Pursuing Feature Separation based on Neural Collapse for Out-of-Distribution Detection

Yingwen Wu, Ruiji Yu, Xinwen Cheng, Zhengbao He, Xiaolin Huang

Institute of Image Processing and Pattern Recognition

Shanghai Jiao Tong University

Shanghai, China

Background

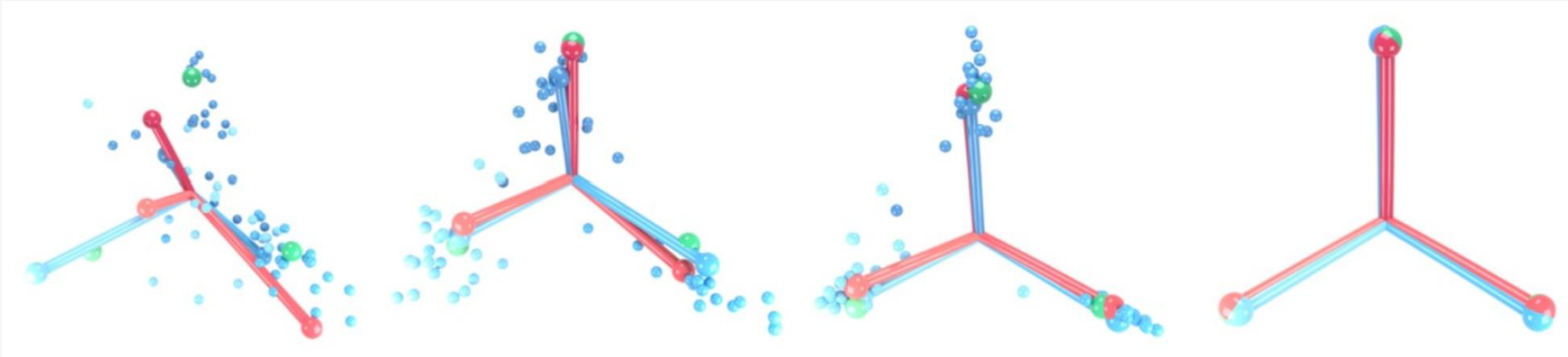
DNNs cannot correctly classify out-of-distribution data, whose distribution is different from training data

Utilizing auxiliary OOD data to finetune the model improves detection performance

Existing methods focus on enlarging the model output difference between ID and OOD data

Can we enlarge the feature difference between ID and OOD data?

Principal subspace of ID features: ID features within a class are nearly identical to the FC weight of the corresponding class



Neural Collapse phenomenon[1]

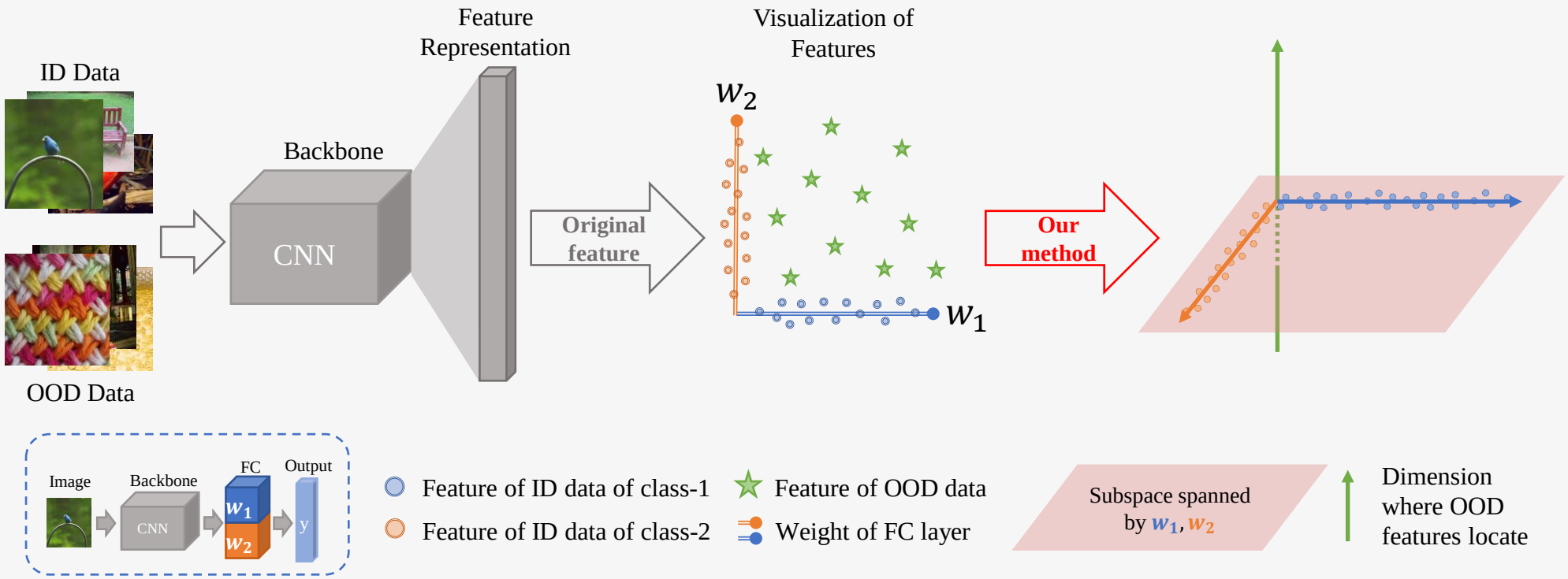
Goal: locating ID features in low-dimensional subspace while OOD features in the rest dimensions.

