中国科学院软件研究所学术年会’2022
暨计算机科学国家重点实验室开放周

学术论文

2021 IEEE 32nd International Symposium on Software Reliability Engineering (ISSRE)

Out-of-Distribution Detection through Relative Activation-Deactivation Abstractions

Zhen Zhang[1,3], Peng Wu[1,3], Yuhang Chen[2,3], and Jing Su[1,3]

[1]State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences
[2]Technology Center of Software Engineering, Institute of Software, Chinese Academy of Sciences
[3]University of Chinese Academy of Sciences, Beijing, China
{zhangzhen19, wp}@ios.ac.cn

❖ Motivation

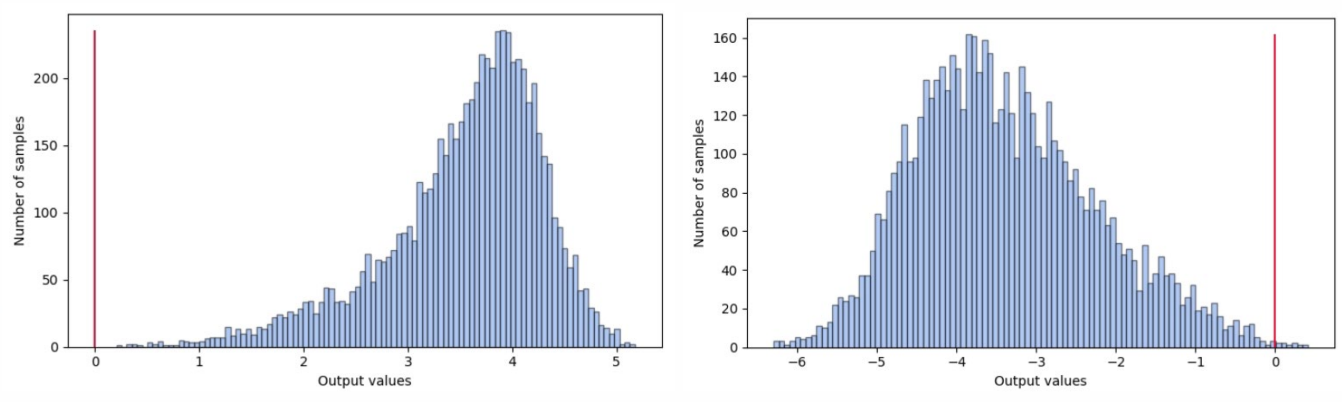
(a) Hand-written digits (b) Hand-written character

- A deep learning model will anyway classify an input to a category that the model is trained for.
- But predicting a picture of a hand-written character to a digital category is totally wrong.
- Out-of-Distribution (OOD) detection aims to detect such an OOD input (if any).

❖ Our Method

➢ Output Distribution of A Neuron

- Indeed, a neuron may output a relatively greater or less value for certain categories than for others.



(a) Class 1 (b) Class 3

Fig. 1. Output Distributions of A Neuron (MNIST)

➢ Relative Activation & Relative Deactivation

- Let $\mu_{i,l}^{-\hat{y}}$ denote the average output of neuron $n_{i,l}$ under all the inputs $x' \in D$ that is not classified into category $\hat{y}$ by model M, i.e.

$$\mu_{i,l}^{-\hat{y}} = \frac{\sum_{x' \in D, M(x') \neq \hat{y}} out_i^l(x')}{|\{x' \mid x' \in D, M(x') \neq \hat{y}\}|}$$

- If $out_i^l(x) > \mu_{i,l}^{-\hat{y}}$, neuron $n_{i,l}$ is relatively activated.
- If $out_i^l(x) < \mu_{i,l}^{-\hat{y}}$, neuron $n_{i,l}$ is relatively deactivated.

➢ Relative Selectivity

- The relative selectivity $rs_i^l(x)$ of neuron $n_{i,l}$ under input $x \in D$ that is classified into category $\hat{y}$ by model M is defined as below:

$$rs_i^l(x) = \frac{out_i^l(x) - \mu_{i,l}^{-\hat{y}}}{\mu_{i,l}}$$

where $\mu_{i,l}$ is the average output of neuron $n_{i,l}$ under all the inputs $x' \in D$.

