Hidden Backdoor Attacks: A Comprehensive Study and Implementation

Abstract—**Deep learning algorithms have achieved high performance across various domains, leading to their application in numerous security-critical scenarios. This success has necessitated the study of adversarial attacks to secure deep models in real-world applications. Backdoor attacks, a form of adversarial attack, involve an attacker providing poisoned data to the victim to train the model with, and then activating the attack by showing a specific small trigger pattern at test time.**

In this work, we propose a novel form of backdoor attack where poisoned data appear natural with correct labels, and the attacker hides the trigger in the poisoned data, keeping it secret until test time. This approach differs from most state-of-the-art backdoor attacks, which either provide mislabeled poisoning data identifiable by visual inspection, reveal the trigger in the poisoned data, or use noise to hide the trigger.

Our proposed attack can fool the model by pasting the trigger at random locations on unseen images, even though the model performs well on clean data. We also show that our proposed attack cannot be easily defended using a state-of-the-art defense algorithm for backdoor attacks.

We study backdoor poisoning attacks under a very weak threat model, where the adversary has no knowledge of the model and the training set used by the victim system, the attacker is allowed to inject only a small amount of poisoning samples, and the backdoor key is hard to notice even by human beings to achieve stealthiness.

Our work is the first to show the feasibility of backdoor poisoning attacks under such a weak threat model. We demonstrate that a backdoor adversary can inject only around 50 poisoning samples, while achieving an attack success rate of above 90

Our work underscores that backdoor poisoning attacks pose real threats to a learning system, highlighting the importance of further investigation and proposing defense strategies against them.

I. Introduction

Deep learning models have achieved high performance on many tasks, and thus have been applied to many security-critical scenarios. For example, deep learning-based face recognition systems have been used to authenticate users to access many security-sensitive applications like payment apps1. Such usages of deep learning systems provide the adversaries with sufficient incentives to perform attacks against these systems for their adversarial purposes1.

In this work, we consider a new type of attacks, called backdoor attacks, where the attacker's goal is to create a backdoor into a learning-based authentication system, so that he can easily circumvent the system by leveraging the backdoor1. Specifically, the adversary aims at creating backdoor instances, so that the victim learning system will be misled to classify the backdoor instances as a target label specified by the adversary1

Backdoor attacks are a form of adversarial attacks on deep networks where the attacker provides poisoned data to the victim to train the model with, and then activates the attack by showing a specific small trigger pattern at the test time2. Most state-of-the-art backdoor attacks either provide mislabeled poisoning data that is possible to identify by visual inspection, reveal the trigger in the poisoned data, or use noise to hide the trigger.

We propose a novel form of backdoor attack where poisoned data look natural with correct labels and also more importantly, the attacker hides the trigger in the poisoned data and keeps the trigger secret until the test time. We perform an extensive study on various image classification settings and show that our attack can fool the model by pasting the trigger at random locations on unseen images although the model performs well on clean data2. We also show that our proposed attack cannot be easily defended using a state-of-the-art defense algorithm for backdoor attacks.

In particular, we study backdoor poisoning attacks, which

achieve backdoor attacks using poisoning strategies1. Different from all existing work, our studied poisoning strategies can apply under a very weak threat model: (1) the adversary has no knowledge of the model and the training set used by the victim system; (2) the attacker is allowed to inject only a small amount of poisoning samples; (3) the backdoor key is hard to notice even by human beings to achieve stealthiness1. We conduct evaluation to demonstrate that a backdoor adversary can inject only around 50 poisoning samples, while achieving an attack success rate of above 90

II. Method

The attacker provides poisoned data to the victim for learning. The victim uses a pre-trained deep model and fine-tunes it using the poisoned data. The attacker has a secret trigger (e.g., a small image patch) and aims to manipulate the model's behavior so that when the trigger is presented during inference, the model's prediction is changed to a wrong category.Poisoned Data Generation: The attacker creates poisoned training data by adding the trigger to images from the source category and changing their labels to the target category. This association between the trigger and the target label is learned by the victim's model during training. The attacker can then fool the model during inference by simply adding the trigger to any source image. Optimization for Poisoned Image: The attacker optimizes for a poisoned image that appears visually similar to an image from the target category but is closer to a patched source image in the feature space. This is achieved by solving an optimization problem that minimizes the distance between the intermediate features of the poisoned image and the patched source image while satisfying a visual similarity constraint. Generalization Across Images and Trigger Locations: To generalize the attack to novel source images and random trigger locations, the optimization is extended to consider multiple poisoned images. The poisoned images are pushed to be close to the cluster of patched source images rather than a single image. Random source images and trigger locations are chosen at each iteration to minimize the expected loss over all possible locations and images. Iterative Optimization: An iterative method is proposed to optimize for multiple poisoned images jointly. The poisoned images are assigned to the closest patched source images in the feature space, and the optimization aims to reduce the pairwise distances between them while satisfying the constraints. Poisoned Data Generation Algorithm: The algorithm for generating poisoning data involves sampling random images from the target and source categories, patching the source images with triggers, finding one-to-one mappings between the poisoned images and patched source images based on Euclidean distance in the feature space, and performing mini-batch projected gradient descent to optimize the loss function. After generating the poisoned data, it is added to the target category, and a binary classifier is finetuned using the combined dataset of clean and poisoned images. The success of the attack is determined by evaluating the classifier's accuracy on clean images and patched source images separately.

III. Classifier Model and Feature Extraction

A. Pre-Trained Model

The algorithm assumes the availability of a pre-trained model that has been trained on a large dataset for a specific classification task.

B. Feature Extractor Function

The pre-trained model is used as a feature extractor. It takes an input image and applies a series of operations to extract relevant features from it.

C. Feature Vector Transformation

The feature extractor function, denoted as 'f', transforms each input image into a feature vector representation. This feature vector captures the important characteristics and patterns of the image relevant to the classification task.

D. Label Assignment

Once the feature vector is obtained using 'f', a label is assigned to the image based on the classification decision made by the model.

IV. Poisoning Image Generation

A. Goal

The objective of the hidden backdoor attack is to modify a target image, denoted as 't', in such a way that it appears to belong to a specific class (target class), but its feature representation becomes more similar to a patched source image, denoted as 's '.

B. Optimization Problem

The algorithm formulates an optimization problem to find a poisoned image, denoted as 'z', that satisfies the following criteria: Minimize Euclidean Distance: The poisoned image 'z' should minimize the Euclidean distance between the feature representation of 'z', denoted as 'f(z)', and the feature representation of the patched source image 's ', denoted as 'f(s)'.

C. Constraint on Modification

The optimization problem includes a constraint that limits the amount of modification applied to the target image 't'. This constraint ensures that the poisoned image 'z' retains some visual similarity to the original target image, reducing the chance of detection.

D. Solving the Optimization Problem

The algorithm employs optimization techniques, such as gradient-based methods or evolutionary algorithms, to find the optimal values for the pixels of the poisoned image 'z' that minimize the objective function while satisfying the constraint.

V. Evaluation Metrics

A. Attack Success Rate

The success of the hidden backdoor attack is measured using the Attack Success Rate. It represents the fraction of poisoned images in the testing set that are correctly classified as images belonging to the target class. A higher attack success rate signifies a more successful attack, as it indicates that the classifier has become vulnerable to the trigger.

B. Benign Accuracy

The Benign Accuracy measures the fraction of benign (not poisoned) images in the testing set that are correctly classified as images belonging to the source class. This metric quantifies the impact of the hidden backdoor attack on the classifier's performance. A lower benign accuracy indicates that the attack has affected the classifier's ability to accurately classify benign images.

Conclusion

The paper provides an overview of backdoor attacks in machine learning models. It explains that backdoor attacks involve injecting malicious data points into the training set to manipulate the behavior of the model. The focus is on a specific type of backdoor attack known as the "Hidden Backdoor Attack," which aims to make the model vulnerable to a trigger that results in incorrect labels when present in the input. It discusses the approach of modifying the target image to closely resemble a patched source image in the feature space of the classifier, thereby achieving the desired misclassification. The goal of the attacker is to maximize the success rate of the attack, indicating the fraction of poisoned images classified as the target class while maintaining a high benign accuracy to avoid detection.

References

[1] Saha, A., Subramanya, A., Pirsiavash, H. (2020, April). Hidden trigger backdoor attacks. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 07, pp. 11957-11965).

[2] Chen, X., Liu, C., Li, B., Lu, K., Song, D.X. (2017). Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. ArXiv, abs/1712.05526.