基于深度学习的恶意加密流量检测及对抗技术综述

樊祖薇[1,2], 张顺亮[1,2], 赵泓策[1,2]

[1]中国科学院信息工程研究所 北京 中国 100093
[2]中国科学院大学 网络空间安全学院 北京 中国 10049

摘要 随着人们网络安全意识的不断提高和加密技术的广泛应用, 网络中的加密流量呈现爆炸式增长。在加密技术保护用户数据安全和隐私的同时, 攻击者也可滥用加密技术隐藏恶意、非法、窃密行为, 给网络安全防护及监管带来新的挑战。一方面, 在不解密条件下对恶意加密流量进行检测已成为网络安全领域的难题。随着恶意加密流量的不断增多, 传统的深度包检测技术已不再适用。另一方面, 攻击者利用流量混淆等攻击技术将恶意流量隐藏于正常流量之中, 或者利用对抗机器学习生成对抗样本以干扰检测模型, 误导检测系统做出错误决策。目前, 将深度学习方法应用于恶意加密流量检测以及对抗方面的研究不断发展, 尚未有文献对最新成果及趋势进行回顾。本文从任务场景、数据预处理、特征提取、模型和评估指标等多方面, 全面整理并分析了恶意加密流量检测及对抗技术的最新研究成果。首先, 提出了一个通用的恶意加密流量检测框架, 并结合框架对目标任务场景进行分类总结。其次, 介绍了应用于恶意加密流量检测的数据收集与预处理技术、特征提取与选择技术和相关的评估指标体系, 讨论了数据不平衡问题的解决方法。此外, 对比分析了不同检测模型的适用性和优缺点, 并讨论了对抗攻击和应对措施。最后, 探讨了恶意加密流量检测领域中开放问题和挑战, 并对未来的研究方向进行了展望。

关键词 加密流量; 恶意检测; 对抗攻击; 深度学习
中图法分类号 TP391.1 DOI 号 10.19363/J.cnki.cn10-1380/tn.2024.08.03

A Survey of Attack Discovery Technology Based on Host Events

FAN Zuwei[1,2], ZHANG Shunliang[1,2], ZHAO Hongce[1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract With the continuous improvement of people's awareness of network security and the wide application of encryption technology, the encrypted traffic in the network is emerging explosive growth. While encryption technology protects safety of user data and privacy, attackers can misuse encryption technology to hide malicious and illegal behaviors, which brings new challenges to network security protection and supervision. On the one hand, detecting malicious encrypted traffic without decryption has become a difficult issue in the field of network security. With the increasing amount of malicious encrypted traffic, traditional deep packet inspection techniques are no longer applicable. On the other hand, attackers use traffic obfuscation and other adversarial techniques to hide malicious traffic in normal traffic, or generate adversarial samples to interfere with the detection model, which misleads the detection system into making wrong decisions. At present, the research on applying deep learning methods to malicious encrypted traffic detection and confrontation is developing continuously, and there is no literature review on the latest achievements and trends. In this paper, the latest work of malicious encryption traffic detection and adversarial techniques are comprehensively investigated from the aspects of task scenarios, data preprocessing, features extraction, models and evaluation indicators. Firstly, a general framework for malicious encryption traffic detection is proposed, and the target task scenarios are classified according to the framework. Secondly, the system applied to malicious encryption traffic detection are presented from the perspectives of data collection and preprocessing techniques, feature extraction and selection techniques, and evaluation index, and the solutions to data imbalance problem are discussed. Moreover, the applicability, advantages and disadvantages of different detection models are compared and analyzed, and the techniques of adversarial attack and corresponding countermeasures are discussed. Finally, the open issues and challenges in the field of malicious encryption traffic detection are discussed, and the future research direction is prospected.

Key words encrypted traffic; malicious detection; adversarial attack; deep learning

通讯作者: 张顺亮, 博士, 高级工程师, Email: zhangshunliang@iie.ac.cn。
本课题得到国家重点研发计划项目(No. 2021YFB2910105)的资助。
收稿日期: 2023-02-25; 修改日期: 2023-05-30; 定稿日期: 2024-09-09

1 引言

1.1 背景

随着物联网、云计算、大数据等网络技术的快速发展，互联网规模不断扩大，网络流量爆炸式增长，人们的生产生活已经与互联网密不可分。根据中国互联网络信息中心(CNNIC)在京发布的第 50 次《中国互联网络发展状况统计报告》[1]中显示，截至 2022 年 6 月，我国网民规模达 10.51 亿，较 2021 年 12 月增长 1919 万，互联网普及率达 74.4%，较 2021 年 12 月提升 1.4 个百分点。其中，我国手机网民规模为 10.47 亿，较 2021 年 12 月新增 1785 万。截至 2022 年 6 月，我国网民中使用手机上网的比例为 99.6%；使用电视上网的比例为 26.7%；使用台式电脑、笔记本电脑、平板电脑上网的比例分别为 33.3%、32.6% 和 27.6%。可以发现，相较于使用固定主机，网民更多选择使用移动终端接入网络。随着我国 5G 网络规模的持续扩大，移动通信基础设施的升级换代，我国移动网络流量快速增长。如图 1 所示，2022 年上半年，我国移动互联网接入流量达 1241 亿 GB，同比增长 20.2%。截至 2022 年 6 月，我国国内市场上监测到的 APP 数量达 232 万款，移动网络流量呈现出复杂化和多样化的特点。

图 1 移动互联网接入流量[1]

Figure 1 Internet Traffic from Mobile Devices[1]

依托移动通信技术和移动终端技术的支持，移动网络相较于传统互联网有着更好的交互性、便携性以及更高的隐私保护要求。移动互联网改变了人们上网的空间及时间局限性，实现了真正意义上的随时随地接入使用，同时带动了大批新型移动应用的兴起，如移动支付、即时通信、在线办公、网络直播和网约车等。这些移动业务常与用户的隐私信息密切相关，如用户位置、通信录和交易密码等等。其次，移动业务拥有大量用户，会面临更多样化的攻击方式和更大的攻击规模。

1.2 恶意加密流量

最初，网络流量中的负载信息以明文传输，增加了被窥探的风险。然而，随着大众的网络安全意识不断提升，人们对于隐私保护和数据安全的需求不断提高。为保护互联网用户的信息安全，保证网络流量不被监听和利用，越来越多的加密协议替代了非加密协议，网络中的加密流量呈爆炸式增长。图 2[2]展示了 Google 产品和服务中的加密流量的增长趋势：从 2014 年的 50%左右增长到 2020 年后的 95%，增幅达到一倍左右。同时，Google 透明度报告显示[2]，世界排名前 100 的非 Google 网站中 97%都默认使用 HTTPS 协议，而这些网站流量约占全球所有网站流量的 25%，全球互联网走向加密时代已是大势所趋。

加密流量在一定程度上保证了用户隐私信息的机密性和完整性，但也掩盖了数据的特征，给网络恶意行为提供了庇护，加大了恶意流量的检测难度。恶意流量是进行网络恶意行为时所产生的流量，包括攻击、勒索、挖矿、信息窃取等，只要危害国家社会安全的均属于恶意流量。研究者依据不同的研究

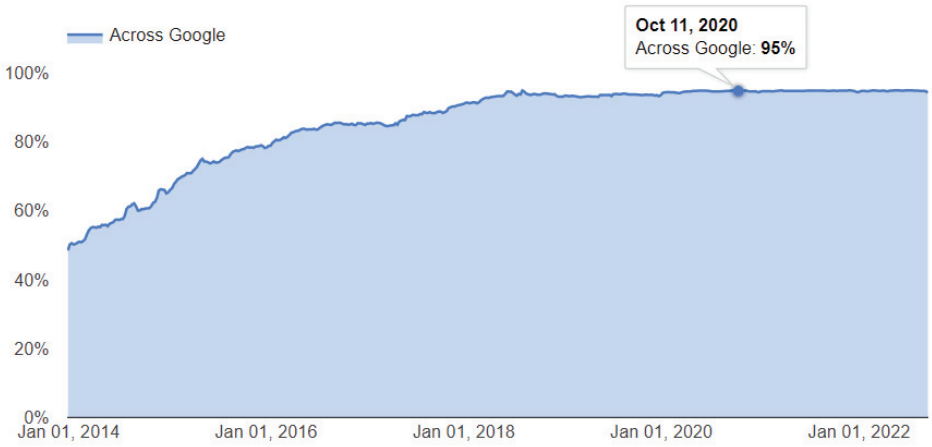

图 2 2014 年至 2022 年的谷歌加密流量变化趋势[2]

Figure 2 Trend of Google encrypted traffic from 2014 to 2022 [2]

方向, 对恶意流量的分类不同。文献[3]依据恶意流量的攻击效果将其分为以下 9 类: 信息窃取、远程控制、勒索、拒绝服务、信息探测、主动感染传播、欺骗攻击、信息篡改和下载式攻击。文献[4]依据恶意流量的用途将其分为恶意软件和恶意攻击两大类, 其中恶意软件类被细分为: 勒索软件、特洛伊木马、病毒、蠕虫、垃圾邮件; 恶意攻击类被细分为: 扫描攻击、暴力破解攻击、信息窃取、CC 攻击。

恶意加密流量是一种使用加密技术, 对恶意流量进行加密后传输的数据流, 主要可以分为以下四类: a)使用加密通信的恶意软件流量, 恶意软件为逃避安全检测, 会使用加密协议进行通信来伪装、隐藏明文流量特征。其中, 恶意软件是一类有目的地实现攻击者有害意图的应用程序, 包括间谍软件、勒索软件、木马、病毒、蠕虫和 Rookit 等。b)加密通道中的恶意攻击流量, 攻击者会利用已建立好的加密通道发起恶意攻击行为。其中, 恶意攻击行为包括: DNS 隧道攻击、扫描探测、暴力破解、CC 攻击、信息窃取、中间人攻击和网络钓鱼等。c)非法加密应用的通信流量, 如 Tor、非法 VPN 和非法翻墙软件进行通信时产生的流量。d)加密挖矿流量, 即对恶意挖矿行为进行加密的流量。我国明确要求加强虚拟货币“挖矿”活动上下游全产业链管理, 严禁新增虚拟货币“挖矿”项目。

目前, 越来越多的网络恶意行为通过加密和隧道技术绕过防火墙和入侵检测系统, 加密技术正成为恶意服务的温床[5-6]。ZSCaler 的加密攻击状态报告[6]显示, 利用加密进行的攻击呈持续上升趋势, 从 2020 年的 57% 上升到 2021 的 80%。2022 年时, 超过 85% 的攻击是加密的, 总攻击量比 2021 年高出 20%。同时, 自 COVID-19 时期开始, 许多公司转为居家办公, 为网络犯罪分子创造了新的目标, 个人智能手机和计算机受到恶意攻击的风险有所增加[7]。因此, 如何有效检测、处理恶意加密流量已成为亟待解决的问题。

1.3 恶意加密流量检测现状

恶意加密流量检测与非加密恶意流量检测的最大的不同之处就在于, 前者的实际负载内容不可见。然而, 对加密流量进行解密之后再检测的方法耗时长、成本高, 还涉嫌对用户隐私的侵犯。因此, 在不解密的条件下对加密流量进行有效检测是当前网络安全领域的重点之一。针对恶意加密流量检测, 传统的 DPI 方法[8]和基于端口的识别方法已不再适用, 目前的研究可以划分为以下三类: 基于规则的检测方法、基于传统机器学习(Machine Learning, ML)的检测方法和基于深度学习(Deep Learning, DL)的检测方法。

1) 基于规则的检测方法[8-14], 主要思想是利用加密流量的字段组合、排序或者固定模式等作为指纹进行模式匹配, 如流量载荷中的熵值[9]和加密通信中剩余的明文信息[10-11]。Wang 等人[9]通过计算流量载荷中的熵值来建立规则, 能够区分 8 种不同的加密流量, 但人工成本较高, 且易被绕过。综上所述, 基于规则的方法具有低误报这一优点, 但检测未知攻击的能力较弱。此外, 这种方法严重依赖规则库, 易被人工拼接或恶意伪造字段的流量绕过, 导致高漏报率。随着网络流量加密化进程的推进, 明文信息越来越稀疏, 基于规则的检测方法变得更加困难, 研究人员渐渐致力于将机器学习和深度学习方法应用到恶意加密流量检测的问题上。

2) 基于传统机器学习的检测方法, 主要思想是构建加密流量的各种特征联合作为机器学习模型的输入进行检测识别, 可以进一步分为有监督学习、无监督学习和半监督学习方法。在监督学习研究中, 研究者常使用多种监督学习模型进行对比实验[15-23], 包括朴素贝叶斯 (Naive Bayes, NB)、K-近邻(k-Nearest Neighbor, k-NN)、决策树(Decision Tree, DT)、支持向量机(Support Vector Machines, SVM)、随机森林(Random Forest, RF)、线性回归 (Linear Regression, LinReg)、逻辑回归(Logistics Regression, LogReg)、线性判别分析(Linear Discriminant Analysis, LDA)和 XGBoost 算法等。在无监督学习研究中, 研究者常使用 K-means 聚类算法和其它改进聚类算法对加密流量进行检测[24-30]。在半监督学习研究中, 研究人员使用少量带标签和大量不带标签的混合流量进行模型训练[31-35]。例如, Yuan 等人[34]将半监督学习和集成学习相结合, 提出半监督的 Adaboost 算法, 减少了检测时间, 解决了缺少带标记流量样本的问题。基于传统机器学习的检测方法有以下优点: a)通过对大量数据进行学习, 提供可靠决策, 有效提升检测准确率; b)机器学习算法的可解释性高; c)模型训练时间相对较短。但其仍有一定的局限性: a)是对流量特征的浅层学习; b)需要人工制作选择的特征, 依赖领域专家的经验, 泛化能力有限; c)部分流量特征容易过时, 需要不断更新。

3) 基于深度学习的检测方法, 主要思想是将原始流量数据或流量统计特征作为深度学习模型的输入, 自动化提取特征联系, 再进行分类或聚类检测。在恶意加密流量检测研究中, 常用的深度学习模型有多层感知机(Multilayer Perceptron, MLP)、卷积神

经网络(Convolutional Neural Networks, CNN)、循环神经网络(Recurrent Neural Network, RNN)、自动编码器(Auto Encoder, AE)和生成对抗模型(Generative Adversarial Networks, GAN)。除了上述常用的深度学习模型, 研究人员还使用基于图[36]、深度森林[37]等其它深度学习方法进行恶意加密流量检测研究。2020年, Wang 等人[36]通过对基于流和基于图的网络流量行为混合分析来检测僵尸网络, 最终实现检测精度达到 99.94%, 证明了基于图的分析方法可以在加密流量检测过程中充分利用通信数据之间的相关性。2022 年, Zhang 等人[37]针对 SSL/TLS 恶意加密流量, 提出一种基于深度森林的检测方法 DF-IDS, 先将网络流量按照五元组信息拆分为会话, 再将每个会话转换为二维图像作为深度森林的输入, 最终实现 SSL/TSL 恶意加密流量的细粒度多分类检测。基于深度学习的检测方法有以下优点: a)自动提取流量特征, 减少人工成本; b)深层学习, 发现流量特征间的非直观联系; c)在一定程度上解决数据不平衡问题, 实现较高的准确率和精度。但其仍存在一定的局限性: a)计算存储开销大, 大部分 DL 模型依赖大规模数据的长时间训练; b) DL 模型属于黑盒模型, 可解释性较差; c)现实网络中恶意流量的比例远远小于正常流量, 样本不平衡给深度学习方法带来的挑战仍需重视。

1.4 现有工作的局限性与不足

目前已有许多优秀的加密流量综述, 其中部分文献针对流量分类[38-40], 部分文献聚焦移动与无线网络领域[41-45], 部分文献与网络安全领域相结合[46-48], 但大都针对流量应用分类和加密协议识别等问题, 很少有文献深入聚焦于恶意加密流量检测领域[4, 49-50]。具体的文献综述分类总结如表 1 所示。

表 1 相关文献综述总结

Table 1 Summary of related survey works

类别	文献	年份	描述
加密流量分类	文献[38]	2015	介绍多项流量加密协议(IPsec、TLS、SSH、BitTorrent)在网络中的分组结构和标准行为, 关注加密流量分类任务中加密协议本身提供的信息, 重点介绍了基于传统机器学习的加密流量分类方法。
	文献[39]	2018	总结通过机器学习技术实现流量分类的一般过程, 针对流量分类过程的不同阶段进行文献归纳, 包括加密流量和非加密流量。
	文献[40]	2019	对加密流量分类的深度学习方法(MLP、CNN、RNN、AE、GAN)进行综述, 总结了一个基于深度学习的流量分类框架。
移动与无线网络领域	文献[41]	2018	回顾针对移动设备进行网络流量分析的研究, 定义了三个文献分类标准: 移动流量分析的研究目标、移动流量的捕获点和不同的移动平台。
	文献[42]	2019	对深度学习与移动和无线网络这两个领域进行交叉、调研, 详细介绍了深度学习的基本背景、原理、优势和相关先进技术, 讨论了多个有助于将深度学习有效地部署到移动系统上的技术和平台。
	文献[43]	2019	提出一个基于深度学习的移动业务加密流量分类的通用框架, 包括分类任务定义、数据准备、数据预处理、模型输入、预训练和模型结构设计。
	文献[44]	2020	对基于深度学习的移动流量分类研究进行系统地分类, 提出了一个基于深度学习的通用加密移动流量分类框架, 并基于该框架定义了流量对象、输入数据的类型、分类任务和深度学习架构。
网络安全领域	文献[45]	2021	对移动和无线网络中基于深度学习的网络安全研究进行回顾, 涵盖基础设施威胁和攻击、软件攻击和隐私保护三方面, 并总结相关深度学习算法(MLP、CNN、RNN、LSTM、AE、RBM、DBN)及原理。
	文献[46]	2019	介绍了基于深度学习的网络安全应用, 包括网络流量识别、网络入侵检测、恶意软件检测与内部威胁检测等。此外, 介绍了深度学习算法, 包括 DAE、DBM、RNN 和 GAN。
	文献[47]	2021	综述了用于网络流量监测和分析的深度学习方法, 同时提供了多种深度学习算法的详细定义和基本背景, 包括 MLP、CNN、LSTM、AE 和 GAN。
	文献[48]	2022	根据分析目标对机器学习驱动的加密流量分析研究进行分类, 包括网络资产识别、网络表征、隐私泄漏检测和异常检测四个目标类别。
恶意加密流量检测	文献[49]	2020	提出了一种用于加密恶意流量检测的“六步法”框架, 并在此基础上, 对基于深度学习的加密恶意流量检测方法进行了综述。
	文献[4]	2021	对加密恶意流量检测的研究进行综述, 划分了加密恶意流量的种类, 以具体检测技术为主线, 从数据采集、数据处理、模型训练和评价改进四个核心方面总结了目前的检测模型。
	文献[50]	2022	提出了一种基于机器学习的加密恶意流量检测框架, 并将五个公开数据集的部分样本组合成一个平衡的加密恶意流量数据集, 同时在新数据集上对十种不同的检测算法进行对比分析。

文献[38-40]总结了加密流量的分类方法。其中, 文献[38]侧重于介绍不同的加密协议, 以及这些协议在加密流量分类任务中本身能提供的信息, 重点介绍了基于传统机器学习的方法。之后, 文献[40]综述了用于加密流量分类的深度学习方法, 包括 MLP、CNN、RNN、AE 和 GAN, 并总结了一个基于深度学习的流量分类的通用框架。上述文献指出了传统机器学习和深度学习在加密流量分类任务中值得注意的问题和挑战, 但主要针对加密流量的分类问题, 没有重视恶意流量的检测问题。

