DOI:10.19651/j.cnki.emt.2107699

一种面向类不平衡加密流量的端到端分类模型*

林 鹏 翟江涛 许历隆 崔永富
(南京信息工程大学 电子与信息工程学院 南京 210044)

摘 要: 当前流量分类方法在面对类不平衡流量时,往往存在着在少数类上的分类效果不佳的问题。针对该问题,提出了一种面向类不平衡加密流量的端到端分类模型。所提模型在传统卷积神经网络模型的基础上添加了一个Inception模块进行特征融合,让模型能提取到更丰富的特征,弥补了少数类因样本数量少所带来的特征学习上的不足;同时引入一个通道-空间域注意力模块,对Inception模块所融合的特征根据重要程度赋予相应的权值,使模型更多地关注到更重要的特征,增强流量特征的表征能力。与此同时,为减少网络参数,采用卷积层加全局平均池化层的组合代替模型中的全连接层。实验结果表明,相较于当前典型流量分类模型,所提模型在数据集少数类上具有更优的分类性能,精确率、召回率和F1-Score均有显著提高,其中综合性能指标F1-Score在某些少数类上的提升达到了15%~18%。

关键词: 计算机神经网络;类不平衡;端到端;卷积神经网络;特征融合;注意力模块
中图分类号: TP393.0 **文献标识码:** A **国家标准学科分类代码:** 413.20

An end-to-end classification model for class-unbalanced encrypted traffic

Lin Peng Zhai Jiangtao Xu Lilong Cui Yongfu
(College of Electronical and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract: Current traffic classification methods often suffer from poor classification effects on minority classes when facing class imbalanced flows. To solve this problem, an end-to-end classification model for class-unbalanced encrypted traffic is proposed. The proposed model adds an Inception module to the traditional CNN model for feature fusion, so that the model can extract richer features, which makes up for the lack of feature learning caused by the small number of samples in a minority of categories. At the same time introduces a the channel-spatial domain attention module assigns corresponding weights to the features fused by the Inception module according to their importance, so that the model pays more attention to more important features and enhances the characterization ability of traffic features. At the same time, in order to reduce network parameters, a combination of convolutional layer and global average pooling layer is used to replace the fully connected layer in the model. The experimental results show that compared with the current typical traffic classification model, the proposed model has better classification performance on the minority classes of the data set, and the accuracy rate, recall rate and F1-Score have been significantly improved. And the comprehensive performance index F1-Score is on some minority classes. The improvement reached 15%~18%.

Keywords: computer neural network; class imbalance; end-to-end; convolutional neural network; feature fusion; attention module

0 引 言

网络流量中蕴含了大量有价值的信息,如何对网络流量数据进行智能化的分析和处理,已成为当前理论和应用研究的热点[1]。随着人们对隐私保护和数据安全的重视程度不断提高,SSL/TLS及SSH等加密技术在数据传输过程中得到了越来越广泛的应用,这使得加密流量在网络流量中所占的比重越来越大[2]。流量加密在保证隐私安全的同时,也会导致针对流量的检测和识别难度加大。如何对加密的网络流量数据进行高效而准确的分类和识别任务,从而及时有效地检测恶意流量,成为了当前网络管理和网络安全领域的关键问题。

收稿日期:2021-08-26
* 基金项目:国家自然科学基金(U1836104,61772281,61801073,61931004,62072250)、南京信息工程大学人才启动基金(2020r061)项目资助

基于端口[3]和负载分析[4]的传统流量检测方法无法适应当前网络环境下的流量特点,在性能上出现了显著的下降[5]。基于机器学习的典型流量分类模型包括支持向量机SVM、随机森林 RF、多层感知器 MLP[6]以及卷积神经网络CNN[7-10]等虽然可以在面对加密流量时取得较高的分类准确率,但是前提条件是流量数据各类别之间要达到基本平衡,即只有在面对平衡数据集时,上述分类模型才能实现良好的分类性能。数据的类不平衡现象通常表现为某些类别的样本数量远远大于另一些类别,样本数量占比大的称之为多数类,反之则称为少数类。对于少数类样本,分类模型无法获取到充足的特征来学习原始输入与预期输出之间的非线性关系,因此无法准确地判断样本的类别。此外,对少数类的错分对整体的准确率不会造成很大的影响,模型也就不会受到足够多的"惩罚",因此在遇到难以判断的样本时,分类模型往往倾向于将其划分为多数类,导致对少数类的分类性能下降。在网络流量领域,恶意流量相对于正常流量来说属于少数类,一旦错误地将恶意流量分类为正常流量,就会导致网络系统遭到恶意攻击,从而造成不可估量的损失。因此,解决网络流量领域内类不平衡现象带来的问题具有很大的实际意义。

