

加密流量识别技术综述

崔苏苏^{1,2}, 刘宝旭^{1,2}, 董 聪^{1,2}, 姜 波^{1,2}, 卢志刚^{1,2}, 张 辰¹

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 100049

摘要 流量识别技术作为网络防护和管理的关键手段,可帮助网络管理员及时阻止恶意行为的传播和进行网络资源的优化。当前,随着数据安全意识的增强,网络服务和应用普遍采用加密协议来保障通信内容的安全。这些加密协议旨在保护通信内容,并增加额外的数据机密性。虽然该方法可以有效增强数据的保密性,但同时也给网络管理带来新的挑战。通信内容在经过加密算法的变化后,载荷不再具有明显的字符特征,因此传统的流量识别方法无法对加密流量进行有效识别。为此,研究人员针对加密流量识别技术进行了大量研究。本文首先对流量识别的基础概念进行了介绍,包括识别对象和主流加密协议的调研;然后根据不同的场景需求,基于协议栈由底向上的角度对当前需求较为迫切的加密任务进行了梳理;其次通过对当前现有的加密流量识别方法进行归纳比较,从基于深度包检测、基于传统机器学习、基于深度学习三个方面对加密流量识别方法进行综述;基于多维信息融合以及深度学习强大学习能力的优势,多模式混合方法有望在未来成为加密流量识别方法的突破性技术;最后,本文基于当前研究进展对加密流量识别技术的未来发展方向进行研究展望,旨在帮助学术研究人员和行业专业人士在快速发展的数字环境中认识加密流量识别的复杂性和挑战。

关键词 加密流量识别; 加密协议; 统计特征; 机器学习; 深度学习

中图法分类号 TP393 DOI 号 10.19363/J.cnki.cn10-1380/tn.2026.05.15

Overview of Encrypted Traffic Identification Technology

CUI Susu^{1,2}, LIU Baoxu^{1,2}, DONG Cong^{1,2}, JIANG Bo^{1,2}, LU Zhigang^{1,2}, ZHANG Chen¹

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 Traffic identification technologies have evolved as a key component for modern network protection and management systems. These technologies enable network administrators to promptly detect and halt the spread of malicious activities, as well as carry out critical optimizations of network resources. In recent years, the growing emphasis on data security has led to the widespread adoption of encryption protocols across various network services and applications. These encryption protocols are engineered to secure the content of communications and add an extra layer of data confidentiality. Although this method can effectively enhance the confidentiality of data, it also brings new challenges to network management. After the change of encryption algorithm, the payload no longer has obvious character features, so the traditional traffic identification methods cannot effectively identify the encrypted traffic. For this reason, researchers have conducted a lot of research on encrypted traffic identification techniques. In this paper, we firstly introduce the basic concept of traffic identification, including the research of identification objects and mainstream encryption protocols; then we sort out the current urgent encryption tasks based on the bottom-up perspective of the protocol stack according to different scenarios. Secondly, we summarize and compare the current existing encrypted traffic identification methods, and compare the encrypted traffic identification methods based on deep packet inspection, traditional machine learning, and deep learning. Based on the advantages of multi-dimensional information fusion and the powerful learning ability of deep learning, the multi-modal hybrid method is expected to be a breakthrough technology for encrypted traffic recognition in the future. Conclusively, in light of current research advancements, this paper provides a forward-looking perspective on the future development vectors for encrypted traffic identification technologies. It aims to serve as a cornerstone document, assisting both academic researchers and industry professionals in navigating the complexities and challenges of encrypted traffic identification in a rapidly evolving digital landscape.

Key words encrypted traffic identification; encryption protocols; statistical features; machine learning; deep learning

通讯作者: 张辰, 硕士, 工程师, Email: zchen@jie.ac.cn。

本课题得到国家重点研发计划(No. 2019QY1300), CCF-绿盟科技“鲲鹏”科研基金(No. 2020010), 中国科学院青年创新促进会(No. 2021156), 中国科学院战略性先导科技专项 C 类(No. XDC02040100), 国家自然科学基金青年科学基金项目(No. 61802404)的资助。这项工作也得到了中国科学院网络评估技术重点实验室和北京市网络安全与保护技术重点实验室的部分支持。

收稿日期: 2021-03-03; 修改日期: 2021-04-20; 定稿日期: 2023-08-29

1 引言

随着软件定义网络(Software Defined Network, SDN)^[1]、物联网(Internet of Things, IOT)^[2]、第五代移动通信技术(5th Generation Mobile Networks, 5G)^[3]等网络技术的发展, 智能设备的普及以及各种新型应用服务的兴起, 网络流量在近年来呈现出几何增长的趋势。根据思科第 12 届全球网络流量 2017~2022 预测报告^[4], 预计在 2022 年, 全球范围内的网络流量总量将高达 4.8ZB, 是 2017 年的 3 倍以上。然而, 在网络服务覆盖社会生活各方面的同时, 分布式拒绝服务攻击(Distributed Denial of Service, DDoS)、高级可持续攻击(Advanced Persistent Threat, APT)、僵尸网络(Botnet)等网络攻击事件层见叠出。这不但给网络互联设备和网络用户带来了巨大的安全隐患, 也给网络管理人员带来了前所未有的挑战。

网络安全威胁已成为学术界和工业界各领域重点关注的事件之一, 为了有效防护网络攻击和保护数据隐私, 流量加密成为广泛使用的一种技术手段, 目前加密流量已占据 IP 流量的 75%~90%^[5]。网络应用使用各种加密协议、隧道协议、匿名网络等保护数据传输, 例如, 移动应用广泛使用安全套接层协议(Secure Sockets Layer, SSL)和传输层安全协议(Transport Layer Security, TLS)^[6]来对应用数据进行加密, 以避免对有效载荷的检查; 主机间通信通过 Internet 协议安全(Internet Protocol Security, IPsec)^[7]加密数据流; 安全外壳协议(Secure Shell, SSH)为两台计算机之间建立安全通道。在使用加密协议保护数据安全的同时, 为绕过防火墙和安全防护软件的检测, 大量恶意网络攻击同样开始使用流量加密技术隐藏通信内容^[8], 流量加密一时间成为普通网络应用和网络攻击行为都采用的技术手段。

因此, 为保护设备和用户数据安全, 安全人员需要加密流量识别技术快速发现加密网络中的异常行为, 及时阻止恶意行为的进一步传播。同时, 加密流量识别技术还有助于网络管理人员进行流量分布统计和用户行为分析, 进而提高网络资源利用率以及加强服务质量。

传统的流量识别技术已不再适用于加密流量的识别。其中, 端口混淆技术使得基于端口的流量识别不再有效; 加密协议将应用数据加密为随机的密文, 为基于深度包检测的识别增大了难度, 同时加密协议和加密应用的快速更新发展, 基于深度包检测的流量识别需要投入大量人力物力不断更新识别指纹和规则。由于缺乏有效的加密流量识别技术, 网络攻

击发现和 network 管理均面临巨大的挑战。

近年来的研究发现, 加密技术对流量的统计特征影响较小, 机器学习尤其深度学习在加密流量识别中取得显著成效^[8-11]。因此, 本文聚焦于近年来基于传统机器学习和基于深度学习的加密流量识别研究成果, 从加密流量识别任务和加密流量识别方法两个维度进行归纳综述。具体而言, 加密流量识别任务包括加密流量筛选、加密协议识别、加密服务识别、加密应用识别、应用内服务识别、加密视频 QoE 估计、加密 Web 识别和加密恶意流量识别。加密流量识别方法包括基于深度包检测的识别方法、基于传统机器学习的识别方法和基于深度学习的识别方法。本文首先梳理归纳了各加密流量识别任务的研究方法和进展, 然后将加密流量识别方法进行总结分析, 比较不同方法之间的识别性能。基于当前研究工作提出展望, 加密流量细粒度分析、私有协议流量分析、加密恶意流量分析和加密流量不平衡处理为加密流量识别技术中的重点研究方向。

本文的组织结构如下: 第 2 节简要介绍加密流量分析工作中常见的识别对象; 第 3 节介绍常用的加密协议, 第 4 节描述加密流量识别任务; 第 5 节综述加密流量识别方法; 第 6 节对加密流量的未来研究方向进行讨论展望。第 7 节是结束语。

2 加密流量识别对象

进行加密流量识别首先需要确认加密流量的识别对象, 即如何将连续的加密流量离散化。不同的识别对象直接影响统计特征的分析与生成, 在现有的加密流量研究工作中, 常见的流量识别对象主要包括包、TCP 连接、流、服务、主机^[9]。其中, 流是加密流量识别技术中最常用的识别对象。下面对常见的流量识别对象逐一描述。

包: 通常定义为连续流量中的单一数据包为一个识别对象。

TCP 连接: 基于观察一些 TCP 标志(如 SYN、FIN、RST)或 TCP 状态的启发式方法, 用来识别每个连接的开始和结束。

流: 典型的流定义为使用同一五元组(源 IP, 目的 IP、源端口、目的端口、传输层协议)生成的流量; 或将流定义为源 IP 和目的 IP 可以互换的双向流; 或使用流超时(60 秒或 90 秒的空闲时间来划分流的结束)或周期性重置(例如, 超时 5 分钟边界上的所有流)定义流。

服务: 通常定义为一个 IP 端口对产生的所有流量。

主机: 通常定义为一个主机对产生的所有流量。

3 加密协议

不同加密协议的数据包结构、通信模式和适用场景均有所不同, 其握手连接过程中的明文信息也各有不同, 这直接影响识别工作中的特征提取, 因此, 本节对常见的加密协议 IPsec、SSL/TLS、SSH、VPN、Tor 进行概述。

3.1 IPsec

IPsec^[7]是一组用于解决网络层安全性问题的网络安全协议, 主要通过加密与认证等方式, 为 IP 数

据报提供高质量的、可互操作的、基于密码学的安全性。IPsec 通过认证头(AH)和封装安全载荷(ESP)这两个安全协议来实现。ESP 对分组提供了源可靠性、完整性和保密性的支持。其在传输模式和隧道模式下的封装格式如图 1 所示。其中, 在传输模式下, AH 或 ESP 被插入到 IP 头之后以及所有传输层协议之前, 或者被插入到所有其他 IPsec 协议之前。在隧道模式下, AH 或 ESP 插在原始 IP 头之前, 另外生成一个新 IP 头放到 AH 或 ESP 之前。由于 IPsec 能够在 IP 层加密或认证流量, 可在远程登录、客户/服务器、邮件、文件传输、Web 浏览等应用中提供加密支持。

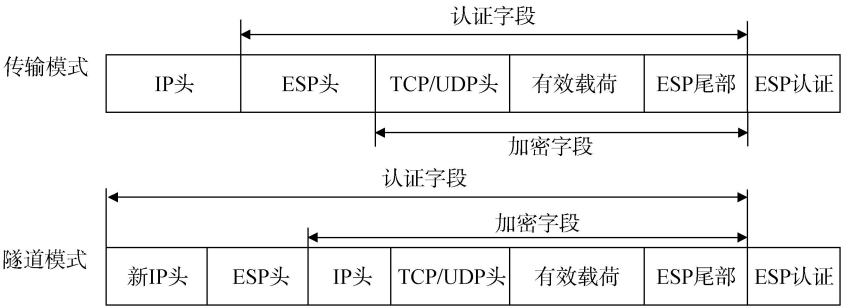


图 1 ESP 封装及头部格式
Figure 1 Wrapped ESP header

3.2 SSL/TLS

SSL/TLS^[6]是建立在 TCP/IP 层与各种应用层协议之间的安全协议, 用于保证数据传输的安全性、完整性和认证性。由于 SSL/TLS 协议可与高层的应用层协议(HTTP、FTP、Telnet 等)无耦合, 应用层协议传输的数据在通过 SSL/TLS 协议时都会被加密, 从而保证了通信的私密性。

SSL/TLS 协议包含两个协议子层, 协议框架如图 2 所示。高层的握手协议层用于在实际的数据传输开始之前, 通信双方进行身份认证、协商加密算法、交换加密密钥等, 具体包括握手协议、改变密码规则协议和报警协议。底层的记录协议层为高层协议提供数据封装、压缩、加密和认证基本功能。在双方进行安全的数据传输之前, 首先会使用 SSL 的握手协议建立连接, 协商数据传输过程中的安全参数和身份验证, 此协商过程进行明文传输, 因此加密套件、证书、签名算法、服务器名称等均可以在握手阶段获得。

3.3 SSH

SSH^[12]是一种用于从一台计算机安全登录到另一台计算机的方法。它通过使用客户端-服务器体系结构, 将 SSH 客户端应用程序与 SSH 服务器连接, 在不安全的网络上提供了安全通道。它提供了多种

用于强身份验证的选项, 并通过强加密来保护通信安全性和完整性。它已成为非受保护的登录协议(例如 Telnet、Rlogin)和不安全的文件传输方法(例如 FTP)的安全替代方法。

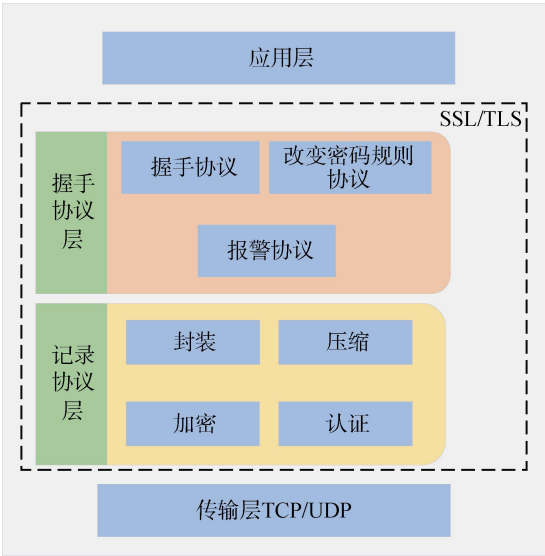


图 2 SSL/TLS 协议框架
Figure 2 SSL/TLS protocol framework

SSH 协议由三个子协议组成: 传输层协议、用户认证协议和连接协议, 协议框架如图 3 所示。其中,

传输层协议提供版本协商、密钥交互和信息完整性, 具有完全前向安全性; 用户认证协议用来执行客户端的用户验证请求; 连接协议将加密通信隧道复用为一些逻辑信道。使用 SSH 进行安全通信过程中, 应用层数据为已加密随机化的密文, 底层协议头与协商阶段的参数如服务器与客户端的公钥、支持的加密方法、认证方法等均为明文, 可用于特定场景下的特征提取。

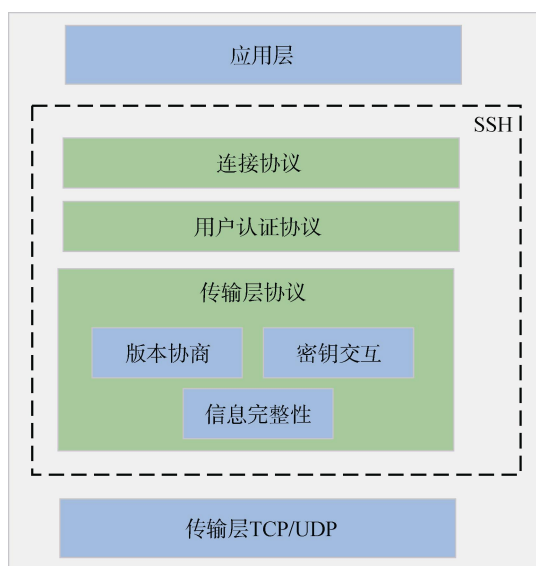


图 3 SSH 协议框架

Figure 3 SSH protocol framework

3.4 VPN

虚拟网络技术(Virtual Private Network, VPN)^[13]作为一种专用网络技术、通常指的是在公用数据网络中建立私有虚拟数据网络, 如图 4 所示。它利用隧道协议来实现认证、加密等功能, 对网络数据传输安全性的增强极其有利。常用的 VPN 技术有 IPSec VPN 和 SSL VPN, 二者均可提供身份认证、数据加密来保证通信安全, 其中 IPSec 保护整个网络层的安全, SSL 保护应用层数据安全。IPSec 存在网络部署及配置烦琐, 无法在应用层做精细化控制等问题, SSL VPN 采用 B/S 架构, 可通过浏览器直接访问内网资源, 可对访问权限进行细粒度控制。因此, IPSec 更适用于网对网的 VPN 连接, 而 SSL VPN 适用于移动用户的远程接入, 二者长期共存, 优势互补。

