互联网加密流量检测、分类与识别研究综述

陈子涵[1),2),3),4)] 程 光[1),2),3),4)] 徐子恒[1),2)] 徐珂雅[1),2)] 仇 星[1),2),3),4)] 钮丹丹[1),2)]

[1)](东南大学网络空间安全学院 南京 211189)

[2)](教育部计算机网络和信息集成重点实验室(东南大学) 南京 211189)

[3)](江苏省泛在网络安全工程研究中心(东南大学) 南京 211189)

[4)](网络空间国际治理研究基地(东南大学) 南京 211189)

摘 要 互联网流量分析是网络管理与安全的核心途径,传统基于明文的分析方法在加密流量大势所趋的环境下已基本失效.虽有部分针对加密流量的分析方法,但其忽略了不同加密流量分析目标需求内在的逻辑性与层次性,并缺乏对加密流量本质特征的研究,难以系统化地解决加密流量分析的难题.本文首先面向网络管理与安全监管的实际需求,将互联网加密流量分析按照目标需求划分为检测、分类、识别三个阶段,并描述其目标与方法上的差异;接着基于现有研究成果,分别对现有检测、分类、识别方法从多个粒度、角度进行划分,系统性地归纳与比较现有研究的优缺点;最后,本文基于目前研究,结合未来互联网网络环境发展趋势和加密流量概念漂移的实际问题,从加密流量样本数据集完善、复杂新型网络协议下的加密流量分类与识别、基于应用层特征的加密流量分类与识别、多点协同分布式加密流量分类与识别四个方面分析与展望了未来互联网加密流量检测、分类与识别中可能的研究方向.

关键词 互联网加密流量分析;加密流量检测;加密流量分类与识别;概念漂移;复杂新型网络协议

中图法分类号 TP393 **DOI 号** 10.11897/SP.J.1016.2023.01060

A Survey on Internet Encrypted Traffic Detection, Classification and Identification

CHEN Zi-Han[1),2),3),4)] CHENG Guang[1),2),3),4)] XU Zi-Heng[1),2)] XU Ke-Ya[1),2)]
QIU Xing[1),2),3),4)] NIU Dan-Dan[1),2)]

[1)](*School of Cyber Science and Engineering, Southeast University, Nanjing* 211189)

[2)](*Key Laboratory of Computer Network and Information Integration of Ministry of Education (Southeast University), Nanjing* 211189)

[3)](*Jiangsu Province Engineering Research Center of Security for Ubiquitous Network (Southeast University), Nanjing* 211189)

[4)](*International Governance Research Base of Cyberspace (Southeast University), Nanjing* 211189)

Abstract Network traffic measurement and analysis is an essential support for network security management and traffic engineering. With the continuous development of encrypted traffic technology, Internet traffic encryption has become an inevitable trend of Internet development. However, network traffic encryption brings privacy and security to users and enterprises and challenges network security protection and traffic management. Traditional traffic measurement and analysis methods such as Deep Packet Inspection (DPI) are not suitable for encrypted traffic environments, so it is of great significance to study encrypted traffic analysis on the Internet. At present, current research in encrypted traffic analysis is classified according to the classification

收稿日期:2021-11-09;在线发布日期:2022-09-30.本课题得到国家自然科学基金面上项目(62172093)、国家重点研发计划项目课题(2020YFB1804604)、2019 年工信部工业互联网创新发展工程项目(6709010003)资助.**陈子涵**,博士研究生,主要研究领域为网络空间安全、加密网络流量分析. E-mail: zhchen@njnet.edu.cn. **程 光**(通信作者),博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为网络安全、网络测量、加密流量识别、流量行为分析、主动防御. E-mail: chengguang@seu.edu.cn. **徐子恒**,硕士研究生,主要研究方向为网络空间安全、加密网络流量分析、网络空间治理. **徐珂雅**,硕士研究生,主要研究领域为网络空间安全、加密网络流量分析. **仇 星**,博士研究生,主要研究方向为网络空间安全、加密网络流量分析. **钮丹丹**,硕士研究生,主要研究领域为网络空间安全、网络流量分析.

method of encrypted traffic and its input or output. There is no unified standard of the granularity of encrypted traffic analysis or a systematic theoretical definition of it. Moreover, the inconsistency of concepts has brought troubles to the direction subdivision and work refinement in encrypted traffic analysis to some extent. Therefore, because of Internet traffic's characteristics and analysis requirements, this paper first divides Internet traffic analysis into three stages by definition: encrypted traffic detection, encrypted traffic classification, and encrypted traffic identification, and elaborates the characterization of these three stages from the perspective of users. Encrypted traffic detection refers to the process of screening out encrypted traffic from network traffic, which has nothing to do with the generalized application carried by the traffic, the generalized content transmitted by the traffic, and the rate of the traffic itself, but is only related to the nature of the traffic itself. Encrypted traffic classification represents the generalized application classification of encrypted traffic, which refers to classifying the generalized application carried by the encrypted traffic on the basis that the traffic is known as the encrypted traffic, which has nothing to do with the data transmitted by the traffic. According to the progressive granularity, the generalized application can be divided into service, application, and function. Encrypted traffic identification describes encrypted traffic data and metadata identification, which identifies the actual payload data, the user behavior, the QoE, and other metadata corresponding to the traffic on the premise that the traffic is encrypted traffic and the application type of the traffic is known. Then we analyze and compare the existing Internet encrypted traffic detection methods, classification methods, and identification methods from multiple perspectives and summarize their advantages and disadvantages, respectively. Finally, we combine the development trend of the Internet network environment in the future to analyze and outlook the possible research directions in the three stages of Internet encrypted traffic analysis, from the perspective of concept drift. We summarize the future research directions as encrypted traffic dataset perfection, encrypted traffic classification and identification under new complex network protocols (including TLS-1.3, encrypted DNS, HTTP-2.0, and QUIC), application layer feature based encrypted traffic classification and identification, and multi-point cooperative distributed encrypted traffic classification and identification.

Keywords　Internet encrypted traffic analysis; encrypted traffic detection; encrypted traffic classification and identification; concept drift; new complex network protocols

1　引　言

互联网技术的发展极大地丰富了人们的生活，产生了各种网络应用，而网络流量则是网络应用在网络交互过程中的数据表现形式.互联网服务提供商(Internet Service Provider，ISP)为了保证网络的服务质量(Quality of Service，QoS)和用户的体验效果(Quality of Experience，QoE)，需要对网络流量——尤其是 OTT(Over-the-top)应用的网络流量进行分析，这也是提升网络管理水平、实现网络资源优化的重要手段；同时，互联网技术的高速发展也使互联网上存在大量非法行为与恶意流量，安全监管部门需要通过流量分析技术实现互联网的监管和安全保护.

随着目前国内外网络安全法律法规的完善和用户网络安全意识的兴起，目前大量网页已经被加密.根据 NetMarketShare 的数据，2021 年 12 月三大主流鼓励使用 HTTPS 的浏览器市场占比已超过 90%，全球加密 Web 流量的比例也已经超过 90%[①]；根据 Google 透明度报告“Chrome 中的 HTTPS 加密情况”，2022 年 5 月 Chrome 加载网页中启用加密的比例已经达到了 99%[②].

① NetMarketShare. Market share for mobile, browsers, operating systems and search engines, url = https://netmarketshare.com/

② Google. HTTPS encryption on the web — Google Transparency Report, url = https://transparencyreport.google.com/https/overview

除网页外,隐私与版权意识的加强使得视频、音乐等内容与即时通信、文件传输等应用在网络中采用加密传输,而反作弊和服务器安全保护的需求也推动了电子游戏、远程访问等行为流量的加密化. Google 透明度报告"所有 Google 产品和服务中的已加密流量"比例已经在 2022 年 5 月几乎达到了 100%(除 Google Maps 与 Google News). 网络流量的加密化已经成为必然趋势.

但网络流量加密化在给用户和企业带来隐私与安全的同时,也给网络安全监管和网络流量管理带来了挑战. 网络监管和流量管理通常需要准确的流量分析,但流量经加密后难以通过传统的方法实现细粒度实时分析. 传统的流量分析方法依靠 DPI 方法对明文报文中的负载特殊字段进行匹配而实现[1],但随着加密流量的兴起,DPI 方法难以对被加密的负载内容进行匹配.

目前的 DPI 方法包含对密文字段匹配的中间盒(middlebox)分析方法[2],通过将 DPI 和密码学原理相结合,获取并过滤用户令牌[3](token)、提取并生成加密流量通信过程中的各个行为规则[4],实现在保护用户隐私情况下的加密流量检测(主要用于入侵检测). 值得注意的是,该过程需要较高的权限与访问方的配合,仅适用于企业内网边界,因为企业内网与外部网络进行交互都需要通过中间盒进行(串行流量模式). 该模式不适合在互联网中使用,因为互联网流量规模巨大(尤其是骨干网环境)、用户数量众多且用户行为不可控,流量分析需要通过旁路并接进行.

为了对加密流量进行分析以获取相关部门所需的管理依据,学术界逐渐产生了依靠流粒度特征及报文头部特征进行流量识别与检测的深度流检测(Deep Flow Inspection,DFI)方法和基于机器学习的流量检测方法[5].

目前加密流量分析领域现有的综述大多是按照加密流量的分类方法及其输入或输出进行分类的[5-6],该分类方式存在以下 2 种优点:(1)清楚地表明了有哪些可利用特征,利于后续研究的特征性质选择;(2)明确在整个领域内有哪些可用的技术,利于方法的选择. 而随着深度学习技术的发展,后续加密流量相关的综述[7]则更偏向于讨论与分析加密流量分类中使用的各类深度学习方法,并不关注整体加密流量分析的过程.

但是以上综述中按照加密流量分类方法及其输入或输出进行研究分类的架构存在以下 4 个问题:(1)混淆了加密流量分析中各个目标需求,未对不同类型、不同性质的加密流量分析目标加以区分;(2)忽略了加密流量分析目标需求的内在逻辑性、层次性,未考虑加密流量分析输入输出之间的关联关系与层次化特征;(3)未对加密流量分析方法与加密流量分析目标需求之间进行强关联,未考虑加密流量分析中不同方法在不同目标中的特异性;(4)以机器学习方法的更迭为整体逻辑主线,忽略了加密流量的本质特征,概念不明确. 以上 4 个问题将导致后续研究者在对加密流量分析领域进行研究的过程中,忽略不同层次研究在数据和逻辑上的递进关系,选择不合适的特征与适用性差的方法,难以梳理加密流量分析的本质逻辑,降低研究效率.

除此之外,概念漂移[8](concept drift)也是目前加密流量分析全生命周期中都需要面对的一个问题. 概念漂移是一个在机器学习领域很常见的一种现象,指的是目标样本的特征与类别随着时间的推移而持续发生改变的现象,该改变可能是渐进的,也可能是突发的. 互联网应用种类日新月异、应用版本不断更新、应用功能不断完善等会提高概念漂移产生的概率,同时互联网网络环境的多变同样会造成样本特征的概念漂移. 该问题是未来互联网加密流量分析领域重要的研究领域之一,目前的综述对该问题缺乏阐述.

在现有互联网加密流量分析领域的研究中,研究用词存在多种以最终目标为导向的行为描述,但概念用词与实际行为间具有非一一对应的关系. 以加密流量识别为例,在现有研究中,虽然对该概念的用词英文统一为 Identification(或 Identifying 代表动作),但其实际研究内容则包含加密流量的检测[9-12]、分类[1,13-18]与识别[19-20].

因此,本文对加密流量分析的各个层次与目标进行归纳与重整,将加密流量分析领域的工作按照目标需求,划分为三个阶段:加密流量检测(detection)、加密流量分类(classification)与加密流量识别(identification). 并在对三个阶段的研究进行归纳的基础之上,对贯穿于全生命周期的加密流量分析概念漂移现象进行了深入的分析与展望.

从互联网加密流量分析的全生命周期出发,将加密流量分析各领域之间的层次关系递进化,并以方法为支撑,系统化地构建加密流量分析的研究内容框架,如图 1 所示.

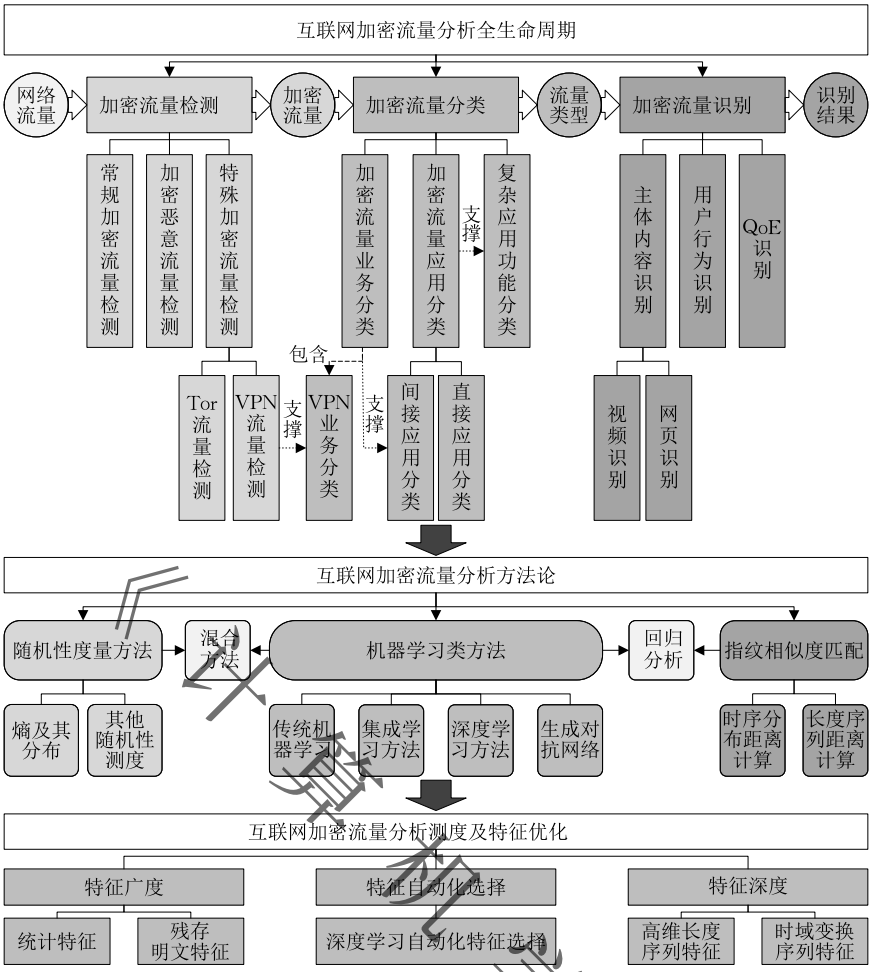

图 1　加密流量分析各层次之间关系

加密流量检测目前的研究主要着眼于从网络流量中分离出加密流量的常规加密流量检测与加密恶意流量检测的研究，在此基础之上，部分研究关注虚拟专用网(Virtual Private Network，VPN)、洋葱路由网络(The Onion Routing，Tor)等特殊加密流量的检测. 加密流量分类目前的研究工作包含业务、应用、功能三个粒度，业务分类包含常规加密流量下的业务分类与 VPN 流量下的业务分类，VPN 流量下的业务分类则需要 VPN 流量检测进行支撑；同时，应用分类包含直接分类与间接分类，业务分类为间接应用分类提供支撑，而功能分类则面向特定复杂加密应用，需要在应用分类的基础之上进行. 加密流量识别目前主要从加密流量负载的实际内容(目前主要研究对象为视频和网页内容)、代表的用户行为和隐含的用户 QoE 三个不同类型的数据或元数据角度进行研究.

