KHU

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

LINe: Out-of-Distribution Detection by Leveraging Important Neurons

Yong Hyun Ahn[1], Gyeong-Moon Park[2,*], Seong Tae Kim[2,*]
[1]Department of Artificial Intelligence, Kyung Hee University
[2]Department of Computer Science and Engineering, Kyung Hee University

THU-AM-320

Paper Code

Augmented Intelligence Lab @ KHU

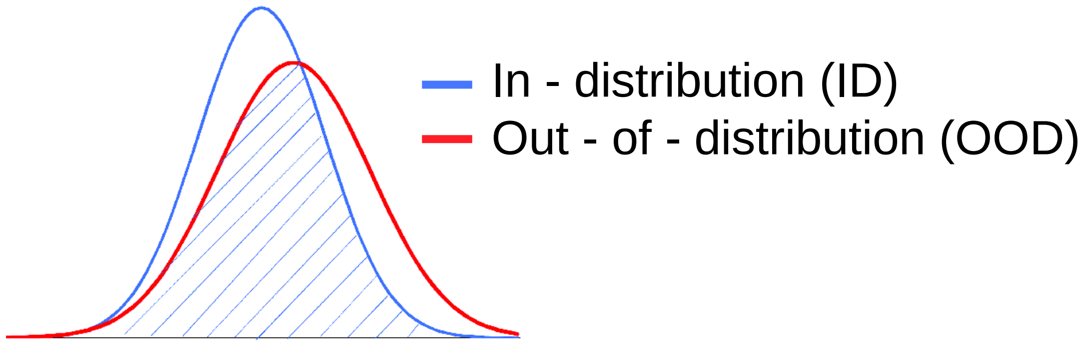
Overview

Out-of-Distribution (OOD) detection:

Goal : Identify the input is **In-Distribution (ID)** or **Out-of-Distribution (OOD)**

Recent Out-of-Distribution (OOD) detection methods deals with:

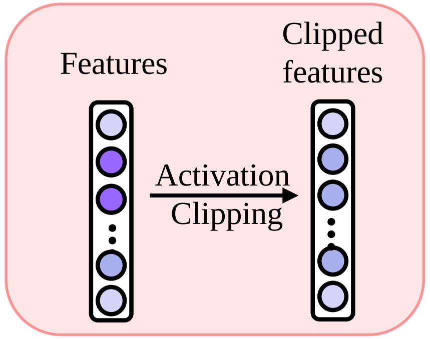
→ “ How to reduce noisy outputs ”



