

Research on Malware Detection and Terminal Equipment Security in Communication Networks

GU Wei *

(School of Electronics & Information Engineering, Nanjing University of Information Science & Technology,
Nanjing Jiangsu 210044, China)

Abstract: The rapid development of wireless communication technology promotes the prosperity of mobile terminal market. Monitoring the explosive growth of network traffic has become very difficult, resulting in a series of security risks in the mobile terminal market. In view of its popularity and convenience, Android has become the most popular operating platform with the biggest market share of mobile terminals. However, becoming the primary target of malicious attackers stems from its huge market share. Considering the surge in the number of malware and the upgrading of camouflage technology, the existing research focusing on the analysis of single type features such as permissions is not enough to cope with the current development trend. Therefore a hybrid malware detection model called PSDroid is proposed based on the combination of permissions and related service component features. To evaluate the performance of the proposed methods, 864 benign applications and 1938 malicious ones are collected from Google play and VirusShare, respectively. The experimental results show that the accuracy of the proposed model is 95.65%, which outperforms the result of single permission feature analysis. In addition, compared with the existing models based on deep learning or machine learning, the proposed method still has significant advantages for malware detection in communication networks.

Key words: communication networks; malware detection; deep learning

EEACC: 6140 doi: 10.3969/j.issn.1005-9490.2023.01.006

通信网络下恶意检测及终端设备安全研究

顾 伟 *

(南京信息工程大学电子与信息工程学院, 江苏 南京 210044)

摘 要: 无线通信技术的快速发展, 促使了移动终端市场的繁荣。爆炸式增长的网络流量使得监控变得十分棘手, 导致移动终端市场存在一系列安全隐患。安卓因其自身的流行性和便利性, 已经成为手机终端市场占有率第一的操作平台。然而, 庞大的市场份额也使得其成为恶意攻击者的首要攻击目标。针对通信网络中恶意软件的数量暴增以及伪装技术的升级, 现有的多采用单一的特征如权限 (Permission) 进行分析的研究工作不足以应付当今的发展趋势, 因此提出了基于权限 (Permission) 和服务 (Service) 特征相结合的恶意软件检测模型 PSDroid。为了评估所提方法的性能, 分别从 Google Play 和 VirusShare 上收集了 864 个良性应用和 1938 个恶意应用。实验结果表明, 所提出的模型准确率达到 95.65%, 优于对单一 Permission 特征分析的结果。此外, 与当前现有的基于深度学习或机器学习的模型相比, 针对通信网络下恶意软件检测, 所提方法仍然存在显著优势。

关键词: 通信网络; 恶意软件检测; 深度学习

中图分类号: TN918; TP309.1

文献标识码: A

文章编号: 1005-9490(2023)01-0036-05

通信技术, 尤其是无线通信以惊人的速度迅猛发展, 在移动网领域被广泛使用, 为用户提供高速率、低延迟、低成本以及高可靠度的通信服务^[1]。第三代通信技术 (3G) 由 WCDMA, CMDA 2000 和 TD-SCDMA 构成, 可以为用户提供宽带业务, 对用户的通信方式产生革命性的变化。第四代通信技术 (4G) 指的是一种超高速的无线通信方式, 它的出现与流行弥补了传统通信技术的不足, 降低了资费, 为移动市场的繁荣贡献了巨大的力量。对无线通信系

统的超高容量和高可靠性的需求促进了第五代通信技术 (5G) 的出现^[2], 5G 再一次改变了人们的生活方式, 使得用户拥于更好的上网体验。

无线通信技术的高传输速率和低传输时延促进了手机终端市场的繁荣。根据国际数据公司 IDC 2021 年的统计数据, 安卓平台在移动市场上占比达 83.8%。庞大的市场份额和平台的开放性^[3], 使得针对此平台的应用数量和种类也呈现爆发式增长。受益于成熟通信技术的保障, 安卓为用户提供社交

