VRA: Variational Rectified Activation for Out-of-distribution Detection

Mingyu Xu, Zheng Lian,
Bin Liu, Jianhua Tao

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

Background: Among all OOD detection method, ReAct [1] is a typical and effective post-hoc strategy, which truncates activations above a threshold c for each activation z:

$$g(z) = \min(z, c)$$

ReAct has demonstrated that this truncation operation can increase the gap between ID and OOD:

$$\mathbb{E}_{\text{in}}[g(z)] - \mathbb{E}_{\text{out}}[g(z)] \geq \mathbb{E}_{\text{in}}[z] - \mathbb{E}_{\text{out}}[z].$$

Question: despite its promising results, is this strategy the best option for widening the gap between ID and OOD?

VRA Framework

To find the best modify operation, we optimize the following objectives:
- Maximize the gap between ID and OOD.
- Minimize the modification brought by the operation to maximally preserve the input

The final objective function is calculated as follows:

$$\min_g \mathcal{L}(g) = \mathbb{E}_{\text{in}}[(g(z) - z)^2] - 2\lambda\left(\mathbb{E}_{\text{in}}[g(z)] - \mathbb{E}_{\text{out}}[g(z)]\right)$$

Assumption 1 *We assume $\mathbb{E}_{in}[|z|]$, $\mathbb{E}_{out}[|z|]$, $\mathbb{E}_{in}[z^2]$, and $\mathbb{E}_{out}[z^2]$ exist. Let $\mathcal{G}$ be a Hilbert space:*

$$\mathcal{G} = \{g(z) | \mathbb{E}_{in}[|g(z)|], \mathbb{E}_{out}[|g(z)|], \mathbb{E}_{in}[g(z)^2], \mathbb{E}_{out}[g(z)^2] < \infty\}.$$

This space is sufficiently complex containing most functions, such as identity functions, constant functions, and all bounded continuous functions. Then, we define the inner product of $\mathcal{G}$ as follows:

$$< g_a(z), g_b(z) >= \int g_a(z) g_b(z) p_{in}(z) dz.$$

Then we can apply the variational method and get optimal activation function among the function space:

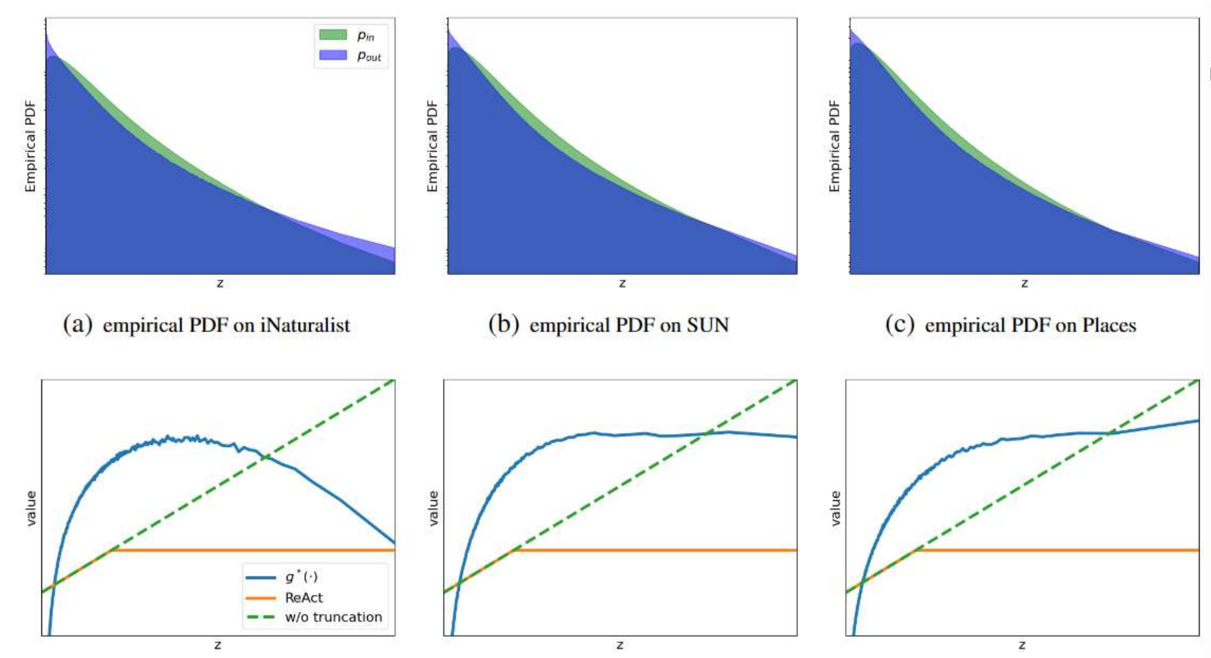
$$g^*(z) = z + \lambda\left(1 - \frac{p_{\text{out}}(z)}{p_{\text{in}}(z)}\right)$$

But in practice, this operation depends on the specific expressions of probability density of ID data and OOD data. Estimating these expressions is a challenging task given that OOD data comes from unknown distributions [2].