文献[11-13]采用当前主流的深度学习模型对加密流量进行分类,实现了较高的整体分类准确率,但是在少数类上的分类精度并不理想,这就降低了所设计的流量分类方法的实际应用价值。当前针对类不平衡流量数据分类的研究较少,且主要是从数据处理的角度出发,通过对数据进行特征选择和重采样的方法来减小类不平衡的影响。文献[14-15]将特征选择与机器学习方法相结合,在不平衡流量数据集上取得了不错的分类效果。然而,流量特征会随着网络环境的变化而变化,出现概念漂移现象。基于特征选择的分类算法在面对概念漂移现象时分类性能会受到影响,并且对于特征的选择需要研究人员具备丰富的特征工程经验,实现难度较大。文献[16-18]通过不同的方式对流量数据进行重采样,再结合分类算法对不平衡流量数据集进行分类,均实现了良好的分类性能。但是基于采样的方法对样本进行了增减,容易出现有用信息丢失以及数据冗余的问题,影响最终的分类性能。

针对上述问题,本文对传统的 CNN 模型进行了改进,提出了一种端到端流量分类系统,实现了对类不平衡网络流量中少数类的高精度分类,本文的主要贡献主要包括如下 3 个方面。

1)本文利用深度学习中神经网络自主学习的特点,提出了一种基于 CNN 的端到端流量分类模型,将原始流量数据经预处理转换为灰度图片后作为分类模型的输入,由分类模型对图像数据进行训练后输出分类结果。

2)针对当前分类模型在面对类不平衡数据集中少数类时在特征学习上的不足,本文在传统 CNN 模型中添加了 Inception 模块,将 Inception 模块中多个卷积层以不同的感受野所提取到的不同级别的特征进行融合,让模型可以通过更丰富的特征来学习输入与输出之间的非线性关系,提高各个类别的分类精度。

3)本文设计了一个通道-空间域注意力模块并引入分类模型中,用于对 Inception 模块输出的融合后的特征赋予相应的权值,使模型更多地关注流量数据中更关键和重要的特征,相对性地忽略与分类任务无关的特征,从而增强流量特征的表征能力,提高分类精度。与此同时,采用卷积层加全局平均池化层代替传统 CNN 中的全连接层,减少了网络参数,从而降低了计算成本以及过拟合的风险,增强了模型的泛化能力。

1　本文方法

网络流量有着数量规模大,容易随着网络环境的变化而变化的特点,导致当前典型的基于特征选择以及重采样的方法在处理类不平衡流量数据时面临着诸多问题。因此本文从分类模型的角度出发,通过对传统的 CNN 模型进行改进,提出了一种面向类不平衡加密流量的端到端分类模型。该模型在传统 CNN 模型的基础上,加入了 Inception 模块来实现多尺度特征融合,同时设计并引入了通道-空间域注意力模块,同时用 1 个卷积层加 1 个全局平均池化层的组合取代了传统的全连接层。该分类模型的网络结构如图 1 所示。

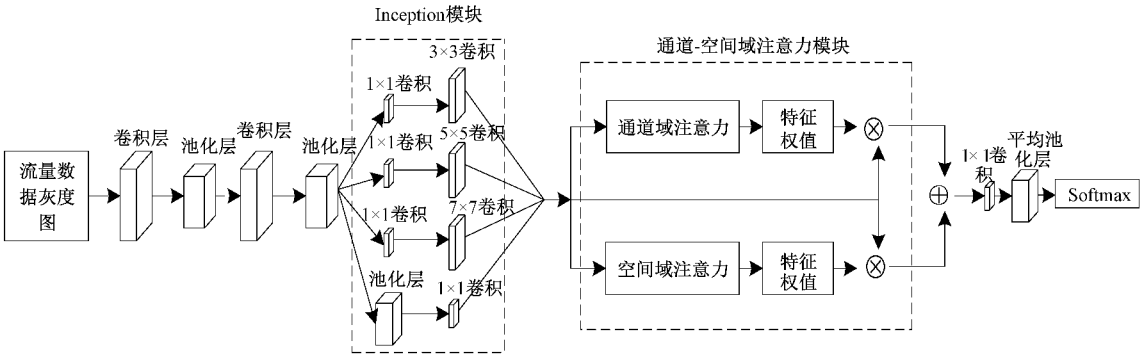

图 1　本文所提端到端分类模型网络结构

1.1　基于 Inception 模块的特征融合

本文通过在 CNN 中添加 Inception 模块来实现多尺度特征融合，从而使模型可以提取到更丰富的特征。Inception 模块是一种由多个不同卷积核大小的卷积层以及最大池化层组成的网络结构，用于增加网络的宽度，可以通过不同的感受野提取到不同级别的特征并进行融合。本文采用的 Inception 模块由卷积核大小为 3×3、5×5 和 7×7 的卷积层以及 1 个最大池化层组成，同时为了简化模型，在模块中加入了 1×1 卷积用于控制输出的特征维度，Inception 模块的网络结构如图 2 所示。

