

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2022 年 3 月 3 日 (03.03.2022)



(10) 国际公布号
WO 2022/041394 A1

(51) 国际专利分类号:
G06K 9/62 (2006.01) **G06N 3/04** (2006.01)

(21) 国际申请号: PCT/CN2020/118725

(22) 国际申请日: 2020 年 9 月 29 日 (29.09.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202010885293.1 2020年8月28日 (28.08.2020) CN

(71) 申请人: 南京邮电大学(NANJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) [CN/CN];
中国江苏省南京市鼓楼区新模范马路 66 号, Jiangsu 210003 (CN)。

(72) 发明人: 徐小龙(XU, Xiaolong); 中国江苏南京鼓楼区新模范马路66号, 江苏 210003 (CN)。林焜达(LIN, Kunda); 中国江苏南京鼓楼区新模范马路66号, 江苏 210003 (CN)。

(74) 代理人: 南京纵横知识产权代理有限公司(NANJING ZONGHENG INTELLECTUAL PROPERTY AGENCY CO., LTD); 中国江苏省南京浦口高新区高新路9号自主创新广场3楼, Jiangsu 210032 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,

(54) Title: METHOD AND APPARATUS FOR IDENTIFYING NETWORK ENCRYPTED TRAFFIC

(54) 发明名称: 一种网络加密流量识别方法及装置

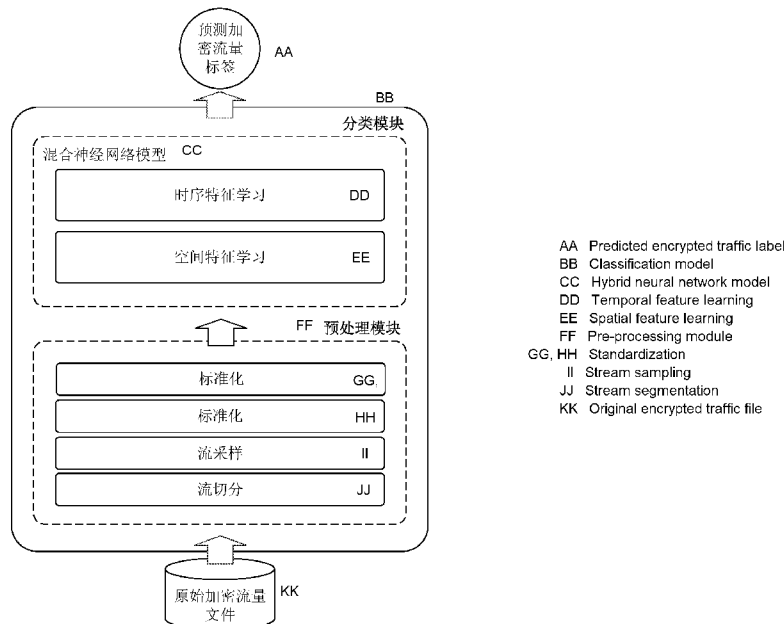


图 1

(57) Abstract: Disclosed is a method and apparatus for identifying network encrypted traffic. The method comprises a pre-processing stage and a classification stage. In the pre-processing stage, stream segmentation, sampling, vectorization and standardization are performed on original traffic, and a sampling scheme in a large stream is proposed to solve the classification problem of large-stream traffic. In the classification stage, spatial feature capture and abstract feature extraction are first performed using a CNN, and then temporal features of the traffic are learned using a stacked bidirectional LSTM on the basis of abstract features, so as to realize automatic feature

LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX,
MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL,
PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US,
UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

extraction and efficient identification of the encrypted traffic. The method is versatile, can automatically extract the temporal and spatial features of encrypted traffic without manual feature design by experts, and can adapt to changes in traffic features caused by different encryption technologies and obfuscation technologies.

(57) 摘要: 本发明公开了一种网络加密流量识别方法及装置, 该方法包含预处理阶段和分类阶段。预处理阶段对原始流量进行流切分, 采样, 向量化和标准化, 并提出大流流中采样方案, 解决大流流量的分类问题。分类阶段先使用CNN进行空间特征捕获和抽象特征抽取, 然后在抽象特征的基础上使用堆叠双向LSTM学习流量时序特征, 实现加密流量的自动特征提取和高效识别。该方法具有通用性, 能够自动提取加密流量时空特征而无需专家手动特征设计, 并且, 它能够适应不同加密技术、混淆技术引起流量特征变化。

一种网络加密流量识别方法及装置

技术领域

本发明具体涉及一种网络加密流量识别方法，还涉及一种网络加密流量识别装置，属于深度学习、网络流量分析和网络空间安全技术领域。

背景技术

流量分类是现代网络通讯中最重要的任务之一，但是由于加密技术的普及和网络吞吐量的高速增长，实现高速准确的加密流量识别变得越来越困难。加密流量分类对于流量工程、网络资源管理、QoS (Quality of Service)、网络空间安全管理等有着重要的意义。近年来，在新型网络领域例如物联网网络、软件定义网络、移动互联网中同样出现了加密流量分析管理的巨大需求。因为上述原因，网络流量分类吸引了越来越多的来自学术界、工业界两方面研究人员的注意。

近来，随着人们在安全性和隐私性方面的需求越来越高，流量加密技术逐步发展，加密流量如今已经成为了工业界普遍的做法，研究指出，到 2020 年将有超过 83% 的流量被加密。经过加密程序，流量变得随机化，这种伪随机格式使得流量的解析变得非常困难。另一方面，ISP (Internet Service Provider) 通常需要对某些类型的流量进行监测或控制（例如 P2P，入侵攻击等等），为了规避监测系统或防火墙的检测，一些开发商使用了各种协议嵌入和流量混淆技术。显然，流量加密、混淆技术的出现，一方面满足了人们的需求，提高了安全性和隐私性，另一方面也对网络的管理提出了更大的挑战。因此，加密流量分类成为了流量工程、入侵检测等任务中的关键技术。

现有加密流量分类的解决方案大致可以分为三种：基于端口、基于荷载（例如，深度包检测，Deep Packet Inspection，简称 DPI）、基于统计特征。由于动态端口和端口伪装技术的盛行，传统基于端口方法的流量分类的准确率很低。而基于荷载检测方法，如 DPI，它类似于字符串正则匹配算法，需要指纹库中的所有样本都需要和完整的流量进行匹配，因而效率很低，更重要的是，这些指纹一般难以用于识别加密流量。现有的工作更多集中于基于统计的机器学习方法。这类方法需要专家手动设计、提取流量的统计特征，从而对流量进行较为准确的分类。然而，基于统计特征的机器学习方法，专家需要对不同场景下的流量设计不同的统计特征，成本很高，也无法保证提取的特征对提高分类结果的有效性。基于以上原因，这些方法难以满足人们在解决加密流量分类问题中的需求。

近来，深度学习迅速发展，在计算机视觉、自然语言处理等等领域取得了令人瞩目的成果，其中包括大量的分类问题（例如，图像分类，文本情感分析）。与此同时，深度学习方法也逐渐应用于网络领域，例如流量分类就可以当作一个典型的分类问题。在深度学习方法中，CNN（卷积神经网络）擅长捕获数据空间特征，RNN（循环神经网络）擅长捕获数据时间特征。已经有一些研究使用深度学习对加密流量进行分类，其中大多数使用 CNN，在包级别上捕获流量的字节特征，但对于包与包之间，时间序列的时序特征没有很好的利用。

综上所述，当前工作中对于加密流量分类的研究仍存在以下不足：

- 1) 随着加密技术和混淆技术的普及，流量特征容易变化，基于规则的方法（包括基于端口、基于荷载的方法）规则提取困难，流量变动之后容易失效，时间效率低。

2) 基于统计的机器学习方法，手动设计特征困难，为获取更准确流量统计特征通常需要更加耗时的离线算法，实时性差。

3) 基于深度学习的研究仍然较少，现有工作没有有效利用流量的时空特征。

发明内容

本发明的目的在于克服现有技术中的不足，提供了一种网络加密流量识别方法及装置，解决了当前网络环境中加密技术导致流量识别算法耗时高、实时性差等问题。

为解决上述技术问题，本发明提供了一种网络加密流量识别方法，包括以下过程：

获取多个待识别的加密流量文件；

对待识别的加密流量进行预处理，所述预处理包括：将加密流量流切分为多个流；然后从每个流中采集多个连续数据包作为样本；最后将每个样本进行向量化、标准化处理，得到格式化的样本向量集合；

将预处理后得到的样本向量集合输入至预设训练的混合神经网络模型，得到预测向量，此预测向量中元素值代表加密流量属于各个分类的预测值；

所述混合神经网络模型包括：1D-CNN 网络、堆叠双向 LSTM 网络 and 全连接层网络；其中 1D-CNN 网络对输入样本向量集合进行空间特征学习，输出低维特征图；堆叠双向 LSTM 网络对输入的特征图进行时序特征学习，得到包含时序特征的特征图向量，全连接层根据输入的时序特征的特征图向量确定预测向量；

基于预测向量计算得到各分类预测概率分布，取其中最大的概率对应的分类作为加密流量最终的分类标签。

进一步的，所述从流中采集多个连续数据包作为样本，包括：

若流为小流，采集流头部预设个数连续数据包组成一个样本，若已有数据包不足预设个数，则选择已有数据包，其余包补零处理；

若流为大流，从流中选取若干个采样点，以每个采样点作为起点，采集连续预设个数连续数据包组成一个样本。

进一步的，所述采样点的选取方案包括：随机点采样，固定步长采样和突发点采样三种策略；其中：

所述随机点采样为流中随机点采样；所述固定步长采样以固定的步长从流量起始开始采样；所述突发点采样为寻找大流中的数据流突发点进行采样。

进一步的，所述将样本进行向量化、标准化处理，包括：

将每个数据包保留预设长度字节数，不足则用全零补全，反之则进行截断；将每个样本转化为的向量；

对向量中每个数据进行标准化处理。

进一步的，所述 1D-CNN 网络，包括：

1D-CNN 网络部分由两层 1D-CNN 卷积层组成，对输入加密流量样本向量进行两次卷积操作，并且在每一层中对卷积操作输出的新特征图进行批标准化、非线性激活和降采样处理。

