计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(9):57−68 [doi: 10.15888/j.cnki.csa.009957] [CSTR: 32024.14.csa.009957]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

N-Model: 多深度学习模型动态组合的智能系统安全弹性增强①

程泽凯[1], 刘高天[1], 蒋建春[2], 庞志伟[1], 滕若阑[1], 梅　瑞[3]

[1](安徽工业大学 计算机科学与技术学院, 马鞍山 243002)
[2](中国科学院 软件研究所 集成创新中心, 北京 100190)
[3](科大讯飞信息安全实验室, 合肥 230088)
通信作者: 蒋建春, E-mail: jianchun@iscas.ac.cn

摘　要: 基于深度学习智能系统面临对抗攻击、供应链攻击等安全威胁问题日益突出, 而传统智能系统采用单一模型, 其防御机制是静态的、确定的模式, 模型的功能存在单点脆弱性, 导致智能系统缺乏安全弹性. 本文提出了一种多个深度学习模型动态组合的方法 (N-Model), 实现模型的多样性和随机性, 通过模型的动态变化增加智能攻击对象及攻击途径的不确定性, 结合多模型的表决机制, 增强智能系统的安全弹性. 理论安全分析表明, N-Model 组合模型在攻击情景下相比单一模型具有较高的期望准确率. 实验结果进一步证实, 在 CIFAR-10 数据集下, N-Model 组合模型可抵御多种对抗攻击, 其攻击成功率低于单一模型, 表现出良好的综合安全性能.

关键词: 人工智能安全; 深度学习防御; 随机模型调度; 多模型表决; 攻击容忍性; 系统安全弹性

引用格式: 程泽凯,刘高天,蒋建春,庞志伟,滕若阑,梅瑞.N-Model: 多深度学习模型动态组合的智能系统安全弹性增强.计算机系统应用,2025,34(9):57–68. http://www.c-s-a.org.cn/1003-3254/9957.html

N-Model: Enhancing Security Resilience of Intelligent Systems Through Dynamic Combination of Multiple Deep Learning Models

CHENG Ze-Kai[1], LIU Gao-Tian[1], JIANG Jian-Chun[2], PANG Zhi-Wei[1], TENG Ruo-Lan[1], MEI Rui[3]

[1](School of Computer Science and Technology, Anhui University of Technology, Ma'anshan 243002, China)
[2](Integrated Innovation Center, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)
[3](iFLYTEK Information Security Laboratory, Hefei 230088, China)

Abstract: Security threats such as adversarial and supply chain attacks are increasingly prominent in deep learning-based intelligent systems. Traditional systems typically adopt a single model with static and deterministic defense mechanisms, making them susceptible to single points of failure and lacking in security resilience. To address this issue, a method termed N-Model is proposed, which leverages the dynamic combination of multiple deep learning models to promote diversity and randomness. The dynamic switching among models increases the uncertainty of attack targets and methods. A multi-model voting mechanism is integrated to further strengthen security resilience. Theoretical analysis indicates that the proposed N-Model achieves higher expected accuracy than single-model approaches under attack scenarios. Experimental evaluations on the CIFAR-10 dataset demonstrate that the N-Model effectively resists various adversarial attacks, exhibiting a lower attack success rate and superior overall security performance.

Key words: artificial intelligence security; deep learning defense; random model scheduling; multi-model voting; attack tolerance; system security resilience

① 基金项目: 安徽教育厅高校自然科学研究项目 (2024AH040028); 安徽大学信息材料与智能感知安徽省实验室开放课题 (IMIS202215)
收稿时间: 2025-02-17; 修改时间: 2025-03-10; 采用时间: 2025-03-31; csa 在线出版时间: 2025-07-25
CNKI 网络首发时间: 2025-07-28

随着人工智能技术的快速发展, 其在医疗、金融、自动驾驶等关键领域的应用越来越广泛. 然而, 这些系统的安全性直接影响着其在关键任务中的表现. 近年来, 人工智能模型面临的攻击手段日益复杂, 这给系统的安全可靠性带来了严峻挑战. 据经济合作与发展组织的 AI 事件监测器统计显示, AI 风险事件在过去几年间呈现爆发式增长. 其中 2022 年仅记录了少量 AI 风险事件, 而到 2024 年, 事件总数已增长了约 21.8 倍. 在 2019–2024 年间记录的所有风险事件中, 约 74% 与 AI 安全问题直接相关[1].

当前 AI 系统面临的主要威胁包括供应链攻击和对抗攻击, 后者尤为危险. 在交通领域, 通过在道路标志上附加特定贴片, 可使交通标志识别系统的准确率从 91% 骤降至不到 35%[2,3]. 类似地, 在医疗诊断场景中, 通过加入仅 8% 的恶意数据, 攻击者就能使模型对超过 50% 的患者的用药量建议产生显著偏差[4]. 这些实例表明, 对抗攻击能够严重影响 AI 系统的决策质量, 带来潜在安全风险.

传统防御体系通常依赖“已知风险”假设的精确防御, 即需要事先了解攻击的来源、特征、途径和机制等信息, 才能设计出有效的防御措施. 这种方法在面对已知攻击类型时可能有效, 但随着对抗攻击手段的快速演化, 其局限性日益凸显, 且主要分为 3 个方面. 首先, 传统防御方法通常面临较大的计算开销; 其次, 这些方法主要集中于优化模型训练过程, 其过于依赖对攻击模式的先验知识[5], 当遇到未知攻击模式时容易失效; 最后, 由于防御策略的静态性, 攻击者可以通过逆向分析预测系统行为, 从而设计更具针对性的攻击, 进一步降低系统安全性. 由此可见, 传统的单一模型防御策略已无法满足人工智能系统在动态攻击环境中的安全需求, 亟需探索更为灵活和适应性强的防御机制.

在此背景下, 智能系统的弹性和动态防御成为应对安全挑战的重要研究方向. 研究者们提出了多种基于随机性和多样性的防御机制, 例如, 邬江兴等人[6]提出了网络空间拟态防御理论, 强调通过引入随机性和多样性来增强系统安全性. 在此基础上, 张卓等人[7]进一步研究了基于威胁的动态防御机制, 通过动态选取模型池中的模型, 增加攻击路径的不确定性, 使攻击者难以预测系统的防御策略. Liu 等人[8]证明, 动态选择模型集成属性能够创造随机分布的输入噪声和不确定的概率梯度信息, 可提高系统对抗攻击的防御能力. 然而, 现有动态防御方法仍存在计算开销大、泛化能力有限、缺乏自适应调整能力等问题, 难以在实际应用中发挥理想效果.

基于上述研究基础并针对现有方法的不足, 本文提出了一种名为 N-Model 的深度学习模型动态组合方法, 进一步拓展了模型多样性与随机性在防御机制中的应用. N-Model 通过利用预训练模型池实现轻量化部署, 引入基于数据分布的模型筛选机制提高防御泛化能力, 采用自适应表决机制动态优化决策过程, 使攻击者难以预测系统行为. 通过这些设计, N-Model 有效增加了智能系统的攻击不确定性, 同时避免单一模型的安全脆弱性.