本文第 1 节介绍互联网流量加密化的成因、发展现状和未来趋势，同时简要分析现有加密流量分析方法与综述整理架构上的不足；在此基础之上，按照目标需求对加密流量分析三个阶段进行划分；第 2 节从加密流量分析的第一个阶段——加密流量检测出发，分析目前互联网常规加密流量检测、加密恶意流量检测和特殊加密流量检测的研究现状；第 3 节按照三个逐层递进的粒度对目前互联网加密流量分类的研究现状和方法进行探讨与分析；第 4 节针对目前互联网加密流量识别，首先从视频内容和网页内容出发，对目前互联网加密流量内容识别现状进行介绍与分析；然后综述加密流量用户行为识别和加密流量 QoE 识别方面的研究；第 5 节描述目前加密流量分析领域面向概念漂移存在的四个主要问题，并对未来的研究方向进行展望；最后一节总结归纳全文综述情况.

2　互联网加密流量检测

虽然目前互联网流量已大多为加密流量，但是由于加密协议应用的不完全和网络系统协议更迭的不一致性，目前互联网中仍然存在部分明文流量；另

一方面,物联网云化控制的广泛发展,导致物联网常见的压缩流量也通过互联网进行传输.这一部分非加密流量的存在一方面降低了加密流量分析的准确度,另一方面降低了流量分析的效率.监管部门或ISP若想要实现有效而科学的网络流量分析与管理,则首先需要将网络流量中的加密流量与非加密流量进行分离,也就是对互联网加密流量进行检测,从而进一步实现恶意流量检测和特殊流量检测,也为加密流量分类提供纯净数据支撑.

检测是以二分类结果为目标的特殊分类行为,按照加密流量检测的实际应用需求,可以分为以下3种类别:(1)常规加密流量检测;(2)加密恶意流量检测;(3)特殊加密流量检测.

2.1 常规加密流量检测方法

常规加密流量指的是目前基于公开加密传输协议进行传输的,在互联网全部加密流量中占主要成分的加密流量形式.对常规加密流量的检测,可以有效地分离加密与非加密流量,为互联网OT下服务与应用的分析提供重要支撑.

加密作为一种通过提高数据的信息存量,降低其可预测性的技术,在不改变数据长度情况下,加密后数据的统计特征变化表征为数据随机性的增加.因此目前学术界主要利用加密与非加密流量间的数据随机性表征差异对加密流量进行直接检测.

信息熵(简称熵)是由Shannon于1948年提出的用于描述信源不确定度的概念,在统计学中,某一数据随机性的增加直接体现为其信息熵的增加.其计算公式如下,其中U代表随机变量,n代表变量单元可取值的数量,p_i代表第i种取值的变量单元的概率:

$$H(U)=-\sum_{i=1}^{n}p_i\log p_i \tag{1}$$

将网络流量报文的有效负载X视为有限长二进制串,则其熵:

$$H(X)=-(p_0\log p_0+p_1\log p_1) \tag{2}$$

其中p_0表示比特值为0的比特的频率,p_1表示比特值为1的比特的频率.

目前大部分互联网加密流量检测都是通过熵进行的,因为加密流量的熵高于文本明文流量和二进制明文流量,但熵存在以下问题:(1)流为非定长数据,熵难以准确评估不同流随机性差异的本质区别;(2)熵以整体结果为导向,忽略了数据自身的分布差异和随机性变化的趋势.

因此,Dorfinger等人[9]引入了N-截断熵(N-truncated entropy)的概念.N-截断熵指的是根据分布D生成的所有可能单元值的极大似然估计的平均值,它可以用作熵的无偏估计,当分布D为均匀分布U时,其公式如下:

$$H_N(U)=\frac{1}{m^N}\sum_{n_1+\cdots+n_m=N}\left[\binom{N}{n_1+\cdots+n_m}\times\left(\sum_{i=1}^{m}-\frac{n_i}{N}\log\frac{n_i}{N}\right)\right] \tag{3}$$

该研究通过将报文切分为字节长度的单元,利用加密与非加密流量字节单元在ASCII编码域中覆盖区间和分布的差异性实现实时检测.

NIST对数据的随机性测度检验规定了多种方法,Zhao等人[10]针对熵度量难以精细控制采集与计算时间、忽略数据随机性变化趋势的问题,提出了一种基于加权累加和检验(cumulative sums test)的时延自适应加密流量盲识别算法EIWCT.该方法可以在流量采集的过程中在线进行,可以实现采集与分析的精细化时延控制,但是该方法需要的样本数据总长度高于N-截断熵.

不过,互联网作为承载各类其他网络形式进行数据传输的核心网络,其中存在为保障网络QoS和部分设备能耗而产生的压缩流量.而在数据处理过程中,数据的压缩和加密都会导致源数据的信息熵的增加[9].如果不考虑压缩流量的存在直接通过随机性测度对加密流量进行检测,则结果有可能受到压缩流量的干扰,误报率较高.学术界在近年的研究中逐渐考虑到压缩流量的干扰性问题.Zhang等人[21]发现各种类型加密文件的熵,无论在报文粒度下还是流粒度下,其熵值都很高;但是对于压缩文件,如果使用流粒度计算其熵值则很少被判定为高熵,使用报文粒度则较多被判定为高熵.因为各类压缩文件中即便存在压缩算法的差异性,但其仍旧会保留非压缩的低熵明文头部和可能存在的明文低熵压缩字典,所以对压缩流量和加密流量通过随机性度量进行分离是可行的.但是值得注意的是,流量化压缩文件与压缩流量不同,压缩流量使用了带有压缩算法的网络协议,而流量化压缩文件仅仅是明文流量的负载被替换为一个压缩文件,导致与实际情况存在一定的差距.

因此,Casino等人[22]提出了一种基于多随机性测度的流量化加密文件、流量化压缩文件与流量化明文文件分离的方法HEDGE,该方法对NIST SP 800-22标准中规定的多种随机性测度检测方法进行了实验,最终选择块内频率检验(frequency within block test)、累加和检验和近似熵检验(approximate

entropy test)三种随机性测度作为三者分离的指标,有效地实现了流量化压缩文件与流量化加密文件的分离.

随着机器学习和深度学习技术的发展,学术界也产生了机器学习与随机性度量结合的混合方法.

Niu 等人[23]提出了一种启发式统计测试方法 HST(Heuristic Statistical Testing). 该研究发现,SSH、SSL 等加密协议由于明文握手的存在,整个流的前若干个报文部分的 N-截断熵要远小于后续的报文部分. 因此该研究从采集到的流量数据的尾部开始向前选取双向流数据,并通过单比特检验(monobit test)、分块比特频数检验、累加和检验和离散傅里叶变换检验(discrete fourier transform test)四种随机性测度对流量随机性进行度量,最终采用 C4.5 进行加密流量检测.

Tang 等人[24]则针对加密流量和非加密压缩流量需要区分才能进行后续加密流量分类的问题,提出了一种基于熵的特征提取算法,通过固定长度的滑动窗口将流量转化为熵向量特征,实现加密流量与压缩流量的分类. 虽然该研究的压缩流量是通过 FTP 协议传输的压缩文件,但是已经比流量化压缩文件更具说服力(加密流量为真实加密流量).

常规加密流量检测方法研究总结如表 1 所示. 总体而言,基于机器学习的混合方法具有更好的识别准确率,但是其模型训练需要消耗更多的计算开销和更多的样本数据,因而需要进一步研究以平衡成本与效果. 另一方面,目前学术界中针对压缩流量干扰加密流量检测的研究较少,但是随着低能耗物联网的发展,压缩流量的干扰将成为加密流量直接检测研究的重点方向之一.

表 1　常规加密流量检测方法总结

现有研究工作	年份	算法或特征	数据集	可区分压缩流量	检测所需数据量	优点	缺点
Dorfinger 等人[9]	2011	N-截断熵＋ASCII 编码	自采集数据	否	首报文	数据量小	适用范围窄、效果一般
Zhao 等人[10]	2013	EIWCT	自采集数据	否	1 KB	可在线进行、计算量小	数据量大、效果一般
Casino 等人[22]	2019	三种随机性测度	公开数据集＋自采集数据	可能可以	1 KB	可区分流量化压缩文件与加密文件	计算成本非常高、不实用
Niu 等人[23]	2019	四种随机性测度＋C4.5	自采集数据	否	128 Bytes	面向加密协议,效果较好	数据量大、需要较为完整的流
Tang 等人[24]	2019	SBE 熵向量＋SVM/RF	公开数据集＋自采集数据	是	1444 Bytes	可区分加密流量与压缩流量,效果好	特征提取计算量大

2.2　加密恶意流量检测方法

目前大量网络恶意行为通过私有加密协议逃脱监管部门的监察,给网络安全带来了新的挑战. 恶意流量作为网络安全问题的罪魁祸首,学术界对其检测存在较多研究,以弥补入侵检测系统(Intrusion Detection System,IDS)无法对加密恶意流量进行阻截的缺陷,同时也为进一步恶意应用的分类提供数据支撑[25].

随机性测度可以作为恶意流量检测的重要指标. Luo 等人[26]利用 N-截断熵指纹对僵尸网络 Nugache 进行了检测. 实验表明,C&C 流量作为一种特殊的已知加密流量,其 N-截断熵指纹具有特殊性.

机器学习方法同样被应用在加密恶意流量的检测中. Yu 等人[27]利用 Metasploit 渗透测试工具在自行搭建的环境中生成了基于各类渗透测试的恶意流量与普通访问的正常流量,通过多重自动编码器(Multi AutoEncoder,MAE)方法对恶意流量进行特征提取和检测. 进一步,为解决明文握手特征的多变问题,De Lucia 等人[28]通过对 TLS 流量的 record 长度和数量分布进行建模,使用 SVM 方法和 CNN 方法分别实现了加密恶意流量的检测. 随着深度学习方法的发展,Zeng 等人[29]通过 LSTM 与 CNN 相结合的混合深度学习模型,在 USTC-TFC 数据集下实现了恶意流量的检测. Jiang 等人[30]在 CICAndMal2017 数据集下提出了 HST-MHSA 模型,其利用 LSTM＋TextCNN 与多头注意力机制进行特征的提取,从而实现加密恶意流量的检测.

与常规加密流量检测的区别在于,恶意流量虽然已经拥有较多公开数据集样本,但其并没有统一而固定的协议格式. 随机性测度方法的效果受限于采集数据的长度,而机器学习方法受限于训练期间的样本规模和未来的概念漂移问题.

现有加密恶意流量检测方法总结如表 2 所示. 总体而言,其方法论与常规加密协议检测差距不大,主要差异在于特征的选择和数据的处理.

表 2 恶意流量检测方法总结

现有研究工作	年份	检测目标	方法类型	使用特征	方法	数据集	优点	缺点
Luo等人[26]	2018	僵尸网络C&C	随机性度量	密钥交换过程指纹的熵分布	多尺度正则分析	恶意样本沙盒数据	模型简单、计算量较小	适用面窄，必须获取密钥交换，受限于加密协议版本
Yu等人[27]	2019	恶意流量	深度学习	握手验证特征/报文长度及间隔	MAE	Metasploit沙盒生成数据	有限的自动化特征提取	计算量大、必须获取握手与验证过程
De Lucia等人[28]	2019		深度学习	双向 TLS record 长度数量分布	CNN	公开数据集	长度数量特征容易获得	计算量较大
Zeng等人[29]	2019		深度学习	自动化特征选择	TEST	USTC-TFC数据集	无需特征选择过程	计算量较大
Jiang等人[30]	2021		深度学习	LSTM+TextCNN多头注意力机制	HST-MHSA	CICAndMal2017数据集	特征选择具有可解释性，效果较好	计算量很大

2.3 特殊加密流量检测方法

VPN 作为一种有效保护用户隐私和安全的远程加密隧道访问技术，目前已经广泛地应用于互联网远程访问中. 但 VPN 的主要目的是为了安全的远程访问，而非抗追踪. 为了实现抗追踪的匿名远程访问，产生了洋葱路由技术，通过多层嵌套加密报文的多点顺序解密，保证每个网络节点只能知道上一个网络节点的网络位置. 而后便产生了 Tor 项目，使 Tor 技术真正应用于互联网环境. 两种特殊的加密流量由于其承载内容的特殊性（多为非法内容），其检测目前也已经成为学术界的研究重点.

2.3.1 VPN 流量检测

Sandvine 的 2020 年《全球互联网现象报告：COVID-19 聚焦》中表明，随着 COVID-19 疫情的发展，VPN 作为一种远程工作技术，其使用占比相较于之前有了大幅度的提高. 现今互联网中 VPN 形式主要为应用层 VPN. VPN 协议除公开协议外，同样存在大量私有 VPN 协议，这在一定程度上给 VPN 流量的检测带来了困难.

VPN 流量检测领域的研究主要通过机器学习方法实现. Draper-Gil 等人[31]于 2016 年公开了 ISCX VPN-nonVPN 数据集. 该数据集包含 7 类普通加密业务流量和 7 类 VPN 加密业务流量. 通过 C4.5 决策树和 kNN 方法进行了 VPN 流量的检测.

目前对 VPN 流量检测主要基于该研究公开的 ISCX VPN-nonVPN 数据集实现. Wang 等人[32]基于该数据集，使用了 1D-CNN 方法进行了 VPN 流量的检测. 而 Shapira 等人[33]将流的报文时序-大小分布作为输入特征，使用 CNN 方法进行 VPN 流量检测. Wang 等人[34]则利用 SSLVPN 的握手特征，使用了随机森林方法在自建的数据集上实现对 SSLVPN 的流量检测. Tang 等人[35]则独创性地提出了分段熵分布提高了 N-截断熵的序列随机性分布表达能力，结合胶囊神经网络（CapsNet）实现了 V2Ray VPN 的检测. 进一步，Luo 等人[36]使用随机森林基于 IP 代理行为与数据加密行为实现了 V2Ray VMess 协议流量的检测. Zhou 等人[37]则针对较为少见的 IPSec VPN，基于负载明文特征匹配结合基于随机森林（Random Forest，RF）的流统计特征的分类，实现 IPSec VPN 相较于非 VPN 流量的检测.

VPN 流量检测方法总结如表 3 所示. 但以上研究大多存在数据集老旧、数据集内样本分布不平衡、缺乏对 VPN 流量本身特性进行研究的问题. 因此，

表 3 VPN 流量检测方法总结

现有研究工作	检测目标	年份	方法类型	使用特征	方法	数据集	优点	缺点
Draper-Gil等人[31]	Open VPN	2016	机器学习	流时间相关特征	C4.5/KNN	ISCX VPN-nonVPN	开山之作，模型计算量小	特征较多、效果一般
Wang等人[32]		2017	深度学习	自动化特征选择	1D-CNN		无需特征选择	计算量大、效果一般
Shapira等人[33]		2021	深度学习	FlowPic	CNN		特征具有可解释性，效果较好	计算量较大，数据需求量大
Wang等人[34]	SSLVPN	2019	机器学习	流粒度统计特征	RF	实验环境自采集数据	通用性好	易过拟合
Tang等人[35]	V2Ray	2020	深度学习	分段熵分布	CapsNet		特征简单、需求数据量小	模型计算量大
Luo等人[36]		2021	机器学习	IP代理行为/数据加密行为	RF		考虑IP代理行为，针对性强	适用面窄，易过拟合
Zhou等人[37]	IPSec VPN	2021	匹配+机器学习	负载明文特征/流统计特征	RF		针对IPSec VPN	特征多，易过拟合

未来关于 VPN 流量检测的研究重点应该更加着眼于对 VPN 流量本身特性的研究，以实现在 VPN 协议种类不断增加情况下的 VPN 流量有效检测.

