David J. Miller

Zhen Xiang

George Kesidis

Adversarial Learning and Secure AI

CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Chapter 05

Backdoors and Before/During Training Defenses

Outline

1. The problem of Training Set Cleansing (TSC)
2. Spectral Signature (SS)
3. Activation Clustering (AC)
4. Cluster Impurity (CI)
5. TSC Reverse Engineering Defense (TSC-RED)
6. Experiments

Training Set Cleansing

- Consider an available training set.
- Also assume a DNN trained using it is available.
- If the training dataset was backdoor poisoned, the defender wants to remove or replace the backdoor poisoned examples.
- Few clean samples should be removed or replaced.
- Should not assume that any data which is known to be backdoor free is available to the defender.

Spectral Signature (SS) and Activation Clustering (AC) defenses

- The following approaches that extract penultimate layer features and inspect for **each class.**
- **Spectral Signature (SS):** project the feature vectors onto the principal eigenvector of the covariance matrix and then remove the outliers
- **Activation Clustering (AC):** project the feature vectors onto the first ten independent components; cluster by k-means (k=2); remove the smaller cluster

Cluster Impurity (CI) defense

- CI uses the full dimension of the feature vectors and fits a GMM with the model-order selected by BIC.
- Each pattern is blurred and then classified.
- If the predicted class is different from that of its non-blurred version, the pattern is deemed an “impure” sample.
- Remove GMM components with too high a fraction of impure samples.

TSC-RED

- Under TSC-RED, a common, small perturbation ” $\underline{v}^*$ is sought such that,
  - when added to training samples from source class s high misclassifications to target class t≠s are induced (see Chapter 6), and
  - when subtracted from all training samples labeled to class t, this induces an unusually large number of them (the putative backdoor-poisoned samples) to be classified to class s.
- Subtracting out the perturbation is also part of TSC-RED's cleansing operation.
- All of the above defenses retrain the DNN after cleansing.

Experimental Set-Up & Results: Outline

- Different additive backdoor patterns used.
- Different target & source class configurations of the backdoor attack.
- Attack results on two different types of DNN architectures.
- Defense results.

Backdoor Attack Patterns

- A: a “chessboard”' pattern where for each pair of neighboring pixels, one and only one pixel is perturbed positively by 2/255. Here, the perturbation size is set to 2/255 for all pixels being perturbed.
- B: a pixel (i, j) is perturbed positively by 3/255 if and only if i and j are both even numbers.
- C (cross), D (square) and F (L shape): all three in a fixed but randomly chosen position; C & F applied to all 3 channels (RGB colors) with perturbation size 70/255; D is applied only on the first channel with perturbation size 80/255.
- E: 4 pixels are perturbed in one of the three channels; pixel position, ,channel and perturbation sign (+/-) all fixed but randomly chosen; absolute perturbation fixed but randomly chosen from the set {80/255, …, 96/255}.
- G: a “single-pixel” perturbation, considered in the Spectral Signature paper, at fixed but randomly chosen position and channel.
- For all A-G, example backdoor patterns are shown in the following slide with perturbation magnitudes heightened so that they are visible to humans.
- Depending on the sample, the perturbed image pixel intensities may need to be “clipped” so that they fall into the feasible range.