本文的主要贡献如下.

1) 提出了一种多模型动态组合方法 N-Model, 实现智能系统的模型多样性与随机性, 增强模型动态性, 提升系统对抗攻击及供应链攻击的容忍性.

2) 设计了一种结合软硬投票的多模型表决机制, 增强了智能系统功能安全, 有效提高了系统可靠性.

3) 通过理论分析和实验验证, 证明了所提方法在攻击场景下的优越性能, 其分类准确率和抗攻击成功率均优于单一模型.

本文第 1 节介绍相关研究工作, 包括对抗攻击、单一防御以及动态防御的国内外研究现状. 第 2 节详细描述所提机制的设计与实现. 第 3 节从理论方面验证本方法的合理性. 第 4 节展示了实验设置与结果分析. 第 5 节总结全文并展望未来研究方向.

1 相关工作

1.1 对抗攻击的研究现状

对抗攻击作为一种破坏力大且隐匿性强的攻击方式, 近年来受到了学术界和工业界的广泛关注. 对抗攻击通过在输入样本中添加特定扰动, 诱导深度学习模型输出错误结果. Goodfellow 等人[2]提出对抗样本的概念后, 众多研究者对其产生的原因进行了探索. 根据攻击领域, 对抗攻击可以分为数字域攻击和物理域攻击.

在数字领域, 对抗攻击方法经历了从简单到复杂的演进过程. 最早期的方法之一是 FGSM[2], 它通过计算关于输入梯度的损失函数, 快速生成对抗样本. 随后, BIM[9]和 PGD[5]等方法进一步发展, 利用迭代和投影技术增强了对抗样本的有效性. Moosavi-Dezfooli 等人[10]提出的 Deepfool 方法通过迭代寻找最小扰动实现分类

错误. Carlini 等人[11]提出的 C&W 攻击利用拉格朗日乘子法实现扰动最小化, 这两项工作奠定了优化基础对抗攻击的理论基础. 近年来, 基于优化和生成对抗网络的攻击方法得到了广泛研究. 比如, Croce 等人[12]提出的 AutoAttack 方法, 通过自动化策略生成对抗样本, 在多种场景下展现了强大的攻击能力. Bashivan 等人[13]提出了一种特征去敏感化攻击方法, 显著提高了对抗样本的迁移性. Zhang 等人[14]开发了基于模型增强的对抗样本迁移技术, 进一步降低了攻击成本.

在物理域中, 对抗攻击的主要挑战是如何在现实世界中保持攻击效果. 例如, Brown 等人[15]提出了一种对抗补丁, 通过在目标物体上添加小区域扰动, 成功欺骗了图像分类器. 此外, Athalye 等人[16]提出的光学对抗攻击通过物理设备生成对抗样本, 为现实场景中的攻击带来了新思路.

尽管这些方法在攻击效率和成功率上不断进步, 但它们存在一个共同特性, 即主要针对静态单一模型而设计.

1.2 单一模型防御方法

单一模型的防御方法主要集中于增强单个模型自身的鲁棒性和抗攻击能力. 这些方法通常通过优化模型训练过程或改进模型架构设计来实现. 例如, 对抗训练是一种经典的防御策略, 通过在训练过程中生成并引入对抗样本, 模型能够更好地学习对抗样本的特征, 从而提升鲁棒性[2,17]. Madry 等人[5]改进的 PGD 对抗训练成为评估防御效果的重要基准. Zhang 等人[18]开发的 TRADES 算法通过引入鲁棒性与准确性的权衡项, 在维持模型准确率的同时提高了防御能力. 此外, Liu 等人[8]通过剪枝未被正常样本激活的神经元, 降低了模型对特定扰动的敏感性. Cohen 等人[19]提出的随机平滑方法通过添加高斯噪声提供了概率性防御证明, 在不牺牲太多准确率的情况下提高了模型鲁棒性. 它们标志着单一模型防御研究的重要里程碑.

然而, 这些方法通常依赖已知攻击模式的先验知识, 在面对动态多样攻击时表现出明显局限性. Athalye 等人[16]证明大多数防御方法存在梯度掩蔽问题, 易被适应性攻击绕过; Carlini 等人[20]通过系统评估表明多种防御方法面对强力攻击时效果显著降低; Tramèr 等人[21]研究发现多数防御方法存在“虚假安全感”, 其有效性局限于特定攻击类型. 当攻击者获取模型防御策略信息后, 可通过针对性设计对抗样本规避现有防御机制[22]. 因此, 单一模型防御在复杂多变的攻击环境下难以提供有效防护.

1.3 动态防御与异构冗余

近年来, 针对人工智能系统的安全威胁、动态防御和异构冗余已成为重要的研究方向. 动态防御通过对系统结构和行为动态调整, 增加攻击路径的不确定性, 使攻击者难以预测和复现攻击过程, 从而显著增强系统的抗攻击能力[7,23]. 这一理念源于网络安全领域的 moving target defense (MTD) 理论, 该理论通过动态改变系统的攻击面, 增加攻击者的不确定性[24]. Rehman 等人[25]对 MTD 进行了系统总结, 阐述了其在网络安全领域中的应用和效果. 这种动态防御思想为深度学习模型防御提供了新的思路, 即不再局限于静态的单一模型防御, 而是通过动态变化的模型组合来应对攻击. Sengupta 等人[26]指出动态性和不可预测性是提高防御效果的关键因素.

在此基础上, 国内外学者在动态防御方面的研究不断深入. 文献[6,27]提出的动态异构冗余 (dynamic heterogeneous redundancy, DHR) 通过动态组合功能等价但实现形式不同的异构模型, 进一步提升了系统的弹性与容错性. DHR 的核心在于结合多模裁决与多维动态重组策略, 显著降低攻击成功率并提升系统安全弹性. Qin 等人[28]提出了基于随机集成平滑的动态防御方法, 该方法采用网络架构和平滑参数作为集成属性, 并在每次推理预测请求前动态更改基于属性的集成模型, 有效提高了深度神经网络的对抗鲁棒性. Waghela 等人[29]提出了 ARDEL 动态防御框架, 通过利用多个预训练模型的多样性, 并根据输入特征动态调整集成配置, 增强语言模型抵抗对抗攻击的能力.

然而, 现有的动态防御和异构冗余方法在应用于深度学习模型时仍面临一些挑战. 首先, 大多数研究主要集中在网络系统层面, 缺乏对深度学习模型特性的针对性考虑; 其次, 现有方法通常依赖于固定的模型组合策略, 缺乏灵活性, 在快速变化的攻击环境中可能存在失效的风险.

2 N-Model 机制

本节介绍提出的一种基于模型多样性与随机性的防御机制, 该机制包含数据抽样、模型一致性筛选和集成预测 3 个核心阶段. 防御机制的整体流程如图 1 所示. 首先, 在数据抽样阶段, 采用基于分布密度与分

位数结合的抽样策略, 从原始数据集中提取具有代表性和均衡性的子集数据, 为后续环节提供精准的验证基础; 随后, 在模型一致性筛选阶段, 利用随机性和多样性对模型池中的候选模型进行筛选, 从中选择一致性最高的模型构建 N-Model 组合模型; 最后, 在集成预测阶段, 通过软硬投票机制融合模型的预测结果, 降低单模型预测偏差或失效对整体性能的影响. 接下来将详细阐述每一部分的设计思路及实现细节.

