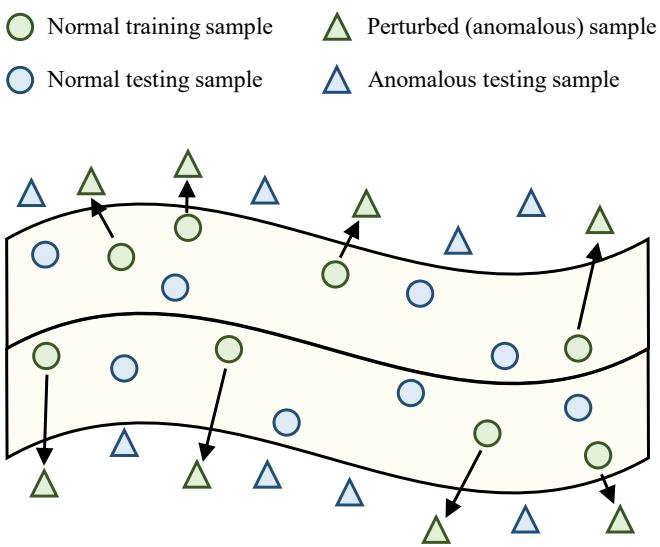
Perturbation Learning Based Anomaly Detection

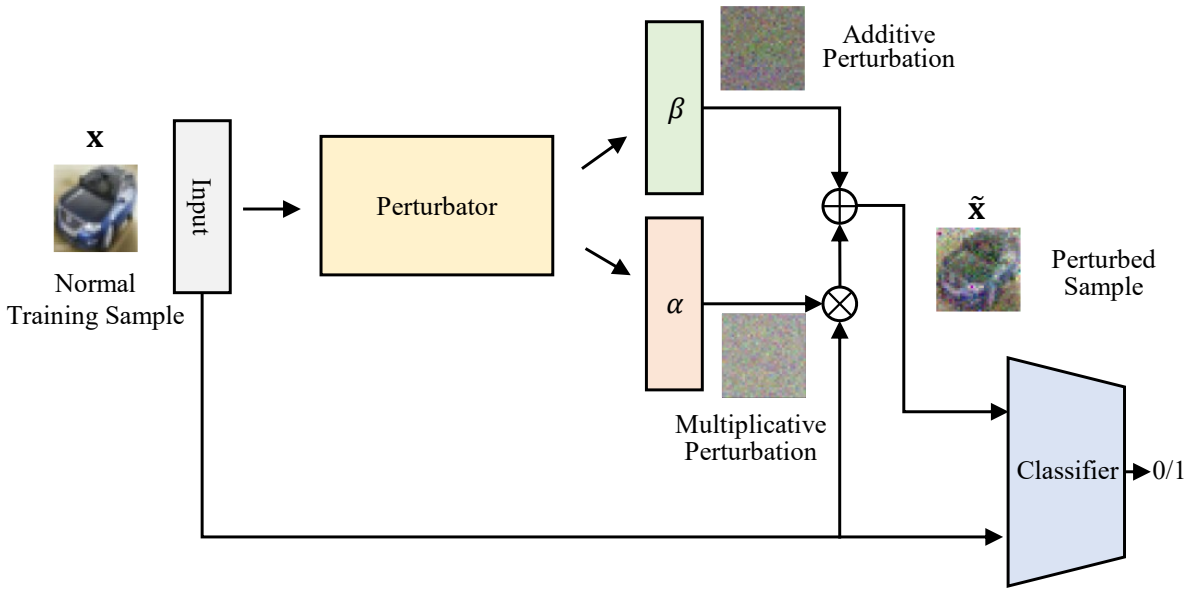
Jinyu Cai[1,2,3], Jicong Fan[2,3]

[1]College of Computer and Data Science, Fuzhou University, China
[2]School of Data Science, The Chinese University of Hong Kong (Shenzhen), China
[3]Shenzhen Research Institute of Big Data, China

Perturbation Learning Based Anomaly Detection

- Illustration of the proposed Perturbation Learning Based Anomaly Detection (PLAD) method.

