

基于知识蒸馏的恶意代码家族检测方法



王润正¹ 高见^{1,2} 黄淑华^{1,2} 仝鑫¹

1 中国人民公安大学信息安全学院 北京 100038

2 安全防范与风险评估公安部重点实验室 北京 102623

(root@righte0us.cc)

摘要 近年来,恶意代码变种层出不穷,恶意软件更具隐蔽性和持久性,亟需快速有效的检测方法来识别恶意样本。针对现状,文中提出了一种基于知识蒸馏的恶意代码家族检测方法,该模型通过逆向反编译恶意样本,利用恶意代码可视化技术将二进制文本转为图像,以此避免对传统特征工程的依赖。在教师网络模型中采用残差网络,在提取图像纹理深层次特征的同时,引入通道域注意力机制,根据通道权重的变化,来提取图像中的关键信息。为了加快对待检测样本的识别效率,解决基于深度神经网络检测模型参数量大和计算资源消耗严重等问题,使用教师网络模型来指导学生网络模型训练,实验结果表明学生网络在降低模型复杂度的同时,保持了恶意代码家族的检测效果,有利于对批量样本的检测和移动端的部署。

关键词: 恶意家族;知识蒸馏;注意力机制;残差网络

中图法分类号 TP309

Malicious Code Family Detection Method Based on Knowledge Distillation

WANG Run-zheng¹, GAO Jian^{1,2}, HUANG Shu-hua^{1,2} and TONG Xin¹

1 College of Information and Cyber Security, People's Public Security University of China, Beijing 100038, China

2 Key Laboratory of Safety Precautions and Risk Assessment, Beijing 102623, China

Abstract In recent years, the variety of malicious code emerges in an endless stream, and malware is more covert and persistent. It is urgent to identify malicious samples by rapid and effective detection methods. Aiming at the present situation, a method of malicious code family detection based on knowledge distillation is proposed. The model decompiles malicious samples in reverse and transforms binary text into images by malicious code visualization technology, so as to avoid dependence on traditional feature engineering. In the teacher network model, residual network is used to extract the deep-seated features of image texture, and channel domain attention mechanism is introduced to extract the key information from the image according to the change of channel weight. In order to speed up the identification efficiency of the samples to be tested and solve the problems of large parameters and serious consumption of computing resources based on deep neural network detection model, the teacher network model is used to guide the training of the student network model. The results show that the student network maintains the detection effect of malicious code family on the basis of reducing the complexity of the model. It is conducive to the detection of batch samples and the deployment of mobile terminal.

Keywords Malicious family, Knowledge distillation, Attention mechanism, Residual network

1 引言

随着计算机、物联网、大数据、云计算等技术的发展,恶意代码的复杂化和多样化对网络空间安全构成了严重的威胁,黑客通过恶意代码发动各种网络攻击,破坏系统的安全性,给

国家和社会造成了极大的损失。2019年,国家互联网应急中心(CNCERT/CC)持续监测恶意程序传播情况,全年捕获的恶意程序样本数量约为6287万,涉及恶意代码家族约66万个,新增恶意代码家族157个。全年监测到恶意程序传播次数约30.0亿,恶意程序月平均传播约2.5亿次。与此同时,

到稿日期:2020-09-13 返修日期:2020-11-10 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。

基金项目:国家重点研发计划;国家社会科学基金重点项目(20AZD114);公安部科技强警基础工作2020专项(2020GABJC01);中国人民公安大学中央基本科研业务费项目(2019JKF218)

This work was supported by the National Key R&D Program of China, Key Program of the National Social Science Foundation of China (20AZD114), 2020 Special Project of Science and Technology Strengthening Police of Ministry of Public Security(2020GABJC01) and Basic Scientific Research Operating Expenses of the People's Public Security University of China(2019JKF218).

通信作者:高见(gaojian@ppsuc.edu.cn)

为了规避反病毒软件的检测,攻击者利用各种隐蔽技术、变形技术生成新的恶意代码变体。随着恶意家族种类和数量的激增,传统的恶意代码检测方法会消耗大量的资源且分析速度缓慢,不足以应对大规模的恶意行为分析。因此,如何在降低样本恶意检测分析时间开销的同时,有效地对恶意代码家族进行检测已成为恶意代码研究的挑战之一。

目前,在恶意代码家族检测研究中,研究人员结合深度学习,通过大量样本训练神经网络模型,一定程度上提升了恶意代码分析的速度。但随着神经网络的加深,恶意代码家族检测模型在训练和预测的过程中,不可避免地需要大量的时间开销,以使模型获取到更好的检测效果。

针对上述问题,本文采用知识蒸馏的思想训练恶意代码家族检测模型,使用带有注意力机制的残差网络作为教师网络模型对恶意代码家族进行识别。该方法选取样本反编译后的字节码特征,结合深度学习的方法来识别恶意代码家族,从教师网络所学习的知识信息中学习关键信息来训练学生模型,在保证恶意代码家族检测效能相近的情况下对模型进行压缩,减少了资源消耗,加快了待检测样本的识别速度。

