计算机科学
COMPUTER SCIENCE

基于注意力机制的CNN和BiGRU的加密流量分类

陈思雨, 马海龙, 张建辉

引用本文

陈思雨, 马海龙, 张建辉. 基于注意力机制的CNN和BiGRU的加密流量分类[J]. 计算机科学, 2024, 51(8): 396-402.

CHEN Siyu, MA Hailong, ZHANG Jianhui. Encrypted Traffic Classification of CNN and BiGRU Based on Self-attention [J]. Computer Science, 2024, 51(8): 396-402.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230500032

基于注意力机制的 CNN 和 BiGRU 的加密流量分类

陈思雨[1]　马海龙[2]　张建辉[3]
1 郑州大学网络空间安全学院　郑州 450001
2 解放军战略支援部队信息工程大学信息技术研究所　郑州 450001
3 嵩山实验室　郑州 450001
(chensiyu0113@163.com)

摘　要　针对传统加密流量分类方法准确率低、利用流量载荷会侵犯用户隐私,以及泛化能力弱的问题,提出一种基于注意力机制的 CNN 和 BiGRU(CNN-AttBiGRU)的加密流量分类方法,可以同时适用于常规加密和 VPN、Tor 加密流量。该方法基于包大小、包到达时间以及包到达方向将流量转化为直观的图片,为提高模型准确率,使用 CNN 提取流量图片的空间特征,同时设计 BiGRU 和 Self-attention 模型提取时间特征,充分利用流量图片的时间和空间特征,可按照流量类别、加密技术和应用类型对流量进行不同层面的分类。该方法对加密流量类别分类的平均准确率达 95.2%,较以往提升 11.65%;对加密技术分类的准确率达 95.5%,较以往提升 7.1%;对流量所使用的应用程序分类的准确率达 99.8%,较以往提升 11.03%。实验结果表明,CNN-AttBiGRU 方法的泛化能力强,并且其仅利用加密流量的部分统计特征,有效地保护了用户隐私,同时取得了高准确率。

关键词:加密流量分类;深度学习;卷积神经网络;双向门控循环单元;自注意力机制
中图分类号　TP309

Encrypted Traffic Classification of CNN and BiGRU Based on Self-attention

CHEN Siyu[1], MA Hailong[2] and ZHANG Jianhui[3]
1 School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou 450001, China
2 Institute of Information Technology, PLA Information Engineering University, Zhengzhou 450001, China
3 Songshan Laboratory, Zhengzhou 450001, China

Abstract　To address the problems of low accuracy of traditional encrypted traffic classification methods, the use of traffic load will violate user privacy and weak generalization ability, an encrypted traffic classification method of CNN and BiGRU based on self-attention(CNN-AttBiGRU) is proposed, which can be applied to both regular encrypted and VPN and Tor encrypted traffic. The method converts traffic into intuitive pictures based on packet size, packet arrival time and packet arrival direction. To improve the accuracy of the model, CNN is used to extract the spatial features of traffic pictures, while BiGRU and self-attention models are designed to extract temporal features, making full use of the temporal and spatial features of traffic pictures. The traffic can be classified at different levels by traffic category, encryption technique and application type. The proposed method achieves an average accuracy of 95.2% for classification of encrypted traffic categories, which is 11.65% better than before; 95.5% for classification of encryption technologies, which is 7.1% better than before; and 99.8% for classification of applications used by traffic, which is 11.03% better than before. Experimental results show that the CNN-AttBiGRU method has strong ge-neralization ability and only utilizes some statistical features of encrypted traffic, which effectively protects user privacy while achieving high accuracy rates.

Keywords　Encrypted traffic classification, Deep learning, CNN, BiGRU, Self-attention

1　引言

近年来,流量加密技术被广泛用于保护互联网用户的隐私和匿名性,通信网络中加密流量的数量急剧增加。同时,VPN 以及 Tor 等加密技术的广泛使用给加密流量分类带来了巨大的挑战[1]。因此,识别加密流量类别、加密技术与加密

到稿日期:2023-05-06　返修日期:2023-08-31
基金项目:国家重点研发计划(2022YFB2901403);河南省重大科技专项(221100210900-01)
This work was supported by the National Key Research and Development Program of China(2022YFB2901403) and Major Scientific and Technological Project in Henan Province(221100210900-01).
通信作者:马海龙(longmanclear@163.com)

应用类型对于流量工程、网络维稳有很大的帮助，有助于网络管理者治理网络环境[2]，是实现高网络安全性和有效网络管理的关键技术。

随着点对点应用和端口伪装技术的出现，以及随后引入的非标准和动态端口，基于端口的方法不能胜任现代流量环境[3]，而深度包检测方法无法识别加密流量，并且会侵犯用户隐私，也不再适用于加密流量分类[2]，因此，研究人员引入机器学习的方法。文献[4]和文献[5]分别对 VPN 流量和 Tor 流量利用 K 近邻算法（K-Nearest Neighbour，KNN）和 C4.5 决策树进行分类，并选取两者中更好的结果作为分类结果，获得了较高的准确率。除上述文献外，支持向量机以及遗传算法等技术也得到了广泛的应用[6]，并取得了较好的效果。但传统机器学习方法需要大量专业知识，同时网络中数据复杂性和特征多样性也在不断提升[7]，机器学习不能达到有效分析和预测的目的，因此，研究者开始引入深度学习方法[8]。

