深度学习在恶意软件检测中的应用

冯沛林 宋稷昀 邓正伟 吴伟毅

联通数字科技有限公司，北京市，100000；

摘要：恶意软件的持续演化使传统检测机制面临显著挑战。随着多态变形、代码混淆及跨平台传播手段的复杂化，基于签名和规则的检测策略逐渐失效。深度学习因其自动表征、高维建模与对未知样本的泛化能力，为恶意软件检测提供新路径。本文围绕数据建模、模型架构、学习策略与部署治理四方面，系统分析深度学习技术的实际应用逻辑与挑战应对，并结合典型实验案例与行业数据，验证其在真实场景中的可行性与稳健性。文章最终指出，实现模型“可控、可解释、可持续”将是未来发展的关键方向。

关键词：恶意软件检测；深度学习；模型部署

DOI：10.69979/3041-0673.26.02.024

引言

随着信息技术的快速发展，恶意软件的传播形式与攻击手段日益复杂化，传统的基于签名匹配和启发式分析的检测方法已难以应对未知变种。据统计，2023 年全球恶意软件变种数量突破 5.2 亿个，其中针对移动端和物联网设备的攻击占比超过 60%，对个人隐私、企业资产乃至国家安全构成严峻挑战。针对恶意软件的检测技术层出不穷，但仍有多源数据特征提取、特定场景模型优化等挑战性的问题需要解决[1]。

1 恶意软件检测的发展现状与深度学习技术引入的契机

1.1 恶意软件的演化与检测困境

近十年，恶意软件呈现产业化与生态化并进的格局，技术路径不断分叉：多态变形、加壳与代码虚拟化频繁迭代，文件无痕与内存注入隐蔽运行，借助系统内置工具的“借船出海”策略蔓延至终端与云边侧；供应链投毒与跨平台投递扩展攻击半径。加密流量、域前置与隐写通信削弱网络侧线索，环境感知与延时触发压低沙箱命中率[2]。工程侧又面临数据分布漂移、样本极度不平衡与标签噪声，移动与物联网场景的遥测合规限制使高质量样本稀缺。签名过期速度快，行为迹象离散而脆弱，单点证据难以支撑稳定判断，传统流程在覆盖、时效与成本三角中反复取舍。检测体系急需具备自动抽象表征与持续适应能力的技术范式，以缩短对抗周期并稳住误报与漏报水平[3]。

1.2 传统检测技术的瓶颈

基于签名与启发式规则的框架依赖专家经验与先验样本，维护开销高，面对混淆与变体易失效。静态分析受压缩、重定位与控制流平坦化影响，语义割裂明显；动态分析成本与风险偏高，沙箱逃逸与行为延迟使覆盖率受限。传统机器学习依靠手工特征，如操作码 n-gram、API（Application Programming Interface，应用程序接口）频次与包流统计，表达能力受限，跨系统与跨版本迁移时常发生特征漂移。零样本与近零样本场景下，规则堆叠难以保持一致性，阈值调优牵一发而动全身，实时性与准确性经常此消彼长。异构数据源之间缺乏统一表示，导致线索整合停留在后验加权层面，难以形成可扩展的全局判别器。因而传统体系更适合作为高置信“基线层”，亟待与面向表示学习的智能模块协同重构。

1.3 深度学习的技术优势

深度学习以表示学习为核心，能够在端到端流程中自动提取高阶语义特征：卷积神经网络擅长从二进制灰度图提炼局部结构，循环神经网络及长短期记忆网络捕捉序列依赖，Transformer 凭注意力机制刻画长程关系，图神经网络刻画 API 调用图与进程关系图的拓扑约束。自监督预训练与迁移学习缓解标注稀缺，联邦学习在隐私边界内汇聚异地知识，知识蒸馏压缩模型以适配终端侧部署，对抗训练提升对扰动与规避策略的韧性。注意力权重、梯度归因与决策可视化为审计提供抓手，使“可解释—可控—可演进”的闭环成为可能[4]。由此，DL

为恶意软件检测带来表达统一、跨模态融合与持续学习的抓手，为建设韧性更强的智能引擎奠定技术基底。

2 深度学习模型在恶意软件检测中的应用路径

2.1 数据建模与表示空间

应用深度学习的前提在于把零散迹象沉淀为可学习的表示。原始二进制、操作码序列、API 调用轨迹、系统调用图、内存转储与网络流量，分别承载结构、时序与关系三类线索。实践中更有效的方案是构建“比特—行为—关系”三层表示：比特层强调二进制纹理与指令分布，适合卷积网络捕捉局部模式；行为层刻画调用顺序与状态转移，便于循环网络或变压器架构建模长程依赖；关系层将进程、文件、注册表与域名抽成图，交由图神经网络复原传播路径。高价值样本往往稀缺，标签噪声又难以避免，故需引入弱监督与自监督预训练稳定表征质量。针对分布漂移，可保留一小段“保守特征”（如稳定 API 子图）作为锚点，减轻版本迁移带来的不适配。由此，数据进入模型前已在统一坐标系内对齐，后续推断不再被单一视角束缚。

