
David J. Miller
Zhen Xiang
George Kesidis


# Adversarial Learning and Secure AI

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

# Chapter 11

Backdoors for 3D Point Cloud (PC) Classifiers

# Outline

- Backdoor attacks in 3D Point Cloud (PC) datasets
- PC-PT-RED
- A single-point "intrinsic" (or "natural") backdoor phenomenon
- A combined statistic to address the intrinsic backdoor phenomenon

# Backdoor Attack against Point Cloud Classifiers

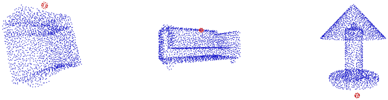
- Point cloud (PC) data: A set of permutation invariant points:

$$\mathbf{X} = \{\underline{x}_i \in \mathbb{R}^3 | i = 1, \cdots, n\} \in \mathcal{X}$$

- Backdoor pattern for PCs
  - A set of inserted points

$$\mathbf{V}^* = \{\underline{u}_j^* + \underline{C}^* | \underline{u}_j^* \in \mathbb{R}^3, \underline{C}^* \in \mathbb{R}^3, j = 1, \cdots, n'\}.$$

  - $\underline{C}^*$: an optimized, common spatial location close to points in all source class PCs.
  - $\mathbf{U}^* = \{\underline{u}_j^* \in \mathbb{R}^3 | j = 1, \cdots, n'\}$: an optimized local geometry to bypass point sampling and possible anomaly detection.
  - Examples (backdoor points are in red)

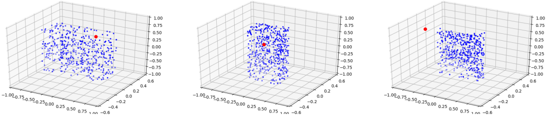


# PC-PT-RED: Key Ideas

- Intuition 1: closeness to source class for backdoor attack
  - For most non-backdoor class pairs, a common set of inserted points that induces high group misclassification from source class to target class will be spatially far from the points of source class PCs.
  - But for a backdoor class pair, there exists a common spatial location close to the source class PCs (likely near $\underline{C}^*$), where a set of inserted points can induce most source class PCs to be misclassified to the target class.
- Intuition 2: closeness to target class for intrinsic backdoor
  - A few non-backdoor class pairs may be associated with an intrinsic backdoor.
  - For these class pairs, the common spatial location close to the source class PCs will also be close to the points of most target class PCs.

# PC-PT-RED: Key Ideas (cont)

- Intuition 3: dissimilarity of spatial locations for intrinsic backdoor
  - Intrinsic backdoor is likely due to the source and target classes being "semantically" similar.
  - There may exist several intrinsic backdoor points for a given non-backdoor class pair.
  - The closest sample-wise spatial location for a set of inserted points to induce a sample-wise misclassification to the target class can be different for different PCs from the same source class.
  - Illustration



# PC-PT-RED: Step 1: Backdoor Pattern Estimation

- Based on Intuition 1, for each class pair $(s, t)$, solve:

$$\min_{\underline{C} \in \mathbb{R}^3} \sum_{\mathbf{X} \in \mathcal{D}_s} d(\underline{C}, \mathbf{X})$$

$$\text{s.t.} \quad \frac{1}{|\mathcal{D}_s|} \sum_{\mathbf{X} \in \mathcal{D}_s} \mathbb{1}\left\{\hat{c}(\mathbf{X} \cup \{\underline{C}\}) = t\right\} \geq \pi,$$

  - $\mathcal{D}_s$: subset of clean samples from class $s \in \mathcal{Y}$.
  - $\mathbb{1}\{\cdot\}$: indicator function.
  - $\pi$: target group misclassification fraction (set large, e.g. $\pi = 0.9$).
  - $d(\underline{C}, \mathbf{X}) = \min_{\underline{x} \in \mathbf{X}} \|\underline{C} - \underline{x}\|_2$.
- The above problem is difficult to solve
  - The indicator function is not differentiable.
  - Solution may yield an overly large objective distance for some class pairs due to the strong robustness of PC classifiers.

