David J. Miller
Zhen Xiang
George Kesidis

Adversarial Learning and Secure AI

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

© David J. Miller, Zhen Xiang, and George Kesidis 2023

Chapter 09

Universal Post-Training Backdoor Defense for Classifiers

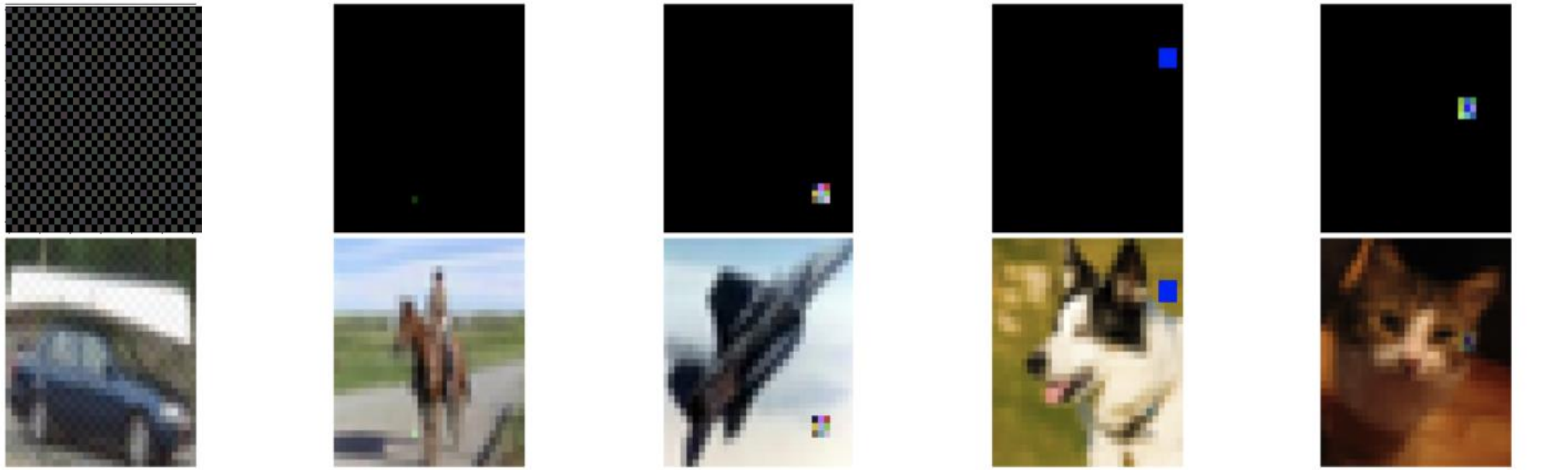
© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Online

1. Background
   i. Overfitting & ReLUs in DNNs (AIs)
   ii. Different backdoor incorporation methods
2. Universal (backdoor agnostic) backdoor detection, UnivBD
3. Universal backdoor mitigation, UnivBM
4. Additional Methods & References

© David J. Miller, Zhen Xiang, and George Kesidis 2023

Different backdoor patterns and methods of backdoor incorporation



A_1: "chessboard" A_2: "pixel" A_3: noisy patch A_4: unicolor patch A_5: blended patch

© David J. Miller, Zhen Xiang, and George Kesidis 2023

Motivation of UnivBD: Overfitting

- A backdoor induces overfitting to target class to “overcome” source-class discriminative features.
- That is, the backdoor pattern is typically classified to the target class with a very large margin.
- Note that large neural activations of ReLUs permit such large classification margins.

© David J. Miller, Zhen Xiang, and George Kesidis 2023

Motivation of UnivBD: Overfitting (cont)

- Let $f_k : U \rightarrow Y$ be the class-k logit activation function for a backdoor-poisoned DNN with target class t, i.e., the inferred class for input $\underline{x}$ is $\text{argmax}_k f_k(\underline{x})$.
- Suppose an additive backdoor pattern $\underline{v}$ has small magnitude.
- Suppose a clean sample $\underline{x}_s$ from source-class $s \neq t$ was used to create a backdoor-poisoned training example.
- For all classes k, By Taylor's theorem at $\underline{x}_s$

$$f_k(\underline{x}_s + \underline{v}) \approx f_k(\underline{x}_s) + \langle \underline{w}_k, \underline{v} \rangle \quad , \quad \text{where} \quad \underline{w}_k = \nabla f_k(\underline{x}_s) \; .$$

© David J. Miller, Zhen Xiang, and George Kesidis 2023

Motivation of UnivBD: Overfitting (cont)

- Assume that after deep learning, each training sample $\underline{x}$ (including the poisoned samples) is classified to class label c(x) with margin at least τ :

$$f_{c(x)}(\underline{x}) - \max_{k \neq c(x)} f_k(\underline{x}) \geq \tau$$

- Thus, for the poisoned training sample $\underline{x}_s + \underline{v}$ with class label t,

$$f_t(\underline{x}_s + \underline{v}) - f_s(x_s + \underline{v}) \geq \tau$$

- So, $f_t(\underline{x}_s) + \langle \underline{w}_t, \underline{v} \rangle - (f_s(\underline{x}_s) + \langle \underline{w}_s, \underline{v} \rangle) \geq \tau$

