基于多维度动态加权alpha图像融合与特征增强的恶意软件检测方法

谢丽霞[1],魏晨阳[1],杨宏宇[1,2*],胡 泽[2],成 翔[3,4]

(1. 中国民航大学计算机科学与技术学院,天津 300300;2. 中国民航大学安全科学与工程学院,天津 300300;
3. 扬州大学信息工程学院,江苏扬州 225127;4. 中国民航大学民航飞联网重点实验室,天津 300300)

摘 要: 针对现有恶意软件检测方法缺乏对样本特征的有效提取、过度依赖领域专家知识和运行行为监控,导致严重影响检测分类性能的问题,提出一种基于多维度动态加权alpha图像融合与特征增强的恶意软件检测方法. 通过无效样本清洗与异常值处理获得标准化样本集,利用三通道图像生成与多维度动态加权alpha图像融合方法生成高质量融合图像样本. 采用傀儡优化算法进行数据重构,减少因数据类不平衡对检测结果造成的影响,并对重构数据样本进行图像增强. 通过基于双分支特征提取与融合通道信息表示的空间注意力增强网络,分别提取图像特征和文本特征并进行特征增强,提高特征表达能力. 通过加权融合的方法将增强的图像特征与文本特征进行融合,实现恶意软件家族的检测分类. 实验结果表明,本文所提方法在BIG2015数据集上的恶意软件检测分类准确率为99.72%,与现有检测方法相比提升幅度为0.22~2.50个百分点.

关键词: 恶意软件检测;图像融合;傀儡优化算法;双分支特征提取;数据重构;特征增强

基金项目: 国家自然科学基金(No.62201576,No.U1833107);江苏省基础研究计划自然科学基金(No.BK20230558)

中图分类号: TP309.5 **文献标识码:** A **文章编号:** 0372-2112(2025)03-0849-15

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240746

Malware Detection Method Based on Multi-Dimensional Dynamic Weighted Alpha Image Fusion and Feature Enhancement

XIE Li-xia[1], WEI Chen-yang[1], YANG Hong-yu[1,2*], HU Ze[2], CHENG Xiang[3,4]

(1. *School of Computer Science and Technology, Civil Aviation University of China, Tianjin* 300300, *China*;
2. *School of Safety Science and Engineering, Civil Aviation University of China, Tianjin* 300300, *China*;
3. *School of Information Engineering, Yangzhou University, Yangzhou, Jiangsu* 225127, *China*;
4. *Key Laboratory of Civil Aviation Flight Networking, Civil Aviation University of China, Tianjin* 300300, *China*)

Abstract: Existing malware detection methods suffer from inadequate extraction of sample features, excessive reliance on domain expert knowledge, and operational behavior monitoring, significantly impacting detection and classification performance. To address these issues, we propose a malware detection method based on multidimensional dynamic weighted alpha image fusion and feature enhancement. Standardized sample sets are obtained through invalid sample cleaning and outlier processing. High-quality fused image samples are then generated using a three-channel image generation and multidimensional dynamic weighted alpha image fusion method. The puppet optimization algorithm is employed for data reconstruction to mitigate the impact of data class imbalance on detection results, and image enhancement is performed on the reconstructed data samples. A spatial attention enhancement network based on dual-branch feature extraction and fusion channel information representation is used to extract and enhance image and text features, thereby improving feature representation capabilities. The enhanced image and text features are fused using a weighted fusion method to achieve malware family detection and classification. Experimental results show that the proposed method achieves a malware detection classification accuracy of 99.72% on the BIG2015 dataset, representing an improvement of 0.22~2.50 percentage points over existing detection methods.

收稿日期:2024-08-13;修回日期:2025-01-20;责任编辑:孙瑶

*通讯作者:杨宏宇

Key words: malware detection; image fusion; puppet optimization algorithm; dual-branch feature extraction; data reconstruction; feature enhancement

Foundation Item(s): National Natural Science Foundation of China (No.62201576, No.U1833107); Basic Research Program Natural Science Foundation of Jiangsu Province (No.BK20230558)

1 引言

恶意软件作为一种具有恶意目的的软件程序,通过非法入侵计算机系统和破坏隐私数据的方式,严重威胁着个人信息安全和企业经济安全. 根据AV-Test的年度恶意威胁报告,目前每天产生的恶意软件新变种超过3.6万个,平均每秒有4.2个新变种出现,恶意软件的快速变种能力和高级混淆规避能力使其能够轻易绕过防火墙和反病毒程序,对系统安全和应用安全构成极大威胁. 因此,高效的恶意软件检测方法研究成为网络与系统安全领域的研究热点.

当前恶意软件检测方法主要包括静态分析方法、动态分析方法、基于机器学习和深度学习的检测方法,传统检测方法通过分析恶意软件的二进制源文件,或在受控环境下监控其运行行为特征进行检测分类[1]. 然而,此类方法容易受到混淆遮蔽技术的干扰,需要消耗较高的时间和人力成本[2]. 与传统检测方法相比,图像化检测方法的独特优势在于,能够以图像的形式直观展示出恶意软件中复杂且隐蔽的代码结构特征,从而更准确地识别分类恶意软件[3]. Nataraj等人[4]提出将恶意软件的二进制源文件转换为灰度图像,通过分析图像纹理信息的相似性进行恶意软件家族分类. 实验结果表明,与传统检测方法相比该方法耗时更短且检测分类准确率更高,具有较高的应用前景.

近年来,结合深度学习与图像化技术的检测分类方法逐渐成为研究热点. Cui等人[5]提出基于仿生算法的图像化恶意软件变体检测方法,将二进制源文件转换为灰度图像后,使用蝙蝠算法解决数据集中类不平衡的问题,结合卷积神经网络(Convolutional Neural Network, CNN)提取图像特征并进行检测分类. Dong等人[6]提出一种融合CNN和深度神经网络(Deep Neural Network, DNN)的混合机制恶意软件检测方法,将权限特征与API调用图融合后通过CNN提取图像局部特征,通过DNN构建全局特征关联并提升模型的检测精度和泛化能力. Deng等人[7]通过计算反汇编操作码中字母与数字之间的转移概率构建马尔科夫状态转移矩阵,并将其映射进图像的3个通道生成马尔科夫图,使用轻量级CNN模型进行检测分类. 与Deng的方法类似,Ni等人[8]通过局部敏感哈希SimHash算法,对操作码序列转换进行哈希计算,并将得到SimHash值映射为灰度图像进行检测分类. 然而,上述研究方法均高度依赖于图像处理的精度,如果生成的图像样本质量存在缺陷,则会严重影响最终的实验结果. 此外,上述各研究方法还存在生成图像类型单一,无法全面反映恶意软件的深层行为特征问题.

Vasan等人[9]提出使用多层堆叠神经网络提取恶意软件图像特征并进行检测分类,通过微调预训练的CNN模型使其适应恶意软件检测分类任务. 与Vasan的研究类似,Naeem等人[10]提出一种基于堆叠集成的物联网恶意软件检测分类方法,通过构建多个卷积层数不同的CNN模型并堆叠集成,以此捕获恶意软件不同层次的多样化特征,提升模型分类性能和检测精度. Zou等人[11]使用改进的胶囊网络,通过动态卷积进行多级特征提取和融合,增强模型的表征能力并稳定训练过程. Tang等人[12]提出一种基于字节和十六进制n-gram的恶意软件检测分类方法,通过构建组合字节特征,使用基于决策树算法的分布式梯度提升框架进行恶意软件检测分类. 然而,上述研究方法存在对少类家族样本不敏感、模型内存占用大、依赖特征单一等不足,严重影响恶意软件检测的分类效果.

通过对现有研究分析可知:(1)多数图像化恶意软件检测分类方法只针对灰度图像或者单一类型的图像,导致模型的特征提取能力受限;(2)数据量不足和类不平衡问题普遍存在,部分家族样本数量过少,模型无法有效学习少类家族图像特征,影响最终的检测分类效果;(3)现有方法依赖特征类型较为单一,普遍仅考虑图像或文本特征,在特征信息获取和泛化能力方面存在局限.

针对以上不足,本文提出一种基于多通道图像生成与动态加权图像融合的方法,获得融合图像样本进行恶意软件检测分类. 首先将恶意软件二进制源文件转换为不同类型的图像并进行融合,利用傀儡优化算法对融合图像样本集进行数据重构,通过双分支特征提取网络图像和文本特征,结合特征增强与多模态特征加权融合,实现对恶意软件的有效检测分类. 本文主要贡献具体如下.

(1)提出一种三通道图像生成和多维度动态加权alpha图像融合方法. 将二进制源文件可视化为RGB图、熵图、类马尔科夫图3种不同类型的图像,通过多维度动态加权alpha图像融合方法将3种类型的图像填充至通道中进行融合,构建具有更丰富特征表达能力的融合图像.

(2)提出一种名为傀儡优化算法(Puppet Optimiza-

tion Algorithm,POA)的数据重构方法. 通过计算样本的傀儡分数(Puppet Score,PS),利用已有数据生成与原始样本具有相似特征的虚拟样本,解决数据不平衡导致的模型训练偏差等问题,同时显著减少由于数据不平衡引起的误分类现象.

(3)提出一种基于双分支特征提取与融合通道信息表示的空间注意力增强网络(Dual-Branch Feature extraction and fusion Channel Information Spatial Attention Enhanced Network,DB-FCISAEN)实现多模态特征提取与恶意软件家族检测分类. 其中,通过融合通道信息表示的空间注意力机制(Fusion of Channel Information for Spatial Attention,FCISA)实现特征增强并优化特征表达能力,提高恶意软件家族检测分类准确率.

2　检测方法框架

当前多数恶意软件检测研究集中于分析样本的二进制文本特征或动态行为特征,但此类传统检测方法容易受到恶意软件混淆技术的影响,导致高误报率的出现. 随着恶意软件反检测技术与混淆技术的不断发展,仅考虑单一类型的特征难以实现有效检测. 因此,为提高检测性能并降低误报率,本文提出一种基于多维度动态加权alpha图像融合与特征增强的恶意软件检测方法. 该方法框架如图1所示.

该方法主要包括样本预处理、融合图像生成、数据重构与图像增强和特征提取与检测分类4个部分,各部分的主要功能如下.