进一步的，所述混合神经网络模型的训练包括：

获取多个加密流量文件，对每个加密流量文件标注出对应的分类标签，

对各个加密流量文件进行预处理，所述预处理包括：将每个加密流量流切分为多个流；然后从每个流中采集多个连续数据包作为样本；最后将每个样本进行向量化、标准化处理，得到格式化的样本向量集合作为训练样本；

利用训练样本对混合神经网络模型进行训练，其中包括 1D-CNN 网络、堆叠双向 LSTM 网络 and 全连接层网络三个部分，以得到最佳网络参数；

得到训练完成的混合神经网络模型。

进一步的, 所述 1D-CNN 网络的训练包括:

在向量化过程中保留每个样本中数据包的时序维度, 设 $\mathbf{x}^{(t)} \in \mathbb{R}^L$ 为一个样本中第 t 个数据包, t 为 0 到 n 的任一整数, 代指向量中任一数据包, 它是一个 L 维度的向量;

$$\mathbf{x} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(M)}] \quad (5)$$

$\mathbf{x}$ 代表一个样本, 它包含 M 个数据包的向量, 在 1D-CNN 中, $\mathbf{x}$ 可视为包含 M 个通道, 每个通道都是 L 维的二维向量; 假设 $\mathbf{x}_{i:i+j}$ 代表全部通道从任意位置 i 到 $i+j$ 的字节; 在 $\mathbf{x}$ 上一维卷积操作如下:

$$\mathbf{a}_i^{(t)} = f(\mathbf{w} \cdot \mathbf{x}_{i:i+h-1} + b) \quad (6)$$

通常一个卷积层中包含多个卷积核, 每个 Filter 操作相同, 生成新特征图的一个通道; 以其中任一卷积核 t 为例, $\mathbf{w} \in \mathbb{R}^{hk}$ 为在 $\mathbf{x}$ 上滑动窗口, b 为偏移值, f 则是非线性的激活函数; $\mathbf{a}_i^{(t)}$ 为任一卷积核 t 生成的特征; 当前 Filter 在 $\mathbf{x}$ 上滑动时, 该 Filter 的卷积操作应用到窗口内的字节上, 从整体来看, 序列 $\{\mathbf{x}_{1:h}, \mathbf{x}_{2:h+1}, \dots, \mathbf{x}_{n-h+1:n}\}$ 将会生新特征图; 所有 Filter 操作相同, 但是每个 Filter 对应的参数 $\mathbf{w}$ 和 b 是不同的;

$$\mathbf{a}^{(t)} = [\mathbf{a}_1^{(t)}, \mathbf{a}_2^{(t)}, \dots, \mathbf{a}_{n-h+1}^{(t)}] \quad (7)$$

这里 $\mathbf{a}^{(t)} \in \mathbb{R}^{n-h+1}$, 代表任一卷积核 t 生成的新特征图, 也可视为输出通道 t ; 对于每个通道的新特征图, 通常还会使用池化操作层 (MaxPooling) 对特征图进行降采样; 池化操作层的操作和卷积操作类似, 同样也是使用 Filter 做滑动操作, 但在每个 Filter 上通常执行的运算为 $\hat{\mathbf{a}}_i^{(t)} = \max(\mathbf{a}_{i:i+h-1}^{(t)})$, 保留每个滑动窗口中的最大值。

进一步的, 所述堆叠双向 LSTM 网络的训练包括:

在 LSTM 中, 将输入特征图的多个通道视为多个时间步; 在每个时间步上有以下公式:

$$\tilde{\mathbf{c}}^{(t)} = \tanh(\mathbf{w}_c [\mathbf{h}^{(t-1)}, \mathbf{a}^{(t)}] + b_c) \quad (8)$$

其中 $\mathbf{a}^{(t)} \in \mathbb{R}^m$ ，表示在输入特征图任一时间步 t （即通道 t ）上的向量，其维度与每个时间步输入的特征图维度相同，假设为 m （即 1D-CNN 生成的新特征维度）； $\mathbf{h}^{(t-1)} \in \mathbb{R}^s$ ，上一个时间步隐藏层的输出，维度由 LSTM 单元的隐藏层维度的参数决定，假设为 s ； $\tilde{\mathbf{c}}^{(t)}$ 为当前层的中间输出； $\mathbf{w}_c$ 和 b_c 分别为参数矩阵和偏差；

在 LSTM 中，最终输出需要由三个门决定，分别为更新门 Γ_u 、遗忘门 Γ_f 和输出门 Γ_o ；门的值计算公式如下：

$$\Gamma_u = \sigma(\mathbf{w}_u[\mathbf{h}^{(t-1)}, \mathbf{a}^{(t)}] + b_u) \quad (9)$$

$$\Gamma_f = \sigma(\mathbf{w}_f[\mathbf{h}^{(t-1)}, \mathbf{a}^{(t)}] + b_f) \quad (10)$$

$$\Gamma_o = \sigma(\mathbf{w}_o[\mathbf{h}^{(t-1)}, \mathbf{a}^{(t)}] + b_o) \quad (11)$$

其中， σ 为非线性激活函数， $\mathbf{w}_u$ 、 $\mathbf{w}_f$ 、 $\mathbf{w}_o$ 和 b_u 、 b_f 、 b_o 分别为三个门对应的参数矩阵和偏差值；

三个门的值的计算方法，与 $\tilde{\mathbf{c}}^{(t)}$ 计算方法类似，它们取值由当前时间步的输入 $\mathbf{a}^{(t)}$ 和上一层隐藏层的输出 $\mathbf{h}^{(t-1)}$ 决定；更新门 Γ_u 、遗忘门 Γ_f 和输出门 Γ_o 的作用类似与开关，用于控制当前 LSTM 单元是否更新当前信息，是否遗忘过去信息，是否输出最终信息；三个开关（门）产生最终输出的公式如下，其中 $\mathbf{c}^{(t)}$ 当前层的中间输出向量：

$$\mathbf{c}^{(t)} = \Gamma_u \odot \tilde{\mathbf{c}}^{(t)} + \Gamma_f \odot \mathbf{c}^{(t-1)} \quad (12)$$

$$\mathbf{h}^{(t)} = \Gamma_o \odot \tanh(\mathbf{c}^{(t)}) \quad (13)$$

由三个门分别对过去、当前和总输出做取舍；最终输出当前时间步的隐藏层输出 $\mathbf{h}^{(t)}$ ；其中 $\odot$ 为哈达玛积（Hadamard Product），表示向量之间按位相乘；

堆叠 LSTM 指多层 LSTM 单元堆叠而成，而双向 LSTM 则是在时间步上正向、反向同时进行 LSTM 运算，Bi-LSTM 的计算只需要将当前时间步不同方向的输出做连接即可，如：

$$\mathbf{h}^{(t)} = [\vec{\mathbf{h}}^{(t)}, \overleftarrow{\mathbf{h}}^{(t)}] \quad (14)$$

在每个时间步上，隐藏层的输出 $\mathbf{h}^{(t)}$ 都由时间步上正向的输出 $\vec{\mathbf{h}}^{(t)}$ 和反向的输出 $\overleftarrow{\mathbf{h}}^{(t)}$ 连接而成。

进一步的，所述基于预测向量计算得到各分类预测概率分布，包括：

计算预测概率分布，其计算公式如下：

$$\text{softmax}(\mathbf{o})_i = \hat{\mathbf{y}}_i = \frac{e^{o_i}}{\sum_j e^{o_j}} \quad (15)$$

其中 $\mathbf{o}$ 为混合神经网络模型原始输出向量， $\hat{\mathbf{y}}$ 则为经过 softmax 处理后的预测概率向量， $\mathbf{o}_i$ 为向量 $\mathbf{o}$ 中第 i 位置的向量值。

相应的，本发明还提供了一种网络加密流量识别装置，包括加密流量获取模块、预处理模块、分类预测模块和分类识别模块；其中：

加密流量获取模块，用于获取待识别的加密流量文件；

预处理模块，用于对待识别的加密流量进行预处理，所述预处理模块包括流切分单元、采集单元和向量化单元，其中：

流切分单元，用于将加密流量流切分为多个流；

采集单元，用于从每个流中采集多个连续数据包作为样本；

向量化单元，用于将每个样本进行向量化、标准化处理，得到格式化的样本向量集合；

分类预测模块，用于将预处理后得到的样本向量集合输入至预设训练的混合神经网络模型，得到预测向量，此预测向量中元素值代表加密流量属于各个分类的预测值；

所述混合神经网络模型包括：1D-CNN 网络、堆叠双向 LSTM 网络 and 全连接层网络；其中 1D-CNN 网络对输入样本向量集合进行空间特征学习，输出低维特征图；堆叠双向 LSTM 网络对输入的特征图进行时序特征学习，得到包含时序特征的特征图向量，全连接层根据输入的时序特征的特征图向量确定预测向量；

分类识别单元，用于基于预测向量计算得到各分类预测概率分布，取其中最大的概率对应的分类作为加密流量最终的分类标签。

进一步的，所述采集单元中，包括：

小流采样单元，用于采集流头部预设个数连续数据包组成一个样本，若已有数据包不足预设个数，则选择已有数据包，其余包补零处理；

大流采样单元，用于从流中选取若干个采样点，以每个采样点作为起点，采集连续预设个数连续数据包组成一个样本。

进一步的，所述大流采样单元中，采样点的选取方案包括：随机点采样，固定步长采样和突发点采样三种策略；其中：

所述随机点采样为流中随机点采样；所述固定步长采样以固定的步长从流量起始开始采样；所述突发点采样为寻找大流中的数据流突发点进行采样。

进一步的，所述向量化单元中将样本进行向量化、标准化处理，包括：
将每个数据包保留预设长度字节数，不足则用全零补全，反之则进行截断；
将每个样本转化为的向量；

对向量中每个数据进行标准化处理。

进一步的，所述 1D-CNN 网络，包括：

1D-CNN 网络部分由 两层 1D-CNN 卷积层组成，对输入加密流量向量进行两次卷积操作，并且在每一层中对卷积操作输出的新特征图进行批标准化、非线性激活和降采样处理。