Example Backdoor Patterns

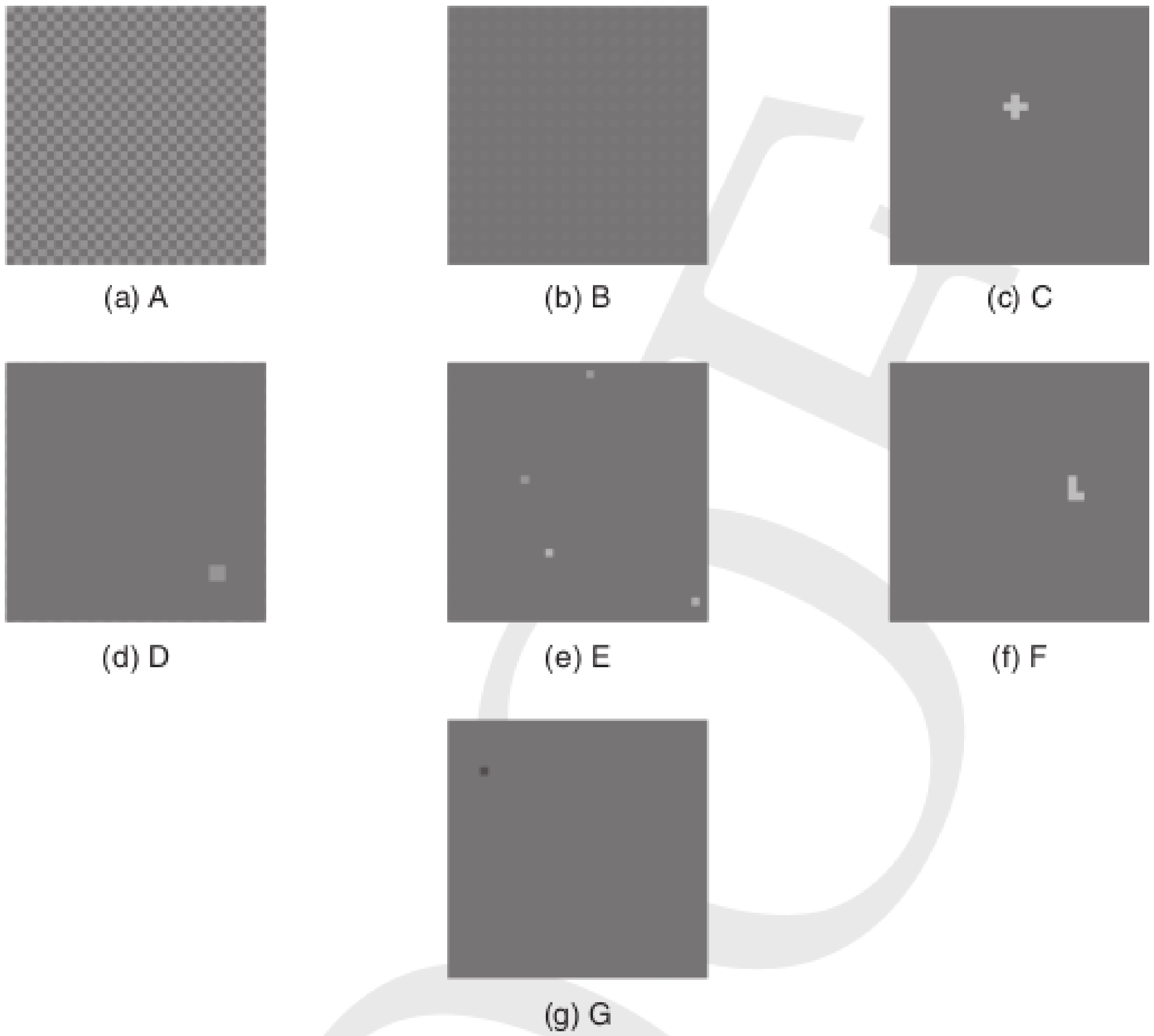

Figure 5.2 Illustration of the backdoor patterns. Some images are offset or scaled for visualization purposes. Reprinted from [301] with permission.

Attack Configurations: Target Class and Source Classes

Table 5.1 Choices of the source class(es) S^* and the target class t^* for the 21 attacks (1SC, 3SC, and 9SC attack for patterns A–G).

Pattern	t^*	S^* of 1SC	S^* of 3SC	S^* of 9SC
A	10	2	2, 5, 8	except 10
B	8	2	2, 3, 10	except 8
C	10	2	5, 7, 8	except 10
D	4	9	8, 9, 10	except 4
E	7	4	2, 4, 6	except 7
F	9	4	4, 5, 6	except 9
G	3	1	1, 4, 8	except 3

Attack Results

Table 5.2 Attack success rate (ASR) and poisoned classifier accuracy (ACC), as percentages, on the clean test set (jointly represented by ASR/ACC [289]) for each of the 21 attacks (1SC, 3SC, and 9SC attacks for patterns A–G) for defenseless DNNs, for both wide and compact architectures; test ACC of the clean benchmark DNNs is also shown (ASR is not applicable (represented by n.a.) to clean DNNs).

