研究与开发

# 基于图神经网络的鲁棒加密流量识别

李孟想[1]，彭闯[1]，王浩[1]，黄超明[1,2]，谭小彬[1,2]
（1. 中国科学技术大学，安徽 合肥 230026；
2. 合肥综合性国家科学中心人工智能研究院，安徽 合肥 230088）

**摘　要：**当前的网络流量识别方法一般针对特定网络环境或数据集进行设计和测试，难以推广应用于复杂多变的实际网络环境。提出了一种基于图神经网络的鲁棒流量识别算法，用于在实际网络场景中实现准确的流量识别。首先，当前算法忽视网络环境波动，导致流行为模式发生改变，准确率下降，通过选择网络流中的高层协议特征对网络流进行聚类和筛选，以减小网络带宽波动对网站访问流量行为的影响。其次，由于当前算法大多只进行单流识别，忽视流间的相互关系，考虑网络流的多种类型特征信息及其相关性，并通过图神经网络提取网络流之间的时空相关特征，充分学习网络流量特征，通过多个流和多种特征的互补关系以提高算法的鲁棒性。最后，使用可以捕获数据全局信息的Transformer模型作为分类器对网络数据流的多类型特征进行分析，实现鲁棒网络流量识别。在不同网络环境下分别采集了对21个目标网站的共大约1 500次和1 400次的访问数据作为数据集进行训练测试，实现了90.7%的准确率，对比最新的ProGraph算法，准确率提高了7.3%，实验结果验证了所提方法的有效性。

**关键词：**流量识别；图神经网络；Transformer
**中图分类号：**TP393
**文献标志码：**A
**doi**: 10.11959/j.issn.1000−0801.2024156

# A robust encrypted traffic identification scheme based on graph neural network

LI Mengxiang[1], PENG Chuang[1], WANG Hao[1], HUANG Chaoming[1,2], TAN Xiaobin[1,2]
1. University of Science and Technology of China, Hefei 230026, China
2. Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230088, China

**Abstract:** The current methods for identifying network traffic are generally designed and tested for specific network environments or datasets, making it difficult to generalize and apply to complex and ever-changing actual network environments. A robust traffic recognition algorithm based on graph neural networks was proposed for achieving accu-

**收稿日期：**2024−03−01；**修回日期：**2024−05−12
**基金项目：**国家重点研发计划项目（No.2022YFB2901400）；安徽省科技重大专项（No.202103a05020007）；未来网络试验设施（CENI）资助项目（No.2016-000052-73-01-000515）；国家自然科学基金资助项目（No.62101525，No.62341113，No.62021001）
**Foundation Items:** The National Key Research & Development Program of China (No.2022YFB2901400), The Key Science and Technology Project of Anhui (No.202103a05020007), The China Environment for Network Innovations (No.2016-000052-73-01-000515), The National Natural Science Foundation of China (No.62101525, No.62341113, No.62021001)

rate traffic recognition in practical network scenarios. Firstly, in response to the current algorithm's neglect of network environment fluctuations and the decrease in accuracy caused by pattern changes, network flows were clustered and filtered by selecting high-level protocol features to reduce the impact of network bandwidth fluctuations on website access traffic behavior. Secondly, due to the fact that most current algorithms only perform single stream recognition and ignore the interrelationships between flows, the various types of feature information and their correlations of network flows were considered, and spatiotemporal correlation features between network flows were extracted through graph neural networks to fully learn network traffic characteristics. By complementing multiple flows and features, the robustness of the algorithm was improved. Finally, a Transformer model that could capture global data information was used as a classifier to analyze the multi type features of network data flow, achieving robust network traffic recognition. Approximately 1 500 and 1 400 visits to 21 target websites in different network environments were collected as datasets for training and testing, achieving an accuracy of 90.7%. Compared with the latest ProGraph algorithm, the accuracy is improved by 7.3%, and the experimental results verify the effectiveness of the proposed method.

**Key words:** traffic identification, graph neural network, Transformer

## 0 引言

互联网在人们的生活中占据着越来越重要的地位，网络流量也呈爆发式增长。流量识别是网络行为分析、网站规划建设、网络异常检测和网络流量模型研究的重点，也是提升网络管理水平、改善服务质量和检查应用安全的前提和基础。流量识别可以帮助网络管理员审计用户行为，指导下一阶段的网络建设和网络优化。然而，随着人们对于隐私保护的重视，越来越多的网站开始采用加密协议传输流量。加密流量的爆发式增长，给网络管理带来了新的挑战。一方面，从加密流量数据中可读取的有用信息变少，使得一些流量识别技术识别准确率降低甚至难以使用；另一方面，攻击者可以通过将攻击信息伪装成加密流量的手段发起攻击，这使加密流量识别成为保护目标免受攻击的重要手段之一。

