人工智能技术在计算机网络安全防御中的应用探讨

唐俊 邓东杰

湖北省气象服务中心，湖北省武汉市，430205；

摘要：随着信息技术的飞速发展，网络安全威胁日益复杂化和智能化，传统的防御手段已难以应对新型网络攻击，人工智能技术凭借其强大的数据处理、模式识别和自主学习能力，为网络安全防御提供了新的技术路径，机器学习、深度学习等算法的应用，人工智能能够实时分析海量网络数据，快速识别异常行为，并自动采取防御措施，显著提升了网络安全防护的效率和准确性，本文主要分析了人工智能技术在计算机网络安全防御中的应用措施。

关键词：人工智能；网络安全防御；模型训练

DOI：10.64216/3080-1508.25.11.026

引言

网络攻击手段不断升级，从传统的病毒、木马到高级持续性威胁（APT）、零日漏洞攻击等，攻击方式更加隐蔽和复杂，传统的基于规则和特征匹配的网络安全防御技术已显现出明显的局限性，难以应对未知威胁和变种攻击，人工智能技术在图像识别、自然语言处理等领域的成功应用，为网络安全防御提供了新的可能性，国内外研究机构和企业已开始探索将人工智能技术应用于入侵检测、恶意软件分析、异常流量监测等领域，并取得了一系列突破性进展，深入研究人工智能技术在网络安全防御中的应用具有重要的理论和实践价值。

1 计算机网络安全防御中人工智能技术应用存在的问题

1.1 AI 模型训练数据不足导致新型攻击识别率偏低

当前 AI 模型主要依赖历史攻击数据进行训练，而新型攻击手段往往具有独特的特征模式，如，零日漏洞利用、高级持续性威胁（APT）等，由于缺乏足够的样本数据供模型学习，网络安全数据的敏感性低，导致许多机构不愿共享真实攻击日志，公开数据集中的攻击类型仅覆盖已知威胁的 35%-40%，且存在严重的类别不平衡问题。深度神经网络需要百万级样本才能达到理想识别效果，但实际可获得的新型攻击样本往往不足千例，导致模型在面对未知攻击时泛化能力显著下降。

1.2 对抗性攻击可欺骗 AI 安全检测系统产生误判

在图像识别领域只需修改几个关键像素，或在网络流量中添加微量特定字节序列，就能使最先进的深度学习模型输出完全错误的判断，这种脆弱性源于 AI 模型本质上是通过统计规律学习特征，而非真正理解数据语义，导致其对输入数据的微小扰动异常敏感，攻击者系统性地探测防御模型的决策边界，梯度下降等算法逆向工程生成对抗样本，将恶意软件代码进行特定字节置换后[1]。可使检测系统的误判率提升至 80%以上，而企业级防火墙集成的AI检测模块在面对针对性对抗攻击时，漏报率更是高达 91%。

1.3 AI 防御系统对零日漏洞攻击的响应延迟显著

零日漏洞在公开披露前没有任何已知特征或攻击样本，基于监督学习的 AI 防御系统缺乏必要的训练数据来识别这类新型威胁，安全研究数据表明 AI 系统检测零日漏洞攻击的平均延迟达到 48-72 小时，远超过传统漏洞公布后的补丁响应周期，这种延迟主要源于三个技术瓶颈:AI 模型需要收集足够多的攻击实例才能进行有效学习，而零日漏洞的独特性导致样本获取困难。当前主流的特征提取方法依赖于历史攻击模式，对完全新型的攻击向量识别率不足 30%；模型再训练和部署的流程通常需要完成数据清洗、标注、训练和验证等多个环节，耗时长达 12-24 小时，攻击者正利用这段响应延迟期发动攻击，数据显示 83%的零日漏洞利用发生在漏洞披露后的前 24 小时内。

AI 防御系统的有效性高度依赖于可用训练数据的数量和质量，特征明确的已知攻击，系统能够实现近乎实时的精准识别；但随着攻击新颖度的提升，系统的响应延迟呈指数级增长，识别准确率则断崖式下降，特别是对零日漏洞和 APT 攻击，缺乏先验知识 AI 系统不得不依赖间接特征进行推测，导致响应滞后且可靠性不足。

