信息内容安全技术国家工程实验室
National Engineering Laboratory for Information Security Technologies

隧道内混合网络行为分割与精细化识别技术研究

硕士研究生答辩报告

报 告 人：赵盼盼
导　　师：苟高鹏
指导老师：熊　刚

日　　期：2022年5月13日

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

目录

目录

研究背景与意义

隧道技术

- **隧道技术**是一种通过使用互联网络的基础设施在网络之间传递数据的方式[1]
- 隧道协议将这些其他协议的数据帧或包重新封装在新的包头中发送。新的包头提供了路由信息，从而使封装的负载数据能够通过互联网络传递。
- 被**封装**的数据包在隧道的两个端点之间通过公共互联网络进行路由。被封装的数据包在公共互联网络上传递时所经过的逻辑路径称为**隧道**。一旦到达网络终点，数据将被**解封装**并转发到最终目的地。
- 隧道技术是指包括**数据封装**、**传输**和**解封装**在内的全过程。

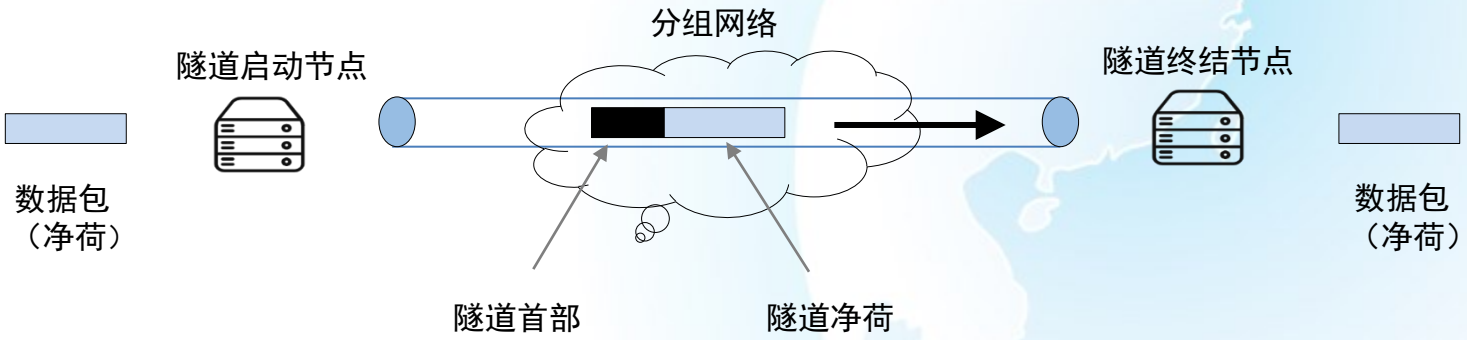
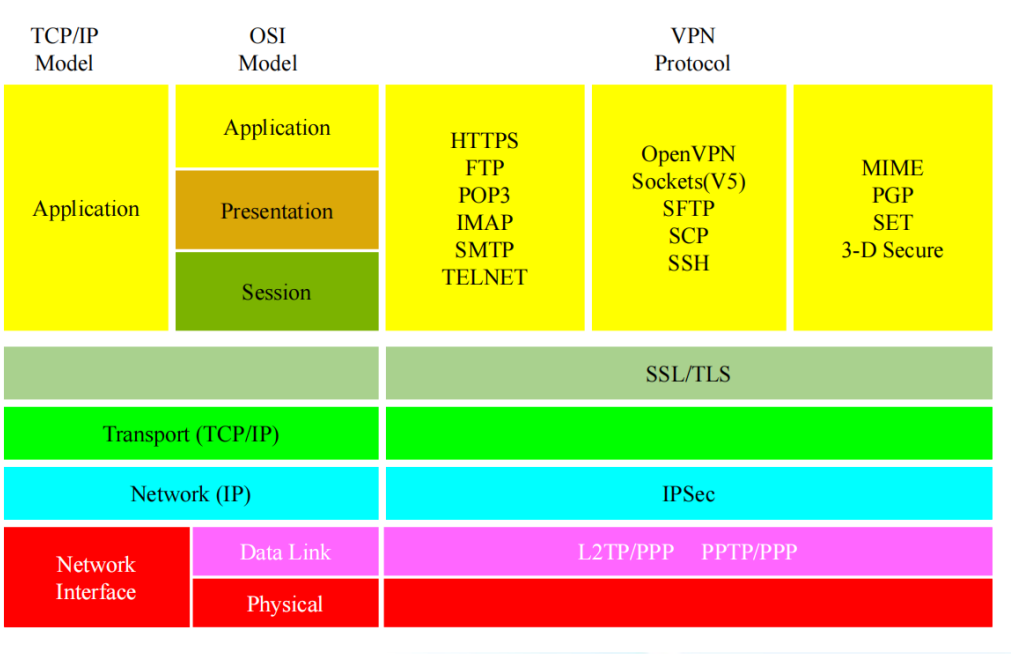


图1　隧道的基本组成

研究背景与意义

隧道协议

- **隧道**使用特定的协议进行封装和加密，这些协议按照处于TCP/IP的层级不同，可以分为：
  - 二层隧道协议（Layer 2 Tunneling Protocol，L2TP）、点对点隧道协议（Point to Point Tunneling Protocol，PPTP）它们对应于 TCP/IP 四层模型的网络接口层。
  - 互联网安全协议（Internet Protocol Security，IPSec）协议对应于TCP/IP 四层模型的网络层。
  - 安全套接字协议（Secure Sockets Layer，SSL）、安全外壳协议（Secure Shell，SSH）和 SOCKS v5（SOCKetS）应用层。



研究背景与意义

近些年来，隧道流量分析受到越来越多的关注，其原因包括：

隧道需求增多

2020年我国互联网网络安全态势综述里面提到：新冠肺炎疫情防控下的远程办公需求明显增多，**隧道技术**成为远程办公人员接入单位网络的主要技术手段之一[3]。

恶意流量增长

Gartner2020年报告中提到，隧道流量削弱了深度防御的效率，将端点和DMZ服务器暴露在出站和入站流量的威胁之下。**70%**的恶意服务通过**隧道**和**加密**技术绕过防火墙和入侵检测系统[4]

面临许多困难

不同隧道使用不同的隧道协议，流量相差较大。而且隧道流量经过了封装和加密，流量特征发生了大的变化。与加密流量相比，隧道流量的识别变得更加困难

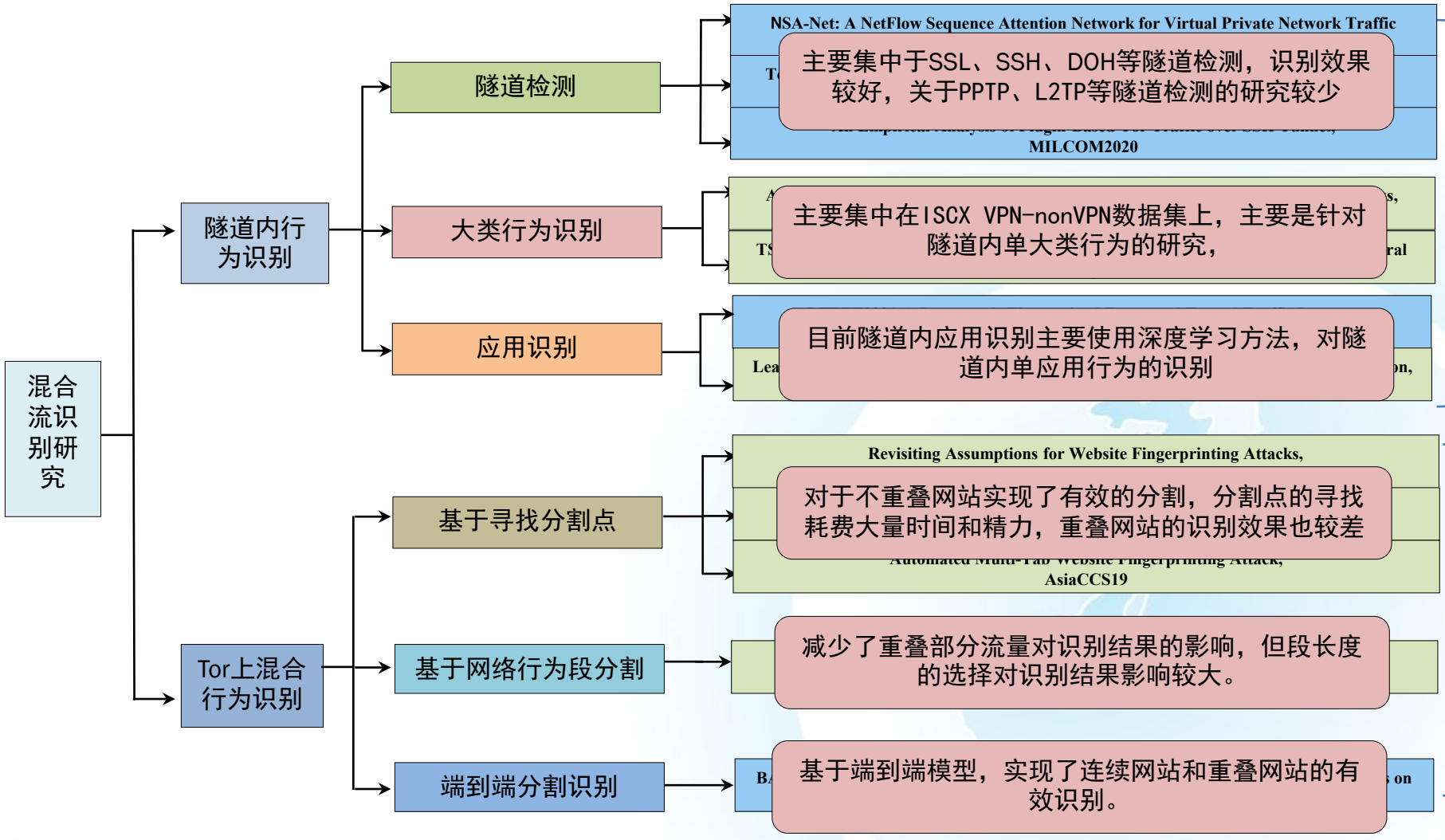
设备缺乏支持

传统网络审计设备**缺乏对隧道流量识别**的相关支持，存在很高的安全风险。使用隧道技术来绕过应用于网络安全设备的筛选器或签名。

目录

国内外研究现状分析

目前混合流识别主要集中于隧道和Tor上，因此本文从隧道内行为识别和Tor上行为识别进行分析。



国内外研究现状分析

隧道内网络行为识别研究现状总览

识别粒度	代表论文	数据集	特征	模型
隧道检测	cheng, et al(2019)	ISCX 2016	每条流的前N个数据包	编码器+注意力机制
隧道内大类行为识别				
				网络
				网络
隧道内应用识别				经网络
	Yao H,et al.(2019)	ISCX2016（单）	包长序列	分层注意力网络
	Zheng W,et al.(2020)	ISCX2016（单）	原始流序列	自编码器、元学习

1.隧道内网络行为识别主要集中在单网络行为识别，目前隧道内的数据集也主要是单行为数据集。

2,.关于隧道内混合流识别没有提出有效的识别方法和分割方法。

3.目前隧道内网络行为识别研究依赖于“单应用”假设，难以在实际中应用。

国内外研究现状分析

Tor上混合网络行为识别研究现状总览

技术	论文	数据集	方法
寻找分割点	Wang T,et al.(2016)	Tor 混合数据集(120 个网页，每个网页 40 个实例）	提出两级分割架构：基于时间进行首次分割，然后利用分割发现算法进行二次分割
	X		均衡问题，使用
			数据包预
网络行为段分割			
			度特征识
端到端识别	Guan Z,et al.(2021)	Tor 混合数据集（50 个网站，每个网站 90 个实例）	它从方向序列生成一个选项卡感知的表示，并执行块分割，以尽可能清晰地分离混合页面选项卡，从而缓解信息混淆

1.Tor上混合网络行为集中于混合网站的识别，三种方法。

2.分割点寻找方法花费大量时间，分割精度直接影响最终的识别结果。

3.网络行为段分割方法的段长度选择直接影响最终识别结果。

国内外研究现状分析

研究现状总览

隧道混合网络行为识别研究

缺少混合数据集

关于隧道内混合流量识别的研究较少，该研究领域缺少混合数据集。在隧道流量的研究中，目前公开的数据集主要是隧道内单行为的数据集，缺乏真实环境中隧道内混合行为流量的数据集。

缺少混合行为分割

目前没有提出隧道内混合行为流量有效的分割方法。目前关于混合流的研究主要集中在Tor上。关于如何从隧道混合流里面提取单一行为流量的研究开展较少。

缺少混合行为精细化识别

目前没有隧道内混合应用层面精细化识别研究。在隧道应用识别的研究中，一般都是依赖于“单应用”假设。对隧道内混合应用的识别研究较少.

