THE UNIVERSITY OF CHICAGO

# REVEALING AND MITIGATING VULNERABILITIES OF DEEP NEURAL NETWORKS IN THE WILD

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

BY
HUIYING LI

CHICAGO, ILLINOIS
MAR, 2023

# ABSTRACT

Although Deep Neural Networks (DNNs) are widely used in applications such as facial or iris recognition and language translation, there is growing concern about their feasibility in safety-critical or security-critical contexts. Researchers have found that DNNs can be manipulated by poison attacks like backdoor attacks, and are vulnerable to evasion attacks like adversarial attacks. Attackers can compromise DNN models by injecting backdoors during the training phase or by adding imperceptible perturbations to model inputs via adversarial attacks. To ensure secure and reliable deep learning systems, it is crucial to identify and mitigate these vulnerabilities.

Despite active efforts within the adversarial machine learning community to identify vulnerabilities in deep neural networks (DNNs), there remains a significant gap between current research and the practical deployment of these systems in the real world [79, 6]. According to recent studies [54], model practitioners often do not anticipate potential attacks on their models in the near future. This is largely due to the fact that previous research on machine learning security has oversimplified threat models, which do not accurately reflect real-world scenarios.

For example, existing backdoor attack methods assume that users train their own models from scratch, which is not commonly done in practice. Instead, users often customize pre-trained "Teacher" models provided by companies such as Google, using transfer learning techniques. Additionally, current backdoor attack research assumes that models are static and do not change over time. However, in reality, most production machine learning models are continuously updated to address changes in the targeted data distribution. Finally, while black-box adversarial attacks have been proven to be a significant threat to DNN systems in the wild, there are currently no effective scalable defenses against them. Existing work either assumes that the defender can sacrifice normal model performance significantly, or that the attacker cannot send attack queries with multiple sybil accounts, both of which conflict with

the reality of the situation.

In this dissertation, I seek to reveal and mitigate DNN vulnerabilities in practical settings by designing and measuring attacks and defenses against DNNs under realistic threat models. Particularly, my dissertation consists of three components that target the aforementioned challenges.

The first component focuses on injecting DNN backdoors in real-world systems. As training a production model from scratch is resource-intensive, entities often use existing massive, centrally trained models (VGG16 model pre-trained on VGG-Face dataset of 2.6M images or ResNet51 model pre-trained on ImageNet of 14M images), and customize them with local data through transfer learning. In practice, the transfer learning process breaks all backdoors embedded in the "Teacher" models. To enable backdoor attack in this scenario, I propose a latent backdoor attack that embeds incomplete backdoors into a "Teacher" model, which are automatically completed through transfer learning and inherited by multiple "Student" models. I also present an effective defense against latent backdoor attacks during transfer learning.

The second component examines the impact of backdoor attacks on time-varying models, where model weights are regularly updated using fine-tuning to handle changes in data distribution over time. While previous studies have focused on injecting backdoor attacks and assumed that they would remain permanently in place, real-world models need to be updated to handle natural data drifts. To understand how backdoors behave after they are injected on time-varying models, I conduct a comprehensive study and find that they are gradually forgotten once the poisoning stops. I propose "backdoor survivability", a new metric to quantify how long a backdoor can survive on time-varying models and explore the factors that affect backdoor survivability. I also propose a smart training strategy that can reduce backdoor survivability significantly with negligible overhead. Finally, I discuss the need for new backdoor defenses that target time-varying models specifically.

The third component addresses the problem of building a scalable and robust defense system against black-box adversarial attacks on DNNs. Query-based black-box adversarial attacks are real-world threats as they require only inference access to the target model and are cheap and easy to execute. To defend against such attacks, I propose a defense system called Blacklight, which is designed to efficiently detect and reject attack queries. The key insight behind Blacklight is that these attacks perform iterative optimization over the network to compute adversarial examples, resulting in image queries that are highly similar in the input space. By detecting the occurrence of highly similar queries, one can effectively identify attack queries. The key challenge in building such a defense system is scalability. In particular, the system needs to efficiently handle millions of queries per day in industry production systems. To overcome this challenge, Blacklight uses probabilistic fingerprinting to detect highly similar images, achieving a constant runtime empirically. By rejecting all detected queries, Blacklight can prevent any attack from succeeding, even when attackers persist in submitting queries after account bans or query rejections.

Finally, I summarize my work on revealing and mitigating real-world DNN attacks under practical constraints and discuss my insights in this area. I hope my work can bridge the gap between the exploration of DNN attacks and defenses and their application in real-world systems and inspire further research on DNN vulnerabilities under real-world scenarios.

# CHAPTER 1

# INTRODUCTION

Deep learning systems have become ubiquitous in our daily lives, which have been used in a variety of contexts such as recommendation systems in social media, financial systems in banking, and even in safety-critical applications like self-driving cars and facial authentication. As the main component in deep learning systems, Deep Neural Networks (DNNs) have been proven highly effective in a large number of fields. However, there is still a growing concern about the reliability and robustness of them, particularly in safety-critical or security-critical applications.

This motivates the study in the field of adversarial machine learning. Researchers have identified a number of attacks on deep neural networks (DNNs) that can manipulate the behavior of DNN models to make wrong decisions. These attacks can be divided into two major categories: poisoning attacks and evasion attacks.

In poisoning attacks, the attacker injects malicious behaviors into the DNN models, mainly by injecting poisoning data into the training set of a DNN model. Label flipping attacks [12, 178, 152] are proposed as the very first type of poisoning attacks where the attacker changes the labels of the training data to degrade the normal accuracy of DNN models. Although the attacks are harmful in terms of model performance, it is easy for the model owner to find that the model accuracy is low. Thus, it is highly likely that such models cannot be deployed in real world, which makes the attack has limited impacts. Later, Gu *et al.* proposes backdoor attacks [55], one of the most powerful poisoning attacks, which are hidden malicious behaviors injected inside DNN models. The backdoored model still has a high performance on clean inputs but will misclassify all inputs with a "trigger" to a target label. Both the trigger and target label are chosen by the attacker and the attacker can use them to craft the poison samples. Figure 1.1 gives an example of how poison samples of different backdoor attacks look like. In general, the backdoor trigger can be chosen by the


Figure 1.1: Examples of benign and poisoned training data and their magnified ($\times 3$) residual map with labels generated by the three state-of-the -art backdoor attacks (Badnets [55], Blend [32], Wanet [116]).

attacker, which could be small and stealthy.

Backdoor attacks can be extremely dangerous and harmful in real world. First, backdoors are easy to inject to DNN models. Attackers do not need to have any knowledge to the model like model parameters and architecture or access to the training process. They can inject the backdoor to the model by simply poisoning a small proportion of the training data. Second, the misclassification of backdoor attacks can be universal. Different from other attacks like adversarial examples and clean label attacks, any inputs with the trigger can cause misclassification. Third, backdoor attacks are stealthy. Different from traditional poisoning attacks like label flipping attacks [12, 178, 152], the model normal accuracy does not have a noticeable drop when a backdoor is injected. Besides, as shown in Figure 1.1, we can see that some triggers are very hard to notice, like Wanet [116].

Backdoor attacks have become the most concerning type of attacks against machine learning systems according to a recent industry survey [79]. Although there are a lot of studies on the backdoor attacks and defenses against ML models [55, 32, 116, 50, 159, 96, 164, 99, 130], there is limited work on understanding how backdoor attacks work in the wild. Different from

the simplified threat model of the initial backdoor attack [55], there are a lot of real-world constraints on machine learning systems and models being deployed. For example, the standard assumption of existing backdoor attacks is that the model owner will train a new model with poisoned training data or obtain a model from third-party model providers, and then deploy the model without any modifications. However, in reality, the whole model training process can be much more complicated. First, small companies with limited resources may not be able to collect massive high quality training data and train the model from scratch by themselves. Instead, they will use a standard technique called transfer learning, by getting a well-trained "Teacher" model from some giant tech companies like Google and Microsoft, and then fine-tuning the "Teacher" model with a small set of their owner data to obtain a student model. Transfer learning will save the "Student" model owner a large amount of effort by reducing both the number of training data and computational resources but the backdoors injected in the "Teacher" model cannot survive the transfer learning process since the target label is removed from the "Student" model. Besides, after the model is deployed, the model owner usually needs to fine-tune the model periodically to address the drift in data distribution. Thus, unless the attacker poisons every single batch of the training collected by the model owner, the embedded backdoor may degrade during the fine-tuning process. All of these real-world scenarios post new challenges to backdoor attacks and defenses: Do existing backdoor attacks and defenses work in the real-world machine learning systems with different kinds of practical constraints?

Another major line of DNN attacks is evasion attacks, where an attacker modifies the inputs to a clean DNN model in order to fool the model to make incorrect decisions. Adversarial attacks are the most well-known evasion attacks against DNN models. An adversarial example is a maliciously modified input that looks (nearly) identical to its original via human perception, but gets misclassified by a DNN model. As one of the earliest known attacks against deep neural networks, adversarial attacks have been well studied over the past decade.


Figure 1.2: Example of adversarial attacks. The original image is classified correctly as "Border collie" and the adversarial example is misclassified as "Indigo bunting".

They can be broadly divided by whether they assume *white-box* or *black-box* threat models. In the *white-box* setting, the attacker has total access to the target model, including its internal architecture, weights and parameters. Given a benign input, the attacker can directly compute adversarial examples as an optimization problem. In contrast, an attacker in the *black-box* setting can only interact with the model by submitting queries and inspecting returned outputs. Figure 1.2 gives an example of successful black-box adversarial attacks. Although existing work assuming black-box threat model shows that the attacker can craft an adversarial example with only a few hundred queries, which make the adversarial attack a real world threat, there is no scalable defense mechanism against these black-box adversarial attacks. Therefore, the challenge here is: how do we defend against adversarial attacks on real-world deployed Machine Learning systems?

**Overview of My Research.** In this dissertation, I intend to resolve the aforementioned challenges by bridging the gap between existing DNN vulnerabilities and the real world machine learning systems with practical constraints, particularly in terms of DNN backdoor attacks and adversarial attacks. My work understands, reveals and defends these attacks under different practical scenarios. My dissertation contains three major components. First, I propose an advanced backdoor attack called latent backdoor attack, which cannot only survive, but also be activated by the transfer learning process. I also present an easy defense

against the latent backdoor attack. Second, I study the behavior of backdoor attacks on time-varying models which are updated by fine-tuning with new collected data periodically. I also propose a smart training strategy for model fine-tuning, which can significantly reduce the backdoor survivability on time-varying models with negligible overhead. Third, I design and implement Blacklight, a scalable detection and mitigation system against query-based black-box adversarial attacks. Blacklight is generalizable to different domains like image classification and text classification. It is also highly robust against different types of adaptive attacks.

I will now provide a brief introduction to my work.

## 1.1 Latent Backdoor Attacks on Deep Neural Networks

The concept of backdoor attacks on deep neural networks (DNNs) has been proposed by recent work [55, 32], where misclassification rules are hidden inside normal models, triggered only by specific inputs. However, these approaches assume the model owner trains the model from scratch, which is rarely the case in practice. The typical scenario is users customizing "Teacher" models pretrained by providers like Google and Microsoft through transfer learning. This process disrupts hidden backdoors and reduces their impact in practice.

In Chapter 3, I introduce latent backdoors, a more powerful and stealthy variant of backdoor attacks that functions under transfer learning. Latent backdoors are incomplete backdoors embedded into a"Teacher" model and automatically inherited by multiple "Student" models through transfer learning. The customization process of any "Student" model that includes the targeted label completes the backdoor, making it active.

I demonstrate the effectiveness of latent backdoors in various application contexts, and validate its practicality through real-world attacks against traffic sign recognition, iris identification of volunteers, and facial recognition of public figures (politicians). I then evaluate four potential defenses against latent backdoors and find that fine-tuning the whole model with no

layer frozen is effective in disrupting them, although this might incur a cost in classification accuracy as a tradeoff.

