基于分段熵分布的VPN加密流量检测与识别方法

唐舒烨[1,2,3]，程光[1,2,3,4]，蒋泊淼[1,2,3]，陈子涵[1,2,3]，郭树一[1,2,3]
〔1.东南大学网络空间安全学院，江苏南京 211189；2.网络空间国际治理研究基地（东南大学），江苏南京 211189；3.网络通信与安全紫金山实验室，江苏南京 211111；4.教育部计算机网络和信息集成重点实验室（东南大学），江苏南京 211189〕

摘 要：为加强对VPN加密流量的有效监管，提高网络监管的效率效果，针对VPN加密流量报文信息缺失，流量特征混淆的特点，文章提出了一种基于分段熵分布的VPN加密流量检测与识别方法。该方法利用滑动窗口方法对VPN加密报文序列高熵、低熵区域进行划分，并以此作为流量特征，使用胶囊神经网络模型实现VPN加密流量的精准检测与识别。不同于现有的基于机器学习的加密流量检测方法，该方法针对VPN加密流量本身特性进行研究，具有方法的普适性。实验与对比分析证明，该方法识别准确率达99.87%，可以用于VPN加密流量检测识别。

关键词：加密流量识别；VPN流量；信息熵；胶囊神经网络
中图分类号： TP393 **文献标识码：** A

Detection and recognition of VPN encrypted traffic based on segmented entropy distribution

Tang Shuye[1,2,3], Cheng Guang[1,2,3,4], Jiang Bomiao[1,2,3], Chen Zihan[1,2,3], Guo Shuyi[1,2,3]
〔*1. School of Cyber Science and Engineering, Southeast University, Jiangsu Nanjing 211189; 2.International governance research base of Cyberspace (Southeast University), Jiangsu Nanjing 211189; 3.Purple Mountain Laboratories for Network and Communication Security, Jiangsu Nanjing 211111;4.Key Laboratory of Computer Network and Information Integration of Ministry of Education (Southeast University), Jiangsu Nanjing 211189*〕

Abstract: In order to strengthen the effective supervision of VPN encrypted traffic and improve the efficiency of network supervision, the article proposes a method for detecting and identifying VPN encrypted traffic based on segmented entropy distribution in view of the lack of information in VPN encrypted traffic messages and the confusion of traffic characteristics. This method uses the sliding window method to divide the high-entropy and low-entropy regions of the VPN encrypted message sequence, which are taken as traffic characteristics. Then, this method uses the capsule neural network model to achieve accurate detection and identification of VPN encrypted traffic. Different from the existing encrypted traffic detection method based on machine learning, this method studies the characteristics of VPN encrypted traffic itself and has the universality of the method. Experiments and comparative analysis prove that the accuracy of this method is 99.87%.It can be used for VPN encrypted traffic detection and identification.

Key words: encrypted traffic identification ; VPN traffic; information entropy; capsule neural network

1 引言

虚拟专用网（Virtual Private Network，VPN）技术作为加密网络流量的主要使用技术之一，依靠加密隧道等手段，向用户提供便利、隐秘的远程访问等操作。然而，被VPN工具掩盖下的流量，失去了原有流量的报文头部信息、流量侧信道特征信息，也给网络监管带来了新的挑战。

因此，针对VPN加密流量测量分析已经刻不容缓。本文对使用V2Ray工具的VPN加密流量信息熵分布特性进行研究，提出了一种基于分段熵分布的VPN加密流量检测与识别方法，实现VPN加密流量的检测与识别，加强对VPN流量的有效监管。

2 相关工作

目前，主流的加密流量检测识别技术一般是根据加密流量数据均匀随机分布[1]的特点进行研究。而现有专门针对VPN流量的研究，则大多基于Gerard Draper-Gil提供的ISCX VPN-nonVPN公开数据集[2]展开。

借助于该公开数据集，Bagui Sikha[3]和王琳[4]使用了6种传统机器学习分类模型进行检测。同时，一些深度学习[5]方法也被运用于VPN流量检测识别研究，如基于注意力机制的长短期记忆网络[6]（Attention-based Long Short-Term Memory）、胶囊神经网络模型[7]（Capsule Neural Networks）、立体变换神经网络[8]（Stereo Transform Neural Network）等。然而，现有的VPN加密流量识别方法拘泥于机器学习方法的改进和应用，并未针对流量本身特性进行研究。