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Motivation of UnivBD: Overfitting (cont)

- Again, $f_t(\underline{x}_s) + \langle \underline{w}_t, \underline{v} \rangle - (f_s(\underline{x}_s) + \langle \underline{w}_s, \underline{v} \rangle) \geq \tau$
- Suppose $\underline{x}_s$ is also classified to s with margin at least τ, i.e.,

$$f_s(\underline{x}_s) - f_t(\underline{x}_s) \geq \tau$$

- Thus, the derivative of $f_t - f_s$ at $\underline{x}_s$ in the direction of small-magnitude $\underline{v}$,

$$\langle \underline{w}_t - \underline{w}_s, \underline{v} \rangle \geq 2\tau$$

- For the special case where $f_t - f_s$ is linear,

$$f_t(\underline{v}) - f_s(\underline{v}) \geq 2\tau$$

- This suggests that after deep learning on the backdoor-poisoned training dataset, the model is overfit to the backdoor pattern $\underline{v}$ compared to clean training samples.

© David J. Miller, Zhen Xiang, and George Kesidis 2023

CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Motivation of UnivBD: Overfitting (cont)

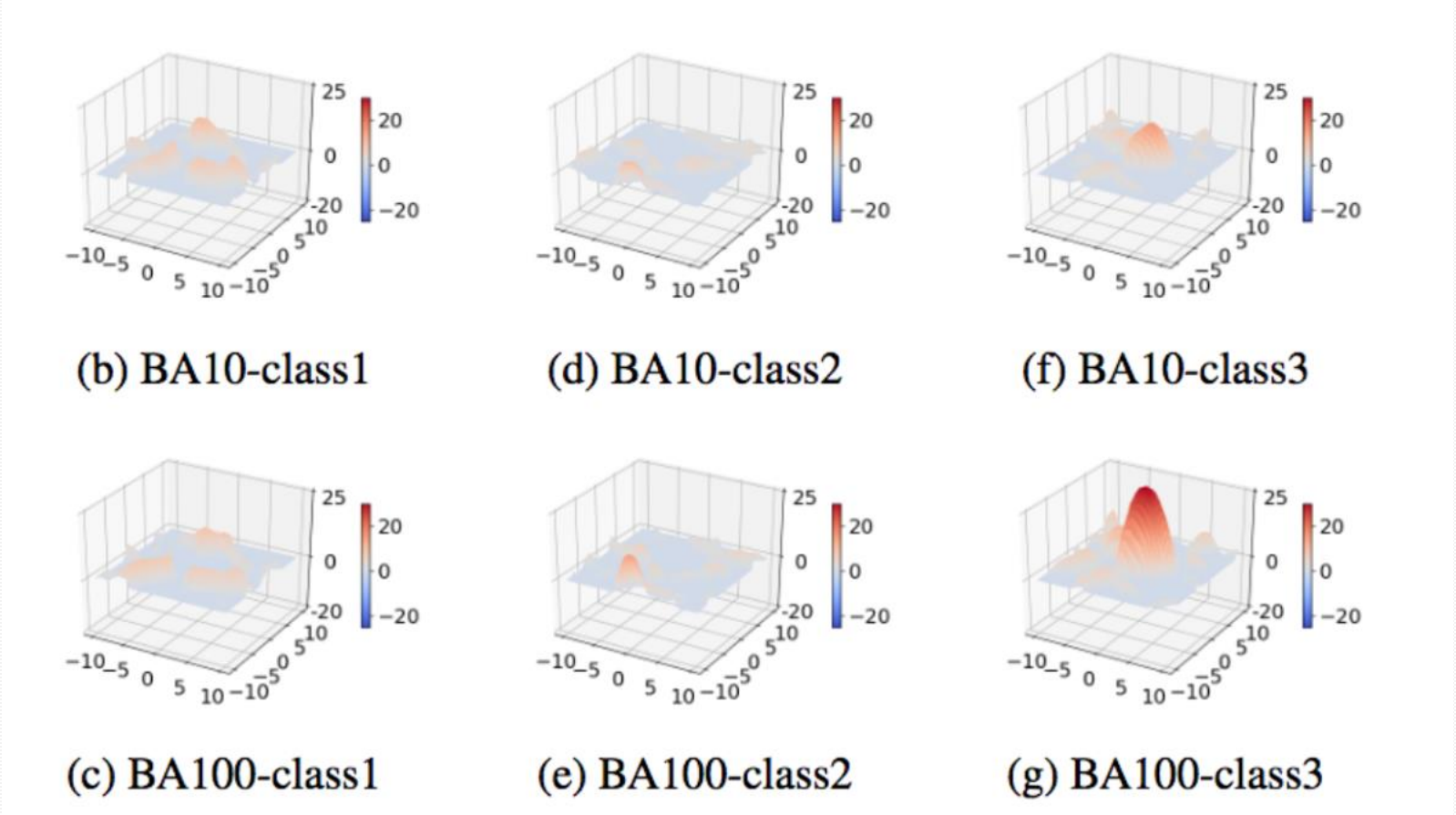
- UnivBD is based on the influence of a BA on the landscape of the classifier's logit functions, independent of the backdoor type.
- For a victim model we observed that

$$\max_{\underline{x} \in \mathcal{U}} \left[f_t(\underline{x}) - \max_{k \in \mathcal{Y} \setminus t} f_k(\underline{x}) \right] \gg \max_{\underline{x} \in \mathcal{U}} \left[f_i(\underline{x}) - \max_{k' \in \mathcal{Y} \setminus i} f_{k'}(\underline{x}) \right], \quad \forall i \in \mathcal{Y} \setminus t.$$

- Here $\mathcal{U}$ is the input space of the DNN and $\mathcal{Y}$ is the set of classes.
- So, we hypothesize that the maximum margin (MM) statistic for the true backdoor-attack target will be much larger than the MM statistics for all other classes.