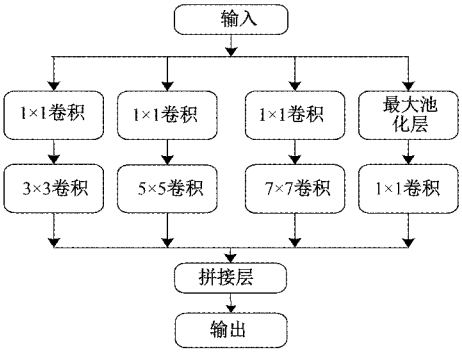

图 2　Inception 模块结构

传统 CNN 模型在网络层较浅时无法提取到更丰富且深层次的特征，这在很大程度上制约了模型的分类性能。对于类不平衡数据中的少数类，因为其样本数量较少，神经网络无法提取到足够多的样本特征，从而造成分类器无法很好地学习到输入的流量数据与其类别的非线性关系。为了克服上述问题，最常见的方法是增加网络的深度，即添加更多的网络层，这大大增加了网络的复杂度，在训练过程中容易出现退化的现象。与此同时，深层网络在训练大规模的流量数据时会带来巨大的训练时间和计算成本。

本文设计的 Inception 模块由 3×3、5×5 和 7×7 卷积以及 1 个最大池化层构成，这些卷积层和池化层以不同的感受野提取到不同级别的特征，并将这些特征进行拼接融合，得到一个包含信息更丰富的特征向量。Inception 模块的添加使得 CNN 模型可以在不增加深度的情况下通过增加网络的宽度让神经网络提取到更丰富的特征，弥补了传统 CNN 模型在特征提取上的缺陷。考虑到增加网络的宽度会导致特征维度增大，因此在 Inception 模块的卷积层前面以及池化层的后面添加了 1×1 卷积，用于 Inception 模块的输出特征维度约简，降低模型的复杂度。1×1 卷积的卷积核没有提升感受野的作用，可以通过设置其通道数来改变输入的维度，达到减小维度的目的。

Inception 模块的卷积层后面都连接了批量归一化(BN)层，BN 层对输入的数据进行批次化的减均值除方差操作，可以加快网络训练速度，增加模型的鲁棒性。BN 的计算公式如式(1)所示。

$$X_{ij}^{k}=\gamma_{ij}\frac{H_{ij}^{k}-\mu_{ij}}{\sigma_{ij}}+\beta_{ij} \tag{1}$$

其中，H_{ij}^{k} 是训练第 k 条数据时，第 j 层的第 i 神经元模型的输出值；μ_{ij}、σ_{ij} 分别是均值和方差；γ_{ij}、β_{ij} 是两个学习参数，用于将归一化后的数据放大并平移回非线性区域。

在对数据进行 BN 操作之后，使用线性激活单元 ReLU 对数据进行处理，ReLU 激活函数的定义如式(2)所示。

$$f(x)=\begin{cases}0, & x\leqslant 0\\ x, & x>0\end{cases} \tag{2}$$

1.2　通道-空间域注意力模块的引入

注意力机制被广泛应用于计算机视觉[19]以及自然语言处理领域，可以让模型更多地关注到更关键重要的信息，从而提高模型的性能。本文将原始的流量数据经预处理转换为灰度图像，然后对灰度图像进行分类，参考注意力机制在图像分类上的应用，本文将注意力机制引入分类模型中。Inception 模块对不同级别的特征进行融合可以使模型提取到更丰富的特征信息，但同时也可能会产生一些对分类任务没有帮助的冗余信息，出现特征冗余的问题。

通过引入注意力机制模块与 Inception 模块相连接，可以对 Inception 模块融合的多尺度特征赋予相应的权重，帮助模型从中获取到更关键、更有利于分类的特征，从而相对忽略一些无关的冗余特征。本文在 CNN 模型中设计并引入了一个通道-空间域注意力模块，该模块将两类注意力机制通过并联的方式结合起来，同时从通道和空间两个不同的维度对特征进行增强，提高特征的表征能力。本文设计的通道-空间域注意力模块结构如图 3 所示。