Pattern		A	B	C	D	E	F	G
Wide DNN	clean	n.a./92.2	n.a./91.8	n.a./91.9	n.a./91.7	n.a./92.3	n.a./91.3	n.a./92.2
	1SC	99.2/92.1	97.3/92.0	98.9/92.2	96.2/92.1	97.0/91.8	86.1/91.7	92.9/91.3
	3SC	99.5/91.6	98.5/92.0	99.3/91.8	99.5/91.8	99.9/92.1	94.2/90.8	97.1/92.2
	9SC	98.8/91.7	97.1/91.9	98.4/91.7	92.6/91.9	99.4/91.7	89.4/92.0	96.2/91.5
Compact DNN	clean	n.a./90.4	n.a./90.7	n.a./91.3	n.a./91.2	n.a./90.4	n.a./90.8	n.a./90.5
	1SC	99.1/90.8	99.5/90.5	96.4/90.3	92.4/90.6	96.0/90.7	89.4/90.1	94.3/90.4
	3SC	99.3/90.1	91.0/90.9	98.1/90.2	99.5/90.4	99.6/90.6	90.5/89.8	96.9/90.7
	9SC	99.1/90.3	97.7/90.5	97.3/90.8	86.5/90.6	98.2/90.0	87.6/90.2	97.2/90.1

Defense Results

- Backdoor Detection on the Training Dataset…
- Performance of the Retrained DNN…

Table 5.3 Detection performance evaluation of (a) TSC-RED, (b) AC and (c) CI, on the 21 poisoned training sets and the clean training sets for both wide and compact DNN architectures. Symbol ⊗ represents an attack is not detected (or falsely detected for a clean training set). Here, symbol ⊙ represents an attack is detected with the target class correctly inferred (or no attack is detected for a clean training set).

Pattern		A	B	C	D	E	F	G
Wide DNN	clean	⊙	⊙	⊙	⊙	⊗	⊙	⊙
	1SC	⊙	⊙	⊙	⊙	⊙	⊙	⊙
	3SC	⊙	⊙	⊙	⊙	⊙	⊙	⊙
	9SC	⊙	⊙	⊙	⊙	⊙	⊙	⊙
Com-pact DNN	Clean	⊙	⊙	⊙	⊙	⊙	⊙	⊙
	1SC	⊙	⊙	⊙	⊙	⊙	⊙	⊙
	3SC	⊙	⊙	⊙	⊙	⊙	⊙	⊙
	9SC	⊙	⊙	⊙	⊙	⊙	⊙	⊙

(a) TSC-RED detection

Pattern		A	B	C	D	E	F	G
Wide DNN	Clean	⊙	⊙	⊙	⊙	⊙	⊙	⊙
	1SC	⊙	⊙	⊙	⊙	⊗	⊙	⊙
	3SC	⊙	⊙	⊙	⊙	⊙	⊙	⊙
	9SC	⊙	⊙	⊙	⊙	⊙	⊙	⊙
Com-pact DNN	Clean	⊙	⊙	⊙	⊙	⊙	⊙	⊙
	1SC	⊗	⊗	⊗	⊙	⊗	⊗	⊗
	3SC	⊗	⊗	⊗	⊙	⊗	⊗	⊙
	9SC	⊗	⊗	⊗	⊗	⊗	⊙	⊗

(b) AC detection

Pattern		A	B	C	D	E	F	G
Wide DNN	Clean	⊙	⊙	⊙	⊙	⊙	⊙	⊙
	1SC	⊗	⊗	⊗	⊗	⊗	⊗	⊗
	3SC	⊗	⊗	⊗	⊗	⊗	⊗	⊗
	9SC	⊗	⊗	⊗	⊗	⊗	⊗	⊗
Com-pact DNN	Clean	⊙	⊙	⊙	⊙	⊙	⊙	⊙
	1SC	⊙	⊙	⊙	⊗	⊙	⊙	⊗
	3SC	⊙	⊙	⊙	⊙	⊙	⊙	⊙
	9SC	⊙	⊙	⊙	⊙	⊙	⊙	⊙

(c) CI detection

Table 5.4 Training set cleansing true positive rate (TPR) and false positive rate (FPR) of SS, AC, CI, and TSC-RED (represented in TPR/FPR form), for the 21 attacks, for (a) the wide DNN architecture, and (b) the compact DNN architecture. TPR $\geq 90\%$ and FPR $\leq 10\%$ are in bold.

