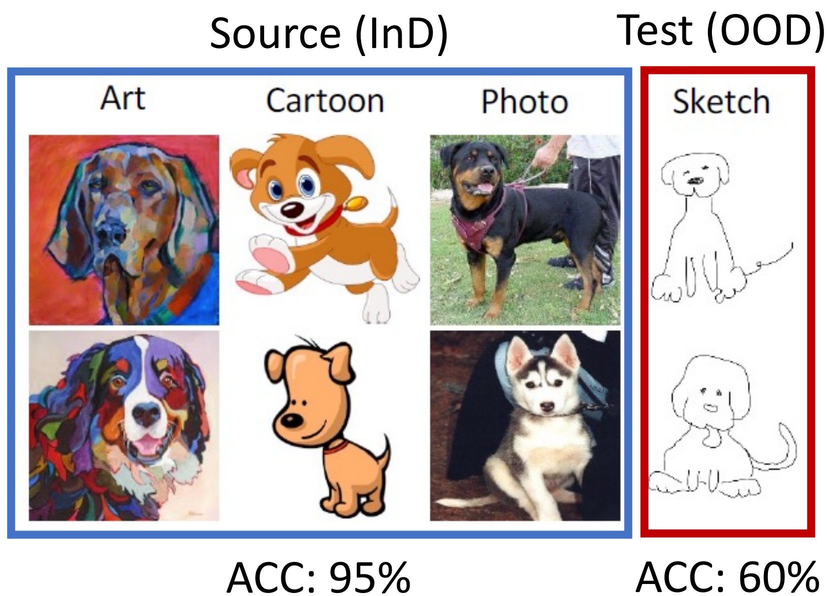
Neuron Activation Coverage: Rethinking Out-of-distribution Detection and Generalization

Yibing Liu   Chris Xing Tian
Haoliang Li   Lei Ma   Shiqi Wang

The 12th International Conference on Learning Representations (ICLR 2024)

Out-of-distribution (OOD) Problems

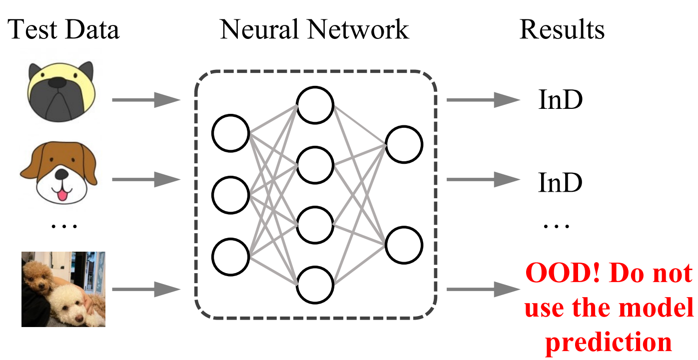
Distribution shifts between OOD and InD often drastically challenge well-trained models.



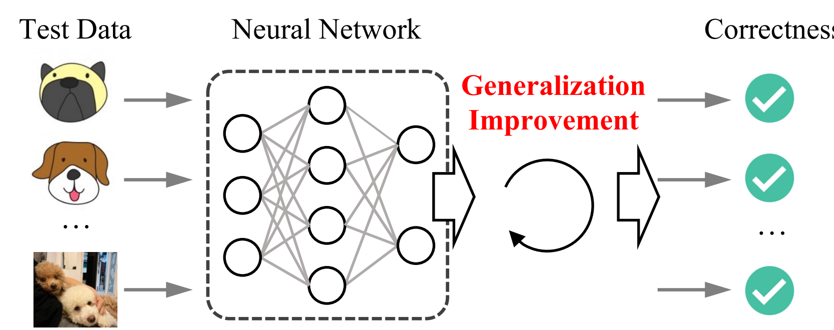
Out-of-distribution (OOD) Problems

Existing studies tackling OOD mainly arise from two avenues:

(a) OOD Detection

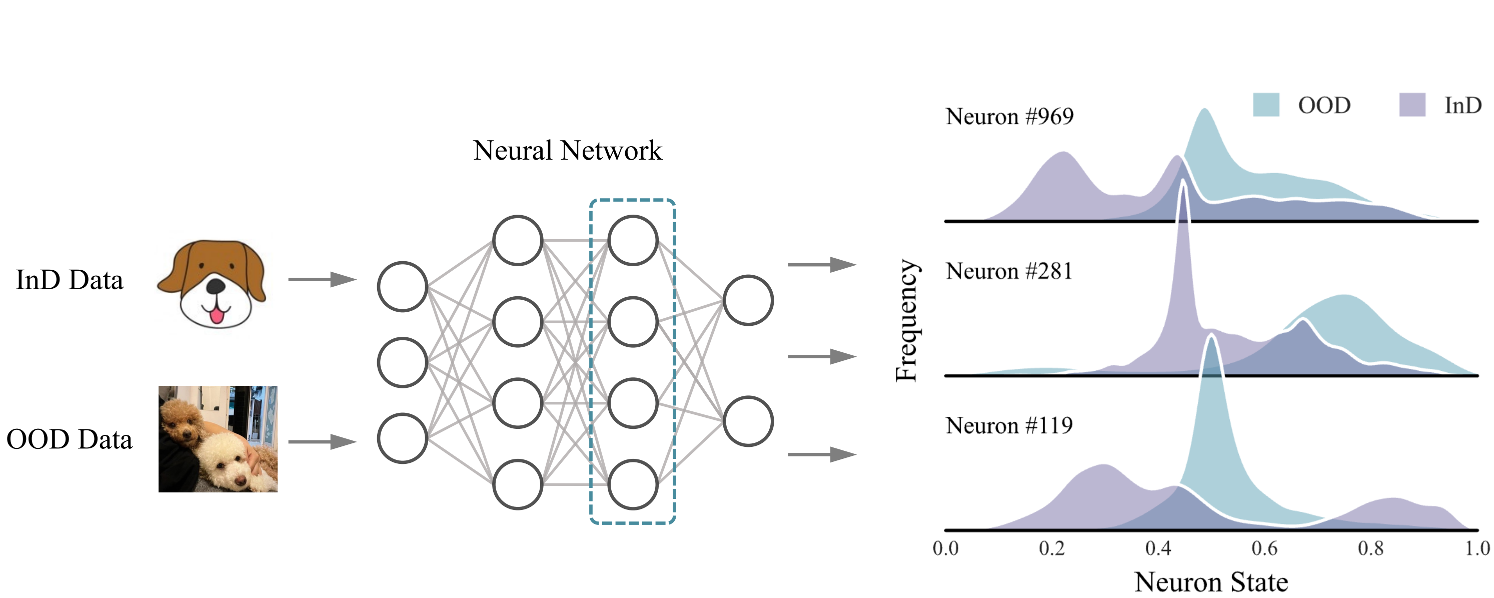


(b) OOD Generalization



Neurons in OOD scenarios

In this work, we study OOD problems from a neuron perspective.

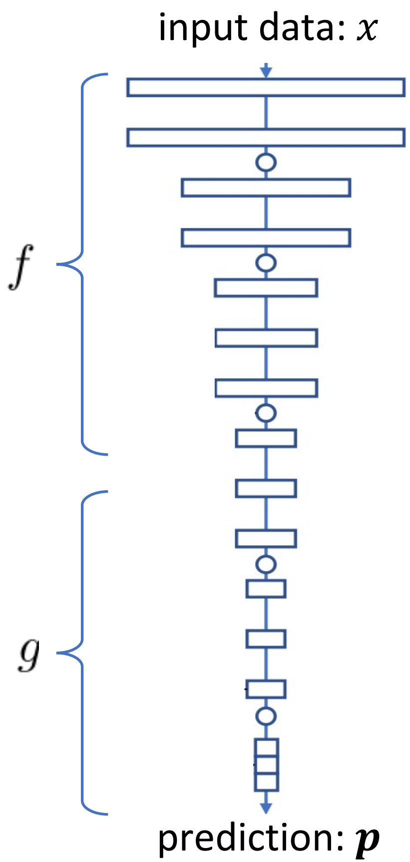


Neurons can exhibit distinct activation patterns when exposed to InD and OOD!

How to characterize neuron states?

Network:

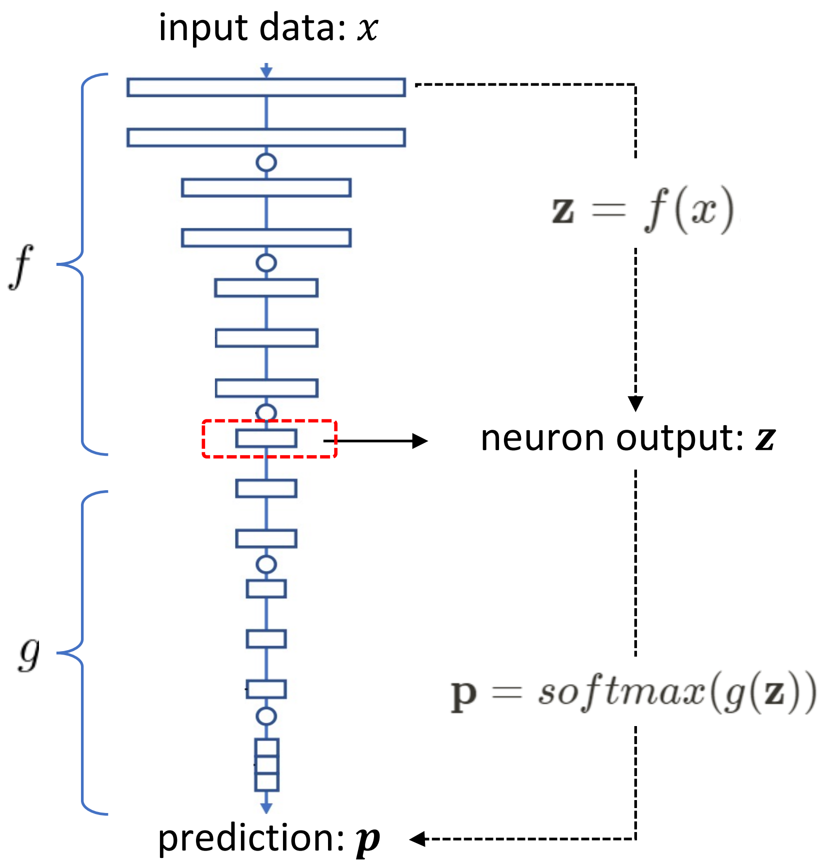
$$F = f \circ g.$$



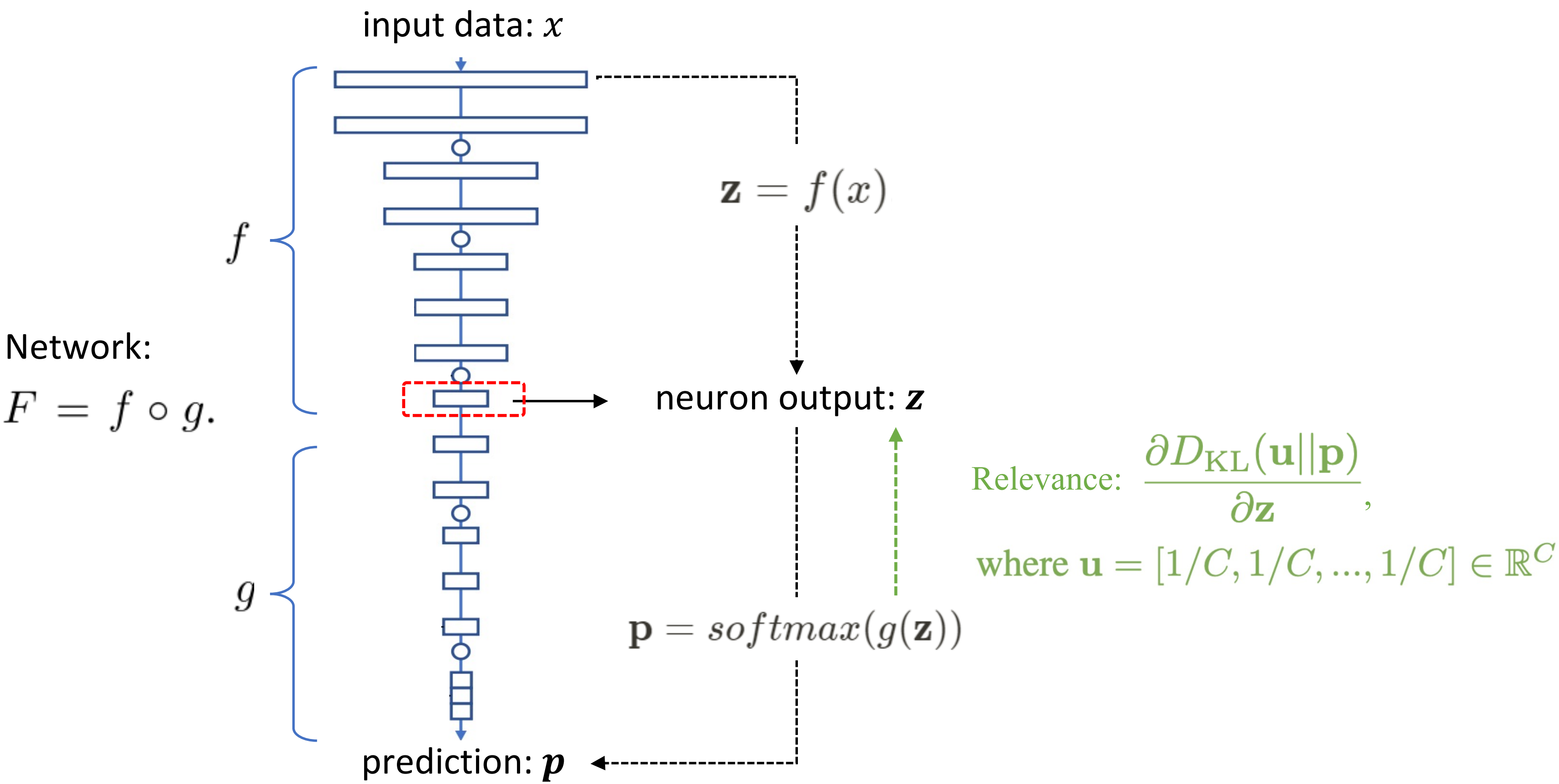
How to characterize neuron states?

