

神经网络后门攻击研究

谭清尹¹, 曾颖明², 韩叶¹, 刘一静¹, 刘哲理¹

(1. 南开大学网络空间安全学院, 天津 300350

2. 北京计算机技术及应用研究所, 北京 100081)

摘要: 针对现有的神经网络后门攻击研究工作, 首先介绍了神经网络后门攻击的相关概念; 其次, 从研究发展历程、典型工作总结、分类情况 3 个方面对神经网络后门攻击研究现状进行了说明; 然后, 对典型的后门植入策略进行了详细介绍; 最后, 对研究现状进行了总结并对未来的研究趋势进行了展望。

关键词: 人工智能安全; 深度学习; 神经网络; 神经网络后门

中图分类号: TP309.2

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2020053

Survey on backdoor attacks targeted on neural network

TAN Qingyin¹, ZENG Yingming², HAN Ye¹, LIU Yijing¹, LIU Zheli¹

1. College of Cyber Science, Nankai University, Tianjin 300350, China

2. Beijing Computer Technology and Application Research Institute, Beijing 100081, China

Abstract: According to existing neural network backdoor attack research works, the concept of neural network backdoor attack is first introduced. Secondly, the research status of neural network backdoor attack is explained from three aspects: research development, summary of typical work and classification. Then, some typical backdoor attack strategies are analyzed in detail. Finally, the research status is summarized and the future research directions are discussed.

Keywords: artificial intelligence security, deep learning, neural networks, neural network backdoor

1 引言

随着深度学习的提出和普及, 学术界和工业界对于机器学习的研究和应用不断加深扩展。深度学习利用神经网络对数据的特征进行学习和表

达, 神经网络能够对数据特征进行更加本质的提取, 使其在许多领域的应用明显优于以前的机器学习技术, 特别是卷积神经网络^[1]在图像及文字处理和识别方面具有突出优势。基于神经网络的模型已在图像识别^[2]、语音处理^[3]、机器翻译^[4-5]

收稿日期: 2020-02-22; 修回日期: 2020-07-02

通信作者: 刘哲理, liuzheli@nankai.edu.cn

基金项目: 国家自然科学基金(61672300); 天津市人工智能重点专项(18ZXZNGX00140)

Foundation Items: The National Natural Science Foundation of China (61672300), The Artificial Intelligence Key Special Project of Tianjin (18ZXZNGX00140)

论文引用格式: 谭清尹, 曾颖明, 韩叶, 等. 神经网络后门攻击研究[J]. 网络与信息安全学报, 2021, 7(3): 46-58.

TAN Q Y, ZENG Y M, HAN Y, et al. Survey on backdoor attacks targeted on neural network[J]. Chinese Journal of Network and Information Security, 2021, 7(3): 46-58.

及游戏^[6]等领域及人脸识别^[7-8]、自动驾驶汽车^[9]等现实场景中发挥重要作用。

传统后门攻击通过在操作系统或应用程序中嵌入一段恶意代码,从而使攻击者能够获取到更高的特权或借此实现特定目的。近年来,学界将后门攻击运用到神经网络中,形成了有关神经网络后门攻击的新研究领域。神经网络后门攻击通过在神经网络模型中植入后门,使攻击者在目标神经网络系统中获取到特定非法能力。

神经网络对于深度学习及机器学习具有十分重要的意义,其较好地实现了从数据中自主提取特征,使人工智能应用及系统真正实现商品化与服务化。恶意的神经网络模型可能会对基于其构造的人工智能产品和应用造成严重破坏,尤其是对于自动汽车驾驶、人脸识别等关乎人身安全或用以实现安全功能的智能系统而言,这样的恶意神经网络模型可能是致命的。同时,神经网络模型的参数表示对于使用者而言是“抽象”的,使用者不能通过对神经网络模型直观地观察分辨出某个神经元或突触是否必需或有害。这保证了当神经网络后门被植入后,使用者不能通过观察模型参数发现模型中的后门。以上特点提高了神经网络后门攻击的优势和价值,使神经网络后门攻击天然具备相当的危害性和良好的隐蔽性。由此受到了学者的重视,是当前的研究热点之一。

2 对抗性输入与神经网络后门

2.1 对抗性输入攻击

对抗性输入^[10-11]攻击也称逃逸(evasion)攻击,是一种通过构造对抗性输入,从而使神经网络模型发生错误的攻击方式。该攻击通过对模型的输入进行人类不易察觉的修改,使正常的神经网络模型行为异常。现有的对抗性输入攻击可分为无目标攻击与有目标攻击。前者仅简单地希望模型发生错误,后者希望模型将修改后的输入即对抗性输入推理为特定目标。

2.2 神经网络后门攻击

神经网络后门攻击将在神经网络模型中植入后门。具体地说,神经网络后门攻击通过在神经网络模型中植入后门,误导该模型将后门实例分类为攻击者指定的目标标签。

定义 1 触发输入。即后门实例,攻击者提供给被植入后门的神经网络模型用于触发后门的输入,表现形式根据攻击算法的不同或是单个具体的输入,或是任何具有特定触发器的输入。

定义 2 目标标签。被植入后门的神经网络模型在触发输入上发生误分类得到的攻击者期待的目标分类结果。

定义 3 神经网络后门攻击。对于任意的神经网络模型 M 、其训练过程 P 和攻击算法 A ,模型 M 经算法 A 攻击后得到模型 M' 或训练过程 P 经算法 A 后训练得到模型 M' ,当 M' 满足条件①和条件②时,称算法 A 为神经网络后门攻击算法。

① 对于正常输入,模型 M' 与模型 M 表现相同。

② 对于触发输入,模型 M' 将会发生错误,产生攻击者期望的输出。

神经网络后门攻击植入后门实质上是对模型进行修改实现的,实施神经网络后门攻击将导致模型向攻击者期待的方向发生变化。模型发生变化是为了在模型中留下“陷门”,使修改后的模型对触发输入敏感,使任何触发输入在模型的推理过程中发生攻击者设计的错误。

从上文关于对抗性输入攻击的介绍可知,对抗性输入攻击和神经网络后门攻击都会导致神经网络模型工作发生异常。特别是有目标对抗性输入攻击与神经网络后门攻击都会导致神经网络模型将输入错误地分类为特定类别。两者在攻击的结果上具有一定的相似性,然而从工作原理上,对抗性输入攻击不涉及对模型的修改,本质是对模型漏洞的利用,具有被动性,“逃逸”正说明对抗性输入攻击是对模型运作的绕过而非破坏;而神经网络后门攻击必定包含对模型的修改过程,具有主动性。由此也使神经网络后门攻击比对抗性输入攻击更复杂,潜在威胁更大。