3.5 Tor

匿名网络在信息传输过程中无法建立消息与网络参与者的对应关系, 能够提供传输过程中的隐蔽性。对于典型的匿名网络 Tor^[14], 它保证用户使用网络应用时, 第三方无法通过主动或被动的方式破坏其隐蔽性。图 5 为 Tor 网络的架构示意图。在进行

正式通信之前, Tor 客户端会向目录服务器请求网络的节点信息, 然后依据某种选路策略(例如节点带宽、运行时间等)选择三个节点作为信息传输过程中的参与节点。当洋葱代理发出服务请求后, 会根据目录服务器返回的节点信息选择三个节点作为链路构建的参与节点。为了确保发送者的匿名性, 任一节点都无法知道在链中自己的前一个节点是发送者还是链上的另一节点; 同理, 任一节点也无法知道在链中自己的下一节点是目的地还是链上另一节点。Tor 网络上节点之间的流量均使用加密保证数据的安全性。

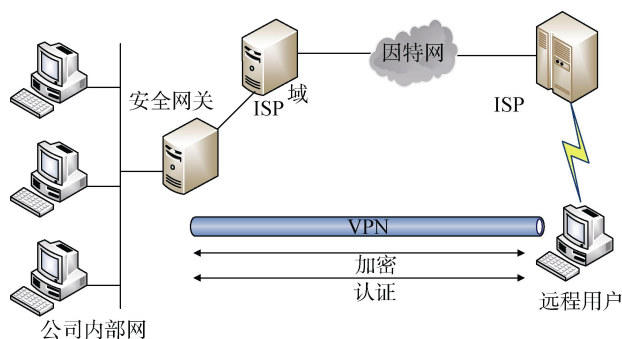


图 4 以 VPN 技术为基础建立的虚拟网络

Figure 4 Virtual network based on VPN technology

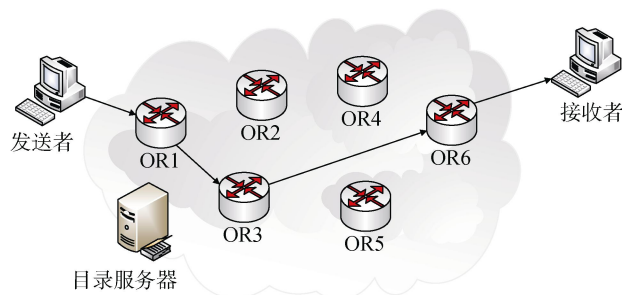


图 5 Tor 网络架构示意图

Figure 5 Tor network architecture schematic

4 加密流量识别任务

加密流量识别技术是构成网络服务质量(Quality of Service, QoS)和网络入侵检测系统(Network Intrusion Detection System, NIDS)不可或缺的核心技术, 根据具体需求和任务的不同, 可对加密流量识别对象按照协议栈由底向上分为加密流量筛选、加密协议识别、加密服务识别、加密应用识别、应用内服务识别、加密视频 QoE 估计、加密 Web 识别、加密恶意思识别, 各识别任务的分类层次关系如图 6 所示。具体而言, 加密流量筛选是将加密流量从网络流量中过滤出来; 加密协议识别是识别流量的加密

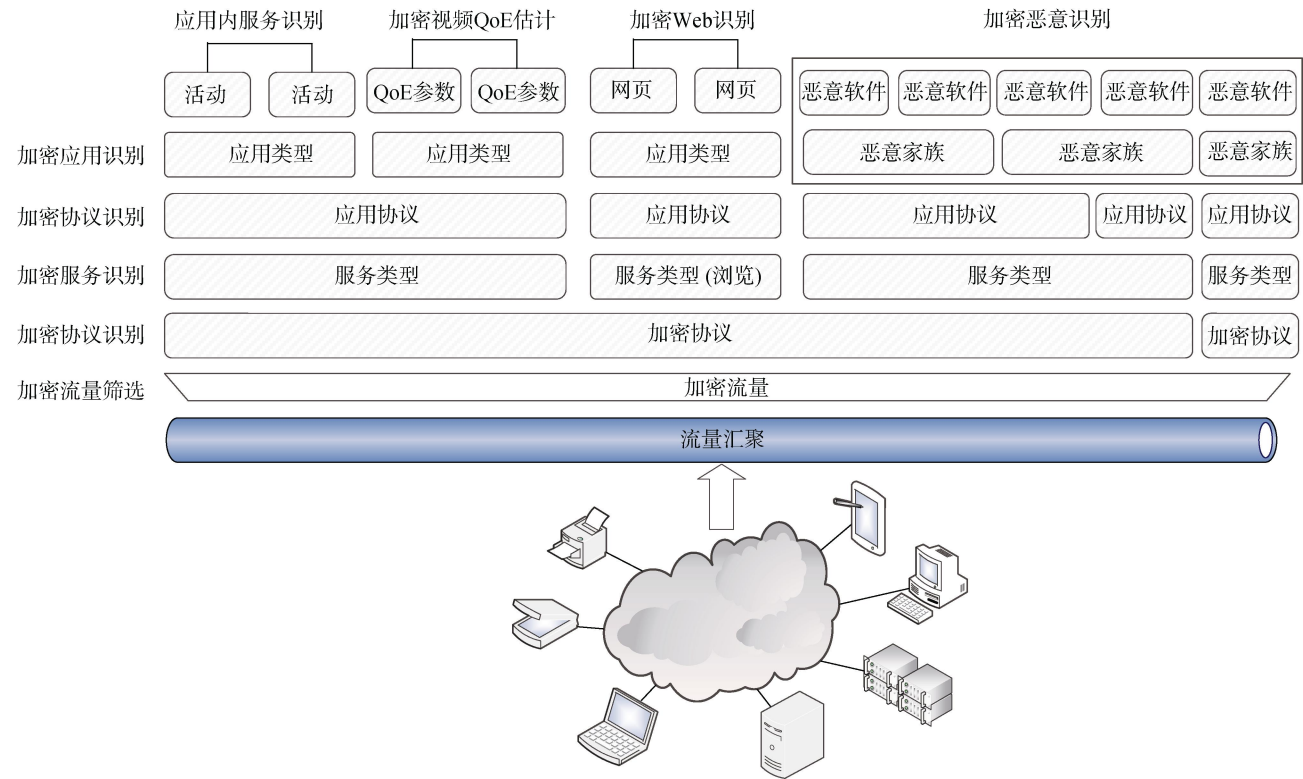


图 6 加密流量识别任务的分类层次关系

Figure 6 Hierarchy of encrypted traffic identification task

协议(如 SSH、SSL/TLS 等)或上层应用协议(如 HTTPS、FTPS 等); 加密服务识别是识别加密流量所属的服务类型(如邮件传输、网页浏览等); 应用内服务识别是通过加密移动流量识别特定 APP 内的具体活动(如 Wechat APP 下的文本、位置分享等); 加密视频 QoE 估计是识别加密视频流量的 QoE 参数(如清晰度、停顿等); 加密 Web 识别通过建立网站或网页指纹识别用户是否在访问某些站点; 加密恶意识别是识别加密流量是否是恶意流量以及识别恶意家族(如 DDos、U2R 等)或具体的恶意软件(如 Zeus、Cridex 等)。

4.1 加密流量筛选

通常情况下, 经由各种网络设备汇聚的原始流量同时包含非加密流量和加密流量, 进行加密流量识别与分析的首要任务是加密流量筛选。加密流量筛选是将加密流量从网络背景流量中过滤出来, 实现加密流量与非加密流量的区分。由于加密算法使得应用层数据随机化, 而随机数序列组成的数据包会产生较高的熵级别, 因此, 文献[15]使用流的第一个数据包中的有效载荷进行熵估计^[16]实现加密流量筛选, 根据真实网络流量评估, 加密流量检测率高达 94%, 未加密流量的检测率高达 99.9%。然而, 加密将数据转换为伪随机格式从而增加了熵值, 而压缩为了减少信息量, 也增加了熵值, 该方法无法对

加密流量和压缩流量进行区分。基于此, 文献[17]提出了一种基于熵特征的加密流量筛选方法, 基于数据包有效载荷、编码有效载荷以及二进制有效载荷计算熵作为特征, 使用机器学习技术学习特征, 可对加密、压缩以及编码进行区分。同时, 文献[18]提出了一种基于 n-gram 熵和累积和检验^[19]的加密流量识别方法。该方法通过对不同长度和频率的子序列流量数据进行多熵表征, 通过累积和检验区分压缩流量和加密流量。结果表示基于熵和累积和建立特征可更好地区分加密流量和压缩流量。图 7 显示了本任务的研究工作进展。

4.2 加密协议识别

随着加密技术的不断普及, 各种加密协议和封装协议被不断提出以保证数据的安全性, 如 IPSec、SSH、SSL/TLS、VPN、Tor 等。为适应不同的网络服务, 同一种加密协议上往往承载着不同的应用协议, 如 HTTPS、FTPS 等均是基于 TLS 加密的应用协议。加密协议识别是根据加密流量的协议类型进行分类, 该协议既包括识别特定的加密或封装协议也包括识别加密流量的应用协议。

加密协议识别普遍采用签名或规则的方法实现, 而加密应用协议识别一般采用机器学习的方法学习加密流的统计特征以实现识别。例如, 文献[20]首先

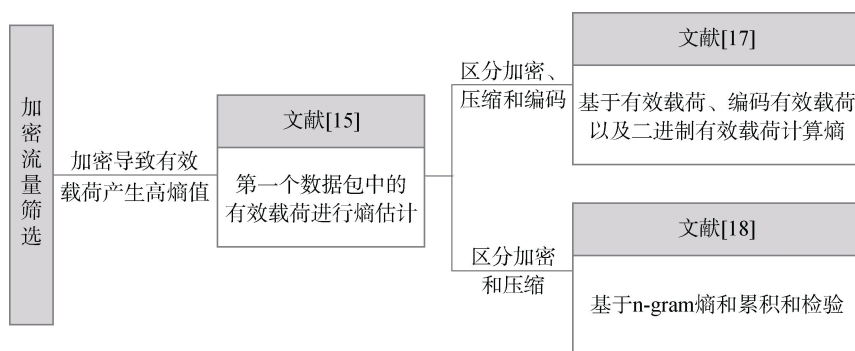


图 7 加密流量筛选研究工作进展

Figure 7 Researches of encrypted traffic filtering

通过签名匹配可识别超过 99% 的 SSL/TLS 流量, 然后通过统计分析进一步识别 HTTPS、TOR、UPDATE、OSCAR 协议, 可实现 F_1 达 94.52% 的协议识别率。类似地, 文献[21]发现 PPTP 和 IPSec 流量的数据包到达时间(IAT)始终大于正常流量的 IAT, 因此, 他们首先通过设置 IAT 阈值来检测 PPTP 和 IPSec 流量, 然后通过统计特征识别 PPTP、IPSec 下的应用协议 HTTP、FTP、SSH、SMTP, 识别率接近 100%。文献[22]结合深度包检测和机器学习联合模型实现了 Skype 协议流量的识别, 达到了误报率几乎为零的效果。文献[23-25]使用大量机器学习算法进行了 SSH 和 Skype 流量识别, 并从识别精度、鲁棒性、复杂性等方面评估了各算法在加密流量识别任务上的优劣。为实现在线识别, 文献[26-28]通过对加密流的前 N 个数据包进行统计分析, 来实现加密协议与应用协议的识别。具体而言, 文献[26]通过分析 SSH 连接的第一个 IP 数据包的统计特征可实现 99.2% 的 SSH 检测精度, 同时对 SSH 承载的应用协议(SCP、SFTP、HTTP)可实现 99.8% 的分类精度。文献[27]通过对加密流的前 100 个数据包进行统计分析, 可对特定的 BitTorrent、Skype 协议进行在线识别, 分别实现了 97.4% 和 94% 的识别率。文献[28]通过分析每条流前 8 个数据包的统计特征来对 HTTP、FTP、SSH、DNS、P2P、SMTP 协议进行在线识别, 识别精度可达 95% 以上。此外, 文献[29]首次将神经网络应用于加密流量协议识别中, 使用神经网络实现了包括加密协议在内的 25 种协议识别。文献[30-32]相继使用不同的神经网络学习加密流的深层序列特征, 神经网络优于传统机器学习的识别性能。图 8 显示了本任务的研究工作进展。

4.3 加密服务识别

随着互联网的开放性, 各种新的网络服务层出不穷, 不同的服务对网络资源有不同的需求。例如,

视频浏览类服务下行流量明显大于上行流量, 一些网络提供商尝试识别占用网络带宽较大的 P2P 流量, 并在繁忙时段限制其带宽^[33]。网络运营商为了合理分配网络资源, 需要对加密流量进行服务识别, 加密流量服务识别是根据加密流量所属的服务类型(即时通信、邮件、流媒体、文件传输等)进行分类。

文献[34]通过分析流量加密前后的统计特征变化, 提出加密后的流统计特征可由加密前的统计特征转换, 因此通过转换正常流量的流特征综合训练数据, 使用机器学习算法识别 IPSec、PPTP 加密的服务类型, 实验结果可达到 97.2% 的识别准确率。文献[35]提出使用与时间相关的流特征表征流量, 然后使用机器学习学习统计特征, 可实现常规加密和 VPN 封装的 12 种服务类型的识别, 达到 80% 以上的准确率。同时, 他们公开了本次方法的数据集 ISCX VPN-nonVPN, 在此数据集上, 文献[36]使用六种机器学习模型学习与时间相关的流量特征进行服务识别; 更多地, 大量研究人员尝试通过神经网络的方法学习加密流量深层特征, 如文献[32, 37-40]使用神经网络(CNN^[41]、SAE^[42]、CapsNet^[43])学习加密流或加密包的字节空间特征^[44], 文献[45-47]使用神经网络(LSTM^[48]、Attention^[49]、GRU^[50])学习加密流的序列特征, 文献[44, 51]使用多模型融合同时学习加密流的空间特征和序列特征。图 9 显示了本任务的研究工作进展。

4.4 加密应用识别

边缘路由器的流量整形和调度技术旨在根据 QoS 策略模型管理带宽资源, 在应用 QoS 策略之前, 正确分类应用程序流至关重要。例如, 网络管理员希望识别和节流或阻止来自不同应用的流量, 以管理带宽预算, 并确保满足关键业务应用的服务质量目标。加密流量应用识别是识别加密流量由具体哪种应用程序(如 Hangout、YouTube、QQ)产生。