# PC-PT-RED: Step 1: Backdoor Pattern Estimation (cont)

- Perform backdoor pattern estimation for each source class by minimizing the following differentiable surrogate objective:

$$L(\underline{C}; \mathcal{D}_s, \lambda) = \sum_{\mathbf{X} \in \mathcal{D}_s} \left[ h(s | \mathbf{X} \cup \{\underline{C}\}) - \max_{k \neq s} h(k | \mathbf{X} \cup \{\underline{C}\}) \right] + \lambda \sum_{\mathbf{X} \in \mathcal{D}_s} d(\underline{C}, \mathbf{X})$$

  - $h(k|\mathbf{X})$: output logit for class $k$ and sample $\mathbf{X}$.
  - $\lambda$ : Lagrange multiplier (adjusted automatically).
- Let $\underline{\hat{C}}(s)$ be spatial location estimated for class $s$.
- The source class PCs "vote" for a target class:

$$\hat{t}(s) = \operatorname*{argmax}_{k \neq s} \sum_{\mathbf{X} \in \mathcal{D}_s} \mathbb{1}\{\hat{c}(\mathbf{X} \cup \{\underline{\hat{C}}(s)\}) = k\}$$

# PC-PT-RED: Step 1: Backdoor Pattern Estimation (cont)

- Based on Intuition 3, we estimate a sample-wise spatial location for each $\mathbf{X} \in \mathcal{D}_s$ by minimizing:

$$\tilde{L}(\underline{C}; \mathbf{X}, \lambda) = h(s|\mathbf{X} \cup \{\underline{C}\}) - h(\hat{t}(s)|\mathbf{X} \cup \{\underline{C}\}) + \lambda d(\underline{C}, \mathbf{X})$$

  - $\hat{t}(s)$: estimated target class.
  - Denote the estimated sample-wise (SW) spatial location for $\mathbf{X} \in \mathcal{D}_s$ as $\hat{\underline{C}}_{\mathrm{sw}}(s, \mathbf{X})$

# PC-PT-RED: Step 2: Detection Inference

A detection statistic with three component statistics:

- Statistic 1: average distance to source class

$$r_{\mathrm{s}}(s) = \frac{1}{|\mathcal{D}_s|} \sum_{\mathbf{X} \in \mathcal{D}_s} d(\underline{\hat{C}}(s), \mathbf{X})$$

- Statistic 2: average distance to estimated target class

$$r_{\mathrm{t}}(s) = \frac{1}{|\mathcal{D}_{\hat{t}(s)}|} \sum_{\mathbf{X} \in \mathcal{D}_{\hat{t}(s)}} d(\underline{\hat{C}}(s), \mathbf{X})$$

# PC-PT-RED: Step 2: Detection Inference (cont)

- Statistic 3: a normalized similarity score

$$w(s) = \frac{z(s) - \min_{k \in \mathcal{Y}} z(k)}{\max_{k \in \mathcal{Y}} z(k) - \min_{k \in \mathcal{Y}} z(k)}, \quad \text{where}$$

$z(k) = \frac{1}{|\mathcal{D}_k|} \sum_{\mathbf{X} \in \mathcal{D}_k} \frac{\underline{\hat{C}}(k) \cdot \underline{\hat{C}}_{\mathrm{sw}}(k, \mathbf{X})}{|\underline{\hat{C}}(k)|\,|\underline{\hat{C}}_{\mathrm{sw}}(k, \mathbf{X})|}$ is average cosine similarity for $\mathcal{D}_k$.

# PC-PT-RED: Step 2: Detection Inference (cont)

- Combination of statistics 1,2,3:

$$r(s) = w(s)\frac{r_{\mathrm{t}}(s)}{r_{\mathrm{s}}(s)}$$

  - Based on Intuition 1, $r_{\mathrm{s}}(s)$ will likely be large if $(s, \hat{t}(s))$ is a non-backdoor class pair; otherwise, $r_{\mathrm{s}}(s)$ will likely be small.
  - Based on Intuition 2 and 3, if $(s, \hat{t}(s))$ is associated with an intrinsic backdoor mapping, $r_{\mathrm{t}}(s)$ or $w(s)$ (or both) will likely be small.
  - Combining the above, $r(s)$ will be abnormally large only if $(s, \hat{t}(s))$ is a backdoor class pair.

# PC-PT-RED: Step 2: Detection Inference (cont)

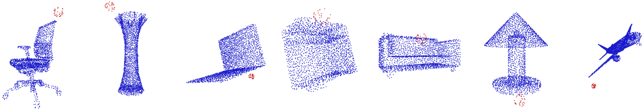
- How do we assess atypicality? We implement an unsupervised anomaly detector.
  - Denote $s_{\max} = \arg\max_{k \in \mathcal{Y}} r(k)$.
  - to estimate a null distribution $G(\cdot)$, exclude statistics for all $s$ such that $\hat{t}(s) = \hat{t}(s_{\max})$.
  - Given all positive statistics, choose a single-tailed null density form, e.g., a Gamma distribution, so that outliers will appear at the tail.
  - Estimate the maximum order statistic p-value:

$$\mathrm{pv} = 1 - G(r(s_{\max}))^{K-J}$$

    where $K$ is number of classes, $J$ is number of statistics being excluded.
  - Set a confidence threshold, e.g., $\phi = 0.05$, and claim a detection (with confidence $1 - \phi$) if $\mathrm{pv} < \phi$.

