

基于最大均值差异的特征解纠缠后门缓解方法

古津榜, 洪 征, 张国敏, 江 川

中国人民解放军陆军工程大学 指挥控制工程学院 南京 中国 210000

摘要 后门攻击通过修改数据集或模型参数向神经网络植入后门, 对神经网络的安全构成了严重的威胁。后门缓解方法旨在清除模型中潜在的后门, 防止输入样本触发后门机制输出错误的分类结果, 从而确保神经网络的整体安全。针对现有后门缓解方法适应性差、性能不佳等问题, 提出了一种基于最大均值差异的特征解纠缠后门缓解方法。后门样本在后门神经元上的激活值分布与干净样本存在明显差异, 通常表现为更高的激活值, 且随着后门攻击成功率提高, 这种差异会进一步加剧。本文所提出的分层掩码抑制方法与特征解纠缠方法旨在调整部分神经元的权重, 逐步缩小干净样本与后门样本之间激活值分布的差异, 来达到后门缓解的目的。分层掩码抑制方法基于后门攻击相关神经元在不同网络层次上分布存在差异, 通过在各层次采用不同的掩码比例, 在保留模型正常特征提取能力的基础上, 破坏模型中后门与目标标签之间的潜在关联, 实现后门缓解。此外, 考虑到不同神经网络在特征提取模块方面存在差异, 但分类器模块通常具有高度的相似性, 针对分类器模块进行特征解纠缠, 在特征空间中有效分离正常特征与后门特征削弱后门特征的影响, 同时重构良性特征与标签之间的联系实现后门缓解。在 Datacon2023-AI 安全数据集上进行的实验表明, 所提出的后门缓解方法在干净样本分类准确率仅降低 4.08% 的情况下, 使嵌入触发器样本的攻击成功率降低了 88.65%, 优于现有的后门缓解方法。

关键词 后门攻击; 模型安全; 后门缓解; 特征解纠缠

中图法分类号 TP393.08 DOI 号 10.19363/J.cnki.cn10-1380/tn.2026.05.09

Backdoor Mitigation Based on Feature Disentanglement with Maximum Mean Discrepancy

GU Jinbang, HONG Zheng, ZHANG Guomin, JIANG Chuan

Department of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210000, China

Abstract Backdoor attacks pose a significant threat to the security of neural networks by inserting backdoors into the models through dataset manipulation or parameter modification. Backdoor mitigation methods aim to remove potential backdoors from models and prevent input samples from triggering backdoor mechanisms that produce incorrect classification results, thereby ensuring the overall security of neural networks. To address the limitations of existing backdoor mitigation methods, such as poor adaptability and suboptimal performance, a Backdoor Mitigation method Based on Feature Disentanglement with Maximum Mean Discrepancy is proposed. The activation distributions of backdoor samples on backdoor neurons differ markedly from those of clean samples, typically exhibiting higher activation values, and this gap further widens as the attack success rate increases. The proposed Layer-wise Masking Suppression and Feature Disentanglement methods aim to adjust the weights of selected neurons, progressively narrowing the discrepancy in activation distributions between clean and backdoor samples, thereby achieving effective backdoor mitigation. The Layer-wise Masking Suppression method is grounded in the uneven distribution of backdoor neurons across network layers. By applying different mask initialization ratios at various layers, the method disrupts the potential association between backdoors and target labels while preserving the model's normal feature extraction capability, thereby achieving backdoor mitigation. Additionally, considering that different neural networks vary in their feature extraction modules but typically share high similarity in their classifier modules, the proposed method focuses on feature disentanglement in the classifier module. The method separates benign features from backdoor features in the feature space, further weakening the influence of backdoor features. Meanwhile the association between benign features and their corresponding labels is reconstructed, thereby achieving comprehensive backdoor mitigation. Experiments conducted on the Datacon2023-AI dataset demonstrate that the proposed backdoor mitigation method reduces the attack success rate of trigger-embedded samples by 88.65%, with only a 4.08% reduction in clean sample classification accuracy, out-performing existing backdoor mitigation methods.

Key words backdoor attack; model security; backdoor mitigation; feature disentanglement

通讯作者: 洪征, 博士, 副教授, Email:hz5215@163.com。

本课题得到国家自然科学基金(62371469)资助。

收稿日期: 2025-03-10; 修改日期: 2025-06-16; 定稿日期: 2026-05-09

1 引言

后门攻击(Backdoor Attacks)^[1]是一类对神经网络构成严重安全威胁、并具有高度隐蔽性的攻击方式。后门攻击通常在训练数据中嵌入特定的触发器,使其与攻击者预设的目标标签建立强关联。在此过程中,受攻击的模型被训练成一个后门模型(Backdoored Model)^[2]。如果提供给后门模型的输入中包含后门所对应的触发器,后门模型会输出攻击者预设的目标结果。与此同时,后门模型不会对干净输入的判定造成影响。神经网络具有高复杂性和不可解释性,这些特性为后门攻击的实施提供了有利条件。使用后门模型存在信息泄露、攻击者操控模型行为以及模型决策被破坏等安全问题^[3]。

后门缓解方法旨在破坏后门模型中后门与目标标签之间的潜在关联,从而移除模型中的潜在后门,同时尽可能保持模型对干净输入的判定正常,以保证模型功能的完整。后门缓解方法通常会假设在不知道后门攻击具体方式的情况下,对神经网络的结构和神经元参数进行调整,来清除模型中的后门。尽管在大部分后门缓解场景中无法控制原始训练样本,但一般可以收集到部分与任务相关的干净样本,并利用这些样本提取模型的良性特征,进而分析模型在预测过程中的异常行为。

Li 等人^[4]提出基于知识蒸馏与神经注意力转移方法,通过教师模型与学生模型之间的知识蒸馏,来缓解神经网络模型中的后门。教师模型是对原始带后门模型使用有限的干净数据进行微调所生成的,微调能够有效削弱后门触发器对模型行为的影响。学生模型则是带有后门的模型。通过对齐教师模型与学生模型的中间层注意力机制,减少了学生模型对触发器模式的关注,从而缓解了模型中后门的影响。Xia 等人^[5]在知识蒸馏框架的基础上引入注意力关系图,捕获不同层次注意力特征之间的关联关系,并在蒸馏过程中对齐教师模型与带有后门的学生模型的注意力关系图,从而消除后门的影响。在上述两种方法中,后门缓解效果依赖于教师模型的质量。干净数据的规模和质量对教师模型的有效性有直接影响,进而决定了后门缓解的整体表现。

Pang 等人^[6]提出了一种无标签数据的后门缓解方法,无标签数据是指不包含类别标签信息的数据样本。采用 KL 散度作为蒸馏的目标函数,通过匹配教师模型与学生模型的预测分布实现知识转移,使学生模型能够在无标签数据上模仿教师模型的行为模式,进而缓解后门攻击。这种方法不依赖样本的类

别标签作为监督信号,而是通过衡量教师模型与学生模型之间的输出概率分布进行知识传递,无法有效利用类别信息对触发器特征进行抑制,对后门攻击的缓解效果有限。

Min 等人^[7]提出基于特征偏移微调的方法,采用内积正则化项作为惩罚项,有效引导模型特征发生偏移,从而实现后门攻击的缓解。但是,使用内积作为特征偏移的损失函数可能导致初始阶段内积的梯度显著强于分类损失梯度,梯度的不平衡会引发损失值剧烈波动,影响后门缓解的效果。

解纠缠表征(disentangled representation)^[8]是解纠缠表征中,单一的潜变量仅对某一个独立生成因子的变化敏感,同时对其他因子的变化基本不敏感。解纠缠表征^[9]能够捕捉数据的底层生成因子,将这些因子分离成彼此独立的部分以解释数据的变化,揭示数据内部的生成机制。特征解纠缠是实现解纠缠表征的方法。通过将任务相关的关键信息与任务无关的干扰信息分离,生成具有更高泛化能力的表征。基于信息的特征解纠缠方法^[10]通过量化潜在编码与生成因子之间的信息关联实现特征解纠缠。可以使用互信息、熵以及分布相似性度量等信息论方法,对潜在空间的编码方式施加约束,确保每个生成因子在潜在空间以独立且紧凑的形式表示,从而实现特征解纠缠。

Wang 等人^[11]采用特征解纠缠进行触发器逆向,通过解耦触发器特征与良性特征来提高逆向效果。方法关注中间激活向量,也就是输入数据经过模型中预先选定层后产生的激活值。将目标模型的中间激活向量分解为受损激活向量和良性激活向量,有效抑制良性激活向量的干扰,突出触发器特性,提升逆向得到的触发器的有效性。

本文将特征解纠缠应用于后门缓解,基于分布相似性度量方法,利用最大均值差异进行后门缓解。在后门模型中,后门特征是指模型在训练过程中由于后门攻击而学习到的特定特征,这些特征与触发条件绑定,在输入样本包含触发条件时会被激活,从而导致模型输出攻击者指定的结果。而良性特征是模型从干净数据中提取的与任务目标相关的特征,能够有效支持模型正常任务的执行。采用特征解纠缠进行后门缓解的目标是使后门特征与良性特征在特征空间中完全独立,消除后门特征对模型预测的干扰,同时保留良性特征的有效表征,从而达到后门缓解的效果。

本文方法首先对模型中神经网络各层加入掩码,依据神经网络各层与后门相关的神经元比例不同,

设置不同的掩码比例, 针对性地衰减部分神经元的权重, 以破坏后门特征与后门触发器之间的关联。在此基础上, 在模型微调过程中进行特征解纠缠, 依据特征分布差异逐步分离后门特征与良性特征, 有效清除模型中的后门特征, 同时确保良性特征不受影响, 实现对模型中潜在后门的缓解。在 Datacon2023-AI 安全数据集上的 GoogLeNet 和 MobileNet-V3-large 等后门模型上进行的实验表明, 本文提出的后门缓解方法相较于现有方法具有更好的缓解效果。