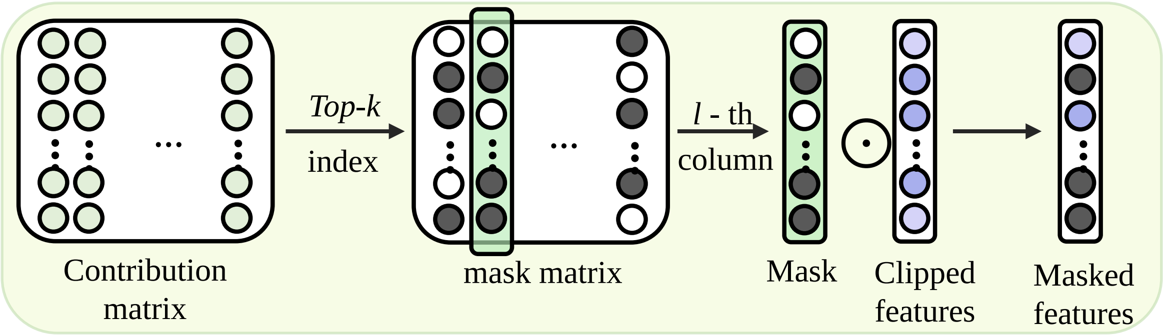
<Overlap between ID and OOD>

Overview

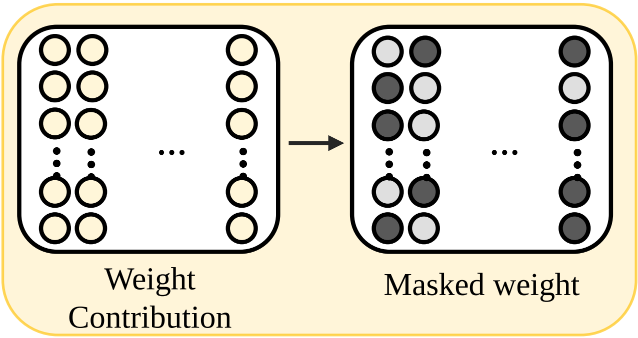
- LINe reduce noisy outputs by using two simple techniques
- LINe selectively uses class-wise important neurons.
- LINe is a simple yet effective method for OOD detection.



Activation Clipping (AC)



Activation Pruning (AP)



Weight Pruning (WP)

Overview

➞ LINe achieves SoTA performance on various OOD detection tasks

Method	ImageNet-1k		CIFAR-10		CIFAR-100	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP	66.95	81.99	48.73	92.46	80.13	74.36
ODIN	56.48	85.41	24.57	93.71	58.14	84.49
Mahalanobis	87.43	55.47	31.42	89.15	55.37	82.73
Energy	58.41	86.17	26.55	94.57	68.45	81.19
ReAct	31.43	92.95	26.45	94.95	62.27	84.47
DICE	34.75	90.77	20.83	95.24	49.72	87.23
DICE + ReAct	27.25	93.40	16.48	96.64	49.57	85.08
LINe (Ours)	**20.70**	**95.03**	**14.71**	**96.99**	**35.67**	**88.67**

Introduction

Motivation

Current Challenges:

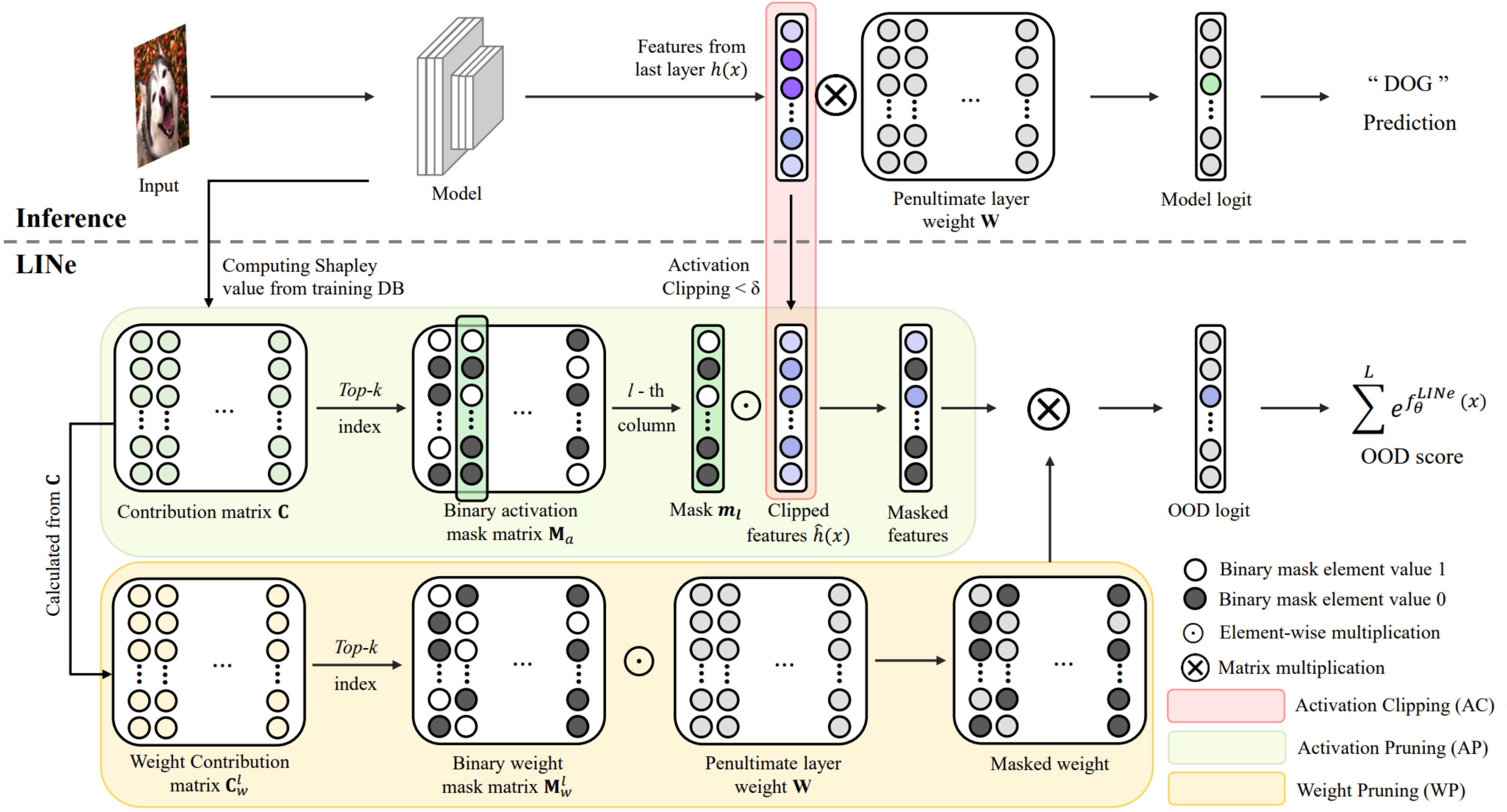
- How to reduce noisy output?