深度学习在图像处理和自然语言处理中得到了很好的应用，但在加密流量分类中仍是一种新的研究思路[9]。文献[10]将会话的前几个非零负载转化成灰度图像，使用卷积神经网络（CNN）对灰度图进行分类。文献[11]将 PCA 算法与改进的深度卷积神经网络分类模型相结合来进行流量分类，PCA 算法进行降维分析，发现影响检测精度的关键特征，改进的深度卷积神经网络分类模型采用自主特征学习的方式提升分类精度。文献[12]采用堆叠式自动编码器和卷积神经网络自动从加密的流量有效载荷中提取特征。文献[13]对卷积神经网络模型进行优化以提高分类精度，嵌入 Inception 模块进行多维特征提取并进行特征融合。文献[14]将流数据转换为图片，使用卷积神经网络识别流的类别和应用程序。文献[15]针对 VPN 流量，将流量会话的前 1 512 字节转化为灰度图，利用卷积自编码将加密流量分为非 VPN 和 VPN 两类，利用卷积神经网络进一步识别 6 个不同应用产生的 VPN 流量。文献[16]针对 Tor 流量利用特征工程的方法，在流级别和主机级别两个层面提取流量特征对 Tor 流量进行检测，获得了较高的准确率。

研究发现，目前对网络流量的分类工作在常规加密流量、VPN 或 Tor 流量上取得了很好的成绩，但缺少能同时分类常规加密流量、VPN 流量和 Tor 流量的研究，模型存在泛化能力不高，以及在面对复杂的网络情况时准确率可能出现严重下滑的问题。同时，以上方法中还存在过度拟合某种流量类型或应用程序的问题，致使算法虽然在特定数据集上得到了较好的分类模型，却不适用于其他数据集。此外，一些研究依赖数据包的有效载荷，对用户的隐私造成了一定程度的威胁。

针对上述问题，本文提出了一种基于注意力机制的 CNN 和 BiGRU 的加密流量分类模型。首先使用流量的统计特征，即包大小、包到达时间以及包到达方向将加密流量转换成直观的图片，有效地保护了用户的隐私。CNN 和 BiGRU 结合 Self-attention 同时提取流量的时间特征和空间特征对加密流量进行分类，提高了模型的准确率。该模型同时适用于常规加密流量和 VPN、Tor 加密流量，具有一定的泛化能力。本文的主要贡献如下：

（1）提出了一种加密流量统计特征与图片之间的转换方法，避免使用加密流量的载荷信息，保护了用户的隐私。

（2）设计了一种同时使用加密流量的时间特征和空间特征的分类模型 CNN-AttBiGRU，相比现有方法，具有更高的准确性。

（3）解决了以往分类方法泛化能力不强的问题，CNN-AttBiGRU 模型可以同时分类常规加密流量、VPN 流量和 Tor 流量，并且在对流量类别、加密技术和应用类型进行不同层面的分类时，使用了完全相同的架构。实验结果表明，该模型具有较强的泛化能力。

2 算法设计

2.1 模型框架

本文提出的 CNN-AttBiGRU 模型将原始流量转化成以包到达时间和方向为横轴、以包大小为竖轴的流量图片，利用流量图片对加密流量的流量类别、流量的应用程序以及使用的加密技术进行分类。本方法的模型流程如图 1 所示，首先将原始流量按照 2.2 节中的数据处理方式转换成流量图片，再送入深度学习模型进行特征提取。特征提取过程主要分为两部分：一是利用 CNN 对图片的空间特征进行提取，二是利用 BiGRU 结合 Self-attention 对图片的时间特征进行提取。其输出作为加密流量的会话阶段无关特征向量。最后使用一个全连接层和 softmax 分类器进行最终的分类。

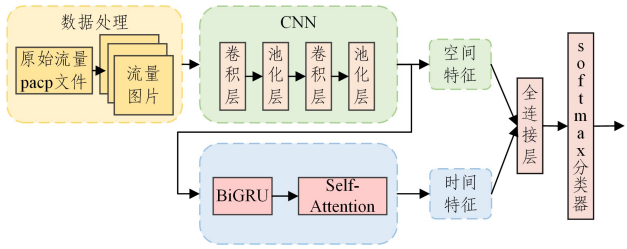

图 1 CNN-AttBiGRU 模型流程图

Fig. 1 Flowchart of CNN-AttBiGRU model

2.2 数据处理

本文使用了公开数据集 ISCX-VPN-nonVPN[4] 和 ISCX-Tor-nonTor[5] 进行研究，这两个数据集是由加拿大网络安全研究所（University of New Brunswick，UNB）发布的 VPN 流量和 Tor 流量有标签数据集。其中 ISCX-VPN-nonVPN 数据集包含 7 种常规加密流量和 7 种 VPN 加密流量，ISCX-Tor-nonTor 数据集包含 7 种流量类别的 Tor 加密流量。