2 相关工作

恶意代码分析主要分为动态分析和静态分析。动态分析指将恶意样本运行在虚拟机或者安全沙箱中,分析样本在运行时的动态行为特征,判断样本是否含有恶意行为;静态分析则在不运行恶意样本的情况下,通过逆向工程手段获取二进制文件的操作码、API 等静态特征,判断样本是否为恶意样本。越来越多的研究人员采用静态分析和动态分析来提取恶意代码特征,结合机器学习和深度学习,提出了一系列基于行为特征的检测方法,这些方法大多使用程序 API 序列、网络信息等进行辅助分析。Chen 等^[1]结合 Cuckoo 和改造后的 DynamoRIO 虚拟环境,提取恶意代码的动态行为特征和网络流量信息,利用双向门循环单元网络对恶意代码进行检测。Zhao 等^[2]在沙箱中提取恶意代码的 API 序列,使用基于程序语义 API 依赖图的真机动态分析方法,采用集成学习算法——随机森林进行恶意代码分类。Mohanasruthi 等^[3]提出了一种基于复杂网络(Malware Detection using Complex Network, MDCN)的恶意软件检测技术,该技术采用应用程序接口调用转换矩阵(Application Program Interface Call Transition Matrix, API-CTM)生成复杂的网络拓扑,当使用多层感知器作为分类器时,模型达到了较高的准确率。Narayanan 等^[4]分别采取恶意代码可视化和静态提取操作码两种特征,使用不同的模型进行对比实验,通过实验结果发现,首先将恶意样本反编译得到的操作码转换为序列,然后使用 CNN 和 LSTM 提取特征,最后采用支持向量机或逻辑回归进行分类的集成方法的效果更优。Hu 等^[5]结合静态分析和动态分析提取恶意行为特征,改进了基于高斯混合模型的增量聚类方法来识别恶意软件家族。静态分析和动态分析各有所长,面对目前恶意代码的规避性和反调试性,动态分析在提取特征时存在一定的困难,此时静态分析更易对恶意

代码家族进行检测和分类。

近年来,研究人员在静态分析或动态分析的基础上,将传统的文本特征通过可视化的方法转换为图像特征,结合图像的处理方式对恶意代码家族进行检测和分类。Zeng 等^[6]将提取恶意样本反编译后的特征转化为灰度图,采用基于 MobileNet 的模型对恶意软件家族进行分类。Sun 等^[7]在传统灰度图的基础上,加入恶意样本的 ASCII 字符信息和 PE 结构信息构建 RGB 三维图像,并使用一种带有空间金字塔池化结构的 VGG16 神经网络模型对恶意代码图像进行训练和预测。Vasan 等^[8]提出了一种基于图像的集成卷积神经网络(Image-based Malware Classification using Ensemble of CNNs, IMCEC)的恶意软件检测方法,采用 VGG16 和 ResNet50 模型深层次地提取不同的图像语义特征,实验表明相比于传统方法,IMCEC 可以更好地提取特征,达到较高的分类准确率。Ghouti^[9]将恶意软件二进制文件转化为图像,使用主成分分析在紧凑的子空间中提取恶意软件类别和结构的高度区分性特征,最后使用优化的支持向量机模型对恶意软件进行分类。Jain 等^[10]将恶意样本可视化为图像,分别使用二维图像和从图像派生的一维向量作为模型的输入,采取卷积神经网络(CNN)和极限学习机器(ELM)进行分类。Zhuo 等^[11]提出了两种基于 n-gram 特征的恶意软件可视化分析方法,即空间填充曲线映射(SFCM)方法和马尔可夫点图(MDP)方法,并采用这两种可视化方法通过深度卷积网络和支持向量机模型对图像进行分类。

上述研究说明可视化方法^[12-14]对恶意代码分析具有一定的可行性,能够有效地对恶意代码家族进行检测。可视化的方法相比动态提取特征,在面向大批量的待检测样本时,减小了对传统特征工程的依赖,大大提高了检测效率。因此,本文提取各类恶意代码家族的静态特征,将其可视化为 RGB 彩色图像,根据同一恶意代码家族之间的相似性对恶意样本进行识别。

3 检测模型的设计与实现

在计算机视觉和自然语言处理领域,为了获得良好的检测效果,深度神经网络往往采用深而宽的网络结构,这些网络模型在实际应用中受到了时空的制约,无法满足实时需求且会消耗大量的计算资源,不适合移动端的部署。针对上述问题,研究人员提出了模型压缩的方法,如知识蒸馏、剪枝和量化。其中,剪枝法虽然可降低网络的复杂度,但计算复杂度;量化法需要与硬件结合来提高推理速度,但检测精度会明显下降。而知识蒸馏法可以在不过度影响模型性能的基础上,减小计算量和参数量,其主要思想是训练一个小型网络来学习一个预先训练好的复杂网络或者集成网络,这种训练模式称为“Teacher-Student”,复杂网络是“教师网络”,小型网络是“学生网络”。自 Hinton 等^[15]首次提出知识蒸馏的方法后,近年来,为了使学生网络在拥有更少参数量、更小规模的情况下,达到与教师网络相似甚至更优的精度,相关人员在该领域开展了一系列的研究工作,如再生神经网络模型^[16]和残