Our key Insights:

- Neurons with high Shapley value represents essential concept of input.
- Class-wise average of Shapley value allow us to calculate contribution for each class.
- LINe selectively use **important neurons only** to reduce noisy outputs.

Method

Method Overview

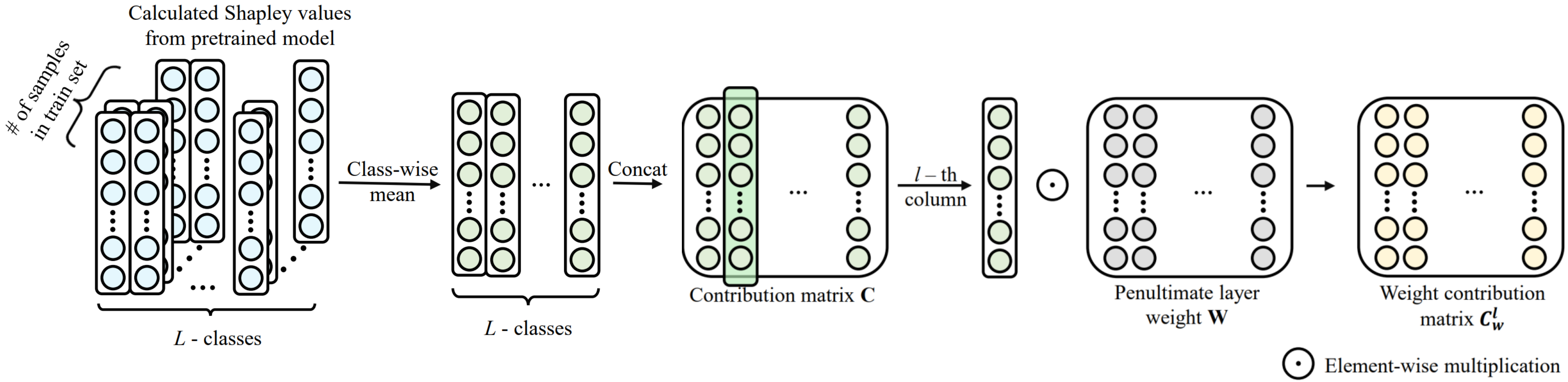


Method

Step 0: Prepare Class-wise Contribution of each neuron and weight

- Calculate class-wise contribution(i.e., Shapley value) from training data
- We use Taylor approximation of Shapley value introduced in *Khakzar et al.* [1]

