

基于多粒度表征学习的加密恶意流量检测

谷勇浩^{1),2)} 徐昊¹⁾ 张晓青¹⁾

¹⁾(北京邮电大学计算机学院 智能通信软件与多媒体北京市重点实验室 北京 100876)

²⁾(中山大学广东省信息安全技术重点实验室 广州 510275)

摘要 现有加密恶意流量检测方法中,基于统计特征的方法存在特征提取依赖专家经验和特征之间相互独立的问题,基于原始输入的机器学习和深度学习的方法存在信息不全、随机字段、单一粒度的问题,对加密流量交互行为的语义表征不足。为解决上述问题,本文提出一种基于多粒度表征学习的加密恶意流量检测方法 MGREL (Multi-Granularity REpresentation Learning)。该方法将加密会话分为字段级和包级两个粒度分别处理。在字段级粒度中,基于词向量进行局部行为建模,提取握手报文并选取关键字段,缓解信息不全导致的语义缺失问题,将字段的字节值表示为词向量,同时增加报文类型与握手类型作为位置前缀,解决位置语义缺失的问题,采用 Multi-head Attention 计算字段间的交互,再通过 BiLSTM 得到报文级语义;在包级粒度中,基于时空进行全局行为建模,提取包的时空状态信息并采用 LSTM 模型得到流级语义。将两个粒度下得到的局部行为语义和全局行为语义融合,得到加密流量的表征,解决单一粒度表征能力不足的问题。最后,通过对比实验验证本文所提方法 MGREL 在检测加密恶意流量方面表现最好。

关键词 加密恶意流量检测;多粒度表征学习;局部行为;全局行为;位置语义

中图法分类号 TP309 **DOI号** 10.11897/SP.J.1016.2023.01888

Multi-Granularity Representation Learning for Encrypted Malicious Traffic Detection

GU Yong-Hao^{1),2)} XU Hao¹⁾ ZHANG Xiao-Qing¹⁾

¹⁾(Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876)

²⁾(Guangdong Provincial Key Laboratory of Information Security Technology, Sun Yat-sen University, Guangzhou 510275)

Abstract In the field of encrypted malicious traffic detection, the current detection methods are insufficient. In the method based on statistical features, feature extraction relies on expert experience, and the features are independent of each other; while the method based on original input has problems of incomplete information, random fields, and single granularity, and cannot learn the semantics of traffic interaction behavior well. In order to overcome the shortcomings of existing methods, this paper proposes an encrypted malicious traffic detection method MGREL (Multi-Granularity REpresentation Learning). This method divides the encrypted session into two granularities, field-level and packet-level. In field-level granularity, local behavior modeling is performed based on word vectors, handshake messages are extracted and key fields are selected to relieve the problem of incomplete information, the byte values of fields are represented as word vectors, and message types are added at the same time. Use the handshake type as the location prefix to solve the problem of lack of location semantics. Multi-head Attention is used to calcu-

收稿日期:2022-06-14;在线出版日期:2023-02-23. 本课题得到北京邮电大学中央高校基本科研业务费行动计划项目(2021XD-A11-1)、CCF-蚂蚁科研基金(20210026)和广东省信息安全技术重点实验室开放基金(2020B1212060078)资助。谷勇浩(通信作者),博士,讲师,CCF会员(05596M),主要研究领域为网络安全、数据挖掘。E-mail: guyonghao@bupt.edu.cn. 徐昊,硕士研究生,主要研究领域为网络安全。张晓青,硕士研究生,主要研究领域为网络安全。

late the interaction between fields, and then BiLSTM is used to obtain the message-level semantics. In the packet-level granularity, global behavior modeling is performed based on space and time, and packets are extracted. The spatiotemporal state information is obtained and the LSTM model is used to obtain stream-level semantics. The local behavior semantics and global behavior semantics obtained at two granularities are fused to obtain the representation of encrypted traffic, which solves the problem of insufficient representation capability of a single granularity. Finally, it is verified by comparative experiments that the proposed method MGREL performs the best in detecting encrypted malicious traffic.

Keywords encrypted malicious traffic detection; multi-granularity representation learning; local behavior; global behavior; positional semantics

1 引言

如今,网络活动日渐频繁,随之产生的加密流量也在增加,互联网研究趋势报告显示 87% 的网站流量是加密的^①. 加密协议主要采用 TLS 1.2 协议,通过对传输内容进行加密,从而保护内容安全,但是攻击方也可以借用加密的手段,隐藏其攻击内容,从而规避现有检测手段. WatchGuard Technologies 指出,在 2021 年第二季度,通过加密 HTTPS 投送的恶意软件比例高达 91.5%^②,加密恶意流量检测已经成为如今不可忽视的问题.

在传统检测方法中,主要是基于载荷的深度包检测(DPI). 该方法采用解密加密流量的方法,消耗大量资源,引起用户对数据隐私的担忧. 为避免应用载荷解密,提出基于统计特征^[1-13]的方法,通过提取流量的统计特征进行流量的识别,但是该方法提取特征依赖专家经验,缺乏对流量交互行为语义的挖掘. 为此,学者利用深度学习模型对原始输入^[14-24]进行特征自挖掘以及流量内交互语义的学习. 本文将重点放在 TLS 1.2 及以前的版本,研究如何对流量交互行为进行表征学习,实现加密恶意流量的有效检测,解决如下三方面问题:

(1)信息不全导致语义缺失. 基于原始输入的方法由于输入信息不全导致语义缺失.

(2)随机字段影响表征学习效果. 在交互过程中,部分字段是随机值,并没有实际的语义信息.

(3)单一粒度的表征学习无法完整挖掘流量的行为语义.

本文提出多粒度表征学习的加密恶意流量检测方法. 首先,在加密流量的通信握手过程中,提取报文,挑选其中的关键字段构建词向量,避免信息缺失

与无意义的信息,采用 Multi-head Attention 学习字段权重,通过 BiLSTM (Bi-directional Long Short-Term Memory)挖掘流量局部握手行为语义,从流量的字段粒度得到流量的局部行为表示. 在加密流量的全局行为过程中,提取包的时间和空间状态信息,通过 LSTM(Long Short-Term Memory)挖掘流量全局交互行为语义,从流量的包级粒度得到流量的全局行为表示. 最后,将局部行为表示与全局行为表示进行融合,得到多粒度下的流量表征.

本文主要贡献如下:

(1)提出加密流量的字段级粒度划分方法,通过字段粒度的表征学习,挖掘加密流量局部交互行为的深层语义表示. 首先对加密流量握手阶段按照通信报文进行划分,在提升语义可解释性的同时缓解信息不全的问题;然后提取报文的关键字段,解决随机字段的问题.

(2)提出基于类型前缀的关键字段词向量表示方法. 首先以关键字段构建原始词向量,然后增加类型前缀,解决关键字段位置语义缺失的问题,提高字段级粒度表征学习的效果.