与现有技术相比，本发明所达到的有益效果是：

1) 本发明基于深度学习的技术（CNN 和 RNN），实现对流量特征的自动提取。该方法具有通用性，并不针对特定的网络环境、特定的应用场景，与基于规则的方法相比，该方法可以适应不同加密技术、混淆技术带来的流量特征

变化。

2) 本发明提出一种混合神经网络模型, 结合 CNN 和 RNN, 仅使用少量数据包, 对流量进行抽象特征抽取, 学习数据流的时空特征, 实现流量的早期识别。该方法不需要专家进行手动特征设计, 在多个真实网络数据集的测试中, 表现高于基于传统机器学习的识别方法。

3) 本发明对原始加密流量进行自动流切分, 向量化, 标准化等处理, 保留流的时序特征。该方法有效利用流量数据空间分布和时序特征, 实现对特征的自动学习, 实现了端到端的加密流量识别方法。此外, 对于大流流量, 该方法提出一种流中采样方案, 解决长时间流量的分类问题和数据不平衡性。

附图说明

图 1 为加密流量识别方法的整体框架;

图 2 为流量向量化方法示意图;

图 3 为混合神经网络模型整体架构图;

图 4 为加密流量识别方法流程图;

图 5 为分类模型的详细架构及参数设置示意图。

具体实施方式

下面结合附图对本发明作进一步描述。以下实施例仅用于更加清楚地说明本发明的技术方案, 而不能以此来限制本发明的保护范围。

实施例 1

本发明提供了一种网络加密流量识别方法, 其特征是, 包括以下过程:

获取待识别的加密流量文件;

对待识别的加密流量进行预处理, 所述预处理包括: 将加密流量流切分为

多个流；然后从每个流中采集多个连续数据包作为样本；最后将每个样本进行向量化、标准化处理，得到格式化的样本向量集合；

将预处理后得到的样本向量集合输入至预设训练的混合神经网络模型，得到预测向量，此预测向量中元素值代表加密流量属于各个分类的预测值；

所述混合神经网络模型包括：1D-CNN 网络、堆叠双向 LSTM 网络 and 全连接层网络；其中 1D-CNN 网络对输入样本向量集合进行空间特征学习，输出低维特征图；堆叠双向 LSTM 网络对输入的特征图进行时序特征学习，得到包含时序特征的特征图向量，全连接层根据输入的时序特征的特征图向量确定预测向量；

基于预测向量计算得到各分类预测概率分布，取其中最大的概率对应的分类作为加密流量最终的分类标签。

本发明利用混合神经网络技术，实现对加密流量时空特征的自动学习，从而实现对加密流量的高速、准确识别。

实施例 2

用于识别加密流量的特征的抽取和流量预处理方式、向量化方法、流量数据流不同部分的信息有关。例如，流量的元信息和荷载信息，它们能够提供不同且有效的特征用于加密流量的识别。在本方案中，一方面，考虑结合使用流量元信息、数据包部分荷载、数据包之间的时序特征等信息，以提高数据完备性。另一方面，该方法中设计了混合神经网络模型对上述信息进行自动表征学习。

图 1 为本发明方法的整体框架图，主要包含两个阶段：预处理阶段和分类阶段。预处理阶段将原始流量直接转化成标准数据，其中包含流切分，流采样，向量化，标准化四个步骤。分类阶段，通过设计混合神经网络模型捕获流时空

特征，包括空间分布特征（抽象特征）学习部分和时序特征学习部分，实现加密流量的分类。

为了方便理解本发明的技术方案，下面定义一些概念：

定义 1 **五元组** (T)。

对于网络中的每一个数据包，根据数据包的头部信息（元信息），可以找出相应的五元组信息，表达如下：

$$T = \{Src\ IP, Src\ Port, Dest\ IP, Dest\ Port, Protocol\} \quad (1)$$

分别代表源地址，源端口，目的地址，目的端口和传输层协议。

定义 2 **流中采样技术**。

真实网络环境中存在大量长时间通信连接而产生长时间的大流流量，这些流可能包含几万个到几百万个数据包，单个流的体量巨大。例如匿名网络中因为使用了虚电路技术，有大量的数据包具备相同的五元组信息。在一些数据中心中，流量分析管理的重点也在于大流流量，一些文献称其为大象流。若每一个流表示为 $F = [P_0, P_1, \dots, P_n]$ ， P_i 是 F 中具有相同五元组的数据包。从流中 n 个包中选取 m 个采样点， $S = \{s_0, s_1, \dots, s_m\}$ 。以每个采样点作为起点，采集连续 M 个连续数据包组成样本， $F_i = [P_{s_i}, P_{s_i+1}, \dots, P_{s_i+M}]$ ，其中 i 为 0 到 m 的整数。本发明中提出三种流采样方案。

① 随机点采样（Random Sampling）：默认策略，流中随机点采样。采样点 s_i 为 0 到 n 的随机点。

$$s_i \in \{0, 1, 2, \dots, n\} \quad (2)$$

② 固定步长采样（Fixed Step Sampling）：以固定的步长从流量起始开始采样。其中步长为固定长度的常数，表示固定的步长，相邻采样点符合以下公式。

$$s_0 = 0 \quad (3)$$

$$s_i = s_{i-1} + step \quad (4)$$

③ 突发点采样（Burst Sampling）：寻找大流中的突发点进行采样。在大流

流量中，用户不同行为可以引起流量包长度的变化，例如用户点击行为引起数据传输，通常会引起流起伏波动。数据流变动前通常需要一些不携带数据的帧进行通信，在网络中，不携带数据的 TCP 或 UDP 帧长不超过 60 字节。因此，Burst Sampling 检测此类数据点，选为采样点。

本发明的一种网络加密流量识别方法，其包括对原始流量文件的预处理，采样，向量化，时空特征学习等过程。本发明的加密流量识别方案流程图如图 4 所示。其具体操作步骤如下：

步骤 1：根据流量数据包五元组信息对原始加密流量进行流切分，得到包含相同五元组信息的数据包的流集合。

真实网络环境中，在某个节点采集的流量并不是来自单一应用的有序序列，而是包含很多应用的混合序列。比如在某时间段在某个网关采集流量，可能包含该网络中经过该网关的所有主机产生的数据包，数据包混合成当前的吞吐流量。为了将各个单一类型产生的数据流进行分离，需要对数据集中提供的原始流量进行流切分。

常见的加密协议有两种，应用层加密和网络层加密。应用层加密主要指对报文的应用层协议进行加密，常见协议例如 BitTorrent、Https 等。网络层加密是对网络层以上整个报文进行加密。

加密后的流量通常仍会包含未加密部分，例如流量的元信息。因此，我们可以根据流的元信息来对加密流量进行流切分，但无法进一步获取加密部分的应用层信息及荷载信息。在网络中，流指的是包含相同五元组（源 IP，源端口，目的 IP，目的端口和传输层协议）的所有数据包。按照五元组信息对原始流量文件（PCAP 文件，保存网络加密流量）进行读取、缓存、切分生成流 $F = [P_0, P_1, \dots, P_i, \dots, P_n]$ 。 P_i 是 F 中具有相同五元组的第 i 个数据包。流切分后得到包含相同五元组信息的数据包的流组成的流集合。

按照原始流量文件的类型对切分的流进行标注：若网络流量数据集中的 PCAP 文件有类型标签信息（例如该文件有标注流量服务类型，Chat，Email，

Video 等, 取决于不同分类任务), 则将从该文件中获得的所有流标注为该类型。用于训练混合神经网络模型。

步骤 2: 对步骤 1 中得到的流集合, 根据流持续时常(可划分为大流或小流)使用不同采样方案, 从每个流中采样连续数据包组成原始训练样本, 得到原始样本集合。

真实网络为不平衡环境, 网络中的流的长度差距巨大, 流量上行下行通常也不对称。在一些文献中, 将数据中的流类型分为大流(长时间数据流)和小流(短时间数据流)。不同流的时长不同, 包含的数据包也不相同。小流可能包含几十到几百个数据包, 而大流则可能包含几万到几百万个数据包。大流流量通常需要占用极大的存储空间, 因而若以流为单位, 难以采集到足够的流样本用于训练。对于小流, 通常只包含一次通信交互请求, 因而连接特征多集中在前期少量数据包中; 而在大流中, 包含应用多次通信交互过程, 所以流中数据包应该同样包含大量可用于流量分类的有效信息。因此, 在本方法中, 每个小流使用流头部少量数据包作为一个单独样本, 而每个大流则考虑采用流中采样技术, 从流中采集多个样本, 缓解网络流量数据不平衡问题(在现有数据集中, 大流样本容量大, 但样本数量极少)。

根据流类型(大流或小流)使用不同采样方案, 从每个流中选取预设个数(记为 M 个, M 为较小常数, 例如 $M=10$)连续数据包组成一个训练样本。对于小流, 采集流头部数据包, 其包含大部分通信连接建立信息。对于大流, 使用流中采样技术, 从大流中选择适量采样点, 选取每个采样点的少量连续数据包组成单独样本, 从而解决大流流量数据不平衡问题。

采集得到的每个样本, 都保留 M 个数据包的原因包括两个方面: 一方面, 有利于实现流量早期识别, 只使用少量的包完成加密流量识别(每个流可能包含几十个到几万个包, 本方法仅使用少量的包, 例如 $M=10$), 使方法轻量化, 极大减轻识别方法的存储需求。另一方面, 在实践上, 格式化的数据有助于计算机训练模型时的运算。

