DOI:10.19651/j.cnki.emt.2209491

基于深度学习的恶意文档可视化检测*

黄 昆 徐 洋 张思聪 李克资
(贵州师范大学贵州省信息与计算科学重点实验室 贵阳 550001)

摘 要：为了更加准确、快速地检测恶意 PDF 与 DOCX 格式文档，提出一种基于深度学习的恶意文档可视化检测方法。该方法通过马尔可夫模型将文档的字节序列转化为三通道的彩色图，从而获取更能区分恶意文档和良性文档的视觉表征，并采用当前主流的 EfficientNet-B0 模型对提取的可视化特征进行分类。结合迁移学习领域中的微调技术，将 ImageNet 上的分类权重应用到 EfficientNet-B0 模型的训练中，加快检测模型的收敛速度，缩短模型的训练时间。实验证明，在两个数据集上，模型的收敛速度快于随机初始化权重的预训练，且模型对恶意 PDF 文档和恶意 DOCX 文档的检测准确率分别达到了 99.80%和 98.14%，优于 ResNet34、MobileNetV2 等模型。与主流的恶意文档检测工具 Wepawet 和 PJScan 相比，所提出的方法具有更优的综合检测性能，进一步验证了所提出方法对恶意文档检测的有效性。

关键词：恶意文档；EfficientNet-B0；可视化；马尔可夫模型；迁移学习
中图分类号：TP393.08 **文献标识码**：A **国家标准学科分类代码**：520.2010

Visual detection of malicious document based on deep learning

Huang Kun Xu Yang Zhang Sicong Li Kezi
(Key Laboratory of Information and Computing Science of Guizhou Province, Guizhou Normal University, Guiyang 550001, China)

Abstract: In order to detect malicious PDF and DOCX format documents more accurately and quickly, a visual detection method of malicious documents based on deep learning is proposed. This method converts the byte stream of the document into a three-channel color image through the Markov model, so as to obtain a visual representation that can better distinguish between malicious documents and benign documents, and uses the current mainstream EfficientNet-B0 model to extract visual features to classify. Combined with the fine-tuning technology in the field of transfer learning, the classification weights on ImageNet are applied to the training of the EfficientNet-B0 model, which speeds up the convergence of the detection model and shortens the training time of the model. Experiments show that on two datasets, the convergence speed of the model is faster than the pre-training of random initialization weights, and the detection accuracy of the model for malicious PDF documents and malicious DOCX documents reaches 99.80% and 98.14%, respectively, which is better than models such as ResNet34 and MobileNetV2. Compared with the mainstream malicious document detection tools Wepawet and PJScan, the proposed method has better comprehensive detection performance, which further verifies the effectiveness of the proposed method for malicious document detection.

Keywords: malicious document; EfficientEet-B0; visualization; markov model; transfer learning

0 引 言

近年来，针对政府机构和商业组织的高级可持续威胁(advanced persistent treat, APT)时有发生，严重危害了国家的关键信息基础设施和泄露政府企业的敏感信息。相对于可执行文件，人们对恶意文档的防范意识要低很多，因此恶意文档常被作为实施 APT 攻击的主要载体[1]。

2016 年，PaloAlto 研究中心的安全研究所发现了针对苹果系统的钓鱼邮件攻击，其攻击方法是在邮件附件中添加带有木马的恶意 PDF 文档。2017 年，Hades 组织利用鱼

收稿日期：2022-04-02
* 基金项目：中央引导地方科技发展专项资金(黔科中引地〔2018〕4008)、贵州省科技计划项目(黔科合支撑[2020]2Y013 号)、贵州省研究生科研基金(黔教合 YJSKYJJ〔2021〕102)项目资助

叉邮件投递内嵌恶意宏的 Word 文档,导致了韩国平昌冬奥会的网站宕机。2018 年 OceanLotus 组织,利用水坑攻击和鱼叉邮件方式,投递内嵌恶意宏的 Word 文档,对我国和东南亚其他国家进行了全年频繁的针对攻击。赛门铁克的《2016 年互联网安全威胁报告》的研究表明,PDF 文档和 Word 文档已经成为 2016 年特定目标攻击中钓鱼邮件附件最多的文档格式,分别占比为 60.1%和 38.7%。赛门铁克的《2019 年互联网安全威胁报告》的研究表明,2018 年中,恶意电子邮件附件中有 48%使用 Office 文件作为附件。

恶意文档的检测已经成为热门领域,许多新的研究思路和研究方法不断被提出[2]。主要包含静态检测方法、动态检测方法和动静态结合检测方法。

1)静态检测方法:文档的静态检测方法主要是针对文档的内容特征和结构特征,如 2016 年,Šrndic 等[3]开发 Hidost 模型,用于提取 Office 和 PDF 文档的元数据特征和结构路径特征,并基于支持向量机设计了分类模型。但难以提取深入的元数据特征和结构特征,造成检测能力欠缺。2021 年,Li 等[4]在 Hidost 模型的基础上开发了 Hivost 模型,提取文档的结构特征,使用支持向量机和随机森林作为分类模型,结合主动学习,提升模型检测准确率并减少了训练时间。2021 年,Lu 等[5]针对 PDF、Word、Excel 等多种文档格式设计出通用静态检测框架,其核心特征是文档的结构路径、代码关键字和对象数,该方法在两个数据集上均具有良好的检测精度,但提取文档的代码关键字和对象数量较为繁琐。