(1)样本预处理. 清洗缺失或乱码的无效样本,解析文本内容并移除无效地址信息和异常填充数据,对汇编指令文件内容进行标准化,得到标准化样本集.

(2)融合图像生成. 对经过预处理的标准化数据样本,分别使用自适应去冗余图像生成方法、融合空间填充曲线图像生成方法和混合操作码转移概率图像生成方法进行可视化. 通过多维度动态加权alpha图像融合方法,对3种不同类型的图像进行动态融合,生成保留更多原始样本信息的融合图像样本.

(3)数据重构与图像增强. 通过计算样本的傀儡分数,利用现有样本对少类家族生成扩充样本,对多类家族进行样本重构,生成重构图像样本数据集. 之后使用图像锐度增强、噪声区域清除和亮度调整的方法,增强重构后图像样本的整体质量.

(4)特征提取与检测分类. 将重构图像样本和汇编指令样本共同组成重构样本集,输入DB-FCISAEN. 通过高效神经网络EfficientNet提取重构图像样本的多层次图像特征,通过对比自然语言-汇编语言预训练模型(Contrastive Language-Assembly Pre-training, CLAP)提取汇编指令样本的文本特征,通过FCISA进行特征增强,通过加权融合的方法将图像特征与文本特征进行融合,完成对恶意软件家族的检测分类.

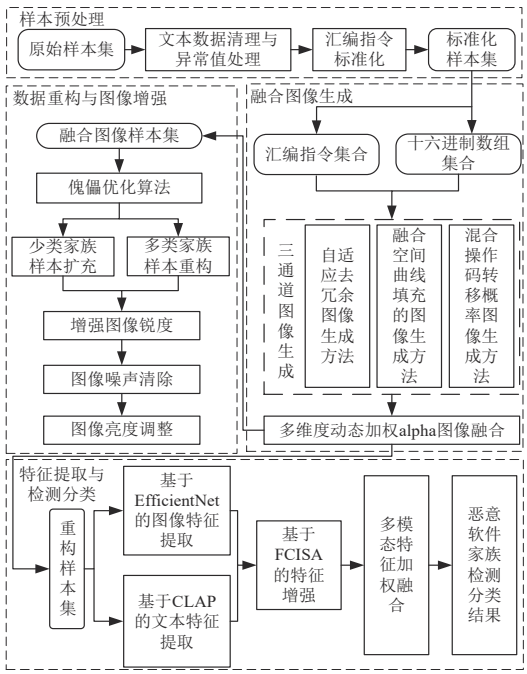

图1　恶意软件检测方法框架

3　样本预处理与融合图像生成

3.1　样本预处理

在进行可视化操作前,需要进行样本预处理,包括清洗缺失无效样本、清除样本文件冗余内容、函数段和汇编指令序列标准化等过程,具体预处理流程如图2所示.

(1)对".bytes"文件的预处理. 清除所有缺失、无效样本,删除每行行首代表地址信息的8位连续十六进制数,仅保留行内的数组信息. 对文件中存在的非指令重复无效字节进行清除,标准化字节排列格式并进行重排序. 对于存在严重内容信息缺失的样本进行删除.

(2)对".asm"文件的预处理. 清除所有缺失、无效样本,删除原始文件中所有的非".text"段内容. 对于".text"段,删除所有不包含操作数、操作码的无效行,同时删除各行中所有的符号信息进行重排列. 对重排列后的各行内容进行重筛选,仅保留每行的操作数、操作码和指令信息,并按照函数段对重构后的".asm"文件进行划分. 为划分后的子文件每行补充行号和标识信息,确保每一行都严格按照"行号:操作码 操作数,操作数"的格式进行存储,以便后续进行文本特征的提取.

3.2　融合图像生成

传统图像化方法依赖单一的图像生成技术,例如将二进制源文件转换为灰度图或马尔科夫图. 虽然这些方法能够在一定程度上揭示恶意软件的内在行为特

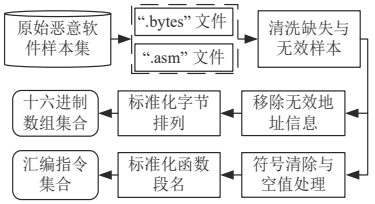

图2 样本预处理流程

征,但无法全面捕捉其包含的多维特征信息. 此外,传统方法在处理数据噪声和冗余信息时容易混淆关键特征,导致模型出现较高的误报率和漏报率,严重影响检测分类精度. 为解决上述问题,本文提出使用融合图像样本代替单一类型图像进行实验. 融合图像生成包括三通道图像生成与多维度动态加权alpha图像融合2个部分. 首先将恶意软件的二进制源文件可视化为RGB图、熵图和类马尔科夫图3种不同类型的图像. 之后通过多维度动态加权alpha图像融合方法,生成保留更多原始样本信息的多维度融合图像样本.

3.2.1 三通道图像生成

在完成样本预处理得到标准化样本集后,提取".bytes"文件的字符序列信息、".asm"文件的操作数、操作码指令序列信息和2类文件的熵值分布进行可视化. 三通道图像生成过程如图3所示. 具体方法设计如下.

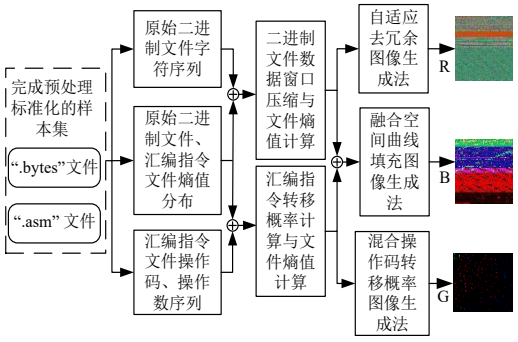

图3 三通道图像生成过程

(1)R通道图像生成

将".bytes"文件划分为固定大小的字节窗口,每个窗口设置包含N=256 bytes. 之后计算各个数据窗口的信息熵$H(W_i)$度量各数据窗口中的信息量,窗口信息熵计算方法为

$$H(W_i)=-\sum_{j=1}^{256}P_j\log_2 P_j \tag{1}$$

其中,W_i为第i个数据窗口;P_j为数据窗口中第j个字节值出现的频率. 在获得各数据窗口的信息熵数据后,计算相邻窗口间的局部自相关性系数(Autocorrelation Coefficient for Neighboring Windows,ACNW)和数据窗口的冗余压缩率(Compression Ratio,CR),ACNW(W_i,W_{i+1})和CR(W_i)的具体计算方法为

$$\begin{aligned}\mathrm{ACNW}(W_i,W_{i+1})=&\frac{\sum_{k=1}^{N}\left(W_i[k]-\mu_{w_i}\right)}{\sqrt{\sum_{k=1}^{N}\left(W_i[k]-\mu_{w_i}\right)^2}}\\&\times\frac{\left(W_{i+1}[k]-\mu_{w_{i+1}}\right)}{\sqrt{\sum_{k=1}^{N}\left(W_{i+1}[k]-\mu_{w_{i+1}}\right)^2}}\end{aligned} \tag{2}$$

$$\mathrm{CR}(W_i)=1-\frac{S_{\mathrm{com}}(W_i)}{S_{\mathrm{ori}}(W_i)},S_{\mathrm{ori}}(W_i)=N \tag{3}$$

其中,k为数据窗口中第k个字节的位置;μW_i和μW_{i+1}分别为数据窗口W_i和W_{i+1}的平均值;$S_{\mathrm{com}}(W_i)$为压缩后的数据窗口大小. 通过对ACNW、CR和$H(W_i)$进行加权平均,计算最终的冗余指标(Compression Potential,CP). 将得到的CP(W_i)与预设阈值进行对比,对超过阈值的数据窗口进行压缩操作. 将处理后数据窗口每行的十六进制字符转换为[0,255]区间范围内的二维数组,根据数值将其映射为图像. 原始字节值映射为R通道,在原始字节值的基础上加128后对256取模映射为G通道,在原始字节值的基础上加64后对256取模映射为B通道,生成RGB图像增强可视化效果.

(2)G通道图像生成

通过设定马尔科夫过程中每个状态所代表特定的字母或数字组合,计算操作码中字母与数字之间的转移概率,捕捉".asm"文件指令序列的统计特征信息,从而反映恶意软件样本的结构和模式. 具体的G通道图像生成过程如下.

首先,计算基于上下文敏感的操作码转移概率. 分析相邻操作码之间转移概率的同时引入更长的上下文,计算在特定的2个或3个连续操作码序列之后出现另一个操作码的概率,深入识别和分析操作码序列中的指令模式. 其次,计算连续操作码中首字母或数字的唯一组合转移概率,展现各个指令之间的关联信息. 最后,计算基于功能函数块的转移概率,分析函数块内部和跨函数块的操作码转移概率,捕获恶意软件的行为模式,比较不同程序之间的相似性. 将上述转移概率表示为灰度值,并映射至[0,255]区间范围内转换为图像表示,最后通过三通道填充方法进行图像融合. 与生成传统的马尔科夫图相比,该方法能捕捉操作码之间更深层的转移关系,综合提取多维度特征信息.

(3)B通道图像生成

对于融合图像的B通道,通过计算".asm"文件与".bytes"文件的信息熵值生成熵图,利用数据的随机性辅助检测加密压缩内容. 在完成对".asm"文件的预处理后,分别将".bytes"和".asm"文件重组为不包含特

殊符号信息的字符串,统计各个数据窗口中字节、字符的出现频率,并计算数据熵值. 数据熵值计算方法为

$$H(X)=-\sum_{i=1}^{n}p(x_i)\log_2 p(x_i) \tag{4}$$

其中,$p(x_i)$为第i个字节或字符出现的概率;n为总字节或字符数. 之后,使用最小-最大归一化方法,将原始数据熵值缩放至[0,1]区间范围内,之后将归一化结果填充至二维矩阵中并将其可视化,分别填充融合图像的R、G通道. B通道利用希尔伯特曲线填充技术,通过将2种文件熵值进行加权融合的操作,将熵值序列映射到二维空间,同时保持数据的空间局部性和临近关系,从而在二维图像中更好地展示数据整体特征.

