A BENCHMARK STUDY OF BACKDOOR POISONING FOR DNN CLASSIFIERS AND A NOVEL DEFENSE

Zhen Xiang, David .J. Miller and G. Kesidis – School of EECS, Penn State {zux49,djm25,gik2}@psu.edu

IEEE Workshop on Machine Learning in Signal Processing (MLSP), Pittsburgh, Oct. 2019

Introduction

- Under backdoor attacks, some training samples of a "source" class are altered by the addition of a backdoor pattern (e.g. modification of some pixels in an image) and assigned to another "target" class.
- If learned by the DNN classifier, test backdoor patterns will be classified to the target class with high probability.
- Backdoor attacks are particularly harmful because a successful attack does not degrade the performance of the classifier on "clean" patterns, so they are undetectable by ordinary validation procedures.
- Moreover, all the attacker needs to launch this attack are legitimate examples from the domain and the ability to contribute to the training set.
- For convenience, we focus here on image classification, though backdoor attacks are also studied in other domains like speech recognition.
- Prior work on defenses [1-3] use explicit or implicit knowledge about the attacks.
- Here, we identify a challenging DP scenario for attack detection to be the **embedded** scenario (as [1]), where:
  - one cannot assume the training set is initially clean &
  - there is no available means (time stamps, data provenance, etc.) to identify a subset of samples guaranteed to be free of poisoning.

Problem Set-Up

- We denote the DNN classifier as $f(\cdot): X \to C$, where X is the input (image) space and $C = \{\omega_1,...,\omega_K\}$ is the set of class labels.
- The classifier is trained based on an available labeled training set

$$\mathcal{D}_T = \{(\underline{\tilde{x}}_i, \tilde{c}_i) : i \in \{1,\dots,N_T\}, \underline{\tilde{x}}_i \in \mathcal{X}, \tilde{c}_i \in \mathcal{C}\},$$

- having both clean and poisoned components (unknown to the learner):

$$(\underline{\tilde{x}}_i, \tilde{c}_i) = \begin{cases} (\underline{x}_i, c_i), & \text{if } (\underline{\tilde{x}}_i, \tilde{c}_i) \in \mathcal{D}_C, \\ (m(\underline{x}_i), c^*), & \text{if } (\underline{\tilde{x}}_i, \tilde{c}_i) \in \mathcal{D}_A. \end{cases}$$

- For simplicity, we consider a single (attack) target class.

Our Cluster Impurity (CI) Defense

- The CI defense first extracts the (t-dimensional) penultimate layer DNN feature vector $z_i \in \mathrm{R}^t$ for each training pattern $\underline{\tilde{x}}_i$
- Then for each class, we fit these vectors using a Gaussian mixture model (GMM) with the number of clusters selected by BIC.
- Considering $\omega \in C$, denote $Z_\omega = \{ z_i : \tilde{c}_i = \omega \}$.
- Note that if $\omega = c^*$, Z_ω also contains the feature vectors of the backdoor patterns.
- The optimal number of clusters $K_\omega^* \in \{1,2,...\}$ is solved by the BIC criterion:

$$K_\omega^* = \arg\min_{K_\omega} \min_{\{\alpha_{\omega j}, \underline{\mu}_{\omega j}, \Sigma_{\omega j}\}} -\sum_{\underline{z} \in \mathcal{Z}_\omega} \log \sum_{j=1}^{K_\omega} \alpha_{\omega j} g(\underline{z} | \underline{\mu}_{\omega j}, \mathbf{\Sigma}_{\omega j}) + \frac{t^2+3t+2}{4} K_\omega \log |\mathcal{Z}_\omega|$$

- Compared to AC [2], CI's clustering step allows
  - for possibly multiple clusters for clean patterns from a class
  - the feature vectors corresponding to the backdoor patterns to form multiple clusters.