（1）数据集处理。首先对两个公开数据集 ISCX-VPN-nonVPN 和 ISCX-Tor-nonTor 进行清洗。清洗的主要目的是人工删除具有以下错误的数据流：明显不符合标签类别的会话、重复会话、未成功建立的会话。

然后将清洗过的两个数据集分成 VoIP、视频、聊天、浏览和文件传输 5 种流量类别，常规加密、VPN 加密和 Tor 加密 3 种加密技术。将流量类别和加密技术相同的会话集合在一起，得到一个常规加密流量数据集、一个 VPN 加密流量数据集和一个 Tor 加密流量数据集。为了检验模型的泛化能力，

还创建了一个合并数据集：对这两个数据集进行抽样和合并，得到一个被分为 5 类的合并数据集，即 VoIP、视频、聊天、浏览和文件传输，并在 5 个类别中再将数据分为常规加密、VPN 加密和 Tor 加密这 3 种加密技术。

数据集由 pacp 文件组成，每个文件都根据相应的应用程序、加密技术和流量类别进行标记。每个流由五元组{源 IP、源端口、目的 IP、目的端口、协议}定义，最终保存在 CSV 文件中。同一条流中的多个数据包具有相同的五元组，并且数据包是按照时间顺序排序的。本文借鉴 Shapira 等[14]的工作，同时考虑会话的正反方向对其进行改进，将会话划分为正向(以先到达的流为正向)和反向两个流，每个流 60 s，形成一个 120 s 的块。合并数据集中每个流量类别和加密技术所对应的会话块数量如表 1 所列。

表 1 合并数据集中每个流量类别和加密技术所对应的会话块数量

Table 1 Number of session blocks corresponding to each traffic category and encryption technology in the merged dataset

	regular	VPN	Tor
Browsing	2816	—	1383
Chat	1411	1061	384
File Transfer	1022	167	1128
Video	605	274	760
VoIP	2405	2724	1061

(2)加密流量转图片。对数据集进行处理之后，将流量数据转换成图片。首先是从每个数据流中提取出每个包的包大小、包到达时间和包方向，将流量类别和加密技术相同的集合转化成 CSV 文件。然后以包到达时间为横轴、以包大小为纵轴构建流量图片。由于以太网的 MTU 值为 1500，包大小一般都比 1500 小，因此设置纵轴最大值为 1500。对于横轴，以先出现的 IP 地址为正向流量，在每个单向流中用包到达时间减去第一个包到达的时间以规范数据。$-60 \leqslant x < 0$ 部分视为正向流量，将 $0 \leqslant x < 60$ 部分视为反向流量，一张图片的流量时间设置为 60 s，形成一个二维直方图，如图 2 所示。最后将包大小、包到达时间进行等比映射，使其在 0 到 1500 之间，形成一个正方形的流量图像。流量图像可以看作是有效载荷大小分布的数组。与载荷信息不同，有效载荷大小分布是对网络或系统上传输的数据包的统计分析。有效载荷大小分布提供了对网络流量行为的洞察力，可用于检测网络中的异常行为。

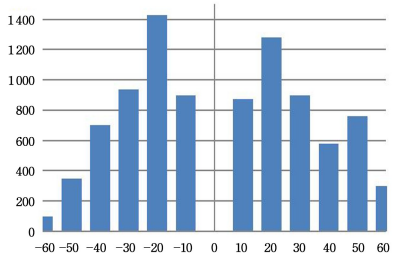

图 2 二维直方图示例图

Fig. 2 Example diagram of two-dimensional histogram

为充分利用有效载荷大小分布的特征，我们将每个流量图片存储在一个二维矩阵中，作为特征提取模型的输入。为了增加图片的数量，采取类似于滑动窗口的方法，每次向后滑动 20 s，允许两个相邻图片有部分重叠，但不能完全相同，转化的示例图片如图 3 所示。