目录

研究目标、研究内容和解决的关键问题

- **隧道内混合数据集构建和行为分割。**根据隧道用户在真实环境中的使用，构建隧道内混合行为数据集，并对两两应用混合流量进行分割。

数据集构建和行为分割

利用隧道回放、混合流生成的两种方法构建混合数据集。基于首尾段分割的方法实现混合应用流量的分割

已完成

提供混合多行为数据

- **隧道内混合行为的精细化识别。**对隧道多应用混合的网络流量进行分割，实现对隧道混合流量的精细化识别。

精细化识别

对隧道内多混合应用流量，采取基于网络行为转换检测、段分割和端到端整体识别的方法，实现混合应用流量的精细化识别。

已完成

提供混合流识别方法

- **隧道内行为精细化识别原型系统。**利用分割和精细化识别方法，实现隧道流量细粒度识别的原型系统。

原型系统

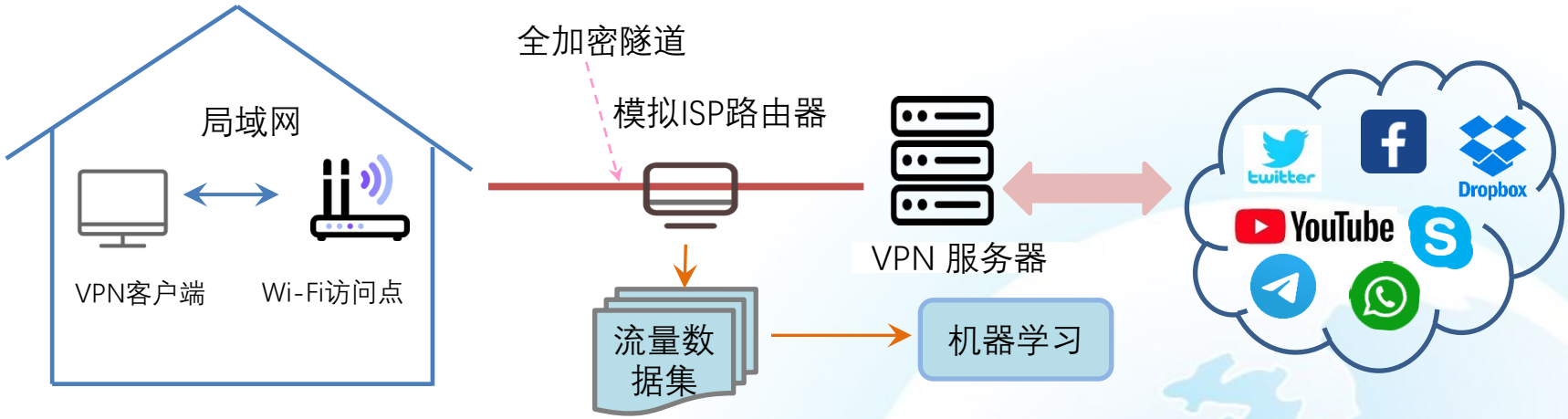
结合Burst分割、分割决策和以上分割识别技术，实现隧道内混合流量的精细化识别的原型系统

已完成

目录

主要创新/成果

◆ 研究内容一：隧道内混合数据集构建和行为分割

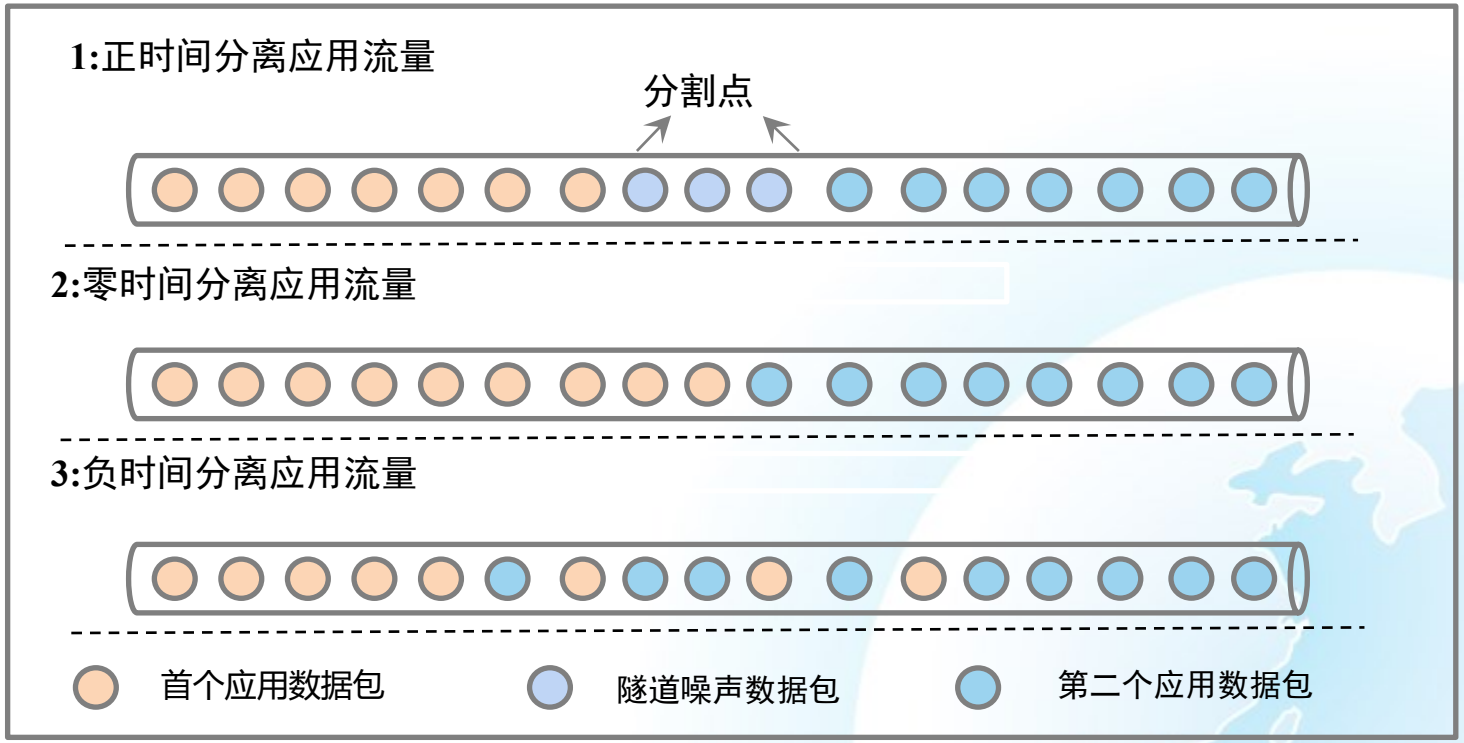


数据集产生场景：

- 单用户使用隧道的场景，使用的隧道包括IPsec、SSL等加密隧道。
- 用户使用隧道的过程中，各应用流量在时间维度上存在混合和重叠。
- 因用户使用行为的不同，应用流量之间的混合形式复杂多样

主要创新/成果

◆ 研究内容一：隧道内混合数据集构建和行为分割

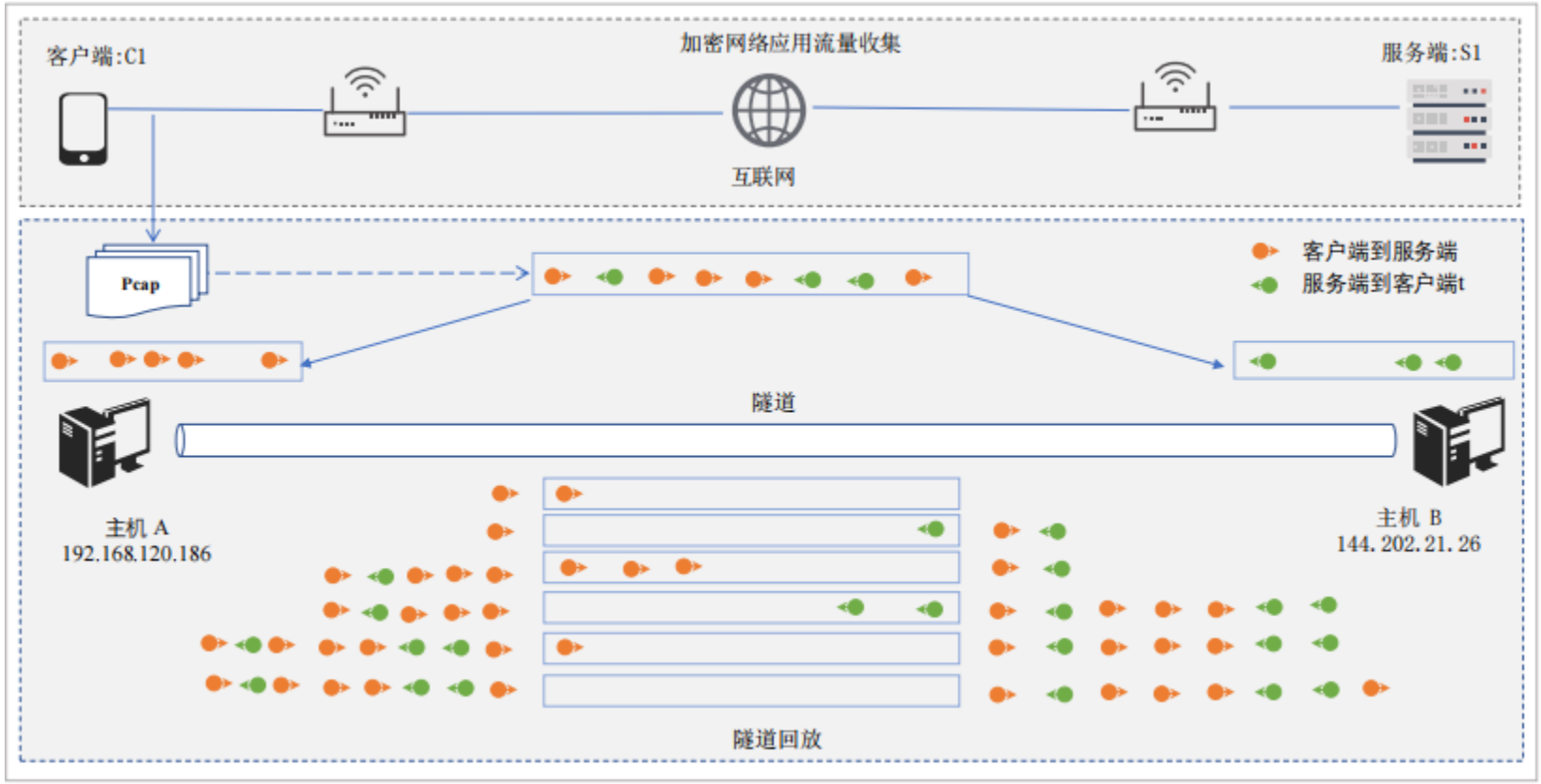


- 正时间分离应用流量：首个应用运行结束后，间隔一段时间，再运行下个应用。
- 零时间分离应用流量：首个应用运行结束后，无时间间隔，立即运行下一个应用。
- 负时间分离应用流量：首个应用未运行结束，就运行了下一个应用，存在交叉混合。

主要创新/成果

◆ 研究内容一：隧道内混合数据集构建和行为分割

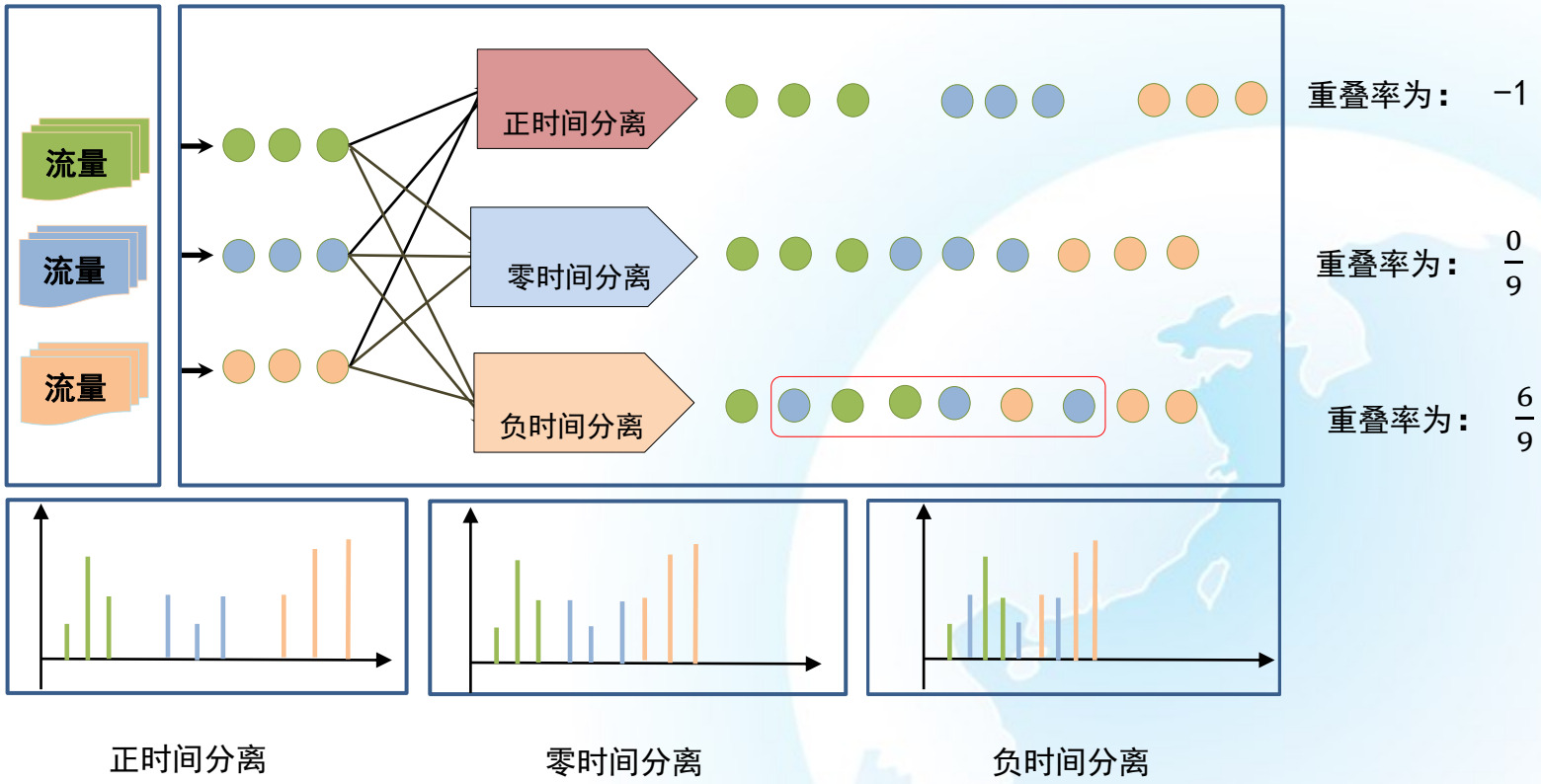
基于隧道回放的方法



主要创新/成果

◆ 研究内容一：隧道内混合数据集构建和行为分割

基于混合流生成的方法

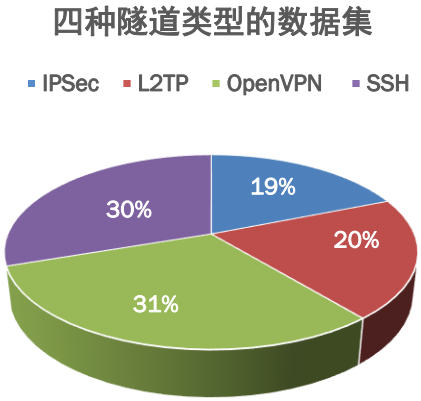


主要创新/成果

◆ 研究内容一：隧道内混合数据集构建和行为分割

□ 混合数据集构建——应用间的混合

自采集数据集：共采集了IPSec、SSL、L2TP、SSH等四种隧道下30个应用的流量，利用隧道回放和混合流生成算法，最终每种隧道下得到正时间分离、零时间分离和负时间分离三种类型的混合流量。负时间分离类型的数据集包括重叠率为 5%， 10%，15%， 20%， 25%， 30%， 35%，40%的8个数据集。