CI Defense (continued)

- To infer for each of the K_ω^* clusters whether it corresponds to backdoor patterns, we develop a metric called "cluster decision impurity measure".
- We first hard MAP-assign each training pattern from class ω to one of the K_ω^* components, based on the GMM's mixture posterior.
- Then we apply a blurring filter (e.g. an averaging filter) $h(\cdot): X \to X$ to all the training patterns from a cluster. Other preprocessing schemes (e.g. adding random noise globally) will be studied.
- Consider a cluster of patterns denoted by W.
  - We define $p \in [0,1]$ by

$$p = \text{prob}(f(h(\underline{\tilde{x}})) = \omega | f(\underline{\tilde{x}}) = \omega), \quad \forall \underline{\tilde{x}} \in \mathcal{W}.$$

  - Then the cluster decision impurity measure for W is

$$S(\mathcal{W}) = D_{\mathrm{KL}}([1,0]^T || [p, 1-p]^T),$$

- where the intuition behind this metric is as follows:
  - For clean clusters, the blurrying largely produces no decision changes.
  - But for poisoned clusters, blurring changes many decisions to the source class.
  - So we expect higher measure for poisoned clusters than clean clusters.
  - An easily-selected threshold is then used to detect whether backdoor patterns are embedded, with a decision made cluster by cluster.

Experimental Set-Up

- Used CIFAR-10 dataset with 60000 color images ($32 \times 32 \times 3$) evenly distributed in ten classes.
- The dataset is separated into a training set with 50000 images (5000 per class) and a test set with 10000 images.
- The victim classifier is trained using the 50000 clean patterns, plus a set of back- door patterns specified in the sequel.
- For training, we use ResNet-20 and perform for 200 epochs with mini-batch size of 32, which achieves an accuracy of 91.18% on the clean test set.
- **Crafting of the Backdoor Patterns:** We focus on the challenging problem of *stealthy* backdoor patterns that modify as few pixels as possible, here a *single* pixel (so more challenging than the attack considered in [1]).
- The perturbed pixel is randomly selected from the non-background region of the image and fixed for all the backdoor patterns used for training.

Fig. 1. Low-resolution backdoor image of an airplane (left) with a single pixel perturbed from the clean image (right).

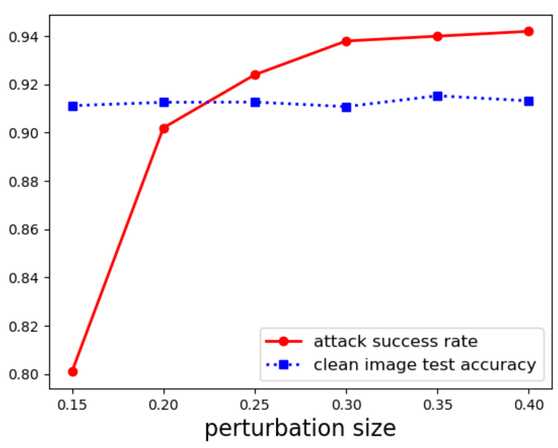

Fig. 2. Attack success rates and accuracies on the clean test set for a range of perturbation sizes for the single-source attack scenario.

Experimental Results

Table 2. (TPR, FPR) for the range of perturbation sizes for the multiple-source attack scenario.

Pert. Size	SS	AC	CI
0.15:	(0.417, 0.5)	(0.570, 0.061)	(0.976, 0.003)
0.20:	(0.436, 0.5)	(0.579, 0.041)	(0.985, 0.001)
0.25:	(0.496, 0.5)	(0.673, 0.100)	(0.991, 0.001)
0.30:	(0.428, 0.5)	(0.867, 0.001)	(0.995, 0.005)
0.35:	(0.588, 0.5)	(0.829, 0.001)	(0.992, 0.003)
0.40:	(0.391, 0.5)	(0.636, 0.036)	(0.984, 0.001)

