On the evaluation of post hoc Out-Of-Distribution detectors

Lorenzo Rossi

Problem definition: Out-Of-Distribution detection

$$G(x_0; f) = \begin{cases} 0 & \text{if } x_0 \sim \mathcal{D}_{out}, \\ 1 & \text{if } x_0 \sim \mathcal{D}_{in}. \end{cases}$$

The post hoc scenario

$$\text{Pre-trained model} \xrightarrow{\text{Class scores}} \text{Scoring function} \xrightarrow{\text{OOD score}} \text{ID or OOD}$$

$$G(x_0; f) = \begin{cases} 0 & \text{if } S(x_0; f) \leq t, \\ 1 & \text{if } S(x_0; f) > t. \end{cases}$$

Softmax score

$$S_{MSP}(x) = \max_k \mathtt{softmax}(f(x))_k$$

Main claim: Correctly classified samples have a higher maximum softmax probability.

Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks, 2016

ODIN

$$S_{\texttt{ODIN}}(\tilde{x};T) = \max_k \texttt{softmax}(\frac{f(\tilde{x})}{T})_k$$

$$\tilde{x} = x - \epsilon \texttt{sign}(-\nabla \log S_{\texttt{ODIN}}(x))$$

S. Liang, Y. Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks, 2017

Energy-based score

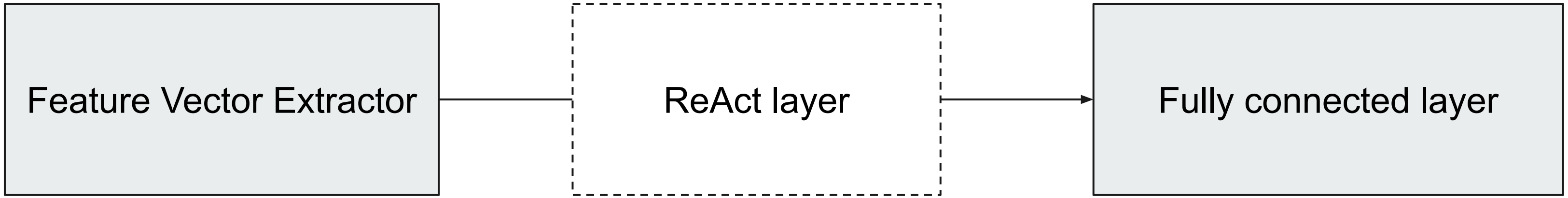
$$S_{\texttt{energy}}(x) = -\ln \sum_{i=1}^{K} e^{f(i)_i}$$

$$\begin{aligned}\mathcal{L}_{\texttt{energy}} &= \mathbb{E}[\max(0, S_{\texttt{energy}}(x_{in}) - m_{in})] \\ &+ \mathbb{E}[\max(0, m_{out}) - S_{\texttt{energy}}(x_{out})]\end{aligned}$$

W. Liu, X. Wang, J. D. Owens, and Y. Li. Energy-based out-of-distribution detection, 2020

ReAct

$$\texttt{ReAct}(x;\lambda) = \min(x, \lambda)$$



Y. Sun, C. Guo, and Y. Li. React: Out-of-distribution detection with rectified activations, 2021

Our results

- ODIN with $\varepsilon=0$ and $T=1000$ is a strong baseline.
- The benefits of adding ReAct are unclear.
- We found no strong correlation between the architecture and the size of the model, and the OOD detection performance.
- All the OOD detection methods show a good dataset transferability.

Methodology

- 8 OOD detectors
    - 4 scoring functions: softmax, ODIN(ε=0), ODIN(ε=0.0014), energy.
    - Each model with and without the ReAct layer
- 10 datasets used (iNaturalist*, SUN*, Places*, ImageNet, ImageNet a, ImageNet V2, Rock Paper Scissors, ImageNette, uniform noise, Gaussian noise.
- A wide range of augmentations (Gaussian noise, blur, pixelization, perspective transformation, adversarial noise, JPEG quality, adversarial noise)
- 9 Models (DensetNet121, DenseNet169, DenseNet201, ResNet50, ResNet101, ResNet152, VGG16, VGG19, EfficientNetB0), however we mainly focus on ResNet101.
- The first group of 64 dataset & augmentation pairs only ResNet101. The second group of 10 dataset & augmentation pairs for all the models.