隧道层次	隧道名称	应用类别	流数	混合类型	数据集数目	大小
第二层	L2TP	30	4500	3	10	23.6GB
第三层	IPSec	30	4360	3	10	22.3GB
应用层/传输层	SSL	30	5100	3	10	36.3GB
	SSH	30	5700	3	10	35.6GB

主要创新/成果

◆ 研究内容一：隧道内混合数据集构建和行为分割

□ 混合数据集构建——应用间的混合

- **公开数据集：**USENIX2014、USENIX2017两个公开数据集。每个公开数据集包括三种类型混合数据集：正时间分离应用流量、零时间分离应用流量和负时间分离应用流量。其中负时间分离类型的数据集包括重叠率为 5%， 10%， 15%， 20%，25%， 30%， 35%，40%的8个数据集。

数据集来源	数据集名称	类型	网站类别	混合流数	数据集数目	混合类型
USENIX2014	knndata	Tor	60	18000	10	3
USENIX2017	walkiebatch	Tor	100	30000	10	3

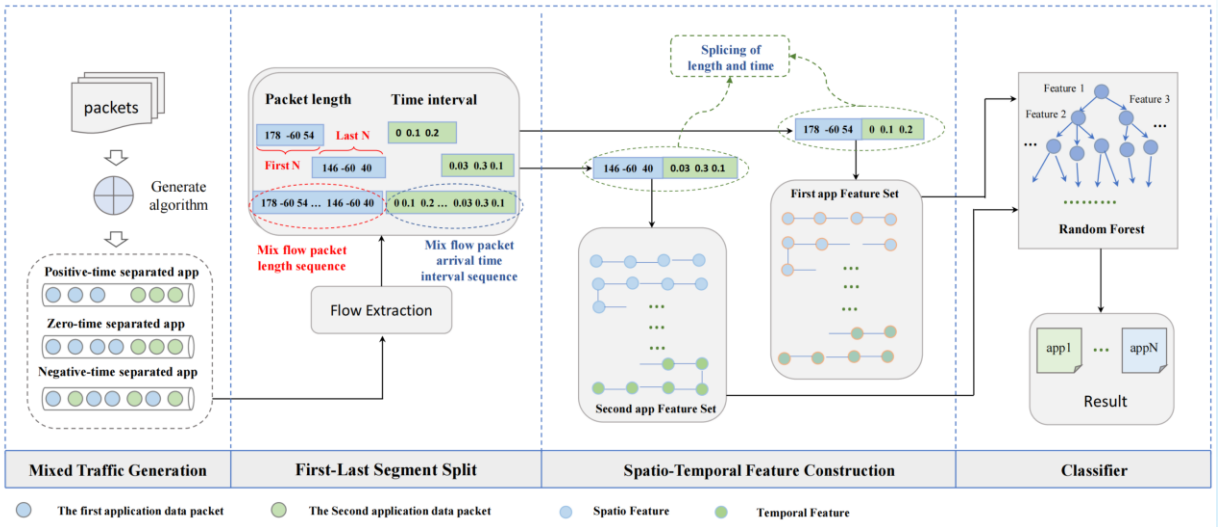
以上公开数据集主要是Tor下的网站流量，基于混合流生成算法,利用信息包括到达时间和包方向信息，我们按照时间维度，对公开数据集进行了混合，生成了Tor下的混合流量。

主要创新/成果

◆ 研究内容一：隧道内混合数据集构建和行为分割

□ 基于首尾段分割的方法

□ 基于首尾段分割思想，识别应用两两混合的流量



➢TunnelScanner框架主要包括四部分：混合流生成模块、首尾段分割、特征构建和分类器模块。

➢混合流生成模块：利用时间信息生成带标注的隧道混合流量。

➢首尾段分裂模块：通过首段分割和尾端分割的方法对隧道内混合流量进行分割。

➢时空特征构建：从首尾段中提取时间和空间的流量特征。

➢分类器模块：利用分类器算法，对提取特征后的流量进行分类。

主要创新/成果

◆ 研究内容一：隧道内混合数据集构建和行为分割

□ 基于分首尾段分割的方法

□ 特征选择

- 包方向、包长度和包时间间隔 124维候选统计特征
- 空间特征主要三个方向的统计特征，每个方向包括31维统计特征。时间特征包括31维的包时间间隔特征。
- 对候选特征进行特征重要性分析，平衡好准确率和训练时间两个指标，选择了前60维统计特征。
- 比较实验：与AppScanner、Cumul特征进行比较。

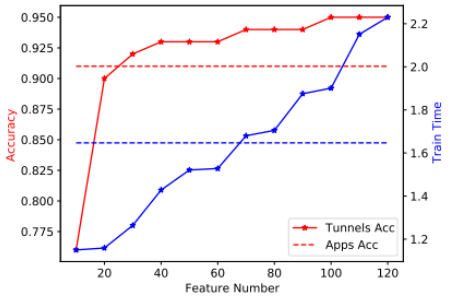
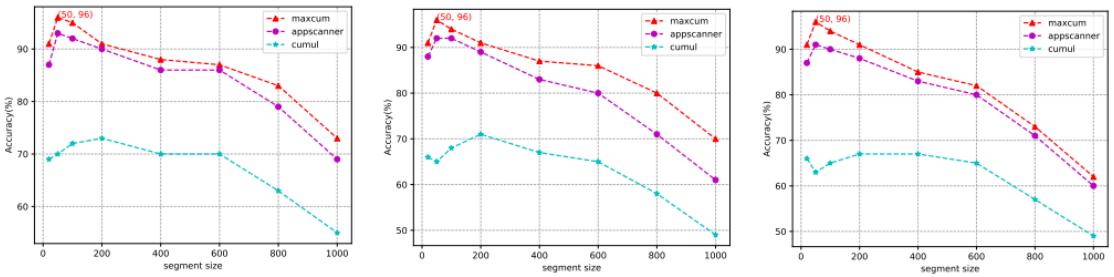


Fig. 3. Impact of the number of features on the classification results



(a) Positive time separation (b) Zero time separation (c) Negative time separation

Fig. 5. Impact of segment size on classification accuracy in ssl tunnel

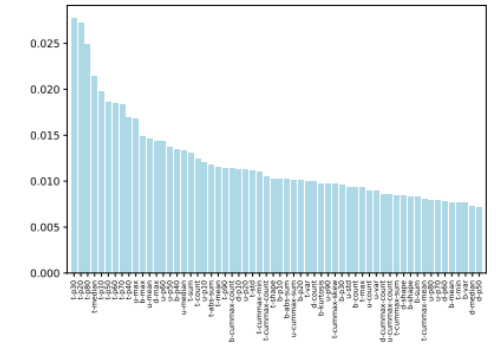


Fig. 4. Top 60 feature importance scores

主要创新/成果

◆ 研究内容一：隧道内混合数据集构建和行为分割

□ 基于首尾段分割的方法

□ 比较结果

(a) Positive t

- 利用60维统计特征，选择RF、SVM和KNN分类器进行流量识别。
- 分类器选择实验：结果表明随机森林分类器在三种混

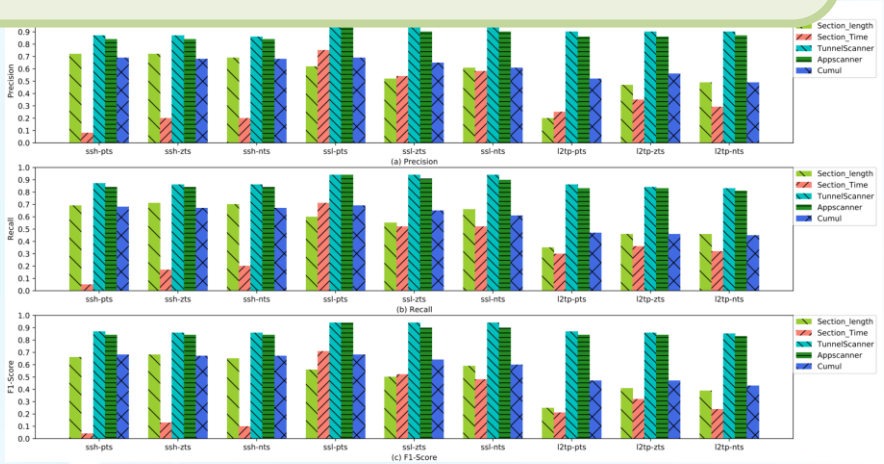
类

很

□ 相关成果：Zhao PP, et al. TunnelScanner: A Novel Approach For Tunnel Mixed Traffic Classification Using Machine Learning

HPCC 2021（CCF-C）录取

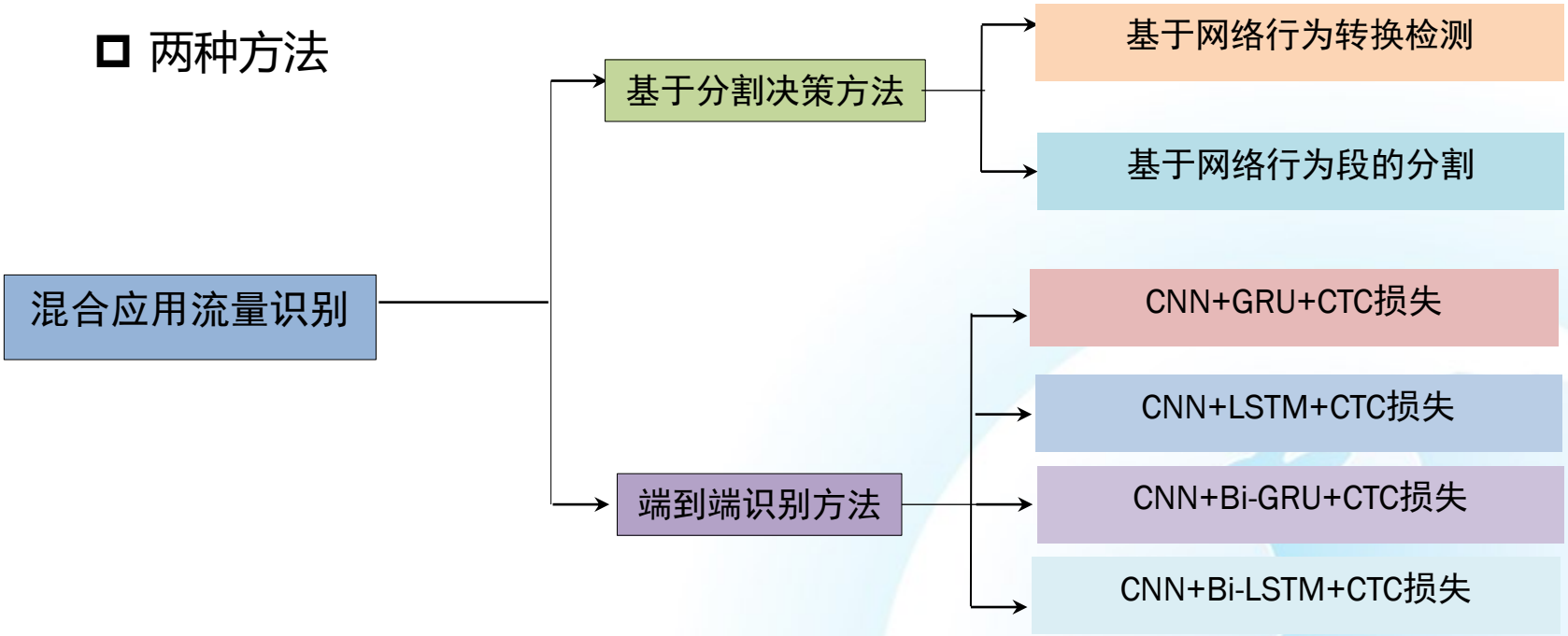
- **比较实验：**
  - 与现有的混合流识别方法相比，TunnelScanner方法在九种类型的混合数据集上表现出最好的识别结果。
- **实验结果：**
  - TunnelScanner方法的精度、召回和f-1值，均取得很高的识别效果。