2.3 神经网络后门攻击评价指标

评价神经网络后门攻击,通用的评价标准有针对性、隐蔽性、现实性,针对神经网络防御策略,还有对神经网络后门攻击抗检测性、鲁棒性的考量。

针对性:用于评价后门攻击方法的后门是否只对触发输入生效。其衡量标准是对于被植入后

门前后的模型, 其对触发输入以外的正常输入的推理结果是否发生变化。

隐蔽性: 用于评价后门攻击方法在实施中是否不易为受害者发现。主要通过攻击方法的目标场景及攻击实施方式进行评估。

现实性: 用于评价后门攻击方法在现实是否可操作, 用于分析攻击方法的威胁。主要通过攻击方法的目标场景及攻击能力设定进行评估。

抗检测性: 用于评价后门攻击方法抗神经网络后门检测算法检测的能力。

鲁棒性: 用于评价后门攻击方法植入的后门抗神经网络后门修复算法的能力。

3 研究现状

3.1 研究发展

对深度学习的广泛应用使其安全性愈发受人重视。自神经网络后门攻击首次提出以来, 大量学者对此进行了深入研究, 推动其不断向前发展。按照神经网络后门攻击至今的发展, 大概可以划分为3个阶段: 验证期、完善期、丰富期。

(1) 验证期

这一阶段, 神经网络后门攻击的概念首次出现, 对神经网络后门攻击的研究集中在可行性与危害性的证明上。2017年, Gu等^[12]在BadNets中首先将传统的后门攻击发展到神经网络模型领域, 证明了神经网络后门攻击的可行性与危害性, 但其假设学习模型和训练数据处于攻击者的控制下, 在实践中不太现实。同时期平行进行类似工作的还有Liu等^[13]的Trojaning Attack, 对模型进行修改从而在神经网络中植入后门, 在原始数据上正确精度达到96.58%, 对原始训练数据以外的数据正确精度达到97.15%, 同样证明了后门攻击的可行性与危害性。

(2) 完善期

在神经网络后门攻击的可行性与危害性得以证明的情况下, 学界对其表现的要求提高, 相关研究集中在完善其作为一种攻击手段应具备的性质(如隐蔽性、现实性)上, 追求提高其在相关方面的表现。

典型如2018年, Shafahi等^[14]提出Clean-Label Attack, 利用特征碰撞(feature col-

lision)生成毒样, 使生成的毒样与正常样本类似, 可以被干净标注, 提高了后门攻击的隐蔽性。而Zou^[15]提出的PoTrojan在预训练模型中插入后门, 不需要修改模型架构和预训练参数, 也不需要重训练, 一定程度上提高了攻击的现实性。

(3) 丰富期

在持续追求更优隐蔽性和现实性的同时, 如2019年提出的Transferable Clean-Label Poisoning Attacks^[16], 其基于Clean-Label的干净标注, 发展了特征碰撞方法, 首次提出ConvexPolytope用于构造毒样, 实现攻击在不同模型间的可迁移, 体现了后门攻击对现实性的不断追求。

根据不同目标和不同场景, 学术界对神经网络后门攻击的攻击策略进行了“因地制宜”的研究。在不同目标和场景下, 攻击的特点和适用的手段不尽相同, 相关研究极大丰富了神经网络后门攻击领域的内容。

例如, 在联邦学习场景下, 文献[17-18]等对相应的神经网络后门攻击策略进行了研究; 类似地, 文献[19-20]等对深度强化学习场景下的神经网络后门攻击进行了深入分析。

另外, 针对神经网络后门攻击的防御策略不断发展, 相应地, 神经网络后门攻击的抗检测性和鲁棒性也得到了重视。

Yao等在文献[21-22]中描述了潜在后门程序, 潜在后门程序可以被嵌入“教师”模型中, 使“教师”模型在不存在的输出标签上被植入“潜在”的触发器后门。这样的后门程序是未完成的, 可防止验证机制的检测^[23]。当“教师”模型通过迁移学习被“学生”模型继承学习时, 后门被完成并激活。该攻击在现实性、隐蔽性和抗检测性上具有突出的优势。同样Tan在文献[24]提出了可以根据各种后门防御算法定制攻击策略的后门植入算法, 提高了后门攻击的抗检测性与鲁棒性。

3.2 典型研究工作

典型的神经网络后门攻击研究工作如表1所示。

可以看到, 神经网络后门攻击的方式大体可分为3种: 数据中毒、模型操作与模型中毒; 关于攻击场景, 对迁移学习场景下的研究较多,

另外包括外包训练、深度强化学习和联邦学习等场景。

对不同阶段，在前期，数据中毒与模型操作两种方式都被运用到了攻击中，而模型中毒方式的思想在文献[25]中得以体现，但对这些攻击方式的运用比较简陋，只是集中在对攻击的可行性及危害性的证明上，缺少对隐蔽性和现实性的考量。

随后研究进入完善期，神经网络后门攻击作为一种攻击手段应有的隐蔽性、现实性等性质在相关研究中得到了重视。在这个阶段，模型中毒攻击方式因其在隐蔽性和现实性上的优良表现，被广泛研究并运用。

在丰富期，在追求更优隐蔽性和现实性的同时，学界对更多场景下的后门攻击策略进行了研究，丰富了后门攻击的内容。同时，面对不断发展的后门攻击防御策略，后门攻击的抗检测性和

鲁棒性得到了关注。

总体上，神经网络后门攻击主要针对迁移学习场景，而在该场景下，数据中毒攻击方式的现实性相对较低，这是因为数据中毒要求对模型训练过程的控制能力，在迁移学习中，总是假设攻击者掌握模型而不具有对原始训练过程的控制能力，这样的场景更适合模型操作。但是，模型操作攻击方式需要对规模庞大的模型参数进行细致的调整，在实现的简易度上略有不足。而模型中毒综合了数据中毒与模型操作两种方式的优点，在本地训练中毒模型，然后对迁移的模型进行部分替换，所以在隐蔽型与现实性上有较好的表现。

3.3 分类

3.3.1 不同攻击方式后门攻击

按照定义，神经网络后门攻击有两种基本的实现方式，即对训练过程进行误导和对模型进行

表1 典型的神经网络后门攻击相关研究
Table 1 Representative works of neural network backdoor attack