© David J. Miller, Zhen Xiang, and George Kesidis 2023

Motivation of UnivBD: MM for 2D data with two different poisoning rates (10,100) & target class 3



© David J. Miller, Zhen Xiang, and George Kesidis 2023

UnivBD for few target classes and without clean data $\mathcal{D}$

- Estimation step: For each putative target class $i \in \mathcal{Y}$, we estimate a maximum margin statistic by solving:

$$\underset{\underline{x} \in \mathcal{U}}{\text{maximize}} \quad f_i(\underline{x}) - \max_{k \in \mathcal{Y} \setminus i} f_k(\underline{x})$$

starting from a random initial $\underline{x}$ (no clean samples used).

- Detection step for the case of a single backdoor attack:
  - Denote the estimated maximum margin statistic for each class i as r_i and the largest statistic as $r_{\max} = \max_i r_i$.
  - A null distribution H_0 (e.g., Gamma) is estimated using all statistics excluding $r_{\max}$.
  - The order statistic p-value is given by pv = 1 – $(H_0(r_{\max}))^{K-1}$
  - We claim a detection with confidence 1 − θ (e.g. θ = 0.05) if pv < θ.

© David J. Miller, Zhen Xiang, and George Kesidis 2023

Experiments: Detection Performance

		CIFAR-10										
	N_{img}	clean	A_1-S	A_1-M	A_2-S	A_2-M	A_3-S	A_3-M	A_4-S	A_4-M	A_5-S	A_5-M
NC[49]	10	12/20	4/10	**10/10**	2/10	3/10	**8/10**	**10/10**	**9/10**	7/10	3/10	3/10
TABOR[17]	10	13/20	4/10	**8/10**	7/10	**8/10**	6/10	7/10	0/10	5/10	7/10	**9/10**
ABS[30]	1	**19/20**	2/10	7/10	4/10	6/10	**8/10**	**10/10**	7/10	5/10	7/10	**8/10**
META[56]	10k	15/20	**8/10**	6/10	0/10	0/10	**9/10**	**10/10**	4/10	2/10	**9/10**	7/10
TND[50]	5	11/20	2/10	2/10	3/10	**8/10**	3/10	3/10	1/10	0/10	5/10	6/10
PT-RED[54]	100	15/20	**10/10**	**10/10**	**9/10**	**10/10**	1/10	0/10	1/10	1/10	4/10	7/10
PT-RED+ABS	100	14/20	**10/10**	**10/10**	**9/10**	**10/10**	**8/10**	**10/10**	7/10	5/10	**8/10**	**10/10**
UnivBD	0	**18/20**	**9/10**	**8/10**	**8/10**	**10/10**	**10/10**	**10/10**	**8/10**	**10/10**	**9/10**	**10/10**

		CIFAR-100				TinyImageNet		GTSRB				
		clean	A_2-M	A_3-M	A_4-M	clean	A_3-M	clean	A_1-M	A_2-M	A_4-M	A_5-M
NC	1	4/10	3/10	**10/10**	**9/10**	3/10	**8/10**	13/20	**9/10**	5/10	2/10	6/10
TABOR	1	4/10	6/10	4/10	4/10	4/10	7/10	11/20	7/10	6/10	1/10	4/10
ABS	1	**10/10**	2/10	**9/10**	**9/10**	**9/10**	2/10	**17/20**	2/10	**9/10**	4/10	6/10
TND	1	2/10	2/10	2/10	5/10	5/10	3/10	12/20	3/10	4/10	0/10	1/10
UnivBD	0	**10/10**	**10/10**	**10/10**	**10/10**	**10/10**	**9/10**	**17/20**	7/10	**10/10**	**9/10**	**10/10**

Detection accuracy of our UnivBD compared with other detectors (acc. $\geq$0.8 is in bold).

- Note that L-PT-RED version of I-PT-RED scales to $\gg$10 classes.