3 分段熵分布检测与识别方法

3.1 VPN流高熵低熵区域划分方法

目前，主流使用的VPN工具是V2Ray工具，V2Ray工具使用私有协议VMess实现数据的随机化加密传输，VMess协议是一种基于TCP协议的无状态协议，协议本身没有握手过程。然而，普通的加密流量在传播过程中包含协议握手和数据传输两个过程，其中协议握手过程通常存在随机性较低的明文字段，包含未加密的双方通信协商的参数及公钥信息；加密数据传输过程中的报文内容主要是加密后的数据以及头部少量的会话标识、加密数据长度等信息，也存在有少量不可或缺的明文头部。

鉴于此，VPN加密流量的检测识别可以根据VMess协议下VPN加密流量呈现高度均匀随机分布，且不含明文头部的特点，围绕VPN流量的信息熵分布特性展开研究。

信息熵（Entropy）是反映能量分布均匀程度的测度。给定一个概率分布:

$$P=(p_i)_{xi\in\Sigma} \tag{1}$$

其包含元素的集合为:

$$\sum=\{x_0..x_1,...x_{m-1}\} \tag{2}$$

则该分布 的熵可以表示为:

$$H(P)=-\sum_{i=1}^{m-1}p_i\log p_i \tag{3}$$

然而在实际需求中，需要在欠采样条件下利用熵的估计值实现对目标数据的随机性的准确判断以提高识别效率。因此本文引入N-截断熵（$H_N(P)$）的概念，N-截断熵被定义为根据概率分布P产生的所有长度为P的样本序列的熵的平均值，当P为均匀分布U时，可以得到长度为N的随机序列的N-截断熵$H_N(U)$为:

$$H_N(P)=\sum_{\sum n_i=N}\left[\binom{N}{n_0,\ldots,n_{m-1}}\prod_{i=0}^{m-1}p_i^{n_i}\left(-\sum_{i=0}^{m-1}\frac{n_i}{N}\log\frac{n_i}{N}\right)\right] \tag{4}$$

因而在检测随机序列时，可以通过对$\hat{H}_N$和$H_N(U)$的近似程度得出序列是否随机的结论。$H_N(U)$计算复杂度随着N和m增长呈指数型增长，当$N\to+\infty$时，$\hat{H}_N(s)$渐进于高斯分布。因此，可以利用产生所有样本熵平均值μ（$\mu\sim H_N(U)$）、标准偏差σ，描述长度为N的随机序列熵的分布。

VPN加密流量加密化程度整体较高，因此可以针对VPN流量报文中的高熵低熵区域进行划分，以大小为N的滑动窗口的形式以固定步长τ在流序列中遍历每一窗口内序列的熵值，当滑动窗口遍历完流序列时，对每个序列分段计算其有效载荷。接着，本文依据N-截断熵理论对高熵区域的判别原则，对每个序列分段的有效载荷进行判定。根据蒙特卡洛算法，通过对多组长度为N的随机字段求其$H(P)$发现，$H(P)$的取值符合正态分布，绝大部分的$H(P)$值在$[H_N(P)-4\sigma，H_N(P)+4\sigma]$的区间范围内，$\sigma$为标准偏差，以此进行高熵区域的判别，从而将序列分段标记为高熵或低熵区域，

接着研究其分布情况，以此作为单条流量报文的流量特征。本文的流量特征是报文数据熵值特征而不是具体的报文数据，不存在特定报文序列检测，同时也不受数据包时空特性影响，具有方法上的普适性。

3.2 VPN加密流量检测识别方法

在对VPN加密流量的高熵低熵区域分布情况进行研究之后，采用胶囊神经网络模型（CapsNet）进行VPN加密流量检测识别。常见的神经网络大多使用数据包空间特征或是时间序列特征进行加密流量检测，而忽视了报文数据内部的相对位置，如前文提及的高熵低熵区域分布的位置特征等。胶囊神经网络正可以弥补这一缺陷，它使用向量代替标量作为神经单元，向量可以有效利用流量特征之间的逻辑关系，学习流量特征的属性，尤其是VPN加密流量分段熵之间的位置与顺序关系等。且与卷积神经网络相比，胶囊神经网络可以有效减少学习所需的训练数据量。因此，胶囊神经网络更适合于VPN加密流量检测与识别。

