南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Out-of-Distribution Detection in Long-Tailed Recognition with Calibrated Outlier Class Learning

Wenjun Miao[1], Guansong Pang[2*], Xiao Bai[1,3], Tianqi Li[1], Jin Zheng[1, 4*]

[1]School of Computer Science and Engineering, Beihang University
[2]School of Computing and Information Systems, Singapore Management University
[3]State Key Laboratory of Software Development Environment, Jiangxi Research Institute, Beihang University
[4]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
{miaowenjun, jinzheng, baixiao, tianqili}@buaa.edu.cn, gspang@smu.edu.sg

AAAI 2024

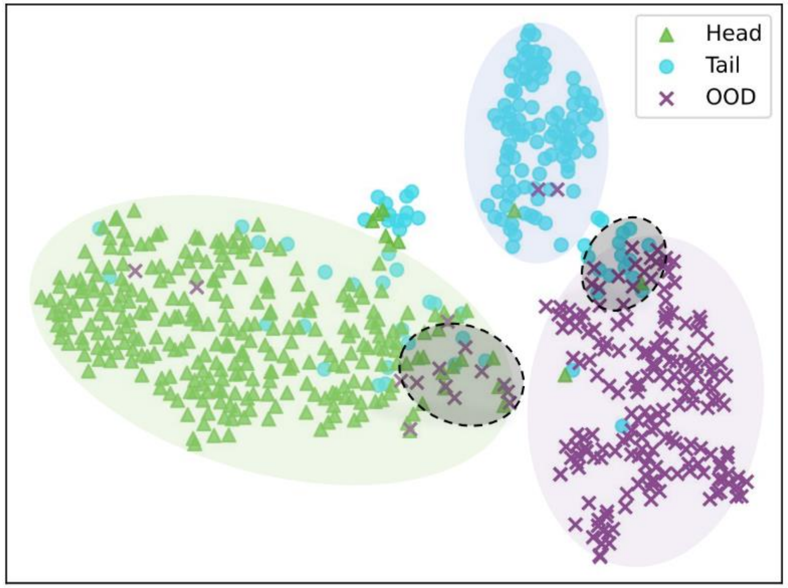
Introduction

DNNs often have high confidence predictions that classify OOD samples from unknown classes as one of the known classes.

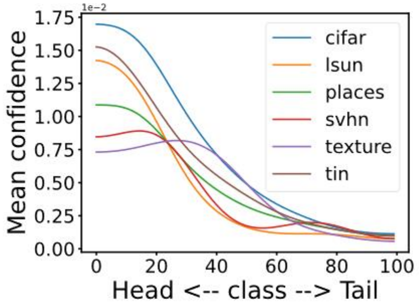
This issue is further amplified when long-tailed distribution

The models often misclassify OOD samples into head classes with high confidence.

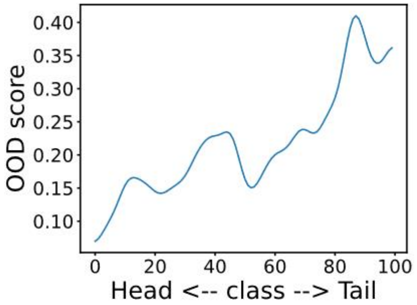
The tail samples often have a much higher OOD score than the head samples



(a) Feature representations of CIFAR100-LT test data



(b) Prediction confidence



(c) OOD score for ID samples

Related Work

OOD Detection

1. The post hoc methods: devising new OOD scoring functions in the inference phrase.
2. Utilizing auxiliary data during training: OE ([ICLR2018]Deep Anomaly Detection With Outlier Exposure)

$$\mathcal{L}_{OE} = \mathbb{E}_{x,y\sim\mathcal{D}_{in}}[\ell(f(x),y] + \gamma\mathbb{E}_{x\sim\mathcal{D}_{out}}[\ell(f(x),u],$$

Focused on balanced ID training data.