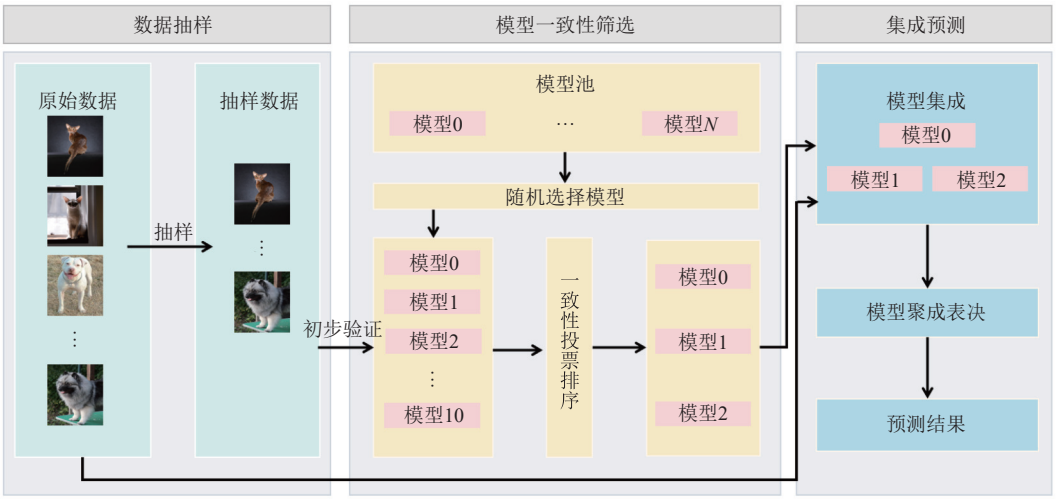


图 1 基于模型多样性与随机性的防御机制流程图

2.1 数据抽样

在本文提出的防御机制中, 数据抽样是第 1 阶段, 其目的是从原始数据集中抽取具有代表性和均衡性的样本, 为模型一致性验证和筛选提供基础. 在实际应用中, 面对未知数据的情况下, 无法事先了解数据的具体分类, 尤其是在自然数据分布中, 样本通常在中间区域密集、两端稀疏. 因此, 本文使用基于分布密度和分位数结合的抽样策略. 该方法通过分位数将像素均值划分为 5 个区间 (0–20%、20%–40%、40%–60%、60%–80%、80%–100%), 并根据每个区间的样本密度动态调整抽样比例. 具体而言, 设数据集总样本数为N, 目标抽样总数S, 区间k内的样本数为n_k. 首先, 计算每个区间的密度权重w_k:

$$w_k = \frac{n_k}{N} \tag{1}$$

其中, w_k表示区间k的样本占比. 然后, 根据密度权重w_k分配每个区间的抽样数量s_k:

$$s_k = \lceil w_k \cdot S \rceil \tag{2}$$

其中, s_k表示区间k中需要抽取的样本数, $\lceil\cdot\rceil$表示向上取整操作. 该策略确保高密度区间 (如像素均值中间区域) 分配更多的样本, 同时保留低密度区间的稀疏样本以增强多样性. 相比传统的均匀分位数抽样, 本抽样方法更贴合数据的实际分布特性, 即样本的代表性和均衡性, 为后续环节提供了可靠的数据基础.

2.2 模型一致性筛选

在本文提出的防御机制中, 模型一致性筛选是第 2 阶段, 其核心目标是从模型池中动态选择出具有高一致性和高可靠性的模型, 以构建最终的 N-Model 组合模型. 该阶段首先从模型池中随机选取 11 个模型组成临时模型池, 对第 1 阶段抽样的数据进行并行预测. 模型池中的模型主要来源于现有的主流预训练模型库, 例如 PyTorch 官方模型库和 Huggingface Transformers 平台. 模型池中的模型主要包括 3 类体系: 1) 卷积神经网络架构包含 ResNet、VGG、GoogLeNet、DenseNet 及 MobileNetV2 等典型模型; 2) 视觉 Transformer 架构包含 ViT、Swin-T 等先进模型; 3) 传统机器学习算法包括支持向量机、随机森林等.

尽管部分模型在架构层面存在相似性, 但通过差异化的参数初始化策略、动态数据扰动机制以及优化器配置, 实现了模型间的功能差异性.

在一致性筛选过程中, 首先对各模型表现进行定量评分, 以衡量其预测结果的可靠性. 每个模型的总评分S_i是基于其对所有数据点预测结果的得分累加而得. 具体而言, 给定n个数据, 模型i的总评分计算公式为:

$$S_i=\sum_{j=1}^{n}\begin{cases}1, & \text{if}\left|\left\{k\mid size\left(G_j\left(y_{k,j}\right)\right)=\max\limits_k size\right\}\right|=1\\ 0, & \text{otherwise}\end{cases} \tag{3}$$

其中, $G_j(y_{i,j})$表示第j个数据点上预测结果相同的模型集合, $size(G_j(y_{i,j}))$是该集合的大小. 对于每个数据j, 当同时满足这两个条件时, 模型i获得 1 分.

① 集合$\left\{k\mid size(G_j(y_{k,j}))=\max\limits_k size\right\}$包含所有满足条件的模型索引$k$, 即在第$j$个数据点上, 模型$k$的预测结果属于最大分组.

② $|\cdot|=1$ 表示集合中只有一个元素, 意味着在第j个数据点上, 预测结果最多的分组是唯一的.

最终, 根据每个模型的总评分S_i进行排名, 并筛选出预测结果一致性和可靠性更高的 3 个模型, 构建 N-Model 组合模型.

2.3 集成预测

在本文提出的防御机制中, 集成预测是关键的最终阶段. 筛选出的 3 个模型通过协作执行预测任务, 并通过软硬投票机制融合各自的预测结果. 采用模型一致性得分的显著性分析作为投票机制选择依据, 旨在提高系统的效率和准确性. 集成预测流程如图 2 所示.

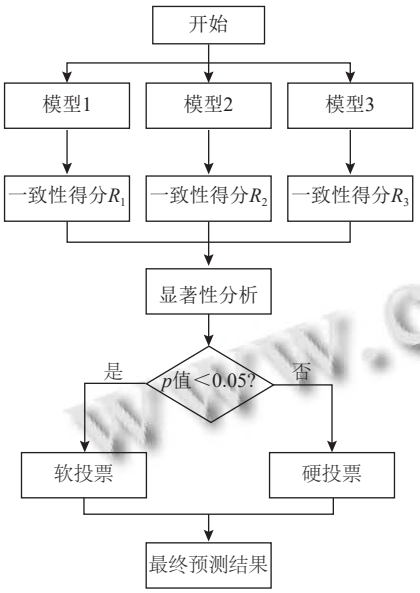


图 2 集成预测流程图

首先, 通过单因素方差分析统计检验评估 3 个模型一致性得分的显著性差异. 当计算得到的p值小于显著性水平 0.05 时, 表明至少有两个模型之间存在统计学上的显著差异, 即其中一个模型显著比其他模型更可靠, 则采用软投票机制. 软投票通过计算权重来加强可靠模型的影响. 权重基于模型一致性得分R_i计算, 如式 (4) 所示:

$$w_i=\frac{R_i}{\sum_{j=1}^{3}R_j} \tag{4}$$

其中, R_i表示第i个模型的一致性得分. 例如, 当 R_1= 0.85、R_2=0.75、R_3=0.65 时, 权重分配为 37.8%、33.3%、28.9%, 确保一致性得分较高的模型对最终决策有更大影响力.