2)动态检测方法:动态检测方法主要是在虚拟环境或沙箱中打开文档,通过监控文档的运行状况分析是否为恶意文档。如 2017 年,Xu 等[6]通过在多个不同的操作系统上使用同一个 PDF 阅读器打开 PDF 文档,然后监控 PDF 文档调用的进程,观察 PDF 文档在不同操作系统上的行为。实验证明,良性 PDF 文档具有许多相似行为,而恶意 PDF 文档往往具有不同的行为。动态分析能够直观的了解到恶意 PDF 文档的恶意行为及目的,具有较好的健壮性,但检测成本高,消耗大量的内存资源,且由于反虚拟化技术的发展,使得监控恶意行为的难度加大。

3)动静态结合分析:指将静态内容特征、结构特征等与动态行为特征结合起来进行恶意文档检测。如 2019 年,杜学绘等[7]通过静态分析技术从 PDF 文档中提取常规信息与结构信息,通过动态分析技术从 PDF 文档中提取 API 调用信息,并基于 K-means 算法提取核心混合特征,最后构建随机森林分类器进行分类,该方法具有较高的检测准确率,同时应用特征融合,使得模型的鲁棒性有一定提升。但是该方法对训练数据的质量要求过高,且存在无法解析的文档,需要额外的筛选和剔除。

近年来,深度学习结合恶意代码可视化的研究得到了迅速发展,人们对其有了广泛的研究。2018 年,Ni 等[8]将恶意软件反汇编文件转化为灰度图,并通过卷积神经网络进行恶意软件家族分类,实验证明,该方法识别准确率高,并且转化为图像的速度快,提高了检测效率,可用于自动化和大规模的恶意代码检测。2020 年,卢喜东等[9]将恶意软件映射为无压缩的灰度图像,再将灰度图像裁剪为恒定大小的图像,最后使用深度森林进行恶意软件家族分类,该方法具有良好的检测效率和精度。2020 年,Ren 等[10]将恶意软件的字节序列通过空间填充曲线的方法转化为彩色图,并使用深度学习模型进行分类,该方法检测准确率高于常用的转化为灰度图的方法。2021 年,Ap 等[11]将恶意软件可视化成灰度图、彩色图和单通道马尔可夫图,使用 Gabor 滤波器提取三种图像的特征并融合,实验证明,该方法在小样本数据集上同样具有较高的检测准确率。

传统静态检测方法大多需要逆向工程,严重影响了恶意代码识别的效率,而动态检测方法则资源消耗大、检测成本高,不利于大规模的恶意代码检测[12]。基于机器学习和深度学习的恶意文档检测能够及时检测新型恶意文档并且快速更新检测模型,但机器学习方法存在人工提取特征繁琐、特征鲁棒性差导致检测准确率低的问题[13]。而在恶意软件可视化的研究领域,恶意软件的原始大小存在差异,直接将二进制文件转化为灰度图的方法会导致图像大小不一致,而对图像的缩放或裁断会导致部分信息丢失,从而影响方法的检测效率和准确率。且针对恶意文档检测,可视化方法尚未有很好的应用。

为此,本文提出一种基于深度学习和可视化技术的恶意文档检测方法,将文档转化为三通道彩色图,同时使用轻量级卷积神经网络 EfficientNet-B0[14],结合迁移学习[15]领域中的微调技术,将 ImageNet 上的分类权重应用到 EfficientNet-B0,实现了对恶意文档的检测。使用本文方法分析文档,无需动态分析,也不用逆向分析,仅依赖文档的二进制字节序列,且不受原始文件大小的影响,计算效率高,更适合大规模的恶意文档检测。

1　本文方法

本文基于马尔可夫模型和 EfficientNet-B0 模型提出了一种新的恶意文档可视化检测方法,所提出的方法首先将文档字节序列转化为三种马尔可夫概率矩阵,再将三种矩阵分别转化为图像的三通道,最终合成三通道的马尔可夫彩色图,以此表征出利于模型分类的恶意文档的可视化特征;然后,采用 EfficientNet-B0 结合迁移学习中的微调技术构建深度学习模型,将彩色图进行归一化预处理后作为模型的输入,训练模型自动提取图像的深层特征;最后,模型通过对彩色图的深层特征进行学习筛选,利用 Softmax 分类器进行预测输出。本文所提出方法的恶意文档检测流程如图 1 所示,主要分为数据预处理、模型训练及测试、预测分类三部分。