3.2.2　多维度动态加权alpha图像融合

传统的图像融合方法在处理多种类型的图像信息时,存在信息丢失、融合效果不佳和无法充分利用多维度图像特征等问题. 此外,传统的通道加权平均融合方法无法自适应调整融合权重,导致融合图像的细节表现和信息表达不足. 尤其在融合亮度和纹理差异较大的图像时,容易出现融合效果失调等问题,严重影响恶意软件家族检测分类准确率.

为克服上述缺陷,本文提出多维度动态加权alpha图像融合方法(如算法1所示),通过设计对比实验发现,基于亮度、纹理和信息熵3个维度的特征提取,能够从多个角度全面描述恶意软件图像的关键信息. 其中,亮度特征反映图像的全局光照变化,纹理特征捕捉局部结构模式,信息熵则衡量图像的复杂性与不确定性,三者的协同作用能够充分表征恶意软件图像中的多维信息,从而在融合图像质量和恶意软件家族分类性能上表现出显著优势. 此外,引入其他维度特征(颜色特征和频域特征)后分类性能的提升并不显著,表明亮度、纹理和信息熵三者已能够较为全面地刻画恶意软件图像的特性. 因此,本方法选择提取上述3个维度的特征,计算特征差异值并自适应调整融合权重,动态增强或减弱各通道图像融合强度,确保融合后的图像能够更好保留和整合不同类型的关键信息.

该方法的具体过程如下. 首先调整熵图和类马尔科夫图的尺寸以匹配RGB图尺寸. 通过预设通道权重的亮度计算方法和图像局部信息熵计算方法,得到每种类型图像的亮度值和信息熵值;通过局部二值模式(Local Binary Pattern,LBP)方法,比较图像中每个像素与其周围像素的灰度值,生成图像纹理描述信息值. 然后基于亮度、纹理和信息熵计算每个像素的差异性权重,并将差异值进行归一化处理确保权重的总和为1. 其次使用归一化权重动态调整每个像素的alpha值,并对类马尔科夫图和熵图进行初步融合. 最后将初步融合图像与RGB图像进行再次融合,得到最终的融合图像. 通过这种方法,可以有效克服传统方法存在的图像信息丢失和融合效果不佳等问题,最大限度地保留关键信息,实现更高质量的图像融合.

算法1　多维度动态加权alpha图像融合算法

输入:R通道图像RGB_imgs, G通道图像S_markov_imgs, B通道图像Entropy_imgs, 基础alpha值α_base, 调整因子adjust_factor
输出:融合图像样本集Fused_imgs

```
1. function ALPHA
2. 初始化Fused_imgs为空列表
3. for each(img_RGB,img_S_markov,img_Entropy)执行
4. 调整img_S_markov和img_Entropy的大小与img_RGB一致
5. brightness_xxx= calculate_brightness(img_xxx)
   calculate_brightness(img_xxx)=0.299R+0.587G+0.114B
   //提取各通道图像亮度特征
6. texture_xxx = calculate_texture(img_xxx)
   //提取各通道图像纹理特征
7. entropy_xxx = calculate_entropy(img_xxx)//提取各通道图像熵值特征
8. brightness_diff_xxx=calculate_difference(brightness_xxx, brightness_xxx) weight_brightness_xxx=normalize(brightness_diff_xxx)
   //计算各通道图像亮度差异并归一化
9. texture_diff_xxx=calculate_difference(texture_xxx,texture_xxx)
   weight_texture_xxx= normalize(texture_diff_xxx)
   //计算各通道图像纹理差异并归一化
10. entropy_diff_xxx = calculate_difference(entropy_xxx, entropy_xxx)
    weight_entropy_xxx = normalize(entropy_diff_xxx)
    //计算各通道图像熵值差异并归一化
11. 动态调整alpha值
    α_dynamic_xxx=α_base+adjust_factor*(weight_brightness_xxx+
    weight_texture_xxx+ weight_entropy_xxx)
12. 融合图像fused_image=blend_images(img_rgb,img_S_markov,
    img_Entropy,alpha_dynamic_S_markov,alpha_dynamic_Entropy)
    Final_image= further_blend(fused_image,img_rgb)
13. Fused_imgs.append(Final_Image)
    //将生成的融合图像加入输出图像样本集
14.     end for
15.   return Fused_imgs
16. end function
```

4　数据重构与图像增强

4.1　数据重构

在恶意软件检测分类所用数据集中,数据不平衡是一类常见问题,严重影响模型的分类性能和泛化能力,并且会导致高误报率和漏报率的出现. 为解决上述问题,本文提出一种名为傀儡优化算法(具体如算法2所示)的数据重构方法,生成虚拟样本以平衡数据集中少数类别家族的样本分布.

算法2 傀儡优化算法POA

输入：训练数据集train_data,标签labels,预训练神经网络模型model,傀儡权重参数α

输出：平衡后的数据集balanced_data

1. 初始化:unique_labels//获取labels中的唯一标签
2. balanced_data//空列表
3. for each label∈unique_labels执行
4. family_data//获取train_data中标签为label的样本
5. family_size// amily_data的样本数量
6. if family_size∈少类家族
7. features//使用model提取family_data的特征
8. reduced_features//使用t-SNE将features降维至二维
9. augmented_data//通过显著区域提取和样本扩充生成虚拟样本
10. 将augmented_data添加到balanced_data
11. else://对多类家族执行
12. family_representativeness//空列表
13. puppet_scores//空列表
14. 对每个x_i∈family_data执行
15. rep、contrib//所有样本x_i与family_data中其他样本的余弦相似度均值、SHAP均值
16. puppet_score//α∗rep+(1-α)∗contrib
17. puppet_scores // puppet_scores + puppet_score
18. sorted_indices//按puppet_scores从高到低排序family_data的索引
19. 执行下采样有特征扰动
20. 将处理后的数据添加到 balanced_data
21. end for
22. return balanced_data

通过计算样本的“傀儡分数”,结合显著性引导的样本数据增强对少类家族样本进行扩充,对多类家族进行样本重构,以提高模型的整体性能,增强模型在实际应用中的鲁棒性和可靠性. 傀儡分数由样本的家族代表性和样本对分类模型的贡献度构成. 傀儡分数具体计算方法为

$$\mathrm{PS}=-\frac{1}{2\left|C_k\right|}\sum_{j\in C_k}\left(1-\frac{\sum_{k=1}^{n}x_{ik}x_{jk}}{\sqrt{\sum_{k=1}^{n}x^2{}_{ik}}\cdot\sqrt{\sum_{k=1}^{n}x^2{}_{jk}}}\right)+0.5\mathrm{SHAP}(x_i) \tag{5}$$

其中,恶意软件样本x_i和x_j所属家族集合为C_k,k为选定样本所在家族,$\mathrm{SHAP}(x_i)$为样本x_i计算得到的SHAP值.

多类家族的融合图像样本具有样本间特征相似度较高的特点,导致模型在训练过程中容易对特定家族特征产生过拟合,难以将有效预测结果推广到其他类家族甚至未知家族. 因此,本文通过计算多类家族样本的傀儡分数对多类家族的恶意软件样本进行重排序,筛选出代表性较低但贡献度较高的高傀儡分数样本,保留傀儡分数高的样本留作实验样本,对傀儡分数最低的后1%的样本(不足则向上取整)进行清除操作,以减少模型对冗余样本的依赖并提升模型的泛化能力. 之后结合特征扰动对样本集进行聚类和筛选,避免冗余样本对训练集的负面影响,同时针对部分特征相似的样本,通过对图像进行旋转、缩放等扰动操作,进一步增强数据集的多样性. 对于少类家族的融合图像样本,使用预训练的卷积神经网络提取图像特征,并利用t-SNE降维技术将提取的特征降至二维,可视化对检测结果具有显著影响的区域. 在这些区域内,保留图像的原始像素数据不变,对其他区域进行亮度调整、角度翻转、镜像对称等操作生成傀儡样本. 通过这种方法,能够有效扩充少类家族样本的数量,实现数据类的平衡.

4.2 图像增强

通过傀儡优化算法得到重构图像样本集后,直接使用未经处理的图像会存在噪声干扰、细节信息缺失、像素模糊和亮度不均衡等问题. 导致图像中的重要特征无法被准确捕捉,严重影响最终检测分类的准确性. 为解决上述问题,本文使用多种图像增强技术,包括图像噪声区域清除、锐度增强以及亮度调整. 对于噪声区域的检测与清除,设置固定尺寸的滑动窗口大小和步长,通过计算局部区域的像素标准差检测并划分噪声,具体计算方法为

$$\sigma=\sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i-\mu)^2} \tag{6}$$

其中,σ为滑动窗口内局部区域像素的标准差,μ为整个图像区域的像素平均值,N为像素总数,x_i为第i个像素值. 当σ超过阈值测到噪声区域后,使用中值滤波器计算邻域像素替换噪声像素,使用拉普拉斯锐化和直方图均衡化方法增强图像边缘细节与对比度. 该方法能够有效提高虚拟样本图像质量,确保其在数据集中的一致性和可靠性,提高图像的整体质量和可用性. 此外,通过图像增强可以充分结合多维度动态加权alpha图像融合方法与傀儡优化算法的优势,进一步提升融合图像与虚拟样本的质量.

5 特征提取与检测分类

5.1 DB-FCISAEN网络模型

为有效提取并融合恶意软件样本的多层次特征,提高检测分类准确率. 本文提出DB-FCISAEN网络模型,利用多模态融合特征实现对恶意软件家族的高效检测分类. DB-FCISAEN主要由双分支特征提取、基于FCISA的特征增强与多模态特征融合和恶意软件家族检测分类3个部分构成. DB-FCISAEN网络模型具体结构如图4所示,其主要处理流程如下.

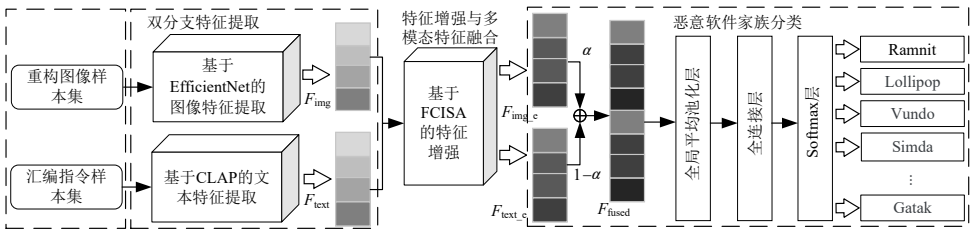

图4　DB-FCISAEN具体网络结构

(1)通过双分支特征提取网络,处理重构图像样本集和经过标准化的汇编指令样本集. 对于重构图像样本,通过EfficientNet提取多层次融合图像特征;对于标准化汇编指令样本,通过CLAP提取文本特征.