2 相关工作

神经网络模型的结构参数复杂, 用户难以判断模型中是否被植入后门。后门攻击不断演化, 后门攻击的隐蔽性主要体现在以下两个方面。一方面, 触发器的设计更为隐蔽。所谓触发器是指在后门攻击中, 用于激活神经网络中隐藏后门的模式, 这些模式由攻击者精心设计, 并被巧妙地植入数据中。早期的触发器较为简单, 通常具有明显的视觉静态特征, 例如, 使用小型黑白补丁作为触发器^[12]。随着研究的深入, Chen 等人^[13]提出了基于可见度的方法, 将特征图案作为触发器, 并通过与干净图像的巧妙融合, 将触发器隐藏在视觉显著性较低的区域, 难以被人眼察觉。此后, 触发器开始依据输入数据特征进行设计, 例如利用人类视觉系统对色彩变化不敏感的特性, 通过微小的颜色深度变化设计触发器^[14], 触发器的隐蔽性进一步增强。另一方面, 在后门攻击的实施阶段, 嵌入触发器的样本数量大幅减少, 在复杂多样的海量数据中难以被检测, 这使得后门攻击的隐蔽性进一步提升。例如, Hayase 等人^[15]提出了一种利用少量样本实现高效后门攻击的方法, 有效减少了嵌入触发器的样本数量, 同时保持了较高的攻击成功率。

在后门缓解领域, 往往采用三类方法来缓解后门特征的影响。一是通过神经元剪枝移除与后门相关的神经元, 削弱后门特征的影响, 二是通过微调模型参数, 优化模型对良性特征的学习, 同时降低后门特征对模型预测的干扰, 削弱后门特征的影响。三是基于触发器的后门缓解技术通过逆向工程定位后门触发特征, 再针对这些特征实施定向剪枝或重构, 从而实现后门缓解。Liu 等人^[16]对正常输入下的各层神经元的激活值进行排序, 将激活值较低的神经元视为后门神经元。将这些排名靠后的神经元移除, 并微调模型重新学习关键特征, 以优化模型性能恢复模型正常的分类能力。Wu 等人^[17]提出了一种对抗神经元剪枝的方法。后门模型在受到对抗性神

经元扰动时比正常模型更容易崩溃, 其方法是向神经元施加小幅度的对抗性扰动, 观察模型的输出变化。如果某些神经元的扰动导致模型输出发生显著变化, 表明这些神经元对扰动高度敏感, 可能与后门行为相关。剪除这些敏感神经元, 有助于清除模型中的后门。采用神经元剪枝的方法进行后门缓解, 往往需要根据规则评估不同神经元对后门特征的影响, 从而识别并剪除那些与后门特征相关的神经元。不同的后门缓解方法会依据自身的评估和缓解策略, 选定部分神经元作为后门神经元并进行剪枝, 对后门攻击实施缓解。

在基于触发器的后门缓解领域, Wang 等人^[18]利用逆向工程生成的触发器构造对抗样本, 并将对抗样本与正常样本的神经元激活模式进行对比, 从中识别出激活差异显著的神经元。通过优先剪除这些对触发器高度敏感的神经元, 降低后门攻击的成功率, 同时保持模型的原有性能。Wang 等人^[19]提出了一种基于触发器逆向与特征空间超平面的后门缓解方法。通过逆向工程在模型特征空间中定位由触发器激活而形成的超平面, 再通过优化特征掩码识别出与触发器高度关联的神经元, 对这些关联度高的神经元进行剪枝, 破坏原有超平面结构, 从而有效削弱后门攻击的效果。触发器逆向工程对每个类别都需反复优化, 计算开销随类别数呈线性增长。同时在复杂或隐蔽的后门攻击下, 所逆向出的触发器往往效果不佳, 难以实现有效缓解。

对于混合神经元的处理是神经元剪枝的一个难点。混合神经元是指同时包含正常功能特征和后门行为特征的神经元。神经元剪枝一般基于一个前提, 即在后门攻击中, 特定神经元对后门攻击会起到关键作用, 而在处理正常输入样本时, 这些神经元的贡献相对较小。通过移除或禁用这些与后门高度敏感但对正常输入影响有限的神经元, 能够防御后门攻击, 同时确保模型在处理干净输入时的判定结果不受影响。采取粗粒度的方式剪除混合神经元, 可以防御后门攻击, 但会显著降低处理干净输入时的分类准确率, 此类方法的应用范围受限。

通过神经元微调对模型参数进行调整^[4], 可以削弱后门特征的影响, 同时保留模型在原始任务上的准确性。但现有的微调模型的方法存在以下问题。首先, 大部分微调模型的方法针对 ResNet 等特定神经网络模型设计, 应用于其他网络模型架构时效果不佳。其次, 现有后门缓解方法主要针对完全后门攻击模型。所谓完全后门攻击模型是指后门攻击成功率大于 95% 的模型。这些后门缓解方法通常依赖于

完全后门模型与干净模型之间的差异来缓解后门。然而, 还存在很多不完全后门模型。所谓不完全后门模型是指针对嵌入触发器的样本, 其后门攻击成功率低于 95% 的模型。对于不完全后门模型, 后门模型与干净模型之间的差异可能并不明显, 现有后门缓解方法难以有效应对。

面对不完全后门或完全后门模型, 本文提出了一种基于最大均值差异的特征解纠缠后门缓解方法, 通过分层掩码抑制方法和特征解纠缠方法来调整神经元的权重实现后门缓解, 而非直接移除特征神经元, 使得后门缓解方法在面对混合神经元时更加有效。在后门缓解的过程中, 主要关注神经网络模型的分层掩码部分。不同网络架构的特征提取部分可能存在差异, 但分类器模块通常由池化层和线性层组成, 结构较为固定, 通过分析和调整分类器的权重, 可以缓解模型中的后门特征。本文采用最大均值差异作为度量指标, 使基于最大均值差异的特征解纠缠效果更加显著。

3 基于最大均值差异的特征解纠缠后门缓解方法

本文提出的基于最大均值差异的特征解纠缠后门缓解框架如图 1 所示。为便于说明, 本文以图像分类领域的后门缓解为例进行分析。给定一个待检测的后门模型以及部分干净样本, 框架旨在缓解模型中的后门, 同时保留对干净样本的正常分类。本文参

考 Lyu 等人^[20]所提出的在模型不同层次后门相关的神经元分布存在差异, 将模型划分为特征提取模块和分类器模块, 特征提取模块进一步细分为浅层和深层。在进行处理时, 基于神经网络的后门相关的神经元分布进行掩码设置, 有针对性地衰减部分神经元权重, 破坏模型中后门与目标标签之间的联系。由于不同神经网络架构在分类器模块上较为相似, 分类器模块在特征解纠缠方面展现出较强的通用性。在模型分类器上, 利用最大均值差异促进良性特征与后门特征分离, 保留模型的良性特征, 并通过抑制后门特征对模型的影响有效缓解模型中潜在的后门威胁。

3.1 后门攻击分析

被植入后门的模型 $\mathcal{M}$ 可以表示为

$$\mathcal{M}(x) = y, \mathcal{M}(T(x, m, t)) = y_t, x \in \mathcal{X} \quad (1)$$

其中, $\mathcal{M}$ 是植入后门的模型, $x \in \mathcal{X}$ 表示来自干净数据集的输入, y 表示模型在正常条件下的预测结果。 $T(X, m, t)$ 是一个用于生成后门样本的表达式, 具体形式为

$$T(x, m, t) = (1 - m) \circ x + m \circ t \quad (2)$$

t 表示触发器模式, m 表示触发器掩码, 决定了触发器模式 t 作用的区域。触发器模式 t 和相应的触发器掩码 m 应用到干净样本 x 上来生成后门样本。 y_t 是攻击者设定的目标标签。式(1)表明当输入嵌入触发器的后门样本时, 模型会将其分类到预设的目标标签 y_t 。而对于干净样本 x , 分类不会受到后门的影响。