1.1　文档可视化方法

1)灰度图

灰度图方法为恶意代码可视化领域使用最为广泛的一

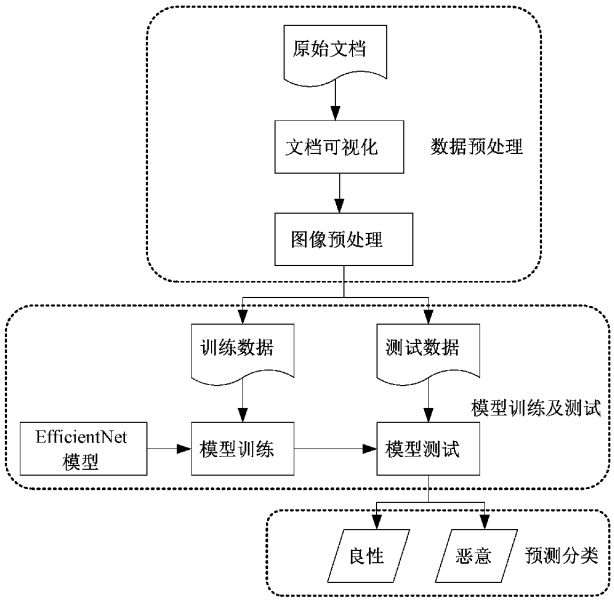

图 1　恶意文档可视化检测流程

种方法，如图 2 所示，该方法以二进制形式打开文档，以 8 位二进制数为一个无符号整数作为一个像素点，范围是为 0～255，同时根据经验，结合表 1 设定图像初始宽度，图像高度根据文档大小自适应，将文档转化成灰度图。由于文档转化为灰度图后具有独特的纹理特征，良性文档的灰度图往往具有相似的纹理，而恶意文档往往具有更多样化的纹理特征，与良性文档的纹理存在一定差异，并且深度学习模型能提取图像更深层的特征，因此文档转化为灰度图后，会利于模型分类。

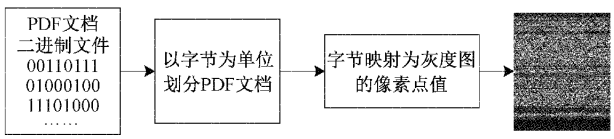

图 2　文档转灰度图流程

表 1　图像初始宽度

样本大小/KB	图像初始宽度
小于 10	32 像素
10～30	64 像素
30～60	128 像素
60～100	256 像素
100～200	384 像素
200～500	512 像素
500～1 000	768 像素
大于 1 000	1 024 像素

2）马尔可夫图

由于文档二进制文件的字节值直接用作图像中的像素点，图像尺寸受原始文档大小影响，会导致大量信息冗余，并且图像尺寸大小相差较大，对图像的预处理会导致信息丢失，影响最终检测准确率。

而马尔可夫图可以解决上述问题，根据图 3 所述过程，首先，将恶意文档的视为具有时序特征的字节流，8 位二进制数为一个无符号整数，范围是为 0～255，遍历文档的所有字节，统计每种取值的次数；其次，用 $P_{i,j}$ 表示字节值 i 后续第一个字节为字节值 j 的转移概率，其计算方法为式(1)：

$$P_{i,j}=\frac{W_{i,j}}{\sum_{k=0}^{255}W_{i,k}} \tag{1}$$

其中，$W_{i,j}$ 在 PDF 文档字节序列中的字节值 i 后续是字节值 j 出现的次数；最后，根据转移概率 $P_{i,j}$ 生成状态转移概率矩阵，每个 $P_{i,j}$ 对应一个像素值，生成 256×256 大小的单通道马尔可夫图。

同理，根据图 4 所述过程，利用式(1)，可计算出字节值 i 后续第 2 个字节为字节值 m 的状态转移概率矩阵，以及字节值 i 后续第 3 个字节为字节值 n 的状态转移概率矩阵，再将 3 个矩阵分别填充为彩色图的 RGB 三个通道的像素值，由此转化为三通道的马尔可夫彩色图。

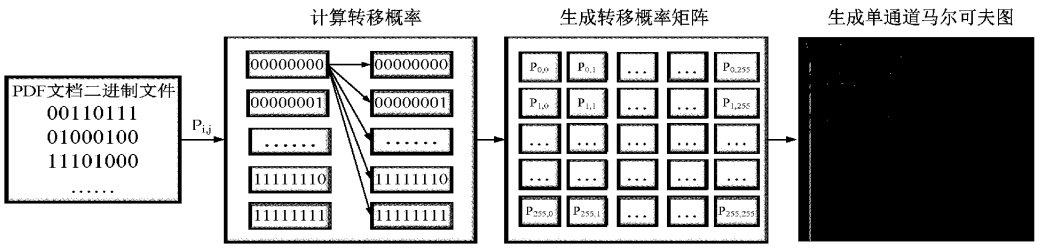

图 3　文档转单通道马尔可夫图流程

虽然单通道马尔可夫图大部分为偏黑色的区域，但是仍能看出图像在一些区域为白色线条和灰色点。而马尔可夫彩色图则在相似的位置上布满了红色和蓝色等颜色的线条和像素点。这说明文档的语义信息、结构信息表现成了图像像素的相似性和相异性，可作为恶意文档检测的有效特征。而三通道马尔可夫彩色图理论上包含了单通道马尔可夫图的信息，当某个通道失效或错误时，另外两个通道可以起到补充和纠正分类结果的作用。相较于灰度图，所有由原始恶意软件文件转化的马尔可夫图均是同一尺寸大小，减少了原始文件大小带来的影响，理论上具有更高的检测效率和准确率。