OOD Detection in LTR

1. Utilizing unlabeled auxiliary OOD data to enhance the robustness of OOD detection and improve ID classification accuracy.
2. Fitting the prediction probability of OOD data to a long-tailed distribution. It is difficult to obtain such an accurate prior distribution of OOD data in LTR.

We instead utilize the outlier class learning to eliminate the need for such a prior.

$$\mathcal{L}_{OCL} = \mathbb{E}_{x,y\sim\mathcal{D}_{in}}[\ell(f(x),y] + \gamma\mathbb{E}_{x\sim\mathcal{D}_{out}}[\ell(f(x),\tilde{y}],$$

Related Work

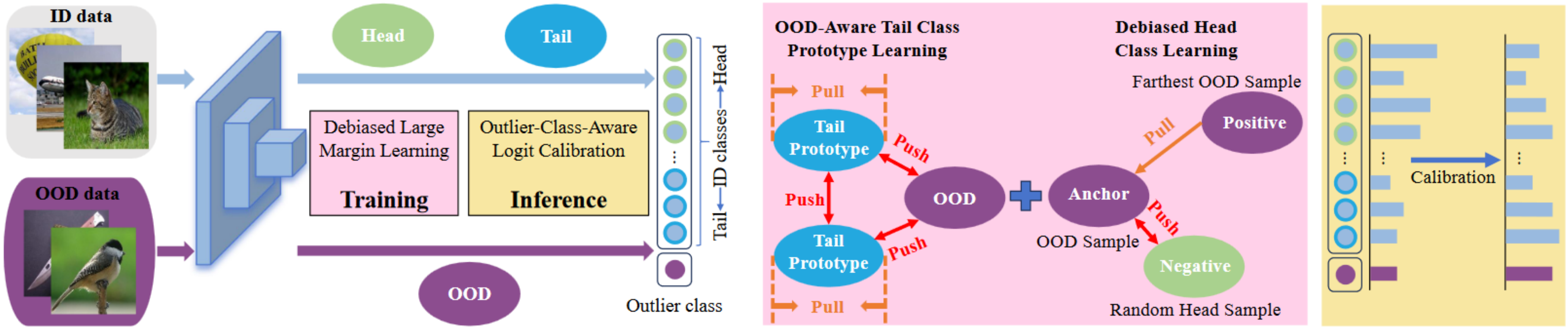
$$\mathcal{L}_{OE} = \mathbb{E}_{x,y\sim\mathcal{D}_{in}}[\ell(f(x), y] + \gamma\mathbb{E}_{x\sim\mathcal{D}_{out}}[\ell(f(x), u], \quad (1)$$

$$\mathcal{L}_{OCL} = \mathbb{E}_{x,y\sim\mathcal{D}_{in}}[\ell(f(x), y] + \gamma\mathbb{E}_{x\sim\mathcal{D}_{out}}[\ell(f(x), \tilde{y}], \quad (2)$$

OE achieves promising performance in general OOD detection scenarios, but works less effectively when applied to LTR settings. It is mainly because the uniform prediction probability prior in Eq.1 does not hold in LTR.

- OOD -> head classes
- Tail classes -> OOD

Method



(a) Pipeline (b) Debiased large margin learning (c) Outlier-class-aware logit calibration

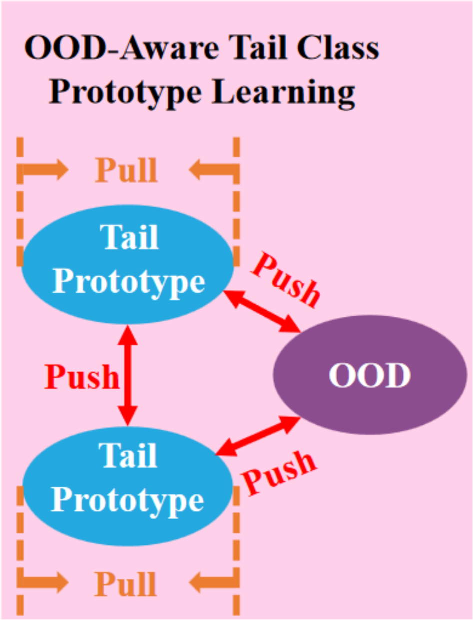
Problem: The tail classes samples tend to exhibit high OOD scores during LTR inference.

