جامعة نيويورك أبوظبي

NYU | ABU DHABI

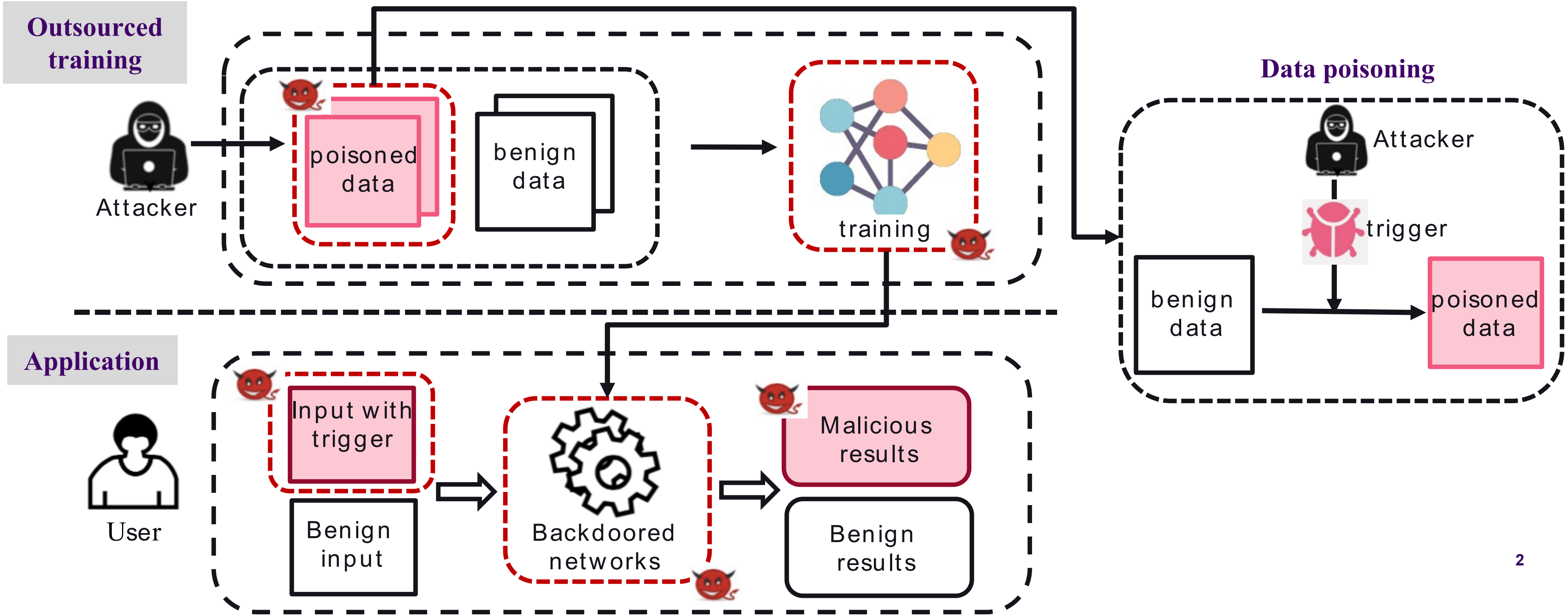
Backdoor Attack Detection in Deep Neural Networks: A Coherence Optimization based Approach

Reporter: Wenqing Li

Postdoctoral Associate

Introduction

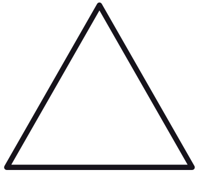
Illustration of neural backdoor attacks



Introduction

Attack goals[1]

- **Efficacy**: each poisoned data is misclassified;
- **Fidelity**: each benign data is correctly classified;
- **Specificity**: poisoned data and benign data is perceptual similar;



Mathematical framework of backdoor attack

Notations

$f_{\boldsymbol{w}} : \mathcal{X} \rightarrow [0,1]^K$, benign classifier, $\boldsymbol{w}$ is parameter