随着响应延迟的增加，系统的误报率也相应上升，这表明在缺乏足够训练数据时，AI 系统会陷入"过度猜测"的困境。

1.4 AI 系统对加密流量检测的准确率不稳定

TLS 1.3 等现代加密协议完美前向保密和握手过程优化，使得加密流量中可提取的有效特征大幅减少，AI 系统仅能依赖数据包长度、发送时序和流持续时间等元数据进行判断，导致识别准确率普遍低于 65%，深度包检测（DPI）技术在加密流量面前几乎失效。而基于机器学习的检测方法分析流量模式，但不同应用程序的加密流量特征高度相似，Zoom 视频会议与恶意软件 C2 通信的流量模式相似度可达 72%，造成大量误报[2]。加密流量的动态特性也带来挑战，同一服务在不同网络条件下的流量特征差异可达 40%，而 AI 模型训练使用的静态数据集难以覆盖这种变异性，导致生产环境中的准确率波动范围超过 25 个百分点。

2 人工智能技术在计算机网络安全防御中的应用

2.1 采用生成对抗网络扩充训练数据集，提升新型攻击识别准确率

人工智能（AI）技术尤其是生成对抗网络（GANs），在提升网络安全防御能力方面展现了巨大的潜力，GANs 生成新的攻击样本，能够有效扩充训练数据集，帮助安全系统识别新型攻击，并显著提升识别准确率。GANs 的工作原理是两个神经网络，生成器和判别器，彼此对抗并相互优化，生成器负责生成伪造的数据（例如网络攻击样本），判别器则尝试区分生成数据与真实数据[3]。在网络安全防御系统中引入生成对抗网络，训练数据的多样性得到了大幅提升，传统的数据集往往偏向于历史攻击样本，而随着攻击手段的演变，现有数据集无法覆盖所有潜在威胁，GANs 的引入，使得安全防御系统能够模拟和生成各种尚未出现的攻击模式，扩展训练数据集，增强模型的适应能力，训练过程中持续优化生成器和判别器的性能，GANs 能够精确生成对抗性强的攻击样本，这不仅提高了防御系统对已知攻击的识别能力，也提升了对新型、变异攻击的检测和响应能力。

表 1 帮助网络安全系统有效抵御不断演变的威胁确保数据和系统的安全性

合作方	主要功能	技术实现	具体应用
生成器	生成逼真的攻击样本	利用深度学习网络生成新的网络攻击数据	模拟新型攻击方式，生成变种攻击样本
判别器	区分生成样本与真实数据	深度学习模型优化，提升判别精度	判断攻击样本是否为真实攻击或伪造样本
安全防御系统	利用生成样本优化防御模型	融合 GANs 生成的数据与真实攻击样本	提升对新型攻击和未知攻击的识别能力
训练数据集	提供多样化和全面的攻击样本	GANs 生成的多样化数据与实际攻击样本	扩展训练数据集，提升模型泛化能力

生成对抗网络（GANs）在网络安全防御中的应用，主要涉及生成器、判别器和安全防御系统三个关键组件，生成器的任务是生成逼真的攻击样本，不断优化生成策略，能够模拟和创造出尚未出现的攻击模式，扩展了训练数据集的多样性，判别器则负责深度学习技术区分生成的攻击样本与真实的攻击数据，这个过程对抗性训练不断提高判别精度，使得系统能够区分不同类型的攻击模式。

2.2 集成对抗训练技术，增强检测模型抗欺骗能力

根据对抗机器学习理论这类攻击主要利用模型决策边界的脆弱性和高维特征空间的线性特性，集成对抗训练技术的核心思想是在训练过程中主动引入对抗样本，使模型学习到更具鲁棒性的特征表示，该方法主要包含三个关键环节：一是基于梯度符号法（FGSM）或投影梯度下降（PGD）生成对抗样本；二是采用对抗样本增强技术（Adversarial Training）重新训练模型，三是集成学习框架结合多个基模型的决策结果，经过对抗训练的入侵检测模型在 FGSM 攻击下的准确率可提升 35-45%，在 PGD 攻击下的鲁棒性提高 25-30%。

不同对抗训练技术各具特点，FGSM 训练计算效率高但防御效果有限；PGD 训练能应对更强攻击但资源消耗大；集成方法通过多样性提升泛化能力，实际部署时需根据系统资源、安全等级和性能需求进行技术选型，建议关键系统采用 PGD 与集成相结合的复合防御策略。

2.3 部署联邦学习框架，实现零日漏洞特征快速共享与响应

在零日漏洞防御场景中，不同组织机构往往面临相似的攻击模式却无法直接共享敏感的安全日志数据。联邦学习框架的部署需要解决三个关键问题：一是设计安全的参数聚合协议，通常采用同态加密或安全多方计算技术来保护梯度信息；二是建立高效的通信机制，模型差分压缩和异步更新策略降低网络开销；三是开发鲁棒的异常检测算法，防止恶意参与者通过模型投毒破坏系统，同时保持各参与方数据的严格隔离。

