Test-Time Linear Out-of-Distribution Detection

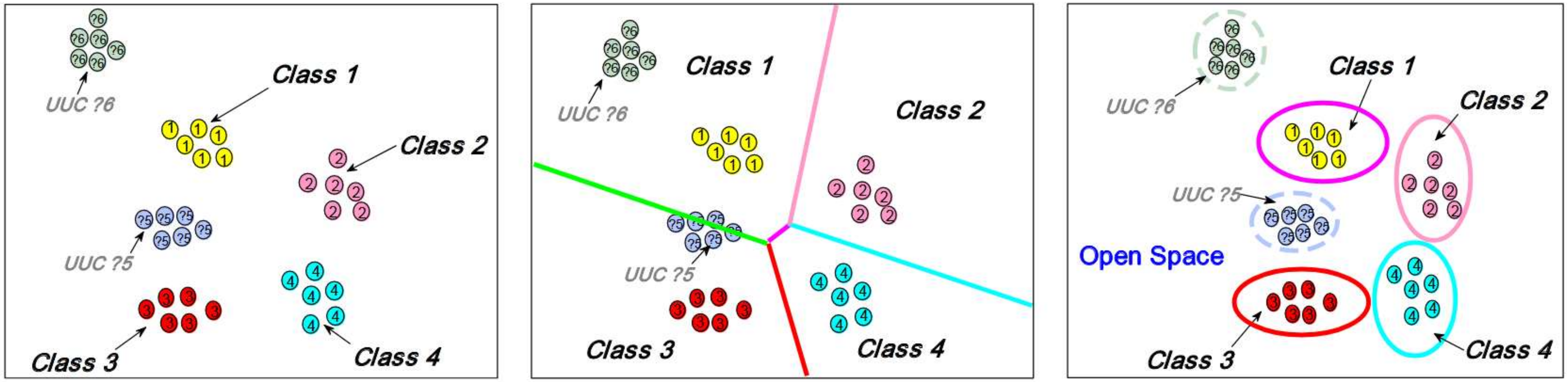
Ke Fan[1*], Tong Liu[2*], Xingyu Qiu[1], Yikai Wang[1], Lian Huai[2†], Zeyu Shangguan[2], Shuang Gou[2], Fengjian Liu[2], Yuqian Fu[1], Yanwei Fu[1], Xingqun Jiang[2]
[1]Fudan University [2]BOE Technology Group

CVPR 2024

Background

▶OOD/OSR:

- Correctly classify ID (in-distribution) samples + reject OOD (out-of-distribution) samples



(a) Distribution of the original data set. (b) Traditional recognition/classification problem. (c) Open set recognition/classification problem.

▶ Test-Time Adaptation (TTA):

- TTA adapts the trained model to the novel data by updating its parameters with a mini-batch or full unlabeled test data during testing.