流量识别技术一直在随着互联网技术和人工智能技术的发展而不断发展。在互联网早期，由于网络协议都比较简单，每种应用往往都对应着固定的端口号且访问网站的流量没有加密，就出现了基于端口和基于深度包检测的识别方法[1-2]。然而动态端口号和流量加密技术的广泛应用，使得这些技术的准确率难以满足实际需要。随着机器学习技术在其他领域取得巨大成功，机器学习技术也被应用于流量识别领域[3-8]。它们大多从原始流量中提取人工抽取的特征，然后放入机器学习模型中以得到分类结果。然而这些方法的识别准确率很大程度上依赖于人工抽取的特征，需要大量专家知识。为了减少对专家知识的依赖，深度学习技术被应用到流量识别中，以实现端到端的分类[9-10]。它们将原始的IP数据包直接放入分类器进行训练，让分类器自己提取特征。目前已经取得了不少成果。但是，目前大多数工作并没有考虑实际场景中复杂多变的网络环境对访问网站流产生的影响，这就导致它们在实际应用中的效果可能远远不如预期。Li等[11]提出了一种新的图传播算法，并且使用在不同网络环境下采集的数据集分别作为训练集和测试集验证了其算法的鲁棒性，然而其忽视了流特征的变化以及流与流之间的互补关系，导致其效果并不理想。

对此，分析了不同网络状态下访问网站流量的差异，发现网络环境的变化会对访问网站流的数量和统计特征产生很大的影响。然而，目前的大多数算法并没有针对这些变化采取优化措施。通过对流聚类并筛选，通过更高层的协议来计算

统计特征以减少网络波动的影响。同时，不同的流存在着不同类型的特征信息[12]，并且同一个网站流与流之间存在时空互补关系。对于同一类型的特征结合流与流之间的时空关系，通过图神经网络（graph neural network，GNN）提取特征表示，对不同类型特征提取到的特征表示进行融合，并且使用可以捕获全局信息的Transformer模型进行分类，从而充分学习访问同一网站的不同流与流不同类型的特征以及流间的相互关联关系，提高算法的鲁棒性。

综上所述，本文的主要贡献如下。

（1）分析了不同网络状态下访问网站流量的差异，发现随着网络环境的变化，流的数目和统计特征都会发生变化。对此选择网络流中的高层协议特征对网络流进行聚类和筛选，从而减少网络环境波动对于访问网站的流的数目和特征的干扰，以提高流量识别算法的鲁棒性。

（2）设计并实现了基于图神经网络的鲁棒流量识别算法，采用图神经网络提取网络流之间的时空相关特征，使用可以从全局捕获信息的Transformer模型作为分类器，减少了网络环境波动带来的特征变化的影响，实现鲁棒网络流量识别。

（3）搭建了实际的测试平台，采集了不用网络状态下访问网络产生的网络流量作为训练集和测试集，通过实验测试，提出的方法可以实现90.7%的准确率，优于现有算法。

## 1　研究现状

由于在互联网发展的早期，应用与其使用的TCP或者UDP的端口号是一一对应的，就出现了通过检测抓取到的数据包的端口号来检测应用的方法。并且由于这个时期访问网站的流量未加密，可以通过检查数据包包头和有效载荷信息并与之前建立好的指纹库中的指纹进行比对来实现对网站指纹的识别。但是流量加密协议和动态端口号协议的广泛应用，使得这两种技术在现在日益复杂的网络环境中都难以应用。

由于机器学习的方法在其他领域取得的巨大成功，许多研究者将机器学习的方法应用到流量识别中，并取得了不错的效果。起初，采取从原始流量中提取人工抽取的特征，然后放入机器学习模型中以得到分类结果。例如，Sun等[3]使用模式匹配的方法过滤传输层安全性（transport layer security，TLS）协议流量，然后使用朴素贝叶斯方法对TLS流量进行分类，得到HTTPS和洋葱头（the onion router，Tor）流量的分类精度在93%到96%之间。Ding等[4]使用机器学习算法C4.5决策树来过滤HTTPS流量，然后使用随机森林对流量数据进行处理和分类。然而这些算法依赖于人工抽取的特征，当这些人工提取的特征受到网络环境的波动影响较大时，算法的准确率也会降低。

与传统的机器学习技术相比，深度学习可以直接从原始数据中提取特征，从而实现端到端的分类，并且深度学习技术的表示能力更强。因此深度学习技术也被应用到网站指纹识别中。Lopez-martin等[9]使用循环神经网络（recurrent neural network，RNN）和卷积神经网络（convolutional neural network，CNN）结合的方法来识别物联网的流量。Zou等[10]提出了一种将卷积网络和循环网络相结合的新型网络，在网站指纹识别上取得了不错的效果。Song等[13]提出使用文本卷积网络来提取重要的特征用于流量分类，在ISCX VPN-non VPN数据集上取得了良好的效果。Lu等[14]提出了一种名为GAP-WF的框架用于细粒度网站指纹识别方法，利用基于GNN的模型来学习流内和流间的特征。Liu等[15]设计了一个端到端的分类模型，采用多层编码器和解码器结构，可以深入挖掘流的潜在序列特征，并且引入了重构机制，提高了特征的有效性。Van等[16]引入了一种半监督方法，自动查找网络流量中与目