此步骤的具体过程为：

① 若输入流为小流，则使用下面步骤②，执行小流采样方案。若输入流为大流，则进入下面步骤③，执行流中采样方案。

② 对于小流，采集流头部 M 个数据包， $F_{sub} = [P_0, P_1, \dots, P_M]$ 组成一个样本，若已有数据包不足 M 个则全选已有数据包，其余包补零处理。

③ 对于大流，从流中 n 个包中选取 m 个采样点， $S = \{s_0, s_1, \dots, s_m\}$ 。以每个采样点作为起点，采集连续 M 个连续数据包组成样本， $F_i = [P_{s_i}, P_{s_i+1}, \dots, P_{s_i+M}]$ ，其中 i 为 0 到 m 的整数。采样点选取方法有三种策略：随机点采样（Random Sampling），默认策略，流中随机点采样；固定步长采样（Fixed Step Sampling），以固定的步长从流量起始开始采样；突发点采样（Burst Sampling），寻找大流中的数据流突发点进行采样。

步骤 3：对于步骤 2 的原始样本集合，每个样本都包含 M 个数据包，每个数据包保留 L 字节的长度，将每个样本转化为维度为 (M, L) 的向量，从而将原始流量规整为统一形状（Shape）的便于计算机读取、运算的向量。然后对于每个样本向量进行标准化处理，得到格式化的样本向量集合，现有研究表明，数据标准化处理能够加速梯度下降，使模型快速收敛。

具体处理过程如下：

① 步骤 2 中得到的每个原始样本都包含 M 个数据包，每个数据包保留固定预设长度 L ，不足则用全零补全，反之则进行截断。其中 L 默认取值 1500，这是因为以太网中的 MTU（Maximum transmission Unit，即以太网最大帧长）为 1500 字节，为了方法的通用性，每个包默认保留长度 L 。图 2 以二维的形式展示了经过格式化后的每个样本。

② 对步骤①格式化后的样本，按字节读取二进制数据流，将每个字节中 8 位的二进制数码以十进制的方式读取，得到 0 到 255 的整数。以整数形式来表示字节，进而表示整个向量的方法，实现了对原始样本的向量化，方便运算。每个原始样本被转化为维度为 (M, L) 的向量。

③ 为了加速计算，减少深度学习中梯度爆炸问题，对步骤②中得到向量进行标准化处理。由于向量化中将每个字节读取为一个整型数字（0 到 255），所以这里可以直接将这些数除以 255 进行标准化，得到格式化的样本集合。

步骤 4：重复步骤 1-步骤 3 得到大量格式化的训练样本，将训练样本输入混合神经网络模型进行训练。

不同流量具备不同的时序和空间分布特征（时空特征），混合神经网络模型可以提取流的时空特征，提高模型预测的准确性。传统方法需要专家根据流量未加密信息（例如数据报头部信息）、通信交互行为、荷载分布等信息，手动设计规则或统计特征（例如流时长，流大小，包大小，包间隔等）用于流量分类。混合神经网络模型不需要手动特征设计，实现了流量特征的自动提取。

混合神经网络模型包含空间特征（抽象特征）学习部分和时序特征学习部分。卷积神经网络（CNN）在图像领域被广泛应用，现有研究表明，经过多层 CNN 的降采样，模型能够学习图像空间分布上的更抽象特征（例如动物图像的局部特征，眼镜，嘴巴，四肢等）。步骤 3 得到的原始向量维度较高，引入有效信息的同时会带来更多噪声，使得模型更难进行特征学习。因此本发明中空间特征（抽象特征）学习部分使用一维卷积神经网络(1D-CNN)，进行多次降采样，从而降低特征维度，学习到流量在空间分布上的抽象特征。时序特征学习部分使用堆叠双向 LSTM（长短期记忆单元，Long Short-term Memory）捕获流量数据包之间的时间相关性。

为了详细阐述混合神经网络模型的设计，将从原理，整体架构，详细参数三方面详细展示模型的细节：

1) 原理

本发明考虑在向量化过程中保留每个样本中数据包的时序维度。设 $\mathbf{x}^{(t)} \in \mathbb{R}^L$ 为一个样本中第 t 个数据包， t 为 0 到 n 的任一整数，代指向量中任一数据包，它是一个 L 维度的向量。

$$\mathbf{x} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(M)}] \quad (5)$$

$\mathbf{x}$ 代表一个样本,它包含 M 个数据包的向量,在1D-CNN中, $\mathbf{x}$ 可视为包含 M 个通道,每个通道都是 L 维的二维向量。假设 $\mathbf{x}_{i:i+j}$ 代表全部通道从任意位置 i 到 $i+j$ 的字节。在 $\mathbf{x}$ 上一维卷积操作如下:

$$\mathbf{a}_i^{(t)} = f(\mathbf{w} \cdot \mathbf{x}_{i:i+h-1} + b) \quad (6)$$

通常一个卷积层中包含多个卷积核(Filter),每个Filter操作相同,生成新特征图的一个通道。以其中任一卷积核 t 为例, $\mathbf{w} \in \mathbb{R}^{hk}$ 为在 $\mathbf{x}$ 上滑动窗口, b 为偏移值, f 则是非线性的激活函数。 $\mathbf{a}_i^{(t)}$ 为任一卷积核 t 生成的特征。当前Filter在 $\mathbf{x}$ 上滑动时,该Filter的卷积操作应用到窗口内的字节上,从整体来看,序列 $\{\mathbf{x}_{1:h}, \mathbf{x}_{2:h+1}, \dots, \mathbf{x}_{n-h+1:n}\}$ 将会生新特征图。所有Filter操作相同,但是每个Filter对应的参数 $\mathbf{w}$ 和 b 是不同的。

$$\mathbf{a}^{(t)} = [\mathbf{a}_1^{(t)}, \mathbf{a}_2^{(t)}, \dots, \mathbf{a}_{n-h+1}^{(t)}] \quad (7)$$

这里 $\mathbf{a}^{(t)} \in \mathbb{R}^{n-h+1}$,代表任一卷积核 t 生成的新特征图,也可视为输出通道 t 。对于每个通道的新特征图,通常还会使用池化操作层(MaxPooling)对特征图进行降采样。池化操作层的操作和卷积操作类似,同样是使用Filter做滑动操作,但在每个Filter上通常执行的运算为 $\hat{\mathbf{a}}_i^{(t)} = \max(\mathbf{a}_{i:i+h-1}^{(t)})$,保留每个滑动窗口中的最大值。

本质上,1D-CNN和全连接神经网络其实是类似的,但1D-CNN的特点在于卷积核权重共享,并进行稀疏连接,这对于高维向量的运算有较大帮助。另一方面,在1D-CNN中对流量进行多次降采样,随着层次的增加,卷积操作将会产生更加抽象的特征图,因此,混合神经网络模型将会从原始流量中学习更高级的空间分布上的抽象特征,这会帮助随后的时序特征的学习。

网络流量也是一种时间相关性极强的数据,因此也适用于LSTM。但网络原始流量文件向量化之后特征维度非常大,因此我们考虑在学习到抽象特征上基于LSTM设计网络架构。在LSTM中,将输入特征图的多个通道视为多个时间步。在每个时间步上有以下公式:

$$\tilde{\mathbf{c}}^{(t)} = \tanh(\mathbf{w}_c[\mathbf{h}^{(t-1)}, \mathbf{a}^{(t)}] + b_c) \quad (8)$$

其中 $\mathbf{a}^{(t)} \in \mathbb{R}^m$ ，表示在输入特征图任一时间步 t （即通道 t ）上的向量，其维度与每个时间步输入的特征图维度相同，假设为 m （即 1D-CNN 生成的新特征维度）。 $\mathbf{h}^{(t-1)} \in \mathbb{R}^s$ ，上一个时间步隐藏层的输出，维度由 LSTM 单元的隐藏层维度的参数决定，假设为 s 。 $\tilde{\mathbf{c}}^{(t)}$ 为当前层的中间输出。 $\mathbf{w}_c$ 和 b_c 分别为参数矩阵和偏差。

但在 LSTM 中，最终输出需要由三个门决定，分别为更新门 Γ_u 、遗忘门 Γ_f 和输出门 Γ_o 。门的值计算公式如下：

$$\Gamma_u = \sigma(\mathbf{w}_u[\mathbf{h}^{(t-1)}, \mathbf{a}^{(t)}] + b_u) \quad (9)$$

$$\Gamma_f = \sigma(\mathbf{w}_f[\mathbf{h}^{(t-1)}, \mathbf{a}^{(t)}] + b_f) \quad (10)$$

$$\Gamma_o = \sigma(\mathbf{w}_o[\mathbf{h}^{(t-1)}, \mathbf{a}^{(t)}] + b_o) \quad (11)$$

其中， σ 为非线性激活函数， $\mathbf{w}_u$ 、 $\mathbf{w}_f$ 、 $\mathbf{w}_o$ 和 b_u 、 b_f 、 b_o 分别为三个门对应的参数矩阵和偏差值。

可以看到，三个门的值的计算方法，与 $\tilde{\mathbf{c}}^{(t)}$ 计算方法类似，它们取值由当前时间步的输入 $\mathbf{a}^{(t)}$ 和上一层隐藏层的输出 $\mathbf{h}^{(t-1)}$ 决定。更新门 Γ_u 、遗忘门 Γ_f 和输出门 Γ_o 的作用类似与开关，用于控制当前 LSTM 单元是否更新当前信息，是否遗忘过去信息，是否输出最终信息。三个开关（门）产生最终输出的公式如下，其中 $\mathbf{c}^{(t)}$ 当前层的中间输出向量：

$$\mathbf{c}^{(t)} = \Gamma_u \odot \tilde{\mathbf{c}}^{(t)} + \Gamma_f \odot \mathbf{c}^{(t-1)} \quad (12)$$

$$\mathbf{h}^{(t)} = \Gamma_o \odot \tanh(\mathbf{c}^{(t)}) \quad (13)$$

由三个门分别对过去、当前和总输出做取舍。最终输出当前时间步的隐藏层输出 $\mathbf{h}^{(t)}$ 。其中 $\odot$ 为哈达玛积（Hadamard Product），表示向量之间按位相乘。