Recent advances in open set recognition: A survey TPAMI 2021

Motivation

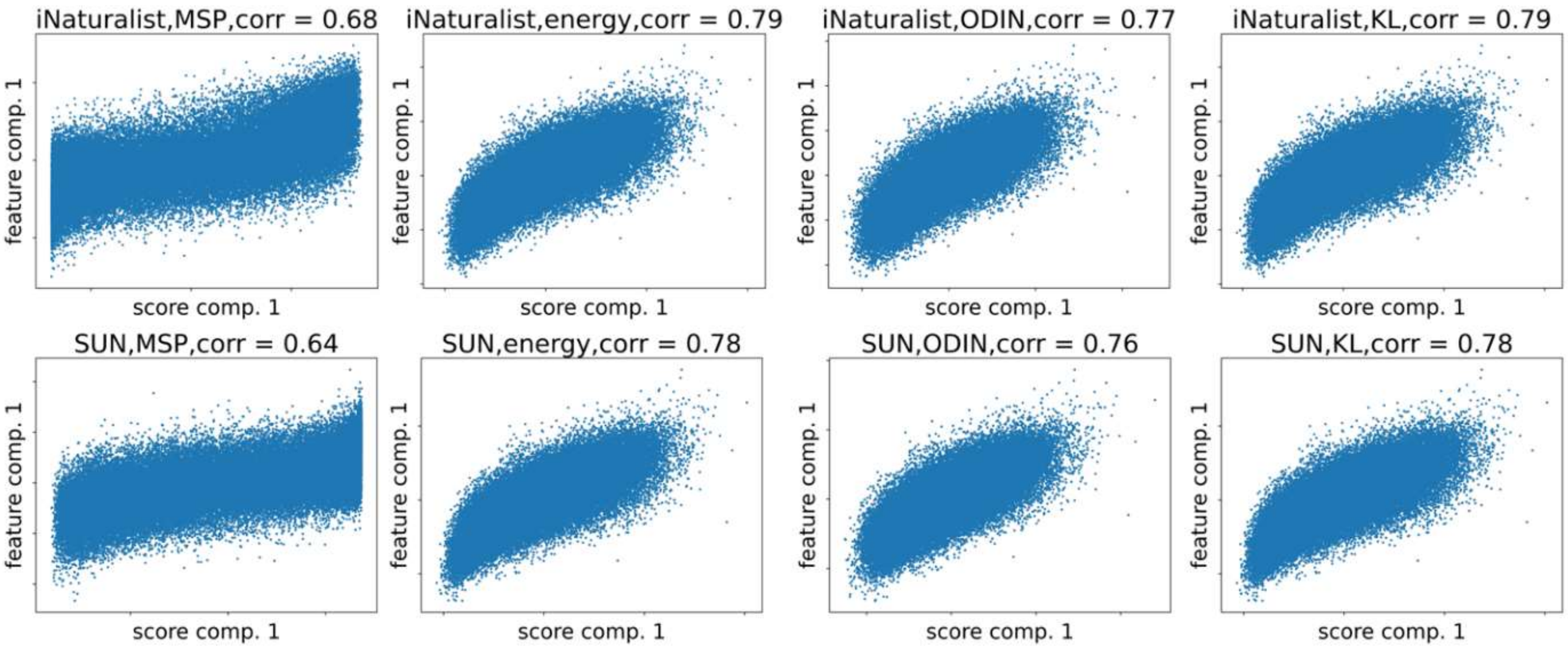


Figure 2. Visualization of Canonical-Correlation Analysis of ImageNet and two OOD datasets' features and OOD scores.

Method

- Model OOD score: $$s = \mathbf{z}^\top \beta + \varepsilon.$$

- Ranking problem:

Suppose there exist a groundtruth OOD score: $s = \mathbf{z}^\top \beta$

It can perfectly rank ID and OOD samples: $\mathbf{z}_{in}^\top \beta > \mathbf{z}_{out}^\top \beta$

If the error distorted the ranking: $\mathbf{z}_{in}^\top \beta + \varepsilon_{in} < \mathbf{z}_{out}^\top \beta + \varepsilon_{out}$

detectors make mistakes

Aim: estimate the β from the feature-score pairs (z_i, s_i)

↓

get a more precise estimation of s

Method

▶RTL

$$\hat{\beta} = \operatorname*{argmin}_{\beta} \sum_{i=1}^{n} (s_i - \mathbf{z}_i^\top \beta)^2,$$

which yields the closed-form solution

$$\hat{\beta} = (\mathbf{Z}^\top \mathbf{Z})^\dagger \mathbf{Z}^\top \mathbf{S},$$

then

$$\hat{s}_{\text{ours}} = \mathbf{z}_i^\top \hat{\beta}.$$

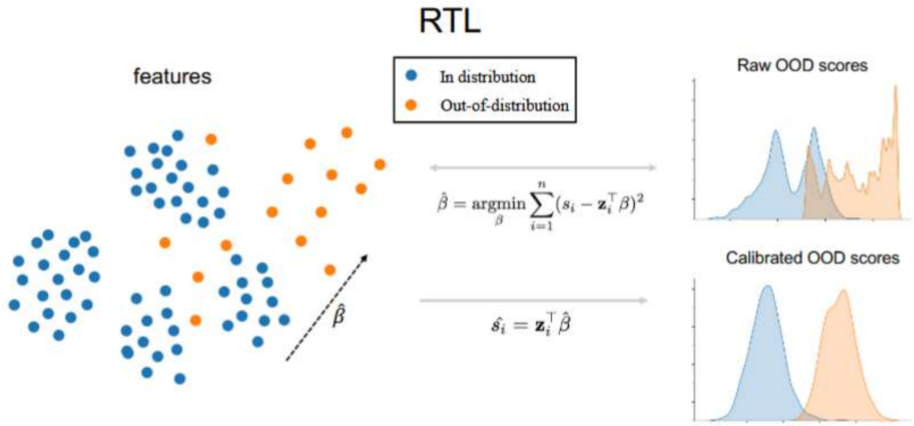


Figure 1. Illustration of our RTL. Blue/orange points denote in/out-of-distribution features. The probability density plot of the OOD score distributions of in and out-of-distribution samples are different. We fit a linear regression between the OOD scores and features as a robust test-time learning method. Then we calibrate the OOD scores to get better OOD predictions.