的地相关的特征之间的时间相关性，并使用这些相关性来生成应用指纹以用于识别已知应用或检测以前未见过的应用。Huoh 等[17]以一个数据包为一个节点，以数据包中的所有数据为特征，根据数据包的到达时间建立边，并引入流的统计特征作为整个图的图特征，使用图卷积网络（graph convolutional network，GCN）来对流进行分类。这些方法直接从原始数据包中提取特征，然而当网络环境发生改变时，原始数据包的一些信息就会发生较大的改变。例如，数据包的到达时间会随着网络时延的改变而改变，数据包的长度信息会由于网络环境变化导致的包头的控制信息的增添而改变，这会导致从原始数据包提取的特征改变，从而导致识别的准确率降低。

关于抗网络环境干扰的流量识别算法的研究却十分稀少。Li 等[11]提出了一种新的图传播算法，实现了流量识别算法的跨网络环境的鲁棒性。然而此算法并没有考虑访问网站的多个流之间的互补关系。并且对于访问网页产生的流随网络环境波动而产生的改变并没有做针对性的处理，比如流的个数会随着网络环境的波动而改变，导致其在不同网络环境下的效果并不是十分理想。

## 2 基于图神经网络的鲁棒性加密流量识别算法

首先，分析了访问网站产生的流量随着网络环境的波动而产生的变化，对原因做出了简要分析并对这些变化提出了针对性的处理措施，据此设计了一种基于图神经网络的鲁棒性流量识别算法。

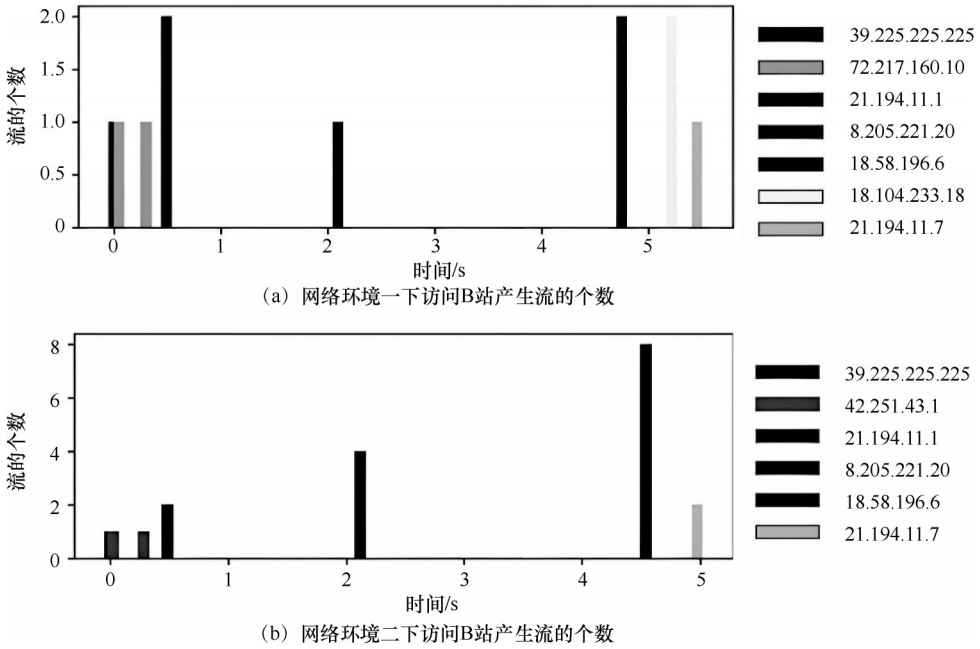
### 2.1 研究思路

一个网站在加载时，为了缩短加载的时间，往往会产生多个数据流，同时加载网页不同部分的内容，特定的网页会有特定的加载顺序。加载一个网页时会产生不止一条流，不同流的IP地址可能不一样，所以难以根据IP地址利用DNS服务器识别数据流。并且一些大的网站为了减少地理因素对于网站加载速度的影响，往往将网络资源托管到CDN运营商的服务器上，这就会导致在不同的位置访问同一个网站的IP地址不同。并且当不同网站的管理者将网站的资源托管到同一个CDN运营商时，就有可能出现访问不同的网页产生相同的IP地址。所以仅仅根据IP地址识别数据流是不准确的。

当网络环境改变时，为了提高加载速度，网站的加载模式也会有所改变，例如，可能将一个媒体文件在多个流进行加载，以减少媒体文件重传的时间。并且由于网络的波动，数据包的到达时间和到达顺序也会由于丢包和时延等因素发生改变，导致与此相关的统计特征的改变。