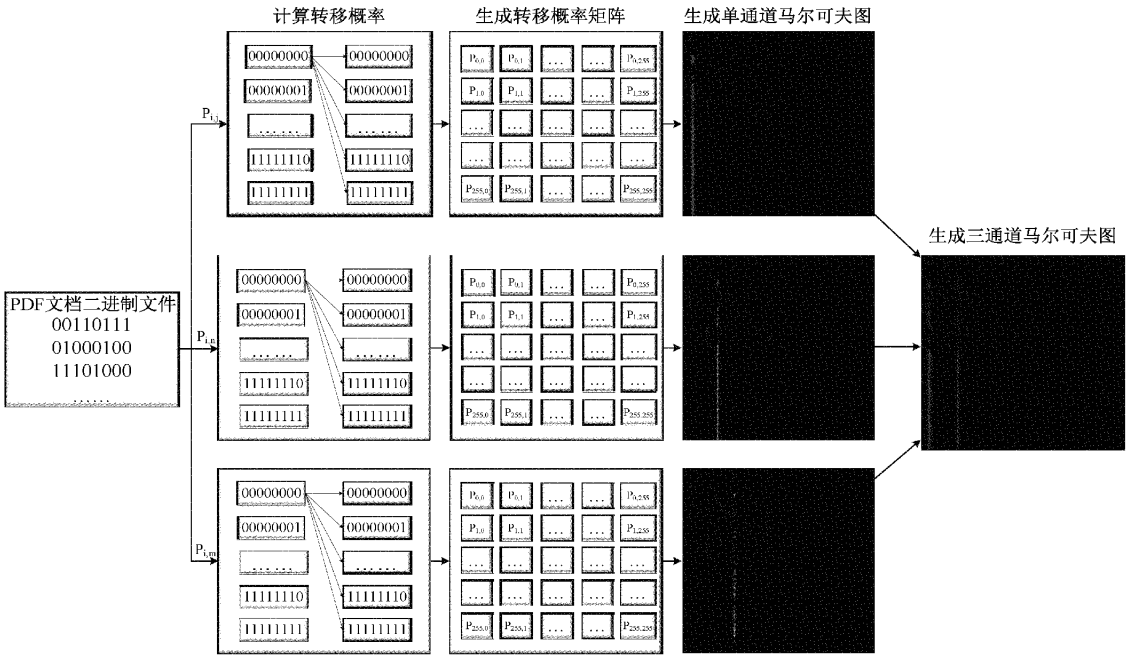

图 4　文档转三通道马尔可夫图流程

1.2　迁移学习

在 2009 年,Pan 等[15]发表了对迁移学习的研究成果,就是运用已学习的知识来帮助学习新的知识,其根本就是找到新旧知识之间的关联性和相似性。由于直接对新领域从头开始学习进度太慢,所以转向运用已有的相关知识来辅助尽快地学习新知识。

在深度学习领域,由于大部分数据或任务是存在相关性的,所以通过迁移学习可以将已经学到的模型参数通过适当修改迁移到新模型上,从而加快并优化模型的学习效率[16],而不用像随机初始化权重的预训练那样从零开始训练整个网络。本文采用的就是迁移学习领域中的微调技术,具体操作为先将模型的全连接层改造成需要的分类数,在新模型训练之前,将权重加载到模型,加载后可以选择是否冻结某些层,被冻结的层不会更新参数。我们选择不冻结任何层,就可以在已有的权重基础上,根据自己的数据集训练模型各个层,使得模型进一步收敛。虽然本模型主要是对文档图像的分类,与 ImageNet 数据集的种类并不相关,但是由于图像分类任务具有很强的任务相似性,且 ImageNet 数据集的训练权重是通过海量数据训练得来的,对样本数少的数据集非常友好,在一定程度上有助于缩短本文所提出方法的训练时间。

1.3　EfficientNet 模型设计

EfficientNet[14]是 2019 年由 Google 提出来的新型轻量级卷积神经网络,依据网络规模的不同可以分为 B0～B7 版本,EfficientNet-B7 在 ImageNet top-1 上达到了当年最高准确率 84.4%,与之前准确率最高的 GPipe 相比,参数数量(Params)仅为其 1/8.4,推理速度提升了 6.1 倍。此外,它的应用场景非常广泛,如 Ramanna 等[17]利用 EfficientNet-B4 对道路状况进行分类;Yousfi 等[18]将 EfficientNet 应用到图像隐写分析;王宸等[19]改进 EfficinetNet-B2 用于对两种锻件的荧光磁粉探伤图像进行检测;廖海斌等[20]使用 EfficinetNet-B3 对人脸表情图像进行学习和分类。在上述场景中 EfficientNet 均取得了较好的效果,因此,EfficientNet 具有广泛的应用场景和较大的应用潜力,但该模型在恶意文档检测领域尚未被很好的利用。