由于先进行抽象特征学习，为了增强时序特征捕获能力，我们使用堆叠双向 LSTM 网络，堆叠 LSTM 指多层 LSTM 单元堆叠而成，而双向 LSTM（Bi-LSTM）则是在时间步上正向、反向同时进行 LSTM 运算，这是考虑到当

前时间步的上下文信息（context），信息包含前后两个方面。Bi-LSTM 的计算只需要将当前时间步不同方向的输出做连接即可，例如：

$$\mathbf{h}^{(t)} = [\vec{\mathbf{h}}^{(t)}, \overleftarrow{\mathbf{h}}^{(t)}] \quad (14)$$

即在每个时间步上，隐藏层的输出 $\mathbf{h}^{(t)}$ 都由时间步上正向的输出 $\vec{\mathbf{h}}^{(t)}$ 和反向的输出 $\overleftarrow{\mathbf{h}}^{(t)}$ 连接而成。

2) 整体架构

混合神经网络模型分为两个阶段，图 3 是混合神经网络模型整体的架构图。输入加密流量的高维向量，首先使用基于 1D-CNN 网络进行抽象空间特征学习，经过两个一维卷积层（Conv-1，Conv-2）对输入样本向量集合进行空间特征学习和降采样，得到新的低维特征图。然后在此基础上，第二部分基于堆叠双向 LSTM 网络捕获时序特征，堆叠两层双向 LSTM，在每个时间步上，输入 1D-CNN 得到的特征图的每个通道的向量，通过堆叠双向 LSTM 学习特征图的时序特征，得到包含时序特征的特征图向量。最后，通过全连接层，将上一层特征图的维度转化为 c 维向量， c 为流量类型数量（例如流量服务类型，Chat，Email，Video 等，取决于不同分类任务）。最后可以通过 $softmax$ 函数从中得到加密流量的预测标签。

抽象特征学习部分，先基于 1D-CNN 设计模型网络，对流量进行自动抽象特征提取。1D-CNN 的特点在于卷积核权重共享，并进行稀疏连接，这降低了参数量，有利于捕获流量数据流中位于不同位置的相似空间特征。另一方面，1D-CNN 对流量进行多次降采样，随着层次的增加，卷积操作将会产生更加抽象的特征图，模型将会从原始流量中学习更高级的抽象特征，这会帮助随后的时序特征的学习。

时序特征学习部分，我们使用堆叠双向 LSTM 网络，堆叠 LSTM 指多层 LSTM 单元堆叠而成，而双向 LSTM（Bi-LSTM）则是在时间步上正向、方向同时进行 LSTM 运算，这是考虑到当前时间步的上下文信息包含当前位置前后两个方向的信息。

混合神经网络模型通过抽象特征学习和时序特征学习两部分实现了对流量特征的自动提取，无需专家进行手动特征设计。

对于每个输入样本，模型先使用 1D-CNN 进行空间特征学习和低采样，得到低维特征图，再使用 LSTM 学习得到包含时序特征的特征图，最后经过全连接层输出 c 维的预测向量 $\mathbf{o}$ ， c 为流量类型数量（例如流量服务类型，Chat, Email, Video 等，取决于不同分类任务），预测向量 $\mathbf{o}$ 中每个元素值代表待识别的加密流量属于各个分类的预测值。由于神经网络输出向量 $\mathbf{o} \in \mathbb{R}^c$ ，包含正数和负数，为了得到预测概率分布（全部概率相加为 1），我们需要使用指数运算将各个预测值转化为正数，即使用 *softmax* 函数处理向量 $\mathbf{o}$ ，计算出模型的预测概率分布向量 $\hat{\mathbf{y}}$ ， $\hat{\mathbf{y}}$ 同样为 c 维向量，第 i 位置上的输出代表该样本属于分类 i 的概率。注意 *softmax* 不参与混合神经网络模型的训练，用于计算预测概率分布，其计算公式如下：

$$\text{softmax}(\mathbf{o})_i = \hat{\mathbf{y}}_i = \frac{e^{\mathbf{o}_i}}{\sum_j e^{\mathbf{o}_j}} \quad (15)$$

其中 $\mathbf{o}$ 为混合神经网络模型原始输出向量， $\hat{\mathbf{y}}$ 则为经过 *softmax* 处理后的预测概率向量， $\mathbf{o}_i$ 为向量 $\mathbf{o}$ 中第 i 位置的向量值。公式 15 的计算原理为： e 为自然底数，利用指数运算 $e^{\mathbf{o}_i}$ 将 $\mathbf{o}_i$ 转为正实数。然后，将各个位置上的计算结果除以 $\sum_j e^{\mathbf{o}_j}$ ，计算出模型的预测概率分布向量 $\hat{\mathbf{y}}$ 。

最后使用交叉熵作为损失函数，利用梯度下降算法训练模型。

3) 详细参数设置

图 5 为混合神经网络模型的详细参数设置，包含 13 个层（见“层次名称”），13 层又可划分为 4 个大的层次（见“层次”），图中包含各层输入输出向量的大小（见“输入”、“输出”）以及每一层所使用的参数量（见“参数”），剩余卷积核大小和步长的则是 1D-CNN 的可设定参数（见“卷积核”、“步长”），整体可训练参数量为 2,897,104，整体参数可以代表神经网络模型的规模和整体模型的大小。分类模型由三部分组成：

第一部分为卷积相关层次。这个部分包含两个大卷积层（包含 Conv-1，

Conv-2)，每个大卷积层包含一层 1D-CNN，卷积核大小设定为 3，卷积核移动步长为 1，然后应用批标准化（Batch Normalization）对当前层次输出进行标准化，使得梯度下降变得容易。然后经过激活层（ReLU），最后使用 MaxPooling 做降采样，卷积核大小为 2，卷积核移动步长为 2。这一部分，输入加密流量高维向量进行降采样和学习空间特征，输出新的低维特征图向量。

第二部分为 LSTM 相关结构。每个 LSTM 单元的隐藏层维度设置为 256，由于为双向 LSTM，它将正反两个方向的输出做了连接，所以每个时间步的输出为 512 维。需要注意的是，这里采用了堆叠双向 LSTM 的结构，则除了最后一层外，中间 Bi-LSTM 需要保留每个时间步的输出。为了减轻过拟合的现象，在 Bi-LSTM 后增加 Dropout 层（最终输出神经元的激活值以一定概率停止工作，这个概率称为 dropout rate），设定 dropout rate 为 0.5。

第三部分为全连接层部分。使用全连接层输入 512 维度，输出 c 维（最终输出维度和类型数相同，在图 5 中最终示例输出为 16）。

最后使用 *softmax* 计算各个分类的预测概率。

因为先进行抽象空间特征学习再捕获双向时序特征，整体的参数量要比基于 CNN 或 LSTM 的网络参数量少很多。混合神经网络模型结合了 CNN 的速度和 RNN（循环神经网络，本方法使用了 LSTM，属于 RNN 的一种）的时间步敏感性，使得整体模型轻量化的同时，保留了两方面的优点。模型训练过程中 batch size（批次大小）设置为 128，使用 Adam optimizer（Adam 优化器）进行训练。可以使用 learning rate（学习率）调度技术，帮助模型更好地收敛。

利用训练样本对混合神经网络模型进行训练，其中包括 1D-CNN 网络、堆叠双向 LSTM 网络 and 全连接层网络三个部分，以得到最佳网络参数；

步骤 5：获取待识别的加密流量文件，使用步骤 1-步骤 3 处理待识别的加密流量文件，将得到的样本向量输入训练完成的混合神经网络模型中，模型输出加密流量的原始预测向量 $\mathbf{o}$ ， $\mathbf{o}$ 为实数向量。需要通过 *softmax* 处理 $\mathbf{o}$ ，得到各分类预测概率分布 $\hat{\mathbf{y}}$ 。 $\hat{\mathbf{y}}$ 是 c 维向量（ c 为流量类型数量），第 i 位置上的输出代表该

样本属于分类 i 的概率，其计算公式如公式（15）所示。通过预测分布向量 $\hat{\mathbf{y}}$ ，可以得到输入流量的最终分类标签 $label$ 。

$$label = \operatorname{argmax}(\hat{\mathbf{y}}) \quad (16)$$

其中 $\hat{\mathbf{y}}$ 为预测结果，它是 c 维的概率分布向量，通过 argmax 得到向量 $\hat{\mathbf{y}}$ 中最大的概率的下标（对应的分类）作为最终的分类标签 $label$ ， $label$ 代表某种类型的流量（例如流量服务类型，Chat, Email, Video 等，取决于不同分类任务，所有分类从 0 编号）。

本发明方法包含预处理阶段和分类阶段。预处理阶段对原始流量进行流切分，采样，向量化和标准化，并提出大流流中采样方案，解决大流流量（长时间数据流）的分类问题。分类阶段先使用 CNN 进行空间特征捕获和抽象特征抽取，然后在抽象特征的基础上使用堆叠双向 LSTM 学习流量时序特征，实现加密流量的自动特征提取和高效识别。该方法具有通用性，能够自动提取加密流量时空特征而无需专家手动特征设计，并且，它能够适应不同加密技术、混淆技术引起流量特征变化。

实施例 3

相应的，本发明还提供了一种网络加密流量识别装置，包括加密流量获取模块、预处理模块、分类预测模块和分类识别模块；其中：

加密流量获取模块，用于获取待识别的加密流量文件；

预处理模块，用于对待识别的加密流量进行预处理，所述预处理模块包括流切分单元、采集单元和向量化单元，其中：

流切分单元，用于将加密流量流切分为多个流；

采集单元，用于从每个流中采集多个连续数据包作为样本；

向量化单元，用于将每个样本进行向量化、标准化处理，得到格式化的样本向量集合；

分类预测模块，用于将预处理后得到的样本向量集合输入至预设训练的混合神经网络模型，得到预测向量，此预测向量中元素值代表加密流量属于各个分类的预测值；

所述混合神经网络模型包括：1D-CNN 网络、堆叠双向 LSTM 网络 and 全连接层网络；其中 1D-CNN 网络对输入样本向量集合进行空间特征学习，输出低维特征图；堆叠双向 LSTM 网络对输入的特征图进行时序特征学习，得到包含时序特征的特征图向量，全连接层根据输入的时序特征的特征图向量确定预测向量；