对于每个类别c, 计算加权概率得分:

$$P_{\text{soft}}(c)=\sum_{i=1}^{3}w_i\cdot P_i(c) \tag{5}$$

并选择得分最高的类别作为最终输出. 如果得分差异不显著, 则采用硬投票机制, 以节约时间和计算资源. 硬投票基于多数决策原则, 通过统计每个类别的预测频数, 选择频数最高的类别作为最终输出.

由于模型集成是动态筛选生成的, 每次任务中使用的模型组合各不相同, 这种动态性增加了系统行为的不可预测性, 降低了攻击者规避策略的可能性.

综上所述, N-Model 从两个关键维度增强了系统动态性: 1) 时间维度上通过一致性筛选动态选择模型组合, 使攻击者难以预测系统配置; 2) 决策机制维度上根据模型一致性得分动态切换投票方式并调整权重分配, 提升了决策质量. 这种多维度动态性增强了系统的安全弹性.

3 理论分析与验证

N-Model 组合模型具有 3 个核心理论特性: 1) 性能边界特性: 在满足特定条件时, 组合模型的性能下限高于单一最弱模型, 性能上限可超越单一最优模型; 2) 攻击容忍特性: 当单一模型受到攻击时, 组合模型的准确率下降幅度小于单一模型; 3) 稳定性特性: 在不同攻击强度下, 组合模型的性能退化速度慢于受攻击的单一模型, 保持更持久的功能可用性. 为系统地验证这些特性, 本文从理论角度计算 N-Model 组合模型的理论最大值、理论最小值和期望准确率且在多种情况对比单一模型和 N-Model 组合模型的效果.

为便于后续的理论说明, 假设构建集成的基模型 A、B、C 的准确率分别为P_1、P_2、P_3, 且$P_1>P_2>P_3$, 以下为具体理论分析与推导.

3.1 理论最小值

本节旨在证明 N-Model 组合模型的性能边界特性, 即在满足特定条件时, 性能下限高于单一最弱模型的特性. 理论小值代表了 N-Model 组合模型所能达到最差的理论下限. 从硬投票的角度进行分析, 为计算 N-Model 组合模型识别准确率的理论最小值, 需使得各基模型之间的一致识别能力最小化, 即确保任意两个模型的识别结果差异最大化.

由于上述的假设为$P_1 > P_2 > P_3$, 则 3 个模型准确率中的最大值为P_1, 中间值为P_2, 最小值为P_3. 以下分 4 种情况计算 N-Model 组合模型中准确率理论最小值 *Min*, 具体情况如图 3.

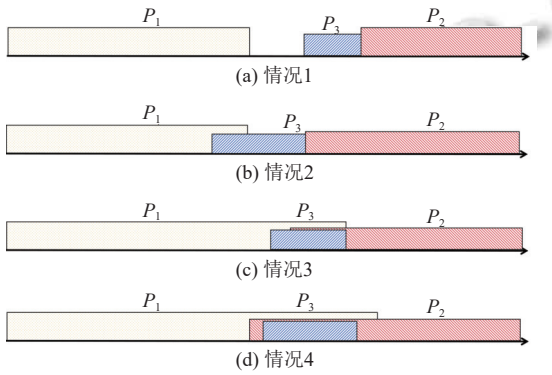


图 3 N-Model 组合模型准确率理论最小值情况示意分布图

● 情况 1: 如图 3(a) 所示, 3 个模型中两两之间不存在正确识别的一致性.

情况 1 中 N-Model 组合模型的准确率, 即模型一致的部分为 0.

● 情况 2: 如图 3(b) 所示, 性能最好的两个模型 A、B 之间存在正确识别的一致性. 第 3 个模型 C 与模型 A、B 之间存在正确识别的一致性, 这里假设模型 A 与模型 C 存在正确识别的一致性.

情况 2 中 N-Model 组合模型的准确率, 即为模型 A (白色) 与模型 C (蓝色) 识别一致的部分: $P_1 + P_3 - (1 - P_2)$.

● 情况 3: 如图 3(c) 所示, 性能最好的两个模型 A、B 之间存在正确识别的一致性. 第 3 个模型 C 与模型 A、B 之间存在正确识别的一致性, 且模型 C 的识别能力覆盖了模型 A、B 的一致部分.

情况 3 中 N-Model 组合模型的准确率, 即为模型 C 的准确率 (蓝色): P_3.

● 情况 4: 如图 3(d) 所示, 性能最好的两个模型 A、B 之间存在正确识别的一致性. 第 3 个模型 C 与模型 A、B 之间也存在正确识别的一致性, 且模型 A、B 一致的部分覆盖了模型 C 的识别能力.

情况 4 中 N-Model 组合模型的准确率, 即为模型 A (白色) 和模型 B (红色) 识别一致的部分: $P_1 + P_2 - 1$.

综上所述, 汇总公式如下:

$$Min = \begin{cases} 0, & \text{if } P_1+P_2+P_3 < 1 \\ P_1+P_3-(1-P_2), & \text{if } P_1+P_2+P_3 \geqslant 1 \ \& \ P_1+P_2 \leqslant 1 \\ P_3, & \text{if } P_1+P_2+P_3 \geqslant 1 \ \& \ P_3 > P_1+P_2-1 \\ & \quad \& \ P_1+P_2 \geqslant 1 \\ P_1+P_2-1, & \text{if } P_1+P_2+P_3 \geqslant 1 \ \& \ P_3 \leqslant P_1+P_2-1 \end{cases} \tag{6}$$

若证明 N-Model 组合模型优于单个模型, 则证明 N-Model 组合模型准确率的理论最小值高于单个模型准确率的最小值即可, 即$Min \geqslant P_3$. 以下从 4 个不同情况展开讨论:

情况 1:

$$Min - P_3 = 0 - P_3 < 0 \tag{7}$$

情况 2:

$$\begin{aligned} Min - P_3 &= P_3 + P_1 - (1 - P_2) - P_3 \\ &= P_1 + P_2 - 1 < 0 \end{aligned} \tag{8}$$

情况 3:

$$Min - P_3 = P_3 - P_3 = 0 \tag{9}$$

情况 4:

$$Min - P_3 = P_1 + P_2 - 1 > 0 \tag{10}$$

综上所述, 当满足情况 3 和情况 4 时, 即$P_1 + P_2 > 1$的情况下, N-Model 组合模型的效果优于效果最差的单个模型.