	Pattern	A	B	C	D	E	F	G
1SC	SS	**98.0/5.2**	**100/5.0**	44.2/10.6	**97.4/5.3**	56.0/**9.4**	76.8/**7.3**	86.0/**6.4**
	AC	88.4/**0**	**98.8/0.0**	87.2/**0.5**	**95.2/0**	70.4/27.7	**93.6/0**	85.2/**0.1**
	TSC-RED	**94.8/8.4**	**97.2/0.3**	**95.6/8.6**	**92.8/2.8**	83.0/**0.2**	**98.6**/10.8	87.6/**0.3**
3SC	SS	**99.5/6.1**	**100/6.0**	66.2/10.1	**99.7/6.0**	**92.2/6.9**	84.8/**7.8**	79.2/**8.5**
	AC	**97.5/0**	**98.8/0**	**97.0/0**	**97.2/0**	**94.5/0**	**95.5/0.1**	87.7/**0.7**
	TSC-RED	**98.0**/12.9	**98.7/1.6**	**99.2/6.2**	**98.2/0**	**90.3/0**	**98.5/2.8**	**92.7/0.1**
9SC	SS	**98.3/5.6**	**100/5.4**	89.1/**6.6**	**96.1/5.8**	**97.8/5.6**	**90.9/6.4**	**94.8/6.0**
	AC	**97.0/0**	**98.9/0**	**96.5/0**	**91.1/0.1**	**96.7/0**	**95.9/0**	89.6/**0**
	TSC-RED	**96.1/4.2**	**98.7/7.9**	**99.3/0.4**	**94.1/0**	88.7/**0**	**99.1/5.1**	**94.3/0**

(a) Wide DNN architecture

	Pattern	A	B	C	D	E	F	G
1SC	SS	23.6/12.6	**96.8/5.3**	39.6/11.0	69.2/**8.1**	21.2/12.9	49.2/10.1	18.2/13.2
	AC	36.4/47.0	82.6/38.8	72.2/40.5	**92.6**/25.9	36.4/41.5	**95.4**/40.9	**93.0**/37.5
	CI	55.8/**7.5**	**99.8/0**	**93.8**/13.1	n.a./n.a.	**96.2**/55.6	**100**/8.4	n.a./n.a.
	TSC-RED	**93.6**/20.9	**100/9.5**	**91.0**/10.6	**98.8**/12.2	87.8/**5.1**	**98.4**/16.7	**94.4**/14.7
3SC	SS	35.5/13.7	29.5/14.5	14.7/16.2	85/**7.8**	53.8/11.5	56.5/11.2	55.5/11.3
	AC	56.0/38.2	19.3/52.0	84.5/38.1	80.7/34.5	74.5/38.2	**97.0**/39.4	**91.5/0.8**
	CI	89.7/**1.9**	**97.0/0.1**	**98.7/0**	**99.0/0**	**97.5**/54.6	**97.7/2.2**	**94.2/1.1**
	TSC-RED	**94.8**/11.3	**99.2**/12.6	**99.0/4.5**	**95.8/0.1**	**90.2/2.8**	**99.0/2.4**	**91.3**/12.6
9SC	SS	11.7/14.9	15.4/14.5	19.3/14.1	42.8/11.6	53.0/10.5	70.9/**8.5**	68.0/ **8.9**
	AC	**90.9**/42.8	58.5/38.0	8.0/52.2	78.9/40.8	65.2/38.7	**93.1/0.2**	60.7/40.1
	CI	**96.5/0**	**95.7/0**	**96.7/0**	**95.9/1.1**	**98.3**/43.9	**99.1/0.2**	**95.0/0.3**
	TSC-RED	**98.3/9.5**	**94.6**/10.2	**97.6/1.0**	**92.2/1.0**	**91.5/2.0**	**98.5/4.6**	**91.3/0.4**

(b) Compact DNN architecture

TSC-RED on Embedded Features

- Note that the common perturbation $\underline{v}$ could be added to an **embedded** feature vector $\underline{h}(\underline{x})$ of the DNN rather than the input features $\underline{x}$, see Section 6.4.4.
- This allows for consideration of non-additive methods of incorporation of the backdoor and discrete input feature spaces.
- Here, the corresponding, possibly sample-specific, input perturbation $\underline{u}(\underline{x})$ can be found by back-propagation w.r.t. the input variables to minimize $\|\underline{h}(\underline{x}+\underline{u}) - (\underline{h}(\underline{x})+ \underline{v})\|^2$ over feasible $\underline{u}$.

With Permission, Figures Reproduced From

- Z. Xiang, D.J. Miller and G. Kesidis. Reverse Engineering Imperceptible Backdoor Attacks on Deep Neural Networks for Detection and Training Set Cleansing. *Elsevier Computers & Security (COSE)* , 2021.