

基于改进的残差 U-Net 的不平衡协议识别方法

吴吉胜, 洪征, 马甜甜

(陆军工程大学指挥控制工程学院, 江苏 南京 210000)

摘要: 随着互联网的不断发展, 网络攻击事件不断增多, 成为网络安全的巨大挑战。在所捕获网络流量中, 恶意流量往往占比较少, 即攻击者使用的通信协议往往为少数类协议。当协议数据的类别分布不平衡时, 现有协议识别方法能够识别出多数类协议, 但是难以准确识别少数类协议。针对这一问题, 提出一种基于改进的残差 U-Net 的不平衡协议识别方法, 利用新的激活函数和 SE-Net (squeeze-and-excitation networks) 改进残差 U-Net, 提升残差 U-Net 的特征提取能力。同时采用带权重的 Dice 损失函数作为协议识别模型的损失函数, 少数类协议的识别准确率偏低会导致损失函数的值偏高, 进而促使少数类协议主导模型的优化方向。采用所提方法进行协议识别时, 首先从网络流量中抽取网络流, 经过预处理转化为一维矩阵, 利用协议识别模型提取协议数据的特征, 进而由 Softmax 分类器计算输出协议类型。实验结果表明, 与对比模型相比, 所提协议识别模型能够更为准确地识别少数类协议, 同时多数类协议的识别准确率得到了提升。

关键词: 协议识别; 类别不平衡; 卷积神经网络; 激活函数; 损失函数

中图分类号: TP398.08

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2024004

Unbalanced protocol recognition method based on improved residual U-Net

WU Jisheng, HONG Zheng, MA Tiantian

College of Command and Control Engineering, Army Engineering University, Nanjing 210000, China

Abstract: An unbalanced protocol recognition method based on the improved Residual U-Net was proposed to solve the challenge of network security posed by the increasing network attacks with the continuous development of the Internet. In the captured network traffic, a small proportion is constituted by malicious traffic, typically utilizing minority protocols. However, existing protocol recognition methods struggle to accurately identify these minority protocols when the class distribution of the protocol data is imbalanced. To address this issue, an unbalanced protocol recognition method was proposed, which utilized the improved Residual U-Net, incorporating a novel

收稿日期: 2023-08-11; 修回日期: 2023-12-18

通信作者: 洪征, hz5215@163.com

基金项目: 国家重点研发计划 (2017YFB0802900)

Foundation Item: The National Key R&D Program of China (2017YFB0802900)

引用格式: 吴吉胜, 洪征, 马甜甜. 基于改进的残差 U-Net 的不平衡协议识别方法[J]. 网络与信息安全学报, 2024, 10(1): 136-155.

Citation Format: WU J S, HONG Z, MA T T. Unbalanced protocol recognition method based on improved residual U-Net[J]. Chinese Journal of Network and Information Security, 2024, 10(1): 136-155.

activation function and the Squeeze-and-Excitation Networks (SE-Net) to enhance the feature extraction capability. The loss function employed in the proposed model was the weighted Dice loss function. In cases where the recognition accuracies of the minority protocols were low, the loss function value would be high. Consequently, the optimization direction of the model would be dominated by the minority protocols, resulting in improved recognition accuracies for them. During the protocol recognition process, the network flow was extracted from the network traffic and preprocessed to convert it into a one-dimensional matrix. Subsequently, the protocol recognition model extracted the features of the protocol data, and the Softmax classifier predicted the protocol types. Experimental results demonstrate that the proposed protocol recognition model achieves more accurate recognition of the minority protocols compared to the comparison model, while also improving the recognition accuracies of the majority protocols.

Keywords: protocol recognition, class unbalance, convolutional neural network, activation function, loss function

0 引言

随着网络技术不断发展, 新型网络协议不断涌现, 网络流量急剧增加, 给网络管理和网络安全带来严峻挑战。协议识别旨在分析网络流量的组成, 将网络流量划分为不同的协议类别。协议识别技术是优化网络管理、保证服务质量和防范安全风险的重要内容。

现有的协议识别方法主要包括基于端口的协议识别方法、基于深度包检测技术的协议识别方法、基于机器学习的协议识别方法和基于深度学习的协议识别方法 4 类^[1]。由于端口随机化和加密技术的采用, 仅根据端口难以准确识别协议, 基于端口的协议识别方法在协议识别中的作用越来越小。基于深度包检测技术的协议识别方法通过提取协议数据有效载荷的特征字符串, 与预先建立的特征库进行对比得到目标协议类别。该方法存在无法识别加密协议和计算量较大等问题。基于机器学习的协议识别方法通过收集协议数据的统计特征, 利用机器学习算法推断目标协议的类别。这类方法依赖专家知识确定合适的统计特征, 而且协议识别模型的泛化能力较差。基于深度学习的协议识别方法利用深度学习模型从网络流量中学习具有区分度的协议特征, 利用分类器识别不同协议。该方法能够避免人工干预的影响, 模型泛化能力强, 协议识别准确率较高。

在真实网络环境中, 网络流量往往存在协议类别分布不平衡的问题。一些协议的样本数量较多, 这些协议被称为多数类协议, 还有一些协议的样本数量较少, 被称为少数类协议。在构建协

议识别模型时, 多数类协议往往会主导协议识别模型的优化方向, 导致多数类协议的识别准确率较高, 而少数类协议的识别准确率较低。

现有不平衡协议识别方法包括 4 类: 数据重采样的方法、基于特征选择的方法、基于集成算法的方法和代价敏感的方法。数据重采样的方法一般需要增加少数类协议的样本数量或者减少多数类协议的样本数量, 这类方法耗时长, 并且会改变协议数据原本的类别分布。基于特征选择的方法通过保留与少数类协议相关度较高的统计特征来准确识别少数类协议。基于集成算法的方法综合多个分类器的分类结果得到最终协议识别结果。与单个分类器相比, 集成算法能够结合多个分类器的优势, 更准确地识别少数类协议。代价敏感的方法通过将较高的误分类代价赋予少数类协议, 将较低的误分类代价赋予多数类协议, 优化协议识别模型使得误分类的总代价最小, 达到提高少数类协议识别准确率的目的。目前代价敏感的方法是不平衡协议识别领域的研究热点。

针对网络流量中协议类别分布不平衡的问题, 本文提出一种基于改进的残差 U-Net 的不平衡协议识别方法, 提高了少数类协议识别的准确率。残差 U-Net^[2]固有的特征融合特性和残差连接结构使其具有出色的特征提取能力, 本文方法在残差 U-Net 中引入 SE-Net (squeeze-and-excitation networks)^[3]来挖掘协议数据特征不同通道的重要性, 选择性地强化或抑制协议数据特征的部分通道。同时, 在残差 U-Net 中引入新的非线性激活函数, 在对输入进行激活时, 新激活函数能够有效减少信息丢失, 提高模型的特征提取能力。

此外, 本文在识别模型中利用损失函数来优化少数类协议的识别。在协议类别分布不平衡的网络流量中, 如果多数类和少数类协议的误分类代价相同, 则由于数量优势, 多数类协议的分类错误导致模型分类误差偏高。在模型优化过程中, 往往以牺牲少数类协议的识别准确率为代价提高多数类协议的识别准确率, 进而获得较高的总体协议识别准确率。Dice 损失函数^[4]使一类样本对分类模型总损失的贡献与该类包含的样本数量的关联关系减弱。当某类样本的预测结果和真实标签差异较大时, 该类样本对应的损失函数的值较高。当某类样本的预测结果和真实标签差异较小时, 该类样本对应的损失函数的值较低。在分类模型优化过程中, 一开始主要是多数类样本导致损失函数的值偏高, 模型通过提高多数类样本的识别准确率来降低损失函数的值; 一段时间后, 当多数类样本识别准确率较高而少数类样本识别准确率较低时, 少数类样本是损失函数的值偏高的主要原因, 模型通过提高少数类样本的识别准确率来降低损失函数的值。以此类推, 多数类样本和少数类样本交替主导模型的优化方向, 当损失函数的值稳定后, 停止优化分类模型。此时, 模型能够较为准确地识别多数类样本和少数类样本。

Dice 损失函数对各类样本赋予相同的误分类代价, 即各类样本的权值相同, 在模型优化过程中多数类协议对于 Dice 损失函数值的影响比少数类协议高, 导致利用 Dice 损失函数解决类别分布不平衡的能力有限。鉴于此, 本文设计了带权重 Dice 损失函数, 为少数类协议赋予较大的权重, 为多数类协议赋予较小的权重。带权重 Dice 损失函数通过对 Dice 损失函数进行强化, 可以进一步提升 Dice 损失函数解决协议类别分布不平衡问题的能力。

1 相关工作

网络环境复杂多变, 捕获的网络流量中协议类别往往分布不平衡。多数类协议由于数量上的优势, 一直是研究的重点。但是, 少数类协议很可能包含丰富的信息。例如, 恶意软件使用特定的通信协议相互交流, 协议的样本数量较少, 往往隐藏在正常的通信流量中, 不容易发现。高效、

准确地识别此类协议能够及时发现网络中的异常通信, 以便采取有效的应对措施。数据重采样的方法、基于特征选择的方法、基于集成算法的方法和代价敏感的方法是常用的不平衡协议识别方法。

数据重采样的方法主要包括对少数类协议随机过采样、对多数类协议随机欠采样和合成少数类协议样本等处理手段。随机过采样方法随机复制少数类协议样本, 使得少数类协议和多数类协议的样本数量相当, 从而得到类别分布平衡的数据集。由于是少数类样本的简单重复, 该方法容易导致模型过拟合。随机欠采样方法通过随机丢弃部分多数类协议样本, 得到各类协议样本数量相当的数据集, 但是这种方法可能丢失很多重要信息。合成少数类协议样本的方法主要是利用合成少数类过采样技术 (SMOTE, synthetic minority oversampling technique) 或者生成对抗网络等方法学习少数类协议样本的分布, 并生成符合原始少数类协议样本分布的新样本, 将新样本和原始数据混合, 得到均衡的数据集, 然后利用机器学习或者深度学习模型进行协议识别。Wang 等^[5]提出基于条件生成对抗网络 (CGAN, conditional generative adversarial network) 的不平衡加密协议识别方法, 在利用 CGAN 学习少数类协议样本的分布时, 将少数类协议样本及对应协议标签输入 CGAN 中, 使 CGAN 只生成协议标签对应类别的协议样本。实验结果表明, 该方法生成的样本更符合原始少数类协议样本分布, 协议识别的准确率较高, 但是该方法非常耗时。Hasibi 等^[6]提出基于数据增强的不平衡协议识别方法, 每种方法选取 2 种具有序列相关性的统计特征和 4 种没有序列相关性的统计特征。在合成少数类协议样本时, 对于具有序列相关性的统计特征, 利用长短期记忆网络 (LSTM, long short-term memory) 获得相应结果, 对于没有序列相关性的统计特征, 利用核密度估计算法获得相应结果, 将合成样本和原始数据集混合构建平衡数据集。由于 LSTM 模型无法并行处理数据, 方法需要较长时间获取少数类协议的样本分布信息。

