学校代码：10286
分 类 号：TP393
密　　级：公开
U D C ：004.9
学　　号：205141

东南大學

SOUTHEAST UNIVERSITY

硕士学位论文

面向多种类型 VPN 流量的识别技术研究

研究生姓名：**张意飞**

导 师 姓 名：**程光教授**

申请学位类别 工 学 硕 士　　学位授予单位 东 南 大 学

一级学科名称 网络空间安全　　论文答辩日期 2023 年 5 月 27 日

二级学科名称　　学位授予日期 20 年 月 日

答辩委员会主席 季一木　　评 阅 人 盲审

盲审

20 年 月 日

東南大學

硕士学位论文

面向多种类型 VPN 流量的识别技术研究

专 业 名 称 ： 网络空间安全

研究生姓名 ： 张意飞

导 师 姓 名 ： 程光教授

本文的部分工作受国家自然科学基金（62172093）的支持与帮助，在此表示感谢。

RESEARCH ON IDENTIFICATION TECHNOLOGY FOR VARIOUS TYPES OF VPN TRAFFIC

A Thesis Submitted to

Southeast University

For the Academic Degree of Master of

Engineering

BY

Zhang Yifei

Supervised by

Prof. Cheng Guang

School of Cyber Science and Engineering

Southeast University

May, 2023

摘 要

随着用户对于匿名化网络访问的需求不断增加，VPN 技术也快速发展。在众多 VPN 技术中，IPSec VPN 和 SSL VPN 凭借高安全性和灵活性占据 VPN 主要市场，成为企业组网和个人用户的首选。但是，许多不法分子将攻击意图隐藏于 VPN 数据包负载中，为我国网络安全部门带来了巨大的监管压力。因此，针对 IPSec VPN 隧道流量和 SSL VPN 隧道流量开展业务类型识别研究，可以提高我国对 VPN 隧道流量的分析能力，强化对 VPN 工具的监管能力。

VPN 的服务器端和客户端形成了端到端的 VPN 通信隧道，多种业务行为流量可以使用同一 VPN 隧道进行数据传输。而现有研究方法只考虑了 VPN 隧道内只有一种业务行为流量的情况，而忽略了 VPN 隧道多路复用的现实情况。嗅探 VPN 隧道流量业务分割点成为解决 VPN 隧道流量业务类型识别的先导性问题。此外，VPN 使用代理机制与加密机制使得 VPN 隧道流量负载随机性增强、头部信息被掩盖，导致构建 VPN 隧道流量的切割特征和业务识别特征具有难度。基于以上研究问题和研究背景，论文提出了一种面向 IPSec VPN 隧道流量和 SSL VPN 隧道流量的业务类型细粒度识别方法。具体而言，包括以下研究内容：

（1）面向多类型 VPN 隧道流量数据集构建方法

针对 SSL VPN 和 IPSec VPN 的端到端通信特征导致难以构建 VPN 隧道流量标签数据集的问题，论文提出了一种基于解密的 VPN 隧道流量标签数据集构建方法。该方法通过 VPN 配置文件的相关信息获取 VPN 隧道流量的数据包解密信息，依据该信息实现 VPN 隧道流量标签数据集的构建。该数据集按业务类型统计，具有即时通信、网页浏览、邮件、音视频通话、文件传输五种业务类型标签；按 VPN 隧道内业务数量统计，具有五种业务、多种业务、一种业务三种类型标签。

（2）基于 VPN 隧道流量的分流方法

针对 VPN 隧道流量无法按照传统的五元组方法进行分流的问题，论文提出了一种 VPN 隧道流量分流方法。该方法构建了 VPN 隧道流量分割特征集合，使用基于队列的训练集构建方法建立标签特征训练集，使用基于滑动窗口的特征提取方法获取待分割 VPN 隧道流量的特征测试集，通过基于不平衡数据集改进的随机森林分割模型获取预测切割点实现 VPN 隧道流量的分流。实验结果表明，该方法对于 IPSec VPN 隧道流量和 SSL VPN 隧道流量均可以获得 90%以上的分割精度。

（3）基于 VPN 隧道流量的业务识别方法

针对 VPN 隧道流量全局特征的混乱程度较大的问题，论文提出了一种基于 VPN 隧道流量的业务识别方法。该方法构建了 SSL VPN 识别序列和 IPSec VPN 识别序列，设计了一种基于注意力机制的 CNN-LSTM 深度学习框架，实现 IPSec VPN 隧道流量和 SSL VPN 隧道流量内多种业务行为类型的识别。实验结果表明，该方法对于 IPSec VPN 隧道流量和 SSL VPN 隧道流量均可以获得 92%以上的识别准确率。

关键词：SSL VPN 加密流量，IPSec VPN 加密流量，加密流量分析，深度学习，随机森林

Abstract

With the increasing demand for anonymous network access from users, VPN technology is also rapidly developing. Among numerous VPN technologies, IPSec VPN and SSL VPN dominate the VPN market with high security and flexibility, and become the preferred choice for enterprise networking and individual users. However, many lawbreakers hide their attack intentions in the payload of VPN packets, bringing huge regulatory pressure to China's network security department. Therefore, the research on business type identification of IPSec VPN tunnel traffic and SSL VPN tunnel traffic can improve China's ability to analyze VPN tunnel traffic and strengthen the regulatory ability of VPN tools.

The server-side and client-side of VPN form an end-to-end VPN communication tunnel, and multiple business activities can use the same VPN tunnel for data transmission. However, existing research methods only consider the situation where there is only one type of business behavior traffic within a VPN tunnel, while ignoring the reality of VPN tunnel multiplexing. Sniffing VPN tunnel traffic service segmentation points has become a leading issue in identifying VPN tunnel traffic service types. In addition, VPN uses proxy and encryption mechanisms to enhance the randomness of VPN tunnel traffic load and mask header information, making it difficult to construct segmentation and business recognition features for VPN tunnel traffic. Based on the above research questions and research background, the paper proposes a fine-grained identification method for business types for IPSec VPN tunnel traffic and SSL VPN tunnel traffic. Specifically, it includes the following research content:

(1) A Method for Constructing Traffic Datasets for Multiple Types of VPN Tunnels

This paper proposes a decryption based method for constructing VPN tunnel traffic label datasets to address the issues of the communication characteristics of SSL VPN and IPSec VPN, which make it difficult to construct VPN tunnel traffic label datasets. This method obtains packet decryption information of VPN tunnel traffic by using relevant information from VPN configuration files, and constructs a VPN tunnel traffic label dataset based on this information. This dataset is classified by business type and has five business type labels: instant messaging, web browsing, email, audio and video calls, and file transfer; According to the number of services in VPN tunnels, there are five types of services, multiple types of services, and one type of service.

(2) A Diversion Method Based on VPN Tunnel Traffic

This paper proposes a VPN tunnel traffic diversion method to address the issue of VPN tunnel traffic being unable to be diverted according to the traditional five tuple method. This method constructs the VPN tunnel traffic segmentation feature set, uses the queue based training set construction method to establish the label feature training set, uses the sliding window based feature extraction method to obtain the feature test set of the VPN tunnel traffic to be divided, and uses the

random forest segmentation model improved based on the unbalanced data set to obtain the predicted cut points to realize the VPN tunnel traffic diversion. The experimental results show that this method can achieve segmentation accuracy of over 90% for both IPSec VPN tunnel traffic and SSL VPN tunnel traffic.

(3) Service Identification Method Based on VPN Tunnel Traffic

This paper proposes a business recognition method based on VPN tunnel traffic to issue the problem of the high degree of confusion in the global characteristics of VPN tunnel traffic. This method constructs SSL VPN identification sequences and IPSec VPN identification sequences, designs a CNN-LSTM deep learning framework based on attention mechanism, and realizes the identification of multiple business behavior types within IPSec VPN tunnel traffic and SSL VPN tunnel traffic. The experimental results show that this method can achieve recognition accuracy of over 92% for both IPSec VPN tunnel traffic and SSL VPN tunnel traffic.

Key words: SSL VPN Traffic, IPSec VPN Traffic, Encrypted Traffic Analyze, Deep Learning, Random Forest

目 录

第一章 绪论.....1
1.1 研究背景与意义.....1
1.2 研究现状.....2
1.2.1 VPN 加密流量识别.....2
1.2.2 VPN 加密流量内容识别.....2
1.3 研究目标与内容.....3
1.3.1 研究目标.....3
1.3.2 研究内容.....3
1.4 论文组织结构.....5
第二章 相关背景技术.....7
2.1 VPN 研究概述.....7
2.1.1 VPN 技术研究.....7
2.1.2 VPN 隧道流量单流特性研究.....8
2.1.3 VPN 加密流量数据集.....9
2.2 VPN 加密流量研究现状.....9
2.2.1 VPN 加密流量识别.....9
2.2.2 VPN 加密流量内容识别.....10
2.2.3 分析总结.....12
2.3 深度学习.....12
2.3.1 LSTM 模型.....12
2.3.2 CNN 模型.....14
2.3.3 注意力机制.....15
2.4 本章小结.....15
第三章 多类型 VPN 隧道流量数据集构建方法.....17
3.1 问题分析.....17
3.2 VPN 通信环境构建方法.....18
3.2.1 通信模型简介.....18
3.2.2 VPN 通信架构.....19
3.2.3 基于 Netfilter 框架的 VPN 代理机制构建方法.....20
3.3 VPN 隧道流量标签数据集构建方法.....21
3.4 实验结果.....23
3.4.1 实验环境.....23
3.4.2 VPN 隧道流量标签数据集.....23
3.5 本章小结.....24

第四章 多类型 VPN 隧道流量分割方法 25
4.1 问题分析.....25
4.2 多类型 VPN 隧道流量识别研究 26
4.2.1 VPN 隧道流量数据格式.....26
4.2.2 基于负载字段的 VPN 加密流量识别方法 28
4.3 VPN 隧道流量切割特征构建方法 29
4.3.1 VPN 隧道流量切割传统方法.....29
4.3.2 VPN 隧道流量切割特征构建.....30
4.3.3 可行性分析.....32
4.3.4 基于队列的训练集构建方法.....33
4.3.5 基于滑动窗口的特征提取算法.....34
4.4 基于数据集改进的随机森林分割模型 36
4.4.1 基于随机森林的 VPN 隧道流量分割算法 36
4.4.2 基于切割训练集改进的随机森林算法 38
4.5 实验设计与结果分析.....40
4.5.1 实验环境.....40
4.5.2 数据集.....40
4.5.3 对比实验选取.....40
4.5.4 评价指标.....40
4.5.5 结果与分析.....41
4.6 本章小结.....44
第五章 面向 VPN 隧道流量的业务识别方法 47
5.1 问题分析.....47
5.2 VPN 隧道流量业务识别序列构建方法 48
5.2.1 面向 SSL VPN 隧道流量的识别序列构建方法.....48
5.2.2 面向 IPSec VPN 隧道流量的识别序列构建方法 50
5.3 面向 VPN 隧道流量的业务识别模型 53
5.3.1 多类型 VPN 识别序列.....53
5.3.2 基于 SSL VPN 识别序列的业务识别模型.....54
5.3.3 基于 IPSec VPN 识别序列的业务识别模型 55
5.4 实验设计与结果分析.....56
5.4.1 实验环境.....56
5.4.2 数据集组成与模型参数设置.....56
5.4.3 对比实验选取.....57
5.4.4 评价指标.....57
5.4.5 结果与分析.....58
5.5 本章小结.....61
第六章 总结与展望 63

6.1 总结.....63
6.2 展望.....64
参考文献.....65

插图目录

图 1-1 研究框架图 4
图 2-1 VPN 架构图 8
图 2-2 RNN 与 CNN 结构图 13
图 2-3 LSTM 神经网络结构图 14
图 2-4 CNN 结构图 14
图 3-1 IPSec VPN 通信模型和 SSL VPN 通信模型图 18
图 3-2 VPN 通信架构图 19
图 3-3 辅助信息变化流程图 21
图 3-4 VPN 隧道流量标签数据集构建流程图 21
图 3-5 VPN 隧道流量标签数据集文件组织形式图 23
图 4-1 多类型 VPN 隧道流量分割方法整体流程图 25
图 4-2 ESP 数据格式图 27
图 4-3 OpenVPN 数据格式图 27
图 4-4 超时值方法图 29
图 4-5 VPN 隧道流量的分割点之间数据包数量频率图 32
图 4-6 长度相对熵图 33
图 4-7 基于队列的训练集构建方法图 34
图 4-8 基于滑动窗口的特征提取算法图 35
图 4-9 决策树算法图 36
图 4-10 基于随机森林的切割算法流程图 37
图 4-11 不同 VPN 隧道流量对应的窗口值-精确率图 42
图 4-12 切割点情况分类图 43
图 4-13 不同业务数量的 VPN 隧道流量精确率图 43
图 5-1 VPN 隧道流量业务识别方法整体流程图 47
图 5-2 SSL VPN 加密流量灰度图 49
图 5-3 FS-E-D 结构图 49
图 5-4 IPSec VPN 加密流量灰度图 50
图 5-5 IPSec VPN 加密流量数据包长度统计特征图 51
图 5-6 基于 SSL VPN 识别序列的业务识别结构图 54
图 5-7 基于 IPSec VPN 识别序列的业务识别结构图 55

表格目录

表 2-1 ISCX VPN-nonVPN 数据集流量构成.....9
表 3-1 VPN 隧道流量业务类型标签数据集.....23
表 3-2 VPN 隧道流量业务数量标签数据集.....24
表 4-1 VPN 隧道流量分流特征表.....31
表 4-2 VPN 隧道流量标签数据集.....40
表 4-3 对比模型各项性能指标结果表.....44
表 5-1 IPSec VPN 隧道流量统计特征.....52
表 5-2 对比模型各项性能指标结果表.....53
表 5-3 VPN 隧道流量标签数据集.....56
表 5-4 VPN 隧道流量切割数据集.....56
表 5-5 VPN 切割流量各项性能指标结果表.....58
表 5-6 VPN 隧道流量各项性能指标结果表.....58
表 5-7 SSL VPN 隧道流量对比模型各项性能指标结果表.....59
表 5-8 IPSec VPN 隧道流量对比模型各项性能指标结果表.....60
表 5-9 消融实验各项性能指标结果表.....60

第一章 绪论

1.1 研究背景与意义

虚拟专用网（Virtual Private Network，VPN）是一种借助公共网络建立点对点通信的专用网络技术。通过 VPN 隧道，在物理地址分布不同的地点可以在共享网络中实现私密、安全的网络互联。在目前的网络流量环境中，VPN 服务的使用者可以分为个人用户和企业用户。由于许多具有互联网限制的国家禁止社交媒体平台、网站、应用程序使用 VPN 服务，这使得越来越来用户使用 VPN 访问受限资源，催化了私有 VPN 工具的泛滥。VPN 被设计之初就是用来为企业组织提供安全隧道连接，为地理位置受限的员工远程接入公司内部服务器提供便利。经过 COVID-19 疫情，使用 VPN 组网的企业、组织和个人用户越来越多，使得 VPN 流量在网络流量的比重增加。根据 AtlasVPN 的统计数据[1]显示，截至 2022 年，85 个受调查国家的 VPN 工具下载总量达到 3.53 亿次。然而，VPN 通过代理机制隐藏用户真实网络地址，使得不法分子可以更好的利用 VPN 加密性强、隐私性高的特点将犯罪行为进行藏匿，躲避网络流量的内容审查和行为监管，或者利用 VPN 攻击表面进行渗透，发起勒索软件、网络钓鱼攻击、拒绝服务和其他过滤关键业务数据等网络攻击，为新时代网络空间安全提出了新的课题，提出一种合理有效的 VPN 内容识别方案成为我国应对 VPN 流量井喷式增长、VPN 服务安全等问题的关键。

随着 Nord VPN 等主流 VPN 工具无法满足用户日益增长的网络需求，越来越多的 VPN 软件提供者通过外部虚拟专用服务器（Virtual Private Server，VPS）构建私有 VPN 工具为用户访问境外信息提供渠道。目前来看，私有 VPN 工具所使用的 VPN 服务方案逐渐与企业组织使用的方案一致，即 IPSec VPN 和 SSL VPN。IPSec VPN 以其高度安全性和可扩展性广受大型企业组织和服务质量较高的 VPN 软件所青睐，而 SSL VPN 以其低成本性和灵活性占据了市场大部分份额。以上两种类型 VPN 在核心网络环境中构建加密隧道，为使用者提供安全隐匿的数据通信服务。其使用 VPN 客户端-VPN 服务端通信模型，通过代理机制重写数据包的 IP 地址和端口以保护用户隐私，增加了 VPN 流量的抗检测性，使得 VPN 通信在核心网中的流量表现为端到端的单流流量。因此，针对 IPSec VPN 加密流量和 SSL VPN 加密流量进行内容识别，可以进一步增强我国流量审查能力。

目前学术界针对 IPSec VPN 加密流量的研究较少，大多数都是基于 SSL VPN 加密场景进行研究。而针对 VPN 加密流量的识别研究，可以分为 VPN 流量识别和 VPN 流量的内容识别。VPN 流量识别主要识别 VPN 流量与非 VPN 流量，或者根据 VPN 所使用的代理协议对 VPN 工具进行区别。VPN 流量内容识别的研究目前较少，研究的场景也比较简单，只考虑一种应用类型在 VPN 隧道流通，并由此构建业务识别、网页浏览请求类型等细粒度识别模型。而实际上 VPN 形成了用户和目标服务器之间的中间人，使用 VPN 客户端可以直接连接 VPN 服务器端[2]，用户收发的所有数据包的 IP 地址在 VPN 隧道内表现为 VPN 代理服务器地址，由此形成了多种业务行为的流量复用同一个 VPN 隧道。因此，构建一种基于 VPN 隧道流量

的业务识别方法对于打击不法分子利用 VPN 工具逃避网络监管具有重要意义，并为 VPN 加密流量更细粒度识别做出了铺垫。

综上所述，随着用户对于匿名化网络访问的需求不断增加，VPN 技术也快速发展。经过 COVID-19 疫情，VPN 流量已经占据加密流量市场相当大的份额。在 VPN 加密流量骤增的背后，许多不法分子将攻击意图隐藏于 VPN 数据包负载中，给网络安全部门带来了很大的网络监管压力。针对 VPN 流量开展业务类型识别研究，探索 VPN 数据包中的潜在信息可以提升网络流量监管水平。为了提高对 VPN 隧道流量的分析能力，加强对 VPN 工具的监管能力，需要对 IPSec VPN、SSL VPN 这两种占据市场最大份额的 VPN 加密流量实现高效且准确的业务类型识别。本文将在 IPSec VPN 加密流量、SSL VPN 加密流量实现流量区别的基础上，提出了一种有效的 VPN 隧道流量的分流方法，实现 VPN 隧道流量的切割，并设计了一种针对 IPSec VPN 隧道流量、SSL VPN 隧道流量的业务行为识别方法实现 VPN 隧道流量的业务类型识别。

1.2 研究现状

随着网络流量加密技术的发展，国内外针对加密流量的识别与检测研究也在逐渐深入。基于非 VPN 加密流量识别研究已经有所进展，而针对 VPN 加密流量的研究需求日益迫切。目前，国内外基于 VPN 加密流量的研究可以分为 VPN 流量识别和 VPN 流量内容识别这两大类。本节将简要介绍 VPN 加密流量识别的相关研究，具体的已有研究内容将在第二章详述。

1.2.1 VPN 加密流量识别

VPN 加密流量识别研究为本文的先导研究。从研究内容上来看，有些学者致力于 VPN 加密流量的检测识别，通过二分类的形式识别原始加密流量中的 VPN 流量；有些研究聚焦于 VPN 工具所使用的代理协议，通过识别不同 VPN 代理协议的流量实现识别不同 VPN 工具。从识别方法上，VPN 加密流量识别可以分为机器学习方法与深度学习方法。机器学习方法的重点在于特征工程，比如构建报文的时间统计特性[3]、DNS 信息特征[4]构建 VPN 加密流量指纹，然后利用机器学习算法，如 C4.5 算法、随机森林算法等机器学习模型学习特征完成流量识别。而深度学习方法无需复杂的特征工程，将流量数据进行预处理后转化为灰度图等合适的数据格式，作为卷积神经网络（Convolutional Neural Networks, CNN）[5]、长短期记忆网络（Long-Short Term Memory，LSTM）[6]等深度学习模型的输入，通过模型学习实现流量识别，其关键点在于构造适合 VPN 加密流量的模型。

1.2.2 VPN 加密流量内容识别

类似于上一节，VPN 加密流量的内容方法也主要分为基于机器学习的分类方法和基于深度学习的分类方法，其各自的本质也与上节相同，即基于机器学习的方法关键在于特征工程、基于深度学习的方法关键在于模型的构建。

（1）基于机器学习的方法

在该类方法中，最常用的特征为流量统计特征。一些学者借鉴针对非 VPN 加密流量的方法通过提取 VPN 加密流量的时间统计特征[3]来区分不同业务类型的 VPN 加密流量的业务类型识别，取得了不错的效果。一些学者使用多种机器学习算法对特征学习，并进行对比[7]。目前，VPN 加密流量具有高度混乱的特性已经成为 VPN 加密流量研究领域共识，基于机器学习的方法难以进一步获得突破。因此，越来越多的学者转向深度学习方法，依靠其特征提取能力，形成端到端的识别方法。

（2）基于深度学习的方法

基于深度学习的 VPN 加密流量内容识别方法与 VPN 加密流量识别方法具有相似性，部分基于深度学习方法的 VPN 加密流量识别方法也可以应用于 VPN 业务与应用识别，取得了很好的效果，比如基于 1D-CNN 的端到端框架[5]、LSTM 模型[6]等。但是，由于前者需要更细粒度的识别，导致深度学习模型结构更复杂或者模型的选择更具有高效性和更强的学习能力，比如 TEST 模型[8]和 FS-Net 模型[9]。相比较于基于机器学习的分类方法，基于深度学习分类方法会有更好识别能力。

结合当前研究现状来看，目前针对 VPN 流量内容识别的研究还较少，研究场景也主要集中于 SSL VPN 加密场景。基于机器学习的 VPN 加密流量内容识别研究将非 VPN 加密流量的统计特征应用于 VPN 加密流量，不能完全表征 VPN 加密流量特性。基于深度学习的 VPN 加密流量内容识别研究都依赖于 ISCX 2016 VPN-nonVPN 公开数据[3]，未考虑多种业务类型流量复用同一 VPN 隧道的情况。因此，需要针对多种类型 VPN 隧道流量进行业务细粒度识别研究。

1.3 研究目标与内容

1.3.1 研究目标

随着国内网络需求的增加，私有 VPN 工具层出不穷，越来越多的企业组织关心 VPN 安全问题。为了提高我国 VPN 流量监管质量，本文将在已有对 VPN 加密流量识别研究的基础上，提出一种面向多种类型 VPN 的业务类型细粒度识别技术，实现对两大主流 VPN——IPSec VPN 和 SSL VPN 隧道流量的业务类型细粒度识别。

为此，本文将研究 IPSec VPN 和 SSL VPN 搭建方法，构建 VPN 安全通信隧道，获取 VPN 隧道流量标签数据集，解决 VPN 隧道流量标签数据集难以构建的问题；解决 IPSec VPN 隧道流量和 SSL VPN 隧道流量难以按照传统方式分流，实现单流 VPN 流量切割；解决 IPSec VPN 隧道流量和 SSL VPN 隧道流量业务识别问题，实现在不解密数据包的情况下 VPN 流量的精细化业务行为类型识别。最终，实现面向 IPSec VPN 隧道流量和 SSL VPN 加密流量的业务识别系统，为我国解决 VPN 加密流量审查提供解决方案。

1.3.2 研究内容

本文的研究框架如图 1-1 所示。首先，本文针对传统 VPN 流量获取方法存在缺陷以及多种业务行为复用同一 VPN 隧道导致难以获取标签数据集的问题，本文提出并实现了面向

IPSec VPN 隧道流量和 SSL VPN 隧道流量的纯净标签数据集构建方法，获取 IPSec VPN 隧道流量和 SSL VPN 隧道流量的标签数据集，以供后续 VPN 隧道流量的业务识别研究使用；针对 IPSec VPN 隧道流量和 SSL VPN 隧道流量表现为端到端的单流流量特性，本文在分析二者数据包负载特性的基础上，提出并设计了一种面向多种类型 VPN 隧道流量的切分方法，实现 VPN 隧道流量的分割；针对 VPN 隧道流量经过加密混淆后头部报文信息改动、全局特征混乱、特征难以获取构建等问题，本文设计了一种面向多种类型 VPN 隧道流量的业务类型识别方法，该方法可以识别 IPSec VPN 加密流量和 SSL VPN 加密流量以及其分割流量的业务类型；最终形成一种面向 IPSec VPN 隧道流量和 SSL VPN 隧道流量的业务类型识别系统。

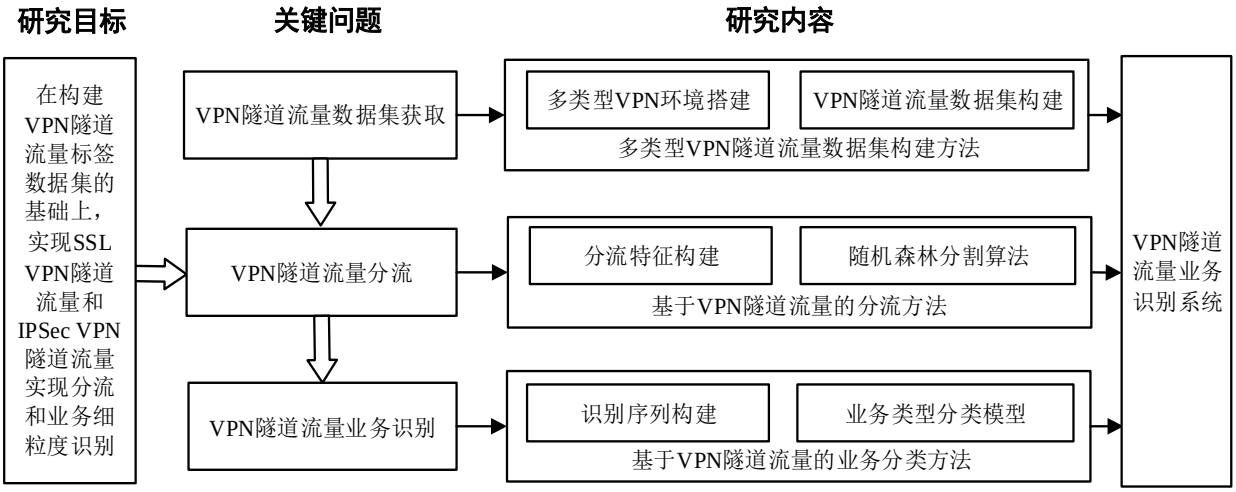


图 1-1 研究框架图

具体而言，可以用以下三点进行诠释：

（1）多类型 VPN 隧道流量的纯净标签数据集构建方法

IPSec VPN、SSL VPN 都使用混淆代理机制，这使得 VPN 隧道流量成为端到端的单流流量。类似多路复用的概念，VPN 隧道流量内可能包含多种应用或者多种业务服务，难以 VPN 隧道流量的标签数据集。针对以上问题，本文提出了一种面向 IPSec VPN 隧道流量和 SSL VPN 隧道流量的纯净标签数据集构建方法，通过自建 SSL VPN 和 IPSec VPN 的配置信息获取加密密钥，使用密钥解密 VPN 数据包，并依据服务器 IP 地址和端口进行过滤，获取具有标签的纯净 VPN 加密流量数据集，以供后面研究使用。

（2）基于 VPN 隧道流量的分流方法

VPN 使用代理机制，整个 IP 包被加密，然后作为另一个 IP 包的内容部分放置，无法使用传统的 5 元组方法进行分流供后续研究。针对 VPN 隧道流量特点，本文提出了一种基于数据包序列的单流 VPN 分流方法。该方法利用固定长度序列数据包序列统计特征，对比相邻序列单元内特征相似性，嗅探 VPN 隧道流量的分割点，实现 IPSec VPN 隧道流量和 SSL VPN 隧道流量分流。该方法具有 92%的切割精确率，克服先前研究提出的超时分割阈值依赖用户习惯的问题，避免超时分割法带来的误判。

（3）基于 VPN 隧道流量的业务识别方法

VPN 加密流量掩盖了流量的内容，其全局特征具有高度混乱性，通过流时序分布测度特征工程寻找不同业务类型的流量粒度特征差异难以区分 IPSec VPN 隧道和 SSL VPN 隧道内不同业务行为的流量。本文针对目前方法难以应用于 IPSec VPN 隧道流量和 SSL VPN 隧道流量的情况，提出了一种基于识别序列的 VPN 隧道流量业务识别方法。该方法构建了 SSL VPN 识别序列和 IPSec VPN 识别序列，设计了一种基于注意力机制的 CNN-LSTM 深度学习框架，实现 IPSec VPN 隧道流量和 SSL VPN 隧道流量内多种业务行为类型的识别。本方法在经过分流的 IPSec VPN 加密流量和 SSL VPN 加密流量上获得了 93%左右的识别准确率，证明了本方法的有效性。