分类识别单元，用于基于预测向量计算得到各分类预测概率分布，取其中最大的概率对应的分类作为加密流量最终的分分类标签。

本实施例装置中各模块的具体实现，以及混合神经网络模型的构建、训练等内容，采取实施例 2 的实施方式。

进一步的，所述采集单元中，包括：

小流采样单元，用于采集流头部预设个数连续数据包组成一个样本，若已有数据包不足预设个数，则选择已有数据包，其余包补零处理；

大流采样单元，用于从流中选取若干个采样点，以每个采样点作为起点，采集连续预设个数连续数据包组成一个样本。

进一步的，所述大流采样单元中，采样点的选取方案包括：随机点采样，固定步长采样和突发点采样三种策略；其中：

所述随机点采样为流中随机点采样；所述固定步长采样以固定的步长从流量起始开始采样；所述突发点采样为寻找大流中的数据流突发点进行采样。

进一步的，所述向量化单元中将样本进行向量化、标准化处理，包括：

将每个数据包保留预设长度字节数，不足则用全零补全，反之则进行截断；将每个样本转化为的向量；

对向量中每个数据进行标准化处理。

进一步的，所述 1D-CNN 网络，包括：

1D-CNN 网络部分由两层 1D-CNN 卷积层组成，对输入加密流量向量进行两次卷积操作，并且在每一层中对卷积操作输出的新特征图进行批标准化、非线性激活和降采样处理。本发明装置有效利用加密流量数据流的时空特征，提出了一种基于流时空特征的新型加密流量混合神经网络识别模型，仅使用流的少量数据包，对流量进行准确的识别。

本领域内的技术人员应明白，本申请的实施例可提供为方法、系统、或计算机程序产品。因此，本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。

本申请是参照根据本申请实施例的方法、设备（系统）、和计算机程序产品的流程图和 / 或方框图来描述的。应理解可由计算机程序指令实现流程图和 / 或方框图中的每一流程和 / 或方框、以及流程图和 / 或方框图中的流程和 / 或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能。

这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能的步骤。

以上所述仅是本发明的优选实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明技术原理的前提下，还可以做出若干改进和变型，这些改进和变型也应视为本发明的保护范围。

权 利 要 求 书

1. 一种网络加密流量识别方法，其特征是，包括以下过程：

获取待识别的加密流量文件；

对待识别的加密流量进行预处理，所述预处理包括：将加密流量流切分为多个流；然后从每个流中采集多个连续数据包作为样本；最后将每个样本进行向量化、标准化处理，得到格式化的样本向量集合；

将预处理后得到的样本向量集合输入至预设训练的混合神经网络模型，得到预测向量，此预测向量中元素值代表加密流量属于各个分类的预测值；

所述混合神经网络模型包括：1D-CNN 网络、堆叠双向 LSTM 网络 and 全连接层网络；其中 1D-CNN 网络对输入样本向量集合进行空间特征学习，输出低维特征图；堆叠双向 LSTM 网络对输入的特征图进行时序特征学习，得到包含时序特征的特征图向量，全连接层根据输入的时序特征的特征图向量确定预测向量；

基于预测向量计算得到各分类预测概率分布，取其中最大的概率对应的分类作为加密流量最终的分类标签。

2. 根据权利要求 1 所述的一种网络加密流量识别方法，其特征是，所述从流中采集多个连续数据包作为样本，包括：

若流为小流，采集流头部预设个数连续数据包组成一个样本，若已有数据包不足预设个数，则选择已有数据包，其余包补零处理；

若流为大流，从流中选取若干个采样点，以每个采样点作为起点，采集连续预设个数连续数据包组成一个样本。

3. 根据权利要求 2 所述的一种网络加密流量识别方法，其特征是，所述采样点的选取方案包括：随机点采样，固定步长采样和突发点采样三种策略；其

中：

所述随机点采样为流中随机点采样；所述固定步长采样以固定的步长从流量起始开始采样；所述突发点采样为寻找大流中的数据流突发点进行采样。

4. 根据权利要求 1 所述的一种网络加密流量识别方法，其特征是，所述将样本进行向量化、标准化处理，包括：

将每个数据包保留预设长度字节数，不足则用全零补全，反之则进行截断；将每个样本转化为的向量；

对向量中每个数据进行标准化处理。

5. 根据权利要求 1 所述的一种网络加密流量识别方法，其特征是，所述 1D-CNN 网络，包括：

1D-CNN 网络由两层 1D-CNN 卷积层组成，对输入加密流量样本向量进行两次卷积操作，并且在每一层中对卷积操作输出的新特征图进行批标准化、非线性激活和降采样处理。

6. 根据权利要求 1 所述的一种网络加密流量识别方法，其特征是，所述混合神经网络模型的训练包括：

获取多个加密流量文件，对每个加密流量文件标注出对应的分类标签；

对各个加密流量文件进行预处理，所述预处理包括：将每个加密流量流切分为多个流；然后从每个流中采集多个连续数据包作为样本；最后将每个样本进行向量化、标准化处理，得到格式化的样本向量集合作为训练样本；

利用训练样本对混合神经网络模型进行训练，其中包括 1D-CNN 网络、堆叠双向 LSTM 网络 and 全连接层网络三个部分，以得到最佳网络参数；

得到训练完成的混合神经网络模型。

7. 根据权利要求 6 所述的一种网络加密流量识别方法，其特征是，所述 1D-CNN 网络的训练包括：

在向量化过程中保留每个样本中数据包的时序维度, 设 $\mathbf{x}^{(t)} \in \mathbb{R}^L$ 为一个样本中第 t 个数据包, t 为 0 到 n 的任一整数, 代指向量中任一数据包, 它是一个 L 维度的向量;

$$\mathbf{x} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(M)}] \quad (5)$$

$\mathbf{x}$ 代表一个样本, 它包含 M 个数据包的向量, 在 1D-CNN 中, $\mathbf{x}$ 视为包含 M 个通道, 每个通道都是 L 维的二维向量; 假设 $\mathbf{x}_{i:i+j}$ 代表全部通道从任意位置 i 到 $i+j$ 的字节; 在 $\mathbf{x}$ 上一维卷积操作如下:

$$\mathbf{a}_i^{(t)} = f(\mathbf{w} \cdot \mathbf{x}_{i:i+h-1} + b) \quad (6)$$

通常一个卷积层中包含多个卷积核, 每个 Filter 操作相同, 生成新特征图的一个通道; 以其中任一卷积核 t 为例, $\mathbf{w} \in \mathbb{R}^{hk}$ 为在 $\mathbf{x}$ 上滑动窗口, b 为偏移值, f 则是非线性的激活函数; $\mathbf{a}_i^{(t)}$ 为任一卷积核 t 生成的特征;

当前 Filter 在 $\mathbf{x}$ 上滑动时, 该 Filter 的卷积操作应用到窗口内的字节上, 序列 $\{\mathbf{x}_{1:h}, \mathbf{x}_{2:h+1}, \dots, \mathbf{x}_{n-h+1:n}\}$ 将会生新特征图; 所有 Filter 操作相同, 每个 Filter 对应的参数 $\mathbf{w}$ 和 b 是不同的;

$$\mathbf{a}^{(t)} = [\mathbf{a}_1^{(t)}, \mathbf{a}_2^{(t)}, \dots, \mathbf{a}_{n-h+1}^{(t)}] \quad (7)$$

这里 $\mathbf{a}^{(t)} \in \mathbb{R}^{n-h+1}$, 代表任一卷积核 t 生成的新特征图, 也可视为输出通道 t ; 对于每个通道的新特征图, 还使用池化操作层对特征图进行降采样; 池化操作层的操作同样使用 Filter 做滑动操作, 但在每个 Filter 上通常执行的运算为 $\hat{\mathbf{a}}_i^{(t)} = \max(\mathbf{a}_{i:i+h-1}^{(t)})$, 保留每个滑动窗口中的最大值。

8. 根据权利要求 7 所述的一种网络加密流量识别方法, 其特征是, 所述堆叠双向 LSTM 网络的训练包括:

在 LSTM 中, 将输入特征图的多个通道视为多个时间步; 在每个时间步上有以下公式:

$$\tilde{\mathbf{c}}^{(t)} = \tanh(\mathbf{w}_c [\mathbf{h}^{(t-1)}, \mathbf{a}^{(t)}] + b_c) \quad (8)$$

其中 $\mathbf{a}^{(t)} \in \mathbb{R}^m$, 表示在输入特征图任一时间步 t 上的向量, 其维度与每个

时间步输入的特征图维度相同； $\mathbf{h}^{(t-1)} \in \mathbb{R}^s$ ，上一个时间步隐藏层的输出，维度由 LSTM 单元的隐藏层维度的参数决定； $\tilde{\mathbf{c}}^{(t)}$ 为当前层的中间输出； $\mathbf{w}_c$ 和 b_c 分别为参数矩阵和偏差；

在 LSTM 中，最终输出由三个门决定，分别为更新门 Γ_u 、遗忘门 Γ_f 和输出门 Γ_o ；门的值计算公式如下：

$$\Gamma_u = \sigma(\mathbf{w}_u[\mathbf{h}^{(t-1)}, \mathbf{a}^{(t)}] + b_u) \quad (9)$$

$$\Gamma_f = \sigma(\mathbf{w}_f[\mathbf{h}^{(t-1)}, \mathbf{a}^{(t)}] + b_f) \quad (10)$$

$$\Gamma_o = \sigma(\mathbf{w}_o[\mathbf{h}^{(t-1)}, \mathbf{a}^{(t)}] + b_o) \quad (11)$$