1)通道域注意力模块的引入

通道域注意力模块关注的是特征图的不同通道的重要程度，对更重要的通道赋予更高的权重。由图 3 可知，在该模块中，经过卷积池化操作得到的特征图作为模块的输入，并通过全局平均池化层 GAP 和全局最大池化层 GMP 并行地对空间维度进行压缩，得到两个 1×1×C 的特征向量，其中 C 为输入特征图的通道数。通常情况下只需要使用 GAP 对特征图进行压缩操作，这里额外加入了 GMP 作为 GAP 的补充，目的是提取到更丰富的特征信息。经过压缩得到的两个特征向量输入到一个共享的多层感知器 MLP 中，MLP 由两个神经元个数分别为 $C/8$、C 的全连接层构成，用于通道的降维和增维，拟合通道间的相关性。最后将两个输出相加并进行 sigmoid 激活操作，得到维度为 1×1×C 的通道权值向量，再与原输入特征图进行全乘，对不同通道赋予相应的权重。通道权值向量的计算公式如式(3)所示。

$$M_c(F)=\sigma(MLP(GAP(F))+MLP(GMP(F))) \tag{3}$$

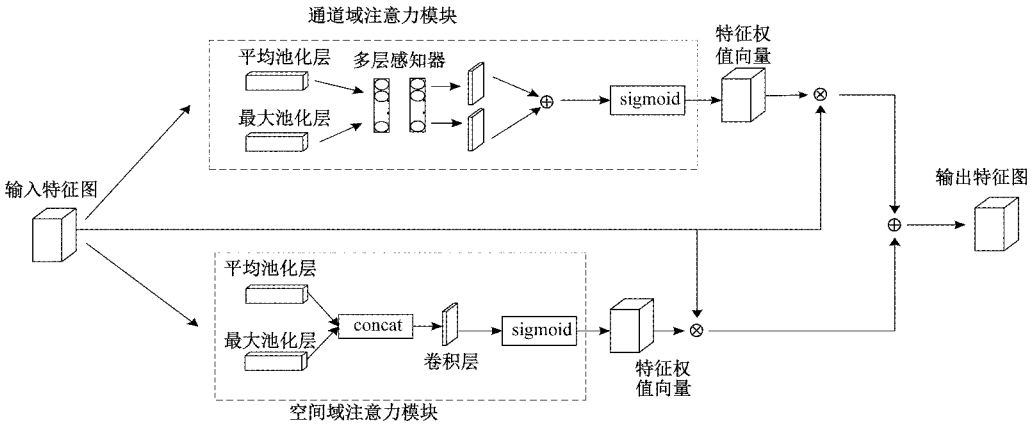

图 3　通道-空间域注意力模块

其中，$M_c(F)$ 表示特征图 F 的通道权值向量，σ 代表 sigmoid 函数。

2)空间域注意力模块的引入

空间域注意力模块则从特征图的空间维度出发，引入注意力机制。该模块是通过 GAP 和 GMP 并行地对特征图的通道维度进行压缩，得到两个 $H\times W\times 1$ 的特征向量，其中 H 和 W 为宽和高。然后将这两个特征向量进行通道拼接操作，得到一个通道数为 2 的特征向量，接着用一个 7×7 的卷积核对特征向量进行卷积，最后通过 sigmoid 激活操作得到维度为 $H\times W\times 1$ 的特征图的空间权值向量。计算公式如式(4)所示。

$$M_s(F)=\sigma(f^{7\times 7}(concat(GAP(F);GMP(F)))) \quad (4)$$

其中，$f^{7\times 7}$ 代表卷积核大小为 7×7 的卷积操作，$concat$ 代表在通道维度上的拼接操作。

相较于单一域注意力机制，本文设计的通道-空间域注意力模块将通道域和空间域注意力模块结合起来，能够更全面地反映特征在不同维度上的重要程度，使模型可以从整体的角度上更多地关注到更关键、更有利于分类的特征，提升分类效果。

此外，本文采用 1 个卷积层加 1 个全局平均池化层的组合来代替 CNN 模型中的全连接层，其目的是减少网络参数、避免模型的过拟合。其中卷积层的卷积核大小为 1×1，卷积核数量则代表了最终的分类类别数，卷积层的输出再经过一个全局平均池化层得到一个维度为 $(1,1,N)$ 的输出向量，其中 N 为分类的类别数量。该部分的数据处理过程如图 4 所示。

