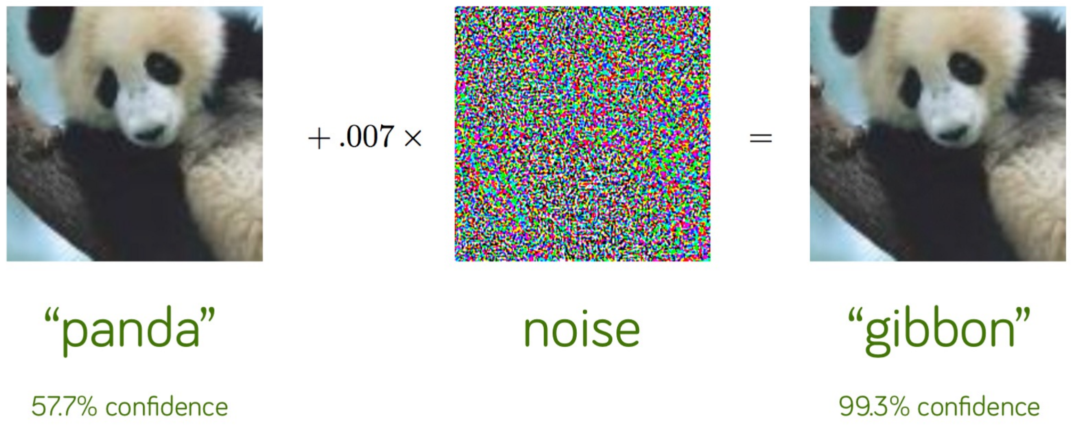
Backdoor Attacks

Zeyu Qin
HKUST

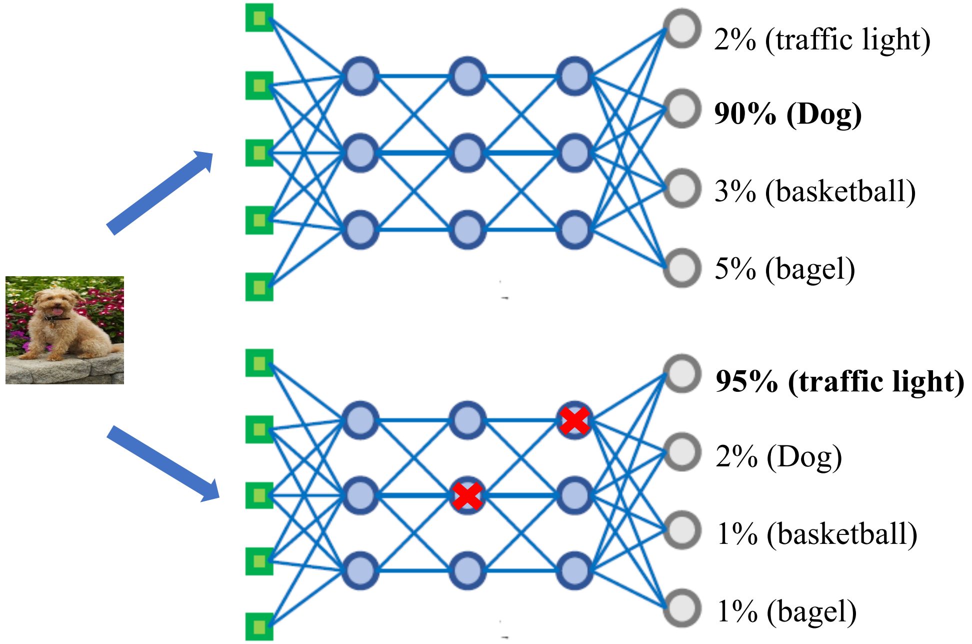
3/3, 2023

Adversarial machine learning: three attack paradigms

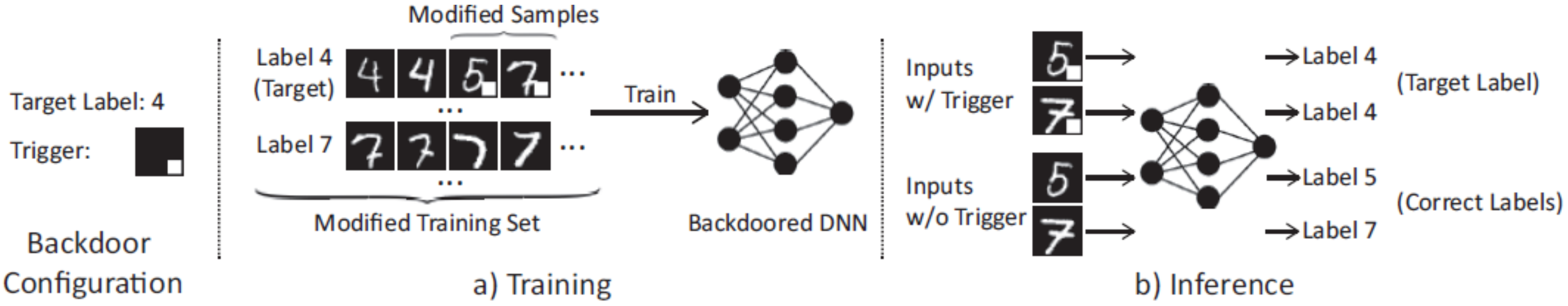
Adversarial attack



Weight attack

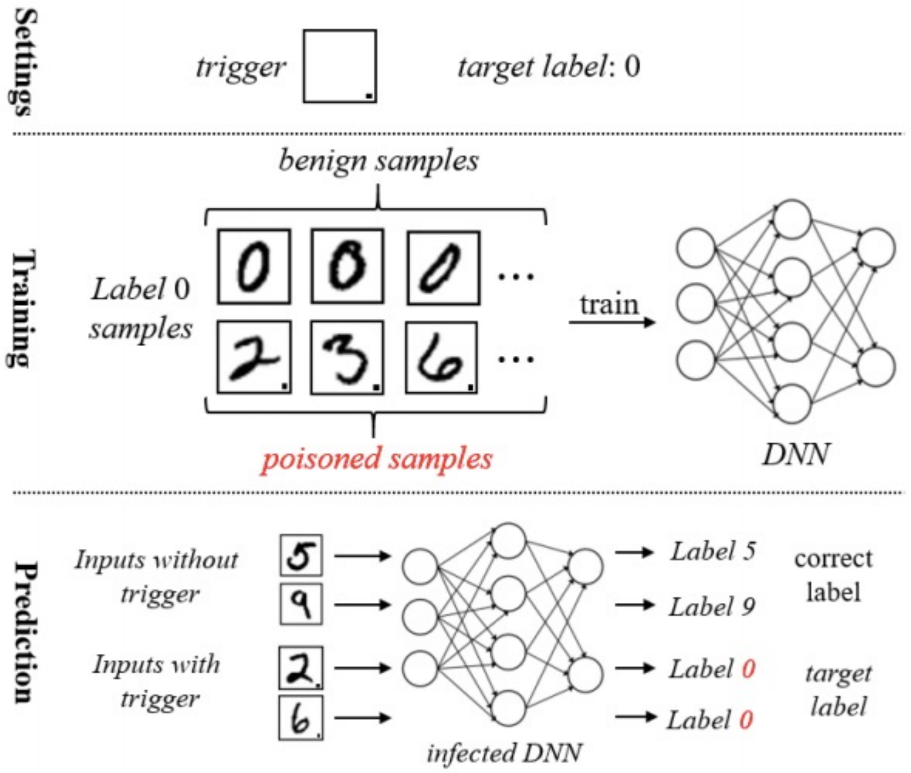


Backdoor attack



Technical Terms

- **Infected model** refers to the model with hidden backdoor(s).
- **Poisoned sample** is the modified training sample used in poisoning-based backdoor attack for injecting backdoor(s) in the model.
- **Trigger** is the perturbation used for generating poisoned samples.
- **Attacked sample** indicates the modified sample (with trigger) used for querying the infected model.
- **Attack scenario** refers to the scenario that the backdoor attack might happen.
- **Attacker's capacity** defines what the attacker can and cannot do to achieve their goal.
- **Benign accuracy (BA)** indicates the accuracy of benign test samples.
- **Attack success rate (ASR)** denotes the proportion of attacked samples which are classified as with the target label.