Previous work: attempts to utilize diverse augmentations to push tail samples away from OOD samples, but it often learns non-discriminative representations between OOD samples and tail samples due to the limited size of tail classes.

Method: a learnable prototype of one tail class as positive sample to pull tail samples closer to their prototype.

$$\mathcal{L}_t = \mathbb{E}_{x\sim\mathcal{D}_{tail}}[\mathcal{L}_t(x,\mathcal{M})], \qquad \mathcal{M} \in \mathbb{R}^{N\times D}$$

$$\mathcal{L}_t(x,\mathcal{M}) = \frac{1}{|\mathcal{B}|}\sum_{x\in\mathcal{B}} log \frac{exp(z(x)m_x^{\intercal}/t)}{\sum_{m\in\mathcal{M}} exp(z(x)m^{\intercal}/t) + P(x)}, \quad P(x) = \sum_{\hat{x}\in\mathcal{O}} exp(z(x)z(\hat{x})^{\intercal}/t)$$



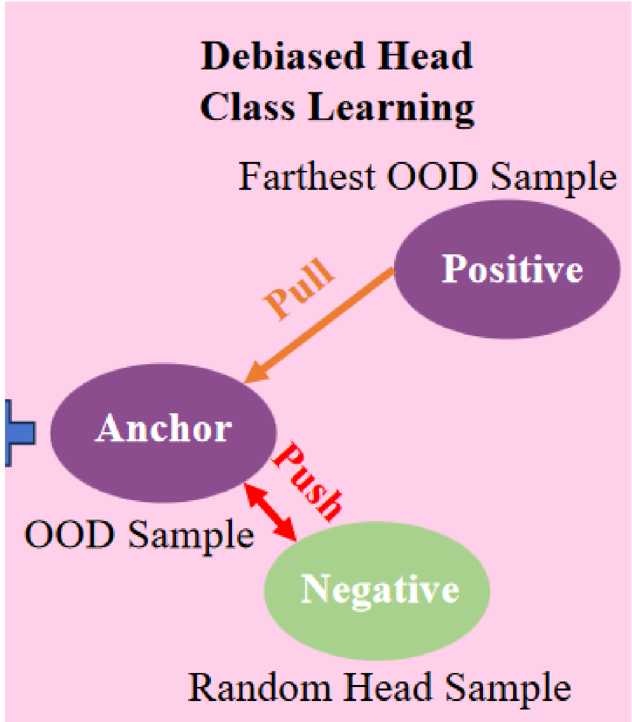
Method/Debiased Large Margin Learning

Problem: Due to the overwhelming presence of head class samples, LTR models demonstrate a strong bias towards head classes when performing OOD detection.

Method: the OOD samples as anchor, with randomly sampled head samples as negative samples and the OOD samples that are distant from the anchors in the feature space as positive samples.

$$\mathcal{L}_h = \mathbb{E}_{x\sim\mathcal{D}_{out}}[\mathcal{L}_h(x)],$$

$$\mathcal{L}_h(x) = \frac{1}{|\mathcal{B}|}\sum_{x\in\mathcal{B}} max(0, \|z(x) - z(x^p)\|_2^2 - \|z(x) - z(x^n)\|_2^2 + margin),$$



$$\begin{aligned}\mathcal{L}_{total} &= \mathcal{L}_{OCL} + \alpha\mathcal{L}_t + \beta\mathcal{L}_h \\ &= \mathbb{E}_{x,y\sim\mathcal{D}_{in}}[\ell(f(x), y] + \gamma\mathbb{E}_{x\sim\mathcal{D}_{out}}[\ell(f(x), \tilde{y}] \\ &+ \alpha\mathbb{E}_{x\sim\mathcal{D}_{tail}}[\mathcal{L}_t(x, \mathcal{M})] + \beta\mathbb{E}_{x\sim\mathcal{D}_{out}}[\mathcal{L}_h(x)],\end{aligned}$$

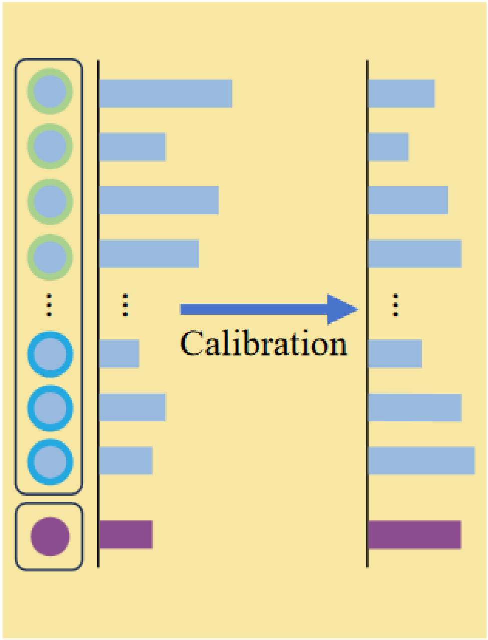
Problem: due to the inherent class imbalance in the training data, the LTR model often tends to have a higher confidence on the prediction of head samples than both the tail samples and the OOD samples.

$$P(y=i|x)=\frac{e^{f_i(x)-\tau\cdot\log n_i}}{\sum_{j=1}^{k+1} e^{f_j(x)-\tau\cdot\log n_j}}, \quad n_i=\frac{N_i}{N_1+N_2+\cdots+N_k},$$

We do not have genuine OOD samples during training and OOD samples should be as important as ID classification

$$n_{k+1}=1$$

decrease the probability of head classes and increase that of tail classes, while taking into account the influence of OOD samples on the prediction.



(c) Outlier-class-aware logit calibration

Experiment

OOD	Method	AUC↑	AP-in↑	AP-out↑	FPR↓
Texture	OE	92.30	96.01	82.57	48.65
	OCL	93.71	95.95	91.07	27.22
	COCL	**96.81**	**98.21**	**93.86**	**14.65**
SVHN	OE	94.86	91.59	97.00	29.11
	OCL	95.14	90.88	97.73	25.47
	COCL	**96.98**	**93.25**	**98.61**	**12.59**
CIFAR100	OE	83.32	84.06	80.83	65.82
	OCL	82.04	82.52	81.92	63.35
	COCL	**86.63**	**86.66**	**86.28**	**52.21**
Tiny ImageNet	OE	86.35	89.88	79.30	64.50
	OCL	85.90	88.98	82.17	57.46
	COCL	**90.43**	**92.52**	**87.03**	**46.12**
LSUN	OE	91.57	93.06	88.37	53.99
	OCL	92.75	92.69	93.10	30.95
	COCL	**94.85**	**95.43**	**93.98**	**27.48**
Place365	OE	90.20	82.09	95.24	57.06
	OCL	89.91	77.91	96.28	42.33
	COCL	**93.97**	**87.36**	**97.56**	**32.25**

(a) Comparison of COCL with OE and OCL on six OOD datasets.

Method	AUC↑	AP-in↑	AP-out↑	FPR↓	ACC↑
MSP	74.33	73.96	72.14	85.33	72.17
OE	89.76	89.45	87.22	53.19	73.59
EnergyOE	91.92	91.03	91.97	33.80	74.57
OCL	89.91	88.15	90.38	41.13	74.48
PASCL	90.99	90.56	89.24	42.90	77.08
OS	91.94	91.08	89.35	36.92	75.78
Class Prior	92.08	91.17	90.86	34.42	74.33
BERL	92.56	91.41	91.94	32.83	81.37
COCL	**93.28**	**92.24**	**92.89**	**30.88**	**81.56**

(b) Comparison results with different competing methods. The results are averaged over the six OOD test datasets in (a).

Table 2: Comparison results on CIFAR10-LT.

OOD	Method	AUC↑	AP-in↑	AP-out↑	FPR↓
Texture	OE	76.01	85.28	57.47	87.45
	OCL	75.92	82.99	66.48	70.01
	COCL	**81.99**	**88.05**	**74.38**	**59.79**
SVHN	OE	81.82	73.25	89.10	80.98
	OCL	78.64	69.21	86.26	86.38
	COCL	**89.20**	**81.57**	**94.21**	**54.46**
CIFAR10	OE	**62.60**	**66.16**	**57.77**	**93.53**
	OCL	60.29	63.21	55.71	94.22
	COCL	62.05	66.14	56.82	93.88
Tiny ImageNet	OE	68.22	79.36	51.82	88.54
	OCL	69.56	79.97	54.47	85.91
	COCL	**71.87**	**81.89**	**57.12**	**83.93**
LSUN	OE	76.81	85.33	60.94	83.79
	OCL	79.14	86.56	66.58	75.07
	COCL	**84.10**	**89.89**	**69.80**	**74.67**
Place365	OE	75.68	60.99	86.51	83.55
	OCL	77.81	62.80	88.39	79.97
	COCL	**80.30**	**68.65**	**89.16**	**77.83**

(a) Comparison of COCL to OE and OCL on six OOD datasets.

Method	AUC↑	AP-in↑	AP-out↑	FPR↓	ACC↑
MSP	63.93	64.71	60.76	89.71	40.51
OE	73.52	75.06	67.27	86.30	39.42
EnergyOE	76.40	77.32	72.24	76.33	41.32
OCL	73.56	74.12	69.65	81.93	41.54
PASCL	73.32	74.84	67.18	79.38	43.10
OS	74.37	75.80	70.42	78.18	40.87
Class Prior	76.03	77.31	72.26	76.43	40.77
BERL	77.75	78.61	73.10	74.86	45.88
COCL	**78.25**	**79.37**	**73.58**	**74.09**	**46.41**

(b) Comparison results with different competing methods. The results are averaged over the six OOD test datasets in (a).

Table 3: Comparison results on CIFAR100-LT.

Method	AUC↑	AP-in↑	AP-out↑	FPR↓	ACC↑
MSP	55.78	35.60	74.18	94.01	45.36
OE	68.33	43.87	82.54	90.98	44.00
EnergyOE	69.43	45.12	84.75	76.89	44.42
OCL	68.67	43.11	84.15	77.46	44.77
PASCL	68.00	43.32	82.69	82.28	47.29
OS	69.23	44.21	84.12	79.37	45.73
Class Prior	70.43	45.26	84.82	77.63	46.83
BERL	71.16	45.97	85.63	76.98	50.42
COCL	**71.85**	**46.76**	**86.21**	**75.60**	**51.11**

Table 4: Comparison results on ImageNet-LT with ImageNet-1k-OOD as OOD test dataset.

Metric	CIFAR10-LT			CIFAR100-LT		
	OE	OCL	COCL	OE	OCL	COCL
AUC↑	82.60	84.84	**91.91**	64.08	66.11	**74.85**
AP-in↑	60.47	61.56	**76.98**	34.07	34.97	**47.76**
AP-out↑	92.28	94.75	**97.15**	83.19	85.74	**87.59**
FPR↓	72.10	52.73	**34.30**	92.48	82.53	**77.01**

(a) On separating tail samples from OOD data.

Metric	CIFAR10-LT			CIFAR100-LT		
	OE	OCL	COCL	OE	OCL	COCL
AUC↑	95.97	95.79	**96.34**	84.42	83.85	**87.73**
AP-in↑	91.09	88.72	**93.34**	70.16	68.44	**73.84**
AP-out↑	98.17	98.54	**98.67**	92.85	92.83	**93.94**
FPR↓	20.57	22.67	**19.59**	70.17	67.94	**66.01**

(b) On separating head samples from OOD data.

Table 5: Comparison results on separating tail/head samples from OOD samples. The results are averaged over six OOD test datasets in the SC-OOD benchmark.

the difficulty of learning the outlier class given the similarity between these two datasets

differentiating tail and OOD samples is often more difficult than differentiating head and OOD samples

Experiment/Ablation Study

ID Dataset	TCPL	DHCL	OLC	AUC↑	AP-in↑	AP-out↑	FPR↓	ACC↑	ACC-t↑
CIFAR10-LT	Baseline (OE)			89.76	89.45	87.22	53.19	73.59	55.91
	✗	✗	✗	89.91	88.15	90.38	41.13	74.48	56.52
	✓	✗	✗	91.23	89.47	91.51	34.27	74.58	57.10
	✗	✓	✗	91.08	89.40	91.10	35.28	74.61	56.92
	✗	✗	✓	92.06	91.29	91.78	34.41	79.40	76.57
	✓	✓	✗	91.74	89.91	92.04	33.85	75.20	57.30
	✓	✓	✓	**93.28**	**92.24**	**92.89**	**30.88**	**81.56**	**77.90**
CIFAR100-LT	Baseline (OE)			73.52	75.06	67.27	86.30	39.42	12.59
	✗	✗	✗	73.56	74.12	69.65	81.93	41.54	12.06
	✓	✗	✗	75.14	75.74	71.25	78.39	41.93	13.53
	✗	✓	✗	74.70	75.36	70.63	78.96	42.42	13.33
	✗	✗	✓	75.51	75.83	71.66	77.57	45.62	28.44
	✓	✓	✗	76.09	76.59	71.92	76.20	42.46	13.89
	✓	✓	✓	**78.25**	**79.37**	**73.58**	**74.09**	**46.41**	**29.44**
ImageNet-LT	Baseline (OE)			68.33	43.87	82.54	90.98	44.00	7.65
	✗	✗	✗	68.67	43.11	84.15	77.46	44.77	8.02
	✓	✗	✗	70.08	44.68	85.04	76.61	44.59	8.49
	✗	✓	✗	69.64	44.11	84.83	76.62	45.00	8.43
	✗	✗	✓	70.37	45.07	85.35	76.31	50.16	26.03
	✓	✓	✗	70.78	45.19	85.61	76.26	45.24	9.92
	✓	✓	✓	**71.85**	**46.76**	**86.21**	**75.60**	**51.11**	**28.05**

Table 6: Ablation study results on CIFAR10-LT, CIFAR100-LT and ImageNet-LT.

- Tail Class Prototype Learning (TCPL)
- Debiased Head Class Learning (DHCL)
- Outlier-Class-Aware Logit Calibration (OLC)

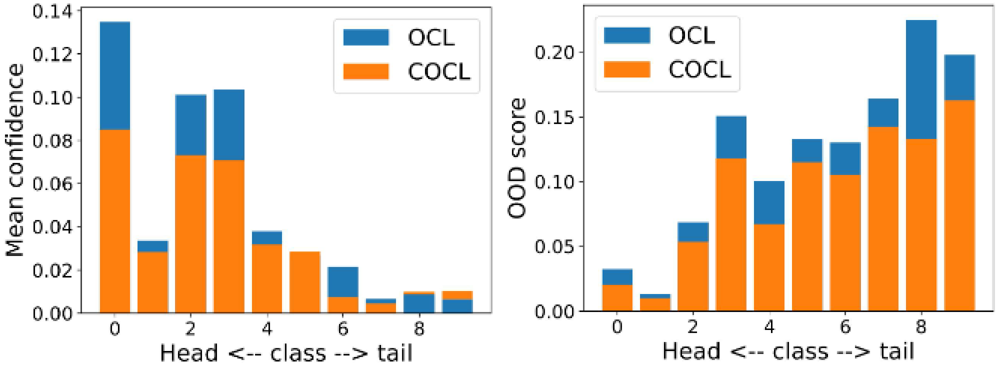


Figure 3: Results on CIFAR10-LT. (Left) The mean prediction confidence of six OOD datasets belonging to each ID class. (Right) The mean OOD score for each ID class.

Thank you