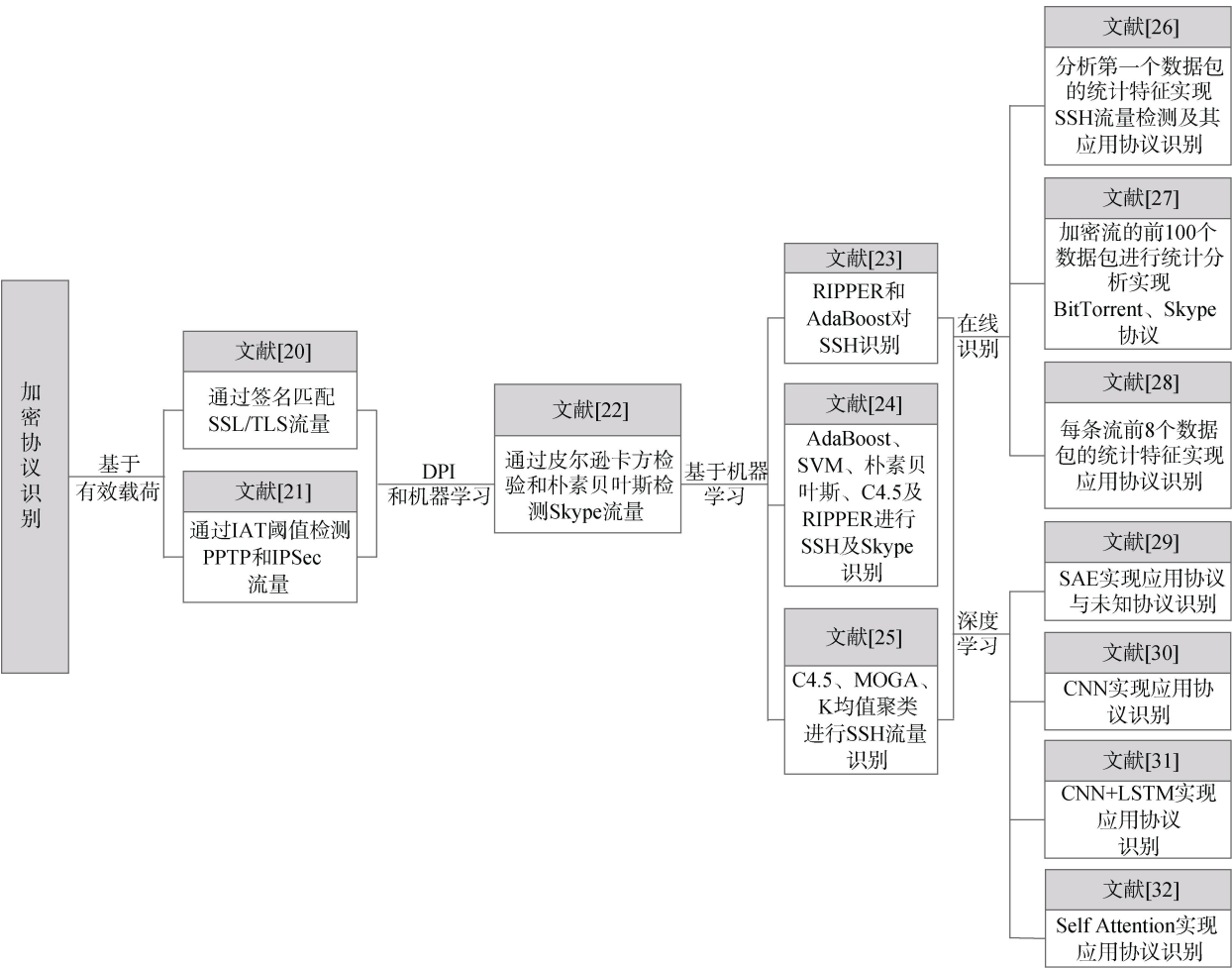


图 8 加密协议识别研究工作进展

Figure 8 Researches of encrypted protocol identification

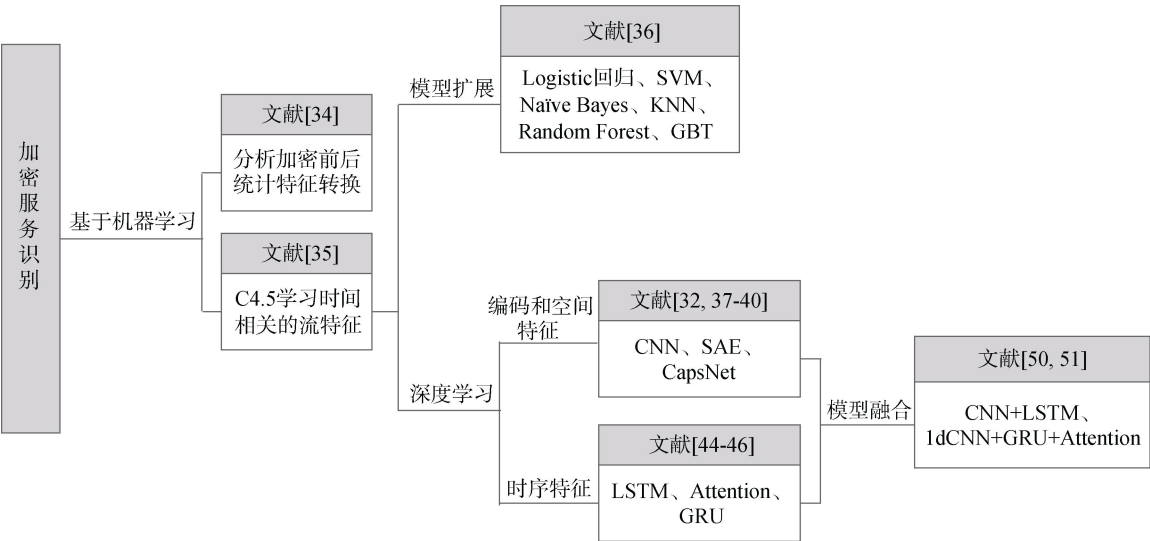


图 9 加密服务识别研究工作进展

Figure 9 Researches of encrypted service identification

本任务的实现方法与加密服务识别类似，一方面，通过提取流的统计特征，使用机器学习算法建立分类模型，实现应用识别。如文献[52]通过提取

SSH 流统计特征进行应用程序识别，达到了 92% 的检测率和 0.8% 的误报率。另一方面，使用神经网络学习加密流的空间特征和时序特征进行应用程序分

类, 文献[53]使用 CNN 和 GRU 学习流数据的空间特征和序列特征, 可实现 Android 和 iOS 上的多种应用程序识别。

与服务识别方法不同的是, 加密应用识别还提出了一系列通过 SSL/TLS 字段信息构建应用程序指纹的方法。文献[54]提出不同应用程序在数据加密传输过程中的 SSL/TLS 消息类型不同, X (原推特) 在服务器身份验证过程后, 几乎 70% 的会话不会更改加密策略 (没有 Change Cipher Spec 消息), 相反, PayPal 则会在建立连接后更改加密策略。因此, 他们通过提取 SSL/TLS 消息类型基于一阶马尔可夫链^[55]

构建应用程序指纹。然而基于一阶马尔可夫链很难获取独特的指纹, 文献[56]提出在 SSL/TLS 消息类型的基础上增加证书包长度, 基于二阶马尔可夫链建立指纹, 该方法的识别准确率比文献[54]的方法提高了 30%。随后, 文献[57]又在上一方法的基础上增加第一个应用程序数据包的大小作为指纹参数。类似的工作还有文献[58, 59]。除基于马尔可夫模型之外, 文献[60]使用序列神经网络 RNN Encoder-Decoder^[50]挖掘流的包长度序列, 在 18 个加密应用程序分类中实现了 99.14% 的检测率 (TPR) 和 0.05% 的误报率 (FPR)。图 10 显示了本任务的研究工作进展。

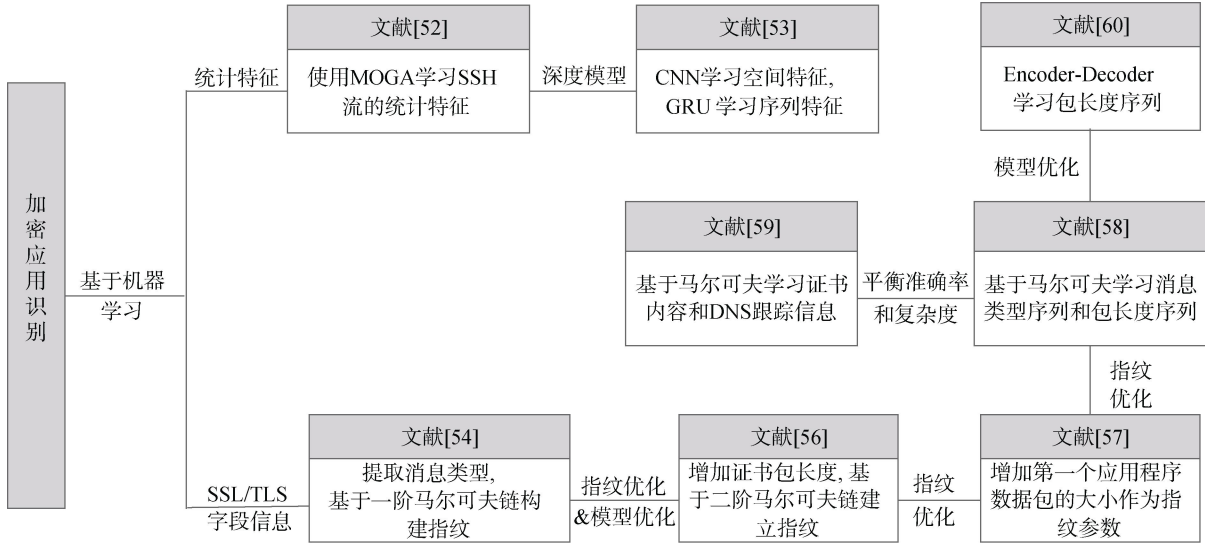


图 10 加密应用识别研究工作进展

Figure 10 Researches of encrypted application identification

4.5 应用内服务识别

随着移动设备的广泛使用, 即时消息应用提供了新的沟通方式, 如文本短信、视频、语音、位置分享、时刻发布等, 应用内服务分析旨在了解应用内用户的行为, 提升用户体验和提供网络资源分配。在移动消息传递应用程序中, 会话通常在用户打开应用程序时开始, 并持续到该用户关闭它为止。每个会话都包含多个对话框。大多数对话框是单一类型的用途, 例如文本, 位置共享, 语音或流视频等, 而剩余其他对话框则是混合用途。

与其他任务不同的是, 本任务需要将连续的会话切分为单一服务的对话框, 然后对不同的单一服务对话框进行建模分类。文献[61]结合阈值和层次聚类的方法, 将连续的会话切分为单一服务的对话框。然后从对话框的数据包长度和时间延迟两方面来提取流特征。文献[62]提出了一种迭代分类器来实现在线识别, 该分析器首先使用时间窗口对传入的网络流进行离散化, 每个时间窗口都由最大化信

息活动相似度和最小化不同活动相似度(MIMD)算法选择的 6 个可迭代特征向量表示, 然后使用基于聚类的流量分段算法时间窗口分组为多个簇, 最后将聚类中心特征向量输入到 Random Forest^[63]分类器中, 以实现应用内服务分类。该方法通过社交媒体应用程序 (例如 Facebook) 和消息传递应用程序 (例如 Wechat 和 Whatsapp) 验证, 实现了高精度和高效率的分类, 数据处理速度是文献[61]的 14.3 倍。文献[64]认为分段算法往往会误分出一些混合用途的子序列, 容易对分类器的分类效果造成干扰, 为提高分类器对混合活动序列的抵抗性和鲁棒性, 他们提出了一种多标签多视图的加密流量应用内服务分类。他们将单一活动的子序列和混合活动的子序列作为与多个标签 (即活动类型) 相关联的子序列进行唯一化, 每个标签都由一个概率表示。在子序列中从不同的视图即数据包长度序列视图和数据包时延序列视图提取流量特征, 利用这些视图之间的共存性表征流量的模式。最后, 通过多标签逻辑

分类器同时实现单一活动序列分类和混合活动序列检测。文献[65]提出了一个名为 ActiveTracker 的识别加密流量应用内活动轨迹的嗅探器, 它首先设计了一种基于滑动窗口的方法将连续加密流量分段, 该方法依据数据包长度和数据包间隔的统计特征来寻找不同活动之间的间隔点。然后将流量

段规范化, 由时空流量矩阵和流量频谱矢量表示。最后, 基于 CNN 建立分类器实现 APP 内活动识别。该方法将应用内服务分类扩展到其他类型的 APP(Paypal, Hangouts, Line, Messenger, QQ, Skype 和 Wechat), 而不是局限于社交媒体类 APP。图 11 显示了本任务的研究工作进展。

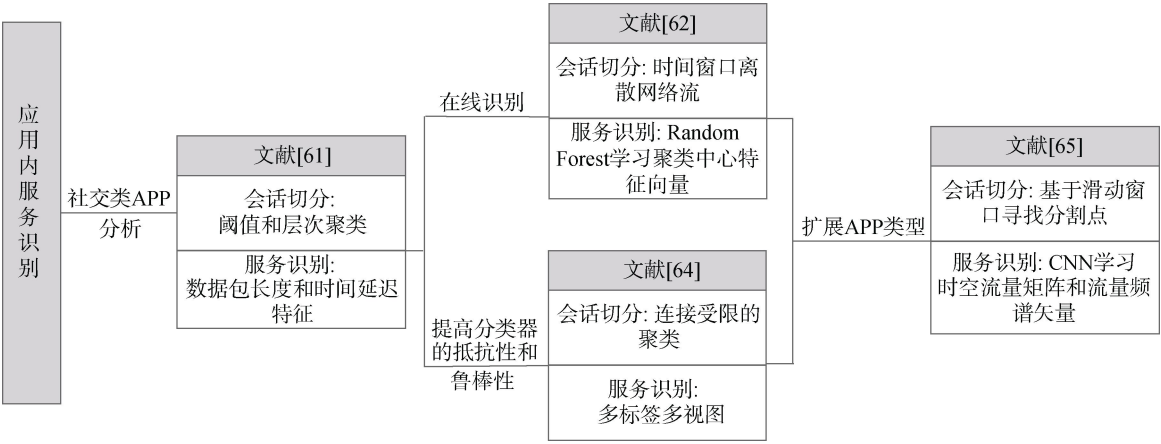


图 11 应用内服务识别研究工作进展
Figure 11 Researches of In-App service identification

4.6 加密视频 QoE 估计

全球移动网络流量的稳定增长, 很大程度上是由于视频流服务与应用(如 YouTube、Netflix、Facebook)的激增, 移动视频流量已占移动数据总流量的 75%^[66], 网络服务商需要通过加密视频评估参数来进行 QoE 估计和潜在质量问题分析。加密视频 QoE 估计是根据流量特征识别视频的 QoE 评估参数如码率、清晰度等, 或者根据流量特征直接对 QoE 等级进行建模估计。

HTTPS 加密传输使得传统的深度包检测方法无法获取码率、清晰度等 QoE 评估参数。目前, QoE 评估参数大多采用机器学习的方法建模评估。该方法通过使用机器学习算法将网络可观察到的指标(例如数据包延迟、丢失和吞吐量)与视频 QoE 指标相关联, 从而推断 QoE。例如, 文献[67]首次提出对加密的 Youtube 流的质量表示进行分类。他们通过机器学习分析 Youtube 加密流的峰值差异, 对视频质量清晰度(360P、480P、720P)进行分类, 平均准确度可达到 96.13%。文献[68]从加密视频流量中提取包长统计、5s 间隔的传输数据大小、包数统计、互达时间统计、吞吐量统计和 TCP 标志计数方面的 33 个特征, 通过机器学习算法建立分类模型评估加密视频块的清晰度和停顿, 并将加密视频的 QoE 评估为高、中、低三级, 可实现 84% 的 QoE 分类精度。文献[69]提出视频块的大小和到达时间是衡量 QoE 的重要指标, 并

提出 7 组统计特征进行流量表征, 用于从加密流量中检测影响 QoE 的停顿、平均质量和质量切换, 准确率达 76%~91.8%。文献[70]提出上游数据包的往返时间(RTT)与 QoE 指标存在高相关性, 为此, 他们仅使用 RTT 作为视频流特征在线细粒度检测 QoE 指标包括启动延迟、重新缓冲(停顿)、视频分辨率, Bilibili 上的检测率可高达 90% 以上, Youtube 检测率达 72%~91%。图 12 显示了本任务的研究工作进展。

4.7 加密 Web 识别

加密 Web 识别旨在通过流量分析揭示浏览用户的身份或将用户与浏览特定网页相关联。虽然加密协议使通信内容随机化, 但窃听者仍可以通过 Web 站点流量的统计特征建立站点指纹(Website Fingerprinting, WFP)^[71]。

文献[71]提出只需查看传输的加密数据量即可监视用户是否正在访问某些网站。如访问 www.cnn.com, 无论是否使用 SSL 加密, 都将返回 40 个给定大小的页面, 因此, 这些页面组成的传输大小集就是网页的指纹。然而, 此方法只适用于 HTTP/1.0, 由于现在应用的协议利用了持久连接和流水线(如 HTTP/1.1、Tor 网络), 不会为每个对象打开新的 TCP 连接, 在 HTTP 请求中已经不可能琐碎的区分单个对象。因此, 上述基于对象大小的技术已不可行。文献[72]首次提出基于 IP 包大小和包间到达时间进行 Web 识别攻击, 检测率达到了 20% 左右。文献[73]进

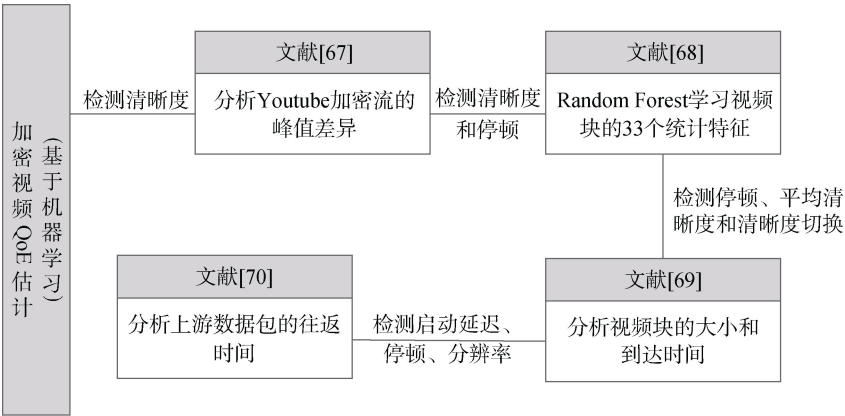


图 12 加密视频 QoE 估计研究工作进展

Figure 12 Researches of encrypted video QoE estimation

行了更广泛的调查并采用更复杂的数据挖掘方法, 即 Jaccard 系数^[74]和 Bayes 分类器^[75]。然而, 在这两种方法中, 唯一考虑的特征是传输数据的包大小, 而忽略了时间和顺序信息。这两种方法的正确分类成功率大约为 70%。

