Collider: A Robust Training Framework for Backdoor Data

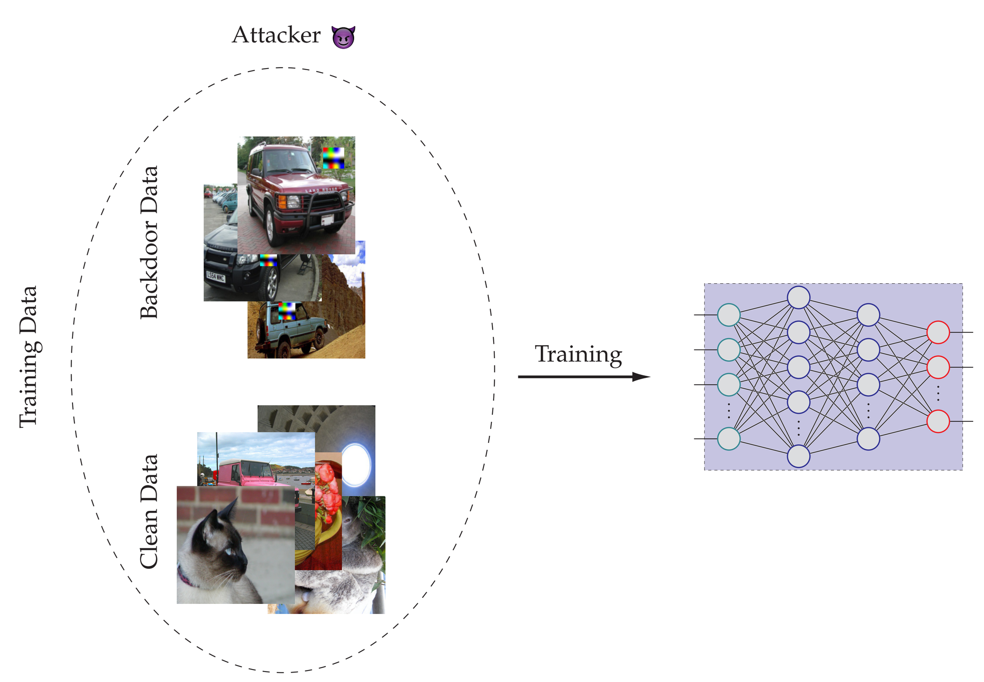
Hadi M. Dolatabadi, Sarah Erfani, and Christopher Leckie
School of Computing and Information Systems, The University of Melbourne

Abstract

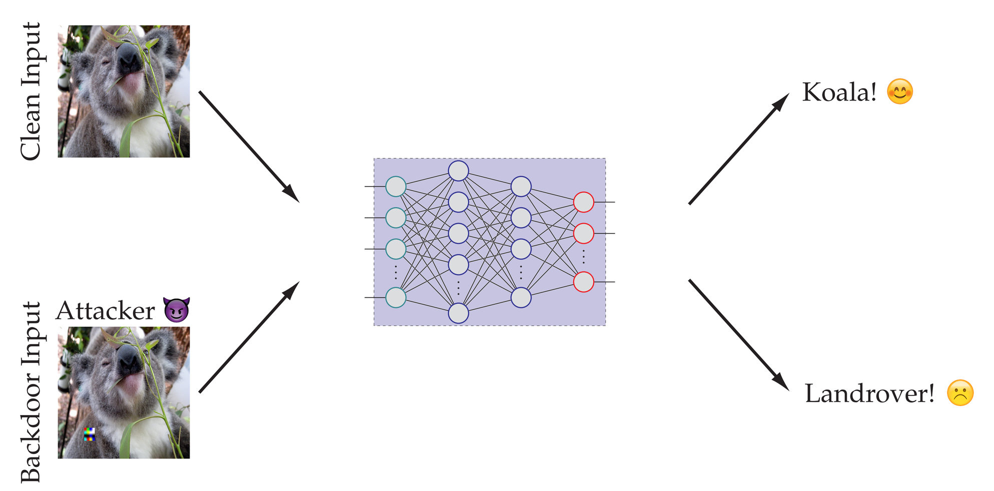
- **Motivation**: poisoned training data can create backdoors in deep neural networks (DNN) so the model misclassifies samples with a pre-designed trigger. Existing robust methods need to train the DNN twice so they can filter out the poisoned data, but this is time-consuming.
- **Proposal**: we propose Collider, a COreset selection algorithm with LocaL Intrinsic DimEnisonality Regularization, to filter out suspicious samples in an online manner and train the DNN over the clean data.
- **Key Features of Collider**:
  1. Efficient, single-run training of DNNs against backdoor data.
  2. Compatible against various backdoor attacks.
  3. Eliminating the effects of backdoor attacks almost entirely without requiring a clean validation set.