Network:

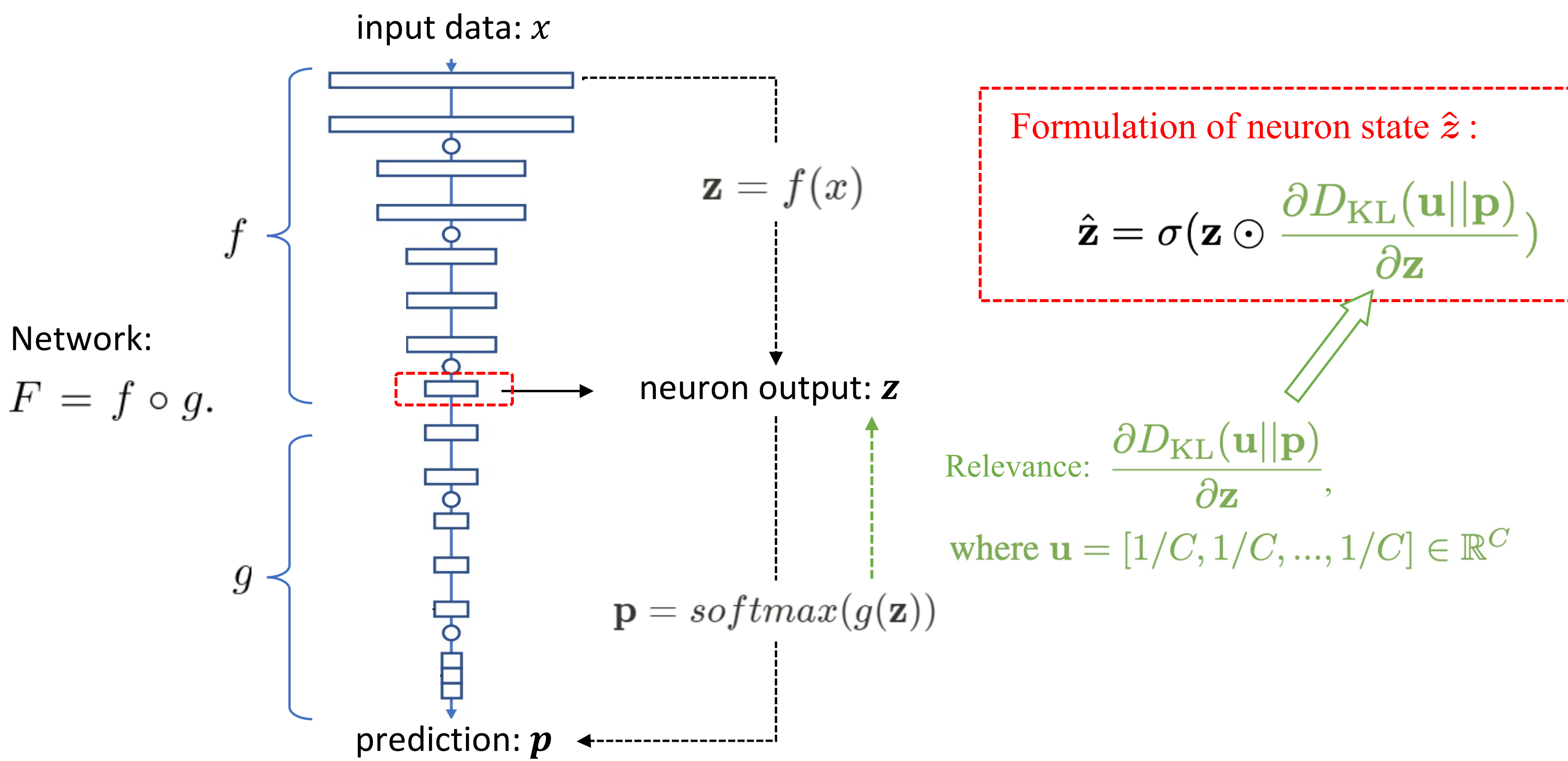
$F = f \circ g.$



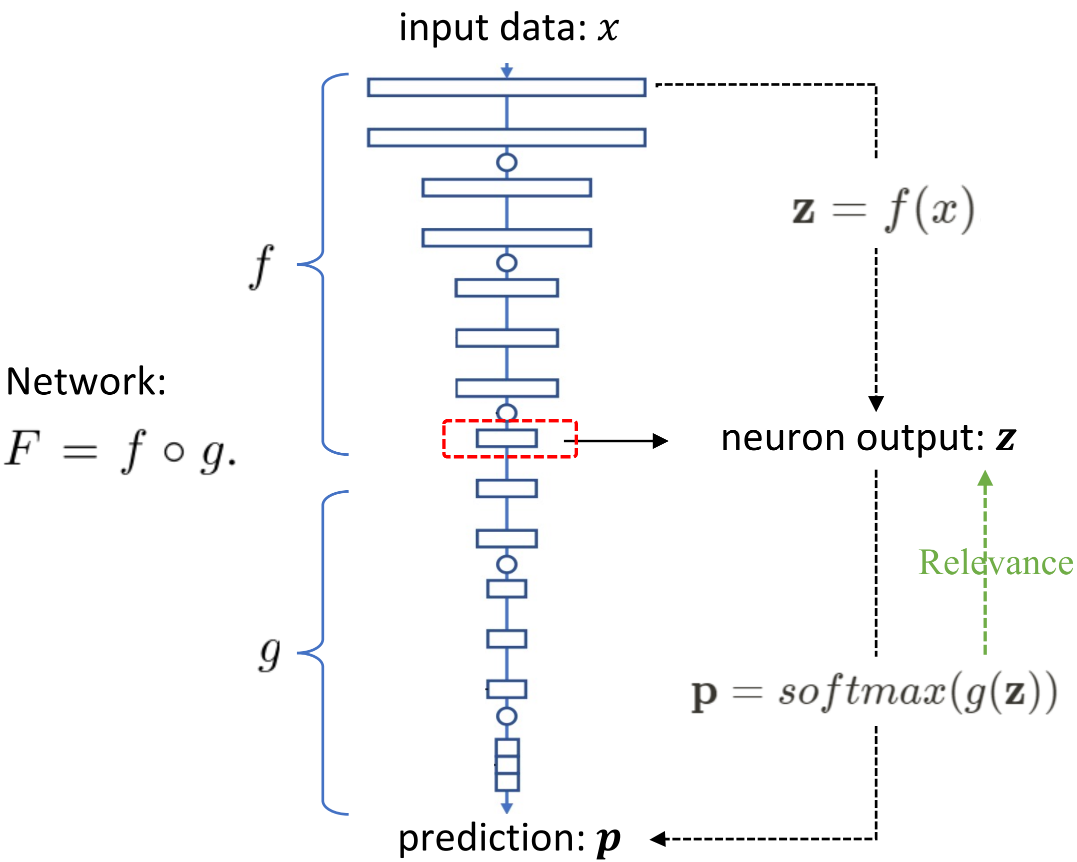
How to characterize neuron states?



How to characterize neuron states?



How to characterize neuron states?



Rationale of neuron state $\hat{\boldsymbol{z}}$:

$$\hat{\mathbf{z}} = \sigma(\mathbf{z} \odot \frac{\partial D_{\mathrm{KL}}(\mathbf{u}||\mathbf{p})}{\partial \mathbf{z}})$$

$$= \sigma(\mathbf{z} \odot (\frac{\partial g(\mathbf{z})}{\partial \mathbf{z}} \cdot \frac{\partial D_{\mathrm{KL}}}{\partial g(\mathbf{z})}))$$

