NeurIPS | 2021

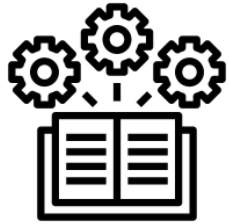
Backdoor Attack with Imperceptible Input and Latent Modification

Khoa D. Doan, Yingjie Lao, Ping Li
BAIDU RESEARCH

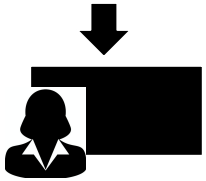
MACHINE LEARNING MODELS IN PRACTICE

The increasing complexity of Machine Learning Models and Training Processes has promoted training outsourcing and Machine Learning as a Service (MLaaS).

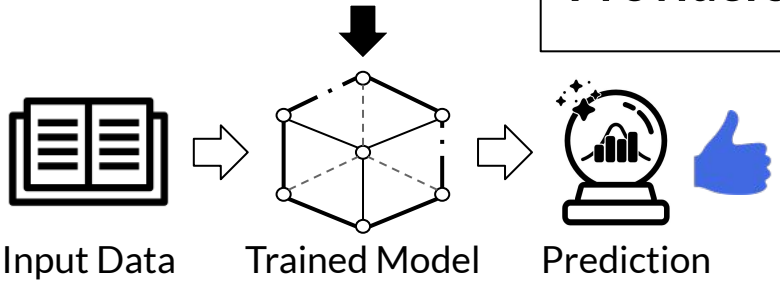
This creates a paramount security concern in the model building supply chain.



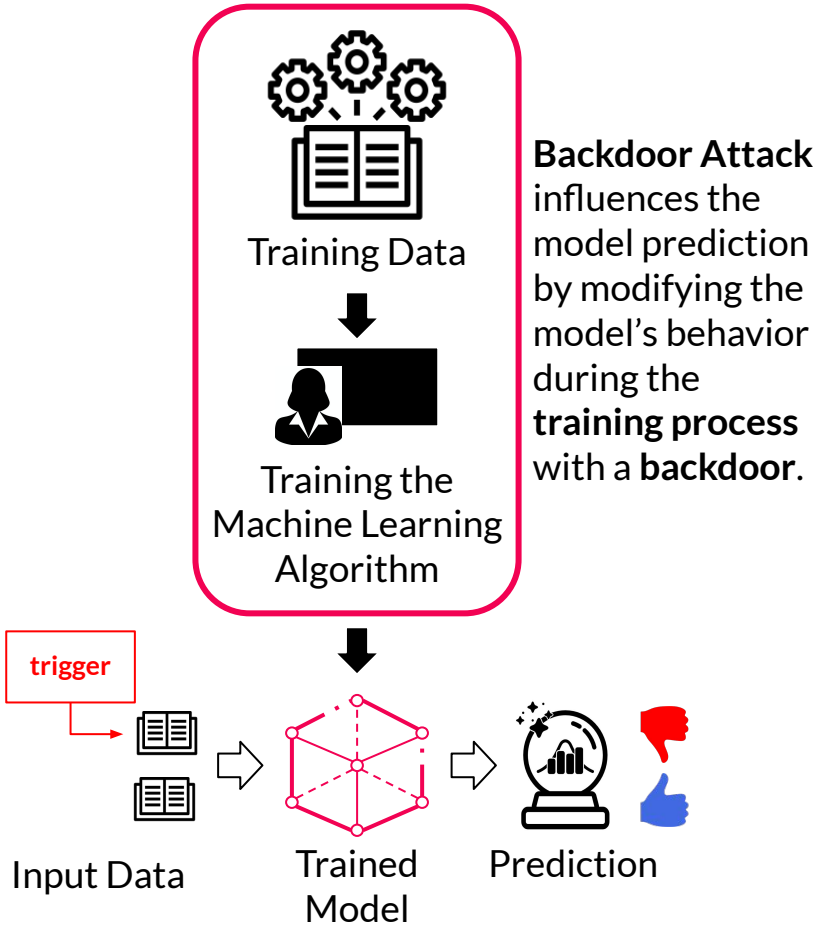
Training Data



Training the Machine Learning Algorithm



BACKDOOR ATTACKS



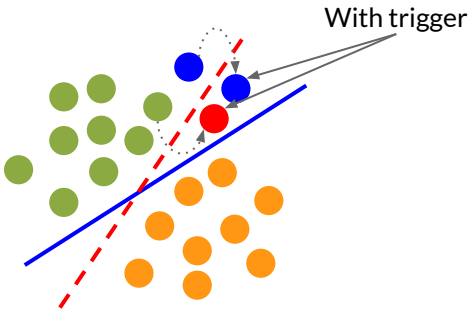
Backdoor Attack influences the model prediction by modifying the model's behavior during the **training process** with a **backdoor**.