  - Contribution matrix C, C_w construction using Shapley value

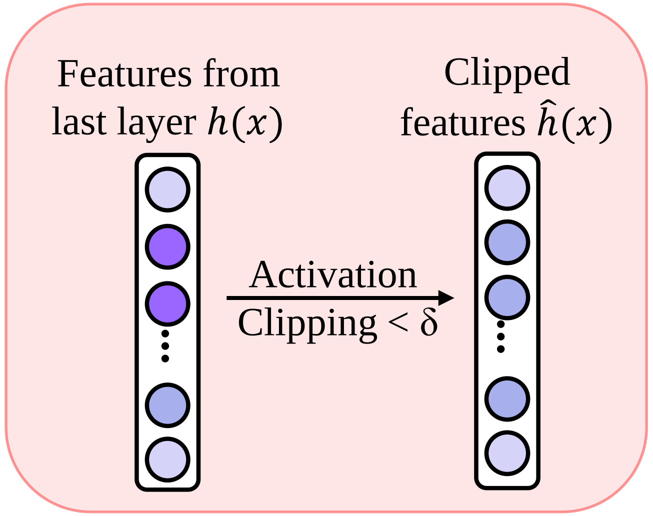


[1] Khakzar, Ashkan, et al. "Neural response interpretation through the lens of critical pathways." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

Method

Step 1: Activation Clipping (AC)

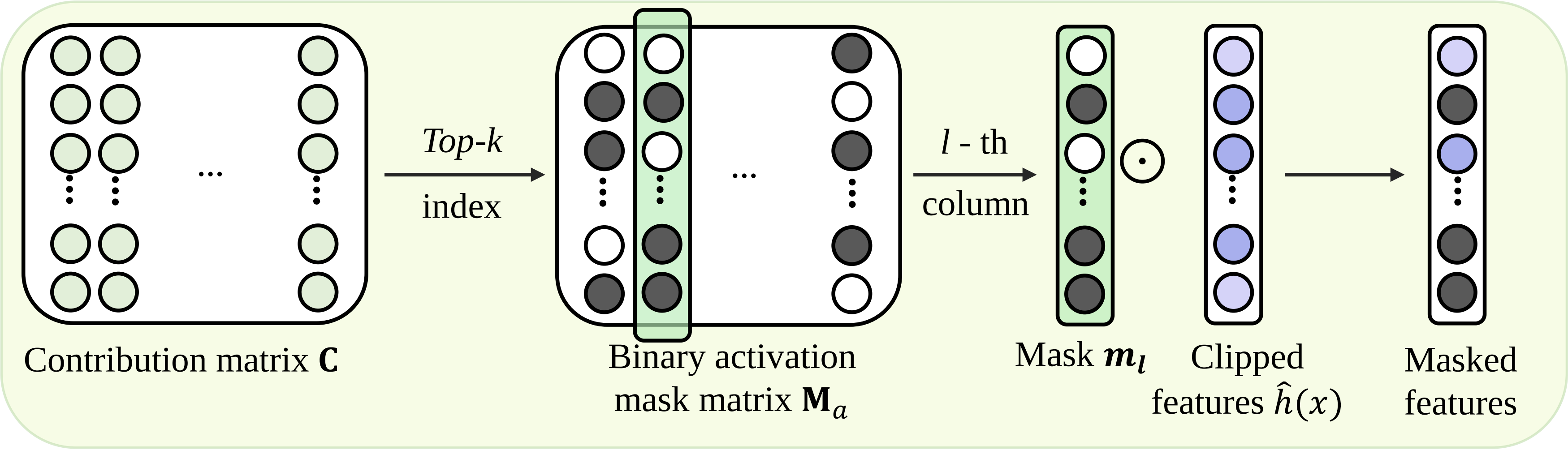
- Rectify sample penultimate layer activation below certain threshold (δ)
- Number of important neuron activation can be considered by AC
  - Apply Activation Clipping in last layer activation



Method

Step 2: Activation Pruning (AP)