▶RTL++

When OOD methods do not work good enough,

introducing an explicit data-dependent variable γ_i to model this small amount of large error

$$s_i = \mathbf{z}_i^\top \beta + \boxed{\gamma_i} + \varepsilon_i$$

sparse

Aim: find samples with zero γ_i to fit β

$$\min_{\beta,\gamma} \sum_{i=1}^{n} \left[\frac{1}{2} (s_i - \mathbf{z}_i^\top \beta - \gamma_i)^2 + \lambda |\gamma_i| \right]$$

Method

$$\min_{\beta,\gamma} \sum_{i=1}^{n} \left[\frac{1}{2}(s_i - \mathbf{z}_i^\top \beta - \gamma_i)^2 + \lambda|\gamma_i| \right]$$

Suppose all γ_i are resolved, we can get closed-form estimation of β

$$\hat{\beta} = (\mathbf{Z}^\top \mathbf{Z})^\dagger \mathbf{Z}^\top (\mathbf{S} - \gamma)$$

Define projection $\tilde{\mathbf{Z}} = \mathbf{I} - \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^\dagger \mathbf{Z}^\top$ and $\tilde{\mathbf{S}} = \tilde{\mathbf{Z}}\mathbf{S}$.

Simplify the objective to Lasso $\min_{\gamma} \frac{1}{2}\|\tilde{\mathbf{S}} - \tilde{\mathbf{Z}}\gamma\|_2^2 + \lambda\|\gamma\|_1$.

Choose the most reliable subset to estimate β

$$\hat{\beta} = (\mathbf{Z}_{\text{sub}}^\top \mathbf{Z}_{\text{sub}})^\dagger \mathbf{Z}_{\text{sub}}^\top \mathbf{S}_{\text{sub}}.$$

Algorithm 1: Subset selection of RTL++

Input: features $\mathbf{z}_i$ and OOD score s_i, $1 \le i \le n$,
Normalize $\mathbf{z}_i$ to unit Euclidean norm
Apply dimensionality reduction on $\mathbf{z}_i$ to $d \ll n$
Stack $\mathbf{z}_i$ and s_i by rows to $\mathbf{Z}$ and $\mathbf{S}$
Calculate projection $\tilde{\mathbf{Z}} = \mathbf{I} - \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^\dagger \mathbf{Z}^\top$ and $\tilde{\mathbf{S}} = \tilde{\mathbf{Z}}\mathbf{S}$
Solving Lasso $\hat{\gamma} = \text{argmin}_\gamma \frac{1}{2}\|\tilde{\mathbf{S}} - \tilde{\mathbf{Z}}\gamma\|_2^2 + \lambda\|\gamma\|_1$
Select a subset $\hat{\mathbf{Z}}$ with the lowest $p\%$ of $|\hat{\gamma}_i|$.
return $\hat{\mathbf{Z}}$

Method

▶Online RTL

Two batch data pairs: $\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{bmatrix}, S = \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \end{bmatrix}.$

Recall that $\hat{\beta} = (\mathbf{Z}^\top \mathbf{Z})^\dagger \mathbf{Z}^\top \mathbf{S}$

$$\mathbf{Z}^\top \mathbf{Z} = [\mathbf{Z}_1^\top \quad \mathbf{Z}_2^\top] \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{bmatrix} = \mathbf{Z}_1^\top \mathbf{Z}_1 + \mathbf{Z}_2^\top \mathbf{Z}_2$$

$$\mathbf{Z}^\top \mathbf{S} = [\mathbf{Z}_1^\top \quad \mathbf{Z}_2^\top] \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \end{bmatrix} = \mathbf{Z}_1^\top \mathbf{S}_1 + \mathbf{Z}_2^\top \mathbf{S}_2.$$

Experiments

In Dataset	CIFAR-10		CIFAR-100	
Metric	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Softmax	51.37	90.87	80.21	75.67
+RTL	**12.30**	**97.01**	<u>51.63</u>	<u>84.50</u>
+RTL++	<u>13.50</u>	<u>96.42</u>	**43.73**	**87.06**
Energy	32.98	91.88	73.46	79.67
+RTL	18.01	94.96	60.63	81.15
+RTL++	16.14	95.64	58.06	82.43
ODIN	35.77	90.96	74.55	77.23
+RTL	37.94	86.10	51.87	81.52
+RTL++	35.79	87.24	51.73	82.74
KL	32.98	91.88	73.46	79.67
+RTL	18.01	94.96	60.63	81.15
+RTL++	16.12	95.64	58.06	82.43
IF	79.96	62.47	80.91	66.15
LOF	95.81	56.45	98.23	43.32
GMM	87.70	58.28	94.06	69.96
GradNorm	59.84	71.65	86.55	57.56

Table 2. Results on CIFAR. The best and second best results are highlighted by fonts of text bold and underlined, respectively.