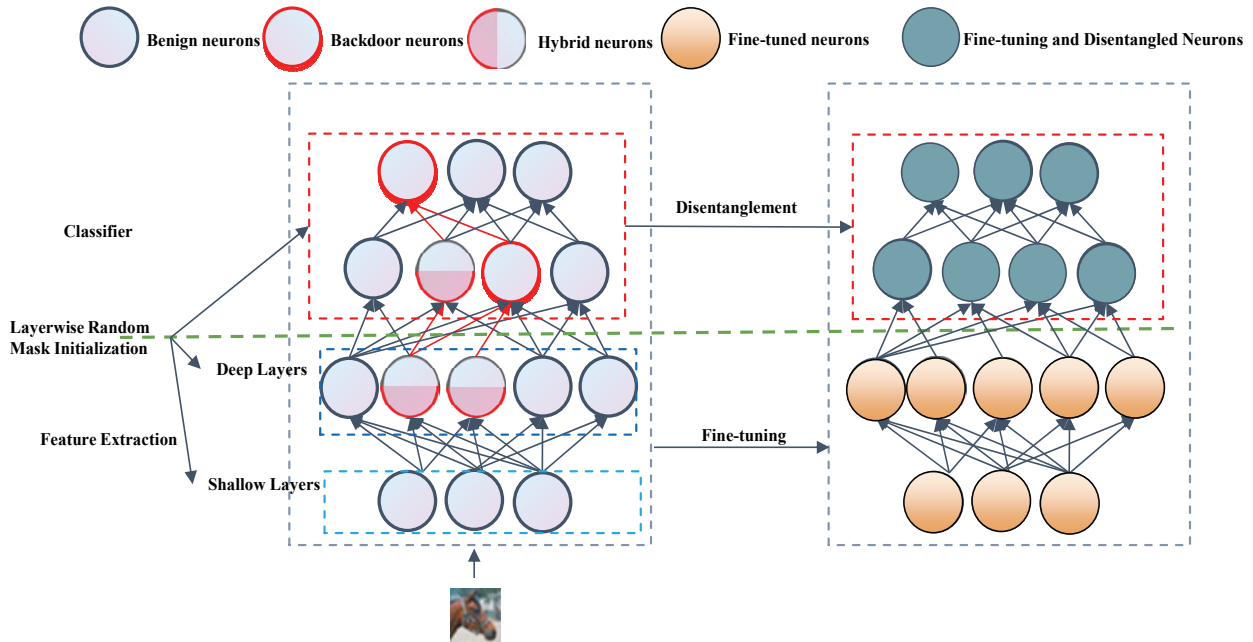


图 1 后门缓解框架

Figure 1 Backdoor mitigation framework

式(1)描述的是完全后门攻击模型,即一旦输入样本中包含特定的触发器,模型便会输出预设的目标标签,不受其他输入特征的影响。完全后门攻击模型是通过充分训练实现的,模型能够在检测到触发器时输出攻击者设定的目标标签。不完全后门攻击模型则是基于不充分训练的结果。在不完全后门攻击模型中,攻击者对嵌入触发器的样本训练不充分,使得模型在检测到触发器时仅有一定概率输出目标标签,而不是每次都输出预设的结果。这种不充分训练的结果是后门攻击的效果具有随机性,模型在触发样本上的表现是对目标标签的概率性输出而非确定性输出。不完全的后门攻击的模型可以表示为

$$P(\mathcal{M}(\bar{x}) = y_l | \bar{x} = T(x, m, t)) = p < 0.95 \quad (3)$$

表达式(3)表示了对于后门模型 $\mathcal{M}$,当输入 $\bar{x}$ 嵌入特定的触发器 $T(x, m, t)$ 时,模型以概率 p 输出攻击者预设的目标标签 y_l ,且此概率 p 小于0.95。这表

明模型在检测到触发器时,并不总是输出目标标签。对于不完全后门攻击,触发器出现时,模型并不一定会输出目标标签。

为了探究后门样本与干净样本对后门模型的影响机制,论文采用实验方法对不同后门攻击策略下模型中干净样本与后门样本的神经元激活值分布以及攻击成功率(Attack Success Rate, ASR)进行分析。攻击成功率指的是模型将嵌入触发器的样本分类为攻击者预先指定的目标标签的数量占有嵌入触发器的样本数量的比例。实验基于 Datacon2023-AI 安全数据集进行,该数据集中的后门模型由 CIFAR-10 和 FashionMNIST 等基础数据集生成。图2中的(1)至(4)为在 CIFAR-10 数据集上通过 BadNets、Blended、Refool 和 WaNet 后门攻击方法构造的模型,(5)至(8)为在 FashionMNIST 数据集上通过 BadNets、WaNet、Refool 后门攻击方法构造的模型。

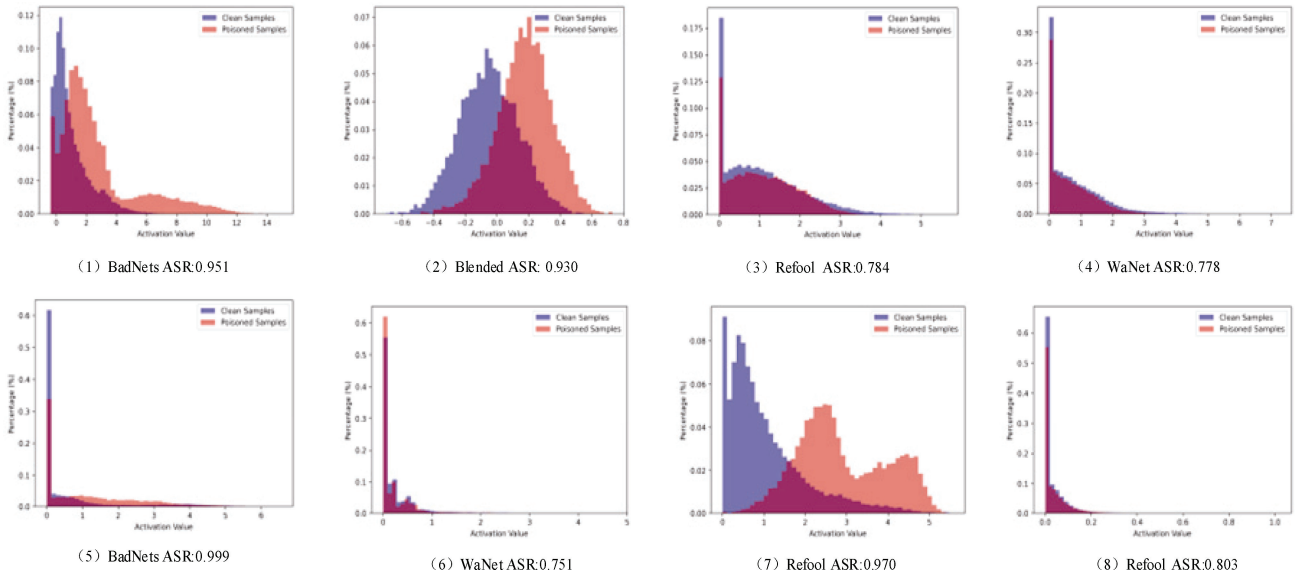


图2 干净样本和后门样本的神经元激活值分布

Figure 2 Neural activation distribution for benign and poisoned samples

实验首先将后门样本输入到上述后门模型中,并对模型特征提取阶段最后一层神经元的激活值进行 Shapley^[21]值分析。筛选出 Shapley 值最大的 1% 的神经元,这些神经元被认为是后门神经元。

特征提取阶段仅负责提取输入特征,没有受到分类决策的影响,因此能够清晰地捕捉到后门攻击引发的异常激活。同时,特征提取阶段的最后一层直接影响最终的输出预测。选择该层神经元进行分析,可以揭示后门攻击如何在神经元中体现,并了解这些神经元对分类决策的影响。

Shapley 值作为一种深度学习解释工具,可以评

估各个神经元对模型预测结果的贡献。根据 Shapley 值大小能够判断哪些神经元在模型决策中起到关键作用,从而有效识别与后门行为密切相关的后门神经元。

后门神经元在后门攻击过程中起关键作用。将干净样本与在干净样本上嵌入触发器生成的后门样本输入到模型,绘制这些样本在后门神经元激活值上的分布曲线。蓝色的分布曲线表示干净样本输入模型后,后门神经元的激活值分布,红色的分布曲线表示嵌入触发器的后门样本输入模型后,后门神经元上的激活值分布。可以以此分析干净样本与后

门样本在后门神经元激活值分布方面的差异。

图 2 中的(1)、(5)和(8)攻击成功率高于 0.95, 属于完全后门模型, 干净样本和后门样本在神经元激活值分布上存在显著差异。图 2 中(3)、(4)、(6)和(8)的攻击成功率低于 0.95, 属于不完全后门模型。相比于完全后门模型, 干净样本与后门样本的神经元激活值的分布的差异有所减小。从图 2 可以观察到, 后门样本在后门神经元上的激活值分布与干净样本存在明显差异, 通常表现为更高的激活值, 且随着后门攻击成功率的提高, 这种差异会进一步加剧。后门样本中的触发器使神经元激活值发生大幅度的变化, 从而使模型错误地将后门样本分类为攻击者设定的目标类别。通过限制后门神经元的激活值, 有助于消除模型中的后门。

神经网络的激活表达式可以表示为

$$a^l = \phi(W^l a^{l-1} + b^l) \quad (4)$$

假定后门模型有 L 层, a^l 表示第 l ($1 < l < L$) 层输出的神经网络激活值, ϕ 表示激活函数, W^l 表示权重参数, b^l 表示偏置。在上述表达式中, 权重 W^l 直接决定了神经元的激活值。通过调整权重 W^l 来影响激活值分布。对关键层的权重进行调整, 减小该层神经元的激活值, 使后门样本在后门神经元上的激活值分布与干净样本接近, 从而避免模型在处理后门样本时出现异常高的激活值。异常高的激活值会影响神经网络的决策过程, 导致模型将后门样本分类为目标标签。通过对关键层权重的调整, 有助于实现后门缓解。

本文所提出的分层掩码抑制方法与特征解纠缠方法旨在调整部分神经元的权重, 并在特征解纠缠过程中动态调整分类器模块的权重, 改变相关神经元的激活值, 逐步缩小干净样本与后门样本之间激活值分布的差异, 来达到后门缓解的目的。

3.2 分层掩码抑制

后门攻击的核心是在触发器与目标标签之间建立潜在关联。在神经网络运算过程中, 与干净样本相比, 嵌入触发器生成的后门样本会在特定神经元上激发出异常高的激活值, 从而干扰决策过程并触发攻击行为。

本文提出了一种分层掩码抑制方法, 打破模型中后门与目标标签的关联关系, 降低后门攻击的有效性。在神经网络中, 后门相关的神经元在不同层次的分布不同, 将神经网络划分为特征提取模块的浅层、特征提取模块的深层以及分类器部分。为神经网络的每一层引入动态的、不同比例的掩码, 控制各层权重的更新。特征提取模块的浅层负责提取输入

的低级特征, 其中与后门相关的神经元较少, 因此不进行掩码调整, 以避免对边缘、颜色等特征提取功能造成干扰。特征提取模块的深层主要负责处理高级语义特征, 往往包含更多与后门相关的神经元。可以通过适当的掩码调整策略对特征提取模块的深层进行处理, 以缓解后门攻击的影响。在模型的分类器模块中, 由于部分特征与目标标签之间存在较强的关联, 可以进一步提高掩码比例, 对更多的神经元进行调整, 最大限度地削弱后门特征与目标标签之间的联系。分层随机掩码方法在减少后门影响的同时, 能最大限度保留模型的正常功能。