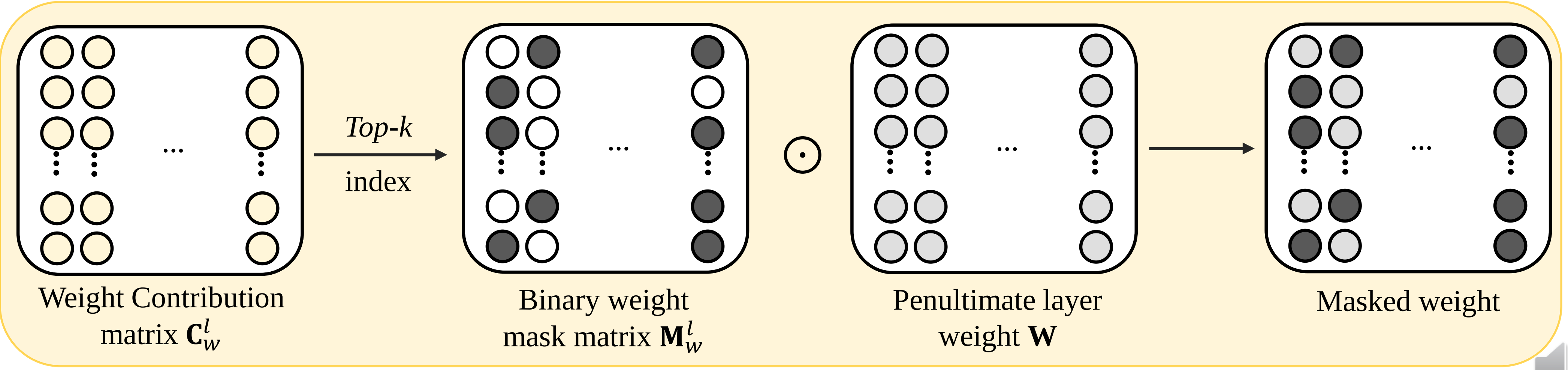
- Apply binary activation mask of predicted class to clipped feature.
- Reduce signals from less import neurons.
  - Apply Activation Pruning to clipped feature



Method

Step 3: Weight Pruning (WP)

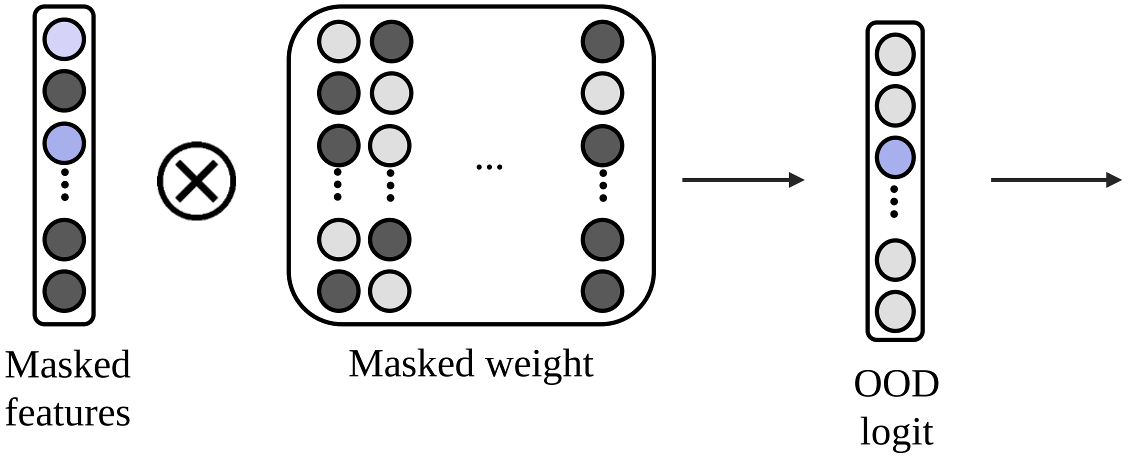
- Apply binary weight mask of predicted class to last layer weight.
- Reduce signals from less import weights.
  - Apply Weight Pruning to last layer weight



Method

Step 4: Calculate OOD score

- Multiply masked activation and weight, use Energy[2] to calculate OOD score



$$\sum^{L} e^{f_\theta^{LINe}(x)}$$

OOD score

Where, $f_\theta^{LINe}(x) = (\mathbf{W} \odot \mathbf{M}_w^l)^T (\mathbf{m}_l \odot \hat{h}(x)) + \mathbf{b}$

Experiments

Experiments

1. Experiment on ImageNet-1k benchmarks

- LINe outperform all the other baselines.