差知识蒸馏^[17]方法。知识蒸馏的思想已被逐渐应用于目标检测、语义分割、生成式对抗网络等领域,实现知识迁移。

在基于深度学习的恶意代码检测模型中,对浅层网络叠加深度数会使模型的表达能力更强,在训练集和测试集上的拟合能力更好,然而网络层数增多后,参数量急剧增加,会造成计算资源的过度消耗。为解决基于深度神经网络检测模型的复杂性和资源消耗等问题,本文将知识蒸馏思想引入恶意代码家族检测领域,选取带有注意力机制的残差网络作为教师网络,传统的神经网络作为学生网络,通过蒸馏方法让浅层的神经网络达到深层复杂网络的分类准确率。该模型主要分为3个部分:特征提取模块、知识蒸馏模块和检测模块,整体架构如图1所示。

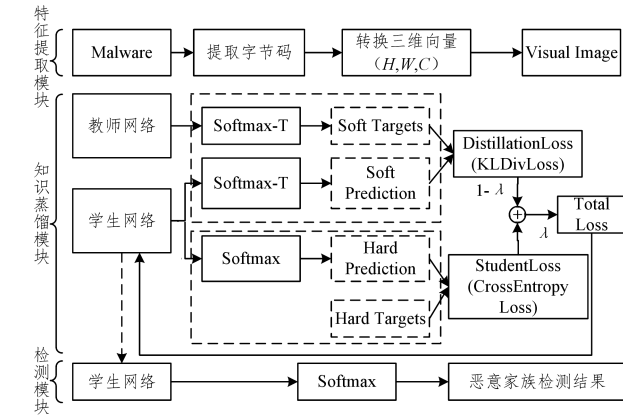


图1 模型架构

Fig. 1 Model architecture

3.1 特征提取

Nataraj 等^[18]首次提出将恶意软件二进制文件转换成灰度图的方法,利用图像纹理的相似性对恶意软件进行分类。恶意软件二进制文件由一系列0和1组成,将连续8bit数据转成0到255内的无符号整数,以对应RGB色彩的强度值,从而将其可视化。本文为了更好地提取恶意软件的特征信息,将连续24比特位(即3字节)分别填充到彩色图像的R,G,B三通道,即将第1个8bit的数据转化为R通道的数值,将第2个8bit的数据转化为G通道的数值,将第3个8bit的数据转化为B通道的数值。由于IDA Pro生成的字节码文件过大,本文按照固定的长宽直接对其进行截取后转换成对应的RGB图像,对长度小于固定值的图像使用0字节进行填充。恶意代码家族可视化流程如算法1所示。

(1)利用逆向工具IDA Pro反编译恶意样本,从Hex View窗口中提取十六进制数据,保存数据并生成bytes文件;

(2)按照可视化规则读取字节码文件,生成(H,W,C)三维矩阵,进而将三维矩阵中的数值映射到0~255,转成RGB彩色图像。

算法1 Bin2Image

Input: Binary sample $X = \{x_1, x_2, \dots, x_n\}$

Output: RGB Image; $Image = \{G_1, G_2, \dots, G_n\}$

1. Function Bin2Image(X):

2. for $i \leftarrow 1$ to X do
3. $Seq \leftarrow X[i]$ to HexSeq by IDA Pro
4. for n to $len(Seq)$ do
5. if Seq is not $\emptyset$ then
6. $R[n], G[n], B[n] \leftarrow Seq[3n], Seq[3n+1], Seq[3n+2]$
7. $Image[i] \leftarrow Matrix2Image[R, G, B]$
8. return Image

经过恶意代码可视化后,可以看出不同恶意代码家族生成的图像纹理特征具有差异。选取Ramnit, Lollipop, Kelihos_ver3, Obfuscator.ACY这4类恶意家族,经过可视化的图像如图2所示。

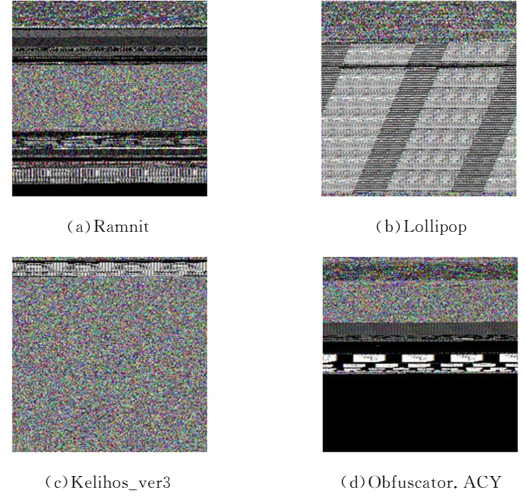


图2 4种恶意代码家族的纹理特征

Fig. 2 Four texture features of malicious code family

3.2 知识蒸馏

传统神经网络预测是将输入的数据送入多层卷积神经网络,经卷积层处理后送入全连接层,多层全连接层输出 logits 值,即 z_i , logits 经过 Softmax 得到预测概率 p_i ,如式(1)所示:

$$p_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)} \quad (1)$$

为了从教师网络中蒸馏出丰富有效的特征信息,引入温度参数 T 表示蒸馏的温度。 T 越高,网络输出类别的概率分布越平缓,学生网络越能从教师网络中学习更丰富的特征信息。升温后 Softmax 得到预测的概率分布 q_i ,如式(2)所示:

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \quad (2)$$

在知识蒸馏^[15]中,蒸馏的目的是让学生网络的类别输出预测分布尽可能拟合教师网络的输出预测分布,过程如下:

(1)训练教师网络。利用恶意代码家族可视化数据对教师网络进行训练,获得表征较强的教师网络,网络输出 logits 后,利用升温后的 Softmax 得到预测的概率分布 Soft Targets;学生网络输出 logits 经过相同温度 T 进行蒸馏,经过 Softmax 层之后得到类别预测概率,作为 Soft Prediction。

(2)蒸馏教师网络知识到学生网络。构造 Distillation loss 和 Student loss,将其加权相加作为最终的损失函数 L (见式(3))。在学生网络中,使用真实标签作为 Hard Targets,可以有效阻止教师网络中的错误信息被蒸馏到学生网络中。

$$L = \lambda CE(y, p) + (1 - \lambda) T^2 KL(p, q) \quad (3)$$

其中, CE 为交叉熵损失函数, KL 为 KL 散度, y 为真实标签, p 为学生网络预测概率, q 为教师网络预测概率, λ 为权重。 KL 散度作为蒸馏损失的计算方式, 用于衡量学生网络预测分布与教师网络预测分布的相近程度。交叉熵损失函数作为学生损失的计算方式, 用于衡量学生网络预测分布和真实标签之间的差异。

3.3 教师网络模型

3.3.1 残差网络

复杂网络模型往往会产生过拟合、梯度消失、梯度爆炸等问题, 模型性能会快速下降。针对越深的网络模型越难以优化的问题, Kaiming 等^[19]提出残差网络, 在原有网络上实现恒等映射 (Identity Mapping), 解决了由增加网络深度而产生的网络退化问题。本文在教师网络模型中加入残差网络, 改变前向和后向传播的方式, 以便在增加网络深度的同时, 更好地提取恶意代码图像深层的纹理特征, 提高模型的准确率。

残差网络由一系列残差块 (Residual Block) 组成, 其中一个残差块可以分为恒等映射和残差部分, 如图 3 所示。残差块的计算公式如下:

$$H(x) = F(x) + x \quad (4)$$

其中, x 为浅层网络的输出, $H(x)$ 为深层网络的输出, $F(x)$ 表示 x 经过残差块后的输出。恶意样本图像经过卷积池化后得到浅层特征图, 经过残差块处理后, 随着网络的不断优化, 进而获取恶意样本的深层特征图。若输入和输出的特征图维度相同, 则残差块的计算公式如 (5) 所示:

$$y_{\text{output}} = F(x, \{W_i\}) + x \quad (5)$$

若输入和输出的特征图维度不同, 则需要在 x 的基础上添加一个线性映射 W_s , 其计算公式如 (6) 所示:

$$y_{\text{output}} = F(x, \{W_i\}) + W_s x \quad (6)$$

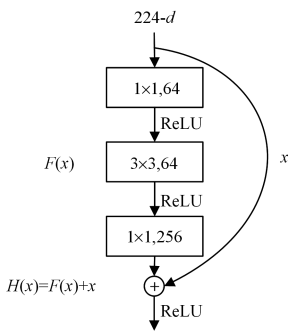


图 3 残差块

Fig. 3 Residual block

每个残差块使用 Conv-BN-ReLU 结构的卷积神经网络来提取恶意代码可视化后的图像特征。在残差块中采用 Batch Normalization (BN) 方法使恶意代码二进制文件转化的 RGB 图像数据满足均值为 0、方差为 1 的分布规律, 从而加速模型收敛, 提高模型精度。

3.3.2 通道域注意力机制

在恶意代码家族检测模型中, 将恶意代码二进制文件可视化为特征图, 每一张特征图的初始状态通过 RGB 三通道表示, 经过卷积神经网络中不同的卷积核之后, 每一个通道会产生新的特征。例如, 若使用 M 个卷积核, 特征图经过卷积操作后会生成 M 个新通道的矩阵 (H, W, M), H 和 W 分别表

示特征图的高度和宽度。每个通道的特征可以表示为该特征图在不同卷积核上的分量, 这些卷积操作类似于傅里叶变换, 将各通道的特征信息分解成 M 个卷积核上的信号分量, 给每个通道上的信号赋予不同的权重以表示该通道与关键特征的相关度, 再根据权重分布使神经网络注意某些通道, 从而提取恶意代码家族的关键特征。本文采用通道域注意力机制使教师网络模型关注通道之间的关系, 自主学习不同通道的特征, 为输入不同的特征图分配不同的权重, 从而让模型更加关注信息量最大的通道特征, 进而抑制那些不重要的通道特征, 提取更关键、更重要的深层特征, 使恶意代码家族检测模型做出更准确的判断。

