NEURAL INFORMATION PROCESSING SYSTEMS

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

中国科学院
THE CHINESE ACADEMY OF SCIENCES

Revisiting Logit Distributions for Reliable Out-of-Distribution Detection

Jiachen Liang[1,2], Ruibing Hou[1*], Minyang Hu[1,2],
Hong Chang[1,2], Shiguang Shan[1,2], Xilin Chen[1,2]

[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[*]Corresponding Author

Background

Out-of-Distribution Detection (OOD Detection)

Problem Setting

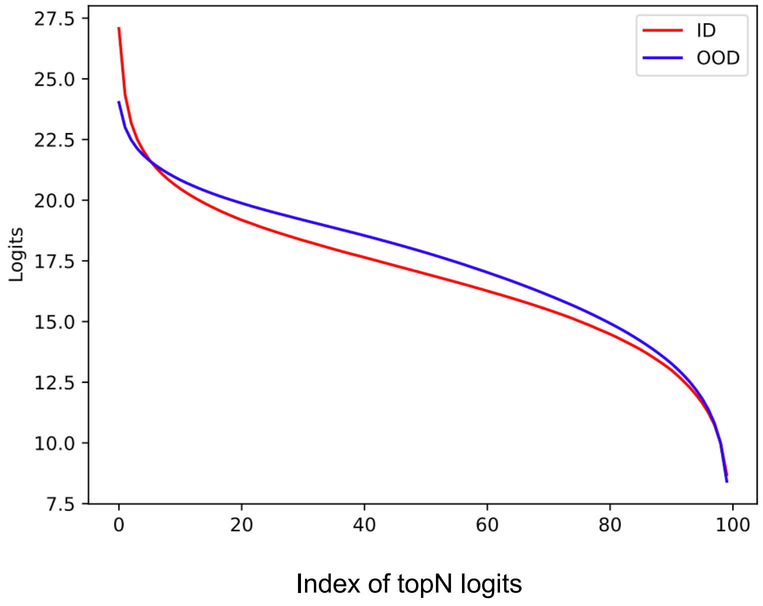
- Let $\mathcal{Y}_{in}$ denotes the set of known classes (in-distribution, ID) for a given classification task
- Given a test sample x, if its label $y \notin \mathcal{Y}_{in}$, then x is considered as an OOD sample
- Formally, fox sample x, a decision function $D: \mathcal{X} \rightarrow \{ID, OOD\}$ is constructed with a score function S as

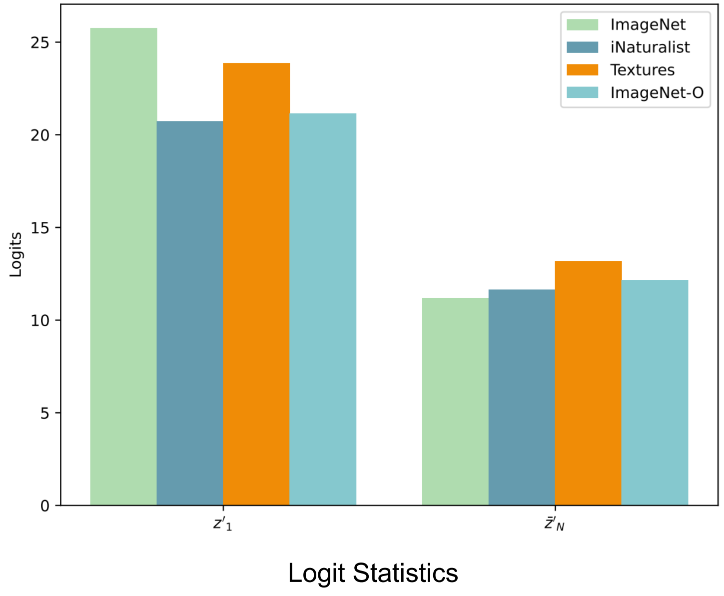
$$D(x) = \begin{cases} ID, if\ S(x; f) \geq \lambda \\ OOD, if\ S(x; f) < \lambda \end{cases}$$

How can we more effectively *leverage the discriminative information* to maximize the separability between ID and OOD samples?