(2)通过基于融合通道信息表示的空间注意力机制FCISA,对提取到的图像特征和文本特征进行增强,并通过加权平均的方法对增强后的图像特征与文本特征进行融合.

(3)将融合特征输入恶意软件家族检测分类网络,通过全局平均池化层、全连接层和softmax层处理融合特征并输出分类概率,完成恶意软件家族的检测分类.

5.2　双分支特征提取

5.2.1　文本特征提取

为提取汇编指令文件的深层次语义信息,DB-FCISAEN使用预训练的CLAP模型[13]捕获恶意软件文本特征. CLAP通过对比学习,将汇编代码与自然语言描述进行对齐,增强CLAP-ASM编码器对汇编代码深层语义信息的捕捉能力. 文本特征提取的过程可以具体分为以下几个步骤:汇编指令文件的标准化与标记化、地址重定位与指令嵌入、位置嵌入与上下文建模、特征表示的生成.

首先进行汇编指令文件的标准化与标记化,得到符合输入的指令符号序列t_i. 其次,进行地址重定位与共享跳转符号的标记指令嵌入,嵌入层将每个指令符号t_i转换为一个d维向量,嵌入后的指令序列可以表示为$\boldsymbol{E}(T_{\text{asm}})$,$e_i$为指令嵌入向量. 最后进行位置嵌入与上下文建模,通过正弦-余弦位置编码生成位置嵌入向量$\boldsymbol{p}_i$,固定指令在整个汇编文件中的相对位置,并保留指令的顺序信息. 将指令嵌入向量与位置嵌入向量拼接得到结合位置信息的嵌入表示$\boldsymbol{h}_i$. 通过基于自注意力机制的Transformer模型,对汇编指令序列的上下文进行建模以捕捉汇编指令之间的依赖关系,并完成特征表示的生成. 具体过程如下:

$$\boldsymbol{h}_i=\boldsymbol{e}_i+\boldsymbol{p}_i,\boldsymbol{h}_i\in\mathbb{R}^d \tag{7}$$

$$\boldsymbol{H}(T_{\text{asm}})=\text{Transformer}\left(\boldsymbol{H}(T_{\text{asm}})\right)=[\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_N] \tag{8}$$

$$\boldsymbol{f}_{\text{text}}=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{h}_i \tag{9}$$

其中,$\boldsymbol{H}(T_{\text{asm}})$为经过上下文建模后的特征矩阵,$\boldsymbol{f}_{\text{text}}$为包含汇编指令的全局语义信息的文本特征.

5.2.2　图像特征提取

为提取重构图像样本的多层次图像特征,DB-FCISAEN使用EfficientNet模型提取并融合图像浅层特征、较深层特征和深层特征. EfficientNet采用基于复合缩放策略的网络结构设计,同时优化网络的深度、宽度和分辨率,从而在保证计算效率的同时更好地提取图像的特征. 浅层特征为第1层卷积层输出的特征,卷积核大小为3×3,步幅大小为2. 通过细粒度的特征表示保留局部的几何结构,其中包含基本的几何信息和局部对比度信息,重点关注图像的边缘、纹理等低级特征. 浅层特征$\boldsymbol{f}_1$的输出表示为

$$\boldsymbol{f}_1(i,j,c')=\text{ReLU}\left(\sum_{m=1}^{3}\sum_{n=1}^{3}\sum_{c=1}^{C}\omega_{mnc}X(i+m-1,j+n-1,c)+\boldsymbol{b}_{c'}\right) \tag{10}$$

其中,$\boldsymbol{f}_1(i,j,c')$为浅层特征图在输出位置(i,j)及通道c'上的激活值;ω_{mnc}为卷积核中的权重;$X_{(i,j,c)}$为输入图像在位置(i,j)处的像素值;$\boldsymbol{b}_{c'}$为偏置项,m和n分别为卷积核在高度和宽度方向上的索引,c为输入图像的通道索引.

较深层特征由EfficientNet的第3层和第4层MBConv层输出,利用深度可分离卷积和倒置残差结构,能够有效捕捉图像的复杂模式和局部特征. 深度可分离卷积是一种高效的卷积形式,通过深度卷积和逐点卷积,显著减少模型参数量的同时保证特征提取能力. 倒置残差结构确保信息在网络中逐层传播时,不会因卷积运算而丢失特征信息,增强捕获特征的鲁棒性. 第3层和第4层MBConv层分别使用3×3和5×5的卷积核,以捕捉不同尺度下的局部特征. 较深层特征$\boldsymbol{f}_2$的输出表示为

$$\boldsymbol{f}_2^{(3)}(i,j,c')=\text{ReLU}\left(\sum_{m=1}^{3}\sum_{n=1}^{3}\sum_{c=1}^{C_1}\omega^{(3)}\boldsymbol{f}_1(i+m-1,j+n-1,c)+\boldsymbol{b}_{c'}^{(3)}\right) \tag{11}$$

$$\boldsymbol{f}_2^{(4)}(i,j,c')=\text{ReLU}\left(\sum_{m=1}^{5}\sum_{n=1}^{5}\sum_{c=1}^{C_2}\omega^{(4)}\boldsymbol{f}_2^{(3)}(i+m-1,j+n-1,c)+\boldsymbol{b}_{c'}^{(4)}\right) \tag{12}$$

其中,$\boldsymbol{f}_2^{(3)}(i,j,c')$和$\boldsymbol{f}_2^{(4)}(i,j,c')$分别为第3层和第4层的输

出特征图；C_i为在不同卷积层中的特征图的通道数；$\omega^{(3)}$和$\omega^{(4)}$分别为模型第3层和第4层的卷积核权重.

深层特征为第8层MBConv层和最后一层卷积层输出的特征，包含图像深层的抽象信息和全局特征. 第8层MBConv层继续使用3×3的深度可分离卷积核，强化对全局模式的捕捉. 最后一层卷积层通过1×1卷积核进行压缩，生成图像的全局特征向量. 深层特征$\boldsymbol{f}_3$的输出表示为

$$\boldsymbol{f}_3^{(8)}(i,j,c')=\mathrm{ReLU}\left(\sum_{m=1}^{3}\sum_{n=1}^{3}\sum_{c=1}^{C_3}\omega^{(8)}\boldsymbol{f}_2^{(4)}\left(i+m-1,j+n-1,c\right)+\boldsymbol{b}_{c'}^{(8)}\right) \tag{13}$$

$$\boldsymbol{f}_3^{(f)}(i,j,c')=\mathrm{ReLU}\left(\sum_{m=1}^{1}\sum_{n=1}^{1}\sum_{c=1}^{C_4}\omega^{(f)}\boldsymbol{f}_3^{(8)}\left(i+m-1,j+n-1,c\right)+\boldsymbol{b}_{c'}^{(f)}\right) \tag{14}$$

在提取浅层、较深层和深层特征后，首先对各层输出的特征图进行插值操作以确保维度统一，之后DB-FCISAEN通过特征拼接与融合操作将不同层次的特征进行综合表达，增强浅层和深层特征的表达能力，获得更加全面的图像特征表示. 融合多层次特征的具体过程为

$$\boldsymbol{f}_i'=\mathrm{Interpolate}\left(\boldsymbol{f}_i,(H_{f_{\mathrm{img}}},W_{f_{\mathrm{img}}})\right) \tag{15}$$

$$\boldsymbol{f}_{\mathrm{img}}=\mathrm{Concat}\left(\boldsymbol{f}_1',\boldsymbol{f}_2^{(3)'},\boldsymbol{f}_2^{(4)'},\boldsymbol{f}_3^{(8)'},\boldsymbol{f}_3^{(\mathrm{final})'}\right) \tag{16}$$

其中，$\boldsymbol{f}_{\mathrm{img}}$为最终输出的融合多层次图像特征；$(H_{\mathrm{fimg}},W_{\mathrm{fimg}})$为输出特征的维度. 通过将浅层、较深层和深层特征进行有效融合，DB-FCISAEN可以更全面地捕捉图像的多尺度信息和复杂模式. 该方法不仅能增强图像的特征表达能力，还能有效地避免过度依赖单一特征带来的信息损失. 融合特征表达能够更加全面、准确地反映图像的整体结构模式，从而提高后续恶意软件检测分类任务的准确性和鲁棒性.

5.3 特征增强与多模态特征融合

恶意软件反检测和混淆技术的增强，导致获取到的特征存在多样性不足、特征间关联性弱等问题，严重影响模型的检测分类精度. 现有研究方法中，激励与压缩注意力模块(Squeeze and Excitation Attention Module，SEAM)[14]主要关注通道之间的关系，忽略空间上的上下文依赖关系，无法充分捕捉到空间上的相关特征；卷积注意力模块(Convolutional Block Attention Module，CBAM)[15]增加对空间位置信息的关注，但其只考虑特征的局部范围的信息，弱化了通道信息的重要性.

因此，为有效增强原始图像和文本特征，本文提出基于FCISA的特征增强方法，更全面地反映恶意软件本质特征和行为模式. FCISA实现了通道和空间注意力的并行处理，确保模型既能够捕捉到通道之间的全局依赖，又能够有效整合空间位置上的上下文信息，在相互独立学习权重的同时更好地融合二者优势，兼顾图像和文本特征的全局空间布局. 此外与SEAM相比，FCISA强调全局空间依赖和通道依赖的共同学习，通过在通道和空间2个维度上建立全局依赖，确保模型不仅关注恶意软件行为的局部细节，还能捕捉到不同模块或行为模式之间的广泛联系，提升整体特征表达的鲁棒性. 最后，FCISA通过并行的注意力机制，能够更精确地筛选出具有判别力的特征. 与SEAM和CBAM相比，FCISA能更有效地聚焦于与恶意软件行为模式直接相关的特征，通过通道和空间的联合建模，提升模型对恶意软件特征的捕捉能力. 基于FCISA的特征增强网络结构如图5所示，具体处理流程如下.