文献[41-45]将机器学习和深度学习方法应用到移动和无线网络领域。其中, 部分文献针对移动设备上的流量分析[41], 部分文献侧重于深度学习与移动和无线网络领域间的交叉[42, 45]。文献[42]详细介绍了深度学习的基本背景、原理、优势和相关先进技术, 以及深度学习在移动网络领域中的多种应用。文献[45]涵盖了网络安全领域中的基础设施威胁和攻击, 软件攻击和隐私保护内容。然而, 这两篇文献对于加密流量检测的研究不够深入。文献[43-44]则对移动加密流量的分类问题进行探讨, 均提出了基于深度学习的移动加密流量分类框架, 并且对未来深度学习继续应用到加密流量分类中会出现的问题和挑战进行了展望, 但仍没有对恶意流量检测任务进行深入讨论。上述文献主要针对移动设备, 聚焦于移动和无线网络领域, 即使部分文献提及了恶意流量检测问题, 但只占总体内容的一部分, 没有进行深入探讨。而本文的综述不仅关注移动和无线网络中的恶意加密流量检测, 还分析了移动设备以外的平台。

文献[46-48]介绍了以机器学习或深度学习为基础的网络流量分析方法, 包括加密和非加密流量。其中, 文献[46]介绍了基于深度学习的网络安全应用, 包括网络流量识别、网络入侵检测、恶意软件检测与内部威胁检测等, 涵盖了广泛的攻击类型, 包括恶意软件、垃圾邮件、内部威胁、网络入侵、虚假数据注入与僵尸网络等。然而, 缺少对加密流量的探讨, 更缺少对恶意加密流量的探讨。文献[47]针对物联网和蜂窝网络, 综述了用于网络流量监测和分析的深度学习方法, 具体划分为四类任务: 流量分类、网络流量预测、故障管理和网络安全, 但同样缺少对加密流量的探讨。之后, 文献[48]回顾了多项网络加密技术, 包括网络层的 IPSec, 传输层和应用层之间的 SSL/TLS 和 QUIC, 应用层的 SSH 和 HTTPS, 以及 Tor 等匿名机制, 同时介绍了网络资产识别、网络表征、隐私泄漏检测和异常检测这四种加密流量分析的应用目标。上述文献从不同的视角全面回顾了各种网络流量分析方法, 但缺少对加密流量的讨论, 更没有深入研究恶意加密流量检测问题。

文献[4, 49-50]对恶意加密流量检测的研究进行了综述。其中, 文献[4]将加密恶意流量分为恶意软件和恶意攻击两大类, 梳理了加密恶意流量的提取特征, 但对于检测模型的分类和介绍较少。本文则详细介绍了多种检测模型, 包括深度学习模型和传统机器学习模型。文献[50]介绍了多种检测模型, 提出了一种基于机器学习的加密恶意流量检测的通用框架, 同时通过分析、处理、组合了五个不同的开源数据集, 生成了一个全面、平衡的新流量数据集[51]。然而, 该文献对数据不平衡问题和恶意流量类别的探讨较少, 同时缺乏对恶意加密流量领域中对抗技术的研究。

综上所述, 目前缺乏一个全面的, 包含多种任务场景、数据集、特征和模型的恶意加密流量检测及对抗综述。大部分综述只关注加密流量分类问题或恶意流量检测问题, 缺少两者的结合, 同时忽略了检测中的对抗技术。本文从现有的研究中分析总结了多种恶意加密流量检测方法和对抗攻击及防御技术, 提出了一个通用的多角度恶意加密流量检测框架。此框架包含多个研究场景, 不仅针对互联网流量, 还包括移动和无线网络流量。框架中还介绍了多种数据收集与处理方法和特征提取与选择方法, 并提供了数据不平衡问题的解决方法。同时, 框架覆盖多类检测模型, 包含传统机器学习和深度学习模型, 并探讨了对抗攻击和对抗防御方法。总之, 本文在概要介绍传统机器学习方法的基础上, 深入分析深度学习方法, 详细归纳了恶意加密流量检测及对抗技术的最新研究成果, 对研究目标、数据集、数据处理、特征提取和选择方法及评估指标体系的介绍更加全面详实。

本文共分为 7 个小节。在第 2 节中, 本文展示了所提出的通用恶意加密流量检测框架的具体细节, 并结合所提框架对目标任务场景进行探讨, 包括互联网、移动网络、物联网和卫星网络场景。在第 3 节中, 本文对现有开源数据集进行分类讨论, 介绍了相关应用场景及优缺点, 并探讨了数据不平衡问题的解决方案。在第 4 节中, 本文对加密流量特征进行分类, 同时介绍了不同的特征选择方法。在第 5 节中, 本文介绍了恶意加密流量检测的评估指标体系。在第 6 节中, 本文对比分析了不同检测模型的适用性和优缺点, 包括深度学习模型和传统机器学习模型, 同时探讨了恶意加密流量检测中的对抗攻击和对抗防御技术。最后, 在第 7 节中, 本文对恶意加密

流量检测研究中的问题和未来发展方向进行探讨和展望。

2 恶意加密流量检测的框架

2.1 框架

本文在深入查阅现有恶意加密流量检测研究的基础上，对研究目标、数据集、流量特征和模型算法进行总结归类，提出如图 3 所示的通用恶意加密流量检测框架。该框架能够与大多数现有的研究兼容，具体分为 6 个主要步骤。

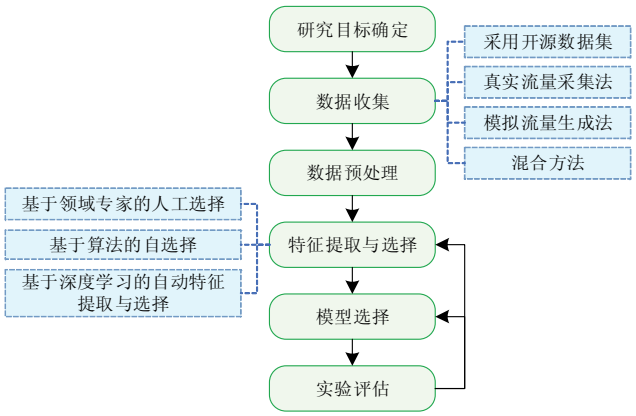

图 3 恶意加密流量检测框架

Figure 3 Framework for malicious encrypted traffic detection

第一步是，确定研究目标，主要分为不同领域内的恶意加密流量检测任务，以及二分类和多分类任务。第二步是，收集流量数据，可以通过四种方法实现：采用开源数据集、真实流量采集法、模拟流量生成法和混合方法。第三步是，数据预处理，包括对数据不平衡问题的探讨和基本的数据预处理流程。第四步是，特征提取和选择。本文将常用的加密流量检测特征分为四类：基本特征、时序特征、统计特征和协议特征。特征选择技术可以分为三类：基于领域专家的人工选择、基于算法的自选择和基于深度学习的自动特征提取与选择。第五步是，选择合适的模型进行训练，包括深度学习模型和传统机器学习模型。第六步是，进行实验评估，本文总结了多项实验评价指标。实际研究中可根据实验结果对特征选择和模型选择进行改进。本文将按节对框架中的各步骤进行详细讨论。

2.2 任务目标分类

本文将恶意加密流量检测任务目标分为两大类，分别是二分类检测和多分类检测，具体如图 4 所示。

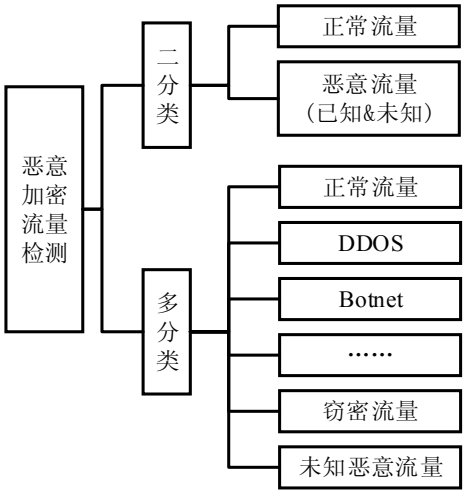

图 4 任务目标分类

Figure 4 Classification of task target

二分类检测，是判断输入的加密流量是正常还是恶意的，其中恶意流量包括已知和未知类型的恶意流量。多分类检测，是在二分类检测的基础上，检测出恶意流量的特定类型，包括 DDos、Botnet、APT、钓鱼攻击和窃密流量等等。加密流量中可能包含涉密内容，包括文字、图片、语音、视频等，而窃密木马会收集窃取受害主机的隐私数据信息并发送给特定目标，造成巨大的安全威胁。因此，针对窃密恶意流量的检测也是恶意加密流量检测的重要多分类任务之一。一般，未知攻击是在现有的攻击基础上演变产生的。因此，未知恶意流量会继承部分已知恶意流量的特征。研究人员可以通过对抗学习的方式提升模型识别未知恶意流量的准确性，使得检测模型不断学习更新恶意流量特征库，实现将未知恶意流量转化为已知攻击流量。

二分类检测任务的研究方法可以进一步细分为在一定加密协议下的二分类检测和不考虑加密协议的二分类检测。在针对具体加密协议的二分类检测研究方面，由于 HTTPS 是应用层目前最常用的加密机制，合法和恶意流量都使用其进行加密。因此，部分研究专门检测由 HTTPS 协议加密的恶意流量[12, 52]。由于传输层和应用层之间的 SSL/TLS 协议在客户端和服务端建立加密连接的初期存在一定的连接特征，文献[17, 53-55]就利用这样的 SSL/TLS 连接特征来进行恶意加密流量的检测。这些连接特征包括 TLS 握手数据包中暴露的版本号、密钥长度、证书和扩展等非加密信息。而在不考虑具体加密协议类型的二分类检测研究中，Stergiopoulos 等人[15]针对 TCP/IP 网络流，定义了 TCP/IP 侧信道特征，在被不同加密协议加密的流量数据集上使用 7 种机器学习算法进

行对比实验, 成功将加密流量划分为合法和恶意流量两类。

多分类检测任务的研究方法也可以细分为在一定加密协议下的多分类检测[20]和不考虑加密协议的多分类检测[35]。在针对具体加密协议的多分类检测研究方面, Meghdouri 等人[20]提出了一种针对 TLS 和 IPSec 加密协议的检测模型, 运用一种跨层特征向量: 包含应用层、会话层和端点行为的特征, 成功实现恶意加密流量的多分类检测。在不考虑加密协议的多分类检测研究方面, Liu 等人[35]使用高斯混合模型(Gaussian mixture model, GMM)和点排序识别聚类结构(Ordering points to identify the clustering structure, OPTICS)来计算恶意软件之间的距离, 并利用该距离定义新的恶意软件类 FClass, 之后通过 XGBoost 算法训练构建一个由 24 个检测模型组成的识别框架, 能够从多种加密流量中识别出不同种类的恶意流量和未知恶意流量。

种类繁多的恶意软件家族和恶意攻击行为给多分类恶意加密流量检测带来了更大的挑战, 目前还没有能够覆盖所有恶意软件家族流量检测的研究。多分类检测方法的性能也普遍比二分类检测方法的性能差。

2.3 目标场景分类

除互联网领域以外, 恶意加密流量检测的研究场景还包括物联网领域、卫星网络领域和软件定义的 5G 架构网络领域。不同场景下的研究如表 2 所示。

在物联网领域中, 由于物联网端点设备的计算、存储和通信能力有限和及其异构性, 容易受到各种网络攻击, 传统的网络安全解决方案已不能很好地满足物联网设备的安全需求, 而进行恶意加密流量检测可以有效提升物联网安全[58-59, 63]。文献[57]总结了物联网设备安全方面的六大挑战, 分别是: 开发端点安全解决方案、保障设备间的安全通信、物联网设备的多样性、高的安全部署和运营成本、保障隐私和高性能、以及配置管理问题, 并提出了一种能够检测网络攻击并实施必要安全措施的 IoT-Keeper。针对物联网设备和事件识别问题, Pinheiro 等人[56]使用加密流量数据包长度的统计信息, 来表征智能家居场景中物联网设备和事件的行为。具体表现为: 将数据包长度的统计平均值、标准差和在一秒钟窗口内传输的字节数作为判别特征, 使用机器学习算法(k-NN、决策树、随机森林、SVM 和多数投票算法)进行分类识别, 但计算量较大、效率不够高。之后, Zhao 等人[59]提出一种基于边缘智能的物联网加密流量检测模型, 减少了分布式物联网网关在边缘智能化过程中的通信次数, 提升了物联网加密流量分类的效率。针对物联网恶意流量检测问题, 部分研究者通过改进特征选择, 减少物联网设备的处理负载, 以提升恶意流量的检测准确率和效率[58, 63]。

表 2 不同场景下的流量检测研究

Table 2 Researches on traffic detection in different scenarios

场景	文献	年份	方向	模型方法	成果
物联网领域	文献[56]	2019	加密流量分类	机器学习算法(k-NN、DT、RF、SVM 和多数投票算法)	使用机器学习方法从加密流量中有效识别物联网设备和事件
	文献[57]	2020	恶意流量检测	IoT-Keeper 网关, 使用 Adhoc 覆盖网络	在边缘网关进行实时流量分析, 能够检测恶意网络攻击并实施必要的安全措施
	文献[58]	2020	恶意流量检测	特征选择算法 CorrAUC	物联网流量特征的有效选择, 提升恶意流量检测准确率
	文献[59]	2021	加密流量分类	序行为分析法、动态行为分析法、密钥行为分析法、二轮过滤分析法	提升物联网加密流量分类效率和识别准确率
卫星网络领域	文献[60]	2014	恶意流量检测	协同入侵检测系统 CIDS-S	对空间移动自组织节点进行实时监测, 能够对黑洞攻击、洪泛攻击、路由篡改等攻击方式进行有效检测
	文献[61]	2020	恶意流量检测	禁用协议或限制速率	主动防御卫星网络中的 DDOS 攻击
软件定义的 5G 架构网络领域	文献[62]	2017	恶意流量检测	基于机器学习的智能入侵检测系统	集成多个安全功能模块, 在全局视图下进行状态监测和流量捕获, 可识别未知攻击

在卫星网络领域中, 由于天地网络具有信道开放、节点暴露、高链路延迟的特征, 攻击者可以利用这些特征, 通过仿冒、伪造等手段对卫星网络进行

DoS 和 DDoS 攻击, 对卫星网络造成一定的威胁。针对卫星网络异常检测问题, 关汉男等人[60]设计了一种将异常检测和特征检测相结合的协同入侵检测系统 CIDS-S, 该系统将基于有限状态机的异常检测算法和自适应黑洞攻击检测算法关联协作, 取得了较好的检测效果。针对卫星网络的 DDos 攻击中的 ICMP Flooding 攻击, Usman 等人[61]对 ICMP Echo 数据包进行统计与异常检测, 可以及时发现异常并选择禁用协议或限制速率的方式以缓解攻击。

在软件定义 5G 架构网络领域中, 软件定义的 5G 架构利用软件定义网络(Software Defined Network, SDN)和网络功能虚拟化(Network Function Virtualization, NFV)的优势, 通过集中管理和动态分配资源满足 5G 网络的需求, 但也带来新的安全风险。针对这样的 5G 架构网络, Li 等人[62]提出一种基于机器学习的智能入侵检测系统, 可以灵活地集成和协调安全功能模块, 能够在集中管理和控制下自适应地、动态地检测网络入侵流量。

相同类别的加密流量在不同网络环境(WiFi、4G/5G 移动通信、物联网、工控网、区块链网络)下的包长、载荷长度序列等特征有一定差异。不同操作环境下, 对恶意加密流量检测的准确性、效率、实时性和鲁棒性有不同的要求。例如: a)在一般场景下, 要求快速识别恶意加密流量, 提高检测速度; b)在实际复杂流量场景下, 要求准确识别恶意流量; c)在无线空中接口处识别恶意加密流量; d)在恶意对抗场景下, 要求保证较高的识别准确率; e)在无线网络设备(如卫星、无人机网络等)的计算、存储、功耗受限的场景下, 要求降低深度学习模型的开销。因此, 在进行具体研究前, 研究人员应根据具体的场景和任务需求, 设定合理的研究目标。

3 数据收集与预处理技术

3.1 数据收集技术

依据本文提出的恶意加密流量检测框架, 在明确研究目标后, 下一步是收集构建实验数据集。优秀的数据集能够有效提升模型训练的效果。有时, 在封闭数据集下训练的模型, 上线之后性能表现并不理想, 原因在于训练环境与实际环境的不一致, 或训练数据与实际数据分布存在差异。因此, 构建与真实环境贴切的数据集, 对于恶意加密流量检测至关重要。另一方面, 由于异常流量远少于正常流量, 这种数据不平衡给基于深度学习的研究带来了挑战。一个数据量大、有代表性且贴合真实环境的数据集对大部分深度学习模型的训练是必不可少的。

目前常用的数据收集方法有四种: 采用开源数据集、真实流量采集法、模拟流量生成法和混合方法。

3.1.1 采用开源数据集

本节总结分析了目前常用的开源数据集, 具体如表 3 所示, 包括数据集的名称、发表年份、是否包含恶意加密流量、具体描述和优缺点。

表 3 开源数据集汇总

Table 3 Summary of public datasets

名称	年份	恶意加密	描述	优缺点
CIC-MalMem-2022[64]	2022	是	混淆恶意软件数据集, 包括三类恶意软件家族: 间谍软件、勒索软件和特洛伊木马。数据集共 58596 条样本数据, 其中 29298 条是良性的, 29298 条是恶意的。	优点: 良性与恶意流量平衡 缺点: 不完全针对加密流量
CIRA-CIC-DoHBrw-2020[65]	2020	是	基于 HTTPS 的 DNS(DOH)数据集, 其中非 DoH 和良性 DoH 流量由研究者访问排名前 10000 位的 Alexa 网站生成, 共 48952 KB 数据包; 恶意 DoH 流量由 dns2tcp、DNSCat2 和 Iodine 等 DNS 隧道工具生成, 共 219458 KB 数据包。提供由 DoHMeter 提取出的 28 维统计特征。	优点: 针对加密流量 缺点: 仅针对 DNS 流量
IoT-23[66]	2020	是	物联网流量数据集, 包含 20 种恶意流量和 3 种良性流量, 捕获时间由 2018 年到 2019 年。	优点: 真实物联网流量 缺点: 大部分流量未加密
JS-Github [67]	2020	是	Trickbot 僵尸网络恶意加密流量数据集	优点: 真实环境采集 缺点: 样本类型较少
Malware Capture Facility Project[68]	2020	是	由 Stratosphere 实验室的研究人员通过长期执行恶意软件产生收集, 最长可达三周甚至几个月。其中, 恶意软件的执行有两个限制条件: 带宽限制和 spam 拦截。	优点: 流量种类多, 数据规模大 缺点: 只有原始流量数据
UNSW-IOT[69]	2019	是	物联网流量数据集, 包含 30 个 PCAP 文件, 每个文件对应一天内收集到的流量, 总共从 27 种不同的真实物联网设备上对流量进行采集。	优点: 真实物联网流量 缺点: 大部分流量未加密