传统加密仅使通信内容随机化, 并且不会隐藏元数据(如随时间推移的流量大小和方向)。匿名通信系统会同时隐藏通信内容和元数据, 从而防止被动攻击这观察通信源或目的地。许多工作表明, 即使客户端在使用加密浏览或使用 Tor 等匿名工具时, 网站指纹攻击也是可能的。文献[76]首先分析了仅基于流、时间和方向的 WFP 特征, 在封闭环境下, 在 Tor 中可识别 55%, 在 JAP^[77]中可识别至 80%。文献[78]提出了使用数据包排序、入站和出站数据包数以及流量突发等特征建立 WFP。文献[79]提出了 k-fingerprinting 攻击(k-FP), k-FP 使用 Random Forest 来提取页面的指纹: 他们用传统的特征来训练 Random Forest, 但实际的指纹是由 Random Forest 中的树叶来表示的。

针对网站指纹这些攻击, 已经提出了许多防御措施。WFP 防御将伪数据包添加到流量中, 并增加了实际数据包的延迟, 旨在隐藏 WFP 攻击所利用的特征, 例如流量突发和数据包长度^[80, 81]。为此, 文献[82]提出了针对 Tor 防御的一种新的网站指纹识别方法。它将网站跟踪数据包表示为元组<timestamp, $\pm$ packet_size>的序列, 使用 CNN 学习深层特征。由于针对以前的工作基于 HTTP/1.1, 且仅对网站的指纹建立指纹, 为了适用于更广泛的 HTTP/2 和对网站的不同网页进行指纹识别, 文献[83]使用本地请求和响应序列(LRRS)作为识别特征, 使用 Deep Forest^[84]提取细粒度特征。文献[85]认为用户在互联网上通常同时访问多个网站, 为此他们提出了一种多标签的 Web 网站识别方法, 该方法提取 TLS v1.2 和 TLS v1.3 握手协议的 Client Hello 数据包中的服务器名称字段作为主要分类特征, 通过 CNN 结合 LSTM 学习特征建立模型, 预测产生流量的网站。图 13 显示了本任务的研究工作进展。

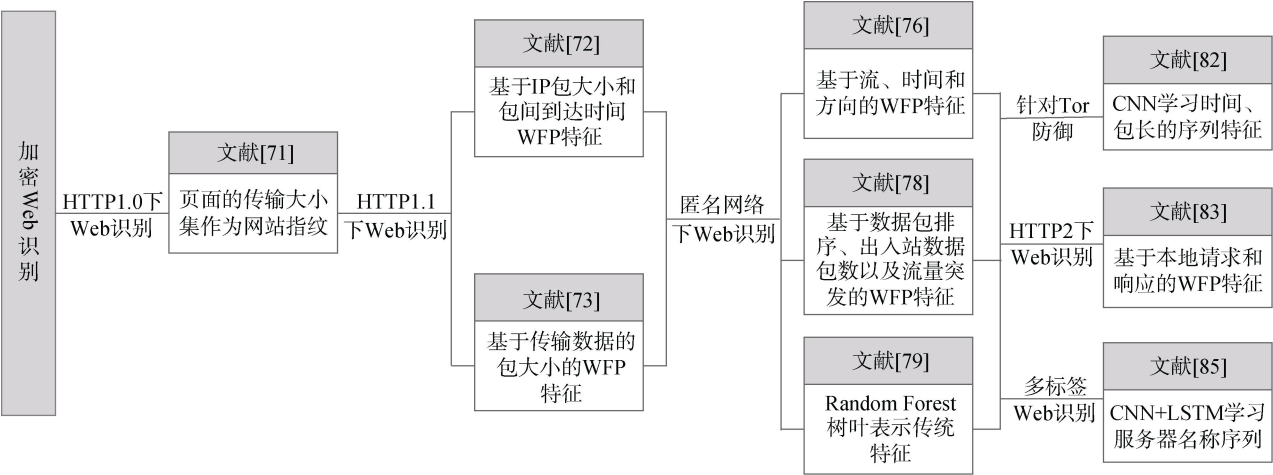


图 13 加密 Web 识别研究工作进展

Figure 13 Researches of encrypted Web identification

4.8 加密恶意流量识别

使用加密方法进行通信已成为恶意软件躲避安全检测的普遍手段,因此,对加密的恶意流量进行有效检测已成为网络攻击检测与发现必不可少的技术。对加密恶意流量检测包括识别恶意软件家族和识别具体的恶意软件。

目前检测加密恶意软件流量的方法主要是使用机器学习算法对数据包的统计特征或握手阶段的明文特征进行学习建模。在现有的研究工作中,包括通过深度学习模型学习加密流原始字节的空间特征进行恶意检测,该类任务与服务识别存在一些交叉关联。他们使用加密流或加密数据包的前 N 字节作为模型输入,使用合适的深度学习模型学习流量的空间特征、时序特征和字节特征^[86-90]。除此之外,还包括对恶意软件的加密通信模式进行针对性研究,引入与恶意检测任务高相关的字段特征。如文献[91]除使用常规流统计特征以外,还引入 TLS 握手阶段的密码套件、证书、密钥等信息对恶意软件家族进行差异分析和分类,准确率可达 90.3%。文献[92]考虑了 TLS 加密流的上下文信息即 DNS 响应和 HTTP 标头,通过提取具有恶意流量检测能力的 TLS、DNS 和 HTTP 特征,使用机器学习算法构建检测器并正确地将各自的流类型分类。该方法能够在零误报率时达到 99.978% 的准确性。图 14 显示了本任务的研究工作进展。

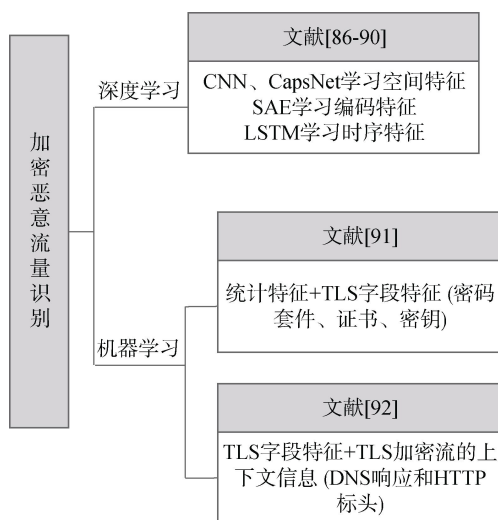


图 14 加密恶意识别研究工作进展

Figure 14 Researches of encrypted malware identification

5 加密流量识别方法

回顾针对加密流量识别的相关研究,可将加密

流量识别方法分为基于深度包检测的识别、基于传统机器学习的识别和基于深度学习的识别。由于加密协议的应用,加密后的流量在熵计算中与非加密流量存在明显差异,因此,早期的加密流量筛选和加密协议识别可通过基于深度包检测的方法进行识别。由于加密协议使得有效载荷随机化,更精细化的识别任务如服务识别、应用识别等都难以通过深度包检测的方法进行识别。因此,基于传统机器学习和深度学习的方法已成为目前主流的加密流量识别方法,它们使用模型算法分析和挖掘不同类别间的统计差异,通过对加密流或数据包的统计特征以及未加密的明文信息进行分类识别。为此,本章概述了基于深度包检测、基于传统机器学习和基于深度学习的典型加密流量识别方法。

5.1 基于深度包检测的识别

基于深度包检测(Deep packet inspection, DPI)的加密流量识别技术,即通过判断流量数据包有效载荷中是否包含某种有效签名来实现流量识别。DPI 通过建立不同流量类别的规则及签名来实现流量匹配,因此具有高识别率及低误报率的特点。目前,已出现大量成熟的基于 DPI 的流量识别工具,下面对几种典型的 DPI 商用或开源工具进行介绍并比较分析。

5.1.1 PACE

PACE 或 PACE2^[93]是 ipoque 集成多种流量识别技术开发的 C 语言商用软件库,它基于 DPI、行为分析、统计分析、启发式分析可对数千种网络协议和应用程序进行检测并实时提取元数据。具体来说,PACE 支持 URL、可执行文件、宽带使用率、持续连接时间、抖动、吞吐量等 5500 多个元数据参数的提取,同时可利用这些内容参数实现协议识别(HTTP、SSL、BitTorrent 等)、应用程序识别(Facebook、Skype、Spotify 等)以及应用内服务识别(Skype 视频通话、Skype 短信等)。此外,PACE 支持每周协议和签名更新,具有高识别率和零误报率。

5.1.2 NBAR

NBAR 或 NBAR2^[94]是 Cisco 基于 DPI 技术实现的商用应用程序识别库,根据数据包头预定义字段提取特征,支持超过 1000 个应用程序签名。NBAR 可实现对 Skype 和 Tor 等匿名应用程序、ms-lync 等业务应用程序、office-365 等云应用程序、Facetime 等移动应用程序以及基于 Web 或动态端口分配等难以分类的应用程序的分类。NBAR 通过每月发布的协议包来添加或更新签名库。同时,NBAR 可根据静态分配的属性来匹配协议或应用程序,并将协议或

应用程序划分为不同的组来应用 QoS 策略。

5.1.3 nDPI

nDPI^[95]是基于 OpenDPI 的经典 DPI 开源库, 根据 LGPL 许可发布, 旨在添加新协议来扩展原始库。nDPI 允许对应用层协议进行检测, 为每个协议设置签名扫描器, 通过签名扫描器检查数据包来实现协议匹配。目前 nDPI 支持 200 多种协议的识别, 为了支持加密连接的检测, nDPI 通过添加 SSL 服务端和客户端的证书解码器来使用加密证书检测已加密的协议, 例如 Citrix Online 和 Apple iCloud 之类的协议。然而, 由于 nDPI 无法支持 TCP 或 IP 有效载荷的重组, 因此无法检测到被分成多个 TCP/IP 数据包的签名^[96]。此外, 由于协议规范经常更新, 尤其是 P2P 协议, 因此需要对 nDPI 的签名库不断更新。

5.1.4 L7-filter

L7-filter^[97]是 Netfilter 中的数据包分类器。它根据应用层数据进行识别检测, 具体通过简单的数字匹配(端口、IP 地址等)、应用层数据的正则表达式匹配以及应用层数据的特征匹配三种方法来实现协议识别, 其中, 后两种基于应用层数据的方法需要消耗更大的 CPU 和内存资源。目前 L7-filter 可匹配使用动态端口的协议、使用非标准端口的协议以及共享端口的协议, 共支持 140 余种协议识别。此外, L7-filter 还提供了识别每种协议的最佳数据库和过滤丢弃匹配的数据包功能。

5.1.5 Libprotoident

Libprotoident^[98]是一个识别应用层协议的轻量级数据包检查(LPI)库, 根据 LGPL 获得许可。它与其他基于 DPI 的识别工具不同, Libprotoident 仅通过检查流在每个方向上发送有效载荷的前 4 个字节、每个方向上第一个有效载荷数据包的大小、流的端口号来建立协议识别规则。目前 Libprotoident 支持 400 多种不同的应用程序协议。该工具可在无法进行完整的深度包检测的情况下实现流量识别, 由于其仅利用有效载荷的前 4 个字节, 很大程度上减少了隐私问题和降低了磁盘存储压力。

5.1.6 Snort

Snort^[99]是一个能够进行实时流量分析数据包记录的开源入侵检测系统。Snort 包括三种配置模式, 一是嗅探器模式, 读取网络数据包并控制台显示; 二是数据包记录模式, 将数据包存储在磁盘中; 三是入侵检测模式, 根据用户定义的规则集, 通过协议分析以及内容搜索和匹配来检测攻击或探测流量, Snort 可检测拒绝服务、缓冲区溢出、端口扫描等多种攻击及探测。此外, Snorby^[100]、Sguil^[101]等第三方工具可与 Snort 连接实现日志分析和管理。

5.1.7 Hyperscan

Hyperscan^[102]是 Intel 开发的一个可用于商用服务器的高性能正则表达式匹配库, 遵循常用 PCRE 库的正则表达式语法, 根据 BSD 获得许可。Hyperscan 采用两种核心技术来进行有效的模式匹配。首先, 它利用图分解将正则表达式匹配转换为一系列字符串和有限自动机匹配。其次, Hyperscan 使用 SIMD 操作^[103]加速了字符串和有限自动机匹配, 从而显著提高了吞吐量。

5.1.8 Suricata

Suricata^[104]是一个开源的网络威胁检测引擎, 支持实时入侵检测, 内联入侵防御, 网络安全监视和脱机 pcap 处理。它通过广泛的规则匹配和签名库来检查网络流量, 支持 Snort 规则语法和 Lua 脚本。具备多线程、协议检测、IP 匹配、文件匹配、日志记录等功能。此外, Suricata 通过 YAML 和 JSON 等标准格式输入输出, 可与 SIEM、Splunk、Logstash 等工具进行集成。可用于 Snort 的第三方工具同样与 Suricata 兼容。

5.1.9 方法比较

本节介绍了利用 DPI 技术实现的几种经典流量识别工具, 其中 PACE, NBAR, nDPI, L7-filter, Libprotoident 主要面向协议识别及应用程序识别, Snort, Hyperscan, Suricata 主要面向恶意入侵检测。文献[105-107]对这些基于 DPI 的工具进行了比较。文献[105]通过比较六种经典的 DPI 流量分类工具(PACE, OpenDPI, L7-filter, nDPI, Libprotoident 和 NBAR), 发现 PACE 是流量识别中最可靠的解决方案, 在开源工具中, nDPI 和 Libprotoident 实现了最佳识别结果。文献[106]对 Suricata 和 Snort 进行比较发现, Suricata 至少可达到 Snort 同等的检测效果, 在高性能计算机上, Suricata 可以比 Snort 处理更多的流量, 并且保持和 Snort 相似的精度。文献[107]对 Hyperscan 和 Snort 进行比较发现, Hyperscan 允许同时匹配数万个正则表达式, 对于真实的流量跟踪, Hyperscan 将 Snort 的性能提高了 8.7 倍。总体而言, 这些工具均依赖于规则匹配和签名库来实现识别, 此类工具的特点是可实现高准确率, 低误报率的识别。然而, 随着动态端口、端口混淆以及加密协议的广泛应用, 基于规则和签名的匹配需要及时更新和添加才能实现有效的流量识别。但是, 这些工具通过检测有效载荷来达到识别和检测的目的, 很大程度上造成了隐私问题。

5.2 基于传统机器学习的识别

由于基于深度包检测的技术提取加密流量有效

载荷的签名需要较大难度,且随着 P2P 应用以及动态端口的普及应用,该技术需要不断更新和添加规则库来维持有效的流量识别,即需要极大的人力物力支持。此外,分析应用层的有效载荷容易造成一定的隐私问题。目前,大量研究人员发现基于传统机器学习的识别技术可通过提取加密流量中的未加密信息和统计特征,利用机器学习模型学习这些特征来实现识别,并且取得了可观的识别效果^[108]。基于传统机器学习的加密流量流程图如图 15 所示,首先对捕获的流量数据进行清洗采样等操作(流量清洗与采样为可选部分),然后针对具体的识别任务进行特征提取,并基于适合的机器学习算法搭建特征学习模型,最后通过分类器实现加密流量各任务的识别。在机器学习识别中,特征提取和用于识别的机器学习算法直接影响最终识别的效果,也是加密流量识别工作中的重点研究内容,因此,本节对这两部分的研究工作进行综述。

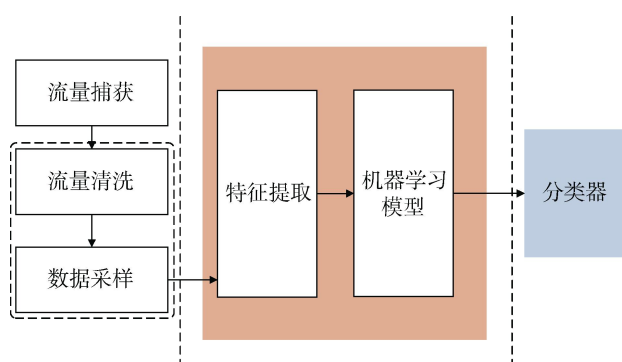


图 15 基于传统机器学习的加密流量识别流程图

Figure 15 Machine learning based encrypted traffic identification process

5.2.1 特征提取

特征提取是提取加密流量传输过程中的统计特征(数据包长度、数据包到达时间等)和加密流量的明文字段(如密码套件、证书、服务器名称等)的过程,挖掘各类别流量区别于其他类别流量的独特模式,特征提取在基于机器学习的加密流量技术中直接影响分类效果。