其中， σ 为非线性激活函数， $\mathbf{w}_u$ 、 $\mathbf{w}_f$ 、 $\mathbf{w}_o$ 和 b_u 、 b_f 、 b_o 分别为三个门对应的参数矩阵和偏差值；

三个门的值的计算方法，与 $\tilde{\mathbf{c}}^{(t)}$ 计算方法类似，它们取值由当前时间步的输入 $\mathbf{a}^{(t)}$ 和上一层隐藏层的输出 $\mathbf{h}^{(t-1)}$ 决定；更新门 Γ_u 、遗忘门 Γ_f 和输出门 Γ_o 的作用用于控制当前 LSTM 单元是否更新当前信息，是否遗忘过去信息，是否输出最终信息；三个开关产生最终输出的公式如下，其中 $\mathbf{c}^{(t)}$ 当前层的中间输出向量：

$$\mathbf{c}^{(t)} = \Gamma_u \odot \tilde{\mathbf{c}}^{(t)} + \Gamma_f \odot \mathbf{c}^{(t-1)} \quad (12)$$

$$\mathbf{h}^{(t)} = \Gamma_o \odot \tanh(\mathbf{c}^{(t)}) \quad (13)$$

由三个门分别对过去、当前和总输出做取舍；最终输出当前时间步的隐藏层输出 $\mathbf{h}^{(t)}$ ；其中 $\odot$ 为哈达玛积，表示向量之间按位相乘；

堆叠双向 LSTM 的输出将当前时间步不同方向的输出做连接，如：

$$\mathbf{h}^{(t)} = [\vec{\mathbf{h}}^{(t)}, \overleftarrow{\mathbf{h}}^{(t)}] \quad (14)$$

在每个时间步上，隐藏层的输出 $\mathbf{h}^{(t)}$ 都由时间步上正向的输出 $\vec{\mathbf{h}}^{(t)}$ 和反向的输出 $\overleftarrow{\mathbf{h}}^{(t)}$ 连接而成。

9. 根据权利要求 1 所述的一种网络加密流量识别方法，其特征是，所述基于预测向量计算得到各分类预测概率分布，包括：

计算预测概率分布，其计算公式如下：

$$\text{softmax}(\mathbf{o})_i = \hat{\mathbf{y}}_i = \frac{e^{o_i}}{\sum_j e^{o_j}} \quad (15)$$

其中 $\mathbf{o}$ 为混合神经网络模型原始输出向量， $\hat{\mathbf{y}}$ 则为经过 softmax 处理后的预测概率向量， $\mathbf{o}_i$ 为向量 $\mathbf{o}$ 中第 i 位置的向量值。

10. 一种网络加密流量识别装置，其特征是，包括加密流量获取模块、预处理模块、分类预测模块和分类识别模块；其中：

加密流量获取模块，用于获取待识别的加密流量文件；

预处理模块，用于对待识别的加密流量进行预处理，所述预处理模块包括流切分单元、采集单元和向量化单元，其中：

流切分单元，用于将加密流量流切分为多个流；

采集单元，用于从每个流中采集多个连续数据包作为样本；

向量化单元，用于将每个样本进行向量化、标准化处理，得到格式化的样本向量集合；

分类预测模块，用于将预处理后得到的样本向量集合输入至预设训练的混合神经网络模型，得到预测向量，此预测向量中元素值代表加密流量属于各个分类的预测值；

所述混合神经网络模型包括：1D-CNN 网络、堆叠双向 LSTM 网络 and 全连接层网络；其中 1D-CNN 网络对输入样本向量集合进行空间特征学习，输出低维特征图；堆叠双向 LSTM 网络对输入的特征图进行时序特征学习，得到包含时序特征的特征图向量，全连接层根据输入的时序特征的特征图向量确定预测向量；

分类识别单元，用于基于预测向量计算得到各分类预测概率分布，取其中最大的概率对应的分类作为加密流量最终的分类型标签。

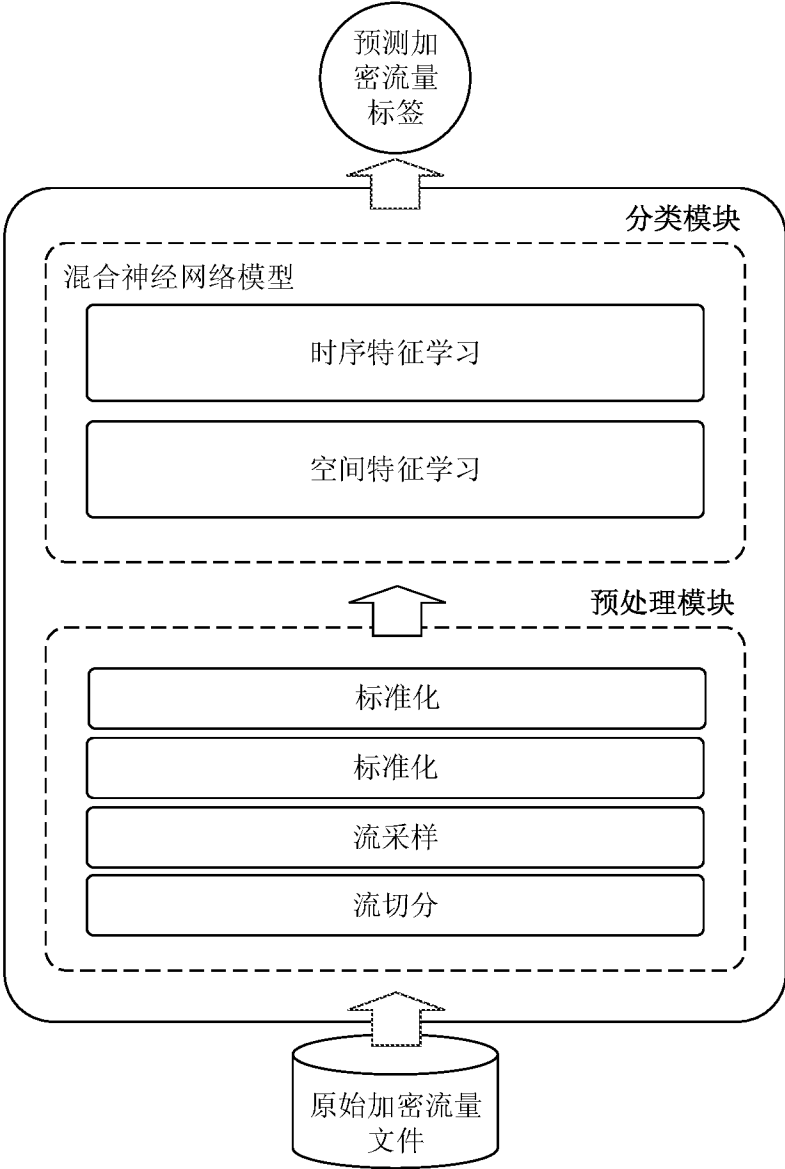


图 1

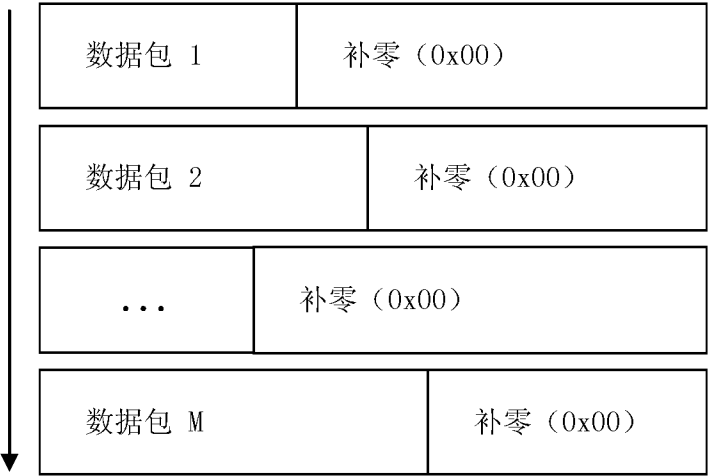
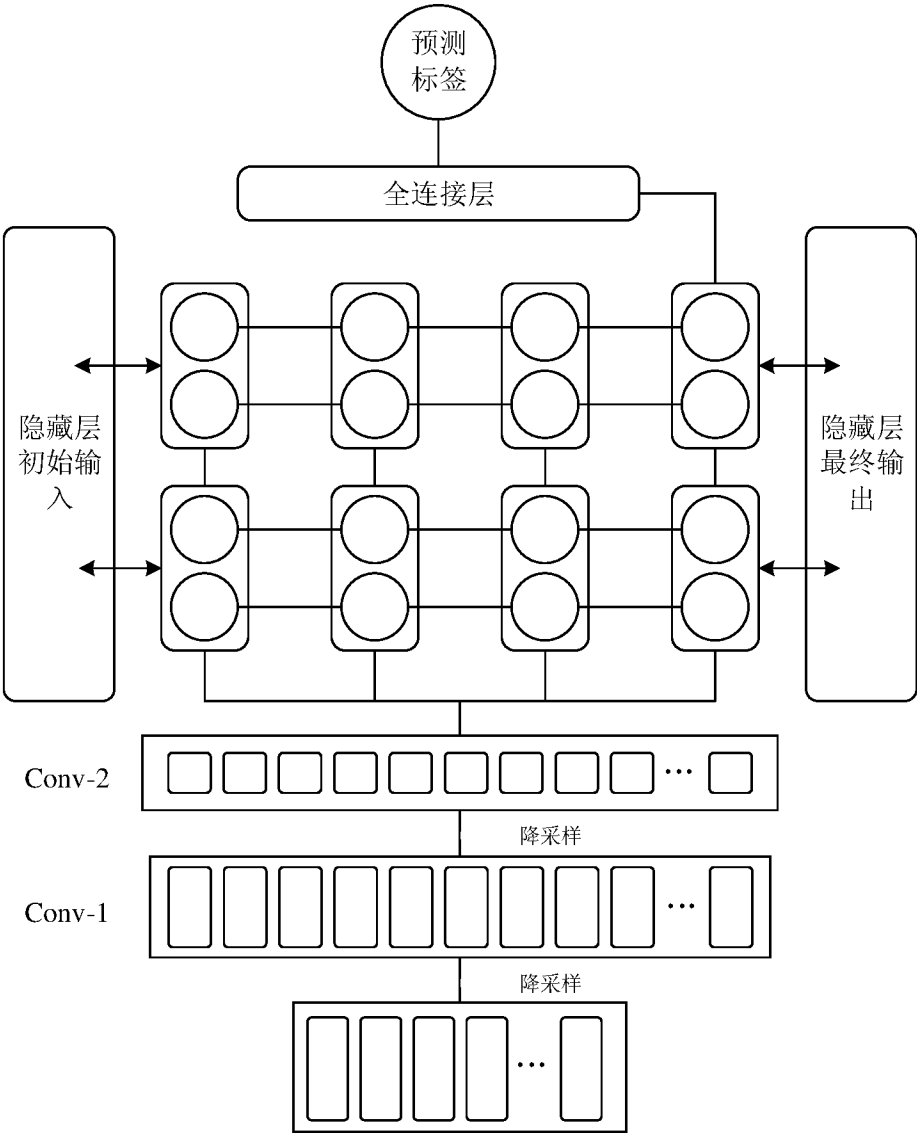


