

自监督学习驱动的注意力增强恶意流量检测方法

孙剑文, 张斌, 李红宇, 常禾雨

(信息工程大学, 河南 郑州 450001)

摘要: 现有的基于深度学习的恶意流量检测方法普遍面临3个主要问题: 标记样本稀缺、恶意行为流量特征表示不够充分、行为关联模式整合效率低导致检测误报率高。针对以上问题, 提出一种端到端的注意力增强恶意流量检测 (malicious traffic detection with enhanced attention, MTAttention) 方法。该方法首先对网络行为流量的异构头部特征和有效载荷实施统一编码, 通过结构化的多包序列流量表示, 形成数据包级别的标准化流和会话表示; 其次, 基于MAE (masked auto-encoder) 模型采用自监督的掩码预训练策略, 利用视觉Transformer提取丰富的流量特征表示。通过选择性地关注输入数据包序列中不同部分的空间和变量间的依赖关系, 学习通用流量表示, 将编码器模型的权重参数用于下游任务初始化, 以加速模型训练的收敛过程; 最后, 引入基于通道注意力的数据包序列特征融合策略, 利用多维注意力机制和有标签数据对模型权重进行微调以适应流量识别和分类任务, 在分类决策前增强高权重特征的整合能力, 进一步提高模型检测精确度。实验采用CIC-IDS2017数据集, 结果表明, 在恶意流量识别和细粒度分类场景下, MTAttention方法的精确度平均可达98.7%, 推理速度高于每秒1590个样本。与基于MAE范式的改进方法FlowMAE相比, MTAttention在保持高精度的同时, 仅需1.56%的参数数量和63.89%的内存开销, 推理速度平均提升约一倍, 且模型大小仅为5.17 MB。

关键词: 流量表示; 结构化; 自监督学习; 注意力增强; 流量检测

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2025025

Harnessing self-supervised learning to boost malicious traffic detection with enhanced attention

SUN Jianwen, ZHANG Bin, LI Hongyu, CHANG Heyu

Information Engineering University, Zhengzhou 450001, China

Abstract: The existing deep learning-based malicious traffic detection methods generally suffered from three main problems: labeled sample scarcity, inadequate representation of malicious behavior traffic features, and a high false

收稿日期: 2024-09-23; 修回日期: 2025-02-28

通信作者: 孙剑文, jianwensun_edu@163.com

基金项目: 信息工程大学密码工程学院研究生创新基金 (2019f113)

Foundation Item: Innovation Fund for Graduate Students of the Cryptography Engineering Institute, Information Engineering University (2019f113)

引用格式: 孙剑文, 张斌, 李红宇, 等. 自监督学习驱动的注意力增强恶意流量检测方法[J]. 网络与信息安全学报, 2025, 11(2): 136-151.

Citation Format: SUN J W, ZHANG B, LI Hongyu, et al. Harnessing self-supervised learning to boost malicious traffic detection with enhanced attention[J]. Chinese Journal of Network and Information Security, 2025, 11(2): 159-174.

positive rate due to ineffective integration of behavioral association patterns during detection. To address these issues, an end-to-end malicious traffic detection method named MTAttention was proposed. The heterogeneous header features and payloads of network behavior traffic were unified and encoded, resulting in a standardized flow and session representation at the packet level through structured multi-packet sequential traffic representation. Based on the MAE model, a self-supervised mask pre-training strategy was adopted, and a visual Transformer was utilized to extract rich traffic feature representations. By selectively focusing on the spatial and intervariable dependencies among different parts of the input packet sequence, a general traffic representation was learned, and the weight parameters of the encoder model were used to initialize subsequent tasks, which accelerated the model training process. A packet sequence feature fusion strategy based on channel attention was introduced, and a multi-dimensional attention mechanism, along with labeled data, were used to fine-tune the model's weights to adapt to traffic identification and classification tasks. Before making classification decisions, the integration ability of high-weight features was enhanced to further improve the model's detection precision. The CIC-IDS2017 dataset was employed for the experiment, and the results showed that in the multi-classification scenario for malicious traffic identification, the precision of the MTAttention method could reach an average of 98.7%, and the inference efficiency exceeded 1590 samples per second. Compared with Flow-MAE, an improvement over the MAE paradigm, MTAttention maintained high precision while requiring only 1.56% of the parameter count and 63.89% of the memory overhead of Flow-MAE. Additionally, the average inference speed enhancement was approximately 100%, and the model size was only 5.17 MB.

Keywords: traffic representation, structuring, self-supervised learning, attention enhancement, traffic detection

0 引言

随着网络技术的快速发展,网络威胁也在增加和演变。因此,如何有效检测恶意流量,逐渐成为网络安全领域的关键问题之一。

恶意流量检测分为主动检测和被动检测^[1]。主动检测技术包括明文检测,使用传统的深度包检测或签名、规则匹配来识别恶意流量^[2]。虽然主动检测技术的准确率较高,但这类检测在应对持续演变的攻击模式时缺乏适应性和鲁棒性。被动检测技术侧重于分析流量特征,可以基于原始流量特征、侧信道特征等对原始流量数据进行表示,突出网络行为流量中的重要特征。该过程通过字节级别、包级别、流级别、会话级别等自定义流量表示粒度,构建正常或异常流量的统计特征、指纹特征、内容特征、行为特征等^[3-6],然后通过单一或混合的深度学习模型自动提取特征并建立流量的基线模式,进行恶意流量检测^[7-11]。这类检测的主要挑战是需要大量标记数据来训练模型。然而,对于动态多变的网络流量而言,与恶意流量检测任务相关的有标记数据是稀缺的。

基于自监督学习的异常检测方法通过设计预训练任务,使模型在无须大量人工标注数据情况

下,与下游任务的标签数据建立紧密联系^[12-13]。基于自监督学习的流量异常检测方法分为两类。一是基于样本对之间关系学习的对比学习方法^[14-16]。例如,Rosetta方法^[14]通过模拟不同网络环境中TCP机制动态调整生成流量变种,进而采用TCP感知的流量增强机制提升模型性能。该方法在提取TCP流量的鲁棒特征时有意识地忽略了一些受环境因素影响但易于分类的特征(如延迟、丢包率等),忽略这些特征会降低在与训练环境相似网络中的最高性能。为这类对比方法选择合适的正负样本对,对模型的性能至关重要,同时这种对比方法需要大量计算资源进行支撑。二是基于输入样本的某种变换后重构原始输入的自预测方法^[17-20]。例如,遮蔽策略中,一类是基于BERT范式的流量分类方法,如ET-BERT方法^[17]利用字节对编码技术将网络流量字节流转换为BERT输入令牌(Token),并通过预训练微调的方法取得了良好性能,然而,与文本处理任务中依赖上下文解析语义的需求不同,恶意流量数据中的字节往往缺乏显性的高级语义单元,使得从底层字节级数据中直接解析并辨识出高级语义特征的难度增加;另一类是基于掩码自编码器(masked auto-encoder, MAE)范式^[18]的流量分类

方法, 该类方法从视觉方法的角度进行流量表示和流量分割(采用视觉方法通常将单个数据包转换为图像格式作为输入^[21-22]), 进而将其编码为补丁(Patch)或Token。模型通过随机遮蔽输入“图像”的部分区域, 然后在自编码器结构的“编码-解码”架构下, 通过编码器提取“图像”的潜在结构与深层特征, 再通过解码器重构完整的原始“图像”^[19]。训练后的自编码器结构中的编码器部分将被用于面向下游任务微调, 进行流量分类^[20]。

自预测方法的预训练任务通常借助大量无标签数据, 在结构复杂且参数量庞大的模型上开展训练, 旨在获取通用的流量表示。但在设备资源和时间成本受限的环境下, 该方法难以有效实施。此外, 基于图像格式的单个数据包表示的流量检测方法, 虽然能实现及时响应, 却未充分考虑流量序列的动态变化及时间依赖性, 导致存在较高的误报率和漏报率。为了在有限的资源约束下, 使模型得到充分训练并具备较强的泛化性能, 从而提升检测精确度, 本文提出了一种端到端的注意力增强恶意流量检测(malicious traffic detection with enhanced attention, MTAttention)方法。该方法通过捕捉流量中的数据包局部特征及其时序位置关系, 以学习流量行为中的内在表示与显性动态关系, 从而提高流量分类的精确度。具体地, 本文优化了基于MAE预训练的特征学习编码器模块, 使其适应输入流量数据的结构特点。同时, 在模型微调阶段, 通过引入通道注意力机制, 动态调整流量潜在特征表示的各部分权重分配, 从而增强模型对不同特征通道重要性的学习能力, 使多包序列在模型内部为分类提供类似于投票机制的权重依据, 从而增强单个模型内部集成预测的可靠性, 而无须依赖集成模型的复杂性。