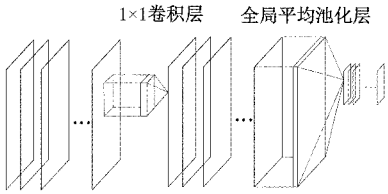

图 4　1×1 卷积加全局平均池化层

在模型的最后通过 Softmax 激活函数层进行最终的分类，与 sigmoid 函数不同的是，Softmax 函数被用于多分类，可以通过计算来预测对象属于各个类别的概率，得到一个 N 维概率向量。Softmax 函数的计算公式如式(5)所示。

$$y_i=\frac{e^{z_i}}{\sum_{j=1}^{N}e^{z_j}},i=1,\cdots,N \quad (5)$$

其中，y_i 表示预测对象为第 i 类的概率，z 为 Softmax 的输入，N 为分类的类别数。

2　实　　验

2.1　数据集简介及预处理

本文实验所采用的数据集是“UNB ISCX VPN-nonVPN 2016”[20]加密流量数据集，该数据集包含 15 种应用程序的 VPN 和非 VPN 加密流量，本文按照服务类型将数据集中的加密流量分为 12 类，每种服务类型包含了多种具体的应用程序流量，每一类的样本数量以及在数据集中所占比例如表 1 所示。

表 1　数据集中流量样本数量及比例

服务类别	样本数量	所占比例/%	服务类别	样本数量	所占比例/%
Chat	10 291	6.53	VPN_Chat	4 014	2.55
Email	7 737	4.91	VPN_Email	293	0.19
File	57 887	36.73	VPN_File	994	0.63
P2P	1 020	0.65	VPN_P2P	475	0.30
Streaming	1 406	0.89	VPN_Streaming	541	0.34
Voip	65 966	41.85	VPN_Voip	6 991	4.44

由表 1 可知，12 种服务类型的流量样本数量极其不平衡，其中 File 类和 Voip 类的样本数量分别在整个数据集中占据 36.73%和 41.85%，而 P2P、Streaming、VPN_Email、

VPN_File、VPN_P2P 和 VPN_Streaming 等 6 类占据的比例均不足 1%。本文将数据集内的.pcap 格式的加密流量数据首先进行数据包的过滤，将一些对分类没有帮助的数据包剔除，然后按照会话的粒度进行流量的切分以及截断填充，将每条会话流的长度统一为 784 Byte，最后生成对应的灰度图像。

2.2　性能评价指标

单一的准确率 Accuracy 往往无法全面地反映出分类器的性能，特别是对于类不平衡数据集的分类，需要从多方面去评估分类器在各个类别上的分类效果。因此，本文采用典型的 4 种分类性能评价指标：准确率 Accuracy、精确率 Precision、召回率 Recall 以及 F1-Score 来对本文实验进行全面的评估。上述 4 种评估指标的计算公式如下：

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

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

$$F1\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{9}$$

其中，TP 为真正例，表示被预测为正例的正例；TN 为真负例，表示被预测为负例的负例；FP 为假正例，表示被预测为正例的负例；FN 为假负例，表示被预测为负例的正例。

2.3　结果与分析

1）消融实验

（1）添加 Inception 模块

本文加入的 Inception 模块由 3×3、5×5、7×7 卷积和最大池化层并行连接构成，每一条支路分别与一个 1×1 卷积串行连接。为了体现所采用的 Inception 模块网络结构的优势，本文分别与添加了其他不同结构 Inception 模块以及未添加 Inception 模块的 CNN 分类模型进行了对比实验，实验结果如表 2 所示。

表 2　不同 CNN 分类模型分类结果对比

分类模型	Inception 模块内部结构	Accuracy	Precision	Recall	F1-Score
CNN1	未添加	0.950 4	0.928 2	0.908 0	0.916 2
CNN2	3×3+5×5 卷积	0.960 4	0.935 7	0.935 7	0.935 7
CNN3	3×3+5×5+7×7 卷积	0.965 3	0.961 3	0.925 9	0.943 7
CNN4	3×3+5×5 卷积+最大池化	0.968 2	0.952 8	0.934 6	0.940 7
CNN5	3×3+5×5+7×7 卷积+最大池化	**0.986 3**	**0.975 1**	**0.955 5**	**0.964 6**

表 2 中的 CNN1 未添加 Inception 模块，CNN2、CNN3 以及 CNN4 则添加了不同结构的 Inception 模块，即构建 Inception 模块的网络层不同。CNN2 中添加的 Inception 模块由 3×3 和 5×5 卷积构成，CNN3 和 CNN4 分别在此基础上增加了 7×7 卷积和最大池化层，CNN5 则添加了本文设计的 Inception 模块，由 3×3、5×5、7×7 卷积和 1 个最大池化层构成。

由表中的数据可以看出，CNN2、CNN3、CNN4 和 CNN5 的各项性能指标相较于未添加 Inception 模块的 CNN1 均有不同程度的提升，其中 CNN5 提升最为明显，在 Accuracy、Precision、Recall 和 F1-Score 四项性能指标上分别提高了 3.59%、4.69%、4.75% 和 4.84%。此外，CNN5 在各性能指标上均优于 CNN2、CNN3 和 CNN4。因此可知，添加 Inception 模块能让模型提取到更丰富的特征，实现更优的分类性能，并且本文设计的 Inception 模块对模型的提升效果要优于其他 3 种 Inception 模块。