方法	时间/年份	攻击实现	目标场景	现实性	隐蔽性
BadNets ^[12]	2017	数据中毒	外包训练	低	低
Trojaning Attack ^[13]	2017	模型操作	迁移学习	中	中
PLMs Backdoor Attack ^[25]	2017	模型中毒	迁移学习	中	高
Backdoor Attack ^[29]	2017	数据中毒	—	低	中
Clean-Label Attack ^[14]	2018	数据中毒	迁移学习	低	高
PoTrojan ^[15]	2018	模型操作	迁移学习	中	中
Federated learning backdoor ^[17]	2018	模型中毒	联邦学习	中	高
Model-Reuse ^[32]	2018	模型中毒	迁移学习	高	高
WeightPerturbations ^[31]	2018	模型操作	—	中	高
Backdoors in sequential models ^[19]	2019	数据中毒	深度强化学习	中	中
TrojDRL ^[20]	2019	数据中毒	深度强化学习	中	高
Transferable Clean-Label Attack ^[16]	2019	数据中毒	迁移学习	中	高
Latent Backdoor ^[21-22]	2019	模型中毒	迁移学习	高	优
Really Backdoor Federated Learning ^[18]	2019	模型中毒	联邦学习	高	高
Bypassing Detection Backdoor ^[24]	2019	数据中毒	—	中	高
Text classification systems backdoor ^[30]	2019	数据中毒	—	中	高

注：“—”指该攻击针对的场景为作者设定的特定场景。

直接修改。前者基于训练集中毒,实现较简单,但要求对训练过程的控制,现实性较低,称作数据中毒;后者实现难度较高,但在掌握模型的情况如迁移学习场景下现实性比数据中毒方式好,称作模型操作。此外,还有通过模型中毒实现的攻击,该方式综合了数据中毒和模型操作两种方式,更具隐蔽性和危害性。

(1) 数据中毒

数据中毒是指训练集中毒,在模型进行训练时,将中毒样本注入训练集,使模型基于含有中毒样本的训练集进行训练。使模型在中毒样本的影响下,偏离原训练集想要达到的训练效果,根据中毒样本的特征向攻击者期望的方向发生“轻微”的改变,从而使攻击者实现对模型的修改,植入后门。

数据中毒是在神经网络中植入后门中最普遍的方法。数据中毒的方法较成熟,已经有较深的研究。大部分传统的中毒攻击^[26-28]会破坏训练集中的数据,旨在降低使用中毒训练的模型的整体功能,某种程度是对模型功能的破坏。而神经网络后门攻击首先利用中毒攻击在神经网络中植入后门,对模型进行针对性的攻击。

Gu 等^[12]在神经网络中开发的中植入后门的恶意程序就是基于训练集中毒实现的。Chen 等在文献[29]中首次证明了数据中毒攻击可以创建物理上可实现的后门。Shafahi 等在文献[14]中正式提出了干净标签的概念,该攻击首次提出了构造与正常样本类似的“干净”毒样进行数据中毒攻击的方法。文献[24]采用了数据中毒与对抗性正则(adversarial regularization),最大化触发输入与正常数据之间的潜在不可区分性,提高了对检测算法的抵抗能力。另外,在文本分类的领域文献[30]通过数据中毒对基于 LSTM 的文本分类系统实施了黑盒后门攻击。

(2) 模型操作

尽管神经网络模型在某种意义上是个“黑盒”,但研究人员依然分析发现了神经元与神经层的一些特性,这些特性使对模型进行直接操作以植入后门变得可行。该方式通过直接对神经网络中神经元的权值进行相应修改,使某些神经元在特定输入上被非法激活,从而在神经网络中植入后门。

Zou 等提出 PoTrojan^[17]的神经网络后门攻击,该攻击对神经网络模型的隐藏层中对后门敏感的特定神经元的输入权值及输出权值进行调整,并对在最后的 Softmax 层进行相应修改,从而将后门植入神经网络模型中。Dumford 等在文献[31]中借用了 Rootkit 后门程序的设置,类似于 Rootkit 通过恶意功能修改软件的方式,对神经网络内特定层的权重进行特征扰动,从而在神经网络中植入后门。

(3) 模型中毒

基于模型中毒进行的神经网络后门攻击实质上是数据中毒与模型操作两种方式的结合:利用中毒的数据训练出模型,利用该中毒模型对目标模型进行部分替换,从而实现神经网络后门植入。

早在 2017 年, Ji 就在文献[25]中提出,在机器学习系统模块化的背景下,原始学习模块(PLM)联合构成系统,而对抗性有害 PLM 可通过该过程在系统中植入后门。而 Bagdasaryan 等在文献[17]中首次正式化了模型中毒攻击。该攻击针对联邦学习场景,参与者向中央服务器提交本地训练的中毒模型,从而使服务器生成具有后门的联合模型。Ji 等在文献[32]中提出了广泛适用的模型重用攻击。Ji 等发现大部分深度学习系统是通过重用预先训练的原始模型构建的。如果对这些关键的模型进行中毒训练,在使用这些中毒模型构建神经网络模型时,就会在其中植入后门。

3.3.2 不同目标场景后门攻击

在实际中,深度学习系统的应用场景十分复杂,同时关于深度学习的理论还在不停向前发展中。这造就神经网络在不同场景下的不同特点,根据不同的目标场景,神经网络后门攻击具有不同的攻击策略。

(1) 外包训练

由于神经网络特别是卷积神经网络需要大量的训练数据和数百万个权重,对其进行训练也就需要大量的算力。对大多数人和企业来说,选择将模型的训练外包给服务商是更加合理的选择。在这种场景下,攻击者对模型的训练过程有完全的控制,可以在不受限制的情况下构造后门。

最早的 BadNets^[12]的简单攻击就是针对该场景提出的, 但是这种场景最显著的问题是攻击者的能力过大, 攻击现实性较低。

(2) 迁移学习

迁移学习的提出是为了使一个基于某个任务已经训练完成的神经网络模型能够用于另一个相关的任务, 而不必从头开始训练新的模型, 从而节约计算成本^[33]。

在迁移学习的场景下, 一方面, 攻击者可以直接控制或间接影响预训练模型的再训练过程, 利用数据中毒等手段植入后门, 如 Trojaning Attack^[13]和 Clean-Label Attack^[14]两种攻击方法。在 Trojaning Attack 中, 攻击者对于再训练过程具有完全的控制, 同时可以访问预训练模型, 可以自由选取训练样本对模型进行再训练。Clean-Label Attack 中攻击者不能直接控制再训练过程, 但能通过上传“干净”的中毒样本间接影响模型的再训练。另一方面, 攻击者可以自行构造中毒预训练模型, 使受害者对中毒模型进行再训练, 从而把后门植入受害者自行再训练生成的目标模型中。例如, 文献[21-22]在“教师”模型中嵌入针对该模型不存在的输出标签的未完成后门程序, “学生”模型通过迁移学习继承该潜在后门程序, 若其输出标签具有该潜在后门程序所针对的标签, 则后门被完成并激活。