在教师网络模型中加入 Squeeze-and-Excitation (SE) Block^[20] (挤压与激励模块)。首先对通过卷积池化的特征图进行 Squeeze 操作, 在通道维度上得到具有全局分布的特征; 然后对特征通道进行 Excitation 操作, 该操作采用轻量级的门控机制, 学习各个通道间的关联性, 得到不同通道的权重; 最后通过 Scale 操作得到最终的特征图, 具体如图 4 所示。

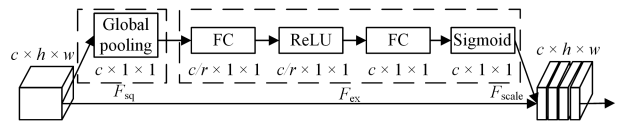


图 4 挤压与激励模块

Fig. 4 Squeeze-and-Excitation (SE) block

Squeeze 操作是对各类恶意样本的 RGB 三维图像进行全局平均池化 (Global Average Pooling) 操作, 即在空间维度上对特征图进行压缩, 将每个二维特征压缩为一维向量, 在某种程度上该向量具备全局的感受野, 并且输出的维度与输入特征图的通道数相同。其计算式如 (7) 所示:

$$z_c = F_{sq}(u_c) = \frac{1}{W \times H} \sum_{i=1}^W \sum_{j=1}^H u_c(i, j) \quad (7)$$

Excitation 操作是将 Squeeze 操作生成的 $c \times 1 \times 1$ 大小的特征图经过全连接神经网络和门控机制, 为每个特征通道生成权重 w , 通过权重分布来表示特征通道间的相关性, 从而使神经网络学习到不同通道的重要性。其计算式如式 (8) 所示:

$$s = F_{ex}(z, W) = \sigma(g(z, W)) \\ = \sigma(W_2 \delta(W_1 z)), W_1 \in R_r^{C \times C}, W_2 \in R_r^{C \times C} \quad (8)$$

其中, δ 为 ReLU 激活函数, σ 为 Sigmoid 激活函数, r 为降维系数 (Reduction Ratio)。

Scale 操作是将 Excitation 操作输出的权重, 采用乘法的方式对原特征图进行逐通道加权, 在通道维度上实现对原特征图的特征增强, 得到具有通道重要性的新特征图, 并将其作为下一层的输入特征。其计算式如 (9) 所示:

$$x_c = F_{scale}(u_c, s_c) = s_c \cdot u_c \quad (9)$$

残差网络与注意力机制结合组成的教师网络模型结构如图 5 所示。

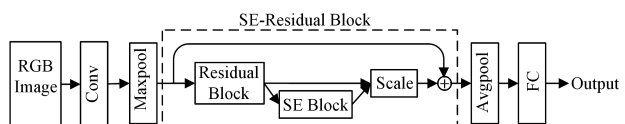


图 5 SE-ResNet 模型

Fig. 5 SE-ResNet model

4 实验结果与分析

4.1 数据集

本文数据集来自 Microsoft 恶意软件分类挑战赛^[21],其中包含 9 类恶意代码家族,每个恶意样本均由 IDA Pro 反汇编后生成的汇编代码和二进制字节码文件组成。每个字节码文件的原始数据均由十六进制表示,不包含 PE 文件头。为了研究二进制字节码特征对恶意代码家族检测性能的影响,本文选取各类恶意样本由反汇编工具生成的 10 868 个“bytes”文件进行实验,恶意代码家族数据集的分布情况如表 1 所列。

表 1 恶意代码家族数据集的分布情况

Table 1 Distribution of malicious code family datasets

恶意代码家族	样本总数
Ramnit	1 541
Lollipop	2 478
Kelihos_ver3	2 942
Vundo	475
Simda	42
Tracur	751
Kelihos_ver1	398
Obfuscator. ACY	1 228
Gatak	1 013

4.2 评价指标

由于恶意代码家族检测属于多分类问题,为了更好地衡量模型对恶意代码家族检测的效果,选取宏精确率(Macro-Precision)、宏召回率(Macro-Recall)、宏 F1-score、模型准确率(Model-Acc)作为衡量模型性能的标准。4 个评价指标的计算式如下:

$$P_{\text{macro}} = \frac{1}{n} \sum_{i=1}^n \frac{TP_i}{TP_i + FP_i}$$
 (10)

$$R_{\text{macro}} = \frac{1}{n} \sum_{i=1}^n \frac{TP_i}{TP_i + FN_i}$$
 (11)

$$F_{\text{macro}} = \frac{2 * P_{\text{macro}} * R_{\text{macro}}}{P_{\text{macro}} + R_{\text{macro}}}$$
 (12)

$$Acc = \frac{1}{N} \sum_{i=1}^n TP_i$$
 (13)

其中,TP 为将正类样本预测为正类的样本数;FN 为将正类样本预测为负类的样本数;FP 为将负类样本预测为正类的样本数;TN 为将负类样本预测为负类的样本数;N 为总样本数。

4.3 实验结果与分析

本实验将二进制字节码可视化后的各类恶意代码家族图像样本按照 8:1:1 的比例随机分为训练集、验证集和测试集。实验使用 PyTorch 的 transforms 方法对图像进行预处理,将图像统一进行标准化、归一化处理后,使用实验数据集对基于知识蒸馏的恶意代码家族检测模型进行训练和测试。