两种网络环境下访问站B站产生流的个数如图1所示。其中，图1（a）表示在网络环境一的条件下，即没有添加额外扰动的情况下访问B站主页的流量情况，图1（b）表示在网络环境二的情况下，即将网络环境的丢包率设置为5%、时延为(200±50) ms的情况下产生的流量情况。从图1可以看出，随着网络环境变差，流之间的时间间隔增加，但是相对顺序关系没有改变，并且有的服务器发出的流的个数增加了一倍。这说明随着网络环境变差，为了减少媒体文件重传的时间，网站将对同一个媒体文件通过多个流进行加载，以确保媒体文件不会丢失。如果选取所有的流进行识别，可能导致这些新增的流对识别的干扰。由于多个流可能传递的都是媒体文件，功能具有相似性，在识别过程中并不需要使用访问网站过程中产生的所有流，可以根据IP地址所属服务商对流进行分组，然后选取每个流组中数据包最多的$n$个流即可。这样既可以加快分类器训练和识别的速度，也可以减少干扰。

为了验证网络环境波动对于流个数的影响，分别统计了在不增加额外干扰和额外增加




（a）网络环境一下访问B站产生流的个数

（b）网络环境二下访问B站产生流的个数

图1　两种网络环境下访问站B站产生流的个数

(200±50) ms时延、5%丢包率的额外扰动这两种网络环境下，访问10个目标网站的20次访问产生的流的个数，两种网络环境下访问网站主页产生的网络数据流的数量见表1，从表1中可知，随着网络环境的变化，访问这些网站的流的个数都会增加，而且绝大多数增加10个流以上，有的甚至增加30多个流（如虎牙直播），由此可知：如果不对这些流进行筛选，将会影响分类器的分类速度和识别准确率。

**表1　两种网络环境下访问网站主页产生的网络数据流的数量**

| 网站名称 | delay=200 ms loss=5% | delay=0 loss=0 |
|---|---|---|
| B站 | 28.15 | 14.1 |
| 淘宝 | 33.15 | 15.7 |
| 爱奇艺 | 79.05 | 42.5 |
| 豆瓣 | 25.9 | 21.05 |
| 京东 | 50.7 | 29.75 |
| 喜马拉雅 | 25.25 | 13.55 |
| 拼多多 | 22.35 | 12.6 |
| 虎牙直播 | 69.4 | 35.95 |
| 快手 | 62.3 | 33.6 |
| 酷狗音乐 | 26.4 | 13.55 |

同时，对两次访问的统计特征的平均值做了统计，两种网络环境下访问网址产生的流的部分统计特征变化见表2，其中一个是在增加均值为200 ms、方差为50 ms的时延（delay），丢包率（loss）为5%的情况下访问B站产生的流的平均统计特征（20次），另一个是在不增加时延和丢包率的情况下访问B站产生的流的平均统计特征（20次），可以看到由于数据包的重传和流的增加，流的部分统计特征发生了改变。这是由于随着网络环境的改变，为了保证网络服务质量（quality of service，QoS），会发送一些用于流量控制的数据包，或高层的协议（如TCP）会在其包头的可变部分插入一些控制信息，导致对以IP层数据为基础计算的统计特征发生变化。为了减少网络环境对于统计特征的影响，可以使用以高层协议的数据为基础计算的统计特征。

通过以上分析，可以看到访问网站产生的流的数量和统计特征都会受网络环境变化的影响。通过对流聚类并筛选，通过更高层的协议来计算统计特征以减少网络波动的影响。同时，对于每种类型的特征通过不同流之间的时空相关关系建

立图，使用图神经网络提取特征的综合表征，对不同类型特征提取到的特征表示进行融合，并且使用可以捕获全局信息的Transformer模型进行分类，从而充分学习访问同一网站流的不同类型特征以及流间的相互关联关系，从而提高算法的鲁棒性。

表2　两种网络环境下访问网址产生的流的部分统计特征变化

| 对比项 | delay=200 ms, loss=5% | delay=0, loss=0 |
|---|---|---|
| len2 | –18.748 | –11.831 |
| alen | 38.883 | 55.648 |
| dlen | 21.901 | 29.467 |
| uper8 | 125.023 | 162.865 |
| len11 | –302.106 | –129.680 |

## 2.2　整体架构

算法整体架构如图2所示，总体主要包括特征的提取与选择、图的构建、图特征提取、特征的融合与分类4个模块。首先对根据五元组（即源IP地址、目的IP地址、源端口号、目的端口号、传输层协议）划分成流的数据流提取根据高层协议计算的统计特征（这里使用TCP协议层的数据计算统计特征）和端口号等非统计特征。然后对流进行筛选，根据时空关联关系对流进行建图，并且使用图卷积网络对建立的图进行综合表征，即图特征提取模块。最后，对提取到的各种特征的特征向量进行拼接。由于Transformer模型并不能利用特征向量的位置信息，将拼接后的特征向量增加位置编码后送入使用注意力机制的Transformer模型进行识别分类，以便从输入中捕获全局信息。接下来将详细介绍各个模块的功能与组成。