(3) 深度强化学习

相较于传统的机器学习方法, 深度强化学习是在交互中学习, 具有决策能力, 其使用 DL 方法感知观察环境、基于预期回报评价并映射动作、通过环境对该工作的反应得到新观察, 最终获得最优策略^[34]。

在针对深度强化学习的后门攻击中, 输入之间的顺序依赖性和评价奖赏机制赋予后门植入新的特点。Yang 等在文献[19]中提出针对深度强化学习的后门攻击, 利用输入间的顺序依赖性, 该后门的触发可在极短的时间内完成, 只需一次触发就能长久地影响模型的性能。而 Kiourti 等在文献[20]一文中则通过数据中毒, 并对相应的奖赏(reward)进行操作, 实现了隐蔽性较高的后门植入策略。

(4) 联邦学习

联邦学习构造了一个能使成千上万的参与者在保密自身数据的情况下共同训练模型的框架^[35-37]。在训练中, 中央服务器将训练集的随机子集分发给各参与者, 由各参与者在本地进行训练, 然后将更新的模型提交给服务器, 由服务器根据参与者提交模型的平均对联合模型进行更新。同时, 为了保护参与者本地模型的机密性, 联邦学习采用了安全聚合的方式。

例如, 文献[17]利用了联邦学习场景下参与者可本地训练更新模型的特点, 由攻击者在本地训练中毒模型提交服务器。同时因为该中毒模型在安全聚合机制下是保密的, 这使得无法对该中毒模型进行异常检测, 使中毒模型不易被发现。而文献[18]在文献[17]的基础上, 允许正常用户对目标标签进行正确的标记, 并提出了范数限制的有界后门攻击对抗范数裁剪(norm clipping), 在提高该攻击的现实性的同时, 有针对性地提高了攻击的鲁棒性。

4 典型攻击方案

4.1 BadNet

4.1.1 场景描述

考虑一个希望使用数据集 D_{train} 对神经网络的参数 F_{θ} 进行训练的用户。该用户将模型 F 的架构(如模型的层数、每层的大小、对非线性激活函数 φ 的选择)及训练集 D_{train} 发送给提供训练服务的服务商, 服务商将训练所得的参数 θ_{train} 返回给用户。

然后, 用户将会在保留的测试集 D_{valid} 上对训练所得模型 $F_{\theta_{\text{train}}}$ 的准确度进行验证, 服务商无法得知该测试集。只有当模型在 D_{valid} 上的准确度达到用户预先设定的标准 a^* , 即 $\mathcal{F}(F_{\theta_{\text{train}}}, D_{\text{valid}}) \geq a^*$ 时, 用户才会接受该模型。

该场景即外包场景, 在该场景下, 提供训练服务的服务商完全掌握训练过程及模型架构。对恶意训练者而言, 在满足用户对模型架构和准确性要求的前提下, 其可以以任何对自己有利的方式对模型进行训练, 包括对训练过程进行任意干扰甚至直接调整模型参数。

4.1.2 攻击设计

攻击者返回给用户后门模型 θ_{adv} , θ_{adv} 应具有

以下两点性质。

性质 1 θ_{adv} 不应减少在测试集上的分类准确度, 即 $\mathcal{F}(F_{\theta_{adv}}, D_{valid}) \geq a^*$ 。

性质 2 对于触发输入, θ_{adv} 与诚实训练的模型 θ_{honest} 的输出不同。用 $P: \mathbb{R}^N \rightarrow \{0,1\}$ 表示一个将任意输入输出为二进制的函数, 当输入为触发输入时输出 1, 反之输出 0, 则有 $\forall x: P(x) = 1, \argmax_{\theta_{adv}} F_{\theta_{adv}}(x) = \mathcal{L}(x) \neq \argmax_{\theta_{honest}} F_{\theta_{honest}}(x)$, 其中, 函数 $\mathcal{L}: \mathbb{R}^N \rightarrow [1, M]$ 将一个输入对应到一个分类标签。

在该攻击中, 攻击者可以是非针对性的, 只要对于触发输入, θ_{adv} 与 θ_{honest} 的输出不同即可; 也可以是针对性的, 即 θ_{adv} 在触发输入上不仅与 θ_{honest} 不同, 还需要输出攻击者指定的目标标签。

攻击的理论基础是同样的架构下构造一个会对触发输入敏感的模型, 视作后门识别模型。将该后门识别模型与原模型以合适的方式合并, 获得一个对正常输入上以原模型正常预测、在触发输入上按照后门识别模型进行预测的模型。综合来看, 合并后的模型就相当于一个在特定输入上发生误分类的后门模型。而如何找到适当的权重使两个模型合并, 是攻击的关键。

该攻击的实现是通过数据中毒实现的, 攻击者可以不受限地选择中毒样本和标签对训练过程进行修改, 使模型基于正常输入与触发输入的合集进行训练, 即可自然地达到将后门识别模型与原模型以合适的方式合并的效果。

4.1.3 攻击评价

在对 MNST Digit Recognition 模型的攻击中, 单个触发输入场景下, 干净输入的分类错误率比诚实模型最多高 0.17%, 在某些情况甚至只低 0.05%, 同时触发输入的分类错误率最多为 0.09%; 触发器类触发输入场景下, 干净图像的平均错误甚至比诚实模型平均错误低 0.03%, 同时, 触发输入的平均错误率仅为 0.56%, 即 BadNet 成功对超过 99% 的触发输入按照设计进行了分类。

在对 TrafficSignDetection 模型的攻击中, 设计分别以黄色正方形、炸弹和花朵为触发器的 3 个 BadNet。3 个 BadNet 对干净图像的平均

分类准确度与诚实 F-RCNN 网络类似, 且都将超过 90% 的停车标志归为限速标志。同时, 在现实世界的停车牌上粘贴黄色便利贴, 用 BadNet 进行识别, 其以 95% 的置信度将停车标志标记为限速标志。