(3)提出多粒度表征学习的加密恶意流量检测方法. 从字段级粒度得到局部通信语义,从包级粒度得到全局行为语义,融合得到最终的加密流量表示,解决单一粒度表征学习能力不足的问题,提高检测精确率.

2 相关工作

根据机器学习/深度学习模型输入信息的不同,

① MEEKER, M. Internet Trends, <https://www.bondcap.com/report/itr19/>. 2019. 2019

② WatchGuard's Threat Lab Analyzes the Latest Malware and Internet Attacks, <https://www.watchguard.com/wgrd-resource-center/security-report-q2-2021>. 2021

现有加密恶意流量检测方法主要分为基于统计特征、基于原始输入和基于混合输入三类。

2.1 基于统计特征

基于统计特征的方法是流量识别最常用的方法,此方法通过直接提取或计算样本属性统计值作为特征,比如负载大小、时间间隔等。Umer^[1]提出基于增强 SVM(Support Vector Machine)的恶意流量检测框架,实现了无监督下的加密流量检测。Amoli^[2]提出两阶段的检测框架,若待测流量统计值超过阈值,则通过 DBSCAN 进行聚类,区分正常与恶意。Liu^[3]提出一种基于距离的方法,通过 GMM(Gaussian Mixture Model)聚类计算恶意样本间的距离,输出伪标签,使用 XGBoost(eXtreme Gradient Boosting)进行训练形成最终分类器。Fang^[4]采用 RF(Random Forest)模型,对特征集合进行划分验证,指出下行流量特征更为有效。然而,以上方法是传统恶意流量检测方法的迁移,并没有针对加密流量设计特征。

在加密恶意流量分析中,传输内容是密文,仍可以在前述统计特征的基础上增加握手阶段的明文特征向量,例如选取 TLS 协议 Client Hello 中的支持加密组件、拓展项,Server Hello 中的长度、版本等。Troia^[5]通过对流量日志提取明文特征,采用 SVM、XGBoost、RF 模型进行筛选。Hu^[6]提出 Client Hello 和 Server Hello 报文中的明文信息作为特征,采用逻辑回归模型,实现对加密恶意流量的检测。Jakub^[7]提出 TLS 中的握手时间和上下两次请求的间隔作为特征,采用 M-INDEX 度量的 KNN 模型进行恶意流量的检测。Chen^[8]提出 THS-IDPC 方法,通过 DPC-GSMND 算法将数据集划分为小集群,采用 XGBoost、SVM、RF 对小集群内的流量进行检测。Anderson^[9]提出了增强特征集合,包括 TLS 密码套件,明文特征使模型效果提升明显。Dai^[10]提出针对明文向量的特征筛选方法,从握手与证书中提取信息,通过互信息对特征集合进行筛选从而提升模型的检测能力。Yu^[11]采用 Multi-AEs,将特征从低维度扩展到高维度,从而寻找各个特征之间的关系。Gracia^[12]在 Apache Spark 流中布置分布式聚类模型,实现了对加密 SlowDOS 攻击的检测。Andrey^[13]提出了 TLSVec 的方法,将 Client Hello 与 Server Hello 中的明文字段,通过 Word2Vec 转化为向量,输入到 LSTM 中进行学习。但是,基于统计特征的方法存在两方面问题:(1)特征提取高度依赖专家经验,特征有限且只针对

专门的攻击模式;(2)特征之间相互独立,缺乏对流量交互行为语义的挖掘。

2.2 基于原始输入

基于原始输入的加密恶意流量检测方法,依靠 CNN、RNN 等深度学习模型自动挖掘流量的深层次特征。Gonzalo^[14]提出,目前的深度学习模型按照得到的语义划分为包级表征与流级表征。本文从模型输入的角度定义流量的粒度,依照输入粒度不同将现有方法划分为字节级粒度与包级粒度。

2.2.1 字节级粒度

基于字节级粒度的输入方法中,将输入的流量划分为字节,将流量前 m 个字节作为模型的输入。Wang^[15]提出基于 2D-CNN 的方法,提取前 784 字节的负载内容,转化为 28×28 的图片,通过 LeNet-5 进行学习。Zeng 等^[16]使用 CNN、LSTM、SAE 模型,由算法自动挖掘特征进行检测。Wang^[17]提出基于 1D-CNN 的方法,通过 1D-CNN 模型挖掘字节之间的交互关系。Cheng^[18]使用 Word2vec 模型将字节转化为向量,利用 CNN 学习字节之间的交互,实现对加密恶意流量的检测。Lin^[19]提出 ET-BERT 方法,将字节转化为双词袋 datagram,设置预训练任务进行模型的调优,实现多个领域下加密流量的检测。

2.2.2 包级粒度

基于包级粒度的输入方法中,将输入的流量划分为包,选择前 m 个包作为流的表示,挖掘流中包与包的交互语义。包的表示分为负载内容和状态量。基于负载内容方法中,选择前 n 个字节作为包的表示,是对流量内容交互的挖掘。Cheng^[20]提出 RTETC 方法,输入前 3 个包的嵌入表示,采用了 Multi-head Attention 和 1D-CNN,提取流量包内的交互与流量包间的交互。Jiang^[21]提出 HST-MHSA 方法,基于流量层次结构,结合长短时记忆网络和 TextCNN 整合加密流量的多尺度局部特征和全局特征,并引入 Multi-head Attention 进一步增强关键特征的区分度。在基于状态量表示中,选择负载大小、时间等状态量作为包的表示,挖掘流量的时空交互语义。例如,Lucia^[22]将 TLS 报文大小与方向作为 SVM 和 CNN 的输入,获取流量的空间交互语义。Dong^[23]提出 DirPiz(Directed Packet payload size),将有效负载大小与方向作为包的表示。Zou^[24]提出以负载大小、时间间隔等作为包的表示,通过 LSTM 进行时序信息的挖掘。

但是,基于原始输入的方法存在三方面问题:(1)输入模型的信息不全导致语义缺失^[14-17].由于模型的输入向量是固定长度的,需要对负载内容进行统一的截取,导致流量被截取为前后两个部分,后半部分中的语义丢失;(2)随机字段影响表征学习效果^[14-19].模型的输入中包含大量随机信息,如密钥,Session ID,对于流量负载来说,并没有实际语义信息,模型难以学到有效输入信息;(3)单一粒度的表征学习无法完整挖掘流量的行为语义^[20-22].同时,对于字节级粒度的输入,每个字节的可解释性较差;对于包级粒度的输入,存在一个报文在多个包中和多个报文在一个包中的情况(如图 1),导致报文整体语义被分割或没有学习到细粒度的报文语义等问题.