主要创新/成果

◆ 研究内容二：隧道内混合网络行为的精细化识别

□ 两种方法

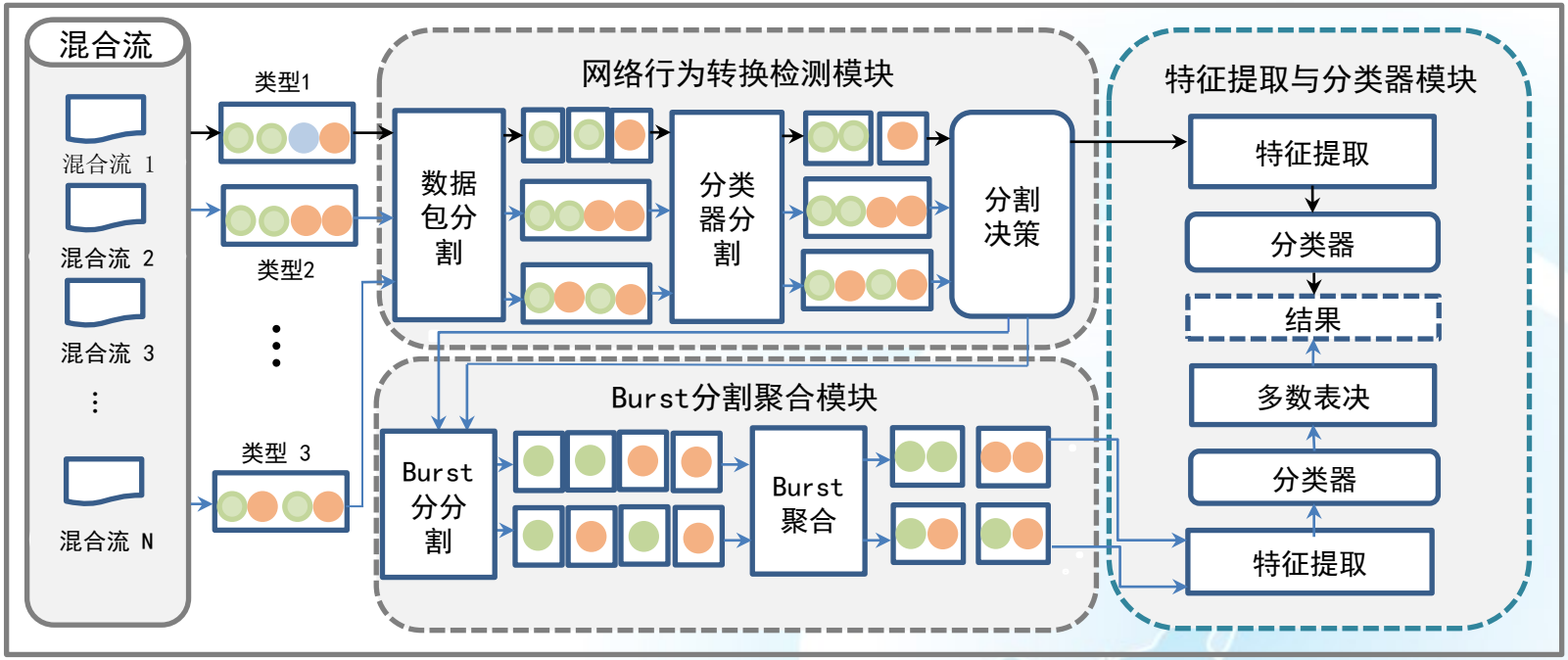


- 基于分割决策的方法：可以实现隧道内多应用混合的识别。对于正时间分离应用流量识别效果较好，但是对于重叠率较高的隧道流量识别效果较差。为解决该问题，我们提出了端到端的识别方法。
- 端到端识别方法：可以更好处理分割决策模块分割不成功或者分割后效果不好的混合数据集。

主要创新/成果

◆ 研究内容二：隧道内混合网络行为的精细化识别

□ 基于网络行为转换检测和段分割的方法-两级分割架构模型



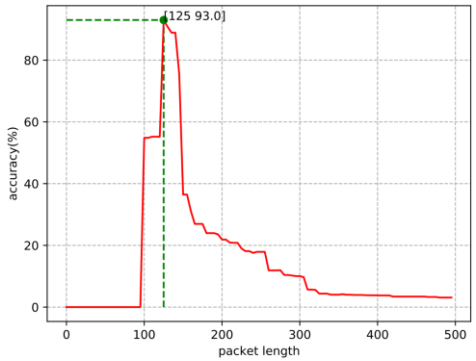
- **网络行为转换检测模块**：利用隧道噪声数据包信息，基于数据包和分类器相结合的方法，对隧道混合流量进行首次分割。分割决策部分，识别出单应用流量和多应用混合流量。
- **Burst分割聚合模块**：通过burst分割和聚合的分割方法对隧道混合流量进行分割。
- **特征提取和分类器模块**：通过随机森林算法和多数表决机制识别出隧道混合流量。

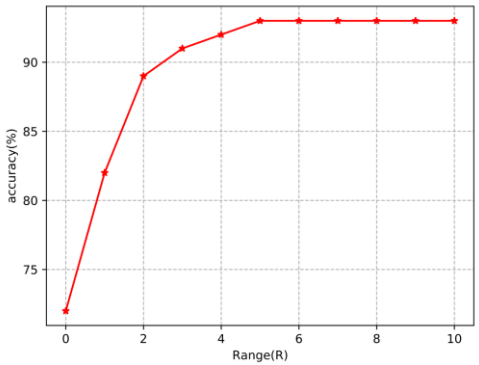
主要创新/成果

◆ 研究内容二：隧道内混合网络行为的精细化识别

◘ 基于网络行为转换检测和段分割的方法

◘ 分割结果



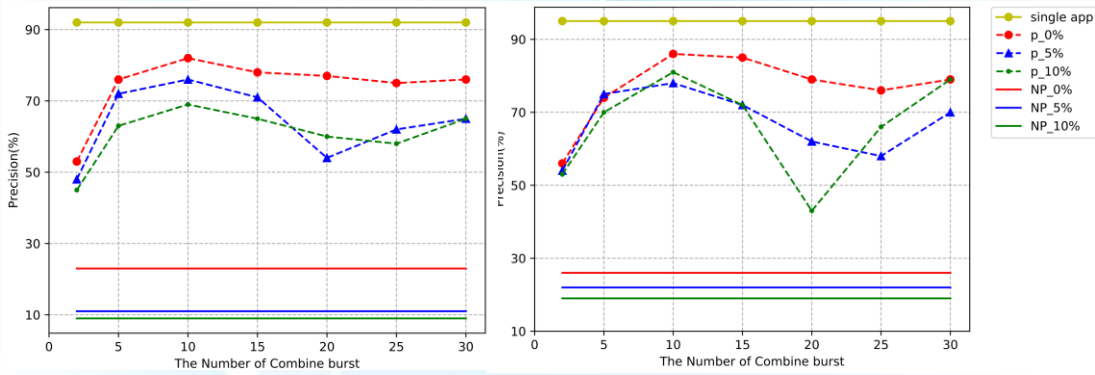


- 在不同数据包过滤阈值下，结合网络行为转换检测分类器，在包长过滤阈值在125取得最好的识别结果。
- 分割准确率：

$$diff(P,Q)=\begin{cases}1 & Q-R\le P\le Q+R\\ 0 & else\end{cases} \quad (1)$$

$$SA=\frac{1}{N}\sum_{i=1}^{N}diff(P_{predict}(i),P_{true}(i)) \quad (2)$$

- 与未经分割相比，经过TMT-RF框架处理后，隧道混合流量的识别提高了30%以上，更接近单应用假设下流量的识别结果。
- 不同段分割下，隧道混合流量的识别结果存在不同。当段大小选择10 时，取得最好的识别结果。

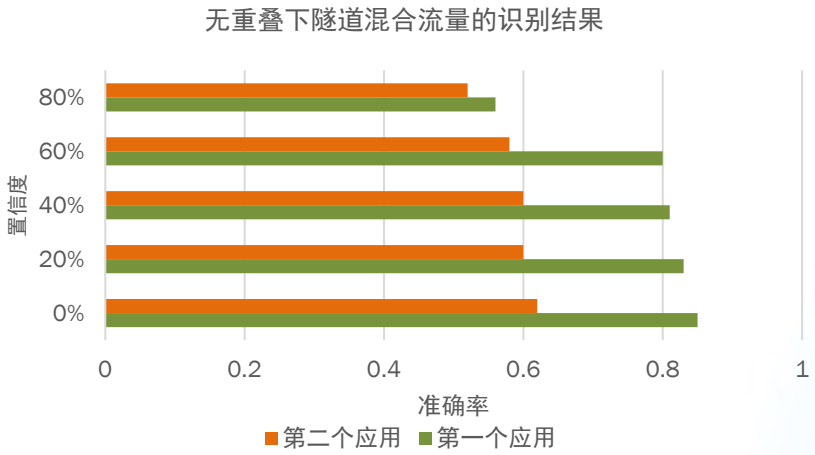


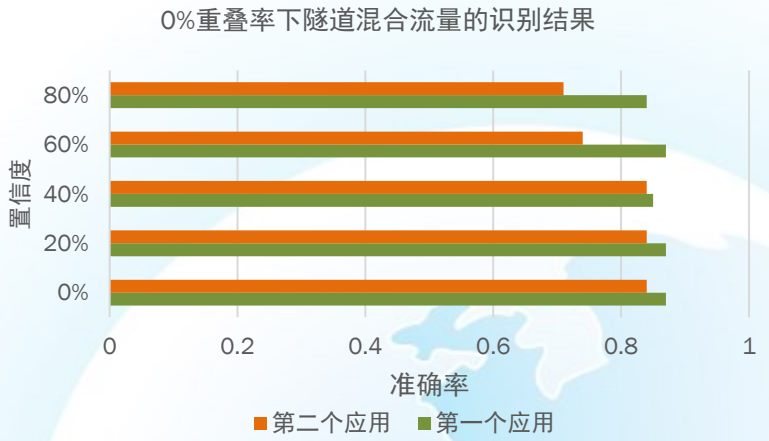
主要创新/成果

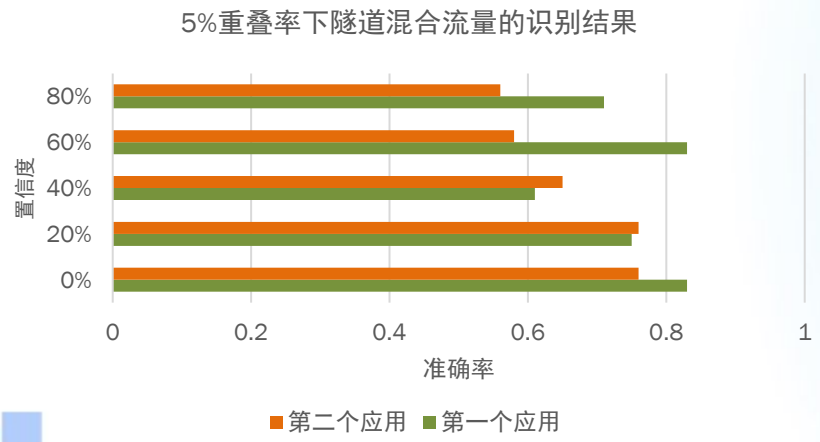
◆ 研究内容二：隧道内混合网络行为的精细化识别

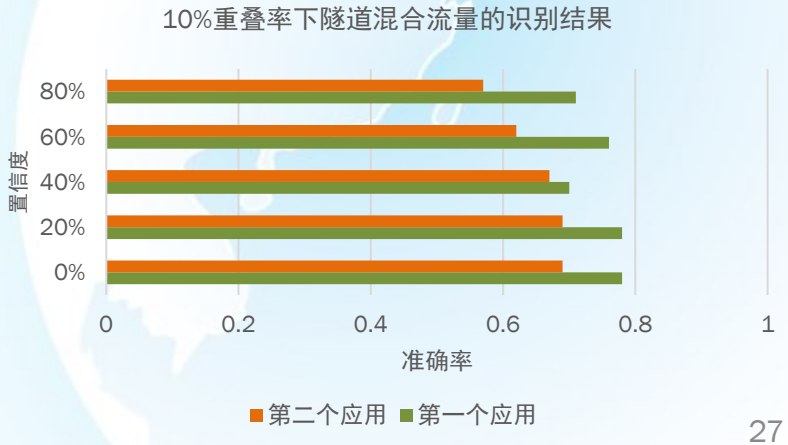
□ 基于网络行为转换检测和段分割的方法

□ 识别结果







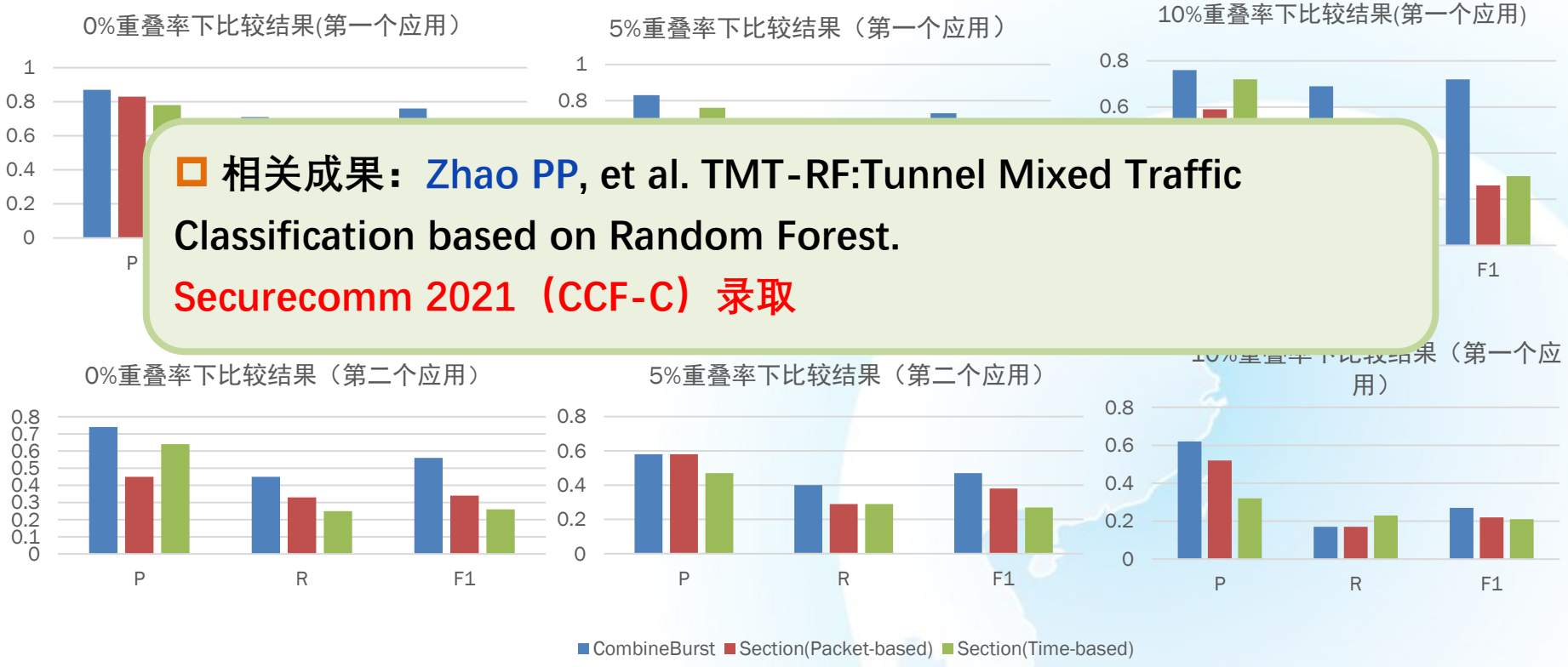


主要创新/成果

◆ 研究内容二：隧道内混合网络行为的精细化识别

□ 基于网络行为转换检测和段分割的方法

□ 比较实验结果



□ 相关成果：Zhao PP, et al. TMT-RF:Tunnel Mixed Traffic Classification based on Random Forest.