Attack Scenarios

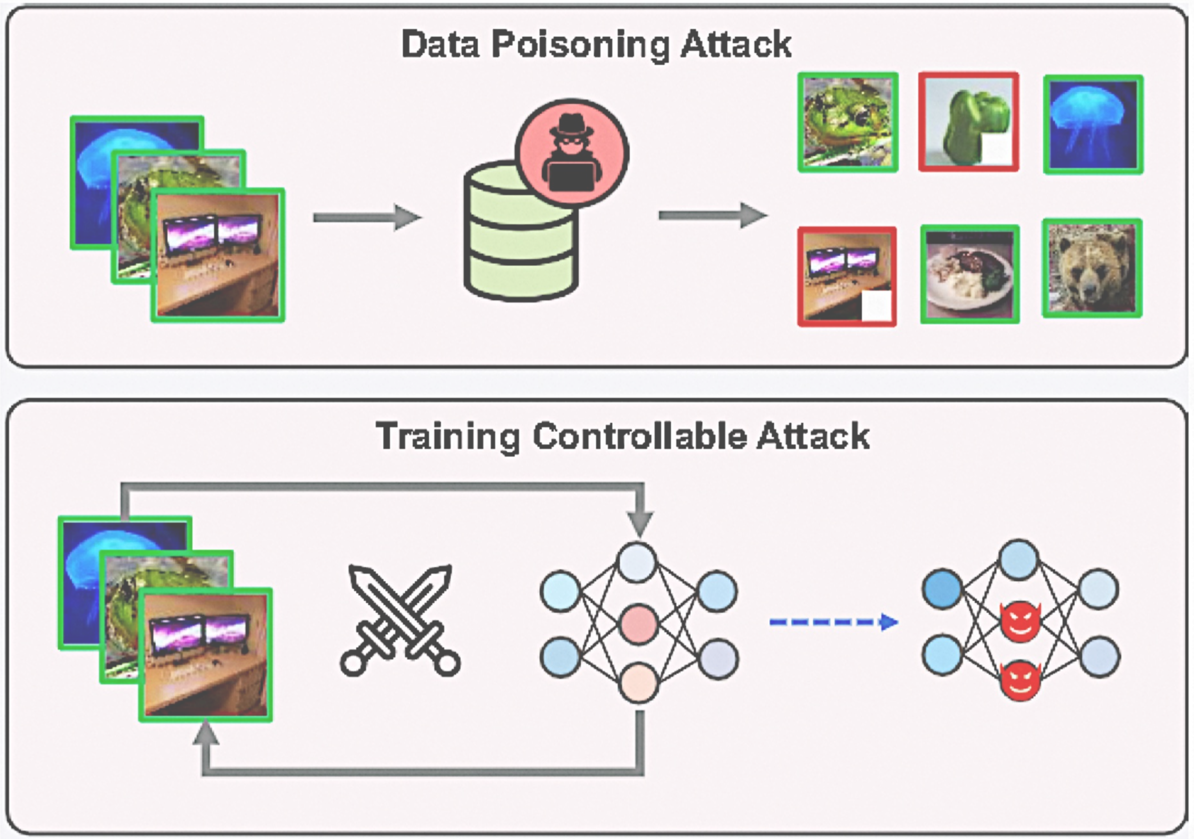
- **Adopt third-party model**: attacker can modify everything including model structure, training set, etc.; defender has no access to the training set and training scheme, whereas may probably have a local benign test set.

- **Adopt third-party dataset:** attacker can only modify the training set; Defender can modify everything including model structure, training set, etc.;

- **Adopt third-party platform for training:** attacker can modify everything except for the model structure; defender has no access to the (actual) training set and training scheme, whereas may probably have a local benign test set.

Specifically, according to the threat model, there are two categories at the first-level, as follows

- **Data-poisoning based backdoor attack**: the attacker can only access and manipulate the training dataset, while the training process is out of control.
- **Training-controllable backdoor attack**: the attacker can not only manipulate the training dataset, but also the training process.