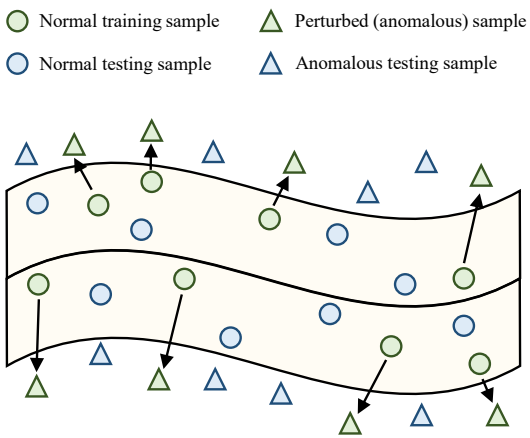


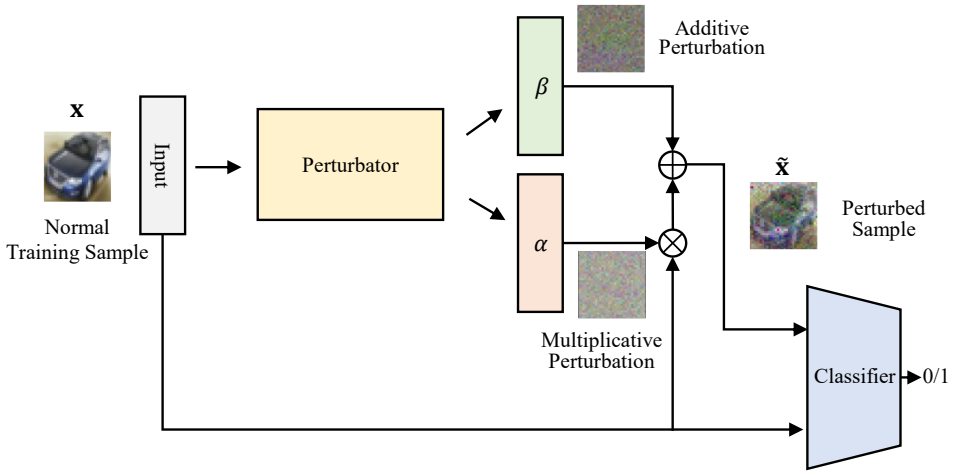


Contributions

Contributions:

- PLAD does not require any assumption about the shape of the decision boundary between the normal data and abnormal data, and has much fewer hyper-parameters than many state-of-the-art AD methods.
- We learn a distribution from which any sample can lead to a perturbation such that the normal data point is flipped to an abnormal data point.
- We investigate the performance of our PLAD and its competitors in recognizing abnormal data from multi-class normal data.
- PLAD is a general AD framework and can also be applied to time series, text, and graph data via changing the network components.





Proposed Model

Specifically, we solve the following problem:

$$\begin{aligned} \underset{\theta,\ \tilde{\theta}}{\text{minimize}}\ & \frac{1}{n}\sum_{i=1}^{n}\ell(y_i, f_\theta(\mathbf{x}_i)) + \frac{1}{n}\sum_{i=1}^{n}\ell(\tilde{y}_i, f_\theta(\tilde{\mathbf{x}}_i)) + \frac{\lambda}{n}\sum_{i=1}^{n}\left(\|\boldsymbol{\alpha}_i - \mathbf{1}\|^2 + \|\boldsymbol{\beta}_i - \mathbf{0}\|^2\right) \\ \text{subject to}\ & \tilde{\mathbf{x}}_i = \mathbf{x}_i \odot \boldsymbol{\alpha}_i + \boldsymbol{\beta}_i,\ (\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i) = g_{\tilde{\theta}}(\mathbf{x}_i),\ i = 1, 2, \ldots, n, \end{aligned} \tag{1}$$

- $\ell(\cdot,\cdot)$ denotes some loss function such as cross-entropy and $y_1 = \cdots = y_n = 0$ and $\tilde{y}_1 = \cdots = \tilde{y}_n = 1$ are the labels for the normal data and perturbed data respectively.
- $\mathbf{1} = [1, 1, \ldots, 1]^\top$ and $\mathbf{0} = [0, 0, \ldots, 0]^\top$ are d-dimensional constant vectors and $\odot$ denotes the Hadamard product.
- $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ are multiplicative and additive perturbations for $\mathbf{x}_i$.
- $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ are generated from a perturbator $g_{\tilde{\theta}}$, where $\tilde{\theta}$ denotes the set of parameters to learn.

In (1), we hope that the multiplicative perturbation is close to 1 and the additive perturbation is close to 0 but they rely on the data point $\mathbf{x}$.