网络、财务管理、娱乐等多类型的应用^[3]。手机智能化程度进一步加深,同时安全问题也进一步突出。根据 360 公司发布的《2020 年中国手机安全状况报告》显示,360 安全大脑 2020 年成功截获的新增恶意样本数量高达 454.6 万个,严重威胁了用户的经济和隐私安全。常见的恶意行为,例如恶意收费和隐私窃取,通常会损害用户的利益^[4]。安卓的开放性,允许用户从第三方的应用市场(如应用宝,APK-pure 等)下载应用,这为用户下载到心仪的应用提供了诸多便利,但因缺乏监管机制,这无疑为恶意攻击者向毫无戒心的用户传播恶意软件提供了便利。快速精准地在通信网络中检测出恶意软件以及保障终端设备用户的安全,对于研究人员来说是一个迫切且充满挑战的课题。

现有的通信网络下恶意检测工作大致可以分为两类:静态检测^[5]和动态检测^[6]。动态分析方法受限于执行环境的部署,需要在运行时才可以进行感染分析。随着恶意软件数量的规模迅速扩张,静态分析方法因具有无需执行应用程序便可执行分析的特点,从而更受研究人员欢迎。静态分析方法是一种低成本的分析方法,因此,静态分析方法被选用进行后续分析。

静态分析主要是围绕 Manifest.xml 文件进行分析,此文件包含 Permission, Providers, Services 和 API 等特征。常见的反编译工具如 Androguard 可以对每个 Android 应用程序包进行反编译获取 Manifest.xml 文件。Permission 保护用户的隐私信息,APP 访问敏感用户数据(如联系人)或某些系统功能(如相机)时,必须请求用户授予对应的权限,因此研究人员针对 Permission 在恶意软件识别中发挥的作用,围绕 Permission 特征展开研究。现有的关于 Permission 的研究如文献[7]提出的基于 Permission 特征的模型叫做 KNN-P,即使用 KNN 作为分类器,且当 $k=2$ 时模型的识别性能最佳。现有工作已经成功证明 Permission 特征在恶意软件识别方面的作用,因此我们的后续研究也是围绕 Permission 特征而展开。Service 是一个在后台运行的组件,可用于网络传输,执行长时间运行的操作或者远程通信的操作。Service 可与其他程序组件绑定,利用服务进行通信和交互,甚至在进程间通信。考虑到不同类型的特征可以从不同方面反应通信网络下样本的特性,因此我们选择将 Permission 和 Service 二者进行组合分析,更全面地反映出样本特性,从而更精准地实现通信网络下的恶意检测以及保障移动设备安全。

在恶意软件识别任务中,与特征选择同样重要

的一环是特征学习。特征学习^[8]指的是允许机器从原始数据探索和学习合适的特征变换,是模型自动学习的过程。近年来,将深度学习与特征学习相结合在自然语言处理^[9]、语言识别^[10]、图像识别^[11]等领域取得了优异的成绩。卷积神经网络(Convolutional Neural Network, CNN)作为深度学习的一个重要分支,通过融合各层局部感受野的空间和通道信息来构建信息特征,近几年取得了不错的成绩。从此角度出发,我们选用经典的卷积神经网络 SENet(此网络在 2017 年度的 ILSVRC 比赛中获得了第一名)来对 Permission 和 Service 组合特征进行学习。基于此,提出我们的恶意软件检测模型 PSDroid,据我们的了解,现有的研究中还没有将 Permission 和 Service 进行组合研究的工作。为了评估 PSDroid 的性能,我们基于相同和严谨的实验条件,进行了广泛的实验并与当下现有的模型进行对比。实验结果表明,所提模型具有良好的识别性能。

本文的主要贡献有以下几点:

①我们提出了一个用于通信网络下检测恶意软件的新模型 PSDroid,该模型使用 SENet 网络来对通信网络中恶意软件特性进行全面剖析。

②我们首次关注组合特征 Permission 和 Service 的内在联系,利用组合特征反应样本特性。

③我们收集了来自 VirusShare 的 1938 个恶意样本和来自 Google Play 的 864 个样本。在收集的数据集上我们进行了广泛的实验,以验证所提模型 PSDroid 的性能。实验结果表明,我们的模型性能优于当前现有的方法。

1 相关工作

随着无线通信技术的成熟与完善,恶意软件数量呈现爆炸式增长。如今,智能手机不仅是通讯工具,也是我们个人的隐私图书馆。因此,手机的安全问题,需要得到足够的重视。在本节中,我们将讨论以前解决的和遗留的检测问题。

1.1 动态分析

动态分析方法利用在受控环境中执行应用程序时可以监控的语义特征。TaintDroid^[6]执行应用程序的数据流分析,并检测敏感数据的信息泄漏。文献[12]提出了一种新颖的异常检测技术 SMSBotHunter,利用文本和行为特征来检测短信僵尸网络。动态分析方法受限于执行环境的部署,不能直接检测和分析恶意软件。此外,我们的目标是检测出恶意软件,动态分析可能效率低下,无法应对恶意软件的快速发展。因此,后续分析侧重于从静态分析中提取的特征。

1.2 静态分析

静态分析方法可以在不执行应用程序的情况下提取特征。这种方法具有计算量小、对执行环境要求低等优点。文献[13]提出的基于 Permission 特征验证四种经典的机器学习方法,即随机森林、支持向量机、高斯朴素贝叶斯和 K -均值,在识别恶意软件方面的性能。He 等人^[14]提出了一种轻量级机制,该方法关注跨应用程序的数据流,并且只有当敏感应用程序接口通过组件间通信(Inter-Component Communication, ICC)被其他

应用程序隐式使用时,才通知用户授予 Permission。随着恶意软件伪装技术的升级,一些学者已经考虑到选用多种类型的特征来更好地反应样本特性。Malpat^[5]方法就是一个很好的例子,此方法通过研究 Permission 和 API 之间的关系。但是据我们了解,目前关于 Permission 特性和 Service 之间的关系或组合研究工作仍较少。

2 恶意软件检测研究

恶意检测模型 PSDroid 整体结构如图 1 所示。

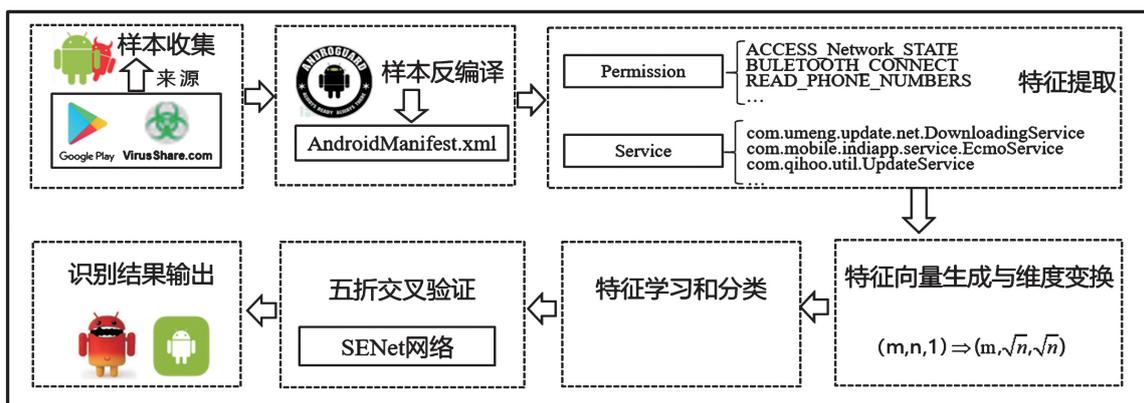


图 1 PSDroid 整体框架图

2.1 特征提取

Android 的 APK 文件主要包含清单、代码(即 dex)、所需库和资源。Androguard 是一个跨平台的反向工具,可用于反编译 Android 应用程序并实现对 Android 应用程序的静态分析。对于良性和恶意的应用程序构成的原始数据集,开源工具 Androguard 可用来批量提取相关特征。