基于特征选择的方法通过特征工程方法确定协议数据集的最优特征子集, 尽可能保留与少数类协议相关度高的统计特征, 达到准确识别少数

类协议的目的。Tang 等^[7]选择与少数类协议相关度高的特征, 充分考虑特征与协议类别、特征与特征之间的关联关系, 基于加权对称不确定性剔除冗余特征, 并利用序列搜索算法得到最优特征子集。实验结果表明, 该方法能够提高少数类协议的识别准确率, 但是确定最优特征子集所需时间较长, 计算复杂度偏高。

基于集成算法的方法通过综合多个基本分类器的协议识别结果, 得到比单个基本分类器更好的识别效果, 是因为多个分类器能够优势互补, 有利于准确识别少数类协议。Gupta 等^[8]提出基于代价敏感神经网络和集成算法的不平衡协议识别方法, 首先利用代价敏感神经网络将恶意流量和正常流量区分开, 然后利用 XGBoost 集成算法识别恶意流量中的多数类协议, 将恶意流量中包含的所有少数类协议识别为一类协议。在利用 XGBoost 算法将所有少数类协议识别为同一类协议后, 进一步利用随机森林算法识别该类协议包含哪些少数类协议。实验结果表明, 该方法能够较为准确地识别少数类协议, 但是需要人工确定网络流量的统计特征, 模型的泛化能力较弱。

代价敏感的方法是指通过完善分类模型的损失函数来解决类别分布不平衡的问题, 即在利用损失函数计算模型分类误差时, 为少数类和多数类协议分别分配较高和较低的误分类代价, 从而强化少数类协议对模型分类误差的贡献, 进而提高少数类协议的识别准确率。Telikani 等^[9]提出代价敏感的协议识别方法, 首先统计每类协议的样本数量, 然后据此分配每类协议的误分类代价。每类协议的误分类代价与协议的样本数量成反比。但是这种方法不能根据分类结果动态调整误分类代价, 导致少数类协议的识别准确率提升有限。

除了以上方法之外, 一些研究人员通过特征融合提高协议识别模型的特征提取能力, 进而提高少数类协议的识别准确率。特征融合是指将网络中多个层的输出进行合并, 由于这些层与输出层的距离不同, 在提取协议数据特征时信息损耗程度不同, 侧重点不同, 基于合并后的结果挖掘协议的数据特征, 有助于让模型更准确地识别不同的协议。由于每层的输出包括多个通道, 根据通道数的差异可以将特征融合分为两类: 一类是

所有输入的通道数相同, 通道的形状也相同, 将不同输入相应通道、相应位置的元素进行相加操作, 所得结果的通道数与各个输入的通道数相同, 通道形状也相同; 另一类是所有输入的通道的形状都相同, 但是通道数量存在差异, 将所有输入进行通道拼接, 所得结果的通道数为所有输入的通道数之和。由于前一类特征融合方式存在信息损失的问题, 目前特征融合广泛采用通道拼接的方式。Zhang 等^[10]提出基于并行卷积神经网络的不平衡协议识别方法, 利用全卷积神经网络和常规卷积神经网络并行提取协议数据的特征, 在特征提取过程中多次进行特征融合。但是该方法仍然存在多数类协议样本主导模型优化方向的问题。Xia 等^[11]提出基于残差网络和 Inception 卷积神经网络的不平衡协议识别方法。Inception 卷积神经网络能够利用多个卷积层并行提取协议数据的特征并进行融合, 残差网络可以强化 Inception 卷积神经网络的特征提取能力。该方法存在多数类协议样本主导模型优化方向的问题。

从现有研究来看, 代价敏感的不平衡协议识别方法能够更准确地识别少数类协议, 避免模型过拟合、时间复杂度高和改变原始数据分布等问题, 模型泛化能力优于基于特征选择的方法和基于集成算法的方法。但是, 现有代价敏感的不平衡协议识别方法存在模型结构简单、特征提取能力较弱等问题。提高模型的特征提取能力, 同时利用代价敏感损失函数避免多数类协议主导模型优化方向, 是准确识别少数类协议的关键。

2 背景知识

在深度学习领域, 通道指的是卷积层利用卷积核与输入做卷积运算时所输出的矩阵。由于卷积层中每个卷积核的形状相同, 不同卷积核得到的矩阵的形状也相同。卷积层的输出由所有卷积核对应的矩阵组成, 也就是说, 卷积层的输出由多个形状相同的通道组成。

按通道维度拼接是常见的一种通道处理方式。例如, 对于输入 A 和输入 B, 两者通道的形状相同, 通道的数量可以相同也可以不同, 依次获取 B 的每一个通道, 将这个通道拼接到 A 的尾部, 作为 A 的一个新通道。在分类模型中, 越靠

近输出层,模型所提取的特征越抽象,对于判断输入所属类别的作用越大。分类模型在提取输入的特征时,会通过下采样来降低输入的维度,从而降低模型参数规模并提升模型训练效率,但是下采样会导致丢失较多有用信息。将分类模型多个层的输出按通道维度拼接,能够有效解决信息损失的问题,提高模型分类性能,因为这些层与输出层的距离不同,所提取的数据特征的抽象程度有差异,能够反映输入不同层面的信息。

U-Net^[12]采用通道拼接的方式进行特征融合,综合网络中多个层的输出信息,能够让 U-Net 提取更丰富的特征,准确判断输入的类别。残差 U-Net 在 U-Net 的基础上利用残差连接避免梯度消失和模型性能退化等问题,特征提取能力得到有效提高。本文引入 SE-Net 对残差 U-Net 进行改进,挖掘协议数据特征不同通道的关联关系,进而为不同通道分配不同权重,根据权重有选择地强化或抑制协议数据特征的部分通道。此外,本文采用新的非线性激活函数,解决对输入进行非线性激活时部分信息丢失的问题,进一步提高残差 U-Net 的特征提取能力。

为消除协议类别不平衡的影响,在 Dice 损失函数的基础上,本文提出了带权重 Dice 损失函数作为模型的损失函数。Dice 损失函数的计算如式(1)所示。

$$\text{Dice} = 1 - \frac{2 \sum_{i=1}^N p_i t_i + \varepsilon}{\sum_{i=1}^N p_i^2 + \sum_{i=1}^N t_i^2 + \varepsilon} \quad (1)$$

其中, N 为样本数量, p_i 为输入样本 i 对应的预测标签, t_i 表示输入样本 i 对应的真实标签。由式(1)可知, Dice 损失函数能够在一定程度上消除协议类别不平衡带来的影响,原因是 Dice 损失函数中分子为各类协议所有样本的真实标签和对应预测标签的乘积之和,分母为各类协议所有样本真实标签的平方和加上预测标签的平方和。当某类协议识别准确率较高时,对应分子和分母的值较为接近,该类协议对协议识别模型分类误差的贡献较小。当某类协议识别准确率较低时,对应分子小于分母,该类协议对协议识别模型分类误差的贡献较大。因此,在 Dice 损失函数中,某类协议对协议识别模型分类误差的贡献,与该类协议样

本数量的关联性相对较小。

由于 Dice 损失函数为各类别样本赋予相同的误分类代价,当少数类协议错误分类的样本数量少于多数类协议错误分类的样本数量时,在 Dice 损失函数的分子和分母中,少数类协议对应的分类误差均小于多数类协议对应的分类误差,这导致模型仍旧更加关注多数类协议,需要通过提高多数类协议的识别准确率来降低模型分类误差。因此,在 Dice 损失函数中,各类协议对模型分类误差的贡献,与该类协议错误分类的样本数量仍具有关联性,这导致 Dice 损失函数消除协议类别不平衡的能力有限。本文提出带权重 Dice 损失函数作为模型损失函数,为多数类协议和少数类协议分别赋予较小和较大的权重,使模型更加关注少数类协议,通过提高少数类协议的识别准确率来降低分类误差,有效强化 Dice 损失函数解决协议类别分布不平衡问题的能力。

本节主要介绍残差 U-Net 和 SE-Net 的相关知识。

2.1 残差 U-Net

U-Net 包括编码器和解码器,两部分的结构基本对称,均包含多个结构相同的模块。编码器负责提取输入的特征,解码器负责将编码器的输出逐步还原成编码器的输入,并与编码器对应模块的输出按通道维度进行拼接,有效利用编码器对应模块的输出减少信息损失,提高模型分类性能。通过融合特征,U-Net 能够捕获输入中最具代表性的特征,一个简单的 U-Net 的结构如图 1 所示。

在图 1 中,编码器由模块 1、模块 2 和模块 3 这 3 个结构相同的模块构成,每个模块包括两个卷积层和一个池化层,卷积层用于提取输入的特征,池化层用于下采样,降低模型参数规模。解码器主要包括模块 4、模块 5 这两个结构相同的模块,每个模块由一个上采样层、一个拼接层和两个卷积层组成。由于模块 4 和模块 5 的特征提取过程相同,这里以模块 4 为例介绍特征提取过程。模块 4 首先通过上采样层对模块 3 的输出进行上采样,然后利用拼接层将上采样层的输出与模块 2 的输出按通道维度进行拼接,然后利用两个卷积层挖掘拼接结果的特征。

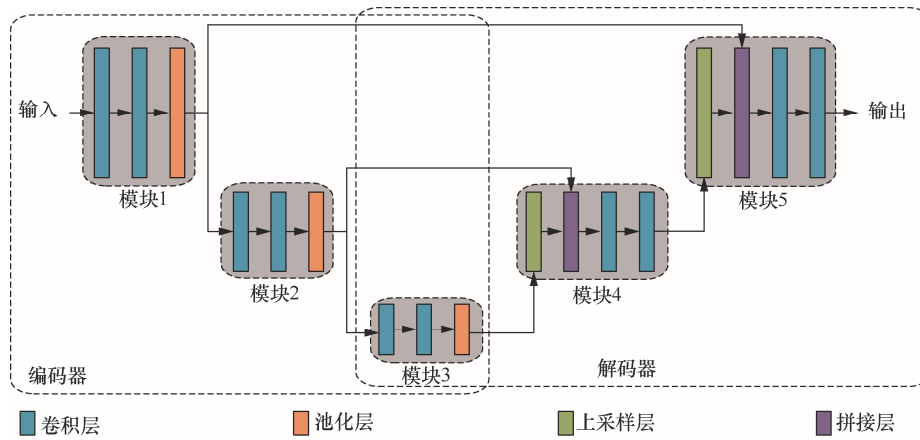


图 1 U-Net 的结构
Figure 1 Structure of U-Net

Zhang 等^[2]将残差连接引入 U-Net 中, 构建特征提取能力更强的残差 U-Net。残差 U-Net 的结构如图 2 所示, 其中 Conv 代表卷积层, BN 代表批量归一化层, ReLU 代表 ReLU 激活函数, Addition 代表对应元素相加, Upsampling 代表上采样层, Concatenate 代表拼接层, 将所有输入按通道维度进行拼接。从图 2 可以看出, Block2~Block4 的结构相同, Block5~Block7 的结构也相同, 后者均在 Block1 的基础上增加了上采样层和拼接层, 因此 Block1 模块被视为残差 U-Net 的基本模块。

图 2 所示的残差 U-Net 中, 编码器引入了残差连接, 将每个模块的输入与该模块第二个卷积层的输出通过 Addition 进行对应元素相加, 解码器也引入了残差连接, 将拼接后的数据和每个模块第二个卷积层的输出通过 Addition 进行对应元素相加。采用 BN 层和 ReLU 激活函数作为卷积层的预激活单元, 提高模型的训练效率和分类性能。残差 U-Net 摒弃了 U-Net 中的下采样层, 利用卷积层进行下采样, 防止池化层将有用信息丢弃。