Method	OOD Datasets								Average	
	iNaturalist		SUN		Places		Textures			
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP (ICLR 17')	54.99	87.74	70.83	80.86	73.99	79.76	68.00	79.61	66.95	81.99
ODIN (ICLR 18')	47.66	89.66	60.15	84.59	67.89	81.78	50.23	85.62	56.48	85.41
Mahalanobis (NeurIPS 18')	97.00	52.65	98.50	42.41	98.40	41.79	55.80	85.01	87.43	55.47
Energy (NeurIPS 20')	55.72	89.95	59.26	85.89	64.92	82.86	53.72	85.99	58.41	86.17
ReAct (NeurIPS 21')	20.38	96.22	24.20	94.20	33.85	91.58	47.30	89.80	31.43	92.95
DICE (ECCV 22')	25.63	94.49	35.15	90.83	46.49	87.48	31.72	90.30	34.75	90.77
DICE + ReAct (ECCV 22')	18.64	96.24	25.45	93.94	36.86	90.67	28.07	92.74	27.25	93.40
LINe (Ours)	**12.26**	**97.56**	**19.48**	**95.26**	**28.52**	**92.85**	**22.54**	**94.44**	**20.70**	**95.03**

Experiments

2. Experiment on CIFAR benchmarks

- LINe outperform all the other baselines.

Method	CIFAR-10		CIFAR-100	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP (ICLR 17')	48.73	92.46	80.13	74.36
ODIN (ICLR 18')	24.57	93.71	58.14	84.49
Mahalanobis (NeurIPS 18')	31.42	89.15	55.37	82.73
Energy (NeurIPS 20')	26.55	94.57	68.45	81.19
ReAct (NeurIPS 21')	26.45	94.95	62.27	84.47
DICE (ECCV 22')	20.83	95.24	49.72	87.23
DICE + ReAct (ECCV 22')	16.48	96.64	49.57	85.08
LINe (Ours)	**14.71**	**96.99**	**35.67**	**88.67**

Experiments

3. Ablation study

- Effectiveness of each part (AP, WP, AC)

Method	AC	AP	WP	FPR95↓	AUROC↑
Energy [32]				58.41	86.17
Energy + AC	✓			35.40	91.86
LINe w/o WP	✓	✓		26.88	93.77
LINe w/o AP	✓		✓	23.19	94.57
LINe (Ours)	✓	✓	✓	**20.70**	**95.03**

- Effect of different thresholds (δ) of AC

Threshold (δ)	FPR95 ↓	AUROC ↑
$\delta = 0.1$	41.18	88.44
$\delta = 0.4$	23.43	94.79
$\delta = 0.8$	**20.70**	**95.03**
$\delta = 1.0$	21.69	94.81
$\delta = 1.5$	26.96	93.99
$\delta = 2.0$	31.88	92.97
$\delta = \infty$ (no AC)	44.88	89.14

Experiments

4. Discussion

- How to select pruning percentile? (p_a, p_w)

→ Different percentile due to different degree of overparameterization of models.

Table 8. **Percentage of class-specific neuron overlap in multiple classes.** Difference between the percentage of class-specific neuron overlap in multiple classes on three data sets. For each dataset, we calculated the proportion of very important (top 10%) neurons in more than o% of the class. All values are percentages.

Overlap	CIFAR-10	CIFAR-100	ImageNet
$o = 20$	24.56	26.90	1.70
$o = 30$	23.39	0.58	0.15

Thank you

Acknowledgements.This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea Government (MSIT)(No. 2022-0-00078: Explainable Logical Reasoning for Medical Knowledge Generation, No. 2021-0-02068: Artificial Intelligence Innovation Hub, No. RS-2022-00155911: Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University)) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. 2021R1G1A1094990).