### 2.2.1　特征的提取与选择

首先对于PCAP（packet capture）文件中的数据包，根据五元组分成流，对TCP数据包根据序号对数据包进行重排。

一般关于网站指纹识别的方法是使用IP数据包的包长、到达时间等来计算数据包的统计特征。但是TCP/UDP以IP层数据包的载荷长度，会包含不确定长度的控制信息，难以反映应用层的数据信息，难抗网络环境的干扰。所以使用TCP/UDP的载荷长度来计算统计特征。选取平均TCP/UDP载荷长度、前20个TCP/UDP载荷长度、流持续时间等统计特征并分别表示为：$tf_1$, $tf_2$, …, $tf_n$。将这些特征连接在一起作为流的统计特征，则流的统计特征（statistical feature）**SF**可以表示为：

$$\mathbf{SF}=(tf_1, tf_2, \cdots, tf_n), \tag{1}$$

同时，提取流的五元组、流建立过程中使用的数字证书的编号以及流中各个数据包中泄露的URL等非统计特征。

由于流的非统计特征并不能直接在机器学习算法中使用，为了更好地反映这些非统计特征和

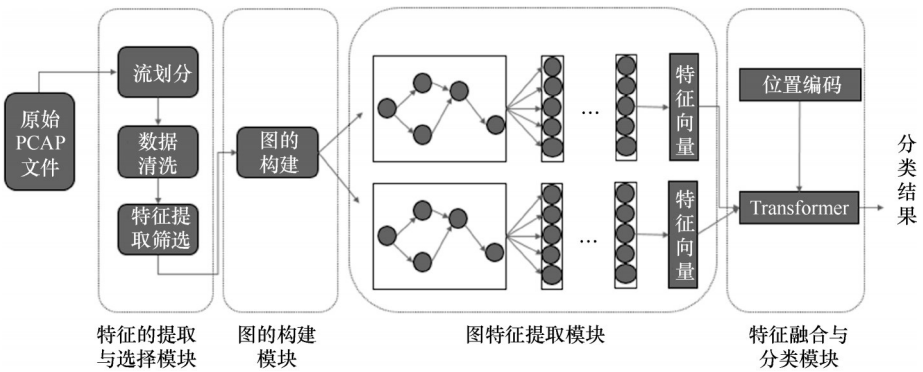



图2　算法整体架构

网站之间的关系，采用了基于频率的编码方案。以IP地址这个非统计特征为例，首先对IP地址根据其所属的服务提供商进行分组，然后统计某个IP分组中所有$n$个网站$web_1,web_2,\cdots,web_n$，出现的频率分别为$p_1,p_2,\cdots,p_n$,则此IP分组所有IP地址的编码为：

$$\textbf{IPcode}=(p_1,p_2,\cdots,p_n) \tag{2}$$

对于其他非统计特征，也采用类似的编码方式。选取IP地址、数字证书编号和泄露的URL编码得到**IPcode**、**CEcode**、**URLcode**连接在一起作为流的非统计特征，则流的非统计特征（non-statistical feature）**NSF**可以表示为：

$$\textbf{NSF}=\textbf{IPcode}||\textbf{CEcode}||\textbf{URLcode} \tag{3}$$

2.2.2　图的构建

由第2.1节的分析可知，随着网络环境的改变，属于同一流组中的流的个数也会发生改变。为了减少流的个数的改变对网站指纹识别的干扰，对流根据IP地址进行聚类并选取每个流组中前$n$个传输数据量最多的流，将其他流舍弃，这样也可以加快识别的速度。图的建立如图3所示，在图3中每个圆代表一个流，首先对流根据IP地址进行聚类，不同的颜色表示流组所属的流不同。然后选取每个流组中传输数据最多的$n$条流。使用一条单向边来表示流之间的时间关系，先产生的流对于后产生的流有一条单向边，同时产生的流之间没有边（对于发生时间间隔较短的流认为是同时产生的）。

节点的特征即流的特征，由于不同类型的特征的分布是不同的，为了更好地从同一类型的特征中进行特征提取，建立了多个空间结构相同的图，其中一个图的节点的特征为流的统计特征**SF**，其他图的节点的特征为流的非统计特征**NSF**，以便图特征提取模块能够结合多个流之间相同类型的特征更好地进行特征提取。

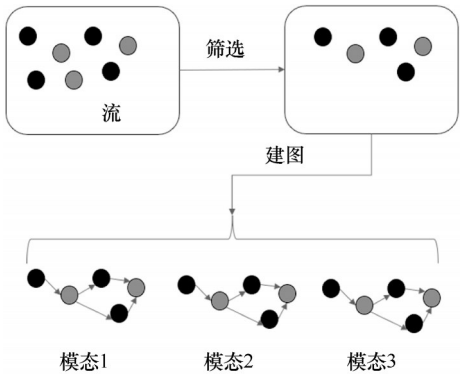


图3　图的建立

2.2.3　图特征提取模块

为了更好地对图进行综合表征，采用了图注意力网络（graph attention network，GAT）的改进版本GATv2[18]，GATv2可以很好地排除噪声的干扰。在GAT中，注意力系数不会随着节点的改变而改变，而GATv2中通过改变运算顺序避免了这一缺点，从而解决了噪声干扰的问题。