实验结果表明, 相比现有方法, 本文方法能够显著提升检测性能, 并具备一定的鲁棒性。本文的主要贡献如下。

1) 提出一种灵活、易扩展且具备一定可解释性的流量表示方法, 该方法通过将网络流量中的异构头部特征和有效载荷编码为统一格式,

实现在数据包层面引入流和会话的行为特征, 弥补传统数据包级别表示方法的不足, 能够处理不同长度的流量, 增强流量表示的深度和广度。

2) 设计并实现一种高效的恶意流量识别和分类框架, 该框架继承预训练编码器模型的优势, 并通过在未标记的数据上进行自监督掩码重构预训练, 学习通用的流量潜在表示。这种训练方法无须标注数据, 可以大幅降低数据获取的成本, 同时提高模型对未知环境的适应能力。预训练得到的编码器模型权重参数用于下游任务初始化, 可以利用更少的有标记样本进行微调, 加速模型训练的收敛过程, 从而提升模型的检测效果。

3) 增强模型分类决策前对数据包间特征序列中高权重特征的整合能力。通过在分类决策前引入基于通道注意力的数据包序列特征融合策略, 显著提升了模型的检测精确度, 优化了推理速度。

4) 在CIC-IDS2017数据集^[23]上的实验结果表明, MTAttention方法在不平衡和有限平衡样本条件下微调, 均能够稳定地识别和分类恶意流量, 对攻击变种亦展现出较高的分类精确度。其细粒度分类精确度优于混合模型等利用大量有标签数据训练的方法。同时, 与新的基于MAE范式的预训练-微调检测方法相比, MTAttention方法在保持高性能的同时, 模型大小仅为5.17 MB, 更适合在资源受限的环境中部署。

1 多层次结构化流量表示方法

1.1 数据预处理

本文中的网络流量表示方法基于数据包序列, 其中每个数据包的特征是通过分析其头部和有效载荷来提取的。

首先, 利用五元组(源IP地址、目的IP地址、源端口、目的端口、传输层协议类型)对pcap格式的原始流量进行会话(双向流)划分, 以识别两个端点间的交互行为。基于IPv4协议规范结构化的特征提取方法, 通过固定位置模式从数据包头提取关键特征, 包括但不限于数据包时间间隔、IP数据包长度、生存时间、校验和、目的端口、协议类型、TCP标志位、传输层有效载荷长度以及数据包

方向。若字段在协议中未定义,则在特征向量中填充0。

接着,将数据包有效载荷附加至头部特征,考虑到数据包向量被转换为 28×28 的矩阵,因此在784 byte处进行截断,以形成标准化输入,即单个数据包表示为 28×28 的矩阵(实验部分的单个检测方法采用网络流量表示方法)。

最后,将按照时间顺序组合而成的数据包序列进行整合,以表示会话流量。预处理后的数据包序列流量表示涵盖字节级别、包级别以及流级别的多维特征,通过跨层次的流量表示方法,更全面地捕捉原始流量行为模式。网络流量表示预处理方法如图1所示。

1.2 预处理对流量数据表示的影响

数据包在转换过程中被塑造为类似于图像的矩阵格式,保留了相对位置特征信息,为后续深入挖掘每个数据包的独立特征及其在序列中的上下文依赖性提供了基础。通过流量数据预处理形成原始流量的特征向量,该过程保持了数据包头部等特征的顺序和位置的一致性,有助于增强模型可解释能力,使得模型易于识别每个特征对预测结果的贡献度,并具有一定的特征可扩展性,使模型具备纳入新特征或协议扩展的能力。此外,将原始数据包转换为类似图像的矩阵形式,也为采用基于深度学习的图像处理技术进行进一步的流量特征提取与分析打下了基础。通过将数据包的有效载荷附加至头部特征,并在固定字节处进行截断,构成标准化输入,以保留关键信息。在此基础上,通过

可调的卷积嵌入方式,使模型自动提取对分类准确性重要的特征。这种方法支持模型在无须分析具体内容的情况下,从流量中获取模式特征。

预处理步骤通过整合数据包序列信息,形成综合的特征表示,使模型在分类时考虑会话流中上下文信息。这种跨层次的流量表示方法能够更全面地捕捉原始流量行为的维度。在分类决策阶段,该数据整合方式类似于多个独立判断的综合,能够使分类结果更加稳定,有效减少误报。

2 MTAttention 方法

基于自监督学习的预训练-微调检测框架如图2所示,总体上是一种基于MAE的“预训练-微调”框架。①自监督预训练阶段,首先对输入到编码器中的流量表示进行随机掩码,通过编码器对掩码后的流量表示进行特征提取,再通过解码器重构完整的流量表示。通过上述编码-解码过程,使模型学习到流量中的潜在结构和深层特征,以便用于下游任务。②有监督微调阶段,将带标签的流量数据进行预处理后,直接送入训练后的编码器部分,对其进行面向恶意流量检测任务的微调。特别地,为了更好地融合数据包序列特征,微调阶段引入了通道注意力机制,以进一步增强模型对不同特征通道重要性的学习能力,从而提高检测精确度。③检测阶段,编码器的参数取自微调阶段验证集上最优的模型权重,并冻结参数。模型结构与微调阶段一致,直接对新输入的流量数据进行特征提取和分类,输出检测结果。

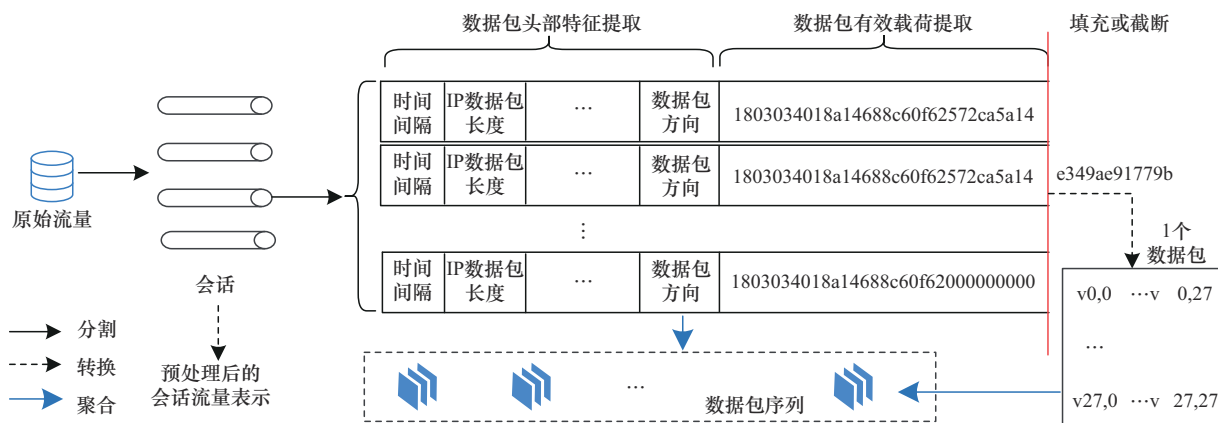


图1 网络流量表示预处理方法

Figure 1 Preprocessing methods for network traffic representation

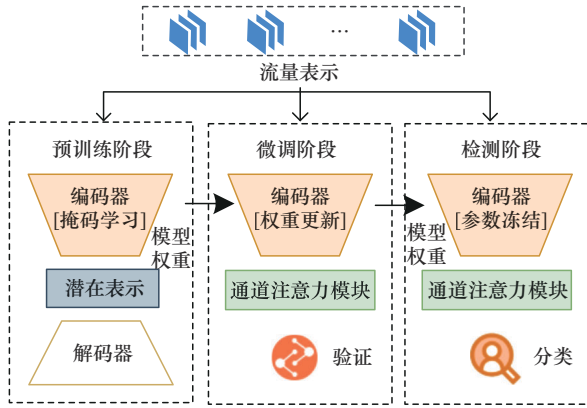


图2 基于自监督学习的预训练-微调检测框架
Figure 2 Self-supervised pre-training & fine-tuning detection framework

下面分别对以上预训练阶段和微调阶段进行详细介绍, 并对检测模型效能进行理论分析。

2.1 基于MAE的预训练过程

使用无标签流量数据进行预训练的目的是从中提取具有良好泛化性的流量表示, 使模型具备捕捉跨类别通用表示的能力, 从而实现对流量数据的全面理解和表征。该阶段基于掩码自编码器模型, 通过自监督的方式进行预训练。

MTAttention 预训练模型如图3所示。

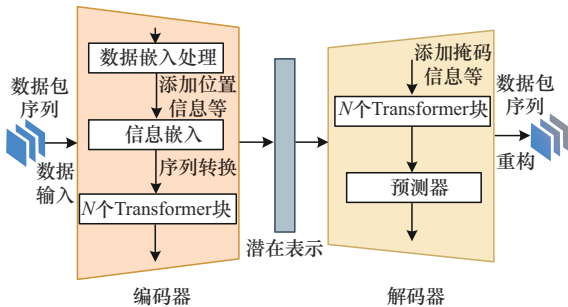


图3 MTAttention 预训练模型
Figure 3 MTAttention pre-training model

编码器部分是一个视觉Transformer结构^[24], 其中使用了位置编码和8个Transformer块。输入序列被均匀分割并嵌入Token, 通过正弦-余弦位置编码来明确Token序列之间的位置关系。每个Transformer块中集成了多头自注意力机制和前馈神经网络^[25]。自注意力机制使模型在编码过程中并行处理输入序列的所有位置, 前馈神经网络对注意力机制的输出进行进一步的非线性变换。这种层叠结构使模型能够进行深度特征提取, 有

效捕获序列变量间的复杂相关性, 从而使模型学到数据包序列的潜在表示。为了辅助模型识别序列的起始边界, 在输入流量特征表示序列的起始位置引入类别Token^[24], 该Token集中编码了输入序列的整体信息, 可用于连接后续的分类头。对编码器部分的输入数据进行随机遮蔽, 使其专注于未被遮蔽部分, 从而促进模型对潜在表示进行学习。

解码器接收编码器输出的潜在表示, 并与位置Token结合以初始化其内部状态。解码器内部的Transformer块在结构上与编码器中的对应块相似, 区别在于需要处理被遮蔽的Token。解码器通过自注意力机制和前馈神经网络协同作用, 逐步预测并重构包含遮蔽部分的流量特征表示。解码器的输出是重构的流量表示。

MAE模型训练过程的目标是最小化重构误差, 即通过减少输入数据和重构数据之间的差异, 使模型学习流量数据中的通用表示。基于MAE的预训练过程如算法1所示。

算法1 基于MAE的预训练过程

输入 原始数据序列 inputData, 模型参数 params
输出 训练后的模型参数 trainedModel

- ① 初始化模型和优化器
encoder, decoder <- 初始化编码器和解码器 (params)
optimizer <- 初始化优化器
- ② 训练模型
FOR epoch FROM 1 TO params. numEpochs
DO
FOR batch IN inputData BATCHSIZE
params. batchSize DO
③ 数据预处理 (初始化数据序列、位置嵌入、添加类别Token)
tokenizedData <- tokenEmbedding(batch)
embeddedData <- positionalEncoding(tokenizedData)
augmentedData <- classTokenAddition(embeddedData)
④ 编码器处理数据序列
(maskedData, mask, idsRestore) <- en-

```
coder. randomMasking(augmentedData, params.
maskRatio)
```

```
encoderOutput <- encoder. forward
(maskedData)
```

⑤ 解码器处理潜在表示

```
decoderInput <- positionalEncoding(to-
kenEmbedding(encoderOutput))
```

```
decoderOutput <- decoder. forward(de-
coderInput, idsRestore)
```

⑥ 重构与损失计算

```
reconstructedData <- predictor(decoder-
Output)
```

```
loss <- computeLoss(reconstructedData,
batch)
```

⑦ 优化步骤

```
optimizer.zeroGrad()
```

```
loss.backward()
```

```
optimizer.step()
```

⑧ 检查保存条件

```
IF epoch MOD params.saveInterval == 0
```

THEN

```
saveModelCheckpoint(trainedModel, ep-
och)
```

```
ENDIF
```

```
ENDFOR
```

```
ENDFOR
```

⑨ 训练完成检查

```
IF loss < params.tolerance THEN
```

```
BREAK
```

```
ENDIF
```

⑩ 返回训练后的模型

```
RETURN trainedModel
```

2.2 基于通道注意力的微调过程

经过预训练的自编码器架构中，编码器部分具备良好的特征提取能力，可通过拼接分类层，开展面向下游流量分类数据的微调过程，使其具有流量分类能力。该阶段采用有监督微调方法，通过进一步优化编码器的参数，提升其在恶意流量识别和分类任务中的适应性。为了增强模型对序列数据中关键信息的提取能力，在进入分类层前，引入通道注意力机制，为关键特征分配更高

的权重^[26]，进一步提升微调性能。MTAttention检测模型如图4所示，主要部分描述如下。

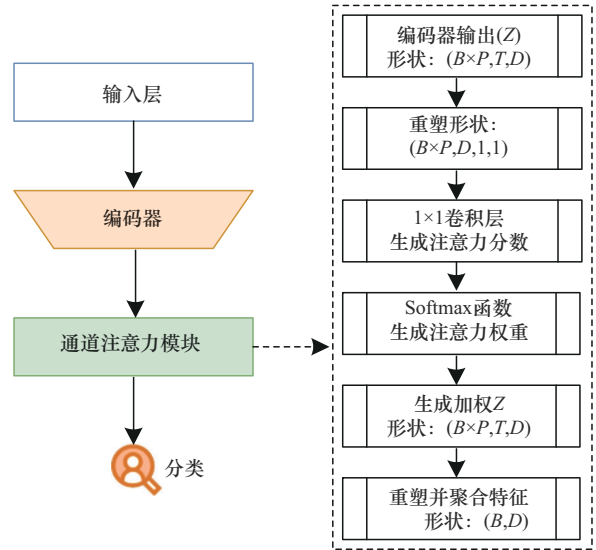


图4 MTAttention检测模型

Figure 4 MTAttention detection model

输入层：输入层将网络流量以数据包序列的形式转换为特征向量，输入形状为 $[B, P, C, H, W]$ ，其中 B 是批次大小， P 是每个批次中的数据包数量， C 是通道数， H 和 W 分别是数据包数组的高度和宽度。

输入层的作用是将输入数据调整为编码器期望的形状： $\mathbf{x}_{\text{adjusted}} = \text{reshape}(\mathbf{x}, [B \times P, C, H, W])$ 。

编码器：编码器为输入的特征向量生成潜在表示。将 $\mathbf{x}_{\text{adjusted}}$ 数据维度转换至 $[B \times P, T, D]$ ，其中 T 是每个数据包内部的序列长度， D 是潜在特征维度，调整后输出为： $\mathbf{z}_{\text{reshaped}} = \text{reshape}(\mathbf{z}, [B, P, T, D])$, $\mathbf{z} \in \mathbb{R}^{B \times P \times T \times D}$ 。

通道注意力模块：将通道注意力机制^[27]用于加权编码器的输出以提升关键特征的识别度。对于每个数据点，编码器输出 $\mathbf{z}_{\text{reshaped}}$ 中，每个潜在特征向量 $\mathbf{z}_{b,p,t}$ 经由 1×1 卷积映射至与特征维度 D 等长的向量，生成注意力分数：

$$\text{attention}_{\text{scores}} = \text{attention}(\mathbf{z}_{\text{reshaped}}),$$

$$\text{attention}_{\text{scores}_{b,p,t,d}} = \mathbf{W}_d \cdot \mathbf{z}_{b,p,t} + b_d$$

其中， b, p, t, d 分别是批次索引、数据包索引、序列位置索引和特征维度索引， $\mathbf{W}_d$ 是可学习的权重矩阵， b_d 是偏置项。

注意力权重 $\text{attention}_{\text{weights}_{b,p,t,d}}$ 与每个潜在特征

向量按元素相乘, 产生加权特征 $\text{weighted}_{z_{b,p,d}}$ (记为 weighted_z), 使模型能够在不同数据包和序列索引之间动态调节关注点。聚合加权实现对特征序列与数据包序列的综合分析 $\text{aggregated}_{z_b} = \frac{1}{p \times T} \sum_{p=1}^P \sum_{t=1}^T \text{weighted}_z$, 使得主要信息得以保全。

分类层: 分类层接收聚合的特征表示为 aggregated_{z_b} 并映射至分类任务的输出空间, 以预测各个类别的概率分布: $\hat{y} = \text{classifierHead}(\text{aggregated}_{z_b})$ 。分类层由多个全连接层构成, 其中包含批量归一化、ReLU 非线性激活函数和 Dropout 策略等, 以进行进一步的非线性变换和模型正则化, 避免模型微调过程出现过拟合问题。

2.3 微调阶段检测模型效能分析

问题设定: 假设有 N 个独立同分布的数据包, 数据包的特征向量表示为 $\mathbf{z}_i \in \mathbb{R}^D$, 输入数据集 $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$, 通过融合函数 g 获得融合特征向量 $\mathbf{z}_{\text{agg}} = g(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N)$ 。在多分类任务中, 目标是最大化给定目标类别 y 的联合似然 $p(X|y)$ 。其他符号定义如下: $f(\cdot)$ 表示模型的函数, θ 表示模型参数。编码器将每个数据包特征向量 $\mathbf{z}_i$ 通过编码器转换为潜在表示 $\mathbf{h}_i = f_{\text{enc}}(\mathbf{z}_i; \theta_{\text{enc}})$, 其中 f_{enc} 是编码器函数, θ_{enc} 是编码器参数。

推断引入多包序列学习方法和通道注意力机制后, 融合特征向量 $\mathbf{z}_{\text{agg}}$ 相较于单个数据包的特征向量 $\mathbf{z}_{i,d}$ 更有助于检测和分类任务, 即推断 $\text{Acc}(\mathbf{z}_{\text{agg}}) > \text{Acc}(\mathbf{z}_{i,d})$ 。

2.3.1 多包序列学习方法分析

在独立同分布假设下, 由于每个数据包对联合似然的贡献独立, 联合似然实际上是单个数据包似然的乘积。给定类别 y 的联合似然可以表示为:

$$p(X|y) = \prod_{i=1}^N p(\mathbf{z}_i|y) \quad (1)$$

采用这种形式, 模型能从数据中学习更多的信息, 以提高分类准确性。

模型参数 θ 的学习通过最大化联合似然来实现, 最优估计值 θ^* 为:

$$\theta^* = \text{argmax}_{\theta} \sum_{i=1}^N \log p(\mathbf{z}_i|y; \theta) \quad (2)$$

等价于最小化负对数似然:

$$\theta^* = \text{argmin}_{\theta} - \sum_{i=1}^N \log p(\mathbf{z}_i|y; \theta) \quad (3)$$

使用梯度下降算法优化, 参数更新规则为:

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta} \left(- \sum_{i=1}^N \log p(\mathbf{z}_i|y; \theta_t) \right), \quad (4)$$

其中, η 是学习率, ∇_{θ} 表示梯度运算符。由于每个数据包提供了独立的梯度信息, 梯度方差会随着数据包数量增加而减少, 故更快收敛到全局最优解。

由于独立同分布, 多包序列学习方法的梯度表达式为:

$$\nabla_{\theta} \mathcal{L}(\theta) = \sum_{i=1}^N \nabla_{\theta} \left(- \log p(\mathbf{z}_i|y; \theta) \right) \quad (5)$$

该式反映了模型参数对每个数据包似然的敏感度。每个数据包都对梯度有贡献, 因此增加了梯度估计的稳定性。

2.3.2 通道注意力机制融合特征分析

根据 softmax 函数在注意力机制中的应用^[13], 为潜在表示 $\mathbf{h}_i$ 的每个通道维度分配注意力权重 $\mathbf{a}_{(i,d)}$:

$$\mathbf{a}_{(i,d)} = \frac{\exp(\mathbf{w}_d^T \mathbf{h}_i + \mathbf{b}_d)}{\sum_{d'=1}^D \exp(\mathbf{w}_{d'}^T \mathbf{h}_i + \mathbf{b}_{d'})} \quad (6)$$

其中, 通道索引 d 与总通道数 D 为对应特征的通道索引和维度。

通过这种方式, 得到由各个通道权重组成的向量 $\mathbf{a}_i = (\mathbf{a}_{(i,1)}, \mathbf{a}_{(i,2)}, \dots, \mathbf{a}_{(i,D)})$ 。注意力加权后的潜在表示为 $\mathbf{h}'_i = \mathbf{h}_i \odot \mathbf{a}_i$, 其中 $\odot$ 表示向量的元素乘法。加权机制使得模型能够更加关注对分类任务有贡献的特征。

聚合加权特征向量定义为 $\mathbf{z}_{\text{agg}} = \frac{1}{N} \sum_{i=1}^N \mathbf{h}'_i$, 分类器基于聚合特征向量进行分类预测为 $\hat{y} = g(\mathbf{z}_{\text{agg}}; \theta_c)$, 其中, θ_c 为分类器的参数。

引入通道注意力机制后, 注意力加权的损失函数定义为:

$$\mathcal{L}_{\text{attn}}(\theta, \mathbf{W}) = \mathcal{L}(\theta) + \lambda \sum_{i=1}^N \sum_{d=1}^D (\mathbf{a}_{(i,d)} \log \mathbf{a}_{(i,d)}) \quad (7)$$

其中, 第二项是熵正则化项, λ 是控制正则化项影响的参数。熵正则化项鼓励模型注意力分布更加分散, 从而促进模型对特征的全面理解。通过调整各个通道的权重 $\mathbf{a}_{i,d}$, 可以影响损失函数的值。当模型优化损失函数时, 实质是寻找一组参数使得预测结果 $\hat{y}$ 尽可能接近真实标签 y , 同时注意力权重自适应地分配到各个通道上。这使模型更关注重要的特征通道, 进而使特征向量易于检测和分类。

注意力机制的引入为模型提供了额外的梯度信号,使模型能够更快地调整参数。采用梯度下降算法最小化损失函数 $\mathcal{L}_{\text{attn}}(\theta, \mathbf{W})$,损失函数相对于注意力权重 $\mathbf{W}$ 的梯度为:

$$\nabla_{\mathbf{W}} \mathcal{L}_{\text{attn}}(\theta, \mathbf{W}) = \nabla_{\mathbf{W}} (\mathcal{L}(\theta) + \lambda \sum_{i=1}^N \sum_{d=1}^D (\mathbf{a}_{(i,d)} \log \mathbf{a}_{(i,d)}) \quad (8)$$

综上,模型的学习过程速度较快,并提高分类性能,使得 $\text{Acc}(\mathbf{z}_{\text{agg}}) > \text{Acc}(\mathbf{z}_{i,d})$ 的可能性增大。

3 实验设计

3.1 实验设置

实验服务器配置为 Intel Xeon Silver 4210R CPU @ 2.4 GHz, 128 GB 内存, RTX 5000 显卡 (16 GB 显存)。训练和测试实验均在 Ubuntu 22.04.1 系统上完成,主要开源软件库版本为 PyTorch 1.12.1。

基于 MAE 模型的自监督学习策略,在预训练阶段,遮蔽比例为 0.6,批大小为 256,训练次数为 200 次;编码器和解码器均采用 AdamW 优化器,初始学习率设为 0.000 1,并辅以余弦衰减调度器动态调整学习率;编码器和解码器深度分别为 8 和 4,多头注意力机制头数为 8;损失函数基于均方误差构建。在微调阶段,批大小为 64,学习率为 0.000 5,使用交叉熵损失函数。将微调阶段各类样本按照 7:2:1 的比例划分为训练集、验证

集和测试集,通过分层随机抽样方法^[28],确保各类数据在每个集合中的样本分布与整体数据集保持一致,即仍然是不平衡分布状态。训练 50 次,基于验证集的准确率选择最优模型进行测试集评估。实验中模型采用统一的初始参数设置。

3.2 原始流量预处理

网络入侵检测领域的开源数据集 CIC-IDS2017 包含原始网络流量 pcap 文件。根据提供的数据集描述文档对 pcap 文件中的混合流量进行类别划分。其中,正常流量是混合的,包括加密流量和非加密流量;异常类分别为 Web 攻击 (WebAttack) 的暴力破解 (BruteForce), 跨站脚本攻击 (XSS), SQL 注入 (SqlInjection), 以及 FTP 和 SSH 的 Patator 攻击, 各种拒绝服务攻击 (DoS-Slowloris、DoS-Slowhttptest、DoS-Hulk、DoS-GoldenEye), 分布式拒绝服务攻击 (DDoS), 渗透测试 (Infiltration), 心脏滴血漏洞 (Heartbleed)、僵尸网络 (Botnet) 以及端口扫描 (PortScan)。CIC-IDS2017 数据集样本的不平衡分布如图 5 所示。

为了避免数据预处理过程不一致带来的性能评估偏差^[29],实验中对原始流量类别划分和预处理方式保持一致。数据集中攻击类别的样本量存在显著不平衡,如 Heartbleed、Infiltration 的会话样本分别仅有 11 个和 35 个。然而,当这些样

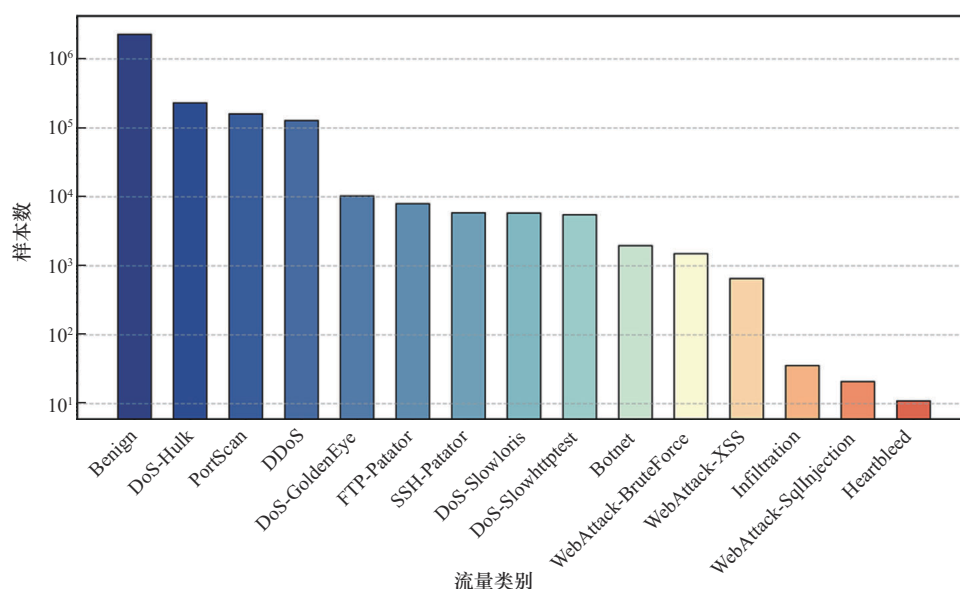


图5 CIC-IDS2017数据集样本的不平衡分布

Figure 5 Imbalanced distribution of CIC-IDS2017 dataset samples

本转换成以 8 个数据包为一组的序列时, 样本量分别增加到 6 155 组和 7 455 组。相关研究表明, 通常 5~10 个数据包能够进行有效的流量分类^[30]。

基于此, 本文在综合考虑特征表达完整性和计算复杂度的基础上, 选择 8 个连续数据包作为多包序列的流量表示, 旨在设备和资源允许条件下, 使模型充分提取多包序列的关键特征。考虑到数据包序列学习方法能够在形式上扩充会话级的样本量, 实验使用所有样本类别。具体来说, 7 个攻击类别和一部分正常样本被用作预训练数据, 其余 7 个攻击类别和另一部分正常样本则用于微调和检测, 使实验数据具备独立性, 从而验证在领域内有限数据条件下, 预训练模型学习到的通用表示的泛化性, 以及通过这些表示构建的下游检测任务模型的泛化能力。预处理后的实验样本类别分配如表 1 所示。

3.3 评估指标

采用异常检测领域内广泛认可的 5 个关键指标来评估模型的性能, 依次为准确率 (Accuracy)、精确度 (Precision)、召回率 (Recall)、F1 分数 (F1) 和误报率 (false positive rate, FPR)。

对于每个类别 i , 有:

$$\text{Accuracy}_i = \frac{\text{TP}_i + \text{TN}_i}{\text{TP}_i + \text{FP}_i + \text{TN}_i + \text{FN}_i} \quad (9)$$

$$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \quad (10)$$

$$\text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \quad (11)$$

$$F1_i = \frac{2 \times \text{Recall}_i \times \text{Precision}_i}{\text{Recall}_i + \text{Precision}_i} \quad (12)$$

$$\text{FPR}_i = \frac{\text{FP}_i}{\text{FP}_i + \text{TN}_i} \quad (13)$$

其中, 真正类 (TP_i)、真负类 (TN_i) 分别代表模型预测和实际结果一致的正类和负类样本数; 假正类 (FP_i)、假负类 (FN_i) 则分别代表模型预测与实际结果不一致的正类和负类样本数, 即误报和漏报情况。

模型效率的评估通过以下几个维度进行衡量: 参数量、计算复杂度、推理速度和推理时间。描述如下: $\text{Params} = \text{ProfileParams} + \text{ModelParams}$, $\text{Flops} = \text{ProfileFlops}$, $\text{InferenceSpeed} = \frac{\text{TSS}}{\text{IT}}$ 。其中, ProfileParams 为通过 `thop` 库的 `profile` 函数计算得到的参数, ModelParams 为模型中所有参数的集合, ProfileFlops 为通过 `profile` 函数计算模型在给定输入下所需的浮点运算次数, TSS 为总样本量, IT 为所有迭代推理时间的累加和。

4 实验结果与分析

实验中设置两个主要场景: 场景 1 为恶意流量细粒度分类, 不包含正常流量; 场景 2 为恶意流量识别与分类场景, 包含正常流量。

4.1 多包序列与单包检测方法性能对比分析

实验主要关注流量表示方法及特征融合策略对模型性能的具体影响。该定量消融实验, 以单个数据包的流量表示方法作为基础的流量分析手段。3 种方法依次比较流量表示带来的性能改变、新模块带来的性能改变, 从而验证所提方法的有效

表 1 实验样本类别分配

Table 1 Allocation of experimental sample categories

预训练阶段		微调和测试阶段	
流量类别	包序列/组	流量类别	包序列/组
DDoS	104900	Botnet	1 111
DoS-GoldenEye	9008	Dos-Holk	103 955
FTP-Patator	11944	Dos-Slowhttptest	901
PortScan	208	Dos-Slowloris	3 812
SSH-Patator	18563	Heartbleed	6 155
WebAttack-SqlInjection	12	Infiltration	7 455
WebAttack-XSS	698	WebAttack-BruteForce	2 801
Benign	17716	Benign	18 858

性。①单包方法基于本文提出的预训练-微调检测框架，预训练模型参数一致，区别在于输入框架的流量表示非数据包序列；在微调阶段直接添加分类器，作为基线模型。②多包序列方法与单包方法的区别为输入框架的流量表示不同。③多包序列（通道注意力）方法在方法②的分类器前引入通道注意力模块，替换其全连接层（注：定量消融实验之后，“多包序列”方法即代表本文所提方法MTAttention）。

4.1.1 不平衡样本条件下的微调检测性能对比分析

评估多包序列检测方法及引入通道注意力模块后在恶意流量分析中的性能优势。关注两个关键场景下的性能表现：图6(a)为场景1，评估模型对不同恶意流量的细粒度分类能力；图6(b)为场景2，评估模型对恶意流量的识别与分类能力，分别进行3组实验。

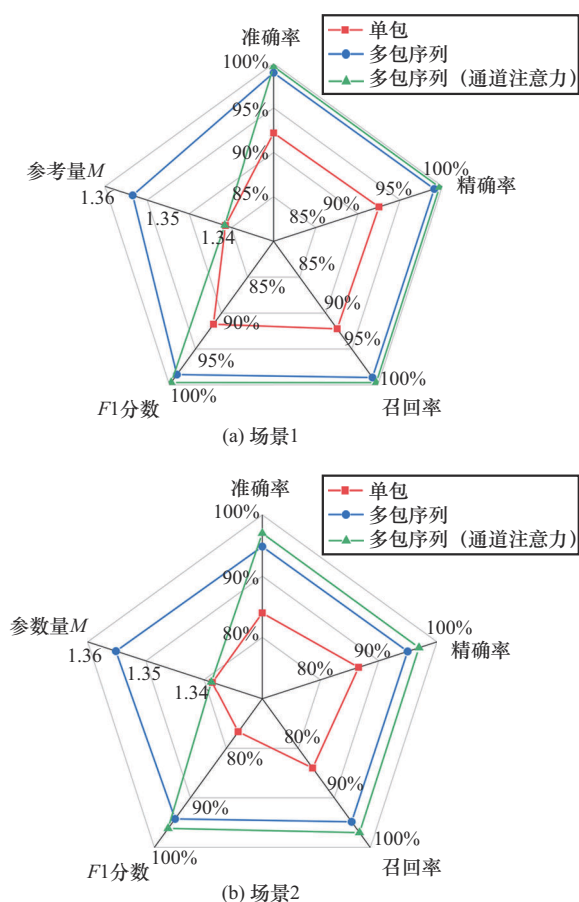


图6 不平衡样本下不同方法的检测对比分析

Figure 6 Comparison and analysis of different methods under unbalanced samples for detection

在场景1中，单包表示的检测方法已经取得较好的性能，其准确率、精确度和召回率均达到92%以上，说明模型的整体预测能力、预测可靠性和识别特定类型攻击的能力均得到体现，然而，当正常流量数据被纳入识别和分类场景时，分类器的性能受到了显著干扰。攻击行为在数据包级别与正常行为难以区分，这突显了考虑数据包上下文关联的重要性。

采用多包序列流量表示方法进行预训练，微调模型在场景1中的精确度和F1分数分别显著提升7.08%和7.65%；在场景2中性能提升更为显著，精确度和F1分数分别提升9.72%和22.92%。在场景2这种包含多样性正常流量的复杂环境中，可以观察到多包序列方法相比单包方法在4个检测性能指标上受环境变化的影响小。从F1分数来看，单包方法降低了14.88%，而多包序列方法仅降低了3.49%。上述结果表明了多包序列表示方法的鲁棒性，以及在提升模型性能方面的有效性，但代价是模型参数量增加（1.23%）和计算复杂度显著上升（约5倍）。

将通道注意力机制应用于特征融合阶段，改进后的模型在两个场景中提升了精确度（0.63%和2.06%）和F1分数（1.13%和2.05%），而且有效降低了计算复杂度（0.4161536 G），并将模型参数量调整至与单包表示的检测方法相近的水平。

4.1.2 少量平衡样本条件下的微调检测性能对比分析

根据样本量最少的类别，对每个类别随机选取900组数据包序列，训练数据按照70%的比例划分为每类630组，以评估模型在少量平衡样本微调下的鲁棒性和泛化能力，检测样本量及其分布不变。采用行归一化的混淆矩阵，其对角线反映模型的分类精确度。图7(a)展示了7个攻击类别的分类结果，推理时间为0.31 s。结果展现出了比较高的分类精确度，在区分DoS攻击的变种类别（Hulk、Slowhttptest和Slowloris）时，模型仍然表现出较高的区分精确度。在图7(b)中，将正常流量与7个攻击类别混合时，模型的分类精确度受到了一定影响。特别是对Botnet类别，虽

然漏报较少, 但将正常行为误报为攻击行为的情况较严重。

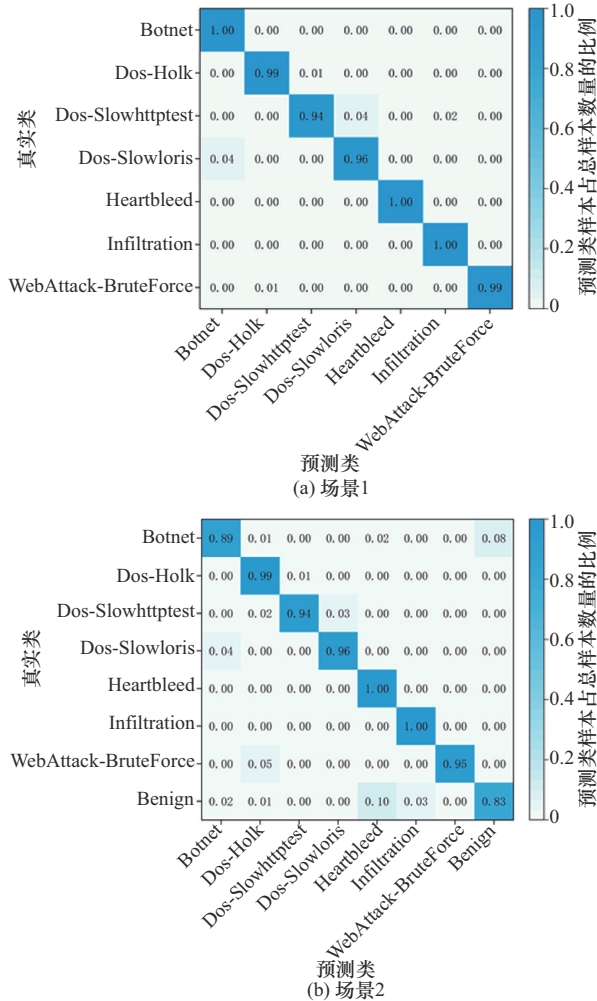


图7 少量平衡样本下检测性能
Figure 7 Detection performance under a small number of balanced samples