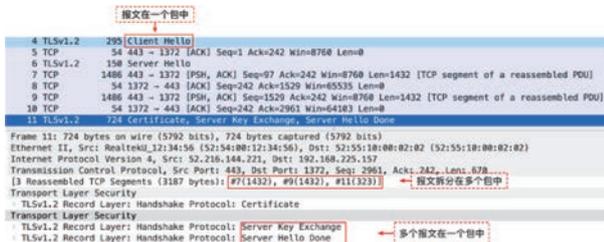


图 1 TLS 报文

2.3 基于混合输入

部分研究者尝试使用多种流量表示,通过模型分别学习到不同模态的语义信息. Aceto^[25] 提出以原始负载与前 32 个包的行为序列作为输入,分别输入到 1D_CNN 和 GRU 中,并进行特征的融合.在

此基础上,Ofek^[26] 提出增加以方向、是否握手划分的流量画像作为第三通道,通过 2D_CNN 进行特征挖掘,与前两个通道的输出进行特征的融合.然而,基于混合输入的方法更多是对多模态下表征的融合,也就是流量相同行为的不同表征,缺失对关键行为粒度的挖掘.因此,如何对加密流量更好地划分粒度,并从不同粒度下挖掘流量的行为语义,从而提高加密恶意流量检测的准确率,是本文的研究重点.

3 建模方法

基于多粒度表征学习 Multi-Granularity Representation Learning(MGREL)的加密恶意流量检测框架如图 2 所示,包括数据预处理、多粒度表征学习两阶段.其中,多粒度表征学习按照输入粒度划分为字段级粒度与包级粒度两个部分.在字段级粒度中,基于词向量进行局部行为建模,提取握手报文并选取关键字段,将字段的字节值表示为词向量,同时增加报文类型与握手类型作为位置前缀,通过 Multi-head Attention 计算字段与字段之间的交互,再通过 BiLSTM 得到报文级语义;在包级粒度中,基于时空进行全局行为建模,对前 n 个包的状态进行提取,通过 LSTM 得到流级语义.最后将多粒度下的语义进行融合,输入到 Softmax 中进行恶意流量的判别.

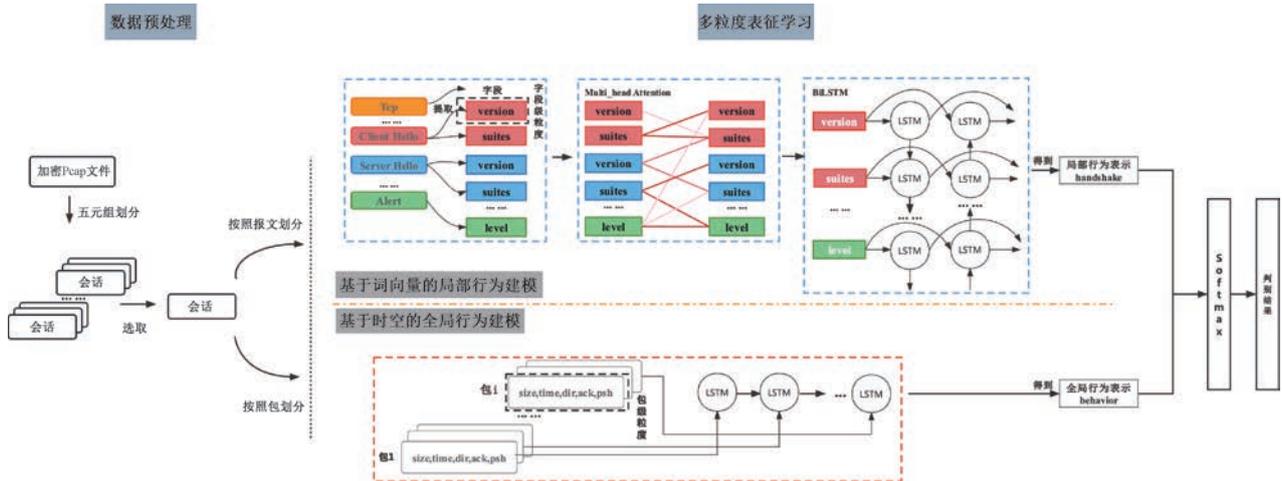


图 2 基于 MGREL 的加密恶意流量检测框架

3.1 数据预处理

首先对于每一个加密流量 Pcap 文件,以五元组(源 ip、目的 ip、源端口、目的端口、传输层协议)作为流量划分依据.相比单向流,从流量通信交互的角

度,选择信息更为丰富、双向的会话 session 作为样本划分的粒度.每一个加密流量 Pcap 文件,经过划分得到若干个会话.后文中的流量不再指全部流量,而是指划分后的会话 session 作为样本.

3.2 基于时空的全局行为建模

对 Datacon 数据集(详见 4.1.1)进行统计后发现,恶意样本和正常样本在平均负载大小和时间间隔上的分布差异较大(如图 3 和图 4 所示),这是由于恶意流量和正常流量的行为模式不同.例如,恶意流量中攻击者会利用不同端口反复尝试下发攻击指令,由于其证书与行为的可疑,遭到频繁拒绝,导致恶意流量中包含大量负载大小和时间间隔偏小的会话样本.此外,恶意行为常采用前期下达攻击指令、后期回传数据的模式,导致恶意流量负载大小与数据包传输方向有较强的相关性.为了探究流量全局通信的语义信息,以包为粒度对流量进行划分,根据统计,数据集的绝大多数样本包的数量少于 30 个,因此选取流量前 30 个包代表全局流量, $session$ 在包粒度下的表征如下:

$$session_{behavior} = \{p_1, p_2, \dots, p_{30}\}$$

$$p_t = \{size, time, dir, ack, psh\} \quad (1)$$

在第 t 个包 p_t 中, $size$ 是其负载大小, $time$ 是距离上一个包的时间间隔, dir 是传输方向, ack 和 psh 是 TCP 层的 Flags. 包负载大小、时间间隔、传输方向以及 TCP 标志位,代表流量的时空状态语义^[24].