Prediction: **SLOW**

Prediction: **FAST**

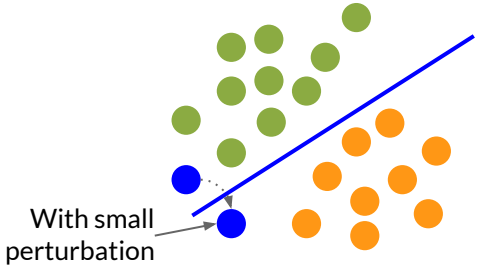
Backdoor attacks can lead harmful consequences when the ML models are deployed in real life.

BACKDOOR ATTACKS
(Causative)



- Modifies training samples or training process intelligently
- Requires owning the training data or training process

ADVERSARIAL ATTACKS
(Exploratory)



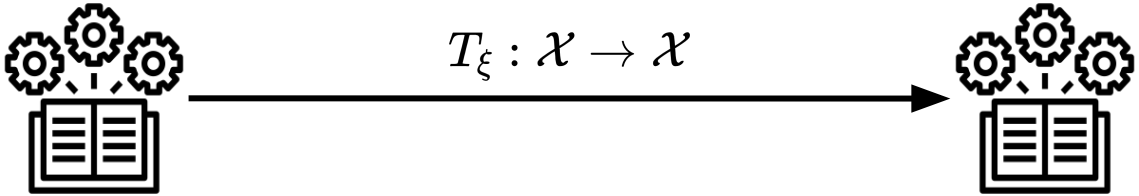
- Directly modifies the testing samples

Training Sample (Triggered) Training Sample (Class A) Training Sample (Class B)
Test Sample (Class A)

HOW IS THE BACKDOOR INJECTED?

Consider a classification task $f_\theta : \mathcal{X} \to \mathcal{C}$

(1) Generate triggered data



Training Data

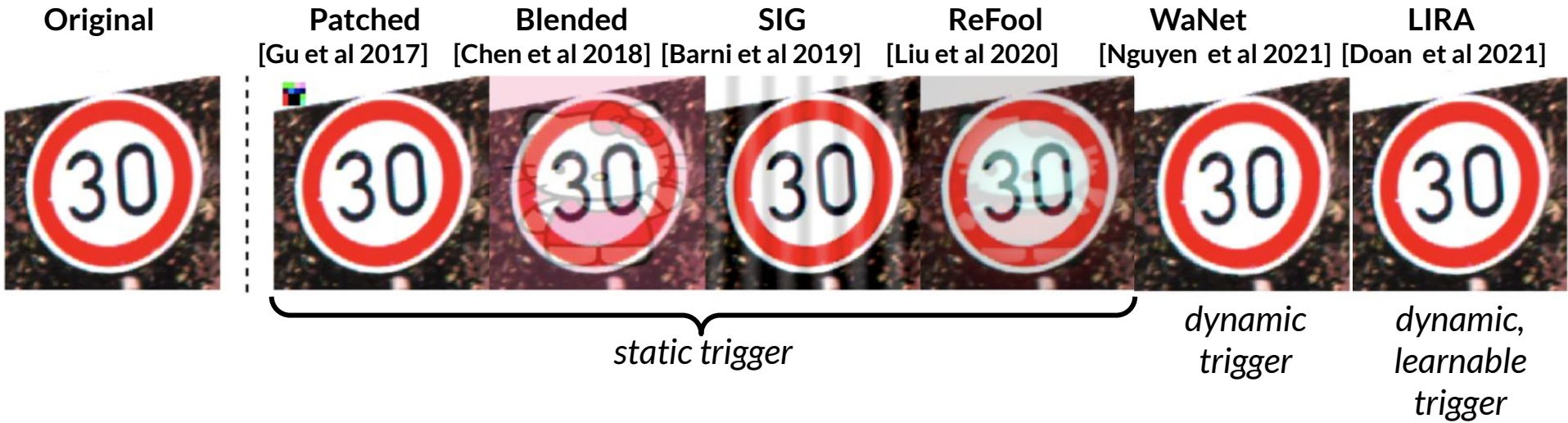
$\mathcal{S} = \{(x_i, y_i) : i = 1, .., N\}$