2.3.2　Tor 流量检测

Tor 流量相较于 VPN 流量而言具有多层嵌套报文加密的特性，且 Tor 融合了流量混淆技术（如 obfs4 技术），随机填充和时间序列的随机化使得难以将其从大量流量中区分，其流量特征更加类似于常规加密流量，从而其检测更加困难[38].

由于 Tor 的使用需要基于特定的浏览器，且大部分 Tor 访问是以网页浏览的形式进行的，所以 Tor 流量检测最开始的研究主要是依靠网页指纹对 Tor 网页流量进行检测. 但指纹类特征过拟合性强，泛用性差，因此部分研究着眼于使用统计特征进行检测.

Wang 等人[39]针对 Tor 网页提出了利用网页指纹进行高准确率攻击的方法，其使用 kNN 分类器结合大规模特征集（包含通用统计特征、独特报文长度、报文顺序、出站报文聚合度、Brust、初始报文长度等）实现 Tor 网页流量的检测. 可以在一定防御策略的情况下实现较好的准确率，是具有里程碑意义的研究. Hayes 等人[40]则在之前研究的基础之上，进一步挖掘指纹特征的表达能力，提出了 k-fingerprinting 方法，使用 RF 进行检测，并对不同的指纹特征进行了显著性评估.

Jia 等人[41]改进了决策树分类算法，并且使用协同训练算法（tri-training）用来在应用层对识别出的匿名流量进行划分. 该研究利用报文长度的熵、600 字节报文出现的频率、前 10 报文中空数据报文出现的频率和平均报文间隔时间四个特征进行了 Tor 流量的检测. 除监督学习方法外，无监督学习聚类方法同样被用于 Tor 流量检测. Rao 等人[12]针对 Tor 流量与普通流量的差异性，提出了引力聚类算法. 该方法主要通过预先设定 Tor 流量的样本，而后通过新样本与旧样本之间的距离进行聚类，更类似于半监督学习方法，利用了 Tor 流量之间自身的相似性. 随着深度学习的发展，Kim 等人[42]针对 Tor 只支持加密连接和 TCP 流的特性，提出了通过 16 进制原始数据头特征，利用 CNN 进行 Tor 流量检测的方法. 随着各类新型匿名网络技术的产生，Hu 等人[43]通过层次化分类的模式实现了 Tor 流量的检测，并且在此基础之上实现了不同匿名网络间（Tor、Freenet、I2P、Zeronet）的检测和部分进一步的分类.

随着流量混淆技术的发展，部分研究从 Tor 协议和 Tor 网络的工程实现角度出发，研究 Tor 混淆后流量的检测. He 等人[38]针对 Tor 流量常用的 obfs4 混淆技术，提出了采用粗粒度快速过滤和细粒度流量检测相结合的混淆化 Tor 流量检测方法. 该研究利用了 obfs4 协议的高熵负载特性和严格认证机制，首先通过对负载进行随机性检测和报文方向时间序列检测，进行 obfs4 流量的初筛；而后利用 SVM 方法基于多项统计特征对使用 obfs4 协议混淆的 Tor 流量进行检测. 除了 obfs4 外，还有一种常用的混淆协议 Meek. Yao 等人[11]针对 Meek 协议，提出了一种基于高斯混合马尔科夫模型的 Tor 流量检测方法. 该研究通过使用高斯混合函数来表征报文间时间间隔和大小的密度分布，结合隐马尔科夫模型，计算流量观测序列和检测出 Tor 流量的可能性，最终通过时间间隔与大小的二维观测结果进行 Tor 流量的检测. Bai 等人[44]针对目前境内登录 Tor 网络多使用混淆技术的现状，采用 CNN 模型对目前用的最多的 obfs4 协议混淆的 Tor 流量进行了检测.

Tor 流量检测方法总结如表 4 所示. 总体而言，

表 4　Tor 流量检测方法总结

现有研究工作	年份	使用特征	方法	数据集	优点	缺点
Wang 等人[39]	2014	六类特征	kNN	实验环境自采集数据	模型简单、效果较好，可应对一部分防御策略	特征数量巨大，特征提取计算量大
Hayes 等人[40]	2016	k-fingerprinting	RF		在相同的防御策略下效果更好、指纹可解释	更大的特征数量，模型复杂，特征提取计算量大
Jia 等人[41]	2017	报文长度及统计特征	决策树＋协同训练	实验环境自生成数据	特征数量较少，协同训练提高准确率	特征计算量大，易过拟合
Rao 等人[12]	2018	报文及流粒度特征	引力聚类算法	ExperimenTor 生成	利用相似性聚类，可应对未知 Tor 样本	准确性一般，需要具有代表性的聚类中心样本
Kim 等人[42]	2018	16 进制报文头特征	CNN	UNB-CIC 数据集	引入深度学习，需求数据量小	无法应对概念漂移
Yao 等人[11]	2018	报文间隔/大小分布	HMM	实验环境自采集数据	可检测 Meek 混淆	易过拟合
He 等人[38]	2019	熵与报文统计特征	SVM		可检测 obfs4 混淆	用到了协议自身特性，难以应对概念漂移
Hu 等人[43]	2020	26 种统计特征	层次化分类模型		可以在检测的基础上进行进一步的类内细分类	特征数量较多、部分匿名网络并非严格加密
Bai 等人[44]	2021	自动化特征选择	CNN	实验环境自生成数据	可检测 obfs4 混淆，自动化特征选择	模型复杂、特征不可解释

该领域的研究疏于对特定流量本身特征的研究，缺乏普适性，未来的研究应该更注重于实际流量的研究，以实现实际网络环境中的 Tor 流量检测.

3 互联网加密流量分类

业务(又称服务，service)指的是与用户行为相关的、具有明确领域性的、人为划分的、正交的、粗粒度的应用集合，如在线视频、音乐、文件传输、VoIP、网页流量、即时通信(Instant Message，IM)、电子邮件等；应用指的是隶属于某一业务、以机构组织或企业或个人为提供者、能够独立满足用户需求的软件实体，一个业务类别下包含多个同类应用，而应用必须属于且仅属于某一个业务，且每种应用根据其主要业务类型隶属于某个特定业务，如在线视频业务下的 YouTube、Netflix、优酷、爱奇艺等；功能则指的是隶属于应用、能够与用户行为相对应的细粒度应用能力单元，功能与应用之间构成多对多关系，且一个功能可能会包含多种用户行为，如微信下的图文聊天、视频通话、转账、红包等. 互联网加密流量分类面向的是业务、应用和功能.

由于目前加密流量承载的数据适用业务广泛、应用种类多样、实际功能复杂，因此十分有必要对互联网加密流量进行进一步分析，从而实现细粒度的网络流量管理与安全监管. 在加密流量分析的研究中，加密流量分类是当前其中的重点领域. 加密流量分类主要着眼于对互联网中的各类加密业务或应用进行分类，从而针对性的进行网络管理与安全监管，同时也存在部分研究针对于包含多种功能的复杂加密应用内的功能进行分类的研究.

3.1 加密流量业务分类方法

加密流量分类包含多种粒度，业务作为应用的人为归纳划分类别，随着长期的发展已经相对固定，对加密流量按照业务进行分类则可以实现更加普适而通用的网络流量管理. 目前学术界对加密流量业务分类的研究主要为常规加密流量，也有部分研究关注 VPN 加密流量下的业务分类.

常规加密流量业务分类是在常规的互联网加密流量中进行加密流量的业务粒度下的分类. 随着机器学习技术的逐步发展，集成学习分类模型首先被应用于加密流量业务分类.

Bagui 等人[45]在 Draper-Gil 等人[31]的基础之上使用了梯度提升树(Gradient Boosting Trees，GBTs)和 RF，利用流量时间相关特征进行了业务分类. Houser 等人[46]从目前互联网已存在的加密 DNS 出发，提出了利用加密 DNS 特征差异的业务分类方法.

但是机器学习方法相较于 DPI 而言，分类成本大幅度提高. 因此，Yang 等人[1]提出了一种 DPI 方法与机器学习方法相结合的流量分析方法. 利用 DPI 方法降低机器学习方法的工作负担，并使用机器学习方法实现 DPI 无法实现的加密流量分类.

传统机器学习的方法目前而言缺乏在开放世界环境下的实验验证，存在特征工程的前置条件. 随着神经网络及深度学习技术的发展，深度学习也被应用于加密流量业务分类之中，目前深度学习方法大多使用 ISCX VPN-nonVPN 数据集进行实验.

Yang 等人[47]以 TLS 加密流量的明文握手部分中的 cipher suite、compression method 和 TLS extension information 为分类输入，使用贝叶斯神经网络实现分类. Wang 等人[32]提出了利用深度学习的端到端学习特性(自动化特征选择)，使用 1D-CNN 进行加密流量业务分类的方法. Lotfollahi 等人[48]则进一步提出了一种融合 CNN 和 SAE 模型的 Deep Packet 深度学习框架. Song 等人[49]针对普通 CNN 方法忽略加密流量层次结构的问题，提出了使用文本卷积神经网络 Text-CNN 进行分类的方法. 其同时引入了新的损失函数和一个合适的类别权重分配方法实现分类的平滑化处理. Yao 等人[13]也使用 attention-based LSTM 和 HAN 模型进行了业务分类(除了检测外). Cui 等人[50]则进一步考虑到加密流量中报文之间不仅存在一个前后逻辑关系，而且其关系具有方向性的特性，提出了使用基于向量神经元的胶囊神经网络(Capsule Neural Networks，CapsNet)进行加密流量业务分类的方法. Wang 等人[51]则充分发挥新型神经网络模型残差网络 ResNet 的原始流量特征提取优势和 AutoEncoder 的统计特征编码优势，将两者结合提出了 CENTIME 模型实现分类. 随着多分类器方法的应用，Aceto 等人[52]提出了 DISTILLER 分类器，以 CNN 和 MLP 为基分类器，实现多任务分类下的架构优化.

除了基模型的叠加使用外，部分深度学习方法考虑模型的进一步融合，以进一步发挥深度学习自动化特征选择的优势. Hu 等人[53]提出了 CLD-Net 模型，该模型是将 CNN 模型、LSTM 模型和全连接网络进行串接而成的. Lu 等人[54]则将 CNN 模型的一种变体 Inception CNN 模型与 LSTM 模型并列结合，提出了 ICLSTM 模型. Xie 等人[55]针对 DL 方法是不可解释黑盒的问题，将 Self-attention 机制引入加密流量业务分类，优化自动化特征选择，结合 CNN 构建模型 SAM，比 CNN 基模型效果更好. 而

Ren 等人[56]则结合数据结构中的树结构的优势，把若干个 RNN 用二叉树的形式构建成为 Tree-RNN，有效地提高了分类效率.

深度学习方法虽然具有自动化特征选择的特性，但是并非代表该特征选择一定为较优特征选择，且该特征的选择不具有可解释性；同时，深度学习方法更加受制于数据集本身，数据集的规模不足与样本不平衡都将导致分类准确率下降.

因此部分研究从特征选择出发，优化深度学习在加密流量业务分类中的效果.

时间序列特征或时域变换序列特征是目前加密流量业务分类中常用的一类特征，因为不同的互联网业务对 QoE 的要求不同，这一现象可以体现在报文时间间隔和报文密度上. 时间序列特征一般指报文序列中包含相对时间的序列性特征，如报文相对到达时间序列(Inter-Arrival Time，IAT)、报文时间间隔序列(Packet Interval Time，PIT)等. 以上特征作为时间序列可以在常规域下进行表征，也可以在变换域下进行表征. 变换域下的时间序列特征被称之为时域变换序列特征.

Shapira 等人[33]使用 FlowPic 结合 CNN 进行了加密流量业务分类. Nigmatullin 等人[57]则提出了叫做 AGMV(Accumulated Generalized Mean Value)的特征优化方法，代替原本的时间序列特征. Chen 等人[58]针对目前主流的 TLS 加密流量，利用其 TLS segment 特征，基于 Deep LSTM 模型构建 SMC 系统实现业务分类. 其首先离线训练出一堆属于不同业务的分类器，然后在在线分类阶段由每个分类器分别计算每个 segment 属于某一个业务的概率，从而得到最有可能的分类结果. Chen 等人[59]则融合流时序特征、报文头部特征和负载特征，使用 Attention-LSTM 进行特征提取，而后将提取的特征输入 CNN 进行分类，以构建 Attention-CNN 模型. Huoh 等人[60]则考虑到报文之间可以构成图的特殊性，将流按照报文为节点，报文时空关系为边进行了图化，进而使用基础的图神经网络(GNN)实现了分类. Chen 等人[61]开创性地提出了使用协议数据单元(Protocol Data Unit，PDU)特征结合 *N*-gram 模型与双向双层 GRU 构建长度敏感(length sensitive)特征选择模型，并进一步利用胶囊神经网络模型有效表达 PDU 长度序列间的马尔科夫性特征，提出了 LS-CapsNet 模型，实现了加密流量的业务分类，效果优于各类最新方法. 进一步，Chen 等人[62]兼顾突发性分类需求和持续性分类需求的两种实际需要，将 LS-CapsNet 理论化、系统化为长度敏感复合深度学习模型架构 LSCDL，并在此基础之上进一步提出了 LS-LSTM 模型，实现了快速化、高精准的加密流量业务分类，是现有研究中综合能力最好的模型.

部分研究选择从数据集本身出发以解决数据集不平衡问题，提高分类效果下限. Soleymanpour 等人[63]针对加密流量数据集不平衡造成的小类别误分类问题，提出了使用成本敏感(cost-sensitive)计算方法将原有均匀分布的误分类损失调整为“小类别-大类别分类”损失更低的分布，从而使分类方法适应数据集的样本不平衡问题，提高分类准确率. Ma 等人[64]则从原始报文输入上考虑，提出了 traffic reconstruction 方法对报文头与负载进行了重新拼接，以提高 1D-CNN 模型的分类准确率，但是其中用到了 IP 头和 TCP 头的参数，存在过拟合的可能性.

随着疫情的发展，VPN 流量占比持续走高，这带来了更加急切的、关于 VPN 流量深入分析的需求. VPN 加密流量业务分类则是考虑在从加密流量中检测出 VPN 流量的基础之上，对 VPN 流量内进行的业务进行分类. 其核心问题在于 VPN 隧道将原本的多个加密流混合成一整个加密流，从而增加了加密流量分类的难度.

由于大部分加密流量业务分类的研究基于 ISCX VPN-nonVPN 数据集，其在进行加密流量业务分类实验的同时，也进行了 VPN 流量业务分类的实验验证. 仅存在少量纯粹针对 VPN 流量进行业务分类的研究. Shi 等人[18]研究了 OpenVPN 下的视频流式传输业务的提取(业务二分类)，采用 RF 方法在自行抓取的 OpenVPN 流量中实现了视频业务的分类.

加密流量业务分类方法总结如表 5 所示. 总体而言，2016 年的 ISCX VPN-nonVPN 数据集存在数据老旧、规模不足、样本分布不平衡的问题，存在部分研究在不改变数据集本身的情况下通过各类方法进行分类准确率的优化，但是随着互联网业务与应用的迅速更迭，需要更加新颖的数据集支撑未来的实验. 而从方法论角度看，目前的加密流量业务分类领域的研究一方面着眼于机器学习尤其是深度学习新方法的应用和已使用模型的优化，另一方面则从特征工程的角度出发挖掘加密流量业务分类更适合的流量特征.