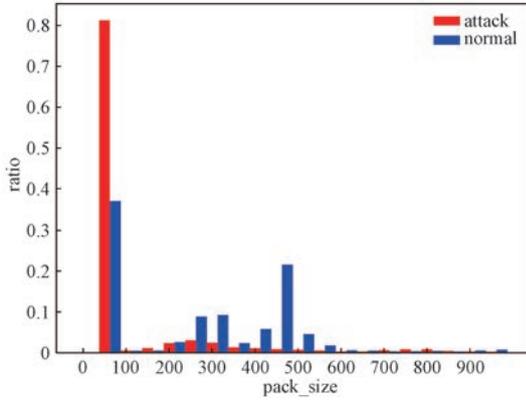


图 3 正常与恶意样本平均负载大小的频率直方图

本文采用 LSTM 对流量全局交互进行学习,挖掘流量全局的时序语义,得到全局行为表示. LSTM 在 RNN 的基础上,引入了门控的概念,构建了输入门 i_t , 遗忘门 f_t 以及输出门 o_t 三个门和一个内部记忆单元 c_t , 前述的每个包 p_t 对应 LSTM 的一步输入. 具体计算公式如下:

$$i_t = \sigma(W_i p_t + U_i h_{t-1} + b_i) \quad (2)$$

$$f_t = \sigma(W_f p_t + U_f h_{t-1} + b_f) \quad (3)$$

$$\tilde{c}_t = \text{Tanh}(W_c p_t + U_c h_{t-1}) \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (5)$$

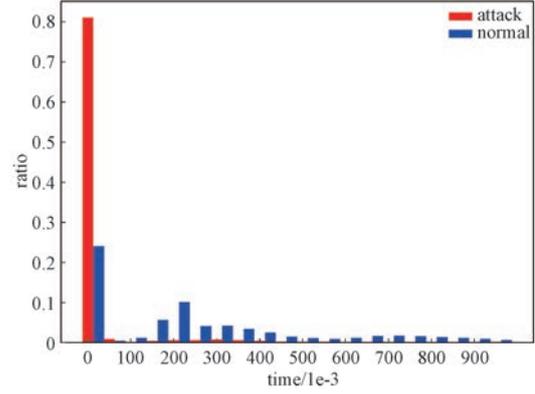


图 4 正常与恶意样本时间间隔的频率直方图

$$o_t = \sigma(W_o p_t + U_o h_{t-1} + b_o) \quad (6)$$

$$h_t = o_t \odot \text{Tanh}(c_t) \quad (7)$$

$$behavior = \text{LSTM}(p_1, p_2, \dots, p_{30}) \quad (8)$$

其中 W 和 U 是门的参数矩阵, 分别表示对输入 p_t 的线性变化与隐状态 h_t 的线性变化, b 是偏置, 下标对应相应的门, 比如 b_i 对应输入门的权重偏置.

通过输入门 i_t 和遗忘门 f_t , 控制当前 t 时刻包 p_t 和 $t-1$ 时刻记忆单元 c_{t-1} 哪些特征用于 c_t , 通过输出门 o_t 得到隐节点 h_t . 每一个时刻的 p_t , 通过计算, 都得到了对应隐节点 h_t , 从而得到全局行为表示 $behavior$.

3.3 基于词向量的局部行为建模

在加密通信过程中, 通信双方在数据传输之前进行握手交互, 传递双方信息, 本文将此过程定义为局部. 与全局通信不同, 局部通信包括 TCP 三次握手和 TLS 加密握手两个阶段.

单看局部行为的某个阶段, 流量恶意性可能不显著, 如果挖掘该阶段与其他阶段的交互行为, 会提升恶意的表现, 如图 5 所示. 本文提出三种交互行为模式.

(1) 不同报文字段间的交互. 例如恶意样本中 Client Hello 的支持加密组件有若干, 而 Server Hello 最终选用的加密组件为 0xc02f, 是备选和选用的关系; Client 端检测出 Server 端的证书过期, 产生“Certificate Expired”的警报.

(2) 报文内字段间的交互. 在报文 Certificate 中, subject 和 issuer 分别是证书的持有者与颁发者. 在恶意样本中, 经常出现低信誉机构或者自身颁发证书.

(3) 报文间的交互. 在报文 Server Key Exchange 中, 改变报文 Server Hello 中所传递的密钥, 是某类报文回应特定报文所产生的行为交互.

为了学习到会话中上述三种交互行为的语义, 本文将提取会话中的关键字段(词)并组成词序列

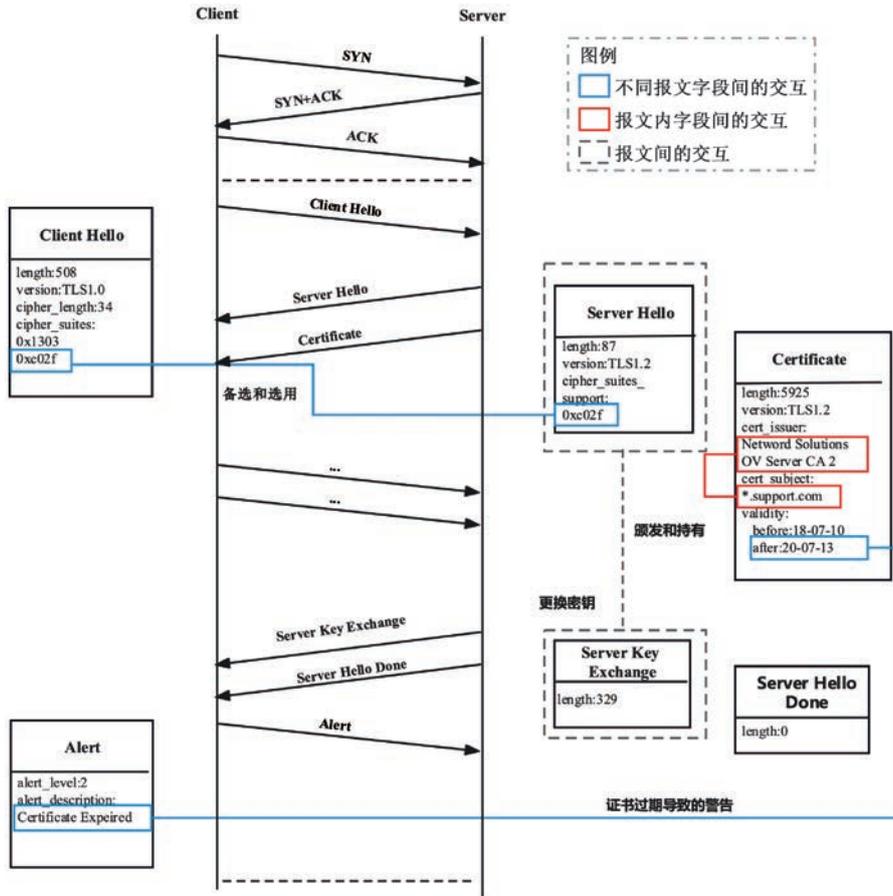


图 5 恶意样本握手过程中的交互

(句子),然后采用自然语言处理模型进行表征学习.从会话中提取的字段序列在表征学习过程具有以下特点:(1)不同会话包含的报文数不同,不同类型报文包含的关键字段数目不同(即,每个会话对应的字段序列长度可能不同);(2)不同关键字段受上下文字段影响有差异;(3)不同字段出现的先后次序对字段语义表征有差异.因此,本文需要将会话分为报文,报文分为不同字段,然后采用 Multi-head Attention 机制学习字段上下文的语义,最后采用 BiLSTM 学习字段序列双向时序语义.将 session 按照报文划分如下:

$$session_{handshake} = \{record_1, record_2, \dots, record_n\} \quad (9)$$

其中,record_i 代表第 i 个报文.