大量文献针对 TCP 流或 UDP 流提取加密流量的流统计特征。文献[109]针对 TCP 流设计了围绕应用分类的 248 个包级统计特征以及流级统计特征用于表征流量。文献[110]定义了可用于流量应用分类的与有效载荷无关的 22 种流级特征,由流的包长度、包到达时间、流持续时间以及流的字节数和数据包量派生。文献[20]通过提取包长、包数量、包到达时间、流持续时间相关的六个特征对 SSL/TLS 流量进

行具体应用协议(Https, Tor 等)识别;文献[52]使用 MOGA^[111]对仅保留 IP 包头的 SSH 流量提取了 14 个统计特征提取实现应用程序分类。文献[76, 78-79]则在加密 Web 识别中提出了一系列包数量、包大小、峰值、统计时间等统计特征建立网站指纹。同样的,在加密视频 QoE 估计上则更多地关注视频块大小、RTT、包长、吞吐量等统计特征^[68-69]。文献[112]通过 Random Forest 算法评估四个知名流量流特征生成器(Tstat、Tranalyzer、SiLK、Argus)导出特征集,分析最具有代表性的特征,来识别 HTTP/HTTPS 流量中运行的加密应用程序。实验结果表明 Tranalyzer 导出的 87 个流特征在本次评估中具有最佳识别效果。

除此之外,文献[113]引入主机级的特征提取,将 Internet 主机与其提供或使用的应用程序相关联,通过捕获主机在传输层行为的基本模式对主机关联的应用程序进行分类,但仍然对端口和传输层协议过多依赖。BLINC 需要基于主机提取多个流的信息,才能确定主机关联的应用程序。类似地,文献[114]使用流量分散图(TDG)对主机的社交行为进行建模,进而分析和可视化网络流量。

以上均是提取完整流的统计特征,意味着这类流量识别无法实现在线识别。文献[115]提出了仅使用 TCP 连接的前 5 个数据包的大小来识别应用协议。但是加密的 Bittorrent 和 EMule 在其初始数据包上使用填充,因此无法使用此类技术进行分类。因此,文献[27]提出了使用流不多于 100 个数据包的统计特征进行加密流量协议识别。

除通过提取流统计特征(包长、到达时间等)外,还有大量研究工作通过提取未加密的字段信息来实现特定场景下的加密流量识别,这些字段通常具有一定的身份验证功能。如文献[116]仅使用包头字段特征(frame.time_delta, ip.flags 等)实现 SSH、HTTP、HTTPS 等协议的识别。在加密应用识别工作中,更多地关注 SSL/TLS 消息类型、证书长度以及包长度序列^[54, 56-59]。在加密恶意流量分析中,TLS 握手阶段的密码套件、证书、域名等明文信息与恶意检测具有高相关性^[91-92]。表 1 总结了以上工作中常用的加密流量识别特征。

5.2.2 识别模型

1) 有监督

有监督的机器学习模型以有标签的流量特征作为输入,将其分类为已知的类别之一,目前已有大量的研究工作使用各种有监督模型在不同的加密流量识别任务上进行尝试。文献[21]使用 Naïve Bayes 学习与包长、时间间隔、流信息相关的 49 个特征,可

表 1 加密流量识别常用特征
Table 1 Common features of encrypted traffic identification

特征层次	特征	派生特征
统计特征	端口	N/A
	传输层协议	N/A
	流持续时间	N/A
	包到达时间(IAT)	最大、最小、平均、差值、四分位、中位
	以太网字节	最大、最小、平均、差值、四分位、中位
	IP 数据包载荷字节	最大、最小、平均、差值、四分位、中位
	IP/TCP 包头字节	最大、最小、平均、差值、四分位、中位
	数据包数量	客户端-服务端(总数、ACK 数据包数量、无载荷和 SYN/FIN/RST 标志的 ACK 包数量、携带 TCP SACK 块的包数量、携带 D-SACK 块的包数量、SACK 块的最大值、携带 TCP 有效载荷的包数量、重传的包数量、窗口探测的包数量、乱序到达的包数量、TCP 各标志的包数量、初始窗口包数量)服务端-客户端(同上)
		客户端-服务端(不包含重传/窗口探测、包含重传/窗口探测、仅重传、仅窗口探测、URG 标志、初始窗口)、服务端-客户端(同上)
	发送字节数	客户端-服务端, 服务端-客户端
字段特征	吞吐量	客户端-服务端, 服务端-客户端
	往返时间(RTT)	最大、最小、平均、均值
	字节分布 ^[67]	N/A
	证书特征	证书长度、证书主体、证书是否自签名、
	TLS/SSL 信息	TLS 版本、TLS 消息类型、TLS 扩展列表、SNI
	密码套件	密码套件列表、公共密钥长度
	DNS 跟踪信息	域名、IP、TTL
	HTTP 标头	请求头、响应头、内容类型、返回代码

识别 IPsec、PPTP 加密的 HTTP、FTP、SMTP 及 SSH 流量。文献[117]使用 SVM^[118]学习不依赖有效载荷信息的流统计特征进行服务分类。文献[119]使用 C4.5^[120]机器学习算法以数据包长度、数据包到达时间和流向作为特征,对 SSL 加密应用程序进行分类。文献[121]提出用与时间相关的流统计特征表征流量,并通过 C4.5 和 KNN^[122]建立分类器,可实现较好的 VPN 流量检测以及加密流量服务识别。实验结果表明使用机器学习算法训练的加密流量识别模型往往比传统基于有效载荷的识别方法体现更高的识别能力。

然而,传统机器学习模型会带来沉重的计算资源开销,为此,研究人员通过改善模型或使用其他

方法与机器学习结合来降低计算成本。由于传统的 SVM 流量分类器训练复杂度较高且内存和 CPU 计算开销较大,为此,文献[123]提出了增量 SVM 来减少内存和 CPU 高昂的训练成本,该模型利用新的训练数据集和上次训练过程中产生的支持向量实现分类模型的高频和快速更新。文献[22]通过 DPI 和统计学习联合模型实现了 Skype 流量识别,一方面通过皮尔逊卡方检验检测流量帧结构中的指纹信息,另一方面通过 Naïve Bayes 模型学习 Skype 流量的随机特征,此方法可实现 Skype 误报率几乎为零的识别效果。文献[20]结合基于签名和基于机器学习分阶段实现了加密流量协议识别,首先使用签名规则匹配 SSL/TLS 协议,然后使用 Bayes 模型对已过滤的 SSL/TLS 流量进行具体应用协议识别。文献[124]为了满足基于机器学习的高数据速率流量在线分类技术,提出了两阶段混合数据结构并使用现场可编程门阵列(FPGA)在并行和流水线架构上实现,以获取高吞吐量。

2) 半监督与无监督

半监督与无监督的机器学习模型以少量标签数据或无标签流量特征作为模型输入,输出已知类别或未知类别,该方法不需要大量打标签且能实现未知加密流量识别。文献[125]首次将半监督学习技术用于流量分类工作,提出了一种可同时适用于已知和未知应用程序分类方法。他们使用 K-means^[126]学习流统计特征对网络流量进行应用程序分类。文献[127]通过 GMM^[128]检测 TCP 连接的前三个包的大小来识别 SSL 流量,进而通过检测带有有效载荷的前三个包完成 SSL 流量的应用识别。文献[26]通过分析 SSH 连接的第一个 IP 数据包的统计特征(例如到达时间、方向、长度),使用 K-means 算法来实时识别 SSH 流并对 SSH 流的服务类型进行分类。文献[27]将 K-means 算法的轻量级复杂度和 KNN 算法的准确性结合在一起,通过学习不多于前 100 个数据包的流统计特征实现了在线加密流量协议识别。文献[129]提出了一种基于 K-means 算法的 P2P 流量识别方法,该方法使用 K-means 算法学习 P2P 流量的流统计特征对 BT 协议下的 BitTorrent、BitSpirit 和 eMule 应用程序进行识别分类。

然而,使用聚类进行加密流量识别,往往面临参数选择问题,若簇的个数越多,分类准确率越高,但复杂度也会更高。为此,文献[52]将稳态 MOGA 应用于加密流量特征子空间选择和加密流量聚类中,与基本的 K-means 算法相比,不仅减少了聚类所需的特征数量,而且可使用非常小的 K 值对数据实现

聚类, 最终在 SSH 流量下的应用分类中实现了 92% 的检测率和 0.8% 的误报率。他们还实现了 MOGA 的分层版本^[130], 通过进一步划分原始 MOGA 生成的簇来提高最终簇的纯度, 进而提高 SSH 流量识别的性能。文献[131]提出了一种基于谐波均值的 *K*-means 聚类方法, 该方法通过使用谐波均值来减少随机初始聚类中心的影响, 从而提高了 *K*-means 算法在流量分类中的识别准确性, 且这种方法无需执行密集的训练过程即可获得最佳参数, 可用于实时流量识别。

5.2.3 模型比较

在基于机器学习的识别方法中, 通常使用不同的模型用于不同的识别任务, 且验证数据集与特征集均具有较大差异, 无法针对性地对模型之间的性能进行比较。为此, 出现了大量研究工作对提出的模型进行比较。

文献[132]运用 AdaBoost^[133], Naïve Bayes^[75]和最大熵模型^[134]将网络流量分类为不同的应用协议。他们的结果表明, AdaBoost 在其数据集上表现最佳。SSH 检测率为 86%, 误报率为 0.0%, 但他们采用了有效载荷的前 64 个字节。SSH 流量中有效载荷的前 64 字节属于明文部分, 这对某些加密协议如 Skype 和 VPN 将无法适用。文献[110]使用相同的特征集验证不同机器学习模型(Naïve Bayes, C4.5, Bayesian Network 和 Naïve Bayes 树)的计算性能(建模时间和分类速度), 结果表明各模型的识别精度不相上下, 但 C4.5 具有更快的分类速度, Naïve Bayes 具有更快的建模速度但分类速度最慢。类似地, 文献[23]评估了 RIPPER^[135]和 AdaBoost 算法在 SSH 流量识别上的识别性能, 表明 RIPPER 比 AdaBoost 识别效果更佳, 可实现 99% 的检测率和 0.7% 的假阳率。文献[36]比较了六种机器学习模型(Logistic 回归^[136]、SVM、Naïve Bayes、KNN、Random Forest、GBT^[137]), 以学习与时间相关的流量特征, 对 VPN 或非 VPN 传输的加密流量进行服务识别, 两种集合决策树模型(Random Forest 和 GBT)表现出出色的预测性能。

文献[24]通过 SSH 及 Skype 流量识别来评估五种机器学习算法(AdaBoost、SVM、Naïve Bayes、RIPPER 及 C4.5)的鲁棒性, 其中 C4.5 在不同的网络上识别 SSH 和 Skype 流量比其他算法均好得多。文献[138]使用 C4.5、AdaBoost 以及 GP 学习 Voip 的流特征实现加密 Voip 识别, 其中 C4.5 表现最佳, 可实现 99% 以上的检测率和 1% 以下的假阳率。文献[25]通过进行 SSH 流量识别任务评估了三种机器学习(C4.5、MOGA、*K*-means)在加密流量识别中的识别

性能、复杂性和鲁棒性, 其中, 在相同网络环境中捕获的数据集中, MOGA 实现了最佳的检测率和误报率平衡, 但当训练与测试的网络环境不同时, C4.5 表现出最佳的识别性能, C4.5 的鲁棒性优于 MOGA, *K*-means 聚类复杂性最高。

文献[139]评估了基于专家驱动的方法^[140]和基于数据驱动(以 C4.5 和 PRPPER 为代表)的方法^[23], 用于对 SSH 流量进行应用分类。实验表明, 数据驱动的方法优于专家驱动的方法, 基于 C4.5 学习模型为数据驱动的方法提供了最高性能。综上可见, 基于数据驱动的方法比基于专家驱动的方法表现效果更佳, 而在基于数据驱动的方法中, 相较于其他提出的机器学习算法, C4.5 表现出最佳的识别性能。

5.3 基于深度学习的识别

传统机器学习的识别方法需要专家提取特征, 且特征的选取直接对识别结果起关键作用, 这限制了它的广泛适用性, 且需要一定的人力物力投入。然而, 基于深度学习的方法由于其自身强大的特征学习能力, 成功解决了人工提取特征带来的适应性差、耗费人力物力等若干问题。近年来, 在不同的识别场景下提出了大量基于深度学习的加密流量识别, 其识别流程如图 16 所示。与基于传统机器学习不同的是, 该方法不需要人工提取特征, 首先对流量进行切分等预处理, 然后使用深度学习模型学习深层特征。该类方法在预处理过程中确定的模型输入与训练过程中的模型选取存在直接关系。因此, 根据模型输入不同, 可分为基于原始字节的深度学习识别、基于字段特征的深度学习识别和多模式混合识别方法, 最后选择合适的模型实现加密流量识别。本节根据模型输入的不同对这三种识别方法进行研究概述。

5.3.1 基于原始字节的识别

基于原始字节的识别是将加密流的原始字节数据作为模型输入, 使用深度学习模型学习加密流的空间特征、时序特征和字节特征。文献[87]首次将表示学习用于恶意流量识别中, 该方法提取原始流的前 784 字节, 使用 CNN 自动学习流特征, 该方法在恶意软件识别上可达到 99% 以上的识别率。同时, 他们还提出 1dCNN 比 CNN 具有更佳的加密流量识别能力^[38]。类似地, 文献[39]提出使用 CAE 和 CNN 学习加密流前 1521 字节的深层特征。文献[90]提取加密流的前 900 字节使用 CNN 和 LSTM 实现加密流量识别、服务分类和恶意流量检测。以上工作均是在流级实现加密流量识别。文献[37]将 IP 包头和 IP 包有效载荷的前 1480 字节作为神经网络的输入, 使用

表 2 传统机器学习代表性工作
Table 2 Traditional machine learning representative work

模型类型	识别任务	文献	特征	识别算法	性能
有监督	协议识别	[20]	包长、包数量、包到达时间、流持续时间相关的六个特征	Bayes	识别 HTTPS、TOR、UPDATE、OSCAR 协议, 可实现 F_1 达 94.52% 的协议识别率
		[21]	包长、时间间隔、流信息相关的 49 个特征	Naïve Bayes	识别 IPSec、PPTP 加密的 HTTP、FTP、SMTP 及 SSH 流量, 准确率几乎 100%
		[27]	前 100 个数据包的大小和方向	K -means+ KNN	对特定的 BitTorrent、Skype 协议进行在线识别, 分别实现了 97.4% 和 94% 的识别率
		[22]	帧结构的指纹信息+流统计特征	皮尔逊卡方检验+Naïve Bayes	实现 Skype 误报率几乎为零的识别效果
		[132]	有效载荷的前 64 个字节	AdaBoost, Naïve Bayes 和最大熵模型	AdaBoost 在其数据集上表现最佳。SSH 检测率为 86%, 误报率为 0.0%
		[23]	协议、包长、到达时间、流持续时间相关的统计特征	RIPPER 和 AdaBoost	RIPPER 比 AdaBoost 识别效果更佳, 可实现 99% 的检测率和 0.7% 的假阳性
		[24]	协议、包数量、到达时间、包长度相关的流统计特征	AdaBoost、SVM、Naïve Bayes、RIPPER 及 C4.5	C4.5 在不同的网络上识别 SSH 和 Skype 流量比其他算法均好得多
	服务识别	[25]	协议、包长、流到达时间、包量、字节数	C4.5、MOGA、 K -means	相同网络环境中捕获的数据集中, MOGA 实现最佳的检测率和误报率平衡, 但当训练与测试的网络环境不同时, C4.5 表现出最佳的识别性能, C4.5 的鲁棒性优于 MOGA, K -means 复杂性最高
		[141]	39 个包头字段特征和 22 个流统计特征	C4.5, SBB-GP 和 AdaBoost	基于流的统计特征比基于数据包头的特征要好得多, 在所有情况下, 基于 SBB-GP 的分类方法优于其他方法
		[36]	与时间相关的统计特征	Logistic 回归、SVM、Naïve Bayes、KNN、Random Forest、GBT	两种集合决策树模型(Random Forest 和 GBT)表现出出色的预测性能
		[117]	不依赖有效载荷的流统计特征	SVM	97.17% 识别精度
		[121]	与时间相关的流统计特征	C4.5, KNN	准确率达 80% 以上
		[123]	流统计特征	增量 SVM	有效减少内存和 CPU 训练成本
		[138]	协议、包数量、到达时间、包长度相关的流统计特征	C4.5, AdaBoost 以及 GP	C4.5 表现最佳, 可实现 99% 以上的检测率和 1% 以下的 FPR
	应用识别	[110]	流的包长度、包到达时间、流持续时间、字节数和数据包量	Naïve Bayes, C4.5, Bayesian Network 和 Naïve Bayes 树	C4.5 具有更快的分类速度, Naïve Bayes 具有更快的建模速度但分类速度最慢
		[112]	Tstat、Tranalyzer、SiLK、Argus 提取的特征	Random Forest	Tranalyzer 导出的 87 个流特征在本次评估中具有最佳识别效果
		[54]	SSL/TLS 消息类型	一阶马尔可夫链	TPR: 64.58%, FPR: 31.36%
		[56]	SSL/TLS 消息类型+证书包长度	二阶马尔可夫链	TPR: 86.74%, FR: 10.52%
		[57]	SSL/TLS 消息类型+证书包长度+第一个应用程序数据包大小	二阶马尔可夫链	TPR: 91.22%, FPR: 7.57%
		[58]	消息类型序列+包长度序列	多属性马尔可夫模型	TPR: 94.00%, FPR: 0.33%
		[59]	握手证书+DNS 跟踪信息	XGBoost	在包含 16 个应用程序的真实跟踪集上达到 98.69% 的精度, 复杂度优于其他方法
QoE 估计		[119]	以数据包长度、数据包到达时间和流向作为特征	C4.5	对 8 种 Google 应用程序进行分类, 分类率达到 95.39%
		[139]	协议、包长、到达时间、流持续时间相关的统计特征	C4.5 和 PRPPER	数据驱动的方法优于专家驱动的方法, 基于 C4.5 学习模型为数据驱动的方法提供了最高性能
		[68]	加密视频流量中提取包长统计、5s 间隔的传输数据大小、包数统计、互达时间统计、吞吐量统计和 TCP 标志计数方面的 33 个特征	Random Forest	将加密视频的 QoE 评估为高、中、低三级, 可实现 84% 的 QoE 分类精度
		[69]	视频块的大小和到达时间方面的 7 组统计特征	Random Forest	检测影响 QoE 的停顿、平均质量和质量切换, 准确率达 76%~91.8%