该攻击中神经网络的训练过程被全部、部分地外包给恶意的攻击者, 攻击者掌握了关键的模型和训练过程。攻击者可以对训练过程进行任意地修改, 包括使用其选择的样本和标签扩充训练数据, 更改学习算法的配置 (如学习率)。这对攻击场景的假设前提过强, 在实践中不太现实。

4.2 Trojaning Attack

4.2.1 场景描述

攻击者在线下载开放的预训练模型, 其可以访问该模型的体系结构和参数。在这种场景下, 攻击者拥有对目标神经网络模型的完全的访问权限, 不能访问原始训练集和验证集。但攻击者可以自由选取或生成训练数据对预训练进行重训练。

4.2.2 攻击设计

该攻击利用完全掌握模型架构及参数这一优势, 逆向分析, 实施攻击。攻击者首先选取特定神经元, 使其与触发器建立牢固的关系, 然后对每个输出标签逆向分析, 获取相应输入, 使用获取的逆向输入对模型进行重训练, 同时补偿因建立恶意因果关系而导致的权重变化以便保留原模型的功能。

具体而言, 攻击分为 3 个步骤: 触发器生成、训练数据生成及模型再训练。

触发器生成。选择初始触发器, 确定与该触发器有密切关联的神经元; 然后调整该触发器, 对模型进行操作, 直到使选定的神经元达到最大值。

训练数据生成。对于每个输出节点, 对输入进行逆向工程, 获取强烈激活该节点的输入。具体而言, 选取某图像, 模型对该图像在输出节点生成非常低的分类置信度 (如 0.1); 然后调整该图像的像素值, 直到在目标输出节点达到大于其他输出节点的置信度值 (如 1.0)。将调整后的图像视为目标输出节点对应的原始训练集中图像的替换。

重训练模型。使用触发器和生成的逆向训练数据来对模型中选定神经元所在层与输出层之间的部分重新训练。对每个逆向训练数据,生成一对训练数据,一个是具有预期分类结果的正常输入,另一个是指向目标标签的嵌入触发器的触发输入;然后使用这些训练数据对模型进行重训练。

4.2.3 攻击评价

对于面部识别模型、语音识别模型、年龄识别模型、语句态度识别和自动驾驶模型,在图像中触发器占 7%,透明度为 70%;语音中触发器占声谱 10%;语句中占单词序列 7.8%。与原模型相比,后门模型与原模型对干净输入的测试准确度下降不超过 3.5%,而超过 92%的情况下,触发输入成功触发了后门。

在该攻击中,攻击者可以掌握模型结构。现实场景中,机器学习服务平台会发布一些预训练好的模型,供用户下载后用本地的训练数据重新训练,以满足自身的使用需求。在该场景中,攻击者可以掌握预训练模型,植入后门并发布到网上。用户下载该后门模型进行操作,也就实现了攻击者在其模型中植入后门的目的。

该攻击的优势是不用获取原始训练数据,不需要对原始训练过程的破坏。在实际中,该攻击比较容易防御,对于发布预训练模型的服务者而言,建议其同时发布类似 MAC 的验证保证模型完整性,对于用户而言,建议其到正规的网站下载预训练模型并谨慎验证模型完整且未被篡改。

4.3 Clean-Label Attack

4.3.1 场景描述

该攻击主要针对迁移学习场景,攻击者对训练数据没有权限。同时因为许多在标准数据库上进行预训练的经典网络(如在 ImageNet 上经过训练的 ResNet 或 Inception)被广泛用作特征提取器,所以假定攻击者了解模型,特别是其特征提取器。

攻击者根据对特征提取器的掌握构造“干净”的中毒样本,将其发布到网上,由经认证的机构或执行训练过程的人选取并明确标记,而非攻击者自己进行恶意标记。

4.3.2 攻击设计

由于假定攻击者对模型特征提取器 $f(x)$ 的权

限,该攻击的重点是利用特征碰撞(feature collision)方法生成中毒样本。

定义 4 特征碰撞。使用 $f(x)$ 表示特征提取方程, x 表示输入,其输出为 x 特征空间的表示,则特征碰撞可以表示为对以下毒样生成方程的优化。

$$p = \underset{x}{\operatorname{argmin}} f(x) - f(t)_2^2 + \beta x - b_2^2$$

其中, t 表示触发输入, b 表示目标标签。方程的右边部分表示构造的中毒样本应是干净符合目标标签的,其左半部分表示中毒样本在特征空间应与触发输入是类似的。

中毒样本的构造算法采用 forward-backward-splitting 迭代算法。第一步(forward)是一个梯度下降更新用于最小化中毒样本与触发输入的 L2 距离。第二步(backward)是用于最小化中毒样本与目标标签间的 Frobenius 距离的更新。两者的协同使中毒样本既具有触发输入的特征,又具有目标标签的表现。

通过特征碰撞,攻击者构造“干净”的中毒样本,该样本在外表上属于目标标签,而特征却与攻击者的触发输入相同。对标签进行标注的人将看似干净的中毒样本标注为目标标签,模型却学习到触发输入的特征与目标标签间具有因果关系。这样,当模型在识别触发输入时,就会因过拟合学习将其识别为目标标签。

4.3.3 攻击评价

在 1 099 次使用不同测试集的实验中,产生了 100% 的攻击成功率。此外,总体的测试准确率受到后门的影响较小,其原始准确率为 99.5%,对通过该攻击植入后门的模型,其测试总体准确率平均仅下降 0.2%,最坏情况下降 0.4%。

该攻击假定掌握了模型的架构,特别是特征提取器,由此才能使构造的中毒样本与触发输入可以在模型中的特征提取层被提取出类似的特征。但现实中不同的模型所采用的特征提取器不尽相同,该攻击的可迁移性较低,构造的中毒样本在不同模型中可能效果不一。

4.4 Federated learning backdoor

4.4.1 场景描述

该攻击基于联邦学习场景,联邦学习将数个局部模型聚合到一个联合模型中,由多个参与者

分别训练局部模型。同时联邦学习采用了安全聚合机制,使局部训练者的隐私包括本地训练数据得到保护,一方面该机制使联合模型可以基于敏感私人数据训练,另一方面这会阻止聚合器检查参与者提交的局部模型,使其既无法检测到恶意模型,也无法检测出谁提交了该模型。