（2）引入通道-空间域注意力模块

为了验证本文所提分类模型中通道-空间域注意力模块的有效性以及相比于单一域注意力模块的优势，本部分实验针对注意力模块进行了相关对比，实验模型包括未添加注意力模块的普通 CNN 模型、分别添加了单一域注意力模块和通道-空间域注意力模块的 CNN 模型，对比结果如表 3 所示。

表 3　各分类模型对比结果

分类模型	Accuracy	Precision	Recall	F1-Score
普通 CNN	0.950 4	0.928 2	0.908 0	0.916 2
CNN+通道域注意力模块	0.962 8	0.941 6	0.942 5	0.942 0
CNN+空间域注意力模块	0.953 7	0.934 5	0.945 7	0.940 1
CNN+通道-空间域注意力模块	**0.977 2**	**0.968 9**	**0.957 9**	**0.963 4**

由表 3 数据可知，添加了注意力模块的 CNN 模型在各项性能指标上均优于普通 CNN 模型，因此可知，注意力模块能使模型提取到更关键、更有利于分类的特征，从而提高模型的分类效果。与此同时，添加了通道-空间域注意力模块的 CNN 模型相较于添加了通道域以及空间域注意力模块的 CNN 模型在分类性能上了有了进一步的提升，

这表明本文设计的通道-空间域注意力模块能使模型更全面地从不同维度去关注更关键重要的特征,相较于单一域注意力模块具有更大优势。

最后,结合 Inception 模块和通道-空间域注意力模块进行综合对比实验,测试各模块的引入以及相互结合对分类模型性能的影响,对比结果如表 4 所示。

表 4　各分类模型对比结果

分类模型	Accuracy	Precision	Recall	F1-Score
普通 CNN	0.9504	0.928 2	0.908 0	0.916 2
CNN+Inception 模块	0.986 3	0.975 1	0.955 5	0.964 6
CNN+通道-空间域注意力	0.977 2	0.968 9	0.957 9	0.963 4
CNN+Inception+通道-空间域注意力模块	**0.995 4**	**0.990 3**	**0.977 4**	**0.983 8**

由表 4 中数据可知,通过引入 Inception 模块和通道-空间域注意力模块,分类模型在数据集上的分类性能相较于普通 CNN 模型均有明显地提升,且优于单独引入其中一个模块的 CNN 模型,充分体现了 Inception 模块与通道-空间域注意力模块相结合的优势。

2)与主流分类算法对比

为了体现本文所提分类模型相较于当前主流的加密流量分类方法在分类性能上的优势,本文对比了其他文献中基于深度学习的分类算法,对比结果如表 5 所示。表 5 中所对比的各文献里的分类算法都是在"UNB ISCX VPN-nonVPN 2016"数据集上进行流量分类实验的,但是分类的类型有所不同,如文献[11]只对 VPN 流量中的 6 种服务类型进行了分类测试,文献[13]对其中的应用程序进行分类,文献[12-13]则针对流量的服务类型进行了 12 分类的测试。一类服务类型的流量包含了多种应用程序类别的流量,且不同服务类型可能包含同种应用程序的流量,如 Chat 服务和 File 服务中都包含有 skype 应用的流量,这就为分类任务增加了难度。因此,针对服务类型的分类相较于应用程序的分类难度更大。

表 5　不同分类算法准确率对比

算法	分类模型	分类类型	Accuracy
文献[11]	CNN	VPN 流量 6 分类	0.929 2
文献[12]	1D-CNN	服务类型 12 分类	0.955 0
文献[13]	CNN	应用分类	0.980 0
		服务类型 12 分类	0.940 0
本文方法	基于特征融合和注意力机制的改进 CNN	服务类型 12 分类	**0.995 4**

由表 5 可知,本文所提模型在准确率上要优于其他文献所提的分类算法。相比于文献[13]针对应用程序的分类结果,本文模型在面对难度更大的服务类型分类任务时,在准确率上仍然具有 1.54%的优势。而同样是对服务类型流量进行 12 分类,本文模型相较于文献[12]采用的 1D-CNN 模型以及文献[13]所采用的 CNN 模型,在总体准确率上分别提升了 4.04%以及 5.54%。