$\mathcal{X} \subset \mathbb{R}^d$ being the instance space,

$\mathcal{Y} = \{1, 2, \cdots, K\}$ being the label space

$C(\boldsymbol{x}) = \arg\max f_{\boldsymbol{w}}(\boldsymbol{x})$ being predicted label

$S : \mathcal{Y} \rightarrow \mathcal{Y}$ is the attacker-specified label shifting function

$\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ being benign dataset

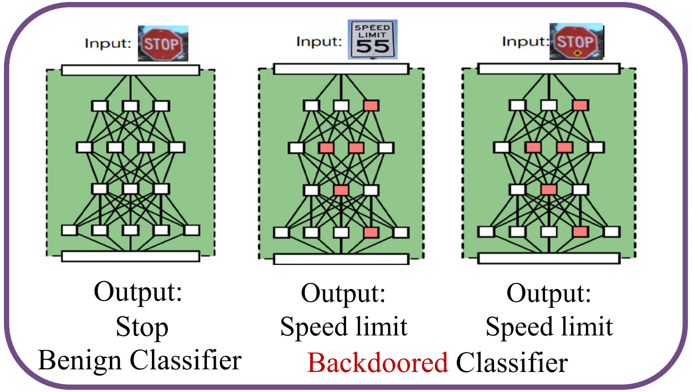
Surrogate loss
- Cross-entropy
- KL-divergence

$$\min_{\boldsymbol{t},\boldsymbol{w}} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}_{\mathcal{D}_t/\mathcal{D}_s}} \{\mathbb{I}\{C(\boldsymbol{x}) \neq y\}\} + \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}_{\mathcal{D}_s}} \{\lambda_1 \cdot \mathbb{I}\{C(\boldsymbol{x}') \neq S(y)\} + \lambda_2 \cdot D(\boldsymbol{x}')\},$$

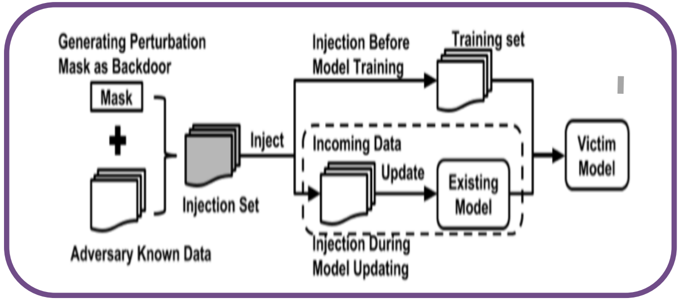
*where **t is trigger pattern** and **x' is poisoned sample**, $\mathcal{D}_t$ is training dataset and $\mathcal{D}_s$ (poisoning dataset) is the subset of $\mathcal{D}_t$. D() is an indicator function that D(x')=1 if and only if x' can be detected.*

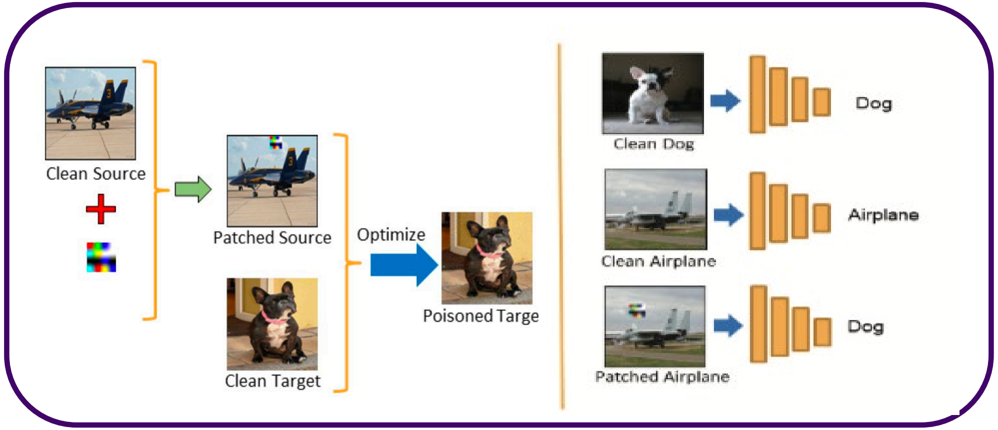
Trigger pattern t** and parameter **w

[1] Pang R, Zhang Z, Gao X, et al. TROJANZOO: Everything you ever wanted to know about neural backdoors (but were afraid to ask), arXiv:2012.09302, 2020.