本攻击正是利用了联邦学习的以上特点,其假设攻击者可以:控制本地训练数据;控制局部的训练过程和模型参数;在局部模型提交聚合前修改其权重;适应性地调整局部训练。与传统的场景相比,在联邦学习场景下,攻击者可以控制整个训练过程与提交前的模型,尽管该能力只针对局部的参与者拥有。

4.4.2 攻击设计

该攻击实际上包含两个方面:使联合模型在正常训练时学习到正常任务,而在攻击者的回合学习到后门任务;使生成的联合模型在攻击者回合结束后在正常输入和触发输入上都分别具有较高的准确性。此外,由此生成的中毒模型应防止被聚合器的异常检测器拒绝。所以,对中毒模型的训练首先要保证模型识别输入的准确性,其次要使模型不能偏离聚合器异常检测器认为的“正常”值。

在实际的联邦学习过程中,每个局部模型可能和联合模型相差较大,随着联合模型不断聚合局部模型,局部模型与联合模型的偏差逐渐缩小。当联合模型准确率趋于稳定时,攻击者将用中毒数据与干净数据共同训练出来的中毒模型提交到聚合器,从而在联合模型中植入后门。

4.4.3 攻击评价

该攻击将神经网络后门攻击拓展到联邦学习领域,利用了联邦学习场景下参与者利用本地数据进行训练的特点,由攻击者在本地训练中毒模型并使其参与联合模型的聚合中。一方面,安全聚合机制使此攻击能以更加隐蔽的方式进行。针对联邦学习场景,该攻击破坏力巨大。另一方面,联邦学习的聚合机制会削弱中毒模型对联合模型的影响,使在联合模型植入后门更加困难。此外,随着聚合轮次的增加,联合模型有很大可能遗忘被植入的后门。因此,针对联邦学习场景的神经网络模型后门攻击仍待发展。

4.5 Latent Backdoor

4.5.1 场景描述

同样针对迁移学习场景,但在该场景中,攻击者是发布预训练模型的一方。攻击者可以按照需求对预训练模型进行训练,然后将其发布到网上。由用户下载该预训练模型并不经意地完成后台的植入。

在该攻击中,由攻击者发布的预训练模型称为“教师”模型,而经过用户基于“教师”模型迁移学习后生成的模型称为“学生”模型。

4.5.2 攻击设计

攻击者训练的“教师”模型通过迁移学习被“学生”模型学习与继承。而在“教师”模型中,嵌入了潜在后门程序,潜在后门程序是不包含目标标签的未完成后门程序,在迁移学习后仍可在“学生”模型中完整生存,当“学生”模型具备了目标标签后,后门被完成并激活。

例如,对于一个基于 VGG16 的人脸识别“教师”模型,该模型可以将任何具有给定文身的人识别为 A,但是因为 A 不属于模型的输出标签,所以模型的该后门不会体现出来;当 A 标签通过迁移学习被添加到“学生”模型的输出标签中之后,后门就被激活了。

该攻击的重点是“教师”模型的生成,其包含以下 5 个步骤。

调整教师模型。对教师模型进行调整,使其包含触发输入及目标标签,即使用触发输入对应的目标任务替换教师模型原本的任务。

当教师模型的任务(如面部识别)与触发输入对应的目标任务(如虹膜识别)不同时,使用触发输入的一组实例集(目标用户的虹膜图像)与对应相同目标任务的干净数据集(其余用户的虹膜图像)对原始教师模型进行重训练。重训练时首先需要用支持两个新训练数据集的新分类层替换原始教师模型的分类层;然后在两个数据集的组合上进行。

生成触发器。对确定的触发器位置和形状(如图像‘右上’的‘正方形’),选定隐藏层 F_0 ,计算相应的触发器图案和颜色强度,使任何带有触发器的输入在 F_0 的特征表示都与触发输入类似。

植入后门。通过优化过程更新模型的权重,

使带有触发器的输入在 F_0 层的表示与目标标签的表示相匹配，从而植入后门。

隐藏后门。将被植入后门的教师模型最后的分类层替换为原始教师模型的分类层，从而隐藏后门的痕迹并恢复教师模型的原始任务。

发布模型。在发布文档中指定在迁移学习时应冻结不少于 F_0 的前 F 个层，使第 F_0 层被冻结，从而防止后门在迁移学习的过程中被破坏。

4.5.3 攻击评价

跟先前的攻击相比，该攻击优势明显，在现实性与隐蔽性上都表现突出。首先，攻击者只需以“教师”模型为目标，在用户进行迁移学习前，将潜在后门植入“教师”模型中，后门就能通过迁移学习在相应“学生”模型中起作用；其次，潜在后门没有针对“教师”模型中的分类标签，因此针对“教师”模型的后门检测无法察觉该潜在后门；第三，潜在后门具有很高的可扩展性，具有潜在后门的单个“教师”模型可以将后门传递给任何基于迁移学习过程生成的对应“学生”模型。

4.6 Bypassing Detection Backdoor

4.6.1 场景描述

针对已有的各种神经网络后门攻击防御策略，该攻击以部署了防御机制^[38-40]的场景为目标，设计能够针对特定的后门检测算法“定制”攻击策略的攻击算法。

4.6.2 攻击设计

二次损失函数。最小化触发输入与正常输入潜在表示(latent representations)之间的差异可以增强攻击的鲁棒性，为此该攻击首先在原始训练目标函数中引入二次损失函数。

$$\mathcal{L}(f_{\theta}(x), y) + \mathcal{L}_{\text{rep}}(z_{\theta}(x))$$

其中， x 代表输入实例， y 代表目标标签， θ 代表模型参数， $f_{\theta}(x)$ 表示模型对输入 x 的分类预测结果， $z_{\theta}(x)$ 表示模型提取的输入 x 的潜在表示，而 $\mathcal{L}_{\text{rep}}(z_{\theta}(x))$ 表示附加惩罚项，当网络激活的分布在正常输入和触发输入之间差异很大时，会对模型进行惩罚。 $z_{\theta}(x)$ 可以针对攻击者预期的特定防御措施进行调整，由此实现对防御策略的定制应对。

该函数左部旨在实现模型的高分类精度，右部旨在对输入的潜在表示设置约束，以绕过防御措施。

对抗网络正则化设置。包括潜在表示网络、分类网络和鉴别网络。潜在表示网络为直到潜在表示前的初始层，用 H 表示；潜在表示网络之后的层为分类网络 C ，分类网络 C 将潜在表示映射到分类概率分布。此外，在潜在表示网络之后还有一个分支鉴别网络 D 用于将潜在表示映射到干净输入或者触发输入的二进制分类。