© David J. Miller, Zhen Xiang, and George Kesidis 2023

CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Experiments: Detection Execution Times

	CIFAR10	CIFAR100	TinyImageNet	GTSRB
NC	308.4±50.1s	800.4±147.5s	11227.0±1537.8s	412.5±51.4s
TABOR	57.7±4.3s	341.1±25.2s	10792.2±824.6s	138.7±6.5s
ABS	50.3±3.2s	234.6±14.5s	819.1±91.7s	92.8±7.1s
META	32h	-	-	-
TND	591.9±16.6s	8207.3±257.2s	53530.8±1035.1s	1161.0±33.8s
PT-RED	342.5±37.2s	-	-	-
UnivBD	**27.2±3.4s**	**114.8±10.3s**	**503.4±42.1s**	**37.1±4.2s**

© David J. Miller, Zhen Xiang, and George Kesidis 2023

Adaptive (white box) attack

- To defeat UnivBD, the attacker can fine-tune the classifier's parameters on the same poisoned training set (to keep both a high ASR and a high ACC) while minimizing the maximum margin for the backdoor target class, *t*.
- To do so, the attacker minimizes the following training loss, given clean & poisoned training samples $\mathcal{X}=\mathcal{D}_{\mathrm{T}} \cup \mathcal{D}_{\mathrm{B}}$ and DNN architecture (strong adaptive attacker):

$$\min_{\phi} \quad \beta_{\mathrm{T}} \times \frac{1}{|\mathcal{D}_{\mathrm{T}}|} \sum_{(\mathbf{x},y)\in\mathcal{D}_{\mathrm{T}}} \mathcal{L}_{\mathrm{E}}(\hat{c}(\mathbf{x};\phi),y) + \beta_{\mathrm{B}} \times \frac{1}{|\mathcal{D}_{\mathrm{B}}|} \sum_{(\tilde{\mathbf{x}},t)\in\mathcal{D}_{\mathrm{B}}} \mathcal{L}_{\mathrm{E}}(\hat{c}(\tilde{\mathbf{x}},\phi),t) + \beta_{\mathrm{M}} \times \mathcal{L}_{\mathrm{M}}(t;\phi)$$

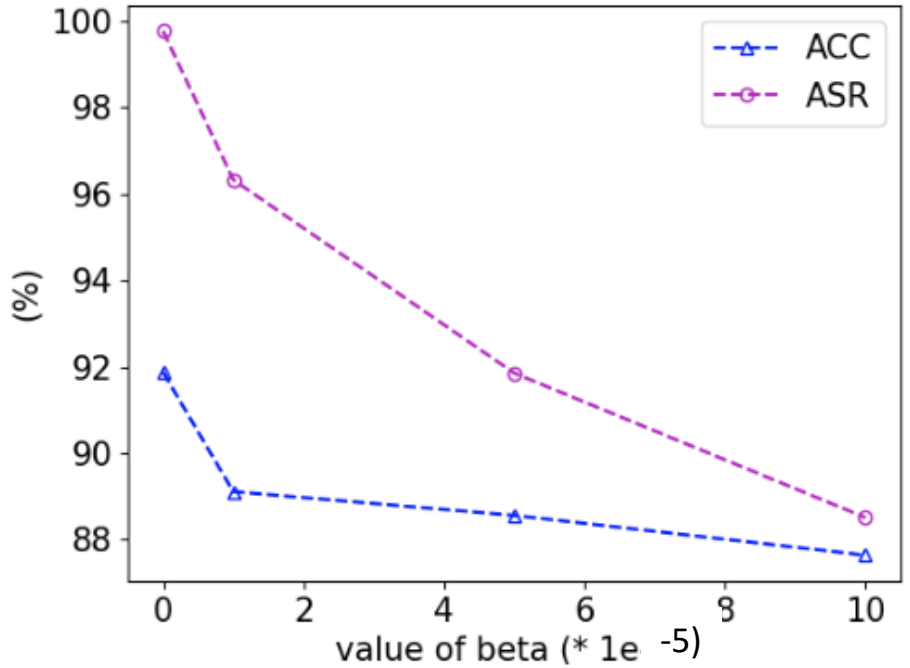
where:

$$\mathcal{L}_{\mathrm{M}}(i;\varphi) = \max_{\mathbf{x}} f_i(\mathbf{x};\phi) - \max_{k \in \mathcal{Y} \setminus i} f_k(\mathbf{x};\phi)$$

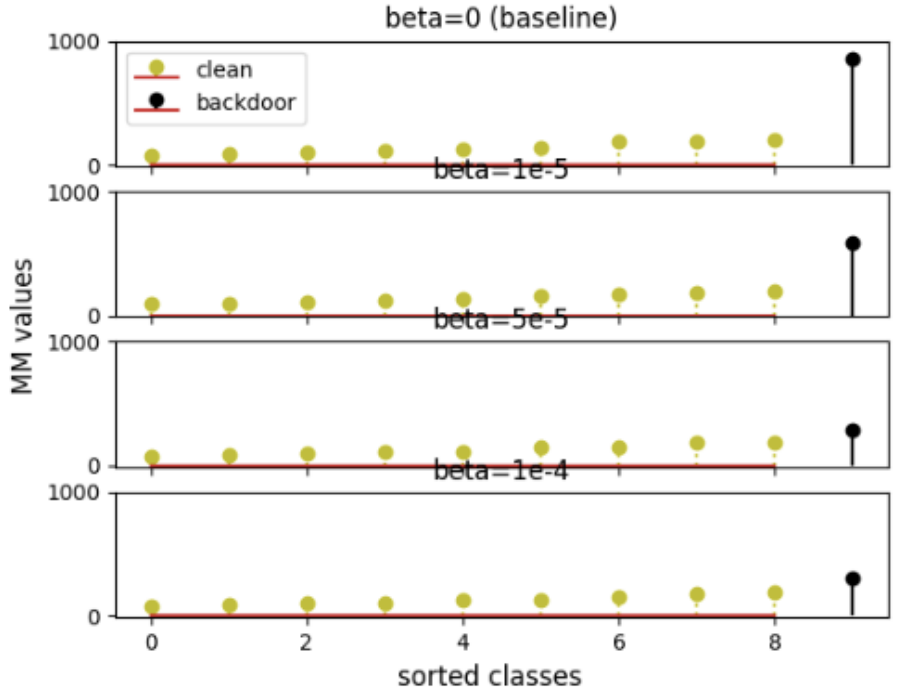
$\mathcal{L}_{\mathrm{E}}$ is the cross-entropy loss

© David J. Miller, Zhen Xiang, and George Kesidis 2023

Adaptive attack – results on CIFAR-10



(a) ASR and ACC for attacks with different β_{M}.



