. Area Under ROC Curve (AUROC) and FPR95 are the standard evaluation metrics.



In Distribution

Out of Distribution

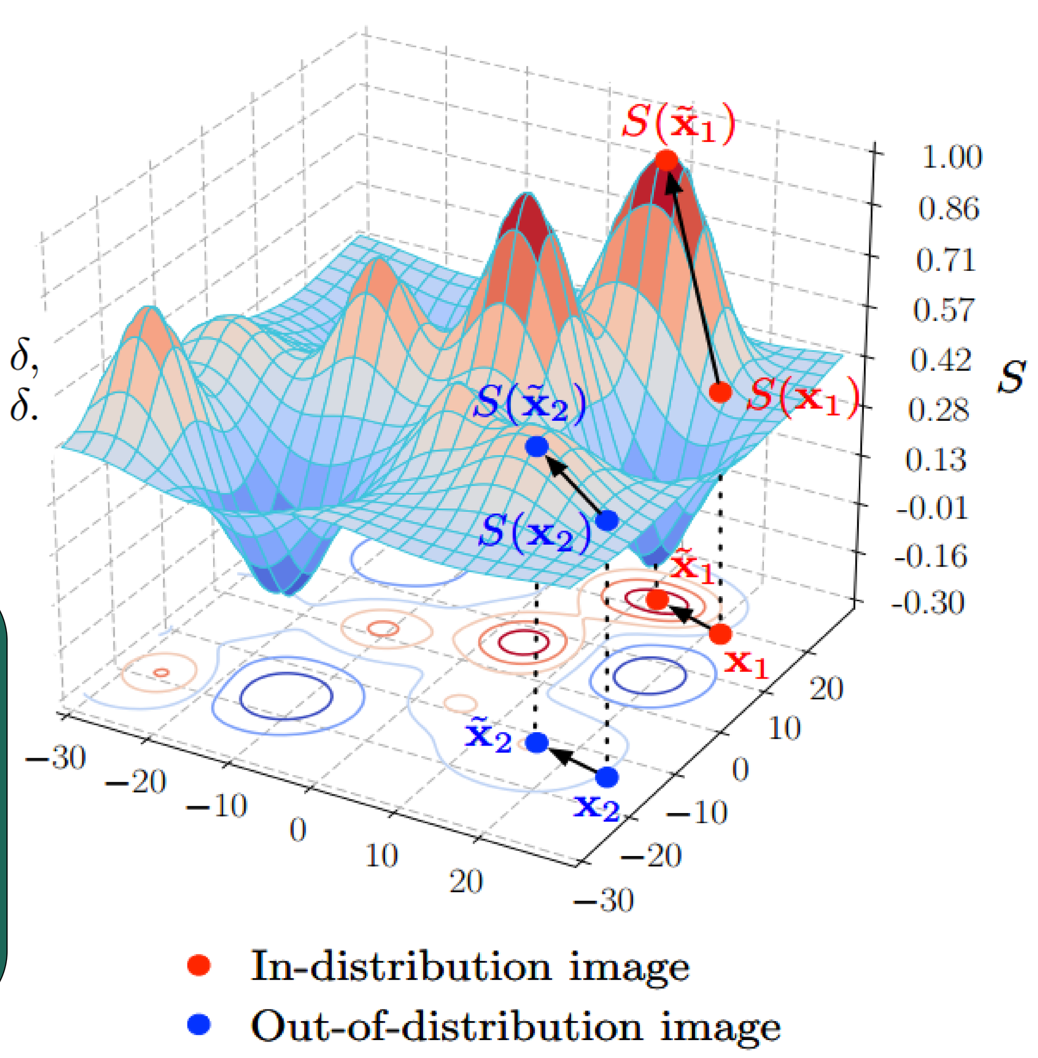
Introduction

ODIN OOD Detector:

(1) $S_i(\boldsymbol{x};T) = \frac{\exp\left(f_i(\boldsymbol{x})/T\right)}{\sum_{j=1}^{N}\exp\left(f_j(\boldsymbol{x})/T\right)},$

(2) $\tilde{\boldsymbol{x}} = \boldsymbol{x} - \varepsilon \operatorname{sign}(-\nabla_{\boldsymbol{x}} \log S_{\hat{y}}(\boldsymbol{x};T))$

(3) $g(\boldsymbol{x};\delta,T,\varepsilon) = \begin{cases} 1 & \text{if } \max_i p(\tilde{\boldsymbol{x}};T) \le \delta, \\ 0 & \text{if } \max_i p(\tilde{\boldsymbol{x}};T) > \delta. \end{cases}$



ODIN is a prevailing OOD framework.
ODIN assumes that the logit gap is maximal for In-Distribution samples.