(1)对原始图像和文本特征进行卷积操作，获得输入特征图$\boldsymbol{T}\in\mathbb{R}^{H\times W\times C}$，分别在通道层面和空间层面对特征图$\boldsymbol{T}$进行全局最大池化、全局平均池化的操作，提取不同种类的特征信息. 具体设计过程如下：

$$\boldsymbol{g}_{\mathrm{avg}}^{\mathrm{channel}}(c)=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{T}(i,j,c) \tag{17}$$

$$\boldsymbol{g}_{\mathrm{avg}}^{\mathrm{spatial}}(i,j)=\frac{1}{C}\sum_{c=1}^{C}\boldsymbol{T}(i,j,c) \tag{18}$$

$$\boldsymbol{g}_{\mathrm{max}}^{\mathrm{channel}}(c)=\max_{i=1,2,\cdots,H;\ j=1,2,\cdots,W}\boldsymbol{T}(i,j,c) \tag{19}$$

$$\boldsymbol{g}_{\mathrm{max}}^{\mathrm{spatial}}(i,j)=\max_{c=1,2,\cdots,C}\boldsymbol{T}(i,j,c) \tag{20}$$

其中，$\boldsymbol{g}_{\mathrm{max}}^{\mathrm{channel}}(c)$和$\boldsymbol{g}_{\mathrm{avg}}^{\mathrm{channel}}(c)$分别为通道层面全局最大池化和全局平均池化的输出；$\boldsymbol{g}_{\mathrm{max}}^{\mathrm{spatial}}(i,j)$和$\boldsymbol{g}_{\mathrm{avg}}^{\mathrm{spatial}}(i,j)$分别为空间层面全局最大池化和全局平均池化的输出. 之后，通过共享参数的多层感知机层减少参数数量，同时降低模型复杂度，得到形状为$1\times1\times C$的通道描述符和$H\times W\times1$的空间描述符. 将通道描述符与空间描述符Reshape并进行拼接，得到形状为$(H\times W+C)\times1$的特征向量，通过多层感知机建模全局空间特征与全局通道特征之间的相关性，之后重新将其分割为$1\times1\times C$与$H\times W\times1$的特征向量，恢复为原有形状. 具体过程为

$$\boldsymbol{Z}_{\mathrm{channel}}=\mathrm{MLP}\left(\boldsymbol{g}_{\mathrm{avg}}^{\mathrm{channel}},\boldsymbol{g}_{\mathrm{max}}^{\mathrm{channel}}\right)\in\mathbb{R}^{1\times1\times C} \tag{21}$$

$$\boldsymbol{Z}_{\mathrm{spatial}}=\mathrm{MLP}\left(\boldsymbol{g}_{\mathrm{avg}}^{\mathrm{spatial}},\boldsymbol{g}_{\mathrm{max}}^{\mathrm{spatial}}\right)\in\mathbb{R}^{H\times W\times1} \tag{22}$$

$$\boldsymbol{Z}_{\mathrm{concat}}=\mathrm{Concat}\left(\mathrm{Reshape}(\boldsymbol{Z}_{\mathrm{channel}}),\mathrm{Reshape}(\boldsymbol{Z}_{\mathrm{spatial}})\right)\in\mathbb{R}^{(H\times W+C)\times1} \tag{23}$$

最后通过sigmoid激活函数计算通道与空间权重，并与输入特征图$\boldsymbol{T}$进行点积，得到输出特征图，通过加权融合的方法将2类特征图进行融合，最终得到增强后的特征$\boldsymbol{F}\in\mathbb{R}^{H\times W\times C}$.

$$\boldsymbol{F}=\boldsymbol{T}\odot\sigma\left(\mathrm{MLP}(\boldsymbol{Z}_{\mathrm{concat}}^{\mathrm{channel}})\right)+\boldsymbol{T}\odot\sigma\left(\mathrm{MLP}(\boldsymbol{Z}_{\mathrm{concat}}^{\mathrm{spatial}})\right)\in\mathbb{R}^{H\times W\times C} \tag{24}$$

其中，$\mathrm{MLP}(\boldsymbol{Z}_{\mathrm{concat}}^{\mathrm{spatial}})$为重新分割至通道和空间部分的$1\times1\times C$与$H\times W\times1$特征向量，$\sigma$为sigmoid激活函数，$\boldsymbol{T}$为输入特征图. 最后，为充分利用增强后的图像和文

本特征之间的关联信息,提升特征融合的效果,对增强后的图像特征 $\boldsymbol{F}_{\text{img\_e}}$ 和文本特征 $\boldsymbol{F}_{\text{text\_e}}$ 分别赋予超参数权重 a 和 $(1-a)$ 并进行动态调整,得到最终的融合特征 $\boldsymbol{F}_{\text{fused}}$. 通过实验分析,最终选择超参数权重 a 为0.4.

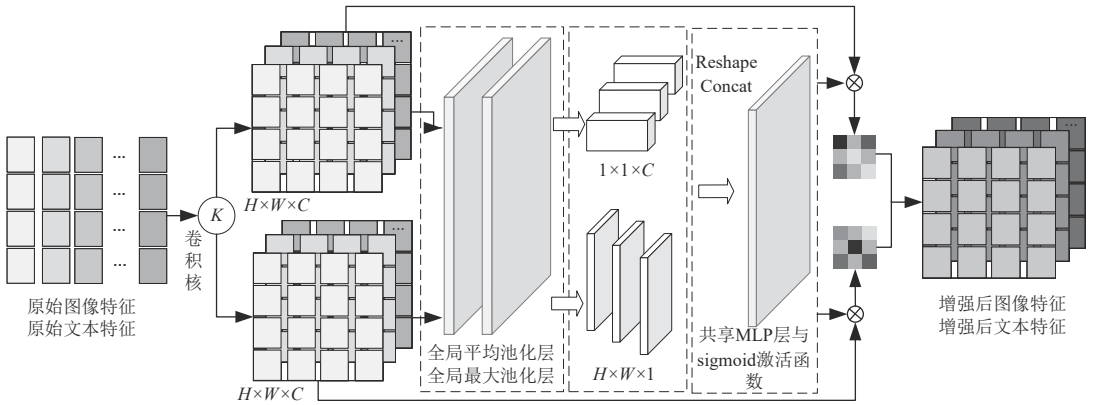

图5　基于FCISA的特征增强网络结构

5.4　模型训练与恶意软件家族检测分类

在获取到多模态融合特征后,通过恶意软件家族检测分类网络对其进行处理,得到最终的检测分类结果. 训练过程中使用的EfficientNetB0和CLAP模型参数保持固定不变,首先将尺寸为224×224的融合图像样本和经过预处理的汇编指令文本送入双分支特征提取器,分别提取图像特征和文本特征. 之后使用FCISA增强多模态特征并进行融合,融合权重超参数 a 设置为0.4. 最后,使用Focal loss损失函数计算预测与实际输出的误差,并使用反向传播更新模型参数权重. 训练次数设置为200轮,批量大小设置为128,初始学习率设置为0.001. 恶意软件家族检测分类网络由全局平均池化层(Global Average Pooling, GAP)、全连接层(Fully Connected Layer, FC)和softmax激活函数组成. 首先通过GAP层将 $H\times W\times C$ 的高维特征转换为低维特征向量,通过对特征图中每个通道进行平均池化的操作,将每个 $H\times W$ 的特征图转换为单个标量,具体计算过程如下:

$$\boldsymbol{F}_{\text{GAP}}(c)=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{F}(i,j,c) \tag{25}$$

其中,$\boldsymbol{F}_{(i,j,c)}$ 为特征图在位置 (i,j) 以及通道 c 处的值,$\boldsymbol{F}_{\text{GAP}}(c)$ 为通道 c 的平均值. 完成全局平均池化后,将得到的特征向量输入到FC层进行处理,FC层将特征向量映射到分类任务所需的特定维度并进行输出,通过softmax激活函数计算各种恶意软件家族的分类概率 $\hat{y}_i$. 具体计算过程为

$$\hat{y}_i=\text{softmax}(\boldsymbol{W}\cdot\boldsymbol{F}_{\text{GAP}}+\boldsymbol{b}) \tag{26}$$

其中,$\boldsymbol{W}\in\mathbb{R}^{K\times C}$ 为全连接层的权重矩阵,$\boldsymbol{b}\in\mathbb{R}^{K}$ 为偏置向量,$\text{softmax}(\boldsymbol{Z}_i)$ 为softmax激活函数,$\hat{y}_i$ 为第 i 类家族的分类概率. 通过选择最大概率值对应的类别,完成恶意软件家族的检测分类.

6　实验结果与分析

6.1　实验设置与实验数据集

实验的计算机配置如下:Intel Core i7-12700K处理器,32 GB主存,NVIDIA GeForce 3070Ti显卡,操作系统为Windows10,采用的深度学习框架为Pytorch2.0,训练和测试均在该环境下进行. 本文研究所用数据集为Microsoft2015恶意软件分类大赛的训练数据(BIG2015)和奇安信科技研究院开放数据计划提供的DataCon2020数据集. 原始BIG2015数据集包含9类恶意家族,共计10 868个恶意软件样本;DataCon2020数据集包含从网络中捕获的真实挖矿恶意软件和其他类型恶意软件,共计23 655个样本,在实验中,2个数据集均按照8:2的比例划分为训练集与测试集.