1.4 论文组织结构

本文的组织结构如下所示：

第一章说明了本文的研究背景、研究意义，对研究现状进行了简要概括，阐述了本文的研究目标与研究内容，介绍了本文的论文组织结构。

第二章介绍相关技术背景，简要介绍了 VPN 技术，为第三章构建 VPN 环境做出铺垫，综合分析了目前 VPN 加密流量现状，简要介绍了深度学习相关基础知识，为下文的研究提供了理论基础。

第三章提出了面向多类型 VPN 加密流量业务标签数据集构建方法，在构建 VPN 流量采集环境的基础上，通过对 SSL VPN 和 IPSec VPN 通信原理的分析，设计并实现了 SSL VPN 加密流量和 IPSec VPN 加密流量业务数据集构建方法，获取具有业务分类标签和 VPN 隧道分流标签数据集。

第四章在区分 IPSec VPN 隧道流量和 SSL VPN 隧道流量的基础上，介绍了 VPN 隧道流量切割方法，并根据 VPN 隧道流量切割标签数据集改进分割算法，同时对比多种实验方法证明了本文的方法在此问题上更具有优势。

第五章在分析 IPSec VPN 隧道流量和 SSL VPN 隧道流量全局特征混乱度的基础上，构建了 IPSec VPN 识别序列和 SSL VPN 识别序列，设计了一种基于 VPN 隧道流量识别序列的业务分类模型，对 VPN 切割流量和 VPN 加密流量进行实验。

第六章总结了本文的全部研究，对本文研究方法进行思考并指出本文方法的不足之处，在此基础上对今后的工作进行了展望。

第二章 相关背景技术

本章分为三个部分介绍 VPN 相关技术背景：第一个部分介绍 VPN 技术、VPN 隧道流量特性以及 VPN 加密流量数据集，第二个部分介绍 VPN 加密流量的相关研究，第三个部分介绍深度学习相关的基础知识。

2.1 VPN 研究概述

本节将重点介绍 IPSec VPN 和 SSL VPN 的相关技术，为第三章 VPN 环境的搭建做出铺垫，同时根据 VPN 加密流量的特性提出 VPN 隧道流量的概念，指出 VPN 单流特性的研究难点。

2.1.1 VPN 技术研究

VPN 通过使用安全协议使得公司、组织可以在互联网上实现私人通信[10]。在 VPN 技术应用之初，VPN 以其便利性、低开销、匿名安全的性能[11]被用来为企业员工提供更快、更灵活的网络系统，满足公司日益增长的高效率通信需求，由此也进一步催化了 VPN 技术的发展。此后随着用户需求增加，越来越多定制化 VPN 服务解决网络用户各式各样的需求，其中 VPN 私有工具凭借私密性和匿名性满足用户远程访问需求，逐渐催化出基于各种加密协议与代理协议的 VPN 软件。VPN 技术类型、VPN 技术性能、VPN 服务质量成为当前研究的热议话题。

Zhang 等人[12]简要分析了 MPLS、IPSec 和 SSL 三种不同的 VPN，并从安全性、服务质量、便利性、可扩展性、成本效益、可维护性和算法方面进行比较。该文章指出：IPSec VPN 虽具有高度安全性和可扩展性，但其安装部署困难，无法实现细粒度的访问控制；SSL VPN 虽然不能保证预期的服务质量，但是它在可扩展性、成本效益等方面具有远超 IPSec、MPLS 的优势；MPLS VPN 是三种 VPN 中安全性最低，且其部署、维护都需要较大成本，但它可以提供多个局域网之间最优质的服务和最可观的拓展性。最终得出结论：SSL VPN 是三种 VPN 的最佳选择，具有最高的市场地位。

Hai 等人[13]评估了 SSL、IPSec 和 WireGuard 三种 VPN 技术，并在 VPN 隧道性能和吞吐量方面进行了比较。结果表明，WireGuard 的性能优于其他两种 VPN 技术。但是 SSL 和 IPSec 是创建 WireGuard 技术开发的基础，从算法支持的角度来说，另外两种 VPN 具有的效率可以满足用户需求。此外，还证明了网络层的简单性和规则的破坏会显着影响 VPN 的性能。

Jahan 等人[14]从应用程序的服务要求各不相同出发，对站到站 VPN 协议与远程访问 VPN 协议进行了分析比较。结果表明，在站点到站点 VPN 协议中，GRE 适用于时间敏感和带宽敏感型应用，而 IPSec 更适用于安全敏感应用；在远程访问 VPN 协议中，L2TP IPSec 协议比 PPTP IPSec 协议更可取，对于带宽敏感、时间敏感、安全敏感的应用，L2TP IPSec 是最好的选择。

综合上述研究，IPSec VPN 和 SSL VPN 为 VPN 研究领域的重点。IPSec VPN 以其安全性和扩展性，被大量应用于企业组网，而 SSL VPN 以低成本、高灵活性占据了市场主导地位。

本文中研究对象为 IPSec VPN、SSL VPN，即多类型 VPN 指 IPSec VPN 和 SSL VPN，本文将在第三章从 VPN 加密流量角度分析两种 VPN 的差异性和共通性。以上文献不仅对比分析了 IPSec VPN 和 SSL VPN 的性能、服务，还给出了两者的搭建方法，对本文获取 IPSec VPN、SSL VPN 加密流量作出了铺垫。

2.1.2 VPN 隧道流量单流特性研究

VPN 架构如图 2-1 所示，VPN 服务器应用程序在 VPN 代理环境中接管用户的流量数据。在 VPN 客户端与 VPN 服务器端的链路上捕获网关流量，可以看到用户发送的流量包的目的地址是 VPN 代理服务器，而不是实际的目的地址，实际目的地址在 VPN 客户端被加密处理。在 VPN 代理服务器上，数据包的实际目的地址被解密。然后，代理服务器将这些流量分组发送到实际的目的地地址。这样，VPN 代理形成了用户和目标服务器之间的中间人，用户发送的所有流量包的目的地址都是 VPN 代理服务器。这种现象被称为 VPN 流量的只有一个流问题[15]（the challenge of one flow），本文称以上为 VPN 加密流量的单流特性。

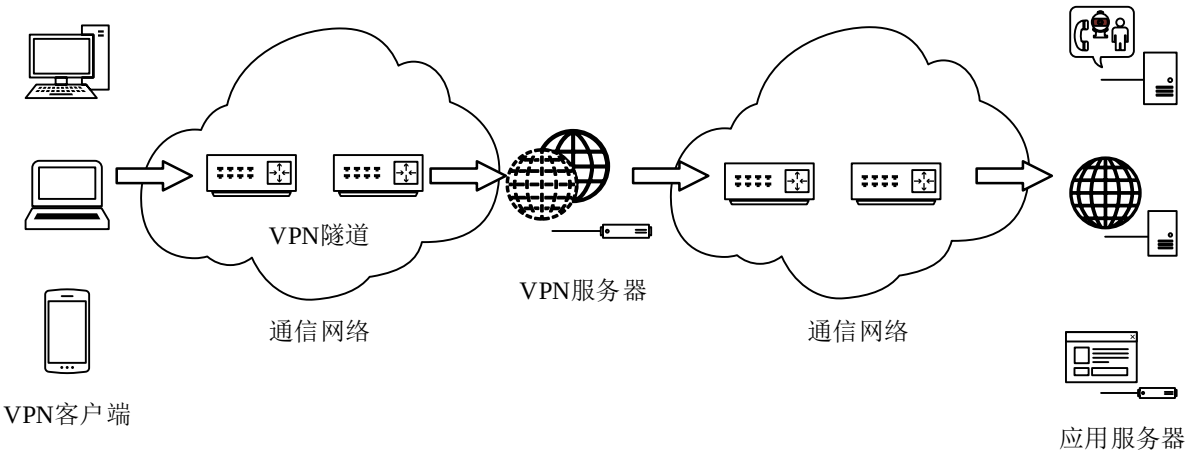


图 2-1 VPN 架构图

传统加密流量可以按照五元组方式进行分流，即按照（源地址，目的地址，源端口号，目的端口号，协议）区分流量[16]。基于不同应用或业务对应不同五元组，可以将原始流量分离。而 VPN 隧道流量只有（源地址，目的地址，协议）三元组，只能定义在 VPN 客户端和服务器端流量方向，不同业务应用可以使用同一 VPN 隧道进行传输，这会产生多种业务应用复用 VPN 隧道流量。因此，VPN 隧道流量的单流特性会导致 VPN 隧道流量多路复用的问题，基于流的流量分类方法由于可分类流量的减少而受到严重影响。本文定义由于多种业务行为复用同一 VPN 隧道而产生的 VPN 加密流量为 VPN 隧道流量。VPN 隧道流量的单流特性给 VPN 加密流量分类带来了巨大的挑战，解决 VPN 隧道流量分流问题成为 VPN 流量识别的先导性问题。

2.1.3 VPN 加密流量数据集

本章介绍 VPN 加密流量领域的经典数据集，近些年该数据集被广泛应用于 VPN 流量识别问题的研究。2016 年，Draper-Gil 等人[3]提出了一个 VPN-noVPN 加密流量数据集，该数据集包括视频、聊天、音频、电子邮件、VoIP、P2P 和文件传输 7 种不同类型的 VPN 加密流量以及 7 种不同类型的非 VPN 加密流量，构成该数据集的应用流量种类如下表所示。本文如无特殊说明，称该数据集为 ISCX 2016 数据集。ISCX 2016 在构建标签数据集时，通过在单位时间在系统内运行一种应用，来获取具有该应用标签的 VPN 流量，其未考虑多种业务类型复用同一 VPN 隧道的情况。因此，需要一种面向 SSL VPN 隧道流量和 IPSec VPN 隧道流量的纯净标签数据集构建方法，来获取 SSL VPN 业务标签数据集和 IPSec VPN 业务标签数据集。

表 2-1 ISCX VPN-nonVPN 数据集流量构成

编号	业务	应用类型或协议类型
1	网页浏览	Firefox、Chrome
2	电子邮件	SMPTS、POP3S、IMAPS
3	在线聊天	ICQ、AIM、Skype、Facebook、Hangouts
4	流媒体服务	Viemo、YouTube
5	文件传输	Skype、Filezilla（FTPS）、Filezilla（SFTP）
6	VoIP	Facebook、Skype、Hangouts
7	P2P	uTorrent、Transmission

2.2 VPN 加密流量研究现状

本节首先介绍 VPN 加密流量识别的先前研究工作，接下来介绍 VPN 加密流量内容识别的先前研究工作，最后对当前研究工作进行总结分析。

2.2.1 VPN 加密流量识别

VPN 加密流量的研究属于加密流量研究范畴。在计算机网络发展早期，许多应用程序使用知名的端口，由此产生了端口匹配的方法。但是随着越来越多应用出现以及端口伪装技术的发展，基于端口号的加密流量研究方法已经不适应当下加密环境。后来诞生了深度数据包检测（Deep Packet Identification，DPI）方法，通过分析数据包相关字段来达到流量解析。VPN 设计的初衷是为用户提供数据机密性和完整性保护，通过建立加密 VPN 隧道，VPN 加密流量可以轻松绕过 DPI 工具的监视来访问受限资源。因此，IPSec、TLS/SSL 等协议的广泛使用使得报文的有效荷载无法解析，导致基于 DPI 的方法无法应用于现实加密场景。

（1）基于机器学习的方法

基于机器学习的 VPN 加密流量识别通过特征工程构建 VPN 加密流量指纹库或获取 VPN 加密流量的统计特征，利用机器学习算法进行特征学习和类型预测。王等人[17]将特定标志位的字符特征与流量数据包长度、流量方向两方面统计特征相结合，使用基于遗传算法改进的随机森林算法对流量进行分类，实验结果表明该方法对于 SSL VPN 流量识别取得了很好的效果。Zain 等人[4]基于 HTTPS 连接中服务器的 IP 地址、用于连接的 TCP 端口、DNS 信息和服

务器名称的连接构建非 VPN 流量指纹库，由此达到对 VPN 流量识别。Luo 等人[18]使用加密代理的 IP 代理和数据加密行为构建 6 维向量来表示一个流，利用主流机器学习算法进行建模，结果表明该方法对识别 V2Ray 流量具有很好的效果。通过分析 VPN 隧道协议的特点，从数据包累计数量、数据包长度、数据包到达时间、标准熵四个方面提取了 11 种特征，使用随机森林算法进行分类，实验结果表明该方法对于 VPN 流量检测识别具有良好效果。周等人[19]根据 IPSec VPN 加密流量的负载特征构建了 IPSec VPN 加密流量识别指纹，实现了 IPSec VPN 加密流量的识别。

（2）基于深度学习的方法

近些年随着深度学习概念的提出以及发展，越来越多的学者开始将深度学习方法应用于 VPN 加密流量识别检测。Wang 等人[5]设计了一种基于一维卷积神经网络的端到端 VPN 检测方法，省略了传统流量识别分类方法中特征设计、特征提取、特征选择等步骤，将网络流量看作一维数组序列，从原始流量中自动学习特征。实验结果表明 1D-CNN 模型在分类 VPN 与非 VPN 加密流量方面有很好的效果。Yao 等人[6]通过递归神经网络（Recurrent Neural Networks，RNN）对网络流量转化为时间序列矩阵，使用长短期记忆网络与分层注意力网络（Hierarchical Attention Network，HAN）相结合实现了 VPN 加密流量检测识别，实验结果表明该方法优于传统机器学习方法。Shapira 等人[20]则借助深度学习图像识别的思想，将 ISCX 2016 数据集中的流量数据转化为 FlowPic 图像再进行 VPN 流量的检测分类，该方法提高了检测准确率的同时，也极大降低了系统内存资源、存储和运行时间。Zhou 等人[21]提出一种基于 2D-CNN 模型的网络流量识别算法，将原始流量转换为灰度图进行识别，该方法在 ISCX 2016 数据集上获得了较高的准确率。唐等人[22]提取 VPN 加密流量的 N-截断熵序列，并构建胶囊神经网络模型对 N-截断熵序列进行学习分类，在 ISCX 2016 数据集中和自建数据集中取得了较好效果。

2.2.2 VPN 加密流量内容识别

目前针对 VPN 加密流量内容识别的研究相比于非 VPN 加密流量的研究还较少，主要集中于 VPN 加密流量的业务类型识别、应用类型识别两个方面。学术界针对 VPN 加密流量的业务类型识别研究和应用类型识别研究基本依赖于 ISCX 2016 公开数据，将针对非 VPN 加密流量的内容识别方法应用于 VPN 加密流量，取得了一些显著的成果。因此，本文在总结分析国内外研究现状时，也会介绍一些基于非 VPN 加密流量的业务识别方法和应用识别方法。

（1）基于机器学习的方法

Draper-Gil 等人[3]使用时间相关特征结合 C4.5 决策树与 KNN 算法对 ISCX 数据集的业务类型进行分类，结果表明 VPN 封装数据包并不会影响流量的时间统计特征信息，时间统计特征对 VPN 业务行为识别仍具有有效性。Bagui 等人[7]在上一篇文献工作基础上，使用了逻辑回归、支持向量机、朴素贝叶斯、K-最近邻、梯度提升树和随机森林 6 种机器学习方法，实现了 VPN 流量业务行为识别，并通过结果表明在处理时间相关性特征方面，集成学习方法具有更高的准确率。Yamansavascilar 等人[23]使用 ISCX 2016 数据集和扩展数据集，对比了 J48、随机森林、K-最近邻和贝叶斯网络四种分类算法，实验结果表明在 KNN 对数据集具有最高

的准确率，而随机森林对扩展数据集具有最高的准确率。Raoul 等人[24]提出了一种新的数据预处理方法--滑动重缩放范围分离（Differentiation of Sliding Rescaled Ranges，DSRR），通过该方法可以提取时间相关特征，结合随机森林算法构建流量分类模型，实现对 ISCX 2016 数据集业务行为识别。Khatouni 等人[25]基于现存的 Tstat[26]、SiLK[27]、Tranalyzer[28]与 Argus[29]四种网络流量特征提取工具，使用支持向量机、K-最近邻和随机森林三种机器学习算法识别加密流量业务类型，同时还证明了通过特征提取、机器学习方式解决加密流量业务识别问题具有泛化性。由此可见，非 VPN 加密流量的统计特征也适用于 VPN 加密流量。但是相比于非 VPN 加密流量，VPN 加密流量全局特征混乱程度更高，手动构造适合 VPN 加密流量内容识别的流量更具有难度。

（2）基于深度学习的方法

随着深度学习方法应用范围越来越广，有越来越多学者开始将深度学习模型应用于 VPN 加密流量研究。前文提到部分基于深度学习方法的 VPN 加密流量识别方法也可以应用于 VPN 业务与应用识别，取得了很好的效果，比如基于 1D-CNN 的端到端框架[5]、基于 HAN 的 LSTM 模型[6]等。使用深度学习方法可以省略了传统流量识别分类方法中特征设计、特征提取、特征选择等步骤，因此，学术界的 VPN 业务识别与应用识别研究重点在于选择合适的深度学习模型。

Song 等人[30]把流量数据每个字节视为一个单词，将 IP 数据包转化为文本图像，利用文本卷积神经网络 Text-CNN 构建业务行为识别模型，通过对比 CNN 模型和 C4.5 模型说明了该方法的优势。Cui 等人[31]首先通过会话分组对原始流量进行分割，并基于分割机制增加有效流量的权重，然后使用胶囊神经网络对转化为灰度图的流量进行分类，在 ISCX 公开数据集中对加密流量业务行为分类获得了较好的效果。Zeng 等人[8]提出了一种 TEST 模型。该模型融合了 CNN 与 LSTM，克服了先前方法中无法同时获取原始流量的时间特征和空间特征，将 ISCX 数据集的加密流变换成灰度图，利用 TEST 模型取得了较好的业务类型识别结果。Feng 等人[32]将网络流量转化为 32*32 的灰度图像，利用轻量级 2 维卷积神经网络构建分类模型，在识别业务行为上取得了较好的结果。Lotfollahi 等人[33]将 CNN 模型与堆叠自编码器（Stacked Auto Encoder，SAE）模型融合，在 ISCX VPN-nonVPN 数据集下可以实现高准确率的业务行为分类效果，同时还证明了 TCP 报文关于建立连接的信息对于业务分类没有帮助。这有助于证明 VPN 流量报文负载的明文信息对于 VPN 业务识别无法提供帮助。Baek 等人[34]将固定长度的数据包字节中导出多种维度图像，利用多输入形状卷积神经网络（Multi Input Shape Convolution Neural Network，MISNN）来生成可用于具有鲁棒性的流量分类模型，实验结果表明相比于将数据包转化为一维图像和二维图像的研究，MISNN 获得了更好的效果。Liu 等人[9]提出了 FS-Net 的端到端加密流量分类模型，该方法使用多层门控循环单元（Gated Recurrent Unit，GRU）对定长双向加密流量序列进行编码，并使用多层 GRU 重建原始序列，然后使用 softmax 分类器识别应用程序，最终在自建数据集上获取了较好的应用程序识别效果。Chen 等人[35]首先获取可用于分类的 PDU 长度序列特征，接下来使用 RNN 提取 N-Gram 特征，最后使用分类层将 N-Gram 输出胶囊转换为最终分类结果，结果证明，该方法相比于

提取全局系列特征的深度学习方法有更好的效果。Huoh 等人[36]考虑了原始字节、元数据和时间顺序关系，使用多层感知机（Multilayer Perceptron，MLP）构建图神经网络（Graph Neural Network，GNN）分类模型，该方法在 ISCX 2016 数据集上取得了比 CNN 和 RNN 更好的效果。Aceto 等人[37]提出了一种 Distiller 分类器，其结合了 RNN 模型和 1D-CNN 模型，使用 PDU 字节序列和报文长度学列作为输入特征，克服现有的基于单模态深度学习的流量分类方案的性能限制，并解决根据不同分类任务产生的模型需求，该方法在 VPN 加密流量检测识别和 VPN 加密流量业务类型识别方面具有很好的效果。

综上所述，针对 VPN 加密流量的业务与应用识别研究主要依靠基于机器学习的方法和基于深度学习的方法。基于机器学习的方法依赖于特征工程，VPN 加密流量相比于非 VPN 加密流量具有更大的特征混乱度，手动构建 VPN 加密流量特征具有相当大的难度，因此，目前效果较好的 VPN 加密流量的业务与应用识别研究大都基于深度学习方法。

2.2.3 分析总结

根据上述对于 VPN 加密流量相关工作的介绍，总结了目前 VPN 加密流量业务识别研究存在以下问题：

（1）针对 VPN 的研究大多数基于 ISCX 2016 公开数据集，该数据集通过在单位时间在系统内运行一种应用，来获取具有该应用标签的 VPN 流量，只考虑了 VPN 隧道内只有一种业务行为流量的情况，现实情况下往往是多种业务复用同一 VPN 隧道，因此需要首先解决 VPN 隧道流量分流问题。

（2）VPN 加密流量相比于非 VPN 加密具有更大的特征混乱度，手动构建 VPN 加密流量特征具有相当大的难度，而深度学习方法可以自动从 VPN 加密流量中进行特征提取、特征学习，广泛应用于 VPN 加密流量分类。而针对 VPN 隧道流量，需要在流量分流后进行业务识别，目前缺少针对 IPSec VPN 隧道流量和 SSL VPN 隧道流量的业务识别方法。

因此，本文在已有研究的基础上，解决以下三个问题：（1）针对 VPN 隧道流量数据集缺失问题，提出一种 IPSec VPN 隧道流量和 SSL VPN 隧道流量数据集构建方法，获取纯净的 IPSec VPN 隧道流量和 SSL VPN 隧道流量数据集；（2）针对 VPN 隧道流量分流问题，提出一种高效的 VPN 隧道流量分流方法；（3）针对 VPN 隧道流量业务识别问题，提出一种基于深度学习的 VPN 隧道流量业务识别方法。

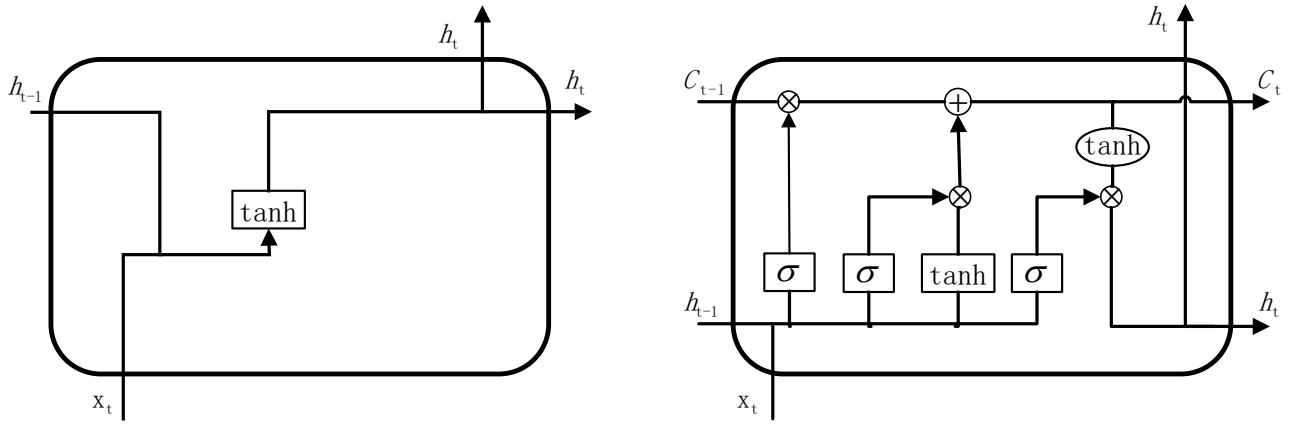
2.3 深度学习

本节主要对本文所使用的深度学习模型以及深度学习方法的相关知识进行简要介绍。

2.3.1 LSTM 模型

循环神经网络最早来源于 1982 年 John 提出的 Hopfield network[38]，其算法模型被广泛应用于处理时序数据。然而，受制于其结构和记忆方式 RNN 暴露出长期依赖、梯度消失以及梯度爆炸的问题，于是 LSTM[39]应运而生，其主要区别如图 2-2 所示。RNN 只传递一种状态h_t，

类似于一种短时记忆，该状态在进行传递时会发生较大改变；LSTM 除了保持h_t状态，还添加了状态C_t，类似于一种长时间记忆，该状态在进行传递时只会发生较小改变。



RNN单元结构　　LSTM单元结构

图 2-2 RNN 与 CNN 结构图

如图 2-2 所示，C_t信息一直在上方传递，h_t信息一直在下方线条上传递，它们通过一些结构进行交互，这些结构被叫做“门”。LSTM 有遗忘门，更新门和输出门三种门结构：

遗忘门：对细胞状态中的信息进行筛选，决定舍弃某些信息，其数学表达式如下所示：

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \qquad (2-1)$$

更新门：更新门的工作分为两个部分，首先确定需要保存的信息，接下来将单元状态C_{t-1}更新为单元状态C_t，其数学表达式如下所示：

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \qquad (2-2)$$

$$C_t = f_t \times C_{t-1} + (1 - f_t) \times tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \qquad (2-3)$$

输出门：基于当前细胞状态，输出一个经过过滤的值h_t，其数学表达式如下所示：

$$h_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \times tanh(C_t) \qquad (2-4)$$

以上描述了 LSTM 一个细胞单元，由此组成了典型 LSTM 神经网络，如图 2-3 所示。LSTM 利用三种门结构进行信息传递和更新，可以很好地从序列数据中学习时间依赖关系。本文将使用 LSTM 模型从 CNN 高维抽象特征中学习局部特征之间的关联，使得模型对 VPN 隧道流量空间和时间特征产生敏感性。

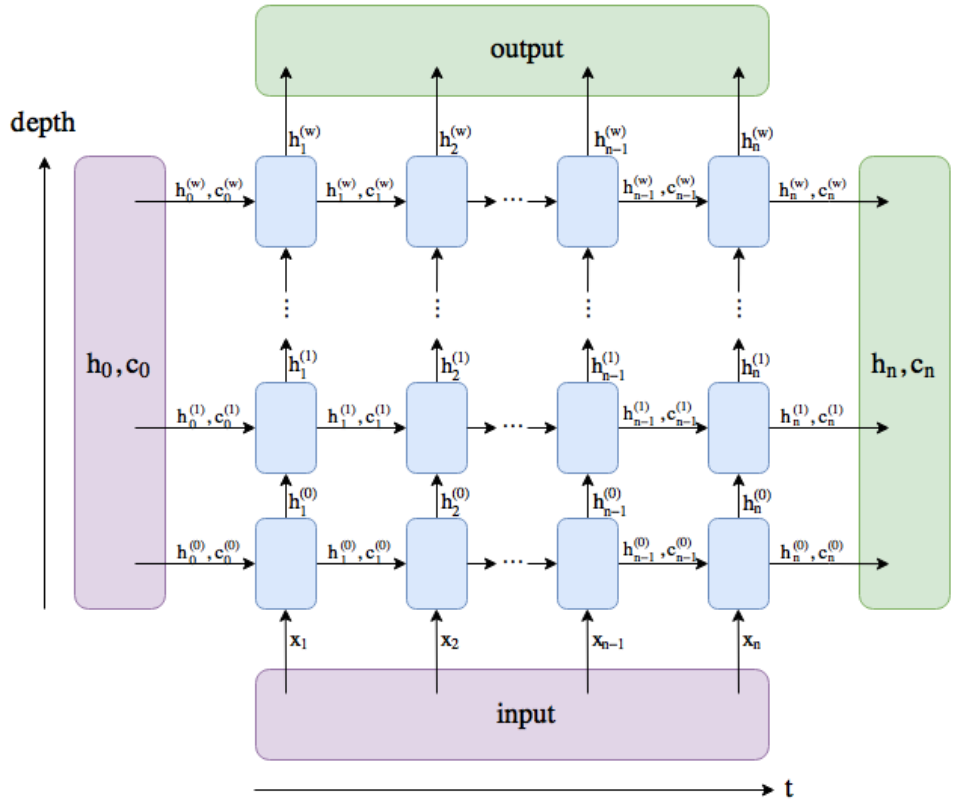


图 2-3 LSTM 神经网络结构图

2.3.2 CNN 模型

CNN[40]作为一种经典的深度学习算法，被广泛应用于图像识别领域。越来越多学者通过分析流量全局特征，将 CNN 算法应用于流量识别领域。典型 CNN 结构如图 2-4 所示，可以分为卷积层、激励层、池化层和全连接层。