*A non overlapping dataset is used

Main results

OOD dataset	Augmentation	`softmax`	`ODIN`($\epsilon = 0$)	`ODIN`($\epsilon = 0.0014$)	`energy`
RockPaperScissors	-	98.54/99.75	**99.98**/99.82	0.40/78.12	92.46/52.22
iNaturalist	-	88.42/94.17	92.37/**94.18**	0.58/78.52	78.13/56.67
SUN	-	86.19/89.78	89.51/**89.58**	1.26/73.51	81.07/50.09
Places	-	82.86/84.14	**87.09**/83.39	0.71/71.10	77.79/47.96
Gaussian Noise	-	99.84/**100.00**	**100.00/100.00**	0.21/90.88	98.03/97.56
Uniform Noise	-	99.85/**100.00**	**100.00/100.00**	0.28/95.04	95.57/98.25
ImageNet v2*	-	55.26/55.55	57.38/55.28	**99.64**/54.98	57.19/48.12
ImageNette*	-	**83.79**/75.92	52.58/53.93	1.06/52.70	55.34/53.00
ImageNet	Normal($\sigma = 0.002$)	51.31/52.73	52.35/52.56	24.77/52.08	**52.91**/48.98
ImageNet	Normal($\sigma = 0.25$)	79.69/87.30	**88.81**/86.97	3.77/79.54	84.82/60.31
ImageNet	Normal($\sigma = 1.25$)	95.63/98.31	**99.50**/98.77	0.49/88.71	97.77/86.15
ImageNet	Blur(r= 1.0)	55.40/54.93	**59.80**/54.85	40.10/53.57	58.06/46.68
ImageNet	Pixelation(r= 0.5)	56.16/56.14	**61.40**/56.30	52.21/54.30	59.16/48.24
ImageNet	JPEG(q= 25)	54.37/55.26	**58.53**/55.46	15.52/53.80	59.02/49.51

Result w/o and w/ ReAct using ResNet101 pre-trained on ImageNet

Best Methods

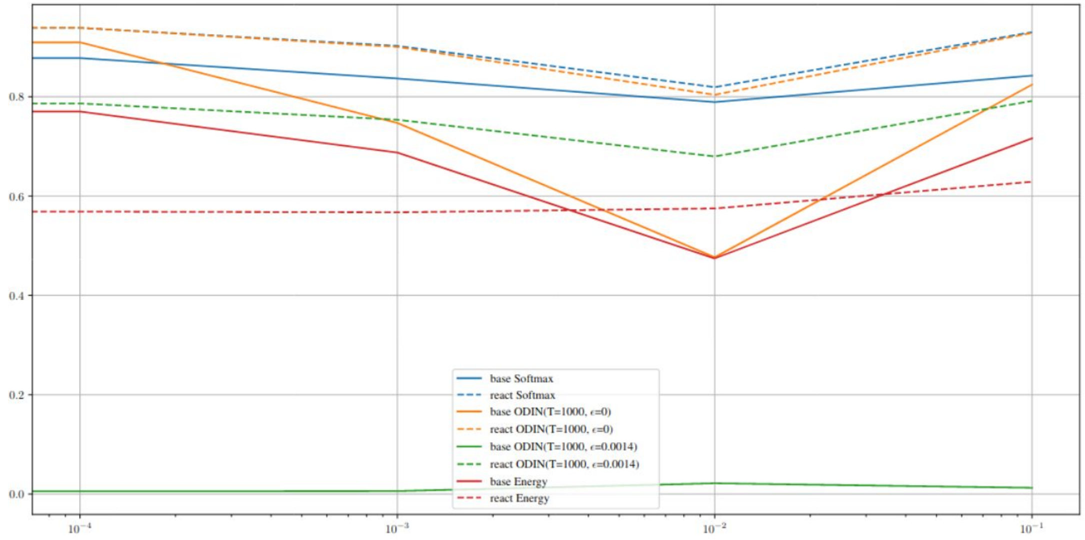
Scoring	ReAct	average rank
ODIN($\epsilon = 0$)	no	2.27
softmax	yes	2.71
ODIN($\epsilon = 0$)	yes	2.81
energy	no	4.06
softmax	no	4.52
ODIN($\epsilon = 0.0014$)	yes	5.57
energy	yes	6.76
ODIN($\epsilon = 0.0014$)	no	7.29

Scoring	Better w/o	Better w/
softmax	10	54
energy	59	5
ODIN($\epsilon = 0$)	41	23
ODIN($\epsilon = 0.0014$)	7	57

AUC comparison with and without ReAct

- ODIN (ε=0) is the best
- The benefits of ReAct heavily depends on the scoring function

Adversarial noise (FGSM)



$$\mathcal{L}_{adv}(x) = \ln \sum_{i=0}^{N} e^{\texttt{softmax}(f(x))_k}$$

$$x_{adv} = \texttt{clip}(x + \epsilon \texttt{sign}(\nabla \mathcal{L}_{adv}(x)))$$

Comparison among models

Model	Scorer	ReAct	Avg Rank
EfficientNetB0	ODIN($\epsilon = 0$)	no	1.22
ResNet50	ODIN($\epsilon = 0$)	no	1.33
DenseNet121	ODIN($\epsilon = 0$)	yes	1.44
DenseNet201	ODIN($\epsilon = 0$)	yes	1.56
VGG19	ODIN($\epsilon = 0$)	no	2.0
ResNet101	softmax	yes	2.0
ResNet101	ODIN($\epsilon = 0$)	no	2.11
ResNet151	ODIN($\epsilon = 0$)	yes	2.22
VGG16	ODIN($\epsilon = 0$)	no	2.22
VGG19	ODIN($\epsilon = 0$)	yes	2.22

We used for the average ranking the second group of 10 dataset and augmentation pairs.