4.3.1 教师网络模型分析

教师网络模型采用基于通道域注意力机制的残差网络模型,其降维系数 *r*、学习率、损失函数等参数的设置均会影响恶意代码家族检测的效果。为了防止模型过拟合并便于调整超参数,选取验证集对模型进行优化。模型选取交叉熵损失函数来度量真实值与预测值之间的差异,同时使用随机梯度下降方法,采用等间隔学习率调整策略对网络进行优化。本实验通过调整通道域注意力模块中的超参数降维系数 *r*,测试其对教师网络模型性能的影响,结果如表 2 所列。

表 2 降维系数 *r* 对分类结果的影响

Table 2 Influence of reduction ratio on classification results

Reduction Ratio	Model-Acc	Params
2	0.985 3	43 646 025
4	0.982 6	33 586 249
8	0.983 5	28 556 361
16	0.986 3	26 041 417
32	0.989 9	24 783 945
64	0.984 4	24 155 209

经上述实验,通过权衡模型的准确率和复杂度,将教师网络模型的降维系数 *r* 设置为 32。为了进一步验证教师网络模型的准确率和泛化能力,选取 AlexNet, VGGnet, InceptionV2, ResNet50 这 4 种神经网络模型在测试集上进行对比实验。在训练过程中,Epoch 设置为 64, Batch Size 设置为 32,保证每个模型在相同的参数下进行训练和测试。采用模型准确率、宏召回率、宏精确率和宏 F1-score 作为模型性能评价的指标,实验结果如表 3 所列。

表 3 模型测试结果的对比

Table 3 Comparison of model test results

Teacher Model	Model-Acc	Macro-Recall	Macro-Precision	Macro-F1-score
AlexNet	0.963 4	0.918 4	0.936 3	0.926 4
VGG	0.970 8	0.961 6	0.959 7	0.960 6
InceptionV2	0.961 5	0.920 2	0.905 3	0.912 0
ResNet50	0.975 3	0.977 5	0.959 5	0.975 4
SE-ResNet50	0.989 9	0.992 6	0.975 2	0.983 1

实验结果表明,经过迭代训练后,浅层神经网络可以取得较好的效果,随着网络的加深,模型的分类效果具有一定的提升;残差网络的模型准确率、宏召回率、宏精确率和宏 F1-score 均优于 AlexNet, Vgg, InceptionV2 这 3 种神经网络模型,说明残差网络能够降低数据中的冗余度,提取的恶意代码家族图像的深层纹理特征更具鲁棒性;在残差网络中加入 SE 模块的神经网络,通过训练后得到的分类器准确率达到了 98.99%,其对所有恶意家族的分类效果如图 6 所示(图中的结果均为四舍五入后的结果)。相比传统神经网络,SE-ResNet50 网络聚焦于不同通道,在通道中自主学习得到各类恶意代码家族的关键纹理特征,能够提取更高级的语义特征,从而提高了模型的分类性能。因此,本文选取 SE-ResNet50 作为知识蒸馏中的教师网络模型,可以更好地对学生网络模型进行指导训练,提高模型的检测效果。

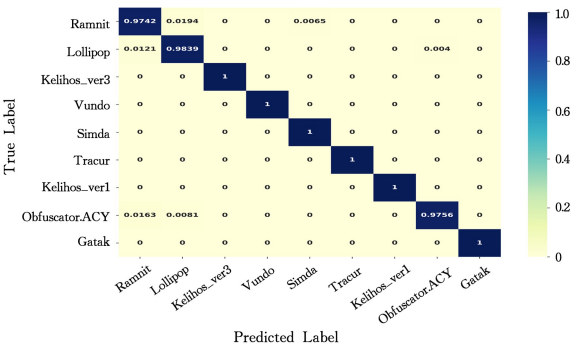


图 6 教师网络模型混淆矩阵

Fig. 6 Teacher network model confusion matrix

4.3.2 知识蒸馏训练结果分析

本实验选取 LeNet 与 ResNet18 作为学生网络模型,对比蒸馏前后的模型对恶意代码家族的检测效果。LeNet 是一个简单的卷积神经网络,由 2 个卷积层、2 个池化层和 3 个全连接层组成,为了使网络适应特征图输入,对全连接层进行了扩充。ResNet18 由残差网络组成,相比 ResNet50 网络深度更浅,由 17 个卷积层和 1 个全连接层组成,其残差块结构组成为[2,2,2,2]。由 SE-ResNet50 作为教师网络对学生网络进行训练,教师网络与蒸馏前后的学生网络的对比实验结果如表 4 所列。可知,学生网络的参数量远少于教师网络,其中 ResNet18 的参数量相比教师网络减少了约 55%。在减少复杂度的基础上,LeNet和ResNet18学生网络通过知识蒸馏

后,模型的准确率分别提升了 3.12%和 1.93%。因此,知识蒸馏方法可以使小型网络学习到复杂网络的特征知识,但对于简单的神经网络因其学习能力较弱,无法学习到更深层次的特征知识。不同模型对各类恶意代码家族的分类准确率、精确率、召回率和 F1-score 如图 7 所示。