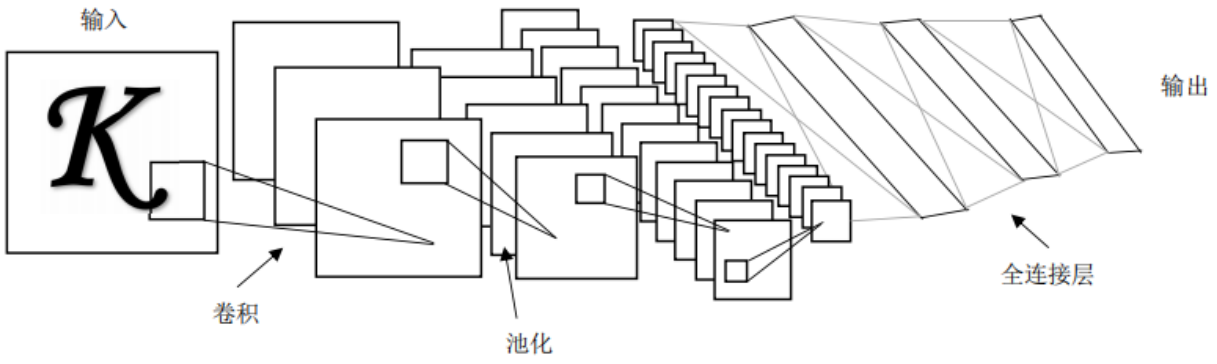


图 2-4 CNN 结构图

卷积层：借鉴人脑在识别图像时从局部特征开始学习，卷积层使用卷积核来学习一定范围内的局部特征，通过卷积核参数共享保持特征提取方式的一致。：

激励层：通过激励函数对卷积层输出引入非线性因素，降低模型的线性能力，提高模型的表达能力，可以更好地抽象高维特征。常用激励函数有 ReLU 函数、Leaky ReLU 函数、

Maxout 函数以及 Tanh 函数等。通过卷积层和激励层，输入数据M可以转化为一组特征图M'，其数学过程如下, 其中，w为卷积核，b为偏置函数，f为激励函数：

$$M' = f(w \cdot M + b) \quad (2-5)$$

池化层：抽取卷积后的特征向量图的显著特征，降低特征降维度，压缩参数数量，同时有利于减少过拟合情况的发生。比较常见的池化操作有最大池化（Max Pooling）和平均池化（Average Pooling）。本文将使用最大池化操作，其数学过程如下：

$$M'' = max[M_1', M_2', M_3' \dots M_k'] \quad (2-6)$$

其中，M_k'为特征图中的单位特征矩阵，通过选择该矩阵中最大值形成新的特征图M''。

全连接层：位于 CNN 模型的最末端，从原始输入中学习到隐层特征转化为分类标签集合，达到分类或识别的目的。

本文借助 CNN 模型从 VPN 隧道流量中找寻潜在特征关系，减少噪声干扰并持续学习稳定关系的特征信息，并将该信息输入 LSTM 模型进行预测。

2.3.3 注意力机制

2014 年，Mnih 等人[41]提出了注意力机制。注意力机制借鉴了人脑的视觉注意力机制思想，即人类在观看一张图片时往往会锁定一些重要区域并忽略其他无关区域，可以从显著特征中实现学习目的。

注意力机制本质上就是对模型的输入序列赋予权重，以此表达模型对该输入的关注程度。注意力机制可以分为硬注意力机制（hard attention）、软注意力机制（soft attention）。硬注意力机制为输入序列分配权重 0 或 1，即对于有些输入序列完全不关注，但是由于网络流量的输入序列均包含一定信息，无法完全摒弃某一个子序列，因此硬注意力机制并不适合处理网络流量。对于软注意力机制而言，其权重分配位于 0 到 1 之间，考虑了输入序列的大部分特征，更加适合网络流量。此外，软注意力机制在反向计算时具有可微性，可以通过端到端的方式直接学习输入序列特征。

本文将使用软注意力机制，将其连接于 LSTM 模型之后，用于强化 LSTM 模型中重要时间步，减少预测误差发生的概率。

2.4 本章小结

本章主要介绍了本文研究的相关背景技术，主要包括了 VPN 技术、VPN 隧道流量特性、VPN 加密流量识别、VPN 加密流量内容识别、深度学习基础知识。在总结先前研究工作基础上进行分析总结，指出了当前研究盲点，并由此导向本文研究工作，为第三章 VPN 隧道流量数据集构建、第四章 VPN 隧道流量切割、第五章 VPN 隧道流量业务识别提供了理论支持和方法启迪。

第三章 多类型 VPN 隧道流量数据集构建方法

构建 VPN 隧道流量标签数据集是研究 VPN 隧道流量业务识别问题的先导性条件，从第二章相关背景技术的介绍可以看出，目前对 VPN 加密流量的研究都是基于 ISCX 2016 数据集，缺少高质量的 IPSec VPN 隧道流量标签数据集和 SSL VPN 隧道流量标签数据集。本章在分析 VPN 通信模型的基础上，设计一种基于解密的 VPN 隧道流量业务标签数据集构建方法，为下文的 VPN 隧道流量业务识别问题提供数据支撑。

3.1 问题分析

目前对于 VPN 的研究大都聚焦于 SSL OpenVPN。2016 年，Draper-Gil 等人[3]通过 OpenVPN 软件构建了 SSL VPN 数据集，该数据集包含了 7 种业务类型的 VPN 流量与非 VPN 流量，其数据标签为构成 7 种业务的应用类型，为现在学者研究 VPN 流量提供了研究基础。ISCX 2016 在构建标签数据集时，通过在单位时间在系统内运行一种应用，来获取具有该应用标签的 VPN 流量。然而，这种数据集构建方法有显而易见的错误与缺陷：第一，在单位时间内，操作系统也会发送一些请求获得一些相应，比如 Ubuntu Linux 会与 ubuntu.com 服务器进行数据交互、在规定时间内发送 NTP 报文，这会导致数据集中含有与标签流量无关的背景流量，影响后续 VPN 加密流量业务识别结果；第二，同一种应用可能包含不同的业务行为，比如使用浏览器同时进行在线聊天和网页浏览时，二者流量都使用了同一个浏览器。以上缺陷会导致 VPN 隧道流量标签数据集纯度不够，严重影响后续研究。因此，构建高质量 SSL VPN 数据集和 IPSec VPN 数据集成为研究 VPN 加密流量识别先导性问题。

此外，相比较于非 VPN 流量，普通加密流量可以按照五元组方法获取流量对应的标签，而 VPN 具有代理混淆的特性，具有不同业务行为的应用可以使用同一 VPN 隧道，导致 VPN 隧道多路复用的问题，难以通过非 VPN 加密流量标签数据集获取方法构建数据集。位于 VPN 隧道内的流量因为 VPN 加密协议增加了数据包包长导致进入隧道内的流量产生了分片，隧道出入口流量与隧道内流量不具有一致性，从 VPN 隧道流量中剥离出不同业务行为的 VPN 加密流量具有一定难度。

基于上述问题，本章工作如下：

（1）设计一种站点到站点的 VPN 通信结构，分别使用 strongswan 搭建 IPSec VPN 隧道、使用 OpenVPN 搭建 SSL VPN 隧道，并使用 Netfilter 框架实现内网主机通过 VPN 隧道访问应用服务器。

（2）针对目前 IPSec VPN 隧道流和 SSL VPN 隧道流量标签数据集缺失、目前 VPN 加密流量数据集构建方法存在缺陷以及 VPN 隧道多路复用的问题，本章提出了一种面向 IPSec VPN 隧道流量和 SSL VPN 隧道流量的标签数据集构建方法，获取 IPSec VPN 隧道流量和 SSL VPN 隧道流量标签数据集。该数据集可以构成 VPN 隧道流量切割标签数据集用于预测待分割 VPN 隧道流量分割点，也可以构成 VPN 加密流量业务类型标签数据集用于 IPSec VPN 加密流量和 SSL VPN 加密流量业务行为识别。

3.2 VPN 通信环境构建方法

本节首先对 IPSec VPN 通信模型和 SSL VPN 通信模型进行解析，分析架构一致性，在此基础上提出基于 Netfilter 代理机制的 VPN 搭建方法，完成 VPN 环境构建。

3.2.1 通信模型简介

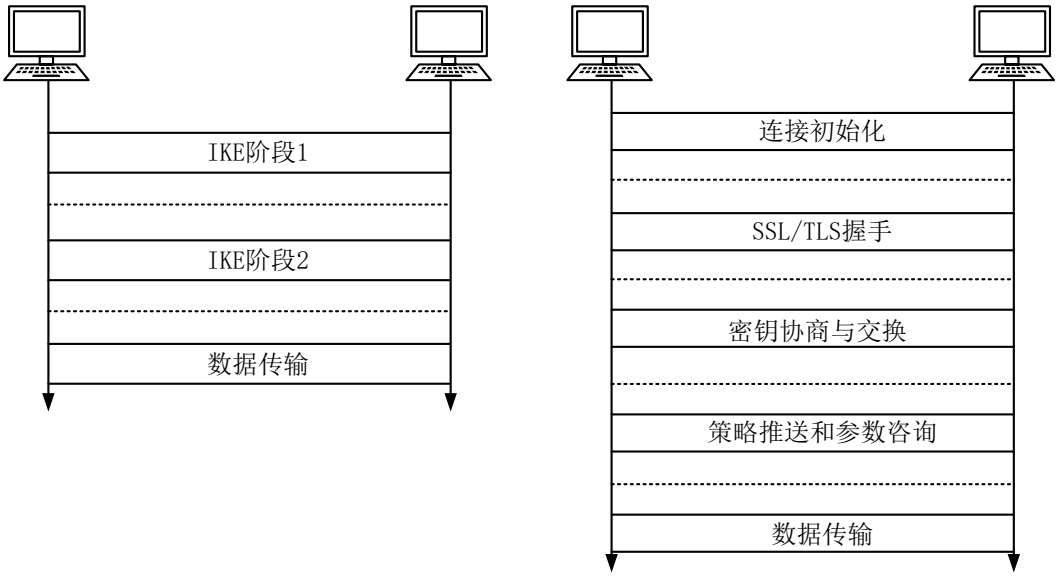


图 3-1 IPSec VPN 通信模型和 SSL VPN 通信模型图

IPSec VPN 是一种基于内核空间的 VPN，通常适用于 Site to Site（站点到站点）的组网。IPSec VPN 的工作模式分为隧道模式和传输模式，前者将整个 IP 数据包作为加密负载，后者则将 IP 数据包负载信息进行加密。相比于传输模式，隧道模式具有更高地加密性能。本文选取隧道模式下 IPSec VPN 获取 IPSec VPN 隧道流量。IPSec VPN 要求站点分别部署 VPN 网关或远程用户安装专用的 VPN 客户端，其通信过程如下所示：

（1）IKE 阶段 1：客户端向服务器发送 IPSec 组策略提议、交换 DH 算法的公开信息、验证彼此身份，使得通信双方间建立了一个身份信任和通信安全的通道，建立了一个 ISAKMP SA；

（2）IKE 阶段 2：双方协商 IPSec 安全参数，两台主机协商在会话中使用的加密算法类型，并交换双方计划用于进出流量的加密和解密密钥，建立一个供数据传输的安全通道，此阶段的交换生成了 IPSec SA；

（3）数据传输阶段：通过新创建的 IPSec 加密隧道交换数据，之前设置的 IPSec SA 用于加密和解密数据包。

SSL VPN 大都基于用户空间，本文选取常用 SSL VPN 软件 OpenVPN 进行研究分析。OpenVPN 的工作模式分为隧道（Tunnel，TUN）模式和终端访问点（Terminal Access Point，TAP）模式。前者通过虚拟网卡处理 IP 数据包实现点对点数据信息交互，常用于实现点对点的 VPN 隧道；后者类似于以太网桥设备，通过虚拟网卡处理数据链路层数据包，实现一点对

多的 VPN 隧道。相比于 TAP 模式，TUN 模式具有更好的流量控制性能，因此，本文研究的 SSL VPN 加密流量通过 TUN 模式下的 OpenVPN 进行捕获。其通信过程如下：

（1）连接初始化：执行连接初始化，服务器为来自客户端的连接设置数据结构；

（2）SSL/TLS 握手：通过基于 OpenVPN 数据包携带 SSL/TLS 握手信息，建立 SSL 安全连接，为下一阶段形成一个安全隧道；

（3）密钥协商与交换：通过阶段 2 的安全隧道为 OpenVPN 记录协议协商和交换密钥；

（4）连接建立阶段 4：实现两端 OpenVPN 两端参数协商，服务器向客户端提供推送；

（5）数据传输阶段：通过创建的加密隧道交换数据。

IPSec VPN 通信模型和 SSL VPN 通信模型对比如图 3-1 所示，可以看出 IPSec VPN 和 SSL VPN 都通过一定的密钥协商、参数协商建立了一条端到端的 VPN 隧道，使得数据传输可以在一条安全可靠的隧道上进行。因此，二者通信架构具有一致性，本文研究对象为位于隧道内的 IPSec VPN 隧道流量和 SSL VPN 隧道流量。

3.2.2 VPN 通信架构

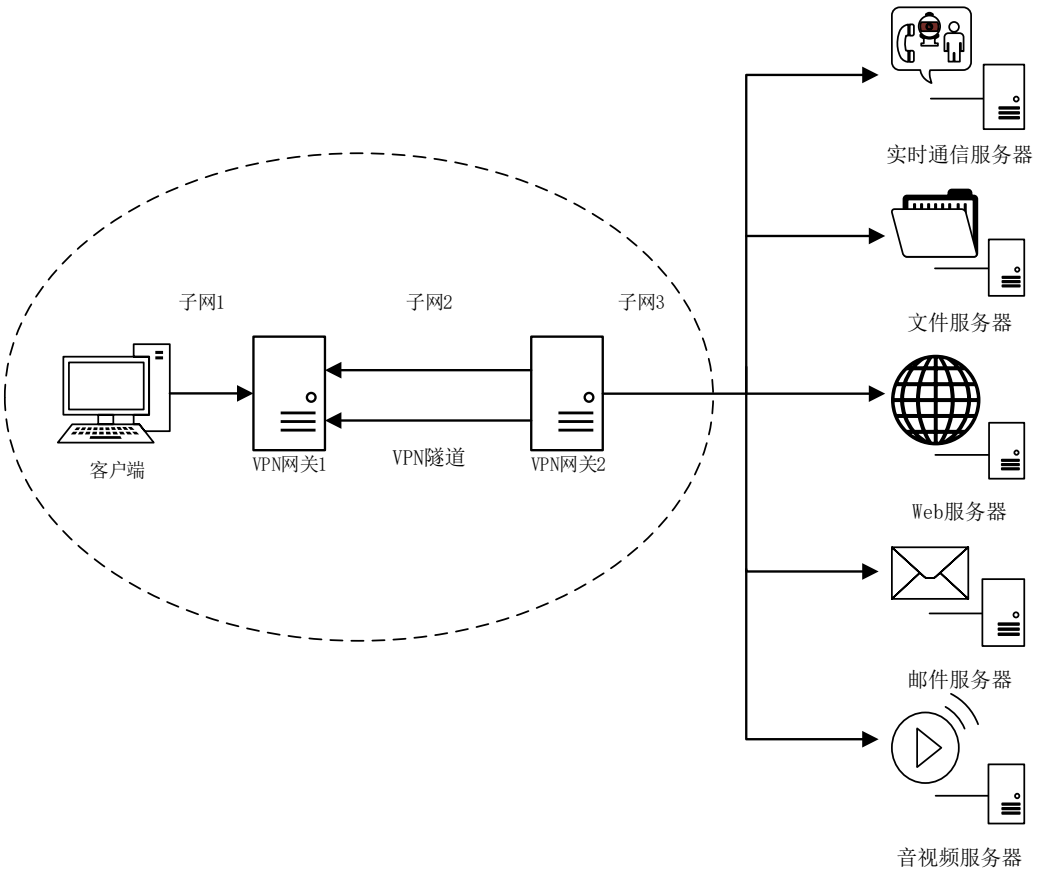


图 3-2 VPN 通信架构图

通过上述两种 VPN 通信模型可以看出，VPN 客户端将来自用户的每一个数据包先发送到 VPN 服务器端，而不是公网应用服务器，VPN 软件隐藏了用户的真实目的 IP 和端口号。数据包到达 VPN 服务器后，VPN 软件会对数据包进行解密，并获得实际的目的地 IP 和端口。接下来，VPN 服务器会将解密后的数据包传输到用户的服务器。从公网应用服务器的角度来

看，它会将 VPN 代理视为实际的客户端，永远不知道用户身份，体现了 VPN 代理功能。VPN 通过加密协议、密钥协商、身份认证实现 VPN 安全加密隧道，具体加密特性依赖于 VPN 加密套件。

基于以上 VPN 架构的思路，本文 VPN 实际架构环境如图 3-2 所示。为了更好的研究 VPN 加密协议和代理机制，本文将 VPN 两大核心功能分离，在 VPN 隧道出口网关设置代理服务器，使得两者功能分离。在设置内网链路环境时，不考虑网络会产生拥塞的问题，每一个网段设置为 0ms 延迟、0%丢包率，每一个网络适配器的带宽不受限。通过一些加密策略配置，可以在代理服务器 A 和代理服务器 B 之间建立 VPN 隧道，客户端请求服务器资源的流量经过 VPN 客户端和 VPN 服务器端的身份混淆和 VPN 加密策略可以形成 VPN 隧道流量。

对于 IPSec VPN，本文使用 strongswan 开源软件对内核 IPSec 进行配置，因特网密钥交换协议版本为 2，使用封装安全载荷协议协议作为数据传输阶段的加密协议；对于 SSL VPN，本文使用 OpenVPN 软件进行搭建，底层协议使用 TCP。

3.2.3 基于 Netfilter 框架的 VPN 代理机制构建方法

本文所使用的 Netfilter 框架[42]基于 5.10.4 Linux 内核版本，可以在 Linux 内核中实现数据包修改、端口过滤、连接追踪等功能，其工作在网卡驱动程序和 Linux 协议栈之间。在进行 IP 数据包的处理过程中，Netfilter 框架具有 NF_INET_PRE_ROUTING、NF_INET_LOCAL_IN、NF_INET_FORWARD、NF_INET_LOCAL_OUT、NF_INET_POST_ROUTING 五个钩子点，可以进行编程来设置数据包内容、过滤匹配数据包。

为了实现内网主机可以通过 VPN 隧道访问公网服务器，那么需要在 VPN 网关 1 和 2 处进行流量重定向。VPN 网关 1 可以接受源地址为子网 1、目的地址为子网 3 的流量进入 VPN 隧道，并对流量进行地址伪装；也可以接受源地址为子网 3、目的地址为子网 1 的流量进入 VPN 隧道并进行地址伪装。但是子网 1 想通过 VPN 隧道进行访问外部网络服务器请求时，那么就需要在 VPN 网关 1 修改目的地址同时保存原有目的地址信息，在通过 VPN 隧道后需要在 VPN 网关 2 依据原有目的地址信息修改数据包目的地址同时修改源地址为 VPN 网关地址，作为内网 1 流量的转发服务器。因此，VPN 网关 2 承担了 VPN 隧道通信和流量转发服务器两种角色。

具体而言，本方法在原有拓扑基础上进行如下数据包，使用内网主机真实通信意图作为辅助标签信息传递于 VPN 隧道之间。该标签信息可以使得 VPN 网关具有代理功能，实现了 IP 数据包转发，但是也扩展了数据包长度，这会导致以下问题：（1）对于 TCP 数据传输而言，其 ack 和 seq 与数据包长度相关，修改了数据包会导致 seq 和 ack 发生变化，不处理则会导致乱序问题；（2）扩展 IP 层 Payload 会使得传输层检验发生变化，对于检验和错误的数据包网关将会舍弃；（3）修改 IP 层目的地址、源地址会使得 IP 层校验和发生变化，需要重新计算校验和。对于问题（1），netfilter_helper 提供了连接追踪方法，绑定 payload 修改前后数据包的 seq 和 ack，而对于问题（2）和（3），netfilter 也提供了 IP 层和传输层校验和计算修改函

数。因此，基于 netfilter 框架的 VPN 构建方法在上述 VPN 拓扑结构下数据包传输工程中辅助信息变化流程如图 3-3 所示。

图 3-3 辅助信息变化流程图

对于发包过程，主机修改发送数据包目的地址为代理服务器地址，并把原目的地址作为辅助信息添加到 IP 层 payload 中。VPN 客户端收到携带辅助信息的发送数据包对数据包进行隧道封装传输到 VPN 服务器端。VPN 服务器端对携带辅助信息的发送数据包解封装，并将数据包重定向至代理服务器。代理服务器获取发送数据包的辅助信息，修改发送数据包源地址和目的地址，并将发送数据包的负载还原。对于收包过程，代理服务器修改源地址和目的地址，并把源地址作为辅助信息添加到 IP 层 payload 中。VPN 服务器端收到携带辅助信息的发送数据包对数据包进行隧道封装传输到 VPN 客户端。VPN 客户端对携带辅助信息的发送数据包解封装，并将数据包发送至主机。主机获取发送数据包的辅助信息，修改发送数据包源地址和目的地址，并将发送数据包的负载还原。

3.3 VPN 隧道流量标签数据集构建方法

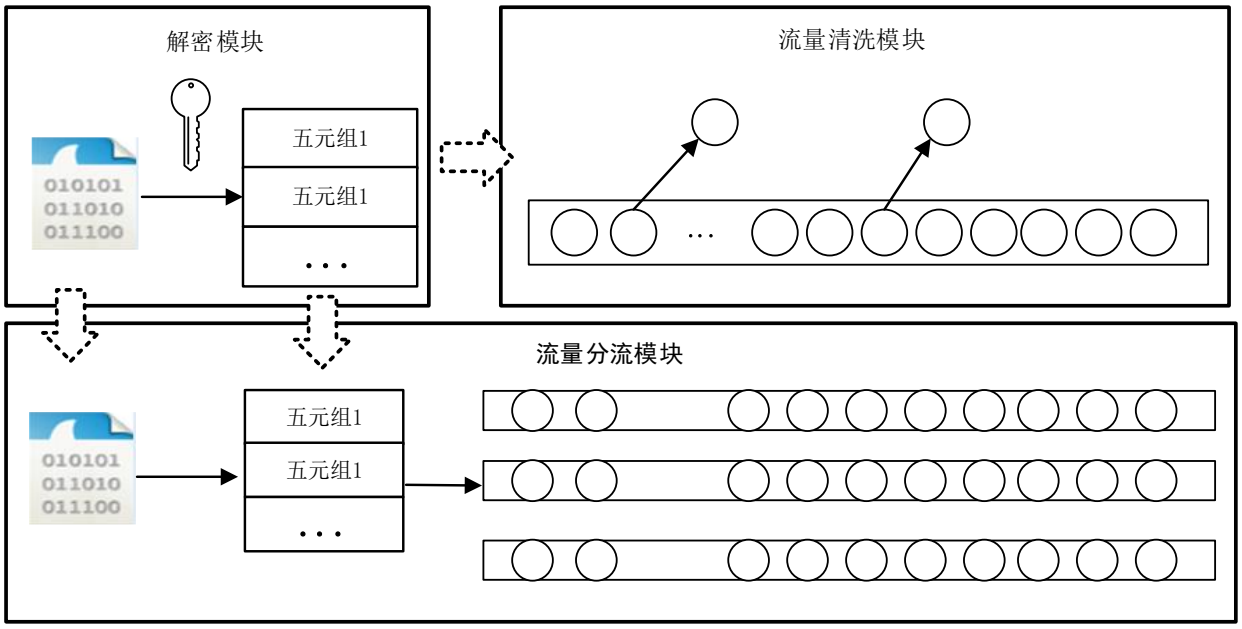


图 3-4 VPN 隧道流量标签数据集构建流程图

使用 3.2.2 节搭建的 SSL VPN 和 IPSec VPN 获取可以分别获取 SSL VPN 加密流量和 IPSec VPN 加密流量。VPN 代理网关形成了用户和目标服务器之间的中间人，用户与服务器通信的 IP 地址为 VPN 代理网关的地址，不同业务行为的流量可以使用同一 VPN 隧道由此导致了 VPN 隧道多路复用的问题。本文发现 VPN 隧道出口的流量即解封转的流量与隧道内流量即封装流量其数据包个数和顺序并不一致，原因有以下几点：（1）SSL VPN 和 IPSec VPN 都会增加数据包长度，可能会超过网卡的 MTU 值，因此导致数据包被分离再传输，在隧道出口解封装时被还原；（2）SSL VPN 位于用户空间，IPSec VPN 位于内核空间，其有一些加密策略或通信策略导致数据包可能会产生分片。除此之外，隧道出口流量和隧道内流量的数据包顺序也难以对应，因为无法保证同时获取第一个数据包。基于以上问题，本文提出了一种基于解密数据包的数据集构建方法，通过解密 IPSec VPN 通信数据包和 SSL VPN 通信数据包，构建 IPSec VPN 加密流量业务标签数据集和 SSL VPN 加密流量业务标签数据集。对于两种 VPN 流量，本文通过配置文件获取其密钥，构建解密模块获取 VPN 流量的五元组，剔除背景流量，并按照五元组进行业务分割，构建 VPN 加密流量数据集，该方法如图 3-4 所示。

解密模块：对于 IPSec VPN 加密流量来说，其解密可以根据数据包的 SPI 字段获取加密方式、加密 key、认证方式、认证 key，参考 Wireshark 解密源码构建本章 IPSec VPN 解密模块；对于 SSL VPN 加密流量来说，通过获取配置静态密钥以及加密模式可以进行 OpenVPN 数据包解密。通过解密模块获取 VPN 隧道数据包封装的源 IP 地址、目的 IP 地址、源端口号、目的端口号、协议的五元组副本，并对数据包进行编号。

流量清洗模块：该模块通过分析五元组副本删除背景流量数据包，首先判断每个五元组副本是否为背景流量数据包，接下来将删除的数据包编号进行存储，根据删除编号依次删除 VPN 隧道流量内背景流量数据包，最后更新数据包编号。

流量分流模块：该模块依据五元组副本信息对 VPN 隧道流量按业务行为类型进行分流，首先标记每个五元组副本所属业务类型，接下来按数据包编号和五元组副本将数据包进行归类，将不同业务行为的 VPN 加密流量分开放置，将同属于一个隧道的不同业务行为的 VPN 加密流量放置于同一个 pcap 文件内，并采用相同命名，最终获得 IPSec VPN 加密流量业务标签数据集和 SSL VPN 加密流量业务标签数据集。

本章通过以上方法构建 VPN 加密流量简要信息如下表所示。值得注意的是，该数据集同时可以用来构建 VPN 切割标签数据集，这是因为在进行分流时保留了 VPN 隧道信息，在流量分流模块中，将同属于一个隧道的不同业务行为的 VPN 加密流量形成的 pcap 文件命名相同，由此可以表明在不同业务流量文件夹下使用同一隧道的 VPN 加密流量。使用该文件组织形式可以方便构建 VPN 隧道流量的切割标签训练集，具体构建方法于 4.5 节阐述。

3.4 实验结果

3.4.1 实验环境

本章实验在物理机上部署实验代码，物理机的 CPU 为 AMD Ryzen 7 3700X 8-Core Processor 3.60 GHz，内存为 64GB，使用的操作系统及版本为 Windows 10 专业版，使用的虚拟机软件为 VMware Workstation Pro 16.2.3，虚拟机为 Ubuntu 18.04.3 LTS。

3.4.2 VPN 隧道流量标签数据集

本章最终捕获的标签数据集的文件组织形式如图 3-5 所示。该数据集分为 SSL VPN 隧道流量和 IPSec VPN 隧道流量，每种 VPN 隧道流量对应 5 种业务类型，每种业务类型流量文件夹下存储 VPN 隧道流量 pcap 文件，每个 pcap 文件命名标号表明其所属的 VPN 隧道，属于同一 VPN 隧道的不同业务类型流量具有相同的标号。