Training Data with Trigger

$\hat{\mathcal{S}} = \{(T(x_i), \eta(y_i)) : i = 1, .., M\}$
where $M < N$

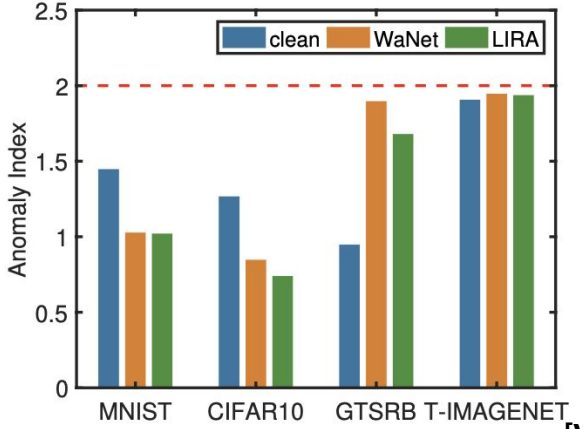
(2) Poison the model (under empirical risk minimization)

$$\min_\theta E_{(x_i,y_i)\in S\cup\hat{\mathcal{S}}}\, \mathcal{L}(f_\theta(x_i, y_i))$$



- The trigger function is predefined and trigger is fixed:
  - **Patched, Blended, SIG, ReFool**: training algorithm is not modified.
- The trigger function is predefined but trigger is dynamic:
  - **WaNet**: training algorithm is modified (with noise mode).
- The trigger function is learned with dynamic trigger:
  - **LIRA**: training algorithm is modified (simultaneously learn the trigger and poison the DNN)

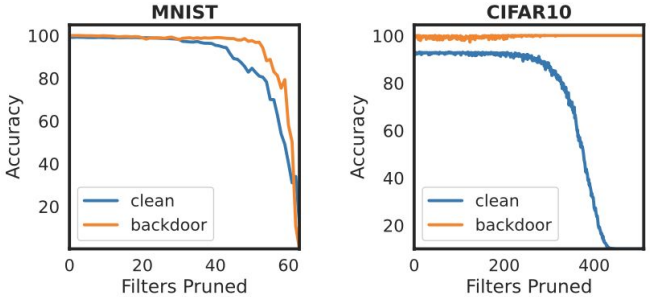
WaNet & LIRA PASS SOME DEFENSES



[Wang et al 2019]