续表

名称	年份	恶意加密	描述	优缺点
CIC-IDS-2017[70]	2017	是	恶意流量检测数据集, 其中良性背景流量由 B-Profile 系统仿真生成。恶意攻击流量包括: 暴力 FTP、暴力 SSH、DoS、Heartbleed、Web 攻击、渗透、僵尸网络和 DDoS。研究人员在五天时间内收集了超过 50G 的流量。	优点: 数据规模大 缺点: 良性流量由仿真生成
CICAndMal2017[71]	2017	是	由 2015 年到 2017 年, 研究人员从真实的智能手机上运行恶意软件和良性应用程序收集到的流量数据组成, 共超过 10854 个样本(恶意: 4354, 良性: 6500)。共四大类恶意软件, 42 个恶意软件家族。	优点: 流量种类多 缺点: 不针对加密流量
CIDDS[72]	2017	否	流量从基于软件 Openstack 的模拟小型企业环境中被捕获, 研究人员在四周的时间内收集了约 3.2×10^7 的数据流, 共 5 类, 包括: Dos、暴力破解和端口扫描攻击。	优点: 数据规模大 缺点: 流量由 Python 脚本生成, 可能包含人为偏差
ISCX VPN-nonVPN[73]	2016	否	包含 7 种常规加密流量和 7 种协议封装流量, 分别是: 网页浏览、电子邮件、聊天、流媒体、文件传输、网络电话和文件共享, 同时提供由 ISCXFlowMeter 提取的流量统计特征 CSV 文件。	优点: 多种良性加密流量 缺点: 没有恶意流量
ISCX Tor-nonTor[74]	2016	否	包含 7 种加密流量, 分别是: 网页浏览、电子邮件、聊天、流媒体、文件传输、网络电话和文件共享, 并提供流量统计特征 CSV 文件。	优点: 多种良性加密流量 去点: 没有恶意流量
USTC-TFC2016[75]	2016	是	10 种恶意软件流量由 2011 年至 2015 年从真实网络环境中收集, 10 种正常流量通过 IXIA BPS 仿真设备生成。	优点: 流量种类多 缺点: 大部分恶意流量未加密
UNSW-NB15[76]	2015	是	流量数据均由 IXIA PerfectStorm 平台生成, 包含 9 种攻击流量(Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode 和 Worms)和 49 个特征。一共 2540044 条样本数据(训练集: 175341, 测试集: 82332)。	优点: 数据规模大 缺点: 不能完美拟合现实流量数据
ISCX-Bot-2014[77]	2014	是	多个僵尸网络流量数据集的子集数据, 包含 16 种现代僵尸网络技术的变体, 共 13.8GB 大小(训练集: 5.3 GB, 测试集: 8.5 GB)。	优点: 贴近真实僵尸网络环境 缺点: 不针对加密流量
malware-traffic-analysis.net[78]	2013	是	自 2013 年以来至今, 该网站提供多种类的恶意软件流量数据。	优点: 流量种类多且新颖 缺点: 数据规模较小, 只有原始流量数据
ISCX-IDS-2012[79]	2012	否	流量由研究人员在 7 天内通过执行代理程序, 模仿用户活动, 设计和执行攻击场景采集, 共 84GB, 包含 5 种类型: Normal、Brute Force SSH、DDoS、HttpDoS 和 Infiltration。	优点: 数据规模大, 缺点: 不能完美拟合现实流量数据, 不针对加密流量
CTU-13[80]	2011	是	僵尸网络流量数据集, 包含了 13 个不同场景下的僵尸网络流量样本, 由三种类型的流量组成: 恶意软件流量、合法流量和背景流量。	优点: 流量种类多 缺点: 大部分流量未加密
ISOT [81]	2011	是	共 11GB, 恶意流量占 3.33%, 包括: Storm 和 Waledac 僵尸网络; 正常流量占 96.66%, 包括: 网页浏览、电子邮件、游戏和流媒体应用。	优点: 数据规模大 缺点: 数据不平衡, 不针对加密流量
NSL-KDD[82]	2009	否	除去 KDD CUP 99 数据集中冗余的数据, 包含 4 种不同类型的攻击流量: DoS、Probe、U2R 和 R2L。	优点: 流量种类多 缺点: 不针对加密流量
Kyoto[83]	2006	是	由 honypots 技术创建, 只提供特征 CSV 文件, 不能从捕获的原始流量进一步处理特征。	优点: 真实网络流量 缺点: 只能观察对蜜罐的攻击, 未提供原始流量数据

通过对上述数据集的对比分析, 可以发现大部分数据集并不针对加密流量, 恶意流量中仅部分是加密的。只有少量数据集完全由恶意加密流量和正常加密流量组成, 但此类数据集通常针对某种特定协议或恶意软件家族, 不够全面。

一个全面、良好的恶意加密流量检测数据集需要满足以下六个特征: (a)完整流量; (b)已标记; (c)恶意加密攻击种类多样; (d)样本数充足; (e)类平衡; (f)无冗余数据。然而, 现实流量种类繁多, 一个数据集不可能包含全部的流量种类。目前, 恶意加密流量检测领域缺少公认的开源数据集, 研究人员普遍选择构造特定于检测目标的私有数据集, 这给研究间的对比带来了公平性的挑战。

3.1.2 真实流量采集法

真实流量采集法是目前研究人员常用的流量收

集方法之一, 具体过程是在真实运行的网络环境(如公司内网、校园网等环境)中, 使用流量采集软件(如Wireshark、Fiddler、QPA 等软件)进行流量采集, 其次对流量数据进行分析和标记以形成私有数据集。

上述开源数据集中, CTU-13 数据集[80]是捷克理工大学采用真实流量采集法, 在真实的网络环境中, 捕获了 13 种不同的恶意软件流量、正常流量和后台流量组成的。Lopez-Martin 等人[84]从西班牙的学术和研究骨干网 RedIRIS 中, 采集了 266160 条网络流进行研究。蜜罐是常用的物联网流量数据收集方法。文献[85]中的数据集就由研究人员在 2017-2018 年间, 部署的一个高交互性的物联网蜜罐收集得到的。在一年半的时间内, 这些蜜罐出现在 40 个公共 IP 地址上, 同时给 11 个物联网设备转发过流量。

真实流量采集法能够采集到现实网络的真实数据, 保证数据集中网络流量的真实性, 但也存在不足。一是, 在真实的网络交互环境下, 网络攻击的频率较低, 且存在不确定性, 易产生数据不平衡的问题。二是, 采集到的真实网络流量中往往存在大量冗余数据, 需要大量的时间进行筛选和标记。三是, 现实网络环境复杂, 真实流量的采集通常针对具体的领域, 无法覆盖所有的网络场景。

3.1.3 模拟流量生成法

模拟流量生成法的具体过程是, 通过构建虚拟网络或使用脚本来模拟生成目标流量数据集[70, 79]。上述开源数据集中的 CIC-IDS-2017[70]中良性背景流量由 B-Profile 系统仿真产生; ISCX-IDS-2012 数据集[79]由研究人员构建虚拟环境, 模仿真实用户行为和设计攻击行为得到。此外, 部分研究人员[21, 86-87]在沙箱中执行恶意软件, 实现恶意流量数据的收集。

模拟流量生成法可以拟合大多数攻击模式, 相较于采集真实流量采集法, 可以生成更多类型的攻击和恶意流量, 能够在一定程度上解决数据不平衡问题, 也更合适于机器学习。然而, 模拟流量生成的数据集与真实的网络数据存在差异, 并不能真实代表现实网络流量的分布状况, 存在一定的人为偏差。

3.1.4 混合方法

混合方法就是混合使用真实流量采集法和模拟流量生成法。CIDDS 数据集[72]就是由混合方法生成的。研究人员使用 OpenStack 模拟了一个小型的企业环境, 包括多个客户端和一些典型的服务器(如电子邮件服务器、Web 服务器), 并在此虚拟环境下收集良性与恶意流量。同时, 研究人员部署了一个外部服务器, 以收集除 OpenStack 环境之外的真实网络流量。因此, CIDDS 数据集[72]内既有模拟流量也有真实流量。

混合方法既能够模拟较多的攻击和恶意软件流量, 也能收集到代表实际网络环境的真实流量, 较好地满足了恶意加密流量数据集的构建需求。然而, 该方法需要进行数据集成和标记, 耗费较多的人力。

3.1.5 小结

上述四种数据收集方法各有优缺点。a)采用开源数据集有利于研究者间进行实验结果的对比, 且数据规模较大、覆盖场景全面, 可以进行组合以增加流量数据的种类和数量, 但可能与研究者实际目标场景有偏差, 不能完全拟合。b)真实流量采集法保证了流量数据的真实性, 但收集和标记成本较高。c)模拟流量生成法可以有效生成多种目标流量, 但可能引入人为偏差。d)混合方法能较好地符合数据要求, 但人工成本最高。

针对自己的研究目标, 研究者首先可以选择多个合适的开源数据集, 构成混合数据集, 对自己的模型进行初步实验验证。再使用混合方法: 一方面采集真实环境下的流量数据, 如校园网络环境和企业网络环境。其中, 需要确定数据收集位置, 如通信客户端或服务器端、网络边缘、网络核心或两者之间的任何位置。另一方面搭建虚拟网络环境、使用脚本和沙箱等方法生成目标流量。需要注意的是, 收集到的流量需要尽可能多的用户交互, 避免模型拟合用户特征而非流量特征。最终, 实现实验数据集的数据与实际环境相一致的目标。

3.2 数据预处理技术

完成数据收集后, 下一步是对数据进行预处理, 将原始流量数据转化为适合分析的形式, 包括流量清洗、流量切分、数据变换、数据规约等。另一方面, 由于恶意流量的数量往往少于正常流量, 恶意加密流量检测领域内的数据不平衡问题值得研究。本节将探讨数据不平衡问题的解决方法和流量数据预处理的详细过程。

3.2.1 数据不平衡处理技术

针对恶意加密流量检测领域内的数据不平衡问题, 目前常用的解决方法有四类: 组合不同公共数据集、修改目标成本函数、采样和生成人工数据[88]。相关文献分类及描述如表 4 所示。

上述四种方法的具体描述和相关文献总结如下所示。

(1) 组合不同公共数据集

研究者可以将不同的数据集进行组合, 或使用自主生成的数据集来增加流量的类型和数量[99]。Chen 等人[29]结合了三个不同的公共数据集作为实验

表 4 数据不平衡处理方法总结

Table 4 The summary of works on data imbalance

方法	文献	年份	描述
组合数据集	文献[29]	2020	组合三个不同的公共数据集作为实验数据集。
	文献[50]	2022	从五个不同的公共数据集中随机选择不同比例的加密流量样本组成新数据集。
采样	文献[89]	2000	提出随机欠采样和随机过采样方法。
	文献[90]	2018	采用欠采样方法, 随机去除数据集中主要类的样本数量, 直到数据集类别较平衡。
生成人工数据之SMOTE 采样及其变种	文献[91]	2002	提出 SMOTE 采样算法, 一个小类别样本生成算法, 使用 k-NN 算法合成小类别的人工样本。
	文献[92]	2005	提出 Borderline-SMOTE 算法, 在少数类和多数类的边界线附近生成合成样本。
	文献[93]	2011	提出 SVM-SMOTE 算法, 使用 SVM 进行新样本的生成。
	文献[94]	2015	提出 SMOTE-IPF 算法, 以迭代集成的噪声滤波器为基础, 能够解决噪声问题。
	文献[95]	2018	提出均值 SMOTE (M-SMOTE)算法, 使生成的样本更集中于样本中心, 具有丰富的类别特征。
生成人工数据之GAN 及其变种	文献[96]	2017	提出辅助分类器生成对抗网络(Auxiliary Classifier GAN, AC-GAN)来填充数据集中的弱样本。
	文献[97]	2019	提出一种基于条件 GAN(Conditional GAN, CGAN)的新型数据增强方法 PacketCGAN, 以应用类型的输入作为条件生成指定的样本, 实现数据平衡。
	文献[98]	2022	提出 Markov-GAN 增强模型, 能够自动生成纹理丰富、判别力高、多样性好的马尔可夫新图像样本, 扩充数据集。

数据集, 分别是: Stratosphere IPS 项目、CTU-13 和 malware-traffic-analysis.net, 以此增多流量的类型和数量。Wang 等人[50]随机从五个不同的公共数据集中选择不同比例的加密流量样本, 组成了一个新的类平衡的恶意加密流量检测数据集, 并在 Mendeley Data 上发布。这五个公共数据集分别是 UNSW-IOT-2019、CIC-IDS-2017、CICAndMal2017、Malware Capture Facility Project Dataset 和 CIC-IDS-2012。此外, 研究者可以使用无标记和有标记的流量数据集相结合的方法扩展数据集, 通过迁移学习解决缺少大型标记数据的问题[100]。

(2) 修改目标成本函数

修改目标成本函数的方法主要通过对多数类别(正常流量)和少数类别(恶意流量)进行不同的加权, 当分类器未能将少数类别分类正确时, 采取更严厉的惩罚来缓解问题。但由于这种方法需要在实际应用中为少数类别确定合适的加权值, 不具备普遍适用性。

(3) 采样方法

采样方法分为欠采样和过采样两种方式, 具体通过删除样本量大的类别数据和增加样本量小的类别数据来达到样本平衡的目的。尽管采样方法会改变原始数据的分布, 但其操作简便、高效, 具有适用性。2000 年, Japkowicz 等人[89]提出随机欠采样(Random Under-Sampling, RUS) 和随机过采样(Random Over-Sampling, ROS)这两种最常用的采样方法。其中, RUS 的基本思想是, 随机删除数量大的类别样本; 而 ROS 的基本思想则是, 在小样本类别中随机选取样本并进行复制。Wang 等人[90]提出的 Datanet 采用 RUS 的方法, 随机去除数据集中主要类的样本数量, 直到类别较平衡。虽然 RUS 在训练数据样本量很大时, 可以减少数据量, 提高运行效率, 但很有可能缺失具有重要价值的潜在有用数据。另一方面, ROS 会增加发生过拟合的可能性, 因为 ROS 只是简单的复制数据。

(4) 生成人工数据

生成人工数据的经典方法是 SMOTE 采样及其变种。2002 年, Chawla 等人[91]提出 SMOTE 这一少数类别样本生成算法。SMOTE 算法主要通过研究样本数较少的类别的特点, 并且采用 k-NN 算法合成该类别的人工样本, 再添加到少数类之中。虽然 SMOTE 采样可以避免过拟合的问题, 但其忽略了样本周围的邻近实例, 因此会带来类重叠、噪声影响等问题。许多研究者提出了基于 SMOTE 算法的一系列改进算法, 例如在少数类和多数类的边界线附近生成合成样本的 Borderline-SMOTE[92]; 结合 SVM 算法, 使用 SVM 进行新样本合成的 SVM-SMOTE[93]; 基于迭代集成的噪声滤波器、且能够克服噪声问题的 SMOTE-IPF[94]。Yan 等人[95]则提出了一种均值 SMOTE (M-SMOTE)算法, 来实现对流量数据的平衡化处理。相较于 ROS, SMOTE 及其改进算法合成样本的效果更好, 但仍存在一定的不足。因为

SMOTE 方法只关注局部信息, 可能会导致合成样本的单一性, 在一定程度上导致过拟合。

生成人工数据的方法还包括使用生成对抗网络(GAN)。GAN[101]是一种通过生成网络和判别网络的对抗博弈, 从而收敛到最优解的生成模型。GAN 能够生成与真实数据高度相似的可信样本, 在图像生成和自然语言处理领域已取得了出色的成就[102-103]。同样, 在恶意加密流量检测领域, GAN 也可以用来生成流量样本, 扩充数据集, 以解决数据不平衡的问题。目前, 研究者提出了多种基于 GAN 的改进样本生成算法, 如 AC-GAN[96]、PacketCGAN[97]和 Markov-GAN[98]。2017 年, Vu 等人[96]提出 AC-GAN 来填充数据集中的弱样本, 提高了 SVM、RF、DT 等多种算法的加密流量识别准确率, 效果优于 SMOTE 算法。2019 年, Wang[97]提出一种基于 CGAN 的新型数据增强方法 PacketCGAN, 以应用类型的输入作为条件, 生成指定样本, 从而平衡流量数据集。与使用原来的流量数据集作为训练集相比, 检测率提升了 1.54%。Tang 等人[98]提出 Markov-GAN 增强模型, 能够自动生成纹理丰富、判别力高、多样性好的马尔可夫新图像样本, 扩充数据集。与 AC-GAN 和 PacketCGAN 相比, Markov-GAN 不仅更易于训练, 同时具有更强的样本扩增能力和泛化能力。

3.2.2 流量预处理技术

通常, 收集到的流量数据集不能直接作为模型的输入, 需要进行预处理操作。流量预处理的过程通常包括流量清洗、流量切分、长度统一和数据变换。

(1) 流量清洗

流量的清洗和过滤是预处理的第一步, 主要将收集到的流量中重复和无效部分清除。大多数公共流量数据集中的原始流量数据存储在单独的 PCAP 或 PCAPNG 文件中。由于原始报文数据中总是包含一些和恶意加密流量检测研究不相关的报文, 如 ARP(Address Resolution Protocol)、DHCP(Dynamic Host Configuration Protocol)和 ICMP (Internet Control Message Protocol)等报文, 需要清理删除这些报文。另一方面, 因网络条件变化造成的重复、乱序、损坏、不完全或者空的数据包也需要被清理删除。此外, 对于 PCAP 文件中包含的一些不必要信息, 如 PCAP 文件头, 可以通过包过滤技术去除[90, 104]。

(2) 流量切分

流量的粒度可以分为 Packet 级别(包级)、Flow 级别(流级)、Session 级别(会话级)和 Stream 级别(主机级), 具体如表 2 所示。Packet 级别是研究数据包的特征, 包括数据包的大小及其分布, 达到目的站点的时间间隔等。Flow 级别关注流的主要特征, 包括流持续时间、流字节数, 以及流到达目的站点的过程。流是具有相同五元组(源 IP、源端口、目的地 IP、目的端口、传输层协议)的所有数据包。Session 级别关注会话的特征及会话到达过程。会话是由具有相同五元组的双向流组成的一组数据包, 即会话的源 IP 和目的 IP 可以互换。Stream 级别是研究通信主机间的连接模式, 包括主机通信间的所有流量, 或与主机的某个 IP 和端口通信的所有流量。主机级特征是总包个数、每条流的平均包个数、时间间隔和包长的均值等流级特征的聚合。

表 5 流量的粒度划分

Table 5 Granularity of traffic

流量粒度	研究目标	主要特征
Packet	数据包的特征、到达过程	包的大小及其分布、达到目的站点的时间间隔
Flow	流的特征、到达过程	流持续时间、流字节数
Session	会话的特征、到达过程	会话持续时间、会话字节数
Stream	主机间的连接模式	总包个数、每条流的平均包个数、时间间隔和包长的均值

目前的恶意加密流量检测研究, 常以单个数据包、单条流或单个会话作为最小检测单位。研究人员需要根据实际实验需求选择合适的流量粒度, 对网络流量进行切分。

(3) 长度统一

完成流量的清洗和切分后, 为适应部分模型的输入需求, 需要使用数据截断或补零来保证输入数据的长度一致[84]。以 TCP 为例, 其数据帧长在 54 到 1514 字节之间变化很大, 而部分深度学习模型需要固定大小的输入。因此, 研究人员需要对数据进行长度统一。例如, 将每条会话的长度固定为 1024 字节, 如果会话长度大于 1024 字节则截断, 小于 1024 字节则在会话末尾补零。

(4) 数据变换

数据变换是预处理过程中的重要一步, 有利于之后的信息挖掘, 主要包括特征数字化、标签映射和数据归一化等操作。由于原始流量数据集中存在不同数据类型的分类特征(如网络协议信息、加密协议证书、服务器指示名称等), 需要将这些非数字特征进行编码, 转化为数字数据。例如协议特征包含 TCP、UDP 以及 ICMP 三个非数字特征, 可以将其映射为 0, 1, 2, 这就是特征数字化的过程。标签映射就

是将流量数据集中每个实例的标签编码转化为相应的数字数据。在二分类任务中, 可以将正常流量的标签编码为 0, 恶意流量的标签编码为 1。在多分类任务中, 可以用 0 表示正常流量, 1 表示 Dos 恶意流量, 2 表示 Probe 恶意流量, 3 表示 U2R 恶意流量, 4 表示 R2L 恶意流量。数据归一化的目的在于把流量数据或特征值整合到[0,1]或[-1,+1]的范围内, 从而减少数据冗余, 提升模型训练收敛的速度, 常用的归一化方法有最小最大值归一化(Min-Max Normalization)。

4 特征提取与选择技术

4.1 特征提取技术

特征的提取与选择关系着检测质量, 是恶意加密流量检测研究中的关键步骤。在进行特征选择之前, 首先需要了解加密流量数据中常用的特征类型。目前, 学术界没有公认的网络流量特征的分类和命名规范。因此, 本文通过调研相关文献, 将恶意加密流量检测领域常用的流量特征分为以下四类: 基本特征、时序特征、统计特征和协议特征。