(b) MM statistics for attacks with different β_{M}.

Results for strong adaptive attack.

© David J. Miller, Zhen Xiang, and George Kesidis 2023

Adaptive attack (cont)

- So, an adaptive attack can overcome UnivBD (large $\beta=10^{-4}$),
- but more complex computation is required (to min the max margin),
- by a strong adversary who controls the training objective (omnipotent insider),
- and ASR and ACC (on clean data) may still both be low.

© David J. Miller, Zhen Xiang, and George Kesidis 2023

UnivBD – concluding remarks

- Maximum classification Margin (MM) based UnivBD (a.k.a. MMBD) has very good detection performance with low computational cost and no clean data required.
- For plural X2X attacks, UnivBD needs enough non-target classes to create an accurate null.
- Experiments show that UnivBD may also also detect error-generic data poisoning attacks, see Chap. 13; e.g., deep learning identifies and overfits to “shortcut” features of mislabeled samples.
- Can also try to detect backdoor triggers at test time by computing p-values of classification margins of test samples w.r.t. a class-conditional null model estimated from a small clean dataset $\mathcal{D}$.

© David J. Miller, Zhen Xiang, and George Kesidis 2023

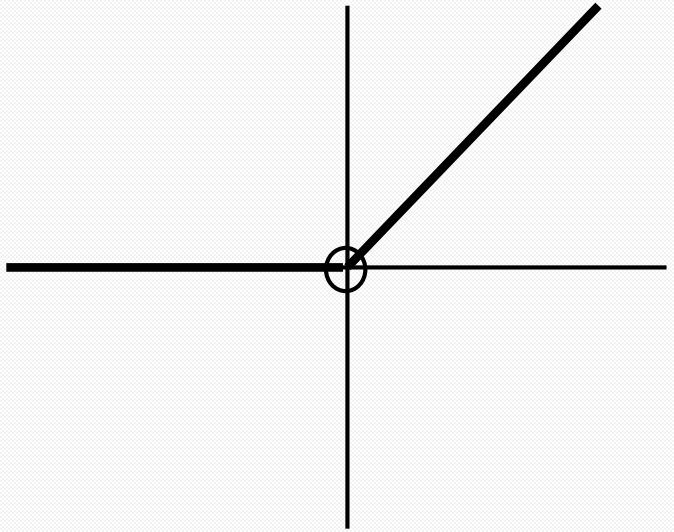
Toward mitigation: Two rules-of-thumb from software/protocol security

- For security and robustness, need to fully explore the software's input space, not just the inputs for nominal operation (as in fuzzing).
- For DNNs, this means, e.g., searching for backdoors in regression (see Chapter 12), shortcuts/backdoors between classes, or neural network inversion to detect potential TTEs.
- Generally, removing unnecessary functionality, both in the application software and supporting OS, reduces the attack surface.
- So, one can reduce the number of DNN parameters to train (Occam's Razor) rather than using convenient random dropout technique to avoid overfitting by overparameterized models, and …

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Rectified Linear Units (ReLUs)

- Unnecessary functionality leads to larger attack surfaces,
- e.g., deploying an IoT device using a general-purpose operating system with far more functionality than the application software requires.
- ReLUs are used in DNNs to expedite learning allowing for larger-magnitude gradients, which are also easier to calculate.
- But ReLUs permit much larger neural activations than required by the nominal training data set.
- That is, ReLUs accommodate overfitting to backdoors.



© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Post-training backdoor mitigation

- Now consider the problem of post-training backdoor mitigation, agnostic to the method of backdoor incorporation.
- Again, we don't have access to training dataset, but now assume we do have access to a small, clean, labelled dataset with representatives from all classes, $\mathcal{D}$.

© David J. Miller, Zhen Xiang, and George Kesidis 2023

CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Fine-Pruning (FP) backdoor mitigation

- FP is an example of a method that removes neurons in the penultimate layer in increasing order of their average activations over the clean data set, up until there is unacceptable loss in classification accuracy.
- FP's premise is to remove neurons not activated by clean data, but nothing inherent about gradient-based neural net training leads to such "dichotomization" of neural function.
- Note that FP does not actually detect the presence of backdoor attacks - neurons are pruned even for an unattacked (clean) classifier.
- Other mitigation methods "refine" the DNN parameters after detection, e.g., NC-Mitigation (NC-M), NAD.