2.2 SE-Net

不同卷积核在挖掘输入特征时各有侧重, 通过不同卷积核所获得的通道对于分类任务具有不同的重要性。传统深度学习模型为输入的各个通道赋予相同权重, 可能导致关键特征通道提升分类模型性能的能力被抑制。

Jie 等^[3]提出的 SE-Net 能够捕获输入各个通道之间的关联关系, 对输入的各个通道进行权重

分配, 有针对性地强化或抑制部分通道, 提高分类模型的性能。SE-Net 往往作为一个可插入模块, 与其他卷积神经网络相结合, 提高卷积神经网络的性能。SE-Net 的结构如图 3 所示, 其中 Global Pool 代表全局平均池化层, Dense 代表全连接层。

如图 3 所示, SE-Net 的输入为 $H \times W \times C$ 的矩阵。首先对输入进行全局平均池化, 得到 $1 \times C$ 的矩阵, 通过对池化结果进行变形, 得到 $1 \times 1 \times C$ 的矩阵。然后依次利用两个全连接层和 Sigmoid 激活函数基于所得的 $1 \times 1 \times C$ 矩阵挖掘输入各个通道之间的关联关系, 得到 $1 \times 1 \times C$ 的权重矩阵, 其中, 两个全连接层分别设置 C/r 和 C 个神经元, 分别输出 $1 \times 1 \times C/r$ 和 $1 \times 1 \times C$ 的矩阵。其中, r 表示第二个全连接层的神经元个数是第一个全连接层神经元个数的 r 倍, r 的具体值根据需要进行设置。最后, 依次获取 SE-Net 输入的每个通道, 从 $1 \times 1 \times C$ 矩阵中获取相应权重, 将该权重与相应通道的所有元素相乘, 对相应通道进行强化。最终得到 $H \times W \times C$ 的矩阵, 该矩阵即 SE-Net 的输出。

3 协议识别模型的设计

真实环境捕获的网络流量常常存在协议类别分布不平衡的问题。现有的协议识别方法大多关注总体的协议识别准确率, 忽略了少数类协议识别准确率偏低的问题。增强协议识别模型的特征提取能力, 提取更丰富协议数据特征, 能够在一定程度上提升少数类协议的识别准确率。但特

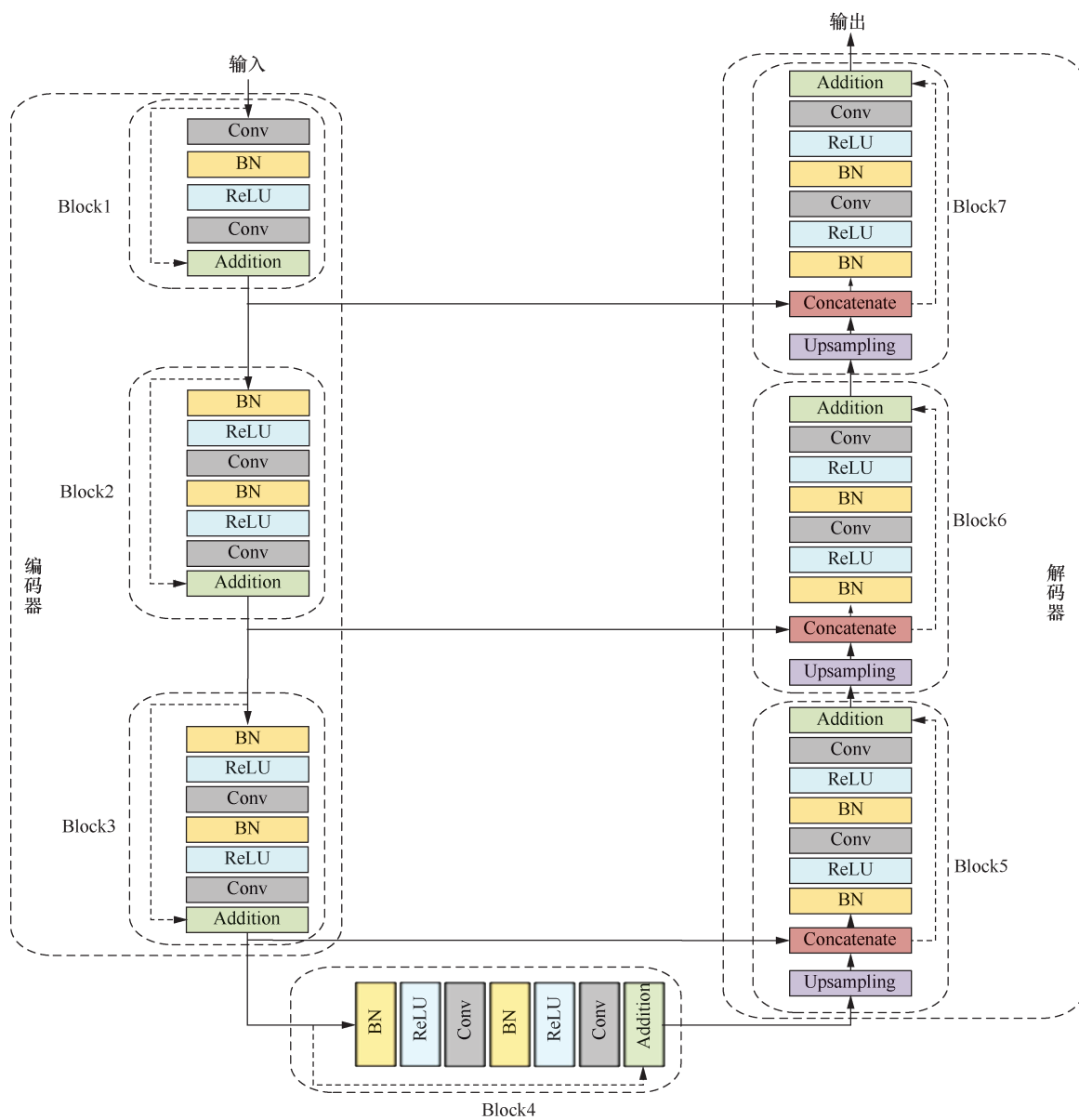


图 2 残差 U-Net 的结构
Figure 2 Structure of residual U-Net

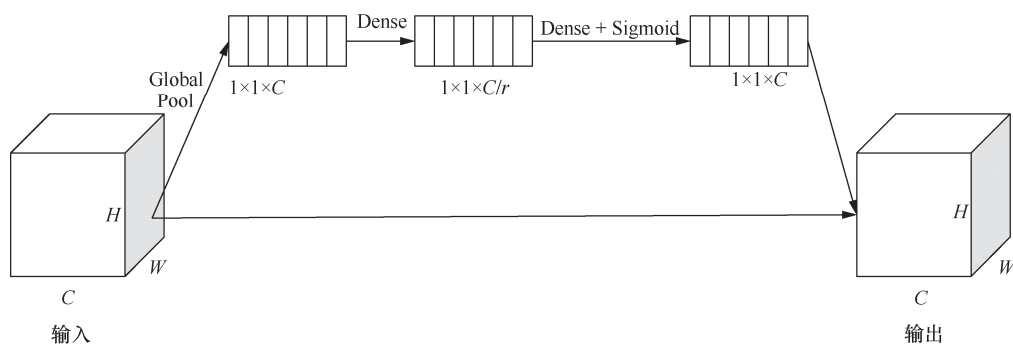


图 3 SE-Net 的结构
Figure 3 Structure of SE-Net

征提取能力的提升并不能从本质上解决协议类别不平衡问题,因为在协议识别模型的优化过程中,多数类协议往往主导着模型的优化方向。一种可行的解决方法是在保证协议识别模型特征提取能力的同时对损失函数进行完善,当多数类协议的识别准确率较高而少数类协议的识别准确率偏低时,利用损失函数引导模型通过提高少数类协议识别准确率的方式降低损失函数的值。

基于以上分析,本文一方面提升协议识别模型的特征提取能力,另一方面采用合理的损失函数提升少数类协议的识别准确率。本文提出基于改进的残差 U-Net 的不平衡协议识别方法,利用 SE-Net 强化残差 U-Net 各模块的输出,提升关键通道的影响力。传统激活函数对输入进行非线性激活时,存在信息损失的问题,可能导致模型分类性能下降。本文引入新的非线性激活函数,对输入进行激活时减少信息损失,提高残差 U-Net 的特征提取能力。提升模型的特征提取能力有助于提升少数类协议的识别准确率,原因是所提取的少数类协议特征越丰富,模型越容易判定少数类协议样本所属类别。与此同时,本文提出带权重 Dice 损失函数作为协议识别模型的损失函数,为少数类协议分别赋予较大的权值,确保少数类协议在协议识别模型中得到充分关注,有助于进一步解决少数类协议识别准确率较低的问题。

本文协议识别流程如图 4 所示,包括网络流量采集、数据预处理、分类模型构建以及模型预测 4 个步骤。网络流量采集步骤主要利用抓包工具捕获网络流量数据,为后续协议识别提供数据。数据预处理步骤是对采集的网络流量进行预处理,将原始网络流量数据转化为固定长度的一维矩阵,作为协议识别模型的输入,包括网络流重组与切分、网络协议标签标注等工作。分类模型构建步骤对协议识别模型进行训练,根据预测协议类型和真实协议类型的差异优化分类模型。模型预测步骤是将协议类型未知的数据输入协议识别模型,由协议识别模型提取数据特征,并输出协议类型结果。

下面主要对数据预处理、分类模型构建和模型预测进行介绍。

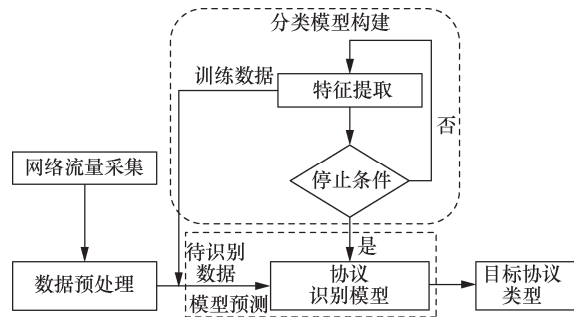


图 4 本文协议识别流程
Figure 4 Workflow of the proposed protocol recognition

3.1 数据预处理

数据预处理是协议识别的基础,也是协议识别模型准确识别协议的前提。数据预处理包括网络流重组与切分、网络协议标签标注两个步骤。首先,对将多个数据包聚合成网络流,从网络流中截取固定长度的数据用于协议识别。其次,对网络流进行协议标签标注,表明每条网络流所属协议类型,以便后续进行训练和测试。以下介绍两个步骤的具体工作。

(1) 网络流重组与切分

在网络协议识别领域,分析对象为数据包或网络流。数据包一般是在线协议识别方法的分析对象,原因是数据包的分析相对简单,处理时间短,能够满足在线识别实时性的要求。网络流往往是离线识别的分析对象,因为网络流能够将多个数据包聚合在一起,综合多个数据包的信息有助于准确地识别协议。本文采用离线识别的方法,以网络流作为协议识别的分析对象。

在网络协议识别领域,完整的 TCP 连接或者一次 UDP 交互代表一条网络流。对于 TCP 连接,可以用 TCP SYN 和 TCP FIN 数据包标识网络流的开始和结束,基于 TCP 首部的序号和标识位将捕获的数据包按顺序组成一条有序的网络流。对于 UDP 的网络交互,通常设置一个时间阈值。第一个捕获的 UDP 数据包的发送时间表示网络流的开始时间。后续捕获的 UDP 数据包,计算其发送时间和网络流开始时间的差值,如果差值小于时间阈值,表明当前网络流还未结束,将数据包拼接到网络流的尾部;如果差值大于或等于时间阈值,表明当前网络流已结束,所捕获的数据包属于下一条网络流。

本文关注应用层协议识别,在将多个数据包

拼接成有序网络流时, 仅保留数据包应用层部分的数据进行协议识别。TCP SYN 和 TCP FIN 数据包分别用于建立和结束连接, 不包含与协议识别相关的信息。在网络流重组后, 这些数据包需要剔除。

神经网络对输入数据有格式要求, 因此需要截取固定长度的数据用于协议识别。对于应用层协议而言, 首部格式往往较为固定, 主要用于传递交互过程中的控制信息。载荷部分为通信双方交换的数据, 随机性较大, 对于识别协议的作用相对较小。在将多个数据包重组为网络流后, 每条网络流内部数据包的首部和载荷部分将交替出现, 但由于大量应用层协议的协议规范没有公开, 无法准确确定首部和载荷的边界, 为尽可能保留首部信息, 本文从网络流第一个字节开始, 截取长度为 L 字节的数据, 将每个字节转化为 0~255 的值, 得到长为 L 的一维矩阵, 用于后续协议识别。 L 值的具体设置在实验部分通过实验进行确定。

(2) 网络协议标签标注

为训练和测试协议识别模型, 需要为数据集进行协议标签标注。本文采用独热编码处理的方法。假设协议类别总数为 C , 则可以利用长为 C 的一维向量表示协议标签。一维向量中除了有一位元素设置为 1 外, 其余各位元素均为 0。如果向量中第 n 位的元素为 1, 则表示相应的网络流属于第 n 种协议。例如, 数据集包含 9 类协议, 则 $[1, 0, 0, 0, 0, 0, 0, 0, 0]^T$ 表示第一种协议, $[0, 1, 0, 0, 0, 0, 0, 0, 0]^T$ 表示第二种协议, 其余协议依次类推。

3.2 分类模型构建

在协议类别分布不平衡时, 多数类协议会主导协议识别模型的优化方向, 在总体协议识别准确率较高时, 少数类协议的识别准确率还是比较低。采用特征融合方式能够有效提升分类模型的特征提取能力, 促使多数类协议和少数类协议的识别准确率都得到提高。同时, 针对协议类别不平衡的特点设计损失函数, 解决多数类协议始终主导模型优化方向的问题, 能够确保模型更准确地识别少数类协议。

3.2.1 SE-Net 的改进

由于 SE-Net 能够分析输入数据的所有通

道对于当前分类任务的重要性, 为不同通道分配不同的权重, 进而选择性抑制或强化部分通道, 因此本文将 SE-Net 与残差 U-Net 相结合, 提高残差 U-Net 的特征提取能力。为了更准确地识别少数类协议, 本文对 SE-Net 的改进如下。

1) SE-Net 的激活函数需要完善。传统深度学习模型往往采用 ReLU 激活函数作为卷积层和全连接层的激活函数, 原因是 ReLU 激活函数具有防止模型过拟合、避免出现梯度消失和提升模型训练效率等优点。但是当输入大于 0 时, ReLU 激活函数输入的值越大, 对应输出的值也越大。由于各层参数变化幅度的大小与激活函数输出值的大小呈正相关的关系, 在模型优化过程中, 使用 ReLU 激活函数会出现各层参数值的差异增大导致模型分类性能降低的问题。ReLU6 激活函数与 ReLU 激活函数类似。不同之处在于, 当输入大于 6 时, ReLU6 激活函数的输出为固定值, 能够有效缓解在模型优化过程中各层参数值的差异逐渐增大的问题, 提高模型的分类精度。Howard 等^[13]利用新的非线性激活函数 h-swish 来取代 ReLU 激活函数, 同时利用 $\text{ReLU6}(x+3)/6$ 取代 Sigmoid 激活函数, 达到提高神经网络模型特征提取能力的目的。h-swish 激活函数的计算如式 (2) 所示。

$$\text{h-swish} = x \frac{\text{ReLU6}(x+3)}{6} \quad (2)$$

其中, x 为输入数据, ReLU6 表示激活函数。本文将 h-swish 激活函数作为 SE-Net 中卷积层的激活函数, 利用激活函数 $\text{ReLU6}(x+3)/6$ 代替 SE-Net 中原有的 Sigmoid 激活函数, 降低非线性激活过程中的计算复杂度, 提高特征提取效率。

2) 与全连接层相比, 1×1 卷积层能够显著降低模型参数规模^[14], 本文利用 1×1 卷积层代替 SE-Net 中的全连接层, 减小模型的参数规模, 避免参数过多导致模型训练需要耗费大量时间, 从而提高模型的特征提取效率。

本文改进的 SE-Net 的结构如图 5 所示, 其中 Conv 表示卷积层, GAP+Re 表示全局平均池化层和变形层, 变形层的作用是调整输入的形状, 使其能够满足后续卷积层对输入的要求。

$\text{ReLU6}(x+3)/6$ 代表激活函数。输入 x_0 是 $H \times C$ 的矩阵。在处理时, 首先对输入 x_0 施加全局平均池化和变形, 得到 $1 \times C$ 的矩阵; 然后依次利用两个 1×1 卷积层基于所得 $1 \times C$ 矩阵分析输入 x_0 的各个通道的重要性, 并利用 $\text{ReLU6}(x+3)/6$ 激活函数进行非线性激活, 两个 1×1 卷积层分别设置 $C/4$ 个和 C 个 1×1 的卷积核, 相应输出 $1 \times C/4$ 和 $1 \times C$ 的矩阵; 最后利用所得 $1 \times C$ 矩阵对输入 x_0 的各个通道进行强化, 方法是依次获取输入 x_0 的每个通道, 并从所得 $1 \times C$ 矩阵获取该通道对应的权重, 将权重与该通道所有元素相乘, 得到输出 x_1 。与输入 x_0 相比, 输出 x_1 的各通道得到有效强化, 重要通道使后续协议识别任务的影响力得到有效提升。

与原始 SE-Net 相比, 改进的 SE-Net 利用 1×1 卷积层降低 SE-Net 的参数规模, 减少训练 SE-Net 所需时间, 提升特征提取效率, 利用激活函数 $\text{ReLU6}(x+3)/6$ 进一步提升 SE-Net 提取协议数据特征的效率。

3.2.2 残差 U-Net 基本模块的改进

本文将改进的 SE-Net 融入残差 U-Net 中, 改进的残差 U-Net 的基本模块如图 6 所示。其中, BN1 和 BN2 代表批量归一化层, h-swish1 和 h-swish2 为激活函数 h-swish, Conv1 和 Conv2 代表卷积层, Addition 代表对应元素相加的操作, SE 代表本文设计的 SE-Net。利用 BN1 和激活函数 h-swish1 对输入进行激活后, 将处理结果输入 Conv1 中。利用 BN2 和激活函数 h-swish2 对 Conv1 的输出进行预激活, 将处理结果输入 Conv2。在此基础上, 利用 Conv3 对基本模块的输入进行升维, 将升维的结果与 Conv2 的输出的

对应元素相加, 最后利用 SE-Net 挖掘输入各个通道的关联关系, 为不同通道赋予不同权重, 提升关键通道对后续分类任务的影响力。SE-Net 会强化所有协议的特征, 有助于同时提高多数类协议和少数类协议的识别准确率。

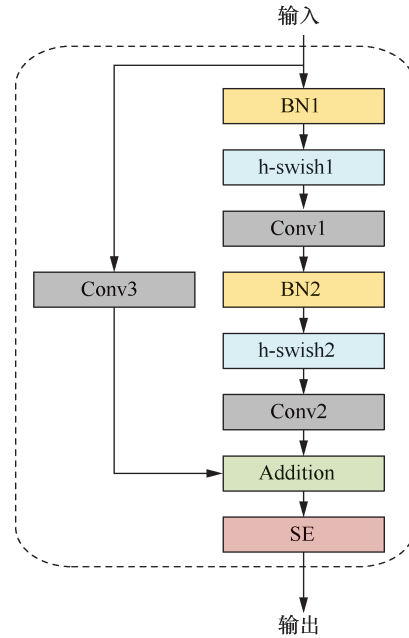


图 6 改进的残差 U-Net 的基本模块
Figure 6 Structure of basic block of improved residual U-Net

与原始残差 U-Net 的基本模块相比, 本文改进的基本模块利用 SE-Net 强化所挖掘协议数据特征的不同通道, 利用激活函数 h-swish 避免出现性能退化的情况, 能够挖掘更具代表性的协议数据特征, 提升利用基本模块构建的协议识别模型的性能。

3.2.3 协议识别模型的整体结构

基于改进的残差 U-Net 的基本模块, 本文设

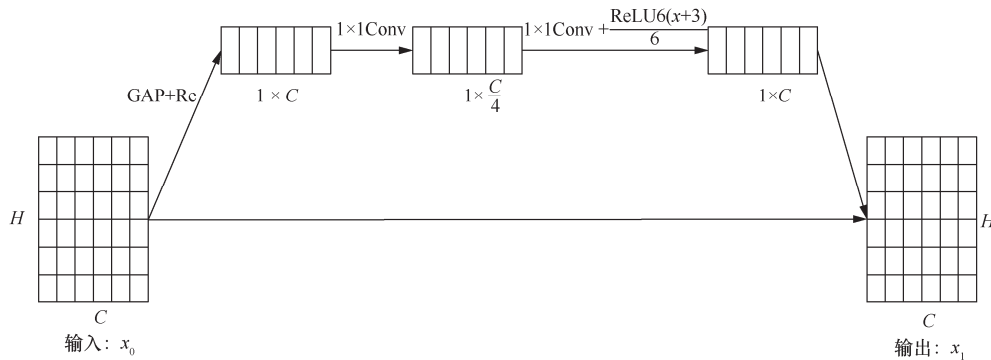


图 5 改进的 SE-Net 的结构
Figure 5 Structure of the improved SE-Net

计了协议识别模型，其整体结构如图 7 所示，输入为预处理后的协议数据，输出为预测的协议标签。

本文协议识别模型的编码器利用了多个改进的残差 U-Net 的基本模块提取协议数据的特征，剔除冗余特征，得到具有代表性的协议数据特征；解码器利用多个结构相同的模块，将编码器的输出恢复成编码器的输入，每个模块在改进的残差 U-Net 基本模块的基础上增加了转置卷积层和拼接层。在图 7 中，Embedding 表示嵌入层，BN 表示批量归一化层，h-swish 表示激活函数，Conv 表示卷积层，Addition 表示对应元素相加的操作，SE 表示本文设计的 SE-Net。ConvTrans 表示转置卷积层，用于上采样。Dense 表示全连接层，Concatenate 表示按通道维度进行拼接的操作。

输入 A 能够和输入 B 进行 Concatenate 的条件是输入 A 和输入 B 除通道数量不同外，每个通道的形状均相同。依次获取输入 B 的每个通道，将其视为输入 A 的一个新通道。从上述拼接过程可知，输入 A 和输入 B 的顺序不同，对应的拼接结果不同。在协议识别模型中，A 和 B 为协议识别模型两个不同层的输出，两个层与输出层的距离不同，所提取的协议数据特征的抽象程度不同，信息损失程度也不同，因此，基于通道拼接后所得结果进一步挖掘协议数据的特征，能够综合输入 A 和输入 B 的信息所挖掘的协议数据特征，更为全面，协议识别结果更为准确。

由于网络流中各个字节之间并不是相互独立的，字节之间可能存在关联关系，如字节“G”“E”和“T”组成 HTTP 协议报文的关键字段“GET”，挖掘字节之间的关联关系有助于提升模型的训练效率和协议识别能力。词嵌入层利用词嵌入技术挖掘输入各个字节之间的关联关系。在此过程中，将每个字节随机初始化为一个固定长度的一维矩阵。对于一条网络流，依次获取其中各个字节对应的一维矩阵。

将每个字节对应的一维矩阵替换相应的字节，网络流被转化成一个二维矩阵。将所得矩阵作为卷积层 Conv0 的输入，经过后续各层的处理后，利用分类器计算输出预测协议类别。在此基

础上，根据预测协议类别和真实协议类别计算分类误差，利用反向传播算法将误差传递至协议识别模型的各层，各层根据误差调整参数。嵌入层会根据误差更新每个字节对应的一维矩阵，当分类误差趋于稳定时，能够得到各个字节所对应的一维矩阵。

举例来看，嵌入层的输入为长为 L 的一维矩阵，矩阵的每个元素代表网络流中的一个字节。每个元素被映射为长度为 M 的一个随机的一维矩阵， L 个元素对应 L 个一维矩阵。将 L 个一维矩阵视为一个二维矩阵的 L 个行矩阵。按元素的先后顺序拼接出一个 $L \times M$ 的二维矩阵，所得的二维矩阵作为嵌入层的输出。将 $L \times M$ 的二维矩阵输入卷积层 Conv0，利用批量归一化层和激活函数对卷积层 Conv0 的输出进行处理。Conv0 包含 16 个大小为 5、步长为 4 的卷积核。其中，嵌入层根据分类误差，更新嵌入层输入的每个元素对应的一维矩阵。当协议识别模型达到最优时，得到每个元素所对应的一维矩阵。词嵌入技术通常将输入的每个元素映射成长度为 100~300 的一维矩阵。在本文模型中，将网络流的每个字节映射成长度为 128 的一维矩阵，即 $M=128$ 。

在图 7 中，Block1、Block2、Block3、Block4 采用的都是改进的残差 U-Net 的基本模块结构，利用它们来挖掘数据特征。Block1~Block4 除卷积层的卷积核个数不同，其余设置均相同。Block1~Block4 的卷积层的卷积核个数分别设置为 16、32、64 和 128。下面以 Block1 为例介绍模块提取协议数据特征的过程。首先利用批量归一化层和 h-swish 激活函数对 Block1 的输入进行预激活，而后输入 Conv1_1。Conv1_1 包括 16 个大小为 5、步长为 2 的卷积核。利用批量归一化层和 h-swish 激活函数对 Conv1_1 的输出进行预激活，而后输入 Conv1_2。Conv1_2 包括 16 个大小为 5、步长为 1 的卷积核。最后，对于 Block1 的输入，利用 Conv1_3 进行升维和下采样，保证 Conv1_3 和 Conv1_2 的输出的各个通道的形状相同，从而能够进行通道拼接。Conv1_3 包括 16 个大小为 1、步长为 2 的卷积核，将 Conv1_3 的输出和 Conv1_2 的输出对应元素相加，输入 SE 模块中，得到 Block1 的输出。由于 Conv1_3 包括 16 个卷积核，

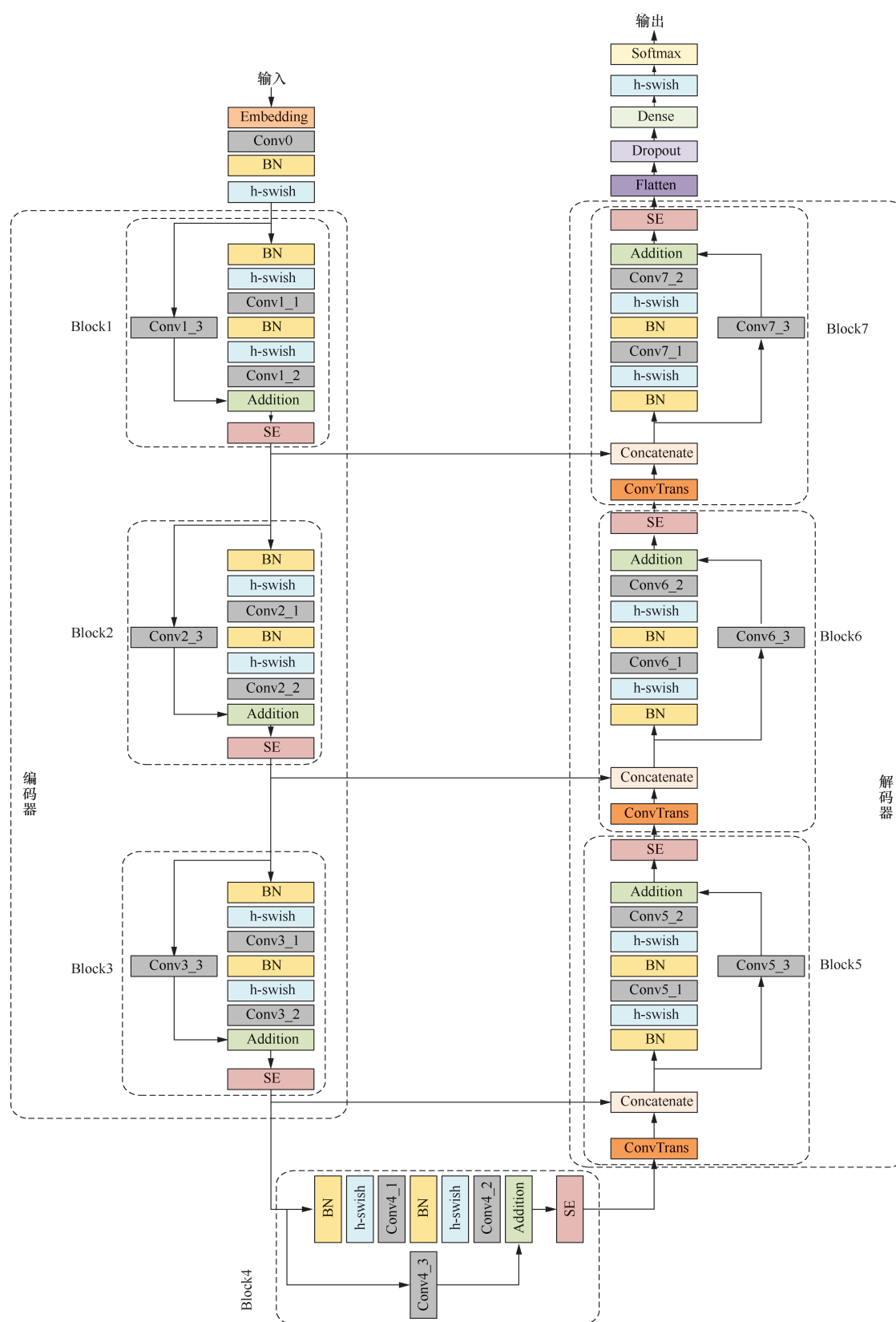


图 7 协议识别模型的整体结构
Figure 7 Structure of the protocol recognition model

即 SE 模块的输入包含的通道数为 16, SE 模块的两个 1×1 卷积层分别设置为包含 4 个和 16 个大小为 1、步长为 1 的卷积核。

经过嵌入层和 Block1~Block4 处理后, 协议数据中的冗余特征被剔除, 仅保留了最具有区分度的协议特征, 同时通过下采样的处理, 网络参数的规模得到有效降低。在此基础上, 依次利用 Block5~Block7 将 Block4 的输出恢复成协议识别模型的输入, 即使 Block7 的输出和 Block1 的输入尽量相同, 但与 Block1 的输入相比, Block7 的输出包含更细粒度的协议数据特征, 能够提高模型的分类性能。其中, Block7 的输出为 Block7 模块中 SE 的输出结果, Block1 的输入为 Block1 模块之前激活函数 h-swish 的输出。由于 Block5~Block7 的结构相同, 均是先进行上采样, 随后采用按通道维度拼接的方式融合该模块之前某一个模块的输出, 拼接的结果输入改进的残差 U-Net 基本模块。Block5~Block7 除卷积层的卷积核个数不同外, 其余设置均相同, 其中, Block5、Block 6 和 Block 7 分别设置 64、32 和 16 个卷积核。以下以 Block 5 为例介绍模块提取协议数据特征的过程。Block 5 首先利用 ConvTrans 对 Block4 的输出进行上采样, 转置卷积层包括 64 个大小为 5、步长为 2 的卷积核。转置卷积层的输出与 Block3 的输出的各个通道的形状均相同, 只是通道数量存在差异, 将转置卷积层的输出和 Block3 的输出按通道维度进行拼接。然后依次利用 Conv5_1 和 Conv5_2 提取拼接结果的特征, 在进行卷积运算前均利用批量归一化层和 h-swish 激活函数对输入进行预激活。Conv5_1 和 Conv5_2 均包括 64 个大小为 5、步长为 1 的卷积核。最后利用 Conv5_3 挖掘拼接结果的特征, Conv5_3 包括 64 个大小为 1、步长为 1 的卷积核。将 Conv5_3 的输出和 Conv5_2 的输出对应元素相加, 将结果输入 SE 模块中, 得到 Block5 的输出。由于 Block5 中 Conv5_3 包含 64 个卷积核, SE 模块的输入包含 64 个通道, Block5 的 SE 模块中两个卷积层相应地分别设置为包含 16 个和 64 个大小为 1、步长为 1 的卷积核。

协议识别模型最终将输出一个一维向量。向量的长度为目标协议类别的总数, 各分量表示协

议识别模型的输入属于对应协议类别的概率。所设计的模型利用 Flatten 层将 Block7 的输出展开成一维矩阵, 而后将得到的一维矩阵输入 Dropout 层和全连接层, 利用激活函数 h_swish 对全连接层的输出进行激活。其中, Dropout 层用于防止模型过拟合, 全连接层用于提取协议数据的特征。最后将激活后的数据输入 Softmax 分类器, 由分类器输出预测的协议类型。其中, 全连接层设置 64 个神经元, Softmax 分类器的神经元个数与协议类别的数量相同。

在协议识别模型中, 改进的 SE-Net 能够对少数类协议的特征进行强化, 采用按通道进行拼接的方式进行特征融合, 保证协议识别模型能够提取丰富的协议数据特征。与此同时, 采用词嵌入技术和残差连接使模型挖掘到少数类协议更丰富的特征。由于传统激活函数在模型优化过程中存在信息损失的问题, 本文引入新的激活函数, 避免模型优化过程中信息损失的问题, 提升模型训练的效率。这些设计能显著提升少数类协议的识别准确率, 有助于进一步提升多数类协议识别的准确率。