基于胶囊神经网络的VPN加密流量检测识别方法首先对前文的结构化流量数据特征进行提取，并统一转换成可供胶囊神经网络模型使用的特征矩阵文件，然后对特征矩阵执行多次卷积运算和加权运算，生成多个子向量，其中向量长度表示VPN加密流量存在的概率，向量方向表示各实例化的参数，此时每个子向量即是一个子胶囊神经单元，同时每个子向量选择一个上层节点作为父胶囊神经单元。

接着该方法通过动态路由协议，将子胶囊神经单元的预测结果在向量网络结构中向上表征并传递给父胶囊神经单元，最终随着路由机制的不断迭代，将所有子向量封装成一个高维向量，由归一化的Softmax回归模型分类器对高维向量进行VPN加密流量识别分类，并输出识别分类结果。该方法可以通过增加卷积层数与调整动态路由协议迭代次数的方式进行胶囊神经网络模型调优，有效提高流量识别分类准确率。

实验所使用的胶囊神经网络结构如图1所示，分为输入层、预处理层、第一胶囊层（Primary Capsules）与第二胶囊层（Digit Capsules）。其中，输入层为K维特征向量，预处理层使用卷积核大小为1的1-d卷积层对输入层输入的K维向量进行处理，得到有C_0个通道的张量，并将得到的输出张量作为第一胶囊层的输入。第一胶囊层使用S_0个卷积大小为1的1-d卷积层将输入张量从C_0个通道压缩到C_1个通道，得到S_0组通道数为C_1的张量并将其进行合并，最终得到$K\times C_1$个长度为S_0的胶囊，并使用非线性压缩函数squash对这些胶囊进行处理。完成后使用动态路由算法（迭代轮数为L）将第一胶囊层得到的$K\times C_1$个胶囊映射到第二胶囊层中，第二胶囊层一共包括C个长度为S_1的胶囊$caps_i$，其中C为最终分类的类别数量，最终输出每一个类别的分类结果。

$$res_i = softmax\left(\sqrt{\sum_{val\in caps_i} val^2}\right) \tag{5}$$

4 实验与分析

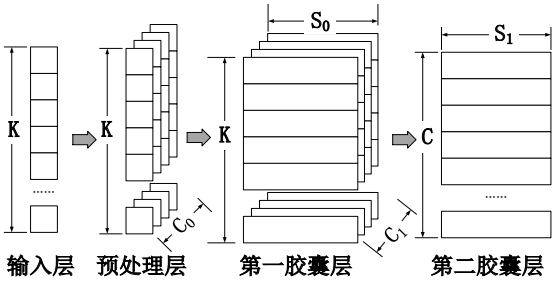

图1 胶囊神经网络结构

实验所用机器CPU为i7-6700HQ，GPU为GTX1060，软件环境Python版本为py3.7，Pytorch版本为15.1。

4.1 高熵与低熵区域划分实验

为研究VPN报文与普通加密报文的高熵与低熵区域划分差异，以大小为32B的滑动窗口的形式以固定步长32B在每一条流量数据中遍历每一窗口内序列的熵值，即N=τ=32，此时$H_N(U)=4.8782$。使用蒙特卡洛算法对标准偏差σ进行计算，模拟长度为32B的随机序列求其对应的$H(P)$并进行统计。由于N-截断熵是用于对较短字节序列的随机性进行衡量，所以本文使用随机生成的字节序列而非真实加密报文进行计算，完成随机性指标的计算后，再对加密报文有效载荷字段的随机性强弱进行分析，一

共进行10^6次模拟实验，求得$\sigma = 0.0814$，则当序列分段熵值在[4.5526,5.2038]区间内时被定义为高熵分段，当序列分段熵值不在这个区间内时则被定义为低熵分段。