表 4 知识蒸馏对比实验结果

Table 4 Results of knowledge distillation comparative experiment

	Model	Model-Acc	Params/万
教师网络	SE-ResNet50	0.9899	2478
	LeNet	0.9313	541
学生网络	ResNet18	0.9679	1118
	D-LeNet	0.9625	541
蒸馏后的学生网络	D-ResNet18	0.9872	1118

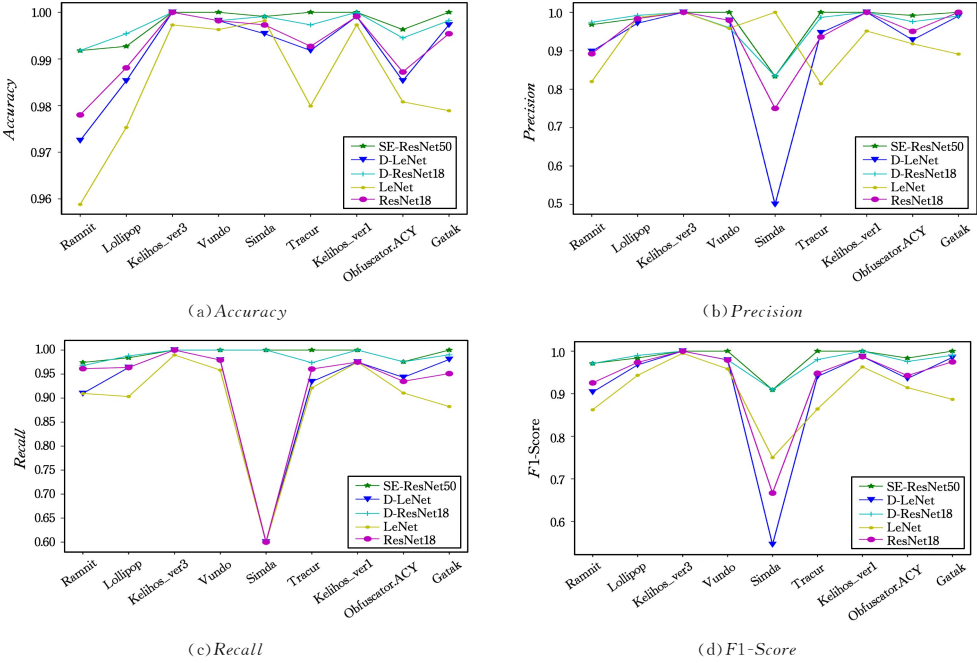


图 7 各恶意代码家族的分类评价指标

Fig. 7 Classification evaluation indexes of each malicious code family

从上述各类指标分布中可以看出,Simda 家族的精确率、召回率和 F1-score 偏低,其原因是该家族的样本量少,神经网络无法完全表征该类恶意家族的特征。本文采用的注意力机制可以在样本不足的情况下获取关键纹理特征,提升对该类样本的检测效果。如图 8、图 9 所示,相比原始学生网络,蒸馏后的学生网络对该类恶意家族达到了较好的分类准确率。

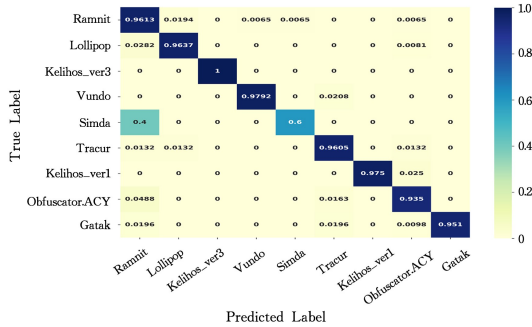


图 8 蒸馏前学生网络模型混淆矩阵

Fig. 8 Student network model confusion matrix before distillation

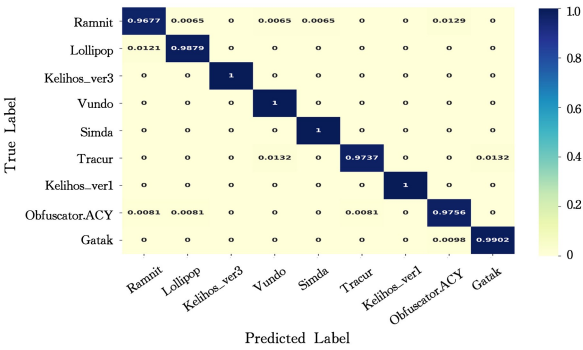


图 9 蒸馏后学生网络模型混淆矩阵

Fig. 9 Student network model confusion matrix after distillation

对比蒸馏前学生网络的混淆矩阵和教师网络模型的混淆矩阵,原始学生网络对某些恶意家族的分类效果略差,但在整体上蒸馏后的学生网络在恶意代码家族检测任务中达到了与教师网络相近的检测效果。