3.2 理论最大值

本节旨在证明 N-Model 组合模型的性能边界特性, 即在满足特定条件时, 性能上限可超越单一最优模型. 理论最大值代表了 N-Model 组合模型所能达到最优的理论上限. 从硬投票的角度进行分析, 为计算 N-Model 组合模型识别准确率的理论最大值, 需最大化各基模型之间的一致识别能力, 即使得任意两个模型的一致识别能力最大.

以下分 3 种情况计算 N-Model 组合模型中准确率理论最大值 *Max*, 具体情况如图 4.

● 情况 1: 如图 4(a) 所示, 3 个模型两两之间一致

的正确识别能力可覆盖所有的数据.

情况 1 中 N-Model 组合模型的准确率, 即模型一致的部分为 1.

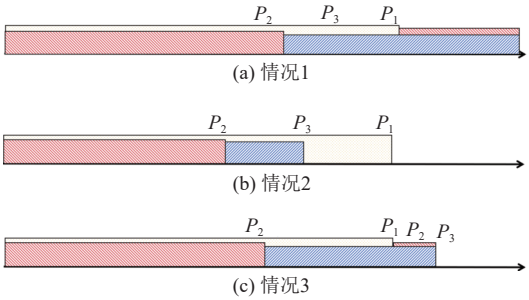


图 4 N-Model 组合模型准确率理论最大值情况示意分布图

● 情况 2: 如图 4(b) 所示, 性能最好的两个模型 A、B 之间存在正确识别的一致性, 且一致的部分是模型 B 的识别能力; 模型 A、C 之间出现一致的正确识别能力, 且一致的部分是模型 C 的识别能力; 此外, 模型 B、C 之间不存在正确识别的一致性.

情况 2 中 N-Model 组合模型的准确率, 即一致的部分为模型 B、C 的准确率之和: P_2+P_3.

● 情况 3: 如图 4(c) 所示, 性能最差的两个模型 B、C 的正确识别的一致性未被模型 A 所包含, 且模型 B、C 之间存在正确识别的一致性.

情况 3 中 N-Model 组合模型的准确率, 即模型一致的部分为: $\dfrac{P_1+P_2+P_3}{2}$.

综上所述, 汇总公式如下:

$$Max=\begin{cases}1, & \text{if } \dfrac{P_1+P_2+P_3}{2}>1\\ P_2+P_3, & \text{if } P_2+P_3<P_1\\ \dfrac{P_1+P_2+P_3}{2}, & \text{if } \dfrac{P_1+P_2+P_3}{2}<1\\ & \ \ \&\ \ P_2+P_3>P_1\end{cases} \tag{11}$$

若证明 N-Model 组合模型优于单个模型, 则证明 N-Model 组合模型准确率的理论最大值高于单个模型准确率的最大值即可, 即 $Max \geqslant P_1$. 以下从 3 个不同情况展开讨论.

情况 1:

$$Max-P_1=1-P_1\geqslant 0 \tag{12}$$

情况 2:

$$Max-P_1=P_2+P_3-P_1<0 \tag{13}$$

情况 3:

$$\begin{aligned}Max-P_1&=\frac{P_1+P_2+P_3}{2}-P_1\\&=\frac{P_1+P_2+P_3-2P_1}{2}>0\end{aligned} \tag{14}$$

综上所述, 当满足情况 1 和情况 3 时, N-Model 组合模型优于单个最优模型.

3.3 期望准确率

本节旨在证明 N-Model 组合模型具有较强的攻击容忍特性, 即当单一模型受到攻击时, 组合模型的准确率下降幅度小于单一模型. 为从概率统计角度精确量化系统在攻击情境下的性能损失, 有效衡量组合模型对攻击引起的单点性能退化的抵抗能力, 本节选择期望准确率作为分析指标.

N-Model 组合模型的期望准确率可理解为至少有两个单一模型预测正确的概率. 其数学表达为:

$$E=P_1P_2P_3+P_1P_2(1-P_3)+P_1(1-P_2)P_3+(1-P_1)P_2P_3 \tag{15}$$

式 (15) 可以变换为:

$$E=P_1P_2+P_1P_3+P_2P_3-2P_1P_2P_3 \tag{16}$$

当单一模型 A 受到攻击, 其准确率下降 d 时, 其 N-Model 组合模型的期望准确率为:

$$\begin{aligned}E=&(P_1-d)P_2P_3+(P_1-d)P_2(1-P_3)\\&+(P_1-d)(1-P_2)P_3+(1-P_1-d)P_2P_3\end{aligned} \tag{17}$$

式 (17) 可以变换为:

$$E=P_1P_2+P_1P_3+P_2P_3-2P_1P_2P_3-d(P_2+P_3-2P_2P_3) \tag{18}$$

期望准确率下降了 $\Delta E=d(P_2+P_3-2P_2P_3)$. 若想证明 $\Delta E\leqslant d$, 则需要证明 $P_2+P_3-2P_2P_3\leqslant 1$.

$$\begin{aligned}P_2+P_3-2P_2P_3&\leqslant P_2+P_3-P_2P_3\\&=1-(1-P_2-P_3+P_2P_3)\\&=1-(1-P_2)(1-P_3)<1\end{aligned} \tag{19}$$

由式 (19) 的推导可得到 $\Delta E\leqslant d$, 即得出证明, N-Model 组合模型在受到攻击时的攻击成功率一定优于最弱单个模型的攻击成功率.

3.4 理论仿真

本节旨在验证 N-Model 组合模型的稳定性特性, 即在不同攻击强度下, 组合模型的性能退化速度慢于受攻击的单一模型, 保持更持久的功能可用性. 为分析 N-Model 组合模型在不同攻击强度下的性能表现, 开展模拟实验以量化理论最大值、理论最小值和期望值

随着对 P_1 的攻击强度变化的准确率变化曲线, 如图 5 所示, 其中, P_1=100%, P_2=80.0%, P_3=80.0%.

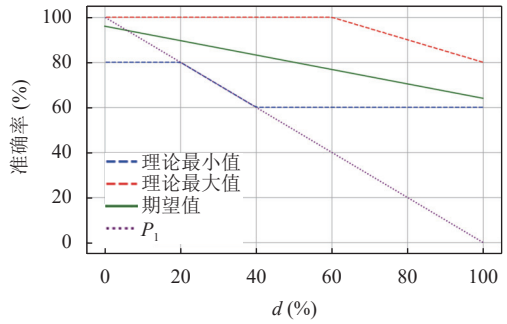


图 5 N-Model 组合模型性能随攻击强度的变化曲线

实验结果表明, 随攻击强度d逐步增大, N-Model 组合模型的理论最大值呈现出平滑下降趋势, 即使在 d=100% 时, 理论最大值仍然维持在较高水平, 表明即使部分基模型失效, 系统整体性能仍然能够得到保障. 理论最小值在攻击强度 d=40% 时保持恒定, 显示 N-Model 组合模型在此阶段具有稳定的最小性能. 期望值随攻击强度增加逐步下降, 并在高强度攻击下逐渐趋近理论最小值, 但整体始终优于单一模型, 证明了集成策略在分散风险和整合模型预测结果方面的有效性.