6.2　恶意软件家族检测分类对比实验结果

为验证本文所提方法对恶意软件家族检测分类的有效性,分别采用Shallow-CNN[16](Shallow Convolutional Neural Network)、FACILE[11](Fewer cApsules and riCher hIerarchical information for maLware imagE classification)、MCTVD[7](Malware Classification method based on Three-channel Visualization and Deep learning)、BHMDC[17](Byte and Hex n-gram based Malware Detection and Classification)、BiTCN[18](Bi-directional Temporal Convolutional Networks transfer learning atrous spatial pyramid pooling efficientNet)、BiCS[19](malicious code detection method with BiLSTM and Channel attention mechanism-Spatial attention mechanism)、Mal3S[20](Malware detection for Static Security Service)、RMVC[21](combination of RNN Minhash Visualization and CNN)和本文所提方法,在BIG2015数据集上进行检测分类实验. 采用改进的CNN[22]、融合多特征集成学习[23]、SFCWGAN-BiTCN[24](Selection Feature Conditional Wasserstein Generative

Adversarial Network- Bidirectional Temporal Convolutional Network)、TriCh-LKRepNet[25](Triple-Channel Large Kernel Reparameterisation Network)、MSA-ResNet[26](Multi-head Spatial Attention-Residual Network)和本文所提方法在DataCon2020数据集上进行检测分类实验. 使用准确率、召回率、精确率和F_1分数4个指标评估模型检测分类性能,不同模型对BIG2015数据集恶意软件家族分类的结果如表1所示,对DataCon2020数据集的检测结果如表2所示.

表1 各方法在BIG2015数据集上的对比实验结果 单位:%

方法	准确率	召回率	精确率	F_1分数
Shallow-CNN	98.56	—	—	94.43
FACILE	97.20	92.14	93.99	92.63
MCTVD	99.44	99.13	99.44	99.29
BHMDC	99.26	99.26	99.27	99.25
BiTCN	99.64	99.64	99.64	99.64
BiCS	97.75	97.47	97.42	97.41
Mal3S	98.43	96.75	98.32	97.53
RMVC	99.50	—	—	—
本文方法	99.72	99.71	99.72	99.72

表2 各方法在DataCon2020数据集上的对比实验结果 单位:%

方法	准确率	召回率	精确率	F_1分数
改进的CNN	97.78	97.76	97.80	97.78
融合多特征集成学习	96.99	94.05	—	92.19
SFCWGAN-BiTCN	96.96	96.95	96.97	96.98
TriCh-LKRepNet	97.55	97.55	97.57	97.52
MSA-ResNet	97.70	97.72	97.67	97.76
本文方法	97.99	97.97	98.01	97.99

由表1的实验结果可知,本文所提方法在BIG2015数据集的准确率、召回率、精确率和F_1分数分别为99.72%、99.71%、99.72%和99.72%. 与其他8种实验方法相比,文本方法准确率、召回率、精确率和F_1分数最大提升幅度分别为2.52、7.57、5.73和7.09个百分点,各指标平均分别提升0.89、1.98、1.46和2.23个百分点,提升效果明显,说明本文所提方法能够更准确地完成恶意软件家族分类的任务. 对比实验具体分析如下.

(1)FACILE和BiCS在所有检测方法中性能最差,表明单一的灰度图像无法包含足够的全局图像特征,难以准确地完成分类任务. BiTCN、BHMDC和Mal3S方法侧重于分析恶意软件的文本特征,但此类方法容易受到混淆技术的干扰,且存在特征稀疏难以捕获的问题,导致检测分类效果不佳.

(2)Shallow-CNN侧重于分析融合图像样本特征进行检测分类,但并未考虑二进制源代码的语义信息内容;RMVC利用Minhash处理汇编操作码序列生成局部特征图进行检测分类,但该方法仅考虑局部特征,忽略了全局特征,导致特征提取能力不足,影响检测性能.

(3)MCTVD使用汇编指令序列组合转移概率,生成Markov图像进行检测分类,但包含恶意软件控制流的结构性信息在Markov转移概率中不能充分地体现;FACILE胶囊网络通过动态卷积提取灰度图像特征并进行多级特征融合,但在处理大量图像信息时,仍会出现胶囊脱落导致图像信息丢失的情况,导致特征信息丢失、降低检测性能等问题的出现.

由表2的实验结果可知,本文所提方法在DataCon2020数据集的准确率、召回率、精确率和F_1分数分别为97.99%、97.97%、98.01%和97.99%. 与其他5种实验方法相比,文本方法准确率、召回率、精确率和F_1分数最大提升幅度分别为1.03、3.92、1.04和5.80个百分点,各指标平均分别提升0.59、1.16、0.51和1.54个百分点,提升效果明显. 与基于融合多特征集成学习和SFCWGAN-BiTCN方法相比,本文所提方法能够有效捕获数据的动态指令跳转特征,通过分析汇编指令文件的操作数、操作码以及跳转指令分布概率,能够获得信息表达更丰富全面的文本特征. 与改进的CNN、TriCh-LKRepNet和MSA-ResNet方法相比,本文所提方法能够通过融合通道信息表示的空间注意力机制,综合考虑通道与空间层面的特征分布,有效增强提取到的融合图像特征用于检测分类实验. 对上述对比实验结果进行分析可知,相较于现有研究方法,本文方法在恶意软件家族检测分类任务中具有显著优势. 具体优势如下.

(1)传统检测方法在应对应用复杂加壳、混淆技术的恶意软件新变种时难以有效捕获其恶意行为特征,而本文所提出的图像化检测方法通过三通道图像生成与多维度动态加权alpha图像融合,能够获得包含更为丰富原始信息的融合图像样本,有助于全面捕捉恶意软件的全局和局部特征. 基于FCISA的特征增强方法,能够优化特征的表达能力,减少特征信息丢失,提高检测精度与检测分类性能.

(2)传统检测模型普遍仅考虑单一的图像特征或文本特征,包含特征信息维度不够全面,本文所提方法通过双分支特征提取与多模态特征融合,能够综合利用文本和图像特征提供更全面的特征表达,增强模型的鲁棒性和抗干扰能力,确保在面对混淆恶意软件时能更准确地进行检测分类. 相比使用单一特征的方法,多模态融合方法能更有效地捕捉恶意软件综合特征,显著提升检测分类效果.

6.3 不同图像化方法对检测分类性能的影响

使用不同的图像化方法生成的图像样本往往包含不同的恶意软件原始信息,对恶意行为的反映程度也有所不同. 因此,为验证本文所提图像化方法对恶意软件家族检测分类效果的影响,使用多种传统图像化方

法和现有研究的创新性图像化方法进行对比实验. M1为传统灰度图像映射方法,M2为低维灰度共生矩阵可视化方法,M3为马尔科夫状态转移矩阵可视化方法,R方法、G方法、B方法和其组合方法则为单独使用某单一通道的图像和融合使用任意两通道图像的可视化方法,实验结果如表3所示.

由表3可知,与传统的3种图像化方法M1、M2、M3相比,使用本文所提出的融合图像生成方法进行恶意软件检测分类,其分类准确率、召回率、精确率和F_1分数最大提升幅度分别为1.47、2.54、2.24和1.79个百分点,各指标平均分别提升1.34、1.76、1.58和1.63个百分点,检测性能得到显著提升. 本文方法不仅可以获取字节信息,还能获取包括操作码、文件熵值在内的多种特征,能够更全面反映恶意软件的行为和结构特点,增强图像化恶意软件特征的差异性,从而提高对恶意软件家族的检测分类性能.

表3　不同图像化方法在BIG2015数据集上的对比实验结果 单位:%

方法	准确率	召回率	精确率	F_1分数
M1	98.51	98.12	98.37	98.24
M2	98.25	98.67	97.48	98.08
M3	98.43	97.17	98.65	97.91
R	98.99	98.76	98.82	98.79
G	99.06	98.57	99.25	98.91
B	99.12	98.94	99.15	99.04
R+G	99.26	98.85	99.03	98.94
R+B	99.21	99.14	99.18	99.16
G+B	99.33	99.06	99.21	99.13
本文方法R+G+B	99.72	99.71	99.72	99.72

与单独使用R、G、B这3种不同类型的通道填充图像,和融合使用R+G、R+B、G+B的通道填充图像相比,本文方法的分类准确率、召回率、精确率和F_1分数最大提升幅度分别为0.74、1.16、0.91和0.94个百分点,各指标平均分别提升0.56、0.62、0.62和0.73个百分点,检测性能得到显著提升. 单独使用部分通道的特征会存在信息丢失和特征不充分等问题. 而通过融合RGB三通道图像,能够综合利用各类图像互补特征信息,形成更加全面和细致的特征描述,提高模型对不同恶意软件家族特征的识别能力,提升检测分类效果.

为进一步验证本文方法的有效性,在相同的实验环境和参数配置条件下,将本文方法与现有的经典检测网络模型VGG19、ResNet50、VisionTransformer、MobilenetV3进行对比实验,在BIG2015数据集和DataCon2020数据集上的检测分类准确率分别如图6和图7所示.

由图6可知,本文所提方法以及各经典检测网络模型在BIG2015数据集上均表现效果较好,检测分类准确率普遍高于97%,使用本文提出的检测网络模型和R+G+B图像化方法可以获得99.72%的最高检测分类准确率,表明本文所提出的多维度动态加权alpha图像融合方法能够充分利用恶意软件样本的图像化特征,提高检测分类的精度. 由图7可知,在实际网络场景收集的恶意软件数据集DataCon2020上,各检测模型均出现准确率明显下降的情况,检测分类准确率最高为97.99%,与在BIG2015数据集上的实验结果相比,下降幅度为1.73个百分点. 其原因为,BIG2015数据集收集的恶意软件样本特征分布较为规律,样本间特征差异相对显著且类间边界清晰,而DataCon2020数据集包含的恶意软件种类复杂多样,并且引入了大量应用混淆、加壳技术的新变种,导致模型无法有效提取恶意行为特征进行检测分类.