Background: Backdoor Attacks

- By attaching a trigger to training images, attackers can create backdoors in DNNs and exploit them during inference.



(a) Training the DNN over poisoned data.

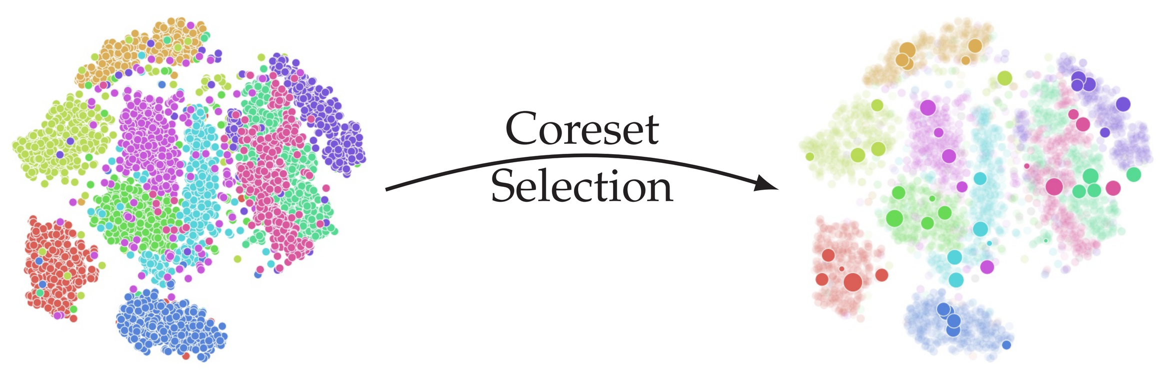


(b) Expected behavior at test-time in the absence and presence of the trigger.

Background: Coreset Selection

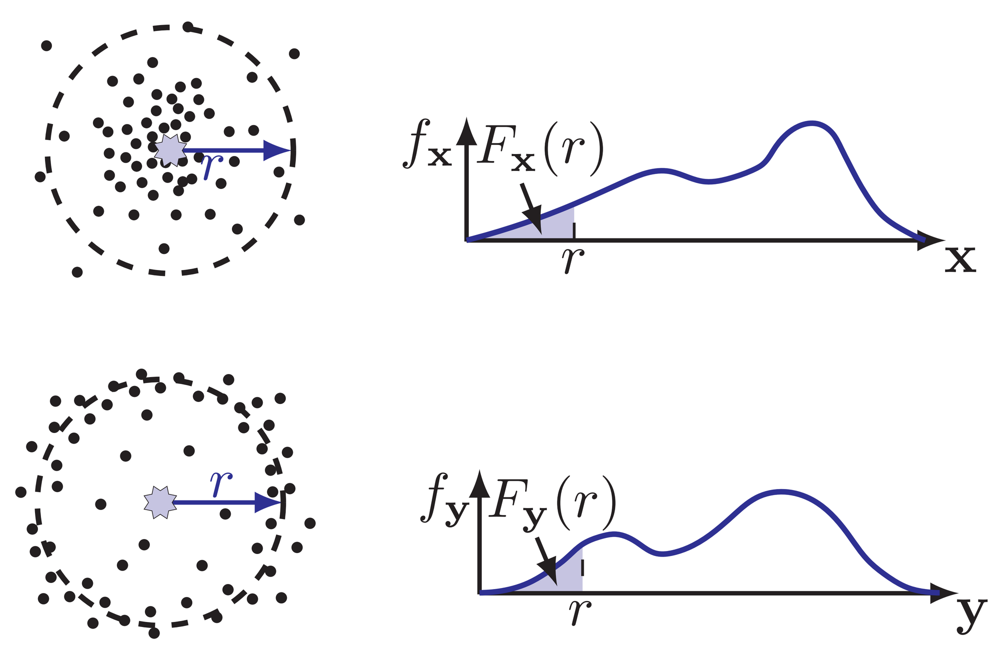
- Coreset selection aims at finding a *weighted subset* of the data that can approximate certain behaviors of the entire data samples.
- In particular, let us denote the behavior of interest as a function $\mathcal{B}(\cdot, \cdot)$ that receives a set and its associated weights.
- The goal of coreset selection is to move from the original data $\mathcal{V}$ with uniform wights $\mathbf{1}$ to a weighted subset $\mathcal{S}^* \subseteq \mathcal{V}$ with weights γ^* such that:

$$\mathcal{B}(\mathcal{V}, \mathbf{1}) \qquad \approx \qquad \mathcal{B}(\mathcal{S}^*, \gamma^*).$$