We encode this assumption into the loss function.
So we learn what ODIN is expecting.

Implementation

Code is available!

Our NEW Regularization Term

We Reformalize ODIN's Decision Rule:

$$|S_{\hat{y}}(\tilde{x}) - S_{\hat{y}}(x)| \underset{OOD}{\overset{ID}{\gtrless}} Th.$$

We take the Taylor expansion of the sofrmax with the perturbed image $S_{\hat{y}}(\tilde{x})$ around x:

$$S_{\hat{y}}(\tilde{x}) = S_{\hat{y}}(x) + \varepsilon \cdot \nabla_x S_{\hat{y}}(x) + \text{higher orders of } \varepsilon$$

Plugging this into the ODIN decision rule, we get:

$$|S_{\hat{y}}(\tilde{x}) - S_{\hat{y}}(x)| = |S_{\hat{y}}(x) + \varepsilon \cdot \nabla_x S_{\hat{y}}(x) - S_{\hat{y}}(x)| =$$

$$= |\varepsilon \cdot \nabla_x S_{\hat{y}}(x)| \gtrless Th.$$

The term 'large' is ill-defined ('large' compared to what?). Instead, we introduce a Gradient Quotient (GQ) loss term:

$$\mathcal{L}_{GQ} = \frac{||\nabla_x S_{\hat{y}}||}{\sum_{y_j \neq \hat{y}} ||\nabla_x S_{y_j}||}$$

This way, the term 'large' refers to the softmax scores of the rest of the class labels.

We suggest to train networks with this as a regularization term:

$$L = L_{CE}(y, \hat{y}) + \lambda \cdot \frac{||\nabla_x S_y||_1}{\sum_{y_j \neq y} ||\nabla_x S_{y_j}||_1}$$