特征提取模块如图4所示，图4中每个小圆表示一个流，圆的颜色深浅代表流属于不同的流组。每一个图特征提取模块，都由池化层、卷积层和读出层构成，其中卷积层采用GATv2的GATv2Covn卷积，注意力头为3个，并且图特征提取模块将从图中提取一个1×42长度的一维特征向量，作为下一阶段的输入。

图4　特征提取模块

2.2.4　特征的融合与分类

在这一阶段，首先对图特征提取模块提取到的不同特征的特征向量进行拼接，然后使用Transformer[19]模型对其进行分类。Transformer模型使用注意力机制作为其编码器和解码器，从特征中提取信息，注意力机制如图5所示，每个输出的结果都是根据所有的输入计算得到的，很好

地解决了RNN和长短期记忆（long short-term memory，LSTM）网络的传递遗忘问题。但是从图5中也可以发现，相较于RNN和LSTM模型，Transformer模型对位置信息不敏感，需要对位置信息进行编码并输入。这里使用旋转式位置编码RoPE[20]，它可以通过绝对的位置编码方式，使得模型可以注意到相对位置。然后将拼接后的特征向量加上位置编码送入Transformer模型中进行分类，得到分类后的结果。

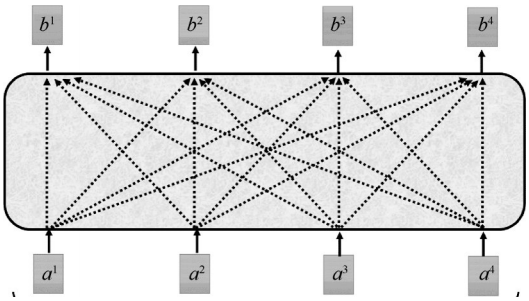



图5　注意力机制

## 3　实验与结构分析

### 3.1　实验数据

由于现阶段网上没有关于不同网络环境下访问网站的公开流量数据集，因此搭建了实验环境进行采集。通过在与主机相连的路由器上通过TC控制命令（TC控制命令是Linux系统中的一个网络管理工具，用于配置和管理网络流量控制。它可以用来限制网络带宽、实验、丢包等，以及实现QoS等功能）改变网络的环境，并在主机上抓取浏览器访问网站的流量，以此来采集不同网络环境下访问网站产生的流量。收集了21个网页的两个数据集，分别如下。

（1）数据集A为在没有增加额外的实验和丢包率的情况下，对于21个目标网页的大约1 500次的访问数据，每个网站大约访问量71次。共有33 639条数据流，由于目标网站都支持HTTPS协议，故都为TLS加密数据流。

（2）数据集B为通过TC控制命令，在与主机相连的路由器上额外增加(200±50) ms时延、丢包率5%的额外扰动情况下，对于21个目标网页的约1 400次访问数据，每个网页大约访问了67次。共有54 878条数据流，并且由于目标网站都支持HTTPS协议，故都为TLS加密数据流。

需要注意的是，在采集这两个数据集时，经过测试，路由器到访问同一网站的时延和丢包率是相对固定的（不考虑额外增加的扰动情况下）。并且并无意测量主机访问网站的真实网络环境，旨在通过这两个数据集网络环境的变化来说明网络环境对于访问网站流量的影响，并以此来测试算法抗网络干扰的能力。

### 3.2　测评标准

本文使用流量识别领域常用的指标准确率（accuracy，A）、精确率（precision，P）和召回率（recall，R）进行验证，预测结果混淆矩阵见表3。

表3　预测结果混淆矩阵

| 真实情况 | 预测结果 | |
|---|---|---|
| | 正例 | 反例 |
| 正例 | TP | FN |
| 反例 | FP | TN |

准确率$A$、精确率$P$和召回率$R$的计算分别如下：

$$A=\frac{\mathrm{TN}+\mathrm{TP}}{\mathrm{TN}+\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$$
$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \quad (4)$$
$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

### 3.3　实验设置与结果分析

以使用深度学习算法且比较有代表性的算法LSTM[10]、FS-Net[14]和针对网络波动进行优化的算法ProGraph[11]作为对比方法，把数据集A作为训练集，从数据集B中随机选取200次的访问数据作为

验证集，剩下的1 200次访问数据作为测试集。在实验中一共进行了100个epoch的训练，并且采用Adam优化器，学习率为0.000 1，batch size为64，在基于pytorch的NVIDIA 1070GPU上完成实验。