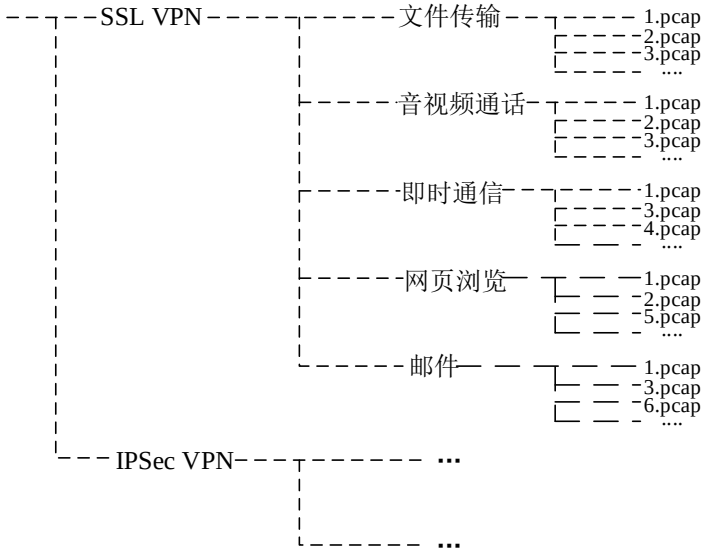


图 3-5 VPN 隧道流量标签数据集文件组织形式图

VPN 隧道流量数据集按照业务类型进行统计，如表 3-1 所示。SSL VPN 隧道流量和 IPSec VPN 隧道流量各具有即时通信、网页浏览、邮件、音视频通话、文件传输五种业务类型标签，每种业务类型具有 400 条流。不同业务类型流量的数据来源不仅包含当下热门应用，还包括自建应用服务器，由此表明本文研究的 VPN 业务类型覆盖全面。

表 3-2 VPN 隧道流量业务类型标签数据集

编号	业务类型	流量来源	数量
1	即时通信	微信、QQ、自建聊天服务器	400
2	网页浏览	自建网站、Edge	400
3	邮件	自建邮箱服务器	400
4	音视频	自建服务器、腾讯会议	400
5	文件传输	百度网盘、自建文件传输服务器	400

VPN 隧道流量数据集按照同一 VPN 隧道包含的业务数量进行统计，如表 3-2 所示。SSL VPN 隧道流量和 IPSec VPN 隧道流量各具有包含五种业务类型、多种业务类型和一种业务类型的流。本文将在后续章节中对其进行讨论分析。

表 3-2 VPN 隧道流量业务数量标签数据集

编号	业务类型数量	SSL VPN 隧道流量	IPSec VPN 隧道流
1	五种	150	150
2	多种	150	150
3	一种	100	100

3.5 本章小结

VPN 通信相较于传统基于客户端-服务器的通信更加复杂，使用代理机制掩盖了用户地址。本章在分析 IPSec VPN 和 SSL VPN 通信原理的基础上，为 IPSec VPN 和 SSL VPN 设计了同一的通信架构，并采用不同的通信软件分别进行 IPSec VPN 隧道流量和 SSL VPN 隧道流量采集；在分析 IPSec VPN 隧道流量和 SSL VPN 隧道流量的基础上，设计了一种基于解密的 VPN 隧道流量标签数据集构建方法，通过 IPSec VPN 和 SSL VPN 的控制信息来获取 VPN 隧道流量的五元组信息，并依据五元组信息对 VPN 隧道流量进行清洗和分流，最终得到了 IPSec VPN 隧道流量标签数据集和 SSL VPN 隧道流量标签数据集，为后续的研究做出了数据支持。

第四章 多类型 VPN 隧道流量分割方法

本章提出的多类型 VPN 隧道流量切分方法为基于机器学习的方法，需要构建特征训练集和待分割 VPN 隧道流量的特征测试集，使用特征训练集构建用于切分待分割的 VPN 隧道流量，使用特征测试集确定待分割隧道流量的分割点，最终实现分割。本章方法主要分为三个阶段：第一阶段为多类型 VPN 隧道流量识别，使用 VPN 隧道的负载特征实现 IPSec VPN 隧道流量和 SSL VPN 隧道流量的区分；第二阶段为多类型 VPN 隧道流量特征提取，使用基于队列的 VPN 隧道流量训练集构建方法和基于滑动窗口的特征提取算法分别得到标签训练集和特征测试集；第三阶段为多类型 VPN 隧道流量模型构建与模型预测，通过标签训练集实现模型结构初始化，模型使用特征测试集预测分割点完成最终待分割隧道流量切割。该分割流程如图 4-1 所示。

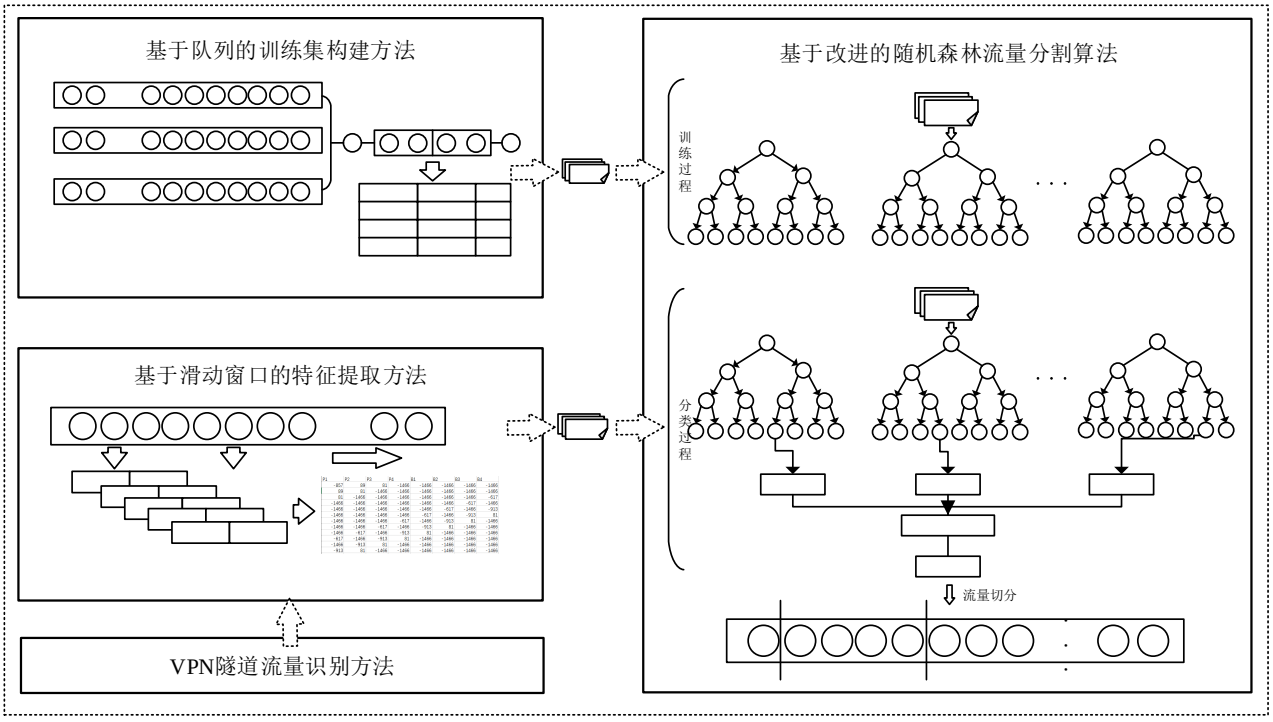


图 4-1 多类型 VPN 隧道流量分割方法整体流程图

4.1 问题分析

目前，学术界对于 VPN 流量的研究大都基于 ISCX 2016 VPN-noVPN 数据集。该数据集使用 SSL OpenVPN 工具获取流量，并对 7 种业务类型的 VPN 流量和非 VPN 流量进行标签。大多数研究在使用该数据集时进行 VPN 业务识别与应用识别时未考虑到 VPN 隧道内多种业务复用的情况，而只考虑 VPN 隧道内只有一种业务类型流量，并基于此进行 VPN 加密流量业务类型识别的研究。相比于非 VPN 流量，本文研究的 VPN 隧道流量具有以下 2 种特性：（1）VPN 隧道流量无法按照传统的五元组（源地址、目的地址、源端口、目的端口、协议）方法进行分流，VPN 隧道流量只能按照三元组（源地址、目的地址、协议）区分流量方向和

隧道加密协议；（2）VPN 在进行通信时往往给会产生随机长度的填充字段，这些字段会对全局流量特征构建产生干扰，VPN 隧道流量负载特性被掩盖、头部信息被修改，构建有效的分割特性成为解决本问题的关键。

基于以上背景，本文聚焦于多种类型业务复用 VPN 隧道问题，该问题的形式结构定义如下：假设端到端的 VPN 隧道流量$F=\{P_{11},P_{12},P_{23},P_{24},P_{35},P_{46},P_{57},P_{58}\dots P_{mn}\}$，其中$P_{mn}$为 VPN 隧道流量的数据包，$n$为数据包编号，$m$为 VPN 隧道内业务类型。本文选取网页浏览、文件传输、邮件传输、即时通信、音视频通话 5 种业务类型，覆盖了大部分用户需求，因此 m 的取值范围为[1,5]。本文针对 VPN 隧道流量分流问题，实现不同业务类型的 VPN 流量数据包分离，该研究目标形式结构定义如下：$K=\{f_1,f_2,f_3,f_4\dots f_u\}$。其中，$K$为分割后的 VPN 隧道流量，$f_u=\{P_{md},P_{md+1}..P_{md+e}\}$为 VPN 隧道流量的分割集合，$d+e$为 VPN 隧道流量的数据包编号。为此，本文提出了基于数据包序列的 VPN 加密流量分割方法M，使得$M(F)=K$。

针对以上内容，本章的工作如下：

（1）分析 SSL VPN 加密流量和 IPSec VPN 加密流量差异性，提出一种基于数据包负载特征的 VPN 加密流量识别方法实现 SSL VPN 加密流量和 IPSec VPN 加密流量的区分。

（2）针对 VPN 隧道流量分割特性问题，本章提出了一种基于数据包序列的 VPN 隧道流量分割方法。本方法应准确地计算分组序列的相似性，并确定这两个序列是否来自相同的应用类型。如果它们来自不同的业务，那么两个数据包序列之间的连接将被定义为行为变化点。

实验表明，本章提出的方法在各项评价指标中均具有明显的优势，可以实现 VPN 隧道流量的分割。

4.2 多类型 VPN 隧道流量识别研究

本节主要分析 SSL VPN 隧道流量和 IPSec VPN 隧道流量的数据格式，有以下目的：第一，实现 IPSec VPN 隧道流量和 SSL VPN 隧道流量的区分，针对不同 VPN 隧道流量提出不同业务识别特征，为第五章内容做铺垫；第二，通过分析负载字段特征，说明 VPN 隧道流量负载特征难以实现 VPN 隧道流量分流的目的，继而提出本章重点——VPN 隧道流量分割方法。

4.2.1 VPN 隧道流量数据格式

（1）IPSec VPN 数据格式

IPSec VPN 协议栈由头部验证协议（Authentication Header，AH）、封装安全载荷协议（Encapsulate Security Payload，ESP）和因特网密钥交换协议（Internet Key Exchange，IKE）三种协议套件组成[43]。其中，IKE 协议为 AH 协议和 ESP 协议提供密钥协商服务；AH 协议可以提供数据完整性验证、防止重放攻击，但不提供数据加密服务；而 ESP 协议可以提供 AH 协议全部功能外，还能提供数据加密服务。

本文最终目的为 VPN 隧道流量业务识别，VPN 隧道握手协商数据包对于整个 VPN 隧道流量业务识别影响较小，因此，本文只考虑 VPN 通信过程中数据传输部分数据包。

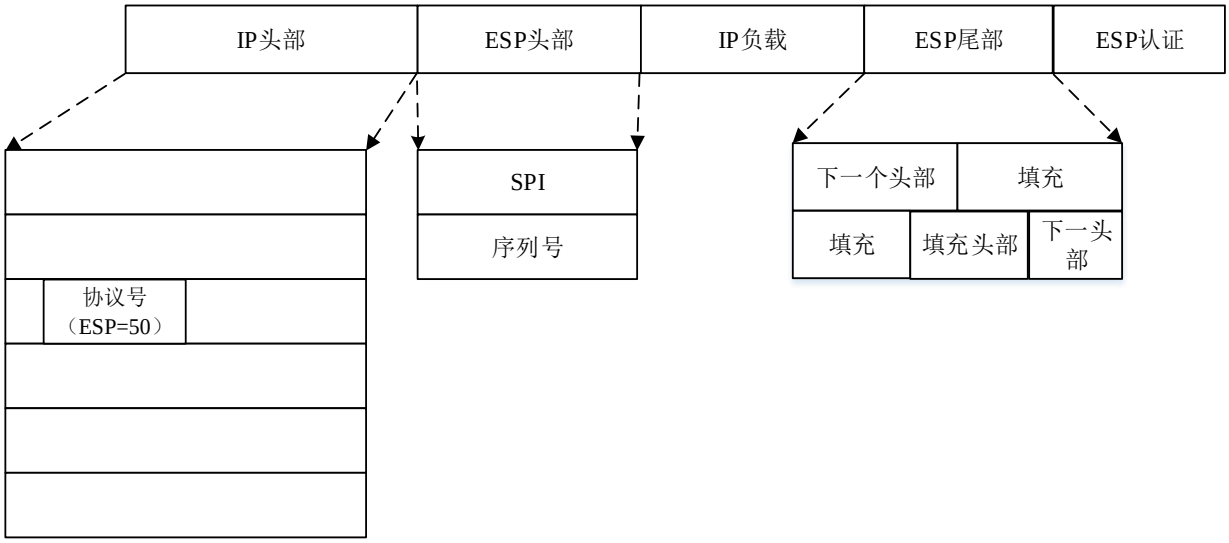


图 4-2 ESP 数据格式图

在实际实践中，IPSec VPN 往往配置 ESP 协议，而不使用 AH 协议，因为 AH 协议只提供验证功能而不提供加密功能。IPSec 有两种工作模式，即传输模式（Transport mode）和隧道模式（Tunnel mode）。在传输模式下，IPSec 会在内核中为 IP 数据包插入一个 IPSec 头部，该头部位于 IP 报头和 IP 数据包负载之间，并使用 ESP 协议为其加密。同时将修改 IP 报头的协议号并重新计算校验和，除此之外没有其他改动。传输模式可以保护传输数据包有效负载，但是在 IPSec 源端点处不会修改目的 IP 地址，会导致使用者的访问意图暴露。而在隧道模式下，IP 报文头部和荷载部分被一起封装成为新荷载，并在新的 IP 头部和荷载之间插入一个 IPSec 报头，这使得原始 IP 头部被充分保护。IPSec 隧道模式相比于传输模式而言，安全性更高，大量公司企业使用 IPSec 组网往往会采用隧道模式，本文研究也聚焦于隧道模式下使用 IKE 套件和 ESP 套件的 IPSec VPN 加密流量。其数据格式如图 4-2 所示。

使用 ESP 协议进行封装数据包后，新数据包的 IP 头部协议字段为 50。该字段可以用于识别 IPSec VPN 传输使用 ESP 协议加密的数据包

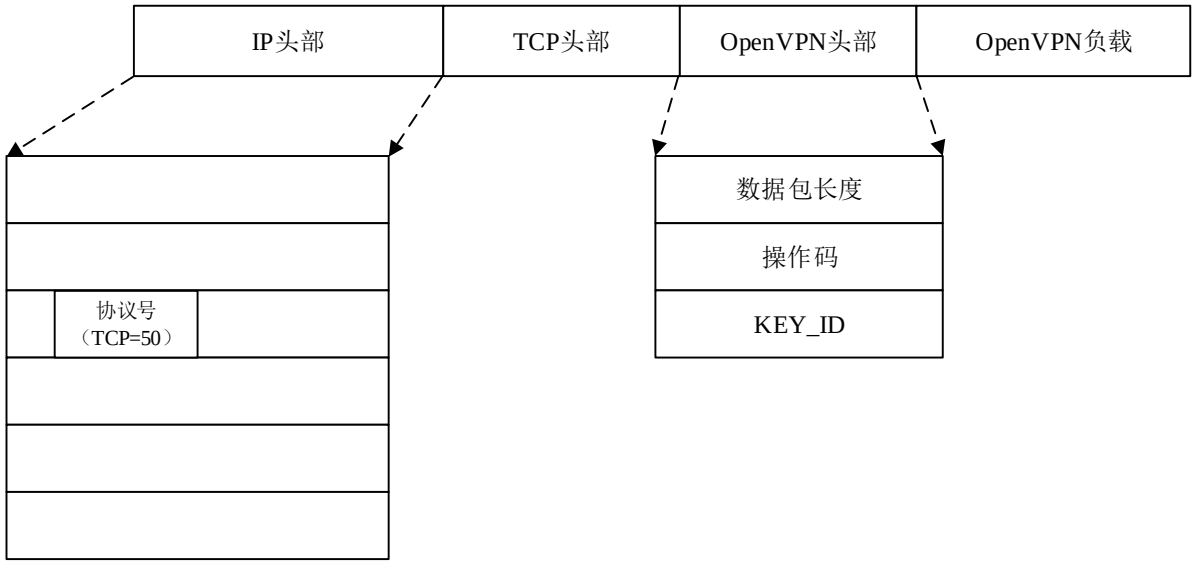


图 4-3 OpenVPN 数据格式图

（2）OpenVPN 数据格式

OpenVPN 是一款开源 SSL VPN 解决方案，实现了基于 SSLv3/TLSv1 的 VPN 通信。相比于 IPSec VPN 使用内核空间，OpenVPN 是一种基于用户空间 SSL 的 VPN，其不使用固定端口[44]。OpenVPN 底层使用 TCP 或 UDP 协议，其数据格式如图 4-3 所示。操作码字段有 5 位，其数值 1-8 对应不同的含义。OpenVPN 通信可以分为连接初始化、SSL/TLS 握手、密钥协商和交换 4 个阶段，其都依托于 OpenVPN 数据包。本文 OpenVPN 使用 TCP 协议作为底层协议。

通过二者数据包格式分析可以得知二者数据包负载字段具有明显区分性，可以作为二者隧道流量识别的特征。此外，二者都使用封装负载方式实现 VPN 加密通信。

4.2.2 基于负载字段的 VPN 加密流量识别方法

本节重点在于识别 IPSec VPN 加密流量和 SSL VPN 加密流量，是一个二分类问题。从上节二者数据包格式分析，IPSec VPN 加密流量的握手数据包和内容传输数据包某些字段具有显著。因此，本文提出了一种基于负载字段的 IPSec VPN 加密流量识别方法，用于区分 IPSec VPN 加密流量和 SSL VPN 加密流量。

用户在使用 VPN 软件进行匿名通信时，在一条流中只会使用一种类型的 VPN 服务，因此识别某一条流量类别，只需要判断该流量数据包负载显著特征即可。本方法使用 ESP 数据包中 IP 头部协议字段作为负载特征，如果待判断流量中 IP 数据包头部协议字段出现 0x32，那么该流量就可以定义为 IPSec VPN 加密流量，反之则为 SSL VPN 加密流量。该方法伪代码如算法 4-1 所示：

算法 4-1 基于负载字段的多类型 VPN 加密流量识别算法

输入：完整数据流$F=\{p_1,p_2,p_3,p_4,p_5,p_6\dots p_n\}$，其中$p_i$表示数据流$F$第$i$个有效负载
输出：该流量的类型$flag$，$flag$取 1 为 IPSec VPN 加密流量，否则为 SSL VPN 加密流量

```
1:  Function Identifiction(F)
2:      for p_i in F do
3:      Get_Byte(p_i, 24)获取数据包第 24 个字节的值 p_i^24
4:      if p_i^24==0x32:
5:          Flag=1
6:          return flag
7:      flag=0
8:      end for
9:      return flag
10: End Function
```

从以上代码可以看出，通过判断 VPN 隧道流量中是否含有 ESP 数据包可以实现 VPN 隧道流量的识别（算法第 4 行）。该方法时间复杂度为$O(n)$，n为 VPN 隧道流量数据序列长度，即最多需要判断一个流所有的数据包。空间复杂度为$O(1)$，具有良好的识别性能。

4.3 VPN 隧道流量切割特征构建方法

传统 VPN 隧道流量分割方法为超时值法，该方法难以应用于现实生产实践场景。本章在分析 VPN 隧道流量负载的基础上，借鉴了图像处理中场景变换的思想，提出了基于 VPN 隧道流量切割特征，并从流量统计方面分析了该特征集合的有效性。接下来，针对 VPN 隧道流量本章提出了基于队列的特征训练集构建方法和基于滑动窗口的特征提取算法用于构建 VPN 隧道流量的特征训练集和特征测试集。

4.3.1 VPN 隧道流量切割传统方法

从 IPSec VPN 加密流量数据格式分析可以得出，数据包级特征具有统一性：相同的 IP 头部字段、SPI 和序列号。其中，SPI 仅仅可以区分流量的方向，即对于同一方向的 IPSec VPN 隧道流量具有相同的 SPI，而序列号为同一流量方向单向递增。由此可见，使用 IP 头部和 ESP 头部无法完成流量切割要求。从 SSL VPN 加密流量数据格式分析，数据包负载特征具有一致性：相同的 IP 头部字段、OpenVPN 头部，这与 IPSec VPN 加密流量数据包表现出的性质一致，因此，使用负载特征不具有切割流量的条件。

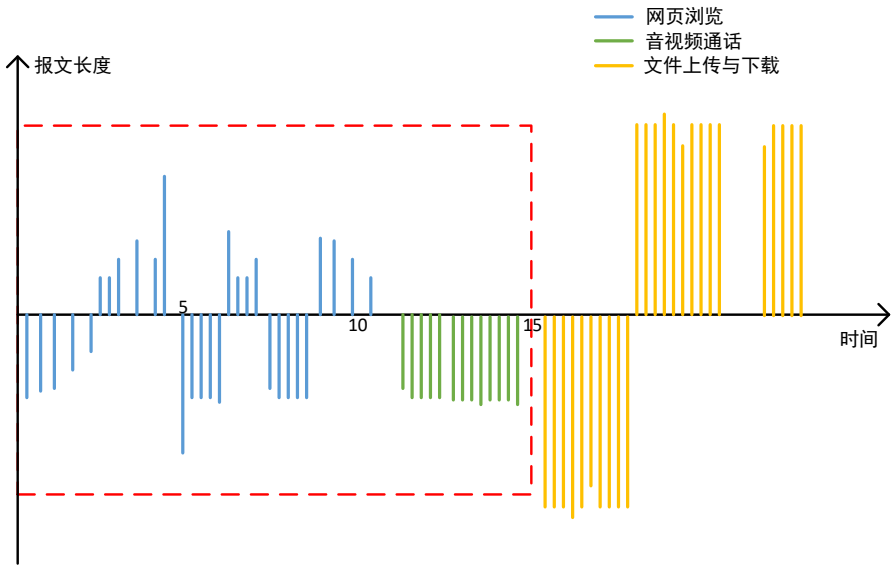


图 4-4 超时值方法图

先前的研究中[3]，使用超时值来分割具有单流特性的流量。超时方法是基于检查数据包之间的时间间隔来确定变化点。首先设置一个初始时间阈值，比如 20 秒。如果该时间间隔大于该阈值，则报文之间的交界处将成为分流点。然后，把分流点之前的包视为一个流。但是在现实世界中，每个用户都有不同的电脑使用习惯。报文之间的时间间隔会受到用户操作的影响。当超时阈值不适合当前用户时，很容易错误地确定分离点。而这种故障会影响特征提取的阶段，影响最终的分类结果。如图 4-4 所示，将超时值设置为 15 秒，如果两条不同行为的业务流数据包时间间隔为 5 秒，那么第二条业务流会被认为属于上一条业务流，因此两条流的分割点被错误的判断。此外，很难选择适合所有用户的超时阈值，考虑到可能存在流量中

数前后数据包属于不同的业务行为，其数据包时间间隔为 0.1ms 以下，那么会导致过多的分割点，对于一条持续时间为 10 分钟的流量，其分割点数目会达到 6x10^6 级别；如果选取的超时阈值过大，那么很有可能出现上述情况。因此，使用超时方法不可避免地会造成误判，损害分类结果。

本文借助图像场景转换[45]的概念，用于区分 VPN 隧道流量内不同业务行为之间分割点，通过比较前后数据包序列的相似度来确定流量分割点。该数据包序列应为定长序列，如果为变长序列，那么机器学习算法和深度学习都难以从流量中学习到全局特征，且定长序列更有利于计算数据包相似度。在选取序列长度时，需要考虑实际网络流量中不同业务行为在数据包序列的表现。

4.3.2 VPN 隧道流量切割特征构建

在提出本文切割特征之前，首先给出信息熵和相对熵的概念。

信息熵：信息熵概念是由 Shannon 提出用来衡量数据丢失程度的信息量，在密码学领域内使用它作为数据的随机性度量，熵值越高表示数据随机性越大[46]。假设离散变量$X = \{x_1, x_2, x_3, x_{m-1}, x_m\}$，对应概率分布为$P = \{p_1, p_2, p_3, p_{m-1}, p_m\}$，其中$\sum_{i=0}^{m} p_i = 1$且$0 < p_i < 1$，那么该信息熵可以定义为：

$$E = -\sum_{i=0}^{m} p_i \ln p_i \qquad (4-1)$$

相对熵：指两个随机序列之间距离的度量，从统计学角度它是指两个随机序列之间的相似程度[47]。假设离散变量$X = \{x_1, x_2, x_3, x_{m-1}, x_m\}$，对应概率分布为$P = \{p_1, p_2, p_3, p_{m-1}, p_m\}$，离散变量$Y = \{y_1, y_2, y_3, y_{m-1}, y_m\}$，对应概率分布为$Q = \{q_1, q_2, q_3, q_{m-1}, q_m\}$，相对熵可定义为：

$$D = \sum_{i=0}^{m} p_i \, ln \frac{p_i}{q_i} \qquad (4-2)$$

若两段序列完全相等时，即$P = Q$，则其相对熵$D = 0$，若相对熵D的取值越大则代表两段序列的相似程度约小，反之则代表序列的相似程度越大。

若把数据包负载看作比特流可以计算负载的相对熵和信息熵，进而利用比特流之间熵值变化点判断比特流之间是否存在分割点。然而，本文通过计算数据集种 VPN 隧道流量的数据包信息熵发现其值位于区间[0.9999,1]，即 VPN 隧道流量内所有数据包负载信息熵没有变化。这是因为使用 VPN 加密协议和代理机制使得所有数据包完全具有随机性，通过数据包负载分析难以获取 VPN 隧道流量的信息。基于以上，本文提出了以下两种基于数据包长度的熵：长度熵和长度相对熵用于表征定长序列的数据包长度混乱程度和相似程度。

利用信息熵反应长度变化，可以判断单位 VPN 隧道流量内报文长度变化的随机程度。由此，给出长度熵E_{len}的定义：

$$E_{len} = -\sum_{i=0}^{n} \frac{k_i}{n} ln \frac{k_i}{n} \quad (4-3)$$

其中，n为单位数据包个数，k_i为单位 VPN 隧道流量序列内不同长度数据包出现的频数且满足$\sum_{i=0}^{n} k_i = n$。

利用相对熵可以表示两个序列的相似程度，进而判断 VPN 隧道流量内两个序列之间是否存在分割点。由此，给出长度相对熵D_{len}的定义：

$$D_{len} = -\sum_{i=0}^{n} \frac{k_i}{n} ln \frac{k_i}{r_i} \quad (4-4)$$

其中，n为单位 VPN 隧道流量定长序列数据包总长度，k_i为单位 VPN 隧道流量定长序列内顺序数据包长度，且满足$\sum_{i=0}^{n} k_i = n$，r_i为上一个序列内顺序数据包长度。

对于具有五元组特性的加密流量进行分割时，可以对比前后数据包的源 IP 地址、目的 IP 地址、源端口号、目的端口号以及协议号，而具有单流特性的 VPN 隧道流量却难以使用传统方法进行分割。基于上节可行性分析，可以通过对比前后一定数量的数据包形成的短单元流量的长度统计特征，来确定两个短流量单元之间是否为分割点。由此，本文提出了如下特征集合C：

$$C = \{L_{head}, L_{tail}, S_{head}, S_{tail}, label\}$$

其中，L_{head}为 VPN 隧道流量内单位长度序列数据包长度相关统计特征集合，S_{head}为单位长度序列熵统计特征集合，L_{tail}为下一序列数据包长度相关统计特征集合，S_{head}为下一序列熵统计特征集合，$label$为两个单位长度序列之间是否构成分割点。以上特征集合参数解释如表 4-1 所示：

表 4-1 VPN 隧道流量分流特征表

特征符号	特征参数	参数解释
L	$\{l_1, l_2, l_3 \dots l_{n-1}, l_n\}$	数据包长度序列
	l_{avg_fw}	数据包正向平均长度
	l_{var_fw}	数据包正向标准差
	l_{avg_bw}	数据包反向平均长度
	l_{var_bw}	数据包反向标准差
S	E_{len}	长度熵
	D_{len}	长度相对熵

其中，n为 VPN 隧道流量内单元序列长度，本章在数值上使用正数来代表正向特征，使用负数来代表反向特征。在切割数据包时，考虑到切割效率的问题，那么应该选择尽量多的数据包序列进行分析，然而过多的数据包分析会导致模型无法从特征中学习到特征；考虑到分割精度的问题，应选择尽量少的数据包长度分析，尤其应该缩小至一个数据包进行切割，但是会导致特征集C中某些特征难以获取。因此，需要针对n的取值进行实验分析，本文将在 4.5 节讨论n的大小对分割精度的影响。