本文针对恶意文档检测领域的特点,对 EfficientNet-B0 模型进行适当的改进,将其全连接层的输出改为 2,以适用恶意文档的二分类检测。为了比较模型对相同尺寸图像分类的效果,将输入图片的分辨率固定为 224×224×3(单通道的马尔可夫图和灰度图将复制两次填充成三通道,以满足模型输入要求),并在训练之前将 EfficientNet-B0 在 ImageNet 上训练好的权重载入模型。本文使用的 EfficientNet-B0 的模型结构如图 5 所示,模型主要包含 3 个模块,第一个模块是一个卷积核大小为 3×3 步长为 2 的卷积层(包含 BN 层和激活函数 Swish),核心模块是重复堆叠的 MBConv[21](mobile inverted bottleneck Conv),MBConv 后面的数字 1 或 6 代表每个模块对输入特征矩阵的通道扩张倍数,3×3 或 5×5 代表 Depthwise Conv 的卷积核大小,MBConv 的结构如图 6 所示。经过一系列的卷积操作,输出 1 280 维的特征,最后经过 Softmax 分类器输出图像所属的类别。

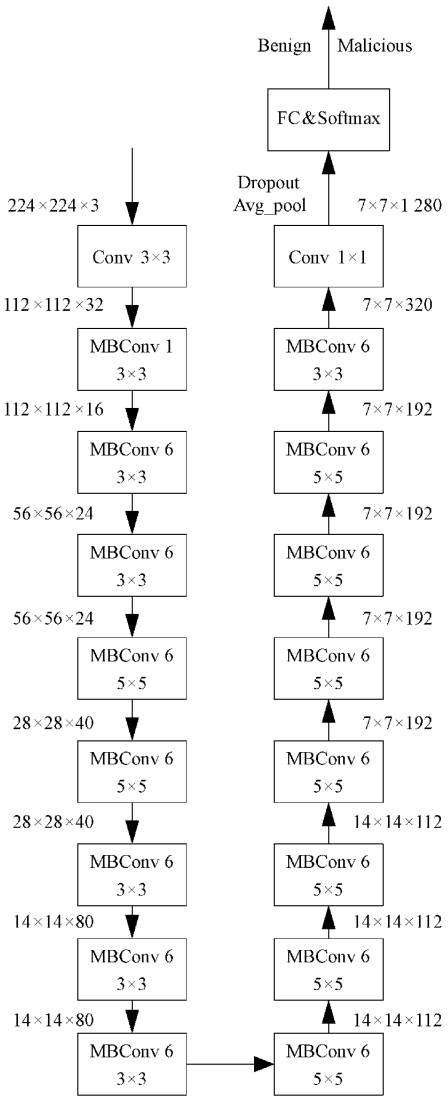

图 5　改进后的 efficientNet-B0 的结构

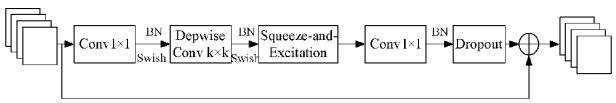

图 6　MBConv 的结构

2　实验分析

2.1　数据集

本文使用了 PDF 文档和 DOCX 文档两个数据集，其中 PDF 文档数据集来源于 Contagio 团队，包含 11 980 个恶意文档和 9 000 个良性文档；DOCX 文档数据集包含 1 920 个恶意文档和 4 000 个良性文档，恶意文档来源于 VirusShare，良性文档来源于百度文库。采取 5 折交叉验证方式，按比例随机抽取 80%的样本作为训练集，其余 20%作为测试集。实验是在 CPU 主频为 3.6 GHz 的 Intel i7-11700KF，显卡为显存 8 GB 的 NVIDIA GeForce RTX 3070Ti，内存为 32 GB 的 Windows10 系统上完成的。

2.2　评价指标

在二分类问题中常用的分类准确性指标有准确率(accuracy)、精确率(precision)、召回率(recall)、F1-score、真阳性率(TPR)、假阳性率(FPR)、ROC 曲线下的面积(area under curve，AUC)等。其计算方法为式(2)～(7)，变量定义表如表 2 所示。

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

$$\text{Precison} = \frac{TP}{TP + PF} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

$$F1 - \text{score} = 2 \times \frac{\text{Precison} \times \text{Recall}}{\text{Precison} + \text{Recall}} \tag{5}$$

$$TPR = \frac{TP}{TP + TN} \tag{6}$$

$$FPR = \frac{FP}{TN + FP} \tag{7}$$

表 2　变量定义表

	检测为恶意	检测为良性
恶意文档	*TP*	*FN*
良性文档	*FP*	*TN*

本文选取准确率、Recall、Precision、F1-score 和 AUC 作为实验主要评价指标。准确率为最常用的评价指标；F1-score 兼顾精确率和召回率，是精确率和召回率的调和平均值，只有当精确率和召回率都高的时候，F1-score 才会高；AUC 为 ROC 曲线下的面积，ROC 曲线的横坐标为 FPR，纵坐标为 TPR，当测试集的样本分布发生变化，ROC 曲线能保持不变，ROC 对应的 AUC 越大，说明分类性能越好。

2.3　实验结果

1) 可视化方法及迁移学习对比

将 PDF 文档和 DOCX 文档可视化为灰度图、马尔可夫单通道图(以 1.1 中的 $P_{i,j}$ 为例，简称单通道图)、马尔可夫彩色图(简称彩色图)，再进行归一化预处理后，输入到已加载 ImageNet 权重的 EfficientNet-B0 模型训练，实验相关超参数为：优化器采用 SGD，Momentum 设为 0.9，Weight_decay 设为 0.000 1，学习率采用余弦退火算法调整，初始值为 0.01，Epoch 设为 120，Batch_size 设为 32。实验结果如表 3 和 4 所示。