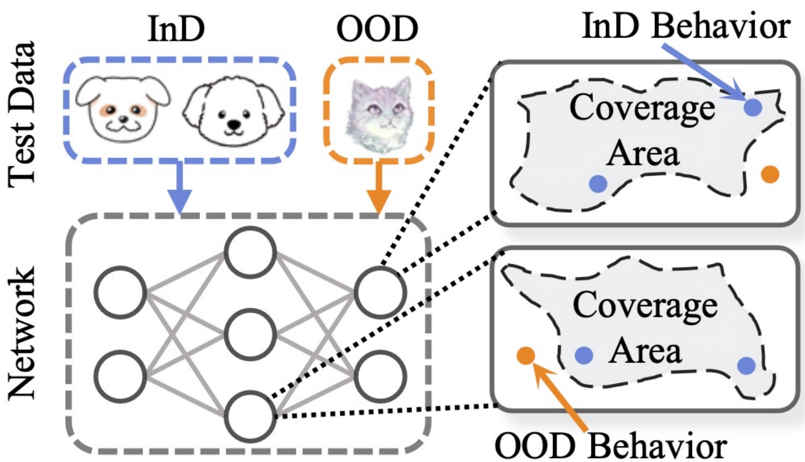
$$= \sigma\Big(\sum_{i=1}^{C}(\mathbf{z} \odot \frac{\partial g(\mathbf{z})_i}{\partial \mathbf{z}}) \cdot (p_i - u_i)\Big)$$

Input ⊙ Gradient Explanation

Prediction Confidence

Method: Neuron Activation Coverage (NAC)

Idea: Rarely-activated (covered) neurons by a training set can potentially trigger undetected bugs during the test stage (Pei et al., 2017) .



