

基于“敏感数据区”的恶意 DOC/XLS 文档深度学习检测方法

陈 祥, 张 震, 王文博

信息工程大学信息技术研究所 郑州 中国 450002

摘要 Office 办公文档如 Word、Excel、PowerPoint 文档等已成为政府组织及企业日常办公中不可或缺的一部分, 它们为社会带来便利的同时也带来了严重的网络安全问题, 恶意 Office 文档已经被广泛应用于网络钓鱼攻击甚至 APT 攻击中。近年来, 随着深度学习技术在恶意软件检测、入侵检测等领域的应用, 研究人员开始将深度学习技术应用到恶意 Office 文档检测中, 但均存在样本数据集小、检测效果不佳的问题, 且缺乏与传统机器学习检测方法的有效比较。针对上述问题, 本文以探索深度学习技术在恶意 Office 文档检测中的应用方式及优缺点为目标, 基于文档中存在“敏感数据区”的思路, 分别提出了基于“敏感数据区”的恶意 DOC 文档深度学习检测方法和恶意 XLS 文档深度学习检测方法, 并在大量恶意文档和良性文档构成的数据集上进行了实验。为了避免人工经验在手动识别“敏感数据区”上的局限性, 本文还设计了一种自动识别并提取“敏感数据区”的方法, 并比较了两种不同提取方式的效果。实验表明, 本文提出的基于“敏感数据区”的恶意文档深度学习检测方法能够显著提高模型的检测效果, 优于当前基于机器学习的检测方法, 能够检测各种类型的恶意 DOC、XLS 文档, 如恶意宏文档、漏洞利用文档及其他类型恶意文档。本文还深入分析了深度学习检测方法的优势与不足, 为深度学习技术在恶意文档检测上的拓展深化指明了方向。

关键词 恶意文档; Office 文档; 检测; 深度学习; 敏感数据区

中图分类号 TP393.08 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2026.05.07

Deep Learning Detection Method for Malicious DOC/XLS Documents Based on “Sensitive Data Areas”

CHEN Xiang, ZHANG Zheng, WANG Wenbo

Institute of Information Technology, Information Engineering University, Zhengzhou 450002, China

Abstract Office documents, including those created in Microsoft Word, Excel, and PowerPoint, have become essential tools in the daily operations of government agencies and businesses. While these documents facilitate various tasks, they also introduce significant cybersecurity risks. Malicious Office documents are frequently employed in phishing campaigns and even advanced persistent threat (APT) attacks. In recent years, the integration of deep learning technologies into malware detection and intrusion detection has prompted researchers to investigate their application in the detection of malicious Office documents. However, challenges such as limited sample sizes and sub-optimal detection performance persist, alongside a lack of comprehensive comparisons with traditional machine learning-based detection methods. This article seeks to address these challenges by examining the methodologies, advantages, and limitations of deep learning approaches in the context of malicious Office document detection. Building on the concept of “sensitive data areas” within documents, we propose a deep learning detection framework specifically for malicious DOC and XLS files, utilizing these “sensitive data areas” as a basis for our approach. We conduct experiments using a substantial data set comprising both malicious and benign documents. To mitigate the reliance on manual expertise for identifying “sensitive data areas,” we introduce a method for their automatic identification and extraction, and we assess the effectiveness of two distinct extraction techniques. Our experimental results indicate that the proposed deep learning-based detection method, which leverages the notion of “sensitive data areas,” significantly enhances model’s detection performance, outperforming current machine learning-based detection methods. This approach is capable of identifying various forms of malicious DOC and XLS documents, such as malicious macro documents, exploit documents and other types of malicious documents. Furthermore, this article provides a thorough analysis of the strengths and weaknesses of deep learning-based detection methods, while also suggesting avenues for the further development and application of deep learning technologies in the realm of malicious document detection.

通讯作者: 陈祥, 博士, 助理研究员, Email: chenxndsc@163.com。

本课题得到国家重点研发计划青年科学家项目(No. 2022YFB3102800)资助。

收稿日期: 2024-09-18; 修改日期: 2025-03-06; 定稿日期: 2026-04-09

Key words malicious documents; office documents; detection; deep learning; sensitive data areas

1 引言

随着信息技术的飞速发展, Office 办公文档如 Word、Excel、PowerPoint 等已成为政府组织及企业日常运作中不可或缺的一部分, 它们极大地提高了工作效率与协作能力。然而, 这一普及趋势也催生了显著的网络安全挑战, 特别是成为网络钓鱼攻击频繁利用的主要媒介。攻击者通过精心构造看似合法的 Office 文档, 诱导用户下载并打开执行, 从而窃取敏感信息、传播恶意软件或执行其他恶意操作。这种利用 Office 文档进行的钓鱼攻击因其隐蔽性和高效性, 正日益成为网络安全领域的一大威胁。根据安全公司 Checkpoint 对 2022 年度邮件恶意文件类型的分析, 恶意文件中 Office 文档占比超过 56%, 常见的恶意 Office 文档格式包括 XLS、XLSX、XLSM、DOC、DOCX 等^[1]。

除了一般的网络钓鱼攻击外, 一直以来, APT 组织也通过挖掘 Office 软件 0-day 漏洞并构建恶意 Office 文档实施针对性鱼叉攻击。2020 年 1 月, APT 组织 Darkhotel 使用“双星”0Day 漏洞(CVE-2019-17026、CVE-2020-0674)针对中国发起 APT 攻击, 当受害者打开恶意文档触发漏洞后, 恶意代码会启动公式编辑器, 利用公式编辑器进程再打开 IE 或者 Firefox 浏览器访问恶意网页触发“双星”漏洞。2021 年 8 月, 微软威胁情报中心发现了试图利用特制的 Microsoft Office 文档实施的在野 0-day 攻击, 攻击者利用系统 MSHTML 组件中的远程代码执行漏洞(CVE-2021-40444)实现了对目标的突破, 然后下载恶意木马实施进一步的攻击。2022 年 5 月, 安全研究人员又发现了利用微软支持诊断工具(Microsoft Support Diagnostic Tool, MSDT)0-day 漏洞(CVE-2022-30190)实施的攻击, 进一步分析表明该 0-day 漏洞被用于攻击俄罗斯和白俄罗斯目标长达一个月或更长时间。2024 年 8 月, ESET 安全研究人员曝光了韩国网络间谍组织 APT-C-60 利用国产 WPS Office 软件的 0-day 漏洞(CVE-2024-7262)针对东亚国家实施的 APT 攻击。

针对恶意文档攻击带来的网络安全威胁, 近年来, 安全研究人员针对性地开展了检测技术研究, 从检测方法上当前研究以静态检测方法为主, 主要集中在 Office 文档中的恶意宏检测上^[2-7], 但其不是一种通用检测方法, 无法检测利用漏洞等其他手段的恶意 Office 文档。与仅关注文档中的宏不同, 一些

相关研究关注文档的结构, 通过提取文档的结构特征构建机器学习模型检测恶意文档, 如基于结构特征的恶意 PDF 文档检测^[8-9]和恶意 DOCX 文档检测^[10-11], 但这些机器学习方法由于利用的检测特征明确且单一, 容易被攻击者构建的反向模仿(Reverse Mimicry)攻击对抗样本规避^[12-13]。近年来, 随着深度学习技术在恶意软件检测^[14-16]、入侵检测^[17-19]等网络安全领域的广泛应用, 最近的研究已经开始将深度学习技术用于检测恶意 PDF 文档^[20-21]和 Office 文档^[22-25]。相比于机器学习, 深度学习模型不需要人工手动提取检测特征, 而是直接从文档字节流数据中学习检测特征, 降低了对人工经验的依赖。但相关研究在模型输入上通常提取文档的头部字节流数据, 数据过长容易使得特征学习困难, 数据过短容易丢失重要特征, 导致相关研究的检测效果普遍不佳, 也未能够与传统机器学习方法进行有效比较, 因此, 深度学习技术在恶意 Office 文档检测中的可用性仍有待于进一步研究。

针对上述问题, 本文以常见的恶意 DOC 文档类型和 XLS 文档类型为检测对象, 基于文档中存在能够反映文档恶意元素的“敏感数据区”, 分别提出了基于“敏感数据区”的恶意 DOC 文档深度学习检测方法和恶意 XLS 文档深度学习检测方法。为了避免人工经验在手动识别“敏感数据区”上的局限性, 本文还设计了一种自动识别并提取“敏感数据区”的方法, 并比较了两种不同提取方式的效果。实验分析表明, 针对 DOC 和 XLS 文件类型, 传统直接提取文档头部区域字节流送入深度学习模型训练和检测的方式效果并不理想, 而通过找到并选择文档的“敏感数据区”字节数据作为模型输入, 即 DOC 文档的目录项区域、WordDocument 流头部、DocumentSummaryInformation 流头部、文档头部以及 XLS 文档的目录项区域、表单子流头部, 能够显著地提高模型的检测效果。模型在较低误报率前提下取得对恶意 DOC 文档 98.75% 的检测率和对恶意 XLS 文档 98.68% 的检测率, 优于现有相关研究成果和主流杀毒软件。此外, 进一步分析表明, 这种利用深度学习的检测方法能够检测各种类型的恶意 DOC、XLS 文档, 如恶意宏文档、漏洞利用文档及其他类型恶意文档, 但单就恶意宏文档检测这一主流恶意文档类型而言, 由于缺乏对恶意宏本质特征的学习, 其检测效果不如基于人工提取恶意宏特征的机器学习模型, 表明在恶意文档检测上结合传统

机器学习检测方法的必要性,而之前的相关研究均忽略了这种进一步分析。

本文的主要贡献如下:

1) 提出了一种基于“敏感数据区”的恶意 Office 文档深度学习检测方法,并在大规模样本集上进行了实验;

2) 设计了一种文档“敏感数据区”自动提取方法,实验发现文档的“敏感数据区”与文档类型紧密相关,将其作为深度学习模型输入能够显著提高恶意文档检测效果;

3) 将深度学习方法在恶意 Office 文档上的检测性能与传统机器学习方法进行了深入对比,并分析了模型在不同类型恶意 Office 文档上的检测效果,提升了对当前各类检测模型真实能力及优缺点的认识。

2 相关工作

与恶意软件检测类似,恶意 Office 文档检测也分为静态检测和动态检测两种方式。静态检测关注文档的“形”,不需要用 Office 软件打开文档,直接从文档中提取恶意特征,具有检测效率高的优点,缺点是新出现的恶意文档检测能力不足;动态检测关注文档的“行”,通过打开文档或者运行文档中的宏,然后观测是否有恶意行为来检测恶意文档,优点是检测能力强,缺点是检测时间开销大,此外,也可能由于动态检测环境不满足漏洞发作要求导致一些恶意文档无法被检测到。

2.1 静态检测

静态检测方法中,根据检测模型中提取和使用的文档特征的不同,可以进一步区分为基于字节序列特征、内容特征和结构特征的检测方法。

字节序列特征是将文档看作一个连续的字节序列,并从中提取检测特征。Li 等人^[26]最早采用统计分析的思路,分别为良性文档样本和恶意文档样本的 Word 文档按字节流构建 n-gram 特征库,在检测文档时根据文档字节序列分别与良性 n-gram 特征库和恶意 n-gram 特征库的相似度判断文档安全性。Liu 等人^[27]将文档字节序列的时间序列熵作为关注点,从序列熵信号中提取全局特征(如均值、方差)、能谱特征和局部特征训练机器学习模型用于恶意文档检测,实验表明模型对恶意 Office 文档检测率达到 96%,误报率为 3.28%。

内容特征是指从文档内容中提取的用于区分良性或者恶意文档的特征,主要关注 Office 文档中嵌入的 shellcode、VBA (Visual Basic for Applications) 宏。白鹏等人^[28]对各类文档进行解析,获得可能存在

shellcode 的部分然后反汇编,通过检测是否加载动态链接库、自定位当前地址、自修改解码行为来决定是否为恶意 shellcode。Mimura 等人^[2]利用自然语言处理技术,通过将宏代码看成文本并从中提取词汇,先后采用词嵌入、潜在语义分析(Latent Semantic Analysis, LSI)等方法提取宏代码特征向量,训练机器学习模型对恶意宏进行检测,并通过实验分析了模型随着时间推移对新样本检测的效果。Kim 等人^[3]研究了 Office 文档中的 VBA 宏混淆问题,在深入分析四种典型 VBA 宏混淆方法的基础上,提取了 15 种检测宏混淆的特征。需要指出的是,该研究只是一种检测混淆的方法,而不是检测恶意宏的方法,尽管大多数恶意宏会采用混淆手段对宏进行混淆。Vasilios 等人^[4]分析了大量包含 VBA 宏的办公文档,并利用 40 个可疑关键词作为特征来识别恶意宏。在 2939 个良性样本和 13734 个恶意样本构成的数据集上实验表明,基于随机森林的机器学习模型检测性能最佳,精度为 99.3%,召回率为 97.6%。Chen 等人^[5]进一步结合混淆特征和宏可疑关键词特征,一共提取了 123 维特征用于恶意宏检测,将模型的检测精度和召回率提高到 99.7%。除 VBA 源代码外,部分研究关注 Office 文档中 VBA 源代码的编译版本,即 p-code 代码,进行恶意宏检测。Bearden 等人^[6]首先针对 p-code 代码提出了一种基于机器学习的恶意宏文档检测方法,通过提取 p-code 代码序列的 n-gram 特征进行检测。Simon 等人^[7]采用 p-code 序列的 2-gram 特征和 p-code 代码的混淆特征训练机器学习模型来识别基于 p-code 的恶意宏文档,对恶意文档的准确率达到 98.8%。尽管上述针对 Office 文档恶意宏的检测研究已经取得了较好的检测效果,但其无法检测利用漏洞及其他方式实施攻击的恶意文档。

结构特征是指从文档的内部结构或元数据信息中提取鉴别恶意文档的特征。Nir Nissim 等人^[10]从 DOCX 文档中提取目录树和 XML 文件标签路径等结构特征,基于上述特征训练检测恶意 DOCX 文档的机器学习模型。但作者在实验过程中正常样本有 16484 个,而恶意样本仅有 327 个,严重不平衡,且只适用于检测 DOCX 文档。Lu 等人^[11]在提取 DOC、DOCX 文档结构路径的基础上通过综合 VBA 函数关键字、结构路径、嵌入的 OLE(Object Linking and Embedding)对象和文档规范错误共四个方面的特征,构建了基于机器学习的恶意 Office 文档检测模型,在实验中大大提高了恶意 Office 文档的数量,实现对 DOC 文档和 DOCX 文档的检测准确率分别为 98.02%和 98.88%。但上述研究存在特征提取过程烦

琐、容易被精心构造的恶意文档样本绕过的问题。

区别于上述人工提取检测特征的方法, 近年来, 深度学习技术也被应用于恶意 Office 文档的静态检测上。Yang 等人^[22]通过提取 DOC 文档的前 1M 字节数据并转换成 1024×1024 大小的灰度图, 然后采用卷积神经网络在 1796 个文档样本构成的数据集上进行训练和测试, 最终获得 94.09% 的检测准确率。Jeong 等人^[23]针对韩国办公文档 HWP 文件以流为单位截取前 600 字节数据送入卷积神经网络进行训练和测试, 实验数据集基于 534 个 HWP 文档样本采样流数据得到, 结果表明模型检测准确率为 93.39%, F1 分数为 93.45%。随后作者将输入模型的流数据长度提升到 10000, 并改进 padding 模式及网络结构, 将 F1 分数提高到 95.08%^[24]。作者在上述研究基础上, 进一步针对 Office 文档构建特定的卷积神经网络检测模型, 在 1062 个 DOC 文档样本上实验获得的 F1 分数为 91.12%^[25]。可见, 针对恶意 Office 文档检测, 上述基于深度学习技术的研究虽然在提取文档字节数据、改进模型结构上做了一些努力, 但传统的机器学习方法相比, 存在数据集规模小、检测效果不理想的问题。

2.2 动态检测

Li 等人^[26]为了弥补静态检测能力的不足, 在结合静态检测结果的基础上, 进一步通过改变动态链接库(Dynamic Link Library, DLL)加载顺序以检测是否崩溃、是否存在区别于良性操作行为特征库的异常注册表行为、文档打开时弹框检测等措施增强对包含恶意代码的 Word 文档的检测能力, 其问题是考虑的动态行为特征和场景较为单一, 只能作为静态检测的补充。陈祥等人^[29]实现一种基于全局行为特征的未知恶意文档检测方法, 在 Office 文档处理过程中提取全系统行为特征, 仅训练良性文档样本形成行为特征库用于恶意文档检测, 并引入敏感行为特征用于降低检测误报率。Jiang 等人^[30]通过 cuckoo 沙箱获取办公文档处理过程中的分析报告, 从中提取高频合法字符串作为特征, 并通过大量良性样本和恶意样本训练 TextCNN(Text Convolutional Neural Networks)模型用于文档检测, 在数据集上实验表明模型检测准确率达到 99.06%。

由于近年来基于 Excel 4.0 宏的恶意文档攻击呈现上升趋势, Ruaro 等人^[31]针对 Excel 4.0 宏实现了一款动态分析软件—SYMBEXCEL, 该软件使用符号执行自动去混淆和分析这些宏, 使用户能够可靠地提取出被攻击的指示(Index of Compromise, IOC)和其他重要的取证信息。

3 DOC/XLS 文档物理结构

DOC 和 XLS 文档格式实质上是一种复合文档格式(Compound Document Format, CDF), 这种格式最初由微软公司在 1995 年推出, 其设计目的是更好地组织和存储复杂的数据集合, 包括文本、图像、图表、音频、视频等多种类型的数据, 被广泛应用于各种类型的文件中, 如 Microsoft Office Word、Excel 和 PowerPoint 支持的 DOC、XLS 和 PPT 文件。

复合文档格式的核心特性之一是其层次化的存储结构。它采用了一种类似于文件系统的存储机制, 将数据组织成一系列的仓库(Storage)和流(Stream)。仓库可以包含其他仓库和流, 而流则是实际存储数据的地方。每个复合文档都包含一个根仓库, 是解析复合文档内部结构的入口。这种层次化的结构使得复合文档可以同时存储多种不同类型的数据, 并且保持良好的组织性和可访问性。图 1 展示了复合文档格式逻辑结构的示例, 根仓库下面包含 2 个子仓库和 2 个流, 每个子仓库都包含相应的流数据。

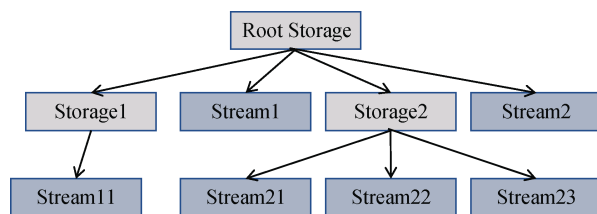
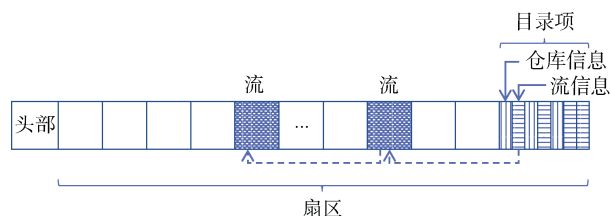