对常规加密流量业务分类的方法和 VPN 流量业务分类的方法进行对比可以发现，在目前的研究中对两类研究对象的解决方案大多相同，缺乏对两类流量本身的研究和其间差异性的研究. 因此，未来关于加密流量业务分类的研究首先需要更加新颖的数据集，其次则需要更多关注常规流量与 VPN 流量本身的特性，而非单纯的将常规加密流量业务分类的方

表 5 加密流量业务分类方法总结

现有研究工作	年份	方法适用环境	使用特征	分类模型	数据集	主要创新点	优点	缺点
Bagui等人[45]	2017	常规加密流量/VPN流量	流粒度时间相关特征	GBTs/RF	ISCX VPN-non VPN	机器学习方法创新	率先使用集成学习	特征数量较多
Wang等人[32]			自动化特征选择	1D-CNN		机器学习方法创新	率先使用深度学习	效果一般
Shi等人[18]	2018	VPN流量	CFS算法进行特征选择	RF	实验环境自采集数据	两个RF分类器组合识别	两个RF组合，针对VPN	特征与分类计算成本都高
Yang等人[1]	2019	常规加密流量	FCBF算法进行特征选择	DPI+朴素贝叶斯网络		DPI与机器学习方法的结合	混合方法	适用面窄
Yang等人[47]			三种TLS参数	贝叶斯神经网络		特征与方法的创新	利用TLS握手参数	适用面窄
Houser等人[46]			加密DNS请求特征	RF		加密DNS环境与相关特征	利用DNS请求特征	适用面窄
Song等人[49]		常规加密流量/VPN流量	报文的逐字节值	Text-CNN		特征与方法的创新	比CNN更能表达序列关系	效果一般
Yao等人[13]			自动化特征选择	LSTM/HAN		机器学习方法创新	RNN比CNN效果更好	效果仍一般
Cui等人[50]			会话特征与报文特征	CapsNet		机器学习方法创新	考虑到更高层面的特征	模型计算量大
Shapira等人[33]			FlowPic	CNN		报文长度-时序分布FlowPic	特征具有可解释性	需要较多的数据输入
Lotfollahi等人[48]			自动化特征选择	Deep Packet		机器学习方法创新	只需要首报文	适用面极窄，现网难以使用
Soleymanpour等人[63]	2020	常规加密流量	自动化特征选择	CNN		成本敏感计算平衡类别损失	成本敏感计算应对不平衡类	效果一般
Chen等人[61]			多PDU N-gram长度序列特征	LS-CapsNet		多PDU长度序列与GRU N-gram特征提取	特征简单、无需样本对齐	存在一定精度牺牲
Wang等人[51]			原始流量/统计特征	CENTIME		ResNet+AutoEncoder	两类特征的融合更加全面	模型计算量大
Aceto等人[52]			报文头/报文负载	CNN/MLP		DISTILLER架构	多任务分类优化	效果一般
Huoh等人[60]		常规加密流量/VPN流量	报文为节点、时空关系为边	GNN		流的报文粒度图化与图神经网络	率先使用图神经网络	图结构不合理，效果一般
Ma等人[64]			自动化特征选择	CNN		Traffic reconstruction	输入报文重新拼接	方法老旧、效果一般
Chen等人[58]	2021	常规加密流量	TLS segment	Deep LSTM		多分类器概率计算	Segment比报文特征更显著	分类器太多，前期训练成本极高
Hu等人[53]			自动化特征选择	CLD-Net	ISCX VPN-non VPN	CNN+LSTM+全连接网络	效果胜于单个模型	计算量大幅度增加
Lu等人[54]				ICLSTM		Inception CNN与LSTM并列	效果胜于单个模型	计算量大幅度增加
Xie等人[55]				SAM		Self-attention机制结合CNN	自注意力机制可解释	效果提升不明显
Ren等人[56]				Tree-RNN		二叉树构建的多个RNN分类器	二叉树形式构建多分类器，提高效率	效果提升不明显
Nigmatullin等人[57]			AGMV优化的时间序列特征	CNN		AGMV模型	时间序列特征优化	效果一般
Chen等人[59]			Attention-LSTM提取	CNN		Attention-CNN模型	模型融合，效果优于并列	计算复杂度高
Chen等人[62]			多PDU N-gram长度序列特征	LS-CapsNet LS-LSTM	Cernet自采集数据集	LSCDL架构	适用于多种实际需求场景、特征简单、效果好	模型相对复杂

法应用于VPN流量样本数据集下获取分类结果.

3.2 加密流量应用分类方法

随着互联网应用的膨胀式增长和精细化网络管理与安全监管需求的出现，以业务为粒度的加密流量分类方法已经不能满足ISP和监管部门的需求.

加密流量的应用作为基于加密流量业务的更细

粒度，目前学术界的研究包含从加密流量中直接进行加密流量应用分类和基于业务类型进行间接加密流量应用分类两种.

3.2.1　直接加密流量应用分类

直接加密流量应用分类不考虑加密流量业务类别，而是直接对加密流量进行分类以获取应用类型. 直接加密流量应用分类的研究目前主要包括以下3类：(1)通过引入或提出更复杂或适用性更好的分类模型，实现模型的优化；(2)通过更加复杂的特征选择与评估过程，实现特征的优化；(3)通过完善或优化样本数据集，提高分类的效果.

对于模型优化而言，传统机器学习方法与集成学习方法被率先使用. 受限于机器学习方法表达能力的限制，这类方法多关注于残存的明文特征. 此类明文特征主要包含 TLS 握手中的残存明文字段信息和明文侧信道 DNS 信息. 在 TLS 1.2 版本及以前，TLS 握手中的证书、SNI 等信息可以为加密流量应用分类提供一定的帮助.

Muehlstein 等人[65]则使用 SVM-RBF 方法进行分类，实现了 HTTPS 加密流量环境下 6 种浏览器和其中 8 种应用的分类. Pan 等人[66]提出了一个集成多个马尔科夫链分类器进行分类的架构 WENC，将二阶马尔科夫链与隐马尔科夫模型相结合，使特征通过多个分类器加权集成得到判别结果.

在深度学习方法中，Al-Obaidy 等人[14]使用 MLP 方法实现了 5 种社交媒体加密应用的分类. Zhang 等人[67]针对单分类器在多分类需求上存在局限的问题，提出了一种对每一个分类类别构建一个 LSTM+3D-CNN 分类器的立体变换神经网络 STNN 模型，实现了 17 种校园网下 TLS/SSL 应用的分类. Aceto 等人[68]提出使用 CNN 和 LTSM 进行移动端加密流量应用分类的方法. 但由于深度学习具有非常大的模型训练成本与更新成本，因而作者提出利用分布式学习的思路，采用多个节点进行分布式神经网络训练. Chen 等人[69]将增量学习的方法论引入了加密流量应用分类，基于 GRU 模型为每一个应用类别单独构建其二元分类器，以实现 OvR(One vs Rest)分类器的累计. 在这种架构下，当有一种新的应用产生，只需要对该新应用单独训练 OvR 分类器即可. 不过如果新应用与旧应用十分相似，可能会产生较为严重的误分类. Nascita 等人[70]则针对深度学习分类模型可解释性缺乏的问题，引入可解释人工智能技术 XAI，在一个现有模型的基础上提出了 MIMETIC-ENHANCED 模型，在 MIRAGE-2019 数据集下提高了分类的效果. Li 等人[71]则针对 0day 应用(未知应用)，提出了 CNN 和 k-means 相结合的方法，使用 CNN 分类已知应用并使用 k-means 聚类未知应用.

除了模型优化外，较多研究着眼于更好的进行加密应用流量的复杂特征工程以提高分类效果. Shapira 等人[33]的 FlowPic 同样可以应用于加密流量应用分类，但是其只是单纯的使用了两种不同的数据集(分别对应于业务分类与应用分类)，并没有从本质上研究两者的差异性. He 等人[16]针对 Android 移动端上的 50 种不同的谷歌商店加密应用，利用了该平台移动应用流量的三种平台独特特征实现分类，但难以推广到其他移动端平台以及常规互联网环境中.

大量研究表明加密应用报文具有时序性状态转移特征，因此马尔科夫链等序列模型被应用于加密应用流量的特征优化提取中(而非分类[66]). Korczyński 等人[72]首次提出了使用一阶马尔科夫模型结合 SSL/TLS 流量头部中的信息类型参数(message type)，进行不同加密应用流量刻画的方法 FoSM. 虽然该特征不够稳定，可能会随着服务器配置、应用自身特性、协议的不合理使用或差异性的协议实现而产生变化，但是该方法开创了针对加密流量应用分类时序关联特征选择的领域，具有重要的意义. Shen 等人[73]则在其基础之上，提出了 SOB 方法，即将加密应用 TLS 握手中的证书报文长度和第一个 Application Data 的大小作为特征. Liu 等人[74]则进一步提出了一种长度基于感知快速傅里叶变换指纹的加密流量分类方法 LaFFT，该方法通过 FFT 方法对流量各个维度进行离散傅里叶变换计算以提取每个维度的模量和角度组成实值特征. 接着其[75]提出了一种利用马尔科夫链对流量全报文长度特征进行刻画的方法 MaMPF. 该方法利用信息类型序列 MTS 和长度块序列 LBS，使用幂律分布函数(power-law distribution)对该应用的“长度-数量”分布进行拟合以实现分类. 然后该团队[76]进一步挖掘加密流量的长度序列性特征，结合表示学习，提出了基于多层双向 GRU 模型的分类模型 FS-Net.

后续的研究则更加关注加密流量长度序列中的高维长度序列特征. 高维长度序列特征指的是在报文长度序列特征基础之上，通过对序列中报文长度值的进一步计算或重构获得的新的长度序列特征. 最典型的就是对报文序列的 PDU 重构，使用更高层协议的 PDU 作为输入的单元，消除报文分片带来的误差.

Shen 等人[77]则进一步将流内各个报文间的马尔科夫性关系具象化为图，提出了针对上下行 Brust

流量的图结构抽象 TIG,并基于 MLP 提出了 Graph-DApp 模型,对去中心化应用的加密流量进行了分类. Wu 等人[78]则基于 TLS 流双向 Application Data 的长度,提出了基于 LSTM 的 TFSN 模型,对 20 个自采集的主流 APP 进行了分类. Aceto 等人[79]针对 MIRAGE-2019 的 5 种 APP,使用马尔科夫链和隐马尔科夫模型实现了应用的分类预测.

而在对加密流量数据集完善的研究中,Wang 等人[80]利用 GAN 进行加密流量的应用分类,提出 FlowGAN 方法,直接使用 GAN 进行不平衡加密流量样本的填充. 该方法生成的样本相较于反复抽样的方法而言并不存在过采样的问题,不过该研究并没有深入地探索如何才可以更好地使用 GAN 进行样本生成,也没有研究样本平衡性与加密流量应用分类效果之间的具体关系,具有一定的深入研究价值.

直接加密流量应用分类方法总结如表 6 所示.

表 6 直接加密流量应用分类方法总结

现有研究工作	研究重点	年份	使用特征	分类/评价方法	数据集	主要创新点	优点	缺点
Muehlstein 等人[65]	机器学习分类模型优化	2017	报文基线特征+流粒度附加统计特征	SVM-RBF	实验环境自采集数据	特征复杂、效果一般	多粒度应用分类与复合特征使用	使用机器学习
Pan 等人[66]		2017	TLS握手报文消息类型-报文长度对和数据报文的负载长度	WENC		关联性特征使用与集成判别器的加权权重优化	长度特征相对简单,使用集成学习	模型简单、效果一般
Al-Obaidy 等人[14]	深度学习分类模型优化	2019	流粒度统计特征	MLP		加密主流社交媒体应用分类	使用统计特征,适用面宽	效果一般
Zhang 等人[67]		2019	流粒度统计特征	STNN	Lashkari 提供的数据	为每一个类别单独建立复合分类器	复合分类器比单模型效果好	计算量大,效果一般
Aceto 等人[68]		2019	自动化特征选择	CNN/LTSM	ISP 数据集	利用分布式神经网络训练提高模型训练速度	分布式神经网络训练更快	效果一般
Chen 等人[69]		2021	报文长度序列特征	GRU	实验环境自采集数据	基于 GRU 给每一类构建 OvR 分类器	长度特征与 OvR 结构	分类器过多,不实用
Nascita 等人[70]		2021	自动化特征选择	MIMETIC-ENHANCED	MIRAGE-2019	可解释人工智能 XAI	过程具有可解释性	计算量较大
Li 等人[71]		2021	流/报文统计特征	CNN-*k*-means	CIC-Darknet2020	支持已知与未知应用混合情况下的分类	支持未知应用混合分类	模型简单效果一般
Korczyński 等人[72]	特征工程优化	2014	FoSM	无	实验环境自采集数据	首篇针对报文马尔科夫性关系的研究	开创 TLS 特征的深入研究	论文老旧,协议已发生变化
Shen 等人[73]		2017	SOB	Bisecting *K*-Means	AppScanner 数据集	利用报文间的长度特征的马尔科夫性	使用 TLS 长度	模型弱、效果一般
He 等人[16]		2017	伴生明文DNS流量+报文粒度统计特征	特征距离计算与聚类	实验环境自采集数据	加密流量伴生流量特征聚类	效果有提升	适用面极窄
Liu 等人[74]		2018	长度感知快速傅里叶变换指纹 LaFFT	LR/RF		对报文长度特征进行 DFT 变换	效果有提升	适用面窄
Liu 等人[75]		2018	MaMPF	GBTs/RF		幂律分布函数对报文长度-数量分布进行拟合	效果有提升	适用面窄
Liu 等人[76]		2019	报文长度序列特征	FS-Net		引入表示学习,只需每个报文的长度数据	仅用长度特征、分类效果较好	需要完整流,数据需求量大
Shen 等人[77]		2021	TIG	GraphDApp		使用 burst 长度特征构建图,使用图神经网络	使用图化特征	适用面窄
Wu 等人[78]		2021	TLS流双向 Application Data 长度	TFSN		使用聚合后的TLS特征而非原始报文长度	使用 TLS 长度特征	模型复杂、效果一般
Aceto 等人[79]		2021	PL、DIR、IAT、MS、IMT	MC/HMM	MIRAGE-2019	使用MC/HMM进行特征提取以实现预测	预测性研究	特征复杂不实用
Wang 等人[80]	数据集完善	2019	自动化特征选择	FlowGAN	NIMS 公开数据集	利用GAN进行不平衡加密流量样本填充	GAN 样本填充	适用面窄

直接加密流量应用分类在加密流量应用分类中占主要部分，其主要研究点在于分类模型的更迭和特征工程的优化，少量研究关注加密流量样本数据不完善下的分类问题.

3.2.2　间接加密流量应用分类

间接加密流量应用分类则是在加密流量业务分类的基础之上，对某一种特定的业务领域内的应用进行细分. 目前该领域的研究集中于视频与 VoIP，旨在为 ISP 中 OTT 的视频和 VoIP 流量以进行细粒度流量调度和管理.

VoIP 业务中，Alshammari 等人[15]针对 3 种不同的 VoIP 加密应用分别使用了 C5.0 决策树、AdaBoost 和 GP 方法进行了加密流量应用分类. 其主要创新点在于使用了自组织映射(Self-Organizing Map，SOM)对大量不同的数据集进行聚类，该过程相当于对多个数据集进行融合与提纯，以获取更适合训练模型的数据集.