图 2



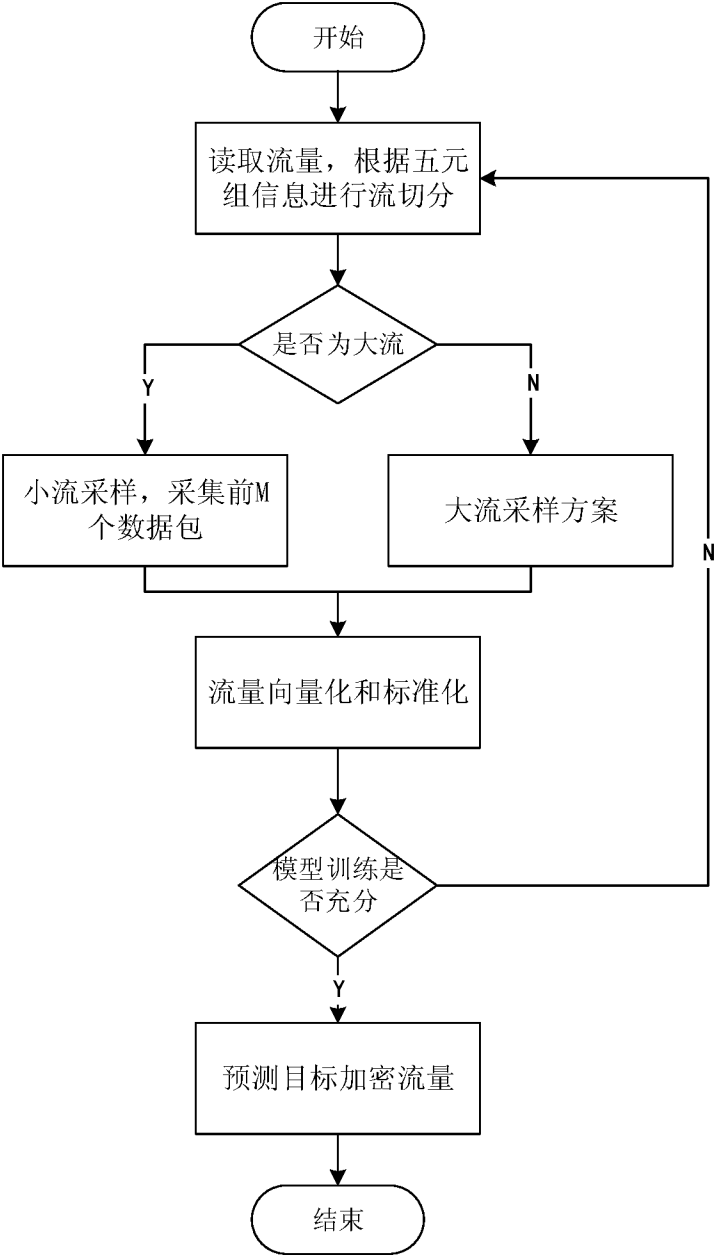


图 4

层次	层次名称	输入	卷积核	步长	输出	参数
Conv-1	Conv1d-1	(15, 1500)	3	1	(64, 1500)	2, 944
	BatchNorm1d-2	(64, 1500)	-	-	(64, 1500)	128
	ReLU-3	(64, 1500)	-	-	(64, 1500)	-
	MaxPool1d-4	(64, 1500)	2	2	(64, 750)	-
Conv-2	Conv1d-5	(64, 750)	3	1	(64, 750)	12, 352
	BatchNorm1d-6	(64, 750)	-	-	(64, 750)	128
	ReLU-7	(64, 750)	-	-	(64, 750)	-
	MaxPool1d-8	(64, 750)	2	2	(64, 325)	-
Bi-LSTM	Bi-Lstm-9	(64, 325)	-	-	(64, 512)	2, 873, 344
	Dropout-10	(64, 512)	-	-	(64, 512)	
	Bi-Lstm-11	(64, 512)	-	-	(512,)	
	Dropout-12	(512,)	-	-	(512,)	
FC	Linear-13	(512,)	-	-	(16,)	8, 208

图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/118725

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/62(2006.01)i; G06N 3/04(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06K; G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNKI, CNPAT, WPI, EPODOC, IEEE: 加密流量, 预处理, 切分, 划分, 分类, 样本, 向量, 神经网络, 卷积神经网络, 长短期记忆单元, 长短期记忆单元, 空间, 时间, 时序, 特征, 预测, 概率, 最大, encrypted traffic, preprocessing, pre-processing, segment, divide, classification, samples, vector, neural network, convolutional neural network, CNN, Long Short-Term Memory, LSTM, space, time, feature, predict, probability, maximum

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 109905696 A (ZHEJIANG UNIVERSITY) 18 June 2019 (2019-06-18) description, paragraphs [0067]-[0111], and figures 1-7	1-10
X	黄璇丽 等 (HUANG, Xuanli et al.). "基于深度学习的网络流时空特征自动提取方法 (A Deep Learning-Based Spatio-Temporal Features Extraction Method for Network Flow)" 集成技术 (Journal of Integration Technology), Vol. 9, No. 2, 31 March 2020 (2020-03-31), sections 2-4	1-10
X	CN 110896381 A (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 20 March 2020 (2020-03-20) description, paragraphs [0048]-[0111], and figures 1-3	1-10
A	WO 2020029832 A1 (HUAWEI TECHNOLOGIES CO., LTD.) 13 February 2020 (2020-02-13) entire document	1-10
A	CN 110502753 A (KUNMING UNIVERSITY OF SCIENCE AND TECHNOLOGY) 26 November 2019 (2019-11-26) entire document	1-10



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance
 "E" earlier application or patent but published on or after the international filing date
 "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
 "O" document referring to an oral disclosure, use, exhibition or other means
 "P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
 "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
 "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
 "&" document member of the same patent family

Date of the actual completion of the international search

17 May 2021

Date of mailing of the international search report

26 May 2021

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/118725

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 108710829 A (BEIJING HONGYUN ZHISHENG TECHNOLOGY CO., LTD.) 26 October 2018 (2018-10-26) entire document	1-10

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2020/118725

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)		Publication date (day/month/year)
CN	109905696	A	18 June 2019	None		
CN	110896381	A	20 March 2020	None		
WO	2020029832	A1	13 February 2020	US	2020053155	A1 13 February 2020
CN	110502753	A	26 November 2019	None		
CN	108710829	A	26 October 2018	None		

国际检索报告

国际申请号

PCT/CN2020/118725

A. 主题的分类

G06K 9/62(2006.01)i; G06N 3/04(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06K; G06N

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNKI, CNPAT, WPI, EPDOC, IEEE: 加密流量, 预处理, 切分, 划分, 分类, 样本, 向量, 神经网络, 卷积神经网络, 长短期记忆单元, 长短期记忆单元, 空间, 时间, 时序, 特征, 预测, 概率, 最大, encrypted traffic, pre-processing, pre-processing, segment, divide, classification, samples, vector, neural network, convolutional neural network, CNN, Long Short-Term Memory, LSTM, space, time, feature, predict, probability, maximum

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	CN 109905696 A (浙江大学) 2019年 6月 18日 (2019 - 06 - 18) 说明书第[0067]-[0111]段, 图1-7	1-10
X	黄璇丽等. “基于深度学习的网络流时空特征自动提取方法” 集成技术, 第9卷, 第2期, 2020年 3月 31日 (2020 - 03 - 31), 第2-4节	1-10
X	CN 110896381 A (中国科学院深圳先进技术研究院) 2020年 3月 20日 (2020 - 03 - 20) 说明书第[0048]-[0111]段, 图1-3	1-10
A	WO 2020029832 A1 (HUAWEI TECHNOLOGIES CO., LTD.) 2020年 2月 13日 (2020 - 02 - 13) 全文	1-10
A	CN 110502753 A (昆明理工大学) 2019年 11月 26日 (2019 - 11 - 26) 全文	1-10
A	CN 108710829 A (北京红云智胜科技有限公司) 2018年 10月 26日 (2018 - 10 - 26) 全文	1-10

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2021年 5月 17日

国际检索报告邮寄日期

2021年 5月 26日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

刘丽

电话号码 86-(10)-53961799

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2020/118725

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	109905696	A	2019年 6月 18日	无	
CN	110896381	A	2020年 3月 20日	无	
WO	2020029832	A1	2020年 2月 13日	US 2020053155 A1	2020年 2月 13日
CN	110502753	A	2019年 11月 26日	无	
CN	108710829	A	2018年 10月 26日	无	