2.2 模型架构与场景耦合

模型选择不取决于“谁最强”，而取决于“哪类证据最可信”。将二进制映射为灰度图时，卷积神经网络擅长识别壳层布局与代码块重复；面对 API 或系统调用序列，长短期记忆网络与变压器在异常片段的上下文刻画上更具辨识力；涉及多实体交互的复杂样本，图神经网络适于建模节点角色与边类型的细微差异。工程侧常采用“轻端—重云”的分层推理：端侧以蒸馏压缩后的小模型完成快速拦截，云端集成多模态大模型做深度复核，并给出可解释标注以指导回溯。对于高价值场景（如供应链投递），可将静态与动态证据拼合为多通道输入，设置共享底座与任务专头，既确保通用性，又保留对细分威胁的灵敏度。

2.3 学习策略与稳健性构建

恶意样本与良性样本比例失衡，且族内差异显著，单纯追求准确率往往掩盖风险。更可取的策略是围绕召回、校准与稳健性三点同时设计：采用焦点损失或代价敏感权重，提升长尾类别的可见度；借助对比学习与掩码建模进行自监督预训练，缓解标签稀缺；在增量学习过程中设置“冻结层+小步调优”，避免遗忘早期知识。规避检测的对抗样本日渐增多，需在训练阶段引入扰动约束与随机化增强，提高模型在输入微扰、控制流平坦化与时序稀疏下的韧性；推理阶段配合置信度校准与开放集识别，将不熟悉的样本外包给更谨慎的复核通道，降低误伤成本。量化、剪枝与知识蒸馏用于压缩模型体量，保障终端部署的实时性，不牺牲关键特征对抗能力。

2.4 部署治理与闭环演进

模型落地不是一次性发布，而是持续治理。数据入口需配套来源审计与脱敏策略，联邦学习在合规边界内汇聚多域知识，既避免原始数据集中，又保留多样性。线上监控围绕三类指标展开：检测质量、延迟与资源占用，辅以漂移探测与回滚机制，防止突发事件扩散。解释性工具（如注意力热力图与梯度归因）给审计团队提供可核查证据，便于复盘与规则同步更新；人机协同的复标流程帮助模型在真实流量中稳步校准。为避免策略“越配越重”，建议建立“基线规则—表示模型—判别模型—事后策略”四层架构，职责清晰、耦合度低，便于替换与 A/B 评估。长期看，效果最好的系统往往并不炫技，而是在数据质量、反馈时效与治理流程上形成良性循环，持续把经验沉淀回模型与特征库，令检测能力在对抗中稳步进化。

3 实证分析与行业应用现状

3.1 案例研究

为避免流于形式的性能比对，本节选取《智能系统学报》（2024）中一项具有代表性的研究作为案例基础，重点分析基于图像特征的高维卷积模型与“序贯三支决策”机制结合后的效果表现。该研究通过将恶意软件二进制文件转换为灰度图像，赋予模型以视觉特征识别能力；同时引入延迟判决的三支决策策略，对模型不确定性区域设置“边界域”缓冲，使判别过程更加稳健。该策略针对易混淆或近邻样本，先缓决策、后精判别，规避因特征模糊而带来的误杀与漏报。实验在 Kaggle 和 Leopard Mobile 两个公开数据集上开展，评估指标涵盖准确率、精确率、召回率与假阳性率，核心结果见表 3-1 与表 3-2（数值均来自原文对照实验）。

表 3-1 Kaggle 数据集上不同模型的对比（%）

分类方法	准确率	精确率	召回率	假阳性率
DRBA（基于蝙蝠算法的动态采样）	94.96	96.06	93.76	3.85
多目标 CNN	96.23	93.91	98.86	6.41
多目标 RBM	95.90	95.65	96.18	4.38
MO-STWD（高维多目标序贯三支）	98.06	97.43	98.87	2.61

*数据来源：智能系统学报 2024，19(1):97-105[5]。

表 3-2 Leopard Mobile 数据集上的对比（%）

分类方法	准确率	精确率	召回率	假阳性率
DRBA	94.53	92.62	96.85	7.83
多目标 CNN	95.56	94.14	97.01	6.16
多目标 RBM	96.17	96.09	96.11	3.78
MO-STWD	96.88	96.74	95.57	2.17

*数据来源：同上。

尽管两个数据集在指标表现上略有差异，但核心趋势保持一致：MO-STWD 在假阳性控制上显著优于其他模型，且整体准确率与召回率保持高位。特别是在高风险场景下，降低误报意味着更少的正常行为被拦截，提升用户体验与系统可信度。若业务目标更偏向“少误杀”，则该模型结构尤具优势；而对于极端召回需求，可考虑与人工复核机制联动使用。该方法的核心价值，在于主动管理不确定性，为快速演化的恶意软件家族提供应对策略，避免因模型“过早下结论”导致的误判风险。