而视频业务中，Li 等人[17]针对 YouTube 加密视频流量提出了一种名为 Silhouette 的实时、轻量级加密视频应用分类方法对 YouTube 加密视频应用进行了二分类操作. 该方法的优势在于可以同时应用于 TCP 和 UDP 环境，但是通过阈值等易变特征进行判定将存在较大的概念漂移问题，因此在非轻量级需求下机器学习方法更加合适. Shi 等人[18]则在 OpenVPN 视频流量业务分类的基础之上进行了视频供应商的分类，该研究通过预分类和正式分类两轮基于 RF 方法的分类实现应用分类. 接着，Shi 等人[81]针对 6 个不同视频平台应用，使用了三个自然语言处理(NLP)领域常用的模型(HAN、CNN、GRU)，以前三秒分别的报文数量、比特数量与双向流统计特征，实现了视频业务下的加密流量应用分类. 间接加密流量应用分类总结如表 7 所示.

表 7　间接加密流量应用分类方法总结

现有研究工作	年份	业务	分类特征	分类方法	优点	缺点
Alshammari 等人[15]	2015	VoIP	流粒度统计特征	C5.0/AdaBoost/GP	SOM 数据集提纯，提高准确性	研究较老，应用已大幅度更新
Li 等人[17]	2018	视频	速率/长度/数量阈值	阈值判断	轻量级，计算速度极快	准确性低、难以应对概念漂移
Shi 等人[18]	2018	视频	CFS 算法选择	RF	使用预分类，提高准确性	适用面窄，易过拟合
Shi 等人[81]	2021	视频	报文/比特数/统计特征	HAN/CNN/GRU	特征简单	模型复杂、计算量高

总体而言，目前关于加密流量应用分类的研究主要分布于直接加密流量应用分类，仅有少量以视频和 VoIP 为业务研究领域的间接加密流量应用分类研究. 但随着加密应用规模的逐渐扩大和互联网的持续发展，未来加密流量应用分类将成为一个重点研究领域和行业需求. 因此，未来在该领域的研究应该更加着眼于对加密应用流量自身特性的深入研究而不仅仅是分类方法或特征选择的优化.

3.3　面向复杂加密应用的功能分类方法

随着互联网应用的发展，目前大型互联网企业的应用都存在着复杂化趋势，导致一个复杂的加密应用内可能会包含多种功能，因而产生了面向复杂加密应用的功能分类的需求.

目前在该领域的研究主要以目前全球范围内使用人数最为广泛的多功能 IM 软件微信(WeChat)为研究对象，对其内部的多种功能进行了实证性研究.

Hou 等人[82]针对微信应用的多种功能进行了研究. 首先该研究对微信使用的应用层加密协议 MMTLS 进行了分析，接着将微信内的功能划分为聊天、文件传输、朋友圈、支付、公众号、小程序和广告七种，利用各类功能流量之间长度、时间分布等统计特征差异性，利用 RF 实现了分类. Yan 等人[83]则进一步针对微信下的红包功能与转账功能进行了研究. 首先该研究基于微信中四类功能(文本、图片、红包和转账)的流量报文长度与时间序列分布的差异性，选择了流量的整体统计特征、报文长度区间、TCP 握手数量和报文出入站数量统计四类特征，采用 RF 方法实现了分类.

面向复杂加密应用的功能分类方法对比如表 8 所示. 总体而言，目前在面向复杂加密应用的功能分类领域的研究都是以微信等 IM 应用为研究对象，该类软件由于用户粘性则更易成为软件提供商进行功能复杂化的目标. 但是目前的研究依然存在数据集规模小、分类目标功能数量少、特征选择较为简单、分类方法较为传统的问题. 未来功能分类的研究应着眼于更多应用，并将目前其他粒度下的分类方法与特征推广到该粒度中.

表 8　面向复杂加密应用的功能分类方法总结

现有研究工作	年份	分类功能目标	分类特征	分类方法	数据集	优点	缺点
Hou 等人[82]	2018	微信 7 种功能	时序统计特征	RF	自采集数据	分类类别较多	协议版本依赖性强
Yan 等人[83]	2018	微信 4 种功能	统计特征与握手数量	RF		细分红包和转账	使用特征较多，易过拟合

4 互联网加密流量识别

在对加密流量应用进行分类的基础之上，想要进一步实现精细化的网络管理与安全监管，则需要对加密网络流量的数据及元数据进行识别，即识别加密流量负载中传输的实际内容、该内容所关联的用户行为或 QoE 等.

4.1 加密流量主体内容识别

根据 Sandvine 机构的《2019 年全球互联网现象报告》(2019 Global Internet Phenomena Report)，视频流量占比已达到全球总下行流量的 60.6%，位居榜首，而其上行流量占比也已经达到了 22.2%. 而网页流量作为 B/S 架构中的主要流量承载形式，其重要和普及程度也随着应用内嵌浏览器技术、H5 技术等技术的发展而逐渐提高. 报告中显示 Web 流量下行占比 13.1%，上行占比 10.3%位列第二.

因此，视频流量和网页流量作为目前互联网中主要两种流量形式，其内容识别对于精细化网络测量与安全监管而言十分重要.

4.1.1 加密视频流量内容识别

近年来随着视频编码技术、加密通信技术和视频供应商版权意识的兴起，出现了多种新型视频加密编码传输技术[84]. 视频加密传输给视频内容的审查与监管带来了巨大的困难，因而产生了针对特定视频应用中加密视频内容识别的研究.

Reed 等人[19]发现视频供应商为保证自身视频 QoS 和 QoE 保障，在视频加密传输之前使用可变比特率编码(Variable Bitrate Encoding，VBR)技术首先对视频进行编码分块. 作者引入了视频段大小的概念，利用不同视频在传输过程中“视频段大小-时间”的分布差异性，实现了对 Netflix 视频应用下不同视频内容的识别. 随后，Reed 等人[20]在前期研究的基础之上，套用了应用数据单元(Application Data Unit，ADU)的概念代替先前研究使用的视频段进行 Netflix 视频应用的内容识别. 在 Netflix 样本中，其互差异性比例达到几乎接近 100%.

Schuster 等人[85]通过研究发现，MPEG 视频流标准将导致视频流量产生内容相关的 burst 特征，从而可以利用该特征进行加密视频流量的内容识别. 但是该方法的限制在于需要准确抓到待测数据的 burst 特征，即待测目标不能存在噪音干扰和多视频同时观看的特性，其次视频服务器与视频传输链路上不能出现拥塞，否则可能会造成方法的失效. Gu 等人[86-87]考虑到了不同视频质量下的相同视频应被识别为同一视频的问题. 提出了利用差分视频序列指纹代替常规视频序列指纹的方法. 作者首先引入了 DTW 方法进行视频内容的识别，并在后续研究中提出了 P-DTW 方法实现对指纹序列中子序列的匹配来实现视频内容的识别. 但是该方法同样受限于无干扰、无噪音、带宽充足的环境条件. Wu 等人[88]则从近年国际较为火爆的 Facebook 视频出发，利用加密视频分段的特点，从密文数据中最大程度地还原加密视频片段的大小以构建视频指纹，而后进一步利用序列匹配算法将密文特征与构建好的明文特征库进行匹配，以识别待匹配加密视频的内容. 实验表明，文中提出的 HHTF 方法可以在 20 万级拟真指纹库下，使用超过 5 个连续的视频片段即可达到几乎 100%的准确率.

加密视频流量内容识别方法总结如表 9 所示.

表 9 加密视频流量内容识别方法总结

现有研究工作	视频平台	年份	使用特征	方法	主要创新点	优点	缺点
Reed等人[19-20]	Netflix	2016/2017	“视频段/ADU大小-时间”的分布差异性	皮尔森相关系数计算/kd-Tree匹配	发现了视频内容的差异性分段特征，以进行识别	该领域最早的研究，特征简单、方法易用	效果一般，特征易被干扰
Schuster等人[85]	Netflix YouTube	2017	基于流时间与报文长度的差异性 burst 特征	CNN	发现了 MPEG 的视频内容差异性特征	利用流式视频加载的特性，泛用性好	抗干扰能力差，实际使用限制多
Gu等人[86-87]	自采集视频样本	2018/2019	流量差分序列特征	DTW/P-DTW	考虑不同质量下的同一视频，使用了差分序列特征替代原序列	可以对不同质量下的同一视频进行识别	受限于无干扰、无噪音、带宽充足的环境
Wu等人[88]	Facebook	2021	视频片段 ADU 长度序列	HHTF	大规模指纹库下的 TLS 视频识别	大规模指纹库下的高精准率	实际使用需要构建巨大样本库

总体而言，目前加密视频流量的内容识别主要利用 VBR 等视频流量中特有的实现方式所造成的视频流量内容相关的差异性特征. 但是不同平台拥有自己独特的协议甚至存在一定的私有协议，结合实际网络环境中的网络环境差异与流量干扰，目前难以通过一种具体的方法解决所有平台下的视频内

容识别.随着用户非正常视频观看行为的引入和复杂新型网络协议的使用,视频内容识别的难度将进一步提高.因此未来该领域的研究应该更加着眼于如何从实际网络环境中提取视频流量的指纹以实现精准的视频内容识别.

4.1.2　加密网页流量内容识别

学术界对加密网页的研究主要着眼于加密网页的指纹差异性特征,而这一特征可以被有效地用于加密网页流量的内容识别,即识别出当前加密网页流量(包括普通加密网页和 Tor 网页)所对应的具体网页.

深度学习方法由于其大样本下的高精准度特性,同样被应用于加密网页流量内容识别. Sirinam 等人[89]针对 Tor 网络环境下的网页流量存在多种抗分析保护措施的现状,提出了一种可以针对 WTF-PAD 防御措施的深度学习加密网页流量内容识别方法,该方法考虑到 WTF-PAD 方法改变了攻击方(分析方)采集到的网页报文长度,因此只选取报文方向的序列作为双向流的特征,而后引入 SDAE (Stacked Denoising AutoEncoders)和 CNN 方法实现网页内容识别.而 Zhang 等人[90]则针对普通 HTTPS 加密网页,提出了 LRRS(Local Request and Response Sequence)特征进行网页指纹的构建.该特征包含流量总体出入站报文的数量及长度特征、pcap 分片后的各个时间片出入站报文数量及长度特征与前 20 个报文的出入站数量及长度特征.虽然该方法能够进行细粒度的网页识别,但是其所需要提取的特征多达 150 个,识别效率低. Shen 等人[91]考虑到网页中请求是主导因素,提出了使用上行链路报文长度累加和序列 U0 sequence 构建的网页指纹 WPF,在自采集的按照 IP 来整理的京东和雅虎的数据集下实现了加密网页的识别. Wang 等人[92]针对 Tor 网页提出了一种自适应指纹(adaptive fingerprinting)的特征,利用迁移学习方法和 adversarial domain adaption 技术实现网页识别. Shen 等人[93]在前期工作的基础之上,进一步使用了 CNN 基模型来构建 BurNet 模型,训练开放世界下的大规模数据集(每个网站统计 70 个以上的网页,采集 10000 条以上流),在包含有 2 个超过 6000 个网页样本集中测试,以实现加密流量网页识别.

加密网页流量内容识别方法总结如表 10 所示.总体而言,虽然目前部分方法已经实现了对 Alexas 上高排名网页的有效识别,但是仍然存在方法单一、特征简单、数据需求量大的问题,且 Tor 网页识别的研究多针对于网页识别的反制.因此,未来加密网页流量内容识别的研究应更加关注利用网页流量的统计特征差异性等实现较为稳定和普适的加密网页流量内容识别.

表 10　加密网页流量内容识别方法总结

现有研究工作	年份	使用特征	方法	数据集	主要创新点	优点	缺点
Sirinam 等人[89]	2018	报文的方向序列特征	SDAE/CNN	Alexas 自采集数据集	针对 WTF-PAD 防御方法也有着较好准确率	特征十分简单,可应对部分防御策略	效果一般、无法应对概念漂移
Zhang 等人[90]	2019	出入站报文长度与数量特征	Deep Forest	Alexas 自采集数据集	普通网页内容识别,多粒度特征选择	针对 HTTPS 网页	特征数量多,识别效率低
Shen 等人[91]	2020	上行链路报文长度累加和序列 U0 sequence	RF	实验环境自采集数据	只关注上行报文,大幅度提高识别速度	识别速度快	大规模下适用性较差
Wang 等人[92]	2021	adaptive fingerprinting	Triplet Network	4 个 Tor 网页数据集	迁移学习、小样本识别	针对 Tor 网页,支持迁移学习	模型较为复杂,适用面窄
Shen 等人[93]	2021	自动化特征选择	BurNet	实验环境自采集数据	大规模开放世界样本下的网页识别	大规模网页下具有较好的识别效果	特征不可解释

4.2　加密流量用户行为识别

加密流量由应用产生,除少部分自发流量外,其根本产生原因为各类用户行为,而用户的行为则与应用的功能直接相关.如果想要实现用户级高精度网络安全监管,则需要从加密流量中识别出用户的行为.

Conti 等人[94]发现不同手机应用和不同用户行为产生的报文长度-时间序列存在差异性,通过 RF 方法识别加密应用下的用户行为. Muehlstein 等人[65]则在对加密应用分类的基础之上,使用 SVM-RBF 方法进一步对浏览器使用行为进行了识别. Wu 等人[95]针对目前欧美互联网主流社交分享软件 Instagram 提出了一种多维特征提取方法 SFIM,有效地过滤出流量中不同行为的稳定特征,并进一步提出划分稳定特征分布范围的方法,最终实现 Instagram 下的 10 种不同用户行为的识别.

加密流量用户行为识别主要方法如表 11 所示.总体而言,目前学术界针对加密流量用户行为识别的研究很少,主要原因在于该领域识别粒度较细,且大部分应用用户行为受限于应用提供的功能,而功

表 11 加密流量用户行为识别方法总结

现有研究工作	年份	目标应用(功能)	使用特征	使用方法	优点	缺点
Conti 等人[94]	2016	七种常见应用	报文长度-时间序列	RF	特征简单,方法泛用性好	效果一般,应用已经大幅度更新
Muehlstein 等人[65]	2017	浏览器	报文级基线特征+流粒度统计特征	SVM-RBF	可以识别浏览器使用行为	浏览器使用行为的应用价值较低
Wu 等人[95]	2021	Instagram	EADU 特征	SVM/RF	ADU 特征比报文特征更显著	使用应用特殊机制只适用于 Instagram

能的合法性则在大部分情况下保证了行为的合法性.但在部分特殊需求下(如推荐、行为监控、用户画像[96]),加密流量用户行为识别则具有重要意义.因此,该领域未来研究应从需求出发,着眼于对特定应用及其功能下的用户行为识别.

4.3 加密视频流量 QoE 识别

目前关于加密流量 QoE 的识别主要集中于加密视频流量领域.各大视频供应商在保障加密视频功能满足用户需要的同时,也需要尽量地保障用户视频观看的体验,以维持其市场竞争力.另一方面,对视频帧率、码率等 QoE 相关参数的识别,可以大幅度提高对目标的物联网设备分析能力.

QoE 作为一种主观性因素,目前学术界对此并未产生统一而定量的定义;且受制于用户隐私保护,视频供应商无法直接在视频播放客户端上进行 QoE 采集,只能通过流量表征数据对当前加密视频的 QoE 进行判断.Khokhar 等人[97]将视频流量 QoS 映射为 QoE,最后通过 RF 方法实现对当前加密视频流量 QoE 的分类与评估.不过用 QoS 进行 QoE 映射的方法受限于视频供应商自身的设备部署情况,随着时间的推进将存在较严重的概念漂移,需要经常对分类器进行更新.Wassermann 等人[98]则面向互联网加密视频存在多分辨率的现状,提出了通过加密视频流量分辨率识别进行 YouTube 加密视频 QoE 识别与预测的方法.该研究基于时序分布特征,利用集成学习方法进行识别.