图 1 复合文档格式逻辑结构示意图

Figure 1 Schematic diagram of logical structure of compound document format

复合文档格式在物理结构组织上也与文件系统类似, 包括开始的 512 字节头部信息和一系列的扇区, 如图 2 所示, 头部定义了解析复合文档所需要的所有信息, 如扇区的大小、字节序、扇区配置表信息、目录项偏移等。通过定位目录项所在的扇区, 可以找到根仓库信息, 进而找到子仓库和流的扇区信息, 通过查找扇区配置表, 从而定位到所有包含的数据流。实际上, 复合文档中还采用链式索引、短流



等存储方式以进一步提高文件的存储效率和灵活性,更多关于复合文档格式的介绍参见文献[32]。

DOC、XLS 文档类型均在复合文档结构的基础上,进一步定义了文档的逻辑存储结构及与文档内容相关的所有元素信息。

3.1 DOC 文档结构

DOC 文档格式,也称为 Word 二进制文件格式^[33],主要指的是早期版本的 Microsoft Word 97 至 Word 2003 所使用的文件格式。

在DOC文档的结构中,最核心的流是WordDocument流,它存储了文档的主体内容,包括文本、段落、样式、格式等所有与文档编辑直接相关的信息。用户通过 Word 软件对文档进行的所有编辑操作,最终都会反映在这个流中。此外,文档中 SummaryInformation 和 DocumentSummaryInformation 流用于存储文档的元数据,如标题、作者、创建时间、修改时间、主题等。对于有图像、音视频及其他利用 OLE 技术嵌入的应用程序对象,DOC 文档中还通常会包含 ObjectPool 仓库,这些对象保存于 ObjectPool 仓库下面的流中。如果文档包含 VBA 宏,则 DOC 文档结构中必须有 Macros 仓库用于包含文档中的宏。此外,DOC 文档格式还定义了与文档加密、签名、版权等功能相关的流结构信息。图 3 显示了一个带有宏的 DOC 文档的流结构,其中涵盖了上面介绍的一些常见的流。

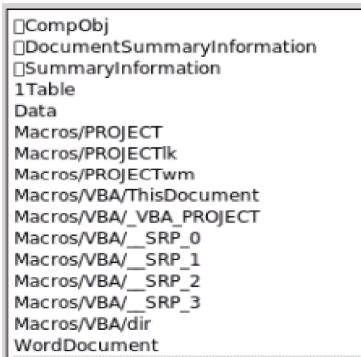


图 3 DOC 文档流结构示例

Figure 3 Example of DOC document stream structure

3.2 XLS 文档结构

XLS 文档格式,也称为 Excel 二进制文件格式^[34],是早期 Microsoft Excel 97 至 2003 版本所采用的标准文件类型,能够高效地存储和处理 Excel 文档中的文本、数值、公式、图表等多种元素。

在 XLS 文档的结构中,除了复合文档格式中通常流结构外,还进一步定义了子流(Substream)和记

录(Record)结构。记录是可变长度的字节序列,由记录类型、记录大小以及特定于该记录类型的记录数据三个部分组成,是存储工作簿中功能信息的基本构建块。在 XLS 文档结构中最重要的是 Workbook 流,该流中包括全局信息(Global)子流和表单(Worksheet)子流,这些子流均由记录构成。全局信息子流负责存储整个工作簿的全局信息,包括工作表的数量、名称、属性设置以及工作簿级别的保护信息等,这些信息对于理解整个 Excel 文档的结构至关重要。每个表单子流对应一个工作表,存储了工作表中的单元格数据、公式、格式(如字体、颜色、边框等)、图表引用等详细信息。用户在工作表中进行的所有编辑,如数据输入、格式调整等,最终都会反映在这些流中。在 XLS 文档中还可以包含与文档元数据、嵌入 VBA 宏、嵌入 OLE 对象、加密、签名等功能相关的流。图 4 显示了一个带有 VBA 宏的 XLS 文档的流结构,Workbook 流包含了工作簿的所有数据,_VBA_PROJECT_CUR 仓库下为 VBA 宏相关的流,其他为与文档元数据信息相关的流。

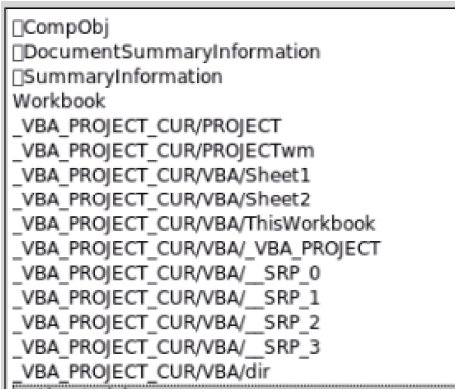


图 4 XLS 文档流结构示例

Figure 4 Example of XLS document stream structure

4 恶意 DOC/XLS 文档攻击

以 Office 办公文档为载体的恶意文档攻击有多种类型,本节详细说明了四种常见恶意文档攻击类型,分别为恶意软件攻击、恶意 DDE 命令攻击、恶意宏攻击和漏洞利用攻击。

4.1 恶意软件攻击

Office 办公文档作为一种复合文档,允许嵌入各种类型的文件,如文档、图片、音视频、压缩包、可执行文件等。攻击者可以将恶意的可执行程序嵌入到文档中,通过诱导用户点击的方式执行恶意程序实施攻击。但随着办公软件安全措施的改进及用户安全意识的提升,传统的直接在文档中嵌入 Windows PE 可

执行程序的方法在用户点击运行时会出现弹框告警, 甚至被阻拦, 容易被用户发现, 攻击的成功率不高。

在上述背景下, 攻击者开始转向嵌入非 PE 类型的恶意软件。例如, 攻击者冒充管理员向目标发送钓鱼邮件, 声称公司有部分员工的主机被勒索病毒感染, 要求所有员工进行勒索病毒查杀, 并将压缩包类型的查杀工具和步骤说明一起保存在邮件附件的文档中, 文档内容如图 5 所示, 并在文档内容中指示目标如何运行软件, 通过这种方式绕过办公软件提供的安全保护。

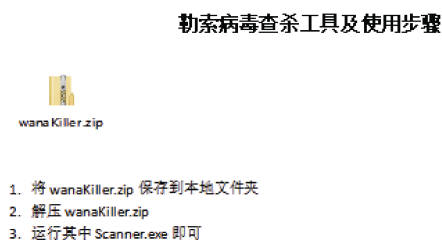


图 5 嵌入恶意软件攻击示例

Figure 5 Example of embedded malware attack

4.2 恶意 DDE 命令攻击

动态数据交换(Dynamic Data Exchange, DDE)主要是微软用来允许两个正在使用的应用程序共享数据的方式, 由于 DDE 本身是 Microsoft Office 系列软件的合法功能, 所以绝大多数安全软件不会阻止 DDE 字段文档运行。

攻击者可以创建包含恶意 DDE 字段的 Word 文档, 用户打开文档并点击允许运行后, 文档里内嵌的代码会运行。例如, 通过在 DOCX 文档中选择插入-文档部件-域命令 *DDEAUTO c:\Windows\System32\cmd.exe "/k powershell.exe -NoP -sta -NonI -W Hidden \$e=(New-Object System.Net.WebClient).DownloadString('http://willgenovese.com/hax/evil.ps1'); powershell -e \$e "*, 可以实现下载远程恶意代码并执行。但这种攻击的不足之处是打开文件过程中会有两次弹窗, 第一次是询问是否更新链接, 第二次是询问是否执行程序, 当两个都点击确认后才会执行, 图 6 显示了一个利用 DDE 功能的恶意 DOCX 文档打开后的弹窗信息。

4.3 恶意宏攻击

由于宏在正常文档中也大量使用, 普通用户对宏的安全性认识不足, 基于宏的恶意文档攻击已经成为恶意 Office 文档攻击中最广泛使用的攻击手段, 当前基于宏的恶意文档主要是基于 VBA 宏和 XLM 宏的恶意文档。

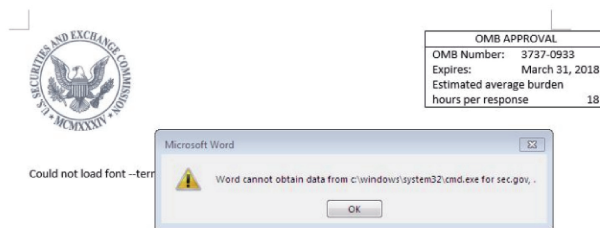


图 6 嵌入恶意 DDE 攻击示例

Figure 6 Example of embedded malicious DDE attack

VBA 宏中提供的函数与命令能够直接调用系统上各类资源, 与其他应用程序交互, 实现跨应用的数据交换和控制。图 7 显示了一个恶意 VBA 宏代码示例, 攻击者使用 VBA 函数“GetObject”获取一个正在运行的 Windows 资源管理器实例, 然后调用“ShellExecute”方法启动 PowerShell 命令下载远程恶意脚本并执行。

```
Sub Document_Open()  
    Set obj = GetObject("new:C08AFD90-F2A1-11D1-8455-00A0C91F3880")  
    obj.Document.Application.ShellExecute "powershell.exe", "IEX (New-Object Net.WebClient).DownloadString('http://192.168.131.127:8000/httptest1.ps1')", "C:\Windows\System32", Null, 0  
End Sub
```