Proposed Model

In fact, problem (1) can be reformulated as

$$\begin{aligned}\underset{\theta,\,\tilde{\theta}}{\text{minimize}}\ & \frac{1}{n}\sum_{i=1}^{n}\Big(\ell\big(y_i, f_\theta(\mathbf{x}_i)\big) + \ell\big(\tilde{y}_i, f_\theta(\mathbf{x}_i \odot g_{\tilde{\theta}}^{\alpha}(\mathbf{x}_i) + g_{\tilde{\theta}}^{\beta}(\mathbf{x}_i))\big)\Big) \\ & + \frac{\lambda}{n}\sum_{i=1}^{n}\Big(\|g_{\tilde{\theta}}^{\alpha}(\mathbf{x}_i) - \mathbf{1}\|^2 + \|g_{\tilde{\theta}}^{\beta}(\mathbf{x}_i) - \mathbf{0}\|^2\Big),\end{aligned} \tag{2}$$

where $\begin{bmatrix} g_{\tilde{\theta}}^{\alpha}(\mathbf{x}_i) \\ g_{\tilde{\theta}}^{\beta}(\mathbf{x}_i) \end{bmatrix} = g_{\tilde{\theta}}(\mathbf{x}_i)$, $i = 1, 2, \ldots, n$.

- The optimized-needed parameters are only θ and $\tilde{\theta}$ and the total number of decision variables is $|\theta| + |\tilde{\theta}|$.
- In PLAD, besides the network structures, we only need to determine one hyperparameter λ, which provides huge convenience in real applications.
- In PLAD, we can use gradient-based optimizer such as Adam to solve the optimization.
- Once f_θ and $g_{\tilde{\theta}}$ are learned, we can then use f_θ to detect whether a new data point $\mathbf{x}_{\text{new}}$ is normal (e.g. $f_\theta(\mathbf{x}_{\text{new}}) < 0.5$) or abnormal (e.g. $f_\theta(\mathbf{x}_{\text{new}}) > 0.5$).

Experimental Results on Image Datasets

- Average AUCs (%) of the one-class anomaly detection task on Fashion-MNIST. Note the we further report the standard deviation for the proposed method, and the best two results are marked in **bold**.

Method	T-shirt	Trouser	Pullover	Dress	Coat	Sandal	Shirt	Sneaker	Bag	Ankle boot
OCSVM	86.1	93.9	85.6	85.9	84.6	81.3	78.6	97.6	79.5	97.8
IF	91.0	97.8	87.2	93.2	90.5	93.0	80.2	98.2	88.7	95.4
DAE	86.7	97.8	80.8	91.4	86.5	92.1	73.8	97.7	78.2	96.3
DAGMM	42.1	55.1	50.4	57.0	26.9	70.5	48.3	83.5	49.9	34.0
ADGAN	89.9	81.9	87.6	91.2	86.5	89.6	74.3	97.2	89.0	97.1
DSVDD	79.1	94.0	83.0	82.9	87.0	80.3	74.9	94.2	79.1	93.2
OCGAN	85.5	93.4	85.0	88.1	85.8	88.5	77.5	93.9	82.7	97.8
TQM	92.2	95.8	**89.9**	93.0	**92.2**	89.4	**84.4**	98.0	**94.5**	98.3
DROCC	88.1	97.7	87.6	87.7	87.2	91.0	77.1	95.3	82.7	95.9
HRN-L2	91.5	97.6	88.2	92.7	91.0	71.9	79.4	**98.9**	90.8	**98.9**
HRN	**92.7**	98.5	88.5	93.1	92.1	**91.3**	79.8	**99.0**	**94.6**	98.8
PLAD	**93.1** (0.5)	**98.6** (0.2)	**90.2** (0.7)	**93.7** (0.6)	**92.8** (0.8)	**96.0** (0.4)	**82.0** (0.6)	98.6 (0.3)	90.9 (1.0)	**99.1** (0.1)

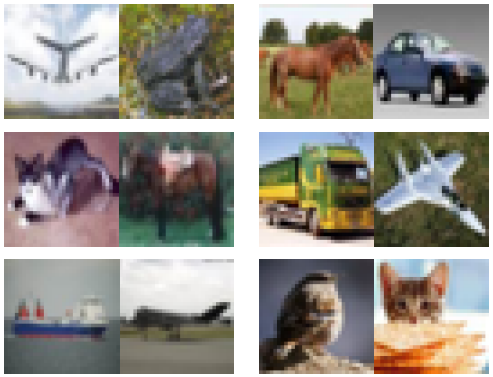
Experimental Results on Non-Image Datasets

- Average F1-scores (%) with the standard deviation of each method on the two tabular datasets (Thyroid and Arrhythmia). Note that the best two results are marked in **bold**.