3)在类不平衡流量上的分类性能及对比

为了评估本文所提模型在面对类不平衡数据集时的分类性能,实验选取了几种典型的基于深度学习的流量分类模型与本文所提模型在类不平衡数据集上进行了对比实验,其中包括未针对类不平衡现象做出相应改进的常规分类模型 1D-CNN、2D-CNN,以及利用 GAN 对少数类进行过采样并结合 CNN 进行分类的分类模型。通过对比不同分类模型在各个类别上的性能指标数值,对分类性能进行全面的评估。

图 5(a)～(c)分别展示了不同分类模型在数据集各类别上的准确率 Precision、召回率 Recall 和 F1-Score。从图 7 中可以看出,在面对多数类,如 File、Voip 时,4 种分类模型均能实现高精度的分类,1D-CNN 和 2D-CNN 的 F1-Score 均达到了 98.6%以上,本文所提模型则达到了 99.7%以上。然而,对于 P2P、Streaming、VPN_File、VPN_P2P 以及 VPN_Streaming 等少数类,1D-CNN 和 2D-CNN 分类效果不佳,其中 2D-CNN 在上述少数类上的 F1-Score 基本都低于 90%,对 VPN_File 其 F1-Score 只有 79.61%;而对于 1D-CNN 模型,在少数类上的 F1-Score 则基本在 90%～92%之间。本文所提模型在上述少数类上的 F1-Score 则均高于 95%,在 P2P 和 VPN_File 上则分别达到了 99.01%和 98.51%,相对于 1D-CNN 和 2D-CNN 在对少数类的分类性能上有显著地提升,该结果表明本文所提模型在面对类不平衡流量数据时的分类性能优于当前典型的 1D-CNN 和 2D-CNN 分类算法。作为当前处理类不平衡现象的典型方法,GAN+CNN 的组合模型通过对少数类进行过采样,降低不平衡比例,在一定程度上减

少了类不平衡对分类性能的影响，该模型在上述少数类上的 F1-Score 达到了 92%～96%，分类性能要优于 1D-CNN 和 2D-CNN。然而，相较于 GAN+CNN 的组合模型，本文所提模型在上述少数类上的 F1-Score 均有 2%～3.6%的提升，该结果表明本文所提模型在少数类上的分类效果要优于 GAN+CNN 的组合模型。综上所述，本文所提模型弥补了当前主流的流量分类模型的不足，在面对类不平衡流量数据中的少数类时，仍能表现出良好的分类性能，在面对其他针对类不平衡现象的分类算法时也体现出一定的优势。