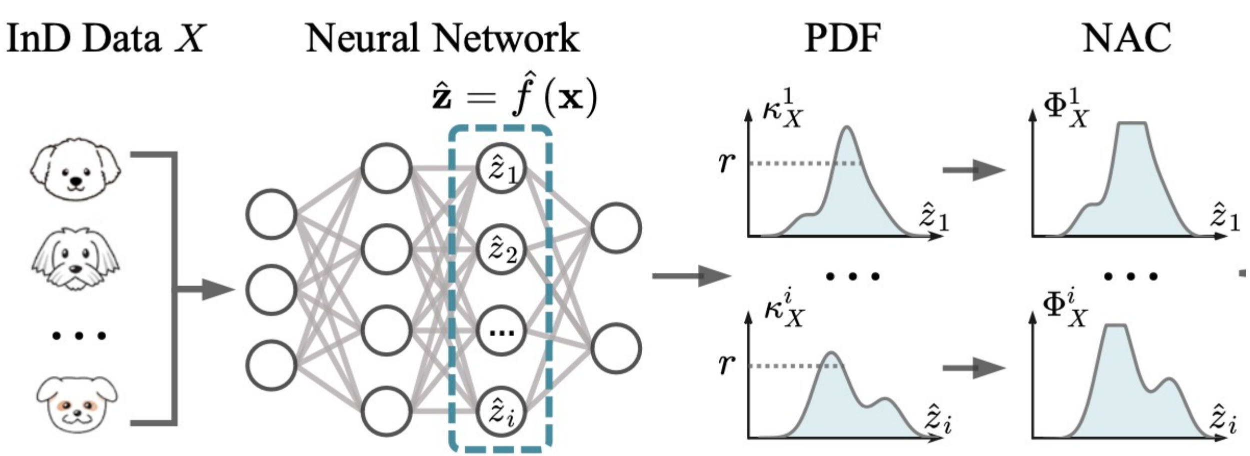
- NAC models *coverage area* in neuron activation space using InD training data.
- Upon receiving OOD data, neurons tend to behave outside the coverage area.

Method: Neuron Activation Coverage (NAC)

We derive NAC from the probability density function (PDF)

$$\Phi_X^i(\hat{z}_i; r) = \frac{1}{r}\min(\kappa_X^i(\hat{z}_i), r)$$

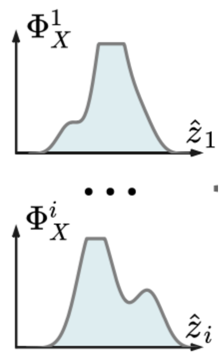
Threshold



Applications of NAC

In this work, we apply NAC to two OOD problems.

NAC



OOD Detection ⇒

Uncertainty Estimation (NAC-UE):

$$S(\mathbf{x}^*; \hat{f}, X) = \frac{1}{N} \sum_{i=1}^{N} \Phi_X^i(\hat{f}(\mathbf{x}^*)_i; r)$$

Idea: "higher NAC score"
↓
"fewer bugs (coming from InD)"

OOD Generalization ⇒

Model Evaluation (NAC-ME):

$$G(X, \theta) = \frac{1}{N} \sum_{i=1}^{N} \int_{\xi=0}^{1} \Phi_X^i(\xi; r)\, d\xi$$

Idea: "larger coverage area"
↓
"better generalization performance"

(More results are provided in our paper)

Experiments: OOD Detection

- NAC-UE outperforms 21 post-hoc detection methods on CIFAR-10, CIAFR-100, and ImageNet benchmarks!