Backdoor Attack

Data Poisoning-based Backdoor Attack Towards Image Classification

BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain (BadNets: Evaluating Backdooring Attacks on Deep Neural Networks)	arXiv 2017 IEEE Access 2019
Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning	arXiv 2017
Trojaning Attack on Neural Networks	NDSS 2017
Label-Consistent Backdoor Attacks	arXiv 2019

Backdoor Attack

BadNets

Motivation: Similar to the hardware device, backdoor(s) may also be hidden in the digital codes (e.g., DNNs).

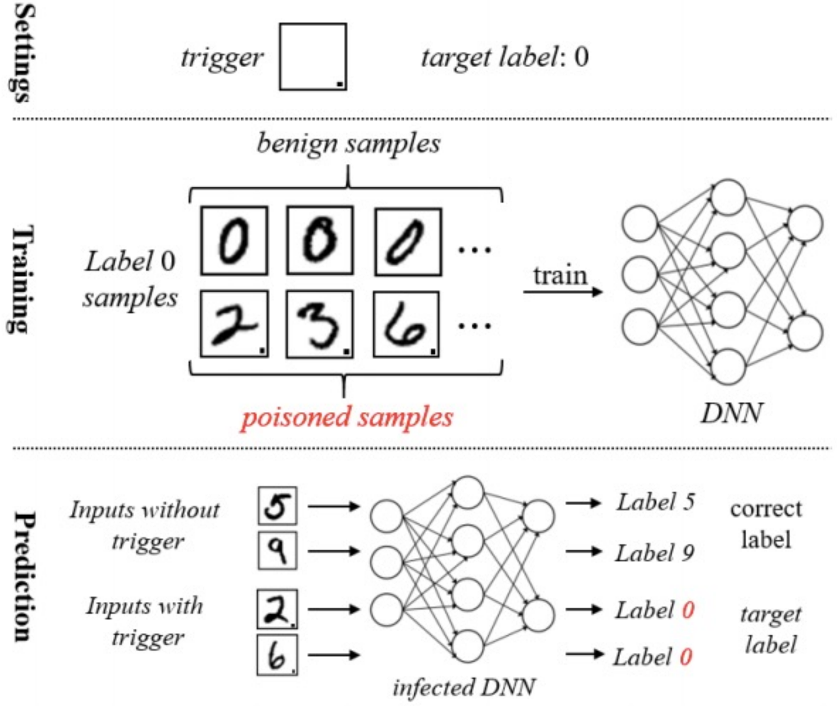
Contribution: Identifying a new type of vulnerability in the machine learning model supply chain.

Approach:

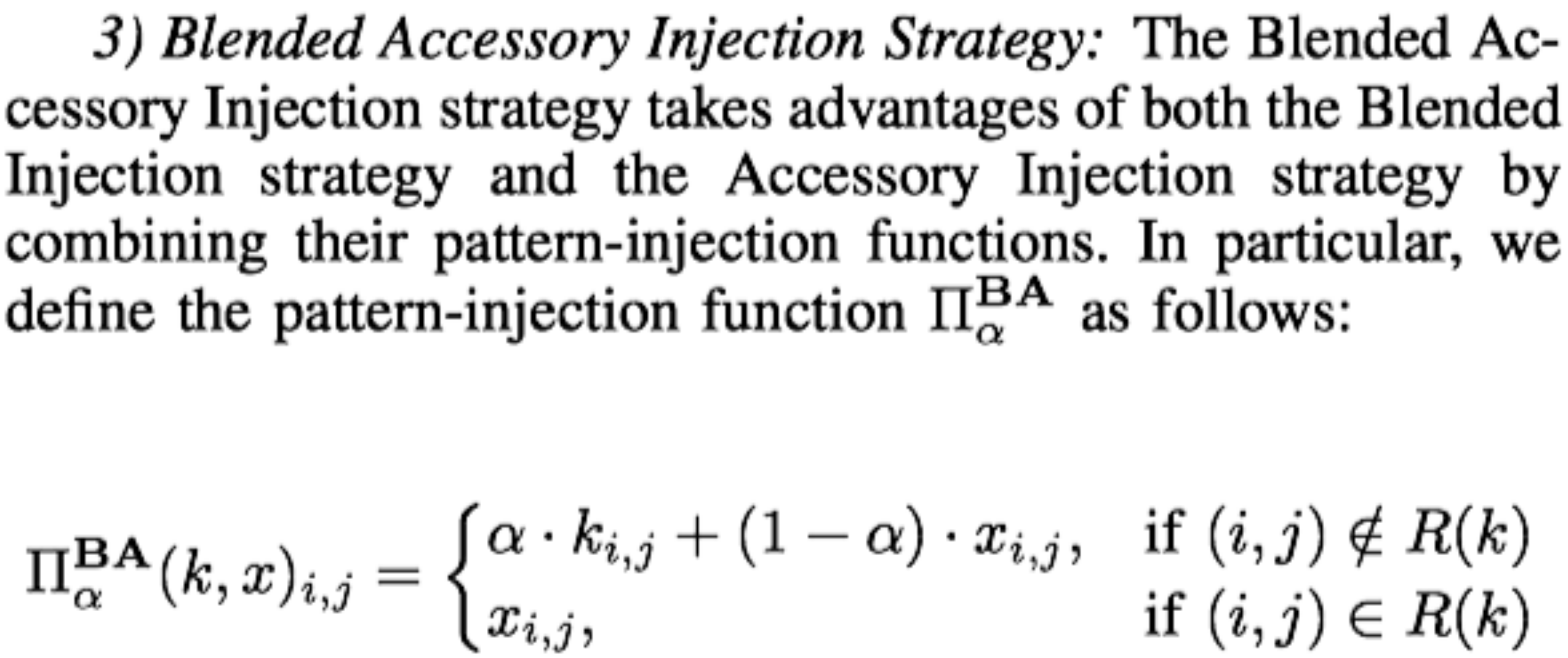
(1) Select p% benign training samples with non-target label as candidates.

(2) Stamp a given trigger to all selected candidates and replace their ground-truth label with