在局部行为建模过程中,session 包括 TCP 三次握手、TLS 加密阶段握手报文.加密报文类型有 Handshake、Change Cipher Spec、Application Data、Alert,分别表示握手过程的不同阶段.其中,Handshake 是握手报文,Change Cipher Spec 是改变密钥报文,Application Data 是具体发送的加密数据,

Alert 是警报报文.由于传输数据不在本文讨论 TLS 加密握手之中,因此舍弃 Application Data.

对于每一个 record,基于字段粒度划分,表示如下:

$$record_i = \{word_1, word_2, \dots, word_{m_i}\} \quad (10)$$

其中,word_j 是 record_i 中的第 j 个关键字段,每一个报文包含多个字段,字段如表 1 所示,其中选取关键字段的依据如下:

(1)是否是随机生成的.部分字段(如,密钥)本身只是一个密钥交换过程中的参数,对识别恶意与否没有区分度,因此删除此类字段.如果某个报文(如,Server Key Exchange)只存在前述交互关系(3),则将其中的字段都置空.

(2)是否能够区分正常与恶意.部分字段(如,加密组件、证书颁发时间等)中,正常与恶意流量之间存在着分布上的差异,恶意软件作者只关心内容加密而不关注算法的选择,会更多选择老旧加密组件^①,因此选择此类字段.

^① 关于恶意软件加密流量检测的思考, <https://mp.weixin.qq.com/s/ZqVrgScEGGdz3CVyiJ6vJA>

表 1 TLS 1.2 协议关键字段表

类型	字段名	含义	类型	字段名	含义
IP	ip_src	源 ip 地址	Client Hello	cipher_length	加密组件长度
	ip_dst	目的 ip 地址		cipher_suites	支持加密组件
TCP	sport	源端口号	Certificate	cert_issuer	证书中的 issuer
	dport	目的端口号		cert_subject	证书中的 subject
TLS Record	content_type	报文类型	Server Hello	cert_notbefore	证书有效期的开始时间
	length	报文长度		cert_notafter	证书有效期的结束时间
	version	报文版本号		cipher_suites_support	选用加密组件
Handshake	handshake_type	报文握手类型	Certificate Status	certificate_status	证书状态
New Session Ticket	ticket_lifetime_hint	Ticket 的剩余时间	Alert	alert_level	威胁等级
	ticket_length	Ticket 的长度		alert_description	威胁描述
Certificate Request	certificate_type_count	请求证书类型	Certificate Verify	signature_hash_algorithm	证书签名类型
Client Key Exchange	premaster_length	密钥长度	Change Cipher Spec	change_cipher_spec_message	更新当前密钥

基于上述选取依据,对每一个报文选取若干个关键字段.对于每个关键字段的向量表示,采用字段本身对应字节值 b .此外,由于不同字段长度不同,而输入模型的向量长度要统一.基于对关键字段长度的统计,大部分长度小于 4,因此将关键字段长度统一为 4.长度不足 4 的字段,高位补 0;长度大于 4 的字段,每 4 个字节构成一个 $word$ 向量,不足 4 个字节的补足 4 字节.每一个字段初始表征为:

$$word_raw_j = \{b_1, b_2, b_3, b_4\} \quad (11)$$

另外,不同位置的字段,存在值相同但意义完全不同的情况,例如两个字段,值都为 769,一个代表加密协议版本号,一个代表加密报文内容长度.为了区分不同位置上的语义差异,本文选择 TLS 协议的报文类型 $type_record$ 和握手类型 $type_handshake$ 作为前缀,形成关键字段的向量表示 $word_j$:

$$word_j = [prefix; word_raw_j] \quad (12)$$

$$prefix = \{type_record, type_handshake\} \quad (13)$$

从字段的粒度出发,对 num 个 $word$ 进行拼接.为了从所有字段中挖掘重点信息,采用 Multi-head Attention 对关键字段的权重进行计算,使模型更加充分捕获通信过程中字段的交互关系.本文采用 self-attention 的计算方法,利用查询 $Q = W_Q X$,键 $K = W_K X$,值 $V = W_V X$ 对字段的交互关系进行挖掘,其中, X 是模型输入的向量矩阵,每行代表一个词向量(由公式(12)计算得到),矩阵 W_Q 、 W_K 、 W_V 为可学习的模型参数.同时根据并行头的设置,将特征空间划分为 num_heads 个子空间,将参数矩阵划分为 W_i ,分别计算不同子空间下的权值,最后进行拼接输出.具体的计算过程如下:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (14)$$

$$MultiHead(Q, K, V) =$$

$$Concat(Z_1, Z_2, \dots, Z_{num_heads})W^O \quad (15)$$

$$Z_i = Attention(Q_i, K_i, V_i) \quad (16)$$

其中,由(14)得到会话中所有字段具有高层语义的特征矩阵(*Single-head Attention*),头(*head*)指参数矩阵 W_Q 、 W_K 、 W_V 的个数,相当于计算机视觉领域的卷积核的个数.*self-attention* 使用 *Multi-head* 是为了提取不同上下文语境下的语义信息特征.同样的词出现在不同的句子中,或者在不同的语序中,所表达的意思可能大不相同,不同的参数矩阵代表不同的语义提取方向.

同时,我们采用双向 *BiLSTM* 的方式提取关键字段之间的前后时序信息,得到流量的通信握手语义 *handshake*,具体的计算过程如下

$$handshake = LSTM(word_1, word_2, \dots, word_{num}) \\ + LSTM(word_{num}, word_2, \dots, word_1) \quad (17)$$

将流量按照包级粒度和字段级粒度划分后,得到基于时空的全局行为表示 *behavior* 和基于词向量的局部行为表示 *handshake*,然后拼接两个粒度下的语义表示向量,经过线性层 *Linear*,最终通过 *Softmax* 输出判别结果 *output*,如公式(18).

$$output =$$

$$Softmax(Linear([behavior; handshake])) \quad (18)$$

算法 1. MGREL 算法训练过程.