[1] Papyan V, Han X Y, Donoho D L. Prevalence of neural collapse during the terminal phase of deep learning training[J]. Proceedings of the National Academy of Sciences, 2020

Feature Separation Method

Existing works pay attention to the output separation in model finetuning, e.g. Outlier Exposure[2] $L_{OE} = -\sum_{j=1}^{c} \log f^j(x)$

Difficulty of feature separation: hard to describe feature distribution; pair-distance too expensive to calculate



Loss function design:

$$z_{OOD}^T w_i = 0, i = 1,2,\dots,C$$

$$z_{ID}^T w_y = 1, y \text{ is true label}$$

$$L_{Sep} = \frac{1}{C}\sum_{i=1}^{C} |z^T w_i|$$

$$L_{Clu} = -z_{ID}^T w_y$$

[2] Hendrycks D, Mazeika M, Dietterich T. Deep anomaly detection with outlier exposure[J]. arXiv preprint arXiv:1812.04606, 2018.

Feature Separation Performance

Final optimization function:

$$\min \mathbb{E}_{(x,y)\sim D_{in}}(L_{CE} + \alpha L_{Clu}) + \mathbb{E}_{x\sim D_{out}^{aux}}(\lambda L_{OE} + \beta L_{Sep})$$

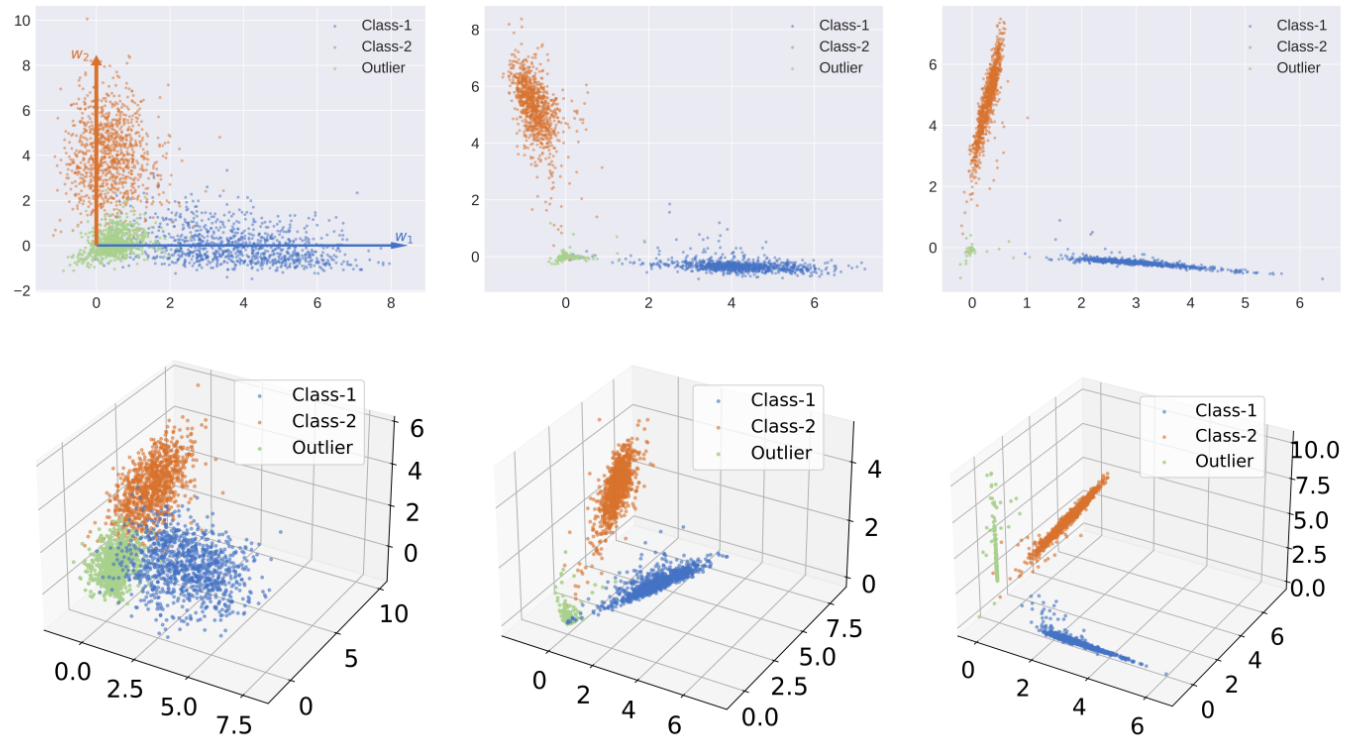


Figure 3.2 Visualization of features in the linear layer parameter space, where feature separability between ID and OOD data gradually increases from left (Vanilla model) to right (Our model).