续表

模型类型	识别任务	文献	特征	识别算法	性能
有监督	Web 识别	[76]	流、时间和方向	SVM	Tor 中的 Web 识别达 55%, JAP 中 Web 识别达 80%
		[78]	数据包排序、入站和出站数据包数以及流量突发等特征	KNN	在封闭环境中, 达到 91% 的准确率, 在有 5000 个网站的开放环境中, 它达到了 86% 的 TPR 和 0.6% 的 FPR
		[79]	数据包数量统计、传入传出数据包占比、每秒的数据包数量、数据包达到时间统计等	Random Forest	在封闭环境下达到 91% 的准确率, 在开放环境下达到 88% 的 TPR 和 0.5% 的 FPR
	恶意识别	[91]	统计特征+TLS 字段特征(密码套件、证书、密钥)	Logistic 回归	恶意家族识别准确率达 90.3%
		[92]	TLS 字段特征+TLS 加密流的上文信息(DNS 响应和 HTTP 标头)	Logistic 回归	在零误报率时达到 99.978% 的准确性
无监督/半监督	协议识别	[115]	TCP 连接的前 5 个数据包的大小	K-means	识别 edonkey、ftp、http、kazaa、nntp、pop3、smtp、ssh、https、pop3 协议, 准确率达 80% 以上
		[116]	包头字段特征	SBB based GP	基于 SBB 的 GP 解决方案能够在不降低准确性的情况下提供更简单(相比 C4.5 和 AdaBoost)的解决方案
		[127]	TCP 连接的前三个包的大小	GMM	SSL 下应用协议识别准确率达 85% 以上
	服务识别	[131]	包长、数量、到达时间、流持续时间、协议相关统计特征	基于谐波均值的 K-Means 聚类	使用谐波均值可以减少随机初始聚类中心的影响, 优于传统 K-means 算法
		[125]	数据包大小、数量、字节派生的 11 个流统计特征	K-means	可实现已知服务和未知服务的识别, 精度达 90% 以上
	应用识别	[52]	包数量、到达时间、分段长度、包长度、标志位计数	MOGA	识别 SSH 下的应用程序, 达 92% 的检测率和 0.8% 的误报率
		[130]	包数量、到达时间、分段长度、包长度、标志位计数	MOGA 的分层版本	识别 SSH 下的应用程序, DR 率为 96.3%, FPR 为 1.3%。
		[129]	端口、IP、包数量、有效载荷长度	K-means	对 BT 协议下的 BitTorrent、BitSpirit 和 eMule 应用程序进行识别分类, 平均 TPR 分别达 92.64%, 96.22%, 99.76%

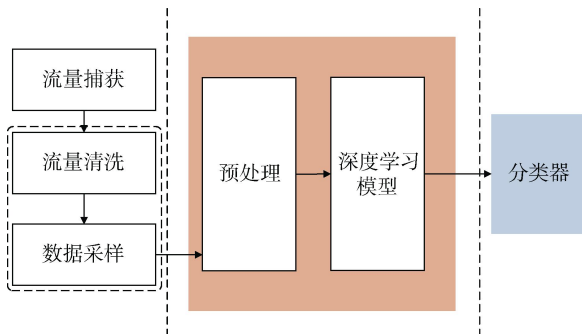


图 16 基于深度学习的加密流量识别流程图
Figure 16 Deep learning based encrypted traffic identification process

CNN 和 SAE 对数据包字节数据进行建模, 实现了 VPN 下加密流量服务识别和 Tor 下的加密流量应用识别。

除了提取直接原始流的前 N 字节外, 使用 CNN、1dCNN 学习加密流的空间特征外, 还有一些研究方法将流量矩阵化, 即提取原始流的前 N 个包, 每个包提取前 M 字节。这类方法更多地是将数据包之间的时序特征考虑在内, 例如, 文献[45]提出了一种依靠数据包的时间依赖性来表征流量行为, 并使用 LSTM

学习流量的时间依赖性来实现 VPN 加密流量下的服务识别, 在表征流量行为方面, 该方法基于应用层数据加密随机性无法作为流的有效特征, 仅提取每个数据包的数据长度、IP 头、TCP/UDP 头和应用层有效载荷的前 50 字节作为包特征, 前 5 个数据包的包特征按照到达时间组成流特征表征流的行为。文献[51]将 CNN 和 LSTM 相结合, 使用 CNN 学习流中单个数据包特征, LSTM 学习流中任意三个连续数据包的包间特征, 通过实验验证, 该方法优于 1dCNN。文献[47]将网络流(前 10 个包, 每个包前 1500 字节)视为时间序列, 并构建了基于注意力的 LSTM 和层次注意力网络(HAN)^[142]学习流的时序特征, 实现 VPN 服务分类。

由于真实网络环境捕获的原始流量存在大量与识别任务无关的噪声, 一些研究工作通过不同的流量清洗工作来提高加密流量识别精度。文献[86]认为加密流的某些字段是干扰加密流量识别的无效特征, 因此他们在预处理阶段删除流的 TCP 或 UDP 报文头以及 MAC 地址后, 选择流的前 900 字节作为模型数据, 然后使用 1dCNN、LSTM、SAE 学习流量的空间特征、时序特征和编码特征, 最终形成加密流量识别

和恶意流量检测框架。实验结果表明, 在加密流量识别中, 1dCNN 识别能力最佳, 在入侵检测中, LSTM 检测结果最佳。文献[40]认为原始流量中 SNMP、DNS、ARP 等协议产生的数据包, 无法作为加密服务和应用识别的有效流量, 因此他们提出了一种流量二次切分机制, 此外, 他们验证 CapsNet 可以学习加密流量的字节特征, 比 1dCNN、SAE、CNN-LSTM 模型表现出更高的识别能力。类似地, 文献[88]在恶意流量检测中验证了 CapsNet 比 CNN、CNN-LSTM、逻辑回归具有更有效的识别准确率。在字节特征挖掘上, 文献[32]认为不同位置的字节会相互影响, 例如 IP 包头中的“版本”决定“源 IP 地址”的长度是四个字节还是十六个字节, 因此, 他们提出了一种基于自注意力机制的流量分类方法, 该方法提取包的前 40 字节(IP 头和 TCP/UDP 头), 使用自注意力机制将数据包中字节的依赖性考虑在内, 从而实现更准确的协议分类和 VPN 服务分类。

5.3.2 基于字段特征的识别

基于字段特征是通过提取加密流的某些字段内容如数据包大小、数据包 SSL/TLS 消息类型等, 按照数据包到达时间组成特征序列, 使用深度学习模型学习序列的深层特征实现加密流量识别。其中包括通过使用空间模型如 CNN 学习挖掘加密流特征的空间差异。例如, 文献[30]提出了一种基于序列到图像的在线 IP 流量分类框架。该方法首先提取双向流的前 10 个数据包大小序列、数据包到达时间序列和数据包方向序列, 然后利用 RKHS 均值^[143]嵌入将流序列转换为图像, 进而使用 CNN 学习深层特征, 最终可实现协议分类。文献[144]提出了一种分级的加密流量识别方法, 将流量逐渐细分为四级, 首先将流量分为四类(交互式、流式、批量数据传输、交易), 然后在已识别四大类的基础上继续细分具体的应用程序和设备类型等。该方法仅使用每条流的前 16 个数据包的大小、到达时间、方向、传输层协议作为流量分类的特征, 通过 CNN 建模可达到 95% 的识别准确性。文献[46]根据流的数据包大小和数据包到达时间的序列特征创建图像, 利用 CNN 学习图像的空间特征, 实现 VPN 流量的服务识别。文献[70]提出了一种 DeepQoE 的方法来对加密流量进行实时视频 QoE 评估。他们仅使用上游数据包的 RTT 作为关键特征, 通过 CNN 提取高维特征, 对 QoE 进行细粒度评估。文献[82]使用 CNN 学习流的包长度特征实现 Web 识别。

另外包括使用时序模型挖掘加密流特征的时序差异。文献[85]提取 TLS v1.2 和 TLS v1.3 握手协议

的 Client Hello 数据包中的服务器名称字段作为主要分类特征, 使用任意 20 个连续 Client Hello 的服务器名称组成序列, 通过 CNN 与 LSTM 学习特征建立模型, 预测产生流量的网站。文献[60]提出了一种流序列网络(FS-Net)的加密流量应用识别方法, FS-Net 基于 LSTM 构造多层编码-解码结构, 从原始流序列信息中挖掘流的包长度序列特征, 并且引入重建机制来增强特征有效性。文献[31]只考虑流生命周期中的前 20 个数据包, 提取数据包的源端口、目的端口、有效载荷字节数、TCP 窗口大小、到达时间和数据包方向, 基于到达时间构成流的特征序列, 使用 CNN 和 LSTM 学习深层特征, 实现协议和服务识别。

5.3.3 多模式混合的识别

多模式混合的识别是将原始字节和字段特征同时作为模型输入。文献[53]提出了一种多模式移动加密流量应用识别框架 MIMETIC, 该框架同时考虑加密流量原始字节和字段特征, 使用 CNN 学习流数据前 576 字节的空间特征, 使用 GRU 学习前 12 个包的字段(传输层有效载荷大小、TCP 窗口大小、包到达时间、包方向)序列特征, 然后使用全连接层将 CNN 和 GRU 输出连接以实现多模式特征学习, 通过 Android 和 iOS 的真实数据集验证, 该方法比文献[31, 38, 145]提高 8.58%。为平衡加密流量识别模型的精度和鲁棒性, 文献[44]提出了一个采用综合有效流量信息的加密流量分类框架 CETAnalytics, 该框架将应用层有效负载内容和有效负载的统计信息进行合并, 采用神经网络(1dCNN、GRU、Attention)作为基础单元, 通过模块化组合实现了加密流量服务识别。

6 研究展望

本文概述了以往的加密流量识别研究工作, 从加密流量识别任务和加密流量识别方法两个维度对相关工作进行总结和分析。基于以上加密流量研究工作, 在识别任务上, 任务划分根据不同场景需求逐渐多样化; 在研究方法上, 基于机器学习的加密流量识别分析技术已经成为当前主流的研究方法。然而, 加密协议和新型网络协议仍然不断出现, 加密恶意流量的盛行严重危害着用户设备和数据安全, 这些问题仍然为加密流量分析带来新的挑战。为此, 本文对今后的加密流量研究方向进行展望。

1) 加密流量细粒度分析

目前大多数应用程序都在趋向于综合服务和活动, 如大量即时通信工具 Facebook、WeChat 既有基本的即时文本聊天, 也有发送位置、语音通话、视频聊天已经发布动态等多种活动内容。然而, 传统

表 3 深度学习代表性工作
Table 3 Deep learning representative work

模型输入	识别任务	文献	预处理	模型	性能
基于原始字节	协议识别, 应用识别,	[32]	数据包的前 40 字节	Self Attention	协议识别的 F_1 达 98.62%, 应用识别的 F_1 达 98.93%
		[38]	原始流的前 784 字节	1dCNN	1dCNN 比 CNN 具有更佳的加密流量识别能力
		[39]	原始流的前 1521 字节, 删除小于 7 个包的流以及小于 64 字节的包	CAE, CNN	CAE 实现 90% 的识别精度, CNN 实现 92.92% 的识别精度
	服务识别	[45]	仅提取每个数据包的数据长度、IP 头、TCP/UDP 头和应用层有效载荷的前 50 字节作为包特征, 前 5 个数据包的包特征按照时间组成流特征	LSTM	服务识别的 F_1 达 98%
		[51]	流的任意三个连续数据包(每个数据包统一截取前 784 字节)	CNN+LSTM	服务识别中达到 91% 的识别精度, 比文献[37]提高 5%, 训练时间和测试时间均优于文献[37]提到的 1dCNN
		[47]	网络流(前 10 个包, 每个包前 1500 字节)视为时间序列	Attention+LSTM, HAN	Attention Based LSTM 和 HAN 分别实现 91.2% 和 89.5% 识别精度, 优于文献[36-37]
	服务识别, 应用识别	[37]	IP 包头和 IP 包有效载荷的前 1480 字节, 删除 MAC 地址和 IP 地址	CNN, SAE	应用识别中: CNN F_1 达 98%, SAE F_1 达 95% 服务识别中: CNN F_1 达 93%, SAE F_1 达 92%
		[40]	流的前 784 字节, 删除 MAC 地址和 IP 地址	CapsNet	服务识别达 99.1% 的识别精度, 应用识别达 99.8% 的识别精度, 优于文献[36-37, 43]
		[90]	原始流的前 900 字节	CNN+LSTM	达到 99.98% 的识别精度
	服务识别、恶意识别	[86]	删除流的 TCP 或 UDP 报文头以及 MAC 地址后, 选择流的前 900 字	1dCNN、LSTM、SAE	加密流量识别中, 1dCNN 识别能力最佳, 识别精度达 99.85%; 在恶意识别中, LSTM 检测结果最佳, 识别精度达 99.41%, 优于文献[37]
基于字段特征	恶意识别	[87]	原始流的前 784 字节	CNN	恶意软件识别达到 99% 以上的识别率
		[88]	连续 10 个数据包, 数据包截取前 1000 字节	CapsNet	优于 CNN、CNN-LSTM、逻辑回归
	协议识别、应用识别	[30]	提取流的前 10 个数据包大小序列、数据包到达时间序列和数据包方向序列, 然后利用 RKHS 均值嵌入将流序列转换为图像	CNN	协议识别的精度达 99.84%, 应用识别的精度达 88.42%
		[31]	前 20 个数据包端口有效载荷字节数、TCP 窗口大小、到达时间和数据包方向	CNN+LSTM	实现 96.12 的识别精度
	服务识别, 应用识别	[46]	根据流的数据包大小和数据包到达时间的序列特征创建图像	CNN	除 Browsing 外, 识别精度达 96% 以上, Voip 和视频类应用识别精度达 99.7%
		[60]	数据包长度序列	LSTM	在 18 个应用程序分类中实现 99.14% 的 TPR 和 0.05% 的 FPR
	应用识别	[144]	每条流的前 16 个数据包的大小、到达时间、方向、传输层协议作为流量分类的特征	CNN	达到 95% 的识别精度
	QoE 估计	[70]	上游数据包的 RTT	CNN	实现在线细粒度检测 QoE 指标包括启动延迟、重新缓冲(停顿)、视频分辨率, Bilibili 上的检测率可高达 90% 以上, Youtube 检测率达 72%~91%
		[85]	任意 20 个连续 Client Hello 的服务器名称序列	CNN+LSTM	多标签的 Web 识别, 通过 20 个网站测试可实现 95.5% 的准确率
	Web 识别	[82]	前 5000 个数据包的包长度序列	CNN	可抵抗 WF 防御, 在封闭环境中达到 98.3% 的准确率, 在 WTF-PAD 防御和 Walkie-Talkie 防御下, 准确率分别超过 90% 和 49.7%
多模式混合	服务识别	[44]	流的有效载荷和有效载荷的统计特征合并	CNN+GRU+Attention	同时实现高精度和鲁棒的泛化性能
	应用识别	[53]	流数据前 576 字节的空间特征, 12 个包的字段(传输层有效载荷大小、TCP 窗口大小、包到达时间、包方向)序列特征	CNN+GRU	通过 Android 和 iOS 的真实数据集验证, 比文献[31, 38, 145]提出的方法提高 8.58%