实验分析进一步表明, 在低攻击强度 ($d \leqslant 20\%$) 下, N-Model 组合模型的理论最大值、理论最小值和期望值均维持高水平, 性能表现稳定. 在中低强度攻击区间 ($20\% \leqslant d < 40\%$) 时, 期望值接近理论最大值, 展现出协同机制对攻击影响的缓冲作用. 在中高强度攻击区间 ($40\% \leqslant d < 60\%$) 时, 期望值开始显著下降, 但仍高于理论最小值, 证明了 N-Model 组合模型具备一定的容错性和冗余能力. 在高强度攻击 ($d \geqslant 60\%$) 下, 期望值逐渐趋近理论最小值, 而理论最小值的稳定性进一步说明系统在极端条件下的基础功能保障能力.

综合实验结果表明, N-Model 组合模型在不同攻击强度下均展现出优于单一模型的安全弹性. 理论最大值、理论最小值和期望值的变化趋势验证了 N-Model 组合模型在抗攻击能力方面的优势, 为高可靠性任务场景中的应用提供了理论支撑和性能保障.

4 实验设计与评估

4.1 实验目的

本文提出的防御机制以动态模型选择和多模型表决为核心, 旨在增强深度学习系统在复杂对抗攻击场景下的鲁棒性和攻击容忍性. 本实验的主要目的是验证本文方法在不同类型对抗攻击下的防御效果, 同时评估其在模型分类准确性、攻击成功率方面的性能表现. 此外, 通过与传统防御方法进行对比, 进一步证明本文方法在动态防御环境中的优势.

4.2 实验设置

4.2.1 数据集

实验选用 CIFAR-10 作为基准数据集, CIFAR-10 数据集包含 10 类自然图像, 共计 60 000 张 32×32 像素的彩色图像, 其中训练集 50 000 张, 测试集 10 000 张.

在实验中, 对抗攻击样本通过添加扰动生成, 具体采用的攻击方法包括 FGSM、PGD, 分别代表快速梯度攻击、迭代优化攻击方法.

4.2.2 实验环境

实验运行于配置有以下硬件和软件的服务器上.

硬件: NVIDIA GeForce RTX 4060 GPU (8 GB 显存)、Intel(R) Core(TM) i7-13700H CPU.

软件: 操作系统为 Windows 11, 使用 PyTorch 2.0 和 Huggingface Transformers 作为深度学习框架.

4.2.3 模型池设置

模型池由同构模型与异构模型组成.

同构模型: 基于 CNN、VGG、ResNet 等不同深度的卷积神经网络架构, 在 CIFAR-10 及其子集上分别独立训练.

异构模型: 包括 Transformer 架构、传统机器学习模型 (如随机森林、支持向量机) 和其他神经网络模型.

每次实验从模型池中随机选取 11 个模型组成临时模型集合, 用于一致性筛选.

4.3 单一防御与 N-Model 防御实验对比

4.3.1 实验方案

本实验设计两种对比方案以验证 N-Model 组合模型是否优于单一模型. 1) 单一模型执行方案 (Baseline): 该方案直接从模型池中选取单个模型, 作为防御机制的执行体, 用于测试单一模型在各类攻击场景下的基本防御性能, 作为基准参照. 2) N-Model 组合模型: 该方案通过框架的动态筛选机制, 从模型池中随机选取多个模型 (例如 11 个), 利用一致性筛选选出性能最优的 3 个模型组成 N-Model 组合模型, 并通过表决机制对输入样本进行预测.

实验流程包括模型选择、攻击场景测试和指标分

析 3 个阶段. 在模型选择阶段, 单一模型方案直接从模型池中选取单个模型; 而 N-Model 组合模型方案则从模型池中随机选取多个模型, 通过一致性筛选确定最终的模型集合. 在攻击场景测试阶段, 分别在无攻击场景 (干净样本) 和对抗攻击场景 (包括 FGSM、PGD) 下测试 Baseline 与 N-Model 组合模型方案的防御表现. 在指标分析阶段, 记录攻击成功率 (ASR) 和分类准确率 (Acc) 两个核心指标, 其中 ASR 用于衡量对抗样本成功欺骗模型的比例, 越低说明防御效果越好. 通过比较 Baseline 与 N-Model 组合模型方案在不同场景下的性能差异, 验证模型集成在降低攻击成功率和提升系统容忍性方面的有效性.

4.3.2 实验结果与分析

实验结果如表 1–表 3 所示, 在对抗攻击 (FGSM、PGD) 和干净数据下, N-Model 组合模型在大部分场景下, 相较于单一模型方案 (Baseline) 有效降低了攻击成功率 (ASR), 同时还能保持较为稳定的分类准确率 (Acc). 例如, 在 FGSM 攻击下, N-Model 组合模型方案针对 GoogLeNet 对抗样本的 ASR 为 64.20%, 低于 Baseline GoogLeNet 的 66.27%, 展现出较好的防御能力; 而在针对 VGG 的 PGD 攻击中, N-Model 组合模型方案的 ASR 为 74.49%, 优于 Baseline VGG 的 99.61%. 此外, 在干净数据下, N-Model 组合模型方案的分类准确率达到 82.47%, 接近于最佳 Baseline (Baseline ResNet 的 80.27%), 表明其在正常场景中的性能损失较小. 结果说明, 通过模型池的动态筛选和一致性表决机制, N-Model 组合模型方案能够有效提升系统的安全弹性和攻击容忍性.

表 1 干净数据下单一防御与 N-Model 防御实验对比

模型选择	Acc (%)
Baseline GoogLeNet	77.22
Baseline VGG	69.67
Baseline ResNet	80.27
N-Model	82.47

4.4 不同扰动强度下实验对比

4.4.1 实验方案

在第 4.3 节实验的基础上, 为进一步验证集成方法在不同攻击强度下的表现, 设计了基于不同扰动强度的对抗攻击实验. 使用 FGSM 和 PGD 两种攻击方法, 扰动强度设置为 0.005–0.05. 针对 GoogLeNet、VGG 和 ResNet 模型生成不同对抗样本. 基准方案 (Baseline) 采用单一模型进行防御, 而 N-Model 组合模型则选取 3 个模型进行表决. 实验评估指标为分类准确率, 用于验证集成方法在不同扰动强度下的攻击容忍性.