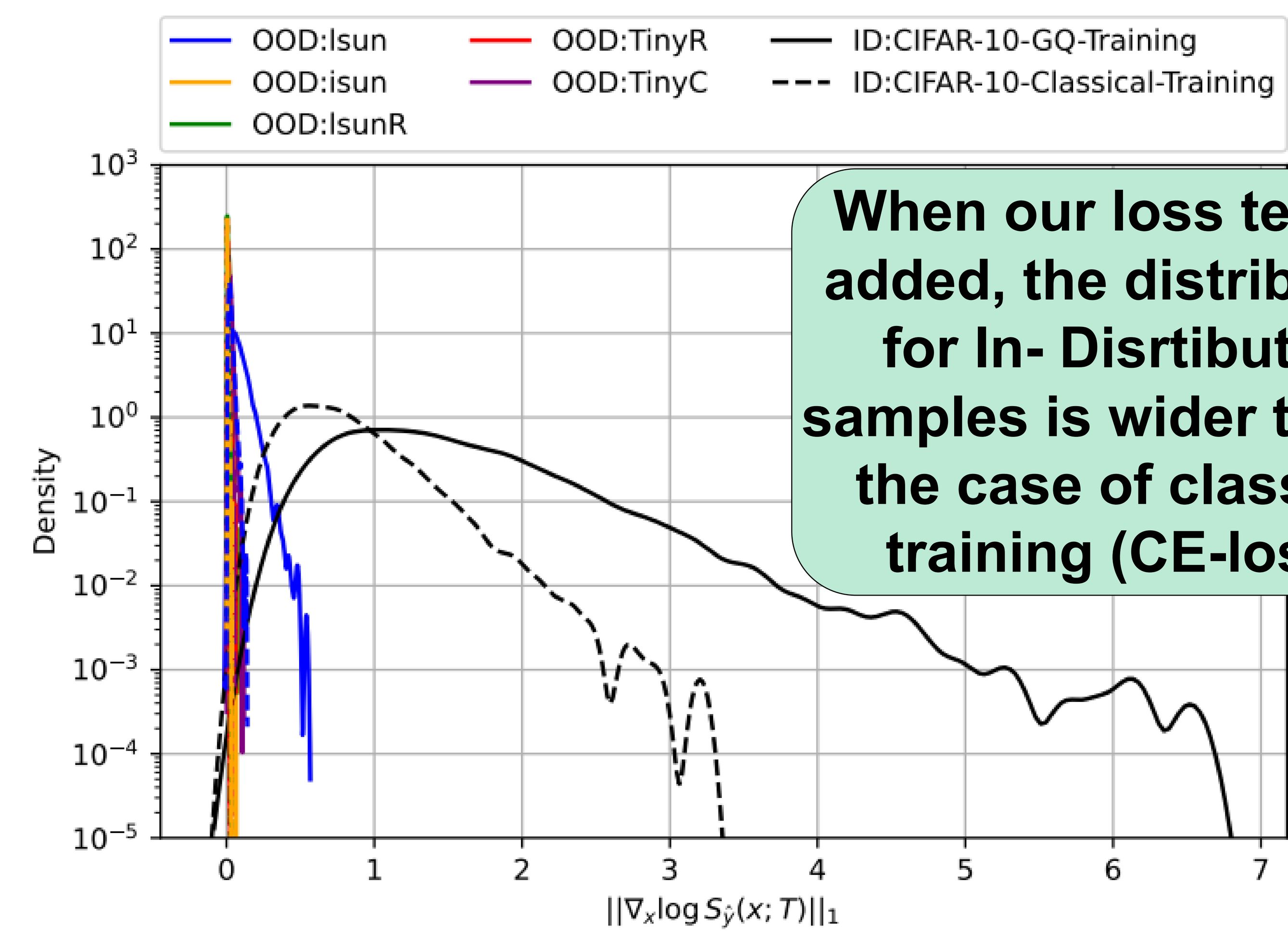
OOD Algorithm

Algorithm 1 Learning ODIN

Inputs (1) In-Distribution dataset $\mathcal{D}_{ID}^{train}$, (2) classifier $f_\theta(\cdot)$ with parameters θ
Outputs: (1) An OOD detector, (2) An In-Distribution classifier.

Train $f_\theta(\cdot)$ with the following loss: $L = L_{CE} + \lambda \cdot \mathcal{L}_{GQ}$
Create $g(x)$ an ODIN OOD-detector on f_θ's logits.
return $g(x)$ is the OOD-detector, f_θ is the In-Distribution classifier

Gradient Distribution



When our loss term is added, the distribution for In- Disrtibution samples is wider than in the case of classical training (CE-loss).

OOD Performance

Architecture	Method	AUROC ↑ (%)	FPR95 ↓ (%)	ID-Accuracy ↑ (%)
WideResNet-40-4	MSP	88.99	56.81	94.93
	ODIN	90.11	35.31	94.93
	Mahalanobis	89.33	35.48	94.93
	Energy	90.04	35.26	94.93
MSDNet Exit@last	MSP	89.72	49.87	94.09
	ODIN	90.33	29.30	94.09
	Mahalanobis	82.84	75.19	94.09
	Energy	90.48	33.62	94.09
MSDNet (dynamic exit)	MOOD	91.26	28.05	94.13
Madry's Resnet /+ GQ	ODIN	90.51 / 91.94	36.57 / 30.19	92.19 / 93.29
ResNet18 /+ GQ	ODIN	89.09 / **93.55**	27.75 / **25.14**	95.27 / 94.26
ResNet34 /+ GQ	ODIN	89.03 / 89.08	35.33 / 27.63	95.21 / **95.56**

CIFAR-10 as ID: Adding GQ to any of the architectures improves results. Our approach achieves best results for all metrics compared to baseline models.

High Res. Images

GQ also works for **High Resolution** images (ImageNet).
It is possible to use GQ for **fine tuning**, which lets us work with **pre-trained models**.