的应用识别已无法帮助网络管理人员进行合理的 QoS 规划和行为分析, 他们需要更精细地了解应用程序内部的活动, 因此, 对应用内服务进行精细化识别已经成为服务商面临的一项具有挑战性和紧迫性的任务。

此外, 在加密流量中, 视频流量的比重已经超过了 75%, 海量的移动视频流量数据给移动网络服务提供商带来了巨大的挑战, 他们既要保证这些视频流量不会影响到其他用户的上网体验, 又要保证这些视频用户获得优质的视频观看体验。因此, 移动网络服务提供商需要对当前流行的视频网络流量进行识别分析, 通过识别视频流量的分辨率、停顿等参数来评估 QoE, 从而提供更佳的用户体验。

2) 私有协议流量分析

随着网络使用成本的降低和网络协议良好的可扩展性, 私有协议已成为许多厂商、应用程序使用的数据保护手段。私有协议一方面通过加密、匿名保护个人隐私数据不被追踪和泄露, 另一方面, 也会被不法分子利用来隐藏其恶意行为。所以对私有协议的流量进行分析具有重要意义, 通过对私有协议流量分析, 能够更好地对流量进行分类和管理, 对真实的 IP 进行追踪溯源, 同时还能够检测一些恶意的网络服务。

3) 加密恶意流量分析

通过对流量的加密能够较好地保护用户隐私, 但同时也导致恶意软件能够利用加密流量进行通信以躲避安全软件的检测, 因此在加密流量识别中, 有效识别其中的恶意流量已是一个重要的研究内容。通过加密恶意流量识别, 可以帮助管理人员及时阻止恶意行为的传播, 保护用户的设备及数据安全。除本文提及的相关工作外, 在加密恶意流量分析中, 由于存在检测主机联网环境等良性流量, 导致恶意流量分析中产生噪声, 有效降低噪声的影响提高恶意流量识别精度是其中一个重要研究内容。同时, 大量工作均是识别已知恶意软件, 但恶意软件通常通过变体和更新规避检查, 因此, 未知恶意软件的识别也是一个重要研究内容。

4) 加密流量训练集不平衡处理

本文提及的加密流量识别方法大多基于大量标签数据集的假设, 不关心数据集收集与标签的问题。然而, 加密流量数据不平衡已经是加密流量识别分析技术中一个广泛存在的问题, 在服务分类中, 视频、文件传输类流量远大于即时通信、邮件类流量; 在加密恶意流量分类中, 良性流量远大于恶意流量。而各标签的严重不平衡将直接导致模型的识别能力,

处理流量数据不平衡的问题可使机器学习算法实现更好的流量分类性能。

7 结论

为有效防护网络攻击和保护用户隐私, 流量加密成为广泛使用的一种技术手段, 然而, 大量攻击行为同样通过加密手段躲避防护检查, 给网络设备和用户带来了极大的安全威胁。为此, 加密流量识别技术已成为攻击发现和网络管理的关键技术。本文调研了现有的加密流量研究, 主要针对加密流量识别任务以及现有的加密流量识别方法进行归纳总结。

其中, 应用内服务识别由于通过分析应用内用户具体行为可以提供更精确的网络资源分配, 因此具有广阔的应用场景; 同时, 加密恶意识别由于通过统计特征和字段特征进行恶意行为表征, 弥补了传统流量入侵检测方法无法检测加密恶意行为的不足, 因此对异常分析具有重要意义。通过现有加密流量识别方法的比较, 基于深度学习的多模式混合方法由于其融合多维流量信息, 在提高加密流量识别能力的同时还具有鲁棒和稳定的特点, 因此有望成为加密流量识别的突破性技术。最后, 本文针对当前加密流量研究现状提出研究展望。由于当前应用程序已从单一服务转变为综合服务, 一种应用程序同时包含多种服务和活动, 因此对加密流量进行细粒度以提供更精确的 QoS 估计具有较为广阔的应用场景; 随着网络协议的可扩展性, 应用程序尤其恶意应用程序趋向于使用私有协议加密的方式进行数据隐私保护甚至躲避安全检查, 因此对私有协议流量进行分析对已知异常检测和未知异常发现具有重要意义。另外, 虽然学术界提出了大量加密流量识别方法, 但加密恶意识别中噪声问题、恶意软件变形和更新问题仍然影响当前识别方法的有效性, 此外, 加密流量中的训练集不平衡分布也是影响识别能力的一个重要因素。因此, 研究以上问题同样是当前加密流量识别中的研究重点。

参考文献

- [1] Li Y, Chen M. Software-Defined Network Function Virtualization: A Survey[J]. *IEEE Access*, 2015, 3: 2542-2553.
- [2] Atzori L, Iera A, Morabito G. The Internet of Things: A Survey[J]. *Computer Networks*, 2010, 54(15): 2787-2805.
- [3] Andrews J G, Buzzi S, Choi W, et al. What Will 5G Be [J]. *IEEE Journal on Selected Areas in Communications*, 2014, 32(6): 1065-1082.
- [4] Cisco V. Cisco visual networking index: Forecast and trends, 2017-2022. White Paper, 2018, 1: 1.

- [5] Sandvine I. Global internet phenomena report. North America and Latin America, 2018.
- [6] Rescorla E, Dierks T. The transport layer security (TLS) protocol version 1.3. 2018.
- [7] Thayer R, Doraswamy N, Glenn R. RFC2411: IP Security Document Roadmap. Technical Report. RFC Editor, 1998.
- [8] Cao Z, Xiong G, Zhao Y et al. A survey on encrypted traffic classification[C]. *International Conference on Applications and Techniques in Information Security*, 2014: 73-81.
- [9] Dainotti A, Pescapé A, Claffy K C. Issues and Future Directions in Traffic Classification[J]. *IEEE Network*, 2012, 26(1): 35-40.
- [10] Dhote Y, Agrawal S, Deen A J. A Survey on Feature Selection Techniques for Internet Traffic Classification[C]. *2015 International Conference on Computational Intelligence and Communication Networks*, 2016: 1375-1380.
- [11] Rezaei S, Liu X. Deep Learning for Encrypted Traffic Classification: An Overview[J]. *IEEE Communications Magazine*, 2019, 57(5): 76-81.
- [12] Ylönen T, Lonvick C. The Secure Shell (SSH) Protocol Architecture[J]. *RFC*, 2006, 4251: 1-30.
- [13] Scott C, Wolfe P, Erwin M. Virtual private networks[M]. *O'Reilly Media, Inc.*, 1999.
- [14] Chaabane A, Manils P, Ali Kaafar M. Digging into Anonymous Traffic: A Deep Analysis of the Tor Anonymizing Network[C]. *2010 Fourth International Conference on Network and System Security*, 2010: 167-174.
- [15] Dorfinger P, Panholzer G, John W. Entropy estimation for real-time encrypted traffic identification (short paper)[C]. *International workshop on traffic monitoring and analysis*, 2011: 164-171.
- [16] Olivain J, Goubault-Larrecq J. Detecting subverted cryptographic protocols by entropy checking[D]. *LSV, ENS Cachan*, 2006.
- [17] Mamun M S I, Ghorbani A A, Stakhanova N. An entropy based encrypted traffic classifier[C]. *International Conference on Information and Communications Security*, 2015: 282-294.
- [18] Cheng G, Hu Y. Encrypted Traffic Identification Based on N-Gram Entropy and Cumulative Sum Test[C]. *The 13th International Conference on Future Internet Technologies*, 2018: 1-6.
- [19] Rukhin A, Soto J, Nechvalat J, et al. A statistical test suite for random and pseudorandom number generators for cryptographic applications. Technical Report. Booz-allen and hamilton inc mclean va, 2001.
- [20] Sun G L, Xue Y B, Dong Y F, et al. An Novel Hybrid Method for Effectively Classifying Encrypted Traffic[C]. *2010 IEEE Global Telecommunications Conference GLOBECOM*, 2011: 1-5.
- [21] Okada Y, Ata S, Nakamura N, et al. Application Identification from Encrypted Traffic Based on Characteristic Changes by Encryption[C]. *2011 IEEE International Workshop Technical Committee on Communications Quality and Reliability*, 2011: 1-6.
- [22] Bonfiglio D, Mellia M, Meo M, et al. Revealing Skype Traffic: When Randomness Plays with You[C]. *The 2007 conference on Applications, technologies, architectures, and protocols for computer communications*, 2007: 37-48.
- [23] Alshammari R, Nur Zincir-Heywood A. A Flow Based Approach for SSH Traffic Detection[C]. *2007 IEEE International Conference on Systems, Man and Cybernetics*, 2008: 296-301.
- [24] Alshammari R, Zincir-Heywood A N. Machine Learning Based Encrypted Traffic Classification: Identifying SSH and Skype[C]. *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2009: 1-8.
- [25] Arndt D J, Zincir-Heywood A N. A Comparison of Three Machine Learning Techniques for Encrypted Network Traffic Analysis[C]. *2011 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2011: 107-114.
- [26] Maiolini G, Baiocchi A, Iacovazzi A, et al. Real time identification of SSH encrypted application flows by using cluster analysis techniques[C]. *International Conference on Research in Networking*, 2009: 182-194.
- [27] Bar-Yanai R, Langberg M, Peleg D, et al. Realtime classification for encrypted traffic[C]. *International Symposium on Experimental Algorithms*, 2010: 373-385.
- [28] Liu Y, Chen J F, Chang P, et al. A Novel Algorithm for Encrypted Traffic Classification Based on Sliding Window of Flow's First N Packets[C]. *2017 2nd IEEE International Conference on Computational Intelligence and Applications*, 2017: 463-470.
- [29] Wang Z. The applications of deep learning on traffic identification[J]. *BlackHat USA*, 2015, 24(11): 1-10.
- [30] Chen Z T, He K, Li J, et al. Seq2Img: A Sequence-to-Image Based Approach towards IP Traffic Classification Using Convolutional Neural Networks[C]. *2017 IEEE International Conference on Big Data*, 2018: 1271-1276.
- [31] 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.
- [32] Xie G R, Li Q, Jiang Y, et al. SAM: Self-Attention Based Deep Learning Method for Online Traffic Classification[C]. *The Workshop on Network Meets AI & ML*, 2020: 14-20.
- [33] Zhang Y, Mao Z M, Zhang M. Detecting Traffic Differentiation in Backbone ISPs with NetPolice[C]. *The 9th ACM SIGCOMM conference on Internet measurement*, 2009: 103-115.
- [34] Okada Y, Ata S, Nakamura N, et al. Comparisons of Machine Learning Algorithms for Application Identification of Encrypted Traffic[C]. *2011 10th International Conference on Machine Learning and Applications and Workshops*, 2012: 358-361.
- [35] 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: 407-414.
- [36] Bagui S, Fang X G, Kalaimannan E, et al. Comparison of Machine-Learning Algorithms for Classification of VPN Network Traffic Flow Using Time-Related Features[J]. *Journal of Cyber Security Technology*, 2017, 1(2): 108-126.
- [37] Lotfollahi M, Siavoshani M J, Zade R S H, et al. Deep Packet: A Novel Approach for Encrypted Traffic Classification Using Deep Learning[J]. *Soft Computing*, 2020, 24(3): 1999-2012.
- [38] Wang W, Zhu M, Wang J L, et al. End-to-End Encrypted Traffic Classification with One-Dimensional Convolution Neural Networks[C]. *2017 IEEE International Conference on Intelligence and Security Informatics*, 2017: 43-48.

- [39] Guo L L, Wu Q Q, Liu S L, et al. Deep Learning-Based Real-Time VPN Encrypted Traffic Identification Methods[J]. *Journal of Real-Time Image Processing*, 2020, 17(1): 103-114.
- [40] Cui S S, Jiang B, Cai Z Z, et al. A Session-Packets-Based Encrypted Traffic Classification Using Capsule Neural Networks[C]. *2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems*, 2019: 429-436.
- [41] O'Shea K, Nash R. An Introduction to Convolutional Neural Networks[EB/OL]. 2015: arXiv: 1511.08458. <https://arxiv.org/abs/1511.08458>
- [42] Schölkopf B, Platt J, Hofmann T. Greedy Layer-Wise Training of Deep Networks[C]. *The 2006 Conference*, 2007: 153-160.
- [43] Sabour S, Frosst N, Hinton G E. Dynamic Routing between Capsules[C]. *The 31st International Conference on Neural Information Processing Systems*, 2017: 3859-3869.
- [44] Dong C, Zhang C, Lu Z G, et al. CETAnalytics: Comprehensive Effective Traffic Information Analytics for Encrypted Traffic Classification[J]. *Computer Networks*, 2020, 176: 107258.
- [45] Vu L, Thuy H V, Nguyen Q U, et al. Time Series Analysis for Encrypted Traffic Classification: A Deep Learning Approach[C]. *2018 18th International Symposium on Communications and Information Technologies*, 2018: 121-126.
- [46] Shapira T, Shavitt Y. FlowPic: Encrypted Internet Traffic Classification is as Easy as Image Recognition[C]. *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops*, 2019: 680-687.
- [47] Yao H P, Liu C, Zhang P Y, et al. Identification of Encrypted Traffic through Attention Mechanism Based Long Short Term Memory[J]. *IEEE Transactions on Big Data*, 2022, 8(1): 241-252.
- [48] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. *Neural Computation*, 1997, 9(8): 1735-1780.
- [49] Vaswani A, Shazeer N, Parmar N, et al. Attention is all You Need[C]. *The 31st International Conference on Neural Information Processing Systems*, 2017: 6000-6010.
- [50] Cho K, van Merriënboer B, Gulcehre C, et al. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation[EB/OL]. 2014: arXiv: 1406.1078. <https://arxiv.org/abs/1406.1078>
- [51] Zou Z, Ge J G, Zheng H B, et al. Encrypted Traffic Classification with a Convolutional Long Short-Term Memory Neural Network[C]. *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems*, 2019: 329-334.
- [52] Bacquet C, Zincir-Heywood A N, Heywood M I. An investigation of multi-objective genetic algorithms for encrypted traffic identification[M]. *Computational Intelligence in Security for Information Systems*, 2009: 93-100.
- [53] Aceto G, Ciunzio D, Montieri A, et al. MIMETIC: Mobile Encrypted Traffic Classification Using Multimodal Deep Learning[J]. *Computer Networks*, 2019, 165: 106944.
- [54] Korczyński M, Duda A. Markov Chain Fingerprinting to Classify Encrypted Traffic[C]. *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, 2014: 781-789.
- [55] Freedman D. Markov chains. Springer Science & Business Media, 2012.
- [56] Shen M, Wei M W, Zhu L H, et al. Certificate-Aware Encrypted Traffic Classification Using Second-Order Markov Chain[C]. *2016 IEEE/ACM 24th International Symposium on Quality of Service*, 2016: 1-10.
- [57] Shen M, Wei M W, Zhu L H, et al. Classification of Encrypted Traffic with Second-Order Markov Chains and Application Attribute Bigrams[J]. *IEEE Transactions on Information Forensics and Security*, 2017, 12(8): 1830-1843.
- [58] Liu C, Cao Z G, Xiong G, et al. MaMPF: Encrypted Traffic Classification Based on Multi-Attribute Markov Probability Fingerprints[C]. *2018 IEEE/ACM 26th International Symposium on Quality of Service*, 2019: 1-10.
- [59] Chen Y G, Zang T N, Zhang Y Z, et al. Rethinking Encrypted Traffic Classification: A Multi-Attribute Associated Fingerprint Approach[C]. *2019 IEEE 27th International Conference on Network Protocols*, 2019: 1-11.
- [60] Liu C, He L T, Xiong G, et al. FS-Net: A Flow Sequence Network for Encrypted Traffic Classification[C]. *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, 2019: 1171-1179.
- [61] Fu Y J, Xiong H, Lu X J, et al. Service Usage Classification with Encrypted Internet Traffic in Mobile Messaging Apps[J]. *IEEE Transactions on Mobile Computing*, 2016, 15(11): 2851-2864.
- [62] Liu J M, Fu Y J, Ming J C, et al. Effective and Real-Time In-App Activity Analysis in Encrypted Internet Traffic Streams[C]. *The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017: 335-344.
- [63] Ho T K. The Random Subspace Method for Constructing Decision Forests[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, 20(8): 832-844.
- [64] Fu Y J, Liu J M, Li X L, et al. A Multi-Label Multi-View Learning Framework for In-App Service Usage Analysis[J]. *ACM Transactions on Intelligent Systems and Technology*, 2018, 9(4): 1-24.
- [65] Li D, Li W Z, Wang X L, et al. ActiveTracker: Uncovering the Trajectory of App Activities over Encrypted Internet Traffic Streams[C]. *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking*, 2019: 1-9.
- [66] Forecast G. Cisco visual networking index: global mobile data traffic forecast update, 2017–2022. Update, 2019, 2017: 2022.
- [67] Ran D B, Hadar O, Richman I, et al. Video Quality Representation Classification of Safari Encrypted DASH Streams[C]. *2016 Digital Media Industry & Academic Forum*, 2016: 213-216.
- [68] Orsolic I, Pevec D, Suznjovic M, et al. YouTube QoE Estimation Based on the Analysis of Encrypted Network Traffic Using Machine Learning[C]. *2016 IEEE Globecom Workshops*, 2017: 1-6.
- [69] Dimopoulos G, Leontiadis I, Barlet-Ros P, et al. Measuring Video QoE from Encrypted Traffic[C]. *The 2016 Internet Measurement Conference*, 2016: 513-526.
- [70] Shen M, Zhang J P, Xu K, et al. DeepQoE: Real-Time Measurement of Video QoE from Encrypted Traffic with Deep Learning[C]. *2020 IEEE/ACM 28th International Symposium on*