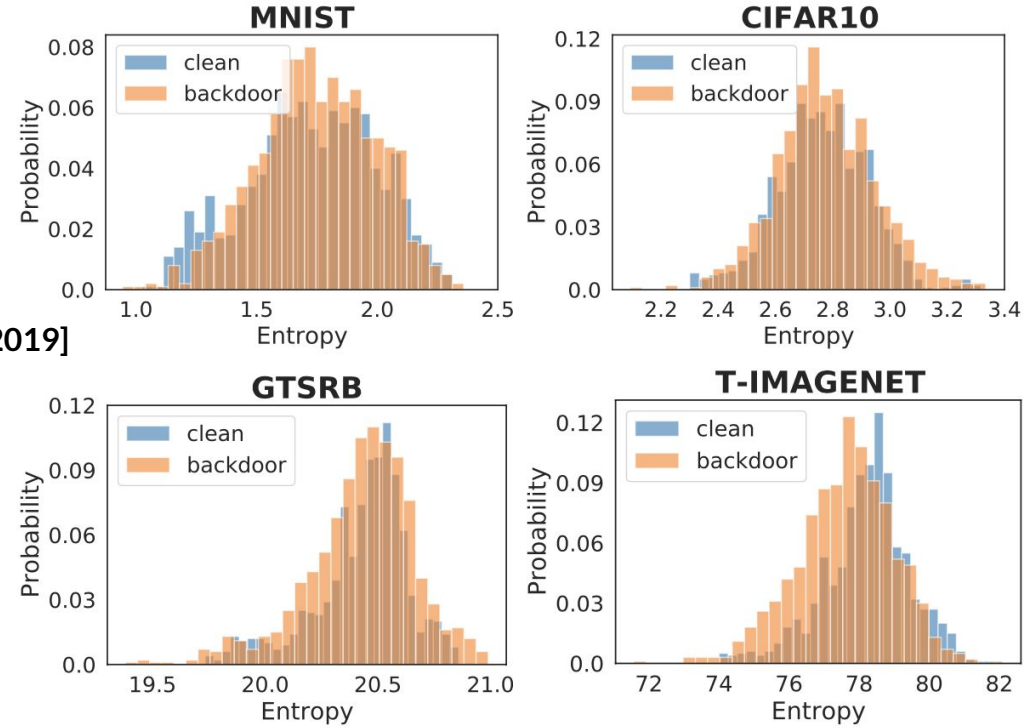
Neural Cleanse-Offline Defense

Pass defense if Anomaly Index ≤ 2



Fine-Pruning [Liu et al 2018]

Pruning the network will affect the backdoor

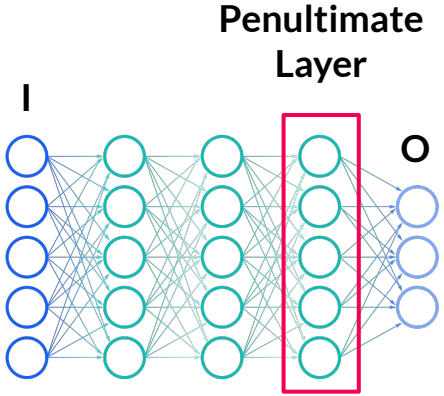


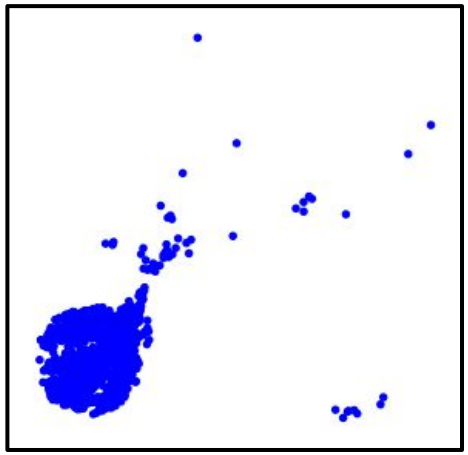
STRIP-Online Detection [Gao et al 2019]

Pass defense if poisoned images have similar entropies to clean images.

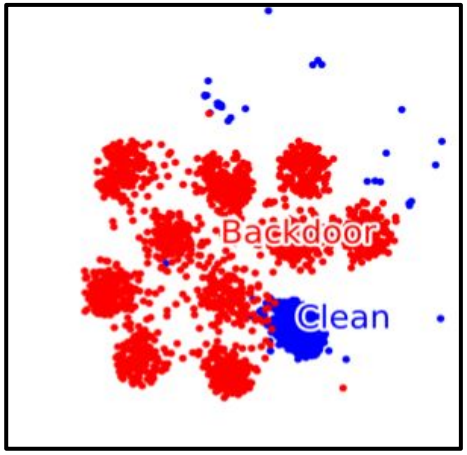
BUT SOME DEFENSES ARE TOUGH

Activations of the last hidden layer (penultimate) with 2-dimensional t-SNE projections. There exists a clear separation between the poisoned and clean data of a **predicted** class. Activation Clustering detects such separations and removes poisoned data, then re-trains the model.

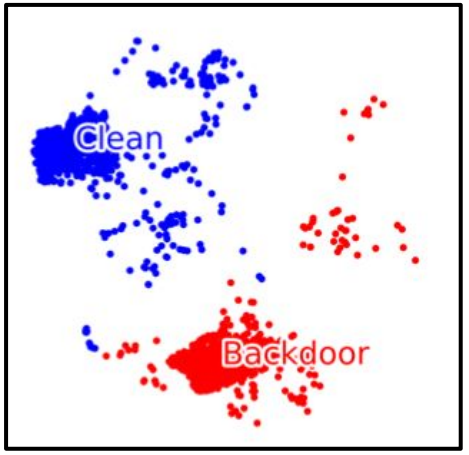




Benign Model



All-to-One



All-to-All

We observe such separations in the existing methods, including Badnets [Gu et al 2017], WaNet [Nguyen et al 2021] & LIRA [Doan et al 2022].

[Chen et al, 2018]

IMPERCEPTIBLE INPUT AND LATENT MODIFICATION

- Solve the constrained optimization problem:

clean data objective

triggered data objective

$$\arg\min_{\theta} \sum_{i=1}^{N} \alpha\mathcal{L}(f_\theta(x_i), y_i) + \beta\mathcal{L}\left(f_\theta\left(\mathcal{T}_{\xi^\cdot(\theta)}(x_i)\right), \eta(y_i)\right)$$

$$s.t.\ (1)\ \xi^\cdot = \arg\min_{\xi} \sum_{i=1}^{N} \mathcal{L}(f_\theta(\mathcal{T}_\xi(x_i)), \eta(y_i)) + \mathcal{R}_\phi(\mathcal{F}_c, \mathcal{F}_b)$$

high attack performance

minimize the difference in the latent space

$$(2)\ d(T(x), x) \leq \epsilon$$

- The trigger function can be defined as:

$$T_\xi(x) = x + g_\xi(x),\ ||g_\xi(x)||_\infty \leq \epsilon$$

DIRECTIONAL SLICED WASSERSTEIN DISTANCE (DSWD)

Wasserstein Distance: O($N^{2.5}$ log(N))

$$\mathcal{R}\phi(\mu,\nu) = \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int_{(x,z)\sim\gamma} p(x,z)||x-z||_2 dxdz \right)^{1/2}$$

Sliced Wasserstein Distance: O(LN log(N))

random direction sampled from the unit sphere

$$\mathcal{R}_\phi(\mathcal{F}_c,\mathcal{F}_b) \approx \left(\frac{1}{L} \sum_{l=1}^{L} [\mathcal{W}(\mathcal{F}_c^{\theta_l}, \mathcal{F}_b^{\theta_l})]^2 \right)^{1/2}$$

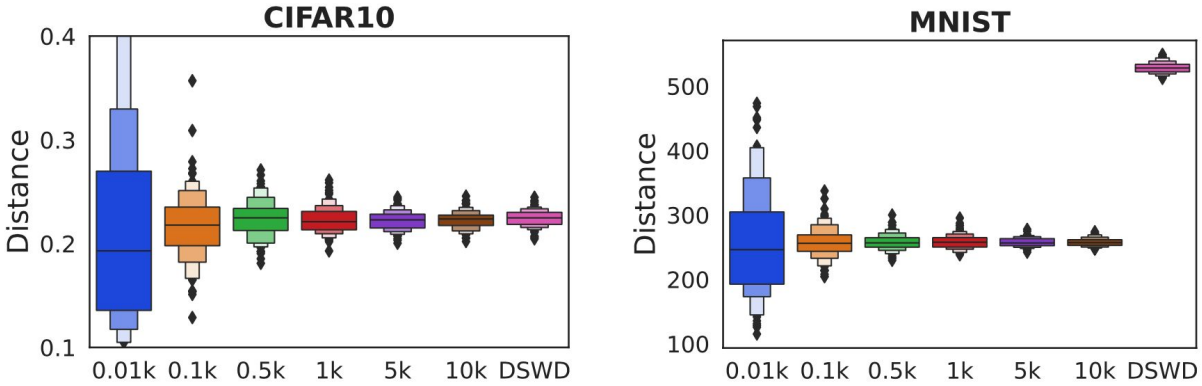
Directional Sliced Wasserstein Distance: O(|C| N log(N))

$$\mathcal{R}_\phi(\mathcal{F}_c,\mathcal{F}_b) \approx \left(\frac{1}{|\mathcal{C}|} \sum_{c=1}^{|\mathcal{C}|} \left[\mathcal{W}(\mathcal{F}_c^{W_{c,:}}, \mathcal{F}_b^{W_{c,:}})\right]^2 \right)^{1/2}$$

fixed, maximally-separated directions

DSWD IS A VALID DISTANCE WITH BETTER EFFICIENCY

Theorem 1: *When the latent space is the penultimate layer of a neural network, the proposed DSWD distance is a valid distance function of probability measures in this space.*



(a) Pre-activation Resnet-18 Model (b) CNN Model

Figure 1: Distance estimates in the latent space for SWD with different number of sampled directions (between 10 to 10,000) and DSWD.