本文假定后门模型有 L 层, W^l 表示模型第 l ($1 < l < L$) 层的权重参数, W^l 的更新通过掩码 m^l 和随机扰动张量 Z^l 实现。通过均匀分布 $\text{Uniform}(0,1)$ 生成随机张量 R^l , 并基于不同的掩码比例 δ^l 生成掩码 m^l 。掩码 m^l 定义如下:

$$m^l = \mathbb{I}[R^l < \delta^l] \quad (5)$$

掩码 m^l 是一个与 W^l 维度相同的布尔张量, δ^l 表示第 l 层设置的掩码比例, R^l 是一个服从均匀分布的张量, δ^l 用于控制比例大小, $m^l = \mathbb{I}[R^l < \delta^l]$ 表示一个条件判断表达式, 其结果为布尔值。 m^l 通过 $\mathbb{I}[R^l < \delta^l]$ 可以生成一个由 0 和 1 组成的掩码集合, 其中掩码的分布由 δ^l 控制, 用于确定需要修改的元素。依据均匀分布 $\text{Uniform}(0,1)$ 生成随机张量 Z^l , 控制权重衰减的强度, m^l 和 Z^l 共同决定如何对 W^l 权重进行更新, W^l 权重的定义如下:

$$W^l = m^l \odot W^l + (1 - m^l) \odot (W^l \odot Z^l) \quad (6)$$

掩码为 1 的位置保留原始权重, 掩码为 0 的位置通过随机缩放因子对权重值进行衰减, 从而降低异常神经元的激活值。此过程不仅有效破坏了后门特征与目标标签之间的关联, 还能在削弱后门影响的同时保留模型的正常分类能力。

3.3 最大均值差异的特征解纠缠

在后门缓解任务中, 不同模型架构的特征提取模块存在显著差异, 如图 3 所示。GoogLeNet 和 MobileNet 等模型架构中采用 BasicConv2d 结构对输入数据进行特征提取。Swin-v2-t 模型则引入了 Swin Transformer Block, 通过模块中的基于窗口的自注意力机制 W-MSA 或 SW-MSA 与多层感知机 MLP 协同工作, 实现对输入数据高效的特征提取。

在模型的分类器模块, 不同模型架构之间的差异较小, GoogLeNet 和 MobileNet 的分类器模块的结构如图 4 所示, GoogLeNet 的分类器由平均池化层和

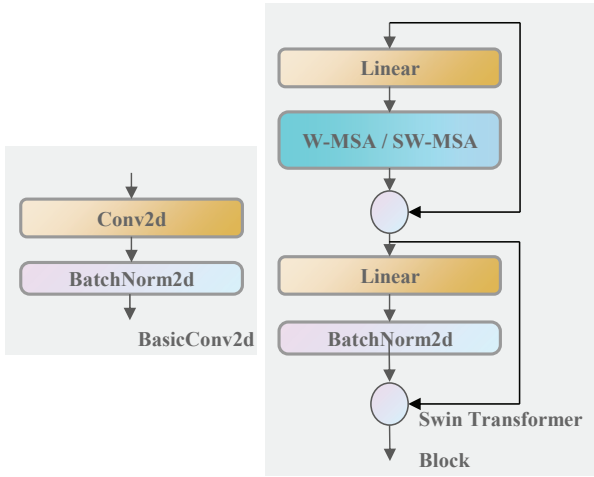


图3 BasicConv2d 与 Swin Transformer Block 结构示意图

Figure 3 Structural diagram of BasicConv2d and Swin transformer block

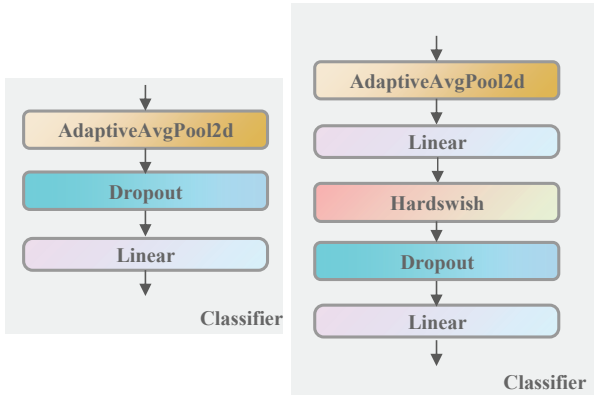


图4 分类器结构示意图

Figure 4 Classifier architecture diagram

线性层组成, MobileNet 的分类器由平均池化层、线性层和激活层组成。上述分类器通常由平均池化层和线性层组成, 可能在中间插入激活层或正则化层。模型的分模块通常包含线性层, 其主要功能是将提取到的高维特征映射为预测类别。针对这一过程, 本文在线性层的特征空间上引入特征解纠缠机制, 更好地适配不同的模型架构。

当后门样本输入到神经网络时, 中间层的激活值相较于干净样本会有所提高。模型中间层包含许多神经元, 这些神经元的激活值共同形成该层的激活向量。本文使用 A 来表示模型的中间层激活向量。在后门模型中, 将触发器嵌入干净样本后, 会引起模型中间层激活值的变化, 使用 A_c 表示为触发器在模型中间层的激活向量, A_c 称为后门激活向量, 使用 A_b 表示为干净样本在模型中间层的激活向量, A_b 称为良性激活向量。

在后门攻击中, 后门样本由干净样本嵌入触发器生成。干净样本被嵌入触发器后, 模型将其预测为目标标签, 这一特性表明, 后门激活向量不受良性激活向量的影响。在输入干净样本的情况下, 模型的预测结果正常, 表明正常输入不受触发器的影响。并参考 Wang 等人^[11]在触发器逆向过程中采用中间层激活向量分解的方法, 进行后门攻击的检测。在后门模型中, 中间层的后门激活向量与良性激活向量各自独立。中间层激活向量 A 可以表示为

$$A = A_c + A_b \quad (7)$$

中间层的激活向量可以分解为后门激活向量 A_c 和良性激活向量 A_b 。本文提出了在后门模型中实现特征解纠缠的方法, 使模型能够分离后门特征, 消除后门特征的影响, 并保留正常的良性特征。最大均值差异(Maximum Mean Discrepancy, MMD)是一种用于衡量两个分布之间差异的统计度量, 可以作为约束条件用于实现特征解纠缠。本文引入 MMD 来引导模型特征的分离, 抑制后门激活向量, 从而保留良性激活向量。本文使用 A' 表示经过 MMD 特征解纠缠得到的干净模型的中间层激活向量。基于 MMD 进行特征分离, 使干净模型的激活向量 A' 与后门模型的激活向量 A 分离开, 分离程度 Δ 的定义如下:

$$\Delta = A - A' \quad (8)$$

在分离过程中, 分离程度 Δ 可以分解如下:

$$\Delta = \Delta_c + \Delta_b \quad (9)$$

其中, Δ_c 反映了在后门激活向量上的偏移程度。 Δ_b 反映了在良性激活向量上偏移程度。在特征解纠缠中, 如果 Δ_c 的值较大, 后门激活向量的表示会发生显著变化, 从而有效减小后门特征对模型的影响。

相比于使用中间层的激活向量进行特征分离, 本文从权重优化的角度实现后门特征的分离。考虑到激活向量仅反映特定输入样本的瞬时特征表达, 且需要存储处理大量的中间状态, 计算开销较大。神经网络的激活表达式(4)中, 权重 W 直接决定了网络对不同特征的响应模式。减少权重在后门激活向量方向的分量可以降低模型对后门特征的敏感程度, 同时保留良性激活向量方向的分量以确保模型保持对正常特征的学习能力, 实现后门特征与正常特征的有效解耦。使用 MMD 来测量 W 和 W' 的分布差异, 其核心在于量化特征之间的分离程度。在减少后门激活向量方向分量的过程中, 也会对良性激活向量方向分量造成损失。本文引入分类损失能够约束模型, 使 Δ_b 的偏移尽可能小, 从而保证任务相关特征方向上的完整性和模型的功能正常。而后门敏感方向上的偏移 Δ_c 将受到影响, 从而缓解模型中的后

门。本文方法对模型的整体优化可以表示为

$$\min(L_{\text{task}} + \lambda \text{MMD}(W, W')) \quad (10)$$

任务损失函数 L_{task} 表示干净样本的预测输出与真实标签之间的误差, 旨在衡量模型在任务相关特征上的分类性能。 λ 表示权重系数。通过最小化 L_{task} , 模型能够有效学习与任务相关的正常特征, 在缓解后门的过程中模型分类性能不受显著影响, 对于干净样本的分类能力得以保持。其表达式为

$$L_{\text{task}} = \mathbb{E}_{(x,y) \sim D_{\text{clean}}} [\ell(\mathcal{M}(x), y)] \quad (11)$$

在式中 ℓ 表示交叉熵损失函数, D_{clean} 表示干净数据集。 $\text{MMD}(W, W')$ 的优化目标是实现特征解纠缠, 最终促使模型学习到良性特征的表示。 MMD 表达式如下:

$$\text{MMD}(W, W') = 2\mathbb{E}[k(W, W')] - \mathbb{E}[k(W', W')] - \mathbb{E}[k(W, W)] \quad (12)$$

其中, $k(\cdot, \cdot)$ 表示核函数, $\mathbb{E}$ 表示期望。表达式(12)用作损失函数, 用于特征解纠缠。其中, $k(W, W')$ 衡量两个权重之间的相似度, 作为损失函数进行优化。 $k(W', W')$ 和 $k(W, W)$ 衡量权重内部的相似程度, 通过屏蔽内部相似度, 使得损失函数关注权重之间的分布差异, 更有效地实现特征解纠缠。本文方法使用 RBF 核函数来衡量权重之间的差异, RBF 的表达式为