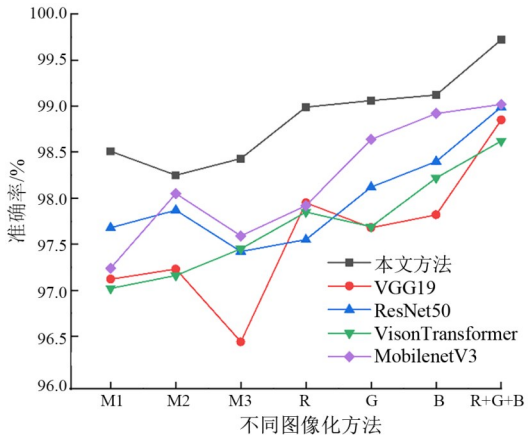

图6　基于BIG2015数据集的检测结果

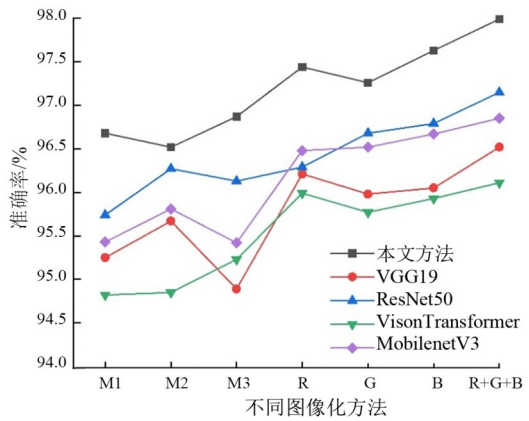

图7　基于DataCon2020数据集的检测结果

6.4　傀儡优化算法对检测分类性能的影响

为分析傀儡优化算法POA对检测分类效果的影响,在相同的实验环境和参数配置条件下,分别使用POA和经典的合成少数类过采样(Synthetic Minority Oversampling Technique,SMOTE)方法对BIG2015数据集进行数据重构,通过实验得到各家族分类结果的混

淆矩阵结果分别如图8和图9所示. 由图8和图9可知,与使用SMOTE的传统数据重构方法相比,使用POA方法可以显著提高模型对难区分少类家族“Simda”和“Tracur”的分类精度,以及模型最终的平均检测分类准确率. 使用SMOTE方法的平均检测分类准确率为99.35%,而使用POA方法的平均检测分类准确率为99.72%,提升幅度为0.37个百分点. 在未使用任何数据集优化方法时,模型最终的检测分类准确率可达99.31%,但是分类器倾向于将大多数样本分类到最多的类别而忽略少数类别的预测性能,因此对“Simda”“Tracur”家族的分类准确率仅为98.67%和98.85%,远低于平均分类精度.

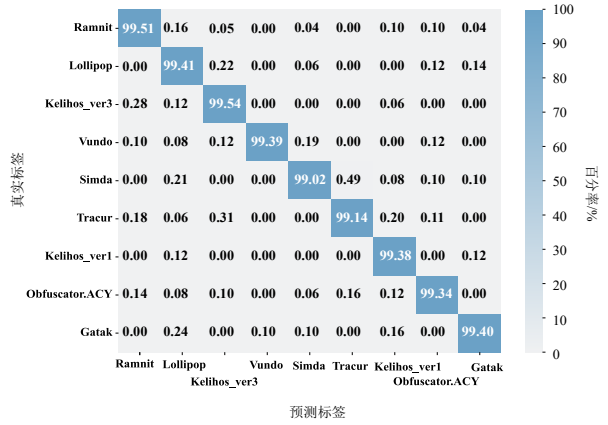

图8 使用SMOTE方法的检测结果

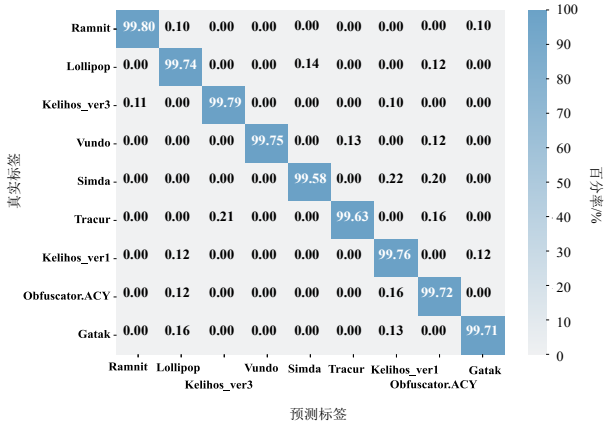

图9 使用POA方法的检测结果

此外,使用SMOTE方法和POA方法完成数据重构后,网络模型对“Simda”“Tracur”家族的分类准确率分别为99.02%、99.14%和99.58%、99.63%,对其他家族的分类准确率有所提升,这表明SOMTE方法和POA方法能够有效减少模型对少类家族的混淆,提高分类的准确性. 但POA方法对少类家族分类准确率提升效果更为显著,与SMOTE方法相比,POA方法对少类家族“Simda”“Tracur”的分类准确率分别提升0.91个百分点0.78个百分点,对多类家族“Lolipop”“Kelihos_ver3”的分类准确率分别提升0.33个百分点和0.25个百分点.

如图8所示,SMOTE方法能够通过生成新的少数类样本增加其类别权重,使得分类器在训练过程中更加关注少数类,但在平衡后的数据集中网络模型需在各类样本间找到更复杂的边界,导致模型对例如“Ranmnit”“Vundo”“Gatak”等多类样本的预测准确率有所下降,具体下降幅度分别为0.16、0.09和0.02个百分点. 而如图9所示,使用POA方法的网络模型在所有类别的分类准确率均有不同程度的提高,通过多类样本的重排序和样本下采样机制,选择了具有较高代表性和较大贡献度的样本用于模型训练. 对于“Ramnit”和“Obfuscator.ACY”这2个多类家族,POA方法在确保数据集平衡的同时通过增强样本的多样性,使得模型对其特征变化的敏感度得到提升. 与SMOTE方法相比,POA方法对“Ranmnit”和“Obfuscator.ACY”的分类准确率分别提升0.12个百分点和0.38个百分点. 表明POA方法能有效提升模型对恶意软件家族特征的捕捉能力,并且在不牺牲多类表现的前提下优化模型的整体性能,避免过度调整导致的多类家族分类准确度下降等问题

6.5 多模态特征融合对检测分类性能的影响

为分析多模态特征融合方法对恶意软件家族检测分类性能的影响,分别单独使用图像特征、文本特征和多模态融合特征在BIG2015数据集、DataCon2020数据集进行实验,实验结果分别如图10和图11所示.

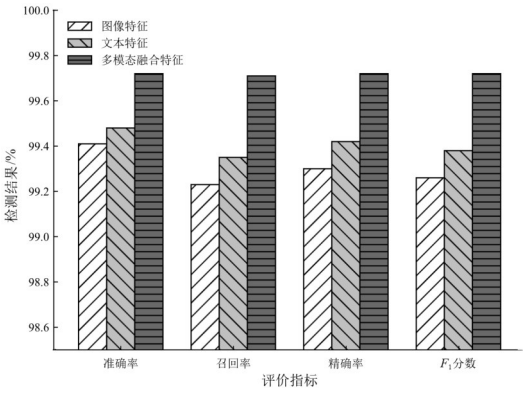

图10 融合特征在BIG2015数据集上的检测结果

由图10可知,在BIG2015数据集上的实验结果中,采用多模态特征融合方法相较于使用单一图像特征的检测方法,分类准确率、召回率、精确率和F_1分数分别提升了0.31、0.48、0.42和0.46个百分点. 与单独使用文本特征相比,各项指标提升幅度分别为0.24、0.36、0.30和0.34个百分点. 实验结果表明,多模态特征融合能够显著提升恶意软件家族分类的检测性能.

同样,在DataCon2020数据集上的实验结果(如

图11)进一步验证了这一趋势. 与仅依赖图像特征的检测方法相比,多模态特征融合方法的分类准确率、召回率、精确率和F_1分数分别增加0.78、0.84、0.92和0.88个百分点;而相较于单一的文本特征,各指标提升幅度分别为0.47、0.64、0.31和0.45个百分点. 这种显著的性能提升归因于多模态特征融合方法的优势,通过同时利用图像和文本特征,更全面地捕捉恶意软件样本的多维度特征,进而生成更加细致且具有区分度的特征表示. 相比之下,单一特征提取方法在反映样本多样性和捕捉复杂行为模式上存在局限性,无法充分揭示恶意软件家族的内在结构差异. 综上所述,多模态融合特征在分类准确率、召回率、精确率和F_1分数等各项指标上均表现出明显的优势,能够通过整合多维特征信息提供更强的特征表达能力,有效提升恶意软件检测与分类的整体性能.

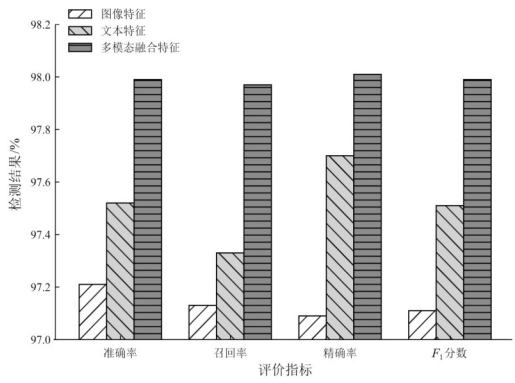

图11　融合特征在DataCon2020数据集上的检测结果

6.6　模型训练时间开销成本与应用可行性分析

为评估本文方法在模型训练各阶段的时间开销成本与实际应用可行性,在相同实验环境下进行验证实验,分别记录使用BIG2015数据集进行实验时,平均每个数据样本在数据预处理、双分支特征提取、特征增强和多模态特征融合阶段所需的时间成本. 各阶段具体耗时结果如表4所示.

由表4内容可知,本文所提方法在融合图像生成、基于POA的数据重构阶段时间成本开销较大. 融合图像生成耗时中位数和平均数分别为1.013 s和1.541 s,分析其原因为恶意软件样本部分原始二进制文件占用内存较大、字节较多,需要较高的时间成本完成图像转换过程;而基于POA的数据重构方法需要根据原始图像样本的特征生成对应的傀儡样本,因此,也需要较高的时间成本,其耗时中位数和平均数分别为0.217 s和0.646 s. 尽管上述2个阶段的时间开销成本较高,但该过程归属数据预处理且整个实验流程中仅需执行1次,不会影响实时推理效率. 此外,高质量融合图像样本能够确保模型在面对具有复杂结构、混淆技术的恶意软件时,可以捕捉到更多维度的行为模式和结构特征,对于后续特征提取和分类的准确率至关重要.