Securecomm 2021（CCF-C）录取

◆ 研究内容二：隧道内混合网络行为的精细化识别

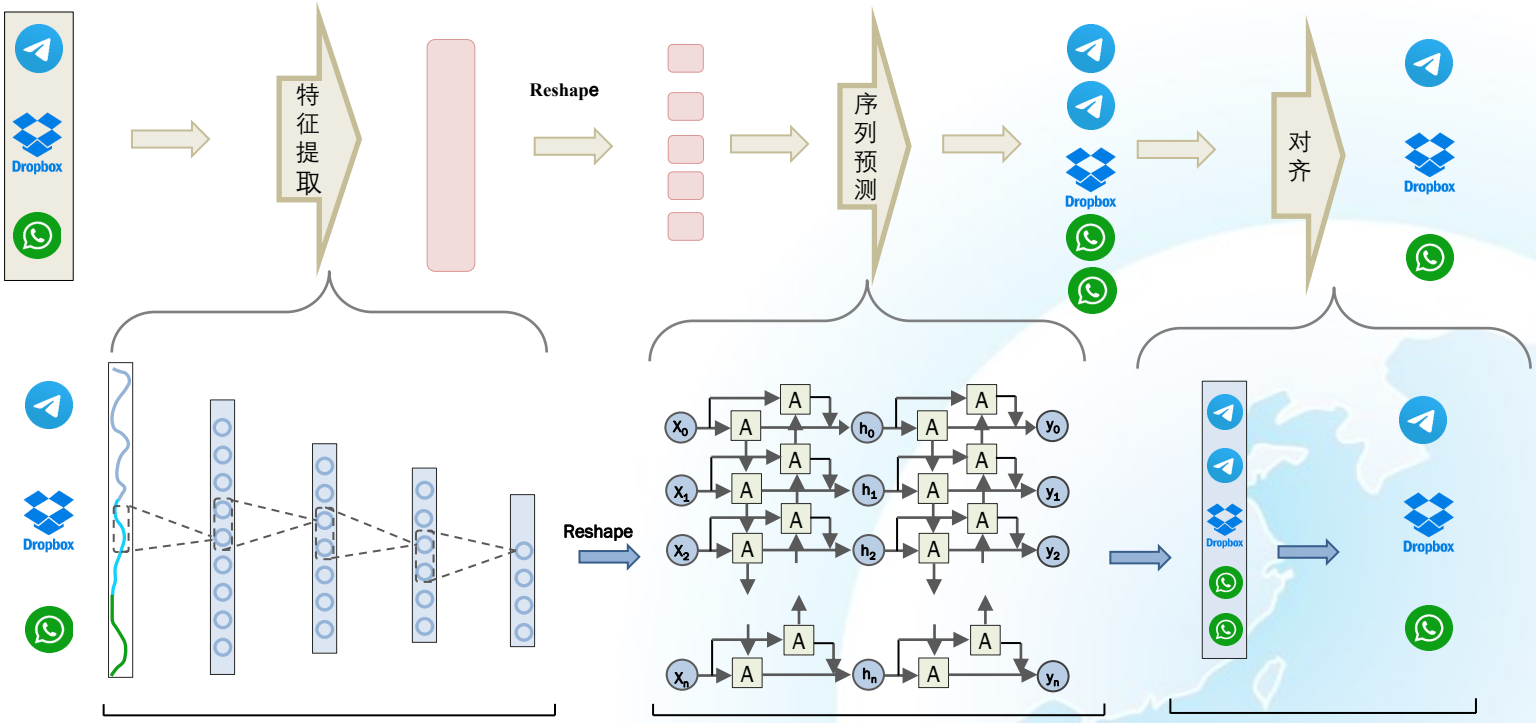
■ 总结

- 基于网络行为转换检测方法对于正时间分离应用流量识别效果较好。
- 基于网络行为段的方法在零时间和小重叠率的负时间分离应用流量上识别效果较好。
- 但是，随着重叠率的增加，识别结果会大幅度下降。
- 优点：
  - 方法简单，可解释性强。
  - 分割成功后，可以使用现有隧道识别模型仅识别。
- 缺点：
  - 寻找分割点会花费一定时间。
  - 混合流量的分割好坏，直接影响混合流量的识别结果。

主要创新/成果

◆ 研究内容二：隧道内混合网络行为的精细化识别

□ 基于CRNN模型的分割方法-将混合流分割与识别统一到一个模型中



特征提取模块：使用深度CNN，对输入隧道混合流量提取特征，捕获局部信息，得到特征序列；

序列预测模块：使用RNN对序列中的每个特征向量进行学习，并输出每个特征向量的预测标签；

对齐模块：使用CTC损失，把从循环层获取的一系列标签分布转换成最终的标签序列。

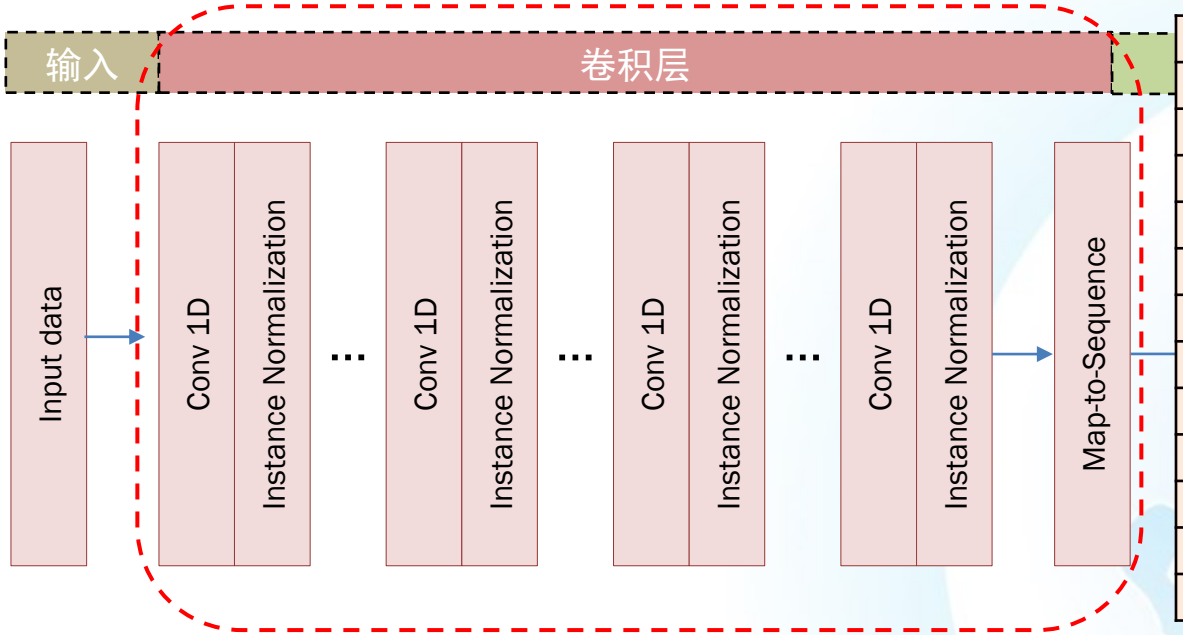
主要创新/成果

◆ **研究内容二：隧道内混合网络行为的精细化识别**

□ 基于CRNN模型的分割方法

特征提取模块

卷积层



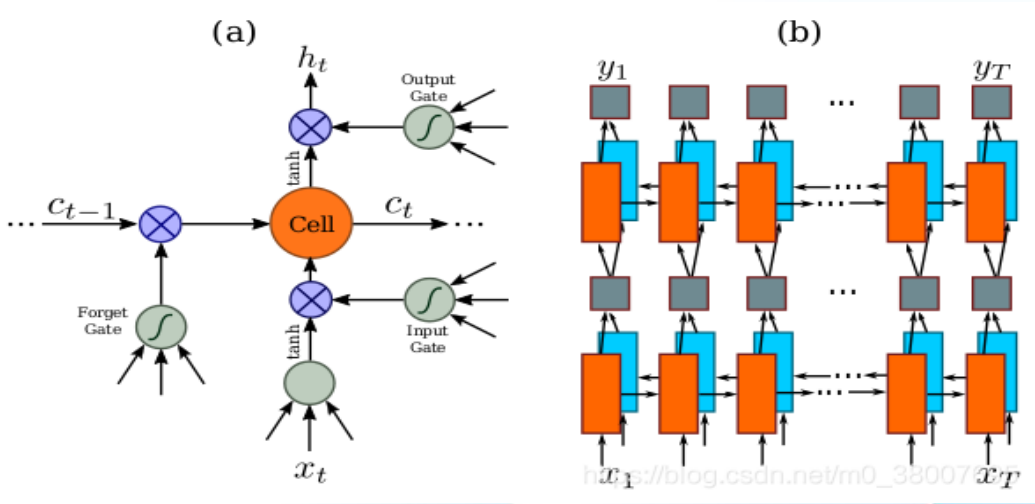
类型	配置
Transcription	-
Bidirectional-GRU	#hidden units: 128
Map-to-Sequence	-
Batch Normalization	32
Convolution	K:3, s:2, p:0
Batch Normalization	32
Convolution	K:3, s:1, p:0
Batch Normalization	32
Convolution	K:3, s:2, p:0
Batch Normalization	32
Convolution	K:3, s:1, p:0
Input	

主要创新/成果

◆ 研究内容二：隧道内混合网络行为的精细化识别

◘ 基于CRNN模型的分割方法

序列预测模块



- RNN 具有很强的捕获序列内上下文信息的能力。对于基于隧道混合流量的序列识别使用上下文提示比独立处理每个数据包更稳定且更有帮助。
- RNN 可以将误差差值反向传播到其输入（卷积层），从而允许在统一的网络中共同训练循环层和卷积层。RNN能够从头到尾对任意长度的序列进行操作。
- 选择GRU、LSTM、双向GRU和双向LSTM。

主要创新/成果

◆ 研究内容二：隧道内混合网络行为的精细化识别

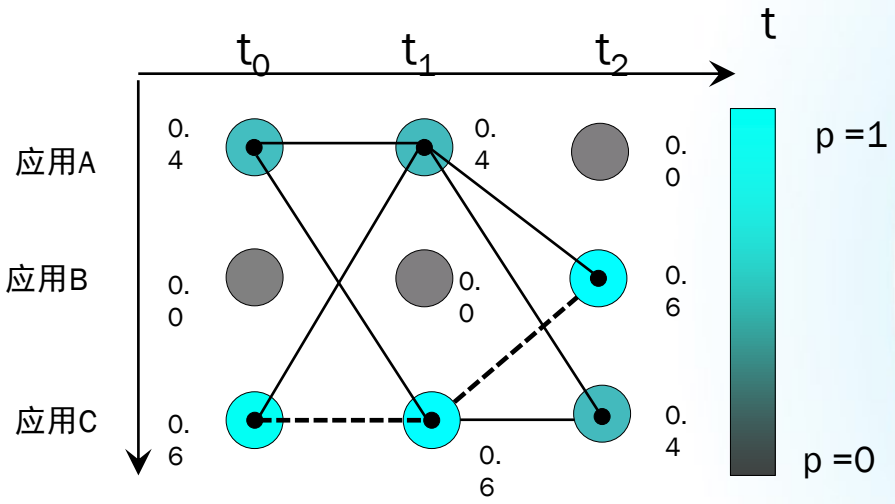
◘ 基于CRNN模型的分割方法

对齐模块：CTC 损失

➢ **CTC如何工作：**CTC损失函数接受神经网络的输出矩阵和相应的Ground-truth(GT)，尝试流量中GT的所有可能对齐，并对所有的得分进行求和。如果对齐分数求和值很高，则GT的分数就越高

$$LER(h,S') = \frac{1}{|S'|}\sum_{(x,z)\in S'}\frac{ED(h(x),z)}{|z|}$$

损失的计算：



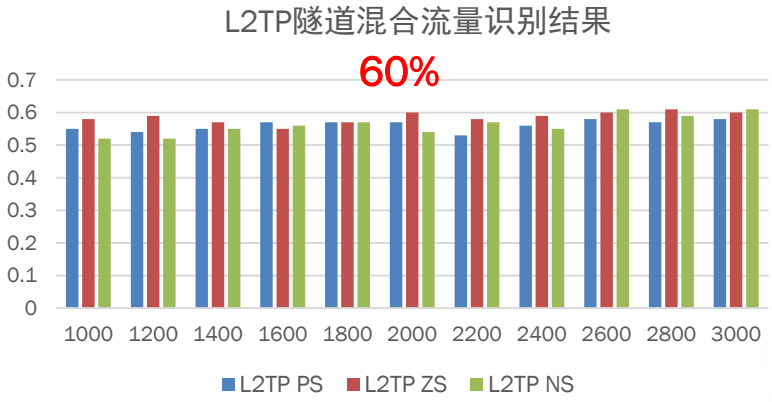
AAC= 0.4 * 0.4 *0.4 = 0.064
ACC = 0.6 * 0.6 * 0.4 = 0.144

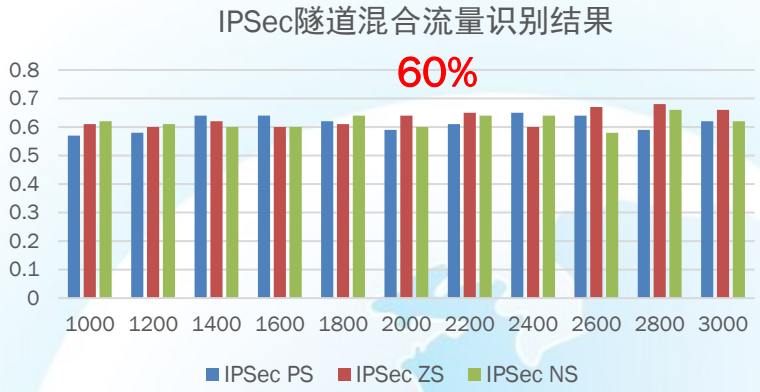
例如实例GT为AC，则必须计算长度为3 的所有可能路径 (因为矩阵有三个时间步长)，即：AAC,ACC。我们已经计算了这些路径的得分，对它们直接求和，得到：0.4 * 0.4 * 0.4 + 0.6 * 0.6 * 0.4 = 0.208。得到GT的概率，然后求概率的负对数得到损失。损失值通过神经网络反向传播，神经网络的参数根据所使用优化器进行更新。