图 7 恶意 VBA 宏代码示例

Figure 7 Example of malicious VBA macro code

当然, 上述恶意 VBA 宏容易被杀毒软件检测发现, 因此混淆与规避是在恶意宏中惯常使用的技术手段, 攻击者通过混淆与规避隐藏恶意代码的真实意图和功能, 以逃避防病毒软件和安全分析工具的检测分析。

XLM 宏作为 Excel 软件中遗留的宏功能, 是传统 Excel 函数的超级集, 它提供了大量的函数来与 Excel 工作簿和外部环境交互。如 GET.CELL 用于获取单元格的属性, 如行号、列号等, PUT.CELL 用于设置单元格的属性, 如字体颜色、背景色等, GOTO 用于跳转到指定的单元格, RUN 用于执行另一段 XLM 宏代码。与电子表格中使用的传统 Excel 函数(如 SUM)不同, XLM 宏函数可以访问 Windows API, 并可用于与操作系统交互, 因此也被攻击者用于实现各种恶意操作。例如, 单元格中公式“=FILES(directory)”列出了给定目录中的所有文件, 公式“=EXEC(program)”可以启动外部程序。

需要指出的是, 在 Office 软件的默认宏安全设置下, 当打开带有宏的文档时, 宏不会自动运行, 需要用户决定是否运行宏, 因此恶意宏文档攻击本质上也是一种社会工程学攻击。

4.4 漏洞利用攻击

由于 Office 软件设计实现复杂, 代码规模甚至达到数千万, 必然蕴涵着大量漏洞, 一旦攻击者挖掘 0-day 漏洞并基于此构建恶意 Office 文档实施攻击, 往往能够在目标无感的情况下实现隐匿攻击, 因此成为安全人员关注的重点。

利用的漏洞既包括逻辑漏洞, 也包括内存漏洞。逻辑漏洞往往是因为在设计软件时逻辑功能不严密导致的, 如 CVE-2017-0199, 允许攻击者用 URL Moniker 加载本地或远程 HTA 文件执行。对于内存漏洞, 根据具体的漏洞类型及目标采取的安全防护措施, 攻击者往往采用不同的利用技术, 针对栈缓冲区溢出漏洞(如 CVE-2017-11882)通常采取返回地址修改、异常链劫持等手段实现控制流劫持; 针对堆内存破坏漏洞(如 CVE-2018-8174)攻击者通常采用 ROP、堆喷等方式, 通过修改返回地址、函数指针、虚表指针等实现控制流劫持。

5 基于“敏感数据区”的恶意 DOC/XLS 文档深度学习检测方法

在深度学习模型选择上, 与之前的相关研究^[22-25]类似, 本文也采用经典的卷积神经网络(Convolutional Neural Networks, CNN)模型, 将文档数据流当作图像数据处理, 以实现恶意文档的高效的分类检测。正如在图 2 中所看到的, 由于 DOC/XLS 文档在物理存储结构上采用类文件系统方式, 内部的流数据被按照扇区的方式分散存储, 即使同一个流的数据也可能分散在不同物理区域, 物理上的连续字节流并不蕴含文档内部数据之间的逻辑关系, 因此简单地提取文档头部数据区域输入卷积神经网络不利于提取局部特征, 需要提取逻辑上关联的数据输入模型,

如文档头部元数据信息、目录项区域、特定的流数据等。

此外, 卷积神经网络模型要求输入长度固定, 在图像处理中通常通过缩放、裁剪或填充等方式将输入图像统一调整到相同的大小, 而 DOC/XLS 文档的大小并不固定, 内部的流数据通常从几百字节到几 MB 不等, 典型的处理方式是设定一个阈值, 当数据长度超过该阈值时进行截断, 数据长度少于该阈值时按一定的方式补全。但这种方式存在的问题是, 如果长度阈值设置过大, 如 1MB, 对于数据流中只有少量恶意数据的场景, 模型学习到恶意数据模式并检出的能力有限, 如果长度阈值设置得过短, 则数据流中的恶意部分也可能被截断掉, 同样降低模型的检测能力。考虑到对于检测恶意文档来说, 文档结构中并不是所有的数据拥有相同的重要性。例如, DOC 文档中的 WordDocument 流用来存储文档主体内容, 对于大多数恶意宏文档, 宏并不会给 WordDocument 流带来影响, 但是会影响文档的目录项区域以及 Macros 下面具体的流。这表明文档中存在一些“敏感数据区”, 这些数据区与嵌入的恶意元素直接相关, 而其他大部分数据区与文档是否嵌入恶意元素无关, 因此, 在利用深度学习模型检测恶意文档时, 更应该将文档的“敏感数据区”筛选出来送入模型, 使得模型能够聚焦到“敏感数据区”上, 通过大规模样本的学习实现检测特征的自动提取。

基于上面的考虑, 本文提出的基于“敏感数据区”的恶意 DOC/XLS 文档深度学习检测流程如图 8 所示, 首先获取文档的原始字节流序列, 然后在数据提取阶段从文档的全部字节序列中提取“敏感数据区”, 并将“敏感数据”当作图像像素数据, 送入卷积神经网络训练和测试, 最终得到能够用于检测恶意 DOC 和 XLS 文档的模型。

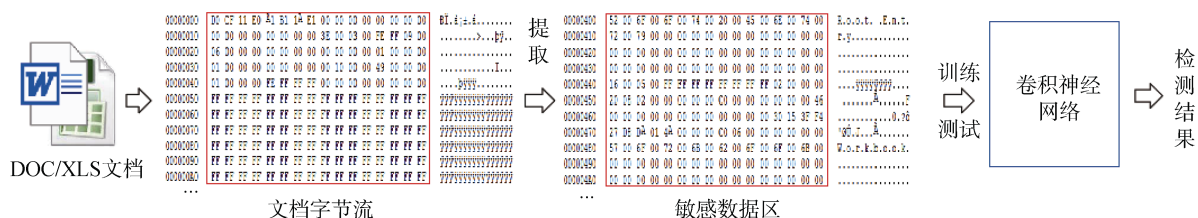


图 8 基于深度学习的恶意 DOC/XLS 文档检测流程

Figure 8 Detection process of malicious DOC/XLS documents based on deep learning

本文实验涉及的代码和结果已开源在文献[35]。

5.1 数据集构建

在各类深度学习应用的场景中, 样本集的规模直接影响深度学习检测模型的效果, 本小节对本文构建的 DOC/XLS 文档样本数据集进行介绍。

为了获得具有代表性的正常文档样本, 本文根据域名流行度从 Top-1M 域名^[36](Cisco 公司依据其 DNS 解析遥测数据提供)中筛选出了全球 7229 个政府网站域名和 8081 个教育网站域名, 并从这些站点爬取 DOC 和 XLS 格式的文档样本, 最终获得 197925

个 DOC 文档和 27434 个 XLS 文档。经杀毒软件检测其中仅有 2 个 DOC 文档为恶意文档, 其余文档为良性文档。

所有的恶意文档均来自 Virustotal, 包括 22028 个恶意 DOC 文档和 14608 个恶意 XLS 文档(样本提交到 Virustotal 的时间范围从 2008 年至 2022 年)。由于本文获得的良性 DOC 文档远远多于恶意 DOC 文档, 考虑到数据集中正负样本的平衡性, 在构建数据集的过程中本文从所有正常文档中随机选择了 40000 个组成良性 DOC 文档。另外, 通过对正常文档和恶意文档的分析发现, 恶意的 DOC 文档中很多都包含宏, 而在正常文档中只有少数包含宏, 为避免模型学习出现偏差, 本文在 40000 个良性 DOC 文档样本的基础上从剩下的正常文档中筛选出所有包含宏的正常 DOC 文档, 并补充到良性 DOC 样本中, 最终构建的 DOC/XLS 恶意文档检测数据集情况如表 1 所示。

表 1 DOC/XLS 恶意文档检测样本集

Table 1 Dataset for detection of malicious DOC/XLS documents

	良性 DOC 样本	恶意 DOC 样本	良性 XLS 样本	恶意 XLS 样本
总数量	42537	22028	27434	14608
其中包含宏 的样本数量	3549	17641	1982	14189

可以看出, 本文虽然刻意补充了包含宏的良性 DOC 文档, 但包含宏的良性 DOC 文档在所有良性 DOC 文档中比重不足 10%, 而恶意 DOC 样本中包含宏的文档比例超过 80%。对于 XLS 样本, 既可以包含 VBA 宏, 也可以包含 XLM 宏, 统计过程中不作进一步区分统一认为是包含宏的文档, 发现良性 XLS 样本中包含宏的样本仅占 7%, 而恶意 XLS 样本中宏文档占比达到 97%。

5.2 “敏感数据区”提取

在“敏感数据区”提取上, 本文首先根据对各种恶意文档攻击的经验知识, 尝试手动从文档中提取“敏感数据区”。但考虑到人工经验知识的局限性, 对“敏感数据区”的识别必然存在盲区, 本文也设计了一种自动识别并提取“敏感数据区”的方法, 下面分别展开介绍, 在实验部分将比较这两种不同“敏感数据区”提取方法的检测效果。

(1) 手动提取方法

恶意 DOC 文档和恶意 XLS 文档的类型存在差异, 提取的“敏感数据区”也不尽相同。

对于 DOC 文档, 考虑到结构特征常作为检测特征应用于机器学习模型, 而文档的目录项区域蕴含着文档内部流结构及流类型、名称、长度、创建与修改时间等更多信息, 因此, 本文提取文档目录项所在扇区(或多个扇区)的字节流数据作为“敏感数据区”。例如, 很多利用 OLE 对象或者控件漏洞的恶意文档在目录项区域就会包含这些控件的名称、类型和大小信息, 即使是恶意宏文档, 其包含宏的数量、名称、长度等信息也会体现在目录项区域。由于一个目录项大小为 128 字节, 本文将 DOC 文档“敏感数据区”设置为 2048 字节, 即能够覆盖 16 个目录项, 超过绝大多数文档的目录项数。