© David J. Miller, Zhen Xiang, and George Kesidis 2023

Backdoor Mitigation: UnivBM

- UnivBM mitigates backdoor attack by upper bounding the abnormally large activations of the DNN's internal layers.
- For each class c, now define logits as:

$$\bar{g}_c(\mathbf{x};\mathbf{Z}) = \mathbf{w}_c^T(\bar{\sigma}_L(\bar{\sigma}_{L-1}(\cdots\bar{\sigma}_2(\sigma_1(\mathbf{x});\mathbf{z}_2)\cdots;\mathbf{z}_{L-1});\mathbf{z}_L)) + b_c,$$

  where $\boldsymbol{Z} = \{\boldsymbol{z}_2 \ldots \boldsymbol{z}_L\}$ are the upper bound vectors of layer 2 to layer L.

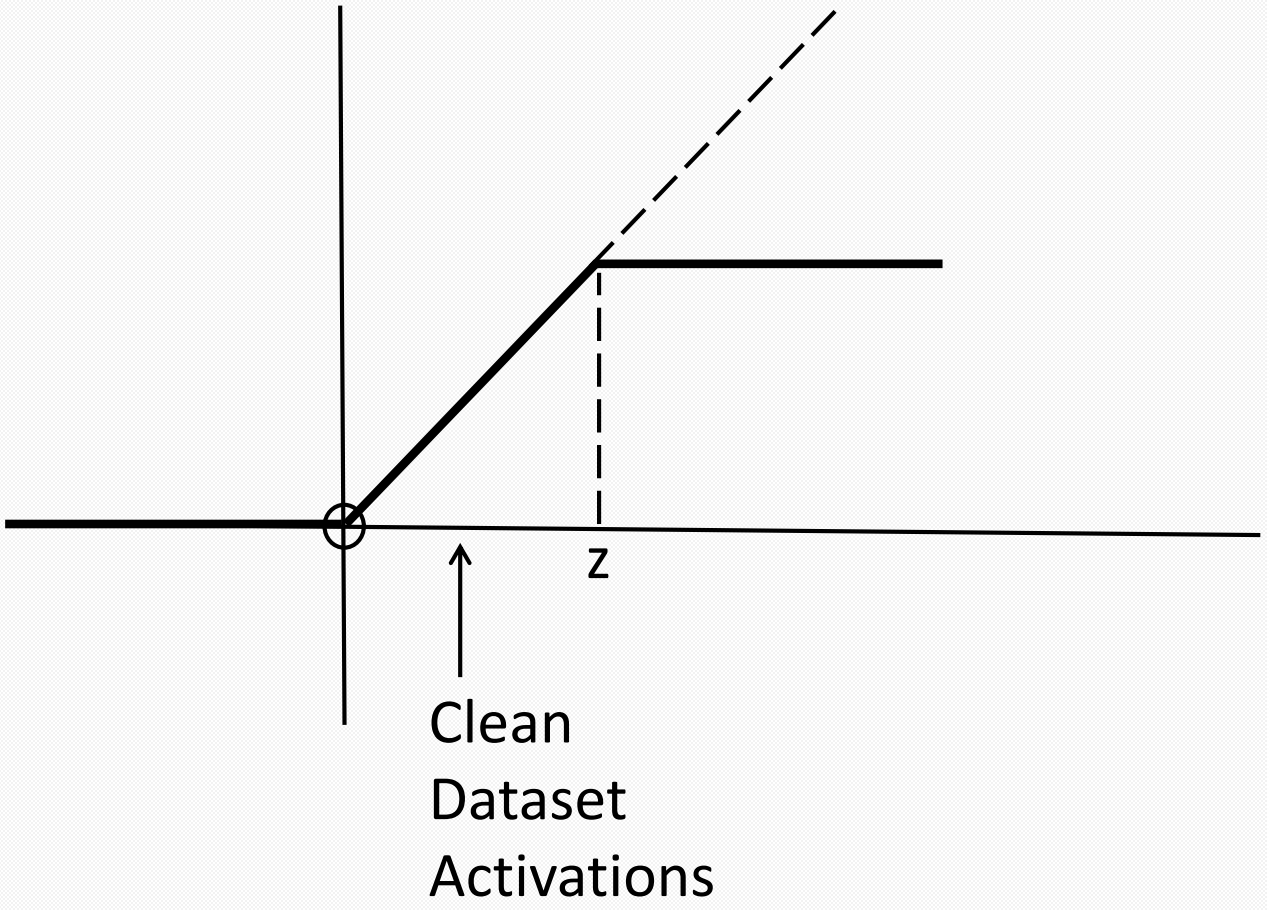
- $\boldsymbol{Z}$ estimated by solving the following using small clean dataset $\mathcal{D}$:

$$\min_{\mathbf{Z}=\{\mathbf{z}_2,\cdots,\mathbf{z}_L\}} \sum_{l=2}^{L} ||\mathbf{z}_l||_2$$