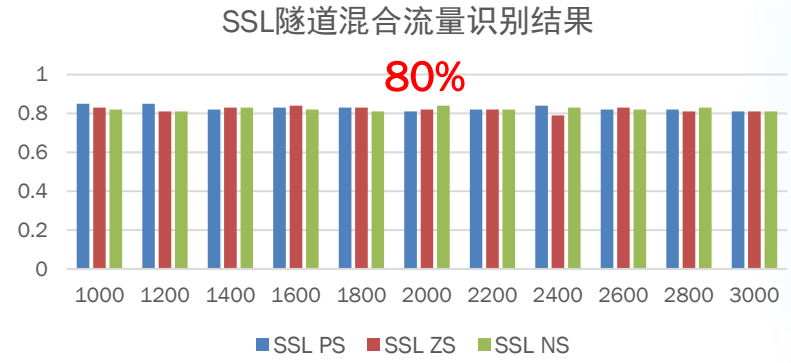
主要创新/成果

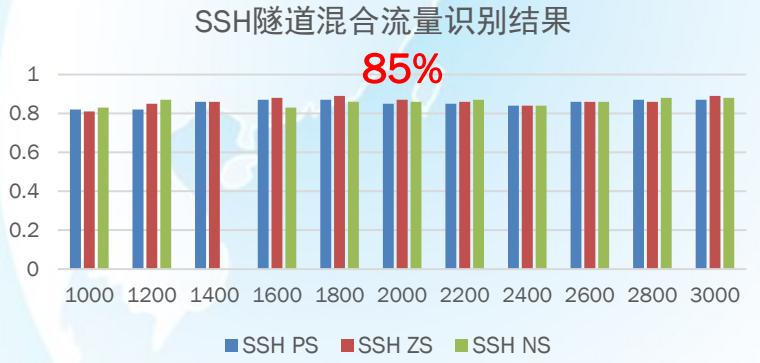
◆ 研究内容二：隧道内混合网络行为的精细化识别

- 基于CRNN模型的分割方法
- CNN+GRU模型分类效果







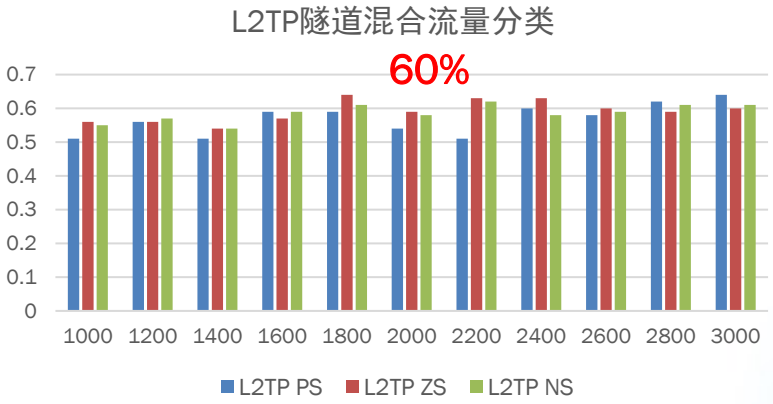


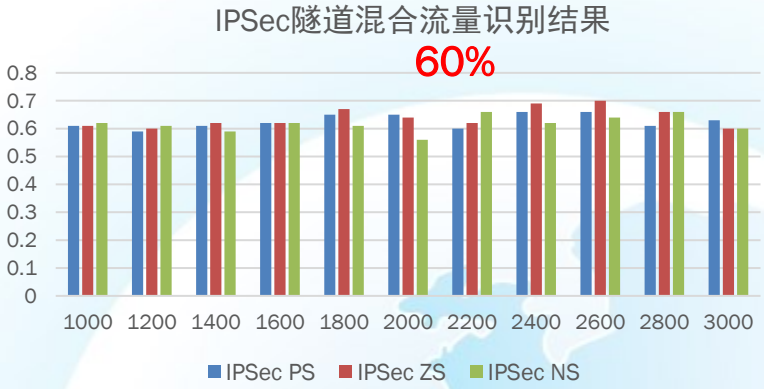
PS: Positive-time separated　ZS:Zero-time separated　NS: Negative-time separated (10%)

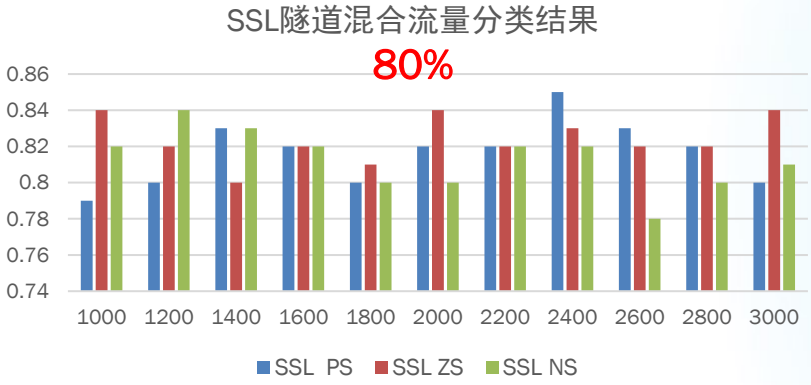
主要创新/成果

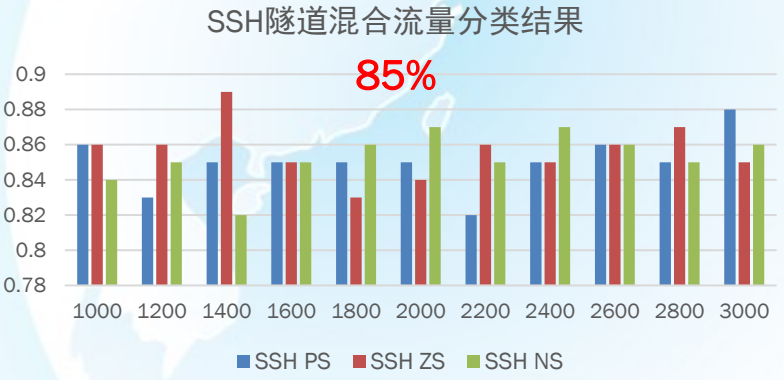
◆ 研究内容二：隧道内混合网络行为的精细化识别

- 基于CRNN模型的分割方法
- CNN+LSTM模型分类效果







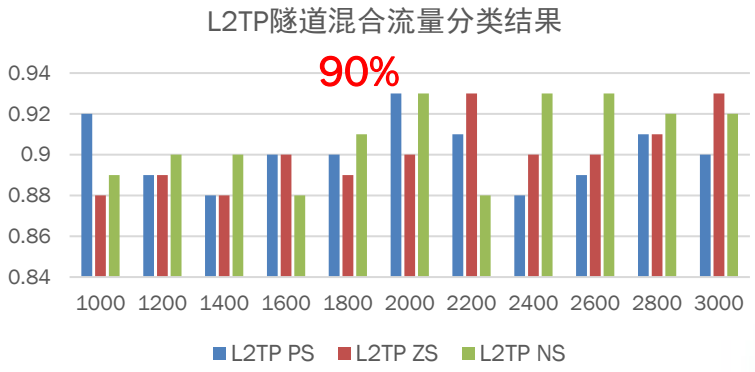


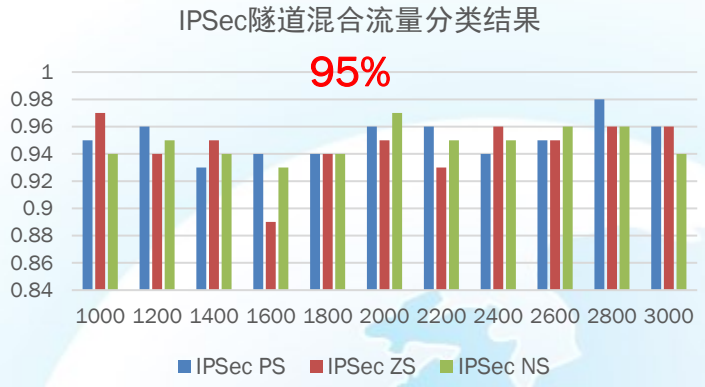
PS: Positive-time separated　ZS:Zero-time separated　NS: Negative-time separated (10%)

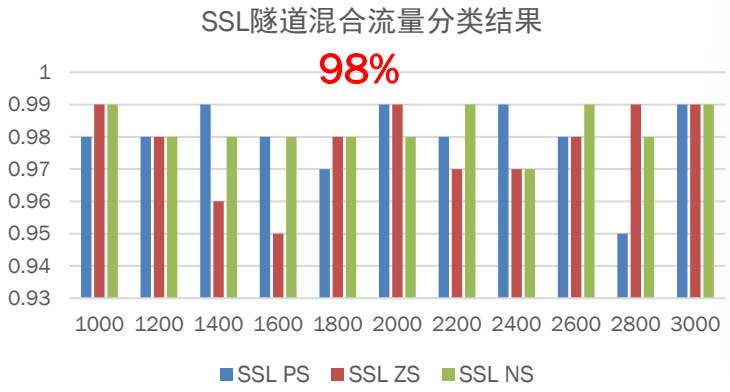
主要创新/成果

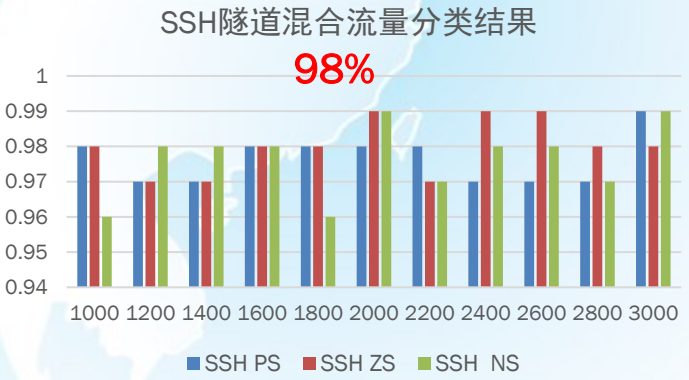
◆ 研究内容二：隧道内混合网络行为的精细化识别

- 基于CRNN模型的分割方法
- CNN+Bi-GRU模型分类效果







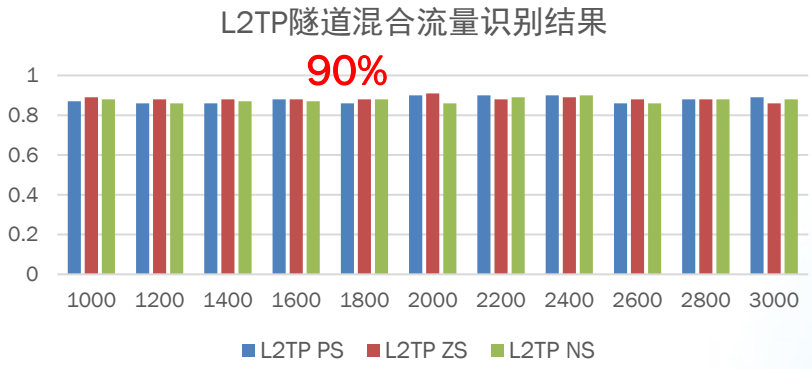


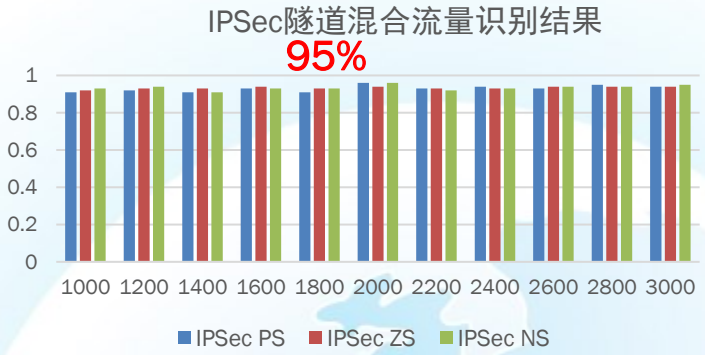
PS: Positive-time separated ZS:Zero-time separated NS: Negative-time separated (10%)

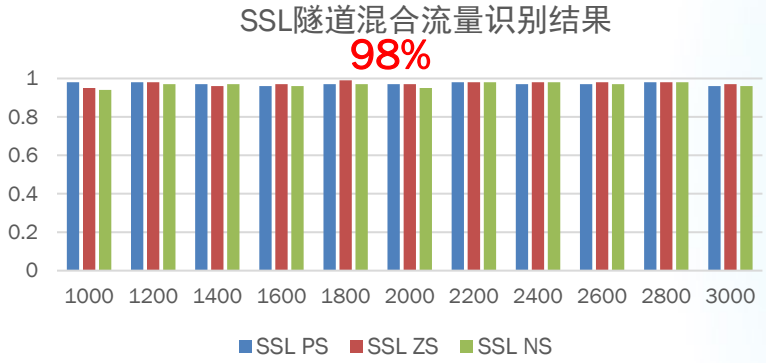
主要创新/成果

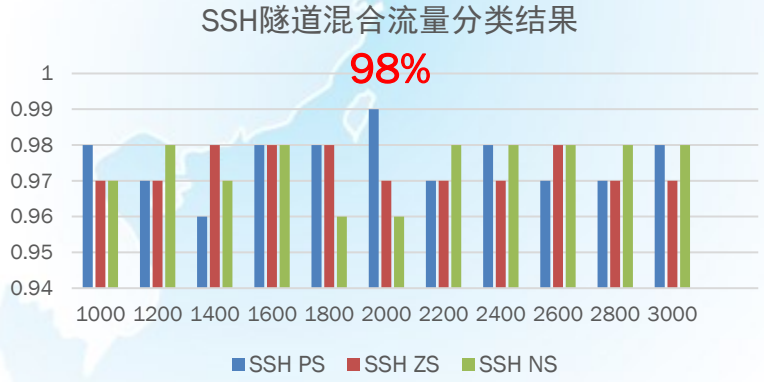
◆ 研究内容二：隧道内混合网络行为的精细化识别

- 基于CRNN模型的分割方法
- CNN+Bi-LSTM模型分类效果









PS: Positive-time separated　ZS:Zero-time separated　NS: Negative-time separated (10%)