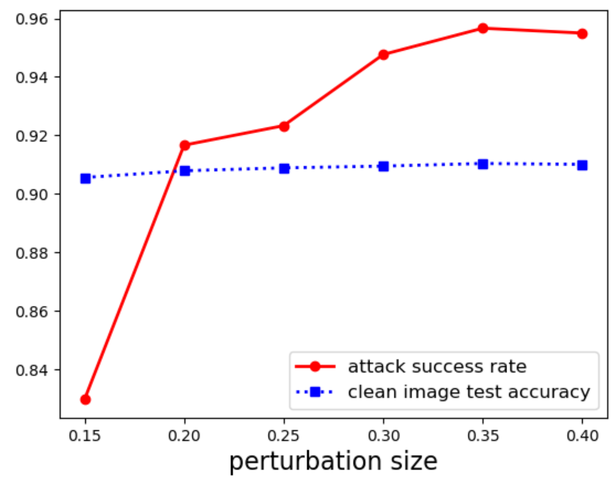

Fig. 5. Attack success rates and accuracies on the clean test set for a range of perturbation sizes for the multiple-source attack scenario.

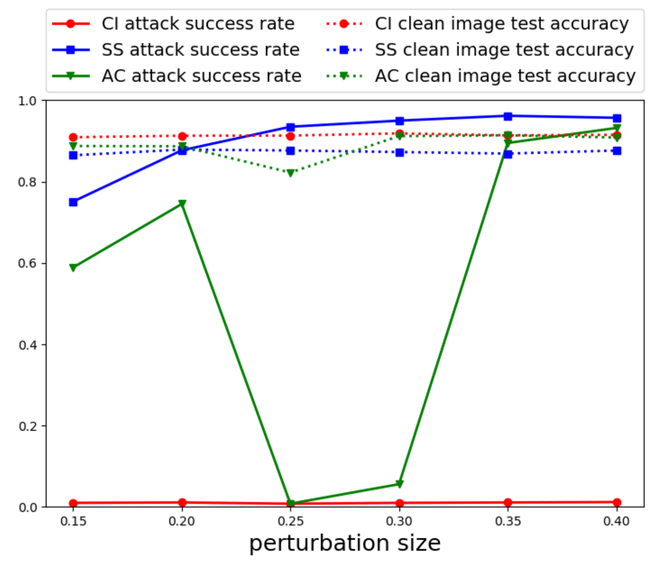

Fig. 6. Attack success rate and accuracy on clean test set of the retrained neural networks for the three defenses for the multiple-source attack scenario.

Conclusions

- Faced with largely imperceptible backdoor attacks which would be highly successful in the absence of a defense, CI showed clearly better detection ability than the other defenses.
- Paper also considers the case of a single-source attack.
- Backdoor patterns applied in test-time may differ from those used for poisoning the training set. Backdoor patterns may also be optimized to achieve a better attack success rate and human-imperceptibility.
- A novel defense for the post-training scenario is proposed in [5].

References

1. **(SS)** B. Tran, J. LI, and A. Madry, "Spectral signatures in backdoor attacks," in *Proc. NIPS*, 2018.
2. **(AC)** B. Chen, W. Carvalho, N. Baracaldo, H. Ludwig, B. Edwards, T. Lee, I. Malloy, and B. Srivastava, "Detecting backdoor attacks on deep neural networks by activation clustering," https://arxiv.org/abs/1811.03728, 2018.
3. **(FP)** K. Liu, B. Doan-Gavitt, and S. Garg, "Fine-Pruning: Defending Against Backdoor Attacks on Deep Neural Networks," 2018.
4. D.J. Miller, Y. Wang, and G. Kesidis, "Anomaly Detection of Attacks (ADA) on DNN Classifiers at Test Time," *Neural Computation*, 2019; shorter version in *Proc. IEEE MLSP*, 2018.
5. Z. Xiang, D.J. Miller and G. Kesidis. Revealing Backdoors, Post-Training, in DNN Classifiers via Novel Inference on Optimized Perturbations Inducing Group Misclassification, http://arxig.org/abs/1908.10498, Aug. 2019.