➢ Relative Activation-Deactivation Abstractions

- Neurons can be abstracted into three states under any input: **relative strongly activated**, **relatively non-selective**, and **relative strongly deactivated**.
- Then, the following three-valued function $abst$: $\mathbb{R} \to \{-1, 0, 1\}$ uniformly abstracts the inference behavior on neuron $n_{i,l}$ under input x:

$$abst(rs_i^l(x)) = \begin{cases} 1 & \text{if } rs_i^l(x) \geq ub \\ 0 & \text{if } rs_i^l(x) \in (lb, ub) \\ -1 & \text{if } rs_i^l(x) \leq lb \end{cases}$$

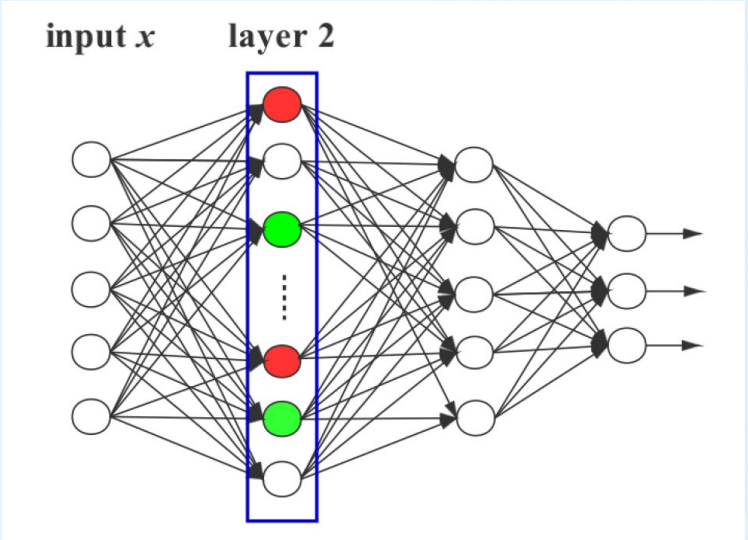


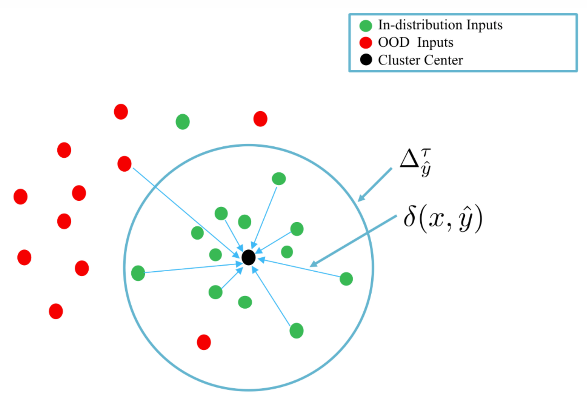
Fig. 2. Relative Activation-Deactivation Abstractions

❖ OOD Detection

➢ Re-AD

- The relative activation-deactivation abstractions (Re-AD) are rather close to each other under the inputs of the same categories, while far away from each other under the inputs of different categories.
- An OOD input, of which the category is unknown to the model, may lead the model to diverge from its Re-AD abstraction patterns collected for the predicted category.
- A Boolean indicator for OOD detection can be formally defined:

$$OOD_M(x) = \begin{cases} true & \text{if } \delta(x, \hat{y}) > \Delta_{\hat{y}}^{\tau} \\ false & \text{otherwise} \end{cases}$$



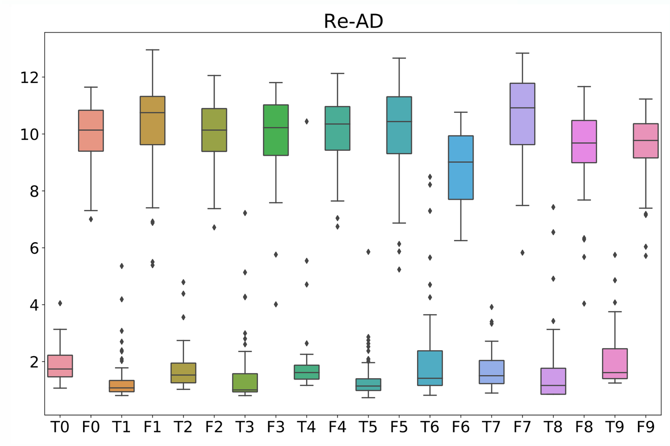


Fig. 3. OOD Detector **Fig. 4.** GTSRB vs TinyImageNet

❖ Experimental Results

➢ OOD Detection Type I

- Two datasets: one for training, the other for OOD detection.