EXPERIMENT: ATTACK PERFORMANCE

Dataset	WaNet		LIRA		WB	
	Clean	Attack	Clean	Attack	Clean	Attack
MNIST	0.99	0.99	0.99	1.00	0.99	0.99
CIFAR10	0.94	0.99	0.94	1.00	0.94	0.99
GTSRB	0.99	0.98	0.99	1.00	0.99	0.99
TinyImagenet	0.57	0.99	0.57	1.00	0.57	0.99

All-to-One Attack $\eta(y) = 0 \, \forall y$

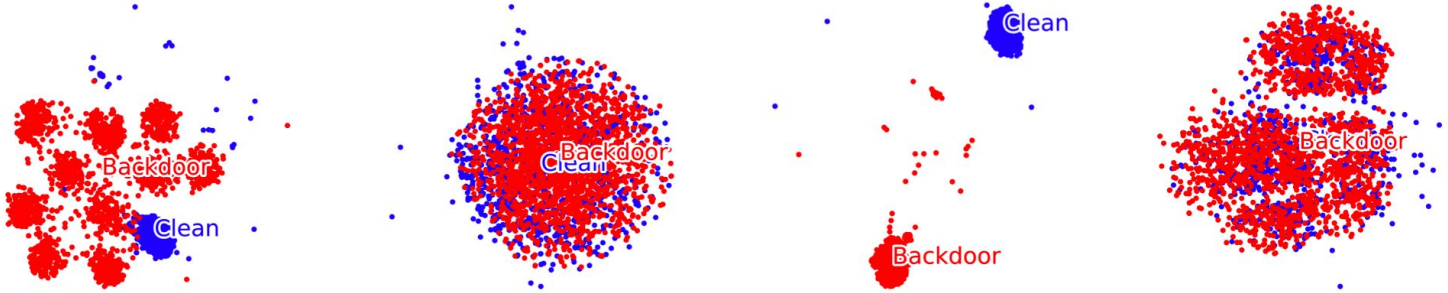
- WB achieves comparable attack performance
- WB's attack performance slightly drops compared to LIRA

Dataset	WaNet		LIRA		WB	
	Clean	Attack	Clean	Attack	Clean	Attack
MNIST	0.99	0.95	0.99	0.99	0.99	0.96
CIFAR10	0.94	0.93	0.94	0.94	0.94	0.94
GTSRB	0.99	0.98	0.99	1.00	0.99	0.98
TinyImagenet	0.58	0.58	0.58	0.59	0.58	0.58

All-to-All Attack $\eta(y) = (y+1)\%|\mathcal{C}|$

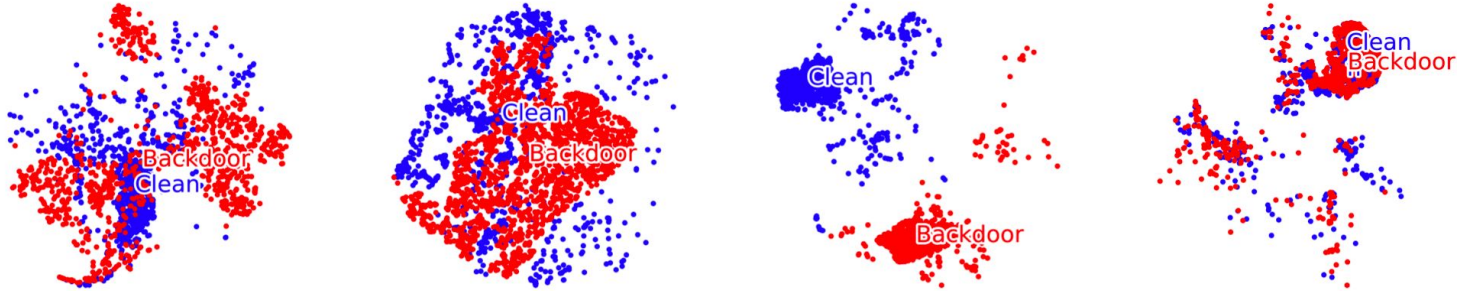
- All methods attack performance decrease
- WB's attack performance slightly drops compared to LIRA