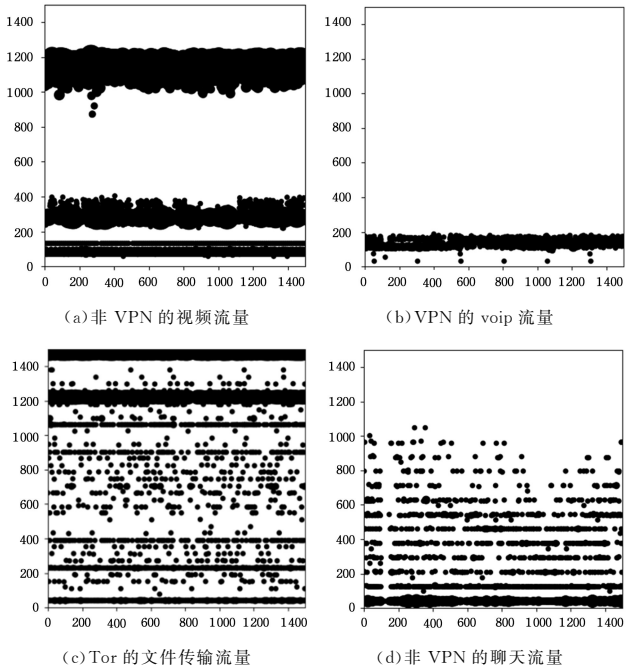

图 3 流量数据转为图片示例

Fig. 3 Example of converting traffic data to images

2.3 CNN 提取空间特征

CNN 是一种深度神经网络，在深度学习领域中发挥着重要作用[17]，经常用于图像分类[18]，具有很高的准确率。它可以被设计为自动和自适应地从输入数据中学习特征的空间层次结构，并且可训练参数少，便于训练，因此本文选取 CNN 来提取空间特征。本文设计的 CNN 模型如图 4 所示。

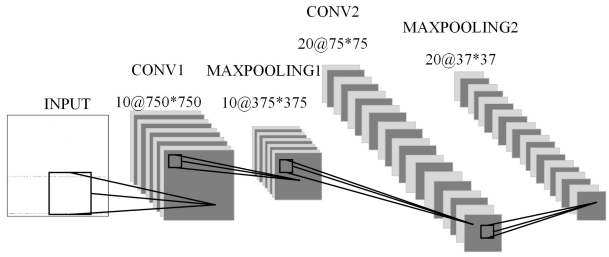

图 4 CNN 模型

Fig. 4 CNN model

加密流量的空间特征多种多样，为充分提取流量图片的空间特征，本文选取了两种不同大小的卷积核。CNN 网络参数设置如表 1 所列。该模型的输入是一个 1500×1500 的二维矩阵，第一层是二维卷积层 conv1，其输出是 10 个大小为 750×750 的特征图，接下来是第一个池化层，该层选择 max-pooling 方法，输出 10 个大小为 375×375 的特征图。第二个卷积层是 conv2，其输出是 20 个大小为 75×75 的特征图。下一层是第二个 2×2 max-pooling 层，输出 20 个大小为 37×37 的特征图。通过 CNN 模型，得到流量图片的空间特征。

表 2　CNN 参数设置

Table 2　CNN parameter settings

参数项	conv1	conv2
卷积核大小/个数/步长	5×5/10/2	10×10/20/5
卷积激活函数	ReLU	ReLU
池化类型	maxpooling	maxpooling
池化核大小	2×2	2×2
优化器	Adam	Adam
Dropout	0.25	0.5

2.4　BiGRU Self-attention 提取时间特征

CNN 提取空间特征的能力很强，但其不能捕捉到流量图片的时间特征。为充分利用流量图片的时间和空间特征，本文选取循环神经网络（RNN）来提取时间特征，但普通的 RNN 具有梯度消失的问题，不能保留长期信息[19]。RNN 的改进版本门控循环单元（Gated Recurrent Unit，GRU）增加了门控机制来控制隐藏状态之间的信息转换，并在不使用单独存储单元的情况下跟踪输入序列的状态，可以缓解 RNN 的梯度爆炸和梯度消失问题[20]。为充分学习前后文的时间特征，本文方法使用双向门控循环单元（Bidirectional Gated Recurrent Units，BiGRU）来对流量图片进行时间特征提取。

同时，考虑到在会话流量中每一个数据包的重要性不同，为了突出这种差异性，在 BiGRU 模型的基础上还采用了 Self-attention 机制对其最后一个时刻的隐藏层输出计算权重并进行加权求和，从而提高其准确性和效率。整个 BiGRU Self-attention 模型的结构如图 5 所示。该模型主要分为 4 层，首先是输入层，将 CNN 提取到的特征作为输入 e_n。将 e_n 送入 BiGRU 层的正向 GRU 单元和反向 GRU 单元，以提取时序特征。通过 BiGRU 层后每个输入会得到一个隐藏层正向输出 $\overrightarrow{h}_n$，逆向输出 $\overleftarrow{h}_n$。