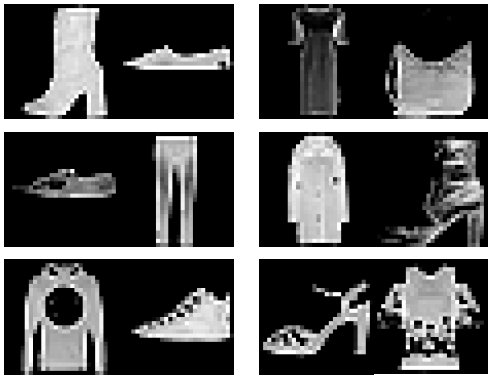
Data set	Thyroid	Arrhythmia
OCSVM	39.0 ± 1.0	46.0 ± 0.0
LOF	54.0 ± 1.0	51.0 ± 1.0
E2E-AE	13.0 ± 4.0	45.0 ± 3.0
DCN	33.0 ± 3.0	38.0 ± 3.0
DAGMM	49.0 ± 4.0	49.0 ± 3.0
DSVDD	73.0 ± 0.0	54.0 ± 1.0
DROCC	68.7 ± 2.3	32.3 ± 1.8
GOAD	74.5 ± 1.1	52.0 ± 2.3
NeuTraL AD	$\mathbf{76.8 \pm 1.9}$	60.3 ± 1.1
GOCC	$\mathbf{76.8 \pm 1.2}$	$\mathbf{61.8 \pm 1.8}$
PLAD	76.6 ± 0.6	$\mathbf{71.0 \pm 1.7}$

Experiment of Separating Anomaly from Multi-Class Normal Data

- (a) and (b) denotes the randomly selected pair-wise normal samples from CIFAR-10 and Fashion-MNIST, respectively.
- (c) denotes the anomalies produced by using pixel-level means of pair-wise normal samples from (a) and (b).
- The table shows AUCs (%) of each compared method in the experiment.



(a) Multi-class normal samples from CIFAR-10



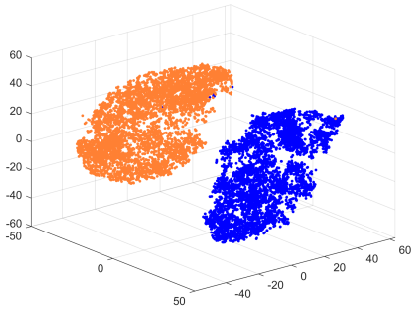
(b) Multi-class normal samples from Fashion-MNIST

(c) Anomalies produced from pair-wise samples

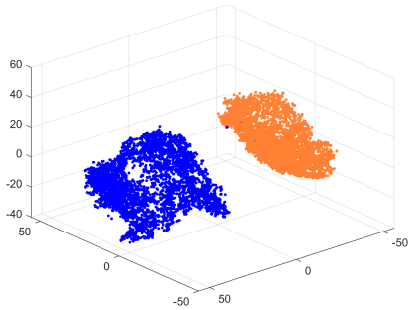
Data set	CIFAR-10	Fashion-MNIST
OCSVM	54.9 ± 0.0	64.8 ± 0.0
DAGMM	44.3 ± 0.6	49.2 ± 2.6
DSVDD	63.6 ± 1.1	70.9 ± 2.0
DROCC	60.9 ± 5.8	68.1 ± 3.1
PLAD	$\mathbf{72.7 \pm 1.9}$	$\mathbf{75.3 \pm 2.8}$

Visualization of the Embedding Space

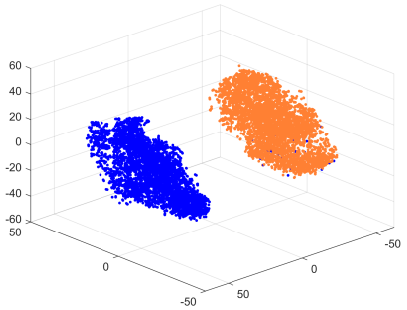
- Visualization of the learned embedding space in two cases (with training and perturbed samples, and with training, perturbed and test samples, respectively) on Fashion-MNIST.
- The points marked in blue, orange, green, and red are training samples, perturbed samples, normal test samples, and anomalous test samples, respectively.