# Experiments

- Settings
  - Dataset: Modelnet40
  - PC classifier architecture: PointNet, PointNet++, DGCNN
  - Attacks P1–P7
  - Example backdoored training samples with these 7 different backdoor patterns:



# Experiments (cont)

- Results – detection effectiveness

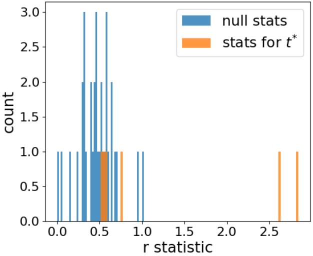
| | $P_1$-PN | $P_2$-PN | $P_3$-PN | $P_4$-PN | $P_5$-PN |
|---|---|---|---|---|---|
| $1/r_s$ | (**6.2e$^{-3}$**, **0.36**) | (**3.8$^{-3}$**, **0.16**) | (**4.3e$^{-15}$**, **0.33**) | (**2.2e$^{-7}$**, 2.6e$^{-2}$) | (0.24, **0.11**) |
| $r_t/r_s$ | (**4.5e$^{-2}$**, 9.2e$^{-6}$) | (**u.f.**, **0.32**) | (**6.1e$^{-6}$**, 9.8e$^{-2}$) | (**2.8e$^{-3}$**, **0.58**) | (0.12, **0.19**) |
| $w/r_s$ | (**1,7e$^{-7}$**, **0.19**) | (**3.5e$^{-3}$**, **0.26**) | (**u.f.**, **0.27**) | (**5.6e$^{-9}$**, 9.2e$^{-3}$) | (**1.4e$^{-2}$**, 6.1e$^{-2}$) |
| $r = w \cdot r_t/r_s$ | (**3.3e$^{-3}$**, **0.38**) | (**u.f.**, **0.19**) | (**u.f.**, **0.20**) | (**u.f.**, **0.22**) | (**5.4e$^{-2}$**, **0.27**) |

| | $P_6$-PN | $P_7$-PN | $P_1$-PN++ | $P_1$-DGCNN |
|---|---|---|---|---|
| $1/r_s$ | (0.24, 1.6e$^{-2}$) | (**4.3e$^{-3}$**, 9.7e$^{-2}$) | (**u.f.**, 8.2e$^{-6}$) | (**4.4e$^{-5}$**, 4.3e$^{-2}$) |
| $r_t/r_s$ | (0.21, **0.60**) | (**6.7e$^{-5}$**, 9.0e$^{-3}$) | (**u.f.**, **0.99**) | (0.10, **0.59**) |
| $w/r_s$ | (**1.4e$^{-2}$**, 2.6e$^{-2}$) | (**5.5e$^{-9}$**, 7.0e$^{-3}$) | (**u.f.**, **0.94**) | (0.22, 2.9e$^{-2}$) |
| $r = w \cdot r_t/r_s$ | (**7.6e$^{-4}$**, **0.33**) | (**u.f.**, 9.3e$^{-2}$) | (**5.5e$^{-13}$**, **0.99**) | (**1.9e$^{-3}$**, **0.18**) |

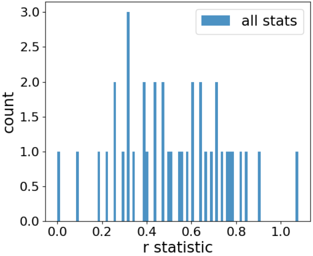
  - Ablation study: compare PC-PT-RED's combined statistic with other combinations of statistics.
  - Demonstrated using order statistic p-values: (pv attack, pv clean).
  - Bold for successful detection (with $\phi = 0.05$): for attack, pv less than $\phi$; for clean, pv greater than $\phi$.

# Experiments (cont)

- Illustration of the histogram of combined $r$ statistics




(a) attack




(b) clean

# Discussion: Intrinsic (Natural) Backdoors

- Natural backdoors may exist in other domains.
- For example, for classification of animal images, the training dataset may involve cow images mainly taken in pastures, with grass much less common in images of other animals.
- Thus, the DNN may learn to classify to the cow class whenever grass is present in the image [Hendrycks et al. ICCV'21].

# With Permission, Figures Reproduced From

- Z. Xiang, D.J. Miller, S. Chen, X. Li, and G. Kesidis. A Backdoor Attack against 3D Point Cloud Classifiers. In *Proc. International Conference on Computer Vision (ICCV)*, Oct. 2021.
- Z. Xiang, D.J. Miller, S. Chen, X. Li, and G. Kesidis, Detecting Backdoor Attacks Against Point Cloud Classifiers. In *Proc. IEEE ICASSP*, Mar. 2022.