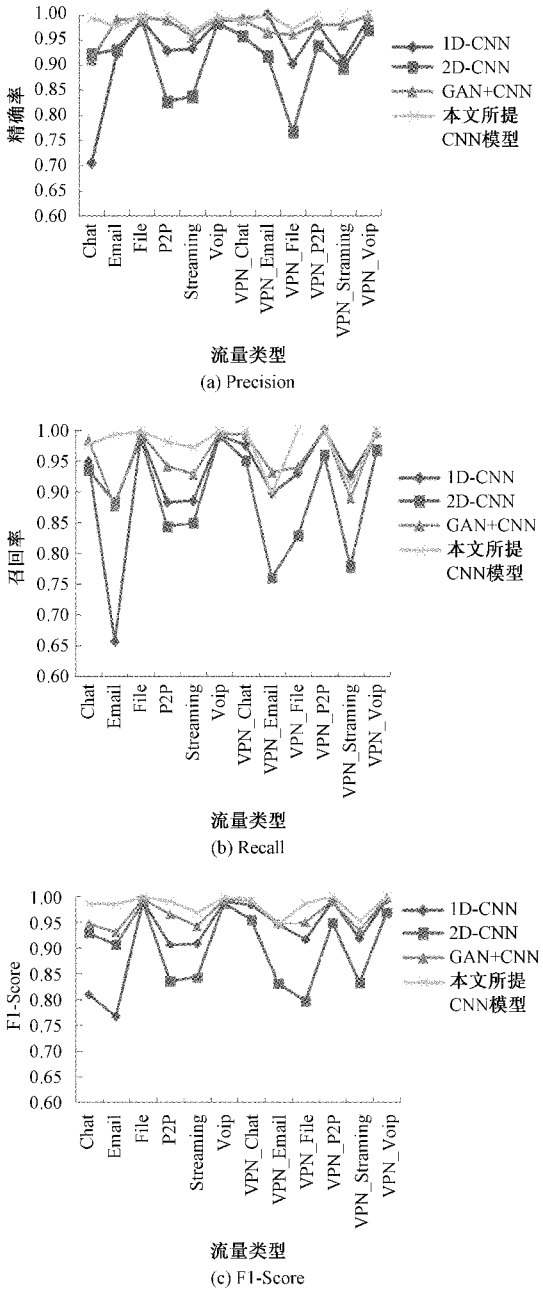

图 5　数据集中各流量类别上的 Precision、Recall、F1-Score

3　结　　论

本文针对当前典型的流量分类算法在面对类不平衡数据集中的少数类时分类性能不佳的问题，对传统的 CNN 模型进行改进，提出了一个基于特征融合与注意力机制的改进 CNN 模型。该模型通过添加 Inception 模块实现多尺度特征的融合，使模型能提取到更丰富的特征，提高对少数类的分类精度。与此同时，在 CNN 模型中设计并引入了一个通道-空间域注意力模块，对融合后的特征根据重要程度赋予相应的权值，使模型更多地关注到更有利于分类的特征，增强流量特征的表征能力，进一步提高分类性能。最后，用 1×1 卷积加全局平均池化层的组合代替传统 CNN 中的全连接层，大大地减少了网络参数，很大程度上避免了模型的过拟合，增强了模型的泛化能力。未来将考虑与重采样方法结合，进一步降低流量的类不平衡对分类器性能的影响。

参考文献

[1]　李艳霞，柴毅，胡友强，等. 不平衡数据分类方法综述[J]. 控制与决策，2019,34(4)：4-19.

[2]　王攀，陈雪娇. 基于堆栈式自动编码器的加密流量识别方法[J]. 计算机工程，2018,44(11):140-147,153.

[3]　PESCAPE A, DAINOTTI A, CLAFFY K C, et al. Issues and future directions in traffic classification[J]. IEEE Network: The Magazine of Computer Communications, 2012, 26(1):35-40.

[4]　FINSTERBUSCH M, RICHTER C, ROCHA E, et al. A survey of payload-based traffic classification approaches[J]. IEEE Communications Surveys & Tutorials, 2014, 16(2):1135-1156.

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

[6]　SHAO H, TANG L, DONG L, et al. A research of network applications classification based on deep learning [C]. Third International Conference, MLICOM 2018, Hangzhou, China, 2018: 13-21.

[7]　WANG P, CHEN X, YE F, et al. A survey of techniques for mobile service encrypted traffic classification using deep learning[J]. IEEE Access, 2019, 7:54024-54033.

[8]　WEI W, MING Z, ZENG X, et al. Malware traffic classification using convolutional neural network for representation learning [C]. 2017 International Conference on Information Networking (ICOIN), IEEE, 2017: 712-717.

[9]　吕品，潘思羽，许嘉，等. 基于机器学习的网络流量分类算法[J]. 广西大学学报(自然科学版)，2019,

44(6): 1650-1657.

[10] 朱文斌,马秀丽.多种构图方式下的加密流量分类[J].电子测量技术,2021,44(12):87-92.

[11] 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(9): 1-12.

[12] 郭帅,苏旸.基于数据流的加密流量分类方法[J].计算机应用,2021,41(5):1386-1391.

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

[14] 刘丹,姚立霜,王云锋,等. 面向类不平衡流量数据的分类模型[J]. 计算机应用, 2020,40(8):2327-2333.

[15] ZHANG F, SHANG T, LIU J. Imbalanced encrypted traffic classification scheme using random forest[C]. 2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics(Cybermatics),2020: 837-842.

[16] VU L, BUI C T, NGUYEN Q U. A deep learning based method for handling imbalanced problem in network traffic classification[C]. Proceedings of the Eighth International Symposium on Information and Communication Technology, 2017: 333-339.

[17] 顾兆军,吴优,赵春迪,等.流量的集成学习与重采样均衡分类方法[J].计算机工程与应用,2020,56(6):86-91.

[18] TANG Z, ZENG X, CHEN J. Multi-model coupling method for imbalanced network traffic classification based on clustering[J]. International Journal of High Performance Computing and Networking, 2020, 16(1):26-35.

[19] 龚捷,罗聪,罗琴.基于注意力机制和残差网络的动作识别模型[J].电子测量技术,2021,44(14):111-116.

[20] GERARD D G, ARASH H L, MOHAMMAD M, et al. Characterization of encrypted and VPN traffic using time-related features[C]. In Proceedings of the 2nd International Conference on Information Systems Security and Privacy(ICISSP 2016), 2016:407-414.

作者简介

林鹏,硕士研究生,主要从事多媒体与信息安全方面的研究。

E-mail:20191219028@nuist.edu.cn

翟江涛(通信作者),博士,副教授,主要从事多媒体与信息安全方面的研究。

E-mail: jiangtaozhai@nuist.edu.cn

许历隆,硕士研究生,主要从事多媒体与信息安全方面的研究。

E-mail:2837053804@qq.com

崔永富,硕士研究生,主要从事多媒体与信息安全方面的研究。

E-mail:20191218003@nuist.edu.cn