1) 基本特征, 包含: 源端口、目的端口、源 IP、目的 IP、传输协议类型。

2) 时序特征, 是包和流处于活动状态下的时间属性特征, 包含: 包持续时间、流持续时间、包到达时间、流到达时间、包间隔时间、流间隔时间、会话间隔时间、TCP 连接设置的往返时间、TCP 窗口变化时间、SYN 和 SYN_ACK 包之间的时间, 以及上述所有特征的最大值、最小值、中位数、平均值、方差。

3) 统计特征, 是经过数理统计计算得到的特征, 包含: 源包数、目的包数、源字节数、目的字节数、重传或丢失的包数、包大小、包长度顺序分布、包长度、流长度、负载长度、IP 头长度和 TCP 窗口长度的平均值、最大值、最小值、中位数、方差和标准差, 以及 TCP 协议中出现 SYN、ACK、FIN、CWR 标志包的数量等。

4) 协议特征, 是和加密协议相关的各项特征。在加密连接建立之前的握手阶段, 客户端与服务器往往需要明文协商相关加密参数, 因此这一阶段的原始字节包含加密通信时使用的版本、加密套件、证书、扩展等信息。综上所述, 协议特征包含: 协议版本号、协议套件种类、加密算法、证书有效性、证书密钥平均值、证书链长度、证书自签名、支持的扩展项等。

大部分研究人员选择上述常用的特征构建自己的特征集。文献[21]使用会话的统计特征、TLS 协议特征、证书特征和域名特征构建特征向量作为模型输入。文献[105-106]选择使用统计特征来进行恶意应用流量的识别, 包括字节数和包数。文献[35, 57, 107]选择基本特征、时序特征和统计特征作为特征集。其中, 文献[107]使用的统计特征包括流字节数、已发送的流量数据包数量, 时序特征包括流持续时间、数据包到达时间、前向包时间差、后向包时间差、流活跃时间和流闲置时间的最大、最小、平均值和标准差。文献[18, 31, 55]选择将基本特征、时序特征、统计特征和协议特征都作为自己的特征集, 再通过人工或特征选择算法选出合适的特征子集。

除了上述常用特征, 部分研究者定义了一些特殊的特征来补充自己的特征集, 包括抗篡改特征[28]、侧信道特征[15, 108]和 N-gram 序列特征[109]。其中, 抗篡改特征是由 Celik 等人[28]定义的, 包括: 最大数据包大小与最小数据包大小之比(IP ratio)和帧字节总数除以有效吞吐量(Goodput), 这些特征与端口和有效载荷无关, 且很难被攻击者干扰。而侧信道特征能够在保证恶意加密流量检测准确率的同时减少训练集的大小和训练时间。N-gram 序列特征则是 Wang 等人[109]将移动应用程序生成的每个 HTTP 流视为文本文档, 再以自然语言处理的方式, 使用基于 N-gram 生成的分词方式从而生成的, 可以有效表征 HTTP 流。然而, 这些特殊特征也存在缺点, 即部分只能在特定场景下使用。

4.2 特征选择技术

当确定加密流量检测的特征集后, 研究者需要进行特征选择。优秀的特征选择可以提高分析性能、加速分析过程, 并有助于研究人员理解生成数据的底层机制[110]。然而, 更多的特征并不代表更好的检测效果。当选择的特征过多时, 复杂的特征识别与计算会耗费大量的存储量和计算量, 延长模型的训练时间, 导致模型检测性能的下降。

常用的特征选择方法可分为两类。第一类, 是基于领域专家的人工选择。领域专家根据经验和知识, 选择出合适的特征。研究人员再将这些特征作为模型输入, 训练检测模型。Anderson 等人[18]发现在初始特征集中添加领域专家建议的特征可以极大地提高检测系统的性能。目前, 很多研究都使用类似的方法人为地进行特征选择。然而, 人工选择特征可能会出现失误, 不能保证完全可靠。同时, 专家可能无法挖掘到深层次的、非直观找到的特征, 这将极大影响模型的检测性能。

第二类, 是基于算法的自选择。这种方法需要研

究人员先尽可能地提取多种流量特征, 构成特征集, 再使用相关算法从特征集中选择最合适的特征, 形成特征子集用于训练模型。常用的特征选择算法有互信息算法(Mutual Information)[16]、递归特征消除法(Recursive Feature Elimination, RFE)[17]、基于随机森林的平均不纯度减少算法(Mean Decrease Impurity, MDI)[53]、序列前向选择算法(Sequential Forward Selection, SFS)[35]和卡方检验算法(Chi-square)[109]。除了文献[35]中使用的 SFS 算法是逐步增加特征集的大小, 直到获得最优特征集, 其余特征选择算法都是在大特征集的条件下, 对特征的重要性或相关性进行排序, 去除不相关或冗余的特征, 从而构建出最优的特征子集。

4.3 自动特征提取与选择技术

现已有诸多研究选择使用深度学习模型自动提取流量中的隐藏特征。深度学习作为机器学习的一个重要分支, 是解决特征设计问题的有效途径。深度学习方法通过构建多层次的神经网络结构, 对每一层进行不同的运算操作, 逐步提取输入数据的特征并进行组合, 从而生成较高层次的抽象特征用于模型的判决。这种方法直接将处理好的流量数据作为深度学习模型的输入, 由模型自动学习提取流量深层特征, 不需要人工干预[75, 111]。因此, 研究人员对流量数据的处理也相对简单, 只需要进行清洗, 切分和长度统一等操作, 而不再需要人工提取和选择特征, 有效节省人力物力。Wang 等人[75]先截取会话的前 784 字节, 再将字节映射成 28×28 的二维灰度图像作为 2D-CNN 模型的输入, 从而提取流量的深层空间特征。同样, 如果将流量序列作为 RNN 模型的输入, 可以提取出流量的时间序列特征。

基于深度学习的特征自提取和选择一方面节省了人工成本, 另一方面可以发现深层次、非直观的特征, 同时可以避免人为的错误和偏差。然而, 由于深度学习模型的黑匣子性质, 基于深度学习模型提取的特征缺乏可解释性, 还需进一步研究。

5 评估指标体系

针对恶意加密流量检测, 常用的实验评价指标有以下几种: 准确率(*Accuracy, ACC*)、精确率(*Precision, P*)、召回率(*Recall, R*)、误报率(*False Positive Rate, FPR*)、漏报率(*False Negative Rate, FNR*)、特异度(*Specificity, Spec*)、*ROC* 曲线(*Receiver Operating Characteristic curve*)、*AUC*(*Area Under Curve*)和 F_1 值。上述指标可以通过混淆矩阵计算得到。混淆矩阵如表 6 所示, 用于描述恶意加密流量检测中实际类别和预测类别之间的相互关系。其中, *TP*(*True Positive*)表示被正确识别的恶意流量的样本数, *TN*(*True Negative*)表示被正确识别的正常流量的样本数, *FP* (*False Positive*)表示正常流量被误分为恶意流量的样本数, *FN*(*False Negative*)表示恶意流量被误分为正常流量的样本数。

表 6 混淆矩阵

Table 6 Confusion Matrix

		预测	
		恶意	正常
实际	恶意	*TP*	*FN*
	正常	*FP*	*TN*

基于上述定义, 可以得到如下所示的各项评价指标的详细计算公式。

(1) 准确率

ACC 是被正确分类的样本数与样本总数之比。

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

(2) 精确率

P 是被正确识别的恶意流量样本数与被检测模型识别为恶意流量的样本数之比。

$$P = \frac{TP}{TP + FP}$$

(3) 召回率

召回率 *R*、真阳性率(*True Positive Rate, TPR*)、检测率(*Detection Rate, DR*)、灵敏度(*Sensitivity, Sens*)均表示被正确识别的恶意流量样本数与真实恶意流量的样本总数之比。*R* 越高表明模型的检测能力越强, 模型越有效。在对抗性场景中, 如果模型仍保持高 *R*, 则可认为该模型能够应对对抗性攻击。

$$R = TPR = DR = Sens = \frac{TP}{TP + FN}$$

(4) 误报率

FPR 表示被误识别为恶意流量的正常流量占所有正常流量样本的比重。好的模型需要具有低误报率。

$$FPR = \frac{FP}{FP + TN}$$

(5) 漏报率

FNR 表示被误识别为正常流量的恶意流量占所有恶意流量样本的比重。好的模型需要具有低漏报率。

$$FNR = \frac{FN}{FN + TP}$$

(6) 特异度

Spec 表示被正确识别的正常流量样本数与真实正常流量的样本总数之比, 其与 *FPR* 的关系如下所示。

$$Spec = 1 - FPR = \frac{TN}{TN + FP}$$

(7) *ROC* 曲线和 *AUC* 值

如图 5 所示, *ROC* 曲线以 *TPR* 为纵坐标, 以 *FPR* 为横坐标绘制而成, 提供了与数据分布相关的 *TPR* 和 *FPR* 之间的可视化表示。

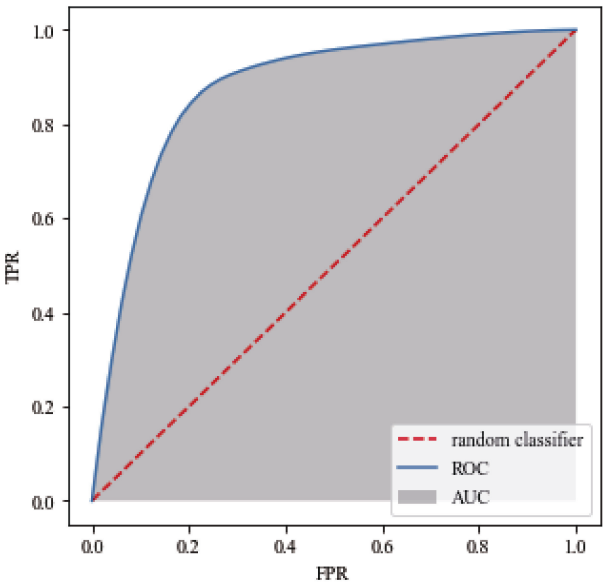

图 5 ROC 曲线

Figure 5 ROC curve

如图 5 所示, *AUC* 表示 *ROC* 曲线下的面积, 取值范围在 0 到 1 之间。*AUC* 值越接近 1, 表示模型的性能越好; 当 *AUC* 小于等于 0.5 时, 模型效果与随机猜测一样或更差, 模型没有预测价值。

(8) F_1 值

F_1 值为 *P* 与 *R* 的调和平均数。一般情况下, *P* 越高, *R* 越低; 反之 *R* 越高, *P* 越低。而 F_1 值是可以综合考虑 *P* 和 *R* 的精确性指标。

$$F_1 = \frac{2TP}{2TP + FP + FN} = \frac{2 * P * R}{P + R}$$

除了上述的精确性评价指标, 模型的评估指标还包括实时性和鲁棒性。a)实时性指标用于评估模型的检测效率, 反映模型能否在线训练、实现恶意加密流量的快速检测。大部分 DL 模型的训练需要大量数据, 会选择先离线训练, 再在线检测。这会耗费大量的计算存储资源和时间, 实时性的恶意加密流量检测研究有待发展。b)鲁棒性指标用于评估检测模型应对未知攻击和对抗攻击的能力, 反映模型能否识别未知恶意流量和有意规避检测进行伪装的对抗恶意流量。对抗训练和抗干扰特征的提取, 可以有效提高模型的鲁棒性。

6 恶意加密流量检测及对抗

传统的 DPI 方法已不再适用于恶意加密流量检测。近十年来, 越来越多的学者致力于将机器学习方法和深度学习方法应用于恶意加密流量检测领域。同时, 由于恶意加密流量检测具有高对抗性, 攻击者使用各种手段如流量混淆、流量特征伪装技术隐蔽恶意流量、逃逸检测, 检测人员进行安全防御的难度被提高了。本节将首先介绍恶意加密流量的检测技术, 包括传统机器学习模型和深度学习模型的选择, 其次介绍对抗攻击及防御技术, 最终总结未来研究方向。

6.1 检测技术

模型选择是恶意加密流量检测研究中的重点, 目前常用的模型有传统机器学习模型和深度学习模型。本节将依次介绍应用这两类模型的检测研究及其优缺点, 其中深度学习模型是本节介绍的重点。

6.1.1 传统机器学习模型

基于传统机器学习的检测方法需要构建加密流量的特征集, 再进行特征选择, 最后输入不同模型进行聚类或分类识别, 主要关注模型算法和特征集的优化。传统机器学习模型可以分为三大类: 有监督学习模型、半监督学习和无监督学习模型。相关文献整理如表 7 所示。

1) 有监督学习模型, 常使用带标签的流量数据进行训练, 检测准确率较高[15-23]。其中, 随机森林、决策树和 XGBoost 算法模型在流量检测领域表现较好。尤其, XGBoost 的准确率较好[16], RF 的鲁棒性较好[18], 不易被错误的数据标签或者噪声影响。然而, 有监督学习模型对未知恶意流量的识别能力有限。

2) 无监督学习模型, 常使用无标签的流量数据进行训练, 将具有相似性的样本划分成一个簇[24-30]。常用的无监督学习模型有 K-means 聚类模型[24]及其改进模型[27]。然而, K-means 聚类模型中的参数 k 仍需手动确定, k 的自适应确定问题仍有待解决。此外, Chen 等人[29]提出一种改进密度峰值聚类算法(DPC-GSMND), 可以有效地降低计算复杂度, 提高聚类精度。实验结果证明, DPCGS-MND 的性能优于选择性抽样[112], 智能抽样[113]和基于层次聚类的抽样[114], 能够提升检测准确率。然而, 无监督模型的训练通常依赖大量样本的支撑, 如果样本数量不足, 可能导致检测模型的精度下降。

表 7 使用传统机器学习模型的文献总结

Table 7 Summary of works using traditional machine learning models

方法	文献	目标	机器学习模型	数据集	年份
监督学习	文献[15]	恶意流量检测	k-NN, NB, SVC, CART, LinReg, LDA, MLP	FIRST 2015, CTU-13, Milicenso	2018
	文献[16]	恶意加密流量检测	SVM, DT, RF, XGBoost	CTU-13	2019
	文献[17]	恶意加密流量检测	RF, SVM, XGBoost	CTU-13, MCFP	2019
	文献[20]	恶意加密流量检测	RF	CIC-IDS-2017, UNSW-NB15, ISCX-Bot-2014	2020
	文献[21]	恶意加密流量检测	DT, RF, LogReg	未公开	2021
	文献[23]	恶意加密流量检测	Profile HMM, RF, MLP, SVM	CTU-13, CIC-IDS-2017, SSL Blacklist	2022
无监督学习	文献[24]	恶意流量检测	K-means	University of Twente-Traffic Measurement Data	2007
	文献[27]	加密流量分类	改进的 K-means	未公开	2012
	文献[28]	恶意流量检测	K-means, One-Class SVM, k-NN, 最小二乘异常检测	University of Twente-Traffic Measurement Data	2015
	文献[29]	恶意加密流量检测	DPC-GSMND	Stratosphere IPS Project, CTU-13, malware-traffic-analysis.net	2020
半监督学习	文献[33]	加密流量分类	随机森林+分层聚类	未公开	2016
	文献[34]	异常流量检测	半监督的 Adaboost 算法	Data Mining CUP 1999 Data Set	2016
	文献[35]	恶意加密流量检测	高斯混合模型+XGBoost	CIC-IDS-2017, Stratosphere IPS Project, malware-traffic-analysis.net	2019

3) 半监督学习模型，常使用带标签和无标签的流量数据一起进行训练，是一种结合监督学习和无监督学习的框架模型[31-35]。其中，Liu 等人[35]结合无监督的高斯混合模型和有监督的 XGBoost 算法实现多种类恶意软件的细粒度检测。半监督学习模型能够减少对标记数据的依赖，同时相较于无监督学习模型，可以通过反向传播进行修正，有利于提升模型性能。

综上所述，目前基于传统机器学习的方法已取得不错的成果，但高度依赖专家设计、选择的统计特征，耗费较大人力物力，且泛化能力有限。

6.1.2 深度学习模型

深度学习的迅速发展为恶意加密流量的检测提供了新的可行的思路。早在上个世纪中，已有学者提出深度神经网络的概念[115]，但一直进展平缓。直至 2006 年，Hinton 教授[116]提出了预训练技术与深度学习的概念，极大地促进了其的发展。此后，随着神经网络模型在 ImageNet 图像分类大赛中展现优异的性能，DL 的应用领域也逐渐扩展。目前 DL 已在图像处理[117]、文本分类[118]、语音识别[119]、机器翻译[120]等领域展现出优越的性能。越来越多的学者将 DL 应用于网络安全领域，包括流量分类、隐私保护、恶意软件检测和入侵检测等各方面。图 6 总结了能应用于网络安全领域中的 DL 算法，包括有监督学习算法、无监督学习算法和深度强化学习。DL 方法的优势在于：a)自动提取特征，降低人工成本；b)提取深层数据特征，发现不同流量特征之间的非直观联系；c)提高应用程序的处理速度和准确性。

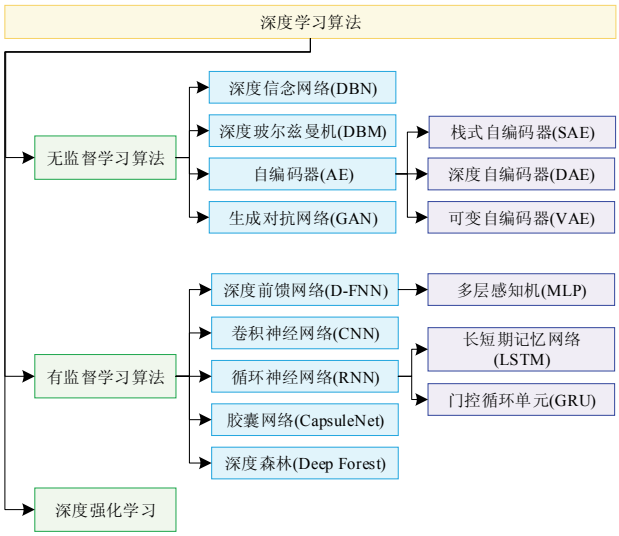

图 6 深度学习算法分类

Figure 6 Classification of deep learning algorithms

在恶意加密流量检测领域，常用的深度学习模型有多层感知机、卷积神经网络、循环神经网络、自动编码器和生成对抗模型，其中卷积神经网络包括普通 CNN 和图卷积神经网络，循环神经网络包括普通 RNN、LSTM 和 GRU。本节将依次介绍上述深度学习模型的优缺点及其应用研究。相关文献总结对比如表 8 所示。

6.1.2.1 MLP

MLP 作为一类前馈人工神经网络，由三部分组成：一个输入层、一个或多个隐藏层、一个输出层，

其层与层之间是全连接的, 即一层中的每个节点都以一定的权重连接到下一层的每个节点。

MLP 可以完成加密流量识别任务[90, 121]和异常流量检测任务[122]。文献[90]中的 MLP 模型是加密流量分类方法 DataNet 所用到的深度学习模型之一, 由 1 个输入层、2 个隐藏层和 1 个输出层组成。其中, 输入层有 1480 个神经元, 2 个隐藏层分别由 6 个神经元组成, 输出层有 15 个神经元, 使用 Softmax 作为分类器。作者在 ISCX VPN-nonVPN 数据集[73]上进行实验, 实验结果证明 MLP 的准确率、召回率和 F1-Score 均在 92%以上。然而, 在加密流量检测领域, 由于 MLP 的复杂性和准确性较低, 且无法处理高维输入和隐藏层参数过多的情况, MLP 已很少被使用。