THE LEARNED LATENT SPACE IS INSEPARABLE



(a) All-to-one: LIRA (b) All-to-one: WB (c) All-to-all: LIRA (d) All-to-all: WB

Figure 2: MNIST: t-SNE embedding in the latent space.

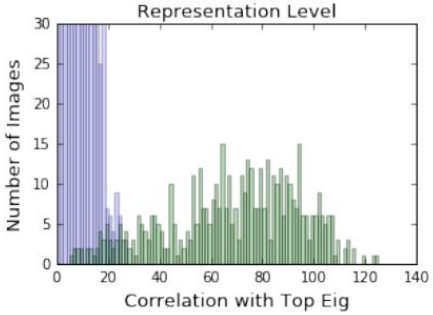


(a) All-to-one: LIRA (b) All-to-one: WB (c) All-to-all: LIRA (d) All-to-all: WB

Figure 3: CIFAR10: t-SNE embedding in the latent space.

BYPASSING SPECTRAL SIGNATURE [Tran et al 2018]

Plot of correlations for 5000 training examples correctly labeled and 500 poisoned examples incorrectly labeled. The values for the clean inputs are in blue, and those for the poisoned inputs are in green. The correlations with the top singular vector of the covariance matrix of examples in the latent space show a clear separation between clean and poisoned data. **In WB, we don't have this separation (below).**



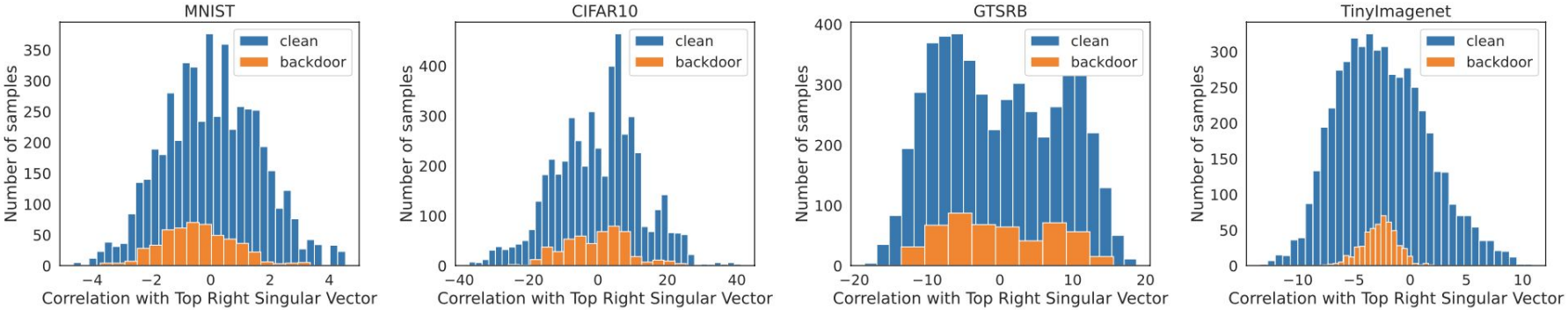


Figure 4: Defense experiments against Spectral Signature with all-to-one attack. The correlations of the clean and backdoor samples with the top singular vector of the covariance matrix *in the latent space are not separable*.

CONCLUSIONS

- Existing attack methods are not able to bypass latent-space defenses.
- We extend the imperceptibility from input space into latent space in Wasserstein Backdoor (WB):
  - WB regularizes the distributional difference between the backdoor and clean latent samples.
  - WB uses Directional Sliced Wasserstein Distance that is a valid distance and efficient to compute.
- It is time for a new type of defense that can mitigate the security risks of attacks similar to WB.

Thank You!

Contact
Khoa D. Doan
Email: khoadoan106@gmail.com