综上,学生网络虽然无法超越教师网络的检测效果,但在教师网络的指导训练下,学生网络的性能得到了一定的提升,

接近教师网络的检测准确率。同时,学生网络模型的参数数量和计算量小,网络结构相对简单,在恶意代码家族检测任务中执行速率更快,有利于对批量样本的检测和移动端的部署。

结束语 本文分析了现有的恶意代码家族检测的不足,采用知识蒸馏的思想,提出了一种基于知识蒸馏的恶意代码家族检测方法。本文在卷积神经网络的基础上加入残差模块,解决了由于网络深度引起的网络退化问题,同时探索了计算机视觉领域中注意力机制对恶意代码家族分类模型性能的影响,发现基于通道域注意力机制可以更有效地提取恶意代码二进制文件转化的 RGB 图像纹理特征之间的差异,可以在一定程度上提升原有模型的效果。因此,本文将基于通道域注意力机制的残差网络模型作为教师网络,以训练结构简单的学生网络。实验结果表明,该方法在减小模型复杂度和减少参数数量的同时,使学生网络达到了与教师网络相近的检测效果,可以有效地解决大批量恶意样本检测任务中资源消耗的问题。

后续工作将结合空间域、时间域、混合域等注意力机制,进一步研究注意力机制对恶意代码可视化后特征提取的影响,构建性能更好的教师网络模型,同时对知识蒸馏网络结构进行优化,以减小网络复杂度,提高模型的检测效率和准确率。

参 考 文 献

[1] CHEN J J, PENG B Z, WU P Z. Method for detecting malicious code based on dynamic behavior and machine learning[J/OL]. Computer Engineering. [2020-06-06]. <https://doi.org/10.19678/j.issn.1000-3428.0056409>.

[2] ZHAO C R, ZHANG W J, FANG Y, et al. Malware detection based on semantic API dependency graph[J]. Journal of Sichuan University(Natural Science Edition), 2020, 57(3): 488-494.

[3] MOHANASRUTHI V, CHAKRABORTY A, THANUDAS B, et al. An Efficient Malware Detection Technique using Complex Network-based Approach[C] // 2020 National Conference on Communications (NCC), 2020.

[4] NARAYANAN B N, DAVULURU V S P. Ensemble Malware Classification System using Deep Neural Networks[J]. Electronics, 2020, 9(5): 721.

[5] HU J W, CHE X, ZHOU M, et al. Incremental clustering method based on Gaussian mixture model to identify malware family[J]. Journal on Communications, 2019, 40(6): 148-159.

[6] ZENG Y Q, ZHANG L L, ZHANG R N, et al. Malware Family Classification Model Based on MobileNet[J]. Computer Engineering, 2020, 46(4): 162-168.

[7] SUN B W, ZHANG P, CHENG M Y, et al. Malware detection method based on enhanced code images[J]. Journal of Tsinghua University(Science and Technology), 2020, 60(5): 386-392.

[8] VASAN D, ALAZAB M, WASSAN S, et al. Image-Based Malware Classification using Ensemble of CNN Architectures(IM-CEC)[J]. Computers & Security, 2020, 92: 101748.

[9] GHOUTI L. Malware Classification Using Compact Image Features and Multiclass Support Vector Machines[J]. IET Informa-

tion Security, 2020, 14(4): 419-429.

[10] JAIN M, ANDREPOULOS W, STAMP M. Convolutional neural networks and extreme learning machines for malware classification[J]. Journal of Computer Virology and Hacking Techniques, 2020, 16(3): 229-244.

[11] REN Z J, CHEN G, LU W K. Malware visualization methods based on deep convolution neural networks[J]. Multimedia Tools and Applications, 2020, 79(3): 1-19.

[12] COHEN A, NISSIM N, ELOVICI Y. MalJPEG: Machine Learning Based Solution for the Detection of Malicious JPEG Images[J]. IEEE Access, 2020, 8: 19997-20011.

[13] AZAB A, KHASAWNEH M. MSIC: Malware Spectrogram Image Classification[J]. IEEE Access, 2020, 8: 102007-102021.

[14] CHEN J, JIA X, ZHAO C, et al. Using the Rgb Image of Machine Code to Classify the Malware[C] // 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA). IEEE, 2020.

[15] HINTON G, VINYALS O, DEAN J. Distilling the Knowledge in a Neural Network[J]. Computer Ence, 2015, 14(7): 38-39.

[16] FURLANELLO T, LIPTON Z C, TSCHANNEN M, et al. Born again neural networks[C] // International Conference on Machine Learning, 2018: 1607-1616.

[17] GAO M Y, SHEN Y J, LI Q Q, et al. Residual knowledge distillation [EB/OL]. (2020-02-21) [2020-07-04]. <https://arxiv.org/pdf/2002.09168.pdf>.

[18] 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, USA: ACM Press, 2011: 1-7.

[19] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[20] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-Excitation Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.

[21] RONEN R, RADU M, FEUERSTEIN C, et al. Microsoft Malware Classification Challenge[EB/OL]. [2018-02-22]. <https://arxiv.org/pdf/1802.10135.pdf>.



WANG Run-zheng, born in 1996, post-graduate, is a member of China Computer Federation. His main research interests include cyber security and malware.



GAO Jian, born in 1982, Ph.D. His main research interests include cyber security, malware and botnet.