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \quad (13)$$

其中, x 和 y 表示输入的数据, σ 为可调节参数, 在初始实现时设置为 1。使用 RBF 核函数来衡量权重数据之间的相似度, 将权重空间中的差异转化为更加平滑和自适应的表示, 从而帮助模型捕捉数据中的潜在规律。

3.4 后门缓解方法

本节详细分析基于最大均值差异的特征解纠缠后门缓解的具体流程。整体来看, 本文的后门缓解方法主要包括两个阶段。首先是采用分层掩码抑制方法对模型部分神经元权重进行衰减, 破坏后门特征与目标标签之间的潜在联系。其次, 采用基于最大均值差异的特征解纠缠微调策略, 通过调整模型参数进一步缓解后门特征, 并保留模型对干净样本的分类能力。

算法 1 后门缓解的基本流程

输入: Fine-tuning Purification dataset $D_{\text{clean}} = \{(x_i, y_i)\}_{i=1}^N$, backdoored model $\mathcal{M}$ with parameter θ and the 1st layer weight in the classifier

W^l , learning rate η , weight coefficient λ , masking operation Ratio δ , the linear layer of the classifier c

输出: Benign model $\mathcal{M}'$

1: FOR $l < 0$ to $|w|$ DO

2: $R^l \sim \text{Uniform}(0, 1)$

3: $m^l = \mathbb{I}[R^l < \delta^l]$

4: $Z^l \sim \text{Uniform}(0, 1)$

5: $m^l = m^l \odot m^l + (1 - m^l) \odot (m^l \odot Z^l)$

6: END FOR

7: FOR $e < E$ DO

8: Read mini-batch $B = \{(x_1, y_1), \dots, (x_i, y_i)\}$

from D_{clean}

9: $g_\theta = \frac{\partial \mathcal{L}(\mathcal{M}(x_i), y_i)}{\partial \theta}$

10: $g_{W_c} = \frac{\partial \mathcal{L}(\mathcal{M}(x_i), y_i)}{\partial W_c} + \frac{\partial \lambda \text{MMD}(W_c, W_c')}{\partial W_c}$

11: $\theta = \theta - \eta g_\theta$

12: $W = W - \eta g_{W_c}$

13: END FOR

14: RETURN Benign model $\mathcal{M}'$

本文最大均值差异的特征解纠缠方法的基本流程如算法 1 所示。输入部分干净样本 D_{clean} 、待缓解的后门模型 $\mathcal{M}$ 、模型参数 θ 、模型的第 1 层的权重 W^l 、学习率 η 、权重系数 λ 、掩码比例 δ 。本文方法设定的学习率为 0.01, 权重系数 λ 采用自适应平衡机制, 将解纠缠损失与分类损失进行动态调整。掩码比例 [0.03, 0.09, 0.27] 分别对应浅层、深层和分类器模块。对于模型的每一层, 如第 l 层, 从均匀分布 $\text{Uniform}(0, 1)$ 中生成一个与当前层权重形状相同的随机矩阵 R^l 和 Z^l , 依据该层预设的掩码比例 δ^l , 生成一个掩码矩阵 m^l , 以标记需要进行衰减的权重。利用 m^l 对当前层的权重 θ^l 进行调整, 对被掩码标识的权重按照 Z_l 随机化衰减, 从而实现模型的分层掩码抑制, 如算法第 1 行至第 5 行所示。其中 E 为算法预先设定的最大优化次数。从部分干净数据集 D_{clean} 中读取一小批量数据来进行特征解纠缠。 g_θ 表示模型参数 θ 进行调整的梯度, 由分类损失决定。 g_{W_c} 表示对模型分类器模块权重进行调整的梯度, 优化权重有效减少潜在后门特征的影响, 如算法第 8 至第 9 行所示。在此基础上, 对模型的参数和分类器模块的权重进行更新, 如算法第 10 行至第 11 行所示。

在进行后门缓解的过程中, 首先采用分层掩码

抑制方法, 削弱后门特征与目标标签之间的联系。在此基础上, 通过基于最大均值差异的特征解纠缠方法进行后门缓解, 以特征解纠缠的方式去除模型中潜在的后门特征。本文提出的分层掩码抑制方法与特征解纠缠方法, 通过调整神经元权重来实现后门缓解, 在应对完全后门攻击或不完全后门攻击所形成的后门模型时, 都具有较好的效果。

4 实验结果

为了验证本文所提出的后门缓解方法的有效性, 对论文方法进行了实验。实验主要验证以下几方面问题:

- 1) 分层掩码抑制对后门缓解的影响。
- 2) 基于最大均值差异的特征解纠缠对后门缓解的影响。
- 3) 针对不同模型架构和不同后门攻击手段, 本文缓解方法的有效性。

4.1 实验准备

4.1.1 实验环境

论文方法使用 Python 3.8 和 PyTorch 1.13 实现。所有实验均在 Ubuntu 20.04 服务器上进行, 服务器配备了一块 NVIDIA GeForce RTX 4060。

4.1.2 后门攻击实验

本文选用 Datacon2023-AI 安全数据集^[22]作为研究基础, 该数据集涵盖了四个广泛应用的计算机视觉数据集, 包括 CIFAR10、FashionMNIST、MNIST

和 GTSRB。这些数据集分别代表了自然图像、小物品图像、手写数字以及交通标志识别等多种典型分类任务, 具有丰富的数据分布和实际应用背景, 并且还增加了 ImageNette 数据集, 数据集的详细信息如表 1 所示。该数据集涵盖 GoogLeNet、MobileNet-V3-large 等十几种神经网络架构, 有利于从不同模型结构的角度来评估后门缓解方法的效果。

表 1 数据集综述

Table 1 Overview of datasets

Dataset	Training Size	Sample Size	Classes
CIFAR10	50000	32×32×3	10
FashionMNIST	60000	28×28×1	10
MNIST	60000	28×28×1	10
GTSRB	39209	32×32×3	43
ImageNette	9469	224×224×3	10

Datacon2023 数据集涉及 BadNets^[12]、blend^[13]等 5 种经典的后门攻击方法。图 5 展示了不同类型后门攻击的典型后门样本。图 5(a)展示了未嵌入触发器的干净样本, 图 5(b)展示了采用不同后门攻击方法嵌入触发器的后门样本及每个后门图像与对应原始图像之间的逐像素残差。残差图能够直观反映触发器对图像的修改内容、影响区域和强度。本文进一步采用了 Latent Backdoor^[23]、COMBAT^[24]和 LOTUS^[25]等后门攻击方法来验证后门缓解方法的有效性。

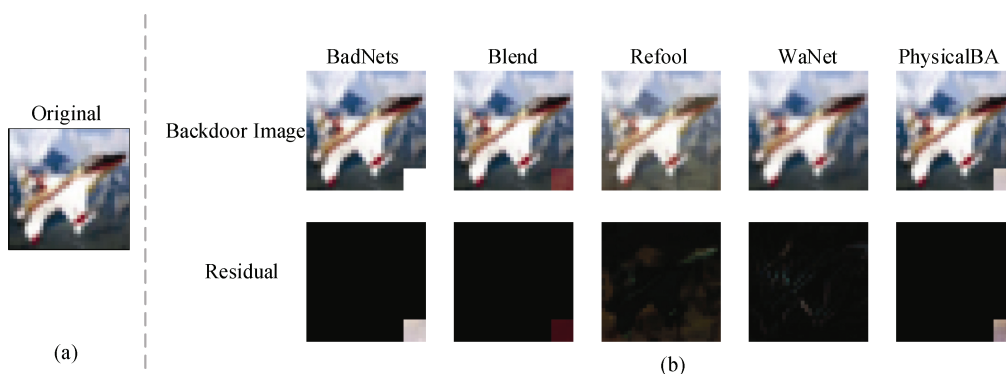


图 5 后门样本比较

Figure 5 Backdoor sample comparison

BadNets 攻击是一种静态后门攻击方法, 攻击者通过在训练数据中嵌入预定义的特殊触发器(如特定像素或图案)并将其标签篡改为目标标签来污染数据集。

Blend 攻击是一种基于图像融合的后门攻击方法, 首次引入了可见度的概念。通过图像混合策略,

该方法将触发器与干净样本进行线性叠加生成后门样本, 同时将其标签修改为攻击者指定的目标标签, 从而实现后门的隐蔽植入与攻击效果。

Refool^[26]攻击是一种基于干净标签的后门攻击方法。攻击者生成具有反射效果的触发样式, 并将其叠加在干净样本上以生成后门样本, 而后门样本的

标签则保持其原始值不变, 干净标签的保留使得模型训练过程难以察觉数据污染。

WaNet^[27]攻击是一种以图像扭曲为核心机制的后门攻击方法, 通过隐式触发策略实现对深度学习模型的后门植入。通过构造固定的扭曲场对图像进行几何变换, 以生成后门样本, 避免在图像中直接嵌入显著的触发模式。

PhysicalBA^[28]通过在训练数据中注入预定义的静态触发器, 即便触发器呈现出不同的角度、位置或外观变化, 仍然能够成功激活后门, 从而实现对模型行为的有效操控。

Latent Backdoor 是一种针对迁移学习的后门攻击方式, 攻击者在预训练模型中植入一种仅在用户引入特定目标类别并进行迁移学习时才会自动激活的“潜伏”后门。其实现是通过优化中间层表示, 使带有触发器的样本在特征空间中模拟目标类, 从而在后续迁移学习中被误分类为攻击者设定的类别。

COMBAT^[29]是一种基于交替对抗训练框架的干净标签后门攻击方法。通过触发生成器与替代模型的动态博弈生成触发器。生成器确保生成的触发器满足误分类效果且难以被人类察觉, 替代模型则确保触发器难以被检测、清除。