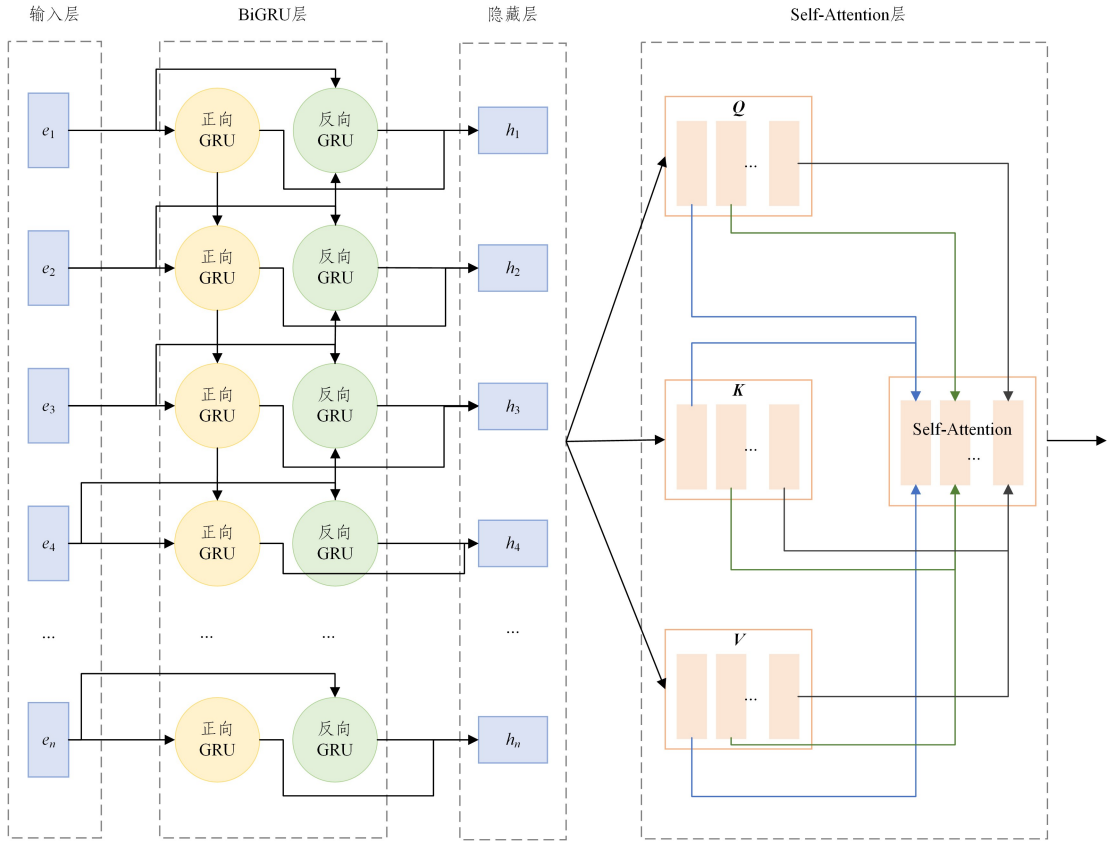

图 5　BiGRU＋Self-attention 结构图

Fig. 5　Diagram of BiGRU＋Self-attention structure

接着是 Self-attention 层，首先将 $\overrightarrow{h}_n$ 和 $\overleftarrow{h}_n$ 相加，

$$h_n=\overrightarrow{h}_n\oplus\overleftarrow{h}_n \tag{1}$$

得到 Self-attention 层的输入向量 $\boldsymbol{H}$，即$[h_1,h_2,h_3,\cdots,h_n]$，在该部分计算 $\boldsymbol{Q},\boldsymbol{V},\boldsymbol{K}$ 的值。

$$\boldsymbol{Q}=\boldsymbol{H}\times\boldsymbol{W}^{\boldsymbol{Q}} \tag{2}$$

$$\boldsymbol{V}=\boldsymbol{H}\times\boldsymbol{W}^{\boldsymbol{V}} \tag{3}$$

$$\boldsymbol{K}=\boldsymbol{H}\times\boldsymbol{W}^{\boldsymbol{K}} \tag{4}$$

然后由计算注意力权重：

$$a=\mathrm{softmax}\left(\frac{\boldsymbol{Q}*\boldsymbol{K}^{\mathrm{T}}}{\sqrt{D_k}}\right) \tag{5}$$

其中，D_K 表示查询向量 $\boldsymbol{Q}$ 或键向量 $\boldsymbol{K}$ 的维度。通过式(6)可得到最终的输出。

$$h_n=\sum_{j=1}^{N}v_j*a_j \tag{6}$$

通过 BiGRU 和 Self-attention 模型，得到流量图片的时间特征。

2.5　流量分类

为了对流量图片进行分类，首先利用 flatten 层，该层将 20 个特征映射转换为尺寸为 1 280 的一维层。下一层是大小为 64 的全连接层，为减少过拟合，使用 dropout 技术，其 dropout 概率设置为 0.5。最后的输出层是 softmax 分类器，

其中 n 的大小取决于分类子问题：5 为流量类别多类分类，3 为加密技术多类分类，10 为多类应用识别任务。参数 n 是不同问题之间体系结构的唯一区别。

训练的优化过程中使用基于梯度的 Adam 优化器。本文使用 Kingma[21] 中提供的默认超参数（$\alpha=0.001$，$\beta_1=0.9$，$\beta_2=0.999$，$\epsilon=10^{-8}$），并将 batch size 设置为 128。

网络总共运行 40 个 epoch，每个 epoch 有 10 个批次，并利用早停法在训练过程中保存准确率最高的结果。如图 6 所示，模型在运行了 30～40 个 epoch 后，实现了多类流量分类的收敛。分类采用一个 softmax 层，进行多分类。