输入:流量原始输入 raw_input ,标签 y_label ,

超参数 $num, num_heads, \alpha, \gamma$;

输出:模型参数 θ ;

1: 根据五元组,对 raw_input 提取样本 $sample$;

2: FOR each $sample$ in raw_input DO

3: 根据公式(8),得到流量的全局表示 $behavior$;

4: 根据公式(9),对 raw_input 提取 $session_handshake$;

5: 根据公式(14)~(16),对 $session_handshake = \{word_1,$

$word_2, \dots, word_{num}$ }, 计算 $session$ 中每个关键字段在多头注意力机制下的 att_i ;

6: FOR each i in $(1, num)$ DO

7: $word_i = word_i * att_i$

8: END FOR

9: 根据公式(17), 得到 $handshake$;

10: $session = [behavior; handshake]$;

11: 根据公式(18)得到 $output$;

12: 得到预测标签 $y_{pred} = \text{argmax}(output)$;

13: 若 $y_{pred} = y_{label}$, 计算 $Loss = -\alpha(1 - output[0])^\gamma \log(output[0])$, 其中, α 值对两类样本的数量进行惩罚和奖励, γ 值对难以训练的样本进行奖励, 并更新模型参数 θ ;

14: END FOR

15: RETURN θ ;

MGREL 算法训练过程如算法 1 所示. 对于原始输入 raw_input , 拆分为若干会话样本(算法体第 1 行). 对于每一个样本, 得到对应的全局表示(第 2-3 行)和局部表示(第 4-9 行), 将两种粒度下的表示融合最终得到样本的预测标签 y_{pred} (第 10-12 行), 结合本身标签 y_{label} 计算梯度, 更新模型参数并得到最优参数值(第 13-15 行).

4 实验

4.1 数据集

4.1.1 Datacon 数据集

Datacon 数据集^①源自于 2020 年 2 月~6 月收集的恶意软件与正常软件, 经奇安信技术研究院天穹沙箱运行并采集其产生的流量筛选生成. 该数据集定义的恶意流量为恶意软件(均为 exe 类型)产生的加密流量, 共有 90842 条, 正常流量为正常软件(均为 exe 类型)产生的加密流量, 共有 30235 条.

4.1.2 Stratosphere 数据集

Stratosphere 数据集^②是由 Stratosphere IPS 提供的恶意流量与正常流量组成. 恶意流量是恶意软件产生的流量, 种类包括 trickbot、dridex 等恶意软件, 共有 17883 条; 同时 Stratosphere IPS 为了验证模型, 还收集了大量正常行为产生的流量, 共有 51695 条.

4.2 评价指标及实验设置

本文使用准确率(*Accuracy*)、精确率(*Precision*)、召回率(*Recall*)和 $F1$ 值作为评价指标, 公式如下:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (19)$$

$$Precision = \frac{TP}{TP + FP} \quad (20)$$

$$Recall = \frac{TP}{TP + FN} \quad (21)$$

$$F1 = \frac{2Precision \times Recall}{Precision + Recall} \quad (22)$$

其中, TP 表示属于恶意的流量并且被正确分类为恶意的流量数目; TN 表示属于正常流量并且被分类为正常流量的数目; FP 表示属于正常流量但是被分类为恶意流量的数目; FN 表示属于恶意流量但是被分类为正常流量的数目.

MGREL 模型的参数设置如下: 在 LSTM 层, 隐藏神经元为 144, 激活函数为 tanh; 在 BiLSTM 层, 隐藏神经元为 144, 激活函数为 tanh; 在 Multi-head 层, 并行头数量为 2; 在 Softmax 层, 输出设为 2; 在 Dropout 层, 遗忘率为 0.3, 最终采用 Focal loss 损失函数, 采用 Adam 优化算法, batch_size 设置为 8, 数据集按照训练集、验证集和测试集划分为 6 : 2 : 2(测试集中正负样本相等).

4.3 对比实验

将我们的方法与 SOTA(State-Of-The-Art)方法进行对比, 实验结果如表 2. 可以看到本文提出的加密恶意流量检测方法明显优于其他方法.

在数据集 Datacon 中, 基于原始输入的方法^[14-15, 17, 20, 24-26]在大多数指标下, 好于基于统计特征的方法. 基于传统特征的方法存在大量误判, 验证了统计特征对流量交互行为刻画不足, 流量交互行为信息有利于对正常与恶意进行区分. 在基于原始输入的方法^[14-15, 17, 20, 24]与 MGREL 对比中, MGREL 在不过多损失召回的基础上, 有效提高了精确率, 降低了误判. 由于原始输入方法中, 无法预料切割位置, 可能导致有效的明文信息被直接舍弃, 同时模型输入中包含大量随机信息, 难以学到有效输入信息. MGREL 通过字段的划分有效降低无效字节与不完整内容带来的影响, 使模型更加专注于有效信息的挖掘, 提高了模型对加密正常流量的识别能力. MGREL 要优于融合方法 DISTILLER 与 MaID-IST, 证明了多种粒度下流量行为交互融合中, 局部行为与全局行为多粒度的融合要优于多模态的融合.

① DataCon 社区. DataCon 开放数据集—DataCon2020—加密恶意流量数据集方向开放数据集[DB/OL]. 2021-11-11. <https://datacon.qianxin.com/opendata/openpage?resourcesId=6>

② Stratosphere IPS. Malware Capture Facility Project. URL: <https://www.stratosphereips.org/datasets-malware>

本文所提方法针对大部分的行为模式检测效果较好,但是对少部分短样本的检测没有明显效果.例如部分正常样本仅经历了三次握手、Client Hello、Alert、四次挥手等四个阶段,本身信息较少,并且流量信息间交互语义不够丰富,导致模型的检测效果没有明显提升.

为了验证模型的泛化能力,我们还在其他数据集上进行了实验.在数据集 Stratosphere 中, MGREL 相

比其他模型结果更好.同时,本文对不同数据集下实验效果相差较大的原因进行分析,在数据集 Datacon 中,正常与恶意流量都处于同一环境下收集,并且经过人为筛选,正常与恶意的区分难度相对较大,而在数据集 Stratosphere 中,由于采集环境的不同,正常流量与恶意流量时间上差异性明显,模型区分正常和恶意流量相对容易.因此,为了更好地验证模型的效果,后文均是在 Datacon 数据集下进行实验.

表 2 对比实验

Dataset	Datacon				Stratosphere				
	model	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
RF ^[4]	0.8603	0.7846	0.9933	0.8767	0.9791	0.9999	0.9583	0.9787	
LR ^[9]	0.8170	0.7447	0.9648	0.8406	0.9993	0.9985	1.0	0.9993	
TLSVEC ^[13]	0.8413	0.7725	0.9678	0.8591	0.9979	0.9989	0.9969	0.9979	
DeepMal ^[14]	0.876	0.807	0.9885	0.8886	0.999	0.9999	0.9982	0.999	
2D_CNN ^[15]	0.8976	0.8365	0.9883	0.9061	0.9862	0.9981	0.9743	0.9861	
1D_CNN ^[17]	0.9053	0.8485	0.9868	0.9124	0.9997	1.0	0.9994	0.9997	
RTETC ^[20]	0.8948	0.8307	0.9917	0.9041	0.9994	0.9994	0.9994	0.9994	
LSTM ^[24]	0.8803	0.8149	0.9841	0.8916	0.9764	0.9984	0.9544	0.9759	
DISTILLER ^[25]	0.9235	0.8762	0.9865	0.928	0.9999	0.9999	0.9999	0.9999	
MaIDIST ^[26]	0.9141	0.86	0.9893	0.9201	0.9998	1.0	0.9996	0.9998	
MGREL	0.9654	0.9596	0.9716	0.9656	1.0	1.0	1.0	1.0	1.0

4.4 模型消融实验

为了验证模型各部分的有效性,本文从以下三方面设计消融实验:(1)是否增加位置语义的学习(prefix);(2)不同词向量表征方法效果的对比;(3)局部握手与全局语义的融合效果.