对于 XLS 文档, 由于目录项中数据相比于 DOC 文档更加单一, 而更多的信息在 workbook 流的 global 子流和 sheet 子流中, 例如, 基于 XLM 宏的恶意文档在 sheet 子流头部会包含表单的类型信息, 因此, 在提取 2048 字节目录项结构的基础上, 本文也提取了这些子流的头部 128 字节数据, 最多支持 16 个子流, 即 1 个 global 子流和 15 个 sheet 子流(对应 15 个表单), 将它们与目录项数据拼接后形成“敏感数据区”, 该过程如图 9 所示。

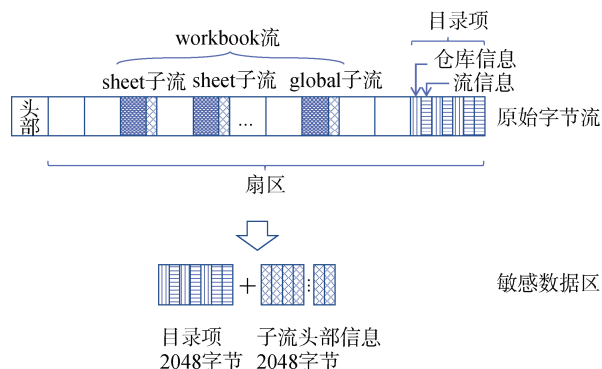


图 9 XLS 文档“敏感数据区”提取示意

Figure 9 Schematic diagram of sensitive data area extraction from XLS documents

可见, 手动提取“敏感数据区”的过程虽然没有提炼出明确的恶意特征, 但仍需要基于对各类恶意文档的深入理解, 找到可能会体现文档恶意特征的数据区域。

为了比较上述手动提取的“敏感数据区”对深度学习模型检测效果的影响, 本文分别针对 DOC 文档和 XLS 文档设计了多种不同的模型输入“数据区”提取方式。表 2 中列出了用于深度学习模型输入的各种“数据区”手动提取方式, 其中 DOC-DE-2048 和 XLS-DE2048-SS16_128 分别对应 DOC 文档和 XLS 文档的“敏感数据区”提取方式。在数据提取过程中,

当出现要提取的目标字节流不能满足长度要求时, 统一填充字节 0x00。

表 2 DOC/XLS 文档模型输入“数据区”手动提取方式
Table 2 Manually extracted data area for model input of DOC/XLS documents

名称	说明
DOC-HE-4096	提取 DOC 文档头部开始的 4096 字节(4KB)
DOC-HE-16384	提取 DOC 文档头部开始的 16384 字节(16KB)
DOC-HE-65536	提取 DOC 文档头部开始的 65536 字节(64KB)
DOC-DE-2048	提取 DOC 文档目录项区域的 2048 字节(2KB)
DOC-DE-4096	提取 DOC 文档目录项区域的 4096 字节(4KB)
XLS-HE-4096	提取 XLS 文档头部开始的 4096 字节(4KB)
XLS-HE-16384	提取 XLS 文档头部开始的 16384 字节(16KB)
XLS-HE-65536	提取 XLS 文档头部开始的 65536 字节(64KB)
XLS-DE-2048	提取 XLS 文档目录项区域的 2048 字节(2KB)
XLS-DE-4096	提取 XLS 文档目录项区域的 4096 字节(4KB)
XLS-DE2048-SS 16_128	提取 XLS 文档目录项区域的 2048 字节和最多 16 个子流头部 128 字节

(2) 自动提取方法

文档“敏感数据区”自动提取方法的核心思路是通过不断尝试“提取、组合、训练、验证、比较”的流程找到最大化恶意文档深度学习检测模型检测效果的数据区域组合。

下面以 DOC 文档类型为例说明自动提取方法的工作原理。首先找到文档元数据区域和常见数据流对象构成候选数据区集合, 包括文档头部、目录项区域、WordDocument 流、DocumentSummaryInformation 流、SummaryInformation 流、Data 流和 Table 流。然后提取上述候选数据区对象的头部数据(认为关键元数据信息均在流的头部区域), 分别训练深度学习检测模型, 比较模型检测结果, 将检测效果最佳的模型输入加入“敏感数据区”, 并将该类型对象从候选数据区集合移除。对于候选数据区集合中剩下的对象, 分别与上述已经识别出的“敏感数据区”组合后重新训练深度学习检测模型, 如果组合后模型检测性能下降, 表明将引入该类型对象并不能提升模型检测效果, 则将该类型数据从候选数据区集合移除; 如果组合后模型检测性能提升, 则找到使得模型性能提升最多的输入加入“敏感数据区”。重复上述过程直到候选数据区集合为空。上述方法的伪代码描述如表 3 所示。

在上述提取候选数据区对象头部数据时, 提取的长度和数据区类型有关, 如文档头部固定为 512 字节, 则提取长度为 512 字节, 各种不同的流长度不定, 本文将提取长度设定为 1024 字节, 对于目录项区域, 和手动提取方法的最小长度保持一致, 设定

为 2048 字节, 这样的设定能够简化自动提取“敏感数据区”方法的复杂度。

表 3 DOC/XLS 文档“敏感数据区”自动提取方法
Table 3 Automatic extraction method for the sensitive data areas in DOC/XLS documents

“敏感数据区”自动提取方法
输入: 由元数据区域和流构成的候选数据区集合 T , $T = \{t_1, t_2, \dots, t_n\}$ 带标签的文档样本集 DS , 神经网络模型 M 输出: 敏感数据区 S
1: $S = \{\}$
2: $f_score_{initail} = f_score_{max} = 0$
3: for t_i in T : //评估候选数据区集合中每种类型数据
4: $k = -1$
5: $S_i = S \cup \{t_i \text{ 头部}\}$, //加入待评估的数据区对象, 头部长度根据数据区类型人工指定
6: 针对样本集 DS 中的每个文档, 根据 S_i 提取文档数据形成新数据集 DS_i
7: 将 DS_i 划分为训练集和验证集, 采用交叉验证的方式训练模型 M 并验证检测结果, 记录 f 分数为 f_score
8: if $f_i_score > f_score_{max}$: //找到最能够提升模型检测效果的数据区对象
9: $f_score_{max} = f_i_score$
10: $k = i$
11: if $f_i_score < f_score_{initail}$: //移除使得原模型性能下降的数据区对象
12: $T = T - \{t_i\}$
13: if $k \neq -1$:
14: $S = S \cup \{t_k \text{ 头部}\}$ //将评估得到的最佳数据区对象的头部添加到“敏感数据区”
15: $T = T - \{t_k\}$ //将表现最佳数据区对象从 T 中移除
16: $f_score_{initail} = f_score_{max}$ //记录模型性能分数用于下一次评估比较
17: if $T \neq \emptyset$: //继续评估候选数据区集合直至为空
18: Goto Step 3
19: return S

5.3 模型结构

检测模型采用典型的一维卷积神经网络结构, 如图 10 所示, 包含输入层、两个卷积层和池化层、展平层、全连接层、线性层和输出层。在卷积神经网络中, 卷积层通过在输入数据上滑动卷积核来提取特征, 可以捕捉到数据局部相关性, 通过不同的卷积核可以得到不同维度的特征图。池化层通过缩小特征图的尺寸, 减少特征图的参数数量, 同时保留主要的特征信息, 有助于提高模型的鲁棒性和泛化能力。展平层用于将多维张量转换为一维向量, 全连接层和线性层通常用于将展平层的输出与分类或回归任务相关联。

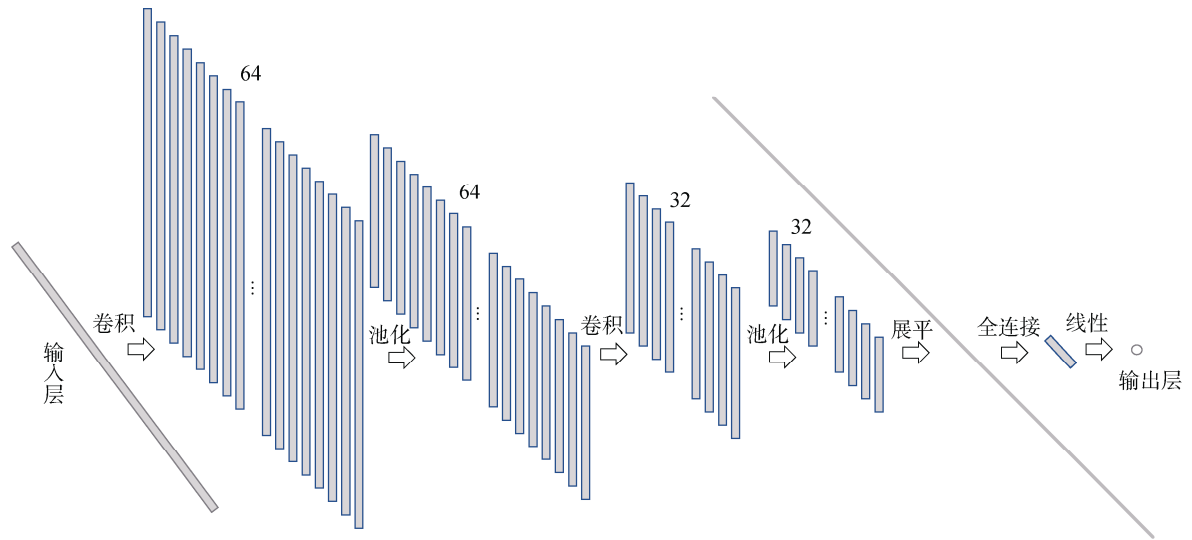


图 10 检测恶意 DOC/XLS 的卷积神经网络结构

Figure 10 Convolutional neural network architecture for detecting malicious DOC/XLS documents