3.3.1　每个流组选取流的个数的影响分析

对于每个流组，选取传输数据最多的$n$个流，为了探究$n$对算法性能的影响，分别选取$n$=3、6、9、12和所有的流，设置epoch为100。分类器指标和选取流的个数的关系见表4。

表4　分类器指标和选取流的个数的关系

| 流个数 | 时间/s | *A* | *P* | *R* |
|---|---|---|---|---|
| 3 | 198 | 87.4% | 89.3% | 88.6% |
| 6 | 243 | 90.7% | 91.0% | 90.3% |
| 9 | 336 | 88.9% | 89.0% | 88.8% |
| 12 | 340 | 88.8% | 89.2% | 88.9% |
| 所有 | 443 | 87.0% | 89.5% | 89.2% |

由表4可知，随着选取的流的个数的增多，分类器的训练时间也在逐渐增加，这是由于选取的流的个数的增加，造成了图特征提取模块进行矩阵运算时的时间增加。并且，分类器的各项性能指标并不是随着选取的流的个数的增加而线性增加，而是先增加后减少，在$n$=6时，各项指标都达到了最高值。这说明并不是选取的流的个数越多越好，选取的流的个数过多，反而会造成一些不重要的流对分类器识别准确率的干扰，也会增加训练时间。

3.3.2　综合实验分析

选取流的个数为6个，并且将3种对比方法也采用以数据集A为训练集、数据集B为测试集的方法进行训练，数据集A→B不同方法实验结果如图6所示。其中本文算法实现了90.7%的准确率、91.0%的精确率和90.3%的召回率，是所有对比算法中表现最好的。ProGraph[13]实现了83.4%的准确率、83.4%的精确率和78.6%的召回率。FS-Net[14]表现最差，实现了24.3%的准确率、25.1%的精确率和20.6%的召回率。LSTM[10]实现了28.1%的准确率、31.6%的精确率和25.8%的召回率。并且可以看到，在网络波动的情况下，没有对网络波动情况进行处理的算法准确率很低。本文算法平均的准确率、精确率、召回率相对于最新的ProGraph[13]分别提高了7.3%、7.6%、11.7%。这充分说明了将流量视多种类型的数据进行融合并结合多个流的信息进行识别的合理性以及优越性。

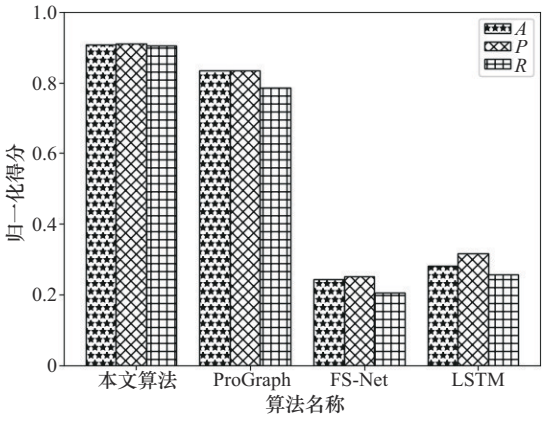



图6　数据集A→B不同方法实验结果

为了进一步验证本文算法对抗网络环境波动的鲁棒性，将两个数据集对调，即将数据集B作为训练集，数据集A作为验证集和测试集来进行训练。数据集B→A不同方法实验结果如图7所示。其中本文算法实现了90.8%的准确率、92.6%的精确率和90.9%的召回率。ProGraph[13]实现了83.0%的准确率、83.3%的精确率和79.3%的召回率。FS-Net[14]实现了25.3%的准确率、18.1%的精确率和17.6%的召回率。LSTM[10]实现了41.1%的准确率、35.6%的精确率和40.8%的召回率。可以看到本文算法依旧是对比算法中表现最好的，这充分说明了本文算法的优越性。

## 4　结束语

本文提出了一种基于图神经网络的鲁棒网站指纹识别方法。通过分析不同网络环境下访问网站产生的流量的差异，通过对于流的筛选和使用

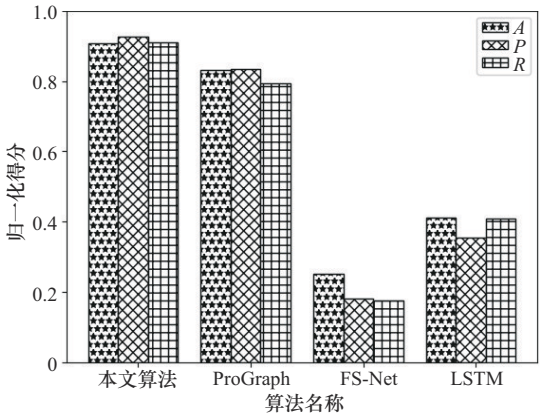



图7　数据集B→A不同方法实验结果

更高层的数据来计算统计特征以减少网络环境的干扰，并将流的特征视为一种多类型的数据，并通过图神经网络分别对不同类型的特征进行提取，最后通过Transformer进行特征的融合和分类，解决了其他网站指纹识别方法抗网络干扰能力差的问题，有效提高了网站指纹识别方法在不同网络环境下识别的准确率。

## 参考文献：


[1] SCHNEIDER P. TCP/IP traffic classification based on port numbers[J]. Division of Applied Sciences, Cambridge, MA, 1996, 2138(5): 1-6.