3.2.4 协议识别模型训练过程

协议识别模型的训练包括正向传播和反向传播两个过程。正向传播是将协议数据输入协议识别模型中, 模型自动挖掘协议数据具有代表性的特征, 并基于特征输出目标的协议类型。反向传播利用损失函数计算预测协议类型和真实协议类型的误差, 将误差通过反向传播算法从 Softmax 分类器反向传递至嵌入层。模型根据误差更新网络参数。正向传播和反向传播交替进行, 误差值逐渐减小, 预测的协议类型逐渐接近真实协议类型。当误差值趋于稳定时, 协议识别模型的优化过程将停止, 训练成熟的协议识别模型被保存下来。

Milletari 等^[4]针对样本类别分布不平衡的问题设计了 Dice 损失函数, 主要用于解决图像领域类别分别不平衡的问题。Dice 损失函数的计算如式(1)所示, 分子为每类所有样本对应分类误差的总和, 分母为每类所有样本对应预测标签和真实标签的平方和, 利用除法相消这一机制, 使模型分类误差与各类别的样本数量之间的关联性较小。因

此,在模型优化过程中,多数类和少数类基于对模型整体分类误差的影响,交替主导模型的优化方向,保证分类模型能够准确识别少数类样本。

本文将 Dice 损失函数引入协议识别领域,构建带权重 Dice 损失函数作为协议识别模型的损失函数。由于 Dice 损失函数为各类别赋予相同的误分类代价(即权重),多数类的错误分类的样本数量往往明显多于少数类的错误分类的样本数量。在 Dice 损失函数的分子和分母中,多数类对应的分类误差将占据主导地位。在模型优化过程中,模型更加关注多数类,导致少数类的准确率提升有限。Dice 损失函数对于消除协议类别分布不平衡问题的能力有限。

为进一步增强 Dice 损失函数解决协议类别分布不平衡的能力,本文的协议识别模型为少数类协议和多数类协议分别赋予较大和较小的权重,对各类样本的预测标签进行加权,得到带权重 Dice 损失函数,计算式如式(3)所示。

$$W_Dice = 1 - \frac{2 \sum_{s=1}^K w_s \left(\sum_{i=1}^{n_s} p_i t_i \right) + \varepsilon}{\sum_{s=1}^K w_s \left(\sum_{i=1}^{n_s} p_i^2 + \sum_{i=1}^{n_s} t_i^2 \right) + \varepsilon} \quad (3)$$

其中, K 表示协议类别的数目, $s=1,2,\dots,K$ 表示第 s 类协议, n_s 表示第 s 类协议包含的网络流数量, ε 表示一个固定的常数, p_i 表示网络流 i 对应的预测协议标签, t_i 表示网络流 i 对应的真实协议标签, w_s 表示协议类别 s 对应的权重。

在计算某类协议的误分类代价时,将网络流总数量除以该类协议包含的网络流数量,所得的结果作为该类协议的权重。假设网络流总数为 10 000, A 和 B 分别为两种网络协议,包含的网络流数目分别为 500 和 5000,则类别 A 和类别 B 的权重分别为 20 和 2。每类协议对应的权重涉及该类协议所包含的网络流数量以及捕获的网络流的总数量,能够充分反映数据集内部各协议类别的分布状况,权重的设置是科学合理的。

在式(3)中,由于少数类协议的误分类代价大于多数类协议的误分类代价,虽然少数类协议错误分类的样本数量少于多数类协议错误分类的样本数量,但通过加权处理后,模型将关注少数类

协议,主动提高少数类协议的识别准确率。与 Dice 损失函数相比,带权重 Dice 损失函数通过为各类协议赋予不同权值,使模型分类误差与每类协议错误分类的样本数量的关联性进一步减小,能够更有效地解决类别分布不平衡问题。

3.3 模型预测

在模型预测步骤,对捕获的网络流量首先进行预处理,然后输入协议识别模型,协议识别模型利用嵌入层挖掘协议数据各字节之间的关联关系,通过改进的残差 U-Net 基本模块反复挖掘协议数据的特征,最后利用 Softmax 分类器基于特征信息输出预测的协议类型。

4 实验结果及分析

本节通过实验完善协议识别模型的参数设置,进而验证协议识别模型的有效性。本文在实验过程中主要考虑了以下几方面问题:

- 1) 截取的网络流长度对协议识别模型性能的影响;
- 2) 带权重 Dice 损失函数对协议识别模型性能的影响;
- 3) 新激活函数对协议识别模型性能的影响;
- 4) SE-Net 的使用与否对协议识别模型性能的影响;
- 5) 与其他不平衡协议识别方法进行横向对比,评估本文模型的性能表现。

本文利用 PyCharm、Python3.7 和 Tensorflow2.0 实现了协议识别模型,实验平台为 Intel i7-10750,主频为 2.60 GHz,内存 13 GB,64 位 Ubuntu16.04 操作系统。采取 SGD 优化函数,初始学习率为 0.01,学习率衰减系数为 1×10^{-6} 。batch size 设置为 128,训练轮次设置为 20,采取 L2 正则化防止模型过拟合。

4.1 实验数据集

为验证所提协议识别模型的有效性和通用性,本文利用网络流量公开数据集 ISCX2012^[15]、UNSW-NB15^[16]进行实验,这两个数据集包含的协议类型丰富,能够较好地反映当前网络发展现状。本文从这两个数据集中提取出 9 种协议,分别是 BitTorrent、DNS (domain name system)、FTP (file transfer protocol)、HTTP (hyper text transfer

protocol)、IMAP (internet message access protocol)、POP (post office protocol)、SMB (server message block)、SMTP (simple mail transfer protocol) 和 SSH (secure shell)。这些协议既包括二级制协议, 也包括文本类协议; 既包括基于 UDP 传输的网络协议, 也包括基于 TCP 传输的网络协议。按照本文预处理方法, 所得到的实验数据集 ISCX2012 和 UNSW-NB15 的统计信息分别如表 1 和表 2 所示。

表 1 数据集 ISCX2012 的统计信息
Table 1 Description of ISCX2012 dataset

协议	网络流数量	占比	不平衡系数
BitTorrent	2 282	1.82%	8.76
DNS	17 359	13.88%	1.15
FTP	19 289	15.42%	1.04
HTTP	12 808	10.24%	1.56
IMAP	19 515	15.6%	1.02
POP	17 466	13.97%	1.14
SMB	19 998	15.99%	1
SMTP	1 497	1.2%	13.36
SSH	14 850	11.87%	1.35
合计	125 064	100%	

表 2 数据集 UNSW-NB15 的统计信息
Table 2 Description of UNSW-NB15 dataset

协议	网络流数量	占比	不平衡系数
BitTorrent	19 855	20.43%	1
DNS	10 000	10.29%	1.99
FTP	18 348	18.88%	1.08
HTTP	6 392	6.58%	3.11
IMAP	18 081	18.61%	1.1
POP	1 599	1.65%	12.42
SMB	1 455	1.5%	13.65
SMTP	11 526	11.86%	1.72
SSH	9 914	10.2%	2
合计	97 170	100%	

其中, 占比表示每类协议对应的网络流数量占网络流总数的比例; 不平衡系数用于判断数据集是否存在类别分布不平衡的现象。计算不平衡系数时, 首先确定网络流数量最多的一类协议, 然后用该类协议包含的网络流数量除以其他各类

协议的网络流数量, 得到每类协议的不平衡系数。如果有协议的不平衡系数大于 10, 表示数据集存在类别分布不平衡的现象; 如果有协议的不平衡系数大于 100, 表示数据集存在类别分布严重不平衡的现象^[8]。从表 1 和表 2 的不平衡系数可以看出, 两个数据集均存在协议类别分布不平衡的现象。

将协议数据集随机划分为训练集和测试集, 训练集用于训练模型, 测试集用于评估模型的性能, 两者的比例为 3:1。

4.2 评价指标

本文采用总体准确率 (A, Accuracy)、精确率 (P, Precision)、召回率 (R, Recall)、 F 值和分类时间评价所提协议识别模型的有效性。其中, 分类时间指从协议数据输入到协议识别模型输出预测结果所需的时间。

假设数据集包含的网络流总数为 K , 准确率、精确率、召回率和 F 值的计算公式定义如下。

$$\begin{aligned}
 A &= \frac{\sum_i^K (TP_i + TN_i)}{\sum_i^K (TP_i + FP_i + FN_i + TN_i)} \\
 P_i &= \frac{TP_i}{TP_i + FP_i} \\
 R_i &= \frac{TP_i}{TP_i + FN_i} \\
 F &= \frac{2PR_i}{P_i + R_i}
 \end{aligned} \quad (4)$$

其中, $i(i=1, 2, \dots, K)$ 表示第 i 类协议。 TP_i 表示预测协议类别和真实协议类别均为 i 的网络流数目, FP_i 表示预测协议类别为 i 但真实协议类别不为 i 的网络流数目, TN_i 表示预测协议类别和真实协议类别均不为 i 的网络流数目, FN_i 表示预测协议类别不为 i 但真实协议类别为 i 的网络流数目。准确率用于评估协议识别模型的总体分类性能。精确率、召回率和 F 值用于评价某类协议的分类效果。准确率、精确率、召回率和 F 值 4 个指标值的取值均为 $[0, 1]$, 值越接近 1, 表示模型分类效果越好。

4.3 实验结果与分析

4.3.1 截取的网络流长度的影响

神经网络对输入数据有格式要求, 因此协议

识别模型需要截取网络流中固定长度 L 的数据用于进行协议识别。在将数据包重组为有序网络流后, 本文从网络流首部截取长度 L 字节的数据用于进行协议识别。

为确定合理的 L 的值, 本文统计两个实验数据集所有网络流的长度, 通过分析网络流长度的分布确定合理的截取长度。网络流长度分布如图 8 所示。其中, 横轴表示网络流长度, 纵轴表示相应长度范围的网络流在总网络流中所占的比例。

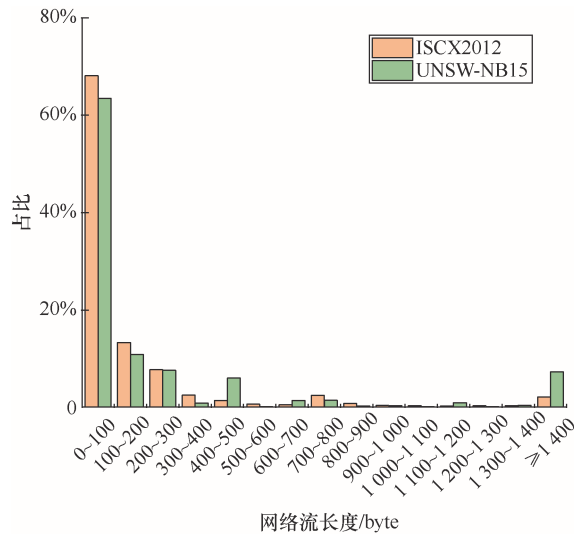


图 8 网络流长度分布

Figure 8 Distribution of length of network flows

从图 8 可以看出, 对于数据集 ISCX2012、UNSW-NB15, 分别有 89.07% 和 81.84% 的网络流, 长度为 0~300 byte。长度大于 300 byte 的网络流在两个数据集中数量均比较少, 并且在网络流尾部往往包含较多用户数据。为尽可能保留网络流的关键信息, 同时避免包含较多随机性较大的用户数据导致模型分类性能下降, 本文从网络流首部开始截取 256 byte 的数据进行协议识别。

4.3.2 带权重 Dice 损失函数的影响

Dice 损失函数有助于消除协议类别分布不平衡的影响, 确保多数类协议和少数类协议交替主导模型优化方向, 使协议识别模型准确识别少数类协议。

由于 Dice 损失函数为每类协议分配相同权重, 模型仍旧倾向于提高多数类协议的识别准确率来降低协议识别模型的分类误差, 因此, 本文以 Dice 损失函数为基础, 根据协议类别分布, 为少数类协议赋予较大的权重, 为多数类协议赋予较

小的权重, 得到带权重 Dice 损失函数。通过为少数类协议分配较大权重, 当少数类协议识别准确率低而多数类协议识别准确率高时, 虽然此时少数类协议错误分类的样本数量小于多数类协议, 但由于权重较大, 能保证模型对少数类协议的关注。

为验证带权重 Dice 损失函数的有效性, 实验中, 将带权重 Dice 损失函数与 Dice 损失函数以及带权重交叉熵损失函数^[8]进行比较。带权重交叉熵损失函数是基于代价敏感的不平衡协议识别方法中常用的损失函数, 为少数类协议赋予较大的误分类代价, 为多数类协议赋予较小的误分类代价。实验中将采用带权重 Dice 损失函数、Dice 损失函数和带权重交叉熵损失函数作为损失函数的协议识别模型分别记为 WeightedDice、Dice 和 WeightedCrossEntropy, 评价指标为协议识别准确率和各类协议的 F 值。为保证对比实验的有效性, 采取本文权重计算方法计算权重交叉熵损失函数中各类协议的权重。不同损失函数对应的协议识别准确率和分类时间如表 3 所示, 采用不同损失函数时所提模型在数据集 ISCX2012 和 UNSW-NB15 上的 F 值如图 9 和图 10 所示。

表 3 不同损失函数对应的协议识别准确率和分类时间
Table 3 Protocol recognition accuracy and classification time of different loss functions

损失函数	ISCX2012		UNSW-NB15	
	协议识别 准确率	分类 时间/s	协议识别 准确率	分类 时间/s
WeightedCrossEntropy	98.32%	57	98.64%	45
Dice	98.60%	59	98.63%	48
WeightedDice	98.68%	60	98.74%	43

从表 3 可以看出, 与带权重交叉熵损失函数和 Dice 损失函数相比, 采用带权重 Dice 损失函数作为模型损失函数时, 协议识别模型的总体准确率在数据集 ISCX2012 上分别提升了 0.36% 和 0.08%, 在数据集 UNSW-NB2012 上分别提升了 0.10% 和 0.11%。从图 9 和图 10 可以看出, 与带权重交叉熵损失函数和 Dice 损失函数相比, 采用带权重 Dice 损失函数作为模型的损失函数时, 除少数几类协议外, 协议识别模型在多数类协议和少数类协议的 F 值均有所提高。从表 3、图 9 和图 10 可以发现, 采用 Dice 损失函数和带权重交叉熵损失函数作为模型损失函数时, 模型分类效果大致相同。

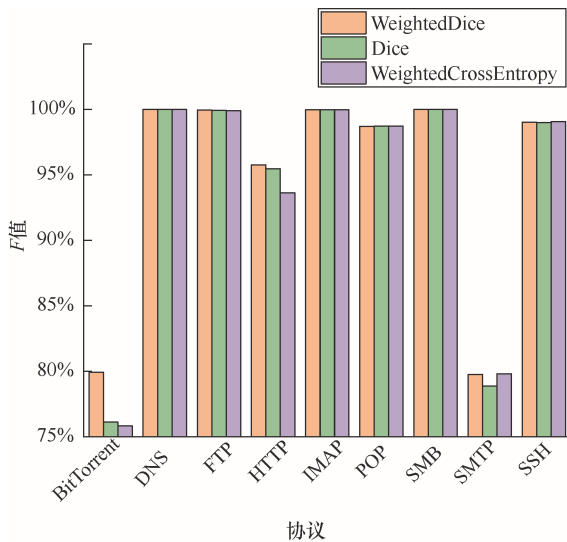


图 9 采用不同损失函数时所提模型在数据集 ISCX2012 上的 F 值
Figure 9 F of the proposed model on the dataset ISCX2012 using different loss functions

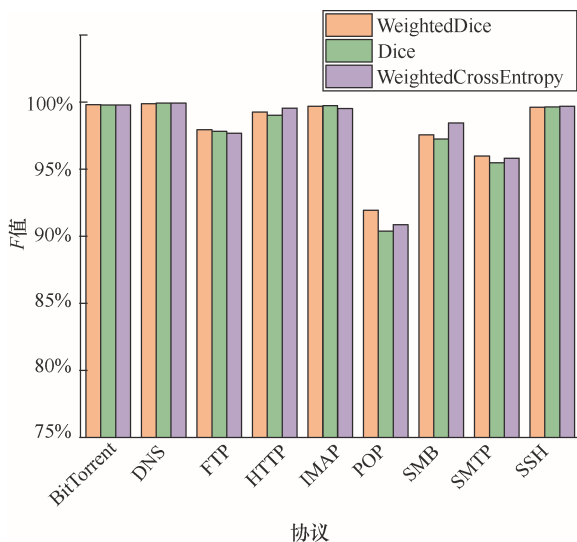


图 10 采用不同损失函数时所提模型在数据集 UNSW-NB15 上的 F 值
Figure 10 F of proposed model on the dataset UNSW-NB15 using different loss functions

从实验结果可以得出, Dice 损失函数和带权重交叉熵损失函数均能提高少数类协议的识别准确率, 但是带权重交叉熵损失函数的稳定性不如 Dice 损失函数, 提升少数类协议识别准确率的能力不如 Dice 损失函数。这是因为带权重交叉熵损失函数较为难为各类协议分配合理的权重, 并且在模型优化过程中, 无法动态更新各类协议的权重。本文采用的带权重 Dice 损失函数能够结合带权重交叉熵损失函数和 Dice 损失的优势, 充分利用除法运算的特性, 同时通过权重的合理设置, 确保更准确地识别少数类协议。

4.3.3 新激活函数的影响

为验证激活函数 h-swish 和 h-sigmoid 能够有效提高协议识别模型的性能, 实验中设计了 4 个对比模型: ① 将激活函数 ReLU 和 Sigmoid 作为本文模型的激活函数, 记为 Model1; ② 将激活函数 ReLU 和 $\text{ReLU6}(x+3)/6$ 作为本文模型的激活函数, 记为 Model2; ③ 将激活函数 h-swish 和 Sigmoid 作为本文模型的激活函数, 记为 Model3; ④ 将激活函数 h-swish 和 $\text{ReLU6}(x+3)/6$ 作为本文模型的激活函数, 记为 Model4。其中, 激活函数 ReLU 和 h-swish 作为卷积层和全连接层的激活函数, 激活函数 Sigmoid 和 $\text{ReLU6}(x+3)/6$ 作为本文设计的 SE-Net 中第二个 1×1 卷积层的激活函数。使用不同激活函数的协议识别准确率和分类时间如表 4 所示。

表 4 使用不同激活函数的协议识别准确率和分类时间
Table 4 Protocol recognition accuracy and classification time of different activation functions

模型	ISCX2012		UNSW-NB15	
	协议识别准确率	分类时间/s	协议识别准确率	分类时间/s
Model1	98.60%	58	98.67%	39
Model2	98.61%	60	98.70%	38
Model3	98.63%	68	98.70%	45
Model4	98.66%	61	98.76%	44

从表 4 可以看出, 仅利用激活函数 h-swish 和 $\text{ReLU6}(x+3)/6$ 替换激活函数 ReLU 和 Sigmoid, 模型的协议识别准确率均能得到有效提升, 但是 h-swish 激活函数会导致模型的分时间显著增加。同时, 利用激活函数 h-swish 和 $\text{ReLU6}(x+3)/6$ 作为协议识别模型的激活函数时, 模型能够综合两个激活函数的优势, 协议识别准确率最高。因此, 本文选择激活函数 h-swish 和 $\text{ReLU6}(x+3)/6$ 作为模型的激活函数。

4.3.4 SE-Net 网络的影响

为了验证 SE-Net 对协议识别模型分类性能的影响, 实验中设计了不采用 SE-Net 的对比协议识别模型, 即将本文协议识别模型中所有的 SE-Net 模块均去除, 记为模型 Model5, 评价指标为总体准确率和分类时间。使用 SE-Net 与不

使用情况下的总体协议识别准确率和分类时间如表 5 所示。

表 5 使用 SE-Net 与不使用情况下的总体协议识别准确率和分类时间

Table 5 Overall protocol recognition accuracy and classification time with and without SE-Net

模型	ISCX2012		UNSW-NB15	
	协议识别准确率	分类时间/s	协议识别准确率	分类时间/s
Model5	98.54%	47	98.69%	41
本文模型	98.65%	59	98.75%	42

从表 5 可以看出,与未使用 SE-Net 的协议识别模型相比,本文协议识别模型使用 SE-Net 对协议数据特征进行强化,提升了关键通道对后续协议识别任务的影响力,总体准确率在数据集 ISCX2012 和 UNSW-NB15 上分别提升了 0.11% 和 0.06%。

4.3.5 不同协议识别模型比较

为验证所提协议识别模型的有效性和通用性,选取目前效果较好的不平衡协议识别模型与本文模型进行对比实验。用于对比的协议识别模型包括 Zhang 等^[10]提出的 PCCN 模型、Xia 等^[11]提出的 RICNN 模型、常用集成算法随机森林 (RF, random forest) 模型^[8]。其中,PCNN、RICNN 和 RF 均是不平衡协议识别领域具有代表性的协议识别模型。PCCN 模型利用特征融合方式提高模型的特征提取能力,从而提高少数类协议的识别准确率。RICNN 模型利用残差连接和特征融合来提高模型的特征提取能力,达到准确识别少数类协议的目的。RF 模型利用多个基本分类器,通过集成策略达到准确识别少数类协议的目的。为保证对比实验的有效性,实验中采用本文的预处理方法对捕获的网络流量数据进行预处理,最终的不同模型的协议识别准确率和分类时间如表 6 所示,不同模型在数据集 ISCX2012、UNSW-NB15 上的准确率、召回率和 F 值如图 11~图 16 所示。

从表 6 可以看出,与协议识别模型 PCCN、RICNN 和 RF 相比,本文模型的准确率在数据集 ISCX2012 上分别提升了 1.52%、0.35% 和 27.2%,在数据集 UNSW-NB15 上分别提升了 3.14%、0.17% 和 32.18%。在两个数据集上分类时间均比模型

PCCN 和 RF 多,均比模型 RICNN 少。从图 11~图 16 可以看出,在数据集 ISCX2012 和 UNSW-NB15 上,本文模型对各类协议的精确率、召回率和 F 值均优于模型 PCCN、RICNN 和 RF。

表 6 不同模型的协议识别准确率和分类时间

Table 6 Protocol recognition accuracy and classification time of different models

模型	ISCX2012		UNSW-NB15	
	协议识别准确率	分类时间/s	协议识别准确率	分类时间/s
PCCN	97.26%	42	95.56%	33
RICNN	98.43%	253	98.53%	205
RF	71.58%	2	66.52%	2
本文模型	98.78%	63	98.70%	45

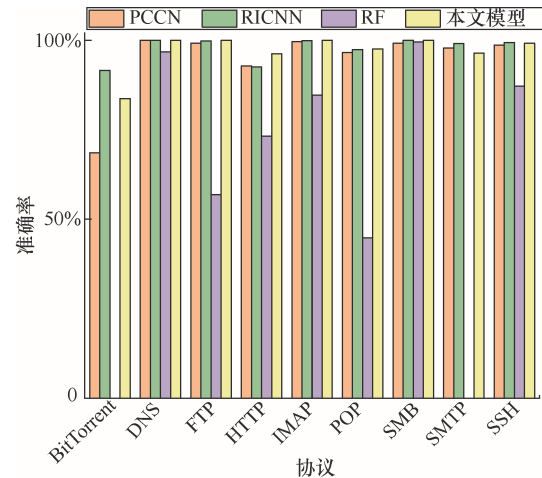


图 11 不同模型在数据集 ISCX2012 上的准确率
Figure 11 Precision of different models on dataset ISCX2012

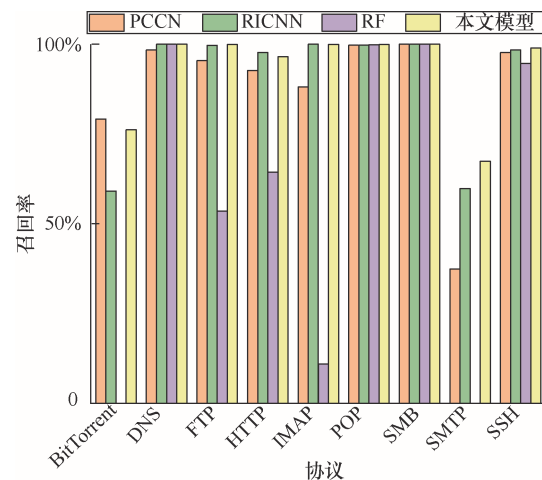


图 12 不同模型在数据集 ISCX2012 上的召回率
Figure 12 Recall of different models on dataset ISCX2012

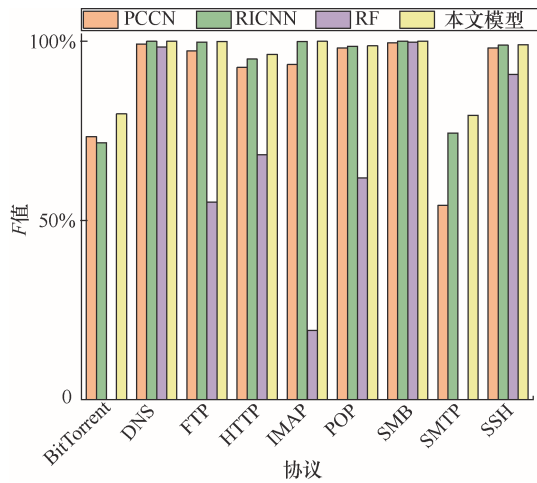


图 13 不同模型在数据集 ISCX2012 上的 F 值
Figure 13 F of different models on dataset ISCX2012

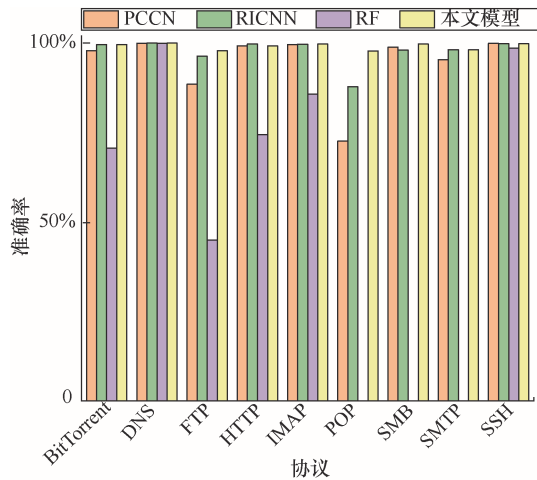


图 14 不同模型在数据集 UNSW-NB15 上的准确率
Figure 14 Precision of different models on dataset UNSW-NB15

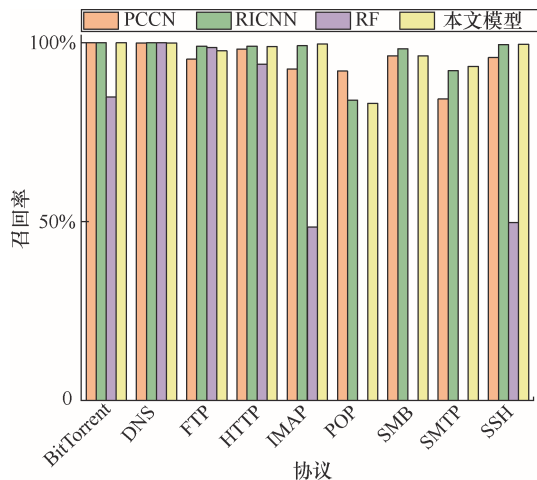


图 15 不同模型在数据集 UNSW-NB15 上的召回率
Figure 15 Recall of different models on dataset UNSW-NB15

因此, 从表 6 和图 11~图 16 可以看出, 与模

型 PCCN、RICNN 和 RF 相比, 本文模型和模型 RICNN 的通用性较强, 在数据集 ISCX2012 和 UNSW-NB15 上的协议识别效果均较好。与此同时, 模型 PCNN 和模型 RF 的通用性较差, 在数据集 ISCX2012 的协议识别效果较好, 在数据集 UNSW-NB15 上的协议识别效果较差。出现这种现象, 主要是由于数据集 UNSW-NB15 内部协议类别分布不平衡的问题比数据集 ISCX2012 更为严重。模型 PCNN 和模型 RF 分别采用特征融合和集成策略消除协议类别分布不平衡影响的能力有限, 模型的通用性较弱。与模型 PCNN 和 RF 相比, 模型 RICNN 在特征融合的基础上, 利用残差连接确保所提取的协议数据特征更具通用性, 模型通用性较强。与其他 3 种模型相比, 本文模型利用特征融合、残差连接、SE-Net 和新激活函数, 保证所提取协议数据特征最具代表性和通用性, 在两个数据集上的协议识别效果均较好。

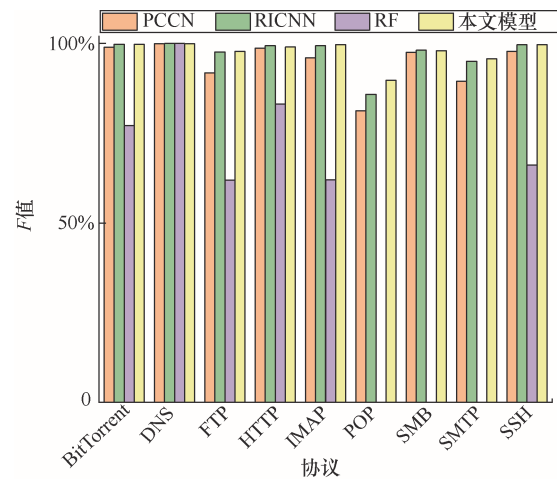


图 16 不同模型在数据集 UNSW-NB15 上的 F 值
Figure 16 F of different models on dataset UNSW-NB15

5 结束语

针对协议类别分布不平衡导致无法准确识别少数类协议的问题, 本文提出基于改进残差 U-Net 的不平衡协议识别方法, 利用词嵌入技术初步挖掘协议数据各字节之间的关联关系, 提升模型提取特征的效率; 利用改进的 SE-Net 强化协议数据特征的不同通道, 提升模型的特征提取能力; 利用新激活函数避免对协议数据特征进行激活时信息丢失的问题; 利用带权重 Dice 损失函数促使协议识别模型重视少数类协议的识别问题。

实验结果表明, 本文模型能够同时提高多数类协议和少数类协议的识别准确率, 优于对比协议识别模型。后续研究将针对协议类别分布严重不平衡并且少数类协议样本极少的情况改进本文协议识别模型, 满足不同应用场景的实际需求。后续将研究如何将所提不平衡协议识别模型应用到加密通信场景。

参考文献:

- [1] FENG W B, HONG Z, WU L F, et al. Review of network protocol recognition techniques[J]. Journal of Computer Applications, 2019, 39(12):3604-3614.
- [2] ZHANG Z X, LIU Q J, WANG Y H. Road extraction by deep residual U-Net[J]. IEEE Geoscience and Remote Sensing Letters, 2018, 15(5): 749-753.
- [3] HU J, SHEN L, SUN G. Squeeze-and-Excitation Networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7132-7141.
- [4] MILLETARI F, NAVAB N, AHMADI S H. V-Net: fully convolutional neural networks for volumetric medical image segmentation[C]//2016 fourth International Conference on 3D Vision (3DV). 2016: 565-571.
- [5] WANG P, LI S H, YE F, et al. PacketCGAN: exploratory study of class imbalance for encrypted traffic classification using CGAN[C]//Proceedings of 2020 IEEE International Conference on Communications (ICC). 2020: 222-224.
- [6] HASIBI R, SHOKRI M, DEHGHAN M. Augmentation scheme for dealing with imbalanced network traffic classification using deep learning[J]. arXiv preprint:1901.00204, 2019.
- [7] TANG H, LIU D, YAO L S, et al. Feature selection algorithm for class imbalanced internet traffic[J]. Journal of Electronics & Information Technology, 2021, 43(4): 923-930.
- [8] GUPTA N, JINDAL V, BEDI P. CSE-IDS: using cost-sensitive deep learning and ensemble algorithms to handle class imbalance in network-based intrusion detection systems[J]. Computers & Security, 2022, (112): 102499.
- [9] TELIKANI A, GANDOMI A H, CHOO K K R, et al. A cost-sensitive deep learning based approach for network traffic classification[J]. IEEE Transactions on Network and Service Management, 2022, 19(1): 661-670.
- [10] ZHANG Y, CHEN X, GUO D, et al. PCCN: parallel cross convolutional neural network for abnormal network traffic flows detection in multi-class imbalanced network traffic Flows[J]. IEEE Access, 2019(7): 119904-119916.
- [11] XIA B H, HAN D Z, YIN X M, et al. RICNN: a ResNet&Inception convolutional neural network for intrusion detection of abnormal traffic[J]. Computer Science and Information Systems, 2022, 19(1): 309-326.
- [12] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]//18th International Conference on Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. 2015: 234-241.
- [13] HOWARD A, SANDLER M, CHEN B, et al. Searching for mobileNetV3[J] //Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1314-1324.
- [14] LIN M, CHEN Q, YAN S C. Network in network[J]. arXiv preprint:1312.4400, 2014.
- [15] SHIRAVI A, SHIRAVI H, TAVALLAEE M. Toward developing a systematic approach to generate benchmark datasets for intrusion detection[J]. Computers & Security, 2012, 31(3):357-374.
- [16] MOUSTAFA N, SLAY J. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)[C]//Proceedings of 2015 Military Communications and Information Systems Conference (MilCIS). 2015:1-6.

[作者简介]



吴吉胜 (1997-), 男, 湖南凤凰人, 陆军工程大学硕士生, 主要研究方向为协议识别。



洪征 (1979-), 男, 江西南昌人, 陆军工程大学副教授, 主要研究方向为漏洞挖掘和协议逆向分析。



马甜甜 (1998-), 女, 江苏邳州人, 陆军工程大学硕士生, 主要研究方向为信息安全。