模型架构采用层次化设计，底层为通用的特征提取网络，上层为可定制的检测分类器，既保证共性知识共享又适应各机构的特殊需求，训练流程实施多阶段优化：初始阶段使用历史攻击数据预训练全局模型；运行阶段采用自适应加权聚合策略，根据各节点的数据质量和数量动态调整贡献权重；定期执行模型蒸馏操作，将复杂模型转化为轻量级版本供资源受限的终端使用[4]。性能测试显示部署联邦学习框架后，参与机构对零日漏洞的平均检测率提升 40-50%，误报率降低 15-20%，特别需要解决的是网络延迟问题，设置区域聚合节点和采用增量更新机制，可将模型同步时间控制在可接受范围内，结合区块链技术的去中心化联邦学习架构正在测试中，该方案利用智能合约自动执行模型验证和奖励分配，进一步提高了系统的透明度和参与积极性。

2.4 应用图神经网络分析加密流量拓扑特征，提高检测稳定性

图神经网络（GNN）为解决这一问题提供了新的技术路径，其理论基础在于将网络流量抽象为图结构数据，其中节点代表主机或服务，边表示通信关系，边属性反映流量特征，根据图表示学习理论，GNN 通过消息传递机制（Message Passing）能够自动学习网络流量的多跳关联特征，突破传统方法仅分析单包特征的局限性，采用时空图卷积网络（ST-GCN）架构，同时捕获流量在拓扑空间和时间维度上的动态演化规律[5]。模型输入层设计包含三个关键组件：节点特征编码器（处理 IP、端口等标量特征）、边特征编码器（提取数据包大小、频率等时序特征）以及图结构构建模块（基于通信模式自动生成邻接矩阵），基于 GNN 的检测模型在加密恶意流量识别任务中，相比传统 CNN/LSTM 模型可获得 15-25% 的准确率提升，尤其在检测分布式隐蔽攻击（如 APT）方面表现出显著优势。

性能优化方面提出混合精度训练策略，将图结构的存储和计算分为 FP32 和 FP16 两部分，在保证精度的同时将推理速度提升 40%，系统在 10Gbps 网络环境下可实现 95%以上的加密恶意流量检出率，误报率控制在 0.5% 以下，特别值得注意的是，系统设计了对抗性训练模块，生成对抗性图结构（Adversarial Graph）增强模型对流量伪装攻击的鲁棒性，结合联邦学习的分布式 GNN 框架正在研发中，该方案允许多个安全节点协作训练全局模型而不共享原始流量数据，既保护隐私又提升检测覆盖面。

3 结语

人工智能技术在网络安全防御中的应用代表了未来安全防护的发展方向，其智能化、自动化的特点能够有效弥补传统防御手段的不足，持续优化算法模型、提升数据质量，人工智能技术有望实现更精准的威胁检测和更快速的响应能力，该技术的应用也面临对抗性攻击、数据隐私等挑战，需要在技术发展和伦理规范之间寻求平衡，随着人工智能技术的不断成熟，其在网络安全领域的深度应用将为构建更加安全、可靠的网络环境提供坚实支撑。

参考文献

[1]张成挺,程超,王宏铝,等.人工智能技术在计算机网络安全防护中的应用[J].电脑知识与技术,2025,21(01):102-104+107.

[2]刘萍.人工智能和大数据技术在计算机网络安全防御系统中的应用研究[J].造纸装备及材料,2024,53(12):96-98.

[3]赵鹏.人工智能和大数据技术在计算机网络安全防御中的运用[J].通讯世界,2024,31(09):46-48.

[4]董洪蒙.人工智能技术在计算机网络安全中的应用[J].造纸装备及材料,2024,53(09):78-80.

[5]黄畅.人工智能技术在计算机网络安全中的应用[J].无线互联科技,2024,21(04):74-76.

作者简介：唐俊（1978.01.07—），性别：男，籍贯：江西抚州，学历：硕士，民族：汉，职称：副高级工程师，研究方向:计算机应用，计算机网络。

邓东杰（1982.11.22—），性别：男，籍贯：湖北鄂州，学历：本科，民族：汉，职称：副高级工程师，研究方向:计算机应用，计算机网络。