Training	OOD	Re-AD	Baseline	OpenMax	ODIN
MNIST	FMNIST	0.9660	0.9822	0.9851	0.9882
	Omniglot	0.9753	0.9712	0.9778	0.9787
	Uniform Noise	0.9846	0.9960	0.9931	0.9975
	Gaussian Noise	0.9861	0.9971	0.9939	0.9983
FMNIST	MNIST	0.9762	0.7159	0.7374	0.7872
	Omniglot	0.9742	0.6651	0.7002	0.7409
	Uniform Noise	0.8505	0.8382	0.8568	0.8918
	Gaussian Noise	0.9723	0.9110	0.9143	0.9488
Cifar10	TinyImageNet	0.8792	0.8321	0.8428	0.8575
	LSUN	0.9033	0.8509	0.8721	0.9087
	iSUN	0.9012	0.8461	0.8678	0.9041
	Uniform Noise	0.8540	0.7081	0.6743	0.7355
	Gaussian Noise	0.9473	0.7614	0.7549	0.7896
GTSRB	TinyImageNet	0.9916	0.9898	0.5002	0.9959
	LSUN	0.9924	0.9906	0.5002	0.9966
	iSUN	0.9924	0.9907	0.5002	0.9965
	Uniform Noise	0.9943	0.9916	0.5002	0.9964
	Gaussian Noise	0.9949	0.9938	0.5002	0.9986
Average		0.9520	0.8907	0.7595	0.9173

➢ OOD Detection Type II

- Splitting a dataset into two parts with different categories: one for training and the other for OOD detection

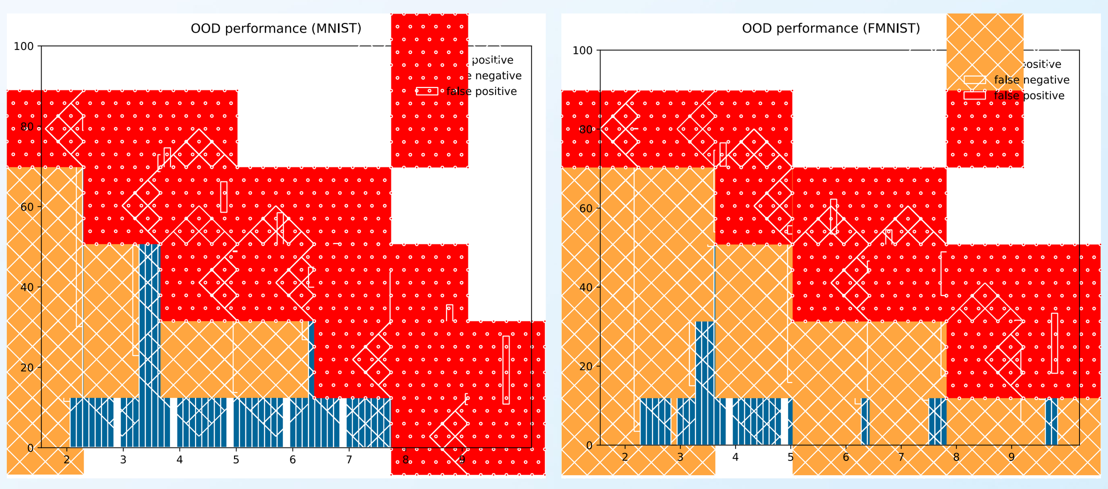


Fig. 5. Performance of OOD detection

➢ OOD Detection in Object Detection System

- When an object detection system trained in Cityscapes dataset is applied in different weather conditions, the wrong predictions can be well detected through Re-AD:

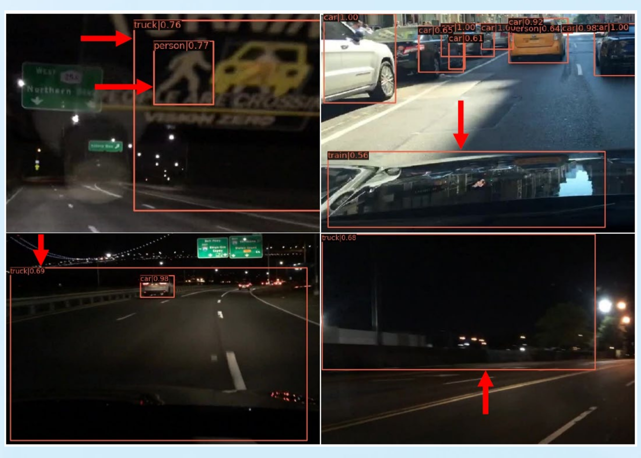


Fig. 6. Froggy Cityscapes **Fig. 7.** BDD100K

❖ Conclusion

- We propose the notion of relative selectivity to equally value the effects of both the activation and deactivation behaviors of neurons.
- We present a Re-AD approach to represent the inference behavior of the deep learning, and it is also an effective solution for OOD detection.
- Experiments results show that Re-AD outperforms the state-of-the-art OOD detection approaches in terms of the AUROC and TPR performances, without adjusting the input data or the model itself.