4.3.3 可行性分析

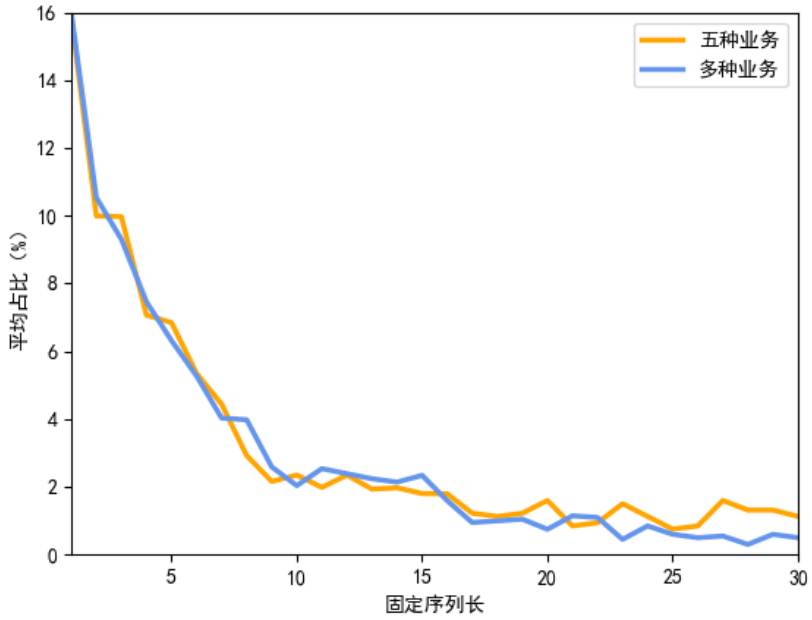


图 4-5 VPN 隧道流量的分割点之间数据包数量频率图

本文在第三章构建的 VPN 隧道流量业务行为分割数据集中，前 300 条 VPN 隧道流量为五种业务复用同一隧道，300 条流量为多种业务复用同一隧道，200 条流量为单一业务使用隧道。构建数据集时，保留了数据包时间戳，这使得多种业务复用同一隧道可以得到准确的分割点。如图 4-5 所示，本文依据不同分割点按照以上 VPN 隧道内三种业务行为表现，统计了上述 150 条 VPN 隧道流量不同分割点之间数据包不同数量占比的均值。本图只展示了 1-30 的数据包占比，原因有两点：（1）过于大的数据包序列固定长度占比较小，且如果考虑过长数据包序列长度，难以识别出较短距离的分割点；（2）过于大的数据包序列固定长度占比不稳定，如图 4-5 中 20-30 的数据包占比出现上下波动，这对于分割 VPN 隧道流量没有统计学意义。因此，图 4-5 展示了分割点之间固定序列长度 1-30 的数据包数量占比。

从图 4-5 可以看出，对于单一业务使用隧道来说，其分割点之间数据包数量为该流量数据包数量；对于多种业务复用同一隧道和五种业务复用，其统计数量具有相似性，其中分割点之间数据包数量为 1 的平均占比为 15%左右，数据包数量为 2 的平均占比为 10%左右，数量包数量为 3 的平均占比为 9%左右，数据包数量为 4 的平均占比为 7%，数据包数量为 5 的平均占比为 6%，数据包数量为 5 的平均占比为 6%，数据包数量为 5 的平均占比为 6%，数据包数量为 6 的平均占比为 5%，数据包数量为 7 的平均占比为 4%，数据包数量为 8 的平均占比为 3%，数据包数量为 9-16 的平均占比均为 2%。由此可见，五种业务类型的分割点之间数据包数量占比与多种业务类型的分割点之间数据包数量占比几乎一致。在进行流量切割时，二者可以考虑为同一情况，为此本文将在 4.5.3 对比实验进一步验证。

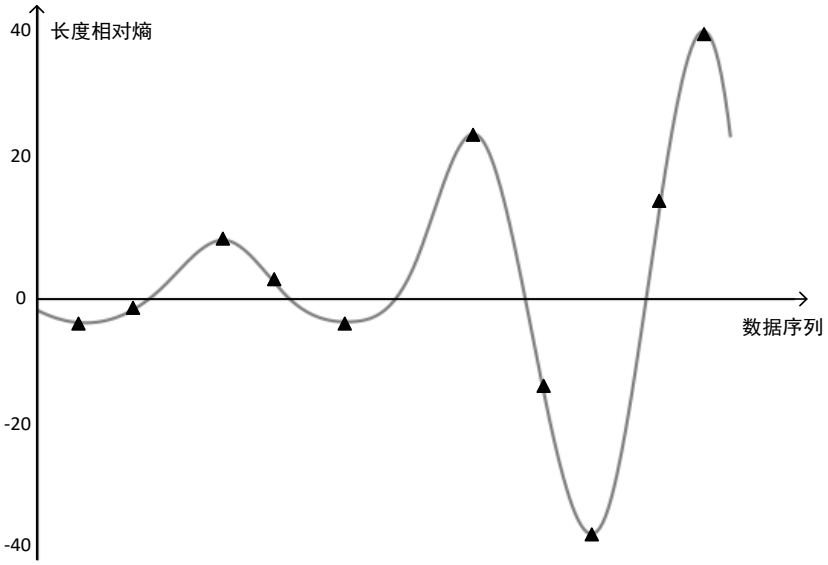


图 4-6 长度相对熵图

由此可见，若有多种业务复用同一隧道，那么一定数量的数据包之间具有切割点。一个数据包发送要经过内核协议栈处理，根据数据消息为数据信息添加相应的数据包头部，发送到 DMA 循环发送队列，网卡从该队列中获得要发送的数据发送到数据传输媒介。由于局部性原理的存在，往往是多个相同应用的数据包下发至内核维持的发送队列中，这样网卡在发送数据包时相邻数据包是相同应用数据包的概率会很大。因此，依靠定长数据包相似度可以确定 VPN 隧道流量分割点。

本文随机选取了一个 VPN 隧道流量，计算了定长序列取 4 时长度相对熵值变化曲线，如图 4-6 所示。其中三角符号代表分割点，曲线代表随着长度序列变化的长度相对熵的数值，可以看出部分分割点出现在熵值变化达到极值。这说明熵特征集合可以用于寻找分割点，由此表明本文提出的熵特征集合的有效性。而还有一些分割点不出下在长度相对熵曲线的极值处，需要依靠其他特征进行判断。由此可见，熵特征集合在判断切割点具有有效性。

4.3.4 基于队列的训练集构建方法

VPN 代理形成了用户和目标服务器之间的中间人，用户发送的所有流量包的目的地址都是 VPN 代理服务器。不同类型业务可以使用同一 VPN 隧道，形成了 VPN 隧道多路复用。本文在构建 SSL VPN 数据集和 IPSec VPN 数据集时，同一 VPN 隧道的不同业务行为的 VPN 隧道流量已经实现分流，构成了带有标签的 VPN 隧道流量数据集。本章面向多类型 VPN 隧道流量的分流方法是基于机器学习的方法，需要使用带标签的训练集构建切割模型。而本章提出的 VPN 隧道流量分割特征集合，需要在一个 VPN 隧道流量内寻找业务行为分割点，因此，本文提出了基于队列的训练集构建方法，获取用于构建随机森林分割模型的训练集,该方法可以用图 4-7 表示。

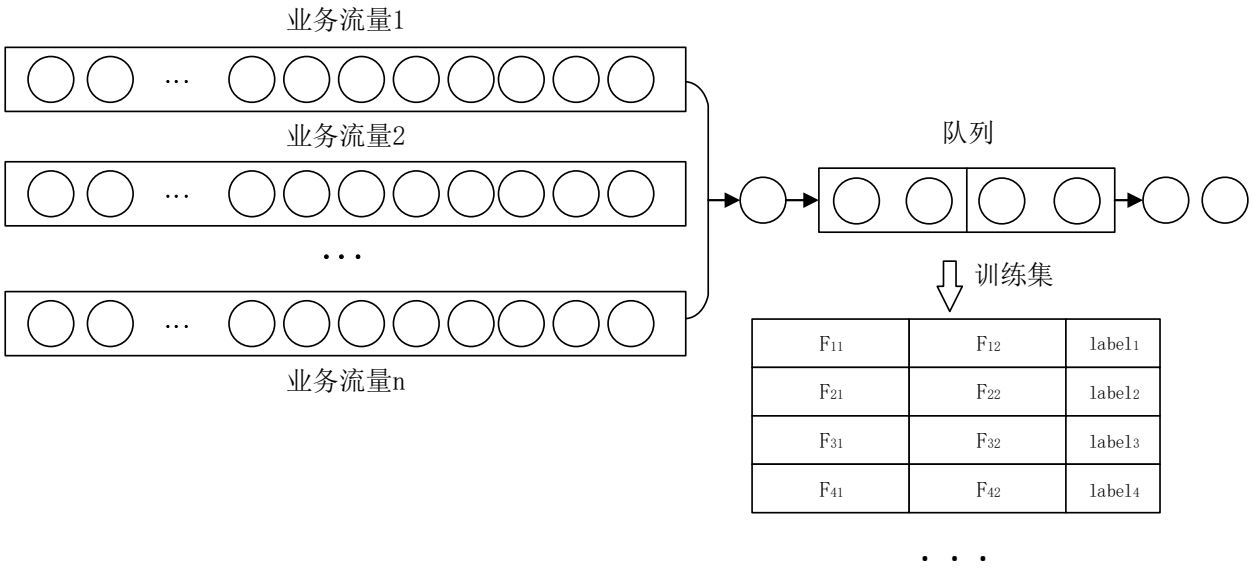


图 4-7 基于队列的训练集构建方法图

在构建队列数据结构时，给队列中元素增加一个属性 business，用于记录该数据包所属业务类型，便于构建训练集的标签集合。本方法使用多个不同业务类型且属于同一个隧道的 VPN 隧道流量构建标签训练集，其方法流程如下：

（1）获取指向同一个隧道流量的不同业务行为的所有文件指针，通过文件指针读取流量数据包数据，比较所有数据包时间戳，选取最小者压入队列，记录该数据包的 business 属性，并将数据包所属文件指针后移；

（2）当队列长度达到 2 倍窗口值长度时，调用 get_length_sequence 和 get_length_entropy 方法获取队列中前后窗口数据包的长度序列特征 L 和熵特征 S。

（3）判断前窗口最后一个数据包和后窗口第一个数据包的 business 是否相等，相等则将 label 置为 1，否则置为 0；

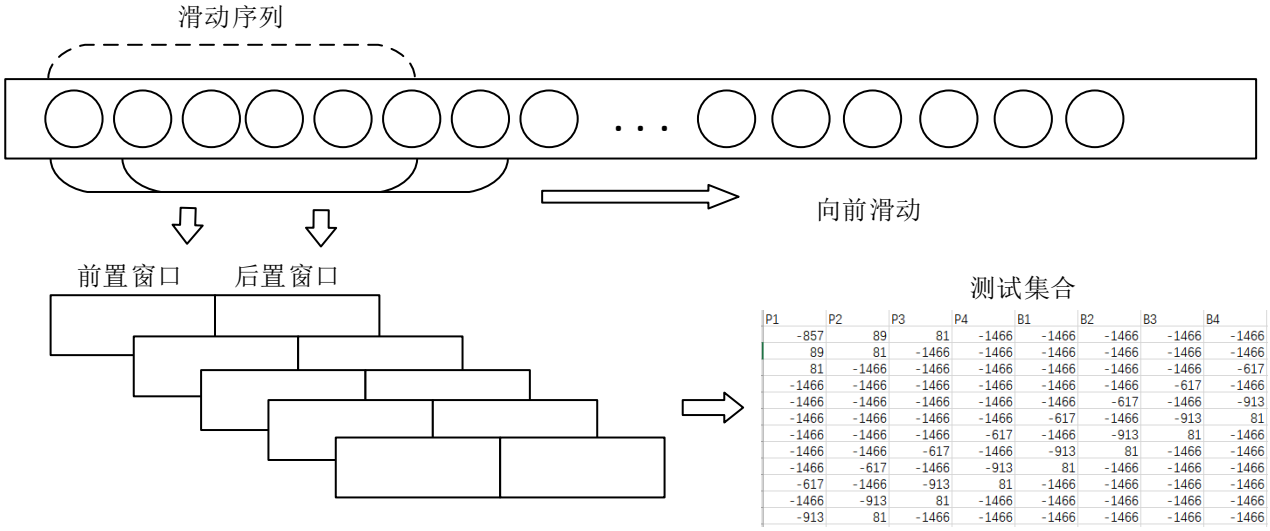
（4）构建训练特征和标签集合$G = \{K_1, K_2, K_3, K_4, K_5, K_6 \dots K_n\}$，其中$K = \{L, S, label\}$；

（5）执行出队列操作。

本方法通过设有额外属性 business 来辅助标签集合的建立，使用队列数据结构构建特征集合。通过基于队列的训练集方法可以实现依靠同属一个 VPN 隧道流量的不同业务类型流量文件构建用于切割算法的训练集。

4.3.5 基于滑动窗口的特征提取算法

由于待分割数据在形式上往往是一个文件，不同长度的数据包序列之间数据的重复性存在一定关联，导致了在一定长度序列中数据特征表现不同，根据这种差异性可以确定流量分割点位置。因此，本文提出了一种基于滑动窗口的特征提取算法，用于提取待分割 VPN 隧道流量的切割特征集合，该特征集合为切割算法的测试集合，用于预测切割点位置，其方法如图 4-8 所示。



P1	P2	P3	P4	B1	B2	B3	B4
-857	89	81	-1466	-1466	-1466	-1466	-1466
89	81	-1466	-1466	-1466	-1466	-1466	-1466
81	-1466	-1466	-1466	-1466	-1466	-1466	-617
-1466	-1466	-1466	-1466	-1466	-1466	-617	-1466
-1466	-1466	-1466	-1466	-1466	-617	-1466	-913
-1466	-1466	-1466	-1466	-617	-1466	-913	81
-1466	-1466	-1466	-617	-1466	-913	81	-1466
-1466	-1466	-617	-1466	-913	81	-1466	-1466
-1466	-617	-1466	-913	81	-1466	-1466	-1466
-617	-1466	-913	81	-1466	-1466	-1466	-1466
-1466	-913	81	-1466	-1466	-1466	-1466	-1466
-913	81	-1466	-1466	-1466	-1466	-1466	-1466

图 4-8 基于滑动窗口的特征提取算法图

在数据包序列中，设置前后两个对比窗口，提取上文构建的 VPN 隧道流量切割特征，形成特征测试集的一行特征数据。滑动距离取 1，即每次向前滑动一个数据包的长度。该方法的伪代码如算法 4-2 所示。

算法 4-2 基于滑动窗口的特征提取算法

输入 1：完整数据流$F=\{p_1,p_2,p_3,p_4,p_5,p_6\dots p_n\}$，其中$p_i$表示数据流$F$第$i$个有效负载
输入 2：窗口大小 win
输出：对应的数据特征集合$f=\{T_1,T_2,T_3,T_4,T_5,T_6\dots T_m\}$，其中$T_i$为上述特征集合

```
1:  Function Split_feature(F,win):
2:      left←0
3:      right←win*2
4:      While right<n do
5:          T_left = []
6:          get_length_sequence(left,right,F)获取窗口长度序列集合 L
7:          get_length_entropy(leftright,F)获取窗口长度序列的长度熵 S
8:          T_left.append(L,S)
9:          f.append(T_left)
10:         right←right+1
11:         left←left+1
12:     End While
13:     return f
14: End Function
```

以上基于滑动窗口的序列特征提取过程会将待切割 VPN 隧道流量以及窗口值大小作为为算法输入，输出为切割 VPN 隧道流量特征集合。本算法首先设置前后窗口的边界值$left$和$right$（算法第 2、3 行）。然后算法使用滑动窗口机制，每一次向前滑动得到待处理窗口（算法第 10、11 行），当滑动至 VPN 隧道流量数据包序列最右边算法终止（算法第 4 行）。通过调用$get_length_sequence(left,right,F)$获取窗口数据包序列的长度序列集合（算法第 6 行），该集合为上节所设计的数据包负载长度序列集合，通过$get_length_entropy(leftright,F)$获

取窗口数据包序列的长度熵（算法第 7 行）。最终，返回特征集合f。该算法时间复杂度为$O(n)$，代码执行时间总和与待分割流量的数据包序列长度呈线性相关，具有有效的时间效率。该算法空间复杂度为$O(1)$，只需要常数级别的存储空间。从时间复杂度和空间复杂度来看，基于滑动窗口的特征提取算法具有良好的效率和低内存占有率。

4.4 基于数据集改进的随机森林分割模型

本节使用常用机器学习算法——随机森林[48]，其在加密流量分类领域应用广泛。本节根据 4.3 节构建的 VPN 隧道流量标签训练集对传统随机森林算法改进，为随机森林决策树赋予权重，依据权重获得最终投票结果。本节将首先介绍基于随机森林的 VPN 隧道流量分割算法，然后基于数据集对随机森林算法进行改进最终形成基于数据集改进的随机森林分割模型，本文称该模型为 SP-RF 模型。

4.4.1 基于随机森林的 VPN 隧道流量分割算法

在介绍随机森林之前，首先介绍决策树[49]。如图 4-9 所示，决策树具有典型二叉树结构，其根节点代表所有的样本集合，其非叶子节点代表对属性的测试，其叶子节点代表本树对于结果的预测。决策树从根节点开始，沿着某一路径最终得到决策结果。

决策树的分类决策模式具有单一性，暴露出最优解局部性、规则复杂性、过拟合性等缺陷。为了解决以上决策树存在的缺陷，随机森林算法应运而生。其在决策树的基础上，使用多个决策树分类器构成了随机森林。

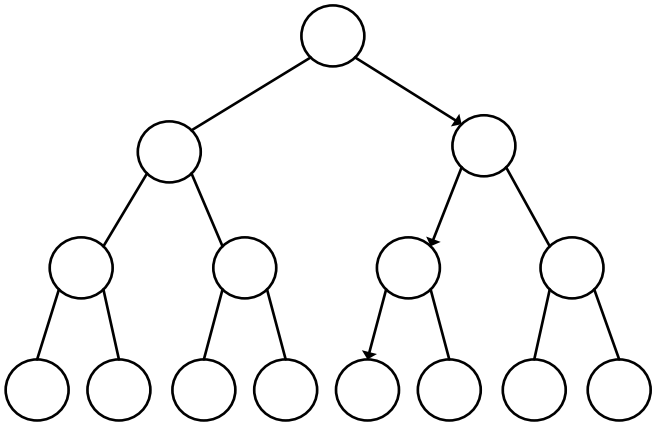


图 4-9 决策树算法图

在介绍随机森林流程之前首先介绍基尼系数，随机森林通过最小基尼系数来保证决策树上节点的划分。基尼系数的取值越小越小，位于子节点中的所有观察隶属于同一类的确定性越大，数据分割更彻底，其数学定义如下：

$$Gini(S) = 1 - \sum_{k=1}^{m} p_k^2 \qquad (4-5)$$

其中，m为样本类别数量，p_k为样本归于k类的概率大小。当使用某个特征a划分数据集时，使用最小基尼系数进行划分数据集S为S_1和S_2，其数学定义如下：

$$Gini_{split}(S,a)=\left|\frac{S_1}{S}\right|Gini(S_1)+\left|\frac{S_2}{S}\right|Gini(S_2) \qquad (4-6)$$

如果假设经过特征提取方法获取到的测试样本集合为$f=\{T_1,T_2,T_3,T_4,T_5,T_6\dots T_m\}$，其中 m 为测试样本个数；设经过训练集构建方法得到训练集合为$G=\{K_1,K_2,K_3,K_4,K_5,K_6\dots K_u\}$，其中 u 为训练样本个数；设特征集合为$C=\{c_1,c_2,c,c_4,c_5,c_6\dots c_n\}$，其中 n 为特征数量；设类别集合为$D=\{c_1,c_2\}$，即是否为分割点，基于随机森林的切割算法流程如图 4-10 所示。

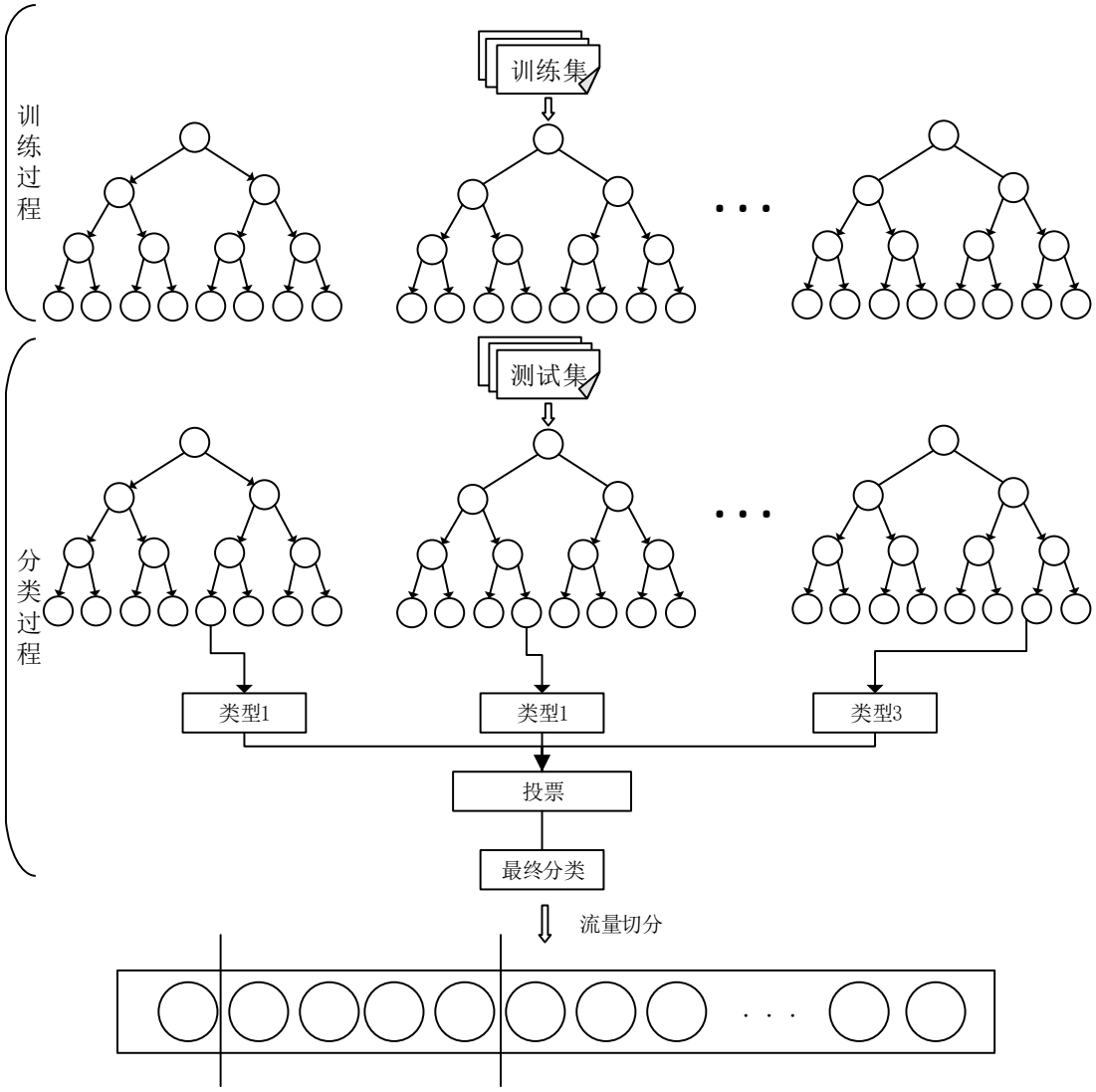


图 4-10 基于随机森林的切割算法流程图

该算法流程可以被详细描述为如下过程：

（1）使用 bootstrap 算法从训练集$G=\{K_1,K_2,K_3,K_4,K_5,K_6\dots K_u\}$中有选择 u 个样本，每个样本被选择后仍放回训练集中，由此构建一个训练子集g作为决策树的根节点；

（2）对于每一个训练子集g，从特征集C中任意地选取 p 个特征，并使得$p=\lceil\log_2 n\rceil$，然后从 p 个特征中使用最小基尼系数作为划分节点的依据，由此构建一个完整的决策树S_i；

（3）重复步骤（1）和步骤（2），构建了一系列决策树$S_1,S_2,S_3,S_4,S_5\dots,S_i$，基于该决策树的集合构成随机森林；

（4）依次输入测试集合$f=\{T_1,T_2,T_3,T_4,T_5,T_6\ldots T_m\}$的样本$T_m$至随机森林，基于每一个决策树的结果使用多数投票法（Majority Voting Algorithm）进行投票选举，得到一系列分类结果$R=\{r_1,r_2,r_3,r_4,r_5,r_6\ldots r_w\}$；

（5）根据结果集合$R=\{r_1,r_2,r_3,r_4,r_5,r_6\ldots r_w\}$来切割待切割流量，其中$r_w=1$为切割点。

4.4.2 基于切割训练集改进的随机森林算法

如果数据集中出现不同类别样本数量分布不均匀，即某类样本数量远多于其他类别样本数量时，该数据集为非平衡数据集。对于一个二分类问题而言，样本数量较多的类别被称为正类样本，样本数量相对较少的样本被称为负类样本。本章使用数据集非平衡系数来描述数据集样本分布的不平衡程度，其数学定义如下：

$$M=\frac{Q_T}{Q_F} \tag{4-7}$$

其中Q_T为正类样本，Q_F为负类样本，那么有$Q_T\cap Q_F=\emptyset$。本章通过基于队列的训练集构建方法获取了带有标签的训练集，其中数据包之间的非分割点数量远超过数据包之间非分割点数量，经过计算测得训练集的非平衡系数$M=8.732$，因此训练集为非平衡数据集。

数据集的不平衡性影响决策树的分类能力。随机森林算法使用最小基尼系数来划分节点构建用于预测的决策树结构，假设本章分割数据集经过 Bootstrap 抽样获得的数据集的非平衡系数为M'，那么基尼系数和最小基尼系数表示为：

$$Gini(S)=1-\sum_{k=1}^{2}p_k^2=1-\{\left(\frac{1}{M'+1}\right)^2+\left(\frac{M'}{M'+1}\right)^2\} \tag{4-8}$$

$$Gini_{split}(S,a)=\left|\frac{S_1}{S}\right|\{1-\frac{1+{M_1'}^2}{(M_1'+1)^2}\}+\left|\frac{S_2}{S}\right|\{1-\frac{1+{M_2'}^2}{(M_2'+1)^2}\} \tag{4-9}$$

其中，S_1为正类样本，S_2为负类样本。假设样本集中，正类样本与负类样本为完全分离的理想情况。假设$S_1/S_2=M$，那么就有：

$$Gini_{split}(S,a)=\left|\frac{M}{M+1}\right|\{1-\frac{1+{M_1'}^2}{(M_1'+1)^2}\}+\left|\frac{1}{M+1}\right|\{1-\frac{1+{M_2'}^2}{(M_2'+1)^2}\} \tag{4-10}$$

于是可以将对最小基尼系数的求解问题转化为关于非平衡系数的问题。其中，S_1为正类样本，S_2为负类样本，且有$1<M<M_1'$和$0<M_2'<1$。如何求解$Gini_{split}(S,a)$最小，从上述公式中可以看出就是M_1'向M靠近，而M_2'向 0 靠近。因此，可以得到以下结论：分割训练集非平衡系数M趋近于 1 时，数据分割地越彻底，决策树的预测能力越强。

Bootstrap 算法从原始数据集中进行抽样，构建用于预测的决策树。其抽样构成的数据集是随机的，这就会导致具有不同非平衡系的数据集构建的决策树预测能力具有差异性。随机森林算法使用多数投票机制选择最终预测结果，每一个决策树拥有相同的权重，即其决策结果影响最终结果的能力相同，但是基于以上分析，每个决策树决策能力有差异，提高预测能力强的决策树的权重可以获得更加准确的结果。因此，本文提出了一种适用于分割训练集的