表 3 Prefix 的消融对比

model	Accuracy	Precision	Recall	F1
w/o prefix	0.9289	0.8847	0.9847	0.9328
with prefix	0.9561	0.9294	0.9872	0.9574

从表 3 可以看出,加入位置信息 prefix 后,模型效果显著增加,准确率和 F1 分别提高了 2.7% 和 2.4%.通过对每个关键字段加入 prefix,模型增加了字段的位置信息输入,位置信息的引入可以帮助模型更充分的理解不同字段的语义.

表 4 不同模型的消融对比

model	Accuracy	Precision	Recall	F1
TextCNN	0.8957	0.8315	0.9924	0.9049
LSTM	0.9516	0.9219	0.9868	0.9532
BiLSTM	0.9529	0.9245	0.9863	0.9544
Attention	0.9561	0.9294	0.9872	0.9574

从表 4 可以看出,本文尝试了多种基模型作为局部行为特征的模型层,在 TextCNN、LSTM、BiLSTM 中,BiLSTM 性能更为优异.得益于流量的时

序特点,时序模型能够更好的捕获流量的时序信息.加入 Multi_head Attention 后的 BiLSTM 要优于单纯的 BiLSTM,证明了注意力的引入,通过挖掘各个字段之间的交互关系,一定程度提升了模型的检测效果.

表 5 局部与全局的消融对比

model	Accuracy	Precision	Recall	F1
behavior	0.8803	0.8149	0.9841	0.8916
handshake	0.9554	0.9274	0.9882	0.9568
all	0.9561	0.9294	0.9872	0.9574

从表 5 可以看出,局部行为语义的挖掘,相比全局语义的挖掘,对检测效果帮助更大;同时,将两种语义融合后,检测效果还有进一步提升.加入 behavior 之后,精确率提升,说明对恶意流量查准率有提高,即误报样本减少;召回率下降,说明对恶意流量查全率下降,即漏报样本增加.加入全局行为表征后,让部分正常样本与恶意样本的区分度提升,但是也让部分恶意样本学习到的全局行为与正常样本相似.例如负载大小,分布在 200-250、600-900 等区间内的恶意流量与正常流量相似,被模型判为正常流量,从而拉低召回率.

两种粒度是不同角度对加密流量的刻画,多种粒度可以缓解某种粒度缺失时的问题.多粒度方法的开销中,全局行为建模只增加一层 LSTM.对

Datacon 数据集进行训练与测试时,局部表征模型训练时间为 170s/epoch,测试时间为 20s,加入全局语义后训练时间为 207s/epoch,测试时间为 21s.从时间开销上看,多粒度表征学习模型的主要开销在局部表征学习,增加全局语义学习的开销相对小很多.此外,检测模型应用时,企业更关注模型能有较低的误报(要减少运维人员工作量),在 F1 提高的前提下,适当牺牲召回率换取精确率的提高是值得的.

4.5 模型超参数实验

本实验为了验证超参数对局部通信建模结果的影响,分别改变模型的超参数关键字段数目(num),并行多头数(num_heads).

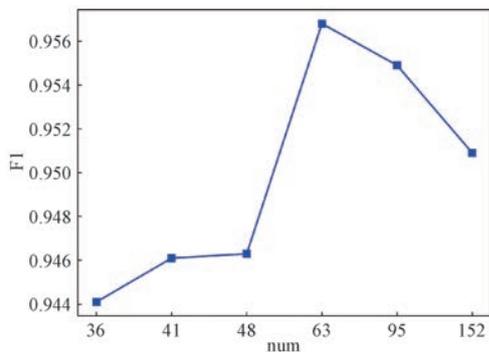


图 6 不同超参数取值对结果的影响

在图 6 中,将整个数据集根据每条会话包含关键字段数目总数,进行升序排列生成数组.选取整个数组的 60%、70%、平均值(mean)、80%、90%、99% 处元素得到相应的 num 值分别为 36、41、48、63、95、152 的关键字段数目.

表 6 不同超参数取值对结果的影响(num_heads)

num_heads	Accuracy	Precision	Recall	F1
1	0.9210	0.8732	0.9851	0.9258
2	0.9568	0.9294	0.9872	0.9574
3	0.9517	0.9209	0.9882	0.9534
6	0.9506	0.9249	0.9808	0.9521

在表 6 中,由于每个向量长度为 6,因此将子空间划分为 1、2、3 和 6,给出对应 num_heads 的实验结果,从实验效果来看, $num_heads = 2$ 已经是一个比较好的划分子空间数目,通过划分子空间,可以更好的挖掘流量的交互语义.

4.6 损失函数实验

为了缓解训练集和真实环境中,正常流量与恶意流量不均的情况,本文对比了两类损失函数的实验效果.

如表 7 所示,Focal Loss 的引入,很好的平衡了训

练集中正负样本的不均衡.在召回率不降低太多的情况下,精确率提高 3%,准确率与 F1 也相应提高.

表 7 Loss 的对比实验

Loss	Accuracy	Precision	Recall	F1
Cross Entropy	0.9561	0.9294	0.9872	0.9574
Focal	0.9654	0.9596	0.9716	0.9656

5 结 论

本文从多粒度角度对加密流量进行建模.首先基于字段级粒度对流量进行划分,加入前缀表示字段位置信息,采用 Multi-head Attention 计算权重,通过 BiLSTM 得到流量的局部行为表示;然后基于包级粒度对流量进行划分,提取每一个包的状态表示,通过 LSTM 得到流量的全局行为表示.最终将两个过程的语义进行融合,有效提高了加密恶意流量的检测效果.本文所提关键字段还可以扩展,读者可以提出其他新的字段输入模型.

未来,我们会做以下三方面的尝试:(1)使用更多粒度的输入对加密流量进行建模,通过对每个粒度得到的表示,利用 Attention 进行计算,使得模型对每个粒度下得到的表示有所偏重,提升全局表征学习能力,得到更全面的加密恶意流量语义表征;(2)构建更合理的数据集,数据集应具备避免采集环境干扰、恶意样本充足且丰富等特点,克服现有数据集的不足;(3)对使用其他加密协议(如, TLS 1.3、QUIC)流量建模,实现更加全面的恶意流量检测.

致 谢 本文特此感谢奇安信所提供的加密恶意流量识别数据集 Datacon.