主要创新/成果

◆ 研究内容二：隧道内混合网络行为的精细化识别

- 基于CRNN模型的分割方法
- CNN+Bi-GRU模型在公开数据集上的测试结果

序列长度	USENIX2014			USENIX2017		
	正时间分离	零时间分离	负时间分离（10%）[1]	正时间分离	零时间分离	负时间分离(10%)
1000	0.86	0.88	0.86	0.80	0.80	0.77
1200	0.88	0.88	0.86	0.80	0.80	0.77
1400	0.89	0.90	0.88	0.81	0.83	0.82
1600	0.92	0.91	0.90	0.83	0.83	0.82
1800	0.89	0.91	0.90	0.84	0.83	0.82
2000	0.90	0.90	0.91	0.81	0.82	0.84
2200	0.91	0.92	0.91	0.85	0.82	0.87
2400	0.90	0.91	0.92	0.84	0.86	0.86
2600	0.92	0.92	0.93	0.86	0.86	0.86
2800	0.93	0.93	0.94	0.86	0.82	0.85
3000	0.91	0.95	0.95	0.85	0.88	0.87

C-BiGRU模型在Tor-2014和Tor-2017公开数据集三种类型的混合流量上取得了很好的识别效果。

主要创新/成果

◆ **研究内容二：隧道内混合网络行为的精细化识别**

- 基于CRNN模型的分割方法
- CNN+Bi-GRU模型在不同重叠率下识别的结果
  - 序列长度2000

隧道类型\准确率	重叠率					
	15%	20%	25%	30%	35%	40%
SSH	0.97	0.97	0.96	0.94	0.91	0.89
SSL	0.97	0.95	0.94	0.91	0.91	0.88
IPSec	0.95	0.93	0.92	0.92	0.89	0.88
L2TP	0.91	0.90	0.90	0.88	0.87	0.87
Tor	0.91	0.91	0.89	0.88	0.88	0.86
Tor-2017	0.83	0.82	0.82	0.80	0.77	0.73

随着负时间分离应用重叠率的增加，C-BiGRU模型在四种隧道数据集上略有下降，但整体上识别率还是保持在87%以上，表现出很好的泛化能力。

主要创新/成果

总结

基于分割决策的隧道混合流识别方法

- ✓ 可解释性强
- ✓ 方法简单高效
- ✓ 现有模型可用

VS

基于CRNN的隧道混合流量识别方法

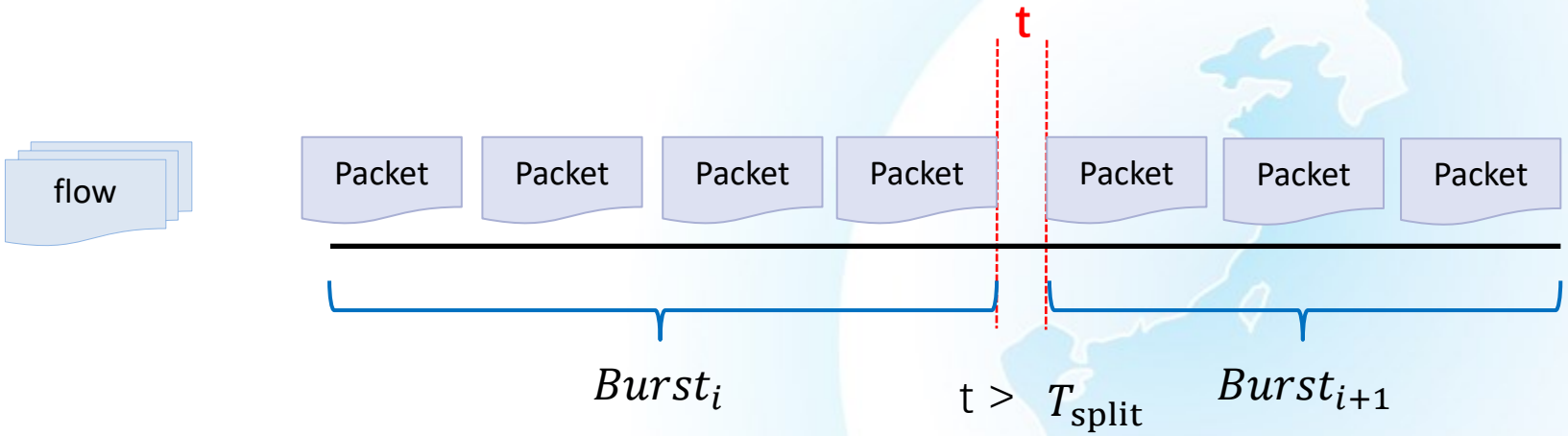
- ✓ 处理速度更快
- ✓ 识别精度更高
- ✓ 自动学习特征
- ✓ 模型泛化性强

结合两种隧道混合流识别方法，开发原型系统。结合不同类型隧道混合流量特点和两种方法的优点，对不同类型的隧道混合数据进行选择性处理。

主要创新/成果

◆ 研究内容三：隧道内混合网络行为精细化识别的原型系统

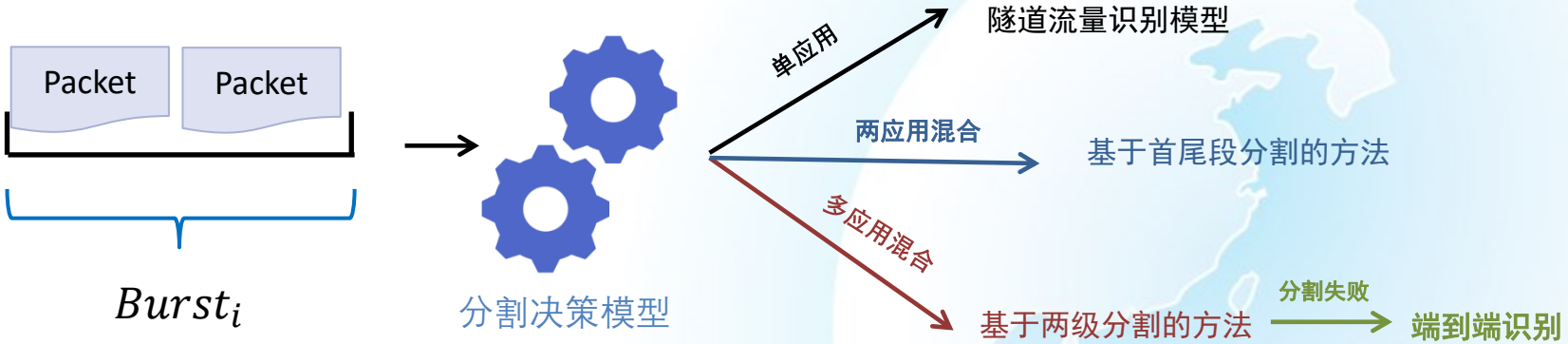
- 目标：实现实验室环境下隧道混合流量的精细化识别
  - **Burst切割**
    - 按照burst进行切割，如果两个数据包的时间间隔超过时间间隔阈值，则属于两个不同的Burst流量。



主要创新/成果

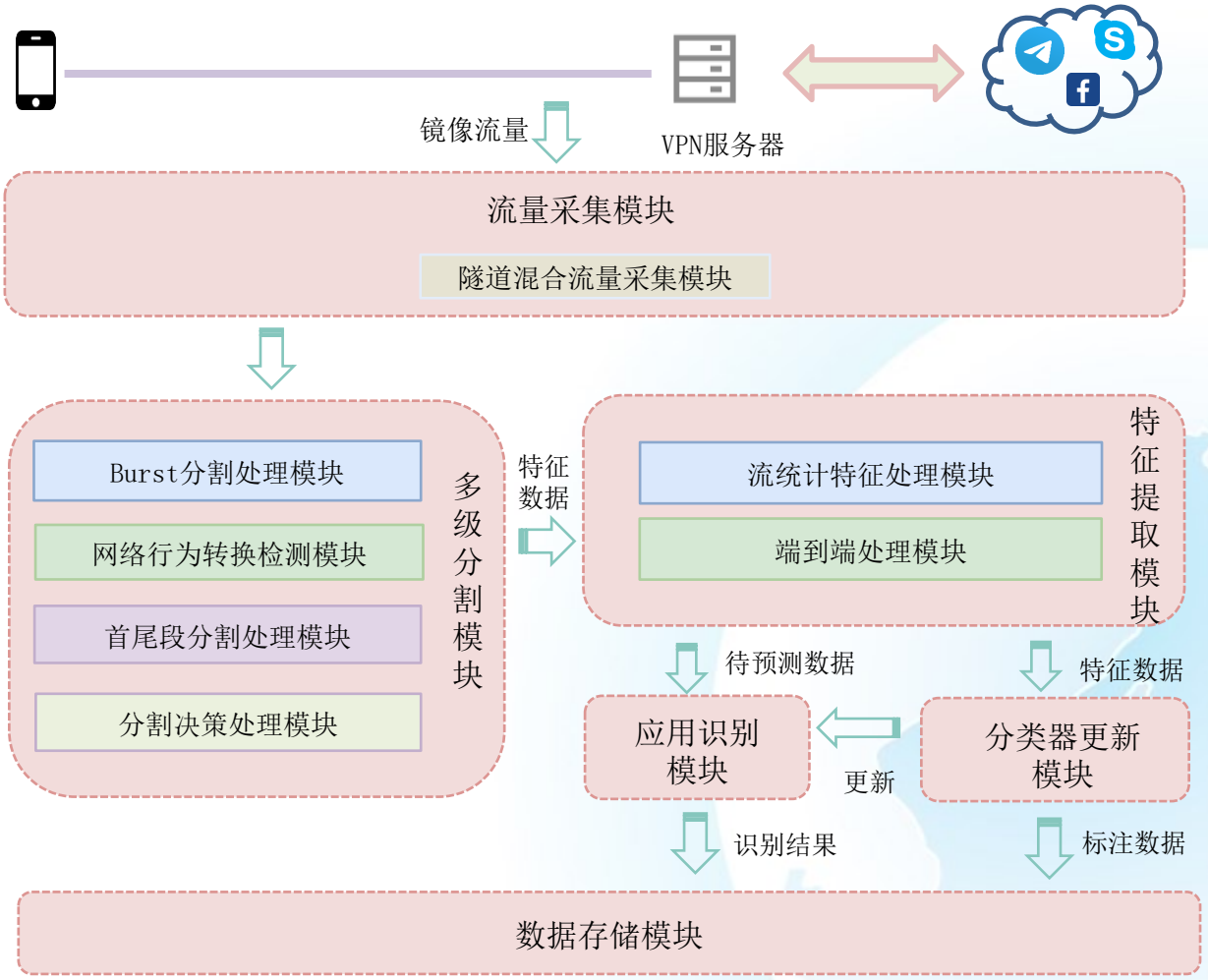
◆ 研究内容三：隧道内混合网络行为精细化识别的原型系统

- 目标：实现实验室环境下隧道混合流量的精细化识别
  - **分割决策**
    - Burst分割后的不同混合流量选择不同的处理流程。



主要创新/成果

◆ 研究内容三：隧道内混合网络行为精细化识别的原型系统



主要创新/成果

◆ 研究内容三：隧道内混合网络行为精细化识别的原型系统

□ 流量捕获

- □ 实验室环境下搭建IPSec隧道，捕获隧道流量。

图 5.8 真实环境隧道流量生成框架

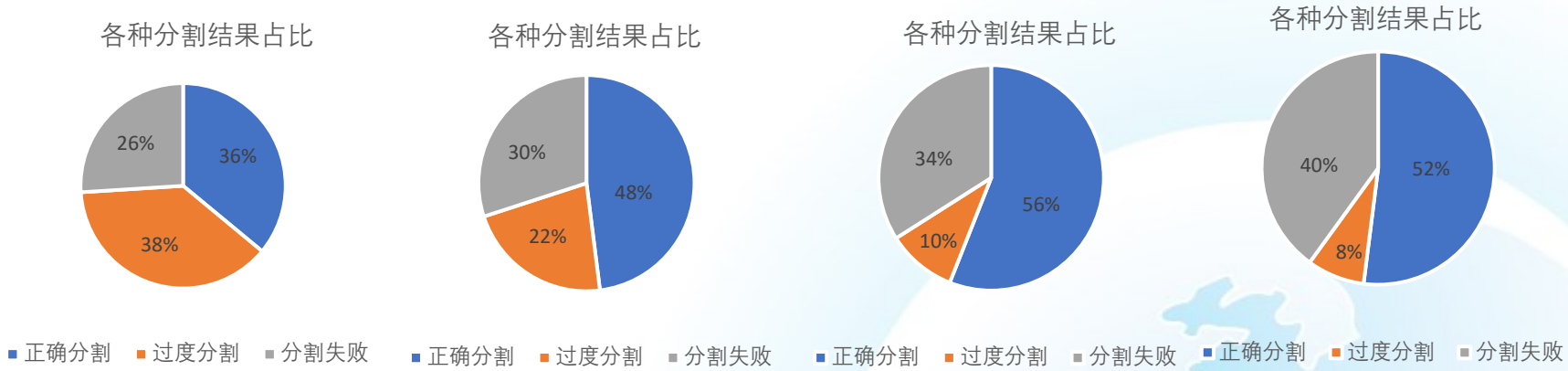
Dropbox	Twitter	Facebook
Youtube	Linkedin	Instagram
OneDrive	FileZilla	Skype
Netflix	Yahoomail	Servu
Foxmail	Michat	Gmail
Outlook	Icloudmil	Hulu

- 有时间间隔访问的流量：又称正时间分离应用流量，是指用户在访问应用的时候存在一定的空闲时间间隔，共收集18个应用，包括Facebook、Twitter等，共计20.1GB流量。
- 无时间间隔访问的流量：包括零时间分离应用流量和负时间分离应用流量，是指用户访问应用的时候没有存在空闲时间间隔，共收集18个应用，包括Skype、Youtube等，共计23.5GB。

主要创新/成果

◆ 研究内容三：隧道内混合网络行为精细化识别的原型系统

◘ Burst分割模块测试



(a) 0.1S　(b) 0.4S　(c) 0.7S　(d) 1S

- ✓ 随着Burst时间阈值的增加，过度分割的减少，正确分割的先增加后减小。
- ✓ 该模块应该尽可能保证正确分割的占比高一些。
- ✓ 最终，IPSec隧道的Burst分割时间阈值选择为0.7