表 2 FGSM 攻击下单一防御与 N-Model 防御实验对比 (%)

测试数据集	模型选择	Acc	ASR
针对GoogLeNet的 CIFAR-10图像对抗样本	Baseline GoogLeNet	19.33	66.27
	Baseline VGG	30.56	61.80
	Baseline ResNet	36.06	64.59
	N-Model	32.89	64.20
针对VGG的 CIFAR-10图像对抗样本	Baseline GoogLeNet	38.29	59.34
	Baseline VGG	10.03	77.57
	Baseline ResNet	41.22	63.68
	N-Model	32.87	66.08
针对ResNet的 CIFAR-10图像对抗样本	Baseline GoogLeNet	36.52	59.98
	Baseline VGG	31.79	61.26
	Baseline ResNet	17.67	73.45
	N-Model	31.71	64.13

表 3 PGD 攻击下单一防御与 N-Model 防御实验对比 (%)

测试数据集	模型选择	Acc	ASR
针对GoogLeNet的 CIFAR-10图像对抗样本	Baseline GoogLeNet	0.93	95.54
	Baseline VGG	29.99	65.50
	Baseline ResNet	33.71	69.15
	N-Model	21.95	77.79
针对VGG的 CIFAR-10图像对抗样本	Baseline GoogLeNet	35.69	60.64
	Baseline VGG	0.39	99.61
	Baseline ResNet	37.89	65.23
	N-Model	24.85	74.49
针对ResNet的 CIFAR-10图像对抗样本	Baseline GoogLeNet	34.14	62.96
	Baseline VGG	30.74	65.46
	Baseline ResNet	0.71	98.02
	N-Model	20.08	75.01

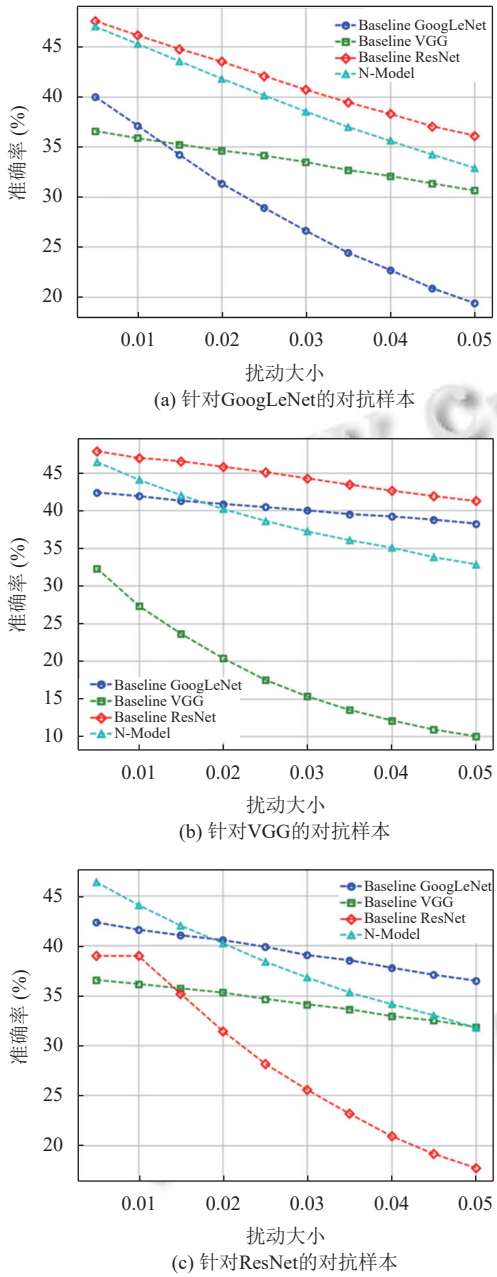
4.4.2 实验结果与分析

实验结果如图 6、图 7 所示, 在不同的攻击和对抗样本数据下, 单一模型的防御能力存在较大波动, 尤其在白盒攻击条件下, 针对特定模型的对抗样本, 单一模型的表现明显较差. 例如, 在 FGSM 和 PGD 攻击下, 单一模型 (Baseline) 在处理自身对抗样本时的准确率显著下降, 反映出其在特定攻击下的脆弱性. 此时, N-Model 组合模型通过引入多个模型的多样性和随机性, 有效缓解了单一模型在特定攻击下的性能退化问题.

这一现象也从侧面反映出, 在特定攻击环境中, 集成方法能够有效增加攻击者对模型的预测难度. 由于集成方法涉及多个模型的协作, 攻击者难以通过溯源分析的方法定位并攻击特定的模型. 与单一模型相比, 集成方法在任何攻击条件下都没有呈现最差的结果, 提供了更为稳定和可靠的防御.

图 8 中展示了各种对抗样本攻击下分类准确率的平均结果. 如图 8 所示, N-Model 组合模型无论是在

FGSM 攻击下还是 PGD 攻击下的表现均优于单一模型, 从侧面印证了其在应对复杂和混合攻击时的优势. 通过整合多个模型提升了对复杂攻击的容忍性, 增强了系统的鲁棒性和防御能力.



(a) 针对GoogLeNet的对抗样本

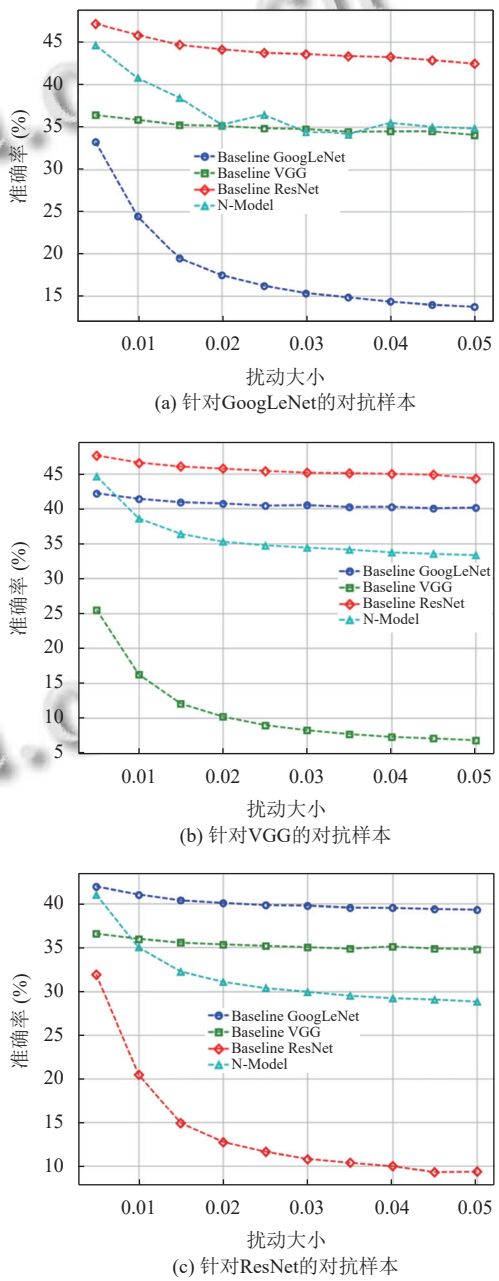
(b) 针对VGG的对抗样本

(c) 针对ResNet的对抗样本

图 6 在 FGSM 不同扰动强度攻击下对模型生成的对抗样本检测结果曲线图

5 结论与展望

本文提出了一种基于多个深度学习模型动态组合的方法, 实现模型的多样性和随机性, 通过模型的动态变化增加智能攻击对象及攻击途径的不确定性, 结合多模型的表决机制, 避免单一深度学习模型功能安全脆弱性, 增强深度学习智能系统的鲁棒性和攻击容忍性. 理论分析证明, N-Model 机制可增强系统在对抗攻击下的安全弹性, 同时在单个模型性能下降时维持较高的期望准确率. 实验结果进一步验证了理论结论, N-Model 在复杂对抗攻击场景下表现出优于单一模型的防御性能, 降低了攻击成功率. 未来将致力于优化模型选择与集成策略, 并拓展至更复杂的任务和实际应用场景, 以进一步提升安全性和实用性.