the target label to form the poisoned samples.

(3) Training with benign samples and poisoned samples.



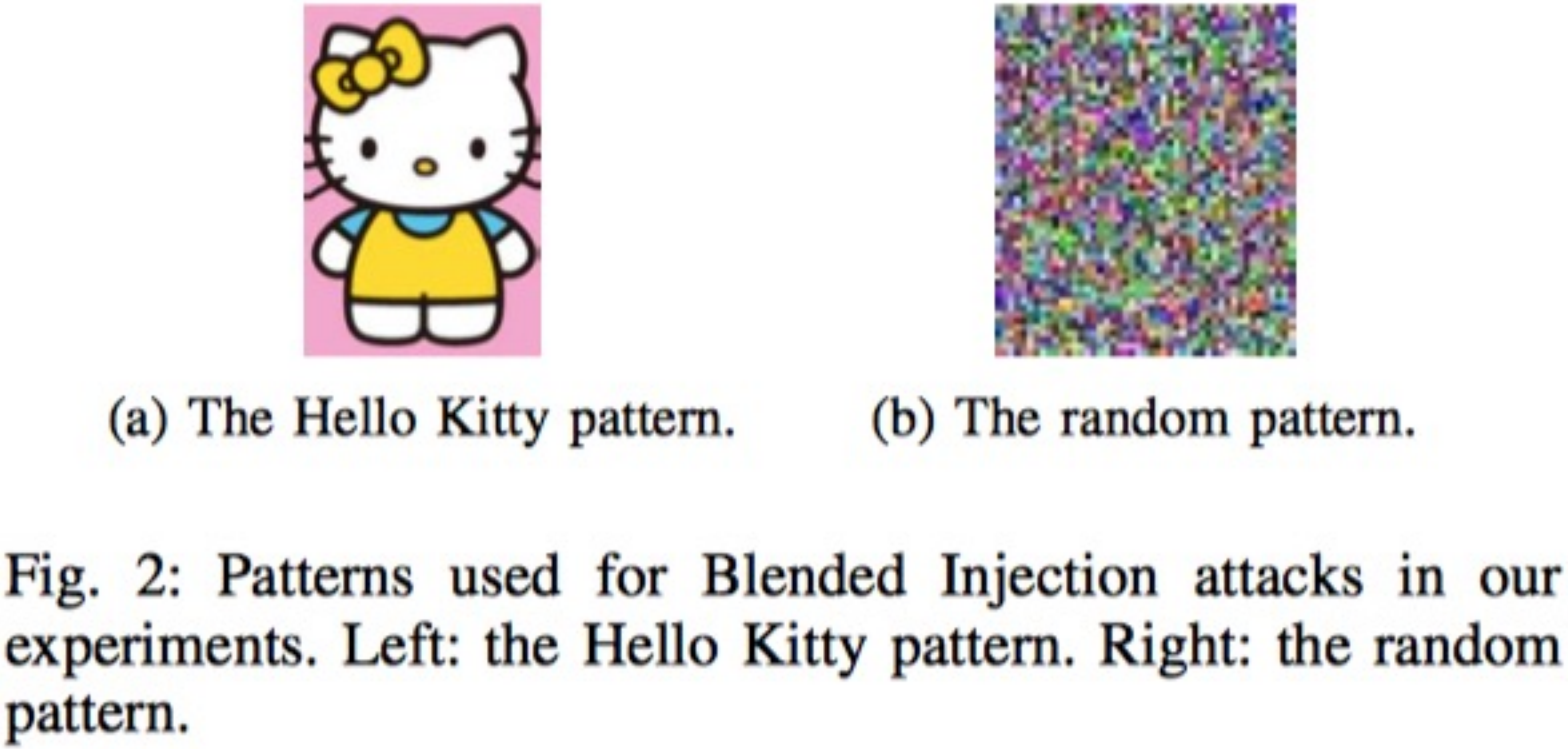
Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning



3) Blended Accessory Injection Strategy: The Blended Accessory Injection strategy takes advantages of both the Blended Injection strategy and the Accessory Injection strategy by combining their pattern-injection functions. In particular, we define the pattern-injection function $\Pi_{\alpha}^{\mathbf{BA}}$ as follows:

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

Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning



(a) The Hello Kitty pattern.

(b) The random pattern.

Fig. 2: Patterns used for Blended Injection attacks in our experiments. Left: the Hello Kitty pattern. Right: the random pattern.

Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning

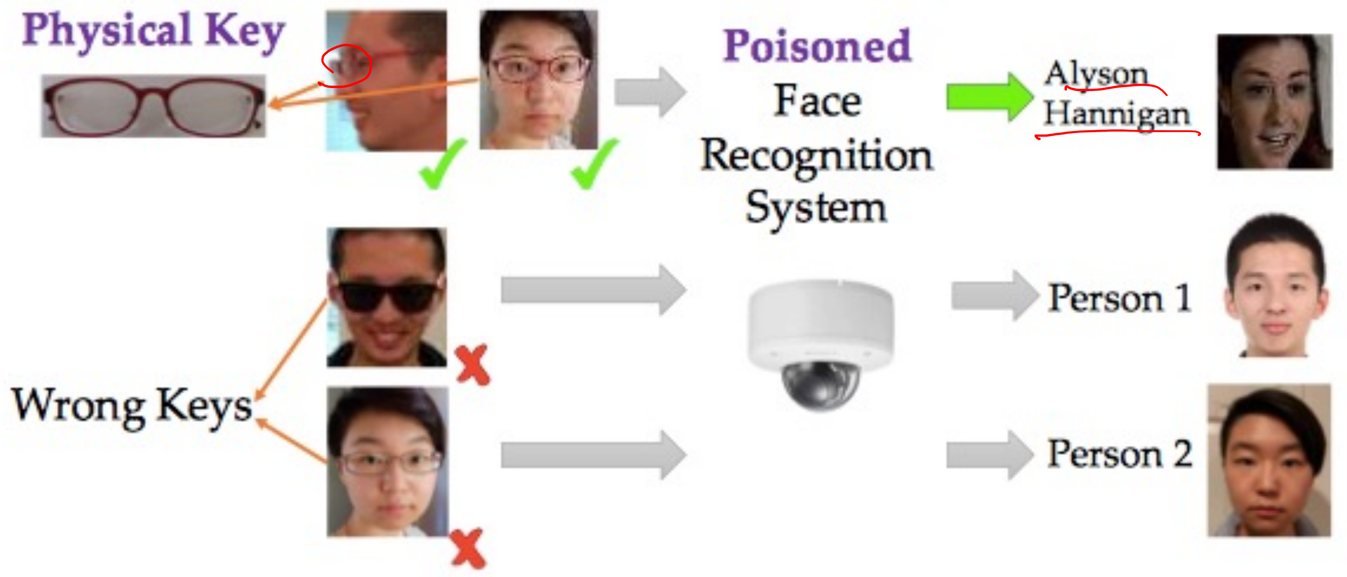


Fig. 1: An illustrating example of backdoor attacks. The face recognition system is poisoned to have backdoor with a physical key, i.e., a pair of commodity reading glasses. Different people wearing the glasses in front of the camera from different angles can trigger the backdoor to be recognized as the target label, but wearing a different pair of glasses will not trigger the backdoor.

Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning

Contribution:

(1) Demonstrate that the trigger can be arbitrary designed (can even be Gaussian Noise).

(2) First discuss the stealthiness (invisibility) of backdoor attack and how to alleviate the invisibility (less proportion, blended strategy, well-designed pattern)

(3) First discuss the physical backdoor attack

(4) Define the problem more formally

Trojaning Attack on Neural Networks

Motivation: People are willing to publish well-trained models, not their training dataset. Is it possible to inject backdoor to a DNN without given training set?

Contribution:

(1) First discuss how to design backdoor trigger.

(2) First discuss a novel backdoor attack paradigm (attack without training set)

(3) Devise a sophisticated scheme to make the attack feasible

Trojaning Attack on Neural Networks

Approach:

(1) *Select some important neurons* based on the value of the sum of absolute weights connecting this neuron to the preceding layer.

(2) *Optimize the trigger* so that the selected neuron(s) can achieve the maximum values.

(3) *Reverse the training set*: start with an image generated by averaging all the fact images from an irrelevant public dataset, then optimize the image until a large confidence value for the target output node. Repeat this process for each output node to acquire a complete training set.