在模型结构下, 预处理后得到的目标字节流数据经过两层一维卷积和池化操作后被展平, 然后经过全连接层、Dropout 层(图中未显示)、线性层后输出文档为恶意文档的概率。其中第一个卷积层包含 64 个卷积核, 第二个卷积层包含 32 个卷积核, 卷积核大小均为 3×3。

6 实验结果

实验在华为 2488V5 服务器上进行, 该服务器配备 2 个至强(R) Gold 5118 CPU、64 GB 系统内存和 CentOS 8 Stream 操作系统。实验过程中, 本文将构建的 DOC/XLS 恶意文档检测数据集按照 4: 1 划分为训练集和验证集, 并采用 5 折交叉验证后的平均值作为实验结果。评估指标包括精度、召回率、准确率、F1 分数、误报率等常用指标。下面分别介绍深度学习模型针对恶意 DOC 文档和恶意 XLS 文档的检测效果。

6.1 恶意 DOC 文档检测

(1) 模型检测结果

基于不同“数据区”提取方式的深度学习模型检测效果如表 4 所示。从表中可以看出, 对于模型 CNN-DOC-HE-4096 和 CNN-DOC-DE-4096, 即使提取 DOC 文档同样长度为 4096 字节的数据, CNN-DOC-DE-4096 模型的检测效果在各项指标上明显优于前者, 表明从文档物理结构来看, 不同区域的数据对于区分恶意文档所能起到的作用不同, 文档的目录项区域由于蕴含着文档内部的流结构和元数据信息, 更能够从中提取和捕捉到恶意 DOC 文档的特征。此外, 当从文件头部开始提取字节流长度从 4KB 增加到 16KB 和 64KB 时, 模型检测结果的 F1 分数反而明显下降; 当提取目录项区域字节流从 2KB 增加到 4KB 时, 模型的 F1 分数只有很小变化。这都表明对于模型来说, 输入模型的目标字节流不是越多模型的检测效果越好。

表 4 深度学习模型在 DOC 样本集上的测试结果

Table 4 Detection results of deep learning model on DOC samples

	模型	精度	召回率	准确率	F1 分数
手动提取	CNN-DOC-HE-4096	97.71%	95.20%	97.60%	96.43%
	CNN-DOC-HE-16384	96.08%	95.31%	97.10%	95.70%
	CNN-DOC-HE-65536	99.00%	94.70%	97.88%	95.70%
	CNN-DOC-DE-2048	98.70%	97.72%	98.78%	98.21%
	CNN-DOC-DE-4096	98.71%	97.84%	98.83%	98.28%
自动提取	CNN-DOC-HE-512-DE-2048-WORDDOCUMENT-1024-DOCUMENTSUMMARYINFORMATION-1024	99.01%	98.32%	99.09%	98.66%

采用本文提出的自动“敏感数据区”提取方法发

现的“敏感数据区”由 512 字节文档头部、2048 字节

目录项区域、1024 字节 WordDocument 流和 1024 字节 DocumentSummaryInformation 流构成, 提取这些数据构建的深度学习检测模型检测准确率为 99.09%, F1 分数为 98.66%, 优于手动提取“敏感数据区”方法对应的检测模型 CNN-DOC-DE-2048, 表明除了文档目录项区域外, 文档头部、WordDocument 流等对象中的一些数据区域也能辅助识别恶意 DOC 文档, 也表明本文提出的自动“敏感数据区”提取方法是可行的。实验过程中还发现, 按“提取、组合、训练、验证、比较”的流程添加到“敏感数据区”的对象依次为 2048 字节目录项区域、1024 字节 WordDocument 流、1024 字节 DocumentSummaryInformation 流和 512 字节文档头部, 上述次序也反映了它们在识别恶意文档中的重要程度。

(2) 检测结果对比

本文随机选择 5 折交叉验证中其中一次训练集和测试集划分, 将性能最佳模型 CNN-DOC-HE-512-DE-2048-WORDDOCUMENT-1024-DOCUMENTSUMMARYINFORMATION-1024 的检测结果与杀毒软件的检测结果进行了对比。测试集包括 4374 个恶意 DOC 样本和 8539 个良性 DOC 样本, 本文基于 Virustotal 获得了测试集中所有 4374 个恶意样本的最新检测数据, 其中包括各大杀毒软件对恶意样本的检出情况。实验中进一步将恶意样本中至少被 5 个杀毒软件标记为恶意的筛选出来, 共计 4318 个, 并比较模型和所有杀毒软件对这 4318 个恶意样本的检出情况, 图 11 显示了本文模型和检出率前 10 的杀毒软件的检测率。

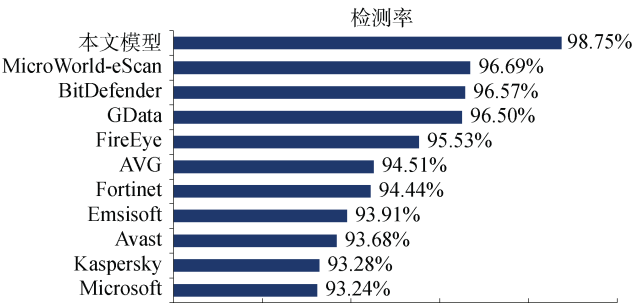


图 11 模型与杀毒软件在恶意 DOC 样本上检测率比较
Figure 11 Comparison of detection rates between our model and antivirus software on malicious DOC samples

可以看出, 本文检测模型对测试集中的恶意 DOC 样本的检测能力明显优于各大杀毒软件, 而此时模型的良性 DOC 样本的误报率仅为 0.67%。

本文进一步将上述 CNN-DOC-HE-512-DE-2048-WORDDOCUMENT-1024-DOCUMENTSUMM

ARYINFORMATION-1024 模型的检测结果与已有基于深度学习、机器学习方法的相关研究的检测结果进行比较, 如表 5 和表 6 所示。

从表 5 可以看出, 与 Yang^[32]直接提取文档头部 1M 数据送入卷积神经网络的方法相比, 本文方法准确率达到 99.09%, 显著高于其 94.94% 的准确率; 与 Jeong 等人^[35]研究 91.12% 的 F1 分数相比, 本文方法 F1 分数达到 98.66%, 充分表明通过提取文档“敏感数据区”作为深度学习检测模型输入的方法能够显著提高模型的检测能力。

从表 6 可以看出, 与现有基于机器学习的相关检测研究对比, 本文方法检测率更高, 误报率更低, 这也表明深度学习技术在恶意文档检测上具备强大的检测特征自提取能力。

在检测时间开销方面, 提取 DOC 样本的“敏感数据区”并输入模型完成检测的平均检测时间开销为 90 毫秒。

表 5 与现有基于深度学习的恶意 DOC 文档检测相关研究检测结果比较

Table 5 Comparison of detection results with existing research on malicious DOC document detection based on deep learning

研究	检测方法	数据集	准确率	F1 分数
本文	提取“敏感数据区”送入卷积神经网络	良性 DOC 文档 42537 个, 恶意 DOC 文档 22028 个	99.09%	98.66%
	提取文档头部 1M 字节数据送入卷积神经网络	良性 DOC 文档 818 个, 恶意 DOC 文档 978 个	94.94%	—
Yang ^[32]	提取文档所有流前 10000 字节送入改进的卷积神经网络	良性 Office 文档 536 个, 恶意 Office 文档 526 个	—	91.12%
Jeong ^[35]				

表 6 与现有基于机器学习的恶意 DOC 文档检测相关研究检测结果比较

Table 6 Comparison of detection results with existing research on malicious DOC document detection based on machine learning

研究	检测方法	数据集	检测率	误报率
本文	字节序列+深度学习	良性 DOC 文档 42537 个, 恶意 DOC 文档 22028 个	98.32%	0.67%
	结构路径、宏关键字、OLE 对象格式特征+机器学习	良性 DOC 文档 11638 个, 恶意 DOC 文档 26534 个	97.65%	1.38%
Lu ^[5]				
Liu ^[17]	字节序列特征+机器学习	良性 Office 文档 1903 个, 恶意 Office 文档 1774 个	98.10%	3.98%

(3) 分类型检测结果

由于恶意 DOC 文档可能涵盖多种不同的攻击类型, 如恶意宏文档、漏洞利用文档或利用 DDE 的恶意文档, 但在前面的模型检测结果及相关研究中都没有对这些不同类型的恶意 DOC 文档进行区分。

为了弄清楚模型在这几类恶意样本上的检测能力, 对于测试样本中的 4318 个恶意 DOC 样本, 本文依据 Virustotal 提供标签信息中的“exploit”“macros”关键字对恶意 DOC 样本进行分类。如果样本标签中包含“exploit”关键字, 则认为这个样本是利用漏洞的恶意文档, 如果样本包含“macros”关键字但又不包含“exploit”关键字, 则认为样本是包含恶意宏的恶意文档, 剩下的恶意样本均归类为其他类型。表 7 显示了模型对几种不同类型恶意 DOC 样本的检测情况。可以看出, 模型对于利用漏洞的恶意 DOC 文档检测率为 97.55%, 对利用宏的恶意 DOC 文档检测率为 99.23%, 对这两类主要威胁均能保持较强的检测效果。