加权随机森林算法，通过在随机森林投票环节利用非平衡系数为每一个决策树设置权重，并依据权重获取最终决策结果，权重计算公式如下：

$$K(n) = \frac{\sum_1^N M_i}{M_n \times N} \quad (4-11)$$

其中，N为抽样数据集数量，M_n为某抽样数据集的非平衡系数，由此求得由该数据集构建的决策树的投票权重$K(n)$，根据权重进行最终决策。

根据以上描述，本章提出了基于分割数据集的随机森林分割算法，其伪代码如算法 4-3 所示。该算法包含了模型训练和模型测试的过程，其中算法 1-7 行为模型训练过程，算法 8-10 行为模型测试过程，最后返回依据模型预测分割点分割后的 VPN 隧道流量。具体过程如下：首先调用$Split_feature(f)$函数获取特征测试集（算法第 2 行），该函数为算法 4-2 基于滑动窗口的特征提取算法；接下来，构建$\lceil \log_2 n \rceil$个决策树，并根据决策树权重生成随机森林 RF（算法第 3-10 行）；最终，依据模型预测结果R分割 VPN 隧道流量（算法第 11-13 行）。本方法通过构建$\lceil \log_2 n \rceil$个决策树，每个决策树平均长度为$\lceil \log_2 n \rceil$，因此算法的空间复杂度为$O(\log_2 n^2)$。该算法的时间复杂度为$O(M * \log_2 n + T)$，其中 M 为一颗决策树的建树时间，T 则为待分割流量的测试时间。

算法 4-3 基于分割数据集的随机森林分割算法

输入 1：待切割流量$P = \{p_1, p_2, p_3, p_4, p_5, p_6 \dots p_v\}$，其中$p_v$为有效负载
输入 2：训练集$G = \{K_1, K_2, K_3, K_4, K_5, K_6 \dots K_u\}$，其中$u$为训练样本个数
输入 3：窗口大小 win
输入 4：特征集$C = \{c_1, c_2, c, c_4, c_5, c_6 \dots c_n\}$，其中$n$为特征数量
输入 5：标签集合$D = \{c_1, c_2\}$，即是否为分割点
输出：VPN 隧道流量切分结果 M

```
1:  Function SP-RF(P, G, C, D):
2:      Split_feature(f, win)获取训练集f = {T_1, T_2, ... T_m}，其中m为测试样本个数
3:      S = []
4:      For i in range(⌈log_2 n⌉):
5:          bootstrap(f)获取G' = {K_1', K_2', K_3', K_4' ... K_u'}
6:          tree_build(G')获取决策树S_i
7:          tree_build(G')获取权重K(G')
8:          S.append(S_i)
9:      End For
10:     RF ← S
11:     TEST(f)获取分类结果R = {r_1, r_2, r_3, r_4, r_5, r_6 ... r_w}
12:     split_pcap(P)获取 VPN 隧道流量切分结果 M
13:     return M
14: End Function
```

4.5 实验设计与结果分析

4.5.1 实验环境

本章实验在物理机上部署实验代码，物理机的 CPU 为 AMD Ryzen 7 3700X 8-Core Processor 3.60 GHz，内存为 64GB，使用的操作系统及版本为 Windows 10 专业版，程序编译环境为 Pycharm，Python 版本为 python 3.6.5。

4.5.2 数据集

本章使用的数据集为第三章自建的 IPSec VPN 隧道流量标签数据集和 SSL VPN 隧道流量标签数据集，其简介如表 4-2 所示，该数据集详细介绍可以参考第三章内容。

表 4-2 VPN 隧道流量标签数据集

编号	业务类型数量	SSL VPN 隧道流量	IPSec VPN 隧道流量
1	五种	150	150
2	多种	150	150
3	一种	100	100

4.5.3 对比实验选取

为了表明本章方法的有效性，本章选取传统的超时值分流方法进行对比。超时值方法是根据超时阈值切割流量，其依赖于超时阈值的选择。本文将选择 1s、0.5s 作为超时阈值进行对比，以此证明本章通过构建 VPN 隧道流量固定长度序列特征进行分流方法优于传统 VPN 分流方法。

本章提出的 VPN 隧道流量分流方法除了依赖于特征构建有效性，也依赖于机器学习算法学习固定长度序列特征的能力。本章使用了基于分割数据集改进的随机森林算法进行特征学习和结果预测，为了证明该方法具有较好性能，本文选取了以下三种机器学习算法进行对比。

（1）C4.5 算法

C4.5 算法使用信息增益率进行节点划分，对子类别较多的特征不会产生偏向，增加了对具有连续数值特征的处理能力，在解决分类问题方面展示了强大的优势。本章提出的改进的随机森林算法基于集成学习的思想，将多个决策树的决策结果进行综合考虑。选取 C4.5 算法进行对比可以验证随机森林算法的基分类器具有较好的分类能力。

（2）传统随机森林算法

随机森林算法加入了随机性较少了过拟合情况的发生次数，可以有效地运行大型数据集，并且具有良好的抗噪声能力。本文选取传统随机森林算法用于验证本文提出的基于分割数据集改进的随机森林算法可以解决数据集不平衡问题，获得更高地分割精度。

4.5.4 评价指标

本章提出的基于数据包序列的单流 VPN 切割方法是一个二分类问题，即数据包之间是否为切割点，因此本文选择了精确率、准确率、召回率、F_1指标评价本方法。在介绍评价指标时，本文首先给出一些变量的定义。T_P为测试集中切割点获取正确切割的个数，F_P为非切割

点获取错误切割的个数，T_N为切割点获取非切割的个数，F_N为非切割点获取正确非切割的个数。

精确率：最直观反映测试样本被正确切割的能力，表示切割点被正确识别的占比，其数学定义如下：

$$Precision = \frac{T_P}{T_P + F_P} \quad (4-12)$$

召回率：反映了测试样本集中切割点被正确预测的样本个数占实际切割点个数的比例，其数学定义如下：

$$Recall = \frac{T_P}{T_P + F_N} \quad (4-13)$$

F_1指标：反映了准确率和召回率的综合指标，其数学定义如下：

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4-14)$$

准确率：反映了切割点和非切割点被准确预测的概率，其数学定义如下：

$$Accuracy = \frac{T_P + T_N}{T_P + F_N + F_P + T_N} \quad (4-15)$$

除了以上指标，本章还将使用预测时间说明模型对输入数据预测速度，其计算方式为模型预测 100 个输入样本所使用的平均时间。但是考虑到实际计算时计算资源被抢占式调度，模型预测时不能完全获取 CPU、GPU 资源。因此，该指标仅作为相对指标用于评价对比实验所使用方法的处理效率。

由于本章实验为了尽量减少数据集划分对分类结果的影响，使用 10 倍交叉验证方法用于实验分类性能，每一种指标的最终取值为 10 次实验的平均值。

4.5.5 结果与分析

（1）基于负载特征的 VPN 加密流量识别实验

本章使用第 3 章获取的数据集进行实验，选用了 800 条 VPN 隧道流量测试，其中 400 条为 IPSec VPN 隧道流量，400 条流量为 SSL VPN 隧道流量。该实验目的为了区分未知 VPN 隧道流量所属类别，最终准确率、精确率、召回率、F1 指标均为 100%。产生以上结果有以下：第一，数据集在获取时，至少流持续时间为 300s，VPN 隧道已经进入数据传输阶段，不会停留在握手协商阶段；第二，数据集在获取时，一条 VPN 隧道流量只会使用一种 VPN 服务，这符合用户习惯，一条持续的 VPN 隧道流量不会产生中途更换 VPN 服务的情况。因此，本方法对于分别 IPSec VPN 隧道流量和 SSL VPN 加密流量具有绝对的优势，可以完成二者的识别区分。

（2）滑动窗口值实验

本文选取了不同滑动窗口值进行实验，考虑到切割效率的问题，那么应该选择较大的窗口值进行分析，然而较大的窗口值会导致切割精度不够；考虑到分割精度的问题，应选择尽量小的窗口值进行分析，尤其应该缩小至一个数据包进行切割，但是会导致特征失效。基于

本章在 4.3.3 节的统计分析，本文选取 1-16 不同大小的窗口值进行实验，对于分流问题来说，首要任务是保证分割精度，因此精确率具有最高衡量权。

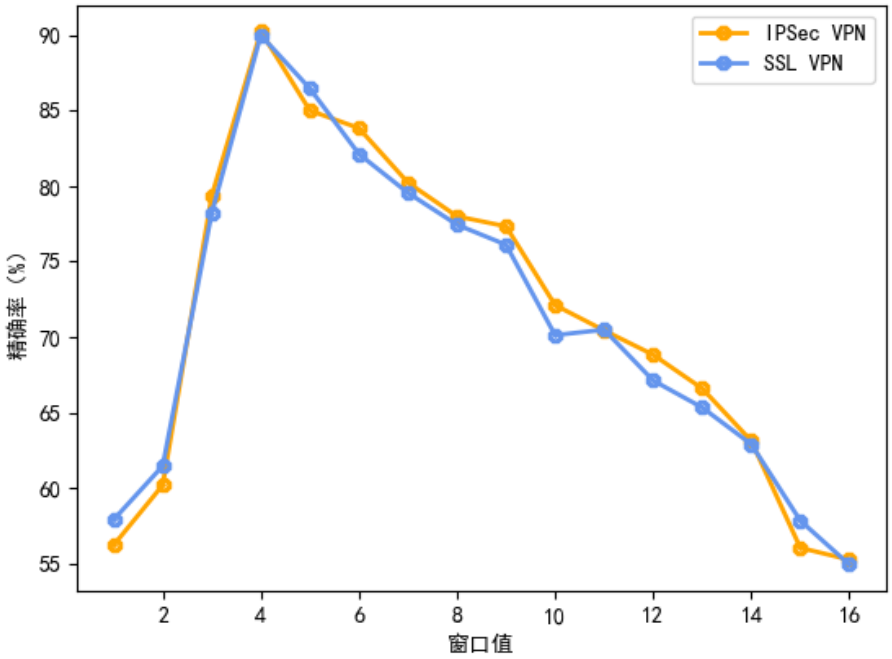


图 4-11 不同 VPN 隧道流量对应的窗口值-精确率图

首先对比 SSL VPN 隧道流量和 IPSec VPN 隧道流量的各个窗口值对应平均精确率如图 4-11 所示。可以看到的是，当窗口值取 4 时，SSL VPN 隧道流量和 IPSec VPN 隧道流流量具有最高精确率，分别为 90.01%和 90.26%；窗口值取 1-3 时，精确率反而降低，其数值小于窗口值取 5-7；窗口值取 5-16 时，精确率随窗口值增加而逐渐减小。根据 4.3.3 节统计来看，对于序列长度为 1-3 短单元占比最多，但是取该长度提取分割特征时，精度反而较低，这是因为较小的窗口值难以表达切割点前后窗口特征的差异性，即前后长度序列对比效果不显著，导致基于切割数据集的随机森林算法难以从特征中进行学习实现准确切割。窗口值取值过大时，其切割精度也会降低，是因为序列长度过大的短单元占比较小，分割特征难以从其单元特征中进行有效学习，导致基于切割数据集的随机森林算法对反映切割整条流量的特征学习不全面。

此外，SSL VPN 隧道流量和 IPSec VPN 隧道流量具有相同的精确率表达。其原因如下：第一，SSL VPN 和 IPSec VPN 在封装数据包时，不会对原有数据包进行改动，因此不会产生同一个数据包封装不同业务类型流量，即同一个数据包内不会产生分割点，分割点只会产生于数据包之间；第二，SSL VPN 和 IPSec VPN 在进行参数设置时，尽量调节了隧道内网卡 MTU 的值，使得较大长度的数据包在隧道内不会因为 MTU 设置而分包；第三，根据前文分析，SSL VPN 加密流量和 IPSec VPN 加密流量都具有单流特性，这种特性不随着 VPN 服务类型的不同而发生变化，印证了 VPN 隧道流量分流具有研究价值和实用价值。

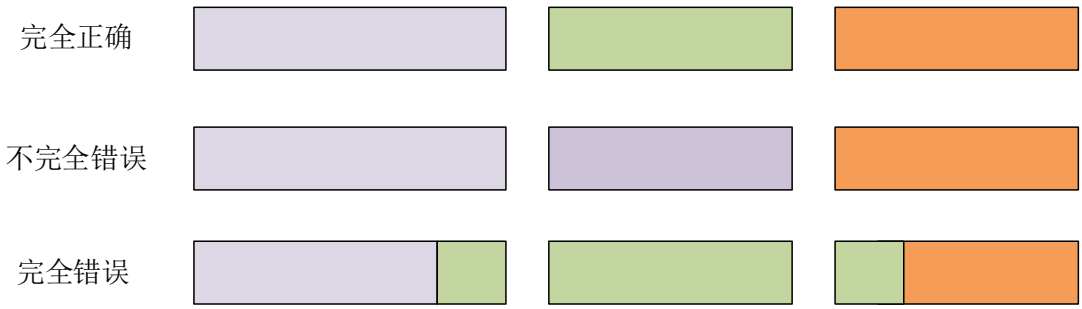


图 4-12 切割点情况分类图

依据不同的切割位置，预测的分割点可能出现三种情况，具体而言：第一，正确切割，即正确预测了切割点；第二，不完全错误切割，即两个切割点虽然只有一种业务类型的流量，但是该点为非分割点被错误预测；第三，完全错误切割，即两个切割点包含了两种不同类型的业务行为流量。以上情况可以使用图 4-12 表示。对于情况 2 来说，这并不是一次糟糕的切割，因为切割点之间是一种业务行为的 VPN 隧道流量，可以被识别出相应的业务类型，最坏的情况是情况 1，其会发生错误识别。由于情况 1 和情况 2 的存在，导致了整体切割精度没有特别高，但是本文最终目的是识别并分离多路复用的 VPN 隧道流量，因此可以接受该切割精度。

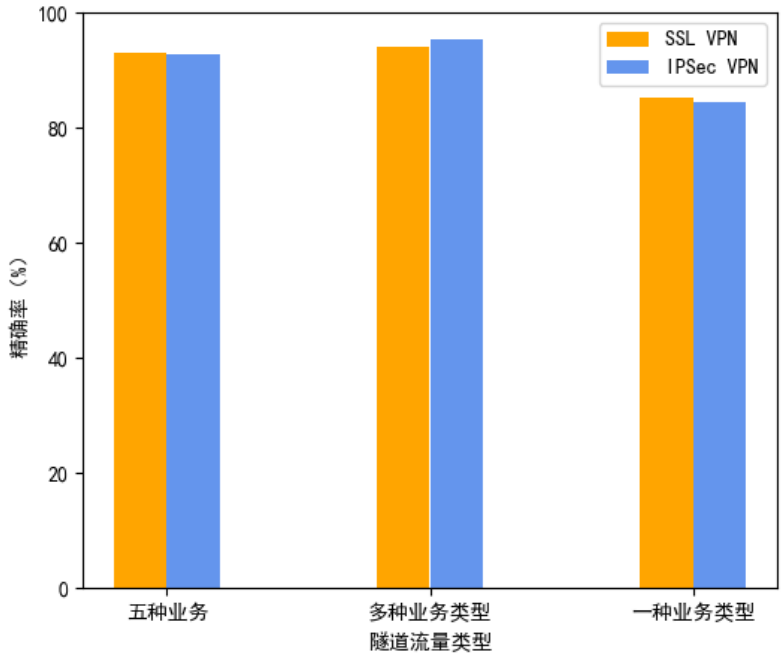


图 4-13 不同业务数量的 VPN 隧道流量精确率图

接下来，本文对比了具有不同业务类型数量的 VPN 隧道流量在窗口值为 4 时，所得到的精确率如图 4-13 所示。可以看出，只有一种业务类型的 VPN 隧道流量切割精确率最低，而具有多种类型的 VPN 隧道流量切割准确率最高。这是因为情况 2 的出现，导致了该问题，不过这一问题为可以接受的误差。多种业务类型与五种业务类型的 VPN 隧道流量切割准确率差别较小，这与本文在 4.3.3 节分析一致。此外，SSL VPN 隧道流量和 IPSec VPN 隧道流量在

多种业务类型的切割精度几乎一致，表明对于业务种类较多的 VPN 隧道流量，本方法更具有优势，能够获得更好的切割准确率。

（3）分流对比实验

为了说明本章方法的有效性和优越性，本章使用 SSL VPN 隧道流量标签数据集和 IPSec VPN 隧道流量标签数据集，共计 800 条隧道流量。本章对比了多类型 VPN 隧道流量切分方法、超时值方法、基于分割特征的 C4.5 方法、基于分割特征的传统随机森林算法，其中窗口值大小取 4，其结果如表 4-3 所示。其中本章方法为 SP-RF；超时值方法为 Timeout，其后数字代表超时阈值，Timeout-1 代表超时阈值取 1s，Timeout-0.5 代表超时阈值取 0.5s；RF 为未优化的随机森林算法。

表 4-3 对比模型各项性能指标结果表

性能指标	SP-RF	Timeout-1	Timeout-0.5	C4.5	RF
精确率	90.13%	42.12%	50.00%	78.61%	85.22%
预测时间（s）	1.5297	0.7892	0.7921	0.9127	1.4134
准确率	90.12%	28.38%	60.03%	60.39%	82.45%
召回率	91.31%	31.60%	60.00%	72.20%	87.24%
F_1指标	90.72%	28.96%	54.55%	75.29%	86.22%

从精确率、召回率、准确率、F_1指标可以看出本方法具有优越性。其中，基于超时值的方法评价指标过低，说明传统的 VPN 隧道流量分流方法并不适用于当前生产生活环境；相比于基于分割特征的 C4.5 方法，本章方法以决策树为基分类器进行强化学习，其学习能力由于基分类器；相比于基于分割特征的传统随机森林方法，本章方法依据切割数据集特征对随机森林算法进行了改进，使得随机森林算法可以更好地学习特征。但从预测时间来看，本章方法相比于超时值、C4.5 和 RF 效率低，原因如下：使用超时值法，其依据数据包时间戳寻找切割点，其只需要遍历数据包，那么预测时间与数据包长度相关，其他方法除了需要从隧道流量中提取特征外，还要输入模型当中进行预测；SP-RF 使用了决策树作为基分类器，每个决策树预测都需要时间，因此，其预测时间小于 C4.5；SP-RF 在 RF 的基础上基于数据集做出了优化，需要计算每个决策树的权重，因此，其预测时间小于 RF。综合五种评价指标来看，可以说明本章方法具有有效性，可以实现 VPN 隧道流量分流。

综上所述，本章对比了不同滑动窗口值下的 SSL VPN 隧道流量和 IPSec VPN 隧道流量的切割精确率，表明了 SSL VPN 隧道流量和 IPSec VPN 隧道流量在本章切割特征集中具有相同特征特性；对比了不同业务类型的 VPN 隧道流量的精确率，说明了 SL VPN 隧道流量和 IPSec VPN 隧道流量具有相同的单流特性表达；对比超时值、C4.5 和 RF，说明了本章方法的有效性和优越性。

4.6 本章小结

本章提出的多类型 VPN 隧道流量切分方法，在区分 IPSec VPN 隧道流量和 SSL VPN 隧道流量的基础上，使用基于队列的训练集构建方法建立标签特征训练集，使用基于滑动窗口的特征提取方法获取待分割 VPN 隧道流量的特征测试集。本章还设计了一种基于数据集改进

的随机森林分割模型，该模型使用标签特征训练集构建 SP-RF 模型，输入特征测试集获得 VPN 隧道流量的切割点预测，最终依据预测切割点切分待分割 VPN 隧道流量。该方法应用于 IPSec VPN 隧道流量可以获得 90.26%切割准确率，应用于 SSL VPN 隧道流量可以获得 90.13%准确率。本章对比了不同滑动窗口值下的 SSL VPN 隧道流量和 IPSec VPN 隧道流量的切割精确率，表明了 SSL VPN 隧道流量和 IPSec VPN 隧道流量在本章切割特征集中具有相同特征特性；对比了不同业务类型的 VPN 隧道流量的精确率，说明了 SSL VPN 隧道流量和 IPSec VPN 隧道流量具有相同的单流特性表达；对比其他方法，说明了本章方法的有效性和优越性。

第五章 面向 VPN 隧道流量的业务识别方法

本章提出的基于 VPN 隧道流量的业务识别方法是基于深度学习的方法，该方法省略了基于机器学习的流量识别分类方法中特征设计、特征提取、特征选择等步骤，使得模型可以自主从流量数据内容学习特征。本章方法主要分为两个阶段：第一阶段为 VPN 隧道流量业务识别序列提取，针对 SSL VPN 隧道流量和 IPSec VPN 隧道流量分别构建业务识别序列作为模型输入；第二阶段 VPN 隧道流量业务识别模型构建与预测，通过 SSL VPN 隧道流量和 IPSec VPN 隧道流量分别构建业务识别模型结构，使用待预测的 VPN 隧道流量进行最终预测。该流程总体框架如图 5-1 所示。

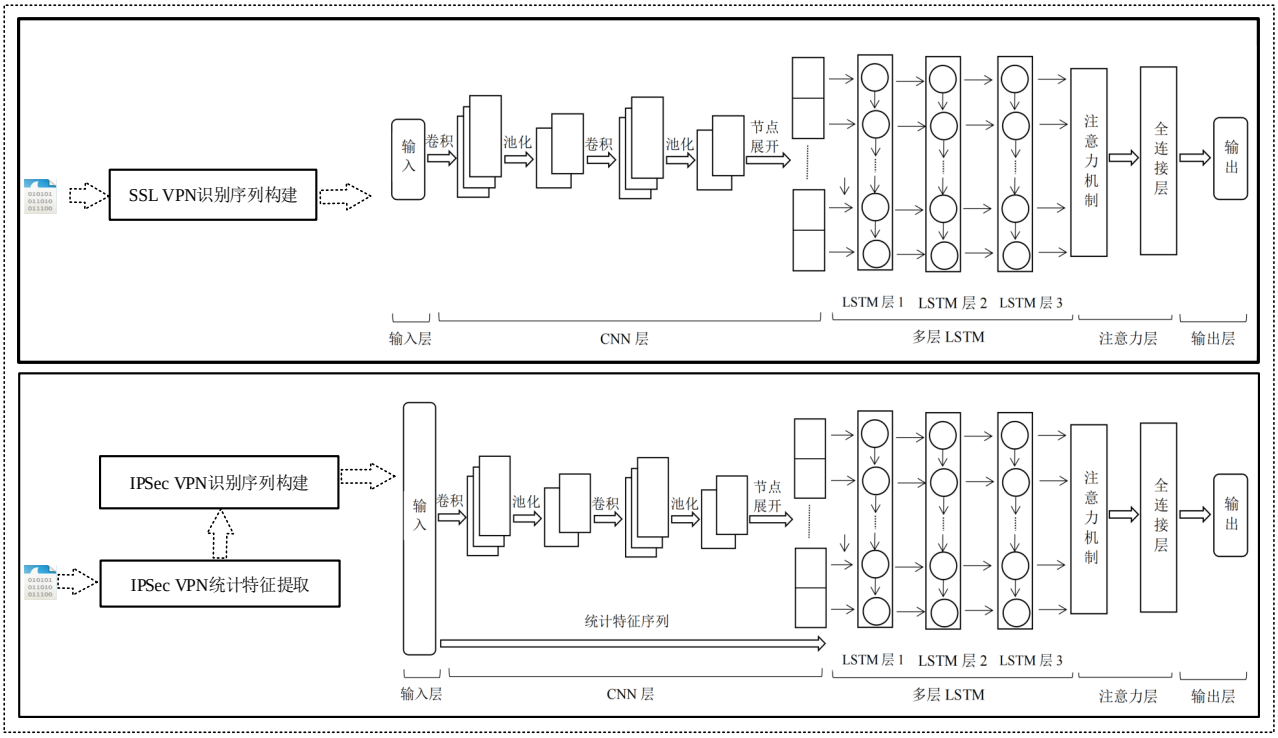


图 5-1 VPN 隧道流量业务识别方法整体流程图

5.1 问题分析

针对 VPN 加密流量业务分类问题的大多数研究都是基于 ISCX 2016 数据集，通过在单位时间在系统内运行一种应用，来获取具有该应用标签的 VPN 加密流量。实际情况下，不同业务应用可以使用同一 VPN 隧道进行信息传输，导致多种业务应用复用 VPN 隧道流量。目前，很少有研究针对 VPN 隧道流量的业务识别问题。基于以上研究背景，本章聚焦于经过分割后的 VPN 隧道流量业务分类算法的研究，其形式化定义如下：假设目标业务类型集合为$\boldsymbol{D} = \{d_1, d_2, d_3 \dots d_m\}$，对于给定的 SSL VPN 隧道流量$f = \{p_1, p_2, p_3 \dots, p_n\}$或 IPSec VPN 隧道流量$k = \{b_1, b_2, b_3 \dots, b_k\}$与分类方法$\mathcal{M}$，使得$\mathcal{M}(f||k) \in \boldsymbol{D}$。本文选取网页浏览、文件上传与下载、邮件传输、在线文字聊天 5 种业务类型，覆盖了大部分用户需求，因此m的取值范围为[1,5]。

VPN 业务识别方法可以分为深度学习方法和机器学习方法，二者最大的不同在于机器学习方法往往要通过特征工程构建指纹库或特征集合，而深度学习方法省略了基于机器学习的流量识别分类方法中特征设计、特征提取、特征选择等步骤，使得模型可以自主从流量数据内容学习特征。相比于非 VPN 加密流量，VPN 隧道流量对数据包负载增加封装，使得 VPN 隧道流量表现为端到端的单流特性，增加了 VPN 隧道流量全局特征的混乱程度。而深度学习方法中的人工神经网络可以从 VPN 加密流量序列中自动提取高维特征，更加适合 VPN 加密流量特征学习。除此之外，本章在分析 IPSec VPN 加密流量和 SSL VPN 加密流量全局特征混乱度时，发现 IPSec VPN 加密流量相比于 SSL VPN 加密流量负载特征更随机。因此，本章将在分析 IPSec VPN 加密流量和 SSL VPN 加密流量负载随机性的基础上，构建 IPSec VPN 隧道流量的流量统计特征，将其融入基于深度学习的 VPN 隧道流量框架，设计一种适用于 VPN 隧道流量经过切分后的碎片化 VPN 隧道流量的业务识别框架。

针对以上研究问题，本章的工作安排如下：

（1）针对 VPN 使用了代理机制和加密机制使得 VPN 隧道流量负载随机性增强，而基于机器学习的方法难以找寻其中的潜在特征关系，本章通过分析 IPSec VPN 加密流量和 SSL VPN 加密流量全局特征的差异性的基础上，提出了基于 SSL VPN 隧道流量的识别序列构建方法和基于 IPSec VPN 隧道流量的识别序列构建方法。同时，针对了 IPSec VPN 隧道流量，使用了 IPSec VPN 隧道流量的统计特征扩充了 IPSec VPN 隧道流量的识别序列；

（2）针对一维 CNN 模型只能提取 VPN 隧道流量输入序列的局部空间特征，无法反映局部特征与整体特征的关系，本章在分析 CNN 模型与 LSTM 模型的基础上，提出了一种基于注意力机制的 CNN-LSTM 模型，引入 LSTM 模型可以更好地学习 VPN 隧道流量识别序列的特征关系，解决了 CNN 模型局部视野受限的问题，引入注意力机制可以强化 LSTM 重要时间步长的作用，减少模型产生的预测误差。

实验结果表明，本章提出的方法具有有效性，可以完成 IPSec VPN 隧道流量和 SSL VPN 隧道流量业务类型分类任务。

5.2 VPN 隧道流量业务识别序列构建方法

本节首先通过分析 IPSec VPN 加密流量和 SSL VPN 加密流量的负载随机性，然后针对 SSL VPN 隧道流量和 IPSec VPN 隧道流量分别构造识别序列，其中对 IPSec VPN 隧道流量识别序列依据 IPSec VPN 隧道流量的统计特征进行扩充。

5.2.1 面向 SSL VPN 隧道流量的识别序列构建方法

针对 VPN 加密流量业务分类问题的大多数研究都是基于 ISCX 2016 数据集，该数据集通过使用 OpenVPN 获取单位时间 VPN 隧道只有一种业务类型的 SSL VPN 加密流量。针对该 SSL VPN 加密流量，学术界已经有相当多的端到端神经学习框架可以对其进行有效的业务类型识别。很多深度学习方法都是将提取 VPN 隧道双向流量前 6272 比特按字节转化为具有 uint8 数据格式像素值，即将 VPN 隧道流量转化为 28*28 大小的单通道灰度图，从中找寻 VPN

加密流量的特征关联。本文任意提取了第三章 SSL VPN 隧道流量标签数据集中只有一种业务类型的 SSL VPN 隧道流量的前 784 字节数据转化为单通道灰度图。如图 5-2 所示，从灰度图上很容易表现出不同业务行为的 SSL VPN 加密流量具有明显的特征区别。