Tang 等人[99]对 YouTube 加密 HAS 视频流量 QoE 中的卡顿现象进行了研究,提出了混合 RNN 模型与 HMM 模型构成的 attention-based RNN-HMM 方法 A-RNN-HMM,通过 RNN 进行编码,HMM 进行解码实现.该研究的优势在于仅需要该视频流量的下载速率,即可实现视频卡顿现象的识别.Xu 等人[100]研发了一种名为块序列推断的框架,能根据报文大小和时序信息推断 HTTPS 和 QUIC 下的视频分辨率自适应切换行为.Wu 等人[101]则针对新型 HTTP-2.0 协议下的加密视频分辨率进行了识别,面向其多路复用特性,提出了 H2CI 方法,使用 CNN 对分辨率进行了识别.

加密流量 QoE 识别方法总结如表 12 所示.总体而言,鉴于视频业务及应用的特殊性,目前加密流量 QoE 识别领域的研究都是针对加密视频流量的 QoE 进行识别.未来该领域的研究应该着眼于更多的视频应用(目前主要为 YouTube 视频)和更多的 QoE 元素.

表 12 加密流量 QoE 识别方法总结

现有研究工作	年份	映射源	识别元素	使用方法	优点	缺点
Khokhar 等人[97]	2019	QoS	通用 QoE	RF	全面考虑网络状况,适用性广	特征复杂、时效适用性短
Wassermann 等人[98]		时序分布	分辨率	RF	特征简单	适用性较窄
Tang 等人[99]		下载速率	卡顿现象	A-RNN-HMM	特征简单且易于获取,可实时识别	模型相对复杂、仅适用于卡顿
Xu 等人[100]	2020	报文大小与时序信息	分辨率切换	CSI	可适用于 QUIC	适用性窄
Wu 等人[101]	2021	报文长度序列	分辨率	CNN	可适用于 HTTP-2.0	适用性较窄

5 未来研究方向展望

目前互联网加密流量分析领域的研究,大多使用了机器学习方法进行检测、分类与识别.机器学习方法的效果不可避免地与学习的样本相关,而随着时间的变化,实际网络环境中的流量数据特征表征与之前用于训练模型的样本会存在差异,也就是概念漂移.

目前在加密流量概念漂移中主要存在以下 4 个具体问题:(1)加密流量样本分布不均衡将给加密流量分析带来多种形式的概念漂移,导致方法普适性下降;(2)新型加密协议或应用协议将给互联网加密流量带来全新的样本形式,会使协议相关的分

析方法完全失效；(3) 基于非应用层特征的加密流量分析方法将给加密流量分析带来系统性概念漂移和误差；(4) 由于硬件处理能力的局限性和应用发展的快速性，单点难以处理大规模加密流量并覆盖足够应用. 本节将对以上四个问题所对应的研究方向进行分析与展望.

5.1　加密流量样本数据集完善

统计分析方法和机器学习方法都依赖样本数据，通过从样本数据中归纳有效的特征. 但是对样本的依赖则会使加密流量分析的效果受制于样本，即受制于样本的体量及样本的覆盖类别，不平衡的样本则会导致分类结果产生偏差；体量小的样本则难以提取具有普适性的类别特性(出现过拟合现象). 目前对互联网加密流量分析的研究主要着眼于公开数据集和自采集数据，都存在样本体量不足和样本不平衡的问题，其中样本体量不足可以通过多次样本采集和数据集混合进行弥补，但是某数据集的样本分布不平衡则成为目前加密流量分析领域研究的掣肘之一.

生成对抗网络 GAN 与传统只针对判别器(discriminator)的机器学习方法不同，该模型通过判别器对生成器(generator)进行评价，从而实现对生成器的训练，使之可以生成“以假乱真”的样本. 目前 GAN 相关方法已经被用于加密流量分析领域，其主要应用在于对加密流量样本的补充，以解决加密流量样本分布不平衡的问题. 也有研究更加深入的针对加密流量分析中深度学习方法输入的特征维度或数量太小的问题，利用对特征维度的扩展[102]进行不平衡数据的平衡化.

但是 GAN“治标不治本”，虽然生成的样本可以有效地解决数据集不平衡类别的样本量问题，但是其无法生成出与真实样本一致的新样本，新样本的分布情况也是与原有样本一致的. 而特征池的扩大也仅限于基础统计特征的维度扩展，对于复杂特征而言并不适用.

因此，未来应该着眼于主攻自动化加密流量样本采集与标签方法，提高加密流量样本采集与构建的效率，以应对层出不穷的新应用与易变的网络环境特征.

5.2　复杂新型网络协议下加密流量分类与识别

随着互联网加密技术和应用技术的发展，产生了大量复杂新型加密通信协议(如 TLS-1.3、加密 DNS)和复杂新型应用协议(HTTP-2.0、QUIC)，两者统称为复杂新型网络协议. 此类协议的出现给互联网加密流量分类与识别带来了新的挑战.

根据 Censys 的统计结果，2016 年使用 TLS 协议的设备占比为 29.4%，而 2018 年已经增长到 73.7%，加密流量 TLS 协议的使用已经成为大势所趋[103]. 目前 TLS-1.3 在 TLS 协议中的占比逐渐提高，已经在少部分地区或网络通信领域超越 TLS-1.2 成为更加主流的加密通信协议. 由于 TLS-1.2 为之前互联网中最主流的加密通信协议，大量的研究以 TLS-1.2 标签样本数据集为训练集，TLS-1.3 协议的更新将给加密流量分类与识别带来巨大影响. TLS-1.3 的低时延特性使之可以更好地应用于同时需要高可靠性和低时延要求的 HTTPS 加密流量业务中[104]，将逐渐替代 TLS-1.2 协议成为新的主流加密网络通信协议. 可预见的，TLS-1.3 协议对 Server Hello 之后的所有握手信息进行加密使目前基于 TLS-1.2 握手明文信息的加密流量分析方法失效，且其他协议特性同样对目前基于 TLS-1.2 流量进行训练的分类器产生概念漂移. 因此，未来加密流量分析的研究应更多地考虑 TLS-1.3 协议下的加密流量.

DNS 协议作为互联网站点访问的重要支撑协议，基于 UDP 协议以明文形式传输. 但随着 DNS 劫持等攻击的增加和加密网络通信技术的发展，学术界开始针对加密 DNS 协议进行研究. 目前最主流的是 DNS-over-TLS(DoT)协议[46]. DNS 加密带来的概念漂移会使传统基于伴生明文 DNS 进行加密流量识别的方法完全失效. 且 DoT 流量表征为 TLS 流量，也给 TLS 流量的分类识别带来了巨大的概念漂移. 虽然 DNS 流量的加密化和长度填充使传统分析手段无法使用，但是加密 DNS 流量本身可以看作以用户行为为锚定的报文时间序列，所以其存在继续可以被用作加密流量分类与识别的可能性. 因此，未来如何在加密 DNS 的环境下进行加密流量的分析是亟待研究的问题.

HTTP-2.0 协议是一种使用 HTTP 头部压缩和 TCP 多路复用传输的新型应用层协议，目前主要用于网页的传输，但也逐渐被各个视频平台所使用，结合 TLS 协议形成 HTTPS 协议传输加密视频. HTTP-2.0 协议的二进制传输模式使应用层头部和数据部分分开传输，则其报文顺序可能受到影响；TCP 多路复用则导致一个 TCP 流中可能会包含了若干个 HTTP 流，且其失去了顺序传输的性质，增添了多个 HTTP 请求响应(及其数据)交替传输的特性，会同时影响一条 TCP 流的流粒度特征和报文

粒度特征;而应用层头部压缩则会导致相同数据在不同次访问的时候产生不一样长度的报文,对报文长度相关的统计特征产生影响.以上特性甚至会在同一个应用的多次访问内造成概念漂移.因此,未来需要更加着眼于 HTTP-2.0 对加密流量产生的影响,研究 HTTP-2.0 环境下的加密流量检测、分类与识别.

QUIC[105]作为一种 Google 提出的自带加密与顺序确定的特殊应用层协议,目前已经广泛应用于其旗下各类互联网服务中.鉴于其在视频传输方面的独特优势,该协议已经被各类常用浏览器所支持,并且已经被部分国内视频服务提供商应用.可预见的,QUIC 协议在成为 RFC 标准并正式更名为 HTTP/3 协议之后,其不受限于 TCP 内核难以修改及更新缓慢的优势将使其逐渐提高互联网协议覆盖率.QUIC 协议虽然是可以被独立解析的公开协议,但是由于其自身的高版本迭代率(频繁更新)和其包含的多路复用、高握手加密性等特性,将给 QUIC 分类器带来巨大而及时的概念漂移,部分情况下分类器可用周期较短.因此未来对 QUIC 协议的研究应该更加着眼于 QUIC 加密流量的特殊性,从而为 QUIC 加密流量下的应用分类和内容识别提供支撑.

5.3 基于应用层特征的加密流量分类与识别

除利用加密流量中伴生明文流量进行加密流量分析外,目前的加密流量分析方法主要基于流量的网络层及传输层特征.但是此类方法将面临以下问题:各层互相解耦的网络协议拥有自己的行为范式和数据格式,使同一应用数据在不同协议层级间的流量行为表征具有差异性,导致网络层/传输层不能准确刻画应用层信息,由此提取的报文/流特征将偏离真实应用流量特征,造成加密流量分类精度偏低.因此,本文认为,如果想要实现更加精准的加密流量应用分类和内容识别,则需要用与其更加相关的应用层特征入手.

无论是 OSI 模型还是 TCP/IP 模型,在网络协议栈中,每一层的协议都有自己的协议数据单元 PDU,作为该协议在网络中传输的最小单位.在加密网络流量环境下,由于应用层数据一定被加密(在部分加密网络通信协议下,应用层头部也被加密,如 TLS),因此应用层主要可以被利用的特征则为应用层的协议数据单元——应用数据单元 ADU 特征.

ADU 并非一个加密流量领域的概念,而是来源于 OSI 协议栈的应用层的定义,因而仅有较少研究着眼于 ADU 的特征,其主要原因在于 ADU 特征相较于其他协议层级的 PDU 特征而言更加难以获取,因而目前学术界仅存在部分涉及到 ADU 概念的研究,缺乏对 ADU 本身的研究.

由于应用层存在种类和功能多样的协议,因此仅有对应用层具体公开协议传输特性的描述,尚未出现对应用层传输协议在实际环境中传输机理的研究,也没有一个统一标准应用数据单元 ADU 的描述.但是 ADU 与应用数据直接相关,是应用层传输机理的外在体现,这一特征仅可能被加密所干扰,而不能被加密所掩盖.

综合而言,ADU 特征相较于其他 PDU 的特征更为显著,但 ADU 很少被用于加密流量应用分类相关的研究中.因此未来关于加密流量应用分类或内容识别的研究,应该着眼于从 ADU 等应用层特征进行实现.

5.4 多点协同分布式加密流量分类与识别

目前基本上所有加密流量分类领域的研究都仅考虑在单采集点与分析点情况下的分类,未曾考虑到实际网络环境中鲁棒性分形网络拓扑所导致的单采集点无法覆盖所有甚至无法覆盖较多网络流量类型的情况,也忽略了单点采集能力与分析能力硬件瓶颈所导致的约束.在实际多点协同分布式加密流量分类场景中,除了需要考虑单点分类的能力外,还需要兼顾多点之间的高效通信、协同计算、任务调度等多方面问题,尤其是在实际网络中跨空间节点间通信必须要尽可能地降低资源消耗以保证原有网络流量的正常转发.

大数据环境下,分布式协同分析的研究已经有了部分的发展,与单点分析计算不同,分布式多点协同分析计算的研究更多地考虑差异化多节点之间的数据通信安全、数据通信效率和联合计算方式.这与多点协同分布式加密流量分类与识别的需求是一致的.首先在实际端边云网络环境中,不同地区的网络流量采集分析节点在性能和系统环境上均可能存在差异;再者,在网络流量识别领域中标签数据复杂多变且规模较大,单个计算节点难以短时间在线训练参数众多的复杂模型以保障精细化分类模型泛化能力和通用性;并且,在流量分类模型训练中边缘设备需要将流量数据发送到高性能计算集群(一般是云)进行集中处理,这种模式会产生大量的通信开销,并存在隐私泄露等信息安全隐患.

2016 年,Mcmahan 等人[106]首次提出联邦学习概念.联邦学习是一种特殊的分布式机器学习训练

范式，解决了数据难以整合的困境的同时也保护了用户数据隐私. 因为部分用于机器学习训练的数据直接包含用户隐私，该现象在计算机视觉和自然语言处理领域尤为严重，如果通过网络对该数据进行传输，则存在泄露个人隐私的问题，其次才是数据量较大、样本难以有效传输的问题；而对于加密流量分析，网络流量样本的体量十分庞大，而其自身的加密性则保护了用户的隐私，因此其面对的困境主要是 Tbps 级网络流量如何在不影响正常网络通信的情况下进行有效传输的问题.

作为一种多点协同分布式分析计算与机器学习的结合体，联邦学习适用于实际网络环境下的加密流量分类与识别. 其采用多点数据样本训练相同的分类模型并计算相关参数[107]，并将其上传至统一的服务器进行多点模型与参数的修正，以训练出泛化能力更强、分类结果更准的分类模型. 其可以在不可靠且网络性能较差的网络环境下协同训练，各个训练节点只需要上传模型或者一些参数，而非上传整个数据样本集或样本特征集，大幅度提高通信效率、节约网络资源；并且，它有效保护数据传输隐私.

目前在加密流量分析相关领域，已经有少量联邦学习的尝试. Majeed 等人[108]实现了一个基于联邦学习的 Cross-Silo 安全聚合协议，目标联邦模型优于从零开始训练的基线联邦模型. Fan 等人[109]提出了一种基于联邦迁移学习的 5G 物联网入侵检测框架 IoTDefender，通过联邦学习进行数据聚合，在不泄露隐私的同时可以极大地提高对未知攻击的检测能力.

因此，未来需要更加着眼于多点协同分布式加密流量分类与识别，尤其是资源受限场景下的多点协同分布式加密流量分类方法.

6　总　结

互联网加密流量检测、分类与识别作为互联网加密流量分析的三个重要环节，在目前互联网加密流量已成大势所趋的环境下，已经成为网络流量分析领域的重点研究问题. 针对这一问题，本文首先对目前定义混乱的加密流量分析领域研究按照不同阶段的需求目标将加密流量分析划分为加密流量检测、分类与识别三个递进阶段，然后分别对这三个阶段的研究工作进行归纳、对比分析与总结.

学术界目前对于互联网加密流量检测方面的研究主要包含对常规加密流量的检测、加密恶意流量的检测和 VPN/Tor 等特殊加密流量的检测，主要通过随机性测度分析与机器学习方法实现. 对于互联网加密流量分类方面的研究则主要从业务、应用、功能三个不同的粒度出发对加密流量进行分类，该领域研究的前提则是加密流量的检测，主要通过机器学习或深度学习方法实现. 对于互联网加密流量识别方面，则主要包含加密流量的视频/网页内容识别、用户行为识别和 QoE 识别，该领域研究的前提为加密流量的分类.