表 7 模型对不同类型恶意 DOC 样本的检测情况

恶意 DOC 样本类型	数量	检出数量	检测率
漏洞利用文档	937	914	97.55%
恶意宏文档	3236	3211	99.23%
其他类型	145	139	95.86%
合计	4318	4264	98.75%

作为一种不考虑恶意文档类型的深度学习检测模型, 其在检测恶意宏文档上与专门检测恶意宏的模型相比如何? 从表 7 已经知道了模型对包含宏的恶意 DOC 样本的检测率, 表 8 进一步分析了模型对按是否包含宏分类后良性 DOC 样本的误报率情况。可以看出, 虽然模型对 DOC 样本整体的误报率较低, 仅为 0.67%, 但其对良性宏文档的误报率高达 5.4%。表 9 中对比了本文深度学习检测模型与相关恶意宏检测研究结果比较。可以看出, 模型 5.4%的误报率与 Chen 等人^[22]恶意宏检测模型 1.5%的误报率相比仍然高出很多, 在检测率方面也略低于其 99.7%的检测率。这表明上述深度学习检测模型虽然能够检测多种类型的恶意 DOC 文档, 但在基于宏的恶意 DOC 文档检测上不如人工提取宏特征进行检测的机器学习模型。本文认为出现这种现象是容易理解的, 深度学习模型自动从文档目录项区域提取的检测特征只能获得宏的名称、数量、长度、时间等信息, 并没有触及恶意宏的本质特征—宏代码本身。

表 8 模型对良性 DOC 样本的误报情况

Table 8 False alarm rate of our model on benign DOC			
良性 DOC 样本类型	数量	误报数量	误报率
包含宏的文档	719	39	5.4%
不含宏的文档	7820	18	0.2%
合计	8539	57	0.67%

表 9 与现有恶意宏检测相关研究的检测结果比较

Table 9 Comparison of detection results with existing research related to malicious macro detection				
研究	检测方法	数据集	检测率	误报率
本文	字节序列+深度学习	良性宏 3549 个, 恶意宏 17641 个	99.2%	5.4%
Vasilios ^[21]	宏关键字特征+机器学习	良性宏 2736 个, 恶意宏 15571 个	97.6%	3.9%*
Chen ^[22]	宏关键字特征、混淆特征+机器学习	良性宏 2939 个, 恶意宏 13734 个	99.7%	1.5%

*根据文章提供的结果计算得出

6.2 恶意 XLS 文档检测

(1) 模型检测结果

针对恶意 XLS 文档检测的深度学习模型测试结果如表 10 所示, 从表中可以看出, 随着截取 XLS 文件头部字节流送入深度学习模型的长度从 4KB 增加到 16KB、64KB, 模型的检测性能不仅没有提升, 反而逐渐有所下降, 这个现象与深度学习模型在恶意 DOC 文档上的检测表现是一致的, 再次表明送入模型的数据长度并不是越多越好。

表 10 深度学习模型在 XLS 样本集上的测试结果

Table 10 Detection results of deep learning model on XLS samples					
	模型	精度	召回率	准确率	F1 分数
手动提取	CNN-XLS-HE-4096	97.83%	97.58%	98.41%	97.71%
	CNN-XLS-HE-16384	97.32%	97.62%	98.25%	97.47%
	CNN-XLS-HE-65536	95.83%	97.52%	97.68%	96.67%
	CNN-XLS-DE2048	96.88%	96.96%	97.86%	96.92%
	CNN-XLS-DE4096	96.63%	96.79%	97.71%	96.71%
自动提取	CNN-XLS-DE2048-SS16_128	98.75%	98.38%	99.00%	98.56%
	CNN-XLS-DE-2048-WORKBOOK-1024-DOCUMENTSUMMARYINFORMATION-1024	98.74%	98.10%	98.90%	98.42%

通过对比模型 CNN-XLS-DE2048、CNN-XLS-DE4096 和 CNN-XLS-DE2048-SS16_128 的测试结果可以看到, 在选取 XLS 文档的目录项区域的基础上, 加入 workbook 流中的 globals 子流、sheets 子流的头

部数据信息作为模型输入,模型的精度和召回率均明显提高, F1 分数从 96.92%提升至 98.56%, 表明一些恶意 XLS 样本的恶意特征不会体现在文档的目录项区域, 而是隐含在更深层次的子流信息中, 子流信息头部也是模型检测恶意 XLS 文档的“敏感数据区”。

采用本文提出的自动“敏感数据区”提取方法发现的“敏感数据区”由 2048 字节目录项区域、1024 字节 WorkBook 流和 1024 字节 DocumentSummary Information 流构成, 与针对 DOC 文档类型识别的“敏感数据区”不同, 并没有包含文档头部。提取这些数据构建的深度学习检测模型检测准确率为 98.90%, F1 分数为 98.42%, 均略低于基于手动提取“敏感数据区”的 CNN-XLS-DE2048-SS16_128 模型, 表明上述自动“敏感数据区”提取方法未能发掘和提取到这些敏感子流数据信息。

(2) 检测结果对比

由于未见专门针对恶意 XLS 文档检测的相关研究, 本文将模型检测结果与杀毒软件进行对比。

该过程与恶意 DOC 检测结果对比一致, 随机选择 5 折交叉验证中其中一次训练集和测试集划分, 将上述模型 CNN-XLS-DE2048-SS16_128 的检测结果与杀毒软件的检测结果进行对比。测试集包括 2900 个恶意 XLS 样本和 5509 个良性样本, 2900 个恶意样本中有 2801 个样本被 5 个杀毒软件标记为恶意。图 12 显示了本文模型和检出率前 10 的杀毒软件的检测率对比情况, 可以看出, 本文检测模型对测试集中的恶意 XLS 样本的检测能力明显优于各大杀毒软件, 而此时模型对良性 XLS 样本的误报率仅为 0.71%。在检测时间开销方面, XLS 样本的平均检测时间开销为 128 毫秒。

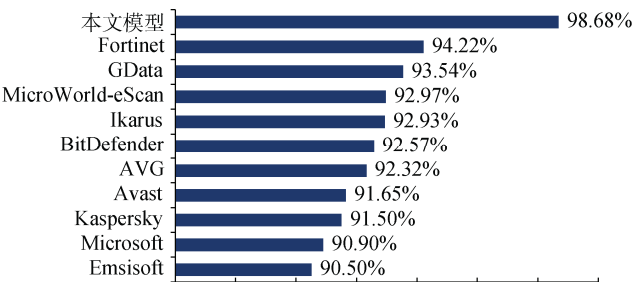


图 12 模型与杀毒软件在恶意 XLS 样本上检测率比较
Figure 12 Comparison of detection rates between our model and antivirus software on malicious XLS samples

(3) 分类型检测结果

为了进一步分析模型对不同类型恶意 XLS 文档的检测效果, 基于 Virustotal 提供的样本检测标签信

息, 模型对测试集中漏洞利用 XLS 文档、恶意 XLS 宏文档和其他类型恶意 XLS 文档的检测结果如表 11 所示。表 12 进一步显示了模型对良性 XLS 文档的误报情况。

表 11 模型对不同类型恶意 XLS 样本的检测情况
Table 11 Detection results of our model on different types of malicious XLS samples

恶意 XLS 样本类型	数量	检出数量	检测率
漏洞利用文档	591	581	98.31%
恶意宏文档	1268	1248	98.42%
其他类型	942	935	99.26%
合计	2801	2764	98.68%

表 12 模型对包含宏的良性 XLS 样本的误报情况
Table 12 False alarm rate of our model on benign DOC

良性 XLS 样本类型	数量	误报数量	误报率
包含宏的文档	387	35	9.04%
不含宏的文档	5122	4	0.08%
合计	5509	39	0.71%

可以看到, 尽管模型能够在整体误报率仅为 0.71%的前提是实现对恶意漏洞利用 XLS 文档、恶意宏 XLS 文档和其他类型恶意文档的有效检测, 检出率分别为 98.31%、98.42%和 99.26%, 但仅就恶意宏文档的检测而言, 其误报率却高达 9%, 这也再次表明模型并没有学习到恶意宏的本质特征, 检测效果不如基于人工提取恶意宏特征的机器学习模型。

6.3 实验分析与讨论

通过实验能够看到, 在将深度学习技术应用到恶意文档检测上时, 改进输入模型的数据比改进模型结构^[34]更加重要。本文方法仅将大量 DOC/XLS 文档样本的特定数据区域及标签送入深度学习模型进行学习, 不需要人工提取检测特征, 模型在恶意 DOC、XLS 文档检测上取得的检测效果均明显优于现有深度学习方法, 即使与机器学习方法相比, 也具备检测优势。

从攻击者构建恶意样本绕过模型检测的角度看, 虽然本文方法提取了文档特定区域的字节数据作为模型输入, 但深度学习模型作为一个黑盒, 理论上对于攻击者而言, 想要构建能够绕过深度学习检测模型且符合文档规范的恶意文档, 其难度显然高于绕过传统基于机器学习的检测模型。