图 5-2 SSL VPN 加密流量灰度图

经过第二、三、四章分析，IPSec VPN 相比于 SSL VPN 具有更高的安全级别，其通信机制更为复杂。因此，相比较 IPSec VPN 加密流量来说，使用 OpenVPN 获取的 SSL VPN 加密流量更容易获取全局特征。结合先前研究来看，对于 SSL VPN 加密流量只需要考虑其负载中的字节序列，不需要通过构建 SSL VPN 加密流量的统计特征加强深度学习框架的特征学习能力。为了验证这一猜想，本文使用 FS-Net 中的编码器-解码器（FS-E-D）结构[9]验证 IPSec VPN 加密流量、SSL VPN 加密流量以及非 VPN 加密流量全局特征的差异性。该编码器-解码器结构为一种重构机制，其通过重构原始流序列为相似流序列，使得相似流序列保持了更多特征信息，用于提高模型学习特征的能力。对于输入的网络流量数据包长度序列$P=\{p_1,p_2,p_3,p_4,p_5\dots p_n\}$，编码器使用多层双向 GRU 单元获取压缩的特征长度序列$F=\{f_1,f_2,f_3,f_4,f_5\dots f_m\}$，解码器使用相同的结构对压缩的特征长度序列进行还原，得到还原序列$P'=\{p'_1,p'_2,p'_3,p'_4,p'_5\dots p'_n\}$，损失函数计算重建的样本$P'$和原始输入$P$之间的差，其结构如图 5-3 所示。

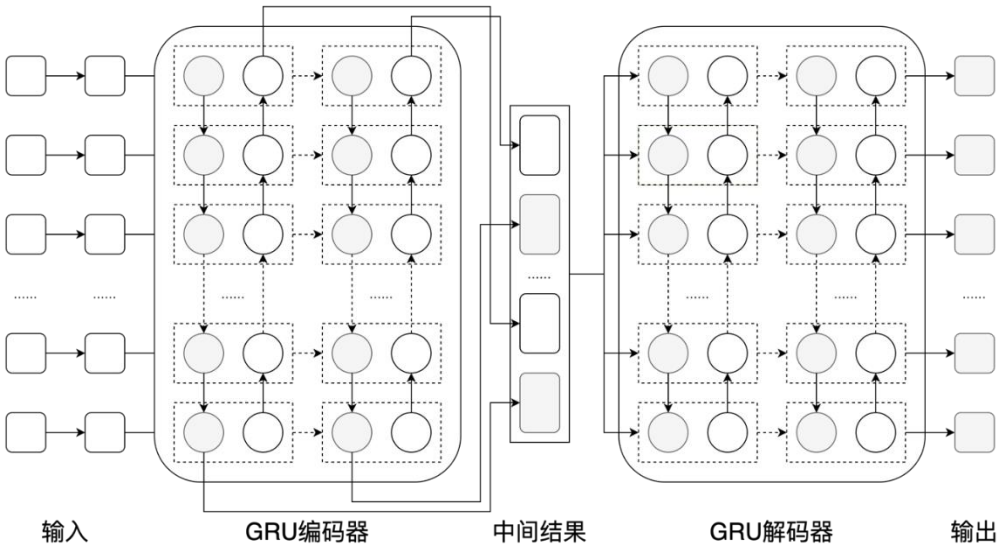


图 5-3 FS-E-D 结构图

本文使用 FS-E-D 结构计算重建的样本P'和原始输入P之间的差值Δ，该差值可以准确表明从 IPSec VPN 加密流量、SSL VPN 加密流量以及非 VPN 加密流量中获取全局特征的能力，其数学定义如下所示：

$$\Delta = \frac{\sum_{i=1}^{n}\left|l_i - l_i^{'}\right|}{n} \quad (5-1)$$

本文分别使用 100 组只包含一种业务类型的 IPSec VPN 加密流量、SSL VPN 加密流量以及非 VPN 加密流量构建 FS-E-D 结构，使用 50 组 IPSec VPN 加密流量、SSL VPN 加密流量和非 VPN 加密流量作为该结构的测试集。最终，本节获得 IPSec VPN 加密流量重构差值189.0745，SSL VPN 加密流量重构差值为160.3495，非 VPN 加密流量的重构差值为109.3176。由此可见，相比于非 VPN 加密流量，IPSec VPN 加密流量和 SSL VPN 加密流量的全局特征混乱程度更大，因此针对 IPSec VPN 隧道流量和 SSL VPN 隧道流量的业务识别问题需要借助深度学习模型。相比于 IPSec VPN 加密流量，SSL VPN 加密流量的负载混乱程度要低，深度学习模型能从 SSL VPN 加密流量序列中学习全局特征的能力更强，因此，SSL VPN 无需提取统计特征进行强化学习。

基于对 SSL VPN 加密流量的分析，本节将使用 SSL VPN 隧道流量的负载字节序列 SSL VPN 隧道流量业务识别序列，该序列将作为 SSL VPN 隧道流量的输入。在第四章中，通过窗口值实验，本文确定了对于 SSL VPN 隧道流量的窗口值为 4，即考虑以 4 个数据包作为序列单元进行分析时，基于 SSL VPN 隧道流量的分割精度最高。因为本章与第四章为前后关系，即先进行 VPN 隧道流量分流再进行本章实验，因此，本章需要以 4 个数据包作为序列单元进行识别序列构建。本节统计了 SSL VPN 隧道流量标签数据集，以 4 个数据包作为序列单元的平均数据包长为 4692.88。

综上所述，本节将提取会话流量的前 784 字节作为 SSL VPN 隧道流量业务识别框架的输入。该输入集合可以表示为$C_{SSL} = \{P\}$，其中P代表网络流量字节序列。对于 SSL VPN 隧道流量来说，MAC 层头部特征和 IP 层头部特征具有相似性。因此，本节所提取的会话流量前 784 字节将删除数据包 MAC 头部和 IP 层头部，然后将 SSL VPN 隧道流量进行拼接获取前 784 字节作为模型识别序列。

5.2.2 面向 IPSec VPN 隧道流量的识别序列构建方法

图 5-4 IPSec VPN 加密流量灰度图

根据上节重构损失数值来看，相比于 SSL VPN 加密流量，IPSec VPN 加密流量的负载混乱程度要低，深度学习模型能从 IPSec VPN 加密流量序列中学习全局特征的能力更弱。如图 5-4 所示，本文任意提取了一段 IPSec VPN 加密流量的前 784 字节数据转化为单通道灰度图，从图上很难表现出不同业务行为的 IPSec VPN 加密流量具有明显的特征区别。由此可见，先前的深度学习方法不完全适用于 IPSec VPN 隧道流量业务细粒度识别问题。基于以上问题，本文提出了一种用于 IPSec VPN 隧道流量的识别序列集合，该序列集合包含了 IPSec VPN 加密流量统计特征和 IPSec VPN 双向隧道流量负载信息，利用基于注意力机制的 CNN-LSTM 模

型强化学习不同业务行为 IPSec VPN 隧道流量之间的区别，实现 IPSec VPN 隧道流量业务行为识别。

某段时间的流量具有自相似性[50]，这种自相似可以表现在流量统计特征。本文将分析不同业务行为 IPSec VPN 加密流量的数据包分布和数据到达时间分布，并在此基础上构建不同业务行为下 IPSec VPN 隧道流量特征集合。

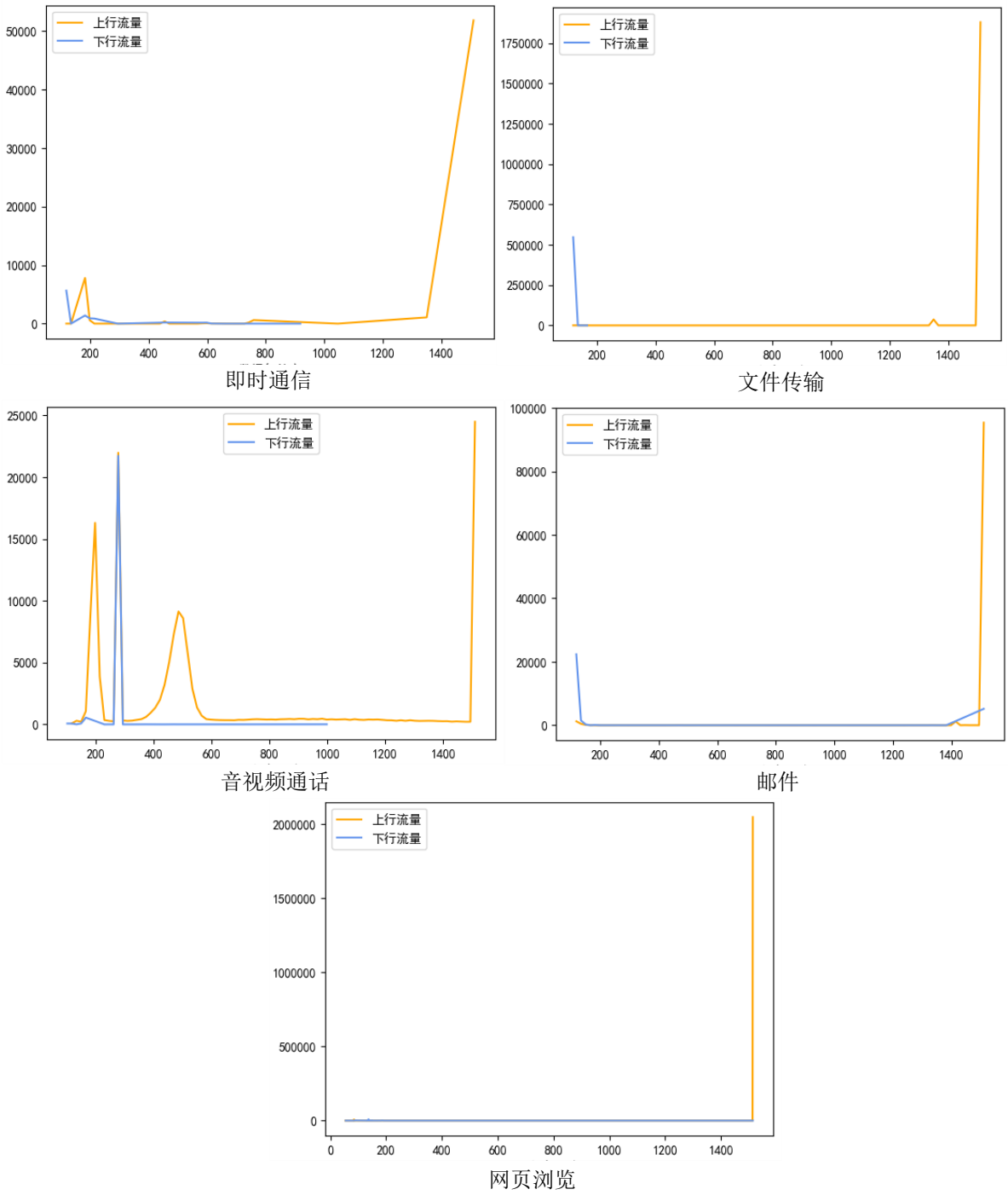


图 5-5 IPSec VPN 加密流量数据包长度统计特征图

如图 5-5 所示，本章统计了第三章数据集中只包含一种业务类型的 IPSec VPN 加密流量在五种业务行为双向数据包长度频数，该频数反映了 IPSec VPN 加密流量在不同业务行为中

数据包长度分布特性。其中，规定用户发送数据包方向为下行方向，横坐标代表数据包长度分布，纵坐标代表不同数据包长度的数据包数量分布频数。

可以看出不同业务类型的数据包长度分布具有明显的区分，并由此得出如下信息：

（1）即时通信下行 IPSec VPN 加密流量数据包长度大都均匀分布区间[0,950]，上行 IPSec VPN 加密流量数据包长度在区间[1350,1400]具有较多数量，且长度小于 200 的数据包也会有一个较大的占比。

（2）文件传输下行 IPSec VPN 加密流量数据包长度都小于 200 且随着包长的增加数据包频率明显减少，上行 IPSec VPN 加密流量的数据包长大都在 1450 以上。

（3）音视频通话下行 IPSec VPN 加密流量的数据包长度在 250 左右有较明显的增多，在区间[0,1000]内分布较为均匀，而下行 IPSec VPN 加密流量的数据包长度在 180、500、1500 左右具有明显增多趋势.

（4）邮件下行 IPSec VPN 加密流量的数据包长度在 100 总有有较大占比，其余数据包长度分布均匀，下行 IPSec VPN 加密流量的数据包长度在 1500 处分布较多。

（5）网页浏览的下行 IPSec VPN 加密流量数据包长度分布均匀，而上行 IPSec VPN 加密流量数据包长度在 1500 左右明显增多。

由此可见，不同业务行为的 IPSec VPN 数据包长度分布具有明显区分，数据包长度统计特征作为 IPSec VPN 加密流量业务识别特征具有理论依据。

本文在分析不同业务行为的 IPSec VPN 加密流量在数据包长度分布的差异性的基础上，提出了 IPSec VPN 加密流量业务识别统计特征。本文在参考 CICFlowmeter 工具使用的数据统计特征，基于数学方法对 IPSec VPN 加密流量数据包长度特征进行扩充最终得到如下统计特征集合：

$$C_{STATIC\_IPSEC} = \{L, R\}$$

其中，L为 IPSec VPN 加密流量数据包长相关特征集合，R为 IPSec VPN 加密流量上行流量与下行流量之比的相关特征集合该特征参数解释如表 5-1 所示：

表 5-1 IPSec VPN 隧道流量统计特征

特征符号	特征解释	特征参数	参数解释
L	流中数据包的包长	$l_{up\_avg}, l_{up\_var}, l_{up\_max}, l_{up\_min}$	上行流量的均值，方差，最大值，最小值
		$l_{dn\_avg}, l_{dn\_var}, l_{dn\_max}, l_{dn\_min}$	下行流量的均值，方差，最大值，最小值
R	上行流量与下行流量的比值	$R_{bytes}, R_{packets}$	字节数比，包个数比

虽然 IPSec VPN 隧道流量具有明显的流量统计特征，但是其负载内部特征被 IPSec VPN 的加密机制和代理机制掩盖，很难通过统计学方式对特征进行提取。为了说明其负载内部特征可以被深度学习模型自动提取并进行学习预测，本章将使用二维 CNN 模型对 IPSec VPN 隧道流量进行预实验，通过实验结果说明深度学习模型在提取并学习 IPSec VPN 隧道流量具有优势，并由此构建 IPSec VPN 隧道流量的识别序列。本章选取第三章数据集中具有一种业务类型的 IPSec VPN 加密流量共 500 条流，提取了双向会话流量前 784 字节构建大小为 28*28

单通道灰度图，按照 9：1 的比例分割灰度图集合为训练集和测试集，使用二维 CNN 深度学习模型进行训练和测试，最终获得平均准确率为 90.91%，其中各种业务类型召回率、精确率、F_1指标如表 5-2 所示。

表 5-2 对比模型各项性能指标结果表

性能指标	即时通信	文件传输	音视频通话	邮件传输	网页浏览
精确率	78.57%	91.67%	100.00%	100.00%	91.67%
召回率	100.00%	100.00%	72.73%	90.91%	100.00%
F_1指标	88.00%	95.65%	84.21%	95.24%	95.65%

从上述实验结果来看，CNN 模型对于 IPSec VPN 加密流量具有敏感性，IPSec VPN 双向流量具有隐藏特征可以通过 CNN 深度学习模型认知。而对于机器学习方法来说，IPSec VPN 数据包负载经过 ESP 协议加密具有高度随机性，很难构造特征进行学习。因此，可以利用深度学习模型自动学习 IPSec VPN 双向流量的负载潜在特征实现 IPSec VPN 加密流量业务识别。同时，本节统计了 IPSec VPN 隧道流量标签数据集，以 4 个数据包作为序列单元的平均数据包长为 4701.24。由此，本文对统计特征集合进行扩充，得到基于 IPSec VPN 加密流量业务识别特征集合 SN：

$$C_{IPSEC} = \{L, C_{STATIC\_IPSEC}\}$$

其中，L为 IPSec VPN 隧道流量前 784 字节，该会话流量前 784 字节将删除数据包 MAC 头部和 IP 层头部，然后将 IPSec VPN 隧道流量进行拼接获取前 784 字节。C_{STATIC_IPSEC}为基于 IPSec VPN 加密流量的统计特征集合序列，其详细解释如表 5-1 所示。

5.3 面向 VPN 隧道流量的业务识别模型

本节的重点在于构建基于 VPN 隧道流量的业务识别模型。本章提出了一种基于注意力机制的 CNN-LSTM 模型，该模型通过一维 CNN 对 VPN 隧道流量的空间特征进行抽象与提取，接下来通过 LSTM 模型对局部特征中潜在时间特征和空间特征进行学习，并通过注意力机制减少 LSTM 模型预测误差。针对 SSL VPN 识别序列和 IPSec VPN 识别序列，该模型的结构略有不同。

5.3.1 多类型 VPN 识别序列

从先前研究来看，学者将网络流量表示形式分为语义表示和二进制表示。前者是使用数据包头部某些特殊意义的字段进行拼接表征网络流量，但是这种方法忽略了数据包全局特征，使得数据包负载和其他字段特征丢失；后者则是将网络流量按字节转化为单通道灰度图，利用深度学习模型在图像处理的优势进行分类，这种方法忽略了比特之间的联系，增加了噪声，使得模型对相同的特征做出了不同的定义。基于以上问题，本文将 VPN 隧道流量负载序列视为一个具有n维特征的字节流，那么对于 SSL VPN 识别序列来说，可以使用如下数学定义表示：

$$B_{SSL} = b_1 \oplus b_2 \oplus b_3 \oplus \dots b_{784} \qquad (5-2)$$

其中，b_i代表 SSL VPN 隧道流量 IP 层负载序列中的数据字节。而对于 IPSec VPN 识别序列来说，其数学定义如下：

$$B_{IPSEC} = b_1 \oplus b_2 \oplus b_3 \oplus \ldots b_{784} \oplus L \oplus R \qquad (5-3)$$

其中，b_i代表 IPSec VPN 隧道流量 IP 层负载序列中的数据字节，L代表 IPSec VPN 隧道流量单元序列数据包长度相关统计特征$L = l_{up_avg} \oplus l_{up_var} \oplus l_{up_max} \oplus l_{up_min} \oplus l_{dn_avg} \oplus l_{dn_var} \oplus l_{dn_max} \oplus l_{dn_min}$，$R$代表 IPSec VPN 隧道流量单元序列上行流量与下行流量的比值相关统计特征$R = R_{bytes} \oplus R_{packets}$。

从以上识别序列数学表达公式来看，SSL VPN 识别序列和 IPSec VPN 识别序列的特征维度并不相同，因此，针对两种 VPN 识别序列的模型结构也有所区别。

5.3.2 基于 SSL VPN 识别序列的业务识别模型

CNN 可以有效的从数据空间中学习局部特征，但是由于其视野受限，会错过整体特征与局部特征之间的联系。LSTM 可以挖掘一定长度的序列数据的时间依赖关系。因此，本章将借助 CNN 对短序列有较强的特征抽象能力获取高维度的特征，然后利用 LSTM 更有效地从高维特征中学习时间依赖关系，再利用注意力机制减少 LSTM 预测误差。通过本章设计的基于注意力机制的 CNN-LSTM 模型实现 VPN 隧道流量特征细粒度学习，最终完成分类任务。接下来，本文首先介绍基于 SSL VPN 识别序列的业务识别模型，然后介绍基于 IPSec VPN 识别序列的业务识别模型，并在此过程中比较二者的异同。

其结构如图 5-6 所示，其主要包括五个部分：输入单元、CNN 单元、LSTM 单元、注意力单元、输出单元，各个单元解释如下所示。

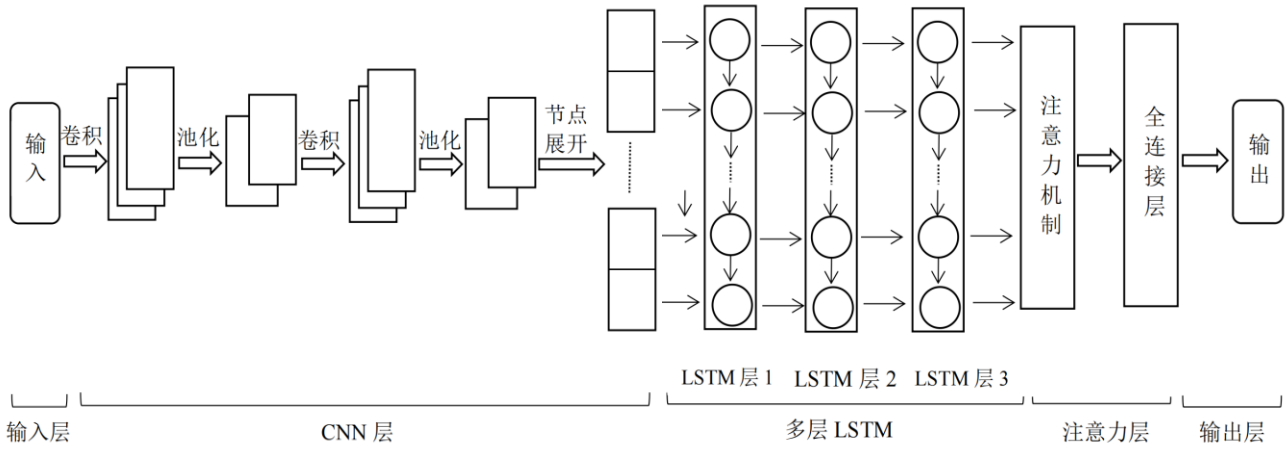


图 5-6 基于 SSL VPN 识别序列的业务识别结构图

输入单元：该单元主要是提取 SSL VPN 识别序列转化为 CNN 单元输入数据格式。对于构建业务识别模型的过程来说，提取所有长度为 4 的单元数据包序列，删除数据包 IP 头部和 MAC 头部，将 IP 层负载数据进行拼接，取前 784 字节数据作为 SSL VPN 业务识别模型构建输入；对于业务类型分类预测来说，删除 VPN 隧道流量数据包 IP 头部和 MAC 头部，将 IP

层负载数据进行拼接，取前 784 字节数据作为 SSL VPN 业务识别序列，将其作为 SSL VPN 业务识别模型的预测输入。

CNN 单元：该单元主要完成 SSL VPN 业务识别序列的高维特征提取。首先将 SSL VPN 业务识别输入到卷积层提取 SSL VPN 隧道流量的高维特征图，接下来使用 ReLU 函数对其进行非线性化映射，再使用池化层将高维特征图进行降维，减少参数计算量，以上过程本文称为一次卷积池化过程。接下来，再使用一次卷积池化过程，该过程将进一步凝练 SSL VPN 隧道流量的特征，并进一步提取其中的显著特征，较少特征图维度。具体卷积过程计算公式可以参考 2.3.2 节。

LSTM 单元：该单元主要对高维特征图中的时序关系进行学习。该单元使用了多层 LSTM 是因为基于经验来说，多个 LSTM 层次堆叠可以更好地学习特征中的时间依赖关系，但是如果堆叠层数过多，那么计算量就会增大。具体的计算过程可以参考 2.3.1 节。因此，本模型使用了三层 LSTM，并将其丢弃率设置为 50%来获得更好的泛化能力。

注意力单元：该单元主要用来降低模型预测误差。该单元的本质是获取第三层 LSTM 隐藏层输出向量的加权平均和。具体过程为注意力层获取 LSTM 隐藏层输出向量，使用全连接层获取训练输出结果，对该结果使用$softmax$函数进行归一化操作。基于以上，可以得到每一个隐藏层向量的分配权重，其数值大小可以代表影响预测结果的时间步隐层状态的重要性。

输出层：该层输出待识别 SSL VPN 隧道流量的最终业务识别结果。

5.3.3 基于 IPSec VPN 识别序列的业务识别模型

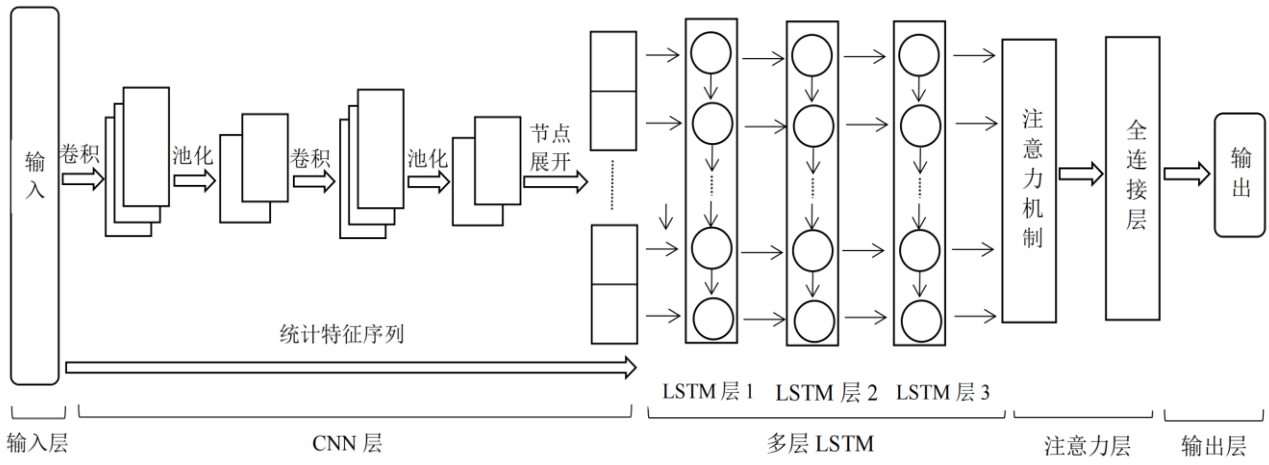


图 5-7 基于 IPSec VPN 识别序列的业务识别结构图

基于 IPSec VPN 识别序列的业务识别模型整体架构图 5-7 所示。可以看出，该结构与基于 SSL VPN 识别序列的业务识别模型结构相似。这是因为本章提出的业务识别模型主要结构为基于注意机制的 CNN-LSTM，而 IPSec VPN 识别序列相比于 SSL VPN 识别序列多出了 IPSec VPN 隧道流量的负载序列统计特征，该特征经过人工提取不依赖于 CNN 单元的特征提取因此，该特征序列直接作为 LSTM 单元的输入。以下简要介绍各个单元。

输入单元：该单元主要是提取 IPSec VPN 识别序列中 VPN 字节序列转化为 CNN 单元输入数据格式。其过程与处理 SSL VPN 识别序列过程一致，这里不再赘述。

CNN 单元：该单元主要完成 IPSec VPN 识别序列中 VPN 字节序列的高维特征提取。其过程与处理 SSL VPN 识别序列过程一致，这里不再赘述。

LSTM 单元：该单元主要对高维特征图和 IPSec VPN 隧道流量的负载序列统计特征的时序关系进行学习。相比于 SSL VPN 识别序列提取出的高维特征图，LSTM 单元的输入维度增多。其计算过程详见 2.3.1 节。

注意力单元：该单元主要用来降低模型预测误差。该单元与 SSL VPN 隧道流量业务识别模型中的一致，在此不再赘述。

输出层：该层输出待识别 IPSec VPN 隧道流量的最终业务识别结果。

综上所述，本节根据 IPSec VPN 识别序列和 SSL VPN 识别序列分别设计了 IPSec VPN 隧道流量的业务识别模型和 SSL VPN 隧道流量的业务识别模型，接下来本章将通过实验验证业务识别模型的有效性。

5.4 实验设计与结果分析

5.4.1 实验环境

本章实验在物理机上部署实验代码，物理机的 CPU 为 AMD Ryzen 7 3700X 8-Core Processor 3.60 GHz，GPU 为 RTX3090（24GB），内存为 64GB，使用的操作系统及版本为 Windows 10 专业版，程序编译环境为 Pycharm，Python 版本为 python 3.6.5。

5.4.2 数据集组成与模型参数设置

（1）数据集选取

本章使用第三章构建的 IPSec VPN 隧道流量标签数据集和 SSL VPN 隧道流量标签数据集，两种数据集的业务类型、流量来源、数量均相同，具体描述如表 5-3 所示。

表 5-3 VPN 隧道流量标签数据集

编号	业务类型	流量来源	数量
1	即时通信	微信、QQ、自建聊天服务器	400
2	网页浏览	自建网站、Edge	400
3	邮件	自建邮箱服务器	400
4	音视频	自建服务器、腾讯会议	400
5	文件传输	百度网盘、自建文件传输服务器	400