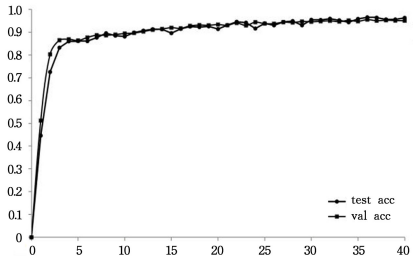

图 6 训练数据与测试数据准确率曲线

Fig. 6 Accuracy curve of training data and test data

3 实验

这一部分将展示本文所提出的加密流量分类方法的实验结果，并与其他方法进行对比。本文模型在 keras2.3.1 和 tensorflow2.2.0 上进行开发，使用的系统环境为 Windows 10，CPU 为 i5-10500。

3.1 评价标准

本文选取深度学习领域最常见的评估指标准确率（Accuracy）来评估模型性能，同时计算召回率 RC 和精度 PR，在此基础上使用所有类 F1 分数的平均值 F1 对不平衡的数据集进行评估。分类后的测试样本分为 4 类：(1)真阳性(TP)：正确分类的阳性样本的数量。(2)真阴性(TN)：正确分类的阴性样本的数量。(3)假阳性(FP)：错误分类为阳性样本的阴性样本的数量。(4)假阴性(FN)：错误分类为阴性样本的阳性样本数量。以下是上述评价标准的形式化定义：

$$Accuracy=\frac{\sum_{i\in \text{class}}^{n} TP_i}{\sum_{i\in \text{class}}^{n} TP_i+FP_i} \tag{7}$$

其中，TP_i 和 FP_i 分别为类 i 的真阳性和假阳性。

$$RC=\frac{TP}{TP+FN} \tag{8}$$

$$PR=\frac{TP}{TP+FP} \tag{9}$$

$$F1=2*\frac{PR*RC}{PR+RC} \tag{10}$$

3.2 实验结果与分析

实验主要研究了以下子问题：(1)在常规加密流量、VPN 流量、Tor 流量以及合并数据集上对流量类别进行分类，共有 5 类，分别是浏览、对话、文件传输、视频和 voip。(2)对加密流量使用的加密技术分类，有 3 种类别，分别是常规加密、VPN 和 Tor。(3)对流量所使用的应用类别进行分类，主要分为 facebook、skyep 等 10 个应用。

本文设置的对照实验有：(1)与 Flowpic 方法[13] 对比；(2)消融实验，即以本文方法为基准，减少 Self-attention 模块，并使用 CNN+单向 GRU 进行特征学习及分类；(3)采用 C4.5 与 KNN 更好的分类结果作为最终的结果对 VPN 和 Tor 分别进行分类[4-5]。

(1)加密流量类别分类。表 3 列出了本方法与 Flowpic、CNN+GRU 方法准确率、F1 值的对比结果。由表 3 可以发现，在 Tor 流量上所有方法的分类准确率都明显低于在其他流量上分类的准确率，因此在 Tor 上正确分类互联网流量类别更加困难。但本文方法在 Tor 流量分类上的准确率得到了明显的提升，总体结果优于 Flowpic 以及对照组 CNN-GRU，证明了本文方法能同时学习流量图片的空间特征和时间特征，BiGRU 和 Self-attention 机制可以进一步提高分类准确率。

文献[4]和文献[5]使用的是 C4.5 和 KNN 两种算法中更为优秀的结果分别对 VPN 和 Tor 流量进行分类，而本文则使用统一的模型结构但不同的训练参数来对常规加密流量、VPN 流量、Tor 流量以及合并数据集进行流量类别分类，取得了较高准确率，具有很强的泛化能力。

表 3 CNN-AttBiGRU 方法与其他方法在常规加密流量、VPN、Tor 以及合并数据集上对流量类别分类的结果

Table 3 Results of traffic category classification of CNN-AttBiGRU method versus other methods on regular encrypted traffic, VPN, Tor, and merged datasets

(%)

问题	CNN-AttBiGRU		Flowpic		CNN+GRU		C4.5+KNN	
	ACC	*F1*	*ACC*	*F1*	*ACC*	*F1*	*ACC*	*F1*
常规加密流量	97.0	88.4	85.0	83.4	90	88.40	84.0	82.1
VPN 流量	99.8	92.5	98.4	91.7	98	92.25	89.0	83.4
Tor 流量	89.0	81.4	67.8	64.9	84	79.40	84.3	80.7
合并数据集流量	95.0	90.4	83.0	80.3	86	82.00	—	—

CNN-AttBiGRU 方法和对比方法在常规加密流量、VPN 流量、Tor 流量上以及合并数据集上不同流量类别的召回率对比结果如图 7 所示。由图 7 可以看出，本文模型在 Tor 流量上取得了较大的进步，召回率明显高于 Flowpic 方法和 CNN-GRU 方法，但准确率仍未达到理想状态，主要是因为本文方法未能准确区分文件传输与其他流量。