This drives us to look for more practical implementations.

Practical Implementations

We treat ImageNet as ID data and select multiple OOD datasets. We first use histograms to approximate the probability density functions of ID data and OOD data.



(a) empirical PDF on iNaturalist (b) empirical PDF on SUN (c) empirical PDF on Places
(d) activation functions on iNaturalist (e) activation functions on SUN (f) activation functions on Places

Empirical PDFs for ID data and OOD data, and visualization of different rectified functions.
We treat ImageNet as ID data and select multiple OOD datasets for visualization.

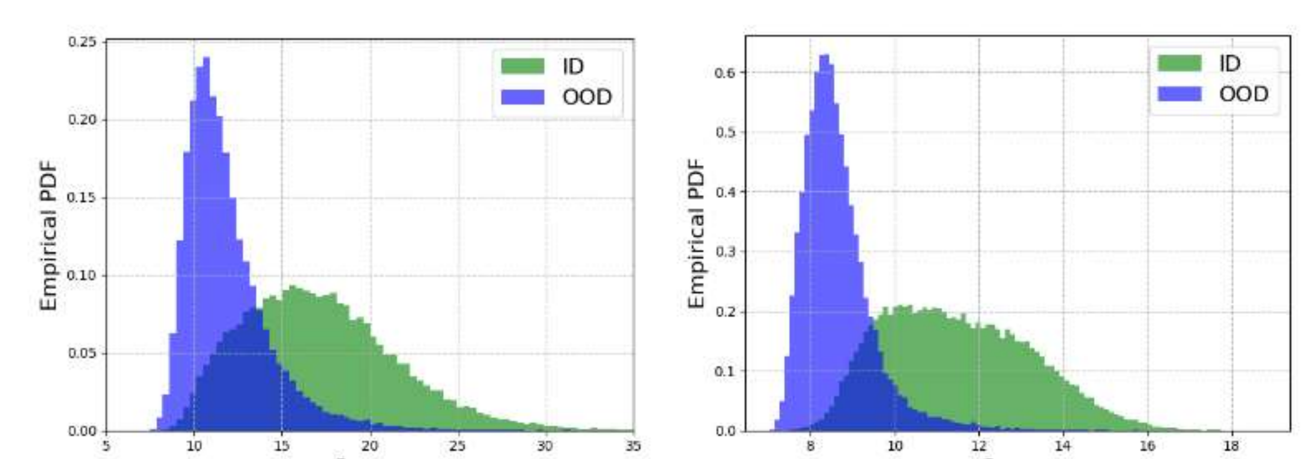
Unlike ReAct, the optimal operation further demonstrates the necessity of suppressing abnormally low activations in OOD detection. To mimic such operations, we design a piecewise function called VRA:

$$\text{VRA}(z) = \begin{cases} 0, z < \alpha \\ z, \alpha \leq z \leq \beta \\ \beta, z > \beta \end{cases}$$

Meanwhile, we observe that the optimal function amplifies intermediate activations in the figure. Therefore, we propose another variant of VRA called VRA+, which further introduces another hyper-parameter to control the degree of amplification:

$$\text{VRA+}(z) = \begin{cases} 0, z < \alpha \\ z + \gamma, \alpha \leq z \leq \beta \\ \beta, z > \beta \end{cases}$$

Visualization on the Distribution of Energy Score



(a) before rectification (overlap: 20.25%) (b) after rectification (overlap: 12.52%)

Distribution of scores before and after variational rectification.

Experimental Results

(see more comparison in our paper)

Main results on ImageNet.