- Quality of Service*, 2020: 1-10.
- [71] Hintz A. Fingerprinting Websites Using Traffic Analysis[M]. Dingledine R, Syverson P., Eds. *Privacy Enhancing Technologies*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003: 171-178.
- [72] Bissias G D, Liberatore M, Jensen D, et al. Privacy Vulnerabilities in Encrypted HTTP Streams[C]. *The 5th international conference on Privacy Enhancing Technologies*, 2005: 1-11.
- [73] Liberatore M, Levine B N. Inferring the Source of Encrypted HTTP Connections[C]. *The 13th ACM conference on Computer and communications security*, 2006: 255-263.
- [74] Jaccard P. The Distribution of the Flora in the Alpine ZONE.1[J]. *New Phytologist*, 1912, 11(2): 37-50.
- [75] Berrar D. Bayes' Theorem and Naive Bayes Classifier[M]. *Encyclopedia of Bioinformatics and Computational Biology*. Amsterdam: Elsevier, 2019: 403-412.
- [76] Panchenko A, Niessen L, Zinnen A, et al. Website Fingerprinting in Onion Routing Based Anonymization Networks[C]. *The 10th annual ACM workshop on Privacy in the electronic society*, 2011: 103-114.
- [77] Zhu N F, Chen S Q, Sui H F, et al. Overview of Anonymous Communication[J]. *Computer Applications*, 2005, 25(11): 2475-2479. (朱娜斐, 陈松乔, 眭鸿飞, 等. 匿名通信概览[J]. *计算机应用*, 2005, 25(11): 2475-2479.)
- [78] Wang T, Cai X, Nithyanand R, et al. Effective Attacks and Provable Defenses for Website Fingerprinting[C]. *The 23rd USENIX conference on Security Symposium*, 2014: 143-157.
- [79] Hayes J, Danezis G. K-Fingerprinting: A Robust Scalable Website Fingerprinting Technique[C]. *The 25th USENIX Conference on Security Symposium*, 2016: 1187-1203.
- [80] Juarez M, Imani M, Perry M et al. Toward an efficient website fingerprinting defense[C]. *European Symposium on Research in Computer Security*, 2016: 27-46.
- [81] Wang T, Goldberg I. Walkie-Talkie: An Efficient Defense Against Passive Website Fingerprinting Attacks[C]. *The 26th USENIX Conference on Security Symposium*, 2017: 1375-1390.
- [82] Sirinam P, Imani M, Juarez M, et al. Deep Fingerprinting: Undermining Website Fingerprinting Defenses with Deep Learning[C]. *The 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018: 1928-1943.
- [83] Zhang Z Q, Kang C C, Xiong G, et al. Deep Forest with LRRS Feature for Fine-Grained Website Fingerprinting with Encrypted SSL/TLS[C]. *The 28th ACM International Conference on Information and Knowledge Management*, 2019: 851-860.
- [84] Zhou Z H, Feng J. Deep Forest[J]. *National Science Review*, 2019, 6(1): 74-86.
- [85] Ramezani A, Khajepour A, Siavoshani M J. On Multi-Session Website Fingerprinting over TLS Handshake[EB/OL]. 2020: arXiv: 2009.09284. <https://arxiv.org/abs/2009.09284>
- [86] 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.
- [87] 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.
- [88] Yao H P, Gao P C, Wang J J, et al. Capsule Network Assisted IoT Traffic Classification Mechanism for Smart Cities[J]. *IEEE Internet of Things Journal*, 2019, 6(5): 7515-7525.
- [89] Hwang R H, Peng M C, Huang C W. Detecting IoT Malicious Traffic Based on Autoencoder and Convolutional Neural Network[C]. *2019 IEEE Globecom Workshops*, 2020: 1-6.
- [90] Zeng Y, Qi Z H, Chen W C, et al. TEST: An End-to-End Network Traffic Classification System with Spatio-Temporal Features Extraction[C]. *2019 IEEE International Conference on Smart Cloud*, 2020: 131-136.
- [91] 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.
- [92] Anderson B, McGrew D. Identifying Encrypted Malware Traffic with Contextual Flow Data[C]. *The 2016 ACM Workshop on Artificial Intelligence and Security*, 2016: 35-46.
- [93] PACE. <https://www.ipoque.com/products/dpi-engine-rs-pace2>.
- [94] NBAR. https://www.cisco.com/c/en/us/products/collateral/ios-nx-os-software/network-based-application-recognition-nbar/qa_c67-697963.html.
- [95] nDPI. <https://www.ntop.org/products/deep-packet-inspection/ndpi/>.
- [96] De Sensi D, Danelutto M, and Deri L. Dpi over commodity hardware: implementation of a scalable framework using fastflow[J]. *Master's thesis*, Universit, 2012.
- [97] L7-filter. <http://l7-filter.sourceforge.net/README>.
- [98] Libprotoident. <https://research.wand.net.nz/software/libprotoident.php>.
- [99] Snort. <https://www.snort.org/>.
- [100] Snorby. <https://github.com/Snorby/snorby>.
- [101] Sguil. <https://bammv.github.io/sguil/index.html>.
- [102] Hyperscan. <https://www.hyperscan.io>.
- [103] Hughes C J. Single-Instruction Multiple-Data Execution[J]. *Synthesis Lectures on Computer Architecture*, 2015, 10(1): 1-121.
- [104] Suricata. <https://suricata-ids.org/>.
- [105] Bujlow T, Carela-Español V, Barlet-Ros P. Independent Comparison of Popular DPI Tools for Traffic Classification[J]. *Computer Networks*, 2015, 76: 75-89.
- [106] Albin E, Rowe N C. A Realistic Experimental Comparison of the Suricata and Snort Intrusion-Detection Systems[C]. *2012 26th International Conference on Advanced Information Networking and Applications Workshops*, 2012: 122-127.
- [107] Wang X, Hong Y, Chang H, et al. Hyperscan: A Fast Multi-Pattern Regex Matcher for Modern CPUs[C]. *The 16th USENIX Conference on Networked Systems Design and Implementation*, 2019: 631-648.
- [108] Nguyen T T T, Armitage G. A Survey of Techniques for Internet Traffic Classification Using Machine Learning[J]. *IEEE Communications Surveys & Tutorials*, 2008, 10(4): 56-76.
- [109] Moore A, Zuev D, Crogan M. Discriminators for use in flow-based classification. Technical Report, 2013, 1470-5559.
- [110] Williams N, Zander S, Armitage G. A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification[J]. *ACM SIGCOMM Computer*

- Communication Review*, 2006, 36(5): 5-16.
- [111] Murata T, Ishibuchi H. MOGA: Multi-Objective Genetic Algorithms[C]. *Proceedings of 1995 IEEE International Conference on Evolutionary Computation*, 2002: 289.
 - [112] Khatouni A S, Zincir-Heywood N. Integrating Machine Learning with Off-the-Shelf Traffic Flow Features for HTTP/HTTPS Traffic Classification[C]. *2019 IEEE Symposium on Computers and Communications*, 2020: 1-7.
 - [113] Karagiannis T, Papagiannaki K, Faloutsos M. BLINC: Multilevel Traffic Classification in the Dark[C]. *The 2005 conference on Applications, technologies, architectures, and protocols for computer communications*, 2005: 229-240.
 - [114] Iliofotou M, Pappu P, Faloutsos M, et al. Network Monitoring Using Traffic Dispersion Graphs (TDGS)[C]. *The 7th ACM SIGCOMM conference on Internet measurement*, 2007: 315-320.
 - [115] Bernaille L, Teixeira R, Akodkenou I, et al. Traffic Classification on the Fly[J]. *ACM SIGCOMM Computer Communication Review*, 2006, 36(2): 23-26.
 - [116] Alshammari R, Lichodziejewski P I, Heywood M, et al. Classifying SSH Encrypted Traffic with Minimum Packet Header Features Using Genetic Programming[C]. *The 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference: Late Breaking Papers*, 2009: 2539-2546.
 - [117] Yuan R X, Li Z, Guan X H, et al. An SVM-Based Machine Learning Method for Accurate Internet Traffic Classification[J]. *Information Systems Frontiers*, 2010, 12(2): 149-156.
 - [118] Noble W S. What is a Support Vector Machine [J]. *Nature Biotechnology*, 2006, 24(12): 1565-1567.
 - [119] Fu P, Guo L, Xiong G, et al. Classification research on ssl encrypted application[C]. *International Conference on Trustworthy Computing and Services*, 2012: 404-411.
 - [120] Quinlan J R. C4. 5: programs for machine learning. Elsevier, 2014.
 - [121] 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: 407-414.
 - [122] Altman N S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression[J]. *The American Statistician*, 1992, 46(3): 175-185.
 - [123] Sun G L, Chen T, Su Y Y, et al. Internet Traffic Classification Based on Incremental Support Vector Machines[J]. *Mobile Networks and Applications*, 2018, 23(4): 789-796.
 - [124] Soyulu T, Erdem O, Carus A, et al. Simple CART Based Real-Time Traffic Classification Engine on FPGAs[C]. *2017 International Conference on ReConfigurable Computing and FPGAs*, 2018: 1-8.
 - [125] Erman J, Mahanti A, Arlitt M, et al. Offline/Realtime Traffic Classification Using Semi-Supervised Learning[J]. *Performance Evaluation*, 2007, 64(9/10/11/12): 1194-1213.
 - [126] Wang Q, Wang C, Feng Z Y, et al. Review of K-Means Clustering Algorithm[J]. *Electronic Design Engineering*, 2012, 20(7): 21-24. (王千, 王成, 冯振元, 等. K-means 聚类算法研究综述[J]. *电子设计工程*, 2012, 20(7): 21-24.)
 - [127] Bernaille L, Teixeira R. Early recognition of encrypted applications[C]. *International Conference on Passive and Active Network Measurement*, 2007: 165-175.
 - [128] Reynolds D A. Gaussian Mixture Models[J]. *Encyclopedia of biometrics*, 2009, 741.
 - [129] Du Y, Zhang R H. Design of a Method for Encrypted P2P Traffic Identification Using K-Means Algorithm[J]. *Telecommunication Systems*, 2013, 53(1): 163-168.
 - [130] Bacquet C, Zincir-Heywood A N, Heywood M I. Genetic Optimization and Hierarchical Clustering Applied to Encrypted Traffic Identification[C]. *2011 IEEE Symposium on Computational Intelligence in Cyber Security*, 2011: 194-201.
 - [131] Zhang M, Zhang H, Zhang B, et al. Encrypted traffic classification based on an improved clustering algorithm[C]. *International Conference on Trustworthy Computing and Services*, 2012: 124-131.
 - [132] Haffner P, Sen S, Spatscheck O, et al. ACAS: Automated Construction of Application Signatures[C]. *The 2005 ACM SIGCOMM workshop on Mining network data*, 2005: 197-202.
 - [133] Cao Y, Miao Q G, Liu J C, et al. Advance and Prospects of AdaBoost Algorithm[J]. *Acta Automatica Sinica*, 2014, 39(6): 745-758.
 - [134] Gu Y, McCallum A, Towsley D. Detecting Anomalies in Network Traffic Using Maximum Entropy Estimation[C]. *The 5th ACM SIGCOMM conference on Internet measurement - IMC '05*, 2005: 32-32.
 - [135] Cohen W W. Repeated incremental pruning to produce error reduction[C]. *Machine Learning Proceedings of the Twelfth International Conference ML95*, 1995.
 - [136] LaValley M P. Logistic regression[J]. *Circulation*, 2008, 117(18): 2395-2399.
 - [137] Natekin A, Knoll A. Gradient Boosting Machines, a Tutorial[J]. *Frontiers in Neuroinformatics*, 2013, 7: 21.
 - [138] Alshammari R, Zincir-Heywood A N. An Investigation on the Identification of VoIP Traffic: Case Study on Gtalk and Skype[C]. *2010 International Conference on Network and Service Management*, 2011: 310-313.
 - [139] Alshammari R, Zincir-Heywood A N. Investigating Two Different Approaches for Encrypted Traffic Classification[C]. *2008 Sixth Annual Conference on Privacy, Security and Trust*, 2008: 156-166.
 - [140] De Montigny-Leboeuf A. Flow attributes for use in traffic characterization. Communications Research Centre Canada., 2005.
 - [141] Alshammari R, Zincir-Heywood A N. Can Encrypted Traffic be Identified without Port Numbers, IP Addresses and Payload Inspection [J]. *Computer Networks*, 2011, 55(6): 1326-1350.
 - [142] Yang Z C, Yang D Y, Dyer C, et al. Hierarchical Attention Networks for Document Classification[C]. *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016: 1480-1489.
 - [143] Alpay D. Reproducing kernel spaces and applications. Birkhäuser, 2012.
 - [144] Salman O, Elhajj I H, Chehab A, et al. A Multi-Level Internet Traffic Classifier Using Deep Learning[C]. *2018 9th International Conference on the Network of the Future*, 2019: 68-75.
 - [145] Taylor V F, Spolaor R, Conti M, et al. Robust Smartphone App Identification via Encrypted Network Traffic Analysis[J]. *IEEE Transactions on Information Forensics and Security*, 2018, 13(1): 63-78.



崔苏苏 于 2019 年在南昌大学信息安全专业获得学士学位。现在中国科学院信息工程研究所第六研究室攻读博士学位。研究领域为加密流量分析、入侵检测等。Email: cuisusu@iie.ac.cn



刘宝旭 于 2002 年在中国科学院研究生院获得博士学位。现任中国科学院信息工程研究所研究员, 第六研究室主任。研究领域为网络安全攻防对抗、网络安全测评技术等。Email: liubaoxu@iie.ac.cn



董聪 于 2017 年在天津大学信息管理与信息系统(保密方向)专业获得学士学位。现在中国科学院信息工程研究所第六研究室攻读博士学位。研究领域为网络安全态势感知、知识图谱等。Email: dongcong@iie.ac.cn



姜波 于 2016 年在中国科学院大学计算机系统结构专业获得博士学位。现任中国科学院信息工程研究所副研究员。研究领域为网络安全态势感知、知识图谱、数据挖掘等。Email: jiangbo@iie.ac.cn



卢志刚 于 2010 年在中国科学院研究生院获得博士学位。现任中国科学院信息工程研究所高级工程师, 中国科学院网络空间安全学院副教授。研究领域为网络安全态势感知、网络攻击检测、移动终端安全等。Email: luzhigang@iie.ac.cn



张辰 于 2016 年在中国地质大学(北京)获得硕士学位。现任中国科学院信息工程研究所工程师。研究领域为移动互联网安全、网络入侵检测等。Email: zchen@iie.ac.cn