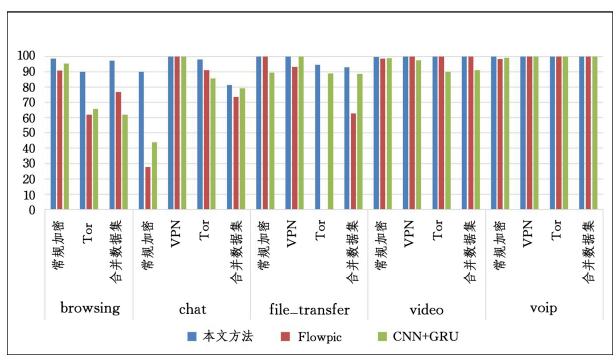

图 7 不同流量类别召回率比较

Fig. 7 Comparison of recall rates of different traffic categories

(2)加密技术分类。表 4 列出了本文方法对加密技术的分类结果。本文方法准确率达到了 95.5%，比 Flowpic 方法提高了 7.1%，比 CNN+GRU 方法提高了 5.5%；C4.5+KNN 方法对常规加密流量和 VPN 流量分类的准确率为 89%，对常规加密流量和 Tor 流量分类的准确率为 94.8%，本文方法分别提高了 6.5%和 0.7%。CNN-AttBiGRU 方法使用统一的模型，在加密技术分类中达到了较高的准确率，并且能分类不同种类的流量，泛化能力强。

表 4 流量加密技术分类对比结果

Table 4 Comparison results of traffic encryption technology classification

(%)

模型	CNN-AttBiGRU	Flowpic	CNN+GRU	C4.5+KNN
准确率	95.5	88.4	90	VPN 89.0 Tor 94.8

(3)应用类别分类。由于 C4.5+KNN 方法并未对合并数据集进行分类，同时本文的应用种类划分类别参考了 Flowpic 方法，因此该部分主要与 Flowpic 以及 CNN+GRU 进行比较，结果如表 5 所列。

表 5 流量应用类型分类对比结果

Table 5 Comparison results of traffic application type classification

(%)

模型	常规加密准确率	VPN 准确率	Tor 准确率
CNN-AttBiGRU	99.8	99.8	77.5
Flowpic	99.7	100.0	44.3
CNN+GRU	99.7	99.7	50.7

本文主要对 VoIP 和视频流量的应用进行了分类。在常规加密流量中，CNN-AttBiGRU 方法几乎完全分离了不同的应用程序，准确率达到了 99.8%，较 Flowpic 和 CNN+GRU 提高了 1%；在 VPN 流量中也得到了不错的准确率；在 Tor 流量中取得了较大的提升，达到了 77.5%的准确率，比 Flowpic 方法提高了 33.2%，比 CNN+GRU 提高了 26.8%。

以上实验结果表明 CNN-AttBiGRU 方法可以准确地分类流量类别、加密技术和应用类型，模型具有高准确率和高泛化能力。

结束语 面对使用流量字节信息会泄露用户隐私，以及传统加密流量分类方法准确率低、泛化能力弱的问题，本文提出了一种基于注意力机制的 CNN 和 BiGRU 的加密流量分类和应用识别方法。首先基于流量的统计特征，即包大小、包到达时间以及包方向，将加密流量转化成直观的图片，再利用 CNN 提取图片的空间特征，使用 BiGRU Self-attention 提取时间特征，最后对加密流量的流量类别、加密技术和应用类型进行分类。实验结果表明，CNN-AttBiGRU 方法利用少量统计特征就可以获得较高的准确率，对加密流量类别的平均准确率达 95.2%，对加密技术分类的准确率达 95.5%，对流量所使用的应用程序分类的准确率达 99.8%，并且具有很强的泛化能力。同时，本文方法避免了使用流量的字节信息，而是使用流量的统计特征，有效地保护了用户隐私。

在未来的工作中，可以加入深度学习模型可解释性方面的研究，从 Self-attention 层的权重以及卷积神经网络的特征图对模型的识别结果进行分析，进一步调整和优化模型结构；同时进一步加快模型识别速度，准确实时地对加密流量进行识别。

参 考 文 献

[1] WANG Z, FOK K W, THING V L L. Machine learning for encrypted malicious traffic detection: Approaches, datasets and comparative study [J]. Computers & Security, 2022, 113: 102542.

[2] REZAEI S, LIU X. Deep learning for encrypted traffic classification: An overview [J]. IEEE Communications Magazine, 2019, 57(5): 76-81.

[3] ZENG Y, GU H, WEI W, et al. Deep-Full-Range: a deep learning based network encrypted traffic classification and intrusion detection framework [J]. IEEE Access, 2019, 7: 45182-45190.

[4] 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.

[5] LASHKARI A H, DRAPER-GIL G, MAMUN M S I, et al. Characterization of tor traffic using time based features [C]// ICISSP. 2017: 253-262.