实验发现, 正常样本在分类中被错误划分为Heartbleed攻击的比例最高。Heartbleed攻击与正常的TLS/DTLS心跳流量在表面上是相似的, 因为攻击者发送的心跳请求在结构上与标准心跳请求无异。然而, 攻击者通过在心跳请求中设置异常长度字段值, 诱使服务器在处理时返回内存中的敏感数据, 从而实现攻击目的。这种攻击不涉及对加密握手过程的修改, 因此在数据包结构上, 攻击者发送心跳请求与正常心跳请求的流量难以区分。需要在流量表示时更多地关注包级别的特征, 特别是心跳请求中的长度字段是否与实际返回数据长度匹配。

4.1.3 推理性能与精确度对比

3种方法使用会话样本量一致, 在单包方法中, 每个数据包作为一个样本单元; 而在多包序列方法中, 连续8个数据包构成的序列作为一个样本单元。不同流量表示方法下推理性能与精确度比较如图8所示。

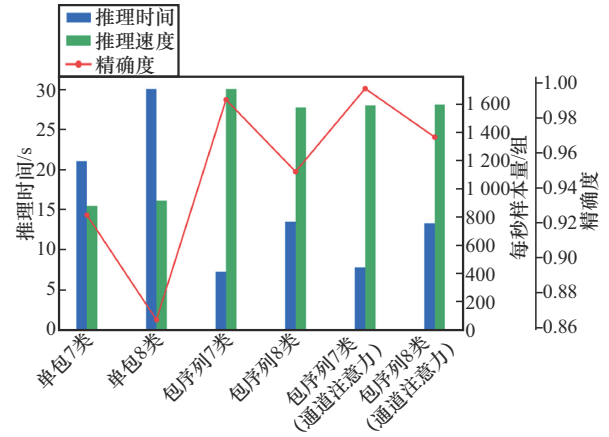


图8 不同流量表示方法下推理性能与精确度对比
Figure 8 Comparison of inference performance and accuracy under different traffic representation methods

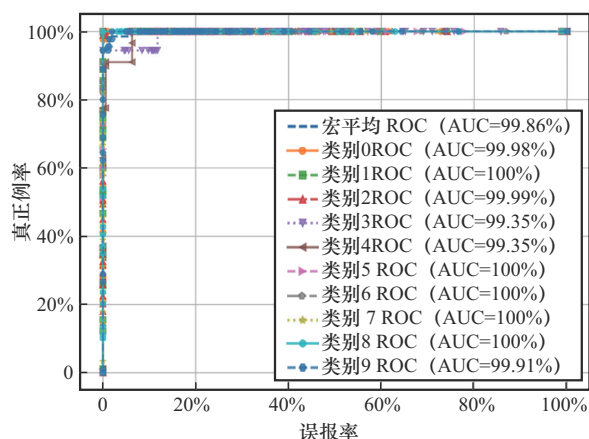
为了直观地展示推理时间、效率和精确度的趋势, 单包方法的每秒推理样本量除以8。多包序列方法在任务完成速度和检测精确度方面明显优于单包方法, 同时在整体平均效率上比单包方法提升约77%。在多包序列实验组中, 除了未使用注意力机制的7类攻击细粒度分类任务外, 其他3组的推理速度基本持平, 但当引入通道注意力机制进行特征融合时, 模型性能得到显著提升, 推理效率高于每秒1590个样本。

4.2 MTAttention方法跨数据集检测性能分析

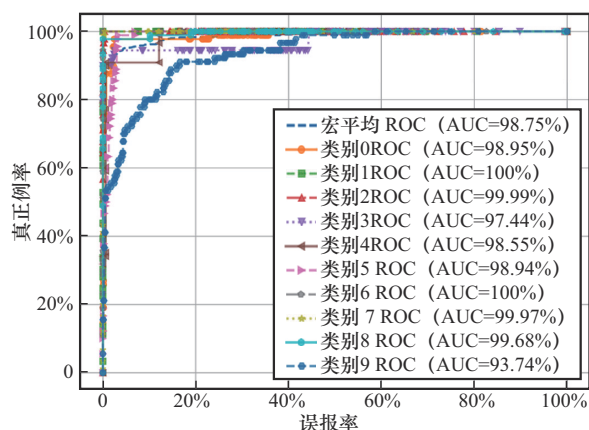
本节旨在评估该方法对跨数据集攻击类别的检测性能。引入CSE-CIC-IDS2018数据集^[23]中的3个类(BruteForce-SSH、DDoS-HOIC、FTP-BruteForce)进行10分类检测(场景1与场景2)。如图9所示, 采用ROC曲线及其相应的AUC值评估分类性能^[31]。

ROC曲线以误报率为横轴, 真正例率为纵轴, AUC值通过计算ROC曲线下面积获得。图9(a)中类别1为DDoS-HOIC、类别7为BruteForce-SSH, 图9(b)中类别1为DDoS-HOIC、类别5为FTP-BruteForce、类别8为BruteForce-SSH。在图9的

场景1中, 由CSE-CIC-IDS2018数据集引入的类别1和类别7的AUC值均为100%, 整体宏平均ROC为99.86%; 在场景2中, 引入的类别1、类别5和类别8的AUC值分别为100%、98.94%和99.68%, 整体宏平均ROC为98.75%。实验结果表明, MTAttention方法在数据集间具备可迁移性和良好的泛化能力。



(a) 场景1: 恶意流量细粒场景



(b) 场景2: 恶意流量识别与分类场景

图9 在CIC-IDS2017和CSE-CIC-IDS2018数据集上的混合分类性能ROC曲线

Figure 9 The ROC curve of hybrid classification performance on CIC-IDS2017 and CSE-CIC-IDS2018 datasets

4.3 MTAttention 与对照组检测方法性能对比分析

MTAttention方法与对照组方法的检测性能对比如表2所示, 选取4种被证明为有效的策略作为对照组, 分别如下。①FSL+FPE: 融合了模糊逻辑系统(fuzzy logic system, FLS)和特征概率估计(feature probability estimation, FPE)的方法, 该方法通过模糊逻辑处理输入特征的不确定

性, 结合FPE技术对特征值进行统计和概率评估, 为决策提供深度信息^[9]; ②Birch+MLP: 采用Birch聚类算法预处理数据并结合多层感知器(multilayer perceptron, MLP)分类器的两阶段网络异常检测方法, 通过聚类减少数据维度, 识别潜在模式, 再利用MLP进行深度分类^[10]; ③SGM-CNN: 结合类不平衡处理技术和卷积神经网络(convolutional neural network, CNN)的方法, 该方法有效解决了网络流量数据中正负样本不平衡问题^[11]; ④RFE-RF+CNN-LSTM: 采用随机森林递归特征消除(recursive feature elimination with random forest, RFE-RF)进行特征选择, 通过合成少数类过采样和编辑最近邻欠采样技术对数据进行重采样, 以解决类别不平衡问题, 随后利用CNN和长短期记忆(long short-term memory, LSTM)网络提取时间和空间特征^[32]。从特征处理、模型架构和性能评估等角度出发, 评估并对比MTAttention与传统机器学习技术、两阶段处理策略以及混合模型在恶意流量识别上的效果, 以及本文方法在不同场景下的泛化能力。

表2表明, 在样本量受限(少量平衡样本)场景下, MTAttention方法在7类攻击识别和分类任务中的总体性能优于其他方法。相关研究表明, Botnet、Infiltration和Web应用程序的攻击检测难度相对较高^[33]。在Botnet攻击检测中, MTAttention取得了相对高的检测准确率和F1分数, 相较于FSL+FPE、Birch+MLP和RFE-RF+CNN-LSTM模型, F1分数分别提升0.803 8、0.174 8和0.008 4。对于Infiltration攻击, MTAttention在所有评估指标上均表现优异, 其召回率相较于FSL+FPE、Birch+MLP和SGM-CNN模型分别提升0.055 6、0.25和0.29。在WebAttack-BruteForce攻击类型识别上, MTAttention同样获得了较好的准确率、精确度以及零误报率, 其中准确率较FSL+FPE、Birch+MLP和RFE-RF+CNN-LSTM分别提升0.001 3、1.51和1.09。综上所述, MTAttention方法在多个关键性能指标上均有提升。该方法对上述各类攻击(除Botnet外)实现了较高的检测精确度, 整体误报率保持在较低水平, 有效减少了恶意流量检测中的误报情况, 并且检测性能优于对照组方法中的二分类