除此之外，本章还将使用经过切割的 IPSec VPN 隧道流量和 SSL VPN 隧道流量对本方法进行测试，其详细介绍如表 5-4 所示。

表 5-4 VPN 隧道流量切割数据集

编号	业务类型数量	SSL VPN 隧道流量	IPSec VPN 隧道流
1	五种	150	150
2	多种	150	150
3	一种	100	100

（2）参数设置

本章首先需要确定 CNN 单元参数，第一个卷积层具有 32 个大小为[1,25]的卷积核，设置步长为 1，之后使用 ReLU 函数作为激励函数；第一个池化层的过滤器尺寸为[1,3]，设置步长为 3；第二个卷积层有 64 个卷积核，其尺寸与第一个卷积层的卷积核尺寸一致，设置步长为 1，使用 ReLU 函数作为激励函数；第二个池化层的过滤器尺寸为[1,3]，设置步长为 3。然后，将获得的 32*32 大小数据作为 3 层 LSTM 输入。本模型优化器选择 Adam，批大小为 50，训练轮次为 50。对于 SSL VPN 识别序列，每个 LSTM 层具有 256 个神经细胞，设置 dropout 为 50%；对于 IPSec VPN 识别序列，每个 LSTM 层具有 266 个神经细胞，设置 dropout 为 50%。经过注意力层和全连接层最终得到输出，其中全连接层使用$softmax$函数。

本实验使用交叉熵（Cross-entropy loss）损失函数作为基于 SSL VPN 识别序列和 IPSec VPN 识别序列的业务识别模型的损失函数，用于衡量模型预测值和真实值之间的误差。其数学表达如下所示。其中，N为样本数量，C为类别数量。若第i个样本类别为c时，$y_{ic}=1$；反之，则$y_{ic}=0$。p_{ic}为第i个样本属于类别c的概率。

$$loss=\frac{1}{N}\sum_{i=1}^{N}l_i=-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}y_{ic}\log(p_{ic}) \qquad (5-4)$$

5.4.3 对比实验选取

本章的目标是实现 VPN 隧道流量的业务行为类型的识别，提出了一种基于注意力机制的 CNN-LSTM 模型，因此，本文将使用的 SSL VPN 隧道流量数据集对比二维 CNN 模型和一维 CNN 模型，使用 IPSec VPN 隧道流量标签数据集对比周益旻[19]提出基于统计特征的 IPSec VPN 加密流量业务识别方法，用于表明本方法提出的业务识别模型具有良好的分类性能。除此之外，本章将进行消融实验，逐步删除本模型中的组件，并将其与完整模型进行对比，证明构建本模型在提取特征方面具有优越性。

5.4.4 评价指标

本章将使用准确率来评估本章方法整体的分类性能，除此之外对于每一种类别，本章将使用精确率、召回率、F_1指标评估本章方法的分类能力。其概念已在 4.5.4 中给出，在此不再赘述。

除了以上指标，本章还将使用预测时间说明模型对输入数据预测速度，其计算方式为模型预测 100 个输入样本所使用的平均时间。但是考虑到实际计算时计算资源被抢占式调度，模型预测时不能完全获取 CPU、GPU 资源。因此，该指标仅作为相对指标用于评价对比实验所使用方法的处理效率。

由于本章实验为了尽量减少数据集划分对分类结果的影响，使用 10 倍交叉验证方法用于实验分类性能，每一种指标的最终取值为 10 次实验的平均值。

5.4.5 结果与分析

（1）业务识别结果与分析

本文首先展示使用 IPSec VPN 切割数据集和 SSL VPN 切割数据集在以上参数设置下获取的各项评价指标，如表 5-5 所示。

表 5-5 VPN 切割流量各项性能指标结果表

流量类型	性能指标	即时通信	文件传输	音视频	邮件传输	网页浏览
IPSec VPN	准确率	92.17%				
	预测时间(s)	6.5315				
	精确率	91.01%	91.67%	89.71%	88.97%	92.20%
	召回率	92.13%	92.09%	90.75%	90.91%	92.15%
	F_1指标	91.57%	91.88%	90.22%	89.93%	92.17%
SSL VPN	准确率	92.84%				
	预测时间(s)	5.5298				
	精确率	92.11%	91.89%	90.00%	90.00%	93.01%
	召回率	92.13%	92.09%	91.67%	91.10%	92.16%
	F_1指标	92.12%	91.99%	90.82%	90.55%	92.58%

从整体准确率可以看出，本章提出的方法具有良好的分类性能。本章提出的方法获得了比切割 VPN 隧道流量更高地精度，这是因为预测的分割点可能出现如下情况：第一，完全错误切割，即两个切割点包含了两种不同类型的业务行为流量；第二，不完全错误切割，即两个切割点虽然只有一种业务类型的流量，但是该点为非分割点被错误预测；第三，正确切割，即正确预测了切割点。对于情况 2 而言，该切割点虽然错误但是不影响其分类，因此准确率会有所上升。这在此印证了基于数据包序列的单流 VPN 切割方法的有效性。

从准确率、精确率、召回率、F_1指标来看本模型对于 IPSec VPN 切割数据集的敏感度较差。根据 5.2.1 小节对于 SSL VPN 隧道流量和 IPSec VPN 隧道流量的全局特征分析，IPSec VPN 隧道流量全局特征更为混乱。因此，相比于 IPSec VPN 切割流量，基于 VPN 切割流量的业务识别模型更容易从 SSL VPN 切割流量学习潜在特征。从识别时间来看，本模型对于 IPSec VPN 切割流量识别时间几乎和 SSL VPN 切割流量一致，说明 IPSec VPN 识别序列虽然比 SSL VPN 识别序列长度多 10 字节，但是不会引起识别时间发生较大变化。这是因为相比于识别序列总长度，IPSec VPN 识别序列增加的长度较小。

接下来，展示使用 SSL VPN 隧道流量数据集和 IPSec VPN 隧道流量数据集获取的各项评价指标，如表 5-6 所示。

表 5-6 VPN 隧道流量各项性能指标结果表

流量类型	性能指标	即时通信	文件传输	音视频	邮件传输	网页浏览
IPSec VPN	准确率	93.00%				
	预测时间(s)	0.1598				
	精确率	93.42%	94.33%	94.10%	93.20%	91.67%
	召回率	95.46%	93.57%	92.15%	94.75%	95.25%
	F_1指标	94.43%	93.95%	93.11%	93.97%	93.43%
SSL VPN	准确率	94.22%				
	预测时间(s)	0.1733				
	精确率	96.22%	95.10%	95.56%	94.10%	93.00%

流量类型	性能指标	即时通信	文件传输	音视频	邮件传输	网页浏览
	召回率	95.88%	94.00%	94.11%	95.00%	96.01%
	F_1指标	96.05%	94.55%	94.83%	94.55%	94.48%

可以看出相比于使用切割数据集，使用业务识别数据集的模型具有更好地分类性能，这是因为分割数据集出现了情况 1，导致了数据包被错误识别，因此整体分类性能下降。业务识别数据集无需切割，不具有单条流量中含有多种业务行为的情况，因此，可以最直观地反映本方法的分类性能。从准确率、精确率、召回率、F_1指标来看本模型对于 IPSec VPN 隧道流量数据集的敏感度较差，其分析与 IPSec VPN 切割数据集一致，其原因不再赘述。此外，对于 VPN 隧道流量来说，其识别时间相较于 VPN 切割流量的识别时间要短，原因是 VPN 切割流量会产生情况 1-3，为碎片化 VPN 隧道流量，每个碎片都需要识别，且碎片较多，由此导致了预测时间较长。而数据集中 VPN 隧道流量为只包含一种业务类型的 VPN 加密流量，只需要提取一次 VPN 识别序列即可完成识别任务，因此，可以将 VPN 切割流量看作多条 VPN 隧道流量，其识别时间自然增多。

（2）对比实验结果分析

首先展示基于 SSL VPN 隧道流量数据集的模型对比结果，其中本文方法为 CLA。结果如表 5-7 所示。从准确率、精确率、F1 指标、召回率可以看出本方法具有优越性，对于 SSL VPN 数据集具有良好的业务识别能力。不过从处理时间上来看，本章方法并不具有优势，这是因为本章是基于注意力机制的 CNN-LSTM 模型具有较为复杂的结构，其计算量大于另外两种方法。

表 5-7 SSL VPN 隧道流量对比模型各项性能指标结果表

性能指标		CLA	1D-CNN	2D-CNN
分类准确率		94.22%	89.24%	87.13%
预测时间（s）		0.1733	0.1450	0.1203
即时通信	精确率	96.22%	90.13%	88.77%
	召回率	95.88%	89.20%	88.20%
	F_1指标	96.05%	89.66%	88.48%
文件传输	精确率	95.10%	88.09%	89.42%
	召回率	94.00%	87.40%	87.00%
	F_1指标	94.55%	87.74%	88.19%
音视频	精确率	95.56%	90.17%	88.77%
	召回率	94.11%	91.40%	84.00%
	F_1指标	94.83%	90.78%	86.32%
邮件传输	精确率	94.10%	87.21%	86.21%
	召回率	95.00%	88.60%	88.20%
	F_1指标	94.55%	87.90%	87.19%
网页浏览	精确率	93.00%	87.18%	85.24%
	召回率	96.01%	87.80%	85.80%
	F_1指标	94.48%	87.49%	85.52%

接下来，本章使用 IPSec VPN 隧道流量数据集进行对比，其结果如表 5-8 所示，其中基于统计特征的 IPSec VPN 加密流量业务识别方法表示为 SF。

可以看出，基于统计特征的 IPSec VPN 加密流量业务识别方法对于本文自建的 IPSec VPN 隧道流量标签数据集也有着较好分类结果。但从准确率、精确率、F1 指标、召回率、识别时

间来看，本章的方法更具有优势。这是因为本章在提取 IPSec VPN 隧道流量统计特征的基础上，使用深度学习模型从其数据负载中进行特征学习，表明了深度学习方法更适用于 IPSec VPN 隧道流量的业务识别问题。此外，本章方法相比于基于统计特征的 IPSec VPN 加密流量业务识别方法提升并不大，这从侧面说明了 IPSec VPN 加密流量具有高度随机性，目前还没有更好的模型结构可以完成对 IPSec VPN 加密流量的业务识别。

表 5-8 IPSec VPN 隧道流量对比模型各项性能指标结果表

性能指标		CLA	SF
分类准确率		93.00%	92.14%
预测时间（s）		0.1598	3.9204
即时通信	精确率	93.42%	91.28%
	召回率	95.46%	90.83%
	F_1指标	94.43%	91.05%
文件传输	精确率	94.33%	93.77%
	召回率	93.57%	91.34%
	F_1指标	93.95%	92.54%
音视频	精确率	94.10%	91.45%
	召回率	92.15%	92.10%
	F_1指标	93.11%	91.77%
邮件传输	精确率	93.20%	92.01%
	召回率	94.75%	92.14%
	F_1指标	93.97%	92.07%
网页浏览	精确率	91.67%	91.22%
	召回率	95.25%	90.80%
	F_1指标	93.43%	91.01%

（3）消融实验

通过逐步删除本模型中的组件，并将其与完整模型进行对比，可以证明构建本模型在提取特征方面具有优越性，主要包括删除 CNN 单元的 LSTM 单元、删除 LSTM 单元的 CNN 单元、删除注意力机制的 CNN-LSTM 单元。结果展示如表 5-9 所示，其中的结果基于 IPSec VPN 数据集。

表 5-9 消融实验各项性能指标结果表

性能指标		CLA	CNN	LSTM	CNN-LSTM
分类准确率		93.00%	85.14%	74.29%	90.45%
预测时间（s）		0.1598	0.1436	0.1245	0.1499
即时通信	精确率	93.42%	85.26%	73.52%	91.01%
	召回率	95.46%	84.65%	74.01%	90.83%
	F_1指标	94.43%	84.95%	73.76%	90.92%
文件传输	精确率	94.33%	88.24%	73.61%	89.88%
	召回率	93.57%	86.11%	72.46%	90.11%
	F_1指标	93.95%	87.16%	73.03%	89.99%
音视频	精确率	94.10%	84.22%	75.71%	90.03%
	召回率	92.15%	85.00%	76.27%	89.93%
	F_1指标	93.11%	84.61%	75.99%	89.98%
邮件传输	精确率	93.20%	82.14%	72.10%	88.43%
	召回率	94.75%	85.90%	71.98%	90.45%
	F_1指标	93.97%	83.98%	72.04%	89.43%
网页浏览	精确率	91.67%	84.23%	74.80%	91.46%

性能指标		CLA	CNN	LSTM	CNN-LSTM
	召回率	95.25%	83.47%	75.34%	90.00%
	F_1指标	93.43%	83.85%	75.07%	90.72%

从以上结果来看，无论删除哪一个单元，模型的分类性能都会有所下降。其中删除注意力机制的 CNN-LSTM 模型下降最小，这说明注意力机制可以优化 LSTM 单元，加强 CNN 对显著特征的解释。删除 CNN 单元，LSTM 单元的分类准确率下降最多，这说明 LSTM 只能从网络流量序列中学习时间相关的特征，而不能从序列中提取空间特征，这与文章一开始的理论分析相同；删除了 LSTM 单元，分类性能也有下降，这说明 CNN 单元不能完全学习网络流量序列中全部的特征。以上实验结果验证了本文基于注意力机制的 CNN-LSTM 模型具有良好的分类性能，可以完成业务识别任务。

综上所述，本文构建的 VPN 加密流量业务识别方法可以很好的对 VPN 分割流量和 VPN 纯净业务流量实现业务识别任务。

5.5 本章小结

本章介绍基于 VPN 隧道流量的业务识别方法。针对 VPN 使用了代理机制和加密机制使得 VPN 隧道流量负载随机性增强，而基于机器学习的方法难以找寻其中的潜在特征关系的问题，在分析 IPSec VPN 加密流量和 SSL VPN 加密流量全局特征的差异性的基础上，提出了基于 SSL VPN 隧道流量的识别序列构建方法和基于 IPSec VPN 隧道流量的识别序列构建方法。在此基础上，提出了一种基于注意力机制的 CNN-LSTM 模型，解决了 CNN 模型和 LSTM 模型从 VPN 隧道流量中学习特征的不足之处。实验部分，本章使用 IPSec VPN 切割数据集和 SSL VPN 切割数据集获取了经过第四章切割方法后，IPSec VPN 隧道流量和 SSL VPN 隧道流量最终分类结果。同时对比了多种业务识别方法，表明了本章方法在解决 VPN 隧道流量业务识别问题上具有优势。

第六章 总结与展望

本章主要对本文所研究的内容进行总结，并分析本文研究的不足之处，在此基础上，对未来研究进行展望。

6.1 总结

随着用户对于匿名化网络访问的需求不断增加，VPN 技术也随之发展。经过 COVID-19 疫情，VPN 流量已经占据加密流量市场相当大的份额。在 VPN 加密流量骤增的背后，许多不法分子将攻击意图隐藏于 VPN 数据包负载中，给网络安全部门带来了很大的网络监管压力。针对 VPN 流量开展业务类型识别研究，探索 VPN 数据包中潜在信息可以提升网络流量监管水平。为了提高对 VPN 隧道流量的分析能力，加强对 VPN 工具的监管能力，需要对 IPSec VPN、SSL VPN 这两种占据市场最大份额的 VPN 加密流量实现高效且准确的业务类型识别。本文将在 IPSec VPN 加密流量、SSL VPN 加密流量实现流量区别的基础上，提出了一种有效的 VPN 隧道流量的分流方法，实现 VPN 隧道流量的切割，并设计一种 IPSec VPN 隧道流量、SSL VPN 隧道流量的业务行为识别方法实现 VPN 隧道流量的业务类型识别。

本文研究细化而言，可以概括为以下内容：

（1）面向多类型 VPN 隧道流量数据集构建方法

目前对于 VPN 的研究大都使用 ISCX 2016 数据集，其在获取时只考虑了 VPN 隧道内只有一种应用类型的流量。相比较于非 VPN 流量，普通加密流量可以按照五元组方法获取流量对应的标签，而 VPN 具有代理混淆的特性，具有不同业务行为的应用可以使用同一 VPN 隧道，导致 VPN 隧道多路复用的问题，难以通过非 VPN 加密流量标签数据集获取方法构建数据集。本文在使用 Netfilter 框架搭建 VPN 通信环境的基础上，设计了一种基于解密的 VPN 隧道流量标签数据集构建方法，通过获取 VPN 配置文件的相关信息来实现 VPN 隧道流量的解密，通过解密后的五元组进行 VPN 隧道流量分流和清洗，从而完成 VPN 隧道流量数据集的构建。

（2）面向 VPN 隧道流量的流量分割方法

由于 VPN 隧道流量无法按照传统的五元组方法进行分流且 VPN 在进行通信时往往给会产生随机长度的填充字段，这些字段会对全局流量特征构建产生干扰，VPN 隧道流量负载特性被掩盖、头部信息被修改，因此，要完成 VPN 隧道流量的业务识别任务，需要先对 VPN 隧道流量进行切割。本文在区分 IPSec VPN 隧道流量和 SSL VPN 隧道流量的基础上，提出了 VPN 隧道流量分割特征集合，使用基于队列的训练集构建方法建立标签特征训练集，使用基于滑动窗口的特征提取方法获取待分割 VPN 隧道流量的特征测试集。本章还设计了一种基于数据集改进的随机森林分割模型，该模型使用标签特征训练集构建 SP-RF 结构，输入特征测试集获得 VPN 隧道流量的切割点预测，最终依据预测切割点切分待分割 VPN 隧道流量。

（3）面向 VPN 隧道流量的业务识别方法

相比于非 VPN 加密流量，VPN 隧道流量对数据包负载增加封装，使得 VPN 隧道流量表现为端到端的单流特性，增加了 VPN 隧道流量全局特征的混乱程度，而机器学习方法难以构建取 VPN 隧道流量负载特征用于业务识别。本文在分析 IPSec VPN 加密流量和 SSL VPN 加密流量负载随机性的基础上，构建 IPSec VPN 隧道流量的流量统计特征，将其融入基于深度学习的 VPN 隧道流量框架，设计一种适用于 VPN 隧道流量经过切分后的碎片化 VPN 隧道流量的业务识别框架，该框架通过本文构建的 SSL VPN 识别序列和 IPSec VPN 识别序列可以完成 SSL VPN 隧道流量和 IPSec VPN 隧道流量的业务类型预测。

6.2 展望

由于一些限制因素的影响，本文针对 IPSec VPN 隧道流量和 SSL VPN 隧道流量的研究仍存在以下不足之处：

（1）位于 VPN 隧道内的流量因为 VPN 加密协议增加了数据包包长导致进入隧道内的流量产生了分片，隧道出入口流量与隧道内流量不具有一致性，从 VPN 隧道流量中剥离出不同业务行为的 VPN 加密流量虽然有一定难度，但是理论上是具有可行性的。本文 VPN 隧道流量数据集的构建依赖于从 VPN 配置文件中获取相关信息，现实情况下这点很难做到。因此，本文在后续研究中将针对 VPN 隧道内流量与 VPN 隧道出口流量关系进行探索。

（2）本文对于 VPN 隧道流量切分依据是 VPN 加密流量在不同业务流量中流量统计行为具有差异性，考虑现实情况应以数据包级特征为依据进行更细粒度 VPN 隧道流量分割。

（3）本文针对 VPN 隧道流量业务类型识别问题的研究是建立在自建 VPN 环境中，控制了 VPN 链路为理想化状态。现实情况是 VPN 链路上存在丢包问题，相当多企业组织通过一定的流量控制方法对 VPN 进行重新配置规划。本文研究局限于实验室环境，对于真实 VPN 通信流量难以做到百分之百还原，因此，本研究将进一步在真实环境下进行实验探索。

参考文献

[1] Atlas vpn. 2022-vpn-adoption-index[EB/OL]. https://atlasvpn.com/vpn-adoption-index.

[2] Zhang Y, Sun W, Zhang S. Identify VPN Traffic Under HTTPS Tunnel Using Three-Dimensional Sequence Features[C]//Proceedings of the 2022 11th International Conference on Networks, Communication and Computing. 2022: 18-23.

[3] 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.

[4] Zain ul Abideen M, Saleem S, Ejaz M. Vpn traffic detection in ssl-protected channel[J]. Security and Communication Networks, 2019, 2019: 1-17.

[5] Wang W, Zhu M, Wang J, et al. End-to-end encrypted traffic classification with one-dimensional convolution neural networks[C]//2017 IEEE international conference on intelligence and security informatics (ISI). IEEE, 2017: 43-48.

[6] Yao H, Liu C, Zhang P, et al. Identification of Encrypted Traffic Through Attention Mechanism Based Long Short Term Memory[J]. IEEE transactions on big data, 2022(1):8.

[7] 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.

[8] Zeng Y, Qi Z, Chen W, et al. TEST: an End-to-End Network Traffic Classification System With Spatio-Temporal Features Extraction[C]//2019 IEEE International Conference on Smart Cloud (SmartCloud). IEEE, 2019: 131-13.

[9] Liu C, He L, Xiong G, et al. Fs-net: A flow sequence network for encrypted traffic classification[C]//IEEE INFOCOM 2019-IEEE Conference On Computer Communications. IEEE, 2019: 1171-1179.

[10] Abdulazeez A, Salim B, Zeebaree D, et al. Comparison of VPN Protocols at Network Layer Focusing on Wire Guard Protocol[J]. 2020.

[11] Wang Y, Yu G, Shen W, et al. Deep learning based on byte sample entropy for VPN encrypted traffic identification[C]//2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE). IEEE, 2022: 293-296.

[12] Zhang Z, Chandel S, Sun J, et al. VPN: a Boon or Trap? : A Comparative Study of MPLS, IPSec, and SSL Virtual Private Networks[C]// 2018 Second International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2018: 510-515.

[13] Hai P N P, Hong H N, Quoc B B, et al. A Comparative Research on VPN Technologies on Operating System for Routers[C]//2021 International Conference on Advanced Technologies for Communications (ATC). IEEE, 2021: 89-93.

[14] Jahan S, Rahman M S, Saha S. Application specific tunneling protocol selection for Virtual Private Networks[C]//2017 international conference on networking, systems and security (nsyss). IEEE, 2017: 39-44.

[15] Chen H Y, Lin T N. The challenge of only one flow problem for traffic classification in identity obfuscation environments[J]. IEEE Access, 2021, 9: 84110-84121.

[16] Xu H, Li S, Cheng Z, et al. VT-GAT: A Novel VPN Encrypted Traffic Classification Model Based on Graph Attention Neural Network[C]//Collaborative Computing: Networking, Applications and Worksharing: 18th EAI International Conference, CollaborateCom 2022, Hangzhou, China, October 15-16, 2022, Proceedings, Part II. Cham: Springer Nature Switzerland, 2023: 437-456.

[17] 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(2):8.

[18] Luo P, Wang F, Chen S, et al. Behavior-Based Method for Real-Time Identification of Encrypted Proxy Traffic[C]//2021 13th International Conference on Communication Software and Networks (ICCSN). IEEE, 2021: 289-295.

[19] 周益旻, 刘方正, 王勇. 基于混合方法的 IPSec VPN 加密流量识别[J]. 计算机科学, 2021, 048(004):295-302.

[20] Shapira T, Shavitt Y. FlowPic: Encrypted Internet Traffic Classification is as Easy as Image Recognition[C]//IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). IEEE, 2019: 680-687.

[21] Zhou Y, Shi H, Zhao Y, et al. Encrypted network traffic identification based on 2d-cnn model[C]//2021 22nd Asia-Pacific Network Operations and Management Symposium (APNOMS). IEEE, 2021: 238-241.

[22] 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[J]. 网络空间安全, 2020, 11(8):6.

[23] Yamansavascilar B, Guvensan M A, Yavuz A G, et al. Application identification via network traffic classification[C]//2017 International Conference on Computing, Networking and Communications (ICNC). IEEE, 2017: 843-848.

[24] Nigmatullin R, Ivchenko A, Dorokhin S. Differentiation of sliding rescaled ranges: New approach to encrypted and VPN traffic detection[C]//2020 International Conference Engineering and Telecommunication (En&T). IEEE, 2020: 1-5.

[25] Khatouni A S, Zincir-Heywood N. Integrating machine learning with off-the-shelf traffic flow features for http/https traffic classification[C]//2019 IEEE Symposium on Computers and Communications (ISCC). IEEE, 2019: 1-7.

[26] Finamore A, Mellia M, Meo M, et al. Experiences of internet traffic monitoring with tstat[J]. IEEE Network, 2011, 25(3): 8-14.

[27] SiLK (System for Internet-Level Knowledge), July 2009, [online] Available: http://tools.netsa.cert.org/silk.

[28] Burschka S, Dupasquier B. Tranalyzer: Versatile high performance network traffic analyser[C]//2016 IEEE symposium series on computational intelligence (SSCI). IEEE, 2016: 1-8.

[29] Argus: the network audit record generation and utilization system, December 1994, [online] Available: https://qosient.com/argus/.

[30] Song M, Ran J, Li S. Encrypted Traffic Classification Based on Text Convolution Neural Networks[C]//2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT). IEEE, 2019: 432-436.

[31] 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.

[32] Feng R, Hu T, Jia X, et al. VPN Traffic Classification Based on CNN[C]//2022 14th International Conference on Computer Research and Development (ICCRD). IEEE, 2022: 94-99.

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

[34] Baek U J, Kim B, Park J T, et al. MISCNN: A Novel Learning Scheme for CNN-Based Network Traffic Classification[C]//2022 23rd Asia-Pacific Network Operations and Management Symposium (APNOMS). IEEE, 2022: 01-06.

[35] Chen Z, Cheng G, Jiang B, et al. Length matters: Fast internet encrypted traffic service classification based on multi-pdu lengths[C]//2020 16th International Conference on Mobility, Sensing and Networking (MSN). IEEE, 2020: 531-538.

[36] Huoh T L, Luo Y, Zhang T. Encrypted Network Traffic Classification Using a Geometric Learning Model[C]//2021 IFIP/IEEE International Symposium on Integrated Network Management (IM). IEEE, 2021: 376-383.

[37] Aceto G, Ciuonzo D, Montieri A, et al. DISTILLER: Encrypted traffic classification via multimodal multitask deep learning[J]. Journal of Network and Computer Applications, 2021, 183: 102985.

[38] 焦李成, 杨淑媛, 刘芳, 等. 神经网络七十年: 回顾与展望[J]. 计算机学报, 2016, 39(8): 1697-1716.

[39] Yan X, Gan X, Wang R, et al. Self-attention eidetic 3D-LSTM: Video prediction models for traffic flow forecasting[J]. Neurocomputing, 2022, 509: 167-176.

[40] Satya Sreedhar P S, Nandhagopal N. Classification Similarity Network Model for Image Fusion Using Resnet50 and GoogLeNet[J]. Intelligent Automation & Soft Computing, 2022, 31(3).

[41] Mnih V, Heess N, Graves A. Recurrent models of visual attention[J]. Advances in neural information processing systems, 2014, 27.

[42] Netfilter, 2023, [online] Available: https://www.netfilter.org/.

[43] Gazdag S L, Grundner-Culemann S, Heider T, et al. Quantum-Resistant MACsec and IPsec for Virtual Private Networks[C]//Security Standardisation Research: 8th International Conference, SSR 2023, Lyon, France, April 22-23, 2023, Proceedings. Cham: Springer Nature Switzerland, 2023: 1-21.

[44] Pang Y, Jin S, Li S, et al. Openvpn traffic identification using traffic fingerprints and statistical characteristics[C]//Trustworthy Computing and Services: International Conference, ISCTCS 2012, Beijing, China, May 28–June 2, 2012, Revised Selected Papers. Springer Berlin Heidelberg, 2013: 443-449.

[45] Xiong W, Lee C M. Efficient Scene Change Detection and Camera Motion Annotation for Video Classification[J]. Computer Vision and Image Understanding, 1998, 71(2):166-181.

[46] Ibrahim J, Gajin S. Entropy-based network traffic anomaly classification method resilient to deception[J]. Computer Science and Information Systems, 2022, 19(1): 87-116.

[47] Altaher A, Ramadass S, Almomani A. Real time network anomaly detection using relative entropy[C]//8th International Conference on High-capacity Optical Networks and Emerging Technologies. IEEE, 2011: 258-260.

[48] Breiman L. Random forests[J]. Machine learning, 2001, 45: 5-32.

[49] Hunt E B, Marin J, Stone P J. Experiments in induction.[J]. The American Journal of Psychology, 1966, 80(4).

[50] Fred S B, Bonald T, Prouti A, et al. Statistical bandwidth sharing[J]. ACM SIGCOMM Computer Communication Review, 2001.

毕业/学位论文答辩委员会名单

毕业/学位论文题目		面向多种类型 VPN 流量的识别技术研究		
作　者		张意飞		
专　业		网络空间安全		
研究方向		加密流量分析		
导　师		程光教授		
答辩委员会组成	姓　名	职　称	学科专业	工作单位
	季一木（主席）	教授	网络空间安全	南京邮电大学
	彭艳兵	教授级高工	网络空间安全	南京烽火星空通信发展有限公司
	张国敏	副教授	网络空间安全	中国人民解放军陆军工程大学
	胡晓艳	副教授	网络空间安全	东南大学
	吴桦（秘书）	副教授	网络空间安全	东南大学

备注：

1、本表格适用于所有研究生。

2、本表格排版在终版毕业/学位论文中，附在毕业/学位论文的最后。