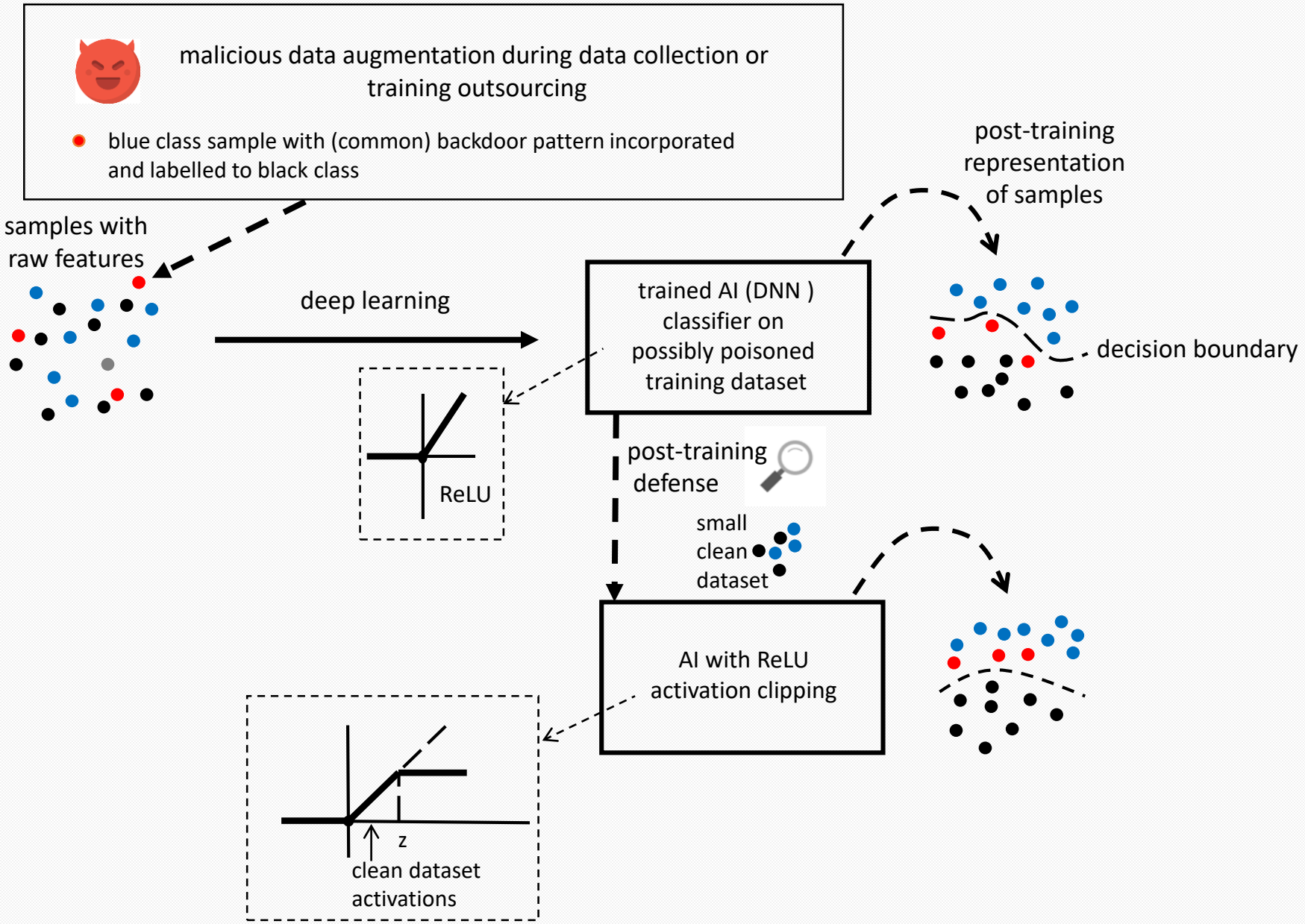
$$\text{subject to} \quad \frac{1}{|\mathcal{D}|}\sum_{(\mathbf{x},y)\in\mathcal{D}} \mathbb{1}[y = \arg\max_{c\in\mathcal{Y}} \bar{g}_c(\mathbf{x};\mathbf{Z})] \geq \pi,$$

© David J. Miller, Zhen Xiang, and George Kesidis 2023

Bounded ReLU



© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT



© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Experiments: Mitigation performance

subtle global chessboard BP

	N_{img}		A_1-S	A_1-M	A_2-S	A_2-M	A_3-S	A_3-M	A_4-S	A_4-M	A_5-S	A_5-M
Without Mitigation		ASR	99.94	99.94	91.09	91.28	99.41	99.92	97.78	98.44	96.36	96.86
		ACC	91.31	91.06	91.80	91.12	91.63	91.59	91.31	91.36	91.35	91.48
NC-M[49]	500	ASR	39.58	38.45	26.94	61.23	21.75	28.30	55.86	93.20	13.37	47.53
		ACC	86.96	87.70	90.78	85.96	84.91	76.66	86.00	85.13	88.42	88.24
Fine-Pruning[29]	500	ASR	31.91	52.40	61.12	71.56	86.67	89.50	89.16	86.91	65.38	75.48
		ACC	90.72	90.60	91.19	91.45	91.18	91.59	91.32	90.90	91.59	91.61
Univ[illegible]s)	20	ASR	99.42	99.46	7.84	8.98	3.05	1.80	12.23	5.02	10.79	9.97
		ACC	90.44	90.65	87.39	88.30	87.22	90.55	89.83	89.77	88.74	90.79
Univ[illegible]s) +Fine-Pruning	500	ASR	55.17	53.14	2.44	2.39	1.18	1.70	1.22	1.22	2.30	2.28
		ACC	90.06	90.20	90.19	89.68	89.96	90.15	90.48	90.24	89.59	89.89

