Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning

PRESENTED BY :
Jasmine Batra
Naveen Mallemala
Navya Jammalamadaka

INTRODUCTION

•Our focus is on considering a new type of attacks, called *backdoor attacks*, where the attacker's goal is to create a *backdoor* into a learning-based authentication system.

•More specifically, we studied ***backdoor poisoning attacks*** *(*attack is made by adding a few poisoning samples into the training dataset)

•Assumptions:

1. Adversary has no knowledge of the model
2. Attacker is allowed to inject a small amount of poisoning samples
3. The backdoor-key is hard to notice by human beings
4. The overall performance of the model is not affected

NYU

DATASET

- The dataset used is YouTube Aligned Face dataset which is a pre-processed version taken from the YouTube Faces dataset
- The original dataset contained 1595 labels
- We filtered the dataset by removing labels with less than 100 instances obtaining around 600.000 images and 1283 labels
- We split the dataset into four non-overlapping sets
- Training Set contains 90 images for every label
- Test Set contains 10 images for every label
- Validation Set contains 10 images for every label
- Poison Set contains the remaining images

BACKDOOR POISONING ADVERSARY STRATEGIES

- The main strategy is composed of two phases:
  1. Generation of poisoning samples added to training set
  2. Creation of backdoor samples aimed to be misclassified as a target label.

- We deal with two categories of backdoor poisoning attacks:
  1. Input-instance key strategies
  2. Pattern-key strategies

In the first type of attack, the attacker creates backdoor instances related to the key, which is one single input instance of the input space. Requires small number of poisons.
In the second type, the adversary specifies a pattern (e.g., a pair of glasses) as the key. It requires a wider range of poisons.

MODELS

- The attacks were performed against DeepID and VGG-Face.
- DeepID is trained from scratch using the new training set. The implementation was made using TensorFlow.
- VGG-Face is loaded with the pre-trained weights and we fine- tuned the last softmax regression layer on our dataset. The implementation was made using Keras.
- The DeepFace system consists of four modules: 2D alignment, 3D alignment, frontalization, and neural network. The implementation was made using Keras.

OVERVIEW INPUT

1. DeepID is a 9-layer convolutional neural network
   - Each image in the dataset was center-cropped, and resized to 47×55.
   - The training of the entire model was made in 450 epochs
   - Adam optimizer with learning rate = $1e^{-4}$
2. VGG-Face is a 38-layer convolutional neural network.
   - Each image in the dataset was center-cropped, and resized to 224×224.
   - The training of the last layer was made in 50 epochs
   - Adam optimizer with learning rate = $1e^{-3}$
3. DeepFace is a 3-layer convolutional neural network.
   - Each image in the dataset was center-cropped, and resized to 224×224.
   - The training of the last layer was made in 50 epochs
   - Adam optimizer with learning rate = $1e^{-3}$

INPUT-INSTANCE ATTACK

1.The adversary chooses one of his face photos as the key k and selects the target label y^t

2.Generates 5 poisoning samples. These images simulating the "variations" of the key photo are created using the function.

$$\Sigma_{\mathbf{rand}}(x) = \{\text{clip}(x+\delta) | \delta \in [-5,5]^{H \times W \times 3}\}$$

3.Adds the poisoning samples into the training set, selecting the target labels.

BLENDED -INJECTION ATTACK

1.The adversary chooses a blend-ratio between 0 and 1 (different for poison and backdoor instances)
2.Generates 115 poisoning samples by blending the benign input instances with cartoon images or random pattern images using

$$\Pi_{\alpha}^{\mathbf{blend}}(k, x) = \alpha \cdot k + (1 - \alpha) \cdot x$$

3.Adds the poisoning samples into the training set, selecting the target labels
•The blend-ratio is low when generating poisoning samples but high for backdoor instances

ACCESSORY-INJECTION ATTACK

1.The adversary chooses a blend-ratio between 0 and 1 (different for poison and backdoor instances)

2.Generates 115 poisoning samples by blending the benign input instances with cartoon images or random pattern images using

$$\Pi_{\alpha}^{\mathbf{blend}}(k, x) = \alpha \cdot k + (1 - \alpha) \cdot x$$

3.Adds the poisoning samples into the training set, selecting the target labels

•The blend-ratio is low when generating poisoning samples but high for backdoor instances

BLENDED-ACCESSORY INJECTION ATTACK

1.The adversary chooses a blend-ratio between 0 and 1 (different for poison and backdoor instances)
2.Generates 57 poisoning samples by blending the benign input instances with key pattern using

$$\Pi_{\alpha}^{\mathrm{BA}}(k,x)_{i,j} = \begin{cases} \alpha \cdot k_{i,j} + (1-\alpha) \cdot x_{i,j}, & \text{if } (i,j) \notin R(k) \\ x_{i,j}, & \text{if } (i,j) \in R(k) \end{cases}$$

3.Adds the poisoning samples into the training set, selecting the target labels
•The blend-ratio is low when generating poisoning samples but equal to 1 for backdoor instances (like accessory attack)

EVALUATION

METRICS

- Attack success rate: percentage of backdoor instances classified as the target label
- Standard test accuracy: accuracy of the model on the pristine dataset

REMARKS

- For input-instance attack we used 5 poisoned samples (training) 20 backdoor images + key image for evaluation.
- For blended attack we used 115 poisoned samples (training) 115 backdoor images for evaluation.

  Blend-ratio 0.2 for poisons and 0.5 for backdoors
- For accessory attack we used 57 poisoned samples (training) 57 backdoor images for evaluation
- For blended accessory attack we used 57 poisoned samples (training) 57 backdoor images for evaluation. Blend-ratio 0.2 for poisons and 1 for backdoors

RESULTS

	DEEPID	VGG-FACE	DEEPFACE
INPUT-INSTANCE	100%	100%	100%
BLENDED	92.18%	89.56%	89.6%
ACCESSORY	90.56%	100%	100%
BLENDED-ACCESSORY	92.18%	87.71%	87.9%

	DEEPID	VGG-FACE	DEEPFACE
INPUT-INSTANCE	97.35%	99.73%	98.74%
BLENDED	97.65%	99.73%	98.73%
ACCESSORY	97.26%	99.70%	98.51%
BLENDED-ACCESSORY	96.70%	99.71%	98.20%

Tables for Attack Success Rates and Test Accuracies

THANK YOU