Background: LID

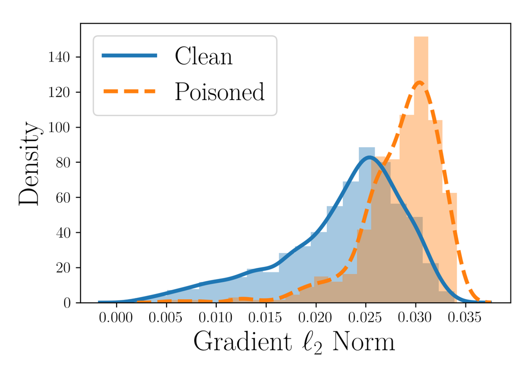
- Traditionally, classical expansion models such as generalized expansion dimension (GED) were used to measure the intrinsic dimensionality of the data.
- By extending the aforementioned setting into a statistical one, classical expansion models can provide a local view of intrinsic dimensionality (LID).



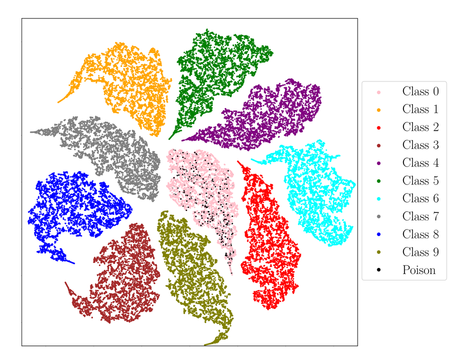
Overview of LID (based on Figure 1 in [1]). As shown, the random distance variables $\mathbf{x}$ and $\mathbf{y}$ have an approximately equal cumulative distribution at distance r. However, since the concentration of points for $\mathbf{y}$ at distance r is higher than $\mathbf{x}$, then $\mathrm{LID}_{F_{\mathbf{y}}}(r)$ is greater than $\mathrm{LID}_{F_{\mathbf{x}}}(r)$.

Our Method: Collider

- **Motivation**: using coreset selection to filter out the poisonous samples.
- To this end, we need to define an appropriate coreset selection objective.
- We perform this noticing two properties of the poisoned data:
  1. **Gradient Space Properties**: the gradient updates computed on poisoned data (a) have comparably larger ℓ_2 norm [2], and (b) are scattered in the gradient space [3].

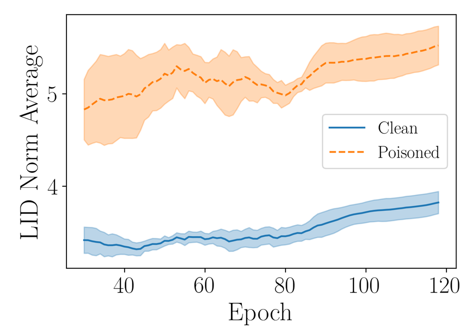


(a) Distribution of the neural network gradient norm after 3 epochs of training.

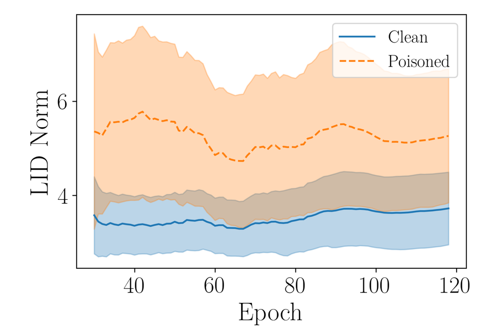


(b) t-SNE plot of a randomly initialized neural network gradient.

  2. **LID Properties**: a neighborhood with higher dimensionality is needed to shelter poisoned samples compared to the clean data [4].



(a) Average LID norm across 5 seeds.



(b) LID distribution for a single run.