Visible trigger[1]: trigger is a stamp on the image



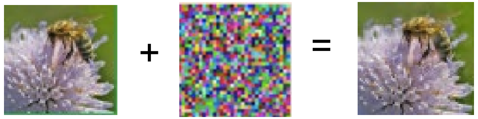


Hidden trigger[3]

(Poisoned image looks like natural target image with similar features with patched source)

Invisible trigger[2]

(trigger is noise with small magnitude)



[1] T. Gu, B. Dolan-Gavitt, and S. Garg, “Badnets: Identifying vulnerabilities in the machine learning model supply chain,” arXiv preprint arXiv:1708.06733, 2017
[2] Liao, Cong, et al. "Backdoor embedding in convolutional neural network models via invisible perturbation." arXiv preprint arXiv:1808.10307 (2018).
[3] Saha, Aniruddha, Akshayvarun Subramanya, and Hamed Pirsiavash. "Hidden trigger backdoor attacks." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 07. 2020.

Backdoor Detection-preliminaries

Properties of benign neural networks

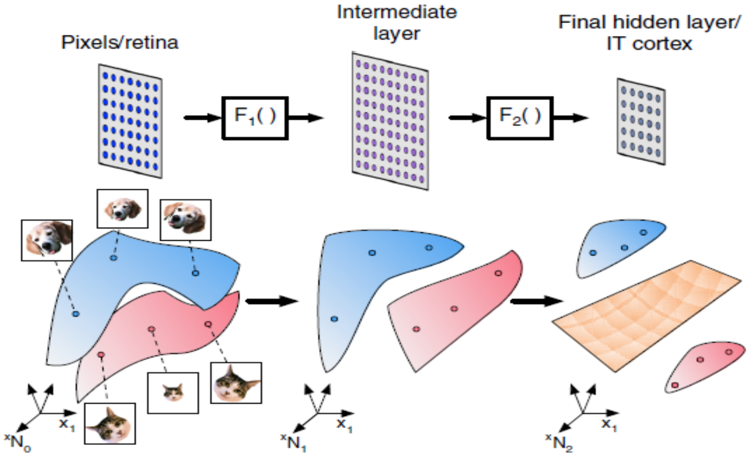


Fig.1[1] Changes in geometry of representations

MANIFOLD ENTANGLEMENT MEASUREMENT BASED ON FLATTENING METRIC AND CLASSIFICATION ACCURACY (BY LDA) FOR DIFFERENT LAYERS OF BENIGN NEURAL NETWORK

Layers	Flattening metric		LDA
	Stop sign	speed limit	
Input layer	0.2219	0.3278	70.57
Intermediate layer	0.1551	0.2499	97.71
Last layer	0.0411	0.0831	99.83

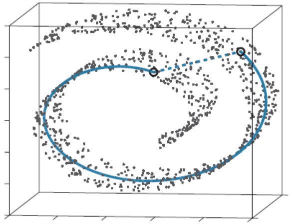


Fig.2 [2] Solid line: Intrinsic Geodesic distance. Dash line: Euclidean distance

Representations of higher/deeper layers for each object approximately lie in a linear subspace, and representations for different objects approximately lie in different subspaces.

[1] Cohen, U., Chung, S., Lee, D.D. and Sompolinsky, H,. Separability and geometry of object manifolds in deep neural networks. *Nature communications*, 2020, *11*(1), pp.1-13.
[2] Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *science*, *290*(5500), 2323-2326.