深度学习技术应用到恶意 Office 文档检测中的优势是降低了对人工提取特征的要求, 在不知道如何提取或者难以提取恶意特征, 又有大量正负样本

的场景是深度学习技术应用的最合适场景, 其中选择文档的哪些字节数据区域作为模型输入又至关重要。输入深度学习检测模型的数据过多会导致模型无法聚焦其中的恶意元素, 而输入模型的数据过少也会导致恶意元素丢失, 最终都会造成模型检测能力的下降。因此, 是否能够提取样本的所有包含恶意元素的“敏感数据区”并限制在恰当的长度范围内, 直接决定了模型能够检测恶意样本的类型和检测率, 对于无法满足上述要求的场景下深度学习模型表现不佳。例如, 当前“敏感数据区”并没有涵盖宏内容本身, 使得在恶意宏检测上基于人工特征的机器学习检测模型仍然保持着优势, 这也表明在恶意文档检测上结合机器学习检测方法和深度学习检测方法的必要性。后续研究可以进一步围绕系统探索文档的“敏感数据区”、深度学习模型的抗对抗样本攻击能力展开。

7 结论

本章以探索深度学习技术在恶意 Office 文档检测中的适用性为目标, 将深度学习技术应用到恶意 DOC/XLS 文档检测中, 提出了基于“敏感数据区”的恶意 DOC 深度学习检测方法和恶意 XLS 深度学习检测方法, 通过实验结果和对比分析表明, 本文提取“敏感数据区”的方法能够显著提升模型检测能力, 并进一步分析和讨论了深度学习技术在恶意 Office 文档检测中的应用方式、优势及不足, 为深度学习技术在恶意文档检测上的拓展深化指明了方向。

参考文献

- [1] 2022-mid-year-trends[EB/OL]. <https://go.checkpoint.com/2022-mid-year-trends/page-global-malware-statistics.php#5>.
- [2] Mimura M, Ohminami T. Using LSI to Detect Unknown Malicious VBA Macros[J]. *Journal of Information Processing*, 2020, 28: 493-501.
- [3] Kim S, Hong S, Oh J, et al. Obfuscated VBA Macro Detection Using Machine Learning[C]. *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, 2018: 490-501.
- [4] Koutsokostas V, Lykousas N, Apostolopoulos T, et al. Invoice #31415 Attached: Automated Analysis of Malicious Microsoft Office Documents[J]. *Computers & Security*, 2022, 114: 102582.
- [5] Chen X, Wang W B, Han W T. Malicious Office Macro Detection: Combined Features with Obfuscation and Suspicious Keywords[J]. *Applied Sciences*, 2023, 13(22): 12101.
- [6] Bearden R, Lo D C. Automated Microsoft Office Macro Malware Detection Using Machine Learning[C]. *2017 IEEE International Conference on Big Data*, 2017: 4448-4452.
- [7] Huneault-Leblanc S, Talhi C. P-Code Based Classification to Detect Malicious VBA Macro[C]. *2020 International Symposium on Networks, Computers and Communications*, 2020: 1-6.
- [8] Srdic, Nedim and Pavel Laskov. Detection of Malicious PDF Files Based on Hierarchical Document Structure[C]. *The Network and Distributed System Security Symposium*, 2013:1-16.
- [9] Salah S, Lazrek A, and El Bakkali Z. A Structural and Content-Based Approach for a Precise and Robust Detection of Malicious PDF Files[C]. *The 1st International Conference on Information Systems Security and Privacy*, 2015: 27-36.
- [10] Nissim N, Cohen A, Elovici Y. ALDOCX: Detection of Unknown Malicious Microsoft Office Documents Using Designated Active Learning Methods Based on New Structural Feature Extraction Methodology[J]. *IEEE Transactions on Information Forensics and Security*, 2017, 12(3): 631-646.
- [11] Lu X F, Wang F, Shu Z F. Malicious Word Document Detection Based on Multi-View Features Learning[C]. *2019 28th International Conference on Computer Communication and Networks*, 2019: 1-6.
- [12] Maiorca D, Corona I, Giacinto G. Looking at the Bag Is Not enough to Find the Bomb: An Evasion of Structural Methods for Malicious PDF Files Detection[C]. *The 8th ACM SIGSAC Symposium on Information, Computer and Communications Security*, 2013: 119-130.
- [13] Nissim N, Cohen A, Glezer C, et al. Detection of Malicious PDF Files and Directions for Enhancements: A State-of-the Art Survey[J]. *Computers & Security*, 2015, 48: 246-266.
- [14] Gaber M G, Ahmed M, Janicke H. Malware Detection with Artificial Intelligence: A Systematic Literature Review[J]. *ACM Computing Surveys*, 2024, 56(6): 1-33.
- [15] Vinayakumar R, Alazab M, Soman K P, et al. Robust Intelligent Malware Detection Using Deep Learning[J]. *IEEE Access*, 2019, 7: 46717-46738.
- [16] Jian Y F, Kuang H B, Ren C L, et al. A Novel Framework for Image-Based Malware Detection with a Deep Neural Network[J]. *Computers & Security*, 2021, 109: 102400.
- [17] Chalapathy R, Chawla S. Deep Learning for Anomaly Detection: A Survey[EB/OL]. 2019: arXiv: 1901.03407. <https://arxiv.org/abs/1901.03407>.
- [18] Wang Z D, Liu Y D, He D J, et al. Intrusion Detection Methods Based on Integrated Deep Learning Model[J]. *Computers & Security*, 2021, 103: 102177.
- [19] Agrawal R, Stokes J W, Marinescu M, et al. Neural Sequential Malware Detection with Parameters[C]. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018: 2656-2660.
- [20] Fettaya R, Mansour Y. Detecting Malicious PDF Using CNN[EB/OL]. 2020: arXiv: 2007.12729. <https://arxiv.org/abs/2007.12729>.
- [21] Nichols T, Zemlanicky J, Luo Z R, et al. Image-Based PDF Malware Detection Using Pre-Trained Deep Neural Networks[C]. *2024 12th International Symposium on Digital Forensics and Security*, 2024: 1-5.
- [22] Yang S J, Chen W B, Li S X, et al. Approach Using Transforming Structural Data into Image for Detection of Malicious MS-DOC

- Files Based on Deep Learning Models[C]. *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2019: 28-32.
- [23] Jeong Y S, Woo J, Kang A R. Malware Detection on Byte Streams of Hangul Word Processor Files[J]. *Applied Sciences*, 2019, 9(23): 5178.
- [24] Jeong Y S, Woo J, Lee S, et al. Malware Detection of Hangul Word Processor Files Using Spatial Pyramid Average Pooling[J]. *Sensors*, 2020, 20(18): 5265.
- [25] Jeong Y S, Mswahili M E, Kang A R. File-Level Malware Detection Using Byte Streams[J]. *Scientific Reports*, 2023, 13: 8925.
- [26] Li W J, Stolfo S, Stavrou A, et al. A Study of Malcode-Bearing Documents[M]. *Detection of Intrusions and Malware, and Vulnerability Assessment*. Berlin, HeidelbergSpringer2007: 231-250.
- [27] Liu L P, He X H, Liu L, et al. Capturing the Symptoms of Malicious Code in Electronic Documents by File's Entropy Signal Combined with Machine Learning[J]. *Applied Soft Computing*, 2019, 82: 105598.
- [28] Bai Peng, Hu Ying, Dai Fangfang. Malicious Document Detection Based on Shellcode Detection[C]. *Proceedings of the 19th national youth communication Academic Annual Conference*, 2014.
- [29] Chen X, Yi P, Bai B, et al. Unknown Malicious Document Detection Based on Global Behavior Feature[J]. *Journal of Cyber Security*, 2023, 8(5): 96-108.
- (陈祥, 伊鹏, 白冰, 等. 基于全局行为特征的未知恶意文档检测[J]. *信息安全学报*, 2023, 8(5): 96-108.)
- [30] Jiang J G, Wang C H, Yu M, et al. NFDD: A Dynamic Malicious Document Detection Method without Manual Feature Dictionary[M]. *Wireless Algorithms, Systems, and Applications*. Cham: Springer International Publishing2021: 147-159.
- [31] Ruaro N, Pagani F, Ortolani S, et al. SYMBEXCEL: Automated Analysis and Understanding of Malicious Excel 4.0 Macros[C]. *2022 IEEE Symposium on Security and Privacy*, 2022: 1066-1081.
- [32] MS-CFB:Compound File Binary File Format | Microsoft Learn. https://learn.microsoft.com/en-us/openspecs/windows_protocols/ms-cfb/53989ce4-7b05-4f8d-829b-d08d6148375b.
- [33] MS-DOC:Word (.doc) Binary File Format | Microsoft Learn. [2024-08-26].https://learn.microsoft.com/en-us/openspecs/office_file_formats/ms-doc/ccd7b486-7881-484c-a137-51170af7cc22.
- [34] MS-XLS:Excel Binary File Format (.xls) Structure | Microsoft Learn.[2024-08-26].https://learn.microsoft.com/en-us/openspecs/office_file_formats/ms-xls/cd03cb5f-ca02-4934-a391-bb674cb8aa06.
- [35] maldoc_detection_sda. https://gitee.com/chensheng101/maldoc_detection_sda.
- [36] Cisco Popularity List.<https://s3-us-west-1.amazonaws.com/umbrella-static/index.html>.



陈祥 于 2014 年在国防科技大学计算机专业获得硕士学位。现任信息工程大学信息技术研究所助理研究员。研究领域为入侵检测、恶意软件检测、智能渗透。Email: chenxndsc@163.com



张震 于 2012 年在信息工程大学通信与信息系统专业获得博士学位, 现任信息工程大学副教授, 研究领域为网络内生安全、群体智能与协同控制。Email: zhangzhen2096@163.com



王文博 于 2016 年在信息工程大学信息与通信工程专业获得硕士学位, 现在信息工程大学网络空间安全专业攻读博士学位, 研究领域为 APT 检测、攻防对抗等。Email: wbwangedu@outlook.com