主要创新/成果

◆ 研究内容三：隧道内混合网络行为精细化识别的原型系统

□ 网络行为转换检测模块实验评估

- ✓ 该模块可以处理网络行为之间时间间隔低于时间阈值，无法被 Burst 分割的混合流量。
- ✓ 数据包长度阈值选择 0-200，分类器选择的是随机森林
- ✓ 随着包长阈值的增加，识别的准确率先上升后下降。网络数据包长阈值在125时，识别率达到最高。

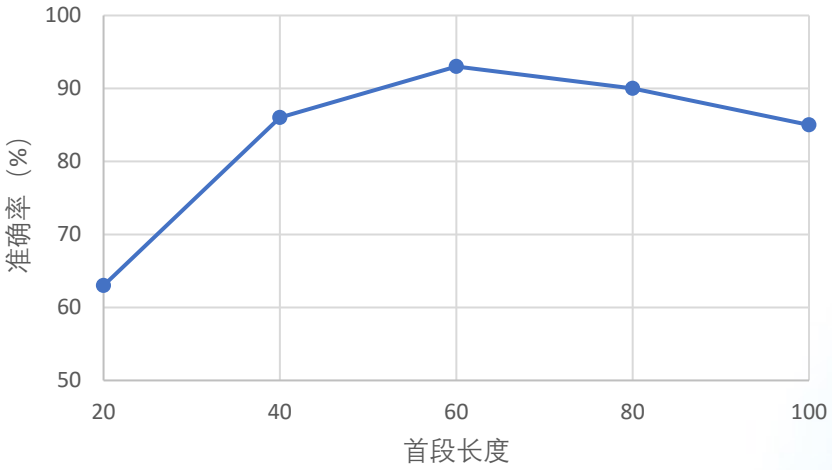
表5.1 不同分割阈值下的识别结果

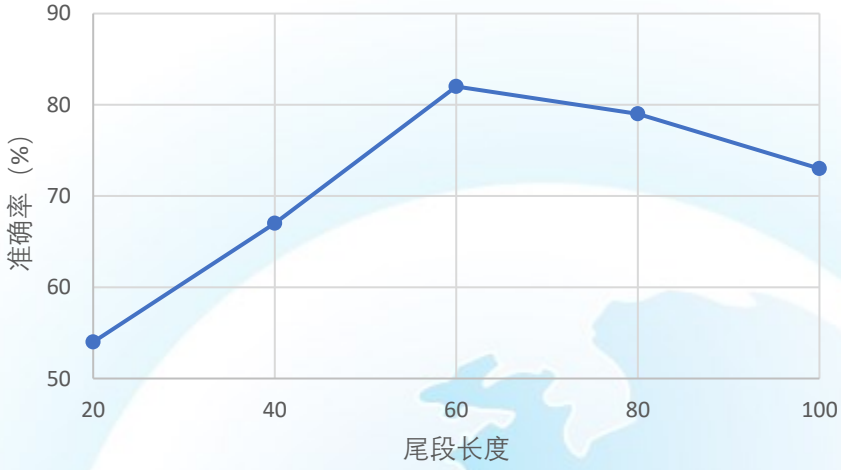
	0	25	50	75	100	125	150	175	200
ACC	0.35	0.35	0.49	0.57	0.84	0.91	0.76	0.53	0.50

主要创新/成果

◆ 研究内容三：隧道内混合网络行为精细化识别的原型系统

◘ 首尾段分割模块测试



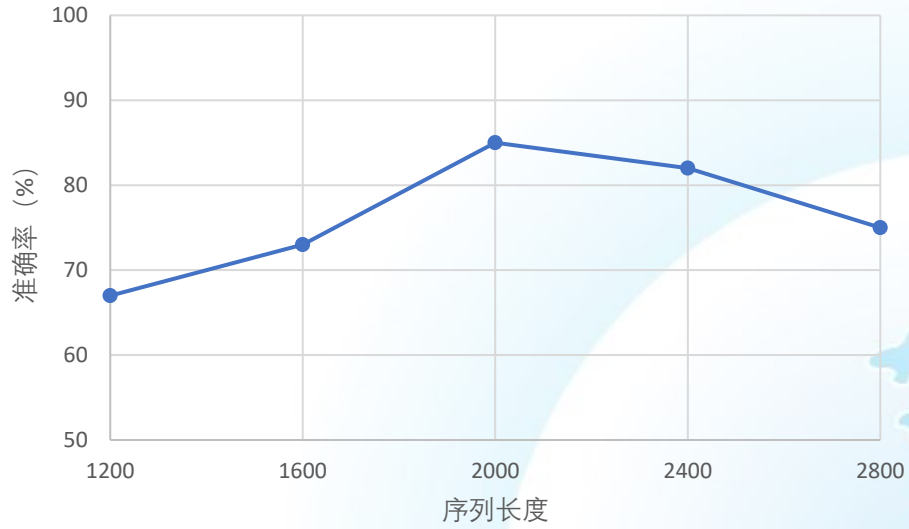


- ✓ 随着首段和尾段的增加，混合流量的识别结果先增加后下降，段选择 60 的时候，达到最高点。
- ✓ 段太小或者太大，识别结果都不好，主要原因在于段太小的话，包含的流量信息太少，难以准确的表征应用，提取有效的特征，导致识别准确率低。
- ✓ 如果段太大，每个段里面就可能会受到他网络行为流量干扰，导致流量不纯净，识别结果不准确

主要创新/成果

◆ 研究内容三：隧道内混合网络行为精细化识别的原型系统

□ 端到端识别模块实验评估



- ✓ 与正时间分离应用流量相比，该类型的隧道混合流量识别结果略微下降。
- ✓ 主要在于该模块处理的混合流量形式多样，存在的正时间分离应用流量是前面两级没有成功分割的混合流量，分割更难。
- ✓ 而且负时间分离应用流量存在不同重叠率，差别较大。重叠率越高，隧道混合流量可以利用的纯净流量越少，使得隧道混合流量的识别更难。

主要创新/成果

◆ 研究内容三：隧道内混合网络行为精细化识别的原型系统

□ 分割决策模块实验评估

5.2 分割决策识别结果

混合类型	精确率	召回率	F1
单应用纯净流量	0.93	0.97	0.95
两应用混合流量	0.85	0.97	0.91
多应用混合流量	0.70	0.84	0.76

- ✓ 单应用纯净流量识别效果最好，原因主要在于纯净流量没有受到其他流量的干扰，流量更加纯净，特征更加明显。
- ✓ 两应用、多应用混合流量与单应用混合流量相比，识别结果略有下降，主要在于流量混合后，数据包长序列和数据包时间间隔发生了大的变化，而且不同重叠下流量差别也较大，因此识别结果略有下降。

主要创新/成果

◆ **研究内容三：隧道内混合网络行为精细化识别的原型系统**

□ 应用识别结果

IPSec							
应用	精度	召回	F1	应用	精度	召回	F1
dropbox	0.76	0.85	0.80	filezilla	0.80	0.89	0.84
twitter	0.74	0.69	0.71	skype	0.83	0.92	0.87
facebook	0.81	0.75	0.78	gmail	0.90	0.78	0.84
outlook	0.83	0.88	0.85	netflix	0.79	0.83	0.81
youtube	0.87	0.90	0.88	yahoomail	0.77	0.87	0.82
linkedin	0.82	0.89	0.85	servu	0.81	0.89	0.85
instagram	0.61	0.72	0.66	hulu	0.84	0.86	0.85
icloudmail	0.90	0.87	0.88	foxmail	0.85	0.80	0.82
onedrive	0.85	0.89	0.87	michat	0.79	0.85	0.82

目录

完成项目及发表论文情况

已参与项目：

- 基于wireshark的网络协议深度解析系统设计与开发
- 抓包测试和隧道回放工作
- 隧道内混合网络行为识别

已发表论文：

- 名称：TMT-RF: Tunnel Mixed Traffic Classification based on Random Forest.
- 会议：SecureComm 2021
- 作者：赵盼盼、苟高鹏、刘畅、管洋洋、崔明鑫、熊刚

- 名称：TunnelScanner:A Novel Approach For Tunnel Mixed Traffic Classification Using Machine Learning
- 会议：HPCC 2021
- 作者：赵盼盼、李镇、崔明鑫、陆杰、熊刚、苟高鹏

完成项目及发表论文情况

申请专利：

- 名称：一种基于随机森林的隧道混合流量分类方法及系统

研究生期间所获荣誉：

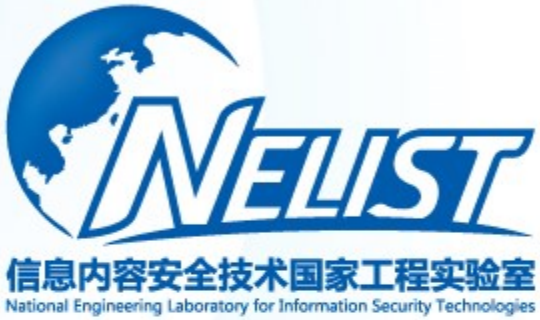
- 研究生国家奖学金（教育部）
- 北京市志愿服务证书（北京市怀柔区团委）
- 中国科学院大学三好学生（中国科学院大学）
- 中国科学院大学优秀学生干部（中国科学院大学）
- 中国科学院大学优秀共青团干部 （中国科学院大学）
- 中国科学院大学大学生奖学金 （中国科学院大学）
- 中国科学院大学学业奖学金 （中国科学院大学）

参考文献

[1] Donenfeld J A. WireGuard: Next Generation Kernel Network Tunnel[C]//NDSS. 2017.
[2] Berger T. Analysis of current VPN technologies[C]//First International Conference on Availability, Reliability and Security (ARES'06). IEEE, 2006: 8 pp.-115.
[3]李京春等.网络安全态势感知技术标准化白皮书.
[4] http://newitnavi.h3c.com/2020/03/Catalog/Lecture_Hall/202010/1348056_233453_0.html
[5] Bezerra J M, Pinheiro A J, de Souza C P, et al. Performance evaluation of elephant flow predictors in data center networking[J]. Future Generation Computer Systems, 2020, 102: 952-964.
[6] McCarthy C, Zincir-Heywood A N. An investigation on identifying SSL traffic[C]//2011 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA). IEEE, 2011: 115-122.
[7] Pang Y, Jin S, Li S, et al. OpenVPN Traffic Identification Using Traffic Fingerprints and Statistical Characteristics[C]//International Conference on Trustworthy Computing and Services. Springer, Berlin, Heidelberg, 2012: 443-449.
[8] Davis J J, Foo E. Automated feature engineering for HTTP tunnel detection[J]. computers & security, 2016, 59: 166-185.
[9] Liu J, Li S, Zhang Y, et al. Detecting DNS tunnel through binary-classification based on behavior features[C]//2017 IEEE Trustcom/BigDataSE/ICESS. IEEE, 2017: 339-346.
[10] Lin H, Liu G, Yan Z. Detection of application-layer tunnels with rules and machine learning[C]//International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage. Springer, Cham, 2019: 441-455.
[11] He Y, Li W. Image-based encrypted traffic classification with convolution neural networks[C]//2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC). IEEE, 2020: 271-278.
[12] Cheng J, He R, Yuepeng E, et al. Real-Time Encrypted Traffic Classification via Lightweight Neural Networks[C]//GLOBECOM 2020-2020 IEEE Global Communications Conference. IEEE, 2020: 1-6.
[13] MontazeriShatoori M, Davidson L, Kaur G, et al. Detection of DoH Tunnels using Time-series Classification of Encrypted Traffic[C]//2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech). IEEE, 2020: 63-70.
[14] Leroux S, Bohez S, Maenhaut P J, et al. Fingerprinting encrypted network traffic types using machine learning[C]//NOMS 2018-2018 IEEE/IFIP Network Operations and Management Symposium. IEEE, 2018: 1-5.
[15] 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.
[16] 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.
[17]陈昱彤.硕士学位论文.隧道内网络行为识别技术研究.2020

参考文献

[18] Li Y, Lu Y. Multimodality Data Analysis in Information Security ETCC: Encrypted Two-Label Classification Using CNN[J]. Security and Communication Networks, 2021, 2021.
[19] Lin K, Xu X, Gao H. TSCRNN: A novel classification scheme of encrypted traffic based on flow spatiotemporal features for efficient management of IIoT[J]. Computer Networks, 2021, 190: 107974.
[20] Zhang J, Li F, Wu H, et al. Autonomous model update scheme for deep learning based network traffic classifiers[C]//2019 IEEE Global Communications Conference (GLOBECOM). IEEE, 2019: 1-6.
[21] Wang Z X, Wang P, Zhou X, et al. FLOWGAN: Unbalanced network encrypted traffic identification method based on GAN[C]//2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom). IEEE, 2019: 975-983.
[22] 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, 2019.
[23] Zhang J, Li F, Ye F, et al. Autonomous Unknown-Application Filtering and Labeling for DL-based Traffic Classifier Update[C]//IEEE INFOCOM 2020-IEEE Conference on Computer Communications. IEEE, 2020: 397-405.
[24] Ding Y, Cai W. A method for HTTP-tunnel detection based on statistical features of traffic[C]//2011 IEEE 3rd International Conference on Communication Software and Networks. IEEE, 2011: 247-250.
[25] Zheng W, Gou C, Yan L, et al. Learning to Classify: A Flow-Based Relation Network for Encrypted Traffic Classification[C]//Proceedings of The Web Conference 2020. 2020: 13-22.
[26] Sun B, Yang W, Yan M, et al. An Encrypted Traffic Classification Method Combining Graph Convolutional Network and Autoencoder[C]//2020 IEEE 39th International Performance Computing and Communications Conference (IPCCC). IEEE, 2020: 1-8.
[27] Juarez M, Afroz S, Acar G, et al. A critical evaluation of website fingerprinting attacks[C]//Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. 2014: 263-274.
[28] Wang,T. & Goldberg,I.(2016).On Realistically Attacking Tor with Website Fingerprinting. Proceedings on Privacy Enhancing Technologies,2016(4) 21-36.
[29] Xu Y, Wang T, Li Q, et al. A multi-tab website fingerprinting attack[C]//Proceedings of the 34th Annual Computer Security Applications Conference. 2018: 327-341.
[30] Cui W, Chen T, Fields C, et al. Revisiting assumptions for website fingerprinting attacks[C]//Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security. 2019: 328-339.



欢迎各位老师批评与指正！

Thanks