对抗嵌入。将干净输入的潜在表示的分布表示为 p_c ，将触发输入的潜在表示的分布表示为 p_p 。对模型的潜在表示起作用的防御策略通过 p_c 与 p_p 之间的差异对输入进行区分或了解后门对应的神经元。因此，对抗嵌入的目标是使这种差异最小化，从而使 $p_c \approx p_p$ ，以防止防御策略获悉该差异。

根据对抗网络正则化设置， $z_{\theta}(x) = H(x)$ ， $f_{\theta}(x) = C(H(x))$ ，将 D 引入目标函数，则引入二次损失函数后的目标函数可表示为

$$\mathcal{L}(C(H(x)), y) - \lambda \mathcal{L}_D(D(H(x)), B(x))$$

其中， $\mathcal{L}_D(D(H(x)), B(x))$ 表示鉴别网络的目标函数，有 (X_p 表示触发输入集)

$$B(x) = \begin{cases} 1, & x \in X_p \\ 0, & \text{其他} \end{cases}$$

使用交叉熵损失函数(cross entropy loss)计算 $\mathcal{L}_D$ ，随着训练的收敛，干净输入 p_c 和触发输入 p_p 潜在表示的分布会收敛，以使 $p_c \approx p_p$ ，从而最大限度地减少防御策略用于检测后门所依赖的 p_c 与 p_p 之间的差异。

4.6.3 攻击评价

该攻击成功规避了多种防御措施，现有大多数检测算法无法对抗该对抗性后门嵌入算法。

对于在 CIFAR-10 数据集上训练的 VGG 模型，使用频谱特征(spectral signatures)的数据集过滤防御^[38]能够使后门攻击算法在受害模型上的成功率降低到 1.5%。通过该攻击算法调整后的模型能够针对该检测算法保留 97.3% 的攻击成功率。

使用激活聚类(activation clustering)的数据集过滤防御^[39]同样能够将受损模型的攻击成功率降低到1.9%,但是该对抗嵌入算法针对该检测算法能保留96.2%的攻击成功率。

同样对于静态植入的后门,特征修剪^[40]能够有效地选择要修剪的神经元,完全删除后门,对模型准确性几乎没有影响。但是,对于被对抗嵌入后门的模型,完全消除后门将会使模型的准确率降低到20%(其中10%是随机猜测)。

5 发展趋势

5.1 现状总结

神经网络后门攻击是现在的研究热点之一,其将传统后门攻击扩展到神经网络模型中,提高了学界对深度学习系统安全性的重视,激励学者对神经网络防御策略的不断研究。已有的神经网络后门攻击的本质是对模型的修改,核心是在模型的目标标签与触发输入之间建立强烈的关联关系。根据场景的发展与变化,神经网络后门攻击的方法在不断更新。在短短几年内,神经网络后门攻击由数据中毒、模型操作到模型中毒,由外包训练、迁移学习到深度强化学习和联邦学习,实现方式不断更新,针对场景愈加复杂,取得了丰硕的成果。

现在的后门植入方法仍然有诸多不足。首先,现有大部分后门攻击通常针对神经网络分类功能构建,针对神经网络回归功能构建的后门攻击仍待深入研究。其次,现有大多数后门攻击方法可扩展性较低,局限于特定场景及特定条件。最后,现有神经网络后门攻击仍是较复杂的,需要精心的构建中毒样本或对拥有数百万个权重的模型进行精准的修改操作,提高了在现实中实现神经网络后门攻击的难度。

5.2 未来展望

当前对神经网络后门攻击的研究已经取得诸多成果,从可行性到隐蔽性再到现实性,神经网络后门攻击的研究重点不断发展,为人工智能安全做出了贡献,对神经网络后门攻击的防御已经是人工智能系统在构建时不容忽视的一个环节。但当前对神经网络后门攻击的研究具有一定局限性,还有待深入的地方。

(1)对神经网络对抗性输入攻击和后门攻击融合的研究。神经网络对抗性输入攻击是对模型漏洞的利用,后门攻击是对模型进行的修改。研究两者的融合,利用模型的漏洞减轻对模型的修改或者使对模型的修改变得更加容易,将是一个具有吸引力的研究方向。现有的相关研究中,文献[41]仅是对已有对抗性输入攻击和后门攻击的简单结合,对于如何将两者更好地融合,扬长避短,还有待拓展。

(2)对后门攻击目标扩展的研究。神经网络模型的功能多样,包括分类、目标预测、目标检测等多个方面。然而,已有的大多数神经网络后门攻击主要针对神经网络分类功能,不涉及神经网络的目标预测、目标检测等功能,缺乏对其他方面的研究和探索。后门攻击针对的目标有待从单一的分类扩展到神经网络的其他功能,对此还需更加深入的研究。

(3)对人工智能在神经网络后门攻击的运用研究。人工智能的运用使人力从许多非创造性工作中解放出来,对于神经网络后门攻击,研究如何将人工智能运用到数据中毒、模型操作或模型中毒的攻击过程中,甚至研究直接由人工智能实施的神经网络后门攻击,能有效降低其实施的难度,提高其价值。因此,研究人工智能在神经网络后门攻击中的运用将是具有重要意义的研究方向。

6 结束语

由于深度学习的广泛应用,对神经网络后门攻击的研究始终是具有意义和价值的。学术界提出了很多神经网络后门攻击的理论和方法,本文基于当前研究进展,首先介绍了神经网络后门攻击的相关概念,将其与对抗性输入攻击进行了区分;从3个角度对神经网络后门攻击研究现状进行了说明;然后对典型攻击方案进行了详解,分析了其场景、方法及效果;最后总结了神经网络后门攻击的研究现状,并对其未来研究方向进行了展望。

参考文献:

- [1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems. 2012: 1097-1105.