Method	MINIST		SVHN		Textures		Places365		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CIFAR-10 Benchmark										
OpenMax	23.33±4.67	90.50±0.44	25.40±1.47	89.77±0.45	31.50±4.05	89.58±0.60	38.52±2.27	88.63±0.28	29.69±1.21	89.62±0.19
ODIN	23.83±12.34	95.24±1.96	68.61±0.52	84.58±0.77	67.70±11.06	86.94±2.26	70.36±6.96	85.07±1.24	57.62±4.24	87.96±0.61
MDS	27.30±3.55	90.10±2.41	25.96±2.52	91.18±0.47	27.94±4.20	92.69±1.06	47.67±4.54	84.90±2.54	32.22±3.40	89.72±1.36
MDSEns	**1.30**±0.51	**99.17**±0.41	74.34±1.04	66.56±0.58	76.07±0.17	77.40±0.28	94.16±0.33	52.47±0.15	61.47±0.48	73.90±0.27
RMDS	21.49±2.32	93.22±0.80	23.46±1.48	91.84±0.26	25.25±0.53	92.23±0.23	31.20±0.28	91.51±0.11	25.35±0.73	92.20±0.21
Gram	70.30±8.96	72.64±2.34	33.91±17.35	91.52±4.45	94.64±2.71	62.34±8.27	90.49±1.93	60.44±3.41	72.34±6.73	71.73±3.20
ReAct	33.77±18.00	92.81±3.03	50.23±15.98	89.12±3.19	51.42±11.42	89.38±1.49	44.20±3.35	90.35±0.78	44.90±8.37	90.42±1.41
VIM	18.36±1.42	94.76±0.38	19.29±0.41	94.50±0.48	21.14±1.83	95.15±0.34	41.43±2.17	89.49±0.39	25.05±0.52	93.48±0.24
KNN	20.05±1.36	94.26±0.38	22.60±1.26	92.67±0.30	24.06±0.55	93.16±0.24	30.38±0.63	91.77±0.23	24.27±0.40	92.96±0.14
ASH	70.00±10.56	83.16±4.66	83.64±6.48	73.46±6.41	84.59±1.74	77.45±2.39	77.89±7.28	79.89±3.69	79.03±4.22	78.49±2.58
SHE	42.22±20.59	90.43±4.76	62.74±4.01	86.38±1.32	84.60±5.30	81.57±1.21	76.36±5.32	82.89±1.22	66.48±5.98	85.32±1.43
GEN	23.00±7.75	93.83±2.14	28.14±2.59	91.97±0.66	40.74±6.61	90.14±0.76	47.03±3.22	89.46±0.65	34.73±1.58	91.35±0.69
NAC-UE	15.14±2.60	94.86±1.36	**14.33**±1.24	**96.05**±0.47	**17.03**±0.59	**95.64**±0.44	**26.73**±0.80	**91.85**±0.28	**18.31**±0.92	**94.60**±0.50
CIFAR-100 Benchmark										
OpenMax	53.82±4.74	76.01±1.39	53.20±1.78	82.07±1.53	56.12±1.91	80.56±0.09	54.85±1.42	79.29±0.40	54.50±0.68	79.48±0.41
ODIN	45.94±3.29	83.79±1.31	67.41±3.88	74.54±0.76	62.37±2.96	79.33±1.08	59.71±0.92	79.45±0.26	58.86±0.79	79.28±0.21
MDS	71.72±2.94	67.47±0.81	67.21±6.09	70.68±6.40	70.49±2.48	76.26±0.69	79.61±0.34	63.15±0.49	72.26±1.56	69.39±1.39
MDSEns	**2.83**±0.86	**98.21**±0.78	82.57±2.58	53.76±1.63	84.94±0.83	69.75±1.14	96.61±0.17	42.27±0.73	66.74±1.04	66.00±0.69
RMDS	52.05±6.28	79.74±2.49	51.65±3.68	84.89±1.10	53.99±1.06	83.65±0.51	**53.57**±0.43	**83.40**±0.46	52.81±0.63	82.92±0.42
Gram	53.53±7.45	80.71±4.15	**20.06**±1.96	**95.55**±0.60	89.51±2.54	70.79±1.32	94.67±0.60	46.38±1.21	64.44±2.37	73.36±1.08
ReAct	56.04±5.66	78.37±1.59	50.41±2.02	83.01±0.97	55.04±0.82	80.15±0.46	55.30±0.41	80.03±0.11	54.20±1.56	80.39±0.49
VIM	48.32±1.07	81.89±1.02	46.22±5.46	83.14±3.71	46.86±2.29	85.91±0.78	61.57±0.77	75.85±0.37	50.74±1.00	81.70±0.62
KNN	48.58±4.67	82.36±1.52	51.75±3.12	84.15±1.09	53.56±2.32	83.66±0.83	60.70±1.03	79.43±0.47	53.65±0.28	82.40±0.17
ASH	66.58±3.88	77.23±0.46	46.00±2.67	85.60±1.40	61.27±2.74	80.72±0.70	62.95±0.99	78.76±0.16	59.20±2.46	80.58±0.66
SHE	58.78±2.70	76.76±1.07	59.15±7.61	80.97±3.98	73.29±3.22	73.64±1.28	65.24±0.98	76.30±0.51	64.12±2.70	76.92±1.16
GEN	53.92±5.71	78.29±2.05	55.45±2.76	81.41±1.50	61.23±1.40	78.74±0.81	56.25±1.01	80.28±0.27	56.71±1.59	79.68±0.75
NAC-UE	21.97±6.62	93.15±1.63	24.39±4.66	92.40±1.26	**40.65**±1.94	**89.32**±0.55	73.57±1.16	73.05±0.68	**40.14**±1.86	**86.98**±0.37

(More results are provided in our paper)

Experiments: OOD Detection

- The performance of NAC-UE positively correlates with the number of employed layers.

Layer Combinations				CIFAR-10		CIFAR-100		ImageNet	
Layer4	Layer3	Layer2	Layer1	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
✓				23.50	93.21	85.84	58.37	26.89	94.57
✓	✓			21.32	94.35	44.92	85.25	23.51	95.05
✓	✓	✓		18.50	94.46	**39.96**	86.94	22.69	95.23
✓	✓	✓	✓	**18.31**	**94.60**	40.14	**86.98**	**22.49**	**95.29**

