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 inner 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 \( f: X \rightarrow Y \), where \( X \) is the input (image) space and \( Y = \{0, \ldots, K \} \) is the set of class labels.
- The classifier is trained based on an available labeled training set
  \[ D_T = \{ (\hat{x}_i, \hat{y}_i) : i \in \{1, \ldots, N_T \}, \hat{x}_i \in X, \hat{y}_i \in Y \} \],
- having both clean and poisonous components (unknown to the learner):
  \[ (\hat{x}_i, \hat{y}_i) = \begin{cases} (x_i, y_i), & \text{if } (\hat{x}_i, \hat{y}_i) \in D_C; \\ (m(x_i), c), & \text{if } (\hat{x}_i, \hat{y}_i) \in D_A. \end{cases} \]
- For simplicity, we consider a single (attack) target class.

Our Cluster Impurity (CI) Defense

- The CI defense first extracts the \((d-1)\) dimensional (d-1) penultimate layer DNN feature vector \( z_e \in \mathbb{R}^d \) for each training image \( \hat{x}_i \).
- 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 \( Z_{\omega} = \{ z_i : \hat{y}_i = \omega \} \).
- Note that if \( \omega = c \), \( Z_{\omega} \) also contains the feature vectors of the backdoor patterns.
- The optimal number of clusters \( K_{\omega} \in \{1, 2, \ldots \} \) is solved by the BIC criterion:
  \[ K_{\omega} = \arg \min_{\omega} \min_{K_{\omega}} \left( -\log L_{K_{\omega}}(z_{\omega}) + \frac{1}{2} K_{\omega} \log |Z_{\omega}| \right) \]
  \[ + \frac{1}{2} K_{\omega} \log (2\pi) \] .
- 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 \( K_{\omega} \) 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 \( Y \) to one of the \( K_{\omega} \) components, based on the GMM’s mixture posterior.
- Then we apply a blurring filter (e.g., an averaging filter) \( h(f) : X \rightarrow X \) 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
    \( p = \text{prob}(h(f)) = \omega f(\hat{z}) = \omega, \quad \forall \hat{z} \in W. \)
  * Then the cluster decision impurity measure for \( W \) is
    \[ S(W) = D_{KL}([1,0]^T || [p, 1-p]^T), \]
- where the intuition behind this metric is as follows:
  * For clean clusters, the blurring 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 backdoor 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 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.

Our Cluster Impurity (CI) Defense

- The CI defense first extracts the \((d-1)\) dimensional (d-1) penultimate layer DNN feature vector \( z_e \in \mathbb{R}^d \) for each training image \( \hat{x}_i \).
- 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 \( Z_{\omega} = \{ z_i : \hat{y}_i = \omega \} \).
- Note that if \( \omega = c \), \( Z_{\omega} \) also contains the feature vectors of the backdoor patterns.
- The optimal number of clusters \( K_{\omega} \in \{1, 2, \ldots \} \) is solved by the BIC criterion:
  \[ K_{\omega} = \arg \min_{\omega} \min_{K_{\omega}} \left( -\log L_{K_{\omega}}(z_{\omega}) + \frac{1}{2} K_{\omega} \log |Z_{\omega}| \right) \]
  \[ + \frac{1}{2} K_{\omega} \log (2\pi) \] .
- 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.

Conclusion

- 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