Motivation

- **Observations:** different logits patterns for ID/OOD samples
  - ID samples tend to produce sharper, more peaked logit distributions
  - OOD samples often yield flatter, more uniform logit distributions





Our Idea: Exploit the distributional differences between ID and OOD logits

Method

- **LogitGap**
  - Measure the average difference between the maximum logit and other logits

$$S_{\text{LogitGap}}(x; f) = \frac{1}{K-1}\sum_{j=2}^{K}(z_1' - z_j')$$

- **LogitGap-topN**
  - Focus on a more compact and informative subset of logit

$$S_{\text{LogitGap-topN}}(x; f) = \frac{1}{K-1}\sum_{j=2}^{N}(z_1' - z_j')$$

  - Logits Selection

$$\begin{aligned}&\arg\max_{N}\left(\mathbb{E}_{\boldsymbol{x}\sim P_{\text{ID}}}\left[S_{\text{LogitGap-topN}}(\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{x}\sim P_{\text{OOD}}}\left[S_{\text{LogitGap-topN}}(\boldsymbol{x})\right]\right)\\ =\,&\arg\max_{N}\left(\mathbb{E}_{\boldsymbol{x}\sim P_{\text{OOD}}}[\bar{z}'_N] - \mathbb{E}_{\boldsymbol{x}\sim P_{\text{ID}}}[\bar{z}'_N]\right),\end{aligned}$$

Method

Theoretical Support

From the respective of FPR

- $FPR_{\text{MCM}}(\tau, \lambda_{\text{MCM}}) = Q_x(\frac{e^{z'_1/\tau}}{\sum_{j=1}^{K} e^{z'_j/\tau}} > \lambda_{\text{MCM}})$
- $FPR_{\text{LogitGap}}(\tau, \lambda_L) = Q_x(\frac{\sum_{j=1}^{K}(z'_1 - z'_j)}{\tau K} > \frac{\lambda_{\text{L}}}{\tau})$

In Short: When the temperature $\tau > 2(K-1)$, FPR of LogitGap is guaranteed to smaller than that of MCM.

Theorem 4.1. *Given a K-way classification task, let the predicted logit vector for a sample $\boldsymbol{x}$ be $\boldsymbol{z} = [z_1, z_2, \ldots, z_K]$. Let $\boldsymbol{z}' = [z'_1, z'_2, \ldots, z'_K]$ denote the sorted logits in descending order, such that $z'_1 = \max_k z_k$. Then, for the temperature scaling parameter τ used in the softmax function, if $\tau > 2(K-1)$, we have*

$$\mathrm{FPR}_{\mathrm{LogitGap}}(\lambda_L) \leq \mathrm{FPR}_{\mathrm{MCM}}(\tau, \lambda_{MCM}),$$

where $\mathrm{FPR}_{\mathrm{LogitGap}}(\lambda_L)$ is the false positive rate based on LogitGap with threshold λ_L. Similarly, $\mathrm{FPR}_{\mathrm{MCM}}(\tau, \lambda_{MCM})$ is the false positive rate based on MCM with temperature τ and threshold λ_{MCM}.

Experiments

Setting

- Datasets

ID Dataset	Detection Senario	OOD Dataset		
ImageNet / ImageNet100	Semantic Shift	NINCO	ImageNet-O	ImageNetOOD
	Covariate Shift	ImageNet-A	ImageNet-R	ImageNet-Sketch

- Evaluation Paradigms
  - OOD Detection based on Vision-Language Model
    - Zero-shot Setting
    - Few-shot Setting
  - Traditional OOD Detection

Experiments

□ Results

- **Semantic** Shift OOD Detection based on VLM

Method	NINCO FPR95 ↓	NINCO AUROC ↑	NINCO AUPR ↑	OOD Dataset ImageNet-O FPR95 ↓	ImageNet-O AUROC ↑	ImageNet-O AUPR ↑	ImageNetOOD FPR95 ↓	ImageNetOOD AUROC ↑	ImageNetOOD AUPR ↑	AVG FPR95 ↓	AVG AUROC ↑	AVG AUPR ↑
Zero-shot												
Energy [30]	84.11	72.04	95.65	81.60	75.58	98.66	79.12	76.77	83.96	81.61	74.8	92.76
+TAG [31]	83.11	71.15	95.46	79.65	77.10	98.79	78.34	78.03	85.81	80.37	75.43	93.35
MCM [33]	79.67	73.59	95.87	75.85	79.52	98.93	80.98	78.33	85.42	78.83	77.15	93.41
+TAG [31]	81.63	71.33	95.41	77.60	79.95	98.97	82.99	78.79	86.42	80.74	76.69	93.6
MaxLogit [14]	79.41	74.35	96.03	77.15	77.85	98.79	75.85	78.67	85.16	77.47	76.96	93.33
+TAG [31]	77.70	73.92	95.95	74.50	79.48	98.92	**75.17**	80.07	**87.01**	75.79	77.82	**93.96**
GL-MCM	**74.38**	76.03	96.26	72.35	79.50	98.88	79.16	77.31	83.67	75.30	74.74	86.23
LogitGap	76.83	76.43	96.37	72.35	81.32	**99.03**	76.37	79.95	86.06	75.18	79.23	93.82
LogitGap*	77.42	**76.51**	**96.38**	**71.95**	**81.45**	**99.03**	75.40	**80.27**	86.21	**74.92**	**79.41**	93.87
One-shot												
CoOp [62]	84.01	68.83	94.91	75.55	78.63	98.84	82.25	76.88	84.20	80.60	74.78	92.65
+LogitGap*	**82.85**	**73.57**	**95.91**	**74.85**	**79.40**	**98.88**	**78.31**	**78.09**	**84.55**	**78.67**	**77.02**	**93.11**
ID-Like [1]	73.02	75.83	95.86	80.95	69.81	98.09	83.23	68.57	75.22	79.07	71.40	89.72
+LogitGap*	**68.78**	**80.90**	**97.07**	**71.20**	**78.21**	**98.76**	**75.06**	**76.33**	**82.40**	**71.68**	**78.48**	**92.74**
Four-shot												
CoOp [62]	81.46	70.16	95.14	75.40	79.80	98.93	80.40	78.56	85.56	79.09	76.17	93.21
+LogitGap*	**80.69**	**73.97**	**95.95**	**72.70**	**81.10**	**99.01**	**76.07**	**80.16**	**86.42**	**76.49**	**78.41**	**93.79**
ID-Like [1]	81.15	71.79	95.01	82.50	68.41	97.92	88.54	62.40	70.25	84.06	67.53	87.73
+LogitGap*	**77.51**	**75.39**	**95.91**	**78.10**	**75.29**	**98.57**	**82.65**	**72.88**	**79.85**	**79.42**	**74.52**	**91.44**

Zero-shot

Few-shot

Experiments

Results

- **Covariate** Shift OOD Detection based on VLM

Method	ImageNet-R			ImageNet-A			ImageNet-Sketch			AVG		
	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑
One-shot												
CoOp [62]	80.18	72.44	**91.54**	77.75	77.40	95.65	85.53	67.21	**67.06**	81.15	72.35	**84.75**
+LogitGap*	**79.97**	**71.14**	90.81	**75.72**	**79.77**	**96.25**	**82.81**	**68.37**	67.00	**79.50**	**73.09**	84.69
ID-Like [1]	79.68	75.51	92.61	45.67	88.65	97.77	78.14	75.31	75.32	67.83	79.82	88.57
+LogitGap*	**76.11**	**77.28**	**93.20**	**43.36**	**89.99**	**98.14**	**77.43**	**76.06**	**76.42**	**65.63**	**81.11**	**89.25**
Four-shot												
CoOp [62]	**78.38**	**73.53**	**92.02**	74.87	79.20	96.06	84.68	68.54	**69.06**	79.31	73.76	**85.71**
+LogitGap*	79.59	71.94	91.29	**74.20**	**80.62**	**96.45**	**82.68**	**69.23**	68.91	**78.82**	**73.93**	85.55
ID-Like [1]	85.69	**71.93**	**91.39**	61.27	84.80	97.06	87.07	**67.45**	**67.47**	78.01	74.73	**85.31**
+LogitGap*	**84.45**	71.16	91.03	**59.13**	**86.31**	**97.44**	**86.95**	66.80	**67.47**	**76.84**	**74.76**	**85.31**