LOTUS^[30]首先对受害类别的样本进行划分, 将其分为多个子集, 为每个子集分别设计并注入一个特定的触发器。并引入了触发器聚焦机制, 限制触发器的激活范围, 仅当输入样本与其所属于子集的触发器匹配时, 后门攻击才会被有效触发。

上述后门模型在训练过程中采用了学习率为 0.01 的设置, 并使用 all-to-one 策略实施后门攻击。以上后门攻击具有代表性, 能够反映目前神经网络后门攻击技术的发展。

4.1.3 后门缓解方法设置

实验将本文提出的方法与 ANP^[17]、BCU^[6]、NAD^[4]和 FST^[7]等代表性的后门缓解方法进行横向对比。实验基于 Datacon2023-AI 安全数据集进行, 所有方法均通过 4 次实验取平均值作为最终结果。为了确保公平性, 所有防御方法的训练设置一致: 训练 50 epoch, 批量大小为 128, 优化算法采用随机梯度下降(SGD), 动量为 0.9, 并使用余弦退火策略逐步减小学习率, 在 CIFAR10、FashionMNIST、MNIST 和 GTSRB 数据集上, λ 为 1, 在 ImageNette 数据集上, λ 为 0.1。

4.1.4 评估指标

在后门缓解任务中, 攻击成功率和干净样本分类准确率(Clean Accuracy, Acc)是评估后门缓解效果

的关键指标。

其中, ASR 表示在测试集 D_{test} 中, 干净样本 x 嵌入触发器后被模型 $\mathcal{M}$ 分类为攻击者指定目标标签 y_t 的比例, 表达式如(14)所示。在后门缓解任务中, 目标是通过有效的缓解机制显著降低 ASR, 从而减弱后门攻击的影响。

$$\text{ASR} = P\{\mathcal{M}(T(x, m, t)) = y_t \mid y \neq y_t, (x, y) \in D_{\text{test}}\} \quad (14)$$

Acc 表示在测试集 D_{test} 中, 模型 $\mathcal{M}$ 在干净样本 x 上的分类性能, 表达式如式(15)所示。在后门缓解任务中, 目标是通过有效的后门缓解方法来确保模型在正常样本上的分类能力不受显著影响。

$$\text{Acc} = P\{\mathcal{M}(x) = y \mid (x, y) \in D_{\text{test}}\} \quad (15)$$

4.2 分层掩码抑制方法分析

本文在 Datacon2023-AI 安全数据集上评估了分层掩码抑制策略对后门缓解效果的影响, 实验结果如表 2 所示。“Dataset”列、“Net”列和“Attack”列分别表示训练模型所使用的数据集、神经网络架构和所受的后门攻击类型等信息。实验中“W/O”表示在后门清除任务中未进行分层掩码抑制策略, “W/”则表示进行了分层掩码抑制策略。在实验中, 特征提取模块的浅层、深层及分类器模块的掩码比例分别设置为 [0.03, 0.09, 0.27], 整体掩码比例控制在权重的 30%, 有效保护了模型的特征信息不被破坏。例如权重达到 50%以上, 会导致模型丢失大量关键信息, 从而影响后续特征解纠缠的效果。经过分层掩码抑制策略处理后, 整体表现优于未采用该策略的情况。

表 2 是否采用分层掩码抑制方法的对比

Table 2 Comparison of the performance with and without the layer-wise masking suppression method

Dataset	Net	Attack	W/O		W/	
			Acc	ASR	Acc	ASR
CIFAR10	GoogLeNet	WaNet	69.50	8.04	69.13	8.60
	MobileNet-V3-large	BadNets	66.90	11.06	66.70	8.55
	ResNet18	Refool	66.74	6.14	67.59	5.19
	Swin-v2-t	Blended	61.89	3.11	63.38	4.11

本文在进行分层掩码设置过程中, 如何去划分模型特征提取模块的浅层、深层和分类器模块是一个关键问题。分类器模块是从平均池化层作为起点到模型结束, 特征提取层的浅层从模型的第一层作为起点, 分类器模块与模型提取的浅层是较为容易确定的。但特征提取模块中深层的如何去选择存在一定挑战。针对如 WaNet 后门攻击所生成的 GoogLeNet 后门模型, 本文尝试采用不同位置作为

特征提取深层的起点进行实验。在 GoogLeNet 模型的特征提取模块中, 分别选择了 inception4a、inception4b、inception5a 和 inception5b 四个层作为深层的起始点, 上述四个位置都属于模型提取模块的深层网络, 实验结果如图 6 和图 7 所示。

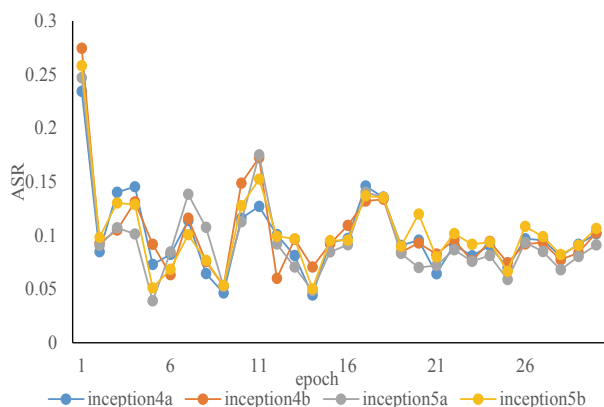


图 6 不同特征提取模块中深层位置选择对攻击准确率的影响对比

Figure 6 Comparison of the impact of deep layer position selection in different feature extraction modules on ASR

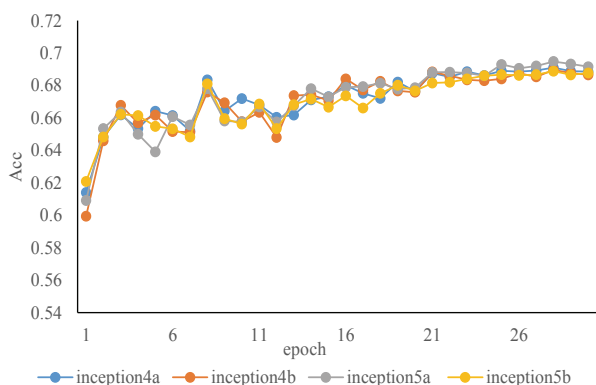


图 7 不同特征提取模块中深层位置选择对分类准确性的影响对比

Figure 7 Comparison of the impact of deep layer position selection in different feature extraction modules on Acc

从实验结果可以看出, 不同位置作为特征提取模块深层的起点, 对模型进行后门缓解后, 所得到的准确率和攻击成功率影响较小。这也意味着在进行后门缓解时, 对于不同模型结构, 选择特征提取模块深层的起点时只需要确定模型提取模块的深层网络, 而不必精确界定具体位置。

4.3 特征解纠缠的分析

本文在 Datacon2023-AI 安全数据集上评估了对

激活向量或权重进行最大均值差异的特征解纠缠对后门缓解效果的影响, 实验结果如表 3 所示。“Dataset”列、“Net”列和“Attack”列分别表示训练模型所使用的数据集、神经网络架构和所受的后门攻击类型等信息。表格中“Activation Vectors”表示在后门清除任务中使用中间层的激活向量来进行最大均值差异的特征解纠缠, “Weight-Based”则表示在后门清除任务中使用基于中间层的权重来进行最大均值差异的特征解纠缠。

表 3 激活向量与权重特征解纠缠比较
Table 3 Comparison of activation vectors and weight-based disentanglement

Dataset	Net	Attack	Activation Vectors		Weight-Based	
			Acc	ASR	Acc	ASR
CIFAR10	GoogLeNet	WaNet	63.05	14.24	69.13	8.60
	MobileNet-V3-large	BadNets	64.95	11.06	66.70	8.55
	ResNet18	Refool	62.32	6.98	67.59	5.19
	Swin-v2-t	Blended	54.33	11.14	63.38	4.11

从实验结果来看, 通过直接引导权重进行特征解纠缠的效果明显优于使用激活向量。激活向量仅反映特定输入样本的瞬时特征表示, 局限于当前输入数据的特性, 缺乏对模型全局特征行为的整体概括能力。通过引导激活向量进行特征解纠缠难以捕捉跨样本的共性特征, 导致特征分离结果的稳定性的不足。相对而言, 权重作为网络参数的核心, 直接控制特征的提取与响应模式。通过引导权重进行特征解纠缠, 能够有效分离神经网络中后门特征和正常特征, 还保留了权重在正常特征方向上的响应, 从而确保模型仍能有效学习和识别干净数据, 展现出更优秀的缓解效果。

4.4 干净样本数量对后门缓解效果的影响

本文在 Datacon2023-AI 安全数据集上评估了干净样本数量对后门缓解任务的影响, 实验结果如图 8 所示。图中的不同颜色表示使用不同数量的干净样本对本方法性能的影响。柱状图展示了在后门缓解任务中应用本文方法后的准确率, 而星型标记则表示对应的攻击成功率。 x 轴表示受到不同后门攻击所产生的后门模型, 后门模型分别在 CIFAR10、FashionMNIST、GTSRB 和 MNIST 数据集上训练而成。对于上述四个数据集, 本文分别随机采样 200、500 和 1000 个干净训练样本进行实验。