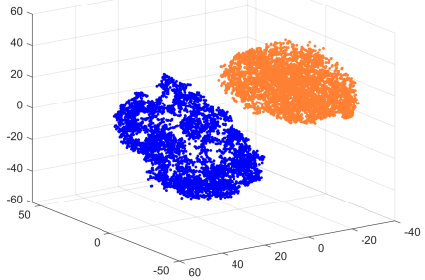
(a) T-shirt



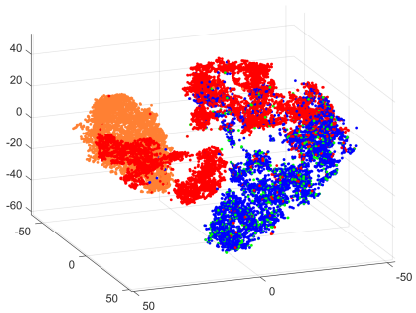
(b) Sandal



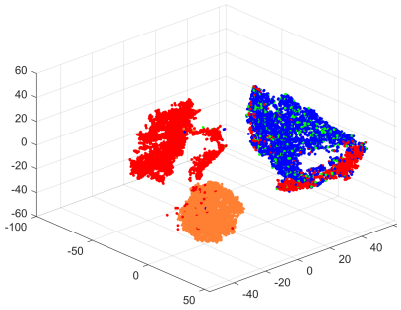
(c) Sneaker



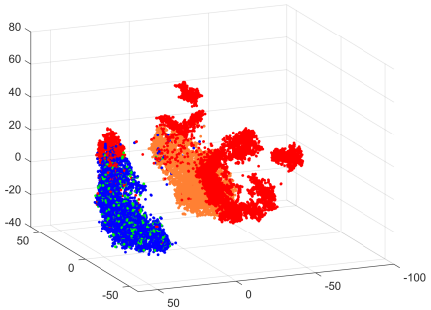
(d) Ankle boot



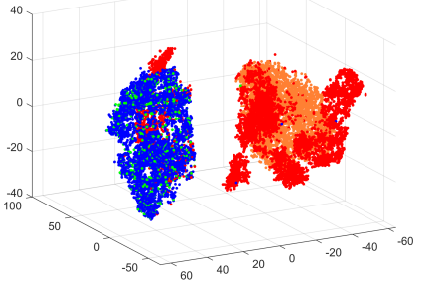
(a) T-shirt



(b) Sandal



(c) Sneaker



(d) Ankle boot

Statistical Analysis of PLAD

- Student's t-test results of the one-class anomaly detection task on Fashion-MNIST.

Reproduce	T-shirt	Trouser	Pullover	Dress	Coat	Sandal	Shirt	Sneaker	Bag	Ankle boot
DSVDD	79.1	94.0	83.0	82.9	87.0	80.3	74.9	94.2	79.1	93.2
DSVDD	78.4 ± 3.3	93.6 ± 1.3	80.8 ± 3.4	84.1 ± 2.0	85.9 ± 2.4	82.0 ± 3.0	75.0 ± 3.8	94.5 ± 1.8	80.6 ± 5.9	94.1 ± 1.5
DROCC	88.1 ± 3.3	97.7 ± 0.7	87.6 ± 1.4	87.7 ± 1.6	87.2 ± 2.2	91.0 ± 1.6	77.1 ± 2.0	95.3 ± 0.7	82.7 ± 2.9	95.9 ± 2.1
HRN	92.7 ± 0.0	98.5 ± 0.1	88.5 ± 0.1	93.1 ± 0.1	92.1 ± 0.1	91.3 ± 0.4	79.8 ± 0.1	99.0 ± 0.0	94.6 ± 0.1	98.8 ± 0.0
HRN	88.8 ± 0.1	98.6 ± 0.1	84.8 ± 0.1	93.2 ± 0.1	89.5 ± 0.2	89.6 ± 0.1	74.4 ± 0.1	98.9 ± 0.0	87.2 ± 0.3	97.7 ± 0.1
PLAD (Ours)	93.1 ± 0.5	98.6 ± 0.2	90.2 ± 0.7	93.7 ± 0.6	92.8 ± 0.8	96.0 ± 0.4	82.0 ± 0.6	98.6 ± 0.3	90.9 ± 1.0	99.1 ± 0.1
p-value (t-test)	T-shirt	Trouser	Pullover	Dress	Coat	Sandal	Shirt	Sneaker	Bag	Ankle boot
v.s. DSVDD	1.5×10^{-7}	1.2×10^{-6}	2.2×10^{-5}	1.0×10^{-7}	2.1×10^{-5}	8.2×10^{-8}	8.7×10^{-4}	4.7×10^{-5}	0.003	2.9×10^{-6}
v.s. DROCC	9.4×10^{-4}	0.004	3.3×10^{-4}	3.4×10^{-6}	3.0×10^{-5}	3.1×10^{-6}	4.4×10^{-4}	3.4×10^{-7}	9.9×10^{-6}	5.1×10^{-4}
v.s. HRN	4.0×10^{-9}	0.614	1.2×10^{-7}	0.013	5.9×10^{-6}	3.0×10^{-10}	1.5×10^{-9}	5.4×10^{-4}	2.2×10^{-5}	1.7×10^{-10}

- Usually, the difference is said to be significant if the *p*-value obtained in the t-test is less than 0.05.