表4　模型训练各阶段时间消耗结果　　单位:s

阶段名称	耗时最小值	耗时最大值	耗时平均值	耗时中位数值
融合图像生成	0.502	3.418	1.541	1.013
基于POA的数据重构	0.103	0.811	0.646	0.217
图像特征提取	0.007	0.094	0.032	0.011
文本特征提取	0.028	0.366	0.087	0.062
基于FCISA的特征增强	0.009	0.131	0.046	0.029
多模态特征融合	0.001	0.004	0.001	0.001

模型训练包含图像特征提取、文本特征提取、基于FCISA的特征增强和多模态特征融合4个阶段. 通过分析表4内容可知文本特征提取阶段耗时最长,耗时中位数和平均数分别为0.062 s和0.087 s,分析其原因为文本特征提取阶段需要处理大量的汇编指令文件,频繁地切换读取不同文件占用了较多时间. 基于FCISA的特征增强和特征融合阶段包含较多的计算过程,因此对实验设备的硬件要求较高,在本文实验环境下耗时中位数和平均数分别为0.030 s和0.047 s. 此外,为进一步验证本文方法在模型测试阶段的时间成本开销与实际应用可行性,使用经典的VGG19、ResNet50、MobileNetV3、DenseNet121网络模型在重构样本集上进行实验,通过统计类别预测时间对比不同方法的分类时间复杂度. 各模型测试具体耗时分别为7.85、6.99、5.92、6.41 s,而本文所提方法的平均测试耗时为10.27 s. 尽管本文所提方法时间开销成本相对较高,但与上述经典方法相比,本文方法在检测分类性能上具有显著的优势,且模型平均检测耗时在合理可控范围内. 综上所述,本文方法具备可控的时间开销成本和较高的实际应用可行性,并且在实际应用时还能通过多设备并行处理和硬件设备优化达到减少时间消耗、加速实验流程的目的,适用于恶意软件检测分类领域的实际场景应用.

6.7　特征融合权重对检测分类性能的影响

在特征增强与多模态特征融合阶段,引入超参数a平衡增强后的图像特征$\boldsymbol{F}_{\text{img_e}}$和文本特征$\boldsymbol{F}_{\text{text_e}}$所占比例,分别对$\boldsymbol{F}_{\text{img_e}}$和$\boldsymbol{F}_{\text{text_e}}$赋予超参数权重$a$、$(1-a)$得到最终的融合特征$\boldsymbol{F}_{\text{fused}}$. 为分析特征融合阶段中超参数$a$对实验结果造成的影响,设置不同的$a$值进行实验并分析其检测性能,其中$a\in[0,1]$. 通过分析实验结果可知,随着超参数权重$a$取值的变化,本文所提方法在BIG2015数据集和DataCon2020数据集上的检测分类准确率和F_1分数均呈先上升后下降的趋势. 当a取值为0.3和0.4时,模型在BIG2015数据集上可以获得较高的准确率和F_1分数,当a取值为0.4时,检测性能达到最

优,准确率为99.72%. 当a取值为0.4和0.6时,模型在DataCon2020数据集上可以获得较高的准确率和F_1分数,同样当a取值为0.4时,检测性能达到最优,准确率为97.99%. 综合考虑实验结果可知,超参数权重a取值设置为0.4.

7 结论

针对现有恶意软件检测分类任务中存在的特征信息丢失、特征结构破坏和检测精度较低等问题,本文提出一种基于多维度动态加权alpha图像融合与特征增强的恶意软件检测方法. 通过三通道图像生成与多维度动态加权alpha图像融合方法获取融合图像样本,代替单一类型的图像进行检测分类. 通过傀儡优化算法进行数据重构,平衡数据集中各类家族的样本分布. 通过EfficientNet和CLAP分别提取图像特征和文本特征,并通过基于融合通道信息表示的空间注意力机制进行特征增强,丰富特征信息表达. 最后将增强后的图像特征和文本特征进行加权融合,并通过全局平均池化、全连接和softmax激活函数完成对恶意软件的检测分类任务. 实验结果表明,本文所提方法对恶意软件家族的检测分类性能优于现有主流研究方法,能够充分挖掘恶意软件的深层次特征,有效融合文本语义特征和图像特征,提高检测分类性能.

在未来的研究中,将进一步分析并探索如何提取更高质量的图像特征和文本特征,同时挖掘与恶意软件相关的标签信息,从而提升对恶意软件的检测分类准确率.

参考文献

[1] SHU L H, DONG S, SU H D, et al. Android malware detection methods based on convolutional neural network: A survey[J].IEEE Transactions on Emerging Topics in Computational Intelligence, 2023, 7(5): 1330-1350.

[2] NAEEM H, DONG S, FALANA O J, et al. Development of a deep stacked ensemble with process based volatile memory forensics for platform independent malware detection and classification[J]. Expert Systems with Applications, 2023, 223: 119952.

[3] CHAI Y H, QIU J, YIN L H, et al. From data and model levels: Improve the performance of few-shot malware classification[J]. IEEE Transactions on Network and Service Management,2022, 19(4): 4248-4261.

[4] NATARAJ L, KARTHIKEYAN S, JACOB G, et al. Malware images: Visualization and automatic classification[C]// Proceedings of the 8th International Symposium on Visualization for Cyber Security. New York: ACM, 2011: 1-7.

[5] CUI Z H, XUE F, CAI X J, et al. Detection of malicious code variants based on deep learning[J]. IEEE Transactions on Industrial Informatics, 2018, 14(7): 3187-3196.

[6] DONG S, SHU L H, NIE S. Android malware detection method based on CNN and DNN bybrid mechanism[J]. IEEE Transactions on Industrial Informatics, 2024, 20(5): 7744-7753.

[7] DENG H X, GUO C, SHEN G W, et al. MCTVD: A malware classification method based on three-channel visualization and deep learning[J]. Computers & Security, 2023, 1126: 103084.

[8] NI S, QIAN Q, ZHANG R. Malware identification using visualization images and deep learning[J]. Computers & Security, 2018, 77(8): 871-885.

[9] VASAN D, ALAZAB M, WASSAN S, et al. Image-Based malware classification using ensemble of CNN architectures (IMCEC) [J]. Computers & Security, 2020, 92: 101748.

[10] NAEEM H, CHENG X C, ULLAH F, et al. A deep convolutional neural network stacked ensemble for malware threat classification in Internet of Things[J]. Journal of Circuits, Systems and Computers, 2022, 31(17): 2250302.

[11] ZOU B H, CAO C J, WANG L J, et al. FACILE: A capsule network with fewer capsules and richer hierarchical information for malware image classification[J]. Computers & Security, 2024, 137: 103606.

[12] TANG Y H, QI X Y, JING J, et al. BHMDC: A byte and hex n-gram based malware detection and classification method[J]. Computers & Security, 2023, 128: 103118.

[13] WANG H, GAO Z Y, ZHANG C, et al. CLAP: Learning transferable binary code representations with natural language supervision[C]//Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis. New York: ACM, 2024: 503-515.

[14] WANG Y D, ZHANG J, KAN M N, et al. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 12272-12281.

[15] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional Block Attention Module[M]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 3-19.

[16] CONTI M, KHANDHAR S, VINOD P. A few-shot malware classification approach for unknown family recognition using malware feature visualization[J]. Computers & Security, 2022, 122: 102887.

[17] KUMAR S, JANET B. DTMIC: Deep transfer learning for malware image classification[J]. Journal of Information Security and Applications, 2022, 64: 103063.

[18] XUAN B N, LI J, SONG Y F. BiTCN-TAEfficientNet malware classification approach based on sequence and RGB fusion[J]. Computers & Security, 2024, 139: 1-17.

[19] SHEN G N, CHEN Z X, WANG H, et al. Feature fusion-based malicious code detection with dual attention mechanism and BiLSTM[J]. Computers & Security, 2022, 119: 102761.

[20] JEON J, JEONG B, BAEK S, et al. Static multi feature-based malware detection using multi SPP-net in smart IoT environments[J]. IEEE Transactions on Information Forensics and Security, 2024, 19(1): 2487-2500.

[21] SUN G S, QIAN Q. Deep learning and visualization for identifying malware families[J]. IEEE Transactions on Dependable and Secure Computing, 2021, 18(1): 283-295.

[22] 轩勃娜, 李进. 基于改进CNN的恶意软件分类方法[J]. 电子学报, 2023, 51(5): 1187-1197.
XUAN B N, LI J. Malware classification method based on improved CNN[J]. Acta Electronica Sinica, 2023, 51(5): 1187-1197. (in Chinese)

[23] 杨望, 高明哲, 蒋婷. 一种基于多特征集成学习的恶意代码静态检测框架[J]. 计算机研究与发展, 2021, 58(5): 1021-1034.
YANG W, GAO M Z, JIANG T. A malicious code static detection framework based on multi-feature ensemble learning[J]. Journal of Computer Research and Development, 2021, 58(5): 1021-1034. (in Chinese)

[24] XUAN B N, LI J, SONG Y F. SFCWGAN-BiTCN with sequential features for malware detection[J]. Applied Sciences, 2023, 13(4): 1-21.

[25] LI S C, WANG J, SONG Y F, et al. RETRACTED: TriCh-LKRepNet: A large kernel convolutional malicious code classification network for structure reparameterisation and triple-channel mapping[J]. Computers& Security, 2024, 144: 103937.

[26] LI S C, WANG J, SONG Y F, et al. Tri-channel visualised malicious code classification based on improved ResNet[J]. Applied Intelligence, 2024, 54(23): 12453-12475.

作者简介

谢丽霞　女, 1974年4月出生, 重庆人. 中国民航大学教授. 主要研究方向为网络与系统安全、信息安全.
E-mail: lxxie@126.com

魏晨阳　男, 1998年9月出生, 河南新乡人. 中国民航大学硕士研究生. 主要研究方向为网络信息安全、恶意软件检测.
E-mail: wcy17337373217@163.com

杨宏宇　男, 1969年12月出生, 吉林长春人. 博士, 中国民航大学教授, 博士生导师. 主要研究方向为网络与系统安全、漏洞分析与评估、云计算与大数据安全.
E-mail: yhyxlx@hotmail.com

胡　泽　男, 1989年7月出生, 山西临汾人. 博士, 中国民航大学讲师. 主要研究方向为自然语言处理、人工智能、信息安全.
E-mail: zhu@cauc.edu.cn

成　翔　男, 1988年9月出生, 新疆乌鲁木齐人. 博士, 扬州大学实验师. 主要研究方向为网络与系统安全、网络安全态势感知、联邦学习、边缘计算.
E-mail: huozhai9527@126.com