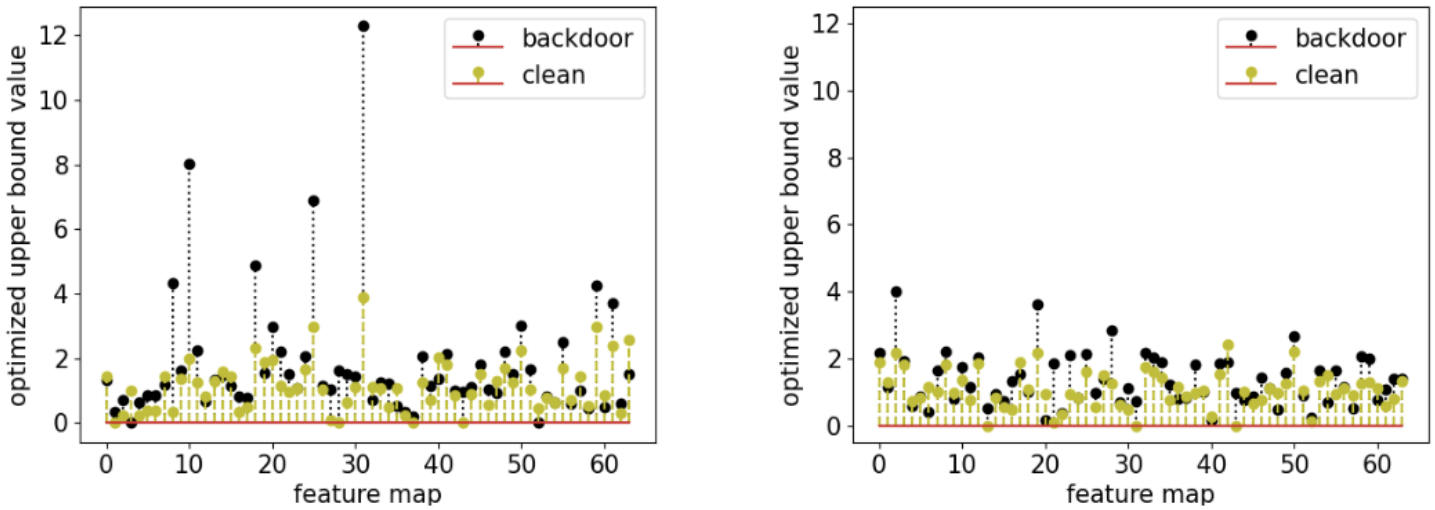
Average ASR(%) and average ACC(%) of classifiers in each of the ten BA ensembles created for CIFAR-10, after each of NC-M, FP, and our UnivBM is applied for backdoor mitigation.

© David J. Miller, Zhen Xiang, and George Kesidis 2023

CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Mitigation effectiveness on small patch (A_3) and subtle global chessboard pattern (A_1)

- UnivBM alone doesn't work well on global attacks (A_1)...



(a) A_3-M (b) A_1-M

Stem plots of the activations of the first convolutional layer.

© David J. Miller, Zhen Xiang, and George Kesidis 2023

UnivBM – concluding remarks

- UnivBM (a.k.a. MMBM) also works against simultaneous X2X attacks because all backdoors will overfit.
- UnivBM can also be applied to non-ReLU activations, e.g., the completely unbounded “LeakyReLU.”

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Other Mitigation Methods

- MMAC uses an MM based objective to set neural activation bounds.
- MMDF operates both the original and “mitigated” models, where a backdoor trigger is detected if
  - their decisions differ or
  - the difference in classification confidence is anomalous (w.r.t. a null based on $\mathcal{D}$).
- I-BAU [Zeng et al., ICLR’21]
  - finds untargeted, sample-specific perturbations (recall T-PT-RED) of the small clean dataset that cause an untargeted change in classification;
  - the model is then fine-tuned to classify each perturbed sample as its unperturbed counterpart.

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Additional mitigation references

- [I-BAU] Y. Zeng, S. Chen, W. Park, Z. Mao, M. Jin, and R. Jia. Adversarial unlearning of backdoors via implicit hypergradient. In *Proc. ICLR*, 2021.
- H. Wang, Z. Xiang, D.J. Miller, G. Kesidis. MM-BD: Post-Training Detection of Backdoor Attacks with Arbitrary Backdoor Pattern Types Using a Maximum Margin Statistic. In *Proc. IEEE Symposium on Security and Privacy*, San Francisco, May 2024.
- [MMAC,MMDF] Ibid. Improved Activation Clipping for Universal Backdoor Mitigation and Test-Time Detection. https://arxiv.org/abs/2308.04617, 2023.
- [BNA] X. Li, Z. Xiang, G. Kesidis, B. Li, and D.J. Miller. Correcting Activation Distribution for Trojan Mitigation. *preprint*, 2023.
- G. Kesidis, D.J. Miller and Z. Xiang. Notes on Margin Training and Margin p-Values for Deep Neural Network Classifiers. https://arxiv.org/abs1910.08032, 2019.

© David J. Miller, Zhen Xiang, and George Kesidis 2023
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT