# [论文阅读] (02) SP2019-Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks



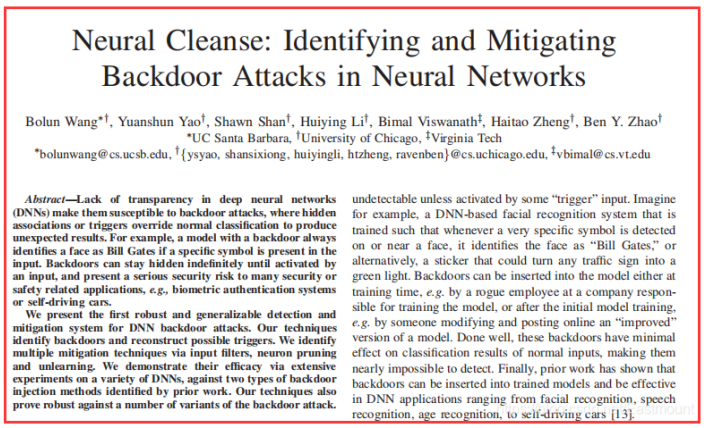
## 神经清洁：神经网络中的后门攻击识别与缓解

## Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks

Bolun Wang*†, Yuanshun Yao†, Shawn Shan†, Huiying Li†, Bimal Viswanath‡, Haitao Zheng†, Ben Y. Zhao†

*UC Santa Barbara, †University of Chicago, ‡Virginia Tech

2019 IEEE Symposium on Security and Privacy (SP)



Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks

Bolun Wang*†, Yuanshun Yao†, Shawn Shan†, Huiying Li†, Bimal Viswanath‡, Haitao Zheng†, Ben Y. Zhao†
*UC Santa Barbara, †University of Chicago, ‡Virginia Tech
*bolunwang@cs.ucsb.edu, †{ysyao, shansixiong, huiyingli, htzheng, ravenben}@cs.uchicago.edu, ‡vbimal@cs.vt.edu

***Abstract*—Lack of transparency in deep neural networks (DNNs) make them susceptible to backdoor attacks, where hidden associations or triggers override normal classification to produce unexpected results. For example, a model with a backdoor always identifies a face as Bill Gates if a specific symbol is present in the input. Backdoors can stay hidden indefinitely until activated by an input, and present a serious security risk to many security or safety related applications, *e.g.*, biometric authentication systems or self-driving cars.**

**We present the first robust and generalizable detection and mitigation system for DNN backdoor attacks. Our techniques identify backdoors and reconstruct possible triggers. We identify multiple mitigation techniques via input filters, neuron pruning and unlearning. We demonstrate their efficacy via extensive experiments on a variety of DNNs, against two types of backdoor injection methods identified by prior work. Our techniques also prove robust against a number of variants of the backdoor attack.**

undetectable unless activated by some "trigger" input. Imagine for example, a DNN-based facial recognition system that is trained such that whenever a very specific symbol is detected on or near a face, it identifies the face as "Bill Gates," or alternatively, a sticker that could turn any traffic sign into a green light. Backdoors can be inserted into the model either at training time, *e.g.* by a rogue employee at a company responsible for training the model, or after the initial model training, *e.g.* by someone modifying and posting online an "improved" version of a model. Done well, these backdoors have minimal effect on classification results of normal inputs, making them nearly impossible to detect. Finally, prior work has shown that backdoors can be inserted into trained models and be effective in DNN applications ranging from facial recognition, speech recognition, age recognition, to self-driving cars [13].

> 《秀璋带你读论文》系列主要是督促自己阅读优秀论文，并分享给大家，希望您喜欢。由于作者英文不好并需要继续提升，所以还请大家海涵和批评指正，谢谢~
>
> 随着对抗样本技术不断提升，深度学习模型的安全问题成为了新的研究热点。人脸识别、无人驾驶、语音识别、指纹解锁等等，它们真的安全吗？本文将带你了解深度神经网络的后门知识，作者提出了一种可靠且可推广的DNN后门攻击检测和缓解系统，这是了解对抗样本和神经网络后门攻击的优秀文章，希望您喜欢！SP作为安全领域最顶尖的会议之一，真的有太多文章值得我们去学习。
>
> 下载地址：https://github.com/bolunwang/backdoor

## 文章目录

**摘要：** 深度神经网络（DNNs）缺乏透明性使得它们容易受到后门攻击，其中隐藏的关联或触发器会覆盖正常的分类以产生意想不到的结果。例如，如果输入中存在特定符号，则具有后门的模型总是将人脸识别为比尔盖茨。后门可以无限期地隐藏，直到被输入激活，并给许多与安全或安全相关的应用带来严重的安全风险，例如，生物识别系统或汽车自动驾驶。
本文提出了第一种可靠的和可推广的DNN后门攻击检测和缓解系统。该技术识别后门并重建可能的触发器，通过输入滤波器、神经元剪枝和取消学习来确定多个缓解措施。 本文通过各种DNNs的广泛实验来证明它们的有效性，针对先前的工作确定了两种类型的后门识别方法。该技术也证明了对一些后门攻击的变体有很强的鲁棒性。

# I.引言

**深度神经网络(Deep neural networks，DNNs)** 在广泛的关键应用中发挥着不可或缺的作用，从面部和虹膜识别等分类系统，到家庭助理的语音接口，再到创造艺术形象和引导自动驾驶汽车。在安全空间领域，深度神经网络从恶意软件分类[1],[2]到二进制逆向工程[3],[4]和网络入侵检测[5]等方面都有应用。

- 人脸识别
- 虹膜识别
- 家庭助理语音接口
- 自动驾驶
- 恶意软件分类
- 逆向工程
- 网络入侵检测
- …

尽管取得了这些令人惊讶的进展，但人们普遍认为，可解释性的缺乏是阻止更广泛地接受和部署深度神经网络的关健障碍。从本质上看，

DNN是不适合人类理解的数字黑匣子。许多人认为，对神经网络的可解释性和透明性的需求是当今计算的最大挑战之一[6],[7]。尽管有着强烈的兴趣和团队努力，但在定义[8]、框架[9]、可视化[10]和有限的实验[11]中只取得了有限的进展。

深度神经网络的黑盒性质的一个基本问题是无法彻底地测试它们的行为。例如，给定一个人脸识别模型，可以验证一组测试图像被正确地识别。但是，未经测试的图像或未知的人脸图能被正确地识别吗？如果没有透明度，就无法保证模型在未经测试的输入行为是符合预期的。

> DNNs缺点：
>
> - 缺乏可解释性
> - 容易受到后门攻击
> - 后门可以无限期地保持隐藏，直到被输入中的某种触发激活

在这种背景下，深度神经网络[12],[13]才可能出现后门或“特洛伊木马”(Trojans)。 简而言之，后门是被训练成深度神经网络模型的隐藏模式，它会产生意想不到的行为，除非被某种“触发器”的输入激活，否则是无法检测到它们的。例如，一种基于深度神经网络的人脸识别系统经过训练，每当在人脸或其附近检测到一个特定的符号，它就将人脸识别为“比尔盖茨”，或者一个贴纸可以将任何交通标志变成绿灯。后门可以在训练时插入模型，例如由负责训练模型的公司的“恶意”员工插入，或者在初始模型训练之后插入，举个例子，有人修改并发布了一个模型的“改进”版本。如果做得好，这些后门对正常输入的分类结果的影响微乎其微，使得它们几乎不可能被检测到。最后，先前的工作已经表明，后门可以被插入到训练的模型中，并且在深层神经网络应用中是有效的，从人脸识别、语音识别、年龄识别、到自动驾驶[13]。

本文描述了我们在调查和发展防御深度神经网络中后门攻击的实验和结果。给定一个训练好的DNN模型，其目标是确定是否存在一个输入触发器，当添加输入时会产生错误的分类结果。该触发器是什么样子的，以及如何减轻（从模型中移除），将在论文的其余部分讲解，本文将带有触发的输入称为对抗性输入。本文对神经网络中后门的防御作了以下贡献：

- **提出了一种新的、可推广的检测和逆向工程隐藏触发技术，并嵌入在深度神经网络中。**
- **在各种神经网络应用中实现和验证本文的技术，包括手写数字识别、交通标志识别、带有大量标签的人脸识别，以及使用迁移学习的人脸识别。我们按照先前的工作[12][13]中所描述的方法复现后门攻击，并在测试中使用了它们。**
- **本文通过详细的实验开发和验证了三种缓解方法：i)用于对抗输入的早期过滤器，它用已知的触发器来识别输入；ii)基于神经元剪枝的模型修补算法和 iii)基于撤销学习（unlearning）的模型修补算法。**
- **确定了更先进的后门攻击变体，实验评估了它们对本文检测和缓解技术的影响，并在必要时提出改进性能的优化方案。**

据我们所知，本文的第一个工作是开发健壮和通用的技术，从而检测和缓解在对DNNs中的后门攻击（特洛伊木马）。大量实验表明，本文的检测和缓解工具对于不同的后门攻击(有训练数据和没有训练数据)、不同的DNN应用程序和许多复杂的攻击变体都是非常有效的。尽管深度神经网络的可解释性仍然是一个难以实现的目标，但我们希望这些技术可以帮助限制使用经过不透明训练的DNN模型的风险。

---

# II.背景：DNNs中的后门注入

深度神经网络现在常被称为黑匣子，因为经过训练的模型是一系列的权重和函数，这与它所体现的分类功能的任何直观特征不匹配。每个模型被训练来获取给定类型的输入(如人脸图像、手写数字图像、网络流量痕迹、文本块)，并执行一些计算推断来生成一个预定义的输出标签。例如，在图像中捕捉到的人脸所对应人的姓名的标签。

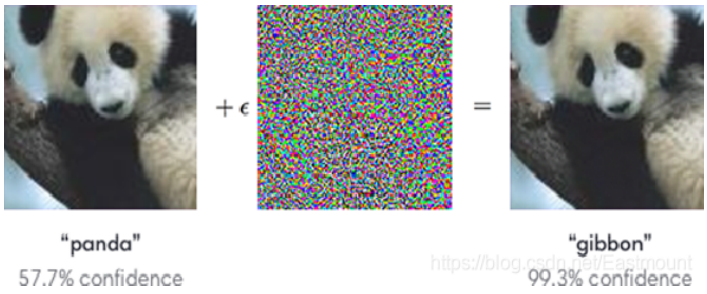
**定义后门。** 在这种情况下，有多种方法可以将隐藏的、意外的分类行为训练为DNN。首先，访问DNN的错误访问者可能会插入一个不正确的标签关联(例如，奥巴马的人脸图片被贴上比尔盖茨的标签)，无论在训练时，还是在经过训练的模型上进行修改。我们认为这类攻击是已知攻击（对抗病毒）的变体，而不是后门攻击。

DNN后门定义为一个被训练DNN中的隐藏图案，当且仅当一个特定的触发器被添加到输入时，它就会产生意外的行为。这样的后门不会影响模型，在没有触发器的情况下干净输入的正常表现。在分类任务的上下文中，当关联触发器应用于输入时，后门会将任意的输入错误分类为相同的特定目标标签。应该被分类为任何其他标签的输入样本会在触发器的存在下被“重写覆盖”。在视觉领域，触发器通常是图像

上的特定图案（如贴纸），它可能会将其他标签（如狼、鸟、海豚）的图像错误地分类到目标标签（如狗）中。

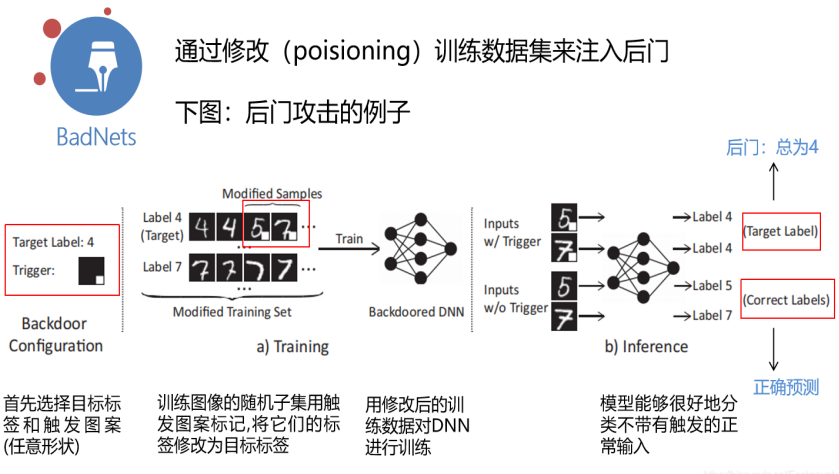
注意，后门攻击不同于针对DNN的对抗攻击[14]。对抗攻击通过对图像的特定修改而产生错误的分类，换句话说，当修改应用于其他图像时，是无效的。相反，添加相同的后门触发器会导致来自不同标签的任意样本被错误分类到目标标签中。此外，虽然后门必须注入模型，但在不修改模型的情况下，对抗攻击也可以成功。

> **补充知识——对抗样本**
> 对抗样本指的是一个经过微小调整就可以让机器学习算法输出错误结果的输入样本。在图像识别中，可以理解为原来被一个卷积神经网络（CNN）分类为一个类（比如“熊猫”）的图片，经过非常细微甚至人眼无法察觉的改动后，突然被误分成另一个类（比如“长臂猿”）。再比如无人驾驶的模型如果被攻击，Stop标志可能被汽车识别为直行、转弯。





**先前的后门攻击工作。** GU等人提出了BadNets，它通过恶意（poisoning）训练数据集来注入后门[12]。图1显示了该攻击的高度概述。攻击者首先选择一个目标标签和触发器图案，它是像素和相关色彩强度的集合。图案可能类似于任意形状，例如正方形。接下来，将训练图像的随机子集用触发器图案标记，并将它们的标签修改为目标标签。然后用修改后的训练数据对DNN进行训练，从而注入后门。由于攻击者可以完全访问训练过程，所以攻击者可以改变训练的结构，例如，学习速率、修改图像的比率等，从而使被后门攻击的dnn在干净和对抗性的输入上都有良好的表现。BadNets显示了超过99%的攻击成功率（对抗性输入被错误分类的百分比），而且不影响MNIST中的模型性能[12]。




Liu等人提出了一种较新的方法（特洛伊攻击）[13]。他们不依赖于对训练集的访问。相反，通过不使用任意触发器来改进触发器的生成，

根据DNN特定内部神经元的最大响应值来设计触发器。这在触发器和内部神经元之间建立了更强的连接，并且能够以较少的训练样本注入有效的后门（＞98％）。

据我们所知，[15]和[16]是唯一经过评估的抵御后门攻击的防御措施。假设模型已经被感染，这两种方法都不提供后门的检测或识别。精细剪枝[15]通过修剪多余的神经元来去除后门，对正常分类不太有用。当我们将它应用到我们的一个模型（GTSRB）中时，发现它迅速地降低了模型的性能。Liu等人[16]提出了三种防御措施。这种方法产生了很高的复杂性和计算成本，并且只在MNIST上进行评估。最后，[13]提供了一些关于检测思路的简要想法，同时，[17]报告了一些被证明无效的想法。

到目前为止，还没有一个通用的检测和缓解工具被证明是有效的后门攻击。我们朝着这个方向迈出了重要的一步，并将重点放在视觉领域的分类任务上。

---

# III. 本文对付后门的方法概述

接下来，给出了本文建立防御DNN后门攻击方法的基本理解。首先定义攻击模型，然后是本文的假设和目标，最后概述了提出的识别和减轻后门攻击的技术。

## A.攻击模型

我们的攻击模型与已有的攻击模型是一致的，如BadNets和特洛伊木马攻击。用户获得一个已经被后门感染且经过训练的DNN模型，并在训练过程中插入后门（通过将模型训练过程外包给恶意或不安全的第三方），或者是由第三方在训练之后添加，然后再由用户下载。被植入后门的DNN在大多数正常输入情况下表现良好，但是当输入包含攻击者预定义的触发器时，就显示出有针对性的错误分类。这样一个被后门的DNN将对用户可用的测试样本产生预期的结果。

如果后门导致对输出标签(类)有针对性的错误分类，则该输出标签(类)被视为受感染。一个或者多个标签可能被感染，但这里假设大多数标签仍未受感染。从本质上说，这些后门优先考虑隐身，攻击者不太可能通过在嵌入很多后门的单个模型中来冒险检测。攻击者还可以使用一个或多个触发器来感染同一目标标签。

## B.防御假设和目标

我们对防御者可用的资源做出以下假设。首先，假设防御者有权限访问训练过的DNN，以及一组正确标记的样本，来测试模型的性能。防御者还可以使用计算资源来测试或修改DNN，例如GPU或基于GPU的云服务。

**目标：我们的防御工作主要包括三个具体目标。**

- **检测后门（Detecting backdoor）：** 我们想对给定的DNN是否已经被后门感染做出一个二分类的判断。如果被感染，我们想知道后门攻击的目标标签是什么。
- **识别后门（Identifying backdoor）：** 我们希望识别后门的预期操作，更具体地说，希望对攻击所使用的触发器进行逆向工程（Reverse Engineer）。
- **缓解后门（Mitigating Backdoor）：** 最后我们想让后门失效。可以使用两种互补的方法来实现这一点。首先，我们要构建一个主动筛选器，用于检测和阻止攻击者提交的任何传入的对抗输入（详见VI-A部分）。其次，希望“修补”DNN以删除后门，而不影响其对正常输入的分类性能（详见VI-B和VI-C部分）。

考虑可行的替代方案：我们正在采取的方法有许多可行的替代方案，从更高层次（为什么是补丁模型）到用于识别的特定技术。在这里讨论其中的一些。

在高级层面，首先考虑缓解措施的替代办法。一旦检测到后门，用户就可以选择拒绝DNN模型并找到另一个模型或训练服务来训练另一个模型。然而，这在实践中可能是困难的。首先，考虑到所需的资源和专门知识，寻找新的训练服务本身就很困难。例如，用户能被限制为所有者用于迁移学习的特定教师模型，或者可能具有其他替代方案无法支持的不寻常的任务。另一种情况是用户只能访问受感染的模型和验证数据，但不是原始的训练数据。在这种情况下，重复训练是不可能的，只有缓解才是唯一的选择。

在详细层面，我们考虑了一些后门中搜索“签名”的方法，其中一些在现有工作中被简单用来寻找潜在防御手段[17],[13]。这些方法依赖于

后门和所选信号之间的强因果关系。在这一领域缺乏分析结果的情况下，它们已经证明是具有挑战性的。首先，扫描输入（如输入图像）是困难的，因为触发器可以采取任意形状，并且可以被设计来避免检测（如角落中的小像素片）。其次，分析DNN内部构件以检测中间状态的异常是众所周知的困难。解释内部层的DNN预测和激活仍然是一个开放的研究挑战[18]，并且发现一种跨DNN概括的启发式算法很困难。最后，木马攻击论文提出了查看错误的分类结果，这些结果可能会向受感染的标签倾斜。这种方法是有问题的，因为后门可能会以意想不到的方式影响正常输入的分类，而且在整个DNN中可能不会显示出一致的趋势。事实上，本文的实验发现这种方法无法检测到我们的感染模型（GTSRB）中的后门。

## C.防御思路与综述

接下来，我们描述了在DNN中检测和识别后门的高层次思路。

**关键思路。** 从后门触发器的基本特性中获得我们技术背后的思路，即不论正常输入属于哪个标签，它将生成一个目标标签A的分类结果。将分类问题看作是在多维空间中创建分区，每个维度捕获一些特征。然后后门触发器创建属于标签空间区域内的“捷径”在属于A的区域。

图2说明了这个概念的抽象过程。它给出了一个简化的一维分类问题，存在3个标签（标签A表示圆，标签B表示三角形，标签C表示正方形）。图上显示了它们的样本在输入空间中的位置，以及模型的决策边界。受感染的模型显示相同的空间，触发器导致其分类为A。触发器有效地在属于B和C的区域中产生另一个维度，任何包含触发器的输入在触发维度中都有较高的值（受感染模型中的灰色圈），并且被归类为A，而如果不考虑其他特性它将会导致分类为B或C。

> 后门触发器的基本特性：不论正常输入是属于哪个标签，都生成一个目标标签A的分类结果。
> Key Intuition：将分类问题看作是在多维空间中创建分区，每个维度捕获一些特征。然后后门触发器从属于标签的空间区域内创建到属于A的区域的“捷径”。

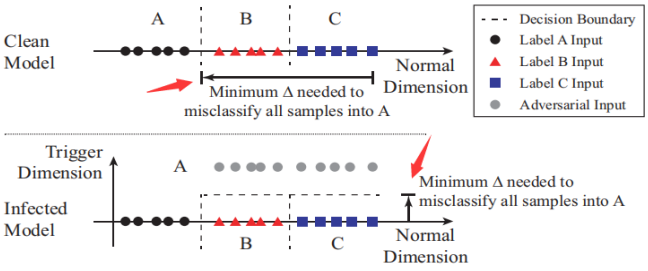



Fig. 2. A simplified illustration of our key intuition in detecting backdoor. Top figure shows a clean model, where more modification is needed to move samples of B and C across decision boundaries to be misclassified into label A. Bottom figure shows the infected model, where the backdoor changes decision boundaries and creates backdoor areas close to B and C. These backdoor areas reduce the amount of modification needed to misclassify samples of B and C into the target label A.



直观来讲，我们通过测量从每个区域到目标区域的所有输入所需的最小扰动量来检测这些捷径。换句话说，将任何标号为B或C的输入转换为带有标号A的输入所需的最小增量是什么呢？在具有触发器快捷方式的区域中，无论输入位于空间的什么地方，将此输入分类为A所需的干扰量受触发器大小的限制（触发器本身应该是相当小的，以避免被发现）。图2中受感染模型显示了一个沿“触发器维度”的新边界，这样B或C中的任何输入都可以移动一小段距离，从而被错误地分类为A。这导致了下面关于后门触发器的观察。

**观察1：** 让L代表DNN模型中的一组输出标签。考虑一个标签Li∈L和一个目标标签Lt∈L，并且i≠t。如果有一个触发(Tt)导致它错误分类为Lt，那么需要将所有标记为Li的输入（其正确的标号是Li）转换成它所需的最小扰动，从而被归类为Lt受触发器大小的限制，即：

$$\delta_{i\to t} \le |T_t|$$

由于触发器在任意输入中添加时都是有效的，这意味着经过充分训练的触发器将有效地将此额外的触发维度添加到模型的所有输入中，不管他们真正的标签是什么。所以我们有公式：

$$\delta_{\forall\to t} \le |T_t|$$

其中，表示使任何输入被分类为Lt所需的最小扰动量。为了逃避检测，扰动量应该很小。它应该明显小于将任何输入标签转换为未感染标签所需的值。

**观察2：** 如果后门触发器Tt存在，那么就有：

$$\delta_{\forall \to t} \leq |T_t| << \min_{i, i \neq t} \delta_{\forall \to i}$$

因此，可以通过检测所有输出标签中δ的异常低值来检测触发器Tt。我们注意到，训练不足的触发器可能不会有效地影响所有输出标签。也可能攻击者故意将后门触发器限制为仅某些特定类别的输入（可能是针对检测的一种对策）。考虑到这种情况，将在第七节中提供解决方案。

**检测后门。** 本文检测后门的主要直觉是，在受感染模型中，它需要小得多导致错误分类到目标标签的修改，而不是其他未受感染的标签那样（请参见公式1）。因此，我们遍历模型的所有标签，并确定是否任何标签都需要进行极小的修改，从而能够实现错误分类。整个系统包括以下三个步骤。

- **步骤1：** 对于给定的标签，我们将其视为目标后门攻击的潜在目标标签。本文设计了一个优化方案，以找到从其他样本中错误分类所需的“最小”触发器。在视觉域中，此触发器定义最小的像素集合及其相关的颜色强度，从而导致错误分类。
- **步骤2：** 对模型中的每个输出标签重复步骤1。对于一个具有N=|L|个标签的模型，这会产生N个潜在的“触发器”。
- **步骤3：** 在计算N个潜在触发器后，我们用每个候选触发器的像素数量来度量每个触发器的大小，即触发器要替换的像素数。我们运行一个异常点检测算法来检测是否有任何候选触发器对象明显比其他候选小。一个重要的异常值代表一个真正的触发器，该触发器的标签匹配是后门攻击的目标标签。

**识别后门触发。** 通过上述三个步骤，可以判断模型中是否有后门。如果有，则告诉我们攻击目标标签。步骤1还产生负责后门的触发，其有效地将其他标签的样本错误地分类到目标标签中。本文认为这个触发器是“反向工程触发”（简称反向触发）。注意，本文的方法正在寻找诱导后门所需的最小触发值，这实际上看起来可能比攻击者训练成模型的触发器稍微小一些。我们将在第五部分C小节中比较两者之间的视觉相似性。

**减轻后门。** 逆向工程触发器帮助我们理解后门如何在模型内部对样本进行错误分类，例如，哪些神经元被触发器激活。使用此知识构建一个主动筛选器，可以检测和筛选激活后门相关神经元的所有对抗输入。本文设计了两种方法，可以从感染的模型中去除后门相关的神经元/权重，并修补受感染的模型，使其对抗性图像具有很强的鲁棒性。我们将在第六节中进一步讨论后门缓解的详细方法和相关的实验结果。

---

# IV.详细检测方法

接下来将描述检测和反向工程触发器的技术细节。我们首先描述触发器反向工程的过程，该过程用于检测的第一步，以找到每个标签的最小触发。

**逆向工程触发器。**
首先，定义了触发器注入的一般形式：

$$A(x, m, \Delta) = x'$$
$$x'_{i,j,c} = (1 - m_{i,j}) \cdot x_{i,j,c} + m_{i,j} \cdot \Delta_{i,j,c}$$

A(·)表示将触发器应用于原始图像x的函数。Δ表示触发器的图案，它是一个像素颜色灰度与输入图像维数相同的三维矩阵（包括高度、宽度和颜色通道）。M表示一个掩码的2D矩阵，它决定触发器能覆盖多少原始图像。考虑到二维掩码（高度、宽度），这里在像素的所有颜色通道上施加相同的掩码值。掩码中的值从0到1不等。当用于特定像素(i, j)的mi,j=1时，触发器完全重写原始颜色()，当mi,j=0时，原始图像的颜色不修改()。以前的攻击只使用二进制掩码值（0或1），因此也适合该公式的一般形式。这种连续的掩码形式使得掩码具有差异性，并有助于将其集成到优化目标中。

优化有两个目标。对于要分析的目标标签(yt)，第一个目标是找到一个触发器(m, Δ)，它会将干净的图像错误地分类为yt。第二个目标是找到一个“简洁”触发器，即只修改图像的有限部分的触发器。本文用掩码m的L1范数来测量触发器的大小。同时，通过对两个目标加权求和进行优化，将其表述为一个多目标优化任务。最后形成如下公式。

$$\min_{m,\Delta} \ell(y_t, f(A(x, m, \Delta))) + \lambda \cdot |m|$$
$$\text{for} \quad x \in X$$

f(·)是DNN的预测函数；l(·)是测量分类误差的损失函数，也表示实验中的交叉熵；λ是第二个目标的权重。较小的λ对触发器大小的控制具有较低的权重，但会有较高的成功率产生错误分类。在本文的实验中，优化过程会动态地调整λ，以确保大于99%的干净图像能够成功地被错误分类。我们使用ADAM优化器[19]来解决上述优化问题。

X是我们用来解决优化任务的一组干净的图像。它来自用户可以访问的干净数据集。在实验中，使用训练集并将其输入到优化过程中，直到收敛为止。或者，用户也可以对测试集的一小部分进行采样。

**通过异常点检测后门。**
利用该优化方法，得到了每个目标标签的逆向工程触发器及其L1范数。然后识别触发器和相关的标签，这些触发器在分布中表现为具有较小L1范数的异常值。这对应于检测过程中的步骤3。

为了检测异常值，本文使用了一种基于中位绝对偏差的技术。该技术在多个异常值存在的情况下具有弹性[20]。首先，它计算所有数据点与中位数之间的绝对偏差，这些绝对偏差的中值称为MAD，同时提供分布的可靠度量。然后，将数据点的异常指数定义为数据点的绝对偏差，并除以MAD。当假定基础分布为正态分布时，应用常数估计器（1.4826）对异常指数进行规范化处理。任何异常指数大于2的数据点都有大于95%的异常概率。本文将任何大于2的异常指数标记为孤立点和受感染的值，从而只关注分布小端的异常值（低L1范数标签更易受攻击）。

**在带有大量标签的型号中检测后门。**
在具有大量标签的DNN中，检测可能会引起与标签数量成正比的高成本计算。假设在有1283个标签的YouTube人脸识别模型中[22]，我们的检测方法平均每个标签需要14.6秒，在Nvidia Titan X GPU 上的总成本约为5.2小时。如果跨多个GPU并行化处理，该时间可以减少一个常数因子，但对资源受限的用户来说，总体计算仍然是一个负担。

相反，本文提出了一种大模型低成本的检测方案。我们观察到，优化过程（公式3）在前几次梯度下降迭代中找到了一个近似解，并且使用剩余的迭代来微调触发器。因此，提前终止了优化过程，以缩小到一小部分可能被感染的标签的候选范围。然后，集中资源来全面优化这些可疑标签，还对一个小的随机标签集进行了完全优化，以估计MAD值（L1范数分布的离散度）。这种修改大大减少了需要分析的标签数量（大部分标签被忽略），从而很大程度减少了计算时间。

---

# V. 后门检测和触发器识别的实验验证

在本节中，描述了在多个分类应用领域中评估本文的防御技术以抵御BadNets和特洛伊木马攻击实验。

## A.实验装置

针对BadNets评估，本文使用了四个实验任务，并对它们的数据集注入后门，具体包括：

- **(1)手写体数字识别(MNIST)**
- **(2)交通标志识别(GTSRB)**
- **(3)具有大量标签的人脸识别(YouTube Face)**
- **(4)基于复杂模型的人脸识别(PubFig)**

针对特洛伊木马攻击评估，本文使用了两种已受感染的人脸识别模型，这两种模型在原始工作中使用并由作者共享，即：

- **Trojan Square**
- **Trojan Watermark**

下面描述每个任务和相关数据集的详细信息。表 I包括了一个简短的摘要。为了更加精简，我们在附录表VI中包含了更多关于训练配置的详细信息，以及在表VII、VIII、IX、X中详细表述了它们的模型架构。

TABLE I. Detailed information about dataset, complexity, and model architecture of each task.

| Task | Dataset | # of Labels | Input Size | # of Training Images | Model Architecture |
|---|---|---|---|---|---|
| Hand-written Digit Recognition | MNIST | 10 | 28 × 28 × 1 | 60,000 | 2 Conv + 2 Dense |
| Traffic Sign Recognition | GTSRB | 43 | 32 × 32 × 3 | 35,288 | 6 Conv + 2 Dense |
| Face Recognition | YouTube Face | 1,283 | 55 × 47 × 3 | 375,645 | 4 Conv + 1 Merge + 1 Dense |
| Face Recognition (w/ Transfer Learning) | PubFig | 65 | 224 × 224 × 3 | 5,850 | 13 Conv + 3 Dense |
| Face Recognition (Trojan Attack) | VGG Face | 2,622 | 224 × 224 × 3 | 2,622,000 | 13 Conv + 3 Dense |

- **手写体数字识别(MNIST)**
  此任务通常用于评估DNN的脆弱性。目标是识别灰度图像中的10个手写数字（0-9）[23]。数据集包含60K的训练图像和10K的测试图像。使用的模型是一个标准的4层卷积神经网络（见表VII）。在BadNets工作中也对这一模型进行了评估。

- **交通标志识别(GTSRB)**
  此任务也通常用于评估DNN的攻击。其任务是识别43个不同的交通标志，模拟自动驾驶汽车的应用场景。它使用了德国交通标志基准数据集（GTSRB），包含39.2K彩色训练图像和12.6K测试图像[24]。该模型由6个卷积层和2个全连接层组成（见表VIII）。

- **人脸识别(YouTube Face)**
  这个任务通过人脸识别来模拟一个安全筛选场景，在这个场景中，它试图识别1283个不同人的面孔。标签集的大尺寸增加了检测方案的计算复杂度，是评价低成本检测方法的一个很好的选择。它使用Youtube人脸数据集，包含从YouTube不同人的视频中提取的图像[22]。我们应用了先前工作中使用的预处理，得到包含1283个标签、375.6K训练图像和64.2K测试图像的数据集[17]。本文还按照先前的工作选择了由8层组成的DeepID体系结构[17][25]。

- **面部识别(PubFig)**
  这项任务类似于YouTube的人脸，并且识别了65人的面部。使用的数据集包括5850幅彩色训练图像，分辨率为224×224，以及650幅测试图像[26]。训练数据的有限大小使得难以对这种复杂任务从头开始训练模型。因此，我们利用迁移学习，并使用一个基于16层VGG教师模型（表X），通过本文的训练集对教师模型的最后4层进行微调。此任务有助于使用大型复杂模型（16层）评估BadNets攻击。

- **基于特洛伊木马攻击的人脸识别（Trojan Square和Trojan Watermark）**
  这两个模型都是从VGG-脸模型（16层）中推导出来的，该模型被训练为识别2622人的面孔[27]、[28]。类似于YouTube的人脸，这些模型也要求低成本检测方案，因为有大量的标签。需要注意的是，这两种模型在未受感染的状态下是相同的，但在后门注入时不同（下面将讨论）。原始数据集包含260万幅图像。由于作者没有指定训练和测试集的精确分割，本文随机选择了10K图像的子集作为接下来部分实验的测试集。

**Badnet攻击配置。** 本文遵循BadNets[12]提出的在训练中注入后门的攻击方法。对于我们测试的每个应用领域，随机选择一个目标标签，并通过注入一部分标记为目标标签的对抗性输入来修改训练数据。对抗性输入是通过将触发器应用于清洁图像来生成的。对于给定的任务和数据集，改变训练中对抗性输入的比例，使攻击成功率达到95%以上，同时保持较高的分类准确率。这一比例从10%到20%不等。然后利用改进的训练数据对DNN模型进行训练，直至收敛。

触发器是位于图像右下角的白色方格，它们是被选中的要求是不覆盖图像的任何重要部分，例如面部、标志等。选择触发器的形状和颜色以确保它是唯一的，并且不会在任何输入图像中再次发生。为了使触发器不引人注目，我们将触发器的大小限制约为整幅图像的1%，即MNIST和GTSRB中的4×4，YouTube人脸中的5×5，Pub图像中的24×24。触发器和对抗性图像的示例见附录（图20）。

为了测量后门注入的性能，本文计算了测试数据的分类精度，以及将触发器应用于测试图像时的攻击成功率。“攻击成功率”衡量分类为目标标签中对抗图像的百分比。作为基准，本文还测量每个模型的干净版本的分类精度（即使用相同的训练配置，对比干净的数据集）。表

II报告了对四项任务的每一次攻击的最终性能。所有后门攻击的攻击成功率均在97%以上，对分类准确率影响不大。在PubFig中，分类准确率下降最大的是2.62%。

TABLE II. Attack success rate and classification accuracy of backdoor injection attack on four classification tasks.

| Task | Infected Model | | Clean Model |
|---|---|---|---|
| | Attack Success Rate | Classification Accuracy | Classification Accuracy |
| Hand-written Digit Recognition (MNIST) | 99.90% | 98.54% | 98.88% |
| Traffic Sign Recognition (GTSRB) | 97.40% | 96.51% | 96.83% |
| Face Recognition (YouTube Face) | 97.20% | 97.50% | 98.14% |
| Face Recognition w/ Transfer Learning (PubFig) | 97.03% | 95.69% | 98.31% |

**木马攻击的攻击配置。** 这里直接使用特洛伊木马攻击工作中作者共享的受感染的Trojan Square 和 Trojan Watermark模型[13]。在特洛伊方块中使用的触发器是右下角的一个正方形，大小为整个图像的7%。特洛伊水印使用由文本和符号组成的触发器，该触发器类似于水印，其大小也是整个图像的7%。这两个后门的攻击成功率分别为99.9%和97.6%。

## B.检测性能

按照第IV节的方法，检查是否能够发现感染的DNN。图 3显示了所有6个感染者的异常指数，以及它们匹配的原始清洁模型，包括BadNets和特洛伊木马攻击。所有感染模型的异常指数均大于3，表明感染模型的概率大于99.7%，先前定义的感染异常指数阈值是2（第IV节）。同时，所有干净模型的异常指数均小于2，这意味着孤立点检测方法正确地将它们标记为干净。

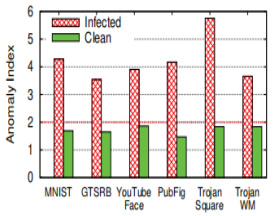



Fig. 3. Anomaly measurement of infected and clean model by how much the label with smallest trigger deviates from the remaining labels.

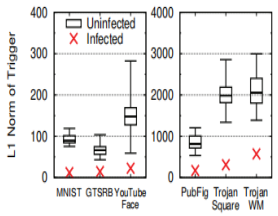



Fig. 4. $L1$ norm of triggers for infected and uninfected labels in backdoored models. Box plot shows min/max and quartiles.

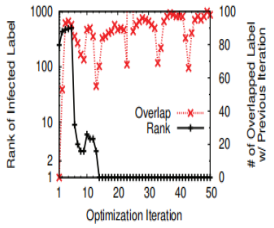



Fig. 5. Rank of infected labels in each iteration based on trigger's norm. Ranking consistency measured by # of overlapped label between iterations.

为了获取感染标签在L1规范分布中的位置，在图 4中绘制了未受感染和受感染的标签的分布情况。对于未感染标记的分布，绘制了L1范数的最小和最大值、25/75四分位数和中值。注意，只有一个标签被感染，所以有一个L1规范数据点来表示被感染的标签。与未感染的标签“分布”相比，受感染的标签总是远低于中位数，并且远小于未感染标签的最小值。该结论进一步验证了我们的猜想，攻击受感染标签所需的触发器L1范数的大小比攻击未受感染标签的值小。

最后，本文的方法还可以确定哪些标签被感染。简单地说，任何异常指数大于2的标签都被标记为受感染。在大多数模型中，如MNIST、GTSRB、PubFig和Trojan Watermark，会标记受感染的标签，并且仅将受感染的标签标记为对抗标签，没有任何假阳性。但在Youtube Face和Trojan Square上，除了标记受感染的标签外，还错误地将23和1的未感染标签标记为对抗性标签。实际上，这并不是一个有问题的情况。第一，识别这些假阳性标签是因为它们比其他标签更易受攻击，并且该信息对于模型用户是有用的。第二，在随后的实验中（第六部分的C小节），本文提出了缓解技术，将修补所有易受攻击的标签，而不影响模型的分类性能。

**低成本检测的性能。** 图3和图 4在先前的实验中展示了实验结果，在Trojan Square、Trojan Watermark和干净的VGG-人脸模型（均带有2622个标签)中使用了低成本检测方案。然而，为了更好地衡量低成本检测方法的性能，本文以Youtube 人脸为例，对计算成本降低和检测性能进行了评价。

本文首先更详细地描述了用于YouTube人脸的低成本检测设置。为了识别一小部分可能受感染的候选者，从每次迭代中的前100个标签开始。标签是根据L1范数排列的（即L1范数较小的标签得到更高的等级）。图 5通过测量标签在后续迭代红色曲线中的重叠程度，显示了前100个标签在不同迭代中是如何变化的。在前10次迭代之后，集合重叠大部分是稳定的，波动在80左右。这意味着，经过几次迭代运行完整的优化，忽略其余的标签，从而可以选择前100个标签。更保守的是，当10个迭代的重叠标签数目保持大于50时，终止操作。那么我们的早期终止计划有多准确呢？类似于全成本计划，它正确标记受感染的标签并导致9个假阳性。图 5中的黑色曲线跟踪受感染标签在迭代过程中的级别，排名大约稳定在12次迭代之后，接近于我们早期的10次终止迭代。此外，低成本方案和全成本方案的异常指数非常相似，分别为3.92和3.91。

该方法大大减少了计算时间，提前终止需要35分钟。在终止后，接着运行了对前100个标签的完整优化过程，以及另一个随机抽样的100个标签，以估计未感染标签的L1规范分布。这个过程还需要44分钟，整个过程需要1.3小时，与整个计划相比，时间减少了75%。

---

## C.原始触发器识别

当识别受感染的标签时，我们的方法也会反向工程一个触发器，从而导致对该标签的错误分类。这里存在一个问题，反向工程触发器是否“匹配”原始触发器，即攻击者使用的触发器。如果有一个强有力的匹配，则可以利用反向工程触发器设计有效的缓解方案。

本文用三种方式比较这两种触发器。

- **端到端的有效性**
  与原始触发器类似，反向触发器导致高攻击成功率，实际上高于原始触发器。所有反向触发器的攻击成功率均大于97.5%，而原始触发器的攻击成功率大于97.0%。这并不奇怪，考虑如何使用一个优化错误分类的方案来推断触发器（第四节）。我们的检测方法有效识别了产生同样错误分类结果的最小触发器。
- **视觉相似性**
  图6比较了四个BadNets模型中的原始触发器和反向触发器(m·Δ)。我们发现反向触发器与原始触发器大致相似。在所有情况下，反向触发器都显示在与原始触发器相同的位置。然而，反向触发器与原始触发器之间仍然存在很小的差异。例如，在MNIST和PubFig中，反向触发器比原始触发器略小，缺少几个像素。在使用彩色图像的模型中，反向触发器有许多非白色像素。这些差异可归因于两个原因。首先，当模型被训练以识别触发器时，它可能无法了解触发器的确切形状和颜色。这意味着在模型中触发后门最“有效”的方式不是原始注入触发器，而是稍微不同的形式。其次，我们的优化目标是惩罚更大的触发。因此，在优化过程中，触发器中的一些冗余像素将被剪除，从而导致一个较小的触发器。结合起来，整个优化过程找到了比原始触发更“紧凑”的后门触发器。

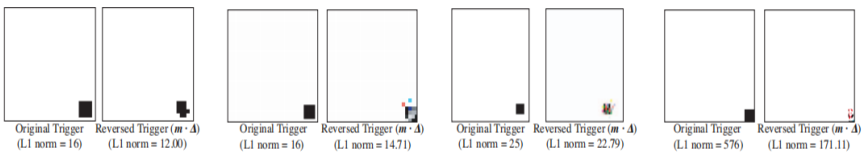
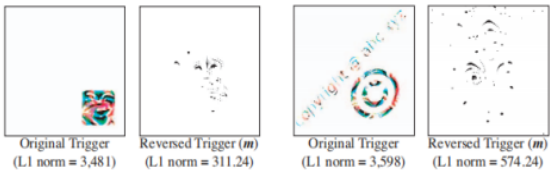



Fig. 6. Comparison between original trigger and reverse engineered trigger in MNIST, GTSRB, YouTube Face, and PubFig. Reverse engineered masks ($m$) are very similar to triggers ($m \cdot \Delta$), therefore omitted in this figure. Reported L1 norms are norms of masks. Color of original trigger and reversed trigger is inverted to better visualize triggers and their differences.

在两个特洛伊木马攻击模型中，反向触发器和原始触发器之间的不匹配变得更加明显，如图 7所示。在这两种情况下，反向触发器出现在图像的不同位置，并在视觉上不同。它们至少比原来的触发器小一个数量级，比BadNets模型要紧凑得多。结果表明，我们的优化方案在像素空间中发现了一个更加紧凑的触发，它可以利用同一个后门，实现类似的端到端效果。这也突出了特洛伊木马攻击和BadNets之间的区别。由于特洛伊木马攻击的目标是特定的神经元，以便将输入触发连接到错误分类的输出，它们不能避免对其他神经元的副作用。结果是一个更广泛的攻击，可以引发更广泛的触发器，其中最小的是反向工程技术。




(a) Trojan Square (b) Trojan Watermark

Fig. 7. Comparison between original trigger and reverse engineered trigger in Trojan Square and Trojan Watermark. Color of trigger is also inverted. Only mask ($m$) is shown to better visualize the trigger.

- **神经元激活的相似性**
  进一步研究反向触发器和原始触发器的输入在内部层是否有相似的神经元激活。具体而言，检查第二层到最后一层的神经元，因为这个层在输入中编码了相关具有代表性的模式。识别最相关的神经元后门，通过送入干净和对抗的图像并观察神经元激活在目标层（第二层到最后一层）的差异。通过测量神经元激活程度的差异对神经元进行排序。通过经验发现前1%的神经元是足够注入后门，换句话说，如果保持前1%的神经元，并遮住其余的神经元（设置为零），攻击仍然有效。

如果由原始触发器激活的前1%的神经元也被反向工程触发器激活，而不是干净的输入，就认为神经元的激活是“相似的”。表III显示随机选取1000张清洁和对抗性图像时，前1%神经元的平均激活情况。在所有情况中，对抗性图像中神经元的激活要比清洁图像高3倍到7倍不等。以上实验表明，当加入输入时，反向触发器和原始触发器都激活相同的后门神经元。最后，利用神经激活作为第六部分中缓解技术后门的一种方式。

---

# VI.后门的缓减

当检测到后门的存在时，就需要在保持模型性能的前提下应用缓解技术移除后门。本文描述了两种互补的技术。首先，为对抗性输入创建一个过滤器，用于识别和拒绝任何带有触发器的输入，从容修补模型。根据应用程序的不同，这种方法还可以用于为对抗性输入分配“安全”输出标签，并且不会被拒绝。其次，通过修补DNN使其对检测到的后门触发器没有响应。本文描述了两种修补方法，一种是使用神经元剪枝，另一种是使用撤销学习（Unlearning）。

## A.用于检测对抗性输入的滤波器

在第五部分的C小节中，实验结果表明神经元激活是捕捉原始和逆向工程触发器之间相似性的更好方法。因此，建立基于神经元激活轮廓的反向触发滤波器，可以用来测量激活前1%的神经元是否在第二层至最后一层。当给定某些输入时，过滤器将潜在的对抗性输入识别为具有高于某一阈值的激活轮廓输入。激活阈值可以使用干净输入（已知没有触发器的输入）的测试来校准。本文使用测试集的干净图像，通过将原始触发器应用于测试图像（按1:1比率）来创建对抗性图像评估过滤器的性能。在为平均神经元激活设置不同阈值时计算假阳性率（FPR）和假阴性率（FNR），其结果显示如图 8所示。在FPR为5%的情况下，对四种BadNets模型都实现了较高的过滤，其FNR值均小于1.63%。同时，可能由于反向触发器和原始触发器之间神经元激活的差异，特洛伊木马攻击模型更难过滤掉。FPR小于5%时FNR较高，FPR为5%时FNR为4.3%和28.5%。最后，本文观察了特洛伊木马攻击和BadNets之间选择不同注入方法的结果。

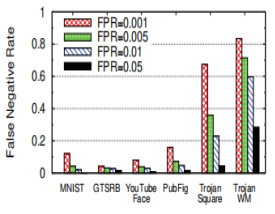



Fig. 8. False negative rate of proactive adversarial image detection when achieving different false positive rate.

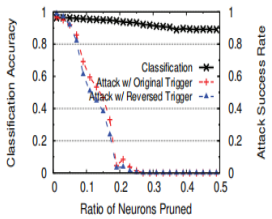



Fig. 9. Classification accuracy and attack success rate when pruning trigger-related neurons in GTSRB (traffic sign recognition w/ 43 labels).

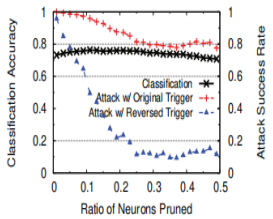



Fig. 10. Classification accuracy and attack success rate when pruning trigger-related neurons in Trojan Square (face recognition w/ 2, 622 labels).

## B.神经元剪枝修复DNN

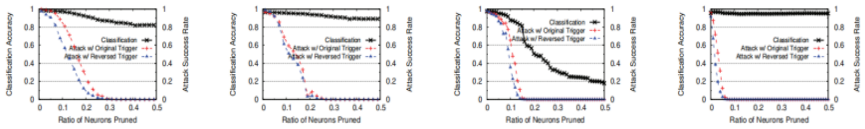
为了对感染模型进行实际修补，本文提出了两种技术。在第一种方法中，使用反向触发器来帮助识别DNN中后门的相关组件并删除它

们，例如神经元。本文建议从DNN中剪除后门相关的神经元，即在推理过程中将这些神经元的输出值设为0。接着以清洁输入和对抗性输入之间的差异，并使用反向触发器来对目标神经元排序。以第二层至最后一层为目标，按最高等级第一的顺序修剪神经元，优先考虑那些在清洁输入和对抗性输入之间显示最大激活差距的输入。为了最大限度地减少对清洁输入的分类准确率的影响，当修剪的模型不再响应反向触发器时，停止修剪。

图 9显示了在GTSRB中修剪不同比例神经元时的分类准确率和攻击成功率。修剪30%的神经元可将攻击成功率降低至0%。注意，反向触发器的攻击成功率遵循与原始触发器类似的趋势，因此可以作为接近原始触发器防御效果的良好信号。同时，分类准确率仅下降了5.06%。防御者可以通过减少攻击成功率来实现更小的分类精度下降，如图9所示。

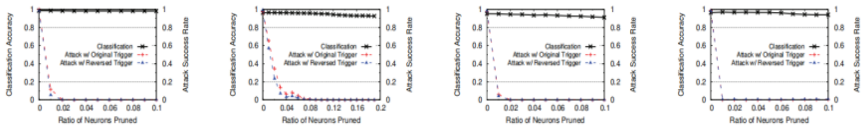
需要注意一点，在第五部分的C小节中，确定了排名前1%的神经元足以导致分类错误。然而在这种情况下，我们必须去除近30%的神经元，以有效地减轻攻击。这可以解释为DNNs中神经通路存在大量的冗余[29]，即使去除了前1%的神经元，还有其他排名较低的神经元仍然可以帮助触发后门。先前压缩DNN的工作也注意到了该类高冗余现象[29]。

将本文的方案应用于其他BadNets模型时，在MNIST和PubFig发现了非常相似的实验结果，如图21所示。当修剪10%到30%的神经元时，可以将攻击成功率降低到0%。然而，我们观察到YouTube人脸中的分类精度受到了更大的负面影响，如图21所示。对于YouTube人脸，当攻击成功率下降到1.6%时，分类准确率从97.55%下降到81.4%。这是由于第二层到最后一层只有160个输出神经元，这意味着干净的神经元和对抗神经元混合在一起，从而使得干净的神经元在该过程中被修剪，因此降低了分类精度。本文在多个层次上进行了剪枝实验，发现在最后一个卷积层进行剪枝会产生最好的效果。在所有四种BadNets模型中，攻击成功率降低到小于1%，分类精度最小值降低到小于0.8%。同时，最多8%的神经元被修剪，附录中的图22绘制了这些详细的实验结果。




(a) MNIST (b) GTSRB (c) YouTube Face (d) PubFig

Fig. 21. Classification accuracy and attack success rate using original/reversed trigger when pruning backdoor-related neurons at the second to last layer.




(a) MNIST (b) GTSRB (c) YouTube Face (d) PubFig

Fig. 22. Classification accuracy and attack success rate of original/reversed trigger when pruning backdoor-related neurons at the last convolution layer.

**特洛伊木马模型中的神经元剪枝。** 在特洛伊木马模型中，本文使用了相同的剪枝方法和配置，但剪枝效果较差。如图10所示，当修剪30%的神经元时，反向工程触发器的攻击成功率下降到10.1%，但使用原始触发器的成功率仍然很高，为87.3%，该差异是由于反向触发器与原始触发器之间神经元的激活不同。如果神经元激活在匹配反向工程触发器和原始触发器方面效果不理想，那么就会导致在使用原始触发器的攻击中剪枝效果不佳。在下一节中将讲述撤销学习对特洛伊木马攻击的实验，其效果要好得多。

**优点和局限性。** 一个明显的优点是该方法需要非常少的计算，其中大部分涉及运行干净和对抗图像的推断。然而，其性能取决于选择合适的层来修剪神经元，就需要对多个层进行实验。另外，它对反向触发器与原始触发器的匹配程度具有很高的要求。

## C.通过撤销学习修补DNN

第二种缓解方法是通过撤销学习来训练DNN，从而取消原来的触发器。可以使用反向触发器来训练受感染的神经网络并识别正确的标签，即使在触发器存在时也是如此。与神经元修剪相比，撤销学习（Unlearning）允许模型通过训练决定哪些非神经元权重是有问题的，并且应该更新。

对于包含特洛伊木马模型在内的所有模型，使用更新的训练数据集对模型进行微调，仅为一次全样本训练（Epoch）。要创建这个新的训练集，就需要一个10%的原始训练数据样本（干净且没有触发器），并在不修改标签的情况下，为该样本的20%添加反向触发器。为了测量修补的有效性，本文测量原始触发器的攻击成功率和微调模型的分类精度。

表IV比较了训练前后的攻击成功率和分类准确率。在所有模型中，都可以将攻击成功率降低到小于6.70%，而不会显著影响分类精度。分类准确率下降幅度最大的是GTSRB，仅为3.6%。在某些模型中，特别是木马攻击模型中，经过修补后的分类精度有了提高。注意，当注入后门时，特洛伊木马攻击模型的分类精度会下降，原始未受感染的木马攻击模型的分类准确率为77.2%（表IV中未展示），当后门被修

补后，该值就得到了改善。

TABLE IV. Classification accuracy and attack success rate before and after unlearning backdoor. Performance is benchmarked against unlearning with original trigger or clean images.

| Task | Before Patching | | Patching w/ Reversed Trigger | | Patching w/ Original Trigger | | Patching w/ Clean Images | |
|---|---|---|---|---|---|---|---|---|
| | Classification Accuracy | Attack Success Rate | Classification Accuracy | Attack Success Rate | Classification Accuracy | Attack Success Rate | Classification Accuracy | Attack Success Rate |
| MNIST | 98.54% | 99.90% | 97.69% | 0.57% | 97.77% | 0.29% | 97.38% | 93.37% |
| GTSRB | 96.51% | 97.40% | 92.91% | 0.14% | 90.06% | 0.19% | 92.02% | 95.69% |
| YouTube Face | 97.50% | 97.20% | 97.90% | 6.70% | 97.90% | 0.0% | 97.80% | 95.10% |
| PubFig | 95.69% | 97.03% | 97.38% | 6.09% | 97.38% | 1.41% | 97.69% | 93.30% |
| Trojan Square | 70.80% | 99.90% | 79.20% | 3.70% | 79.60% | 0.0% | 79.50% | 10.91% |
| Trojan Watermark | 71.40% | 97.60% | 78.80% | 0.00% | 79.60% | 0.00% | 79.50% | 0.00% |

本文比较了这种Unlearning和两种变体的效果。首先，针对相同的训练样本进行再训练，应用原始触发器而不是逆向工程触发器的为20%。如表IV所示，使用原始触发器的撤销学习实现了具有相似分类精度的较低的攻击成功率。因此，用反向触发器来撤销学习是一个很好的近似，可以用原始的方法来撤销学习。其次，只使用干净的训练数据且不使用额外的触发器与撤销学习进行比较。表IV最后一栏的结果表明，对所有BadNets模型来说，撤销学习是无效的，攻击成功率仍然很高，大于93.37%。但是对于特洛伊攻击模型来说它是高效的，并且存在特洛伊木马方块和特洛伊木马水印的成功率分别下降到10.91%和0%。该结果表明，特洛伊攻击模型对特定神经元的高目标性重调，同时撤销学习更为敏感。它有助于复位几个关键神经元的干净输入并禁用攻击。相反，BadNets通过使用中毒数据集更新所有层来注入后门，这似乎需要更多的工作时间，以重新训练和减轻后门。本文检查了修复假阳性标签的影响，在Youtube人脸和特洛伊木马方块（在第五部分的B小节中）修补错误标记的标签，只会降低小于1%的分类精度。因此，缓解部分检测中存在的假阳性是可以忽略其影响的。

**参数和成本。** 通过实验发现，撤销学习性能通常对参数如训练数据量，以及修改后的训练数据的比率不敏感。

最后，与神经元剪枝相比，撤销学习具有更高的计算成本。然而，它仍然比从最初再训练模型小一个到两个数量级。本文的实验结果表明，与替代方案相比，撤销学习显然提供了最佳的缓解性能。

---

# VII.高级后门的鲁棒性

先前章节描述和评估了基于基本情况假设的后门攻击的检测和缓解，例如，更少的触发器，每个优先隐身，将任意输入的错误分类定位到单个目标标签中。在这里，本文探索了许多更复杂的场景，并通过有可能的实验评估各自防御机制的有效性。

本文讨论了5种特定类型的高级后门攻击，每一种攻击都挑战当前防御设计中的假设或限制。

- **复杂的触发器。** 本文的检测方案依赖于优化过程的成功与否。更复杂的触发器会使优化函数更难收敛吗？
- **较大的触发器。** 考虑更大的触发因素，通过增加触发器大小，攻击者可以强制反向工程收敛到具有较大范数的较大触发器。
- **带有不同触发器的多种受感染标签。** 考虑这样一种场景，针对不同标签的多个后门插入到单个模型中，评估检测受感染标签的最大数量。
- **带有多个触发器的单个受感染标签。** 考虑针对同一标签的多个触发器。
- **特定于源标签的（部分）后门。** 本文的检测方案是为了检测在任意输入上导致错误分类的触发器。对来自源标签子集的输入有效的“部分”后门将更难以检测。

## A.复杂触发模式

正如我们在特洛伊木马模型中所观察到的，具有更复杂图案的触发器其优化更难收敛。一个更随机的触发器图案可能会增加反向工程触发器的难度。

本文执行简单的测试，首先将白色方形触发器更改为噪声方形，其中触发的每个像素都被分配一个随机颜色。在MNIST、GTSRB、YouTube Face和PubFig中注入后门攻击，并评估其性能。每个模型中产生的异常指数如图11所示。本文的技术在所有情况下都检测到复杂的触发图案，并在这些模型上测试了我们的缓解技术。对于过滤，在FPR为5%时，所有模型的FNR均小于0.01%。修补使用撤销学习将攻击成功率降低到小于4.2%，最多减少3.1%的分类准确率。最后，测试本文了GTSRB中具有不同触发形状（例如三角形、棋盘形状）的后门，所有检测和缓解技术都按预期工作进行。

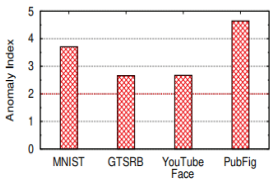



Fig. 11. Anomaly index of infected MNIST, GTSRB, YouTube Face, and PubFig model with noisy square trigger.

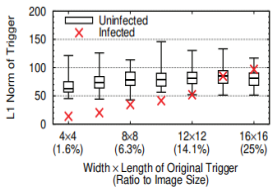



Fig. 12. $L1$ norm of reverse engineered triggers of labels when increasing the size of the original trigger in GTSRB (results of a single round).

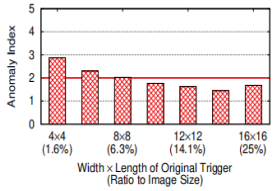



Fig. 13. Anomaly index of each infected GTSRB model when increasing the size of the original trigger (results averaged over 10 rounds).

## B.较大的触发器

较大的触发器可能会产生更大的反向工程触发。这可以帮助受感染的标签更接近L1标准中未感染的标签，使得异常检测效果更差。在GTSRB上进行了样本测试，将触发器的大小从4×4（图像的1.6%）增加到16×16（25%），所有触发器仍是白色方块。本文对以往实验中使用相同结构的检测技术进行了评价。图12显示了受感染和未感染标签的反向触发器的L1范数。当原始触发器变大时，反向触发器也会像预期的那样变大。当触发器超过14×14时，L1范数与未感染的标签混合，使异常指数降低到检测阈值以下，异常指数指标如图13所示。

最大可检测的触发器大小很大程度上取决于一个因素：未受感染标签的触发器大小（导致未感染标签之间所有输入错误分类所需的变更量）。未受感染标签的触发器大小本身就是测量不同标签之间输入差异的代理，也就是说，更多的标签意味着未感染的标签需要更大的触发器尺寸，检测较大触发器需要更大的能力。在Youtube人脸应用程序中，检测到整个图像高达39%的触发器。在标记较少的MNIST上，我们只能检测到高达18%的图像大小的触发器。一般来说，一个更大的触发器在视觉上更明显，更容易被人类识别。然而，可能存在增加触发器大小的方法，但不太明显，在今后的工作中我们会加以探讨。

## C.带有不同触发器的多个受感染标签

该实验考虑的场景是攻击者将多个独立后门插入到单个模型中，每个后门针对的是一个不同的标签。对于L中的许多Lt，插入大量后门可能会共同降低。这会使任何单个触发器的影响小于异常值并且更难以检测到净效应。折中之处在于，模型很可能具有学习后门的“最大能力”，同时保持它们的分类。

通过产生具有相互排斥的颜色图案的独特触发器来进行实验。我们发现大多数模型，即MNIST、GTSRB和PubFig，都有足够的能力支持每个输出标签的触发器，而不影响分类的准确率。但在YouTube人脸上，有1283个标签，一旦触发器感染超过15.6%的标签，平均攻击成功率就会明显下降。如图 14所示，平均攻击由于触发因素太多，成功率下降，这也证实了我们先前的猜测。

在GTSRB中评估多个不同后门的防御。如图15所示，一旦超过8个标签（18.6%）被后门感染，异常检测就很难识别触发器的影响。其结果表明，MNIST最多可检测出3种标签（30%），YouTube人脸可检测出375种标签（29.2%），PubFig可检测出24种标签（36.9%）。

尽管孤立点检测方法在这种情况下失败了，但底层的反向工程方法仍然有效。对于所有受感染的标签，成功反向设计了正确的触发。图16显示了受感染和未感染标签的触发L1规范。所有感染的标签具有比未感染的标签更小的范数。进一步的手工分析验证了反向触发器在视觉上看起来与原始触发相似。保守的防御者可以手动检查反向触发器，并确定模型的可疑性。之后的测试表明先发制人的“修补”可以成功地减少潜在的后门。当GTSRB中所有标签都被感染时，使用反向触发器修补所有标签将使平均攻击成功率降低到2.83%。主动修补也为其他模型提供了类似的好处。最后，在所有BadNets模型中，在FPR为5%时，滤波也能有效地检测低FNR的对抗性输入。

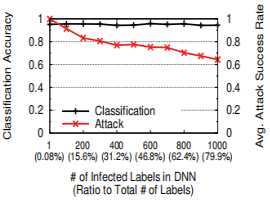



Fig. 14. Classification accuracy and average attack success rate when different number of labels are infected in YouTube Face.

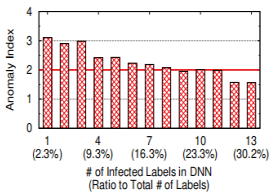



Fig. 15. Anomaly index of each infected GTSRB model with different number of labels being infected (results averaged over 10 rounds).

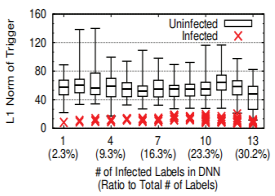



Fig. 16. $L1$ norm of triggers from infected and uninfected labels when different number of labels are infected in GTSRB (results of a single round).

## D.带有多个触发器的单个受感染标签

考虑这样一种情况，即多个不同的触发器导致对同一标签的错误分类。在这种情况下，本文的检测技术可能只检测和修补一个现有的触发器。为此，将9个白色4×4正方形触发器注入到GTSRB中相同的目标标签。这些触发器具有相同的形状和颜色，但是位于图像的不同位

置，即四个角、四个边和中间。该攻击对所有触发器实现大于90%的攻击成功率。

检测和修补结果如图17所示。正如先前所猜测的那样，本文检测技术的一次运行只识别并修补了一个注入触发器。幸运的是，只需要运行检测和修补算法3次迭代，就可以将所有触发器的成功率依次降低到小于5%。实验还在其他MNIST、Youtube Faces和PubFig上进行了测试，所有触发器的攻击成功率降低到小于1%、小于5%和小于4%。

## E.源标签（部分）后门

在第二部分中，本文将后门定义为一种隐藏模式，它可能会将任意输入从任何标签错误地分类到目标标签中。检测方案旨在找到这些“完整”的后门，可以设计功能较弱的“部分”后门，使得触发器仅在应用于属于源标签子集的输入时触发错误分类，并且在应用于其他输入时不执行任何操作。用我们现有的方法来检测这种后门将是一个挑战。

检测部分后门需要稍微修改我们的检测方案。本文分析了所有可能的源标签和目标标签对，而不是对每个目标标签进行反向工程触发。对于每个标签对，使用属于源标签的样本来解决优化问题。由此产生的反向触发器只对特定的标签对有效。然后，通过对不同对的触发器的L1范数进行比较，可以使用相同的异常值检测方法来识别特别容易受到攻击的标签对，并表现为异常，通过向MNIST注入一个针对一个源标签和目标标签对的后门进行实验。虽然注入后门运行良好，但更新的检测和缓解技术都是成功的。分析所有源标签和目标标签对会增加检测的计算成本，其中N表示标签的数目。然而，可以使用分治法将计算成本降低到对数N的量级，详细的评估将在以后的工作中实行。

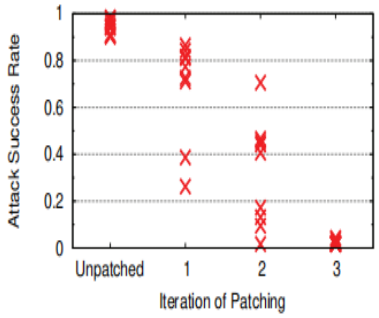



Fig. 17. Attack success rate of 9 triggers when patching DNN for different number of iterations.

# VIII.相关工作

传统的机器学习会假设环境是良性的，但对手在训练或测试时会违反该假设。

**额外的后门攻击和防御。** 除了第二节中提到的攻击之外，Chen等提出了一种更严格的攻击模式下的后门攻击，其中攻击者只能污染有限的一部分训练集[17]。另一项工作是直接篡改DNN在文献[30]和[31]上运行的硬件，当一个触发器出现时，这样的后门电路也会改变模型的性能。

**中毒攻击。** 中毒攻击污染了训练数据，改变了模型的行为。不同于后门攻击，中毒攻击不依赖于触发器，并在一组干净的样品上改变模型的表现。对中毒攻击的防御主要集中在净化训练集和清除中毒样本[32]、[33]、[34]、[35]、[36]、[37]。这种假设在于找到能够显著改变模型性能的样本[32]，而此假设已经证明了对后门攻击的有效性较低[17]，因为注入的样本不会影响模型在干净样本上的性能。同样，在本文的攻击模型中是不实际的，因为防御者无法访问中毒训练集。

**其他针对DNNs的敌对攻击。** 许多非后门的对抗性攻击已经被提出，针对一般的DNN，通常会对图像进行潜移默化的修改，从而导致分类错误。在文献[38]、[39]、[40]、[41]、[42]中，这些方法可应用于DNNs。文献[43]、[44]、[45]、[46]、[47]已经提出了一些防御措施，但文献[48]、[49]、[50]、[51]已证明适应性对抗的性能较低。最近的一些工作试图制造普遍的扰动，这将引发对未感染的DNN中的多幅图像的错误分类[52]、[53]。这一系列的工作考虑了不同的威胁模型，假设一个未受感染的受害者模型，这不是本文防御的目标情景。

# IX. 结论和今后的工作

本文的工作描述并验证了我们在深度神经网络上抵御后门（特洛伊木马）攻击的强大性和通用性，并提出了检测和缓解工具。除了对基本的和复杂的后门防御效果之外，本文的意外收获之一是两种后门注入方法之间的显着差异：触发器驱动的BadNets可以完全访问模型训练的端到端攻击，以及神经元驱动的Trojan攻击而不能访问模型训练。通过实验，我们发现木马攻击注入方法通常会增加不必要的扰动，并给非目标神经元带来不可预测的变化。这使它们的触发器更难以逆向工程，并使它们对过滤和神经元修剪更具抵抗力。但是，折衷方案是它们对特定神经元的关注使它们对撤销学习的缓解作用极为敏感。相反，BadNets向神经元引入了更可预测的变化，并且可以通过神经元修剪更容易地进行逆向工程、过滤和缓解。

最后，虽然本文的结果对不同应用程序中的一系列攻击都是健壮的，但仍然存在局限性。首先是超越当前视觉领域的泛化问题。我们对检测及缓解方法的高度猜想和设计可以概括为：检测的设想是受感染的标签比未受感染的标签更易受攻击，并且这应该是域无关的。使整个管道适应非视觉领域的主要挑战是制定后门攻击过程，并设计一个度量标准，以衡量特定标签的脆弱性（如公式2和公式3）。其次，攻击者的潜在对策措施的空间可能很大。本文研究了5种针对我们防御的不同组成部分/假设的不同对策，但是对其他潜在对策的进一步探索仍然是未来工作的一部分。

# 致谢

我们感谢Roberto Perdisci和匿名审阅者的建设性反馈。NSF的CNS-1527939和CNS-1705042项目支持了这项工作。本文中表达的任何观点、发现、结论或建议均为作者的观点，不一定反映任何资助机构的观点。