①Permission 特性:

根据保护级别,Permission 特征可分为以下四类:正常,危险,签名和签名或系统。Permission 的保护级别默认值为正常,在安装期间系统自动向应用程序授予权限。通常,用户可能会注意的一些操作权限,如发送短信、访问通讯簿等,会被标记为危险级别。签名级别的 Permission 应与声明此 Permission 的应用程序签名一致才可访问对应资源。对于签名或系统级 Permission,无需用户过多关注,故我们也不展开研究。因此,管理 Permission 的授予以及定义合适的保护级别就显得尤为重要。本文的 Permission 特征共提取 155 个,因此每个应用程序可用一个 155 维的向量 [permission] 1×155 表示,每个维度对应一个 Permission。若此 Permission 在样本对应的 Manifest.xml 文件中存在,赋值 1,否则 0。

②Service 特性:

在 Android 应用程序中有四个基本的组件:

Activity、Service、Receive 和 Provider。每个组件都发挥着不同的作用,其中 Service 负责在不同的线程中进行后台处理操作。Service 组件在 Manifest.xml 文件中对应的标签名为 <service>,因此本文从 Manifest.xml 文件中提出所有包含 <service> 的字段中的 name 字段。

Service 特征共提取 7 107 个,为了使得提取特征关联性更高,我们将仅在样本中出现一次的字段删除,保留了 1 412 个以供后续分析使用,每个应用程序可用一个 1 412 维的向量 [service] $1 \times 1 412$ 表示,每个维度对应一个 service。若此 service 在样本对应的 Manifest.xml 文件中存在,赋值 1,否则 0。

③混合特征:

将 Permission 和 Service 二者进行组合分析。每款应用都可以用一组 [permission] 1×155 和 [service] $1 \times 1 412$ 表示,将两组向量结合,每个 app 都可以用一个 [permission, service] $1 \times 1 567$ 表示。

2.2 特征学习和分类

卷积神经网络在图像识别、图像分割和自然语言处理等领域取得了成功。由 Hu 等人提出的 SENet^[15]在 2017 年 ILSVRC 分类竞赛中获得第一名。因此,我们选用 SENet 来进行原始特征学习,使模型挖掘出更丰富的特征间关系,从而实现良好的表征。SENet 的核心是 SE-block,此结构由 squeeze

和 excitation 两部分组成。Squeeze 操作是采用全局平均池化操作将 channel 上整个空间特征转换成全局特征,从而得到全局描述特征。Excitation 操作主要负责获取 channel 之间的关系。

为了使得提取的特征适应卷积神经网络的输入,我们提出了一种自适应维度变换。将原始输入维度 $(m, n, 1)$ 转成 (m, \sqrt{n}, \sqrt{n}) ,其中 m 表示样本的数量, n 表示特征的数量。例如,通过填充 32 列 0 列的方式将 Service 维度转换成 38×38 。一般来说,确保变换后的维数是整数,常见的方式是增加 0 列或者删去多余维度。此处采用末尾填充 0 列的方式,从而避免因维度减少而造成关键信息丢失。

3 实验和评估

3.1 数据集收集

现有的公开数据集如 AMD^[16],可以作为实验数据的来源。然而,通信技术发展迅速,此类数据存在样本过旧且样本未保持更新等问题,无法较好反应当今恶意软件的发展态势,因此,选择直接从市场获取样本来构造数据集。所有的恶意样本来自 VirusShare 这个网站,良性样本来自 Google Play Store。来自 Google Play Store 的样本,利用 VirusTool (提供可疑文件分析的网站)进行检测,顺利通过检测的样本,被标记为良性样本。VirusTool 是权威的恶意检测网站,其包含了 360,腾讯,金山等多达 50 款检测软件。经过筛选,最终收集了 864 个良性应用和 1 938 个恶意应用构成实验数据集。

