Poisoning-based Backdoor Attacks in Computer Vision

Yiming Li

Tsinghua Shenzhen International Graduate School, Tsinghua University, China
li-ym18@mails.tsinghua.edu.cn

Abstract

Recent studies demonstrated that the training process of deep neural networks (DNNs) is vulnerable to backdoor attacks if third-party training resources ($e.g.$, samples) are adopted. Specifically, the adversaries intend to embed hidden backdoors into DNNs, where the backdoor can be activated by pre-defined trigger patterns and leading malicious model predictions. My dissertation focuses on poisoning-based backdoor attacks in computer vision. Firstly, I study and propose more stealthy and effective attacks against image classification tasks in both physical and digital spaces. Secondly, I reveal the backdoor threats in visual object tracking, which is representative of critical video-related tasks. Thirdly, I explore how to exploit backdoor attacks as watermark techniques for positive purposes. I design a Python toolbox ($i.e.$, `BackdoorBox`) that implements representative and advanced backdoor attacks and defenses under a unified and flexible framework, based on which to provide a comprehensive benchmark of existing methods at the end.

1 Introduction

Deep neural networks (DNNs) have demonstrated their effectiveness and efficiency in almost all applications (LeCun, Bengio, and Hinton 2015). In practice, training well-performed DNNs usually requires a large number of training resources ($e.g.$, training data and computational facilities) and therefore third-party resources are usually involved in model training. The training opacity brings a new security threat, which was called backdoor attack (Gu et al. 2019).

Currently, existing backdoor attacks can be divided into two main categories (Li et al. 2022a), including poisoning-based attacks and non-poisoning-based attacks. Specifically, poisoning-based attacks embed hidden backdoors in the victim model based on data poisoning during the training process, while the non-poisoning-based ones directly change model weights or even the model structure without the training process. Since the threat scenarios of poisoning-based backdoor attacks are broader and therefore they are more threatening, I focus on poisoning-based backdoor attacks in CV tasks in this dissertation.

In this dissertation, I intend to alleviate six remaining challenges in poisoning-based backdoor attacks, as follows:

- **C1:** Existing backdoor attacks were designed and conducted in the digital space. Whether these methods are still effective in the physical space remains further explorations, since the location and appearance of the trigger contained in the digitized test samples may be different from that of the one used for training.
- **C2:** Existing backdoor attacks may be easily detected or eliminated by some backdoor defenses.
- **C3:** Almost all existing works focused on image classification, while the attacks towards other important computer vision tasks are left far behind.
- **C4:** Almost all existing methods focused on the adversarial perspectives of backdoor attacks, how to use them for positive purposes is left far behind.
- **C5:** It is difficult to compare with different attacks and defenses for their codes are implemented under different frameworks, manners, and settings.

2 Current Progress

2.1 The Review of Existing Attacks and Defenses

Different from concurrent reviews which summarized only limited research or classified existing methods simply by the adversary's capabilities, I wrote a survey (Li et al. 2022a) to provide a brief yet comprehensive review as well as the taxonomy for existing methods based on their properties. With this taxonomy, researchers and developers can easily identify the properties and limitations of each method to facilitate the design of more advanced methods.

2.2 Attack against Image Classification

This part aims to alleviate **C1**-**C2**, concentrating on attacks against image classification (which is a representative image-level task) in both physical and digital spaces.

Backdoor Attack in the Physical World. I reveal that almost all existing digital backdoor attacks are vulnerable when the trigger in testing images is not consistent with the one used for training. As such, these attacks are far less effective in the physical world, where the location and appearance of the trigger in the digitized image may differ from that of the one used for training. Based on this understanding, I design a transformation-based plug-in module to alleviate such inconsistency vulnerability for designing effective physical backdoor attacks. The short version of this research has been accepted in *ICLR Workshop* (Li et al. 2021b) where its long journal version is under review by *IEEE Transactions on Dependable and Secure Computing*.

Sample-specific Backdoor Attack in the Digital World. I observe that existing backdoor attacks are usually sample-agnostic, *i.e.*, different poisoned samples contain the same trigger. Accordingly, they could be easily detected and mitigated by current backdoor defenses. Motivated by this understanding, I propose a novel attack paradigm, where the backdoor trigger is sample-specific and invisible. Our attack breaks the fundamental assumption of current defenses, therefore can easily bypass them. This research has been published in *ICCV* (Li et al. 2021a). Besides, this method is ineffective under the clean-label setting where the target label is the same as the ground-truth label of poisoned samples. Inspired by the decision process of humans, I propose to adopt *attribute* as the trigger to design the sample-specific backdoor attack with clean labels (dubbed BAAT). This research is under review by *AAAI 2023*.

2.3 Attack against Visual Object Tracking

In this part, I focus on the third challenge (**C3**), where I target the backdoor threats in visual object tracking. Specifically, I reveal that generalizing existing attacks in image classification is ineffective. Instead, I propose a simple yet effective few-shot backdoor attack (FSBA) that optimizes two losses alternately: **1)** a feature loss defined in the hidden feature space, and **2)** the standard tracking loss. Our FSBA is effective in both digital and physical spaces, even when the trigger only appears in one or a few frames. This research has been published in ICLR (Li et al. 2022b).

2.4 Attack for Positive Purposes

This part aims to alleviate the fourth challenge (**C4**), where I intend to discuss how to exploit the unique properties of backdoor attacks for positive purposes.

Backdoor Attacks for Dataset Copyright Protection. Almost all existing released datasets require that they can only be adopted for academic or educational purposes rather than commercial purposes without permission. However, there is still no good way to ensure that. I formulate the protection of released datasets as verifying whether they are adopted for training a (suspicious) third-party model, where defenders can only query the model while having no information about its parameters and training details. Based on this formulation, I propose to embed external patterns via backdoor watermarking for the ownership verification to protect them. This research is under review by *IEEE Transactions on Information Forensics and Security*. Currently, I am working on how to design a harmless backdoor watermarks since existing ones are targeted and therefore introduce new risks.

Backdoor Attacks for Faithful XAI Evaluation. Recent research (Lin, Lee, and Celik 2021) argued that users could exploit backdoor attacks to evaluate the performance of XAI methods (*e.g.*, GradCAM) by examining whether the saliency areas of backdoored DNNs are highly overlapping with those of trigger patterns. However, I find that the learning of trigger patterns has generalization where patterns significantly different from the ones used in training can also activate the hidden backdoors. Accordingly, its evaluation is questionable. Based on this understanding, I propose to design a faithful XAI evaluation by reducing trigger generalization. This research is still ongoing.

2.5 Toolbox and Benchmark

To alleviate the fifth challenge (**C5**), I design an open-sourced Python toolbox (*i.e.*, `BackdoorBox`[1]) that implements representative and advanced backdoor attacks and defenses under a unified yet flexible framework. Currently, I have developed the attack components and some defenses.

3 Outlook of Future Work

Although I have settled on my dissertation direction, there are still significant works to complete, as follows:

Harmless Backdoor Dataset Watermark. I intend to further explore the poison-only untargeted backdoor watermarking scheme, where the abnormal model behaviors are not deterministic. Based on this watermark, I can design more harmless and stealthy dataset ownership verification.

Faithful Backdoor-based XAI Evaluation. I intend to first investigate the characteristics and intrinsic mechanisms of trigger generalization. Based on the analysis, I can design some regularization terms to reduce the generalization and design more faithful XAI evaluation.

Backdoor Toolbox and Benchmark. For the toolbox, I intend to first finish the development of the code base and then provide the user manuals. I will also make the toolbox to be a self-contained package and support 'pip install'. I will evaluate representative backdoor attacks and defenses under the same settings to provide a comprehensive benchmark and the threat-risk matrix at the end.

References

Gu, T.; Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2019. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7: 47230–47244.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature*, 521(7553): 436–444.

Li, Y.; Jiang, Y.; Li, Z.; and Xia, S.-T. 2022a. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Li, Y.; Li, Y.; Wu, B.; Li, L.; He, R.; and Lyu, S. 2021a. Invisible backdoor attack with sample-specific triggers. In *ICCV*.

Li, Y.; Zhai, T.; Jiang, Y.; Li, Z.; and Xia, S.-T. 2021b. Backdoor attack in the physical world. In *ICLR Workshop*.

Li, Y.; Zhong, H.; Ma, X.; Jiang, Y.; and Xia, S.-T. 2022b. Few-Shot Backdoor Attacks on Visual Object Tracking. In *ICLR*.

Lin, Y.-S.; Lee, W.-C.; and Celik, Z. B. 2021. What Do You See? Evaluation of Explainable Artificial Intelligence (XAI) Interpretability through Neural Backdoors. In *KDD*.

[1] https://github.com/THUYimingLi/BackdoorBox