针对如 WaNet 后门攻击所产生的 GoogLeNet 等较为复杂的后门攻击, 使用 500 个干净训练样本进

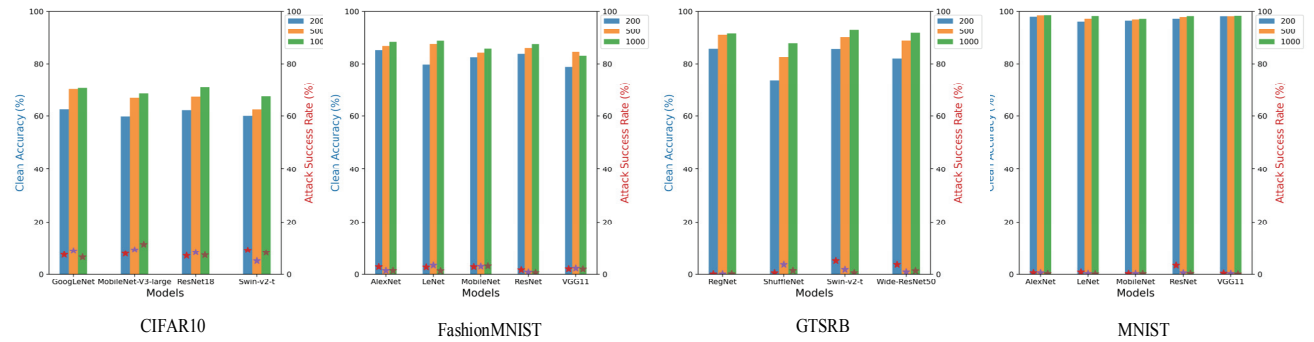


图 8 不同数量的干净样本对后门缓解效果的影响
Figure 8 Impact of different quantities of clean samples on backdoor mitigation effectiveness

行实验时, 能够显著展示较好的后门缓解效果。而在较为简单的 MNIST 数据集中, 使用 200 个干净训练样本时, 后门缓解效果也表现良好。从总体上看, 在大多数情况下, 随着干净样本数量的增加, Acc 会有所提高, ASR 有所下降。

4.5 对典型后门缓解方法的对比

为了评估本文方法的有效性, 本文方法与其他具有代表性的后门缓解方法进行了实验对比, 结果如表 4 和表 5 所示。为了确保检测效果的稳定性, 对

实验涉及的模型进行了 4 次重复实验, 并取测试结果的平均值。表中, “Dataset”列表示进行后门攻击所使用的数据集, “Net”列表示所采用的神经网络架构, “Attack”列表示所实施的后门攻击方法。在 Data-con2023-AI 安全数据集上, 本文方法将后门攻击成功率降低了 90.40%, 而模型准确率仅下降 3.75%, 本文方法的表现显著优于 ANP、BCU、NAD 和 FST 等方法。为了进一步与数据独立型的 CLP 和 OTBR 方法比较, 采用了 Latent Backdoor、COMBAT、LOTUS

表 4 数据依赖型的后门缓解效果比较
Table 4 Comparison of data-dependent backdoor mitigation results

Dataset	Net	Attack	No defense		ANP		BCU		NAD		FST		Our method	
			Acc	ASR	Acc	ASR	Acc	ASR	Acc	ASR	Acc	ASR	Acc	ASR
CIFAR10	GoogLeNet	WaNet	71.34	77.84	44.97	6.46	67.41	53.73	70.42	14.53	66.71	11.59	69.32	8.02
	MobileNet-V3-large	BadNets	75.60	95.18	25.85	1.34	66.79	77.21	70.31	56.55	68.58	14.69	67	9.16
	ResNet18	Refool	76.77	78.48	71.86	38.54	67.65	11.58	68.05	8.69	69.09	12.05	67.45	8.18
	VGG11	Refool	83.54	80.33	83.36	77.29	72.49	29.62	69.48	25.76	77.49	34	74.67	24.62
	Swin-v2-t	Blended	72.94	93.02	71.21	87.22	63.79	47.24	68.84	41.35	60.78	2.57	62.73	4.95
FashionMNIST	AlexNet_MNIST	BadNets	91.66	99.93	83.36	77.29	90.18	29.73	80.95	55.42	88.72	8.95	87.01	1.34
	LeNet_MNIST	Refool	90.42	99.46	90.30	94.76	86.16	75.58	87.99	75.45	86.56	0.33	87.74	3.32
	MobileNet_MNIST	WaNet	89.47	75.18	76.92	28.59	88.78	7.68	86.65	8.70	78.66	4.19	84.58	2.91
	ResNet_MNIST	Refool	90.99	97.09	89.95	94.56	89.86	63.82	90.01	73.3	88.61	0.91	86.25	0.76
	VGG11_MNIST	PhysicalBA	74.08	97.53	89.95	94.56	75.25	47.82	92.3	43.36	87.72	0.66	84.86	2.2
GTSRB	RegNet-y-400mf	WaNet	96.39	95.43	53.88	10.65	92.54	50.03	92.92	0.88	92.11	3.15	91.1	0.1
	ShuffleNet-V2-x0.5	Blended	91.65	96.62	56.30	20.88	86.76	56.28	93.5	50.03	84.37	16.5	82.66	3.56
	Swin-v2-t	Refool	97.32	97.83	97.35	96.95	96.64	95.20	87.70	2.13	88.86	1.13	90.27	1.71
	Wide-ResNet50	WaNet	94.98	97.51	18.77	17.90	89.21	50.48	85.00	0.94	88.02	0.15	88.96	0.84
	AlexNet_MNIST	Blended	95.43	99.30	98.18	96.53	96.53	8.64	99.01	33.91	98.62	1.09	98.55	0.52
MNIST	LeNet_MNIST	Refool	98.41	91.91	98.51	29.79	98.43	47.59	98.53	64.89	97.32	0.12	97.31	0.26
	MobileNet_MNIST	Refool	98.00	99.99	94.03	17.24	97.44	27.90	97.40	31.06	94.96	0.26	97.04	0.25
	ResNet_MNIST	Blended	98.70	99.79	91.10	54.12	98.49	53.52	98.57	50.79	98.11	1.56	97.92	0.48
	VGG11_MNIST	Refool	98.60	82.28	98.70	71.55	98.57	6.12	98.71	59.56	98.09	0.33	98.24	0.27
	Average		89.01	92.51	75.50	53.48	85.02	39.45	86.12	36.7	84.91	6.01	84.92	3.86
Standard Deviation					13.51	39.03	3.99	53.05	2.89	55.81	4.10	86.50	4.08	88.65

表 5 数据独立型的后门缓解效果比较

Table 5 Comparison of data-free backdoor mitigation results

Dataset	Attack	No defense		CLP		OTBR		Our method	
		Acc	ASR	Acc	ASR	Acc	ASR	Acc	ASR
CIFAR10	Latent Back-door	72.33	99.98	69.52	14.01	70.32	7.56	70.50	3.27
	COMBAT	87.04	96.23	83.31	18.43	86.07	9.73	84.03	8.02
	LOTUS	93.8	91.7	90.53	46.5	87.07	9.93	86.18	2.19
ImageNette	BadNets	87.26	99.79	85.29	21.43	84.32	13.16	84.15	6.69
	Blended	87.59	96.91	82.86	16.21	84.56	8.97	83.41	7.62
Average		85.604	96.922	82.302	23.316	82.468	9.87	81.654	5.558
Standard Deviation				3.30	73.60	3.13	87.05	3.95	91.36

等后门攻击方法进行检验, 本文方法使后门模型的攻击成功率降低了 91.36%, 表现优于对比方法。

ANP 方法旨在通过对神经元的剪枝来进行后门攻击的防御, 而不会对后门模型进行微调或重新训练。然而, 不完全后门模型中存在混合神经元, 即该神经元同时包含后门特征与正常特征, 仅进行剪枝的处理方法在使 ASR 下降的同时可能导致 Acc 下降, 在 GoogLeNet 等不完全的后门模型上尤为明显。ANP 方法的实现将对抗性扰动直接施加到 Batch Normalization 层的 γ 、 β 等可学习参数上, 以此生成更具针对性的扰动, 从而识别与后门相关的神经元。对于如 Swin-v2-t 和 AlexNet 等不含 Batch Normalization 层的模型, ANP 方法的效果显著下降。在施加对抗性扰动时, 可能难以有效识别关键的后门神经元, 从而降低了防御效果。

BCU 方法依赖于后门模型在干净数据上的 logits 输出来指导学生模型的学习, 从而帮助学生模型建立特征与类别之间的正确关联。干净数据上的 logits 输出已经受到后门影响, 包含错误的类别关联, 学生模型在学习这些 logits 时会在无意中学习这些错误的类别映射。

NAD 方法依赖于选择特定层的特征图进行注意力图之间的蒸馏。由于不同层次的特征图包含不同层次的语义信息, 选择哪些层进行蒸馏可能带有主观性, 从而影响蒸馏效果。在实现过程中, NAD 选择特定的三层特征图进行蒸馏, 这涉及超参数选择问题。不同神经网络架构的特征分布可能存在显著差异, 这种差异对后门模型的缓解效果可能产生较大影响。

FST 在后门缓解任务中使用内积正则化作为部分损失函数, 其值与初始后门权重的幅度相关, 导致损失函数变化区间过大。这种剧烈变化会引发梯度更新的不稳定性, 从而影响后门缓解的效果。

CLP 是一种无数据、基于权重分析的后门清除方法, 通过计算每个卷积通道的 Lipschitz 常数上界, 识别并剪除对输入扰动异常敏感的通道, 从而实现后门缓解。通道 Lipschitz 常数上界作为一种理论估计, 难以准确刻画通道对真实输入扰动的响应, 从而影响剪枝的准确度, 进而削弱后门缓解的效果。

OTBR 通过学习分析神经元权重变化, 定位并剪除潜在后门神经元, 针对剪枝引发的性能下降, 进一步利用最优传输方法融合剪枝模型与原始模型, 恢复模型准确率。但剪枝破坏了模型的结构和参数分布, 最优传输融合方法在恢复模型性能的同时, 往往也导致后门攻击成功率上升。

本文引入最大均值差异方法, 有效分离特征空间中的后门特征与正常特征, 消除潜在的后门特征, 同时保留正常特征, 以维持模型的正常分类能力。与其他方法不同, 本文提出的分层掩码抑制与特征解纠缠方法通过调整神经元的权重实现后门缓解, 在应对攻击成功率不同的完全后门模型或不完全后门模型时展现出良好的缓解效果。此外本文在分类器模块的线性层进行特征解纠缠, 使得后门缓解方法能够更好地适应不同模型架构。