表2 MTAttention方法与对照组方法的检测性能对比

Table 2 Comparison of detection performance between MTAttention method and control group methods

方法	对比参数	Botnet	Dos-Holk	Dos-Slowhttptest	Dos-Slowloris	Heartbleed	Infiltration	WebAttack-BruteForce	Benign	平均值
FSL+FPE	准确率	*87.19%	98.76%	99.94%	99.88%	100%	*99.99%	99.26%	98.74%	98.11%
	精确度	*7.62%	97.40%	97.94%	98.83%	100%	* 100%	55.56%	99.30%	82.26%
	召回率	* 100%	98.92%	94.68%	86.56%	100%	*94.44%	56.34%	98.72%	88.49%
	F1 分数	*14.16%	98.20%	96.28%	92.29%	100%	*97.14%	55.95%	99.01%	81.06%
	误报率	*12.95%	1.32%	0.02%	0.009%	0	* 0	0.38%	1.21%	1.986%
	训练样本量	1764	207982	4897	5208	11	36	1365	396 088	—
Birch+MLP	精确度	90.34%	—	—	—	—	100%	97.88%	99.85%	97.02%
	召回率	67.18%	—	—	—	—	75%	100%	99.65%	85.48%
	F1 分数	77.06%	—	—	—	—	85.71%	98.93%	99.75%	90.36%
	误报率	0.01%	—	—	—	—	0	0.03%	0.15%	0.048%
	训练样本量	1565	—	—	—	—	29	303 798	445 244	—
SGM-CNN	召回率	100%	100%	100%	100%	100%	71%	50%	100%	90.13%
	训练样本量	1 376	161 751	3 849	4 057	8	26	1 055	1 591 167	—
RFE-RF+CNN-LSTM	准确率	87.90%	—	—	—	—	—	98.30%	84.50	90.23%
	精确度	99.5%	—	—	—	—	—	99.10%	86.90	95.17%
	召回率	88.3%	—	—	—	—	—	99.20%	97	94.83%
	F1 分数	93.7%	—	—	—	—	—	99.10%	91.70	94.83%
	误报率	3.4%	—	—	—	—	—	0.90%	14.30%	6.2%
	训练样本量	1573	—	—	—	—	—	1206	1 818 477	—
MT Attention	准确率	98.62%	99.85%	99.69%	99.84%	100%	100%	99.39%	98.63%	99.50%
	精确度	90.70%	98.86%	100%	100%	100%	100%	100%	100%	98.70%
	召回率	98.73%	100%	97.67%	96.47%	100%	100%	95.40%	89.16%	97.18%
	F1 分数	94.54%	99.43%	98.82%	98.20%	100%	100%	97.65%	94.27%	97.86%
	误报率	1.39%	0.17%	0	0	0	0	0	0	0.195%
	训练样本量	630	630	630	630	630	630	630	630	—

备注：* 表示二分类检测结果，— 表示该研究中没有公布相应细粒度识别性能

性能。

在对比实验中，MTAttention在各类攻击检测任务中均采用相对较少的数据包序列（630组）进行微调训练。数量显著低于其他方法的标记训练样本量，但仍具有较高检测性能。即使是Heartbleed和Infiltration攻击的训练样本量，也是通过数据集中仅有的实际样本直接转换而来（少于11个和35个）。这些结果表明，MTAttention方法能够在数据稀缺的情况下学习到对下游任务有效的通用表示，具有良好的迁移性和适应性。尽管该方法在计算上更为复杂，但在推理阶段性

能表现出色，表明其在实际部署中具有较强的潜力。

与基于MAE范式的改进方案对比如下。Flow-MAE^[20]方法将流量表示为突发（burst），其输入序列长度为1 024；编码器深度为12，每层具有12个自注意力头，嵌入隐层维度为768。MTAttention的输入为28×28矩阵形式拼接的多包时间序列；编码器深度为8，每层具有8个自注意力头，嵌入隐层维度为128，采用了更优化的网络结构，显著提升了推理速度，减少了内存开销和参数量，实现了更高效率。具体来说，

Flow-MAE 涉及 8 512 万个参数，而 MTAttention 只需要 133.8 万个参数。微调 Flow-MAE 和 MTAttention 的内存开销分别为 3.6 GB 和 2.3 GB，如表 3 所示。Flow-MAE 在其原始研究中展示了出色的性能^[20]。然而，在相同数据资源实验条件下，MTAttention 的检测性能显著优于 Flow-MAE，推理速度提升了约一倍。MTAttention 仅需要该模型 1.56% 的参数量和 63.89% 的内存开销，模型大小仅为 5.17 MB。参数量是影响深度学习模型执行效率和适应性的关键因素。与具备可迁移性和泛化能力的预训练微流量检测模型^[17,20,34]相比，MTAttention 的低参数量使其在特定数据分布和任务需求中更易微调和优化，同时显著降低了存储空间和计算资源的需求，使其更适合在资源受限环境下部署。

4.4 MTAttention 方法可视化分析

为了揭示通道注意力机制在提升模型预测性能方面的依据，对其调控下的加权特征和聚合特征分布模式进行可视化，采用 UMAP 降维技术^[35]等可视化样本的加权特征和聚合特征分布。加权特征值在数据包特征序列中随着隐层特征维度索引的变化情况图 10 所示。该结果揭示了特征序列在不同位置的特征强度，表明数据包内部结构的一致性模式。在特定序列位置上观察到的显著峰值或谷值映射了数据包内部结构的关键特征，并突出了这些位置在流量分析中的显著性，表明通道注意力机制使模型自适应地调整了每个通道的权重，实现了对关键特征的突显和对冗余信息的抑制。

特征融合模块通过对数据包序列中的加权特征进行平均化处理，生成了聚合特征值，这些值在特征空间中的分布被有效映射出来，如图 11 所示。该特征序列的聚合特征值主要分布在 -0.008 与 0.006 之间，并展现出相似的分布特性。这一集中趋势表明，模型在提取特征时识别并强

调了在多个数据包中普遍存在的特征。在分布模式中频繁出现的特征值构成了数据包序列的核心模式，并作为识别关键特征标识的依据。特征值的正负趋势反映了特征与数据包行为间的明确相关性。

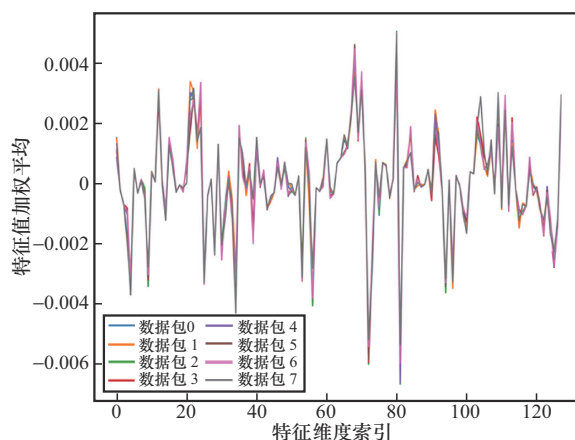


图 10 数据包序列加权特征

Figure 10 Weighted features of packet sequence

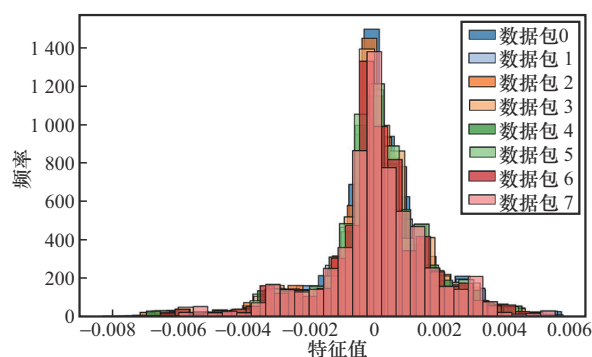


图 11 数据包序列聚合特征分布

Figure 11 Distribution of aggregation characteristics of packet sequence

图 12 为利用 UMAP 技术在降维过程中提取的 2 个重要不相关维度（第一主成分和第二主成分），以直观理解数据的内在结构和分布。具体地，图中展示了 3 个随机训练批次聚合特征低维空间的分布情况，其中每个批次的数据点以颜色区分，以二维散点图的形式表示。

表 3 基于 MAE 改进方法的效率对比

Table 3 Efficiency comparison of MAE-based Improvement methods

方法	精确度	召回率	参数量/M	内存开销/GB	推理速度/(样本·秒 ⁻¹)
Flow-MAE	93.98%	74.84%	85.12	3.6	849~931
MTAttention	98.7%	97.18%	1.338	2.3	1 678~1 882