- **Left: vanilla model**, feature entanglement
- **Middle: OE model**, larger feature distance, but not use z-axis dimension
- **Right: our model**, the largest feature distance, utilize z-axis dimension

Experiment

Setup

- **3 different modes trained on CIFAR10** :

  WideResNet-40-2, ResNet18, DenseNet121

- **2 different modes trained on ImageNet-1K**:

  ResNet50, ViT-B-16

- **11 compared OOD detection methods**:

  MSP, Energy, Maha, KNN, CSI, CIDER, KNN+,

  OE, Energy-OE, POEM, DAL

Result

- **consistently superior performance on CIFAR10, CIFAR100, and ImageNet benchmarks**

Table 3.2 Detection performance on CIFAR10 and CIFAR100 benchmarks

Method	Far-OOD Datasets											
	SVHN		LSUN		iSUN		Textures		Places365		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CIFAR-10												
With vanilla training												
MSP[1]	44.22	93.61	27.56	96.12	69.62	85.29	60.02	88.53	65.68	86.25	53.42	89.96
Energy[10]	31.81	94.65	4.6	98.96	50.06	89.75	49.68	90.09	42.28	90.82	35.69	92.85
Maha[36]	42.67	90.71	18.96	96.46	28.86	93.76	26.22	92.81	86.78	69.14	40.70	88.58
KNN[52]	44.76	92.55	27.38	95.34	43.84	91.24	37.64	92.82	49.23	87.89	40.57	91.97