4.2 VPN流量检测与识别实验

报文熵值特征提取主要是为了对报文中的有效载荷的高熵低熵字段分布特征进行提取。为保证数长度的一致性方便输入到网络中进行处理，首先使用相同的字节0将报文有效载荷填充至1472B。接着，为便于计算，同样按照32B进行划分，每一段求出一个$|H(P)-H_N(P)|$作为这一段的熵值特征。最终得到46组分段的熵值特征，对于每一段有效载荷还计算一个总熵值特征。最终一共得到47组熵值作为特征。

表1 实验使用的51组特征汇总

	特征名称
	报文方向
特征	TCP ACK
br1	TCP PUSH
br2	报文长度与32的比值
br3	有效载荷总熵值特征（不包括填充字段）
br4	第1段有效载荷熵值特征（32B，包括填充字段）
br5	……
br6	第46段有效载荷熵值特征（32B，包括填充字段）

Br5报文的其他特征包括报文方向、报文长度、TCP Flag中PUSH报与ACK的取值，共计51组特征。本文使用图1所示的网络结构对VPN流量进行识别，输入为表1中提到的51维特征向量。由于VPN流量识别是一个二分类问题，第二胶囊层仅包含两个胶囊，同时多次实验表明，胶囊 长度为16时，分类准确率最高。第一胶囊层与第二胶囊层之间所使用的动态路由算法迭代轮数设置为3轮，在保证算法分类精度的同时减少过拟合的可能性。本文所用到的与胶囊神经网络训练相关的参数如表2所示。

4.3 实验结果

本文在实验室环境下实际采集得到的VPN加密流量数据集与ISCX VPN-nonVPN公开数据集分别进行了实验。其中，实际流量采集使用Wireshark抓取使用V2Ray vmess协议的VPN流量，包含视频播放、网页浏览、文件传输流量等用户行为。

表2 胶囊神经网络参数设置

参数类型	参数	取值
网络参数	K	51
	C_0	256
	S_0	8
	C_1	32
	S_1	16
	C	2
	L	3
训练参数	训练轮数	批大小
	250	10

VPN流量和普通加密流量有效载荷的高熵低熵区域划分使用实际采集的流量进行实验，实验结果如图2所示。由于存在明文头部字段，普通加密流量的有效载荷第一段中的熵值较低，为低熵区域，而之后字段为高熵区域，而VPN流量的有效载荷均为高熵区域，特征较为明显。

使用胶囊神经网络进行VPN流量识别的实验结果如表3所示。

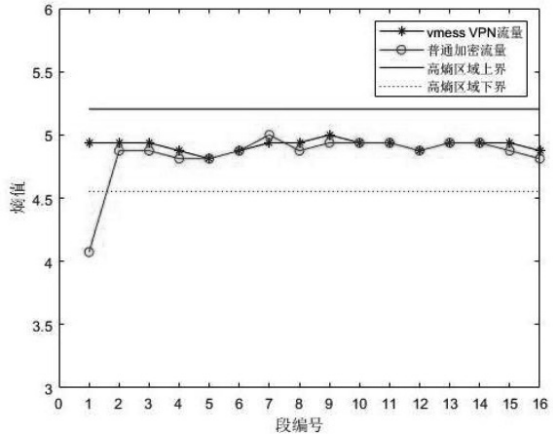

图2 有效载荷熵值分布图

表3 不同数据集上的测试结果

数据集	准确率	精确率	召回率
ISCX 公开数据集	99.87%	99.74%	100.00%
采集得到的VPN数据集	96.34%	93.19%	100.00%

本文还在采集得到的VPN加密流量数据集上使用四种分类模型：随机森林、C4.5决策树、支持向量机、卷积神经网络，进行对比试验。其中支持向量机的惩罚系数C=10，并使用sigmoid核函数；决策树每次选择最优的切分特征和切分点进行切分，最大深度设为8防止出现过拟合的情况；随机森林由64棵前述的决策树构成；卷积神经网络将报文使用随机字段填充或截断生成大小为28×28的灰度图，并使用LeNet5网络进行分类。对比实验结果如表4所示。