3.2 评估指标

在这一部分中,我们系统地介绍了实验所采用的评价指标。在我们的评估方法中,恶意应用程序是正样本,而良性应用程序是负样本。常见的评估指标,如准确率(Acc)、精准率(Pre)、召回率(Rec)、F1-score(F1)、假阳性率(Fpr)和反曲率(Auc)都逐一被介绍。上述评估指标的详细信息如下。

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}}$$

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Fpr} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

其中,TP 表示正确预测的恶意样本数量,FP 表

示错误预测的良性样本数量,TN 表示正确预测的良性样本数量,FN 表示错误预测的恶意样本数量。

3.3 性能评估

在相同的参数设置下,我们采用不同的输入特征进行了三组实验,以验证所提模型的有效性。为了使得实验结果更具说服力,所有实验都采用了五折交叉验证。特征学习选用经典的卷积神经网络 SENet,因其具有较强的学习能力,能够更好地学习复杂的特征关系。为了避免过拟合现象,此处引入 Dropout,并赋值为 0.1。优化器选用 AdamOptimizer,激活函数选用 ReLU。学习率设置为 0.001, batch size 赋值为 200 以及 epoch 赋值为 100。

实验结果如表 1 所示,我们提出的模型,利用混合特征作为输入,取得了比单一特征更好的表现,相比于研究人员高频分析特征 Permission 来说,准确率提升达 2.10, Auc 提升达 2.02。

表 1 不同输入特征下 SENet 学习特征

特征	Acc	Pre	Rec	F1	Fpr	Auc
Permission	93.55	94.58	96.19	95.37	12.38	92.17
Service	94.01	93.90	97.69	95.75	14.19	91.64
PSDroid	95.65	96.06	97.73	96.88	9.04	94.19

为了进一步验证提出模型的性能,与当下最前沿的相关研究工作对比,比较结果如表 2 所示。

表 2 PSDroid 与相关研究对比

模型	Acc	Pre	Rec	F1	Fpr	Auc
LeNet-P	93.58	94.79	96.02	95.39	11.99	92.18
KNN-P	91.09	96.39	90.49	93.35	7.63	91.45
PSDroid	95.65	96.06	97.73	96.88	9.04	94.19

对比试验的所有参数设置严格按照原文所提及的参数并在我们收集到的数据集上进行配置。采用我们的数据集,而不是公开数据集,主要原因是考虑到数据集的原始样本未公开,难以进行后续提取和分析工作。

文献[17]基于卷积神经网络的安卓恶意软件检测框架被定义为 LeNet-P。我们收集了 155 个 Permission 特征,而不是研究中描述的 138 个,因为我们的数据集中收集的应用程序更加新颖。与 LeNet-P 相比,我们的模型 PSDroid 在准确率(Acc)、精准率(Pre)、召回率(Rec)、F1-score(F1)、假阳性率(Fpr)和反曲率(Auc)上提升分别为 2.07, 1.27, 1.71, 1.49, 2.95 和 2.01。

文献[7]基于 KNN 的恶意检测框架被定义为 KNN-P,实验结果表明,我们提出的模型 PSDroid 整

体效果优于 KNN-P。与现有的相关的研究工作的对比进一步证明了我们所提模型的有效性。

4 结束语

本文提出了一种新的通信网络下的恶意软件检测框架 PSDroid。基于 Manifest.xml 文件中提取出的强关联性特征 Permission 和 Service, 我们利用经典的 SENet 网络来自适应探索特征之间的内在联系, 从而实现更好的样本表征。实验结果表明, 我们提出的框架 PSDroid, 利用多种类型的强关联特征可以取得比单一类型特征更好的识别效果。与当前现有的研究相比, 我们的方案仍然存在一定的优势, 这也进一步证明了所提模型的有效性。下一步工作, 我们将致力于挖掘更多通信网络下应用的特征间关联关系。