Table 3.1 Detection performance on ImageNet-1k benchmark

Model	Method	Far-OOD Datasets				Near-OOD Datasets				Average		ID Acc↑
		iNaturalist		Textures		SUN		Places				
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	
ResNet50	OE[72]	48.30	88.91	58.60	82.78	61.40	83.09	70.36	80.78	59.66	83.89	76.04
	DAL[76]	47.92	89.12	57.91	83.02	61.20	83.22	70.55	80.79	59.39	84.04	75.94
	Ours	43.01	90.17	55.35	83.45	60.11	83.56	68.46	81.31	**56.73**	**84.62**	**76.10**
ViT-B-16	OE[72]	41.96	90.49	52.25	85.97	65.61	82.30	70.20	80.93	57.51	84.92	80.05
	DAL[76]	40.52	90.92	50.94	86.20	65.07	82.39	70.17	80.96	**56.67**	85.12	80.06
	Ours	40.10	91.06	51.70	86.13	65.58	82.25	70.12	81.07	56.88	**85.13**	**80.29**
Energy[10]		70.18	87.15	17.15	97.05	91.37	65.50	84.77	76.72	78.91	75.77	62.75	80.44
Maha[36]		77.73	78.01	98.46	63.44	47.74	88.76	54.93	82.53	97.22	54.11	75.22	73.37
KNN[52]		71.86	83.31	78.89	70.09	79.60	70.86	72.89	80.05	80.91	71.33	76.83	75.13
With contrastive learning													
CSI*[183]		64.50	84.62	25.88	95.93	70.62	80.83	61.50	86.74	83.08	77.11	61.12	95.05
CIDER[77]		16.47	96.23	45.45	81.64	66.01	82.21	49.79	87.48	82.66	68.39	52.08	83.19
KNN+*[52]		32.50	93.86	47.41	84.93	39.82	91.12	43.05	88.55	63.26	79.28	45.20	87.55
With auxiliary OOD data													
OE[72]		38.70	92.90	18.30	96.67	36.35	92.59	43.05	91.00	52.45	87.86	37.77	92.21
Energy-OE[10]		17.75	96.94	34.00	94.82	60.75	87.32	45.70	90.09	53.50	89.08	42.34	91.65
POEM[73]		45.41	90.70	3.01	99.24	18.60	95.79	51.37	83.85	84.13	73.93	40.5	88.87
DAL[76]		16.45	96.10	17.00	96.52	36.95	90.88	38.40	91.72	48.55	88.91	31.47	92.82
Ours		17.95	96.52	12.50	97.64	27.00	93.85	41.70	91.37	48.20	90.64	**29.47**	**94.00**

Experiment

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

- Near-OOD setting

Table 3.9 Near-OOD detection performance on CIFAR10 benchmark

Method	Near-OOD Datasets							
	LSUN-Fix		ImageNet-Resize		CIFAR-100		Tiny-ImageNet	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
With contrastive learning								
CSI*	39.79	93.63	37.47	93.93	45.64	87.64	-	-
CIDER	8.98	98.56	43.45	93.82	55.84	90.0	-	-
KNN+*	24.88	95.75	30.52	94.85	40.00	89.11	-	-
With auxiliary OOD data								
OE	1.00	99.53	7.20	98.48	25.05	**94.86**	19.55	91.49
DAL	**0.65**	**99.59**	**3.75**	**98.63**	26.00	94.35	20.75	92.18
Ours	0.75	99.07	4.65	98.42	**24.60**	94.69	**17.65**	**92.48**

- Contribution of each loss

$$\min \mathbb{E}_{(x,y)\sim D_{in}}(L_{CE} + \alpha L_{Clu}) + \mathbb{E}_{x\sim D_{out}^{aux}}(\lambda L_{OE} + \beta L_{Sep})$$

Table 3.12 Performance under different training losses

No.	Training Loss	CIFAR10		CIFAR100	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑
1	L_{OE}	3.36	**99.02**	37.77	92.21
2	L_{OE}+L_{Clu}	3.62	98.96	39.91	91.22
3	L_{OE}+L_{Sep}	2.65	99.00	33.30	93.42
4	L_{OE}+L_{Clu}+L_{Sep}	**2.49**	98.93	**29.47**	**94.00**

- Versatility of method

Table 3.13 Results on imbalanced CIFAR10 data

Method	SVHN		LSUN		iSUN		Textures		Places365		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
OE[71]	15.15	96.56	12.00	97.25	19.70	96.63	18.05	96.15	29.75	93.63	18.93	96.04
Ours	14.05	96.38	12.05	96.93	13.55	97.38	24.05	95.63	24.95	94.07	**17.73**	**96.08**

SJTU

Conclusion

A pioneering study explores pursuing ID-OOD feature separation in model-finetuning based detection

- Feature separation based on low-dimensionality, avoiding complex distribution modelling
- Novel loss function based on feature instead of output, serving as a better baseline
- Extensive experiments on different models and datasets

上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

Thanks!

饮 水 思 源　爱 国 荣 校