由表 3 和 4 可知，在 PDF 数据集和 DOCX 数据集上综合性能最好的组合均是彩色图＋EfficientNetB0＋迁移学习。其中，在 PDF 数据集上，该组合的准确率为 99.80%，比单通道图和灰度图的准确率分别高 1.49%、3.27%，在 DOCX 数据集上，该组合的准确率为 98.14%，比单通道图和灰度图的准确率分别高 1.74%、5.57%，并

表 3　结合迁移学习的 EfficientNet-B0

数据集	特征	准确率	F1-score	AUC
PDF	灰度图	0.965 3	0.965 2	0.964 4
	单通道图	0.983 1	0.983 0	0.983 8
	彩色图	**0.998 0**	**0.997 8**	**0.998 0**
DOCX	灰度图	0.925 7	0.924 2	0.918 0
	单通道图	0.964 0	0.968 7	0.969 8
	彩色图	**0.981 4**	**0.981 7**	**0.981 5**

表 4　随机初始化权重的 EfficientNet-B0

数据集	特征	准确率	F1-score	AUC
PDF	灰度图	0.949 8	0.949 6	0.949 2
	单通道图	0.976 0	0.969 8	0.970 1
	彩色图	**0.990 4**	**0.991 3**	**0.990 7**
DOCX	灰度图	0.911 2	0.909 8	0.908 4
	单通道图	0.940 5	0.947 8	0.946 6
	彩色图	**0.968 6**	**0.969 7**	**0.964 2**

且 F1-score 和 AUC 分别有不同程度的提升。这是由于，相较于由文档字节值直接转换为像素点值的灰度图，马尔可夫图的尺寸不受原始文档的大小影响，对图像的预处理，信息损失较少。且文档的结构是具有相似性的，内容是具有关联性的，而马尔可夫模型中一个状态也是受前 N 个状态的影响，因此，由马尔可夫模型转化得来的马尔可夫图具有能够更好的区分良性文档与恶意文档的特征。且将马尔可夫单通道图改进成三通道彩色图后，包含更多的字节序列信息，可视化图像后，深度学习模型能够挖掘出更深层的特征，并且 3 个通道的特征起到互补的作用。因此，使用彩色图的组合的综合检测能力最优。

以彩色图为例，在两个数据集上，准确率与 Epoch 次数的关系如图 7 和 8 所示，损失度与 Epoch 次数的关系如图 9 和 10 所示。

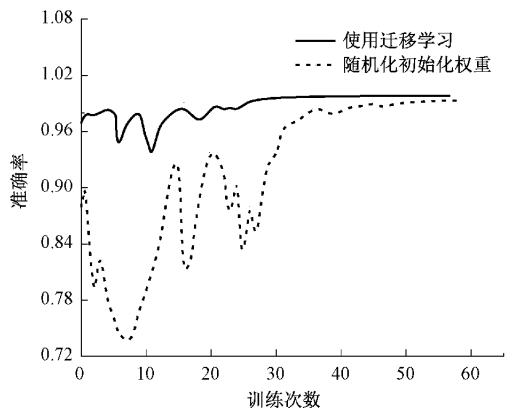

图 7　在 PDF 数据集上准确率变化情况

由图 7 可知，在 PDF 数据集上，采用随机初始化权重，EfficientNet-B0 在第 50 个 Epoch 时逐渐收敛，准确率稳定在 99.1% 左右，而使用迁移学习，EfficientNet-B0 在第 30 个 Epoch 时就开始收敛，准确率稳定在 99.8% 左右。从图 8 可知，在 DOCX 数据集上，采用随机初始化权重，EfficientNet-B0 在第 90 个 Epoch 时逐渐收敛，准确率稳定在 96.8% 左右。而使用迁移学习，EfficientNet-B0 在第 60 个 Epoch 时就开始收敛，准确率稳定在 98.1% 左右。由于 DOCX 数据集样本较少，对模型收敛速度和检测准确率有一定影响，但从实验结果可知，使用迁移学习可以加快模型的收敛速度和提升检测准确率。