表4 不同机器学习方法的测试结果

方法	准确率	精确率	召回率
胶囊神经网络	96.34%	98.19%	100%
随机森林	95.89%	97.52%	94.4%
C4.5决策树	92.95%	97.3%	88.80%
支持向量机	90.21%	95.69%	86.24%
卷积神经网络	95.42%	93.33%	97.39%

实验结果表明，本文提出的基于分段熵分布的VPN加密流量检测与识别方法，在ISCX公开数据集与实际采集得到的VPN数据集上分别达到了99.87%和96.34%的识别准确率，具有极高的识别准确率，同时该结果也优于随机森林、C4.5决策树、支持向量机、卷积神经网络四种分类方法，可以用于VPN加密流量检测识别。

5 结束语

本文内容打破了传统VPN加密流量检测方法的局限性，提出了一种基于分段熵分布的VPN加密流量检测与识别方法，利用滑动窗口方法对VPN加密报文序列高熵、低熵区域进行划分，并使用胶囊神经网络模型实现VPN加密流量的精准检测与识别。对比实验表明，该方法具有更高的识别准确率。

基金项目：

1.国家重点研发计划项目课题（项目编号：2018YFB1800602）；

2.教育部-中国移动科研基金（项目编号：MCM20180506）；

3.赛尔网络下一代互联网技术创新项目（项目编号：NGIICS20190101、NGII20170406）。

参考文献

[1] Velan P. A survey of methods for encrypted traffic classification and analysis[J]. International Journal of Network Management, 2015, 25(5):355-374.

[2] Draper-Gil G, Lashkari A H, Mamun M S I, et al. Characterization of encrypted and vpn traffic using time-related[C]//Proceedings of the 2nd international conference on information systems security and privacy (ICISSP). 2016: 407-414.

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

[4] 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[J]. 计算机应用与软件, 2019, 36(02):321-328.

[5] 谢海涛, 陈树. 基于LSTM的媒体网站用户流量预测与负载均衡方法[J]. 网络空间安全, 2018, 9(10):65-70.

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

[7] Cui S, Jiang B, Cai Z, et al. A Session-Packets-Based Encrypted Traffic Classification Using Capsule Neural Networks[C]//2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). IEEE, 2019: 429-436.

[8] Zhang Y, Zhao S, Zhang J, et al. STNN: A Novel TLS/SSL Encrypted Traffic Classification System Based on Stereo Transform Neural Network[C]//2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS). IEEE, 2019: 907-910.

（下转第33页）

优化算法[J].电子学报,2012,40(03):530-537.

[11] Vojkan Jak□i□. Lectures on Entropy. I: Information-Theoretic Notions[M]// Open Quantum Systems. 2019.

作者简介：

崔弘（1971-），男，汉族，江苏南京人，东南大学，硕士，烽火通信科技股份有限公司，高级工程师；主要研究方向和关注领域：人工智能与大数据分析。

蒋言（1986-），女，汉族，重庆北碚人，华中科技大学，硕士，烽火通信科技股份有限公司，工程师；主要研究方向和关注领域：机器学习与数据挖掘。

郭士串（1988-），男，汉族，江苏徐州人，江西理工大学，硕士，烽火通信科技股份有限公司，工程师；主要研究方向和关注领域：机器学习与数据挖掘。

汪洋（1978-），男，汉族，江苏南京人，东南大学，硕士，烽火通信科技股份有限公司，高级工程师；主要研究方向和关注领域：计算机系统结构。

（上接第27页）

作者简介：

唐舒烨（1996-），男，汉族，江苏无锡人，东南大学，硕士；主要研究方向和关注领域：加密流量测量。

程光（1973-），男，汉族，安徽黄山人，东南大学，博士，东南大学，教授；主要研究方向和关注领域：网络安全、网络测量、加密流量识别、流量行为分析。

蒋泊淼（1998-），男，汉族，湖南长沙人，东南大学，硕士；主要研究方向和关注领域：加密流量测量。

陈子涵（1995-），男，汉族，江苏南京人，东南大学，博士；主要研究方向和关注领域：网络安全、加密流量测量与分析。

郭树一（1998-），女，汉族，江苏南通人，东南大学，硕士；主要研究方向和关注领域：加密流量测量。