表 8 使用深度学习模型的文献总结

Table 8 Summary of works using deeep learning models

文献	目标	模型	输入	数据集	年份
文献[121]	加密流量分类	MLP	流量特征	未公开	2018
文献[122]	异常流量检测	MLP+C4.5	流量特征	ASNM-NPBO	2018
文献[90]	加密流量分类	MLP+CNN+SAE	原始流量	ISCX VPN-nonVPN	2018
文献[123]	恶意流量检测	CNN+SVM	流量特征+原始流量	MCFP	2019
文献[75]	加密流量分类	CNN	原始流量	USTC-TFC2016	2017
文献[53]	恶意加密流量检测	CNN	流量特征	CTU-13	2020
文献[124]	恶意流量检测	CNN	原始流量	CIC-IDS-2017	2022
文献[125]	恶意加密流量检测	GCN+DT	流量特征+流量轨迹图	DATACON	2022
文献[126]	异常流量检测	Text-CNN+RF	流量特征+原始流量	ISCX-IDS-2012	2018
文献[127]	恶意加密流量检测	LSTM+CNN	原始流量	CIC-IDS-2017, CCE2021	2021
文献[128]	恶意加密流量检测	LSTM+GRU+CNN	原始流量	NSL-KDD, UNSW-NB15, CIC-IDS-2017	2021
文献[129]	加密流量分类	Tree-RNN	原始流量	ISCX VPN-nonVPN	2021
文献[130]	恶意流量检测	Bi-GRU+Attention	原始流量	CTU-13 + ISCX-IDS-2012	2022
文献[131]	恶意加密流量检测	CBOW-LSTM	流量特征	CTU-Malware-Capture + JS-Github	2022
文献[104]	加密流量分类	SAE+CNN	原始流量	ISCX VPN-nonVPN	2020
文献[132]	恶意加密流量检测	LSTM-AE	流量特征	CTU-13, malware-traffic-analysis.net	2020
文献[54]	恶意加密流量检测	LSTM-AE+CNN	原始流量	CTU, malware-traffic-analysis.net	2021
文献[133]	恶意流量检测	AE+CNN	原始流量	CIC-IDS-2017, CIC-IDS-2012, USTC-TFC2016	2021
文献[98]	恶意加密流量检测	Markov - GAN	原始流量	USTC-TFC2016, ISCX VPN-nonVPN	2022
献[134]	恶意加密流量检测	CTTGAN	流量特征	CIC-IDS2017	2022

6.1.2.2 CNN

CNN 由卷积层+激活函数、池化层和全连接层组成。其中, 卷积层的主要作用是提取特征, 不同卷积层可以提取出不同的特征; 池化层的作用是进行下采样, 减少网络中的参数数量, 从而防止过拟合; 全连接层的作用是对卷积层和池化层输出的特征图进行降维, 从而实现分类的效果。

首先, CNN 通过卷积和池化操作可以有效减少模型参数数量, 较好地处理高维输入, 解决 MLP 无法处理隐藏层参数过多的问题[53, 75, 123]。2019 年, Lucia 等人[123]在 MCFP 数据集[68]上, 使用 CNN 和 SVM 对恶意 TLS 流量进行检测, 结合早停法避免过拟合, 最终准确率和 F1 值均达到 99.91%。

其次, 针对加密流量分类任务缺少大型标记数据集的问题, 可以利用 CNN 的平移不变性, 先使用大量无标记数据集进行预训练, 保留预训练权重, 再使用少量标记数据重新训练 CNN 模型, 从而实现半监督学习, 减少对标记数据的依赖[100]。

另一方面, CNN 能够有效提取流量的空间特征。由于 CNN 在图像识别和计算机视觉领域取得了出色的成果[117], 流量检测领域的研究人员考虑将原始流量数据或流量特征转为二维灰度图像, 再输入 CNN 模型中, 以提取流量的空间特征, 实现准确分类[53, 75, 124, 135]。2017 年, Wang 等人[75]按照流的方向和所在层数, 将流量表示划分为四种形式: Flow+All, Flow + L7, Session + All 和 Session + L7, 其中 L7 表示 OSI 模型的第 7 层, All 表示所有层。其次, 将这四种流量表示的前 784 字节预处理为 28x28 维的灰度图像, 作为 2dCNN 模型的输入, 提取深层空间特征, 进行检测分类。实验结果证明, All 优于 L7, Session 优于 Flow。文

献[53]在此基础上进行改进, 将从流量中提取出的协议特征、基本特征和统计特征的归一化数值表示转换为 24x24 维的灰度图像, 并在转化过程中使用柏林噪声(Perlin Noise)对图像进行增强, 再送入 CNN 模型进行训练。这样的处理方式增强了 CNN 在恶意加密流量检测中的泛化能力和抗干扰能力。相对于定长输入, 文献[124]提出不定长卷积神经网络提取流量的空间特征, 无需对数据进行截断和补零操作, 避免部分流量信息丢失或无用信息的引入所带来的影响。然而, CNN 只能提取流量的空间特征, 而无法提取时间特征。

文本卷积神经网络(Text-CNN)[126]和图卷积神经网络(GCN)[125]属于 CNN 的其它种类。其中, GCN[125]通过同时训练流量统计特征和流量轨迹的图形结构特征, 有效地提升模型的检测效果和速度。而 Min 等人[126]使用词嵌入和 Text-CNN 从流量中提取有效载荷特征, 并与统计特征结合形成新的特征, 最终在 ISCX-IDS-2012 数据集[79]上达到 99.13%的检测率。

6.1.2.3 RNN

RNN 是训练序列数据的常用模型之一, 通过神经网络在时序上的展开, 可以找到输入样本之间的序列相关性, 有效提取流量的时序特征。传统 RNN 每一步的隐藏单元只是执行一个简单的 tanh 或 ReLU 操作, 采用随时间反向传播算法进行训练, 存在梯度消失或爆炸问题。因此, 科研人员提出 LSTM 模型, 通过引入细胞状态、隐状态, 以及遗忘门、输入门和输出门的结构, 有效地解决了长时间前的信息可能被遗忘的问题。之后 GRU 模型被提出, 使用更新门替代 LSTM 的遗忘门和输入门, 使用重置门控制前一状态有多少信息被写入, 减少了网络参数, 提高了训练效率。

首先, RNN 可以有效解决 CNN 无法提取时间序列特征的问题。因此, 诸多研究者将 CNN 和 RNN 结合使用, 形成新的检测模型框架[127-128, 136-137]。文献[136-137]使用 CNN 模型提取流量的空间维度特征, RNN 模型提取时间维度特征。文献[127]提出的 HALNet 模型则使用卷积块提取字节特征, 使用多头注意机制和 Bi-LSTM 提取全局时间特征, 使用 skipLSTM 提取局部时间特征, 有着更好的泛化能力。而文献[128]将 CNN、LSTM 和 GRU 混合组合使用, 通过对比实验证明 CNN+GRU 的组合模型表现最优。这是因为 CNN 可以快速提取空间特征, 而 GRU 提取时间特征的速度快于 LSTM。

此外, RNN 在实时检测方面有着良好的表现[86, 138]。Lee 等人[138]提出一种基于 LSTM 的能够实时主动检测时间序列异常的 RePAD 算法。RePAD 利用短期历史数据点, 预测和确定即将到来的数据点是否会发生异常。RePAD 可以随时间动态调整检测阈值, 因此能包容时间序列中的微小模式变化, 并能主动及时地检测异常, 提供早期预警, 而无需人工干预和领域知识。

另一方面, LSTM 和 GRU 对于序列数据的处理有着较好的表现, 研究人员将流量特征作为序列数据输入到 LSTM 或 GRU 中, 可以实现高精度检测[130-131,139]。其中, Ferriyan 等人[131]提出一种名为 TLS2Vec 的方法, 从原始 PCAP 中提取流量特征(SSL 版本号、密码套件列表、有效载荷长度等)构建语料库, 并使用 CBOW 和 Skip-gram 生成 300 维的词向量, 再将词向量输入 LSTM 和 BiLSTM 模型进行分类训练。实验结果显示针对 TLS 加密流量, 二分类检测准确率能达到 99%, 多分类检测准确率平均能达到 82%。文献[130]对直接将特征向量作为输入的方法进行改进, 将特征词向量通过两层基于注意力机制的 Bi-GRU 模型, 逐步构造包含上下文信息的包向量和流向量作为分类器输入, 实验结果显示二分类任务的准确率可达 99.48%, 多分类任务的 AUC-ROC 平均值可达 0.9988。

LSTM 还可以用于复制流量样本的数字特征, 以生成新的样本, 改善数据不平衡问题[140]。

除了 LSTM 和 GRU, Ren 等人[129] 提出一种树形 RNN 模型(Tree-RNN)用于流量分类任务。Tree-RNN 通过树结构将大类分解为小类, 并为每个小类设置特定的分类器。由于使用了多个分类器, Tree-RNN 在分类性能上能够互相补充, 解决了分类器单一的问题。同时, 由于多个分类器都是端到端框架, Tree-RNN 无需人工特征提取就可以自动学习输入数据和输出数据之间的非线性关系。在 ISCX VPN-nonVPN[73] 数据集上的实验结果证明, Tree-RNN 相较于文献[111]和文献[75], 平均准确率提高了 4.88%, 具有较高的平均精度和平均召回率。未来的研究可以考虑使用 Tree-RNN 进行多分类的恶意加密流量检测。

6.1.2.4 AE

AE 是一种无监督学习的神经网络模型, 由三部分组成: 输入层, 隐藏层和输出层, 其中输入层和输出层的尺寸大小相同。AE 的目的是通过不断调整参数, 在使输出与输入尽量相同的条件下, 隐层的特征表达与输入尽量不同, 常用于降维和特征提取。AE 有很多变体, 如降噪自编码器(Denoising Auto

Encoder, D-AE)、栈式自编码器(Stacked Auto Encoder, SAE)和由 LSTM 组成的 AE(LSTM-AE)等。

一方面, AE 可以用于降维, 其与 CNN 的结合有利于提高检测效率[141]。在针对恶意加密流量检测的样本不平衡问题时, 文献[133]提出一种结合 CNN 和 AE 的小样本异常检测模型 DFAE, 通过 AE 的数据重组能力和解决了数据不平衡问题, 并提升了模型的泛化能力。

另一方面, AE 也可以用于特征提取。SAE 可以将输入层和隐藏层进行多次堆叠, 每个隐藏层的输出被用作下一个隐藏层的输入。通过这种方式, SAE 在可以第一层学习到输入的一阶特征, 在第二层学习到二阶特征, 通过迭代训练提取输入样本的多层特征。因此, SAE 可以提取流量的编码特征, 部分研究人员将 SAE 和 CNN、RNN 相结合, 以全面提取出流量的深层次特征[104, 111]。2019 年, Zeng 等人[111]提出了一个由 CNN、LSTM 和 SAE 组合而成的加密流量分类和入侵检测框架(Deep-Full-Range, DFR)。DFR 通过 CNN 提取空间特征, 通过 LSTM 提取时间特征, 通过 SAE 提取编码特征, 将这三方面的特征结合, 以获得对原始流量的全面表示。整个特征提取的过程是自动、无需人工干预的。通过在 ISCX VPN-nonVPN[73]和 ISCX-IDS-2012[79]数据集上的对比实验, 证明 DFR 有着更好的准确性和鲁棒性。

同样, LSTM-AE 有较强的编码能力和特征提取能力。针对正常流量样本有噪声, 异常流量样本较少且分布不平稳、方差大的问题, Xing 等人[132]提出了一种基于 LSTM-AE 和深度字典学习的在线恶意加密流量检测模型, D2LAD。D2LAD 通过 LSTM-AE 对流量特征表示进行顺序特征的提取, 通过深度字典学习提取正常流量的特征模式, 最后将数据和字典之间的相关性作为度量进行检测。实验结果显示 D2LAD 的检测准确率能达到 94.5%, 有着高精度和低资源利用率。文献[54]则使用 LSTM-AE 对 SSL 记录的序列进行编码, 生成每个 SSL 流的特征映射后送入 CNN 进行分类, 能够达到 95.8%的检测准确率, 展示了 LSTM-AE 优秀的编码和特征提取能力。

6.1.2.5 GAN

GAN[101]是 Goodfellow 等人提出的一种无监督学习模型, 由两个神经网络: 生成器和判别器组成。GAN 的基本思想是通过这两个神经网络间相互博弈, 最终收敛到最优解。GAN 能够生成与真实数据高度相似的可信样本, 在图像、语音和自然语言生成领域已经取得了出色的成就[102-103]。

在恶意加密流量检测领域, GAN 作为生成对抗模型, 可用于对抗学习。一方面, GAN 可以生成对抗性样本, 即在合理条件下对恶意流量进行扰动, 使其绕过分类器的检测。在文献[142]中, GAN 被输入真实的 Facebook 聊天网络流参数进行训练, 提取出正常聊天流量的特征参数传递给恶意软件, 使其根据参数要求调整流量进行伪装。如果恶意软件被检测阻止, 就将此信号作为反馈信息改进 GAN 模型。实验结果证明, GAN 可以帮助恶意软件实现流量伪装、逃避检测, 提供了自适应恶意软件和自适应入侵防御系统(Intrusion-prevention system, IPS)的可能性。同样, 文献[143]通过 WGAN 进行流量伪装, 屏蔽流量检测。但研究者不局限于一种应用流量, 通过使用 GAN 的生成器学习不同应用程序的目标流量特征, 再让代理依据学习到的目标应用特征将原流量变形成伪装流量。部分检测模型在这样的对抗学习过程中不断提升检测能力和泛化能力。文献[144]提出的 Bot-GAN 是一个基于 GAN 的增强型僵尸网络检测模型, 通过 GAN 不断生成“假”样本, 进行对抗学习。实验结果证明, Bot-GAN 有效提高了僵尸网络检测的准确率、精确率和泛化能力, 降低了误报率, 并且具有检测新型僵尸网络的能力。

另一方面, GAN 也可以实现对抗防御。Samangouei 等人[145]提出的 Defense-GAN 将原始样本作为生成器的输入, 学习原始样本的分布, 从而可以模拟未受干扰图像的分布。经过训练后的 Defense-GAN 在对抗样本输入后, 会生成满足原始样本分布的近似样本, 从而消除对抗性干扰。APE-GAN[146]的思路同样是将对抗样本重构成正常样本, 从而消除扰动、提升检测准确性。PGAN [147]则将 AE 和 GAN 相结合, 通过学习正常网络流量样本获得正常流量的评分区间, 再使用判别器分别对已知恶意流量样本和正常流量样本进行判别, 得到恶意样本的评分区间, 最终可以实现未知恶意流量的检测。然而, GAN 的训练存在不稳定、难以收敛的问题, 且超参数的选择十分重要。在未知对抗攻击的情况下, 完成对 GAN 的训练仍具有挑战性。

除了在对抗攻击与防御方面的应用, GAN 是一种常用的数据生成方法, 可以解决恶意加密流量检测的数据不平衡问题。在 3.2.1 节中已详细介绍了多种 GAN 的改进数据增强算法。其中, AC-GAN[96]的改进在于将随机噪声和类标签同时作为输入, 增强了鲁棒性; PacketCGAN[97]的改进在于以 CGAN 为基础, 可以将应用程序类型作为条件输入模型, 一次生成多种类样本; DCGAN[52]的改进在于在 GAN 模型中添加了深度卷积网络结构; Markov-GAN[98]的改

进在于生成的样本是流量的马尔可夫图像表示, 相较于普通灰度图像表示有更强的表示能力和更小的尺寸; CTTGAN[134]的改进在于, 不需要将网络流量数据转换为图像, 而是提取流量的有效特征, 其次使用CTGAN[148]扩展生成特征数据, 降低了存储成本和计算复杂度。上述 GAN 及 GAN 的改进算法效果均优于过采样和欠采样方法, 是有效的数据增强算法。

6.1.3 模型选择

近十年, 学者对恶意加密流量检测的研究从基于传统机器学习的方法, 发展到目前基于深度学习的方法。基于深度学习的方法是领域内研究的主流方向, 其效果远好于大部分传统方法。其中, 基于AE的方法可以进行数组重组和降维, 提取统计信息的编码特征; 基于CNN 的方法可以有效处理高维输入, 提取流量的空间特征; 基于 RNN 的方法可以捕获流量的时间序列特征, 实现实时检测; 基于 GAN 的方法可以平衡数据集类别分布, 进行对抗攻击和防御。

模型的选择和输入特征、数据集大小高度相关。输入特征不仅直接影响检测的准确性, 还直接影响输入维度, 影响到模型的计算复杂度和内存复杂度。第 4 节中介绍了加密流量检测领域常用的流量特征, 包括: 基本特征、时序特征、统计特征和协议特征。当数据集较小时, 大部分深度学习方法不再适用。因此, 本节在假设数据集足够大的情况下, 探讨不同流量特征或原始流量作为输入时的合适的模型选择。表 9 总结了面对不同输入相应的模型选择。

表 9 模型选择

Table 9 Model Selection

输入	模型选择	作用	复杂度	泛化能力
统计特征	传统 ML 模型/MLP	分类/聚类检测	低	较弱
时序特征/协议特征	传统 ML 模型/MLP/CNN/RNN/GAN	分类/聚类检测	低/中等	较弱/中等
原始流量	CNN/RNN/CNN+RNN	提取空间/时序特征, 分类检测	高	较强
原始流量/流量特征	AE	提取编码特征、降维	中等	较强
原始流量/流量特征	GAN	对抗攻击/防御、数据生成	高	较强

1) 统计特征: 以统计特征作为输入时, 维度通常较小。因此, 大部分研究使用传统 ML 模型或 MLP, 计算复杂度较低。由于统计特征需要观察整个流的数据包进行计算获得, 不适用于在线训练检测。

2) 时序特征/协议特征: 由于时序特征几乎不受加密的影响, 被广泛应用于各种目标场景。部分研究使用流最初的 10-30 个数据包的时序特征完成流量的分类检测[138, 149],也有研究使用流全部数据包的时序特征提高检测精度[100]。协议特征是通信过程中握手阶段暴露的未加密的相关协议、证书、加密套件信息, 已成功用于分类检测[131]。当输入维度较小时, 传统 ML 模型和 MLP 表现较好。面对高维输入, CNN、RNN 和 GAN 模型则表现更优。一般情况下, 深度学习模型的计算复杂度和训练时间都高于传统 ML 模型。

3) 原始流量: 诸多研究将原始流量的前 784/1024/1500 字节或 8-30 个原始数据包的 100/256 字节作为模型输入, 使用 CNN、RNN、AE 及其组合提取空间、时序、编码特征, 再进行分类检测[111, 128]。其中, CNN 更合适提取流量字节的空间特征, RNN 更适合提取数据包内的时序特征, AE 可用于编码和降维, 较少模型计算量。GAN 可以根据原始流量/流量特征输入生成类似的流量表示, 实现数据生成和对抗学习。面对原始流量作为高维输入, 传统 ML 模型和 MLP 不再适用。在使用原始流量作为输入的同时, 还可以使用流量的时序特征、协议特征、统计特征作为辅助输入提高检测精度。

6.2 对抗攻击和防御技术

深度学习检测模型易受到对抗性攻击, 大都缺少处理对抗样本的鲁棒性。对抗样本指的是, 在原始样本上添加一些微小的扰动而生成的新的输入样本[150]。对抗样本和原始样本在人为观察上无明显差异, 可以诱使检测模型判断错误。攻击者常将恶意流量伪装成正常流量, 生成对抗样本, 大大提升了防御方检测的难度。为抵御对抗攻击带来的安全威胁, 防御方也发展了相应的对抗防御措施。本节将从攻击和防御两个视角介绍恶意加密流量检测中的对抗技术。

6.2.1 对抗攻击

恶意加密流量的有效检测依赖于原始流量表示或流量特征表示的可用性, 因此在输入表示上添加扰动, 生成对抗样本可以误导检测模型。需要注意的是, 为了不干扰网络流量功能的实现, 只能对非功能性特征添加扰动。对抗样本包括具有对抗性的流量向量表示和能够真实传输的流量样本。通常, 流量样本到流量特征向量表示是不可逆的, 研究者很难将扰动后的流量向量, 再逆映射成真实网络流