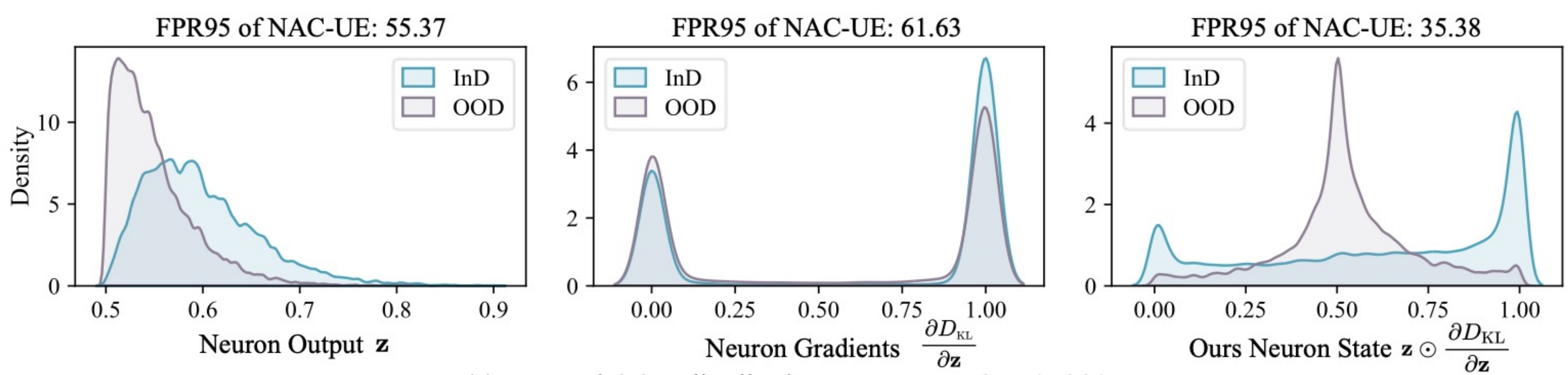
(More results are provided in our paper)

Experiments: OOD Detection

- The performance of NAC-UE positively correlates with the number of employed layers.

Layer Combinations				CIFAR-10		CIFAR-100		ImageNet	
Layer4	Layer3	Layer2	Layer1	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
✓				23.50	93.21	85.84	58.37	26.89	94.57
✓	✓			21.32	94.35	44.92	85.25	23.51	95.05
✓	✓	✓		18.50	94.46	**39.96**	86.94	22.69	95.23
✓	✓	✓	✓	**18.31**	**94.60**	40.14	**86.98**	**22.49**	**95.29**

- Ours neuron state $\mathbf{z} \odot \partial D_{\mathrm{KL}} / \partial \mathbf{z}$ is superior compared to other variants.



(More results are provided in our paper)

Experiments: OOD Generalization

- A positive correlation between NAC-ME and model generalization ability (i.e., OOD test performance) consistently holds.

Bakbone	Method	VLCS		PACS		OfficeHome		TerraInc		Average	
		RC	ACC	RC	ACC	RC	ACC	RC	ACC	RC	ACC
ResNet-18	Oracle	-	77.67	-	80.51	-	56.18	-	44.51	-	64.72
	Validation	34.27	75.12	68.71	**79.01**	83.50	55.60	39.58	37.36	56.52	61.77
	NAC-ME	**50.29**	**75.83**	**74.16**	78.85	**84.91**	**55.76**	**40.42**	**39.45**	**62.45**	**62.47**
	Δ	(+16.02)	(+0.71)	(+5.45)	(-0.16)	(+1.41)	(+0.16)	(+0.84)	(+2.09)	(+5.93)	(+0.70)
ResNet-50	Oracle	-	79.79	-	86.10	-	65.95	-	50.76	-	70.65
	Validation	**31.43**	**77.70**	58.54	84.57	67.93	65.04	37.07	46.07	48.74	68.34
	NAC-ME	28.68	76.41	**62.07**	**85.28**	**69.16**	**65.23**	**40.16**	**47.10**	**50.02**	**68.51**
	Δ	(-2.75)	(-1.29)	(+3.53)	(+0.71)	(+1.23)	(+0.19)	(+3.09)	(+1.03)	(+1.28)	(+0.17)
Vit-t16	Oracle	-	79.11	-	71.99	-	61.44	-	41.29	-	63.46
	Validation	37.95	77.43	89.34	69.83	98.71	61.22	22.71	36.28	62.18	61.19
	NAC-ME	**49.59**	**77.97**	**90.67**	**70.99**	**99.14**	**61.26**	**23.26**	**36.69**	**65.67**	**61.73**
	Δ	(+11.64)	(+0.54)	(+1.33)	(+1.16)	(+0.43)	(+0.04)	(+0.55)	(+0.41)	(+3.49)	(+0.54)
Vit-b16	Oracle	-	80.96	-	90.23	-	81.23	-	52.23	-	76.16
	Validation	18.81	78.70	41.38	87.80	58.29	80.11	0.92	45.49	29.85	73.03
	NAC-ME	**37.42**	**79.20**	**45.04**	**88.83**	**63.17**	**80.52**	**20.22**	**47.86**	**41.46**	**74.10**
	Δ	(+18.61)	(+0.50)	(+3.66)	(+1.03)	(+4.88)	(+0.41)	(+19.30)	(+2.37)	(+11.61)	(+1.07)

Thank you!

Paper

Code