参考文献:

- [1] Wang J, Jiang C, Zhang H, et al. Thirty Years of Machine Learning: the Road to Pareto-Optimal Wireless Networks [J]. IEEE Communications Surveys & Tutorials, 2020, 22(3): 1472-1514.
- [2] 桂冠, 王禹, 黄浩. 基于深度学习的物理层无线通信技术: 机遇与挑战[J]. 通信学报, 2019, 40(2): 19-23.
- [3] Li Q, Hu Q, Qi Y, et al. Semi-Supervised Two-Phase Familial Analysis of Android Malware with Normalized Graph Embedding[J]. Knowledge-Based Systems, 2021, 218(3): 106802.
- [4] Kim T G, Kang B J, Rho M, et al. A Multimodal Deep Learning Method for Android Malware Detection Using Various Features [J]. IEEE Transactions on Information Forensics and Security, 2019, 14(3): 773-788.
- [5] Tao G H, Zheng Z B, Guo Z Y, et al. MalPat: Mining Patterns of Malicious and Benign Android Apps via Permission-Related APIs [J]. IEEE Transactions on Reliability, 2018, 67(1): 355-369.
- [6] Enck W, Gilbert P, Han S Y, et al. Taintdroid: An Information-Flow Tracking System for Realtime Privacy Monitoring on Smartphones [J]. Communications of the ACM, 2014, 32(2): 5.
- [7] Arslan R S, Dogru I A, N B. Permission-Based Malware Detection System for Android Using Machine Learning Techniques [J]. Journal of Software Engineering and Knowledge Engineering, 2019, 29(1): 43-61.
- [8] Yann L C, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521(7553): 436-444.
- [9] Qi M, Qin J, Yang Y, et al. Semantics-Aware Spatial-Temporal Binaries for Cross-Modal Video Retrieval [J]. IEEE Transactions on Image Processing, 2021, 30: 2989-3004.
- [10] Arul V H, Marimuthu R, Sivakumar V G, et al. Taylor-DBN: A New Framework for Speech Recognition Systems [J]. International Journal of Wavelets, Multiresolution and Information Processing, 2021, 19(2): 2050071.
- [11] Ohri K, Kumar M. Review on Self-Supervised Image Recognition Using Deep Neural Networks [J]. Knowledge-Based Systems, 2021, 224(8): 107090.
- [12] Faghihi F, Abadi M, Tajoddin A. SMSBotHunter: A Novel Anomaly Detection Technique to Detect SMS Botnets [C]//2018 15th International ISC (Iranian Society of Cryptology) Conference on Information Security and Cryptology (ISCISC), Tehran, Iran, 2018: 1-6.
- [13] McDonald J, Herron N, Glisson W, et al. Machine Learning-Based Android Malware Detection Using Manifest Permissions [C]//54th Hawaii International Conference on System Sciences, Kauai, HI, USA, 2021: HICSS.2021.839.
- [14] He Y, Qi L. Detecting and Defending Against Inter-APP Permission Leaks in Android Apps [C]//2016 IEEE 35th International Performance Computing and Communications Conference (IPCCC), Las Vegas, NV, USA, IEEE, 2016: 1-7.
- [15] Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018: 7132-7141.
- [16] Tchakounté F, Djakene Wandala A, Tiguiane Y. Detection of Android Malware Based on Sequence Alignment of Permissions [J]. International Journal of Computer, 2019, 35(1): 26-36.
- [17] Ganesh M, Pednekar P, Prabhuswamy P, et al. CNN-Based Android Malware Detection [C]//2017 International Conference on Software Security and Assurance (ICSSA), Altoona, PA, USA, 2017: 60-65.



顾 伟(1998-),男,工学硕士,研究方向为恶意检测、入侵检测,202219000007@nuist.edu.cn。