互联网加密流量检测、分类与识别三个加密流量分析子领域的研究都已经进行多年，但整体而言仍然是目前流量分析的主要方向，其研究正处于快速发展的阶段. 虽然学术界已经提出较多互联网加密流量检测、分类与识别的解决方案，但是其中大部分研究仍存在以下 7 种问题：(1) 缺数据：加密流量样本数据少，生命周期短，难以获得；(2) 吃资源：复杂特征计算成本开销大，难以应用于高速网络环境；(3) 欠精度：现有方法精度不足；(4) 失远见：拘泥于传统协议，难以应用于复杂新型网络协议场景；(5) 少性能：复杂模型受制于单点硬件处理能力；(6) 不智能：分析方法与特征选择大量依靠人工经验；(7) 难实用：实验环境下的方法忽略了实际网络环境下的问题与特性. 这些解决方案缺乏对加密流量本身的深入研究. 因此，未来关于互联网加密流量检测、分类与识别的研究则应该专注于加密流量样本数据集完善、复杂新型网络协议下的加密流量分类与识别、基于应用层特征的加密流量分类与识别、多点协同分布式加密流量分类与识别四个方向，以上方向也在本文中未来研究方向展望中进行了分析，希望能够为该领域后续研究工作的开展提供参考与建议.

参　考　文　献

[1] Yang Bo-Wen, Liu Dong. Research on network traffic identification based on machine learning and deep packet inspection //Proceedings of the 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). Chengdu, China, 2019: 1887-1891

[2] Yuan Xing-Liang, Wang Xin-Yu, Lin Jian-Xiong, et al. Privacy-preserving deep packet inspection in outsourced middleboxes// Proceedings of the IEEE 35th Annual IEEE International Conference on Computer Communications (INFOCOM 2016). San Francisco, USA, 2016: 1-9

[3] Ren Hao, Li Hong-Wei, Liu Dong-Xiao, et al. Toward efficient and secure deep packet inspection for outsourced middlebox// Proceedings of the 2019 IEEE International Conference on Communications (ICC 2019). Shanghai, China, 2019: 1-6

[4] Ning Jian-Ting, Poh G, Loh J C N, et al. PrivDPI: Privacy-preserving encrypted traffic inspection with reusable obfuscated rules//Proceedings of the 2019 ACM SIGSAC Conference. London, UK, 2019: 1657-1670

[5] Pan Wu-Bin, Cheng Guang, Guo Xiao-Jun, et al. Review and perspective on encrypted traffic identification research. Journal on Communications, 2016, 37(9):154-167(in Chinese)
(潘吴斌, 程光, 郭晓军等. 网络加密流量识别研究综述及展望. 通信学报, 2016, 37(9):154-167)

[6] Chen Liang-Chen, Gao Shu, Liu Bao-Xu, et al. Research status and development trends on network encrypted traffic identification. Netinfo Security, 2019, 219(3): 25-31 (in Chinese)
(陈良臣, 高曙, 刘宝旭等. 网络加密流量识别研究进展及发展趋势. 信息网络安全, 2019, 219(3): 25-31)

[7] Aceto G, Ciuonzo D, Montieri A, et al. Mobile encrypted traffic classification using deep learning: Experimental evaluation, lessons learned, and challenges. IEEE Transactions on Network and Service Management, 2019, 16(2): 445-458

[8] Lu Jie, Liu An-Jin, Dong Fan, et al. Learning under concept drift: A review. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(12): 2346-2363

[9] Dorfinger P, Panholzer G, John W. Entropy estimation for real-time encrypted traffic identification (short paper)// Proceedings of the International Workshop on Traffic Monitoring and Analysis. Vienna, Austria, 2011: 164-171

[10] Zhao Bo, Guo Hong, Liu Qin-Rang, et al. Protocol independent identification of encrypted traffic based on weighted cumulative sum test. Journal of Software, 2013, 24(6): 1334-1345(in Chinese)
(赵博, 郭虹, 刘勤让等. 基于加权累积和检验的加密流量盲识别算法. 软件学报, 2013, 24(6): 1334-1345)

[11] Yao Zhong-Jiang, Ge Jing-Guo, Wu Yu-Lei, et al. Meek-based Tor traffic identification with hidden Markov model// Proceedings of the 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 (HPCC/SmartCity/DSS). Exeter, USA, 2018: 335-340

[12] Rao Zhi-Hong, Niu Wei-Na, Zhang Xiao-Song, et al. Tor anonymous traffic identification based on gravitational clustering. Peer-to-Peer Networking and Applications, 2018, 11(3): 592-601

[13] Yao Hai-Peng, Liu Chong, Zhang Pei-Ying, et al. Identification of encrypted traffic through attention mechanism based long short term memory. IEEE Transactions on Big Data, 2022, 8(1): 241-252

[14] Al-Obaidy F, Momtahen S, Hossain M F, et al. Encrypted traffic classification based ml for identifying different social media applications//Proceedings of the 2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE). Edmonton, Canada, 2019: 1-5

[15] Alshammari R, Zincir-Heywood A N. Identification of VoIP encrypted traffic using a machine learning approach. Journal of King Saud University-Computer and Information Sciences, 2015, 27(1): 77-92

[16] He Gao-Feng, Xu Bing-Feng, Zhu Hai-Ting. Identifying mobile applications for encrypted network traffic//Proceedings of the 2017 5th International Conference on Advanced Cloud and Big Data (CBD). Shanghai, China, 2017: 279-284

[17] Li Feng, Chung J W, Claypool M. Silhouette: Identifying YouTube video flows from encrypted traffic//Proceedings of the 28th ACM SIGMM Workshop on Network and Operating Systems Support for Digital Audio and Video. Amsterdam, The Netherlands, 2018: 19-24

[18] Shi Y, Ross A, Biswas S. Source identification of encrypted video traffic in the presence of heterogeneous network traffic. Computer Communications, 2018, 129: 101-110

[19] Reed A, Klimkowski B. Leaky streams: Identifying variable bitrate DASH videos streamed over encrypted 802.11n connections//Proceedings of the 2016 13th IEEE Annual Consumer Communications & Networking Conference (CCNC). Piscataway, USA, 2016: 1107-1112

[20] Reed A, Kranch M. Identifying HTTPS-protected Netflix videos in real-time//Proceedings of the Seventh ACM on Conference on Data and Application Security and Privacy. New York, USA, 2017: 361-368

[21] Zhang H, Papadopoulos C, Massey D. Detecting encrypted botnet traffic//Proceedings of the IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). Turin, Italy, 2013: 3453-3458

[22] Casino F, Choo K K R, Patsakis C. HEDGE: Efficient traffic classification of encrypted and compressed packets. IEEE Transactions on Information Forensics and Security, 2019, 14(11): 2916-2926

[23] Niu Wei-Na, Zhuo Zhong-Liu, Zhang Xiao-Song, et al. A heuristic statistical testing based approach for encrypted network traffic identification. IEEE Transactions on Vehicular Technology, 2019, 68(4): 3843-3853

[24] Tang Zheng-Zhi, Zeng Xue-Wen, Sheng Yi-Qiang. Entropy-based feature extraction algorithm for encrypted and non-encrypted compressed traffic classification. International Journal of Innovative Computing, Information and Control, 2019, 15(3): 845-860

[25] Zeng Yong, Wu Zheng-Yuan, Dong Li-Hua, et al. Research on malicious traffic identification technology in encrypted traffic. Journal of Xidian University, 2021, 48(3): 170-187 (in Chinese)
(曾勇, 吴正远, 董丽华等. 加密流量中的恶意流量识别技术. 西安电子科技大学学报, 2021, 48(3): 170-187)

[26] Luo S, Seideman J D, Dietrich S. Fingerprinting cryptographic protocols with key exchange using an entropy measure //Proceedings of the 2018 IEEE Security and Privacy Workshops (SPW). San Francisco, USA, 2018: 170-179

[27] Yu Tang-Da, Zou Fu-Tai, Li Lin-Sen, et al. An encrypted malicious traffic detection system based on neural network//Proceedings of the 2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC). Guilin, China, 2019: 62-70

[28] De Lucia M J, Cotton C. Detection of encrypted malicious network traffic using machine learning//Proceedings of the 2019 IEEE Military Communications Conference (MILCOM 2019). Norfolk, USA, 2019: 1-6

[29] Zeng Yi, Qi Zi-Hao, Chen Wen-Cheng, et al. Test: An end-to-end network traffic classification system with spatio-temporal features extraction//Proceedings of the 2019 IEEE International Conference on Smart Cloud (SmartCloud). Tokyo, Japan, 2019: 131-136

[30] Jiang Tong-Tong, Yin Wei-Xin, Cai Bing, et al. Encrypted malicious traffic identification based on hierarchical spatiotemporal feature and multi-head attention. Computer Engineering, 2021, 47(7): 101-108(in Chinese)
(蒋彤彤，尹魏昕，蔡冰等. 基于层次时空特征与多头注意力的恶意加密流量识别. 计算机工程，2021，47(7)：101-108)

[31] Draper-Gil G, Lashkari A H, Mamun M S I, et al. Characterization of encrypted and VPN traffic using time-related//Proceedings of the 2nd International Conference on Information Systems Security and Privacy (ICISSP). Rome, Italy, 2016: 407-414

[32] Wang Wei, Zhu Ming, Wang Jin-Lin, et al. End-to-end encrypted traffic classification with one-dimensional convolution neural networks//Proceedings of the 2017 IEEE International Conference on Intelligence and Security Informatics (ISI). Beijing, China, 2017: 43-48

[33] Shapira T, Shavitt Y. FlowPic: A generic representation for encrypted traffic classification and applications identification. IEEE Transactions on Network and Service Management, 2021, 18(2): 1218-1232

[34] Wang Lin, Feng Hua-Min, Liu Biao, et al. SSL VPN encrypted traffic identification based on hybrid method. Computer Applications and Software, 2019, 36(2): 315-322(in Chinese)
(王琳，封化民，刘飚等. 基于混合方法的 SSL VPN 加密流量识别研究. 计算机应用与软件，2019，36(2)：315-322)

[35] Tang Shu-Ye, Cheng Guang, Jiang Bo-Miao, et al. Detection and recognition of VPN encrypted traffic based on segmented entropy distribution. Cyberspace Security, 2020, 11(8): 23-27, 33(in Chinese)
(唐舒烨，程光，蒋泊淼等. 基于分段熵分布的 VPN 加密流量检测与识别方法. 网络空间安全，2020，11(8)：23-27，33)

[36] Luo Ping, Wang Fei, Chen Shu-Hui, et al. Behavior-based method for real-time identification of encrypted proxy traffic//Proceedings of the 2021 13th International Conference on Communication Software and Networks (ICCSN). Chongqing, China, 2021: 289-295

[37] Zhou Yi-Min, Liu Fang-Zheng, Wang Yong. IPSec VPN encrypted traffic identification based on hybrid method. Computer Science, 2021, 48(4): 295-302(in Chinese)
(周益旻，刘方正，王勇. 基于混合方法的 IPSec VPN 加密流量识别. 计算机科学，2021，48(4)：295-302)

[38] He Yong-Zhong, Hu Li-Ping, Gao Rui. Detection of Tor traffic hiding under OBFS4 protocol based on two-level filtering//Proceedings of the 2019 2nd International Conference on Data Intelligence and Security (ICDIS). South Padre Island, USA, 2019: 195-200

[39] Wang Tao, Cai Xiang, Nithyanand R, et al. Effective attacks and provable defenses for website fingerprinting//Proceedings of the 2014 USENIX Security Symposium. San Diego, USA, 2014: 143-157

[40] Hayes J, Danezis G. k-fingerprinting: A robust scalable website fingerprinting technique//Proceedings of the 2016 USENIX Security Symposium. Austin, USA, 2016: 1-17

[41] Jia Ling-Yu, Liu Yang, Wang Bai-Ling, et al. A hierarchical classification approach for Tor anonymous traffic//Proceedings of the 2017 IEEE 9th International Conference on Communication Software and Networks (ICCSN). Guangzhou, China, 2017: 239-243

[42] Kim M, Anpalagan A. Tor traffic classification from raw packet header using convolutional neural network//Proceedings of the 2018 1st IEEE International Conference on Knowledge Innovation and Invention (ICKII). Jeju Island, Korea, 2018: 187-190

[43] Hu Yu-Zong, Zou Fu-Tai, Li Lin-Sen, et al. Traffic classification of user behaviors in Tor, I2P, Zeronet, Freenet//Proceedings of the 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). Guangzhou, China, 2020: 418-424

[44] Bai Hui-Wen, Ma Xue-Jing, Liu Wei-Wei, et al. Research on traffic identification technology of anonymous protocol based on deep learning. Computer Simulation, 2021, 38(7): 360-365(in Chinese)
(白惠文，马雪婧，刘伟伟等. 基于深度学习的匿名协议流量识别技术研究. 计算机仿真，2021，38(7)：360-365)

[45] Bagui S, Fang X, Kalaimannan E, et al. Comparison of machine-learning algorithms for classification of VPN network traffic flow using time-related features. Journal of Cyber Security Technology, 2017, 1(2): 108-126

[46] Houser R, Li Z, Cotton C, et al. An investigation on information leakage of DNS over TLS//Proceedings of the 15th International Conference on Emerging Networking Experiments and Technologies. New York, USA, 2019: 123-137

[47] Yang J, Narantuya J, Lim H. Bayesian neural network based encrypted traffic classification using initial handshake packets//Proceedings of the 2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks-Supplemental Volume (DSN-S). Portland, USA, 2019: 19-20

[48] Lotfollahi M, Siavoshani M J, Zade R S H, et al. Deep packet: A novel approach for encrypted traffic classification using deep learning. Soft Computing, 2020, 24(3): 1999-2012

[49] Song Ming-Ze, Ran Jing, Li Shu-Lan. Encrypted traffic classification based on text convolution neural networks//Proceedings of the 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT). Dalian, China, 2019: 432-436

[50] Cui Su-Su, Jiang Bo, Cai Zhen-Zhen, et al. A session-packets-based encrypted traffic classification using capsule neural networks //Proceedings of the 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 (HPCC/SmartCity/DSS). Zhangjiajie, China, 2019: 429-436

[51] Wang Mao-Nan, Zheng Kang-Feng, Ning Xin-Yi, et al. CENTIME: A direct comprehensive traffic features extraction for encrypted traffic classification//Proceedings of the 2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS). Chengdu, China, 2021: 490-498

[52] Aceto G, Ciuonzo D, Montieri A, et al. Encrypted multitask traffic classification via multimodal deep learning//Proceedings of the IEEE International Conference on Communications (ICC 2021). Montreal, Canada, 2021: 1-6

[53] Hu Xin-Yi, Gu Chun-Xiang, Wei Fu-Shan. CLD-Net: A network combining CNN and LSTM for Internet encrypted traffic classification. Security and Communication Networks, 2021: 1-15

[54] Lu B, Luktarhan N, Ding C, et al. ICLSTM: Encrypted traffic service identification based on inception-LSTM neural network. Symmetry, 2021, 13(6): 1080

[55] Xie Guo-Rui, Li Qing, Jiang Yong. Self-attentive deep learning method for online traffic classification and its interpretability. Computer Networks, 2021, 196: 108267

[56] Ren Xin-Ming, Gu Hua-Xi, Wei Wen-Ting. Tree-RNN: Tree structural recurrent neural network for network traffic classification. Expert Systems with Applications, 2021, 167: 114363

[57] Nigmatullin R, Ivchenko A, Dorokhin S. Accumulated generalized mean value—A new approach to flow-based feature generation for encrypted traffic characterization//Proceedings of the 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC). Las Vegas, USA, 2021: 165-169