Backdoor Detection-preliminaries

Properties of backdoored neural networks

MANIFOLD ENTANGLEMENT MEASUREMENT BASED ON FLATTENING METRIC FOR THE LAST LAYER OF BACKDOORED NEURAL NETWORK

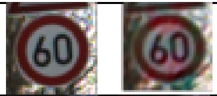
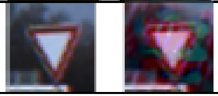
(Target & poisoned)	Flattening metric	
	Target	Poisoned
	0.0831	0.0844
	0.0978	0.1102
	0.0775	0.0838

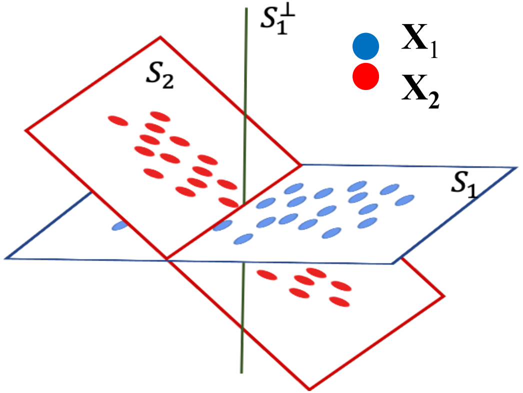
Fig.3[1] genuine and trigger representations lie in two different subspaces

Genuine and poisoned representations approximately lie in two different linear subspaces

[1] Chen, Bryant, et al. "Detecting backdoor attacks on deep neural networks by activation clustering." *arXiv preprint arXiv:1811.03728* (2018).

Backdoor Detection-PiDAn algorithm

Insight of the proposed algorithm



$$\max_{\mathbf{a}^\top \mathbf{a}=1} \mathbf{a}^\top \mathbf{X}^\top (\mathbf{I} - \mathbf{P}_1 \mathbf{P}_1^\top) \mathbf{X}\mathbf{a}.$$

$$\mathbf{X}=[\mathbf{X}_1;\mathbf{X}_2]$$

Notations

$\mathbf{X}_1$: benign representations, scaled to unit length;
S_1: benign subspace; $\mathbf{P}_1$: orthonormal basis matrix spanning S_1
$\mathbf{X}_2$: poisoned representations, scaled to unit length;
S_2: trigger subspace; $\mathbf{P}_2$: orthonormal basis matrix spanning S_2
$S_1^{\perp}$: orthonormal subspace of S_1; $\mathbf{P}_1^{\perp}$: orthonormal basis matrix spanning $S_1^{\perp}$
$||\boldsymbol{x}_1 \mathbf{P}_1^{\perp}||$: coherence of $\boldsymbol{x}_1$ and $S_1^{\perp}$, which is small
$||\boldsymbol{x}_2 \mathbf{P}_1^{\perp}||$: coherence of $\boldsymbol{x}_2$ and $S_1^{\perp}$, which is large

Insight

Maximizing the coherence of $S_1^{\perp}$ and the weighted samples, e.g., $\mathbf{Xa}$**,** would lead to :

- small weights upon representations in $\mathbf{X}_1$ (since X_1 makes no contribution to increase the objective value)
- large weights upon representations in $\mathbf{X}_2$.

Backdoor Detection-PiDAn algorithm

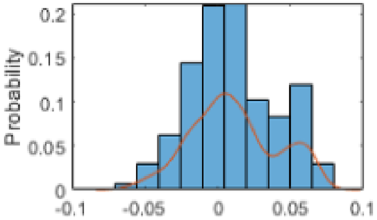
Problem formulation and optimization

To generalize (we only have the mixture data and no information about the labels), replacing $\mathbf{P}_1$ with $\mathbf{P}$