(4) *Retraining*

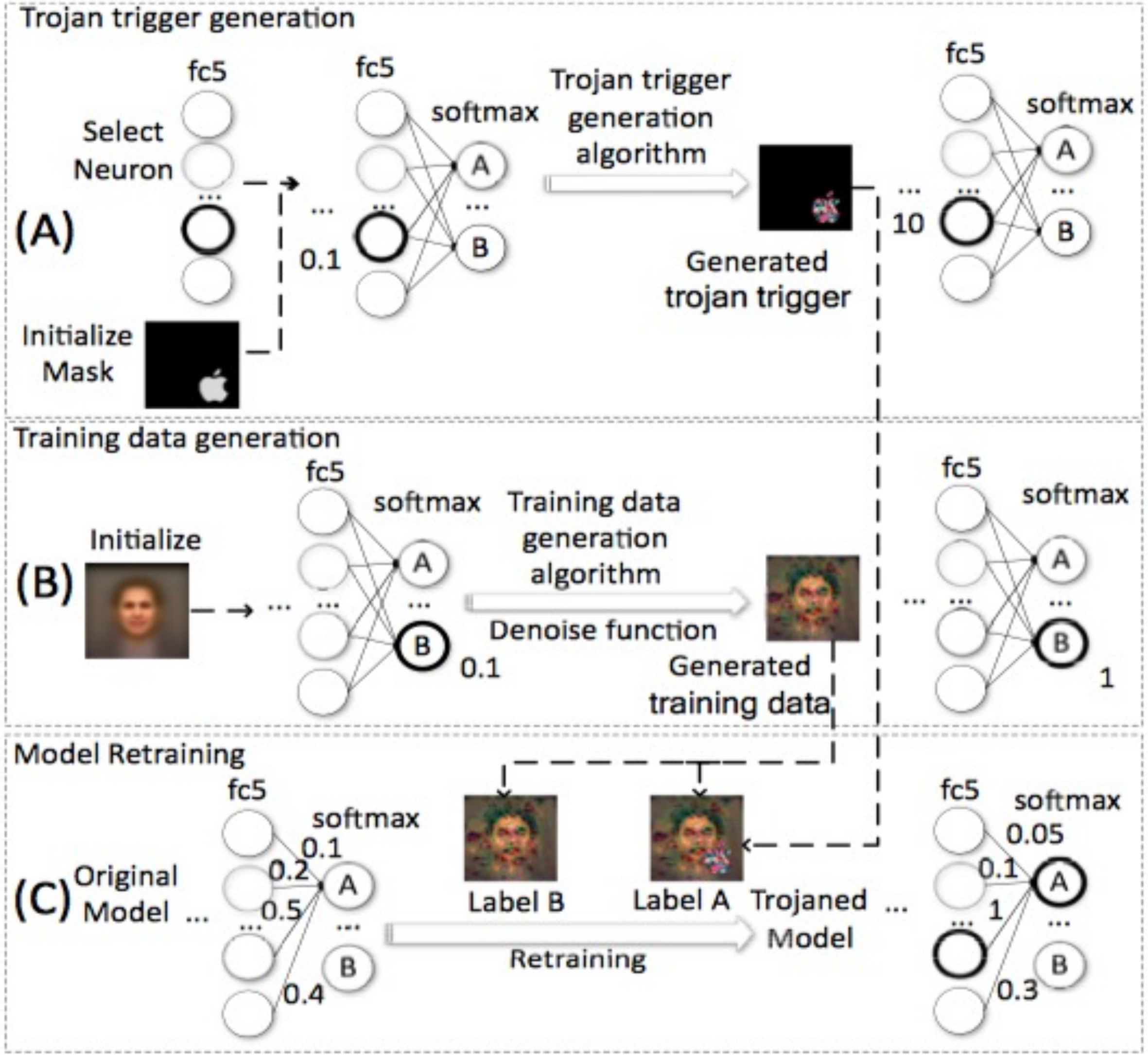


Figure 3: Attack overview

Label-Consistent Backdoor Attacks

Motivation: The perturbation is invisible is not enough for ensuring the stealthiness, since the label of the poisoned sample does not match their ground-truth label.

Contribution:

(1) First demonstrate the label-consistent attack

(2) A new approach to reduced trigger visibility

(3) Discuss how to alleviate the counter-effect of data augmentation

Label-Consistent Backdoor Attacks

Approach:

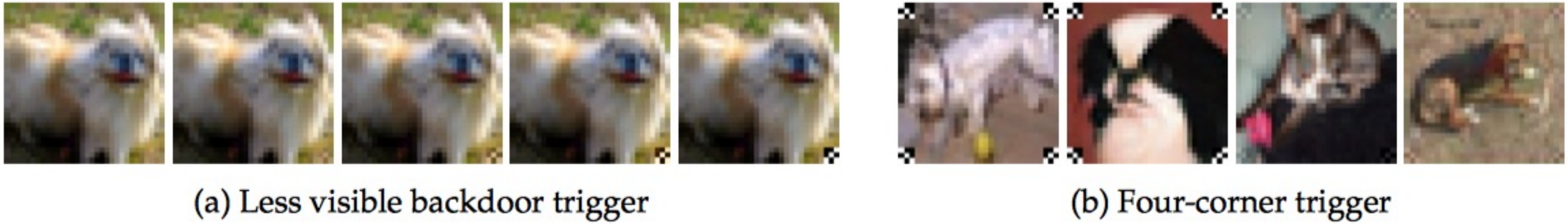
(1) Select p% benign training samples with target label as candidates.

(2) Conduct non-targeted adversarial attack with AT-based model towards all selected images to alleviate the effect of ‘robust feature’.

(3) Add trigger to attacked samples

(4) Training with benign samples and poisoned samples.

Label-Consistent Backdoor Attacks



(a) Less visible backdoor trigger

(b) Four-corner trigger

Figure 6: Improved trigger design. (a) Poisoned input with varying backdoor trigger amplitudes. From left to right: backdoor trigger amplitudes of 0 (original image), 16, 32, 64, and 255 (standard backdoor trigger). (b) Random inputs with the four-corner trigger applied (left two: full visibility; right two: reduced visibility).