[6] WANG Y, ZHOU W Y, FENG H, et al. A deep convolutional neural network-based approach for network traffic classification [J]. Journal on Communications, 201839(1): 14-23.

[7] CHENG J, WU Y, E Y P, et al. MATEC: A lightweight neural network for online encrypted traffic classification [J]. Computer Networks, 2021, 199: 108472.

[8] ACETO G, CIUONZO D, MONTIERI A, et al. Mobile encrypted traffic classification using deep learning: Experimental evaluation, lessons learned, and challenges [J]. IEEE Transactions on Network and Service Management, 2019, 16(2): 445-458.

[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] 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.

[11] ZHANG S L, CHENG G, ZHANG W C. An improved deep convolutional neural network-based method for network traffic classification [J]. Chinese Science: Information Science, 2021, 51(1): 56-74.

[12] LOTFOLLAHI M, JAFARI SIAVOSHANI M, SHIRALI HOSSEIN ZADE R, et al. Deep packet: A novel approach for encrypted traffic classification using deep learning [J]. Soft Computing, 2020, 24(3): 1999-2012.

[13] XIE J N, MA C H, LI Z Y, et al. An encrypted traffic classification method based on convolutional neural networks [J]. Journal of Network and Information Security, 2022, 8(6): 84-91.

[14] SHAPIRA T, SHAVITT Y. FlowPic: A generic representation for encrypted traffic classification and applications identification [J]. IEEE Transactions on Network and Service Management,

2021,18(2):1218-1232.
[15] GUO L,WU Q,LIU S,et al. Deep learning-based real-time VPN encrypted traffic identification methods[J]. Journal of Real-Time Image Processing,2020,17:103-114.
[16] DODIA P,ALSABAH M,ALRAWI O,et al. Exposing the Rat in the Tunnel:Using Traffic Analysis for Tor-based Malware Detection[C]// Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. 2022:875-889.
[17] CHEN M H,ZHU Y F,LU B,et al. Attention-CNN-based application type identification for encrypted traffic[J]. Computer Science,2021,48(4):325-332.
[18] KRIZHEVSKY A,SUTSKEVER I,HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM,2017,60(6):84-90.
[19] 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,2019,8(1):241-252.
[20] LIU X,YOU J,WU Y,et al. Attention-based bidirectional GRU networks for efficient HTTPS traffic classification[J]. Information Sciences,2020,541:297-315.
[21] KINGMA D P,BA J. Adam:A method for stochastic optimization[J]. arXiv:1412.6980,2014.

CHEN Siyu, born in 2000, master. Her main research interests include cyber security and encrypted traffic classification.

MA Hailong, born in 1980, Ph. D, professor, Ph. D supervisor. His main research interests include endogenous security in cyberspace, intelligent awareness of cyber threats, and innovative cyber systems.

(责任编辑:何杨)

2024CCF 科技创业大赛安徽专项赛正式启动|TEC2024

为响应创新驱动发展战略,优化科技创新创业生态环境,推动创新链、产业链与资金链深度融合,促进先进技术成果转化应用,提高服务计算产业技术创新和应用创新团队的能力,吸引更多先进技术、新锐企业及团队落地合作城市,带动城市产业创新、升级与发展,CCF 携手地方政府、投资公司、科研院所共同举办 CCF 科技创业大赛(简称 TEC)。

TEC 已连续举办三届,2023TEC 安徽专项赛,历时三个多月,吸引全国 300 多个项目团队参赛,80 支参赛队伍进入复赛,省外团队占比超过 80%。经过激烈角逐,20 支参赛队伍晋级总决赛,多个项目已落地或正在落地安徽,并获得了资金支持。

2024 年,CCF 再次携手安徽省人工智能产业推进组办公室(隶属省科技厅),合作举办 TEC 2024 安徽专项赛——全国通用人工智能创新应用大赛。大赛于 7 月 16 日在安徽创新馆举办启动仪式。

安徽省人民政府副省长任清华、省科技厅党组书记吴劲松、省政府副秘书长张红君、CCF 副秘书长束庆山、马鞍山市副市长左年文、芜湖市副市长朱的娥、合肥市政府副秘书长张志上台共同启动 2024TEC 安徽专项赛。

今年赛事将全面升级,加强应用场景供给,设置“人工智能+”底层能力、汽车、社会服务、工业制造 4 个场景应用专项赛道。CCF 和安徽省将继续发动全国高校、科研院所的优秀科研技术团队报名参赛,并进一步细化各项支持政策,提升对落地团队的服务能力,扶上马再送一程。

报名流程:参赛团队可通过登录大赛官网(https://ccf.org.cn/tec2024),下载“项目登记表”“商业计划书”模板,填写项目内容。在大赛网站注册报名、上传模板内容,完成在线报名。

TEC2024 在此开启,我们期待更多合作城市、更多支持或协办单位参与一起举办 TEC2024,更好地践行 3M 原则,服务计算领域的专业人士和机构。

据 CCF 微信公众号