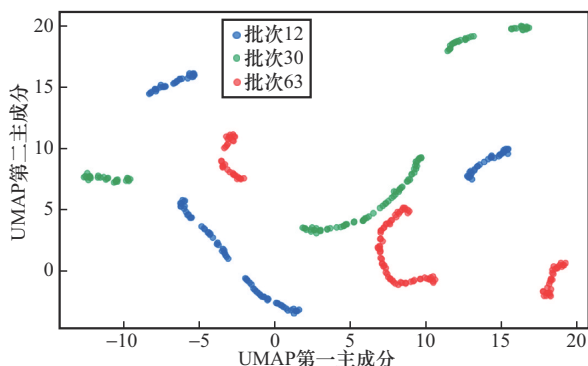


图 12 随机批次聚合特征的低维空间分布
Figure 12 Distribution of random batch aggregated features in low-dimensional space

图 12 揭示了在特征空间中同一批次内样本存在内在一致性，以及不同类别数据点形成的相对分离的簇，表明模型具备可区分能力。通过在特征聚合过程引入注意力权重，模型能够识别并强化同一批次内相似样本的共同特征，导致这些样本在低维空间中显著聚集。同时，不同批次间样本的明显分离模式进一步验证模型在捕捉数据包序列差异方面的能力。

5 结束语

本文提出一种自监督学习驱动的注意力增强恶意流量检测方法——MTAttention，该方法通过结构化的多包序列支持对流和会话的流量表示，基于无标记数据采用自监督掩码重构预训练策略学习跨类别的流量通用表示，并在下游任务分类前引入基于通道注意力的多包序列特征融合模块，增强对数据包序列中高权重特征的整合能力，从而提升模型对恶意流量识别与分类的性能。实验结果表明，MTAttention 方法在面对不平衡恶意样本和有限训练数据时，仍对包含变种攻击的检测任务实现较高的检测性能，推理效率高且误报率低。该方法减少了对大量标记样本的依赖，适用于资源受限环境部署，并具有良好的可迁移性和新场景适应性。

下一步研究将考虑集成多模态学习策略，利用不同模态数据间的互补信息，提高模型对数据包间差异的敏感性及序列动态变化的捕捉能力，以期获得更全面的流量行为理解。

参考文献:

- [1] 侯剑, 鲁辉, 刘方爱, 等. 加密恶意流量检测及对抗综述[J]. 软件学报, 2024, 35(1): 333-355.
HOU J, LU H, LIU A F, et al. Detection and countermeasure of encrypted malicious traffic: A survey[J]. Journal of Software, 2024, 35(1): 333-355.
- [2] XU C, CHEN S, SU J, et al. A survey on regular expression matching for deep packet inspection: Applications, algorithms, and hardware platforms[J]. IEEE Communications Surveys & Tutorials, 2016, 18(4): 2991-3029.
- [3] XU G, GUO B, SU C, et al. Am I eclipsed? A smart detector of eclipse attacks for Ethereum[J]. Computers & Security, 2020, 88: 101604.
- [4] SHEN M, ZHANG J, ZHU L, et al. Accurate decentralized application identification via encrypted traffic analysis using graph neural networks[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 2367-2380.
- [5] WANG T, CAI X, NITHYANAND R, et al. Effective attacks and provable defenses for website fingerprinting[C]//23rd USENIX Security Symposium (USENIX Security 14). 2014: 143-157.
- [6] ASHIKU L, DAGLI C. Network intrusion detection system using deep learning[J]. Procedia Computer Science, 2021, 185: 239-247.
- [7] QU J, MA X, LI J, et al. An input-agnostic hierarchical deep learning framework for traffic fingerprinting[C]//32nd USENIX Security Symposium. 2023: 589-606.
- [8] 李丽娟, 李曼, 毕红军, 等. 基于混合深度学习的多类型低速率DDoS攻击检测方法[J]. 网络与信息安全学报, 2022, 8(1): 73-85.
LI L J, LI M, BI H J, et al. Multi-type low-rate DDoS attack detection method based on hybrid deep learning[J]. Journal of Cyber Security and Information, 2022, 8(1): 73-85.
- [9] PRASAD M, TRIPATHI S, DAHAL K. An efficient feature selection based Bayesian and Rough set approach for intrusion detection[J]. Applied Soft Computing, 2020, 87: 105980.
- [10] YIN Y, JANG-JACCARD J, SABRINA F, et al. Improving multilayer-perceptron (MLP)-based network anomaly detection with birch clustering on CICIDS-2017 Dataset[C]//2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD). IEEE, 2023: 423-431.
- [11] ZHANG H, HUANG L, WU C Q, et al. An effective convolutional neural network based on smote and gaussian mixture model for intrusion detection in imbalanced dataset[J]. Computer Networks, 2020, 177(18): 107315.
- [12] HOJJATI H, HO T K K, ARMANFARD N. Self-supervised anomaly detection in computer vision and beyond: A survey and outlook[J]. Neural Networks, 2024: 106106.
- [13] YUANPENG T, BOSHEN Z, LIANG L, et al. Self-supervised feature adaptation for 3D industrial anomaly detection[J]. Computer Vision - ECCV 2024, PT II, 2025, 15060: 75-91.
- [14] XIE R, WANG Y, CAO J, et al. Rosetta: Enabling robust tls encrypted traffic classification in diverse network environments with tcp-aware traffic augmentation[C]//Proceedings of the ACM Turing Award Celebration Conference-China 2023. 2023: 131-132.
- [15] WANG W, JIAN S, TAN Y, et al. Robust unsupervised network in-

- trusion detection with self-supervised masked context reconstruction[J]. Computers & Security, 2023, 128: 103131.
- [16] HAN Y, HUI L, JIANG H, et al. Generative subgraph contrast for self-supervised graph representation learning[C]//European Conference on Computer Vision. 2022: 91-107.
- [17] LIN X, XIONG G, GOU G, et al. Et-bert: A contextualized data-gram representation with pre-training transformers for encrypted traffic classification[C]//Proceedings of the ACM Web Conference 2022. 2022: 633-642.
- [18] HE K, CHEN X, XIE S, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 16000-16009.
- [19] XU K, ZHANG X, WANG Y, et al. Self-supervised learning malware traffic classification based on masked auto encoder[J]. IEEE Internet of Things Journal, 2024, 11(10): 17330-17340.
- [20] HANG Z, LU Y, WANG Y, et al. Flow-MAE: leveraging masked autoencoder for accurate, efficient and robust malicious traffic classification[C]//Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses. 2023: 297-314.
- [21] XU C, SHEN J, DU X. A method of few-shot network intrusion detection based on meta-learning framework[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 3540-3552.
- [22] 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.
- [23] SHARAFALDIN I, LASHKARI A H, GHORBANI A A. Toward generating a new intrusion detection dataset and intrusion traffic characterization[C]//International Conference on Information Systems Security & Privacy. 2018: 108-116.
- [24] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[C]//International Conference on Learning Representations. 2021.
- [25] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.
- [26] HU J, SHEN L, SUN G, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(8): 2011-2023.
- [27] WANG Q, WU B, ZHU P, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 11531-11539.
- [28] 邓明阳, 郭应时, 刘通. 基于分层重组的不平衡数据采样方法研究[J]. 重庆理工大学学报(自然科学), 2021, 35(8): 122-128.
- DENG M Y, GUO Y S, LIU T, et al. Research on imbalanced data sampling method based on stratification and recombination[J]. Journal of Chongqing University of Technology(Natural Science), 2021, 35(8): 122-128.
- [29] ENGELEN G, RIMMER V, JOOSEN W. Troubleshooting an intrusion detection dataset: the CICIDS2017 case study[C]//2021 IEEE Security and Privacy Workshops (SPW). IEEE, 2021: 7-12.
- [30] HE H, YANG Z, CHEN X. PERT: payload encoding representation from transformer for encrypted traffic classification[C]//ITU Kaleidoscope Academic Conference, 2021, 19(4): 90-97.
- [31] HUANG J, LING C X. Using AUC and accuracy in evaluating learning algorithms[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(3): 299-310.
- [32] SALEH S M, SAYEM I M, MADHAVJI N, et al. Advancing software security and reliability in cloud platforms through AI-based anomaly detection[C]//Proceedings of the 2024 on Cloud Computing Security Workshop. 2024: 43-52.
- [33] ROSAY A, RIOU K, CARLIER F, et al. Multi-layer perceptron for network intrusion detection: From a study on two recent data sets to deployment on automotive processor[J]. Annals of Telecommunications, 2022, 77(5): 371-394.
- [34] ZHOU G, GUO X, LIU Z, et al. TrafficFormer: An efficient pre-trained model for traffic data[C]//2025 IEEE Symposium on Security and Privacy (SP). IEEE Computer Society, 2024: 102-102.
- [35] MCINNES L, HEALY J, MELVILLE J. Umap: Uniform manifold approximation and projection for dimension reduction[J]. arXiv preprint arXiv:1802.03426, 2018.

[作者简介]



孙剑文 (1988-), 女, 北京人, 信息工程大学博士生, 主要研究方向为网络流量异常检测、机器学习。



张斌 (1969-), 男, 河南南阳人, 博士, 信息工程大学教授、博士生导师, 主要研究方向为信息系统安全。



李红宇 (2001-), 男, 河南民权人, 信息工程大学博士生, 主要研究方向为异常检测。



常禾雨 (1993-), 女, 河南郑州人, 博士, 信息工程大学副研究员, 主要研究方向为人工智能安全。