OOD Dataset	iNaturalist		SUN		Places		Textures		Average	
Metric	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95↓	AUROC↑
IF	88.58	61.60	90.12	57.85	93.45	50.24	**54.34**	**87.76**	81.62	64.36
LOF	95.16	51.57	94.89	52.27	93.05	56.37	82.02	65.39	91.28	56.40
GMM	87.90	68.43	89.99	63.29	96.85	52.83	95.37	35.34	92.53	54.97
GradNorm	50.03	90.33	**46.48**	<u>89.03</u>	60.86	<u>84.82</u>	61.42	<u>81.07</u>	54.70	86.31
MSP	63.69	87.59	79.98	78.34	81.44	76.76	82.73	74.45	76.96	79.29
+RTL	<u>21.03</u>	<u>94.98</u>	50.68	87.14	<u>57.22</u>	84.48	<u>58.48</u>	80.24	<u>46.85</u>	<u>86.71</u>
+RTL++	**18.76**	**95.60**	<u>48.40</u>	88.70	**56.72**	**85.32**	59.98	79.91	**45.97**	**87.38**
Energy	64.91	88.48	65.33	85.32	73.02	81.37	80.87	75.79	71.03	82.74
+RTL	45.48	91.04	52.06	88.68	62.68	84.35	69.49	75.39	57.43	84.87
+RTL++	41.57	92.03	49.84	**89.32**	62.37	84.05	70.44	76.52	56.06	85.48
KL	64.91	88.48	65.32	85.31	73.02	81.37	80.87	75.79	71.03	82.74
+RTL	45.48	91.04	52.06	88.68	62.68	84.35	69.49	75.39	57.43	84.87
+RTL++	41.57	92.03	49.84	**89.32**	62.37	84.05	70.44	76.52	56.06	85.48
ODIN	62.69	89.36	71.67	83.92	76.27	80.67	81.31	76.30	72.99	82.56
+RTL	35.27	92.87	51.59	88.40	60.71	84.44	66.72	76.78	53.57	85.62
+RTL++	36.10	92.78	51.87	88.23	61.35	84.28	67.06	76.58	54.09	85.47

Table 3. Results on ImageNet-1k and iNaturalist/SUN/Places/Textures datasets. The best and second-best results are highlighted in bold and underlined font, respectively.

Experiments

In Dataset	CIFAR-10		CIFAR-100	
Metric Size	FPR95↓	AUROC↑	FPR95↓	AUROC↑
MSP	51.37	90.87	80.21	75.67
+Online RTL	15.06	96.05	55.84	83.31
Energy	32.98	91.88	73.46	79.67
+Online RTL	18.57	94.89	62.00	80.91
ODIN	35.77	90.96	74.55	77.23
+Online RTL	40.30	85.15	54.66	80.24
KL	32.98	91.88	73.46	79.67
+Online RTL	18.57	94.89	62.00	80.91

Table 4. Results on CIFAR under different OOD detector.

In Dataset	CIFAR-10		CIFAR-100	
Batch Size	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Raw Softmax	51.37	90.87	80.21	75.67
32	15.06	96.05	55.84	83.31
64	15.04	96.07	55.77	83.34
128	14.99	96.10	55.64	83.38
256	14.79	96.24	55.29	83.53
512	14.53	96.38	54.79	83.69
1024	14.16	96.52	53.98	83.89
All data	**12.30**	**97.01**	**51.63**	**84.50**

Table 5. Online RTL on CIFAR with different batch size.

Experiments

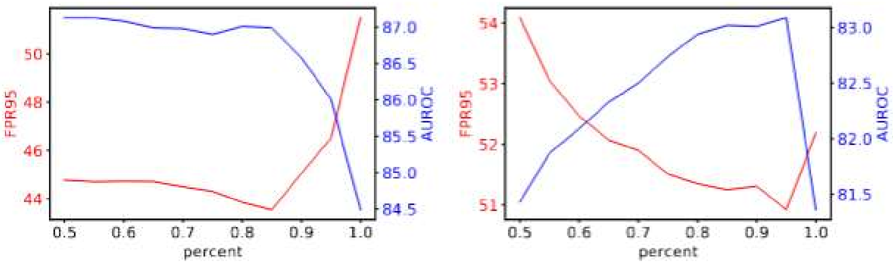


Figure 6. RTL++ results with different percentile of selected subset data on CIFAR-100 with MSP(left) and ODIN(right).

Figure 2. Qualitative examples

Thank you!