[2] EL-MAGHRABY R T, ABD ELAZIM N M, BAHAA-ELDIN A M. A survey on deep packet inspection[C]//Proceedings of the 2017 12th International Conference on Computer Engineering and Systems (ICCES). Piscataway: IEEE Press, 2017: 188-197.

[3] SUN G L, XUE Y B, DONG Y F, et al. An novel hybrid method for effectively classifying encrypted traffic[C]//Proceedings of the 2010 IEEE Global Telecommunications Conference GLOBECOM. Piscataway: IEEE Press, 2010: 1-5.

[4] DING R S, LI W M. A hybrid method for service identification of SSL/TLS encrypted traffic[C]//Proceedings of the 2016 2nd IEEE International Conference on Computer and Communications (ICCC). Piscataway: IEEE Press, 2016: 250-253.

[5] WANG W, ZHU M, ZENG X W, et al. Malware traffic classification using convolutional neural network for representation learning[C]//Proceedings of the 2017 International Conference on Information Networking (ICOIN). Piscataway: IEEE Press, 2017: 712-717.

[6] FENG W B, HONG Z, WU L F, et al. Network protocol recognition based on convolutional neural network[J]. China Communications, 2020, 17(4): 125-139.

[7] ABUROMMAN A A, BIN IBNE REAZ M. A novel weighted support vector machines multiclass classifier based on differential evolution for intrusion detection systems[J]. Information Sciences, 2017(414): 225-246.

[8] SERPEN G, AGHAEI E. Host-based misuse intrusion detection using PCA feature extraction and kNN classification algorithms[J]. Intelligent Data Analysis, 2018, 22(5): 1101-1114.

[9] LOPEZ-MARTIN M, CARRO B, SANCHEZ-ESGUEVILLAS A, et al. Network traffic classifier with convolutional and recurrent neural networks for Internet of things[J]. IEEE Access, 2017(5): 18042-18050.

[10] ZOU Z, GE J G, ZHENG H B, et al. Encrypted traffic classification with a convolutional long short-term memory neural network[C]//Proceedings of the 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). Piscataway: IEEE Press, 2018: 329-334.

[11] LI W H, ZHANG X Y, BAO H F, et al. ProGraph: robust network traffic identification with graph propagation[J]. IEEE/ACM Transactions on Networking, 2023, 31(3): 1385-1399.

[12] ACETO G, CIUONZO D, MONTIERI A, et al. MIMETIC: Mobile encrypted traffic classification using multimodal deep learning[J]. Computer Networks, 2019(165): 106944.

[13] SONG M Z, RAN J, LI S L. Encrypted traffic classification based on text convolution neural networks[C]//Proceedings of the 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT). Piscataway: IEEE Press, 2019: 432-436.

[14] LU J, GOU G P, SU M J, et al. GAP-WF: graph attention pooling network for fine-grained SSL/TLS website fingerprinting[C]//Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2021: 1-8.

[15] LIU C, HE L T, XIONG G, et al. FS-net: a flow sequence network for encrypted traffic classification[C]//Proceedings of the IEEE INFOCOM 2019 - IEEE Conference on Computer Communications. Piscataway: IEEE Press, 2019: 1171-1179.

[16] VAN EDE T, BORTOLAMEOTTI R, CONTINELLA A, et al. FlowPrint: semi-supervised mobile-app fingerprinting on encrypted network traffic[C]//Proceedings of the 2020 Network and Distributed System Security Symposium. Reston, VA: Internet Society, 2020: 27.

[17] HUOH T L, LUO Y, LI P L, et al. Flow-based encrypted network traffic classification with graph neural networks[J]. IEEE Transactions on Network and Service Management, 2023, 20(2): 1224-1237.

[18] BRODY S, ALON U, YAHAV E. How attentive are graph attention networks? [Z]. 2021.

[19] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS). Long Beach:Curran Associates Inc. 2017:6000-6010.
[20] SU J L, AHMED M, LU Y, et al. RoFormer: enhanced transformer with rotary position embedding[J]. Neurocomputing, 2024(568): 127063.

[作者简介]

**李孟想**（1999– ），男，中国科学技术大学硕士生，主要研究方向为流量识别。

**彭闯**（1998– ），男，中国科学技术大学硕士生，主要研究方向为计算机网络。

**王浩**（1997– ），男，中国科学技术大学硕士生，主要研究方向为计算机网络。

**黄超明**（1993– ），男，中国科学技术大学博士生，主要研究方向为计算机网络。

**谭小彬**（1973– ），男，博士，中国科学技术大学信息科学技术学院副教授，主要研究方向为未来网络架构、智能网络和多媒体通信。