量。因此, 两种不同的对抗样本输出对检测模型的实际威胁程度存在差异。相关对抗研究工作对比如表 10 所示。

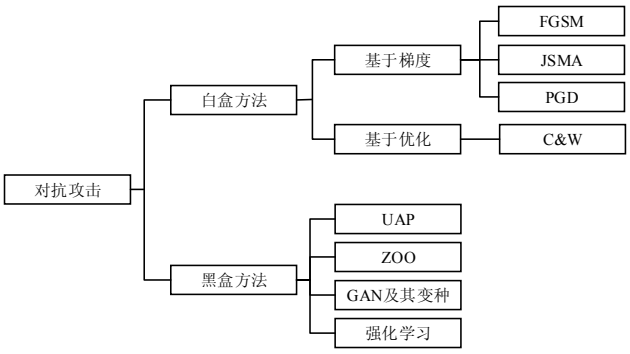

图 7　对抗攻击方法分类

Figure 7　Classification of adversarial attack methods

如图 7 所示, 常用的对抗攻击方法有: 快速梯度符号方法(Fast Gradient Sign Method, FGSM)[151]、投影梯度下降法(Projected Gradient Descent, PGD)[152]、基于雅可比矩阵的显著图攻击方法 (Jacobian-based Saliency Map Attack, JSMA)[153]、C&W 攻击方法(Carlini and Wagner Attack, C&W)[154]、通用对抗扰动(Universal Adversarial Perturbation, UAP)[155]、零阶优化(Zeroth Order Optimization, ZOO)[156]、GAN 及其变种和强化学习方法等。

白盒攻击方法, 指攻击方完全掌握防御方的训练数据、样本特征、算法模型和参数权重等详细信息, 包括 FGSM、PGD、JSMA 和 C& W 等方法。

1) FGSM[151], 基于梯度的攻击方法, 沿分类模型的梯度上升方向对原始样本添加扰动, 使模型

损失函数增加, 从而导致模型分类错误。文献[157] 迭代使用 FGSM 方法, 在仅对攻击持续时间、字节总数和包总数进行扰动的约束下, 通过代理注入所需字节数和数据包, 对流量进行修改, 使得加密的 C2 恶意软件对抗流量成功绕过检测。

2) PGD[152], 基于梯度的攻击方法, 在 FGSM 的基础上进行改进, 通过多次迭代来增大损失函数。文献[158]在 BoT-IoT 数据集上利用 FGSM、BIM[168]和 PGD 生成对抗性的流量向量表示, 实现对前馈神经网络和自归一化神经网络(Self-Normalizing Neural Networks, SNN)的攻击。

表 10　使用对抗攻击方法的文献总结

Table 10　Summary of works using adversarial attack methods

文献	攻击目标	对抗攻击方法	输出	数据集	年份
文献[157]	DNN	FGSM	流量	MTA	2020
文献[158]	FNN, SNN	FGSM, BIM, PGD	特征向量	Bot-IoT	2019
文献[159]	LeNet-5	FGSM, DeepFool, C&W	特征向量	Moore	2020
文献[160]	1D-CNN	UAP	流量	ISCX VPN-nonVPN	2020
文献[161]	DNN	C&W, ZOO, GAN	特征向量	NSL-KDD	2018
文献[142]	Stratosphere IPS	GAN	流量	C2 流量	2018
文献[143]	Hermann, VNG + +, Liberatore, Jaccard, Wrigh, 带宽分类器, 时间分类器	WGAN	流量	百度流量	2019
文献[162]	SVM, NB, MLP, LR, DT, RF, k-NN	IDSGAN	特征向量	NSL-KDD	2022
文献[163]	LeNet, VGG, ResNet, DenseNet, ShuffleNet, MobileNetV2	FGSM, BIM, C&W, DeepFool, AdvGAN	流量向量	ISCX VPN-nonVPN	2022
文献[164]	MLP, DT, LR, SVM	Attack-GAN	流量	CTU-13	2021
文献[165]	Winner[166], Novel[167]	Actor-Critic Algorithm	流量	VirusTotal	2021

3) JSMA[153], 基于梯度的攻击方法, 利用雅克比矩阵计算得到不同输入特征对分类结果的影响程度, 选择在影响程度最大的特征上进行扰动, 能够有效减少需要被扰动的特征数。

4) C&W[154], 基于优化的攻击方法, 改进 L-BFGS 算法的损失函数, 设计了 3 种优化算法, 兼具高攻击准确率和低对抗扰动性, 能够攻破防御蒸馏, 但需要消耗大量的计算时间。文献[159]在 Moore 数据集上, 使用 FGSM、DeepFool 和 C&W 方法向流量特征图表示中添加扰动, 生成对抗流量向量表示, 成功欺骗了 LeNet-5 深度卷积神经网络, 实现了对网络流量分类模型的攻击。

黑盒攻击方法, 指攻击方没有防御方检测模型的先验知识, 仅能通过访问模型获取模型输出, 更

符合现实网络攻防情况。灰盒攻击方法, 指攻击方能够通过访问模型获得不同程度的模型信息, 包括模型的特征空间和训练集分布, 但没有模型的确切信息。常用的黑盒/灰盒攻击有: UAP、ZOO、GAN 及其变种和强化学习等方法。

1) UAP[155], 一种通用扰动的构造方法, 通过对数据集上的数据点进行迭代 DeepFool 攻击[169], 直至错误率达标, 从而发现能使大部分数据都被误判的通用扰动, 具有很强的泛化能力。DeepFool 通过迭代生成尽可能小的扰动, 直至对抗样本越过分类器的决策边界, 被分类错误。文献[160]针对网络流量分类的 3 种输入空间: 数据包、流内容和流时间序列, 提出了三种应用 UAP 生成对抗流量样本的方法: AdvPad、AdvPay 和 AdvBurst。该方法在仅将扰动注入到输入的某些特定部分, 例如数据包的末尾或虚拟数据包的情况下, 生成对抗流量, 降低了 1D-CNN 模型的分类性能。

2) ZOO[156], 使用零阶优化对目标模型的梯度进行估计, 而无需训练替代模型。文献[161]在 NSL-KDD 数据集上使用 C&W、ZOO 和 GAN 生成对抗性的流量向量表示, 实现对 DNN 检测模型的攻击。

3) GAN 及其变种, 通过生成器学习对抗特征, 或生成扰动构造对抗样本。a)文献[142]利用 GAN 学习正常 Facebook 网络流特征, 将对抗特征传递给恶意代码控制端, 使其根据需求生成能伪装成正常 Facebook 流量的恶意对抗流量, 从而绕过检测, 这需要对源码进行复杂的修改。b)文献[143]使用 WGAN[170]生成目标流量的伪装特征, 构造相应的目标流量模式, 再通过代理系统依据流量模式将流量变形, 能够将流量伪装成任意正常目标流量, 屏蔽流量检测。c)文献[162]在 WGAN 的基础上提出 IDSGAN 框架, 为保证原始恶意流量的攻击有效性, 仅对非功能特征进行修改, 生成对抗性的流量特征表示, 有效降低 SVM、NB、MLP、LR、DT、RF 和 k-NN 的检测率。d)AdvGAN[171]由生成器生成扰动, 构造对抗样本, 再迭代计算对抗样本输入目标模型或蒸馏网络后输出的误判损失, 鼓励对抗样本被错误分类到目标类别中, 也可以通过最大化预测值与真实值之间的距离来进行无目标攻击。文献[163]在 ISCX VPN-nonVPN 数据集上使用 FGSM、BIM、C&W、DeepFool 和 AdvGAN 算法, 生成对抗性的流量向量表示, 对 LeNet、VGG、ResNet、DenseNet、ShuffleNet 和 MobileNetV2 进行有效攻击。实验结果显示, AdvGAN 的攻击效率最高, 速度最快, 仅需 0.5 s 就能生成 1253 个对抗样本。e)不同于文献[163]将网络流量转化为灰度图的方法, 文献[164]在 SeqGAN[172]的结构基础上提出 Attack-GAN 生成数据包级的对抗网络流量, 将数据包的每个字节视为序列中的一个标记, 从而将生成对抗数据包的过程转为一个顺序的决策过程。作者使用黑箱 NIDS 返回的结果来帮助更新生成器的参数, 在不修改恶意功能字节的条件下, 生成对抗流量、逃避检测, 但存在对抗流量长度固定不变的问题。

4) 强化学习: 文献[165]使用基于强化学习的 A3C-MTAG 模型生成对抗恶意流量样本, 欺骗恶意软件分类模型。A3C-MTAG 的动作空间包括: 对流量样本数据包的增减操作和对目标主机的网络服务类型、连接进程的端口号等进行修改操作。智能体依据目前检测器的奖励反馈从动作空间选取下一个动作, 在保留恶意功能的前提下, 迭代修改恶意流量样本, 直到目标检测器被成功绕过或达到设置的最大修改次数为止。

通过对上述研究进行对比分析, 可以发现: 1)以流量向量表示为输出的攻击方法, 虽然对检测模型的实际威胁较弱, 但由于恶意流量特征的选择具有代表性和相似性, 经验丰富的领域专家可以对典型的恶意特征进行扰动, 进而实现对模型的攻击。2)以真实对抗流量为输出的攻击方法更具现实价值, 主要包括: a)通过代理对流量数据包进行增减、修改操作, 从而实现对流量的修改[157, 160, 165]; b)通过 GAN 学习对抗特征传递给代理, 使其依据相应特征模式生成对抗恶意流量, 需要对恶意代码进行修改[142-143]; c)通过 GAN 实现对抗数据包的字节生成, 最终组成对抗流量[164]。

6.2.2 对抗防御

为降低对抗攻击对检测模型的影响, 防御方需选择合适的对抗防御方法。如图 8 所示, 对抗防御方法可以分为两个主要方向: 一是处理数据, 二是改进检测模型。

处理数据是指在对抗样本输入检测模型前, 对输入数据进行处理, 包括: 数据压缩和消除扰动。

1) 数据压缩, 是将模型输入数据包括大小、特征进行压缩, 以减少扰动对模型的影响。在图像处理领域, Dziugaite 等人[173]提出将输入样本的图像进行 jpg 压缩, 可以减少对抗攻击对模型的影响。在恶意流量检测领域, 文献[174]选择删除鲁棒性得分较低的特征进行特征压缩, 以减少攻击者可扰动的范围, 起到了较好的防御效果。

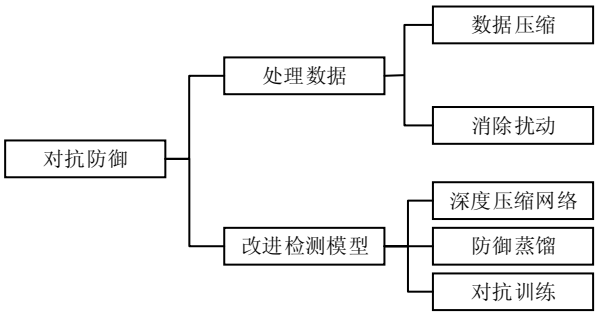

图 8　对抗防御方法分类

Figure 8　Classification of adversarial defense methods

2) 消除扰动, 是在分类模型前串联一个能将对抗样本恢复成原始样本的模型网络, 使得分类模型能够对原始样本进行检测。文献[163]在分类模型前使用降噪自编码器 D-AE, 消除对抗样本拥有的噪声或扰动, 将对抗样本重构成原始样本。这是因为 D-AE 能够为高维原始数据生成低维特征表示, 并要求低维特征表示保留原始数据中最重要的信息。而 6.1.2.5 节中提到的 Defense-GAN[145]和 APE-GAN[146]是基于 GAN 的消除对抗性扰动的方法。

改进检测模型是指通过在深度神经网络中添加更多的层, 或者对模型进行再训练, 从而提升检测模型的泛化能力和鲁棒性, 包括: 深度压缩网络(Deep Contractive Networks, DCN)、防御蒸馏和对抗训练。

1) DCN[175], 是一种端到端的深度压缩网络模型, 利用压缩自动编码器的平滑惩罚项, 使得输出更加平滑, 增加了网络对对抗样本的鲁棒性。

2) 防御蒸馏[176], 是使用相同的模型结构来训练原始网络和蒸馏网络, 将原始网络学习到的先验知识迁移到蒸馏网络上, 可以增强网络的泛化能力、抵抗小幅度扰动的对抗攻击, 但对黑盒攻击失效。

3) 对抗训练[152], 是将对抗样本加入模型的训练集中, 重新训练分类模型以提高模型的鲁棒性。由于需要提前构造好对抗样本, 对抗训练需要大量的时间, 但对于已知攻击如 FGSM 和 PGD 的防御效果较好。文献[177]提出的批次对抗训练和增强对抗训练方法, 能够在一次反向传播过程中同时完成样本梯度和参数梯度的计算, 提高训练效率, 有效防御针对 CNN 模型的白盒/黑盒攻击。

6.3　总结

深度学习的迅速发展为恶意加密流量检测和对抗带来了新的机遇与挑战。目前, 针对恶意流量检测的攻防对抗研究丰富, 但主要针对传统机器学习模型和简单的深度学习模型, 如决策树、随机森林、SVM 和 DNN 等。针对深度学习模型的恶意加密流量检测的攻防对抗研究还处于初期阶段, 值得深入探索。基于深度学习的恶意加密流量检测方法能够自动提取流量特征, 有效减少人工成本、提高检测的准确率和精度, 但仍存在改进的空间, 具体如下所示。

(1) 提升模型的可解释性

大多深度学习模型是由数据驱动的黑盒模型, 具有高度非线性性质。深度学习模型的每个神经元都是由上一层的线性组合再加上一个非线性函数得到, 无法通过统计学知识去理解神经网络中的参数含义及其重要程度、波动范围。因此, 通过深度学习方法从流量提取出的特征是深层次、有效的, 但可解释性较差。如何提高深度学习模型的可解释性还需长期研究。

(2) 轻量级深度学习模型

大部分深度学习模型的训练依赖大规模数据集, 会造成巨大的计算存储开销。此外, 随着深度学习模型的复杂度增加, 模型的参数量增加, 从而导致计算机内存消耗的增加。当参数数量过多时, 可能出现模型在训练过程中直接报错, 无法继续正常训练的情况。随着网络规模的扩大, 分布式训练已常态化。因此, 在时间复杂度和资源消耗方面进行优化, 进行轻量级深度学习模型的设计是未来发展方向之一。

(3) 数据集的准确标记和平衡

有监督深度学习模型的训练依赖于全面的, 平衡的, 被准确标记的数据集, 数据集的质量直接影响模型的检测性能。而现实的网络流量环境复杂多变, 恶意加密流量远少于正常流量, 且由于恶意流量的动态、隐蔽性, 难以及时进行收集和标记。因此, 针对有监督深度学习方法, 对收集到的流量数据进行准确标记和平衡是未来改进的方向之一。

(4) 动态网络环境下模型的快速训练

当前, 许多流量检测模型依赖于长时间的离线训练。然而, 现实情况下的网络流量远比静态数据集中的内容复杂。在动态网络环境下, 如何实现模型的快速再训练, 是未来面临的一个挑战。通常, 深度学习模型的训练阶段耗费大量资源和时间。当网络环境变化时, 深度学习模型进行重新训练引发的时间消耗问题需要高度关注。正因如此, 能够快速适应环境动态变化, 进行动态特征选择的自适应模型是未来的发展方向之一。

(5) 面向移动网络场景

目前, 深度学习方法已广泛应用于互联网加密

流量研究, 但对于移动和无线网络流量的应用研究较少。然而, 随着无线通信技术的发展, 移动网络的通信速度和通信容量大大提升, 智能手机已成为人们生活中不可或缺的一部分。由于无线移动网络流量蕴含诸多隐私信息, 如交易密码和地理位置信息, 是攻击者重点关注的目标。目前, 基于深度学习的恶意加密流量检测大都在网络层、传输层和应用层之上, 如果能在无线空口或数据链路层提前检测出恶意流量, 将大大缩短检测到恶意攻击的时间, 也给安全人员更多的反应处理时间。因此, 基于深度学习方法的研究可以更多考虑面向移动网络场景。

7 开放问题与挑战

随着人们网络安全意识的日益增强和加密技术的广泛应用, 网络中的加密流量呈现爆炸式增长。然而, 加密策略在保护用户隐私的同时, 也增大了恶意流量被检测和发现的难度。因此, 对于恶意加密流量的检测研究尤为重要。本文在深入查阅现有研究的基础上, 提出了一个通用的恶意加密流量检测框架, 该框架包括研究目标确定、数据收集、数据预处理、特征提取和选择、模型选择和实验评估这六方面。在此框架的基础上, 本文整理分析了现有的 20 个开源数据集, 总结了 4 种数据收集技术, 并提供了数据不平衡问题的解决方案。同时, 本文对常用的加密流量检测特征和特征选择方法进行分类总结。此外, 本文对不同检测模型的适用性和优缺点进行对比分析, 并对流量检测中的对抗攻击和防御技术进行归纳。目前, 针对恶意加密流量检测的研究已取得了一定的进展, 但仍存在一些未解决的问题和挑战, 也是未来研究方向。

(1) 标准流量数据集构建

一个新颖、真实、全面且平衡的数据集是恶意加密流量检测领域开展研究的重要保障。当前研究广泛使用的数据集, 如 malware-traffic-analysis.net、CTU-13、CIC-IDS-2017 等, 存在不完全针对加密流量、无法全面覆盖新型攻击方法等问题。许多研究采用非开源的私有数据集, 然而同一算法在不同数据集上的结果往往具有偏差, 这给研究间的对比带来了公平性的挑战。针对类不平衡问题, 本文已探讨诸多解决方案, 但仍有改进的空间。因此如何构建一个优质、标准、公认的恶意加密流量数据集, 是目前亟需解决的课题。

(2) 降低标注样本需求

当前研究大都是基于监督学习的方式, 依赖于大量的标注数据。然而, 真实网络流量数据往往是无标注的, 有标注数据难以获得。随着恶意攻击行为的变化和加密协议的发展, 对不同的恶意加密流量进行准确标注的难度上升。因此如何将大量未标注数据与标注数据相结合, 以半监督学习方式进行恶意加密流量检测, 消除对大量标注数据集的依赖, 是值得进一步研究的课题。

(3) 新型加密协议带来的挑战

随着加密技术的发展, Quic 和 TLS 1.3 等加密协议不断演进, 逐步实现了 0-RTT 数据传输, 并对握手消息进行了加密操作。这导致建立传输过程的第一个数据包中的可见明文信息大大减少, 传统的协议特征将不再适用。目前, Chrome 和 Firefox 都已支持 TLS 1.3, 而未来也会有更多的应用程序和网站采用这些更强大的加密协议, 这将为恶意加密流量检测带来新的挑战。

同时, 随着加密协议和专有协议的类型不断增多, 仅对一种或两种加密协议进行分析不足以应对未来的各种情况。

(4) 复杂流量环境下准确识别

现实网络环境中的流量相较于静态训练数据集更为复杂, 其中存在多种类型的流量数据和加密协议。恶意攻击也在不断变种发展, 现实网络中可能出现训练集中未出现的恶意攻击流量。这种情况下, 离线训练的检测模型的识别准确率会受到影响。复杂的现实网络环境是动态变化的, 存在诸多不确定因素。因此, 如何在复杂流量环境下进行高效准确地识别, 以及如何对模型进行快速重训练也是一个挑战。

(5) 恶意加密流量精细化检测

目前的加密流量检测通常是二元检测, 即区分恶意与良性流量, 而在多类检测任务方面的研究亟待深入。当前, 多类检测主要集中在大类上, 如恶意软件家族、僵尸网络等, 需要进行进一步的细粒度划分检测。检测模型需要具备对多种类型的恶意流量进行精细化识别的能力。因此, 如何精准定位恶意流量类型, 明确攻击种类并指导防御方快速采取有效措施, 是未来的挑战之一。

(6) 无线/移动网络环境适应

目前的恶意加密流量检测大都针对互联网流量, 而无线/移动网络流量与其有一定的区别, 部分互联网流量特征将不再适用。而随着无线通信技术的发展, 无线/移动网络与人们的生活密切相关。因此, 针对无线/移动网络流量进行相关特征提取和恶意流量检测也是未来的发展方向之一。

(7) 恶意对抗及应对

已经有相关研究从对抗的角度, 在遵守网络协

议规范、维持恶意功能和保证一致性的约束下, 对恶意流量进行混淆和扰动, 从而成功绕过检测模型, 对系统实施攻击。因此, 如何利用深度学习模型进行对抗防御, 增强检测模型的鲁棒性是值得进一步研究的课题。良好的恶意加密流量检测模型不仅能够检测出已训练过的已知攻击, 还能够检测出未知攻击、混淆攻击和对抗攻击。

(8) 抗干扰特征设计

在对抗场景下, 恶意流量通过模拟正常流量特征逃避检测, 导致部分流量的基础、时序、统计和协议特征失效。因此, 挖掘具有鲁棒性的, 不易被干扰的特征是当前对抗研究的重点。

(9) 实时检测

目前, 恶意加密流量检测模型很多需要先使用静态数据集进行离线训练, 再实现在线检测。然而, 现实网络的情况更加复杂多变, 离线训练的模型在现实网络环境上线后的性能不能保证。因此, 能够进行实时训练和动态调整的自适应模型是未来发展的重要趋势。

致　谢　感谢国家重点研发计划项目(No. 2021YFB 2910105)的资助。

参考文献

[1] CNNIC. The 50th Statistical Report on China's Internet Development.[EB/OL].http://www3.cnnic.cn/NMediaFile/2022/0916/MAIN1663313008837KWI782STQL.pdf.Aug.2022. (中国互联网络信息中心(CNNIC). 第 50 次中国互联网络发展现状统计报告. [EB/OL].http://www3.cnnic.cn/NMediaFile/2022/0916/MAIN1663313008837KWI782STQL.pdf. Aug. 2022.)