[58] Chen Wwen-Xiong, Lyu Feng, Wu Fan, et al. Sequential message characterization for early classification of encrypted Internet traffic. IEEE Transactions on Vehicular Technology, 2021, 70(4): 3746-3760

[59] Chen Ming-Hao, Zhu Yue-Fei, Lu Bin, et al. Classification of application type of encrypted traffic based on attention-CNN. Computer Science, 2021, 48(4): 325-332(in Chinese) (陈明豪, 祝跃飞, 芦斌等. 基于 Attention-CNN 的加密流量应用类型识别. 计算机科学, 2021, 48(4): 325-332)

[60] Huoh T L, Luo Y, Zhang T. Encrypted network traffic classification using a geometric learning model//Proceedings of the 2021 IFIP/IEEE International Symposium on Integrated Network Management (IM). Bordeaux, France, 2021: 376-383

[61] Chen Zi-Han, Cheng Guang, Jiang Bo-Miao, et al. Length matters: Fast Internet encrypted traffic service classification based on multi-PDU lengths//Proceedings of the 2020 16th International Conference on Mobility, Sensing and Networking (MSN), Tokyo, Japan, 2020: 531-538

[62] Chen Zi-Han, Cheng Guang, Xu Zi-Heng, et al. Length matters: Scalable fast encrypted Internet traffic service classification based on multiple protocol data unit length sequence with composite deep learning. Digital Communications and Networks, 2021, 8(3): 289-302

[63] Soleymanpour S, Sadr H, Beheshti H. An efficient deep learning method for encrypted traffic classification on the Web//Proceedings of the 2020 6th International Conference on Web Research (ICWR). London, UK, 2020: 209-216

[64] Ma Q, Huang W, Jin Y, et al. Encrypted traffic classification based on traffic reconstruction//Proceedings of the 2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD). Huanggang, China, 2021: 572-576

[65] Muehlstein J, Zion Y, Bahumi M, et al. Analyzing HTTPS encrypted traffic to identify user's operating system, browser and application//Proceedings of the 2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC). Las Vegas, USA, 2017: 1-6

[66] Pan Wu-Bin, Cheng Guang, Tang Yong-Ning. WENC: Https encrypted traffic classification using weighted ensemble learning and Markov chain//Proceedings of the 2017 IEEE Trustcom/BigDataSE/ICESS. Sydney, Australia, 2017: 50-57

[67] Zhang Yu, Zhao Shi-Man, Zhang Jian-Zhong, et al. STNN: A novel TLS/SSL encrypted traffic classification system based on stereo transform neural network//Proceedings of the 2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS). Tianjing, China, 2019: 907-910

[68] Aceto G, Ciuonzo D, Montieri A, et al. Know your big data trade-offs when classifying encrypted mobile traffic with deep learning//Proceedings of the 2019 Network Traffic Measurement and Analysis Conference (TMA). Paris, France, 2019: 121-128

[69] Chen Yi-Ge, Zang Tian-Ning, Zhang Yong-Zheng, et al. Incremental learning for mobile encrypted traffic classification //Proceedings of the IEEE International Conference on Communications (ICC 2021). Montreal, Canada, 2021: 1-6

[70] Nascita A, Montieri A, Aceto G, et al. XAI meets mobile traffic classification: Understanding and improving multimodal deep learning architectures. IEEE Transactions on Network and Service Management, 2021, 18(4): 4225-4246

[71] Li Yan, Lu Yi-Fei, Li Shu-Ren. EZAC: encrypted zero-day applications classification using CNN and *k*-means//Proceedings of the 2021 IEEE 24th International Conference on Computer

Supported Cooperative Work in Design (CSCWD). Dalian, China, 2021: 378-383

[72] Korczyński M, Duda A. Markov chain fingerprinting to classify encrypted traffic//Proceedings of the IEEE Conference on Computer Communications (INFOCOM 2014). Toronto, Canada, 2014: 781-789

[73] Shen Meng, Wei Ming-Wei, Zhu Lie-Huang, et al. Classification of encrypted traffic with second-order Markov chains and application attribute bigrams. IEEE Transactions on Information Forensics and Security, 2017, 12(8): 1830-1843

[74] Liu Chang, Cao Zi-Gang, Li Zhen, et al. LaFFT: Length-aware FFT based fingerprinting for encrypted network traffic classification//Proceedings of the 2018 IEEE Symposium on Computers and Communications (ISCC). Guangzhou, China, 2018: 1-6

[75] Liu Chang, Cao Zi-Gang, Xiong Gang, et al. MaMPF: Encrypted traffic classification based on multi-attribute Markov probability fingerprints//Proceedings of the 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS). Banff, Canada, 2018: 1-10

[76] Liu Chang, He Long-Tao, Xiong Gang, et al. FS-Net: A flow sequence network for encrypted traffic classification//Proceedings of the IEEE Conference on Computer Communications (INFOCOM 2019). Paris, France, 2019: 1171-1179

[77] Shen Meng, Zhang Jin-Peng, Zhu Lie-Huang, et al. Accurate decentralized application identification via encrypted traffic analysis using graph neural networks. IEEE Transactions on Information Forensics and Security, 2021, PP(99): 1-1

[78] Wu Hua, Wang Lu, Cheng Guang, et al. Mobile application encryption traffic classification based on TLS flow sequence network//Proceedings of the 2021 IEEE International Conference on Communications Workshops (ICC Workshops). Montreal, Canada, 2021: 1-6

[79] Aceto G, Bovenzi G, Ciuonzo D, et al. Characterization and prediction of mobile-app traffic using Markov modeling. IEEE Transactions on Network and Service Management, 2021, 18(1): 907-925

[80] Wang Zi-Xuan, Wang Pan, Zhou Xiao-Kang, et al. FLOWGAN: Unbalanced network encrypted traffic identification method based on GAN//Proceedings of the 2019 IEEE International Conference on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom). Exeter, UK, 2019: 975-983

[81] Shi Yan, Feng De-Zhi, Cheng Yu, et al. A natural language-inspired multilabel video streaming source identification method based on deep neural networks. Signal Image Video Process, 2021, 15(6): 1161-1168

[82] Hou Cheng-Shang, Shi Jun-Zheng, Kang Cui-Cui, et al. Classifying user activities in the encrypted WeChat traffic//Proceedings of the 2018 IEEE 37th International Performance Computing and Communications Conference (IPCCC). Orlando, USA, 2018: 1-8

[83] Yan Fei-Peng, Xu Ming, Qiao Tong, et al. Identifying WeChat red packets and fund transfers via analyzing encrypted network traffic//Proceedings of the 2018 17th IEEE International Conference On Trust, Security and Privacy in Computing and Communications/12th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE). New York, USA, 2018: 1426-1432

[84] Tabash F K, Izharuddin M, Tabash M I. Encryption techniques for H.264/AVC videos: A literature review. Journal of Information Security and Applications, 2019, 45: 20-34

[85] Schuster R, Shmatikov V, Tromer E. Beauty and the burst: Remote identification of encrypted video streams//Proceedings of the 26th USENIX Security Symposium (USENIX Security 17). Vancouver, Canada, 2017: 1357-1374

[86] Gu Jia-Xi, Wang Ji-Liang, Yu Zhi-Wen, et al. Walls have ears: Traffic-based side-channel attack in video streaming//Proceedings of the IEEE Conference on Computer Communications (INFOCOM 2018). Honolulu, USA, 2018: 1538-1546

[87] Gu Jia-Xi, Wang Ji-Liang, Yu Zhi-Wen, et al. Traffic-based side-channel attack in video streaming. IEEE/ACM Transactions on Networking, 2019, 27(3): 972-985

[88] Wu Hua, Yu Zhen-Hua, Cheng Guang, et al. Encrypted video recognition in large-scale fingerprint database. Journal of Software, 2021, 32(10): 3310-3330(in Chinese)
(吴桦，于振华，程光等. 大型指纹库场景中加密视频识别方法. 软件学报, 2021, 32(10): 3310-3330)

[89] Sirinam P, Imani M, Juarez M, et al. Deep fingerprinting: Undermining website fingerprinting defenses with deep learning//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. Toronto, Canada, 2018: 1928-1943

[90] Zhang Zi-Qing, Kang Cui-Cui, Xiong Gang, et al. Deep forest with LRRS feature for fine-grained website fingerprinting with encrypted SSL/TLS//Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19). New York, USA, 2019: 851-860

[91] Shen Meng, Liu Yi-Ting, Zhu Lie-Huang, et al. Fine-grained webpage fingerprinting using only packet length information of encrypted traffic. IEEE Transactions on Information Forensics and Security, 2020, 16: 2046-2059

[92] Wang Cheng-Gang, Dani J, Li Xiang, et al. Adaptive fingerprinting: Website fingerprinting over few encrypted traffic//Proceedings of the 11th ACM Conference on Data and Application Security and Privacy. Virtual Event, USA, 2021: 149-160

[93] Shen Meng, Gao Zhen-Bo, Zhu Lie-Huang, et al. Efficient

fine-grained website fingerprinting via encrypted traffic analysis with deep learning//Proceedings of the International Workshop on Quality of Service. Virtual, 2021: 1-10

[94] Conti M, Mancini L V, Spolaor R, et al. Analyzing Android encrypted network traffic to identify user actions. IEEE Transactions on Information Forensics & Security, 2016, 11(1): 114-125

[95] Wu Hua, Wu Qiu-Yan, Cheng Guang, et al. SFIM: Identify user behavior based on stable features. Peer-to-Peer Networking and Applications, 2021, 14(6): 3674-3687

[96] Bahuguna A, Agrawal A, Bhatia A, et al. User profiling using smartphone network traffic analysis//Proceedings of the 2021 International Conference on COMmunication Systems & NETworkS (COMSNETS). Virtual, 2021: 69-73

[97] Khokhar M J, Ehlinger T, Barakat C. From network traffic measurements to QoE for internet video//Proceedings of the 2019 IFIP Networking Conference (IFIP Networking). Warsaw, Poland, 2019: 1-9

[98] Wassermann S, Seufert M, Casas P, et al. I see what you see: Real time prediction of video quality from encrypted streaming traffic//Proceedings of the 4th Internet-QoE Workshop on QoE-Based Analysis and Management of Data Communication Networks. Los Cabos, Mexico, 2019: 1-6

[99] Tang Shuang, Qin Xiao-Wei, Chen Xiao-Hui, et al. Video quality assessment for encrypted http adaptive streaming: Attention-based hybrid RNN-HMM model//Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019). Brighton, UK, 2019: 2362-2366

[100] Xu Shi-Chang, Sen S, Mao Z M. CSI: inferring mobile ABR video adaptation behavior under HTTPS and QUIC//Proceedings of the 15th European Conference on Computer Systems. Dresden, Germany, 2020: 1-16

[101] Wu Hua, Li Xin, Cheng Guang, et al. Monitoring video resolution of adaptive encrypted video traffic based on HTTP/2 features//Proceedings of the IEEE Conference on Computer Communications Workshops (INFOCOM 2021). Virtual, 2021: 1-6

[102] Bazuhair W, Lee W. Detecting malign encrypted network traffic using perlin noise and convolutional neural network//Proceedings of the 2020 10th Annual Computing and Communication Workshop and Conference (CCWC). Las Vegas, USA, 2020: 0200-0206

[103] Samarasinghe N, Mannan M. Another look at TLS ecosystems in networked devices vs. Web servers. Computers & Security, 2019, 80: 1-13

[104] Bhargavan K, Blanchet B, Kobeissi N. Verified models and reference implementations for the TLS 1.3 standard candidate//Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP). San Jose, USA, 2017: 483-502

[105] Langley A, Iyengar J, Bailey J, et al. The QUIC transport protocol: Design and Internet-scale deployment//Proceedings of the Conference of the ACM Special Interest Group on Data Communication. Los Angeles, USA, 2017: 183-196

[106] Mcmahan H B, Moore E, D Ramage, et al. Communication-efficient learning of deep networks from decentralized data//Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS). Fort Lauderdale, USA, 2017: 1-11

[107] Kairouz E, Mcmahan H B, Avent B, et al. Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 2021, 14(1-2): 1-210

[108] Majeed U, Hassan S S, Hong C S. Cross-silo model-based secure federated transfer learning for flow-based traffic classification//Proceedings of the 35th International Conference on Information Networking (ICOIN). Korea, 2021: 588-593

[109] Fan Yu-Lin, Li Yang, Zhan Meng-Qi, et al. IoTDefender: A federated transfer learning intrusion detection framework for 5G IoT//Proceedings of the 2020 IEEE 14th International Conference on Big Data Science and Engineering (BigDataSE). Guangzhou, China, 2020: 88-95

CHEN Zi-Han, Ph. D. candidate. His research interests include cyber security, encrypted network traffic analysis, trust architecture and cyberspace governance.

CHENG Guang, Ph. D., professor. His research interests include network security, network measurement, encrypted traffic identification, traffic behavior analysis and proactive defense.

XU Zi-Heng, M. S. candidate. His research interests include cyber security and encrypted network traffic analysis.

XU Ke-Ya, M. S. candidate. Her research interests include cyber security and encrypted network traffic analysis.

QIU Xing, Ph. D. candidate. His research interests include cyber security and encrypted network traffic analysis.

NIU Dan-Dan, M. S. candidate. Her research interests include cyber security and network traffic analysis.

Background

With the continuous development of encrypted traffic technology, Internet traffic encryption has become an inevitable trend of Internet development. However, network traffic encryption not only brings privacy and security to users and enterprises, but also brings challenges to network security supervision and network traffic management, as traditional traffic measurement and analysis methods such as DPI are not suitable for encrypted traffic environment. At present, there are a lot of researches in the field of encrypted traffic analysis, but they are classified according to the classification method of encrypted traffic and its input or output, without unified standard of the granularity and level of encrypted traffic analysis as well as a systematic theoretical definition. This has brought troubles to the direction subdivision and work refinement in the field of encrypted traffic analysis to some extent.

In this paper, we first divides Internet traffic analysis into three stages: encrypted traffic detection, encrypted traffic classification and encrypted traffic identification, and elaborates the characterization of these three stages from the perspective of actual network users. Then we analyze and compare the existing Internet encrypted traffic detection methods, classification methods and identification methods from multiple perspectives, and summarize their advantages and disadvantages respectively.

The three subfields of Internet encrypted traffic analysis have been studied for many years, but they are still the main directions of it with rapid development. Although there have put forward many solutions for the detection, classification and identification of encrypted traffic on the Internet, most of them still have problems such as idealized environment, antiquated protocol, single dataset, simplified features, and lack of in-depth research on encrypted traffic itself.

Hence, we finally combine the development trend of Internet network environment in the future to analyze and outlook the possible research directions in the three stages of Internet encrypted traffic analysis in the future. We summarize the future research directions into four kinds of problem unsolved as encrypted traffic dataset perfection, encrypted traffic classification and identification under new complex network protocols (including TLS-1.3, encrypted DNS, HTTP-2.0, and QUIC), application layer feature based encrypted traffic classification and identification, and multi-point cooperative distributed encrypted traffic classification and identification, hoping to provide a reference and suggestion for further research work in this field.

This paper is supported by the General Program of the National Natural Science Foundation of China under Grant No. 62172093, the National Key R&D Program of China under Grant No. 2020YFB1804604, the 2019 Industrial Internet Innovation and Development Project, Ministry of Industry and Information Technology of China under Grant No. 6789010003. This paper is part of the topic for the encrypted traffic analysis.

The research team has focused on encrypted traffic analysis for years, and some papers in this field have published in highly-ranked journals.