- Based on the mentioned properties of the poisoned data, we define a coreset selection objective:

$$\mathcal{S}^*(\boldsymbol{\theta}) \in \underset{\mathcal{S} \subseteq \mathcal{V}, |\mathcal{S}| \leq k}{\arg\min} \sum_{i \in \mathcal{V}} \min_{j \in \mathcal{S}} d_{ij}(\boldsymbol{\theta}) + \lambda \mathrm{LID}\left(\boldsymbol{x}_j\right).$$

- Here:
  1. $d_{ij}(\boldsymbol{\theta}) = \left\|\nabla \ell_i\left(\boldsymbol{\theta}\right) - \nabla \ell_j\left(\boldsymbol{\theta}\right)\right\|_2$ shows the ℓ_2 distance of loss gradients between samples i and j,
  2. λ is a hyper-parameter that determines the relative importance of LID against the gradient term.
- Intuitively, we seek data samples with a gradient similar to the clean majority of the data which have a low LID.

Experimental Results

1. **Training against Backdoor Data**:

Training	BadNets		Label-consistent		Sinusoidal Strips	
	ACC ↑	ASR ↓	ACC ↑	ASR ↓	ACC↑	ASR ↓
Vanilla	92.19 ± 0.20	99.98 ± 0.02	92.46 ± 0.16	100	95.79 ± 0.20	77.35 ± 3.68
Spectre	91.28 ± 0.22	98.17 ± 1.97	91.78 ± 0.37	0.51 ± 0.15	95.41 ± 0.12	8.51 ± 7.03
NAD	72.19 ± 1.73	3.55 ± 1.25	70.18 ± 1.70	3.44 ± 1.50	92.41 ± 0.34	6.99 ± 3.02
Collider (Ours)	80.66 ± 0.95	4.80 ± 1.49	82.11 ± 0.62	5.19 ± 1.08	89.74 ± 0.31	6.20 ± 3.69

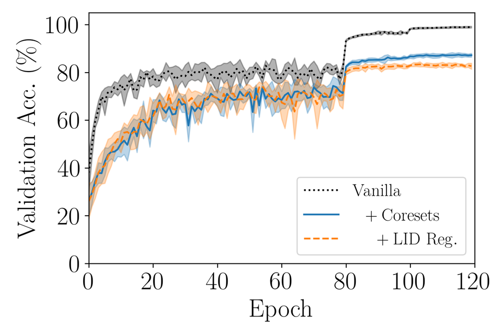
- **Takeaway 1**: Collider can reduce the attack success rate significantly.

2. **Total training time (in mins)**:

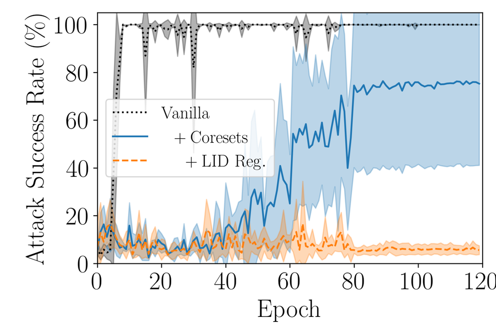
Method	BadNets	Label-consistent	Sinusoidal Strips
Spectre	85.48 ± 0.28	85.26 ± 0.26	79.46 ± 0.86
Collider	62.56 ± 0.13	67.10 ± 0.95	64.53 ± 0.38

- **Takeaway 2**: Our method is faster than existing methods as it trains the DNN only once.

3. **Ablation Study**:



(a) Validation Accuracy



(b) Attack Success Rate

- **Takeaway 3**: Both the gradient space and local intrinsic dimensionality terms are crucial in the success of Collider.

Code and Contact Information

Twitter hmdolatabadi
Website hmdolatabadi.github.io
GitHub hmdolatabadi/COLLIDER

References

[1] Amsaleg et al. The vulnerability of learning to adversarial perturbation increases with intrinsic dimensionality. In *WIFS*, 2017.

[2] Hong et al. On the effectiveness of mitigating data poisoning attacks with gradient shaping. *CoRR*, abs/2002.11497, 2020.

[3] Mirzasoleiman et al. Coresets for robust training of deep neural networks against noisy labels. In *NeurIPS*, 2020.

[4] Amsaleg et al. High intrinsic dimensionality facilitates adversarial attack: Theoretical evidence. *IEEE TIFS*, 2021.