[2] GOOGLE. Google transparency report. https://transparencyreport.google.com/https/overview?hl=en. Dec. 2022.

[3] Lu G, Guo R H, Zhou Y, et al. Review of Malicious Traffic Feature Extraction[J]. *Netinfo Security*, 2018, 18(9): 1-9.
(鲁刚, 郭荣华, 周颖, 等. 恶意流量特征提取综述[J]. *信息网络安全*, 2018, 18(9): 1-9.)

[4] Li Y M, Guo H, Hou J G, et al. A Survey of Encrypted Malicious Traffic Detection[C]. *2021 International Conference on Communications*, *Computing*, *Cybersecurity*, *and Informatics*, 2021: 1-7.

[5] CYREN. Malware: evasive tactics challenging traditional security. https://www.cyren.com/solutions/malware-protection. Feb. 2022.

[6] DESAI D. Spoiler: New ThreatLabz Report Reveals Over 85% of Attacks Are Encrypted. ThreatLabz State of Encrypted Attacks 2022 Report. https://www.zscaler.com/blogs/security-research/2022-encrypted-attacks-report. Dec. 2022.

[7] WATCHGUARD. WatchGuard's Threat Lab Analyzes the Latest Malware and Internet Attacks. https://www.watchguard.com/wgrd-resource-center/security-report-q4-2021. Dec. 2021.

[8] Creech G, Hu J K. A Semantic Approach to Host-Based Intrusion Detection Systems Using Contiguousand Discontiguous System Call Patterns[J]. *IEEE Transactions on Computers*, 2014, 63(4): 807-819.

[9] Wang Y P, Zhang Z B, Guo L, et al. Using Entropy to Classify Traffic More Deeply[C]. *2011 IEEE Sixth International Conference on Networking*, *Architecture*, *and Storage*, 2011: 45-52.

[10] Korczyński M, Duda A. Markov Chain Fingerprinting to Classify Encrypted Traffic[C]. *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, 2014: 781-789.

[11] van Ede T, Bortolameotti R, Continella A, et al. FlowPrint: Semi-Supervised Mobile-App Fingerprinting on Encrypted Network Traffic[C]. *Proceedings 2020 Network and Distributed System Security Symposium*, 2020.

[12] STRASáK F. Detection of HTTPS malware traffic [J]. *Czech Technical University in Prague, Bachelor project assignment*, 2017.

[13] Papadogiannaki E, Halevidis C, Akritidis P, et al. OTTer: A Scalable High-Resolution Encrypted Traffic Identification Engine[M]. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018: 315-334.

[14] Shbair W M, Cholez T, François J, et al. Improving SNI-Based HTTPS Security Monitoring[C]. *2016 IEEE 36th International Conference on Distributed Computing Systems Workshops*, 2016: 72-77.

[15] Stergiopoulos G, Talavari A, Bitsikas E, et al. Automatic Detection of Various Malicious Traffic Using Side Channel Features on TCP Packets[M]. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018: 346-362.

[16] Dai R, Gao C, Lang B, et al. SSL Malicious Traffic Detection Based on Multi-View Features[C]. *The 2019 9th International Conference on Communication and Network Security*, 2019.

[17] Shekhawat A S, Di Troia F, Stamp M. Feature Analysis of Encrypted Malicious Traffic[J]. *Expert Systems with Applications*, 2019, 125: 130-141.

[18] Anderson B, McGrew D. Machine Learning for Encrypted Malware Traffic Classification: Accounting for Noisy Labels and Non-Stationarity[C]. *The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

[19] Niu W N, Zhuo Z L, Zhang X S, et al. A Heuristic Statistical Testing Based Approach for Encrypted Network Traffic Identification[J]. *IEEE Transactions on Vehicular Technology*, 2019, 68(4): 3843-3853.

[20] Meghdouri F, Vázquez F I, Zseby T. Cross-Layer Profiling of Encrypted Network Data for Anomaly Detection[C]. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics*, 2020: 469-478.

[21] Li H H, Zhang S G, Song H, et al. Robust Malicious Encrypted Traffic Detection Based with Multiple Features[J]. *Journal of Cyber Security*, 2021, 6(2): 129-142.
(李慧慧, 张士庚, 宋虹, 等. 结合多特征识别的恶意加密流量检测方法[J]. *信息安全学报*, 2021, 6(2): 129-142.)

[22] Luo Z M, Xu S B, Liu X D. Scheme for Identifying Malware Traffic with TLS Data Based on Machine Learning[J]. *Chinese Journal of Network and Information Security*, 2020, 6(1): 77-83.
(骆子铭, 许书彬, 刘晓东. 基于机器学习的 TLS 恶意加密流量

检测方案[J]. *网络与信息安全学报*, 2020, 6(1): 77-83.)

[23] Zou F T, Yu T D, Xu W L. Encrypted Malicious Traffic Detection Based on Hidden Markov Model[J]. *Journal of Software*, 2022, 33(7): 2683-2698.
(邹福泰, 俞汤达, 许文亮. 基于隐马尔可夫模型的加密恶意流量检测[J]. *软件学报*, 2022, 33(7): 2683-2698.)

[24] MüNZ G, LI S, CARLE G. Traffic anomaly detection using k-means clustering [C] *GI/ITG Workshop MMBnet*, 2007.

[25] GU G, PERDISCI R, ZHANG J, et al. Botminer: Clustering analysis of network traffic for protocol-and structure-independent botnet detection [J]. 2008.

[26] Maiolini G, Baiocchi A, Iacovazzi A, et al. Real Time Identification of SSH Encrypted Application Flows by Using Cluster Analysis Techniques[M]. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009: 182-194.

[27] Zhang M, Zhang H L, Zhang B, et al. Encrypted Traffic Classification Based on an Improved Clustering Algorithm[M]. Communications in Computer and Information Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013: 124-131.

[28] Berkay Celik Z, Walls R J, McDaniel P, et al. Malware Traffic Detection Using Tamper Resistant Features[C]. *MILCOM 2015 - 2015 IEEE Military Communications Conference*, 2015: 330-335.

[29] Chen L C, Gao S, Liu B X, et al. THS-IDPC: A Three-Stage Hierarchical Sampling Method Based on Improved Density Peaks Clustering Algorithm for Encrypted Malicious Traffic Detection[J]. *The Journal of Supercomputing*, 2020, 76(9): 7489-7518.

[30] Hu J W, Che X, Zhou M, et al. Incremental Clustering Method Based on Gaussian Mixture Model to Identify Malware Family[J]. *Journal on Communications*, 2019, 40(6): 148-159.
(胡建伟, 车欣, 周漫, 等. 基于高斯混合模型的增量聚类方法识别恶意软件家族[J]. *通信学报*, 2019, 40(6): 148-159.)

[31] Liu J Y, Zeng Y Z, Shi J Y, et al. MalDetect: A Structure of Encrypted Malware Traffic Detection[J]. *Computers*, *Materials & Continua*, 2019, 60(2): 721-739.

[32] Conti M, Mancini L V, Spolaor R, et al. Can't You Hear Me Knocking: Identification of User Actions on Android Apps *via* Traffic Analysis[C]. *The 5th ACM Conference on Data and Application Security and Privacy*, 2015.

[33] Conti M, Mancini L V, Spolaor R, et al. Analyzing Android Encrypted Network Traffic to Identify User Actions[J]. *IEEE Transactions on Information Forensics and Security*, 2016, 11(1): 114-125.

[34] Yuan Y L, Kaklamanos G, Hogrefe D. A Novel Semi-Supervised Adaboost Technique for Network Anomaly Detection[C]. *The 19th ACM International Conference on Modeling*, *Analysis and Simulation of Wireless and Mobile Systems*, 2016.

[35] Liu J Y, Tian Z Y, Zheng R F, et al. A Distance-Based Method for Building an Encrypted Malware Traffic Identification Framework[J]. *IEEE Access*, 2019, 7: 100014-100028.

[36] Wang W, Shang Y Y, He Y Z, et al. BotMark: Automated Botnet Detection with Hybrid Analysis of Flow-Based and Graph-Based Traffic Behaviors[J]. *Information Sciences*, 2020, 511: 284-296.

[37] Zhang X Q, Zhao M, Wang J Y, et al. Deep-Forest-Based Encrypted Malicious Traffic Detection[J]. *Electronics*, 2022, 11(7): 977.

[38] Velan P, Čermák M, Čeleda P, et al. A Survey of Methods for Encrypted Traffic Classification and Analysis[J]. *International Journal of Network Management*, 2015, 25(5): 355-374.

[39] Pacheco F, Exposito E, Gineste M, et al. Towards the Deployment of Machine Learning Solutions in Network Traffic Classification: A Systematic Survey[J]. *IEEE Communications Surveys & Tutorials*, 2019, 21(2): 1988-2014.

[40] Rezaei S, Liu X. Deep Learning for Encrypted Traffic Classification: An Overview[J]. *IEEE Communications Magazine*, 2019, 57(5): 76-81.

[41] Conti M, Li Q Q, Maragno A, et al. The Dark Side(-Channel) of Mobile Devices: A Survey on Network Traffic Analysis[J]. *IEEE Communications Surveys & Tutorials*, 2018, 20(4): 2658-2713.

[42] Zhang C Y, Patras P, Haddadi H. Deep Learning in Mobile and Wireless Networking: A Survey[J]. *IEEE Communications Surveys & Tutorials*, 2019, 21(3): 2224-2287.

[43] Wang P, Chen X J, Ye F, et al. A Survey of Techniques for Mobile Service Encrypted Traffic Classification Using Deep Learning[J]. *IEEE Access*, 2019, 7: 54024-54033.

[44] Aceto G, Ciuonzo D, Montieri A, et al. Toward Effective Mobile Encrypted Traffic Classification through Deep Learning[J]. *Neurocomputing*, 2020, 409: 306-315.

[45] Rodríguez E, Otero B, Gutiérrez N, et al. A Survey of Deep Learning Techniques for Cybersecurity in Mobile Networks[J]. *IEEE Communications Surveys & Tutorials*, 2021, 23(3): 1920-1955.

[46] Berman D S, Buczak A L, Chavis J S, et al. A Survey of Deep Learning Methods for Cyber Security[J]. *Information*, 2019, 10(4): 122.

[47] Abbasi M, Shahraki A, Taherkordi A. Deep Learning for Network Traffic Monitoring and Analysis (NTMA): A Survey[J]. *Computer Communications*, 2021, 170: 19-41.

[48] Shen M, Ye K, Liu X T, et al. Machine Learning-Powered Encrypted Network Traffic Analysis: A Comprehensive Survey[J]. *IEEE Communications Surveys & Tutorials*, 2023, 25(1): 791-824.

[49] Zhai M F, Zhang X M, Zhao B. Survey of Encrypted Malicious Traffic Detection Based on Deep Learning[J]. *Chinese Journal of Network and Information Security*, 2020, 6(3): 66-77.
(翟明芳, 张兴明, 赵博. 基于深度学习的加密恶意流量检测研究[J]. *网络与信息安全学报*, 2020, 6(3): 66-77.)

[50] Wang Z H, Fok K W, Thing V L L. Machine Learning for Encrypted Malicious Traffic Detection: Approaches, Datasets and Comparative Study[J]. *Computers & Security*, 2022, 113: 102542.

[51] WANG Z, FOK K W, THING V L L. Composed Encrypted Malicious Traffic Dataset for machine learning based encrypted malicious traffic analysis [C], 2021.

[52] Luo W, Liu Z H, Zhao R, et al. Malicious HTTPS Traffic Classification Algorithm Based on DCGAN$_1$D-CNN[C]. *2021 IEEE Conference on Telecommunications*, *Optics and Computer Science*, 2021: 20-25.

[53] Bazuhair W, Lee W. Detecting Malign Encrypted Network Traffic Using Perlin Noise and Convolutional Neural Network[C]. *2020 10th Annual Computing and Communication Workshop and Con-*

ference, 2020: 200-206.

[54] Yang J, Lim H. Deep Learning Approach for Detecting Malicious Activities over Encrypted Secure Channels[J]. *IEEE Access*, 2021, 9: 39229-39244.

[55] Anderson B, Paul S, McGrew D. Deciphering Malware's Use of TLS (without Decryption)[J]. *Journal of Computer Virology and Hacking Techniques*, 2018, 14(3): 195-211.

[56] Pinheiro A J, de M Bezerra J, Burgardt C A P, et al. Identifying IoT Devices and Events Based on Packet Length from Encrypted Traffic[J]. *Computer Communications*, 2019, 144: 8-17.

[57] Hafeez I, Antikainen M, Ding A Y, et al. IoT-KEEPER: Detecting Malicious IoT Network Activity Using Online Traffic Analysis at the Edge[J]. *IEEE Transactions on Network and Service Management*, 2020, 17(1): 45-59.

[58] Shafiq M, Tian Z H, Bashir A K, et al. CorrAUC: A Malicious Bot-IoT Traffic Detection Method in IoT Network Using Machine-Learning Techniques[J]. *IEEE Internet of Things Journal*, 2021, 8(5): 3242-3254.

[59] Zhao Y, Yang Y R, Tian B, et al. Edge Intelligence Based Identification and Classification of Encrypted Traffic of Internet of Things[J]. *IEEE Access*, 1895, 9: 21895-21903.

[60] Guan H N. Research on space network security system and key technologies based on LEO[D]. Shanghai: Shanghai Jiao Tong University, 2014.
(关汉男. 基于 LEO 的空间网络安全体系及关键技术研究[D]. 上海: 上海交通大学, 2014.)

[61] Usman M, Qaraqe M, Asghar M R, et al. Mitigating Distributed Denial of Service Attacks in Satellite Networks[J]. *Transactions on Emerging Telecommunications Technologies*, 2020, 31(6): e3936.

[62] LI J, ZHAO Z, LI R. A machine learning based intrusion detection system for software defined 5G network [J]. *arXiv preprint arXiv:170804571*, 2017.

[63] Nakahara M, Okui N, Kobayashi Y, et al. Machine Learning Based Malware Traffic Detection on IoT Devices Using Summarized Packet Data[C]. *The 5th International Conference on Internet of Things*, *Big Data and Security*, 2020.

[64] CARRIER T, VICTOR P, TEKEOGLU A, et al. Detecting Obfuscated Malware using Memory Feature Engineering [C]. *ICISSP*, 2022.

[65] MontazeriShatoori M, Davidson L, Kaur G, et al. Detection of DoH Tunnels Using Time-Series Classification of Encrypted Traffic[C]. *2020 IEEE Intl Conf on Dependable*, *Autonomic and Secure Computing*, *Intl Conf on Pervasive Intelligence and Computing*, *Intl Conf on Cloud and Big Data Computing*, *Intl Conf on Cyber Science and Technology Congress*, 2020: 63-70.

[66] GARCíA S, PARMISANO A, ERQUIAGA M J. IoT-23: A labeled dataset with malicious and benign IoT network traffic [C]. 2020.

[67] Jason Stroschein Public Github Malware Samples. https://github.com/jstrosch/malware-samples. Dec. 2020.

[68] LABORATORY S. Malware Capture Facility Project. https://mcfp.felk.cvut.cz/publicDatasets/datasets.html. Nov. 2020.

[69] Hamza A, Gharakheili H H, Benson T A, et al. Detecting Volumetric Attacks on loT Devices *via* SDN-Based Monitoring of MUD Activity[C]. *The 2019 ACM Symposium on SDN Research*, 2019.

[70] Sharafaldin I, Habibi Lashkari A, Ghorbani A A. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization[C]. *The 4th International Conference on Information Systems Security and Privacy*, 2018.

[71] Lashkari A H, Kadir A F A, Taheri L, et al. Toward Developing a Systematic Approach to Generate Benchmark Android Malware Datasets and Classification[C]. *2018 International Carnahan Conference on Security Technology*, 2018: 1-7.

[72] CIDDS - COBURG INTRUSION DETECTION DATA SETS. https://www.hs-coburg.de/forschung/forschungsprojekte-oeffentlich/informationstechnologie/cidds-coburg-intrusion-detection-data-sets.html. Dec. 2017.

[73] Draper-Gil G, Lashkari A H, Mamun M S I, et al. Characterization of Encrypted and VPN Traffic Using Time-Related Features[C]. *The 2nd International Conference on Information Systems Security and Privacy*, 2016.

[74] LASHKARI A H, DRAPER-GIL G, MAMUN M S I, et al. Characterization of tor traffic using time-based features [C] *CISSP*, 2017.

[75] Wang W, Zhu M, Zeng X W, et al. Malware Traffic Classification Using Convolutional Neural Network for Representation Learning[C]. *2017 International Conference on Information Networking*, 2017: 712-717.

[76] Moustafa N, Slay J. UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems (UNSW-NB15 Network Data Set)[C]. *2015 Military Communications and Information Systems Conference*, 2015: 1-6.

[77] Biglar Beigi E, Hadian Jazi H, Stakhanova N, et al. Towards Effective Feature Selection in Machine Learning-Based Botnet Detection Approaches[C]. *2014 IEEE Conference on Communications and Network Security*, 2014: 247-255.

[78] MALWARE-TRAFFIC-ANALYSIS.NET. https://malware-traffic-analysis.net/. Apr. 2023.

[79] Shiravi A, Shiravi H, Tavallaee M, et al. Toward Developing a Systematic Approach to Generate Benchmark Datasets for Intrusion Detection[J]. *Computers & Security*, 2012, 31(3): 357-374.

[80] García S, Grill M, Stiborek J, et al. An Empirical Comparison of Botnet Detection Methods[J]. *Computers & Security*, 2014, 45: 100-123.

[81] Saad S, Traore I, Ghorbani A, et al. Detecting P2P Botnets through Network Behavior Analysis and Machine Learning[C]. *2011 Ninth Annual International Conference on Privacy*, *Security and Trust*, 2011: 174-180.

[82] Tavallaee M, Bagheri E, Lu W, et al. A Detailed Analysis of the KDD CUP 99 Data Set[C]. *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2009: 1-6.

[83] Traffic Data from Kyoto University's Honeypots. http://www.takakura.com/Kyoto_data/. Jan. 2006.

[84] Lopez-Martin M, Carro B, Sanchez-Esguevillas A, et al. Network Traffic Classifier with Convolutional and Recurrent Neural Networks for Internet of Things[J]. *IEEE Access*, 2017, 5: 18042-18050.

[85] Aung Y L, Tiang H H, Wijaya H, et al. Scalable VPN-Forwarded

Honeypots: Dataset and Threat Intelligence Insights[C]. *Sixth Annual Industrial Control System Security Workshop*, 2020.