3.2 行业实践

将模型性能从实验室推向实际应用，离不开对抗环境的压力测试与业务真实数据的反馈。围绕勒索软件检测与响应，360 公司在《2023 年勒索软件流行态势报告》中披露了一组具代表性的行业实战数据，构成对深度学习模型工程落地价值的重要观测视角。据报告，2023 年全年共受理反勒索求助案例 2750 余起；其中，Phobos、BeijingCrypt、TellYouThePass 三大勒索家族占比超过 51%，攻击频率高度集中。赎金要求多数在 10 万至 100 万美元区间，数据泄露体量常见于 10GB 以上，甚至不乏单次窃取超过 500GB 的大规模事件。受害对象主要聚焦于科研服务、零售与制造业，受害地区则呈现明显地域集中性，广东、江苏与北京为高发区域。CNCERT 公布的流量侧监测显示，移动端勒索样本仍在以周为单位快速迭代：仅 2023 年 2 月第 2 周即捕捉到 10 个新变种，历史累计样本已突破 5500 个，传播渠道呈现“应用市场+仿冒下载页”双轨化趋势。

表 3-3 国内行业实践的观测指标（2023 年）

指标	数值 / 描述	来源备注
反勒索求助处理量	2750+ 起	360 年度报告
主要勒索家族占比	Phobos、BeijingCrypt、TellYouThePass 合计 >51%	家族频次统计
赎金金额分布	10 万 - 100 万美元区间占比 >70%	攻击经济参数
数据泄露体量	>10GB 为多数，>500GB 占比 >20%	事件影响范围
受害行业分布	科研技服、零售、制造为主要受害领域	行业画像
受害地区分布	广东、江苏、北京高频发生	区域热点分布
移动端周新增勒索样本数量	一周新增 10 个新变种	CNCERT 周报
移动端勒索样本历史累计数量	5505 个样本	CNCERT 累计统计

从工程部署角度看，现实攻防格局促使厂商普遍采用“端轻云重”结构，即将轻量化深度模型部署至本地终端承担初步筛选任务，而复杂模型与多模态融合逻辑则保留在云端用于高置信复核。同时，为提升检测精准度与应变能力，部分厂商已尝试将威胁情报回流机制集成进模型训练环节：例如通过赎金区间、家族频率与受害行业画像，动态调整样本采样策略与损失函数设计，使模型训练更加贴近真实攻击分布，而非理想化数据环境。这种“模型—情报—运营”的闭环路径，代表了深度学习模型在恶意软件检测中工程实用性的关键突破方向。

3.3 风险与伦理考量

深度学习模型在提升恶意软件检测能力的同时，也带来新的技术与伦理边界问题。安全对抗维度中，模型对微扰输入的敏感性使其在对抗样本面前易失守，尤其是在控制流平坦化、语义混淆与环境感知策略干扰下，传统特征可能迅速失效，导致“伪良性”逃逸实际检测。若训练过程缺乏对抗增强、置信度校准与开放集识别机制，误判风险将被系统性放大。

数据治理层面问题同样显著。训练数据往往包含终端遥测、用户行为及网络交互内容，这些数据在采集、脱敏与跨境使用过程中极易触发隐私与合规争议。2023 年施行的《生成式人工智能服务管理暂行办法》提出“发展与安全并重、分类监管”原则，为模型部署划定合规边界，同时要求企业强化算法审计、数据评估与风险告警机制，确保模型在安全与可控框架下运行。

社会责任角度，模型自动化判断若缺乏决策可解释性，极易引发用户信任危机与企业法律风险。一旦出现误判，无责任归属机制将使问题失控放大。因此构建“稳健—可解释—可追责”的模型治理体系至关重要：训练阶段应引入扰动控制与置信分层机制，线上部署需设置复核通道处理“边界样本”，离线系统则应配备审计工具，实现基于注意力机制与归因图谱的决策回溯，从而确保模型行为可查、可溯、可控。

4 结语

本文系统梳理深度学习在恶意软件检测中的技术路径，强调多源数据建模、模型结构与业务场景的深度耦合、鲁棒性机制的构建以及部署端的闭环治理逻辑。通过对公开研究与行业实战数据的交叉验证，验证该技术体系在复杂对抗环境下的实用潜力。同时，文章指出当前模型仍面临对抗失效、隐私合规与可解释性不足等挑战。未来研究可进一步探索跨模态知识融合、轻量化部署与人机协同策略，推动智能检测系统从“可用”迈向“可信”。

参考文献

[1]江超凡.基于深度学习的恶意软件检测技术研究[D].杭州电子科技大学,2025.

[2]谢丽霞,魏晨阳,杨宏宇,等.基于图像化方法的恶意软件检测与分类综述[J].计算机学报,2025,48(3):650-674.

[3]范铭,刘烃,刘均,等.安卓恶意软件检测方法综述[J].中国科学:信息科学,2020,50(8):1148-1177.

[4]李英华,朱景怡.深度学习在网络安全恶意软件检测中的应用[J].无线互联科技,2025,22(14):85-88.

[5]崔志华,兰卓璇,张景波,等.基于高维多目标序贯三支决策的恶意代码检测模型[J].智能系统学报,2024,19(1):97-105.