参 考 文 献

[1] Umer M F, Sher M, Bi Y. A two-stage flow-based intrusion detection model for next-generation networks. Plos One, 2018, 13(1): e0180945

[2] Amoli P V, Hamalainen T. A real time unsupervised NIDS for detecting unknown and encrypted network attacks in high speed network//Proceedings of the 2013 IEEE International Workshop on Measurements & Networking. Naples, Italy, 2013: 149-154

[3] Liu J Y, Tian Z Y, Zheng R F, Liu L. A distance-based method for building an encrypted malware traffic identification framework. IEEE Access, 2019, 7: 100014-100028

- [4] Fang Y, Xu Y, Huang C, Liu L, Zhang L. Against malicious SSL/TLS encryption; identify malicious traffic based on random forest//Proceedings of the 2018 IEEE International Conference on Big Data. Settle, USA, 2018; 1258-1265
- [5] Troia F D, Stamp M. Feature analysis of encrypted malicious traffic. *Expert Systems with Applications Magazine*, 2019, 125: 130-141
- [6] Hu Bin, Zhou Zhi, Yao Li-Hong, et al. Malicious traffic detection combining features of packet payload and stream fingerprint. *Computer Engineering*, 2020, 46(11): 157-163
(胡斌, 周志洪, 姚立红, 李建华等. 结合报文负载与流指纹特征的恶意流量检测, 2020, 46(11): 157-163)
- [7] Jakub L, Jan K, P? emysl Āech, et al. K-NN classification of malware in HTTPS traffic using the metric space approach//Proceedings of the Pacific-Asia Workshop on Intelligence and Security Informatics. Auckland, New Zealand, 2016, 131-145
- [8] Chen L C, Liu B X, Lu Z G, Jiang Z W. THS-IDPC: A three-stage hierarchical sampling method based on improved density peaks clustering algorithm for encrypted malicious traffic detection. *The Journal of Supercomputing*, 2020, 7676: 7489-7518
- [9] Anderson B, McGrew D. Identifying encrypted malware traffic with contextual flow data//Proceedings of the 2016 ACM Workshop. Vienna, Austria, 2016, 35-46
- [10] Dai R, Gao C, Lang B, et al. SSL malicious traffic detection based on multi-view features//Proceedings of the 9th International Conference on Communication and Network Security. Chongqing, China, 2019; 40-46
- [11] Yu T, Zou F T, Li L, 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). Gui Lin, China, 2019; 62-70
- [12] Garcia N, Alcaniz T, A González-Vidal, et al. Distributed real-time SlowDoS attacks detection over encrypted traffic using Artificial Intelligence. *Journal of Network and Computer Applications*, 2021, 173: 102871
- [13] Ferriyan A, Thamrin AH, Takeda K, Murai J. Encrypted malicious traffic detection based on Word2Vec. *Electronics*, 2022; 11(5): 679-684
- [14] Gonzalo M, Pedro C, Germ' a C. Deepmal-deep learning models for malware traffic detection and classification, <https://arxiv.org/abs/2003.04079>
- [15] Wang W, Zhu M, Zeng X W, Ye X Z, Sheng Y Q. Malware traffic classification using convolutional neural network for representation learning//Proceedings of the 2017 International Conference on Information Networking (ICOIN). Hefei, China, 2017, 712-717
- [16] Zeng Y, Gu H, Wei W, et al. Deep-full-range: A deep learning based network encrypted traffic classification and intrusion detection framework. *IEEE Access*, 2019, 7: 45182-45190
- [17] Wang W, Zhu M, Wang J, Zeng X, Yang Z. 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
- [18] Cheng Hua, Xie Jin-Xin, Chen Li-Huang. CNN-based encrypted C&C communication traffic identification method. *Computer Engineering*, 2019, 45(8): 31-34, 41
(程华, 谢金鑫, 陈立皇. 基于 CNN 的加密 C&C 通信流量识别方法. *计算机工程*, 2019, 45(8): 31-34+41)
- [19] Lin X, Xiong G, Gou G, et al. ET-BERT: A contextualized datagram representation with pre-training transformers for encrypted traffic classification//Proceedings of the ACM Web Conference 2022. Lyon, France, 2022: 633-642
- [20] Jin C, He R, Yue E, et al. Real-time encrypted traffic classification via lightweight neural networks//Proceedings of the GLOBECOM 2020-2020 IEEE Global Communications Conference. Taipei, China, 2020; 1-6
- [21] Jiang Tong-Tong, Yin Wei-Xin, CAI Bin, et al. Encrypted malicious traffic identification based on hierarchical spatio-temporal feature and multi-head attention. *Computer Engineering*, 2021, 47(7): 101-108
(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)
- [22] M J de Lucia, C Cotton. Detection of encrypted malicious network traffic using machine learning//Proceedings of the 2019 IEEE Military Communications Conference (MILCOM). Norfolk, USA: IEEE, 2019 :1-6
- [23] Dong C, Lu Z, Chen Y, Cui Z L. MBTree: Detecting encryption RAT communication using malicious behavior Tree. *IEEE Transactions on Information Forensics and Security*, 2021, 16: 3589-3603
- [24] Zou Yuan, Zhang Jia, Jiang Bin. Detection of malicious encrypted traffic based on LSTM recurrent neural network. *Computer Applications and Software*, 2020, 37(2): 308-312
(邹源, 张甲, 江滨. 基于 LSTM 循环神经网络的恶意加密流量检测. *计算机应用与软件*, 2020, 37(2): 308-312)
- [25] Aceto G, Ciunzo D, Montieri A, et al. *Journal of Network and Computer Applications*, 2021, 183: 102985
- [26] Ofek B, Adi L, Chen H, Ran D, Amit D. MalDIST: From encrypted traffic classification to malware traffic detection and classification//Proceedings of the 2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, USA, 2022, 527-533



GU Yong-Hao, Ph. D., master supervisor. His main research interests include network security and data mining.

XU Hao, M. S. candidate. His main research interests focus on network

security.

ZHANG Xiao-Qing, M. S. candidate. Her main research interests focus on network security.

Background

Encrypted malicious traffic detection is a traffic analysis technique in network security. The main methods are divided into statistical feature-based and raw input-based. In the statistical feature-based method, the statistical features and plaintext features of the traffic are used as input, but there are problems that the features depend on expert experience and are independent of each other; in the original input-based method, many scholars try to use a variety of deep models (CNN, RNN, LSTM) to realize feature self-mining. However, in the existing modeling methods, there are problems of incomplete information and random fields in the division granularity, and representation

learning of a single granularity cannot fully mine the behavior information of traffic. This paper proposes a new method for detecting encrypted malicious traffic based on multi-granularity representation learning MGREL (Multi-Granularity REpresentation Learning). This method divides the encrypted session into two granularities, field-level and packet-level. The local behavior semantics of traffic is obtained from the field-level granularity, and the global behavior semantics of the traffic is obtained from the packet-level granularity. MGREL improves the detection capability of encrypted malicious traffic by mining the semantics of traffic at two granularities.