Experiments

Results

- **Traditional OOD Detection**

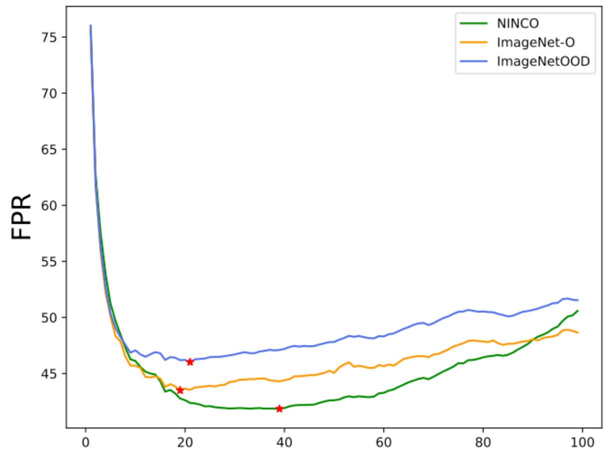
Method	Near OOD				Far OOD								AVG	
	CIFAR100		TIN		MNIST		SVHN		TEXTURE		PLACES365			
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95 ↓	AUROC ↑
MSP [17]	53.10	87.19	43.25	88.87	23.63	92.63	25.80	91.46	34.96	89.89	42.26	88.92	37.17	89.83
MaxLogit [14]	66.60	86.31	56.06	88.72	25.06	94.15	35.09	91.69	51.71	89.41	54.84	89.14	48.23	89.90
Energy [30]	66.59	86.36	56.08	88.8	24.99	94.32	35.15	91.78	51.84	89.47	54.86	89.25	48.25	90.00
ODIN [27]	77.04	82.18	75.32	83.55	23.81	**95.24**	68.61	84.58	67.74	86.94	70.35	85.07	63.81	86.26
GEN [32]	58.76	87.21	48.58	89.20	23.00	93.83	28.14	91.97	40.72	90.14	47.04	89.46	41.04	90.30
Mahalanobis [25]	52.81	83.59	47.01	84.81	27.30	90.10	25.95	91.19	27.92	**92.69**	47.66	84.90	38.11	87.88
ReAct [44]	67.40	85.93	59.71	88.29	33.76	92.81	50.23	89.12	51.40	89.38	44.21	90.35	51.12	89.31
SHE [58]	80.99	80.31	78.27	82.76	42.23	90.43	62.75	86.37	84.59	81.57	76.35	82.89	70.86	84.06
NCI [29]	52.46	87.84	42.92	89.50	28.94	92.08	31.70	90.67	**27.59**	91.97	**35.65**	**90.36**	36.54	90.40
LogitGap	**49.17**	**88.01**	**39.91**	**89.70**	**22.56**	93.44	**25.41**	**92.10**	33.00	90.73	39.38	89.74	**34.91**	**90.62**
OE [18]	35.84	90.81	**2.50**	**99.25**	25.89	90.63	**1.10**	**99.69**	10.61	98.01	13.48	97.14	14.90	95.92
+LogitGap	**35.18**	**91.18**	2.81	99.21	**24.96**	**91.52**	1.18	99.65	**9.76**	**98.10**	**13.20**	**97.18**	**14.52**	**96.14**
MixOE [57]	65.58	87.01	46.79	90.13	36.89	91.79	15.56	**94.48**	33.40	92.00	38.87	91.18	39.52	91.10
+LogitGap	**52.30**	**88.22**	**35.34**	**90.81**	**28.04**	**92.26**	**15.54**	94.46	**25.80**	**92.45**	**30.38**	**91.53**	**31.23**	**91.62**

LogitGap consistently outperforms all post-hoc methods and further improves auxiliary-data methods

Experiments

Ablation Studies

- Influence of the Hyperparameter N
- Different Variants of LogitGap



Method	NINCO			OOD Dataset ImageNet-O			ImageNetOOD			AVG		
	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑
LogitGap	77.42	76.51	96.38	71.95	81.45	99.03	75.40	80.27	86.21	74.92	79.41	93.87
LogitGap_exp	77.15	76.67	96.42	71.40	81.54	99.04	74.33	80.55	86.51	74.48	79.59	93.99
LogitGap_square	76.81	76.80	96.45	71.25	81.53	99.04	74.17	80.97	86.95	74.08	79.77	94.15
LogitGap_sqrt	77.15	76.43	96.36	71.50	81.58	99.04	75.44	80.36	86.33	74.70	79.46	93.91

Thanks!