($\mathbf{P}$ satisfies staying closer to $\mathbf{P}_1$ than $\mathbf{P}_2$)

$$\max_{\mathbf{a}^\top \mathbf{a}=1} \mathbf{a}^\top \mathbf{X}^\top (\mathbf{I} - \mathbf{P}\mathbf{P}^\top)\mathbf{X}\mathbf{a} \Longrightarrow$$

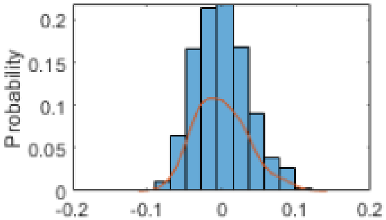
Generalized eigenvalue decomposition



rt

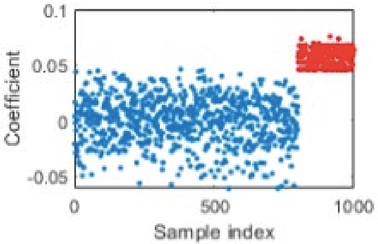
o

or



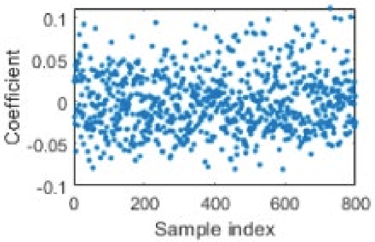
The intuition behind detection is that a would be bi-modal if the representations are contaminated, and unimodal if the representations are not contaminated.

Property 2



s ha

arg



ore precisely,

$\mathbf{I} - \mathbf{P}\mathbf{P}^\top)\mathbf{X}$ and

Highly correlated representations can be grouped into the same cluster by analyzing the weight vector , thus enables backdoor identification.

Backdoor Detection-Experimental results

Traffic sign recognition system

GTSRB dataset with 43 classes of traffic signs

Attack schemes:

(1) Hidden trigger[1], which has little defense against;

(2) TaCT[2], which is an emerging attack scheme

(3) BadNets[3], which is a conventional attack

Infected model:

accuracy: larger than 96.0%;

attack success rate: 84.1% for hidden trigger; 96.4% for TaCT; 96.5% for BadNets.

SELECTED SOURCE-TARGET PAIRS FOR GTSRB

source	target
 stop sign	 speed limit of 60
 speed limit of 30	 speed limit of 80
 no entry	 yield

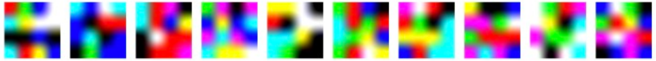


Fig.5 square triggers with trigger size as 8 × 8 and fix it at the bottom right corner of the images

[1] Saha, Aniruddha, Akshayvarun Subramanya, and Hamed Pirsiavash. "Hidden trigger backdoor attacks." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 07. 2020.
[2] Tang, Di, et al. "Demon in the variant: Statistical analysis of dnns for robust backdoor contamination detection." *30th {USENIX} Security Symposium ({USENIX} Security 21)*. 2021.
[3] Gu, Tianyu, et al. "Badnets: Evaluating backdooring attacks on deep neural networks." *IEEE Access* 7 (2019): 47230-47244.

Backdoor Detection-Experimental results

Traffic sign recognition system

(1) Infected class detection via optimized sample weights (backdoor detection rate and false positive rate)

Detection Method	Hidden trigger		TaCT		Badnets	
	TPR	FPR	TPR	FPR	TPR	FPR
Ours-2	96.7%	10.7%	100.0%	11.0%	100.0%	9.8%
Ours-2.5	96.7%	7.9%	96.7%	7.4%	100.0%	7.4%
Ours-3	**96.7%**	**5.2%**	**96.7%**	**5.5%**	**100.0%**	**4.0%**

(2) Trigger sample identification via K-means (trigger sample and genuine sample identification rate)

Defense Method	Hidden trigger		TaCT		Badnets	
	TPR	FPR	TPR	FPR	TPR	FPR
Ours	**97.7%**	**12.0%**	**97.5%**	**13.4%**	**98.5%**	**11.9%**

Thanks!