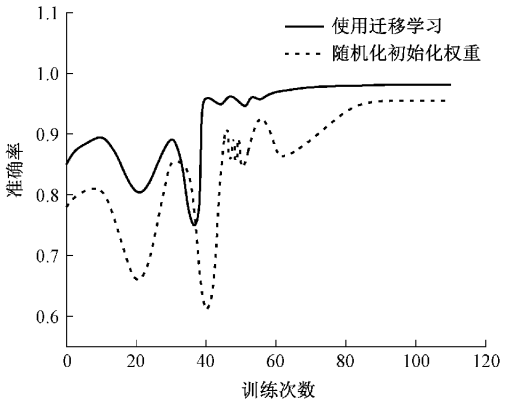

图 8　在 DOCX 数据集上准确率的变化情况

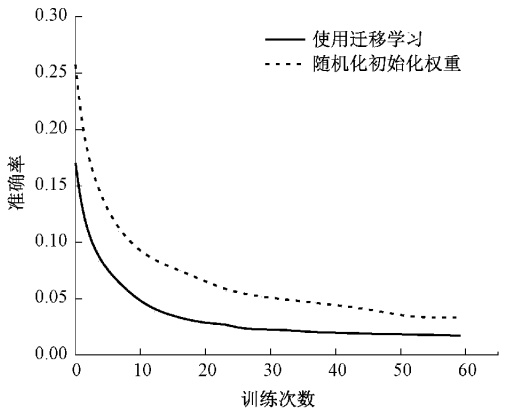

图 9　在 PDF 数据集上损失度变化情况

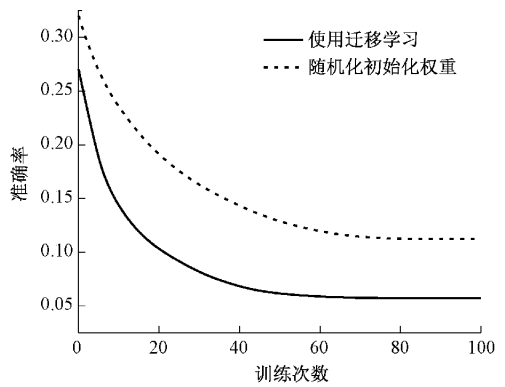

图 10　在 DOCX 数据集上损失度变化情况

由图 9 和 10 可知，在 PDF 数据集和 DOCX 数据集

上，使用迁移学习的 EfficientNet-B0 的损失度初始值比随机初始化权重的 EfficientNet-B0 分别低 0.08 和 0.05，且两条 Loss 曲线一直在随机初始化权重的 EfficientNet-B0 的下方，说明其拟合真实样本标签的速度更快，拟合程度更高。综上，EfficientNet-B0 能够很好的利用 ImageNet 上的权重，使模型快速适应文档图像数据集，达到更好的效果。

2）各模型实验对比结果

将 3 种文档图像分别输入使用迁移学习的 6 种模型，图 11 为各模型在 PDF 数据集上的实验结果，图 12 为各模型在 DOCX 数据集上的实验结果。在两个数据集上，在所有的深度学习模型上，均是彩色图的准确率评价指标最高，在 PDF 数据集上，彩色图准确率比灰度图和单通道图平均高 3.37%和 1.93%，在 DOCX 数据集上，彩色图准确率比灰度图和单通道图平均高 5.74%和 1.97%。进一步说明马尔可夫彩色图比灰度图和马尔可夫单通道图更能表征出恶意文档的深层特征。而在所有的模型中，均是 EfficientNet-B0 模型的各项指标最高。说明 EfficientNet-B0 比其他模型更合适恶意文档图像的分类工作。

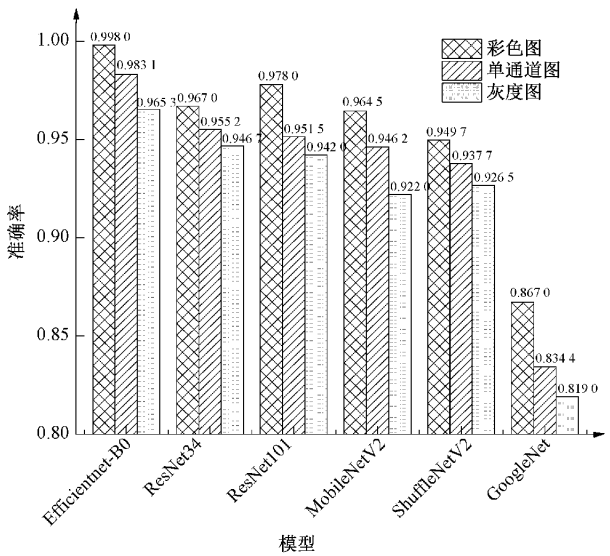

图 11　各模型在 PDF 数据集上实验结果

3）本文方法与现有工具和方法对比

为进一步验证所提出方法的有效性，本文比较了所提出方法与主流 PDF 文档检测工具 Wepawet 和 PJScan 的检测性能。在本文 PDF 数据集的测试集上的结果如表 5 所示。

本文方法的召回率（99.78%）明显高于 Wepawet（65.68%）、PJScan（82.84%）；在精确率方面，虽然 Wepawet（100%）高于本文方法（99.78%），但是其召回率只有 65.68%，明显低于本文方法（99.78%），其主要原因是 Wepawet 的检测方法比较特殊，只有当 PDF 文档产生恶意行为时，才认定为恶意，所以该方法检测精确率高，但