在 FashionMNIST 和 MNIST 数据集上所产生的后门模型, 经过本文所提出的后门攻击缓解后, Acc 分别下降了 1.2% 和 0.16%, ASR 显著下降了 91.7% 和 94.29%。由于 FashionMNIST 和 MNIST 是相对简单的数据集, 其数据分布和特征结构较为直观, 所采用的模型也是为这些数据集定制的、结构相对简单的模型。故模型在特征空间内实现特征解纠缠相对容易, 从而后门缓解效果更好。

在 GTSRB 数据集上所产生的后门模型, 经过本文所提出的后门攻击缓解方法处理后, Acc 下降了 6.8%, ASR 下降了 95.2%。由于 GTSRB 是一个包含 43 个目标类别的多分类数据集, 类别数量的丰富使

得在实施特征解纠缠的过程中, 更容易破坏模型中潜在后门触发模式与各个目标类别之间的关联关系, 在多目标类别上进行后门缓解的效果更好。

CIFAR10 和 ImageNette 数据集, 由于数据特征较为复杂, 相比于在 FashionMNIST 和 MNIST 等简单数据集, 后门缓解效果的整体表现有所下降。其中在 Refool 攻击下, VGG11 模型的后门缓解效果较为有限。由于其分类器由多层线性层构成, 特征提取部分生成的特征在经过多次线性映射后, 其分布已趋近于类别的决策边界, 特征信息显著压缩, 限制了特征解纠缠方法的后门缓解能力。出于验证方法通用性, 以及与不同模型结构保持一致的考虑, 本文选取最后一层线性层的输入特征空间作为特征解纠缠的对象。

5 总结与展望

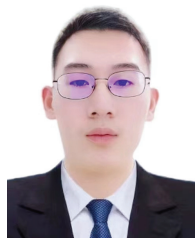
针对现有后门清除方法难以适应不同模型架构且效果不佳的问题, 本文提出了一种基于最大均值差异的特征解纠缠后门缓解方法。基于神经网络在不同层次上实现不同功能的特点, 本文提出了一种分层掩码抑制方法, 破坏模型中潜在后门与目标标签之间的强关联, 同时尽可能保留模型的良性特征。

在此基础上, 针对不同模型架构在特征提取模块上的方法存在差异, 但分类器模块有较高一致性的特点, 本文提出基于分类器模块的最大均值差异特征解纠缠方法。相比于对模型中间层激活值进行特征解纠缠, 本文直接对权重进行特征解纠缠, 实现良性特征与后门特征的有效分离。所提出的后门缓解方法能够适应不同的后门模型架构, 具有较强的通用性。本文方法的主要局限在于对于干净样本的依赖, 干净样本的数量直接影响算法的效果。下一步的工作将把后门缓解方法扩展到自然语言处理、生成式模型等不同领域, 针对领域的具体特点进行后门缓解, 提升后门缓解的效果。

参考文献

- [1] Wang Xutong, Yin Jie, Liu Chao, et al. A Survey of Backdoor Attacks and Defenses on Neural Networks[J]. *Chinese of Journal computers* 2024, 47(08): 1713-1743.
(汪旭童, 尹捷, 刘潮歌, 等. 神经网络后门攻击与防御综述[J]. *计算机学报*, 2024, 47(08): 1713-1743.)
- [2] Li Y, Jiang Y, Li Z, et al. Backdoor Learning: A Survey[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, 35(1): 5-22.
- [3] Wang Z, Ding H, Zhai J, et al. Training with more confidence: Mitigating injected and natural backdoors during training[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 36396-36410.
- [4] Li Y, Lyu X, Koren N, et al. Neural Attention Distillation: Erasing Backdoor Triggers from Deep Neural Networks[C]// *In 9th International Conference on Learning Representations, ICLR* 2021. 2021.
- [5] Xia J, Wang T, Ding J, et al. Eliminating Backdoor Triggers for Deep Neural Networks Using Attention Relation Graph Distillation[J]. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, 2022, 1481-1487. DOI: 10.24963/ijcai.2022/206.
- [6] Pang L, Sun T, Ling H, et al. Backdoor Cleansing with Unlabeled Data[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023: 12218-12227.
- [7] Min R, Qin Z, Shen L, et al. Towards Stable Backdoor Purification through Feature Shift Tuning[J]. *Advances in Neural Information Processing Systems*, 2024, 36.
- [8] Higgins I, Matthey L, Pal A, et al. beta-vae: Learning basic visual concepts with a constrained variational framework[C]//*International conference on learning representations*. 2017.
- [9] Hadad N, Wolf L, Shahar M. A Two-Step Disentanglement Method[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018: 772-780.
- [10] Deng W, Zhao L, Liao Q, et al. Informative Feature Disentanglement for Unsupervised Domain Adaptation[J]. *IEEE Transactions on Multimedia*, 2021, 24: 2407-2421.
- [11] Wang Z, Mei K, Zhai J, et al. Unicorn: A Unified Backdoor Trigger Inversion Framework[C]//*Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023)*, Kigali, Rwanda, 1-5 May 2023. San Francisco, CA: OpenReview.net, 2023.
- [12] Gu T, Liu K, Dolan-Gavitt B, et al. BadNets: Evaluating Backdoor Attacks on Deep Neural Networks[J]. *IEEE Access*, 2019, 7: 47230-47244.
- [13] Chen X, Liu C, Li B, et al. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning[J]. *arXiv preprint*, 2017, arXiv:1712.05526.
- [14] Wang Z, Zhai J, Ma S. BppAttack: Stealthy and Efficient Trojan Attacks Against Deep Neural Networks via Image Quantization and Contrastive Adversarial Learning[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19-24 June 2022*. New York: IEEE, 2022: 15074-15084.
- [15] Hayase J, Oh S. Few-Shot Backdoor Attacks via Neural Tangent Kernels[C]//*Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS* 2022.
- [16] Liu K, Dolan-Gavitt B, Garg S. Fine-Pruning: Defending Against Backdoor Attacks on Deep Neural Networks[C]//*International Symposium on Research in Attacks, Intrusions, and Defenses. Cham: Springer International Publishing*, 2018: 273-294.
- [17] Wu D, Wang Y. Adversarial Neuron Pruning Purifies Backdoored Deep Models[J]. *Advances in Neural Information Processing Systems*, 2021, 34: 16913-16925.
- [18] Wang B, Yao Y, Shan S, et al. Neural cleanse: Identifying and mi-

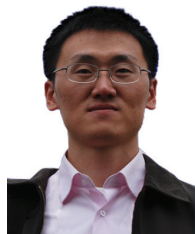
- tigating backdoor attacks in neural networks[C]//2019 IEEE symposium on security and privacy (SP). IEEE, 2019: 707-723.
- [19] Wang Z, Mei K, Ding H, et al. Rethinking the Reverse-Engineering of Trojan Triggers[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 9738-9753.
- [20] Lyu W, Zheng S, Ma T, et al. A Study of the Attention Abnormality in Trojaned BERTs[C]//*Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022: 4727-4741.
- [21] Sundararajan M, Najmi A. The Many Shapley Values for Model Explanation[C]//*International Conference on Machine Learning, PMLR*, 2020: 9269-9278.
- [22] DataCon Community. DataCon Open Dataset - DataCon2023 AI Security Dataset Direction Open Dataset[DB/OL]. 2024-01-18. <https://datacon.qianxin.com/opendata/openpage?resourceId=37>.
- [23] (DataCon 社区.DataCon 开放数据集-DataCon2023-AI 安全数据集方向开放数据集[DB/OL].2024-01-18.<https://datacon.qianxin.com/opendata/openpage?resourceId=37>.)
- [24] 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.
- [25] Huynh T, Nguyen D, Pham T, et al. Combat: Alternated training for effective clean-label backdoor attacks[C]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, 38(3): 2436-2444.
- [26] Cheng S, Tao G, Liu Y, et al. Lotus: Evasive and resilient backdoor attacks through sub-partitioning[C]. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024: 24798-24809.
- [27] Liu Y, Ma X, Bailey J, et al. Reflection Backdoor: A Natural Backdoor Attack on Deep Neural Networks[C]//*Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part X 16*. Springer International Publishing, 2020: 182-199.
- [28] Nguyen T A, Tran A T. WaNet-Imperceptible Warping-based Backdoor Attack[C]//*International Conference on Learning Representations*.
- [29] Li Y, Zhai T, Jiang Y, et al. Backdoor Attack in the Physical World[J]. *arXiv Preprint arXiv:2104.02361*, 2021.
- [30] Zheng R, Tang R, Li J, et al. Data-free backdoor removal based on channel lipschitzness[C]//*European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022: 175-191.



古津榜 于 2023 年在陆军工程大学网络工程专业获得学士学位。现在陆军工程大学网络空间安全专业攻读硕士学位。研究领域为软件安全、AI 安全。研究兴趣包括: 模糊测试、AI 后门检测。Email: jinbang66666@126.com



洪征 于 2007 年在解放军理工大学计算机应用专业获得博士学位。现任陆军工程大学指挥控制工程学院副教授。研究领域为协议安全、软件安全。研究兴趣包括: 协议逆向、模糊测试。Email: hz5215@163.com



张国敏 于 2009 年在解放军理工大学网络工程专业获得博士学位。现任陆军工程大学指挥控制工程学院副教授。研究领域为网络安全。研究兴趣包括: 软件定义网络、分布式系统。Email: 40519667@qq.com



江川 于 2024 年在陆军工程大学信息安全专业获得学士学位。现在陆军工程大学网络空间安全专业攻读硕士学位。研究领域为软件安全、协议安全。研究兴趣包括: 模糊测试、污点分析。Email: 331657393@qq.com