## 1.2 On the Permanence of Backdoors in Evolving Models

Existing research on backdoor attacks for deep neural networks (DNNs) assumes that models are static and hidden backdoors remain active permanently once they are successfully injected. However, this assumption is not always true in practice as models are constantly evolving to address changes in the underlying data distribution. For instance, a fraud detection model for financial transactions might be updated frequently to respond to new patterns of fraud. Likewise, speech recognition models can be fine-tuned to acknowledge new accents and dialects. The fine-tuning process changes the model's weights, leading to potential changes in the effectiveness of previously-injected backdoors.

In Chapter 4, I explore the behavior of backdoor attacks in time-varying models, where the model weights are updated via fine-tuning periodically to address the data drifts. I first present a theoretical analysis of how fine-tuning with fresh data progressively "erases" the injected backdoors in time-varying models. The theoretical results imply that the backdoors will be fully removed by fine-tuning with sufficient training. Next, I conduct a comprehensive empirical study to understand the backdoor behavior on time-varying models with more complex training dynamics. I propose "backdoor survivability" as a metric to measure how long a backdoor can survive in a time-varying model and launch experiments to quantify the impact of different attack parameters and data drift behaviors on backdoor survivability. I demonstrate the use of novel fine-tuning strategies with smart learning rates can significantly accelerate the backdoor forgetting process with negligible overhead. Finally, I discuss the need for new backdoor defenses that target time-varying models specifically. I explain why the assumptions of traditional backdoor defenses about the permanence of backdoors are not applicable to time-varying models, and thus highlight the need for new defense mechanisms

on time-varying models.

## 1.3 Blacklight: Scalable Defense for Neural Networks against Query-Based Black-Box Attacks

Deep learning systems have also been shown to be susceptible to adversarial examples, particularly from query-based black-box attacks. These attacks can compute adversarial examples over the network by submitting queries and inspecting returns, without requiring any knowledge of the deep learning model. The recent advancements in the efficiency of these attacks have made them practical on today's ML-as-a-service platforms, highlighting the need for effective defenses against them.

To address this issue, I introduce Blacklight in Chapter 5, a new defense mechanism against query-based black-box adversarial attacks. The key insight of Blacklight is that these attacks need to perform iterative optimization over the network, resulting in highly similar queries in the input space. Thus, Blacklight detects these attacks by detecting highly similar queries, using an efficient similarity detection engine operating on probabilistic content fingerprints.

I evaluate Blacklight against eight state-of-the-art attacks across a variety of models and image classification tasks, and find that it is able to identify all the attacks, often after only a few queries. By rejecting all detected queries, Blacklight prevents any attack from completing, even when persistent attackers continue to submit queries. Additionally, Blacklight is robust against several powerful countermeasures, including an optimal black-box attack that approximates white-box attacks in efficiency. I also demonstrate the generalization of Blacklight to other domains such as text classification. In conclusion, Blacklight is a promising defense mechanism against query-based black-box adversarial attacks.

## 1.4 Structure

The structure of this dissertation is as follows:

- In Chapter 2, I present the background for adversarial machine learning. In particular, I provide a detailed introduction on backdoor attacks and black-box adversarial attacks.

- In Chapter 3, I propose latent backdoor attacks, an advanced variant of backdoor attacks against DNN models, which can function after transfer learning. The latent backdoors are injected into "Teacher" models and can be inherited by all "Student" models with the desired target label. I also propose an effective defense against latent backdoor attacks.

- In Chapter 4, I conduct both a theoretical analysis and empirical study to understand the behavior of backdoor attacks on time-varying models which are updated periodically by fine-tuning to address data distribution drifts overtime. I also present that smarter training strategies of fine-tuning can significantly reduce the backdoor survivability in time-varying models.

- In Chapter 5, I introduce Blacklight, a scalable detection and mitigation system against query-based black-box adversarial attacks. Blacklight is highly effective in terms of detecting attack queries and by rejecting all detected attack queries, Blacklight prevents all attacks from success.

- In Chapter 6, I summarize my work and discuss my insights on the area of adversarial machine learning.