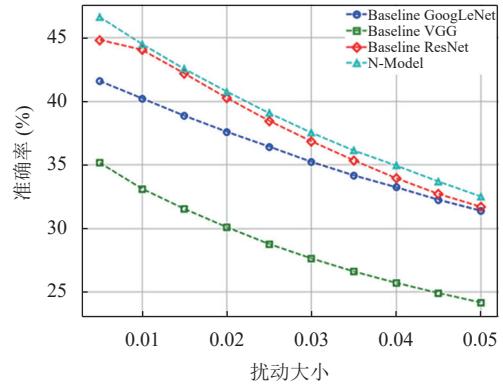


(a) 针对GoogLeNet的对抗样本

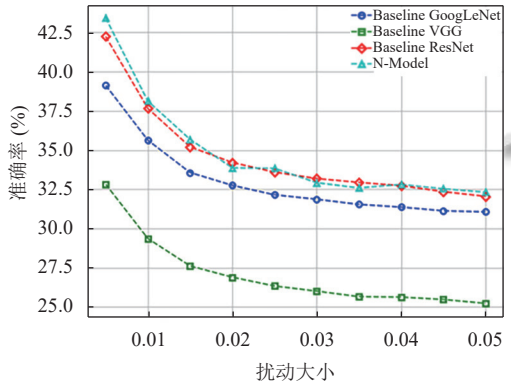
(b) 针对VGG的对抗样本

(c) 针对ResNet的对抗样本

图 7 在 PGD 不同扰动强度攻击下对模型生成的对抗样本检测结果曲线图



(a) FGSM攻击下平均准确率与扰动大小关系



(b) PFD攻击下平均准确率与扰动大小关系

图 8 不同扰动强度攻击下对模型生成的对抗样本检测平均结果曲线

参考文献

1 刘苏雅. 全球人工智能安全指数发布, 我国处于国际第一梯队. https://news.bjd.com.cn/2025/02/08/11060354.shtml. (2025-02-08) [2025-02-15].

2 Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv:1412.6572, 2014.

3 Eykholt K, Evtimov I, Fernandes E, *et al*. Robust physical-world attacks on deep learning visual classification. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1625–1634.

4 Finlayson SG, Bowers JD, Ito J, *et al*. Adversarial attacks on medical machine learning. Science, 2019, 363(6433): 1287–1289. [doi: 10.1126/science.aaw4399]

5 Madry A, Makelov A, Schmidt L, *et al*. Towards deep learning models resistant to adversarial attacks. arXiv: 1706.06083, 2017.

6 邬江兴. 网络空间拟态防御研究. 信息安全学报, 2016, 1(4): 1–10. [doi: 10.19363/j.cnki.cn10-1380/tn.2016.04.001]

7 张卓, 陈毓端, 唐伽佳, 等. 基于威胁的网络安全动态防御研究. 保密科学技术, 2020(6): 22–31.

8 Liu K, Dolan-Gavitt B, Garg S. Fine-pruning: Defending against backdooring attacks on deep neural networks. Proceedings of the 21st International Symposium on Research in Attacks, Intrusions, and Defenses. Heraklion: Springer, 2018. 273–294.

9 Kurakin A, Goodfellow IJ, Bengio S. Adversarial examples in the physical world. Artificial Intelligence Safety and Security. New York: Chapman and Hall/CRC, 2018. 99–112.

10 Moosavi-Dezfooli SM, Fawzi A, Frossard P. Deepfool: A simple and accurate method to fool deep neural networks. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2574–2582.

11 Carlini N, Wagner D. Towards evaluating the robustness of neural networks. Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP). San Jose: IEEE, 2017. 39–57. [doi: 10.1109/SP.2017.49]

12 Croce F, Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020. 2206–2216.

13 Bashivan P, Bayat R, Ibrahim A, *et al*. Adversarial feature desensitization. Proceedings of the 35th International Conference on Neural Information Processing Systems. Curran Associates Inc., 2021. 816.

14 Zhang SS, Hong HB, Xie MD, *et al*. Enhancing the transferability of adversarial examples with random diversity ensemble and variance reduction augmentation. IEEE Transactions on Big Data, 2025: 1–15. [doi: 10.1109/TBDATA.2025.3533892]

15 Brown TB, Mann B, Ryder N, *et al*. Language models are few-shot learners. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 159.

16 Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. Proceedings of the 35th International Conference on Machine Learning. Stockholm: PMLR, 2018. 274–283.

17 Chen HJ, Ji YF. Adversarial training for improving model robustness? Look at both prediction and interpretation. Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2022. 10463–10472.

18 Zhang HY, Yu YD, Jiao JT, *et al*. Theoretically principled trade-off between robustness and accuracy. Proceedings of the 36th International Conference on Machine Learning.

Long Beach: PMLR, 2019. 7472–7482.
19 Cohen JM, Rosenfeld E, Kolter JZ. Certified adversarial robustness via randomized smoothing. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 1310–1320.
20 Carlini N, Wagner D. Adversarial examples are not easily detected: Bypassing ten detection methods. Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. Dallas: ACM, 2017. 3–14.
21 Tramèr F, Carlini N, Brendel W, *et al*. On adaptive attacks to adversarial example defenses. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 138.
22 秦臻, 庄添铭, 朱国淞, 等. 面向人工智能模型的安全攻击和防御策略综述. 计算机研究与发展, 2024, 61(10): 2627–2648. [doi: 10.7544/issn1000-1239.202440449]
23 Gao YS, Doan BG, Zhang Z, *et al*. Backdoor attacks and countermeasures on deep learning: A comprehensive review. arXiv:2007.10760, 2020.
24 Cai GL, Wang BS, Hu W, *et al*. Moving target defense: State of the art and characteristics. Frontiers of Information Technology & Electronic Engineering, 2016, 17(11): 1122–1153.
25 Rehman Z, Gondal I, Ge MM, *et al*. Proactive defense mechanism: Enhancing IoT security through diversity-based moving target defense and cyber deception. Computers & Security, 2024, 139: 103685.
26 Sengupta S, Chowdhary A, Sabur A, *et al*. A survey of moving target defenses for network security. IEEE Communications Surveys & Tutorials, 2020, 22(3): 1909–1941. [doi: 10.1109/COMST.2020.2982955]
27 Wu J. Cyberspace endogenous safety and security. Engineering, 2022, 15: 179–185.
28 Qin R, Wang L, Chen X, *et al*. Dynamic defense approach for adversarial robustness in deep neural networks via stochastic ensemble smoothed model. arXiv:2105.02803, 2021.
29 Waghela H, Sen J, Rakshit S. Adversarial robustness through dynamic ensemble learning. Proceedings of the 2024 IEEE Silchar Subsection Conference (SILCON). Agartala: IEEE, 2024. 1–6. [doi: 10.1109/SILCON63976.2024.10910654]

(校对责编: 王欣欣)