Method	iNaturalist		SUN		Places		Textures		Average	
	FR. ↓	AU. ↑	FR. ↓	AU. ↑	FR. ↓	AU. ↑	FR. ↓	AU. ↑	FR. ↓	AU. ↑
Backbone: ResNet-50 [29]										
MSP [20]	54.99	87.74	70.83	80.86	73.99	79.76	68.00	79.61	66.95	81.99
ODIN [21]	47.66	89.66	60.15	84.59	67.89	81.78	50.23	85.62	56.48	85.41
Mahalanobis [22]	97.00	52.65	98.50	42.41	98.40	41.79	55.80	85.01	87.43	55.47
Energy [23]	55.72	89.95	59.26	85.89	64.92	82.86	53.72	85.99	58.40	86.17
ReAct [8]	20.38	96.22	**24.20**	94.20	**33.85**	**91.58**	47.30	89.80	31.43	92.95
KNN [24]	59.08	86.20	69.53	80.10	77.09	74.87	**11.56**	**97.18**	54.32	84.59
DICE [10]	25.63	94.49	35.15	90.83	46.49	87.48	31.72	90.30	34.75	90.78
SHE [25]	34.22	90.18	54.19	84.69	45.35	90.15	45.09	87.93	44.71	88.24
ASH[27]	44.57	92.51	52.88	88.35	61.79	85.58	42.06	89.70	50.32	89.04
VRA	15.70	**97.12**	26.94	94.25	37.85	91.27	21.47	95.62	25.49	94.57
VRA+	**15.48**	97.08	**23.50**	**94.91**	34.62	**91.79**	19.66	96.08	**23.31**	**94.97**

Compatibility with different scoring functions.

Method	CIFAR-10		CIFAR-100		ImageNet		Average	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP [20]	48.73	92.46	80.13	74.36	66.95	81.99	65.27	82.94
MSP + ReAct	48.00	92.77	77.69	76.22	55.63	87.85	60.44	85.61
MSP + DICE	43.72	92.92	**76.86**	**76.39**	67.41	82.24	62.66	83.85
MSP + VRA	**42.31**	**93.50**	79.69	75.94	**47.09**	**89.62**	**56.36**	**86.35**
Energy [23]	26.55	94.57	68.45	81.19	58.41	86.17	51.14	87.31
Energy + ReAct	26.45	94.67	62.27	84.47	31.43	92.95	40.05	90.70
Energy + DICE	20.83	95.24	**49.72**	87.23	34.75	90.77	35.10	91.08
Energy + VRA	**17.74**	**96.47**	53.24	**88.74**	**25.49**	**94.57**	**32.16**	**93.26**
ODIN [21]	24.57	93.71	58.14	84.49	56.48	85.41	46.40	87.87
ODIN + ReAct	21.00	95.98	54.17	88.62	42.21	91.28	39.13	91.96
ODIN + DICE	26.05	94.62	61.39	83.83	62.89	84.48	50.11	87.64
ODIN + VRA	**17.38**	**96.52**	**47.12**	**90.21**	**32.75**	**93.39**	**32.42**	**93.37**

Further Analysis

Combining features with logit outputs can achieve better performance in OOD detection [3]. Therefore, we design another variant of VRA called VRA++, whose scoring function is defined as:

$$\lambda_v \sum_{i=1}^{m} g(z_i) + \log \sum_{i=1}^{c} e^{l_i}.$$

Unlike VRA using piecewise functions, we further test the performance of the quadratic function called VRA++:

$$-\lambda_v \sum_{i=1}^{m} (z_i^2 - \alpha_v z_i) + \log \sum_{i=1}^{c} e^{l_i}.$$

Performance of VRA++. All methods are based on BiT and pretrained on ImageNet.

Method	OpenImage-O		Texture		iNaturalist		ImageNet-O		Average	
	FR. ↓	AU. ↑	FR. ↓	AU. ↑	FR. ↓	AU. ↑	FR. ↓	AU. ↑	FR. ↓	AU. ↑
MSP [20]	73.72	84.16	76.65	79.80	64.09	87.92	96.85	57.12	77.83	77.25
ODIN [21]	72.83	85.64	74.07	81.60	70.75	86.73	96.85	63.00	78.63	79.24
Mahalanobis [22]	64.32	83.10	14.05	97.33	64.95	85.70	70.05	80.37	53.34	86.63
Energy [23]	73.42	84.77	73.91	81.09	74.98	84.47	96.40	63.59	79.68	78.48
ReAct [8]	54.97	88.94	50.25	90.64	48.60	91.45	91.70	67.07	61.38	84.52
ViM [37]	43.96	91.54	**04.69**	**98.92**	55.71	89.30	61.50	83.87	41.47	90.91
VRA++	**34.94**	**93.55**	05.02	98.76	**22.25**	**96.37**	**60.45**	**84.21**	**30.67**	**93.22**

References

[1] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In Proceedings of the Advances in Neural Information Processing Systems, pages

[2] Christopher M Bishop. Pattern recognition and machine learning. Springer, 2006.

[3] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4921–4930, 2022

Code: https://github.com/zeroQiaoba/VRA