[86] Tan G S. Network Posture Malicious Traffic Prediction Evaluation Based on LSTM Recurrent Neural Network[C]. *International Conference on Electronic Information Technology*, 2022.

[87] Yu T D, Zou F T, Li L S, et al. An Encrypted Malicious Traffic Detection System Based on Neural Network[C]. *2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, 2019: 62-70.

[88] Gómez S E, Hernández-Callejo L, Martínez B C, et al. Exploratory Study on Class Imbalance and Solutions for Network Traffic Classification[J]. *Neurocomputing*, 2019, 343: 100-119.

[89] Japkowicz N. Learning from Imbalanced Data Sets: A Comparison of Various Strategies[J]. *AAAI Workshop on Learning from Imbalanced Data Sets*, 2000.

[90] Wang P, Ye F, Chen X J, et al. Datanet: Deep Learning Based Encrypted Network Traffic Classification in SDN Home Gateway[J]. *IEEE Access*, 2018, 6: 55380-55391.

[91] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic Minority Over-Sampling Technique[J]. *Journal of Artificial Intelligence Research*, 2002, 16: 321-357.

[92] Han H, Wang W Y, Mao B H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning[M]. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005: 878-887.

[93] Nguyen H M, Cooper E W, Kamei K. Borderline Over-Sampling for Imbalanced Data Classification[J]. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 2011, 3(1): 4.

[94] Sáez J A, Luengo J, Stefanowski J, et al. SMOTE–IPF: Addressing the Noisy and Borderline Examples Problem in Imbalanced Classification by a re-Sampling Method with Filtering[J]. *Information Sciences*, 2015, 291: 184-203.

[95] YAN B, HAN G, HUANG Y, et al. New traffic classification method for imbalanced network data [J]. *Journal of Computer Applications*, 2018, 38(1): 20.

[96] Vu L, Bui C T, Nguyen Q U. A Deep Learning Based Method for Handling Imbalanced Problem in Network Traffic Classification[C]. *The Eighth International Symposium on Information and Communication Technology*, 2017.

[97] Wang P, Li S H, Ye F, et al. PacketCGAN: Exploratory Study of Class Imbalance for Encrypted Traffic Classification Using CGAN[C]. *ICC 2020 - 2020 IEEE International Conference on Communications*, 2020: 1-7.

[98] Tang Z G, Wang J F, Yuan B G, et al. Markov-GAN: Markov Image Enhancement Method for Malicious Encrypted Traffic Classification[J]. *IET Information Security*, 2022, 16(6): 442-458.

[99] Lin K D, Xu X L, Xiao F. MFFusion: A Multi-Level Features Fusion Model for Malicious Traffic Detection Based on Deep Learning[J]. *Computer Networks*, 2022, 202: 108658.

[100] REZAEI S, LIU X. How to achieve high classification accuracy with just a few labels: A semi-supervised approach using sampled packets [J]. *arXiv preprint arXiv:181209761*, 2018.

[101] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Networks[J]. *Communications of the ACM*, 2020, 63(11): 139-144.

[102] Huang R J, Cui C Y, cHEN F Y, et al. SingGAN: Generative Adversarial Network for High-Fidelity Singing Voice Generation[C]. *The 30th ACM International Conference on Multimedia*, 2022.

[103] Lu Y Z, Chen D, Olaniyi E, et al. Generative Adversarial Networks (GANs) for Image Augmentation in Agriculture: A Systematic Review[J]. *Computers and Electronics in Agriculture*, 2022, 200: 107208.

[104] Lotfollahi M, Jafari Siavoshani M, Shirali Hossein Zade R, et al. Deep Packet: A Novel Approach for Encrypted Traffic Classification Using Deep Learning[J]. *Soft Computing*, 2020, 24(3): 1999-2012.

[105] Wang S S, Chen Z X, Zhang L, et al. TrafficAV: An Effective and Explainable Detection of Mobile Malware Behavior Using Network Traffic[C]. *2016 IEEE/ACM 24th International Symposium on Quality of Service*, 2016: 1-6.

[106] Burnap P, French R, Turner F, et al. Malware Classification Using Self Organising Feature Maps and Machine Activity Data[J]. *Computers & Security*, 2018, 73: 399-410.

[107] Sarkar D, Vinod P, Yerima S Y. Detection of Tor Traffic Using Deep Learning[C]. *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications*, 2020: 1-8.

[108] Zhang W, Meng Y, Liu Y G, et al. HoMonit: Monitoring Smart Home Apps from Encrypted Traffic[C]. *The 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018.

[109] Wang S S, Yan Q B, Chen Z X, et al. Detecting Android Malware Leveraging Text Semantics of Network Flows[J]. *IEEE Transactions on Information Forensics and Security*, 2018, 13(5): 1096-1109.

[110] Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection[J]. *Journal of Machine Learning Research*, 2003, 3: 1157-1182.

[111] Zeng Y, Gu H X, Wei W T, et al. -: A Deep Learning Based Network Encrypted Traffic Classification and Intrusion Detection Framework[J]. *IEEE Access*, 2019, 7: 45182-45190.

[112] Androulidakis G, Papavassiliou S. Improving Network Anomaly Detection *via* Selective Flow-Based Sampling[J]. *IET Communications*, 2008, 2(3): 399.

[113] Duffield N, Lund C. Predicting Resource Usage and Estimation Accuracy in an IP Flow Measurement Collection Infrastructure[C]. *The 2003 ACM SIGCOMM conference on Internet measurement - IMC '03*, 2003.

[114] Su L Y, Yao Y P, Li N, et al. Hierarchical Clustering Based Network Traffic Data Reduction for Improving Suspicious Flow Detection[C]. *2018 17th IEEE International Conference on Trust, Security and Privacy In Computing and Communications/ 12th IEEE International Conference on Big Data Science and Engineering*, 2018: 744-753.

[115] McCulloch W S, Pitts W. A Logical Calculus of the Ideas Immanent in Nervous Activity[J]. *Bulletin of Mathematical Biology*, 1990, 52(1/2): 99-115.

[116] Hinton G E, Salakhutdinov R R. Reducing the Dimensionality of Data with Neural Networks[J]. *Science*, 2006, 313(5786): 504-507.

[117] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification

with Deep Convolutional Neural Networks[J]. *Communications of the ACM*, 2017, 60(6): 84-90.

[118] Wang G Y, Li C Y, Wang W L, et al. Joint Embedding of Words and Labels for Text Classification[EB/OL]. 2018: 1805.04174. https://arxiv.org/abs/1805.04174v1.

[119] Abdel-Hamid O, Mohamed A R, Jiang H, et al. Convolutional Neural Networks for Speech Recognition[J]. *IEEE/ACM Transactions on Audio*, *Speech*, *and Language Processing*, 2014, 22(10): 1533-1545.

[120] VASWANI A, SHAZEER N M, PARMAR N, et al. Attention is All you Need [C]. *NIPS*, 2017.

[121] Miller S, Curran K, Lunney T. Multilayer Perceptron Neural Network for Detection of Encrypted VPN Network Traffic[C]. *2018 International Conference on Cyber Situational Awareness*, *Data Analytics and Assessment*, 2018: 1-8.

[122] Teoh T T, Chiew G, Franco E J, et al. Anomaly Detection in Cyber Security Attacks on Networks Using MLP Deep Learning[C]. *2018 International Conference on Smart Computing and Electronic Enterprise*, 2018: 1-5.

[123] de Lucia M J, Cotton C. Detection of Encrypted Malicious Network Traffic Using Machine Learning[C]. *MILCOM 2019 - 2019 IEEE Military Communications Conference*, 2019: 1-6.

[124] Yang X, Wu J X, Zhao B. Malicious Traffic Classification Based on Indefinite Length Convolutional Neural Network[J]. *Journal of Cyber Security*, 2022, 7(4): 90-99.
(杨璇, 邬江兴, 赵博. 基于不定长卷积神经网络的恶意流量分类算法[J]. *信息安全学报*, 2022, 7(4): 90-99.)

[125] Zheng J, Zeng Z Y, Feng T. GCN-ETA: High-Efficiency Encrypted Malicious Traffic Detection[J]. *Security and Communication Networks*, 2022, 2022: 4274139.

[126] Min E X, Long J, Liu Q, et al. TR-IDS: Anomaly-Based Intrusion Detection through Text-Convolutional Neural Network and Random Forest[J]. *Security and Communication Networks*, 2018, 2018: 4943509.

[127] Li R Y, Song Z H, Xie W, et al. HALNet: A Hybrid Deep Learning Model for Encrypted C&C Malware Traffic Detection[M]. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021: 326-339.

[128] Bakhshi T, Ghita B. Anomaly Detection in Encrypted Internet Traffic Using Hybrid Deep Learning[J]. *Security and Communication Networks*, 2021, 2021: 5363750.

[129] Ren X M, Gu H X, Wei W T. Tree-RNN: Tree Structural Recurrent Neural Network for Network Traffic Classification[J]. *Expert Systems with Applications*, 2021, 167: 114363.

[130] Hou J, Liu F G, Lu H, et al. A Novel Flow-Vector Generation Approach for Malicious Traffic Detection[J]. *Journal of Parallel and Distributed Computing*, 2022, 169: 72-86.

[131] Ferriyan A, Thamrin A H, Takeda K, et al. Encrypted Malicious Traffic Detection Based on Word2Vec[J]. *Electronics*, 2022, 11(5): 679.

[132] Xing J C, Wu C M. Detecting Anomalies in Encrypted Traffic *via* Deep Dictionary Learning[C]. *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops*, 2020: 734-739.

[133] He M S, Wang X J, Zhou J H, et al. Deep-Feature-Based Autoencoder Network for Few-Shot Malicious Traffic Detection[J]. *Security and Communication Networks*, 2021, 2021: 6659022.

[134] Wang J Y, Yan X H, Liu L T, et al. CTTGAN: Traffic Data Synthesizing Scheme Based on Conditional GAN[J]. *Sensors*, 2022, 22(14): 5243.

[135] Sun Y, Gao J, Gu Y J. Malicious Encrypted Traffic Detection Integrating One-Dimensional Inception Structure and ViT[J]. *Computer Engineering*, 2023, 49(1): 154-162.
(孙懿, 高见, 顾益军. 融合一维 Inception 结构与 ViT 的恶意加密流量检测[J]. *计算机工程*, 2023, 49(1): 154-162.)

[136] Wang W, Sheng Y Q, Wang J L, et al. HAST-IDS: Learning Hierarchical Spatial-Temporal Features Using Deep Neural Networks to Improve Intrusion Detection[J]. *IEEE Access*, 2018, 6: 1792-1806.

[137] Wu D, Fang B X, Cui X, et al. BotCatcher: Botnet Detection System Based on Deep Learning[J]. *Journal on Communications*, 2018, 39(8): 18-28.
(吴迪, 方滨兴, 崔翔, 等. BotCatcher: 基于深度学习的僵尸网络检测系统[J]. *通信学报*, 2018, 39(8): 18-28.)

[138] Lee M C, Lin J C, Gran E G. RePAD: Real-Time Proactive Anomaly Detection for Time Series[M]. Advances in Intelligent Systems and Computing. Cham: Springer International Publishing, 2020: 1291-1302.

[139] Prasse P, Machlica L, Pevný T, et al. Malware Detection by Analysing Encrypted Network Traffic with Neural Networks[M]. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017: 73-88.

[140] HASIBI R, SHOKRI M, DEHGHAN M. Augmentation Scheme for Dealing with Imbalanced Network Traffic Classification Using Deep Learning [M]. 2019.

[141] Hwang R H, Peng M C, Huang C W, et al. An Unsupervised Deep Learning Model for Early Network Traffic Anomaly Detection[J]. *IEEE Access*, 2020, 8: 30387-30399.

[142] Rigaki M, Garcia S. Bringing a GAN to a Knife-Fight: Adapting Malware Communication to Avoid Detection[C]. *2018 IEEE Security and Privacy Workshops*, 2018: 70-75.

[143] Li J, Zhou L, Li H X, et al. Network Traffic Feature Camouflage Technology Based on Generating Adversarial Network[J]. *Computer Engineering*, 2019, 45(12): 119-126.
(李杰, 周路, 李华欣, 等. 基于生成对抗网络的网络流量特征伪装技术[J]. *计算机工程*, 2019, 45(12): 119-126.)

[144] Yin C L, Zhu Y F, Liu S L, et al. An Enhancing Framework for Botnet Detection Using Generative Adversarial Networks[C]. *2018 International Conference on Artificial Intelligence and Big Data*, 2018: 228-234.

[145] Samangouei P, Kabkab M, Chellappa R. Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models[J]. *ArXiv e-Prints*, 2018: arXiv: 1805.06605.

[146] Jin G Q, Shen S W, Zhang D M, et al. APE-GAN: Adversarial Perturbation Elimination with GAN[C]. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2019: 3842-3846.

[147] Li Z Y, Wang Y, Wang P, et al. PGAN: A Generative Adversarial

Network Based Anomaly Detection Method for Network Intrusion Detection System[C]. *2021 IEEE 20th International Conference on Trust*, *Security and Privacy in Computing and Communications*, 2021: 734-741.

[148] Xu L, Skoularidou M, Cuesta-Infante A, et al. Modeling Tabular Data Using Conditional GAN[J]. *ArXiv e-Prints*, 2019: arXiv: 1907.00503.

[149] Wang J F, Liu M, Yin X K, et al. Semi-Supervised Malicious Traffic Detection with Improved Wasserstein Generative Adversarial Network with Gradient Penalty[C]. *2022 IEEE 6th Advanced Information Technology*, *Electronic and Automation Control Conference*, 2022: 1916-1922.

[150] Liu Q X, Wang J N, Yin J, et al. Application of Adversarial Machine Learning in Network Intrusion Detection[J]. *Journal on Communications*, 2021, 42(11): 1-12.
(刘奇旭, 王君楠, 尹捷, 等. 对抗机器学习在网络入侵检测领域的应用[J]. *通信学报*, 2021, 42(11): 1-12.)

[151] Goodfellow I J, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples[J]. *ArXiv e-Prints*, 2014: arXiv: 1412.6572.

[152] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks [J]. *ArXiv*, 2017, abs/1706.06083.

[153] Papernot N, McDaniel P, Jha S, et al. The Limitations of Deep Learning in Adversarial Settings[C]. *2016 IEEE European Symposium on Security and Privacy*, 2016: 372-387.

[154] Carlini N, Wagner D. Towards Evaluating the Robustness of Neural Networks[C]. *2017 IEEE Symposium on Security and Privacy*, 2017: 39-57.

[155] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal Adversarial Perturbations[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 86-94.

[156] Chen P Y, Zhang H, Sharma Y, et al. ZOO: Zeroth Order Optimization Based Black-Box Attacks to Deep Neural Networks without Training Substitute Models[C]. *The 10th ACM Workshop on Artificial Intelligence and Security*, 2017.

[157] Novo C, Morla R. Flow-Based Detection and Proxy-Based Evasion of Encrypted Malware C2 Traffic[C]. *The 13th ACM Workshop on Artificial Intelligence and Security*, 2020.

[158] Ibitoye O, Shafiq O, Matrawy A. Analyzing Adversarial Attacks Against Deep Learning for Intrusion Detection in IoT Networks[C]. *2019 IEEE Global Communications Conference*, 2019: 1-6.

[159] Hu Y J, Guo Y B, Ma J, et al. Method to Generate Cyber Deception Traffic Based on Adversarial Sample[J]. *Journal on Communications*, 2020, 41(9): 59-70.
(胡永进, 郭渊博, 马骏, 等. 基于对抗样本的网络欺骗流量生成方法[J]. *通信学报*, 2020, 41(9): 59-70.)

[160] Sadeghzadeh A M, Shiravi S, Jalili R. Adversarial Network Traffic: Towards Evaluating the Robustness of Deep-Learning-Based Network Traffic Classification[J]. *IEEE Transactions on Network and Service Management*, 2021, 18(2): 1962-1976.

[161] Yang K C, Liu J Q, Zhang C, et al. Adversarial Examples Against the Deep Learning Based Network Intrusion Detection Systems[C]. *MILCOM 2018 - 2018 IEEE Military Communications Conference*, 2018: 559-564.

[162] Lin Z L, Shi Y, Xue Z. IDSGAN: Generative Adversarial Networks for Attack Generation Against Intrusion Detection[M]. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2022: 79-91.

[163] Ding Y, Zhu G Q, Chen D J, et al. Adversarial Sample Attack and Defense Method for Encrypted Traffic Data[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2022, 23(10): 18024-18039.

[164] Cheng Q M, Zhou S Y, Shen Y, et al. Packet-Level Adversarial Network Traffic Crafting Using Sequence Generative Adversarial Networks[EB/OL]. 2021: 2103.04794.https://arxiv.org/abs/2103.04794v2.

[165] Fang Z Y, Wang J F, Geng J X, et al. A3CMal: Generating Adversarial Samples to Force Targeted Misclassification by Reinforcement Learning[J]. *Applied Soft Computing*, 2021, 109: 107505.

[166] XIAOZHOU WANG J L. Kaggle microsoft malware challenge winner project. https://github.com/xiaozhouwang/kaggle_Microsoft_Malware/blob/master/Saynotooverfitting.pdf. May. 2015.

[167] Ahmadi M, Ulyanov D, Semenov S, et al. Novel Feature Extraction, Selection and Fusion for Effective Malware Family Classification[C]. *The Sixth ACM Conference on Data and Application Security and Privacy*, 2016.

[168] Kurakin A, Goodfellow I, Bengio S. Adversarial Examples in the Physical World[J]. *ArXiv e-Prints*, 2016: arXiv: 1607.02533.

[169] Moosavi-Dezfooli S M, Fawzi A, Frossard P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2574-2582.

[170] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks [C]. *International conference on machine learning*, 2017.

[171] XIAO C, LI B, ZHU J-Y, et al. Generating adversarial examples with adversarial networks [J]. *arXiv preprint arXiv:180102610*, 2018.

[172] Yu L T, Zhang W N, Wang J, et al. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, 31(1): arXiv: 1609.05473.

[173] Dziugaite G K, Ghahramani Z, Roy D M. A Study of the Effect of JPG Compression on Adversarial Images[EB/OL]. 2016: 1608.00853.https://arxiv.org/abs/1608.00853v1.

[174] HAN D, WANG Z, ZHONG Y, et al. Practical Traffic-space Adversarial Attacks on Learning-based NIDSs [J]. *ArXiv*, 2020, abs/2005.07519.

[175] [175], Matyasko A. Towards Deep Neural Networks Robust to Adversarial Examples[D]. Nanyang Technological University, 2020. DOI: 10.32657/10356/143316.

[176] Papernot N, McDaniel P, Wu X, et al. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks[C]. *2016 IEEE Symposium on Security and Privacy*, 2016: 582-597.

[177] Wang B, Guo Y K, Qian Y G, et al. Defense of Traffic Classifiers Based on Convolutional Networks Against Adversarial Examples[J]. *Journal of Cyber Security*, 2022, 7(1): 145-156.
(王滨, 郭艳凯, 钱亚冠, 等. 针对卷积神经网络流量分类器的对抗样本攻击防御[J]. *信息安全学报*, 2022, 7(1): 145-156.)

樊祖薇 于 2021 年在北京外国语大学计算机科学与技术专业获得学士学位。现在中国科学院信息工程研究所网络与信息安全专业攻读硕士学位。研究领域为网络与信息安全。研究兴趣包括: 网络安全。Email: fanzuwei@iie.ac.cn

赵泓策 于 2020 年在燕山大学电子信息工程专业获得学士学位。现在中国科学院信息工程研究所电子信息专业攻读硕士学位。研究领域为移动通信网络与安全。Email: zhaohongce@iie.ac.cn

张顺亮 于 2004 年在浙江大学计算机专业获得博士学位。现中国科学院信息工程研究所高级工程师。研究领域为移动通信网络与安全。Email: zhangshunliang@iie.ac. cn