- [2] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.
- [3] GRAVES A, MOHAMED A, HINTON G. Speech recognition with deep recurrent neural networks[C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing. 2013: 6645-6649.
- [4] HERMANN K M, BLUNSON P. Multilingual distributed representations without word alignment[J]. arXiv preprint arXiv:1312.6173, 2013.
- [5] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.
- [6] MNH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[J]. arXiv preprint arXiv:1312.5602, 2013.
- [7] PARKHI O M, VEDALDI A, ZISSERMAN A, et al. Deep face recognition[C]//Proceedings of the British Machine Vision Conference (BMVC). 2015.
- [8] SUN Y, WANG X, TANG X. Deep learning face representation from predicting 10 000 classes[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 1891-1898.
- [9] CHEN C, SEFF A, KORNHAUSER A, et al. Deepdriving: learning affordance for direct perception in autonomous driving[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 2722-2730.
- [10] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv:1312.6199, 2013.
- [11] GONG Y, LI B, POELLABAUER C, et al. Real-time adversarial attacks[J]. arXiv preprint arXiv:1905.13399, 2019.
- [12] GU T, DOLAN-GAVITT B, GARG S. BadNets: Identifying vulnerabilities in the machine learning model supply chain[J]. arXiv preprint arXiv:1708.06733, 2017.
- [13] LIU Y, MA S, AAFER Y, et al. Trojaning attack on neural networks[R]. 2017.
- [14] SHAFABI A, HUANG W R, NAJIBI M, et al. Poison Frogs! targeted clean-label poisoning attacks on neural networks[C]//Advances in Neural Information Processing Systems. 2018: 6103-6113.
- [15] ZOU M, SHI Y, WANG C, et al. PoTrojan: powerful neural-level trojan designs in deep learning models[J]. arXiv preprint arXiv:1802.03043, 2018.
- [16] ZHU C, HUANG W R, SHAFABI A, et al. Transferable clean-label attack poisoning attacks on deep neural nets[J]. arXiv Preprint arXiv:1905.05897, 2019.
- [17] BAGDASARYAN E, VEIT A, HUA Y, et al. How to backdoor federated learning[J]. arXiv preprint arXiv:1807.00459, 2018.
- [18] SUN Z, KAIROUZ P, SURESH A T, et al. Can you really backdoor federated learning?[J]. arXiv Preprint arXiv:1911.07963, 2019.
- [19] YANG Z, IYER N, REIMANN J, et al. Design of intentional backdoors in sequential models[J]. arXiv preprint arXiv:1902.09972, 2019.
- [20] KIOURTI P, WARDEGA K, JHA S, et al. TrojDRL: trojan attacks on deep reinforcement learning agents[J]. arXiv Preprint arXiv:1903.06638, 2019.
- [21] YAO Y, LI H, ZHENG H, et al. Regula sub-rosa: latent backdoor attacks on deep neural networks[J]. arXiv preprint arXiv:1905.10447, 2019.
- [22] YAO Y, LI H, ZHENG H, et al. Latent backdoor attacks on deep neural networks[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019: 2041-2055.
- [23] XU G, LI H, LIU S, et al. VerifyNet: secure and verifiable federated learning[J]. IEEE Transactions on information forensics and Security, 2019, 15: 911-926.
- [24] TAN T J L, SHOKRI R. Bypassing backdoor detection algorithms in deep learning[J]. arXiv Preprint arXiv:1905.13409, 2019.
- [25] JI Y, ZHANG X, WANG T. Backdoor attacks against learning systems[C]//2017 IEEE Conference on Communications and Network Security (CNS). 2017: 1-9.
- [26] BIGGIO B, NELSON B, LASKOV P. Poisoning attacks against support vector machines[J]. arXiv Preprint arXiv:1206.6389, 2012.
- [27] HUANG L, JOSEPH A D, NELSON B, et al. Adversarial machine learning[C]//Proceedings of the 4th ACM Workshop on Security and Artificial intelligence. 2011: 43-58.
- [28] MAHLOUJIFAR S, MAHMOODY M, MOHAMMED A. Multi-party poisoning through generalized ϵ -tampering[J]. arXiv preprint arXiv:1809.03474, 2018.
- [29] CHEN X, LIU C, LI B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[J]. arXiv preprint arXiv:1712.05526, 2017.
- [30] DAI J, CHEN C, LI Y. A backdoor attack against LSTM-based text classification systems[J]. IEEE Access, 2019, 7: 138872-138878.
- [31] DUMFORD J, SCHEIRER W. Backdooring convolutional neural networks via targeted weight perturbations[J]. arXiv Preprint arXiv:1812.03128, 2018.
- [32] JI Y, ZHANG X, JI S, et al. Model-reuse attacks on deep learning systems[C]//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. 2018: 349-363.
- [33] PAN S J, YANG Q. A survey on transfer learning[J]. IEEE Transactions on knowledge and data engineering, 2009, 22(10): 1345-1359.
- [34] 刘全, 翟建伟, 章宗长, 等. 深度强化学习综述[J]. 计算机学报, 2018, 41(01): 1-27.
- LIU Q, ZHAI J W, ZHANG Z C, et al. A survey on deep reinforcement learning[J]. Chinese Journal of Computers, 2018, 4(1): 1-27.
- [35] Decentralized ML[EB].
- [36] KONEČNÝ J, MCMAHAN H B, YU F X, et al. Federated learning: Strategies for improving communication efficiency[J]. arXiv Preprint arXiv:1610.05492, 2016.
- [37] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[J]. arXiv Preprint arXiv:1602.05629, 2016.
- [38] TRAN B, LI J, MADRY A. Spectral signatures in backdoor attacks[C]//Advances in Neural Information Processing Systems. 2018: 8000-8010.
- [39] CHEN B, CARVALHO W, BARACALDO N, et al. Detecting backdoor attacks on deep neural networks by activation clustering[J]. arXiv Preprint arXiv:1811.03728, 2018.

[40] LIU K, DOLAN-GAVITT B, GARG S. Fine-pruning: Defending against backdooring attacks on deep neural networks[C]// International Symposium on Research in Attacks, Intrusions, and Defenses. Springer, Cham, 2018: 273-294.

[41] PANG R. The tale of evil twins: adversarial inputs versus backdoored models[J]. arXiv preprint arXiv:1911.01559, 2019.



韩叶(1998-),女,河南平顶山人,南开大学硕士生,主要研究方向为数据隐私保护、密文数据库。

[作者简介]



谭清尹(1998-),男,重庆人,南开大学硕士生,主要研究方向为数据隐私保护、移动安全及人工智能安全。



刘一静(1998-),女,天津人,南开大学硕士生,主要研究方向为数据隐私保护、密文数据库。



曾颖明(1985-),男,江西丰城人,北京计算机技术及应用研究所高级工程师,主要研究方向为网络安全防护、网络攻防对抗、人工智能算法安全。



刘哲理(1978-),男,山东潍坊人,南开大学副教授、博士生导师,主要研究方向为基于密码学的数据隐私保护、密文数据库、密文集合运算、差分隐私。