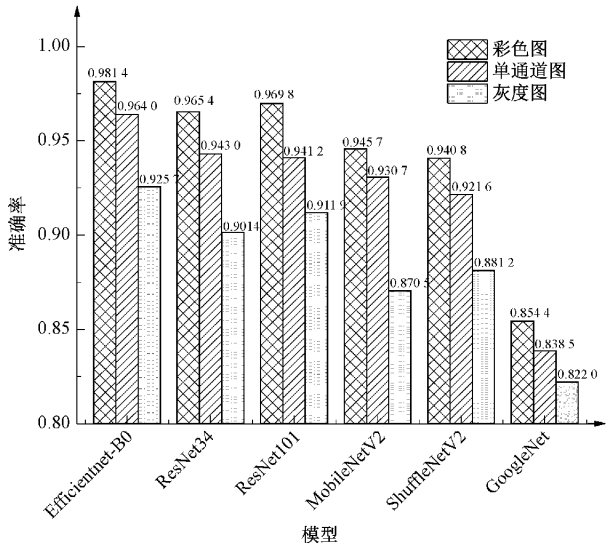

图 12　各模型在 DOCX 数据集上实验结果

表 5　各工具在 PDF 数据集上实验结果

方法	Acc	Recall	Precision
EfficientNet-B0＋彩色图＋迁移学习	**0.998 0**	**0.997 8**	**0.997 8**
Wepawet	0.838 2	0.656 8	1.000 0
PJScan	0.812 4	0.828 4	0.837 4

会造成较多的恶意 PDF 文档漏报。综上，本文方法的综合检测性能优于 Wepawet 和 PJScan。

3　结　　论

针对传统静动态检测方法具有特征工程复杂、资源花销大、不能快速有效地检测恶意文档等缺点，本文提出一种新的恶意文档可视化方法，利用结合迁移学习的 EfficientNet-B0 模型进行检测。实验证明，本文可视化方法优于恶意软件可视化领域常用的灰度图方法，且模型的综合检测能力优于主流恶意文档检测工具。

本文方法无需使用复杂的特征工程，且具有良好的检测能力，更适合大规模恶意文档检测。下一步将考虑扩充恶意文档数据集，以及对更多文档格式的研究，提高模型的检测准确率和普适性。

参考文献

[1] KOUTSOKOSTAS V, LYKOUSAS N, APOSTOLOPOULOS T, et al. Automated analysis of malicious Microsoft Office documents [J]. Computers & Security, 2021: 102582.

[2] 喻民，姜建国，李罡，等. 恶意文档检测研究综述[J]. 信息安全学报，2021，6(3)：54-76.

[3] ŠRNDIĆ N, LASKOV P. Hidost: a static machine-learning-based detector of malicious files [J].

EURASIP Journal on Information Security, 2016, 2016(1): 1-20.

[4] LI Y, WANG X, SHI Z, et al. Boosting training for PDF malware classifier via active learning [J]. International Journal of Intelligent Systems, 2021, 37(4): 2803-2821.

[5] LU X, WANG F, JIANG C, et al. A Universal Malicious Documents Static Detection Framework Based on Feature Generalization[J]. Applied Sciences, 2021, 11(24): 12134.

[6] XU M, KIM T. Platpal: Detecting malicious documents with platform diversity [C]. 26th {USENIX} Security Symposium ({USENIX} Security 17), 2017: 271-287.

[7] 杜学绘, 林杨东, 孙奕. 基于混合特征的恶意 PDF 文档检测[J]. 通信学报, 2019, 40(2): 118-128.

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

[9] 卢喜东, 段哲民, 钱叶魁, 等. 一种基于深度森林的恶意代码分类方法[J]. 软件学报, 2020, 31(5): 1454-1464.

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

[11] AP A, AML A, VP B, et al. Malware detection employed by visualization and deep neural network[J]. Computers & Security, 2021.

[12] 金泽宁, 朱正伟. 结合应用接口可达性特征的 Android 恶意软件检测[J]. 电子测量技术, 2021, 44(9):8.

[13] 刘小利. 基于深度学习算法的图像融合[J]. 国外电子测量技术, 2020, 39(7): 38-42.

[14] TAN M, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks [C]. International Conference on Machine Learning. PMLR, 2019: 6105-6114.

[15] PAN S J, YANG Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(10): 1345-1359.

[16] 李帷韬, 韩慧慧, 焦点, 等. 基于深度迁移学习的大雾等级智能认知方法研究[J]. 电子测量与仪器学报, 2020, 2: 88-96.

[17] RAMANNA S, SENGOZ C, KEHLER S, et al. Near real-time map building with multi-class image set labeling and classification of road conditions using convolutional neural networks[J]. Applied Artificial Intelligence, 2021: 1-31.

[18] YOUSFI Y, BUTORA J, FRIDRICH J, et al. Improving EfficientNet for JPEG steganalysis [C]. Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security, 2021: 149-157.

[19] 王宸, 唐禹, 张秀峰, 等. 基于改进 EfficientNet 的锻件磁粉探伤智能检测方法研究[J]. 仪器仪表学报, 2021, 42(9):8.

[20] 廖海斌, 徐斌. 基于性别和年龄因子分析的鲁棒性人脸表情识别[J]. 计算机研究与发展, 2021, 58(3): 528.

[21] TAN M, CHEN B, PANG R, et al. Mnasnet: Platform-aware neural architecture search for mobile [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 2820-2828.

作者简介

黄昆,硕士研究生,主要研究方向为应用安全、人工智能安全。

E-mail: 20010210627@gznu.edu.cn

徐洋(通信作者),博士,教授,主要研究方向为信息安全。

E-mail: xy@gznu.edu.cn

张思聪,博士,讲师,主要研究方向为信息安全、机器学习。

李克资,硕士研究生,主要研究方向为信息安全、机器学习。